Uber Dataset Analysis

Al-Hussain Hakim Shaikh

Department of Computer Engineering, Vishwakarma Institute of Technology,
Savitribai Phule Pune University, Pune- 411037

I. INTRODUCTION

Uber Technologies, Inc., commonly known as **Uber**, is an American technology company. Its services include ride-hailing, food delivery, package delivery, couriers, freight transportation, and, through a partnership with Lime, electric bicycle and motorized scooter rental.

We will mainly use data regarding **Uber ride**. We will use Python to analyse data from Uber. You will gain a hands-on experience with Python in analysing data.

This project will explain the working of an Uber dataset, which contains data produced by Uber for New York City. Uber is defined as a P2P platform. The platform links you to drivers who can take you to your destination. The dataset includes primary data on Uber pickups with details including the date, time of the ride as well as longitude-latitude information, Using the information, the paper explains the use of the k-means clustering algorithm on the set of data and classify the various parts of New York City. Since the industry is booming and expected to grow shortly. Effective taxi dispatching will facilitate each driver and passenger to reduce the wait time to seek out one another. The model is employed to predict the demand on points of the city.

II. OBJECTIVE

Depending on how much data you have and features, the analysis can go on and on. The following questions are useful to do our analysis:

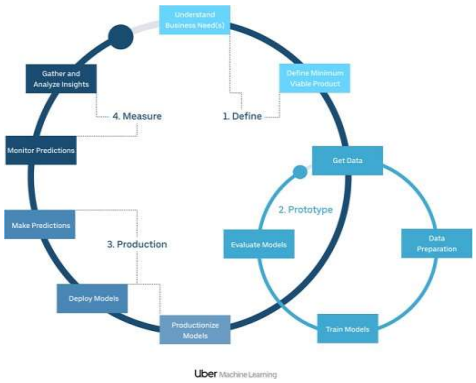
We will use Python to:

- What Hour Do Most People Take Uber to Their Destination?
- Check The Purpose of Trips.
- Which Day Has the Highest Number of Trips
- What Are the Number of Trips Per Each Day?
- What Are the Trips In The Month
- The starting points of trips. Where Do People Start Boarding Their Trip from Most?
- How many times have I travelled in the past?
- How many trips were completed and cancelled?
- Where did most of the layoffs take place?
- What type of product is most often selected?
- What a measure. fare, distance, amount, and time spent on the ride?
- What a measure. fare, distance, amount, and time spent on the ride?
- Which days of the week have the highest fare?
- Which is the longest / shortest and most expensive / cheapest ride?
- What is the average lead time before requesting a trip?

III.BACKGROUND

The day-to-day effect of rising prices varies depending on the location and pair of the Origin-Destination (OD pair) of the Uber trip: at accommodations/train stations, daylight hours can affect the rising price; for theatres, the hour of the important or famous play will affect the prices; finally, attractively, the price hike may be affected by certain holidays, which will increase the number of guests and perhaps even the prices; Finally, at airports, the price of escalation will be affected by the number of periodic flights and certain weather conditions, which could prevent more flights to land and land.

The weather is likely to have a significant impact on the rise in prices of Uber fares and airports as a starting point, as departure and accommodation of aircraft depending on the weather at that time. Different weather conditions will certainly affect the price increase in different ways and at different levels: we assume that weather conditions such as clouds or clearness do not have the same effect on inflation prices as weather conditions such as snow or fog. As for the day of the week, one thing that really matters is to distinguish between weekends and weekends: people often engage in different activities, go to different places, and maintain a different way of traveling during weekends and weekends. With forecasting in mind, we can now, by analysing marine information capacity and developing graphs and formulas, investigate whether we have an impact and whether that increases their impact on Uber passenger fares in New York City.



IV. HARWARE AND SOFTWARE REQUIREMENT

A. Uber Dataset

The dataset for this project I have used from the Kaggle website.

