

# LIVE PROJECTS IN BUSINESS ANALYTICS USING R

---

**Edited and Co-authored by**

**DR SHEETAL MAHENDHER**

Head, Department of Business Analytics  
and Quantitative Techniques

# **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

<b>S. No.</b>	<b>Title - Names of Students</b>	<b>Page No.</b>
1.	An Evaluation on E Gaming in the Wake of the Pandemic <ul style="list-style-type: none"><li>- Humani</li><li>- Shashi</li></ul>	1
2.	Customers Preferences on Dinning Out in Restaurants During Covid-19 <ul style="list-style-type: none"><li>- K Jaswanth Achari</li><li>- Gopu Anifa Bala</li><li>- Bhagyasree Darsi</li><li>- Pediboina Srikant</li><li>- Nagalla Narmada</li></ul>	15
3.	Public Transport in Mumbai <ul style="list-style-type: none"><li>- Mahaveer Shukla</li><li>- Shamali Patil</li><li>- Somya Bansal</li><li>- Saumya Ranjan</li></ul>	21
4.	COVID-19 Awareness of People and its Impact on Lifestyle and Various Sectors of the Economy <ul style="list-style-type: none"><li>- Nikhil Rao</li><li>- Arindam Debnath</li><li>- Abhirupa Maiti</li></ul>	27
5.	Impact of Covid-19 on Factors Affecting Employee Engagement <ul style="list-style-type: none"><li>- Prachet Kulkarni</li><li>- Anirban Dasgupta</li><li>- Riya Ganguly</li></ul>	34
6.	Distance Learning and Keeping Connection for Students During the Coronavirus Outbreak <ul style="list-style-type: none"><li>- Richa Yaduka</li><li>- Lokesh Doda</li><li>- Himaja Reddy</li></ul>	46

<b>S. No.</b>	<b>Title - Names of Students</b>	<b>Page No.</b>
7.	Analysis of Risk of Heart Disease <ul style="list-style-type: none"><li>- Sanjana Kunjar</li><li>- Muthulakshmi Shunmugham</li><li>- Vignesh Krishnamoorthy</li></ul>	54
8.	Employee Satisfaction in Hospitality Industry <ul style="list-style-type: none"><li>- Shelaj Sharma</li><li>- Ritom Das</li><li>- Meghana Kalapala</li><li>- Harmanjeet Kaur</li></ul>	65
9.	Analysis of Viewing Movies and Series <ul style="list-style-type: none"><li>- Siddharth T</li><li>- Surendra Prasath S</li><li>- Komathisha K R</li></ul>	75
10.	The Impact of Employee Engagement in an Organization <ul style="list-style-type: none"><li>- Sunny Singh</li><li>- Taha Aktar</li><li>- Sagar Gadhawe</li></ul>	82
11.	Preferred Mode of Transportation Used by Different Segments of People <ul style="list-style-type: none"><li>- Utkarsh Kumar Gupta</li><li>- Diti Ghosh</li><li>- Reshma Chaudhary</li></ul>	89

# An Evaluation on E Gaming in the Wake of the Pandemic

Submitted By-  
Humani (PG19057)  
Shashi (PG19116)

## **Abstract:**

In this modern era, Technology plays an important role in the human's life. People are using the brand new technologies for information and entertainment which are providing wide ranges of happiness to the human community. Online gaming is becoming a form of both entertainment and socialization for youth, which becomes an inherent part of all their lifestyle activity. This study discusses on "An evaluation on online gaming in the wake of the pandemic". Online gaming addiction is a behavioral problem that has been classified and explained in numerous ways. In this Survey, The Respondents collected are 210 is the sample size. Also, we conclude that people prefer the online mode of gaming more over offline as Men are more likely to play Action gaming and married people do not prefer to participate in the online gaming competition. Online gaming is also more popular than social networking.

## **Introduction:**

Nowadays, web gaming is a noteworthy inclination on the planet. Anyone can play if he/she has gotten to the web. Web gaming is also more notable than long-range relational correspondence during this pandemic. The Electronic Game Industry is seeing a giant headway discreetly in the pandemic. The examination reveals that adolescents get to know web gaming through Advertisements, colleagues, family, and match social events. Normal, games could be brought from shops, as regularly as conceivable as a plate for use on a PC or console to play.

However, electronic games can similarly be downloaded in mobiles. Games are played on various plans. Web network in a game incorporates other lucky opportunities for gamers as it licenses players to find and play against, or with, various players around the world.

### **Online Gaming is Defined Based on:**

- The device used to play
- Games have categorization like a puzzle, action Strategy, adventure, sports, pulsate, skill-based.

### **Statement of the Problem:**

We mainly aim to study the level of online gaming addiction during this pandemic. Now because of this pandemic, teenagers are more addicted to the system which includes online games and they just get into their world of fantasies, and then they become less socialize.

### **Objective:**

- To study do People prefer online gaming rather than offline?
- To find out mostly which age group people are spending more time ingames?
- To find out which age group is spending more time playing games?
- To know which games mostly played online?

### **Research Methodology:**

#### **Research Design:-**

The Research Design followed for this research study is a descriptive research design, where we find a solution to an existing problem. Descriptive research is used to depict the presence of the business condition.

#### **Method of Data Collection:-**

- The data needed for the research study was collected by primary data.
- The method used for collecting data was a survey questionnaire.

**Sampling Design:-**

**Sample Size:** The Respondents collected are 210 is the sample size.

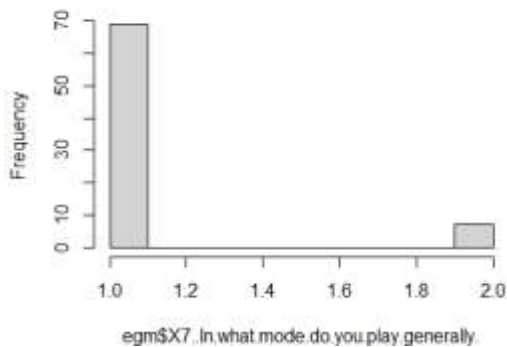
**Sample Method:** A method of sampling we used is **Convenience Sampling**. The main advantage of this type of sampling is the availability and the quickness with which data can be gathered.

**Source of Data:** Primary Data

**Analysis And Interpretation:**

**1) Do People prefer online gaming over offline?**

Histogram of egm\$X7..In.what.mode.do.you.play.generally



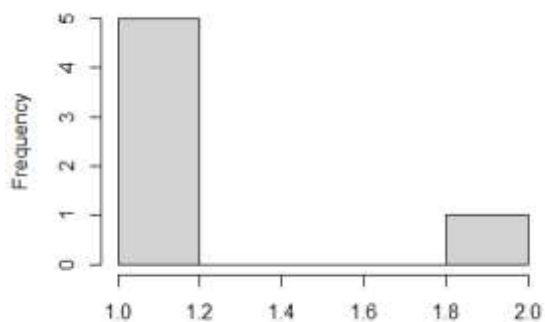
**INTEPRETATION:**

In a sample of 210 responses, People are more likely towards online gaming over offline.

**Source: Primary Data**

**2) Mostly which gender is playing Action games?**

..Gender[egm\$X13..Which.Category.of.the.games.do



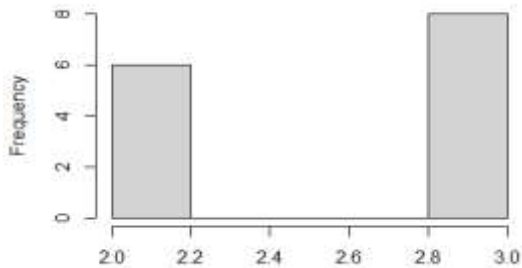
**INTEPRETATION:**

In a sample of 210 responses, Men are more likely to play the action games.

**Source: Primary Data**

### 3) Do Married people prefer to participate in online gaming competitions?

.Do.you.participate.in.e.gaming.competitions.[egm\$X4]



9. Do you participate in e.gaming.competitions [egm\$X4. Relationship]

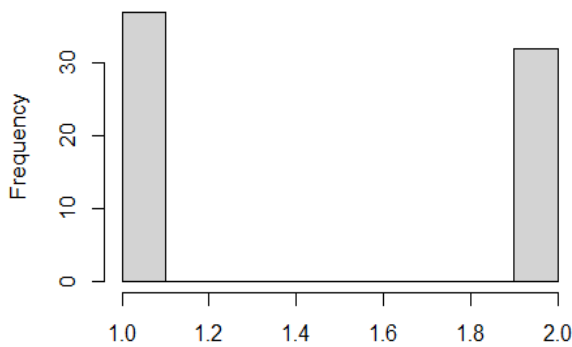
#### **INTEPRETATION:**

From the 210 responses, it is inferred that Married people do not prefer to participate in e gaming competition.

Source: Primary Data

### 4) Which Gender prefers online gaming more?

egm\$X2..Gender[egm\$X7..In.what.mode.do.you.play]



egm\$X2..Gender[egm\$X7..In.what.mode.do.you.play.generally. ==

#### **INTEPRETATION:**

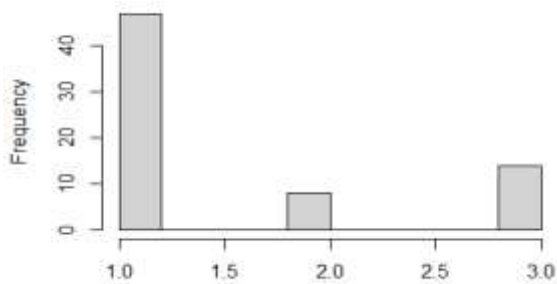
In a sample of 210 responses, it is inferred that Male prefers online gaming more.

Source: Primary Data



### 5) Which Profession prefers online gaming more?

egm\$X5..Occupation[egm\$X7..In.what.mode.do.you.p



egm\$X5..Occupation[egm\$X7..In.what.mode.do.you.play generally. =

#### INTEPRETATION:

From the sample of 210, it is inferred that Students prefer e gaming more.

**Source: Primary Data**

#### **Cleaning The Data:**

**Data Cleaning** is the process of transforming raw **data** into consistent **data** that **can** be analyzed. It **is** very important to understand how **you can** import **data** into **R** and save it as a **data** frame. It is aimed at improving the content of statistical statements based on the **data** as well as their reliability. **Data cleaning** may profoundly influence the statistical statements based on the **data**.

## LIVE PROJECTS- Predictive Analysis Using R

```

complete.cases(game)

game<-na.omit(game)
summary(game)

##           X2           X3           X4           X5
## Female:34  0-12 : 1  Married:14  Own Business : 9
## Male :42  18-24 :54  Single :62  Student :53
##           25-34 :19           Working Professional:14
##           Dec-17: 2
##
##
##
##
X6
## Have never played before but started since pandemic
:10
## I have always been a gamer and it has increased since pandemic
: 9
## I have always been a gamer but my habit is in control
:41
## I used to play, but stopped at some point, but started again since pandemic:16
##
##
##
##           X7           X8           X9           X10           X11
## Offline: 7  Console: 4  1 :28  Not sure:28  Heavy : 5
## Online :69  Mobile :67  2 :28  Weekdays:21  Light :45
##           PC : 5  3 : 9  Weekends:27  Medium:26
##           4 : 5
##           0 : 2
##           5 : 2
##           (Other): 2
##
##           X12           X13
## No, Never bothered :34  Casual Games :26
## Yes, and I play games suitable for my age:42  Battle Royale Games:17
##           Fighting Games : 7
##           Sport Games : 7
##           Racing Games : 6
##           Action Games : 4
##           (Other) : 9
##
##           X17           X18           X19           X20           X21           X2
## Friends and family:49  Maybe:27  Maybe:32  Maybe:40  Maybe:10  Maybe
: 9
## Solo Player :19  No :44  No :38  No :14  No :22  No

```

## LIVE PROJECTS- Predictive Analysis Using R

```
:13
## Strangers      : 8   Yes   : 5   Yes   : 6   Yes   :22   Yes   :44   Yes
:54
##
##
##
##
##      X23                      X24
## No   :17   No, I don't allow them. :18
## Yes :59   Yes, I allow them.       58
##
##
##
##
##
```

### Factoring The Data:

Factors represent a very efficient way to store character values, because each unique character value is stored only once, and the **data** itself is stored as a vector of integers. Because of this, read.table will automatically convert character variables to factors unless the as.is= argument is specified.

```
game$X2=as.factor(game$X2)
game$X3 =as.factor(game$X3)
game$X4 =as.factor(game$X4)
game$X5 =as.factor(game$X5)
game$X6 =as.factor(game$X6)
game$X7 =as.factor(game$X7)
game$X8 =as.factor(game$X8)
game$X9 =as.factor(game$X9)
game$X10 =as.factor(game$X10)
game$X11 =as.factor(game$X11)
game$X12 =as.factor(game$X12)
game$X13 =as.factor(game$X13)
game$X17 =as.factor(game$X17)
game$X18 =as.factor(game$X18)
game$X19 =as.factor(game$X19)
game$X20 =as.factor(game$X20)
game$X21 =as.factor(game$X21)
game$X22 =as.factor(game$X22)
game$X23 =as.factor(game$X23)
game$X24 =as.factor(game$X24)
```

X2= Gender,  
X3= Age  
X4= Relationship  
X5= Occupation  
X6= Please choose one of the following:  
X7= In what mode do you play generally?  
X8= On which platform do you generally play?  
X9= How many hours a day do you spend on gaming?  
X10= Are you a weekend player or weekdays player?  
X11= What type of gamer do you define yourself?  
X12= 12. Do you see the PG rating of the gaming before playing?  
X13= Which Category of the games do you preferably play?  
X17= With whom do you generally play?  
X18= Are you willing to spend on in-game purchases?  
X19= Do you participate in e-gaming competitions?  
X20= Are you likely to continue playing video game post covid situation?  
X21= Does playing an E-game seem more convenient to you other than playing outside?  
X22=Do you see e-sport and e-gaming taking over traditional games and sports?  
X23= Does kids at your home play e-games?  
X24= Do you allow kids to play e-games as physically playing outside is not possible?

```
str(game)
```

```
## 'data.frame': 76 obs. of 20 variables:
## $ X2 : Factor w/ 2 levels "Female","Male": 2 2 2 1 1 2 1 2 2 2 ...
## $ X3 : Factor w/ 4 levels "0-12","18-24",...: 2 2 2 2 2 2 2 2 2 3 ...
## $ X4 : Factor w/ 2 levels "Married","Single ": 2 2 2 2 2 2 2 2 2 2 ...
## $ X5 : Factor w/ 3 levels "Own Business",...: 2 2 2 2 2 1 2 2 2 2 ...
## $ X6 : Factor w/ 4 levels "Have never played before but started since pan
demic",...: 4 4 3 4 4 4 4 3 3 3 ...
## $ X7 : Factor w/ 2 levels "Offline","Online": 2 2 2 2 2 2 2 2 2 2 ...
## $ X8 : Factor w/ 3 levels "Console","Mobile",...: 2 1 2 2 2 2 2 2 2 2 ...
## $ X9 : Factor w/ 8 levels "0","1","2","3",...: 8 3 2 2 2 1 2 2 2 3 ...
## $ X10: Factor w/ 3 levels "Not sure","Weekdays",...: 2 3 3 3 1 1 1 1 1 1 ..
..
## $ X11: Factor w/ 3 levels "Heavy","Light",...: 1 3 3 2 2 2 2 3 2 2 ...
## $ X12: Factor w/ 2 levels "No, Never bothered",...: 2 2 1 2 1 1 2 2 1 1 ..
.
## $ X13: Factor w/ 11 levels "Action-Adventure Games",...: 4 4 2 5 11 6 6 4
4 4 ...
## $ X17: Factor w/ 3 levels "Friends and family",...: 1 1 1 1 1 2 1 1 3 1 ..
.
## $ X18: Factor w/ 3 levels "Maybe","No","Yes": 1 1 2 2 1 2 1 2 2 1 ...
## $ X19: Factor w/ 3 levels "Maybe","No","Yes ": 1 1 2 2 2 2 2 2 2 2 ...
## $ X20: Factor w/ 3 levels "Maybe","No","Yes": 3 3 3 2 2 2 1 2 1 1 ...
## $ X21: Factor w/ 3 levels "Maybe","No","Yes ": 3 1 3 2 1 2 2 2 1 1 ...
## $ X22: Factor w/ 3 levels "Maybe","No","Yes ": 3 3 3 3 3 2 3 1 1 3 ...
## $ X23: Factor w/ 2 levels "No","Yes ": 1 2 2 2 2 1 2 2 1 1 ...
## $ X24: Factor w/ 2 levels "No, I don't allow them. ",...: 2 2 2 1 2 1 2 2
1 2 ...
```

## PARTITION THE DATA TO TRAIN AND TEST:

```
library(caret)
```

```
partition<-createDataPartition (y=game$X12
,p=0.50, list=FALSE)training<-game[partition,]
test<-game[-partition,]
```

## MODEL BUILDING:

### MODEL 1:

```
model1<-glm(X12 ~X2 ,data=training, family =binomial())
summary(model1)
```

## LIVE PROJECTS- Predictive Analysis Using R

```
##
## Call:
## glm(formula = X12 ~ X2, family = binomial(), data = training)
##
## Deviance Residuals:
##   Min       1Q   Median       3Q      Max
## -1.354 -1.215  1.011  1.141  1.141
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.4055     0.5270   0.769   0.442
## X2Male      -0.3185     0.6723  -0.474   0.636
##
## (Dispersion parameter for binomial family taken to be 1)
##
##   Null deviance: 52.257  on 37  degrees of freedom
## Residual deviance: 52.032  on 36  degrees of freedom
## AIC: 56.032
##
## Number of Fisher Scoring iterations: 4
```

### MODEL 2:

```
model2<-glm(X12 ~X2 +X7 ,data=training, family =binomial())
summary(model2)
```

```
##
## Call:
## glm(formula = X12 ~ X2 + X7, family = binomial(), data = training)
##
## Deviance Residuals:
##   Min       1Q   Median       3Q      Max
## -1.354 -1.315  1.011  1.046  1.665
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.01160     1.35167  -0.748   0.454
## X2Male      -0.08701     0.70263  -0.124   0.901
## X7Online     1.41707     1.24468   1.138   0.255
##
## (Dispersion parameter for binomial family taken to be 1)
##
##   Null deviance: 52.257  on 37  degrees of freedom
## Residual deviance: 50.553  on 35  degrees of freedom
## AIC: 56.553
##
## Number of Fisher Scoring iterations: 4
```

**MODEL 3:**

```
model3<-glm(X12 ~X2 +X7 +X10 ,data=training, family =binomial())
summary(model3)
```

```
##
## Call:
## glm(formula = X12 ~ X2 + X7 + X10, family = binomial(), data = training)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5564  -1.0867   0.8410   0.9066   1.4692
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.7766     1.6152  -1.100   0.271
## X2Male         0.1195     0.7681   0.156   0.876
## X7Online       1.4399     1.3571   1.061   0.289
## X10Weekdays   0.9931     0.9244   1.074   0.283
## X10Weekends   1.0747     0.8142   1.320   0.187
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 52.257  on 37  degrees of freedom
## Residual deviance: 48.364  on 33  degrees of freedom
## AIC: 58.364
##
## Number of Fisher Scoring iterations: 4
```

**MODEL 4:**

```
model4<-glm(X12 ~X2 +X7 +X10 +X17 ,data=training, family =binomial())
summary(model4)
```

```
##
## Call:
## glm(formula = X12 ~ X2 + X7 + X10 + X17, family = binomial(),
##      data = training)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6643  -1.1334   0.6855   0.8369   1.7943
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3050     1.7133  -0.762   0.446
## X2Male         0.1030     0.8017   0.128   0.898
## X7Online       1.3854     1.4529   0.954   0.340
## X10Weekdays   1.1971     1.0215   1.172   0.241
## X10Weekends   0.7885     0.8652   0.911   0.362
```

## LIVE PROJECTS- Predictive Analysis Using R

```
## X17Solo Player -1.3819 0.9137 -1.512 0.130
## X17Strangers -0.1807 1.4457 -0.125 0.901
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 52.257 on 37 degrees of freedom
## Residual deviance: 45.792 on 31 degrees of freedom
## AIC: 59.792
##
## Number of Fisher Scoring iterations: 4
```

### MODEL 5:

```
model5<-glm(X12 ~X2 +X7 +X10 +X17 +X22 ,data=training, family =binomial())
summary(model5)
```

```
##
## Call:
## glm(formula = X12 ~ X2 + X7 + X10 + X17 + X22, family = binomial(),
## data = training)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -1.7130 -1.0303 0.2904 0.9274 1.9233
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 16.8840 2723.4030 0.006 0.995
## X2Male 0.1480 0.8902 0.166 0.868
## X7Online 1.3476 1.5437 0.873 0.383
## X10Weekdays 1.5088 1.1102 1.359 0.174
## X10Weekends 0.5825 0.9460 0.616 0.538
## X17Solo Player -1.8637 1.1724 -1.590 0.112
## X17Strangers -0.3425 1.5114 -0.227 0.821
## X22No -18.3554 2723.4023 -0.007 0.995
## X22Yes -18.1928 2723.4022 -0.007 0.995
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 52.257 on 37 degrees of freedom
## Residual deviance: 42.157 on 29 degrees of freedom
## AIC: 60.157
##
## Number of Fisher Scoring iterations: 16
```

