CE 262: ANALYSIS OF TRANSPORTATION DATA

Mini-Project Report

Statistical Analysis of UberPool and UberX Pricing in Berkeley, CA

Group 6

Contributors	Student ID
Ariel Jarvis	3034314911
Jeremy Kho	3034285479
Zining Wang	3034314469

Statement of Purpose

Ride sharing services such as Uber are becoming increasingly popular. Though convenient, these services can be expensive when taken at the wrong times. Our objective is to analyze UberPool and UberX prices between two fixed locations to determine the best time to take an Uber. We will also analyze both UberPool and UberX data to determine which type of Uber is the most economical for the time.

Data Collection

For our mini project we looked at Uber trips within Berkeley, California originating and ending at two specific locations. We used Uber's API for data collection. A Python script was created that made a request to Uber every minute for a week. The script essentially masks itself as a customer that would like to request a ride without actually confirming the ride. This provides us with information on the price, duration, distance, and wait time of the proposed trip. This script was then loaded onto a Virtual Machine on Google Cloud to ensure continuous operation. Using this process, we gathered two separate datasets for the different trip types: UberPool and UberX. A fixed pickup and dropoff location was used in order to limit our independent variables to trip type and time of day. The dependent variable of interest was price.

Define Parameter/Random Variables

Certain conditions affected the completeness and accuracy of our dataset. First of all, the API provided by Uber does not give exactly the same pricing and time estimate as the estimates provided by the Uber application. Uber's API would also periodically kick us off of the service resulting in gaps in our data. We ran the script on a online virtual machine provided by Google. This virtual machine also periodically went offline which also resulted in gaps in our dataset. Below is a graph showing the price fluctuation of the UberPool and UberX data over the period of interest.

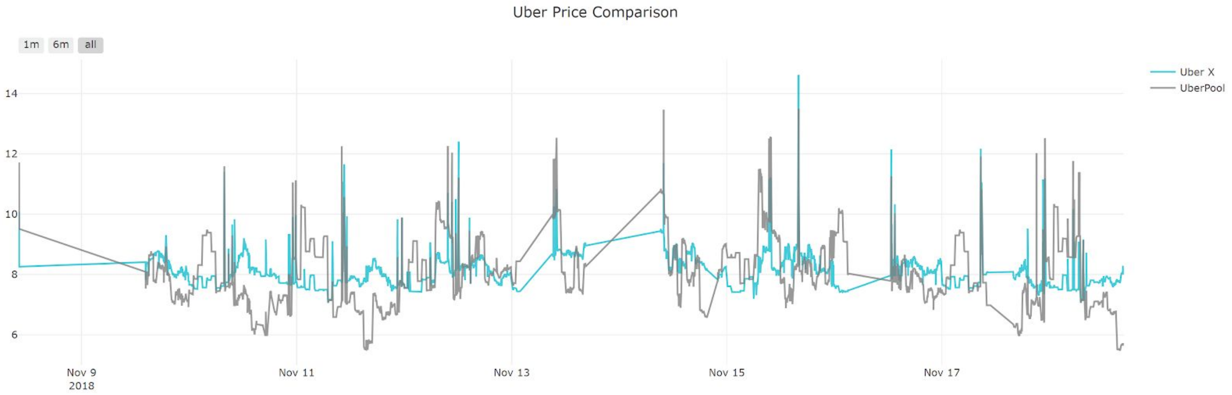


Figure 1 - Uber Price Comparison

Times with highest and lowest prices are summarized below.

	UberPool	UberX
Low	Sunday 3-5 PM	Sunday 2-3 AM, 4-8 AM Monday 1-4 AM Tuesday 1-2 AM Thursday 1-4 AM Friday 1-3 AM
High	Sunday 1-2 AM Monday 7-8 AM Wednesday 9-10 AM Thursday 7-8 AM	Wednesday 9-10 AM

Table 1 - Low and High Price Comparison

Any straight lines found in the graph can be attributed to periods of time where either Uber's API was unresponsive or Google's Virtual Machine crashed. There are also periods where UberPool prices appear to be greater than UberX price estimates. Out of 8902 paired data points, UberX was cheaper 3767 times or 42.3% of the time. UberX was more than $2 cheaper Sunday 1-2 AM, Friday 1-2 AM, and Thursday 7-8 AM. UberX being cheaper does not make intuitive sense and could be attributed to a bug in Uber's API. Regardless, we were still able to collect over 8,000 data points per trip type over the week, which for the purposes of this project is a complete dataset.

We also plotted our trip prices in a histogram as can be seen in the figures below.

