UBER: BUSINESS IN BOSTON

COMP4462: Data Visualization

Accelerate

CHENG, Ho Sing (20513950)
LIU, Yat Long (20520953)
WAN, Kwok Kit (20519679)
YEUNG, Man Yin Michael (20603418)

TABLE OF CONTENTS

1. Abstract

Uber is one of the largest for-hire vehicle operators in the world. In this project, a variety of tasks is conducted to analyze customers' behaviours and Uber's performance, by looking into customers' expectations and the correlation between factors and the frequencies of trips. A multitude of visualizations are used to show the findings. Customers favor professional service without technical and monetary issues. The number of trips depends on time and weather but not location. The price does not depend on location as well, and is not a primary factor in cross-operators comparison.

2. Introduction

As Uber expands its business, there are a total of 75 million users. Such figures reflect the wide use of Uber globally. We would like to take this opportunity to see if there are any insights related to its success today. In this project, we will focus on analyzing the rides in Boston, MA, United States, and how customers view their Uber experiences in general.

2.1. Overview of Dataset

Two datasets are obtained from Kaggle and were used in this project: "Uber Ride Reviews" (Uber_Ride_Reviews.csv), and "Uber and Lyft Dataset in Boston, MA" (rideshare_kaggle.csv).

Uber Ride Reviews is text data. It contains the reviews that customers submitted and the corresponding ratings they gave after the ride.

The Uber and Lyft Dataset is spatial and time-series data. It contains some operational data of Uber and its competitor - Lyft. It includes the time, origin and destination of each ride, the weather condition and other information with respect to that period of time.

In order to get the specific locations of the origins (pick-up locations / “source”) and destinations, the data is pre-processed to include the latitude and longitude of the pick-up and drop-off locations. Also, to allow users to have more flexibility in visualizing the data in a different time scale, a column of “Weekday” is added, where “1” equals Monday and “7” equals Sunday.

2.2. Limitation of the Datasets

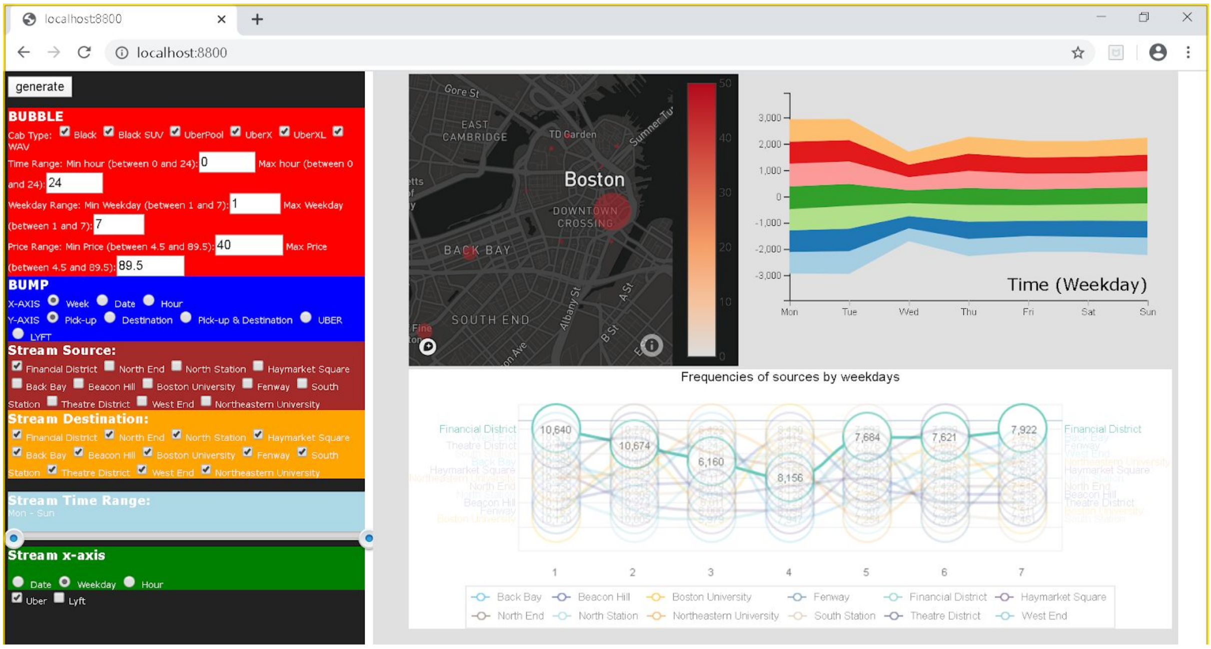
It is believed that the “Uber and Lyft dataset in Boston” was a dataset with some flaws. There are missing data in terms of time-range and locations. Also, we believed that the dataset may have been modified that the total number of rides from different locations and car types are similar. However, the dimensionality of the dataset is very rich. For example, it has weather, price, distance, temperature, windspeed, etc. We still think that there are lots of insights that can be concluded from this dataset, and hence, we decided to use it in our project.

3. Visualizations

There are 7 visualizations prepared for this project in total. We will explain below: the use of data, and the ways we generated the visualizations.

3.1. Dynamic Visualizations

All dynamic visualizations will be put together in the same-page view complying with the concept of “eye beats memory”. Each graph can complement one another by choosing corresponding filters or choices. For example, the Uber “Bubble” cannot show the popularity of each district at a time period. This can be solved by viewing the Uber “Stream” at the same time. Also, Uber “Stream” can show the quantity of each car type in different weekdays but it cannot show the ranking of each car type. This can be solved by looking at the Uber bump simultaneously. The quantity and ranking information can be viewed by users concurrently. The design can be viewed in "dyanamic_designs.html".



localhost:8800

generate

BUBBLE

Cab Type: ☑ Black ☑ Black SUV ☑ UberPool ☑ UberX ☑ UberXL ☑ WAV

Time Range: Min hour (between 0 and 24): 0 Max hour (between 0 and 24): 24

Weekday Range: Min Weekday (between 1 and 7): 1 Max Weekday (between 1 and 7): 7

Price Range: Min Price (between 4.5 and 89.5): 40 Max Price (between 4.5 and 89.5): 89.5

BUMP

X-AXIS ● Week ○ Date ○ Hour

Y-AXIS ● Pick-up ○ Destination ○ Pick-up & Destination ○ UBER ○ LYFT

Stream Source:

☑ Financial District □ North End □ North Station □ Haymarket Square □ Back Bay □ Beacon Hill □ Boston University □ Fenway □ South Station □ Theatre District □ West End □ Northeastern University

Stream Destination:

☑ Financial District ☑ North End ☑ North Station ☑ Haymarket Square ☑ Back Bay ☑ Beacon Hill ☑ Boston University ☑ Fenway ☑ South Station ☑ Theatre District ☑ West End ☑ Northeastern University

Stream Time Range:

Mon - Sun

Stream x-axis

○ Date ● Weekday ○ Hour

☑ Uber □ Lyft

Figure 1: An overview of Financial District

3.1.1. Uber "Stream"

Lodash.js is used to process the data based on users' requirements on "Date", "Weekday" and "Hour", filtering "Source" and "Destination". The Uber "Stream" is constructed by utilizing d3.js.

A stream is used to visualize the relationship between the number of trips and time. It follows the principle of "Overview first, Zoom and Filter, Details on Demand". It shows the stream of Uber and Lyft first. When users specify the operator (by clicking the stream), they can compare different car types of Uber and Lyft respectively. Uber has 7 car types, while Lyft has 6 car types.

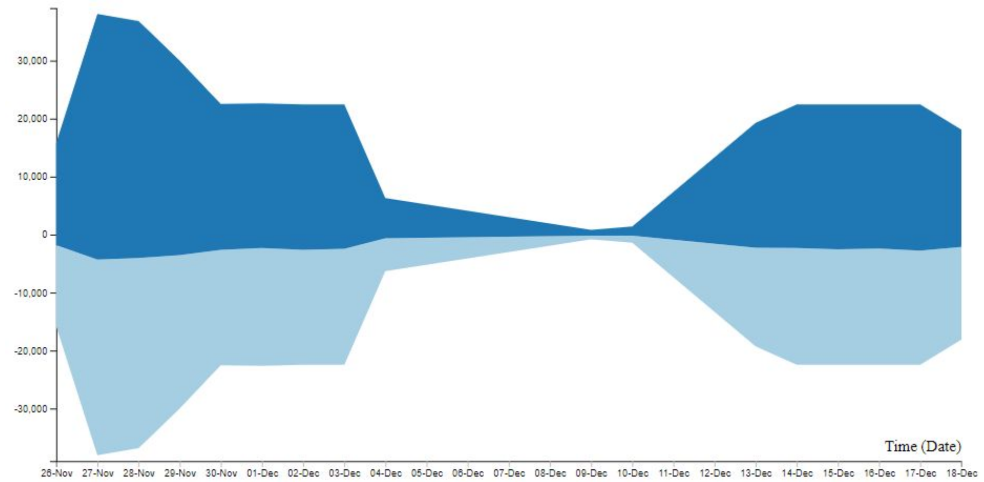


Figure 2: The Overview of Uber "Stream"

