Taxi trips in New York

Datavisualization course

1]Nicklas Marc Pedersen 2]Dinh Phu Luu 3]Ahmadullah Naibi 4]Casper Fenger Jensen

Prepared for 28/12/2023
Link to dashboard: https://datavis.anusmountain.com/
GitHub: https://github.com/Nicklason/datavisualization

Abstract

This report is a study of how we visualize data using R and knowledge from the data visualization course. The chosen data is from the NYC government website specifically about taxi trips with yellow caps. To visualize how the data can be utilized in graphs for the most impactful way to display data and to generate as much information without any prior knowledge of the data. For data processing, the chosen dataset was constrained into a single month due to the large number of records. To the dataset, some questions were made which were answered using visualizations, including graphs like heatmap, flowmap, box plots, and more.

Contents

1 Introduction

This project analyses the TLC Trip Record Data dataset [1]. The dataset is made up of yellow taxi trip records from the New York City taxi zones. The dataset includes various fields, such as pick-up and drop-off times and locations, trip distances, payment amount, payment type, and passenger counts. This dataset has been made available by the NYC Taxi and Limousine Commission (TLC).

The dataset consists of files that contain a collection of taxi trips for each month. This project focuses on the dataset from January 2023.

The reason for this dataset to be chosen is for various reasons.

- The dataset has millions of trips for each month
- Different types of data, most notably geospatial data
- GeoJSON file to view NYC taxi zones

The dataset was discovered during the "Big Data and Data Science Technologies" course. The couse teaches about different technologies in handling big data, where some of the technologies used this dataset as an example [2].

Many interesting questions can be derived from the dataset, and the decision to choose this dataset was also heavily influenced by the geospatial aspect of the dataset [3], and the possibilities in visualizing the movement of people in a big city like New York City. New York City consists of five boroughs [4], and each borough is also classified in the dataset. To put it simply, the dataset consists of taxi trips for the five boroughs, which are made up of 265 zones and have over three million trips for January 2023. Pick-up and drop-off locations are described as location IDs that correspond to a specific zone. Because of this, it is not possible to know the exact address of a pick-up or drop-off.

[1] https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
[2] https://docs.stackable.tech/home/stable/demos/trino-taxi-data.html
[3] https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc
[4] https://da.wikipedia.org/wiki/New_York_City

2 Project objectives

With the dataset, the group wishes to answer a list of questions. Each question is listed below together with an explanation for why it was chosen to be investigated. Some of the questions might not have been answered during the project, but this will be explained in the conclusion/discussion section of the report.

2.1 Difference in amount of taxi trips at different times

This question was created from the idea that the dataset consists of many months of data, and that different periods should show different patterns in taxi usage. It would be interesting to see the difference in the amount of taxi rides at different times of the day, different weekdays, different months, and even by zone. However because the dataset was chosen to only be of January 2023, then the different months will not be investigated.

Visualizing the amount of taxi trips at different times could be done using heat maps, bar charts, scatter plots, and line charts.

2.2 Correlation between different variables

The dataset consists of many columns. Each column has some sort of relation to each other for each row of the dataset. It would be interesting to find, and visualize, the correlation between trip distance, trip time, amount of passengers, time of day, taxi fare, and more. This wants to answer if it is possible to find a correlation by visualizing it. There should be a correlation between the different variables and the taxi fare because taxi trips are usually priced by time and distance.

2.3 Money spent on taxi trips

Various columns are available regarding the price of the taxi fare. This data should be visualized, such as viewing how much money is spent in each zone, how much money is spent each day, and each hour. These visualize could be interesting for a taxi company to efficiently distribute their taxi around the clock and according to zones.

2.4 Time spent on taxi trips

The time spent on a taxi trip can be calculated by subtracting the pick-up time from the drop-off time. The amount of time spent on taxi will increase depending on the trip distance and the amount of traffic, but traffic is not part of the dataset. It would be interesting to correlate the taxi trips and the duration of the trip with the traffic at a particular time.

