Mining Uber Dataset

Abhilash Mysore Somashekar
Raviraj Prakash Wani
Satya Akhil Chowdary Kuchipudi
Sanil Jain

1 April 2017

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

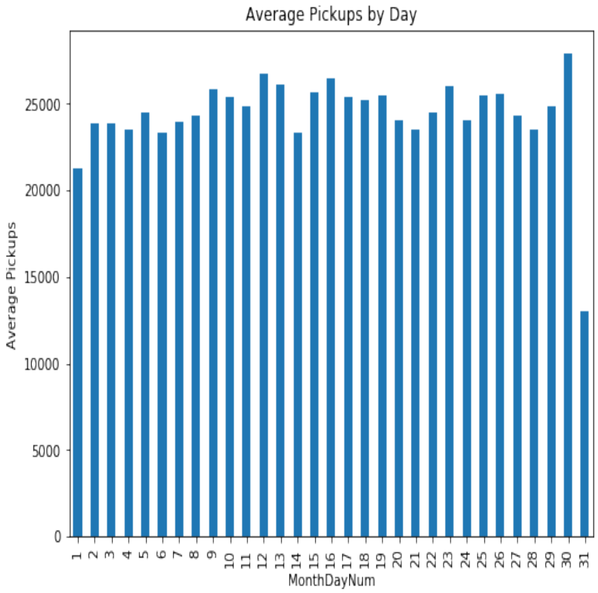
Exploratory Analysis Goals

- 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 thecoverage of NYC
- Analyze 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

Graphical Analysis:

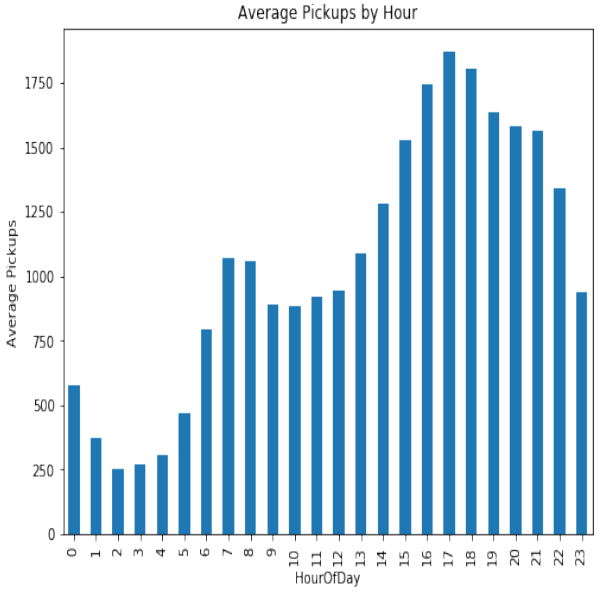
Average Pickups by Day (for all 6 months)

Comparable bars over all days of 6 months suggest that Uber data is continuous over the data set. Also similar graphs have been observed for each individual month



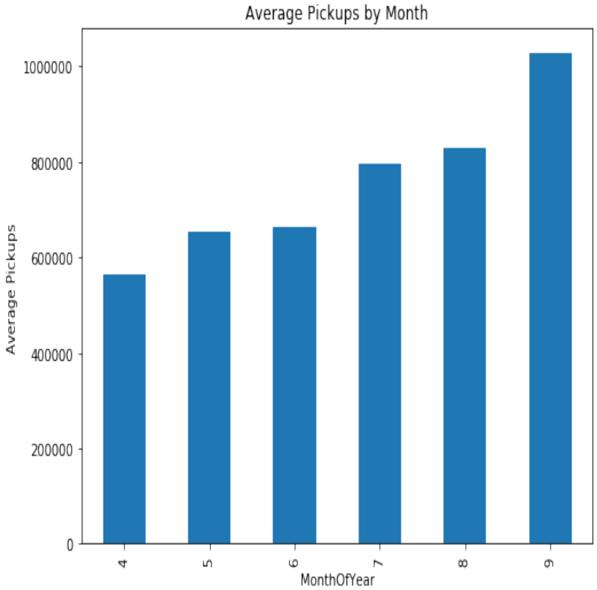
Average Pickups over a day

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.



