Module 6: Final Project – Big Research

**Black Friday Analysis**

Consumer Purchasing Behavior Case Study

AY6015 CRN 81055– Intermediate Analytics - Spring 2019

Instructor: Valeriy Shevchenko

June 28, 2019

**Group Delta**

Rahman Shavahatli

Binbing Zhao

Priyanka Balaji

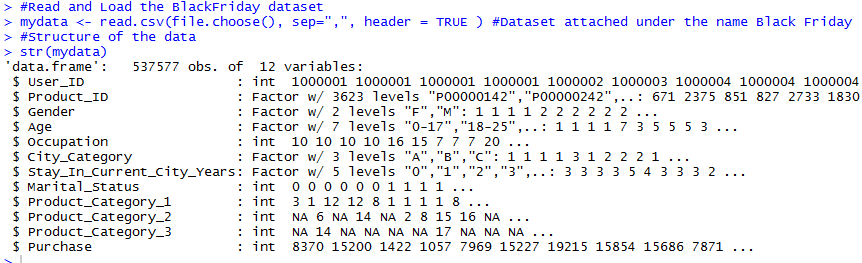
Northeastern University

**Introduction**

Black Friday is a day that is termed to describe the day after Thanksgiving during the month of November-December. It is traditionally considered as one of the busiest shopping seasons of the year. Black Friday season is a very important period for the market and economy. Thirty percent of the annual retail sales occur during this holiday season. The main purpose of our project is to understand how better the customer purchase behavior changes against different products.

**Analysis**

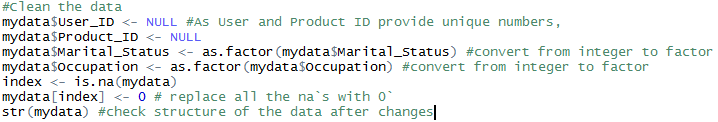
Black Friday dataset is retrieved from the Kaggle.com and our dataset have 537577 observations and 12 variables. The dataset here is a sample of the transactions made in a retail store. The store wants to know better the customer purchase behavior against different products. When we check the structure of the dataset it is clearly visible that our variables are either integer or categorical.

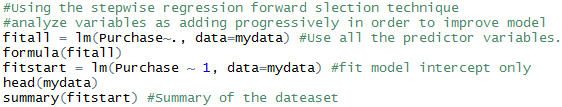
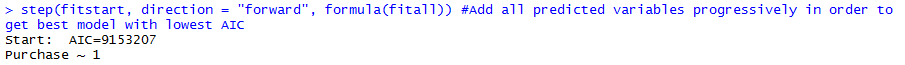


Data quality is an important issue which has been addressed and recognized in research communities such as data warehousing, data mining and information systems. It has been agreed that poor data quality will impact the quality of results of analyses and that it will therefore impact on decisions made on the basis of these results (Liebchen, 2010). So, in order to check data quality, we always need to look for whether dataset has any missing values, anomalies, and typos or not.

Turning to the case, in first step we determined error types such as missing values, and data type. In first stage, we started to clean our data set with “User\_ID” and “Product\_ID” column as they provided the unique numbers and we cannot be able to use mentioned variables further on our modeling part. The next steps were to convert “Occupation” and “Marital\_Status” variables from integer to factor. For example, 0 and 1 in Marital\_Status are an integer number, while them actually indicates where female or males married or unmarried. The final cleaning step was to convert all NAN values to zero under the “Product\_Category\_2” and “Product\_Category\_3” variables and convert them to integer. We considered that the customer purchased products not from all the product category.

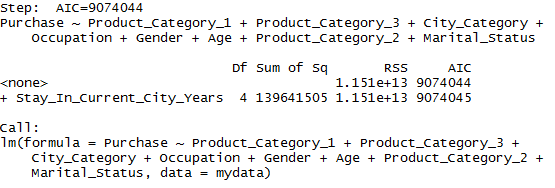
The following R code is used:

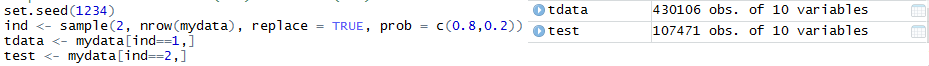
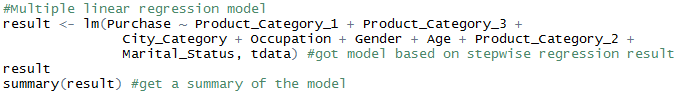
f



