**UBER DATA ANALYSIS**

Project Report

Project\_24 Team members:

Divija Y (17WH1A0425)

Deepthimayi P (17WH1A0445)

BVRIT Hyderabad College of Engineering for Women

Hyderabad, Telangana, India

**INTRODUCTION**

1. **1 Overview:**

Data storytelling is an important component of *Machine Learning* through which companies are able to understand the background of various operations. With the help of visualization, companies can avail the benefit of understanding the complex data and gain insights that would help them to craft decisions.

* 1. **Purpose:**

The purpose of this project is to analyze the Uber Pickups in New York City dataset. This is a data visualization project where we use visualisation libraries for understanding the data and for developing an intuition for understanding the customers who avail the trips. We are building a business model using which Uber can promote and develop their business.

* 1. **Dataset Description:**

This directory contains data on over 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015. All the files are as they were received on August 3, Sept. 15 and Sept. 22, 2015.

FiveThirtyEight obtained the data from the NYC Taxi & Limousine Commission(TLC) by submitting a Freedom of Information Law request on July 20, 2015.

**The Data:**

The dataset contains, roughly, three 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.
* Aggregate ride and vehicle statistics for all FHV companies (and, occasionally, for taxi companies).

**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.

**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 (Taxi & Limousine Commission) 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.

**The Base codes are for the following Uber bases:**

B02512 : Unter B02598 : Hinter B02617 : Weiter B02682 : Schmecken B02764 : Danach-NY B02765 : Grun B02835 : Dreist B02836 : Drinnen.

For coarse-grained location information from these pickups, the file taxi-zone-lookup.csv shows the taxi Zone (essentially, neighborhood) and Borough for each locationID.

**Aggregate Statistics:**

The file Uber-Jan-Feb-FOIL.csv contains aggregated daily Uber trip statistics in January and February 2015.

**THEORITICAL ANALYSIS**

**Setting up the development environment:**

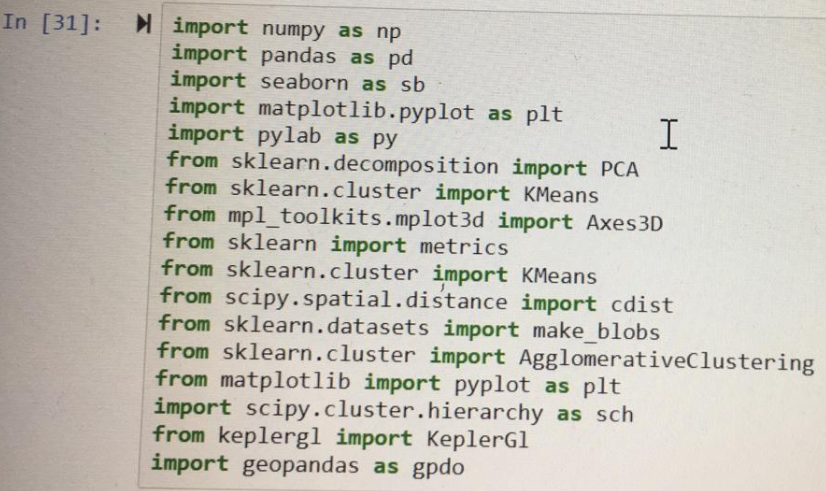
1. Install Anaconda from the link <https://www.anaconda.com/products/individual>.

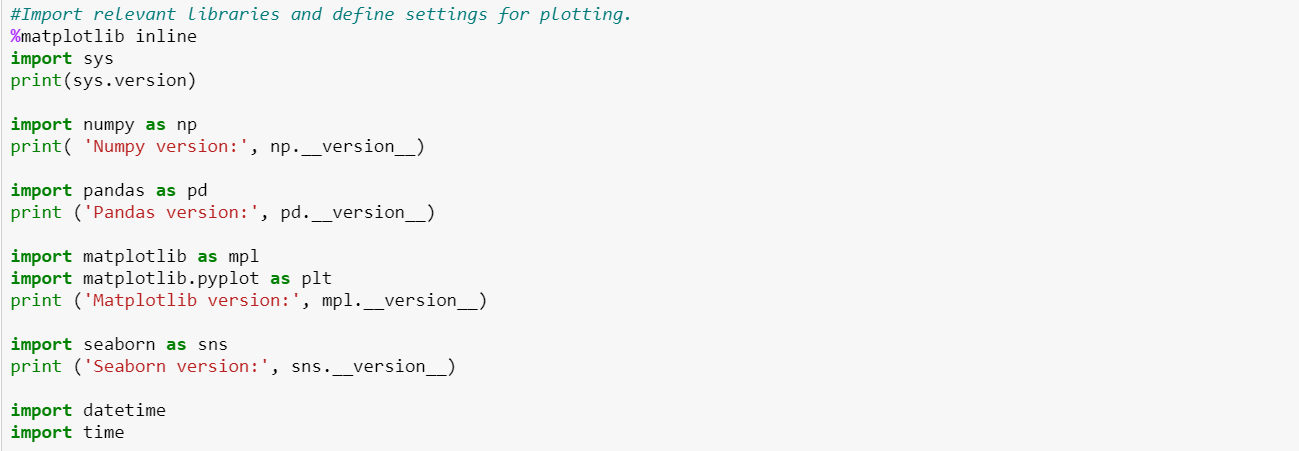
2. Install Jupyter notebook for Anaconda.

**Importing the necessary packages:**

In order to visualize the uber data, we have used python libraries. Below are the few packages and libraries that need to be installed and imported to run our project.

Pandas, numpy, seaborn, matplotlib, keplergl, pylab, sklearn, mpl\_toolkits, scipy, geopandas.



