EAS 503 Final Project Report

Group 8:

Anish C Shah

Martin Jia

Yashas J Shamaraju

Title: NYC Uber Taxi Data Analysis

Abstract:

In this project we've tried to explore and analyze this interesting dataset using various quantitative data analysis methods with the use of Python libraries such as 'Pandas', 'Numpy', 'Matplotlib', 'Seaborn' and 'PyMySQL' for MySQL connectivity and 'MySQL server'. In our analysis we tried to do various tasks such as characterizing the demand based on identified patterns in the time series, estimating the value of the NYC market for Uber and its revenue growth, getting other insights about the usage of the service and also visualizing Uber's ridership growth in NYC during that period.

Introduction:

As people nowadays are using For-Hire Vehicles (FHVs) more than ever, it can give us various patterns for its usage and meaningful insights. This project aims to visualize, explore and analyze the NYC Uber Data.

Data Description:

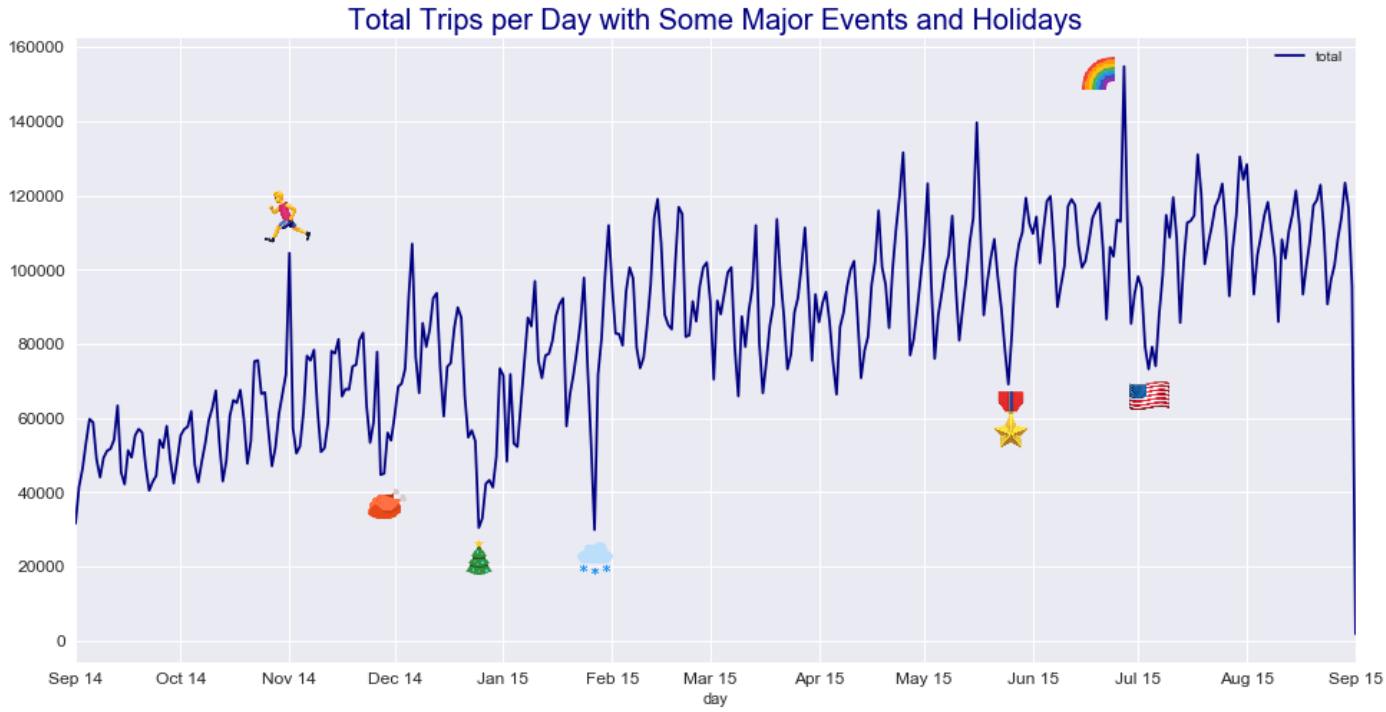
In early 2017, a dataset about Uber's ridership between September 2014 and August 2015 was released NYC Taxi and Limousine Commission (TLC). This dataset contains features such as trip distance, destination and duration that were not available in other sets released before. With a total of about 31 million entries, the data comprises one complete year of trips.