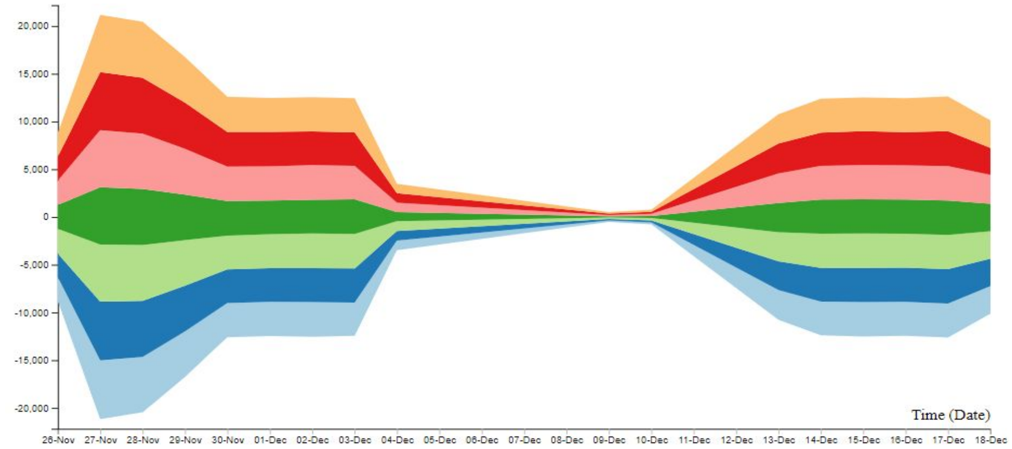
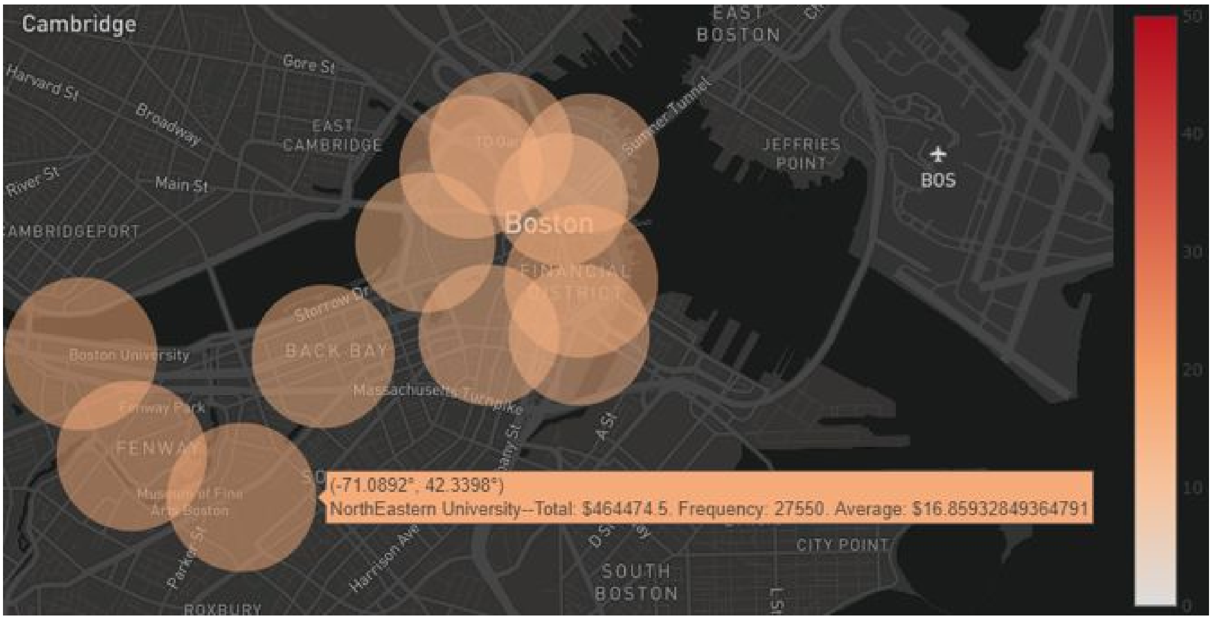


Figure 3: Date View of Uber "Stream" with different car types

3.1.2. Uber “Bubble”

Javascript was used for processing the data based on users’ requirements on “Price”, “Source”, “cab_type”, and filter “Weekday” and “Hour” for time. Plotly mapbox was used for generating a bubble map.

A bubble map is used to visualize the money expenditure of different locations. The size of the circle represents the total money spent in a district and the color of the circle indicates the average money spent in the district. Users can make obvious comparisons between different districts.



(-71.0892°, 42.3398°)
NorthEastern University--Total: $464474.5. Frequency: 27550. Average: $16.85932849364791

Figure 4: Overview of Uber trips in Boston

3.1.3. Uber “Bump”

Javascript was used for processing the data based on users’ requirements. Koolchart library was used for generating a bump chart.

The Uber “Bump” has different choices for the axes listed below:

X-axis:	date, week, hour
Y-axis:	pick-up, destination, pick-up and destination, Uber, Lyft

The choice of X-axis specifies the grouping method of data, i.e. how they are divided and categorized. Y-axis specifies the role of each bump. When it is set to [“pick-up”, “destination” or “pick-up and destination”], each bump represents a district in Boston, while their counts equals the frequency of trips [start / end / sum both start and end] in that district.

Bump charts are good for visualizing the ranking for time series data. Popularity of each district in different time periods (weekdays/dates/hours) can be easily viewed by the Uber “Bump”.

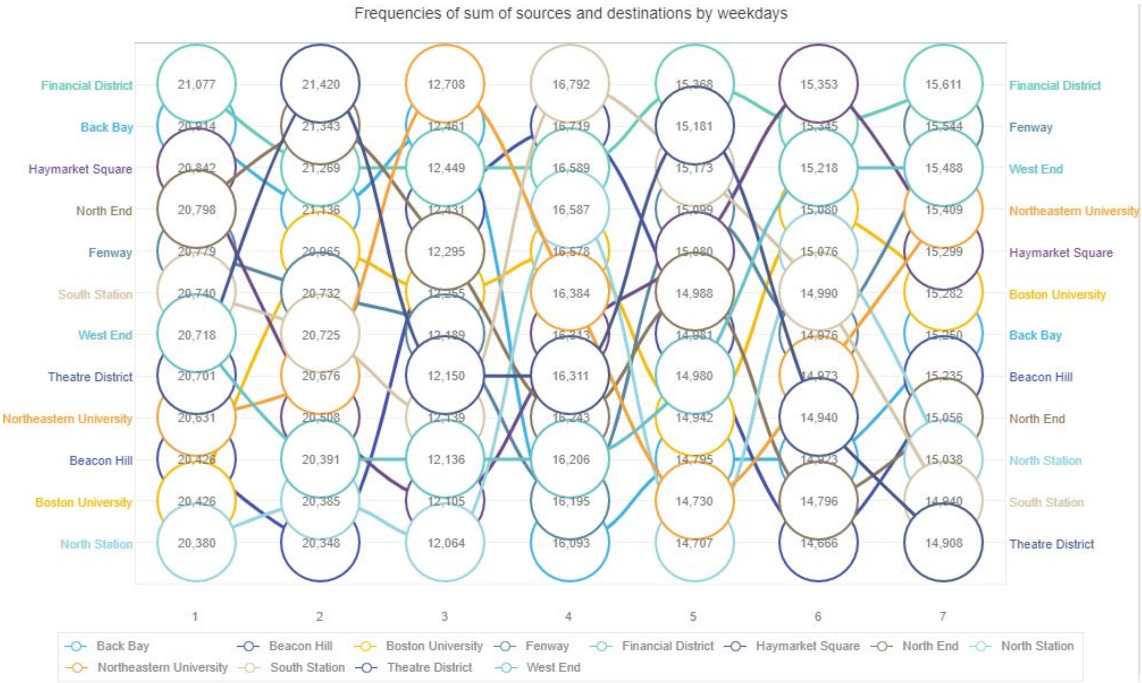


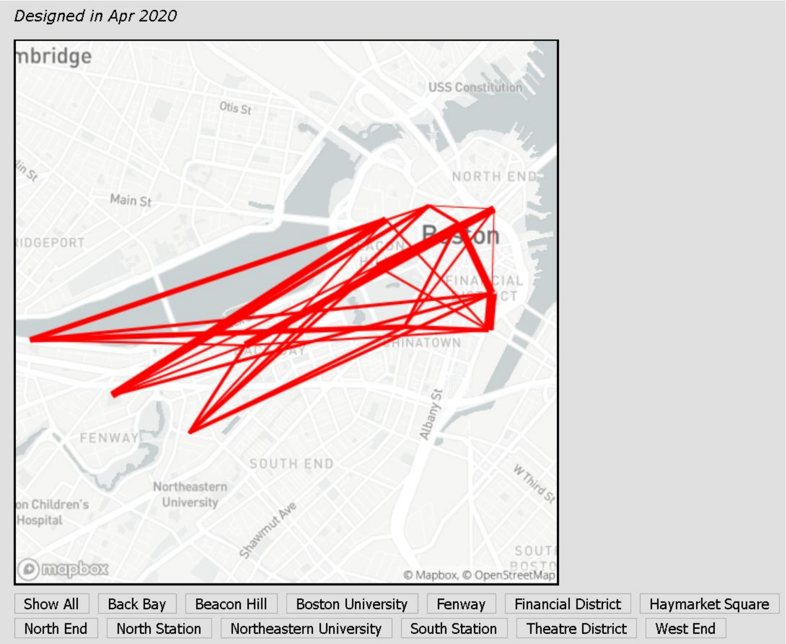
Figure 5: Overview of Uber “Bump”

3.1.4. Uber Map

Python 3 (map_frequency.py) was used for processing the data to obtain the number of trips to and from the "Source" and "Destination". The processed data is then manually transferred to the JavaScript (map.js). Javascript was used for the dynamic part in the map design, which is presented in an html canvas object. This design is not presented together with the other 3 dynamic designs, and it cannot be controlled by the panel. It can be found individually in "whiteboard.html".

