IIT School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

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

Table of Contents

1. Introduction

Black Friday is an informal name for the Friday following Thanksgiving Day in all the States in USA. Usually it's been celebrated on the fourth Thursday of November. The day after Thanksgiving has been regarded as the beginning of America's Christmas Shopping season. It has routinely been the busiest shopping day of the year in the United States.

Black Friday Sales relies on a few simple retail strategies that, with tons of customer data and forecasting software, have become precise and planning for Black Friday is key for many retailers, particularly in predicting consumer interest in product ranges, which many retailers got wrong last year. It must be carefully planned for every year to ensure orders can be fulfilled without compromising on the level of customer service and seamless delivery.

The purpose of this project is to find out the what are the reasons or factors influencing the sales during Black Friday sales and design models which will help the retailers to understand what changes are required to achieve maximum profit and better promotions.

2. Data

The dataset has sample of the transactions made in a retail store. The store wants to know better the customer purchase behavior against different products where we are trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables.

To work on this project, we have chosen the dataset provided by Mehdi Dagdoug to predict Black Friday sales based on various parameters such as Product category, Customer age, gender and location etc.

There are more than half a million (550 000) records available to train the models which we would be using to test and predict the sales with a 95% confidence level. The attribute details are as follows.

Attribute Name	Description	Attribute Data Type	
User_ID *	ID assigned to the customer	Quantitative	Discrete
Product_ID *	ID assigned to the product	Qualitative	Nominal
Gender	Gender of the Customer	Qualitative	Binary
Age	Age group to which of the customer	Qualitative	Nominal
Occupation	Conveys how long the customer has been working	Quantitative	Discrete
City_Category	Category of city where the retail store Resides	Qualitative	Nominal
Stay_In_Current_City_Years	Conveys how long the customer resides in current city	Quantitative	Discrete
Marital_Status	Conveys whether the customer is married or not	Qualitative	Binary
Product_Category_1	Quantity of products bought in category 1 by a customer	Quantitative	Discrete
Product_Category_2	Quantity of products bought in category 2 by a customer	Quantitative	Discrete
Product_Category_3	Quantity of products bought in category 3 by a customer	Quantitative	Discrete
Purchase	Total cost of expenditure of a customer during black Friday	Quantitative	Discrete

***** excluded in our analysis as it based on generic features and not on case specific variables

This dataset has been retrieved from the link: https://www.kaggle.com/mehdidag/black-friday

3. Problem to be Solved

- To analyze, learn the customer behavior on Black Friday sales and build regression models to predict the sales.
- Research on the purchases based on Gender of customers.
- Research on sales among various product categories and its quantity.
- Predict the age group of customers based on sales record.
- Predict the Marital status of customers.
- Predict customer's location.

4. Solution

To address the above problems, we wish to work towards achievement of following solutions:

- Using Linear Regression technique to learn the customer behavior and confirm if all the predictors (categorical and numerical features) have a significant effect on the Purchase on Black Friday sales.
- Comparing the purchases made by the customers during black Friday sales based on Gender using hypothesis testing – One sample
- Comparing the number of products bought by customers from various product categories during black Friday sales using hypothesis testing – Two sample
- Predict the age group of customers based on sales record using Naive Bayes Classification Technique.
- Predict the Marital status of customers using Logistic Regression Technique.
- Predict customer's location (city category) using K- Nearest Neighbor Classification Technique.

5. Experiments and Results

5.1 Methods and Process

5.1.1 Hypothesis Testing

I) One sample – One Tailed Hypothesis Testing

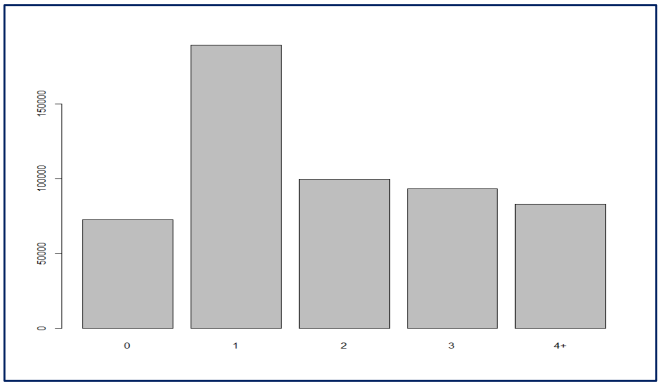
HYPOTHESIS

Null Hypothesis: H0: Average stay in current city is equal to **2.86** ➔ μ=**2.86**