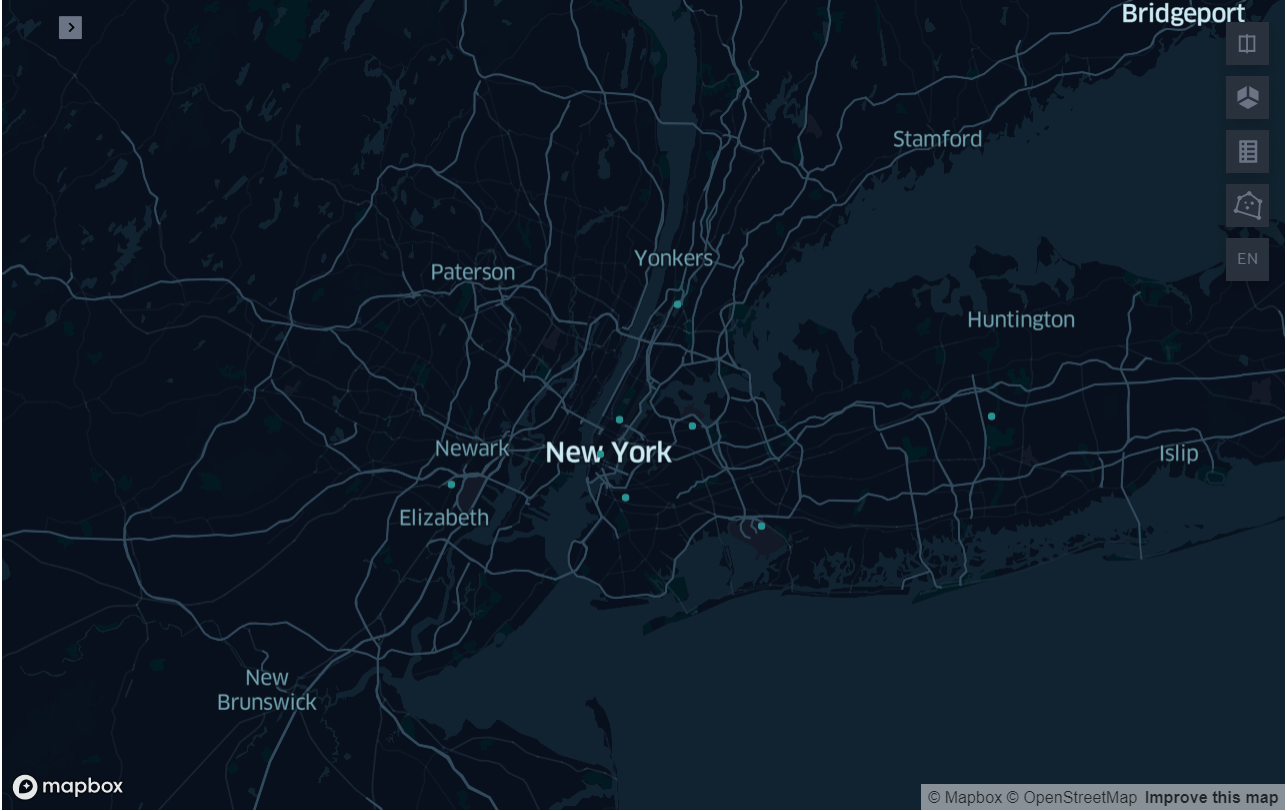


We have used Kmeans clustering and divided the huge data into different clusters. We then found the centroids of these clusters and marked these latitudes and longitudes on the New York city map using keplergl. By this we have predicted the zones with the most number of pickups. Uber can place more number of cabs in these zones to gain more profits.

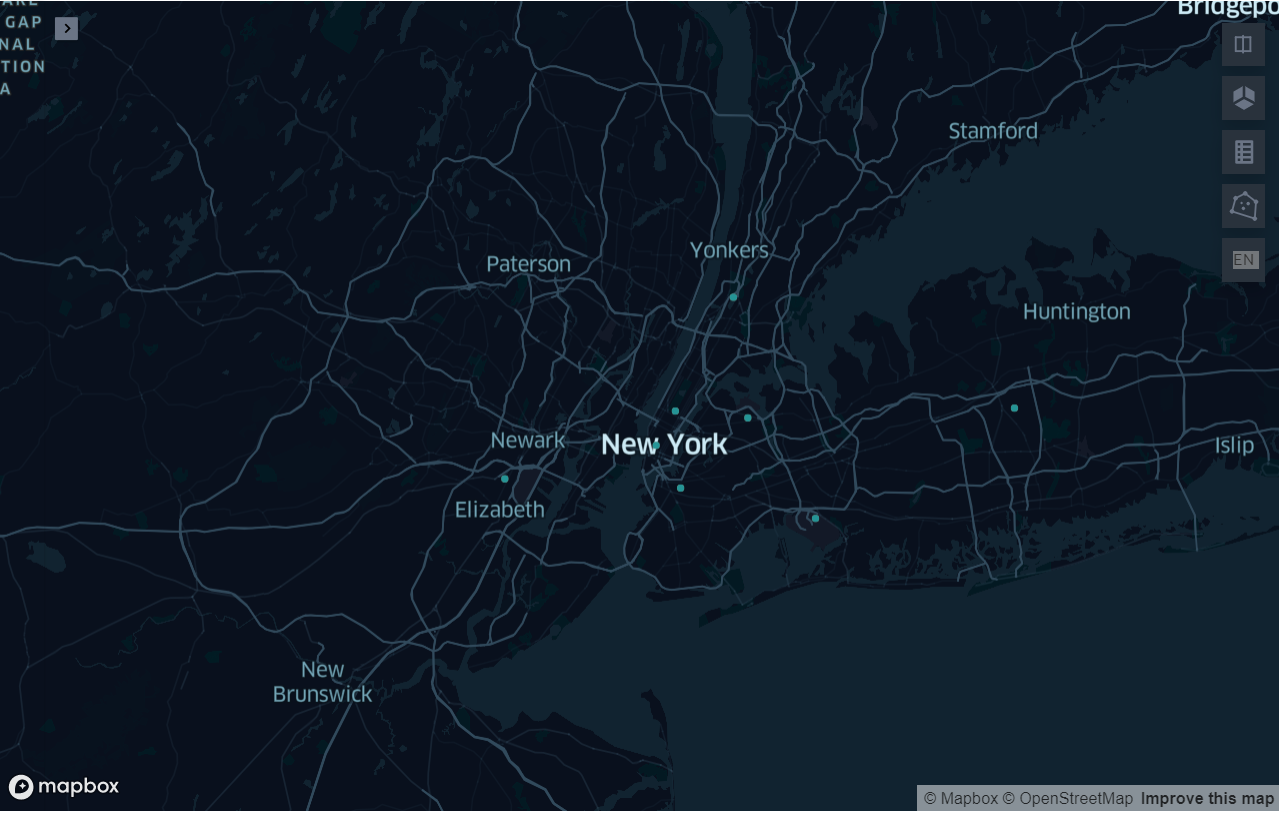
**EXPERIMENTAL ANALYSIS**

The centroid plots for different months can be seen below:

1. May



1. September



1. June

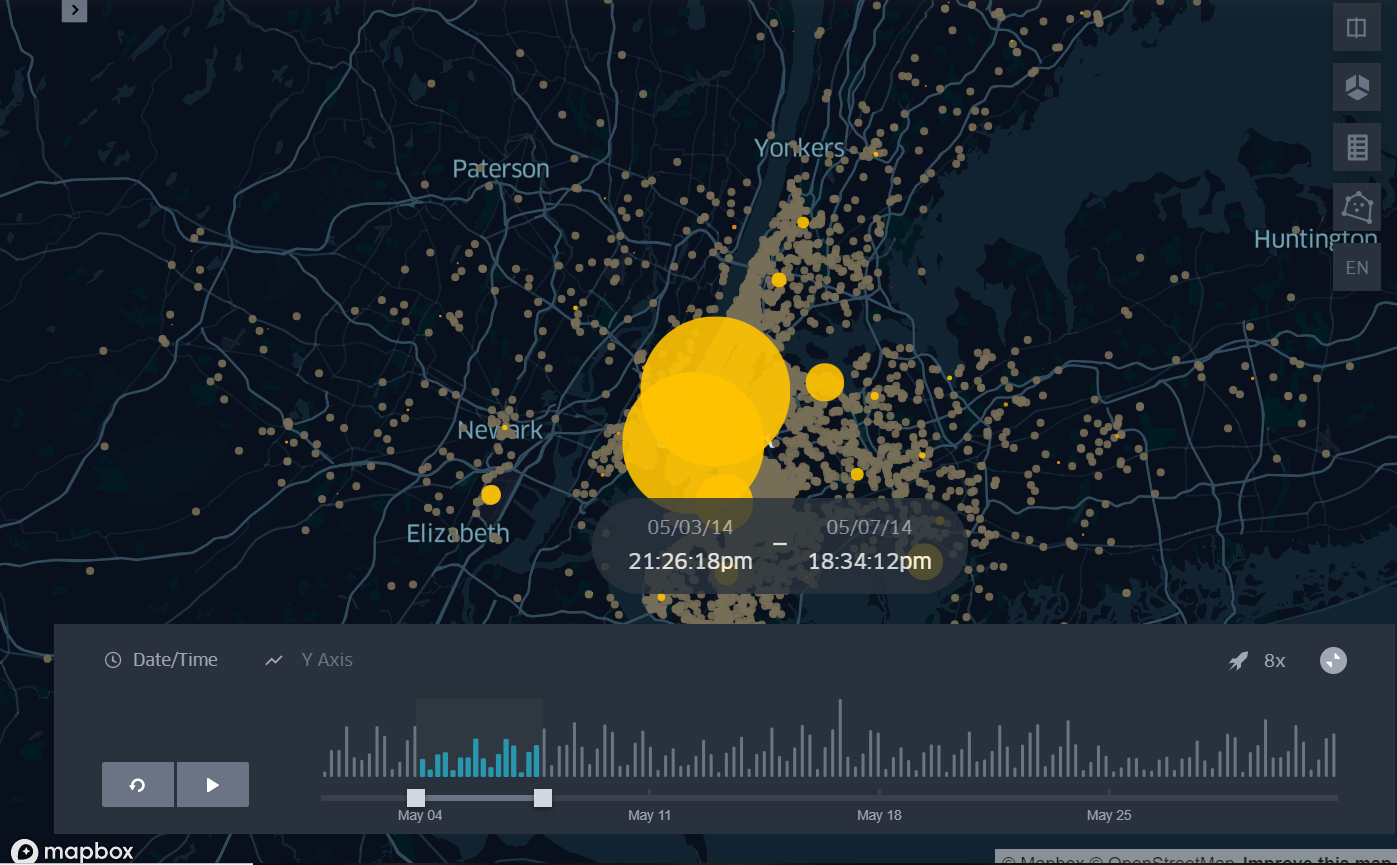


1. April

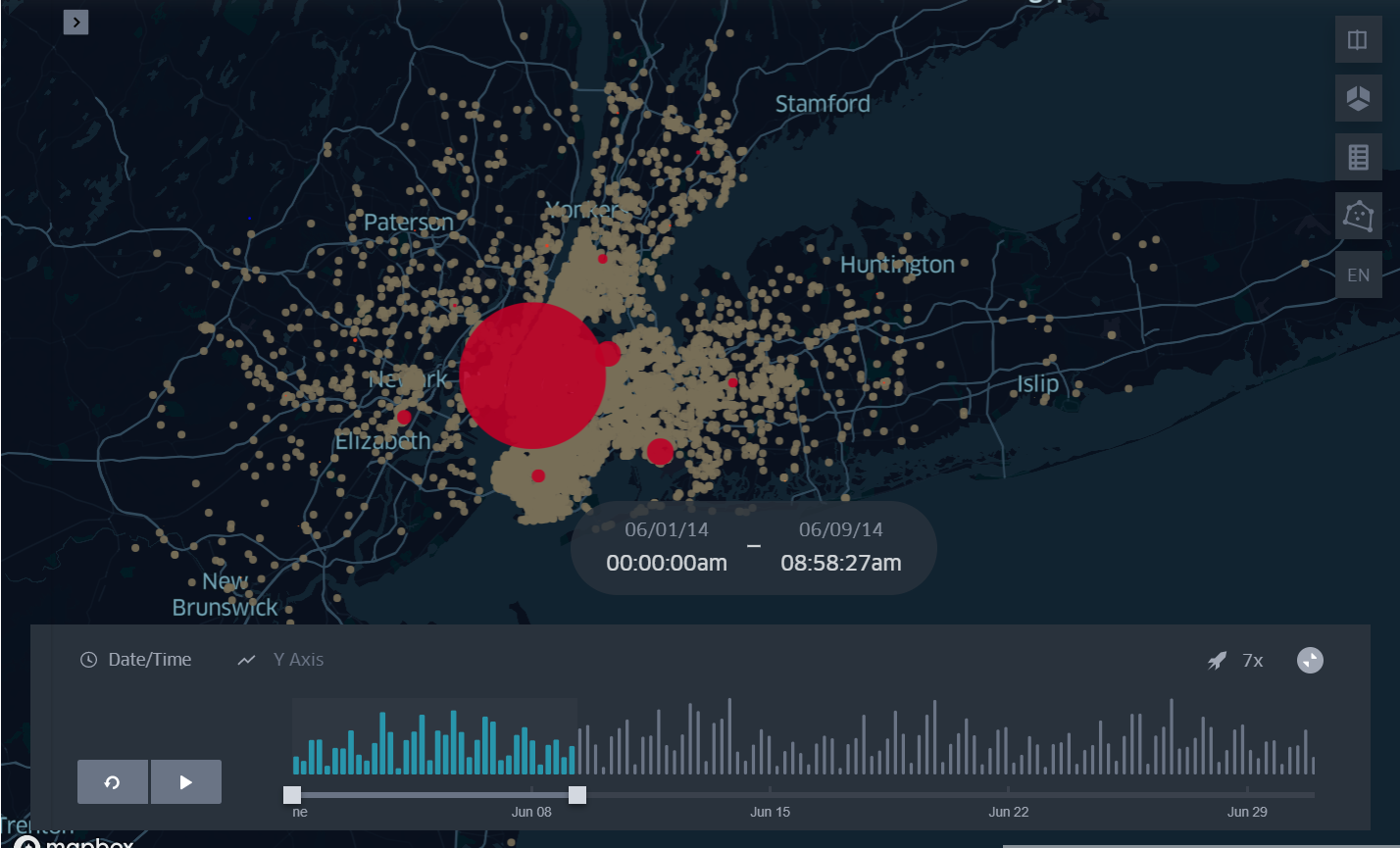


We have also plotted all the pickups in a particular month on the New York map. We added a slider widget using which the user can check the day to day pickup locations on the map. Below are the screenshots of the same.

