Dorcas Munoko

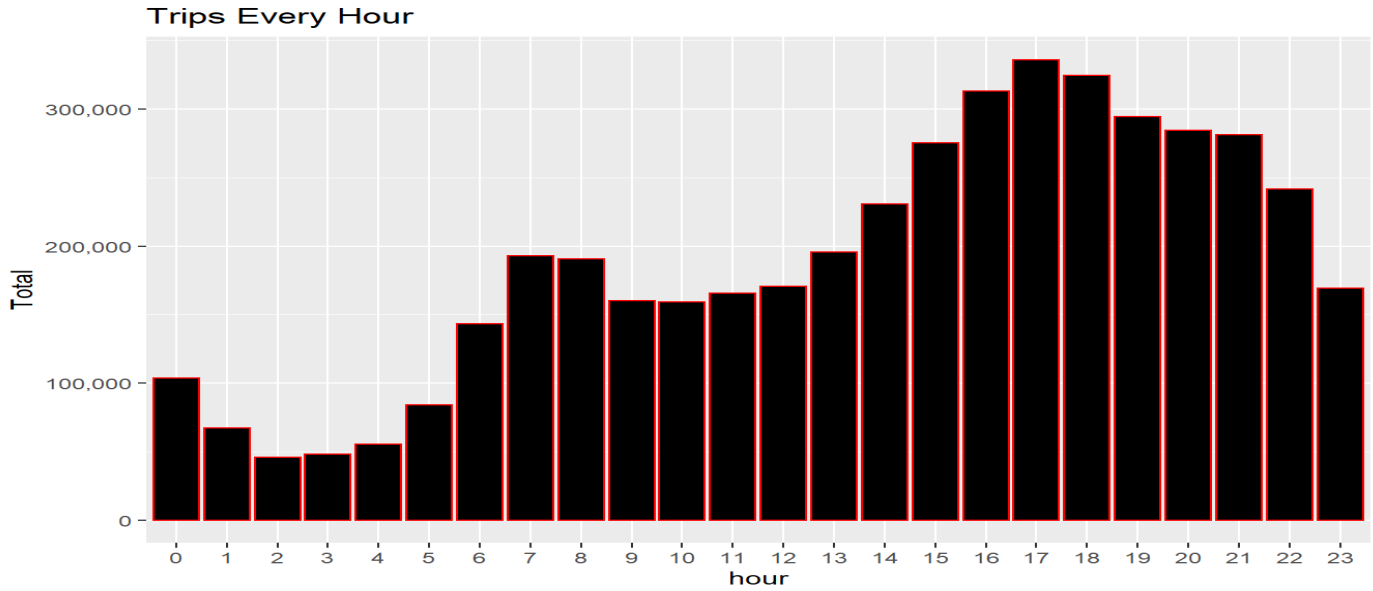
Uber Analysis

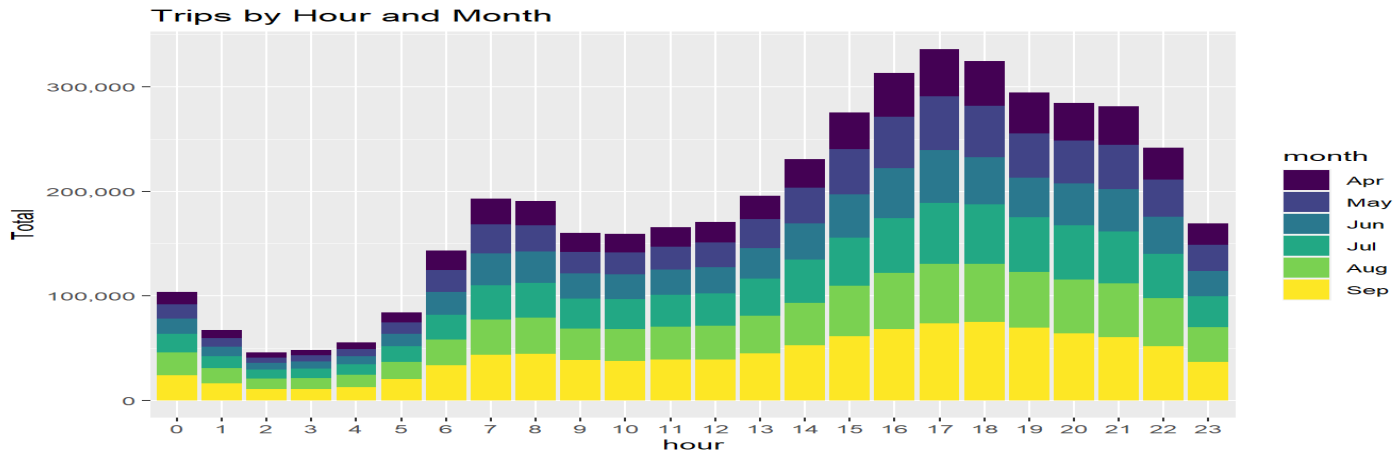
Introduction

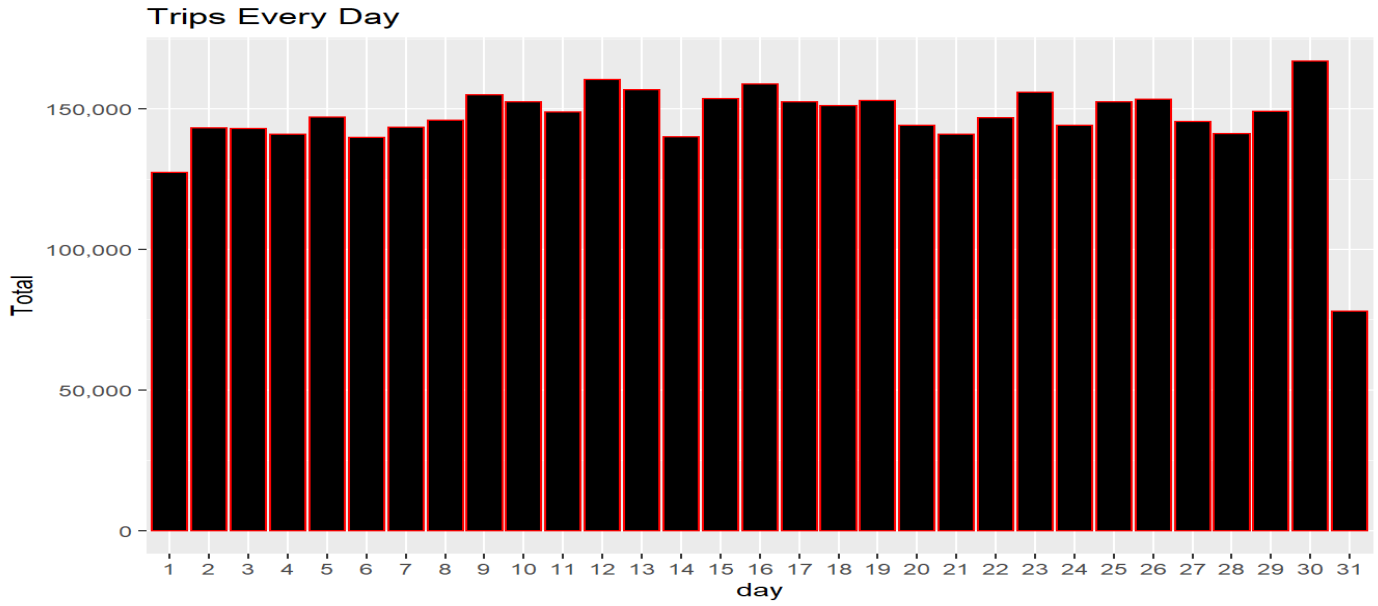
The aim of the project was to analyze Uber pick ups in New York city dataset and create appealing visualization plots. The dataset contained 4,534,327 observations with main features including date and time, latitude, longitude, and base. The first step taken in data preprocessing was to merge 6 dataframes ranging from April to September each containing daily observations. Appropriate formatting of Date.Time column was also conducted to create factors of time objects like day, month, year, hour, minute, and second.

