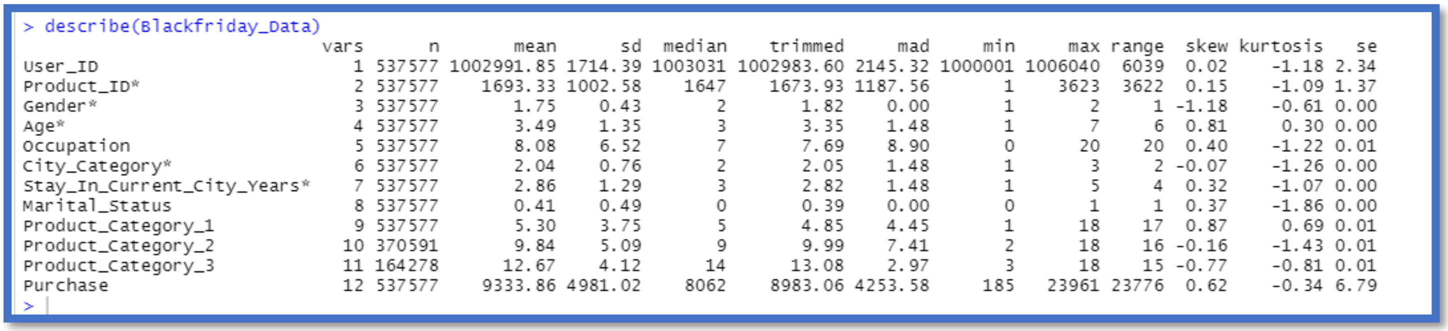
# GROUP 242: BLACK FRIDAY SALES

| CWID | First name | Last Name | IIT Email |
|---|---|---|---|
| **A20401921** | Anusha | Satish | athattehalli@hawk.iit.edu |
| **A20436206** | Sivaranjani | Prabasankar | sprabasankar@hawk.iit.edu |

# DESCRIPTION OF BLACK FRIDAY DATASET

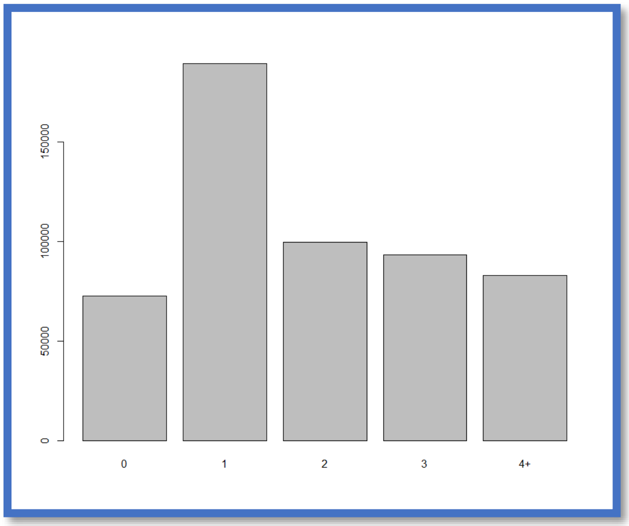


```
> describe(Blackfriday_Data)
                            vars      n       mean      sd  median    trimmed     mad     min     max range  skew kurtosis   se
User_ID                        1 537577 1002991.85 1714.39 1003031 1002983.60 2145.32 1000001 1006040  6039  0.02    -1.18 2.34
Product_ID*                    2 537577    1693.33 1002.58    1647    1673.93 1187.56       1    3623  3622  0.15    -1.09 1.37
Gender*                        3 537577       1.75    0.43       2       1.82    0.00       1       2     1 -1.18    -0.61 0.00
Age*                           4 537577       3.49    1.35       3       3.35    1.48       1       7     6  0.81     0.30 0.00
Occupation                     5 537577       8.08    6.52       7       7.69    8.90       0      20    20  0.40    -1.22 0.01
City_Category*                 6 537577       2.04    0.76       2       2.05    1.48       1       3     2 -0.07    -1.26 0.00
Stay_In_Current_City_Years*    7 537577       2.86    1.29       3       2.82    1.48       1       5     4  0.32    -1.07 0.00
Marital_Status                 8 537577       0.41    0.49       0       0.39    0.00       0       1     1  0.37    -1.86 0.00
Product_Category_1             9 537577       5.30    3.75       5       4.85    4.45       1      18    17  0.87     0.69 0.01
Product_Category_2            10 370591       9.84    5.09       9       9.99    7.41       2      18    16 -0.16    -1.43 0.01
Product_Category_3            11 164278      12.67    4.12      14      13.08    2.97       3      18    15 -0.77    -0.81 0.01
Purchase                      12 537577    9333.86 4981.02    8062    8983.06 4253.58     185   23961 23776  0.62    -0.34 6.79
> 
```

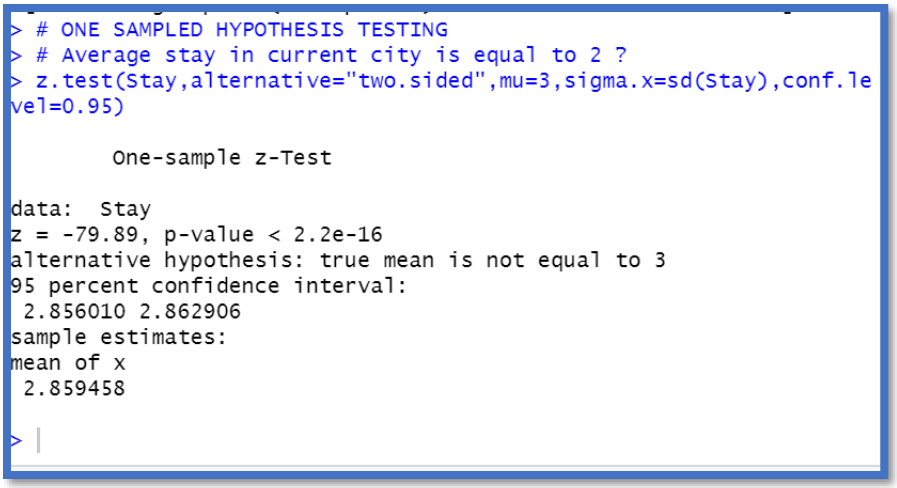
## HYPOTHESIS

### I) One sample – One Tailed Hypothesis Testing on STAY IN THE CURRENT CITY

BAR GRAPH – STAY IN THE CURRENT CITY (IN YEARS)

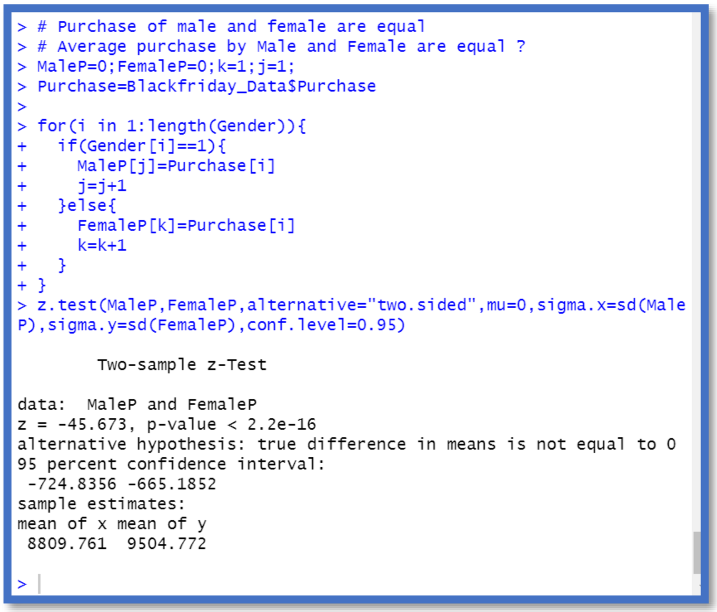


Z.TEST



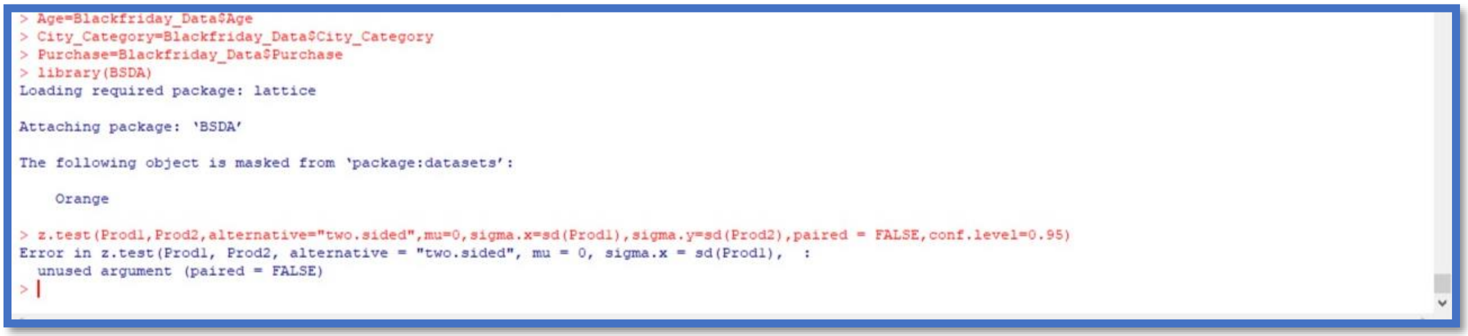
```
> # ONE SAMPLED HYPOTHESIS TESTING
> # Average stay in current city is equal to 2 ?
> z.test(Stay,alternative="two.sided",mu=3,sigma.x=sd(Stay),conf.le
vel=0.95)

        One-sample z-Test

data:  Stay
z = -79.89, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.856010 2.862906
sample estimates:
mean of x
 2.859458

> 
```

## II) One sample – Two Tailed Hypothesis Testing on GENDER

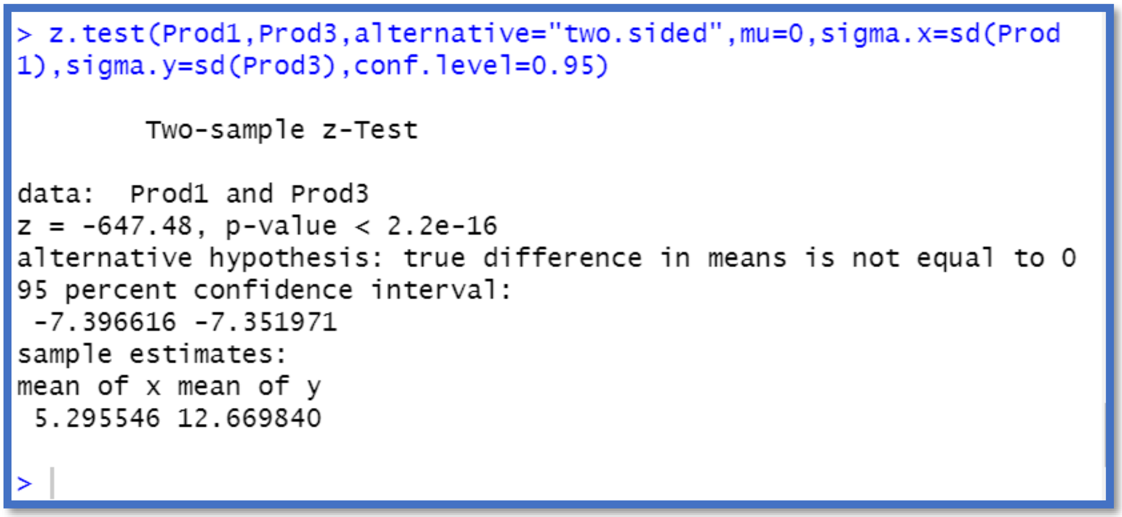


```
> # Purchase of male and female are equal
> # Average purchase by Male and Female are equal ?
> MaleP=0;FemaleP=0;k=1;j=1;
> Purchase=Blackfriday_Data$Purchase
>
> for(i in 1:length(Gender)){
+   if(Gender[i]==1){
+     MaleP[j]=Purchase[i]
+     j=j+1
+   }else{
+     FemaleP[k]=Purchase[i]
+     k=k+1
+   }
+ }
> z.test(MaleP,FemaleP,alternative="two.sided",mu=0,sigma.x=sd(Male
P),sigma.y=sd(FemaleP),conf.level=0.95)

        Two-sample z-Test

data:  MaleP and FemaleP
z = -45.673, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -724.8356 -665.1852
sample estimates:
mean of x mean of y
 8809.761  9504.772

> |
```

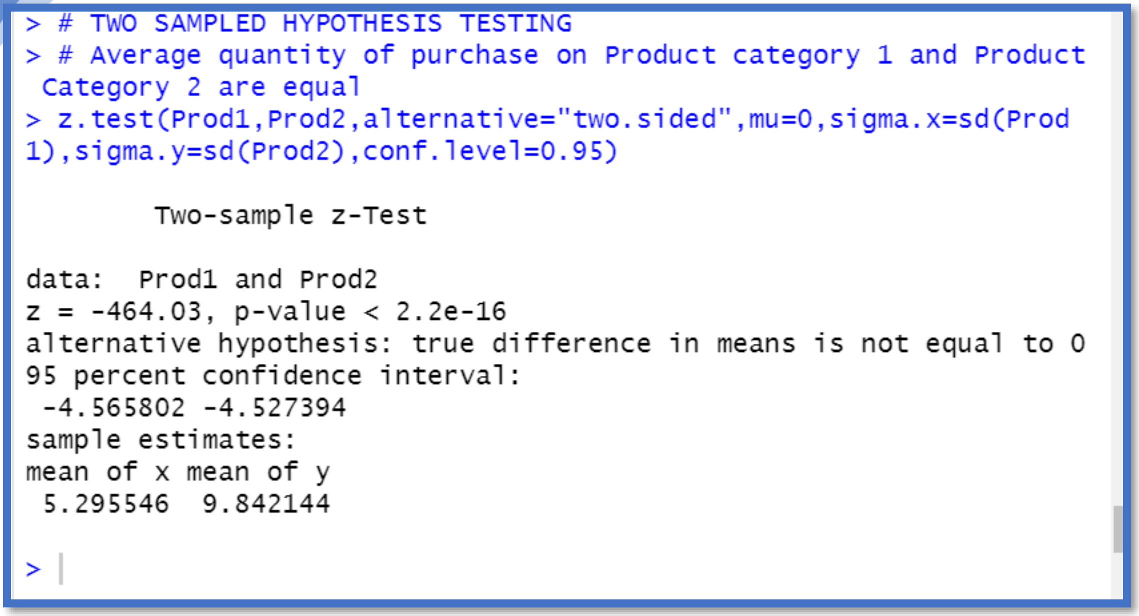
## III) Two Sampled Hypothesis Testing on PRODUCT CATERORIES



