

MAIB CSC 101

Graphs for Competitive Analysis on New York City's Private For Hire Vehicle Companies

Python Dashboard

Abbiegael Klara Go Chu
4-25-2023

Table of Contents

I.	Problem Statement	4
II.	Dataset	4
III.	Data Cleaning and Feature Engineering	5
IV.	Plotly Dash Graphs	5
A.	Pie Chart.....	5
B.	Box Plot	5
C.	Time Series Chart	6
D.	Scatter Plot	6
E.	Heatmap	6
F.	Choropleth Map	7
V.	References	8
VI.	Appendices.....	9

Table of Figures

Figure 1 Pie Chart Image 1	9
Figure 2 Pie Chart Image 2	9
Figure 3 Pie Chart Image 3	10
Figure 4 Boxplot Image 1	10
Figure 5 Boxplot Image 2	11
Figure 6 Boxplot Image 3	11
Figure 7 Time Series Image 1	12
Figure 8 Time Series Image 2	13
Figure 9 Time Series Image 3	14
Figure 10 Correlation Heatmap	15
Figure 11 Scatterplot Image 1	16
Figure 12 Heatmap Image 1	17
Figure 13 Heatmap Image 2	18
Figure 14 Heatmap Image 3	19
Figure 15 Choropleth Map Image 1	20
Figure 16 Choropleth Map Image 2	21

I. Problem Statement

New York City is well known for its classic subways and yellow-colored taxis, but with the advancement of technology, transportation was also upgraded, giving us ridesharing companies, like Uber, Lyft, and Via. Ridesharing companies are convenient and easy to use for an average consumer. With just a tap on one's phone screen, one could find a ride and be off to where they need to go. Since there is apparent competition among the ridesharing companies, this project aims to aid competitive analysis on metrics, such as trip counts, trip miles, trip time, etc., and provide key insights to help support campaigns. This project seeks to give an overview of where the ridesharing market is at and provide a deep dive on popular locations, consumer behavior and more, using Plotly Dash Graphs.

II. Dataset

The dataset used for this project was from the NYC Taxi and Limousine Commission, a public service and government agency. A parquet dataset was retrieved from this site:

<https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

Due to the sheer number of entries in each file, only January 2021's data was used as a sample for this project. Furthermore, since the file contained 11.9 million entries, a stratified random sampling was conducted on the dataset, splitting the groups into 3, Uber, Lyft, and Via, using the companies' license numbers. Only 10% of the data was randomly sampled, leaving 1,190,845 entries to be studied.

The dataset contains 24 columns, which are:

- | | |
|------------------------|------------------------|
| - hvfhs_license_num | - tolls |
| - dispatching_base_num | - bcf |
| - originating_base_num | - sales_tax |
| - request_datetime | - congestion_surcharge |
| - on_scene_datetime | - airport_fee |
| - pickup_datetime | - tips |
| - dropoff_datetime | - driver_pay |
| - PULocationID | - shared_request_flag |
| - DOLocationID | - shared_match_flag |
| - trip_miles | - access_a_ride_flag |
| - trip_time | - wav_request_flag |
| - base_passenger_fare | - wav_match_flag |

Shuheng_Mo from Kaggle provides the description of these columns and a taxi zone lookup file in this link:

<https://www.kaggle.com/datasets/shuhengmo/uber-nyc-forhire-vehicles-trip-data-2021>

The GeoJSON file of New York city was retrieved from:

<https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm>

III. Data Cleaning and Feature Engineering

From the dataset's columns, only the following were retained:

- hvfhs_license_num
- pickup_datetime
- PULocationID
- DOLocationID
- Trip_miles
- Trip_time
- Base_passenger_fare
- Tolls
- Congestion_surcharge
- Tips
- Driver_pay

Feature engineering was conducted on the hvfhs_license_num, changing the license numbers into the ride companies' names and changing the column name to Ride Company. A ride count column was added, which only contains 1s, to help keep track of a company's ride count. The pickup_datetime column was changed into a date only format and renamed as Date. Lastly, the columns were renamed to exclude underscores in the title and were capitalized in preparation for the dashboards.

IV. Plotly Dash Graphs

A. Pie Chart

The pie chart shows the distribution among Uber, Lyft, and Via based on the three options provided, the total number of rides provided, the total miles of their trips, and the total minutes of travel. People using this dashboard can choose options using the dropdown list as well as exclude any of the three ride-sharing companies mentioned for a one-to-one comparison between two companies. This chart shows a quick overview of the market distribution in the industry.

The code used to make this graph were based on Plotly's guide on making pie charts. A dropdown list was added to provide the viewer a quick view on other metrics without the need to look at other graphs. The Dash app is running on <http://127.0.0.1:8060/>.

Figure 1, Figure 2, and Figure 3 are images of the pie chart.

B. Box Plot

Box plots are useful tools to see the minimum, maximum, mean, and outlier values of a dataset as well as where its distribution lies in comparison to other companies. The viewer can use the dropdown menu to choose between trip miles, trip time, congestion surcharge, base passenger fare, driver pay, and tips to have a side-by-side comparison with the companies. Additionally, hovering a mouse over the parts of the box plots shows the properties of the data, like the quantile values and outliers. The graph is hosted on <http://127.0.0.1:8061/>.

Figure 4, Figure 5, and Figure 6 are images of the boxplots.

C. Time Series Chart

The time series graph shows the values of metrics chosen over a certain period. Three controls were coded into this graph. First is the radio item, where the viewer can choose between seeing the sum of the values or the mean. Although the sum of values gives a more accurate representation of what happened in real life, the average aggregation method gives us an insight on what the numbers look like if all three companies were on the same footing. An example is, based on the pie chart above, Via seems to have the lowest share among the companies. However, if we look at the average trip time of all three companies, we can see that Via users have highest times compared to Lyft and Uber users. This could suggest that Via users are more inclined to take farther trips, or Via drivers are more willing to drive farther away. The second widget is a dropdown menu on metrics. Multiple numerical values were included in this list in case there are time-based patterns in the data. Lastly, the third widget is a date range slider, where the user can change the starting and ending dates. In the code, the minimum and maximum values were used for the ticks. This dashboard is hosted on <http://127.0.0.1:8062/>.

