NORTHEASTERN UNIVERSITY

A Project Report on

"BLACK FRIDAY SALES PURCHASE ANALYSIS"

By
Group 26

Akash Chitrey **001888848**
Darshan Durve **001898887**

Under the guidance of
Professor. Xuemin Jin

Executive Summary:

Black Friday is one of the most popular and busiest times of the year for shoppers as well as retailors. Every year, millions of people head out shopping to retail stores which offer the best discounts or explore the internet to get the best deals searching long and wide. This project is an attempt to uncover insights on consumer behavior against various products which in turn affects the total purchase made at a store. This study will be particularly helpful to both stores as well as e-commerce businesses as they can get a better understanding on which products are making the most money with different sectors of the buyers. This in turn can help them identify their target audience as well as position and market their products better.

The data is taken from Kaggle which is a sample of transactions made at a retail store. The course of this report will start by discussing the background of Black Friday and why we chose this particular dataset, followed by explaining the variables used in our dataset by plotting various graphs and observing the relationships between them. We have then described how we preprocessed our data in order for it to be fit and clean to run various algorithms on.

Considering the purpose of this project we have chosen to design and run Linear Regression and Random Forest for the prediction of total purchase and K – means Clustering to segment the data into a set of homogenous clusters of records for generating the insight of how many people belonged to each city and how that ranked in terms of total purchase. Based on our observations and the model which generated the highest accuracy, the age group 26-35 contained the highest buyers, the majority of buyers were unmarried and the most important factors for retailors and shop owners are the gender of consumers, occupation of consumers, marital status and age of the buyers.

I. Background and Introduction

Black Friday as we know it today is an extravaganza of sales, promotions, and long lines outside of stores. Although the term "Black Friday" originally represented the pitfalls of two Wall Street businessmen and the mayhem of downtown Philadelphia following Thanksgiving, it is now familiarly known today as the busiest shopping day of the year. Retailers such as Target, Best Buy, Amazon, and many others look forward to this day every year with the hopes that consumers will take advantage of door-busting deals. The term "Black Friday" has also spawned other retail holidays such as "Cyber Monday", "Small-Business Saturday", and "Giving Tuesday." However, to maximize sales retailers are always overwhelmed by several factors surrounding customer purchase behavior against their products and thus the intent behind this study is to drive insights on how the purchase outcome is determined by the various attributes of consumers. The predictor and classifier designed in this report could be utilized by retailors to help better product positioning, marketing as well as determining the total purchase expected during Black Friday.

Step-Wise Solution:

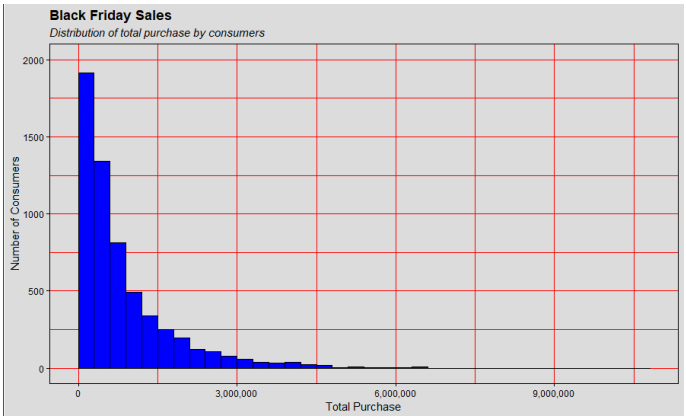
- Perform exploratory data analysis to establish the relationship between the indicators and outcome.
- Perform feature selection, data cleaning, dimension reduction and identify outliers.
- Visualize the data using bar graphs and PCA to find any existing trends and patterns.
- Create a predictive model using Linear Regression and Random Forest to predict the total purchase cost.

