New York City Taxi Data Analysis

DSC 530- Data Visualization

Abhishek Manoj Kumar

ID: 01675536

Contents:

New York City taxi dataset

I obtained the datasets for different taxi services in New York from NYC for Different taxi services. The cab services have 2 categories, Yellow cabs and Green cabs.

URL:

cab:http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Semantics:

Date and time: (Quantitative) States the date and time at which the cab was booked.
Latitude and Longitude: (Quantitative) specifies the exact location of pickup and drop-off.
Passenger: (Quantitative) specifies the number of passengers.
Trip distance: (Quantitative) distance travelled by the passengers.
Fare amount: (Quantitative) Amount charged for the particular trip.
Tip amount: (Quantitative) Tip given by the customer.

Dataset processing:

I used R to process the data and the R code I used is available in the fallowing link:

https://github.com/abhi080194/Data-Visualization-Project

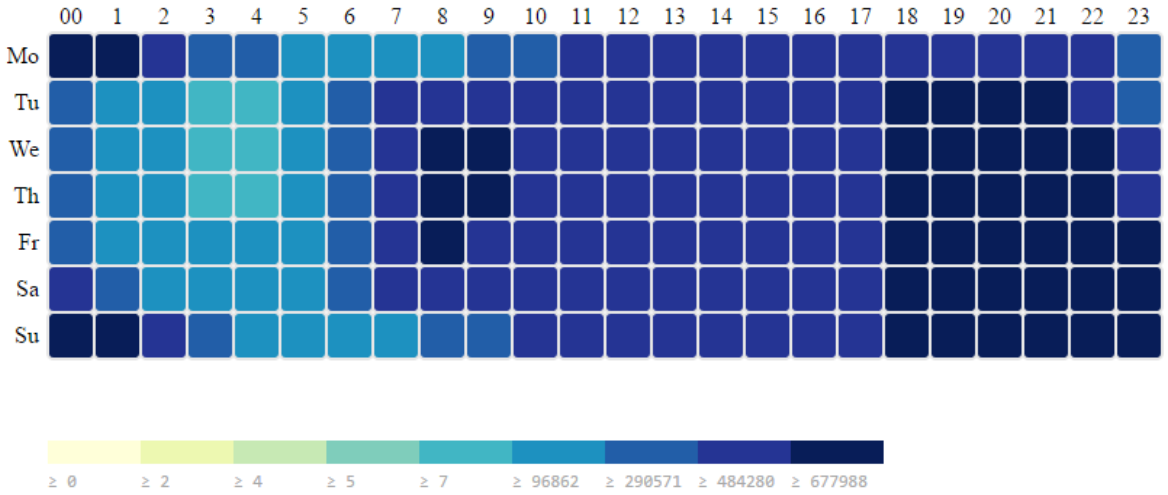
Questions I have analysed through Visualization:

1. How many cabs are booked based on hours and days of the week i.e. are there certain times with high cab usage then others
2. What cab services are most used based on part of the city i.e. is there a relation between the cab service and parts of New York.
3. Does the distance effect people's choice of cab service i.e. certain cab services might be cheaper for longer distances and certain for shorter distances which can have an impact on the cab service chosen.
4. What are the trends observed in tips given to cab drivers i.e. are certain times and days of the week observe better tips than the rest of the week.
5. Does time/day have an impact on passenger count? I.e. when do people travel alone and in groups?

Visualizations:

Question 1:

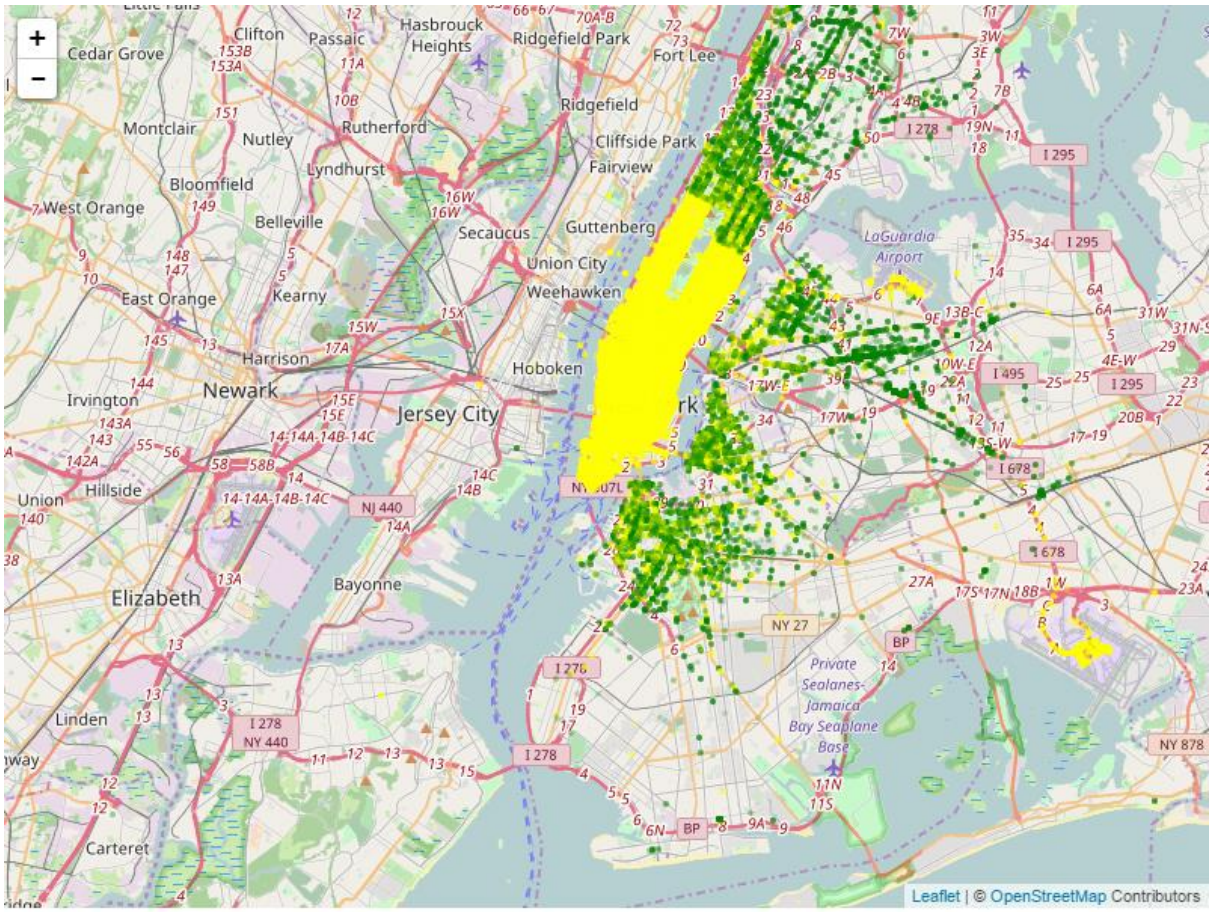
To answer question 1 of finding the frequency of the cabs based on specific days and hours, I made use of heatmaps. Here I made use of all 6 months of data since I had to aggregate the counts and I was able to upload the dataset in GitHub. The trends are clearly seen. There are more cabs during evening and around midnight and morning as well. Reason for this can be that these are usual office hours and cabs are used for transportation. I made use of Sequential colour to show this relation. Heatmap was a suitable for showing the frequency of cabs since it is very convenient to show data for 7 days and 24 hours of the week. I have also made a interaction using this heatmap with New York city map where when you click on a particular square of a heatmap, Cab pickup points are show on the map. In this interaction only columns are relevant which shows the hours. There is no interaction based on days.



Question 2:

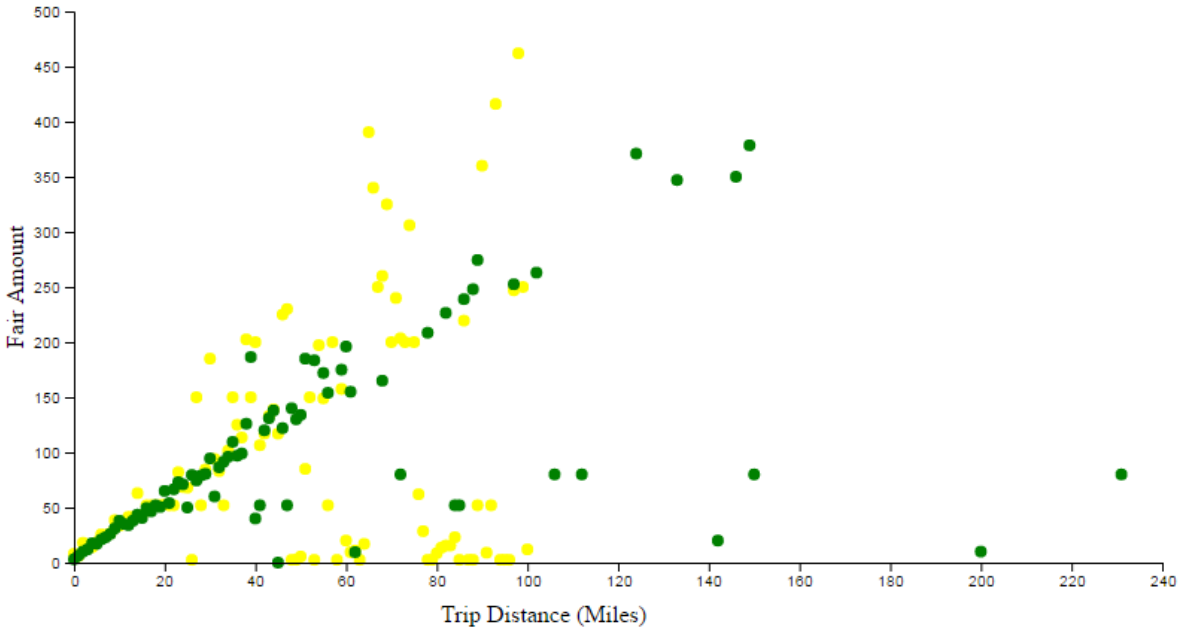
Along with the interaction which shows the density of cabs per hour, I have used maps to answer question 2 as well, which shows if there is a relationship between parts of New York and the cab service used. The visualization shows that there is a relation. Most of yellow cabs are around the Manhattan area and around the airports and the green cabs are spread out around other parts of the city. There is an interesting trend that yellow cabs are used around the airports as well because people will enter Manhattan from the airport. I have used a leaf let map and I differentiate the cabs using categorical colours and the colour choice was natural in this case since the cab names are colours. I had previously planned to analyse 6 months of data, but later realized that it's too much which will require using some

data handling technology in the backend and also map will get too cluttered. So I have randomly sampled 25 % of data for One day for yellow and green cabs.



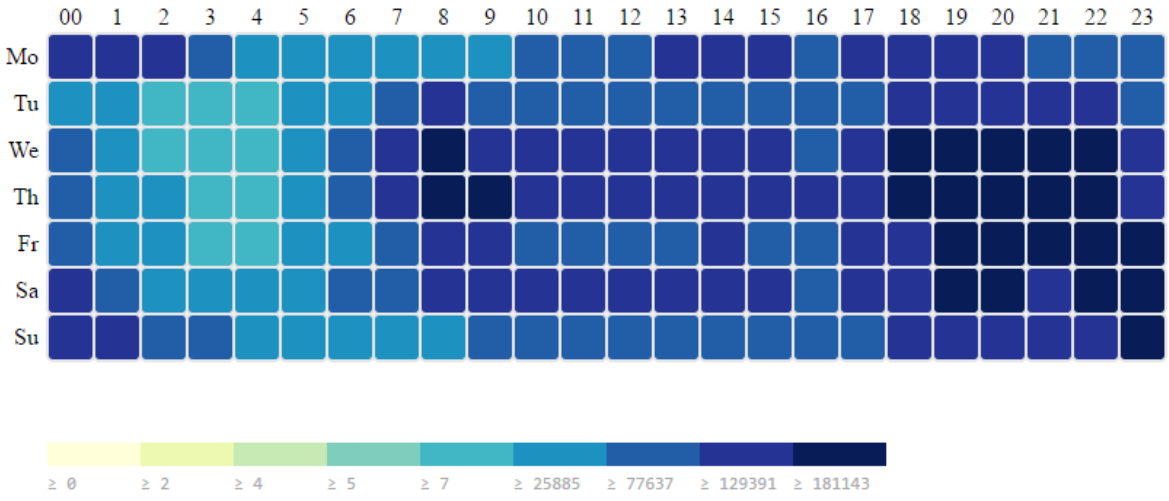
Question 3:

In this visualization, I answer the 3rd question where I see if there is a co-relation between the distance travelled by the cabs and the fair amount. Scatter plot is used to show this co-relation and same colours are used. From this plot, it can be observed that that green is cheaper for long distances. Also yellow cabs don't seem to travel more than 100 miles on a single trip. I have taken a sample of unique fair amount and distance to avoid cluttering. There are some odd points here appearing in the plot and this might be just null values in the dataset or there might be reasons like cancellation of a ride which can cost a certain amount as well.



