

Springboard Data Science Career Track

Capstone Project

NYC Taxi Trip Durations

Created By: Andrew Milne

<b>Problem Definition</b>	<b>3</b>
<b>Client:</b>	<b>3</b>
<b>Data Cleaning/Wrangling:</b>	<b>3</b>
<b>Other Datasets:</b>	<b>3</b>
<b>Summary of Key Findings:</b>	<b>4</b>
<b>Exploratory Data Analysis &amp; Inferential Statistics Discussion:</b>	<b>4</b>
Fig. 1 - Trip Data Distributions	4
Key Finding: Statistical Distribution	4
Fig. 2 - Average Trip Time by Vendor	5
Key Finding: Hypothesis Test	5
Fig. 3 - Trips & Duration by Hour	6
Fig. 4 - Trips & Duration by Month	6
Fig. 5 - Trips & Duration by Weekday	6
Fig. 6- Trip Duration Boxplots: Hour & Weekday	7
Fig. 7 - Trip Duration Boxplots: Month	7
Key Finding: Data Trends	7
Fig. 8 - Daily Trips Profile	8
Key Finding: Weather Event	8
Fig. 9 - Trip Speed	9
Fig. 10 - Passengers	9
Key Finding: Passenger Count	9
<b>Machine Learning Regressions</b>	<b>10</b>
Clustering with K-Means Outcomes:	10
Fig. 11 - Neighbourhoods of NYC	10
Fig. 12 - Optimum Cluster Number	11
Key Finding: Cluster Optimization	11
Fig. 13 - NYC Boroughs	11
Key Finding: Cluster Number	11
Regression Models:	12
Table 1 - Regression Performance Metrics	12
Key Finding: Random Forest	12
Fig. 14 - Random Forest Residuals	13
Fig. 15 - Relative Importance of Features in Determining Ride Time	13
<b>Conclusions &amp; Recommendations:</b>	<b>13</b>
EDA & Inferential Statistics	13
Machine Learning Regressions	14

# NYC Taxi Data - Capstone 1 Final Report

## Problem Definition

The kaggle competition behind this is to build a model that predicts the total ride duration of taxi trips in New York City. The primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.

Source:

<https://www.kaggle.com/c/nyc-taxi-trip-duration>

## Client:

The literal client is Kaggle but the figurative client would be a taxi company looking to optimize their route selection. This kind of service provider could be an Uber/Lyft or traditional taxi company...perhaps even a technologically inclined rickshaw driver.

## Data Cleaning/Wrangling:

The data needed to be cleaned to remove the following:

- Extremely long ride times, eliminated data outside +/- 3 std of mean or < 5 hrs duration
- Limited passenger count to <=6 as most cabs don't have capacity of a mini-van
- Direct route information, filled blanks with average durations and steps

Feature engineering was performed to incorporate the following

- Conversion of timestamp data into day, week, month, hour, year, date
- Checking of trip duration & date time values for consistency
- Haversine distances between pickup and drop off latitude & longitude
- Directional bearings calculated based upon latitude & longitude of pickups
- Average speed of trips
- Time series plots

I used a variety of tool bags and libraries to assist in this analysis. These included the following:

- Python 3.6.1 run in Jupyter Notebook
- Plotting capability provided by matplotlib, seaborn
- Analysis capability from pandas, numpy, datetime
- Machine Learning provided by SciKit Learn: MiniBatchKMeans, Linear Regression, Decision Tree Regressor, Random Forest Regressor

## Other Datasets:

- Direct route information incorporated from an Open Source Routing Machine such as one found at [project-osrm.org](http://project-osrm.org)
- Weather related data could have been incorporated to review outliers for snow storms or flooding (*not included*)

## Summary of Key Findings:

Through exploratory analysis and investigation we have determined the following:

- The average trip durations between each vendor are similar but statistically different
- The average trip lasts 840 seconds or ~14 minutes.
- The variables are generally independent of each other, meaning there are few pairs that demonstrate strong correlation to each other, save for trip duration and distance.
- Time of day, week, month can impact trip durations
- The lowest amount of traffic in the city occurs in the early morning hours and late evening.
- Machine Learning algorithms for Linear Regression and Clustering were applied to reveal important features and neighbourhoods
- There is a strong response in trip duration based upon the following three variables: direct distance, hour of the day, direction and dropoff cluster.