Alternate Hypothesis: Hα: Average stay in current city is greater than **2.86**➔ μ> **2.86**

Confidence level = 95% = 0.95

Level of Significance ɑ = 1- Confidence level ➔ ɑ = 1- 0.95 = 0.05



Bar Graph – Stay in the current city (In years)

```
> z.test(Stay,alternative="greater",mu=2.86,sigma.x=sd(Stay),conf.level=0.95)

        One-sample z-Test

data:  Stay
z = -0.30794, p-value = 0.6209
alternative hypothesis: true mean is greater than 2.86
95 percent confidence interval:
 2.856565       NA
sample estimates:
mean of x
 2.859458

> 
```

Z test for hypothesis

INTERPRETATION

- ✓ P-value implies area under normal curve based on test statistics. As P-value (0.6209) > α (0.5), we don't have enough evidence to reject NULL Hypothesis (H0) with 95% confidence level.
- ✓ With 95% confidence we can conclude that Average stay of customers in current city is 2.86 yrs.

II) One sample – Two Tailed Hypothesis Testing

HYPOTHESIS

Null Hypothesis: H0: Average purchases made by Male and Female are equal ➔ $\mu_f = \mu_m$

Alternate Hypothesis: Hɑ: Average purchases made by Male and Female are not equal ➔ $\mu_f \neq \mu_m$

Confidence level = 95% = 0.95

Level of Significance ɑ = 1- Confidence level = 1- 0.95 = 0.05

```
> # 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

>
```

One – sample Z-Test (Two tailed)

INTERPRETATION

- ✓ As P-value ($2.2e^{-16}$) < α (0.05), we don't have enough evidence to accept NULL Hypothesis (H0) with 95% confidence level.
- ✓ With 95% confidence level, we can conclude that Average purchases made by Male and Female are not equal.

III) Two Sampled Hypothesis Testing

HYPOTHESIS

Null Hypothesis: H0: Average No. of products bought from category 1 and category 2 are equal

Alternate Hypothesis: Hα: Average No. of products bought from category 1 and category 2 are not equal.

Confidence level = 95% = 0.95

Level of Significance α = 1- Confidence level = 1- 0.95 = 0.05

```
> 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)
>
```

We tried Z test by passing Paired value as FALSE, but R showed error hence we proceeded by removing paired parameter in Z test.

```
> 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

>
```

INTERPRETATION

- ✓ As P-value < α, we don't have enough evidence to accept NULL Hypothesis (H0) with 95% confidence level
- ✓ Average quantity of purchase made on Product category 1 and category 2 are not equal.

Similarly,

- ✓ Average quantity of purchase made on Product category 2 and category 3 are not equal.

- ✓ Average quantity of purchase made on Product category 1 and category 3 are not equal.

5.1.2 ANOVA

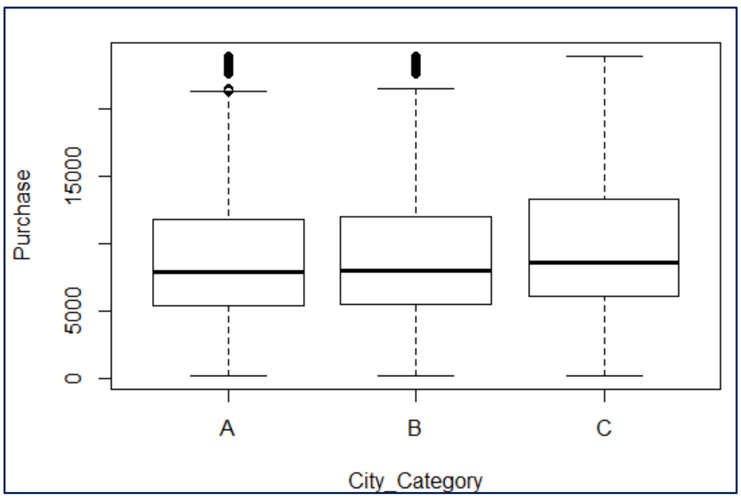
OBJECTIVE : Compare group means among more than two groups by analyzing the variances.

I) Comparing group means of City category

HYPOTHESIS

Null Hypothesis: H0: Average purchases across all city categories are equal ➔ μA = μB = μC

Alternate Hypothesis: Hα: Average purchases across all city categories are not equal ➔ Not all μ's is equal



FURTHER ANALYSIS

1. F- test

   Null Hypothesis: No X variable is significant in predicting Y
   Alternate Hypothesis: At least one X variable is significant in predicting Y

```
> 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

> 
```

INTERPRETATION

- ✓ As P-value ($2.2e^{-16}$) < α (0.05), we don't have enough evidence to accept Null hypothesis.
- ✓ From the F-test results we can conclude that "With 95% confidence level atleast one Independent feature has significant effect in predicting Purchase".

