Adam Dulloo 995318

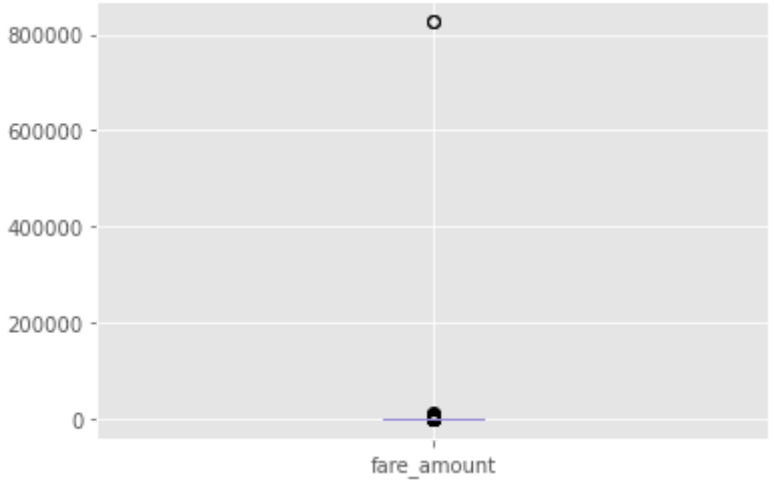
Predicting Fare Amount in New York Yellow Taxi Trips

Goal

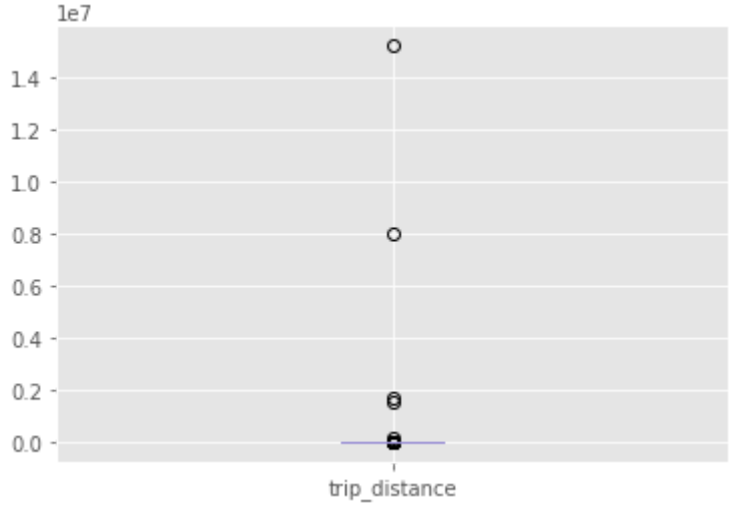
The aim of this project is to build a model to predict the fare amount in New York City's yellow taxi cabs in the year 2015 and month of December. The attributes that will be analysed are tpep_pcikup_datetime, passenger count, trip_distance, pickup_longitude, pickup_latitude and fare_amount. These attributes were selected as they logically correlate directly to the fare_amount or could have some correlation. Trip distance obviously has a direct correlation to the fare, the time of day and which day of the week it is could have a correlation to the fare and the pickup location also could have a correlation to the fare.

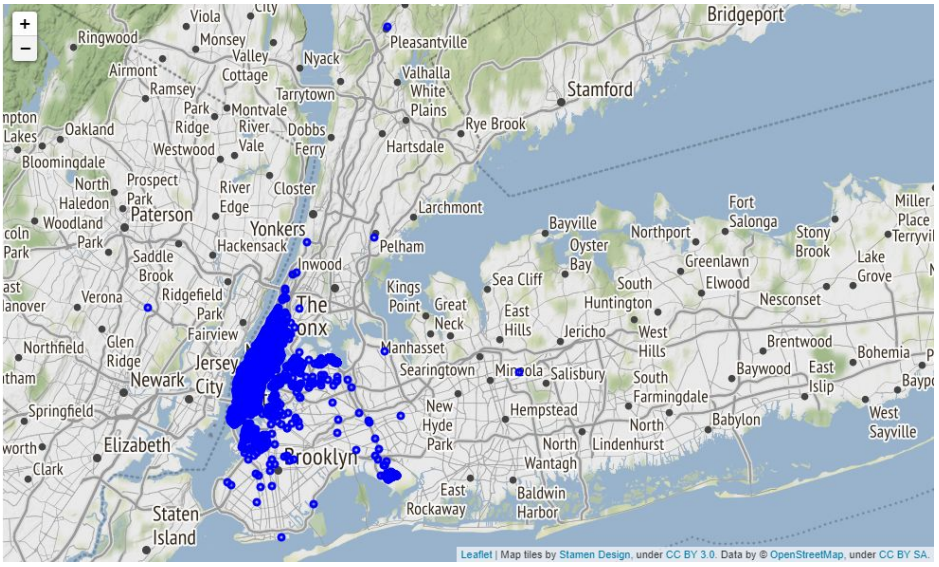
Pre-Processing/Data Cleansing Steps

Looking at the distribution of fare_amount via a boxplot, there are clearly outliers that skew the plot.



