LIVE PROJECTS IN BUSINESS ANALYTICS USING R

Edited and Co-authored by

DR SHEETAL MAHENDHER
Head, Department of Business Analytics
and Quantitative Techniques

ISBR Business School
REAL WORLD REAL LEARNING

International School of
Business & Research
Bangalore

Live Projects- Business Analytics Using R

Edited and Co- Authored by

Dr. Sheetal Mahendher

Head, Department of Business Analytics and Quantitative Techniques

Batch 2019-2021

Contents

Perception of Consumers Towards Digital Payment

Submitted By-
Abhinav Kumra (PG19003)
Kamalesh Kar (PG19064)
Akhil Sharma (PG19008)

Introduction

It has been said that every disruption creates opportunities and one such disruption was the announcement of demonetization by Prime Minister Mr. Narender Modi on 08 November 2016. Demonetization created huge growth opportunity for digital payment in India and the digital wallet companies garbed the opportunities with both the hands to expand their market share. Demonetization has presented a unique platform for adoption of digital payment, as an alternative to cash for Indian consumers.

Adoption of cashless transaction has been significantly pushed by Prime Minister Mr. Narender Modi as part of government reforms after demonetization of high value currency of Rs. 500 and 1000 (86% of cash circulation). The demonetization resulted in unprecedented growth in digital payment. By February this year, digital wallet companies had shown a growth of 271 percent for a total value of US$2.8 billion (Rs. 191 crores) [1], Indian government and private sector companies such as **Paytm**,**G-Pay**,**PayPa**l had been aggressively pushing several digital payment applications, including the Aadhaar Payment app, the **UPI app**, and the National Payments Corporation of India (NPCI) developed the Bharat Interface for Money (**BHIM**) app. Digital transfers using apps has brought behavioural change and helped in the adoption of digital payment. This has resulted in ease of transfer of money in rural areas which was not touched earlier by the digital payment method. Now many foreign investors want to invest in digital payment industry which is new attractive destinations because of scope of tremendous expansion in India.

The ongoing spread of **COVID-19** has become one of the biggest threats to the global economy

and financial markets. To contain the impact of the coronavirus outbreak, India, like many countries across the globe, is taking several measures, including a nationwide lockdown; limiting movement of the entire population; shutting down public places and transport; and urging the public to stay indoors, maintain social distance, and work from home. The resulting economic disruption is huge and the short- term decline in activity for businesses, both large and small, considerable. The contactless approach in order to protect from disease boost online payment method to remain touchless.

There are number of facilitators which are leading to the growth of digital payment and transition from cash economy to less cash economy. These facilitators include penetration of internet connectivity on smart phones, non-banking financial institution facilitating digital payment, one touch payment, rise of financial technology sector and push by government either by giving incentives or tax breaks.

Literature Review

Bamasak carried out study in Saudi Arabia found that there is a bright future for m-payment. Security of mobile payment transactions and the unauthorized use of mobile phones to make a payment were found to be of great concerns to the mobile phone users. Security and privacy were the major concerns for the consumers which affect the adoption of digital payment solutions. Doan illustrated the adoption of mobile wallet among consumers in Finland as only at the beginning stages of the Innovation-Decision Process.

Doing payments via mobile phones has been in use for many years and is now set to explode. Also, mobiles are increasingly being used by consumers for making payments. “Digital Wallet “has become a part of consumers which are nothing but smart phones which can function as leather wallets. Digital wallet offered many benefits while transferring money such as convenience, security and affordability. Growth in technology has opened many modes of payments through which consumers can do transactions which are more convenient, accessible and acceptable, consumers have an inclination towards mobile payment apps usage. Offering various benefits such as flexi payment digital wallet brands are providing extra convenience to consumers. Major factor in adoption of digital wallet is convenience in buying products online without physically going from one location to another location. There has been many studies conducted in past on mobile payment application to find consumer interest and they found consumer has positive inclination for the same.

The factors such as perceived ease of use, expressiveness and trust affect adoption of digital wallet as payment method. These factors are termed as facilitators and plays crucial role in adoption of digital payment solution. Usage of digital wallet among youth in the state of Punjab was found to be associated with societal influence and usefulness, controllability and security,

and need for performance enhancement. Premium pricing, complexity, a lack of critical mass, and perceived risks are the barriers to adoption of digital payment systems.

A comprehensive model ‘Payment Mode Influencing Consumer Purchase Model’ was proposed by Braga and Mazzon. This model considered factors such as temporal orientation

and separation, self-control and pain of payment constructs for digital wallet as a new payment mode. Consumer perspective of mobile payments and mobile payment technologies are two most important factors of mobile payments research. Mallat studied consumer adoption of mobile payments in Finland. Study found that mobile payment is dynamic and its adoption depends on lack of other payments methods and certain situational factors.

Digital wallet payments bring extra convenience to shoppers by offering flexible payment additions and accelerating exchanges. Shin and Ziderman tested a comprehensive model of consumer acceptance in the context of mobile payment. It used the unified theory of acceptance and use of technology (UTAUT) model with constructs of security, trust, social influence, and self-efficacy. The model confirmed the classical role of technology acceptance factors (i.e., perceived to users' attitude), the results also showed that users' attitudes and intentions are influenced by perceived security and trust. In the extended model, the moderating effects of demographics on the relations among the variables were found to be significant. Digital wallets offer the consumers the convenience of payments without swiping their debit or credit cards. Instant Cash availability and renders seamless mobility is also a unique feature of these digital apps, for instance the balance in your Paytm wallet can be very easily transferred to your bank account as and when you want. Following are some other advantages of making transactions through e wallets:

Saves time: digital wallets hold the amount in the electronic form so as to ease the payment process where users can make online payments without entering any card details.

Ease of use: As digital wallet is like one click pay without filling details about card viz card number and passwords every time, It allows user to link digital wallet to accounts and pay right away so that the consumers face no issues to enter the details every time a transaction happen.

Security: There is a good amount of security when payments are made through e wallets since the wallet does not pass the payment card details to the website. These virtual wallets allow users to lock their wallet.

Convenient and information stored under one roof: As digital wallets helps to eliminate need to carry the physical wallet they are highly convenient. Also a better management is possible as there is synchronization of data from multiple platforms like bank accounts, credit and debit cards, mobile accounts and billing portals.

Attractive discount: Cash back and discounts are being offered by most of the players along with providing offline wallet balance top up known as 'Cash Pickup' service. This service is being offered by Mobikwik that will facilitate cash to be directly added to MobiKwik wallet where consumers of even smaller towns can be benefited.

As per Ministry of Finance Report (December 2016) on Digital payment, financial inclusion is one of the foremost challenge facing India. 53 percent of India population had access to formal financial services. In this context, digital payment can act as accelerator to financial inclusion. Increasing availability of mobile phone, availability of data network infrastructure, rollout of 3G and 4G networks and large merchant eco system are the critical enablers of digital payment in

India. It is further supported by the coordinated efforts of industry, regulator and government. As per RBI's report 'Vision 2018' four pronged strategy focusing on regulation, robust infrastructure, effective supervisory mechanism and customer centricity has been adopted to push adoption of digital payment in India.

The percentage of cash for transactions has seen a rapid decline in the past few years in India. In 2010, the percentage of cash in all payments was 89% compared with 78% in 2015. This rapid decline is a result of an increased adoption of non-cash instruments such as cards and

digital payments such as mobile wallets, electronic transfers, etc. Stored value instruments like mobile wallets (Paytm, G-pay, BHIM, etc.) and prepaid and gift cards have made payments though internet devices convenient and easy. India represents one of the largest market opportunities for digital payments. With a population of 1.25 billion, India accounts for roughly 18% of the global population. The two key drivers of digital payments-mobile phones and internet users are already well established in India. To date, India has about 1.0 billion mobile phone subscribers and 300 million internet users, ranking 2nd on both metrics globally.

Objectives

The objective of the study was to find out the customer perception and impact of demographic factors on adoption of digital mode of payment:

Research Methodology

The current study is based on primary data collected from 108 respondents from the different parts of INDIA. A well-structured questionnaire was designed to collect the information from the respondents the questionnaire was designed to study perception of consumer towards adoption of digital payment mode. Likert five point scales were used for obtaining responses via social media platforms and E-mail as well.

Sampling Plan

Sampling unit: This call is for defining the target population to be surveyed. In this research the sampling unit was the customers who have been using the digital payment modes.

Sample size: In this survey the sample size decided was 108.

Sampling procedure: We adopted digital media platform for collection of primary data, as it is not possible to take appointment from a large number of respondents. Purpose of this research was told to respondents and questions were explained to them in Google form for understanding any particular question. There had been no personal bias or distortions were allowed while recording the responses.

Research and Statistical Tools Employed

The research and statistical tools employed in this study are model preparation, sensitivity analysis then confusion matrix and finally accuracy of our model to see weather our model is satisfying the hypothesis or not. Confusion Matrix were used to visualize important predictive analysis recall, specificity, accuracy. Confusion Matrices are useful because they give direct

comparison of values like True positives, False Positives, True Negatives and False Negatives. Then we find out accuracy of our model to conclude the output of our observation.

Analysis: -

We have named our dataset as "digi"

There are two levels of our dependent variable (Profession) i.e Yes=1 and No= 2

1. CLEANING DATA: To clean data, we need to remove all missing values i.e. NAs or Nas present in the data.

   Code used: **complete.cases(...)** => Return a logical vector indicating which cases are complete, i.e., have no missing values (Na or NaN)

   Usage of the code: **table(complete.cases(digi))**

⇨ In our data, we used this code to know whether there any missing values in our data that should be further removed. The False values indicate number of rows with missing values and True values indicate number of the rows which don't have missing values. We used this code and got the result as TRUE – 200 and FALSE – 0 which means our data doesn't contain any missing value i.e. NA or NaN values. We proceed further.

2. **CONVERTING INTO FACTORS**: The next step of our analysis was to convert all the variables we had into factors.

   Code used: **as. factor(x)** => which is used to convert the data type of a variable to a factor variable. The function factor is used to encode a vector as a factor. If argument ordered is TRUE, the factor levels are assumed to be ordered. For compatibility with S there is also a function ordered. as. factor is one of coercion functions for these classes.

   Usage of the code in our analysis: **digi$Paytm = as. factor(digi$Paytm)** (same for all variables)

⇨ This code was used on all the variables of our data which were in characters and numeric so as to convert all of them into categorical variable for further analysis. In our data, "Paytm" was numeric and other variables like 'Profession', 'GPay', 'Bhim', 'Bank own app', 'East ro use', 'only cash' , 'Purpose', 'PayPal', 'Phone Pe', 'User Friendly' were having character data type.

3. **MODEL BUILDING**: We build models to predict the value of the dependent variable for independent variables for whom some is available and to estimate the effect of some independent variables on the dependent variable.

   Code used: **glm (x~…., data =, family = binomial ())** => x is dependent variable and stands for independent variables. glm is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution. AIC is considered as the parameter to select the best model out of all. When we run "summary (model name), we get AIC and other information related to that model. The model with least AIC is considered as best for further analysis. If incase the selected model does not give a good

accuracy and validation, we need to work more on that model or look for other closest one.

Usage of code in our analysis =>

· **model3 <- glm(formula = Profession ~ Age + PayPal + PhonePe + Bhim + make.life.easy.. +**

Time.Savings. + epidemic.increased.the.demand + offer.wide.range. + traditional.methods., family = binomial(), data = digi)

· **summary(model3)**

· In our analysis, we used the model building code for logistic regression and built seven models and out of all those model 3 has the least AIC i.e. 36.

· We worked with all three models and got the best results from model 3. So, we considered model 7 as the best model.

4. **PREDICTION**: We need to make predictions for our data so as to work on Predictive Analysis. We want to know the future impact of our independent variables on our dependent variable for which predictions are necessary to make.

Code used: **predict(object..)** => Predict is a generic function for predictions from the results of various model fitting functions. The function invokes particular methods which depend on the class of the first argument. Object is a model for which prediction is required and is the additional arguments affecting the predictions produced.

Head(x$y)=> Returns the first or last parts of a vector, matrix, table, data frame or function. x is the name of the dataset and y is the name of column storing predicted values.

Table(x$z) => table uses the cross-classifying factors to build a contingency table of the counts at each combination of factor levels. X is the name of dataset and z is the dependent variable

ifelse(test, yes, no)=> ifelse returns a value with the same shape as test which is filled with elements selected from either yes or no depending on whether the element of test is TRUE or FALSE.

Usage of code in our data ->

digi$pred=predict (model3, type="response")

head(digi$pred)

table(digi$Profession)

digi$pred=ifelse(digi$pred>0.5,"Y","N")

· In our analysis, we made predictions on model7 using above mentioned codes. With the help of predict code we got a new column called ‘pred’ in our dataset with all the predicted values for all variables. Further head code gave us certain values that helped in estimating one threshold

cutoff that came out to be 0.5. Using table code, we got to know that the responses for "Harmful" carries 99 Yes and 101 No. Using ifelse, code, we divided the predicted values into two categories 'Y' and 'N'. The predicted values more than 0.5 were categorized into 'Y' and values less than 0.5 were stored as 'N' into the 'pred' column. This was done to make confusion matrix and further analysis more convenient.

5. **CONFUSION MATRIX**: a **confusion matrix is** a very useful tool for calibrating the output of a model and examining all possible outcomes of your predictions (true positive, true negative, false positive, false negative). It helps in ascertaining the performance of our model

   Code used: **confusionMatrix(actual, predicted, cutoff = 0.5)** => Creates a confusion matrix given a specific cutoff.

   Usage of code in our analysis =>

   con. matrix=confusionMatrix(digi$Profession,digi$pred)

   con. matrix

   · In our analysis, the values are □

   True Negative = 89 (states that actual was "N" and predicted came out to be "N" as well)

   False Negative= 0 (states that actual was "Y" and predicted came out to be "N") False Positive=0(states that actual was "N" and predicted came out to be "Y") True Positive = 18 (states that actual "Y" and predicted came out to be "Y" as well) (where Y= Yes and N=No)

· we named our confusion matrix as con.matrix and got the results of sensitivity (TP/(TP+FN)) = 1 and specificity (TN/(TN+FP)) = 1 which depicted that our model 3 is good as both there parameters should be more than 0.50 for the model to be good, we are getting good values. The accuracy of our model on the other hand came out to be 100% which is good. This means that are model 3 is performing good in our analysis.

6. AREA OF UNDER CURVE (AUC) OF RECEIVER OPERATOR CHARACTERISTIC (ROC): **AUC - ROC curve** is a performance measurement for classification problem at various thresholds settings. **ROC** is a probability **curve** and **AUC** represent degree or measure of separability. It tells how much model is capable of distinguishing between classes. An **AUC** of 0.5 suggests no discrimination, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.

   An **ROC curve** shows the relationship between clinical sensitivity and specificity for every possible cut-off. The **ROC curve** is a **graph** with: The x-axis showing 1 – specificity i.e. false positive rate. The y-axis showing sensitivity i.e. true positive rate.

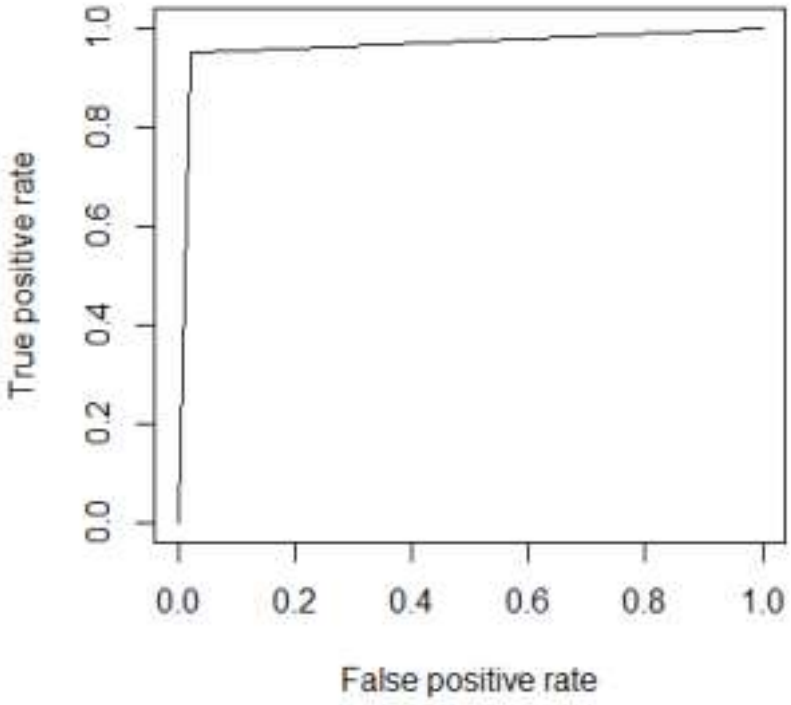
   Usage of codes in our analysis =>

   i) For roc curve

   pred = prediction (as. numeric(digi$pred),as.numeric(digi$Profession))

roc.pred=performance(pred,measure = "tpr",x.measure = "fpr")

plot(roc.pred)



ii) For auc value

auc=performance (pred,measure = "auc")

auc@y.values[{1}]

⇨ In our analysis, we used prediction – performance codes to determine roc curve which turned out to be towards y axis i.e. tpr (true positive rate). This means that our roc curve is good. The area covered by roc i.e area under cover value came out to be 1. This means that our model shows no discrimination and is close to acceptable as it is 1.

Result & Discussion

Profile of Respondents

The respondent profile as observed replicate the professional population generally engaged in use of digital payment. Most of the respondents are male (72%), mostly are graduates (90%) and in the age group of 22-35 years (100%) or 31-40 years (28%). Their annual income is Rs. 1 to 5 Lacs (33%). This is the ideal profile for user of digital mode and who are educated, employed and having decent income.

Conclusion

Present study has made an attempt to understand customer perception regarding digital payment. It was found that Majority of professional respondents agree that mobile wallet/digital payment provides benefits to individual for purchase of products, improve the quality of decision, helpful in buying products as compared to traditional methods, they offer a wide range of banking services and payment options. They also agree that interaction with mobile wallet is helpful and that they trust the service providers.

Impact of Covid 19 on Digital Entertainment

Submitted By-
Aman Hans (PG19162)
Akshita Sharma (PG19011)
Pragya Chhibber (PG19088)

Abstract

During this whole pandemic, government of India has taken various unprecedented and drastic measures to curb the spread of novel Coronavirus (COVID-19). These measures include imposing lockdown in the whole country, prohibitions under Section-144 and imposing various guidelines and government advisory on social distancing. The novel coronavirus 2019 currently designated as COVID-19 is an infectious disease caused by a newly discovered corona virus. The rapid spread of new corona viruses throughout China and the world in 2019–2020 has had a great impact on social development like the entertainment industries where various activities like movies and sporty activities are being suspended all over the world. Due to whole this situation, people need their entertainment so they have moved on to the digital platforms like Netflix, Amazon prime video and Hotstar or many other platforms, they found these platforms more convenient and pocket friendly. This industry has flourished with the time, but in this situation, this has been flourishing with a great impact.

Introduction

- The COVID-19 pandemic has changed the way people consume media and entertainment. Due to strict national lockdowns around the world people have been forced to stay at home, changing consumer behaviour on a large scale. As movie theatres, museums, events, and other external entertainment consumption models have been banned, social lives have moved online, and entertainment consumption has increased significantly for online gaming and over-the-top (OTT) services.

- Traditional media services such as television and newspapers have also been side lined as drastic cuts in ad spends of large companies have severely dented revenues of traditional media giants. Even government advertising has taken a hit post the pandemic. To a large extent viewership has been limited to consumers looking for live news updates about the coronavirus.
- In contrast, services like Hotstar, Amazon Prime and Netflix in India have seen an 82.63% increase in time spent. Similarly, YouTube has seen a 20.5 percent surge in subscribers in the country. It garnered over 300 billion views in the first quarter of 2020 and has been growing at a rate of 13 percent since the fourth quarter of 2019

Objective of the Research

In the research we have a dependent variable "Harmful" and other independent variable. We have done the research how our dependent variable is impacted from other independent variables and which independent variable impacts our dependent variable the most. We have a question "Is the OTT platforms harmful or not?" So, we have done the research according to that.

Analysis

• **Predictive Analysis**:

We have named our dataset as "OTT"

There are two levels of our dependent variable (Harmful) i.e Yes=1 and No= 2

1. CLEANING DATA: To clean data, we need to remove all missing values i.e. NAs or NaNs present in the data.

   Code used: **complete.cases(...)** => Return a logical vector indicating which cases are complete, i.e., have no missing values (Na or NaN)

   Usage of the code: **table(complete.cases(ott))**

⇨ In our data, we used this code to know whether there any missing values in our data that should be further removed. The False values indicate number of rows with missing values and True values indicate number of the rows which don't have missing values. We used this code and got the result as TRUE – 200 and FALSE – 0 which means our data doesn't contain any missing value i.e. NA or NaN values. We proceed further.

2. CONVERTING INTO FACTORS: The next step of our analysis was to convert all the variables we had into factors.

   Code used: **as. factor(x)** => which is used to convert the data type of a variable to a factor variable. The function factor is used to encode a vector as a factor. If argument ordered is TRUE, the factor levels are assumed to be ordered. For compatibility with S there is also a function ordered. as. factor is one of coercion functions for these classes.

Usage of the code in our analysis: **ott$Age = as. factor(ott$Age)** (same for all variables)

- This code was used on all the variables of our data which were in characters and numeric so as to convert all of them into categorical variable for further analysis. In our data, "Age" was numeric and other variables like 'Profession', 'Binge Watch', 'Platform', 'Pocket Friendly', 'Digital Preference', 'Theatres Preference' , 'Show enjoyment', 'Online Stream', 'Preference Post Pandemic', 'User Friendly', 'Child Friendly' ,'Parental Control', 'Measures and locks' and 'Harmful' were having character data type.

3. CORRELATION (r) : A correlation between two variables measures the strength and direction of their linear relationship. The value of r lies between -1 to 1. The more it is closest to 1 the high the correlation is and vice versa. If the value of r is in positive, this shows that with increase in one variable the other variable also increases and if it is negative means with increase in one variable, other variable decreases.

   Code used: **cor.test (x, …)** => Test for association between paired samples, using one of Pearson's product moment correlation coefficient, Kendall's tau or Spearman's rho. X stand for a numeric vector of data values.

   Usage of code in our analysis: **cor.test (as. numeric(ott$Harmful), (ott$Age))** (same for all variables)

   - In our analysis, we correlated all the independent variables with that of our dependent variable i.e. harmful. Correlation requires the data type of variables to be numeric only.

   Null Hypothesis – No correlation between two variables

   Alternate Hypothesis – correlation between two variables exist.
   If the p value is greater than 0.05, we reject null hypothesis and accept the alternate hypothesis which means that the correlation between two variables exist.

   The result we for correlation between:

   - Harmful and age => p value = 0.666 i.e. 0.05 . So, correlation exists. r is 0.0306, less positive correlation of 3.065% between harmful and Age, as one increases, other also increases

- Harmful and Binge. Watch => p value = 0.905 i.e. > 0.05. So, correlation exists. r is 0.0085, less positive corr of 0.85% between harmful and binge. Watch, as one increases, other also increases
- Harmful and Profession => p value = 0.8015 i.e > 0.05 . So, correlation exists. r is 0.0179, less positive coll of 1.79% between harmful and profession, as one increases, other also increases
- Harmful and Platform => p value = 0.2704 i.e > 0.05 . So, correlation exists. r is 0.0783, less positive coll of 7.83% between harmful and platform, an one increases, other also increases
- Harmful and Pocket Friendly => p value = 0.6727 i.e > 0.05 . So, correlation exists. r is 0.0301, less positive coll of 3.01% between harmful and Pocket friendly, an one increases, other also increases
- Harmful and Digital preference => p value = 0.1705 i.e > 0.05 . So, correlation exists. r is -0.0973, less negative coll of -9.73% between harmful and Digital Preference, an one increases, other decreases
- Harmful and theatre preference => p value = 0.791 i.e > 0.05 . So, correlation exists. r is 0.0189, less positive coll of 1.89% between harmful and Theaters preference, an one increases, other also increases
- Harmful and show enjoyment => p value = 0.6464 i.e > 0.05 . So, correlation exists. r is 0.3264, less positive coll of 32.64% between harmful and Show enjoyment, an one increases, other also increases
- Harmful and Online Stream => p value = 0.5204 i.e > 0.05 . So, correlation exists. r is 0.0457, less positive coll of 4.57% between harmful and Online stream, an one increases, other also increases
- Harmful and Preference post pandemic => p value = 0.09371 i.e > 0.05 . So, correlation exists. r is 0.1188, less positive coll of 11.88% between harmful and preference post pandemic, an one increases, other also increases
- Harmful and User friendly => p value = 0.2475 i.e > 0.05 . So, correlation exists. r is -0.0821, less negative coll of -8.21%% between harmful and user friendly, an one increases, other decreases
- Harmful and child friendly => p value = 0.1479 i.e > 0.05 . So, correlation exists. r is 0.1027, less positive coll of 10.27% between harmful and child friendly, an one increases, other also increases
- Harmful and Parental control => p value = 0.8606 i.e > 0.05 . So, correlation

exists. r is 0.0125, less positive coll of 1.25% between harmful and parental control, and one increases, other also increases

- Harmful and measures and locks => p value = 0.4801 i.e > 0.05 . So, correlation exists. r is -0.0502, less negative coll of -5.02% between harmful and measures and locks, an one increases, other decreases.

4. MODEL BUILDING: We build models to predict the value of the dependent variable for independent variables for whom some is available and to estimate the effect of some independent variables on the dependent variable.

Code used: **glm (x~…., data =, family = binomial ())** => x is dependent variable and stands for independent variables. glm is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution. AIC is considered as the parameter to select the best model out of all. When we run "summary (model name), we get AIC and other information related to that model. The model with least AIC is considered as best for further analysis. If incase the selected model does not give a good accuracy and validation, we need to work more on that model or look for other closest one.

Usage of code in our analysis =>

- **model7 <- glm(Harmful~Binge.Watch + Pocket.friendly, data=ott, family=binomial())**
- **summary(model7)** (same for other models as well)

⇨ In our analysis, we used the model building code for logistic regression and built seven models and out of all those model 4 has the least AIC i.e 262.48 . After that, model3 has an AIC of 262.84 and then model7 with AIC of 262.93.

⇨ We worked with all three models and got the best results from model 7. So, we considered model 7 as the best model

5. PREDICTION: We need to make predictions for our data so as to work on Predictive Analysis. We want to know the future impact of our independent variables on our dependent variable for which predictions are necessary to make.

Code used: **predict(object..)** => Predict is a generic function for predictions from the results of various model fitting functions. The function invokes particular methods which depend on the class of the first argument. Object is a model for which prediction

is required and is the additional arguments affecting the predictions produced.

Head(x$y)=> Returns the first or last parts of a vector, matrix, table, data frame or function. x is the name of the dataset and y is the name of column storing predicted values.

Table(x$z) => table uses the cross-classifying factors to build a contingency table of the counts at each combination of factor levels. X is the name of dataset and z is the dependent variable

ifelse(test, yes, no)=> ifelse returns a value with the same shape as test which is filled with elements selected from either yes or no depending on whether the element of test is TRUE or FALSE.

Usage of code in our data ->

ott$pred=predict (model7, type="response")

head(ott$pred)

table(ott$Harmful)
ott$pred=ifelse(ott$pred>0.5,"Y","N")

⇨ In our analysis, we made predictions on model7 using above mentioned codes. With the help of predict code we got a new column called 'pred' in our dataset with all the predicted values for all variables. Further head code gave us certain values that helped in estimating one threshold cutoff that came out to be 0.5. Using table code, we got to know that the responses for "Harmful" carries 99 Yes and 101 No. Using ifelse, code, we divided the predicted values into two categories 'Y' and 'N'. The predicted values more than 0.5 were categorized into 'Y' and values less than 0.5 were stored as 'N' into the 'pred' column. This was done to make confusion matrix and further analysis more convenient.

6. CONFUSION MATRIX: a **confusion matrix is** a very useful tool for calibrating the output of a model and examining all possible outcomes of your predictions (true positive, true negative, false positive, false negative). It helps in ascertaining the performance of our model

Code used: **confusionMatrix(actual, predicted, cutoff = 0.5)** => Creates a confusion matrix given a specific cutoff.

Usage of code in our analysis =>

con. matrix=confusionMatrix(ott$Harmful,ott$pred)

con. matrix

⇨ In our analysis, the values are ->

True Negative = 62 (states that actual was “N” and predicted came out to be “N” as well)

False Negative= 29 (statesthat actual was “Y” and predicted came out to be “N”)

False Positive=37 (statesthat actual was “N” and predicted came out to be “Y”)

True Positive = 72 (statesthat actual “Y” and predicted came out to be “Y” as well) (where Y= Yes and N=No)

⇨ we named our confusion matrix as con.matrix and got the results of sensitivity (TP/(TP+FN)) = 0.713 and specificity (TN/(TN+FP)) = 0.626 which depicted that our model 7 is good as both there parameters should be more than 0.50 for the model to be good, we are getting good values. The accuracy of our model on the other hand came out to be 62% which is also quite good. This means that are model 7 is performing good in our analysis.

7. AREA OF UNDER CURVE (AUC) OF RECEIVER OPERATOR CHARACTERISTIC (ROC): **AUC - ROC curve** is a performance measurement for classification problem at various thresholds settings. **ROC** is a probability **curve** and **AUC** represent degree or measure of separability. It tells how much model is capable of distinguishing between classes. An **AUC** of 0.5 suggests no discrimination, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent,
and more than 0.9 is considered outstanding.

An **ROC curve** shows the relationship between clinical sensitivity and specificity for every possible cut-off. The **ROC curve** is a **graph** with: The x-axis showing 1 – specificity i.e. false positive rate. The y-axis showing sensitivity i.e. true positive rate.

Usage of codes in our analysis =>

i) For roc curve

```
pred = prediction (as.

numeric(ott$pred),as.numeric(ott$Harmful))

roc.pred=performance(pred,measure = "tpr",x.measure = "fpr")

plot(roc.pred)
```

ii) For auc value

```
auc=performance (pred,measure = "auc")

auc@y.values[{1}]
```

⇨ In our analysis, we used prediction – performance codes to determine roc curve which turned out to be tilted towards y axis i.e. tpr (true positive rate). This means that our roc curve is good. The area covered by roc i.e area under cover value came out to be 0.67. This means that our model shows no discrimination and is close to acceptable as it is above 0.50 and close to 0.70.

Conclusion

Our Model 7 is best with most accuracy. The AUC being 67% i.e. decent. This means that the digital platforms being harmful will be mostly impacted by Binge watching by the consumers and platforms being Pocket friendly for them.

Behaviour of Consumers Towards Online Shopping

Submitted By-
Anit Dhillon (PG19017)
Garima (PG19048)
Nicole Fernandes (PG19081)

Introduction

This live project was conducted to study the behaviour of consumers towards online shopping.. The apparel buying behaviour of Indian consumers through five dimensions viz. consumer characteristics, reference groups, store attributes, promotion and product attributes. The results show that the store attributes promotion and reference groups are the important dimensions of apparel buying behaviour. The demographic aspects namely occupation of the consumer and social class of the consumer has no effect on the consumer buying behaviour in choosing private label brands. The purpose of this study is to understand the consumer perspective towards online shopping, their liking, disliking, and satisfaction level.

Objective of the Study

- To build a regression model in order to find out which independent variables influence the dependent variable ie. frequency of online shopping and to what extent.
- To understand the consumer pattern and the awareness among the consumers regarding the e-commerce platform, to analyse the factors influencing online shopping, to find about the variety of products purchased by the customers through online shopping.

Research Methodology

In this research, random sampling method was used, where a survey was conducted among 200 respondents. In order to carry out the survey, we used an online questionnaire to collect the responses of the sample and this data was used for further analysis. The questionnaire included

open and close ended questions. Only primary data collected from the survey was used to arrive at conclusions at the end of the study. A descriptive and quantitative research was carried in order to analyse the results from the questionnaire. We used R Analytics as a tool to carry out a detailed analysis from the data obtained.

```
getwd()

## [1] "C:/Users/DORSLYN FERNANDES/Desktop/nicole/R Classes"

shopp=read.csv("C:/Users/DORSLYN FERNANDES/Desktop/nicole/R Classes/shop.csv")
View(shopp)
str(shopp)

## 'data.frame':    200 obs. of  21 variables:
##  $ Age.                                                : int  58 56 26 59 17 55 59 20 25 24 ...
##  $ Gender.                                             : chr  "Female" "Female" "Male" "Female" ...
##  $ Occupation.                                         : chr  "Employed" "Employed" "Student"
"Business" ...
##  $ Income..per.month..                                 : int  20000 120000 90000 50000 50000 50000
500000 100000 140000 20000 ...
##  $ Device.Used                                         : chr  "Personal Computer (Website)" "Smart
Phone (Application)" "Smart Phone (Application)" "Personal Computer (Website)" ...
##  $ Motivation.Saves.time                               : chr  "Agree" "Agree" "Neutral" "Agree" ...
##  $ Motivation.Broad.variety.of.goods                   : chr  "Agree" "Agree" "Agree"
"Disagree" ...
##  $ Motivation.Best.price.with.difference.schemes       : chr  "Neutral" "Highly Agree"
"Agree" "Disagree" ...
##  $ Motivation.Some.products.are.not.available.in.retail.store  : chr  "Agree" "Highly Agree"
"Agree" "Disagree" ...
##  $ Motivation.Home.Delivery                            : chr  "Neutral" "Highly Agree" "Agree"
"Neutral" ...
##  $ Do.you.check.the.reviews.of.a.product.              : chr  "No" "Yes" "Yes" "Yes" ...
##  $ Problem.Encountered.Delay.in.Delivery               : chr  "Sometimes" "Very Often"
"Sometimes" "Rarely" ...
##  $ Problem.Encountered.Quality                         : chr  "Very Often" "Always" "Rarely"
"Sometimes" ...
##  $ Problem.Encountered.Product.Damage                  : chr  "Sometimes" "Sometimes"
"Never" "Rarely" ...
##  $ Problem.Encountered.Payment.not.successful          : chr  "Sometimes" "Rarely"
"Never" "Sometimes" ...
##  $ Problem.Encountered.Difference.between.displayed.or.delivered.product: chr  "Very Often"
"Very Often" "Rarely" "Rarely" ...
##  $ Mode.of.payment                                     : chr  "Cash on Delivery" "Cash on Delivery"
"Debit/Credit Card" "Cash on Delivery" ...
##  $ Website                                             : chr  "Flipkart" "Amazon" "Myntra" "Decathlon" ...
##  $ Product.shopped.for.the.most                        : chr  "Footwear" "Apparels" "Apparels"
"Apparels" ...
##  $ Checking.the.offline.shops.before.online.purchase   : chr  "Yes" "Yes" "No" "Yes" ...
```

```
##  $ Frequency.of.Online.Shopping.                    : chr  "Occasionally/On the basis of
Requirement" "Occasionally/On the basis of Requirement" "Routine" "Routine" ...
```

From the output, it was observed that all the data was in integer or characters. Therefore, the data needed to be converted to factors since most of the data was in a categorical format having levels.

Converting into Factors

```
shopp$Gender.=as.factor(shopp$Gender.)
shopp$Occupation.=as.factor(shopp$Occupation.)
shopp$Device.Used=as.factor(shopp$Device.Used)
shopp$Motivation.Saves.time=as.factor(shopp$Motivation.Saves.time)
shopp$Motivation.Broad.variety.of.goods=as.factor(shopp$Motivation.Broad.variety.of.goods)
shopp$Motivation.Best.price.with.difference.schemes=as.factor(shopp$Motivation.Best.price.with.
difference.schemes)
shopp$Motivation.Some.products.are.not.available.in.retail.store=as.factor(shopp$Motivation.Som
e.products.are.not.available.in.retail.store)
shopp$Motivation.Home.Delivery=as.factor(shopp$Motivation.Home.Delivery)
shopp$Do.you.check.the.reviews.of.a.product.=as.factor(shopp$Do.you.check.the.reviews.of.a.prod
uct.)
shopp$Problem.Encountered.Delay.in.Delivery=as.factor(shopp$Problem.Encountered.Delay.in.Deli
very)
shopp$Problem.Encountered.Quality=as.factor(shopp$Problem.Encountered.Quality)
shopp$Problem.Encountered.Product.Damage=as.factor(shopp$Problem.Encountered.Product.Dam
age)
shopp$Problem.Encountered.Payment.not.successful=as.factor(shopp$Problem.Encountered.Paym
ent.not.successful)
shopp$Problem.Encountered.Difference.between.displayed.or.delivered.product=as.factor(shopp$P
roblem.Encountered.Difference.between.displayed.or.delivered.product)
shopp$Mode.of.payment=as.factor(shopp$Mode.of.payment)
shopp$Website=as.factor(shopp$Website)
shopp$Product.shopped.for.the.most=as.factor(shopp$Product.shopped.for.the.most)
shopp$Checking.the.offline.shops.before.online.purchase=as.factor(shopp$Checking.the.offline.sho
ps.before.online.purchase)
shopp$Frequency.of.Online.Shopping.=as.factor(shopp$Frequency.of.Online.Shopping.)

str(shopp)

## 'data.frame':    200 obs. of  21 variables:
##  $ Age.                                          : int  58 56 26 59 17 55 59 20 25 24 ...
##  $ Gender.                                         : Factor w/ 2 levels "Female","Male": 1 1 2 1 2 1 1 2
2 1 ...
##  $ Occupation.                                       : Factor w/ 4 levels "Business","Employed",..: 2 2 4
1 4 1 1 1 1 1 ...
##  $ Income..per.month..                                    : int  20000 120000 90000 50000 50000 50000
500000 100000 140000 20000 ...
##  $ Device.Used                                       : Factor w/ 2 levels "Personal Computer
(Website)",..: 1 2 2 1 2 1 2 2 1 1 ...
```

```
## $ Motivation.Saves.time                                  : Factor w/ 5 levels "Agree","Disagree",..: 1 1
5 1 1 1 2 5 2 1 ...
## $ Motivation.Broad.variety.of.goods                      : Factor w/ 5 levels "Agree","Disagree",..:
1 1 1 2 3 2 2 5 2 1 ...
## $ Motivation.Best.price.with.difference.schemes          : Factor w/ 5 levels
"Agree","Disagree",..: 5 3 1 2 3 2 2 1 2 1 ...
## $ Motivation.Some.products.are.not.available.in.retail.store : Factor w/ 5 levels
"Agree","Disagree",..: 1 3 1 2 3 5 5 2 2 5 ...
## $ Motivation.Home.Delivery                               : Factor w/ 5 levels "Agree","Disagree",..: 5
3 1 5 3 2 5 5 5 1 ...
## $ Do.you.check.the.reviews.of.a.product.                 : Factor w/ 2 levels "No","Yes": 1 2 2 2
2 2 1 2 1 2 ...
## $ Problem.Encountered.Delay.in.Delivery                  : Factor w/ 5 levels
"Always","Never",..: 4 5 4 3 3 5 5 4 3 5 ...
## $ Problem.Encountered.Quality                            : Factor w/ 5 levels "Always","Never",..: 5
1 3 4 4 3 4 3 5 4 ...
## $ Problem.Encountered.Product.Damage                     : Factor w/ 5 levels
"Always","Never",..: 4 4 2 3 3 4 5 4 3 5 ...
## $ Problem.Encountered.Payment.not.successful             : Factor w/ 5 levels
"Always","Never",..: 4 3 2 4 2 3 5 4 3 5 ...
## $ Problem.Encountered.Difference.between.displayed.or.delivered.product: Factor w/ 5 levels
"Always","Never",..: 5 5 3 3 4 5 4 1 5 4 ...
## $ Mode.of.payment                                        : Factor w/ 4 levels "Cash on Delivery",..: 1 1 2
1 2 2 2 1 2 4 ...
## $ Website                                                : Factor w/ 7 levels "Amazon","Decathlon",..: 4 1 5 2
4 3 4 1 5 5 ...
## $ Product.shopped.for.the.most                           : Factor w/ 8 levels "Accessories",..: 7 2 2
2 3 3 2 2 3 8 ...
## $ Checking.the.offline.shops.before.online.purchase      : Factor w/ 2 levels "No","Yes": 2
2 1 2 2 2 2 2 2 1 ...
## $ Frequency.of.Online.Shopping.                          : Factor w/ 2 levels "Occasionally/On the
basis of Requirement",..: 1 1 2 2 2 2 1 2 1 1 ...
```

It is observed that the data is converted into factors and further analysis can be performed

accurately.

```
View(shopp)
```

Building Models

Frequency of online shopping is the dependent variable and the remaining variables are taken as independent variables. Factors regarding online shopping like it is time saving, variety of products available, home delivery, problems encountered during a purchase, shopping website, etc. play an influence on a person's buying behavior on an online site. All these factors determine whether a person makes a purchase online or not or determines how frequently he shops online.