2. Individual parameter test

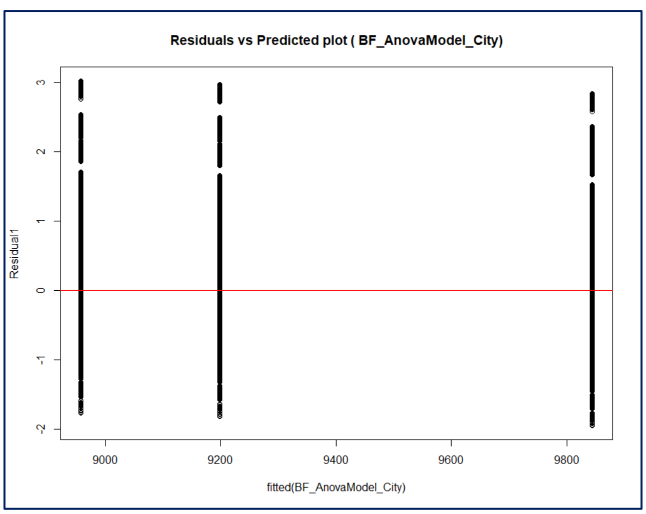
   Null Hypothesis: X variable is not significant in predicting Y
   Alternate Hypothesis: X variable is significant in predicting Y

INTERPRETATION

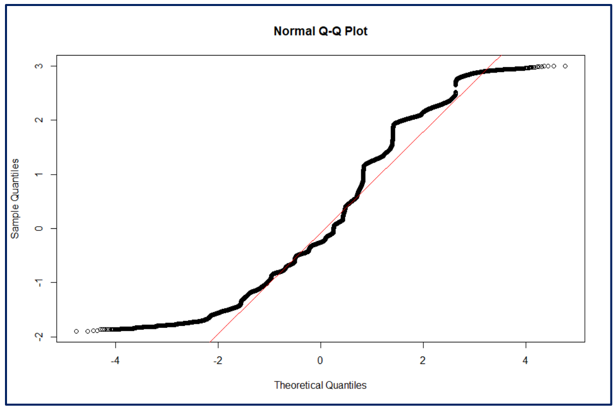
- ✓ For all the X variables, P value < α (0.05), we don't have enough evidence to accept Null hypothesis.
- ✓ From the Individual parameter test result we can conclude that "With 95% confidence level all the Independent feature has significant effect in predicting Purchase".

3. Residual Analysis

   - Constance Variance



- 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
> 
```

EQUATION

$\bar{Y}$(City) = 8958.01 + 240.65*(City Cat B) + 886.43*(City Cat C)

INTERPRETATION

✓ The F-test statistic is F = 1377 with p-value 2.2e-16 (< 0.05).

✓ As P-value < α, we don't have enough evidence to accept NULL Hypothesis (H0) with 95% confidence level.

✓ With 95% confidence level we can conclude that Average purchases across all city categories are not equal.

II) Comparing group means of Age Group

HYPOTHESIS

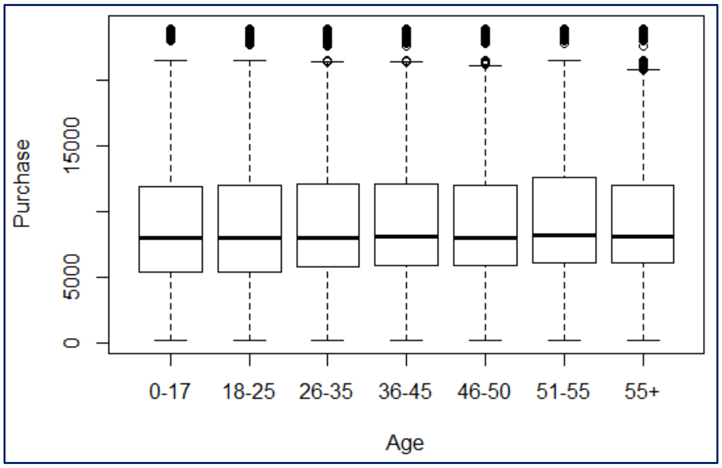
Null Hypothesis: H0: Average purchases made over different age groups are equal

$\mu_{0\text{-}17} = \mu_{18\text{-}25} = \mu_{26\text{-}35} = \mu_{36\text{-}45} = \mu_{46\text{-}50} = \mu_{51\text{-}55} = \mu_{55+}$

OR There is no difference in means ➔ β 1 = β 2 = β3 = β 4 = β 5 = β 6 = 0

Alternate Hypothesis: Hα: Average purchases made over different age groups are not equal ➔ Not all μ's is equal

OR There is some difference in means βi ≠ 0



FURTHER ANALYSIS

1. F- test

   Null Hypothesis: No X variable is significant in predicting Y

   Alternate Hypothesis: At least one X variable is significant in predicting Y