After completing the data cleaning part, we switched to using the stepwise regression which is a modification of the forward selection. It is a technique which a variable adding progressively and all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level in order to improve model. In forward selection we started to fit linear model intercept only, which means there is no particular predictor variables in the model and we progressively add variables for improving the model. As we already mentioned our response variable is “Purchase” and fitted all predictive variables to analyze. An in-depth study stepwise regression made me conclude that it gives us best fit model with the lowest Akaike information criterion (AIC). AIC is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

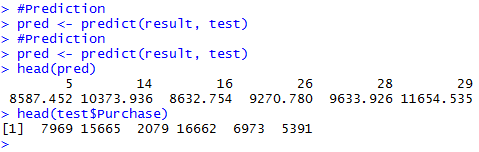
It is clearly visible that our model started with AIC value by 9153207 and after analyzing it gave us the best formula with the lowest AIC.





As we got the best formula to our multiple linear model, we started to divide our model into train and test data to make a prediction. We decided to analyze 80% of observations in training stage and the rest 20% in testing stage. Based on insights, we have two new data sets as we already mentioned. Train data has about 430106 observations and we specified it as 1. Moreover, the second data set called train which consist of 107471 observations and specified as 2.

After all, we started to make a prediction:





As it can be seen from the table, there big differences between actual purchase and in our prediction. Based on the summary of our model, R squared is just about 14%, which means the independent variables failed to explain dependent variable in a perfect way.

**Hypothesis Testing**

First, we have to install all the useful R packages and load them into the system:

# Install and load the packages

install.packages("tidyverse")

library(tidyverse)

install.packages("psych")

library(psych)

install.packages("plotly")

install.packages("data.table")

library(plotly)

library(dplyr)

library(scales)

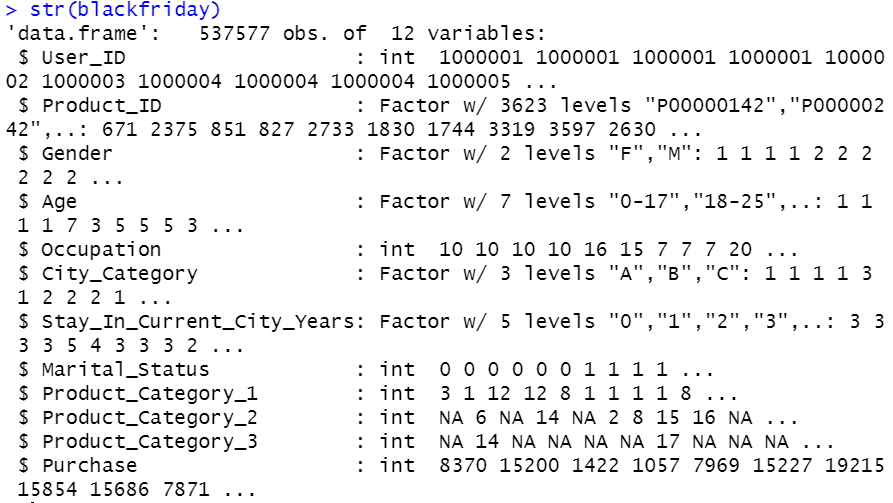
library(data.table)

# Read the blackfriday dataset by choosing csv file

blackfriday <- read.csv(file.choose(),header = T)

# Structure the data

str(blackfriday)



Because there are some blank spaces left in this data set, we must use sapply() function to checking the missing value in order to carry out more accurate hypothesis testing.

# Checking missing value

sapply(blackfriday , function(x) sum(is.na(x)))

> sapply(blackfriday , function(x) sum(is.na(x)))

User\_ID Product\_ID

0 0

Gender Age

0 0

Occupation City\_Category

0 0

Stay\_In\_Current\_City\_Years Marital\_Status

0 0

Product\_Category\_1 Product\_Category\_2

0 166986

Product\_Category\_3 Purchase

373299 0

**Hypothesis Testing 1: The Gender Difference**

In this part, we assume that women are more likely to spend much more money on the Black Friday than men.

