Uber Pickups Analysis

Tequed Labs
Invent - Innovate - Iterate

DONE BY: B DEVA DEEKSHITH

INTERNSHIP ID: TLS21A1200

INDEX

INTRODUCTION

Uber Technologies, Inc., commonly known as Uber, is an American technology company. Its services include ride-hailing, food delivery (Uber Eats and Postmates), package delivery, couriers, freight transportation, and, through a partnership with Lime, electric bicycle and motorized scooter rental. The company is based in San Francisco and has operations in over 900 metropolitan areas worldwide. It is one of the largest firms in the gig economy.

Uber is estimated to have over 93 million monthly active users worldwide. In the United States, Uber has a 71% market share for ride-sharing and a 22% market share for food delivery. Uber has been so prominent in the sharing economy that changes in various industries as a result of Uber have been referred to as uberisation, and many startups have described their offerings as "Uber for X".

MOTIVE

To analyze the data of the customer rides and visualize the data to find insights that can help improve business. Data analysis and visualization is an important part of data science. They are used to gather insights from the data and with visualization you can get quick information from the data.

System requirement

1. Hardware requirements:

- **Processor : Intel core i7**
- **Ram size: 8 GB(minimum)**
- **Hard disk: 500GB**

2. Software Requirements:

- **Operating system: Windows 10**
- **Platform: IDLE shell**
- **Language used : Python**
- **Domain: Data science**

My Contribution To The Group Project

THIS PROJECT WAS DEVELOPED BY MY TEAM (MYSELF AND JASWANTH), WE ARE EQUALLY RESPONSIBLE FOR THESE PROJECT . WE HAVE BEEN USING ALL THE CONCEPTS THOUGHT IN THE INTERNSHIP.

WHEN WE FIRST BEGAN WORKING AS A COLLECTIVE NONE OF US WERE VERY SURE ABOUT HOW TO PROCEED; WE KNEW THAT FOR THIS TO WORK, EVERYONE WOULD NEED TO TAKE THE ROLES WITHIN THE GROUP.

MY ROLE WAS TO ANALYSIS THE DATASET FROM APRIL 2014 TO SEPTEMBER 2014.

UBER TRIP DATA FROM 2014 ANALYSIS

- **DATA ACQUISITION**
- **DATA CLEANING**
- **DATA INFORMATION**
- **VISUALIZATION**
- **COMPARISON , MEASUREMENT AND ANALYSIS**

EXPLANATION OF DATASET

The dataset contains, roughly, TWO 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.**

Uber trip data from 2014

There are six files of raw data on Uber pickups in New York City from April to September 2014. The files are separated by month and each has the following columns:

- `Date/Time` **: The date and time of the Uber pickup**
- `Lat` **: The latitude of the Uber pickup**
- `Lon` **: The longitude of the Uber pickup**
- `Base` **: The TLC base company code affiliated with the Uber pickup.**

These files are named:

- `uber-raw-data-apr14.csv`
- `uber-raw-data-aug14.csv`
- `uber-raw-data-jul14.csv`
- `uber-raw-data-jun14.csv`
- `uber-raw-data-may14.csv`
- `uber-raw-data-sep14.csv`

Uber trip data from 2015

Also included is the file `uber-raw-data-janjune-15.csv` This file has the following columns:

- **`Dispatching_base_num` : The TLC base company code of the base that dispatched the Uber.**
- **`Pickup_date` : The date and time of the Uber pickup**
- **`Affiliated_base_num` : The TLC base company code affiliated with the Uber pickup.**
- **`locationID` : The pickup location ID affiliated with the Uber pickup**

These files are named:

- **uber-raw-data-janjune-15.csv**

Uber Basic Data Analysis for April to September 2014.

