Analyzing Cab Data to Streamline Transportation

Pranjali Deshmukh
Computer Science
Illinois Institute Of Technology
Chicago, USA
pdeshmukh3@hawk.iit.edu

Naman Rajendra Jangid
Computer Science
Illinois Institute Of Technology
Chicago, USA
njangid@hawk.iit.edu

Akshay Jain
Computer Science
Illinois Institute Of Technology
Chicago, USA
ajain91@hawk.iit.edu

Sarvesh Shroff
Computer Science
Illinois Institute Of Technology
Chicago, USA
sshroff3@hawk.iit.edu

CSP 571 - Data Preparation and Analysis

Professor: Jawahar Panchal

Illinois Institute of Technology

Index

1. ABSTRACT **2**

2. INTRODUCTION **2**

3. PROBLEM STATEMENTS **3**

4. DATA PROCESSING **3**

4.1 Data Sources 3

4.2 Issues in data and changes made 4

4.3 Observations 5

4.4 Final Data Description 5

5. EXPLORATORY DATA ANALYTICS **6**

5.1 Visualizations 6

6. DATA MODELLING **15**

6.1 Linear Regression Model 15

6.2 Decision Tree 16

7. MODEL EVALUATION **19**

8. OBSERVATIONS & CONCLUSION **20**

9. PLANNED FUTURE WORK **21**

10. REFERENCES **21**

➢ABSTRACT

During the past few years, Uber and Lyft have dominated the ride-sharing industry. The secret to Uber and Lyft success is convenience. They provide prompt, affordable transportation services to any area upon request from their clients. The rates for each trip vary depending on the type of service, travel distance, pickup location, etc., making it clear that each firm has a unique pricing strategy. In this project, we'll use datasets from Uber and Lyft to learn more about the dynamics of pricing and the differences between the two companies' approaches to setting prices. This project will teach us more about a range of advanced data visualization tools that may be used to identify patterns in the data of this enormous organization. We will first prepare the data by processing it, then use several data analysis visualization techniques, before training the model to predict prices.

➢INTRODUCTION

In modern cities, personal transportation has become an essential feature of daily life, providing convenience, speed, and reliability to urban dwellers. Taxicab services have been a dependable mode of transportation for many years, but in recent years, new online ride-sharing services such as Uber, Lyft, and others have emerged and are beginning to gain momentum, threatening to displace traditional taxi services. The market expansion of these ride-sharing services has been quite domineering due to several factors.

One of the primary reasons people are switching to online ride-sharing services is the convenience they offer. Traditional taxis cannot be reserved for a specific time, and you must phone in advance to ensure it arrives on time. In contrast, online taxi services allow you to schedule your pick-up in advance, increasing the driver's dependability as they are closer to the pickup location when you need it. Moreover, online taxi services tend to be less expensive than traditional taxis, with predetermined flat rates that do not increase even if the travel takes longer than anticipated. This makes it easier for passengers to estimate the cost of their ride and budget accordingly. Additionally, the "share my ETA" feature, which allows others who have been granted access by the passenger to view the precise route that is being taken in real-time, enhances safety and peace of mind. As online ride-sharing services operate primarily online, they generate vast amounts of data daily, and this data can be analyzed to evaluate pricing dynamics and differences between various ride-sharing services. In this project, a basic dataset of Uber and Lyft trips from late 2018 was preprocessed to remove irrelevant data and increase the data's dependability. The preprocessed data was then used to build a model that analyzed the pricing dynamics and differences in special rates between Uber and Lyft, providing insights that could be useful in the transportation industry. Thus, as the ride-sharing services continue to expand and gain momentum, understanding their dynamics and differences can be crucial to their continued growth and success.

➢PROBLEM STATEMENTS

Some of the problem statements that we are going to answer in this modeling evaluation are as follows-

- Are any features correlated with one another?
- Which features are most relevant?
- Comparison of performance among models.
- The most rides are offered by which company?
- Most common pick-up and drop locations?
- What time of day do most rides happen?
- What are the cheapest and most expensive ride prices?
- How much does each route generally cost?
- How are rides affected by the weather?

➢DATA PROCESSING

○ Data

Kaggle Dataset Link

We will analyze a dataset comprising of Uber and Lyft trips that occurred between November 26, 2018, and December 18, 2018, to gain insight into the factors that influence dynamic pricing and to compare the special prices offered by Uber and Lyft. The dataset used in this project has been sourced from Kaggle and consists of 693,071 rows and 57 columns.

```
60 ```{r}
61 #Checking the shape of the dataset
62 row = nrow(cabDataSet)
63 col = ncol(cabDataSet)
64 sprintf("The rows and colums are: %s %s",row,col)
65 ```
```

```
[1] "The rows and colums are: 693071 57"
```

○ Data Cleaning and Data Pre-processing

We have undertaken the following the data pre-processing steps in order to get reliable results and run the model.

- The dataset contains numerous missing values and column label inconsistencies across various files.
- The source and destination names have a high number of missing values, which could affect the accuracy of the analysis.
- Different files have varying date-time formats, making it challenging to merge and analyze the data.
- The same features have different data type values in different files, potentially leading to discrepancies when integrating and analyzing the data.
- Changes made:
    - Used started time and ended time of the trip to calculate trip duration and added it wherever it was missing.
    - Formatted date time and other feature values in all files to have a single standard format.
    - Added additional columns with Day of the week from the date field
    - Renamed columns among all files to common labels
    - Dropped features that were not present among all files (SunriseTime, SunsetTime, MoonPhase etc)
    - Used values of latitude longitude to populate missing source and destination values among the files.

- Weather Data:
    - The date format is different from the datasets.
- Changes made:
    - Date format has been changed to reflect the same format as dataset.

```{r}
#See whether missing values or not
sapply(cabDataSet, function(x) sum(is.na(x)))
cabDataSet_distinct <-  na.omit(cabDataSet)
cabDataSet_distinct <- cabDataSet_distinct %>% distinct()
print(paste("The number of records removed : ", nrow(cabDataSet) - nrow(cabDataSet_distinct)))
```

```
                        id                 timestamp                      hour                       day                     month 
                         0                         0                         0                         0                         0 
                  datetime                  timezone                    source                   destination                  cab_type 
                         0                         0                         0                         0                         0 
                product_id                      name                     price                  distance           surge_multiplier 
                         0                         0                     55095                         0                         0 
                  latitude                 longitude               temperature         apparentTemperature             short_summary 
                         0                         0                         0                         0                         0 
              long_summary            precipIntensity          precipProbability                  humidity                 windSpeed 
                         0                         0                         0                         0                         0 
                 windGust              windGustTime                 visibility           temperatureHigh       temperatureHighTime 
                         0                         0                         0                         0                         0 
           temperatureLow         temperatureLowTime     apparentTemperatureHigh   apparentTemperatureHighTime      apparentTemperatureLow 
                         0                         0                         0                         0                         0 
 apparentTemperatureLowTime                      icon                  dewPoint                  pressure                 windBearing 
                         0                         0                         0                         0                         0 
                cloudCover                   uvIndex               visibility.1                     ozone                 sunriseTime 
                         0                         0                         0                         0                         0 
                sunsetTime                 moonPhase          precipIntensityMax               uvIndexTime             temperatureMin 
                         0                         0                         0                         0                         0 
       temperatureMinTime             temperatureMax         temperatureMaxTime      apparentTemperatureMin  apparentTemperatureMinTime 
                         0                         0                         0                         0                         0 
    apparentTemperatureMax  apparentTemperatureMaxTime 
                         0                         0 
[1] "The number of records removed :  55095"
```

○ Observations

- Collecting and manipulating data is a time-consuming and effort-intensive process, particularly when identifying discrepancies in file formats, column names, and missing data. In this project, a significant amount of time has been dedicated to this task.
- While working with the current dataset, we did not experience any issues with the laptops we were using. However, as the dataset size grows, it becomes crucial to optimize the code for data manipulation. To achieve this, we rewrote the script using pandas and treated the columns as vectors.

