

DATA SCIENCE PROJECT REPORT

**EXPLORATORY DATA ANALYSIS ON UBER**

**PICKUPS**

**GROUP NUMBER :**

**GROUP MEMBERS:**

* HARSHA SATHISH
* ARVIND KUMAR
* ABISHEK S
* NAVNEET KUMAR

**TABLE OF CONTENTS**

* ABSTRACT
* INTRODUCTION
* BROAD CONTEXT
* STUDY SYSTEM
* METHODS
* RESULTS
* DISCUSSION
* LITERATURE CITED
* CONCLUSION

**ABSTRACT**

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 And Time Comparison With Every Other Field Of The Dataset. Then We Apply Different Data Science Models To Complete The Analysis. After This, The Results Of Applied Data Science Models Were Compared And Analysed On The Basis Of Accuracy.

**AIM AND SCOPE**

Complete Analysis And Exploration Of Uber Data.

**EXPECTED OUTCOME**

After The Complete Analysis Of The Data We Get Those Factors Or Features Which Enhance The Business Of Uber Company**.**

**INTRODUCTION**

Uber Technologies commonly known as Uber, is an American technology company. The company is based in San Francisco and has operations in over 900 metropolitan areas worldwide. Uber is estimated to have over 93 million monthly active users worldwide. So it has a large number of customers and its been one of the most used taxi worldwide. In our project we are going to comparing business vs personal trips, the frequency for the purpose of the trip, checking how many are the round trips, evaluating the frequency of the trip in each month, and so on.

Due to dynamic pricing models, prices for the same route may vary based on the supply and demand for rides at the time the ride is requested. When rides are in high demand in a certain area and there are not enough drivers in such area, fares increase to get more drivers to that area. So by the end of our project we will be able to find the best way to all the demands.

**BROAD CONTEXT**

Here in this project Looking at Data find that the data is increasing day by day and approximate 2.5 quintillion bytes of data generate every day. Now, from this data analysis and get useful information which is most important and to understand that here we perform data analysis on UBER data.

In this project, we compare and analyse different data of uber analysis.

We study different columns of the table and try to co-relate them with others and find a relation between those two.

We try to find and analyse those key factors like date, month etc which helps Uber Company to enhance their business by focusing on those services and make required changes.

**STUDY SYSTEM**

**WHAT IS DATA ANALYSIS ?**

The process of cleaning, transforming, manipulating data into useful information that is Data analysis. When we take a particular decision based on previous data that is data analysis. We can make future decisions using data analysis.

**EXPLORATORY DATA ANALYSIS ?**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns to spot anomalies to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

It is a good practice to understand the data first and try to gather as many insights from it.

Exploratory Data Analysisis all about making sense of data in hand.

**WHY WE USE DATA ANALYSIS ?**

All the business has lots of data. To grow business, sometimes data analysis required. By analysing data we get important topics on which work out and make our plan for the future through which made perfect future decisions. Most of the businesses going online where the data generate increases day by day. To grow business with this competitive environment data analysis is necessary.

This project contains many libraries like:

-> Pandas

-> NumPy

-> Matplotlib

-> Seaborn

**Pandas:**

The Pandas library is used to import different types of datasets.

**NumPy:**

The NumPy library is used for manipulating and performing on the arrays.

**Matplotlib:**

The Matplotlib library is used for plotting the graphs.

**Seaborn:**

The seaborn library is used for plotting the heatmap.

**METHODS AND RESULTS**

~ Defining the problem statement

~ Collecting the data

~ Exploratory data analysis

**DEFINING THE PROBLEM STATEMENT :**

In this project, we study the data of Uber which is present in tabular format in which we use different libraries like numpy , pandas and matplotlib and different algorithms.

**COLLECTING THE DATA :**

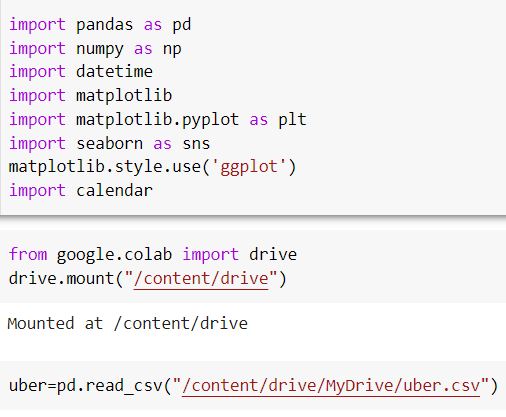
Using The Dataset Provided Lets Collect The Information Inside It.