## Exploratory Data Analysis & Inferential Statistics Discussion:

The following represents a few key features about the data.

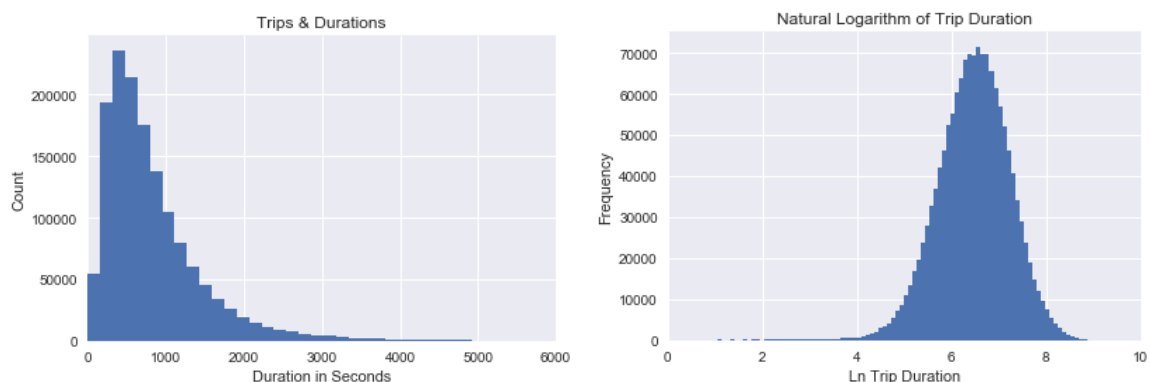


Fig. 1 - Trip Data Distributions

### *Key Finding: Statistical Distribution*

The trip duration is normally distributed around of mean of 840 seconds or 14 minutes.

Split by vendor this makes for an interesting opportunity to test means using *hypothesis test for two samples*

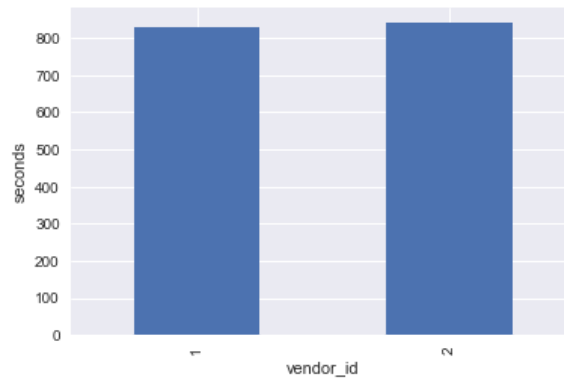


Fig. 2 - Average Trip Time by Vendor

These are huge samples at roughly 600k per vendor. As such we can determine the means and standard deviations to be representative of the populations. The Central Limit Theorem holds that since sample size is  $>40$ .

$$H_0: \bar{x} - \bar{y} = \Delta_0$$

$$H_a: \bar{x} - \bar{y} < \Delta_0$$

This is to say that for  $H_a$ , the mean trip duration of vendor 1 is less than mean trip duration of vendor 2. This inequality indicates a lower tailed test where we reject the null hypothesis if  $z \leq -z_\alpha$ .

For an alpha value of 0.95 the  $z_{critical}$  value is 1.65 (Probability & Statistics for Engineers - Table A.3) meaning that in order to reject the null hypothesis our calculated statistic has to be less than or equal to -1.65.

$$z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)}}$$

Equation of Z Statistic

In this case our z statistic is -10.7 which is less than -1.65 and so we must reject the null hypothesis that the mean trip durations are equal and accept that vendor 1 has a lower mean trip duration than vendor 2.

The P-value is also 0 meaning  $H_0$  should be rejected for any reasonable significance level.

It is also fair to say that although the means are different, they are not different by much and this will likely not be a significant factor in determining trip duration.

#### Key Finding: Hypothesis Test

Vendor will probably not impact ride time significantly.

We investigate other features of the data including pickup times and distances.

