

UBER DATA VISUALIZATION

CSCI 6406 VISUALIZATION
PROJECT REPORT

Naina Nijher
nn696043@dal.ca
B00812651



ABSTRACT


A city can be said to be as good at the transportation system of the city, that is the connectivity across the various parts of the city. Taxis and cabs are hence, of utmost importance. Because of they offer high flexibility and availability, taxi services are very popular along all age groups. Taxi service is an ever-increasing industry, that has generated huge volumes of data, thus, giving the opportunity to analyze the data and draw conclusions. Based on the Uber Dataset in New York city, in this paper, I have put together a few visualizations in a dashboard that would be helpful from both the business point of view as well as the consumer point of view. First, from the records a comparative graph has been plotted for the number of uber trips completed against the days of the month and number of uber trips completed against days of the week. Pattern analysis can be done from these graphs to conclude which days of the week and the month would be busier than the others. Second, study has been done to observe a few neighborhoods of New York for the busiest hours along the week. Thus, every locality has its own pulse, that makes the taxi system behave in a predictable fashion.[1] Third, an analysis has been done for a few locations in New York for the increase in the number of pickup requests along a time period of 6 months. This is helpful to make conclusions and predict increase in demand of the taxis in those locations.

INTRODUCTION

Travelling is one of the major components of our lives. We travel every day to and from school/ work/ shopping centres. Taxis are an integral part of our daily commutations. Because of their high reachability and availability round-the-clock, cabs are a popular choice amongst the people. In the last few years, a boost has been observed in these taxi services, as a result of increase outreach of mobile applications. The increase can be observed both, in terms of the number of taxis on the road as well as the number of trips being served by each taxi. For instance, by 2011, there were about 14,000 taxis in New York City, serving around half a million passengers on an average day. [2]

There is a huge amount of taxi trace data, that has been possible because the GPS technology that is now so commonly available. This data has been made available, so that research and analysis can be performed on this data that might help in the improvement of the business. For instance, observing the pattern, can help in prediction of surges in taxi demand and hence, help to pull in more business. Consumers can plan accordingly, keeping in mind wait time when there is such a surge.

Such studies assist in understanding the cab-system. Transportation patters of a city can be characterized by looking into the supply-demand relationship and the movement of the taxis around the city. Such studies are also helpful to better the urban transportation system. Studying the impact of any urban activities like concerts, holidays, games on the taxi demand can also



provide insight into the operations of the taxis around the city. Thus, observing all these factors can give you an idea regarding the city. [6]

In this paper, I have presented a dashboard consisting of 4 interactive graphs. The first graph shows the number of completed uber trips against a particular day of the month. This is a typical statistical data representation and the trends can be seen varying across the selected month. The second graph shows a waterfall chart that shows the cumulative result for the number of uber trips completed across days of the week for a particular time period. This graph can again be used for trend analysis and prediction to assess the increase in demand of the taxi services over the week. The third graph is a heatmap that represents the trends in the number of pickup requests from a particular locality (substituted from the Base Id in the dataset) against days of the week on an hourly basis. This graph can help provide precise insights on the demand of taxi services at a particular time of the day. Thus, there is a strong periodicity that can be observed from this graph. The periodicity is governed by specificities of a location, the time of the day – that is around beginning and ending of school and work timings. Given a more detailed dataset, a pattern can also be observed in relation with the weather and any cultural/social events like concerts, festivals, games. The fourth graph in the dashboard represents in trends in the increase of number of pickup requests for uber along 6 months for various locations in the city. Some locations see quite a jump in business which may be the result of the opening of a new shopping centre or a similar event. Predictions can be made again by observing the trend and help cater to the increase demand in advance.

TOOLS AND TECHNOLOGY

The following tools and technology have been used to create the dashboard:

- Tools: Microsoft visual code for HTML, CSS, D3.js, Anaconda (Jupyter Notebook) for pre-processing of the data and Microsoft Excel to view the data
- Web Server for Chrome to host all the files and data locally
- Technologies: HTML, CSS, D3.js v3, Python

DATASET

- DATA COLLECTION:

After some exploration, I came across the dataset for New York for uber and a few other cab services on Kaggle. I downloaded the data from this link:

<https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city#uber-raw-data-apr14.csv>

The dataset was very voluminous, so I concluded to work with 4 months of data to draw the visualizations. Each month, on an average had 5.75 lakh trip records. The dataset downloaded was a zip file and comprised of several files with different headers. After

inspecting the dataset, there were 5 files that were mainly required. 1 file per month having the trip records for each month with the column headers as Date and Time, Latitude, Longitude and Base ID. The 5th file that was downloaded had an aggregated data for all months with the column headers as Dispatching base Number, Pickup Date and Time, Affiliated Base Num and Location ID.