Description of Features

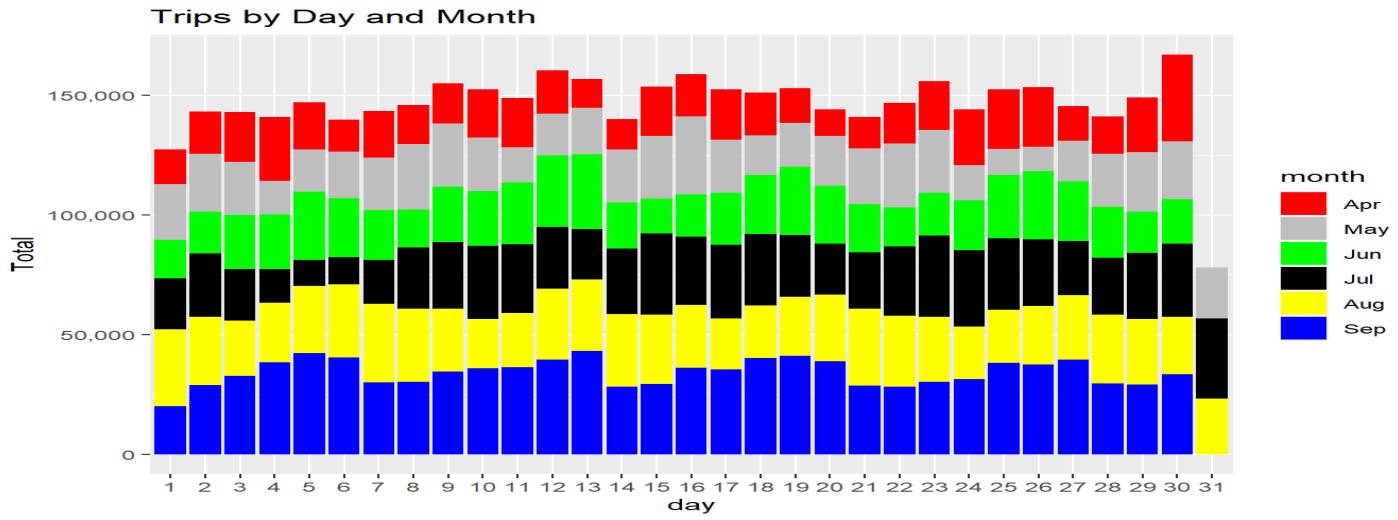
The number of trips per day were plotted. It was observed that the number of trips is higher in the evening around 5:00 and 6:00 PM. Data was also plotted based on every day of the month. It was observed that 30th of the month had the highest trips in the year which was mostly contributed by the month of April.