- **DATA ACQUISITION**

1. Importing Modules

```
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as ss
```

2. Loading Data

```
apr14=pd.read_csv("uber-raw-data-apr14.csv")
may14=pd.read_csv("uber-raw-data-may14.csv")
jun14=pd.read_csv("uber-raw-data-jun14.csv")
jul14=pd.read_csv("uber-raw-data-jul14.csv")
aug14=pd.read_csv("uber-raw-data-aug14.csv")
sep14=pd.read_csv("uber-raw-data-sep14.csv")
data=apr14.append([may14,jun14,jul14,aug14,sep14],ignore_index=True)
```

3. Preparation of data

```
data['Date/Time'] = pd.to_datetime(data['Date/Time'])
data['Month'] = data['Date/Time'].dt.month_name()
data['Weekday'] = data['Date/Time'].dt.dayofweek
data['Day'] = data['Date/Time'].dt.day
data['Hour'] = data['Date/Time'].dt.hour
data['Minute'] = data['Date/Time'].dt.minute
```

4. Data information

```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4534327 entries, 0 to 4534326
Data columns (total 9 columns):
 #   Column     Dtype
---  ------     -----
 0   Date/Time  datetime64[ns]
 1   Lat        float64
 2   Lon        float64
 3   Base       object
 4   Month      object
 5   Weekday    int64
 6   Day        int64
 7   Hour       int64
 8   Minute     int64
dtypes: datetime64[ns](1), float64(2), int64(4), object(2)
memory usage: 311.3+ MB
```

- **Visualization**

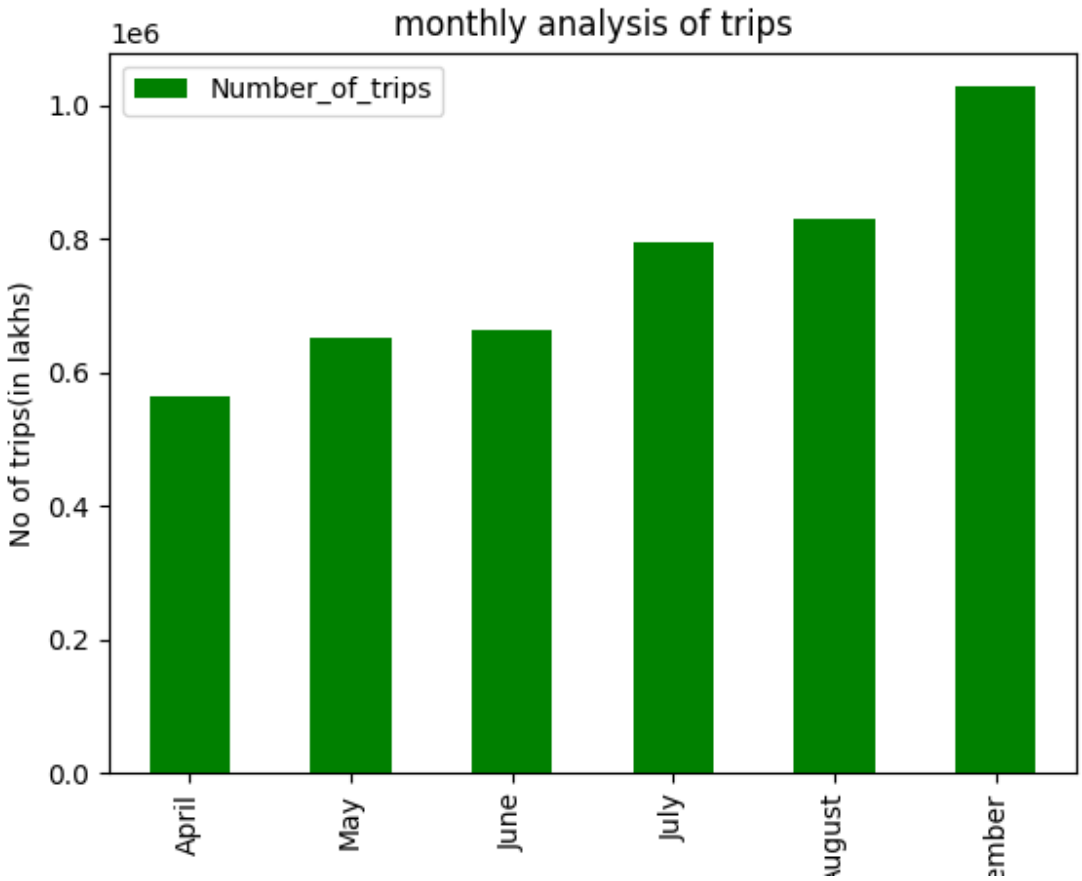
  Through our exploration we are going to visualize and analyse:

  - The number of trips by month
  - The number of trips by day
  - The number of trips by hour and month
  - The number of trips by hour and day

1. Trips by month

```
data_month = data.groupby(['Month'], sort=False).count()
d_month = pd.DataFrame({'Number_of_trips':data_month.values[:,0]}, index =
data_month.index)
d_month.plot(kind='bar',color="green")
plt.ylabel('No of trips(in lakhs)')
plt.title('monthly analysis of trips')
plt.show()
```

Plotting the result



Analysing the results

```
number_trips_apr= d_month.loc['April'].values
number_trips_sep = d_month.loc['September'].values
r_month = (((number_trips_sep - number_trips_apr) / number_trips_apr) *
100)[0]
r_month = round(r_month)
print('The ratio of the increase from April to September is {} %.'.format(r_month))
```

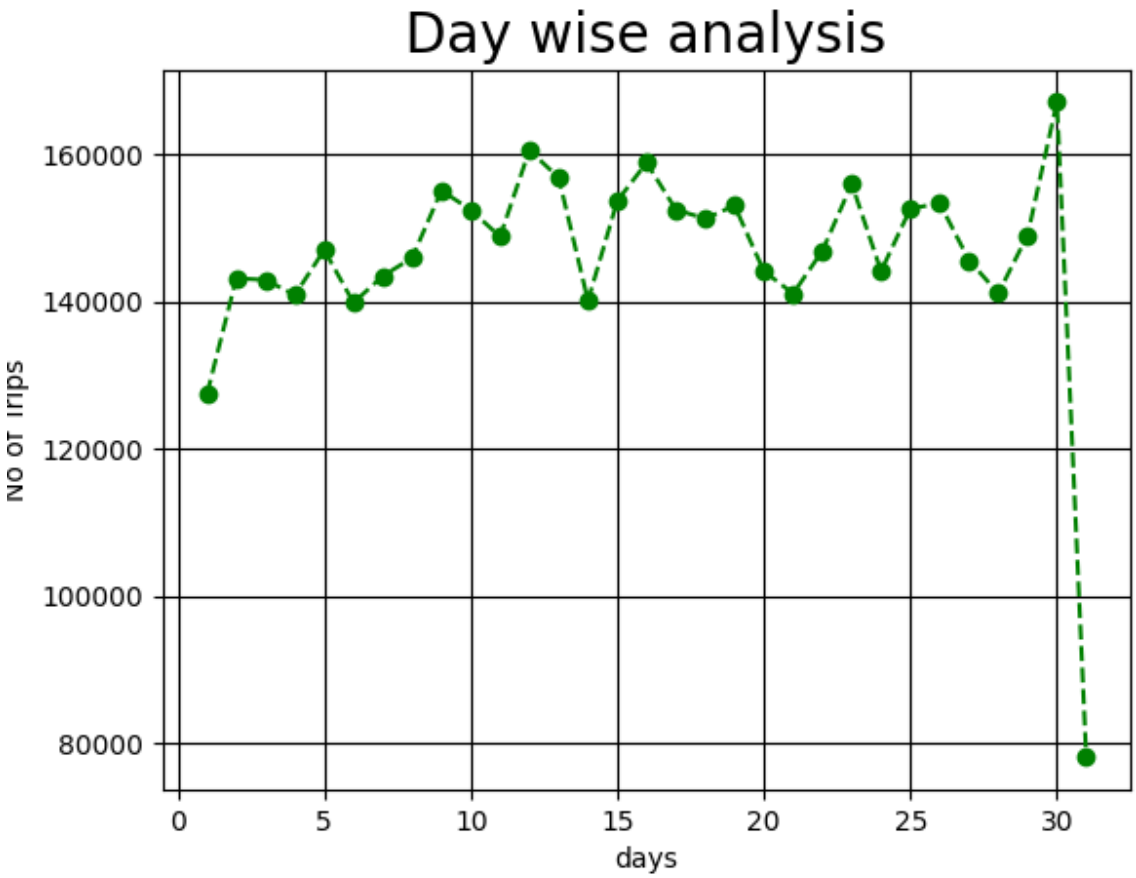
The ratio of the increase from April to September is 82 %.

