参赛队员姓名：　巢骏至（Junzhi Chao）

中学：　The Lawrenceville School

省份：　New Jersey

国家/地区：　USA

指导教师姓名：　刘哲

论文题目：Modeling and Analysis of Uber's Rider Pricing

Modeling and Analysis of Uber's Rider Pricing

Junzhi Chao

Abstract

A bustling financial center and a diverse cultural cosmopolitan, New York City (NYC)'s transportation system has always been an interesting topic for academics and various industries. The patterns and features of the transportation system, including traditional means of travel such as taxis and subways as well as innovative tools like ride-hailing platforms (Uber, Lyft, etc.), are important research topics in economics, transportation, and operational research fields. Thanks to Uber Developer's family of APIs, we now have a precious opportunity to acquire real-time operational data (price, ETA etc.) to further our analysis. This project aims to analyze the data of different locations, weathers, times, and dates (intraday and mid-week), using the acquired Uber operational data in New York City and applying time series analysis, statistical regression and prediction in econometrics. By calculating and analyzing the impact of these factors on Uber riders' payment amounts, we obtain conclusions that are instructive and beneficial in practice.

Key words: sharing economy, ride-hailing systems, dynamic pricing, econometrics, big data analysis

Table of contents

1. Introduction

As the process of modernization and urbanization in the late twentieth century and early twenty-first century rapidly unfolds, 55% of the world population now reside in urban space and that number is expected to grow in the following decades, according to a 2018 article published by the United Nations' Department of Economic and Social Affairs ("68% of the world population projected to live in urban areas by 2050, says UN | UN DESA Department of Economic and Social Affairs"). With the wave of urbanization comes the "Sharing Economy", a necessary compromise for the limited space and resources in urban areas, online car services such as Uber and Lyft being one of them. This new era, while seeing unprecedented changes and improvements in many aspects, faces many challenges in the same and new theories and models in economics need to be made in order to cope with these pending challenges. It is thus important to engage in field studies and set up a new system of formulas and models to evaluate to what extent and in what ways the urban lifestyle, rush hours, and commuting patterns affect online car services. Packed with residential neighborhoods, commercial centers, traffic junctions, tourist attractions, the city of New York is a perfect experiment ground due to its enormous market for online car services, complicated traffic systems, and its extraordinary diversities culturally, economically, and functionally. Investigating the Uber pricing and the various factors behind it in NYC is not an isolated study but rather a quest providing insights for urban areas around the world, which will remain to be vitally important as urban development and population cease to increase.

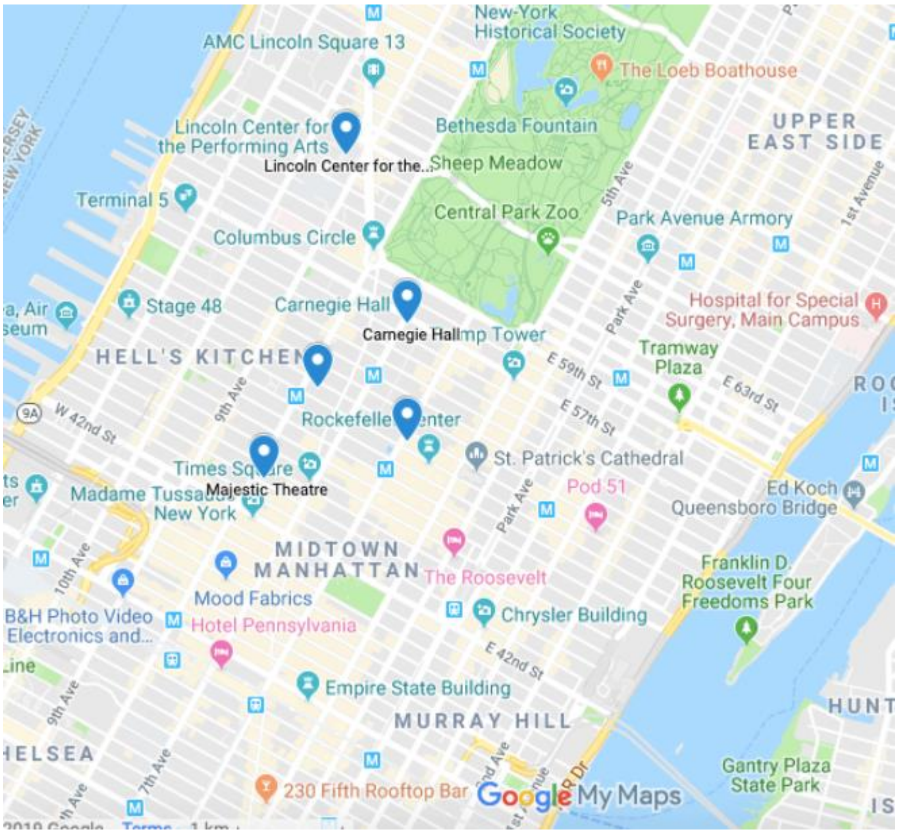
In order to understand the exact pricing model for Uber rides in NYC, we will first have to know Uber's regular pricing standards; note that there are many types of Uber vehicles and each type has its own pricing standard. These types include the regular UberX, which is the most regularly used and cheapest type of Uber vehicle; Uber XL, which uses SUVS and provides larger spaces and more comfortable rides; UberBLACK. which uses upscale cars and professional drivers (Majaski, "Uber vs. Yellow Cabs in New York City: What's the Difference?", 2019). This research, however, is based mostly on the pricing of the most commonly used UberX and did not take into considerations of carpooling. For UberX, according to the current standard in NYC, it has a base fare of 2.55 dollars, minimum fare of 7 dollars, and charges up 0.35 dollars per minute + 1.75 dollars per mile. The basic pricing model for Uber rides in NYC should thus be $P(X,Y) = 2.55 + 0.35X + 1.75Y$, where $ is the unit of measurement, X stands for the total amount of minutes driven, and Y stands for the total amount of miles travelled (Majaski, "Uber vs. Yellow Cabs in New York City: What's the Difference?", 2019). While this model is a viable way of knowing the NYC Uber pricing in certain ideal conditions, it is not sufficient and complete enough to be used as a precise model in this study as Uber imposes something called surge pricing, which charges higher fares during times of high demands (Majaski, "Uber vs. Yellow Cabs in New York City: What's the Difference?", 2019). Uber's claim is that drivers might need additional monetary incentive to ride passengers during various conditions——such as rush hours or bad weathers——which is why surge pricing is implemented during these times. This leads to the ultimate question and the goal of our study: what factors can possibly lead to surge pricing, and, to what extent they might determine the price.

Uber's big competitor in NYC is none other than the iconic yellow cabs, or taxis. The basic fare for these yellow cabs is 2.5 dollars, plus an additional 0.5 dollars for each mile travelled. Furthermore, no surge pricing is added to the riding fares for yellow cabs, which seems to make the yellow cabs more economically friendly than even the most basic UberX. However, an additional surcharge is usually added to the taxi bill due to the evening and morning rush hours. Nonetheless, without the surge pricing (which can be unreasonably high during some specific time periods), taxis seem to be a better option in times of rush hours, traffic jams, or other extreme conditions that can lead to surge pricing in Uber (Silverstein, "These Animated Charts Tell You Everything About Uber Prices In 21 Cities", 2014). Despite Uber's surge pricing, the fact that Uber still keeps a considerable market share NYC deserves more investigations on how surge pricing works in real life occasions.

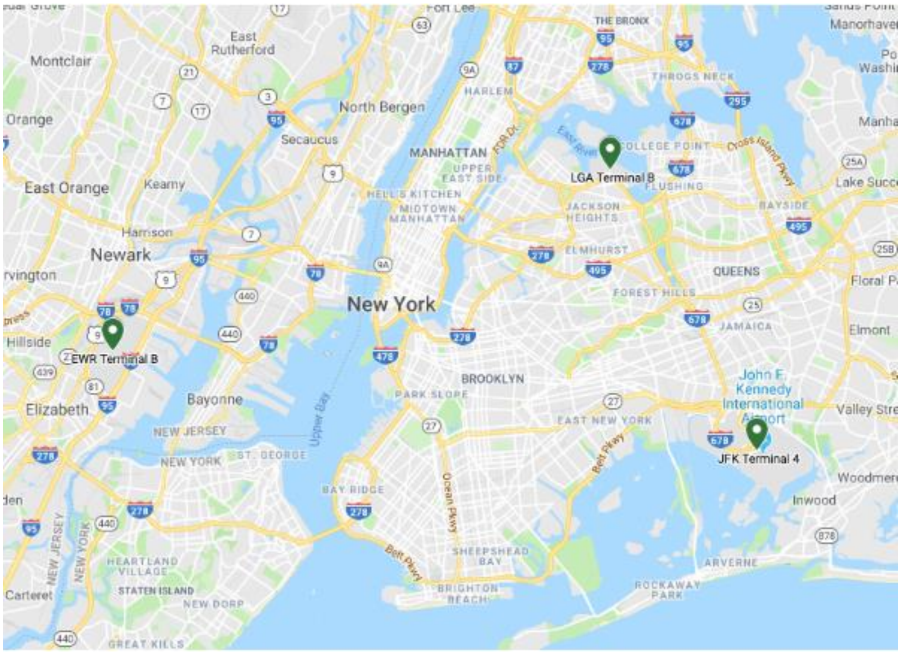
This paper presents a method and several approaches, including linear and logistic regression, to effectively estimate the different variables that affect Uber pricing in New York City and find out to what extent they impact the pricing of Uber rides during the time period when the data is collected and documented. We have made several observations about the possible factors behind Uber's surge pricing, such as the weather condition, the specific date (day of week), and the hour of day when rides take place. The hour of day's effect on surge pricing varies according to the starting location and Origin-Destination pair (O-D pair) of the Uber ride: for residential/train station locations, the rush hours of a day can really affect the surge pricing; for theatre locations, the hour of an important or popular play will affect the surge pricing; finally, for attractions, surge pricing might be affected by specific holidays, which will definitely increase the number of visitors and presumably the surge pricing; finally, for airports, surge pricing will be affected by the number of flights coming in at certain times as well as certain weather conditions, which might prohibit a number of flights from landing and taking off. The weather condition presumably has the most significant impact on the surge pricing of Uber rides with airport as starting location, as the taking off and landing of flights are dependent on the weather condition at the time. Different weather conditions will certainly affect surge pricing in different ways and to different degrees: we presume that weather conditions such as clouds or clear do not have the same impact on surge pricing as weather conditions such as snow or fog have. As for day of week, one important factor is definitely the weekday-weekend distinction: people often engage in different activities, go to different locations, and keep a different travelling pattern during weekdays and weekends. With the predictions in mind, we can now, through analyzing the sea load of data and constructing graphs and formulas, investigate whether they have an effect and to what extend their impact are to the surge pricing of Uber rides in New York City.

1.1 Hotspot locations shown on a map

In this research, we divide our data, according to the starting locations of the Uber rides, into five categories: theatres, residential area, train/bus stations, airport terminals and tourist attractions. As shown on the Figure 1 below, theatre locations include Lincoln center for the performance art, carnegie hall, majestic theatre, gershwin theatre, and radio city. Airport terminals include Laguardia Airport Terminal B, EWR Terminal B and JFK Terminal 4. Residential areas include 2nd avenue and 82nd street, Christopher street and Bleecker street, Columbus street and west 72nd street. Train stations include Grand Central Terminal, Pennsylvania station, Columbus circle (note that columbus circle is not a train station but a bus station/traffic junction; in this study, however, it is classified as a train station due to its similar function and similar effects in terms of surge pricing). Finally, tourist attractions include times square and empire building, two of the arguably most popular tourist destinations in NYC.

