Submitted by-
Name- Shawgat Ashrafi
ASTU Roll number- 170610107063
College- Assam Engineering College
Semester- 6th semester
Session- 2017-2021

Objective: The main objective was to analyse a given dataset of Uber rides using machine learning in order to:

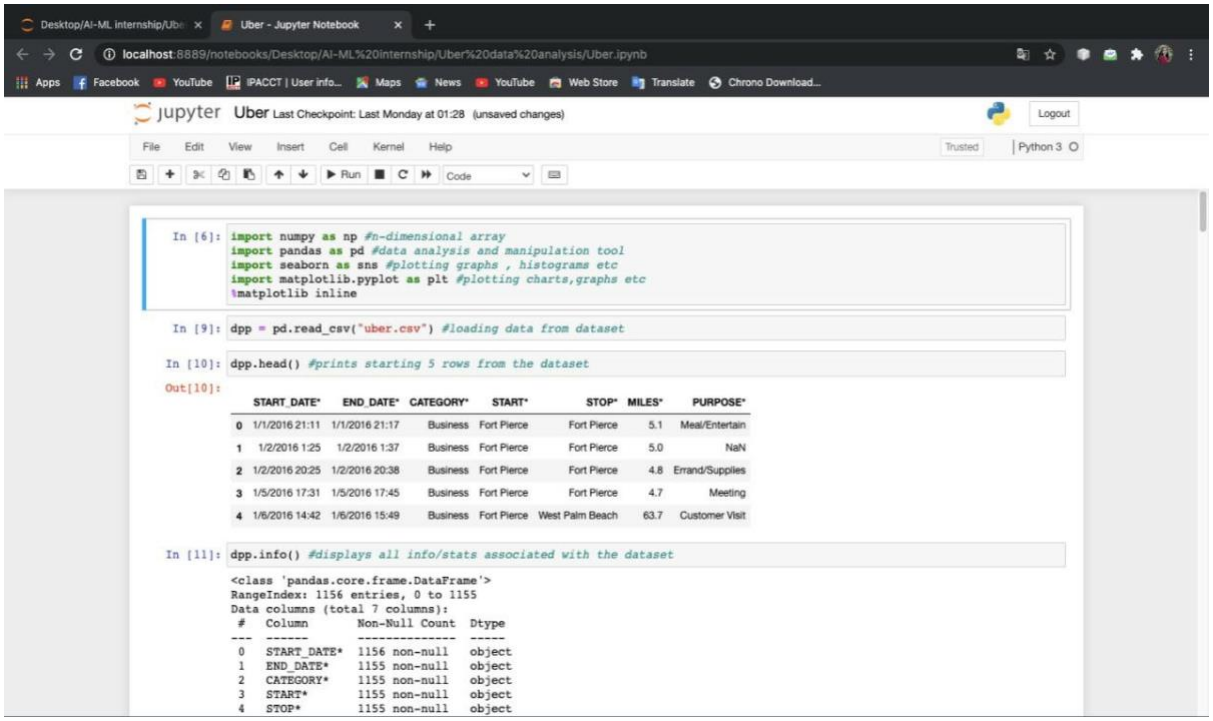
- Find traveling time and calculate the average speed of the trip.
- Visualize the data in terms of trips per hour of the day, per day of the week, and per month of the year.
- From the above step find out in which month highest trips are made.

***Strategies* used**: Initial strategy was to search the loaded dataset for columns with null values and drop them for a meaningful analysis. Then the columns with timestamp were converted according to convenience to find out travelling time, speed and rides based on hourly, weekly, monthly basis resp.

Finally the database was updated and the outcomes were plotted into bar graph.

Execution of the model:

1. Importing the libraries:

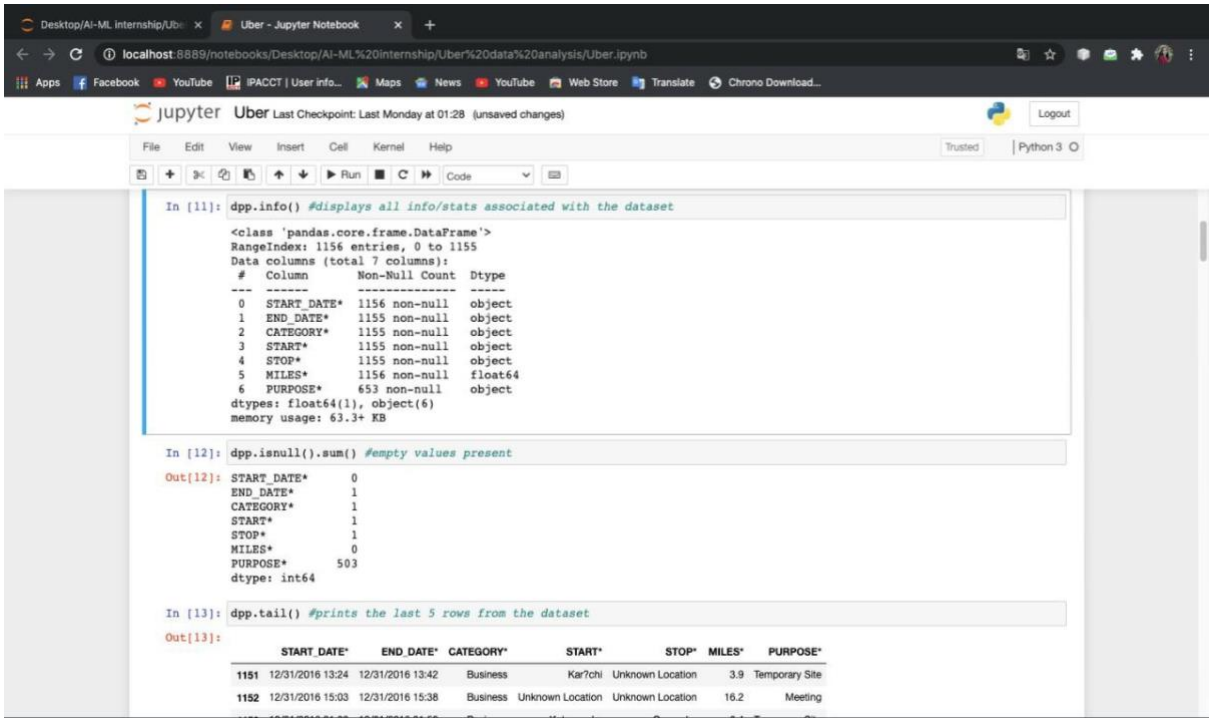


```
In [6]: import numpy as np #n-dimensional array
        import pandas as pd #data analysis and manipulation tool
        import seaborn as sns #plotting graphs , histograms etc
        import matplotlib.pyplot as plt #plotting charts,graphs etc
        %matplotlib inline

In [9]: dpp = pd.read_csv("uber.csv") #loading data from dataset

In [10]: dpp.head() #prints starting 5 rows from the dataset
Out[10]:
```

	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

```
In [11]: dpp.info() #displays all info/stats associated with the dataset
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1156 entries, 0 to 1155
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   START_DATE*  1156 non-null   object
 1   END_DATE*    1155 non-null   object
 2   CATEGORY*    1155 non-null   object
 3   START*       1155 non-null   object
 4   STOP*        1155 non-null   object
```

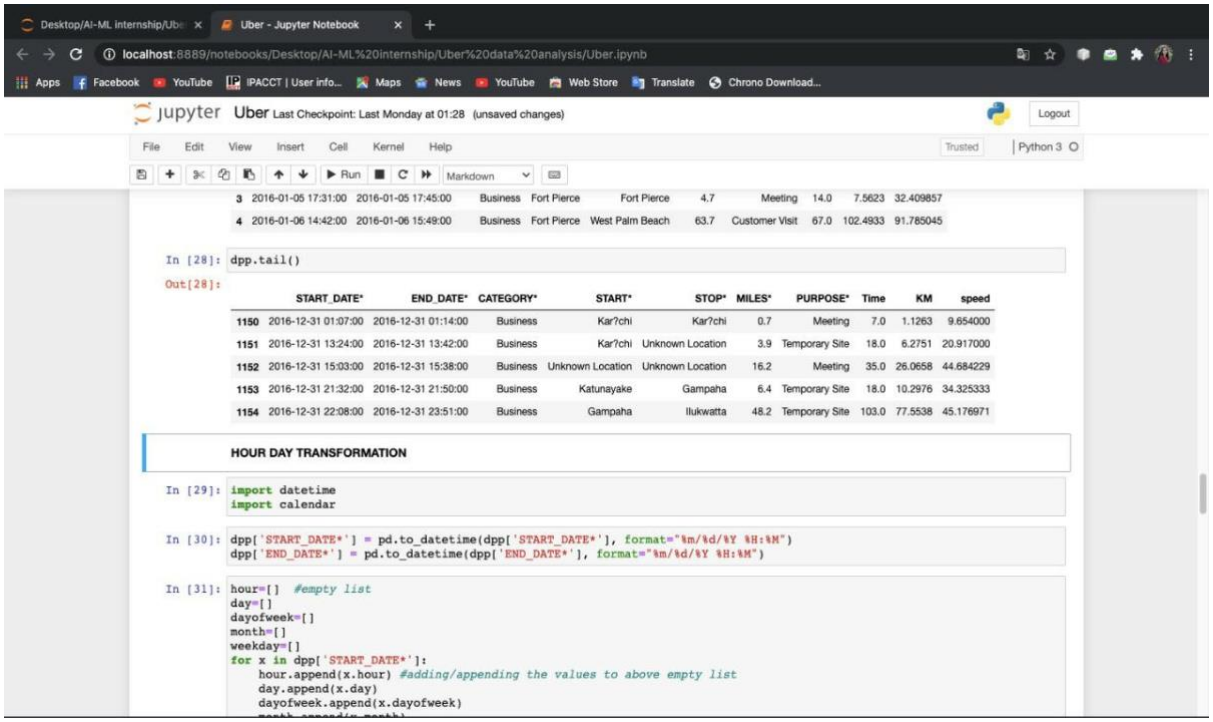
2. Display the information associated with the dataset:



```
In [11]: dpp.info() #displays all info/stats associated with the dataset
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1156 entries, 0 to 1155
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   START_DATE*  1156 non-null   object
 1   END_DATE*    1155 non-null   object
 2   CATEGORY*    1155 non-null   object
 3   START*       1155 non-null   object
 4   STOP*        1155 non-null   object
 5   MILES*       1156 non-null   float64
 6   PURPOSE*     653 non-null    object
dtypes: float64(1), object(6)
memory usage: 63.3+ KB

In [12]: dpp.isnull().sum() #empty values present
Out[12]: START_DATE*      0
         END_DATE*        1
         CATEGORY*        1
         START*           1
         STOP*            1
         MILES*           0
         PURPOSE*       503
         dtype: int64

In [13]: dpp.tail() #prints the last 5 rows from the dataset
Out[13]:
```

