Name: Chandrashekhar.

Reg No: 11802214.

Course Code: INT248.

Course Name: Advanced Machine Learning.

Topic: Uber Data Analysis.

Submitted to: Md. Imran Hussain.

CONTENTS:

INTRODUCTION

1.1 Overview:

Uber Technologies, Inc., commonly known as Uber, was a ride-sharing company and offers vehicles for hire, food delivery (Uber Eats), package delivery, couriers, freight transportation, and, through a partnership with Lime, electric bicycle and motorized scooter rental. Travis Kalanick and Garrett Camp, a successful technology entrepreneur, founded it in 2009. After selling his first start up to eBay, Camp decided to create a new start up to address San Francisco's serious taxi problem.

Together, the pair developed the Uber app to help connect riders and local drivers. The service was initially launched in San Francisco and eventually expanded to Chicago in April 2012, proving to be a highly convenient great alternative to taxis and poorly funded public transportation systems. Over time, Uber has since expanded into smaller communities and has become popular throughout the world. In December 2013, USA Today named Uber its tech company of the year.

In Supervised learning, we have a training set and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. We applied machine-learning algorithms to make a prediction of Price in the Uber Dataset of Boston. Several features will be selected from 55 columns. Predictive analysis is a procedure that incorporates the use of computational methods to determine important and useful patterns in large data.

1.2 Objective

The objective is to first explore hidden or previously unknown information by applying exploratory data analytics on the dataset and to know the effect of each field on price with every other field of the dataset. Then we apply different machine learning models to complete the analysis. After this, the results of applied machine learning models were compared and analysed based on accuracy, and then the best performing model was suggested for further predictions of the label 'Price'.

1.3 Issues and Problem Faced:

1. **Overfitting in Regression Problem:-** Overfitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables. This problem occurs when the model is too complex. In regression analysis, overfitting can produce misleading R-squared values. When this occurs, the regression coefficients represent the noise rather than genuine relationships. However, there is another problem. Each sample has its unique quirks. Consequently, a regression model that becomes tailor-made to fit the random quirks of one sample is unlikely to fit the random quirks of another sample. Thus, overfitting a regression model reduces its generalizability outside the original dataset.

2. **Strip-plot and Scatter diagram:-** One problem with strip plots is how to display multiple points with the same value. If it uses the jitter option, a small amount of random noise is added to the vertical coordinate and if it goes with the stack option it increments the repeated values to the vertical coordinate, which gives the strip plot a histogram-like appearance.

Scatter plot does not show the relationship for more than two variables. In addition, it is unable to give the exact extent of correlation.

3. **Label Encoding:-** It assigns a unique number(starting from 0) to each class of data which may lead to the generation of priority issues in the training of data sets. A label with high value may be considered to have high priority than a label having lower value but actually, there is no such priority relation between the attributes of the same classes.

4. **Computational Time:-** Algorithms like support vector machine(SVM) don't scale well for larger datasets especially when the number of features are more than the number of samples. In addition, it sometimes runs endlessly and never completes execution.

1.5 Organization of the Project Report

The first section of this paper presents the concept of exploratory data analysis, which told general information about the dataset. Then from the next section feature engineering part was started in which we plot many charts and deal with columns to extract the features helpful for our predictions in many ways. In the last part, we did modeling and testing in which we apply different models to check the accuracy and for further price prediction.

DATASET USED

For this project, I have taken the dataset from kaggle in which there are 7 columns and approx. 1156 rows and the file is in csv format. you can view or download the dataset through this link: https://www.kaggle.com/zusmani/uberdrives

PROPOSED ARCHITECTURE

4.1 Data Preparation

The data i used for our project is provided on the www.kaggle.com website. The original dataset contains 1156 rows and 7 columns, which contain the data of under drives. The dataset has many fields that describe us about the time, geographic location, and climatic conditions when the different Uber cabs opted.