From our results, we can say that from April to September 2014, Uber was in a continuous improvement process.

2. Trips by days

```
d_day = data.groupby(['Day']).count()
dj_day = pd.DataFrame({'Number_of_trips':d_day.values[:,0]}, index =
d_day.index)
plt.plot(dj_day, color = 'g', linestyle = 'dashed',marker = 'o')
plt.ylabel('No of Trips')
plt.xlabel('days')
plt.grid(color='black')
plt.title('Day wise analysis', fontsize = 20)
plt.show()
```

Plotting the result



Analysing the results

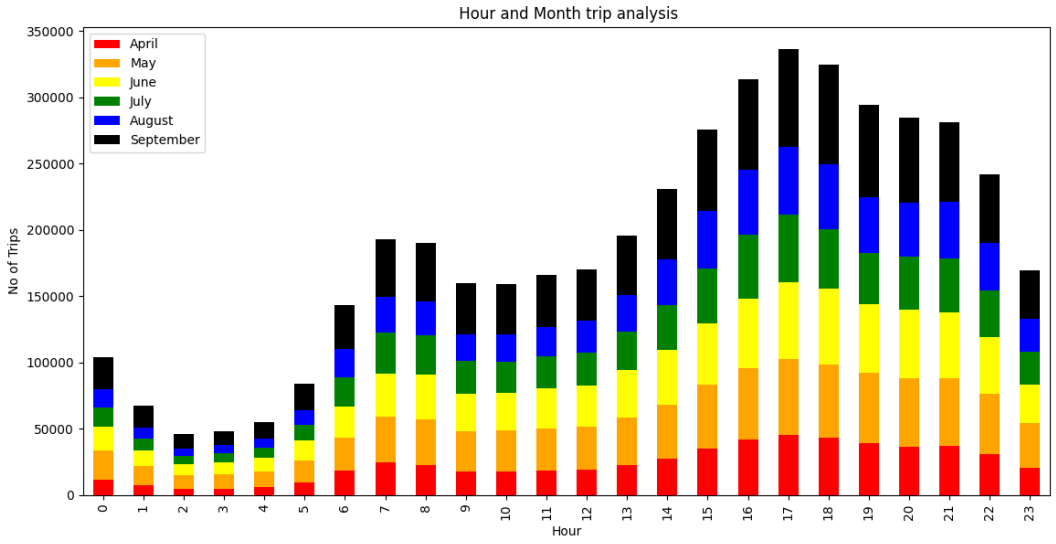
The overall observation of the graph depicts that the highest number of trips was observed in day 30 however, day by day no of trips differ marginally.

The lowest number of trips is on the 31st day because April, June and September have 30 days.