Data Cleaning:

First step which we carried out is data cleaning as this is a very large dataset. In this step we deleted all the null values present which were there maybe because of the system errors or some other factors such as if someone cancelled the ride the trip duration would be zero but there would be pickup and drop-off location. We also counted the number of 'extreme' trips i.e. the trips which were longer than 50 hours and assigned them an average duration value of trips which were less than 10 hours to simplify the analysis. We also took account only 365 days for convenience purposes which only excluded about 1852 cases in such a huge dataset.

Data Analysis:

- In our first analysis task we tried to analyze the demand based on identified patterns in the time series.

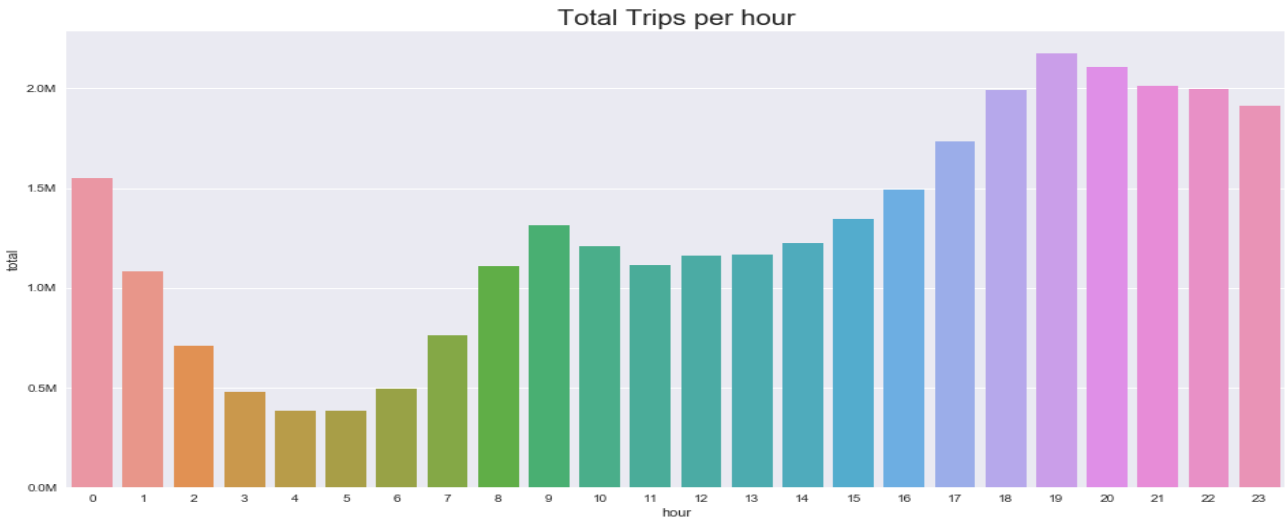


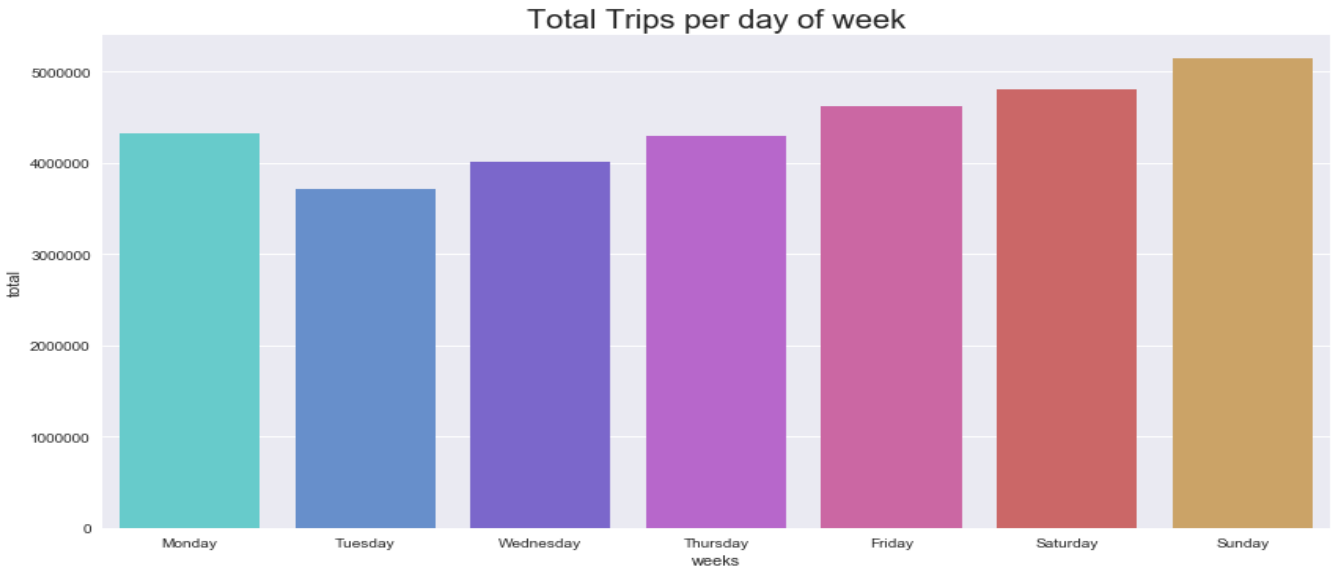
Interesting insight from the plot above we got is the effect on the number of trips the major events have. The negative impacts are related to the Thanksgiving, Christmas, Memorial Day and the Independence Day.

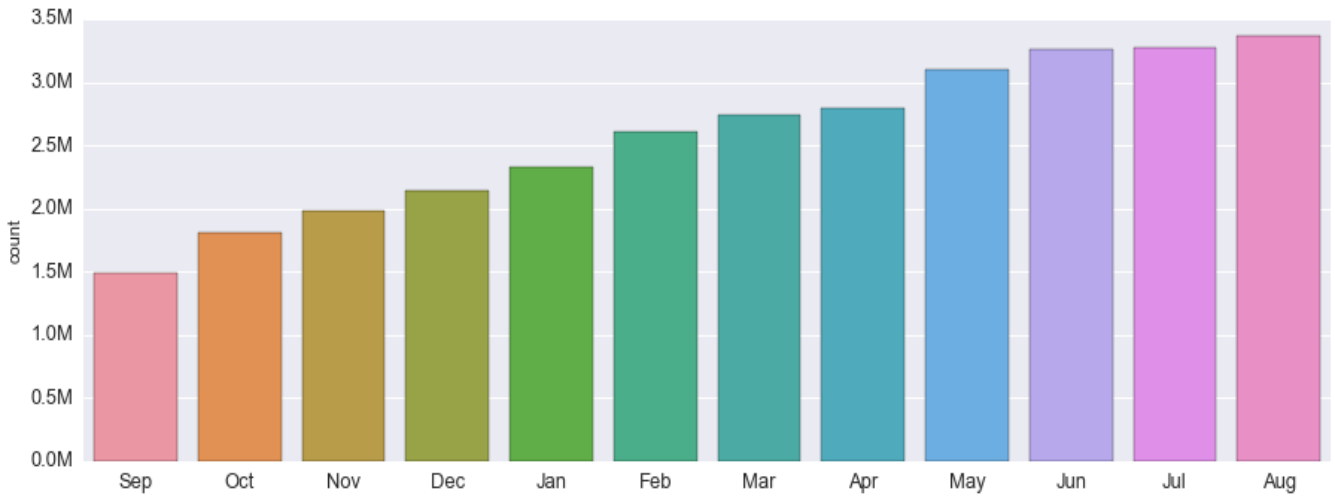
We can also see an apparently odd and very significant drop in the number of trips which is shown on January 27th. There was a curfew imposed by the mayor of NYC for preparation for a blizzard.