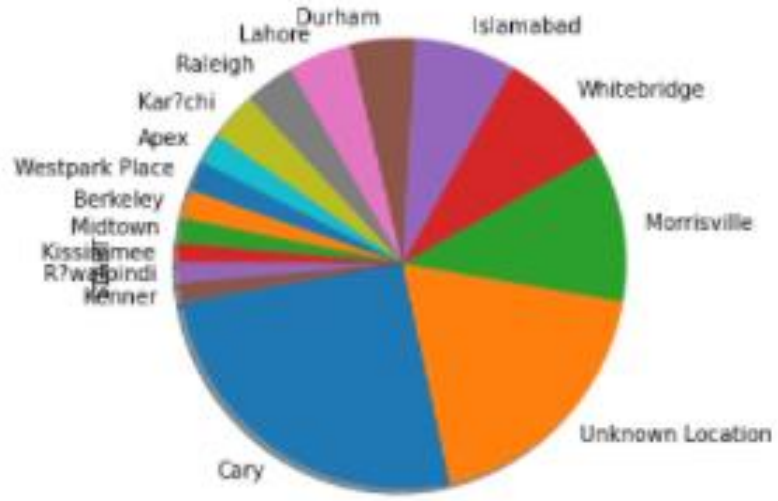
Data has 3 types of data-types which were as follows:- integer, float, and object. The dataset is not complete which means we have also null values in a column named price of around 503.

	START_DATE*	END_DATE*	CATEGORY*	START*	STOP*	MILES*	PURPOSE*
0	1/1/2016 21:11	1/1/2016 21:17	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain
1	1/2/2016 1:25	1/2/2016 1:37	Business	Fort Pierce	Fort Pierce	5.0	NaN
2	1/2/2016 20:25	1/2/2016 20:38	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies
3	1/5/2016 17:31	1/5/2016 17:45	Business	Fort Pierce	Fort Pierce	4.7	Meeting
4	1/6/2016 14:42	1/6/2016 15:49	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit

4.2 Data Visualization

Data visualization is a graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

For the same purpose, we have to import matplotlib and seaborn library and plot different types of charts like strip plot, scatter plot, and bar chart.



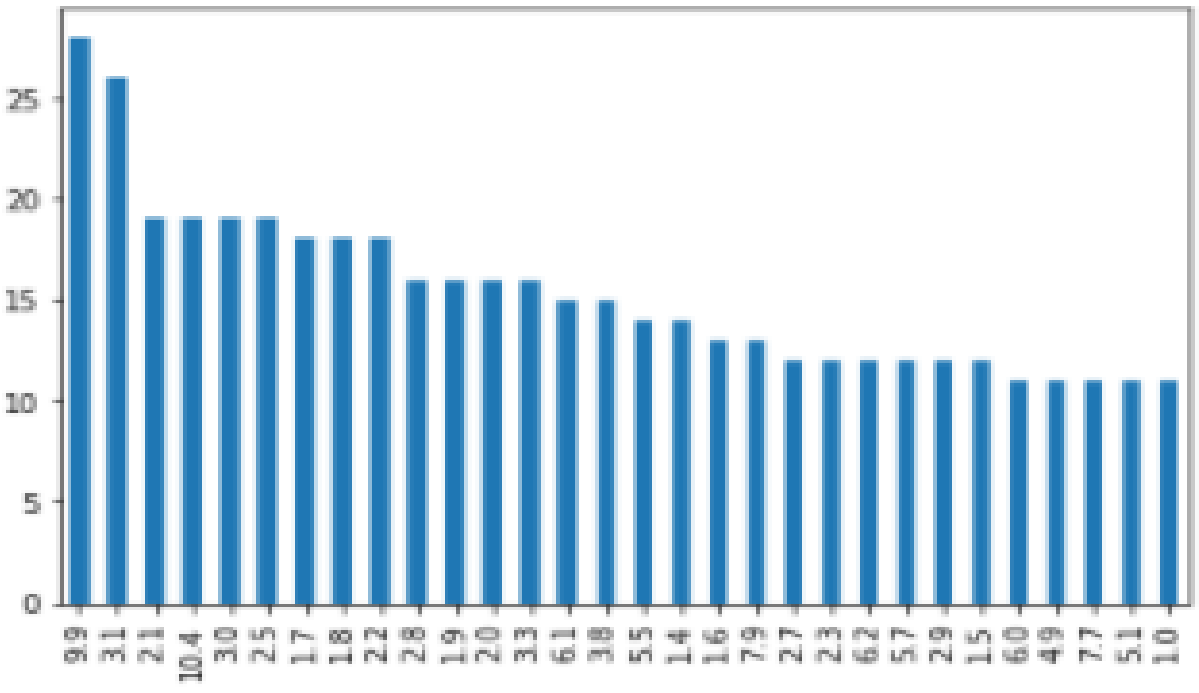


Fig : cabe rides with respect to miles

4.3 Feature Engineering

Feature engineering is the most important part of the data analytics process. It deals with, selecting the features that are used in training and making predictions. All machine-learning algorithms use some input data to create outputs. This input data comprise features, which are usually in the form of structured columns. Algorithms require features with some specific characteristics to work properly. A bad feature selection may lead to a less accurate or poor predictive model. To filters out all the unused or redundant features, the need for feature engineering arises. It has mainly two goals:

- Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
- Improving the performance of machine learning models.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1155 entries, 0 to 1154
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   START_DATE  1155 non-null   datetime64[ns]
 1   END_DATE    1155 non-null   datetime64[ns]
 2   CATEGORY    1155 non-null   object
 3   START       1155 non-null   object
 4   STOP        1155 non-null   object
 5   MILES       1155 non-null   float64
 6   PURPOSE     1155 non-null   object
dtypes: datetime64[ns](2), float64(1), object(4)
memory usage: 72.2+ KB
```

Coversion of data and time object data type to datetime

Improved dataset by reducing the NULL Values and coverting data and time datatype object to datetime and made the data compatible with the machine learning algorithm requirements.

Results and Experimental Analysis:

I have made exploratory data analysis by using the dataset of uber drives which contains 1156 rows and 7 columns and visualized the null values for each attribute and found 43% missing values in the purpose attribute which is around 502 Null Values. After filling the null values we converted date and time of start_date and end_date attribute of data type object to datetime and the different analysis by considering certain parameter such as highest stop points, most miles travelled,places where cab drives are popular, purpose of cab ride vary with distance and time etc.

We have found the highest start point for the cab drives is in the lo cations of Islamabad, whitebridge,morris ville,Cary and lowest start points are Washington,East Austin, Lower Garden District.