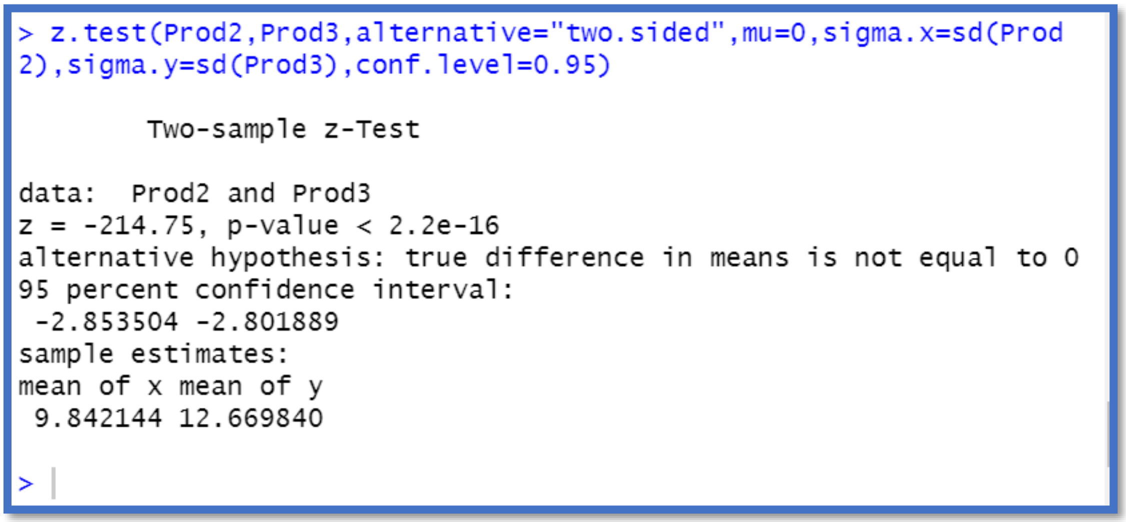
```
> Age=Blackfriday_Data$Age
> City_Category=Blackfriday_Data$City_Category
> Purchase=Blackfriday_Data$Purchase
> library(BSDA)
Loading required package: lattice

Attaching package: 'BSDA'

The following object is masked from 'package:datasets':

    Orange

> z.test(Prod1,Prod2,alternative="two.sided",mu=0,sigma.x=sd(Prod1),sigma.y=sd(Prod2),paired = FALSE,conf.level=0.95)
Error in z.test(Prod1, Prod2, alternative = "two.sided", mu = 0, sigma.x = sd(Prod1),  :
  unused argument (paired = FALSE)
> |
```



```
> z.test(Prod1,Prod3,alternative="two.sided",mu=0,sigma.x=sd(Prod
1),sigma.y=sd(Prod3),conf.level=0.95)

        Two-sample z-Test

data:  Prod1 and Prod3
z = -647.48, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.396616 -7.351971
sample estimates:
mean of x mean of y
 5.295546 12.669840

> |
```

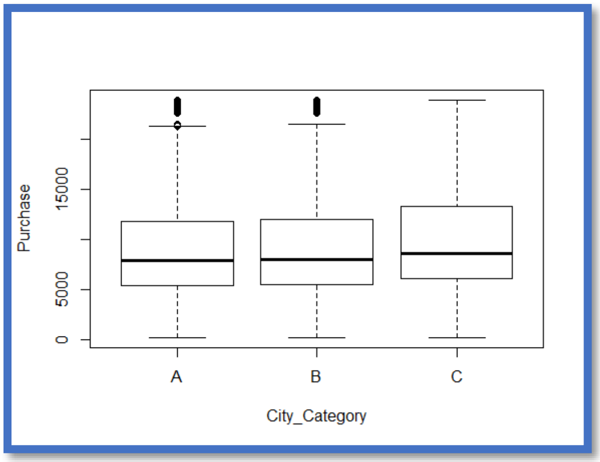


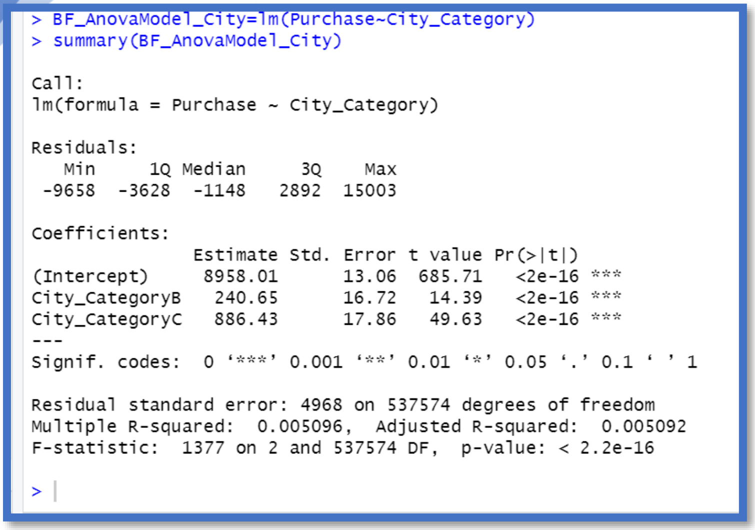
```
> # TWO SAMPLED HYPOTHESIS TESTING
> # Average quantity of purchase on Product category 1 and Product
 Category 2 are equal
> z.test(Prod1,Prod2,alternative="two.sided",mu=0,sigma.x=sd(Prod
1),sigma.y=sd(Prod2),conf.level=0.95)

        Two-sample z-Test

data:  Prod1 and Prod2
z = -464.03, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.565802 -4.527394
sample estimates:
mean of x mean of y 
 5.295546  9.842144 

> 
```



```
> z.test(Prod2,Prod3,alternative="two.sided",mu=0,sigma.x=sd(Prod
2),sigma.y=sd(Prod3),conf.level=0.95)

        Two-sample z-Test

data:  Prod2 and Prod3
z = -214.75, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.853504 -2.801889
sample estimates:
mean of x mean of y 
 9.842144 12.669840 

> 
```

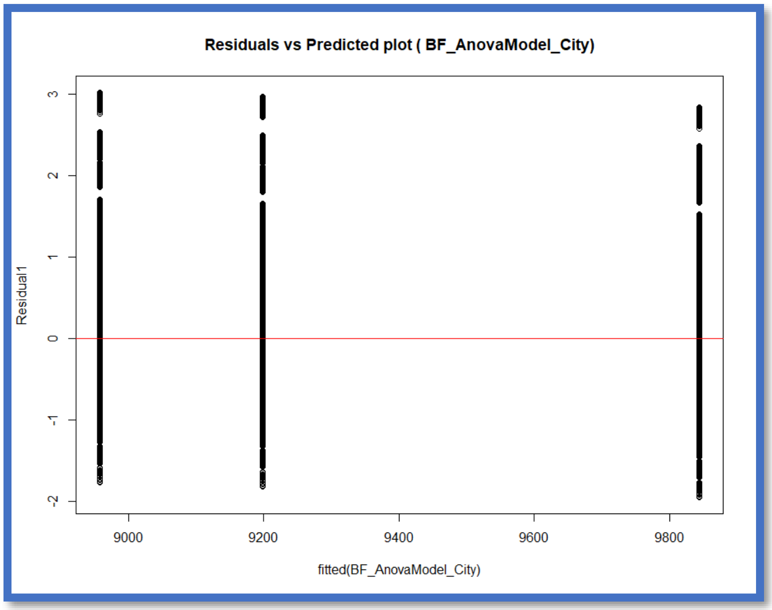
## ANOVA - City category



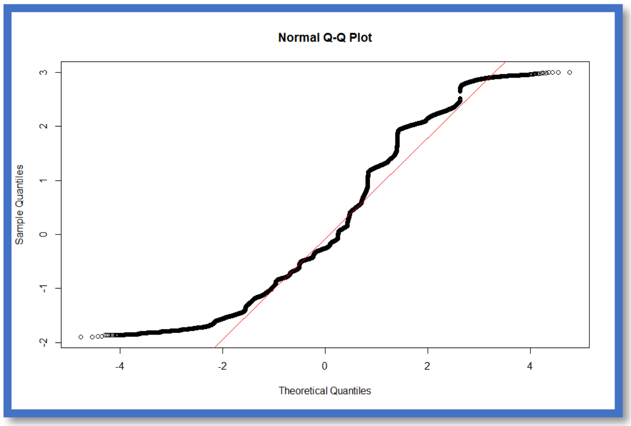


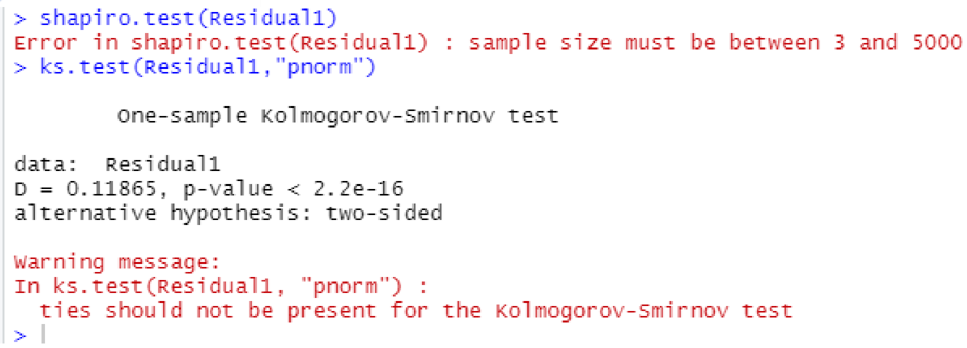
```
> BF_AnovaModel_City=lm(Purchase~City_Category)
> summary(BF_AnovaModel_City)

Call:
lm(formula = Purchase ~ City_Category)

Residuals:
   Min     1Q  Median     3Q    Max 
 -9658  -3628  -1148   2892  15003 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     8958.01      13.06  685.71   <2e-16 ***
City_CategoryB   240.65      16.72   14.39   <2e-16 ***
City_CategoryC   886.43      17.86   49.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4968 on 537574 degrees of freedom
Multiple R-squared:  0.005096,  Adjusted R-squared:  0.005092 
F-statistic:  1377 on 2 and 537574 DF,  p-value: < 2.2e-16

> 
```

RESIDUAL ANALYSIS

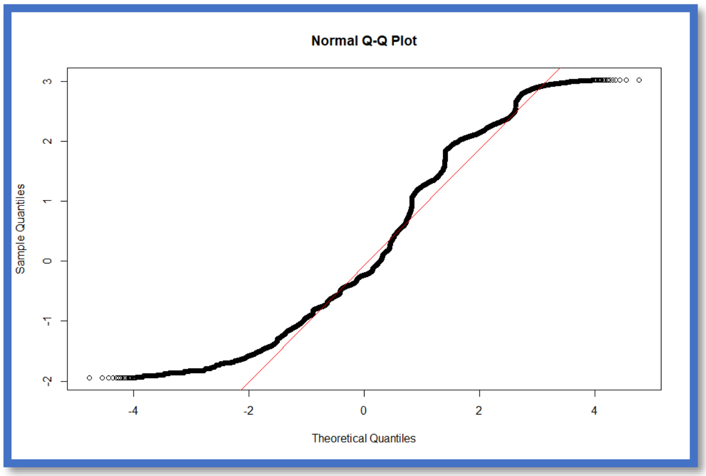


Normality test

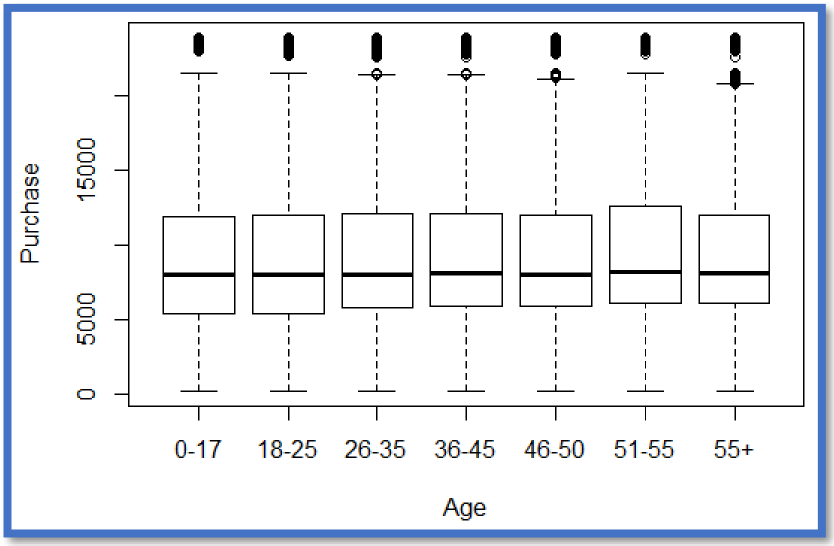




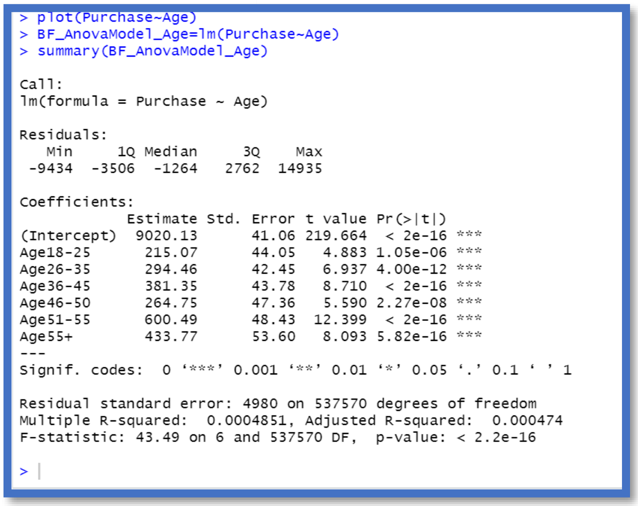
```
> shapiro.test(Residual1)
Error in shapiro.test(Residual1) : sample size must be between 3 and 5000
> ks.test(Residual1,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  Residual1
D = 0.11865, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(Residual1, "pnorm") :
  ties should not be present for the Kolmogorov-Smirnov test
> 
```