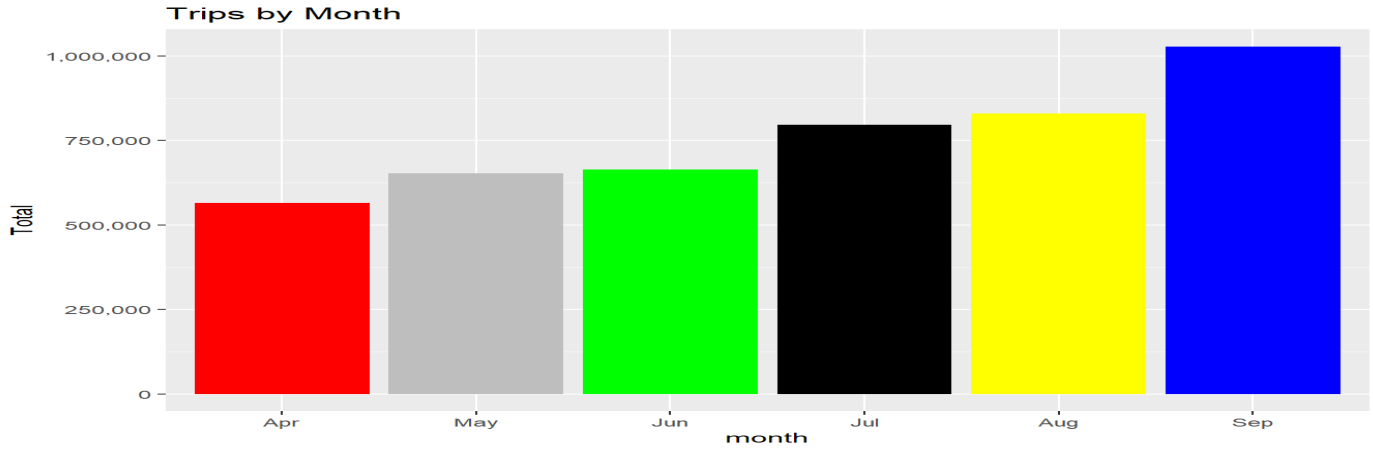


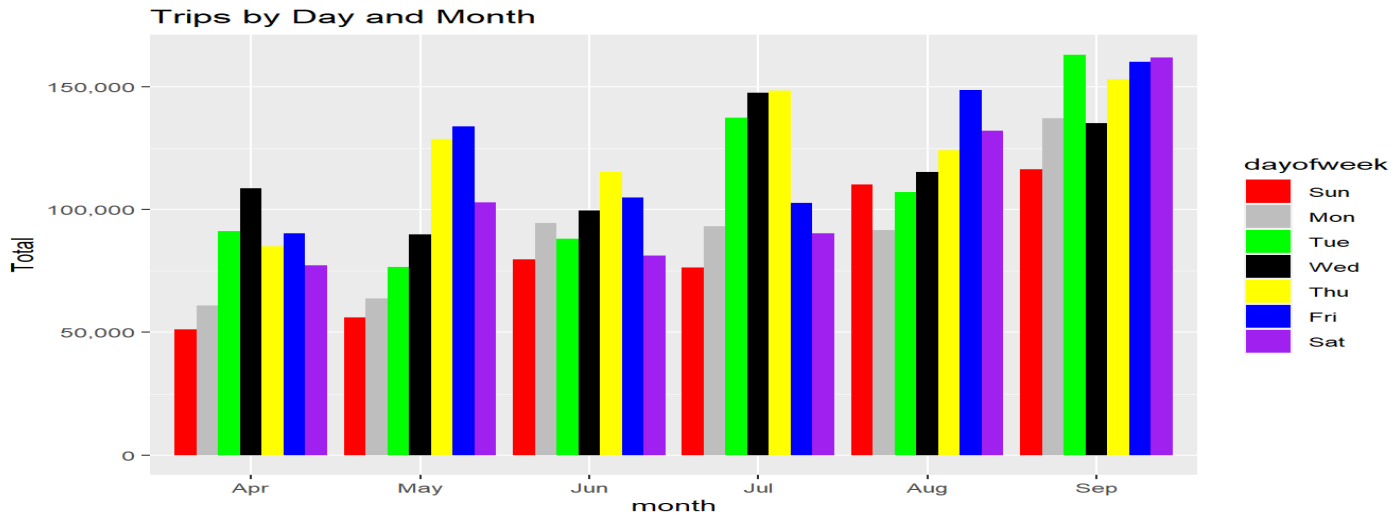




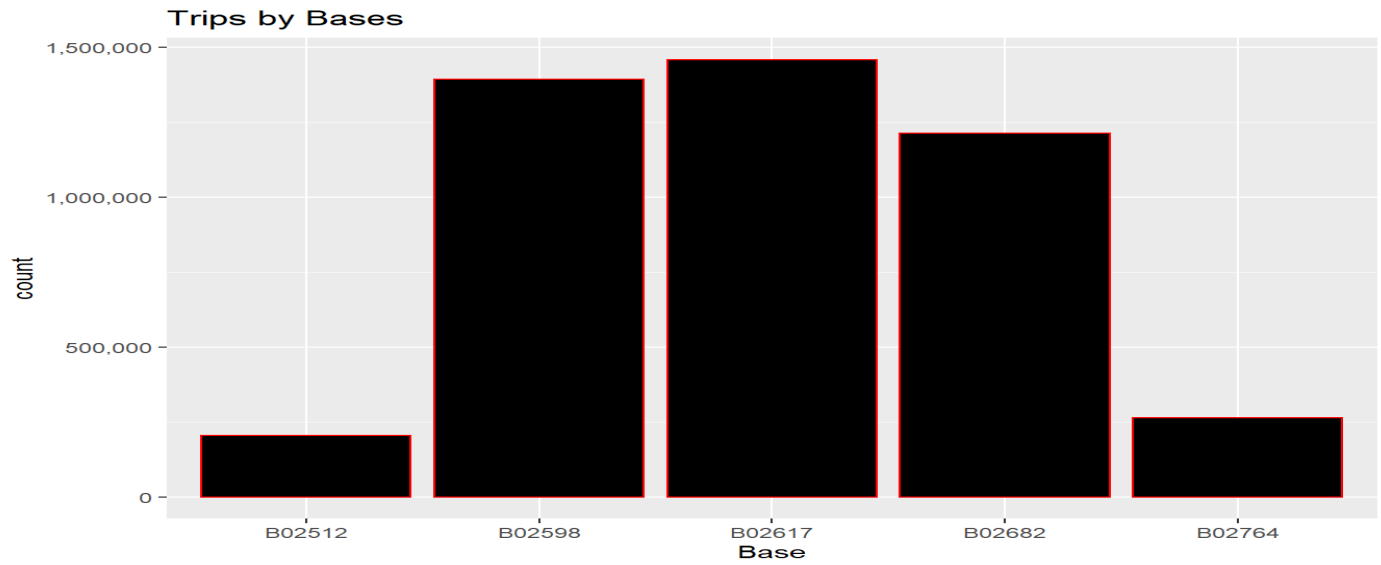
Further analysis was carried out to determine the number of trips that took place each month of the year. It was observed that most trips were made during the month of September. Visual reports of the number of trips that were made on every day of the week were also obtained.

