Introduction

On March 28th, 2019, San Francisco based ride hailing company Lyft became a public company via IPO. It is expected that Lyft's major competitor Uber will follow the suit to be IPO soon. Peer-to-peer ridesharing become an important part of people's life everywhere. Ridesharing company implements a dynamic pricing model when calculating the trip cost. As an active ridesharing service customer, I have been interested in the pricing model those companies implemented for a while. The goal of this project is to develop a ridesharing service pricing model that allows customer like me to estimate the trip cost.

Dataset

During my research for related data, I was lucky enough to discover that City of Chicago is the first city to release the ride hailing related data to public. Therefore, I was able to obtain Trip information from Chicago Data Portal. The actual dataset contains 17.4 million rows as the portal started to collect data from November, 2018. For this project, a random sample of 1 million rows were selected and loaded into Jupyter notebook. Dataset contains the following 21 features:

Trip ID: A unique identifier for the trip.
Trip Start Timestamp: When the trip started, rounded to the nearest 15 minutes.
Trip End Timestamp: When the trip ended, rounded to the nearest 15 minutes.
Trip Seconds: Time of the trip in seconds.
Trip Miles: Distance of the trip in miles.
Pickup Census Tract: The Census Tract where the trip began.This column often will be blank for locations outside Chicago.
Dropoff Census Tract: The Census Tract where the trip ended. This column often will be blank for locations outside Chicago.
Pickup Community Area: The Community Area where the trip began. This column will be blank for locations outside Chicago.
Dropoff Community Area: The Community Area where the trip ended. This column will be blank for locations outside Chicago.
Fare: The fare for the trip, rounded to the nearest $2.50.
Tip: The tip for the trip, rounded to the nearest $1.00. Cash tips will not be recorded.
Additional Charges: The taxes, fees, and any other charges for the trip.
Trip Total: Total cost of the trip. This is calculated as the total of the previous columns, including rounding.
Shared Trip Authorized: Whether the customer agreed to a shared trip with another customer, regardless of whether the customer was actually matched for a shared trip.
Trips Pooled: If customers were matched for a shared trip, how many trips, including this one, were pooled. All customer trips from the time the vehicle was empty until it was empty again contribute to this count, even if some customers were never present in the vehicle at the same time. Each trip making up the overall shared trip will have a separate record in this dataset, with the same value in this column.

Pickup Centroid Latitude: The latitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy. This column often will be blank for locations outside Chicago.
Pickup Centroid Longitude: The longitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy. This column often will be blank for locations outside Chicago.
Pickup Centroid Location: The location of the center of the pickup census tract or the community area if the census tract has been hidden for privacy. This column often will be blank for locations outside Chicago.
Dropoff Centroid Latitude: The latitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy. This column often will be blank for locations outside Chicago.
Dropoff Centroid Longitude: The longitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy. This column often will be blank for locations outside Chicago.
Dropoff Centroid Location: The location of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy. This column often will be blank for locations outside Chicago.

Since position features are outside the project scope, they will not be included in the analysis. Trip Total is the sum of Fare, Tip, Additional Charge, however, the goal of this project is to analyze the pricing model, therefore, only Fare is the related feature that is needed in the analysis. A modified dataset is created by eliminating Trip ID, position features and money features except Fare. This narrows down the dataset to 9 features: **Trip Start Timestamp,Trip End Timestamp, Trip Seconds, Trip Miles, Pickup Community Area, Dropoff Community Area, Fare, Shared Trip Authorized,Trips Pooled.**

Data Cleaning

The initial data type of the modified dataset is shown as below:

Trip Start Timestamp	object
Trip End Timestamp	object
Trip Seconds	float64
Trip Miles	float64
Pickup Community Area	float64
Dropoff Community Area	float64
Fare	float64
Shared Trip Authorized	bool
Trips Pooled	int64

Since Pickup Community Area and Dropoff Community Area are actually categorical variables, those two features were converted to categorical with astype ('category'). Trip Start Timestamp and Trip End Timestamp were converted to datetime variables with to_datetime. The updated data type of the modified dataset is shown here:

Trip Start Timestamp	datetime64[ns]
Trip End Timestamp	datetime64[ns]
Trip Seconds	float64
Trip Miles	float64
Pickup Community Area	category
Dropoff Community Area	category
Fare	float64
Shared Trip Authorized	bool
Trips Pooled	int64

After checking the missing data, Trip Seconds, Pickup Community Area, Dropoff Community Area has 155, 75005, 84427 missing data respectively. Based on the definition of Pickup and Dropoff Community Area, all the blanks indicated the area outside Chicago. Blanks will be replaced by 0 to represent area outside Chicago. Imputation was performed with missing values of Fare feature with average fare amount.

A new column which convert the dates to day of the week was created to find some interesting patterns throughout the week.

Exploratory Data Analysis

One of the important features of the dataset is whether the trip is carpooled or not. For this random selected dataset, there were more trips not carpooled as shown below graph.

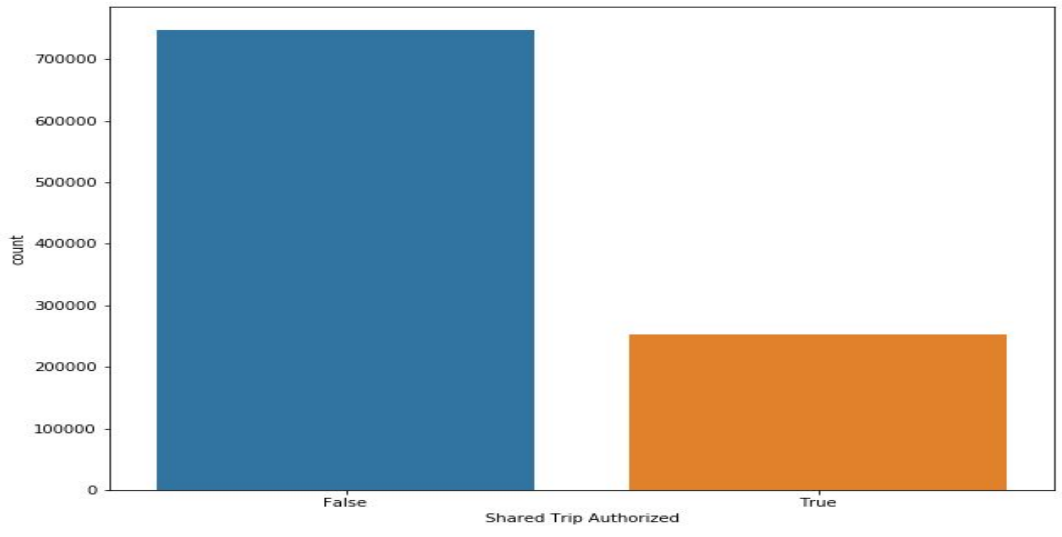


Figure 1. Count of whether the trip was shared trip or not.