We use **logistic regression** to build the model since the dependent variable is categorical has two levels.

```
mod1=glm(Frequency.of.Online.Shopping.~.,data=shopp, family=binomial())
summary(mod1)

step(mod1)
```

This code automatically adds relevant variables to the model and builds a strong model with the lowest AIC as seen below.

```
mod1=glm(formula = Frequency.of.Online.Shopping. ~ Motivation.Saves.time +
        Motivation.Broad.variety.of.goods + Motivation.Best.price.with.difference.schemes +
        Motivation.Some.products.are.not.available.in.retail.store +
        Motivation.Home.Delivery + Do.you.check.the.reviews.of.a.product. +
        Problem.Encountered.Delay.in.Delivery + Problem.Encountered.Product.Damage +
        Problem.Encountered.Payment.not.successful + Mode.of.payment +
        Product.shopped.for.the.most + Checking.the.offline.shops.before.online.purchase,
      family = binomial(), data = shopp)
summary(mod1)

##
## Call:
## glm(formula = Frequency.of.Online.Shopping. ~ Motivation.Saves.time +
##     Motivation.Broad.variety.of.goods + Motivation.Best.price.with.difference.schemes +
##     Motivation.Some.products.are.not.available.in.retail.store +
##     Motivation.Home.Delivery + Do.you.check.the.reviews.of.a.product. +
##     Problem.Encountered.Delay.in.Delivery + Problem.Encountered.Product.Damage +
##     Problem.Encountered.Payment.not.successful + Mode.of.payment +
##     Product.shopped.for.the.most + Checking.the.offline.shops.before.online.purchase,
##     family = binomial(), data = shopp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2791  -0.6893  -0.2257   0.5696   2.4996
##
## Coefficients: (1 not defined because of singularities)
##                                                  Estimate
## (Intercept)                                      -18.16980
## Motivation.Saves.timeDisagree                     -4.72152
## Motivation.Saves.timeHighly Agree                  1.03589
## Motivation.Saves.timeHighly Disagree              16.75843
## Motivation.Saves.timeNeutral                      -0.07733
## Motivation.Broad.variety.of.goodsDisagree          1.11060
## Motivation.Broad.variety.of.goodsHighly Agree      2.32184
## Motivation.Broad.variety.of.goodsHighly Disagree  -35.34535
## Motivation.Broad.variety.of.goodsNeutral           1.10581
## Motivation.Best.price.with.difference.schemesDisagree      3.14893
## Motivation.Best.price.with.difference.schemesHighly Agree -0.67082
## Motivation.Best.price.with.difference.schemesHighly Disagree 57.97335
## Motivation.Best.price.with.difference.schemesNeutral       0.20573
```

```
## Motivation.Some.products.are.not.available.in.retail.storeDisagree         -0.94440
## Motivation.Some.products.are.not.available.in.retail.storeHighly Agree      1.63384
## Motivation.Some.products.are.not.available.in.retail.storeHighly Disagree   4.39477
## Motivation.Some.products.are.not.available.in.retail.storeNeutral          -0.23763
## Motivation.Home.DeliveryDisagree                                            1.77579
## Motivation.Home.DeliveryHighly Agree                                       -1.70126
## Motivation.Home.DeliveryHighly Disagree                                    -2.82968
## Motivation.Home.DeliveryNeutral                                            -1.01344
## Do.you.check.the.reviews.of.a.product.Yes                                  -2.38062
## Problem.Encountered.Delay.in.DeliveryNever                                 -3.96947
## Problem.Encountered.Delay.in.DeliveryRarely                                -4.97583
## Problem.Encountered.Delay.in.DeliverySometimes                             -3.53212
## Problem.Encountered.Delay.in.DeliveryVery Often                            -5.37398
## Problem.Encountered.Product.DamageNever                                    22.85057
## Problem.Encountered.Product.DamageRarely                                   22.72881
## Problem.Encountered.Product.DamageSometimes                                21.74989
## Problem.Encountered.Product.DamageVery Often                               23.36100
## Problem.Encountered.Payment.not.successfulNever                             1.04113
## Problem.Encountered.Payment.not.successfulRarely                           -1.12839
## Problem.Encountered.Payment.not.successfulSometimes                        1.73632
## Problem.Encountered.Payment.not.successfulVery Often                           NA
## Mode.of.paymentDebit/Credit Card                                           0.36632
## Mode.of.paymente-Wallet                                                   -3.71517
## Mode.of.paymentNet Banking                                                 0.60292
## Product.shopped.for.the.mostApparels                                       0.80592
## Product.shopped.for.the.mostBooks                                          1.57740
## Product.shopped.for.the.mostConsumer Durables                             20.28231
## Product.shopped.for.the.mostCosmetics                                     -0.23160
## Product.shopped.for.the.mostElectronics                                    0.77518
## Product.shopped.for.the.mostFootwear                                      -0.61886
## Product.shopped.for.the.mostHome Appliances                               -0.25033
## Checking.the.offline.shops.before.online.purchaseYes                      -0.80737
##                                                          Std. Error
## (Intercept)                                            3956.18072
## Motivation.Saves.timeDisagree                             2.14551
## Motivation.Saves.timeHighly Agree                         0.76050
## Motivation.Saves.timeHighly Disagree                   3956.18060
## Motivation.Saves.timeNeutral                              0.78676
## Motivation.Broad.variety.of.goodsDisagree                 1.28960
## Motivation.Broad.variety.of.goodsHighly Agree             0.93183
## Motivation.Broad.variety.of.goodsHighly Disagree       4534.96373
## Motivation.Broad.variety.of.goodsNeutral                  0.64515
## Motivation.Best.price.with.difference.schemesDisagree     1.64369
## Motivation.Best.price.with.difference.schemesHighly Agree 0.83708
## Motivation.Best.price.with.difference.schemesHighly Disagree 6018.07815
## Motivation.Best.price.with.difference.schemesNeutral      0.57701
## Motivation.Some.products.are.not.available.in.retail.storeDisagree 1.05287
## Motivation.Some.products.are.not.available.in.retail.storeHighly Agree 0.90073
## Motivation.Some.products.are.not.available.in.retail.storeHighly Disagree 1.24430
## Motivation.Some.products.are.not.available.in.retail.storeNeutral 0.73598
## Motivation.Home.DeliveryDisagree                          1.87498
```

```
## Motivation.Home.DeliveryHighly Agree                                0.73160
## Motivation.Home.DeliveryHighly Disagree                          4534.96405
## Motivation.Home.DeliveryNeutral                                     0.95836
## Do.you.check.the.reviews.of.a.product.Yes                           0.91756
## Problem.Encountered.Delay.in.DeliveryNever                       5594.88470
## Problem.Encountered.Delay.in.DeliveryRarely                      5594.88467
## Problem.Encountered.Delay.in.DeliverySometimes                   5594.88463
## Problem.Encountered.Delay.in.DeliveryVery Often                  5594.88442
## Problem.Encountered.Product.DamageNever                          3956.18088
## Problem.Encountered.Product.DamageRarely                         3956.18090
## Problem.Encountered.Product.DamageSometimes                      3956.18091
## Problem.Encountered.Product.DamageVery Often                     3956.18077
## Problem.Encountered.Payment.not.successfulNever                     1.28801
## Problem.Encountered.Payment.not.successfulRarely                    1.29447
## Problem.Encountered.Payment.not.successfulSometimes                1.28005
## Problem.Encountered.Payment.not.successfulVery Often                    NA
## Mode.of.paymentDebit/Credit Card                                    0.57882
## Mode.of.paymente-Wallet                                             1.44073
## Mode.of.paymentNet Banking                                          0.77412
## Product.shopped.for.the.mostApparels                                0.68023
## Product.shopped.for.the.mostBooks                                   0.81133
## Product.shopped.for.the.mostConsumer Durables                    1600.11400
## Product.shopped.for.the.mostCosmetics                               0.92452
## Product.shopped.for.the.mostElectronics                             0.67621
## Product.shopped.for.the.mostFootwear                                1.14304
## Product.shopped.for.the.mostHome Appliances                         0.87436
## Checking.the.offline.shops.before.online.purchaseYes                0.53850
##                                                                      z value
## (Intercept)                                                           -0.005
## Motivation.Saves.timeDisagree                                         -2.201
## Motivation.Saves.timeHighly Agree                                      1.362
## Motivation.Saves.timeHighly Disagree                                   0.004
## Motivation.Saves.timeNeutral                                          -0.098
## Motivation.Broad.variety.of.goodsDisagree                              0.861
## Motivation.Broad.variety.of.goodsHighly Agree                          2.492
## Motivation.Broad.variety.of.goodsHighly Disagree                      -0.008
## Motivation.Broad.variety.of.goodsNeutral                               1.714
## Motivation.Best.price.with.difference.schemesDisagree                  1.916
## Motivation.Best.price.with.difference.schemesHighly Agree             -0.801
## Motivation.Best.price.with.difference.schemesHighly Disagree           0.010
## Motivation.Best.price.with.difference.schemesNeutral                   0.357
## Motivation.Some.products.are.not.available.in.retail.storeDisagree    -0.897
## Motivation.Some.products.are.not.available.in.retail.storeHighly Agree  1.814
## Motivation.Some.products.are.not.available.in.retail.storeHighly Disagree 3.532
## Motivation.Some.products.are.not.available.in.retail.storeNeutral     -0.323
## Motivation.Home.DeliveryDisagree                                       0.947
## Motivation.Home.DeliveryHighly Agree                                  -2.325
## Motivation.Home.DeliveryHighly Disagree                               -0.001
## Motivation.Home.DeliveryNeutral                                       -1.057
## Do.you.check.the.reviews.of.a.product.Yes                             -2.595
## Problem.Encountered.Delay.in.DeliveryNever                            -0.001
```

```
## Problem.Encountered.Delay.in.DeliveryRarely                          -0.001
## Problem.Encountered.Delay.in.DeliverySometimes                        -0.001
## Problem.Encountered.Delay.in.DeliveryVery Often                       -0.001
## Problem.Encountered.Product.DamageNever                               0.006
## Problem.Encountered.Product.DamageRarely                              0.006
## Problem.Encountered.Product.DamageSometimes                           0.005
## Problem.Encountered.Product.DamageVery Often                          0.006
## Problem.Encountered.Payment.not.successfulNever                       0.808
## Problem.Encountered.Payment.not.successfulRarely                      -0.872
## Problem.Encountered.Payment.not.successfulSometimes                   1.356
## Problem.Encountered.Payment.not.successfulVery Often                  NA
## Mode.of.paymentDebit/Credit Card                                     0.633
## Mode.of.paymente-Wallet                                              -2.579
## Mode.of.paymentNet Banking                                           0.779
## Product.shopped.for.the.mostApparels                                 1.185
## Product.shopped.for.the.mostBooks                                    1.944
## Product.shopped.for.the.mostConsumer Durables                        0.013
## Product.shopped.for.the.mostCosmetics                                -0.251
## Product.shopped.for.the.mostElectronics                              1.146
## Product.shopped.for.the.mostFootwear                                 -0.541
## Product.shopped.for.the.mostHome Appliances                          -0.286
## Checking.the.offline.shops.before.online.purchaseYes                 -1.499
##                                                                      Pr(>|z|)
## (Intercept)                                                          0.996336
## Motivation.Saves.timeDisagree                                        0.027761
## Motivation.Saves.timeHighly Agree                                    0.173158
## Motivation.Saves.timeHighly Disagree                                 0.996620
## Motivation.Saves.timeNeutral                                         0.921699
## Motivation.Broad.variety.of.goodsDisagree                            0.389130
## Motivation.Broad.variety.of.goodsHighly Agree                        0.012714
## Motivation.Broad.variety.of.goodsHighly Disagree                     0.993781
## Motivation.Broad.variety.of.goodsNeutral                             0.086522
## Motivation.Best.price.with.difference.schemesDisagree                0.055395
## Motivation.Best.price.with.difference.schemesHighly Agree            0.422911
## Motivation.Best.price.with.difference.schemesHighly Disagree         0.992314
## Motivation.Best.price.with.difference.schemesNeutral                 0.721435
## Motivation.Some.products.are.not.available.in.retail.storeDisagree   0.369731
## Motivation.Some.products.are.not.available.in.retail.storeHighly Agree  0.069693
## Motivation.Some.products.are.not.available.in.retail.storeHighly Disagree 0.000413
## Motivation.Some.products.are.not.available.in.retail.storeNeutral     0.746792
## Motivation.Home.DeliveryDisagree                                     0.343588
## Motivation.Home.DeliveryHighly Agree                                 0.020050
## Motivation.Home.DeliveryHighly Disagree                              0.999502
## Motivation.Home.DeliveryNeutral                                      0.290299
## Do.you.check.the.reviews.of.a.product.Yes                            0.009472
## Problem.Encountered.Delay.in.DeliveryNever                           0.999434
## Problem.Encountered.Delay.in.DeliveryRarely                          0.999290
## Problem.Encountered.Delay.in.DeliverySometimes                       0.999496
## Problem.Encountered.Delay.in.DeliveryVery Often                      0.999234
## Problem.Encountered.Product.DamageNever                              0.995392
## Problem.Encountered.Product.DamageRarely                             0.995416
```

```
## Problem.Encountered.Product.DamageSometimes                    0.995613
## Problem.Encountered.Product.DamageVery Often                   0.995289
## Problem.Encountered.Payment.not.successfulNever                0.418902
## Problem.Encountered.Payment.not.successfulRarely               0.383370
## Problem.Encountered.Payment.not.successfulSometimes            0.174957
## Problem.Encountered.Payment.not.successfulVery Often                 NA
## Mode.of.paymentDebit/Credit Card                              0.526817
## Mode.of.paymente-Wallet                                       0.009918
## Mode.of.paymentNet Banking                                    0.436067
## Product.shopped.for.the.mostApparels                          0.236109
## Product.shopped.for.the.mostBooks                             0.051869
## Product.shopped.for.the.mostConsumer Durables                 0.989887
## Product.shopped.for.the.mostCosmetics                         0.802194
## Product.shopped.for.the.mostElectronics                       0.251647
## Product.shopped.for.the.mostFootwear                          0.588221
## Product.shopped.for.the.mostHome Appliances                   0.774645
## Checking.the.offline.shops.before.online.purchaseYes          0.133796
##
## (Intercept)
## Motivation.Saves.timeDisagree                                 *
## Motivation.Saves.timeHighly Agree
## Motivation.Saves.timeHighly Disagree
## Motivation.Saves.timeNeutral
## Motivation.Broad.variety.of.goodsDisagree
## Motivation.Broad.variety.of.goodsHighly Agree                 *
## Motivation.Broad.variety.of.goodsHighly Disagree
## Motivation.Broad.variety.of.goodsNeutral                      .
## Motivation.Best.price.with.difference.schemesDisagree         .
## Motivation.Best.price.with.difference.schemesHighly Agree
## Motivation.Best.price.with.difference.schemesHighly Disagree
## Motivation.Best.price.with.difference.schemesNeutral
## Motivation.Some.products.are.not.available.in.retail.storeDisagree
## Motivation.Some.products.are.not.available.in.retail.storeHighly Agree    .
## Motivation.Some.products.are.not.available.in.retail.storeHighly Disagree ***
## Motivation.Some.products.are.not.available.in.retail.storeNeutral
## Motivation.Home.DeliveryDisagree
## Motivation.Home.DeliveryHighly Agree                          *
## Motivation.Home.DeliveryHighly Disagree
## Motivation.Home.DeliveryNeutral
## Do.you.check.the.reviews.of.a.product.Yes                     **
## Problem.Encountered.Delay.in.DeliveryNever
## Problem.Encountered.Delay.in.DeliveryRarely
## Problem.Encountered.Delay.in.DeliverySometimes
## Problem.Encountered.Delay.in.DeliveryVery Often
## Problem.Encountered.Product.DamageNever
## Problem.Encountered.Product.DamageRarely
## Problem.Encountered.Product.DamageSometimes
## Problem.Encountered.Product.DamageVery Often
## Problem.Encountered.Payment.not.successfulNever
## Problem.Encountered.Payment.not.successfulRarely
## Problem.Encountered.Payment.not.successfulSometimes
```

```
## Problem.Encountered.Payment.not.successfulVery Often
## Mode.of.paymentDebit/Credit Card
## Mode.of.paymente-Wallet                                    **
## Mode.of.paymentNet Banking
## Product.shopped.for.the.mostApparels
## Product.shopped.for.the.mostBooks                           .
## Product.shopped.for.the.mostConsumer Durables
## Product.shopped.for.the.mostCosmetics
## Product.shopped.for.the.mostElectronics
## Product.shopped.for.the.mostFootwear
## Product.shopped.for.the.mostHome Appliances
## Checking.the.offline.shops.before.online.purchaseYes
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.37  on 199  degrees of freedom
## Residual deviance: 165.26  on 156  degrees of freedom
## AIC: 253.26
##
## Number of Fisher Scoring iterations: 16
```

Result □ AIC: 253.26

This is the **best model** as it gives us the lowest AIC:

```
mod1=glm(formula = Frequency.of.Online.Shopping. ~ Motivation.Saves.time
+Motivation.Broad.variety.of.goods + Motivation.Best.price.with.difference.schemes +
Motivation.Some.products.are.not.available.in.retail.store + Motivation.Home.Delivery +
Do.you.check.the.reviews.of.a.product. + Problem.Encountered.Delay.in.Delivery +
Problem.Encountered.Product.Damage + Problem.Encountered.Payment.not.successful +
Mode.of.payment + Product.shopped.for.the.most
+Checking.the.offline.shops.before.online.purchase, family = binomial(), data = shopp)
```

Prediction □

```
shopp$pred=predict(mod1,type = "response")
table(shopp$Frequency.of.Online.Shopping.)

##
## Occasionally/On the basis of Requirement
##                                     121
##                                 Routine
##                                      79

head(shopp$pred)

## [1] 0.389446084 0.002369736 0.711782252 0.775410548 0.723848139 0.836325563

View(shopp)
```

If predicted value > 0.5 then take it as Routine, otherwise, Occasionally/On the basis of Requirement

```
shopp$pred=ifelse(shopp$pred > 0.5,"Routine","Occasionally/On the basis of Requirement")
table(shopp$pred)

##
## Occasionally/On the basis of Requirement
##                                     127
##                                 Routine
##                                      73

str(shopp)

## 'data.frame':    200 obs. of  22 variables:
##  $ Age.                                                : int  58 56 26 59 17 55 59 20 25 24 ...
##  $ Gender.                                             : Factor w/ 2 levels "Female","Male": 1 1 2 1 2 1 1 2
2 1 ...
##  $ Occupation.                                         : Factor w/ 4 levels "Business","Employed",..: 2 2 4
1 4 1 1 1 1 1 ...
##  $ Income..per.month..                                 : int  20000 120000 90000 50000 50000 50000
500000 100000 140000 20000 ...
##  $ Device.Used                                         : Factor w/ 2 levels "Personal Computer
(Website)",..: 1 2 2 1 2 1 2 2 1 1 ...
##  $ Motivation.Saves.time                               : Factor w/ 5 levels "Agree","Disagree",..: 1 1
5 1 1 1 2 5 2 1 ...
##  $ Motivation.Broad.variety.of.goods                   : Factor w/ 5 levels "Agree","Disagree",..:
1 1 1 2 3 2 2 5 2 1 ...
##  $ Motivation.Best.price.with.difference.schemes       : Factor w/ 5 levels
"Agree","Disagree",..: 5 3 1 2 3 2 2 1 2 1 ...
##  $ Motivation.Some.products.are.not.available.in.retail.store  : Factor w/ 5 levels
"Agree","Disagree",..: 1 3 1 2 3 5 5 2 2 5 ...
##  $ Motivation.Home.Delivery                            : Factor w/ 5 levels "Agree","Disagree",..: 5
3 1 5 3 2 5 5 5 1 ...
##  $ Do.you.check.the.reviews.of.a.product.              : Factor w/ 2 levels "No","Yes": 1 2 2 2
2 2 1 2 1 2 ...
##  $ Problem.Encountered.Delay.in.Delivery               : Factor w/ 5 levels
"Always","Never",..: 4 5 4 3 3 5 5 4 3 5 ...
##  $ Problem.Encountered.Quality                         : Factor w/ 5 levels "Always","Never",..: 5
1 3 4 4 3 4 3 5 4 ...
##  $ Problem.Encountered.Product.Damage                  : Factor w/ 5 levels
"Always","Never",..: 4 4 2 3 3 4 5 4 3 5 ...
##  $ Problem.Encountered.Payment.not.successful          : Factor w/ 5 levels
"Always","Never",..: 4 3 2 4 2 3 5 4 3 5 ...
##  $ Problem.Encountered.Difference.between.displayed.or.delivered.product: Factor w/ 5 levels
"Always","Never",..: 5 5 3 3 4 5 4 1 5 4 ...
##  $ Mode.of.payment                                     : Factor w/ 4 levels "Cash on Delivery",..: 1 1 2
1 2 2 2 1 2 4 ...
##  $ Website                                             : Factor w/ 7 levels "Amazon","Decathlon",..: 4 1 5 2
4 3 4 1 5 5 ...
##  $ Product.shopped.for.the.most                        : Factor w/ 8 levels "Accessories",..: 7 2 2
```

```
2 3 3 2 2 3 8 ...
##  $ Checking.the.offline.shops.before.online.purchase               : Factor w/ 2 levels "No","Yes": 2
2 1 2 2 2 2 2 1 ...
##  $ Frequency.of.Online.Shopping.                                   : Factor w/ 2 levels "Occasionally/On the
basis of Requirement",..: 1 1 2 2 2 2 1 2 1 1 ...
##  $ pred                                                            : chr  "Occasionally/On the basis of Requirement"
"Occasionally/On the basis of Requirement" "Routine" "Routine" ...
```

As we can see, the predicted variables are in character which needs to be converted in factors for further analysis.

```
shopp$pred=as.factor(shopp$pred)
str(shopp)
```

```
## 'data.frame':    200 obs. of  22 variables:
##  $ Age.                                                            : int  58 56 26 59 17 55 59 20 25 24 ...
##  $ Gender.                                                         : Factor w/ 2 levels "Female","Male": 1 1 2 1 2 1 1 2
2 1 ...
##  $ Occupation.                                                     : Factor w/ 4 levels "Business","Employed",..: 2 2 4
1 4 1 1 1 1 1 ...
##  $ Income..per.month..                                             : int  20000 120000 90000 50000 50000 50000
500000 100000 140000 20000 ...
##  $ Device.Used                                                     : Factor w/ 2 levels "Personal Computer
(Website)",..: 1 2 2 1 2 1 2 2 1 1 ...
##  $ Motivation.Saves.time                                           : Factor w/ 5 levels "Agree","Disagree",..: 1 1
5 1 1 1 2 5 2 1 ...
##  $ Motivation.Broad.variety.of.goods                               : Factor w/ 5 levels "Agree","Disagree",..:
1 1 1 2 3 2 2 5 2 1 ...
##  $ Motivation.Best.price.with.difference.schemes                   : Factor w/ 5 levels
"Agree","Disagree",..: 5 3 1 2 3 2 2 1 2 1 ...
##  $ Motivation.Some.products.are.not.available.in.retail.store      : Factor w/ 5 levels
"Agree","Disagree",..: 1 3 1 2 3 5 5 2 2 5 ...
##  $ Motivation.Home.Delivery                                        : Factor w/ 5 levels "Agree","Disagree",..: 5
3 1 5 3 2 5 5 5 1 ...
##  $ Do.you.check.the.reviews.of.a.product.                          : Factor w/ 2 levels "No","Yes": 1 2 2 2
2 2 1 2 1 2 ...
##  $ Problem.Encountered.Delay.in.Delivery                           : Factor w/ 5 levels
"Always","Never",..: 4 5 4 3 3 5 5 4 3 5 ...
##  $ Problem.Encountered.Quality                                     : Factor w/ 5 levels "Always","Never",..: 5
1 3 4 4 3 4 3 5 4 ...
##  $ Problem.Encountered.Product.Damage                              : Factor w/ 5 levels
"Always","Never",..: 4 4 2 3 3 4 5 4 3 5 ...
##  $ Problem.Encountered.Payment.not.successful                      : Factor w/ 5 levels
"Always","Never",..: 4 3 2 4 2 3 5 4 3 5 ...
##  $ Problem.Encountered.Difference.between.displayed.or.delivered.product: Factor w/ 5 levels
"Always","Never",..: 5 5 3 3 4 5 4 1 5 4 ...
##  $ Mode.of.payment                                                 : Factor w/ 4 levels "Cash on Delivery",..: 1 1 2
1 2 2 2 1 2 4 ...
##  $ Website                                                         : Factor w/ 7 levels "Amazon","Decathlon",..: 4 1 5 2
4 3 4 1 5 5 ...
##  $ Product.shopped.for.the.most                                    : Factor w/ 8 levels "Accessories",..: 7 2 2
2 3 3 2 2 3 8 ...
```

```
##  $ Checking.the.offline.shops.before.online.purchase : Factor w/ 2 levels "No","Yes": 2
2 1 2 2 2 2 2 2 1 ...
##  $ Frequency.of.Online.Shopping.                     : Factor w/ 2 levels "Occasionally/On the
basis of Requirement",..: 1 1 2 2 2 2 1 2 1 1 ...
##  $ pred                                              : Factor w/ 2 levels "Occasionally/On the basis of
Requirement",..: 1 1 2 2 2 2 1 1 1 1 ...
```

Confusion Matrix □

We used two methods to arrive at the confusion matrix

#1

```
table(shopp$Frequency.of.Online.Shopping.,shopp$pred)
```

```
##
##                                    Occasionally/On the basis of Requirement
##   Occasionally/On the basis of Requirement                              106
##   Routine                                                                21
##
##                                    Routine
##   Occasionally/On the basis of Requirement     15
##   Routine                                      58
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(lattice)
library(ggplot2)
```

#2

```
cm1=confusionMatrix(shopp$Frequency.of.Online.Shopping.,shopp$pred)
cm1
```

```
## Confusion Matrix and Statistics
##
##                                Reference
## Prediction                           Occasionally/On the basis of Requirement
##   Occasionally/On the basis of Requirement                              106
##   Routine                                                                21
##                                Reference
## Prediction                           Routine
##   Occasionally/On the basis of Requirement     15
##   Routine                                      58
##
##               Accuracy : 0.82
```

```
##               95% CI : (0.7596, 0.8706)
##    No Information Rate : 0.635
##    P-Value [Acc > NIR] : 8.113e-09
##
##                  Kappa : 0.6184
##
##  Mcnemar's Test P-Value : 0.4047
##
##             Sensitivity : 0.8346
##             Specificity : 0.7945
##          Pos Pred Value : 0.8760
##          Neg Pred Value : 0.7342
##              Prevalence : 0.6350
##          Detection Rate : 0.5300
##    Detection Prevalence : 0.6050
##       Balanced Accuracy : 0.8146
##
##       'Positive' Class : Occasionally/On the basis of Requirement
##
```

Accuracy of the Model : 0.82 = 82%

True Positive=106, True Negative=58, False Positive=15, False Negative=21

Sensitivity = TP/(TP+FN)

```
sensi = 106/(106+21)
sensi
```

```
## [1] 0.8346457
```

Hence there is 83.46% sensitivity

Specificity = TN/(TN+FP)

```
speci = 58/(58+15)
speci
```

```
## [1] 0.7945205
```

Hence there is 79.45% specificity

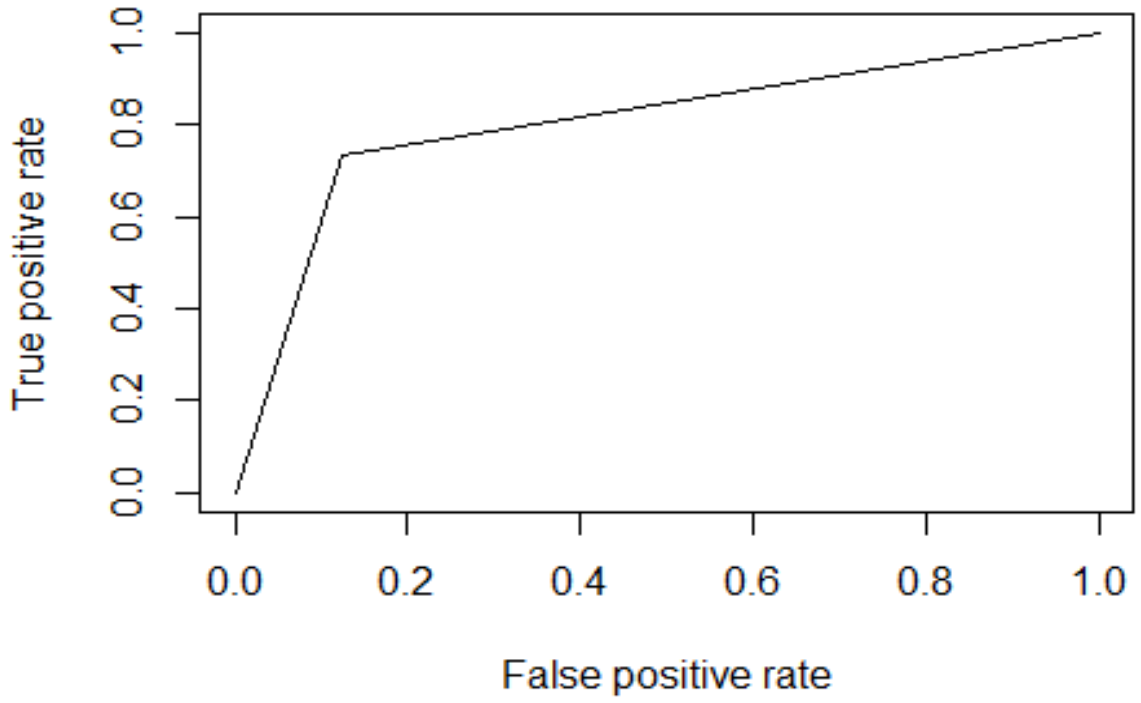
ROCR Curve and AUC (Area Under Curve) □

ROC determines the accuracy of a classification model at a user defined threshold value. It determines the model's accuracy using Area Under Curve (AUC). The area under the curve (AUC), also referred to as index of accuracy (A) or concordant index, represents the performance of the ROC curve. Higher the area, better the model. ROC is plotted between True Positive Rate (Y axis) and False Positive Rate (X Axis).

```
library(ROCR)
pred=prediction(as.numeric(shopp$pred),as.numeric(shopp$Frequency.of.Online.Shopping.))
roc.pred=performance(pred,measure='tpr',x.measure='fpr')
```

tpr is true positive rate and fpr is false positive rate

```
plot(roc.pred)
```



It can be observed that the graph is tilting towards the Y-axis which is a good sign as it shows that the forecasts of the model ie. the predicted values are accurate to a great extent.

```
auc=performance(pred,measure ="auc")
auc@y.values[{1}]

## [[1]]
## [1] 0.8051051
```

AUC of the model is 80.51%. Hence, the accuracy of the model is good.

Conclusion:

We have built a logistic regression model where ‘Frequency of online shopping’ is the dependent variable. The frequency of shopping online by the consumers is highly influenced by the following independent variables:

- Time – saving
- Broad variety of goods
- Best price with difference schemes
- Products not available offline
- Home delivery of the product
- Reviews of a product
- Problems encountered during the purchase like delayed delivery, product damage and unsuccessful payments
- Mode of payment
- Products shopped for the most
- Checking offline stores before making an online purchase

Therefore, the regression model built helps us predict that a person is an occasional shopper or routine shopper based on the above mentioned independent variables. All these factors have a good correlation with the ‘frequency of online shopping’ and are highly significant factors in predicting whether a consumer is a regular or not.

Predictive Modelling to Prescribe Optimal Sleep Patterns for Ensuring Higher Productivity Levels

Submitted By-
Anjali Ambekar (PG19018)
Rohith M S (PG19103)
Toshith Sastry (PG19140)
Yashus G (PG19151)

Abstract:

This study is focused on the collection of data to determine the correlation, between sleep patterns and the productive capabilities of individual and develop a prescriptive model which explores the relationship for expected sleep patterns based on required levels of productivity on various controllable personal characteristics. Through our work, we aim to create recommendation model which can be used to classify individuals as per their respective levels of productivity, to use the model to prescribe the necessary standards of control variables to ensure the an individual achieves the desired levels of their productivity.

Introduction:

This study has been evoked, with the simple idea of enhancing human genius and productivity, for those who wish to tap into their most optimum capacity for a sustainably enhanced performance. A comprehensive field of study has been generated by gathering primary data as a source of evidence to conclude upon. The main purpose of gathering this initial primary evidence is to determine whether various sleep variables have a direct result on the productivity of the individual concerning their work requirements and overall performance. Studies have proven to show that sleep does have various physiological impacts ranging from mental health to immunity, muscle building to improved cognition.

People who can effectively control their sleep cycle essentially command the power to structure their life best required to face their external situations. As a result of which an enhanced understanding of their own physical and mental faculties are pertinent. This research will facilitate in making human beings capable and foster room for holistic growth and development. This specific research is focused on the elimination of any unrequited awakenings and its impact on productivity levels, to recommend to our diverse population if their sleep patterns are ill advised against their best interests. As many people suffering from regular sleep loss are not fully aware of it, and many do not realize that they are victims of lack of sleep and continue to remain blind as to what it will cost.

Objectives:

1. It focuses on the resolution of a pertinent aspect of human life. Aim at reducing the mismanagement of sleep deprivation, which can be associated with shortening span of attention, slower motor functions, higher reaction time, memory loss, extended periods of information processing.
2. Empowers better decision-making which can negate degenerative diseases such as Dementia & Alzheimer in advanced stages with improved lifestyles. With more instability in the matters of sleep, individuals demonstrate worsening performance despite their best efforts for which they may claim indifference or ignorance towards the outcomes of their sub-standard performance levels.
3. On the other hand, more than required sleep invokes lethargy and laziness and as a result, also tends to negatively impact productivity.
4. Thus, we aim to optimize a state which can be achieved by just the right amount of sleep requirements and consistently sustained sleep cycles.

Methodology

To do complete justice to the categorical and numerical variable dataset which has been collected through the Primary Data Collection method, a Factor Analysis has been performed in order to categorize variables based on similarity of their characteristics, which are then grouped into two main clusters entailing:

a) High Productivity Levels

b) Low Productivity Levels

K-means clustering algorithm has been used in this study to classify the data points into the aforementioned groups based on a variety of individual characteristics. This enables us to efficiently analyze the grouped significant variables which must be considered in order to facilitate the creation of an effective recommendation based on a logistical regression model. The obtained results have been analyzed to determine the accuracy of the classification model using various evaluation metrics.

Data Analysis & Model building

```
dat=read.csv("C:/Users/admin/Desktop/R
Project/SleepDataPrepared(revised).csv") dat2=dat
#str(dat)
#View(dat)
#Converting categorical variables in to factors
dat$Gender=as.factor(dat$Gender)
dat$Age.Group=as.factor(dat$Age.Group)
dat$Profession=as.factor(dat$Profession)
#dat$Activity=as.factor(dat$Activity)
#dat$Workout=as.factor(dat$Workout)

dat_pca=prcomp(dat[,c(-1,-2,-3,-8,-10,-11)],center = TRUE,
scale. = TRUE) #removing all the categorical variables
#Removing the dependant variables to perform factor analysis

summary(dat_pca)

## Importance of components:
## PC1 PC2 PC3 PC4 PC5 ## Standard deviation 1.3629
1.2464  0.9564  0.65417  0.49625  ##  Proportion  of
Variance 0.3715 0.3107 0.1830 0.08559 0.04925 ##
Cumulative Proportion 0.3715 0.6822 0.8652 0.95075
1.00000

plot(dat_pca,type="l")
```

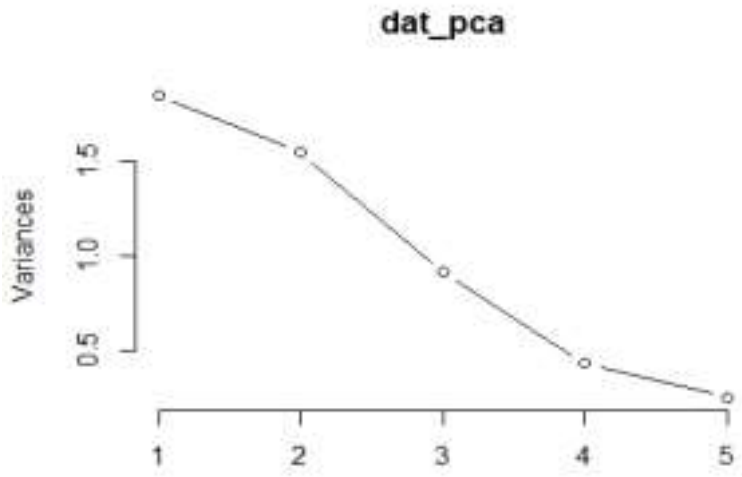


Figure 1

```
dat.fact=factanal(dat[,c(-1,-2,-3,-8,-10,-
11)],2,rotation="varimax") dat.fact
```

```
##
## Call:
## factanal(x = dat[, c(-1, -2, -3, -8, -10, -11)], factors = 2,
rotation =  "varimax")
##
## Uniquenesses:
## Wkday_Slp WkdayQ WkndSlp WkndQ Workout
## 0.005 0.664 0.505 0.005 0.879
##
## Loadings:
## Factor1 Factor2
## Wkday_Slp 0.997
## WkdayQ 0.229 0.533
## WkndSlp 0.691 0.136
## WkndQ 0.997
## Workout 0.221 -0.269
##
## Factor1 Factor2
## SS loadings 1.571 1.371
## Proportion Var 0.314 0.274
## Cumulative Var 0.314 0.588
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 12.38 on 1 degree of freedom.
## The p-value is 0.000435
```

```
print(dat.fact,digits=2,cutoff=0.35,sort=TRUE)
```

```
##
## Call:
## factanal(x = dat[, c(-1, -2, -3, -8, -10, -11)], factors = 2,
rotation =  "varimax")
##
## Uniquenesses:
## Wkday_Slp WkdayQ WkndSlp WkndQ Workout
## 0.00 0.66 0.50 0.00 0.88
##
## Loadings:
## Factor1 Factor2
## Wkday_Slp 1.00
## WkndSlp 0.69
## WkdayQ 0.53
## WkndQ 1.00
## Workout
##
## Factor1 Factor2
## SS loadings 1.57 1.37
## Proportion Var 0.31 0.27
## Cumulative Var 0.31 0.59
##
## Test of the hypothesis that 2 factors are sufficient.
```

```
## The chi square statistic is 12.38 on 1
degree of freedom. ## The p-value is 0.000435

############################################################
############# #Loadings:
# Factor1 Factor2
#Wkday_Slp 1.00
#WkndSlp 0.70
#WkdayQ 0.53
#WkndQ 1.00
#Workout

#########################################################################
#

dat$SlpScr=(dat$Wkday_Slp+dat$WkndSlp)/2
dat$SlpQScr=(dat$WkdayQ+dat$WkndQ)/2


#dat$Activity=dat2$Activity
#str(dat)
dat$ActScr=(dat$Workout+dat$Activity)/2

dat$Prod=(dat$PrbSol+dat$Decision)/2

dat$Prod=round(dat$Prod)

############################################################
############# #############To determine the optimal
number of clusters

#c$prod=0
#View(c)
c=dat[,c(12,13,14)]
c$prod=ifelse(dat$Prod<3,1,2)

#View(c)
library(NbClust)
dat_clust=NbClust(c[,-4],distance='euclidean',min.nc = 2,max.nc =
6,method = "a verage")

## [1] "Frey index : No clustering structure in this data set"
```

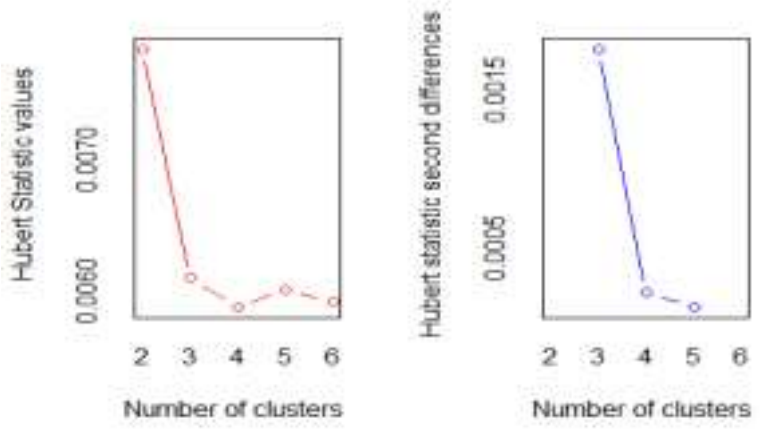


Figure 2

```
## *** : The Hubert index is a graphical method of determining the
number of cl usters.
## In the plot of Hubert index, we seek a significant knee that
corresponds to a
## significant increase of the value of the measure i.e the sig
nificant peak in Hubert
## index second differences plot.
##
```

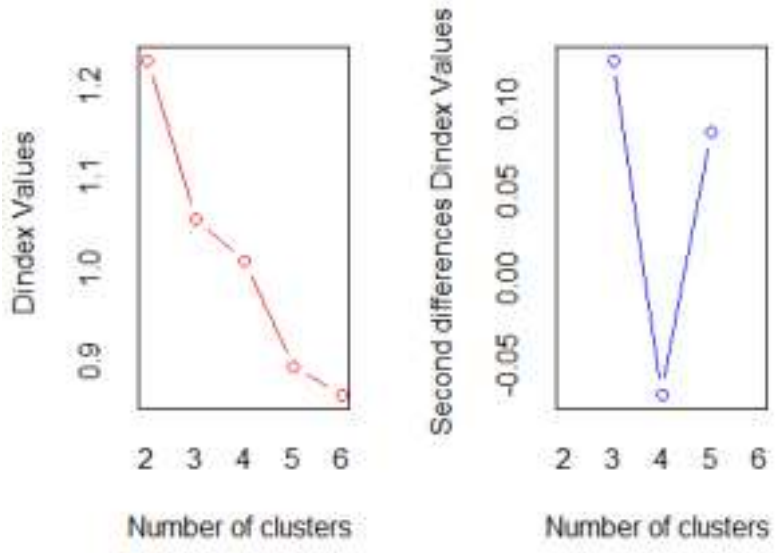


Figure 3

```
## *** : The D index is a graphical method of determining the
number of cluster s.
## In the plot of D index, we seek a significant knee (the sign
```

```
ificant peak in Dindex
## second differences plot) that corresponds to a significant i
ncrease of the value of
## the measure.




##
##
*******************************************************
*********  ## * Among all indices:
## * 11 proposed 2 as the best number of clusters
## * 3 proposed 3 as the best number of clusters
## * 6 proposed 5 as the best number of clusters
## * 3 proposed 6 as the best number of clusters
##
## ***** Conclusion *****
##
## * According to the majority rule, the best number of
clusters is 2  ##
##
##
*******************************************************************
```

```
#Apply K-means cluster using 2 clusters
scl_C=scale(c)
Ckn=kmeans(scl_C,2,nstart = 20)
Ckn$size
```

```
## [1] 20 190
```

```
#knbio$centers ########################explore more
#table(biopsy$class)

#Create Custer plot
library(cluster)
clusplot(scl_C, Ckn$cluster,color = T, label=4,cex = 1,main="K-
clustering")

#as.factor(dat$Prod)
#as.factor(Ckn$cluster)
str(Ckn$cluster)
```

```
## int [1:210] 2 2 2 2 2 1 2 2 2 2 ...
```

```
tabulation=table(c$prod,Ckn$cluster)
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R
```

```
version 4.0.2 ## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4

randIndex(tabulation)

## ARI
## 1

tabulation
```

```
##
## 1 2
## 1 20 0
## 2 0 190

Ckn$cluster[210]

## [1] 2
```