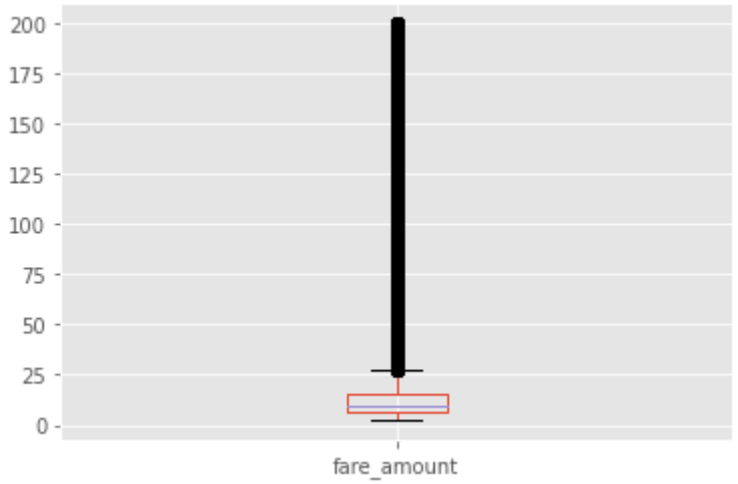
A similar observation can be made for trip_distance. There are extreme outliers that need to be taken care of. Distances of thousands of miles are recorded which are clearly incorrect, these outliers can be seen in the boxplot below. These outliers can also be seen on a pickup map as shown on the map below.