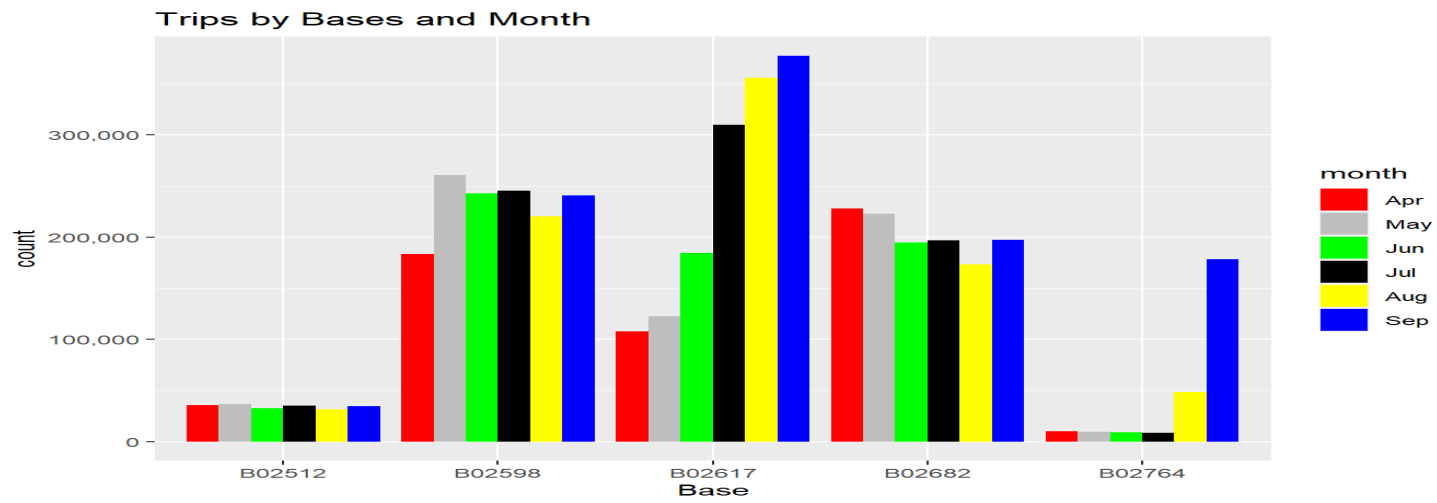


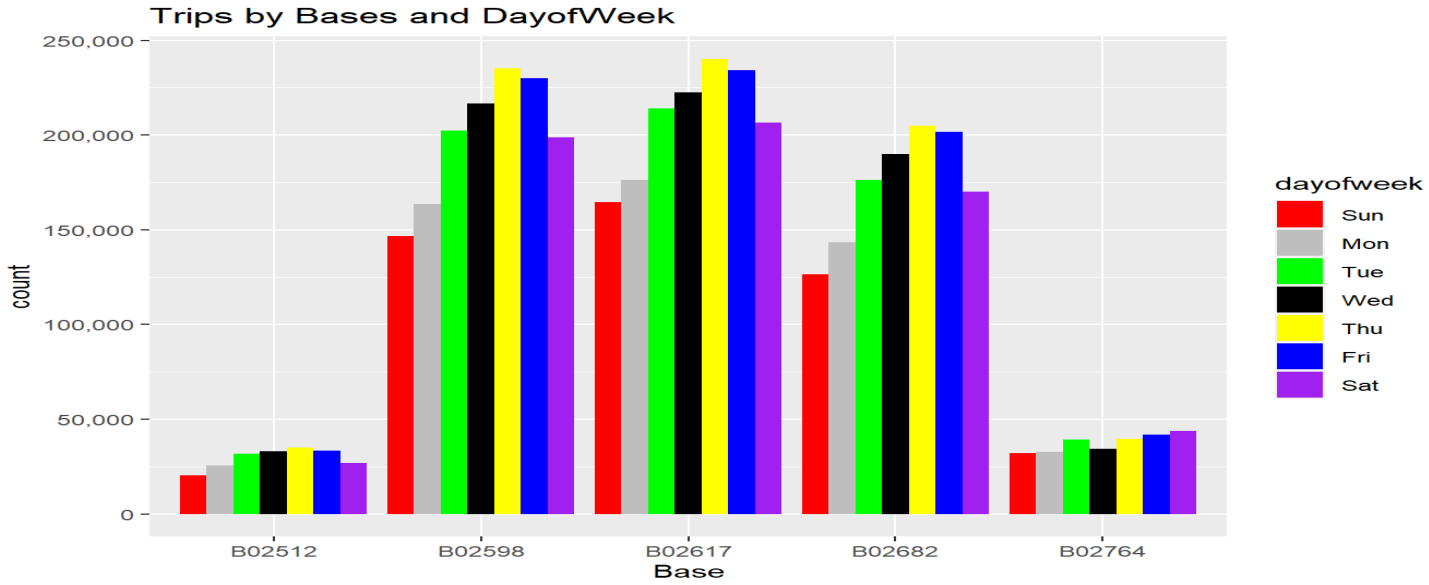




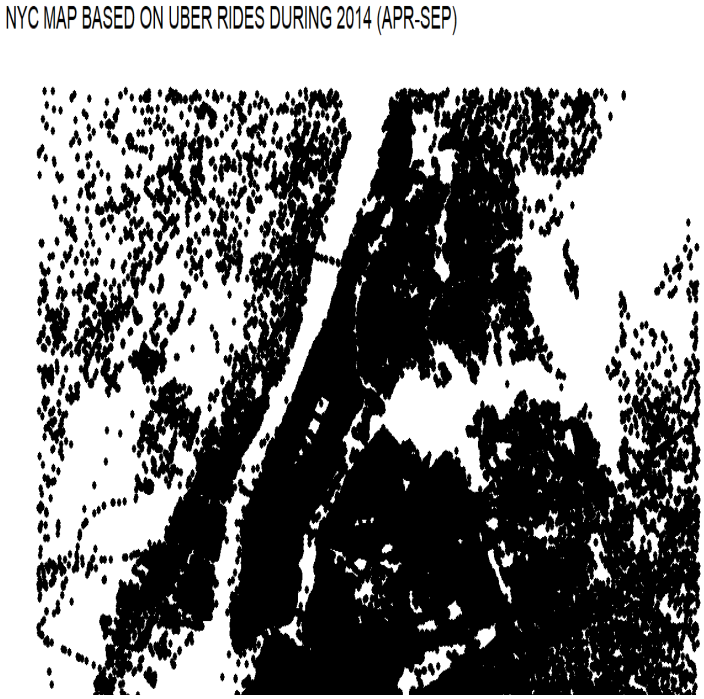
The number of trips that were taken by the passengers from each of the bases were also plotted. It was observed that B02617 had the highest number of trips. Furthermore, this base had the highest number of trips in the month. Thursday observed highest trips in the three bases – B02598, B02617, B02682.

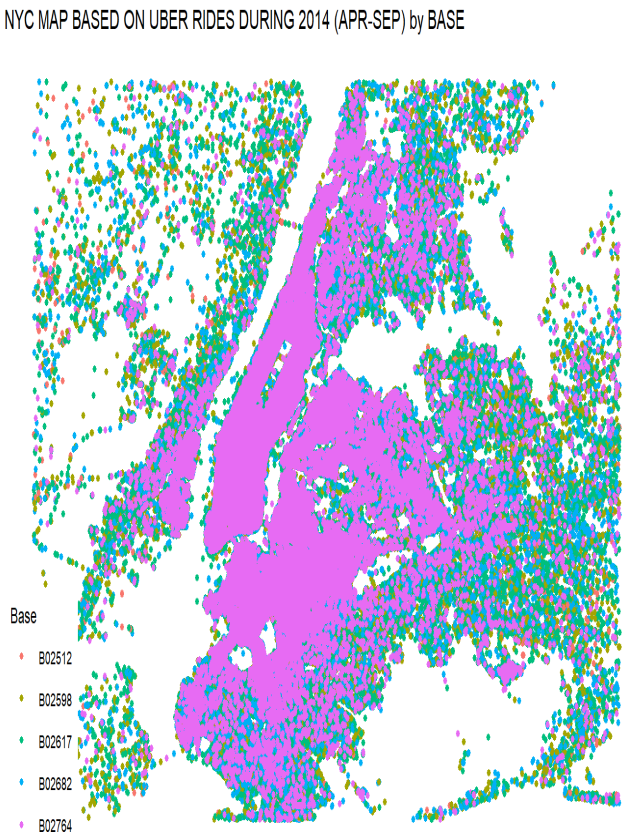






Finally, the rides in New York city were also visualized by creating a geo-plot that visualized the rides during the 2014 (April – September) and by the bases in the same period.





Dorcas Munoko

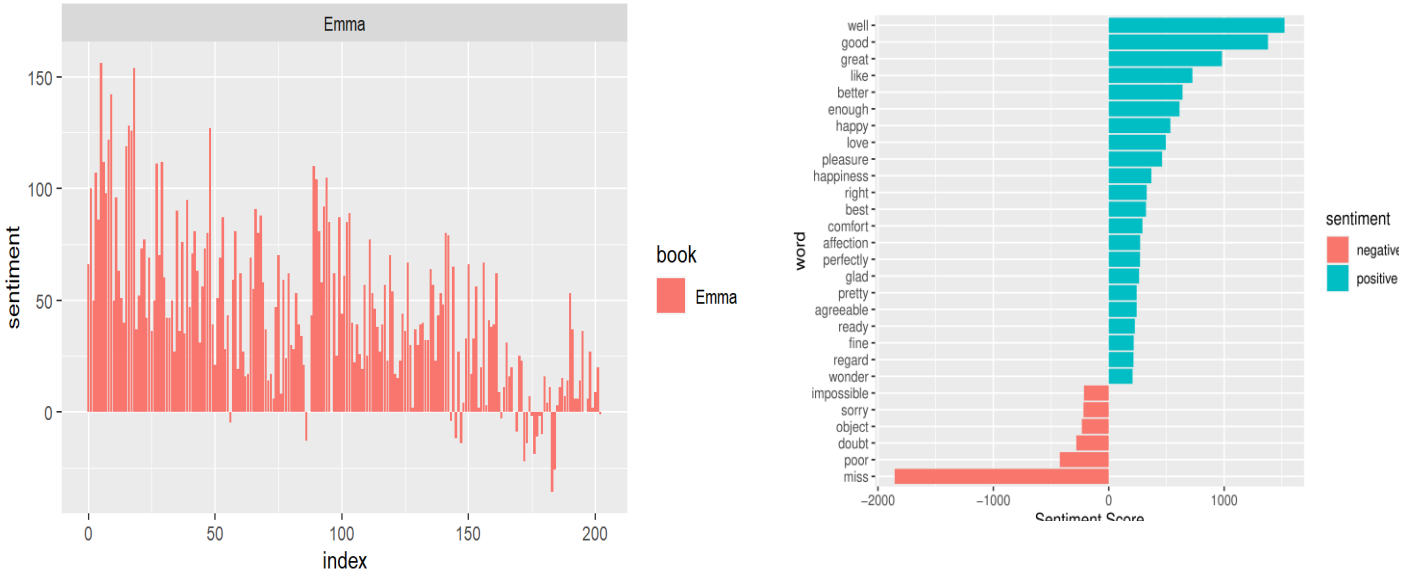
SENTIMENT ANALYSIS

Introduction

Sentiment analysis is a type of classification where data is classified into different classes which are binary in nature or can have multiple classes. It is a process of extracting opinions that have different polarities, that is, positive, negative, or neutral. This is also known as opinion mining and polarity detection. The aim of the project was to build a sentiment analysis model that would allow us to categorize words based on their sentiments, that is, whether they are positive, negative, and the magnitude of it. The dataset used was provided by the R package 'janeaustenR'. This package provides the textual data in the form of books authored by the novelist Jane Austen.