## ANOVA – AGE Group

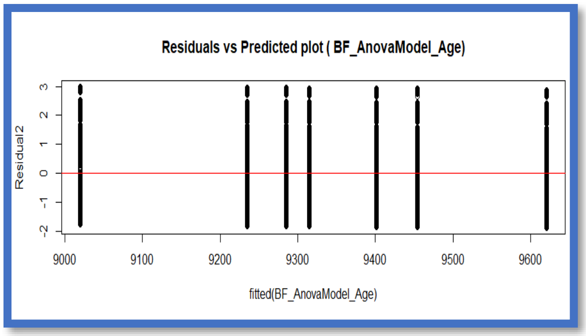


## F TEST

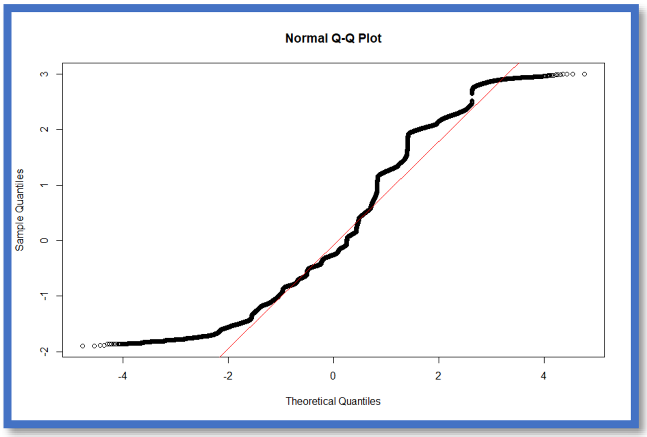


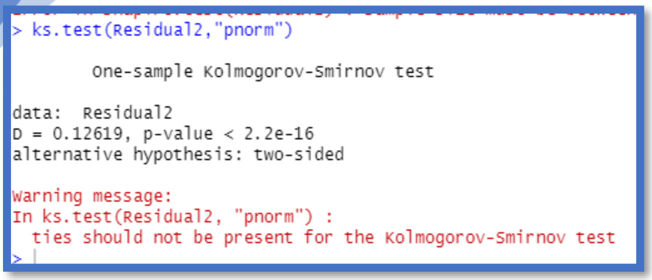
```
> plot(Purchase~Age)
> BF_AnovaModel_Age=lm(Purchase~Age)
> summary(BF_AnovaModel_Age)

Call:
lm(formula = Purchase ~ Age)

Residuals:
   Min     1Q Median     3Q    Max 
 -9434  -3506  -1264   2762  14935 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9020.13      41.06 219.664  < 2e-16 ***
Age18-25      215.07      44.05   4.883 1.05e-06 ***
Age26-35      294.46      42.45   6.937 4.00e-12 ***
Age36-45      381.35      43.78   8.710  < 2e-16 ***
Age46-50      264.75      47.36   5.590 2.27e-08 ***
Age51-55      600.49      48.43  12.399  < 2e-16 ***
Age55+        433.77      53.60   8.093 5.82e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4980 on 537570 degrees of freedom
Multiple R-squared:  0.0004851,	Adjusted R-squared:  0.000474 
F-statistic: 43.49 on 6 and 537570 DF,  p-value: < 2.2e-16

> |
```

## RESIDUAL ANALYSIS



## NORMALITY TEST

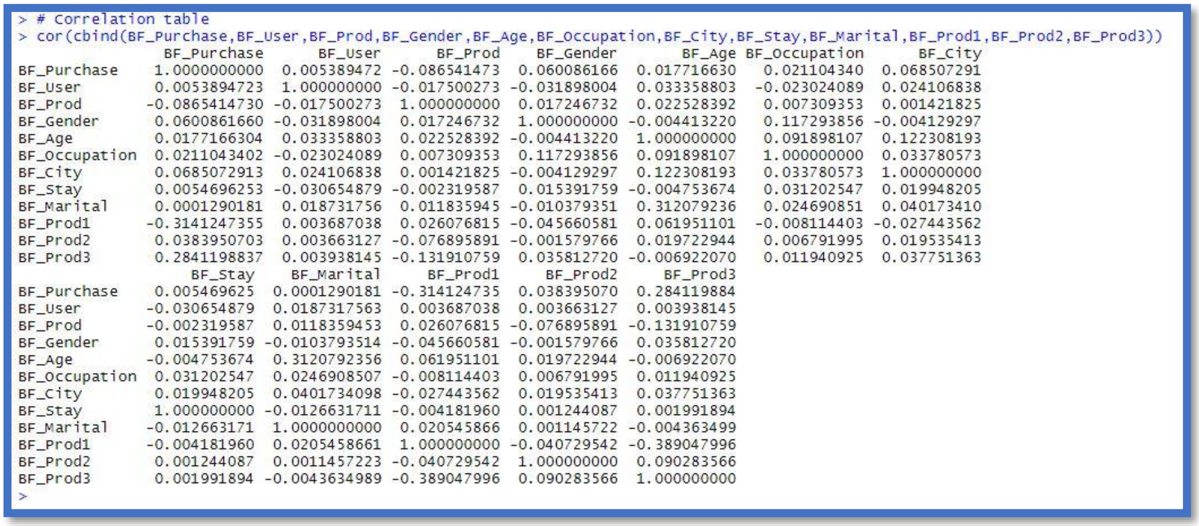




```
> ks.test(Residual2,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  Residual2
D = 0.12619, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(Residual2, "pnorm") :
  ties should not be present for the Kolmogorov-Smirnov test
>
```

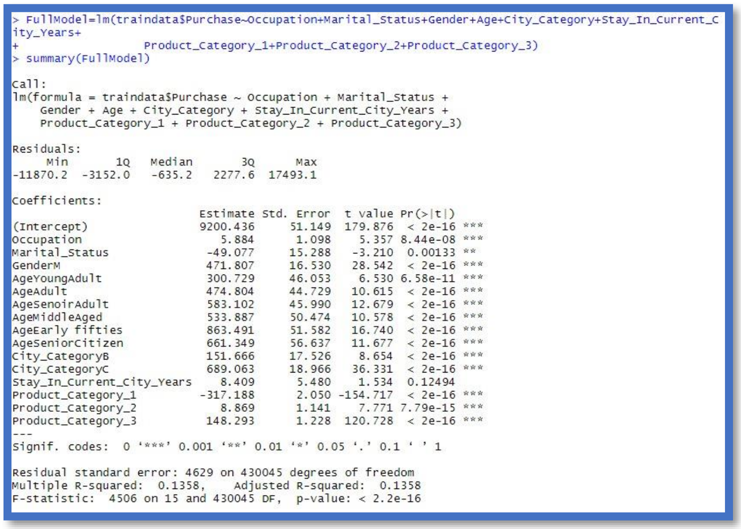
# Linear Regression Model

## CORRELATION

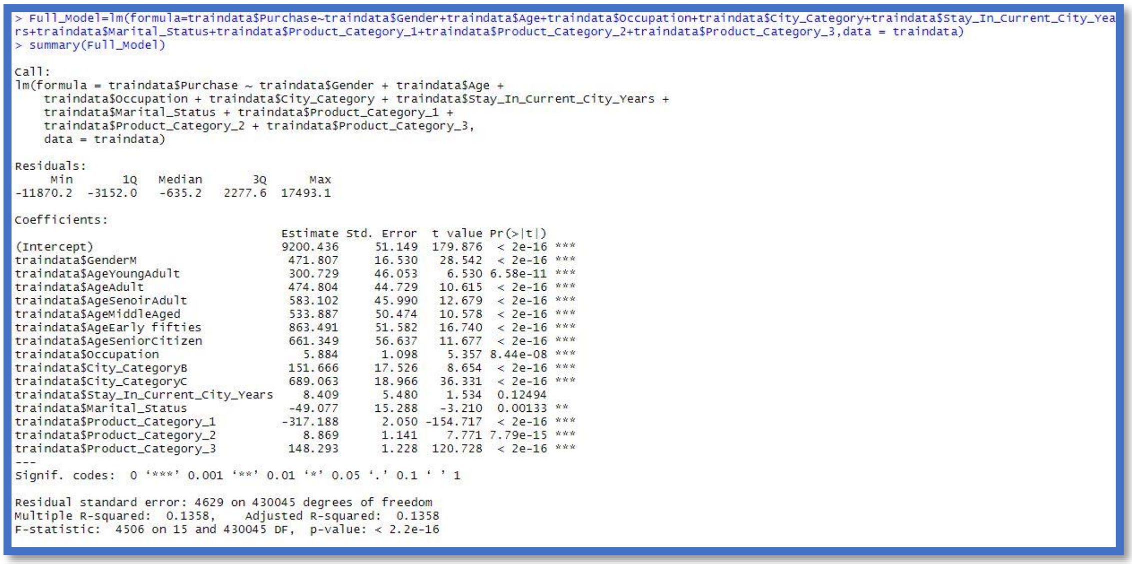


```
> # Correlation table
> cor(cbind(BF_Purchase,BF_User,BF_Prod,BF_Gender,BF_Age,BF_Occupation,BF_City,BF_Stay,BF_Marital,BF_Prod1,BF_Prod2,BF_Prod3))
                BF_Purchase      BF_User      BF_Prod    BF_Gender       BF_Age BF_Occupation      BF_City
BF_Purchase    1.0000000000  0.005389472 -0.086541473  0.060086166  0.017716630   0.021104340  0.068507291
BF_User        0.0053894723  1.000000000 -0.017500273 -0.031898004  0.033358803  -0.023024089  0.024106838
BF_Prod       -0.0865414730 -0.017500273  1.000000000  0.017246732  0.022528392   0.007309353  0.001421825
BF_Gender      0.0600861660 -0.031898004  0.017246732  1.000000000 -0.004413220   0.117293856 -0.004129297
BF_Age         0.0177166304  0.033358803  0.022528392 -0.004413220  1.000000000   0.091898107  0.122308193
BF_Occupation  0.0211043402 -0.023024089  0.007309353  0.117293856  0.091898107   1.000000000  0.033780573
BF_City        0.0685072913  0.024106838  0.001421825 -0.004129297  0.122308193   0.033780573  1.000000000
BF_Stay        0.0054696253 -0.030654879 -0.002319587  0.015391759 -0.004753674   0.031202547  0.019948205
BF_Marital     0.0001290181  0.018731756  0.011835945 -0.010379351  0.312079236   0.024690851  0.040173410
BF_Prod1      -0.3141247355  0.003687038  0.026076815 -0.045660581  0.061951101  -0.008114403 -0.027443562
BF_Prod2       0.0383950703  0.003663127 -0.076895891 -0.001579766  0.019722944   0.006791995  0.019535413
BF_Prod3       0.2841198837  0.003938145 -0.131910759  0.035812720 -0.006922070   0.011940925  0.037751363
                   BF_Stay   BF_Marital     BF_Prod1     BF_Prod2     BF_Prod3
BF_Purchase    0.005469625  0.0001290181 -0.314124735  0.038395070  0.284119884
BF_User       -0.030654879  0.0187317563  0.003687038  0.003663127  0.003938145
BF_Prod       -0.002319587  0.0118359453  0.026076815 -0.076895891 -0.131910759
BF_Gender      0.015391759 -0.0103793514 -0.045660581 -0.001579766  0.035812720
BF_Age        -0.004753674  0.3120792356  0.061951101  0.019722944 -0.006922070
BF_Occupation  0.031202547  0.0246908507 -0.008114403  0.006791995  0.011940925
BF_City        0.019948205  0.0401734098 -0.027443562  0.019535413  0.037751363
BF_Stay        1.000000000 -0.0126631711 -0.004181960  0.001244087  0.001991894
BF_Marital    -0.012663171  1.0000000000  0.020545866  0.001145722 -0.004363499
BF_Prod1      -0.004181960  0.0205458661  1.000000000 -0.040729542 -0.389047996
BF_Prod2       0.001244087  0.0011457223 -0.040729542  1.000000000  0.090283566
BF_Prod3       0.001991894 -0.0043634989 -0.389047996  0.090283566  1.000000000
>
```