- **DATA CLEANING AND PRE-PROCESSING:**

The objectives behind data pre-processing are:

- To handle inconsistent corrupt data for increased accuracy
- To trim the columns that are not required to increase the processing time

The data present in each of the monthly files was not uniform. Thus, to clean and prepare the data for implementation, I used panda dataframe in python.

```
import pandas
import os
from datetime import datetime

#df_april = pandas.read_csv("E:/Visualisation/uber-raw-data-apr14_clean.csv", sep = ",")
#print("April read")
df_april = pandas.read_csv("E:/Visualisation/uber-raw-data-apr14.csv", sep = ",")
#print("May read")
#df.head()
#df[["date"]]
#df.date.dtype
#df_april[["date"]].astype(str)
#df.date.dtype
#df.date.head()
df_april['date'] = df_may['date'].dt.strftime('%x')
#frames = [df_may, df_june]
#result = pandas.concat(frames)
print(df_april)
#result.to_csv(r"E:/Visualisation/uber-raw-data-may14_clean.csv", index = None, header=True)
```

Image 1: Python code snippet to clean data

04-01-2014 00:11	40.7630	-73.9945	B02512	1	April 1st, 2014	40.7630	-73.9945	B02512	1
04-01-2014 00:17	40.7267	-74.0345	B02512	1	April 1st, 2014	40.7267	-74.0345	B02512	1
04-01-2014 00:21	40.7316	-73.9873	B02512	1	April 1st, 2014	40.7316	-73.9873	B02512	1
04-01-2014 00:28	40.7588	-73.9776	B02512	1	April 1st, 2014	40.7588	-73.9776	B02512	1
04-01-2014 00:33	40.7594	-73.9722	B02512	1	April 1st, 2014	40.7594	-73.9722	B02512	1
04-01-2014 00:33	40.7383	-74.0403	B02512	1	April 1st, 2014	40.7383	-74.0403	B02512	1
04-01-2014 00:39	40.7223	-73.9887	B02512	1	April 1st, 2014	40.7223	-73.9887	B02512	1
04-01-2014 00:45	40.7620	-73.9790	B02512	1	April 1st, 2014	40.7620	-73.9790	B02512	1
04-01-2014 00:55	40.7524	-73.9960	B02512	1	April 1st, 2014	40.7524	-73.9960	B02512	1

Image 2: Snippet of the original data

Image 3: Snippet of the cleaned data

There were a few anomalies in the data in the date column, wherein a few entries had the year written as 14 instead of 2014 and the format of writing the date was not consistent (in some records the fields in the date were separated by – and in some with /). This issue was fixed by simply replacing all these errors using the replace function in Excel.

For the second graph, aggregated data according to the day of the week was required. After exploring the data nest() function for the first graph, I conclude that D3.js is not built to handle such a large dataset. Hence, I aggregated the data using python and exported the dataset to a csv file which was then used to build the graph.

```

from datetime import datetime
#df["Day"] = [d.date() for d in df["Pickup_date"]]
#df['Time'] = [d.time() for d in df["Pickup_date"]]
df['date'] = pd.to_datetime(df.date)
df['day_of_week'] = df['date'].dt.day_name()

df2 = df.groupby(["day_of_week"]).size().reset_index(name='count')

df2

df2.to_csv(r"E:/Visualisation/uber_weekdays.csv", index = None, header=True)

```

Image 4: Pre-processing for graph 2

For the third and fourth graph, the location is required. But in the downloaded dataset, only the field Location Id is mentioned. Another table having a mapping between the Locations and Location ID was absent from the dataset. Hence, I have created an arbitrary mapping as I am not being able to find an actual mapping and will be using that for the implementation of the third and fourth graph.