Fare, Trip Seconds, and Trip Miles has been drawn against the day of week to find some patterns in below three boxplots with adjusted y scale (Figure 2, Figure 3, Figure 4) . No conclusion regarding the mean of the fare, trip seconds and trip miles against day of week can be drawn from below boxplot. More sophisticated statistical test will be needed to determine whether the difference among the mean is significant against day of week.

One-way ANOVA test has been applied to those pairs of features via below result tables (Table 1, Table 2, Table 3). For Fare against day of the week, overall the model is significant as F = 227.1 and p-value = 5.00e-291 ≈ 0. This indicated that there is significant difference among the average Fare for at least one of the week day. For Trip miles against day of the week, overall the model is also significant as F = 374.8 and p-value = 0. This indicated that there is significant difference among the average trip distance for at least one of the week day. Last, for Trip seconds against day of the week, overall the model is significant as F = 227.1 and p-value = 5.00e-291 ≈ 0. This indicated that there is significant difference among the average trip duration for at least one of the week day.

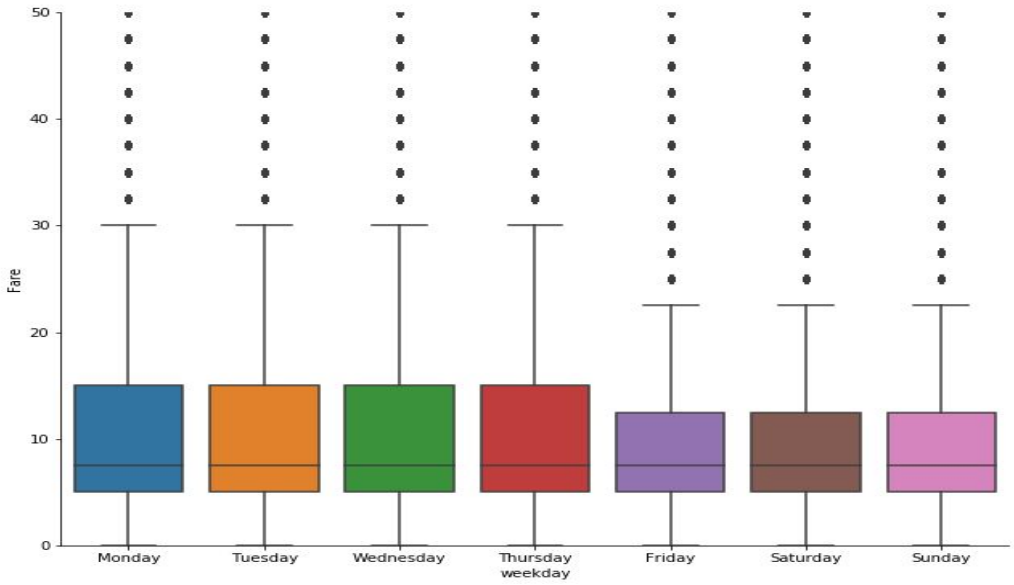


Figure 2. Boxplot of Fare in day of week

OLS Regression Results

Dep. Variable:	Fare	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	227.1
Date:	Mon, 27 May 2019	Prob (F-statistic):	5.00e-291
Time:	18:27:07	Log-Likelihood:	-3.7242e+06
No. Observations:	1000000	AIC:	7.448e+06
Df Residuals:	999993	BIC:	7.449e+06
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.4707	0.024	473.177	0.000	11.423	11.518
C(weekday)[T.Monday]	0.4088	0.039	10.605	0.000	0.333	0.484
C(weekday)[T.Saturday]	-0.6592	0.033	-19.684	0.000	-0.725	-0.594
C(weekday)[T.Sunday]	-0.3159	0.035	-8.969	0.000	-0.385	-0.247
C(weekday)[T.Thursday]	0.3578	0.036	10.008	0.000	0.288	0.428
C(weekday)[T.Tuesday]	0.1588	0.039	4.086	0.000	0.083	0.235
C(weekday)[T.Wednesday]	0.1155	0.038	3.060	0.002	0.042	0.190

Omnibus:	896728.677	Durbin-Watson:	1.926
Prob(Omnibus):	0.000	Jarque-Bera (JB):	133641985.008
Skew:	3.780	Prob(JB):	0.00
Kurtosis:	59.127	Cond. No.	7.31