So, we use 95% confidence interval for µ1 - µ2, alpha = 0.5 to test the hypothesis that µ1 > µ2:

|  |  |  |  |
| --- | --- | --- | --- |
| **Null Hypothesis Ho:** | ***µ*1 -*µ*2** | **≤** | **0** |
| **Alternative Hypothesis Ha:** | ***µ*1 -*µ*2** | **>** | **0** |

The R code to test the hypothesis is as follows:

### Hypothesis Testing 1: Gender Difference

female\_purchase <- blackfriday %>%

filter(Gender == "F") %>%

select(Purchase)

#Define the purchase data of men

male\_purchase <- blackfriday %>%

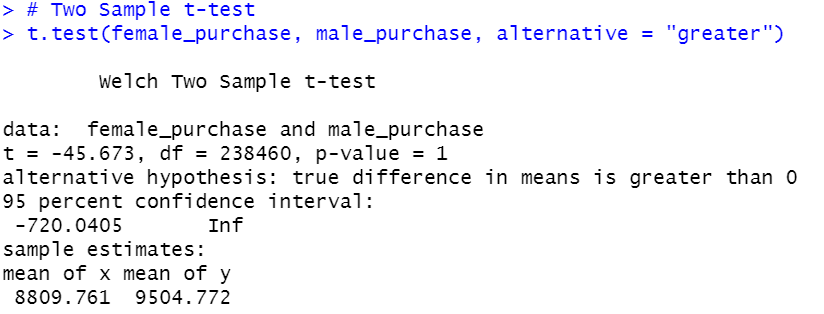
filter(Gender == "M") %>%

select(Purchase)

For the statistical analysis, the two-sample test is to test the performance on the data of two different samples, and to determine whether there is difference between these two populations.

# Two Sample t-test

t.test(female\_purchase , male\_purchase , alternative = "greater")



Based on the result above from RStudio, we can see the large p-value. We use this method to decide whether reject H0 or not (Reject H0 if P value ≤α). Because the p-value is 1, which is much greater than the significance level, and there is no sufficient evidence to prove that the mean of female purchase amount $ Is greater than the Male purchase amount $. We can conclude that women are not likely to purchase more products than men on black Friday.

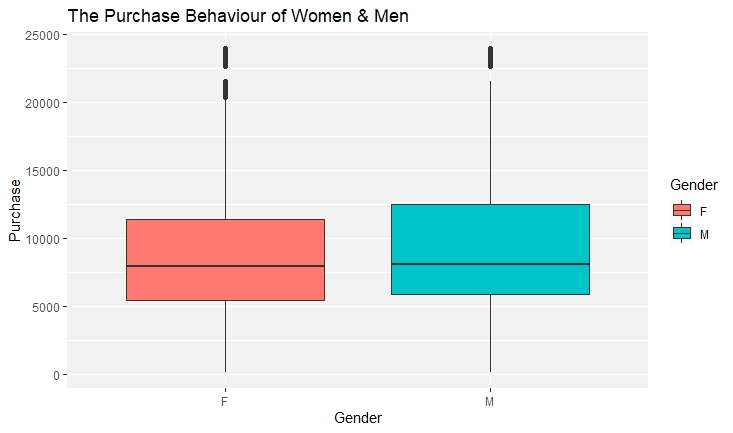
Then we make a boxplot to visualize this dataset to present our findings in the hypothesis testing again. We input the ggplot() function for the plotting below.

# Plotting the gender difference

ggplot(aes(x = Gender , y = Purchase , fill = Gender),data = blackfriday) +

geom\_boxplot() +

ggtitle("The Purchase Behaviour of Women & Men")



We can easily see from the plot that the purchase amount of male is larger than that of female. The interquartile range of men is from approximately 6500 dollars to 12500 dollars, while the IQR for purchase amount of female is from about 6000 dollars to 11150 dollars. Even though there are some outliers on the top of the plot, from the IQR range value of different gender and we can see the purchase amount of male is negatively skewed because the median is greater than the mean. That means men bought much more expensive items than the relatively cheaper stuff in that retail store on black Friday. And they are probably more willing to spend money on better quality and technology products on black Friday than women according to this observation.

# Sampling method to estimate the population

t.test(blackfriday$Purchase , mu = 967.13)

Because this dataset is for 550 000 consumers in a retail store on black Friday, if we have to make the inference for the population, we need to use data from the balance web to examine the sample mean.

> # Sampling method to estimate the population

> t.test(blackfriday$Purchase , mu = 967.13)

One Sample t-test

data: blackfriday$Purchase

t = 1231.6, df = 537580, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 967.13

95 percent confidence interval:

9320.545 9347.175

sample estimates:

mean of x

9333.86

Based on the result of estimating, we can see a very similar mean between sample purchasing data and the populating dataset. So, we can conclude this situation is not just in this retail store but happening in most of the stores operating marketing strategies to attract customers on the black Friday.

**Hypothesis Testing 2: The Marital Status Difference**

In this part, we assume that married consumers are more likely to spend more or less on the Black Friday than single customers due to the family expenditures, living standards and marriage life costs. To test the difference, we use 95% confidence interval for µ1 - µ2, alpha = 0.5 to test the hypothesis that µ1 ≠ µ2:

|  |  |  |  |
| --- | --- | --- | --- |
| **Null Hypothesis Ho:** | ***µ*1 -*µ*2** | **=** | **0** |
| **Alternative Hypothesis Ha:** | ***µ*1 -*µ*2** | **≠** | **0** |

The R code to test the hypothesis is as follows:

### Hypothesis Testing 2: Marriage Status Difference

ma\_purchase <- blackfriday %>%

filter(Marital\_Status == 1) %>%

select(Purchase)

unma\_purchase <- blackfriday %>%

filter(Marital\_Status == 0) %>%

select(Purchase)

# Two Sample t-test

t.test(unma\_purchase, ma\_purchase)

> t.test(unma\_purchase , ma\_purchase)

Welch Two Sample t-test

data: unma\_purchase and ma\_purchase

t = -0.094631, df = 473150, p-value = 0.9246

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-28.38174 25.76732

sample estimates:

mean of x mean of y

9333.325 9334.633

Based on the result and the p-value (0.9246), which is greater than the alpha, we can conclude that there is no enough evidence to prove that the marital status has been an impacting factor for consumers’ purchase behavior.

We can also testify the conclusion on the plot below:

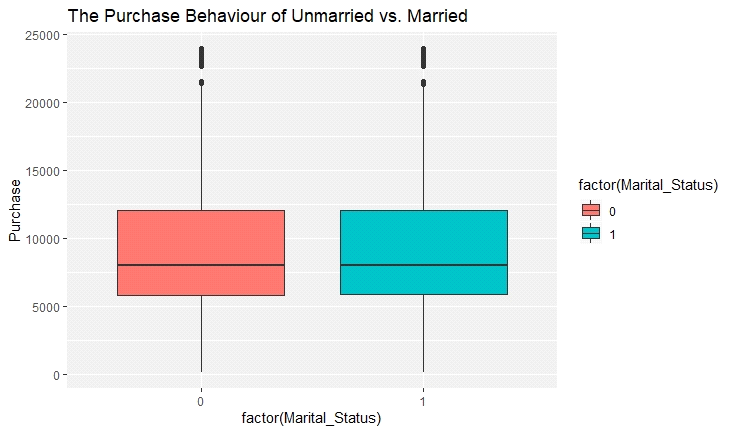
# Plotting the marital status difference

blackfriday %>%

ggplot(aes(x = factor(Marital\_Status) , y = Purchase , fill = factor(Marital\_Status)))+

geom\_boxplot() +

ggtitle("The Purchase Behaviour of Unmarried vs. Married")



According to the result in R and the box plot above, there is no significant difference between married consumers and single consumers of purchasing behavior on black Friday.

**Hypothesis Testing 3: The Age Group Difference**

### Hypothesis Testing 3: Age Difference

# The analysis of variance

age.aov <- aov(Purchase ~ Age , data = blackfriday )

# Summarizing the analysis of variance model

summary(age.aov)

# Pairwise comparisons using t tests

pairwise.t.test(blackfriday$Purchase , as.vector(blackfriday$Age))

> summary(age.aov)

Df Sum Sq Mean Sq F value Pr(>F)

Age 6 6.471e+09 1.078e+09 43.49 <2e-16 \*\*\*

Residuals 537570 1.333e+13 2.480e+07

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> # Pairwise comparisons using t tests

> pairwise.t.test(blackfriday$Purchase , as.vector(blackfriday$Age))

Pairwise comparisons using t tests with pooled SD

data: blackfriday$Purchase and as.vector(blackfriday$Age)

0-17 18-25 26-35 36-45 46-50 51-55

18-25 1.0e-05 - - - - -

26-35 5.2e-11 0.00026 - - - -

36-45 < 2e-16 6.4e-13 2.7e-05 - - -

46-50 2.5e-07 0.24329 0.32753 0.00026 - -

51-55 < 2e-16 < 2e-16 < 2e-16 2.9e-12 < 2e-16 -

55+ 9.3e-15 9.9e-08 0.00052 0.32753 0.00031 0.00052

P value adjustment method: holm

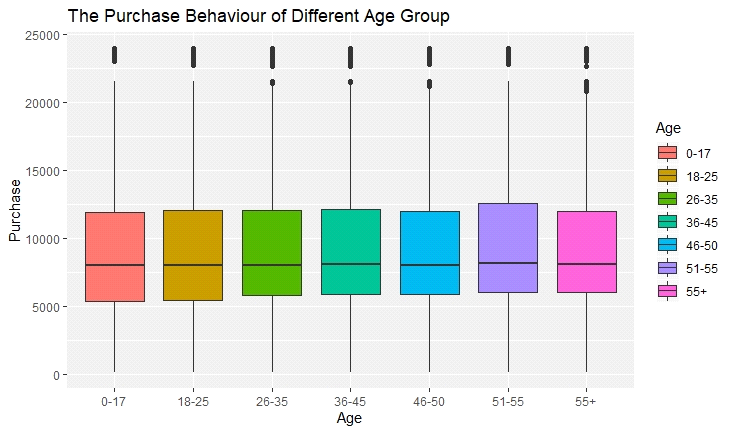
# Make a plot to visualize the age difference

blackfriday %>%

ggplot(aes(x = Age ,y = Purchase , fill = Age)) +

geom\_boxplot()+

ggtitle("The Purchase Behaviour of Different Age Group")



**Hypothesis Testing 4**

We assume women between 26-35 who are married are more likely to make purchase on Black Friday than the same age and marital status group of Men.

So, we use 95% confidence interval for µ1 - µ2, alpha = 0.5 to test the hypothesis that µ1 > µ2:

|  |  |  |  |
| --- | --- | --- | --- |
| **Null Hypothesis Ho:** | ***µ*1 -*µ*2** | **≤** | **0** |
| **Alternative Hypothesis Ha:** | ***µ*1 -*µ*2** | **>** | **0** |

### Hypothesis Testing 4

women <- blackfriday %>%

filter(Marital\_Status == 1 , Age == "26-35", Gender == "F") %>%

select(Purchase)

men <- blackfriday %>%

filter(Marital\_Status == 1 , Age == "26-35", Gender == "M") %>%

select(Purchase)

# Two sample t-test

t.test(women , men , alternative = "greater")

> t.test(women , men , alternative = "greater")

Welch Two Sample t-test

data: women and men

t = -10.976, df = 34843, p-value = 1

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-494.839 Inf

sample estimates:

mean of x mean of y

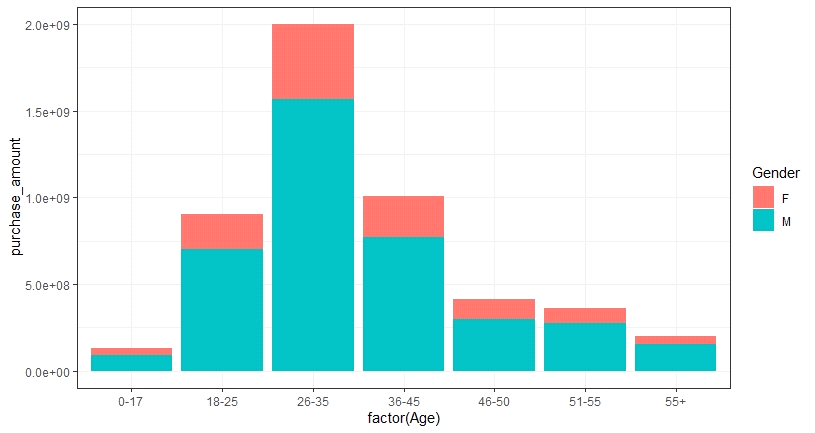
8988.077 9418.422

The result shows that the p-value is 1, which is much greater than the significance level, and there is no sufficient evidence to prove that women between 26-35 who are married are more likely to make purchase on Black Friday than the same age and marital status group of Men.

In order to make all the hypothesis testing in a visualized way and easier to interpret. We will plot multiple groups of bar charts and line plot as follows:

# Plotting the bar chart to visualize two groups

blackfriday %>% group\_by(Age, Gender) %>% summarize(purchase\_amount = sum(as.numeric(Purchase))) %>% ggplot(mapping = aes(x = factor(Age), y = purchase\_amount, fill = Gender)) + geom\_col() + theme\_bw()



This bar chart shows the highest purchasing volume is male group at 26-35, and the second and the third group are 36-45 and 18-25. We think retailers should focus more potential consumers to boost their sales in sales volume by creating personalized promotion strategies through different channels based on their preferences and lifestyles for male age group like 36-45, 18-25 and female between 18-35.

# Define the multiline chart

df\_multiline <- blackfriday[c(3,4,8,12)]

df\_multiline <- df\_multiline %>% group\_by(Gender,Marital\_Status,Age) %>%

summarise(meanPurchase= mean(Purchase))

df\_casted<-dcast(df\_multiline, Age ~ Gender + Marital\_Status, value.var = c("meanPurchase"))

df\_casted['place'] <- seq(1,7,1)

# Plotting the multiple line chart to visualize the hypothesis testing

ggplotly(

ggplot(df\_casted,aes(x=place))+

geom\_line(aes(y=M\_0,color='Male\_Single'))+

geom\_point(aes(y=M\_0),color='forestgreen',fill='white',size=3,shape=21,stroke=1)+

geom\_line(aes(y=M\_1,color='Male\_Married'))+

geom\_point(aes(y=M\_1),color='red',fill='white',size=3,shape=21,stroke=1)+

geom\_line(aes(y=F\_0,color='Female\_Single'))+

geom\_point(aes(y=F\_0),color='forestgreen',fill='white',size=3,shape=21,stroke=1)+

geom\_line(aes(y=F\_1,color='Female\_Married'))+

geom\_point(aes(y=F\_1),color='red',fill='white',size=3,shape=21,stroke=1)+

scale\_color\_manual( values = c('Male\_Single'='turquoise',

'Male\_Married'='royalblue1',

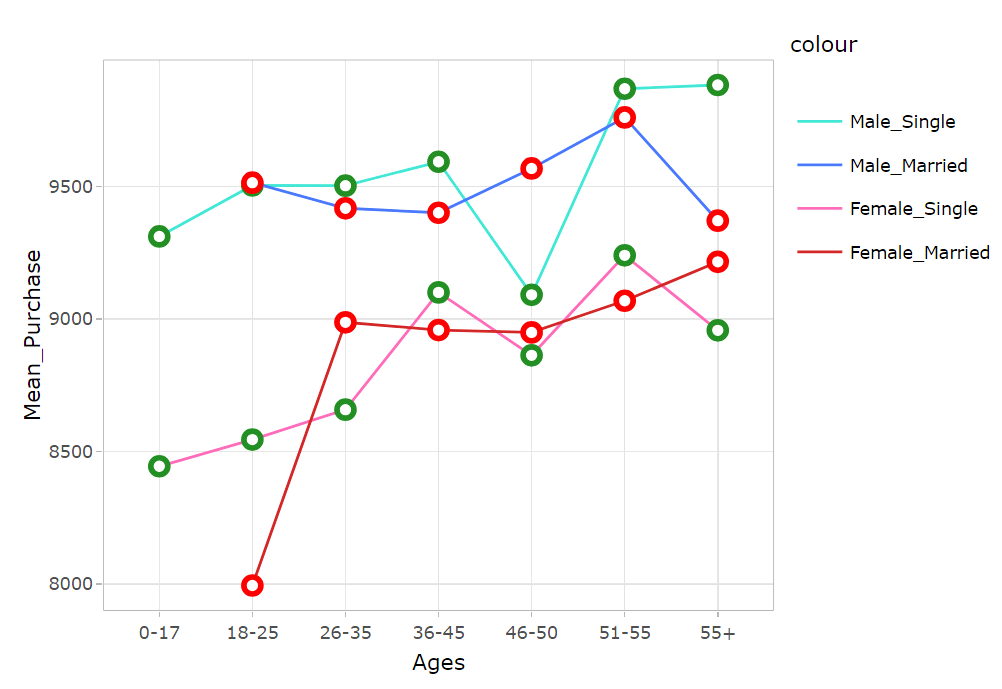
'Female\_Single'='hotpink',

'Female\_Married'='firebrick3'))+

labs(x="Ages",y="Mean\_Purchase")+

scale\_x\_discrete(limits=df\_casted$Age)+

theme\_light())



We can clearly see the gender, age group and marital status differences within this one single multiline plot. The highest point of the mean of purchase amount is single male group aged 55+, and the lowest point of the mean of purchase amount is married female group age at 18-25. The younger generation, for instance at the age of 26 to 35, the mean for both married and single men are greater than all the female group. This is a signal for the retailers to go deeper in observing those specific products sold to single male who are at age group of 36-45 and both married and single male age group at 51-55. It is a phenomenon to show that different income level, consumption ability or living standards will also be important affecting factors on the average expenditure towards shopping. We have to analysis more detailed information based on the data and dig the useful insight behind rather than only look at the overall amount figure.

**Data Mining Technique : Clustering**

* Clustering is the term coined to explain the exploration of data, where the similar points are grouped into a cluster.
* Marketing segmentation is a specific case that would investigate the sales of Black Friday retail store data for determining the ability of consumer zones to make affordable and reliable marketing strategies.
* K-Means clustering method is used in this problem for clustering the total number of people by Gender.
* Since there is no response variable, this is considered as an unsupervised method that seeks to find relationships between the n observations without being trained by a response variable.

We begin by loading the required packages in R. Next, we determine the number of clusters that are required, and we map the data to compute multiple models with different k values using the map\_dbl function by selecting the variable Purchase. And then we determine the actual number of clusters required by utilizing the elbow method.

#Load the required libraries tidyverse and DT

> library(tidyverse)

> library(DT**)**

#Then we read and load the dataset into R after which, we store it in the variable "sales\_df"

**>** sales\_df <- read.csv("D:/BlackFriday.csv")

> BlackFridayForClustering <- sales\_df %>%

select(Purchase)

# Utilize map\_dbl function to execute multiple models with different k values (centers)

tot\_withinss <- map\_dbl(1:10, function(k){

model <- kmeans(x = BlackFridayForClustering, centers = k)

model$tot.withinss

})

# Generates a data frame that contains both k and tot\_withinss

elbow\_df <- data.frame(

k = 1:10,

tot\_withinss = tot\_withinss

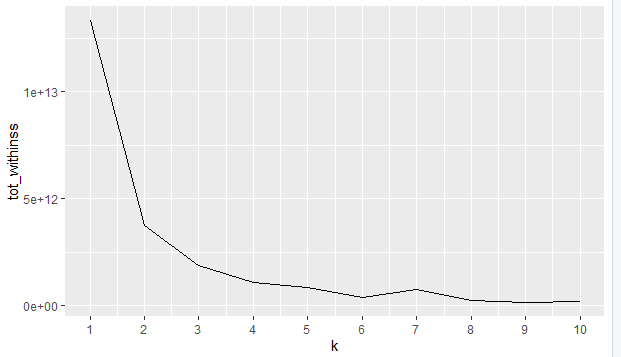
)

# Draws graphical representation-elbow plot

ggplot(elbow\_df, aes(x = k, y = tot\_withinss)) +

geom\_line() +

scale\_x\_continuous(breaks = 1:10)



The elbow method is used to obtain the optimal number of clusters for K-means clustering.

For the range of values of k from 1 to 10 , we observe that there is a clear elbow at k=3 , indicating that 3 is the best number of clusters to be considered**.**

Now, we begin our clustering with centroid = 3. Next, we obtain the cluster assignment vectors and create a new data frame by appending the cluster vector assignment. Summarise the clusters based on the minimum, maximum and average purchases. And our final cluster results are obtained.

# Constructs a k-means Model with 3 cluster centers

model\_km3 <- kmeans(BlackFridayForClustering, centers = 3)

# Draws out the vectors acquired by the k-means model

clust\_km3 <- model\_km3$cluster

# Generates a latest data frame including the appended cluster assigned

BlackFriday\_Clust <- mutate(sales\_df, cluster = clust\_km3)

# Outlines Clustering

BlackFriday\_Clust\_Note <- BlackFriday\_Clust %>%

group\_by(cluster) %>%

summarise(min\_purchase = min(Purchase),

max\_purchase = max(Purchase),

avg\_purchase = round(mean(Purchase),0))

# Number of people in every cluster

BlackFriday\_Clust %>%

group\_by(Gender, cluster) %>%

summarise(n = n()) %>%

ggplot(aes(x=Gender, y = n)) +

geom\_col(aes(fill = as.factor(cluster))) +

theme\_linedraw() +

theme(legend.box.background = element\_rect(colour = "black"),

legend.background = element\_rect(fill = "gainsboro"),

panel.background = element\_rect(fill = "gainsboro", colour = "white", size = 0.5, linetype = "solid"), #theme panel settings

plot.background = element\_rect(fill = "gainsboro"), #theme panel settings

panel.grid.major = element\_line(size = 0.5, linetype = 'solid', colour = "white"),

# Settings of Theme Panel

panel.grid.minor = element\_line(size = 0.25, linetype = 'solid', colour = "white"), #Settings of Theme Panel

plot.title = element\_text(hjust = 0, face = 'bold',color = 'black'), # Setting of Title

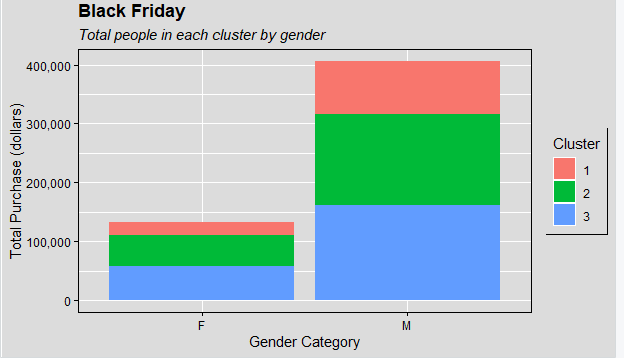
plot.subtitle = element\_text(face = "italic")) + #Setting of Subtitle

labs(x = 'Gender Category', y = 'Total Purchase (dollars)', title = "Black Friday", #Label Title and Axes

subtitle = "Total people in each cluster by gender") + #Label subtitle

guides(fill=guide\_legend(title = "Cluster")) + # Delete Legend color

scale\_y\_continuous(labels = scales::comma) #prevent scientific number in x-axis



The results show 3 clusters 1, 2 and 3 that is being clustering by Gender, Male (M) and Female(F). The graph is being plotted against the Total Purchase value. We observe from the F that cluster 3 has the maximum number of female population rate that have purchased up to $60,000 and cluster 1 indicating minimum number of female populations in that region. But resulting cluster 1 have purchased the highest ranging from $110,000 to $140,000. From the M, it is indicated that cluster 3 has the maximum number of people and cluster1 has the least number of male populations. But the purchase rate is higher for the cluster 1 and lower for cluster 3.

Overall expenditure made by Female is $140,000 and the total purchase made by the Male population is $410,0000.

**Conclusion**

In conclusion, from all our obtained results, we can conclude that the Male population have a higher purchase rate and do a greater amount of shopping than the Female population on a Black Friday day. The age group of the consumer who make the maximum purchase is between 25-38. Also, based upon the model created, we have predicted least accuracy to predict purchase behavior against different products, but we believe that by optimizing the model and future engineering we can improve the performance on the model by predicting more accurately.

**References**

1. Kaggle (2018) Black Friday dataset Retrieved from <https://www.kaggle.com/mehdidag/black-friday>

2. J. Scott Armstrong (1991). "Prediction of Consumer Behavior by Experts and Novices". Journal of Consumer Research. 18 (2): 251–256. Retrieved from <https://web.archive.org/web/20100620213825/http://marketing.wharton.upenn.edu/documents/research/Prediction%20of%20consumer%20behavior.pdf>

3. Liebchen,G.A. (2010). Data Cleaning Techniques for Software Engineering Data Sets.