NYC Taxi Data - Inferential Statistics Report - Capstone 1

Background:

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.

The data needed to be cleaned to remove the following:

- Extremely long ride times - > 100min (some as long as 1+ days)
- Trips with >5 passengers was deemed impossible and removed
- Trips with zero distance but more than 5 minutes of duration

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
- Bearings calculated based upon latitude & longitude of pickups

Note that the competition data contains only the first 6 months of the year.

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

Summary:

Through my exploratory analysis and statistical investigation I have determined the following:

- The average trip durations between each vendor are similar but statistically different
- The average trip lasts 834 seconds or ~15 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.
- There is a strong response in trip duration based upon the following three variables: distance, month of the year and passenger count.

Discussion:
The following represents a few key features about the data.

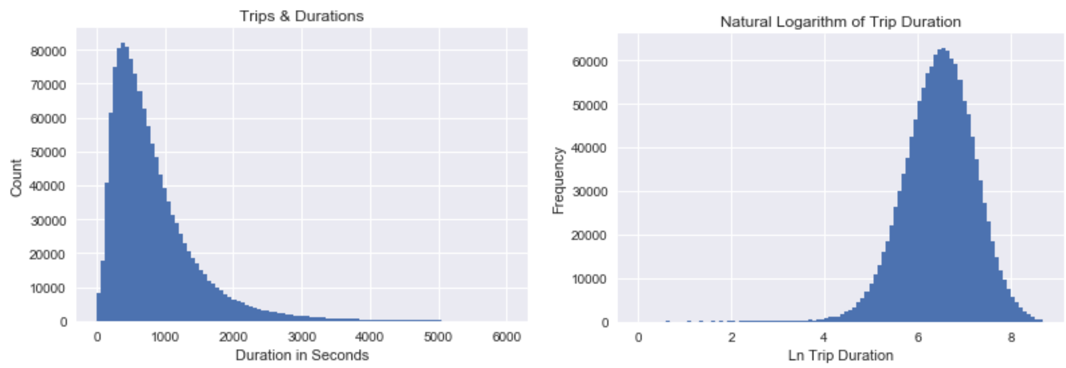


Fig. 1 - Trip Data Distributions

We can see that trip duration is normally distributed around of mean of 840 seconds of 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.

Ho: mu1 - mu2 = deltao
Ha: mu1 - mu2 < deltao

This is to say that for Ha, 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<= -zalpha

For an alpha value of 0.95 the Zcritical value is 1.65 meaning that in order to reject the null hypothesis our calculated statistic has to be z<= -1.65.

In this case our z statistic is -10.2 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 Ho 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.