II. Data Exploration and Visualization

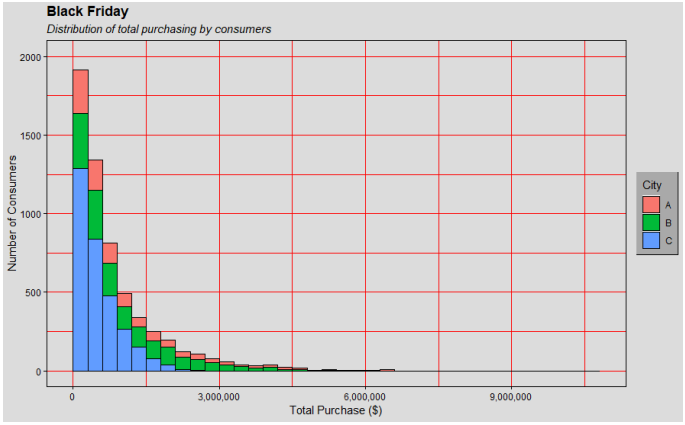
Exploratory Data Analysis

We analyzed the distribution of various predictors with the outcome variable, **Purchase()** to find the trends and relationship between them.

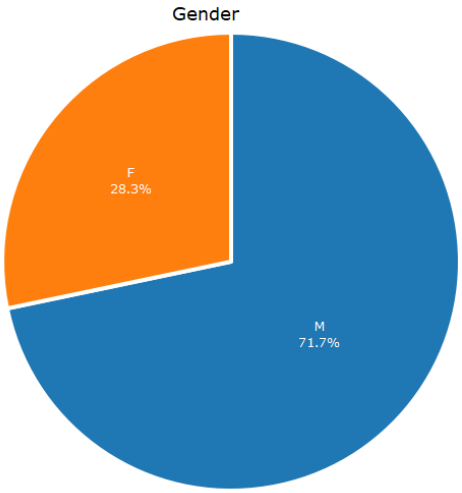
1) ***Total Purchase Distribution:*** We analyzed the distribution of the total sum of purchase by the total number of consumers during the Black Friday Sales to understand the purchase trend by most consumers.



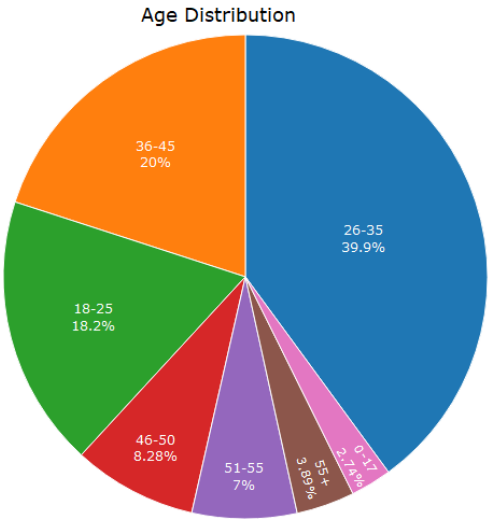
2) ***Total Purchase by City:*** We analyzed the variation between the total sum of purchase distributed by city (City A, City B, City C) and we can see that City C had the highest number of consumers.



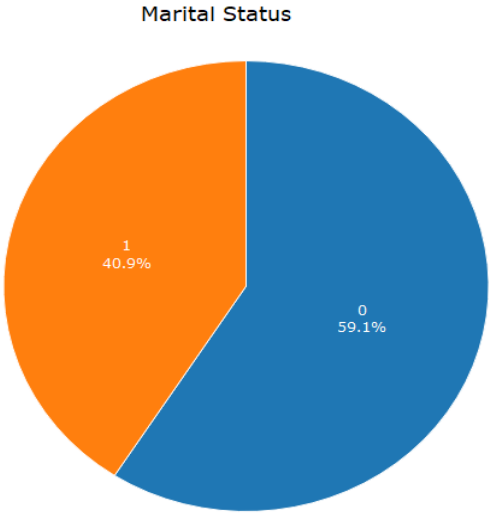
3) ***Total Purchase by Gender:*** We analyzed the proportionality between the number of male and female consumers with respect to the purchase. We can see that there is a substantially higher ratio of male consumers versus female.