IMPLEMENTATION [3,4,5]

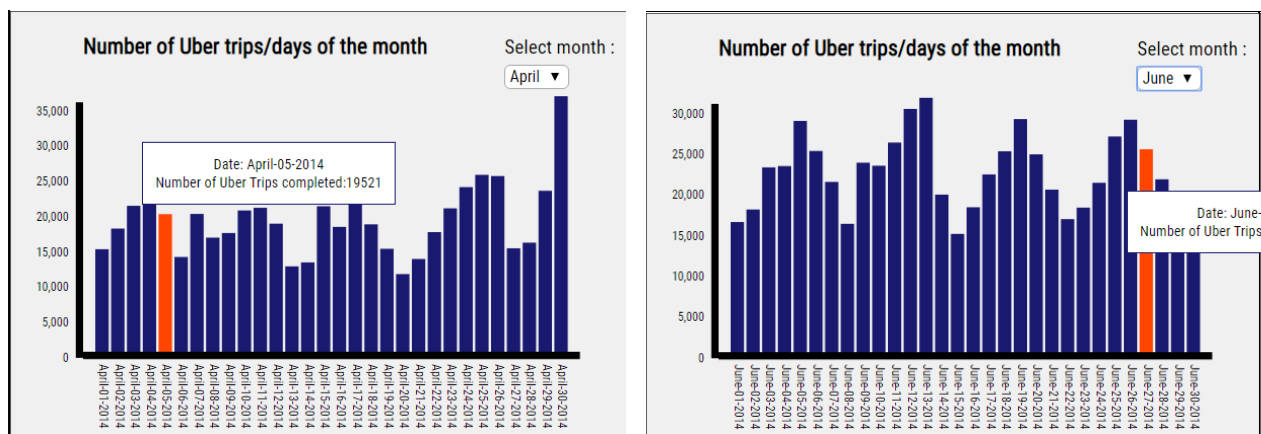


Image 5 & 6: Graph 1 Number of Uber trips/ days of the month

Graph 1 represents a count of the number of trips against the days of the month. The user can view at the data for several months by selecting the month from the dropdown list. The graph has been implemented by using the `d3.nest()` function and data is being filtered according to the month given as the input. The y-axis is being updated as per the maximum values for the count of the trips for that month.

For the second graph, I have implemented the Waterfall model that models the data for each day of the week over a span of 4 months for the total number of completed Uber trips. The graph also shows a total number of the Uber trips over the said time period. Waterfall chart was chosen to implement this graph to show the cumulative results for better analysis.

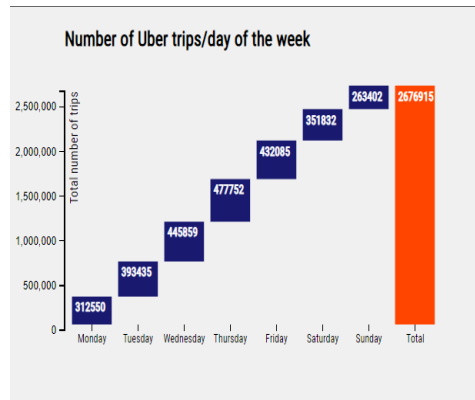


Image 7: Number of trips/ days of the week

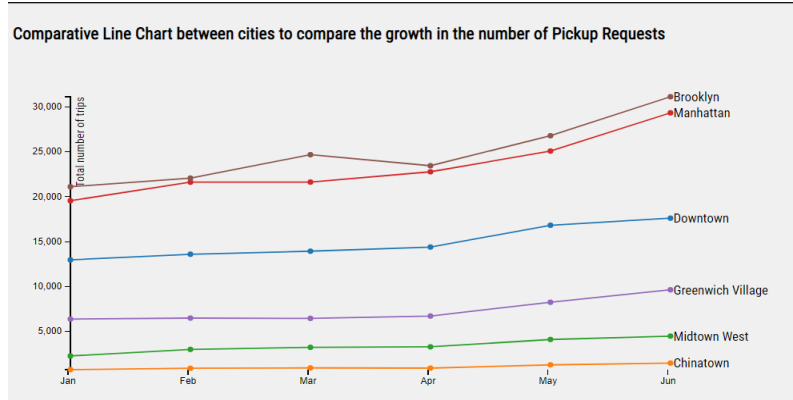


Image 8: Trends in number of pickup requests over 6 months

The fourth graph has been implemented to compare and visualize the trends in increase in the number of pickup requests for a few locations in New York. Some locations, like Chinatown, have barely shown an increase over a span of 6 months, whereas some locations like Brooklyn and Manhattan have shown a drastic increase from January to June.