3. Trips by hours and month

```
d_hourmonth = data.groupby(['Hour','Month']).count()
h_month = pd.DataFrame({'Number_of_trips':d_hourmonth.values[:,1]}, index =
d_hourmonth.index)
h_month.reset_index(inplace= True)
data_hour_month = h_month['Number_of_trips'].values.reshape(24,6)
h_month = pd.DataFrame(data = data_hour_month, index =
h_month['Hour'].unique(), columns = data['Month'].unique())
c=['red','orange','yellow','green','blue','black']
h_month.plot(kind='bar', stacked=True,color=c)
plt.xlabel('Hour')
plt.ylabel('No of Trips')
plt.title(' Hour and Month trip analysis')
plt.show()
```

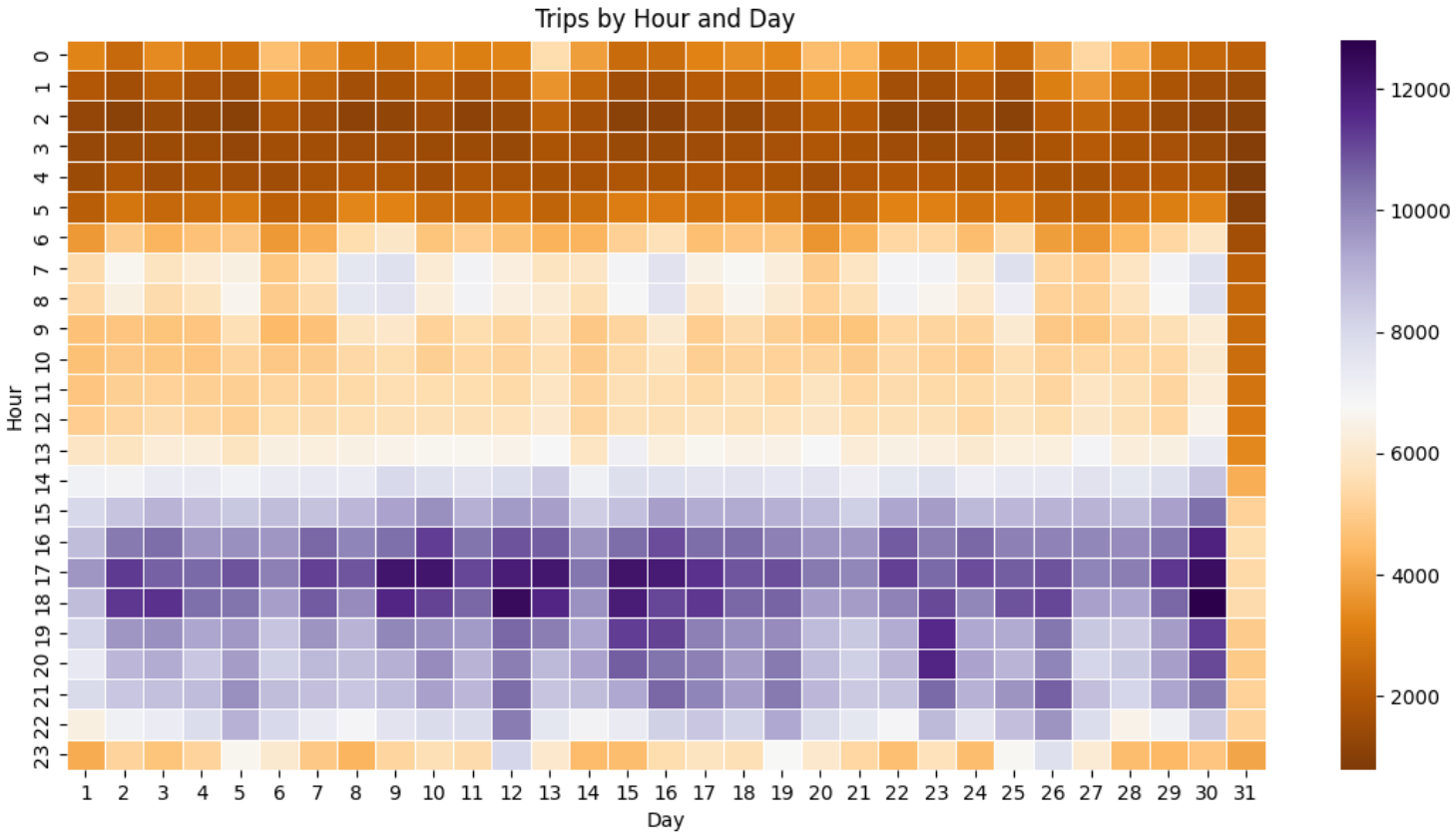
Plotting the result



5. Trips by hour and days

```
def c(row):
    return len(row)
dd = data.groupby('Hour Day'.split()).apply(c).unstack()
plt.figure(figsize = (12,8))
from matplotlib import cm
ax = ss.heatmap(dd, cmap=cm.PuOr, linewidth = .5)
ax.set(title="Trips by Hour and Day")
plt.show()
```

Plotting the result



Analysing the results

We see that the number of trips is increasing throughout the day, with a peak demand in the evening between 16:00 and 18:00.

It corresponds to the time where employees finish their work and go home.

Uber Basic Data Analysis for January to June 2015