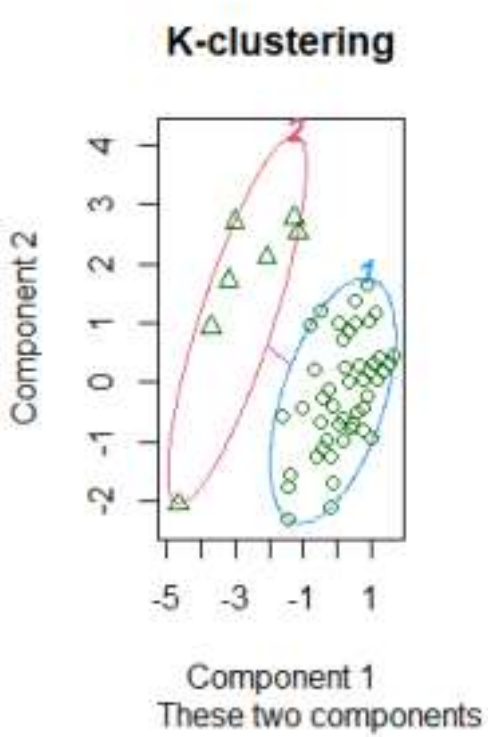


Figure 4

```
##############################################################
```

```
#LOGISTIC REGRESSION


#View(c)
library(caret)

## Warning: package 'caret' was built under R version 4.0.2

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.2

set.seed(100)

c$prod=ifelse(c$prod==1,0,1)

part=createDataPartition(c$prod,p=0.8,list = FALSE)
train=c[part,]




test=c[-part,]

#####################################
#####################################
#View(da)
da=dat[,c(1,2,12,13,14)]
da$prod=c$prod
da$Gender=ifelse(da$Gender=="Male",0,1)
part=createDataPartition(da$prod,p=0.8,list = FALSE)
train=da[part,]
test=da[-part,]
mod2=glm(prod~.,data=train,family=binomial())

#prod=-10.6600+Gender*-
1.2580+Age.Group*1.0347+SlpScr*0.6892+SlpQScr*1.3816+Act
Scr*1.5959

summary(mod2)

##
## Call:
## glm(formula = prod ~ ., family = binomial(),
data = train) ##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -2.6836 0.1585 0.2231 0.3916 1.3725
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.6600 3.5486 -3.004 0.002664 **
## Gender -1.2580 0.7126 -1.765 0.077493 .
## Age.Group2 1.0347 0.9880 1.047 0.294976
```

```
## SlpScr 0.6892 0.2993 2.303 0.021288 *
## SlpQScr 1.3816 0.3909 3.535 0.000408 ***
## ActScr 1.5959 0.6619 2.411 0.015901 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1 ##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 105.669 on 167 degrees of freedom
## Residual deviance: 76.248 on 162 degrees of freedom
## AIC: 88.248
##
## Number of Fisher Scoring iterations: 6
```

```
test$pred1=predict(mod2,type = 'response', newdata = test )
View(test)
#test$pred1Conv=0
test$pred1Conv=ifelse(test$pred1<0.8639745,0,1)

library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 4.0.2
```

```
table(test$pred1Conv)
```

```
##
## 0 1
## 8 34
```

```
confusionMatrix(as.factor(test$pred1Conv),as.factor(test$prod))
```

```
## Confusion Matrix and Statistics
##
## Reference
## Prediction 0 1
## 0 3 5
## 1 1 33
##
## Accuracy : 0.8571
## 95% CI : (0.7146, 0.9457)
## No Information Rate : 0.9048
## P-Value [Acc > NIR] : 0.8999
##
## Kappa : 0.4273
##
## Mcnemar's Test P-Value : 0.2207
##
## Sensitivity : 0.75000
## Specificity : 0.86842
## Pos Pred Value : 0.37500
## Neg Pred Value : 0.97059
```

```
## Prevalence : 0.09524
## Detection Rate : 0.07143
## Detection Prevalence : 0.19048
## Balanced Accuracy : 0.80921
##
## 'Positive' Class : 0
##

#Accuracy : 0.8571
#Sensitivity : 0.75000
#Specificity : 0.86842

pred=prediction(test$pred1Conv,test$prod)
roc_pred=performance(pred,measure = "tpr",
x.measure = "fpr") plot(roc_pred)
```

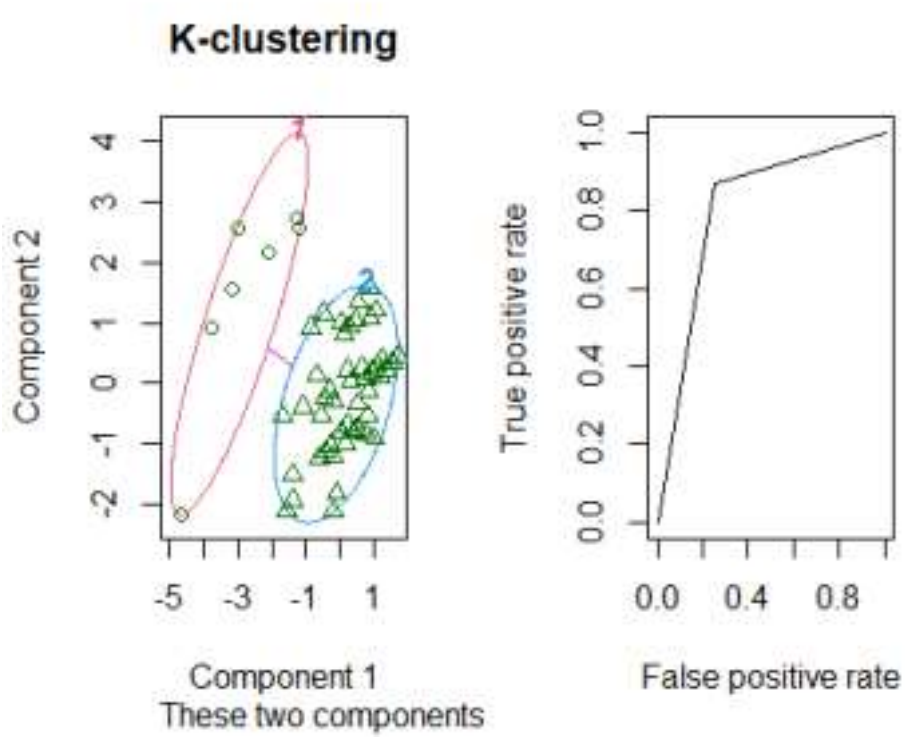


Figure 5

```
auc=performance(pred,measure = "auc")
auc@y.values[{1}]

## [[1]]
## [1] 0.8092105

# 0.8092105

#####################################
#####################################
```

Conclusion:

Upon computation of various analysis, we conclude that productivity is affected by their

respective quantity and quality of sleep and their activity levels as well. Based on the primary data collated it was observed that there exists a significant association between the determined "Sleep Quality Score" and "Productivity Levels".

The principle component analysis (*Fig1.*) excludes the dependent and categorical variables in order to determine the appropriate factors . The *NbClust package* is used to visualize the two graphs (*Fig 2. & Fig 3.*), which depicts the optimum number of clusters to be used for the K-Means Clustering

method, we conclude the optimum number of clusters to be two. The Cluster plot *(Fig 4.*) is used to visualize the output of the productivity categorization model classified into High & Low productivity levels, where in *cluster 1* represents all the samples which lie within Low Productivity and *cluster 2* points out all the samples which have been categorized into High Productivity Levels based on the variety of individual productivity scores calculations.

With the help of this cluster classification, we have primarily developed a categorization model which helps us evaluate the current status and productivity levels of an individual. Furthermore, based on the above classification a logistic regression model is developed for fututre prescriptions to make changes to controllable variables aimed at improving sleep patterns which will then enable an individual transcend from lower to higher productivity levels.

Impact of Vegetarianism on Millennials and Gen Z

Submitted By-
Alstrin Cyrus
Anusha M
Dharshini M

Abstract:

Vegetarianism is a practice of not eating meat or fish. It has become a polarizing subject in today's culture. The main objective of this survey is to understand the impact of vegetarianism on Millennials and Gen Z. This study was done based on the data collected from 150 people where the majority were millennials and Gen Z. Various factors such as age, cultural upbringing, environmental concerns, etc., were analysed to come to a conclusion. There is a rising knowledge on veganism and its benefits and the evidence shows a slow and steady increase in vegetarianism among this generation. The data collected from individuals of various backgrounds shows us that the rise in the mentioned concerns has had its effects on people's minds. The result obtained in the end tends to show higher accuracy towards people of younger age groups moving towards veganism for a healthier option.

Key Words:

Vegetarianism, Veganism, Millennials, Gen Z, Balanced diet, Environment effects

Introduction:

"Appropriately planned vegetarian diets, including total vegetarian or vegan diets, are healthful, nutritionally adequate, and may provide health benefits in the prevention and treatment of certain diseases." - American Dietetic Association

Millennial are dieting just like older generations. But their reasons for doing so seem to differ. More Millennial are changing their diets in pursuit of both physical and mental wellness and a desire to reduce their climate footprints, than are members of older generations. We have

analyzed the data to find the impact of Veganism on millennials and gen Z. The data has been cleaned and sorted. The relationship between the different factors are been found by correlation and regression. The best model is built by taking age and their choice of diet. Tests are done to confirm that the model built is the best model.

Purpose:

- To understand the impact of Vegetarianism and Veganism, and the factors that help them make their diet choices feasible for Millennials and Gen Z
- To build a best model to determine the factors influencing their choice of diet

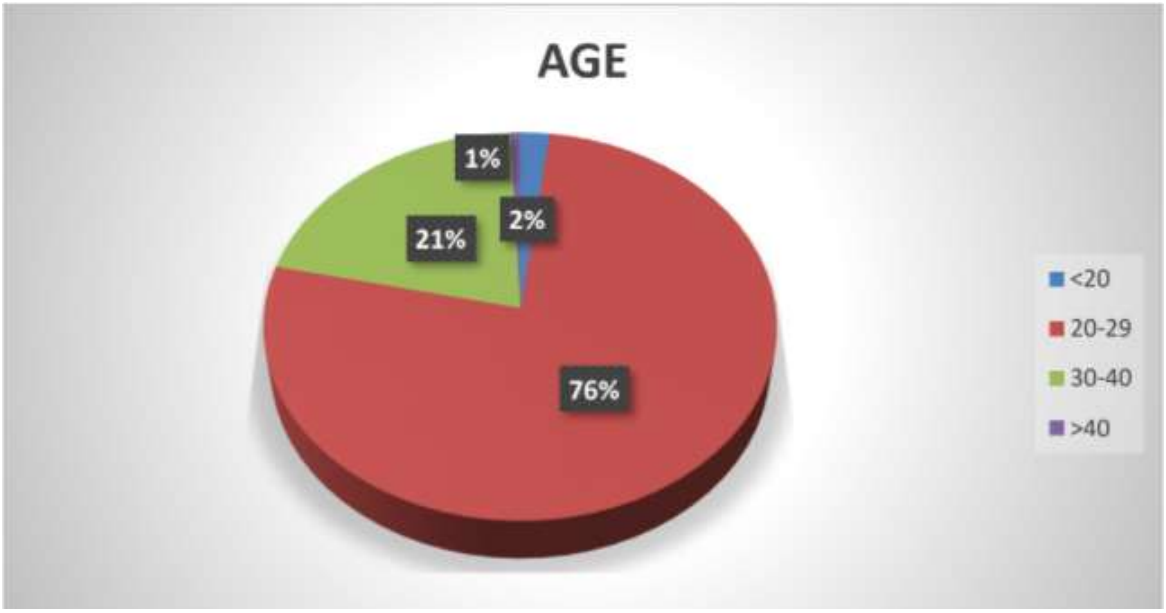
Methodology:

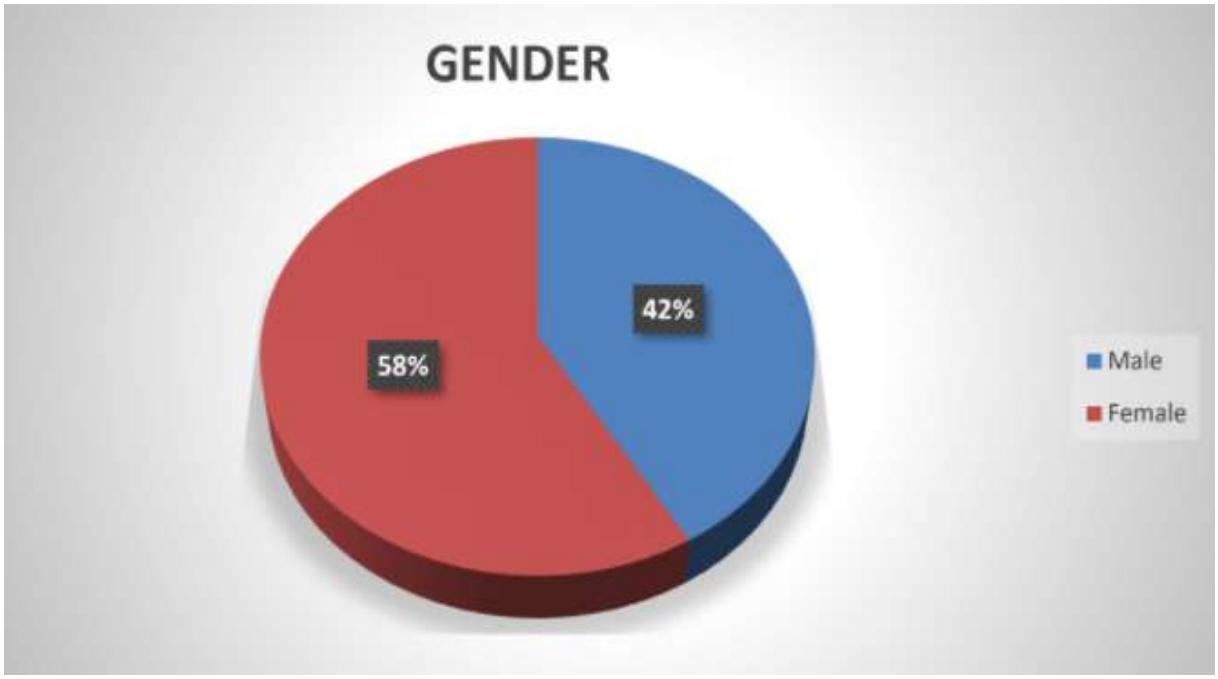
Qualitative Analysis has been done by making use of Predictive analytics. Literature review of the available past research papers on this topic has been referred to understand the impact of vegetarianism on Gen Z.

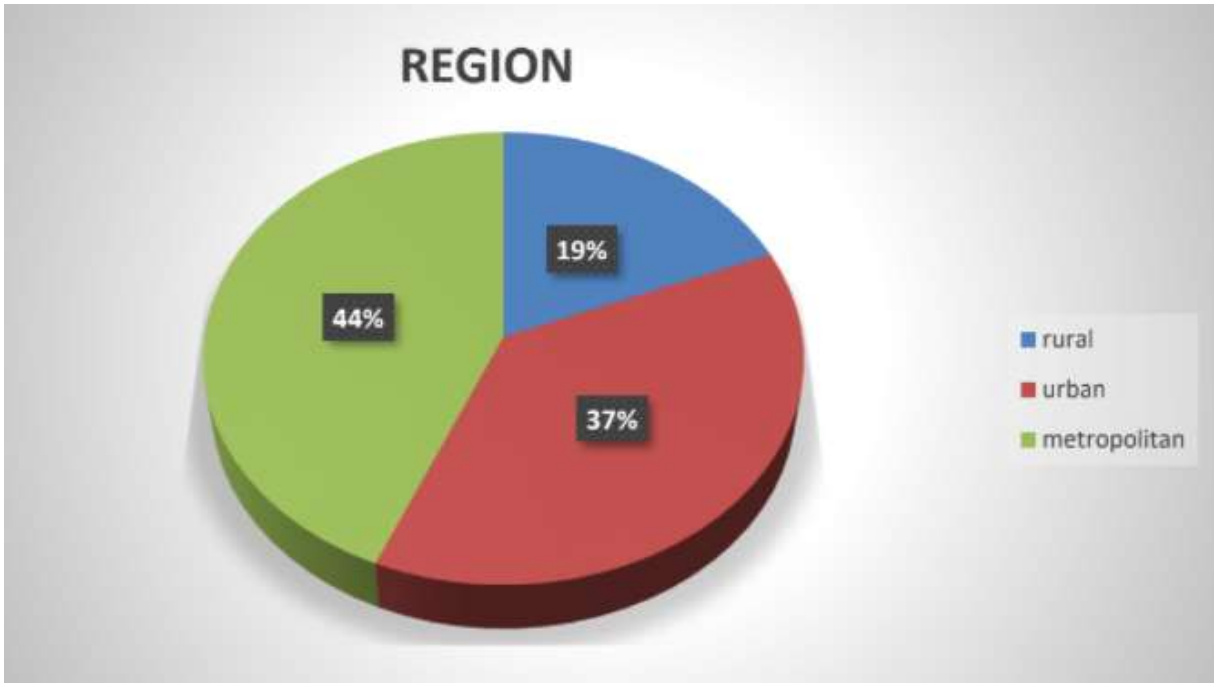
Methodology:

- In this research we have conducted a google form online survey among 150 people to understand the impact of Vegetarianism and veganism and factors that help them make their diet choices feasible for the Millennial and Gen Z
- Literature review has been done to understand the past research which has been done in esteemed research papers and to understand the impact of vegetarianism on Millennials and Gen Z.

Demographic Profile:




GENDER
42%
58%
Male
Female


REGION
19%
44%
37%
rural
urban
metropolitan

Predictive Analysis:

```
getwd()
setwd("C:/Users/Anusha/Documents/R 4th trimester")
vegan<-read.csv("Impact of vegetarianism.csv")
str(vegan)
```

Step 1: Converting as factors

```
vegan$Gender<-as.factor(vegan$Gender)
vegan$Region<-as.factor(vegan$Region)
vegan$Diet.<-as.factor(vegan$Diet.)
vegan$X.Meat.consumption<-as.factor(vegan$X.Meat.consumption)
vegan$X.Culture.upbringing<-as.factor(vegan$X.Culture.upbringing)
vegan$X.Culture.upbringing<-
as.factor(vegan$X.Weight.management)
vegan$X.Animal.welfare<-as.factor(vegan$X.Animal.welfare)
vegan$X.Environmental.concerns<-
as.factor(vegan$X.Environmental.concerns) vegan$X.Allergies<-
as.factor(vegan$X.Allergies)
vegan$X.Conversation.about...diet.<-
as.factor(vegan$X.Conversation.about...diet.) vegan$X.Availability<-
as.factor(vegan$X.Availability)
vegan$X..Healthier.Alternative<-as.factor(vegan$X..Healthier.Alternative)
vegan$Availability.in.Restaurants.<-
as.factor(vegan$Availability.in.Restaurants.) vegan$X.Cost<-
as.factor(vegan$X.Cost)
vegan$X.Maintainig.balanced.diet.<-as.factor(vegan$X.Maintainig.balanced.diet.)
```

Step 2: No need to clean data as there are no missing values.

```
table(complete.cases(vegan))
```

TRUE

149

Step 3: Creating models

From the dataset, Diet is taken as the dependent data where 1 suggests that they follow vegan diet and 2 suggests non-vegan diet.

Logistic Regression Analysis

```
library(caret)

set.seed(100)

part<-createDataPartition(vegan$Diet., p=0.80, list=FALSE)

part

trainlog<-vegan[part,]

testlog<-vegan[-part,]


library(MASS)

fitall<-glm(Diet.~.,data=trainlog,family=binomial())

stepAIC(fitall)


model<-glm(formula = Diet. ~ Age + X.Meat.consumption +

 X.Maintainig.balanced.diet.,

 family = binomial(), data = trainlog)

summary(model)
```

The best model was found with least **AIC** of **148.96**

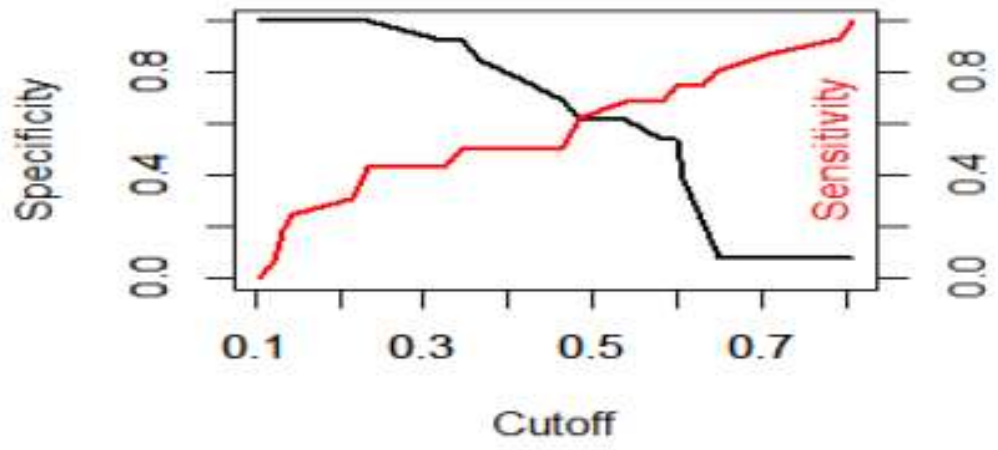
Predicting values for test data:

```
pred<-predict(model, newdata = testlog, type = "response")
```

Now since the predicted values are in decimals, the values needs to be converted to 1s and 2s for which we need to find the cut off value.

Cut off Value:

```
library(ROCR)

predictions<-prediction(pred,testlog$Diet.)

roc.pred=performance(predictions,measure='tpr',x.measure='fpr')


plot(unlist(performance(predictions, "sens")@x.values),
unlist(performance(predictions,  "sens")@y.values),

 type="l", lwd=2, ylab="Specificity", xlab="Cutoff")

par(new=TRUE)

plot(unlist(performance(predictions, "spec")@x.values),
unlist(performance(predictions,  "spec")@y.values),

 type="l", lwd=2, col='red', ylab="", xlab="")

axis(4, at=seq(0,1,0.2))

mtext("Sensitivity",side=4, padj=-2, col='red')
```



The cut off value is 0.5.

Confusion Matrix:

```
convert<-ifelse(pred<0.5,"1","2")
conf<-data.frame(predicted=convert, actual=testlog$Diet.)
conf$predicted=as.factor(conf$predicted)
res<-confusionMatrix(conf$predicted,conf$actual)
res
```

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2
         1 11  5
         2  5  8

               Accuracy : 0.6552
                 95% CI : (0.4567, 0.8206)
    No Information Rate : 0.5517
    P-Value [Acc > NIR] : 0.1756

                  Kappa : 0.3029

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.6875
            Specificity : 0.6154
         Pos Pred Value : 0.6875
         Neg Pred Value : 0.6154
             Prevalence : 0.5517
         Detection Rate : 0.3793
   Detection Prevalence : 0.5517
      Balanced Accuracy : 0.6514

       'Positive' Class : 1
```

Reference

Prediction 1 2

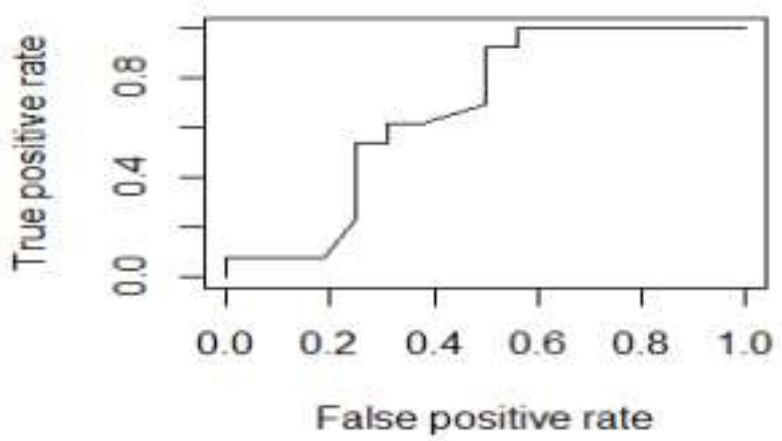
1 11 5

2 5 8

Accuracy : 0.6552

Area Under the Curve:

```
plot(roc.pred)
```




```
auc=performance(predictions,measure='auc')

auc@y.values[{1}]