```
> 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

> |
```

INTERPRETATION

- ✓ As P-value ($2.2e^{-16}$) < α (0.05), we don’t have enough evidence to accept Null hypothesis.
- ✓ From the F-test results we can conclude that “With 95% confidence level at least one Independent feature has significant effect in predicting Purchase”.

2. Individual parameter test

   Null Hypothesis: X variable is not significant in predicting Y

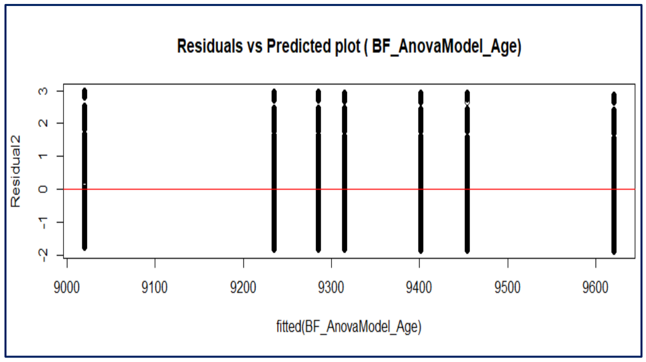
   Alternate Hypothesis: X variable is significant in predicting Y

INTERPRETATION

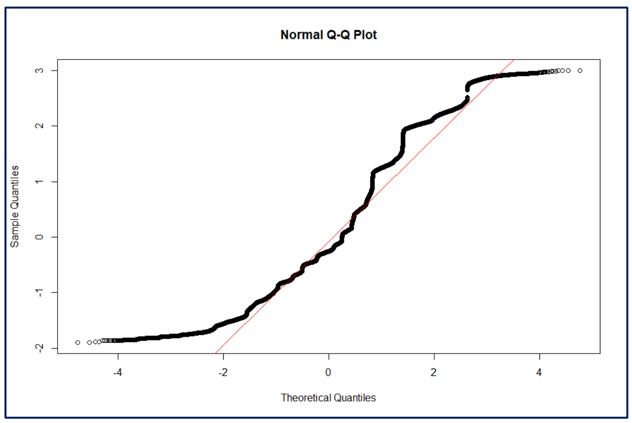
- ✓ For all the X variables, P value < α (0.05), we don’t have enough evidence to accept Null hypothesis.
- ✓ From the Individual parameter test results, we can conclude that “With 95% confidence level all the Independent feature has significant effect in predicting Purchase”.

3. Residual Analysis

   a. Constance Variance



   b. 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
>
```

EQUATION

$\bar{Y}$(Age) = 9020.13 + 215.07*(Age 18-25) + 294.46*(Age 26-35)+ 381.35*(Age 36-45)+ 264.75*(Age 41-50)+ 600.49*(Age 51-55) + 433.77 * (Age 55+)

INTERPRETATION

- ✓ The F-test statistic is F = 43.49 with p-value 2.2e-16 (< 0.05)
- ✓ As P-value < α, we don't have enough evidence to accept NULL Hypothesis (H0) with 95% confidence level
- ✓ With 95 % confidence level, we can conclude that average purchases across all age groups are not equal.

5.1.3 Linear Regression technique

OBJECTIVE: To Predict purchases on Black Friday Sales

Step 1: Loading Data in R

Dataset stored in CSV format has been loaded in R using read.table function

Step 2: Identify Dependent and Independent variables

To predict the Numerical dependent variable (Purchase) based on values of Independent features (Gender, Age, Occupation, Stay in Current city, Marital Status, City category, Product_Category_1, Product_Category_2, and Product_Category_3) we planned to build models using linear regression techniques and improvise the same using feature selection.

Step 3: Correlation between the variables

```
> # 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
>
```

Step 4: Data Pre-processing

a) Replace Missing values

We have some missing values in Product category 2 and 3. Hence, replaced the NULL values with '0' as our dataset represents NULL in these features if customer didn't purchase any products of the respective category.