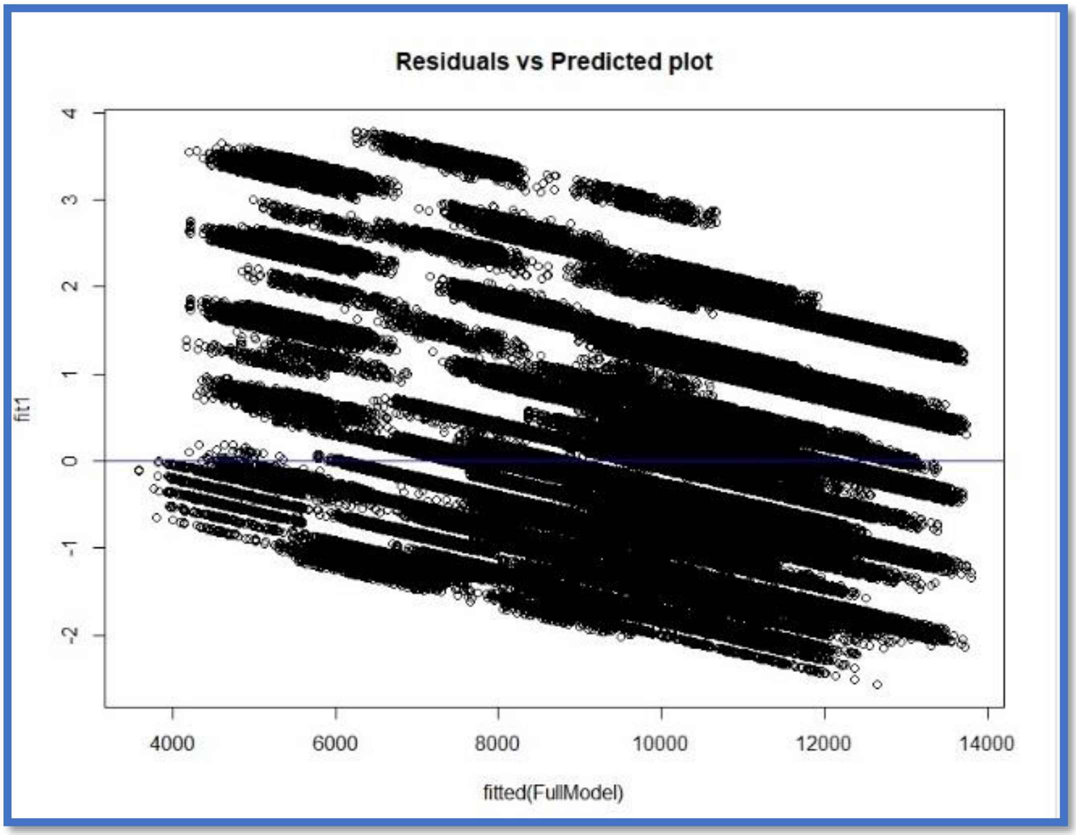
## FULL MODEL



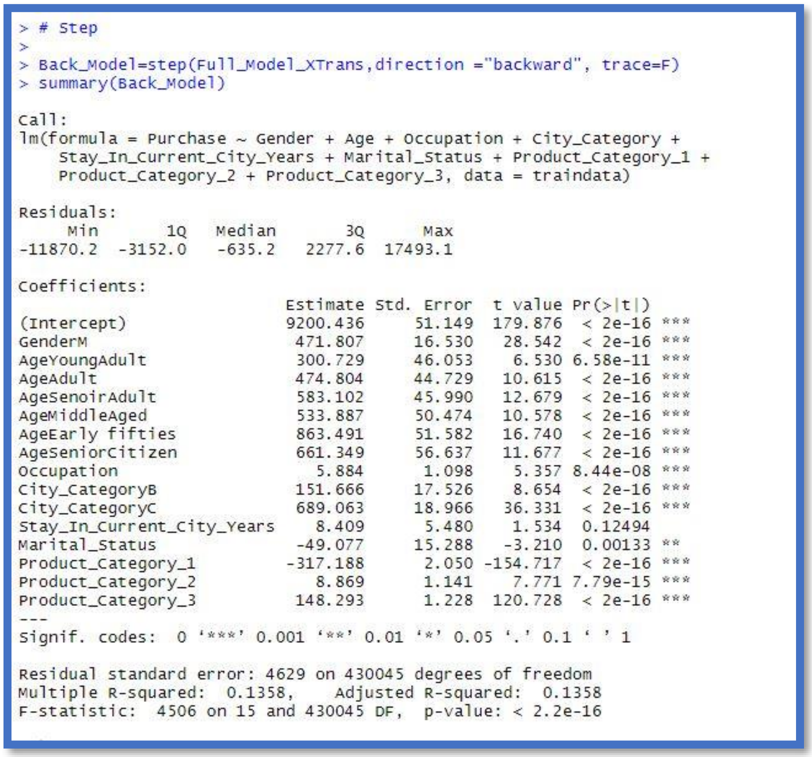
```
> FullModel=lm(traindata$Purchase~Occupation+Marital_Status+Gender+Age+City_Category+Stay_In_Current_C
ity_Years+
+                    Product_Category_1+Product_Category_2+Product_Category_3)
> summary(FullModel)

Call:
lm(formula = traindata$Purchase ~ Occupation + Marital_Status + 
    Gender + Age + City_Category + Stay_In_Current_City_Years + 
    Product_Category_1 + Product_Category_2 + Product_Category_3)

Residuals:
     Min       1Q   Median       3Q      Max 
-11870.2  -3152.0   -635.2   2277.6  17493.1 

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                 9200.436     51.149  179.876  < 2e-16 ***
Occupation                     5.884      1.098    5.357 8.44e-08 ***
Marital_Status               -49.077     15.288   -3.210  0.00133 ** 
GenderM                      471.807     16.530   28.542  < 2e-16 ***
AgeYoungAdult                300.729     46.053    6.530 6.58e-11 ***
AgeAdult                     474.804     44.729   10.615  < 2e-16 ***
AgeSenoirAdult               583.102     45.990   12.679  < 2e-16 ***
AgeMiddleAged                533.887     50.474   10.578  < 2e-16 ***
AgeEarly fifties             863.491     51.582   16.740  < 2e-16 ***
AgeSeniorCitizen             661.349     56.637   11.677  < 2e-16 ***
City_CategoryB               151.666     17.526    8.654  < 2e-16 ***
City_CategoryC               689.063     18.966   36.331  < 2e-16 ***
Stay_In_Current_City_Years     8.409      5.480    1.534  0.12494    
Product_Category_1          -317.188      2.050 -154.717  < 2e-16 ***
Product_Category_2             8.869      1.141    7.771 7.79e-15 ***
Product_Category_3           148.293      1.228  120.728  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4629 on 430045 degrees of freedom
Multiple R-squared:  0.1358,	Adjusted R-squared:  0.1358 
F-statistic:  4506 on 15 and 430045 DF,  p-value: < 2.2e-16
```



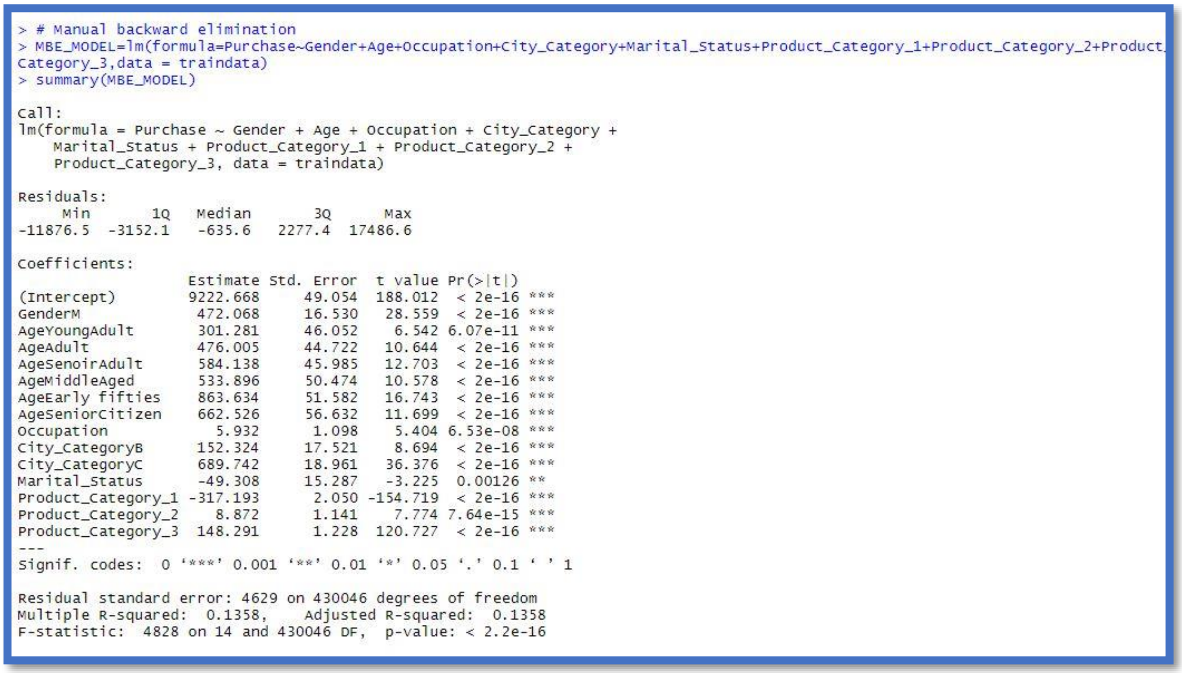
```
> Full_Model=lm(formula=traindata$Purchase~traindata$Gender+traindata$Age+traindata$Occupation+traindata$City_Category+traindata$Stay_In_Current_City_Yea
rs+traindata$Marital_Status+traindata$Product_Category_1+traindata$Product_Category_2+traindata$Product_Category_3,data = traindata)
> summary(Full_Model)

Call:
lm(formula = traindata$Purchase ~ traindata$Gender + traindata$Age +
    traindata$Occupation + traindata$City_Category + traindata$Stay_In_Current_City_Years +
    traindata$Marital_Status + traindata$Product_Category_1 +
    traindata$Product_Category_2 + traindata$Product_Category_3,
    data = traindata)

Residuals:
     Min       1Q   Median       3Q      Max
-11870.2  -3152.0   -635.2   2277.6  17493.1

Coefficients:
                                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)                         9200.436     51.149  179.876  < 2e-16 ***
traindata$GenderM                    471.807     16.530   28.542  < 2e-16 ***
traindata$AgeYoungAdult              300.729     46.053    6.530 6.58e-11 ***
traindata$AgeAdult                   474.804     44.729   10.615  < 2e-16 ***
traindata$AgeSenoirAdult             583.102     45.990   12.679  < 2e-16 ***
traindata$AgeMiddleAged              533.887     50.474   10.578  < 2e-16 ***
traindata$AgeEarly fifties           863.491     51.582   16.740  < 2e-16 ***
traindata$AgeSeniorCitizen           661.349     56.637   11.677  < 2e-16 ***
traindata$Occupation                   5.884      1.098    5.357 8.44e-08 ***
traindata$City_CategoryB             151.666     17.526    8.654  < 2e-16 ***
traindata$City_CategoryC             689.063     18.966   36.331  < 2e-16 ***
traindata$Stay_In_Current_City_Years   8.409      5.480    1.534  0.12494
traindata$Marital_Status             -49.077     15.288   -3.210  0.00133 **
traindata$Product_Category_1        -317.188      2.050 -154.717  < 2e-16 ***
traindata$Product_Category_2           8.869      1.141    7.771 7.79e-15 ***
traindata$Product_Category_3         148.293      1.228  120.728  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4629 on 430045 degrees of freedom
Multiple R-squared:  0.1358,	Adjusted R-squared:  0.1358
F-statistic:  4506 on 15 and 430045 DF,  p-value: < 2.2e-16
```



## BACKWARD ELIMINATION



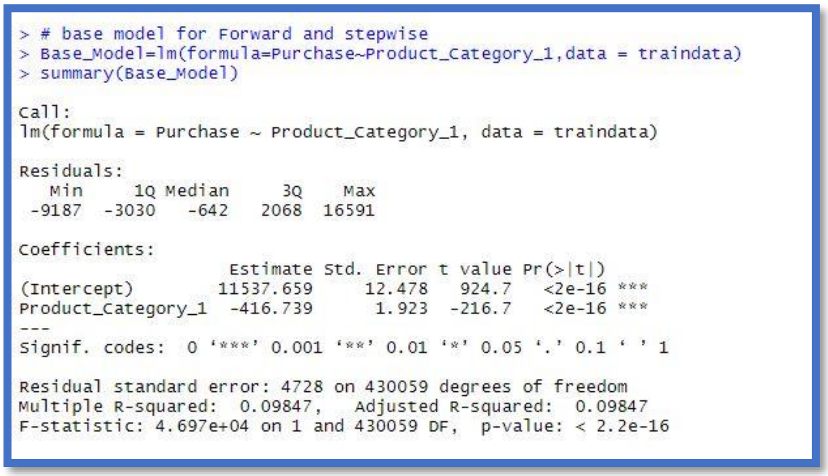
```
> # Step
>
> Back_Model=step(Full_Model_XTrans,direction ="backward", trace=F)
> summary(Back_Model)

Call:
lm(formula = Purchase ~ Gender + Age + Occupation + City_Category +
    Stay_In_Current_City_Years + Marital_Status + Product_Category_1 +
    Product_Category_2 + Product_Category_3, data = traindata)

Residuals:
     Min       1Q   Median       3Q      Max
-11870.2  -3152.0   -635.2   2277.6  17493.1

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)                9200.436     51.149  179.876  < 2e-16 ***
GenderM                     471.807     16.530   28.542  < 2e-16 ***
AgeYoungAdult               300.729     46.053    6.530 6.58e-11 ***
AgeAdult                    474.804     44.729   10.615  < 2e-16 ***
AgeSenoirAdult              583.102     45.990   12.679  < 2e-16 ***
AgeMiddleAged               533.887     50.474   10.578  < 2e-16 ***
AgeEarly fifties            863.491     51.582   16.740  < 2e-16 ***
AgeSeniorCitizen            661.349     56.637   11.677  < 2e-16 ***
Occupation                    5.884      1.098    5.357 8.44e-08 ***
City_CategoryB              151.666     17.526    8.654  < 2e-16 ***
City_CategoryC              689.063     18.966   36.331  < 2e-16 ***
Stay_In_Current_City_Years    8.409      5.480    1.534  0.12494
Marital_Status              -49.077     15.288   -3.210  0.00133 **
Product_Category_1         -317.188      2.050 -154.717  < 2e-16 ***
Product_Category_2            8.869      1.141    7.771 7.79e-15 ***
Product_Category_3          148.293      1.228  120.728  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4629 on 430045 degrees of freedom
Multiple R-squared:  0.1358,    Adjusted R-squared:  0.1358
F-statistic:  4506 on 15 and 430045 DF,  p-value: < 2.2e-16
```



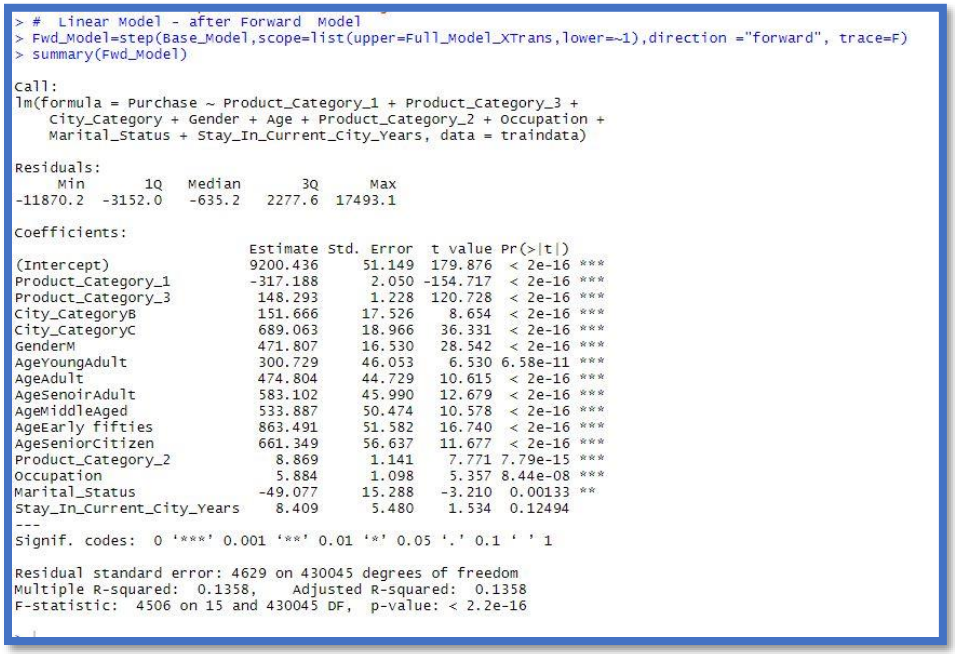
```
> # Manual backward elimination
> MBE_MODEL=lm(formula=Purchase~Gender+Age+Occupation+City_Category+Marital_Status+Product_Category_1+Product_Category_2+Product_
Category_3,data = traindata)
> summary(MBE_MODEL)

Call:
lm(formula = Purchase ~ Gender + Age + Occupation + City_Category +
    Marital_Status + Product_Category_1 + Product_Category_2 +
    Product_Category_3, data = traindata)

Residuals:
     Min       1Q   Median       3Q      Max
-11876.5  -3152.1   -635.6   2277.4  17486.6

Coefficients:
                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)        9222.668     49.054  188.012  < 2e-16 ***
GenderM             472.068     16.530   28.559  < 2e-16 ***
AgeYoungAdult       301.281     46.052    6.542 6.07e-11 ***
AgeAdult            476.005     44.722   10.644  < 2e-16 ***
AgeSenoirAdult      584.138     45.985   12.703  < 2e-16 ***
AgeMiddleAged       533.896     50.474   10.578  < 2e-16 ***
AgeEarly fifties    863.634     51.582   16.743  < 2e-16 ***
AgeSeniorCitizen    662.526     56.632   11.699  < 2e-16 ***
Occupation            5.932      1.098    5.404 6.53e-08 ***
City_CategoryB      152.324     17.521    8.694  < 2e-16 ***
City_CategoryC      689.742     18.961   36.376  < 2e-16 ***
Marital_Status      -49.308     15.287   -3.225  0.00126 **
Product_Category_1 -317.193      2.050 -154.719  < 2e-16 ***
Product_Category_2    8.872      1.141    7.774 7.64e-15 ***
Product_Category_3  148.291      1.228  120.727  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4629 on 430046 degrees of freedom
Multiple R-squared:  0.1358,    Adjusted R-squared:  0.1358
F-statistic:  4828 on 14 and 430046 DF,  p-value: < 2.2e-16
```

