NYC Taxi & Limo

Exploratory Analysis & Modelling

Created by Andrew Milne

A typical NYC taxi travels 70,000 miles a year

20% of all trips are <1 mile

600,000 passengers a day

There are 50,000 taxi drivers in NYC

Can we model how long it takes?

Data Wrangling

Extremely long trips > limited to +/- 3 std dev or 5hrs

Way too many passengers > limited to minivan capacity

Trips with no duration > dropped from data

Feature Building

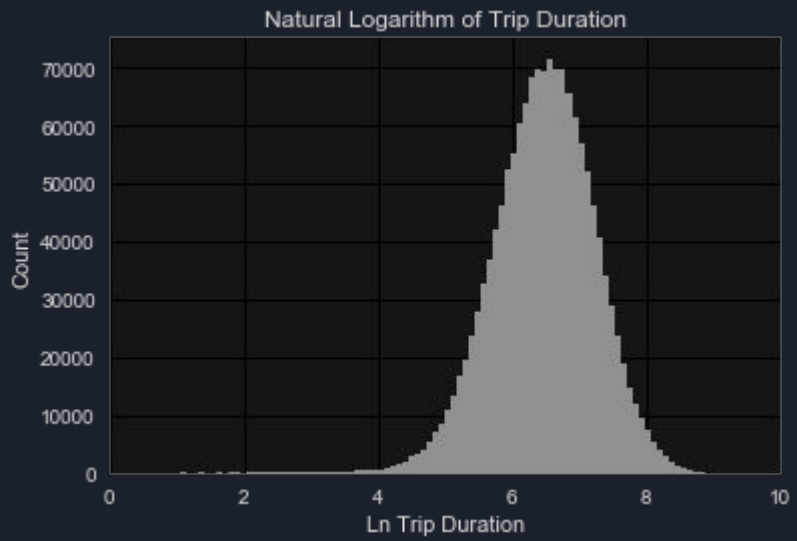
Distance > Geographic distance travelled & Bearing

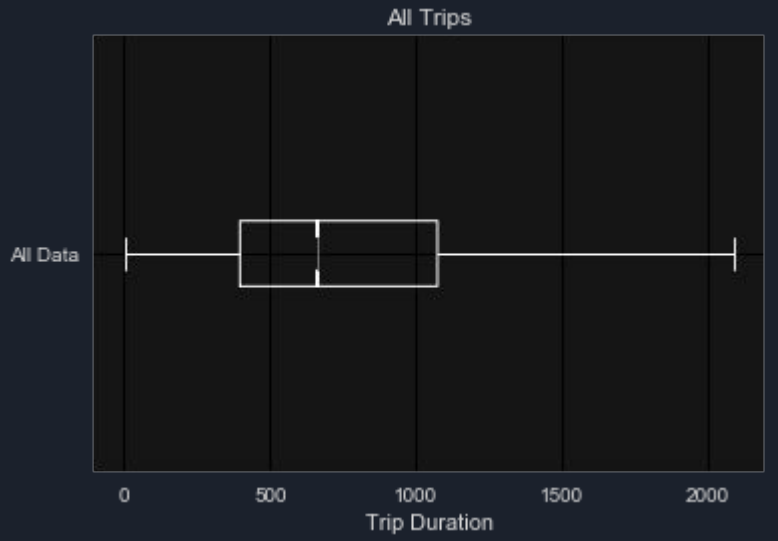
Time > Time Series Data conversions

Speed > Average velocity

Routes > Open Source Route Machine Data

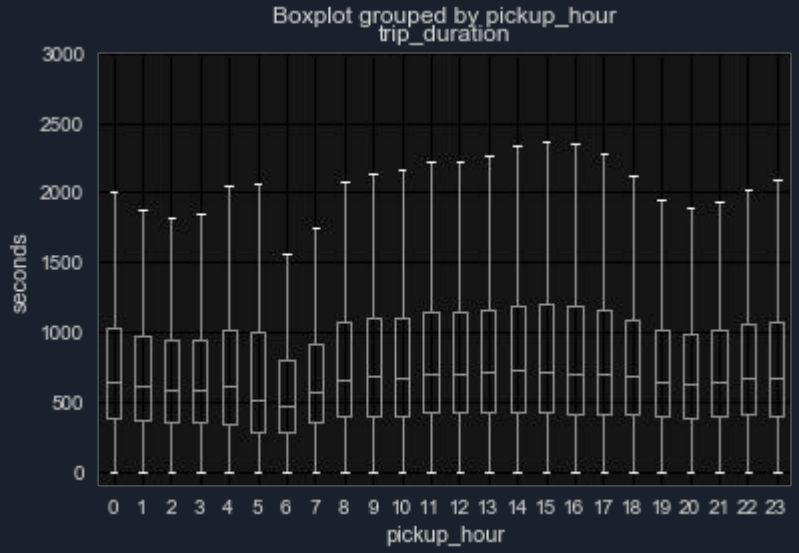
Are we there yet? - An Exploratory Analysis

