Final project: Data Analysis of Uber & Lyft Rides in Boston

Xiaoyu Liu, Hanqi Liu, Xingya Wang

5/8/2020

Abstract

Predicting how could Uber & Lyft drivers make profits is challenging because it needs us to identify the demand shifters. In this project, we use a data set of rides and related weather conditions, including a few hot locations in Boston from November 26th to December 18th in 2018 and also different weather indexes, to deal with this problem.

We investigate what factors influence trip volumes in Uber & Lyft by plotting inspiring figures, then set up a regression model by the backward method using AIC index as an evolution measurement and figure out which factors have statistically significant influence on the prices. Lastly, we apply the K-Means Clustering algorithm to our data set, using the latitude and longitude to identify the different boroughs within Boston and figure out which cluster will be the best district for taxi drivers to maximize their profits. This paper also brings insight into Uber & Lyft driver picking-up strategies.

Introductions

With the development of modern technology, a demand in transportation efficiency emerges and the need for cabs has been increasing significantly in the recent decade. Uber, a technology company founded in 2009, created a smart phone application that matches and handles payments between consumers seeking rides and Uber's "driver-partners." Later, a similar app called Lyft was founded in 2012. These two apps provide people with excellent ride experience and also an easy way to make some money in their spare time, or even make a living by driving Uber or Lyft.

As we all know, these online ride-hailing apps will algorithmically match drivers to consumers and help them seeking ride matches. Another critical feature of those apps is that they use real-time pricing (i.e. "surge" pricing) to equilibrate local, short-term supply and demand. We are interested in how the price of ride-sharing in Uber & Lyft is affected by different factors and whether there is a specific pattern of daily ride requests. This paper offers both some pieces of advice for consumers to get cheaper rides and for drivers to maximize their profits during work.

In this project, we use the dataset of Uber & Lyft in Boston to make some interesting plots and figure out how the prices of these ride-sharing apps changed. Besides, we apply the K-Means Clustering to find the best location to wait for new orders.

Specifically, in order to investigate what factors contribute to the demand for cabs, we use some ggplots to answer questions like: how days of a week will influence the demand? Do weekdays have more demand than weekends at 9 am? How would different weather conditions influence the demand for cabs? Which types of riding would people more likely to use in different weather/times? The data visualization part will solve the questions above.

After we are clear about the demand shifters, we build regression models with different variable selections and see how well they fit the original data. In addition, we would like to know whether there is a specific area that has more orders than other places. A lot of researchers used a density-based spatial clustering of application with noise named DBSCAN, however, it turns out that the method doesn't apply to our data set. The results of DBSCAN don't make any sense. So we just come back to the typical K-Means Clustering,

using the data of longitude and latitude. We hope we can suggest the Uber & Lyft drivers where to go to receive more requests in a single day.

Data and Methods

Our dataset contains the real-time data using Uber & Lyft API queries and corresponding weather conditions and a few hot locations in Boston. The collectors queried cab ride estimates every 5 minutes and weather data every 1hour. The data goes from the end week of November and to a few weeks in December in 2018. It not only covers the various types of cabs for Uber & Lyft and their price for a given location, but also the related weather condition, including temperature, rain, cloud, etc. The full dataset contains nearly 700,000 pieces of data, which definitely allows us to do some research on the ride-hailing situation in Boston.

To be more specific, it contains variables related to cabs: the hour, day, the month of each trip; the source and destination; cab types and the subdivision; price and distance; the latitude and longitude. What else, it also provides weather information: temperature, a summary of the weather (cloudy, rain, clear, etc.), visibility, humidity, precipitation probability, and so on.

Given that demand is one of the most important factors that affects cab drivers' profits, we firstly find the variables that influence total demand, for the later use of price modeling. To accomplish that, we make some plots (see Appendix) to show them more intuitively.

After considering the general situation, as shown above, we set up a linear model by backward regression method using the main variables, specifically, the distance, the wind speed, and different names of cabs. The result shows that they have the statistically significant influence on the price of rides.

Considering the uniqueness of our Uber & Lyft dataset, which contains the information about the address of the pick-up location of every order, i.e. the longitude and the latitude. We choose to use the K-Means Clustering method. K-Means Clustering is a centroid based clustering. It is an iterative clustering algorithm in which the notion of similarity is derived by how close a data point is to the centroid of the cluster. Though sometimes it's more reasonable to choose the number of clusters based on prior pieces of knowledge, we have to find the "elbow" on a plot of SSEW versus K.

Applying the K-Means Clustering method on the latitude and longitude data has been widely used by many researchers as well as entrepreneurs, i.e. eBay, Amazon. Etc. Following that literature, we choose to use the K-Means Clustering method on the latitude and longitude data of every order. After that, we get the descriptive statistics based on the clustering result and find the number of orders are particularly high in the area of The Great Boston Area, showing that the Uber & Lyft drivers may have a higher probability to get a ride in those areas, details are in the Results section.

Results

1. Plots

The demand for cabs and cab types might be influenced by different factors including weather, hours of a day, and so on. Let's focus on several figures to see how factors contribute to demand.

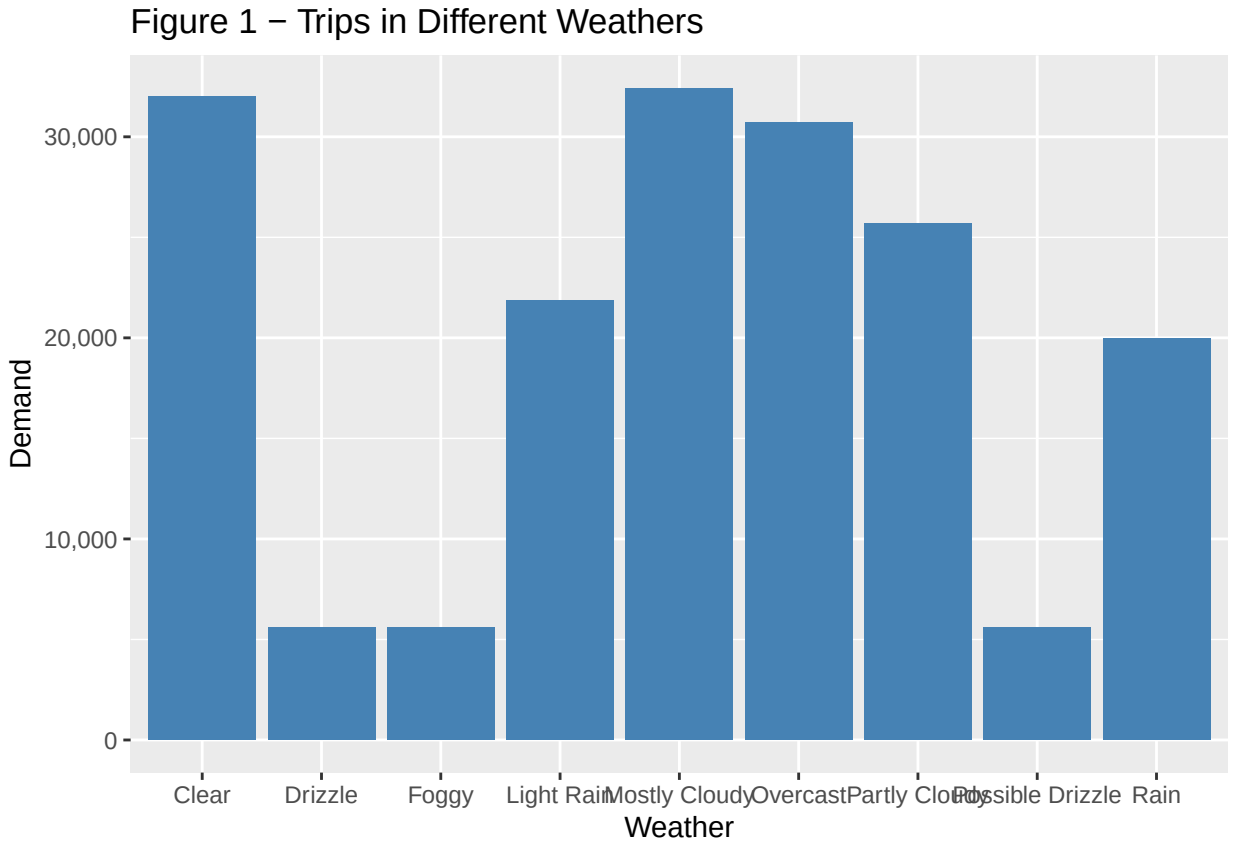


Figure 1 summarizes the trips, which also represent the demand for cabs, in different weathers. Since total order size might be misleading, which means more same weather will lead to higher trips in total, we separate the data into each day in order to find the real demand in different weathers. It is obvious that clear, cloudy, and overcast days provide larger demand; followed by rainy days; and foggy, drizzle days have less demand compared with others.

It might because people do not prefer to go out during "bad" weathers including days with poor visualization and would like to do so if the weather does nothing to their real life, such as the sunny and cloudy days.

Figure 2 − Trips by Hour and Cab Types

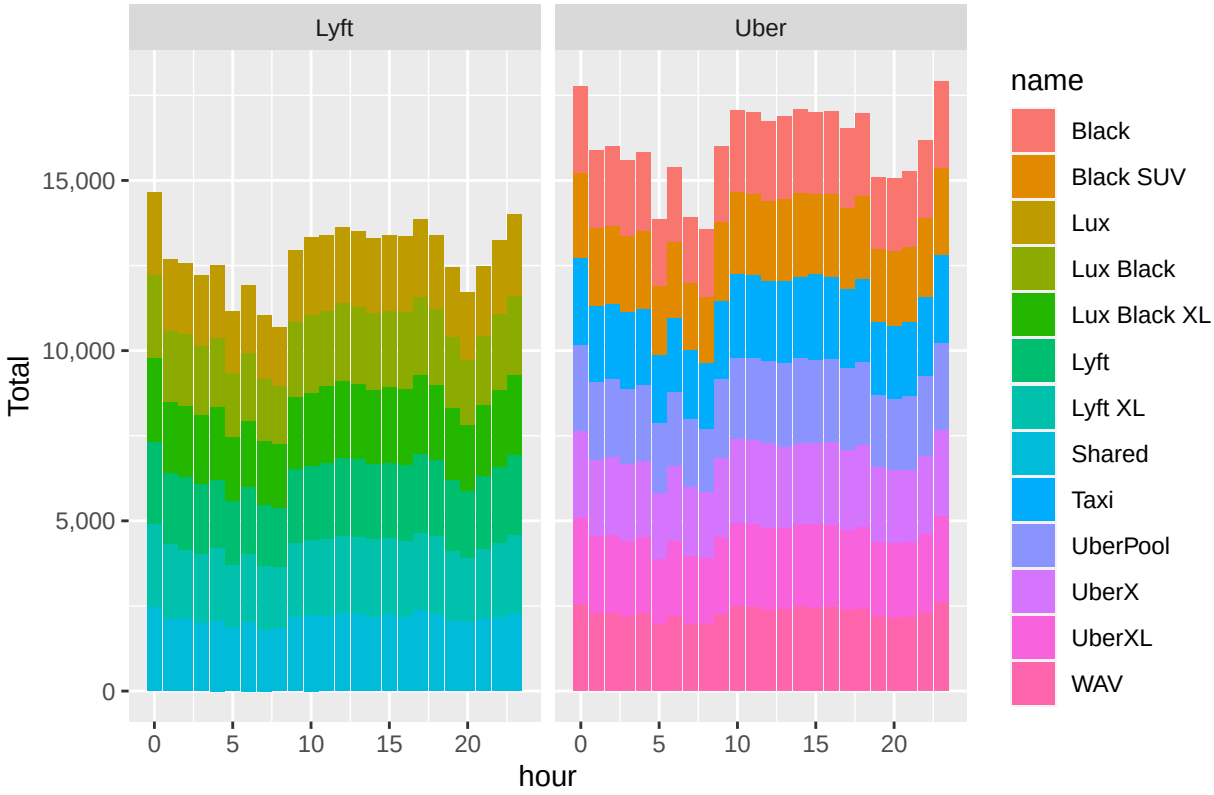


Figure 2 shows a plot of the relationship in demand and hours, combined with different type names faceted in Uber and Lyft. We can see hours do not really change how people choose the cab types, which is, cab types changes here only based on the total hour demand difference. Also notice that Uber has higher demand than Lyft at any time, however, the influences of hours to demand are still almost the same.

What else, if we only consider the hour's demands here, it also shows some frequencies in Figure 2. From 10 a.m. to 6 p.m., and also the two hours in the end and the beginning of a day, there are more trips and kind of less at other times.

Now, we consider the average temperature of a day rather than of any time, since the demand for the same cab types does not vary too much in each hour. See Figure 3.

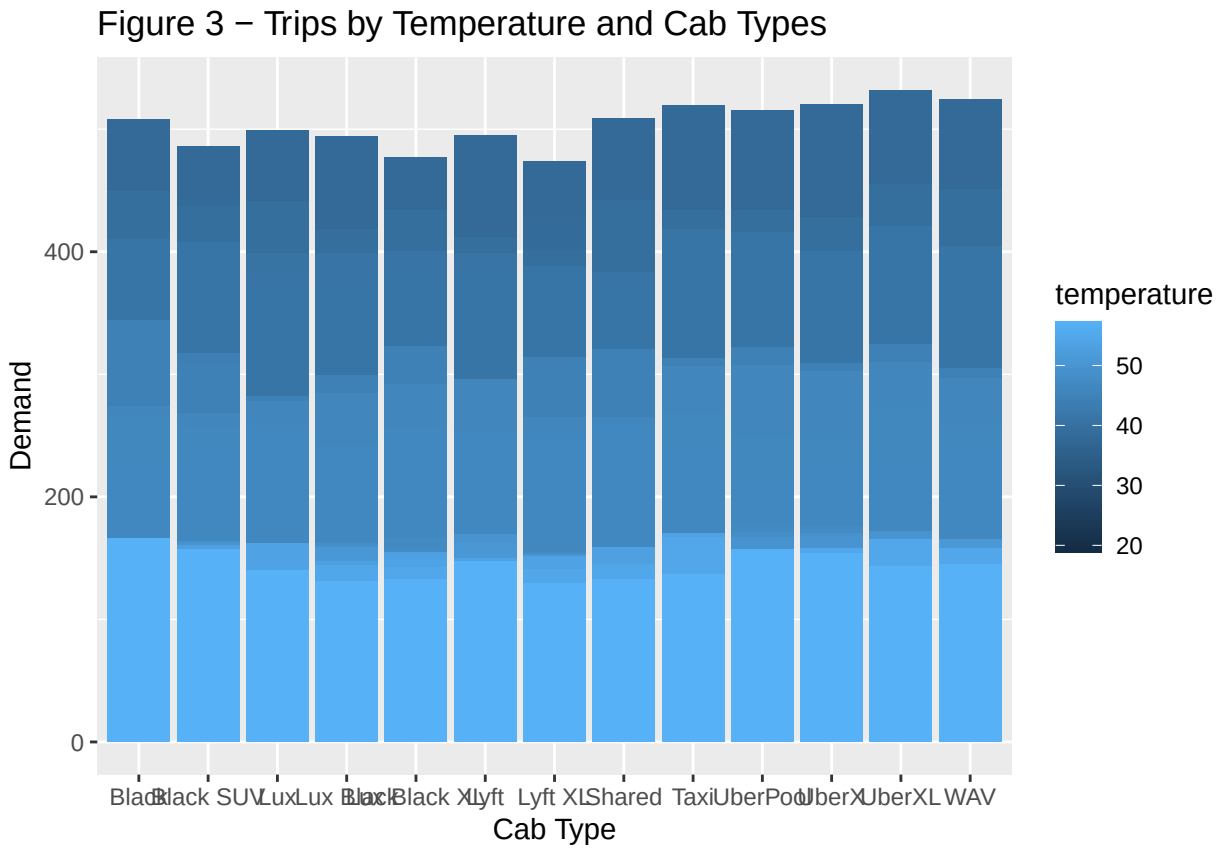


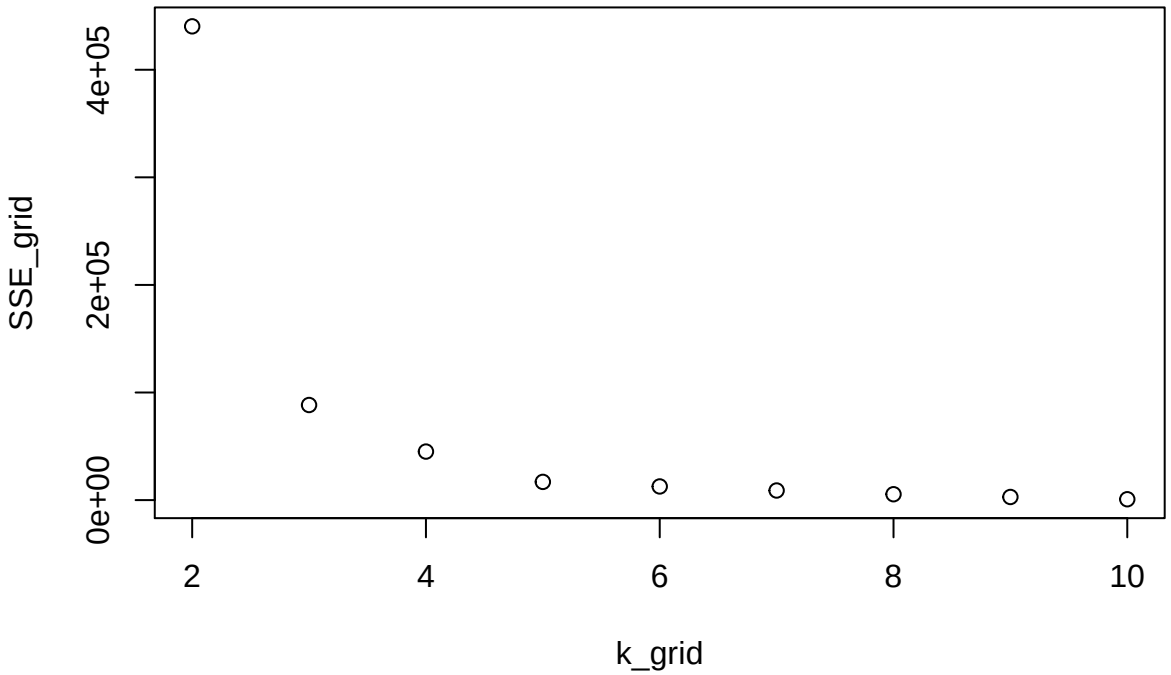
Figure 3 shows something more interesting. When the temperature is slightly warm (55°F or more), Uber cabs have a little bit higher demand than Lyft as we find before. However, when it's around 45°F, people show less preference in Lyft cab types except for Lux Black XL, Lyft XL, and Shared. Then, when it's below 45°F, Lux, Lux Black, Lyft, and Shared increase more than other types in Lyft.

What happened here could be translated as when the weather is pretty good in that month, people do not vary too much in the cab types, however, if the temperature went done, people will move to other choices due to waiting time and prices or other factors.

2. Clustering

Regarding the Clustering, we firstly use the K-means method introduced in the method section to do the clustering on the latitude and longitude data, with k = 3, which is chosen because k=3 is the elbow point in the graph. The result gives us 3 clusters, of the address centered in (42.36040, -71.06075), (42.34703, -71.09374), and (42.21480, -71.03300), which are all around the Boston City downtown area. They are all located in the center of the City of Boston, which is consistent with common sense that the demand is high in the downtown area.

Figure 4 − Elbow Plot



After this, we showed the order summation in the period on the 3 clustering areas, as shown in Table 1. As shown in Table 1, the second clustering point, which is centered in the City Hall of Boston City, have the most orders in the given period. This leads to the result that compared to the point of (42.36040, -71.06075) and the point of (42.21480, -71.03300), more Uber & Lyft driving orders are emerging in the second area, which is centered at (42.34703, -71.09374).

Table 1: Sum of Orders in the selected cluster area

Cluster 1	Cluster 2	Cluster 3
88546	416630	187895

3. Model

Firstly, after reading the literature and knowing the basic principles of Economics, we consider the price of ride-sharing is related to the types of cars, the distances of the ride, and the hour of a day when the consumers take the ride.

Considering there is a close correlation between the names of cabs and the prices, we set up the dummy variables of the name of the cars into the model. From the dataset, we know there are 13 different names of the ride-hailing cars, they are Black, Black Suv, Lux, Lux Black, Lux Black XL, Lyft, Lyft XL, shared, Taxi, UberPool, UberX, UberXL, and WAV. Also, we have to see that there is no data about the taxi here.Then we set up and the names of the ride-sharing cars as dummy variables and add them up to the original dataset to help regress the model.

We build the model to analyze the relation of price and different independent variables and selecting the hour, distance, temperature, windspeed, and the dummy variables of different names of the cars as the explanatory variables. The coefficient estimations and statistic measure of these this model named fit1 were reported as below. In our report, we take the broadly-used indicator p-value to test the significance of explanatory variables. Most of the explanatory variables were found to be statistically significant at the 99%

level (p-value<0.01), especially for the explanatory variables for the different names of ridesharing cars.

From the summary of model fit1, we know the variable hour, temperature and windspeed are not statistically significant on 90% level, it may be because of our source data, which shows that some hours in the day have the same amount of demand, in this way, considering the bias of selection, it will negatively influence of hour and temperature on the ridesharing prices. Furthermore, we can see that the size of the cars, to be specific, the existence of Uber Black, Uber Black XL, and Black SUV will increase the price of riding, holding all else factors as constant. Also, R-squared equals 0.9244, which performed great.

After analyzing this, we used the backward method to select variables and compared the AIC index of these two models. By taking this measurement, we received the model fit2, which deleted the insignificant variables hour, the temperature of the model fit1. Then we checked the AIC index of two models, which shows that AIC of fit1 model is -556784.9, while the AIC of fit2 model is -556789.4, which is less than the first one. Therefore, we plan to take model fit2 as our final model, showing that distance and the names of different cars have a statistically significant effect on the prices.

```
                                (Intercept)
                                2.6292059362
                                   distance
                                0.1667460449
                                  windSpeed
                                0.0001097308
           factor.newridershare.name.UberX
                               -0.7438055279
        factor.newridershare.name.UberPool
                               -0.8521265673
       factor.newridershare.name.Black.SUV
                                0.4041603043
             factor.newridershare.name.Lux
                               -0.1574954157
            factor.newridershare.name.Lyft
                               -0.7624182173
          factor.newridershare.name.Lyft.XL
                               -0.3059182158
           factor.newridershare.name.Shared
                               -1.3436105871
              factor.newridershare.name.WAV
                               -0.7438077829
           factor.newridershare.name.UberXL
                               -0.2796559717
    factor.newridershare.name.Lux.Black.XL
                                0.4611896268
       factor.newridershare.name.Lux.Black
                                0.1094109356
distance:factor.newridershare.name.Shared
                                0.0378107821
```

Table 2: AIC of Our Selected Models

Model 1	Model 2
-556784.9	-556789.4

Conclusion

By summarizing the previous plots, we could determine several factors that influence the demand for cabs. First, during the clear, cloudy, and overcast days, the demand would be higher than others. Second, hours and days of the week are not significant in choosing cab types (see Appendix for days of week part), what people care about is the temperature and would not worry about the higher prices during cold days. Third, which is pretty obvious, people prefer Uber to Lyft.

For the model regression part, we know that the distance, and the types of different car names, to be specific, whether the ride is a shard or not, will statistically significantly influence the ridesharing prices on 99% level. However, considering the selection bias of the original data and the nature of Boston in November and December, we can't figure out the relatively significant correlation between price and weather situation, for example, the temperature and the precipitation intensity this time.

We apply the K-Means Clustering method on the address data (longitude and latitude) for every order of this dataset. With k = 3, we get the result of 3 areas, be centered at the point of (42.36040, -71.06075), the point of (42.34703, -71.09374), and the point of (42.21480, -71.03300) respectively. We next count the order number based on the clustering result, and it shows that the area centered at (42.34703, -71.09374) has the most Uber & Lyft orders in that period, attracted more drivers and customers compared to the areas centered at the point of (42.36040, -71.06075) and the point of (42.21480, -71.03300).

Reference

[1] Barahona, Diego, et al., "Exploring the taxi and Uber demand in New York City: An Empirical and spatial modeling", Conference Paper, 2017

[2] Cohen P, Hahn R, Hall J, et al. Using big data to estimate consumer surplus: The case of uber[R]. National Bureau of Economic Research, 2016.

[3] Correa D, Xie K, Ozbay K. Exploring the taxi and Uber demand in New York City: An empirical analysis and spatial modeling[C]//96th Annual Meeting of the Transportation Research Board, Washington, DC. 2017.

[4] Cramer J, Krueger A B. Disruptive change in the taxi business: The case of Uber[J]. American Economic Review, 2016, 106(5): 177-82.

[5] Uber and Lyft Dataset Boston, MA: https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma

Appendix

To give the reason why we think temperature is an important variable, we put two related figures below. It is really obvious that they gather in different tiles in two plots under different conditions (with or without temperature). In Figure A1, Thursday, Friday, and Saturday have more trips, and in Figure A2, only Thursday shows higher demand. Thus, the temperature will actually influence demand and we have to see how it works.

Figure A1 − Demand Map by Cab types and Day of Week(with temp

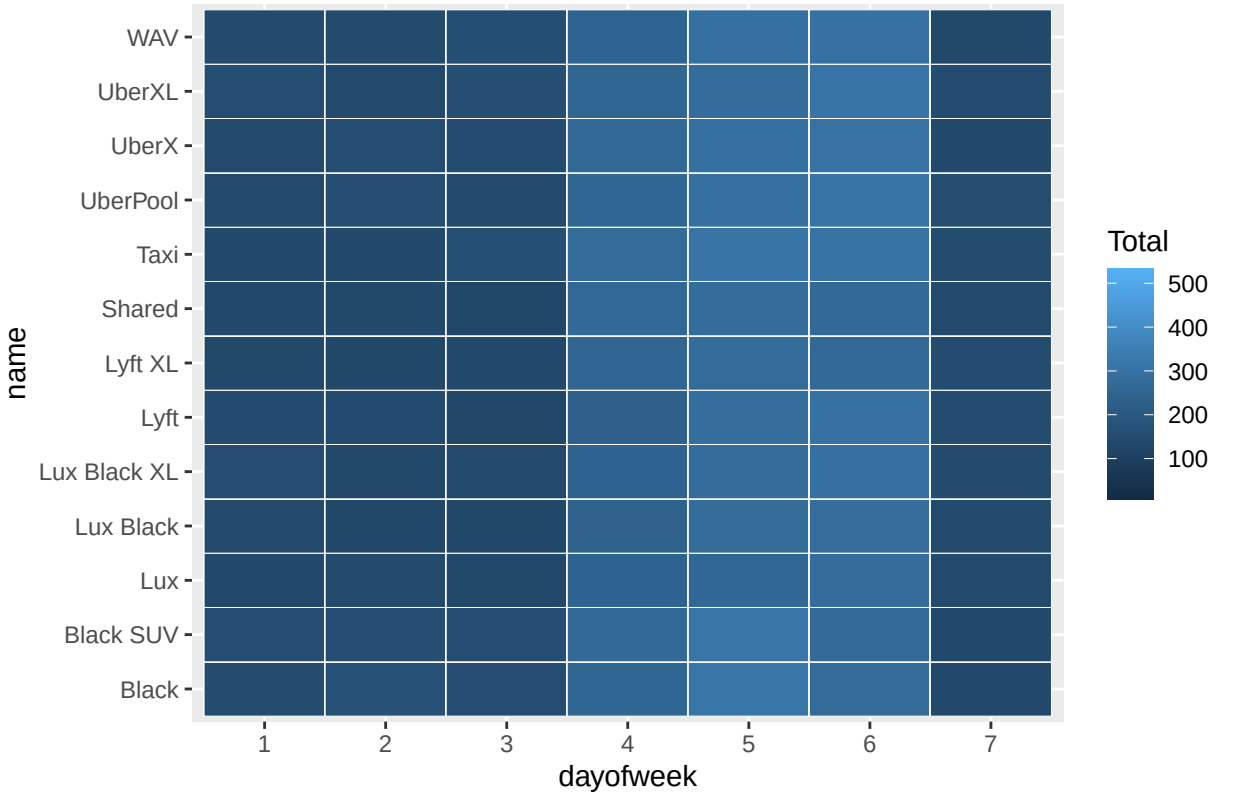


Figure A2 − Demand Map by Cab types and Day of Week(without te