In the Python 3 program, the number of all combinations of trips are counted, no matter its direction ("A to B" and "B to A" are considered equivalent). The numbers, stored in "lists", are then manually transferred to a variable in the JavaScript. The "red and gray" filter is written in JavaScript. The widths of the lines are also adjusted in order to generate a better visual effect on the difference.

The design below is used, in order to show the patterns of the trips and the popular locations.

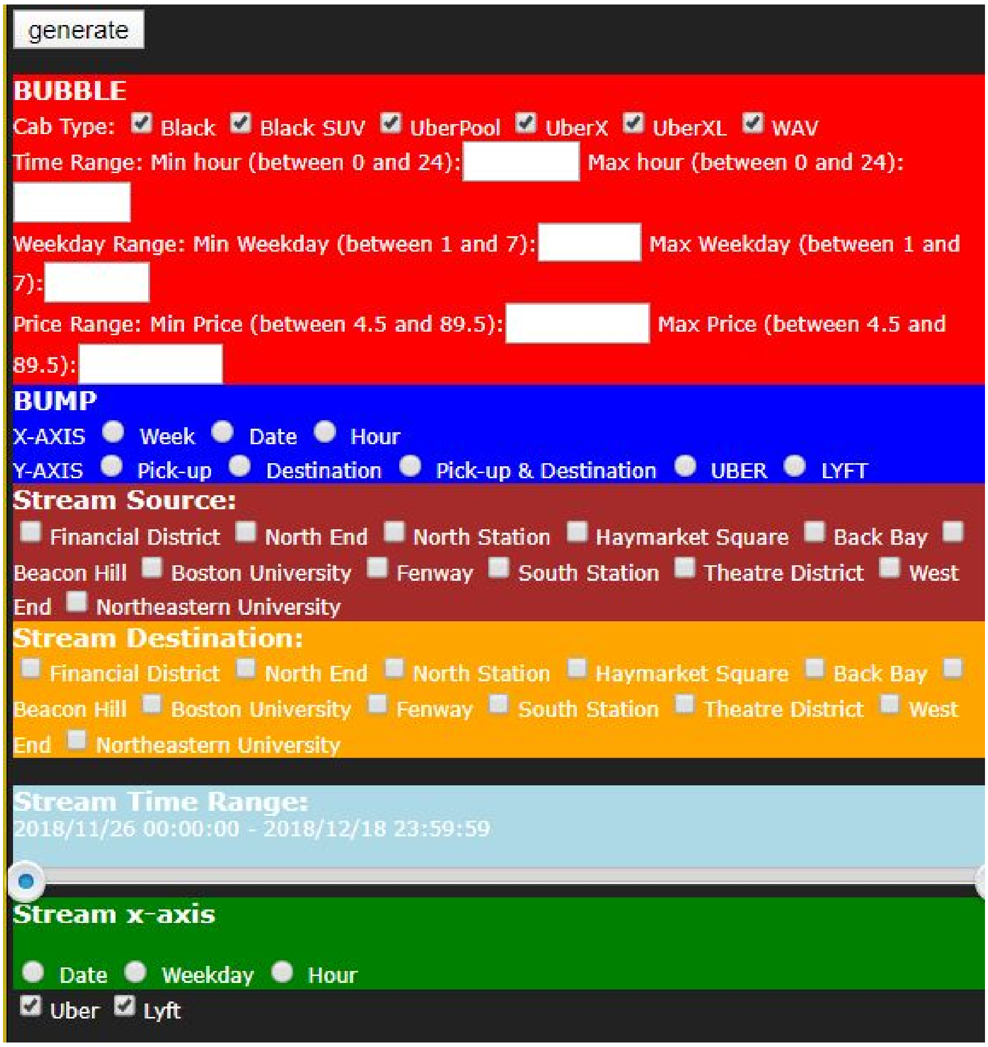


Designed in Apr 2020

Figure 6: Uber Map

3.1.5. Control Panel

There is a control panel allowing users to specify the parameters of the data used for different visualizations. The users can generally select particular car types, time range, as well as the source and destination for all dynamic visualizations. Tooltips will be shown when the mouse hovers on the points inside each of the visualizations.



generate

BUBBLE

Cab Type: ☑ Black ☑ Black SUV ☑ UberPool ☑ UberX ☑ UberXL ☑ WAV

Time Range: Min hour (between 0 and 24): Max hour (between 0 and 24):

Weekday Range: Min Weekday (between 1 and 7): Max Weekday (between 1 and 7):

Price Range: Min Price (between 4.5 and 89.5): Max Price (between 4.5 and 89.5):

BUMP

X-AXIS ○ Week ○ Date ○ Hour

Y-AXIS ○ Pick-up ○ Destination ○ Pick-up & Destination ○ UBER ○ LYFT

Stream Source:

□ Financial District □ North End □ North Station □ Haymarket Square □ Back Bay □ Beacon Hill □ Boston University □ Fenway □ South Station □ Theatre District □ West End □ Northeastern University

Stream Destination:

□ Financial District □ North End □ North Station □ Haymarket Square □ Back Bay □ Beacon Hill □ Boston University □ Fenway □ South Station □ Theatre District □ West End □ Northeastern University

Stream Time Range:

2018/11/26 00:00:00 - 2018/12/18 23:59:59

Stream x-axis

○ Date ○ Weekday ○ Hour

☑ Uber ☑ Lyft

Figure 7: Control Panel of the dynamic visualizations

3.2. Static Visualizations

3.2.1. Word Clouds

We utilized a python 3 program (textana.py) to find frequent words in the reviews, and prune two types of words: stopwords and “common words”.

We first generate five lists of frequent words using Python 3, each corresponds to a rating (1-5) and each list is 30 words long. All stopwords are ignored and will not appear in the 30-word lists. Then, search for all words that appear in 3 or more lists. We define these words as “common words” (intersection) and we prune them away (ie. these words will not appear in any of the lists again). The pruning will occur repetitively (recursively) until there are no words that appear in 3 or more lists. These are the five “final lists” of words, and we export the lists to wordclouds.com for word clouds generation.

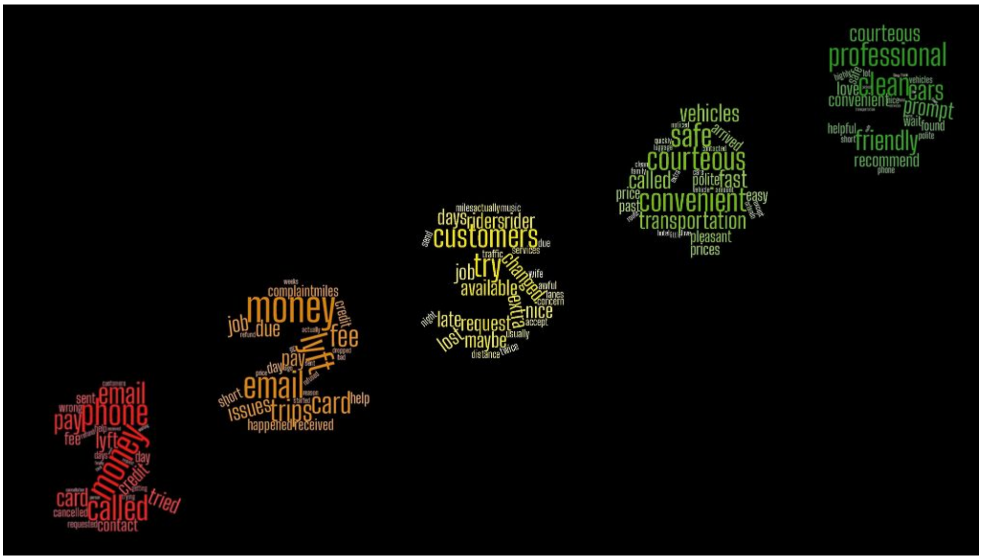


Figure 8: All clouds of frequent words in customer reviews per ratings

Figure 8 visualizes the words in the five “final lists”, and it shows how the words used by customers in the reviews relate to the ratings that they gave. From the visualization, it can be easily inferred that the types of words differ among the ratings. The color hue and shapes encode the rating scale, while the font size and color saturation of each word encode the frequencies.

3.2.2. Humidity-Temperature Scatter

Python 3 (temp-humid-isRain.py), with "matplotlib.pyplot" imported, was utilized to both process and visualize the relationship between "Temperature", "Humidity" and Time ("Date" and "Hour"), in terms of frequency.

Figure 9 shows the frequencies of trips in different hours encoded in color hue. Brown shows the highest frequency level, while green shows the lowest frequency level (order: brown, red, orange, yellow, green). The two axes represent humidity and temperature respectively. The crosses show the hours which are raining (or alternatively in figure 10).

In order to show the distributions of data points and their clusters effectively, we used a scatter to visualize the processed data points.

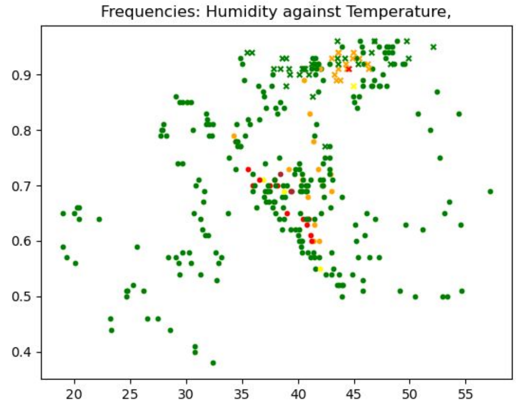


Figure 9: The frequencies of trips per hour, with humidity against temperature. Rainy hours shown in crosses.