Step 5: Data Split

We have two ways of splitting data 1) Hold-Out Evaluation and 2) N-Fold cross validation. We have used Hold out evaluation Techniques to split my data as data size is large. We have used 80 % of total rows to train our model and 20 % of total rows to test our model.

Step 6: Build Models

i) Building full model without Transforming any features

```
> 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
```

ii) Building full model after Transforming some of the X variables

```
> Full_Model_XTrans=lm(formula=Purchase~Gender+Age+Occupation+City_Category+Stay_In_Current_City_Years+Marital_Status+Product_C
ategory_1+Product_Category_2+Product_Category_3,data = traindata)
> summary(Full_Model_XTrans)

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
```

Step 7: Feature Selection

i) 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
```

Backward Elimination for X Trans Model

ii) Forward Selection

```
> #  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
```

Forward Selection for X Trans Model

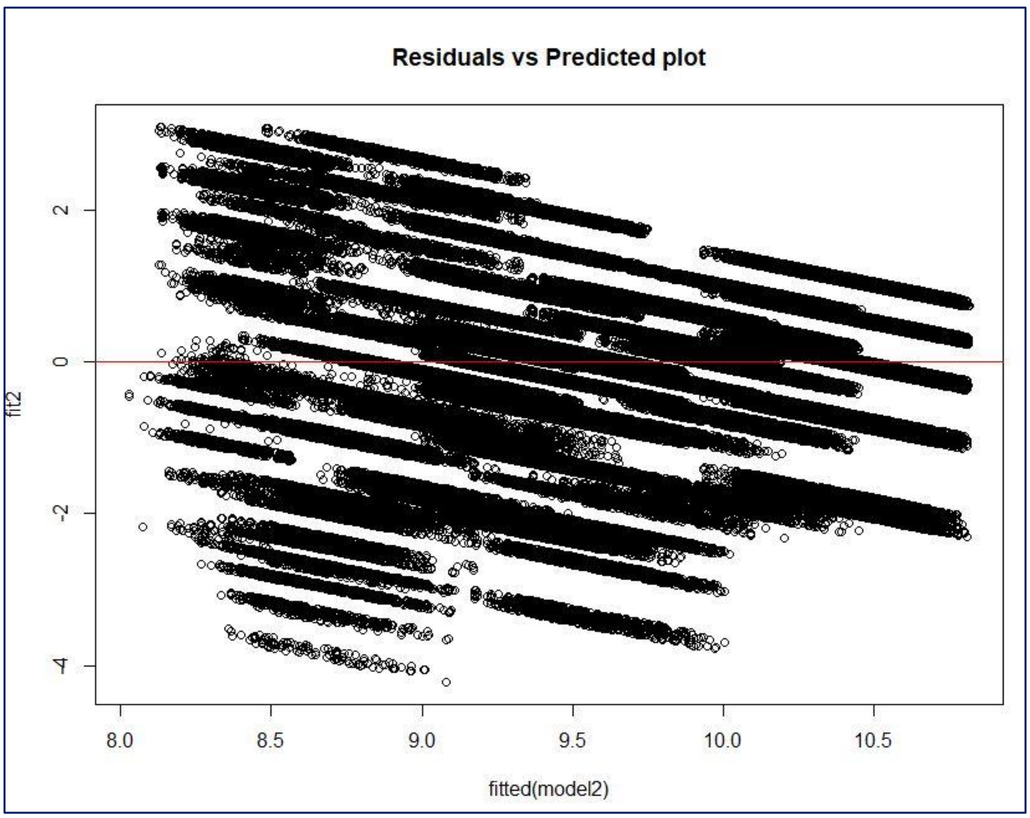
iii) Stepwise Selection