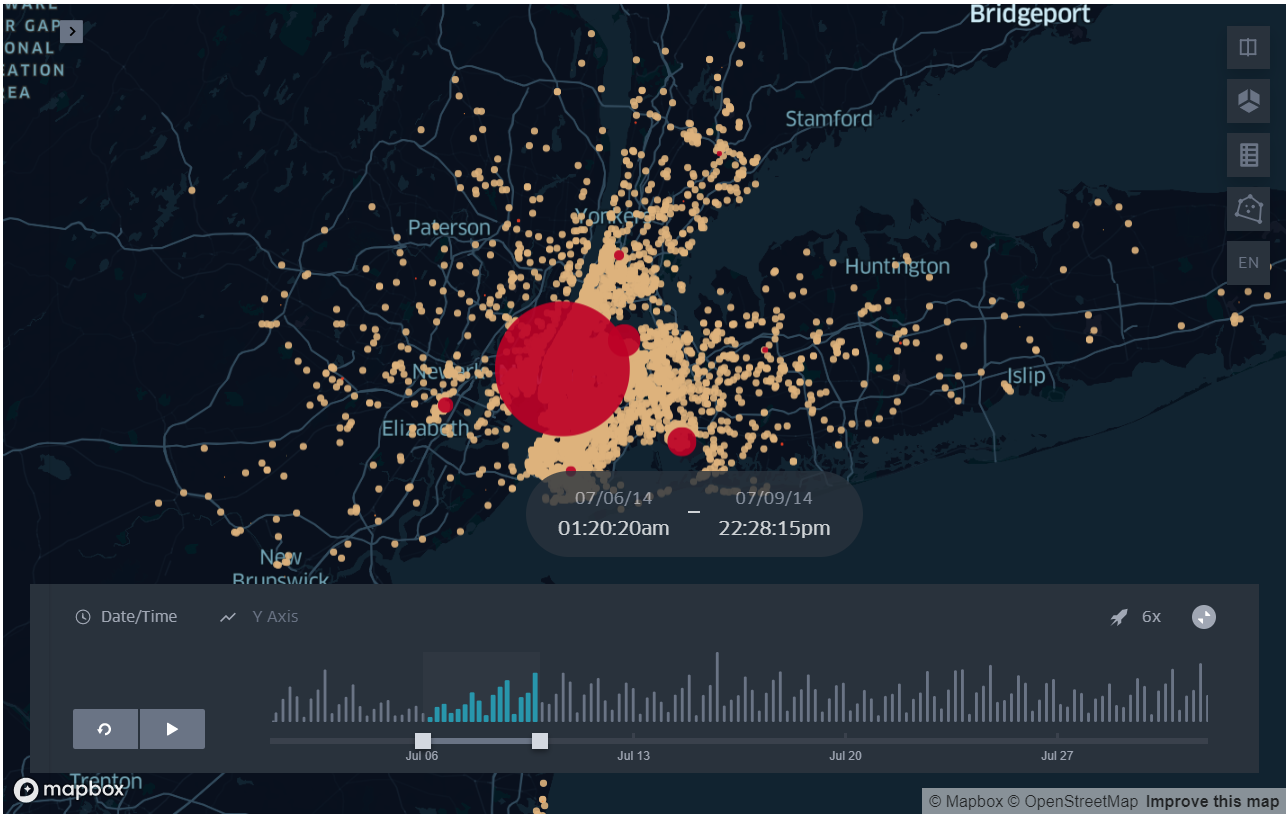
1. May



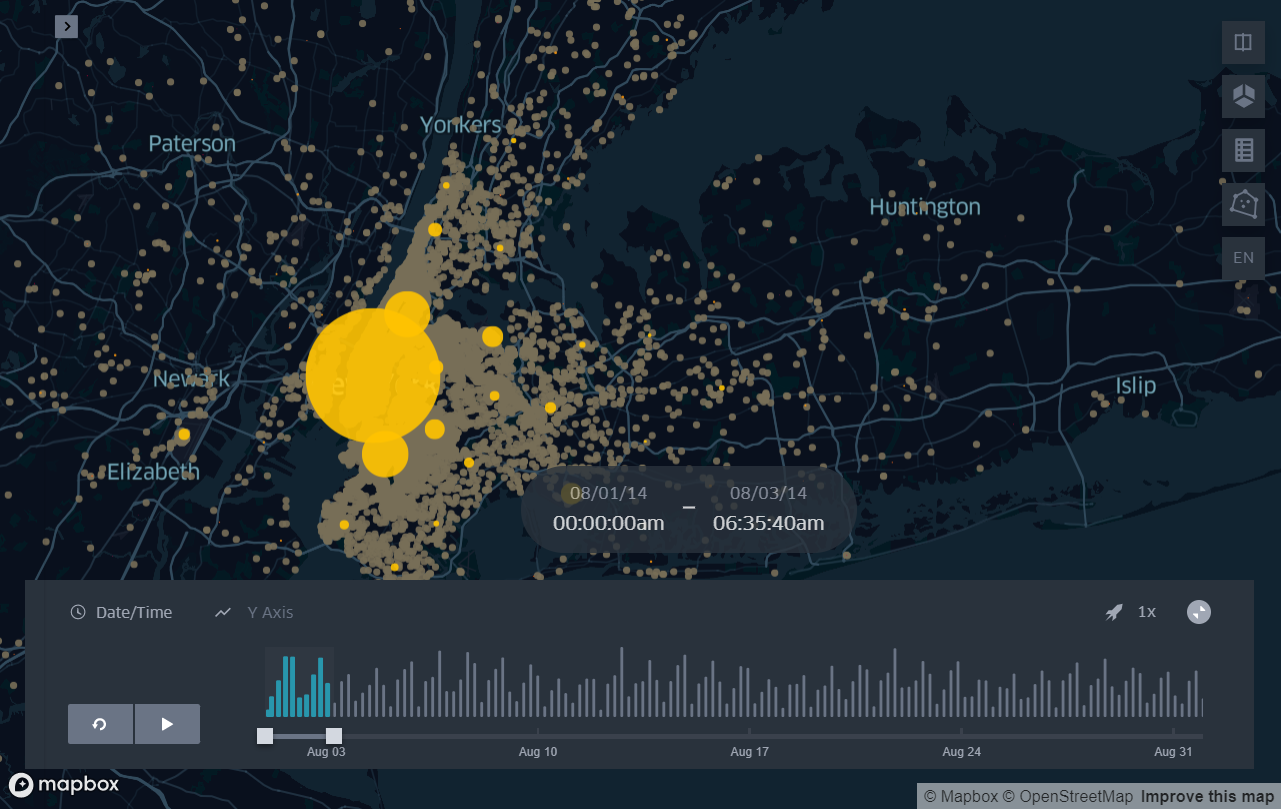
1. June



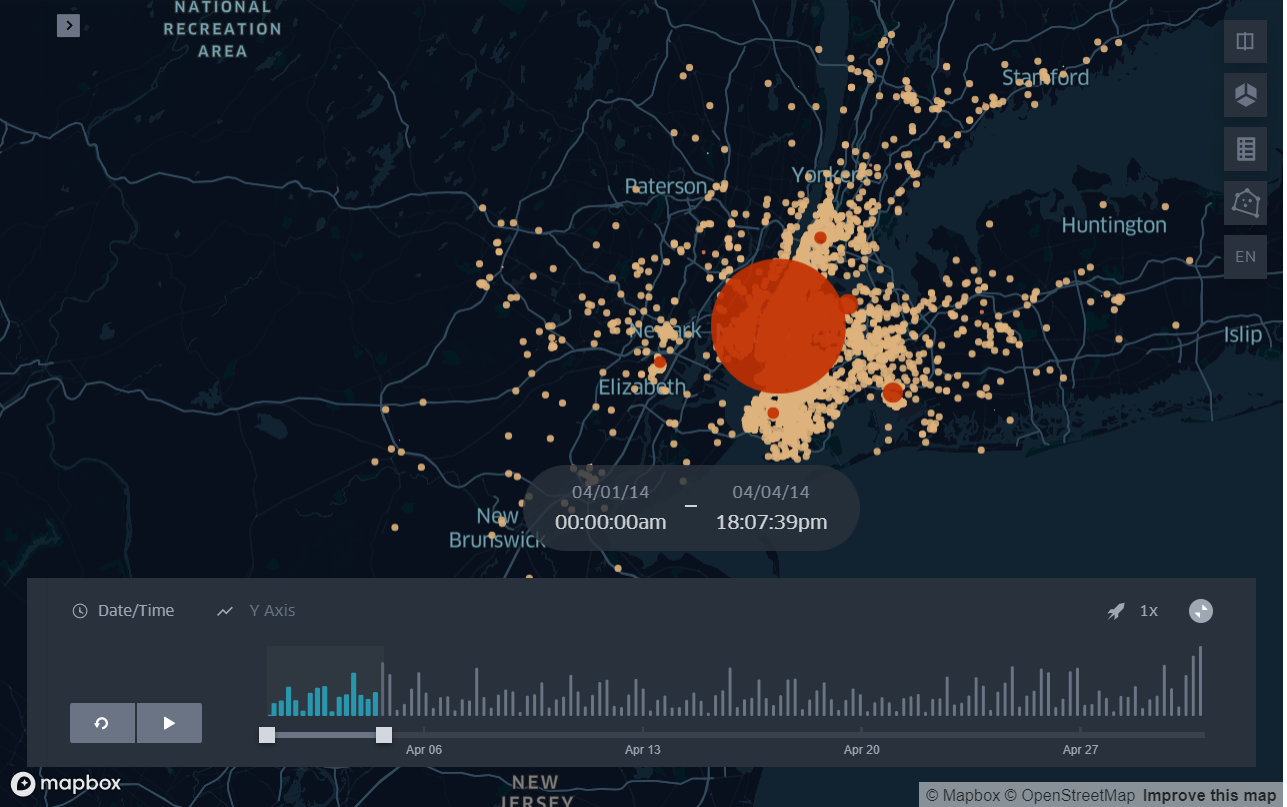
1. July



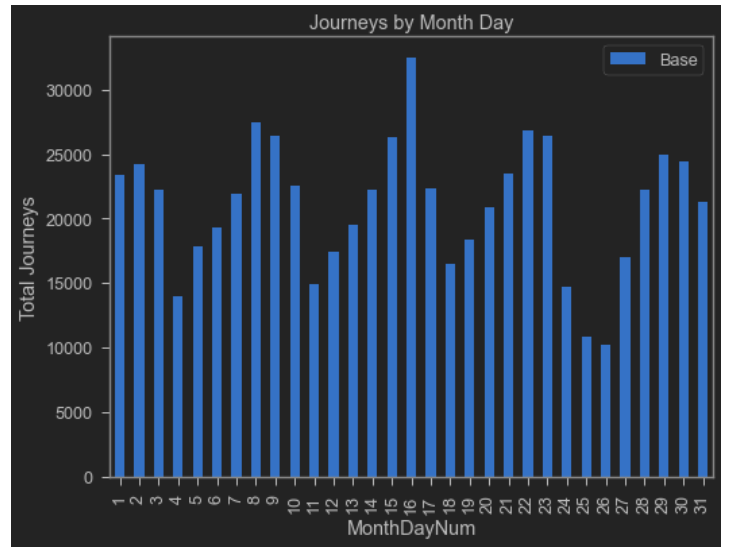
1. August



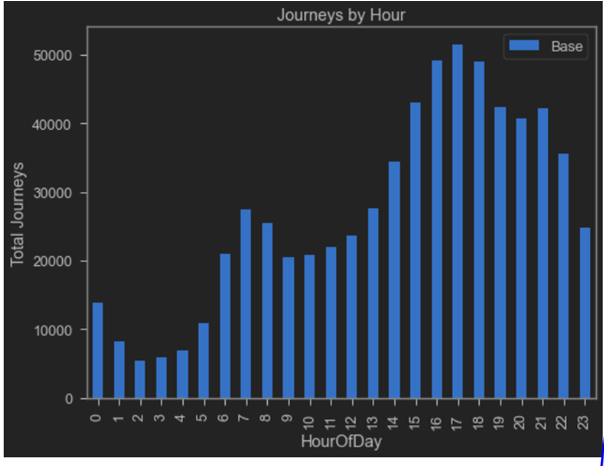
1. April