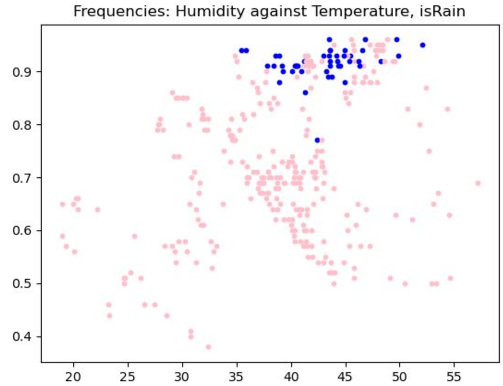


Figure 10: The frequencies of trips per hour, with humidity against temperature. Rainy hours shown in blue.

3.2.3. Price-Distance Scatter

Price and distance (in miles) attributes are used for the scatter plot. We construct this scatter plot through python 3 program (price-distance.py), using “matplotlib”. As this is a 2-dimensional relationship, so we chose a scatter plot for obtaining clear observation.

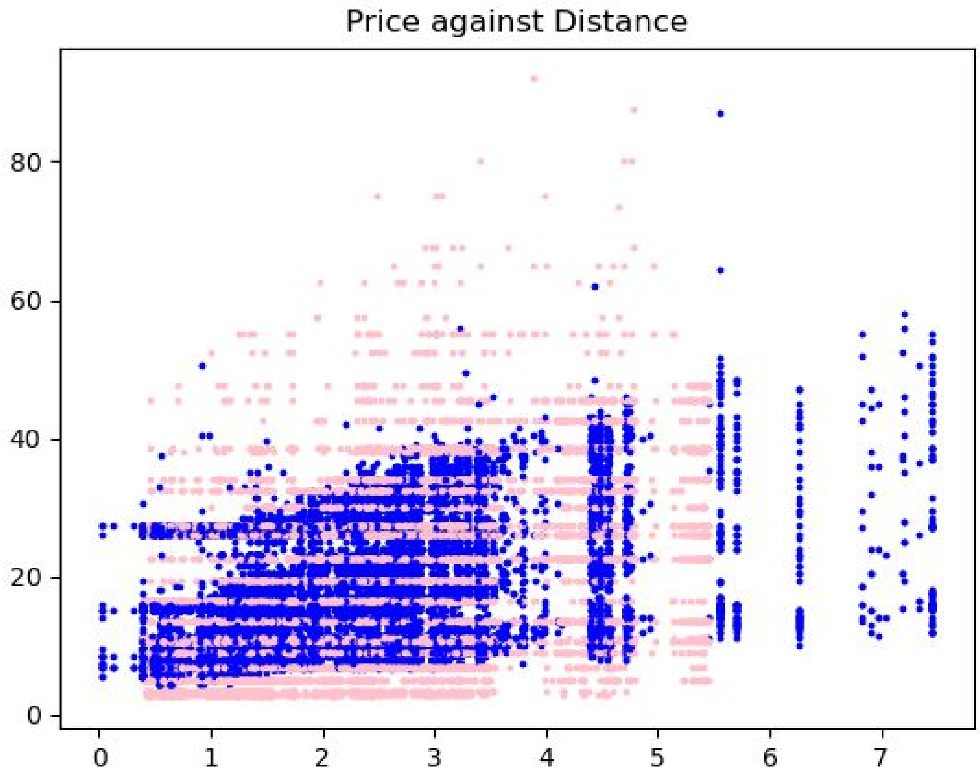


Figure 11: Price-Distance Scatter Plot

4. Tasks and Insights

We are having different tasks to analyze the performance of Uber as well as its competitors Lyft, while finding trends on customer taking Uber.

4.1. Task I: Analysis of Uber Customer Reviews

To better understand the performance of Uber, it is important to look at how customers view Uber. Text analysis is conducted to investigate the rationale behind each rating and how Uber and its drivers can improve their services in the future, in order to enhance their business. Word cloud is used to visualize the results of the text analysis.

To visually see the relationship between the customer ratings and (frequent words used in) their reviews, the dataset "Uber Ride Reviews" (Uber_Ride_Reviews.csv) was used in this task.

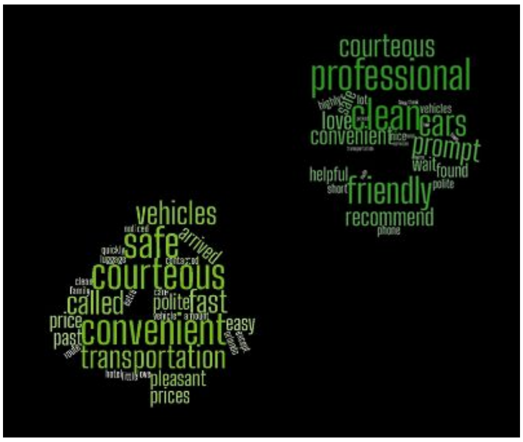


Figure 12: Clouds of positive ratings

When we look at positive ratings (4-5), we can see that customers tend to use positive adjectives in their reviews (figure 12), including "courteous", "prompt", "friendly", "professional", "convenient", etc. These words reveal that customers value both good attitudes of the drivers and the efficiency of the trip. They wanted professional services to be provided to them.

When we look at negative ratings (1-2), we can see that customers tend to state the issues (problems) in their reviews (figure 13), including “money”, “pay”, “email”, “called”, “card”, “Lyft” etc. We conclude that technical and money issues are occurring the most frequently, and dissatisfied customers are the most concerned about them.

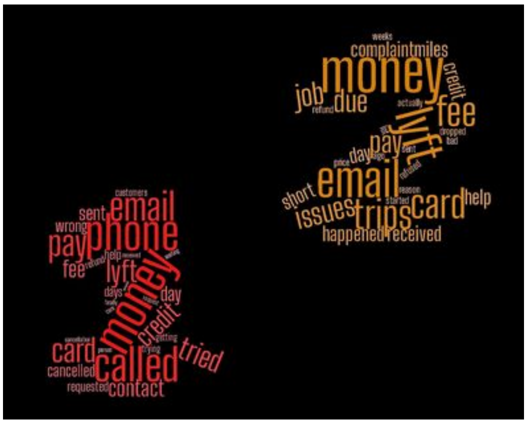


Figure 13: Clouds of negative ratings

We also consider that dissatisfied customers are comparing the services provided by Lyft, and may tend to use Lyft’s services in future trips.

4.2. Task II: Analysis of Uber Trips in Boston

To have a comprehensive view on the Uber trips, we separated it into 5 different sub-tasks: Peak Time, Money Expenditure, Popular Route, Popular locations, and Weather.

4.2.1. Peak Time

In this part, we utilized the control panel to sort data based on different time measures: Date, Weekday, and Hour.

In the “Date” view, (same with figure 3 in section 3.1.1), even with missing data (from Dec 5 to Dec 8) and the limitation of data (similar number of car types), it is observed that the number of trips in the period of Late November and Early December is greater than that in Mid December.

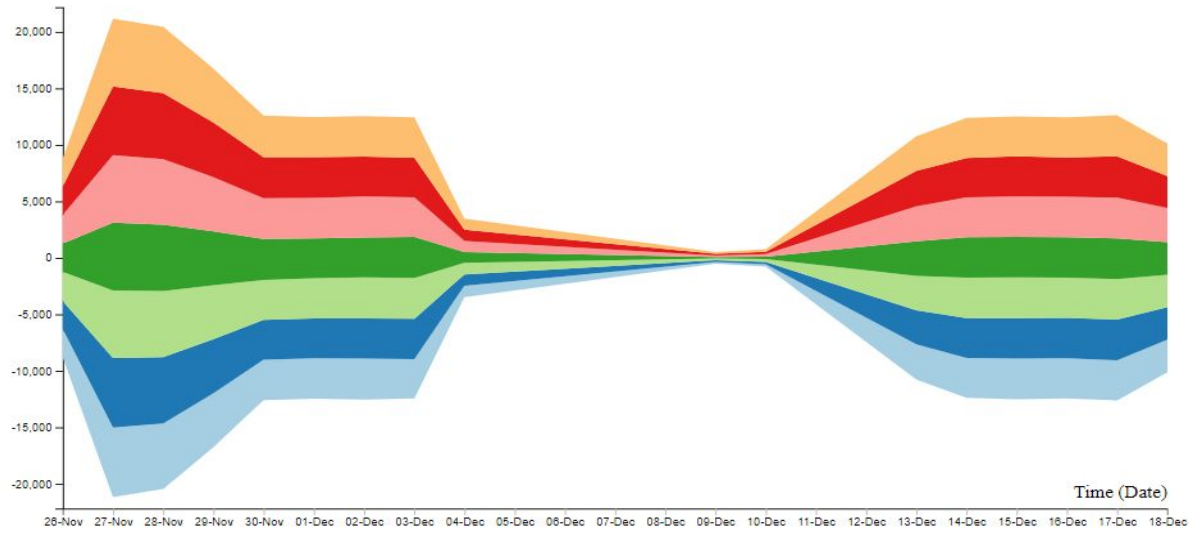


Figure 3: Date View of Uber "Stream" with different car types

In "Weekday" view, even with the missing data (from Dec 5 to Dec 8, Wednesday to Saturday) and the limitation of data (similar numbers of car types), we can still deduce that the number of trips in weekdays (Monday and Tuesday) is greater than that in weekends (Sunday), exempting the period of Wednesday to Saturday.