```
> #  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
```

Stepwise Selection for X Trans Model

Step 8: Residual Analysis

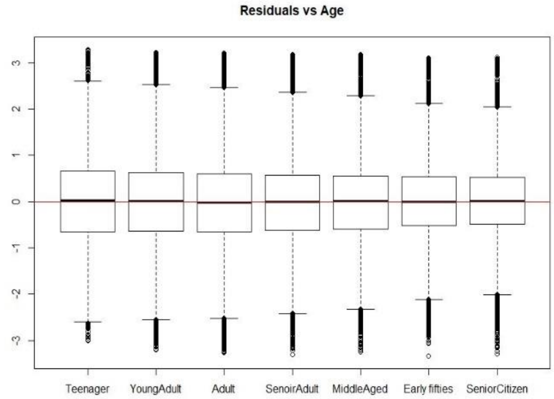
a) Residuals vs Predicted

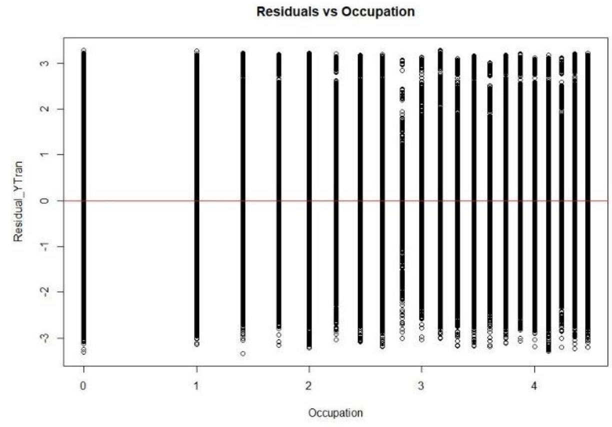


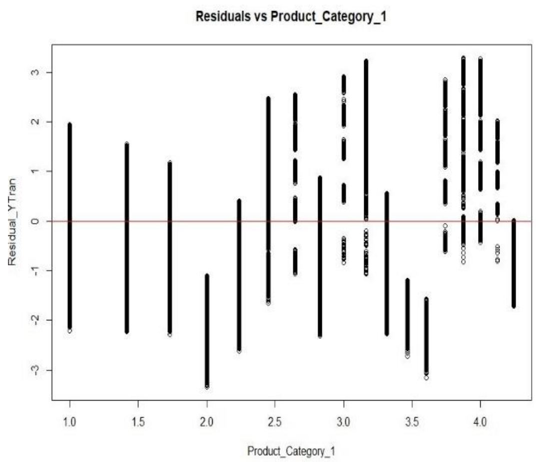
Interpretation

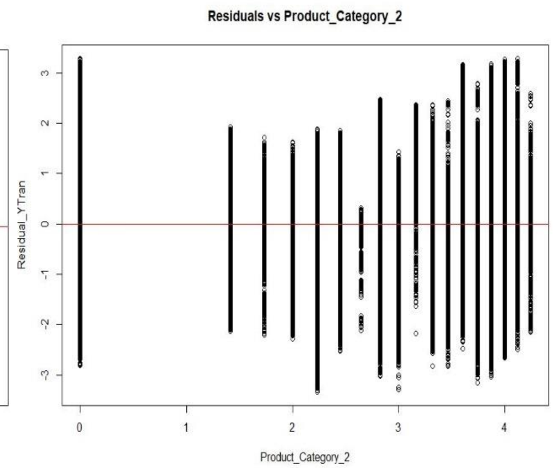
From the residual plot we can understand that there is Constance and Variance in model. Hence no transformation required for Y variables.

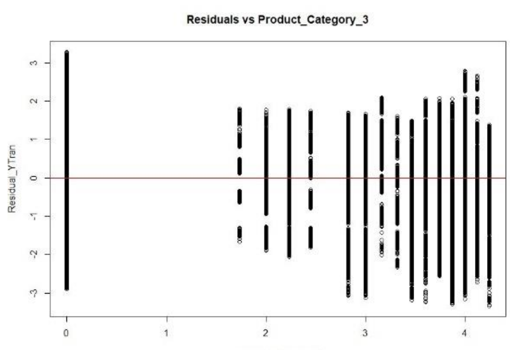
b) Residual vs Individual X variables

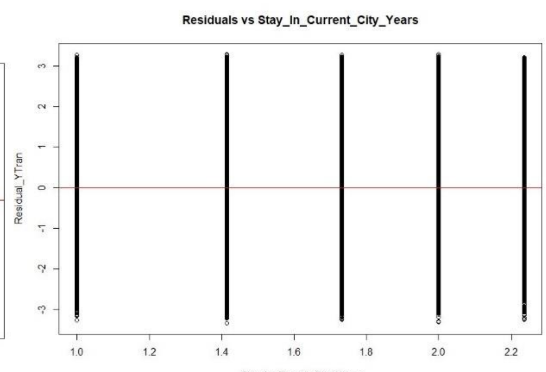






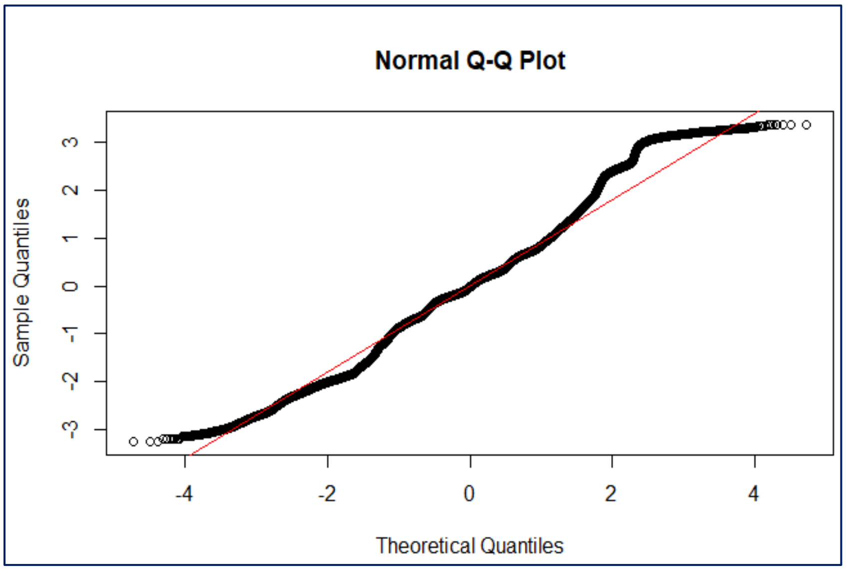






c) Normality Check