2.5 Amount of taxi trips for each zone

The dataset consists of three million trips that are all distributed over 265 zones. It would be interesting to create a heatmap that displays the number of taxi trips for each zone using the GeoJSON that contains shapes for each of the taxi zones. Other visualizations can also be made, but a heatmap was the main interest.

2.6 Taxi trips to and from airports

The NYC taxi zones contain three major airports, Newark Airport, JFK Airport, and La-Guardia Airport. Each airport is its own zone. It would be interesting to do visualizations based on these zones. The airports are the main gateways for tourists to NYC. With this, it would be interesting to show the movement of people to and from airports. This could also be done for each zone, to see where people move to and from. It could also be combined with the different times, to see the movement of people at different times and dates.

2.7 COVID-19 pandemic affect on taxi trips

The complete dataset starts from January 2009. It would be interesting to see how the COVID-19 pandemic affected taxi trips, such as the movement of passengers, and the amount of trips. The focus is on January 2023, but it would be interesting to include different months and see the data for different months of different years. It could also be used to answer questions about how seasons affect taxi trips.

2.8 Taxi trips speed

With this question, we want to look into the speed of taxi trips. Such as what is the average speed of a taxi trip in general and in comparison with the number of passengers and distance. It could be imagined that longer trips would use a highway and achieve higher average speed than shorter trip distances. This could also be combined with the trip distance, assuming they are trying to make trips more efficient by having larger taxis and collecting more passengers for longer trips.

2.9 Distribution of payment types

The distribution of payment types, or the percentage of cash and credit card payments would be interesting to see to find out whether people use cash or credit card most often in each zone and the total amount of each type.

3 Data

3.1 Data source

Data was taken from the TLC Trip Record Data dataset https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page .

3.2 Description

The data describes taxi fares for a specific month in NYC, where data contains the number of passengers in the taxi, price of the fare, tip amount, pickup location, destination, distance to destination, and zones. It consists of parquet files with NYC taxi trips for each month, a GeoJSON file for visualizing the taxi zones of NYC, and a CSV file for mapping the ID of zones with their name and borough. The focus is on one month January 2023 which makes the dataset a collection of 3066766 records and 22 columns, to answer some of the questions additional columns will need to be calculated or combined with other datasets.

3.3 Data processing

The data contains many weird values, such as trips of negative distance and negative time. Values like these are filtered out where appropriate, such as for calculating the average speed of a taxi during a taxi fare. While the distance or time might be wrong, we can't prove or deny that the taxi fare amount is not correct.

4 Visualization

4.1 Design

Figure 1: Map of NYC with overlay of taxi zones (source)

The dataset has geospacial aspects. The pick-up and drop-off locations can be referenced in a GeoJSON file that can then be plotted to create shapes, an example of this can be seen on fig. 1. Heatmaps can be created, where each zone can have a different color depending on different values. This works well for visualizing zones relative to each other, but to visualize the movement of passengers then you need a flow map.

All trips have a start and end location. This can be used to create a flow map. There are $265 * 265 = 70225$ possible ways to move between zones. Consider two zones, A and B, you can move from zone A to A, B to B, A to B and B to A. It is directional, meaning the total amount of trips between two zones A and B is $A_B + B_A$. It would be confusing to attempt to understand 70225 connections and a color or thickness to represent the amount of trips. Because of this, the use of filtering would be required. The same can be said about scatter plots. There are over three million rows in the dataset and it would work very poorly to have a point for each trip. Making the user able to filter is a necessity to be able to make sense of the data when there is too much of it.

Other types of plots, such as bar charts, line charts, and pie charts can also be used to efficiently convey the dataset. Line charts can be used to see trends over time, bar charts can be used to quickly compare different values, such as values for zones, and pie charts. Pie charts should only be used if you have a good reason because it is difficult to understand the area of slices and compare it, while a bar chart you can easily tell that one bar is larger than another.

4.2 Requirements

For this project were some must-have requirements as part of the report skeleton. Besides these, we made some ourselves which we also consider must-haves and are therefore required for us to consider the project to be a success.