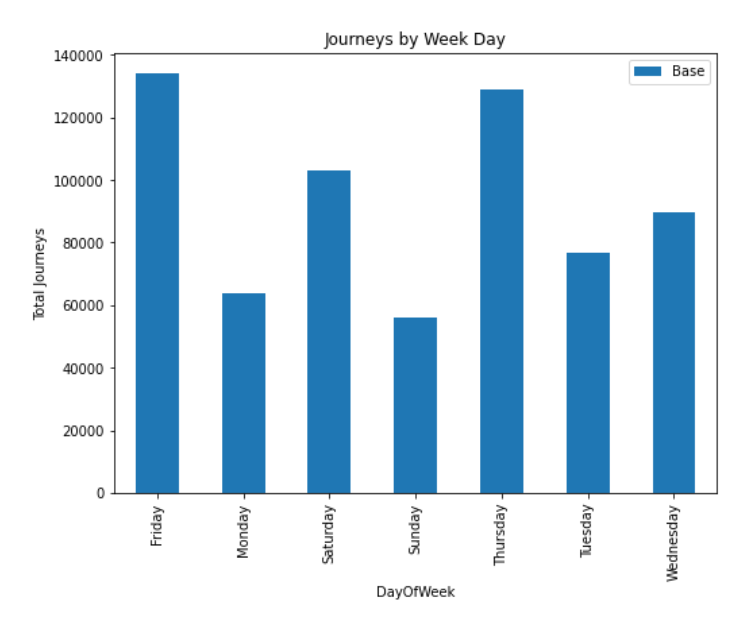
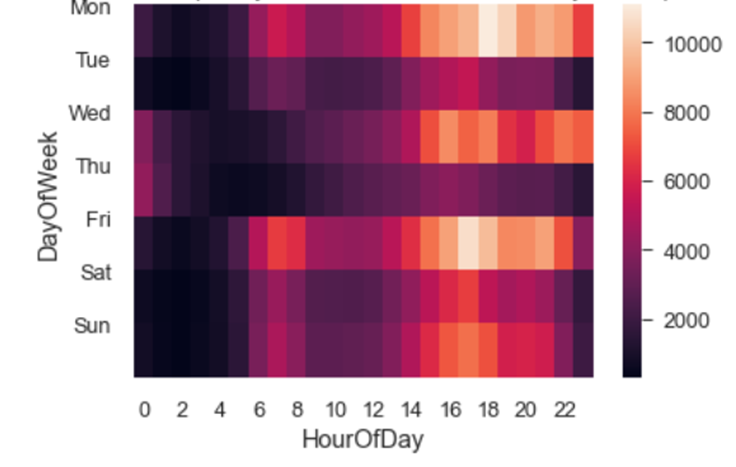
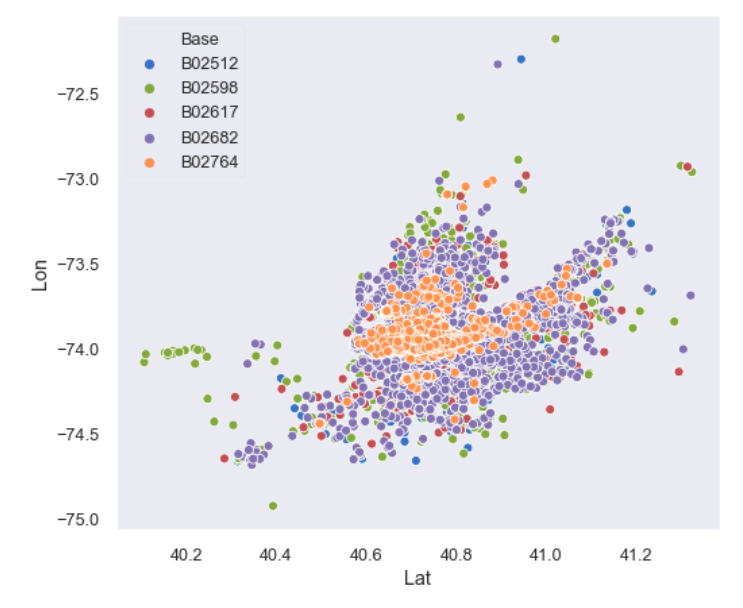
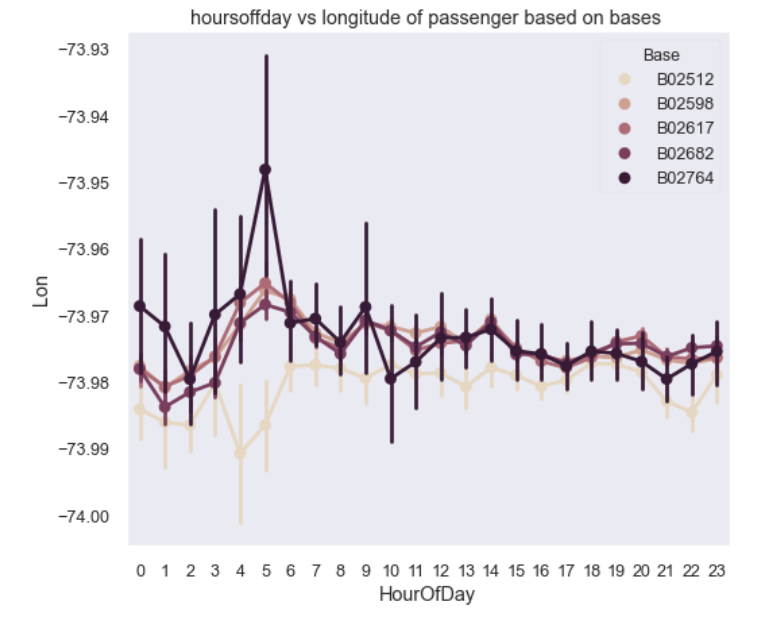
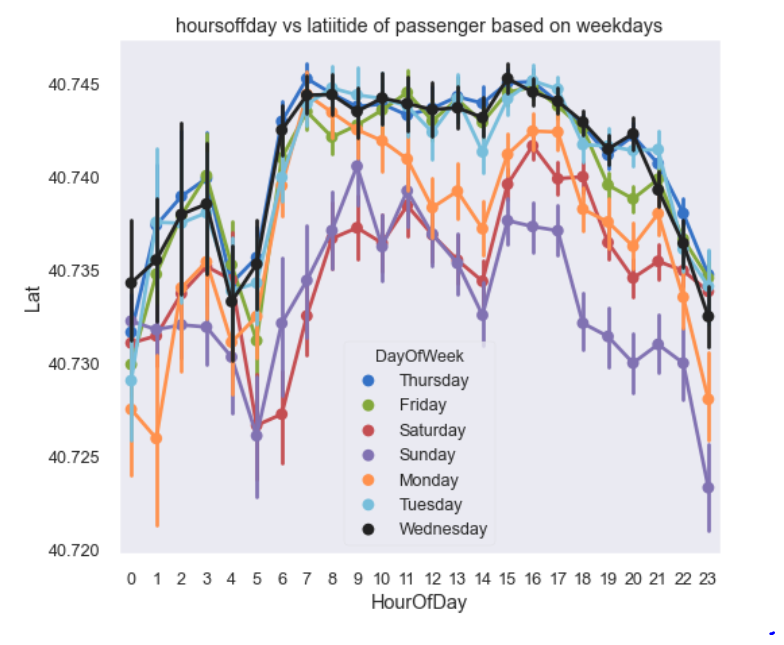
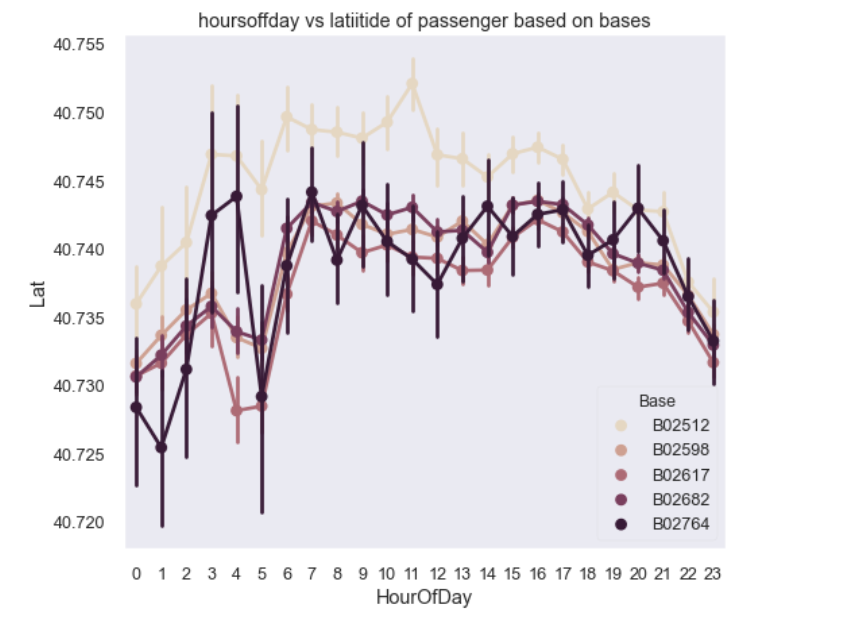
We have plotted many graphs for visualization. For example, below are the graphs plotted for the month of May 2014:



From the above graph, it can be observed that 16th of May had the highest number of pickups. It can be seen that the number of pickups kept increasing during the weekends.