Analysis

The three general purpose lexicons used include AFINN (scores words from a range of -5to5), bing (classifies the sentiment into a binary category of negative or positive), and Loughran (performs analysis of the shareholder's reports). Words present in the book 'Emma' were visualized based on their corresponding positive or negative scores. Sentiment scores were also visualized and plotted along the axis that is labeled with positive and negative words. Finally, a wordcloud was created that would delineate the most recurring positive and negative words.




```
## # A tibble: 6,786 x 2
##    word        sentiment
##    <chr>       <chr>
##  1 2-faces     negative
##  2 abnormal    negative
##  3 abolish     negative
##  4 abominable  negative
##  5 abominably  negative
##  6 abominate   negative
##  7 abomination negative
##  8 abort       negative
##  9 aborted     negative
## 10 aborts      negative
## # ... with 6,776 more rows
```

```
## # A tibble: 668 x 2
##    word          n
##    <chr>     <int>
##  1 well        401
##  2 good        359
##  3 great       264
##  4 like        200
##  5 better      173
##  6 enough      129
##  7 happy       125
##  8 love        117
##  9 pleasure    115
## 10 right        92
## # ... with 658 more rows
```

Dorcas Munoko

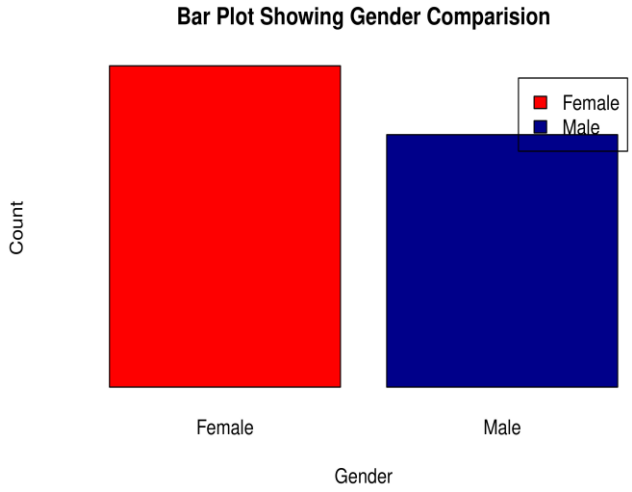
Customer Segmentation Analysis

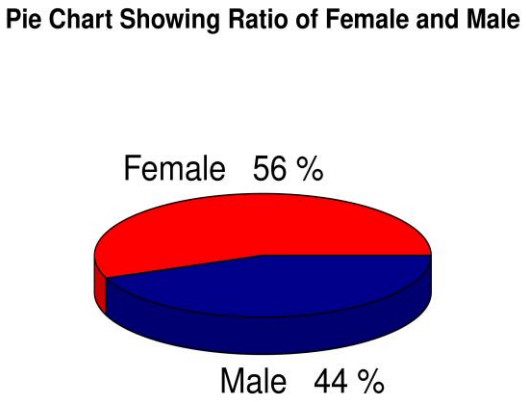
Introduction

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. The dataset contained 200 observations with main features including customer ID, gender, age, annual income, and spending score.

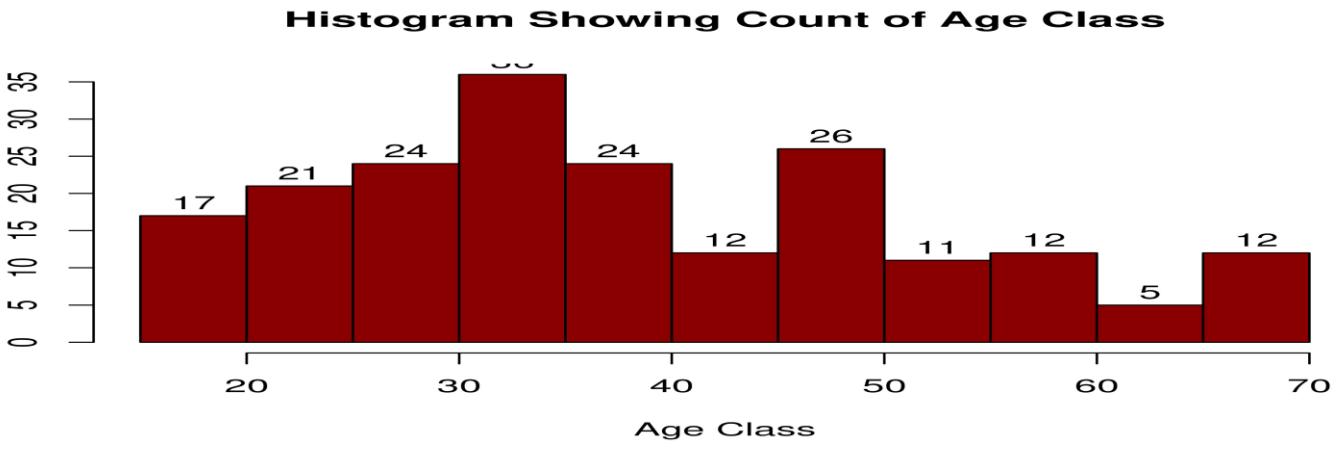
Description of Features

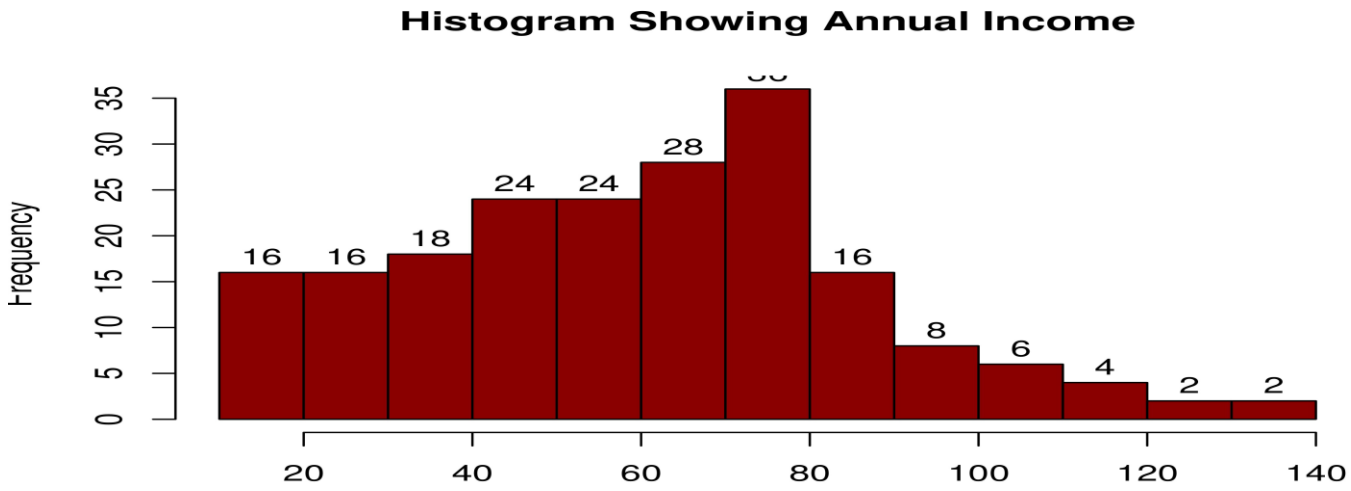
Using bar plot to display gender comparison, it was observed that the number of females was higher as compared to males. Females accounted for 56% of customers in the dataset while male accounted for 44%. The maximum customer ages were between 30 and 35. The minimum age of customers was 18, whereas the maximum age was 70.

