Analyzing Factors of Uber and Lyft Pricing

Ivan Jutamulia
MIT Course 6-3 Undergraduate
ivanj@mit.edu

Jack Snowdon
MIT Course 6-14 Undergraduate
jsnowdon@mit.edu

Patricia Lu
MIT Course 6-2 Undergraduate
pjlu@mit.edu

Zach Roberts
MIT Course 1 Undergraduate
zroberts@mit.edu

IDS.012 Final Project - December 11, 2019

Abstract

Over the past few years, we have seen Uber and Lyft overtake the ride-sharing industry. Uber and Lyft thrive on convenience, giving their customers reasonable prices for a ride to wherever they need to go within a matter of minutes, on demand. Each company's prices fluctuate based on a variety of parameters, and clearly their pricing algorithms are different. We seek to identify which factors most influence the pricing of Uber and Lyft trips, considering variables such as distance, weather, time of day, traffic etc. We use a variety of techniques to evaluate and quantify how much of an effect each of these potential factors has on price.

CONTENTS

I. Introduction

With the recent surge in popularity of using ride-sharing with companies such as Uber and Lyft, we are seeing more and more people resort to the convenience that it offers, willing to pay good money for the service. Anyone who has used one of these apps before knows that the price is not always the same, and will fluctuate largely on a per ride basis. But it isn't necessarily obvious how these prices are determined, or what parameters they might depend on.

As consumers, it is in our best interest to understand the underpinnings of why services or products are priced the way they are. This is especially relevant when we consider the situation of Uber and Lyft, as they basically operate as a duopoly in the ride-share industry. As such, the companies have a lot of pricing power, and we don't know exactly how they are doing this pricing. Furthermore, this understanding of pricing will also give us as consumers a better idea of what conditions result in lower prices, and when it might be worth to use ride-share versus not.

While it may be extremely difficult, if not impossible, to fully recreate Uber and Lyft's pricing model, it is well within reason to figure out what kinds of variables affect how they price each individual ride. Though this doesn't give us their entire method of pricing, doing this will at least allow us to gain a better understanding of what may constitute expensive rides or cheap rides, and will paint a clearer picture for ride-share consumers.

II. Research Question

In thinking about this problem, we proposed the following research question that this paper will focus on addressing: *what kinds of factors affect pricing for Uber/Lyft ride-share the most?*

More specifically, we hypothesized and tested a few potential variables that we thought could have an effect on price. Namely, we tested distance, various weather factors, time of day, destination location, and to a lesser extent traffic conditions, using a variety of statistical analysis techniques to determine if they actually do have relationships to price. We chose these specific variables to test because we believe intuitively that they have a strong correlation to demand, which in turn relates to price. Also, they are variables that are within the scope of our data to test.

To address this research question, we plan on doing the following. For the first couple weeks, we intend on doing some more background research behind Uber and Lyft pricing, and see if any related work has been done to analyze them. Furthermore, we want to spend this time to gather whatever data we will need, including ride data, weather data, traffic data, etc. Once we have this data, we will spend some time cleaning it and preprocessing it, perhaps needing to merge datasets together, before doing some preliminary exploratory analysis with it. A few weeks into the project, we will start conducting more rigorous statistical analyses based on what we can find in our preliminary work, such as diving deeper into which variable we think might have a large effect on price. This is where we plan on incorporating some modules from the class such as hypothesis testing or network analysis. Finally, we will aggregate our results and evaluate if any definitive conclusions can be drawn from our analysis.

III. Data

We utilized three primary datasets for our analysis, coming from two distinct sources. Two datasets come from Kaggle, consisting of ride data along with weather data over the same time period [1]. The other dataset consists of traffic data in the general Boston area, and is provided by the Boston Department of Transportation through Analyze Boston [2].

A. *Ride Data*

Within the Kaggle source of data we utilize two separate datasets, which both contain data collected for about a one-week time period in November of 2018.

Uber	Lyft
UberPool	Shared
UberX	Lyft
UberXL	Lyft XL
Black	Lux
Black SUV	Lux Black
WAV	Lux Black XL

TABLE I
RIDE TYPES. A LIST OF UBER AND LYFT RIDE TYPES SORTED IN GENERALLY INCREASING LEVEL OF LUXURY.

One of the datasets which we will refer to as the Cab Ride Dataset, consists of information for all the instances of recorded rides through Uber and Lyft over the course of that week. Importantly, the rides are categorized into different ride types, meaning the different levels of "luxury" that Uber and Lyft offer. The full list of ride types is shown in Table I above. This dataset contained approximately 1.1 million rides that happened in Boston over that time period, giving us important information about the prices, distance traveled, etc. The exact schema of the table is depicted in Fig. 1 below.