- Able to fulfill the project objectives and answer the questions
- Interactive elements in plots
- Intuitive visualisations

- Create map of the taxi zones

The plots should be interactive to allow the user to personally explore the visualizations, for example to only see data for relevant zones because it is infeasible to display it all at once. It can be done using hovering to show more info, brushing to select an area, filtering by time, or selecting specific taxi zones etc.

4.3 Optional features

Drag the mouse over an area of a visualization to get detailed information about the marked area, or get a summary of the data within the marked area, such as points on a scatter plot, or taxi zones on a map. For some visualizations it might be necessary, but it is mostly seen as a nice to have, to be able to gain more insights and to make the visualizations more interactive.

5 Results

5.1 Difference in amount of taxi trips at different times

Answering the question from section 2.1.

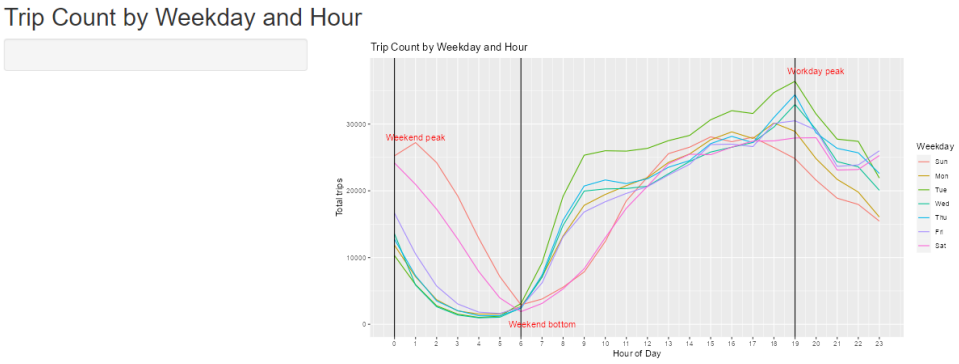

Figure 2: Trip Count by Weekday and Hour

A line chart (fig. 2) is used to display the taxi trips at different times of the day on each weekday. The trend is similar for all the days with around 6 AM being the most quiet time

all days. In the weekend you can see that the the bottom has been moved two hours later. This is probably due to people going out on Friday / Saturday night and need to take a taxi home later. The difference between weekends and workdays early on the day was expected, but the close similarity throughout the rest of the day being the same for all days was odd. The visualization works well because it clearly shows the trends for different weekdays and the annotations makes it easy to explain the results to the viewer.

5.2 Correlation between different variables

Answering the question from section 2.2.