[[1]]
[1] 0.6730769
```

Findings:

Findings done in the case show us that the best model constructed would be the one where the age affecting their dietary choices and also that if people are having meat often or not, does not affect their health in any way to keep a balanced diet. While there is no correlation between gender and their choice of diet, there seems to be some significance in peoples demographic and their ease in finding vegan food. The model predicting that age has a correlation with people eating vegan food is of significant importance to the case and its results. Even though there is a rising knowledge on veganism and its benefits, the availability of vegan products to maintain a healthy diet should increase. The case provides data to back its suggestion that the awareness and use of vegan products are higher in millennials and gen z.

Practical Implications:

The practice of vegetarianism in the generation that is going to be the future identifies itself as an environmentally concerned society with empathy for animals. It increases our chances for the fight against few of the environmental damages we create on ourselves. This case shows the growing concern of issues that is being fought in a major part of our future society.

Conclusion:

The question arising in this growing population and the concerns on the future of the decaying environment is trending in this society and is not about to die soon. The data collected from individuals of various backgrounds shows us that the rise in the mentioned concerns has had its effects on people's minds. The result obtained in the end tends to show higher accuracy towards people of younger age groups moving towards veganism for a healthier option. Veganism does not only seem to be a trend but a healthier option when it comes to saving the resources of the planet. This practice in everyday lives tends to keep increasing in our world, and seems to form the base in fighting the killing of innocent animals. Various organizations are seemed to be forming to spread this practice for the betterment of the future and also as a proof to having a balanced and healthy diet. Without having to slaughter animals and without having to spend too much money, one can easily live a healthy life with empathy and financial stability. The analysis done and the results obtained give us the knowledge that the society of the future have started to follow and spread the word on veganism.

A Study of Customer Attributes and Various Mobile Phone Features Influencing Purchase Decisions

Submitted By-
Aishwarya Ajith (PG19007)
Ayan Paul (PG19034)
Pallavi Choudhary (PG19085)

Introduction

Customer retention refers to the ability of a company or product to retain its customers over some specified period. Higher customer retention means customers of the product or business tend to return to, continue to buy or in some other way does not defect to another product or business, or to non-use entirely. The study we did on topic customer retention on different brands of mobiles.

The study is aimed towards understanding the likelihood of a customer choice from the different brands of mobile and also to know whether they are using the same brand or changing to another brand while buying next time. It will help us understand customer retention of ability of the concerned companies.

The various parameters we are using for testing are Name, Gender, Age, Profession, Income, Criteria of buying, Customer Rating across various attributes, how frequently a user has changed his devices.

Methodology

The study is based on the quantitative data and we are collecting the primary data from the people who are using smartphone and also people who are above age 15. The data we have gather is descriptive data because we are gathering data without any intervening.

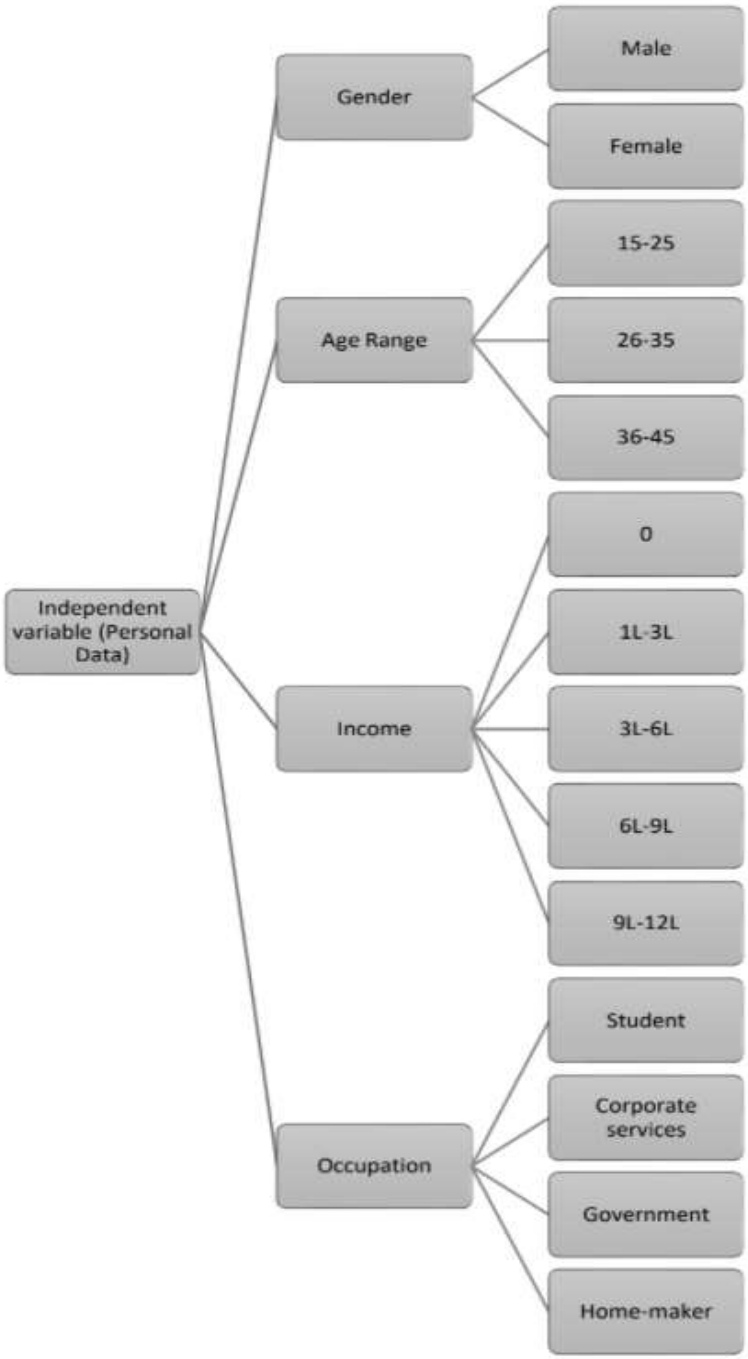
The methods we used for data collection is quantitative method. Survey was conducted by creating Google form to take the responses of the respondent. The questions we design in our google form are multiple choice and Likert scale. The data consists of 90 entries.

We created a form of two section. In first section we are having personal information of the respondent and in the second section we are having the questions which helps us to understand the smartphone preference of respondent and also the criteria of buying the smartphone.

The methods of analysis used in study is quantitative. The data is prepared before doing the analysis is by checking for the missing data and then removed the outliers. The software we use to analyse the data is R . We used Multinomial Logistic Regression for predicting the dependent variable which is the switching time (Time taken by them in switching from old to new phone). The various demographic factors are used as the predictors. The model was accordingly built and validated.

Analysis And Interpretation

The analysis and interpretation have been done using Histogram Charts ,Hypothesis testing (preferably Chi-Square test) and Regression Models(Multiple logistic regression) . The Histogram charts depicts the level of agreeableness when two independent data are plotted against each other. The Hypothesis testing tests the influence of one independent variable on other by giving us a p-value on the basis of which we can assess the influence. The Logistic models shows how well the demographics can predict the switching time.

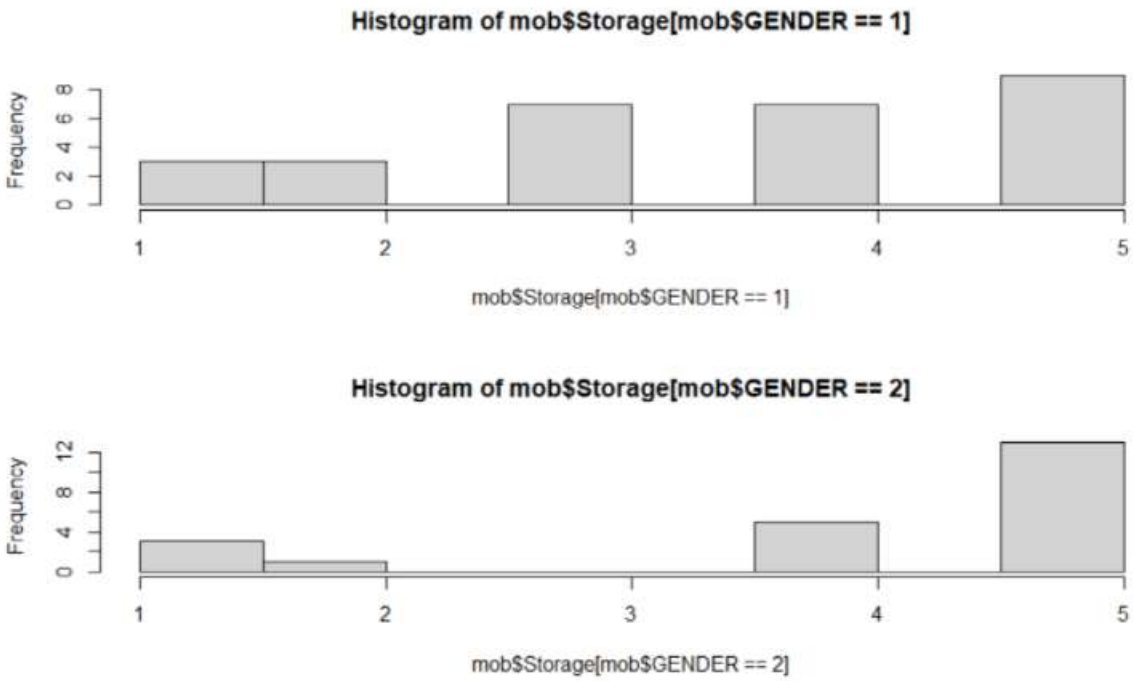


Independent variable (FACTORS)	LEVEL OF AGREEABLINESS
Storage Cost Battery Features Display OS	1- Strongly Disagree 2- Disagree 3- Neutral 4- Agree 5- Strongly Agree

1) Correlation

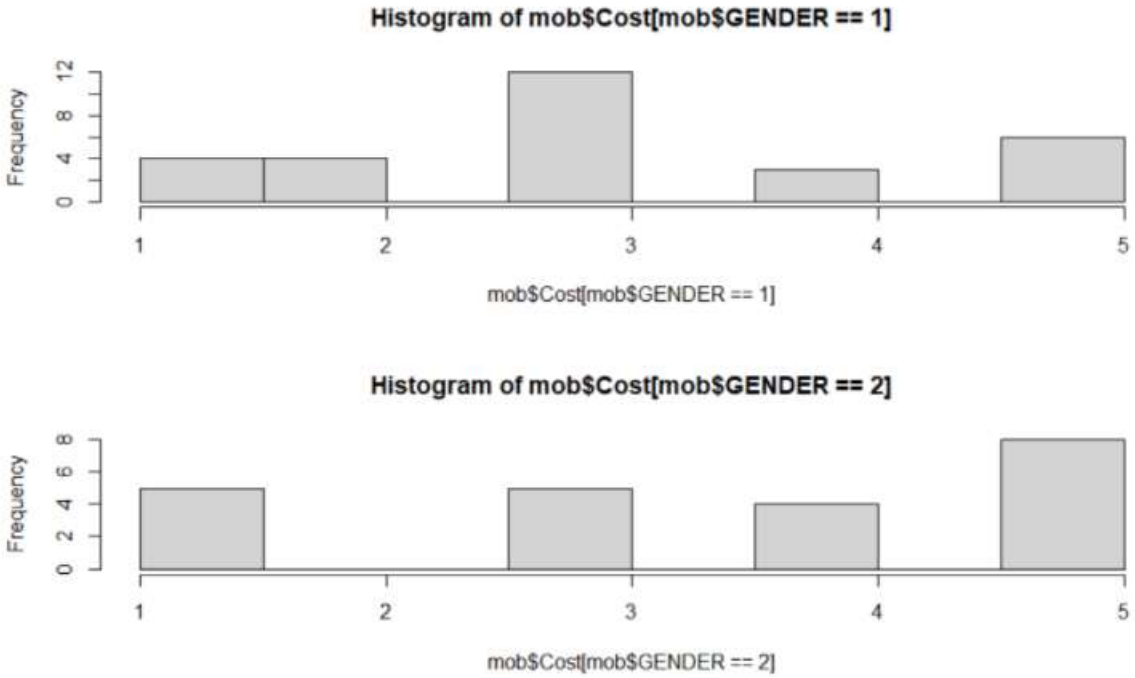
GENDER

Storage



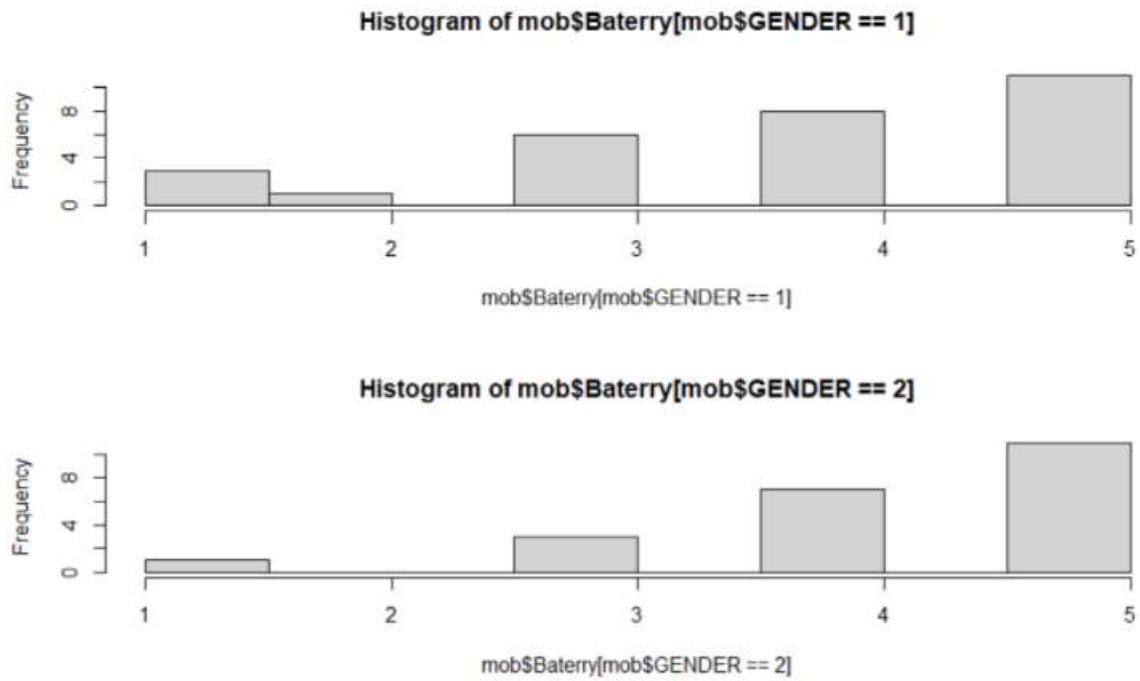
- Therefore, females, majority of them (8) Strongly Agree that storage is an important factor while switching into a new phone.
- In males, majority of them (10) Strongly Agree that storage is an important factor while switching into a new phone.

Cost



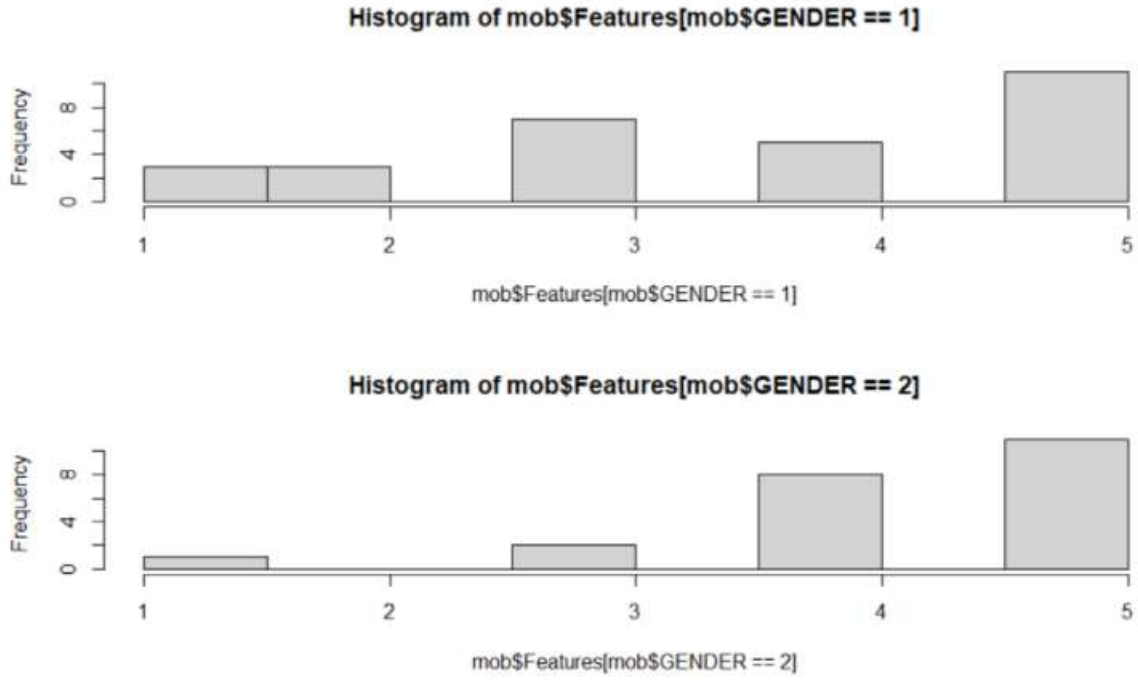
• Therefore, in females, most of them (12) remained neutral that cost factor is important.
• In males, majority of them (8) Strongly Agreed that cost factor is important.

Battery life



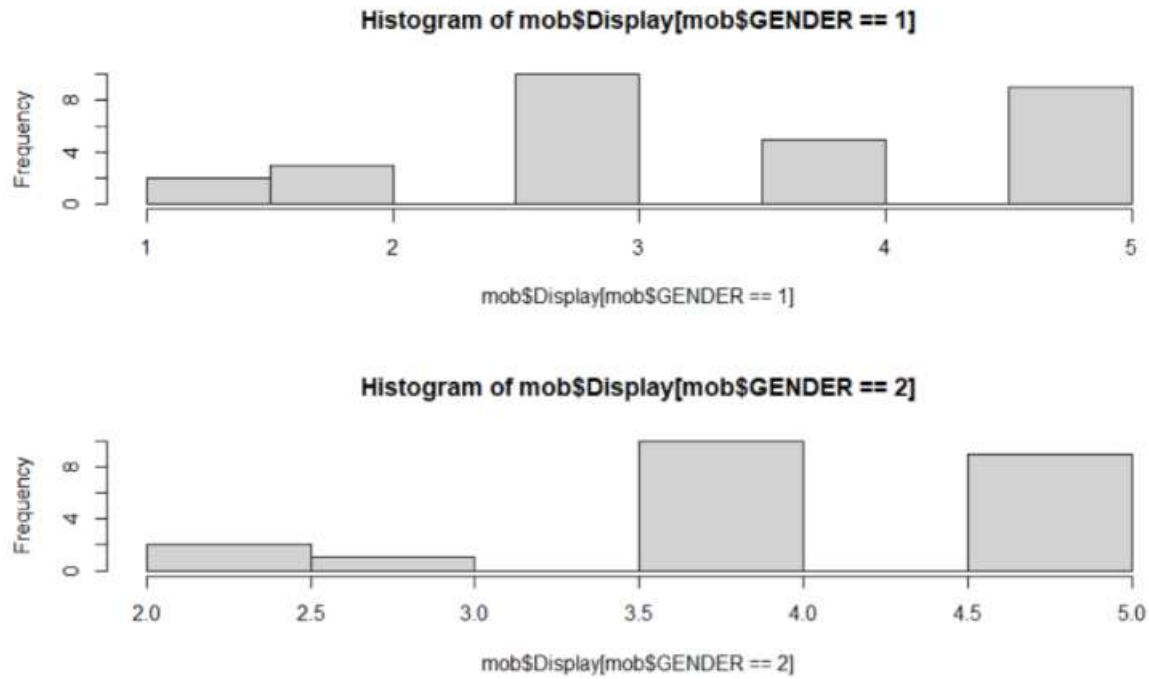
• Here both males and females, strongly agreed (10) on the significance of battery life.

Features



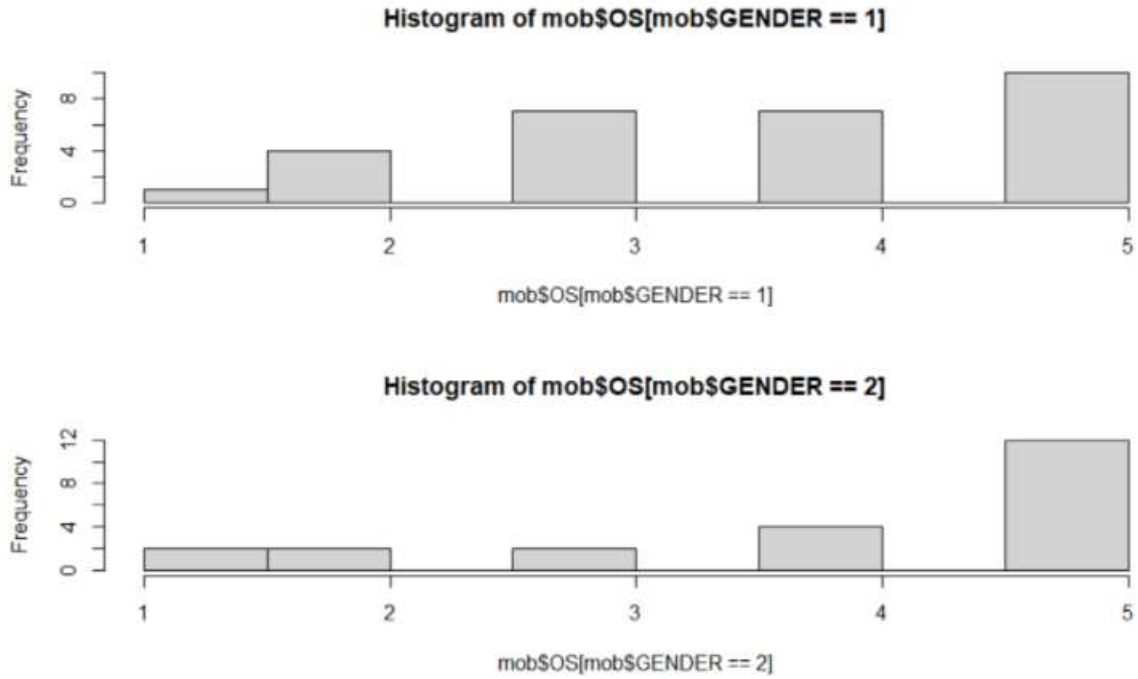
• Both male and female, Strongly Agree on the significance of Features.

Display



• Females tend to stay Neutral whereas males Agree.

OS



• Both the gender Strongly Agree that OS is an important factor.

1. Does gender influence on how much they value storage factor?

Null hypothesis (H0): Gender does not influence on how much they value storage factor. *Alternate Hypothesis (H1):* Gender does influence on how much they value storage factor

X-squared = 11.475, df = 8, p-value = 0.1762

Here, P- value > 0.05

Therefore, we accept H0. Gender does not influence on how much they value storage factor.

2. Does gender influence on how much they value cost factor?

Null hypothesis (H0): Gender does not influence on how much they value cost factor. *Alternate Hypothesis (H1):* Gender does influence on how much they value cost factor

X-squared = 12.251, df = 8, p-value = 0.1404

Here, P-value > 0.05

Therefore, we accept the H0. Gender does not influence on how much they value cost factor.

3. Does gender influence on how much they value battery life?

Null hypothesis (H0): Gender does not influence on how much they value battery life. *Alternate Hypothesis (H1):* Gender does influence on how much they value battery life.

X-squared = 3.4553, df = 8, p-value = 0.9026

Here, P-value > 0.05

Therefore, we accept the H0. Gender does not influence on how much they value battery life.

4. Does gender influence on how much they value features?

Null hypothesis (H0): Gender does not influence on how much they value features. *Alternate Hypothesis (H1):* Gender does influence on how much they value features.

X-squared = 8.0312, df = 8, p-value = 0.4304

Here, P-value > 0.05

Therefore, we accept the H0. Gender does not influence on how much they value features.

5. Does gender influence on how much they value display?

Null hypothesis (H0): Gender does not influence on how much they value Display. *Alternate Hypothesis (H1):* Gender does influence on how much they value Display.

X-squared = 12.75, df = 8, p-value = 0.1207

Here, P-value > 0.05

Therefore, we accept the H0. Gender does not influence on how much they value Display.

6. Does gender influence on how much they value OS?

Null hypothesis (H0): Gender does not influence on how much they value OS. *Alternate Hypothesis (H1):* Gender does influence on how much they value OS.

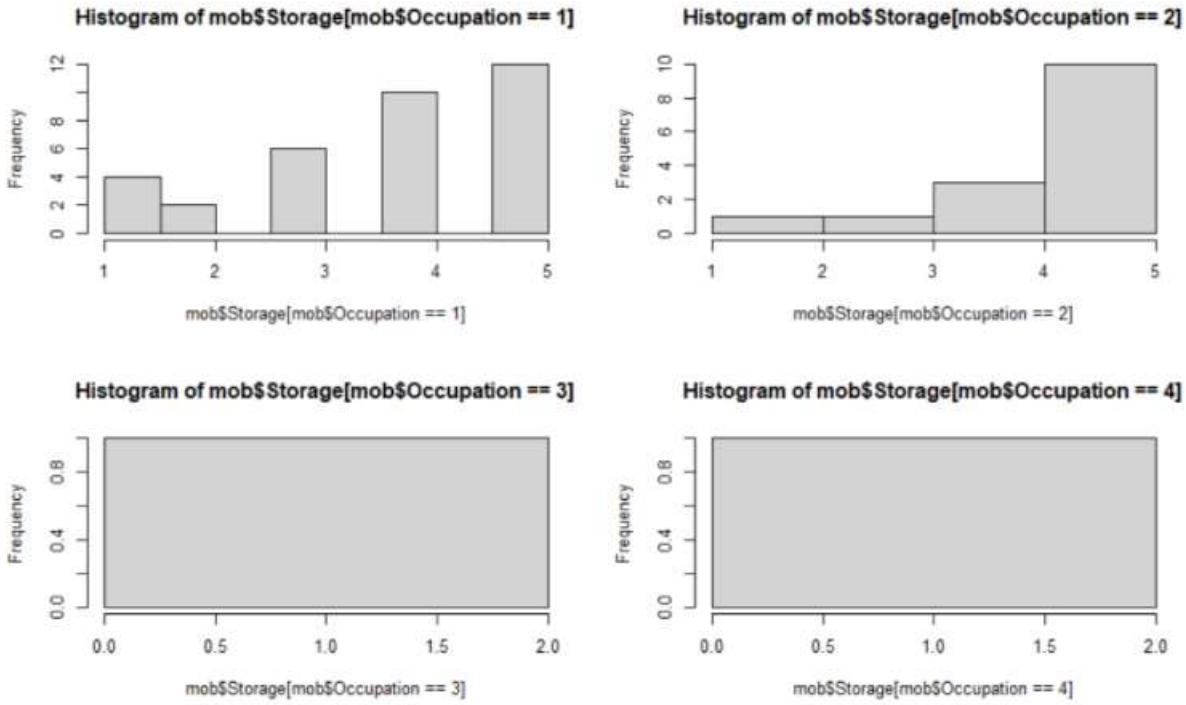
X-squared = 7.3478, df = 8, p-value = 0.499

Here, P-value > 0.05

Therefore, we accept the H0. Gender does not influence on how much they value OS.

OCCUPATION

Storage



Majority of the Students as well as the one working in Corporate Services strongly agree to storage being an important criterion whereas Government Employees and Home-makers disagree to storage being an important criterion.

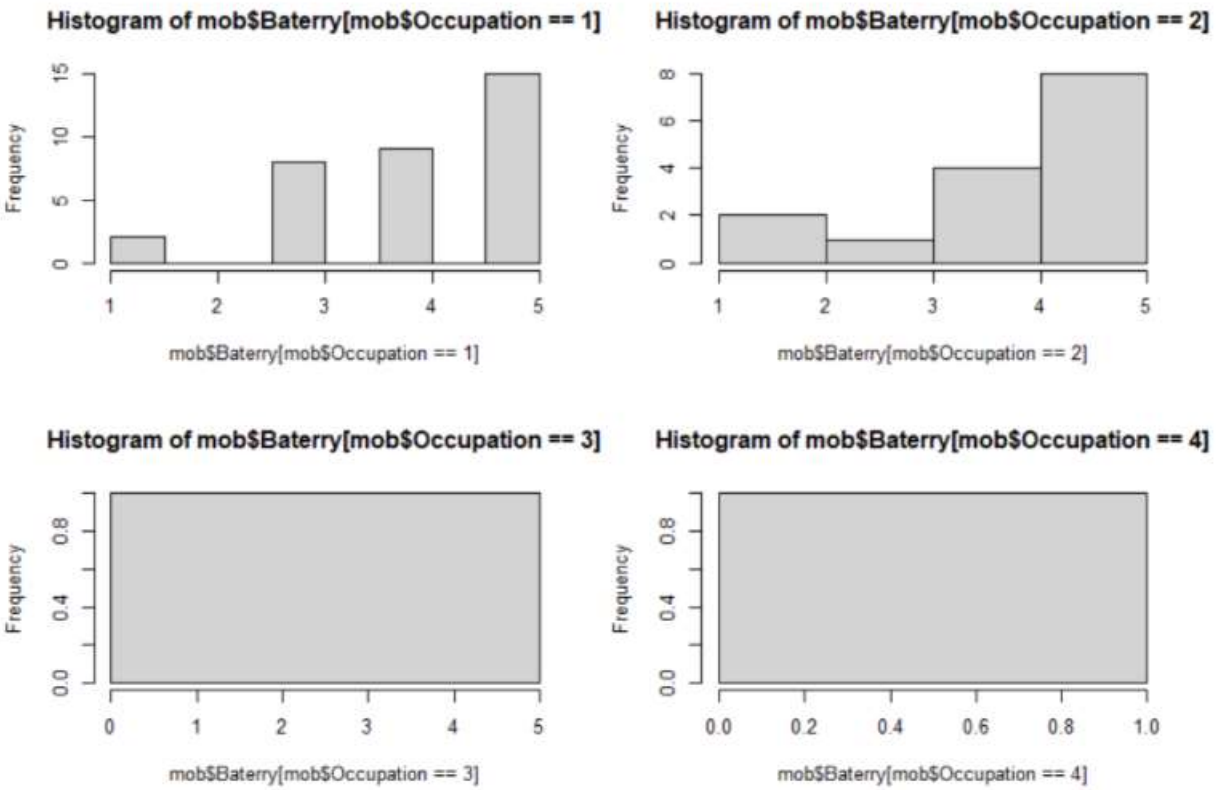
Does Occupation influence how much they value storage factor?

Null hypothesis (H0): Occupation does not influence how much they value storage factor. *Alternate Hypothesis (H1):* Occupation does influence how much they value storage factor. X-squared = 37.146, df = 16, p-value = 0.002

Here, P-value < 0.05

Therefore, we reject the H0. Occupation does influence how much they value storage factor.

Battery



Majority of the Students as well as the one working in Corporate Services strongly agree to battery being an important criterion whereas Government Employees strongly agree and Home-makers disagree to battery being an important criterion.

Does Occupation influence how much they value battery life?

Null hypothesis (H0): Occupation does not influence how much they value

Battery life. *Alternate Hypothesis (H1):* Occupation does influence how much

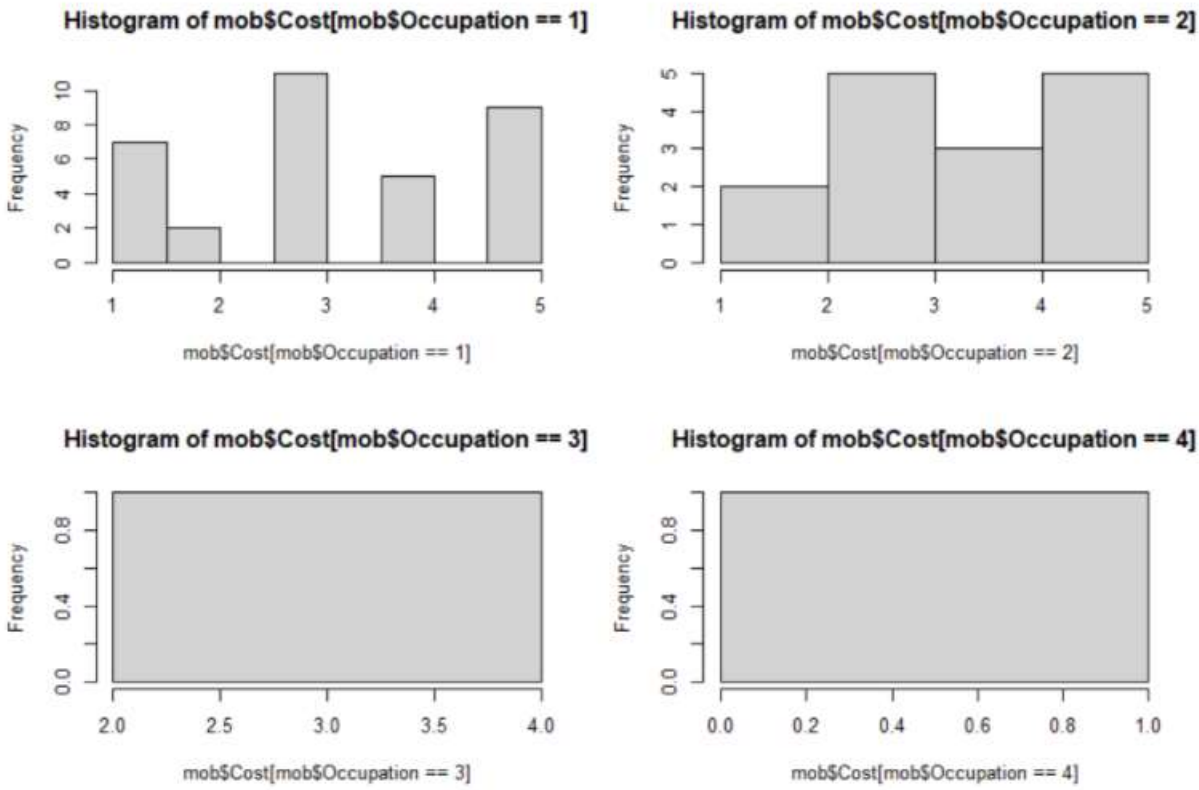
they value Battery life. X-squared = 21.622, df = 16, p-value = 0.1558

Here, P-value > 0.05

Therefore, we accept H0. Occupation does not influence how much they value

Battery life.

Cost



Majority of the Students are neutral for cost. The one working in Corporate Services may strongly agree as well as disagree to cost being an important criterion. Government Employees agree and Home makers strongly disagree to cost being an important criterion.

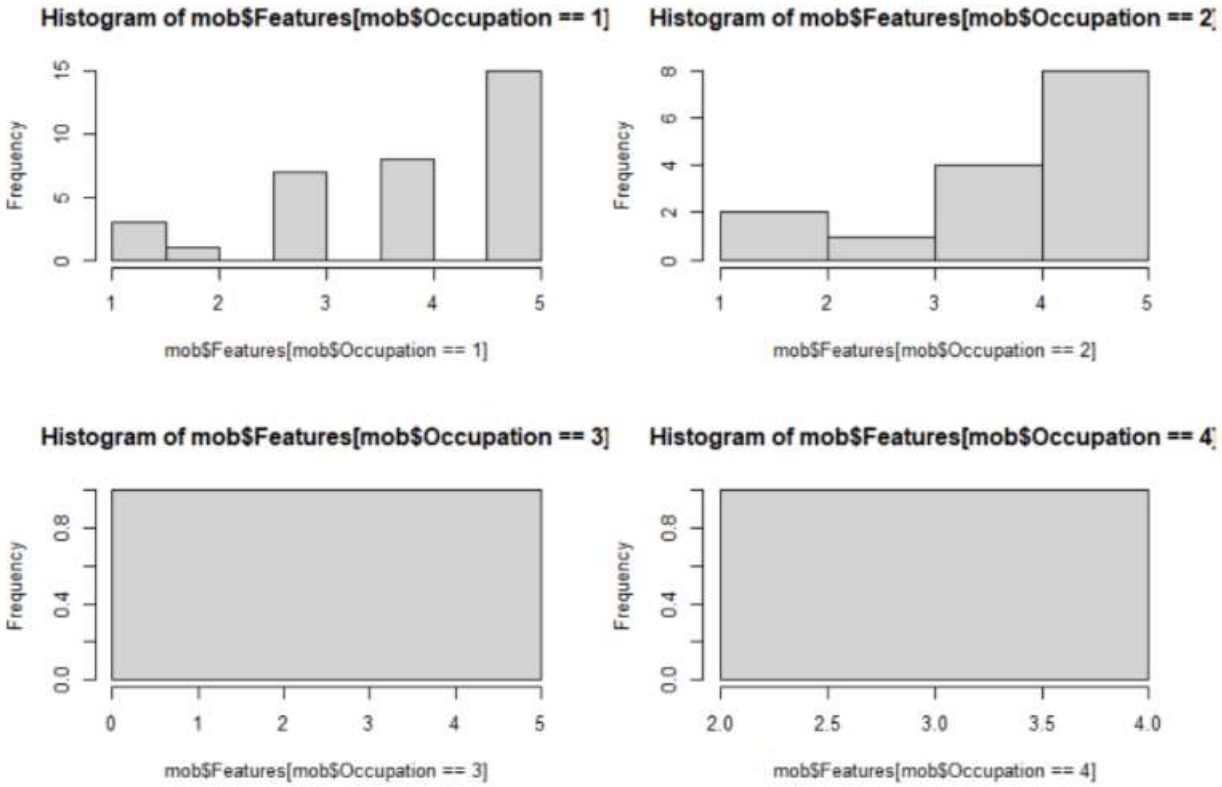
Does Occupation influence how much they value Cost factor?

Null hypothesis (H0): Occupation does not influence how much they value Cost factor. Alternate Hypothesis(H1): Occupation does influence how much they value Cost factor. X-squared = 20.647, df = 16, p-value = 0.1925

Here, P-value > 0.05

Therefore, we accept the H0. Occupation does not influence how much they value Cost factor.

Features



Majority of the Students as well as the one working in Corporate Services strongly agree features being an important criterion. Government Employees agree and Home-makers also agree to features being an important criterion.

Does Occupation influence how much they value Features?

Null hypothesis (H0): Occupation does not influence how much they value

Features. *Alternate Hypothesis (H1):* Occupation does influence how much

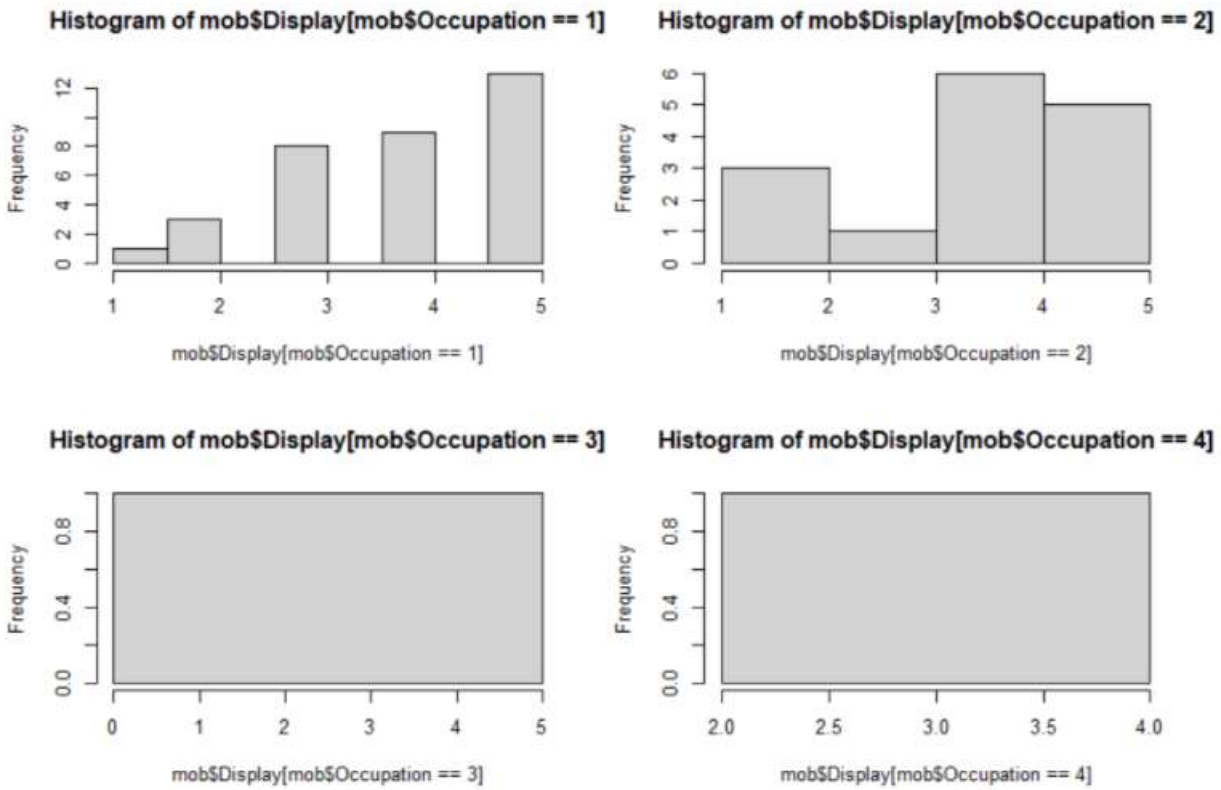
they value Features. X-squared = 26.2, df = 16, p-value = 0.05127

Here, P-value ≥ 0.05.

Therefore, we accept H0. Occupation does not influence how much they value

Features.

Display



Majority of the Students strongly agree Display is important as well as the one working in Corporate Services agree display being an important criterion. Government Employees strongly agree and Home makers also agree to display being an important criterion.

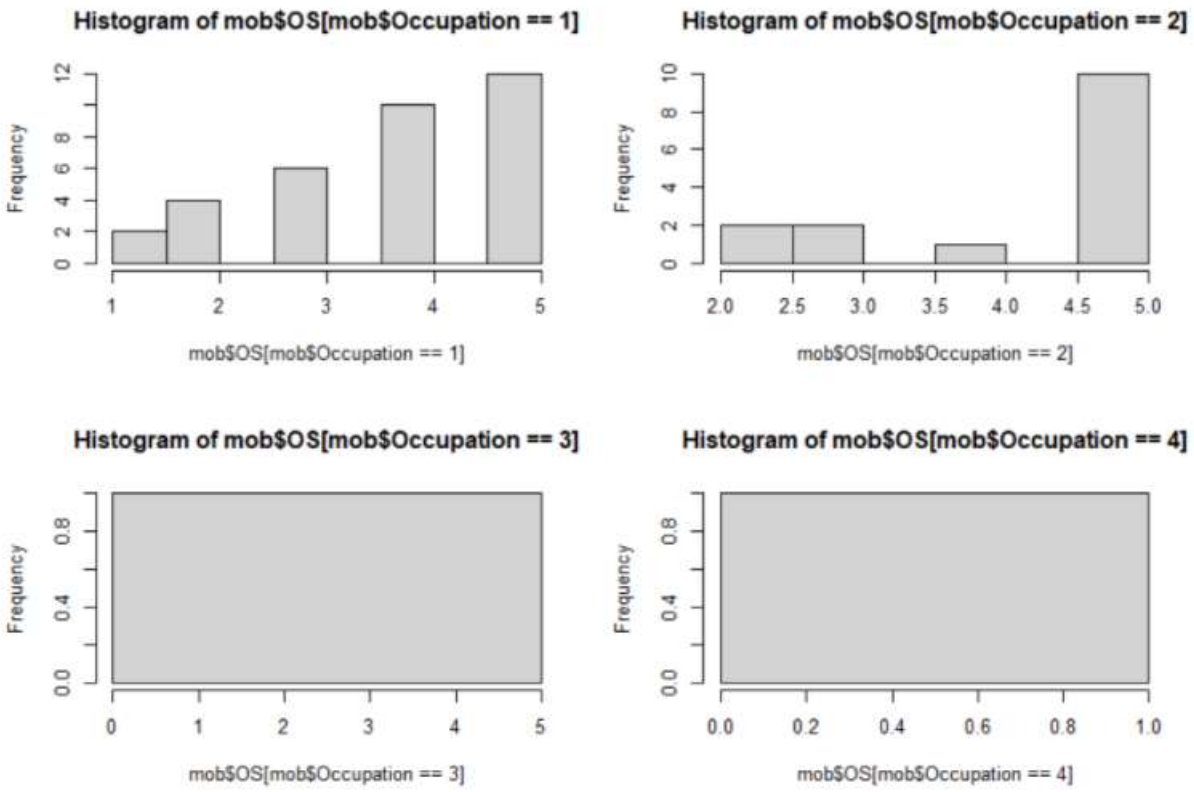
Does Occupation influence how much they value Display?

Null hypothesis (H0): Occupation does not influence how much they value Display. *Alternate Hypothesis (H1):* Occupation does influence how much they value Display. X-squared = 12.659, df = 16, p-value = 0.6975

Here, P-value > 0.05.

Therefore, we accept H0. Occupation does not influence how much they value Display.

OS



Majority of the Students strongly agree OS is important as well as the ones working in Corporate Services strongly agree OS being an important criterion. Government Employees strongly agree and Home-makers strongly disagree to OS being an important criterion.

Does Occupation influence how much they value OS?

Null hypothesis (H0): Occupation does not influence how much they

value OS. *Alternate Hypothesis (H1):* Occupation does influence how

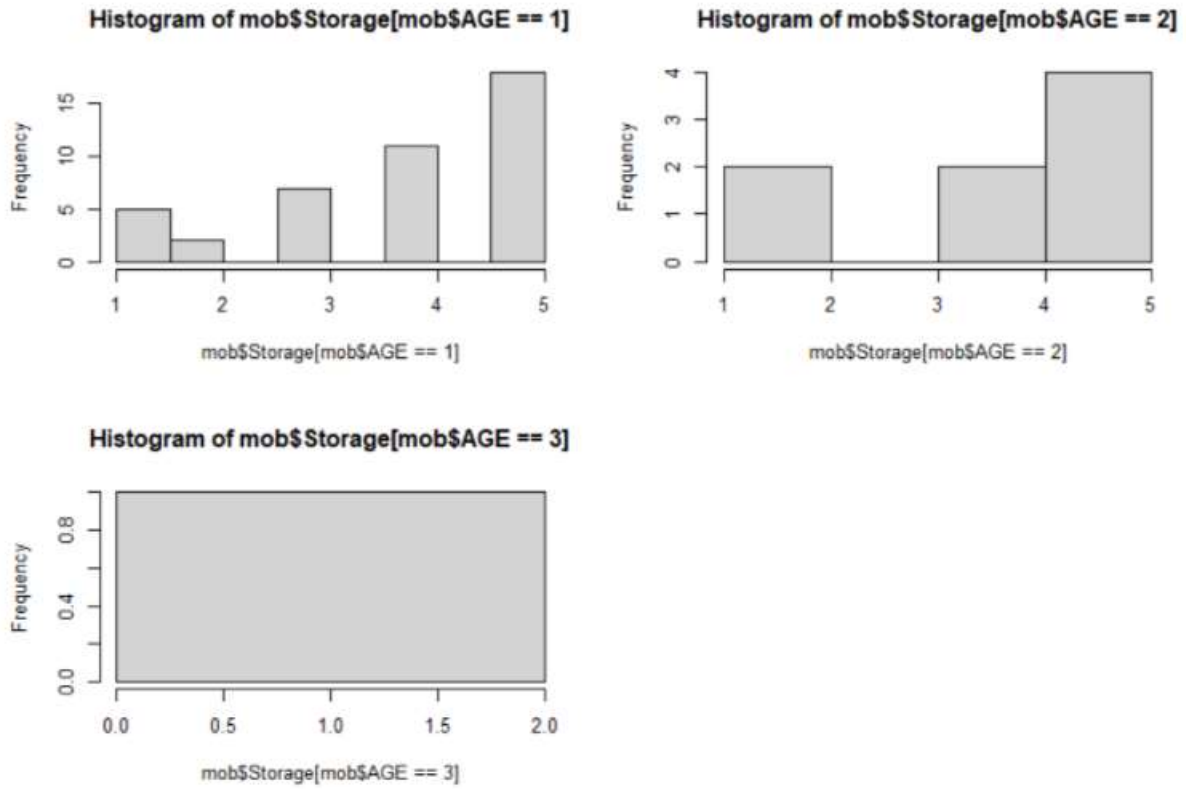
much they value OS. X-squared = 30.334, df = 16, p-value = 0.01635

Here, P-value < 0.05.

Therefore, we reject H0. Occupation does influence how much they value OS.

AGE

Storage



Age Group (15-25): Strongly Agree ; Age group (26-35): Strongly Agree ; Age group (36-45): Disagree

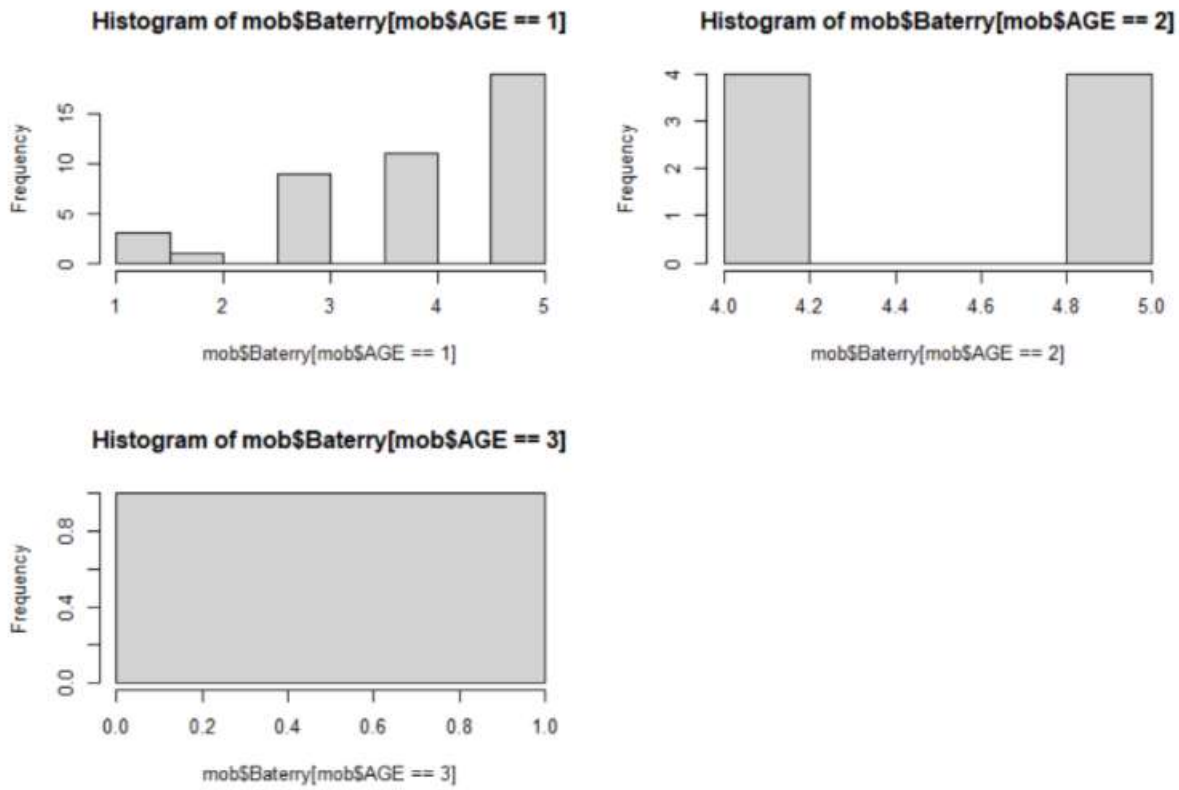
Does Age influence how much they value storage factor?

Null hypothesis (H0): Age does not influence how much they value storage factor. Alternate Hypothesis(H1): Age does influence how much they value storage factor. X-squared = 14.214, df = 8, p-value = 0.07635

Here, P-value > 0.05.

Therefore, we accept H0. Age does not influence how much they value storage factor.

Battery



Age Group (15-25): Strongly Agree ; Age group (26-35): Agree-Strongly Agree ; Age group (36-45): Strongly Disagree

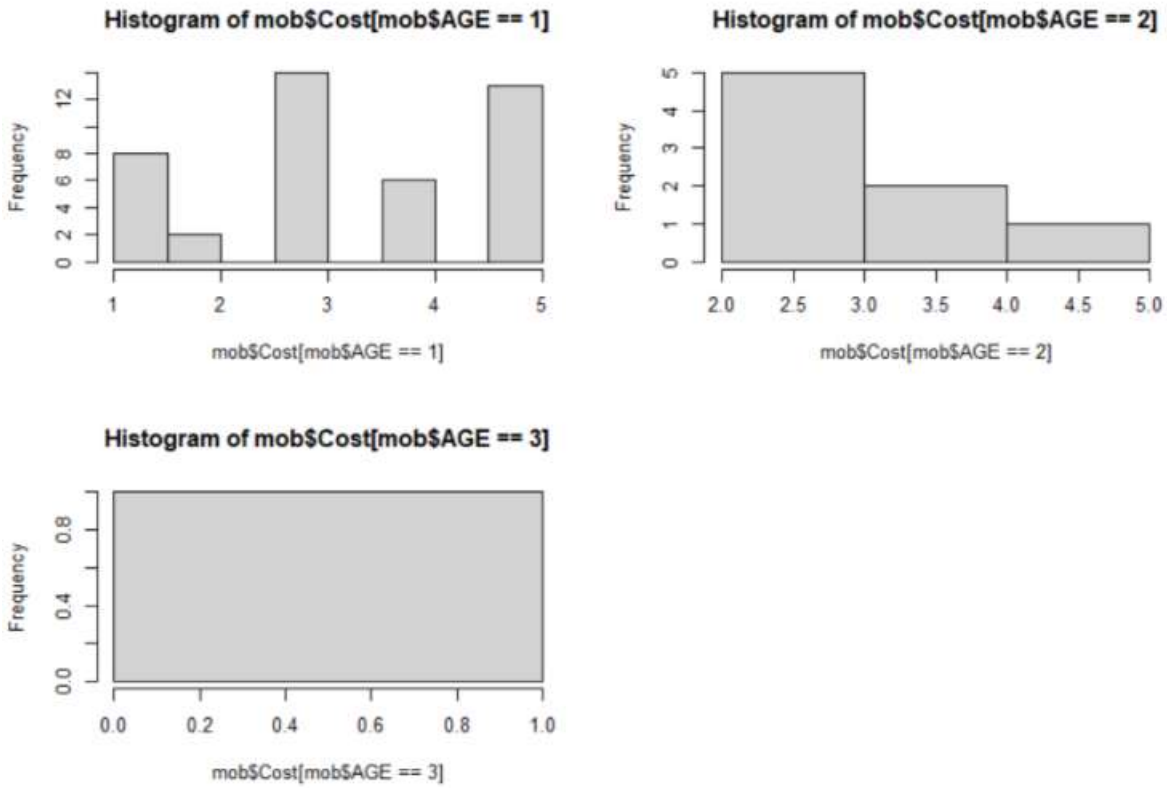
Does Age influence how much they value Battery life?

Null hypothesis (H0): Age does not influence how much they value Battery life. *Alternate Hypothesis (H1):* Age does influence how much they value Battery life. X-squared = 16.005, df = 8, p-value = 0.04231

Here, P-value < 0.05.

Therefore, we reject H0. Age does influence how much they value Battery life.

Cost



Age Group (15-25): Neutral; Age group (26-35): Disagree-Neutral;

Age group (36-45): Strongly Disagree

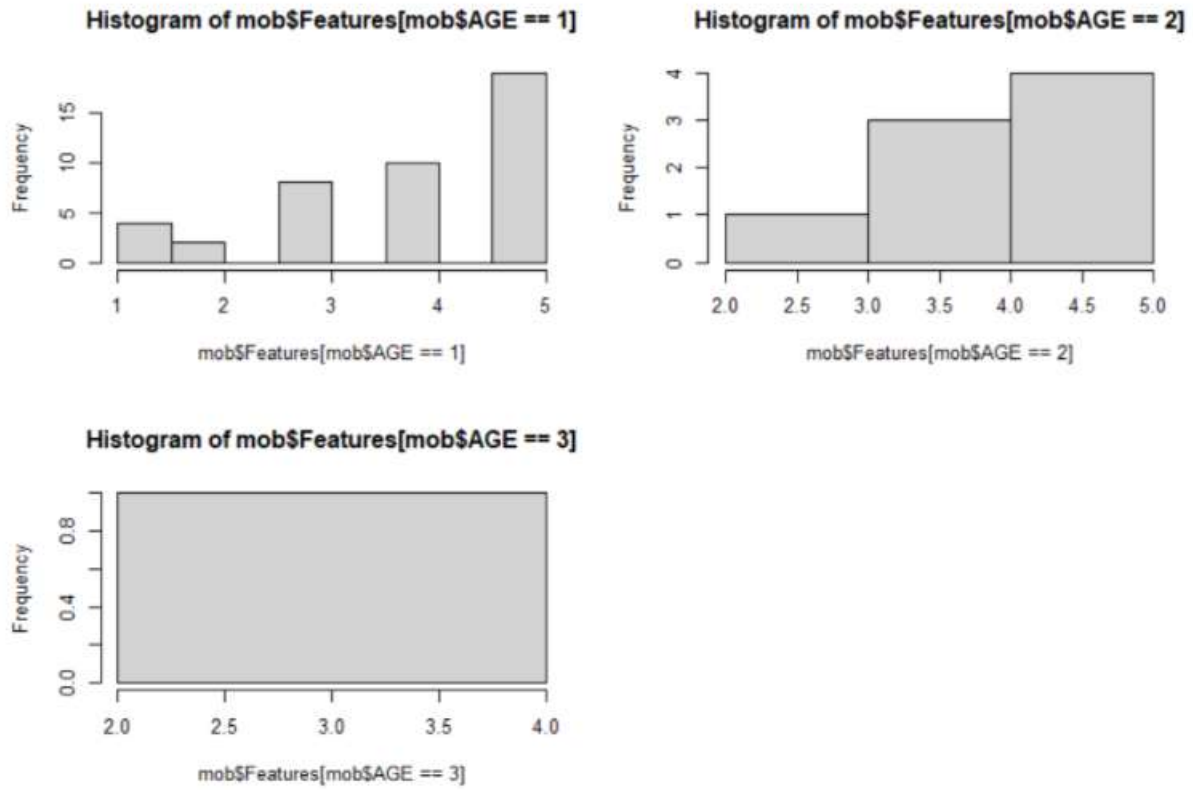
Does Age influence how much they value Cost factor?

Null hypothesis (H0): Age does not influence how much they value Cost factor. *Alternate Hypothesis (H1):* Age does influence how much they value Cost factor. X-squared = 11.224, df = 8, p-value = 0.1893

Here, P-value > 0.05.

Therefore, we accept H0. Age does not influence how much they value Cost factor.

Features



Age Group (15-25): Strongly Agree; Age group (26-35): Strongly Agree;

Age group (36-45): Agree

Does Age influence how much they value Features?

Null hypothesis (H0): Age does not influence how much they value Features.

Alternate Hypothesis (H1): Age does influence how much they value

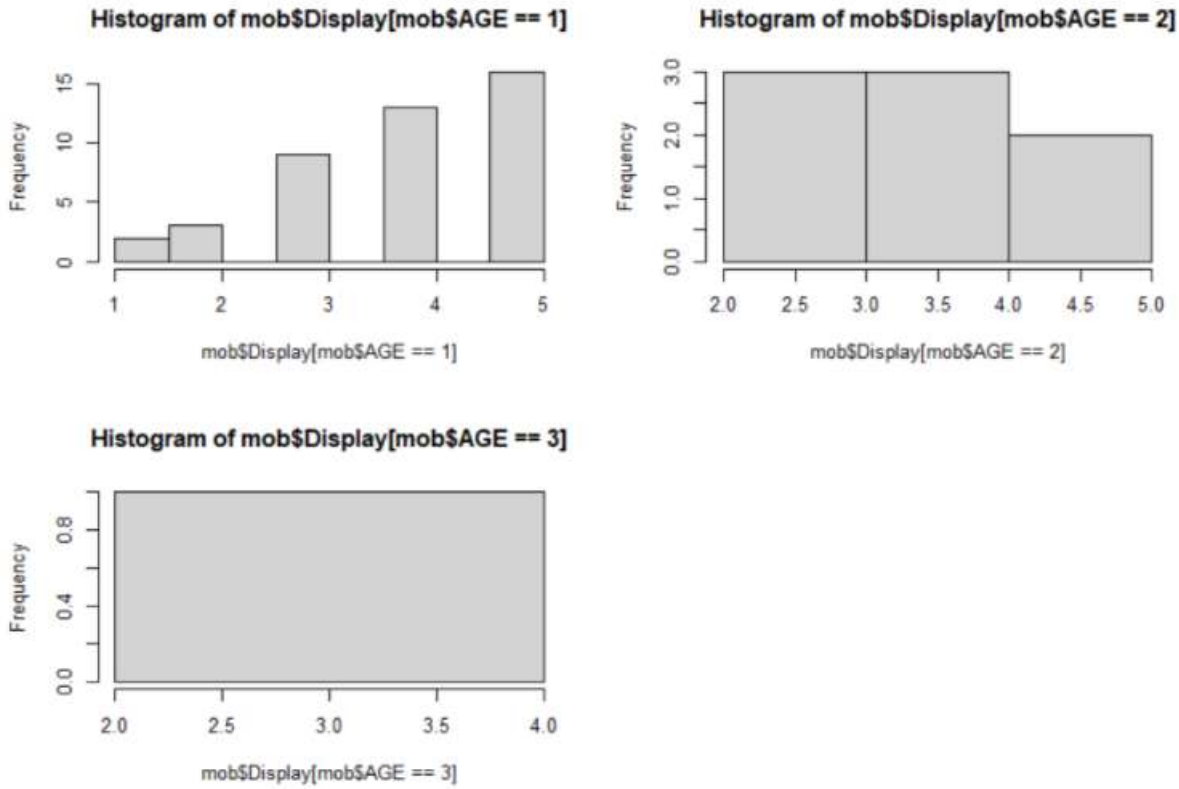
Features. X-squared = 8.2984, df = 8, p-value = 0.4049

Here, P-value > 0.05.

Therefore, we accept H0. Age does not influence how much they value

Features.

Display



Age Group (15-25): Strongly Agree; Age group (26-35): Disagree-Strongly Agree; Age group (36-45): Agree

Does Age influence how much they value Display?

Null hypothesis (H0): Age does not influence how much they value Display.

Alternate Hypothesis (H1): Age does influence how much they value Display.

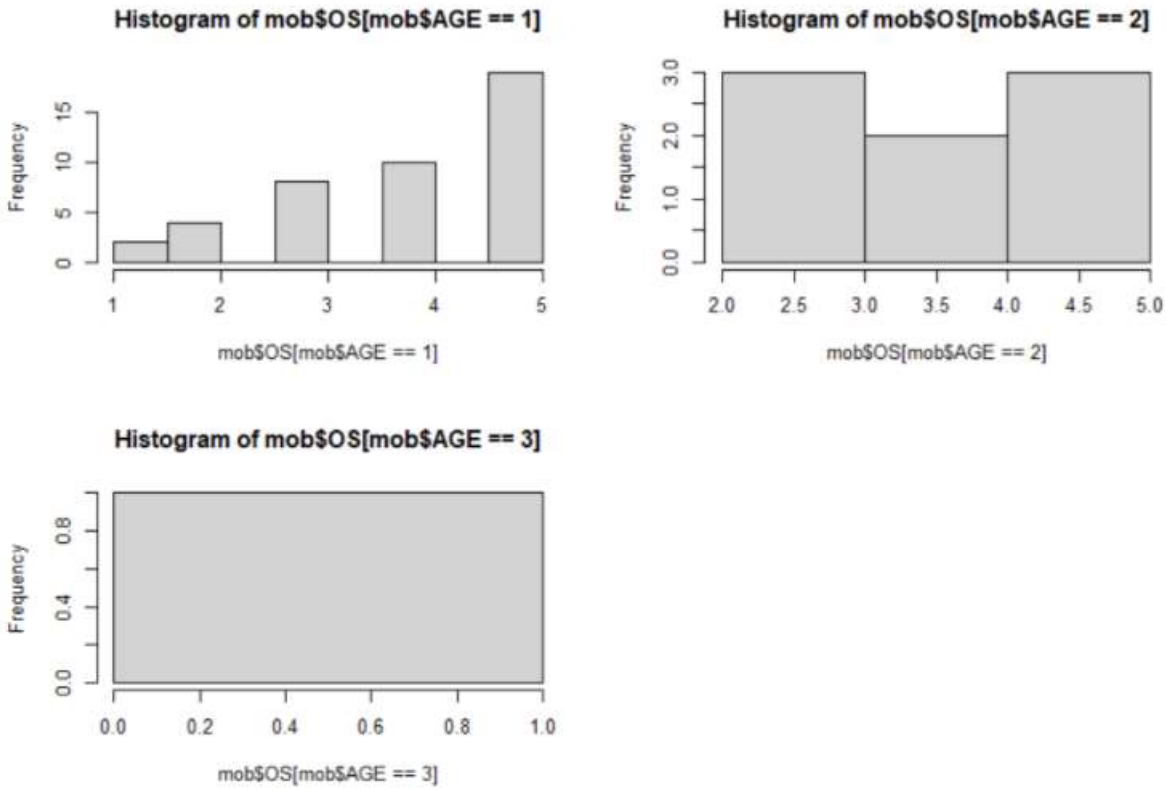
X-squared = 7.0913, df = 8, p-value = 0.5268

Here, P-value > 0.05.

Therefore, we accept H0. Age does not influence how much they value

Display.

OS



Age Group (15-25): Strongly Agree; Age group (26-35): Disagree-

Strongly Agree; Age group (36-45): Strongly Disagree

Does Age influence how much they value OS?

Null hypothesis (H0): Age does not influence how much they value OS.

Alternate Hypothesis (H1): Age does influence how much they value OS.

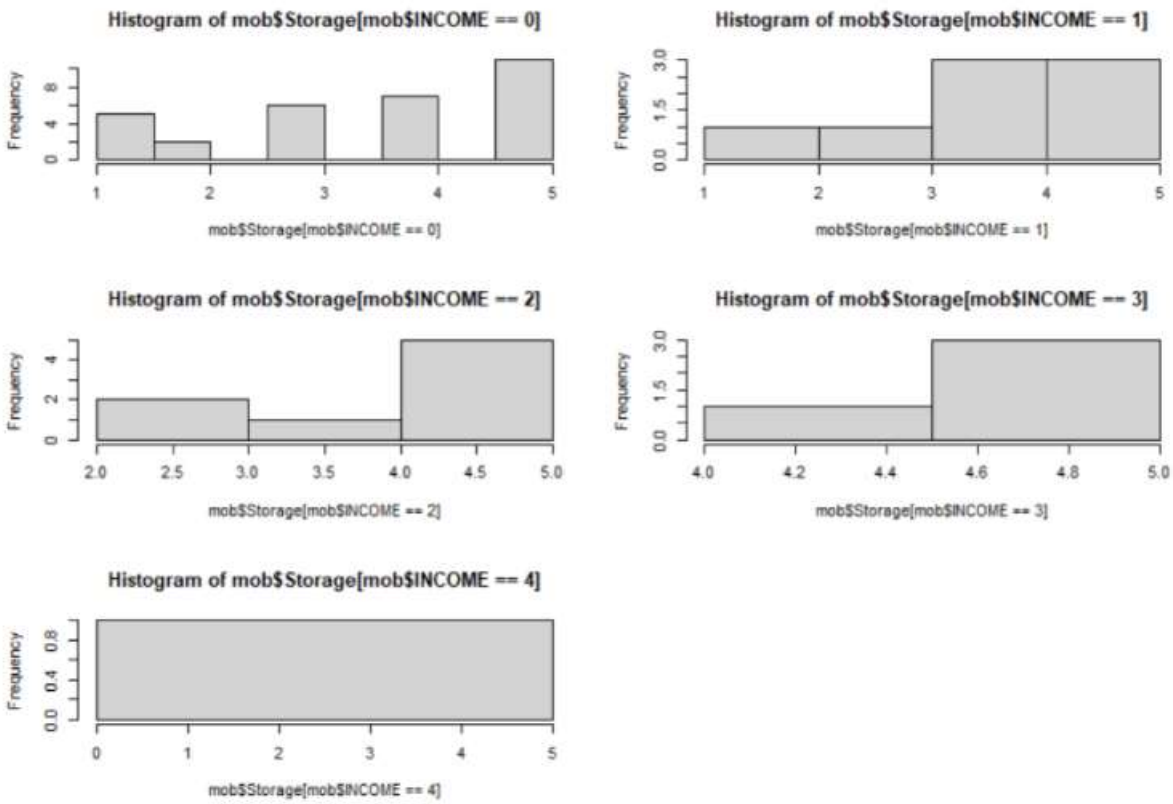
X-squared = 18.572, df = 8, p-value = 0.01732

Here, P-value < 0.05.

Therefore, we reject H0. Age does influence how much they value OS.

INCOME-LEVEL

Storage



Income Levels Level of Agreeableness
0 Strongly Agree
1L-3L Agree-Strongly Agree
3L-6L Strongly Agree
6L-9L Strongly Agree
9L-12L Strongly Agree

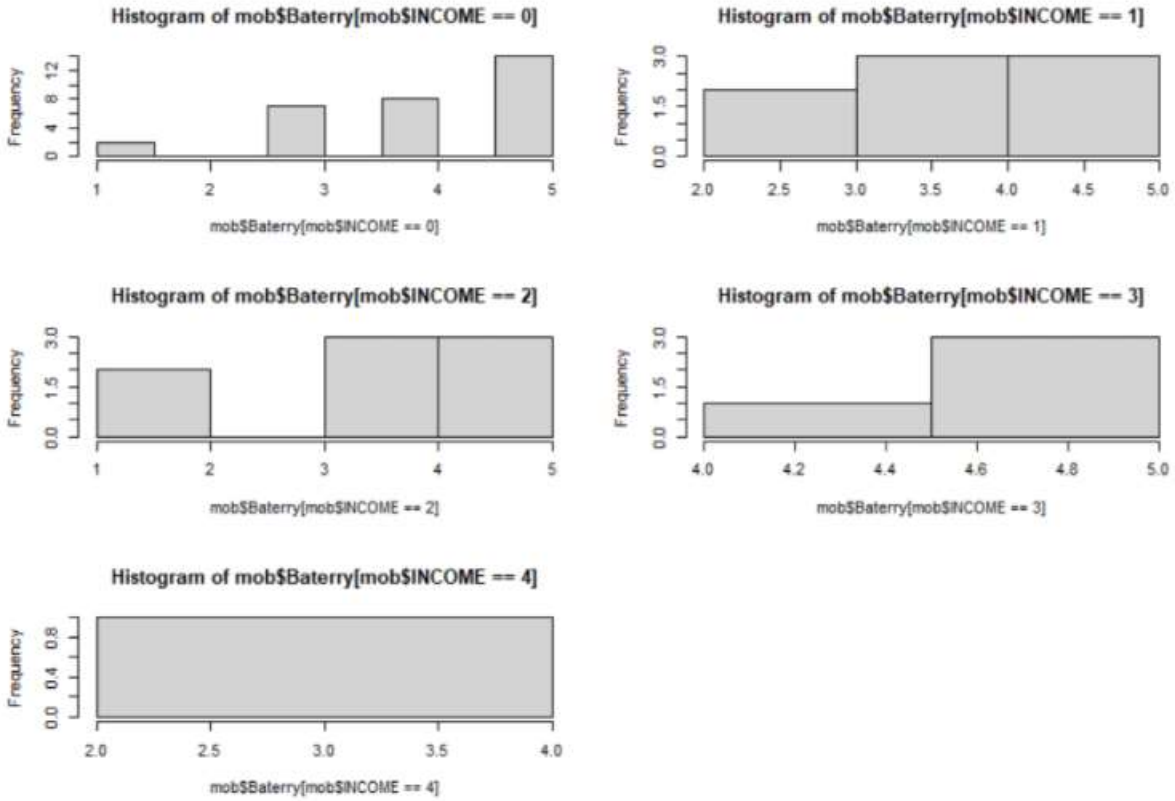
Does Income level influence how much they value storage factor?

Null hypothesis (H0): Income level does not influence how much they value storage factor. *Alternate Hypothesis (H1):* Income level does influence how much they value storage factor. X-squared = 14.717, df = 16, p-value = 0.5454

Here, P-value > 0.05

Therefore, we accept H0. Income level does not influence how much they value storage factor.

Battery



Income Levels Level of Agreeableness
0 Strongly Agree
1L-3L Neutral-Strongly Agree
3L-6L Neutral-Strongly Agree
6L-9L Agree-Strongly Agree
9L-12L Agree

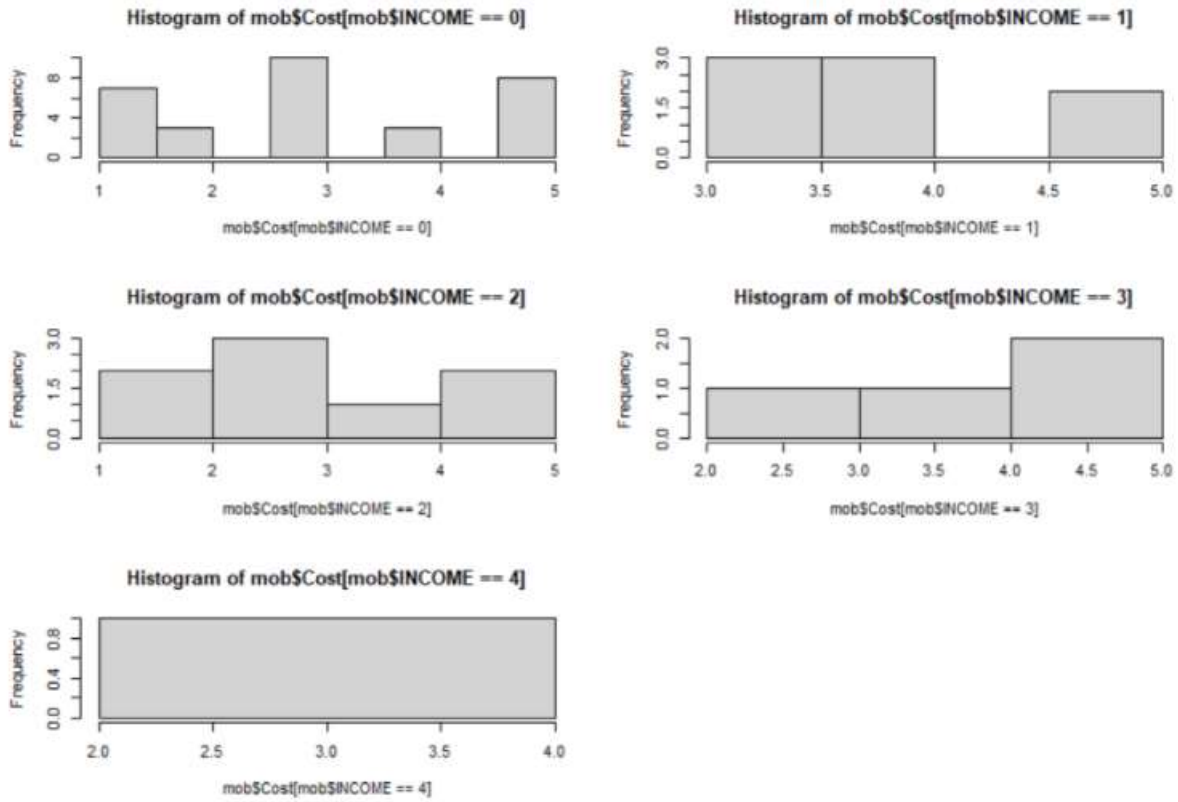
Does Income level influence how much they value Battery life?

Null hypothesis (H0): Income level does not influence how much they value Battery life. *Alternate Hypothesis (H1):* Income does influence how much they value Battery life. X-squared = 18.602, df = 16, p-value = 0.2898

Here, P-value > 0.05.

Therefore, we accept H0. Income level does not influence how much they value Battery life.

Cost



Income Levels Level of Agreeableness

0 Neutral
1L-3L Neutral-Agree
3L-6L Neutral
6L-9L Agree-Strongly Agree
9L-12L Agree

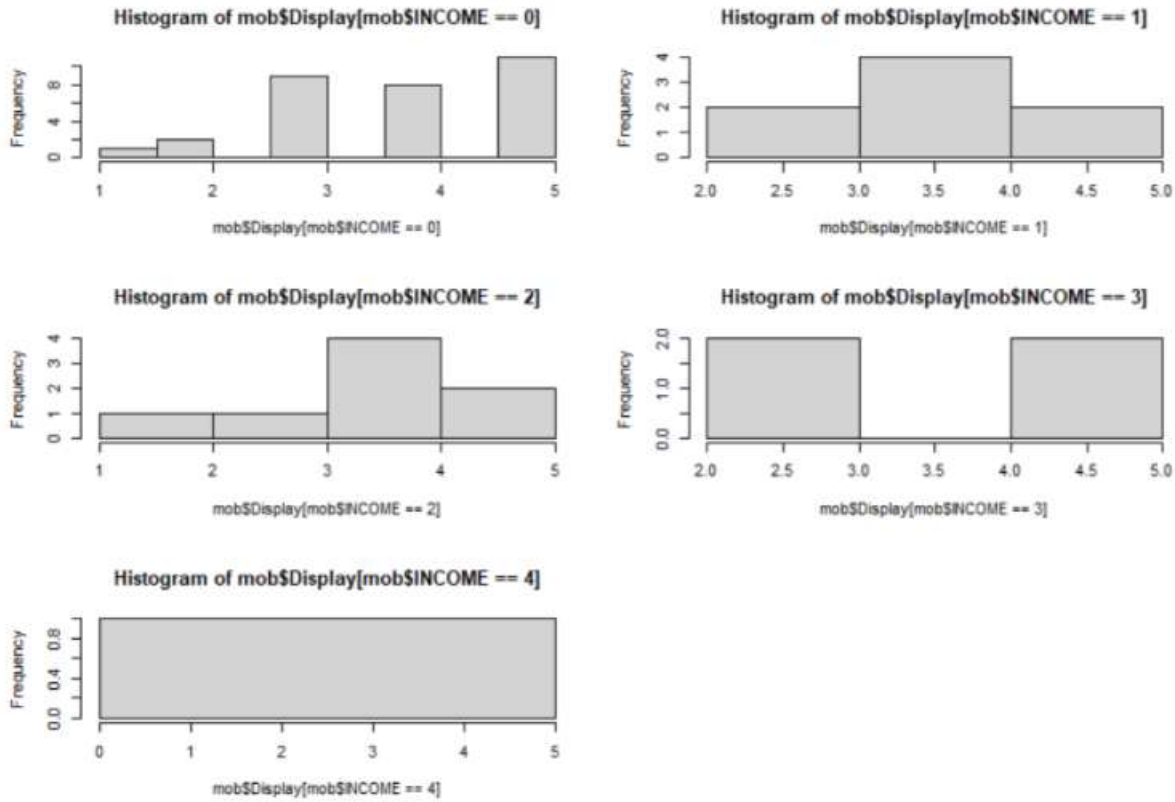
Does Income level influence how much they value cost factor?

Null hypothesis (H0): Income level does not influence how much they value cost factor. *Alternate Hypothesis (H1):* Income does influence how much they value cost factor. X-squared = 13.588, df = 16, p-value = 0.6294

Here, P-value > 0.05.

Therefore, we accept the H0. Income level does not influence how much they value cost factor.

Display



Income Levels Level of Agreeableness

0 Strongly Agree
1L-3L Neutral-Agree
3L-6L Neutral
6L-9L Agree-Strongly Agree
9L-12L Agree

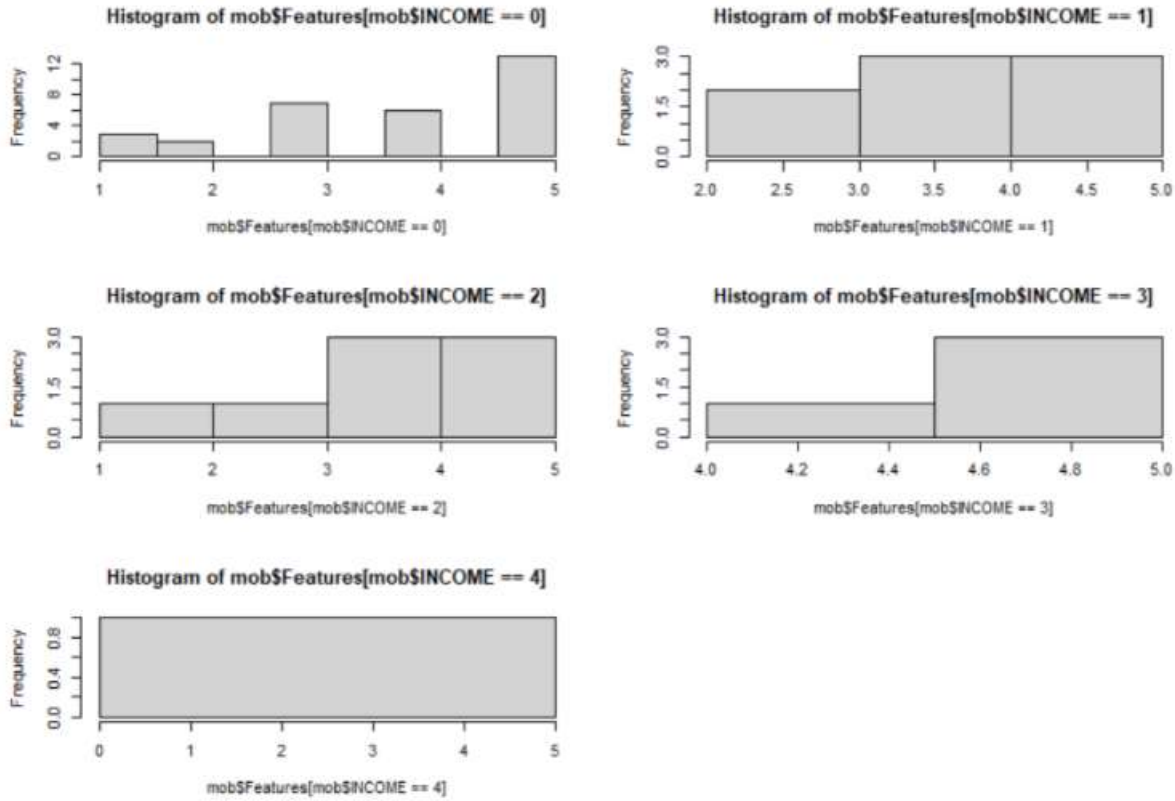
Does Income-level influence how much they value Display?

Null hypothesis (H0): Income-level does not influence how much they value Display. *Alternate Hypothesis (H1):* Income-level does influence how much they value Display. X-squared = 18.317, df = 16, p-value = 0.3057

Here, P-value > 0.05.

Therefore, we accept H0. Income-level does not influence how much they value Display.

Features



Income Levels Level of Agreeableness
0 Strongly Agree
1L-3L Neutral- Strongly Agree
3L-6L Neutral- Strongly Agree
6L-9L Strongly Agree
9L-12L Strongly Agree

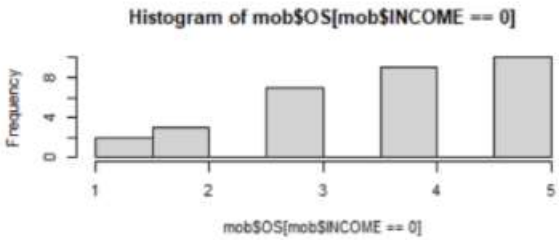
Does Income-level influence how much they value features?

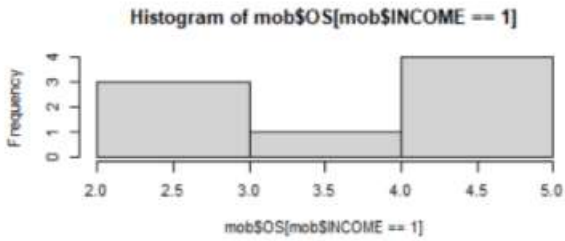
Null hypothesis (H0): Income-level does not influence how much they value features. *Alternate Hypothesis (H1):* Income-level does influence how much they value features. X-squared = 7.7848, df = 16, p-value = 0.955

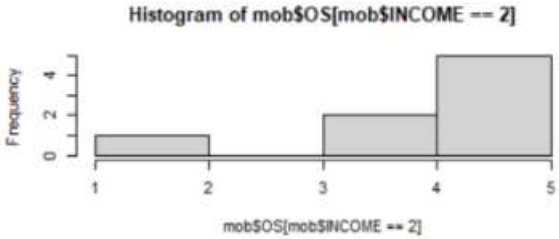
Here, P-value > 0.05.

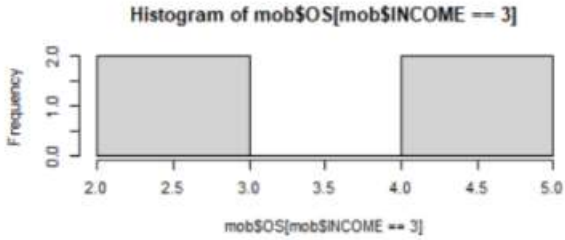
Therefore, we accept H0. Income-level does not influence how much they value features.

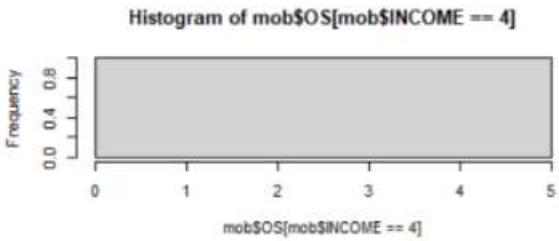
OS











Income Levels Level of Agreeableness
0 Strongly Agree
1L-3L Strongly Agree
3L-6L Strongly Agree
6L-9L Equally Disagree & Strongly Agree
9L-12L Strongly Agree

Does Income-level influence how much they value OS?

Null hypothesis (H0): Income-level does not influence how much they value OS. *Alternate Hypothesis (H1):* Income-level does influence how much they value OS. X-squared = 15.187, df = 16, p-value = 0.511

Here, P-value > 0.05.

Therefore, we accept H0. Income-level does not influence how much they value OS.

2) Regression (Multinomial Logistic Regression)

***Preparation of Dummy Variables*:**

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. Here, the categorical variables are recoded into a set of separate binary variables. This recoding is called "dummy coding".