### MODEL 6:

```
model6<-glm(X12 ~X2 +X7 +X10 +X17 +X22 +X24 ,data=training, family =binomial(
```

## LIVE PROJECTS- Predictive Analysis Using R

```
))
summary(model6)
```

```
##
## Call:
## glm(formula = X12 ~ X2 + X7 + X10 + X17 + X22 + X24, family = binomial(),
##      data = training)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68919  -0.88648   0.00012   1.01298   1.83448
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    36.4265  5225.4236  0.007  0.994
## X2Male          0.1280    0.9306  0.138  0.891
## X7Online        1.5307    1.5797  0.969  0.333
## X10Weekdays    1.3593    1.1932  1.139  0.255
## X10Weekends     0.3168    0.9689  0.327  0.744
## X17Solo Player  -2.0166    1.3468 -1.497  0.134
## X17Strangers    -0.1623    1.5169 -0.107  0.915
## X22No           -37.3742  5225.4232 -0.007  0.994
## X22Yes          -19.5890  4410.5844 -0.004  0.996
## X24Yes, I allow them. -18.4131  2802.1050 -0.007  0.995
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 52.257  on 37  degrees of freedom
## Residual deviance: 36.856  on 28  degrees of freedom
## AIC: 51.241
##
## Number of Fisher Scoring iterations: 17
```

### **BEST MODEL**

```
library(car)
```

```
outlierTest(model6) #data is normal as p<.05
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 36 2.513233          0.011963      0.4546
```

```
durbinWatsonTest(model6)
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.2372469 1.445228 0.096
## Alternative hypothesis: rho != 0
```



## LIVE PROJECTS- Predictive Analysis Using R

```
#p - value >.05, hence autocorrelation does not exist
```

```
#multicollinearity  $y=m_2x_1 + m_2x_2+m_3x_3+\dots m_nx_n+c$   
#VIF(>10, bad  
#(or)  $\sqrt{\text{vif}()}>2$  returns True, not good, all should be false.  
vif(model6)
```

```
##          GVIF Df GVIF^(1/(2*Df))  
## X2  1.281492e+00  1      1.132030  
## X7  1.441817e+00  1      1.200757  
## X10 1.691873e+00  2      1.140491  
## X17 1.484057e+00  2      1.103729  
## X22 6.864196e+06  2      51.185556  
## X24 6.864196e+06  1     2619.961013
```

```
 $\sqrt{\text{vif}(\text{model6})}>2$ 
```

```
##          GVIF    Df GVIF^(1/(2*Df))  
## X2  FALSE FALSE                FALSE  
## X7  FALSE FALSE                FALSE  
## X10 FALSE FALSE                FALSE  
## X17 FALSE FALSE                FALSE  
## X22 TRUE  FALSE                TRUE  
## X24 TRUE  FALSE                TRUE
```

```
#multicollinearity assumption is met
```

```
# use predict()  
pred1<-predict(model6, data = test, type = "response")  
summary(pred1)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
## 0.0000  0.4785  0.5676  0.5526  0.7540  1.0000
```

```
View(test)
```

```
# convert the predicted values by rounding them  
convert<-ifelse(pred1<0.5, "neg", "pos")  
head(convert)
```

```
##      3      4      5      6      7     10  
## "pos" "pos" "neg" "neg" "neg" "pos"
```

```
table(convert)
```

```
## convert  
## neg pos  
## 13 25
```

Model 6 comes out to be the best model.

**Conclusion:**

First, we clean the data and then factor the data. Then start building the model. We examine each model for testing and with the interpretation and testing; we conclude that the model 6 is the best model from all. The variables it contains are Gender, mode of play, weekend players, with whom they play, Effect on traditional games, playing online rather than physically. AIC Value of Model 6 is 51.241 and the data is normal as  $p < .05$ . As, Model 6 has the lowest AIC value. So, it becomes the best Model.

# Customers Preferences on Dinning Out in Restaurants During Covid-19

Submitted By-  
K Jaswanth Achari (PG19061)  
Gopu Anifa Bala (PG19090)  
Bhagyasree Darsi (PG19168)  
Pediboina Srikant (PG19086)  
Nagalla Narmada (PG19079)

## **Abstract:**

The outbreak of coronavirus all over the world had hit many changes in all the sectors, Especially the food industry had need to change their operations and daily activities restaurants to suit these post COVID-19 demands of their customers. People lifestyle and preferences had been changed drastically, This live project we would like to predict what are the demographic factors that will determine whether the people had any change in preferences post COVID.

This study helps restaurants to undergo changes with the help of these factors according to the people preferences so that they can retain customers, can get their business on track, can overcome the loss which occurred due to pandemic. We had done predictive analysis in this project by using the primary data based on survey, done a regression test with analysis and the best model which in terms give us the result.

## **Introduction:**

The Indian restaurant industry is rapidly transforming before our eyes, and restaurants will have to think their daily operations to suit these post COVID-19 demands of their customers. While demand will return rapidly as millions of Indians were craving their favorite's dishes it is as crucial to make necessary changes to restore consumer confidence and trust by rapidly

## LIVE PROJECTS- Predictive Analysis Using R

evolving the restaurant's approach in usage of technology. since the lockdown had been implemented in India people stopped going to restaurants as the fear made many people to cook by themselves instead of having food outside. The Restaurants industry had hit the huge loss after unlock people slowly started dinning out in restaurants but still people were expecting few changes in order to dine out in the restaurants. In this project, we would like to predict if demographic factors like gender, age, profession etc. will determine whether the people prefer to dine out in restaurants even if they dine out the variables they may expect like health and safety, hygiene, digitization post COVID- 19 we had taken inputs of the primary data which was collected across various parts of the country, irrespective of gender, age, profession etc. In this study we do analysis and predict preferences of customer preferences based upon the resulted outputs which help the restaurant to engage customers.

**2Analysis:**

```

getwd()
setwd("C:/Users/sai/Desktop/Desktop/R/live")

library(datasets)
hotel=read.csv("hotel Resp.csv")

View(hotel)
scaled_hotel=scale(hotel)
table(complete.cases(scaled_hotel))
summary(scaled_hotel)
table(is.na(hotel))

#library(caret)
library(nnet)
set.seed(205)

#Building Models

#model1
model1=glm(hotel$Dine.out.during.COVID. ~
hotel$Education+hotel$Income+hotel$Age+hotel$Gender , data = hotel)
summary(model1) #AIC = 295

```

```

Call:
glm(formula = hotel$Dine.out.during.COVID. ~ hotel$Education
+
  hotel$Income + hotel$Age + hotel$Gender, data = hotel)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8034 -0.5343  0.3042  0.4038  0.6607

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.14467    0.17054   6.712 1.94e-10 ***
hotel$Education  0.06216    0.04086   1.521  0.1298
hotel$Income    0.03320    0.02154   1.542  0.1248
hotel$Age      -0.01670    0.04129  -0.404  0.6863
hotel$Gender    0.11600    0.06907   1.679  0.0946 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2395
346)

    Null deviance: 49.756 on 204 degrees of freedom
Residual deviance: 47.907 on 200 degrees of freedom
AIC: 295.75

Number of Fisher Scoring iterations: 2

```

## LIVE PROJECTS- Predictive Analysis Using R

#As the p-value for all education, income, age, gender is more than 0.05. So, we consider them as insignificant to people in Dine out during COVID

```
#model2
```

```
model2 = glm(hotel$Take.children.to.dinning ~  
hotel$Marital.Status+hotel$Education+hotel$Age+hotel$Gender,  
data = hotel)summary(model2) #AIC = 255.31
```

```
Call:  
glm(formula = hotel$Take.children.to.dinning ~ hotel$Marital  
.Status +  
hotel$Education + hotel$Age + hotel$Gender, data = hotel  
)  
Deviance Residuals: 0.2303 0.2866 0.5562  
Coefficients:  
* hotel$Marital.Status 1.46868 0.14975 9.808 < 2e-16 **  
hotel$Education 0.06949 0.09840 0.706 0.48088  
hotel$Age 0.11564 0.03748 3.085 0.00232 **  
hotel$Gender -0.10513 0.04814 -2.184 0.03012 *  
---  
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for gaussian family taken to be 0.1966  
499)  
  
Null deviance: 41.59 on 204 degrees of freedom  
Residual deviance: 39.33 on 200 degrees of freedom  
AIC: 255.31  
  
Number of Fisher Scoring iterations: 2
```

#As the p-value for education and age is less than 0.05. So, we consider them as significant to people preferring to take children to dining during COVID

**5#model3**

```
model3 = glm(hotel$Team.out.dinning.during.COVID ~
hotel$Marital.Status+hotel$Education+hotel$Age+hotel$Income,
data = hotel)summary(model3) #AIC = 284.26
```

```
Call:
glm(formula = hotel$Team.out.dinning.during.COVID ~ hotel$Marital
.Status +
hotel$Education + hotel$Age + hotel$Income, data = hotel)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7789 -0.5594  0.2879  0.3592  0.5902
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.377145   0.142374   9.673  <2e-16 ***
hotel$Marital.Status -0.003107   0.105513  -0.029  0.9765
hotel$Education  0.102479   0.040226   2.548  0.0116 *
hotel$Age       -0.058541   0.051281  -1.142  0.2550
hotel$Income    0.017835   0.020960   0.851  0.3958
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 0.2264854)
Null deviance: 47.005  on 204  degrees of freedom
Residual deviance: 45.297  on 200  degrees of freedom
AIC: 284.26
Number of Fisher Scoring iterations: 2
```

#As the p-value for marital status, income, age, is more than 0.05. So, we consider them as insignificant to people prefer to go to team outing during COVID, only Education is contributing to it as the p-value is less than 0.05

**#model4**

```
model4 = glm(hotel$Celebrate.party.s.attend.during.COVID. ~
hotel$Age+hotel$Education+hotel$Marital.Status+hotel$Income,
data = hotel)summary(model4) #AIC = 278.4
```

```
Call:
glm(formula = hotel$Celebrate.party.s.attend.during.COVID. ~
hotel$Age + hotel$Education + hotel$Marital.Status + hotel$Income,
data = hotel)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8267 -0.5380  0.2531  0.3360  0.5985
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.60565   0.14107  11.382  <2e-16 ***
hotel$Age      -0.13203   0.05081  -2.599  0.0101 *
hotel$Education  0.08137   0.03986   2.041  0.0425 *
hotel$Marital.Status 0.07685   0.10455   0.735  0.4631
hotel$Income   -0.02462   0.02077  -1.185  0.2372
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 0.22235)
Null deviance: 46.712  on 204  degrees of freedom
Residual deviance: 44.470  on 200  degrees of freedom
AIC: 280.48
Number of Fisher Scoring iterations: 2
```

## LIVE PROJECTS- Predictive Analysis Using R

#As the p-value for education and age is less than 0.05. So, we consider them as significant to people preferring to celebrate and attend during COVID.

```
6 hotel1=data.frame(hotel)str(hotel1)
```

```
head(hotel1)
```

```
hotel1=glm(model2) library(car) library(carData)
```

```
#H0: There are no outliers in the data#H1: There are outliers in the data outlierTest(hotel1)
```

```
#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 = hotel1)) #Hence the data is normally distributed
```

```
#Multicollinearity
```

```
#Variance inflation factor should be less than 10 #squareroot of Variable inflation factor  
should be less than 2vif(hotel1)
```

```
sqrt(vif(hotel1))>2
```

```
#VIF is less than 10 for independent variables
```

```
#scatterplot matrix pairs(hotel$Take.children.to.dinning ~
```

```
hotel$Marital.Status+hotel$Education+hotel$Age+hotel$Gender, main = "Scatterplot  
Matrix")
```

### **Conclusion:**

As we conclude that, we were able to find the most significant variables which will determine whether Customer preferred to dine out during COVID – 19. As per analysis, The Education and Age factors are more influencing dinning out during COVID. These were the factors where they are more concerned about the taking children to dine out during COVID and Celebrating party's and attending them during COVID. These two factors are the most concerned ones and most influencing factors on Customer preferences to Dine out during COVID – 19.



# Public Transport in Mumbai

Submitted By –  
Mahaveer Shukla (PG 19071)  
Shamali Patil (PG 19115)  
Somya Bansal (PG 19166)  
Saumya Ranjan (PG 19111)

## Abstract

We have conducted a short study on local transport in Mumbai. Few of the aspects that we have focused are – Factors considered for travel, most affordable and convenient mode of commute, travel during pandemic, type of fuel used etc. For the purpose of this study, our method is qualitative data which we have taken through questionnaire. We have received 200 responses by this questionnaire.

## Introduction

Public transport in Mumbai involves the transport of millions of its citizens by train and road. As of 2015, 52% of commuters use public transport. Mumbai has the largest organized bus transport network among major Indian cities. Mumbai's public transport consists primarily of rapid transit on exclusive suburban railway lines augmented by commuter rail on main lines serving outlying suburbs, the bus services of the three municipalities making up the metropolitan area, public taxis and auto rickshaws, as well as ferry services.

## Local Transport In Mumbai

```
getwd() mumbai <-read.csv(choose.files()) View(mumbai) colnames(mumbai) mumbai  
summary(mumbai) str(mumbai)
```

## To check if there are any NAs in the data—

```
table(complete.cases(mumbai)) table(is.na(mumbai)) #203 True values, which means  
that there are no NA values .
```

## LIVE PROJECTS- Predictive Analysis Using R

```
mumbaiSex = as.factor(mumbaiSex)

mumbaiAgeGrp = as.factor(mumbaiAgeGrp)
mumbaiProfession = as.factor(mumbaiProfession)
mumbaiPreferred.mode = as.factor(mumbaiPreferred.mode)
mumbaiFactor.considered = as.factor(mumbaiFactor.considered)
mumbaiAffordable.mode...Local.Train. =
as.factor(mumbaiAffordable.mode...Local.Train.)
mumbaiAffordable.mode...Bus. = as.factor(mumbaiAffordable.mode...Bus.)
mumbaiAffordable.mode...Cab. = as.factor(mumbaiAffordable.mode...Cab.)
mumbaiAffordable.mode...Metro. = as.factor(mumbaiAffordable.mode...Metro.)
mumbaiConvenient.mode...Local.Train. =
as.factor(mumbaiConvenient.mode...Local.Train.)
mumbaiConvenient.mode...Bus. = as.factor(mumbaiConvenient.mode...Bus.)
mumbaiConvenient.mode...Cab. = as.factor(mumbaiConvenient.mode...Cab.)
mumbaiConvenient.mode...Metro. = as.factor(mumbaiConvenient.mode...Metro.)
mumbaiTrip.length = as.factor(mumbaiTrip.length) mumbaiFrequency =
as.factor(mumbaiFrequency.)
mumbaiFacilities..Platform. = as.factor(mumbaiFacilities..Platform.)
mumbaiFacilities..Ticket.Counter. = as.factor(mumbaiFacilities..Ticket.Counter.)
mumbaiAlternative.for.a.two.wheeler. =
as.factor(mumbaiAlternative.for.a.two.wheeler.)
mumbaiTransportation.during.pandemic =
as.factor(mumbaiTransportation.during.pandemic)
mumbaiWill.you.prefer.going.in.an.auto.late.night.in.Mumbai. =
as.factor(mumbaiWill.you.prefer.going.in.an.auto.late.night.in.Mumbai.)
mumbaiWhich.one.do.you.prefer.most. =
as.factor(mumbaiWhich.one.do.you.prefer.most.)
mumbaiWhy.do.you.prefer.it. = as.factor(mumbaiWhy.do.you.prefer.it.)
mumbaiAre.you.planning.to.change.what.you.ve.been.using. =
as.factor(mumbaiAre.you.planning.to.change.what.you.ve.been.using.)

str(mumbai) ## Correlation
cor.test(as.numeric(mumbaiAgeGrp), as.numeric(mumbaiPreferred.mode)) #p value
= 0.4778 i.e > 0.05 . So, correlation exists #r is -0.0501, less negative corr of -5.01%
cor.test(as.numeric(mumbaiProfession), as.numeric(mumbaiFactor.considered)) #p
value = 0.1251 i.e > 0.05 . So, correlation exists #r is 0.108007, less positive corr of
10.80%
cor.test(as.numeric(mumbai
Monthly.Income..yours.family.), as.numeric(mumbaiPreferred.mode)) #p value =
0.4691 i.e > 0.05 . So, correlation exists #r is -0.0511, less negative corr of -5.11%
cor.test(as.numeric(mumbai
Monthly.Income..yours.family.), as.numeric(mumbaiFactor.considered)) #p value
= 0.8414 i.e > 0.05 . So, correlation exists #r is -0.1413, less negative corr of -14.13%
cor.test(as.numeric(mumbaiAgeGrp), as.numeric(mumbaiAffordable.mode...Local.Tr
ain.)) #p value = 0.8904 i.e > 0.05 . So, correlation exists #r is 0.00973, less positive
correlation of 0.97%
cor.test(as.numeric(mumbaiAgeGrp), as.numeric(mumbaiConvenient.mode...Bus.))
#p value = 0.2075 i.e > 0.05 . So, correlation exists #r is -0.0888, less negative corr of -
8.88% cor.test(as.numeric(mumbaiAgeGrp), as.numeric(mumbaiTrip.length)) #p
```

```
value = 0.008437 i.e < 0.05 . So, correlation does not exist
cor.test(as.numeric(mumbaiProfession), as.numeric(mumbaiFacilities..Platform.)) #p
value = 0.4317 i.e > 0.05 . So, correlation exists #r is -0.0555, less negative corr of -
5.55%
cor.test(as.numeric(mumbaiProfession), as.numeric(mumbaiFacilities..Ticket.Counte
r.)) #p value = 0.2192 i.e > 0.05 . So, correlation exists #r is -0.0866, less negative corr of -
8.66%
cor.test(as.numeric(mumbaiMonthly.Income..yours.family.), as.numeric(mumbai
Which.one.do.you.prefer.most.)) #p value = 0.6278 i.e > 0.05 . So, correlation exists #r is
0.0342, less positive correlation of 3.42%
```

## Models

### Preferred mode is dependent variable

```
mumbai$Name<-NULL View(mumbai) model1 <-
glm(Preferred.mode~.,data=mumbai,family=binomial()) summary(model1) ## AIC of
the model is 261.77

model2 <- glm(Preferred.mode~AgeGrp+Profession+Sex+Trip.length, data=mumbai,
family=binomial()) summary(model2) ## AIC of the model is 265.69

model3 <-
glm(Preferred.mode~AgeGrp+Profession+Sex+Factor.considered+Trip.length,
data=mumbai, family=binomial()) summary(model3) ## AIC of the model is 264.37

model4 <-
glm(Preferred.mode~AgeGrp+Profession+Factor.considered+Trip.length+Transportati
on.during.pandemic+Will.you.prefer.going.in.an.auto.late.night.in.Mumbai.+Are.you.plan
ning.to.change.what.you.ve.been.using., data=mumbai, family=binomial())
summary(model4) ## AIC of the model is 260.36

model5 <- glm(Preferred.mode ~ Profession + Convenient.mode...Metro. + Frequency. +
Facilities..Platform. + Transportation.during.pandemic +
Will.you.prefer.going.in.an.auto.late.night.in.Mumbai. + Why.do.you.prefer.it. +
Are.you.planning.to.change.what.you.ve.been.using., family = binomial(), data =
mumbai) summary(model5) ## AIC of the model is 229.23
```

**Therefore, model 5 is the best with least AIC of 229.23**

### Prediction

```
mumbai$pred5=predict(model5, type="response") head(mumbai$pred5)
table(mumbai$Preferred.mode) View(mumbai)

#if pred > 0.35, take it as Personal Vehicles else Local Trains / Metr0

mumbaiconvert5 = ifelse(mumbaipred5>0.35,"Personal Vehicles","Local Trains /
Metr0") table(mumbai$convert5)

str(mumbai) View(mumbai)
```

```
####Confusion matrix---
```

```
table(actual=mumbaiWhich.mode.do.you.prefer.most.to.travel,predicted =
mumbaiconvert5)
```

```
#TN = 92 #FN = 17 #TP = 55 #FP = 39
```

```
#Sensitivity = tp/(tp+fn) = 0.76 #Specificity = tn/(tn+fp) = 0.70
```

**both specificity and sensitivity depicts that model 5 is good.**

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

```
mumbaiconvert5 = as.factor(mumbaiconvert5)
```

```
con.matrix5=confusionMatrix(mumbai
```

```
Which.mode.do.you.prefer.most.to.travel,mumbaiconvert5) con.matrix5 ##
```

```
Acuracy = 72.41%
```

## Validation of Model

```
install.packages("InformationValue") library(InformationValue)
```

```
somersD(mumbaiWhich.mode.do.you.prefer.most.to.travel,mumbaiconvert5) #
should be > 0.6
```

## Area under curve (auc) of receiver operator characteristic(roc)

```
install.packages("ROCR") library(ROCR) convert5=
```

```
prediction(as.numeric(mumbaiconvert5),as.numeric(mumbaiPreferred.mode))
```

```
roc.pred5=performance(convert5,measure = "tpr",x.measure = "fpr") #tpr is true
positive rate and fpr is false positive rate
```

```
plot(roc.pred5) #the graph has moved towards the y axis
```

```
auc=performance(convert5,measure = "auc") auc@y.values[1]
```

```
#n auc > 0.7 is good (cutoff) #0.7330 is the area under the curve. #it is a good model.
```

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

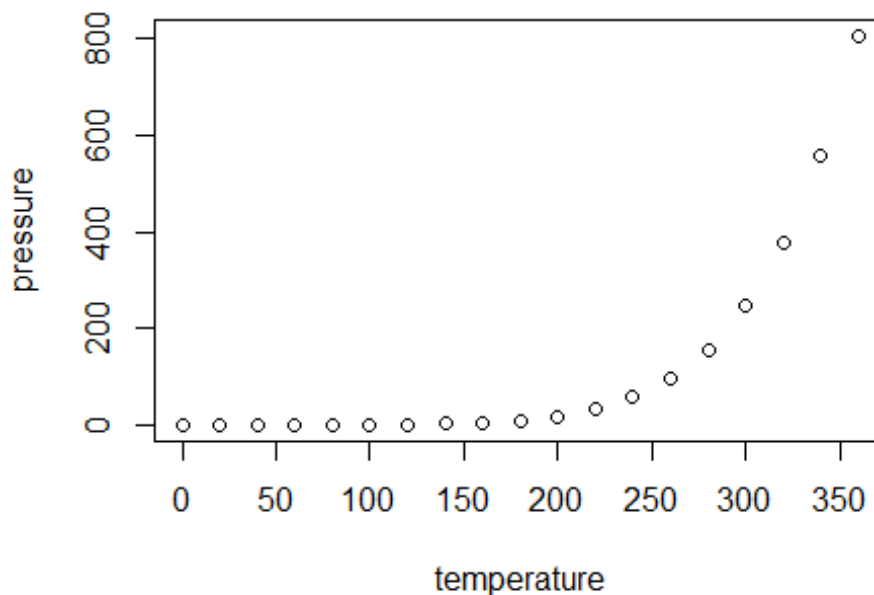
```
summary(cars)
```

```
##      speed          dist
##  Min.   : 4.0      Min.   :  2.00
##  1st Qu.:12.0     1st Qu.: 26.00
```

```
## Median :15.0   Median : 36.00
## Mean   :15.4   Mean    : 42.98
## 3rd Qu.:19.0   3rd Qu.: 56.00
## Max.   :25.0   Max.    :120.00
```

## Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

## Conclusion

We have represented the preferred mode of travel with income, age group, profession in order to understand how the preferred mode of travel is affected by these factors. On viewing our data, we saw that income, age group and profession affect the preferred mode of travel.