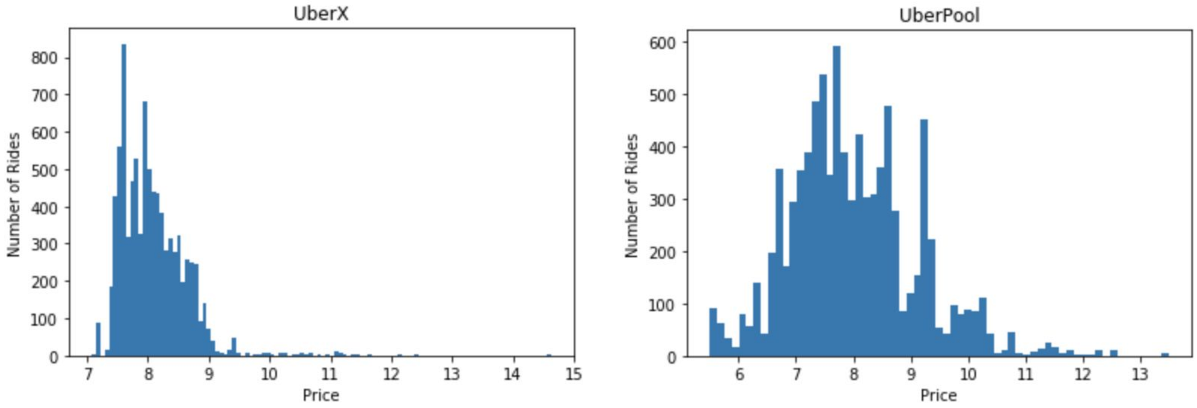


Figure 2 - Uber Price Histograms

Both plots appear to show that price is normally distributed. However the UberPool dataset has a larger variance compared to the UberX dataset which could point to greater price fluctuation.

	UberPool	UberX
Sample Mean	$8.00	$8.07
Sample Standard Deviation	1.12	0.55
Sample Variance	1.25	0.31
Minimum	$5.50	$7.08
Maximum	$13.49	$14.62

Table 2 - Dataset Observations

The highest price for both UberPool and UberX occurred on November 16 at 4 PM. It can be assumed that more Uber trips were requested during the evening rush that day due to the unhealthy AQI.

Hypothesis Testing

We performed two hypothesis tests on our data. A t-test was used to determine whether the means of UberPool prices and UberX prices were significantly different. The Welch's t-test was used because the two samples have unequal variances. We assumed the observations in each sample were independent and identically distributed and normally distributed. The null hypothesis was that the means of the two samples are equal. The Welch's t-test returned a p-value of 3.61×10^{-9}, which means there is a close to 0% chance we would get the same mean if the two groups tested were identical. This value is much smaller than 0.05, so we rejected the null hypothesis using a confidence interval of 95%. The means are significantly different.

A Z-test was used to determine whether the mean of UberPool prices was significantly different than the cost of taking public transportation. The cost of public transportation was calculated by adding the standard bus fare of $2.35 to the value of the extra time it takes to use public transportation. The average travel time using UberPool is 11.25 minutes while the travel time using the bus is 34 minutes. Assuming the value of time is $15 per hour, the time value of the additional 22.75 minutes is $5.68. Therefore the estimated cost of taking public transportation is $8.03.

Because the sample size was large, we could substitute the sample variance for the population variance, giving a plug-in test. This is not an exact Z-test since the uncertainty in the sample variance is not accounted for, however it is still a good approximation. The null hypothesis was that the mean of UberPool prices was equal to the cost of taking the bus. The Z-test returned a p-value of 0.00566, which means there is a 0.566% chance we would get the same mean if the two groups tested were identical. This value is smaller than 0.05, so we rejected the null hypothesis using a confidence interval of 95%. The prices are significantly different.

Regression Analysis

So far in this project, we have been working on two datasets: UberPool Data and UberX Data. For the purpose of this section, we focused on UberX Data to perform a regression analysis. Further specifying the scope of analysis, we extracted weekday data to perform an Ordinary Least Squares (OLS) regression.

Before we began analysis, we took a look at the available features to see how they would fit into the model. Duration and Distance are expected to have positive correlation over Price. The longer and farther the trip, the more expensive due to increased fuel cost and value of time of the driver. However, it is important to note that Distance was a fixed input in our repeated API requests to Uber, and therefore cannot be used as an independent variable. Moreover, the relationship of Time and Price was further dissected by the group. This is visualized below:

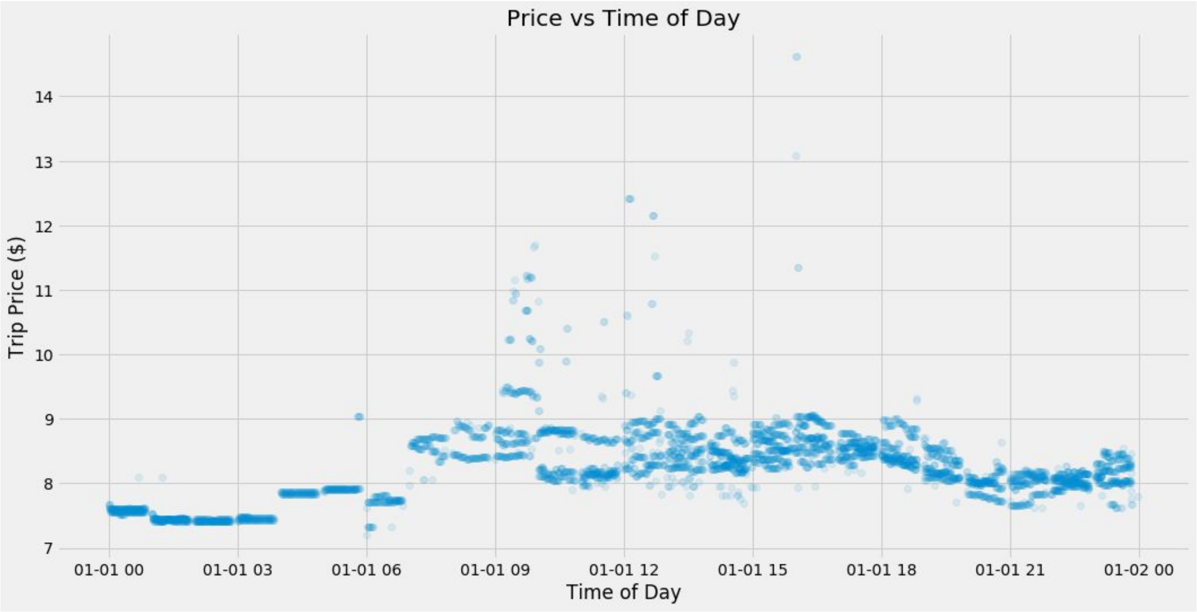
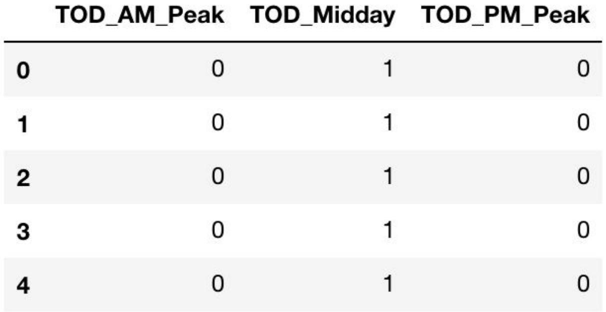