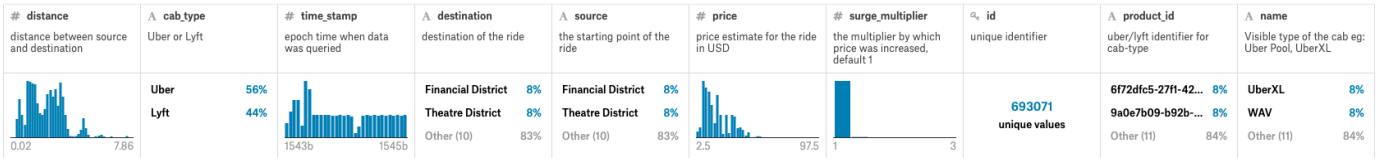

Fig. 1. **Cab Ride Dataset Schema**. This dataset consists of ride information for about 1.1 million rides over a one week period in Boston.

The Kaggle source also provided us with a dataset about weather conditions during that span of time. Various weather measurements were made at 1 hour intervals throughout the week at 12 specific locations in the Boston area. The exact information provided by this dataset is shown in the schema depicted in Fig. 2 below.

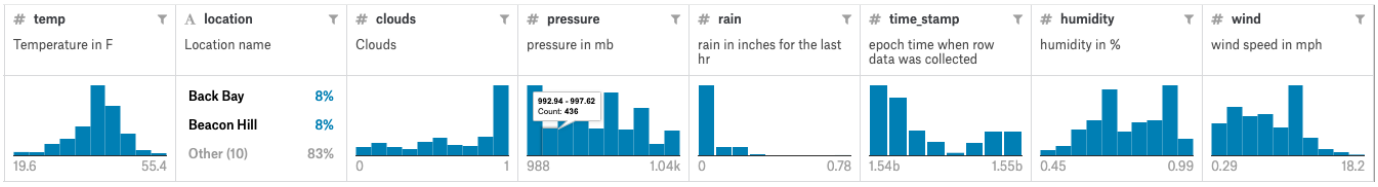

Fig. 2. **Weather Dataset Schema**. This dataset consists of weather conditions over the same one week period in Boston, measured at one-hour intervals.

B. Traffic Data

The traffic data used was compiled through Analyze Boston, the city of Boston's open source data hub. This dataset has intersection data from different areas of the city. The intersection data includes the number of vehicles that go through the intersection in each 15 minute interval between 7am and 6pm. While this includes bicycle and pedestrian counts, we were only interested in car and heavy vehicle counts for the scope of this project.

There are 11 different neighborhoods, in which we take 10 intersections from each neighborhood. When selecting the intersection, we tried to select the widest range of points as possible within each neighborhood. In doing so, we avoid intersections that appear next to each other since which would result in many of the same vehicles passing through, leading to skewed data. Due to lack of intersection data in certain neighborhoods, there ended up being a total of 103 intersections accounted for.

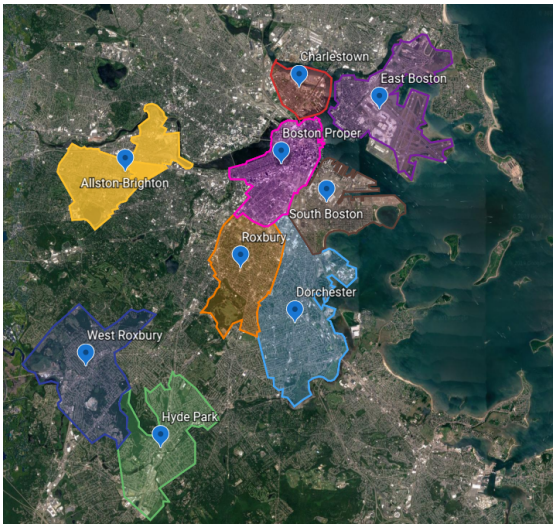

Fig. 3. **Neighborhood Map**. A map depicting the neighborhoods that we collected intersection data for.

C. *Strengths and Limitations*

Before diving into our analysis, it is important to first make some commentary about the data we are using, specifically identifying its strengths and limitations.

The Kaggle dataset is very clean with very few missing values or noisy data. This makes performing analysis on it and data preprocessing much easier, not having to write separate code to handle null values, or figure out some other method of populating missing data.

That being said, one glaring issue with these datasets is that it only spans a one-week period. Although we have over 1.1 million data points to analyze, this one-week period doesn't show a lot of variance in weather factors such as temperature, or variables that only change over longer seasonal periods of time. As such, we aren't necessarily able to make any conclusions about those particular variables because we only observe them take on a particular value.

This characteristic can also be seen as a positive thing though. Since those variables will stay mostly constant, we don't have to worry about them being confounding variables when we test other factors' effect on price. So for example, if we're testing rain's impact on price, we don't have to concern ourselves with temperature's confounding effect because it will stay mostly the same for all our data points.

The Analyze Boston dataset also has its strengths and limitations. There are a plethora of intersections covered that could certainly map out the city well when all compiled together. However, it was extremely difficult to pull the data in, as the data was laid out in different formats for most of the intersections, creating a lot of manual labor for calling multiple intersection CSV's into one function. It was for this reason, and for time constraints, that we were only able to sample each neighborhood for 10. Also, many of the files, namely ones collected before 2015, are in PDF format rather than CSV, making it impossible for us to analyze them, which further restricted the dataset.