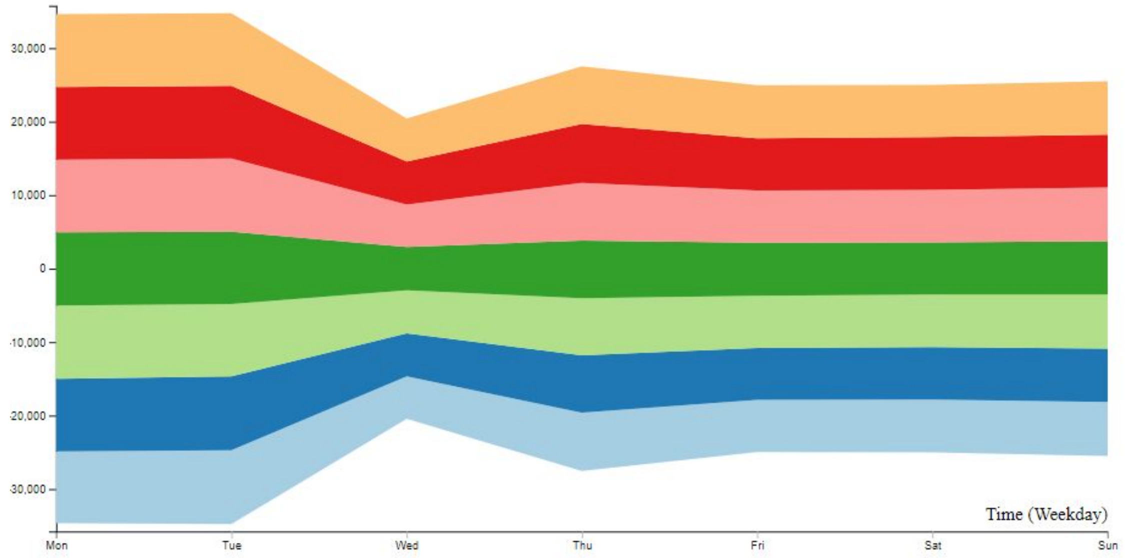


Figure 14: Weekday View of Uber "Stream"

In the "Hour" view, even with the missing data (from Dec 5 to Dec 8) and the limitation of data (similar car types), there will not be a significant impact on the insight, as data of each existing date is complete (in terms of hours). It is observed that the peak hour of the Uber trips is in the afternoon (10 am - 7 pm) and at midnight (11 pm - 1 am).

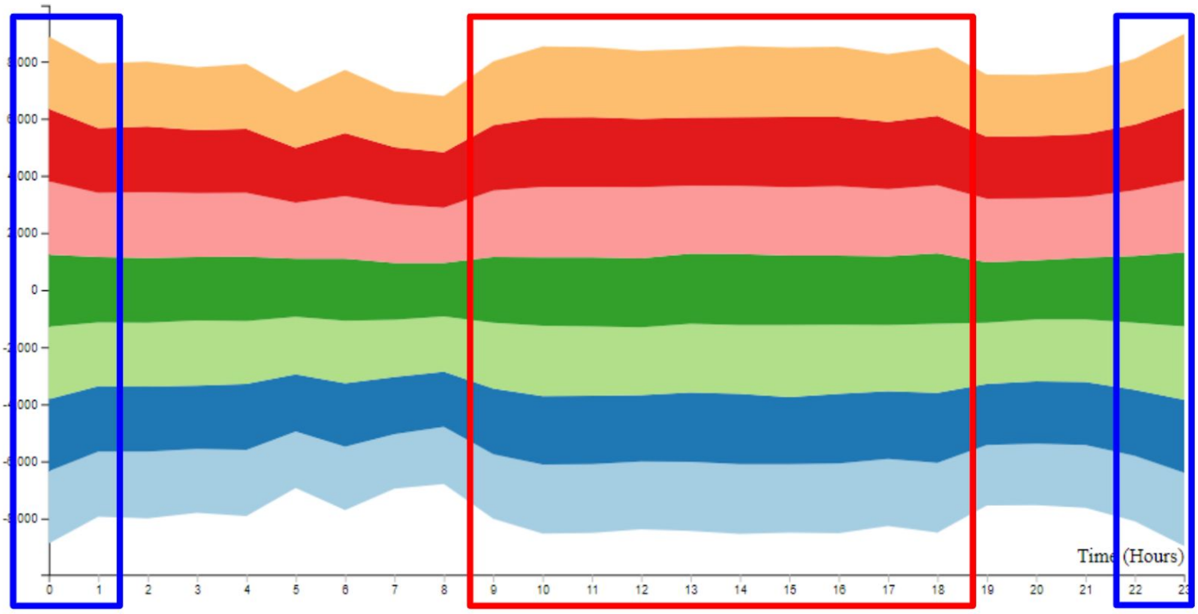


Figure 15: Hour View of Uber "Stream"

Similar analyses are conducted to and from different locations (using different settings in the panel). Yet, similar patterns are observed in different choices of combinations of locations.

We can conclude that the number of trips is dependent on time, but is independent to the location.

4.2.2. Money Expenditure

The money expenditure renders users the option of selecting different price ranges and also different car types for further investigation.

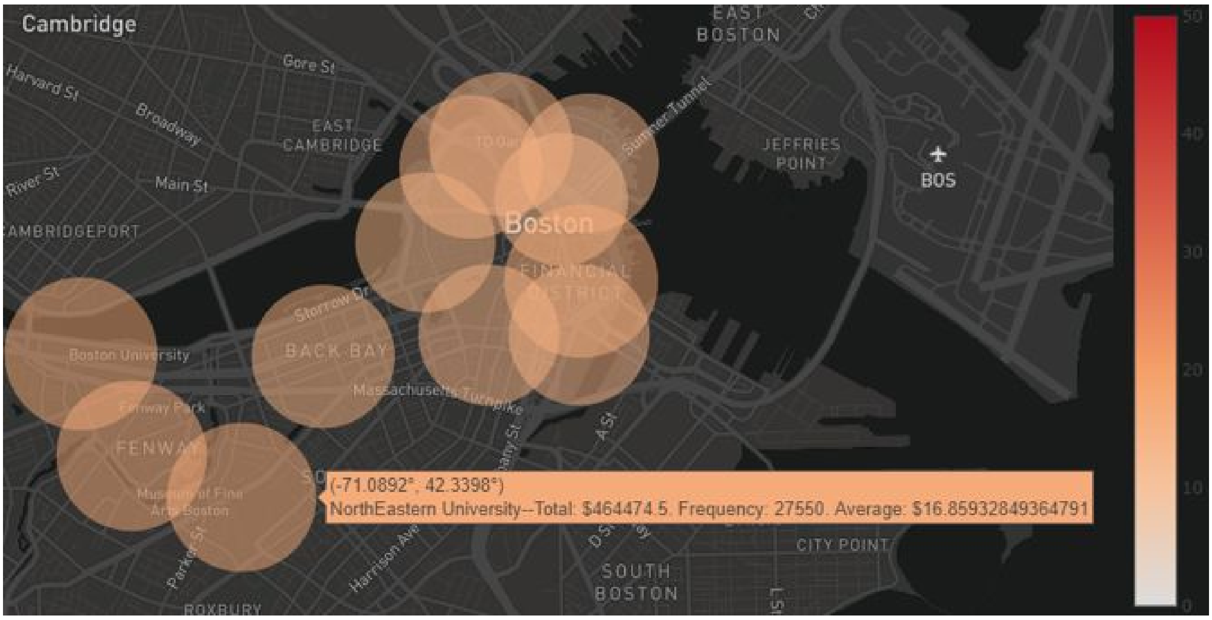


Figure 4: Overview of Uber trips in Boston

Although the limitation of the dataset (similar number of rides in every district and cartype) will affect the size of the bubble, the average money spent on each trip is not affected. It is observed that the bubble colors are very close in every district in different cartypes views (Figure 16). Even though each car type has a different average money spent, the standard deviations of the average money spent in different districts are also very small and close to 0. For example, UberX has the smallest standard deviation among the cartypes, most of the districts' average money spent is $9, only a few districts' average money spent varies around $10-$11. We can see that the average money spent in each trip is independent from the locations. Locations will not affect the expected money spent on each trip.