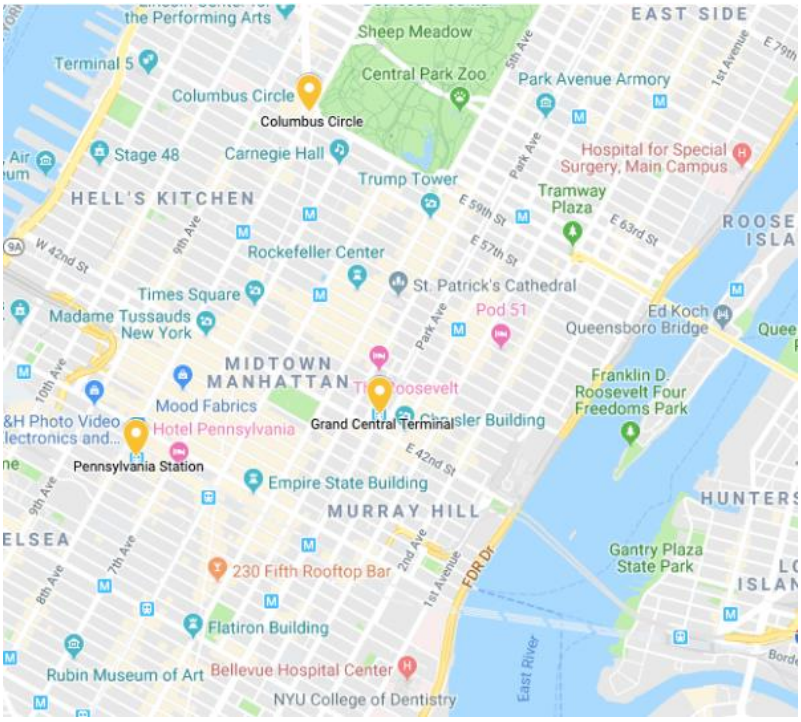


a. NYC theatre locations on the map



b. NYC airport locations on the map

c. NYC residential locations on the map



d. NYC train station locations on the map

e. NYC attraction locations on the map

Figure 1. 5 types of starting points on the map

2. Data analysis

2.1 Overview of data

The raw datasets for each location include several different columns, such as "weather" ("clear", "cloud", "thunderstorm", etc), which is the weather type when the Uber ride takes place; “x_low" , "x_high" (from which we can obtain"x_mid"), which determines the price range of the

Uber ride; "time", which indicates the date and the time when the Uber ride takes place. Other columns include "x_duration", the duration of the ride, and "x_distance", the distance of the ride. Note that although the duration and distance of a Uber ride from a given point to another should remain the same, the traffic in NYC and other factors such as road renovation and extreme weathers might affect the duration and distance. The sizes of different datasets vary, but are mostly centered around 149000 rows (minutes). For example, the carnegie_hall-second_e82 data set has 149087 rows, and the empire_building-grand_central dataset has 149175 rows, etc. The time range of the data sets is from around Feburary 17th, 2019 to around June 5th, 2019, which is about three and a half months. In the datasets we studied, the price/ETA is sampled at a frequency of approximately once per minute. In this particular research, we have collected datas on 10 different individual O-D pairs and we can group them into different categories: "airport-attraction" category, such as the "JFK_T4-times_square" pair; "attraction-train station" category, such as the "times_square-columbus_circle" pair; "theatre-residential" category; "theatre-train station" category; "residential-train station" category; "train_station-residential" category. These O-D pairs are all carefully selected and clearly represent the urban life patterns in NYC: for example, the "residential-train station" and "train-station-residential" O-D pair clearly represent the habit of the commuters in NYC. The basic Uber pricing model is price = minimum price + time price/unit time + distance price/unit distance.

2.2 Data visualization

2.2.1 Price variation throughout a day

These graphs below demonstrate the correlations between the hour of a day and the mean price (x_mid) of an Uber ride from a specific location in NYC to another. We can clearly see from the graph how the type of the location (the starting point of the Uber ride) affect the pricing. Through horizontal comparison, we can see for almost all locations, the pricing of the ride peaks around hour 17 to 18, which elapses with the evening rush hour. For type residential (exp. second_e82-penn_station), the pricing also peaks around hour 7 to hour 10, which elapses with the morning rush hour.

Figure 2 demonstrates the relationships between the hour of a day and the mean price (x_mid) of an Uber ride from Second Avenue and East 82nd Street, an intersection inside a residential neighborhood to Grand Central Terminal, a train station. This O-D pair belongs to the "residential-train station" pair. We can observe that the mean price drops to bottom between hour 3 and hour 5, peaks between hour 7 and hour 8, and rises again between hour 16 and hour 18. As hour 7 and hour 17 are both during rush hours, we can see how rush hours have significant effects on mean prices with "residential-train station" pair.

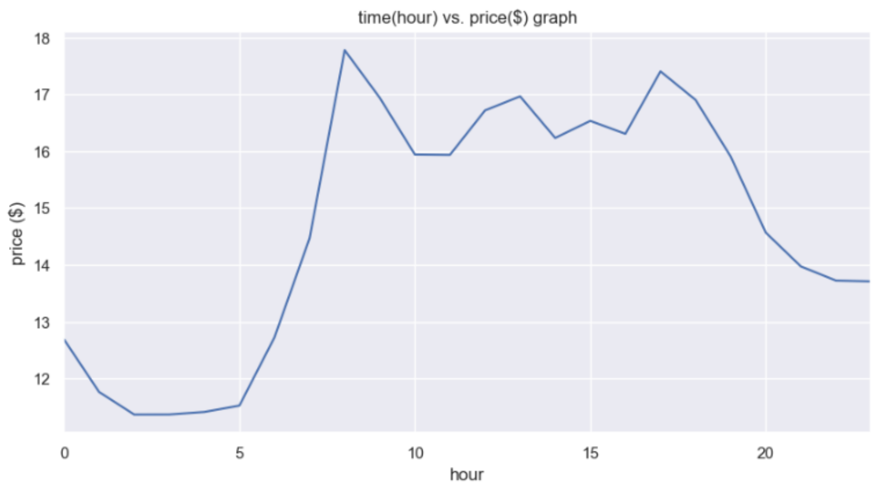


Figure 2. Time (hour) versus price ($) graph, "residential-train station" pair

Figure 3 demonstrates the relationships between the hour of a day and the mean price (x_mid) of an Uber ride from Terminal B, LaGuardia Airport, to Pennsylvania Station, a train station in NYC. This O-D pair belongs to the "airport-train station" pair. The surge pricing for this kind of O-D pair is not affected to the same degree as it is in the "residential-train station" O-D pair. Such a trend is logical because the amount of visitors to an airport is usually affected by the number of flights arriving and taking off during the time of the Uber ride, which fluctuates according to factors like the weather conditions.

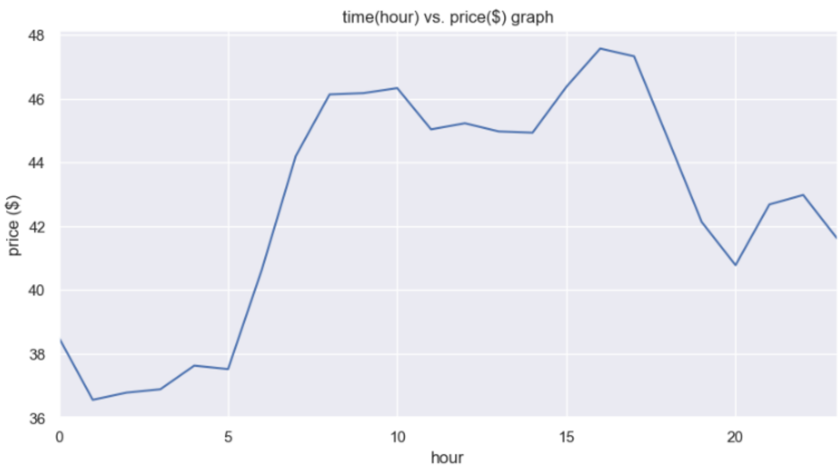


Figure 3. Time (hour) versus price ($) graph, " airport-train station " pair

Figure 4. demonstrates the relationships between the hour of a day and the mean price (x_mid) of an Uber ride from Gershwin Theatre, a theatre in NYC, to Pennsylvania Station, a train station. The O-D pair for this graph belongs to the "theatre-train station" pair. We can see the graph clearly has a different shape than the graph of the "residential-train station" O-D pair. This is because the surge pricing near a theatre is greatly affected by the schedule of popular shows and performances, not the rush hours or commuting patterns. The mean price for this O-D pair peaks at around hour 17, which is logical because most theatres put on shows during afternoons.

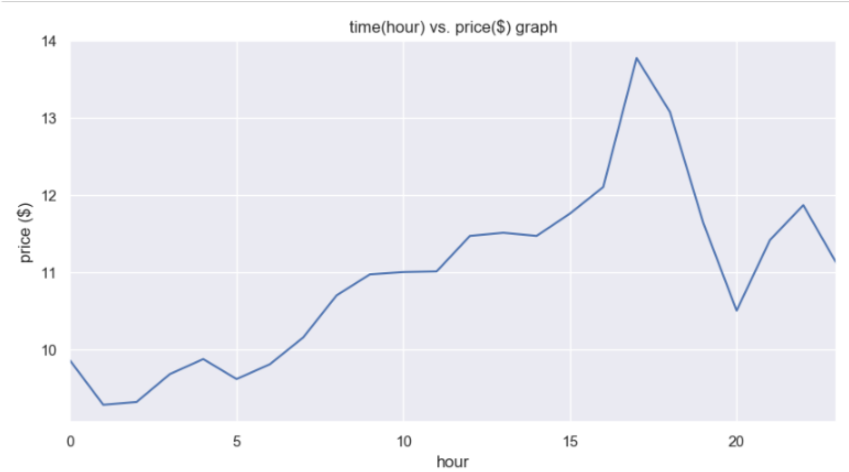


Figure 4. Time (hour) versus price ($) graph, "theatre-train station" pair

Figure 5. demonstrates the relationships between the hour of a day and the mean price (x_mid) of an Uber ride from Times Square, a tourist attraction, to Columbus Circle, a train/bus station. The Columbus Circle is actually a bus station and transportation center; however, its surge pricing pattern is quite similar to that of the train stations, which is why it is classified as a train station for the sake of this research. The O-D pair of this graph belongs to the "attraction-train station" pair. We can see that the trend of the mean price in this graph is completely different from that of the mean prices in the other graphs. This is because the surge pricing of attractions is highly dependent on the number of tourists visiting during the time period which is affected by factors like weather and date rather than rush hours. We can also see a relatively high mean price at hour 17 and again at hour 21. This is logical because the Times Square is known for its fabulous billboards, which looks more beautiful and stunning with the lightings at night.

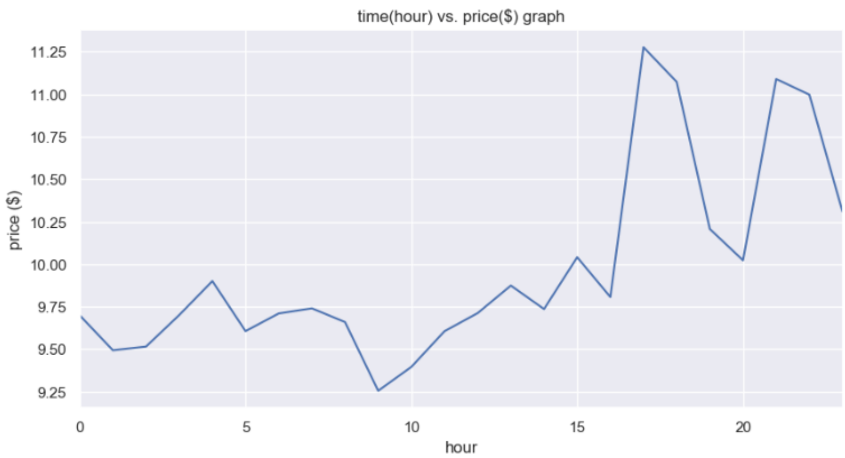


Figure 5. Time (hour) versus price ($) graph, "attraction-train station" pair

Figure 6. demonstrates the relationships between the hour of a day and the mean price (x_mid) of an Uber ride from Pennsylvania Station, a train station, to Christopher Street and Bleecker Street, an intersection inside a residential neighborhood. The O-D pair of the graph below is the "train station-residential" pair which is similar to the "residential-train station" pair

graphed previously. The two O-D pairs are very similar because they are both affected significantly by rush hours. However, a key difference is that the mean price for the "train station-residential" pair usually peaks during evening rush hour, because people usually take trains to get from their workplaces to the stations and then order Uber rides to get from the stations to their residences in the evening. On the contrary, the mean price for the "residential-train station" pair usually peaks during morning rush hours, because people usually order Uber rides to get from their residences to train stations and then take the trains to go to their workplaces in the morning. This graph provides the evidences as the mean price peaks between hour 17 and hour 18 which elapses with the time period of evening rush hours.