Other things the plot highlights which events have positively impacted the number of trips that year, with the Gay Pride Week and International Marathon standing out as the strongest contributors.

- In the next analysis task we explored the data to find out the trends in the demand for rides in the City.
  Here we can see from the bar charts below that the demand for Uber is higher from 4 PM until around midnight. Sunday has the highest demand. Seasonal effects are masked by the consistent month-to-month growth when looking at the total demand per month along the period of time analyzed.

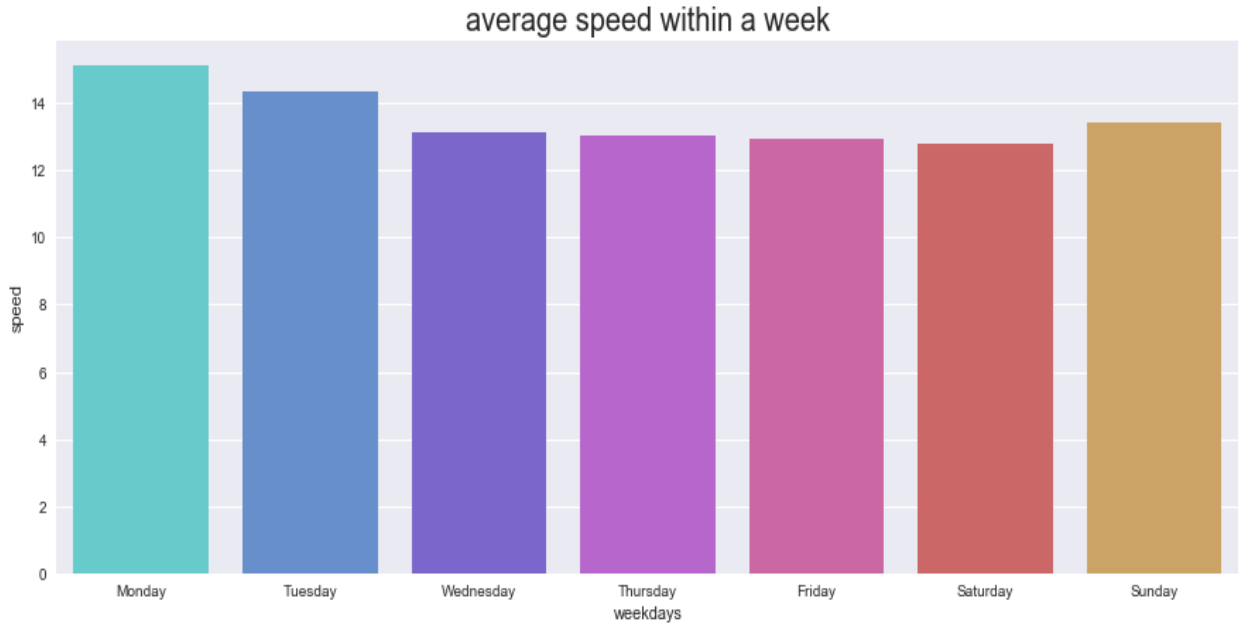

Total Trips per hour
total
2.0M
1.5M
1.0M
0.5M
0.0M
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
hour