**EXPLORATORY DATA ANALYSIS :**

Here, we are going to perform data analysis task in four steps.

**STEP 1:**

In step 1 we are going to import the libraries and read the data



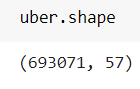
After importing the libraries read the data :



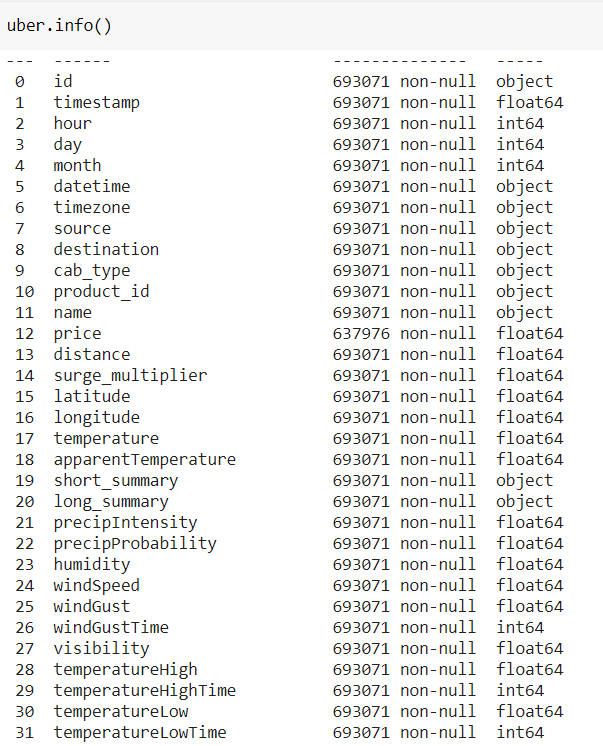
**STEP 2:**

In step 2 we need to clean the data

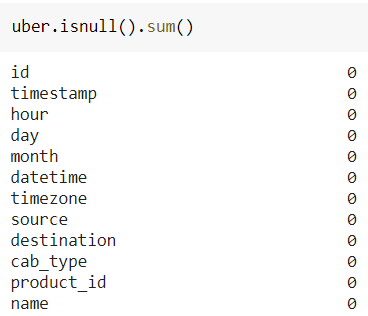
First find shape of the dataset :



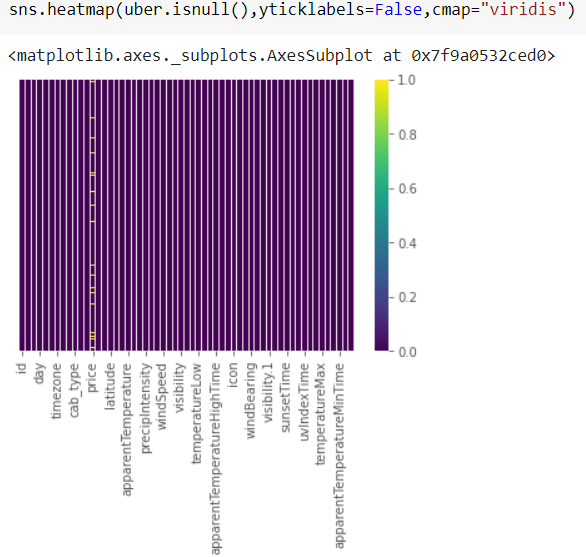
After that find the information about the data :



Next step is to check for null values :



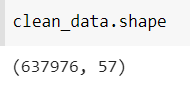
As graph :



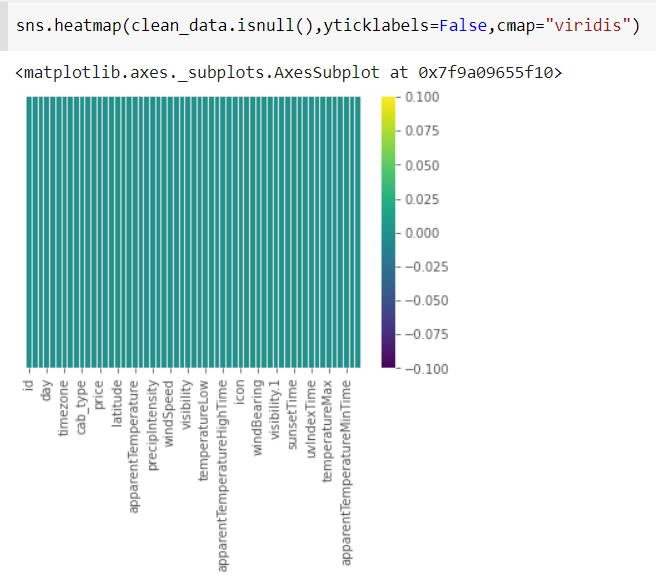
Next we have to clean the null data:



After that check the new shape of data set :



Now the data set is clean without null values :



**STEP 3:**

In the next step we have to transform the data.