## FORWARD SELECTION

BASE MODEL

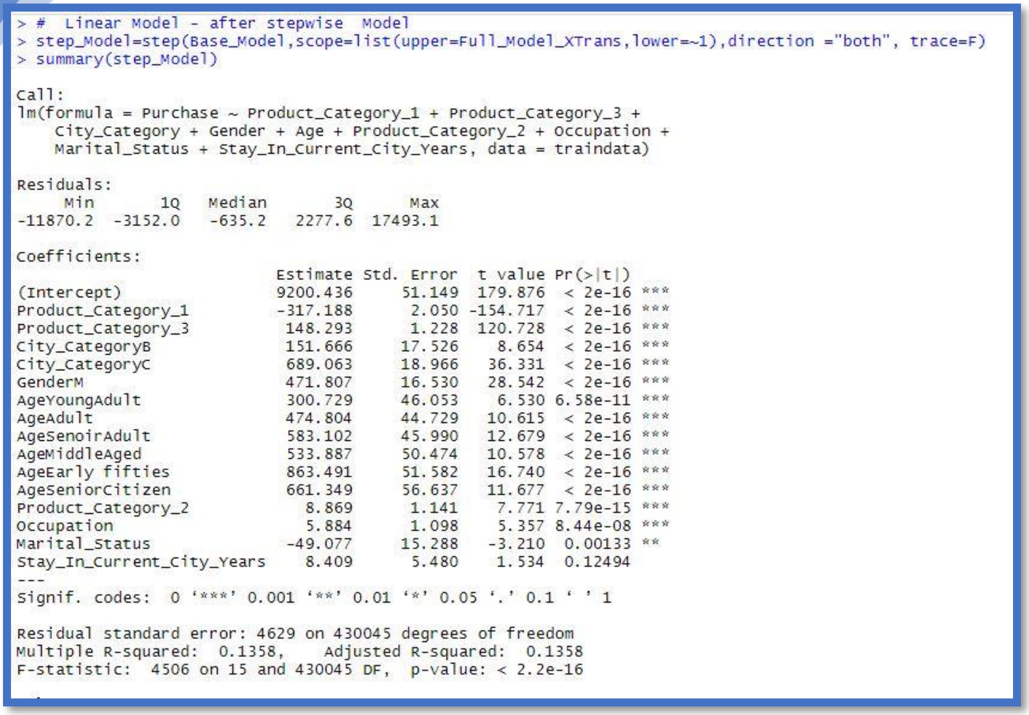


```
> # base model for Forward and stepwise
> Base_Model=lm(formula=Purchase~Product_Category_1,data = traindata)
> summary(Base_Model)

Call:
lm(formula = Purchase ~ Product_Category_1, data = traindata)

Residuals:
   Min     1Q Median     3Q    Max 
 -9187  -3030   -642   2068  16591 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        11537.659     12.478   924.7   <2e-16 ***
Product_Category_1  -416.739      1.923  -216.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4728 on 430059 degrees of freedom
Multiple R-squared:  0.09847,	Adjusted R-squared:  0.09847 
F-statistic: 4.697e+04 on 1 and 430059 DF,  p-value: < 2.2e-16
```



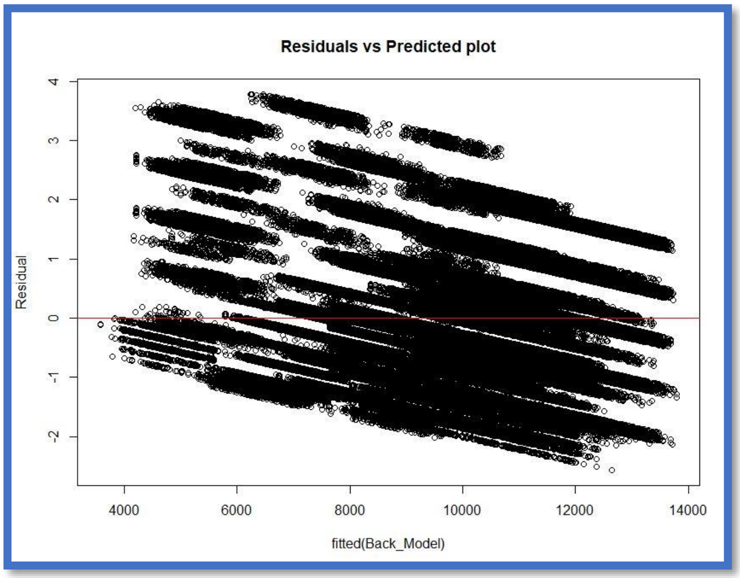
```
> #  Linear Model - after Forward  Model
> Fwd_Model=step(Base_Model,scope=list(upper=Full_Model_XTrans,lower=~1),direction ="forward", trace=F)
> summary(Fwd_Model)

Call:
lm(formula = Purchase ~ Product_Category_1 + Product_Category_3 + 
    City_Category + Gender + Age + Product_Category_2 + Occupation + 
    Marital_Status + Stay_In_Current_City_Years, data = traindata)

Residuals:
     Min       1Q   Median       3Q      Max 
-11870.2  -3152.0   -635.2   2277.6  17493.1 

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                9200.436     51.149  179.876  < 2e-16 ***
Product_Category_1         -317.188      2.050 -154.717  < 2e-16 ***
Product_Category_3          148.293      1.228  120.728  < 2e-16 ***
City_CategoryB              151.666     17.526    8.654  < 2e-16 ***
City_CategoryC              689.063     18.966   36.331  < 2e-16 ***
GenderM                     471.807     16.530   28.542  < 2e-16 ***
AgeYoungAdult               300.729     46.053    6.530 6.58e-11 ***
AgeAdult                    474.804     44.729   10.615  < 2e-16 ***
AgeSenoirAdult              583.102     45.990   12.679  < 2e-16 ***
AgeMiddleAged               533.887     50.474   10.578  < 2e-16 ***
AgeEarly fifties            863.491     51.582   16.740  < 2e-16 ***
AgeSeniorCitizen            661.349     56.637   11.677  < 2e-16 ***
Product_Category_2            8.869      1.141    7.771 7.79e-15 ***
Occupation                    5.884      1.098    5.357 8.44e-08 ***
Marital_Status              -49.077     15.288   -3.210  0.00133 ** 
Stay_In_Current_City_Years    8.409      5.480    1.534  0.12494    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4629 on 430045 degrees of freedom
Multiple R-squared:  0.1358,	Adjusted R-squared:  0.1358 
F-statistic:  4506 on 15 and 430045 DF,  p-value: < 2.2e-16
```

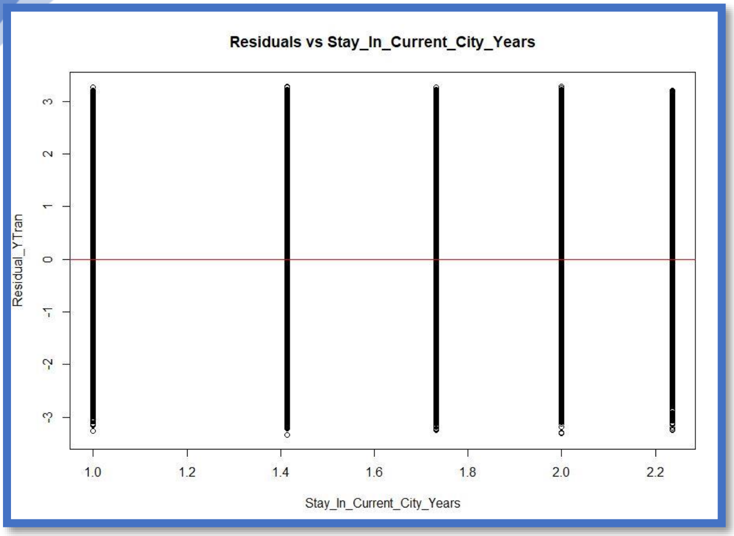
STEPWISE METHOD

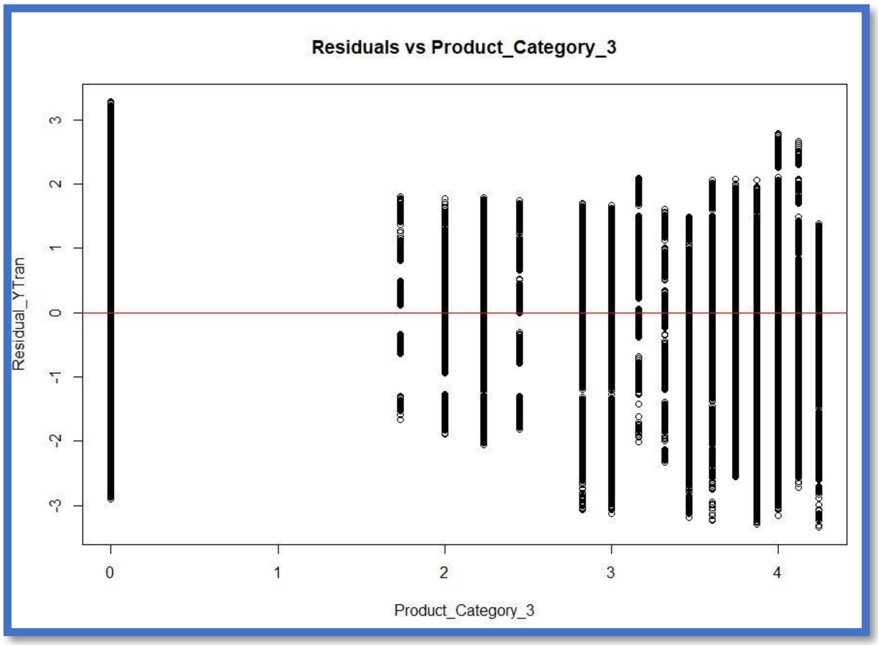


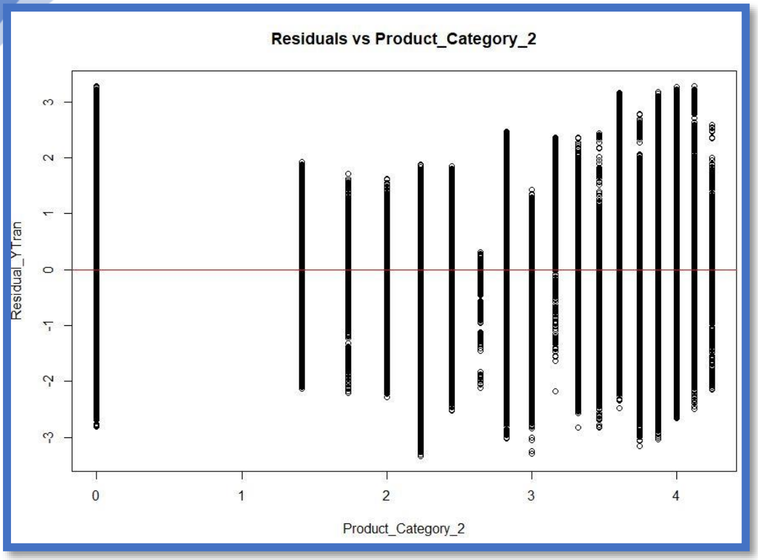
```
> #  Linear Model - after stepwise  Model
> step_Model=step(Base_Model,scope=list(upper=Full_Model_XTrans,lower=~1),direction ="both", trace=F)
> summary(step_Model)

Call:
lm(formula = Purchase ~ Product_Category_1 + Product_Category_3 + 
    City_Category + Gender + Age + Product_Category_2 + Occupation + 
    Marital_Status + Stay_In_Current_City_Years, data = traindata)

Residuals:
     Min       1Q   Median       3Q      Max 
-11870.2  -3152.0   -635.2   2277.6  17493.1 

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                9200.436     51.149  179.876  < 2e-16 ***
Product_Category_1         -317.188      2.050 -154.717  < 2e-16 ***
Product_Category_3          148.293      1.228  120.728  < 2e-16 ***
City_CategoryB              151.666     17.526    8.654  < 2e-16 ***
City_CategoryC              689.063     18.966   36.331  < 2e-16 ***
GenderM                     471.807     16.530   28.542  < 2e-16 ***
AgeYoungAdult               300.729     46.053    6.530 6.58e-11 ***
AgeAdult                    474.804     44.729   10.615  < 2e-16 ***
AgeSenoirAdult              583.102     45.990   12.679  < 2e-16 ***
AgeMiddleAged               533.887     50.474   10.578  < 2e-16 ***
AgeEarly fifties            863.491     51.582   16.740  < 2e-16 ***
AgeSeniorCitizen            661.349     56.637   11.677  < 2e-16 ***
Product_Category_2            8.869      1.141    7.771 7.79e-15 ***
Occupation                    5.884      1.098    5.357 8.44e-08 ***
Marital_Status              -49.077     15.288   -3.210  0.00133 ** 
Stay_In_Current_City_Years    8.409      5.480    1.534  0.12494    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4629 on 430045 degrees of freedom
Multiple R-squared:  0.1358,	Adjusted R-squared:  0.1358 
F-statistic:  4506 on 15 and 430045 DF,  p-value: < 2.2e-16
```