IV. Analysis Methodology

A. *Preprocessing and Preliminary Exploration*

The first step in conducting our analysis was cleaning the data and doing some preliminary visualizations to summarize the data.

After merging the weather data and ride data together so that each ride also had weather conditions attached to it in addition to the ride info itself, then removing null values (which there were only less than 100 of), we conducted a series of preliminary tests to visualize the data that we had.

*1) **Distance**:* We first addressed the somewhat more obvious and intuitive variable of distance, plotting the average price of rides as a function of distance for each of the different ride types. The resulting plots are shown in Fig. 4 below, separated by Lyft rides and Uber rides.

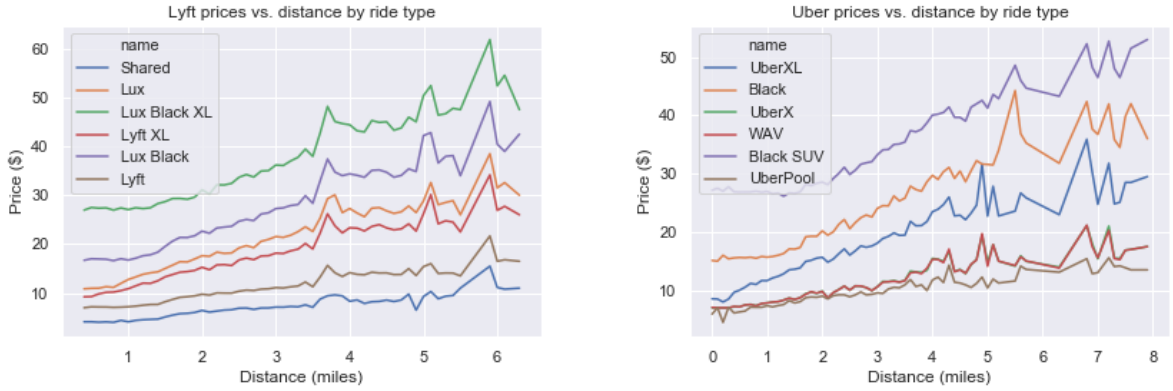

Fig. 4. **Distance vs. Price**. [*Left*] Prices as a function of distance for different Lyft ride types. [*Right*] Same plot but for Uber ride types.

Clearly, we see an upward linear trend of price relative to distance, which agrees with our intuition and experience of how Uber and Lyft rides are priced. What is also interesting to note however is the fact that for all ride types, we observe a relatively constant price if the ride is less than 1.5 miles. This indicates to us that it is likely both Uber and Lyft charge baseline prices for short rides less than that distance threshold. We also observe that the variance grows as distance increases. This makes sense as well as long rides will likely have many more confounding factors affect the price, and thus makes it more volatile.

*2) **Time of Day**:* We then looked at how the time of day affected prices. This was more of an intuition about how higher demand might result in higher average prices. As such, we used the number of rides by hour as a proxy for demand, and looked at how that changed throughout the day. Shown in Fig. 5 is a comparison of the number of rides on an hourly basis with the average price per mile using UberX and Lyft as the ride types, which are by far the most popular choices.

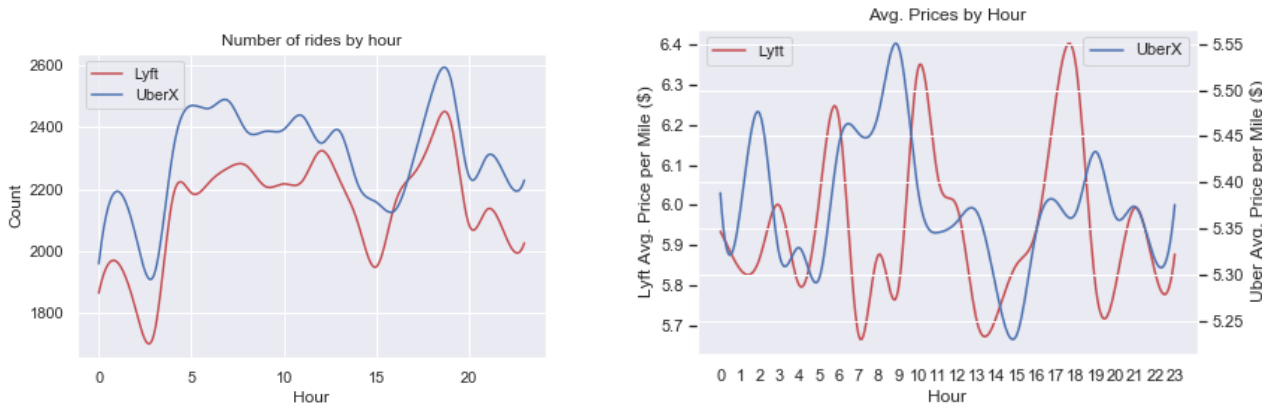

Fig. 5. **Per hour demand analysis**. [*Left*] The number of Lyft and UberX rides throughout the day per hour, a proxy for demand. [*Right*] The average price per mile by hour.