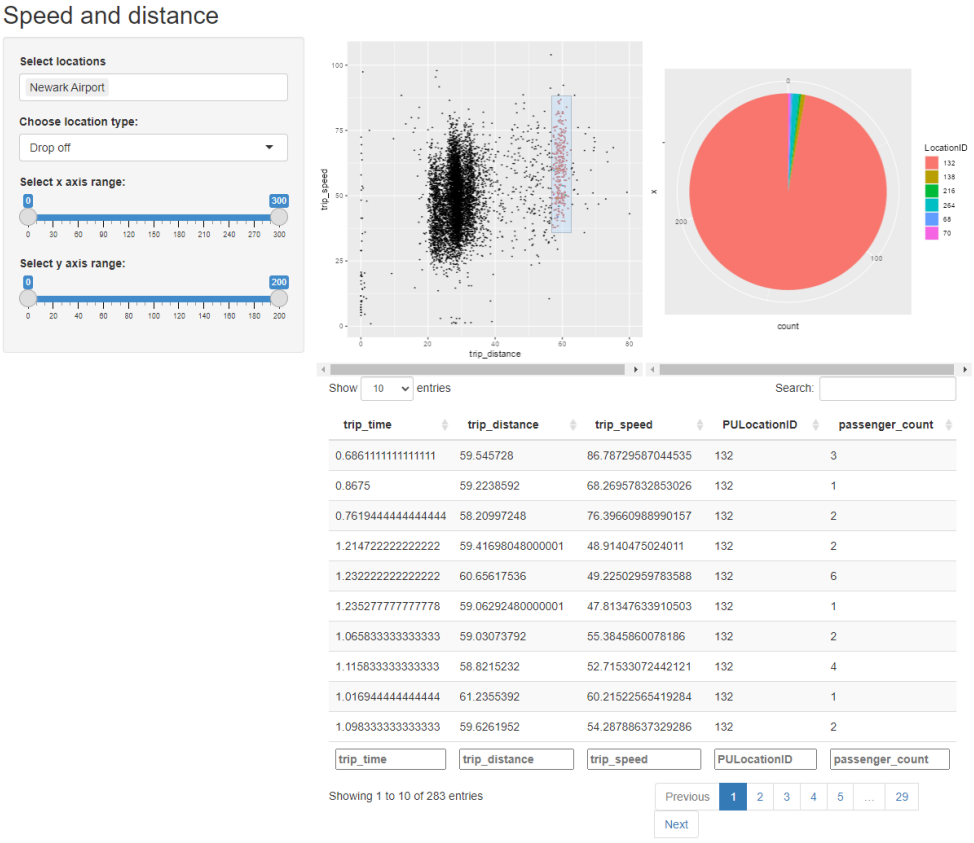

Figure 3: Speed and Distance

Plotting trip speed and trip distance in a scatter plot (fig. 3) gave interesting results. While creating the graph we noticed that the points are creating clusters at roughly the same

distances (x-axis) but different speeds (y-axis). The hypothesis was that the trips might be going to/from the same locations inside these clusters. To test the hypothesis, we made it possible to select points on the plot and show it in a table. It showed that many of the trips were indeed to/from the same location. A piechart was created to show the distribution of locations within the selected points and the color of the selected points were changed to match the piechart. Almost all of the trips inside the selected area start at zone 132, which is the JFK Airport. We can see a correlation between distance and pick-up/drop-off zone, which makes sense because the trip distance should roughly be the same between the pick-up and drop-off zones. It should be noted that just because you start and stop in the same zones as other trips, it does not mean the distances are the same because you could start and stop at different locations inside a zone which would make the distances different depending on how big the zones are. It should also be noted that just because the trips are of similar length does not mean it is to the same zones, because it is just a radius from the starting point, where the end locations could be in widely different directions but still be the same distance away. You could be in a situation where one cluster is made up of many different locations, but this can be seen using the piechart and the coloring of the points when selecting them. A major improvement would be that the units should be explicitly stated to not confuse the reader. For example, in fig. 3, the columns does not include units. The trip_time column is in hours, the trip_distance column is in kilometers, and the trip_speed is in kilometers per hour.