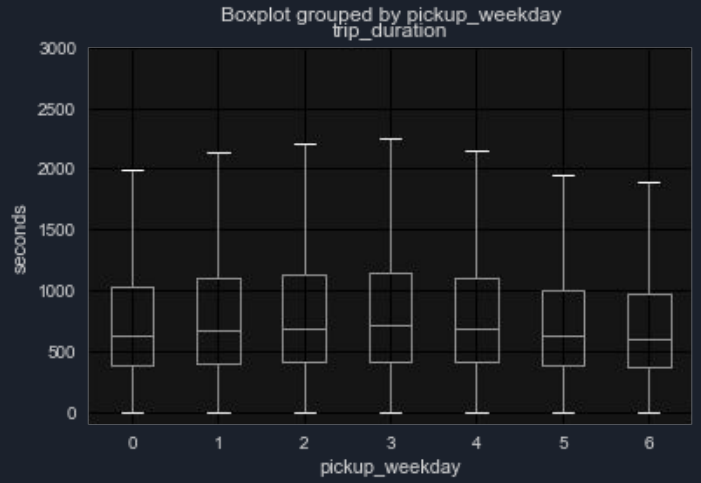


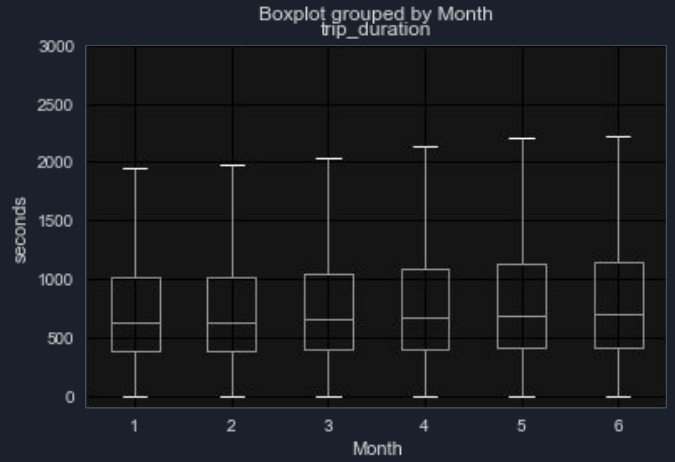


Normally distributed data with a mean of 14 minutes and median of 11 minutes

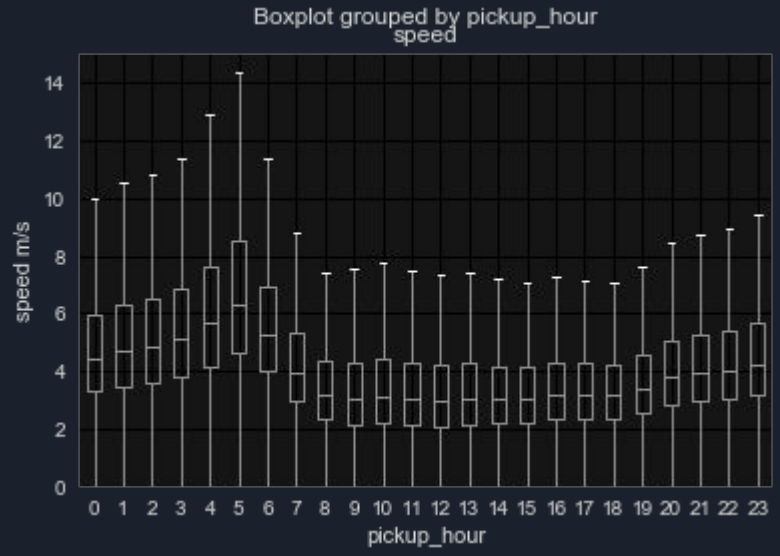
EDA - Duration

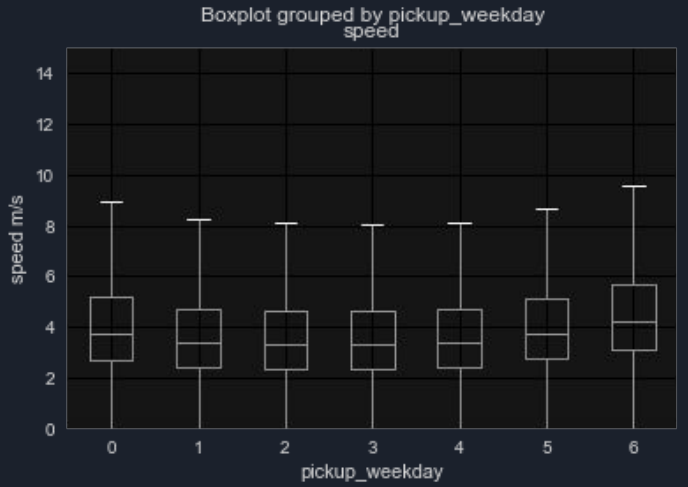


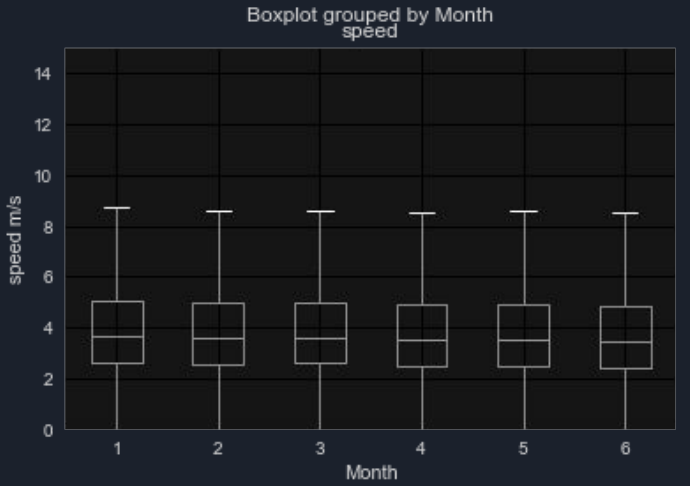




EDA - Speed


Boxplot grouped by pickup_hour
speed
speed m/s
0 2 4 6 8 10 12 14