- **DATA ACQUISITION**

1. **Importing Modules**

```
import seaborn as sns
from matplotlib import cm
import matplotlib.pyplot as plt
import pandas as pd
```

2. **Loading Data**

```
data=pd.read_csv("uber-raw-data-janjune-15.csv")
```

3. Preparation of data

```
data['Pickup_date'] = pd.to_datetime(data['Pickup_date'])
data['Month'] = data['Pickup_date'].dt.month_name()
data['Weekday'] = data['Pickup_date'].dt.dayofweek
data['Day'] = data['Pickup_date'].dt.day
data['Hour'] = data['Pickup_date'].dt.hour
data['Minute'] = data['Pickup_date'].dt.minute
```

4. Data information

```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14270479 entries, 0 to 14270478
Data columns (total 9 columns):
 #   Column              Dtype
---  ------              -----
 0   Dispatching_base_num  object
 1   Pickup_date         datetime64[ns]
 2   Affiliated_base_num   object
 3   locationID          int64
 4   Month               object
 5   Weekday             int64
 6   Day                 int64
 7   Hour                int64
 8   Minute              int64
dtypes: datetime64[ns](1), int64(5), object(3)
memory usage: 979.9+ MB
```

- ## Visualization

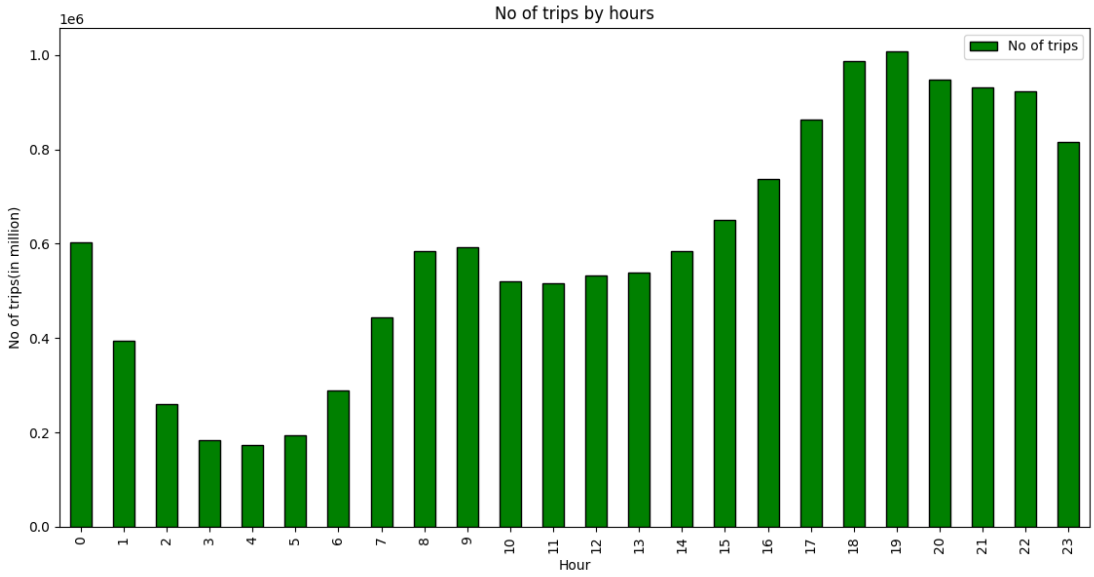
Through our exploration we are going to visualize and analyse:

- The number of trips by hours
- The number of trips by hours and weekdays
- The number of trips by months

1. Trips by hours

```
data_hour = data.groupby(['Hour']).count()
dat_hour = pd.DataFrame({'No of trips':data_hour.values[:,0]}, index =
data_hour.index)
import matplotlib.pyplot as plt
dat_hour.plot(kind='bar', figsize=(8,6),color='green', edgecolor='black')
plt.ylabel('No of trips(in million)')
plt.title("No of trips by hours")
plt.show()
```

Plotting the result



Analysing the results

The highest number of trips per hour is over 1 million trips, which corresponds to the peak hour 19:00.

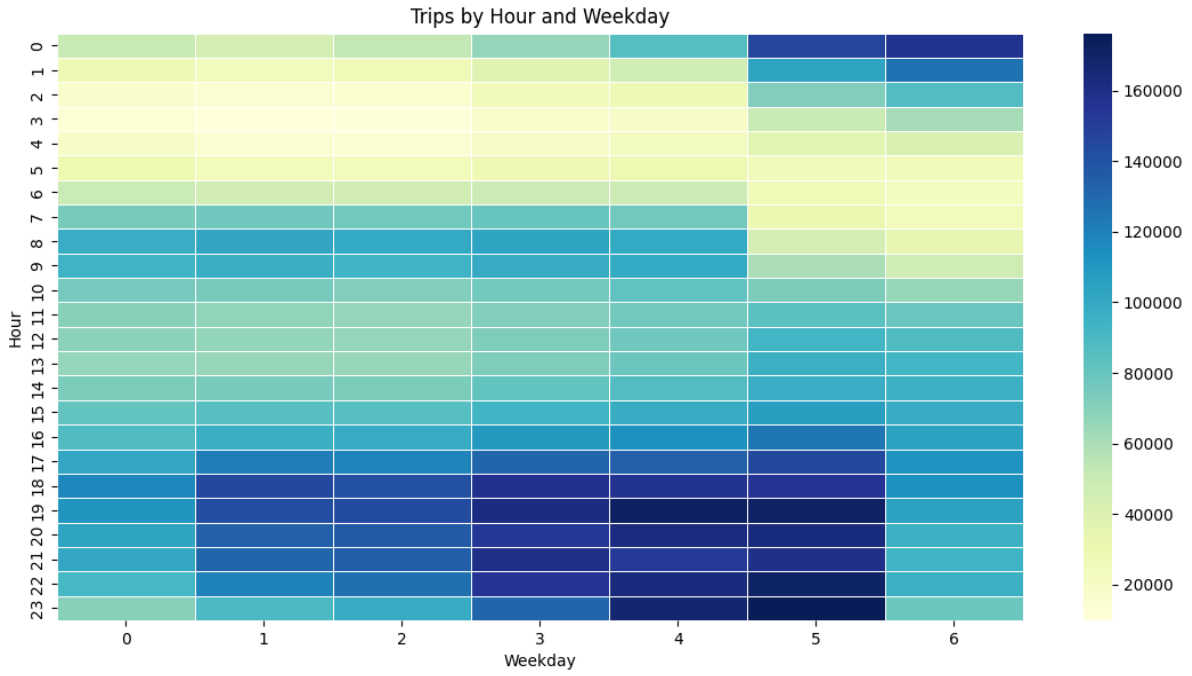
We observe that the number of trips are higher around 17:00 and 22:00, with a spike at 17:00. It matches the end of a working day in the United States (16:30), the time when the workers go home.