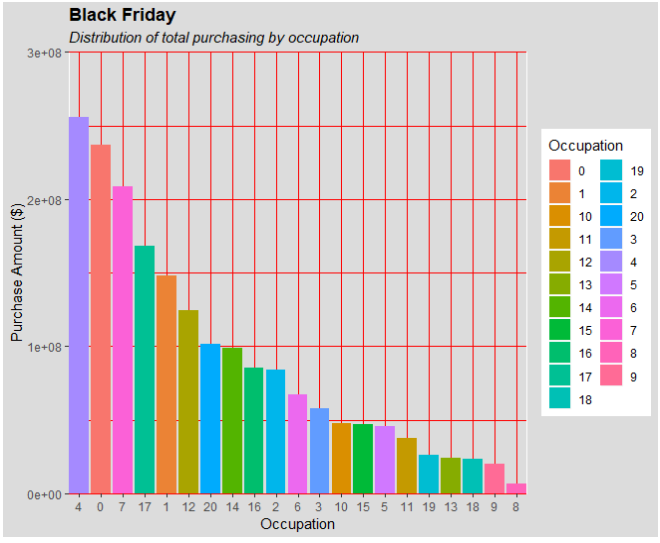
4) ***Total Purchase by Age***: We analyzed the variation of all the age groups present with the total purchase. Most of the consumers were between 26-35, followed by 36-45 and 18-25 respectively.



5) ***Total Purchase by Marital Status:*** We analyzed whether the majority of the consumers are married or single. Since, the values were masked, let's assume 0 = Single and 1 = Married. We can see that most of the consumers were single.



6) ***Total Purchase by Occupation:*** We analyzed the distribution of the total purchase by the occupation of all the consumers. Occupations 4, 0 and 7 were the top 3 occupations respectively.



III. Data Preparation and Preprocessing

i. ***Data Summary:*** The Black Friday sales dataset from Kaggle contains 550,000 instances with 13 attributes:

Attributes	Description
User_ID	Unique identifier assigned to each buyer.
Product_ID	Unique identified assigned to each product sold.
Gender	The gender of each consumer who made a purchase, M = Male and F = Female.
Age	The age bracket of each buyer.
Occupation	The occupation each buyer belonged to. Occupations range from 0-20.
City_Category	The city each buyer was from, City A, B or C.
Stay_In_Current_City_Years	The duration each buyer has stayed in their current city.
Marital_Status	The relationship status of each buyer, whether married or single.
Product_Category_1	The parent category of a product.
Product_Category_2	The parent category of a product.
Product_Category_3	The parent category of a product.
Purchase	Total purchase amount ($)

```
   User_ID            Product_ID     Gender         Age          Occupation
Min.   :1000001   P00110742:  1591   F: 36932   0-17 : 4789   Min.   : 0.00
1st Qu.:1001497   P00025442:  1586   M:127346   18-25:30889   1st Qu.: 2.00
Median :1003053   P00112142:  1539              26-35:65916   Median : 7.00
Mean   :1003001   P00057642:  1430              36-45:32758   Mean   : 8.18
3rd Qu.:1004418   P00184942:  1424              46-50:13135   3rd Qu.:14.00
Max.   :1006040   P00046742:  1417              51-55:11018   Max.   :20.00
                  (Other)   :155291             55+   : 5773
City_Category Stay_In_Current_City_Years Marital_Status   Product_Category_1
A:40848       0 :22061                   Min.   :0.0000   Min.   : 1.000
B:68185       1 :57297                   1st Qu.:0.0000   1st Qu.: 1.000
C:55245       2 :31040                   Median :0.0000   Median : 1.000
              3 :28886                   Mean   :0.4022   Mean   : 2.742
              4+:24994                   3rd Qu.:1.0000   3rd Qu.: 4.000
                                         Max.   :1.0000   Max.   :15.000

Product_Category_2 Product_Category_3    Purchase
Min.   : 2.000     Min.   : 3.00      Min.   :  185
1st Qu.: 2.000     1st Qu.: 9.00      1st Qu.: 7871
Median : 6.000     Median :14.00      Median :11757
Mean   : 6.896     Mean   :12.67      Mean   :11661
3rd Qu.:10.000     3rd Qu.:16.00      3rd Qu.:15627
Max.   :16.000     Max.   :18.00      Max.   :23959
```

ii. Data Cleaning:

We utilized the various functionalities of R to clean, balance and normalize our dataset. Our raw dataset contained 537,577 observations belonging to 12 variables out of which we had 166,986 and 373,299 NA values in 2 of the variables which we replaced with the mean of the respective columns.

a. **Uncleaned Data Set**

User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
1000001	P00069042	F	0-17	10	A	2	0	3	NA	NA	8370
1000001	P00248942	F	0-17	10	A	2	0	1	6	14	15200
1000001	P00087842	F	0-17	10	A	2	0	12	NA	NA	1422
1000001	P00085442	F	0-17	10	A	2	0	12	14	NA	1057
1000002	P00285442	M	55+	16	C	4+	0	8	NA	NA	7969
1000003	P00193542	M	26-35	15	A	3	0	1	2	NA	15227
1000004	P00184942	M	46-50	7	B	2	1	1	8	17	19215
1000004	P00346142	M	46-50	7	B	2	1	1	15	NA	15854
1000004	P0097242	M	46-50	7	B	2	1	1	16	NA	15686
1000005	P00274942	M	26-35	20	A	1	1	8	NA	NA	7871
1000005	P00251242	M	26-35	20	A	1	1	5	11	NA	5254
1000005	P00014542	M	26-35	20	A	1	1	8	NA	NA	3957
1000005	P00031342	M	26-35	20	A	1	1	8	NA	NA	6073
1000005	P00145042	M	26-35	20	A	1	1	1	2	5	15665
1000006	P00231342	F	51-55	9	A	1	0	5	8	14	5378
1000006	P00190242	F	51-55	9	A	1	0	4	5	NA	2079
1000006	P0096642	F	51-55	9	A	1	0	2	3	4	13055
1000006	P00058442	F	51-55	9	A	1	0	5	14	NA	8851
1000007	P00036842	M	36-45	1	B	1	1	1	14	16	11788
1000008	P00249542	M	26-35	12	C	4+	1	1	5	15	19614
1000008	P00220442	M	26-35	12	C	4+	1	5	14	NA	8584