```
> 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

> |
```



Step 9: Calculating RMSE

Model	ADJ –R2	RMSE
Full Model without Transformation	0.1358	4636.44
Full Model with X Transformation	0.1358	4644.651

INTERPRETATION

By building various model and analyzing the Adj-R2 and RMSE values we can conclude that Feature selection has not improved the model much. Similarly, there is no significant improvement in AdjR2 and RMSE by transforming X variables.

So, our best model was able to explain 13.58% transformation of Purchases using X variables with RMSE of 4636.

5.1.4 K- Nearest Neighbor Classification Technique

K-NN algorithm is one of the simplest classification algorithms and it is used to identify the data points that are separated into several classes to predict the classification of a new sample point. K-NN is a non-parametric, lazy learning algorithm. It classifies new cases based on a similarity measure (e.g. distance functions).

OBJECTIVE: To predict city category

Step 1: Loading Data in R

Dataset stored in CSV format has been loaded in R using read.table function

Step 2: Deciding Dependent and Independent Variables

To forecast or deduce the categorical Multi Class dependent variable (CITY Category) based on values of Independent features (Gender, Age, Occupation, Stay in Current city, Marital Status, Product_Category_1, Product_Category_2, and Product_Category_3) we planned to build 7 models for KNN classification with different K values.

Step 3: Pre-processing Data

Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues

a. Replace Missing values

We have some missing values in Product category 2 and 3. Hence, replaced the NULL values with '0' as our dataset represents NULL in these features if customer didn't purchase any products of the respective category.

b. Data Transformation - Convert Categorical values to Numerical values

For KNN our dependent variable should be categorical and independent variable should be numeric. Most of the independent features are categorical hence, it must be converted to dummy or representation variables. We used dummy.data.frame function in R to transform data.

c. Normalize the values

Feature normalization is used to convert values in a feature to the same or similar scales with values in other features. In KNN all independent features should be numeric, and it should be of same scale range. So, we normalize only the independent variables as dependent variable is categorical in KNN. We used lapply function in R to transform data.

Step 4: Data split

We have two ways of splitting data 1) Hold-Out Evaluation and 2) N-Fold cross validation. We have used Hold out evaluation Techniques to split my data as data size is large. We have used 30 % of total rows to train our model and 70 % of total rows to test our model.

Step 5: Deciding K Values

We have decided to use K values from 1,3,5,101,299,399 and 499 for my data set to check the accuracy.

Step 6: Building Model

```
> 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
```

Step 7: Finding Accuracy

Model	Accuracy
K =1	0.3736128
K = 3	0.376148
K = 5	0.3794406
K = 101	0.4138542
K = 299	0.4215395
K = 399	0.4220976
K = 499	0.422156

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

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

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

```
>
> 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
```

```
> accuracy(testdef,knn_model399)
[1] 0.4220976
> knn_model299=knn(trainset,testset,traindef,k=299)
> accuracy(testdef,knn_model299)
[1] 0.4215395
```