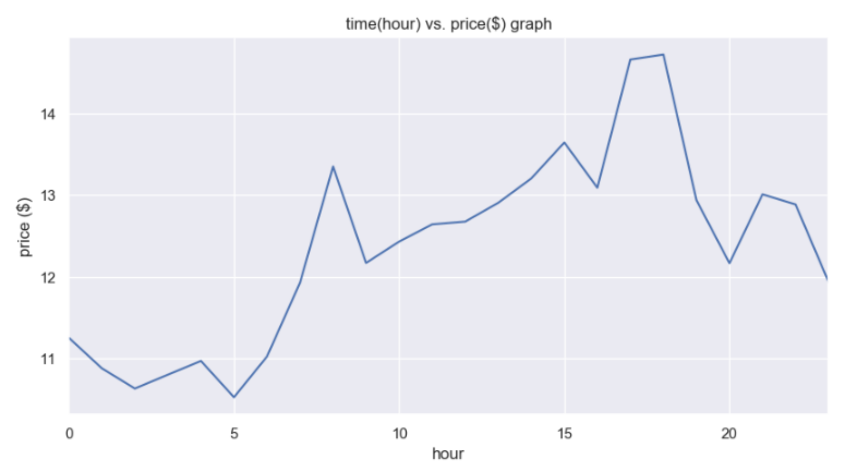


Figure 6. Time (hour) versus price ($) graph, "train station-residential" pair

2.2.2 Price variation over a week

This series of graphs demonstrate the mean price (x_mid) of an Uber ride from a specific location in NYC to another on a specific date. Here we choose one from each of the five categories of locations to compare their individual average value, 25%-value, 75%-value, the difference between days of a week, the difference between weekdays and weekends. By comparing these values and observing the trends vertically and horizontally, we can tell the how the day of week when the Uber ride takes place can impact the pricing and how different locations and O-D pairs can influence the pricing.

Figure7. shows the mean price of an Uber ride from second Avenue and East 82nd st., an intersection in a residential location, to the Grand Central Terminal, a train station location. The O-D pair of the graphed ride belongs to the "residential-train station" pair. From the graph you can observe the mean prices during weekdays are slightly higher than the mean prices during weekends, which is logical because the O-D pair is the "residential-train station" pair. Usually during weekdays, commuters will leave their residences and head to train stations (using Uber) for transportation to their workplaces, especially during the rush hours. During weekends, however, commuters no longer need to get to train stations or other transportation centers and Uber rides of these O-D pairs are thus no longer in high demands, consequently lowering the mean price (x-mid).

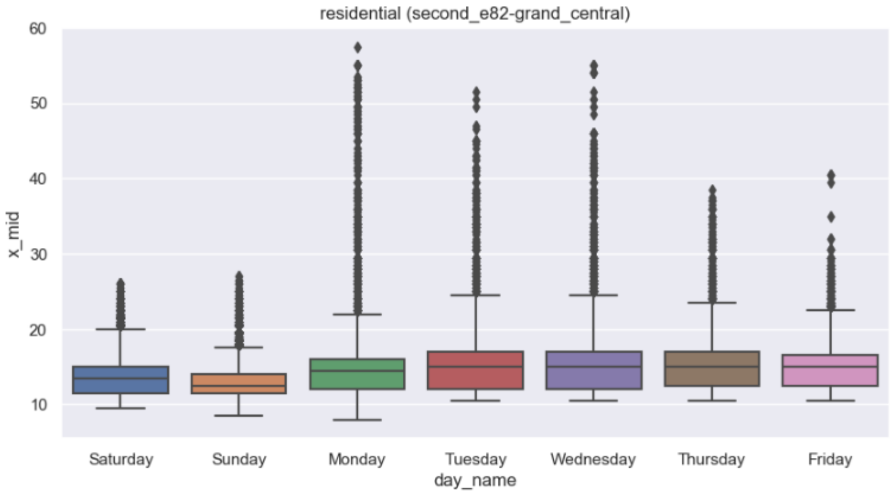


Figure 7. Mean price by day of a week, "residential-train station " pair

Figure 8. shows the mean price of an Uber ride from Terminal B, LaGuardia Airport, to times square, a tourist attraction. The O-D pair of the graphed ride belongs to the "airport-attraction" pair. From the graph you can observe that the mean prices during weekdays and during weekend are about the same which is logical because the O-D pair is the "airport-attraction" pair and both locations do not conform to commuting patterns or rush hour effects. One thing one might notice is that the mean prices of the "airport-attraction" pair are significantly higher than the means prices of other O-D pairs. The reason is that the distances between the three airports in NYC and the city center are considerably higher than the distances in other O-D pairs.

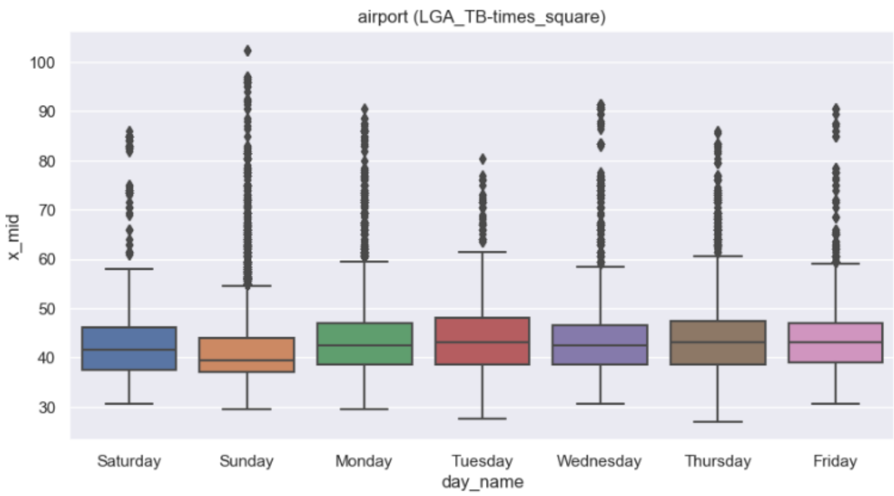


Figure 8. Mean price by day of a week, "airport-attraction" pair

Figure 9. shows the mean price of an Uber ride from Gershwin theatre, a theatre location, to Penn station, a train station. The O-D pair of the graphed ride belongs to the "theatre-train station" pair. From the graph you can o bserve that the mean prices during weekdays and during weekend are about the same which is logical because the O-D pair is the "theatre-train station" pair and both locations do not conform to commuting patterns or rush hour effects.

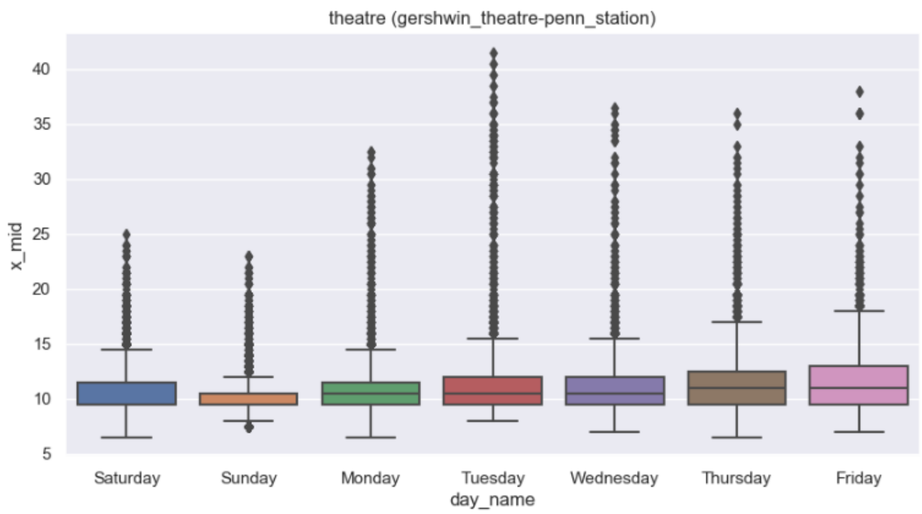


Figure 9. Mean price by day of a week, "theatre-train station" pair

Figure 10. shows the mean price of an Uber ride from times square, an attraction, to Columbus Circle, a train/bus station. The O-D pair of the graphed ride belongs to the "attraction-train station" pair. From the graph you can observe that the mean prices for all days in a week except Friday stay the same whereas the mean price on Friday is significantly higher. There are no clear fluctuations from mean prices in weekdays to mean prices in weekends, for attraction type locations are not usually affected by commuting patterns or rush hours. Beware that the same mean prices from Saturday to Thursday might be caused by an insufficient sum of data or other errors during the process of research or field studies.

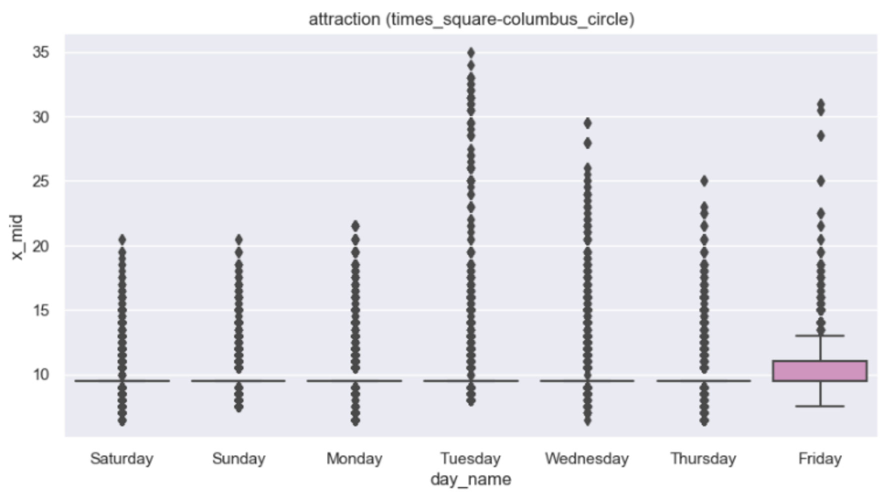


Figure 10. Mean price by day of a week, " attraction-train station " pair

Figure 11. shows the mean price of an Uber ride from Pennsylvania Station, a train station, to Christopher st. and Bleecker st., an intersection in a residential location. The O-D pair of this Uber ride belongs to the "train station-residential pair". From the graph you can observe that the mean prices are generally higher during weekdays and lower during weekends, the bottom being Sunday. This is because commuters use Uber to get to their residences from train stations much more frequently during rush hours on weekdays.

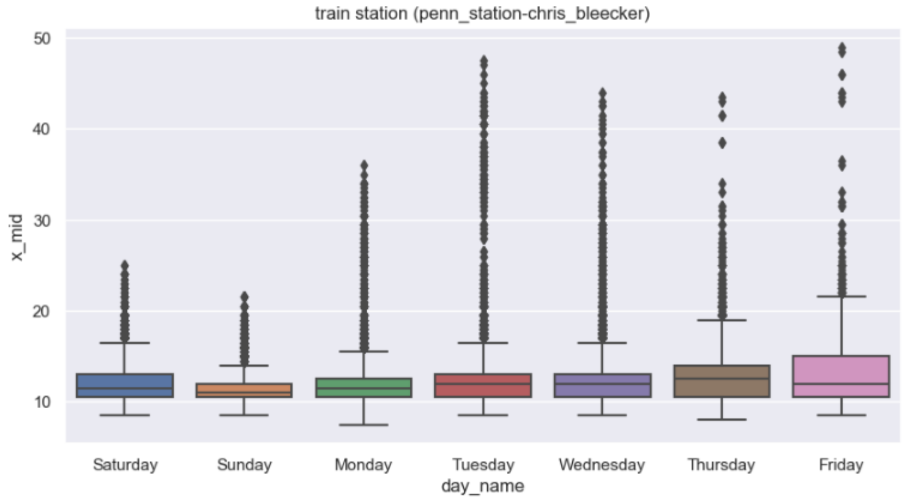


Figure 11. Mean price by day of a week, " train station-residential pair " pair

2.2.3 A finer look into the price variation

Figure 12. demonstrates the correlations between the time (minute) and the mean price (x_mid) of an Uber ride from a specific location (e.g Second Avenue and East 82nd street) in NYC to another (e.g Grand Central Terminal). To get a finer look into the price variation, the figure below shows the minute-level average price for a ride between Second Avenue and East 82nd street, an intersection in a residential neighborhood and the Grand Central Terminal over a day and we can find the peak of the mean price is around 400 to 600 minutes. This O-D pair belongs to the "residential-train station" pair which is usually more influenced by rush hours.