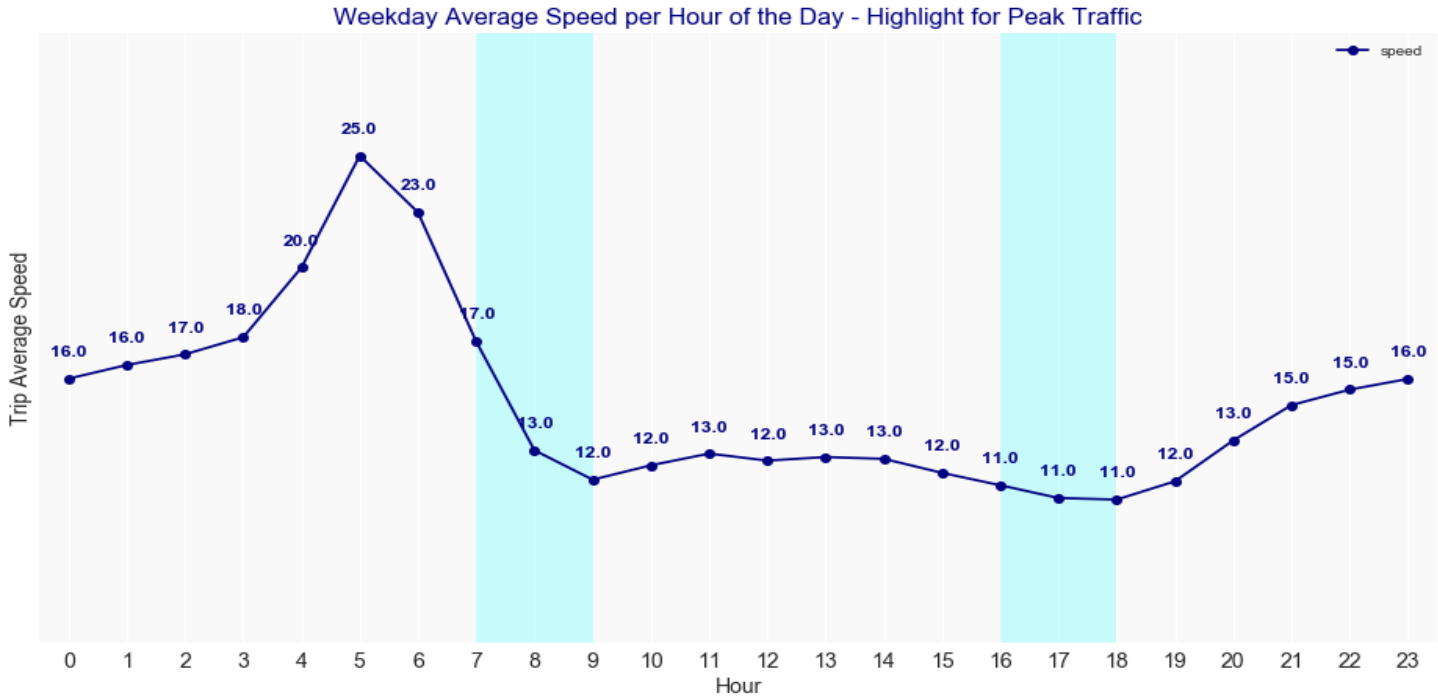

Total Trips per day of week
total
5000000
4000000
3000000
2000000
1000000
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
weeks


count
3.5M
3.0M
2.5M
2.0M
1.5M
1.0M
0.5M
0.0M
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug

- **Traffic Analysis:**

  In the next task we analyzed the traffic by calculating the speed of each ride using available attributes in the dataset. So by finding the average speed for each day of the week we tried to visualize how traffic will be on a particular day. As expected Monday has the better flow compared to other days of the week followed by Tuesday and then Sunday.