```
Edgehill Farms       10
New Orleans          10
Kenner               10
Emeryville            9
Central               9
                     ..
Daytona Beach         1
Sand Lake Commons     1
Sky Lake              1
Vista East            1
Ilukwatta             1
Name: STOP, Length: 173, dtype: int64
```

Lowest Start Points

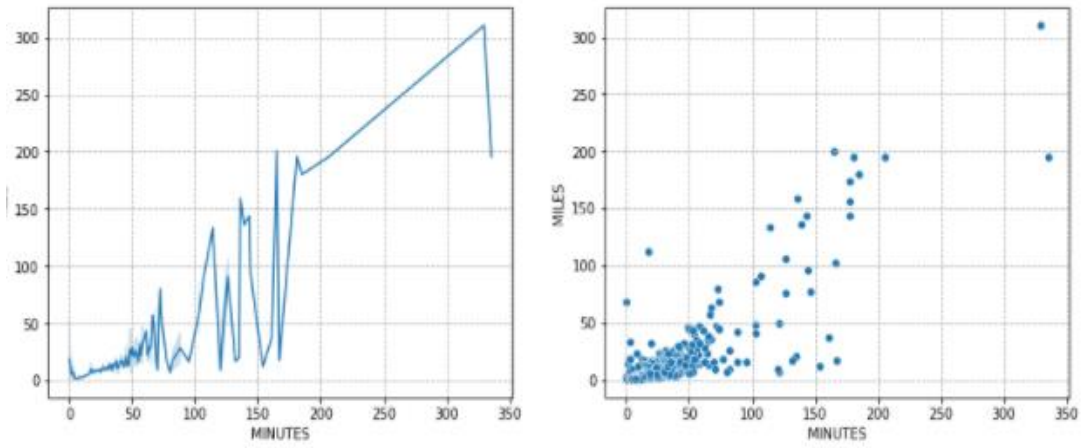
```
Cary                 203
Unknown Location     149
Morrisville           84
Whitebridge           65
Islamabad             58
Durham                36
Lahore                36
Raleigh               29
Kar?chi               26
Apex                  17
Berkeley              16
Westpark Place        16
R?walpindi            13
Kissimmee             12
Midtown               11
Edgehill Farms        10
New Orleans           10
Kenner                10
Name: STOP, dtype: int64
```

Highest Start Points

We also the also found the reasons for picking the cab rides and tried to find the certain pattern by grouping it with miles attribute and found most people pick fort fierce as start point for different purposes such as meal and entertainment errand/supplies,meeting or customer visit.