For the next process we need to find the average , maximum, minimum values of the data set.

Then we can use .describe method for finding almost everything we needed like mean, max, min, average, 25%, 50%, 75%.



By using this technique we can find the required :

=> Average hour, day, month usage of uber taxi.

=> Average price and distance .

=> Max and min cost of a journey.

=> Max and min distance travelled.

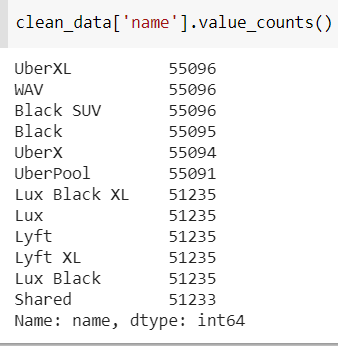
=> Max and min hours of uber running.

**STEP 4:**

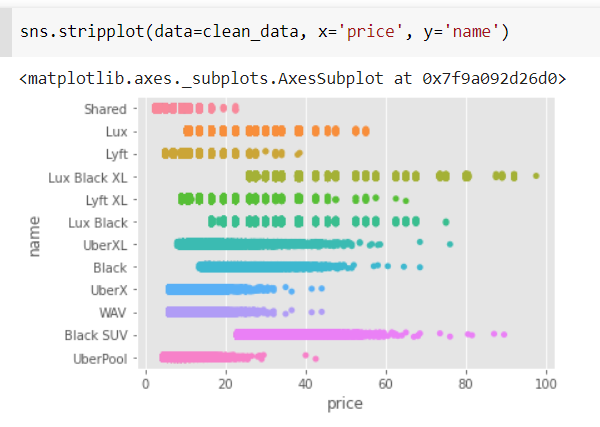
Here we are going to visualize the data.

Different categories of data. From data, we can find the types of people and their usage of UBER.

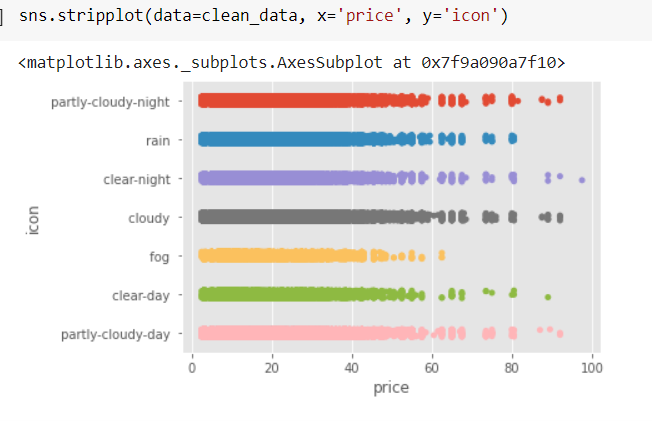
First we are finding the values each variety of car is having :



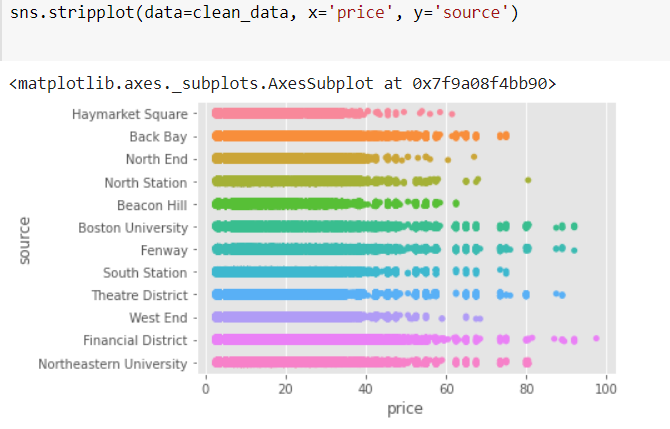
We are now plotting the relation between price and car variety(taxi name):



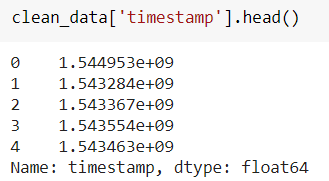
Now we are plotting the relation between the climate condition and price :