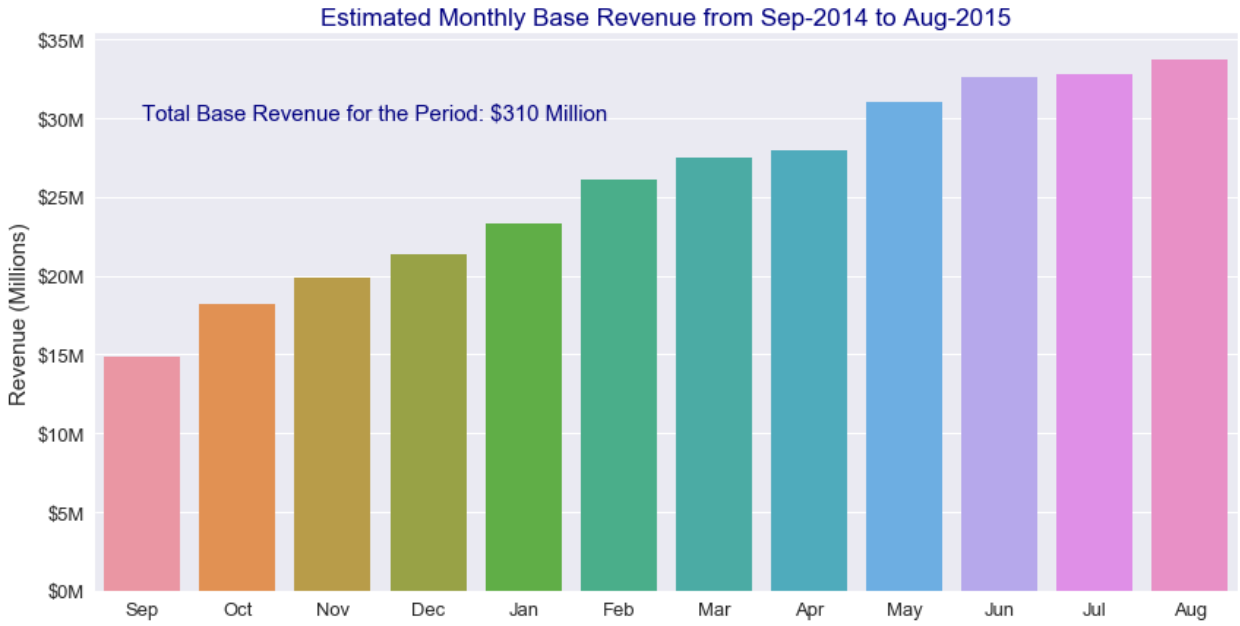


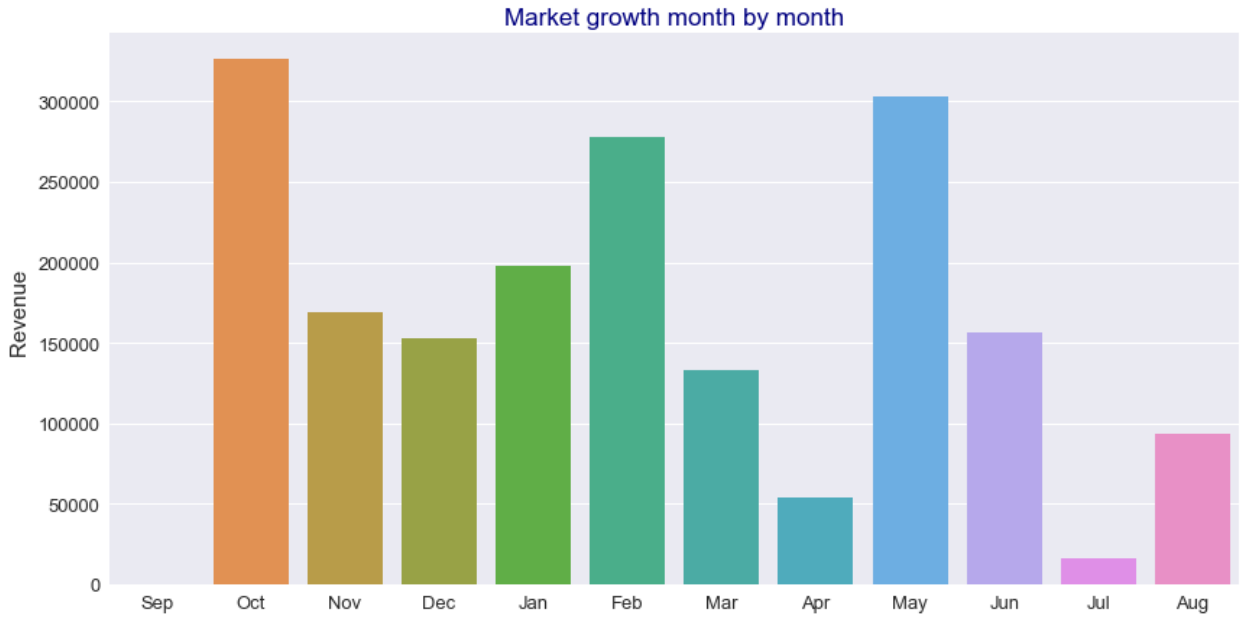


We allotted the times between 7-9 AM in the morning and 4-6 PM in the evening as the peak hours as we can see the traffic increasing exponentially during these times, the traffic is almost constant throughout the day and only improves after 8 PM in the evening.