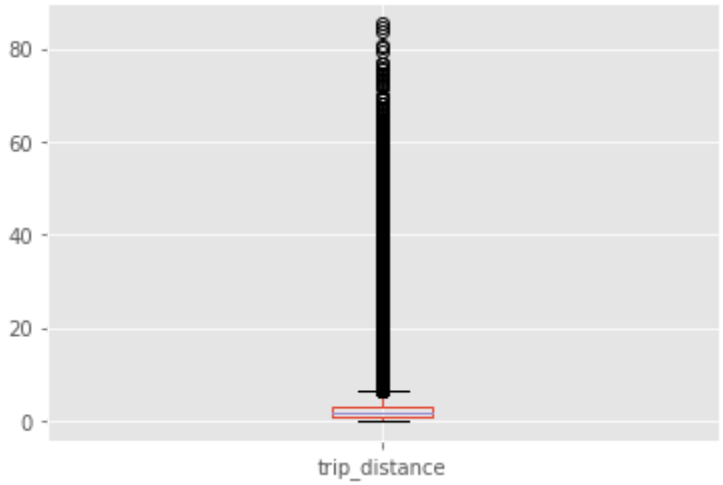


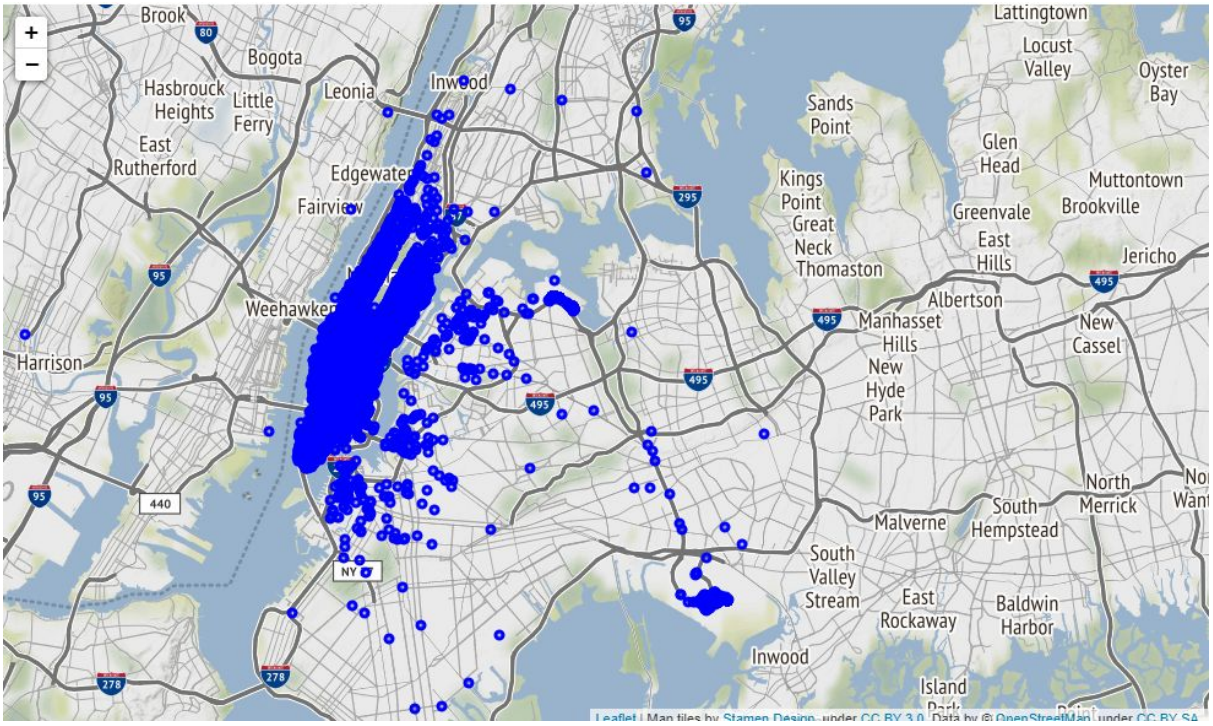


Therefore, the data can be cleaned by first; setting boundary restrictions via the longitude and latitude coordinates that remove any observations that are not in New York City, secondly; restricting the data to 1 to 6 passengers as that is the minimum and maximum number of passengers allowed in a taxi trip, thirdly; setting the fare_amount to $2.50 to $200 as the base rate is $2.50 and $200 is an appropriate upper threshold, lastly; setting the trip distance to 0 to 90 miles as those are the minimum and maximum distances for a taxi trip in New York City, with the 90 miles being an educated judgement. Null values are removed entirely as the dataset is large, so there is no major loss of information.

With these restrictions, the boxplots and pickup map look much better as shown below.

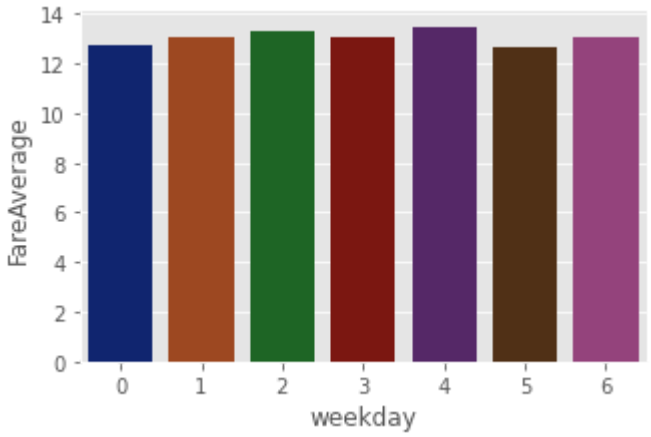


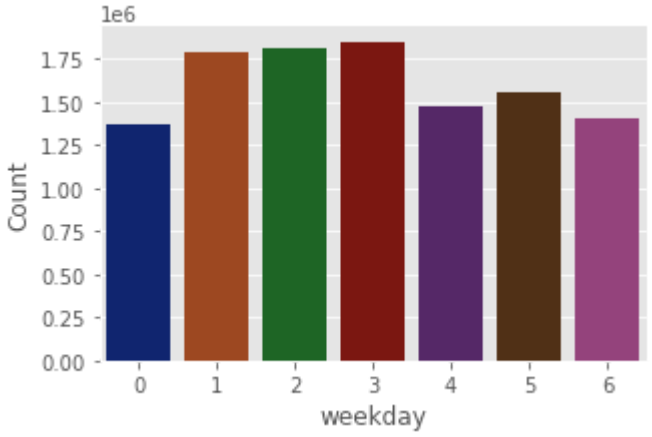


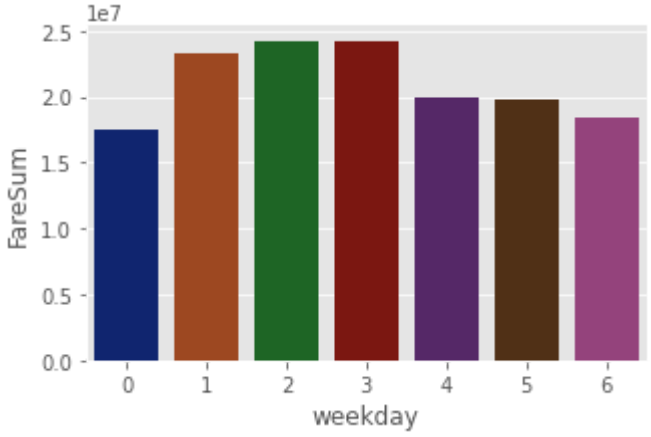


Feature Engineering and Analysis

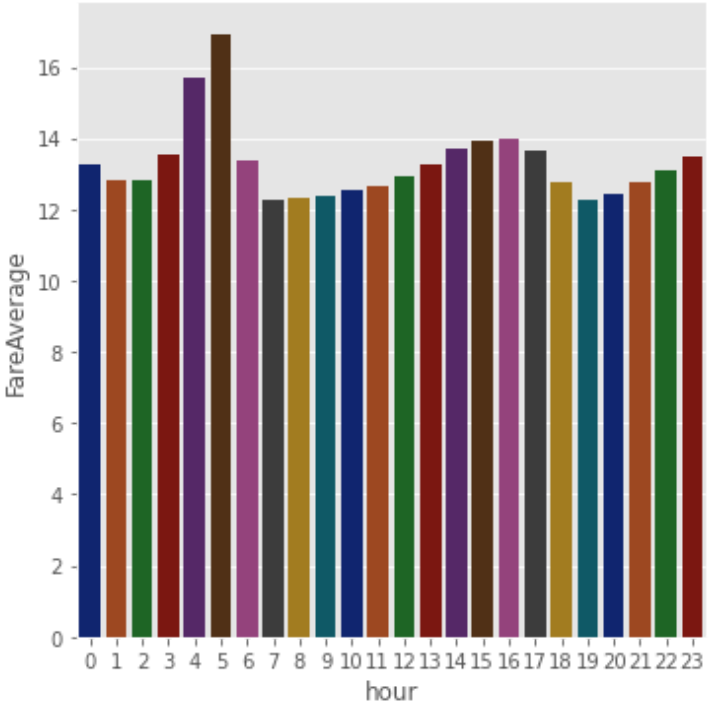
Since the day of the week and time of day were of interest, 2 new features “weekday” and “hour” were created using the tpep_pickup_datetime attribute. “weekday” is an integer from 0 to 6 corresponding to Monday to Sunday and “hour” is an integer from 0 to 23 corresponding to the time of day.



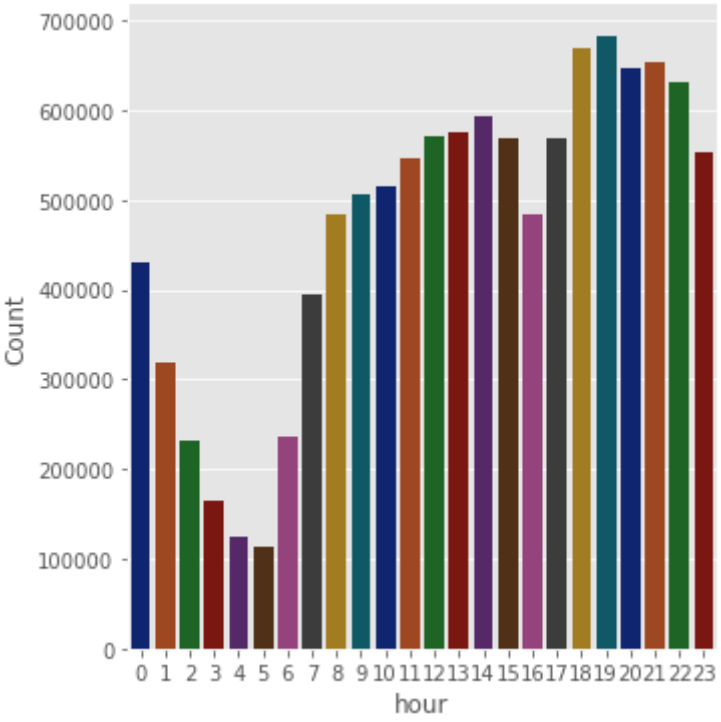


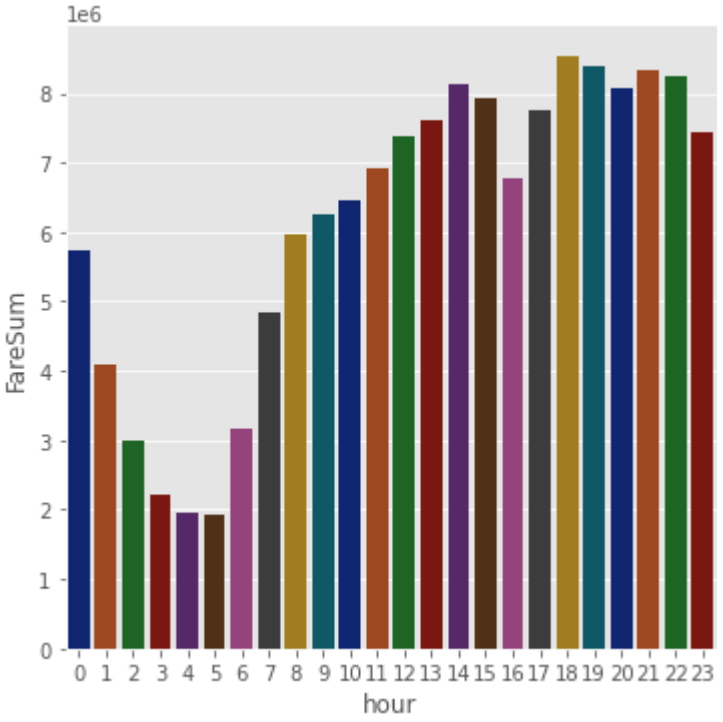


The average fare across the week is the same, so naturally the number of rides correspond to the total sum of fares. Most revenue is generated during the weekdays probably because a lot of people need to travel to and from school/work.



When plotting the time of day, there is evidence to suggest that during rush hour in the morning, the fares are significantly higher and during rush hour in the evening, the fares are slightly higher than the average.

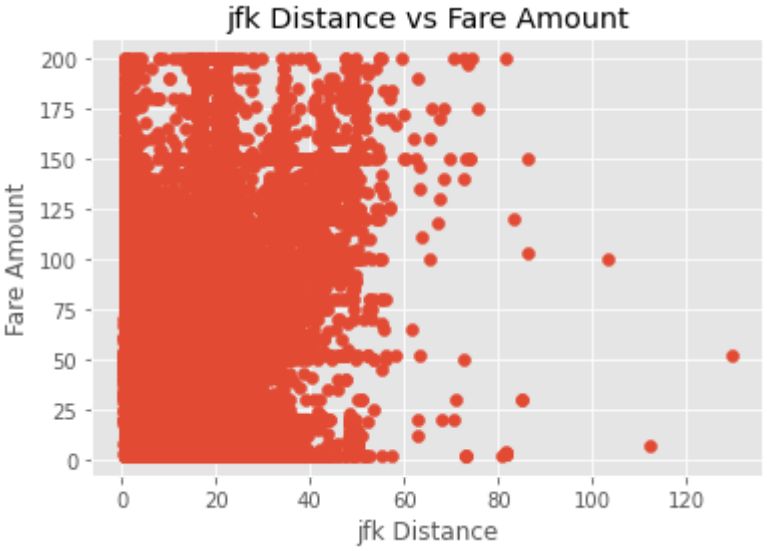


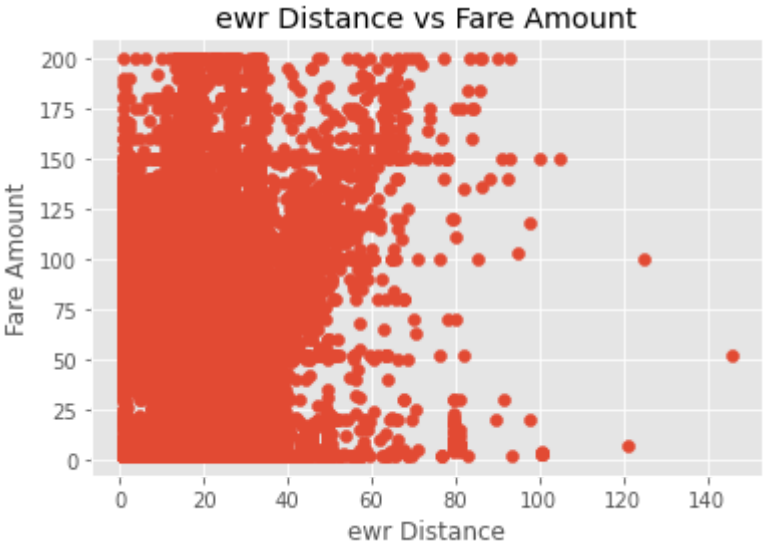


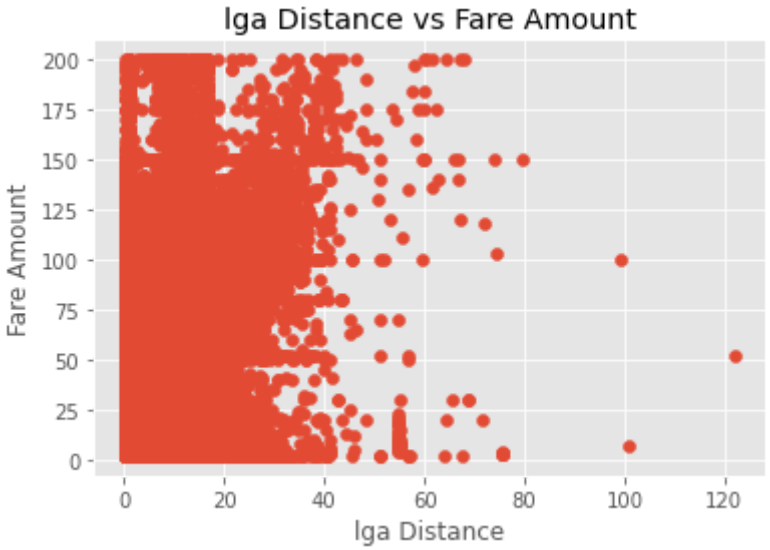
As for the number of trips, there are significantly less trips during the early morning rush hour and more trips during the evening/night time. This makes sense as people generally prefer to take taxi's during the night when they are going out rather than in rush hours. This is because public transport and other forms of transport are a lot quicker during 5-6 a.m. and 4 p.m. so people prefer to not use taxis as there would be a lot of traffic.

Looking at the pickup map from before, there seem to be hotspots. This can be attributed to the 3 major airports located in New York City. These airports are John F. Kennedy International Airport (JFK), LaGuardia (LGA) and Newark (EWR). So, 3 extra features were made to accommodate for this. These features are jfk_dist, ewr_dist and lga_dist which correspond to the distances between the airports and the pickup locations.
Using a scatter plot to first plot the trip_distance and the fare_amount, there is obviously a clear correlation of the higher the distance, the higher the fare as shown below.

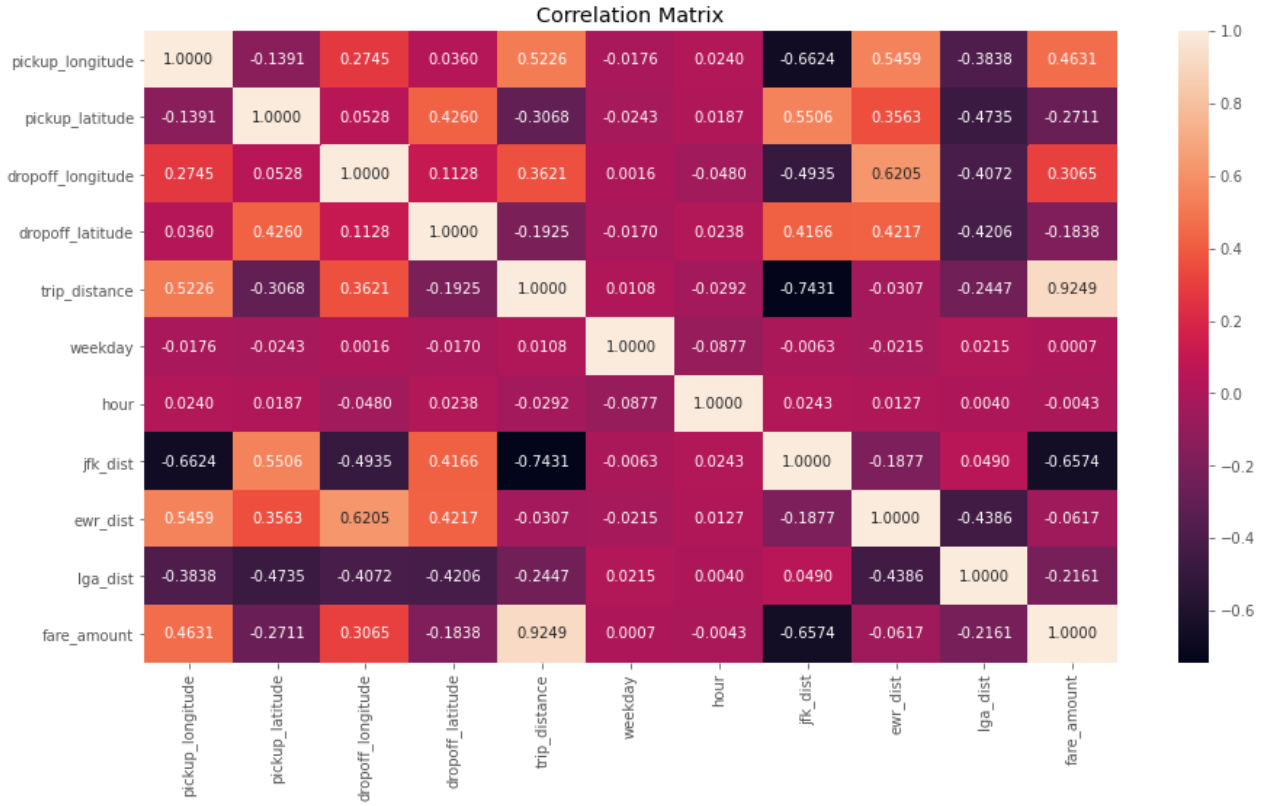
However, when plotting the respective airport's distances (shown below), they are vastly different to the normal trip_distance attribute.



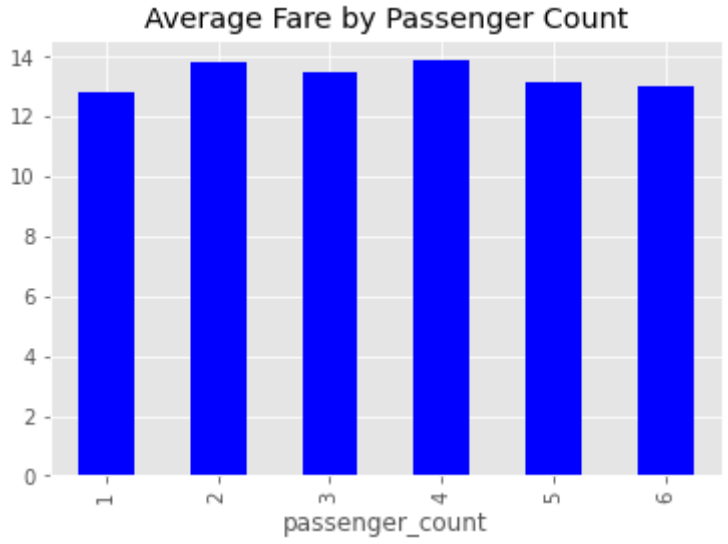




Each airport's distance vs fare amount look similar and suggest that there is a set fee for airport pickups/dropoffs. A correlation matrix show below is plotted to see the correlation between each attribute and the fare amount.



Clearly, the trip distance has the highest correlation of 0.925 with fare amount, meaning it has a very strong positive relationship in determining fare amount. jfk_dist has a moderately strong negative relationship with fare_amount. ewr_dist and lga_dist don't have a strong relationship probably due to JFK airport being the most popular airport in New York where most people go to. The other attributes don't have a strong relationship so they are dropped.
Plotting the average fare by passenger count as shown below, there is no evidence to suggest that it affects the fare amount so it is also dropped.



Modelling

A linear regression model with the ordinary least square method was first used with all the attributes, the results are shown below.

OLS Regression Results

Dep. Variable:	fare_amount	R-squared:	0.859
Model:	OLS	Adj. R-squared:	0.859
Method:	Least Squares	F-statistic:	6.214e+06
Date:	Fri, 09 Oct 2020	Prob (F-statistic):	0.00
Time:	11:06:47	Log-Likelihood:	-3.1734e+07
No. Observations:	11264333	AIC:	6.347e+07
Df Residuals:	11264321	BIC:	6.347e+07
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.0630	0.024	45.192	0.000	1.017	1.109
passenger_count[T.2]	0.1219	0.003	35.257	0.000	0.115	0.129
passenger_count[T.3]	0.1179	0.006	19.788	0.000	0.106	0.130
passenger_count[T.4]	0.2026	0.008	24.851	0.000	0.187	0.219
passenger_count[T.5]	-0.1046	0.005	-19.420	0.000	-0.115	-0.094
passenger_count[T.6]	-0.0992	0.007	-14.671	0.000	-0.112	-0.086
trip_distance	2.8829	0.001	4906.014	0.000	2.882	2.884
jfk_dist	0.1798	0.001	328.517	0.000	0.179	0.181
ewr_dist	-0.0628	0.001	-100.075	0.000	-0.064	-0.062
lga_dist	0.0580	0.001	92.224	0.000	0.057	0.059
weekday	-0.0460	0.001	-72.983	0.000	-0.047	-0.045
hour	0.0365	0.000	193.760	0.000	0.036	0.037

Omnibus:	16845364.813	Durbin-Watson:	1.875
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28933955056.887
Skew:	8.704	Prob(JB):	0.00
Kurtosis:	250.678	Cond. No.	608.

A R squared of 85.9% isn't bad but it isn't great. Despite this, the p values are all 0 which suggest that they are significant in predicting the fare amount despite the small correlation from the correlation matrix. To test this, passenger count was first taken out and the model results are shown below. The R-squared value has decreased to 85.8% which suggests a poorer fit.

OLS Regression Results

Dep. Variable:	fare_amount	R-squared:	0.858
Model:	OLS	Adj. R-squared:	0.858
Method:	Least Squares	F-statistic:	1.366e+07
Date:	Fri, 09 Oct 2020	Prob (F-statistic):	0.00
Time:	11:06:59	Log-Likelihood:	-3.1738e+07
No. Observations:	11264333	AIC:	6.348e+07
Df Residuals:	11264327	BIC:	6.348e+07
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.9332	0.023	39.820	0.000	0.887	0.979
trip_distance	2.8829	0.001	4905.502	0.000	2.882	2.884
jfk_dist	0.1797	0.001	328.155	0.000	0.179	0.181
ewr_dist	-0.0628	0.001	-100.057	0.000	-0.064	-0.062
lga_dist	0.0574	0.001	91.325	0.000	0.056	0.059
hour	0.0380	0.000	202.796	0.000	0.038	0.038

Omnibus:	16841795.071	Durbin-Watson:	1.874
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28876717040.802
Skew:	8.701	Prob(JB):	0.00
Kurtosis:	250.432	Cond. No.	603.

Taking out all attributes except for trip_distance and jfk_distance, the R-squared value decreases again to 85.7%. Both the AIC and BIC increases as well. Therefore, the full model will use these attributes; passenger_count, trip_distance, jfk_dist, ewr_dist, lga_dist, weekday, and hour.

A simple baseline model along with a linear regression model with the ordinary least square method was first used to predict the fare amount. The baseline model only used the means as the predicted value. The RMSE of the baseline model was 10.76% and the full model was 4.0%. This means that the predicted fare amount was on average +- 4.0% of the actual fare amount. The summary of the full model is shown below.

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.944
Model:	OLS	Adj. R-squared (uncentered):	0.944
Method:	Least Squares	F-statistic:	5.782e+06
Date:	Fri, 09 Oct 2020	Prob (F-statistic):	0.00
Time:	11:54:36	Log-Likelihood:	-3.1587e+07
No. Observations:	11264333	AIC:	6.317e+07
Df Residuals:	11264300	BIC:	6.317e+07
Df Model:	33		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
0	248.1285	0.033	7580.414	0.000	248.064	248.193
1	-0.2032	0.004	-53.866	0.000	-0.211	-0.196
2	24.6455	0.041	598.114	0.000	24.565	24.726
3	-8.1694	0.054	-151.276	0.000	-8.275	-8.064
4	10.1053	0.054	185.717	0.000	9.999	10.212
5	0.1295	0.003	37.953	0.000	0.123	0.136
6	0.1133	0.006	19.245	0.000	0.102	0.125
7	0.1826	0.008	22.691	0.000	0.167	0.198
8	-0.1048	0.005	-19.709	0.000	-0.115	-0.094
9	-0.1221	0.007	-18.299	0.000	-0.135	-0.109
10	-0.2661	0.009	-28.913	0.000	-0.284	-0.248
11	-0.3892	0.010	-38.286	0.000	-0.409	-0.369
12	-0.4303	0.011	-37.449	0.000	-0.453	-0.408
13	-0.4624	0.013	-36.315	0.000	-0.487	-0.437
14	-0.5802	0.013	-43.848	0.000	-0.606	-0.554
15	-0.5852	0.010	-57.791	0.000	-0.605	-0.565
16	0.0972	0.009	11.197	0.000	0.080	0.114
17	0.8818	0.008	107.058	0.000	0.866	0.898
18	1.1272	0.008	138.390	0.000	1.111	1.143
19	1.2022	0.008	148.318	0.000	1.186	1.218
20	1.4311	0.008	179.069	0.000	1.415	1.447
21	1.4752	0.008	186.384	0.000	1.460	1.491
22	1.4395	0.008	182.216	0.000	1.424	1.455
23	1.5581	0.008	198.621	0.000	1.543	1.573
24	1.7133	0.008	216.258	0.000	1.698	1.729
25	1.6914	0.008	205.588	0.000	1.675	1.707
26	1.7298	0.008	218.162	0.000	1.714	1.745
27	1.5245	0.008	198.984	0.000	1.509	1.539
28	1.0757	0.008	141.071	0.000	1.061	1.091
29	0.6086	0.008	79.002	0.000	0.593	0.624
30	0.4052	0.008	52.679	0.000	0.390	0.420
31	0.4207	0.008	54.275	0.000	0.406	0.436
32	0.2832	0.008	35.540	0.000	0.268	0.299

Omnibus:	17061117.144	Durbin-Watson:	1.924
Prob(Omnibus):	0.000	Jarque-Bera (JB):	32037231919.225
Skew:	8.910	Prob(JB):	0.00
Kurtosis:	263.656	Cond. No.	35.7

Taxi trips with the following contributed to a higher fare amount; trips with a higher distance and trips with a lower distance to the airports especially JFK airport (pickups and dropoffs to the airports). There is also no or very little correlation for the amount of passengers, time of day and which day of the week it is to the fare amount. The R-squared value is 94.4% which is significantly better than 85.9%. Therefore, the model does a good job predicting the fare amount. However, with a condition number of 35.7 suggests that there is some multicollinearity which means that one or more columns are close to a linear combination of the rest of the columns. Also a high kurtosis of 263.7 suggests that the data has some heavy tails which mean that there are likely to be extreme outliers that are still in the dataset. The large Jarque-Bera value indicates that the errors are not normally distributed and the Durbin-Watson value indicates that there is strong positive autocorrelation, this means that an error of a given sign tends to be followed by an error of the same sign.

Conclusion

The full model with an R-squared value of 94.4% is good but is probably due to overfitting of the dataset. With a condition number of 35.7, independent variables are actually dependent in the full model which causes problems in fitting the model and interpreting the results. In the future, a multicollinearity test such as a VIF test should be used on the dataset before fitting a model to get rid of dependent variables. There are also some extreme outliers in the dataset as indicated by a kurtosis of 263.7. A more extensive measure should be taken to remove all possible outliers. This would in turn make the residuals non-normal. In the future, more tests for outliers should be taken and tests for multicollinearity should also be taken to avoid having dependent variables. A possible reason for having dependent variables is the airport distances that were added. A possible solution for this could be to separate the trips to/from airports and normal trips. This would mean that the trips to/from airports would be a set amount as seen in the scatter plots before and the model could predict just the normal trips.

Resources

TLC Trip Record Data - Yellow Taxi Trip Records (December 2015) [ONLINE]
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

TLC Trip Record Data - Yellow Trips Data Dictionary [ONLINE]
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page