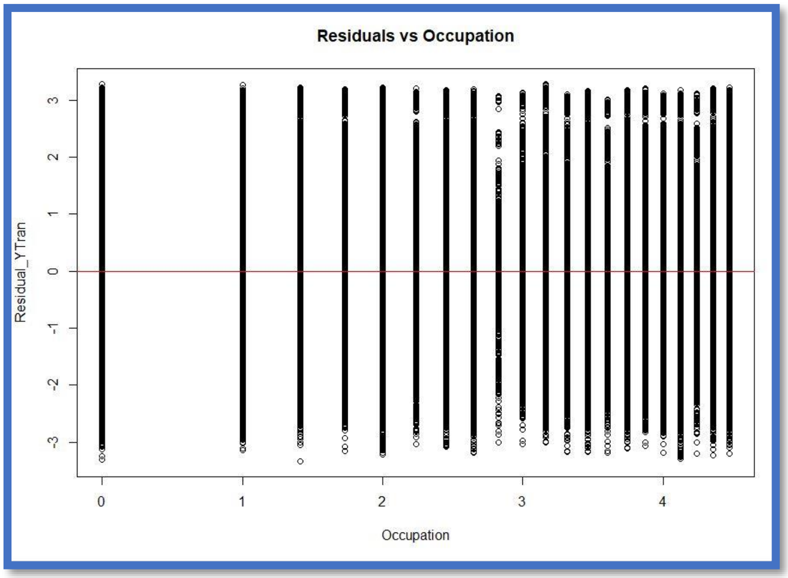
RESIDUAL ANALYSIS

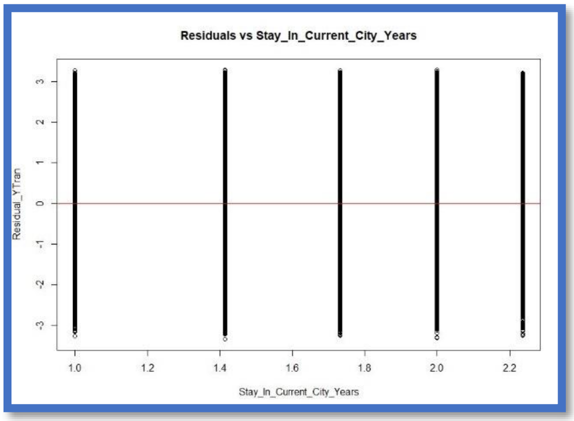


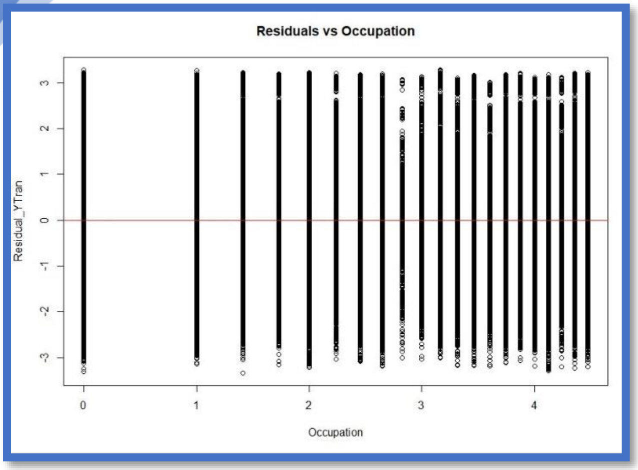

Residuals vs Stay_In_Current_City_Years
Residual_YTran
3
2
1
0
-1
-2
-3
1.0
1.2
1.4
1.6
1.8
2.0
2.2
Stay_In_Current_City_Years

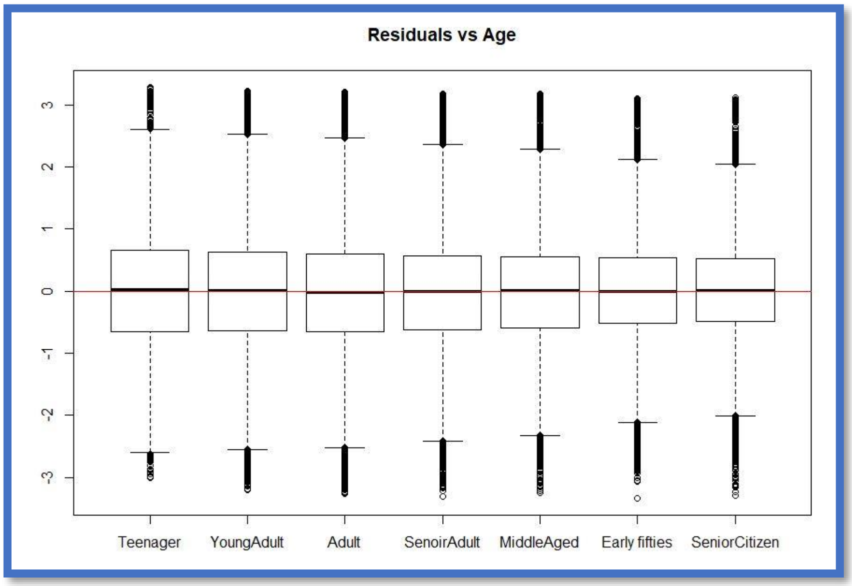

Residuals vs Product_Category_3
Residual_YTran
3
2
1
0
-1
-2
-3
0
1
2
3
4
Product_Category_3


Residuals vs Product_Category_2
Residual_YTran
3
2
1
0
-1
-2
-3
0
1
2
3
4
Product_Category_2

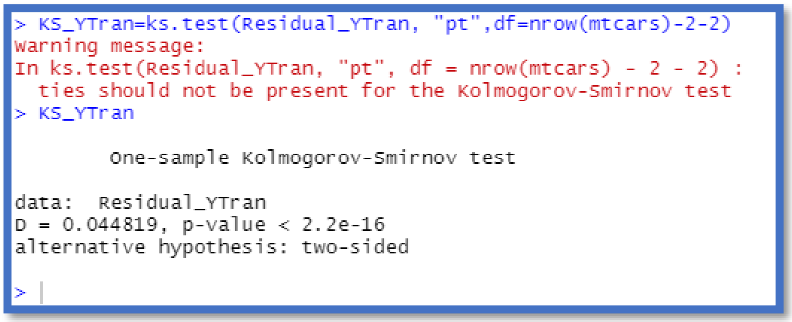

Residuals vs Occupation
Residual_YTran
3
2
1
0
-1
-2
-3
0
1
2
3
4
Occupation


Residuals vs Stay_In_Current_City_Years
Residual_YTran
3
2
1
0
-1
-2
-3
1.0
1.2
1.4
1.6
1.8
2.0
2.2
Stay_In_Current_City_Years

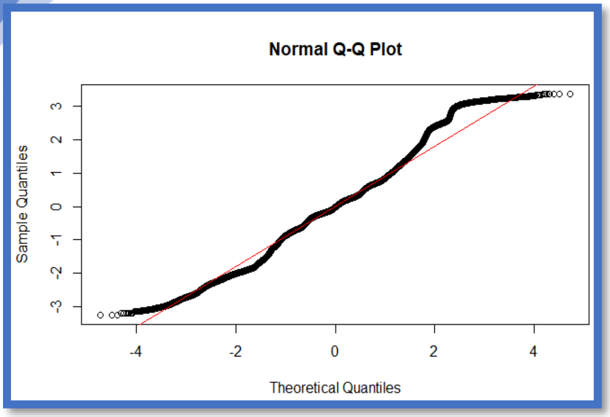




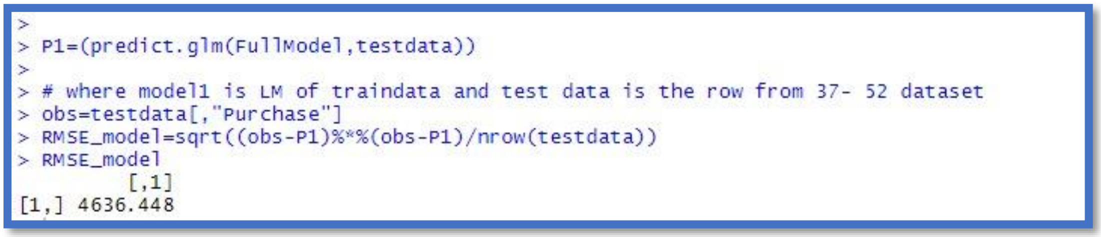
## NORMALITY TEST



```
> KS_YTran=ks.test(Residual_YTran, "pt",df=nrow(mtcars)-2-2)
Warning message:
In ks.test(Residual_YTran, "pt", df = nrow(mtcars) - 2 - 2) :
  ties should not be present for the Kolmogorov-Smirnov test
> KS_YTran

        One-sample Kolmogorov-Smirnov test

data:  Residual_YTran
D = 0.044819, p-value < 2.2e-16
alternative hypothesis: two-sided

> 
```



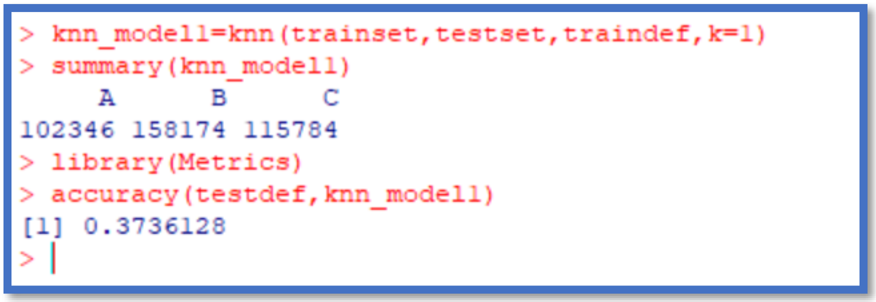
RMSE



```
>
> P1=(predict.glm(FullModel,testdata))
>
> # where model1 is LM of traindata and test data is the row from 37- 52 dataset
> obs=testdata[,"Purchase"]
> RMSE_model=sqrt((obs-P1)%*%(obs-P1)/nrow(testdata))
> RMSE_model
         [,1]
[1,] 4636.448
```

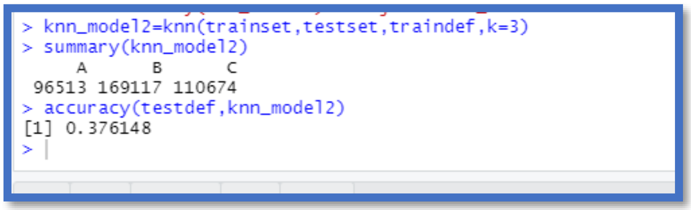
## K- Nearest Neighbor Classification Technique

K=1



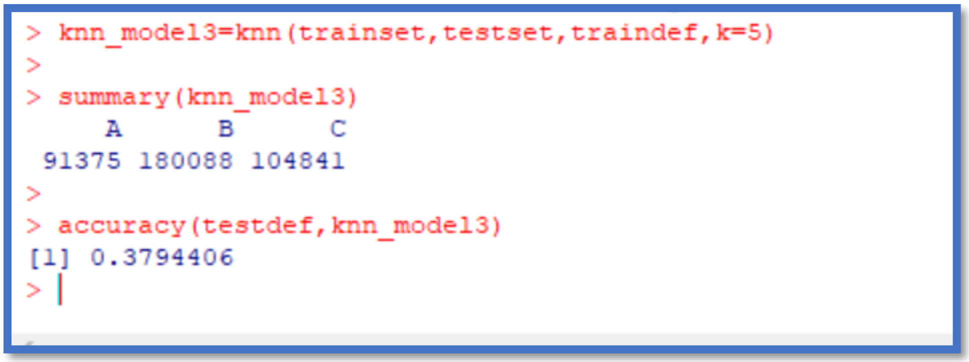
```
> knn_model1=knn(trainset,testset,traindef,k=1)
> summary(knn_model1)
     A      B      C 
102346 158174 115784 
> library(Metrics)
> accuracy(testdef,knn_model1)
[1] 0.3736128
> 
```

K=3



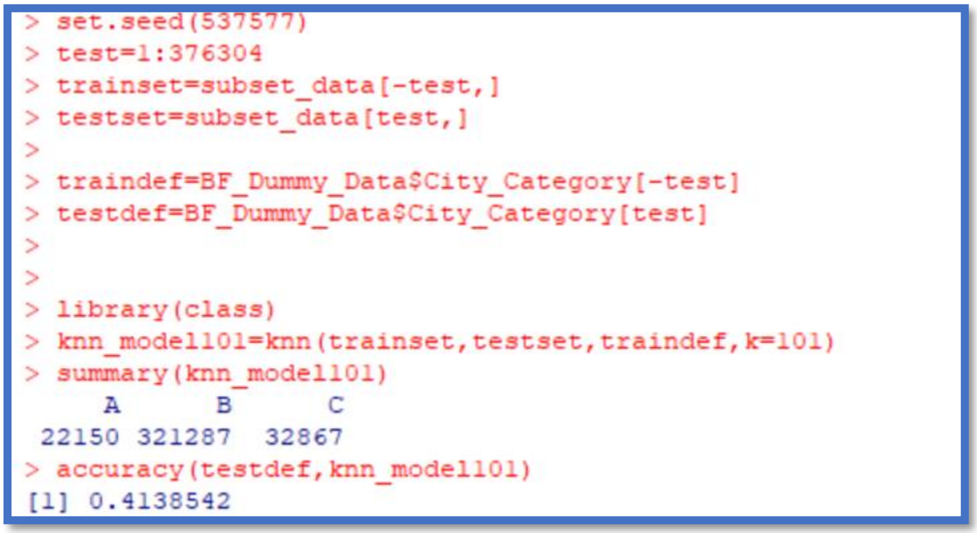
```
> knn_model2=knn(trainset,testset,traindef,k=3)
> summary(knn_model2)
     A      B      C 
 96513 169117 110674 
> accuracy(testdef,knn_model2)
[1] 0.376148
> 
```

K=5



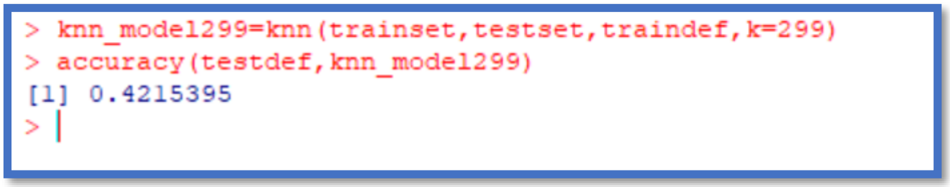
```
> knn_model3=knn(trainset,testset,traindef,k=5)
>
> summary(knn_model3)
     A      B      C
 91375 180088 104841
>
> accuracy(testdef,knn_model3)
[1] 0.3794406
> |
```

K=101



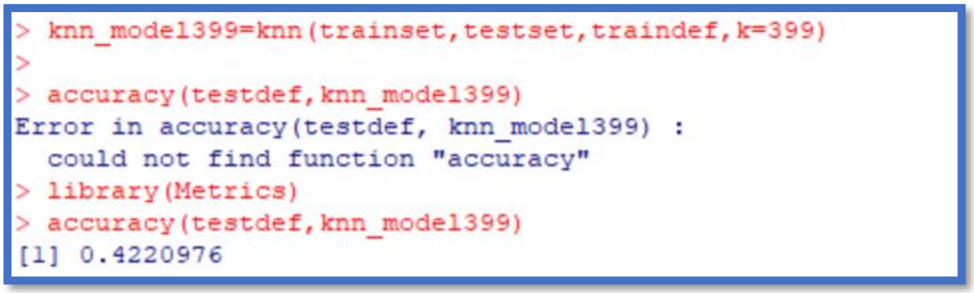
```
> set.seed(537577)
> test=1:376304
> trainset=subset_data[-test,]
> testset=subset_data[test,]
>
> traindef=BF_Dummy_Data$City_Category[-test]
> testdef=BF_Dummy_Data$City_Category[test]
>
>
> library(class)
> knn_model101=knn(trainset,testset,traindef,k=101)
> summary(knn_model101)
     A      B      C
 22150 321287  32867
> accuracy(testdef,knn_model101)
[1] 0.4138542
```