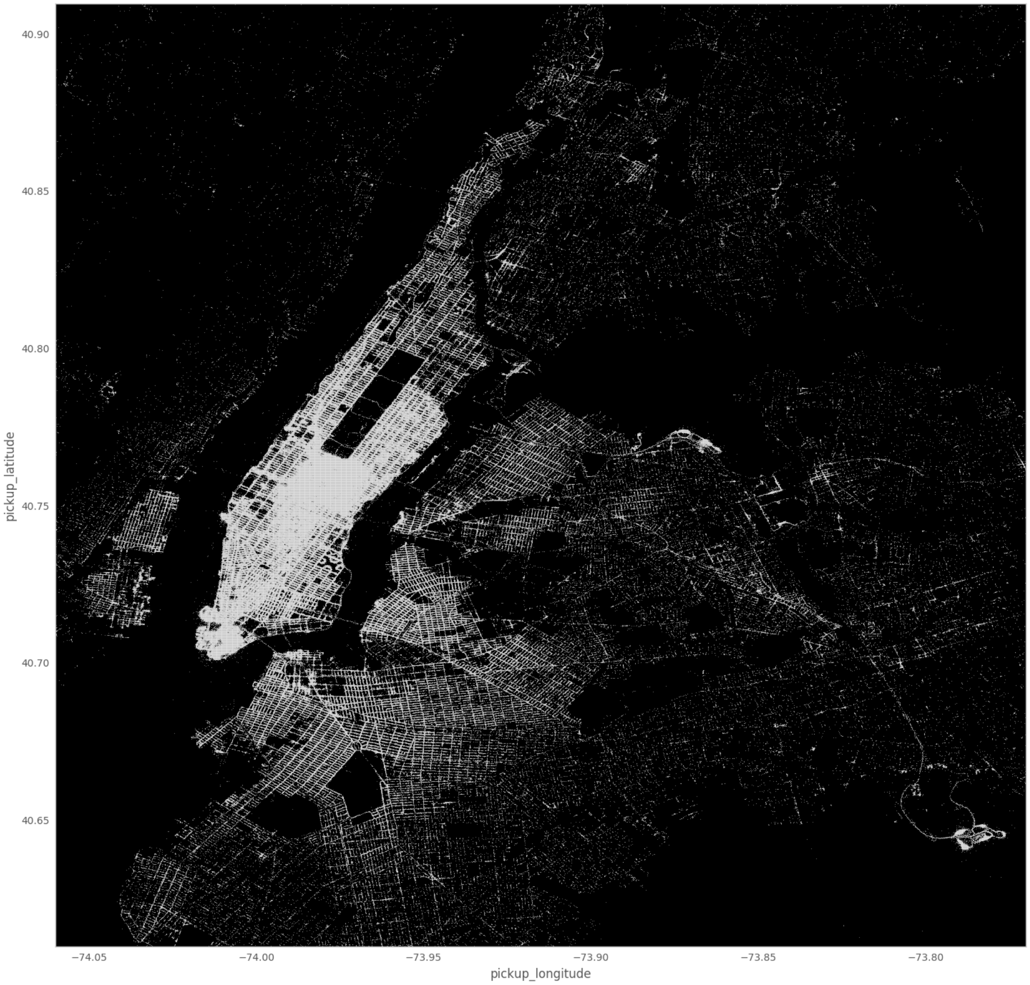
Average Pickups over 6 months (April to September)

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)

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.


40.90
40.85
40.80
40.75
40.70
40.65
pickup_latitude
−74.05
−74.00
−73.95
−73.90
−73.85
−73.80
pickup_longitude

Milestone 2
Following were the actions items decided for milestone 2

- Binning
- Decide algorithmic approach for mining the pickup pattern
- Understanding GPFlow

Binning : The uber-dataset for all the 6 months contains around 4.5 millions records. Binning helps in reducing the number of the input required for model creation. Also it helped us in decided the label required for regression. Binning was done based on following creitera.
Time based binning : All the pickup rides are divided into 24 hours slot as per pickup time.
Location based binning : Entire NYC areas was divided into 'n' clusters. We used spatial DBSCAN for finding this each area center point.

Spatial DBSCAN:
Kernel

$$K_{nm} = 1/\lambda \left(\phi\phi^T\right)$$

$$\tfrac{1}{\lambda}\overrightarrow{\phi}\left(\overrightarrow{X_n}\right)^T\overrightarrow{\phi}\left(\overrightarrow{X_m}\right) = \tfrac{1}{\lambda}l\left[\overrightarrow{X_n}, \overrightarrow{X_m}\right]$$

$$K := \phi\phi^T$$

As per Linear Rigression ::

$$\theta = (XX^T)^{-1}X^TY$$

Hence for Kernel Ridge Regression

$$E\left[w/y\right] = A^{-1}\phi^TY$$

$$A = \left(\phi^T \phi + \lambda I\right)$$

$$\phi = \phi\left(X\right)$$

$$K = 1/\lambda\left(\phi\phi^T\right)$$

$$A^{-1}\phi^T = \phi^T\left(K + \lambda I^{-1}\right)$$

$$E\left[f\left(X\star\right)/y\right] = \phi\left(x\star\right)^T E\left[w/y\right]$$