In looking at the left plot, we see the demand spike at times of day when we expect it to, namely in the morning when people are going to work, and around 5-6pm during rush hour when people are leaving work. It is also worth noting that in general UberX is more popular than Lyft.

Looking at the right plot relates the time of day to average prices. We see spikes at similar times in the day, in the 8-10am window and in the 5-7pm window for both UberX and Lyft. The fact that this lines up with the number of rides is no coincidence, and leads us to believe that in fact average prices are sensitive to demand fluctuations throughout the day.

B. Linear Regression

As a first piece of statistical analysis, we wanted to conduct a relatively simple test to figure out what kinds of factors *might* have important relationships with price. We decided to fit a linear regression, using distance and weather factors as variables and regressing on price. The idea of this analysis was to look at the regression coefficients of the best linear fit, and pick out ones that were high, meaning they had relative high correlation with price.

The variables that we used as regressors were distance, temperature, cloud cover percentage, pressure, rain, humidity, and wind. Before fitting the regression, we scaled the features so that they all lied on the same scale, as to not bias any particular feature. We then fit an ordinary least squares linear regression model that minimizes sum of squared error to each of the different ride types. We show the coefficients of Lyft vs. UberX as an example in Fig. 6, plotted on a log scale. All other ride types follow a similar structure in relative coefficient values.

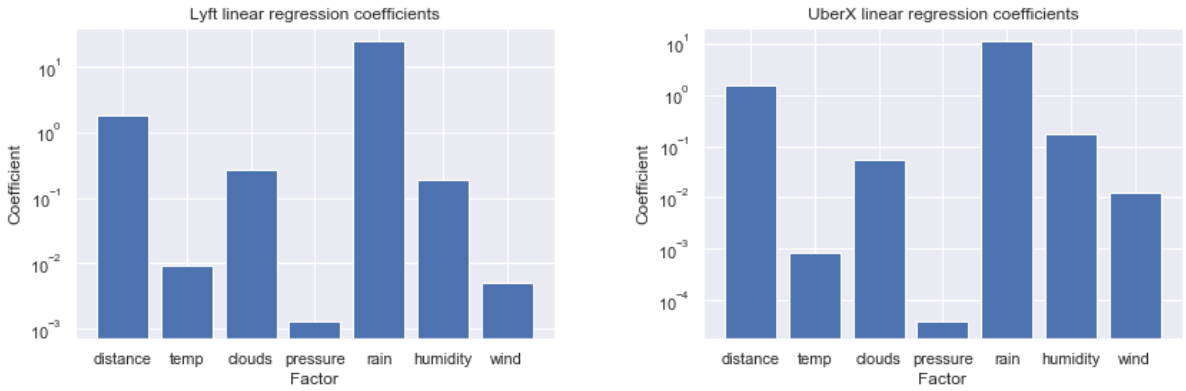

Fig. 6. **Regression coefficients**. [*Left*] Linear regression coefficients for Lyft rides. [*Right*] Linear regression coefficients for UberX rides.

We notice that the only significant factors based on the linear regression are rain and distance, which seems intuitively correct. To completely remove distance as a confounding variable, we also applied the same method on just the weather factors, regressing on price per mile instead of just price. The results for Lyft vs. UberX are shown in Fig. 7.

Again we see that rain seems to be the only relevant weather factor in determining price. This serves as an indication to us that perhaps we need to dig deeper into exactly what the relationship between rain and price are.

It is important to emphasize that we conducted this linear regression not as an extensive method of comparing factors, but rather to identify potential ones. If we wanted to be more rigorous in our linear regression analysis, we would have to compute p-values. However, we were only interested in finding indications that certain factors might have a relationship with price before conducting more rigorous analysis.

C. Hypothesis Tests

We wanted to test the effects of a variable on pricing in a more rigorous manner, so we split the data on the variable of interest and conducted two-sample t-tests, also known as Welch's test. We make the