```
#For Gender -dummies

mob$Gender=ifelse(mob$Sex=="Female",1,0)


#For age - dummy

mob$Aged1=ifelse(mob$Age=='15-25',1,0)

mob$Aged2=ifelse(mob$Age=='26-35',1,0)


#For Prof - dummies

mob$Prof1=ifelse(mob$Profession=='Student',1,0)

mob$Prof2=ifelse(mob$Profession=='Corporate
Services',1,0)

mob$Prof3=ifelse(mob$Profession=='Business',1,0)

mob$Prof4=ifelse(mob$Profession=='Government
Employee',1,0)
```

Here, dummies have been created for the Independent variable which are:

- Gender
- Age
- Profession
- Income

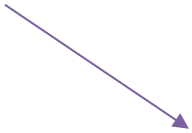
Making dummies makes analysis easier for the algorithm.

Preparing of Factors:

Factors are the data objects which are used to categorize the data and store it as levels. They are useful in data analysis for statistical modelling.

```
mob$income2=ifelse(mob$Income=='1L-3L',1,0)
```

```
mob$income3=ifelse(mob$Income=='3L-6L',1,0)
 mob$Switching_time=as.factor(mob$Switching_time)

mob$income4=ifelse(mob$Income=='6L-9L',1,0)
 nlevels(mob$Switching_time)

 ## [1] 3

 levels(mob$Switching_time)

 ## [1] "1-5 yrs" "More than 5 yrs" "Within an year."\
```



The Dependent variable i.e., Switching time has been factored and turned into levels

Predicting Switching time using Multinomial Logistic Regression:

A Multinomial regression is used to predict the Switching Time on the basis of Customer Data provided. Multinomial regression an extension of binomial logistic regression. The algorithm allows us to predict a categorical dependent variable which has more than two levels. Like any other regression model, the multinomial output can be predicted using one or more independent variable.

The dataset contains information about 90 variables divided into three categories which are represented by 1 to 3 numbers. The dependent variable here is Switching Time.

Splitting:

The dataset is split into train and test using *sample_frac()* function from *{dplyr}* package.

```
library(dplyr)
```

```
 ## Warning: package 'dplyr' was built under R version 4.0.2

 ##
 ## Attaching package: 'dplyr'

 ## The following objects are masked from 'package:stats':
 ##
 ## filter, lag

 ## The following objects are masked from 'package:base':
 ##
 ## intersect, setdiff, setequal, union
```

```
# Using sample_frac to create 60 - 40 slipt into test and train
train <- sample_frac(mob, 0.6)
sample_id <- as.numeric(rownames(train)) # rownames() returns character so
as.numeric
test <- mob[-sample_id,]
```

Relevelling:

When we build logistic models, we need to set one of the levels of the dependent variable as a baseline. We achieve this by using **relevel()** function.

```
train$Switching_time <- relevel(train$Switching_time, ref
=3) nlevels(train$Switching_time)
```

```
## [1] 3
```

```
str(train$Switching_time)
```

```
## Factor w/ 3 levels "Within an year.",..: 1 2 2 2 3 1 2 1 3 2 ...
```

Training the Dataset:

After the baseline we load the {nnet} package which contains the multinomial function. Following which we use **multinom()** function to fit the model and then use **summary()** function to explore the beta coefficients of the model.

```
# Loading the nnet package
require(nnet)
```

```
## Loading required package: nnet
```

```
# Training the multinomial model
#Model1
multinom.fit <- multinom(Switching_time ~ Sex-1, data = train) #Sw time
wi th gender
```

```
## # weights: 12 (6 variable)
## initial value 59.325064
## iter 10 value 42.091990
## iter 20 value 42.065166
## iter 30 value 42.063353
## final value 42.063352
## converged
```

```
# Checking the model
summary(multinom.fit)
```

```
## Call:
## multinom(formula = Switching_time ~ Sex - 1, data =
train) ##
## Coefficients:
##  SexFemale  SexMale  SexPrefer not to say  ##  1-5 yrs
1.029619 2.4423518 -10.07913 ## More than 5 yrs -0.223145
0.9162963 -10.07913 ##
## Std. Errors:
##  SexFemale  SexMale  SexPrefer not to say  ##  1-5 yrs
0.5209880 0.7372113 154.4126 ## More than 5 yrs 0.6708206
0.8366613 154.4126 ##
## Residual Deviance: 84.1267
## AIC: 96.1267
```

The output of the model is the log of odds. To get the relative risk IE odds ratio, we need to exponentiate the coefficients.

```
exp(coef(multinom.fit))

##  SexFemale  SexMale  SexPrefer not to say  ##  1-5 yrs
2.7999999  11.500055  4.194585e-05  ##  More  than  5  yrs
0.7999988 2.500014 4.194585e-05
```

```
head(probability.table <- fitted(multinom.fit))

## Within an year. 1-5 yrs More than 5 yrs
## 1 0.21739137 0.6086958 0.1739128
## 2 0.21739137 0.6086958 0.1739128
## 3 0.21739137 0.6086958 0.1739128
## 4 0.06666636 0.7666668 0.1666668
## 5 0.06666636 0.7666668 0.1666668
## 6 0.06666636 0.7666668 0.1666668
```

The model is checked for accuracy by building classification table. So, a classification table is made for training dataset and the model accuracy is calculated.

```
#Predicting accuracy of the model
# Predicting the values for train dataset
train$precticed <- predict(multinom.fit, newdata = train, "class")

# Building classification table
ctable <- table(train$Switching_time, train$precticed)
```

```
# Calculating accuracy - sum of diagonal elements divided by total
obs round((sum(diag(ctable))/sum(ctable))*100,2)
```

```
## [1] 70.37
```

We now repeat the above on the unseen dataset that tests dataset. ***Testing the Dataset***:

```
#Testing

multinom.fit <- multinom(Switching_time ~ Sex-1, data = test) #Sw time
with gender


# Checking the model

summary(multinom.fit)

exp(coef(multinom.fit))

head(probability.table <- fitted(multinom.fit))


#Predicting accuracy of the model


# Predicting the values for train dataset

test$precticed <- predict(multinom.fit, newdata = test, "class")


# Building classification table

ctable <- table(test$Switching_time, test$precticed)