Figure 3 - Price vs Time of Day

We can observe that time of day has a cyclical relationship with price. Generally, prices go up during the day, and go down deep at night when people are home or asleep. Hence, it would not make sense to directly utilize this as an independent variable in an Ordinary Least Squares analysis. Instead, we categorized the day into bins and treat them as categorical data. Based on general information and the plot above, we can divide the data into four mutually exclusive categories - AM Peak (6AM to 9AM), Midday (9AM to 4PM), PM Peak (4PM to 8PM), and Night (8PM to 6AM).

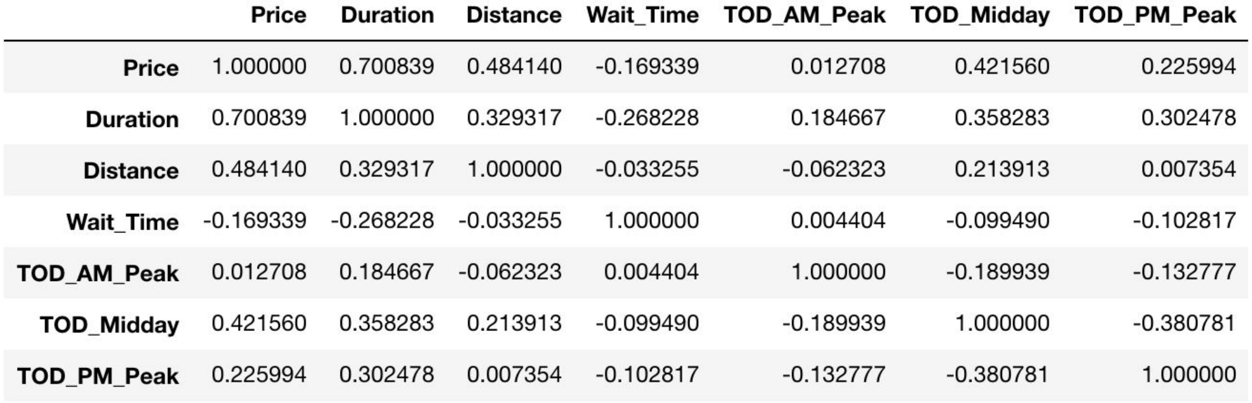
To prepare the data, these four categories were one-hot encoded to three columns ‘TOD_AM_Peak’, ‘TOD_Midday’, and ‘TOD_PM_Peak’. Because the categories are mutually exclusive, the night time of day category is represented when all three of these columns have a zero value. A quick view of how this looks is shown below:



	TOD_AM_Peak	TOD_Midday	TOD_PM_Peak
0	0	1	0
1	0	1	0
2	0	1	0
3	0	1	0
4	0	1	0

Figure 4 - Quick View of One-Hot Encoded Categorical Time of Day

To begin the analysis, we obtained the correlation matrix of the different variables in our dataset, shown below.



	Price	Duration	Distance	Wait_Time	TOD_AM_Peak	TOD_Midday	TOD_PM_Peak
Price	1.000000	0.700839	0.484140	-0.169339	0.012708	0.421560	0.225994
Duration	0.700839	1.000000	0.329317	-0.268228	0.184667	0.358283	0.302478
Distance	0.484140	0.329317	1.000000	-0.033255	-0.062323	0.213913	0.007354
Wait_Time	-0.169339	-0.268228	-0.033255	1.000000	0.004404	-0.099490	-0.102817
TOD_AM_Peak	0.012708	0.184667	-0.062323	0.004404	1.000000	-0.189939	-0.132777
TOD_Midday	0.421560	0.358283	0.213913	-0.099490	-0.189939	1.000000	-0.380781
TOD_PM_Peak	0.225994	0.302478	0.007354	-0.102817	-0.132777	-0.380781	1.000000

Table 3 - Correlation Matrix

From this table, we observe that Duration, Distance, and Time of Day (Midday) have the highest correlation coefficients to Price, in that order. However, in the methodology of collecting the data, distance was a fixed input to the API requests. Hence, Duration and Time of Day are the candidate predictors to the regression model. Although the correlation coefficients vary greatly among the three Time of Day columns (0.01 to 0.42), we will first include all three in this exercise and will improve the model as we go along. As an additional check, we note that the two predictors, Duration and Time of Day, have low correlation with each other.