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

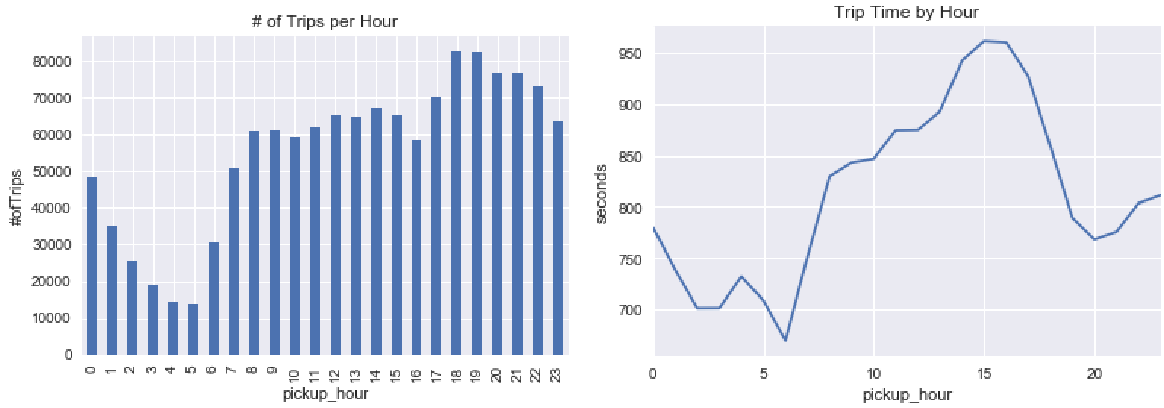


Fig. 3 - Trips & Duration by Hour

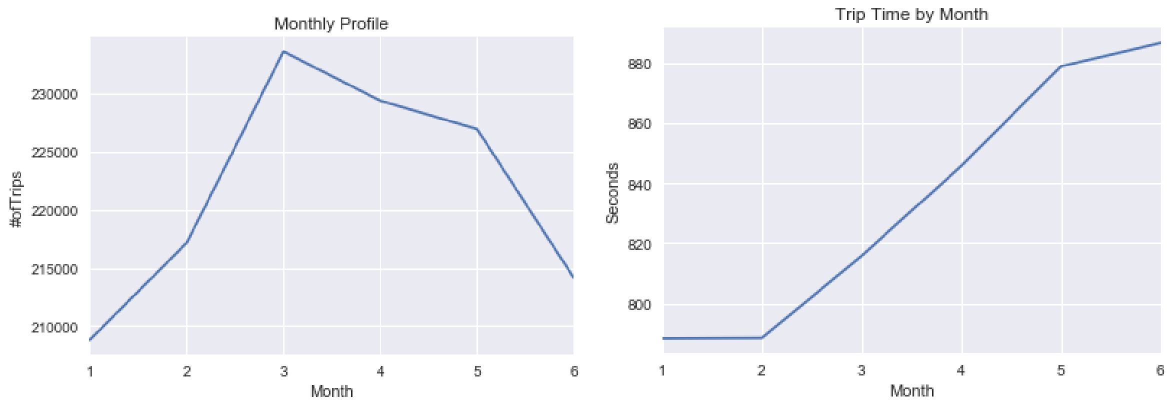


Fig. 4 - Trips & Duration by Month

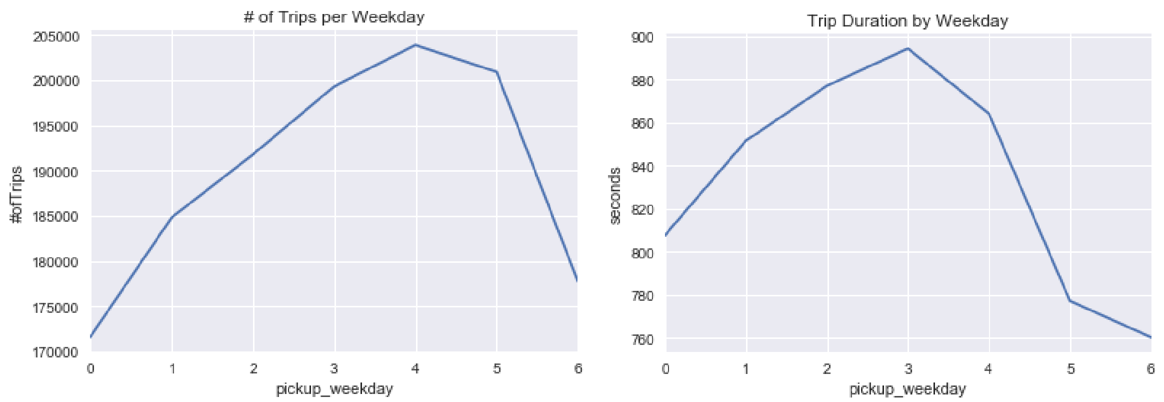


Fig. 5 - Trips & Duration by Weekday

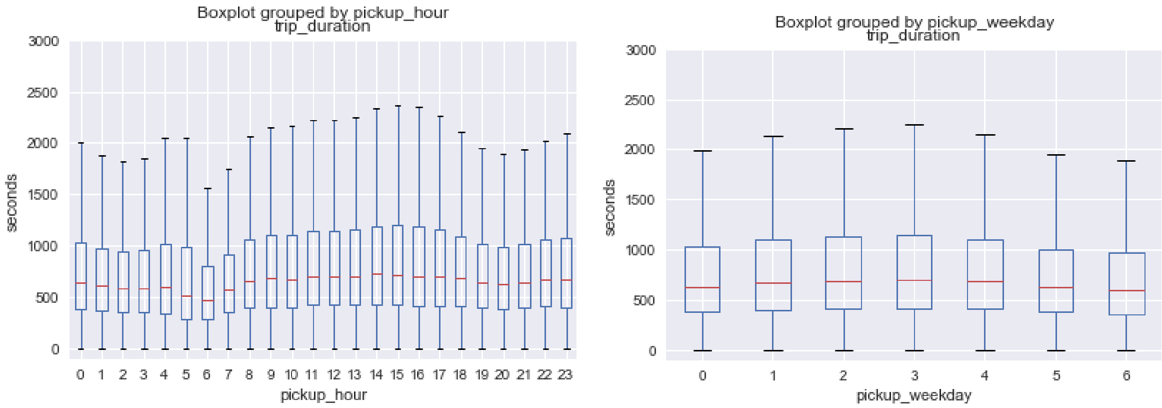


Fig. 6- Trip Duration Boxplots: Hour & Weekday

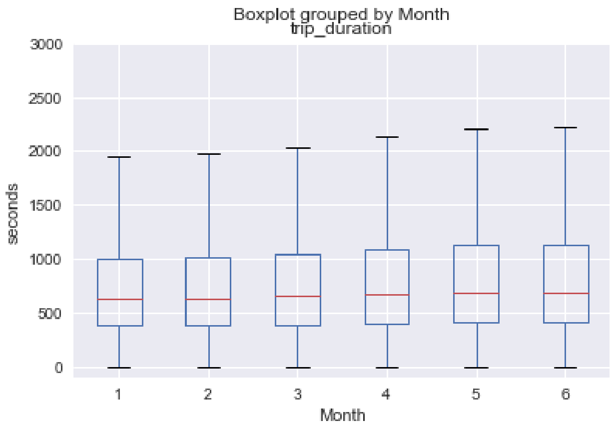


Fig. 7 - Trip Duration Boxplots: Month

In general we see that the profiles of figures 3,4 and 5 are supported by the trends in the boxplots for trip duration. We see an increase in trip duration for the evening hours, the summer months and peak trip durations on Wednesday nights.