# Calculating accuracy - sum of diagonal elements divided by total

obs round((sum(diag(ctable))/sum(ctable))*100,2)
```

```
## [1] 58.33
```

```
#15.7% less
```

The accuracy of the test dataset turns out to be *15.7%* less as compared to training dataset. So, we have a problem of overfitting here. We keep running models until we find the least difference between the training accuracy and testing accuracy

The results of the test are represented in the table below.

MODEL	FACTORS	TRAINING DATA ACCURACY	TESTING DATA ACCUARCY	DIFFERENCE
M1	Gender	74.04%	58.33%	15.7%
M2	Gender, Age, Income	79.63%	61.11%	18.5%
M3	Age, Profession	72.22%	61.11%	11%
M4	Gender , Age	74.02%	61.11%	13%

Model M3 containing factors variables Age and Profession comes out to be the best model.

Customer Retention

	Apple	Motorola	OnePlus	Oppo	RealMe	Samsung	Vivo	Xiaomi	
	2	1	0	0	0	0	1	0	0
Apple	0	4	0	2	0	0	1	0	0
Asus	0	0	0	0	0	0	1	0	0
Motorola	0	0	2	0	0	0	1	0	1
OnePlus	0	0	0	0	0	0	0	0	1
Oppo	0	0	0	0	1	0	0	0	0
RealMe	0	0	0	0	0	1	0	0	0
Samsung	0	4	0	3	0	0	6	0	4
Vivo	0	0	0	1	0	0	1	2	0
Xiaomi	1	0	0	3	0	2	4	2	0

- Current Apple users:
  4 people had old and new phone as Apple, while other 4 had old as Samsung and current apple

- Current Motorola users:
  2 people have Motorola as old and current phone.

- Current One plus users
  None had same brand before, 9 in total shifted to one plus from brands like Apple (2), Samsung (3), Vivo (1) and Xiaomi (3)

- Current Oppo users:
  1 person - stayed with the same brand

- Current Realme users
  1 person stayed with the same brand, 2 of them shifted from Xiaomi to Realme

• Current Samsung users
  6 of them stayed with the same brand
  9 shifted from other brands to Samsung

• Current vivo users
  2 of the remained
  2 of them shifted from Xiaomi

• Current Xiaomi users
  4 from Samsung ,1 from Motorola , 1 from OnePlus

Company	Retention Rate(%)
Samsung	76.923
Apple	85.47
Vivo	96.153
RealMe	96.153
OnePlus	0
Oppo	48.07
Xiaomi	0

Conclusion

Based on the responses obtained from our survey and after applying the testing and interpretations, we could come up with a general idea regarding the extent to which people value each of the factors like storage, cost, battery life, display, Operating system and other factors respectively while switching over to a new phone. Upon testing the dependence of these factors on gender, occupation, age and income levels of the respondents, various conclusions were drawn for the correlations existing between the variables.

As for Regression, we took Switching time as the dependent variable and the demographic factors as the predictors. Since, switching time has more than two levels, we used multinomial logistic regression. Various models where built and the testing and training accuracies were checked. The best model we found was model M3. Hence, it can be concluded that the combination of age with profession is the best predictor of Switching time.

Finally, we created a table to understand the retention trend to see which mobile brand has the highest retention rate. As seen from the table, brands Vivo and RealMe have the highest retention rates. And Xiaomi and OnePlus has the least retention rate.

Livelihood of Indian Households Before and During Lockdown Due to Covid-19 Pandemic

Submitted By-
Benedict Robin
Maheshwaran

Abstract:

The lockdown due to the pandemic has affected the lives of people in unimaginable ways. Data has been collected about the livelihood of people before and during the lockdown. Predictive analysis techniques have been used to Understand the effect of the lockdown due to Covid-19 pandemic on the livelihood of Indian people and how the survival instincts kick in varying the attributes of people's lives.

Introduction:

This pandemic has a greater impact on the livelihood of Indian households. We decided to do a study on them by collecting nominal data with respective to the lockdown status (i.e. prior to lockdown & during lockdown). We have analyzed this data to find the relationship between them and how far it is affecting the livelihood of the people. Data has been cleaned and sorted. The relationship between the different factors are been found by correlation and regression. The best model is built taking power difference and mental health as the dependent variable. Various tests are done to confirm that the model built is the best model.

Research Objective:

- To Understand the effect of the lockdown due to Covid-19 pandemic on the livelihood of Indian people and how the survival instincts kick in varying the attributes of people's lives.

- To build the best model to determine the factors influencing the difference in Power Consumption and the State of Mental Health of the people before and during the lockdown.

Methodology:

Quantitative analysis has been done by making use of Predictive Analytics. Since the pandemic was all of a sudden and something which hasn't happened in the recent past, there hasn't been many past research papers in the subject. Literature Review of the available past research papers on the topic has been referred to understand the impact of the lockdown on the lives of people.

Findings:

The Linear Regression finding includes the best model where the change in Power Consumption before and during the lockdown depends on the number of people at home, their AC, mobile and laptop consumption, their OTT usage, the preference of work and their petrol dependency.

The Logistic Regression finding includes the best model where the Mental Health of the people before and during lockdown depends on their OTT usage, petrol dependency, number of people working remotely from home, the fear of losing their job and the tendency to stock groceries.

It is also found that the data can be put into three relevant clusters and can be factored into two relevant factors.

Practical Implications:

Knowing how lives are affected when an unforeseen circumstance is encountered, helps individuals as well as organizations to plan ahead for the contingency.

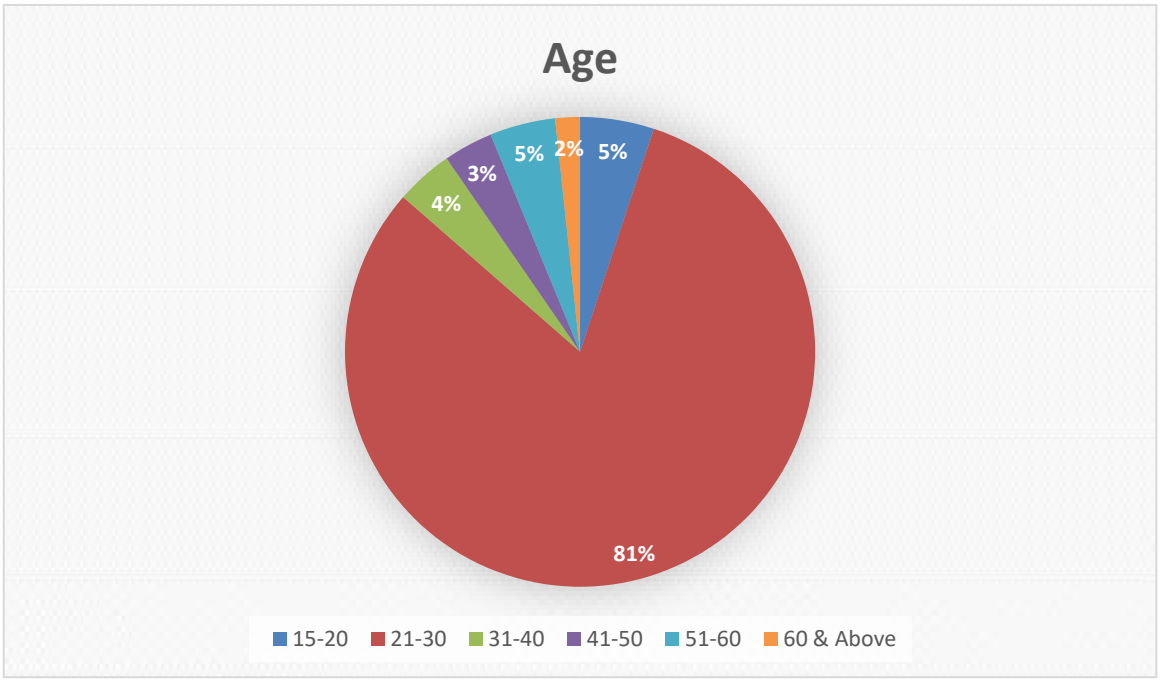
This study demonstrates the following research questions:

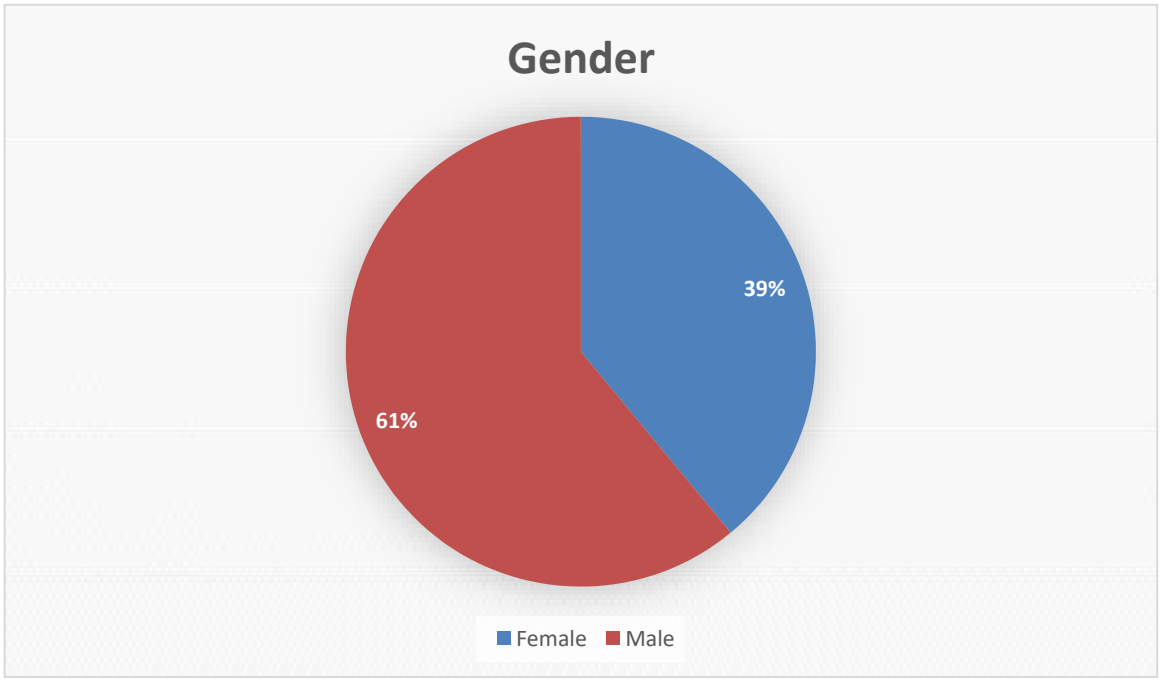
- What is the impact of lockdown on the lives of Indian households?
- How the best model that determines the changes in livelihood data and what affects thesame?

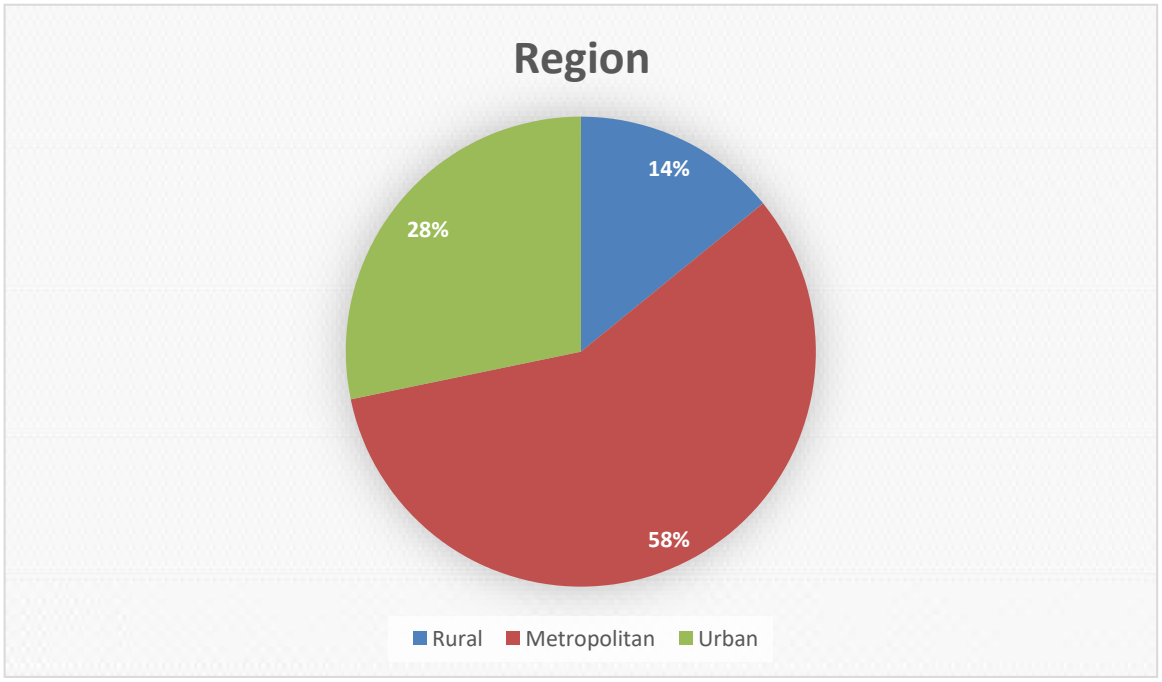
Research Methodology:

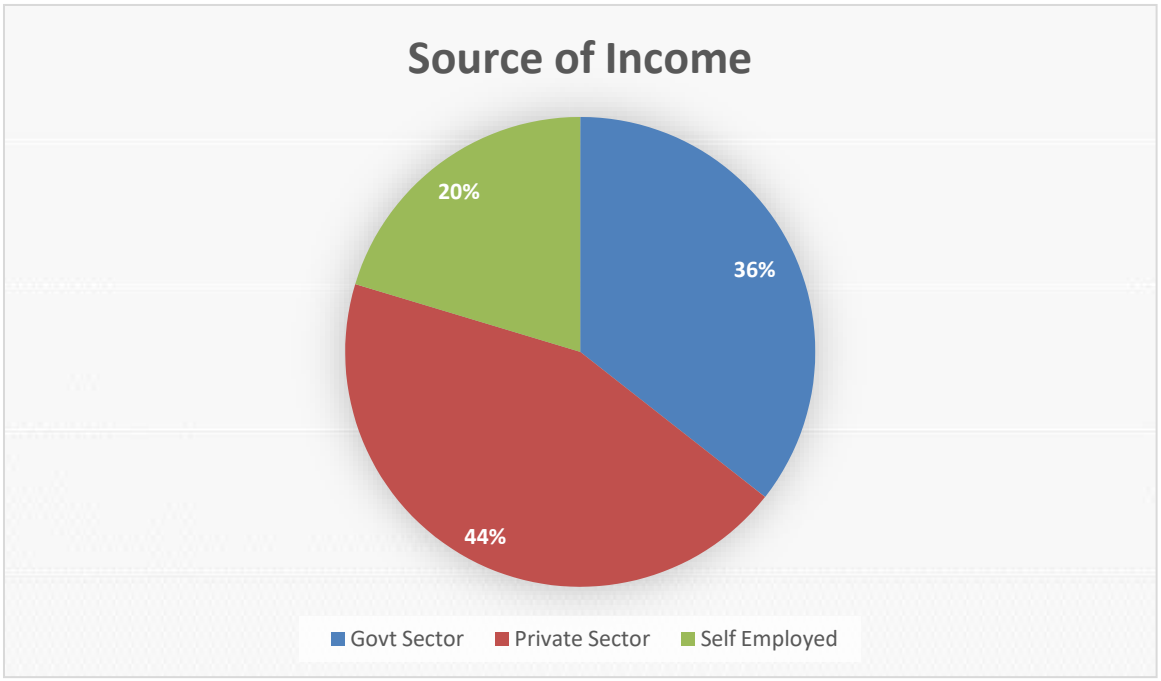
In this research we have conducted a google form online survey among 177 Indian households with both open ended and closed ended questions enquiring about the household data of the time before lockdown and during lockdown.

Demographic Profile:






Region
14%
28%
58%
Rural
Metropolitan
Urban


Source of Income
20%
36%
44%
Govt Sector
Private Sector
Self Employed

Predictive Analysis

Step 1: The data is read as corona and subsequent removal of columns are done based on thetype of study and the same has been named coronaa and coronanum.

corona<-read.csv("Master Data.csv")

coronanum<-corona[c(-1,-2,-3,-5,-9,-22,-23,-25:-30)]

coronaa<-corona[c(-6,-7,-10,-11,-13,-14,-16,-17,-19,-20)]

Step 2: There is no need to clean the data as there is no missing values.

table(complete.cases(coron

a))TRUE

177

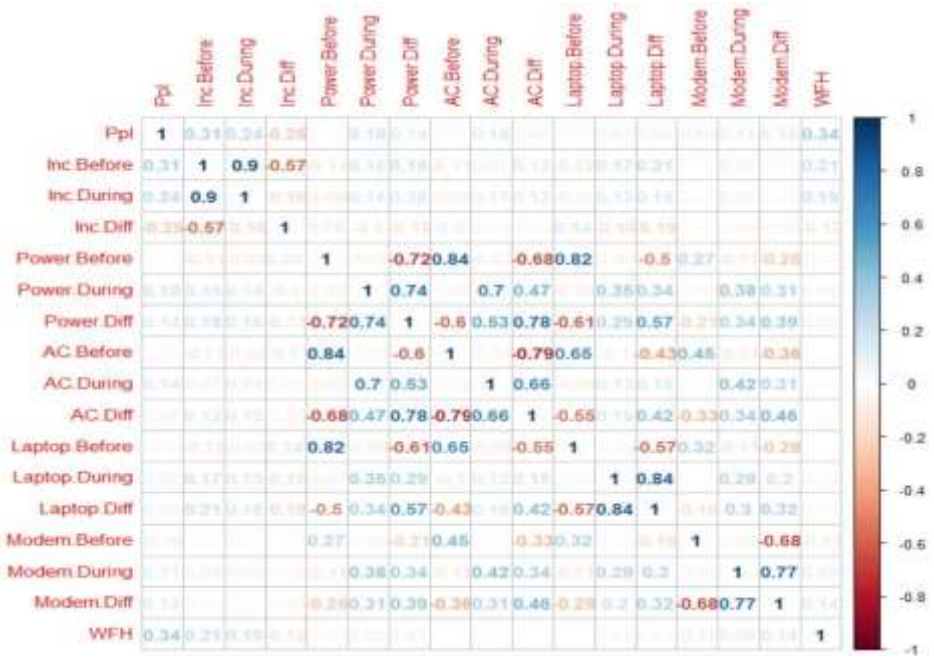
Step 3: Creating the models to find the best model.

Linear Regression: Here the difference in power consumption is taken as the dependent variable. All the possible combinations of dependent variables were tried and the best model with Adjusted R-square of 73.11% was found.

fit1<-
lm(Power.Diff~Ppl+OTT++AC.Diff+Laptop.Diff+Petrol+Work.pref,data=trai
ning)

The correlation of all the numerical datatype is found out.

corrplot(cor(coronanum),method='number')



Outlier Check:

#Ho: There are no outliers in the given data/model.

#H1: There are outliers in the given data.

outlierTest(fit1)

The 14 outliers were removed from the training data.

No Studentized residuals with Bonferroni p <

0.05Largest |rstudent|:

rstudent unadjusted p-value

Bonferroni p 169 3.627233

0.00042059

0.054677

Shapiro-Wilk Test for Normality

#Ho: The data is normally distributed.

#H1: The data is not normally distributed.

shapiro.test(residuals(object=fit1))

Shapiro-Wilk normality test

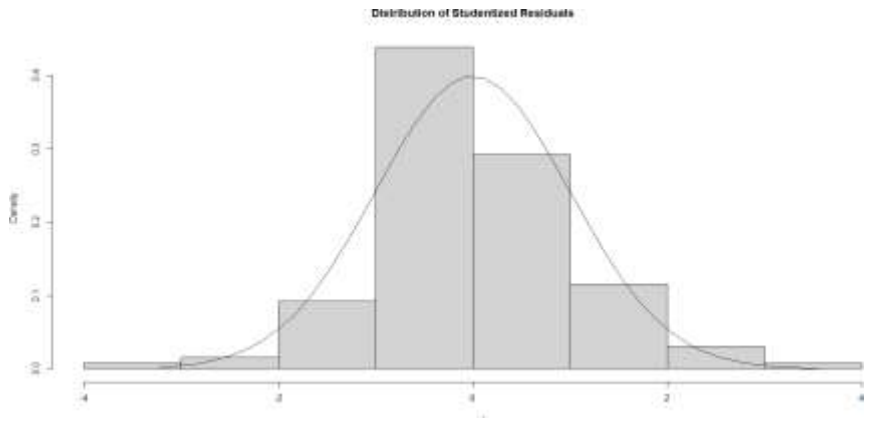
data: residuals(object =

fit1)

W = 0.98094, p-value = 0.06453

#Pvalue > 0.05. Thus, Ho is accepted. The Data is normal.

The plot of the residuals after removal of outliers:



Independence of Errors:

#Ho: Autocorrelation does not exist.

#H1: Autocorrelation exists.

durbinWatsonTest(fit1)

lag Autocorrelation D-W Statistic p-

value1 0.09113894

1.777567 0.188

Alternative hypothesis: rho != 0

#Pvalue >0.05. Hence autocorrelation does not exist.

Homoscedasticity:

#Ho: The variance of error is same across all the independent variables.

#H1: The variance of error is not same across all the independent variables.

ncvTest(fit1)

Non-constant Variance Score

TestVariance formula: ~

fitted.values

Chisquare = 8.49522, Df = 1, p = 0.0035608

Even though the null hypothesis is not accepted, fit1 is the best model considering all theother tests.

Multicollinearity:

#Ho: The independent variables aren't highly correlated with each other.

vif(fit1)

sqrt(vif(fit1

))>2

#The variance inflation factor of all the variables is less than 10. Thus, independent variablesaren't highly correlated with each other and the multicollinearity assumption is met.

Validation:

```
training$pred<-predict(fit1)

training$res<-residuals(fit1)

test$pred1<-predict(fit1,newdata =

test)test$res1<-test$Power.Diff-

test$pred1
```

Logistic Regression:

Here the data where people believe if their mental health has improved post lockdown is taken as the dependent data where 2 suggests that their mental health has improved while 1 suggests otherwise.

The best model was found with the least AIC of 152.45 and the same is named as

fitlog. *fitlog<-glm(formula = Mental.Health ~ OTT + Petrol + WFH + Fear +*

Groceries,family = binomial(), data = trainlog)

Predicting the values for test data:

pred<-predict(fitlog, newdata = testlog, type = "response")

Now the predicted values are in decimals of 0 to 1, but the dependent variable is categorical. Thus, we need to convert the pred values to 1s and 2s for which we need to find the cut off value.

Cut off Value:

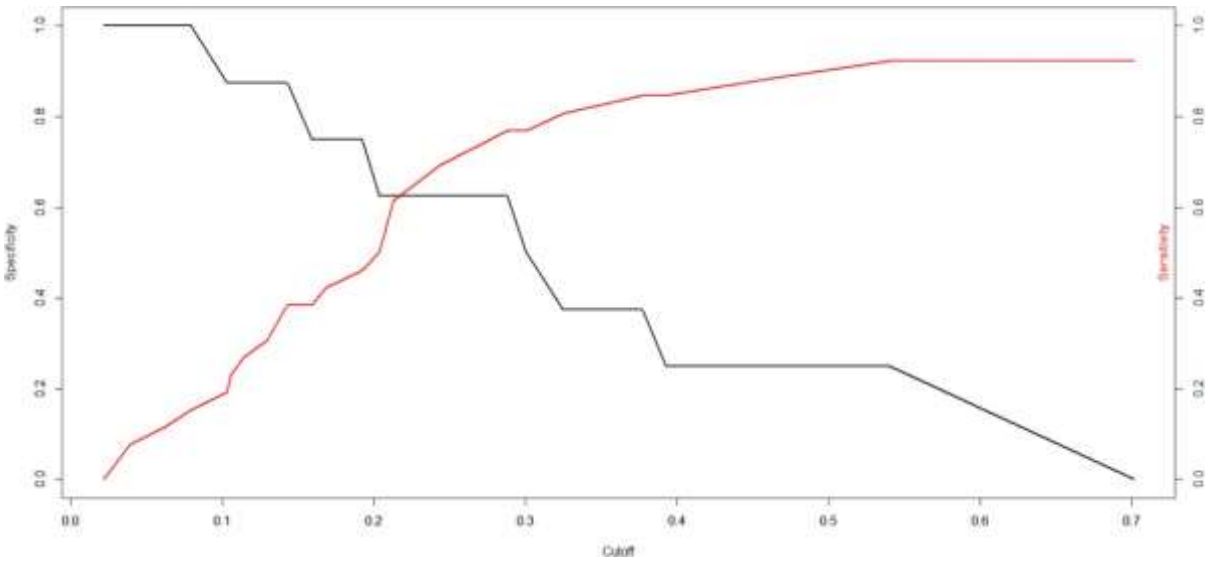
```
install.packages('R

OCR')

library(ROCR)

predictions<-prediction(pred,testlog$Mental.Health)

roc.pred=performance(predictions,measure='tpr',x.measure='

fpr') dist<-rep(9999, length(roc.pred@x.values[[1]]))

for(i in 1:

length(roc.pred@x.values[[1]])){

cur_x<- roc.pred@x.values[[1]][i]

  cur_y<- roc.pred@y.values[[1]][i]
```

```
  dist[i]<-(0-cur_x)(0-cur_x)+(1-cur_y)(1-cur_y) }

ideal_cutoff<-

roc.pred@alpha.values[[1]][dist==min(dist)]*10

ideal_cutoff
```

The cut off is found to be 0.2176732.

The cut off can also be found with the intersection of specificity and the sensitivity plots.

```
plot(unlist(performance(predictions,     "sens")@x.values),
unlist(performance(predictions,"sens")@y.values), type="l", lwd=2, ylab="Specificity",
xlab="Cutoff")

par(new=TRUE)


plot(unlist(performance(predictions,     "spec")@x.values),
unlist(performance(predictions,"spec")@y.values),type="l", lwd=2, col='red', ylab="",
xlab="")

axis(4, at=seq(0,1,0.2))

mtext("Sensitivity",side=4, padj=-2,

col='red')
```



Thus, the cut off value is 0.22.

Confusion Matrix:

convert<-ifelse(pred<0.22,"1","2")

conf<-data.frame(predicted=convert,

actual=testlog$Mental.Health)

conf$predicted=as.factor(conf$predicted)

res<-

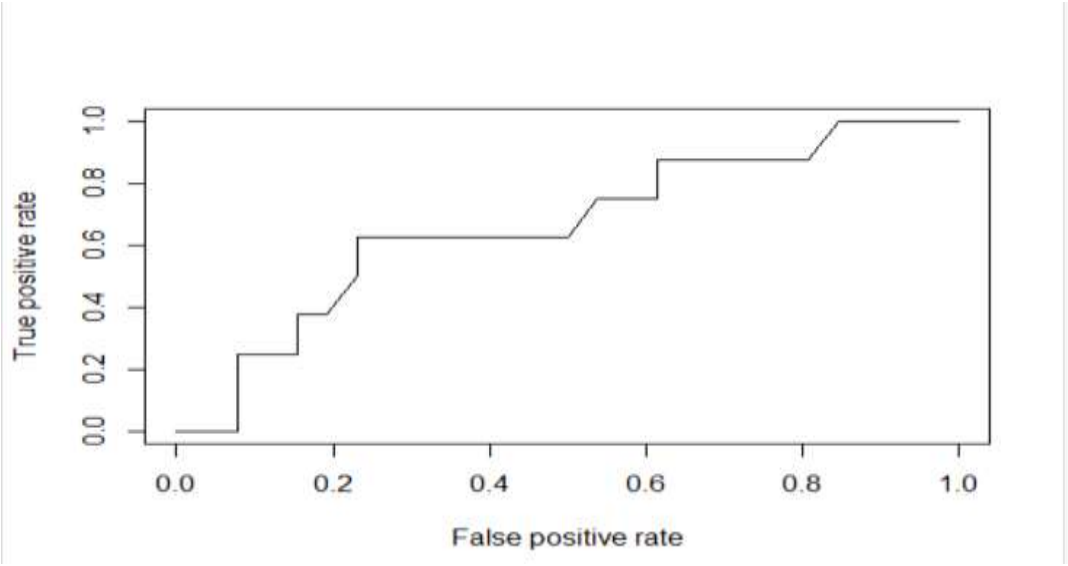
confusionMatrix(conf$predicted,conf$actual)

res

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2
         1 18  3
         2  8  5

               Accuracy : 0.6765
                 95% CI : (0.4947, 0.8261)
    No Information Rate : 0.7647
    P-Value [Acc > NIR] : 0.9174

                  Kappa : 0.2609

 Mcnemar's Test P-Value : 0.2278

            Sensitivity : 0.6923
            Specificity : 0.6250
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.3846
             Prevalence : 0.7647
         Detection Rate : 0.5294
   Detection Prevalence : 0.6176
      Balanced Accuracy : 0.6587

       'Positive' Class : 1
```

Area under the Curve:

plot(roc.pred)



auc=performance(predictions,measure='

auc')auc@y.values[{1}]

[[1]]

[1] 0.6610577

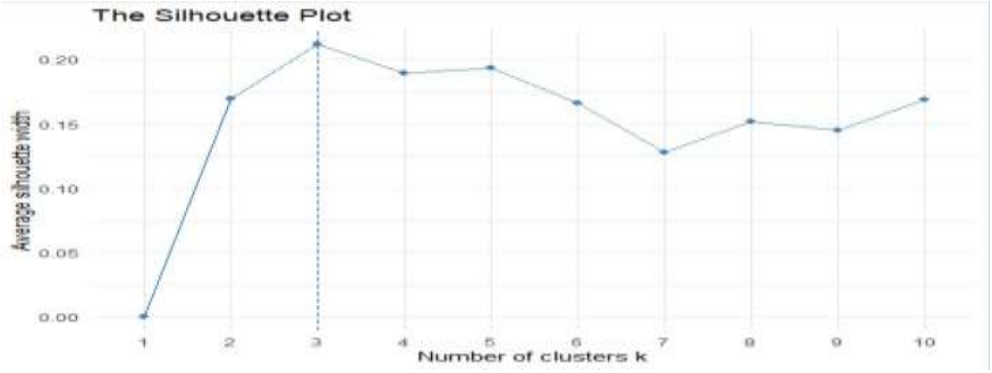
Cluster Analysis:

The data is scaled which helps to normalise the data within a particular range.

corscale<-scale(coronanumm)

Finding the ideal number of clusters:

fviz_nbclust(corscale, kmeans, method = "silhouette", k.max = 10) +theme_minimal() + ggtitle("The Silhouette Plot")



Thus, the ideal number of clusters is 3.

kmcor<-kmeans(coronanumm,3,nstart = 25)

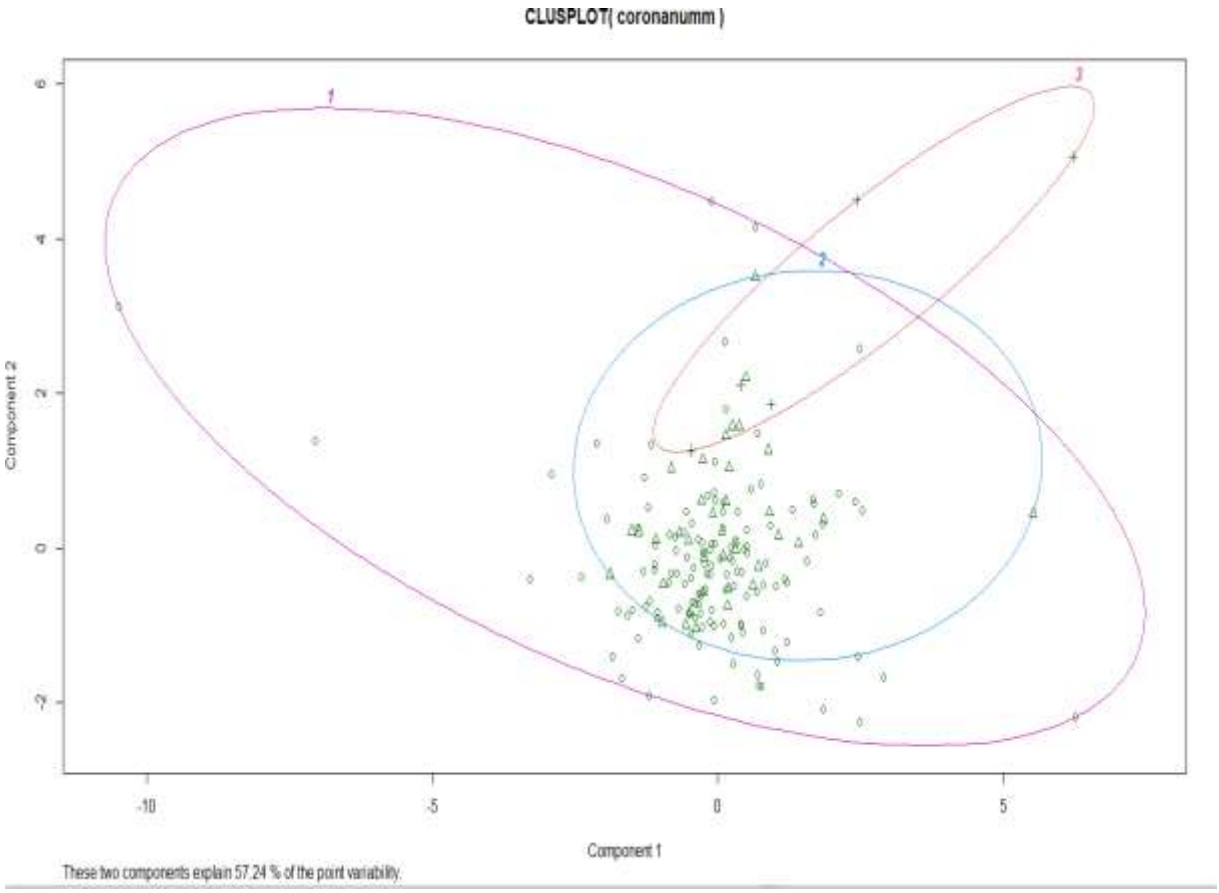
Cluster and it's components:

table(coronaa$Mental.Health,kmcor$cluster)

	Cluster 1	Cluster 2	Cluster 3
NO	101	30	3
YES	34	7	2

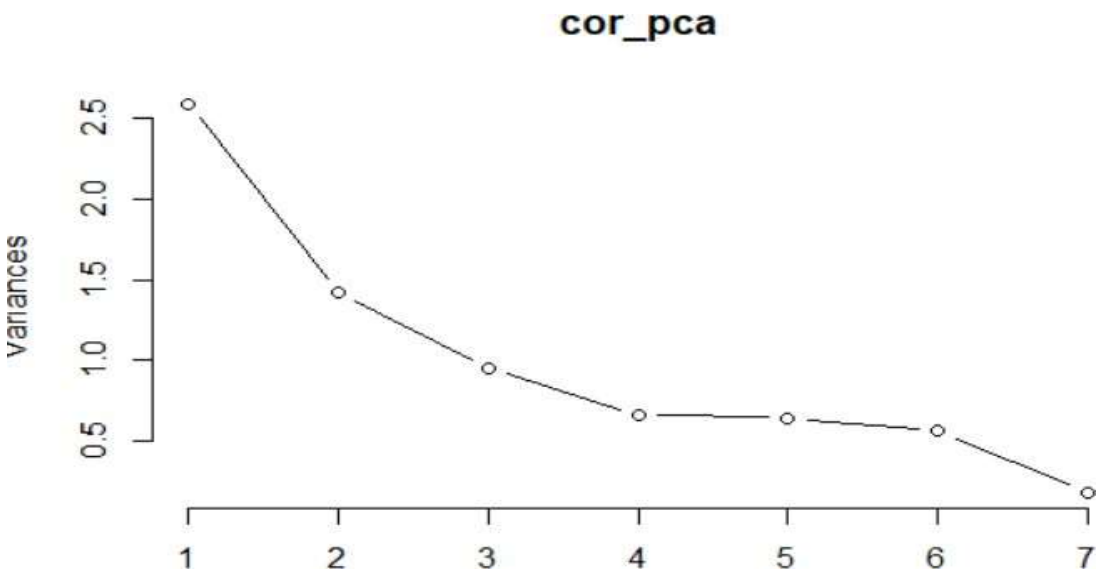
Cluster Plot:

clusplot(coronanumm,kmcor$cluster,color = T,labels = 4)



Factor Analysis of the Numeric datatype:

```
cor_pca<-
prcomp(coronanumm,center=TRUE,scale=TRUE)
summary(cor_pca)
plot(cor_pca, type="l")
```



Three components explain up to 70% of the variance.

```
cor_fact<-factanal(coronanumm,factors=3,rotation =
"none",trace=T)print(cor_fact,digit=2,cutoff=0.3,sort=TRUE)
```

Further it is reduced to 2 factors from the output.

1. Electricity Consumption(factor 1)
    - Power Diff
    - Laptop Diff
    - AC diff
    - Modem Diff
2. Work Related(factor 2)
    - Income Diff
    - WFH

Conclusion:

Thus, the models built show how the respective dependent variables, difference in power consumption and the mental health of the people before and during lockdown vary with respect to all the components. Whenever an uncertainty strikes and people are forced to lockthemselves at home, the predicted models can help us understand how each household component would vary and how it would affect the overall livelihood. This enables both individuals and organizations including governments to prepare the contingency plans.

Impact of Knowledge Management on IT Sector Performance

Submitted By-
Doel Bhattacharya
Kaushik Jyoti Talukdar

Abstract

Purpose:

An effective knowledge management is the key tool for driving organizational effectiveness and forms a key driver for organizational survival in long run, competitiveness and profitability.

- To understand the impact of knowledge management on employees' performance.
- To build best model and understanding the factors which are influencing the employees'performance.

Methodology:

Quantitative analysis has been done by making use of Predictive Analytics. Literature review of past research papers on this field to understand the holistic concept of knowledge management inenhancing IT sector performance.

Finding:

Finding includes that best model of knowledge management is depended on developing decisionmaking capability in an employee by proper coaching, senior leadership support provides an essential significance in developing knowledge management mechanisms in an organization, enhancing the scope of new learning, efficient knowledge transfer, focus shall be given on proper selection of tools, various challenges such as technical problems

shall be resolved. All these factors help in enhancing productivity of an employee which in turn impacts overall productivity of an organization. Budget allocation should be done wisely in order to make effective use of the organization's knowledge management resources.

Practical Implication:

Knowledge management also helps employees to become more flexible and enhances their job satisfaction. It also enhances employee adaptability and they are more likely to accept the change. Employees' market value is enhanced in relation to other organizations' employees.

Originality/Value:

The paper presents concepts of knowledge management and it is identified as the framework for crafting and designing an organization's strategy, and processes in order to create economic and social value for its employees as well as customers.

Keywords:

Knowledge management (KM), Senior leadership support, Productivity, Knowledge Transfer

Introduction:

An effective knowledge management has been used as the critical pillar for an IT organization which seeks to ensure sustainable strategic competitive advantage. It is the key driver of organizational effectiveness and forms a critical tool for organizational survival in long run, competitiveness and profitability. Therefore, organizations have realized the importance of creating, managing, sharing and utilizing knowledge effectively. It is identified as the framework for crafting and designing an organization's strategy, and processes in order to create economic and social value for its employees as well as customers. Successful organizations have now understood the importance of managing knowledge, developing plans so as to accomplish this objective and devoting time and energy to these efforts. (Omotayo, 2015) Knowledge depends on the action of human and results from the interaction among insights, judgement and intuition regarding information, which is being influenced by the innovation and the user experience. (Rodrigo Valio Dominguez Gonzalez1, 2014) Knowledge management also helps to enhance talent management which basically deals with attracting, developing and retaining the key talent of an organization. This concept of talent management has been a great value addition, employee retention and employee engagement. (Mohammed A Abusweilem, 2019) The study reveals a positive association between constructive feedback and customer oriented service as well as relation between organizational strategy and customer focused strategy. In this research we have incorporated quantitative analysis by performing various hypothesis testing in order to understand the impact of various attributes such as virtual platform experience, liberty to access details from said department, senior leadership support, constructive feedback, customer service, new learning,

business strategy, knowledge transfer and self-upskilling on IT employees' performance as well as organizational effectiveness.

Significance:

The goal of Knowledge Management is to enable organizational learning and to create a learning culture, in which the sharing of knowledge is increased. When thinking about knowledge management, it is important to consider that specialized knowledge of employees should not leave with them or remain unutilized. It boosts an efficiency of an organization's decision – making capability. It helps in building smarter workforce, who are quick and able to make informed decisions. IT Organizations begin the process of knowledge management for following reasons such as encouraging teams to share expertise, the retirement of key individual employee could lead to capture their knowledge. It is also used in training of new employees. (Valamis, 2020) Various sources of knowledge management include gamification used in training employees, expert knowledge transfer sessions, tutorials, collaborative environment, learning and development environment, case studies, webinars etc. Knowledge management process takes place in four main steps which involves: the discovery process is understanding the knowledge flow of an organization, capturing knowledge by making use of technology, process which incorporates how knowledge can be best folded into the structure of an organization which includes establishing and

promoting a shift towards sharing of knowledge and developing employees as innovators. Knowledge sharing and learning which enhances better decision making.(Valamis, 2020) Processof knowledge transfer at different level of analysis are

- Individual level: Human resource is agent of learning.
- Network level: Structural position of firm relative to other network members.

Business strategy factors that drive knowledge management are competitor knowledge advantage, learning cycles and rate of dynamic learning and competitor learning cycles.

Various challenges are:

- Improper selection of knowledge management tool
- Technical problems
- Lack of experience for conducting knowledge transfer session
- Lack of Senior leadership support

This study demonstrates the following research questions:

- What is the impact of knowledge management on IT employees?
- How the best model of knowledge management does enhances the efficiency of an organization?

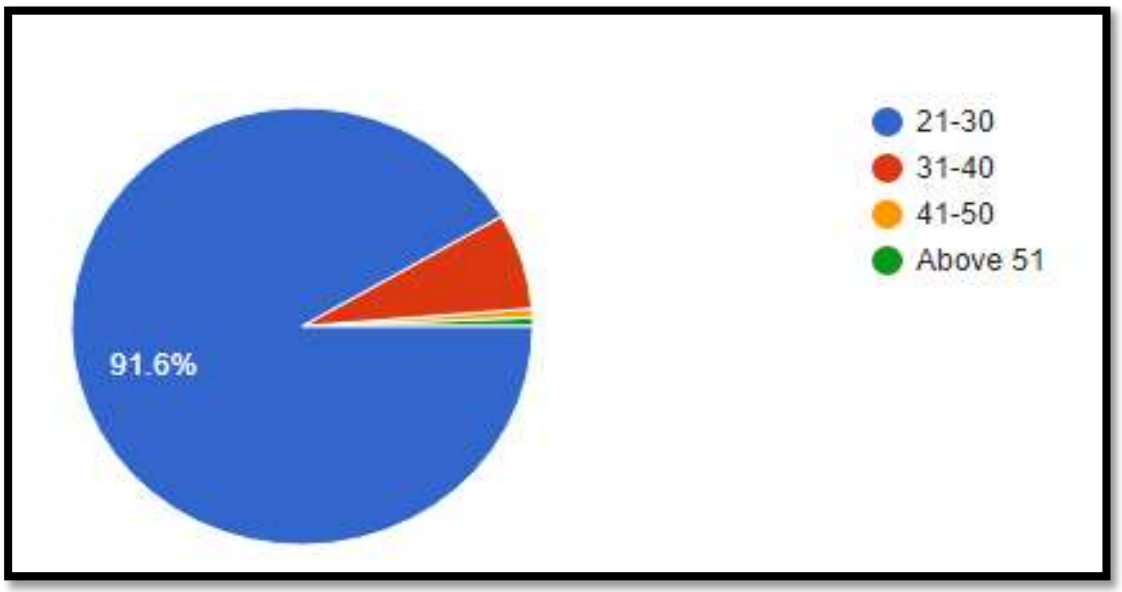
Research Objective:

- To understand the impact of knowledge management on employees' performance.
- To build best model and understanding the factors which are influencing the employees' performance.

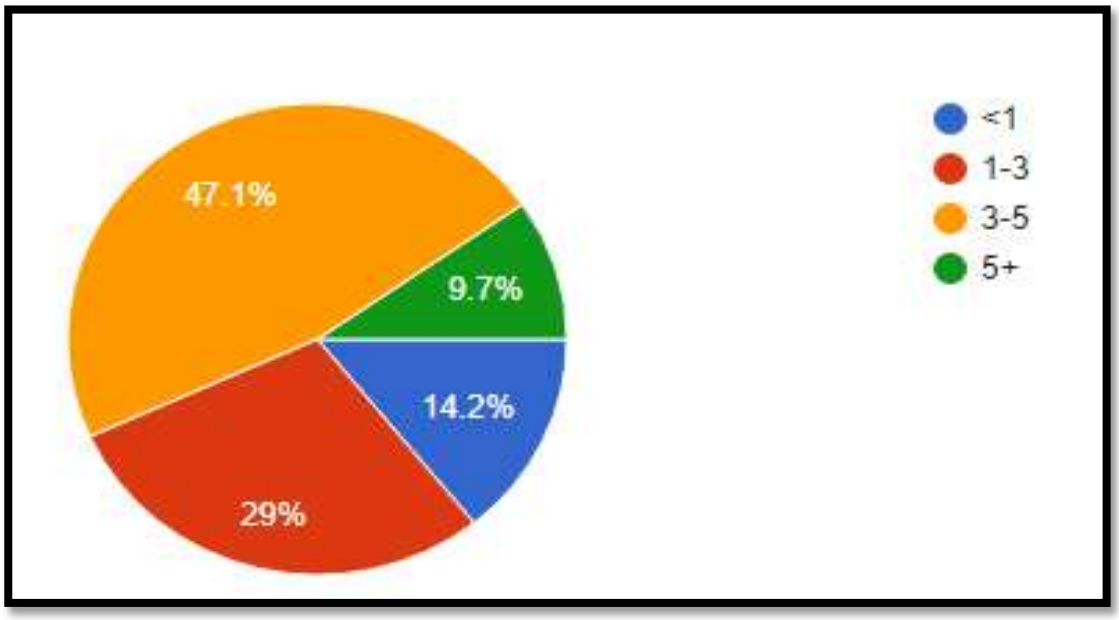
Methodology:

- In this research we have conducted a google form online survey among 150+ IT employees to understand the status of knowledge management in their respective organizations and how it is beneficial to enhance both organizational effectiveness and boosting employees' performance.
- Literature review has been done to understand the past research which has been done in esteemed research papers and to come up with innovative strategies in order to foster growth of an IT organization as well as upskilling of its employees.

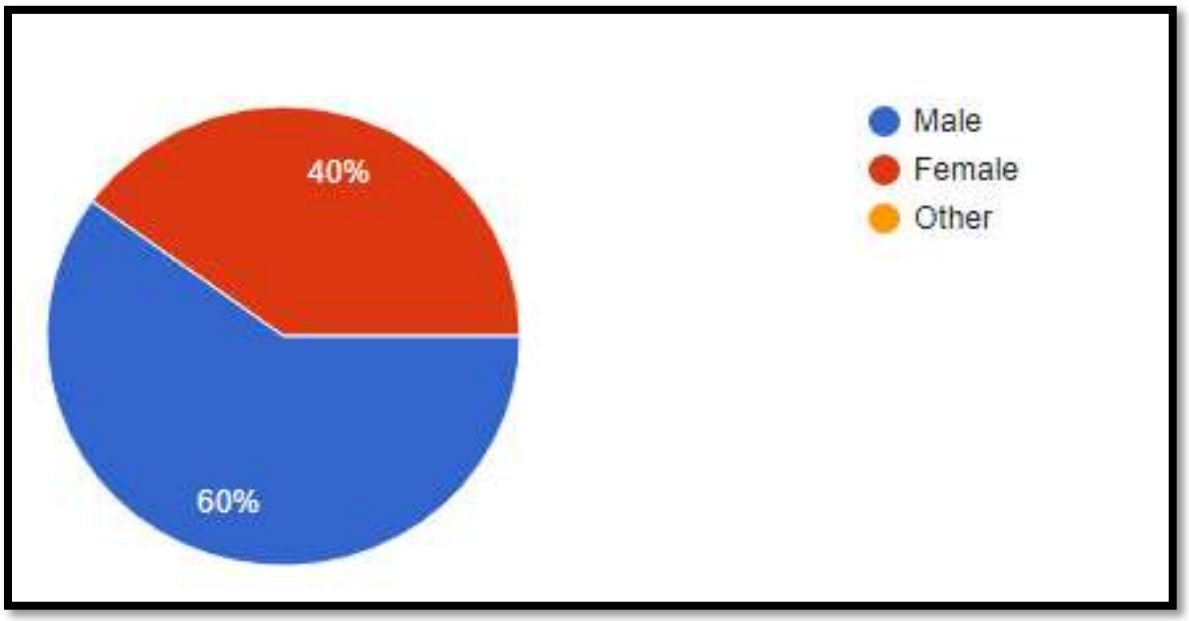
Demographic Profile:



AGE



YEARS OF EXPERIENCE



GENDER

Analysis and Finding:

```
getwd()
setwd("C:/Users/Doel bhattacharya/Documents/R
folder")km<-
read.csv("Knowledge_management.csv")
str(km)
summary(km)
```

Predictive Analytics

In this project KMservice which emphasizes Knowledge management services provided byan organization is taken as a dependent variable.

Step-1:

There is no requirement of cleaning of data as we don't have NAs in our dataset.

Step-2:

Converting categorical variables into dummy variables.