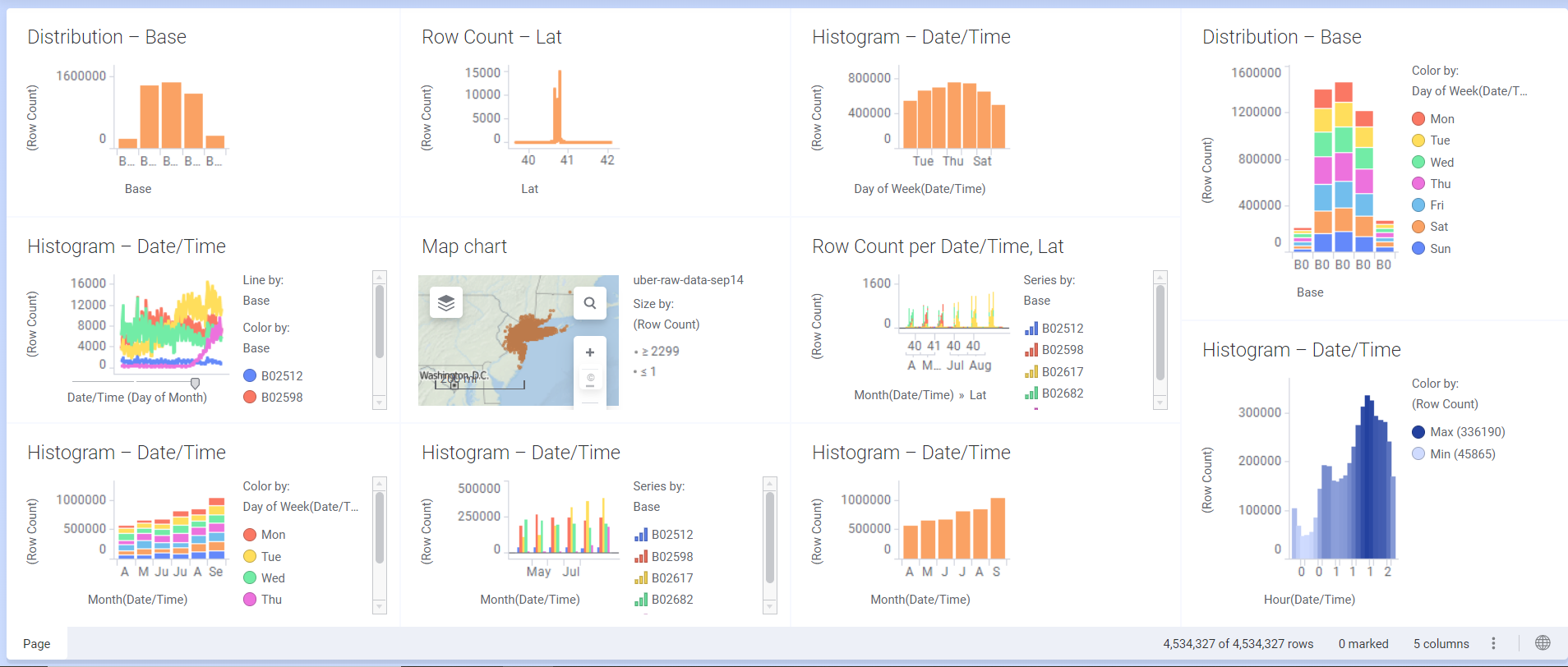
From this graph, it can be observed that the highest number of pickups is between 5 pm – 7 pm. This is the closing time for the schools, offices etc., so there might be increase in the traffic.



Different seasonalities and trends can be defined from these graphs.

Similarly, we have plotted these graphs and visualized the data for all the months. This is for the year of 2014. To make it a little more interactive and easy to understand, we have done a dashboard using Tibco software.

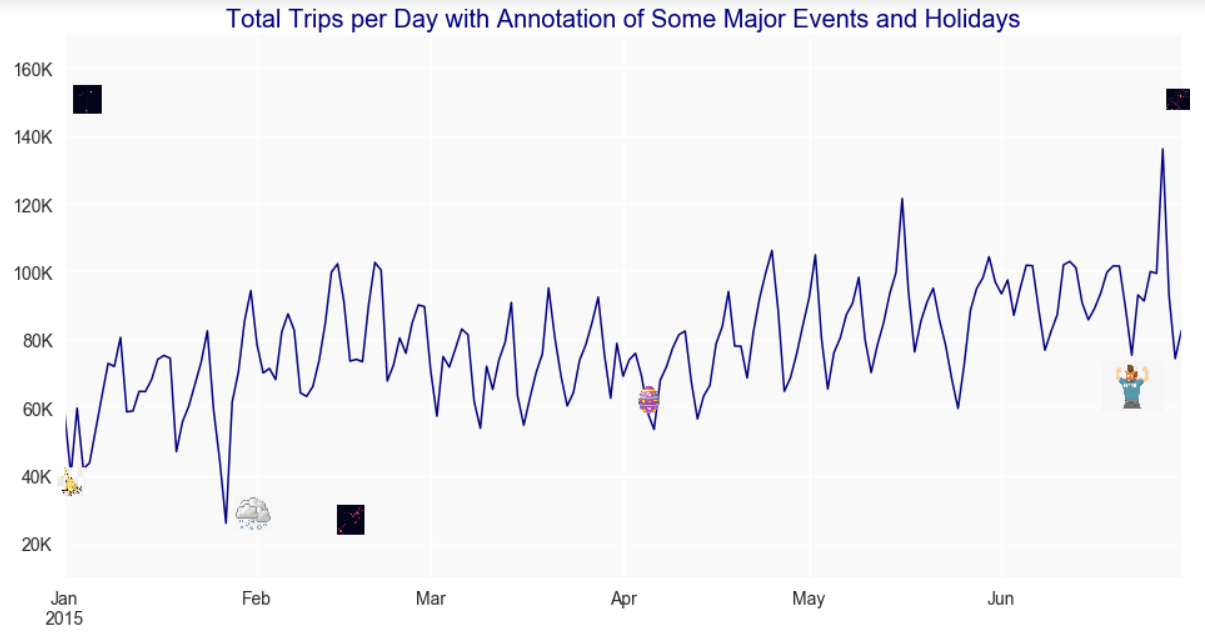
The dashboard can be seen below:



This dashboard has maps and graphs plotted for the combined data of the year 2014.

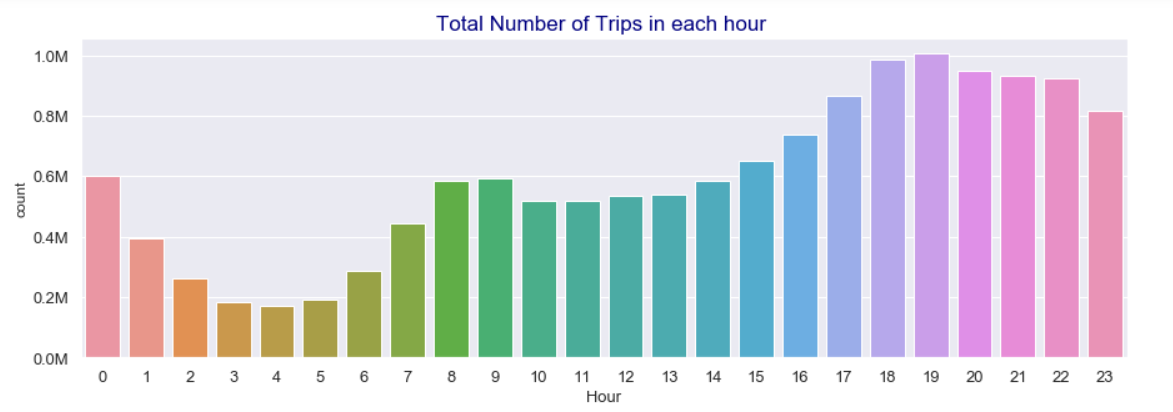
For the year 2015, we have done something different. The dataset of 2015 contains the combined data from the month of January 2015 to June 2015. We have plotted a wave graph, showing the peak and off peak hours and analysed the reasons for the rise or fall.