Apart from age, income and profession, we also checked whether choice of fuel also affects the preferred mode of travel. To conclude, there are several aspects that affect the mode of travel in Mumbai. According to our views, most people, irrespective of their gender, age, profession face some or the other issue while travelling and their choices also vary.

We have built a general linear model where 'Preferred mode' is the dependent variable. The frequency of preferred mode for travel by people is highly influenced by the following independent variables:

- AgeGrp

## LIVE PROJECTS- Predictive Analysis Using R

- Profession
- Trip length
- Factor considered
- Transportation during pandemic

Therefore, the general linear model built helps us predict that preferred mode of travel is local trains , metro, cab services or personal vehicle based on the above mentioned independent variables.

# **COVID-19 Awareness of People and its Impact on Lifestyle and Various Sectors of the Economy**

Submitted By-  
Nikhil Rao (PG19082)  
Arindam Debnath (PG19025)  
Abhirupa Maiti (PG19004)

## **Abstract**

### **Purpose:**

We all know a new respiratory disease called COVID-19 is spreading across the world and India is one of the most affected countries of COVID-19 pandemic. The government is trying to contain the spread of the disease but till now the affected and death rate has been increasing day by day in India. This pandemic has not only been taking human lives but also affecting Indian Economy poorly. Hence, we have conducted a research survey to understand the awareness of people about the pandemic and COVID-19 impact on their lifestyles and their views on how the economy and various sectors of the society will be affected by the pandemic in an Indian Context.

### **Design/Methodology:**

The study employed descriptive research method to understand the awareness among people regarding COVID-19 pandemic and how the pandemic impacted their lifestyle and Indian economy. The primary data collection method was used through a semi-structured questionnaire using Google form. A sample of 200 people which was the total population of participants was selected for this study including people from different states, age and profession.

R-studio 4.0.2 is used for hypothesis testing and data interpretation.

### **Finding:**

There were total twelve questions were used for hypothesis testing based on people awareness of COVID-19 spread and if there were any effects on lifestyle and Indian Economy due to COVID-19. The variables were mostly categorical variables. The most of the test results rejected null hypothesis as ( $p\text{-value} < 0.05$ ). So we concluded that Indian people were aware of COVID-19 spread and most of them agreed that this pandemic affected their lifestyles as well as Indian Economy.

### **Originality/Value:**

The paper provides original data on how people of India from different states, age groups and professions are thinking about COVID-19 spread, whether they are aware of the precaution to prevent Covid, how much their lifestyle has been impacted and impact on Indian Economy due to this pandemic.

### **Practical Implication:**

The paper will be informative to Government and Research Scholars who are further researching on impact of Covid-19 on people's lifestyle and economy.

**Paper Type:** Research Paper

**Key Words:** COVID-19, Lifestyle, Economy, Awareness

## **Introduction**

A new respiratory virus called the COVID-19 has been making headlines from 2019 end for causing an outbreak of respiratory illness throughout the world. The outbreak began in Wuhan, Hubei Province, China and quickly spread internationally. Millions of people have become sick and public health officials are keeping a close watch on how the virus is spreading. India is one of the most affected countries of COVID-19 pandemic. The government is trying to contain the spread of the disease but till now the infected and death rate has been increasing day by day in India. This pandemic has not only been taking human lives but also affecting Indian Economy poorly. Hence, we have conducted a research survey to understand the awareness of people about the pandemic and COVID-19 impact on their lifestyles and their views on how the economy and various sectors of the society will be affected by the pandemic in an Indian Context.

## **Research Purpose and Objectives**

The study demonstrates the following research questions:

- Whether the people of India are aware of Covid19 spread and government precaution?
- Whether people are taking any self-precaution (wearing mask, hand wash, avoiding crowded place, taking vitamins etc.) or not?



- If Covid-19 has any impact on lifestyle (Purchase decision, going out for movies, restaurant, pub, travelling)?
- If Covid-19 has any impact on various sectors of Economy (Large and small organization, unemployment, Train, Airways etc.)?

**Based on our research questions, the research objective is developed:**

- To study the awareness of people regarding Covid19 spread and Government precaution
- To understand what type of self-precaution people are taking
- To study the impact of Covid19 on people lifestyle and various sectors of Economy

**Methodology**

**Research Approach:** Quantitative and Descriptive research method was considered the most suitable for the purpose of investigation, which could provide the necessary insights into a new area of research. Quantitative research was concerned with the responses of participants. The primary data collection method was used through a semi-structured questionnaire using Google form.

**Research Participants:** A sample of 200 people which was the total population of participants was selected for this study including people from different states, age and profession. Probability sampling method was applied using systematic sampling method. An analysis of the demographic profile of respondents revealed that 57.7% of the respondents were male and 42.3% of the respondents were female. After analysing the respondent's age, it emerged that the largest group of respondents (59.7%) were aged between 18 to 25 years. The second highest group of respondents (22.4%) were aged between 26 to 35 years. Additionally, approximately 52.2% of respondents were from metropolitan cities and lowest 2.5% of respondents were from rural areas. It was also seen that 55.3% of total respondents were students and 20.1% were private sector employees by profession.

**Data Analysis:** R-studio 4.0.2 is used for hypothesis testing and data interpretation. Initially, demographic data of the subjects, factors as well as central tendency were established. Following this, a series of multivariate statistical procedures included hypothesis testing, correlation analysis and linear regression were performed on all the variables.

**Result & Discussion**

Our Questionnaire consisted of twelve questions. We did hypothesis test first. We also performed correlation and regression analysis to understand how much one variable can effect another. The result are as below:

Q1. Does location of stay has impact on people washing hand regularly?

We performed Chi-square test between location and washing hand data as both were two categorical variables and found out that location has impact on people habit of washing hand regularly in Covid time.

Q.2 If there is any decrease in buying food from outside for male population than female?

## LIVE PROJECTS- Predictive Analysis Using R

We performed Chi-square test between gender and online food order as both were two categorical variables. It was observed that Gender does not have significant effect on buying food from outside.

To understand better if there is any decrease in buying food from outside for male population than female, we use histogram and from the histogram we clearly understand that there is decrease in buying food from outside for male population than female

Q.3 Are people of India concerned about spread of covid 19 in their localities?

We performed Chi-square test between location and spreading concerned data as both were two categorical variables.

As per our test, location has impact on people spreading concerned.

Q4. Are people of India avoiding social gathering or crowded place?

Yes, according to our research study, people of India are avoiding crowded place.

Q5. Does location has an impact on safety concern adopted by people in buying products from stores?

We performed Chi-square test between location and safety concerned data as both were two categorical variables

We proved statistically that location has impact on people's safety concerned regarding purchasing product from stores.

Q6. Does large business/corporation will be affected according to region of stay.

We performed Chi-square test between location and large business data as both were two categorical variables and found that large business/corporation will be affected according to region of stay

Q7. Does gender have an effect on unemployment in the pandemic situation?

Gender does have a significant effect on unemployment in the pandemic situation

Q8. Does Small business/corporation will be effected according to region of stay?

Our testing revealed small business/corporation will be effected according to region of stay.

Q9. Is there a change in use of digital payment apps by men and women in last 10-15 days?

We performed Chi-square test between gender and digital payment data as both were two categorical variables and our inference is there is a significant increase of people using digital payment methods during COVID crisis.

Q10 .Does location of stay has impact on people availing out of home entertainment facilities?

There is a significant change of people availing out of home entertainment facilities, i.e. - it has decreased.

Q11. Does location of stay has impact on people availing food delivery options?

Thus there is no significant change of people availing food delivery.

Q12. Is there a difference in travel by gender in comparison to pre-Covid crisis?

Thus there is a significant decrease of people travelling right now.

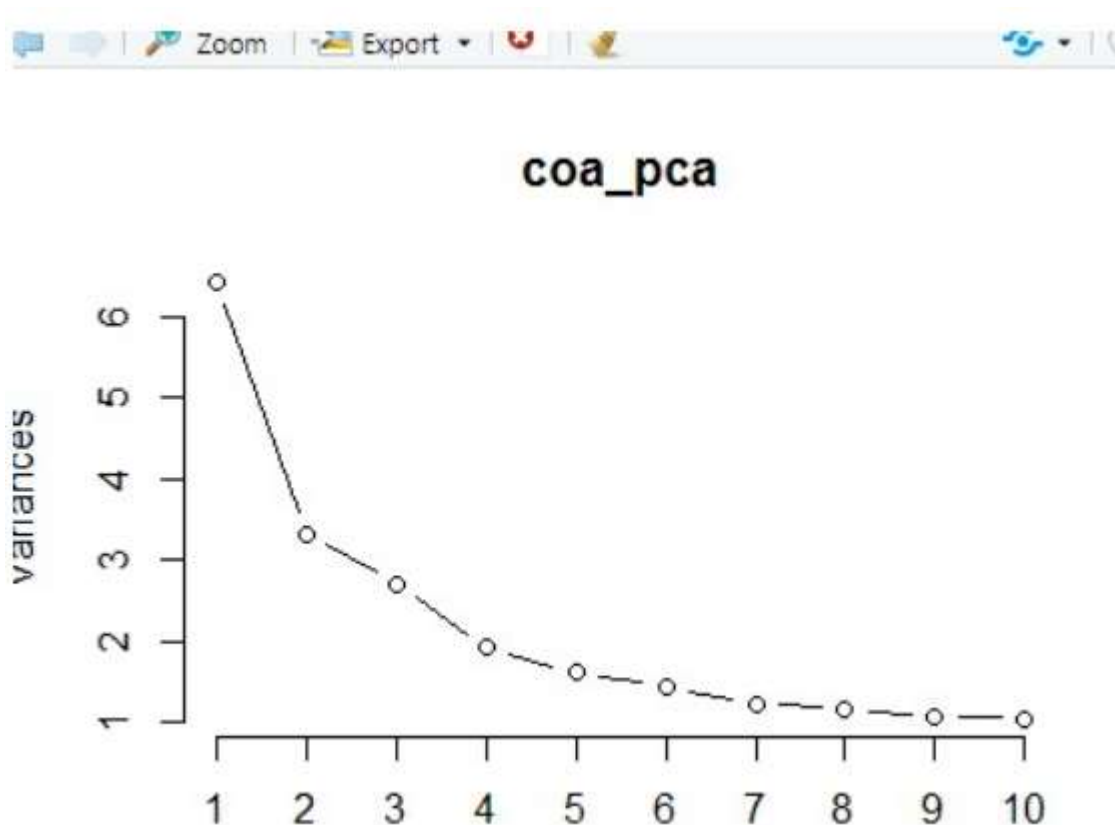
Regression: We have done logistic regression and created model for each section. The main sections are: General Awareness, Health Awareness, Purchase Decision and Economic Effect. For general awareness in the best model we saw that Region , Gender , healthcare confident, Stable duration and mask usage has best effect on general awareness of people regarding covid-19. As AIC is 116.26, lowest among other model, we have considered it as best model. The significant components are more here. We have done outlier test and Durbin Watson test. P-value is less than 0.05 for outlier test. So data is normal, no residuals are there. p -value  $>.05$  for Durbin Watson test, hence autocorrelation does not exist. Multicollinearity assumption is met too.

For Purchase Decision, we have created four models; the best model is model three. We can clearly see safety concerns while purchasing, wait for sale and avoid purchase has more effect on Purchase decision people during pandemic. As AIC is 135.71, lowest among other model, we have considered it as best model. The significant components are more here. We have done outlier test and Durbin Watson test. P-value is less than 0.05 for outlier test. So data is normal, no residuals are there. p -value  $>.05$  for Durbin Watson test, hence autocorrelation does not exist. Multicollinearity assumption is met too.

For economic conditions we have created four models, the best model is model four. We can clearly see large business affected, Finance expenditure, recession and Healthcare service has more effect on economic condition in our country during pandemic. As AIC is 136.47, lowest among other model, we have considered it as best model. The significant components are more here. We have done outlier test and Durbin Watson test. P-value is less than 0.05 for outlier test. So data is normal, no residuals are there. p -value  $>.05$  for Durbin Watson test, hence autocorrelation does not exist. Multicollinearity assumption is met too.

For Lifestyle usage habits we have created four models, the best model is model four. We can clearly see Digital pay app usage, cloths shopping, Gadget usage, and travelling rate has best effect on People lifestyle changes during pandemic. As AIC is 133.98, lowest among other model, we have considered it as best model. The significant components are more here. We have done outlier test and Durbin Watson test. P-value is less than 0.05 for outlier test. So data is normal, no residuals are there. p -value  $>.05$  for Durbin Watson test, hence autocorrelation does not exist. Multicollinearity assumption is met too.

We have also done factor analysis and here is the graph:



Upto 5 components has variance. There is an effect of components of best models as we can observe.

### **Practical Implication**

The paper will be informative to Government and Research Scholars who are further researching on impact of Covid-19 on people's lifestyle and economy. As Covid-19 is new in India and there are no vaccine available right now, containing the spread of the disease by taking preventive measures are only way to make people safe from this virus. Government is trying their best to create awareness but still a lot of people are not maintaining social distance and other rules. This study will give a clear idea on people awareness and what kind of measures they are taking. This study will give idea about the various sectors of economy which are affected due to pandemic which will help the government and organisation to overcome and revive business strategy. Also, this study gives idea about purchasing decision about different products which will help the companies to decide what kind of products should be in market right now.

### **Limitations and Future scope of the Study**

The main limitation of the study is the sample size. Due to time constraint, we could only collect 200 respondents. The main objective of our study is to understand the awareness of people regarding Covid19 spread and how much it has affected their lifestyle in Indian context. A sample size of 200 is very less for data analysis as India is a country where population is 135 crore. Also most of the samples are collected from urban and metropolitan cities and most of them are young population and are student or private sector employees by profession. To get a real scenario, we must have to collect data from different demographic variables too. In Future we want to continue our study and collect more samples for a better result.

### **Conclusion**

This report has discussed the COVID19 awareness of people and impact of lifestyle and various sectors of the economy in the pandemic situation. The objectives of this research survey to understand the awareness of people about COVID-19 impact on their lifestyles and their views on how the economy and various sectors of the society will be affected by the pandemic,

The objective was met by adopting an exploratory research study for our findings via Google form survey to have a better understanding of the pre- COVID-19 and existing lifestyles of people, this report includes interpretation of the result, including the new findings from the research, with proven hypothesis testing and data interpretation results, the result does support the hypothesis.

Finally, the overall significance of the project is to understand the awareness of people and impact of lifestyle and various sectors of the economy in the pandemic situation and as stated in the hypothesis that there will significance change in future.

### **Acknowledgement**

Presentation inspiration and motivation have always played a key role in the success of any venture

I would like to express my special thanks of gratitude to our research guide Dr. Sheetal Mahender for her constant support and guidance in completing the live project successfully.

It was a great experience working in a team with Arindam Debnath, Abhirupa Maiti and Nikhil Rao in the live project on: COVID-19 awareness of people and its impact on lifestyle and various sectors of the economy.

# Impact of Covid-19 on Factors Affecting Employee Engagement

Submitted By-  
Prachet Kulkarni (PG19066)  
Anirban Dasgupta (PG19016)  
Riya Ganguly (PG19102)

## Abstract

The main objective of the case study is to understand the impact of COVID-19 on employee engagement in an organisation. Successful employee engagement strategy creates a community at a workplace and not just a work force. The major study has been done by collecting information from employees who are working in different organisations and also from students in order to understand their viewpoints that how they should be treated while they will be working in an organisation. The main aim was to understand how the pandemic situation has affected the motivating factors of employee engagement. The researcher adopted descriptive research and the data is collected from the employees and students through convenience sampling method with the help of personally administrated questionnaire containing close ended questions having 5 pointer scale and the sample size is 160. This case study will make you understand how the employees mindsets are changing, in these new circumstances and what all they need to be more engaging towards work, post COVID-19; also students who have responded to this research topic will make a clearer view what millennials are actually expecting from companies which will make them work more smoother and faster as they are the generation of learning new things every day. This case study illustrates detailed focus on employee engagement with an impact of COVID-19 in order to give a broader perspective towards finding the problem, identifying challenges, analysis and solution by finding out feasible employee engagement for both employees and students.

## **Introduction**

“Employee Engagement is the state in which individual are emotionally and intellectually dedicated to the organization or company as measured by three primary behaviours: say, stay and strive”.

To become successful in today's world one requires a good bit more and good attendance. The Employees play a vital role in each and every organization. Likes and dislikes of employee will assist to achieve organizational objectives. The limit to which an employee believes in the mission, purpose and believes and values of an organization and demonstrates that commitment through their action as an employee and their attitude towards their employer and customer is Employee Engagement.

In the past 20 years companies had been trying to realize the benefit of empowerment, teamwork, recognition, people development, performance management and new leadership style. It is not the same that putting in place initiatives that have a goal of increasing employee engagement and truly seeing the payoffs whereas one might easily attribute low engagement to persistent downsizing, which lead to an erosion of loyalty and commitment. The working definitions of engagement largely defined in terms of how a person “feels inside”. But, when we ask people if the level of workplace engagement would be readily apparent to a visitor from the outside, they would say yes. One can observe levels of excitement and energy, observes people going to extra length to solve customer issues, and one can see an ethic of quality and continuous improvement. Similarly, workplace behaviours indicative of low dedication to work result in whining, low energy, passive-aggressive behaviour, lack of teamwork etc are also visible. With the dynamic changes as brought by COVID-19, we’re trying to focus on how much it has impacted the factors affecting employee engagement.

## **Review of Literature**

1. Sudhesh Venkatesh, HHR at TESCO HSC views employee engagement as a psychological association. Success is because to a corporate culture that support individual creativity as well as team work, paradox studies measure employee engagement term two dimensions: how employees feel (their emotion towards the company, the leadership, the work environment) and for how they intend to cut in the future (will they stay, give extra efforts).
2. Ken scarlet, president and CEO of scarlet international: Employee engagement will make employee more contributed, more empowered, more loyal and will give the benefits such as high morale, happy environment and lower attrition rates. Organization can achieve employee bliss through employee engagement.
3. The conference board New York: author (JOHN GIBBONS) published 2006: This summarizes what is known on the topic of employee employment and what companies can do to foster true engagement in the work place. It provides a review of current research on their important and timely topic when workers feel mentally and emotionally connected to their jobs, they are willing to apply discretionally effort to their company success.

4. Scottish Govt. publication's 2007 (May) There is no discernible difference between the dynamics of engagement within the public sector rather difference in engagement level is result from organization characteristics, which level sectors that organizational site.

5. Human capital strategy volume-9; No.3 August 2005: This article summarized engaged employee begets satisfied customers. This in turn improves the profitability of the organization. HR should help in identification and reengagement of disengaged employee by launching special initiatives directed towards bringing this group of employees into the maintenance.

The Study conducted by A. Marcus and Namitha M. Gopinath, following are the findings:

The various drivers of employee engagement are organization, management, superior, career development, reward and

recognition, performance appraisal, training and monetary benefits. However, the recent studies conducted in employee engagement hints that there is a change in the relevance of the drivers of employee engagement. The impact of superior, reward and recognition and performance appraisal on employee engagement have been highly discussed worldwide. According to the 2013 Blessing White report on employee engagement, there is an increase in trust of employees in superiors. The top contribution drivers according to this survey were superior, reward and recognition and performance appraisal. When an employee contributes, and is recognized for his/her contribution, it naturally drives employee engagement.