As a point of interest, the distance feature is 'direct distance' not accounting for one way streets and turns. We obtained the average velocity information.

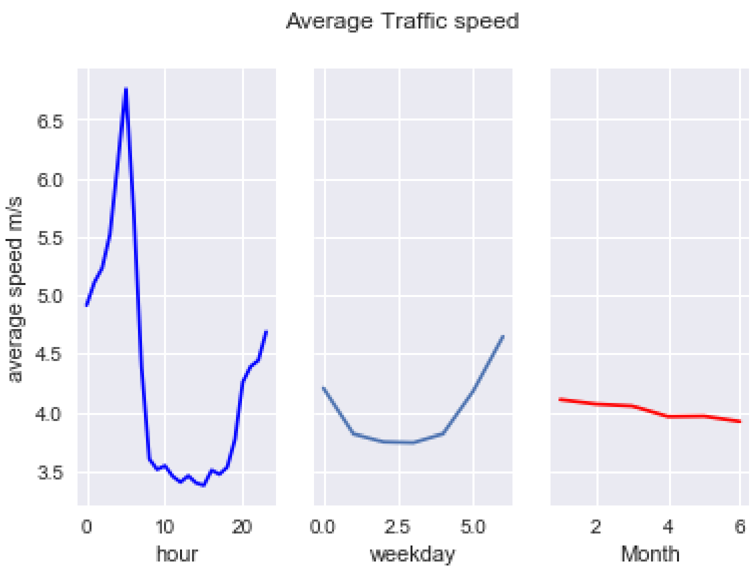
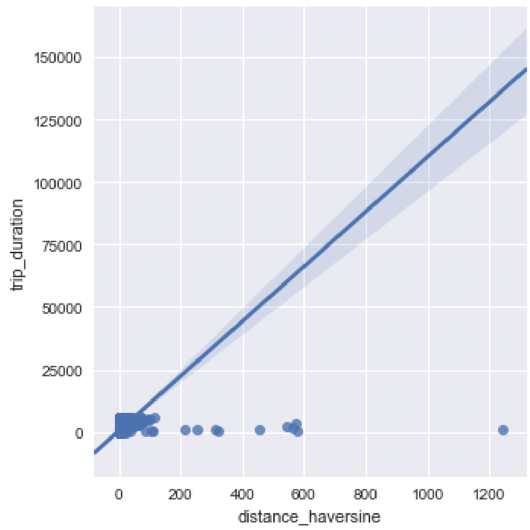


Fig. 8 - Trip Speed

We can then move on to a linear regression model using SciKit Learn. The following parameters were included in the model:

- Month
- Pickup_hour,
- Pickup_weekday,
- Direct Distance,
- Vendor,
- passenger_count

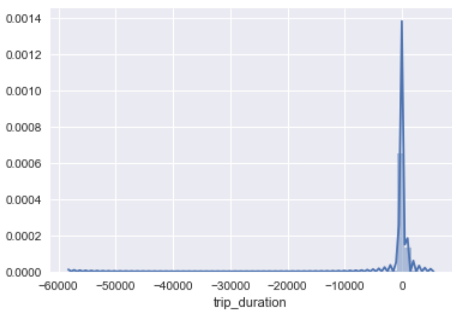
Logically, distance is likely to be the largest contributing factor to trip duration. The following plot demonstrates a strong positive correlation.



A model was created using SciKit Learn's 'TrainTestSplit'. The model produced the following errors and R-squared values:

MAE: 297.663310569
MSE: 185540.117878
RMSE: 430.743680021
R Squared: 0.56

While not impressive, the R-squared value does indicate a degree of fit.
Residual Plot shows Normal Distribution, a positive sign.



The coefficients of the model confirmed suspicions regarding direct distance.

	Coefficient
Distance	105
Month	18
Passenger Count	15
Pickup Hour	4
Vendor	3
Pickup Weekday	-12

Conclusion:

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 we have more passengers, when it is later in the day and if we travel with Vendor 2. As the week goes on the trips generally decrease in duration.

In further analyses I would choose to incorporate more features such as weather and the routes through the city. This would provide greater insight on external factors and also actual distance travelled.

Statistically we can do more investigation also. For example, ANOVA tests could be performed on features to determine correlation with an outcome. We could also perform a clustering algorithm and include the cluster numbers as features for our regression model.