Figure 7, Figure 8, and Figure 9 are images of the time series chart.

D. Scatter Plot

Scatter plots can help with examining the relationship between a dependent and an independent variable. In this project, a quick correlation analysis was plotted on a heatmap to guide the viewer which variables are statistically correlated to each other. After the heatmap, a scatter plot dashboard was made with three dropdown menus. The first one let's the user choose which ride company should be examined. The second and third menus let's the user choose the independent and dependent variables, respectively. The dashboard is hosted here: <http://127.0.0.1:8063/>.

Figure 10 is the correlation heatmap, while Figure 11 is the scatterplot image.

E. Heatmap

In this project, the heatmap shows which pickup and drop off zones are most frequently seen among consumers. This dashboard has 2 single dropdown menus, and 2 multiple dropdown menus. The 2 single dropdown menus are for selecting which company to examine and for selecting the value, whether it would be ride count or congestion surcharge. The 2 multiple dropdown menus are for the pickup and drop off zones. Since there are a total of 265 zones in New York, the graph was hardcoded to initially show all zones together, but the user can choose the top 50 zones or individually select the areas they have in mind. With this graph, the user will be able to see which pickup and drop off zones are hotspots, while also checking which zones have the highest surcharge among the companies. The graph is hosted on <http://127.0.0.1:8064/>.

Figure 12, Figure 13, and Figure 14 are images of the location heatmap.

F. Choropleth Map

The final graph is a choropleth map of New York with each zone separated. The graph aims to visualize which zones of New York are “hot” based on ride count and congestion surcharge. The user can change the data with three dropdown menus. The first dropdown menu chooses which company to analyze. The second menu selects if it is a pickup or a drop off event. And the third menu selects if the data to be visualized is ride count or congestion surcharge. With the choropleth map, the user can determine hotspots among the city’s zones to advise more drivers there or target more prospects in that area. This graph is hosted on <http://127.0.0.1:8065/>.

Figure 15 and Figure 16 are images of the choropleth map.

V. References

<https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

<https://www.kaggle.com/datasets/shuhengmo/uber-nyc-forhire-vehicles-trip-data-2021>