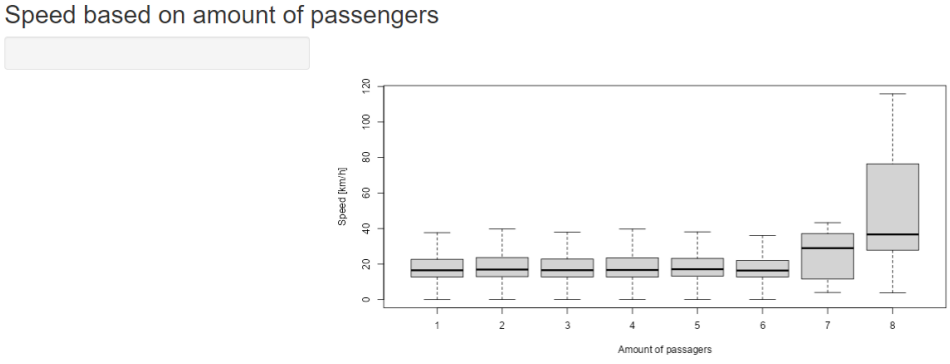

Figure 4: Speed Based on Number of Passengers

Visualizing for speed and passengers would be chaotic with plot points or brings no real impact on information with a bar chart, making them into a box plot helps to visualize in terms of the speed on the y-axis will show the percentiles taxi drivers speed, these values go from being numerical continuous to numerical discrete, with an x-axis showing the number of passengers in a categorical sequence. The initial expectation for the graph would be that driving with fewer people would be fast, however with the chart plotted shows the contrary, that the more onboard passengers the faster they drive, this would likely be the case that in 5-seat taxis they drive slower in the cities stopping a lot with traffic, whereas the people driving 8-seat vans goes through the highway from Newark to New York City. It would be interesting to see the amount of passengers, speed, and distance, in a single visualization. It would also be an improvement to show the amount of rides in each box. There might be very few trips with 8 passengers but they might drive very far, making the average speed higher.

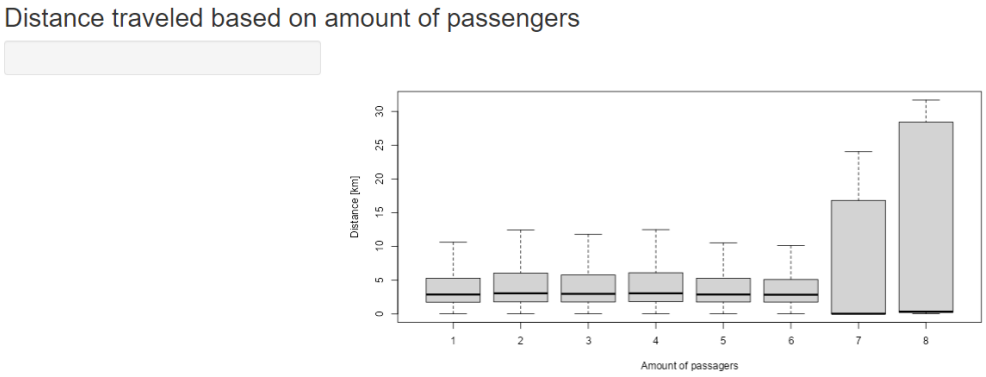

Figure 5: Distance Based on Number of Passengers

The same goes for visualizing based on distance over the number of passengers, for the y-axis the dataset already has the distance accounted for in numerical discrete values, and the x-axis is the same for passengers being categorical like fig. 4. By the same assumptions, the initial expectation was that the fewer people the shorter the distance they drive, which is correct in this case that one to six people take a taxi less than five km but the distance traveled for an 8-seater has a median under one km, meaning that one-half of the 8-seater

trips has gone over one km distance.

5.3 Money spent on taxi trips

Answering the question from section 2.3.