These two independent variables are plotted against Price in the charts below. These plots visually validate the correlation coefficients of Duration and Time of Day, as we observe the general shape of the data.

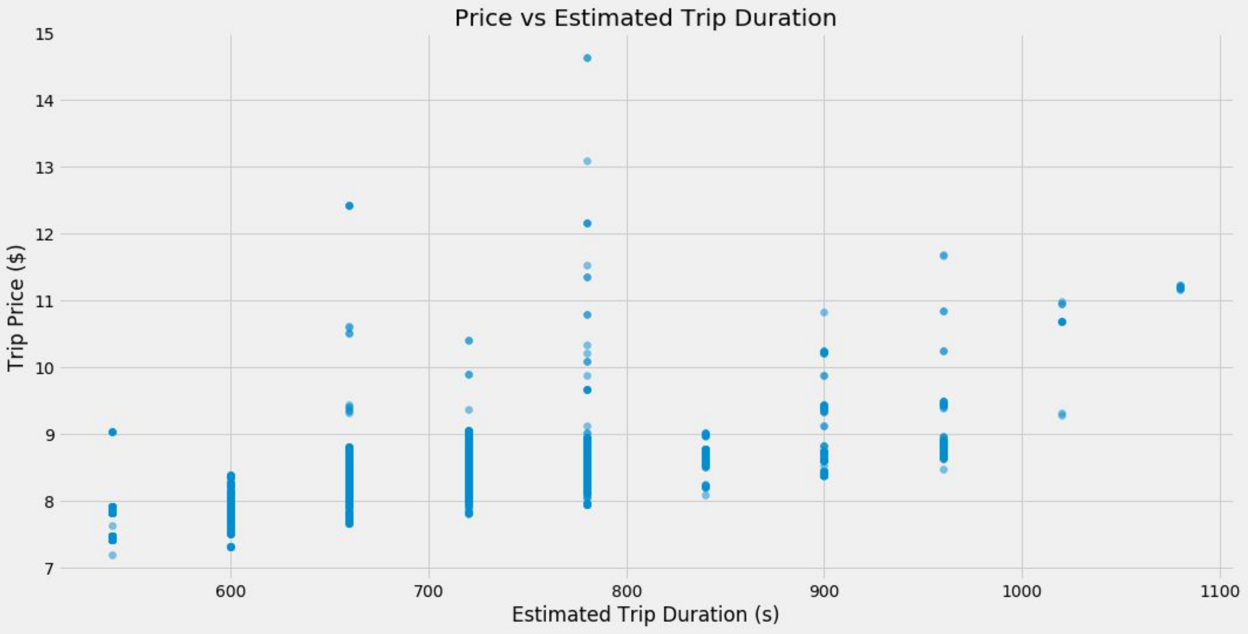


Figure 5 - Price vs Estimated Trip Duration

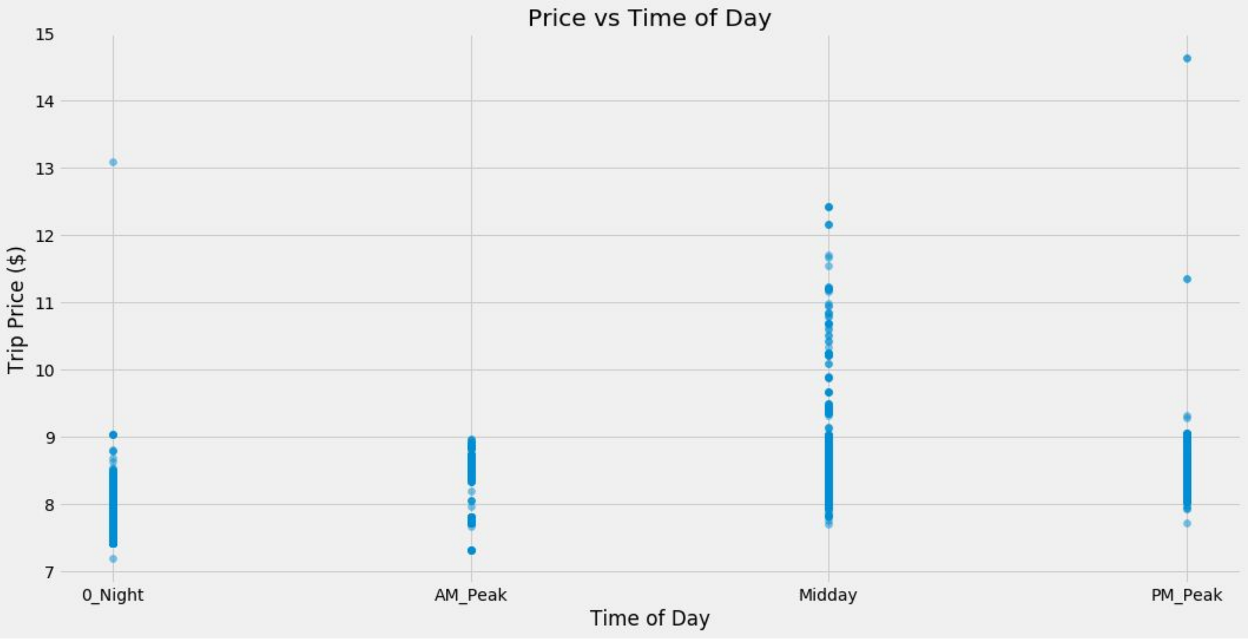
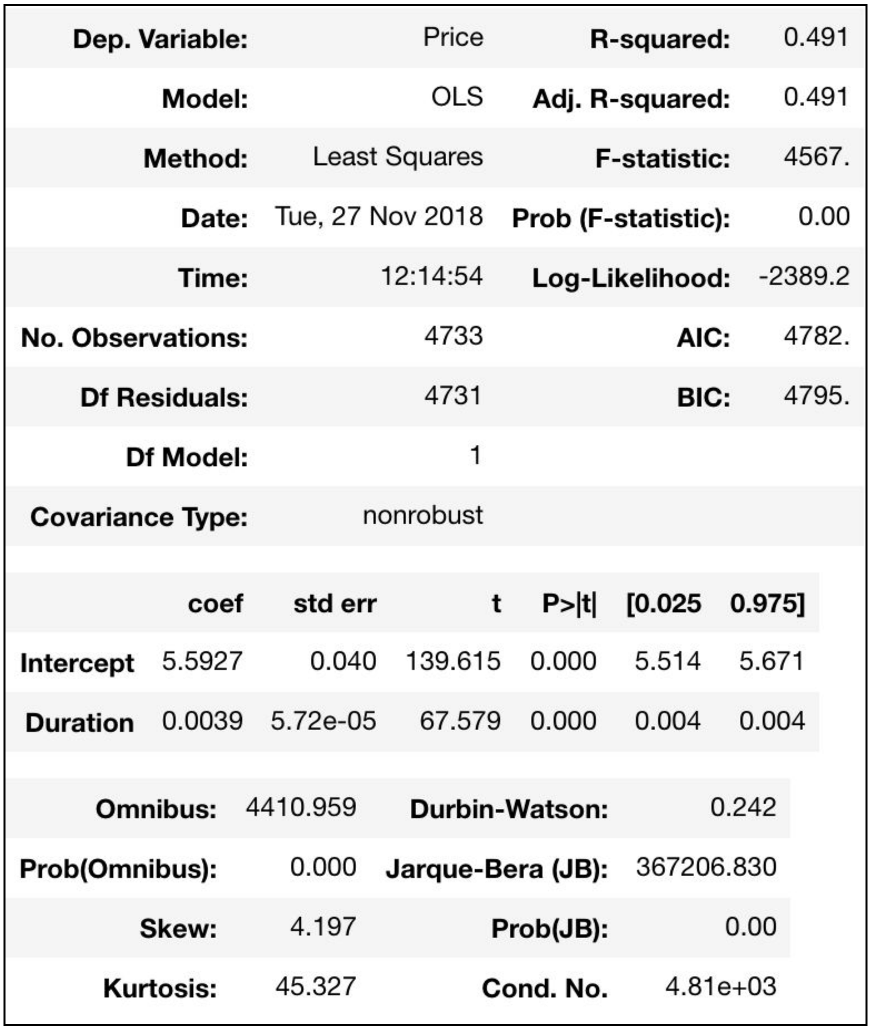


Figure 6 - Price vs Time of Day

Because of this, in building the OLS model, we first utilize the Estimated Trip Duration as the sole independent variable, with Trip Price as the dependent variable. The summary of the results of the model is shown in the table below.



Dep. Variable:	Price	**R-squared:**	0.491
Model:	OLS	**Adj. R-squared:**	0.491
Method:	Least Squares	**F-statistic:**	4567.
Date:	Tue, 27 Nov 2018	**Prob (F-statistic):**	0.00
Time:	12:14:54	**Log-Likelihood:**	-2389.2
No. Observations:	4733	**AIC:**	4782.
Df Residuals:	4731	**BIC:**	4795.
Df Model:	1		
Covariance Type:	nonrobust		