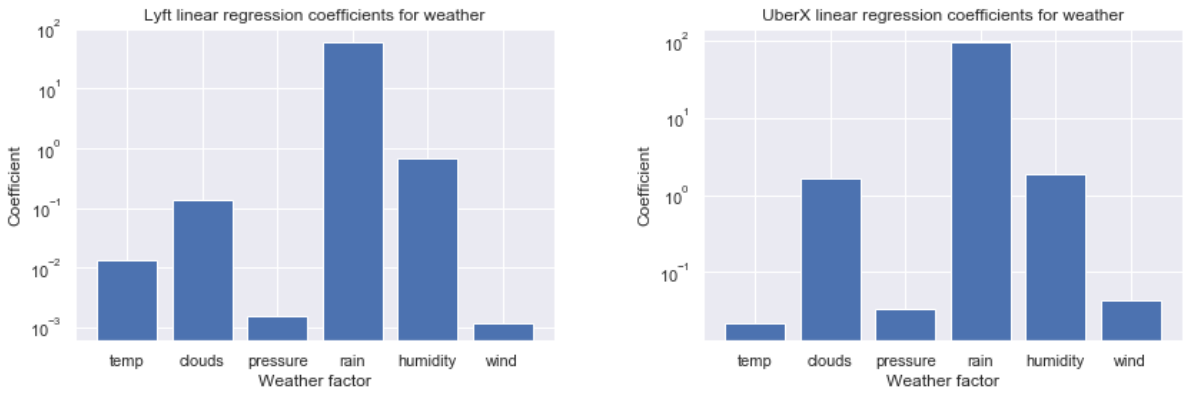

Fig. 7. **Weather regression coefficients**. Linear regression coefficients for Lyft and UberX rides only considering weather factors.

assumption that the external factors that impact ride pricing are relatively constant amongst the rides sampled on weekdays between 1PM and 3PM. This attempts to ensure that the two samples are capturing the true effect of the variable being investigated. Fortunately, the data contained over 1.1 million rides, so even after the filters were applied, the samples always consisted of many hundred rides.

*1) **Welch's Test**:* Welch's test is used to test the null hypothesis that two population means are equal. In particular, Welch's test assumes that the two populations have normal distributions, but importantly doesn't assume that the two population variances have to be equal. The test statistic is defined as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1}$$

where $\bar{X}_1$, $\bar{X}_2$ represent the sample means of the respective populations 1 and 2, σ_1^2, σ_2^2 the sample variances, and n_1, n_2 the respective sizes of the populations. With this test statistic we can analyze our hypotheses:

$$H_0: \quad \mu_1 = \mu_2 \qquad\qquad H_A: \quad \mu_1 \neq \mu_2$$

by applying a two-tailed test to calculate p-values [3].

*2) **Effect of Rain**:* After applying the weekday afternoon filter to the ride data, we split the rides based on the presence of rain. With this dataset, the number of rides that occurred in the presence of rain was about 190,000, a significant amount to test with. This analysis was carried across various partitions of the ride types. At a significance level of 0.05, we found Uber Black, UberXL, BlackSUV, WAV, UberPool, UberX, Lyft, Lux, Lux Black, Lyft XL, and all rides collectively exhibited a significant difference in means between the empirical distributions of the price of rides in dry and wet conditions. The only two ride types not to exhibit a statistically significant difference in price were Lyft Shared, the cheapest Lyft option, and Lux Black XL, usually the most expensive Lyft option. Given their extreme price points, it seems the prices of these services are resistant to the presence of rain, which makes intuitive sense. The resulting p-values for each ride type are summarized in Table II below.

Fig. 8 is a plot showing the empirical counts of all rides based on price per mile, separated by rain and no rain. We clearly see a shift in the distribution, indicating a structural difference. This was confirmed by the p-value of 0.002. Similar plots were generated for each of the different ride types, which gave the results described above.

*3) **Uber vs. Lyft**:* We separated Uber rides from Lyft ones to compare their pricing across various ride types. At a significance level of 0.05, we found that the mean of the empirical distribution of all of Uber's services grouped together significantly differed from that of all of Lyft's services grouped together.

Ride Type	p-value
UberPool	0.0271
Shared	0.1896
UberX	0.0046
Lyft	0.0428
UberXL	0.0025
Lyft XL	0.031
Black	0.0159
Lux	0.004
Black SUV	0.0117
Lux Black	0.009
WAV	0.0134
Lux Black XL	0.078

TABLE II

RAIN P-VALUES. P-VALUES OF HYPOTHESIS TESTS FOR RAIN VS. NO RAIN BY RIDE TYPE. RED VALUES ARE NOT DISCOVERIES.

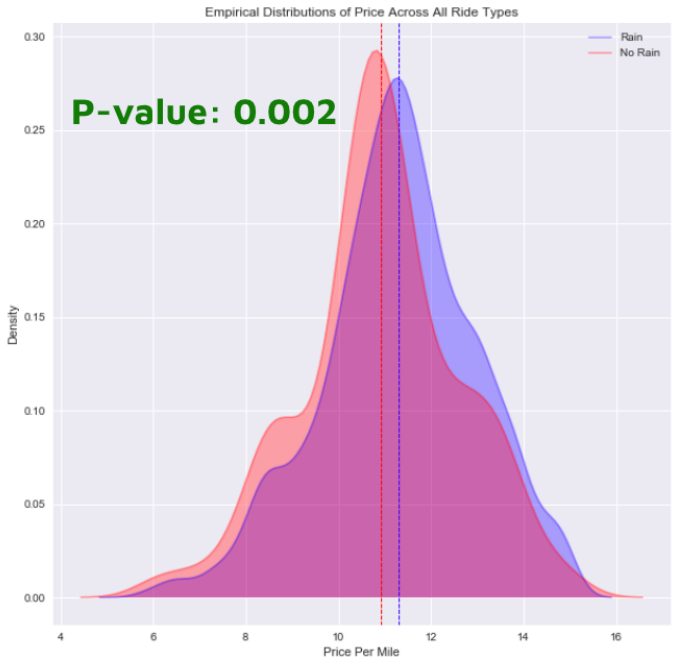

Fig. 8. **Rain vs. No Rain.** Distribution of rides by price per mile distinguished by presence of rain.