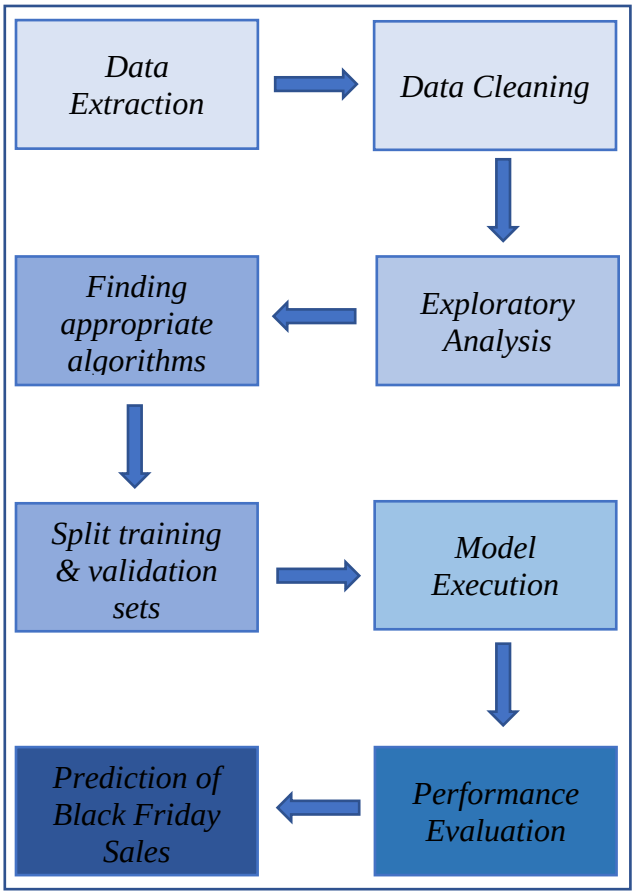
b. **Cleaned Data Set**

User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
1000001	P00069042	F	0-17	10	A	2	0	3	9.842144	12.66984	8370
1000001	P00248942	F	0-17	10	A	2	0	1	6.000000	14.00000	15200
1000001	P00087842	F	0-17	10	A	2	0	12	9.842144	12.66984	1422
1000001	P00085442	F	0-17	10	A	2	0	12	14.000000	12.66984	1057
1000002	P00285442	M	55+	16	C	4+	0	8	9.842144	12.66984	7969
1000003	P00193542	M	26-35	15	A	3	0	1	2.000000	12.66984	15227
1000004	P00184942	M	46-50	7	B	2	1	1	8.000000	17.00000	19215
1000004	P00346142	M	46-50	7	B	2	1	1	15.000000	12.66984	15854
1000004	P0097242	M	46-50	7	B	2	1	1	16.000000	12.66984	15686
1000005	P00274942	M	26-35	20	A	1	1	8	9.842144	12.66984	7871
1000005	P00251242	M	26-35	20	A	1	1	5	11.000000	12.66984	5254
1000005	P00014542	M	26-35	20	A	1	1	8	9.842144	12.66984	3957
1000005	P00031342	M	26-35	20	A	1	1	8	9.842144	12.66984	6073
1000005	P00145042	M	26-35	20	A	1	1	1	2.000000	5.00000	15665
1000006	P00231342	F	51-55	9	A	1	0	5	8.000000	14.00000	5378
1000006	P00190242	F	51-55	9	A	1	0	4	5.000000	12.66984	2079
1000006	P0096642	F	51-55	9	A	1	0	2	3.000000	4.00000	13055
1000006	P00058442	F	51-55	9	A	1	0	5	14.000000	12.66984	8851
1000007	P00036842	M	36-45	1	B	1	1	1	14.000000	16.00000	11788
1000008	P00249542	M	26-35	12	C	4+	1	1	5.000000	15.00000	19614
1000008	P00220442	M	26-35	12	C	4+	1	5	14.000000	12.66984	8584
1000008	P00156442	M	26-35	12	C	4+	1	8	9.842144	12.66984	9872

iii. Dimension Reduction:

Irrelevant attributes can mislead the classifier into building an incorrect model for predicting accuracy. Attributes with unique values such as Product_ID can build a model that would be based on that attribute with unique value and predict with maximum accuracy. It would not help us in analyzing the data set or get any insight. Such an attribute is also called a False Predictor.

IV. Data Mining Techniques and Implementation



After selecting our variables, we selected three algorithms for our experimental design. The three algorithms we selected to run on our dataset are:

- ***Linear Regression***
- ***Random Forest***
- ***k – Means Clustering***

i. Linear Regression:

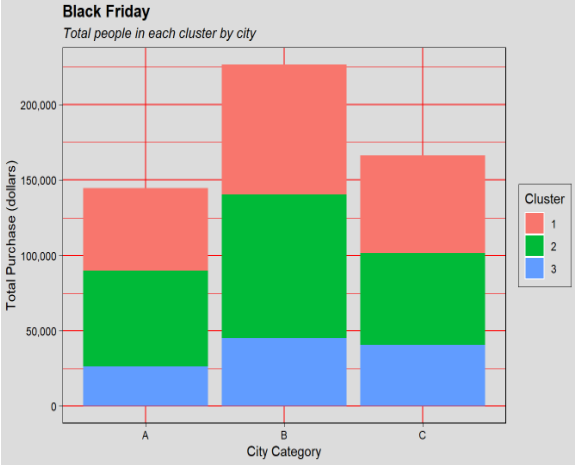
This is a predictive analysis is to study the relationship between one continuous independent variable and two or more independent variables. In this project we are trying to analyze the relationship between the Gender, Age, Occupation, Marital_Status, City_Category, Stay_In_Current_City_Years, Product_Category vs Purchase. This helps us to understand how much the dependent variable will change when we change the independent variables. For instance, a multiple linear regression preformed in this project can tell us how much Purchase amount is expected to increase (or decrease) with a significant increase (or decrease) in Age.

ii. Random Forest:

In this approach we ensemble methods for classification and regression by creating many decision trees. The ideal outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model. For larger data sets, Random Forest is most accurate learning algorithm it produces an accurate and highly efficient results. In this project we have performed Regression Analysis using Random Forest to measure the % variance in order to compare it with the Linear Regression Model.

iii. k – Means Clustering:

Cluster analysis is used to form groups or clusters of similar records based on several measurements made on these records. The key idea is to characterize the clusters in ways that would be useful for the aims of the analysis. In this project we have analyzed the total people in each cluster by city with respect to the total purchase.