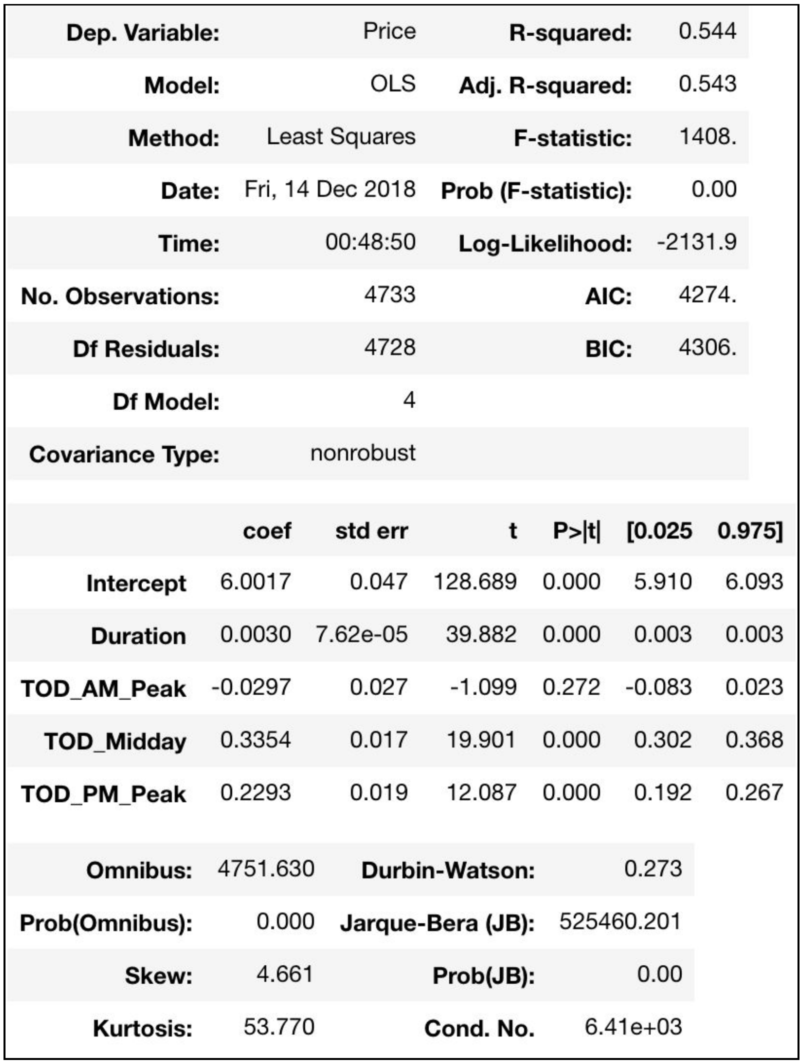
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	5.5927	0.040	139.615	0.000	5.514	5.671
Duration	0.0039	5.72e-05	67.579	0.000	0.004	0.004

Omnibus:	4410.959	**Durbin-Watson:**	0.242
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	367206.830
Skew:	4.197	**Prob(JB):**	0.00
Kurtosis:	45.327	**Cond. No.**	4.81e+03

Table 4 - OLS Regression Results (Duration Only)

With Estimated Trip Duration as the sole predictor, the regression model performs decently with very low p-values for the estimators of the model coefficients, indicating a high level of confidence for these parameters. However, R-squared is relatively low with a value of 0.491. This could indicate that the relationship of the independent and dependent variables is not linear and/or that there are other factors affecting price that have not been captured in this model.

In order to address the latter, another OLS model was constructed adding the three columns of Time of Day. The summary of the results of the new model is shown in the table below.



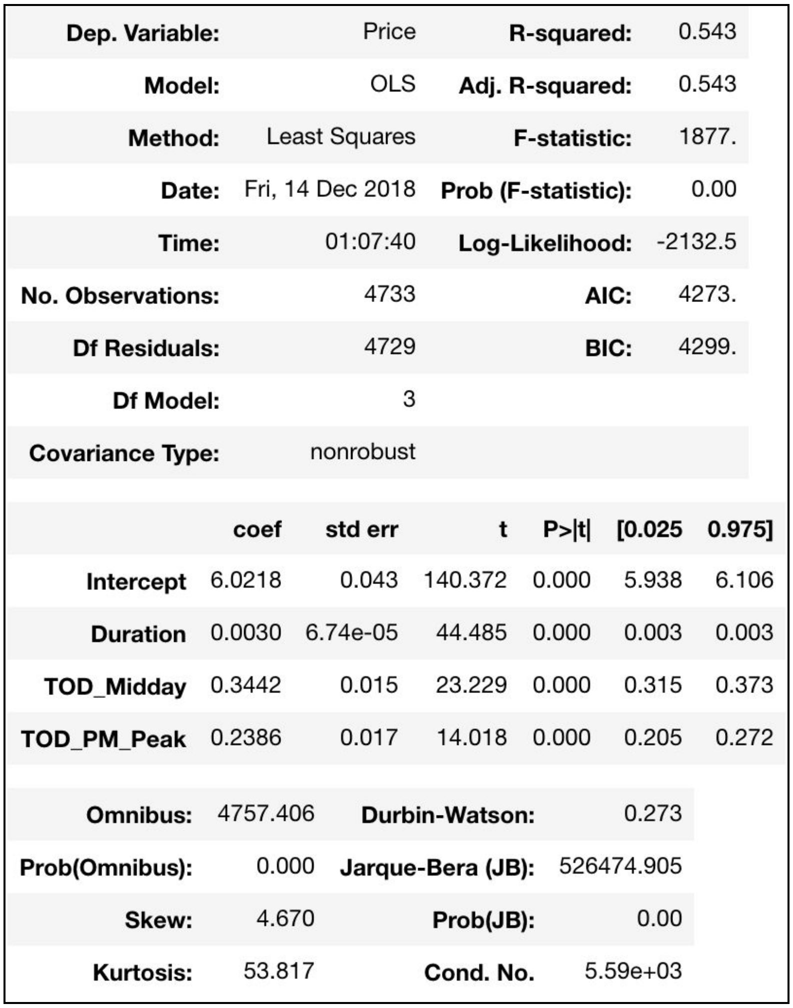
Dep. Variable:	Price	**R-squared:**	0.544
Model:	OLS	**Adj. R-squared:**	0.543
Method:	Least Squares	**F-statistic:**	1408.
Date:	Fri, 14 Dec 2018	**Prob (F-statistic):**	0.00
Time:	00:48:50	**Log-Likelihood:**	-2131.9
No. Observations:	4733	**AIC:**	4274.
Df Residuals:	4728	**BIC:**	4306.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	6.0017	0.047	128.689	0.000	5.910	6.093
Duration	0.0030	7.62e-05	39.882	0.000	0.003	0.003
TOD_AM_Peak	-0.0297	0.027	-1.099	0.272	-0.083	0.023
TOD_Midday	0.3354	0.017	19.901	0.000	0.302	0.368
TOD_PM_Peak	0.2293	0.019	12.087	0.000	0.192	0.267

Omnibus:	4751.630	**Durbin-Watson:**	0.273
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	525460.201
Skew:	4.661	**Prob(JB):**	0.00
Kurtosis:	53.770	**Cond. No.**	6.41e+03

Table 5 - OLS Regression Results (Duration and 3 Time of Day Columns)

We observe that the adjusted R-squared fit improved marginally, now at 0.543. We also observe that most of the p-values remain very low, indicating strength in the estimators for the coefficients of the regression. However, the p-value for the slope of 'TOD_AM_Peak' is very high at 0.272. This means that we cannot reject the null hypothesis that the slope of 'TOD_AM_Peak' is zero. Intuitively, this makes sense as the correlation coefficient we obtained for this variable in the previous section was low at 0.01. Hence, we could opt to remove the variable from the model, which is shown below.



Dep. Variable:	Price	**R-squared:**	0.543
Model:	OLS	**Adj. R-squared:**	0.543
Method:	Least Squares	**F-statistic:**	1877.
Date:	Fri, 14 Dec 2018	**Prob (F-statistic):**	0.00
Time:	01:07:40	**Log-Likelihood:**	-2132.5
No. Observations:	4733	**AIC:**	4273.
Df Residuals:	4729	**BIC:**	4299.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	6.0218	0.043	140.372	0.000	5.938	6.106
Duration	0.0030	6.74e-05	44.485	0.000	0.003	0.003
TOD_Midday	0.3442	0.015	23.229	0.000	0.315	0.373
TOD_PM_Peak	0.2386	0.017	14.018	0.000	0.205	0.272

Omnibus:	4757.406	**Durbin-Watson:**	0.273
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	526474.905
Skew:	4.670	**Prob(JB):**	0.00
Kurtosis:	53.817	**Cond. No.**	5.59e+03

Table 6 - OLS Regression Results (Duration and 2 Time of Day Columns)

After removing 'TOD_AM_Peak' as a variable, the adjusted R-squared stayed the same, but other indicators like AIC and BIC improved showing a reduction in overfitting and a better bias-variance balance than the previous model, without reducing the quality of the regression.

Looking deeper into the characteristics of the final model, we observe that the value of the intercept is quite high. This value of intercept means that if all the variables have a value of 0, the price would be $6.02. Because the AM Peak and Night Time of Day categories are not included in the model, the price values for these two times of day are incorporated in the intercept. This could also indicate a base price that Uber imposes and/or that there are other factors not accounted for in the model, as the value of the R-squared metric suggests.

Furthermore, the coefficient for Duration is very low at 0.003. This is primarily due to the fact that the values of the Duration variable are very high, with its units expressed in seconds. Depending on the purpose of the regression, one way to address this is to standardize all the variables.

On the other hand, TOD_Midday and TOD_PM_Peak are directly comparable since their values are only either 0 or 1. Looking at the values of their coefficients, we can surmise that whether a trip happens during midday has a bigger impact to price than whether it happens during the PM peak. Looking back to the distribution of price vs time of day, this makes sense as the most high-priced trips seem to occur during midday.

With the question of what time of day is best to take an Uber, the model suggests that night time and AM peak are the cheapest times. Although some might intuit that the AM peak should be more expensive than the middle of the night, the model suggests that the difference in price is not statistically significant.

Intuitively, the characteristics of the regression model make sense. The pricing algorithm of Uber is complex and is protected by intellectual property rights. While trip duration and distance are clearly important factors in price determination, Uber has also expressed that their pricing model is based on the real time supply of drivers and real time rider demand, which are datasets unlikely to be obtained by the group. In this analysis, Time of Day merely serves as a proxy for supply and demand distribution.

In order to improve model construction, other regression methodologies can be explored and other forms and sources of data can be obtained to improve approximations of supply and demand.

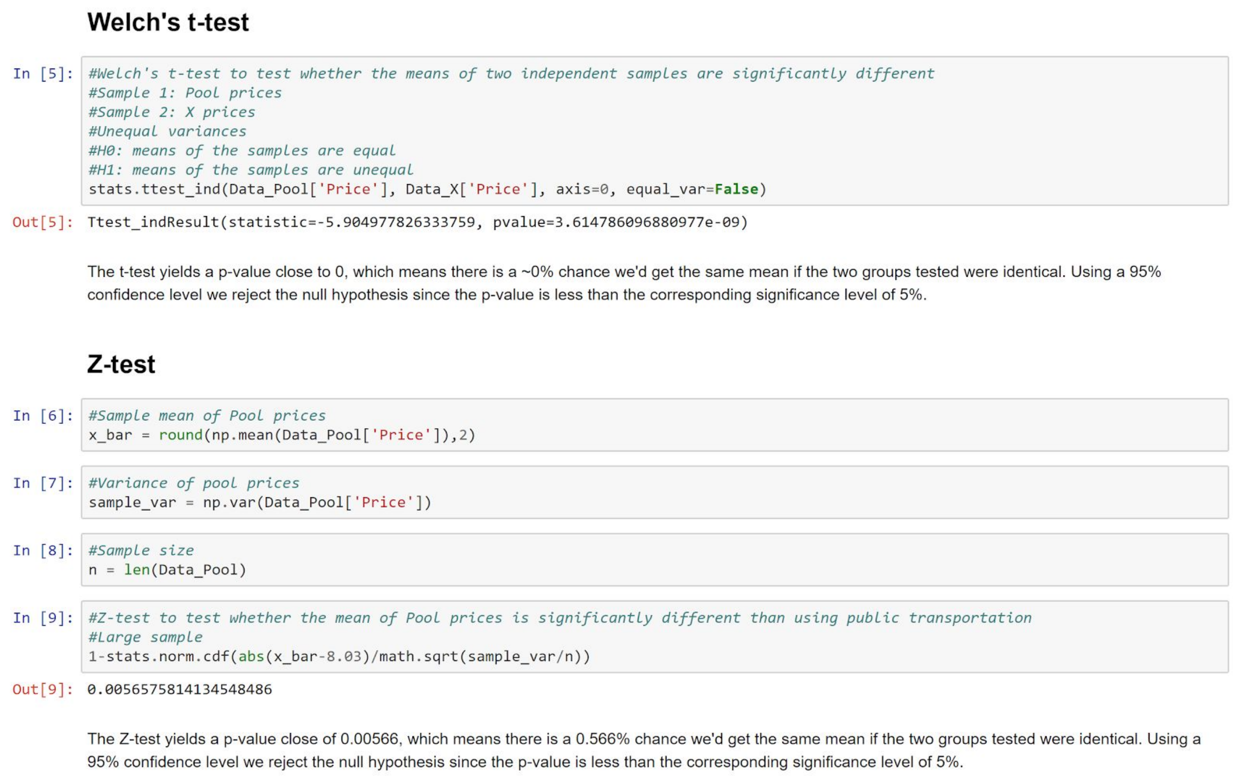
Conclusion

The purpose of this mini project was to analyze UberPool and UberX prices between two fixed locations in Berkeley. Our objective was to determine the best time to take an Uber as well as which type is most economical for a given time. We used Uber's API to collect UberPool and UberX trip data every minute for one week. With price as our dependent variable of interest, we plotted an Uber price comparison (Figure 1). This plot shows the lowest and highest prices and when they occurred, as summarized in Table 1. The plot also reveals when UberX was cheaper, which occurred 42% of the time. To further analyze the datasets, we performed two hypothesis tests. A Welch's t-test comparing the means of UberPool and UberX prices returned a p-value close to 0, meaning the means were significantly different. A Z-test comparing the mean of UberPool prices and the cost of taking the bus returned a p-value below 0.05, meaning the prices are significantly different. We also performed a one variable OLS regression using trip duration and a two variable OLS regression using both duration and time of day, with the time of day one-hot encoded from four categories into three columns. The two variable model returned a higher R-squared value suggesting a better fit. However, the p-value for the slope of the AM Peak Time of Day was high at 0.27. Hence, the model is not statistically confident that the relationship exists between price and whether a trip is taken during the AM peak. After removing the AM Peak column, the R-squared fit of the model stayed the same, but reduced the potential for overfitting to an irrelevant variable.

Sources of error include Uber's API being unresponsive and Google's Virtual Machine crashing for short periods of time. Another possible source of error was data collection occuring when Berkeley was experiencing an unhealthy AQI, which likely resulted in higher than normal demand. Further work could include repeating our analysis with a larger dataset, different sources of data, and different regression methods.

Appendix

Python code for performing hypothesis tests is shown below.



Welch's t-test

```
In [5]: #Welch's t-test to test whether the means of two independent samples are significantly different
        #Sample 1: Pool prices
        #Sample 2: X prices
        #Unequal variances
        #H0: means of the samples are equal
        #H1: means of the samples are unequal
        stats.ttest_ind(Data_Pool['Price'], Data_X['Price'], axis=0, equal_var=False)

Out[5]: Ttest_indResult(statistic=-5.904977826333759, pvalue=3.614786096880977e-09)
```

The t-test yields a p-value close to 0, which means there is a ~0% chance we'd get the same mean if the two groups tested were identical. Using a 95% confidence level we reject the null hypothesis since the p-value is less than the corresponding significance level of 5%.

Z-test

```
In [6]: #Sample mean of Pool prices
        x_bar = round(np.mean(Data_Pool['Price']),2)

In [7]: #Variance of pool prices
        sample_var = np.var(Data_Pool['Price'])

In [8]: #Sample size
        n = len(Data_Pool)

In [9]: #Z-test to test whether the mean of Pool prices is significantly different than using public transportation
        #Large sample
        1-stats.norm.cdf(abs(x_bar-8.03)/math.sqrt(sample_var/n))

Out[9]: 0.0056575814134548486
```

The Z-test yields a p-value close of 0.00566, which means there is a 0.566% chance we'd get the same mean if the two groups tested were identical. Using a 95% confidence level we reject the null hypothesis since the p-value is less than the corresponding significance level of 5%.

The analysis code and dataset can be found at the following GitHub repository in its entirety:

https://github.com/ZiningW/CE262Project