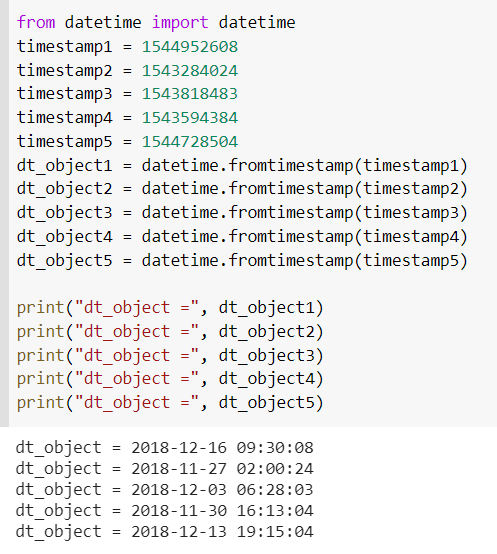


Then lets plot the relation between price and the location(America):



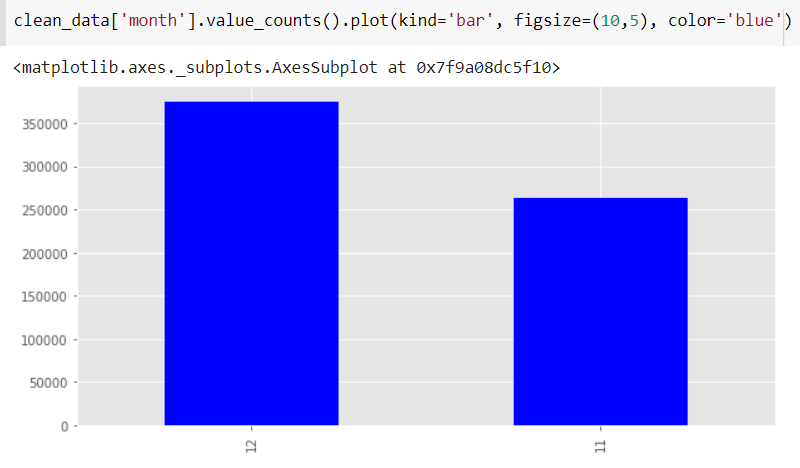
Here the next important step is to convert the timestamp into datetime value:





So by this timestamp to datetime conversion we get to know that, our data is of the year 2018 and in the month of November and December only.

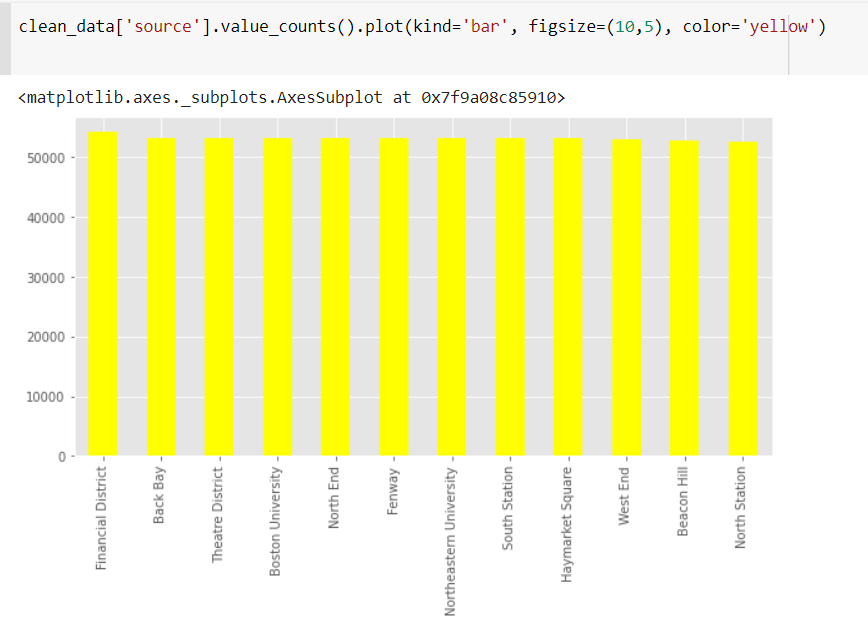
Next let us visualise the difference between November and December months :



This tells us that we are Having more data in the month of December when compared to November.

So like this we need to visualise the data we needed and the relation between them.

After knowing the difference in months now let us find in which location we are having more data values :



**CONCLUSION**

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, due to which we understand the relationship between values in data set.

Co-lab link : https://colab.research.google.com/drive/1OrY2at\_xwpBWMAcJwwLzd9diOarYOFQa?usp=sharing