V. Performance Evaluation

The accuracy of our Linear Regression model turned out to be **~75%** while the accuracy of our Random Forest model turned out to be **~55%.** The Mean Absolute Percentage Error are are as follows:

	MAPE	
	Training	Validation
Linear Regression	0.6902	0.6908
Random Forest	0.3003	0.3363

VI. Discussion and Recommendation

Based on our analysis, we can conclude that our Linear Regression model provided better results and is a better fit. According to our dataset and overall data mining task, we can state the following:

- The disadvantage of using Random Forest is that it has less accuracy and high sensitivity to outliers compared to Linear Regression.
- The error for both Linear Regression and Random Forest can be boosted by Cross-Validation and Bootstrap Aggregating.
- The insights from this dataset could have been even more meaningful had most of the values not been masked.
- The advantage of Linear Regression was that it has the ability to determine the relative influence of one or more predictor variables to the criterion value.

VII. Summary

This case study shows the prediction of Purchase() using the Linear Regression and Random Forest models. We implemented Linear Regression and Random Forest algorithms to train the data and predict the purchase for the given test data. This prediction would be helpful for retail stores with better positioning and marketing of their products based on the consumer behavior statistics. From our estimation it is evident that the Linear Regression model has a higher accuracy comparatively to the Random Forest model. Variables like Gender, Age, Occupation, Marital Status and Product Category are the most vital for this study.

Appendix

Loading the data and cleaning the data

```
library(readr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(plotly)

## Loading required package: ggplot2

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout

library(ggplot2)
library(purrr)
library(readr)

blackfriday <- read.csv(file.choose())
summary(blackfriday)
colSums(is.na(blackfriday))
blackfriday$Product_Category_2[is.na(blackfriday$Product_Category_2)] <- mean(blackfriday
$Product_Category_2, na.rm = T)
blackfriday$Product_Category_3[is.na(blackfriday$Product_Category_3)] <- mean(blackfriday
$Product_Category_3, na.rm = T)
colSums(is.na(blackfriday))
str(blackfriday)
```

Exploratory Data Analysis

Total Purchase Distribution

```
blackfriday %>%
  group_by(User_ID) %>%
  summarise(total_purchase = sum(Purchase)) %>%
  ggplot(aes(x = total_purchase)) +
  geom_histogram(col = 'black', fill = 'blue', binwidth = 300000, center = 150000) +
                theme_linedraw() +
                theme(panel.background = element_rect(fill = "gainsboro", colour = "white", size =
0.7, linetype = "solid"),
                      plot.background = element_rect(fill = "gainsboro"),
                      panel.grid.major = element_line(size = 0.5, linetype = 'solid', colour = "red"),
                      panel.grid.minor = element_line(size = 0.25, linetype = 'solid', colour = "red"),
                      plot.title = element_text(hjust = 0, face = 'bold',color = 'black'),
                      plot.subtitle = element_text(face = "italic")) +
                labs(x = 'Total Purchase', y = 'Number of Consumers', title = "Black Friday Sales",
                     subtitle = "Distribution of total purchase by consumers") +
                scale_y_continuous(limits = c(0,2000), breaks = c(0,500,1000,1500,2000)) +
                scale_x_continuous(labels = scales::comma)
```

Total Purchase by City

```
blackfriday %>%
  group_by(User_ID, City_Category) %>%
  summarise(total_purchase = sum(Purchase)) %>%
  ggplot(aes(x = total_purchase, group = City_Category)) +
  geom_histogram(aes(fill=City_Category),col = 'black', binwidth = 300000, center = 150000) +
  theme_linedraw() +
  theme(legend.box.background   = element_rect(colour = "black"),
        legend.background = element_rect(fill = "darkgrey"),
        panel.background = element_rect(fill = "gainsboro", colour = "white", size = 0.5, linetype
= "solid"),
        plot.background = element_rect(fill = "gainsboro"),
        panel.grid.major = element_line(size = 0.5, linetype = 'solid', colour = "red"),
        panel.grid.minor = element_line(size = 0.25, linetype = 'solid', colour = "red"),
        plot.title = element_text(hjust = 0, face = 'bold',color = 'black'),
        plot.subtitle = element_text(face = "italic")) +
  labs(x = 'Total Purchase ($)', y = 'Number of Consumers', title = "Black Friday",
       subtitle = "Distribution of total purchasing by consumers") +
  guides(fill=guide_legend(title = "City")) +
  scale_y_continuous(limits = c(0,2000), breaks = c(0,500,1000,1500,2000)) +
  scale_x_continuous(labels = scales::comma)
```

