Predicting Cab Prices in Boston against Weather

Xu Zhou, Yuhan Wang, Oum Parikh, Valerie Robert

Northeastern University – DS4200 Information Visualization

ABSTRACT

Uber and Lyft have changed the face of taxi ridership, making it more convenient and comfortable for riders. But there are times when customers are left unsatisfied because of a shortage of vehicles which ultimately led to adopting surge pricing. It's a very difficult task to forecast the price ridership at different locations in a city at different points in time. This gets more complicated with changes in weather. In this project, we attempt to estimate the price of trips in Boston. We add an exogenous factor, weather, to this analysis to see how it impacts the changes in price of trips. We fetched data from Uber and Lyft rides in Boston from Kaggle. We also gathered weather data for Boston in the same time period. In this project, we attempted to analyze cab data and weather data together to estimate the price of trips per borough due to the changing weather conditions. Multiple models were built to predict the price due to the variations in the weather.

Keywords: cab, weather, price.

1 INTRODUCTION

This visualization uses uber and lyft data to analyze the impact of weather factors for the prime areas in boston. Considering the uncertain rains in Boston, the visualization tries to model how rain impacts the prices in different areas. We hypothesized that changes in weather would affect the price of Uber and Lyft rides. Location and prices are the 2 most important factors to the cab user because those are the ones that impact their decision to ride which is why we have decided to use those to create a tool that links both of those factors. We are adhering to a very broad user base where anyone who uses lyft or uber would find it helpful. The dataset compiled for this project serves as a foundation for additional research. Analyzing at least a year's worth of data will bring further insights. Demand can be more accurately predicted if the actual price of rides requested information is available along with the price of live rides.

2 RELATED WORK

2.1 First Related Work

The first related work explores and visualizes the spatial effects and patterns in ride-sourcing trip demand and characteristics [1]. This work focused on analyzing the complexity of the demand pattern of ride-sourcing and how this is a challenge for transportation modeling practitioners. It is a challenge for them due to how dynamic the ride-sourcing system is.

This work had several interesting visualizations that we could definitely extend to be useful for our use cases. The work included visualizations pointing to the Median Transportation Network Companies Trip LDPC at Pick-up and drop-off Census Tracts. The work also has a visualization showing the city of Chicago's empirical spatial mean for daily pick-ups and drop-offs in 2019. Both of these visualizations provide an example on how we can use maps for our visualization to showcase numbers of trips within a city based of duration.

2.2 Subsection Two

This second paper looks at how temperature affects transportation network companies in the city of Toronto [2]. This research was based on historical data of only Uber trips from September 2016 to 2018.

This work had several intriguing visualizations that we can draw inspiration from as we create the visualization for our project. One visualization highlights the Trip generation in both 2018 and 2017. This visualization does a comparison. This type of visualization can be useful for the use cases we have. We can compare the ways in which weather has effect different types of Ubers and Lyfts. Another visualization shows a variation Transportation Network Companies in 2017 and 2018. This type of visualization is a map which helps identify and visualize the wide variation of TNCs. This type of visualization would be especially useful as the data we have divides the city of Boston up into different neighborhoods.

3 USE CASE

3.1 Subsection One

The visualization tool would provide complete picture which would allow the user to use the best cab app and schedule their cabs on the best time given the weather. First, the user would see a simple visualization which shows the 12 most popular pick-up and drop off areas in Boston. This would help them understand the relative frequency of their trip. There are 2 paths that the user could choose next. They could either select their pickup and drop off location, after which they would further be prompted with 2 parameters: rain or temperature. After they make a choice, 2 visualizations, one for uber and the other for Lyft. Both visualizations show the respective prices as a function of the parameter that they chose. They could then compare to evaluate what cab they would choose for a particular time or what should be saving them the most money.

If they want a more generalized visualization which does not take areas into consideration, they could choose another path, where they would be asked to choose either temperature or rain. After choosing, the tool would display 2 visualizations, but instead of total prices, it would show price per km for both apps.

3.2 Subsection Two

The visualization would support several domain tasks which would allow the user to get all the necessary information they need.

3.2.1 Subsection One

The visualization would quantify the amount of trip taken to and from all the popular areas in Boston. For this task, we plan to have

2 bar charts, one for all the pickup locations and the other for the drop off location. This would give generic information about frequencies of cab rides in different areas to check if their pickup and drop off location is among popular ones. This would segregate the 2 pathways that the user could choose 2 gain specific information of the weather, since, if their locations are not present, they would have to choose to use the prices per km only.