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
pickup_hour


Boxplot grouped by pickup_weekday
speed
speed m/s
0 2 4 6 8 10 12 14
0 1 2 3 4 5 6
pickup_weekday


Boxplot grouped by Month
speed
speed m/s
0 2 4 6 8 10 12 14
1 2 3 4 5 6
Month

Is one taxi faster than another?

Yes, but it’s not worth waiting...

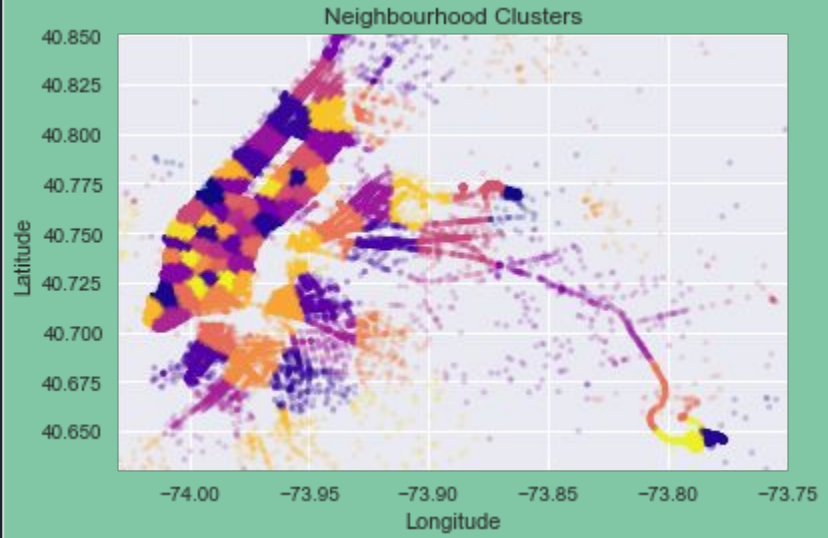
Performed a hypothesis test for two samples