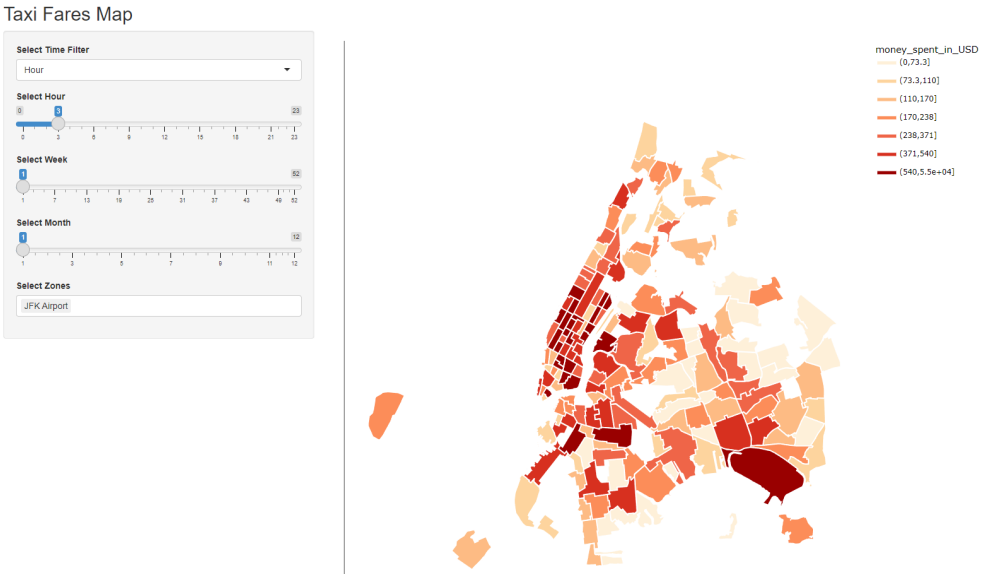

Figure 6: Taxi Fare Map

To visualize the money spent, a heatmap was made (fig. 6). For selectors are hour, week, month, and zones. The figure money spent at 3 AM on trips to and from JFK Airport. It shows that many of the zones has trips going to/from the JFK Airport zone. When changing from 3 AM to 1 AM or 7 PM when there are busy hours, all numbers are written scientifically because of the size of the numbers, which makes it hard to read. An improvement that should be made is to use full writing of numbers, so the length of the numbers can be compared visually. The month slider was also not implemented because it would use too many resources to look at all months of the entire dataset. The tooltip only shows on the edge of zones, which is annoying because edges can consist of more than one zone.

5.4 Amount of taxi trips for each zone

Answering the question from section 2.5.

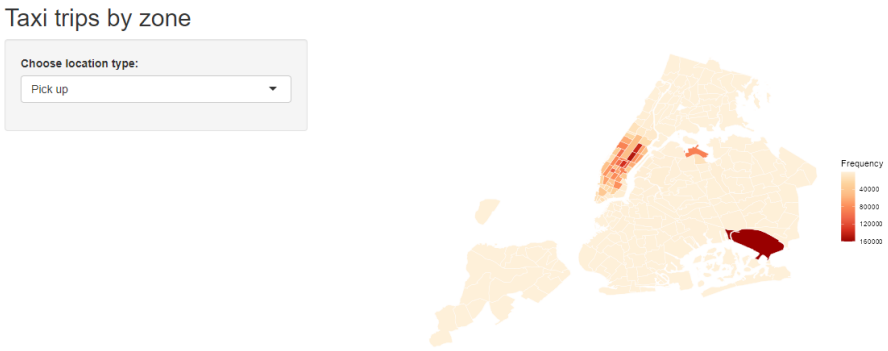

Figure 7: Taxi Trips by Zone

For this question, two visualizations were made, the first being a heatmap to compare all the zones fig. 7 this having the selector of pick-up or drop-off showing that the JFK Airport is the zone with most pick-ups. All the other zones have far fewer pick-ups and it is hard to compare these with each other. An improvement would be to use a non-linear scale, such as logarithmic, to get a larger difference in color between zones.

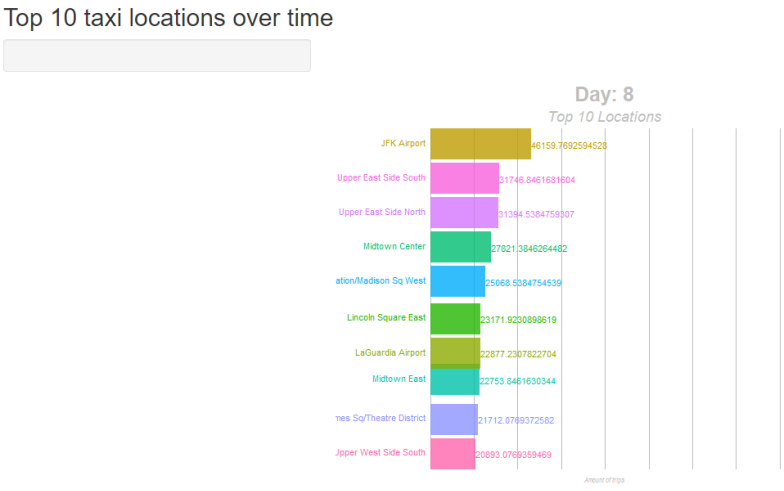

Figure 8: Top 10 Taxi Locations Over Time

To get a clearer difference between zones, and to fulfill the animated visualization requirement, the top 10 total pick-ups and drop-offs are plotted in an animated bar chart over the whole month with each step being a full day, a frame from this animation is shown in

fig. 8 where LaGuardia Airport is switching positions with Midtown East. This only shows the top 10, but allows the viewer to get more detailed of how different these zones are. An improvement is a selector with the range allowing for custom ranges to be explored like the bottom 10 or middle 5, and maybe even display the zones on a map using the same colors in the bar chart to make it easy to see where the zones are located.