	START_DATE*	END_DATE*	CATEGORY*	START*	STOP*	MILES*	PURPOSE*
1151	12/31/2016 13:24	12/31/2016 13:42	Business	Kar?chi	Unknown Location	3.9	Temporary Site
1152	12/31/2016 15:03	12/31/2016 15:38	Business	Unknown Location	Unknown Location	16.2	Meeting

3. Hour-day transformation:



jupyter Uber Last Checkpoint: Last Monday at 01:28 (unsaved changes)

3	2016-01-05 17:31:00	2016-01-05 17:45:00	Business	Fort Pierce	Fort Pierce	4.7	Meeting	14.0	7.5623	32.409857
4	2016-01-06 14:42:00	2016-01-06 15:49:00	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit	67.0	102.4933	91.785045

In [28]:
```
dpp.tail()
```

Out[28]:

	START_DATE*	END_DATE*	CATEGORY*	START*	STOP*	MILES*	PURPOSE*	Time	KM	speed
1150	2016-12-31 01:07:00	2016-12-31 01:14:00	Business	Kar?chi	Kar?chi	0.7	Meeting	7.0	1.1263	9.654000
1151	2016-12-31 13:24:00	2016-12-31 13:42:00	Business	Kar?chi	Unknown Location	3.9	Temporary Site	18.0	6.2751	20.917000
1152	2016-12-31 15:03:00	2016-12-31 15:38:00	Business	Unknown Location	Unknown Location	16.2	Meeting	35.0	26.0658	44.684229
1153	2016-12-31 21:32:00	2016-12-31 21:50:00	Business	Katunayake	Gampaha	6.4	Temporary Site	18.0	10.2976	34.325333
1154	2016-12-31 22:08:00	2016-12-31 23:51:00	Business	Gampaha	Ilukwatta	48.2	Temporary Site	103.0	77.5538	45.176971

HOUR DAY TRANSFORMATION

In [29]:
```
import datetime
import calendar
```

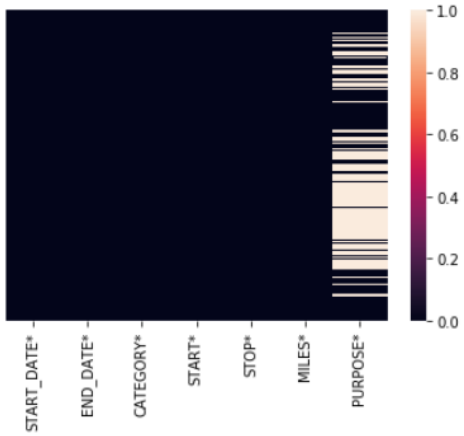
In [30]:
```
dpp['START_DATE*'] = pd.to_datetime(dpp['START_DATE*'], format="%m/%d/%Y %H:%M")
dpp['END_DATE*'] = pd.to_datetime(dpp['END_DATE*'], format="%m/%d/%Y %H:%M")
```