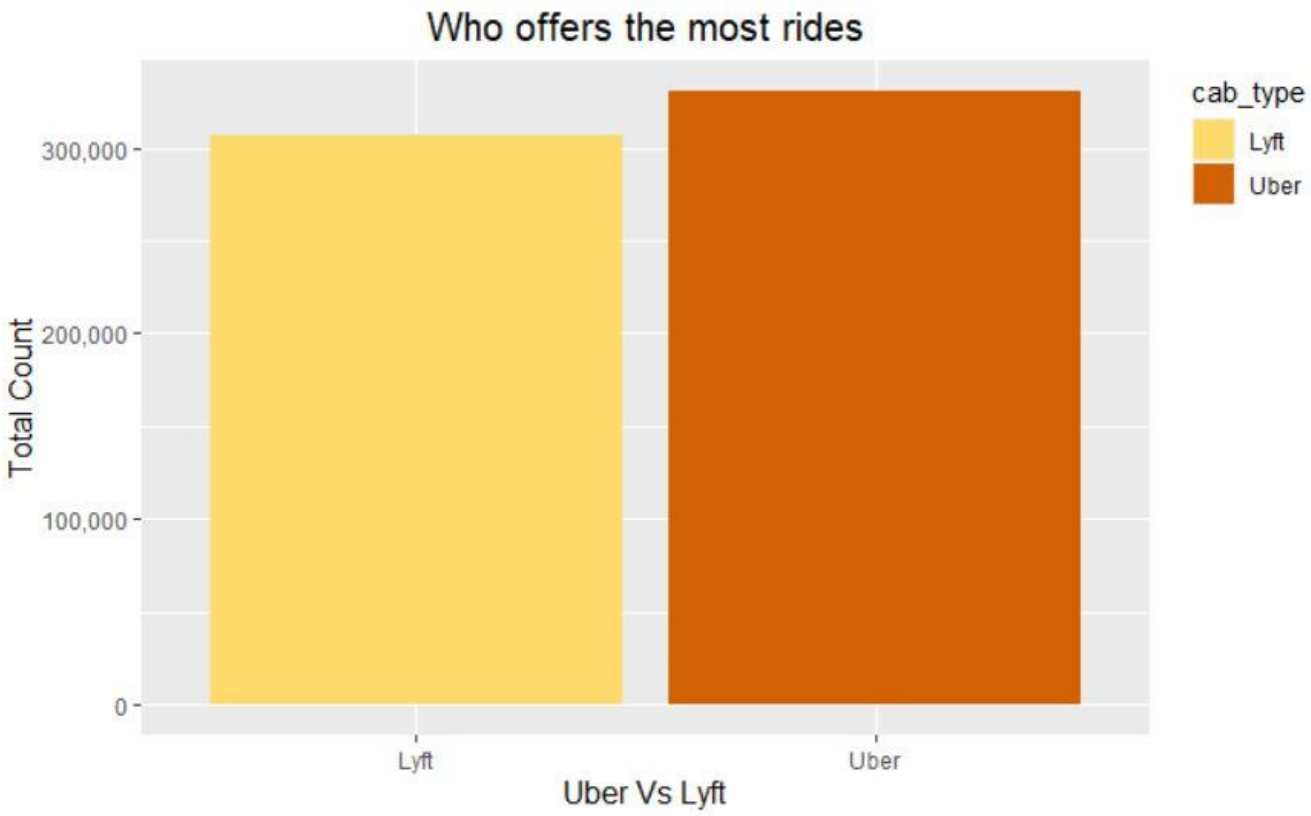
○ Final Data Description

- We ended up with the following 23 features after all of the data cleansing and preparation.

Features selected	
distance	UberPool
surge_multiplier	UberXL
Fri	Black
Sat	Black SUV
Sun	WAV
Shared	Possible Drizzle
Lyft XL	Rain
Lux Black XL	Partly Cloudy
LUX	Overcast
Lux Black	Light Rain
Mostly Cloudy	Foggy
Drizzle	