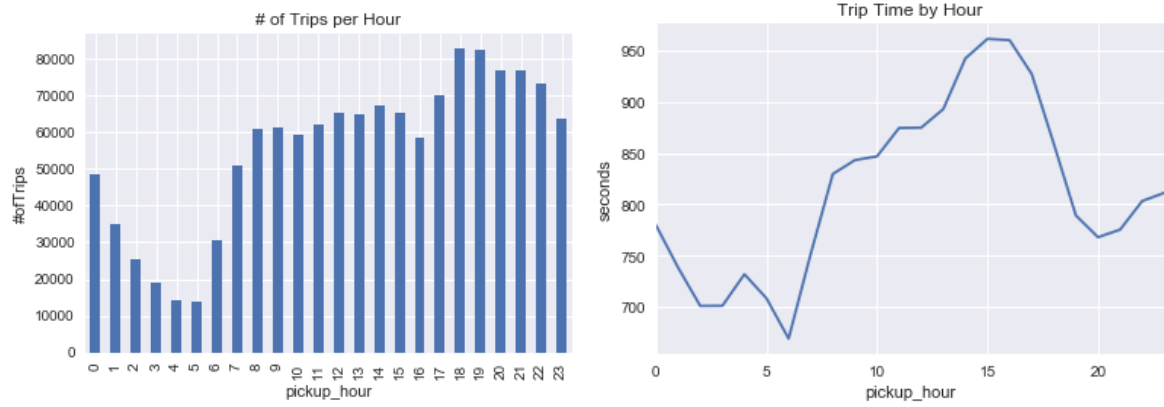


Fig. 3 - Trips & Duration by Hour

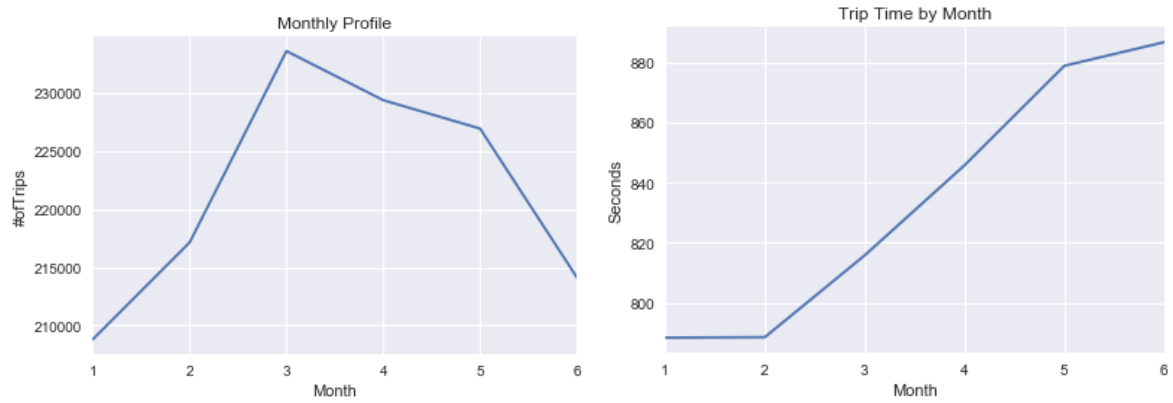


Fig. 4 - Trips & Duration by Month

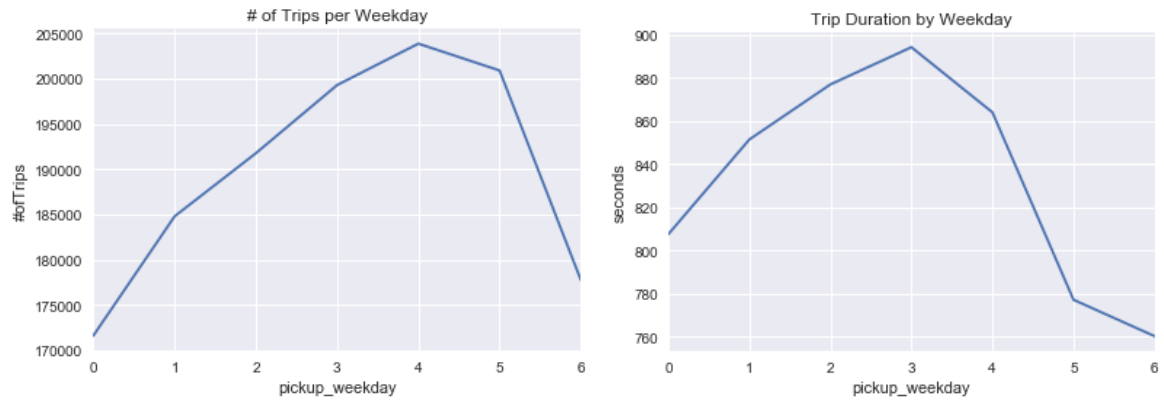


Fig. 5 - Trips & Duration by Weekday

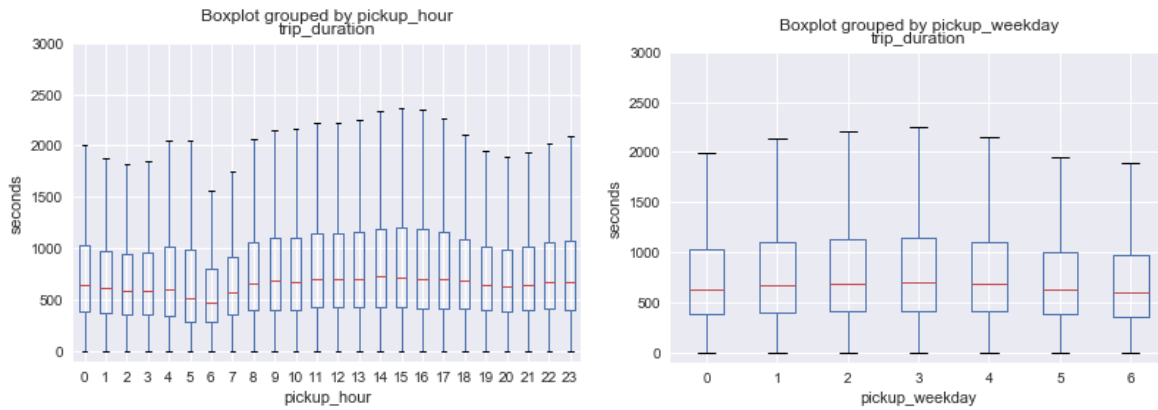


Fig. 6- Trip Duration Boxplots: Hour & Weekday

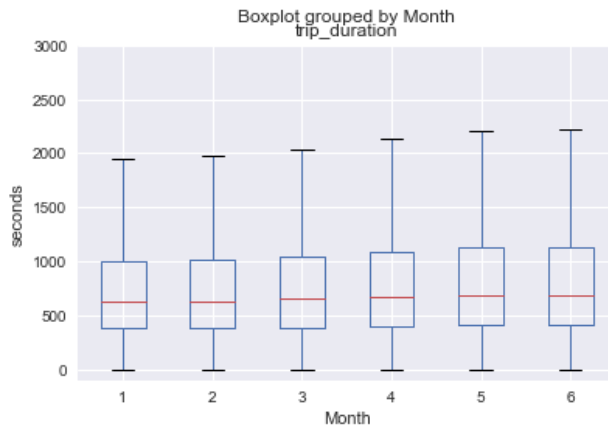


Fig. 7 - Trip Duration Boxplots: Month

In general we see that the profiles of figures 3,4 and 5 are supported by the trends in the boxplots for trip duration.

#### Key Finding: Data Trends

We see an increase in trip duration for the evening hours, the summer months and peak trip durations on Wednesday nights. We will incorporate these in our models.

Let us perform a time series plot on number of trips by date. The figure below shows a relatively consistent traffic pattern except for late January 2016.