Converting categorical variables into dummy variables.

```
km$Kmdept<-ifelse(km$Kmdep=="Yes",1,0)

km$decmaking1<-ifelse(km$Decisionmaking=="Strongly Agree" | km$Decisionmaking
==  "Agree",1,0)
km$libertytoaccess1<-ifelse(km$libertytoaccess=="Strongly Agree" | km$libertytoaccess
==  "Agree",1,0)
km$Virtualplatformexp1<-ifelse(km$Virtualplatformexp=="Strongly Agree"
|  km$Virtualplatformexp == "Agree",1,0)
km$Seniorleadershipsupport1<-ifelse(km$Seniorleadershipsupport=="Strongly Agree"
|  km$Seniorleadershipsupport == "Agree",1,0)
km$Constructivefeedback1<-ifelse(km$Constructivefeedback=="Strongly Agree"
|  km$Constructivefeedback == "Agree",1,0)
 == "Agree",1,0)
km$Customerservice1<-ifelse(km$Customerservice=="Strongly Agree" |
km$Customerservice  == "Agree",1,0)
```

```
km$Newlearning1<-ifelse(km$Newlearning=="Strongly Agree" | km$Newlearning
==  "Agree",1,0)

km$Businessstrategy1<-ifelse(km$Businessstrategy=="Strongly Agree" |
km$Businessstrategy  == "Agree",1,0)

km$Knowledgetransfer1<-ifelse(km$Knowledgetransfer=="Strongly Agree"
|  km$Knowledgetransfer == "Agree",1,0)

km$Customerfocus1<-ifelse(km$Customerfocus=="Strongly Agree" | km$Customerfocus
==  "Agree",1,0)

km$Selfupskilling1<-ifelse(km$Selfupskilling=="Strongly Agree" | km$Selfupskilling
==  "Agree",1,0)

km$Improperselection1<-ifelse(km$Improperselection=="Strongly Agree"
|  km$Improperselection == "Agree",1,0)

km$Technicalproblem1<-ifelse(km$Technicalproblem=="Strongly Agree"
|  km$Technicalproblem == "Agree",1,0)

km$productivity1<-ifelse(km$productivity=="Strongly Agree" | km$productivity
==  "Agree",1,0)

km$KMservice1<-ifelse(km$KMservice=="Strongly Agree" | km$KMservice ==
"Agree",1,0)

km$overallproductivity1<-ifelse(km$overallproductivity=="Strongly Agree"
|  km$overallproductivity == "Agree",1,0)

View(km)
```

Step-3: Creation of models

```
m1<-lm(KMservice1~decmaking1, data = km)

summary(m1)
```

Multiple R-squared: 0.2541, Adjusted R-squared: 0.2417

```
m2<-lm(KMservice1~decmaking1+libertytoaccess1,

data = km) summary(m2)
```

Multiple R-squared: 0.2742, Adjusted R-squared: 0.2496

```
m3<-lm(KMservice1~decmaking1+Seniorleadershipsupport1, data =  km)

summary(m3)
```

Multiple R-squared: 0.3094, Adjusted R-squared: 0.286

```
m4<-lm(KMservice1~decmaking1+Seniorleadershipsupport1+Newlearning1, data = km)

summary(m4)
```

Multiple R-squared: 0.4039, Adjusted R-squared: 0.3731

```
m5<-
```

```
lm(KMservice1~decmaking1+Seniorleadershipsupport1+Newlearning1+Knowledgetransf
er1, data = km)

summary(m5)
```

Multiple R-squared: 0.4263, Adjusted R-squared: 0.386

```
m6<-
lm(KMservice1~decmaking1+Seniorleadershipsupport1+Newlearning1+Knowledgetransfer1+
I mproperselection1, data = km)

summary(m6)
```

Multiple R-squared: 0.4453, Adjusted R-squared: 0.3958

```
m7<-
lm(KMservice1~decmaking1+Seniorleadershipsupport1+Newlearning1+Knowledgetransfer1+
I mproperselection1+Technicalproblem1, data = km)

summary(m7)
```

Multiple R-squared: 0.4557, Adjusted R-squared: 0.3963

```
m8<-
lm(KMservice1~decmaking1+Seniorleadershipsupport1+Newlearning1+Knowledgetransfer1+
I mproperselection1+Technicalproblem1+productivity1, data = km)
summary(m8)
```

Multiple R-squared: 0.5679, Adjusted R-squared: 0.5119

```
m9<-
lm(KMservice1~decmaking1+Seniorleadershipsupport1+Newlearning1+Knowledgetransf
er1+I
mproperselection1+Technicalproblem1+productivity1+overallproductivity1+budgetalloca
tion, data = newkm6)

summary(m9)
```

Multiple R-squared: 0.8555, Adjusted R-squared: 0.8284

```
str(km)
```

Output: Model9 is the best

modelStep – 4: Assumption

testing library(car)

Removal of Outliers

```
outlierTest(m9)

newkm=km[-41,]

newkm1=newkm[-18,]
```

```
newkm2=newkm1[-7,]
newkm3=newkm2[-5,]

newkm4=newkm3[-59,]
newkm5=newkm4[-28,]
newkm6=newkm5[-11,]
```

Checking the plot for

outliers

```
plot(m9,4)
```

Data is normal no

residual

```
r1<-
residuals(object =
m9)
shapiro.test(x=r1)
```

p<0.05, p-value = 2.101e-07

Independence of error

```
durbinWatsonTest(m9)
```

p>0.05, p=0.236 hence auto correlation does not exist

Homoscedasticity, ncv()

```
ncvTest(m9)
```

p<0.05, p = 3.6949e-12, hence the assumption of Homoscedasticity is not met

```
sqrt(vif(m9))>2
```

All are false so it is good. Multi Collinearity assumption is met

Step – 5 Partitioning of data

```
library(caret)
set.seed(1000)
partition<-createDataPartition(y=newkm6$KMservice1, p=0.8, list =
FALSE) training<-newkm6[partition,]
test<-newkm6[-partition,]
```

```
m9<-
lm(KMservice1~decmaking1+Seniorleadershipsupport1+Newlearning1+Knowledgetransf
er1+I
mproperselection1+Technicalproblem1+productivity1+overallproductivity1+budgetalloca
tion, data = training)

summary(m9)
```

Multiple R-squared: 0.871, Adjusted R-squared: 0.8379

Step – 6 : Training is done on

modeltraining$pred<-

predict(m9) training$resd<-

residuals(m9)

Step – 7 : Validate

test$pred1=predict(m9,

newdata=test)

test$resd1=test$KMservice1 -

test$pred1View(training)

View(test)

Step-8: Factor

Analysis For

Principal component

library(MASS)

str(newkm6)

View(newkm6)

newkm7=newkm6[,c(-1:-21)]

View(newkm7)

newkm7_pca<-prcomp(newkm7, center = TRUE, scale = FALSE)

Cannot rescale a constant/zero column to unit variance. Hence taken scale = FALSE

summary(newkm7_pca)

5 components are explaining up to 90% of variance

```
plot(newkm7_pca, type = "l")
```

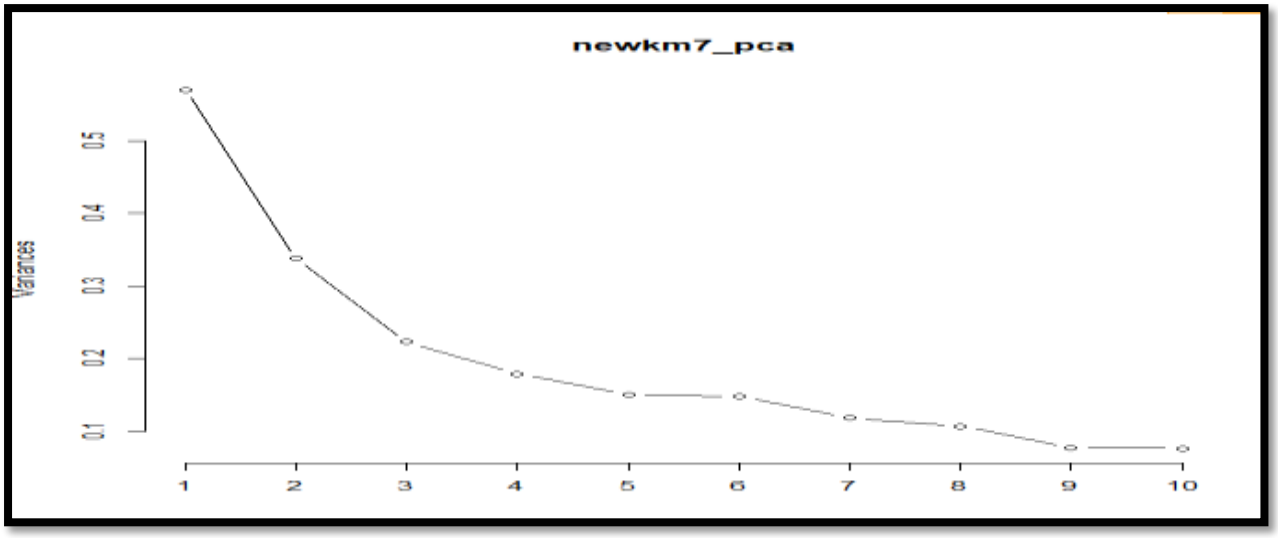


Figure 1: Principle component analysis

5 main factors are clearly identified from

the plotFactor analysis

newkm7.fact<-factanal(newkm7, 4 , rotation = "varimax")

newkm7.fact

Applying cutoff

newkm7.fact<- factanal(newkm7[], 4 , rotation = "varimax", scores = "regression", cutoff=0.3)

newkm7.fact

Diagrammatic Representation of three main factors

Inference:

Through the Principal component analysis we observed three main components which makes the model more efficient in order to enhance the overall productivity of an employee as well as leadsto an organizational effectiveness.

- To attain organizational effectiveness main parameters are decision making, senior leadership support, new learning which leads to enhancement in productivity of an employee which in turn leads to productivity of the organization.
- To ameliorate knowledge management services it is important for an organization to keep a track on liberty to access knowledge material from said department, constructive feedback from both employees as well as trainers, efficiency in providing customer services and self-upskilling.
- Challenges to knowledge management services are improper selection of technology and technical problems. There is need to rebuilt business strategy, making knowledge transfer sessions more efficient and enhancing customer focus by creating visibility.

Managerial Implication and future trends:

IT organizations adopt knowledge management processes in order to enhance employees' as well as customers' satisfaction. It helps in retention of expertise and increasing profits or revenues. (Becerra-Fernandez, 2007) Knowledge management also helps employees to become more flexible and enhances their job satisfaction. It also enhances employee adaptability and they are more likely to accept the change. Employees have a great experience because of their motivation, knowledge acquisition and enhancement of skills. Employees' market value is enhanced in relation to other organizations' employees. Providing tried and tested results or better defined as solutions amplifies effectiveness of employees in performing their jobs. This process helps to keep employees always motivated. (Becerra-Fernandez, 2007) Human capital knowledge is the major organizational capability and is taken as base for all the competitive advantages. IT organizations can acquire and sustain an apt competitive advantage through strategic resource. (Ehsan Zargar1, 2013) Technology has a great impact on knowledge management, motivating and inspiring the development of software platforms to leverage various innovative strategies. The software continues to evolve in response to upcoming demands and challenges. Some of the latest innovations in knowledge management segment are as follow: First Social media is one of the biggest outlook for knowledge management. Advanced search indexing helps to smoothen the process of internal search indexes and makes navigation easy and quicker. Seamless tools of collaboration such as Gantt charts used in IT organizations helps easy scheduling and transparency. Concept of digital workplace is the new phenomenon which is actively used. Image-focused and simple understandable layouts are the new trends. Intranet

software eliminates the problem of log into several applications. Organizing content through specified tags helps in refined content categorization. Digital workplaces and scope of segmented groups helps in improving user friendliness, consistent and immediate notification. Superior senior and customer support. Cloud based software and automated content to provide support to the end user from time to time. (Eisenhauer, 2020)

Conclusion

Companies can gain a strategic competitive advantage through knowledge management as there are lot of technological changes happening in a continuous manner. This paper deals with the importance of knowledge management and how it can lead to a better business productivity. The results of the study conducted confirmed that companies utilize knowledge management when there is a need for a strategic change of the business to gain competitive advantage over its competitors. The study has provided a set of suggestions for the managers. Organizations can motivate their employees to acquire knowledge through online platforms or workshops. Organizations can also provide knowledge through proper training to the employees on a regular basis. This will ultimately lead to improved organizational productivity and performance of the employees will be elevating rapidly. The proper implementation of these knowledge management practices will lead to a better performing organization and a satisfied employee.

Works Cited

Becerra-Fernandez. (2007). Organizational impacts of knowledge management. *2007 Dekai*

Wu, 45.Becerra-Fernandez. (2007). Organizational impacts of knowledge management. *2007*

Dekai Wu, 45.

Ehsan Zargar1, M. R. (2013). The Study of Knowledge management effect on performance rate of employees. *European Online Journal of Natural and Social Sciences 2013*, 1-6.

Eisenhauer, T. (2020). *17 hot knowledge management trends for 2021.* Axero blog.

Inkinen, H. (2016). Review of empirical research on knowledge management practices and firm performance. *Journal of Knowledge Management*, 230 - 257.

Kaminska-Labbé, B. a. (2011). Step-in or step-out: Supporting innovation through communities of practice.

Journal of Business Strategy 32(3):29-36.

Masa'deh. (2017). The Impact of Knowledge Management on Job Performance in Higher Education: The Case of the University of Jordan. *Journal of Enterprise Information Management*.

Mohammed A Abusweilem, S. A. (2019). The impact of knowledge management process and businessintelligence on organizational performance . *Management Science Letters*, 15.

Omotayo, F. O. (2015). Knowledge Management as an important tool in Organizational Management: A Review of Literature. *DigitalCommons@University of Nebraska - Lincoln*.

Rodrigo Valio Dominguez Gonzalez1, M. F. (2014). Knowledge Management: an Analysis From theOrganizational Development. *Journal of Technology Management and Innovation*, 17.

Valamis. (2020). *Knowledge management*. Retrieved from Valamis.com:https://www.valamis.com/hub/knowledge-management

Valamis. (2020). *Knowledge management*. Retrieved from Valamis:https://www.valamis.com/hub/knowledge-management

Yang, J.-t. (2007). The impact of knowledge sharing on organizational learning and effectiveness. *Journalof Knowledge Management*.

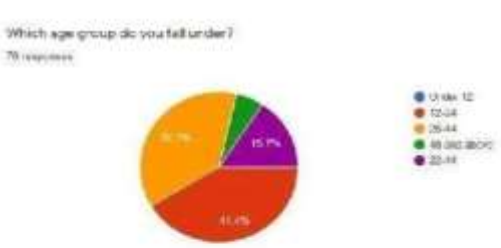
Analysis on Global Pandemic- COVID-19

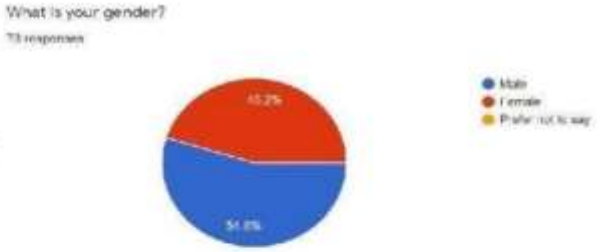
Submitted By-
B. Devi Prasad (PG19035)
Gaurav Maurya (PG19049)
Nikunj Marda (PG19083)

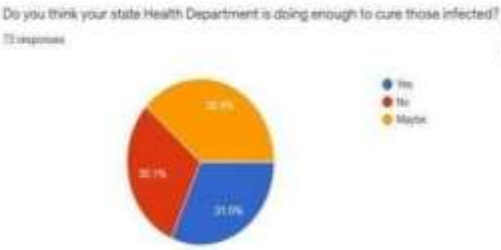
Coronavirus officially called as COVID-19, which was noticed during December 2019 (Wuhan) China, became a major public health problem leading to pandemic affecting worldwide and causing morbidity and mortality, despite various control measures. This Virus is spreading through contact and the symptoms are fever, cough, cold, tiredness, sneezing headache loss of taste, etc and these are visible in people within 10-12 days after they come in contact with the person affected by it.

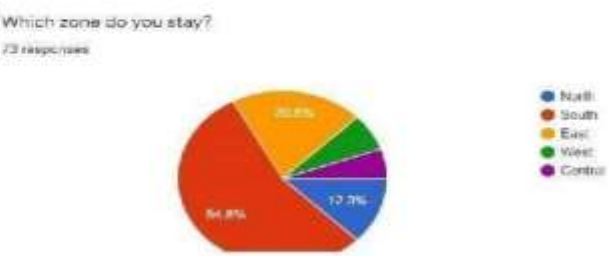
Everyone now is aware of the pandemic and suffering a lot due to it and the sudden lockdown created a disastrous impact on the life of the middle and lower middle-class people who have to migrate thousands of kilometres to their native place from cities as they had no work and could not survive. When the lockdown was there, the cases were in control but once the lockdown was called off, the people started to move out of their homes and the cases increased rapidly. This research was undertaken to assess the level of awareness of coronavirus disease (COVID-19) among locality. We prepared a small set of Questionnaire regarding the awareness of Corona Virus and circulated it to the society consisting of all kinds of people and recorded their views on the same. We got to know the views of people on the basic questions that we asked them.

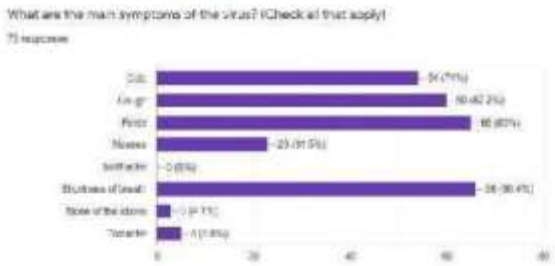
The questions and the results of the reactions are shown below

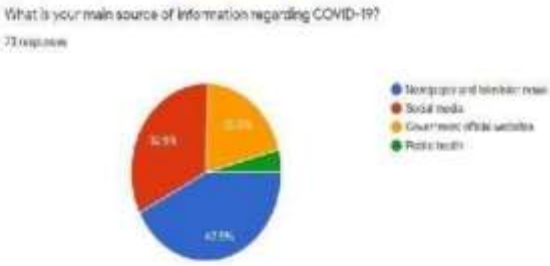

Which age group do you fall under?
73 responses

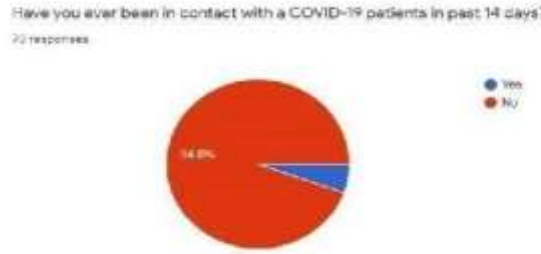

What is your gender?
73 responses
Male
Female
Prefer not to say

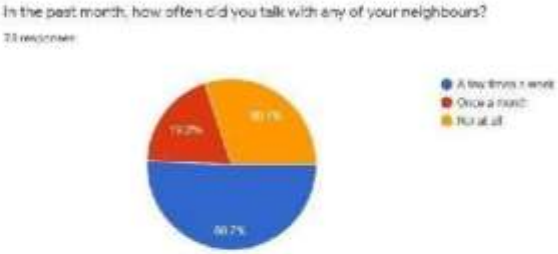

Do you think your state Health Department is doing enough to cure those infected?
73 responses
Yes
No
Maybe

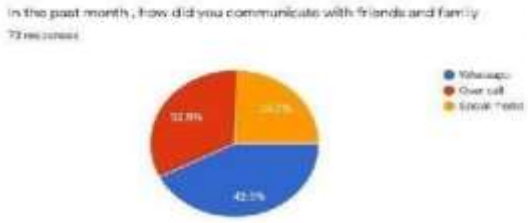

Which zone do you stay?
73 responses
North
South
East
West
Central

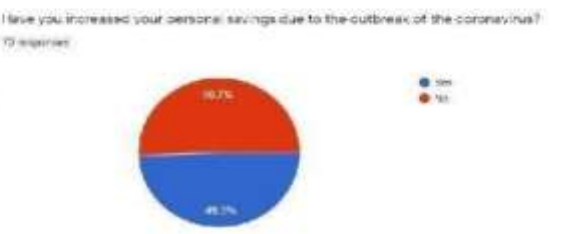

What are the main symptoms of the virus? (Check all that apply)
73 responses

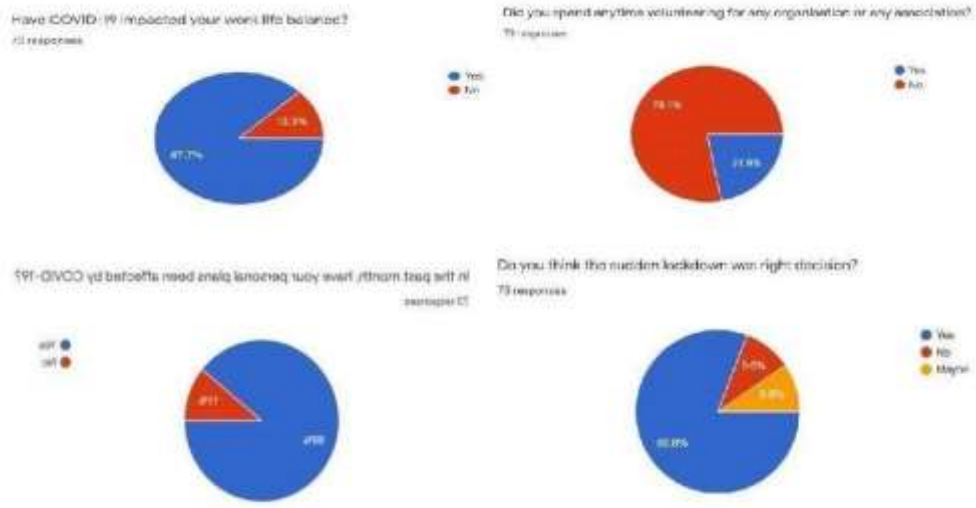

What is your main source of information regarding COVID-19?
73 responses
Social media


Have you ever been in contact with a COVID-19 patients in past 14 days?
72 responses
Yes
No


In the past month, how often did you talk with any of your neighbours?
73 responses
Once a week
Not at all


In the past month, how did you communicate with friends and family
73 responses


Have you increased your personal savings due to the outbreak of the coronavirus?
73 responses



Based on the output we have made certain Models using R studio using predictive analysis.

Methodology

We prepared a questionnaire on which we received close to 170 responses. Based on these responses we have used Logistic Regression to find whether Outbreak (Dependent Variable) creates an impact on other variables or not. Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. In Linear Regression these two variables are related through an equation, where exponent (power) of both these variables is 1. We have used Generalizedlinear model (glm) function, so as to arrive on a desired model. For this we have verified this model byusing Sensitivity, Specificity, confusion matrix, ROCR. Since we got good accuracy with this model so we went ahead with the same.

(We have labelled the heading so as to reduce the length of the variable.)

```
res$Which.age.group.do.you.fall.under.= as.factor(res$Which.age.group.do.you.fall.under.)
res$gender=as.factor(res$gender)

res$zone=as.factor(res$zone)

res$Department=as.factor(res$Department)

res$Symptoms=as.factor(res$Symptoms)

res$Symptoms1=as.factor(res$Symptoms1)

res$information=as.factor(res$information)

res$contact=as.factor(res$contact)

res$neighbours=as.factor(res$neighbours)

res$communicate=as.factor(res$communicate)

res$outbreak=as.factor(res$outbreak)

res$lockdown=as.factor(res$lockdown)

res$personal=as.factor(res$personal)

res$work.life.balance=as.factor(res$work.life.balance)
```

```
model1 =glm(outbreak ~ .,data=res,family=binomial())
summary(model1) # AIC : 214.94

library(MASS)
stepAIC(model1)
```

To arrive at best model, we have encountered 10 models shown as below: -

```
Step:  AIC=214.23
outbreak ~ which.age.group.do.you.fall.under. + gender + zone +
    Department + Symptoms + Symptoms1 + information + contact +
    neighbours + communicate. + lockdown + personal + work.life.balance
```

```
Step:  AIC=206.12
outbreak ~ which.age.group.do.you.fall.under. + gender + zone +
    Department + information + contact + neighbours + communicate. +
    lockdown + personal + work.life.balance
```

```
Step:  AIC=199.49
outbreak ~ which.age.group.do.you.fall.under. + gender + Department +
    information + contact + neighbours + communicate. + lockdown +
    personal + work.life.balance
```

```
Step:  AIC=194.57
outbreak ~ gender + Department + information + contact + neighbours +
    communicate. + lockdown + personal + work.life.balance
```

```
Step:  AIC=190.48
outbreak ~ gender + Department + contact + neighbours + communicate. +
    lockdown + personal + work.life.balance
```

```
Step:  AIC=188.25
outbreak ~ Department + contact + neighbours + communicate. +
    lockdown + personal + work.life.balance
```

```
Step:  AIC=186.33
outbreak ~ Department + contact + communicate. + lockdown + personal +
    work.life.balance
```

```
Step:  AIC=184.65
outbreak ~ Department + contact + communicate. + lockdown + personal
```

```
Step:  AIC=183.33
outbreak ~ Department + contact + lockdown + personal
```

```
Step:  AIC=180.99
outbreak ~ Department + contact + personal
```

```
Call:  glm(formula = outbreak ~ Department + contact, family = binomial(),
    data = res)
```

The above model is chosen for further analysis after selecting the lowest AIC.

Here we take the value as 0 and 1 to determine the values as there are many responses and we cannotdetermine all separately.

```
res$pred = predict(model2 , type = "response")
table(res$outbreak)
head(res$pred)
View(res)
```

```
No Yes
51  89
> head(res$pred)
[1] 0.6985579 0.5446694 0.6985579 0.5446694 0.4455509 0.5446694
> View(res)
```

So the values above 0.5 is taken to be as 1 and below 0.5 is taken to be 0. This will help to determinethe predictive analysis.

```
#if predicted value >1 then take it as 1 or else 0
res$pred <- ifelse(res$pred > 0.5, 1,0)
table(res$pred)
```

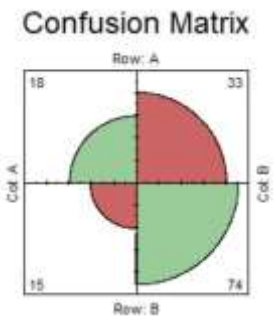
```
 0   1
33 107
```

A confusion matrix is a table that is often used to describe the performance of a classification model(or "classifier") on a set of test data for which the true values are known.

```
table(res$outbreak,res$pred)
str(res)
```

```
     0  1
 No  18 33
 Yes 15 74
>
```

```
#Confusion Matrix
ctable <- as.table(matrix(c(18, 33, 15, 74), nrow = 2, byrow = TRUE))
fourfoldplot(ctable, color = c("#CC6666", "#99CC99"),
             conf.level = 0, margin = 1, main = "Confusion Matrix")
```



The confusion matrix itself is relatively simple to understand, but the related terminology canbe confusing

These are the basic terms used to understand the matrix

- **True positives (TP):** These are cases in which we predicted yes, and they are actuallyyes
- **True negatives (TN):** We predicted no, and they are actually no
- **False positives (FP):** We predicted yes, but they are actually no
- **False negatives (FN):** We predicted no, but they are actually yes.

```
#TP=74   # true positive
#TN=18   # true negative
#FP=33, # false positive
#FN=15    # false negative  yes
```

Sensitivity measures the proportion of positives that are correctly identified (e.g., the percentage of sick people who are correctly identified as having some illness)

```
# finding the sensitivity

#Sensitivity = TP/(TP+FN)
sensi = 74/(74+15)
sensi
#Hence there is 83.15% sensitivity
```

```
> #Sensitivity = TP/(TP+FN)
> sensi = 74/(74+15)
> sensi
[1] 0.8314607
>
```

The result of the Sensitivity analysis is **83.15%** which is highly accurate and almost

Sensitivity generally measures how apt the model is to detecting events in the positive class(yes) so, if the personal saving increased during the outbreak is correctly predicted as yes they are increased then we use sensitivity. so out of **100 %, 83.15%** were increased their saving during outbreak.

Specificity measures the proportion of negatives that are correctly identified (e.g., the percentage ofhealthy people who are correctly identified as not having some illness).

```
# finding the specificity

#Specificity = TN/(TN+FP)
speci = 35/(35+16)
speci
#Hence there is 68.62% specificity
```

```
> #Specificity = TN/(TN+FP)
> speci = 35/(35+16)
> speci
[1] 0.6862745
>
```

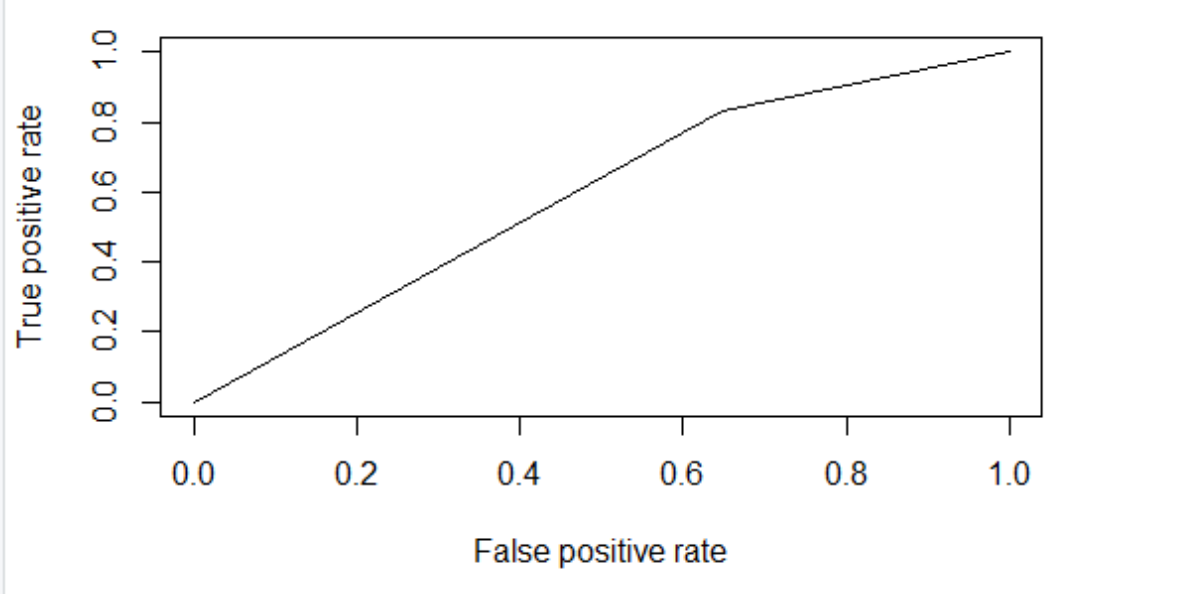
Specificity generally measures how apt the model is to detecting events in the negative class(no) so, if the personal saving is not increased during the outbreak is correctly predicted as no they are not increased then we use specificity. so out of **100 %, 68.62%** were not increased their saving during outbreak

Receiver Operating Characteristic (ROC) curves summarize the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for a predictive model using different probability thresholds.

ROC curves are appropriate when observations are balanced between each class in the dataset.

```
# finding ROC and AUC

library(ROCR)

pred1 <- prediction(res$pred,res$outbreak)
roc.pred1 <- performance(pred1 , measure = 'tpr' , x.measure = 'fpr')
#tpr is true positive rate and fpr is false positive rate

plot(roc.pred1)
#the graph has moved towards the y axis

auc = performance(pred1 , measure ="auc")
auc@y.values[[1]]
#n auc > 0.7 is good (cutoff)
#is the area under the curve
# 60.94%
#its a good model
```

```
> auc@y.values
[[1]]
[1] 0.5922009
```



The curve shows that when the curve is above **0.5** (our threshold point), it means that the model is better. In other words, the Area under the curve is not a good model and more over and above **0.5**, the better is the model.

Conclusion

The coronavirus disease continues to spread across the world following a trajectory that is difficult to predict**.** After we overcome the pandemic, which will surely happen, we must carry out a comprehensive evaluation of the world's ability to maintain stability when faced with similar challenges in the future. We must also craft measures to cope with these challenges together**.** Our study concludes that Outbreak has a great impact of COVID-19 on other variables. More emphasis should be put on updating their knowledge regarding the diagnosis and treatment component of the COVID-19 disease. Interactive educational webinars should be conducted to create awareness. Moreover, the studies can focus upon the psychosocial impact of the disease on individuals.

Customers View on Changing Trends of Mobile Operators

Submitted By-
D Sai Goutham (PG19040)
Sirigina Sushmitha (PG19105)
Setty Rajeswari (PG19167)
Srikar Burra (PG19130)
RVL Soujanya (PG19128)

Abstract

Purpose:

To understand how much the customers are satisfied and are willing tocontinue with the network they are currently using by various factors. To analyze and build a model to understand what improvements can bemade.

Methodology:

Quantitative analysis has been done by making use of Predictive Analytics. Data was collected through a questionnaire about various factors required for the study like qualification, occupation, no of sims, minimum balance enquiry, kind of mobile, minimum balance, satisfaction levels and continuing with network. Related research papers and literature review has been referred to understand the requirements of the study.

Findings:

The 1st linear regression model findings includes the satisfaction levels with other effecting factors.

The 2nd linear regression model findings includes the continuity with current network with

other effecting factors.

We have taken satisfaction levels and continuing with network as dependent variables on independent variables - Kind of mobile, no ofsims and so on.

The presence of other variables which leads to their job satisfaction. The linear regression also finds out the best model by adjusted R squareand the tests - outlier and shapiro are used to check the normality and oytliers of the models. Independence or errors and multicollinearity are checked for the models.

Introduction:

Mobile Phone Operators:

A mobile phone operator, wireless provider, mobile telecommunications company that provides wireless Internet GS or carrier is services for mobile device users. The operator gives a SIM card to the customer who inserts it into the mobile device to gain access to the service. There are two types of mobile operators:

- A mobile network operator (MNO) which owns the underlying network and spectrum assets required to run the services.

- A mobile virtual network operator (MVNO) which buys wholesale service from an MNO and sells on to its own customers.

The Role of Mobile Operators:

Mobile operators have the capabilities, the experience and the track record to provide fast and secure authentication. For more than two decades, mobile operators have been authenticating consumers'devices on their networks, securely providing voice calls, messaging, Internet access and other services, while safeguarding consumers' privacy and personal data.

Telecom Companies in India:

- **BSNL:** The Bharat Sanchar Nigam Limited, country's largest cellular service operator was set up in the year 2000. It is a state- owned telecom company with its headquarters located in New Delhi. BSNL is also the largest land line telephone establishment in India.

- **AIRTEL**: Also known as Bharti Airtel Limited was started in July 1995, with its head office based in New Delhi. Airtel runs its operations in as many as 19 countries across the world and is also ranked fifth as telecom service provider globally. As of April 2011, figures show that Airtel has over 164.61 million users which make it the biggest mobile service operator in India. Its service includes both 2G and 3G facilities.

- **RELIANCE JIO**: Also known as Jio was set up in 2016, with its head office in Navi Mumbai. Reliance Communications as of now has more than 128 million users all across the world.

- **VODAFONE:** Vodafone was founded in 1994 with its head office at

Mumbai. Vodafone provides services to 23 telecom circles across India. Idea was started in 1995, with its head office in Mumbai. It also provides 3G services to its subscribers. And later they Merged.

- **TELENOR**: This Company is a joint venture between Telenor Group and Unitech Group and was started in 2009.

Dual Sims:

Dual SIM refers to mobile phones that support use of multiple SIM cards. Dual SIM phones are mainstream in many countries where phones are normally sold unlocked. Dual SIMs are popular for separating personal and business calls in locations where lower prices apply to calls between clients of the same provider, where a single network may lack comprehensive coverage, and for travel across national and regional borders.

Minimum Balance:

The telecom companies, backed by TRAI, have made it mandatory to keep a minimum balance of at least Rs 35 to prevent SIM deactivation. To put it simply, you will not only be required to maintain a minimum balance in your bank account but also in your phone's main balance. Of course, the minimum balance recharge will have to be maintained only by prepaid users.

Problem of the Study:

The mobile operators have pulled up warning subscribers of certain plans that their SIM cards would be deactivated if they do not recharge their pre-paid accounts with a fixed minimum balance. So, this makes the users to recharge with a larger amount than usual which is creating problem to many users. Mobile Operators like Airtel, Vodafone Idea have introduced this minimum balance which is indirectly making the users to switch to other networks.

Scope of the Study:

The main aim of the study is to establish a platform to examine the customer preferences for the selected mobile networks.

Code Analysis :

```
getwd()
setwd("E:/R studio")
rep=read.csv("E:/R studio/report.csv")
str(rep)
View(rep)
summary(rep)
table(is.na(rep))
#to find if there are any missing values
#there are no missing values so false
list(rep)

rep$SEX=NULL
rep$AGE=NULL
rep$Qualification=NULL
rep$Occupation=NULL
names(rep)
View(rep)

#Models
#Model1
model1=lm(satisfaction~Kind.of.mobile+Min..Balance+Min..Balance.Enquiry+Method.of.recharg
e, data = rep)
summary(model1)

rept1=data.frame(rep)
str(rept1)
head(rept1)

rept1=lm(model1)
library(car)
library(carData)
#H0: There are no outliers in the data
#H1: There are outliers in the data
outlierTest(rept1)
#p value< 0.05 so, There are no outliers in the data


#Shapiro Wilk Test
#H0: The data is normally distributed
#H1: The data is not normally distributed
shapiro.test(residuals(object = rept1))
#Hence the data is normally distributed

#Independence of Errors
#H0: Autocorrelation doesnot exist
#H1: Autocorrelation exist
durbinWatsonTest(rept1)
#0.02<0.05 so autocorrelation exists
```

```
#Multicollinearity
#Variance inflation factor should be less than 10
#squareroot of Variable inflation factor should be less than 2
vif(rept1)
sqrt(vif(rept1))>2
#VIF is less than 10 for independent variables

#Model2
model2=lm(Cont..with.network ~
Kind.of.mobile+Min..Balance+Min..Balance.Enquiry+Method.of.recharge, data = rep)
summary(model2)

rept2=data.frame(rep)
str(rept2)
head(rept2)

rept2=lm(model2)
#H0: There are no outliers in the data
#H1: There are outliers in the data
outlierTest(rept2)
#There are no outliers in data.

#Shapiro Wilk Test
#H0: The data is normally distributed
#H1: The data is not normally distributed
shapiro.test(residuals(object = rept2))
#this model is normally distributed

#Independence of Errors
#H0: Autocorrelation doesnot exist
#H1: Autocorrelation exist
durbinWatsonTest(rept2)
#-0.09177<0.05 so autocorrelation exists


#Multicollinearity
#Variance inflation factor should be less than 10
#squareroot of Variable inflation factor should be less than 2
vif(rept2)
sqrt(vif(rept2))>2
#VIF is less than 10 for independent variables
```

Output And Conclusions:

```
setwd("C:/Users/HP/OneDrive/Desktop")
> rep=read.csv("C:/Users/HP/OneDrive/Desktop")
rep=read.csv("C:/Users/HP/OneDrive/Desktop/rep.csv")
> str(rep)
'data.frame':  233 obs. of  12 variables:
 $ Sno.                 : int  1 2 3 4 5 6 7 8 9 10 ...
 $ SEX                  : int  1 2 2 1 2 2 2 1 2 2 ...
 $ AGE                  : int  60 34 50 57 44 45 18 22 30 23 ...
 $ Qualification        : int  5 5 5 5 5 5 2 3 4 4 ...
 $ Kind.of.mobile       : int  2 2 2 2 1 2 1 1 2 1 ...
 $ No..of.sims          : int  2 1 1 2 1 2 2 2 1 1 ...
 $ Min..Balance         : int  1 2 2 1 1 2 1 2 2 1 ...
 $ Occupation           : int  2 1 1 1 1 1 2 2 1 2 ...
 $ Method.of.recharge   : int  1 2 1 1 2 2 1 2 1 2 ...
 $ Min..Balance.Enquiry: int  4 4 5 2 0 4 5 0 5 1 ...
 $ satisfaction         : int  3 2 3 1 3 3 1 3 2 1 ...
 $ Cont..with.network   : int  1 1 2 2 2 1 2 1 1 2 ...
> View(rep)
> summary(rep)
      Sno.          SEX             AGE        Qualification   Kind.of.mobile
No..of.sims     Min..Balance
 Min.   :  1   Min.   :1.000   Min.   :15.00   Min.   :1.000   Min.   :1.000
Min.   :1.000   Min.   :1.000
 1st Qu.: 59   1st Qu.:1.000   1st Qu.:26.00   1st Qu.:4.000   1st Qu.:1.000
1st Qu.:1.000   1st Qu.:1.000
 Median :117   Median :1.000   Median :36.00   Median :5.000   Median :2.000
Median :1.000   Median :2.000
 Mean   :117   Mean   :1.489   Mean   :36.85   Mean   :4.283   Mean   :1.528
Mean   :1.476   Mean   :1.545
 3rd Qu.:175   3rd Qu.:2.000   3rd Qu.:49.00   3rd Qu.:5.000   3rd Qu.:2.000
3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :233   Max.   :2.000   Max.   :60.00   Max.   :5.000   Max.   :2.000
Max.   :2.000   Max.   :2.000
   Occupation    Method.of.recharge Min..Balance.Enquiry satisfaction
Cont..with.network
 Min.   :1.000   Min.   :1.000      Min.   :0.000        Min.   :1.00   Min.
:1.000
 1st Qu.:1.000   1st Qu.:1.000      1st Qu.:1.000        1st Qu.:1.00   1st
Qu.:1.000
 Median :1.000   Median :2.000      Median :3.000        Median :2.00   Median
:2.000
 Mean   :1.352   Mean   :1.567      Mean   :2.652        Mean   :2.03   Mean
:1.528
 3rd Qu.:2.000   3rd Qu.:2.000      3rd Qu.:4.000        3rd Qu.:3.00   3rd
Qu.:2.000
 Max.   :2.000   Max.   :2.000      Max.   :5.000        Max.   :3.00   Max.
:2.000
> table(is.na(rep))

FALSE
 2796
> #to find if there are any missing values
> #there are no missing values so false
> list(rep)
[[1]]
   Sno. SEX AGE Qualification Kind.of.mobile No..of.sims Min..Balance
Occupation Method.of.recharge
1     1   1  60             5              2           2             1
2                  1
2     2   2  34             5              2           1             2
1                  2
3     3   2  50             5              2           1             2
1                  1
4     4   1  57             5              2           2             1
1                  1
5     5   2  44             5              1           1             1
1                  2
```

6	6	2	45	2	5	2	2	2	1
7	7	2	18	1	2	1	2	1	2
8	8	1	22	2	3	1	2	2	2
9	9	2	30	1	4	2	1	2	1
10	10	2	23	1	4	1	1	1	2
11	11	1	34	2	5	1	1	2	1
12	12	2	32	1	5	2	1	2	1
13	13	2	21	2	4	2	1	1	2
14	14	2	49	1	5	2	2	2	1
15	15	2	56	1	5	1	2	1	1
16	16	1	23	2	4	1	1	2	2
17	17	1	44	2	5	2	2	1	1
18	18	2	42	1	5	1	2	1	1
19	19	1	17	1	1	1	2	1	2
20	20	1	60	1	5	2	2	2	2
21	21	2	57	1	5	2	1	1	1
22	22	2	15	2	1	1	2	1	2
23	23	2	57	2	5	2	1	2	1
24	24	1	20	2	3	2	2	1	2
25	25	1	23	1	4	1	2	2	2
26	26	1	46	1	5	2	2	2	1
27	27	1	29	1	4	1	1	1	1
28	28	2	31	2	5	2	1	1	1
29	29	2	16	2	1	2	1	1	2
30	30	1	57	1	5	1	1	2	1
31	31	1	59	2	5	1	1	2	1
32	32	1	31	1	4	1	1	2	1
33	33	1	24	1	4	1	1	2	2
34	34	2	28	2	4	2	1	2	1
35	35	2	50	1	5	2	2	1	1
36	36	1	28	2	4	1	2	2	1
37	37	1	38	2	5	2	2	1	1
38	38	2	36	1	5	2	1	2	1
39	39	1	21	2	3	1	1	2	2
40	40	2	28	1	4	1	1	1	1