```
> library(Metrics)
> accuracy(testdef,knn_model499)
[1] 0.422156
> |
```

Check for overfitting problem

We don't have any overfitting problem in KNN, as we don't have any learning process.

INTERPRETATION

At K=499, maximum accuracy achieved is 42%.

5.1.5 Naive Bayes Classification Technique

OBJECTIVE: To predict Age group of customers

Step 1: Loading Data in R

Dataset stored in CSV format has been loaded in R using read.table function.

Step 2: Deciding Dependent and Independent Variables

We planned to build a model to forecast or deduce the Multi Class variables Categorical dependent variable (AGE Category) based on values of Independent features (Gender, City category, Occupation, Stay in Current city, Marital Status, Product_Category_1, Product_Category_2, and Product_Category_3).

Step 3: Data preprocessing

a) Replace Missing values

We have some missing values in Product category 2 and 3. Hence, replaced the NULL values with '0' as our dataset represents NULL in these features if customer didn't purchase any products of the respective category.

b) Data Transformation

For Naïve Bayes our dependent & independent variable should be categorical. Most of the independent features are categorical but some are represented as numerical data in data set. Hence transformation from numerical to categorical required for variables in our data set as part of Naïve Bayes classification. And categorical details are already mentioned in Attribute Information. We can use cut function in R to transform data.

c) Imbalance Issue

It is the problem in data set where the total number of a class of data is far less than the total number of another class of data. We don't have imbalance issue in our data set and all the classes are distributed fairly.

Step 4: Data Split

We have two ways of splitting data 1) Hold-Out Evaluation and 2) N-Fold cross validation. We have used Hold out evaluation Techniques to split my data as data size is large. We have used 80 % of total rows to train our model and 20 % of total rows to test our model.

Step 5: Building Models

```
> 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)
```

Step 6: Finding Accuracy

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

Step 7: Check for Overfitting Problem

We don't have any overfitting problem in Naïve Bayes, as we don't have any learning process.

INTERPRETATION:

- ✓ Maximum accuracy achieved is 31.2 % by using conditional probability.

5.1.6 Logistic Regression Technique

Logistic regression is kind of like linear regression but is used when the dependent variable is not a number, but something else (like a Yes/No response). It's called Regression but performs classification as based on the regression it classifies the dependent variable into either of the classes.

Objective: To predict the martial status of customers

Step 1: Loading Data in R

Dataset stored in CSV format has been loaded in R using read.table function

Step 2: Deciding Dependent and Independent Variables

To forecast or deduce the categorical Binary Class dependent variable (Marital Status) based on values of Independent features (Gender, Age, Occupation, Stay in Current city, City category, Product_Category_1, Product_Category_2, and Product_Category_3) we planned implement logistic regression to build models and improvise the same using feature selection techniques.

Step 3: Pre-processing Data

Replace Missing values

We have some missing values in Product category 2 and 3. Hence, replaced the NULL values with '0' as our dataset represents NULL in these features if customer didn't purchase any products of the respective category.

Step 4: Splitting of Data

We have two ways of splitting data 1) Hold-Out Evaluation and 2) N-Fold cross validation. We have used Hold out evaluation Techniques to split my data as data size is large. We have used 80 % of total rows to train our model and 20 % of total rows to test our model.

Step 5: Building 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

> 
```

Step 6: Feature Selection

a) 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
```

b) Forward Selection

```
> 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
```

c) 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
```

Step 7: Finding accuracy

```
> 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_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_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
```

```
> 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
>
```

Model	**AIC**	**Accuracy**
Full Model	525591	0.668747
Backward Elimination	525589	0.668747
Forward Selection	525589	0.668747
Stepwise Selection	525589	0.668747

5.2 Evaluations and Results

Model	Prediction Attribute	Accuracy/ RMSE
Linear Regression	Purchase	4636.44
K – Nearest Neighbor	City Category	42%
Naïve Bayes	Age group	31.2%
Logistic Regression	Marital Status	66.9%

5.3 Findings

⇒ By using Linear regression technique to predict customer purchase we achieve maximum adjusted R2 of 13.58%.

⇒ By using KNN technique to predict city category we got maximum accuracy of 42% when K =499.

⇒ Utilizing Naïve Bayes technique to predict customer's Age group we achieved 31.2% of accuracy at maximum.

⇒ We achieved 66.9% accuracy while using Logistic Technique to predict marital status of customer.

6. Conclusion

Thus, these models will help retailers to determine the sales during black Friday, Age group of the customers, their marital status of the customers and the City where the customer resides. Hence, they can use this information to achieve maximum profit, promote their products across different kind of customers.

7. Limitation

Due to huge data set we are not able

- To find multi collinearity between Independent feature using VIF
- To find influential factors for our models.

8. Future Work

- Implement few more classification methods like decision trees, Random Forest which may give better results.
- Use hypothesis testing to determine the best model among the models created in logistic regression