MINING UBER DATASET

Project Github Link: https://github.com/msabhi/mining-uber-dataset

Objective:

The dataset collected by FiveThirtyEight contains data on over 4.5 million Uber pickups in New York City from April to September 2014. The data set is divided across six files i.e. one file per month.

Exploratory Analysis Goals:

- Preprocessing of data (rounding of the Latitude and Longitudes)
- To find out whether 4.5 million data points over 6 months are continuous or not
- To plot counts in histograms as well as plot Uber pickups on actual map and look at the coverage of NYC
- Analysis Uber pickup data over various months
- Finding the hotspot locations in the data. These are the locations where there are pickups more than a specified threshold

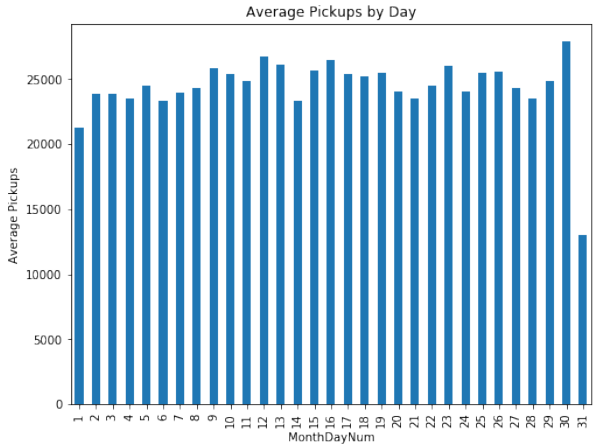
Pre-processing of raw data:

Grouping the data by pickup_latitude, pickup_longitude combination (precision upto 4 places of decimal)

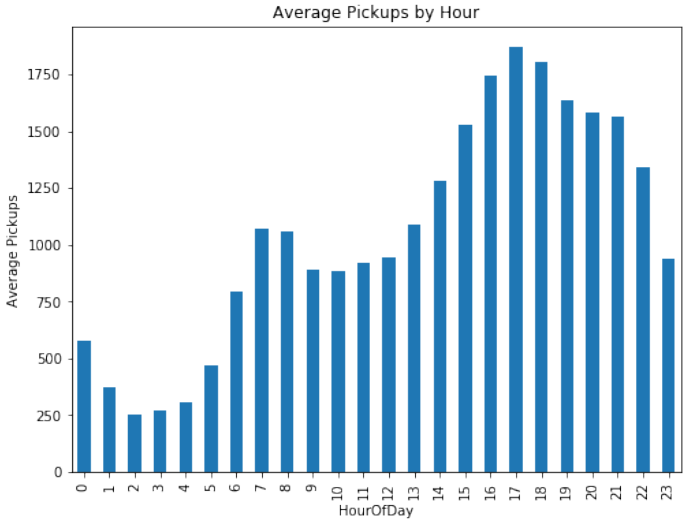
```
#!/bin/bash

echo ",pickup_latitude,pickup_longitude,num_pickups" > "$2"
awk -F "," '{printf (",%.4f,%.4f\n",$2,$3) }' "$1" | awk -F "," '{a[$0]++} END{for(i in a) print i","a[i]}' >> "$2"
```

Graphical Analysis:

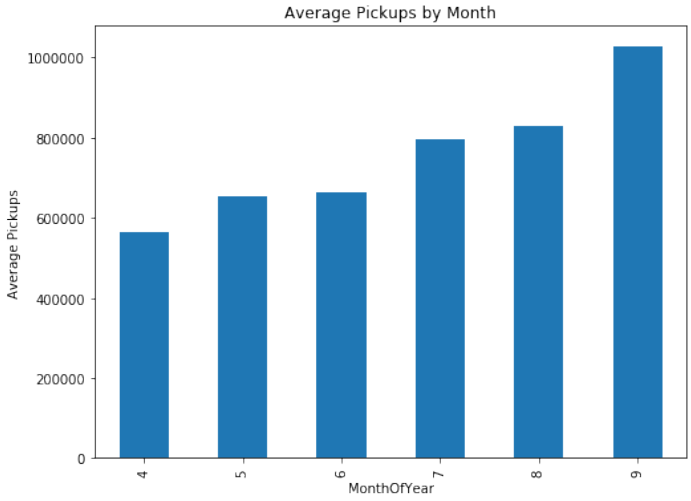
- **Average Pickups by Day (April)**
  Comparable bars over all days of a month suggest that Uber data is continuous over the data set. Also similar graphs have been observed for rest of the months