The study also suggested the re-assessment of performance appraisal strategies to ensure that it does not hinder the engagement efforts.

### **Methodology: -**

We had conducted a primary research by taking data from about 160 respondents of varying age groups, belonging to either categories of being employed, unemployed or students.

The type of research we conducted was primarily qualitative in nature. Qualitative research involves collecting and analysing non-numerical data (e.g., text, video, or audio) to understand concepts, opinions, or experiences. It can be used to gather in-depth insights into a problem or generate new ideas for research. Responses were collected based on a 5-pointer scale.

Various tests were conducted after assuming necessary regression models, correlation, confusion matrix, ROCR curve and AUC and lastly factor analysis for the various aspects of employee motivation which were then measured against the demographic factors, to arrive at our necessary inferences.



## Analysis

Since COVID-19 has created a great impact in our lives, we have decided to find the relation in the Motivation of Employees pre COVID and post COVID and the Employee Engagement. After Data Smoothing, we have taken Employee Motivation in Pre and Post COVID as a Dependent Variable.

We have then created a model using 'glm' function. For which the AIC has come to 169.11 which was least.

Now to get a better understanding of the variables involved in the model we have given them alphabetical codes which goes as follows: -

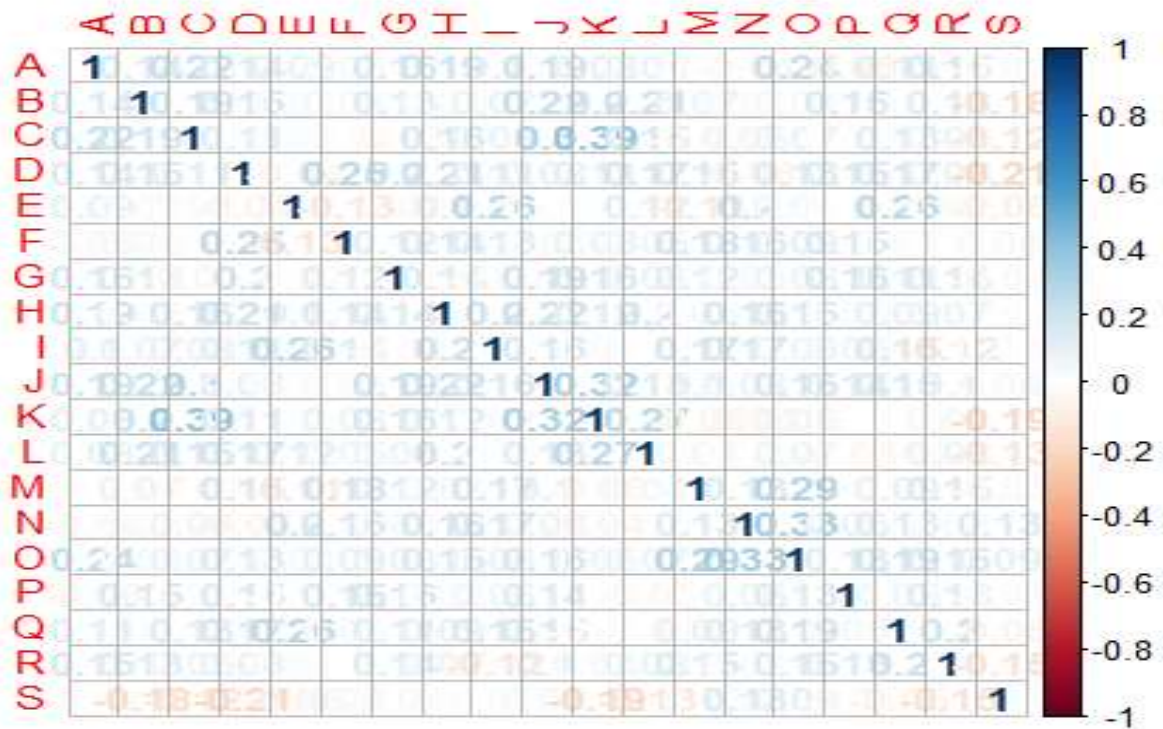
Sr no.	Variable	Code
1	Communication	A
2	Potentiality	B
3	Team Work	C
4	Formal Feedback	D
5	Informal Feedback	E
6	Coaching Feedback	F
7	Managers Motivating Employees	G
8	Team Leads Motivating Employees	H
9	Customers Motivating Employees	I
10	Employee Engagement through Celebrating Culture	J
11	Employee Engagement through Annual Programmes	K
12	Employee Engagement through Seminars and Webinars	L
13	Problems faced - Remote Working	M
14	Problems faced – Less use of Skills	N
15	Problems faced – Disengagement	O
16	Workplace Culture	P
17	Brand Name affecting perception	Q
18	Effect of Favouritism on Employee Engagement	R
19	Motivation same now as pre COVID	S

Table 1.1

**Correlation: -**

Plot of Correlation of all the above variables was done using the function – ‘`corplot(cor(dataset))`’

The plot is as follows-



Few components which stand out are there is a moderate positive relation (0.39) in between C and K which is Team Work and Annual Programmes resulting in increasing of Employee Engagement. Which explains how Team work can be enhanced in Annual Programmes. Another component which shows the relation amongst them are N and O which are Less use of Skills and Disengagement. They have a moderate positive correlation of 0.33 which will explain the relation when there is less use of skill there is Disengagement which affects the Employee Engagement.

Similarly, it can be checked for other factors as well.

### Regression Models: -

Here 'S' being categorical we have to use 'Logistic Regression' to build the model. With the help of StepAIC function we can find the best model of the data keeping 'S' as the Dependent Variable.

$$\text{Mod.1} = \text{glm}(S \sim ., \text{data} = \text{res}, \text{family} = \text{binomial}())$$

Here in Model 1 the AIC is 176.5 which can be further reduced as the model develops

With the help of this Mod.1 we can find out all the significant factors. Another method of doing the same is StepAIC – in which R itself gives us the best model with the least AIC.

$$\text{Mod.2} = \text{glm}(S \sim \text{Age} + \text{Gender} + \text{Occupation} + A + B + C + D + E + F + G + J + K + L + M + N + O + P + Q + R, \text{family} = \text{binomial}(), \text{data} = \text{res})$$

Here in Model 2 the AIC has come down and reduced to **169.11**.

---

### Prediction: -

One of the main functions of performing logistic regression in R studio is that we can predict the outcomes and validate them with the help of some tests and plots.

Here in the study we have predicted whether the employee has the same Motivation post COVID or not. For which we performed certain tests and used certain functions such as 'predict' to predict the outcome.

After performing the function an additional column is created in our dataset which implies the predicted values but they are in decimals. Hence in order to convert them in 0 and 1 we use 'ifelse' function.

Once this is done we can get the predicted values in the format of 0 and 1. Hence our actual values i.e res\$\$ and our predicted values i.e res\$pred are in 0 and 1.

---

### Validation: -

Now that we have predicted our values as to if the employee will have the same motivation as pre COVID or not we have to validate our outcomes and check for the accuracy using

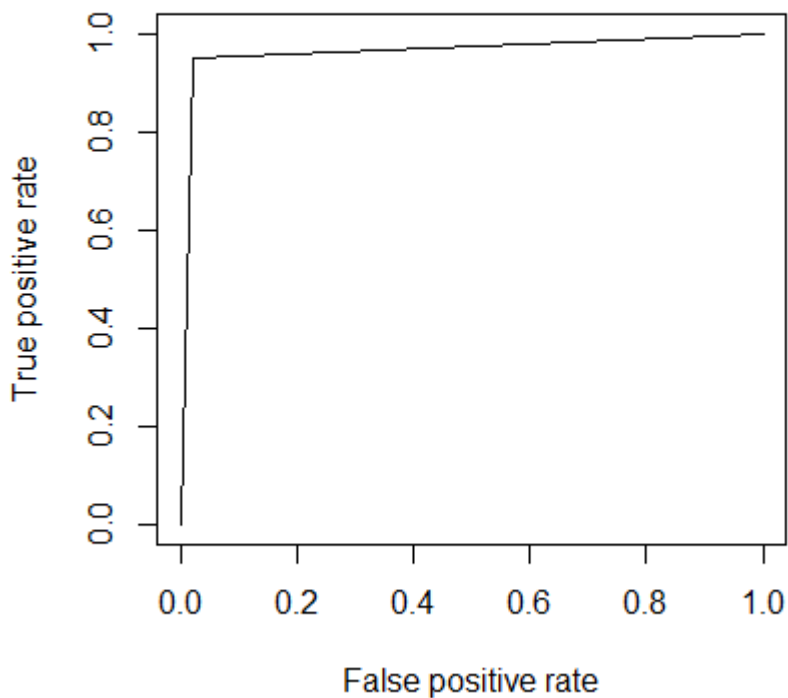
## LIVE PROJECTS- Predictive Analysis Using R

Confusion Matrix. In confusion Matrix we have used the function 'table' in which we put the Actual values and the predicted values and this is the outcome we have got –

		Predicted	
		0	1
Actual	No	95	2
	Yes	3	60

From the above table we can analyse that 95 observations were true negative which means we predicted it as No and actually they are No and 60 observations were actually Yes and we predicted as Yes. This will lead us to Specificity and Sensitivity which is 97.93% and 95.23% respectively. This indicates the accuracy of the model.

Another way of validation of the model is ROCR Curve and Area Under Curve which can be obtained by using 'performance' function. And the Area Under curve was 96.58% and the Plot is as follows: -



Since the curve is very close to Y-Axis and the AUC is 96.58% this indicates that the model is very good and accurate.

**Summary Analysis for Model 2: -**

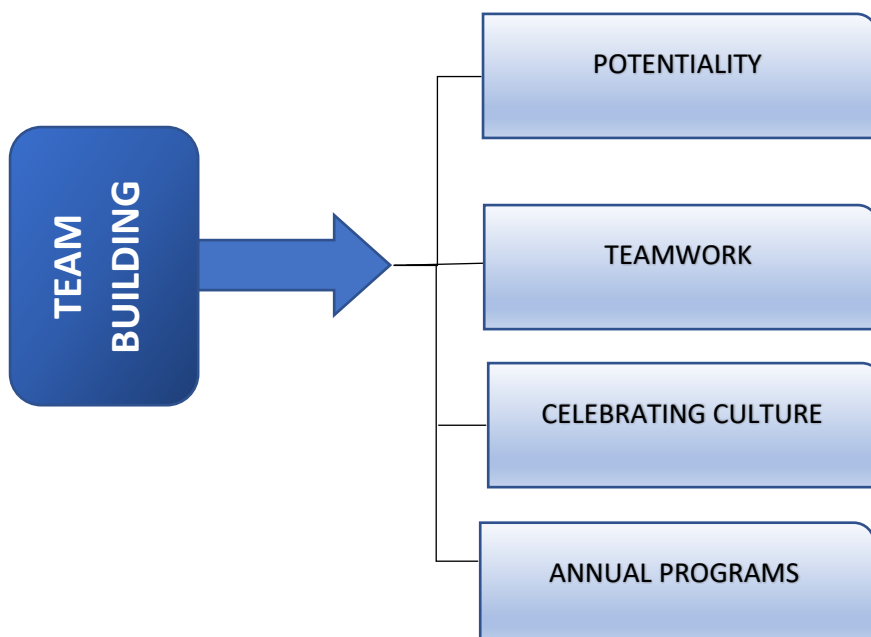
Sr No.	Parameter	Output
1	AIC	169.11
2	True Negative	95
3	True Positive	60
4	Specificity	97.93 %
5	Sensitivity	95.23 %
6	Area Under Curve	96.58 %

**DISCUSSION**

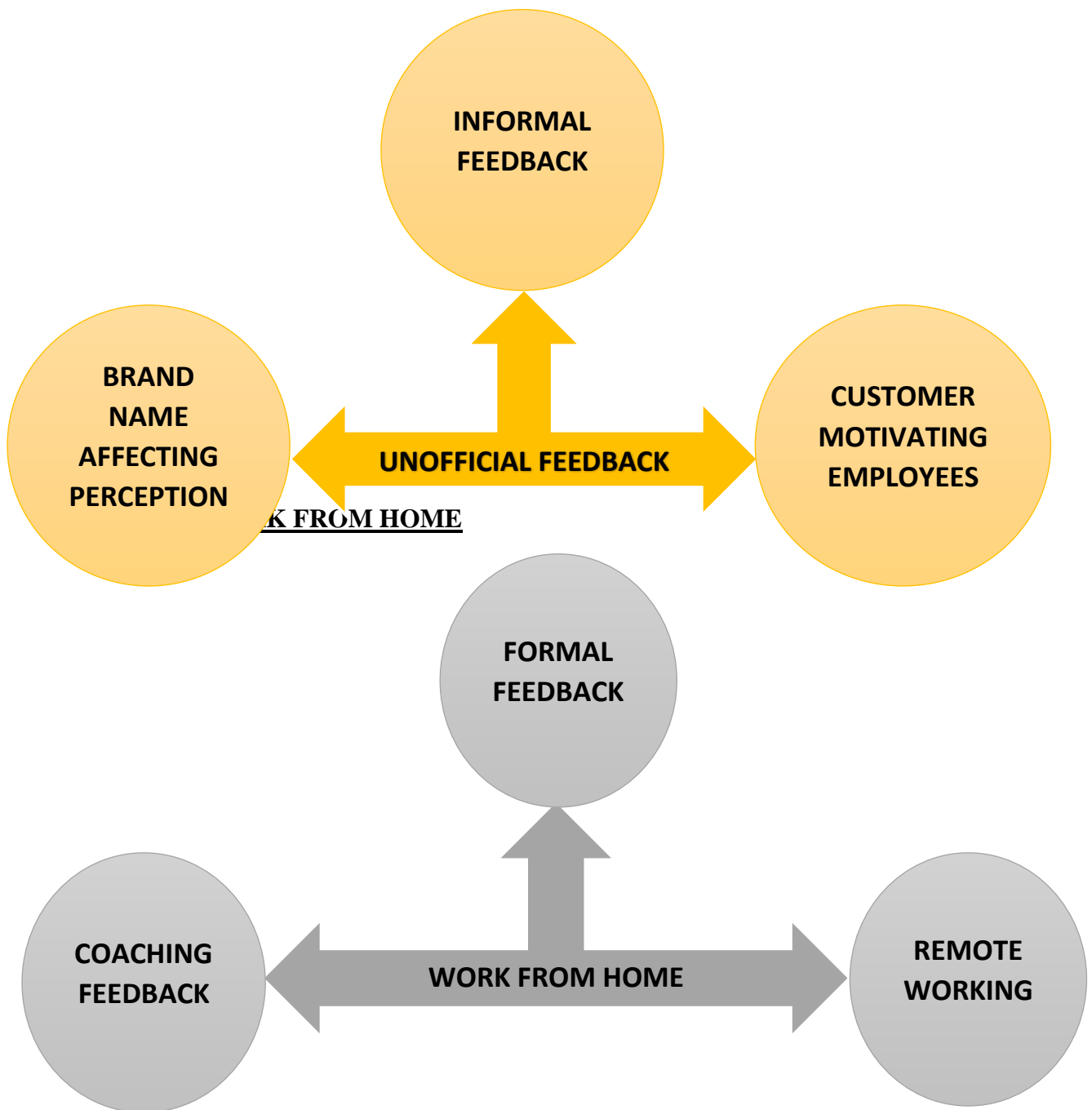
We conducted a factor analysis and characterized the attributes into 4 factors:

**Diagrammatic Representation**

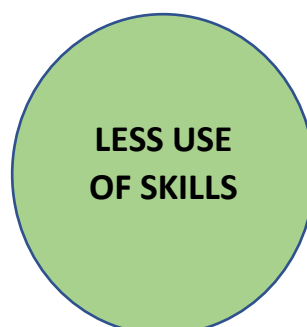
**Factor 1 – TEAM BUILDING**



**FACTOR 2 – UNOFFICIAL FEEDBACK**



**FACTOR 4 – STAGNATION**





### **Factor 1: Team Building**

1. Team work (C)
2. Celebrating Culture (J)
3. Annual Programme (K)
4. Potentiality (B)

The first factor involves attributes towards team building, which positively affects employee engagement. While employees are working from home, team work becomes all the more difficult. The absence of the physical presence, is a new element in everybody's lives. Thus, team work becomes an essence, for effective team building. Potentiality of each member also suggests the calibre and how well the member syncs in with team; also forms an essential part towards team building. Nowadays teams are cross-cultural in nature, having employees from different caste and creed; and with the ensuing lockdown, it has become important to boost their morale from preventing the work becoming mundane. Thus, celebrating cultural and conducting annual programme, not only helps in effective team building; but also makes the employees feel inclusive, even when they are miles apart; and additionally, boosts their morale and keeps them positively engaged towards their work.

### **Factor 2: Unofficial feedback**

1. Informal feedback (E)
2. Customer motivating employees (I)
3. Brand name affecting perception (Q)

While formal employee feedback is desirable, there are situations in virtually every employee's workday where informal feedback can be surprisingly effective. That's why managers need to adopt certain guidelines about providing feedback, so the process becomes comfortable for both them and their employees - and is aimed at improving performance, not inhibiting it with an onslaught of negative comments. The feedback from customers, helps employees understand the degree of their efforts and a good feedback reinforces the employees to become more engaged towards their work. The

brand name affecting perception of employees acts as another element; as employees feel privileged and positively fortify the employees to do better at their job.

These acts as an added incentive towards engaging employees towards their obligations and moreover, at a time when employees are confined to their respective homes, such can prove as an effective tool towards employee engagement.

### **Factor 3: Work from home**

1. Formal feedback (D)
2. Coaching feedback (F)
3. Remote working (M)

As employees have been engaged in remote working for the past months, feedback mechanism has been one of the most effective tools, to keep them constantly motivated and positively building towards effective employee engagement. To make feedback more effective, managers can reinforce it with coaching. When managers feed their employees with information on goals and objectives and expected behavior, it is necessary that they can empower them with the right skills and knowledge. This can be achieved through ongoing feedback and coaching. Regular coaching with regular feedback can double their performance and improve their confidence. Managers must not just identify gaps in performance but ensure that they fill it with the use of coaching. Effective feedback and coaching increase employee motivation and initiative; thus, positively attributing towards employee engagement.

### **Factor 4: Stagnation**

1. Less use of skill (N)
2. Disengagement (O)
3. Motivation during COVID-19 (S)

While we have discussed the positive impacts that has impacted employee engagement, stagnation at work, is one of the attributes that has negatively impacted employee motivation. The initial period of the lockdown had kept the employees confined for a seemingly indefinite period of time. As a result of this, their scope for using their skills (especially in case of physical labour) declined considerably, thus rendering them in a stagnated state. Such stagnation also comes from disengaged employees, as lack of physical contact and continuous guidance can lead to employees being disengaged and oblivious from their duties.

These also results in a downfall of their motivation, creating a pitfall in their mental health. So steps must be taken by the organisation to ensure that such stagnation doesn't continue for a long time, lest it would become harder and harder to keep the employees engaged.



## Conclusion

Humans are the most vital resource of any organisation. Effective management of them has a major impact on the success of any organisation. Thus, to get the maximum output from any employee, it is essential to engage them efficiently towards their jobs.

There are many factors which forms as different aspects of employee engagement. Now due to the volatile and unpredictable nature of humans, we can never generalize the impact of any factor over motivation that how it will be post covid. Our study thus strives to understand the relation in the Motivation of Employees pre COVID and post COVID and the Employee Engagement., to understand the perspective of each individual in regards to the factors and to understand how can we train them in this condition to make them back into normal life which can make them easy to develop their motivation. Our analysis was done by regression and correlation analysis, created models to understand which is the best AIC and also, we did factor analysis in order to find out best 4 factors in order to understand how the overall factors can be named under 1 factor to get analysis. Thus, we can state that overall, the employee motivation has been arising 4 factors and specificity of our models which states that employee motivation is being affected due to post COVID-19.

## References

- Gibbons, J. (2006). Employee Engagement: A Review of Current Research and Its Implications. The Conference Board, New York, pp. 1-21.
- Human capital strategy volume-9; No.3 August 2005
- Ken scarlet. (2009). Impact of demographic factors on employee engagement. MPRA Paper No. 39768
- Marcus, A & Gopinath, Namitha. (2017). IMPACT OF THE DEMOGRAPHIC VARIABLES ON THE EMPLOYEE ENGAGEMENT - AN ANALYSIS. ICTACT Journal on Management Studies. 03. 502-510. 10.21917/ijms.2017.0068.
- [https://mpr.ub.uni-muenchen.de/39768/1/MPRA\\_paper\\_39768.pdf](https://mpr.ub.uni-muenchen.de/39768/1/MPRA_paper_39768.pdf)
- Sudhesh Venkatesh. (2018) Guest Editorial ;Volume: 11 issue: 1, page(s): 5-6
- Scottish Govt. publication's 2007 (May)

# Distance Learning and Keeping Connection for Students During the Coronavirus Outbreak

Submitted by-  
Richa Yaduka (PG19097)  
Lokesh Doda (PG19068)  
Himaja Reddy (PG19054)

## Abstract

In the wake of the coronavirus (COVID-19) pandemic, businesses, schools and colleges have had to dramatically shift on how they operate. In fact, nearly all students currently enrolled in higher education programs had in-person classes cancelled because of coronavirus (COVID-19).

