

Examining Factors that Determine Pricing of Ridesharing Services

“University of Nebraska at Omaha”

*“Allie Bray, Bikram Maharjan, Brendan Schutte, Mackenzie Maschka, Maryam Raftari”
December 12, 2019*

Abstract

Ridesharing services, like Uber and Lyft, have quickly become popular forms of commute for users all across the world. Time of day, weather, price, trip source and destination are all driving factors that influence the usage of these ridesharing services. By exploring the relationship between these factors, and understanding the usage patterns, we can improve the quality of the service for both the ridesharing drivers and the users of the platform. This ensures that the drivers would want to be on the road and the users would want to utilize the platform for their every day commute. Using R, we explored a dataset of Uber and Lyft prices, trip sources and destinations, and distances that were collected from Boston, MA for roughly two weeks from November to December 2018 at regular five minute intervals. This data was cross-analyzed with a dataset of the weather conditions in Boston, MA at the same time the ride was taken. We sought to identify the most influential factor that drives price fluctuation for ridesharing services.

1 Introduction

Ride sharing has taken over urban areas as a replacement for traditional taxi services. In the last 5 years, companies such as Uber and Lyft have seen substantial growth, both in the number of drivers and the number of users. According to Uber’s website, in 2018, 14 million Uber rides were taken each day in over 600 countries. The concept of ride sharing is simple: the ride sharing company develops a user-friendly phone app which connects users to drivers who will take them to their desired destination using the driver’s personal vehicle. The appeal of driving for one of these companies stems from the fact that the driver is allowed to work whenever he or she wants. There are no set hours. The appeal for users comes from the convenience of an app on their phone combined with an average lower price when compared with taxis.

As ride sharing continues to gain momentum, it is important to explore factors affecting the experiences of users and drivers. The most notable factor that affects users is price. Both Uber and Lyft have varying prices that can fluctuate depending on the time of day, location, etc. For this report, data gathered from Uber and Lyft rides in Boston will be analyzed to explore these possible relationships. The results of this study can give insights into how the ride sharing companies price their services which can be beneficial to both the users and the drivers.

1.1 R Packages Used

This project was imported into the RStudio environment and compiled by researchers wishing to reproduce this work for future datasets. In the first step of our R project, we imported the essential packages that we used in this Uber and Lyft data analysis project. To construct this report we used following packages:

List of Packages
dplyr
ggmap
ggplot2
knitr [R-knitr]
lubridate
reshape2
sqldf
tidyverse
kable Extra
formattable

Table: List of packages used in this report.

- **dplyr**
This package is the lingua franca of data manipulation in R. dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges.
- **ggmap**
The ggmap package extends the ever-amazing ggplot core features to allow for spatial and geographic visualization. This package is a collection of functions to visualize spatial data and models on top of static maps from various online sources (e.g Google Maps and Stamen Maps). It includes tools common to those tasks, including functions for geolocation and routing.
- **ggplot2**
This is the backbone of this project. ggplot2 is the most popular data visualization library that is most widely used for creating aesthetic visualization plots. It is a system for declaratively creating graphics, based on The Grammar of Graphics. ggplot2 is now over 10 years old and is used by hundreds of thousands of people to make millions of plots.
- **knitr**
This package provides a general-purpose tool for dynamic report generation in R using Literate Programming techniques. The R package knitr is a general-purpose literate programming engine, with lightweight API's designed to give users full control of the output without heavy coding work.
- **lubridate**
Functions to work with date-times and time-spans: fast and user friendly parsing of date-time data, extraction and updating of components of a date-time (years, months, days, hours, minutes, and seconds), algebraic manipulation on date-time and time-span objects. The 'lubridate' package has a consistent and memorable syntax that makes working with dates easy. Our dataset involves various time-frames. In order to understand our data in separate time categories, we used the lubridate package.
- **reshape2**
The reshape2 package allows for data frames to be reshaped into a 'wide' or 'long' format.
- **sqldf**
This packaged allows for the use of SQL queries to be used to organize and manipulate data frames within R.
- **tidyverse**
The tidyverse is an opinionated collection of R packages designed for data science. The basic principle of it is to tidy the columns where each variable is present in a column, each observation is represented by a row and each value depicts a cell.

- kable Extra

kableExtra is NOT a table generating package. It is a package that can “add features” to a kable() output using a syntax that every userR loves - the pipes %>%. We see similar approaches to deal with plots in packages like ggvis and plotly. There is no reason why we cannot use it with tables. Most functionalities in kableExtra can work in both HTML and PDF. In fact, as long as you specifies format in kable() (which can be set globally through option knitr.table.format), functions in this package will pick the right way to manipulate the table be themselves. As a result, if users want to left align the table, kable(...) %>% kable_styling(position = “left”) will work in both HTML and PDF.

- formattable

This package is designed for applying formatting on vectors and data frames to make data presentation easier, richer, more flexible and hopefully convey more information. Atomic vectors are basic units to store data. Some data can be read more easily with formatting. A numeric vector, for example, stores a group of percentage numbers yet still shows in the form of typical floating numbers. This package provides functions to create data structures with predefined formatting rules so that these objects store the original data but are printed with formatting. The package provides several typical formattable objects such as percent, comma, currency, accounting and scientific. These objects are essentially numeric vectors with pre-defined formatting rules and parameters.

2 About the Data

As neither Uber or Lyft publicly shares data of rides or prices, the data used in this analysis stems primarily from a relevant dataset found on the website, Kaggle (1). Data regarding cab rides and weather was collected from twelve popular locations in the Boston area from November 26, 2018 to December 18, 2018. The owner of the dataset created an application in Scala to query data at regular intervals; cab rides were queried every five minutes and weather every hour (2). The data collected regarding cab rides includes the following variables:

- the unique ID of each cab,
- the ride source and destination,
- whether the rides was facilitated by Uber or Lyft,
- what type of vehicle completed the ride,
- the estimated price of the ride,
- the distance of the ride, and
- a surge multiplier that determines how the price was increased.

It is important to note the demographics of each location in the dataset. These observations could prove as justification to the results of our analysis. The following table introduces each location, that could be a source or destination in the data set, and describes the area.

Location	Description
Back Bay	office space, upscale shopping, hotels, upscale housing
Beacon Hill	upscale housing
Boston University	private research university
Fenway	baseball park
Financial District	government offices, downtown
Haymarket Square	produce market
North End	residential
North Station	major transportation hub
Northeastern University	private reserach university
South Station	major transportation hub
Theatre District	entertainment, location of plays/performances, night life
West End	office space, hospital, government offices

Table: Locations used as sources and destinations in dataset.
The data regarding weather included the following variables:

- temperature,
- precipitation,
- humidity,
- wind, and
- a short summary of the weather type (i.e. Rain, Clear, Overcast, Sunny, etc).

The weather dataset lacked the unique ID of each cab needed to merge the cab rides dataset and the weather dataset. This was remedied by means of a separate data collection process described further in coming sections. A discontinuity of data was also discovered from December 5-8 and December 11-12. A method of imputation was employed to aid in the visualization of this missing data.

2.1 Preparing the Data

Data preparation was one of the most challenging and as well as the fun part of the project. We worked on transportation data, which included the data of ridesharing companies such as Lyft and Uber. This dataset was collected in Boston, MA from mid-November 2018 to mid-December 2018. We selected this dataset because we could examine the cost of the rides between Lyft and Uber on various factors such as distance, cost per distance, volume of ride requests at different hours of the day, how weather affects ride requests and several other factors between Lyft and Uber. Please note that these data are not actual trips taken but an API requests at the various interval of time for about a month.

The dataset we used was available in [Kaggle.com](#) as a Public Domain License via the Kaggle user [RaviMunde](#). As per [Wikipedia](#)(3), “Kaggle is an online community of data scientists and machine learners, owned by Google LLC. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.”

Using Kaggle we could easily view parts of the dataset, most importantly the column head. This is especially useful in larger dataset. The dataset we used had 2 files. The first one is the actual dataset itself called `cab_rides.csv` with 10 columns and 693K rows. The second one was the weather data associated with the `cab_rides.csv` called `weather.csv` with 8 columns and 6276 rows. The `cab_rides.csv` dataset’s row is uniquely identified by a column `id`. However, there was no unique relationship between the `cab_rides.csv` and `weather.csv`. While `cab_rides.csv` itself could give us insights, we felt it would be interesting to understand if weather do play a significant role in pricing a ride, so we decided to find an alternative way to get the weather data that match the current dataset location and time the ride was requested. This weather data could provide us an invaluable insight while we examine on why a particular ride request was expensive or cheaper. We could also find out if the weather really does affect the pricing model of the ride.

Search for a Historical Weather Data dataset

We found several websites that offers the historical weather data for our analysis, such as:

- <https://openweathermap.org/history>
- <https://www.weatherbit.io/api/weather-history-hourly>
- <https://darksky.net>
- <https://developer.accuweather.com/>

While all the above websites provide historical weather data, some were difficult to configure and the process to fetch data for our particular needs with our dataset was cumbersome. Others were only “premium” version which required payment to use their service. Of all the websites we found, we selected **DarkSky API** (<https://darksky.net>) for our analysis because it was easy to use and fit right into our current data schema. As per the website <https://darksky.net> **DarkSky API** (<http://darksky.net/dev>) “The Dark Sky Company specializes in weather forecasting and visualization” (4)

Gathering the Weather data from DarkSky API

The DarkSky website provides an API, also known as Application Programming Interface which allows for communication between a client (meaning the API user) and a server (meaning the Dark Sky API) intended to get a data, given the minimum proper request parameters. The Dark Sky website is free to use for the first 1000 API requests per 24-hour period. After that, the users needed to pay a fee. This was a suitable platform for us because even though we had around 693K ride data, the entire dataset when we convert the timestamp was 332 hours.