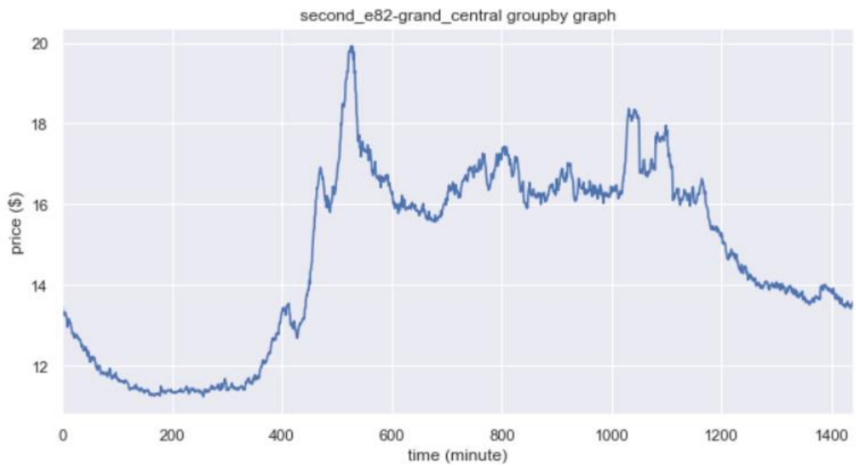


Figure 12. Time and mean price at a specific location

2.2.4 Weather

We use NYC's public weather data to keep track of the 10 weather types shown in the following table 1, together with their basic statistics. Note that some weather types may prevail at the same time.

Table 1. Descriptive Weather condition data

	Clear	Clouds	Drizzle	Fog	Haze	Mist	Rain	Snow	Squall	Thunderstorm
count	149157	149157	149157	149157	149157	149157	149157	149157	149157	149157
mean	0.372	0.261	0.100	0.078	0.020	0.286	0.212	0.042	0.000	0.009
std	0.483	0.439	0.301	0.269	0.139	0.452	0.409	0.200	0.020	0.097
min	0	0	0	0	0	0	0	0	0	0
25%	0	0	0	0	0	0	0	0	0	0
50%	0	0	0	0	0	0	0	0	0	0
75%	1	1	0	0	0	1	0	0	0	0
max	1	1	1	1	1	1	1	1	1	1

3. Model and analysis

3.1 Regression analysis

In this section, we will look closely to the data collected and study the impact of weather, location, time, date, and other variables on price using regression analysis.

3.1.1 Linear regression models

The regression model is

$$price=\beta0+\beta1[Hour]+\beta2[Weather]+\beta3[Day] \quad (1)$$

where β0 is the constant and βi, i=1,2,3, are coefficient vectors of Hour, Weather and Day features (dummy variables), respectively.

3.1.2 Findings

The data shown in the graphs below are the average coefficients of each type of locations for different days of a week, different weather types, and different hours of a day. Through regression analysis, we can discover many trends and features from the average coefficients of each types of locations in NYC. From comparing the average coefficients for each days of a week, we can notice that the average coefficients of weekdays, Monday through Friday, are higher on average than the average coefficients of weekends, Saturday and Sunday for all locations except Airport. This is logical because the amount of Uber users in locations such as residential and train station vary greatly due to the effects of rush hours during weekdays, while locations like airport are genuinely unaffected by rush hour effects.

From comparing the average coefficients under different weather conditions, we can notice that the average price coefficients are considerably lower under weather conditions such as mist and clouds and significantly higher under weather conditions like squall and thunderstorm. The relatively low price coefficients for mist and clouds are reasonable, due to their limited effects on traffics. The higher price coefficients for thunderstorm are also very logical, because thunderstorms can pose a danger on traffic and city transportation, cutting down the number of available Uber drivers and thus increasing the average price coefficients. And note that the unusually large price coefficient found for squall might be the result of the lack of research data.

From comparing the average coefficients during different hours of a day, we can notice that for all locations except airports, especially residential and attraction locations, the average coefficients peak at around hour 7 to hour 8 and again at around hour 17 to hour 18. Clearly, these time periods elapse with the time periods for morning and evening rush hours, which give rise to higher demands for Uber rides and higher Uber pricing, consequently. Another trend we can see is that for theatre locations, the average coefficients stay relatively low during hour 0 to hour 6 and again during hour 9 to hour 12. This trend is reasonable because most shows in theatres take place in the afternoons and at nights, sparking high demands and high pricing for Uber rides during those times and leaving no demand for the rest of the time when no show is put on stage.

3.1.3 Regression results

We find the R-squared value for each types of locations, the highest being the residential location example (0.601), and the lowest being the attraction location example (0.206). The higher the R-squared value is, the more consistent and closer to the line of best fit it is. The regression results are shown in Table 2. Detailed regression tables are listed in Appendix part.

Table 2. Results of 5 regression models

Dataset	O-D Pair	Starting point	R Squared
LGA_TB-times_square	airport-attraction	airport	0.565
Gershwin_theatre-Penn_station	theatre-train station	theatre	0.434
second_e82-grand_central	residential-train station	residential	0.601
penn_station-chris_bleecker	train station-residential	train station	0.353
times_square-columbus circle	attraction-train station	attraction	0.206

3.1.4 Linear regression charts

Table 3. Day of week vs. Price Coefficient For Different Locations in NYC

Day of week	Theatre	Residential	Train station	Airport	Attraction
Sunday	0	0.37	0.27	4.04	0.44
Monday	0.35	0.6	0.61	4.62	0.68
Tuesday	0.69	0.56	0.82	3.24	0.9
Wednesday	0.7	0.65	1.02	2.42	1.06
Thursday	0.48	0.38	0.94	2.11	1.01
Friday	0.68	0.73	1.26	2.08	1.19
Saturday	0.07	0.49	0.38	2.18	0.48

Table 4. Weather vs. Coefficient For Different Locations in NYC

Weather	Theatre	Residential	Train station	Airport	Attraction
Clear	0.23	0.3	0.01	0.28	0.02
Clouds	-0.03	0.03	-0.19	0.25	-0.08
Drizzle	-0.51	-0.44	-0.4	-0.28	-0.33
Fog	0.17	0.15	0.11	0.44	0.06
Haze	0.17	0.15	0.17	**1.09**	0.02
Mist	0.06	0	-0.1	-0.07	-0.05
Rain	0.64	0.61	0.61	0.18	0.5
Snow	0.71	0.93	0.57	**1.24**	0.45
Squall	**16.45**	**16.48**	**15.09**	**2.48**	**11.65**
Thunderstorm	**4.41**	**4.47**	**3.99**	**1.05**	**3.15**

Table 5. Hour of day vs. Price Coefficient For Different Locations in NYC

Time (hour)	Theatre	Residential	Train station	Airport	Attraction
0	-0.33	-0.17	-0.2	2.48	0.05
1	-0.67	-0.48	-0.46	0.43	-0.11
2	-0.71	-0.48	-0.53	-1.06	-0.09
3	-0.5	-0.34	-0.36	-1.41	0.1
4	-0.28	-0.27	-0.13	-0.1	0.28
5	-0.38	-0.4	-0.35	1.52	0.01
6	-0.05	0.39	-0.11	0.61	0.06
7	0.22	1.08	0.14	0.27	0.04
8	0.5	2.35	0.42	-0.25	0.09
9	-0.53	-0.36	-0.52	-1.3	-0.5
10	-0.62	-0.76	-0.6	-1.62	-0.42
11	-0.43	-0.57	-0.42	-1.42	-0.29
12	-0.19	-0.24	-0.23	0.55	-0.16
13	0.08	0.14	0.08	1.25	0.06
14	0.19	0.19	0.25	1.59	0.05
15	0.63	0.67	0.61	2.55	0.3
16	0.57	0.51	0.52	1.75	0.2
17	1.89	0.99	1.97	1.61	1.4
18	1.54	0.71	2.02	1.41	1.46
19	0.08	0.18	0.38	0.36	0.29
20	-0.28	-0.38	-0.03	1.45	0.06
21	0.83	0.22	1.17	2.4	1.17
22	0.99	0.39	1.15	3.9	1.13
23	0.4	0.4	0.53	3.75	0.57

3.1.5 Price coefficients graphical analysis

Figure 13. demonstrates the relationship between the time of a day (hour) and the change of Uber ride price for different locations in NYC according to the type of the location. As shown on the graph above, the y-value of attraction category is significantly higher than other categories from hour 20 to 24; one explanation will be that many tourists visit these attractions during night time, thus boosting the coefficient during the same time period. Another trend, as shown on the graph, is that the y-value of train station category peaks around hour 8 to 10; one explanation will be that this time period elapses with the time of the rush hour, prompting more Uber users to go to locations near a train or bus station.

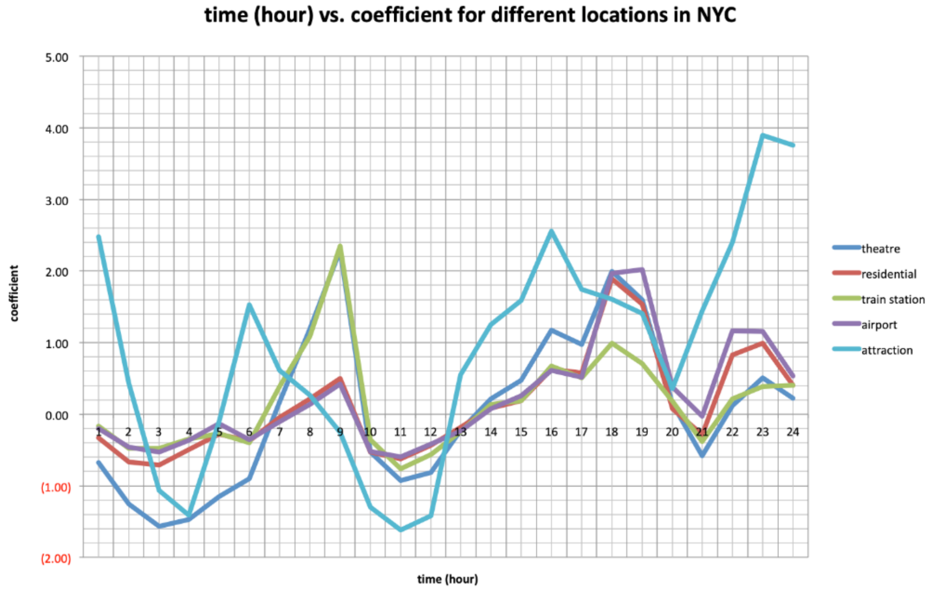


Figure 13. Time (hour) versus coefficient for different locations

Figure 14. shows the relationship between the day of a week and the change of Uber ride price for different locations in NYC according to the type of the location. Noting that the distance travelled (x_distance) was not considered in this graph, we can only carry out horizontal comparison at each data point. One specific trend we can notice from the graph is that for all categories except airport, the highest point of their y-values occur on day 6 (Friday). One possible explanation will be that people often follow their work-day patterns (going to work) during the day and follow their off-day patterns (engaging in social activities) during the night on Friday, because it is the last day before a weekend. Such patterns increase the amount of activities during Friday, thus increasing the y-values for most categories; the airport is an exception because the density of flights arriving, and departing does not usually correlate with the day of a week.