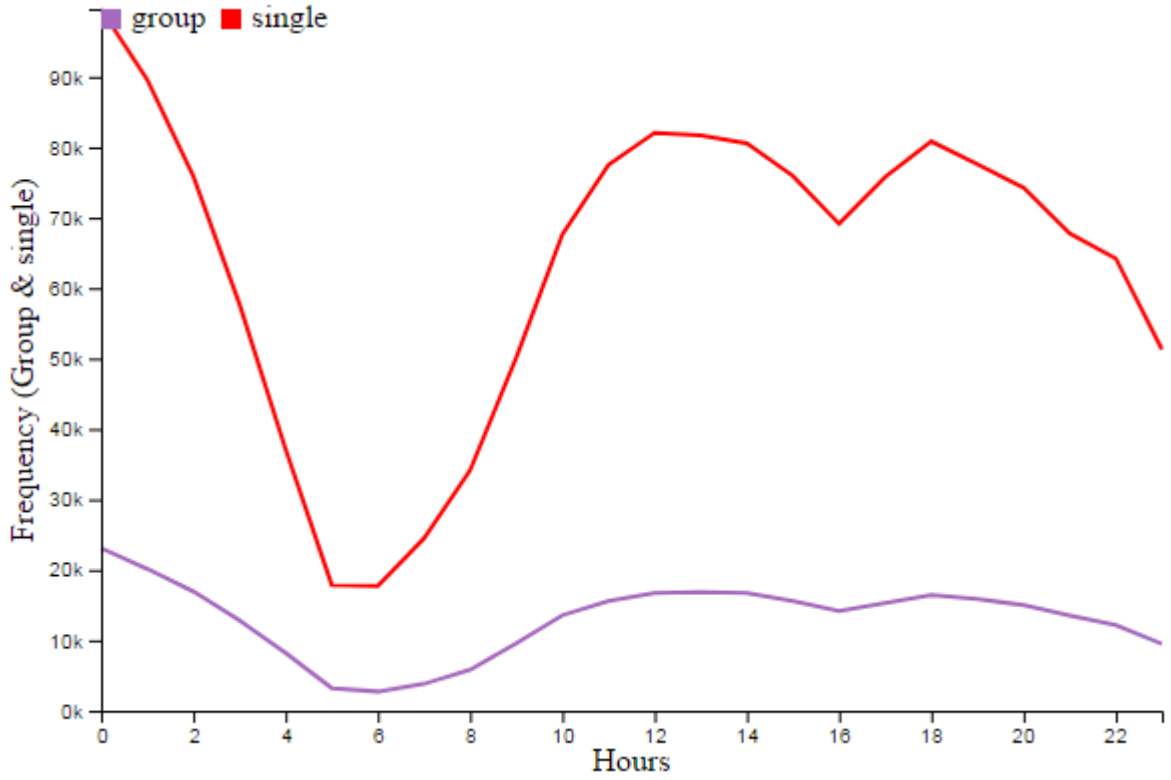
Question 4:

I made use of heatmaps again to show the trend of tips given based on hours and days of the week. Here too I could use data of 6 months since I could aggregate the tip amounts based on hours and days of the week. Some interesting trends are observed like week days have more tips than weekends, since most working professionals use cabs on weekdays. I also show an interaction with this heatmap and line chart showing passengers traveling in groups and in individual. Here the interaction is based in rows i.e. the days and not the hours.



Question 5:

To answer the 5th question of whether there is a relationship between people travelling in groups and people traveling alone based on days and hours of the week, I make use of a line chart. The categorical colours of the lines are randomly chosen. The reason of analysing this relation has some possible applications like, the cabs companies can promote pool services during times when more people travel alone and suggest SUV type vehicles when they travel alone.



References:

1. http://bl.ocks.org/tjdecke/5558084 (Heatmap)
2. http://bl.ocks.org/d3noob/9267535(Leaf let map)
3. http://bl.ocks.org/weiglemc/6185069 (Scatter plot)
4. https://bl.ocks.org/d3noob/4db972df5d7efc7d611255d1cc6f3c4f (Line chart)