Using DarkSky API

The Dark Sky API URL looks like this:

`https://api.darksky.net/forecast/[API_KEY]/[LATITUDE],[LONGITUDE],[TIME]`

The URL has a minimum of 3 required parameters:

- **API_KEY:** This is the private key that a user account is assigned after the account is created and email verified.
- **LATITUDE:** Latitude is the geographic coordinate that specifies the north-south position of a point on the Earth's surface for a particular location. Eg: 41.258543
- **LONGITUDE:** Longitude is a geographic coordinate that specifies the east-west position of a point on the Earth's surface for a particular location. Eg: , -96.012878
- **TIME:** Time is in epoch format per second. Eg: 1571769494

If we combine a valid Latitude and Longitude, separated by a comma(,), we can locate a place on Earth. For example 41.258543, -96.012878 is Durham Science Center, Omaha, NE 68132. These parameters, in our case would be variables that Dark Sky API requires to return a weather data. For example, if we want to know the weather in **Durham Science Center, Omaha, NE 68132** at a given time, we have to provide the **latitude,longitude** and **timestamp** to get the historical weather details.

A detailed and updated uses document can be found in the Dark Sky API docs <https://darksky.net/dev/docs>

Formatting our data to match Dark Sky API

Our `cab_rides.csv` dataset has useful columns that we can use to get the data from the Dark Sky website such as a `timestamp` in *epoch milliseconds*, `source` and `destination` meaning the physical location. With those parameters we could find the exact weather data. However, there were a couple of issues we needed to solve before we could get the weather data. We need to convert `source` and `destination` into the Dark Sky API readable `latitude` and `longitude` format and convert the `timestamp` from epoch milliseconds to epoch seconds.

We converted the `milliseconds` to `seconds` by dividing the `timestamp` column which is on `milliseconds` by 1000.

We solved the `coordinate` issue by separating out the unique location in the `source` column from our `cab_rides.csv` and got 12 location returned.

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

Table: The table containing unique **source** column.

The table containing unique ‘source’ column

Using the location from the table above, we used **Google Maps**(5) to get the corresponding **coordinates** and applied it to the **source** column. Since we are only examining the cost from the **source** for the weather analysis, we will not get the coordinates for the **destination**.

Now that we had all the required parameters to use it with **DarkSky API**, such as **latitude**, **longitude** and **timestamp** in epoch **seconds**, we were ready to start fetching the weather data. However, we had one last backlog that we needed to fix first. Our dataset had 693K rows but as we already know that **DarkSky API** only gave us 1000 API requests per 24 hours else we would need to pay for the extra API requests. Even though we request 1000 data points per day using the free 1000 available API requests, it would have taken us ~697 days to fetch all the data points and we actually do not need weather data for each and every minute or seconds. We could just go with the hourly weather data which should give us a good weather estimate. We fixed this issue by converting our **seconds** based timestamp into **hour** by creating a new column labeled **hour**. We also created additional column called **month**, and **day** so that we could group it by all three parameters and have unique hour for each day and location. We realize that while components of weather vary by infinitesimal time periods, the error of lumping the infinitesimal weather data into hour time differences will be absorbed by the fact that the difference in weather within hourly time frames is minimal. Our final **hour** data set was 332 rows vs the 693K rows, which was low enough for us to use the amazing **Dark Sky API** to fetch the corresponding weather data.

Now that we have the hourly dataset, we used the **Dark Sky API** to get all the corresponding weather values using their free 1000 API requests per day.

2.2 Data limitations

Throughout our research into the data we have identified a few qualities of our dataset that are not desirable. After looking further into the method of data accumulation, we learned that the author who compiled all the data for the original set could not acquire precise values for the price of each ride. In fact, the value supplied by the dataset is the estimate provided by the rideshare service, which itself can vary based on variables such as time and distance. The true cost of any ride is calculated at the end of the trip using those variables, and both services claim that prices will increase during heavy demand. All of this considered means we can only draw conclusions based on what each app *thinks* price for a given ride will be, not the actual cost of that particular ride.

This quality, unfortunately, undermines many potential analysis that we would want to make. Ideally, we would like each record to contain the data from a completed ride, and initially we believed this was the case. However, since each record is effectively just the information that the app feeds back when requesting a ride, but nothing further. Since this is the case, we based most of our analysis over the one bit of data that matters the most: the price. We can try to extrapolate potential factors that could affect the price that the application will return based on what data we have, but not much more than that can be analyzed.

However, another big issue with the data we later discovered was the timeframe. The amount of time encompassed by this data is effectively 2 weeks of winter. This is a huge pitfall, since the driving experience is greatly impacted by the changing seasons. This means that we cannot examine how cold weather might affect the price vs warm weather since each day represented by our data is only what would be considered cold. Not only is this an issue, but the time itself is not continuous, but this issue will be addressed below. In the future, we recognize the need to consider a wider range of time data for these types of analysis.

We also discovered a discontinuity in our time series data. We learned that data for December 5th thru 8th, and along with December 11th and 12th did not exist in the data set, which poses a problem when trying to visualize how time correlates with other data. We explored a possible solution through imputing usable data to help fill the gap in our time data.

We chose to use the bootstrap methodology in order to impute missing data. The first challenge was deciding which bootstrapping method to use. We could randomly select the entire data set, but we wanted an approach that maintained the integrity of our data. In our approach, we theorized that each week day should appear similar to itself in our dataset, e.g. data from Monday should look like every other Monday data in the dataset. So, first, we computed how many records each day had on average, call it x , then we grouped all of the data by weekday and bootstrapped x records from that group. We repeated this for each missing day, adding the bootstrapped data to the dataset.

The Figure 1 below displays the results of bootstrapping to impute missing data. The blue points represent what our data looked like before bootstrapping. The red points indicate the new records that were bootstrapped. It appears that the boot strap procedure was effective but not perfect, there are still a few gaps inbetween the timeline. Even still, boot strapping new data has repaired a giant hole in our time series by referencing similar records in our data.



Figure 1: Imputed Data. Blue denotes prior data while red denotes new, imputed data.

3 Methods

The rideshare data breaks down over 600,000 potential rides taken in Boston by characteristics including original location, destination, price, time, etc. In addition, weather data for the day each ride was taken was added to the original data set. The main focus of this report is to determine which factors most strongly correlate to price fluctuation. Ridesharing services, like most other services, charge based on supply and demand. It is known that when more users are looking for rides, the prices increase. Likewise, at times when fewer people are looking for rides, rides become cheaper since there is a surplus of drivers. Thus, in this study, we are looking for factors that might increase the number of people looking for rides. In addition, we are also looking for factors that might not fit into the simple supply and demand formula that might also be influencing pricing.

The data products and figures in this report are utilized in the analysis of each factor's effect on price. They are presented so as to give readers a full understanding of the conclusions that are later drawn based on the data. They in no way encompass every aspect that could possibly affect the price of a ridesharing service.

One factor that is closely analyzed in this report is geography. Geography can play a role in both the location of pickup and the destination. It is obvious that the higher the volume of users, the higher the prices will be for ridesharing. However, it is unclear whether the number of people looking for ridesharing services correlates with the total number of people at each location. Thus, we can analyze which types of locations (school, train station, etc.) typically have the highest prices for both departures and destinations. This is accomplished through maps of the area which illustrate the locations where prices are the highest. We can then determine if it is advantageous for a rider to walk to a nearby location for a better price.

The weather might also factor into the supply and demand pricing structure. This report explores when people are most likely to use a ridesharing service based on the weather. We can examine whether higher prices correlate with certain types of weather. We can visualize the pricing compared to weather patterns using graphs. Then, we can explain why pricing discrepancies exist during certain days.

Results

Geographic Data

In the figures below, the average cost per mile to go to each destination is displayed. In Figure 2, the destinations are represented by dots on the graph. The red dots signify destinations with higher costs while the blue dots represent the destinations with the lower costs. It is clear by looking at this data that South Station is the most expensive destination while Fenway is the cheapest. This could be due to a variety of factors. It could be that since South Station is a hub for both intracity and intercity travel, many people visit throughout the day. Along with this, those who are traveling to South Station typically don't want to pay for parking. Thus, many visitors would opt for ridesharing as they do not have to worry about leaving their cars while they travel. Fenway's cheap price can be explained by evaluating the time of the year the data was collected. Since the data was collected in late November, there likely were no baseball games occurring. Thus, there was little demand to go to the stadium.



Figure 2: Cost per Mile for Each Destination. This figure shows the average cost per mile to go to each destination. The averages are calculated using all source locations for all destinations. The more expensive locations are in red.

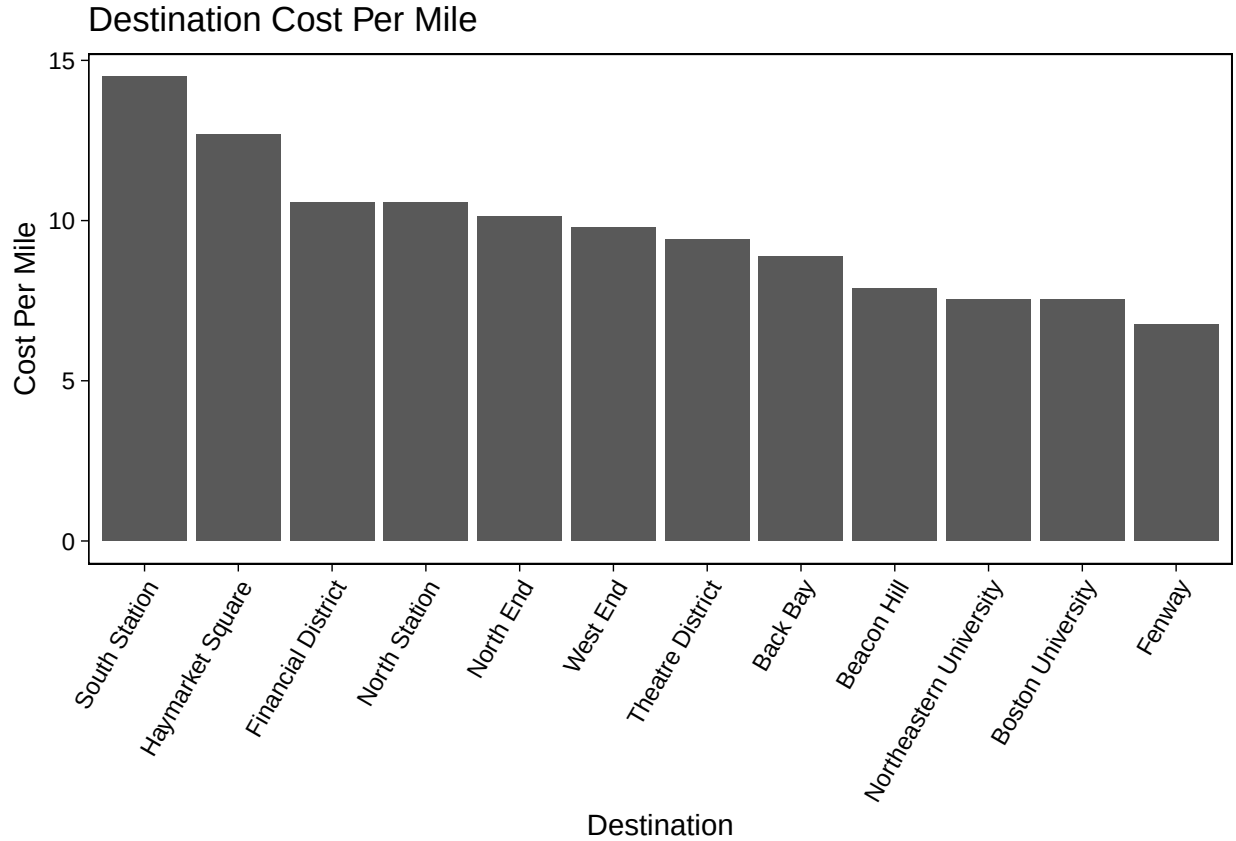


Figure 3: Cost per Mile for Each Destination. This figure is a different representation of the same data from Figure 2. This shows the numerical values for the average cost to travel to each destination.

Figure 4 shows the average price from each source to each destinations. Some trips from a source to a destination appear to have abnormally large costs per mile when compared to other trips. For example, the average price to go from the Financial District to South Station is much higher than other destinations. The same is true for the opposing trip. However, it is not true that all trips to or from the Financial District have abnormally large costs per mile. This observation could be caused by excess demand and distance for these trips as we would expect trips with higher demand and higher distance to have a higher cost.

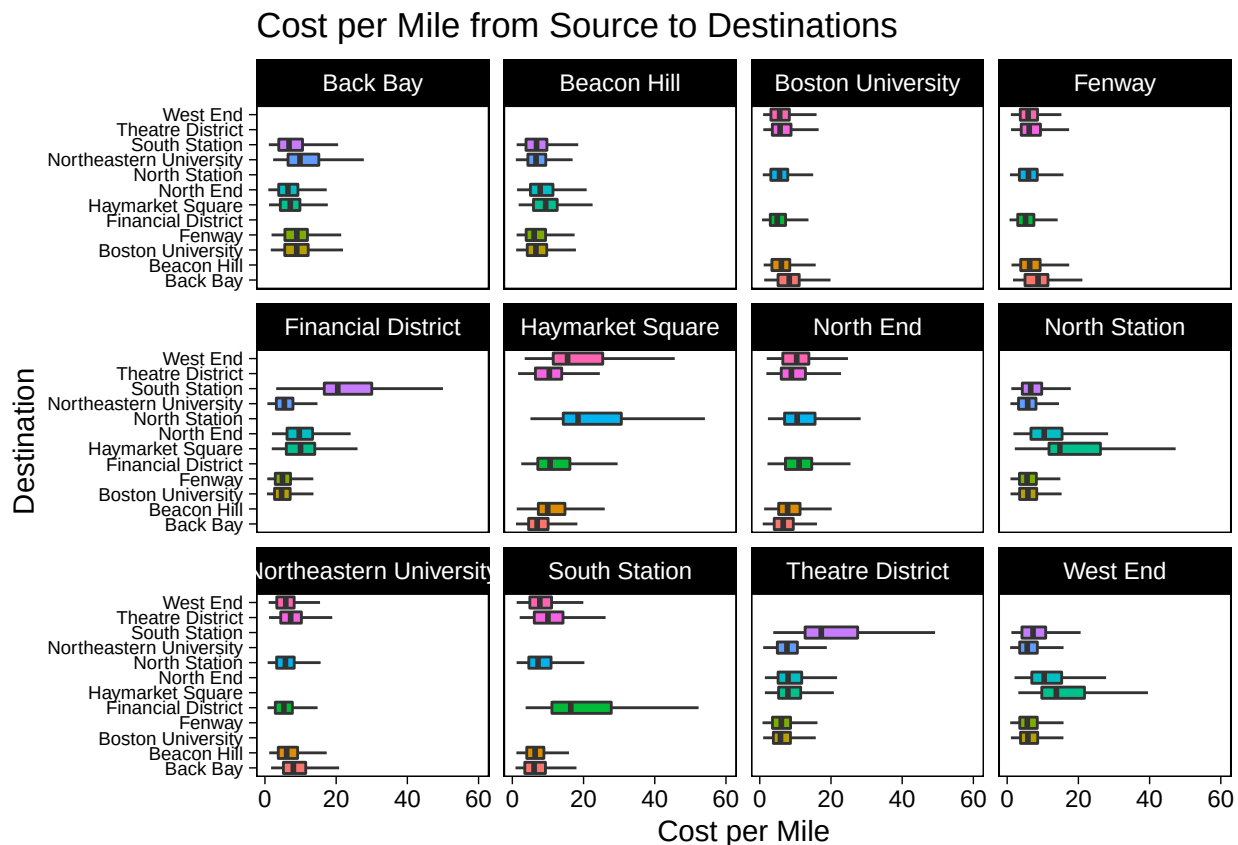


Figure 4: Cost per Mile from Source to Destination. This graph shows the minimum, first quartile, average, third quartile, and maximum cost from each source to (the title of each individual plot) to each destination. The price of the ride greatly depends on where the ride begins and where the ride ends.

Figure 5 below shows the density of prices per mile from each source to each destination. Sources are shown in green and destinations in red. The fact that not all sources have data for all destinations is clear here. Using this visual, it is determined what is the most concentrated price per mile from each source to each destination. Higher peaks represent higher concentrations. It appears that Boston College, as both a source and destination, has high instances of ride prices around \$5-\$10 per mile. Conversely, the prices per mile from any source to North End are not concentrated to a clear range of prices, but rather spread over an extended range of prices.

`## Warning: Removed 567 rows containing non-finite values (stat_density).`

Density of Prices per Mile from Source to Destination

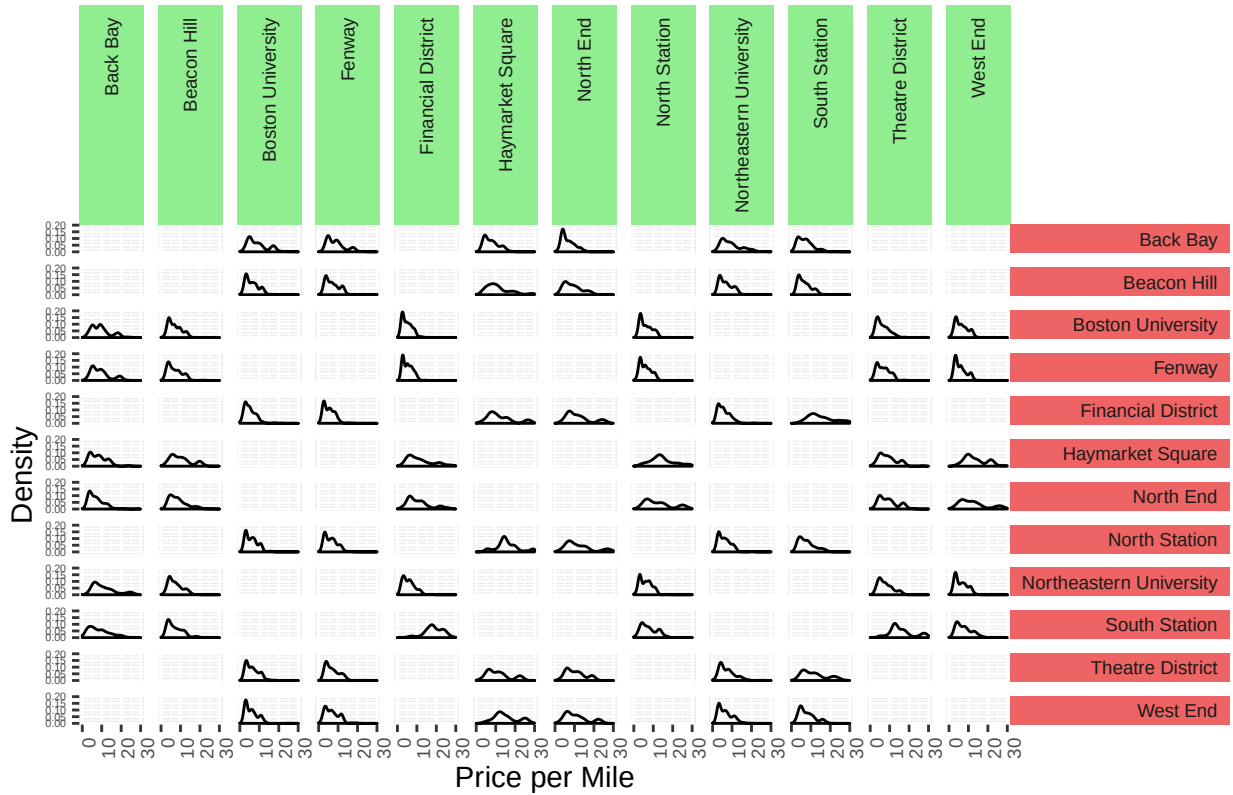


Figure 5: Density of Prices from Source to Destination. This figure depicts the density of prices from each sources (green) to each destination (red). The most common price range is between 0–20.

Cab Type

Each service provides different cab types, each with unique costing and benefits. Both provide options for large groups, car pooling, luxury vehicles, and basic ride services. Below is a table comparing the various cab types provided by each service and the average price per mile for each. Here, we can make the determination that Lyft services are typically cheaper when compared to Uber’s pricing. Two cab services were omitted due to their presence in only one of either Uber or Lyft. The two services were WAV, Uber’s wheelchair accessible ride service, and Lux, which was another Luxury offering from Lyft. Each service provided options for car pooling (at a discount compared to the standard rate), the standard service, the luxury service (in this table both are luxury *black* vehicle services), and group services which allowed up to 6 passengers in one vehicle.

Service	Carpool	Luxury	Luxury Group	Standard	Standard Group
Lyft	\$3.30	\$12.76	\$18.77	\$5.37	\$8.30
Uber	\$5.34	\$12.37	\$19.43	\$5.94	\$9.13

Table: A comparison between like services offered by both services

We can glean some further information on how the prices for each cab type compare to each other in the graphic below. It contains a histogram distribution of a sampling of pricings for each cab type. It also displays two colored vertical lines. The red vertical line represents the average price of the entire sample, while the blue vertical line measures the average price for that particular cab type.

Histogram of Price/Mile for each Service Offered

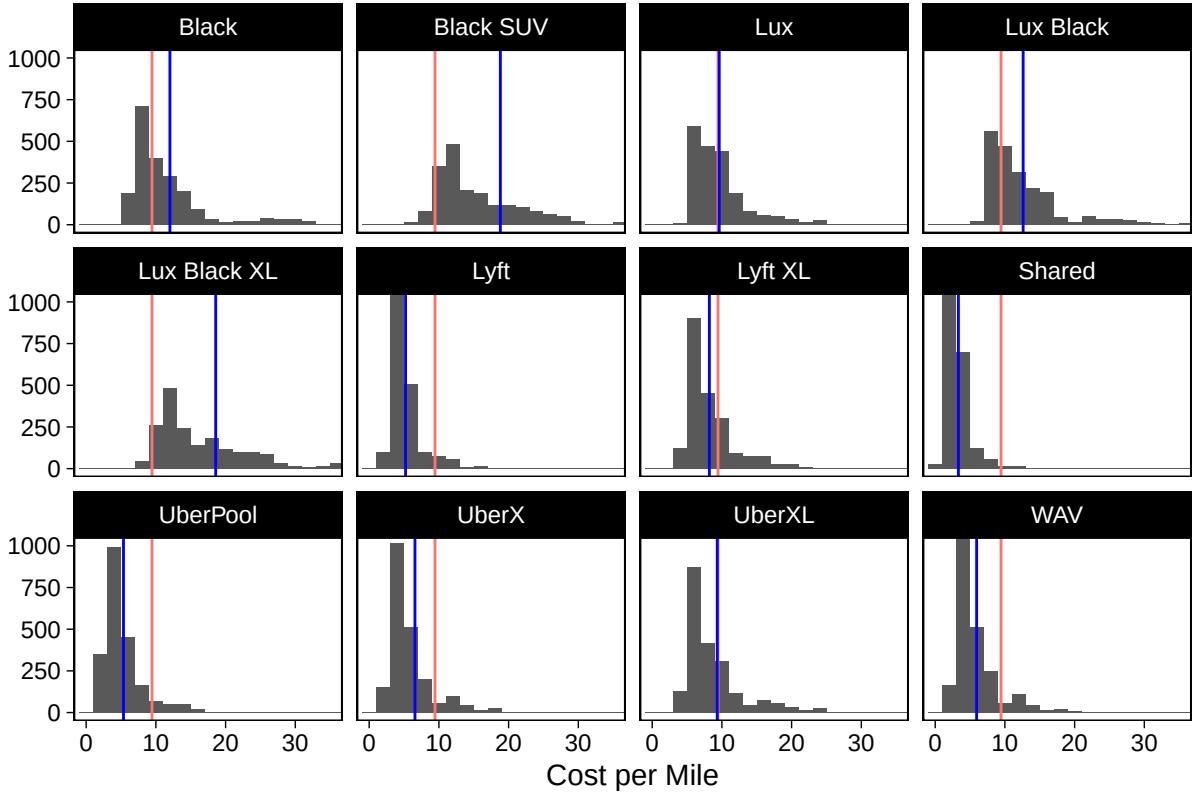


Figure 6: Price per Mile by Cab Type. The red verticle line indicates the average price per mile of the entire data set. The blue verticle line indicates the average price per mile for that particular ride service.

We notice that luxury cab types tend to have a longer tail when compared the more economy versions, which is what we would expect since a luxury cab should be on average more expensive. We also notice that each graph has a large spike of data initially, and then a general downward trend. This is likely due to a minimum fee that both services charge people, so if there are a significant amount of short distance ride requests then it would make sense that we would see such a high spike in record count towards the beginning. The advantage of ride sharing is also made apparent by this data as shared rides are significantly cheaper on average.

Time Analysis

It is interesting to know when people used Uber and Lyft services how much was the average price per mile in any hour of a day and compare the average price per mile for Uber and Lyft in each hour. To compare the average cost per mile between each service in each hour of a day, we generated a plot according to the average price per mile that passengers paid in each hour of a day. However, our data are included just 7 directions in Boston and it is just for 2 months (November and December), so our data is not completed and all measurements and conclusions are related to our data. With data visualization, we can understand how the average price per mile rates throughout the hours. We observed in figure 7 that during a day the rates of the average cost per mile are depended on the hours. So figure 7 shows small fluctuation for the rates of average price per mile during a day for Lyft. On the other hand, figure 7 presents a large fluctuation for the rate of average price per mile around 5 PM and between 7 PM to 11 PM for Uber. The rates of average price per mile was changed just \$1 during all hours of a day for Lyft, but it was changed about \$2.5 around 5 PM and about \$3.5 between 7 PM to 11 PM for Uber and the maximum rate for average price/mi was happened at 9 PM (\$11.25) for uber. Therefore we can conclude that the price/mi was more for Lyft from 10 AM to

6 PM than Uber, and more passengers used these services in December. We also checked the number of trips during each weak day of the month. As it is clear from the resulting visualization that Monday had the highest trips for both services.

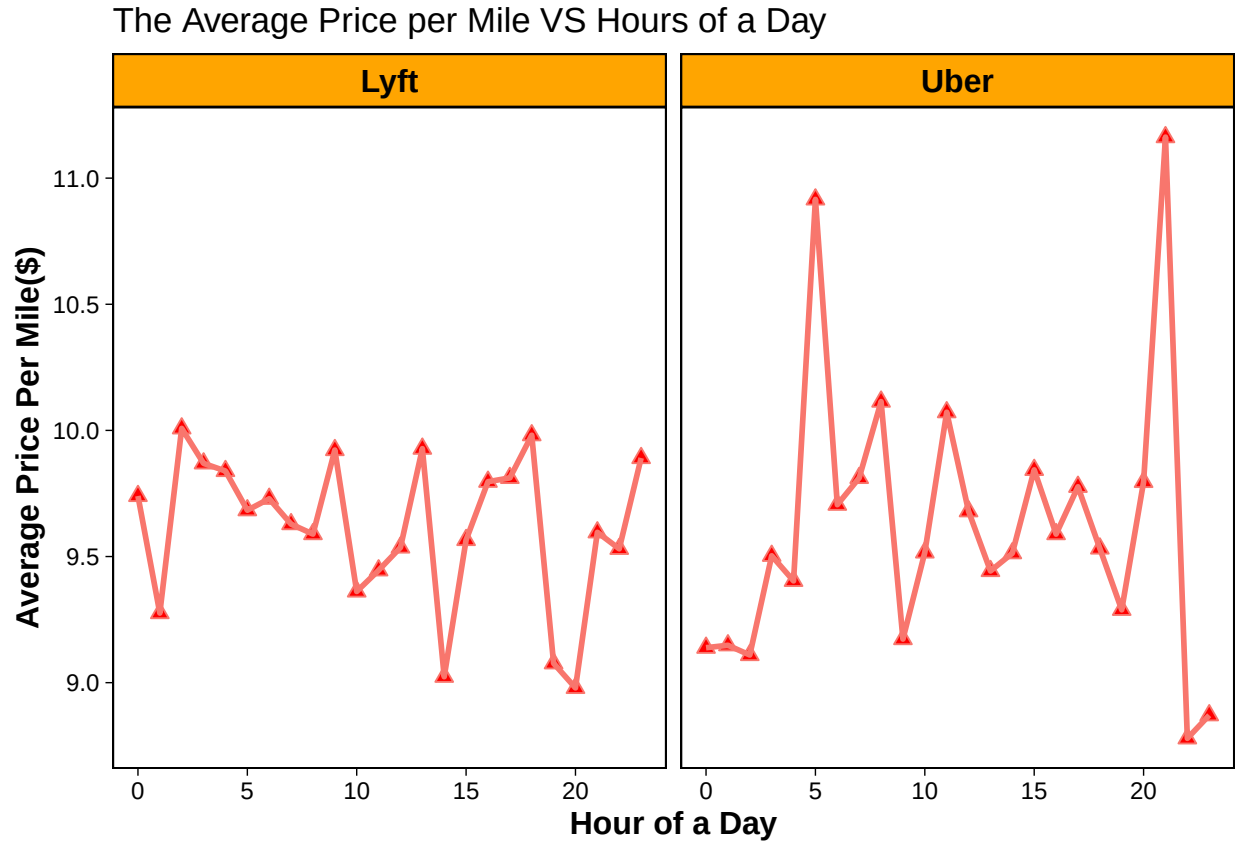


Figure 7: Average Cost Per Mile for Each Hour of a Day by Cab Type. This figure shows fluctuation for the numerical value of the average price per mile during all hours of a day by Cab Type. The X axis is demonstrated each hour of a day and Y axis is demonstrated the average cost per distance to travel to each destination.

In this part, we compared different data values for each service versus each hour of a day by month. Figure 8 and 9 shows the relation between the average price per mile and total cost with the average cost and the total number of trips in each hour. Figure 8 demonstrates the relation between the total cost that passengers paid and the total number of trips, which were happened in each hour of a day in two months (November and December). According to the results in figure 8, the total number of trips in December is bigger than November, which could be, because December is the last month of a year and many events happened in this month and also the weather is colder in December. Therefore, more people needed to use Uber or Lyft service in December. However, it is clear in these plots that the number of trips with Uber is more than Lyft but the total price that people paid to the Lyft is more than Uber. According to these results, we can see the maximum of the total cost that passengers paid to Uber were around \$140000 at 1:30 PM in December, but the maximum of the total cost that passengers paid to Lyft were around \$150000 at 2 PM in December. So we can say that Lyft service is more expensive than Uber and Uber is more popular than Lyft.

Figure 9 represents the average cost per mile versus the average cost that passengers paid for each service (Uber and Lyft) in each hour of a day, by different months (November and December). In this plot, we would like to understand the relation between the average price per mile and the average cost. Figure 9 shows that Uber has more fluctuation in the average of the price and the maximum value for the average of the price in

Uber is more than maximum value in Lyft.

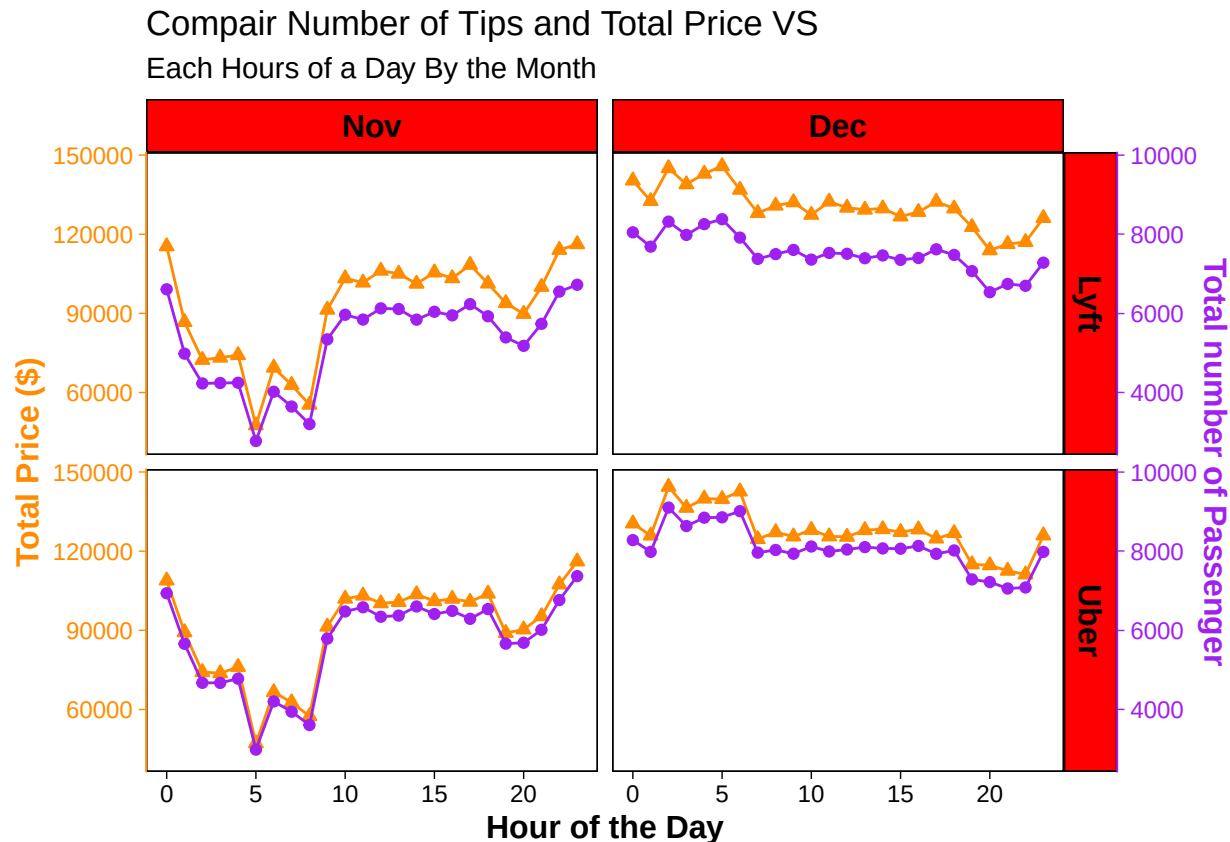


Figure 8: Comparisson the Number of Passenger and Total Price that Passengers Paied in Each Hour of a Day in the different Month by Cab Type. The orange line indicates the total price that passenger paid in each hours a day for November and December. The purple line indicates the total number of passenger that they used each service in each hours of a day for November and December. The Y axis in right demonstrate the values of number of passenger and Y axis in left demonstate the values for total price per.

Compair Average Cost Per mile and Average Cost in Each Hours of Day By the Month

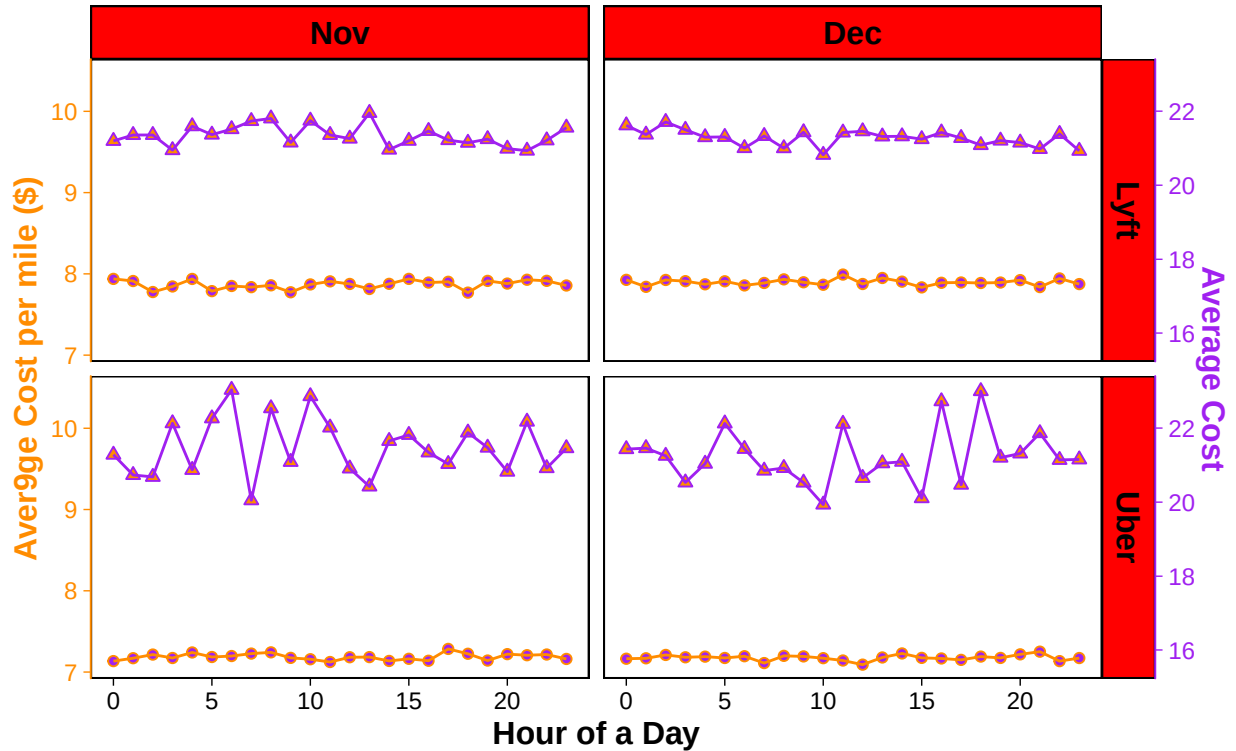


Figure 9: Comparisson the Average Price per Mile and Average cost that passenger paid to The Uber and Lyft in Each hours of a Day and facet By Months. Total Price that Passengers Paied in Each Day of a Week in the different Month by Cab Type. The X axis demonestrate each hour of a day. The Y axis in right demonestrate the average cost that passenger paied and Y axis in left demonestrate the average price per mile.

In following plot we cheked the total price that passengers paied in each days of a week in different month (November and December). In this plot we would like to understand that is the total paied price is deppended to the day of a week. Accourding to this plot the total price that was paied to each service is more on Tuesday in November however these services made more money on Saturday, Sunday and Monday in December. Also Uber made less money on Saturday in December, which could be related to the difference of price between Uber and Lyft.

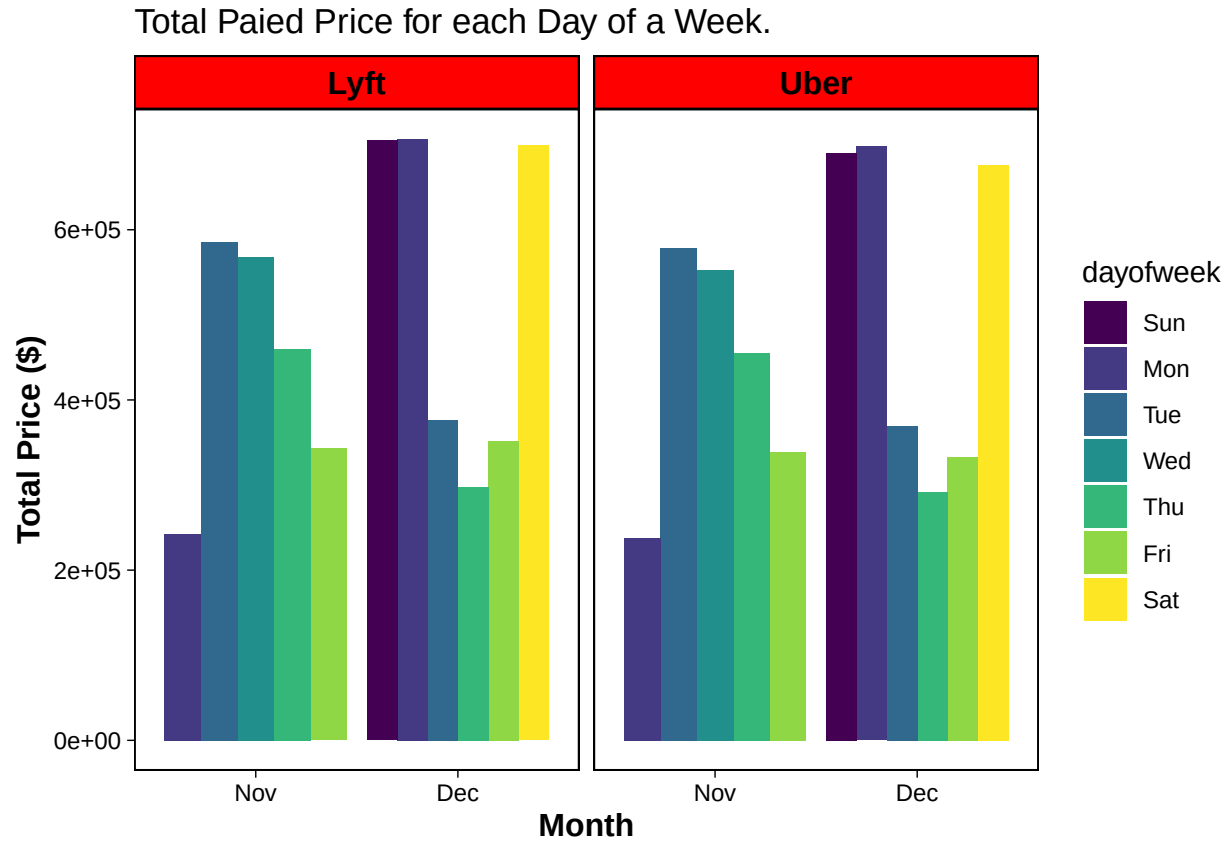


Figure 10: Comparisson the Total Price that Passengers Paied in Each Day of a Week in the different Month by Cab Type. The X axis demonestrate the month (November and December). The Y axis in demonestrate the values for total paied price. Each color in legend demonestrate a day of a week.

In this psrt we checked the number of trips to each destination in our data for each service. As the results shows higher number of trips went to the Financial District, which is the downtown's business nerve center in Boston and crowded with stately banks, high-rise offices and sleek condos. So it could be make sence that total trips to the Financial District destination be more than other. Then we tried to check the number of trips by Uber and Lyft together to each destination in each day of week, to see which destination has more trips in each day of week. This result shows that more trips went to Back Bay on Tuesday.

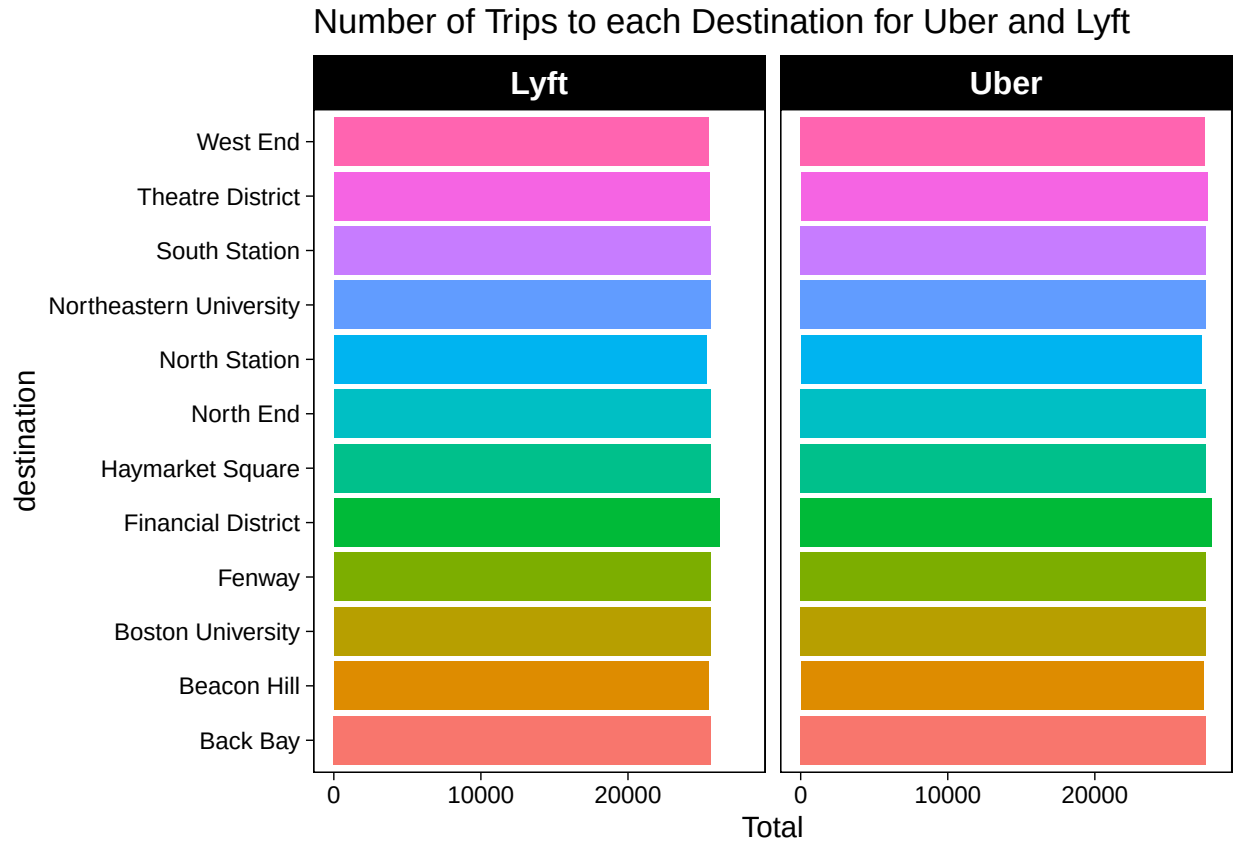


Figure 11: The Total Number of Trips to each destination in Boston by Cab Type. The X axis demonstrates the total number of trips. The Y axis demonstrates all destinations in Boston that we have in our data.

Figure 11 demonstrates the average price per mile for each destination in different days of a week. As it is clear in this plot, the highest value for average price per mile was for South Station and it was on Monday. South Station is the historic train terminal, opened in 1898, now serves thousands of commuters, travelers, shoppers, and diners each day. South Station is a transfer station on the MBTA rapid transit Red Line and bus rapid transit Silver Line, located at Summer Street and Atlantic Avenue in downtown Boston, Massachusetts. So it makes sense that on Monday everyone would like to go to their job, school, shopping etc. therefore more people need to get Uber or Lyft to go to the transfer station, and whenever some area of a city is busy Uber and Lyft are multiplied a constant value to their price. Also the largest value for average price per mile for Boston University was on Monday too, because Monday is the first day of the week and more students use these services. Haymarket Square is the historic name of a former town square in Boston, located between the North End, Government Center, the Bulfinch Triangle, and the West End. It is downtown Boston and there is many bars and public markets and etc. there. Also according to our results the maximum average price per mile in Haymarket Square was on Saturday, which more people go to the downtown to have fun.

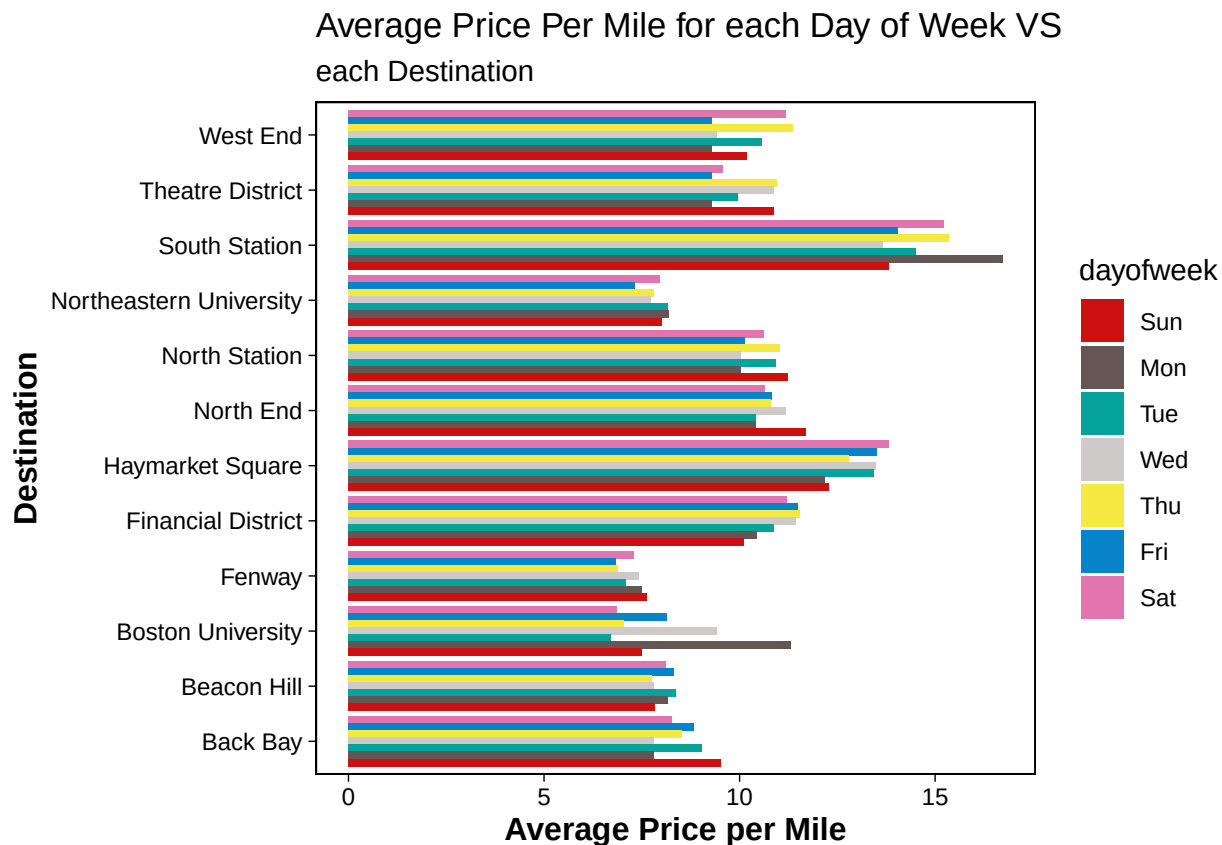


Figure 12: Comparisson the Average of Price Per Mile that Passengers Paied in Each Day of a Week in the different Month by Cab Type. The X axis demonstrate the month (November and December). The Y axis in demonstrate the values for total paied price. Each color in legend demonesterate a day of a week.

Weather

We looked into the following 4 questions with the weather data to examine if weather do impact on pricing the ride

- Cost comparision with weather types, such as Sunny, Foggy, Overcast, Rain, etc
- Cost comparision with specific to high/low windy days
- Cost comparision with specific to rain
- Cost comparision with specific to temperature

Cost comparision with weather types

In this question, we are examining how would various weather types affect the price of average rides taken during those time.

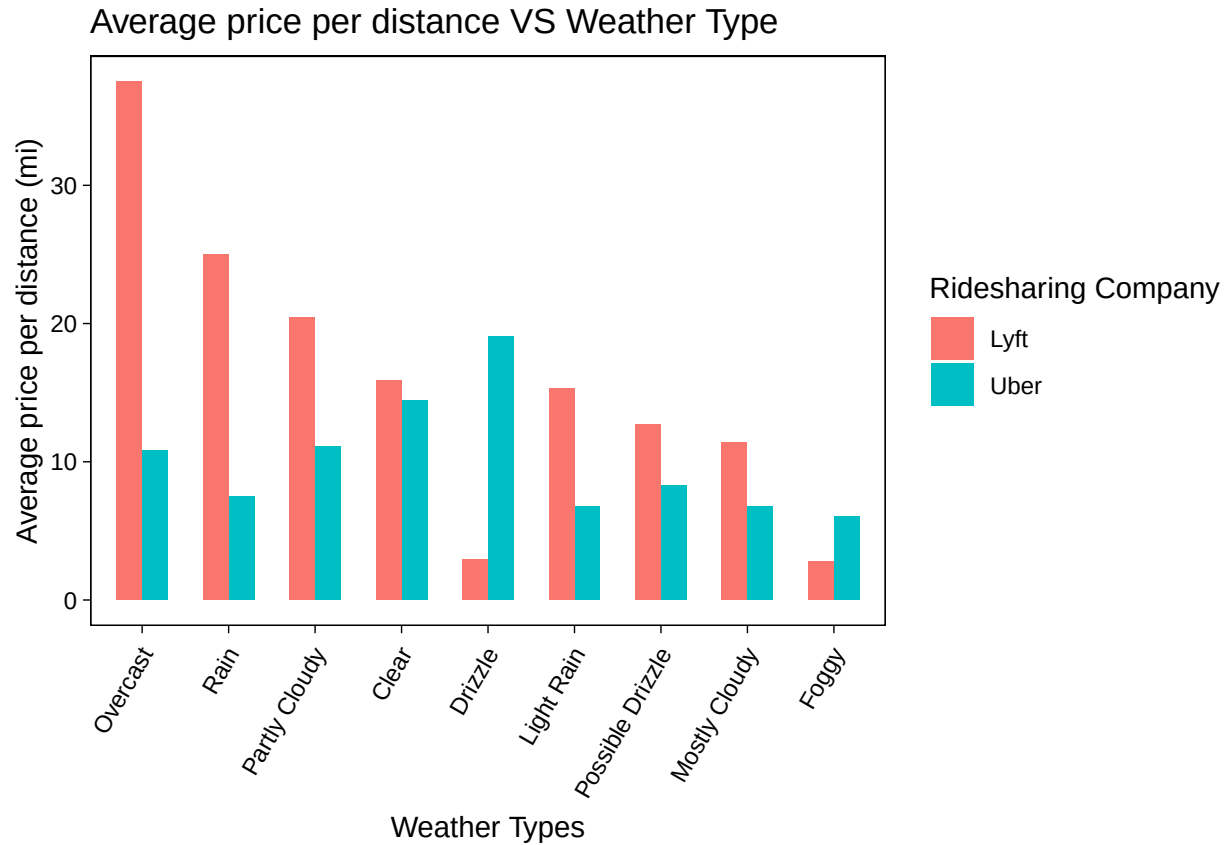


Figure 13: Average cost vs different weather types

In the figure above we are calculating the **average cost** which is price per distance and the **weather types** for **Lyft** and **Uber** and it turns out that **Lyft** is on average expensive than **Uber** during Overcast and also cheaper during Drizzle and Foggy weather. Uber seems to have mostly uniform pricing over different types of the weather.

Cost comparison with specific to high/low windy days

In this question we wanted to find out if the price changes during high and low wind gust.

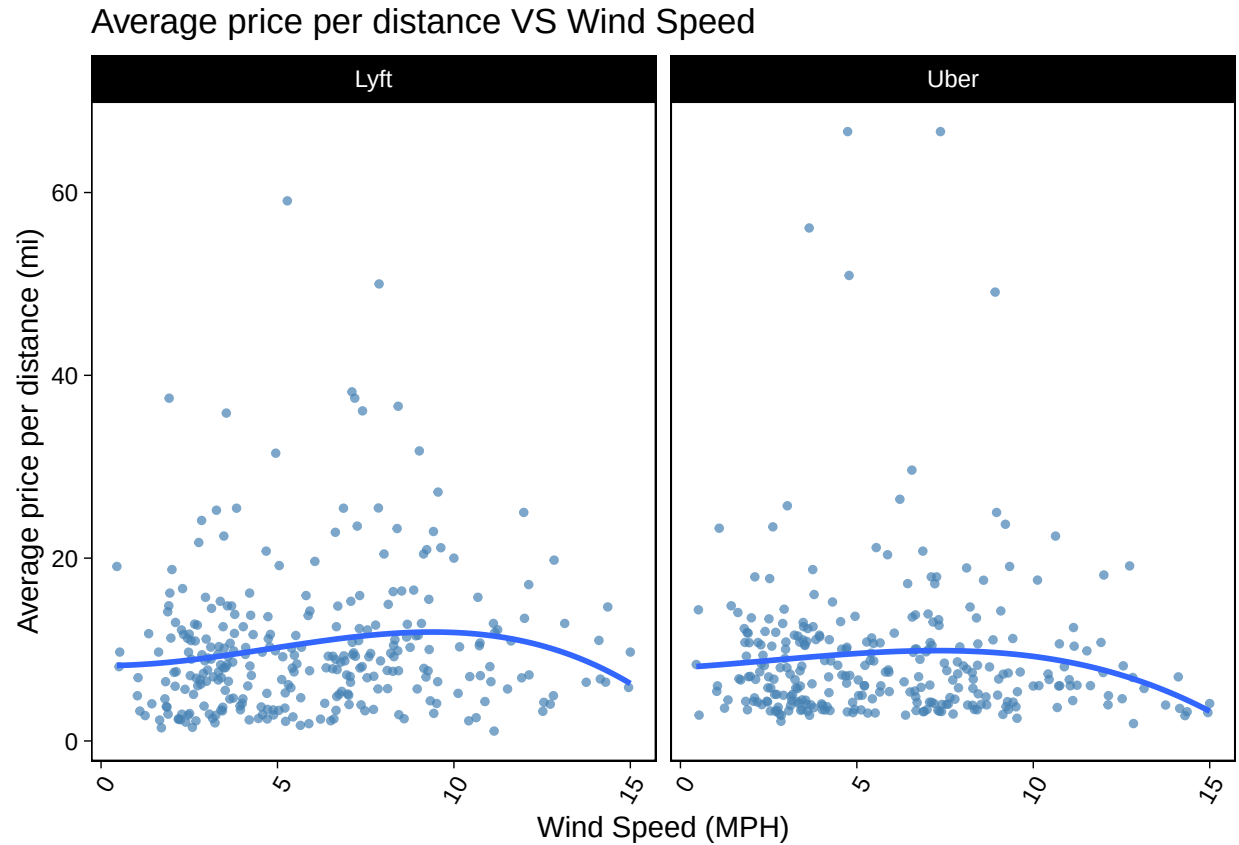


Figure 14: Average cost vs different weather types

Lyft and Uber have the similar price upto around the wind speed of 5MPH but Lyft starts to get expensive with wind speed but the decreases if its more than around 12 MPH. Uber seems to have consistent price through out the wind variation but similar to Lyft, the price per distance is decreasing with the increase in MPH.

As per <https://weatherspark.com> (6) the average wind in Boston, MA around November - December is between 6 MPH to 10 MPH

Cost comparision with specific to rain

In this question we are examining if rain can factor in increase or decrease in price for ride requests.

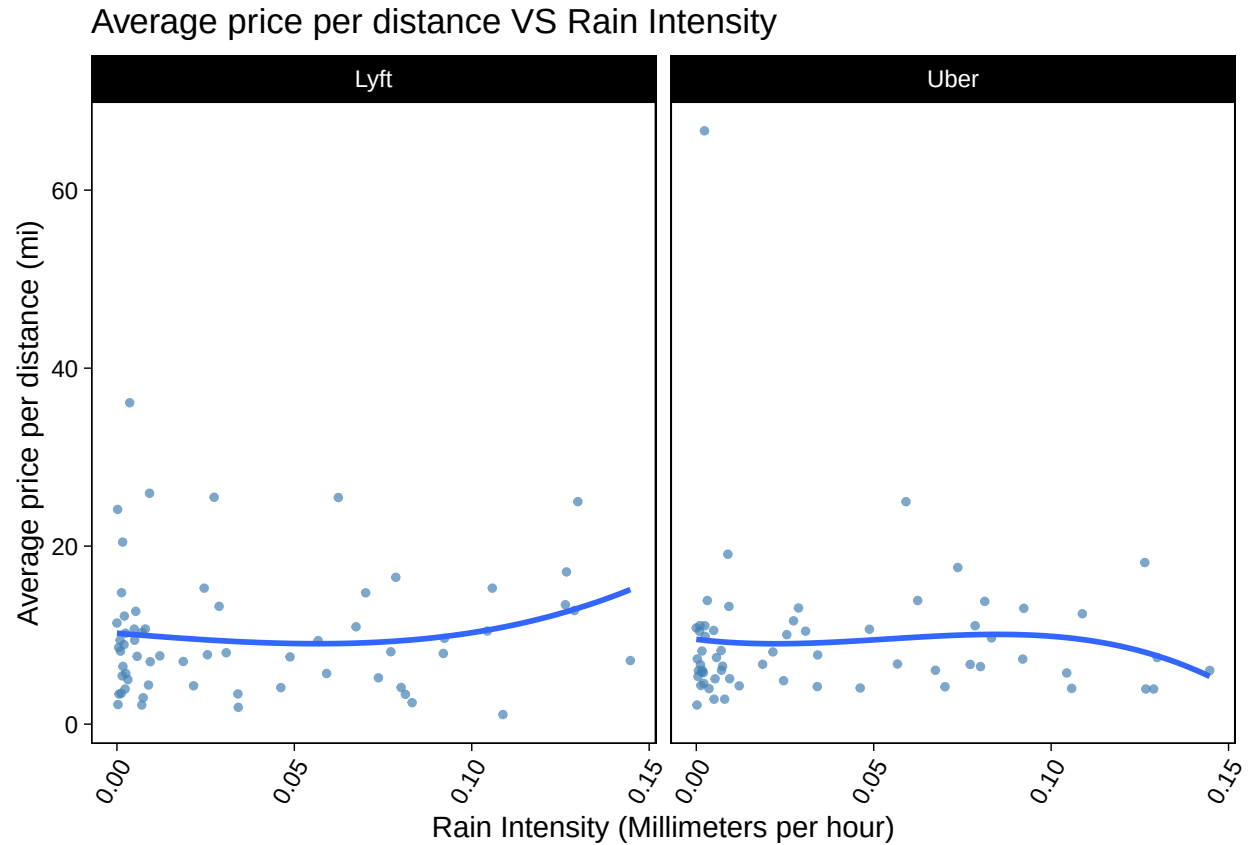


Figure 15: Average cost vs different weather types

Lyft get expensive as the Rain intensity increases while **Uber** seems to lower the price with rain.

Cost comparison with specific to temperature

In this question, we are examining if **temperature** can factor in increase or decrease in price for ride requests.

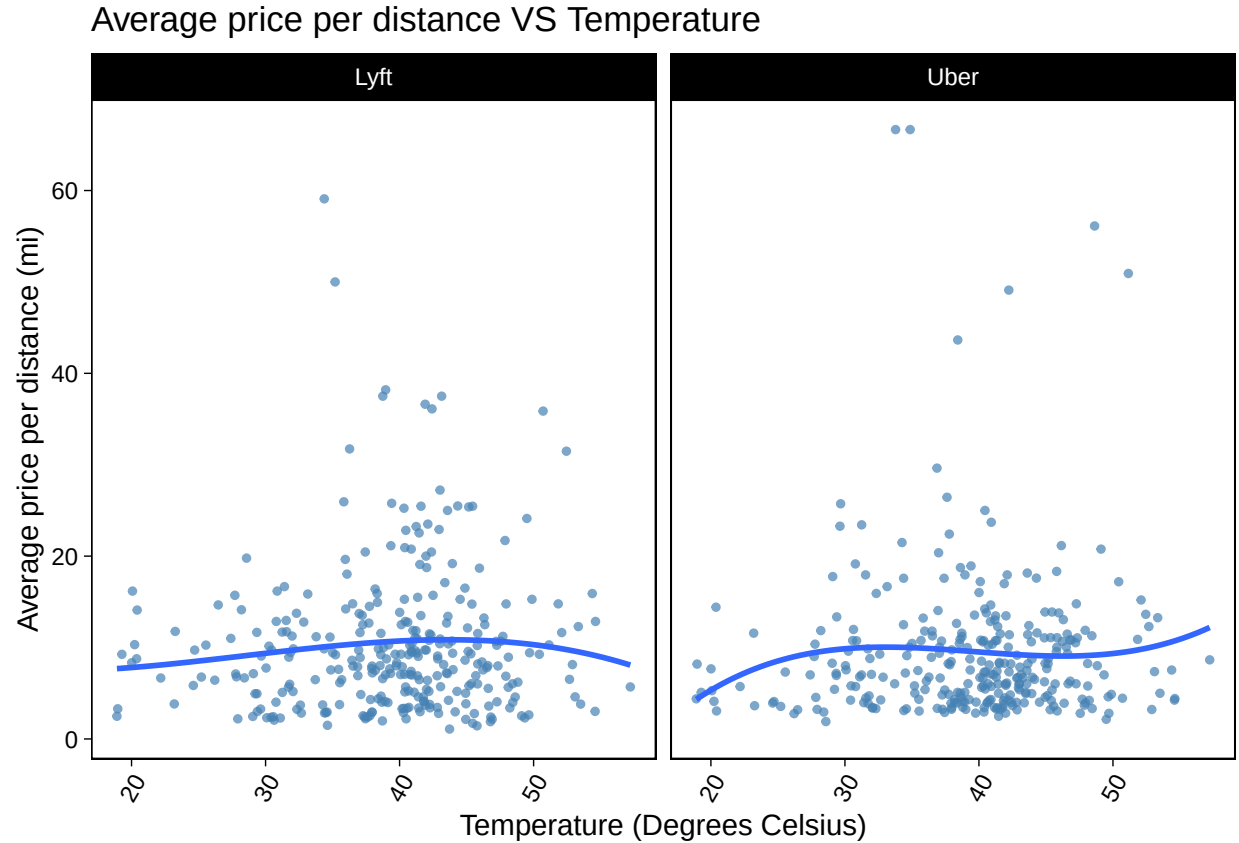


Figure 16: Average cost vs different weather types

Both **Lyft** and **Uber** price increases with increase in temperature and decreases with drop in temperature. However the price of **Lyft** drops after 55 degree but **Uber** keeps increasing as far as our dataset goes.

Weather Linear Model

In order to determine the significance of each weather factor on price, a linear model was created using three variables: temperature (X), the probability of precipitation (Y), and wind speed (Z). The following equation was obtained.

$$Price = 9.487 + 0.0063X + 0.1225Y + -0.011Z$$

As is shown in the model, none of the variables have a large impact on the price. It would appear initially that the probability of precipitation has the largest impact on the price. However, since the probability is a value between 0 and 1, the most the price would change due to this variable is \$0.12 per mile.

Conclusion

As ride sharing continues to gain momentum, it is important to explore factors affecting the experiences of users and drivers. The most notable factor that affects users is price. Both Uber and Lyft have varying prices that can fluctuate depending on the time of day, location, etc. For this report, data gathered from Uber and Lyft rides in Boston is analyzed to explore these possible relationships. The results of this study can give insights into how the ride sharing companies price their services which can be beneficial to both the users and the drivers.

The geography of the area is a great determination in price. It is clear that there is a discrepancy between destinations in price. This correlates with the theory that ridesharing companies charge more to travel into busy areas. The destinations with the larger demand may have a higher premium. This finding was also quantified in Figure 4, where it was observed that the price per miles was not consistent from trip to trip.

The type of cab type will have an effect on the expected price. When comparing like services to like, Lyft typically provided cheaper services when compared to Uber. Our data also confirms that using group cab types significantly reduces the cost per mile on average.

Time and date have a direct effect on price. Uber has large fluctuation on price depending to the time and date. Price for both Uber and Lyft is more on December because of the weather and holidays. Also price will increase during weekend at downtown and on Monday at university area, business area and transportation area. Because these places will be more busy in weekend and Monday, so Uber and Lyft multiply a constant value to the price and increase the price to get motivation to the driver.

Further research of this subject could include additional variables that may affect pricing a demand of rides, such as traffic patterns, and noting popular events for days and times within the time frame (i.e. baseball games, musicals, etc.). If possible we would like to fit our findings to a model to demonstrate the various relationships we've found. We hope that this research will prove useful in further predicting demand for rideshare service, guiding the driver to be in the right spot at the right time and the consumer to decrease the cost of a ride.

References

- 1: RaviMunde | Kaggle. Uber & Lyft Cab Prices. <https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices#weather.csv>
- 2: Munde, Ravi. Ravi72munde/Scala-Spark-Cab-Rides-Predictions. 2018. 2019. GitHub, <https://github.com/ravi72munde/scala-spark-cab-rides-predictions>
- 3: "Kaggle." Wikipedia, 1 Dec. 2019. Wikipedia, <https://en.wikipedia.org/w/index.php?title=Kaggle&oldid=928690534>.
- 4: "Weather." | Dark Sky - About. <https://darksky.net/about>
- 5: "Google Maps." Google Maps, <https://www.google.com/maps>
- 6: The Typical Weather Anywhere on Earth - Weather Spark <https://weatherspark.com/y/26197/Average-Weather-in-Boston-Massachusetts-United-States-Year-Round>