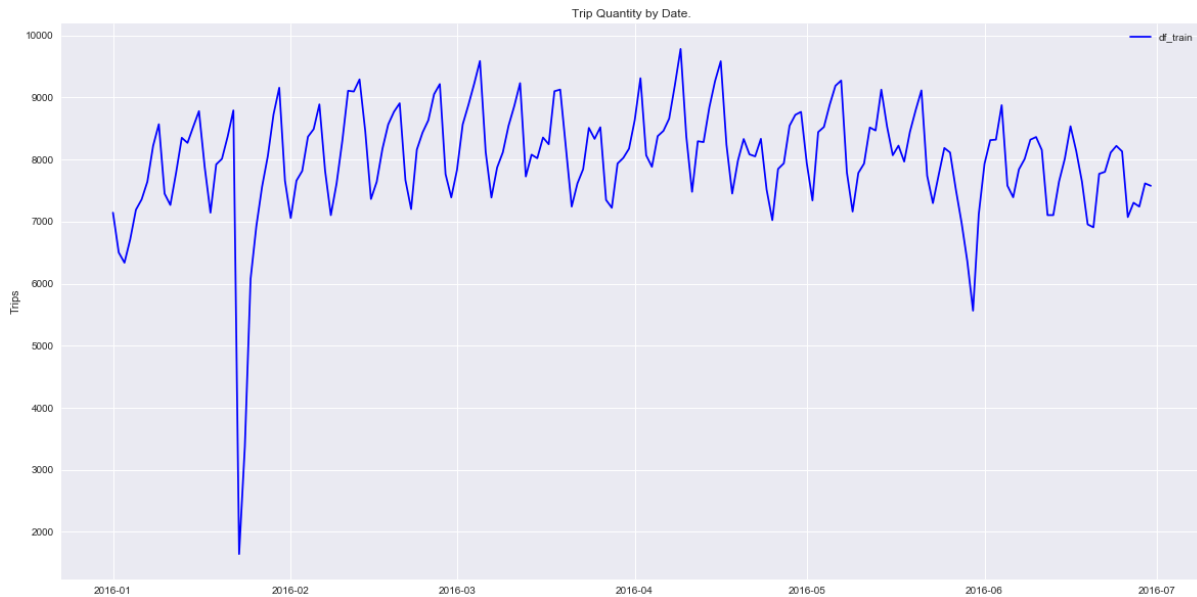


Fig. 8 - Daily Trips Profile

Since this date in late January is such an outlier we ask the question, what could have happened to cause such a drop in the number of taxi rides? My first thought was some form of disaster caused by humans or nature. As it turns out a quick google search for weather events yields:

*Key Finding: Weather Event*

The fact that there was a [crippling blizzard](#) event caused traffic to stop for an extended period.

As this is isolated we have elected to not pursue further investigation.

As a point of interest, the distance feature is 'direct distance' not accounting for one way streets and turns. Data related to shortest route was added as a bit of feature enrichment using the 'Open Source Route Machine'. More on this later.

We obtained the average velocity information by incorporating haversine distance calculations and dividing by units of time.



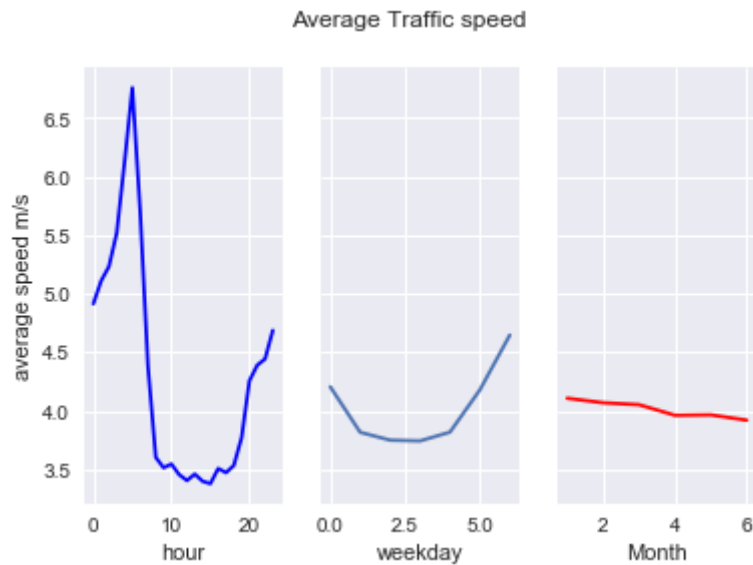


Fig. 9 - Trip Speed

The trip speed data reflects what we saw earlier in the plots of trip duration, Fig 3-5. These will be left out of the model as they are functions of other independent variables.

The number of passengers per trip may prove significant in our modelling. Let us plot it first to see what the distribution is like.