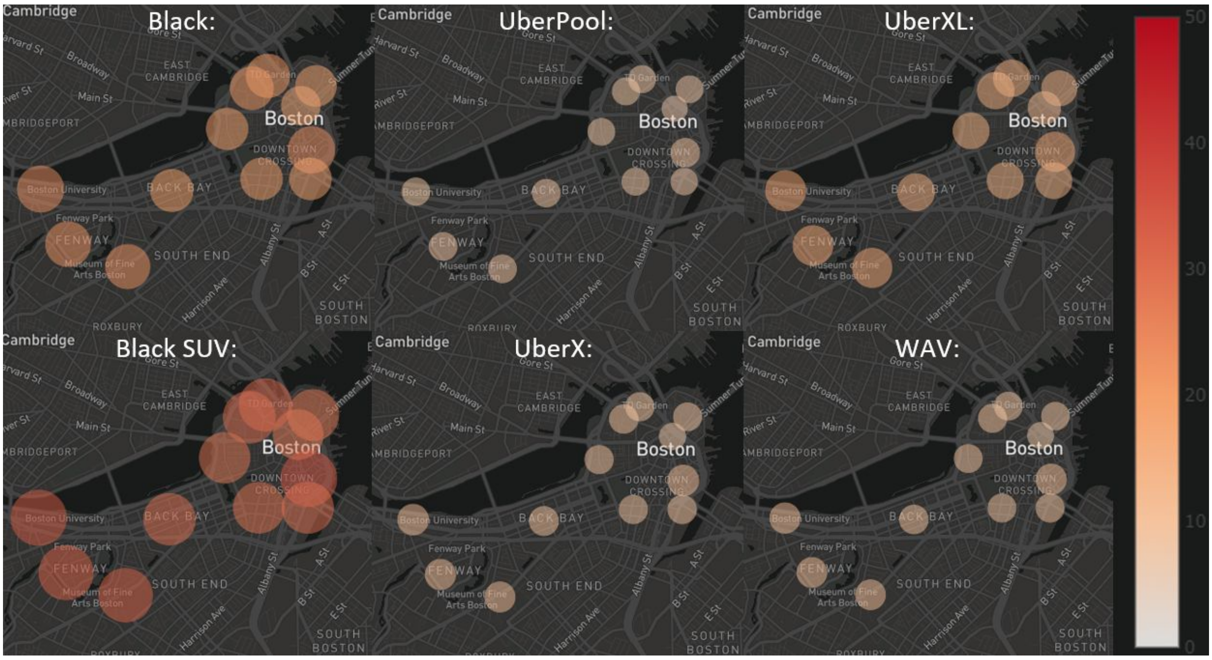


Figure 16: Bubble maps of money spent on different Uber car types

Another interesting findings are found in Figure 17. It is the bubble map of Uber trips in Boston with only trips that spent more than $40. As we can see that there is only 1 bubble that is big in the map which locates in the Financial District. It shows that a large proportion of “expensive trips” occurs in the Financial District and “expensive trips” occurs only frequently in the Financial District.

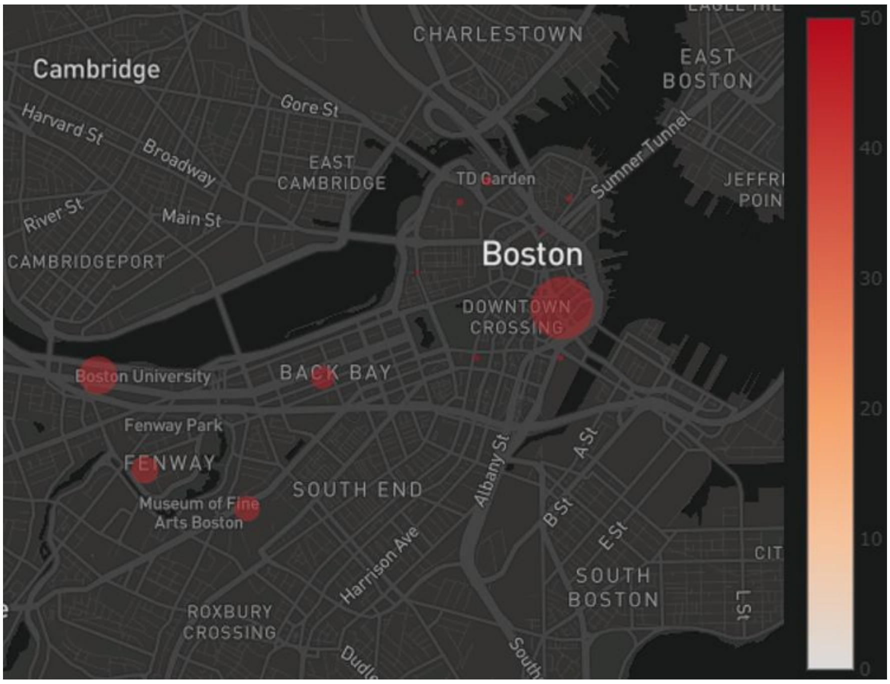


Figure 17: Bubble map of >$40 money spent in Boston

4.2.3. Popular Route

Figure 6 is the visualization of the Uber trips in Boston. The width of the lines encodes the frequency of trips, while the two ends of the lines show the “pick-up location” or “destination” of the trip combinations. It is not to scale, so the representation is more ordinal than cardinal.

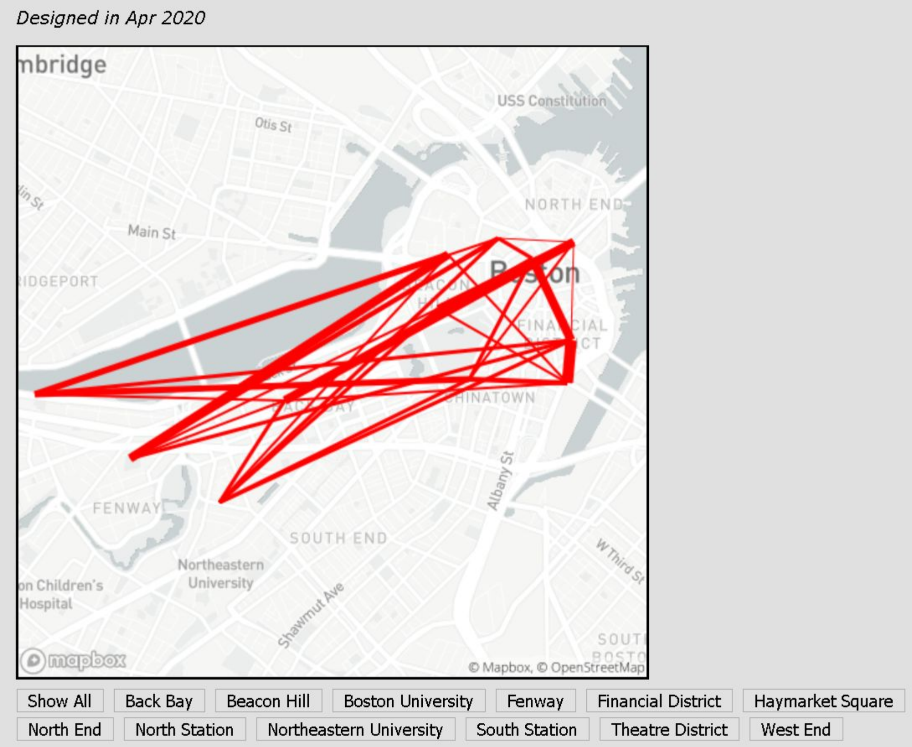


Figure 6: Map of Trips in Boston, widths not to scale

As we can see from the map above, thick lines always occur horizontally, meaning in an east-west direction, and also long-haul comparatively. After researching, the eastern area mainly contains transportation facilities and the financial center, while the western area mainly contains universities and other residential sub-areas. We suggest that Uber customers may possibly travel for the purpose of: long-haul trips going from stations or downtown, to universities or residential areas (or the opposite direction).

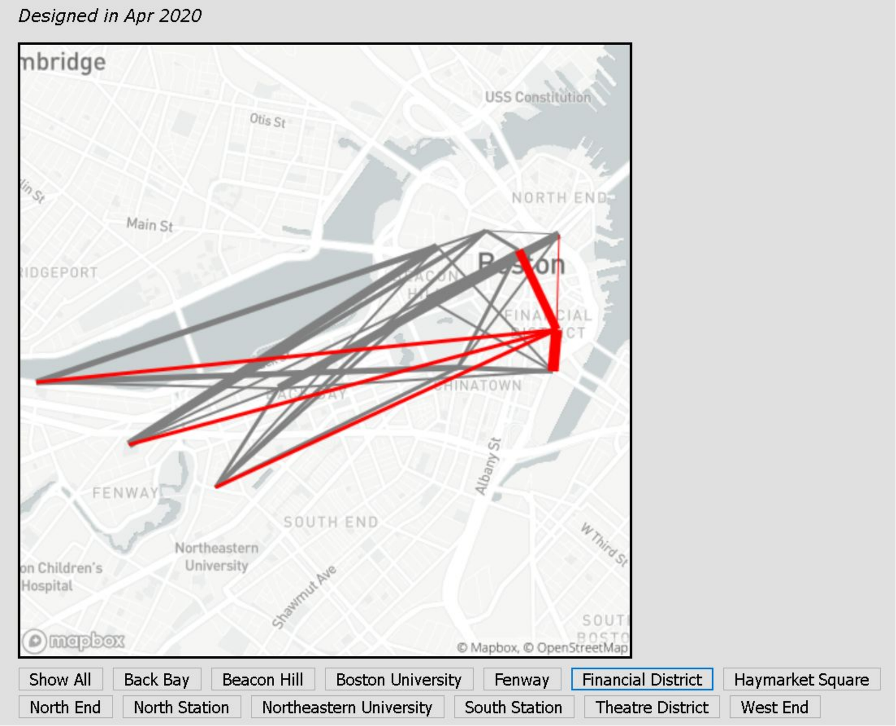


Figure 18: Map of Trips in Boston with filter: Financial District, widths not to scale

Figure 18 shows the map design with a filter "Financial District". The trips to and from the Financial District are shown in red. We observe that these trips are more short-haul (short lines are thicker). This short-haul phenomenon only occurs in the "Financial District" filter. We suggest that trips to and from the Financial District are an exception of the long-haul east-western trend we found above. Uber customers to and from the Financial District are possibly (financial) office workers, in which they have trips and working areas only inside the Financial District.

4.2.4. Popular Locations

Figure 5 shown in section 3.1.3 visualises the popularity of different locations in Boston by weekdays.