Yet, the learning hasn't stopped; students are still being assigned coursework from home. We wanted to learn how this transition is going and what support students feel they need right now. The major study has been done by collecting information from students who are studying in different schools/colleges. The main aim was to understand how the pandemic situation has affected the learning and their connection with family and instructors.

## Introduction

As the COVID-19 pandemic has closed schools and led to a rapid transition to online classes, teachers have been working diligently to adapt lesson plans to support virtual learning. While in school buildings, teachers can see students and talk to them to gauge how they are doing, in virtual classrooms, it is considerably harder for a teacher to assess a student's mental health or state of mind.

Previous studies have reported that students may use various technologies for e-learning in

their chosen settings, while some of the assigned technologies may sometimes be neglected in favor of their own mobile technologies. Whereas technologies-in-practices are seen to be changeable over time as students' knowledge, experiences, contexts, and technology itself might undergo changes through human action. Although extensive Covid research has been carried out on open distance learning, no single study exists which deals about the good, the bad and the ugly of distance learning in higher education.

### Objective of the Study

An effective connection is the key tool for driving organizational effectiveness and forms a key driver for one's own survival in long run, competitiveness and profitability.

- To understand the impact of Covid on students.
- To build best model and understanding the factors which are affecting Distance learning.

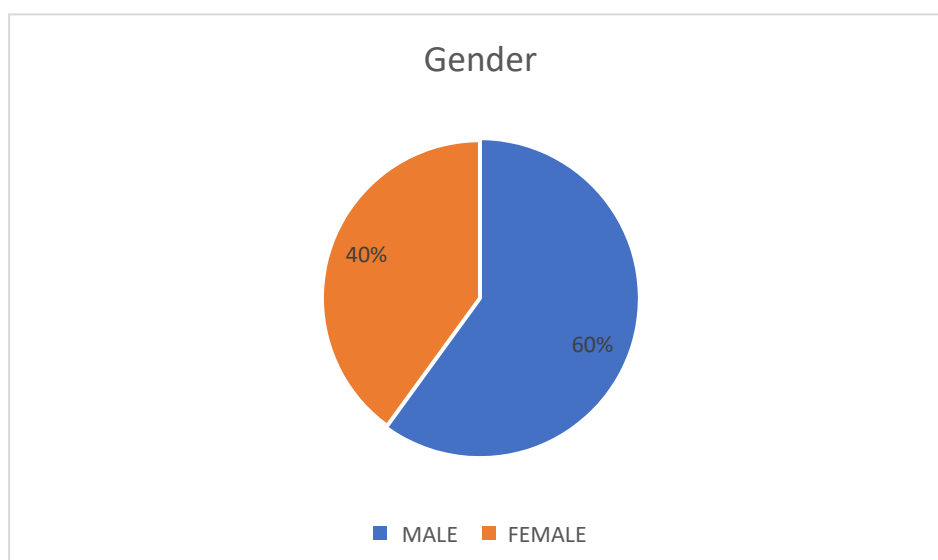
### Methodology

We had conducted a primary Covid research by taking data from about 160 Covid respondents of varying age groups, belonging to either categories of being employed, unemployed or students.

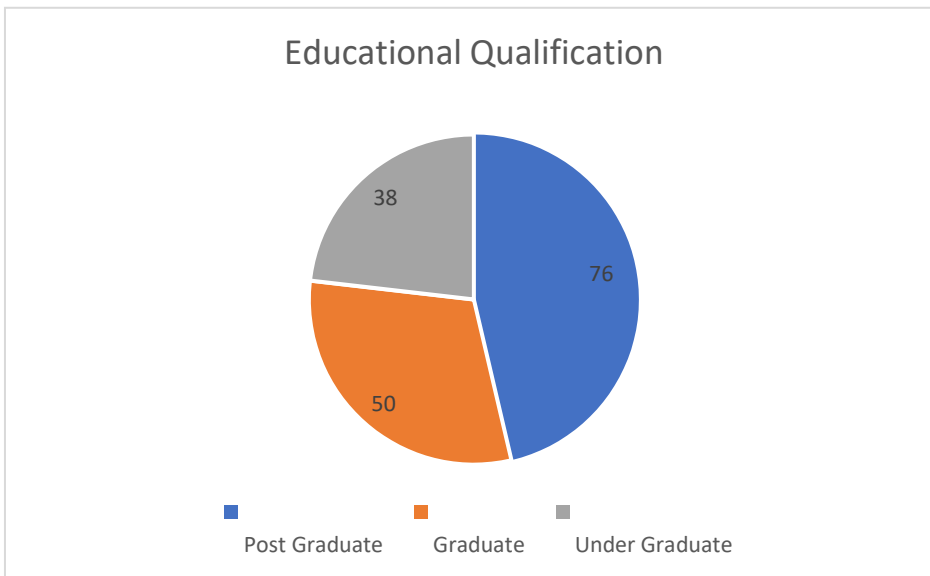
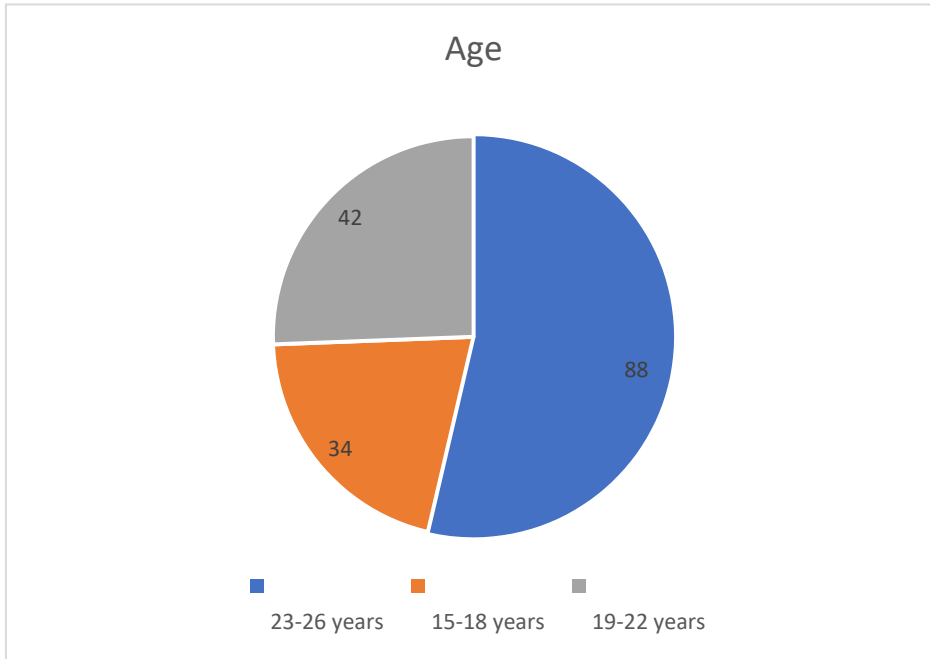
The type of Covid research we conducted was primarily qualitative in nature. Qualitative Covid research involves collecting and analysing non-numerical data (e.g., text, video, or audio) to understand concepts, opinions, or experiences. It can be used to gather in-depth insights into a problem or generate new ideas for Covid research. Covid responses were collected based on a 5-point scale.

Various tests were conducted after assuming necessary regression models, correlation, confusion matrix, ROC curve and AUC and lastly factor analysis for the various aspects of employee motivation which were then measured against the demographic factors, to arrive at our necessary inferences.

### Demographic Profile



# LIVE PROJECTS- Predictive Analysis Using R



## Analysis

We have created a model using 'glm' function.

The variables involved in the model we have given them alphabetical codes which goes as follows :

A=Following News  
B=News Sources  
C=Which is the most used way of your connection?  
D=Do you feel connected with your family?  
E=College Updates  
F=Disruptive  
G=Preparation  
H=Connection with instructors  
I=Student Interaction  
J=Real/virtual

## Converting the data into factor

```
Covid$Gender= as.factor(Covid$Gender)
Covid$Occupation = as.factor(Covid$Education)
Covid$A = as.factor(Covid$A)
Covid$B = as.factor(Covid$B)
Covid$C = as.factor(Covid$C)
Covid$D = as.factor(Covid$D)
Covid$E = as.factor(Covid$E)
Covid$F = as.factor(Covid$F)
Covid$G = as.factor(Covid$G)
Covid$H = as.factor(Covid$H)
Covid$I = as.factor(Covid$I)
Covid$J = as.factor(Covid$J)
```

## Regression Models

We are choosing dependent variable as the connection with instructor (H) that is whether students are feeling connected with their instructor or not. Here 'H' being categorical we have to use 'Logistic Regression' to build the model. With the help of StepAIC function we can find the best model of the data keeping 'H' as the Dependent Variable.

```
model1=glm(H ~ . , data= Covid, family = binomial())
```

Here in model1 the AIC is 190.2 which can be further reduced as the model develops. With the help of this model1 we can find out all the significant factors. Another method of doing the same is StepAIC – in which R itself gives us the best model with the least AIC.

```
model2 = glm(H ~ Age + Gender + Education+ A + B + C + D + F + G + J+I, family
= binomial(), data = Covid)
```

Here in Model 2 the AIC has come down and reduced to 140.11.

## Prediction

One of the main functions of performing logistic regression in R studio is that we can predict the outcomes and validate them with the help of some tests and plots.

Here in the study we have predicted whether the students are feeling connected with their instructor or not. For which we performed certain tests and used certain functions such as 'predict' to predict the outcome.

```
Covid$pred = predict(m , type = 'Covidponse')
table(Covid$H)
head(Covid$pred)
```

```
Covid$Output = ifelse(Covid$pred>0.5,'Good','Bad')
table(Covid$Output)
```

After performing the function an additional column is created in our dataset which implies the predicted values but they are in decimals. Hence in order to convert them in 0 and 1 we use 'ifelse' function. Once this is done we can get the predicted values in the format of 0 and 1. Hence our actual values i.e Covid\$H and our predicted values i.e Covid\$pred are in 0 and 1.

## Validation

Now that we have predicted our values as to if the students have connection with their instructor or not we have to validate our outcomes and check for the accuracy using Confusion Matrix. In confusion Matrix we have used the function 'table' in which we put the Actual values and the predicted values and this is the outcome we have got –

```
library(caret)
Covid$Output = as.factor(Covid$Output)
str(Covid)
newtable<data.frame(CI=Covid$H,pr=Covid$Output)
cm = confusionMatrix(newtable$CI,newtable$pr)
cm
```

```
#Sensitivity = TP/(TP+FN)
sensi = 61/(61+6)
sensi
#Hence there is 91.044% sensitivity
```

```
#Specificity = TN/(TN+FP)
speci = 94/(94+6)
speci
#Hence there is 94% specificity
```

	0	1
No	94	3
Yes	6	61

Predicted	Actual
-----------	--------

From the above table we can analyse that 94 observations were true negative which means we predicted it as No and actually they are No and 61 observations were actually Yes and we predicted as Yes. This will lead us to Specificity and Sensitivity which is 94% and 91.044% Covidpectively. This indicates the accuracy of the model.

Another way of validation of the model is ROCR Curve and Area Under Curve which can be obtained by using 'performance' function. And the Area Under curve was 94.58%.

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

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

Since the curve will be very close to Y-Axis and the AUC is 94.58% this indicates that the model is very good and accurate.

### Summary Analysis for Model

```
AIC =140.11
True Negative = 94
True Positive = 61
Specificity = 94%
Sensitivity = 91.044%
Area Under Curve =94.58%
```

## Factor analysis

We conducted a factor analysis and categorized it under 3 heads  
The factanal() function produces maximum likelihood factor analysis.

```
#factor analysis
Covida<-Covid
Covida<-Covida[-1]
Covida<-Covida[-1]
Covida<-Covida[-1]
Covida<-Covida[-1]
Covida<-Covida[-20]
str(Covida)
View(Covida)
summary(Covida)

Covida$A=as.numeric(as.factor(Covida$A))
Covida$B=as.numeric(as.factor(Covida$B))
Covida$C=as.numeric(as.factor(Covida$C))
Covida$D=as.numeric(as.factor(Covida$D))
Covida$E=as.numeric(as.factor(Covida$E))
Covida$F=as.numeric(as.factor(Covida$F))
Covida$G=as.numeric(as.factor(Covida$G))
Covida$H=as.numeric(as.factor(Covida$H))
Covida$I=as.numeric(as.factor(Covida$I))
Covida$J=as.numeric(as.factor(Covida$J))

#Principle component
Covida_pc <- prcomp(Covida[] , center = TRUE , scale. = TRUE)
summary(Covida_pc)
plot(Covida_pc , type="l")

#Factor Analysis
Covida.fact=factanal(Covida[], 4 ,rotation = "varimax")
Covida.fact

#applying cutoff
Covida.fact = factanal(Covida[] , 4 , rotation = "varimax" , scoCovid = "regression")
print(Covida.fact , digits = 2 , cutoff=0.4 , sort=TRUE)
```

Social media	Connecting with people	Mental state
<ul style="list-style-type: none"> <li>• A=Following News</li> <li>• B=News Sources</li> <li>• C=Which is the most used way of your connection?</li> </ul>	<ul style="list-style-type: none"> <li>• D=Connection with your family?</li> <li>• H=Connection with instructors</li> <li>• I=Student Interraction</li> </ul>	<ul style="list-style-type: none"> <li>• F=Disruptive</li> <li>• G=Preparation</li> </ul>



## **Inference**

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

- To know about one's personal doing the main parameters are whether they follow yes or not, if they follow what sources do they use the most and which source they use to connect with their family or friends which helps to know about a student's activity.
- To know how connected students feel with the people around the main parameters are finding connection with their family members, their instructors and how often they interact with their friends or peers.
- The last factor includes disruption i.e., how much they have deviated from all and how prepared they are for this distance learning.

## **Conclusion**

As the COVID-19 coronavirus continues to spread, schools around the globe are shifting to online learning in an effort to slow the spread of the disease. Studies on online students indicate that building a

sense of community among students enhances student learning, retention and student satisfaction with their online experience. To minimize the challenges experienced by distance learning, e-learning should be encouraged. Infrastructure can be updated by introducing modern technology, fast Internet connection, continuous power supply, security, regular maintenance, and efficient administration of distance learning. Lecturers and students should also have skills and confidence to use electronic equipment, and to have the necessary knowledge about the method in which the information is delivered. People are finding it difficult to stay connected but then also keeping connection with the family members, instructors and friends are very important.

Our analysis was done by regression and correlation analysis, created models to understand which is the best AIC and also, we did factor analysis in order to find out best 3 factors in order to understand how the overall factors can be named under 1 factor to get analysis.

# Analysis of Risk of Heart Disease

Submitted By-  
Sanjana Kunjar  
Muthulakshmi Shunmugham  
Vignesh Krishnamoorthy

## Abstract

In a fast-moving world, many fall prey to heart conditions that makes life difficult or may even be fatal. The objective of this research is to help factor in people's daily practices to help deduce how susceptible they are to heart diseases.

To help assist with the same, after consultation with an expert, 22 dimensions of a person's daily routines that have shown prominent influence over developing heart diseases were taken into account. In accordance, a questionnaire enquiring about the aspects of the study was created for the population under study. The data so obtained was processed by various hypothesis testing using R Programming to find a degree to which a person is at a risk of heart failure.

A natural safety limit exists **23** which was determined by consulting with a doctor and various research journals, which when exceeded can put one at the risk of heart failure. An ideal case, where the person's practices falls under the permissible threshold, has been taken as our overall safety limit. The population has responded with their information, age ranging from 18 to 65 years, most of whom fall under the age group of 20-30. For every respondent, each aspect of their life is taken and evaluated relative to the safety limit and the cumulative value of all these risks is obtained.

## Methodology:

### Linear Regression-

Linear regression is a linear approach to modeling the relationship between a scalar response

(or dependent variable) and one or more explanatory variables (or independent variables).

Dependent Variable- Column 23(Number)

Independent Variables- Column 1-22

Best Model- **Mod18**- 4 factors out of 22 have been removed accounting to a fall in Adjusted R Squared. The remaining 18 factors display positive correlation with dependent variable (Number).

### **Logistic Regression-**

Logistic Regression, also known as Logit Regression or Logit Model, is a mathematical model used in statistics to estimate the probability of an event occurring having been given some previous data. Logistic Regression works with binary data.

Dependent Variable- Column 24(Number1)- Categorical with 2 factors “Yes” and “No”

Independent Variables- Column 1-22

Best Model- **m2**- generalized linear model is used in calculating StepAIC. Least value of StepAIC is considered the best fit

### **Cluster Analysis-**

The purpose of cluster analysis is to place objects into groups, or clusters, suggested by the data, not defined a priori, such that objects in a given cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar.

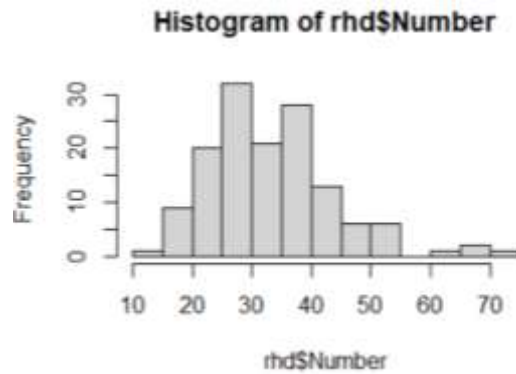
```
setwd("C:/Users/Sanjana/Desktop/ISBR/Trimester 5/R")
rhd<-read.csv("Analysis of risk of heart disease.csv")
View(rhd)
str(rhd)
```

```
# Data cleaning
# Check for NA
```

```
table(is.na(rhd))
#There are no NAs in the dataset
```

```
# Graphical representation of data
hist(rhd$Number)
```

## LIVE PROJECTS- Predictive Analysis Using R

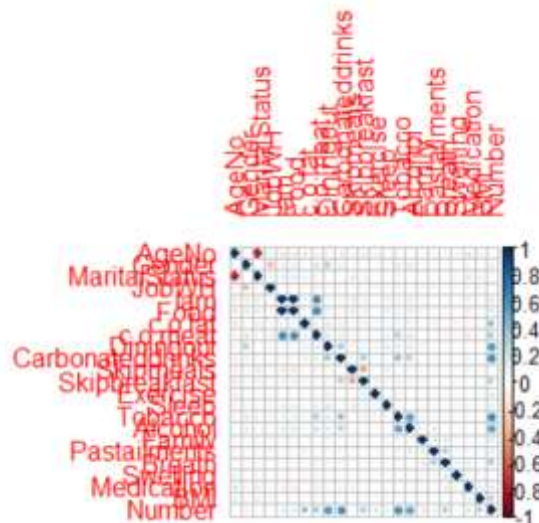


```
table(rhd$Number1)
# No Yes
# 21 119
```

# The above histogram represents the cumulative scores measured in terms of frequency and it is inferred that 119 people fall within the danger zone of contacting heart diseases.

```
plot(rhd$Number, rhd$BMI)
```

```
# Correlation between variables
library(corrplot)
corrplot(cor(rhd[,1:23]))
```



# The factors with positive correlation have been highlighted by shades of blue while those with less/negative correlation have been highlighted in shades of red.

```
cor(rhd$Number, rhd$Tobacco)
#0.6023803
```

```
cor(rhd$Number, rhd$Carbonateddrinks)
#0.5919225
```

```
cor(rhd$Number, rhd$c.o.meat)
#0.350482
```

## LIVE PROJECTS- Predictive Analysis Using R

```
cor(rhd$Number, rhd$MaritalStatus)
# -0.03694158

#####

# Linear regression
library(caret)

set.seed(1)

partition<- createDataPartition(y=rhd$Number,p=0.70 , list = FALSE)
training<- rhd[partition,]
test<- rhd[-partition,]

#Building Models
mod1 = lm(Number~ Gender, data = training)
summary(mod1)
#Adjusted R-squared: 0.01826

mod2= lm(Number~ Gender ,Food, data = training)
summary(mod2)
# Adjusted R-squared: 0.06263

mod3= lm(Number~ Gender ,Food,c.o.fat, data = training)
summary(mod3)
#Multiple R-squared: 0.2425, Adjusted R-squared: 0.2188

mod4= lm(Number~ Gender ,Food,c.o.fat,c.o.meat , data = training)
summary(mod4)
# Multiple R-squared: 0.2582, Adjusted R-squared: 0.227

mod5= lm(Number~ Gender ,Food,c.o.fat,c.o.meat,Diningout, data = training)
summary(mod5)
# Multiple R-squared: 0.4568, Adjusted R-squared: 0.4279

mod6= lm(Number~ Gender ,Food,c.o.fat,c.o.meat,Diningout,Carbonateddrinks,
data = training)
summary(mod6)
# Multiple R-squared: 0.5609, Adjusted R-squared: 0.5326

mod7= lm(Number~ Gender
,Food,c.o.fat,c.o.meat,Diningout,Carbonateddrinks,Skipmeals, data =
training)
summary(mod7)
# Multiple R-squared: 0.5724, Adjusted R-squared: 0.5398

mod8= lm(Number~ Gender
,Food,c.o.fat,c.o.meat,Diningout,Carbonateddrinks,Skipmeals,
Skipbreakfast, data = training)
summary(mod8)
#Multiple R-squared: 0.6112, Adjusted R-squared: 0.577
```