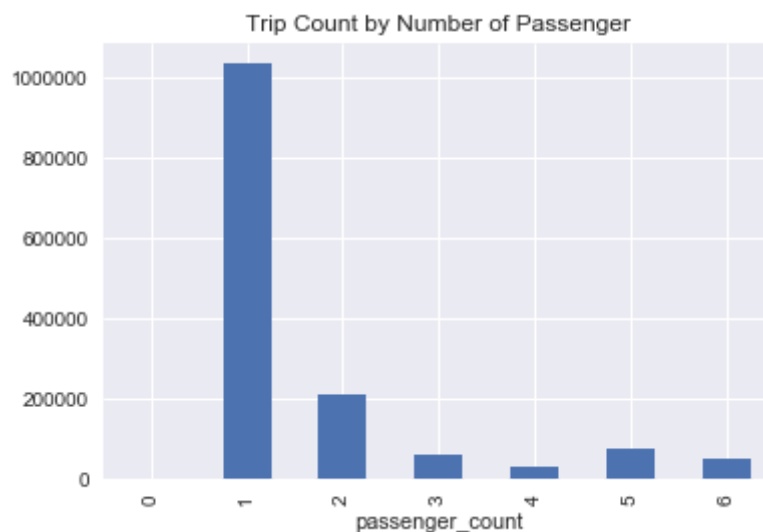


Fig. 10 - Passengers

Figure 10 shows that the majority of trips are for one passenger only. Relatively few trips occur with groups of riders.

#### *Key Finding: Passenger Count*

Logically, it follows that passenger count is not likely to be a big factor in determining ride time.

## Machine Learning Regressions

### Clustering with K-Means Outcomes:

As the data represents the cumulative set of taxi trips within a defined geographic area it makes sense to try and do some clustering algorithms. These clusters can be treated as analogues to the city neighbourhoods.

The first attempt at clustering was to use a number of clusters roughly equal to the number of neighbourhoods within NYC, ~100. In order to perform this analysis it was necessary to first create a numpy array of the available coordinates defined by the pickup and dropoff latitudes and longitudes.

We then created a random sample of the coordinates data and fit a small batch k-means clustering algorithm. The next step was to map the pickup and dropoff clusters by predicting the cluster off the actual latitude and longitude of the pickup and dropoff locations. This meant that the k-means model had been fit to a random sample and then used to predict clusters of unseen data.

An intuitively beautiful figure was produced by the first pass at clustering. While slightly disproportional due to density of the data we can see distinct similarities to a map of the city.

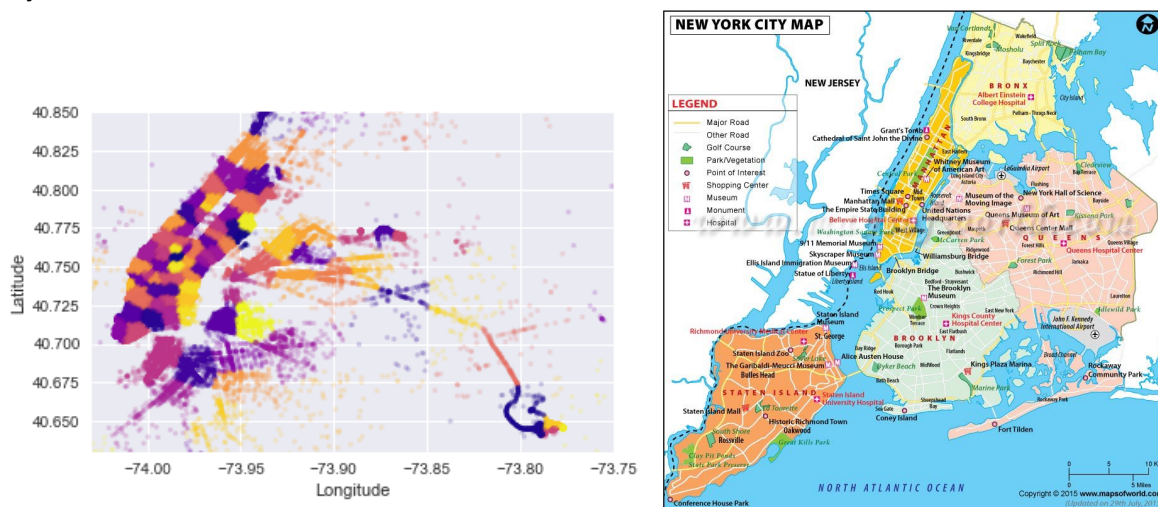


Fig. 11 - Neighbourhoods of NYC

We repeated the process but first employed the 'elbow method' to determine an optimum number of clusters. The point at which the sum of squared error begins to flatten is deemed the optimal value for clusters on the data.