The wave plot can be seen below:

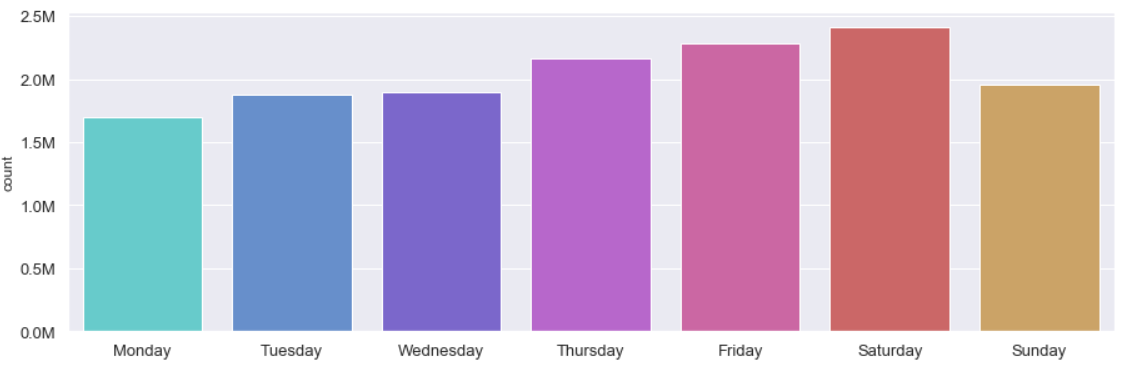


Here, the icons present symbolize the major events. The snowy cloud symbolizes the snow season. Because of the snow, the roads might not be suitable for travelling and so we can see that there is a significant drop in the number of trips. The egg icon symbolizes Easter day. The icon of a happy man(in June) symbolizes father’s day.

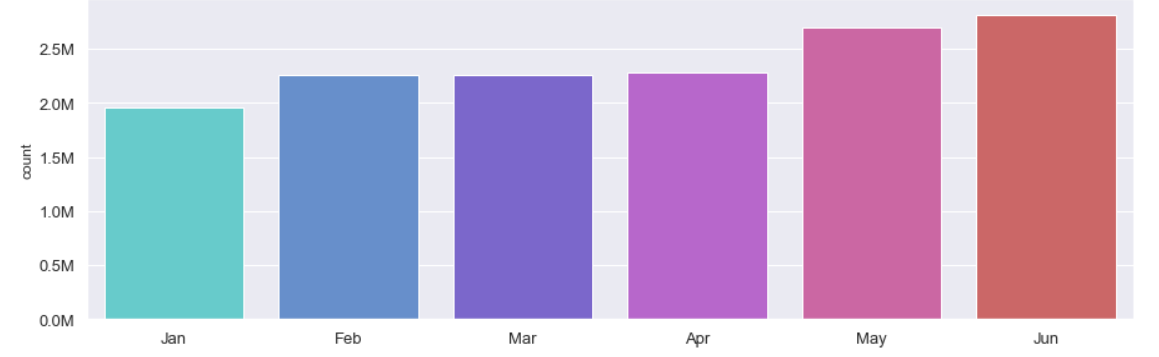
Next we have plotted bar graphs for the total number of trips during each day, hour, month, week, and in each borough just like we have done for the 2014 dataset. These graphs can be seen below:



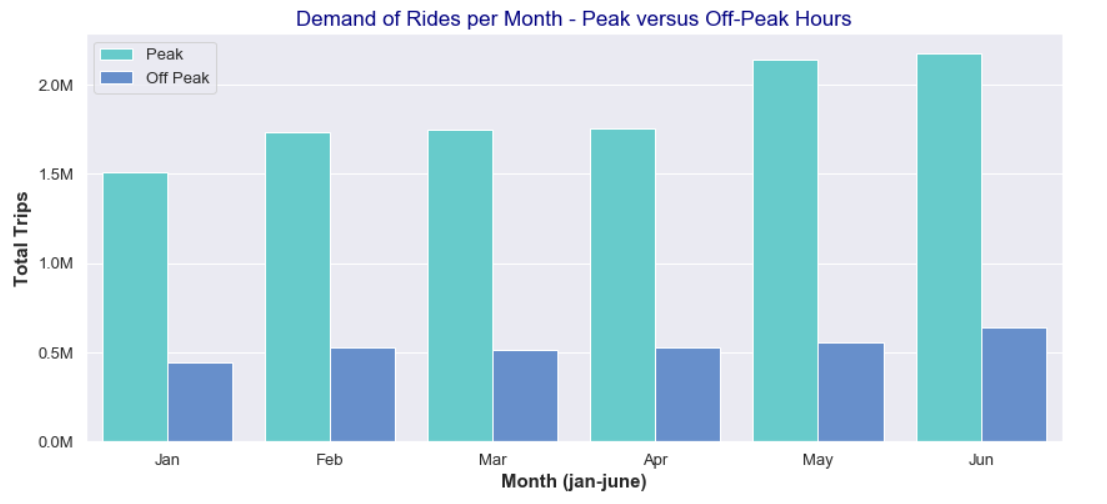
It can be seen that there is high demand in between 6pm and 7pm.



From the above graph, it can be observed that the demand is high during the weekends.



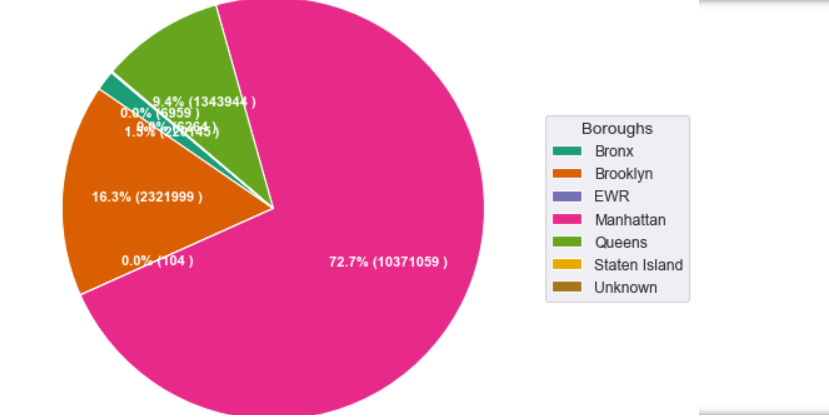
From the above graph, it can be seen that the demand is high during the months May and June. This is the spring season which is pleasant for travelling and so there might be a hike in the demand.



The below graph shows the number of pickups in each borough. It can be seen that Manhattan is the city with the most number of pickups.



The same thing can be seen in the form of a pie chart below:



**APPLICATIONS:**

Using our visualized data, Uber can increase their profits and grow their business. They can know when, where and why there is a high demand or a low demand and take measures accordingly.

**CONCLUSION**

In this project, we have learnt how to use different python libraries and visualize the data. We have learnt how machine learning helps in growing a business.

Git link to our project:

<https://github.com/Divija819/Uber-Data-Analysis>