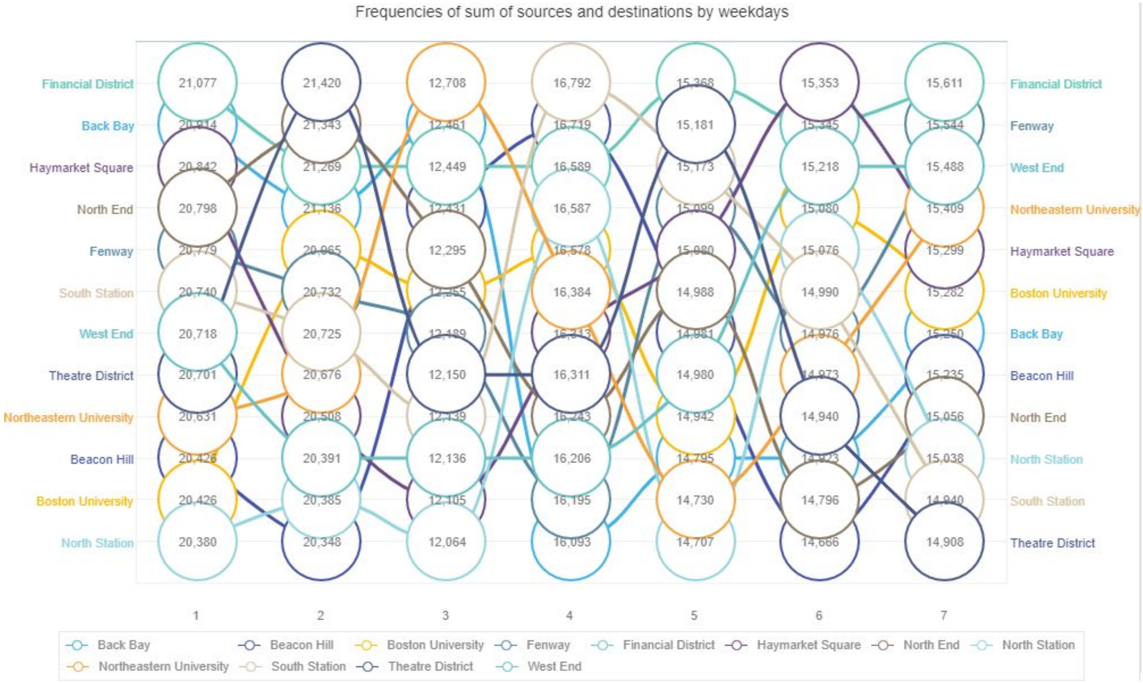


Figure 5: Overview of Uber "Bump"

When we zoom into the bump chart by weekdays (Figure 19). We find that some locations are popular throughout the week, for example the Financial District, while some locations are only popular during weekends like the West End.

When we zoom into the bump chart by hours (Figure 20). We find that no locations are popular in a particular time period. There is no clear dependence on hours. However, some destinations have comparatively higher rank throughout.

We can conclude that popularity of locations mainly depends on weekdays, while the Financial District is a popular location throughout the week. Also popularity of locations do not depend observably on hours in a day, the rank majorly fluctuates within a day.

However, the above conclusion is relatively less well-founded. The numbers in each bump are close possibly because of the limitation of the dataset. For example, in Figure 5, the highest rank bump in Monday has total 21077 rides while the lowest rank bump has total 20380. The difference of lowest and highest rank is not significant.

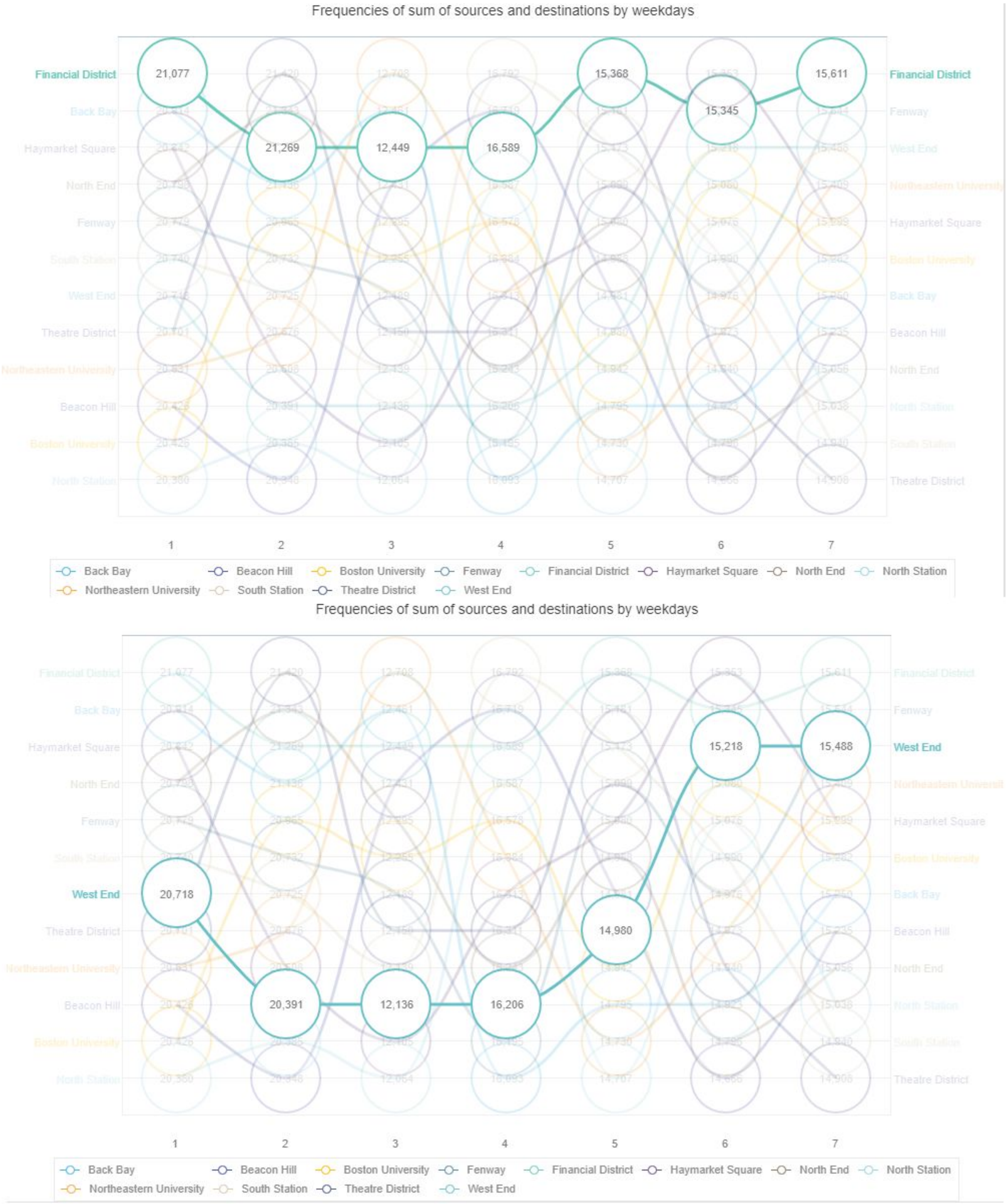


Figure 19: Bump chart of the popularity of in the Financial District and West End (X-axis:weekdays, Y-axis:Locations - Pick-up and Destination)