3.2.2 Subsection Two

The tool would calculate the price per km for different areas for different weather conditions depending on the location that is chosen by the user. We might do a scatter plot with points with distinct colors showing all the weather conditions. From this we would extrapolate a general trend which would be used by the user to evaluate how much difference it would make to travel in their current condition.

4 Data

4.1 Assess and Explore Data

Data linked here.

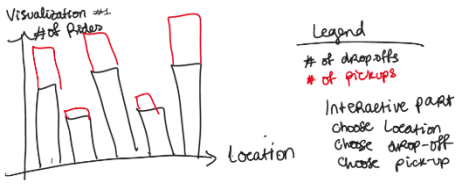
The data chosen for this project is simulating Uber and Lyft rides using real prices. These real prices were determined by considering if someone were to take a ride on Uber or Lyft it would cost that price. Neither Uber nor Lyft do not make their data public. This cab ride data was collected by using Uber & Lyft API queries and corresponding weather conditions. The data was collected for a singular week in the month of November in 2018. This data attributes include distance (distance between rider and destination), the type of cab (Uber or Lyft), time stamp (when the data was queried), destination, source (starting point of the ride), price (in USD), surge (multiplier by which price was increased), id, product id (uber/lyft identifier for the type of cab), name (type of cab). The weather data was collected every hour. The weather data has attributes including temperature, location, clouds, pressure, rain, the time, humidity, and wind.

There were no explicit biases in the collection of the data. The only possible data would be in regard to the locations (neighborhoods) that were chosen in Boston. We cleaned the data provided, by dropping any null values in the cab-ride data. We converted all null values in the weather data to 0 since they represented raindrop levels. We also added a new column “date” which converted Unix time stamps in both the carbide dataset and weather dataset. We did not find any outliers nor inconsistencies in both datasets.

5 Design Process

5.1 Three rough sketches

Visualization #1



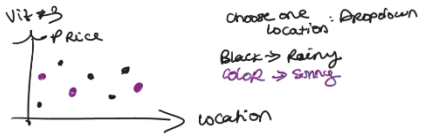
This visualization is a stacked bar chart. X-axis represents the location and Y-axis represents the number of rides. The black part of the visualization highlights the number of drop offs, and the red part represents the number of pickups for one location. This graph is interactive since the user can choose a location, drop-off, pick-up.

The visualization would quantify the number of trips taken to and from all the popular areas in Boston.

Visualization #2

This visualization is a bar chart. X-axis represents the location and Y-axis represents rain level. This graph is interactive since the user can choose to view information at a specific time.

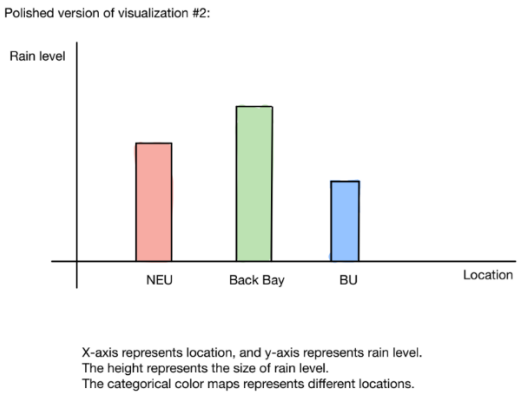
Visualization #3



This visualization is a scatter plot. X-axis represents the location and Y-axis represents prices. The black part of the visualization highlights when it was raining, and the purple part represents when it was sunny. This graph is interactive since the user can choose a specific location.

The visualization tool would calculate the price per km for different areas for different weather conditions depending on the location that is chosen by the user.

5.2 Polished visualization



This polished version of the second sketch remains the same for axis: x-axis for location, y-axis for level of rains in the last hour.

The graph is interactive because users can see the rain at what location at a certain time. The example shows the rain level at Northeastern University, Back Bay, and Boston University at a certain time.

Marks:

- Area: height to represent the size of the rain levels

Channels:

- Horizontal position: various positions represent different physical locations
- Color: different hues of color represent physical locations
- Size of area: the larger the area of the bar, the higher the rain level

5.3 Explanation of final Sketch

Our visualization evolved several times throughout this process. Our visualization changed into only having two visualizations one boxplot and one scatterplot. We originally wanted to also include an interactive map of Boston where a user could select a location to view similar information that our current visualizations show. We also wanted to include a timeline graph, to highlight how the price varies based on time for a specific location.

Throughout the implementation process we encountered some difficulty with the initial dataset that we had found. This dataset was composed of two separate sub datasets (data on weather and data on cab prices). This posed some difficulty when it came time to link the visualizations we had. We also had difficulty creating a map with all of the location (both source and destination) due to how the location data was presented.