Table 1. ANOVA for Fare Against Weekday

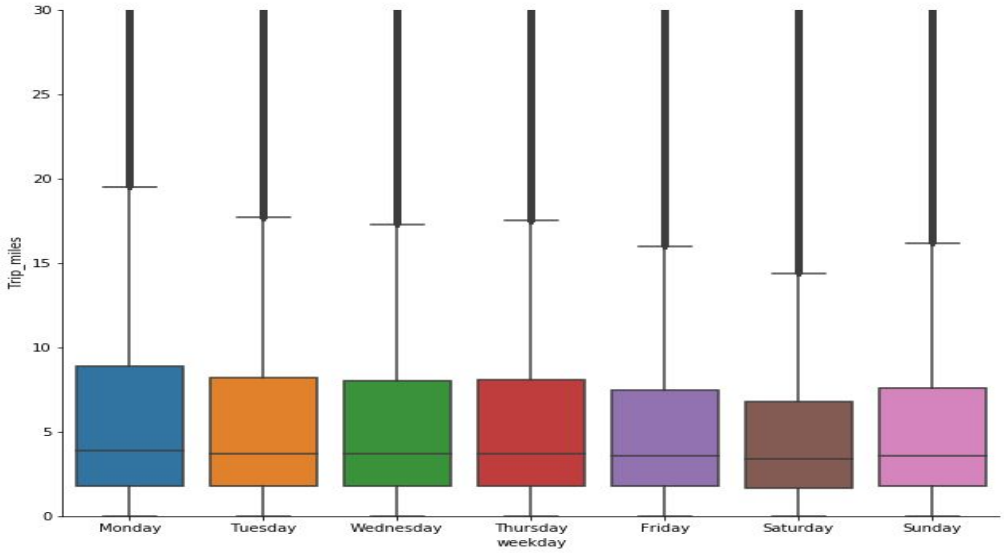


Figure 3. Boxplot of Trip Miles in day of week

OLS Regression Results

Dep. Variable:	Trip_miles	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.002
Method:	Least Squares	F-statistic:	374.8
Date:	Mon, 27 May 2019	Prob (F-statistic):	0.00
Time:	18:28:28	Log-Likelihood:	-3.3731e+06
No. Observations:	1000000	AIC:	6.746e+06
Df Residuals:	999993	BIC:	6.746e+06
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.0293	0.017	353.325	0.000	5.996	6.063
C(weekday)[T.Monday]	0.7158	0.027	26.379	0.000	0.663	0.769
C(weekday)[T.Saturday]	-0.3712	0.024	-15.747	0.000	-0.417	-0.325

C(weekday)[T.Sunday]	0.1690	0.025	6.817	0.000	0.120	0.218
C(weekday)[T.Thursday]	0.3988	0.025	15.845	0.000	0.349	0.448
C(weekday)[T.Tuesday]	0.4363	0.027	15.949	0.000	0.383	0.490
C(weekday)[T.Wednesday]	0.3384	0.027	12.734	0.000	0.286	0.391

Omnibus:	840651.198	Durbin-Watson:	1.906
Prob(Omnibus):	0.000	Jarque-Bera (JB):	72648318.970
Skew:	3.567	Prob(JB):	0.00
Kurtosis:	44.142	Cond. No.	7.31

Table 2. ANOVA for Trip Distance against Weekday

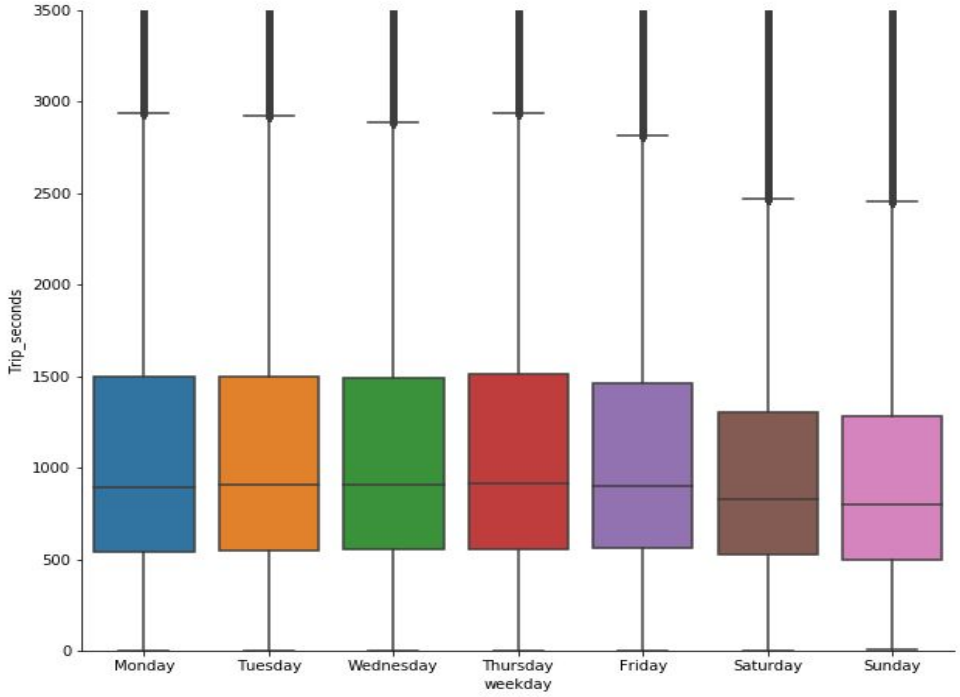


Figure 4. Boxplot of Trip Seconds in day of week

OLS Regression Results

Dep. Variable:	Trip_seconds	R-squared:	0.007
Model:	OLS	Adj. R-squared:	0.007
Method:	Least Squares	F-statistic:	1150.
Date:	Mon, 27 May 2019	Prob (F-statistic):	0.00
Time:	18:27:56	Log-Likelihood:	-8.0925e+06
No. Observations:	1000000	AIC:	1.618e+07
Df Residuals:	999993	BIC:	1.619e+07
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1127.6919	1.913	589.537	0.000	1123.943	1131.441
C(weekday)[T.Monday]	-1.6480	3.042	-0.542	0.588	-7.610	4.314
C(weekday)[T.Saturday]	-116.0739	2.642	-43.926	0.000	-121.253	-110.895

C(weekday)[T.Sunday]	-136.0845	2.779	-48.973	0.000	-141.531	-130.638
C(weekday)[T.Thursday]	27.8426	2.821	9.869	0.000	22.313	33.372
C(weekday)[T.Tuesday]	18.2530	3.066	5.953	0.000	12.243	24.263
C(weekday)[T.Wednesday]	13.9122	2.979	4.670	0.000	8.073	19.751

Omnibus:	612885.938	Durbin-Watson:	1.954
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30620121.981
Skew:	2.287	Prob(JB):	0.00
Kurtosis:	29.720	Cond. No.	7.31

Table 3. ANOVA for Trip Duration against Weekday

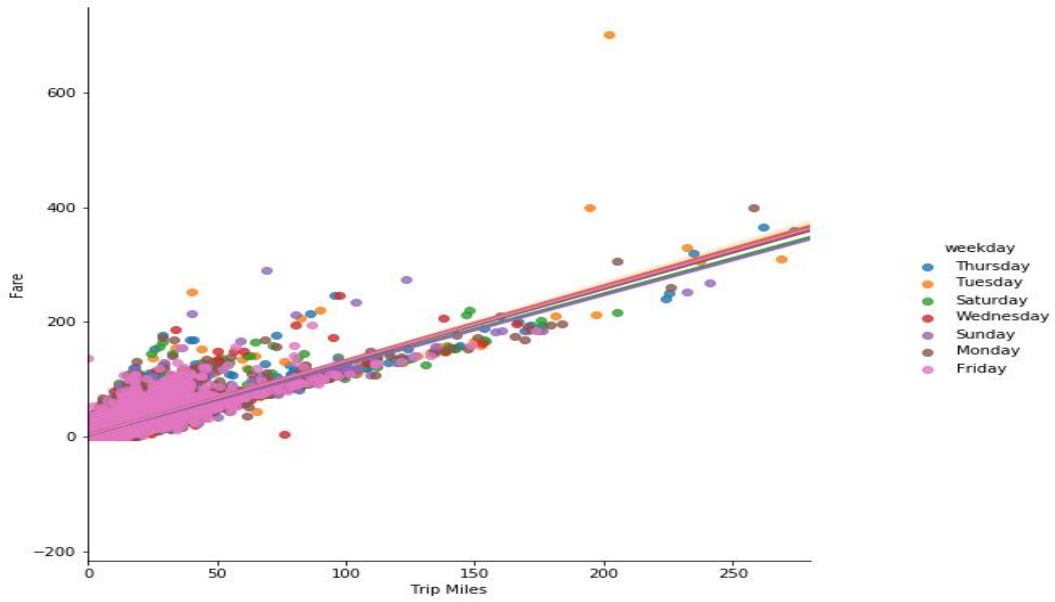
Correlation Analysis

A scatter plot with regression line between Fare and Trip Miles was created to verify the linearity between those two variables. Day of the week was incorporated with different colors. From the graph shown below, there was no significant difference among day of the week. The linear relationship between those two variables could be verified. The Pearson Correlation Coefficient was calculated and shown here:

The Pearson Correlation Coefficient is: 0.8929456863732214

The p-value is: 0.0

As the p-value is very small < α = 0.05, we could reject the null hypothesis. The Pearson Correlation Coefficient showed a strong positive correlation between Fare and Trip Miles which was consistent with the regression plot. Generally as customer travel longer distance, the trip fare also goes up. Since there was strong evidence showing the correlation between Fare and Trip Miles, Trip Miles needed to be added to the model



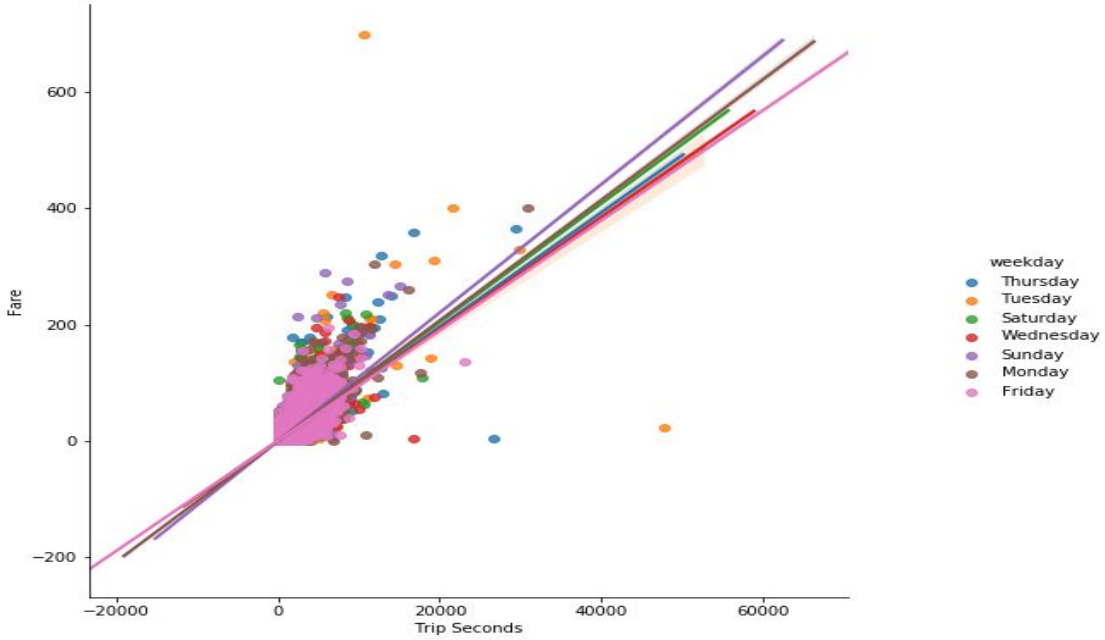
Graph 5. Regression plot between Fare and Trip Miles. Day of Week in colors

Similar to Fare and Trip Miles, the linear relationship could also be observed between Fare and Trip Seconds from the regression plot below. Again, the plot incorporated day of week as different color to find any potential difference among the days. From the graph, there was no big difference among them. They all showed a positive correlation which could be verified with the Pearson Correlation Coefficient calculation results shown here.

The Pearson Correlation Coefficient is: 0.7850275437258065

The p-value is: 0.0

As the p-value is very small < α = 0.05, we could reject the null hypothesis. The Pearson Correlation Coefficient showed a strong positive correlation between Fare and Trip Seconds which was consistent with the regression plot.Generally as trip times goes up, the trip fare also goes up. Since there was strong evidence showing the correlation between Fare and Trip Seconds, Trip Seconds needed to be added to the model.



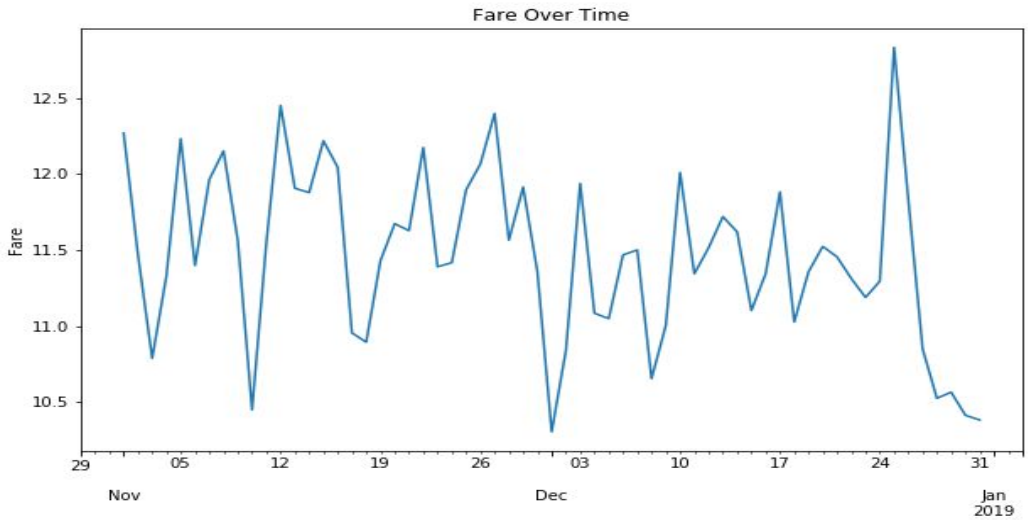
Graph 6. Regression plot between Fare and Trip Seconds. Day of Week in colors

To analyze the correlation between the categorical variables, chi-square test has been implemented among the three categorical variables (Shared Trip Authorized, Pickup Community Area, and Dropoff Community Area) verse weekday. The results has been summarized in below table. The p-values for all the pairs were 0, therefore, we can reject the null hypothesis. Those pairs of variables are indeed dependent on one another.

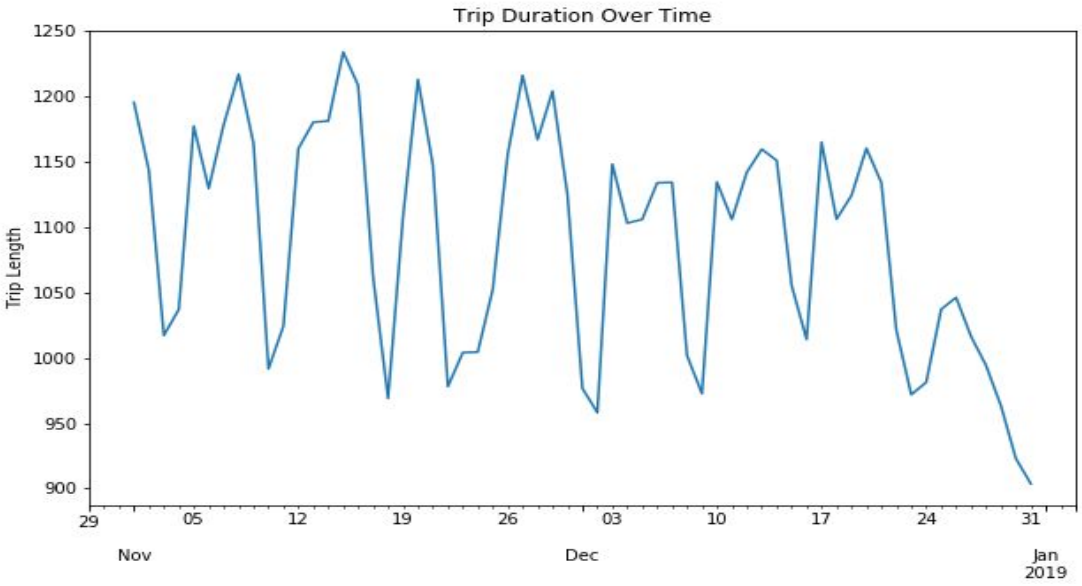
Weekday vs	Chi-Square Test value	p-value	Degree of Freedom
Shared Authorized Trip	3884.10	0.0	6
Pickup Community Area	16563.67	0.0	462
Dropoff Community Area	20776.08	0.0	462

Table 4. Chi-Square Test Results

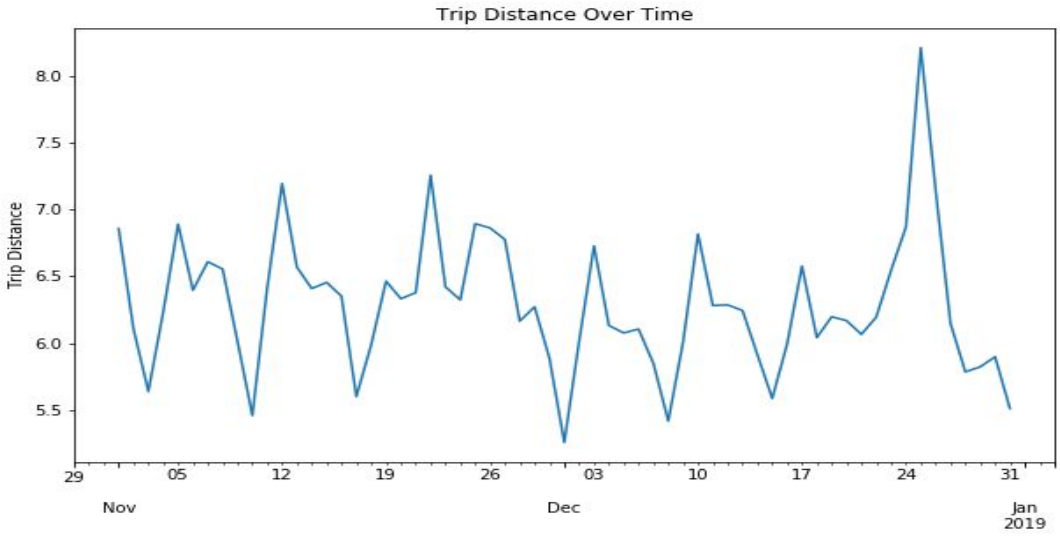
To analyze the impact of time on Fare, Trip Seconds and Trip Miles, those variables are plotted against time that aggregated by 'Day' frequency using Python resample attribute. The result graphs are shown in Graph 7, Graph 8, Graph 9 respectively. For fare over time, no overall downward or upward trend was observed. The fluctuation of the fare over time seems random. However, there was a big spike right after Christmas Eve. For trip duration over time, a slightly overall downward trend was observed. Like fare over time, the trip duration also showed a random fluctuation throughout the period. For trip distance over time, no overall downward or upward trend can be found from the graph. Again, the fluctuation of the trip distance also seems random. However, the same big spike was also found right after Christmas Eve.



Graph 7. Fare over time



Graph 8. Trip Duration over time



Graph 9. Trip Distance over time

Model

Since the goal of this project is to estimate the fare of the trip, it would be a regression problem. Three model method were used: Linear Regression, Ridge Regression, and Elastic Net.

Pre-processing

Since the dataset has several categorical variables, one hot encoding was performed to create dummy variable columns. Pickup Community Area, Dropoff Community Area, Shared Trip Authorized and weekday are the four categorical variables being transformed. After one hot encoding, the number of columns increased from 10 to 174.

Fare was the dependent variable. For the remaining feature data frame (173 column), Trip Start Timestamp, Trip End Timestamp ,Pickup Community Area, Dropoff Community Area, Shared Trip Authorized, and weekday has been excluded from the analysis. Since time series analysis was outside the scope of this project, Trip Start Timestamp and Trip End Timestamp was eliminated. Pickup Community Area, Dropoff Community Area, Shared Trip Authorized, and weekday all had their corresponding dummy columns, therefore, those will not be included as well. The feature dataframe decreased to 167 columns

An OLS was performed to calculate the p-value of each variable in feature dataframe. The result table indicated the following features had a p value > α = 0.05 which indicated that those variables are not statistically significant. Therefore, those variables will be excluded from the model. This leaves the feature dataframe with 138 columns to build model.

	coef	std err	t	P>|t|	[0.025	0.975]
Pickup Community Area_36.0	0.1345	0.124	1.086	0.277	-0.108	0.377
Pickup Community Area_39.0	0.0550	0.061	0.895	0.371	-0.065	0.175
Pickup Community Area_40.0	0.0108	0.086	0.125	0.900	-0.158	0.180
Pickup Community Area_45.0	0.0903	0.135	0.670	0.503	-0.174	0.355
Pickup Community Area_46.0	0.1369	0.093	1.470	0.141	-0.046	0.319
Pickup Community Area_48.0	0.1169	0.128	0.916	0.359	-0.133	0.367
Pickup Community Area_49.0	0.0832	0.067	1.244	0.214	-0.048	0.214
Pickup Community Area_50.0	0.0402	0.142	0.283	0.777	-0.238	0.319
Pickup Community Area_51.0	-0.0823	0.130	-0.631	0.528	-0.338	0.173
Pickup Community Area_52.0	-0.1259	0.191	-0.660	0.510	-0.500	0.248
Pickup Community Area_54.0	-0.0452	0.184	-0.246	0.806	-0.406	0.316
Pickup Community Area_55.0	-0.2934	0.255	-1.153	0.249	-0.792	0.206
Pickup Community Area_57.0	0.1169	0.101	1.156	0.248	-0.081	0.315
Pickup Community Area_62.0	0.1883	0.115	1.634	0.102	-0.038	0.414
Pickup Community Area_65.0	0.1750	0.089	1.976	0.048	0.001	0.349

Pickup Community Area_70.0	0.1297	0.084	1.538	0.124	-0.036	0.295
Pickup Community Area_73.0	0.0595	0.089	0.672	0.501	-0.114	0.233
Dropoff Community Area_36.0	-0.1112	0.129	-0.859	0.390	-0.365	0.143
Dropoff Community Area_39.0	0.0316	0.065	0.488	0.626	-0.095	0.158
Dropoff Community Area_40.0	0.0220	0.090	0.244	0.807	-0.154	0.198
Dropoff Community Area_43.0	0.0928	0.056	1.671	0.095	-0.016	0.202
Dropoff Community Area_47.0	0.3760	0.238	1.577	0.115	-0.091	0.843
Dropoff Community Area_48.0	0.1635	0.130	1.256	0.209	-0.092	0.419
Dropoff Community Area_50.0	0.2346	0.139	1.688	0.091	-0.038	0.507
Dropoff Community Area_54.0	0.2439	0.183	1.332	0.183	-0.115	0.603
Dropoff Community Area_55.0	0.2670	0.241	1.109	0.267	-0.205	0.739
Dropoff Community Area_58.0	0.1075	0.073	1.474	0.140	-0.035	0.250
Dropoff Community Area_60.0	0.1116	0.060	1.854	0.064	-0.006	0.229
Dropoff Community Area_70.0	0.1321	0.087	1.524	0.127	-0.038	0.302

Then the feature data frame has been split into training data (70%) and test data (30%).

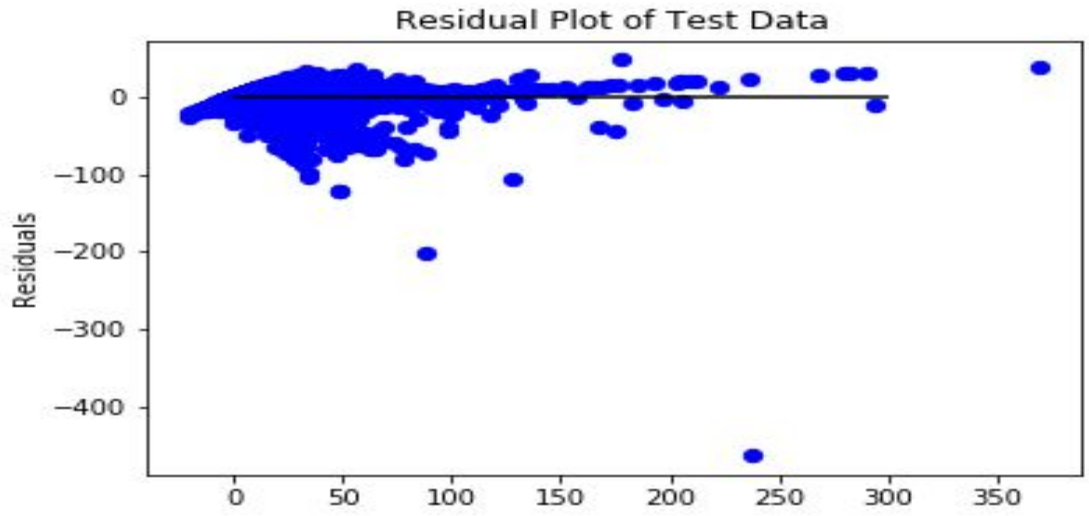
Modeling

Linear regression, Ridge regression and ElasticNet has been used to fit the data and make prediction. The result Mean Square Error, Variance Score and Accuracy Score has been summarized in below table. There is no significant difference among the model.

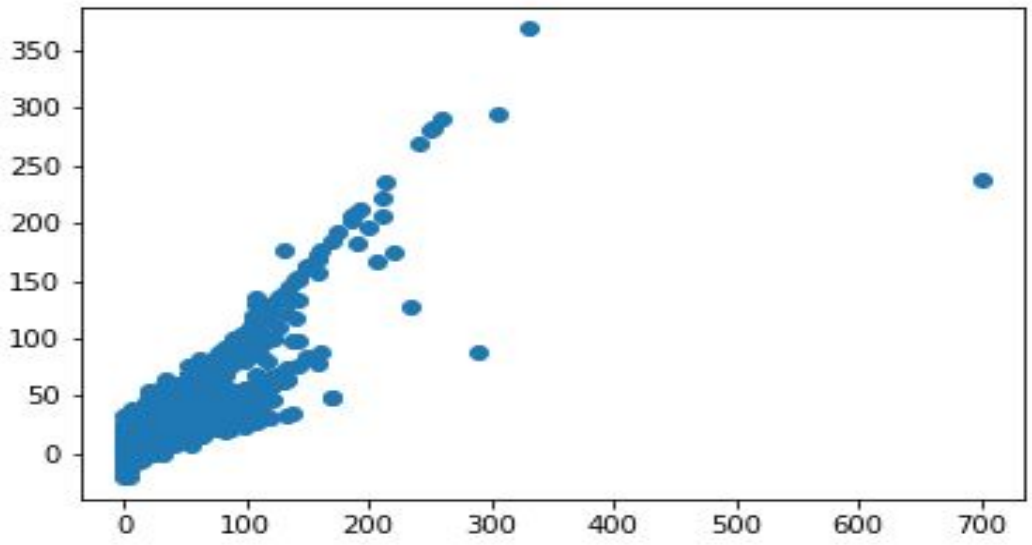
Model	Mean Square Error	Variance Score	Accuracy Score
Linear	15.248	0.852	0.852
Ridge	15.248	0.852	0.852
Elastic Net	15.306	0.852	0.852

The Residual Plot of Test Data of each method were shown in Graph 10 (Linear Regression), Graph 12 (Ridge Regression), Graph 14 (Elastic Net). In all three graphs, the residual points are scattered randomly around the 0 line with no trend indications. All three models made fairly accurate predictions.

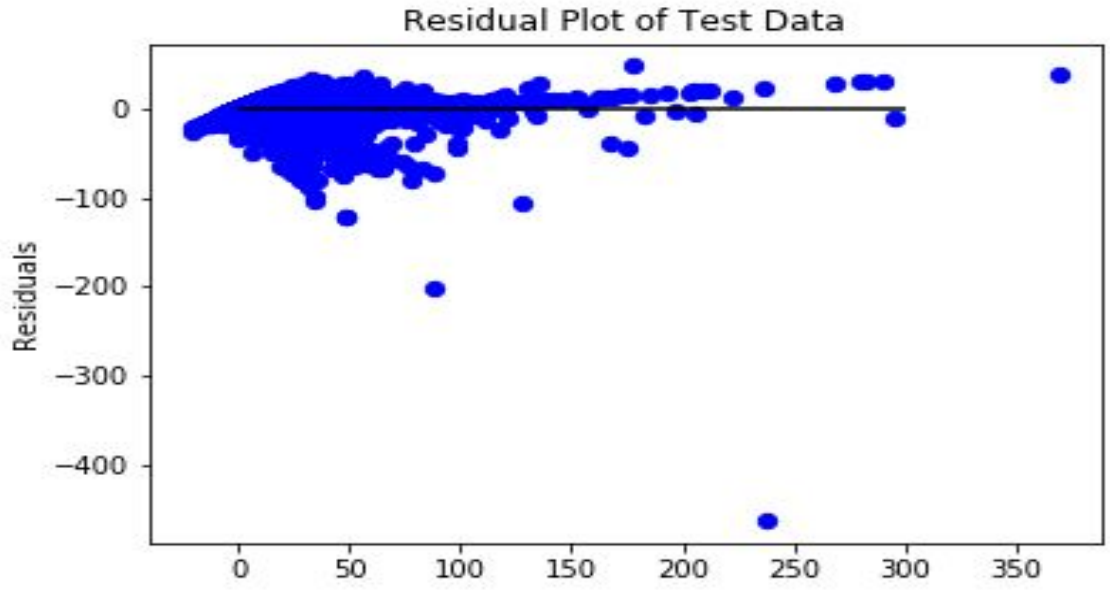
The Prediction Value and Actual Value graphs of each method were shown in Graph 11(Linear Regression), Graph 13 (Ridge Regression), Graph 15 (Elastic Net). All of them showed a fairly linear relationship which also help to verify the accuracy of the predictions.



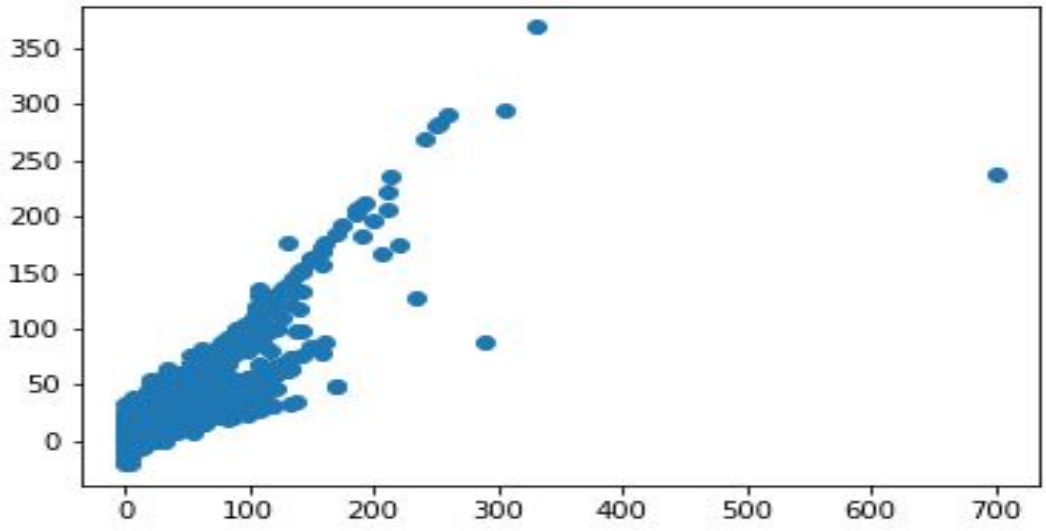
Graph 10. Residual Plot of Test Data Linear Regression



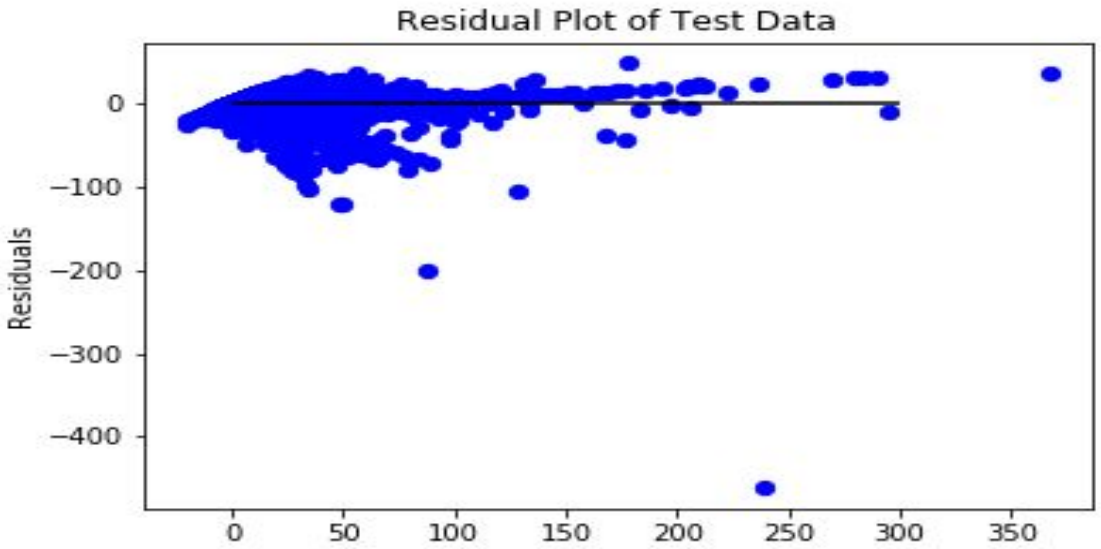
Graph 11. Predict and Actual Value Linear Regression



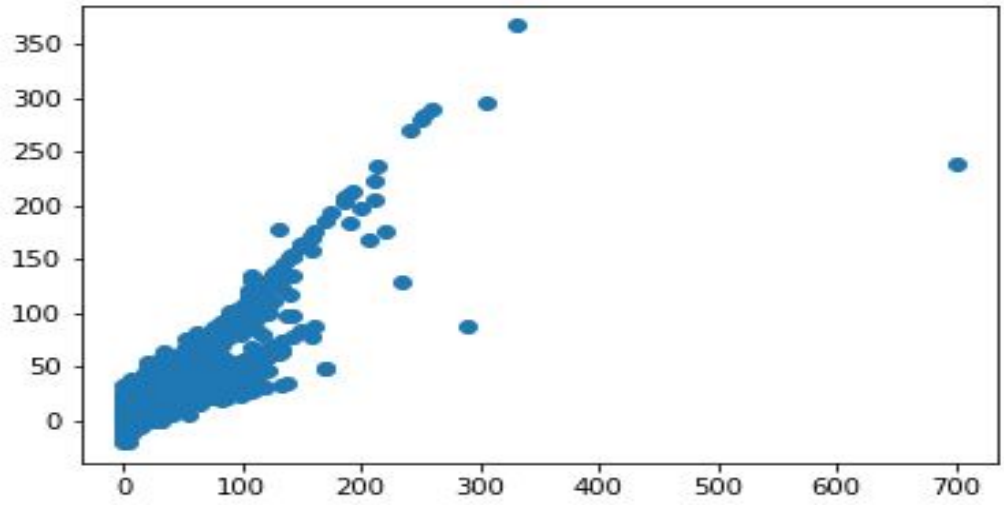
Graph 12. Residual Plot of Test Data Ridge Regression



Graph 13. Predict and Actual Value Ridge Regression



Graph 14. Residual Plot of Test Data Elastic Net



Graph 15. Predict and Actual Value Elastic Net

In summary, the cost of ride sharing service can be estimated using a regression model with variables such as trip distance, trip duration, pickup location, drop off location, day of the week and carpool or not. Residual plot and prediction & actual plot for all three types of regression model (Linear, Ridge, Elastic Net) indicated those three types model provided a fairly accurate prediction. All of them had an accuracy score of at least 0.85.

Conclusion

The Ride Sharing data from the City of Chicago has been processed and fitted into several regression models to estimate the trip fare. As there is no significant difference among the model accuracy, a simple linear regression with an accuracy score of 0.852 can be used. Spatial variables has been converted into categorical variables to incorporate them into the model. Since time variables were not directly included in model building, it will be interesting to include them in future analysis to access the impact of time on fare.