<https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm>

<https://dash.plotly.com/>

VI. Appendices

Market Distribution

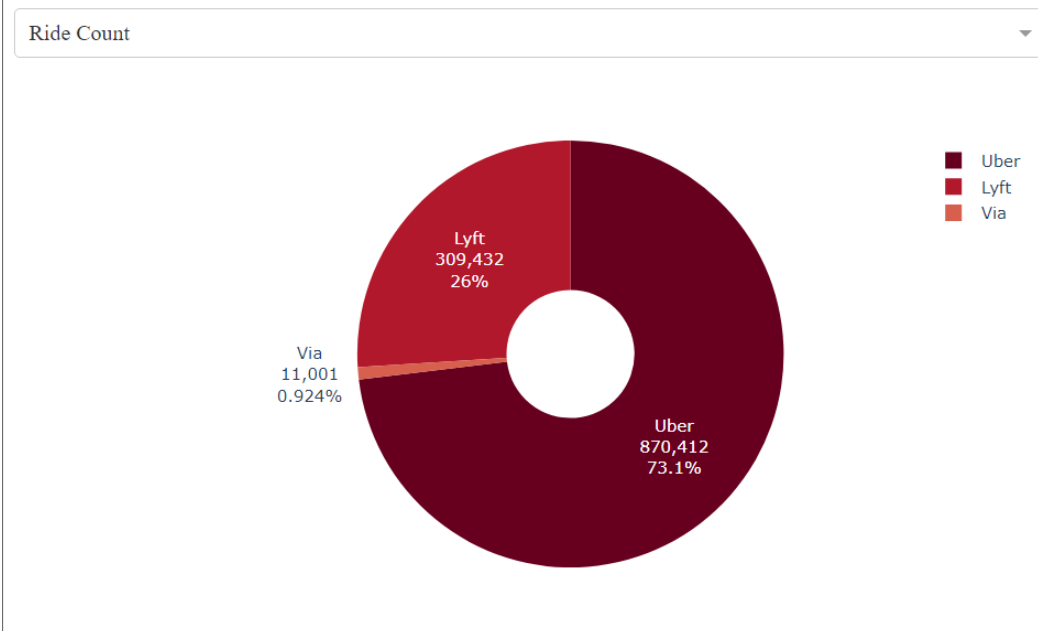


Figure 1 Pie Chart Image 1

Market Distribution

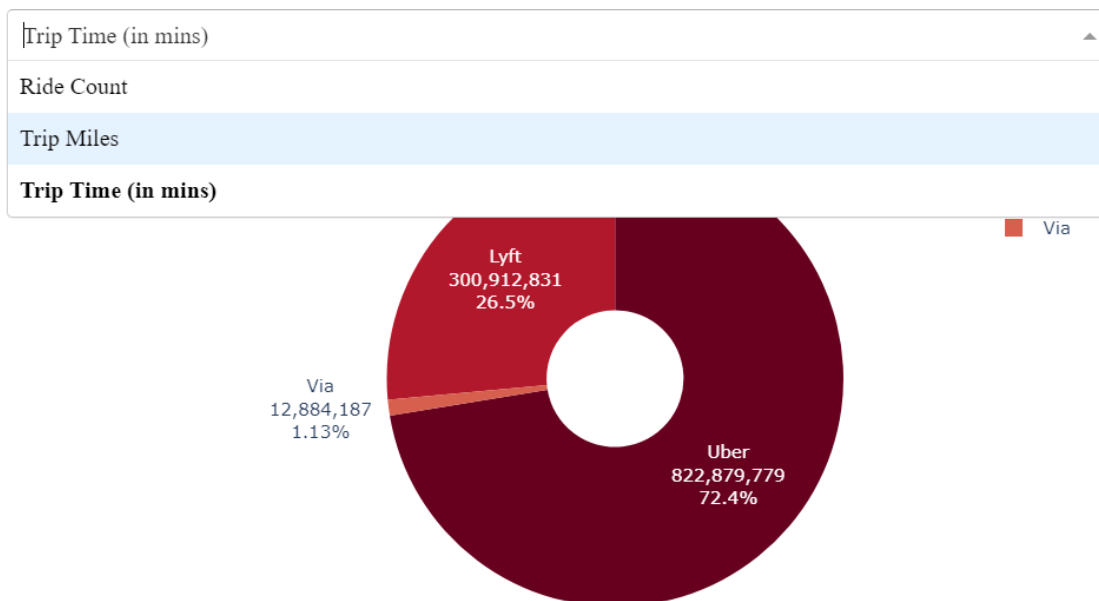


Figure 2 Pie Chart Image 2

Market Distribution

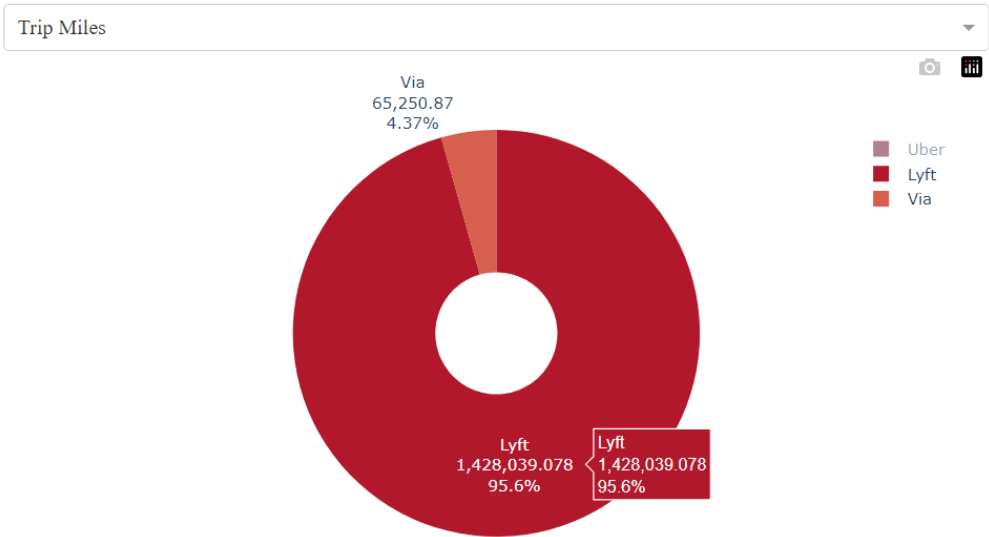