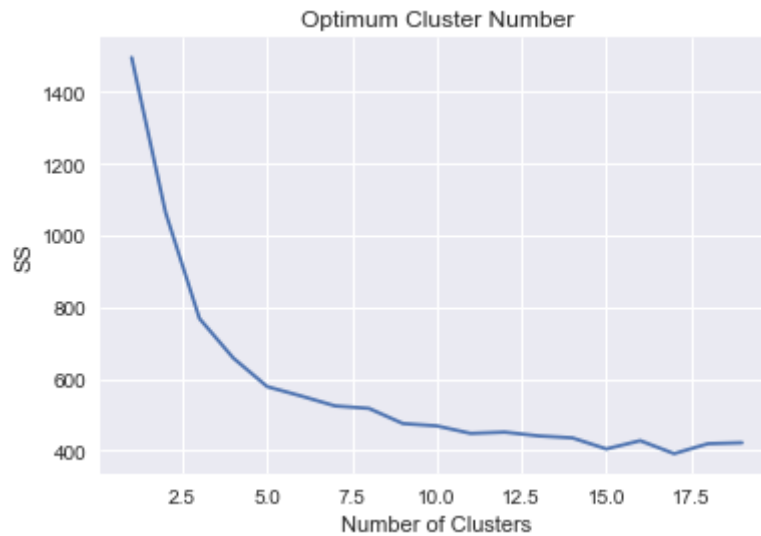


Fig. 12 - Optimum Cluster Number

*Key Finding: Cluster Optimization*

The optimum value for cluster quantity was determined to be 5.

This is a convenient result because there are 5 boroughs in New York City - Queen's, Harlem, The Bronx, Staten Island and Manhattan.

Here are the results of 5 clusters on the same plot of the pickup and dropoff locations.

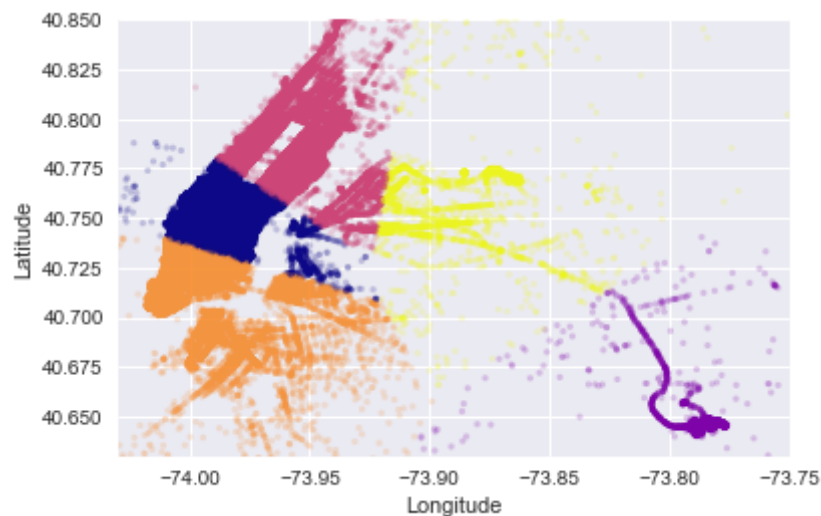


Fig. 13 - NYC Boroughs

Again, we take note of the fact that this silhouette bears resemblance to the city. We can see distinct areas such as central park, the rivers, the airport and the five boroughs.

*Key Finding: Cluster Number*

We will use the cluster number as a feature in our models.

## Regression Models:

We can then move on to regression models using SciKit Learn. The following parameters were included in the model based on the key findings of our EDA:

- Month (1-6)
- Pickup hour of the day (0 to 23),
- Pickup weekday (0-6) with Monday as zero,
- Direct Distance in kilometres,
- Total Distance based upon OSRM in metres,
- Number of steps in ride, i.e. turns, based upon OSRM data,
- Pickup cluster & Dropoff cluster (0 to 4),
- Passenger count (0 to 6),
- Vendor,
- Log Trip Duration

The 'log trip duration' variable was used as the dependent outcome as we are trying to predict trip durations. The natural logarithm was used as it was required for the submission to Kaggle and also for ease of calculations.

The regressions were carried out with and without dummy variables for categorical features. It proved to be true that the models performed better without dummies. This is most likely caused by the categorical features having numerical values to begin with.