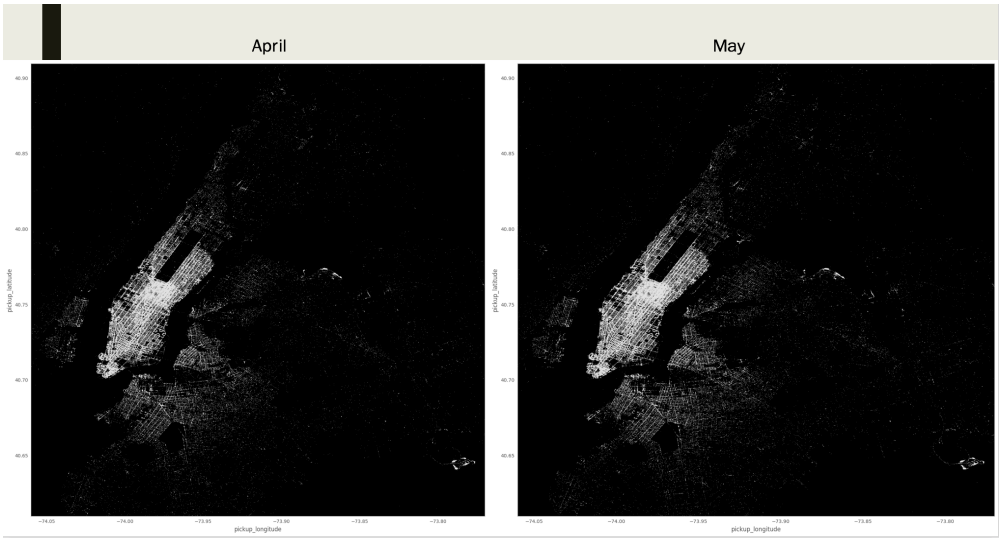
- **Average Pickups over a day**
  This graph is an average of all the days data over 6 months. This suggest the evening 5:00 pm are the peak hours for Uber.

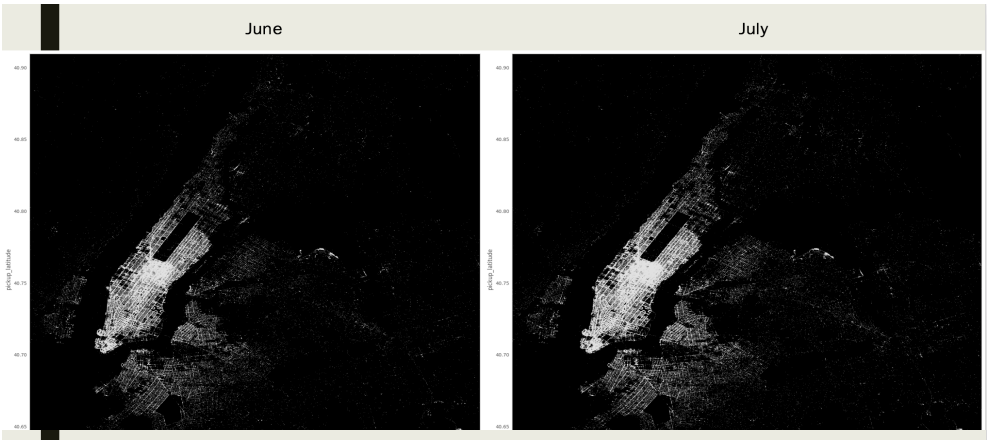


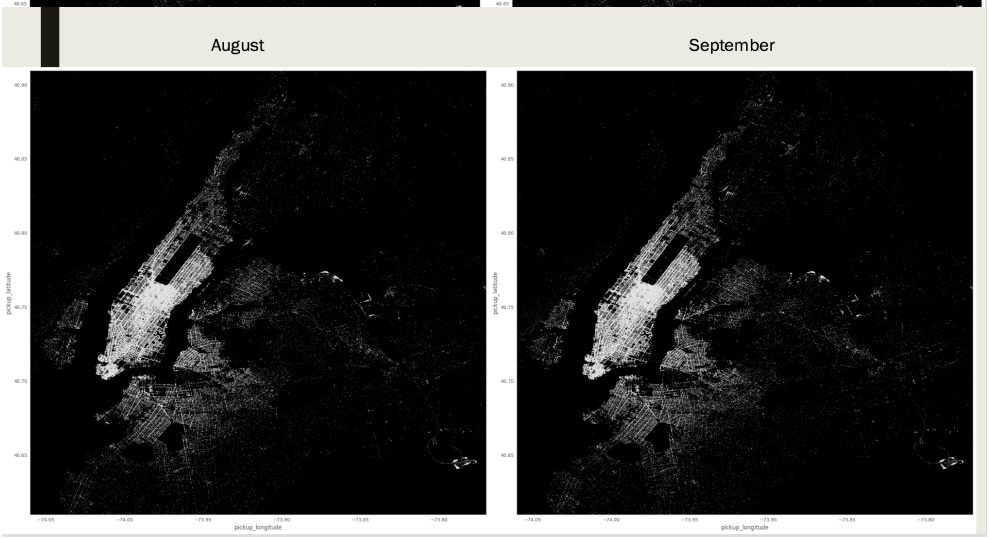
- **Average Pickups over 6 months (April to September)**
  The plot of cumulative pickups suggest that Uber business has grown over these set of months since average number of pickups have almost doubled from April to September



- **Geographical Plot (Density Plot)**
  Six density plots over each month consistently suggest that Manhattan is the area where Uber is most popular. Others incluse JFK Airport and LaGuardia Airport







- **Plotting 4.5 million data points:**
  The complete density plot carves out the map of New York City on a black canvas by geography. This further strengthens the range of geographical data in dataset.

Conclusions:

- We have sufficient data and it is continues in time
- Geographic plot actually carves out the map of NYC on a black canvas so the data set covers New York cities with its outer boroughs
- Plot of pickups over months suggest that monthly pickups are comparable and can be relatively analyzed to mine patterns

Team:

- Satya Akhil Chowdary Kuchipudi
- Sanil Jain
- Abhilash Mysore Somashekar
- Ravi Raj Wani