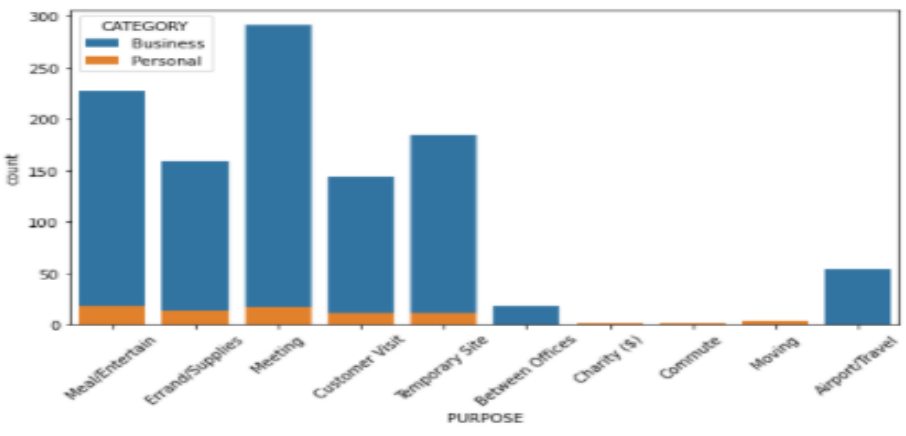
	START_DATE*	END_DATE*	CATEGORY*	START*	STOP*	MILES*	PURPOSE*
0	1/1/2016 21:11	1/1/2016 21:17	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain
1	1/2/2016 1:25	1/2/2016 1:37	Business	Fort Pierce	Fort Pierce	5.0	NaN
2	1/2/2016 20:25	1/2/2016 20:38	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies
3	1/5/2016 17:31	1/5/2016 17:45	Business	Fort Pierce	Fort Pierce	4.7	Meeting
4	1/6/2016 14:42	1/6/2016 15:49	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit

It also results to find the round trips in which start and end locations are the same and to evaluate the frequency of trips in every month and when are the cab rides became more popular.the data analysis also to calculate the duration of cab rides took to reach the destination with respect to distance.



we see that our conventional logic, that distance is proportional to time, is challenged as some cab rides took more time for less distance.

The analysis helps to Distribute the cab rides based on category so that the company can provide the comfortable rides to the customers and insights to cab aggregators to decide which sector to introduce new cabs in.