Further, there was a significant difference in the pricing of each company's basic service, UberX and Lyft normal. What we found was that Uber tended to have much higher variability in pricing than Lyft, but was on average cheaper. Fig. 9 shows this difference.

D. Network Analysis

We thought a network could better inform our analysis of traffic flow in Boston, and give us more insights on how ride prices relate to destinations.

First, we created a network where the nodes are the 12 pickup/drop-off zones from our Uber/Lyft rides dataset, and the directed edges represent rides from a source zone to a destination zone. The edges are weighted by the total number of rides between the specific source/destination pair. Using this network, we saw that the Financial District has both the highest in-degree and out-degree, meaning it is the busiest zone, measured purely by number of rides. Furthermore, we looked at betweenness centrality. Depicted in the visualization of this network in Figure 10, the node sizes are the inverses of their betweenness

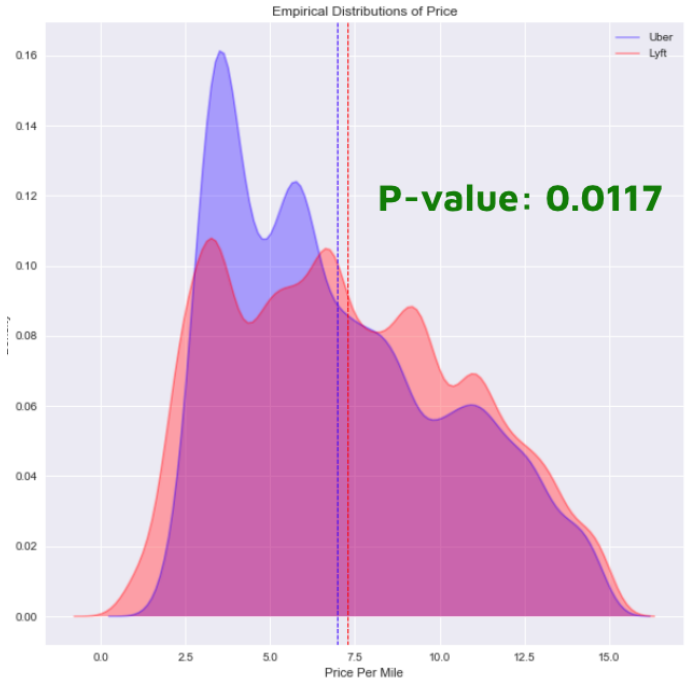

Fig. 9. **Uber vs. Lyft.** Distributions of all Uber and Lyft rides by price per mile.

centrality values (larger node for lower betweenness centrality) because when edges are weighted by number of rides, a lower centrality value actually signifies a more popular location. As expected, we see that the Financial District has a large node, as well as Back Bay. On the other hand, North Station does not have many rides passing through it, leaving it with a small node.

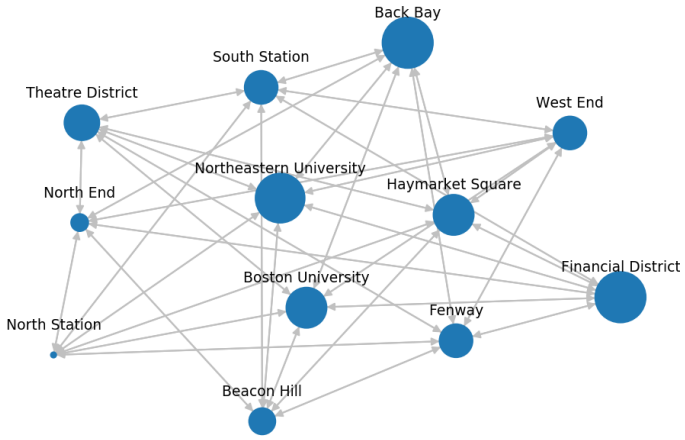

Fig. 10. **Popularity network.** Inverse betweenness centralities indicating the popularity of a particular destination.

We wanted to use centrality as a measure of traffic and see if held any relationship with price, but through plotting average price per mile vs betweenness centrality for each of the 12 locations seen in Figure 11, we saw no relationship.

Even though we did not see much of a relationship between average price per mile and centrality, we still wanted to see if certain destinations have particularly high prices. We created a second graph with the same 12 nodes as before, but instead weighted the edges with the number of rides with price per mile

Fig. 11. **Price vs. centrality.** Plot showing average price per mile vs. betweenness centrality. Looks to be random.