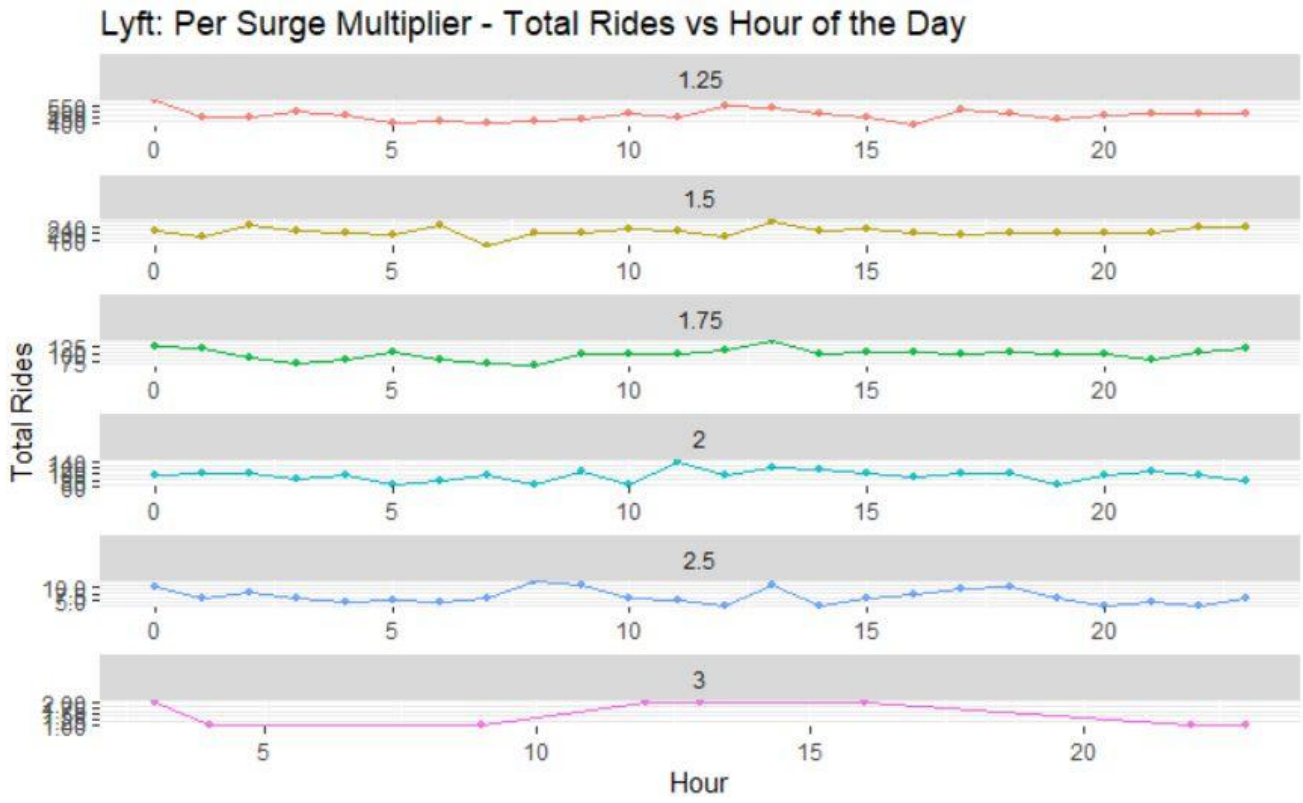
➢EXPLORATORY DATA ANALYTICS

- **Visualizations**
  - Who offers the most rides, Uber or lyft?



As we can observe from the above graph more than 50% of the rides were for uber and the rest are from lyft.

- Lyft: Per Surge Multiplier - Total Rides vs Hour of the Day

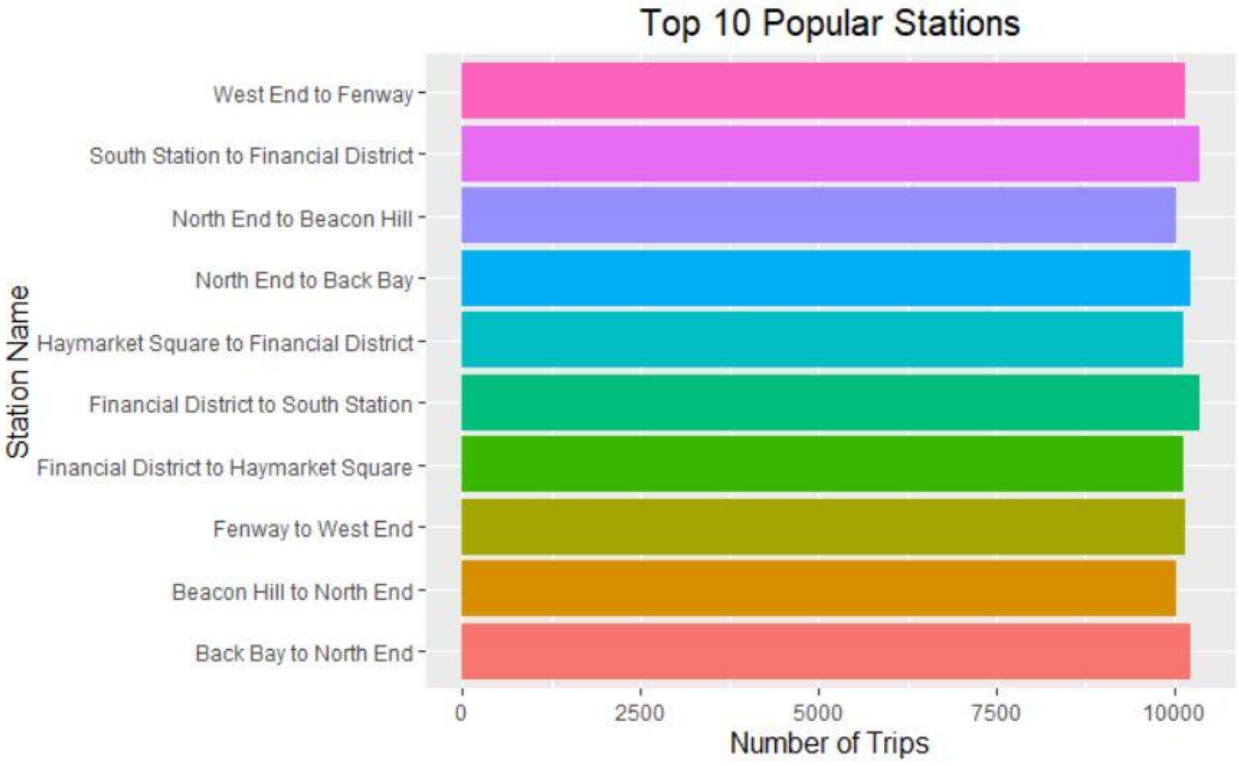


The figure analyzed the total count of Lyft rides for each day of the week along with their corresponding surge multiplier. Irrespective of the surge multiplier, the maximum number of rides were recorded on Tuesday and Wednesday.

- Top 10 most Popular Stations

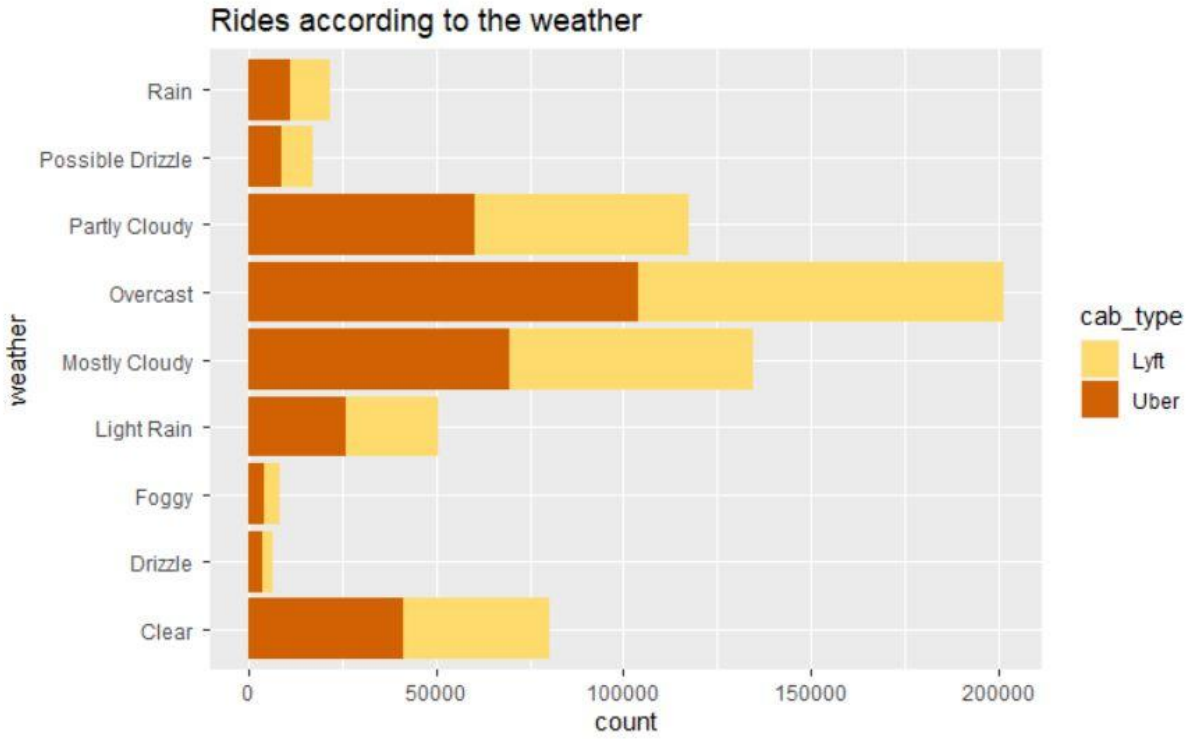
The below plot examines the maximum number of times a cab was booked (Uber and Lyft) for a particular source to a particular destination.