## LIVE PROJECTS- Predictive Analysis Using R

```
mod9= lm(Number~ Gender
,Food,c.o.fat,c.o.meat,Diningout,Carbonateddrinks,Skipmeals,
      Skipbreakfast,Exercise , data = training)
summary(mod9)
#Multiple R-squared:  0.6316, Adjusted R-squared:  0.5947

mod10= lm(Number~ Gender
,Food,c.o.fat,c.o.meat,Diningout,Carbonateddrinks,Skipmeals,
      Skipbreakfast,Exercise,Sleep , data = training)
summary(mod10)
#Multiple R-squared:  0.6403, Adjusted R-squared:  0.5999

mod11= lm(Number~ Gender
,Food,c.o.fat,c.o.meat,Diningout,Carbonateddrinks,Skipmeals,
      Skipbreakfast,Exercise,Sleep,Tobacco , data = training)
summary(mod11)
#Multiple R-squared:  0.7207, Adjusted R-squared:  0.6967

mod12= lm(Number~ Gender
,Food,c.o.fat,c.o.meat,Diningout,Carbonateddrinks,Skipmeals,
      Skipbreakfast,Exercise,Sleep,Alcohol,Tobacco , data =
training)
summary(mod12)
#Multiple R-squared:  0.7493, Adjusted R-squared:  0.7256

mod13= lm(Number~ Gender ,Food,c.o.fat,c.o.meat,Family
,Diningout,Carbonateddrinks,
      Skipmeals,Skipbreakfast,Exercise,Sleep,Alcohol,Tobacco , data
= training)
summary(mod13)
#Multiple R-squared:  0.7905, Adjusted R-squared:  0.7689

mod14= lm(Number~ Gender ,Food,c.o.fat,c.o.meat,Family
,Pastailments,Diningout,
Carbonateddrinks,Skipmeals,Skipbreakfast,Exercise,Sleep,Alcohol,Tobacco,
      data = training)
summary(mod14)
#Multiple R-squared:  0.8053, Adjusted R-squared:  0.7835

mod15= lm(Number~ Gender ,Food,c.o.fat,c.o.meat,Family ,Breath
,Pastailments,Diningout,
Carbonateddrinks,Skipmeals,Skipbreakfast,Exercise,Sleep,Alcohol
      ,Tobacco , data = training)
summary(mod15)
#Multiple R-squared:  0.8448, Adjusted R-squared:  0.826

mod16= lm(Number~ Gender ,Food,c.o.fat,c.o.meat,Family ,Swelling,Breath
,Pastailments,
Diningout,Carbonateddrinks,Skipmeals,Skipbreakfast,Exercise,Sleep,
      Alcohol,Tobacco , data = training)
summary(mod16)
```

## LIVE PROJECTS- Predictive Analysis Using R

```
#Multiple R-squared:  0.8498, Adjusted R-squared:  0.8303

mod17= lm(Number~ Gender ,Food,c.o.fat,c.o.meat,Medication,Family
,Swelling,Breath,
Pastailments,Diningout,Carbonateddrinks,Skipmeals,Skipbreakfast,Exercise,
Sleep,Alcohol,Tobacco , data = training)
summary(mod17)
#Multiple R-squared:  0.8829, Adjusted R-squared:  0.8661

mod18= lm(Number~ Gender ,Food,c.o.fat,BMI,c.o.meat,Medication,Family
,Swelling,Breath,Pastailments,Diningout,Carbonateddrinks,Skipmeals,Skipbre
akfast,Exercise,Sleep,Alcohol,Tobacco , data = training)
summary(mod18)

#Multiple R-squared:  0.9568, Adjusted R-squared:  0.9472
#Model 18 is the best Model as Adjusted R-squared is the highest.

#Normality test
install.packages("car")
library(car)

outlierTest(mod18)
# No residual for  $P < 0.05$  hence data is normal and there are no outliers

shapiro.test(residuals(object=mod18))
# p-value = 0.05112
# Since P-Value is  $> 0.05$ , data is normally distributed

durbinWatsonTest(mod18)
# P value  $> 0.05$ 
# Since P-value  $> 0.05$ , auto correlation does not exist

# Multi colinearity test
table(sqrt(vif(mod18))>2)
# All table values are false. Therefore the independent variable are not
highly correlated with each other.

#####

training$pred<-predict(mod18)
training$res<-residuals(mod18)

#Validation
test$pred<-predict(mod18,newdata = test)
test$res<-test$Number-test$pred
```

## LIVE PROJECTS- Predictive Analysis Using R

```
View(training)
View(test)

rhd.pred<- predict(mod18, newdata= test, type='response')
summary(rhd.pred)
head(rhd.pred)
View(rhd.pred)

converted<-ifelse(rhd.pred<23, "No", "Yes")
head(converted)
table(converted)

# The threshold number is determined to be 23. Therefore, people with
# calculated number greater than 23 are prone to a risk of heart disease.

confm<-data.frame(predicted=converted, actual= test$Number1)
confm$predicted=as.factor(confm$predicted)
confm$actual=as.factor(confm$actual)
str(confm)

library(caret)
View(confm)
resi<-confusionMatrix(confm$predicted,confm$actual)
resi

#           Reference
# Prediction No Yes
#           No    6  0
#           Yes   0 34

# Accuracy : 1

#####

# Logistic Regression Analysis

rhd$Number1<-as.factor(rhd$Number1)
str(rhd)
library(caret)

set.seed(100)
partition1<- createDataPartition(y=rhd$Number1,p=0.70 , list = FALSE)
training1<- rhd[partition1,]
test1<- rhd[-partition1,]

m1<-glm(Number1~., data=training1,family=binomial())
library(car)
library(MASS)
stepAIC(m1)

m2<-glm(formula = Number1 ~ JobWH , c.o.meat , Diningout ,
Carbonateddrinks , Skipmeals , Family , Breath , Medication , BMI, family
= binomial(), data = training1)
```



## LIVE PROJECTS- Predictive Analysis Using R

```
summary(m2)
```

```
# AIC = 20(least) m2 is the best model
```

```
pred1<- predict(m2, newdata= test1, type='response')
```

```
summary(pred1)
```

```
View(test1)
```

```
head(pred1)
```

```
View(pred1)
```

```
library(ROCR)
```

```
predictions<-prediction(pred1,test1$Number1)
```

```
roc.pred1=performance(predictions,measure='tpr',x.measure='fpr')
```

```
plot(roc.pred1)
```



```
dist<-rep(9999, length(roc.pred1@x.values[[1]]))
```

```
for(i in 1: length(roc.pred1@x.values[[1]])){
```

```
  cur_x<- roc.pred1@x.values[[1]][i]
```

```
  cur_y<- roc.pred1@y.values[[1]][i]
```

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

```
}
```

```
ideal_cutoff<- roc.pred1@alpha.values[[1]][dist==min(dist)]
```

```
ideal_cutoff
```

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

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

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

```
par(new=TRUE)
```

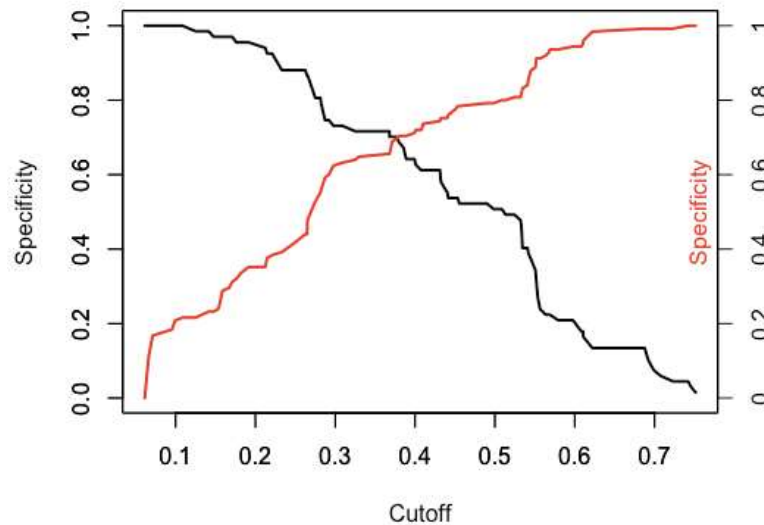
```
plot(unlist(performance(predictions, "spec")@x.values),
```

```
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')
```



**# Intersection- 0.7**

```
convert<-ifelse(pred1<0.7, "No","Yes")
head(convert)
table(convert)
```

```
cm<-data.frame(predicted=convert, actual= test1$Number1)
head(cm)
cm$predicted=as.factor(cm$predicted)
str(cm)
View(cm)
res<-confusionMatrix(cm$predicted,cm$actual)
res
```

```
#           Reference
# Prediction No Yes
#           No  4  2
#           Yes  2 33
```

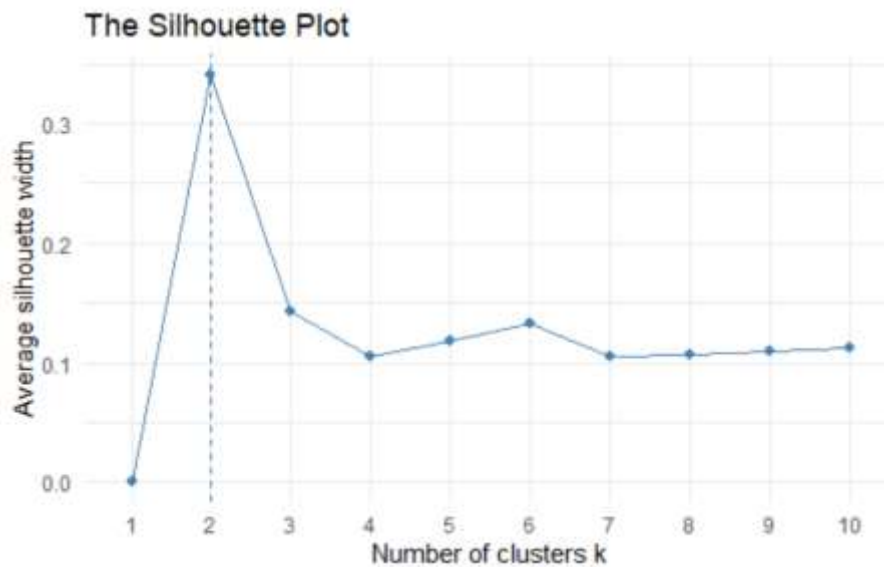
**# Accuracy : 0.9024**

#####

**# Cluster analysis**

```
new.rhd<-rhd[,-24]
View(new.rhd)
rhdscale<- scale(new.rhd)
```

```
install.packages("factoextra")
library(factoextra)
fviz_nbclust(rhdscale, kmeans, method = "silhouette", k.max = 10) ,
theme_minimal() , ggtitle("The Silhouette Plot")
```



# From the graph, it is inferred that the number of clusters = 2

```
krhd<- kmeans(new.rhd,2,nstart =5)
```

```
krhd
```

```
krhd$size
```

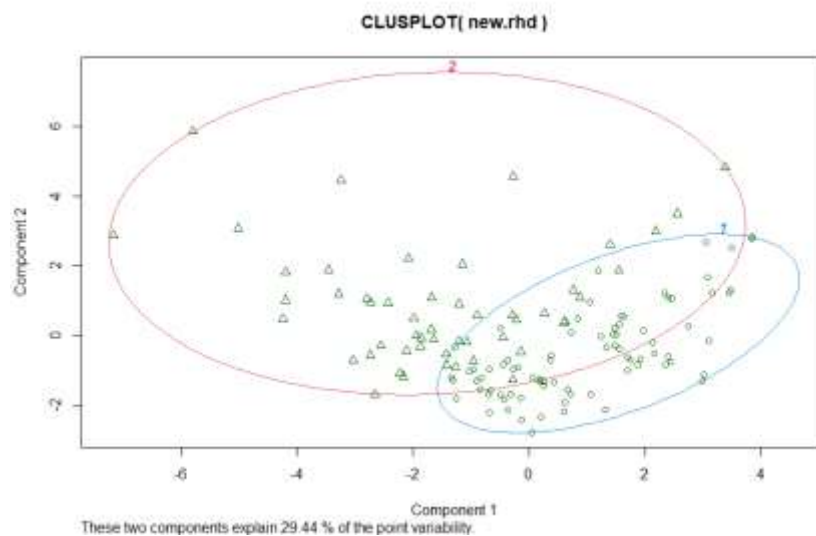
```
# 83, 57
```

# This means that the 83 people from the data set fall under cluster 1 and 57 under cluster 2 based on their everyday habits

# creating cluster plots

```
library(cluster)
```

```
clusplot(new.rhd, krhd$cluster, color = T, labels = 4)
```



```
rhd.cluster<-table(rhd$Number1, krhd$cluster)
```

```
rhd.cluster
```

	1	2
No	21	0
Yes	62	57

## Conclusion

With changes in our routine practices, it has been found that more and more people are now susceptible to heart diseases. From general factors taken in the study, it was found that the following factors have high impact on risk of heart diseases.

- Gender
- Food(vegetarian/non-vegetarian)
- Consumption of fat
- BMI
- Consumption of meat
- Medication
- Family history of heart diseases
- Swelling (legs, hands and abdomen)
- Shortness of breath
- Past ailments
- Dining out
- Consumption of carbonated drinks
- Skipping meals
- Skipping breakfast
- Hours of exercise
- Hours of sleep
- Consumption of alcohol and tobacco

This analysis exhibits the aspects of our lives that may be damaging our hearts. For example, if your busy schedule does not allow for a sound 6-8 hours of sleep, regular meals or exercise, it could, in the long term endanger your heart.

The study enables us to set limits and priorities to ensure a healthy heart.

# Employee Satisfaction in Hospitality Industry

Submitted By-  
Shelaj Sharma  
Ritom Das  
Meghana Kalapala  
Harmanjeet Kaur

## Abstract

### Purpose:

- To understand how employees in hospitality industry are effected by various factors.
- To analyze and build a model to understand what improvements can be made in order to attract more talent and youth to take up opportunities in hospitality industry.
- To understand the personnel management needs of the hospitality industry.

### 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 job appreciation, pay package, use of paid leaves, working schedule, work life balance and medical benefits. Related research papers and literature review has been referred to understand the requirements of the study.

### Findings:

The logistic regression model findings include the work role of an employee and the factors that affect their job.

We have taken work role of an employee as our dependent variable and found the how the

employee in a particular work role: *manager, chef, executive and student* have appreciated The presence of other variables which leads to their job satisfaction.

The logistic regression also finds out the best model that compares a combination of dependent and independent variables that can be implemented to support our purpose.

### **Practical Implications:**

To execute the best model which helps in increasing the employee satisfaction and in understanding the employment needs and requirements in the hospitality industry.

### **Introduction:**

Hospitality industry can be defined and understood as an industry which provides facility for stay, food and complete related services for the comfort of the travelers and visitors. The hospitality industry refers to a variety of businesses and services linked to leisure and customer satisfaction. We decided to do a study on the employees in the hospitality industry to understand how different factors affect the employee satisfaction in hospitality sector. Employee satisfaction is of utmost importance for employees to remain happy and also deliver their level best. Satisfied employees are the ones who are extremely loyal towards their organization and stick to it even in the worst scenario. We have taken different factors which affects employee satisfaction and have analyzed this data to find the relationship between the level of satisfaction and their work role. The relationship between the different factors are been found by correlation and regression. The best model is built taking work role as dependent variable. Various tests are done to confirm that the model built is the best model.

### **Research Objective:**

- To understand and identify various factors that determine the employee satisfaction in hospitality industry.
- To build the best model to determine the factors affecting employee satisfaction and what improvements can be made in order to attract more talent to take up opportunities in hospitality industry.

### **Research Methodology:**

In this research we have conducted a google form online survey among household data of the time 133 employees and students from various hotels in different segments of food and beverages, travel and tourism, lodging and recreation.

### **Predictive Analysis:**

**Step 1:** The data is read as HOSP and subsequent removal of columns are done based on the type of study and the same has been named HOSP and summary was taken to analyze mean, median and standard deviation for each variable and the entire data.

```
HOSP=read.csv("F:/Data/My Documents/R studio/HOSPP.csv")str(HOSP)
```

View(HOSP) summary(HOSP)

**Step 2:** There is no need to clean the data as there is no missing values.  
table(complete.cases(scaled\_hosp))

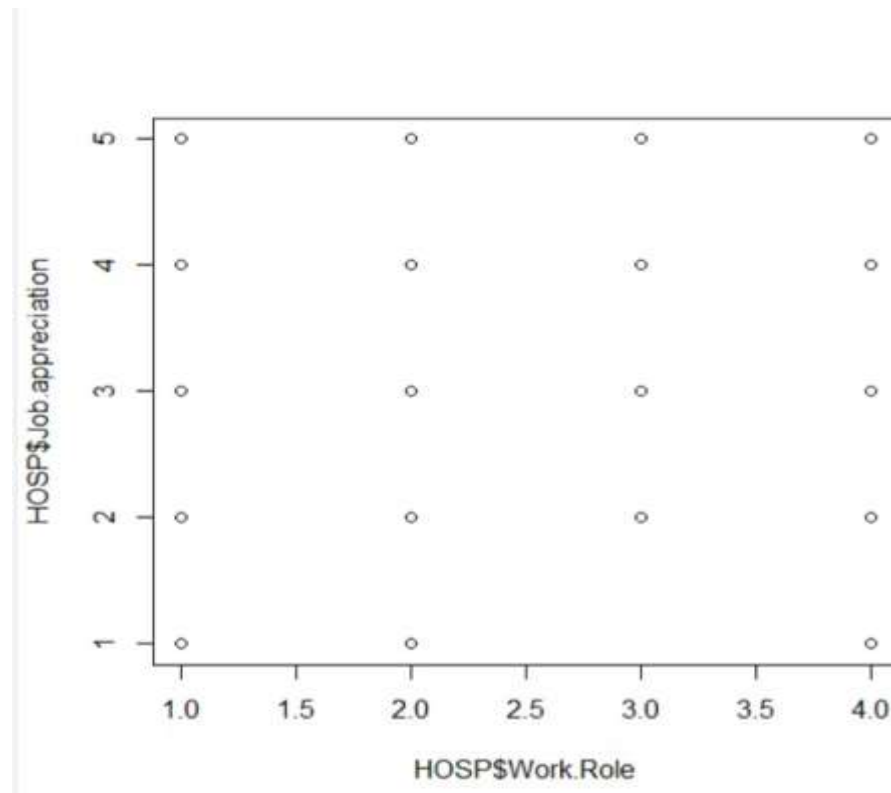
*TRUE*

*133*

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

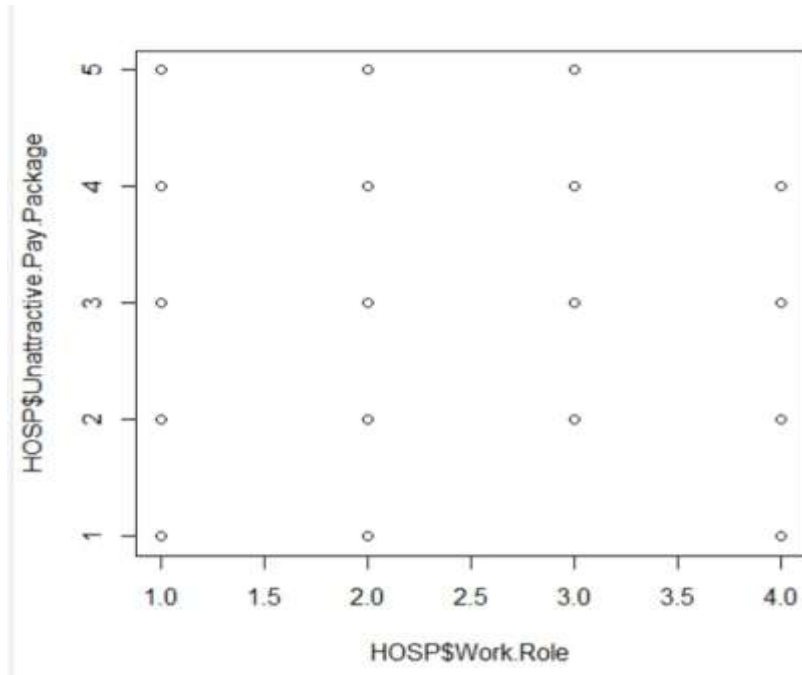
**Plots:**

plot(HOSP\$Work.Role,HOSP\$Job.appreciation)