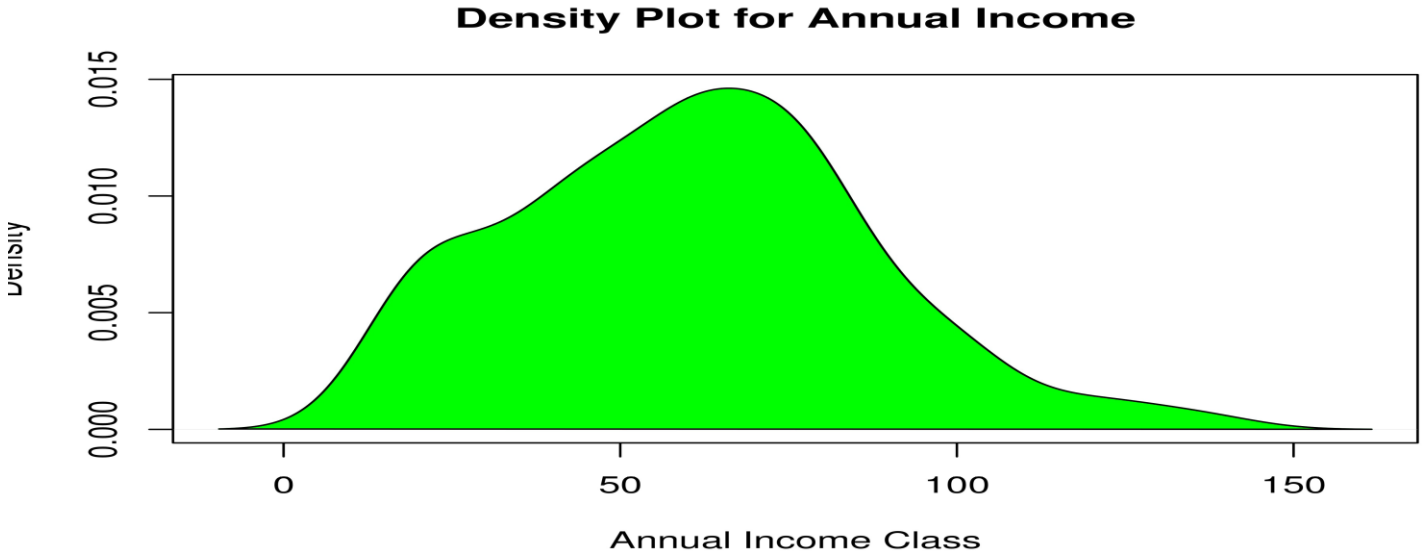




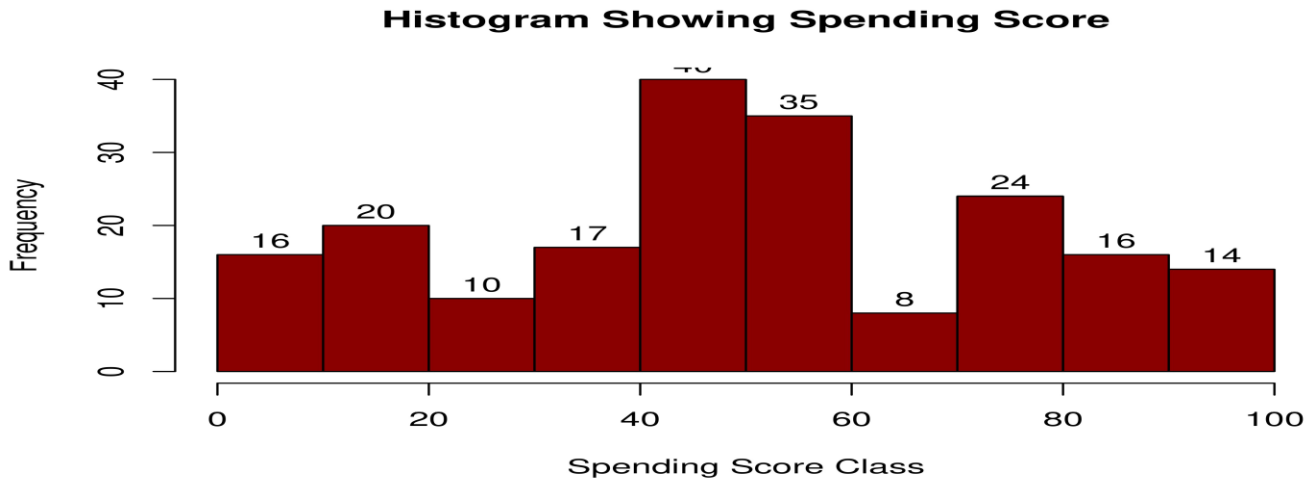
Further analysis was carried out to determine the income of the customers. It was observed that the minimum annual income of the customers was 15 and the maximum income was 137. People earning an average income of 70 had the highest frequency count. The average salary of all the customers was 60.56. The Kernel Density Plot displayed indicates that the annual income has a normal distribution.







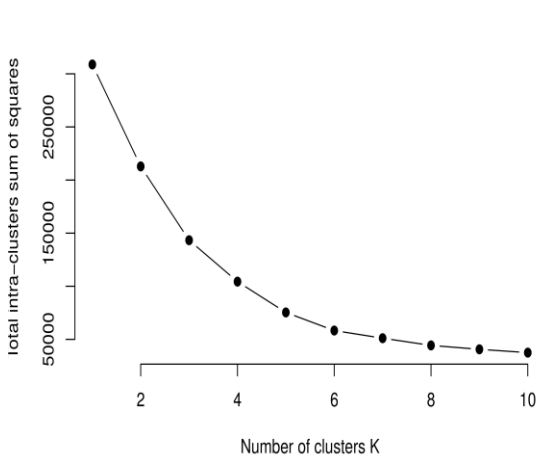
Spending score was also analyzed. It was observed that the minimum was 1, Max was 99 and average was 50.20. From the histogram, we conclude that customers between class 40 and 50 had the highest spending score among all the classes.

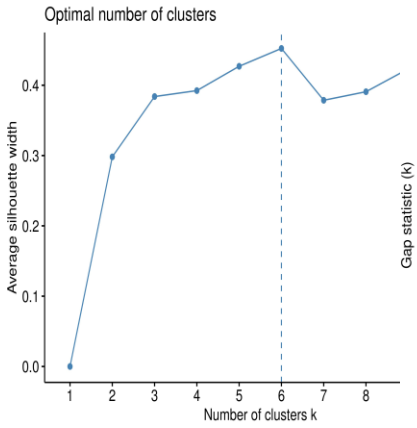


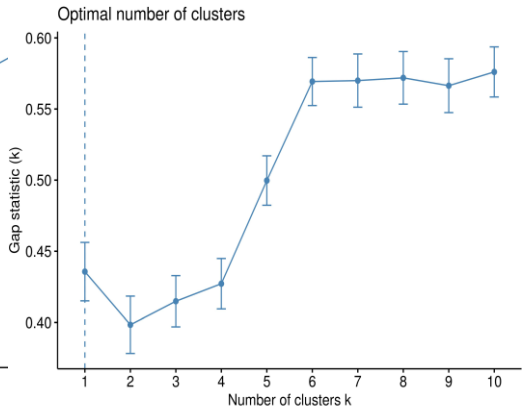
Dorcas Munoko

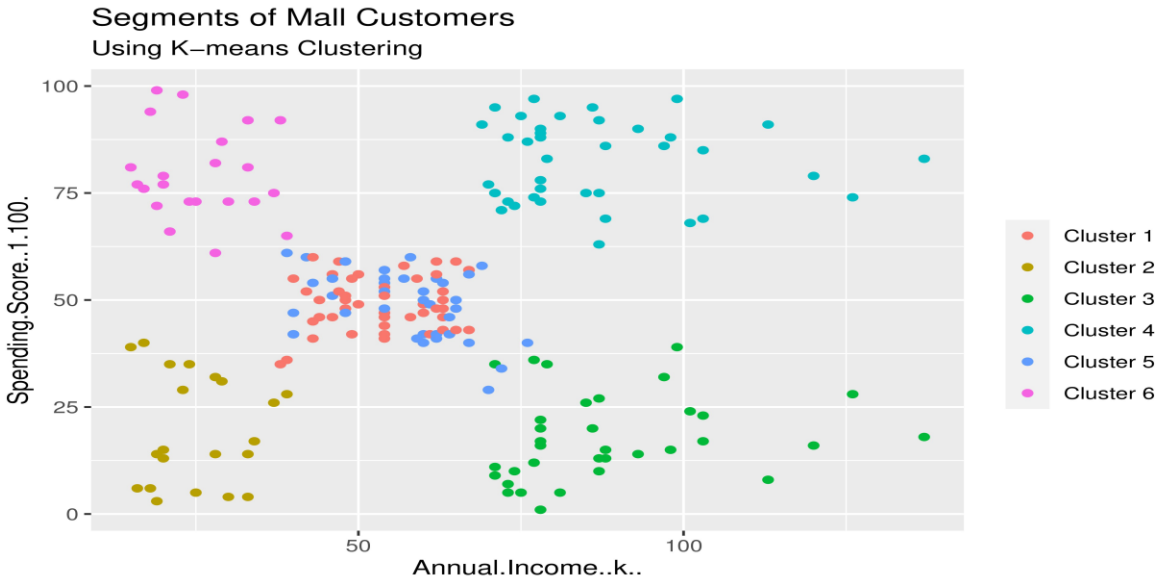
Models, Results, and Conclusion

K-means clustering algorithm was used. The elbow, silhouette, and gap statistic methods were used to determine the optimal clusters. With the elbow method, it was observed that 4 was the appropriate number of clusters since it appeared at the bend in the elbow plot. The average silhouette method determines how well within the cluster is the data object. A high average silhouette means that we have a good clustering. Finally, with the gap statistic method, it was observed that 6 was the appropriate number of clusters.