In [31]:
```
hour=[]  #empty list
day=[]
dayofweek=[]
month=[]
weekday=[]
for x in dpp['START_DATE*']:
    hour.append(x.hour) #adding/appending the values to above empty list
    day.append(x.day)
    dayofweek.append(x.dayofweek)
```

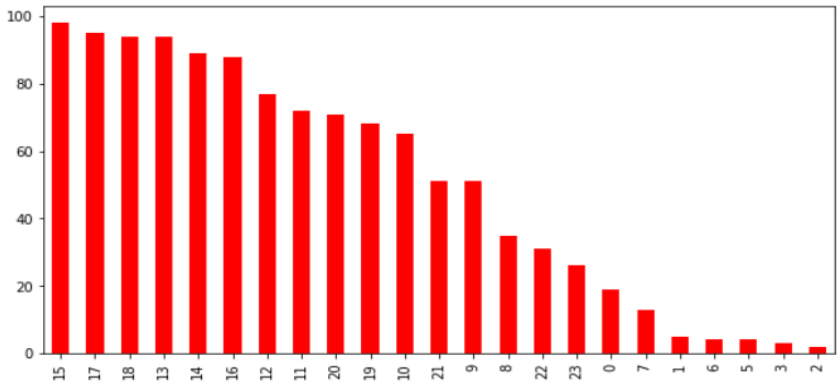
Output of plotted graphs:

Heatmap:

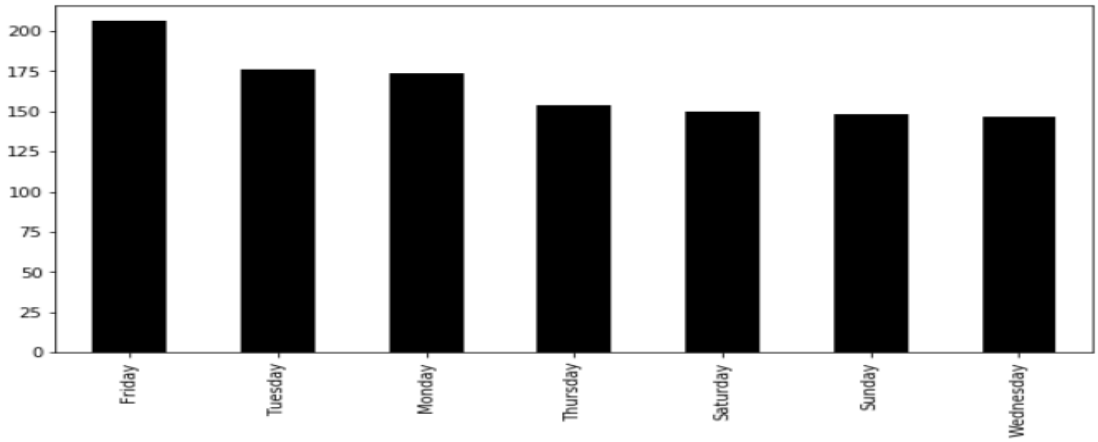
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x170503e1780>



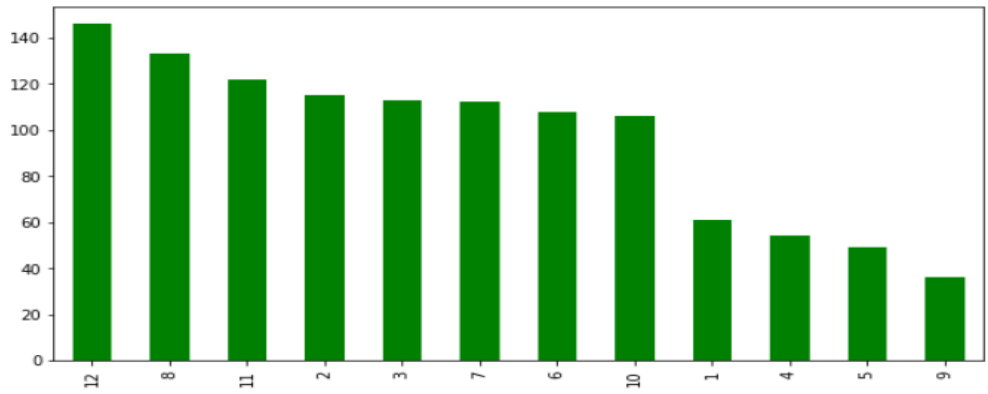
Hour graph:



Weekday graph:



Month of trip graph:



From the above graph we find the highest number of trips take place in December.