We can say that the majority of Uber's clients are workers.

2. Trips by hours and weekdays

```
def ows(rows):
    return len(rows)
hour_weekday = data.groupby('Hour Weekday'.split(), sort =
True).apply(ows).unstack()
import seaborn as sns
from matplotlib import cm
plt.figure(figsize = (12,8))
dd = sns.heatmap(hour_weekday, cmap=cm.YlGnBu, linewidth =
.5)
dd.set(title="Trips by Hour and Weekday");
plt.show()
```

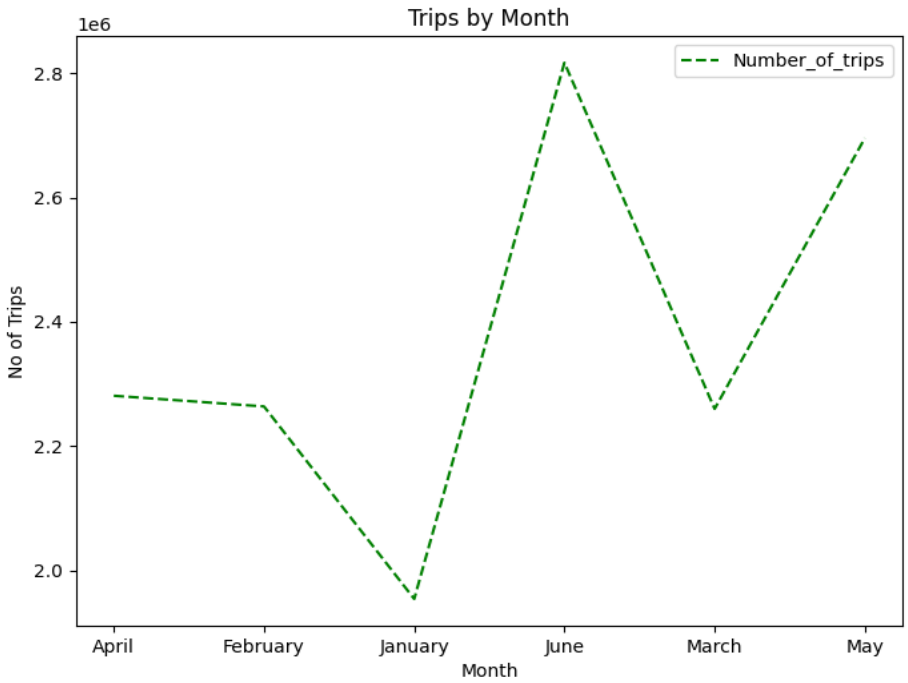
Plotting the result



3. Trips by months

```
nth_grouped = data.groupby(['Month'], sort=True).count()
dd_month = pd.DataFrame({'Number_of_trips':nth_grouped.values[:,0]},
index = nth_grouped.index)
dd_month.plot(kind='line', figsize=(8,6),color="green",linestyle =
'dashed')
plt.ylabel('No of Trips')
plt.title('Trips by Month')
plt.show()
```

Plotting the result



CONCLUSION

Through our examination of the Uber Pickups in New York City informational collection in 2014 and 2015, we figured out how to get the accompanying informations:

1. **From our outcomes, we can say that from April to September 2014, Uber was in a ceaseless improvement measure.**
2. **The general perception of the diagram portrays that the largest number of excursions was seen in day 30 in any case, step by step no of outings contrast imperceptibly.**
3. **The largest number of excursions each hour is more than 1 million outings, which relates to the pinnacle hour 19:00.**
4. **Individuals will in general utilize Uber late around evening time (around 12 PM) during ends of the week.**