As we can see from the graph, the trip from Financial District to South Station has the highest number of cab bookings.



- Weather affects the rides

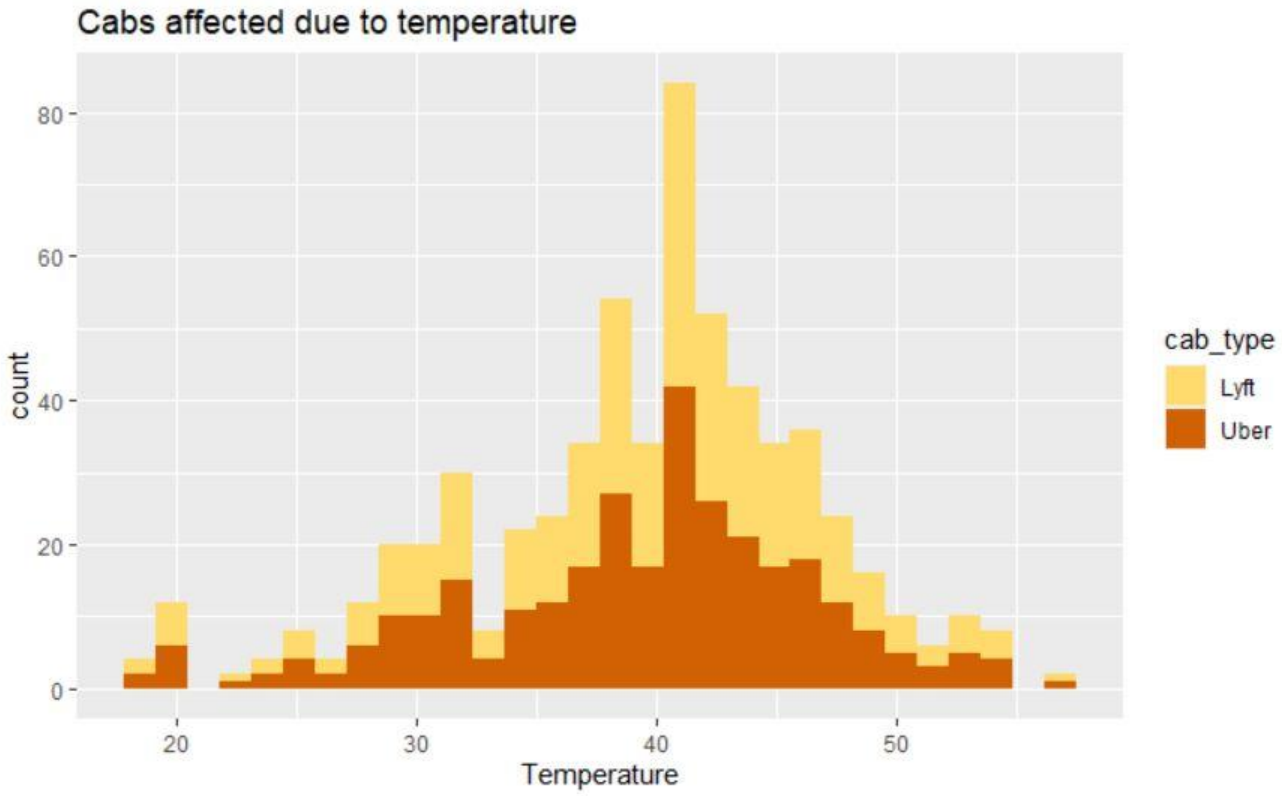
  There is a clear correlation between weather conditions and the number of rides. The count of rides during overcast, clear, mostly and partly cloudy weather conditions is significantly different from those during rainy, foggy, possible drizzle, and drizzle conditions. The number of rides during overcast, clear, mostly and partly cloudy conditions is significantly greater compared to those during rainy, foggy, possible drizzle, and drizzle conditions.