Here we analyze the demand of rides per month during peak hours and off-peak hours. We can see that over 1/4th of the demand is the peak hours for Uber.

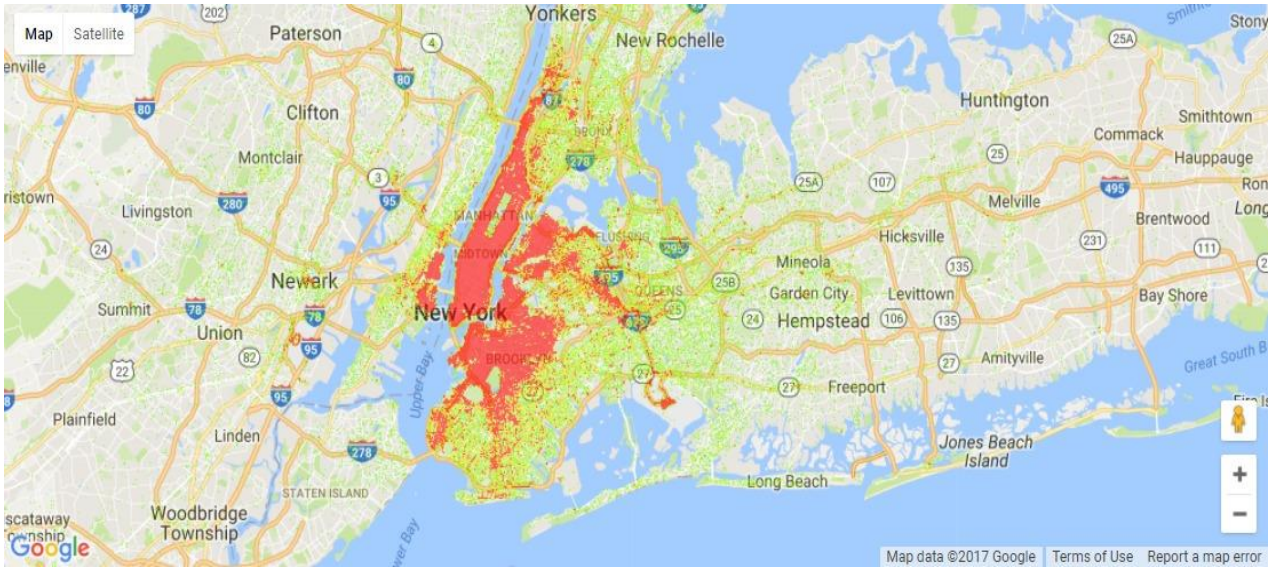
- In the next step of analysis we determine the value of the NYC market for Uber and its growth during the period. We calculated the base revenue for the period using the base value we got from http://uberestimate.com/prices/New-York-City/ for Uber X which is the basic Uber ride. So this is just the minimum estimate of the revenue. Uber released a statement some time ago that the average fare for this period in NYC was $27 so this makes the revenue approximately $595million.





This plot shows us the market growth for Uber during that period taking September as the initial parameter. We can see that the October 2014 recorded the highest growth for Uber in the following year.

- Last step we did was visualizing the pickup locations for finding which area was very good in business for Uber. For doing this we plotted a heatmap by using pickup co-ordinates. Below is the heatmap:



From the above heatmap we can see that Manhattan is the most popular area for Uber, Brooklyn being the second most popuar location. Some major pickups happening outside of the city seem to be in Jersey City, La Guardia airport and John F Kennedy airport which is far away from the city.

Conclusion:

So these were the meaningful insights we got after exploring, visualizing and performing various analysis methods on the Uber NYC data.

Future Directions:

After learning these many change points from the data we can move towards more machine learning stuff in the future such as forecasting demand for Uber and using other sources of data such as weather data with this dataset to get more insights which affect number of trips taken on a particular day.

References:

1. https://www.kaggle.com/dotman/data-exploration-and-visualization/notebook
2. http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
3. Dataset Link: https://s3.amazonaws.com/nyc-tlc/misc/uber_nyc_data.csv