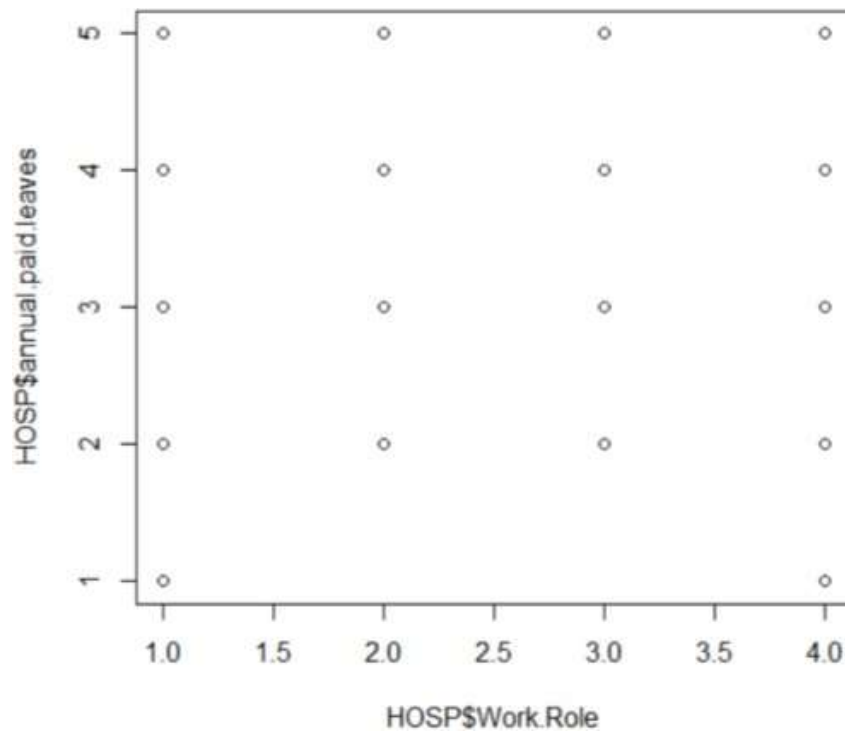


# LIVE PROJECTS- Predictive Analysis Using R

```
plot(HOSP$Work.Role,HOSP$Unattractive.Pay.Package)
```



```
plot(HOSP$Work.Role,HOSP$annual.paid.leaves)
```





**Correlation:** Here we found out how our dependent variable i.e. is correlated to the independent variables that make our best model.

```
cor(HOSP$Work.Role,HOSP$Job.appreciation)
-0.006250428
cor(HOSP$Work.Role,HOSP$annual.paid.leaves)
-0.0125706
cor(HOSP$Work.Role,HOSP$Unattractive.Pay.Package)
-0.06459428
```

All the three independent variables; job appreciation, annual paid leaves and unattractive pay package are negatively correlated to our dependent variable; work role. These independent variables have a direct effect on our dependent variable,

**Non-Linear Assumption:**

**#Ho:** There is no direct relationship between dependent and independent variable

**#H1:** There is a relationship between dependent and independent variable

```
cor.test(HOSP$Work.Role,HOSP$annual.paid.leaves)
Pearson's product-moment correlation
data: HOSP$Work.Role and HOSP$annual.paid.leaves
t = -0.14389, df = 131, p-value = 0.8858
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
```

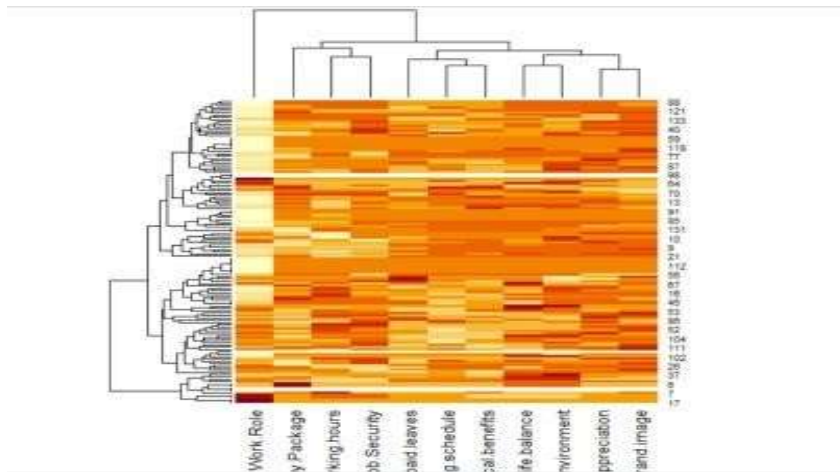
```
cor.test(HOSP$Work.Role,HOSP$annual.paid.leaves,method = "kendall")
Kendall's rank correlation tau
data: HOSP$Work.Role and HOSP$annual.paid.leaves
z = -0.0064504, p-value = 0.9949
alternative hypothesis: true tau is not equal to 0
sample estimates:
```

```
tau
-0.0004751851
```

```
x<-as.matrix(HOSP)
```

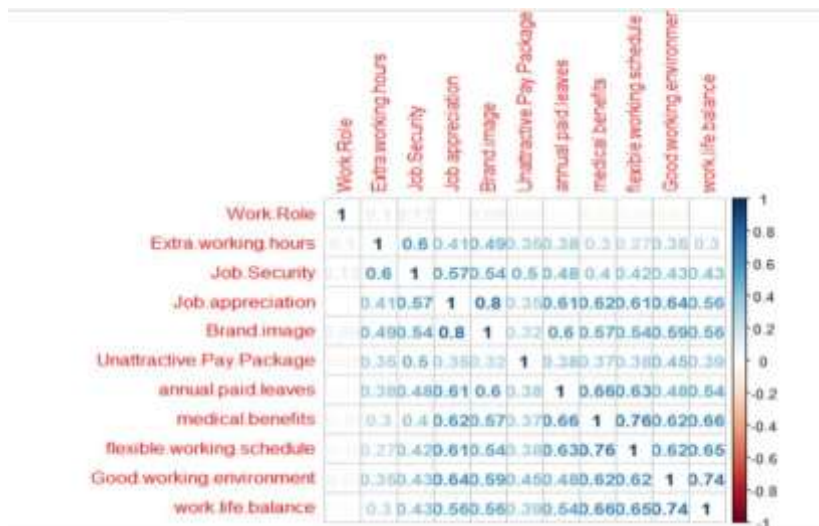
We have run a heatmap to understand the results of our predictions

```
heatmap(x)
```



In our correlation plot, all the possible combinations of dependent variables were tried and the best model with Adjusted R-square was found.

```
library(corrplot)
corrplot(cor(HOSP), method = "number")
```



**Models:**

We have taken multinomial regression for our models as our dependent variable has four factors. In our first model, we have analyzed our dependent variable (*work role*) with all our independent variables (*extra working hours, job security, job appreciation, brand image, unattractive pay package, annual paid leaves, medical benefits, flexible working hours, working environment and work life balance*)

We have taken 7 models to predict our dependent variable with combinations of independent variables and have chosen the best model as the 5<sup>th</sup> one as it has the lowest AIC value of 357.46 compared to all other models.

We have predicted all our models to understand how the factors of our dependent variable were affected in each model.

```
#model1=multinom(Work.Role ~ . ,data = HOSP,family = binomial())
```

## LIVE PROJECTS- Predictive Analysis Using R

```
model1=multinom(Work.Role ~ . ,data = HOSP,family =
binomial())summary(model1)      #AIC=375.96
library(MASS)
stepAIC(model1)
pred1 = predict(model1 , type = "class")
table(pred1)
head(pred1)
View(pred1)

#model2
model2=multinom(HOSP$Work.Role~HOSP$Extra.working.hours      +
HOSP$Job.appreciation , data = HOSP, family=binomial())
summary(model2)  #AIC = 361.72
pred2=predict(model2 , newdata = HOSP , type = "class")
summary(pred2)

#model3
model3=multinom(HOSP$Work.Role~HOSP$Job.Security      +
HOSP$Unattractive.Pay.Package + HOSP$Good.working.environment , data = HOSP,
family=binomial())
summary(model3)  #AIC = 362.17
pred3=predict(model3 , newdata = HOSP , type = "class")
summary(pred3)

#model4
model4=multinom(HOSP$Work.Role~HOSP$Extra.working.hours      +
HOSP$Job.appreciation + HOSP$work.life.balance , data = HOSP, family=binomial())
summary(model4)  #AIC = 365.84
pred4=predict(model4 , newdata = HOSP , type = "class")
summary(pred4)

#model5
model5=multinom(HOSP$Work.Role~HOSP$Job.appreciation      +
HOSP$Unattractive.Pay.Package + HOSP$annual.paid.leaves , data = HOSP,
family=binomial())
summary(model5)  #AIC = 357.46
pred5=predict(model5 , newdata = HOSP , type = "class")
summary(pred5)

#model6
model6=multinom(HOSP$Work.Role~HOSP$Job.Security + HOSP$annual.paid.leaves +
HOSP$medical.benefits , data = HOSP, family=binomial())
summary(model6)  #AIC = 359.44
pred6=predict(model6 , newdata = HOSP , type = "class")
summary(pred6)

#model7
model7=multinom(HOSP$Work.Role~HOSP$Good.working.environment      +
HOSP$annual.paid.leaves + HOSP$Extra.working.hours, data = HOSP, family=binomial())
summary(model7)  #AIC = 358.04
```

## LIVE PROJECTS- Predictive Analysis Using R

```
pred7=predict(model7 , newdata = HOSP , type = "class")
summary(pred7)
```

```
#model5 has the lowest AIC
```

### Converting into factors:

We have then converted our data into factors with each variable.

```
HOSP$Work.Role<-as.factor(HOSP$Work.Role)
HOSP$Job.appreciation<-as.factor(HOSP$Job.appreciation)
str(HOSP)
fit<-
glm(Work.Role~Job.appreciation+Extra.working.hours+Job.Security+Brand.image+Unattrac
tive.Pay.Package+annual.paid.leaves+medical.benefits+flexible.working.schedule+Good.wor
king.environment+work.life.balance,data = HOSP,family = binomial())
summary(fit)
HOSP$pred<-predict(fit,type="response")
table(HOSP$Job.appreciation)
View(HOSP)
head(HOSP$pred)
HOSP$satis<-ifelse(HOSP$pred>0.1,1,0)
table(HOSP$satis)
```

### Somer's D:

**#Ho:** The data is normally distributed.

**#H1:** The data is not normally distributed.

```
library(InformationValue)
somersD(HOSP$Job.appreciation,HOSP$pred)
#if somerd value>.6, then it is good. The data is normal.
```

### Confusion Matrix:

```
library(caret)
newdata<-data.frame(ac=HOSP$Job.appreciation,sat=HOSP$satis)
confusionMatrix(as.factor(HOSP$Work.Role),as.factor(HOSP$Job.appreciation))
```

### Shapiro – Wilk Test:

**#Ho:** The data is normally distributed.

**#H1:** The data is not normally distributed.

```
shapiro.test(residuals(object=fit))
```

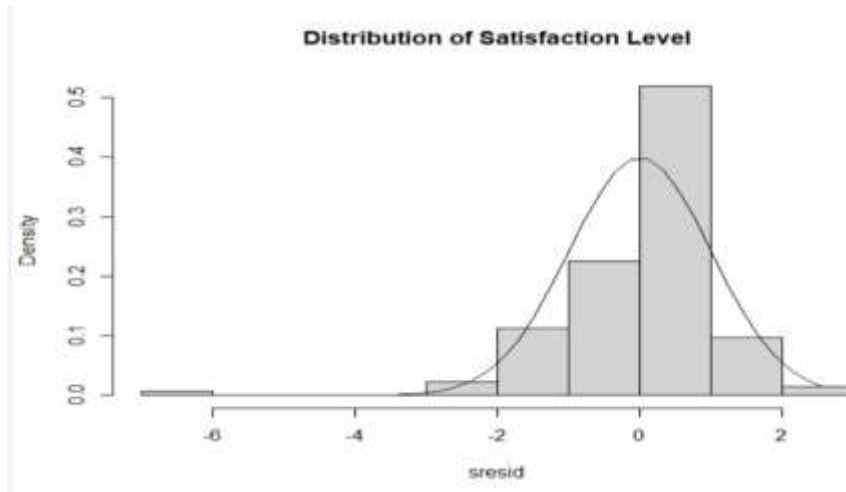
#Pvalue > 0.05. Thus Ho is accepted. The Data is normal.

**Histogram:** We have built a histogram showing the distribution of satisfaction levels of employees.

```
library(MASS)
sresid <- studres(fit)
```

```
hist(sresid, freq=FALSE,
```

```
  main="Distribution of Satisfaction Level"
```



```
xfit<-seq(min(sresid),max(sresid),length=40)
```

```
yfit<-dnorm(xfit)
```

```
lines(xfit, yfit)
```

### **Durbin – Watson Test:**

**#Ho:** linear regression residuals of time series data are uncorrelated

**#H1:** Auto Correlation exists

```
durbinWatsonTest(fit)
```

#Pvalue >0.05. Hence autocorrelation does not exist.

### **Multi Collinearity:**

The independent variables aren't highly correlated with each other. Variance inflation factor should be less than 10 or if  $\sqrt{\text{variable inflation factor}}$  should be less than 2

```
vif(fit)
```

```
sqrt(vif(fit))>2
```

The independent variables aren't highly correlated with each other. Thus the multi collinearity assumption is met.

### **Hosmer Lemeshow Test:**

**#Ho:** the observed and expected proportions are the same across all doses.

**#H1:** the observed and expected proportions are not the same.

```
install.packages("ResourceSelection")
```

```
library(ResourceSelection)
```

```
hoslem.test(HOSP$Job.appreciation,HOSP$pred)
```

## LIVE PROJECTS- Predictive Analysis Using R

#Ho: actual & predicted values are very close and model is good.

The employee satisfaction in hospitality industry is an important issue. Knowing the results of various model above we can improve the working environment in hospitality industry.

Most related variables [Job Appreciation, Attractive pay-packages, Annual paid leaves]

There are obvious significant of Job appreciation and perfect perks on employee satisfaction. If supervisors are appreciating and providing the best pays and perks for the work done, the employees seems to be more satisfied.

### **Conclusion:**

Employee satisfaction is a key factor to analyze for the present hospitality industry because it has currently become a necessity for today's world as it ultimately leads to customer satisfaction. Employee satisfaction will directly affect the quality of services that customer will be provided with. Our study has helped to us to understand the different variables involved to improve the satisfaction levels of employees in hospitality industry. Based on our research when employees are appreciated and rewarded with perks for the work they have done, they feel satisfied. These employees feel more productive and thus help in providing the best service. This establishes an opportunity for all the hotels in the segments of food and beverages, travel and tourism, lodging and recreation to improve their work environment where employees feel valued in providing the best service.

# Analysis of Viewing Movies and Series

Submitted By-  
Siddharth T (PG19124)  
Surendra Prasath S (PG19133)  
Komathisha K R (PG19163)

## **Abstract:**

The amount of time spent in watching movies and series has undergone a change during this pandemic situation. In this live project, we would like to predict what are the demographic factors that will determine whether the people will be engaged to spend the same amount of time in watching them post covid 19; also we would like to find out if those demographic factors had an impact in the time spent for watching them during the lockdown period. A regression test and a correlation test was done to figure out the answers we needed.

## **Introduction:**

The emergence of the covid 19 virus has made such an impact in our lives. It has created a paradigm shift in all of our daily activities. Lockdown has been implemented in India since the end of March and is being extended in some way till date. People had a choice in entertainment through both outdoor and indoor activities but the lockdown disabled the option of outdoors. This made people to shift towards indoor activities for entertainment. Movies and series have always been an important priority for people to spend their leisure time. With the OTT platforms on the rise, it has become quite easy for people to watch them. With being indoors and able to spend some time and money for relaxation, people have turned into this habit. In this project, we would like to predict if demographic factors like gender, age, occupation will determine whether the people will be engaged to spend the same amount of time in watching them post covid 19; also we would like to find out if those demographic factors had an impact in the time spent for watching them during the lockdown period. Primary Data was collected across people of different cities, age group, occupation and gender. The data was cleansed for removing NA values and unnecessary information was removed. Various regression models were developed to find the best one.

## **Coding:**

```
getwd()

setwd("F:/ISBR/T-4/rscript")

primary<-read.table("Questionnaire Responses .csv",header = T,fill = TRUE)

data<-read.csv("Questionnaire Responses .csv")

View(data)

primary<-subset(data,select = -
c(LK, City,LW,Favorite.Movie.Series.Genre,In.which.languages.would.you.like.to.watch.mo
vies.series.,Languages.Known))

View(primary)

#Data Cleaning

is.na(primary)

nprimary<-na.omit(primary)

View(nprimary)

complete.cases(nprimary)

table(complete.cases(nprimary))

gendersplit<-split(nprimary, nprimary$Gender)

head(gendersplit$Male)

library("VIM")

library(glmnet)

library(MASS)

library(nnet)

library(e1071)

library(devtools)

m1<-
multinom(Post.Lockdown..Will.you.continue.watching.movies.series.the.same.amount.of.tim
e.like.you.watch.during.lockdown~Gender ,data=nprimary)

summary(m1)
```



## LIVE PROJECTS- Predictive Analysis Using R

```
m2<-
multinom(Post.Lockdown..Will.you.continue.watching.movies.series.the.same.amount.of.time.like.you.watch.during.lockdown~Age ,data=nprimary)

summary(m2)

m3<-
multinom(Post.Lockdown..Will.you.continue.watching.movies.series.the.same.amount.of.time.like.you.watch.during.lockdown~Occupation ,data=nprimary)

summary(m3)

m4<-
multinom(Post.Lockdown..Will.you.continue.watching.movies.series.the.same.amount.of.time.like.you.watch.during.lockdown~Gender+Age ,data=nprimary)

summary(m4)

m5<-
multinom(Post.Lockdown..Will.you.continue.watching.movies.series.the.same.amount.of.time.like.you.watch.during.lockdown~Gender+Occupation ,data=nprimary)

summary(m5)

m6<-
multinom(Post.Lockdown..Will.you.continue.watching.movies.series.the.same.amount.of.time.like.you.watch.during.lockdown~Age+Occupation ,data=nprimary)

summary(m6)

m7<-
multinom(Post.Lockdown..Will.you.continue.watching.movies.series.the.same.amount.of.time.like.you.watch.during.lockdown~Gender+Age+
Occupation ,data=nprimary)

summary(m7)

cor.test(nprimary$ts,nprimary$Gender1)
cor.test(nprimary$ts,nprimary$Age1)
cor.test(nprimary$ts,nprimary$Occupation1)
hist(nprimary$ts[nprimary$male==1],main= "timespend",xlab = "male",ylab = "time spend")
hist(nprimary$ts[nprimary$female==1],main= "timespend",xlab = "female",ylab = "time spend")
par(mfcol=c(1,2))
```

**Analysis:**

As it can be seen, a total of seven models were developed in order to find out the best one. Of these, the model number 5 seems to have the lowest AIC value of all. Hence the model was accepted.

```
call:
multinom(formula = Post.Lockdown..will.you.continue.watching.movies.series.
e.amount.of.time.like.you.watch.during.lockdown ~
  Gender + occupation, data = nprimary)

Coefficients:
  (Intercept) GenderFemale GenderMale OccupationHousewife
No    11.431131  -11.249121 -11.762774          -18.00067
Yes   -4.153199   3.134751  4.346815          -16.95986
  occupationSelf Employed OccupationStudent
No           -4.717197      0.47280794
Yes          18.099519      -0.02714088

Std. Errors:
  (Intercept) GenderFemale GenderMale OccupationHousewife
No    301.128151  301.128458 301.128621      4.016930e-06
Yes     4.739713   4.746458  4.743419      3.568315e-06
  occupationSelf Employed OccupationStudent
No           1.641280e-08      0.5558711
Yes          2.062881e-06      0.6603602

Residual Deviance: 143.9733
AIC: 167.9733
> |
```

From this, it was concluded that Gender and Occupation are the demographic factors that will determine whether the people will be engaged to spend the same amount of time in watching them post covid 19.

The correlation test was performed to know the relation between gender, age, occupation to time spent in watching movies and series during lockdown.

```
> cor.test(nprimary$ts,nprimary$Gender1)
```

```
      Pearson's product-moment correlation
```

```
data: nprimary$ts and nprimary$Gender1
t = 0.40297, df = 77, p-value = 0.6881
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1770323  0.2643033
sample estimates:
      cor
0.0458738
```

The correlation test conducted between gender and time spent in watching movies and series shows that there is no correlation between them.

```
> cor.test(nprimary$ts,nprimary$Age1)
```

```
      Pearson's product-moment correlation
```

```
data: nprimary$ts and nprimary$Age1
t = -1.1967, df = 77, p-value = 0.2351
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.34590132  0.08863289
sample estimates:
      cor
-0.1351257
```

The correlation test conducted between age and time spent in watching movies and series shows that there is no correlation between them.