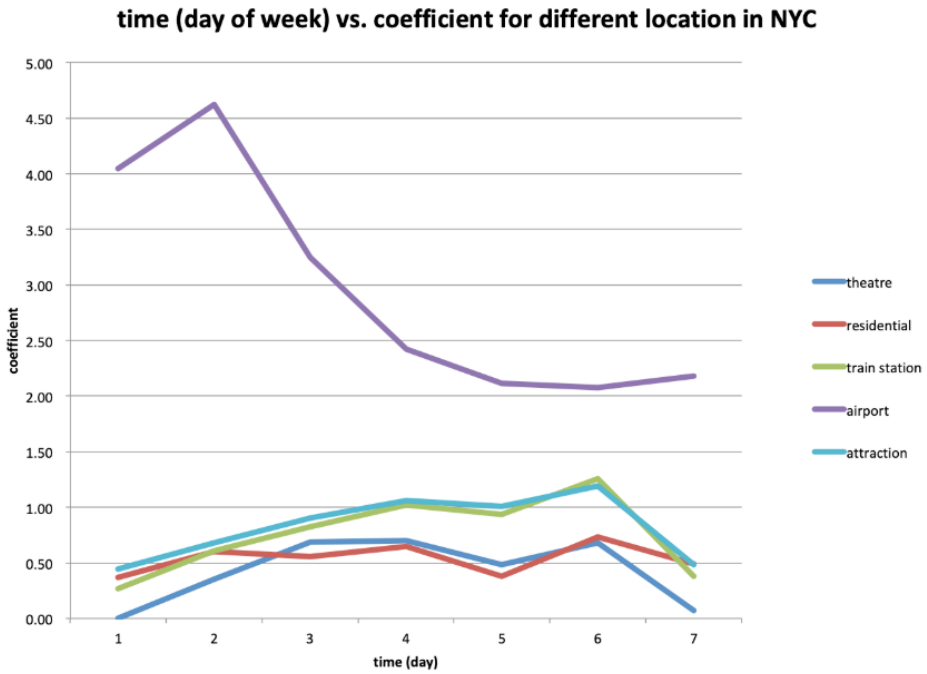


Figure 14. Time (day of week) versus coefficient for different locations

3.1.6 Logistic regression model

The logistic regression model is given by

$$\text{Prob(surge)} = \frac{1}{1+e^{-(\beta_0+\beta_1[\text{Hour}]+\beta_2[\text{Weather}]+\beta_3[\text{Day}])}} \tag{2}$$

The following two tables show the regression results for the logistic regression model we constructed above. Table 6 which shows the relationship between the hour of a day and the surge value for different locations in NYC, we can observe that for train stations, higher values occur at hour 8 and again between hour 17 to hour 19. Similarly, for residential locations, higher values occur at hour 8 and again at hour 17. Both locations peak during times of morning and evening rush hours, which create high demands in a short amount of time and thus affect the surge pricing. The same trend does not occur for airports, attractions, or theatres, because the surge pricing is applied in different conditions and during different times for these locations. For example, surge pricing might be applied to an attraction during national holidays when many visitors flood into the sites and create higher demands consequently. Another example might be that surge pricing for theatres is only applied during a popular show or performance when more audiences show up and the demand for Uber rides goes up at the same time.

Table 6. Hour of day vs. Prob (surge) For Different Locations in NYC

Time (hour)	Theatre	Residential	Train station	Airport	Attraction
0	-11.58	-34.81	-34.84	-6.9	-48
1	-39.73	-54	-71.25	-10.73	-47.5
2	-44.47	-54.62	-72.4	-16.1	-48.28
3	-32.58	-36.3	-52.69	-13.01	-47.49
4	-12.45	-35.77	-34.86	-11	-15.78
5	-17.89	-21.58	-52.73	-7.67	-47.51
6	-9.87	-10.7	-33.87	-4.35	-14.68
7	-7.12	-6.48	-12.11	-0.69	-28.92
8	-4.37	-3.08	-8.9	-0.2	-8.78
9	-3.8	-7.95	-13.24	-0.83	-30.45
10	-2.4	-11.74	-50.78	-2.56	-49.13
11	-2.1	-12.28	-50.31	-3.69	-48.49
12	-3.22	-10.26	-31.56	-1.95	-48.26
13	-1.17	-8.31	-29.59	-0.65	-32.71
14	-2.16	-7.89	-12.13	-0.53	-12.22
15	-0.62	-6.59	-10.46	2.66	-10.94
16	-1	-7.08	-12.38	2.98	-13.73
17	2.56	-4.88	-6.78	3.57	-8.22
18	1.53	-6.85	-7.09	1.9	-7.28
19	-2.06	-8.64	-9.78	-3.56	-8.79
20	-4.5	-14.24	-32.3	-4.34	-47.62
21	-1.97	-13.55	-28.89	-3.08	-47.35
22	-2.34	-12.42	-29.04	-2.8	-12.15
23	-5.05	-15.35	-31.36	-4.84	-11.9

Table 7. shows the relationship between the weather conditions when the Uber rides take place and the surge value for different locations in NYC. We can observe from the chart that the surge values during "Clear", "Cloud", or "Drizzle" are significantly lower than those during "Thunderstorm" or "Haze". This is a reasonable outcome because the earlier three weather conditions typically don't affect traffics as much as the latter two do. Drivers are less likely to go out and seek for passenger during extreme weather conditions such as thunderstorms; to encourage drivers to continue working in times of severe weathers, Uber typically uses surge pricing during these times to attract drivers. Also, during extreme weather, the demands of the passengers usually exceed the supplies of the drivers, which is also why Uber will utilize surge pricing to increase their profits.

Table 7. Weather Condition vs. Prob (surge) For Different Locations in NYC

weather	Theatre	Residential	Train station	Airport	Attraction
Clear	-4.37	-9.12	-15.44	-2.65	-19.81
Clouds	-4.71	-9.72	-13.61	-2.38	-18.31
Drizzle	-0.74	-2.66	-27.88	-0.13	-28.62
Fog	-0.99	-2.26	-2.62	-1.03	-3.61
Haze	3.5	-2.75	-6.04	2.13	-2.54
Mist	-4.12	-5.17	-7.31	-1.98	-11.43
Rain	-0.07	-3.21	-5.39	-0.71	-5.46
Snow	1.98	-2.39	-6.36	-0.45	-4.38
Squall	81.73	21.92	23.53	10.58	105.27
Thunderstorm	2.71	3.59	2.85	0.59	5.55

Figure 15. showcases the relationship between the time of a day (in hours) and the price coefficient for Uber rides starting from different types of locations: theatre, residential, train station, airport, and attraction. We can identify some specific trends in this graph for some types. For the train station category, for example, the y-value reaches its peak around hour 9 and again around hour 18 to 19, which correlates with the time of morning and evening rush hours. The trend is logical because train station is used more frequently during rush hours. With the sudden influx of people, the Uber prices also increase correspondingly, adhering to the laws demand and supply in economics.

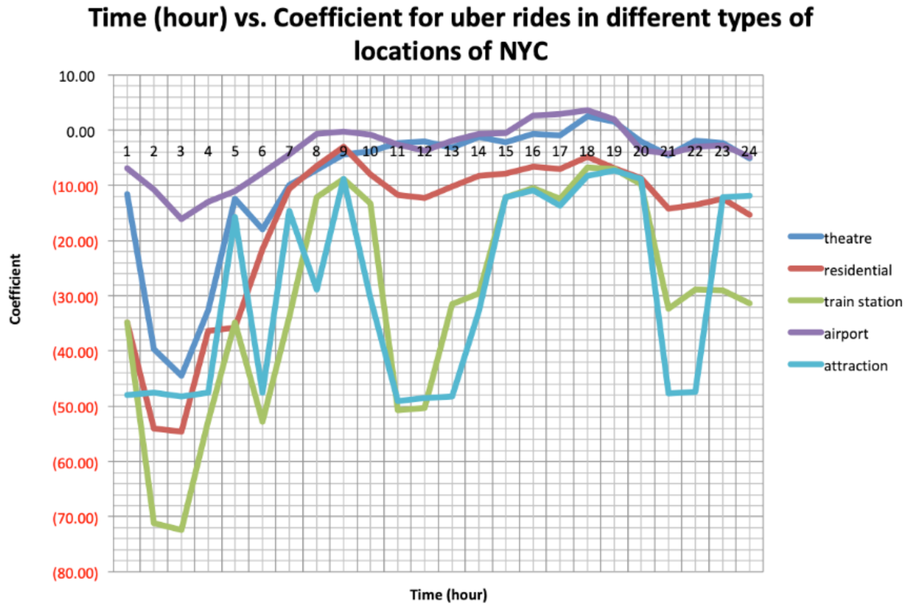


Figure 15. Time (hour) vs. coefficient for uber rides in different types of locations of NYC

Figure 16. showcases the relationship between the weather and the price coefficient for Uber rides starting from the five different types of locations listed above. We can notice that according to the graph, weather condition squall can have a significant impact on price coefficient for all locations. One possible explanation is that squall might be an outlier in this experiment, because it happens rarely and thus making the coefficient very extreme. Overall, we have not been able to see a clear trend in the data sets we gathered about weather condition's effect on Uber pricing for different locations in NYC.

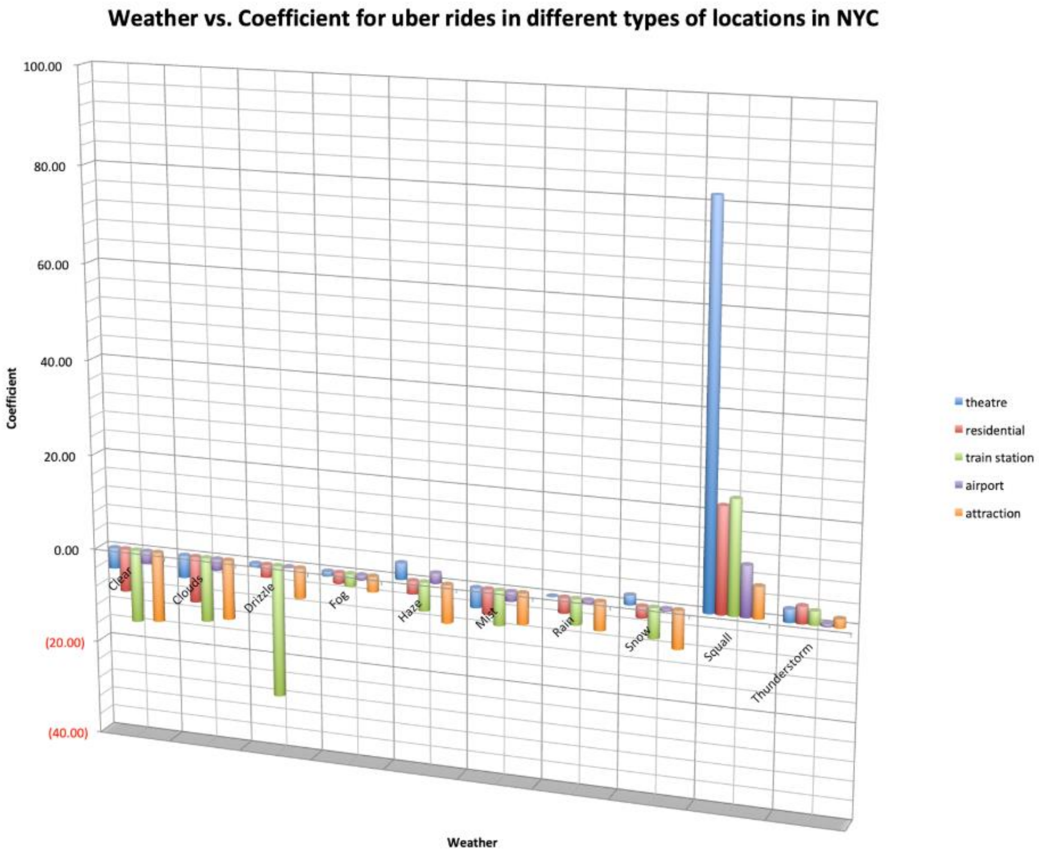


Figure 16. Weather vs. Coefficient for uber rides in different types of locations in NYC

3.2 Time series analysis

The series of graphs below illustrate the correlations between the day of a week (1 represents Sunday, ..., 6 represents Monday) and the mean price (x_mid) of an Uber ride from a specific location in NYC to another for all five given categories of locations. we plotted charts with different methods, weekly mean resample and rolling average (moving window), to show the basic ideas of the data: seasonality and trend.

3.2.1 Airport

Figure 17. shows the time series for Uber rides starting from an airport location. This example shows the daily mean, weekly mean, and rolling average of an Uber ride from LaGuardia Airport to times square, which is an "airport-attraction" O-D pair. From graph 1, we can observe that both the weekly mean and daily mean reach their bottom between April 2nd and April 9th and peak between May 14th and May 21th. We can also observe that for each week, the daily mean usually peak at Friday, or the fifth day of the week. For example, for one week from May 14th to May 20th, the daily mean reaches its relative maximum on May 18th, which is a Friday. Similarly, for one week from April 16th to April 22nd, the daily mean reaches its relative maximum on May 20th, a Friday as well.