The corpus of the column named “long summary” is as follows-

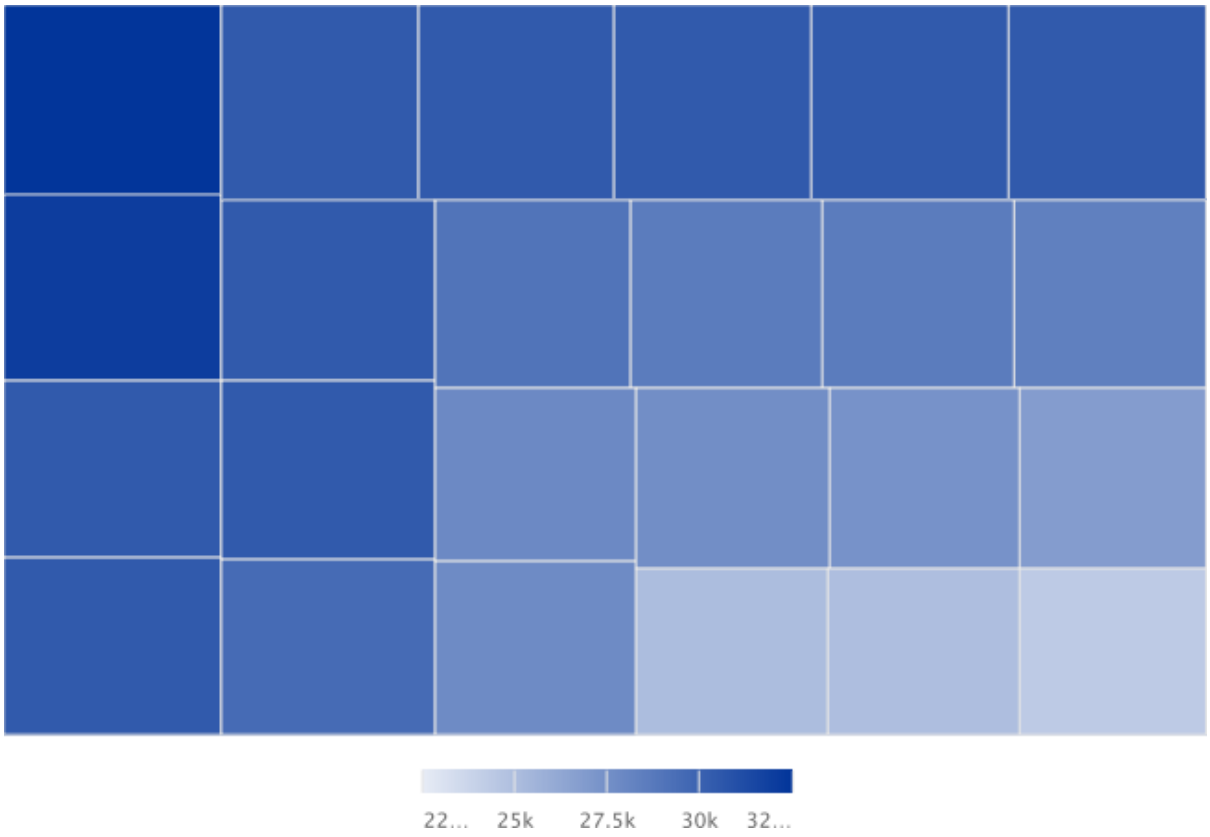
- Temperature affects the ride's price

  Similar to the above graph and comparison, the number of rides fluctuate with the temperature. This fluctuation is realized in the graph below.



- Time division on basis of hour

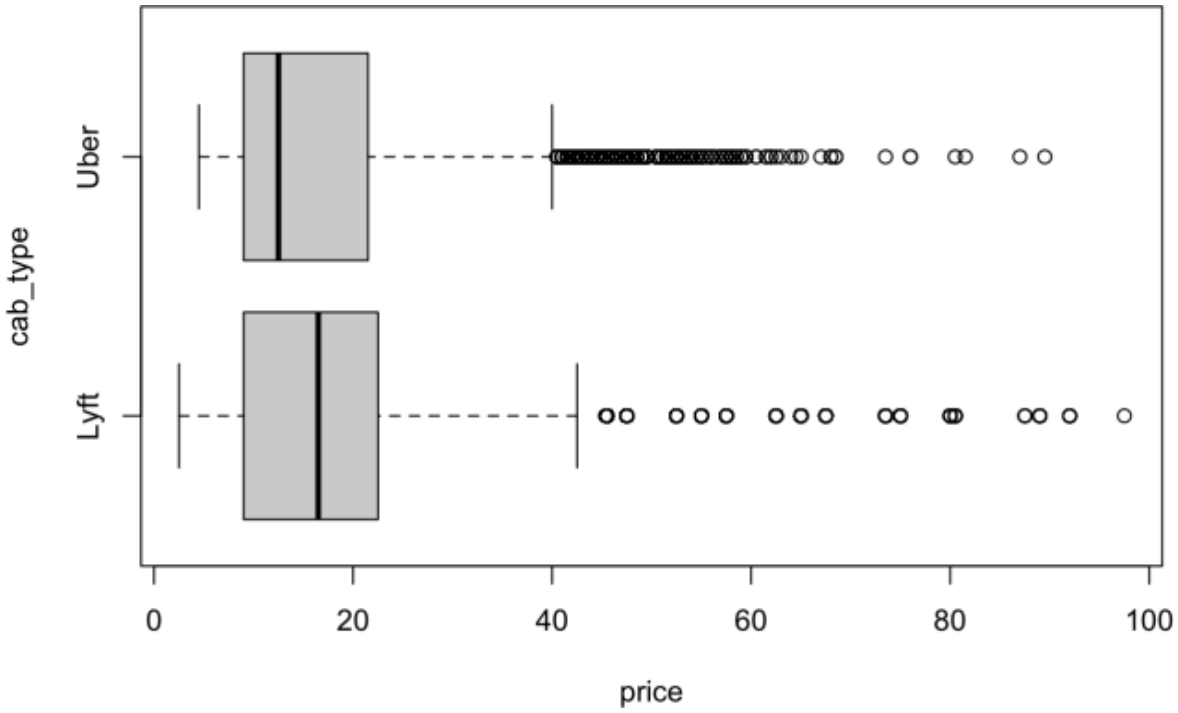
  The below Treemap shows the cab booking interval in a 24 hour time period. Darker the shade of blue, more the number of rides were booked in that 1 hour interval.



- Price range between Uber and Lyft

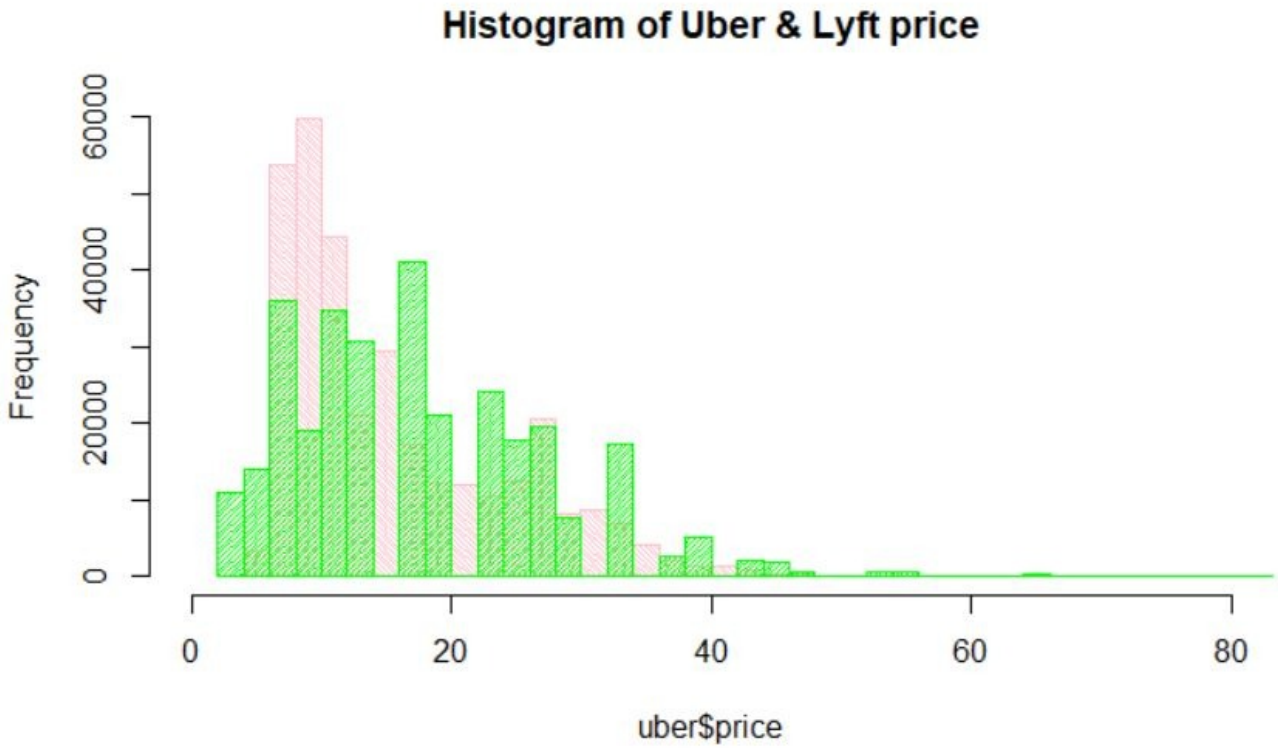
  The Price range for Uber and Lyft rides were then plotted, using a Boxplot and a Histogram. The results of the plottings are as follows-