We performed three iterations of regression. The performance results are as follows:

Table 1 - Regression Performance Metrics

	Linear Regression	Decision Tree	Random Forest
R-Squared	0.47	0.72	0.73
MAE	0.42	0.29	0.28
MSE	0.31	0.17	0.16
RMSE	0.56	0.41	0.39

### *Key Finding: Random Forest*

Random Forest is our best model

It is demonstratively clear that the Random Forest regression model performed best when tested on unseen data. The plot of the residuals for the Random Forest model is shown below and profiles a normal distribution, indicating good fit.

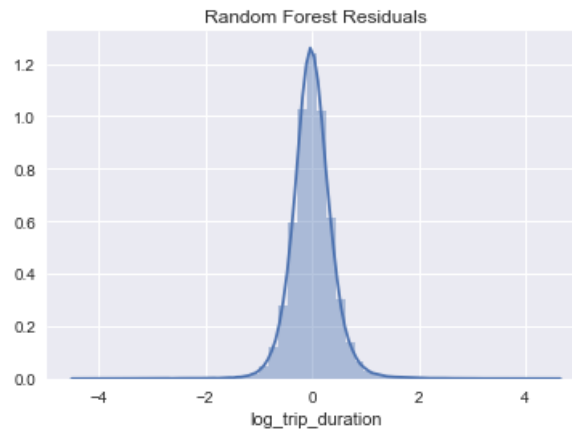


Fig. 14 - Random Forest Residuals

The maximum depth of the decision tree and random forest regressors was set to ten based upon n-1 number of features. This was done to limit the calculations and also avoid over fit of the data.

Lastly, the relative importance of features in the model are shown below:

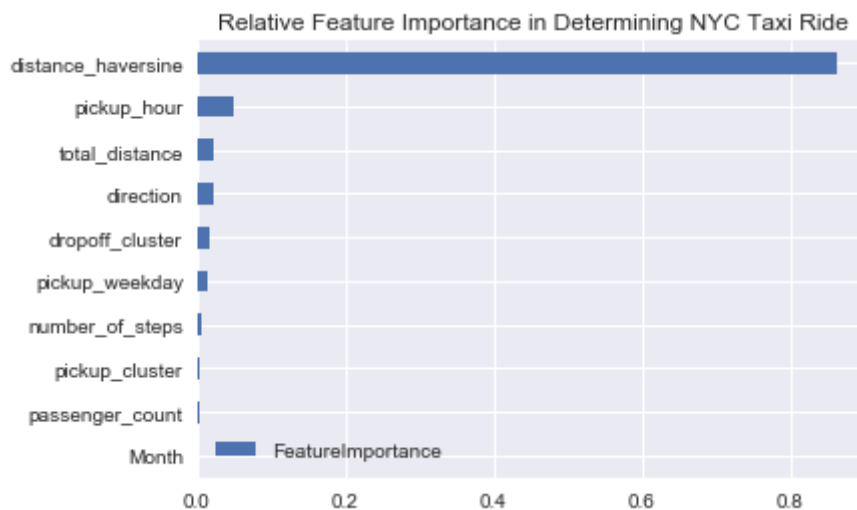


Fig. 15 - Relative Importance of Features in Determining Ride Time

The most prominent factor in determining ride time is the direct geographical distance. This is followed by the time of day, total distance and general direction of route.

## Conclusions & Recommendations:

### EDA & Inferential Statistics

The data for trip durations was normally distributed before and after transforming it to the natural logarithm.

What is likely to play the biggest role in determining length of the time spent riding in a cab is the total distance travelled. Beyond that obvious conclusion we can expect to take longer rides in the summer months, when it is later in the day and if it is during rush hour. As the week goes on the trips generally decrease in duration.

After incorporating the distance calculations and the OSRM we were able to add a bit more context to the data.

Further exploration could be performed on the directional bearing of the trips, weather related data, annual city events, durations by landmark etc.

Statistically we could do more investigation. For example, ANOVA tests could be performed on features to determine correlation with an outcome.

### **Machine Learning Regressions**

The regression models proved to be fairly successful after using the Random Forest model. Further analysis would probably take the shape of an XGBoost or perhaps some Principal Component Analysis.

While drastically improved, the Random Forest model is not yet suitably accurate. Ideally the R-Squared value would be higher than 0.8. Shortcomings of the model are likely centred around the ability of SciKitLearn and my machine to handle the dummy variables. If the dummy variables were included this would help differentiate a ride on a Friday night from one on Monday when trips are much slower.