5.5 Taxi trips to and from airports

Answering the question from section 2.6.

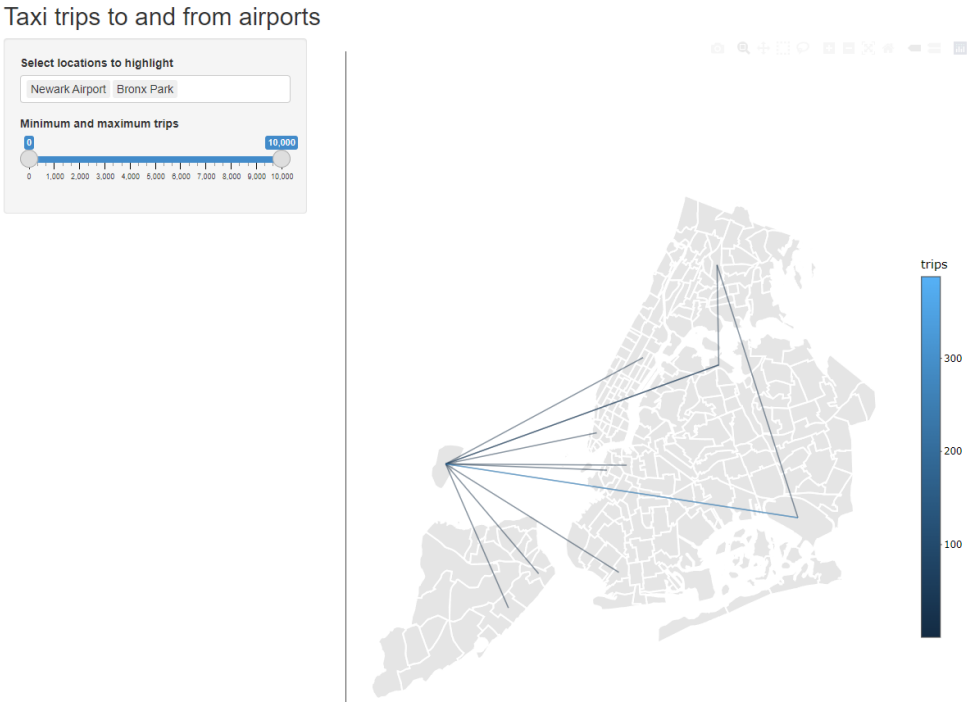

Figure 9: Taxi Trips to and from Airports

To see which zones have trips to and from airports a flow map is used (see fig. 9). The lines shows the start and end location, and the color is the number of trips. This chart has two selectors, the first is for zones allowing to plot multiple zones at once, and the second is the range of the number of trips, this is great when highlighting zones that have trips going to many different zones, such as the JFK Airport zone. Improvements to the chart could be

to try to replace the color distinction between the number of trips with the size of the line being relative to the number of trips. Arrow heads should also be used to better indicate the direction of the trips.

5.6 Taxi trip speed

Answering the question from section 2.8.