K = 299



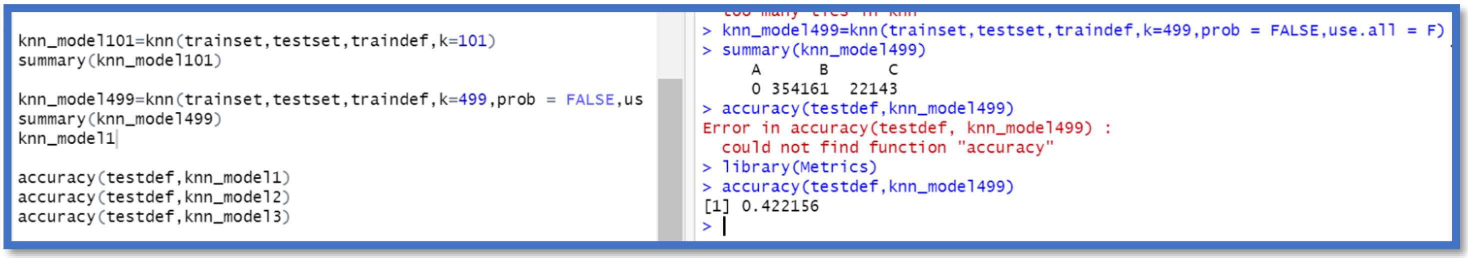
```
> knn_model299=knn(trainset,testset,traindef,k=299)
> accuracy(testdef,knn_model299)
[1] 0.4215395
> |
```

K=399



```
> knn_model399=knn(trainset,testset,traindef,k=399)
>
> accuracy(testdef,knn_model399)
Error in accuracy(testdef, knn_model399) :
  could not find function "accuracy"
> library(Metrics)
> accuracy(testdef,knn_model399)
[1] 0.4220976
```

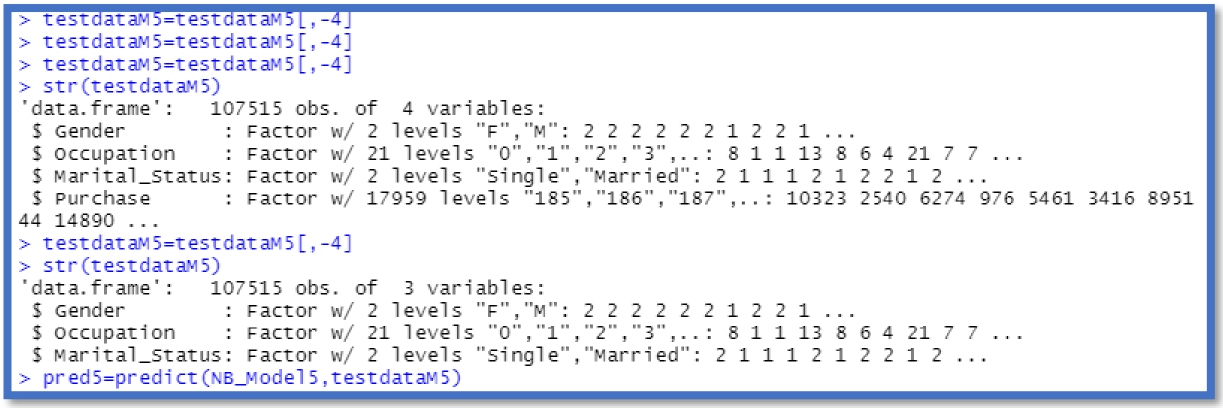
K=499



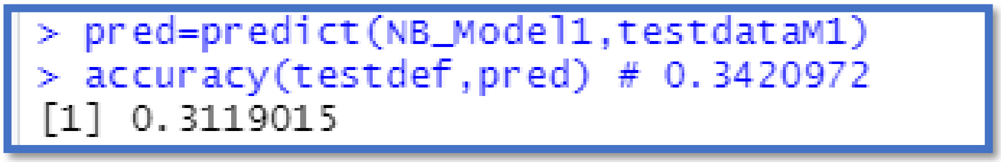
```
knn_model101=knn(trainset,testset,traindef,k=101)
summary(knn_model101)

knn_model499=knn(trainset,testset,traindef,k=499,prob = FALSE,us
summary(knn_model499)
knn_model1

accuracy(testdef,knn_model1)
accuracy(testdef,knn_model2)
accuracy(testdef,knn_model3)
```

```
  too many ties in knn
> knn_model499=knn(trainset,testset,traindef,k=499,prob = FALSE,use.all = F)
> summary(knn_model499)
     A      B      C
     0 354161  22143
> accuracy(testdef,knn_model499)
Error in accuracy(testdef, knn_model499) :
  could not find function "accuracy"
> library(Metrics)
> accuracy(testdef,knn_model499)
[1] 0.422156
> 
```

## Naive Bayes Classification Technique



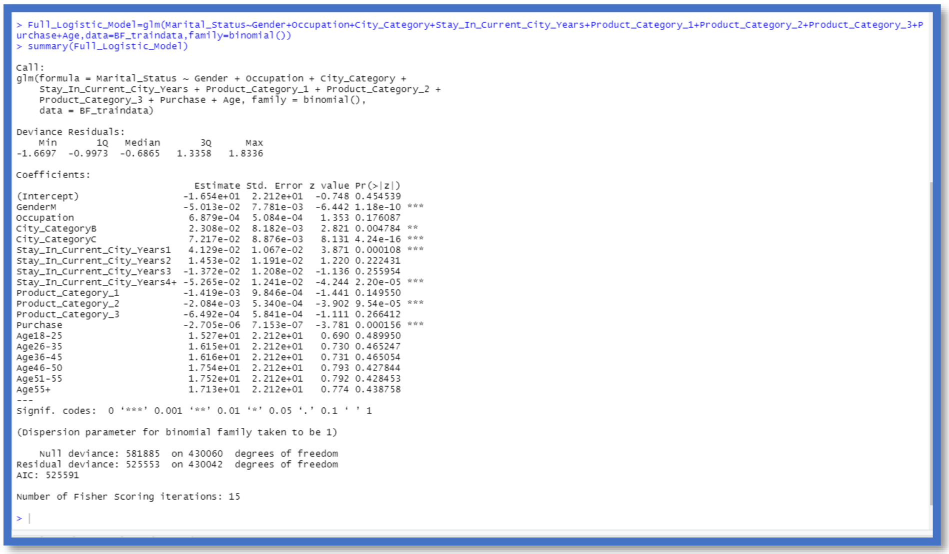
```
> testdataM5=testdataM5[,-4]
> testdataM5=testdataM5[,-4]
> testdataM5=testdataM5[,-4]
> str(testdataM5)
'data.frame':   107515 obs. of  4 variables:
 $ Gender        : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 2 2 1 ...
 $ Occupation    : Factor w/ 21 levels "0","1","2","3",..: 8 1 1 13 8 6 4 21 7 7 ...
 $ Marital_Status: Factor w/ 2 levels "Single","Married": 2 1 1 1 2 1 2 2 1 2 ...
 $ Purchase      : Factor w/ 17959 levels "185","186","187",..: 10323 2540 6274 976 5461 3416 8951
44 14890 ...
> testdataM5=testdataM5[,-4]
> str(testdataM5)
'data.frame':   107515 obs. of  3 variables:
 $ Gender        : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 2 2 1 ...
 $ Occupation    : Factor w/ 21 levels "0","1","2","3",..: 8 1 1 13 8 6 4 21 7 7 ...
 $ Marital_Status: Factor w/ 2 levels "Single","Married": 2 1 1 1 2 1 2 2 1 2 ...
> pred5=predict(NB_Model5,testdataM5)
```



```
> pred=predict(NB_Model1,testdataM1)
> accuracy(testdef,pred) # 0.3420972
[1] 0.3119015
```

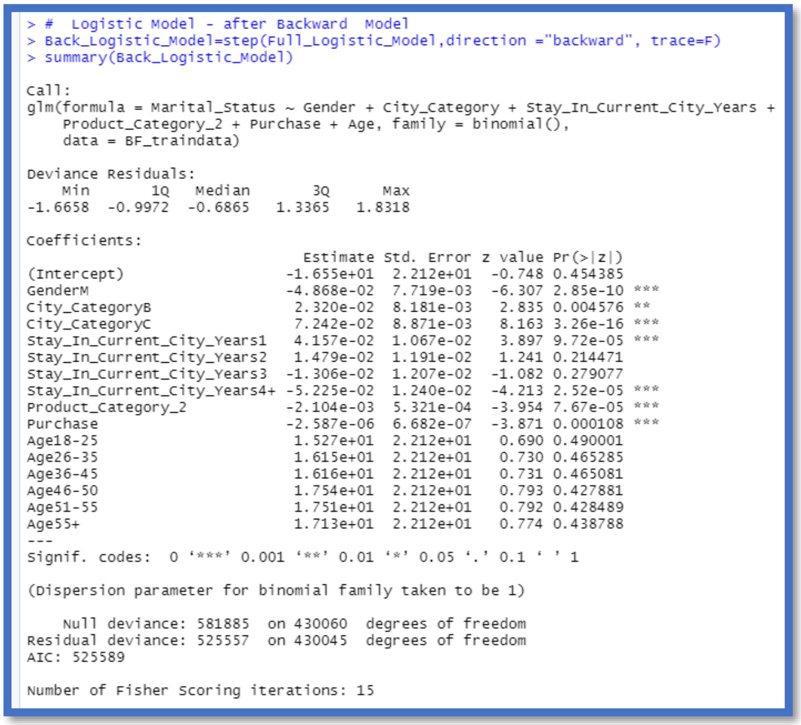
## Logistic Regression Technique

FULL MODEL



```
> Full_Logistic_Model=glm(Marital_Status~Gender+Occupation+City_Category+Stay_In_Current_City_Years+Product_Category_1+Product_Category_2+Product_Category_3+P
urchase+Age,data=BF_traindata,family=binomial())
> summary(Full_Logistic_Model)

Call:
glm(formula = Marital_Status ~ Gender + Occupation + City_Category + 
    Stay_In_Current_City_Years + Product_Category_1 + Product_Category_2 + 
    Product_Category_3 + Purchase + Age, family = binomial(), 
    data = BF_traindata)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6697  -0.9973  -0.6865   1.3358   1.8336  

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 -1.654e+01  2.212e+01  -0.748 0.454539    
GenderM                     -5.013e-02  7.781e-03  -6.442 1.18e-10 ***
Occupation                   6.879e-04  5.084e-04   1.353 0.176087    
City_CategoryB               2.308e-02  8.182e-03   2.821 0.004784 ** 
City_CategoryC               7.217e-02  8.876e-03   8.131 4.24e-16 ***
Stay_In_Current_City_Years1  4.129e-02  1.067e-02   3.871 0.000108 ***
Stay_In_Current_City_Years2  1.453e-02  1.191e-02   1.220 0.222431    
Stay_In_Current_City_Years3 -1.372e-02  1.208e-02  -1.136 0.255954    
Stay_In_Current_City_Years4+ -5.265e-02  1.241e-02  -4.244 2.20e-05 ***
Product_Category_1          -1.419e-03  9.846e-04  -1.441 0.149550    
Product_Category_2          -2.084e-03  5.340e-04  -3.902 9.54e-05 ***
Product_Category_3          -6.492e-04  5.841e-04  -1.111 0.266412    
Purchase                    -2.705e-06  7.153e-07  -3.781 0.000156 ***
Age18-25                     1.527e+01  2.212e+01   0.690 0.489950    
Age26-35                     1.615e+01  2.212e+01   0.730 0.465247    
Age36-45                     1.616e+01  2.212e+01   0.731 0.465054    
Age46-50                     1.754e+01  2.212e+01   0.793 0.427844    
Age51-55                     1.752e+01  2.212e+01   0.792 0.428453    
Age55+                       1.713e+01  2.212e+01   0.774 0.438758    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 581885  on 430060  degrees of freedom
Residual deviance: 525553  on 430042  degrees of freedom
AIC: 525591

Number of Fisher Scoring iterations: 15

> 
```