- Boxplot



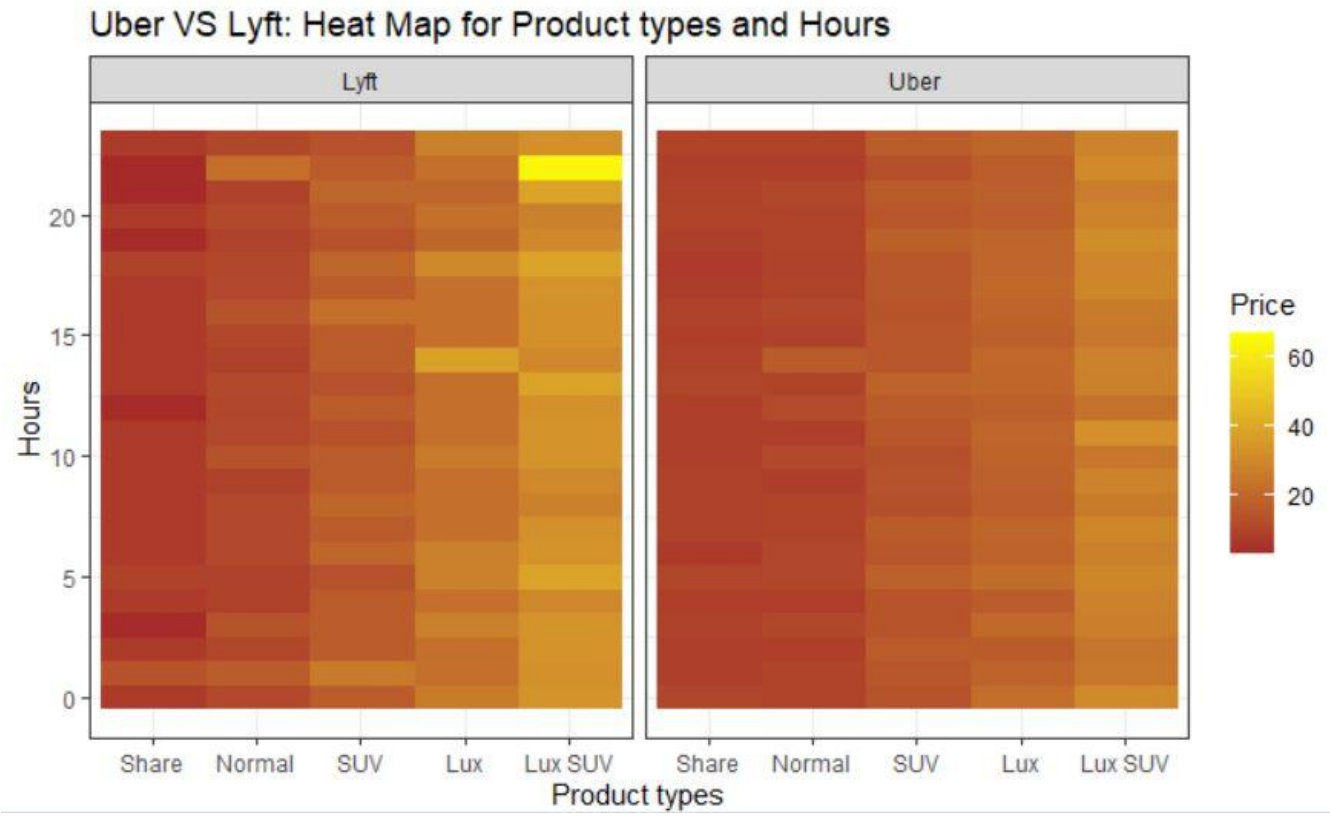
As we can see that there are pretty distinct differences in the pricing of both Uber and Lyft, and both the plots have some outliers.

- Histogram



The chart illustrates the frequency distribution of price ranges for Uber and Lyft. The blue color represents the data for Uber, while the green color represents the data for Lyft. The minimum price for Lyft is 2.5, whereas for Uber it is 4.5. The maximum price for Lyft is 97.5 and for Uber, it is 89.5.

- Heatmap for specific location and hours

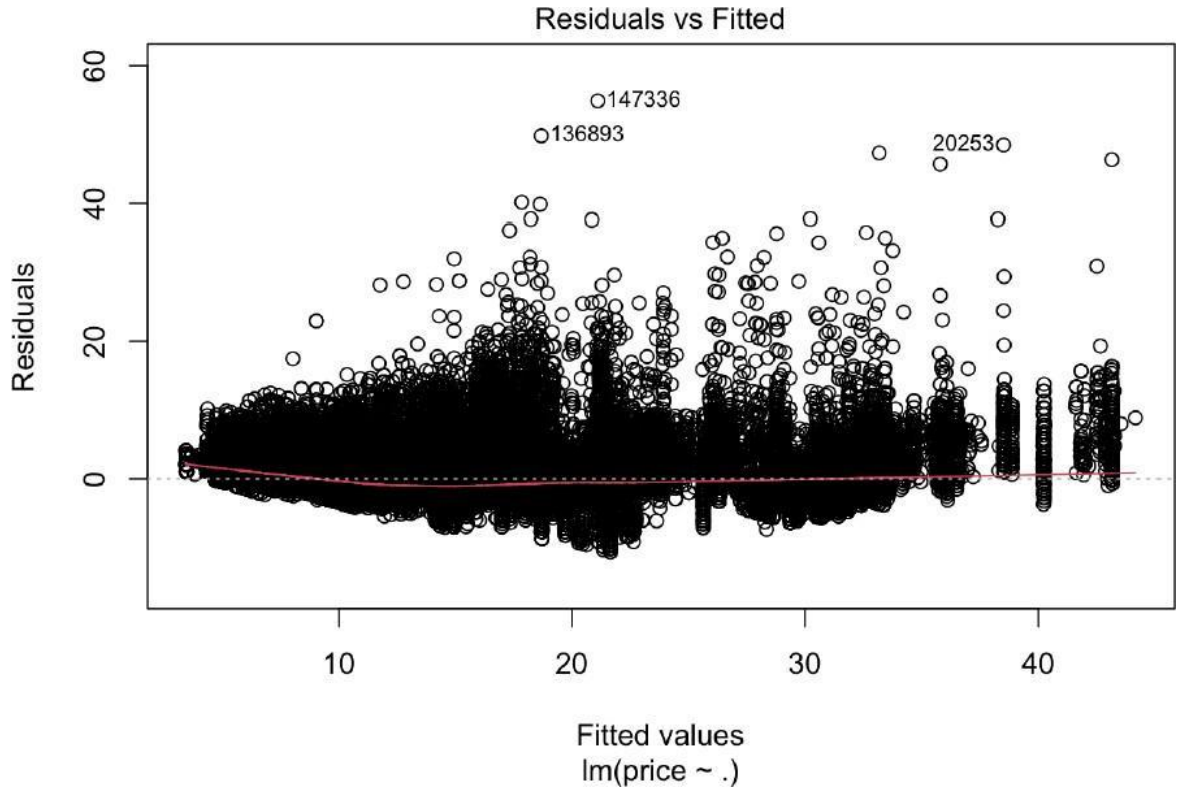


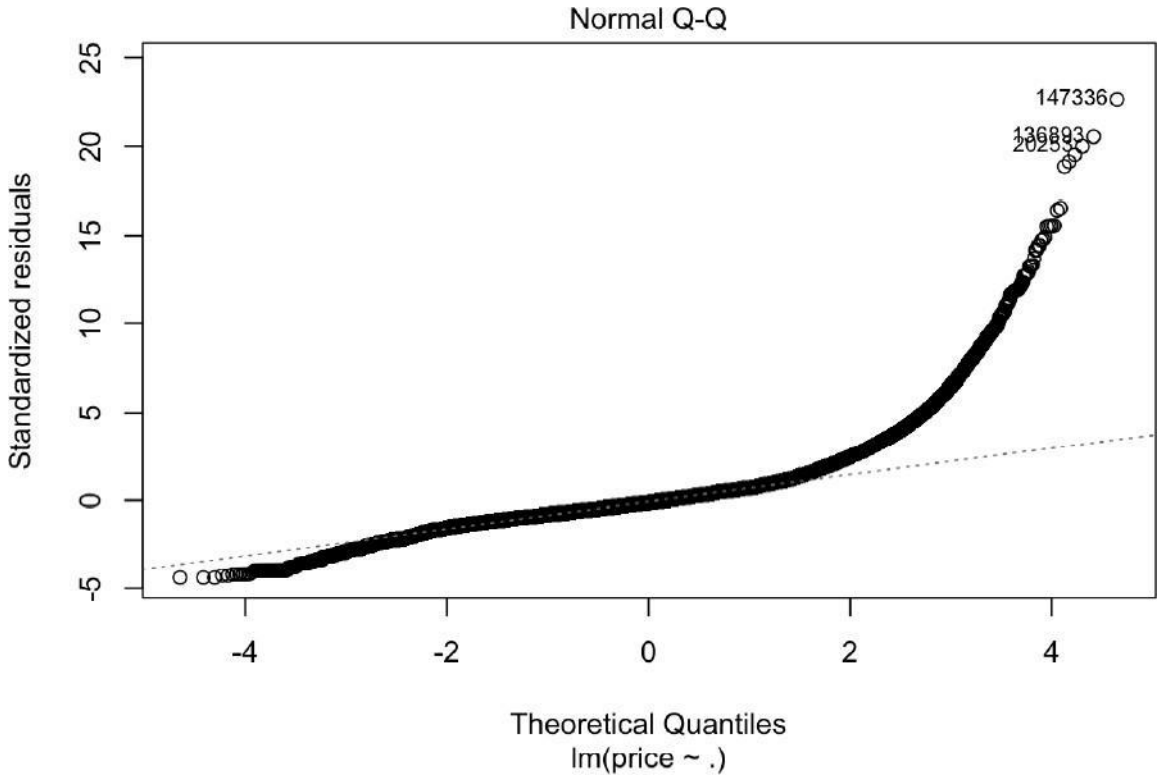
For Lyft, the price of “Share” products does not change with Hours. However, for “Normal” products, the price at 1pm is slightly higher than other hours. For “SUV” , 5am is higher and for “Lux”, 1–2pm is higher. For “Lux SUV”, 10am is higher. There is no consistent pattern in Lyft so that certain hours have higher prices for all types of products. Therefore, the relationship between hours and price may not have an universal rule but depends on specific circumstances for Lyft. As for Uber, the price of “Share” products is higher at night from 10pm to 12am. For “Normal” products, the price at 10am, 1pm, and 3pm are higher than other hours, and it is interesting to see that the price at 1pm is also higher for “SUV”, “Lux”, and “Lux SUV”.

➢DATA MODELLING

○ Linear Regression Model

Linear regression analysis is a statistical technique used to predict the value of one variable based on another variable. The variable that is being predicted is called the dependent variable, while the variable that is being used to make the prediction is called the independent variable. The scatter plot below shows the residual versus fitted plot, which is generated by removing some predictors based on their p-values. In this case, we predicted prices using predictors such as temperature, wind, snow, day of the week, year, and rain. The analysis showed that the R-squared value for these predictors is 0.9365.

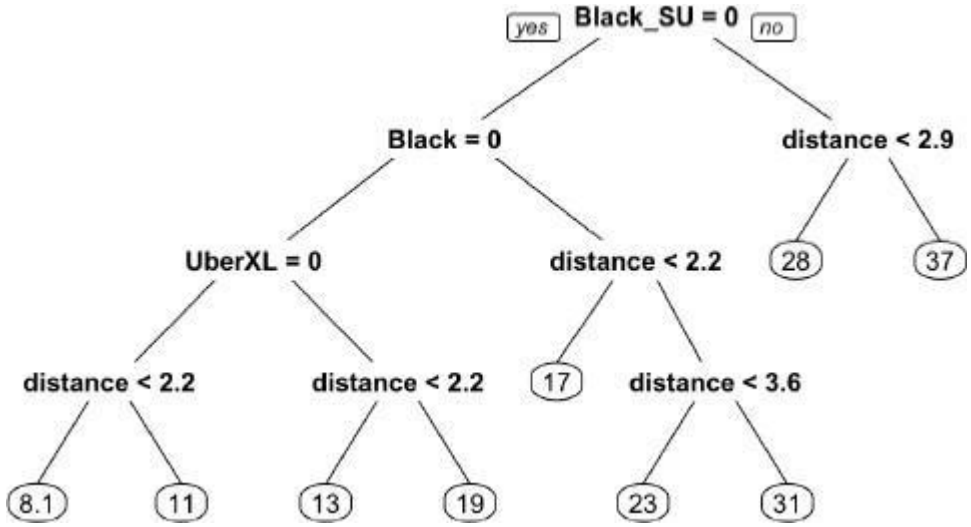




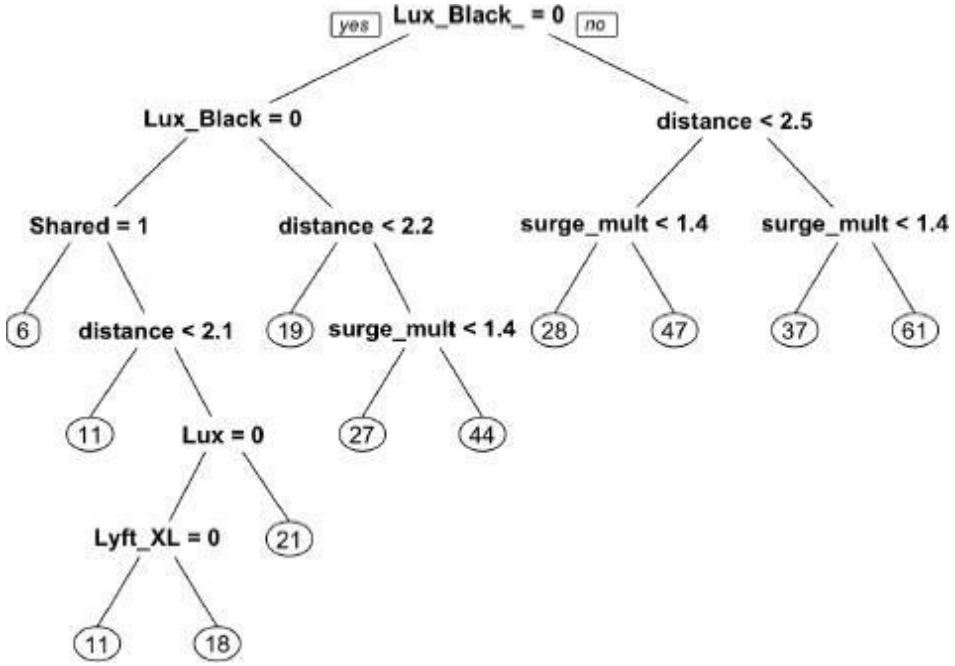
- **Decision Tree**

A decision tree is a type of flowchart that shows a direct path to a decision. It is an algorithm used for categorizing data that employs conditional statements. Starting from a single node, the decision tree branches out into multiple paths. Each branch represents different possible outcomes, considering a range of choices and variables until a conclusive decision is made. Decision trees are highly useful in data analytics and machine learning as they break down complex data into more manageable parts.

Decision Tree model for Uber -



Decision Tree model for Lyft -



- **Random Forest**

In order to compare our linear models, we used a random forest model as a baseline. This model creates splits to reduce the overall error rate, allowing it to fit the data more closely. By using this model as a baseline, we can compare the performance of other models against it..

The Random Forest model was applied to the dataset and various parameters were evaluated.

We got the below results -

```
#Evaluation
mat_rf_uber<- regr.eval(uberTesting[,19], uberPrediction_rmforest)#, stats = c('mape','rmse'))
print(mat_rf_uber)
```

```
       mae       mse      rmse      mape
1.7313295 6.5205389 2.5535346 0.1290741
[1] "The Accuracy of Random Forest for Uber :87.092593"
```

```
#Evaluation
mat_rf_lyft<- regr.eval(lyftTesting[,19], lyft_pred_rmforest)#, stats = c('mape','rmse'))
print(mat_rf_lyft)
```

```
     mae      mse     rmse     mape
1.666995 5.217596 2.284206 0.128877
[1] "The Accuracy of Random Forest for Lyft :87.112304"
```

➢MODEL EVALUATION

To evaluate the performance of different strategies in predicting trip outcomes, we employ various assessment criteria. The following statistics are used to measure their prediction accuracy–

Analysis of all implemented models:

```
[1] "------------------ Uber Statitics ------------------"
         Linear Regression Decision Tree Random Forest
MAE              1.6697108     1.7995829     1.7313295
MSE              5.8347082     6.8132767     6.5205389
RMSE             2.4155141     2.6102254     2.5535346
MAPE             0.1191017     0.1230596     0.1290741
Accuracy        88.0898340    87.6940363    87.0925928
[1] "------------------ lyft Statitics ------------------"
         Linear Regression Decision Tree Random Forest
MAE              1.8068009     2.6463307     1.6669953
MSE              6.2327136    12.5978293     5.2175959
RMSE             2.4965403     3.5493421     2.2842057
MAPE             0.1493054     0.1969981     0.1288770
Accuracy        85.0694620    80.3001895    87.1123040
```

As we can observe from the above table that the best score for UBER data is achieved by the Linear regression model by 88.089834 and that for the LYFT is achieved by the Random Forest model by 86.733867. The overall accurate model is Random Forest.

Based on different metrics, we can conclude that the Random Forest model is the best model and thus we could consider this as our final model.

➢ OBSERVATIONS & CONCLUSION

From the exploratory data analysis, we noticed that:

- We performed EDA to gain a better understanding of the data.
- For both Uber and Lyft, the overall model and all the predictors used were statistically significant.
- Our model explained 91.31% of the variability in Log(Fare) for Lyft and 80.67% for Uber.
- Uber is more affordable, but Lyft provides fair competition.
- Long-distance travel is more expensive, but the relationship is not linear. However, a spike in demand at a certain time can affect the price.
- Lyft's base pricing for shared rides is cheaper than their premium "Black" and "Black XL" vehicles.
- The Random Forest model's modified results proved to be the most effective.

➢ PLANNED FUTURE WORK

There are multiple ways to tackle this problem and different perspectives to consider. One approach is to enhance the models to improve the accuracy of the forecasts. For example, we can explore the correlations between variables such as the predictors for distance and types of cabs. We can also analyze the Random Forests prediction errors by optimizing the parameter values. Additionally, we plan to include external data such as traffic conditions and time to our analysis. We aim to investigate the reasons behind the high variability of fares for a given source and destination. With the insights gained from this project, we aim to develop an enhanced model with more useful features and experiment with new data formats such as Avro or Parquet to improve the efficiency of row/column operations.

> CODE AND PRESENTATION LINK

GitHub code link

Presentation Link

➢ REFERENCES

- Shashank H, "Data Analysis of Uber and Lyft Cab Services," International Journal of Interdisciplinary Innovative Research & Development, 2022,ISSN: 2456-236XVol.05 Issue 01 | 2020 http://ijiird.com/wp-content/uploads/050144.pdf

- Mrinalini Sunder, (2021, July 27). “Uber and Lyft Cab Prices Data Analysis and Visualization”Retrievedfrom https://medium.com/mindtrades-consulting/uber-and-lyft-cab-prices-data-analysis-and-visualization-93aca4596f20.

- *ACM Digital Library*. (n.d.). ACM Digital Library. Retrieved December 4, 2022, from https://dl.acm.org/doi/fullHtml/10.1145/3178876.3186134

- Leo, M. S. (2021, January 24). *New York Taxi data set analysis*. Medium. RetrievedDecember4,2022,from https://towardsdatascience.com/new-york-taxi-data-set-analysis-7f3a9ad84850

- Liu, L., Qiu, Z., Li, G., Wang, Q., Ouyang, W., & Lin, L. (2019). Contextualized Spatial-Temporal Network for Taxi Origin-Destination Demand Prediction.