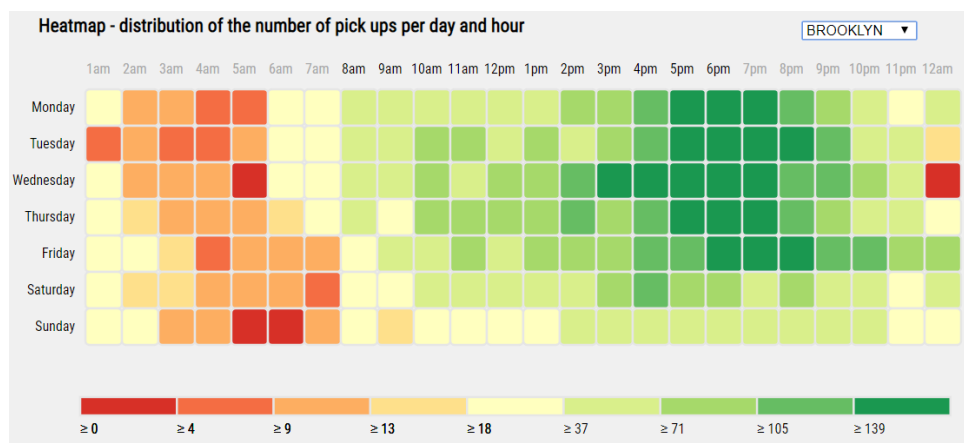


Image 9: Trends of the number of pickup requests per hour per day of the week

The third graph has been carried out with a heatmap that shows the trends of the Uber pickup requests across the day over the week. A few locations show an increase in the number of requests around 9am and 6pm. Thus, pointing to a conclusion that these locations may have a number of offices.[7]