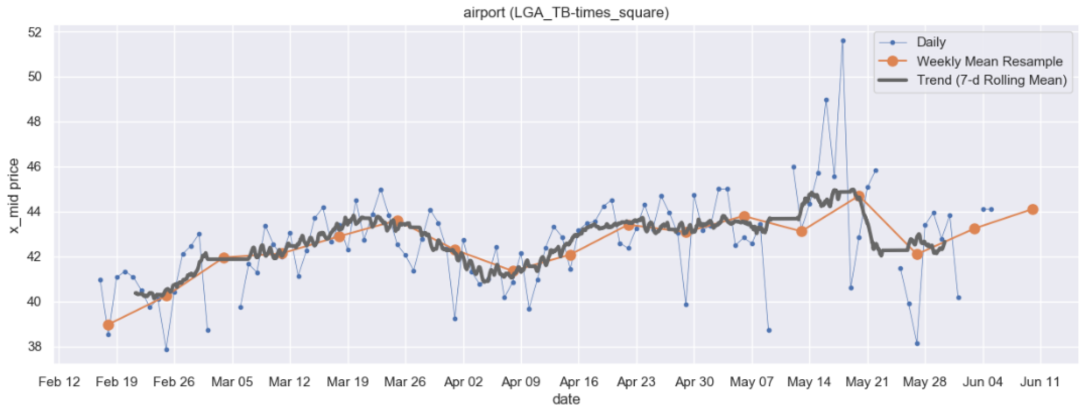


Figure 17. Time series for Uber rides starting from an airport location

3.2.2 Train station

Figure 18. shows the time series for Uber rides starting from a train station in NYC. This example shows the daily mean, weekly mean, and rolling average of an Uber ride from Pennsylvania Station, a train station, to Christopher Street and Bleecker Street, an intersection in a residential neighborhood. This is a "train station-residential" O-D pair, which is significantly affected by rush hours and commuting behaviors. From graph 2, we can observe that both the weekly mean and daily mean reach their bottom between March 26th and April 2nd and peak between May 14th and May 21st. We can also observe that for each week, the daily mean usually peak at Fridays, or the fifth day of the week. For example, for one week from March 12th to March 18th, the daily mean reaches its relative maximum on March 16th, which is a Friday. Similarly, for one week from April 23rd to April 29th, the daily mean reaches its relative maximum on April 27th, a Friday as well. The fact that the Uber ride's daily mean peaking on Fridays is logical because on Friday morning and evening rush hours, commuters go in and out of the train stations in great magnitude, increasing the demands and pricing of Uber rides from the train stations. Furthermore, after work on Friday, many residents of NYC engage in social activities, increasing the usage of the train/bus stations and the pricing once again. Under both effects, the daily mean of Uber pricing reaches its greatest on Friday every week.

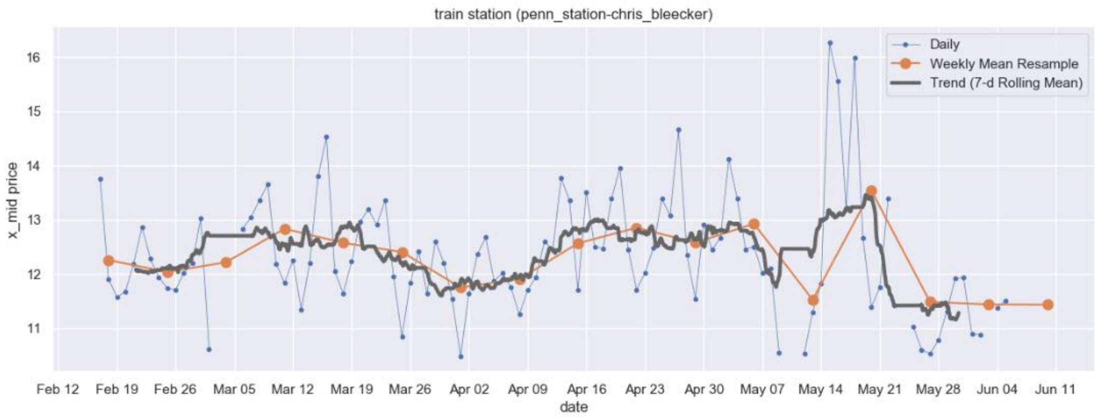


Figure 18. Time series for Uber rides starting from a train station in NYC

3.2.3 Theatre

The chart below illustrate the time series for Uber rides starting from a theatre location in NYC. This example shows the daily mean, weekly mean, and rolling average of an Uber ride from Lincoln Center, a theatre and center for performance art, to Christopher street and Bleecker street, an intersection in a residential neighborhood. This is a "theatre-residential" O-D pair, which is not as significantly affected by rush hours and commuting behaviors as the "train station-residential" pair. From graph 3, we can observe that both the weekly mean and daily mean reach their bottoms between May 7th and May 14th and peak between May 14th and May 21st. We can also observe that for each week, the daily mean usually peak at Fridays, or the fifth day of the week. For example, for one week from March 12th to March 18th, the daily mean reaches its relative maximum on March 16th, which is a Friday. Similarly, for one week from April 23rd to April 29th, the daily mean reaches its relative maximum on April 27th, a Friday as well.

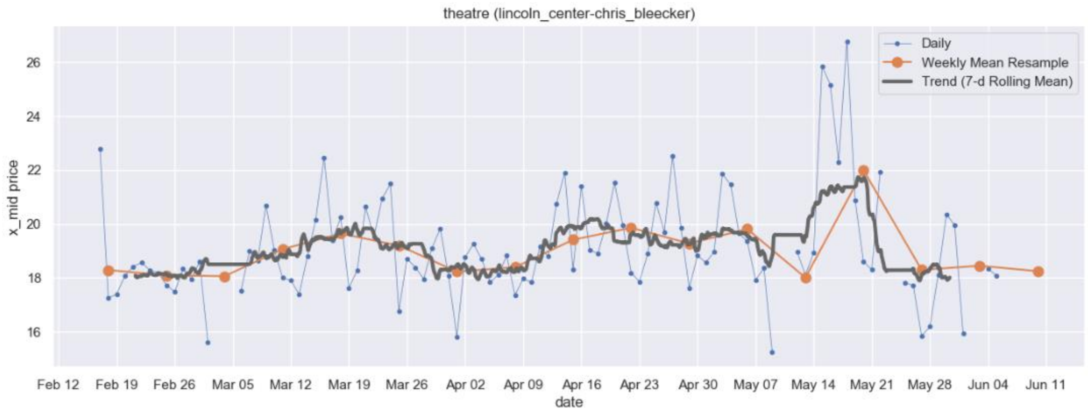


Figure 19. Time series for Uber rides starting from a theatre in NYC

3.2.4 Attraction

Figure 20. is the time series graph for Uber rides starting from an attraction location in NYC. This example shows the daily mean, weekly mean, and rolling average of an Uber ride from times square, a tourist attraction, to Columbus Circle, a train/bus station. This is a "attraction-train station" O-D pair, which is not as significantly affected by rush hours and commuting behaviors as some other O-D pairs. The effect of "attraction-train station" O-D pair, an unique feature of graph 4, is that the highest point of each week is not necessarily Friday. Another feature is that the weekly mean of this time series graph do not fluctuate as much as the weekly mean of other locations and O-D pairs, possibly because of the steady current of visitors to attraction locations. Finally, the weekly and daily mean of this graph is significantly higher than those of other graphs of other locations and O-D pairs between March 5th and March 26th possibly due to the effect of spring break, which has a similar time span. Visitors to these attractions naturally increase during school breaks and holidays, thus increasing the demand and Uber pricing.

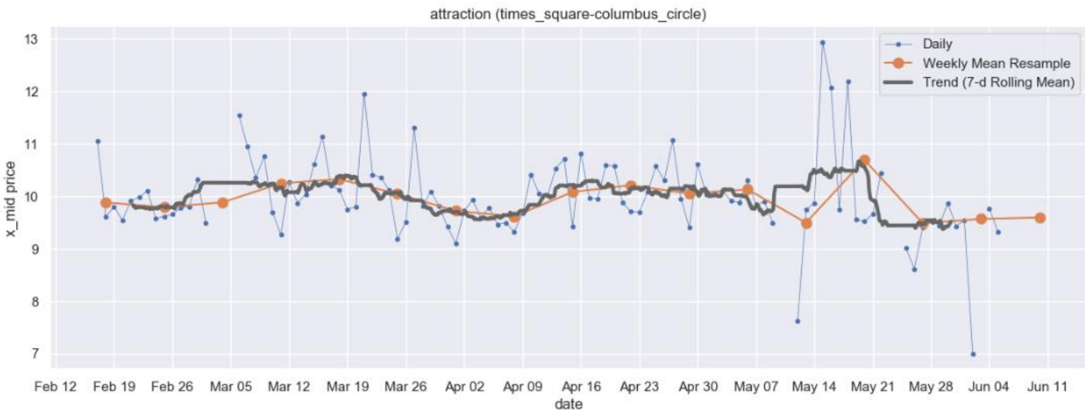


Figure 20. Time series for Uber rides starting from an attraction in NYC

3.2.5 Residential

Figure 21. is the time series graph for Uber rides starting from a residential location in NYC. This example shows the daily mean, weekly mean, and rolling average of an Uber ride from Second avenue and 82nd street., an intersection inside a residential neighborhood, to Grand Central Terminal, a train station. This belongs to the "residential-train station" O-D pair, which, unlike many other O-D pairs, is significantly affected by rush hours and commuting behaviors. From graph 5, we can observe that the daily and weekly mean both reach their bottom between March 26th and April 2nd and peak between May 14th and May 21st. One unique feature of this graph is that the daily means for weekdays are significantly higher than the daily means for weekends, for almost all weeks recorded on this graph. This feature occurs because that rush hours during weekdays increase the demand for Uber rides between residences and train/bus stations, increasing the prices consequently.

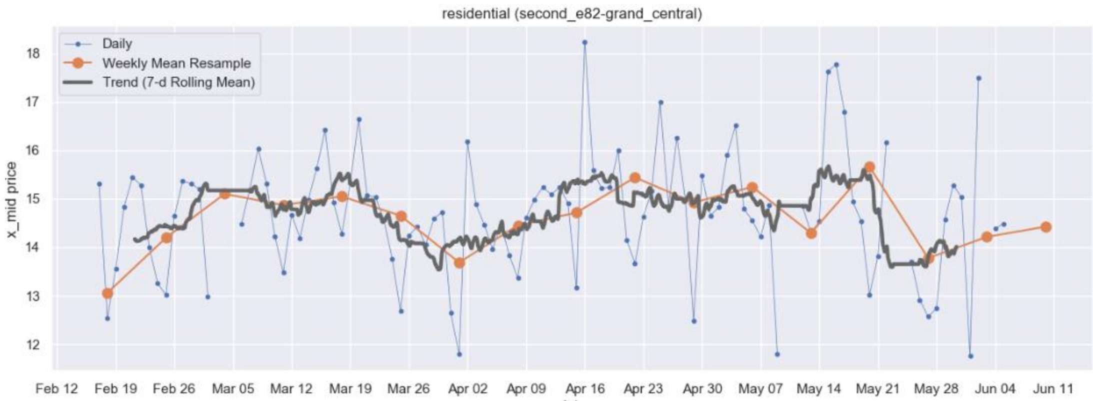


Figure 21. Time series for Uber rides starting from an residential in NYC

3.3 Prediction

Based on our analysis of regression and time series and the data we collected and compared, we are able to evaluate the effects of the variables on Uber's pricing and make predictions. One obvious conclusion we can reach, for example, is that the Uber pricing in NYC goes hand in hand with the commuting patterns, a fairly common phenomenon in today's urban areas: Uber prices rise correspondingly during the rush hours of a typical workday in places association with the working men and women, such as residences and train stations. In other locations such as airports and attractions, these patterns are not always correct due to the time-randomness of arriving tourists and incoming flights. We are able to observe many other different trends in Uber pricing as well, which are most likely the combined product of multiple variables.

Another discovery we can find is that we have seen various degrees of rises in Uber prices for all locations except airport on Friday, which is logical in a real-life context. First of all Friday is the margin between the weekdays and the weekend and genuinely a good time for social activities. During Friday mornings and evenings, the amount of traffic rises considerably due to the effect of rush hours, which is a common feature for weekdays across the board. During Friday nights, however, traffic rises again due to the increases in the amount of people engaging in social activities, such as dining, clubbing, and watching shows and movies, things people wouldn't do during weekdays. Thus, in Fridays, traffic increases not only during rush hours but also during nights, which enables surge pricing during multiple time periods and increases the mean prices overall.

In this paper, We also research the effects of different weather types and fail to see a very clear connection to Uber pricing aside from weather condition squall. If it is not an outlier, There can be an indication that Uber prices increase significantly during squalls. In conclusion, using the regression model formula, we are now able to make predictions about Uber pricing based on the day of a week, the hour of a day, or the concurrent weather when the Uber ride starts.