Total Purchase by Gender

```
gender <- blackfriday %>%
  group_by(Gender) %>%
  distinct(User_ID) %>%
  summarise(Total=n())
```

```
plot_ly(gender, labels = ~Gender, values = ~Total, type = 'pie',
        textposition = 'inside',
        textinfo = 'label+percent',
        insidetextfont = list(color = '#FFFFFF'),
        hoverinfo = 'text',
        text = ~paste(Total, 'People'),
        marker = list(colors = colors,
                      line = list(color = '#FFFFFF', width = 4)), showlegend = FALSE) %>%
  layout(title = 'Gender', titlefont = list(size = 18, color = 'black'),
         xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
         yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
```

Total Purchase by Age

```
age <- blackfriday %>%
  group_by(Age) %>%
  summarise(Total=n())
plot_ly(age, labels = ~Age, values = ~Total, type = 'pie',
        textposition = 'inside',
        textinfo = 'label+percent',
        insidetextfont = list(color = '#FFFFFF'),
        hoverinfo = 'text',
        text = ~paste(Total, 'People'),
        marker = list(colors = colors,
                      line = list(color = '#FFFFFF', width = 1)), showlegend = FALSE) %>%
  layout(title = 'Age Distribution', titlefont = list(size = 18, color = 'black'),
         xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
         yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
```

Total Purchase by Marital Status

```
marital_status <- blackfriday %>%
  group_by(Marital_Status) %>%
  summarise(Total=n())
plot_ly(marital_status, labels = ~Marital_Status, values = ~Total, type = 'pie',
        textposition = 'inside',
        textinfo = 'label+percent',
        insidetextfont = list(color = '#FFFFFF'),
        hoverinfo = 'text',
        text = ~paste(Total, 'People'),
        marker = list(colors = colors,
                      line = list(color = '#FFFFFF', width = 1)), showlegend = FALSE) %>%
  layout(title = 'Marital Status', titlefont = list(size = 18, color = 'black'),
         xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
         yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
```

Total Purchase by Occupation

```
customers_total_purchase_amount = blackfriday %>%
  group_by(User_ID) %>%
```

```
  summarise(Purchase_Amount = sum(Purchase))

customers_Occupation =  blackfriday %>%
  select(User_ID, Occupation) %>%
  group_by(User_ID) %>%
  distinct() %>%
  left_join(customers_total_purchase_amount, Occupation, by = 'User_ID')


totalPurchases_Occupation = customers_Occupation %>%
  group_by(Occupation) %>%
  summarise(Purchase_Amount = sum(Purchase_Amount)) %>%
  arrange(desc(Purchase_Amount))

totalPurchases_Occupation$Occupation = as.character(totalPurchases_Occupation$Occupation)
typeof(totalPurchases_Occupation$Occupation)

occupation = ggplot(data = totalPurchases_Occupation) +
  geom_bar(mapping = aes(x = reorder(Occupation, -Purchase_Amount), y = Purchase_Amoun
t, fill = Occupation), stat = 'identity') +
  scale_x_discrete(name="Occupation", breaks = seq(0,20, by = 1), expand = c(0,0)) +
  scale_y_continuous(name="Purchase Amount ($)", expand = c(0,0), limits = c(0, 300000000))
+
  theme(panel.background = element_rect(fill = "gainsboro", colour = "white", size = 0.5, linety
pe = "solid"),
        plot.background = element_rect(fill = "gainsboro"),
        panel.grid.major = element_line(size = 0.5, linetype = 'solid', colour = "red"),
        panel.grid.minor = element_line(size = 0.25, linetype = 'solid', colour = "red"),
        plot.title = element_text(hjust = 0, face = 'bold',color = 'black'),
        plot.subtitle = element_text(face = "italic")) +
  labs(x = "Occupation", y = "Purchase Amount", title = "Black Friday",
       subtitle = "Distribution of total purchasing by occupation")
```

Data Clustering

```
blackfridayclustering <-  blackfriday %>%
                    select(Purchase)
```

Determine the Number of Clusters

```
tot_withinss <- map_dbl(1:10, function(k){
  model <- kmeans(x = blackfridayclustering, centers = k)
  model$tot.withinss
})
```

Generating a data frame containing both k and tot_withinss

```
elbow_df <- data.frame(
  k = 1:10,
  tot_withinss = tot_withinss
)

ggplot(elbow_df, aes(x = k, y = tot_withinss)) +
  geom_line() +
  scale_x_continuous(breaks = 1:10)
```

Cluster Model

```
model_km3 <- kmeans(blackfridayclustering, centers = 3)
clust_km3 <- model_km3$cluster
blackfriday_clust <- mutate(blackfriday, cluster = clust_km3)
blackfriday_clust_node <- blackfriday_clust %>%
                    group_by(cluster) %>%
                    summarise(min_purchase = min(Purchase),
                          max_purchase = max(Purchase),
                          avg_purchase = round(mean(Purchase),0))
```

Determining how many people in each cluster

```
blackfriday_clust %>%
group_by(City_Category, cluster) %>%
summarise(n = n()) %>%
ggplot(aes(x=City_Category, 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"),
      plot.background = element_rect(fill = "gainsboro"),
      panel.grid.major = element_line(size = 0.5, linetype = 'solid', colour = "red"),
      panel.grid.minor = element_line(size = 0.25, linetype = 'solid', colour = "red"),
      plot.title = element_text(hjust = 0, face = 'bold',color = 'black'),
      plot.subtitle = element_text(face = "italic")) +
labs(x = 'City Category', y = 'Total Purchase (dollars)', title = "Black Friday",
```

```
    subtitle = "Total people in each cluster by city") +
guides(fill=guide_legend(title = "Cluster")) +
scale_y_continuous(labels = scales::comma)
```

Linear Regression

```
blackfriday <- blackfriday[,-c(1,2)]
blackfriday <- as.data.frame(unclass(blackfriday))
str(blackfriday)
```

Splitting the data

```
library(caTools)
set.seed(123)
split=sample.split(blackfriday$Purchase, SplitRatio = 0.6)
training_set = subset(blackfriday, split == TRUE)
test_set = subset(blackfriday, split == FALSE)
```

Designing the model

```
regressor = lm(formula = Purchase ~ ., data = training_set)
summary(regressor)
y_pred = predict(regressor, test_set)
prediction_lm = data.frame(cbind(actual = test_set$Purchase, predicted = y_pred))
correlation_accuracy_lm = cor(prediction_lm)
n = sum(correlation_accuracy_lm)
diag = diag(correlation_accuracy_lm)
accuracy_lm = sum(diag)/n
accuracy_lm
mape(training_set$Purchase, y_pred_train)
mape(test_set$Purchase, y_pred_test)
```

Random Forest

```
library(randomForest)
library(Metrics)
library(MLmetrics)
regressor_rf = randomForest(x = training_set[,-10], y = training_set$Purchase, ntree = 100)
summary(regressor_rf)
y_pred_rf = predict(regressor_rf, test_set)
prediction_rf = data.frame(cbind(actual = test_set$Purchase, predicted = y_pred_rf))
correlation_accuracy_rf = cor(prediction_rf)
n = sum(correlation_accuracy_rf)
diag = diag(correlation_accuracy_rf)
accuracy_rf = sum(diag)/n
accuracy_rf
y_pred_train_rf = predict(regressor_rf, newdata = training_set)
rmse(training_set$Purchase,y_pred_train_rf)
mape(training_set$Purchase,y_pred_train_rf)
mape(test_set$Purchase,y_pred_test_rf)
```