OUTPUT SCREENSHOTS:

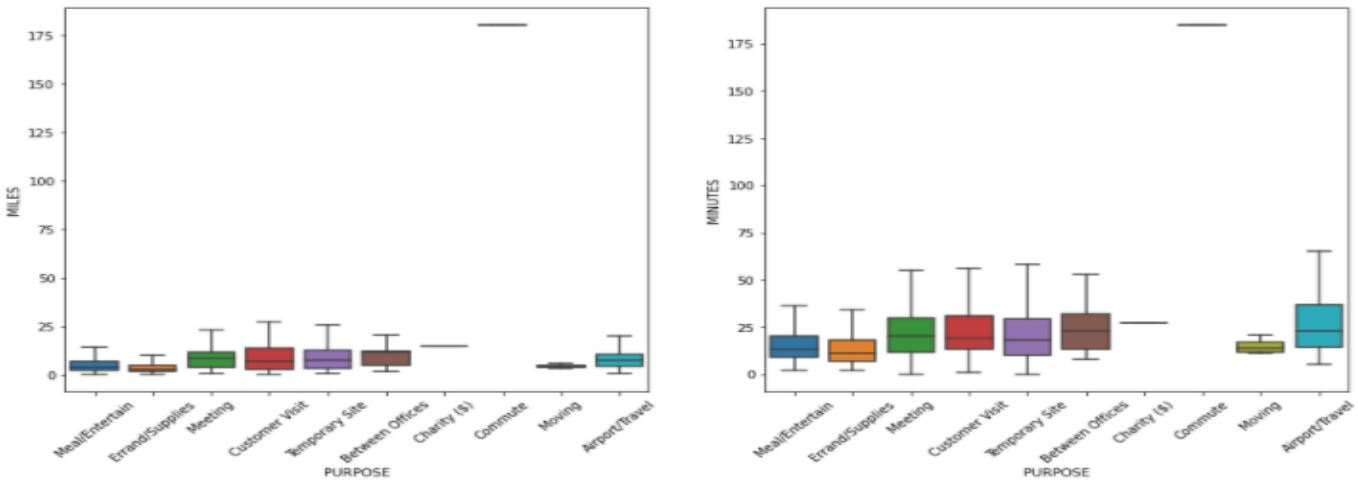


Fig 1.5 PURPOSE of Cab ride vary with time and distance

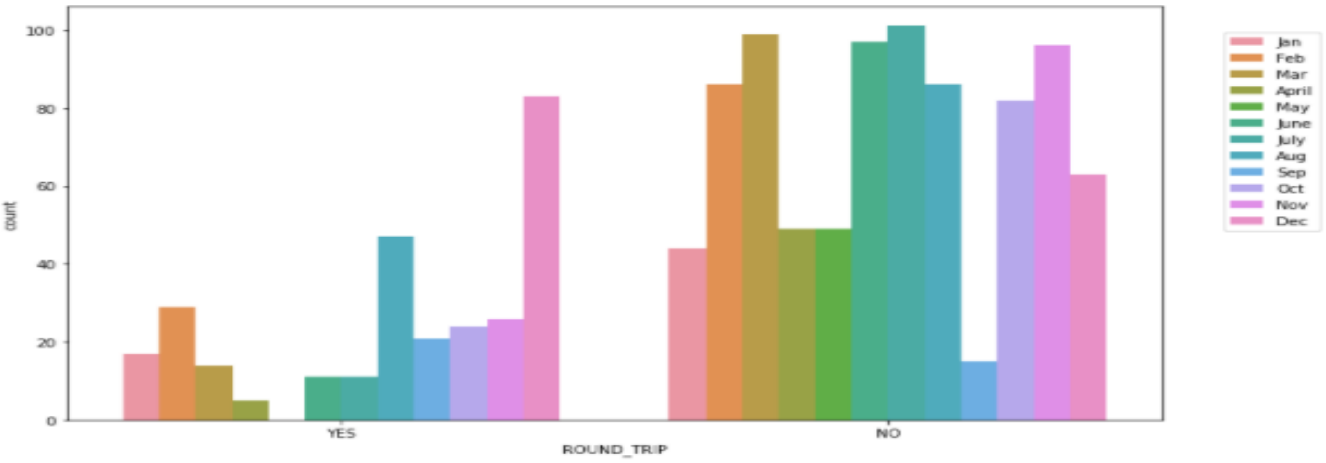
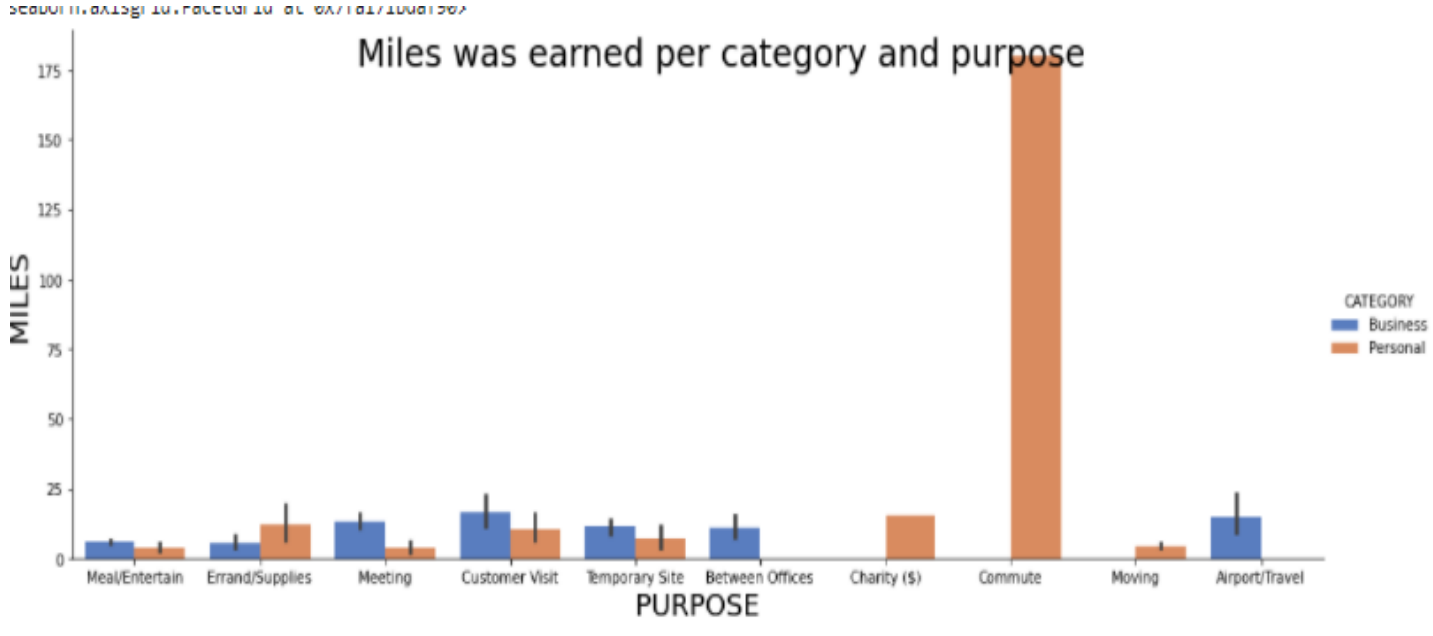


Fig 1.6 Round Trips against Months



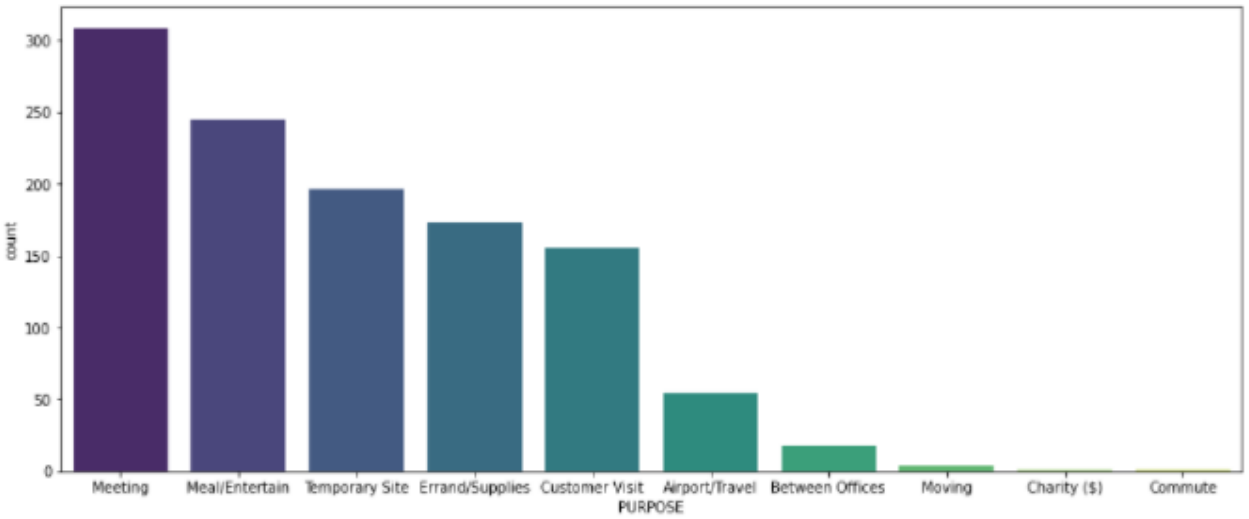


Fig : Purpose for Travel

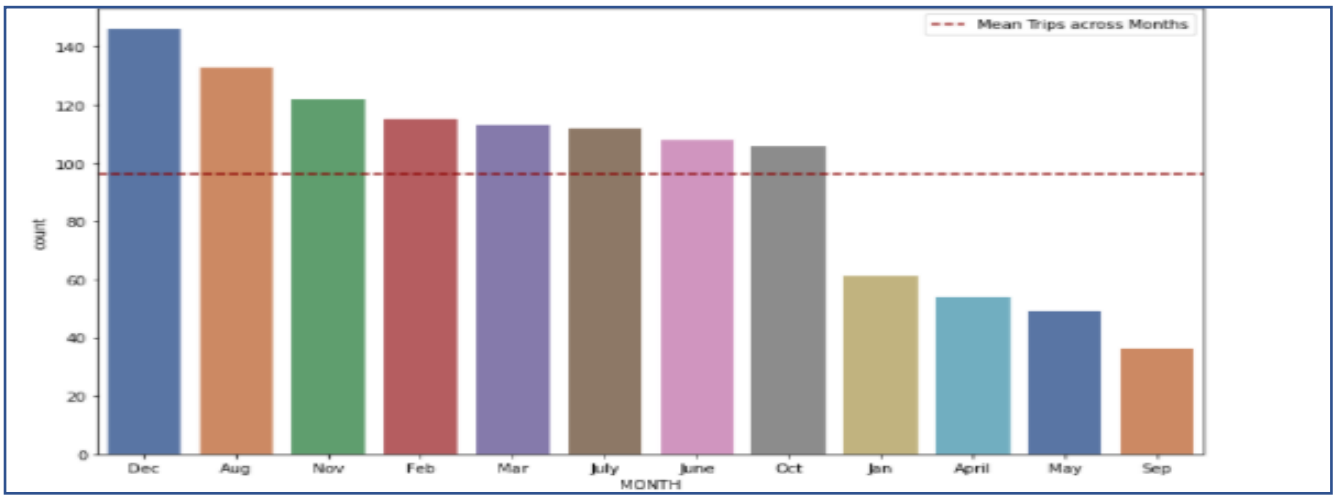
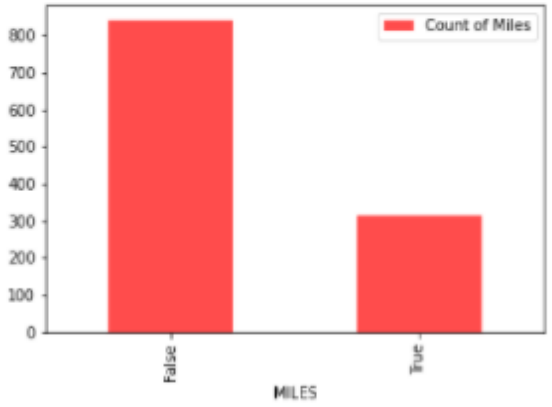
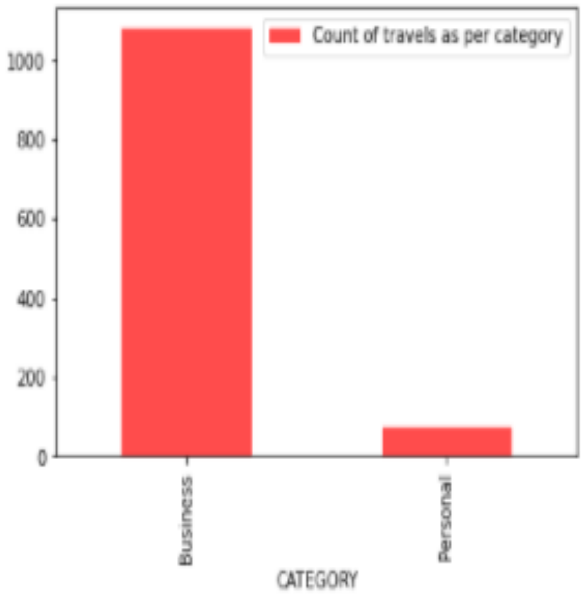


Fig : Cab Rides popular Months

col_0	Count of Miles
MILES	
False	840
True	315





CONCLUSION AND FUTURE SCOPE

Before working on features first we need to know about the data insights which we get to know by EDA. Apart from that, we visualize the data by drawing various plots,.After this, we convert all categorical values into continuous data type and found out 502 NULL and filled values by the median of other values and also done analysis by exploring the data and plotting different graphs such purpose of cabs drives,highest stop point, cab drives rounds trips against months etc and I found different conclusions for the questionarie.

- Most of the cab rides are within a distance of 31 miles taking about 34 minutes.
- Business Cab rides are not only more in volume, but also in distance travelled.
- Main uses of cab rides are Meal/Entertainment, Customer visit, Meeting, Errand/Supplies.
- Cab traffic is mostly concentrated in 5 cities or localities.
- Ac seasonal pattern of cab ride volume exists, which is highest on December.

We can use this data for training a model using ML and building a smart AI based predictive system. Model can automatically send the insights to the authorities or drivers related to areas having most trips and passenger count in certain areas and improves GPS data and its own algorithms which can make alterations based on the time of the journey and provide statiscal analysis of estimating fares, showing up surge prices and heat maps to the drivers on where to position themselves within the city.

REFERENCES

- https://matplotlib.org/1.3.1/users/legend_guide.html
- https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston
- https://archive.ics.uci.edu/ml/machine-learning-databases/00339/
- https://growvation.com/paritoshsankhla/project/uber-data-analysis/5e95ee80-9455- 4473-acaf-b670fe2abc8b
- https://github.com/geoninja/Uber-Data-Analysis