4. Conclusion

4.1 Insights

The insight of our study is far beyond researching Uber pricing in NYC. In this study, we are able to produce a conclusion on how Uber's rider payment or waiting time is affected by the various factors using big data, applying regression analysis, and formulating economic models. By using these formulas and models, we will be able to test in other cities (exp. Chicago) and expand our study to other popular internet car services (exp. Lyft). More importantly, lying beneath this study can we have a better idea on city dwellers' lifestyle as well as their social and economic needs, prompting more studies in other fields in social sciences, such as psychology, sociology, and anthropology. The topic of online car services is a comprehensive study of economics and computer science, as well as the practical application of economics and financial modeling and big-data-analysis-related technologies. As researchers, we will utilize the basics of econometric modeling and acquire skills in big data analysis using Python, thereby enhancing our understanding of the shared economy represented by online cars. For commuters and city dwellers in NYC, our research findings can also help them design a travelling plan that is more cost efficient. The significant effects that rush hours have on surge pricing for Uber rides can provide some suggestions for Uber users and commuters: in times of rush hours, traffic jams, or severe weather conditions (when demand of rides increases abruptly), it is better not to choose Uber and other online car services but choose the traditional yellow cab, as the surge pricing will increase Uber ride fare in a greater multitude; in other times when demand is steady and normal, Uber and online car services might indeed be a better option, as no surge pricing is applied during these times.

4.2 Methodology

Using big data and economic models, we have finally reached various conclusions about the effect that the weather when the Uber ride takes place, the hour of a day, and the day of a week have on Uber pricing in different types of locations. Such discoveries can bring new light to further studies in fields such as sharing economy or econometrics. By comparing the two research methods we used in the research, time series and regression analysis, we are able to find a more suitable tool for this topic. The time series analysis produces considerable results and offers some visual presentations of our data. However, the regression analysis yields by far the most conclusions and is much more time consuming. As it is more manageable and cost-efficient, regression analysis method can be a better and more suitable option for future researchers in the same field.

5. References

68% of the world population projected to live in urban areas by 2050, says UN | UN DESA Department of Economic and Social Affairs. (n.d.). Retrieved from https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html.

Chen, K., & Sheldon, M. (December 11, 2015). Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the Uber Program. Ucla.edu. Retrieved August 18, 2019, from https://www.anderson.ucla.edu/faculty_pages/keith.chen/papers/SurgeAndFlexibleWork_WorkingPaper.pdf.

Cohen, P., Hahn, R., Hall, J., Levitt, S., & Metcalfe, R. (august 30, 2016). Using Big Data to Estimate Consumer Surplus: The Case of Uber. Retrieved August 18, 2019, from http://www.datascienceassn.org/sites/default/files/Using Big Data to Estimate Consumer Surplus at Uber.pdf

Majaski, C. (2019, June 25). Uber vs. Yellow Cabs in New York City: What's the Difference? Retrieved from https://www.investopedia.com/articles/personal-finance/021015/uber-versus-yellow-cabs-new-york-city.asp

Silverstein, S. (2014, October 16). These Animated Charts Tell You Everything About Uber Prices In 21 Cities. Retrieved from https://www.businessinsider.com/uber-vs-taxi-pricing-by-city-2014-10

6. Appendix

6.1 "airport-attraction" O-D pair regression results

```
                            OLS Regression Results

===================================================================
=============
Dep. Variable:                x_mid   R-squared:
0.565
Model:                          OLS   Adj. R-squared:
0.565
Method:               Least Squares   F-statistic:
 4020.
Date:              Wed, 21 Aug 2019   Prob (F-statistic):
  0.00
Time:                      19:56:23   Log-Likelihood:           -3.55
97e+05
No. Observations:            127076   AIC:                       7.12
0e+05
Df Residuals:                127034   BIC:                       7.12
4e+05
Df Model:                        41

Covariance Type:          nonrobust

===================================================================
===============
                coef    std err          t      P>|t|      [0.025
0.975]
-------------------------------------------------------------------
---------------
const         13.1793      0.141     93.314      0.000      12.902
 13.456
x_distance      1.2159      0.016     77.893      0.000       1.185
   1.246
x_duration      0.0080   3.93e-05    203.990      0.000       0.008
    0.008
Clear          0.3235      0.057      5.649      0.000       0.211
  0.436
Clouds          0.2136      0.059      3.612      0.000       0.098
  0.330
Drizzle        -0.2286      0.043     -5.308      0.000      -0.313
  -0.144
```

Fog	0.3839	0.047	8.150	0.000	0.292	0.476
Haze	1.2445	0.085	14.721	0.000	1.079	1.410
Mist	-0.4340	0.049	-8.922	0.000	-0.529	-0.339
Rain	0.3326	0.042	7.906	0.000	0.250	0.415
Snow	0.5617	0.057	9.847	0.000	0.450	0.673
Squall	1.7657	0.523	3.373	0.001	0.740	2.792
Thunderstorm	0.7965	0.118	6.761	0.000	0.566	1.027
0	0.5822	0.056	10.341	0.000	0.472	0.693
1	-0.6141	0.057	-10.740	0.000	-0.726	-0.502
10	0.5192	0.058	8.992	0.000	0.406	0.632
11	-0.5143	0.057	-9.004	0.000	-0.626	-0.402
12	0.4673	0.056	8.332	0.000	0.357	0.577
13	0.5186	0.056	9.289	0.000	0.409	0.628
14	0.8011	0.056	14.410	0.000	0.692	0.910
15	1.5651	0.057	27.673	0.000	1.454	1.676
16	1.4535	0.058	24.932	0.000	1.339	1.568
17	0.7687	0.059	13.034	0.000	0.653	0.884
18	-0.1164	0.056	-2.079	0.038	-0.226	-0.007
19	0.1475	0.053	2.758	0.006	0.043	0.252
2	-0.4789	0.057	-8.384	0.000	-0.591	-0.367
20	0.4867	0.055	8.926	0.000	0.380	0.594
21	2.0647	0.055	37.550	0.000	1.957	2.173

```
22               2.6380      0.055     48.219      0.000       2.531
  2.745
23               2.5509      0.055     46.239      0.000       2.443
  2.659
3               -0.3570      0.057     -6.285      0.000      -0.468
  -0.246
4                0.3735      0.057      6.562      0.000       0.262
  0.485
5               -0.1451      0.056     -2.575      0.010      -0.256
  -0.035
6               -0.3226      0.055     -5.839      0.000      -0.431
  -0.214
7                0.7066      0.055     12.839      0.000       0.599
  0.814
8                0.2776      0.058      4.779      0.000       0.164
  0.391
9               -0.1935      0.059     -3.285      0.001      -0.309
  -0.078
Friday           1.3727      0.038     35.659      0.000       1.297
   1.448
Monday           2.8808      0.032     89.919      0.000       2.818
   2.944
Saturday         1.1703      0.036     32.415      0.000       1.100
    1.241
Sunday           2.2581      0.033     67.759      0.000       2.193
   2.323
Thursday         1.4066      0.035     39.881      0.000       1.337
    1.476
Tuesday          2.2623      0.033     68.381      0.000       2.197
   2.327
Wednesday        1.8286      0.035     52.190      0.000       1.760
    1.897
==================================================================
=============
Omnibus:                 126750.652   Durbin-Watson:
 0.131
Prob(Omnibus):                0.000   Jarque-Bera (JB):        10278
630.627
Skew:                         4.816   Prob(JB):
0.00
Kurtosis:                    45.994   Cond. No.                   2.3
9e+18
```

6.2 "theatre-train station" O-D pair regression results

```
OLS Regression Results
==================================================================
=============
Dep. Variable:                  x_mid   R-squared:
0.434
Model:                            OLS   Adj. R-squared:
0.433
Method:                 Least Squares   F-statistic:
 2371.
Date:                Wed, 21 Aug 2019   Prob (F-statistic):
   0.00
Time:                        19:56:48   Log-Likelihood:          -2.60
29e+05
No. Observations:              127032   AIC:                      5.20
7e+05
Df Residuals:                  126990   BIC:                      5.21
1e+05
Df Model:                          41

Covariance Type:            nonrobust

==================================================================
===============
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------
---------------
const          2.7719      0.073     38.040      0.000       2.629
   2.915
x_distance      2.1178      0.058     36.526      0.000       2.004
    2.231
x_duration      0.0072   4.51e-05    160.212      0.000       0.007
     0.007
Clear          0.1700      0.027      6.305      0.000       0.117
   0.223
Clouds        -0.0210      0.028     -0.753      0.451      -0.076
    0.034
Drizzle       -0.3708      0.020    -18.275      0.000      -0.411
   -0.331
Fog            0.0387      0.022      1.743      0.081      -0.005
   0.082
Haze           0.1437      0.040      3.614      0.000       0.066
   0.222
```

Mist	0.0247	0.023	1.073	0.283	-0.020	0.070
Rain	0.4798	0.020	24.215	0.000	0.441	0.519
Snow	0.7259	0.027	27.150	0.000	0.673	0.778
Squall	12.5254	0.247	50.734	0.000	12.042	13.009
Thunderstorm	3.3028	0.056	59.453	0.000	3.194	3.412
0	-0.3061	0.026	-11.810	0.000	-0.357	-0.255
1	-0.4143	0.026	-15.907	0.000	-0.465	-0.363
10	-0.6144	0.026	-23.395	0.000	-0.666	-0.563
11	-0.5800	0.026	-22.224	0.000	-0.631	-0.529
12	-0.3153	0.026	-11.958	0.000	-0.367	-0.264
13	-0.1944	0.026	-7.390	0.000	-0.246	-0.143
14	-0.0478	0.026	-1.828	0.067	-0.099	0.003
15	0.2565	0.026	9.804	0.000	0.205	0.308
16	0.4813	0.026	18.356	0.000	0.430	0.533
17	1.5318	0.027	56.972	0.000	1.479	1.585
18	0.9829	0.027	36.739	0.000	0.931	1.035
19	-0.2013	0.026	-7.733	0.000	-0.252	-0.150
2	-0.1729	0.026	-6.543	0.000	-0.225	-0.121
20	-0.5460	0.025	-21.414	0.000	-0.596	-0.496
21	0.6370	0.025	25.157	0.000	0.587	0.687
22	0.9697	0.026	38.003	0.000	0.920	1.020
23	0.4062	0.026	15.787	0.000	0.356	0.457

```
3              0.2339     0.026     8.870     0.000      0.182
  0.286
4              0.3621     0.026    13.786     0.000      0.311
  0.414
5              0.2896     0.027    10.880     0.000      0.237
  0.342
6              0.2582     0.026     9.821     0.000      0.207
  0.310
7              0.2076     0.026     8.011     0.000      0.157
  0.258
8              0.0589     0.026     2.284     0.022      0.008
  0.109
9             -0.5111     0.026   -19.479     0.000     -0.562
  -0.460
Friday          0.3834     0.019    20.029     0.000      0.346
   0.421
Monday          0.3986     0.016    24.905     0.000      0.367
   0.430
Saturday       -0.1531     0.017    -9.138     0.000     -0.186
   -0.120
Sunday          0.1761     0.016    11.098     0.000      0.145
   0.207
Thursday        0.5575     0.017    32.524     0.000      0.524
    0.591
Tuesday         0.7992     0.016    48.828     0.000      0.767
   0.831
Wednesday       0.6102     0.017    35.405     0.000      0.576
    0.644
==================================================================
=============
Omnibus:                 65387.736   Durbin-Watson:
 0.126
Prob(Omnibus):               0.000   Jarque-Bera (JB):        1093
652.067
Skew:                        2.090   Prob(JB):
0.00
Kurtosis:                   16.753   Cond. No.                 3.3
3e+17
==================================================================
=============
```

6.3 "residential-train station" O-D pair regression results

```
OLS Regression Results
```

```
==============================================================================
=============
Dep. Variable:                  x_mid   R-squared:
0.601
Model:                            OLS   Adj. R-squared:
0.601
Method:                 Least Squares   F-statistic:
 4668.
Date:                Wed, 21 Aug 2019   Prob (F-statistic):
  0.00
Time:                        19:58:03   Log-Likelihood:          -2.78
41e+05
No. Observations:              127033   AIC:                      5.56
9e+05
Df Residuals:                  126991   BIC:                      5.57
3e+05
Df Model:                          41

Covariance Type:            nonrobust

==============================================================================
===============
                  coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
---------------
const           2.7137      0.066     41.163      0.000       2.584
  2.843
x_distance       1.4353      0.023     63.017      0.000       1.391
   1.480
x_duration       0.0091   5.28e-05    172.687      0.000       0.009
    0.009
Clear           0.3191      0.031     10.260      0.000       0.258
  0.380
Clouds          0.0276      0.032      0.860      0.390      -0.035
  0.091
Drizzle        -0.2842      0.023    -12.143      0.000      -0.330
  -0.238
Fog             0.1221      0.026      4.768      0.000       0.072
 0.172
Haze            0.2097      0.046      4.572      0.000       0.120
  0.300
Mist            0.1515      0.026      5.730      0.000       0.100
  0.203
```