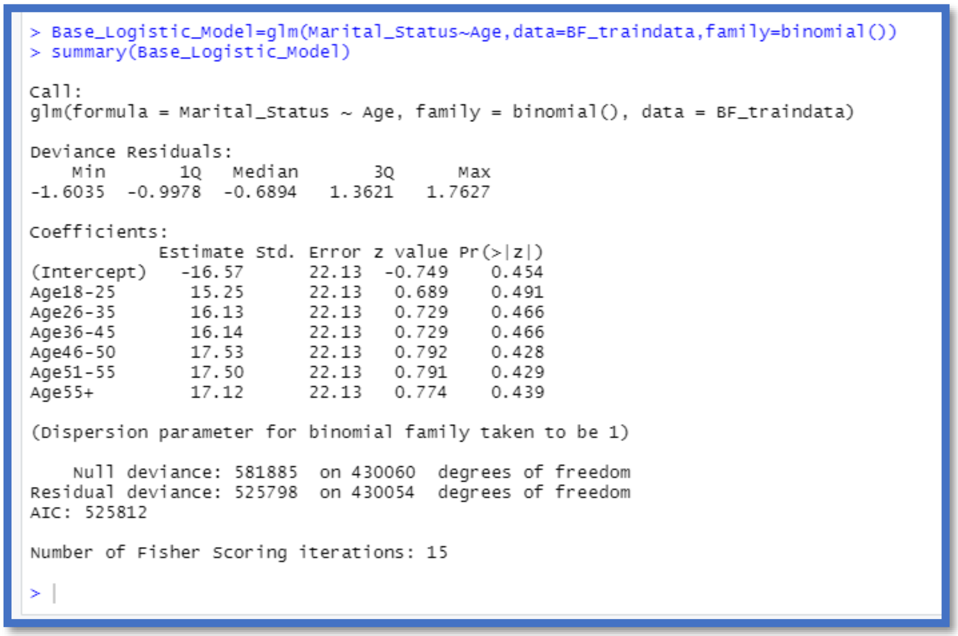
## BACKWARD ELIMINATION



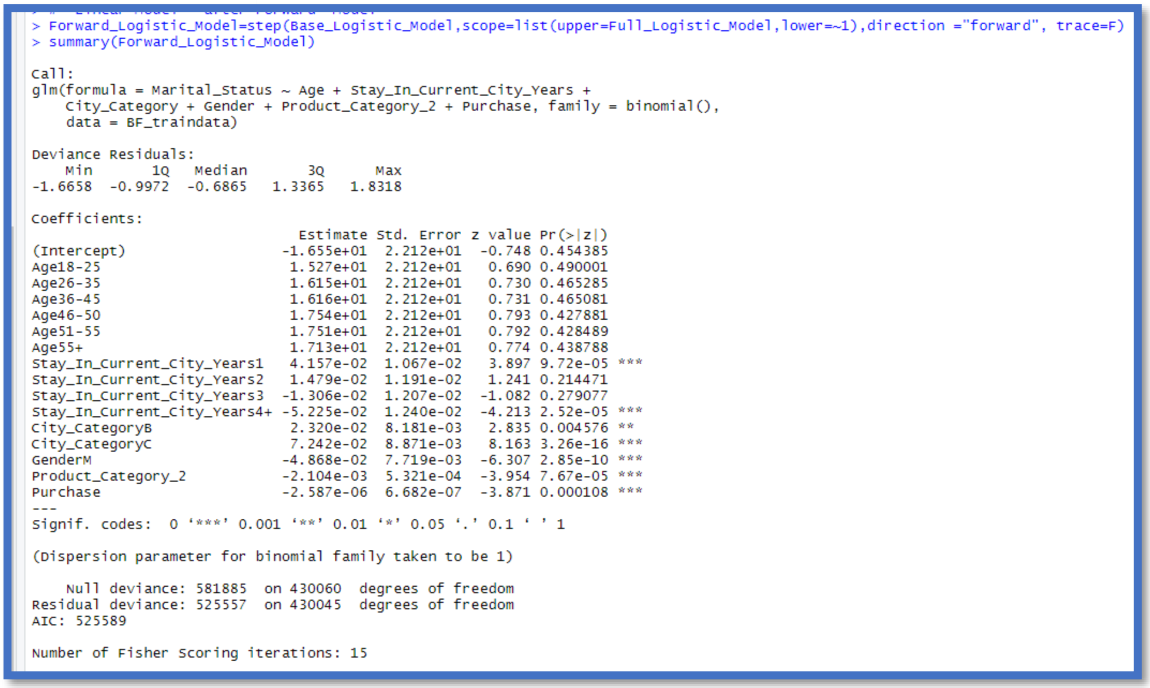
```
> #  Logistic Model - after Backward  Model
> Back_Logistic_Model=step(Full_Logistic_Model,direction ="backward", trace=F)
> summary(Back_Logistic_Model)

Call:
glm(formula = Marital_Status ~ Gender + City_Category + Stay_In_Current_City_Years +
    Product_Category_2 + Purchase + Age, family = binomial(),
    data = BF_traindata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6658  -0.9972  -0.6865   1.3365   1.8318

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.655e+01  2.212e+01  -0.748 0.454385
GenderM                     -4.868e-02  7.719e-03  -6.307 2.85e-10 ***
City_CategoryB               2.320e-02  8.181e-03   2.835 0.004576 **
City_CategoryC               7.242e-02  8.871e-03   8.163 3.26e-16 ***
Stay_In_Current_City_Years1  4.157e-02  1.067e-02   3.897 9.72e-05 ***
Stay_In_Current_City_Years2  1.479e-02  1.191e-02   1.241 0.214471
Stay_In_Current_City_Years3 -1.306e-02  1.207e-02  -1.082 0.279077
Stay_In_Current_City_Years4+ -5.225e-02  1.240e-02  -4.213 2.52e-05 ***
Product_Category_2          -2.104e-03  5.321e-04  -3.954 7.67e-05 ***
Purchase                    -2.587e-06  6.682e-07  -3.871 0.000108 ***
Age18-25                     1.527e+01  2.212e+01   0.690 0.490001
Age26-35                     1.615e+01  2.212e+01   0.730 0.465285
Age36-45                     1.616e+01  2.212e+01   0.731 0.465081
Age46-50                     1.754e+01  2.212e+01   0.793 0.427881
Age51-55                     1.751e+01  2.212e+01   0.792 0.428489
Age55+                       1.713e+01  2.212e+01   0.774 0.438788
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 581885  on 430060  degrees of freedom
Residual deviance: 525557  on 430045  degrees of freedom
AIC: 525589

Number of Fisher Scoring iterations: 15
```

## FORWARD SELECTION

## BASE MODEL

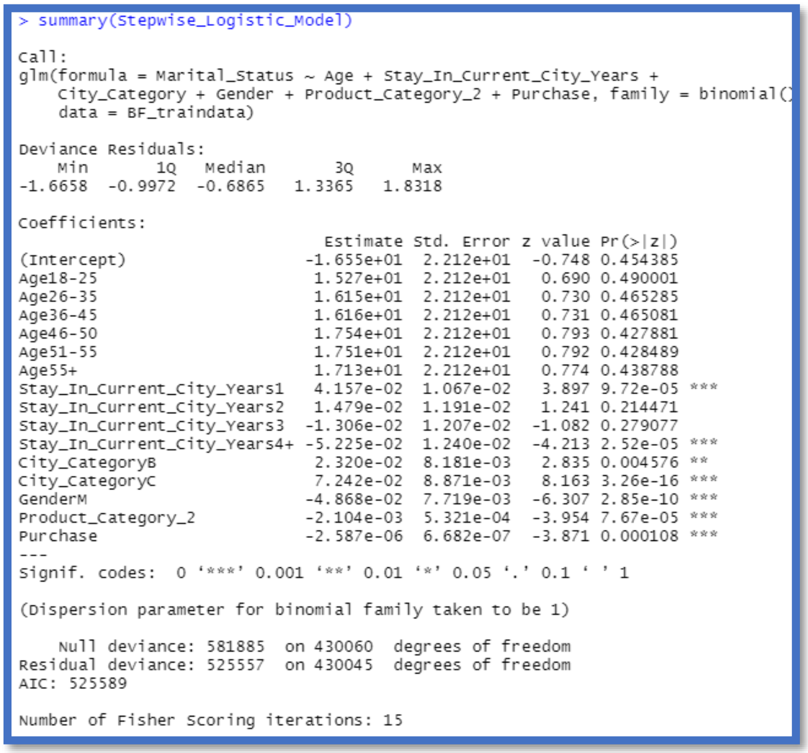


```
> Base_Logistic_Model=glm(Marital_Status~Age,data=BF_traindata,family=binomial())
> summary(Base_Logistic_Model)

Call:
glm(formula = Marital_Status ~ Age, family = binomial(), data = BF_traindata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6035  -0.9978  -0.6894   1.3621   1.7627

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -16.57      22.13  -0.749    0.454
Age18-25       15.25      22.13   0.689    0.491
Age26-35       16.13      22.13   0.729    0.466
Age36-45       16.14      22.13   0.729    0.466
Age46-50       17.53      22.13   0.792    0.428
Age51-55       17.50      22.13   0.791    0.429
Age55+         17.12      22.13   0.774    0.439

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 581885  on 430060  degrees of freedom
Residual deviance: 525798  on 430054  degrees of freedom
AIC: 525812

Number of Fisher Scoring iterations: 15

> |
```



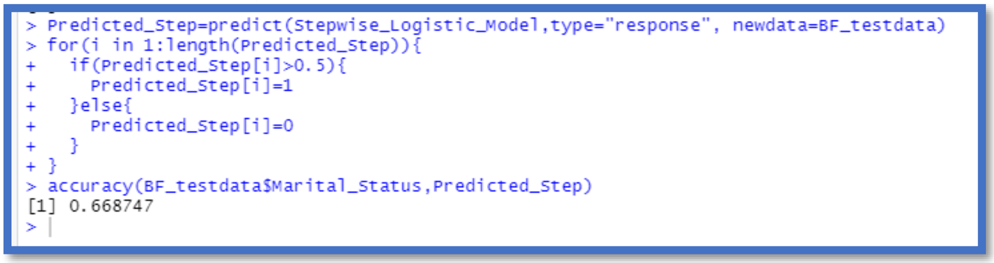
```
> Forward_Logistic_Model=step(Base_Logistic_Model,scope=list(upper=Full_Logistic_Model,lower=~1),direction ="forward", trace=F)
> summary(Forward_Logistic_Model)

Call:
glm(formula = Marital_Status ~ Age + Stay_In_Current_City_Years +
    City_Category + Gender + Product_Category_2 + Purchase, family = binomial(),
    data = BF_traindata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6658  -0.9972  -0.6865   1.3365   1.8318

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.655e+01  2.212e+01  -0.748 0.454385
Age18-25                     1.527e+01  2.212e+01   0.690 0.490001
Age26-35                     1.615e+01  2.212e+01   0.730 0.465285
Age36-45                     1.616e+01  2.212e+01   0.731 0.465081
Age46-50                     1.754e+01  2.212e+01   0.793 0.427881
Age51-55                     1.751e+01  2.212e+01   0.792 0.428489
Age55+                       1.713e+01  2.212e+01   0.774 0.438788
Stay_In_Current_City_Years1  4.157e-02  1.067e-02   3.897 9.72e-05 ***
Stay_In_Current_City_Years2  1.479e-02  1.191e-02   1.241 0.214471
Stay_In_Current_City_Years3 -1.306e-02  1.207e-02  -1.082 0.279077
Stay_In_Current_City_Years4+ -5.225e-02  1.240e-02  -4.213 2.52e-05 ***
City_CategoryB               2.320e-02  8.181e-03   2.835 0.004576 **
City_CategoryC               7.242e-02  8.871e-03   8.163 3.26e-16 ***
GenderM                     -4.868e-02  7.719e-03  -6.307 2.85e-10 ***
Product_Category_2          -2.104e-03  5.321e-04  -3.954 7.67e-05 ***
Purchase                    -2.587e-06  6.682e-07  -3.871 0.000108 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 581885  on 430060  degrees of freedom
Residual deviance: 525557  on 430045  degrees of freedom
AIC: 525589

Number of Fisher Scoring iterations: 15
```

STEPWISE SELECTION

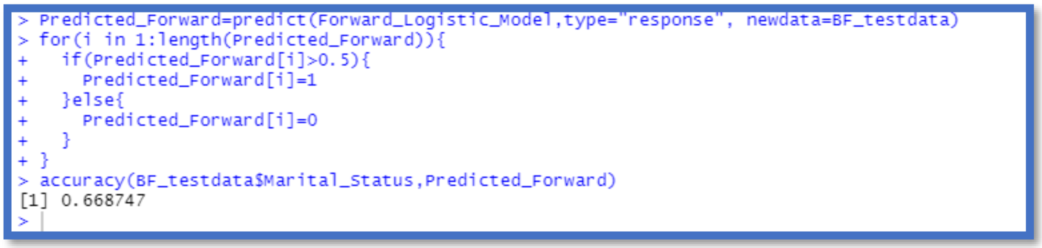


```
> summary(Stepwise_Logistic_Model)

Call:
glm(formula = Marital_Status ~ Age + Stay_In_Current_City_Years +
    City_Category + Gender + Product_Category_2 + Purchase, family = binomial(
    data = BF_traindata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6658  -0.9972  -0.6865   1.3365   1.8318

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.655e+01  2.212e+01  -0.748 0.454385
Age18-25                     1.527e+01  2.212e+01   0.690 0.490001
Age26-35                     1.615e+01  2.212e+01   0.730 0.465285
Age36-45                     1.616e+01  2.212e+01   0.731 0.465081
Age46-50                     1.754e+01  2.212e+01   0.793 0.427881
Age51-55                     1.751e+01  2.212e+01   0.792 0.428489
Age55+                       1.713e+01  2.212e+01   0.774 0.438788
Stay_In_Current_City_Years1  4.157e-02  1.067e-02   3.897 9.72e-05 ***
Stay_In_Current_City_Years2  1.479e-02  1.191e-02   1.241 0.214471
Stay_In_Current_City_Years3 -1.306e-02  1.207e-02  -1.082 0.279077
Stay_In_Current_City_Years4+ -5.225e-02  1.240e-02  -4.213 2.52e-05 ***
City_CategoryB               2.320e-02  8.181e-03   2.835 0.004576 **
City_CategoryC               7.242e-02  8.871e-03   8.163 3.26e-16 ***
GenderM                     -4.868e-02  7.719e-03  -6.307 2.85e-10 ***
Product_Category_2          -2.104e-03  5.321e-04  -3.954 7.67e-05 ***
Purchase                    -2.587e-06  6.682e-07  -3.871 0.000108 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 581885  on 430060  degrees of freedom
Residual deviance: 525557  on 430045  degrees of freedom
AIC: 525589

Number of Fisher Scoring iterations: 15
```

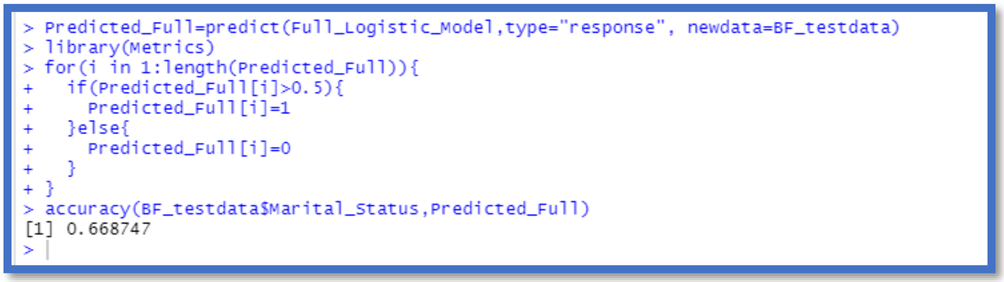
## ACCURACY



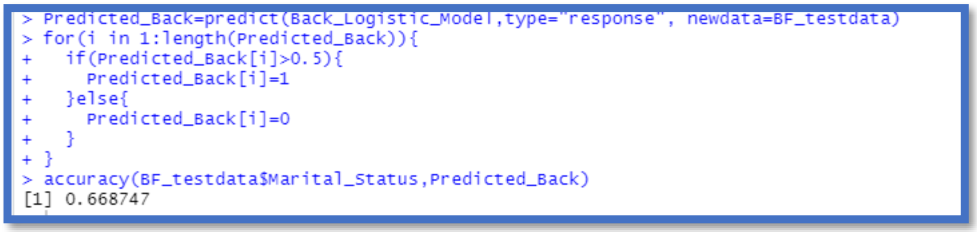
```
> Predicted_Step=predict(Stepwise_Logistic_Model,type="response", newdata=BF_testdata)
> for(i in 1:length(Predicted_Step)){
+   if(Predicted_Step[i]>0.5){
+     Predicted_Step[i]=1
+   }else{
+     Predicted_Step[i]=0
+   }
+ }
> accuracy(BF_testdata$Marital_Status,Predicted_Step)
[1] 0.668747
> 
```



```
> Predicted_Forward=predict(Forward_Logistic_Model,type="response", newdata=BF_testdata)
> for(i in 1:length(Predicted_Forward)){
+   if(Predicted_Forward[i]>0.5){
+     Predicted_Forward[i]=1
+   }else{
+     Predicted_Forward[i]=0
+   }
+ }
> accuracy(BF_testdata$Marital_Status,Predicted_Forward)
[1] 0.668747
> 
```



```
> Predicted_Full=predict(Full_Logistic_Model,type="response", newdata=BF_testdata)
> library(Metrics)
> for(i in 1:length(Predicted_Full)){
+   if(Predicted_Full[i]>0.5){
+     Predicted_Full[i]=1
+   }else{
+     Predicted_Full[i]=0
+   }
+ }
> accuracy(BF_testdata$Marital_Status,Predicted_Full)
[1] 0.668747
> 
```



```
> Predicted_Back=predict(Back_Logistic_Model,type="response", newdata=BF_testdata)
> for(i in 1:length(Predicted_Back)){
+   if(Predicted_Back[i]>0.5){
+     Predicted_Back[i]=1
+   }else{
+     Predicted_Back[i]=0
+   }
+ }
> accuracy(BF_testdata$Marital_Status,Predicted_Back)
[1] 0.668747
```