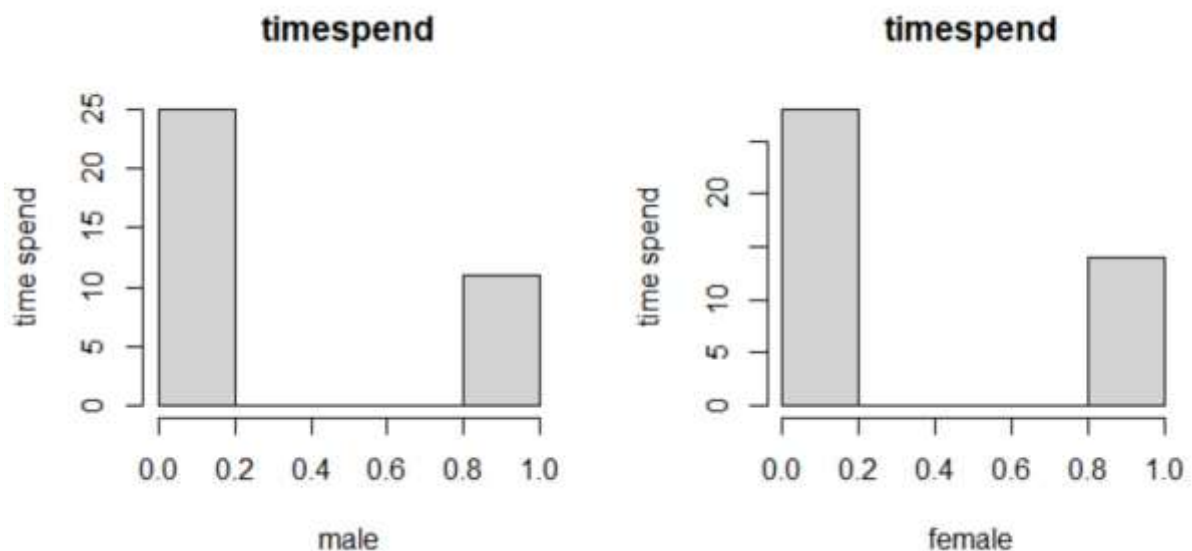
```
> cor.test(nprimary$ts,nprimary$Occupation1)

Pearson's product-moment correlation

data:  nprimary$ts and nprimary$Occupation1
t = 0.83381, df = 77, p-value = 0.407
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1292181  0.3092374
sample estimates:
      cor
0.09459501
```

The correlation test conducted between occupation and time spent in watching movies and series shows that there is no correlation between them.

A histogram was displayed to know on an average, how much time do men and women spend in watching movies and series.



We have concluded that more men and women spend about 1-3 hours on an average for watching movies and series.

**Conclusion:**

Through this report, we were able to identify the significant demographic factors that will determine whether the people will be engaged to spend the same amount of time in watching them post covid 19. We were also able to determine an average time men and women spend in watching movies and series.

# The Impact of Employee Engagement in an Organization

Submitted By-  
Sunny Singh (PG19153)  
Taha Aktar (PG19136)  
Sagar Gadhave (PG19107)

## **Abstract:**

The purpose of this paper is to identify the impact of employee engagement on organization. It also studies the impact of employee engagement on employee performance. It is the level of commitment and involvement an employee has towards their organization and its values and beliefs. An engaged employee is aware of business context, and works with colleagues to improve performance within the job for the benefit of the organization. It projects the impact of employee engagement on organization's productivity and presents the factors influencing the employee engagement and organizational outcomes

This paper is a presentation of findings of a research carried out to explore the impact of employee engagement on organization. The research methodology employed an explanatory-descriptive survey design. We have focused on the qualitative analysis method for collecting data in form of responses through online survey form. An online survey was conducted from 1 to 30 June 2020 to collect the information. A structural questionnaire link using 'Google form' was sent to students' through WhatsApp and E-mail. A total of 105 students provided complete information regarding the survey. This questionnaire design for study the factors of employee engagement and performance. Tracking employee engagement is important in determining whether or not your employees are happy and how long they'll stay with your company.

Employee engagement was measured using Hypothesis testing with the help of R language on R studio and Questionnaire. Results of the study revealed that only 33.8% of the employees were actively engaged while the remaining 66.2 % showed low levels of

engagement. It is highly important to know the level of Engagement in an Organization and act upon the results for growth of the Organization

### **Introduction:**

Mone and London (2010) defined employee engagement is “a condition of employee who feels involved, committed, passionate, and empowered and demonstrates those feelings in work behavior”. It is thus the level of commitment and involvement an employee has towards their organization and its values. The organization must work to develop and nurture engagement, which requires a two-way relationship between employer and employee. Thus, employee engagement is a barometer that determines the association of a person with the organization.

Rothbard (2001) defines engagement as psychological presence and, furthermore, states that it involves two critical components: attention and absorption. Attention refers to cognitive availability and the amount of time one spends thinking about a role, while absorption means being engrossed in a role and refers to the intensity of one's focus on a role.

Saks (2006) argues that one way for individuals to repay their organisation is through their level of engagement. In other words, employees will choose to engage themselves to varying degrees and in response to the resources they receive from their organisation. Bringing oneself more fully into one's work roles and devoting greater amounts of cognitive, emotional, and physical resources is a very profound way for individuals to respond to an organization's actions.

### **Objective:**

1. To study the impact of Employee Engagement on organization
2. To Synthesize the outcomes associated with employee engagement

### **Methodology:**

Quantitative analysis has been done by conducting various hypothesis testing. Data collection has been done by making use of questionnaire survey from 106 employees with the help of google form survey.

Buckingham and Coffman (2005) said, pay and benefits are equally important to every employee, good or bad. A company's pay should at least be comparable to the market average. However, bringing pay and benefits package up to market levels, which is a sensible first step, will not take a company very far- they are like tickets to the ballpark, they can get the company into the game, but can't help it win.

Saks (2006) argues that one way for individuals to repay their organisation is through their level of engagement. In other words, employees will choose to engage themselves to varying degrees and in response to the resources they receive from their organisation. Bringing oneself more fully into one's work roles and devoting greater amounts of cognitive, emotional, and physical resources is a very profound way for individuals to respond to an organization's actions.

## LIVE PROJECTS- Predictive Analysis Using R

According to Holbeche and Springett (2003), people's perceptions of 'meaning' with regard to the workplace are clearly linked to their levels of engagement and, ultimately, their performance.

According to Maslach et al. (2001), six areas of work-life lead to either burnout or engagement: workload, control, rewards and recognition, community and social support, perceived fairness and values. They argue that job engagement is associated with a sustainable workload, feelings of choice and control, appropriate recognition and reward, a supportive work community, fairness and justice, and meaningful and valued work

Robinson et al. (2004) define employee engagement as "a positive attitude held by the employee towards the organization and its value. An engaged employee is aware of business context, and works with colleagues to improve performance within the job for the benefit of the organization. The organization must work to develop and nurture engagement, which requires a two-way relationship between employer and employee.

### **Analysis:**

Quantitative analysis has been done by conducting various hypothesis testing. Data collection has been done by making use of questionnaire survey from 106 employees with the help of google form survey.

```
getwd()
setwd("C:/Program Files/R_files")
EE<-read.csv("EmployeeEngagement_Sgr.csv")
str(EE)
summary(EE)
View(EE)

#Predictive Analytics
#In this project Employee which emphasizes EmployeeEngagement services provided by an
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.
EE$EEdep1<-ifelse(EE$EEdep=="Yes",1,0)
EE$Effort1<-ifelse(EE$Effort=="Strongly Agree" | EE$Effort == "Agree",1,0)
EE$feedback1<-ifelse(EE$feedback=="Yes",1,0)
EE$success1<-ifelse(EE$success=="Yes",1,0)
EE$Time1<-ifelse(EE$Time=="Yes",1,0)
EE$Employee1<-ifelse(EE$Employee=="Strongly Agree" | EE$Employee == "Agree",1,0)
EE$encourage1<-ifelse(EE$encourage=="Strongly Agree" | EE$encourage == "Agree",1,0)
EE$Important1 <-ifelse(EE$Important == "Strongly Agree" | EE$Important == "Agree",1,0)
EE$Communicating1<-ifelse(EE$Communicating=="Strongly Agree" | EE$Communicating
```



## LIVE PROJECTS- Predictive Analysis Using R

```
== "Agree",1,0)
EE$Leadership1<-ifelse(EE$Leadership=="Strongly Agree" | EE$Leadership ==
"Agree",1,0)
EE$Policies1<-ifelse(EE$Policies=="Strongly Agree" | EE$Policies == "Agree",1,0)
EE$Personal1<-ifelse(EE$Personal=="Strongly Agree" | EE$Personal == "Agree",1,0)
EE$Amout1<-ifelse(EE$Amout=="Strongly Agree" | EE$Amout == "Agree",1,0)
EE$TruelyDrives1<-ifelse(EE$TruelyDrives=="Strongly Agree" | EE$TruelyDrives ==
"Agree",1,0)
EE$Teamwork1<-ifelse(EE$Teamwork=="Strongly Agree" | EE$Teamwork == "Agree",1,0)
EE$Work.life1<-ifelse(EE$Work.life=="Exreamly satisfies" | EE$Work.life ==
"satisfied",1,0)
View(EE)
```

#Step-3: Creation of models

```
m1<-lm(Employee1~Effort1, data = EE)
summary(m1)
#Multiple R-squared: 0.2142, Adjusted R-squared: 0.2025
m2<-lm(Employee1~Effort1+feedback1, data = EE)
summary(m2)
#Multiple R-squared: 0.2985, Adjusted R-squared: 0.2345
m3<-lm(Employee11~Effort1+Time1, data = EE)
summary(m3)
#Multiple R-squared: 0.3542, Adjusted R-squared: 0.275
m4<-lm(Employee11~Effort1+Time1+Important 1, data = EE)
summary(m4)
#Multiple R-squared: 0.4145, Adjusted R-squared: 0.3845
m5<-lm(Employee1~Effort1+Time1+Important 1+Leadership1, data = EE)
summary(m5)
#Multiple R-squared: 0.4325, Adjusted R-squared: 0.398
m6<-lm(Employee1~Effort1+Time1+Important 1+Leadership1+Amountl1, data = EE)
summary(m6)
#Multiple R-squared: 0.4548, Adjusted R-squared: 0.3945
m7<-lm(Employee11~Effort1+Time1+Important 1+Leadership1+Amountl1+TruelyDrives1,
data = EE)
summary(m7)
#Multiple R-squared: 0.4685, Adjusted R-squared: 0.3485
m8<-lm(Employee1~Effort1+Time1+Important
1+Leadership1+Amountl1+TruelyDrives1+Teamwork1, data = EE)
summary(m8)
#Multiple R-squared: 0.5785, Adjusted R-squared: 0.5125
```

```
library(car)
```

```
newEE=EE[-45,]
```

```
newEE1=newEE[-19,]
```

```
newEE2=newEE1[-5,]
```

```
newEE3=newEE2[-4,]
```

```
newEE4=newEE3[-69,]
```

```
newEE5=newEE4[-26,]
```

```
newEE6=newEE5[-15,]
```

## LIVE PROJECTS- Predictive Analysis Using R

```
outlierTest(m9)

m9<-lm(Employee1~Effort1+Time1+Important
1+Leadership1+Amount1+TruelyDrives1+Teamwork1+overallTeamwork1+budgetallocatio
n, data = newEE6)
summary(m9)
#Multiple R-squared:  0.7554,      Adjusted R-squared:  0.8145
str(EE)

#Assumption testing

plot(m9,4)
#Data is normal no residual
r1<-residuals(object = m9)
shapiro.test(x=r1)
#p<0.05, p-value = 3.201e-07

#independence of error
durbinWatsonTest(m9)
#p>0.05,p=0.369 hence auto correlation does not exist

#Homoscedasticity, ncv()
ncvTest(m9)
#p<0.05, p = 3.
3.7949e-12, hence the assumption of Homoscedasticity is not met

#multicollinearity
#y=m1x1+m2x2+m3x3+.....mnxn+c
#If variance inference factor(VIF)>10,bad
#orsqrt(VIF())>2 returns true, not good, all should be false
vif(m9)#As<10, good
sqrt(vif(m9))>2
#All are false so it is good. Multi Collinearity assumption is met

library(caret)
set.seed(1000)
partition<-createDataPartition(y=newEE6$Employee1, p=0.8, list = FALSE)
training<-newEE6[partition,]
test<-newEE6[-partition,]

m9<-lm(Employee1~Effort1+Time1+Important 1+Leadership1+Amount1
+Teamwork1+overallTeamwork1, data = training)
summary(m9)
#Multiple R-squared:  0.871, Adjusted R-squared:  0.8379

#Training is done on model
training$pred<-predict(m9)
training$resd<-residuals(m9)
```

## LIVE PROJECTS- Predictive Analysis Using R

```
#Validate
test$pred1=predict(m9, newdata=test)
test$resd1=test$Employee1 - test$pred1
#resd is error difference between actual and outcome
View(training)
View(test)

#Factor Analysis
#new dataset
#For Principal component
library(MASS)
str(newEE6)
View(newEE6)
newEE7=newEE6[,c(-1:-21)]
View(newEE7)
newEE7_pca<-prcomp(newEE7, center = TRUE, scale = FALSE)

#cannot rescale a constant/zero column to unit variance. Hence taken scale = FALSE
summary(newEE7_pca)
plot(newEE7_pca, type = "l")
#9 components are explaining upto 90% of variance

#factor analysis
library(psych)
library(GPArotation)
install.packages("GPArotation")
newEE7.fact<-factanal(newEE7,5, rotation = "varimax")
newEE7.fact

#applying cutoff
newEE7.fact<- factanal(newEE7[,5, rotation = "varimax", scores = "regression", cutoff=0.5)
newEE7.fact
print(newEE7, digits=2, cutoff=0.5, sort=TRUE)

#install.packages("dummies")
#library(dummies)
#EE<-read.csv("Knowledge_management.csv")
#EE.df <- data.frame(EE)

#EE.new <- dummy.data.frame(EE, sep = ".")
#names(EE.new)
#dummy(EE.new$`Effort.Strongly Agree`, sep = ".")
#EE.new.decmaking <- dummy.data.frame(EE.new$Effort, names = c("Strongly
Agree", "Agree", "Neutral", "Disagree", "Strongly Disagree"), sep = ".")
#View(EE.new)
#str(EE.new)
#attach(`Work life.Strongly Agree`)
#model6<-lm(Work life ~ ., data=EE.new)
#summary(model6)
```

**Findings:**

Most of the respondents were 21-25 years old, and 70.5% of the respondents were male.

78.1% of the respondents said that their supervisors recognize their efforts when they perform well and 63.8% of the respondents stated that they are able to give a fair amount of time to their family.

Most of the respondents agreed that their employee valuation process is fair and some were neutral, whereas very less of them disagreed to it.

High percentage of respondents agreed that the employee engagement is important.

Most of the respondents stated that their manager is professional and cordial while communicating and their team participate and encourage in task.

60% of the respondents agreed that the leadership in their organization treats all employees fairly.

Only 16.2% of the respondents disagreed to the fact that their allotted amount of work is reasonable. And 15.2% of the respondents stated that the work causes unwanted tension in their personal life.

When the respondents were asked about what truly drives engagement in an organization, so 32.4% of the respondents said constructive feedback, 26.7% said senior leadership, 22.9% said customer oriented, 9.5% said encourage flexibility and 8.6% said manual error.

40% of the respondents were Satisfied with the organization's supportive healthy work-life balance, 23.8% of the respondents were Extremely Satisfied, 22.9% were Neutral, 7.6% were Dissatisfied and 5.7% were Extremely Dissatisfied.

**Conclusion:**

Our objective for this research was to know the impact of employee engagement in an organization. And hence it can be certainly concluded that employee engagement leads to improved employee commitment & involvement towards job and thus creating a motivated workforce, that works together to achieve the common goals of the organization.

Acquiring skilled workforce is just not enough in today's changing economy like ours; instead a lot needs to be done to retain, involve and make them committed to the organization and its goals. Thus, engagement is a state where an individual is not only intellectually committed but has great emotional attachment with his/her job that goes above and beyond the call of duty so as to further the interest of the company.

The organizations should not only provide their employees with good infrastructure and other facilities but also freedom to make their work exciting and also provide them an environment wherein they can say good-bye to a monotonous work. They should focus on retention and thus working in a safe and cooperative environment adds to the engagement level of an employee.

# Preferred Mode of Transportation Used by Different Segments of People

Submitted By-  
Utkarsh Kumar Gupta (PG19143)  
Diti Ghosh (PG19042)  
Reshma Chaudhary

## Abstract:

Travel mode choice prediction of individuals is important in planning new transportation projects. This study examines travel patterns and identifies factors that influence commuters' choice of travel mode. The presented methods use individuals' characteristics, transport mode specifications and data related to places of work and residence. The dataset analysed comes from a national survey. It contains information on the daily mobility (e.g., from home to work) of individuals who either live or work in Luxembourg. We extracted individual characteristics to relate daily movements (journeys between home and work, in particular) to the characteristics of working individuals. We used the information about public transportation and some geographical location of the residential and work places. We compare the rates of successful prediction obtained by neural networks and several alternative approaches for predicting the travel mode choice using cross-validation. The results show that the artificial neural networks perform better compared to other alternatives. Our analysis can be used to support management decision-making and build predictions under uncertainty related to changes in people's behaviour, economic context or environment and transportation infrastructure. The results highlight that travel time and distance are the most influential predictors of public transport use, indicating that areas with better provision of transport infrastructure associated with higher public transport use. However, the response to transport fare policies varies amongst different user groups. Public transport is more popular for medium to long distance commuters, while student users are more vulnerable than staff to the increase

of public transport fare. Therefore, policy intervention targeting specific user groups would be more effective to encourage public transport use.

### **Introduction:**

Our topic is “preferred mode of transportation used by different segments of people. Transport modes are designed to their carry passengers or freight, but most modes can carry a combination of both. For instance, an automobile has the capacity to carry some freight while a passenger plane has a belly hold that is used for luggage and cargo. Each mode is characterized by a set of technical, operational, and commercial characteristics. Technical characteristics relate to attributes such as speed, capacity, and motives technology, while operational characteristics involve the context in which modes operated, including speed limits, safety conditions or operating hours. The demand for transport and the ownership of modes are dominant commercial characteristics. We consider the following travel modes: private car, public transport (bus or train) or soft mode (walking or cycling). By modal split we mean the composition (percentages) of commuters who use each of these travel modes.

We this topic because we can survey almost anyone who is studying or working and use a transportation mode to commute to work, even it is be walking. We have created a questionnaire for collection of our data. Which a gain valuable insight and to make our analysis with R. We want data insights on thinks like which mode of transport is most common in a metropolitan city which mode of transportation costs less, how many percentages of people use public transport. One of the most important strategies to combat the environment impacts is to encourage the use of active transport. Active transport modes as those transport forms that can encourage physical activities. According to this definition, public transport provides an active transport mode because it involves physical activities at both ends of travel. And thus, encouraging public transport use will not only benefit the environment through reducing the demand for parking spaces, but also improve public health by promoting more active lifestyle. The benefits that public transport offers make it an imperative issue to understand factors that motivate mode changes in automobile-dominant cities. The globalization of the economy and the development of transport and telecommunication technologies has led to an increasing concentration of knowledge-intensive employment and global firms in metropolitan regions.

Understanding travel behaviour is essential to informing transport management and planning. Behaviour survey is often adopted as a method to understand individual travel behaviour. Public transportation has undoubtedly played a vital role in commuting passengers to work or to places they desire, and more importantly, to reduce traffic congestion. It is undeniable that the role of public transport is to provide users with reasonably priced fares that cater to several individuals at the same time, in ensuring less congestion and pollution. his study investigates users’ expectations towards the services provided by public transportations and its relationships to customer satisfaction, loyalty and environmental factors. Additionally, it attempts to determine the most preferred mode of public transport. Though this project we also try to understand that how much money people are want to spend on transportation. Are people comfortable on public transport or not.

## Method

### Data collection :

For data collection we have used primary data collection method which include different segment of people including collage students, family people, working professional etc. For our data collection we have used following questionnaire:

- Name
- Age
- Gender
- Occupation
- Monthly Income
- Where do you live ?
- Type of Accomodation
- Most preferred transport method
- You use Private 4 Wheeler frequently
- You use Private Bike frequently
- You use 3 or 4 wheeler rental services like Ola or Uber frequently
- You use Rental Bike or Scooty like Bounce or Vogo frequently
- You use private or rental cycle frequently
- You walk frequently for work purpose
- If you have enough time you will walk rather than using vehicle for work
- If you will have enough money will you buy expensive and luxurious vehicle
- You are happy with your current daily transportation mode
- You feel your daily travelling is expensive
- If you will buy new vehicle which one you prefer
- Mode of transport you use to commute to work
- How many bikes do you have at home
- How many 4 wheeler you have at home
- How many Bicycle do you have at home
- On average, monthly how much money do you spend to travel to work? (include fare charges, petrol/diesel. vehicle servicing) in Rupees
- On average, monthly how many times you use rental services? (Ola, Uber, Rapido, Bounce, Vogo etc)

## Analysis

**Heading Replacement:** Since my data has big headings as questionnaire so I have replace them with small headings for easy analysis.

**Data Cleaning:** As my data have missing values so first I have used function “naniar” function to find out percentage of missing values in each column. Then I have replaced missing values in columns with high missing values by most frequent variable. After replacing higher

## LIVE PROJECTS- Predictive Analysis Using R

percentage of missing values I have removed missing values from remaining columns by using “omit” function.

**Data Compatibility:** For making data compatible for linear regression step wise I have converted all column with factors in to factors and assigned numeric values to columns on the basis of their intensity.

### Model Building:

```
m1 <- glm(M.P.T_method ~.,data = nres1, family = binomial())
summary(m1)
```

```
library(MASS)
stepAIC(m1)
```

```
m2 <- glm(formula = M.P.T_method ~ Gender + Inhabitant + T_Acc + RentS +
          Rental_BS + PR_cycle + Walk_W + Walk_P + EM_ExpL Veh + H_DTM +
          Feel_Exp + N_Pref + MoT_WP + N_Bikes + N_Wh + Mon_Exp + N_RentSev,
          family = binomial(), data = nres1)
```

### Confusion Matrix

```
cm <- table(nres1$M.P.T_method,nres1$pred)
```


```
cm
```

```
# TP = 129, FP = 59, TN = 0, FN = 0
```

```
summary(cm)
```

	0	1
0	129	0
1	0	59





**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