Rain	0.4103	0.023	17.938	0.000	0.365	0.455
Snow	0.7591	0.031	24.547	0.000	0.699	0.820
Squall	15.3292	0.285	53.851	0.000	14.771	15.887
Thunderstorm	4.2504	0.064	66.197	0.000	4.125	4.376
0	-0.2761	0.030	-9.164	0.000	-0.335	-0.217
1	-0.3687	0.031	-11.912	0.000	-0.429	-0.308
10	-0.7250	0.032	-22.508	0.000	-0.788	-0.662
11	-0.5215	0.032	-16.393	0.000	-0.584	-0.459
12	0.0084	0.032	0.260	0.795	-0.055	0.072
13	0.4463	0.032	13.929	0.000	0.383	0.509
14	0.1639	0.031	5.234	0.000	0.103	0.225
15	0.5297	0.031	16.917	0.000	0.468	0.591
16	0.3595	0.031	11.519	0.000	0.298	0.421
17	0.9784	0.031	31.159	0.000	0.917	1.040
18	0.6766	0.031	21.706	0.000	0.616	0.738
19	0.5056	0.030	16.939	0.000	0.447	0.564
2	-0.4880	0.032	-15.378	0.000	-0.550	-0.426
20	-0.1299	0.030	-4.349	0.000	-0.188	-0.071
21	-0.2905	0.030	-9.818	0.000	-0.349	-0.233
22	-0.3511	0.030	-11.820	0.000	-0.409	-0.293
23	-0.1004	0.030	-3.339	0.001	-0.159	-0.041
3	-0.5395	0.031	-17.157	0.000	-0.601	-0.478

```
4              -0.3747      0.032    -11.800      0.000      -0.437
  -0.312
5              -0.2527      0.032     -7.939      0.000      -0.315
  -0.190
6               0.4443      0.031     14.356      0.000       0.384
  0.505
7               1.2156      0.030     40.646      0.000       1.157
  1.274
8               2.0507      0.031     66.044      0.000       1.990
  2.112
9              -0.2471      0.034     -7.350      0.000      -0.313
  -0.181
Friday          0.4563      0.020     22.637      0.000       0.417
   0.496
Monday          0.5098      0.018     28.954      0.000       0.475
   0.544
Saturday        0.6311      0.016     39.670      0.000       0.600
    0.662
Sunday          0.1902      0.016     12.039      0.000       0.159
   0.221
Thursday        0.1366      0.020      6.919      0.000       0.098
   0.175
Tuesday         0.3465      0.019     18.519      0.000       0.310
   0.383
Wednesday       0.4431      0.020     22.301      0.000       0.404
    0.482
==================================================================
=============
Omnibus:                 105126.323   Durbin-Watson:
 0.125
Prob(Omnibus):                0.000   Jarque-Bera (JB):        5280
551.649
Skew:                         3.676   Prob(JB):
0.00
Kurtosis:                    33.718   Cond. No.                 3.7
8e+17
==================================================================
=============
```

6.4 "train station-residential" O-D pair regression results

```
OLS Regression Results
==================================================================
=============
```

```
Dep. Variable:                  x_mid   R-squared:                       
0.353
Model:                            OLS   Adj. R-squared:                  
0.353
Method:                 Least Squares   F-statistic:                     
 1693.
Date:                Wed, 21 Aug 2019   Prob (F-statistic):              
   0.00
Time:                        19:59:58   Log-Likelihood:           -2.78
79e+05
No. Observations:              127084   AIC:                        5.57
7e+05
Df Residuals:                  127042   BIC:                        5.58
1e+05
Df Model:                          41                                         

Covariance Type:            nonrobust                                         

==================================================================
===============
                 coef    std err          t      P>|t|      [0.025      
0.975]
------------------------------------------------------------------
---------------
const           6.3072      0.157     40.181      0.000       6.000       
   6.615
x_distance       0.1841      0.097      1.903      0.057      -0.006       
    0.374
x_duration       0.0067   6.26e-05    106.717      0.000       0.007       
     0.007
Clear          -0.0502      0.031     -1.611      0.107      -0.111       
   0.011
Clouds          -0.2456      0.032     -7.630      0.000      -0.309      
   -0.183
Drizzle         -0.3694      0.023    -15.750      0.000      -0.415      
   -0.323
Fog              0.0564      0.026      2.200      0.028       0.006       
  0.107
Haze             0.1153      0.046      2.509      0.012       0.025       
   0.205
Mist           -0.1997      0.027     -7.533      0.000      -0.252      
  -0.148
Rain             0.6210      0.023     27.130      0.000       0.576       
   0.666
```

Snow	0.5258	0.031	16.881	0.000	0.465	0.587
Squall	14.2286	0.285	49.873	0.000	13.669	14.788
Thunderstorm	3.8707	0.064	60.217	0.000	3.745	3.997
0	-0.2840	0.030	-9.418	0.000	-0.343	-0.225
1	-0.4350	0.031	-13.978	0.000	-0.496	-0.374
10	-0.5015	0.031	-16.128	0.000	-0.562	-0.441
11	-0.3661	0.031	-11.680	0.000	-0.428	-0.305
12	-0.2694	0.031	-8.627	0.000	-0.331	-0.208
13	0.1050	0.031	3.392	0.001	0.044	0.166
14	0.3559	0.031	11.360	0.000	0.294	0.417
15	0.8466	0.031	27.072	0.000	0.785	0.908
16	0.5193	0.031	16.941	0.000	0.459	0.579
17	1.9437	0.030	64.064	0.000	1.884	2.003
18	2.1812	0.030	72.305	0.000	2.122	2.240
19	0.4121	0.030	13.886	0.000	0.354	0.470
2	-0.5165	0.032	-16.143	0.000	-0.579	-0.454
20	-0.0897	0.030	-3.039	0.002	-0.148	-0.032
21	1.0802	0.029	36.746	0.000	1.023	1.138
22	1.1430	0.030	38.595	0.000	1.085	1.201
23	0.3600	0.030	12.068	0.000	0.302	0.418
3	-0.3051	0.033	-9.339	0.000	-0.369	-0.241
4	-0.1262	0.033	-3.864	0.000	-0.190	-0.062

```
5              -0.5586     0.032   -17.260     0.000      -0.622
 -0.495
6              -0.1781     0.032    -5.639     0.000      -0.240
 -0.116
7               0.3443     0.030    11.328     0.000       0.285
 0.404
8               1.0215     0.030    33.978     0.000       0.963
 1.080
9              -0.3754     0.030   -12.386     0.000      -0.435
 -0.316
Friday          1.4059     0.028    50.372     0.000       1.351
  1.461
Monday          0.7296     0.027    26.860     0.000       0.676
  0.783
Saturday        0.2854     0.027    10.574     0.000       0.233
   0.338
Sunday          0.3356     0.026    12.831     0.000       0.284
  0.387
Thursday        1.2862     0.027    47.024     0.000       1.233
   1.340
Tuesday         0.9684     0.027    35.516     0.000       0.915
  1.022
Wednesday       1.2960     0.028    46.996     0.000       1.242
   1.350
==================================================================
=============
Omnibus:                  79920.478   Durbin-Watson:
 0.095
Prob(Omnibus):                0.000   Jarque-Bera (JB):        1879
470.053
Skew:                         2.636   Prob(JB):
0.00
Kurtosis:                    21.087   Cond. No.                  5.2
6e+17
```

6.5 "attraction-train station" O-D pair regression results

```
OLS Regression Results
==================================================================
=============
Dep. Variable:                x_mid   R-squared:
0.206
```

```
Model:                            OLS   Adj. R-squared:
0.205
Method:                 Least Squares   F-statistic:
 801.3
Date:                Wed, 21 Aug 2019   Prob (F-statistic):
   0.00
Time:                        20:01:24   Log-Likelihood:           -2.38
57e+05
No. Observations:              127057   AIC:                       4.77
2e+05
Df Residuals:                  127015   BIC:                       4.77
6e+05
Df Model:                          41

Covariance Type:            nonrobust

==========================================================================
===============
                 coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------
---------------
const          4.5428      0.085     53.554      0.000       4.377
   4.709
x_distance      4.2228      0.104     40.662      0.000       4.019
    4.426
x_duration      0.0011   6.56e-05     17.401      0.000       0.001
     0.001
Clear          0.0122      0.023      0.536      0.592      -0.032
   0.057
Clouds        -0.0857      0.023     -3.656      0.000      -0.132
   -0.040
Drizzle       -0.3705      0.017    -21.657      0.000      -0.404
   -0.337
Fog            0.0735      0.019      3.931      0.000       0.037
  0.110
Haze           0.0582      0.034      1.736      0.083      -0.008
   0.124
Mist          -0.0636      0.019     -3.291      0.001      -0.101
  -0.026
Rain           0.4643      0.017     27.802      0.000       0.432
   0.497
Snow           0.4441      0.023     19.522      0.000       0.400
   0.489
```

Squall	9.6550	0.208	46.430	0.000	9.247	10.063
Thunderstorm	2.8463	0.047	60.711	0.000	2.754	2.938
0	0.0227	0.022	1.030	0.303	-0.020	0.066
1	-0.1209	0.022	-5.396	0.000	-0.165	-0.077
10	-0.5338	0.023	-23.649	0.000	-0.578	-0.490
11	-0.2950	0.023	-13.076	0.000	-0.339	-0.251
12	-0.1696	0.023	-7.505	0.000	-0.214	-0.125
13	0.0355	0.023	1.577	0.115	-0.009	0.080
14	-0.0623	0.022	-2.813	0.005	-0.106	-0.019
15	0.2788	0.022	12.736	0.000	0.236	0.322
16	0.0164	0.022	0.745	0.456	-0.027	0.059
17	1.2936	0.022	58.539	0.000	1.250	1.337
18	1.2356	0.022	56.064	0.000	1.192	1.279
19	0.3159	0.022	14.524	0.000	0.273	0.359
2	-0.1248	0.023	-5.526	0.000	-0.169	-0.081
20	0.1917	0.022	8.896	0.000	0.149	0.234
21	1.2834	0.021	59.770	0.000	1.241	1.325
22	1.2055	0.022	55.931	0.000	1.163	1.248
23	0.5593	0.022	25.911	0.000	0.517	0.602
3	0.0614	0.023	2.725	0.006	0.017	0.106
4	0.2287	0.022	10.316	0.000	0.185	0.272
5	-0.0462	0.023	-2.045	0.041	-0.091	-0.002

```
6              -0.0022      0.022     -0.100      0.920      -0.045
   0.041
7              -0.0029      0.022     -0.135      0.892      -0.046
   0.040
8              -0.1848      0.022     -8.479      0.000      -0.228
  -0.142
9              -0.6431      0.022    -29.038      0.000      -0.687
  -0.600
Friday           0.9659      0.017     56.163      0.000       0.932
   1.000
Monday           0.5762      0.016     36.037      0.000       0.545
   0.608
Saturday         0.3012      0.017     18.083      0.000       0.269
    0.334
Sunday           0.1926      0.016     11.861      0.000       0.161
   0.224
Thursday         0.8146      0.016     49.573      0.000       0.782
    0.847
Tuesday          0.7756      0.016     48.740      0.000       0.744
   0.807
Wednesday        0.9168      0.017     54.823      0.000       0.884
    0.950
==================================================================
=============
Omnibus:                  72188.161   Durbin-Watson:
 0.064
Prob(Omnibus):                0.000   Jarque-Bera (JB):        1474
440.467
Skew:                         2.324   Prob(JB):
0.00
Kurtosis:                    19.028   Cond. No.                 1.1
1e+17
==================================================================
=============
```

本参赛团队声明所提交的论文是在指导老师指导下进行的研究工作和取得的研究成果。尽本团队所知，除了文中特别加以标注和 致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过 的研究成果。若有不实之处，本人愿意承担一切相关责任。

参赛队员： 巢骏至（Junzhi Chao） 指导老师： 刘哲

2019 年 9 月 2 日