Figure 3 Pie Chart Image 3

Boxplots by Ride Company



Figure 4 Boxplot Image 1

Boxplots by Ride Company

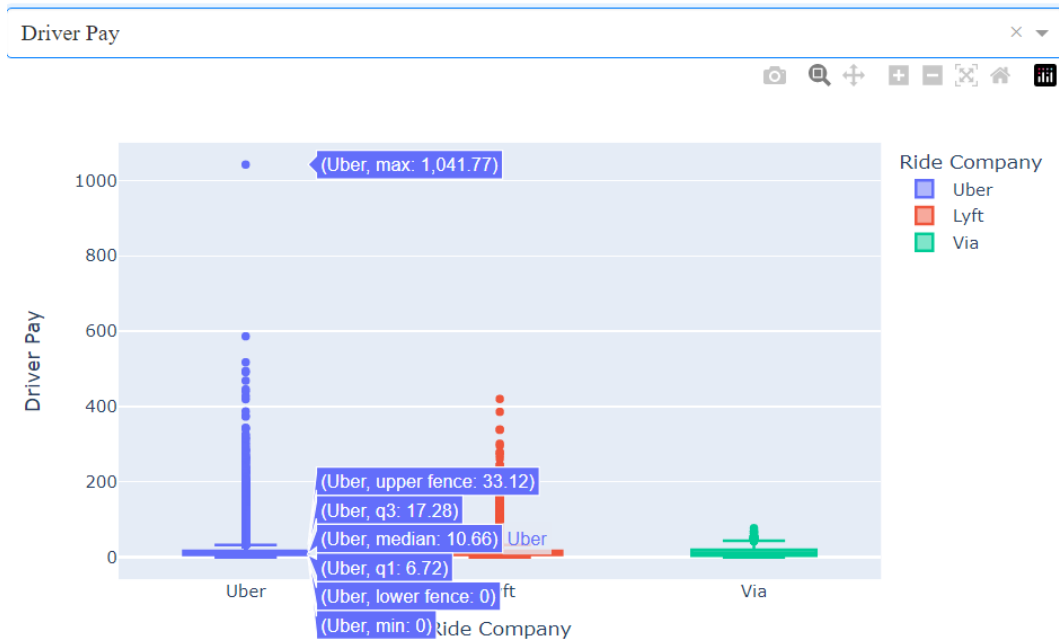


Figure 5 Boxplot Image 2

Boxplots by Ride Company

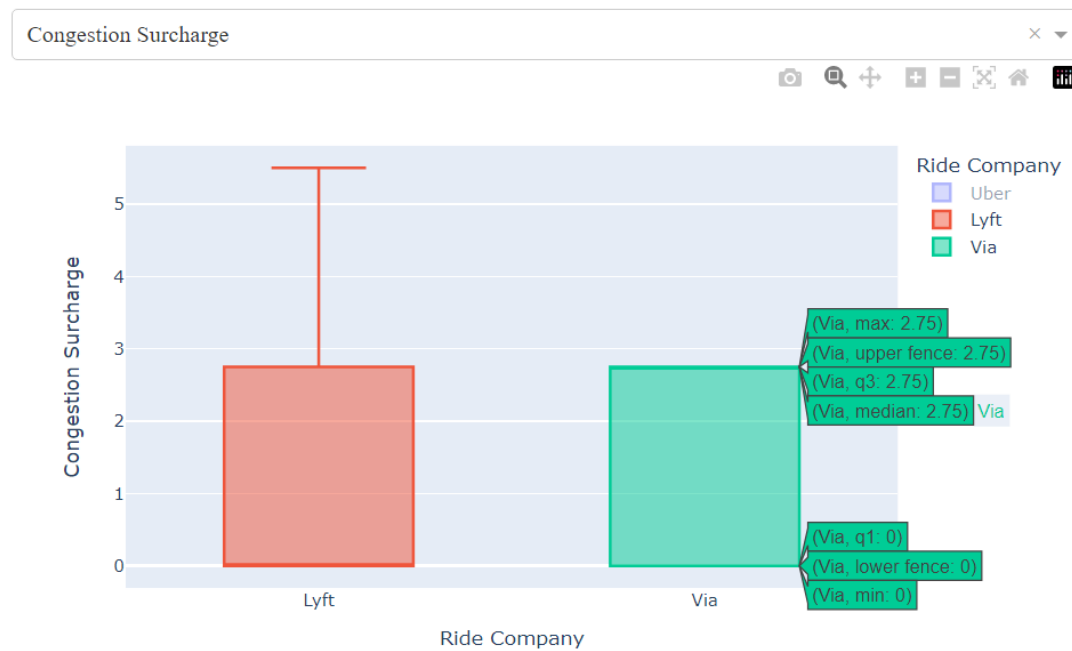


Figure 6 Boxplot Image 3

Time Series Graph

Aggregation method:

- ☒ Sum
- ☐ Average

Metric:

Trip Time

×

▼

Date range:

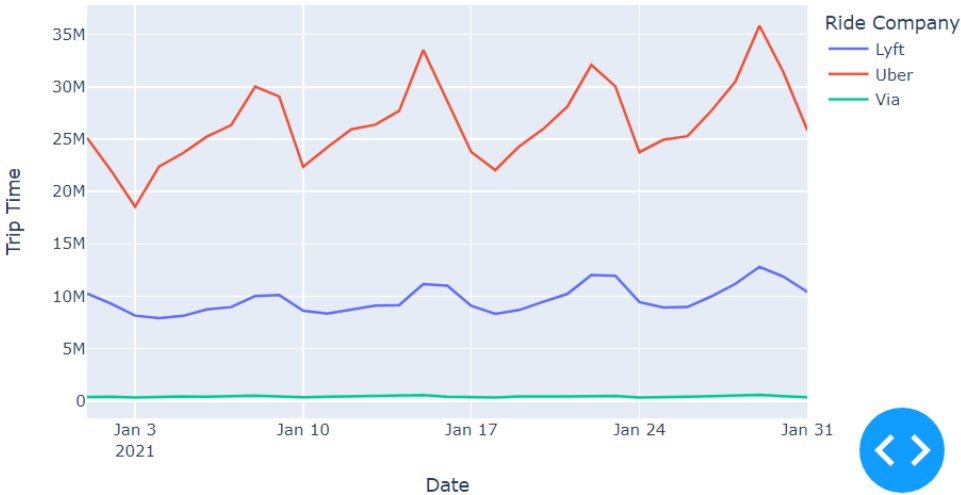


Figure 7 Time Series Image 1

Time Series Graph

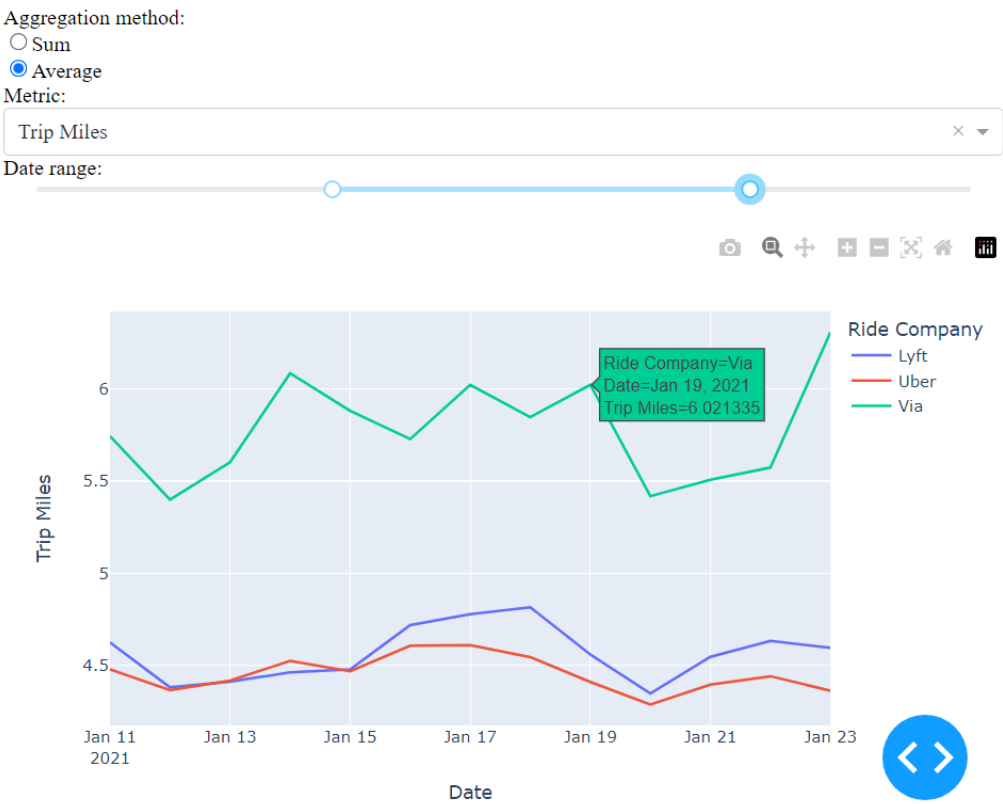


Figure 8 Time Series Image 2

Time Series Graph



Figure 9 Time Series Image 3

Correlation Heatmap

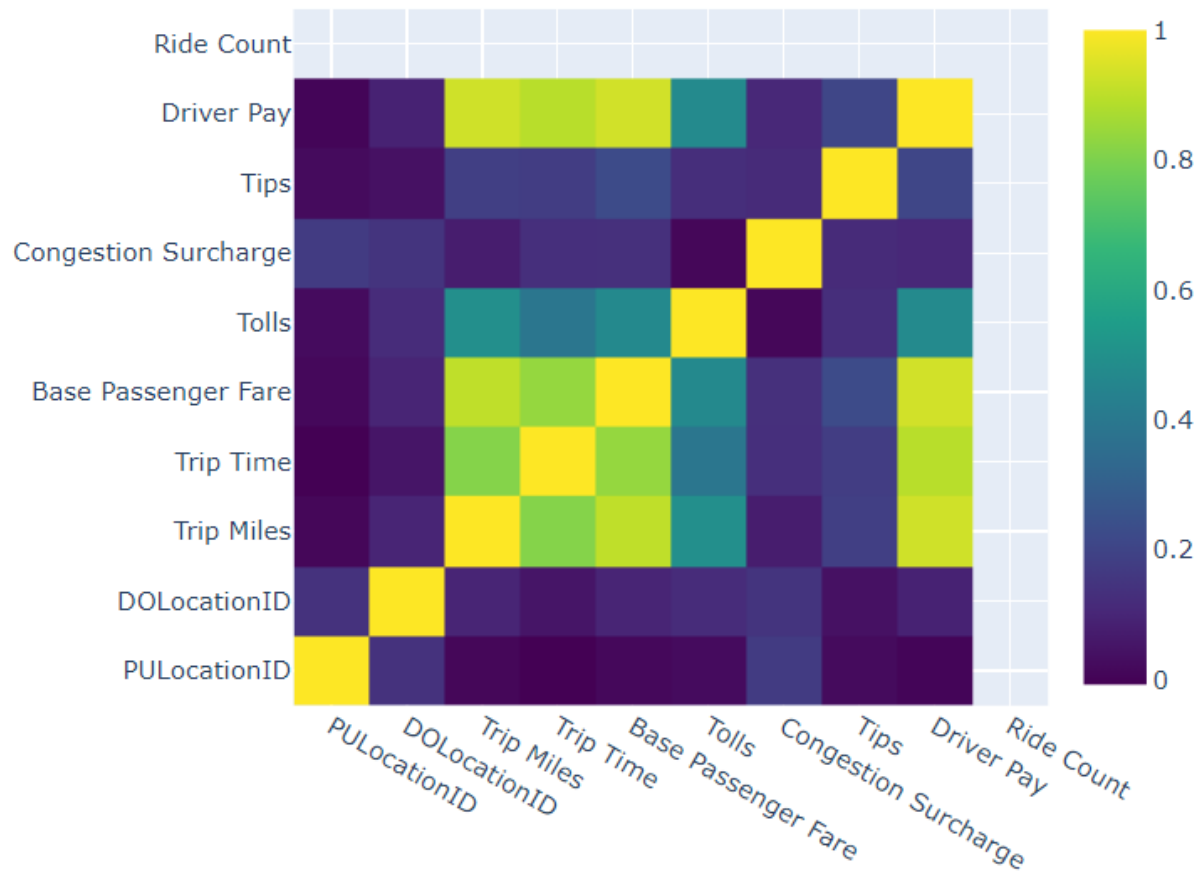


Figure 10 Correlation Heatmap

Scatterplot

Ride company:


Uber

Independent variable:

Trip Time

Dependent variable:

Base Passenger Fare



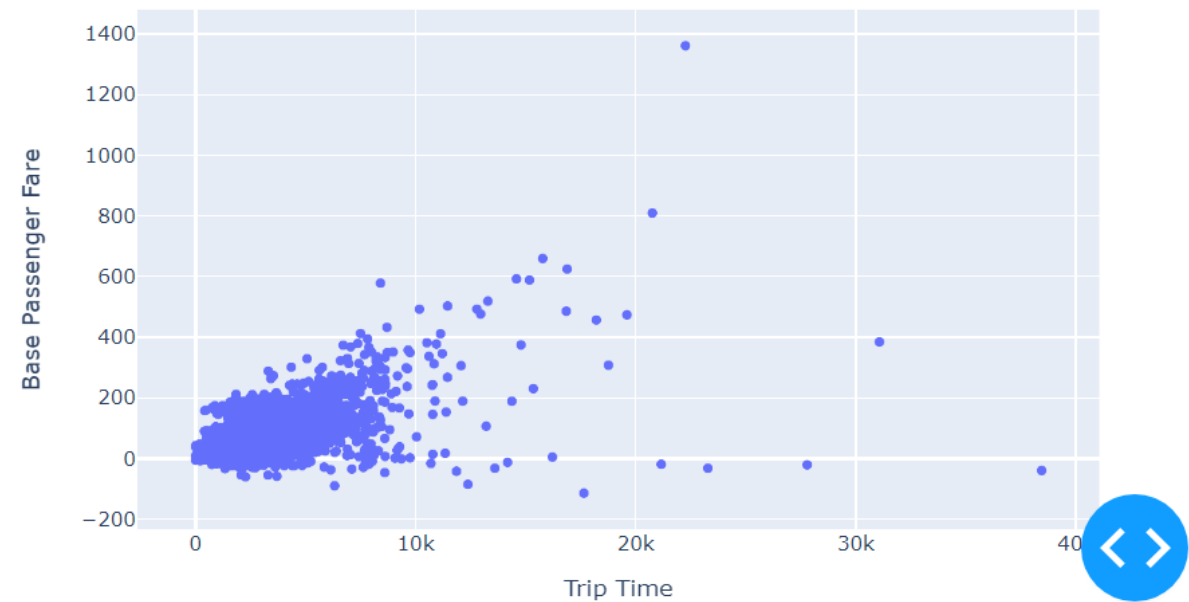


Figure 11 Scatterplot Image 1

Heatmap of Pickup and Dropoff Locations

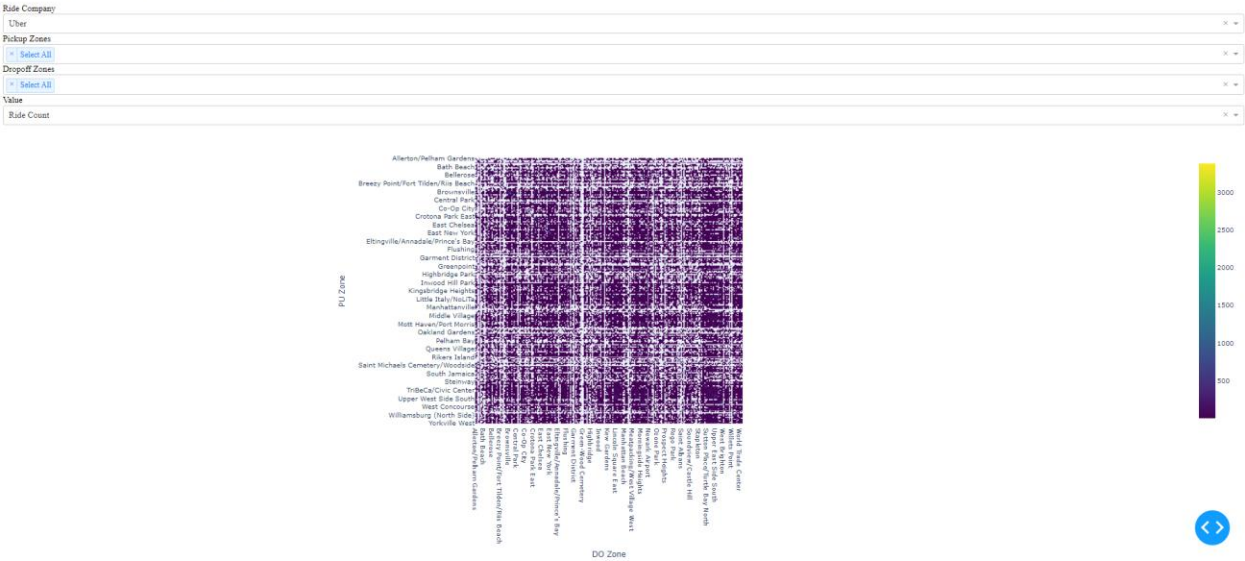


Figure 12 Heatmap Image 1

Heatmap of Pickup and Dropoff Locations