Determined:

the means are statistically different

Outlier! - Late January snowstorm 2016

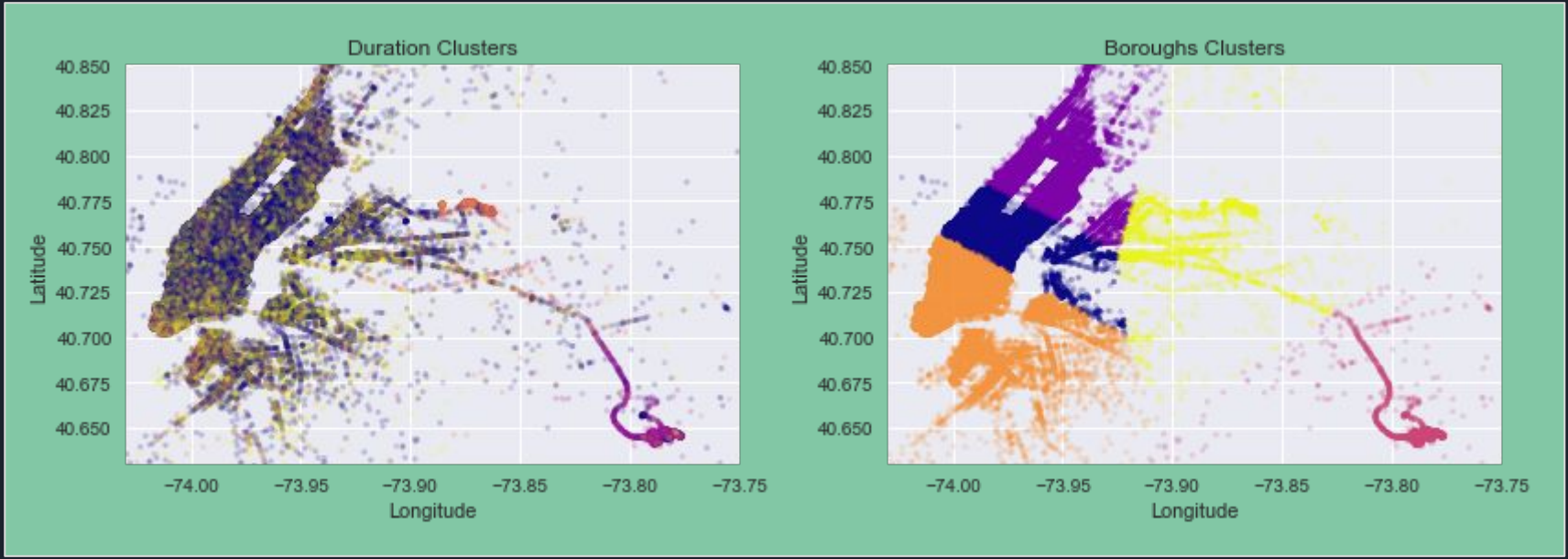
Modelling - Clusters



100 neighbourhoods...and the airport

A map of town...how cool is that?!

Modelling - Clusters

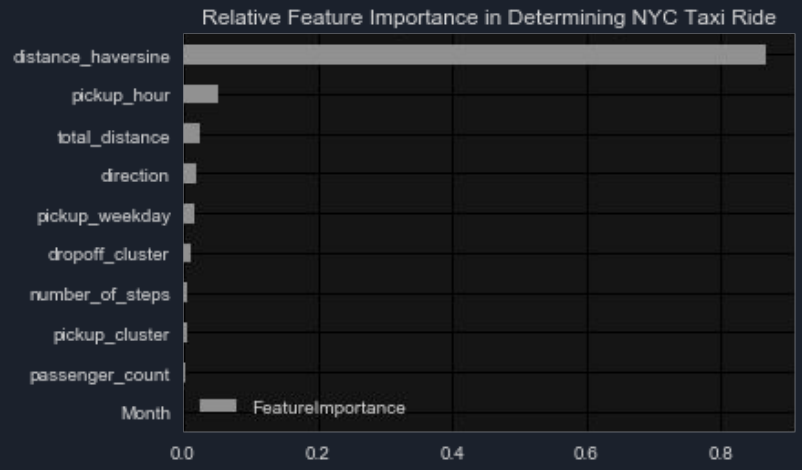


By duration - faster areas exist

The Five Boroughs

Let's Make a Model

	Linear Regression	Decision Tree	Random Forest
R-Squared	0.53	0.73	0.74
MAE	0.39	0.29	0.28
MSE	0.28	0.16	0.16
RMSE	0.53	0.40	0.39



3 iterations performed and tweaked Random Forest with GridSearchCV

Finally, let’s test the model

The model predicted a mean trip duration of 792 seconds or 13 minutes or 1 standard deviation.

The test data had never been evaluated before.

Further tweaks might include the use of a Randomized Search, XG Boost and other features.

Backup

https://github.com/andrewcmilne/capstone1_taxi