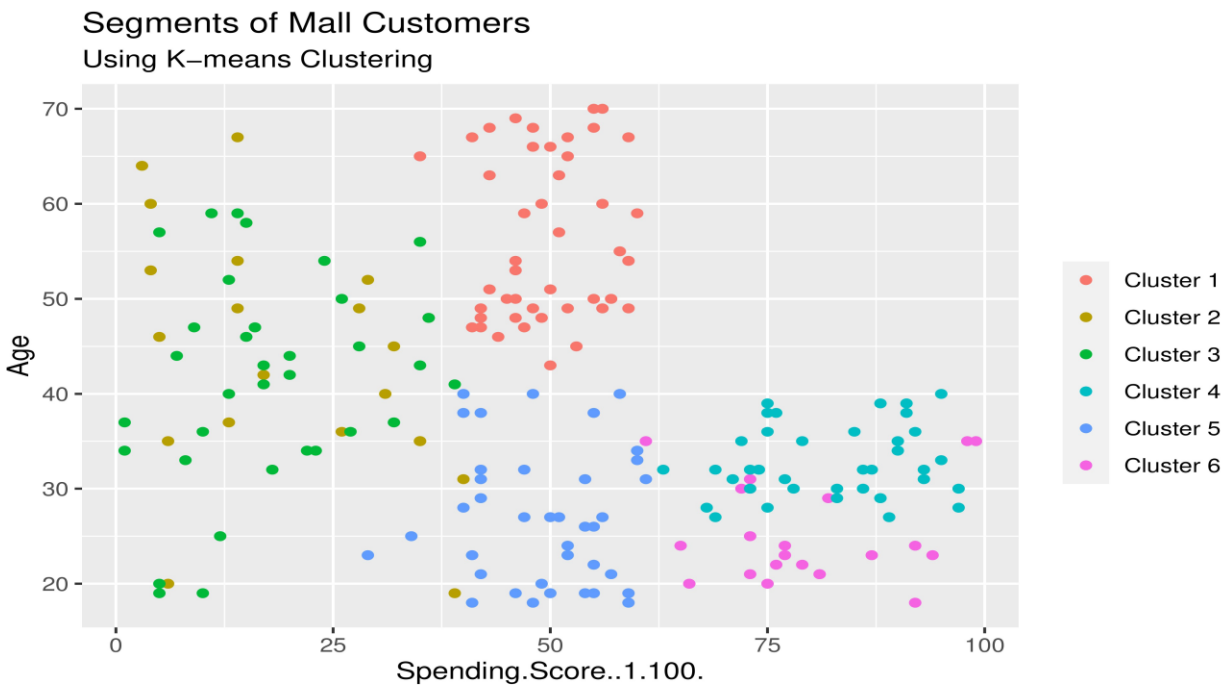
Cluster 6 and 4 – Represent customers with the medium income salary as well as the medium annual spend of salary.

Cluster 1 – Represents customers having a high annual income as well as a high annual spend.

Cluster 3 – Represents customers with low annual income as well as low yearly spend of income.

Cluster 2 – Represents a high annual income and low yearly spend.

Cluster 5 – Represents a low annual income but its high yearly expenditure.



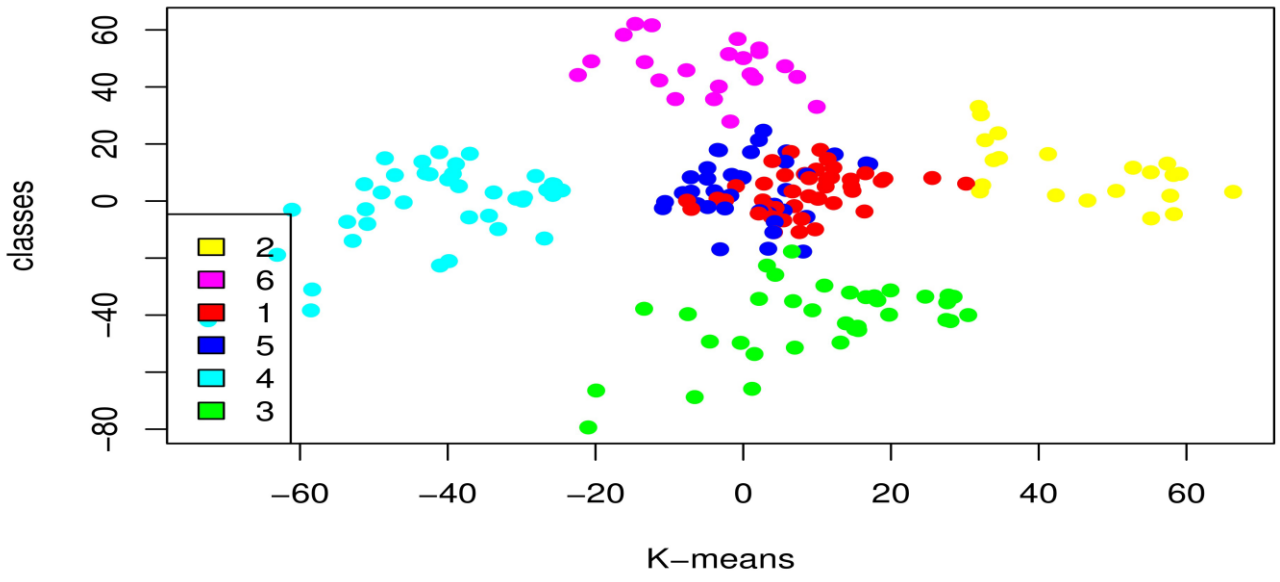
Cluster 4 and 1 – These two clusters consist of customers with medium PCA1 and medium PCA2 score.

Cluster 6 – Represents customers having a high PCA2 and a low PCA1.

Cluster 5 – Represents customers with a medium PCA1 and a low PCA2 score.

Cluster 3 – Represents customers with a high PCA1 income and a high PCA2.

Cluster 2 – Represents customers with a high PCA2 and a medium annual spend of income.