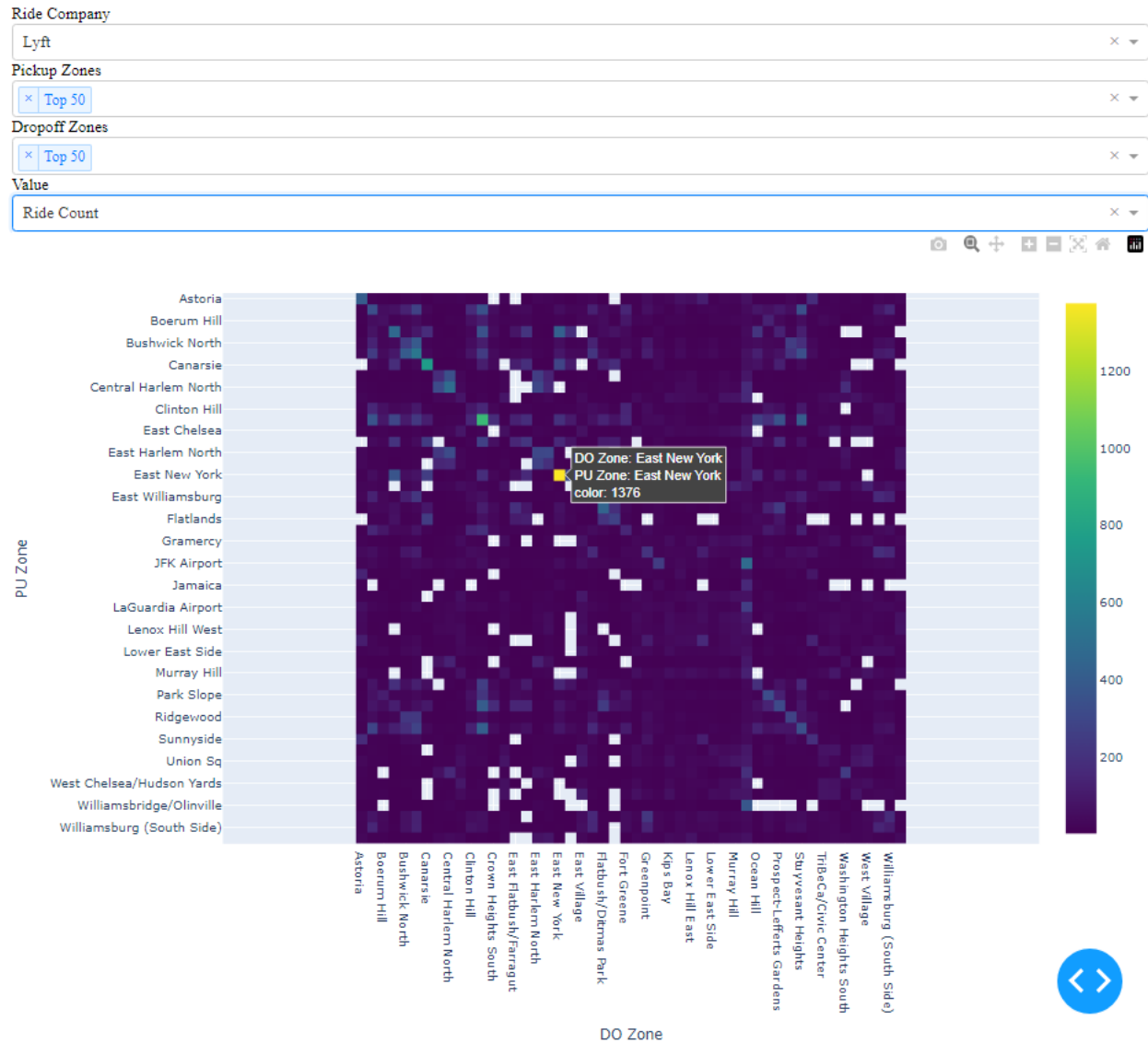


Figure 13 Heatmap Image 2

Heatmap of Pickup and Dropoff Locations

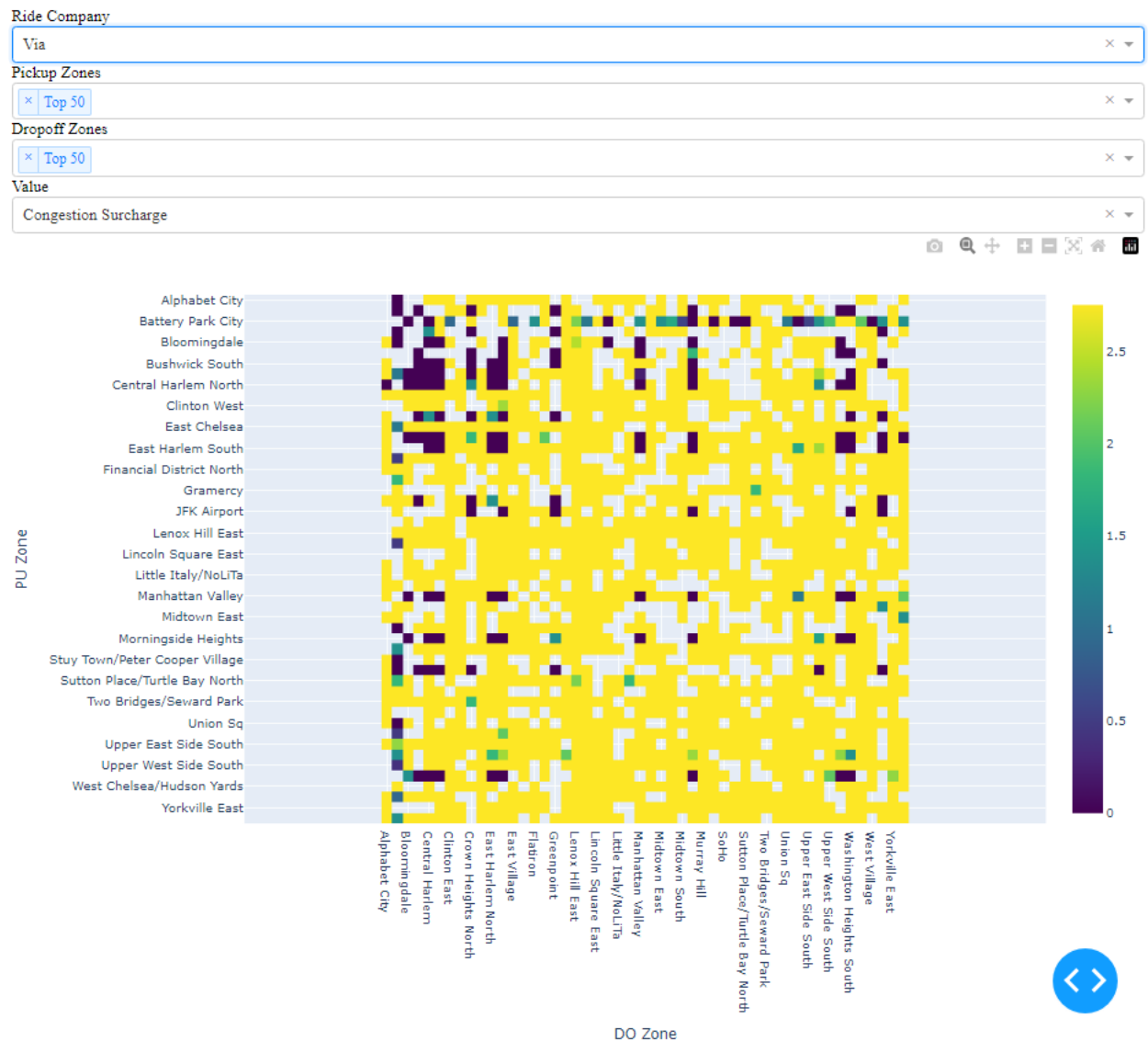


Figure 14 Heatmap Image 3

Choropleth Map of New York by Ride Company