41	41	2	35		5	1	1	1	1
2									
42	42	1	52		5	1	1	2	1
1									
43	43	1	47		5	1	2	2	1
1									
44	44	1	46		5	1	2	1	1
2									
45	45	1	49		5	1	2	2	1
1									
46	46	1	32		5	2	2	1	1
2									
47	47	2	26		4	2	1	2	1
2									
48	48	1	18		2	1	1	1	2
2									
49	49	2	15		1	1	1	1	2
2									
50	50	1	25		4	1	1	2	1
2									
51	51	2	53		5	2	1	2	1
1									
52	52	2	39		5	2	2	2	1
2									
53	53	2	47		5	2	2	1	1
2									
54	54	1	30		4	2	1	2	1
1									
55	55	2	50		5	1	2	1	1
2									
56	56	2	27		4	2	2	2	1
1									
57	57	1	29		4	1	1	2	1
2									
58	58	1	57		5	1	1	1	1
2									
59	59	2	56		5	1	1	2	1
1									
60	60	1	57		5	2	1	1	1
1									
61	61	2	56		5	2	1	2	1
2									
62	62	1	31		5	1	1	2	1
2									
63	63	1	50		5	2	1	2	1
2									
64	64	2	48		5	1	2	2	1
2									
65	65	2	27		4	1	1	1	1
1									
66	66	2	53		5	1	2	2	1
2									
67	67	1	40		5	2	2	2	1
1									
68	68	2	28		4	1	1	2	1
2									
69	69	1	21		3	1	1	1	2
2									
70	70	1	29		4	2	2	1	1
2									
71	71	2	47		5	1	1	1	1
2									
72	72	2	36		5	2	2	2	1
1									
73	73	1	36		5	1	1	2	1
2									
74	74	2	31		5	1	2	1	1
2									
75	75	2	16		1	1	1	1	2
2									

```
76   76   1  54              5              1              1              1
1                  1
77   77   2  21              3              2              2              1
2                  2
78   78   2  25              4              2              2              2
1                  1
79   79   2  44              5              1              1              1
1                  2
80   80   1  31              5              2              2              1
1                  1
81   81   2  48              5              2              2              2
1                  1
82   82   1  20              3              1              1              2
2                  2
83   83   1  58              5              2              1              2
1                  1
   Min..Balance.Enquiry satisfaction Cont..with.network
1                     4            3                  1
2                     4            2                  1
3                     5            3                  2
4                     2            1                  2
5                     0            3                  2
6                     4            3                  1
7                     5            1                  2
8                     0            3                  1
9                     5            2                  1
10                    1            1                  2
11                    1            2                  1
12                    2            3                  2
13                    4            3                  2
14                    5            2                  2
15                    2            1                  2
16                    4            3                  1
17                    2            2                  1
18                    3            1                  1
19                    1            1                  2
20                    3            1                  2
21                    5            2                  2
22                    0            1                  2
23                    0            2                  1
24                    1            2                  1
25                    5            1                  2
26                    3            2                  2
27                    0            1                  2
28                    1            1                  1
29                    2            3                  2
30                    5            3                  2
31                    4            2                  2
32                    5            1                  1
33                    3            2                  2
34                    3            2                  2
35                    5            1                  1
36                    2            2                  1
37                    5            1                  1
38                    2            2                  2
39                    5            1                  2
40                    4            1                  1
41                    1            1                  2
42                    3            2                  1
43                    0            2                  2
44                    3            2                  2
45                    4            1                  1
46                    3            2                  1
47                    1            2                  1
48                    5            3                  2
49                    1            1                  1
50                    4            3                  2
51                    1            2                  2
52                    0            1                  1
53                    0            1                  2
54                    3            1                  1
```

```
55                      1               2                   2
56                      1               3                   2
57                      1               2                   2
58                      2               3                   1
59                      3               3                   1
60                      0               1                   2
61                      3               3                   1
62                      5               1                   1
63                      0               3                   1
64                      1               3                   2
65                      0               1                   1
66                      2               1                   2
67                      1               2                   2
68                      5               2                   1
69                      3               3                   2
70                      5               1                   1
71                      2               3                   2
72                      1               1                   2
73                      0               3                   1
74                      0               1                   1
75                      1               3                   1
76                      1               2                   1
77                      5               3                   2
78                      4               2                   2
79                      1               1                   1
80                      0               2                   2
81                      4               2                   2
82                      1               3                   2
83                      1               3                   1
 [ reached 'max' / getOption("max.print") -- omitted 150 rows ]

> rep$SEX=NULL
> rep$AGE=NULL
> rep$Qualification=NULL
> rep$Occupation=NULL
> names(rep)
[1] "Sno."                "Kind.of.mobile"      "No..of.sims"
"Min..Balance"
[5] "Method.of.recharge"  "Min..Balance.Enquiry" "satisfaction"
"Cont..with.network"
> View(rep)
> #Models
> #Model1
>
model1=lm(satisfaction~Kind.of.mobile+Min..Balance+Min..Balance.Enquiry+Method.
of.recharge, data = rep)
> summary(model1)

Call:
lm(formula = satisfaction ~ Kind.of.mobile + Min..Balance +
Min..Balance.Enquiry +
    Method.of.recharge, data = rep)

Residuals:
     Min       1Q   Median       3Q      Max
-1.36852 -0.85079  0.00646  0.83832  1.36335

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.26167    0.31616   3.991 8.89e-05 ***
Kind.of.mobile        0.18091    0.10932   1.655   0.0993 .
Min..Balance          0.06408    0.11006   0.582   0.5610
Min..Balance.Enquiry  0.07138    0.03094   2.307   0.0219 *
Method.of.recharge    0.12998    0.11058   1.175   0.2410
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.832 on 228 degrees of freedom
Multiple R-squared:  0.04234,  Adjusted R-squared:  0.02554
F-statistic:  2.52 on 4 and 228 DF,  p-value: 0.04202
```

```
> rept1=data.frame(rep)
> str(rept1)
'data.frame':  233 obs. of  8 variables:
 $ Sno.                 : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Kind.of.mobile       : int  2 2 2 2 1 2 1 1 2 1 ...
 $ No..of.sims          : int  2 1 1 2 1 2 2 2 1 1 ...
 $ Min..Balance         : int  1 2 2 1 1 2 1 2 2 1 ...
 $ Method.of.recharge   : int  1 2 1 1 2 2 1 2 1 2 ...
 $ Min..Balance.Enquiry: int  4 4 5 2 0 4 5 0 5 1 ...
 $ satisfaction         : int  3 2 3 1 3 3 1 3 2 1 ...
 $ Cont..with.network   : int  1 1 2 2 2 1 2 1 1 2 ...
> head(rept1)
  Sno. Kind.of.mobile No..of.sims Min..Balance Method.of.recharge
Min..Balance.Enquiry satisfaction
1    1               2           2            1                  1
4             3
2    2               2           1            2                  2
4             2
3    3               2           1            2                  1
5             3
4    4               2           2            1                  1
2             1
5    5               1           1            1                  2
0             3
6    6               2           2            2                  2
4             3
  Cont..with.network
1                  1
2                  1
3                  2
4                  2
5                  2
6                  1
> rept1=lm(model1)
> library(car)
Loading required package: carData
> library(carData)
> library(car)
> library(carData)
> #H0: There are no outliers in the data
> #H1: There are outliers in the data
> outlierTest(rept1)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
     rstudent unadjusted p-value Bonferroni p
196 -1.670482           0.096202           NA
> #Shapiro Wilk Test
> #H0: The data is normally distributed
> #H1: The data is not normally distributed
> shapiro.test(residuals(object = rept1))

        Shapiro-Wilk normality test

data:  residuals(object = rept1)
W = 0.90852, p-value = 9.51e-11

> #Independence of Errors
> #H0: Autocorrelation doesnot exist
> #H1: Autocorrelation exist
> durbinWatsonTest(rept1)
 lag Autocorrelation D-W Statistic p-value
   1      0.02435859        1.9397   0.634
 Alternative hypothesis: rho != 0
> #Multicollinearity
> #Variance inflation factor should be less than 10
> #squareroot of Variable inflation factor should be less than 2
> vif(rept1)
      Kind.of.mobile         Min..Balance Min..Balance.Enquiry
Method.of.recharge
            1.002590             1.011136             1.002319
1.010900
```

```
> sqrt(vif(rept1))>2
      Kind.of.mobile         Min..Balance Min..Balance.Enquiry
Method.of.recharge
               FALSE                FALSE                FALSE
FALSE
> #Model2
> model2=lm(Cont..with.network ~
Kind.of.mobile+Min..Balance+Min..Balance.Enquiry+Method.of.recharge, data =
rep)
> summary(model2)

Call:
lm(formula = Cont..with.network ~ Kind.of.mobile + Min..Balance +
    Min..Balance.Enquiry + Method.of.recharge, data = rep)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6223 -0.5123  0.3781  0.4570  0.5657

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.4699248  0.1904521   7.718 3.68e-13 ***
Kind.of.mobile       0.0346921  0.0658522   0.527    0.599
Min..Balance         0.0776017  0.0662991   1.170    0.243
Min..Balance.Enquiry 0.0003444  0.0186364   0.018    0.985
Method.of.recharge  -0.0739515  0.0666138  -1.110    0.268
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5012 on 228 degrees of freedom
Multiple R-squared:  0.01383,  Adjusted R-squared:  -0.003473
F-statistic: 0.7993 on 4 and 228 DF,  p-value: 0.5268

> rept2=data.frame(rep)
> str(rept2)
'data.frame':  233 obs. of  8 variables:
 $ Sno.                : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Kind.of.mobile      : int  2 2 2 2 1 2 1 1 2 1 ...
 $ No..of.sims         : int  2 1 1 2 1 2 2 2 1 1 ...
 $ Min..Balance        : int  1 2 2 1 1 2 1 2 2 1 ...
 $ Method.of.recharge  : int  1 2 1 1 2 2 1 2 1 2 ...
 $ Min..Balance.Enquiry: int  4 4 5 2 0 4 5 0 5 1 ...
 $ satisfaction        : int  3 2 3 1 3 3 1 3 2 1 ...
 $ Cont..with.network  : int  1 1 2 2 2 1 2 1 1 2 ...
> head(rept2)
  Sno. Kind.of.mobile No..of.sims Min..Balance Method.of.recharge
Min..Balance.Enquiry satisfaction
1    1               2           2            1                    1
4             3
2    2               2           1            2                    2
4             2
3    3               2           1            2                    1
5             3
4    4               2           2            1                    1
2             1
5    5               1           1            1                    2
0             3
6    6               2           2            2                    2
4             3
  Cont..with.network
1                  1
2                  1
3                  2
4                  2
5                  2
6                  1
> rept2=lm(model2)
> #H0: There are no outliers in the data
> #H1: There are outliers in the data
> outlierTest(rept2)
No Studentized residuals with Bonferroni p < 0.05
```

```
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
9 -1.258087            0.20965           NA
> #Shapiro Wilk Test
> #H0: The data is normally distributed
> #H1: The data is not normally distributed
> shapiro.test(residuals(object = rept2))

        Shapiro-Wilk normality test

data:  residuals(object = rept2)
W = 0.73724, p-value < 2.2e-16

> #Independence of Errors
> #H0: Autocorrelation doesnot exist
> #H1: Autocorrelation exist
> durbinWatsonTest(rept2)
 lag Autocorrelation D-W Statistic p-value
   1     -0.09177298      2.174782    0.16
 Alternative hypothesis: rho != 0
> #Multicollinearity
> #Variance inflation factor should be less than 10
> #squareroot of Variable inflation factor should be less than 2
> vif(rept2)
      Kind.of.mobile         Min..Balance Min..Balance.Enquiry
Method.of.recharge
            1.002590             1.011136             1.002319
1.010900
> sqrt(vif(rept2))>2
      Kind.of.mobile         Min..Balance Min..Balance.Enquiry
Method.of.recharge
               FALSE                FALSE                FALSE
FALSE
>
> #VIF is less than 10 for independent variables
```

WFH Analysis- A Short Research on the Work from Home Culture Adopted During the Pandemic and its Various Results

Submitted By -
Harshit Maheshwari
Ritika Goyal
Annu Shukla

Introduction:

In the current situation, there have been many changes in the work and study culture. From going to the office, school, college daily, we have adapted to working from our homes.Be it a student, a teacher, an engineer, or any other profession we can think of, the base location of the work has changed. No one thought we would be having exams from home via online platforms, nor did we think that we would be having meetings, internships, jobs from home in different platforms like Google Meet, Cisco Webex meetings, and so on.

Before the COVID-19 situation came, most of us were not even aware of these platforms, many did not even exist. But now, the entire circumstances have changed.

We have conducted a short study on the effects of this shift from office/school/college to home on productivity, time management, work life balance, and several other aspects to find out what opinion people of different age groups, gender, occupation, etc. have aboutthis shift.

Abstract:

This paper focuses on the effect of various things like demographics including gender, age, experience, problems faced in online connection, productivity and work life balance, and some more independent variables on the choice of people in selecting work from home or a

hybrid setup (combination of both work from home and in-office) as their medium of workor study. We have done a predictive analysis to find out how much these factors are correlated with the dependent variable, that is preference of people, and predict the preference of people if we have a certain set of independent known variable. Using the logistic regression model, confusion matrix, ROCR and AUC curves in R programming, we've arrived at the conclusion that while some of these factors have a great effect on the dependent variable, some have insignificant influence on choice of people.

Analysis:

The following code in R represents our analysis of the data that we collected from various students and professions in the form of a questionnaire, in which we got 147 responses. The gist of this analysis is that, firstly we've factorized our data, that is converting the characters into factors, as for building models it is a necessity. Then we applied various combinations of independent variables with our dependent variable to build models using the logistic regression model, the reason for using the GLM model is that our dependent variable is categorical in nature. After the model building, we check thew accuracy of each model by applying various checks, namely, confusion matrix, ROCR curve, and AUC. Two of our models had similar accuracy so we have demonstrated the both here.

```
wfh= read.csv("WFH.csv")
View(wfh)

#converting characters into factors
wfh$Gender = as.factor(wfh$Gender)
wfh$Residency = as.factor(wfh$Residency)
wfh$Profession = as.factor(wfh$Profession)
wfh$Preference = as.factor(wfh$Preference)
wfh$Time_Management= as.factor(wfh$Time_Management)
wfh$Productivity = as.factor(wfh$Productivity)
wfh$WLB = as.factor(wfh$WLB)
wfh$Family_Time = as.factor(wfh$Family_Time)
wfh$New_Start = as.factor(wfh$New_Start)
wfh$Engagement = as.factor(wfh$Engagement)
wfh$Night_Shifts = as.factor(wfh$Night_Shifts)
wfh$Connectivity_Issues = as.factor(wfh$Connectivity_Issues)
wfh$Interruptions = as.factor(wfh$Interruptions)
wfh$Involvement = as.factor(wfh$Involvement)
wfh$Unavailability_PC = as.factor(wfh$Unavailability_PC)

str(wfh)

## 'data.frame':    142 obs. of  16 variables:
##  $ Age                : int  22 26 20 20 21 42 19 20 20 23 ...
  ##  $ Gender             : Factor w/ 3 levels "Female","Male",..: 3 1 2 2 1 1
  1 2 2 1 ...
  ##  $ Residency          : Factor w/ 3 levels "Rural area","Semi-Urban",..: 3
  3 3 3 3 3 2 3 2 3 ...
  ##  $ Profession         : Factor w/ 5 levels "Business","Employed",..: 4 4 4
  4 4 4 4 4 4 4 ...
  ##  $ Preference         : Factor w/ 2 levels "Hybrid (combination of
  both)",..: 1 1 1 1 1 1 1 1 1 1 ...
  ##  $ Time_Management    : Factor w/ 5 levels "Equally efficient",..: 3 3 5 4
  5 1 4 1 4 1 ...
  ##  $ Productivity       : Factor w/ 5 levels "Equally productive",..: 1 2 3
  1 1 1 2 5 5 1 ...
  ##  $ WLB                : Factor w/ 5 levels "Completely agree",..: 1 5 1 4
  1 1 4 4 4 4 ...
  ##  $ Family_Time        : Factor w/ 5 levels "Negatively affects",..: 3 4 5
  2 2 2 5 4 4 4 ...
  ##  $ New_Start          : Factor w/ 5 levels "Completely agree",..: 2 3 1 4
  4 2 3 4 1 2 ...
  ##  $ Engagement         : Factor w/ 5 levels "Completely disengaged",..: 3 3
  2 5 5 3 5 4 3 4 ...
  ##  $ Night_Shifts       : Factor w/ 5 levels "Completely agree",..: 4 1 1 4
  1 1 3 4 1 1 ...
  ##  $ Connectivity_Issues: Factor w/ 5 levels "1","2","3","4",..: 3 5 4 3 5 2
  3 4 5 5 ...
  ##  $ Interruptions      : Factor w/ 5 levels "1","2","3","4",..: 3 5 4 3 5 2
  3 4 4 5 ...
```

```
##  $ Involvement        : Factor w/ 5 levels "1","2","3","4",..: 3 5 5 2 4 3
2 4 3 5 ...
##  $ Unavailability_PC  : Factor w/ 5 levels "1","2","3","4",..: 1 3 1 1 3 1
1 1 1 4 ...
```

```
library(MASS)
#finding best models
model1 <- glm(Preference ~ ., data= wfh, family = binomial())
summary(model1)
```

```
##
## Call:
## glm(formula = Preference ~ ., family = binomial(), data = wfh)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8817  -0.5189  -0.1877   0.5289   3.0207
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -0.23630    5.52867  -0.043  0.96591
## Age                                 0.03799    0.11309   0.336  0.73689
## GenderMale                          1.79678    0.80763   2.225  0.02610 *
## GenderPrefer not to say           -17.17161 2399.54625  -0.007  0.99429
## ResidencySemi-Urban                 0.35305    1.52458   0.232  0.81687
## ResidencyUrban                     -0.19456    1.24955  -0.156  0.87627
## ProfessionEmployed                  3.89468    1.77089   2.199  0.02786 *
## ProfessionInterior                 22.70907 1696.73875   0.013  0.98932
## ProfessionStudent                   2.46282    1.70481   1.445  0.14856
## ProfessionUnemployed                3.54678    2.12971   1.665  0.09584 .
## Time_ManagementInefficient          1.97779    3.07325   0.644  0.51987
## Time_ManagementLess efficient       3.23516    1.16465   2.778  0.00547
**
## Time_ManagementSomewhat efficient   1.09654    0.95760   1.145  0.25217
## Time_ManagementVery efficient       3.95308    1.34441   2.940  0.00328
**
## ProductivityLess productive        -2.60654    1.22708  -2.124  0.03365 *
## ProductivityMore productive        -1.23242    1.04388  -1.181  0.23776
## ProductivityNot productive          2.60123    3.21176   0.810  0.41799
## ProductivitySomewhat productive    -1.01390    1.12997  -0.897  0.36957
## WLBcompletely disagree              1.85644    2.17624   0.853  0.39363
## WLBno difference                    1.56504    1.59076   0.984  0.32520
## WLBsomewhat agree                  -0.69601    1.11625  -0.624  0.53294
## WLBsomewhat disagree                0.84259    1.24959   0.674  0.50013
## Family_TimePositively Affects      -0.36664    1.74722  -0.210  0.83379
## Family_TimeSame                    -1.12464    1.93718  -0.581  0.56154
## Family_TimeSome bad effects        -1.22353    1.64858  -0.742  0.45798
## Family_TimeSome good affects        0.21252    1.58484   0.134  0.89332
## New_Startcompletely disagree        0.10669    1.32164   0.081  0.93566
## New_StartNo Difference              2.30418    1.91365   1.204  0.22856
```

```
## New_Startsomewhat agree               1.58187    1.13142   1.398  0.16207
## New_Startsomewhat disagree            1.84820    1.24623   1.483  0.13807
## EngagementFull engagement            -3.58832    3.48893  -1.028  0.30372
## EngagementNeutral                    -1.36703    3.36220  -0.407  0.68431
## EngagementSomewhat disengaged        -2.97204    3.30296  -0.900  0.36822
## EngagementSomewhat engagement        -4.05834    3.49916  -1.160  0.24613
## Night_Shiftscompletely disagree      -5.35243    3.31216  -1.616  0.10610
## Night_ShiftsNo Difference            -2.09725    1.27811  -1.641  0.10082
## Night_Shiftssomewhat agree            1.31364    0.90273   1.455  0.14562
## Night_Shiftssomewhat disagree         0.21667    1.11341   0.195  0.84571
## Connectivity_Issues2                 -1.65853    1.05514  -1.572  0.11598
## Connectivity_Issues3                 -1.37483    1.12596  -1.221  0.22207
## Connectivity_Issues4                 -3.08340    1.30347  -2.366  0.01800 *
## Connectivity_Issues5                 -3.38638    1.29472  -2.616  0.00891
**
## Interruptions2                       -1.91458    1.21908  -1.571  0.11630
## Interruptions3                       -2.03527    1.30921  -1.555  0.12005
## Interruptions4                       -0.45038    1.39181  -0.324  0.74625
## Interruptions5                       -2.34142    1.66951  -1.402  0.16078
## Involvement2                         -0.75397    1.19125  -0.633  0.52679
## Involvement3                         -1.01601    1.35836  -0.748  0.45448
## Involvement4                         -1.30174    1.45761  -0.893  0.37182
## Involvement5                         -3.81764    1.92393  -1.984  0.04722 *
## Unavailability_PC2                    0.79536    0.89162   0.892  0.37237
## Unavailability_PC3                    2.56375    1.29633   1.978  0.04796 *
## Unavailability_PC4                   -1.15382    1.28320  -0.899  0.36856
## Unavailability_PC5                    2.11240    1.26795   1.666  0.09571 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 190.47  on 141  degrees of freedom
## Residual deviance: 108.31  on  88  degrees of freedom
## AIC: 216.31
##
## Number of Fisher Scoring iterations: 15
```

```r
#AIC = 216.31

model2 = glm(Preference~ Gender + Profession + Time_Management, data = wfh,
family = binomial())
summary(model2)
```

```
##
## Call:
## glm(formula = Preference ~ Gender + Profession + Time_Management,
##     family = binomial(), data = wfh)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.7147  -0.9489  -0.5030   1.1997   2.0641
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -2.6732     1.2736  -2.099   0.0358 *
## GenderMale                           0.7381     0.3988   1.851   0.0642 .
## GenderPrefer not to say            -15.7709  2399.5448  -0.007   0.9948
## ProfessionEmployed                   1.5181     1.2611   1.204   0.2287
## ProfessionInterior                  16.8753  1696.7348   0.010   0.9921
## ProfessionStudent                    0.6695     1.2491   0.536   0.5920
## ProfessionUnemployed                 1.9345     1.4457   1.338   0.1809
## Time_ManagementInefficient           2.2943     0.9496   2.416   0.0157 *
## Time_ManagementLess efficient        1.2085     0.5577   2.167   0.0302 *
## Time_ManagementSomewhat efficient    0.7011     0.5360   1.308   0.1908
## Time_ManagementVery efficient        1.6259     0.6431   2.528   0.0115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##   Null deviance: 190.47  on 141  degrees of freedom
## Residual deviance: 164.08  on 131  degrees of freedom
## AIC: 186.08
##
## Number of Fisher Scoring iterations: 15

#AIC = 186.08

model3 = glm(Preference~Gender + Profession + Time_Management +
Productivity,data = wfh, family = binomial())
summary(model3)

##
## Call:
## glm(formula = Preference ~ Gender + Profession + Time_Management +
##     Productivity, family = binomial(), data = wfh)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6441  -0.9388  -0.4714   1.0608   2.1223
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -2.1657     1.2700  -1.705  0.08814 .
## GenderMale                           0.8262     0.4257   1.941  0.05226 .
## GenderPrefer not to say            -16.7539  2399.5448  -0.007  0.99443
## ProfessionEmployed                   1.6201     1.2381   1.309  0.19069
## ProfessionInterior                  17.1341  1696.7348   0.010  0.99194
## ProfessionStudent                    0.7498     1.2231   0.613  0.53983
```

```
## ProfessionUnemployed                  2.2958     1.4422   1.592  0.11141
## Time_ManagementInefficient            2.6028     1.1185   2.327  0.01997 *
## Time_ManagementLess efficient         1.6037     0.6179   2.595  0.00945 **
## Time_ManagementSomewhat efficient     0.6058     0.5635   1.075  0.28228
## Time_ManagementVery efficient         1.5974     0.6908   2.312  0.02076 *
## ProductivityLess productive          -1.5069     0.6260  -2.407  0.01607 *
## ProductivityMore productive          -0.8259     0.5564  -1.484  0.13768
## ProductivityNot productive           -0.4135     1.4575  -0.284  0.77662
## ProductivitySomewhat productive      -0.7251     0.6021  -1.204  0.22849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 190.47  on 141  degrees of freedom
## Residual deviance: 157.25  on 127  degrees of freedom
## AIC: 187.25
##
## Number of Fisher Scoring iterations: 15

#AIC  = 187.25

model4 = glm(Preference ~ Gender + Profession + Time_Management +
Connectivity_Issues, data = wfh, family = binomial())
summary(model4)

##
## Call:
## glm(formula = Preference ~ Gender + Profession + Time_Management +
##     Connectivity_Issues, family = binomial(), data = wfh)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8461  -0.8857  -0.4683   0.9868   2.0039
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -2.4584     1.2931  -1.901  0.05728 .
## GenderMale                           0.7172     0.4166   1.722  0.08513 .
## GenderPrefer not to say            -16.1477  2399.5448  -0.007  0.99463
## ProfessionEmployed                   2.1190     1.2873   1.646  0.09974 .
## ProfessionInterior                  17.6684  1696.7349   0.010  0.99169
## ProfessionStudent                    1.1905     1.2710   0.937  0.34891
## ProfessionUnemployed                 2.2632     1.4420   1.569  0.11653
## Time_ManagementInefficient           2.5741     1.0134   2.540  0.01108 *
## Time_ManagementLess efficient        1.4451     0.5851   2.470  0.01351 *
## Time_ManagementSomewhat efficient    0.7276     0.5525   1.317  0.18787
## Time_ManagementVery efficient        1.7877     0.6800   2.629  0.00856 **
## Connectivity_Issues2                -0.6623     0.6239  -1.062  0.28846
## Connectivity_Issues3                -0.5957     0.6671  -0.893  0.37188
```

```
## Connectivity_Issues4                -1.1489     0.5891  -1.950  0.05114 .
## Connectivity_Issues5                -1.6280     0.6471  -2.516  0.01188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 190.47  on 141  degrees of freedom
## Residual deviance: 156.28  on 127  degrees of freedom
## AIC: 186.28
##
## Number of Fisher Scoring iterations: 15

#AIC = 186.28

model5= glm(Preference~Gender + Time_Management + Connectivity_Issues,data =
wfh, family = binomial())
summary(model5)

##
## Call:
## glm(formula = Preference ~ Gender + Time_Management + Connectivity_Issues,
##     family = binomial(), data = wfh)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6694  -0.9358  -0.5196   1.0620   1.8648
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -1.0946     0.5309  -2.062  0.03924 *
## GenderMale                         0.7576     0.4038   1.876  0.06064 .
## GenderPrefer not to say          -14.2482   882.7436  -0.016  0.98712
## Time_ManagementInefficient         2.5784     0.9891   2.607  0.00914 **
## Time_ManagementLess efficient      1.2275     0.5529   2.220  0.02641 *
## Time_ManagementSomewhat efficient  0.6147     0.5270   1.166  0.24351
## Time_ManagementVery efficient      1.7787     0.6177   2.879  0.00398 **
## Connectivity_Issues2              -0.3336     0.5874  -0.568  0.57012
## Connectivity_Issues3              -0.4509     0.6552  -0.688  0.49135
## Connectivity_Issues4              -0.8529     0.5552  -1.536  0.12446
## Connectivity_Issues5              -1.4412     0.6210  -2.321  0.02030 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 190.47  on 141  degrees of freedom
## Residual deviance: 165.18  on 131  degrees of freedom
## AIC: 187.18
```

```
##
## Number of Fisher Scoring iterations: 13

#AIC = 187.18

model6 = glm(Preference~Gender + Time_Management,data = wfh, family =
binomial())
summary(model6)

##
## Call:
## glm(formula = Preference ~ Gender + Time_Management, family = binomial(),
##     data = wfh)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8069  -0.9110  -0.6007   1.1161   1.8980
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -1.6208     0.4140  -3.915 9.04e-05 ***
## GenderMale                         0.8108     0.3894   2.082  0.03732 *
## GenderPrefer not to say          -13.9011   882.7435  -0.016  0.98744
## Time_ManagementInefficient         2.2249     0.9264   2.402  0.01632 *
## Time_ManagementLess efficient      0.9558     0.5166   1.850  0.06428 .
## Time_ManagementSomewhat efficient  0.5349     0.5098   1.049  0.29401
## Time_ManagementVery efficient      1.5958     0.5908   2.701  0.00691 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 190.47  on 141  degrees of freedom
## Residual deviance: 171.81  on 135  degrees of freedom
## AIC: 185.81
##
## Number of Fisher Scoring iterations: 13

#AIC = 185.81

model7 = glm(Preference~Gender + Time_Management +Connectivity_Issues
+Productivity, data = wfh, family = binomial())
summary(model7)

##
## Call:
## glm(formula = Preference ~ Gender + Time_Management + Connectivity_Issues
+
##     Productivity, family = binomial(), data = wfh)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.7310  -0.8924  -0.5177   1.0303   2.0285
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -0.4561     0.6178  -0.738  0.46032
## GenderMale                           0.7877     0.4240   1.858  0.06322 .
## GenderPrefer not to say            -16.0236  1455.3978  -0.011  0.99122
## Time_ManagementInefficient           2.8774     1.1748   2.449  0.01431 *
## Time_ManagementLess efficient        1.6077     0.6089   2.640  0.00828 **
## Time_ManagementSomewhat efficient    0.6009     0.5518   1.089  0.27622
## Time_ManagementVery efficient        1.8036     0.6753   2.671  0.00756 **
## Connectivity_Issues2                -0.2494     0.6147  -0.406  0.68493
## Connectivity_Issues3                -0.6941     0.6895  -1.007  0.31415
## Connectivity_Issues4                -0.8885     0.5637  -1.576  0.11500
## Connectivity_Issues5                -1.5147     0.6612  -2.291  0.02196 *
## ProductivityLess productive         -1.3646     0.6339  -2.153  0.03135 *
## ProductivityMore productive         -0.8904     0.5821  -1.530  0.12607
## ProductivityNot productive          -0.2204     1.5052  -0.146  0.88358
## ProductivitySomewhat productive     -0.7705     0.5718  -1.348  0.17782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 190.47  on 141  degrees of freedom
## Residual deviance: 159.20  on 127  degrees of freedom
## AIC: 189.2
##
## Number of Fisher Scoring iterations: 14
```

```
#AIC = 189.2

#we check all models for best fit
#model3
#predicting the accuracy of model
wfh$pred1 = predict(model3,type = "response")

table(wfh$Preference)
```

```
##
## Hybrid (combination of both)                    In-Office
##                           86                           56
```

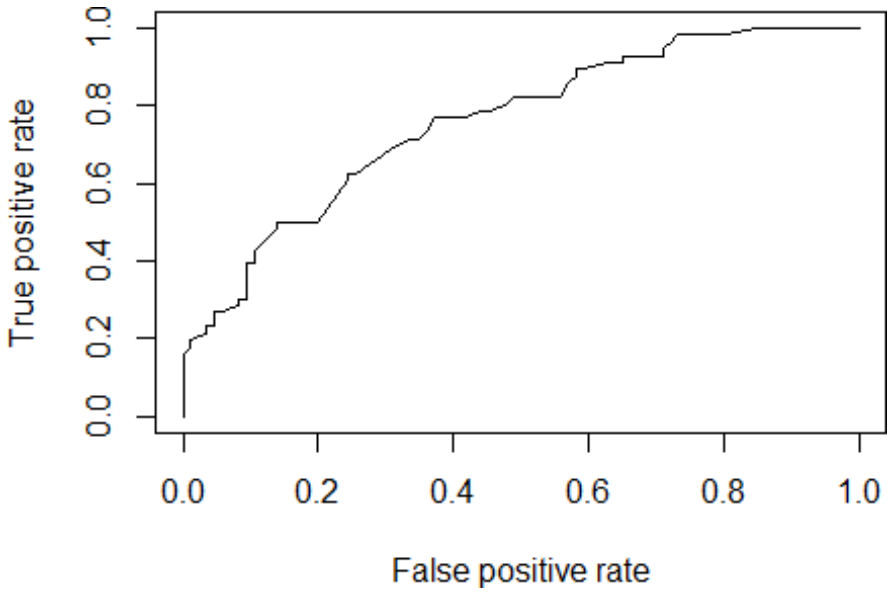
```
#0-86, 1-56
head(wfh$pred1)
```

```
## [1] 6.389220e-08 2.109710e-01 5.453254e-01 5.040470e-01 5.452557e-01
## [6] 1.953069e-01
```

```
View(wfh)
```

```
wfh$prefer1 = ifelse(wfh$pred1 > 0.5,"In-Office","Hybrid (combination of
both)")
table(wfh$prefer1)

##
## Hybrid (combination of both)                    In-Office
##                           87                           55

# 0-87, 1-55

#confusion matrix for model3
newwfh1 = data.frame(predicted=wfh$prefer1,actual=wfh$Preference)
newwfh1$predicted=as.factor(newwfh1$predicted)
str(newwfh1)

## 'data.frame':    142 obs. of  2 variables:
##  $ predicted: Factor w/ 2 levels "Hybrid (combination of both)",..: 1 1 2
2 2 1 1 1 1 1 ...
##  $ actual   : Factor w/ 2 levels "Hybrid (combination of both)",..: 1 1 1
1 1 1 1 1 1 1 ...

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

res1 = confusionMatrix(newwfh1$predicted,newwfh1$actual)
res1

## Confusion Matrix and Statistics
##
##                                Reference
## Prediction                      Hybrid (combination of both) In-Office
##   Hybrid (combination of both)                           65         22
##   In-Office                                              21         34
##
##                Accuracy : 0.6972
##                  95% CI : (0.6145, 0.7714)
##     No Information Rate : 0.6056
##     P-Value [Acc > NIR] : 0.01485
##
##                   Kappa : 0.3641
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.7558
##             Specificity : 0.6071
##          Pos Pred Value : 0.7471
##          Neg Pred Value : 0.6182
##              Prevalence : 0.6056
##          Detection Rate : 0.4577
```

```
##    Detection Prevalence : 0.6127
##       Balanced Accuracy : 0.6815
##
##        'Positive' Class : Hybrid (combination of both)
##

#confusion matrix for model3
#            hybrid   in-office
#hybrid       65         22
#in-office    21         34
#accuracy = 0.6972
#sensitivity = 0.7558
#Specificity = 0.6071

library(ROCR)

## Warning: package 'ROCR' was built under R version 4.0.3

#ROCR curve for model 3
pred1 <- prediction(wfh$pred1 , wfh$Preference)
roc.pred1 <- performance(pred1 , measure = 'tpr' , x.measure = 'fpr')
plot(roc.pred1)
```



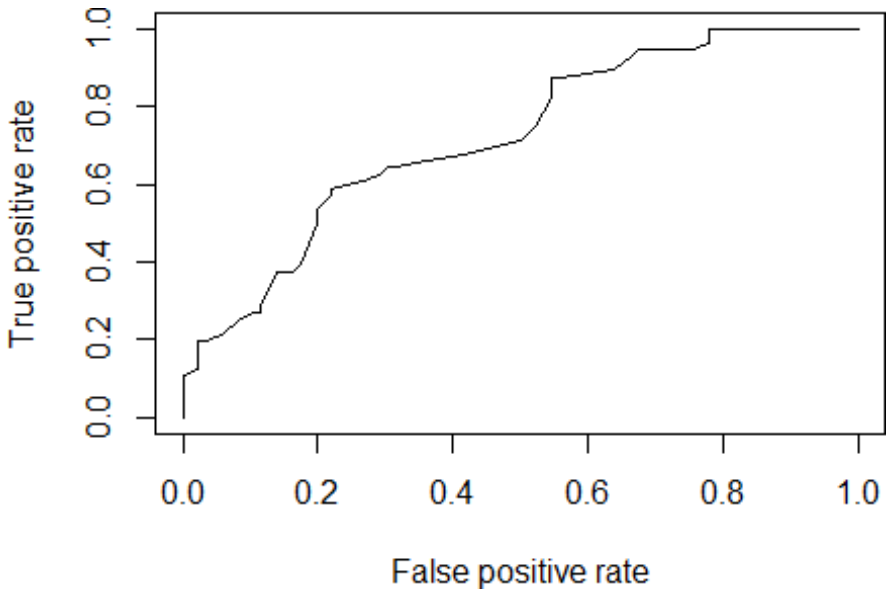

```
#our curve is close to y axis than x axis, therefore this model is good

#AUC for model 3
auc = performance(pred1 , measure ="auc")
auc@y.values[{1}]
```

```
## [[1]]
## [1] 0.7585133

#auc >0.7, hence this model is good (auc = 0.7585)

wfh$pred2 = predict(model5,type = "response")

wfh$prefer2 = ifelse(wfh$pred2 > 0.5,"In-Office","Hybrid
(combination ofboth)")
table(wfh$prefer2)

##
## Hybrid (combination of both)                    In-Office
##                           95                           47

#0-95, 1-47

#confusion matrix for model5
newwfh2 =
data.frame(predicted=wfh$prefer2,actual=wfh$Preference)
newwfh1$predicted=as.factor(newwfh2$predicted)
str(newwfh2)

## 'data.frame':    142 obs. of  2 variables:
##  $ predicted: chr  "Hybrid (combination of both)" "Hybrid (combination
ofboth)" "In-Office" "Hybrid (combination of both)" ...
##  $ actual   : Factor w/ 2 levels "Hybrid (combination of both)",..:
1 1 11 1 1 1 1 1 1 ...

library(caret)
res2 =
confusionMatrix(newwfh1$predicted,newwfh1$actual)
res2

## Confusion Matrix and
Statistics##
##                                Reference
## Prediction                      Hybrid (combination of both) In-
Office##                          Hybrid (combination of both)        69    26
##   In-Office                                                      17        30
##
##                Accuracy : 0.6972
##                  95% CI : (0.6145, 0.7714)
##     No Information Rate :
0.6056##  P-Value [Acc > NIR]
: 0.01485##
##                   Kappa : 0.3478
##
##  Mcnemar's Test P-Value :
0.22247##
##             Sensitivity : 0.8023
##             Specificity : 0.5357
##          Pos Pred Value : 0.7263
##          Neg Pred Value : 0.6383
```

```
##              Prevalence : 0.6056
##          Detection Rate : 0.4859
##    Detection Prevalence : 0.6690
##       Balanced Accuracy : 0.6690
##
##        'Positive' Class : Hybrid (combination of both)
##
```

```
#confusion matrix for model5
#            hybrid   in-office
#hybrid       69         26
#in-office    70         30
#accuracy = 0.6972
#sensitivity = 0.8023
#Specificity = 0.5357
```

```
#ROCR curve for model 5
pred2 <- prediction(wfh$pred2 , wfh$Preference)
roc.pred2 <- performance(pred2 , measure = 'tpr' , x.measure = 'fpr')
plot(roc.pred2)
```




```
#our curve is close to y axis than x axis, therefore this model is good

#AUC for model 5
auc = performance(pred2 , measure ="auc")
auc@y.values[{1}]
```

```
## [[1]]
## [1] 0.7210341

#auc >0.7, hence this model is good (auc = 0.7210)

#Model 3 has better AUC, so we choose model 3 as the best model
```

Interpretation:

By the above analysis, we interpret the following things:

1. The best model, which can accurately predict the preference of a person is the one which combines the dependent variable preference with independent variables gender, profession, time management and productivity. This combination gives usan AIC of 187.25.
2. The confusion matrix formed gives us true and false positive and negative rates. The TPR and TNR should be high and the FPR and FNR should be low. This would mean that our accuracy of predicting the preference is good. Currently, we have an accuracy of 69.72.
3. Sensitivity and specificity are two factors which also help us in predicting the accuracy of the model. Sensitivity depicts how accurate the model is in predicting the positive class. That is, it depicts how many predicted Hybrid preferences are actually true. Specificity measures how exact the assignment to the positive class is, that is how many preferences are predicted incorrectly. In model 3, sensitivity is 0.7558, that means 75% predictions are true and specificity is 60.71 that is about 40% WFH predictions are actually hybrid.
4. Next is the ROCR curve, the better ROCR curve is one where the true positive rate is higher, that is the graph formed is more towards the Y-axis, in our case, it is true so our model has a good ROCR curve.
5. Lastly, AUC that is area under curve is considered good if the value is above 70%. Iour model 3, the AUC is 75.85%, so the model is a good one.

Conclusion:

With this Research, we can conclude that more people prefer a combination of work from home and in-office work culture than just working from home in this pandemic, the reason analytically might be several like facing connectivity issues, decrease in productivity or anything else, but in our opinion, there are other things also that influence their choice. One could be the screen time, which is extensive in WFH only, other could be family relations which can be maintained better in a combination of both, another could be that in only WFH a person gets the feeling of lack of freedom and in a hybrid, he can have the cake and eat it too. So, this concludes our study on Work from home culture that has been adapted extensively during this COVID-19 pandemic.

REAL WORLD.
REAL LEARNING.

ISBR BUSINESS SCHOOL BANGALORE CAMPUS

107, Near INFOSYS, Behind BSNL Telephone
Exchange, Electronic City - Phase I,
Bangalore - 560 100
Phone: 080-4081 9500