UBER STATISTICS

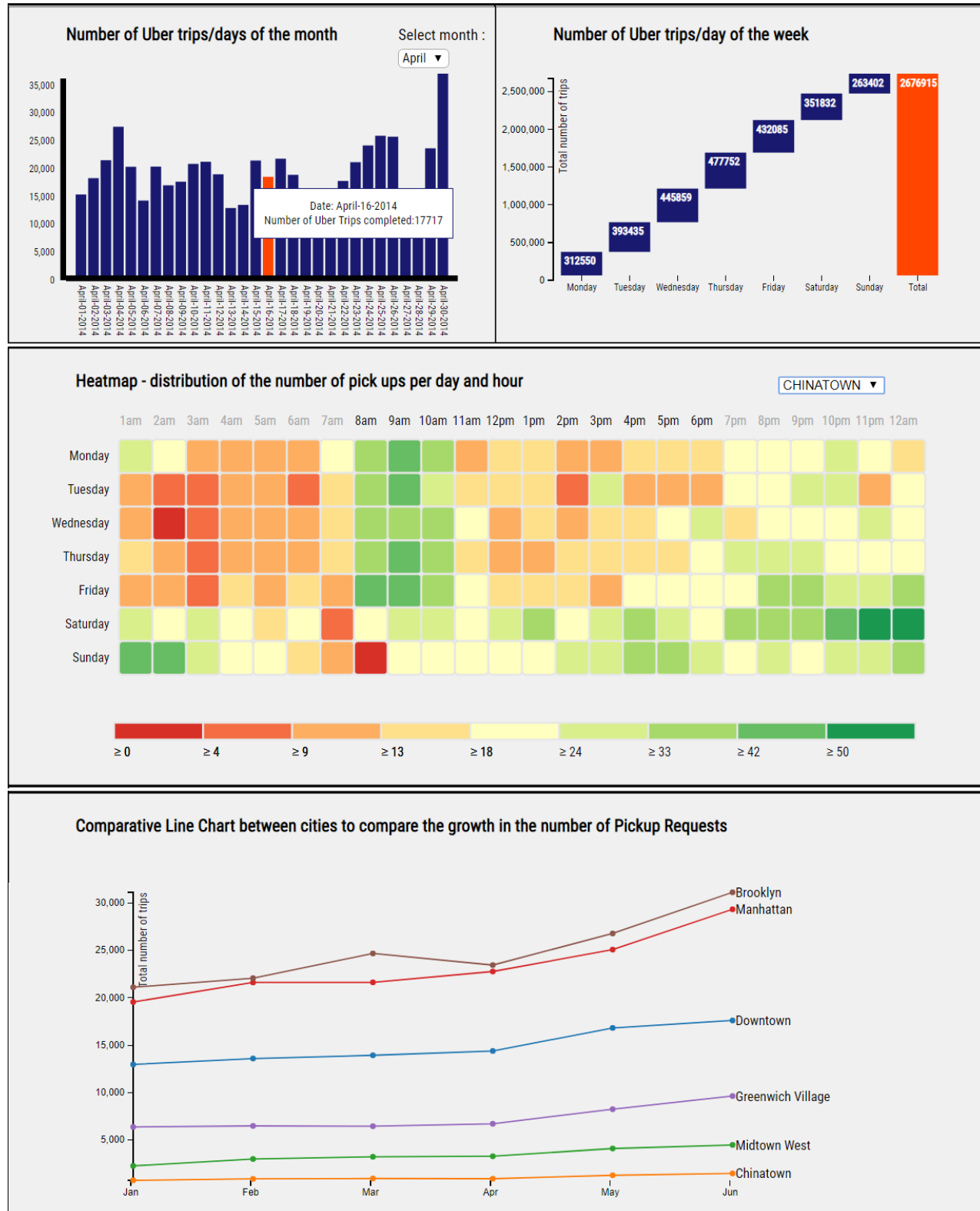


Image 10: Dashbaord



ISSUES FACED

One of the major issues that I faced while implementing this bar chart was that I found no practical way to take a count of all the grouped data. So, as a last resort, I added an extra column to the dataset, that has the value 1 for each of the records and then, just took the sum over that column for all the records that were grouped together.

Another issue that I faced was that there were no mappings provided in the dataset for the Location IDs and the Locations that were being referenced to. As a solution, I created arbitrary mappings between the Location IDs and the Locations.

The third issue, which I am still facing is that the nesting function implemented in the first graph is taking some time to load. This might be due to the large amount of data that is approximately 2.5 million records.

FUTURE WORK

- Research on methods to increase the processing time for the first graph.
- While working on the project, another visualization idea that I came up was mapping the number of pickup requests for the same destination over a particular time span. I am currently researching on ways of implementing this.

SUMMARY


Through this project, a detailed analysis has been done of the Uber taxi service. Several insights can be obtained from the dashboard that can benefit both the business owner as well as consumers. Conclusions can be drawn which can help consumers save time by planning in case of a time that might experience a surge in the number of pickup requests. The business owner can foresee if there would be an increase in the number of pickup requests at a particular time for a location and manage the fleet accordingly.

REFERENCES

[1] B. Prabhakar and C. Zhu, "MEASURING THE PULSE OF A CITY VIA TAXI OPERATION: A CASE STUDY", *Microsoft.com*, 2019. [Online]. Available: <https://www.microsoft.com/en-us/research/wp-content/uploads/2017/11/trb2017.pdf>. [Accessed: Apr- 2019].

[2] Y. Zheng, Y. Liu, J. Yuan, X. Xie. Urban computing with taxicabs. In Proceedings of the 31st international conference on Ubiquitous computing. ACM, 2011.

[3] Kaggle.com. (2019). *UBER STATISTICS / Kaggle*. [online] Available at: <https://www.kaggle.com/dhaval8895/uber-statistics/code> [Accessed Feb. 2019].



[4] Kaggle.com. (2019). *Data Exploration and Visualization / Kaggle*. [online] Available at: <https://www.kaggle.com/dotman/data-exploration-and-visualization> [Accessed Feb.2019].

[5] Kaggle.com. (2019). *Uber Plots, Heatmaps and Tables / Kaggle*. [online] Available at: <https://www.kaggle.com/ikleiman/uber-plots-heatmaps-and-tables> [Accessed Feb.2019].

[6] Qian, Xinwu & Zhan, Xianyuan & V. Ukkusuri, Satish. (2015). Characterizing Urban Dynamics Using Large Scale Taxicab Data. 10.1007/978-3-319-18320-6_2.

[7] C. Zhu. Analysis and modeling of large-scale systems: taxis and social polling. Ph.D. 39 dissertation, Stanford University, 2015.

[8] "Popular Blocks - bl.ocks.org", *Bl.ocks.org*, 2019. [Online]. Available: <http://bl.ocks.org/>. [Accessed: Feb- 2019].

[9] Y. Holtz, "The D3 Graph Gallery - Simple charts made in d3.js", *D3-graph-gallery.com*, 2019. [Online]. Available: <https://www.d3-graph-gallery.com/index.html>. [Accessed: Feb- 2019].

[10] "D3.js Tips and Tricks", *D3noob.org*, 2019. [Online]. Available: <http://www.d3noob.org/>. [Accessed: Feb- 2019].