The dataset contains, roughly, four groups of files:

- Uber trip data from 2014 (April - September), separated by month, with detailed location information
- Uber trip data from 2015 (January - June), with less fine-grained location information
- Non-Uber FHV (For-Hire Vehicle) trips. The trip information varies by company, but can include day of trip, time of trip, pickup location, driver's for-hire license number, and vehicle's for-hire license number.
- Aggregate ride and vehicle statistics for all FHV companies (and, occasionally, for taxi companies)

B. Python

Python is an object oriented, scripted and interpreted language for both learning and real-

world programming. Python is a powerful high-level language created by Guido van Rossum. In this paper, we will provide an introduction to the main Python software tools used for Data science, Machine learning techniques and IOT. Briefly, this paper will first introduce Python as a language, and give introduction about Data science, Machine learning and IOT, and then describe packages that are popular in the Data science and Machine learning communities, such as NumPy, SciPy, TensorFlow, Keras, Matplotlib etc. From there, we will move to show the importance of python for building IOT applications. We will use different code examples throughout. To aid the learning experience, execute following examples contained in this paper interactively using Jupiter notebooks.

- **NumPy:**

  NumPy will help us to manage multi-dimensional arrays very efficiently. Maybe it is difficult to do that directly, but since the concept is a crucial part of data science, many other libraries (well, almost all of them) are built on NumPy. Simply to say, without NumPy it is difficult to use Pandas, Matplotlib, SciPy or Scikit-Learn.

- **3-dimensional NumPy array:**

  But on the other hand, it also has a few well-implemented methods. It's quite to use NumPy random function, which is found slightly better than the random module of the standard library. And when it comes to simple predictive analytics tasks like linear or polynomial regression, NumPy Polyfit function will be favourite.

- **Matplotlib:**

  Data visualization is very important. Data visualization helps us to better understand the data, discover things that wouldn't discover in raw format and communicate findings more efficiently to others. The best and most well-known Python data visualization library is Matplotlib. It is not easy to use, but usually it provides many functionslikebarchart, scatterplot, pie chart, etc which are useful for projecting many dimensions of data.

- **Pandas:**

  Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named NumPy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution, from those that come with your operating system to commercial vendor distributions.

- **Seaborn**

  Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and colour palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

Algorithm

- **K-mean Clustering:**

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be

created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

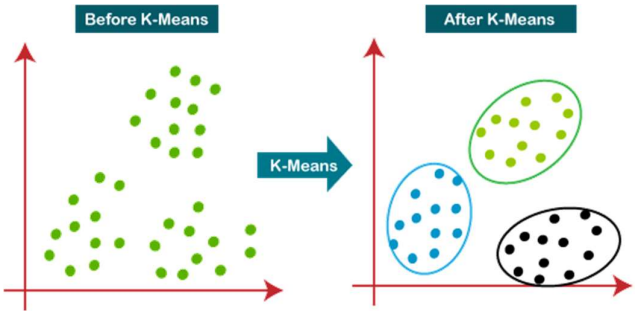
The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K centre points or centroids by an iterative process.
- Assigns each data point to its closest k-centre. Those data points which are near to the particular k-centre, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

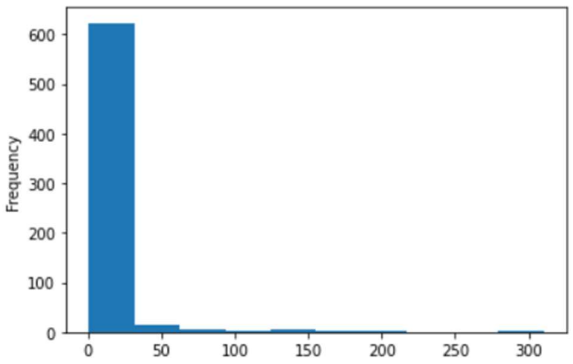
The below diagram explains the working of the K-means Clustering Algorithm:

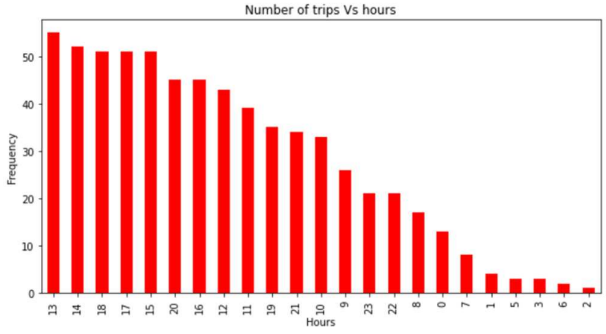


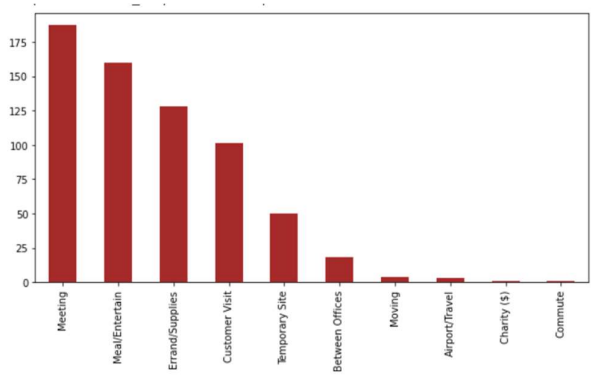
V. CODE

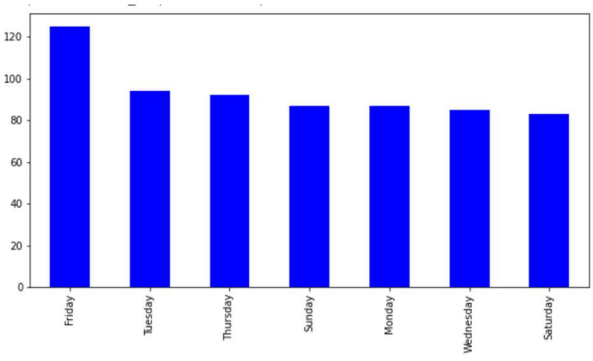
```
import pandas as pd
import numpy as np
import datetime
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import calendar

data=pd.read_csv('/content/Uber Drives -
 .csv')

data.head()

data.isnull().any()

data.isnull().sum()

data=data.dropna()

data.isnull().sum()

data['START_DATE*'] = pd.to_datetime(data
['START_DATE*'], format="%m/%d/%Y %H:%M")
data['END_DATE*'] = pd.to_datetime(data['
END_DATE*'], format="%m/%d/%Y %H:%M")

hour=[]
day=[]
dayofweek=[]
month=[]
weekday=[]
for x in data['START_DATE*']:
    hour.append(x.hour)
    day.append(x.day)
    dayofweek.append(x.dayofweek)
    month.append(x.month)
    weekday.append(calendar.day_name[dayo
fweek[-1]])
data['HOUR']=hour
data['DAY']=day
data['DAY_OF_WEEK']=dayofweek
data['MONTH']=month
data['WEEKDAY']=weekday

data.head()

data['CATEGORY*'].value_counts()

sns.countplot(x='CATEGORY*',data=data)
```

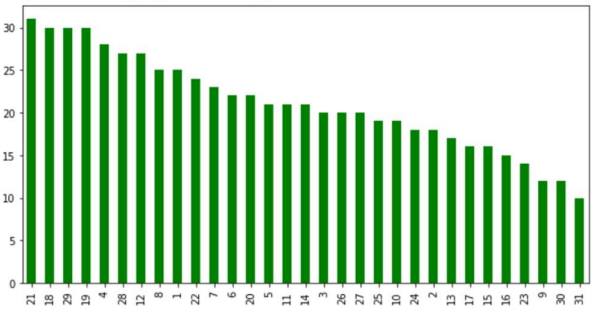
```
data['MILES*'].plot.hist()

hours = data['START_DATE*'].dt.hour.value
_counts()
hours.plot(kind='bar',color='red',figsize
=(10,5))
plt.xlabel('Hours')
plt.ylabel('Frequency')
plt.title('Number of trips Vs hours')

data['PURPOSE*'].value_counts().plot(kind
='bar',figsize=(10,5),color='brown')

data['WEEKDAY'].value_counts().plot(kind=
'bar',figsize=(10,5),color='blue')


data['DAY'].value_counts().plot(kind='bar
',figsize=(10,5),color='green')

data['MONTH'].value_counts().plot(kind='b
ar',figsize=(10,5),color='black')

data['START*'].value_counts().plot(kind='
bar',figsize=(25,10),color='blue')
```