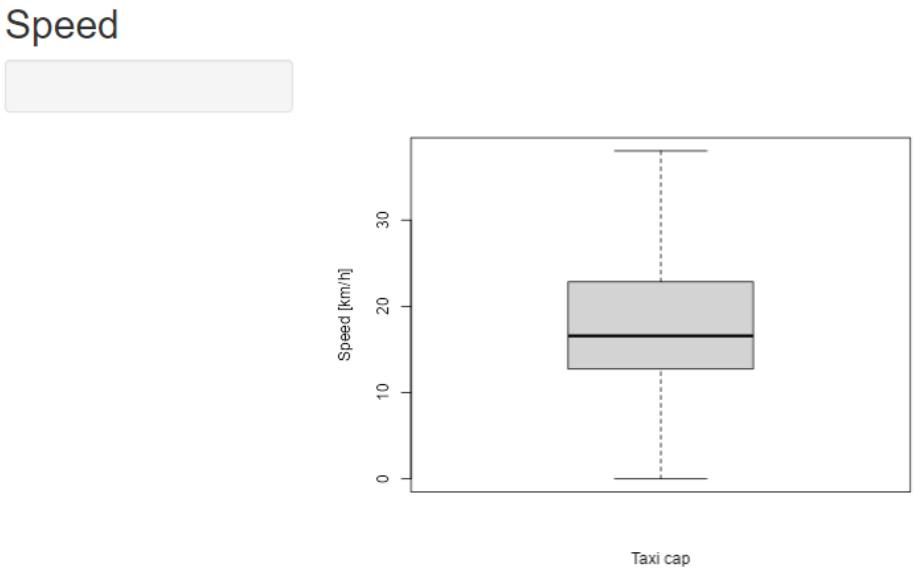

Figure 10: Taxi trip Speed

For curiosity's sake looking for speed alone to show a box plot for how the average speed a taxi driver would go, would show how the flow of traffic is to the busy streets of New York, using only calculated speed from taking the distance which is included in the dataset over the duration of the trip by calculating the estimated time between from pickup to drop-off, it was initially expected to be about 45km/h based on common assumptions of the flow of traffic and traffic laws, but the visualized data shows that the 50% percentile drives about 17 km/h. This shows that the median driver is more than 2 times slower than expected.

5.7 Distribution of payment types

Answering the question from section 2.9.

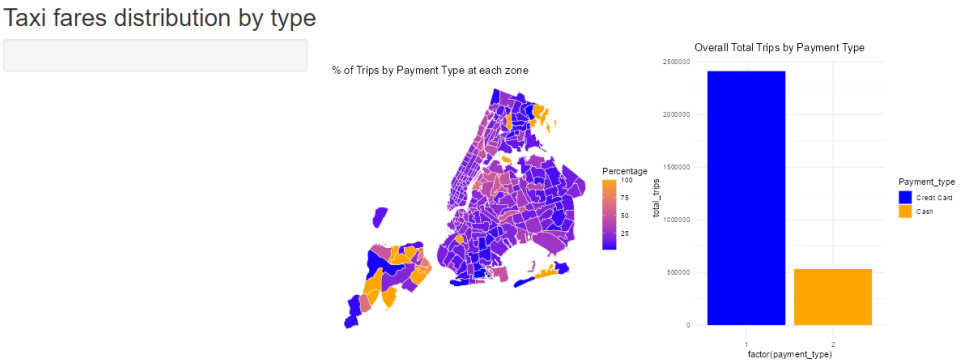

Figure 11: Payment Type Distribution of Taxi trip

To visualize the distribution of payments, the plot on fig. 11, was made. It is a heatmap visualizing whether credit card or cash is the major or equal payment method in different zones. Beside this is a bar chart showing the overall distribution. It shows that airports and other tourist places are move towards credit cards whereas some outer zones majorly use cash. The bar chart shows the frequency of cash only being a fifth of credit card use. It was expected that credit cards would be more frequent.

6 Conclusion & Discussion

For this project, 7 types of graphs were made (line chart, scatter plot, pie chart, box plot, heat map, flow map, and bar chart) across 12 graphs, where one of them is animated.

Most of the challenges came from the size of the dataset. The size made it slow to process the data, but it also made it possible to create interesting visualizations because of the amount of data that was available. One of the major downsides was that the app crashes when using shinyapps because it runs out of memory on the free plan (1GB memory). While it might have been possible to download all datasets and create visualizations to answer the COVID-19 question, it would have been very inefficient. Focus was therefore put on filtering to optimize the speed at which graphs were created.

Filtering was a big part of the project. It was not feasible to plot three million points, and it was not visually possible to gain any valuable insights from it. But by using filtering, the user was made able to personalize the visualizations.