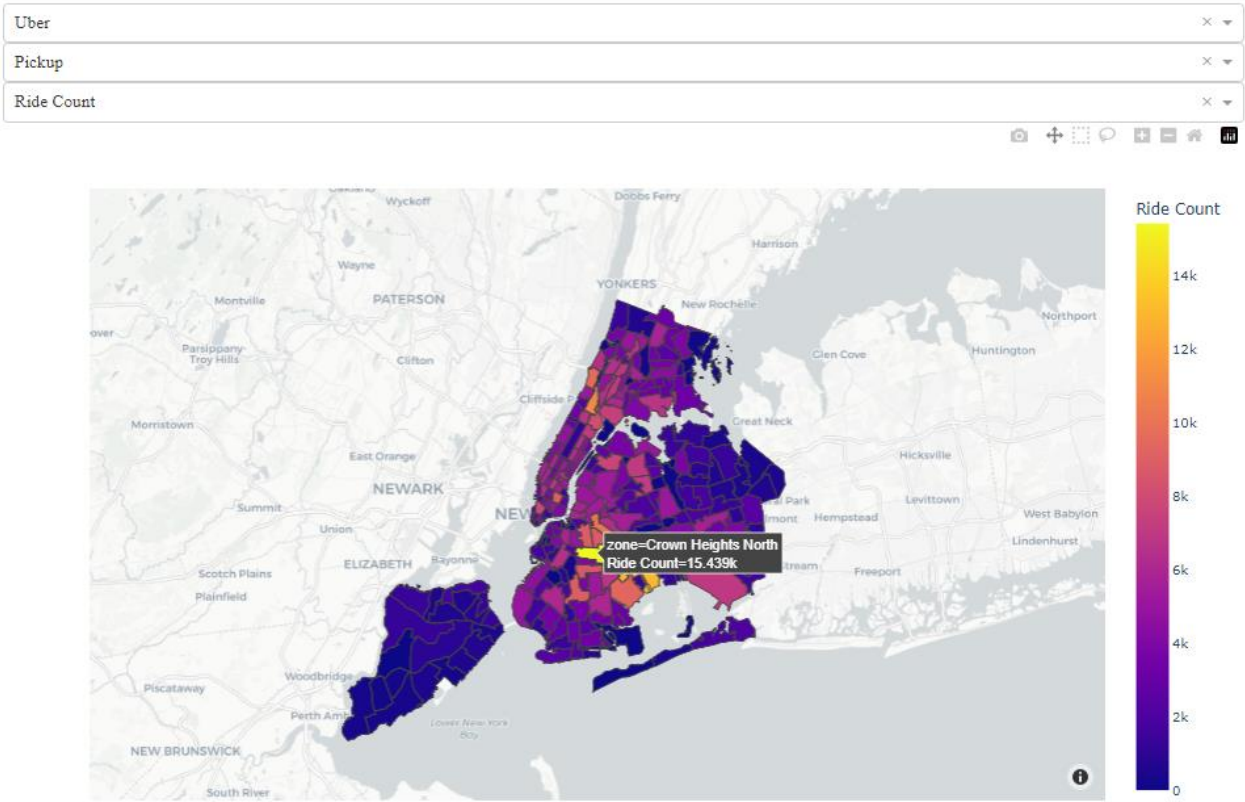


Figure 15 Choropleth Map Image 1

Choropleth Map of New York by Ride Company

Via	x	▼
Dropoff	x	▼
Congestion Surcharge	x	▼

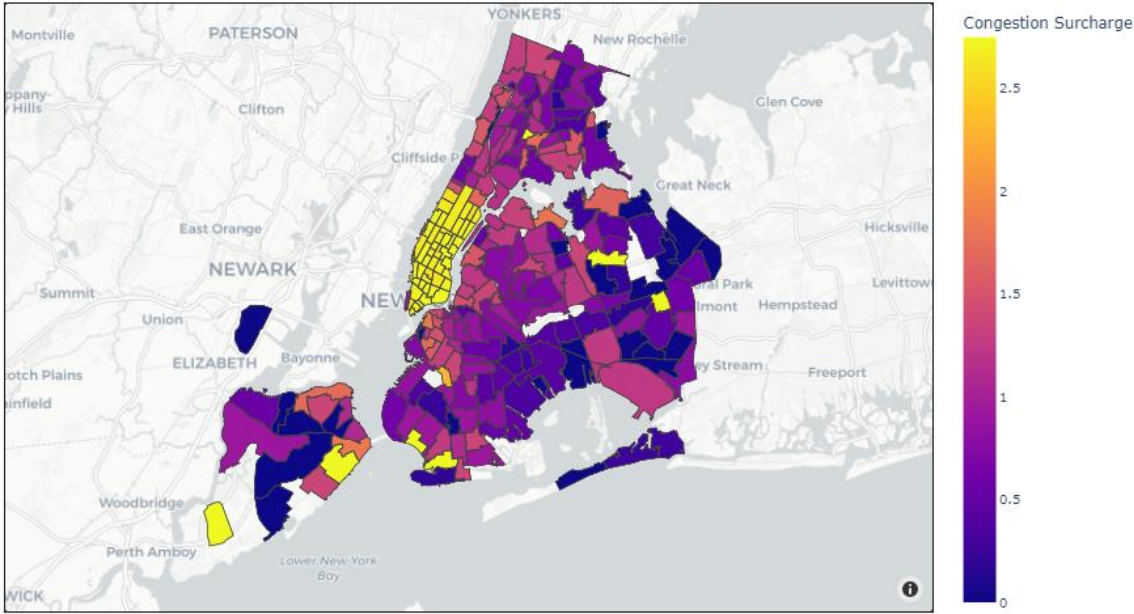


Figure 16 Choropleth Map Image 2