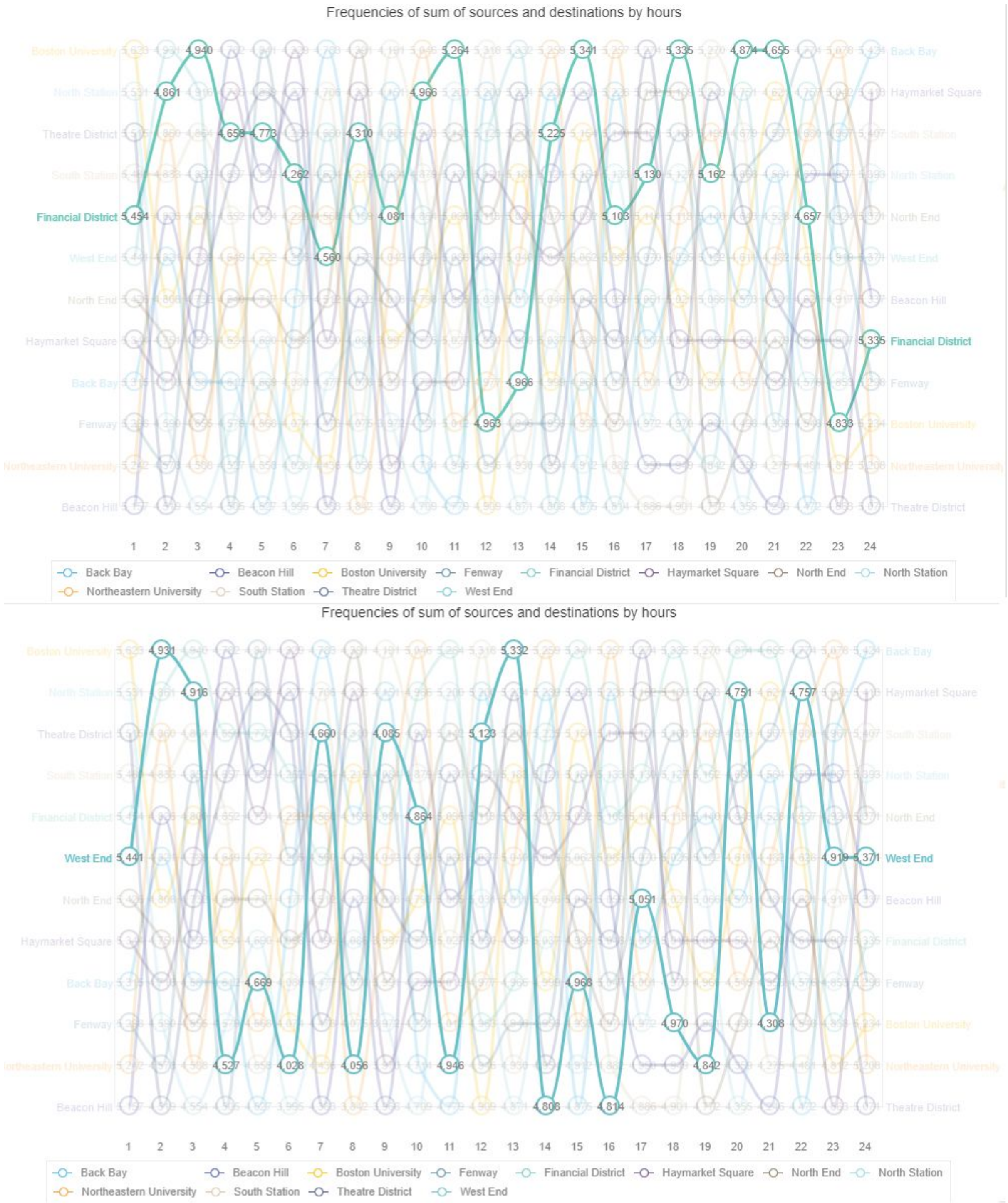


Figure 20: Bump chart of the popularity of in the Financial District and West End (X-axis: hours, Y-axis: Locations - Pick-up and Destination)

4.2.5. Weather

In this task, we attempted to analyze the effect of weather on Uber ridership.

From figure 9 in section 3.2.2, we observe that there are two clusters, one located near the top, and are mainly “crosses”; while the other one located near the center of the scatter. This can be inferred that frequent trips occur on two occasions: rainy hours (with high humidity), and medium temperature and humidity (abbreviated mTH). We suggest that, during rainy hours, more people are using Uber because walking or talking public transport is comparatively inconvenient in rainy days. Also, we suggest that people tend not to go out under 32 degrees Fahrenheit (0 degrees Celsius) because there may be snow or hail, or it is just simply too cold for external activities. We can see that the mTH occurs just above 32 degrees Fahrenheit.

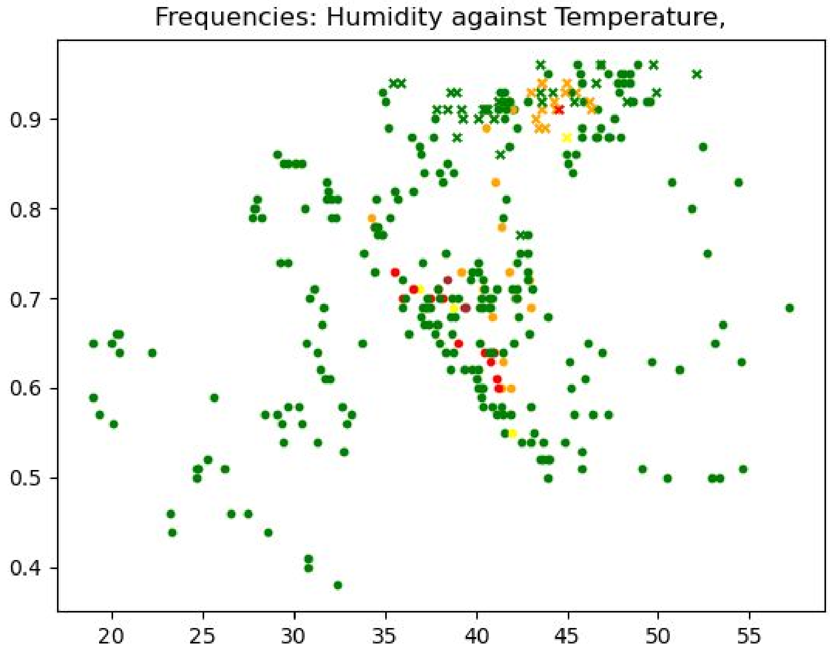


Figure 9: The frequencies of trips per hour, with humidity against temperature. Rainy hours shown in crosses.

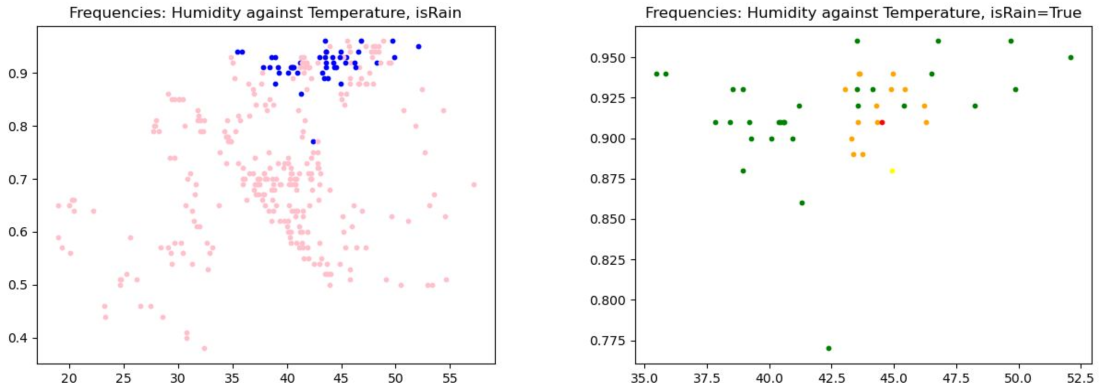


Figure 21: [left] Rainy hours shown in blue,
[right] Rainy hours (blue dots of left) are shown

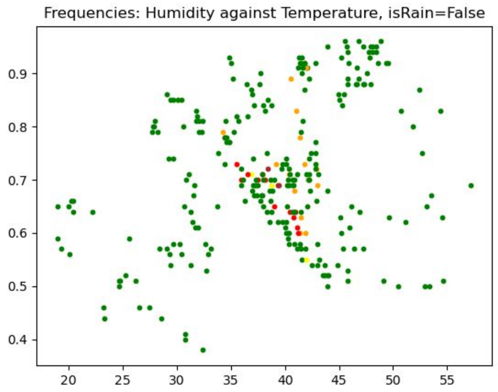


Figure 22: Non-rainy hours are shown

As figure 21 shows, about one-third to half of the dots “are orange or above”, compared to about one-fifth in non-rainy hours (mTH in figure 22). That infers that Uber drivers have a higher chance to get customers during the rainy days.

4.3. Task III: Battle Between Uber and Lyft in Boston

Both Uber and Lyft operate in Boston, MA. They compete with each other to expand their market share. In this task, we will focus on the price attribute to see which operator will win in the battle.

A scatter plot is prepared by Python and is used to visualize the relationship between price and distance (in miles). The x-axis is the distance (in miles) and the y-axis is the price. Different colors are used to encode the operators: blue represents Uber, while pink represents Lyft.

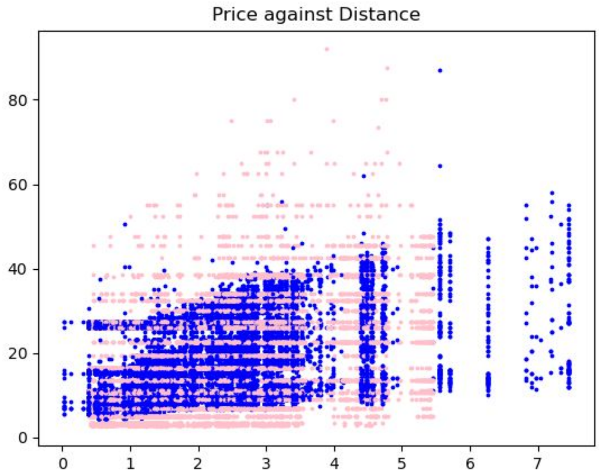


Figure 10: Price-Distance Scatter Plot

From the figure 10 appeared in Section 3.2.3, it is observed that Lyft is generally having lower minimum charges regardless of distance, but a wider range of fare rate than those in Uber.

Although Lyft has a lower price than Uber, Uber is still the priority of customers' choices. It is believed that customers value the professional services provided by the drivers more, while price is not the primary factor.

5. Conclusion

Uber is one of the largest for-hire vehicle operators in the world. Its trips can be affected by various factors. In general, customers are satisfied with drivers' attitude, professional and prompt services, and avoiding technical and monetary issues. Narrowing down to Boston, trips tend to appear frequently at noon, in the Financial District and during rainy hours. Expensive rides tend to exist in the Financial District as well. Compared with Lyft, Uber has a higher minimum charge but a lower fare rate. Only when Uber fully understands the customers' behavior and magnifies its edge over Lyft, can Uber thrive and make more profits in this era.

6. Future Improvements

A more detailed data collection can be proceeded in the future, to obtain a more structured and complete dataset. The visualization used in this project can well adapt to the detailed dataset and reveal more insightful and well-founded trends.

7. References

Dataset: Uber Ride Reviews
https://www.kaggle.com/purvank/uber-rider-reviews-dataset

Dataset: Uber and Lyft Dataset about Boston
https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma