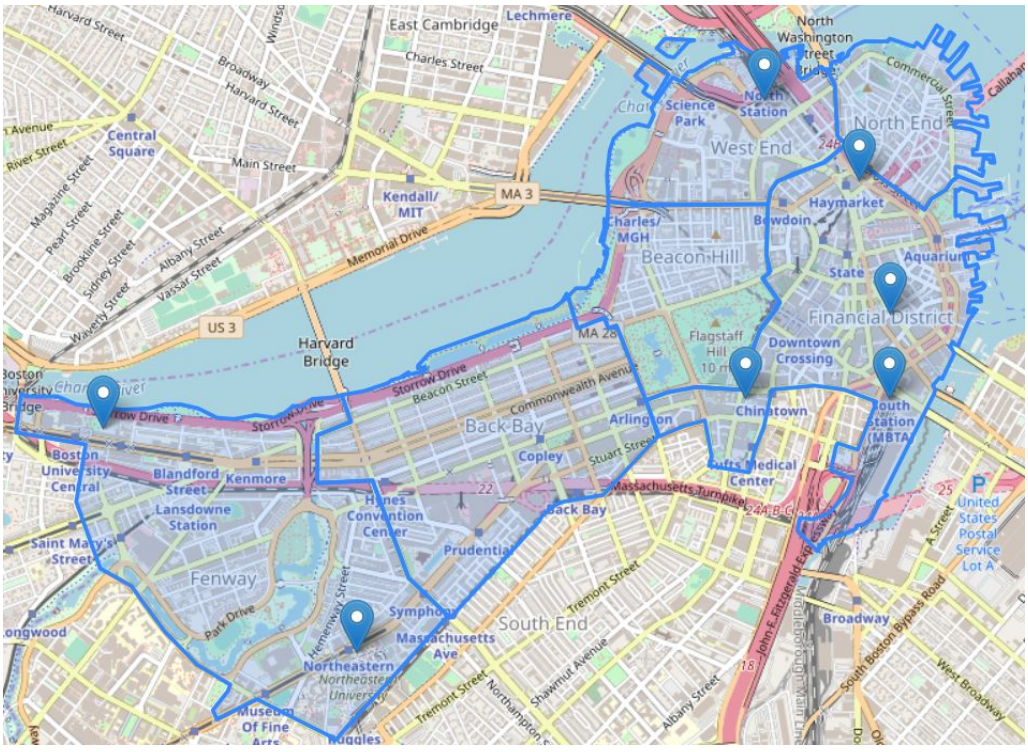
Milestone Report

The purpose of this project, for the prediction of a price of an Uber or Lyft ride based on a variety of predictors including weather data and other ride information, is a supervised regression problem (1), because the target variable is quantitative. Predicting how the price is determined, or what promotes the surge-multiplication of rides (resulting in higher than normal ride prices) would be valuable to clients such as Uber and Lyft as the purpose of this project is to understand what factors drive the demand and supply of rides due to differing conditions such as time, distance, and weather. Other ride-sharing companies or cab companies may use this information to rival Uber and Lyft by providing discounted prices for their rides to increase ridership.

The data was collected using Uber and Lyft API queries (ride data collected in time intervals differing by 10000000 epoch units, or roughly 1 hour, 30 minutes apart), containing 10 features including: ride application used (Uber, Lyft), distance between pickup and dropoff, timestamp of ride, pickup location of ride and destination of ride located in twelve areas around Boston (five main neighborhoods, and seven smaller areas as seen by the map below), price of ride, surge multiplier of ride (over much price was increased), specific type of ride (e.g. Uber Black, Lyft Lux XL), and corresponding id is included. There are over 690000 ride data instances recorded for this dataset. Weather data, containing shared features with the ride dataset such as the location of the ride pickup, at timestamps ranging between November 26 to December 18, 2018, contains unique features such as rain, clouds, humidity, wind, and pressure measurements. All datasets acquired will be combined with the ride and weather data into a csv file. Ride and weather data is provided from: https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices and was relatively clean.

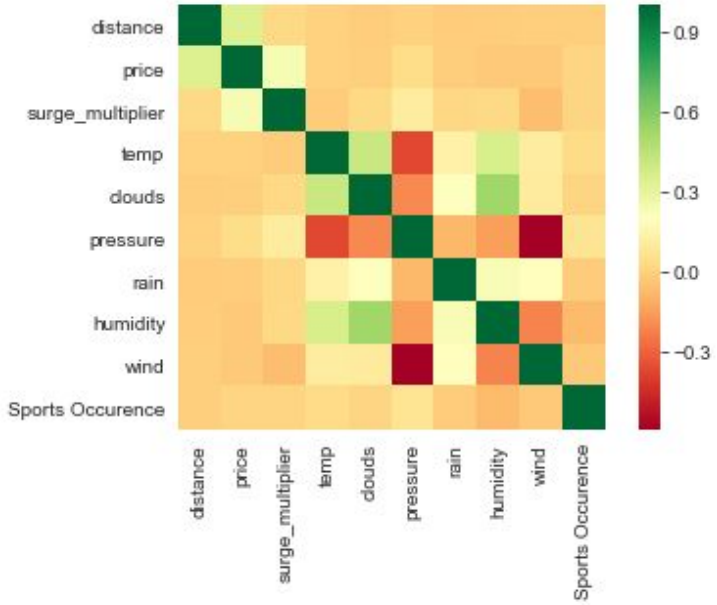


To prepare for the data wrangling process, both the ride and weather data, from Kaggle, were saved as a local copy, in csv file format, for reference. The ride and weather data were read in, using the pandas package in Python, as two separate dataframes. The timestamp values, or columns, for the rides and weather dataframes were in epoch time format, where ride timestamps were in milliseconds and weather timestamps were in seconds, so they were converted into datetime objects. Both dataframes were then sorted by their timestamps, in which the column name of the location in the weather data was renamed to 'source' to match the source column (representing the pickup location of the ride). The rides dataset contained approximately 50000 missing values for the price column (representing the target variable price of the ride). All of these instances with the missing price values come from an Uber ride type: Taxi. Because there is not any recorded price for the Taxi ride type, these ride instances were removed from the dataframe.

Once all of the missing values were filled in, the weather dataframe was then merged, using the merge_asof method in pandas, into the ride dataframe, matching weather rows to ride rows by the 'source' columns on the 'time_stamp' column. A timedelta tolerance of one hour is specified, which will match weather data occurring within one hour of the ride pickup time. The merged dataframe now contains missing weather column values due no weather data occurring within 1 hour of the ridetime. For example, multiple ride instances on November 26th, 2018 occur at 2:30 A.M., whereas the weather data begins at 3:40 A.M on the same day, yielding instances on this day with null weather values. To deal with this, these null weather value ride instances are dropped, for a final dataframe containing 633,296 ride instances and their appropriate weather data appended. A local copy of the dataframe is saved by creating a csv file containing all of the instances.

A dummy variable column (2) is created representing an occurrence, either 1 for 'did occur', or 0 for 'not occur' of a sports game, such as an NHL game by the Boston Bruins, or an NBA game by the Boston Celtics, occuring at the TD Garden stadium, located in the vicinity of North Station. The ride must have occurred within two hours of the game starting, or two hours of the game stopping, and have a source or destination located in North Station.

To understand what independent variables are highly correlated with the target variable, price, or with other independent variables to handle multicollinearity, we compute the correlation coefficient for each pairing of variables. The following heatmap of correlations for all quantitative variables in the dataset, and correlation coefficient table of values for the independent variables (not price) are shown below:

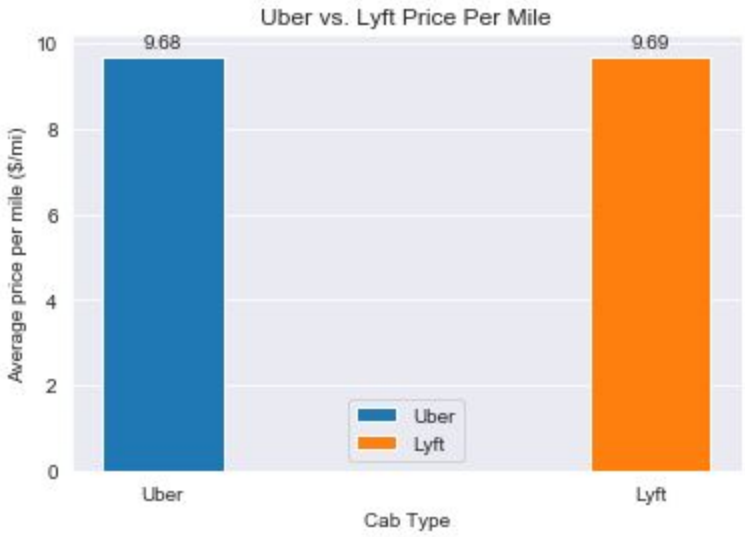


feature	distance	surge_mult	temp	clouds	pressure	rain	humidity	wind
distance	1	0.026	0.003	-0.006	0.002	-0.012	-0.011	-0.003
surge_ mult	0.026	1	-0.013	0.028	0.112	0.022	0.031	-0.059
temp	0.003	-0.013	1	**0.414**	**-0.367**	0.135	**0.364**	0.113
clouds	-0.006	0.028	**0.414**	1	-0.202	0.207	**0.533**	0.101
pressure	0.002	0.112	**-0.367**	-0.202	1	-0.079	-0.146	**-0.589**
rain	-0.012	0.022	0.135	0.207	-0.079	1	0.235	0.205
humidity	-0.011	0.031	**0.364**	**0.533**	-0.146	0.235	1	-0.213
wind	-0.003	-0.059	0.113	0.101	**-0.589**	0.205	-0.213	1

According to the heatmap, distance is positively correlated with the price of the ride, as well as, surge-multiplier, which is intuitive, as the greater the distance of the ride results in a higher priced ride, and the lack of supply of rides or rides at a certain time of day, causing a surge-multiplier > 1.0, causes an increase in the price of the ride. A few of the weather features strongly correlate with other weather features, such as the negative correlation with temp and pressure, wind and pressure. A moderate positive correlation exists with clouds and humidity, temp and clouds, temp and humidity. Looking at the actual values, coefficients values in bold display either a moderately positive or negative linear relationship between their respective

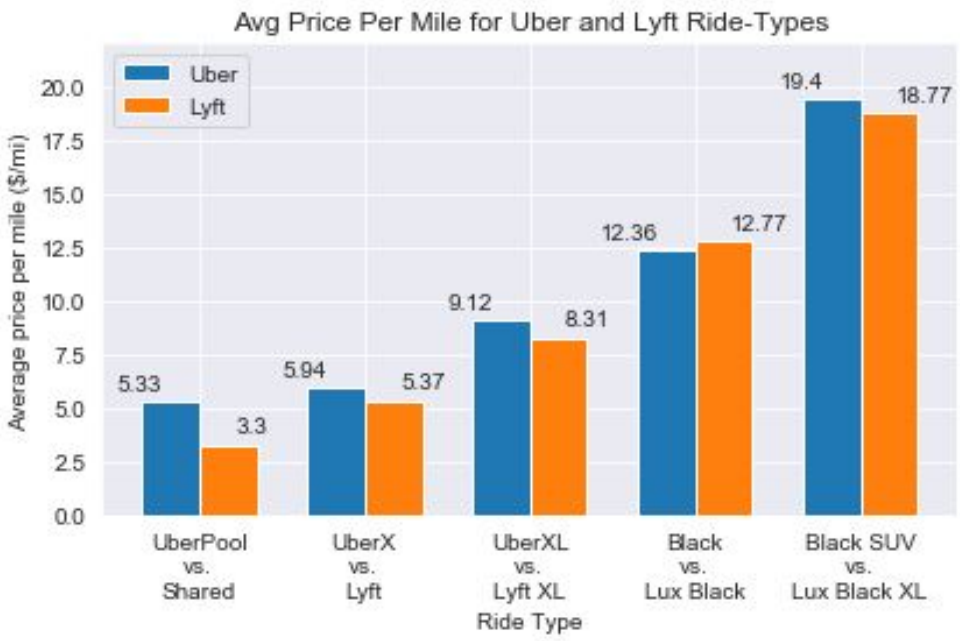
independent variables. This may affect the interpretation of the regression model, and will be monitored for multicollinearity when modelling process occurs. The pairs of independent variables with a high degree of correlation occur in weather variables: temp and clouds, temp and pressure, temp and humidity, clouds and humidity, pressure and wind.

When comparing rides under both the Uber and Lyft applications, the average price per mile was calculated:



According to the barplot, when comparing the Uber and Lyft ride applications, both ride applications, on average, cost around the same per mile, differing by 1 cent. For riders deciding between the two applications, both apps provide the same value for rides towards their destination. It may be through the option of different ride types that we see a difference in value between each ride application.

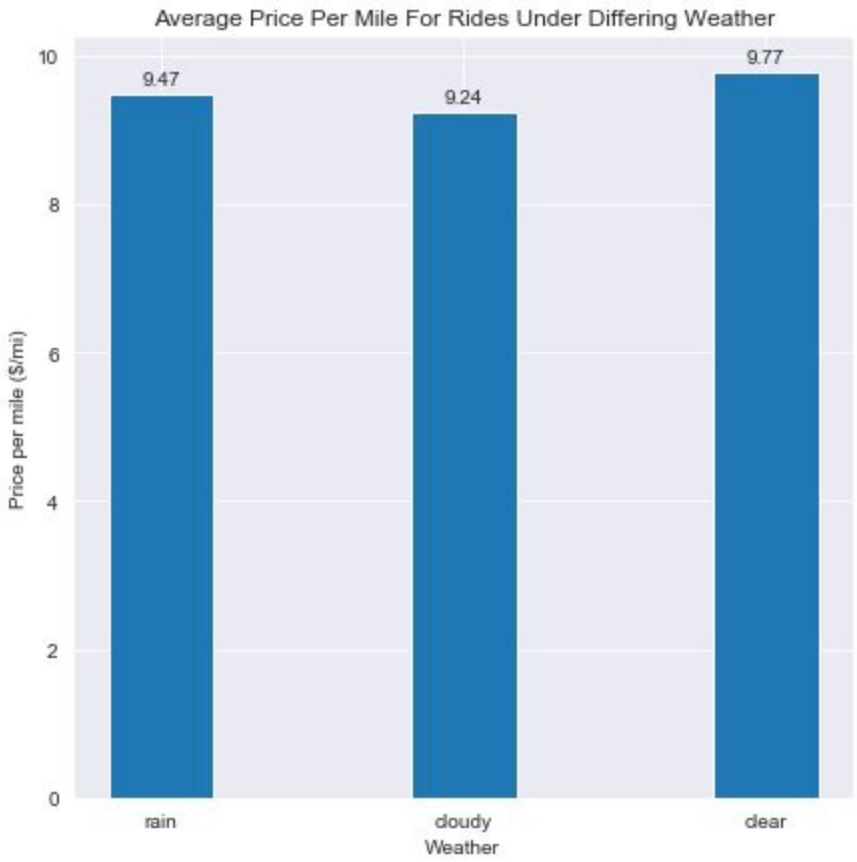
When comparing similar ride types between Uber and Lyft, such as UberPool vs. Shared, in which the rider will share their ride with other riders towards their destination, or Uber Black vs. Lux Black, in which the rider will expect premium black car service, for up to four riders, towards their destination, the average price per one mile was calculated and visualized on the barplot below:



Based on the average price per one mile of each ride-type in Boston from Uber and Lyft, if the rider would like to share their ride, with other riders, to their destination, splitting the cost, the Shared ride type by Lyft would be preferable, as on average, the Shared ride is 2 dollars less per mile than UberPool. If the rider requested their own car for transportation, for up to 3 riders, Lyft is preferable over UberX slightly (60 cents difference), and Lux Black is preferred over Uber Black (40 cents difference). For transportation of more than 4 riders, Uber Black or Lux Black XL is preferable over their counterparts as their is a slight difference in their average price per mile (80 cents and 70 cents difference respectively). These prices differences can add up if the destination is many miles farther than the pickup, so riders may want to choose accordingly which ride-type provides the best value per mile towards their destination. To statistically justify this, we performed five independent samples t tests, to compare the means of the Uber and Lyft ride-type's prices per mile, using a 0.05 significance level (3).