Thus, we searched and found a new dataset that was simpler to manipulate. We truncated and cleaned the data set by removing null values and from there were able to create two simpler visualizations: a scatterplot and boxplot.

In terms of usability testing, the user had an understanding of both visualizations. They recommended that we switched the placement of the two explanatory paragraphs which we did. The user also recommended that we add the units to the tooltip box.

6 Final Design

This tool is composed of two visualizations. The first visualization is a Boxplot highlighting the rain level at a given location. The second visualization is a scatterplot that displays the value of the price and rain level for a given location. This given location is selected in the first visualization.

With the first visualization a user can select the source location they are located at. After selecting the source location, they are at, the user can analyze the rain level distribution at this specific location. Then the User can look at the second visualization to view by how much their ride would surge for a give rain level at the selected location (selected location is from first visualization).

7 Discussion

The visualization tool currently links weather to location while also modeling how rain impacts price for each location. The first visualization would show rain pattern for all the locations, while the other would be linked to the first one. By clicking bars on the first one, the user could see how rain affects prices for that place. The visualization, thus, addresses 2 of the most key factors that are affecting the price. Even if these factors are the most useful ones, the data was gathered 3 years ago which makes it outdated since pricing algorithms would have changed by now. However, 2022 datasets did not have all the required features which is why we chose a past dataset, giving more importance to rain data and locations instead of time since it was collected.

As users of these apps, we also know that price and surge multipliers are affected by many more factors like time. However, the data set did not have explicit time value and instead they were timestamps which needed to be processed. Even if we thought time was a key factor, the visualization was limited to modeling 1 factor since putting 2 factors on the same graph with the price would require 3 axis and hence, a 3-dimensional graph. We felt like rain was more important which is why we chose that over time.

In the future, we aim to make the tool more comprehensive by adding more visualizations that model different factors like time and distance. These would be used to calculate price per km for various levels of rain and time to analyze what time would be the best value for travelling from a certain location to the other.

8 Conclusion

This visualization gives an overview of the rain levels over a period at locations in Boston and the prices for specific locations at these rain levels. This would help the user decide the cheapest locations to ride to and rain levels for those places. Even though it gives information about crucial factors, there is some scope of improvement to make the model more comprehensive and help users of this visualization make more real time decisions.

Everyone put in their best efforts to turn in the milestones of the project on time. Yuhan and Valerie made the visualization containing weather at various locations while Bill made the 2nd visualization and linked it with Yuhan help. Oum helped in setting up the webpage and in descriptive parts of the webpage while also trying to help in debugging. The report was a collective effort by everyone.

References

[1] Kelleny Bishoy, Ishak Sherif. - Exploring and visualizing spatial effects and patterns in ride-sourcing trip demand and characteristics. In *Journal of Sustainable Development of Transport and Logistics,* volume 6, pages 6-24.2021

[2] Md Sami Hasnine, Jason Hawkins, Khandker Nurul Habib. Effects of built environment and weather on demands for transportation network company trips. In *Transportation*

Research Part A: Policy and Practice, volume 150, Pages 171-185.2021.

APPENDIX

Data Abstraction

Cab-Ride Dataset – Abstraction

1. Row Data Type – Items
2. Column Data Type – Attributes
    a. Id
        i. Categorical
    b. Name
        i. Ordinal
    c. Distance
        i. Quantitative
    d. Cab_type
        i. Categorical
    e. Time_Stamp
        i. Quantitative
    f. Destination
        i. Quantitative
    g. Source
        i. Categorical
    h. Price
        i. Quantitative
    i. Surge_multiplier
        i. Ordered
    j. Product_id
        i. Categorical
    k. Date
        i. Ordinal

Weather Dataset – Abstraction

1. Row Data type – Items
2. Column Data Type – Attributes
    a. Temperature
        i. Ordinal
    b. Location
        i. Ordinal
    c. Clouds
        i. Ordinal
    d. Pressure
        i. Ordinal
    e. Rain
        i. Ordinal
    f. Time_stamp
        i. Ordinal
    g. Humidity
        i. Ordinal
    h. Wind
        i. Ordinal
    i. Date
        i. Ordinal

Task Abstraction

Task 1: The visualization would quantify the number of trips taken to and from all the popular areas in Boston.
Actions:
1. High: Consume - Present
2. Medium: Lookup
3. Low: Summarize

Task 2: The visualization tool would calculate the price per km for different areas for different weather conditions depending on the location that is chosen by the user.
Actions:
1. High: Consume - Discover
2. Medium: Lookup
3. Low: Compare