A study of sales through consumer behaviors on Black Friday

Group ID : 242
Anusha Satish – A20401921
Sivaranjani Prabasankar – A20436206

CONTENT

ILLINOIS INSTITUTE OF TECHNOLOGY

- Introduction
- Scope
- Hypothesis Test
  - One Sample Hypothesis Test
  - Two Sample Hypothesis Test
  - ANOVA
- Regression Models
  - Linear Regression
- Classification Models
  - K – Nearest Neighbor
  - Naïve Bayes
  - Logistic Regression
- Purpose
- Conclusion & Future Work

INTRODUCTION

- Black Friday is an informal name for the Friday following Thanksgiving Day in all the States in USA. Usually its 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.

- Data Source https://www.kaggle.com/mehdidag/black-friday

SCOPE

- 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.

ATTRIBUTE DETAILS

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 has been residing 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 product of category 1 by a customer	Quantitative	Discrete
Product_Category_2	Quantity of products bought in product of category 2 by a customer	Quantitative	Discrete
Product_Category_3	Quantity of products bought in product of category 3 by a customer	Quantitative	Discrete
Purchase	Total cost of expenditure of a customer during black Friday sales	Quantitative	Discrete

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

DESCRIPTIVE STATISTICS

ILLINOIS INSTITUTE OF TECHNOLOGY

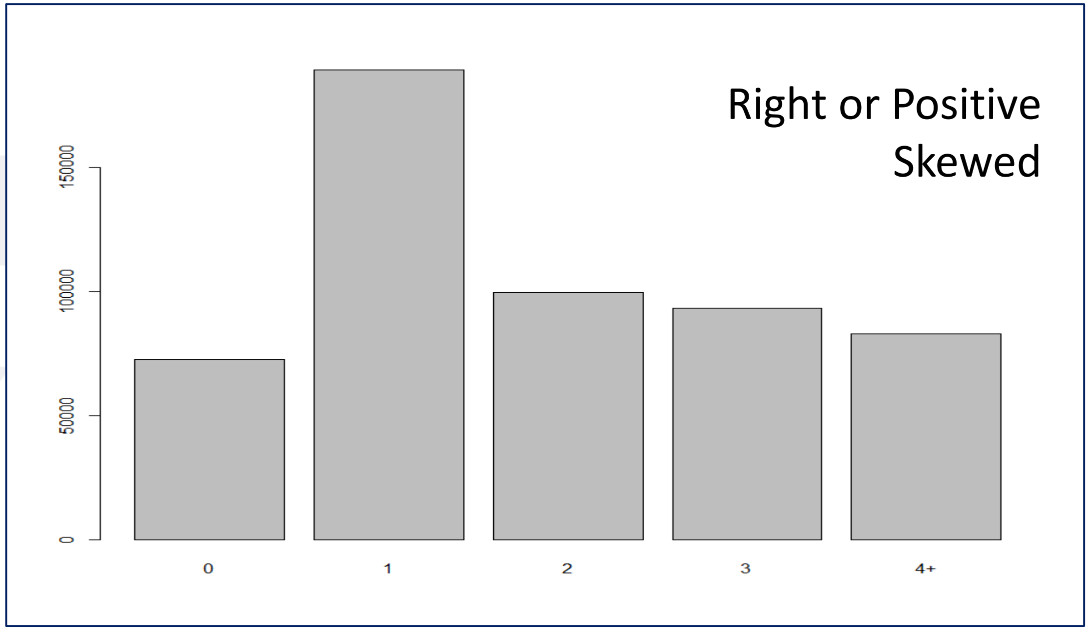
```
> describe(Blackfriday_Data)
                           vars      n       mean      sd  median   trimmed     mad     min     max range  skew kurtosis   se
User_ID                       1 537577 1002991.85 1714.39 1003031 1002983.60 2145.32 1000001 1006040  6039  0.02    -1.18 2.34
Product_ID*                   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
> 
```

ONE SAMPLE HYPOTHESIS TEST

ILLINOIS INSTITUTE OF TECHNOLOGY

HYPOTHESIS

- H0: Average stay in current city is equal to 3 ➔ $\mu=3$
- Hα: Average stay in current city is greater than 3➔ $\mu>3$



CALCULATION

Confidence level = 95% = 0.95

Level of Significance α = 1- Confidence level

α = 1- 0.95 = 0.05

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

        One-sample z-Test

data:  Stay
z = -79.89, p-value = 1
alternative hypothesis: true mean is greater than 3
95 percent confidence interval:
 2.856565       NA
sample estimates:
mean of x
 2.859458

>
```

INTERPRETATION

- ONE TAILED hypothesis test
- P-value implies area under normal curve based on test statistics
- As P-value > α, we don't have enough evidence to reject NULL Hypothesis (H0) with 95% confidence level

ONE SAMPLE HYPOTHESIS TEST

ILLINOIS INSTITUTE OF TECHNOLOGY

HYPOTHESIS

- H0: Average purchases made by Male and Female are equal

  $\mu_f = \mu_m$

- Hα: Average purchases made by Male and Female are not equal

  $\mu_f \neq \mu_m$

```
> # 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

>
```

INTERPRETATION

- TWO TAILED hypothesis test
- As P-value < α, we don't have enough evidence to accept NULL Hypothesis (H0) with 95% confidence level
- Average purchases made by Male and Female are not equal with 95% confidence level.

TWO SAMPLE HYPOTHESIS TEST

HYPOTHESIS

- H0: Average No. of products bought from category 1 and category 2 are equal .
- Hα: Average No. of products bought from category 1 and category 2 are not equal.

PRE-PROCESSING

✓ Ignore missing values

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

```
> # 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
```

```
> 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
```

ANOVA

ILLINOIS INSTITUTE OF TECHNOLOGY

OBJECTIVE

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

HYPOTHESIS

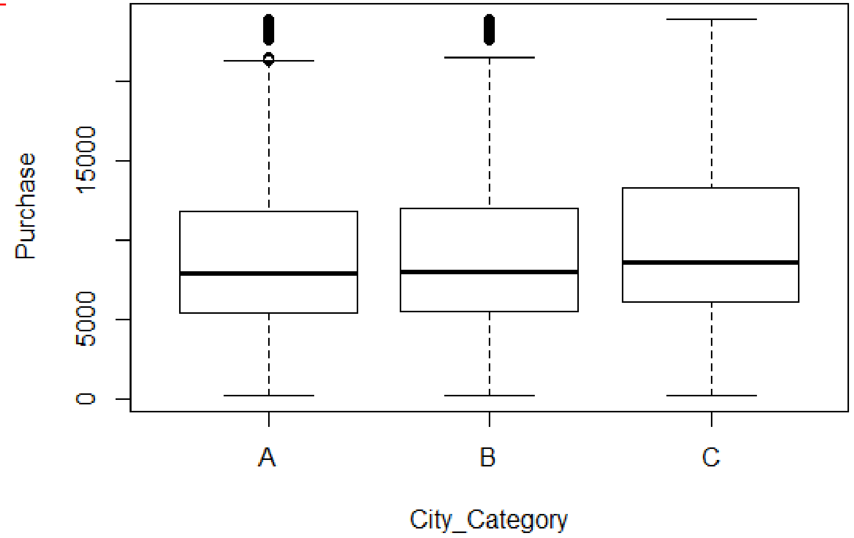
- H0: Average purchases across all city categories are equal ➔ $\mu_A = \mu_B = \mu_C$
- Hα: Average purchases across all city categories are not equal ➔ Not all μ's are equal

FURTHER ANALYSIS

1. F- test
2. Individual parameter test
3. Co- efficient of determination
4. Residual Analysis
    - Constance Variance
    - Normality Test

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.
- ➢ Average purchases are not equal among all city categories.




```
> 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

> |
```

ANOVA

ILLINOIS INSTITUTE OF TECHNOLOGY

Hypothesis

H0: Average purchases made over different age groups are equal

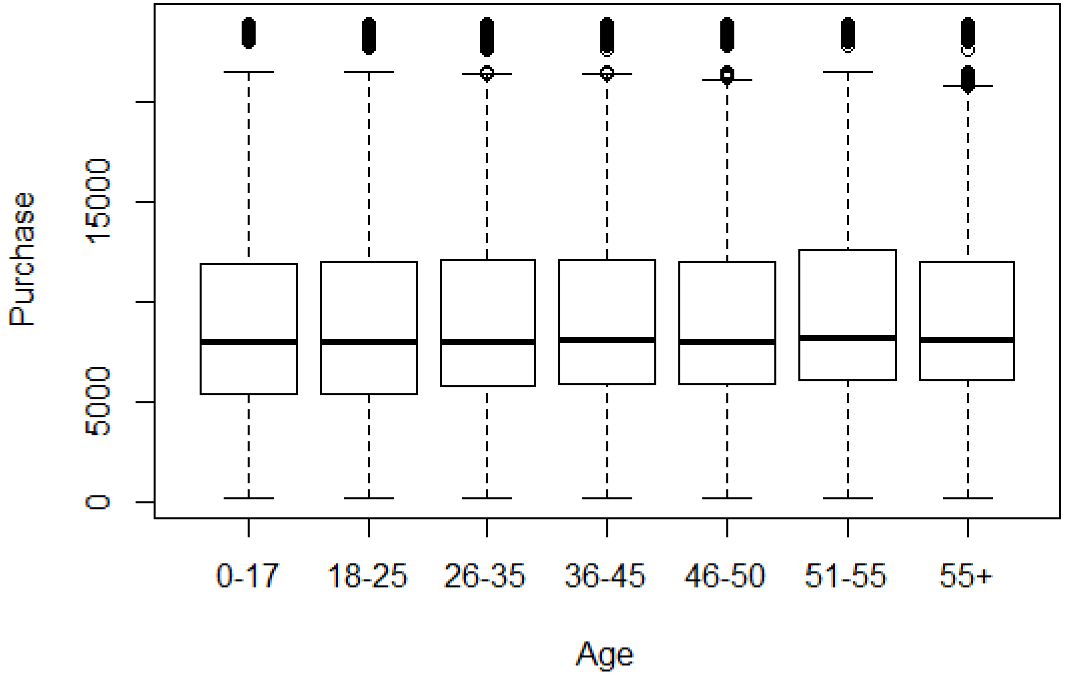
$\mu_{0-17} = \mu_{18-25} = \mu_{26-35} = \mu_{36-45} = \mu_{46-50} = \mu_{51-55} = \mu_{55+}$

OR There is no difference in means ➔ $\beta 1 = \beta 2 = \beta 3 = \beta 4 = \beta 5 = \beta 6 = 0$

Hα: Average purchases made over different age groups are not equal

➔ Not all μ's are equal

OR There is some difference in means $\beta i \neq 0$

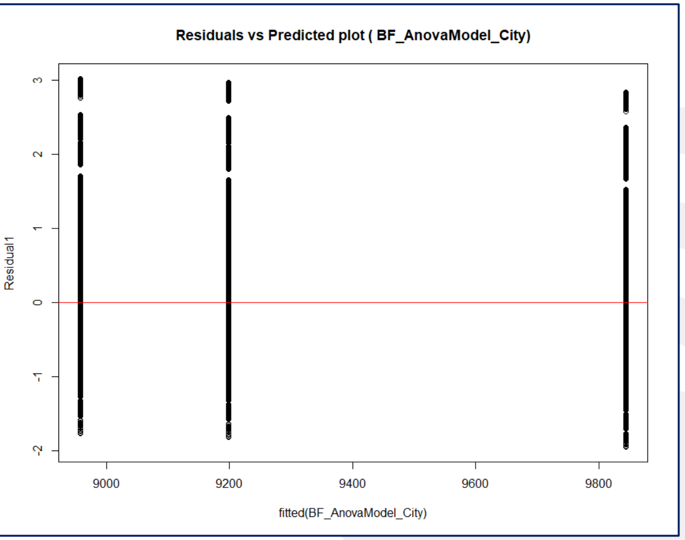


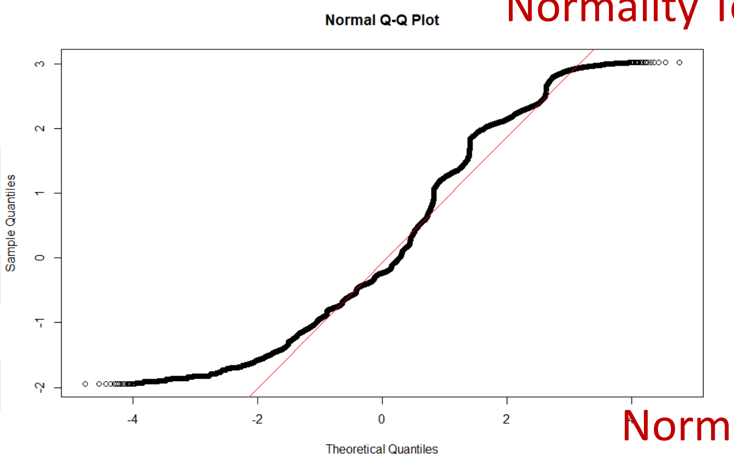

```
> plot(Purchase~Age)
> BF_AnovaModel_Age=lm(Purchase~Age)
> summary(BF_AnovaModel_Age)

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

- 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
- Average purchases are not equal across all age groups.

Residual Analysis for City Category



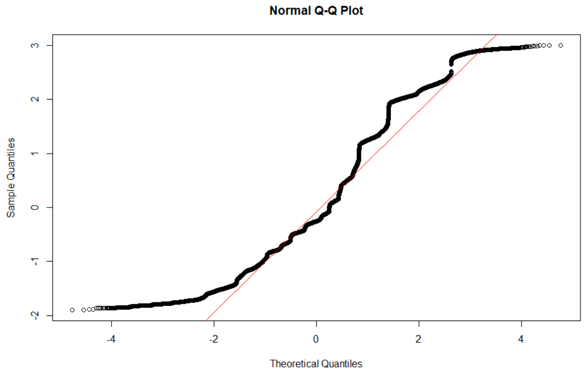


Normality Test for Age Groups

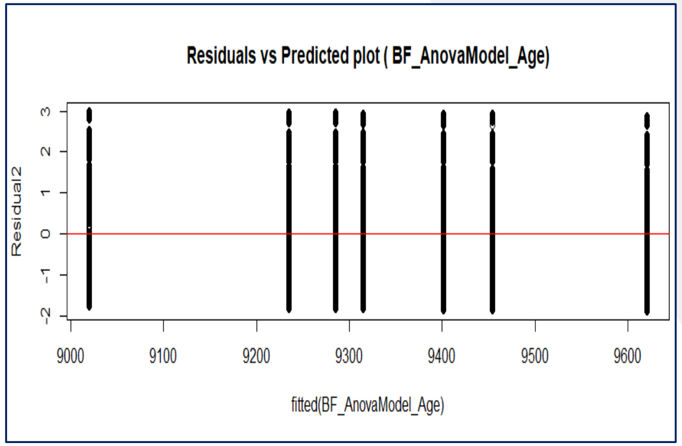
```
> shapiro.test(Residual1)
Error in shapiro.test(Residual1) : sample size must be between 3 and 5000
> ks.test(Residual1,"pnorm")

        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
>
```

Normality Test for City Category

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

        One-sample Kolmogorov-Smirnov test

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



Residual Analysis for Age Groups



INTERPRETATION

Equation 1 :

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

Equation 2 :

$\bar{Y}$(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+)

PREDICT PURCHASE

ALGORITHM USED : Linear Regression Technique

- ✓ Identify Dependent and Independent variables
- ✓ Find Correlation
- ✓ Data preprocessing
  - Replace Missing values
  - Transformation of X Variables
- ✓ Data Split – Hold Out evaluation
  - 80% for Train data
  - 20% for Test data
- ✓ Build a model using Train data
- ✓ Feature Selection
  - Backward Elimination
  - Forward Selection
  - Stepwise Selection

Correlation

```
> # Correlation table
> cor(cbind(BF_Purchase,BF_Us
                BF_Purchase
BF_Purchase    1.0000000000
BF_User        0.0053894723
BF_Prod       -0.0865414730
BF_Gender      0.0600861660
BF_Age         0.0177166304
BF_Occupation  0.0211043402
BF_City        0.0685072913
BF_Stay        0.0054696253
BF_Marital     0.0001290181
BF_Prod1      -0.3141247355
BF_Prod2       0.0383950703
BF_Prod3       0.2841198837
```

```
> Full_Model=lm(formula=traindata$Purchase~traindata$Gender+traindata$Age+traindata$Occupation
rs+traindata$Marital_Status+traindata$Product_Category_1+traindata$Product_Category_2+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
```

Full Model

PREDICT PURCHASE

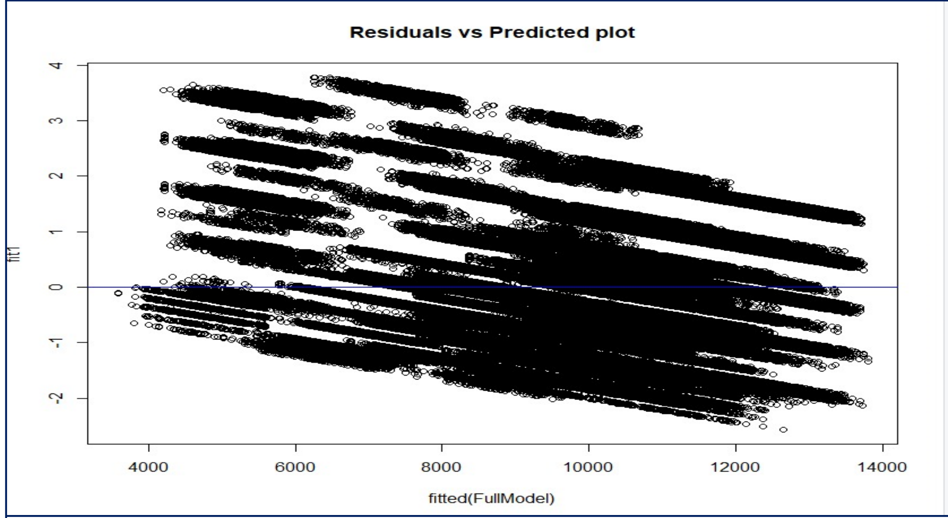
ILLINOIS INSTITUTE OF TECHNOLOGY

- ✓ Residual Analysis
  - Residuals vs Predicted
  - Residuals vs X variables
- ✓ Goodness of Fit Test
  - F Test
  - Individual Parameter Test
  - Co-efficient of determination R2
- ✓ Normality Test
  - Kolmogorov Smirnov (KS) Test
  - QQ Plot
- ✓ Evaluate the model performance

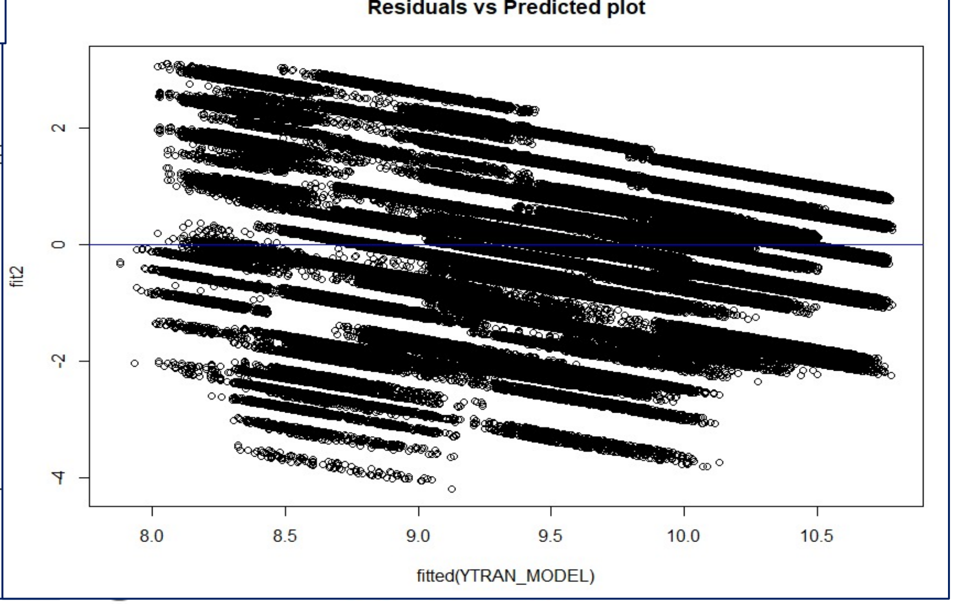
Y Transformed Model

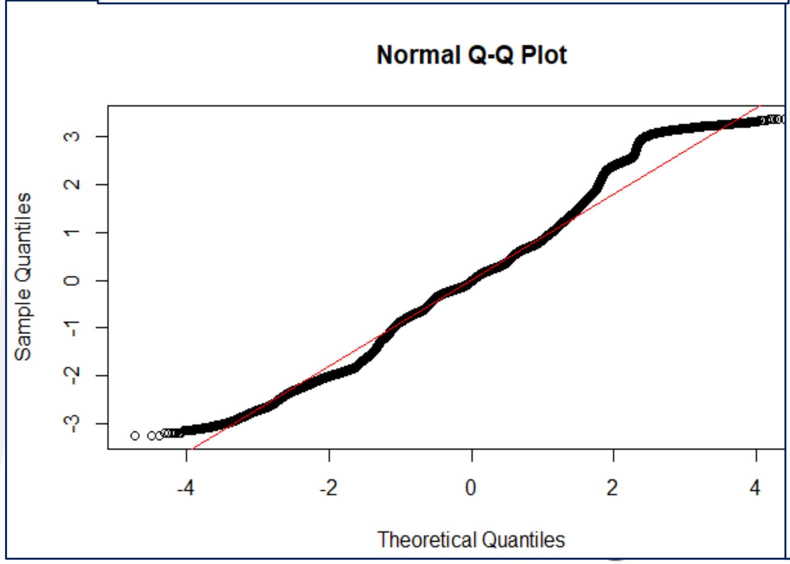
```
> summary(modelback)

Call:
lm(formula = ssqrt_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
-5.4287 -0.7132  0.0618  0.7582  4.0251

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                10.414529   0.015656  665.215  < 2e-16 ***
Occupation                  0.008787   0.001433    6.132 8.67e-10 ***
Marital_Status             -0.017618   0.004268   -4.127 3.67e-05 ***
GenderM                     0.081103   0.004614   17.578  < 2e-16 ***
AgeYoungAdult               0.114191   0.012855    8.883  < 2e-16 ***
AgeAdult                    0.181971   0.012492   14.567  < 2e-16 ***
AgeSenoirAdult              0.218362   0.012840   17.007  < 2e-16 ***
AgeMiddleAged               0.220181   0.014092   15.625  < 2e-16 ***
AgeEarly fifties            0.308188   0.014401   21.401  < 2e-16 ***
AgeSeniorCitizen            0.275335   0.015811   17.414  < 2e-16 ***
City_CategoryB              0.037605   0.004894    7.685 1.54e-14 ***
City_CategoryC              0.174494   0.005296   32.946  < 2e-16 ***
Stay_In_Current_City_Years  0.002703   0.001530    1.767   0.0773 .
Product_Category_1         -0.606375   0.002666 -227.471  < 2e-16 ***
Product_Category_2         -0.011158   0.001281   -8.713  < 2e-16 ***
Product_Category_3          0.092591   0.001362   67.986  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.292 on 430045 degrees of freedom
Multiple R-squared:  0.1801,    Adjusted R-squared:   0.18
F-statistic:  6296 on 15 and 430045 DF,  p-value: < 2.2e-16
```

Residual Plot for Full Model



Residual Plot for Y Transformed Model






```
> 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

>
```

KS Normality Test

Model	ADJ –R2
Full Model without Transformation	0.1358
Full Model with X Transformation	0.1358
Full Model with X Transformation – Backward Elimination	0.1358
Full Model with X Transformation – Forward Selection	0.1358
Full Model with X Transformation – Stepwise	0.1358
After Residual Analysis	
Full Model with Y Transformation	0.18
Full Model with Y Transformation – Backward Elimination (Eliminated Product Cat2)	0.18
Full Model with Y Transformation – Forward Selection	0.1424
Full Model with Y Transformation – Stepwise	0.1424

Model	ADJ –R2	RMSE
Full Model without Transformation	0.1358	4636.44
Full Model with Y Transformation	0.1358	4644.651
Full Model with Y Transformation – Backward Elimination	0.18	4635.448

INTERPRETATION

After backward elimination **Full Model with Y Transformation** can explain 18 % of variations in Purchase using Independent variables.

PREDICT CITY CATEGORY

ILLINOIS INSTITUTE OF TECHNOLOGY

ALGORITHM USED : K- Nearest Neighbor Classification Technique

STEPS :

- ✓ Identify Dependent and Independent variables
  - City category ⇒ Multi Class Classification
- ✓ Data preprocessing
  - Replace Missing values
  - Create Dummy variables
  - Normalize values
- ✓ Data Split – Hold Out evaluation
  - 30% for Train data
  - 70% for Test data
- ✓ Choice of K ⇒ Odd Number to avoid ties
- ✓ Distance Metrics ⇒ Euclidean Distance by default in R
- ✓ Training a model on data
- ✓ Evaluate the model performance
- ✓ Checks 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

INTERPRETATION:

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

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

PREDICT AGE GROUP

ILLINOIS INSTITUTE OF TECHNOLOGY

ALGORITHM USED : Naive Bayes Classification Technique

STEPS :

- ✓ Identify Dependent and Independent variables
  - Age Group ⇒ Multi Class Classification
- ✓ Data preprocessing
  - Replace Missing values
  - Data Transformation
- ✓ Data Split – Hold Out evaluation
  - 80% for Train data
  - 20% for Test data
- ✓ Check for Imbalance Issues
- ✓ Training a model on data
- ✓ Evaluate the model performance
- ✓ Checks accuracy

Model	Independent Attributes	Accuracy
Model	Using all Independent Variables	0.312

```
> NB_Model2=naive_bayes(Age~.,traindata)
> pred2=predict(NB_Model2,testdata)
> accuracy(testdef,pred2) #  0.3109482
[1] 0.3119015
```

INTERPRETATION:

Maximum accuracy achieved is 31.2 % by using conditional probability.

PREDICT MARITAL STATUS

ALGORITHM USED : Logistic Regression Technique

STEPS :

- ✓ Identify Dependent and Independent variables
  - Marital Status ⇒ Binary Classification
- ✓ Data preprocessing
  - Replace Missing values
- ✓ Data Split – Hold Out evaluation
  - 80% for Train data
  - 20% for Test data
- ✓ Build a model on train data
- ✓ Feature Selection to improve the model
- ✓ Evaluate the model performance
  - Set CUT-OFF value ⇒ 0.5
- ✓ Find accuracy

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

INTERPRETATION:
Using Binomial function, Maximum accuracy achieved is 66.9%.

```
> 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
>
```

Forward Selection

```
> Forward_Logistic_Model=step(Base_Logistic_Model,scope=list(upper=Full_Logistic_Mod
> 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

>
```

PURPOSE

⇒ On predicting the purchase, retails can plan estimate their profit during Black Friday sales.
Example: Target to achieve 15% profit during Black Friday sales

⇒ Age group of customers, retails can plan to introduce new products to grab the attention of customers in certain age category
Example: Smart phones for senior citizens

⇒ On predicting the city category, retails can plan to certain products are on high demand in certain areas.
Example: Promoting St. Patricks' costume and goodies for Chicago residents.

⇒ On predicting the marital status of customers, retails can plan to promote new products and discounts.
Example: Promoting home décor and house hold products

CONCLUSION

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

The models built helps to determine

- The sales during black Friday
- Age group of the customers
- Marital status of the customers
- City where the customer resides

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.

THANK YOU !!!