$>$ price per mile threshold t. We chose t to be two standard deviations above the mean, about \$30, which we believe will only include very high priced rides. We also normalized each of the edge weights by the total number of rides between the two respective locations.

$$t = \bar{\mu} + 2\bar{\sigma} \tag{2}$$

While this threshold was chosen relatively arbitrarily, we tried different thresholds which gave similar results.

This network is shown in Figure 12. The node sizes are based on the node's degree, so we can get an idea of which locations have high ride prices. We expected to see locations like South Station and Fenway have large nodes due to people traveling and attending events at Fenway (e.g. Red Sox game). We see that the South Station node is large, matching our expectation, the Financial District node is also large, and that actually the Fenway node is extremely small. We hypothesize that the there are many people going to and from work in the Financial District, causing higher prices. Fenway not being significant was a surprising result to us that we try and provide an explanation for later.

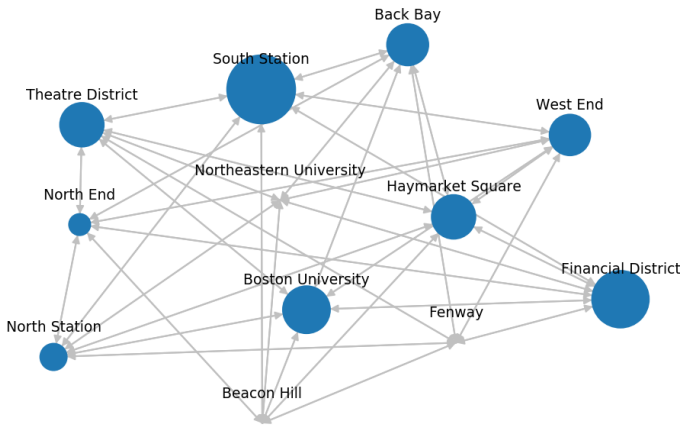

Fig. 12. **High price destinations.** Destinations with large number of high price per mile rides.

E. Traffic Analysis

We were unable to pursue our traffic analysis as much as we would have liked due to time limitations, but we did perform some exploratory analysis of traffic in the Boston area to be used as a stepping stone for future work.

With our intersection data, at each 15 minute interval, we summed up the number of vehicles moving through the intersections in all directions and divided by 15 (minutes) at all times t to get the cars per minute (cpm) as a metric to compare over time. All intersections in a neighborhood were averaged together at each 15 minute interval. Fig. 13 shows the cpm plot over time for each of the 11 neighborhoods.

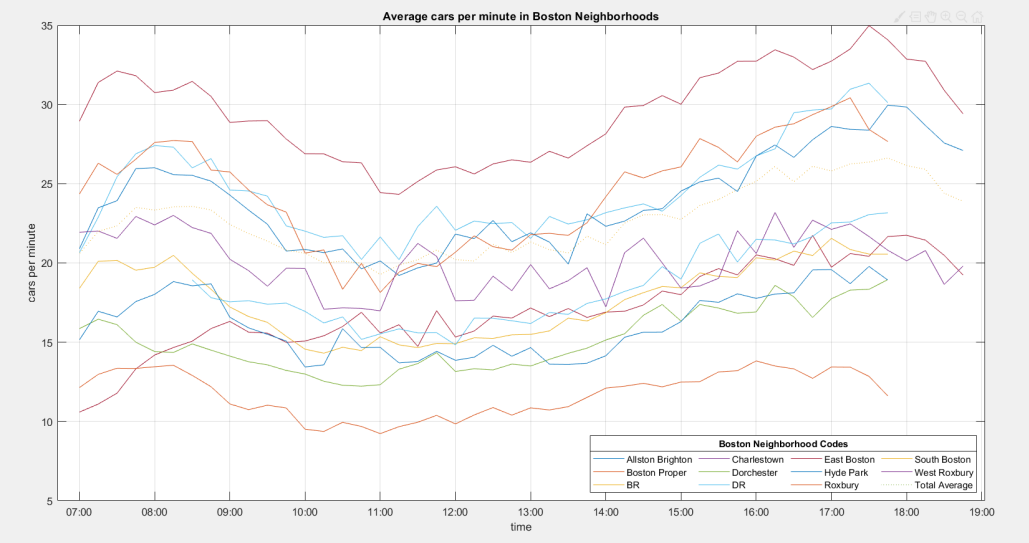

Fig. 13. **Average traffic**. Average cars per minute throughout the day by neighborhood.

Note with this plot that the shapes of all the neighborhoods are generally consistent, following a U-shape, and that we see peaks in each neighborhood cpm at approximately 8:00 and 17:30. This parallels the "time of day effect" for Uber/Lyft given in the first dataset that showed prices tended to be higher at rush hour. The correlation further supports the relationship between demand and average prices, as this traffic is also observed during rush hour periods.

In addition, we found traffic scores for each intersection and plotted them on a map to visualize the traffic in the city. Traffic scores were calculated by taking the mean of the cars per minute (cpm) values for the given intersection. A larger circle on the map, which is shown in Fig. 14 represents a higher traffic score.

We were not able to incorporate much of this in our analysis of prices, though we believe this analysis is something that could be easily extended to do so.

V. Results

We summarize our results as follows. We identified and tested for the significance of a few different factors that affect the pricing of Uber and Lyft rides. These factors include:

- Distance
- Time of day
- Presence of rain
- Destination location

We showed the effect of distance by plotting price as a function of distance for all ride types, showing that there was a clear linear trend in the relationship. We also saw that Uber and Lyft seem to charge baseline prices for rides under 1.5 miles, something surprising that popped out of our analysis.

We showed the time of day effect by plotting the average prices as a function of the hour in the day, and using number of rides as a proxy for demand to justify the relationship. We concluded that prices vary according to the demand of rides at a given point in the day, with spikes around rush hour times of 8-10am and 5-7pm.

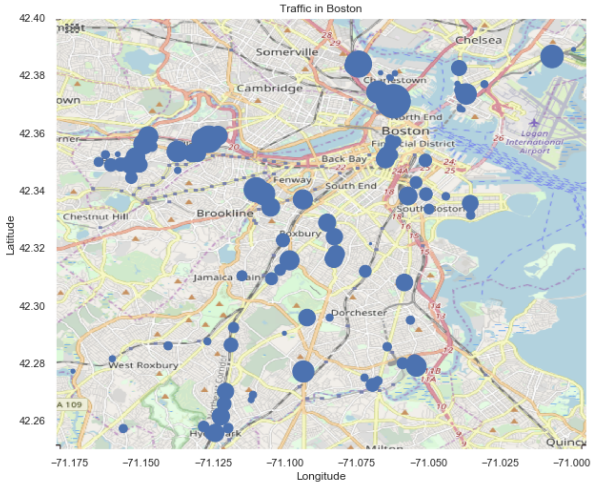

Fig. 14. **Traffic map**. Map showing generally where traffic exists in the Boston area.

To show the effect of rain on pricing, we compared all rides with rain and without rain for various ride types, looking at the distribution of prices per mile. We performed Welch's hypothesis test to test for equal means and found that for almost ride types, rain had a significant effect of the mean price per mile, increasing the prices on average.

Lastly, we used a network analysis to show that certain destinations have higher average prices than others. In particular, we calculated centralities for certain destinations in our network that equated to having a large number of rides at high prices. We concluded destinations such as South Station and the Financial District were examples of such destinations, explained by the consumer price inelastic nature of needing to travel or get to work. A surprising part of these results, however, is the fact that Fenway wasn't a high average price destination, which we had initially thought might be true due to large events such as concerts or sports games. A potential explanation is that the Fenway area has very congested and narrow roads already which people are aware of, and thus many people elect not to use ride-share to get to these events. In fact, based on our observation, when there are events at Fenway, most people end up walking or taking public transit like the T to get there.

VI. Conclusion

We believe that with our work we have showed relationships between certain factors of ride-sharing rides and their prices. In particular, we have confirmed some intuitive beliefs as well as uncovered some more surprising results. As a consumer, we now have a better understanding of when prices tend to be cheaper, how inflated prices are when it is raining outside, and how certain destinations may affect how much you pay.

This study was far from perfect, however. We had a very limited dataset that was confined to a one-week period only in Boston. While we claim that using Boston is a good approximation for general metropolitan areas, a more rigorous study would have to be done with other city's data to confirm this.

We believe that we have just scratched the surface with the analysis that can be done for analyzing these sorts of factors. In reality we know that Uber and Lyft have much more complex pricing models,

and are certain that there are more factors that go into their pricing scheme than just the factors that we explored in this project. We started to look at traffic, but didn't have time in the end to fully relate it to pricing. Moreover, we did not have yearly data so weren't able to compare across extreme and variable weather conditions. In the future, more of these variables can be analyzed as the data gets better. This project does serve however, as an excellent stepping stone for such work in the future.

ACKNOWLEDGMENT

We would first like to thank the TA who helped supervise our project, Genevieve Flaspohler. She was a tremendous help for guiding us through and supporting our ideas. We would also like to thank Professor Munther Dahleh and Professor Stefanie Jegelka for instruction of the IDS.012 course this semester.

REFERENCES

[1] Munde, Ravi. (2019, May). Uber & Lyft Cab prices, Version 4. Retrieved October 4, 2019 from https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices.

[2] Analyze Boston. Traffic-Related Data. 28 Mar. 2017. Sceris BTD Traffic Data Collection. Retrieved October 4, 2019 from https://data.boston.gov/dataset/traffic-related-data.

[3] "Welch's t-Test." Wikipedia, Wikimedia Foundation, 26 Nov. 2019, https://en.wikipedia.org/wiki/Welch's_t-test.