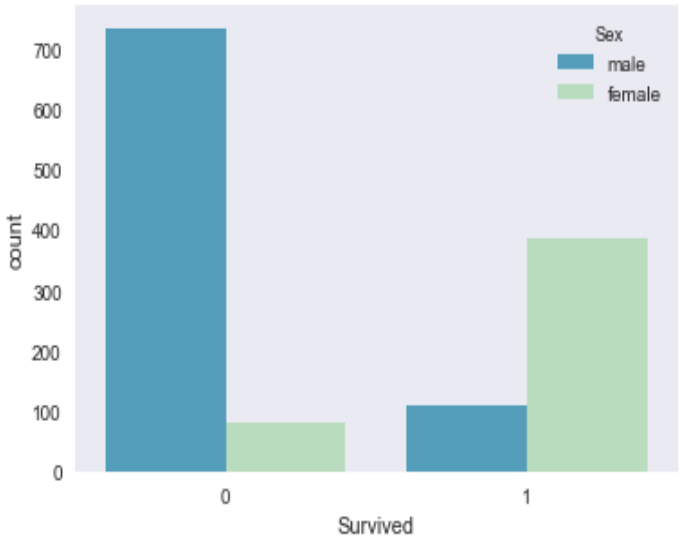


Dorcas Munoko
Big Data and Forecasting
2/12/2022

Introduction

The aim of the project was to build a tree classifier that predicts the probability of death in the Titanic. The train data contains 891 observations while the test data contains 418 observations, with main features including passenger id, survived, passenger class, age, cabin, sibling and spouse, parent and child, fare, sex, and embarked. The first step that I took in data pre-processing was to combine train and test datasets. I then checked for missing values in the combined dataset and replaced them with mean for numeric dataset and mode for categorical dataset. I also dropped features like cabin, name, ticket, and passenger Id, that were poor predictors of survival. I used log to transform the fare data column to uniform data distribution because the data was positively skewed. The last step in pre-processing was converting data type to integer to enable the model to run well.

Description of Features



Third class had the highest number of passengers (700), followed by first class (330), while second class had the lowest number of passengers (290). The ship had more male (843) than female (450), while more passengers embarked from city S (916). Approximately 800 did not survive, majority being male (730). Most of these deaths occurred in third class. I created a correlation matrix, and it indicates that survival had a negative correlation with age, passenger id, and passenger class. On the other hand, it had a positive correlation with fare, parent child, and sibling child.

Model, Results, and Conclusion

I built a decision tree model to estimate probability of survival in the Titanic. The variables I used include age, embarked, parent child, passenger class, sex, and sibling spouse. Sex (57%) and age (21%) were the strongest predictors of survival. I also carried out an accuracy analysis, and the model was 82% accurate in estimating the probability of survival in the Titanic. I carried out a 10-fold cross validation of the model, and it had a score of 82%. I compared the model to a simple probit classifier. Results indicate that the two classifier models had the same accuracy score in predicting the probability of survival (82%).

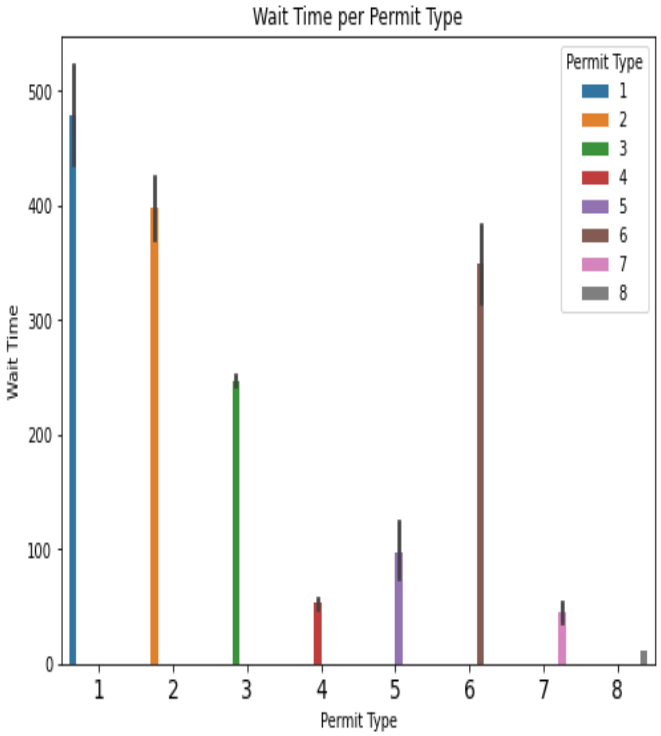
Dorcas Munoko
Big Data and Forecasting

Introduction

The aim of the project was to explore issues that lead to an increase in processing times of building permits in San Francisco. The data contained 198,900 observations with dates ranging from the years 2013-2018. Top features include permit type, filed date, issued date, number of proposed stories, revised cost, proposed units, existing construction type, proposed construction type, supervisor district, zip code, and plansets. The first step I took in data pre-processing was to find out the missing values in the dataset. Since the dataset contained a lot of missing values, deleting them would lead to loss of massive critical data. I created a threshold of columns containing at least 80% missing data to be deleted, dropped columns that were highly colinear, and those that had no effect on our outcome variable.

Description of Features

Permit type 1 had the highest number of wait time (approximately 480 days) as compared to permit type 8 which took approximately 10 days to be issued. Permit type 8 had the highest number of records (with otc alteration as the most popular) followed by permit type 3. The rest (2,4,5,6,7) had very few records. More permits had already been issued (80,000) as compared to revoked and incomplete permits. Construction Type 5 had the highest number of permits. I created a correlation matrix, and it indicates that existing construction and revised costs are positively correlated with the outcome variable while permit type, number of proposed stories, and proposed units are negatively correlated with outcome variable.



Models, Results, and Conclusion

I built different models to explore issues that lead to an increase in processing time of building permits. The features that I used include permit type, filed date, issued date, number of proposed stories, revised cost, proposed units, existing construction type, proposed construction type, supervisor district, zip code, and plansets. The regression tree model had the highest accuracy score (81%), Multinomial Naïve Bayes' (62%), LDA (56%), Gaussian Naïve Bayes' (53%), while Ridge regression had the lowest accuracy score. According to Ridge regression results, the best estimators of a longer wait time when alpha equals 2 are existing and proposed construction types. Finally, existing construction type appears to be the best predictor of delay in issuance of building permits in San Francisco.

Dorcas Munoko
Big Data and Forecasting
2/3/2022

Predicting Top Quality Wine

Introduction

The aim of the project was to provide a classifier that can predict top-quality wines based on the observable characteristics. The data contains 1599 rows and 12 columns. I created data visualizations to know the data and patterns.

Data Preparation and Visualization

My first step was to clean the data to prepare it for analysis. I checked data type and missing values in dataset. Finally, I checked for outliers in each column in the dataset. I created a count plot to determine the number of values for each quality. The quality value starts from 3 to 8, but the number of values is greater for wine qualities 5 and 6. I also created bar plots to determine which values are related to quality. The output results indicate that volatile acidity and chloride are inversely proportional to quality, while citric acid, alcohol, and sulphates are directly proportional to quality. Fixed acidity, residual sugar, free chloride dioxide, and total sulfur dioxide show no significant relationship with quality. I ran a correlation analysis of the variables against quality. The highest correlation is between alcohol and quality (0.48), volatile acidity and alcohol (-0.2), sulphates (0.25), citric acid (0.23), total sulfur dioxide (-0.19), density (0.019), chlorides (-0.041), fixed acidity (-0.021), pH (-0.0011), and free sulfur dioxide (-0.049).

Algorithms, Methodology, and Results

I began by separating out quality data to be used in machine learning. I used the Naïve Bayes classifier technique to classify wine quality. This technique holds an assumption of independence. It assumes that a class feature is unrelated to any other feature. The accuracy results for the three models are 54.3% (Gaussian), 43.6% (Multinomial), 41.5% (Bernoulli). The 10-fold cross-validation score for the models is 56.8% (Gaussian), 44.7% (Multinomial), and 43.7% (Bernoulli). In conclusion, the Gaussian model did great in predicting wine quality. It has the highest accuracy rate (54.3%) and score (56.8%). It is also the best model for continuous data unlike Bernoulli which is best for binary data, and multinomial which is best for categorical data.