Because all of the calculated p-values were either 0, or less than the alpha level of 0.05, we reject the null hypothesis for all five cases (no equal average), and conclude that UberPool, UberX, UberXL, and Black SUV rides are, on average, charged more compared to their Lyft counterparts, based on the significantly positive t-statistics. Lux Black rides are, on average, charged more than its Uber counterpart, based on significantly negative t-statistic generated.

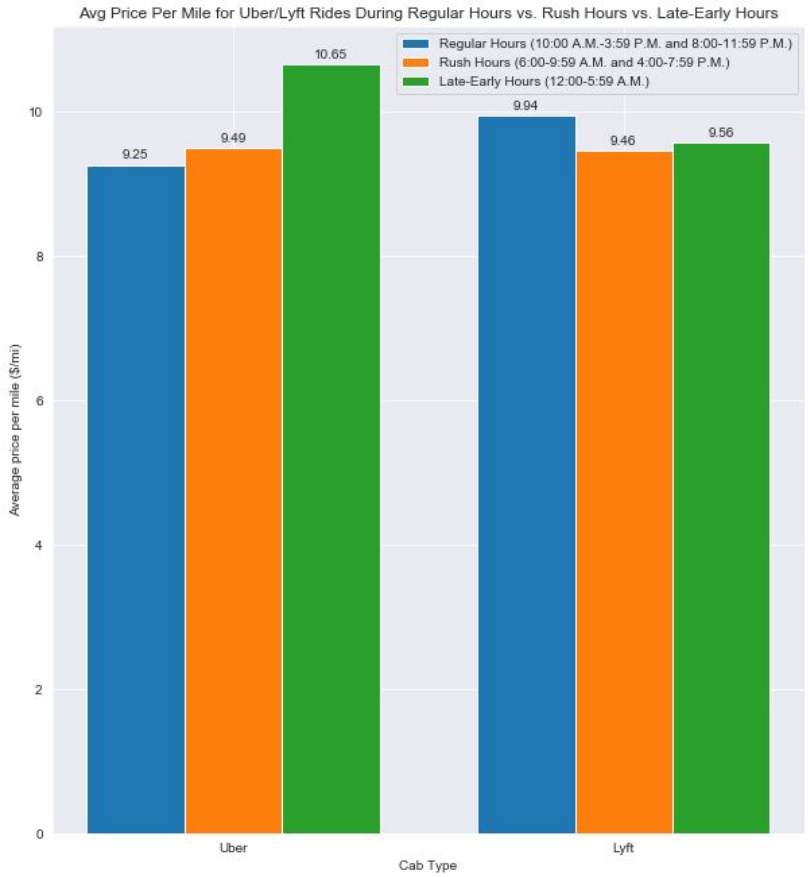
We compare the average price per mile for rides under differing weather conditions, rainy, cloudy, and clear:



According to prices per mile under rainy, cloudy, or cool weather, there is only a slight increase in price for rides under clear weather (0.30 cents increase) over rainy weather, which is only slightly more pricier (0.25 cents increase) than cloudy weather. Surprisingly, rides, on average, under clear weather cost more than under rainy and cloudy, however, this could be attributed to the fact that more people may be inclined to stay indoors during rainy weather, and more people are inclined to travel when the weather is clear. To statistically justify this, we performed three independent samples t-test for each of the three pairings: rainy and cloudy, rainy and clear, clear and cloudy.

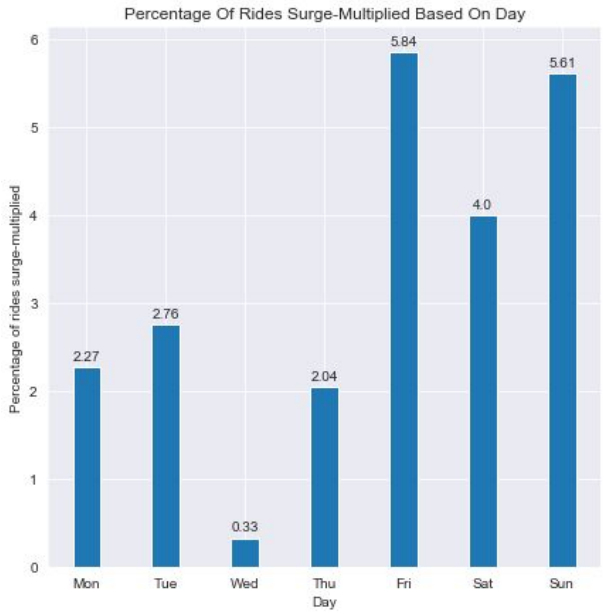
The results of the t-test indicate that prices are different depending on the weather, as we would reject the null hypothesis in all three cases, because the p-value is less than 0.05. The statistic indicates that prices under rainy weather are, on average, lower than under cloudy and cool weather, and prices under cloudy weather are, on average, lower than under cool weather, according to the negative statistics calculated for all three cases.

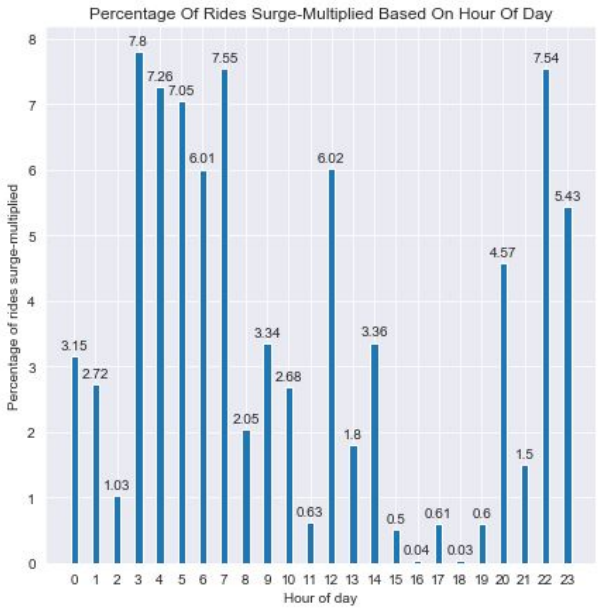
Looking at Uber and Lyft rides are differing time ranges: during rush hours, regular hours, and late to early hours, we compare the average price per one mile:



Based on the visual, Uber rides during late to early hours, on average, cost more than rides at both rush hour and regular hours, based on the 1 to 1.50 dollar price increase during these times, and, Uber rides during rush hour times cost, on average, only slightly more than regular hours. However, this is not the case for Lyft rides, in which rides at regular hours, on average, cost slightly more than both rides at rush hours and late to early hours. For Uber, the prices at late to early hours could be linked to the low supply of drivers at laters hours or the high demand of riders who may have gone to areas with nightlife and did not transport themselves. For Lyft rides, surprisingly, this is not the case as rides during rush hours and late to early hours are cheaper than rides during regular hours, which may point towards other factors that may influence the price at these hours. Overall, riders can expect higher prices during late hours for Uber, or regular hours for Lyft when deciding which application, at a specific time frame, provides the best value per mile for their rides.

Looking at rides during specific days of the week or specific hours in the day, the percentage of rides that were surge-multiplied > 1.0 was calculated:

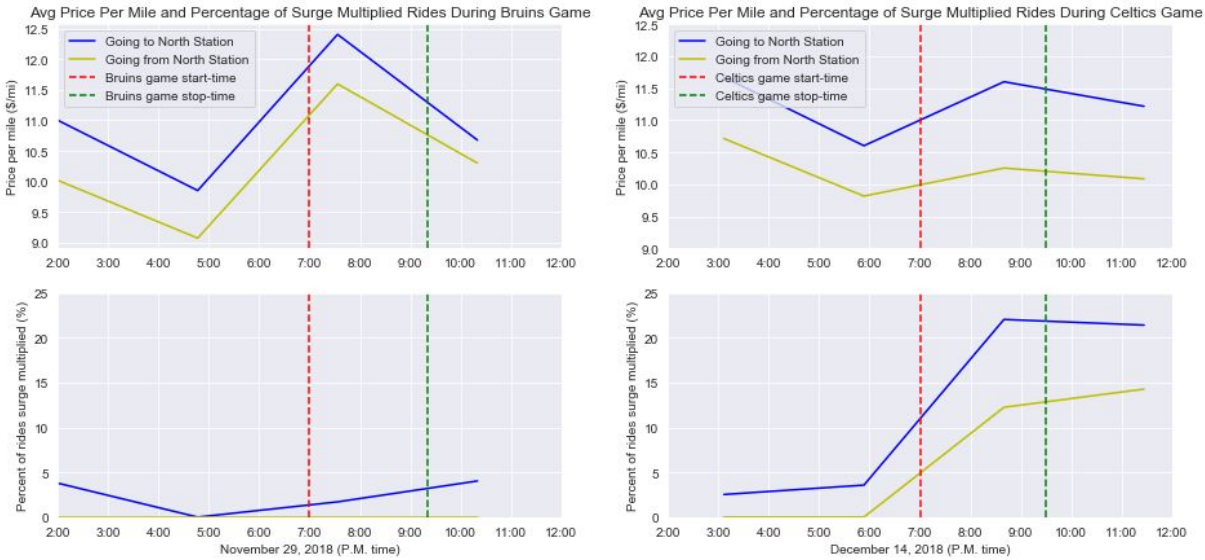




During the first four weekdays (Mon-Thur), a small percentage of rides are surge multiplied greater than 1.0, as Wednesday shows a particularly small percentage of rides that are surge-multiplied. Friday, Saturday and Sunday almost doubles the percentage of rides that are surge multiplied, compared to weekdays, indicating that these days are prone to higher than normal pricing of rides from either Uber or Lyft. This could be attributed to a higher demand of rides during the weekends, when most people are free from work and require transportation for daytime activities or nightlife in downtown Boston for areas prominent with bars and clubs, which should see a higher number of people that don't transport themselves and require Uber or Lyft.

The hours that are more prone to higher than normal pricing, by being surge-multiplied greater than 1.0, are between 3:00 to 7:00 A.M., as well as, 12:00 P.M, 8:00 P.M, and 11:00 to 12:00 P.M. With the exception of rides at 12:00 and 8:00 P.M, the majority of rides that are surge-multiplied, greater than 1.0, have occurred at late to early hours. Most rides occuring during normal hours and rush-hours seem to not be as surge-multiplied to the extent to these late to early hours. This as well could be attributed to a higher demand of rides towards the late night and early morning, for which the number of drivers may not be substantial, and the transportation of riders who may have gone downtown or areas prominent with nightlife and did not drive themselves. The increase in surged rides at 12:00 and 8:00 P.M. may be targeted towards riders on lunch breaks or dinner breaks that require transportation if they do not drive themselves.

Finally, we investigate the effect of a sports occurrence on the price of a ride, specifically, a Bruins game occuring on November 29, 2018, and a Celtics game occuring on December 14, 2018, on the average price per mile and percentage of rides surge-multiplied at certain times prior, during, and after the game. The time-series plots for each game are shown below:



For the Boston Bruins game occuring on November 29, 2018, the average price per mile increases slightly more for rides going towards and coming from North Station (in which the TD Garden arena is slightly north of) once the game started at 7:00 P.M. As opposed to the Celtics game, the percentage of rides that are surge-multiplied greater 1.0 have slightly increased once the game began for rides going towards the game, but none for rides coming from the area. Rides, on average, after the game finished still had higher pricing per mile than rides before the game started, which can be attributed to the demand of rides from a concentration of people who come out of North Station, near where the stadium is location, to their destination, rather than the concentration of riders coming from multiple source towards North Station. This may also attribute to the slight increase of surged rides during and after the game, as the demand of rides increases for game attendees.

For the Celtics game occuring on December 14, 2018, the average price per mile increase slightly more for rides going towards and coming from North Station (in which the TD Garden arena is slightly north of) once the game started at 7:00 P.M. The percentage of rides that are surge-multiplied greater 1.0 have significantly increased once the game began, which may indicate a greater demand for rides going toward and coming from the area. Rides, on average, after the game finished had both higher pricing per mile and percentage of surged rides than rides before the game started, which can be attributed to the demand of rides from a concentration of people who come out of North Station, near where the stadium is location, to their destination, rather than the concentration of riders coming from multiple source towards North Station.

Sources:

1. See “Capstone 1 Project Proposal” for further detail

2. See “Capstone 1 Data Wrangling Report” for further detail
3. See “Capstone 1 Statistical Method Report” for further detail