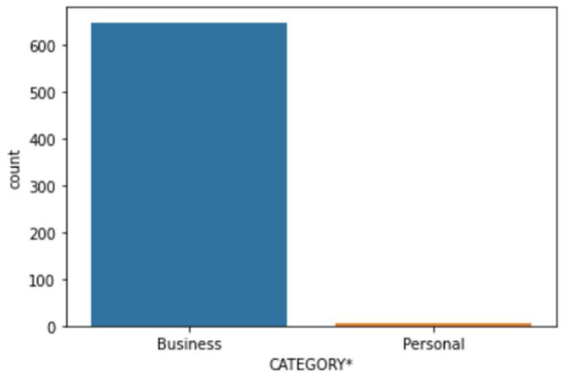


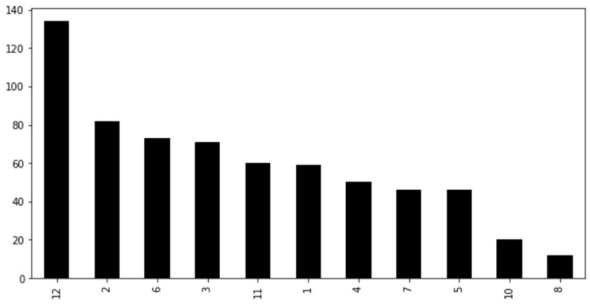




VI. OUTPUT

	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





VII. FUTURE SCOPE

With more than 8 million users, 1 billion Uber trips and 160,000+ people driving for Uber across 449 cities in 66 countries – Uber is the fastest growing start-up standing at the top of its game. Tackling problems like poor transportation infrastructure in some cities, unsatisfactory customer experience, late cars, poor fulfilment, drivers denying to accept credit cards and more –Uber has "eaten the world" in less than 5 years and is a remarkable name to reckon when it comes to solving problems for people in transportation.

If you have ever booked an Uber, you might know how simple the process is –just press a button, set the pickup location, request a car, go for a ride and pay with a click of a button. The process is simple but there is a lot going on behind the scenes. The secret key driving growth of the $51 billion start-up, is the big data it collects and leverages for insightful and intelligent decision making. While Uber moves people around the world without owning any cars, data moves Uber. With the foundation to build the most intelligent company on the planet by completely solving problems for riders –Big Data and Data Science are at the heart of everything Uber does - surge pricing, better cars, detecting fake rides, fake cards, fake ratings, estimating fares and driver ratings. Read on to understand how Uber makes clever use of big data and data science to reinvent transportation and logistics globally.

CONCLUSIONS

Here, we see the following topics:

- 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.Use of 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.

- UBER data analysis in Python.

VIII. REFERENCES

[1] Aguinaldo Bezerra, Gisliany Alves, Ivanovitch Silva, Pierangelo Rosati (2019). "A Preliminary Exploration of Uber Data as an Indicator of Urban Liveability". Research Gate. DOI: 10.1109/CyberSA.2019.8899714.

[2] Widdoes, Kaylene, "Case Study of Uber Data in the Central London Area" (2016). Honors Research Projects. 411.

[3] Uber Technologies, Inc., "Facts and figures," 2018.

[4] A. Ley and P. Newton, "Creating and sustaining liveable cities," in developing living cities: From analysis to action. World Scientific, 2010, pp. 191–229.

[5] R. Cervero, Transit-oriented development in the United States: Experiences, challenges, and prospects. Transportation Research Board, 2004, vol. 102.