Capstone Project

NYC Taxi Trip Time Prediction

Sandesh Gaikwad

Content

- Problem Statement
- Exploratory Data Analysis
- Feature Engineering
- Modeling and comparing results
- Conclusions

Problem Statement

Problem : Build a model that predicts the total ride duration of taxi trips in New York City.

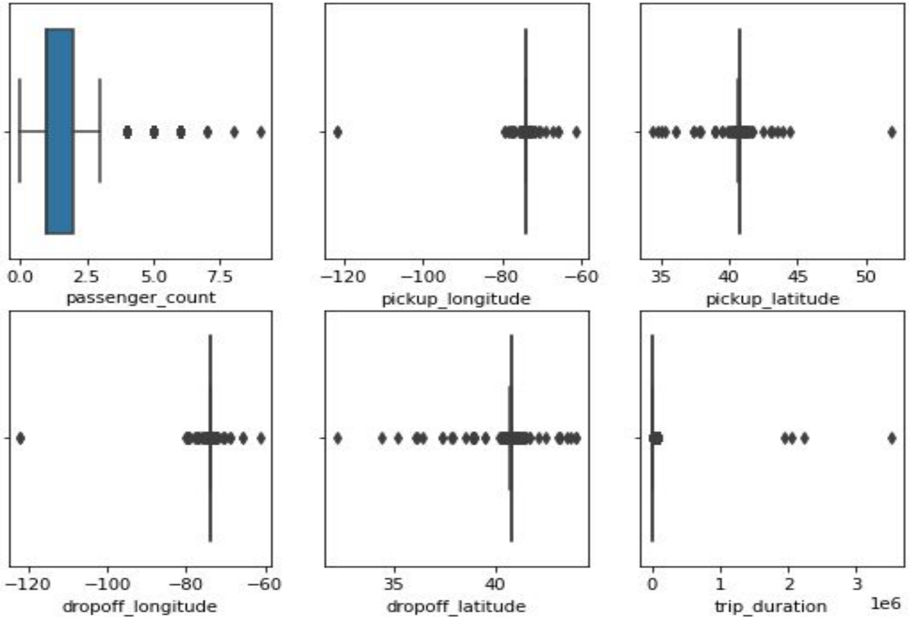
Data Description - The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this project. Based on individual trip attributes, we need predict the duration of each trip in the test set.

Data Summary

- Id : A unique identifier for each trip
- Passenger count : Number of passengers in the vehicle (driver entered value)
- Location : Four columns with longitude and latitude of pickup and drop Locations
- Time: Two columns with pickup and dropoff date time
- Trip duration : duration of the trip in seconds

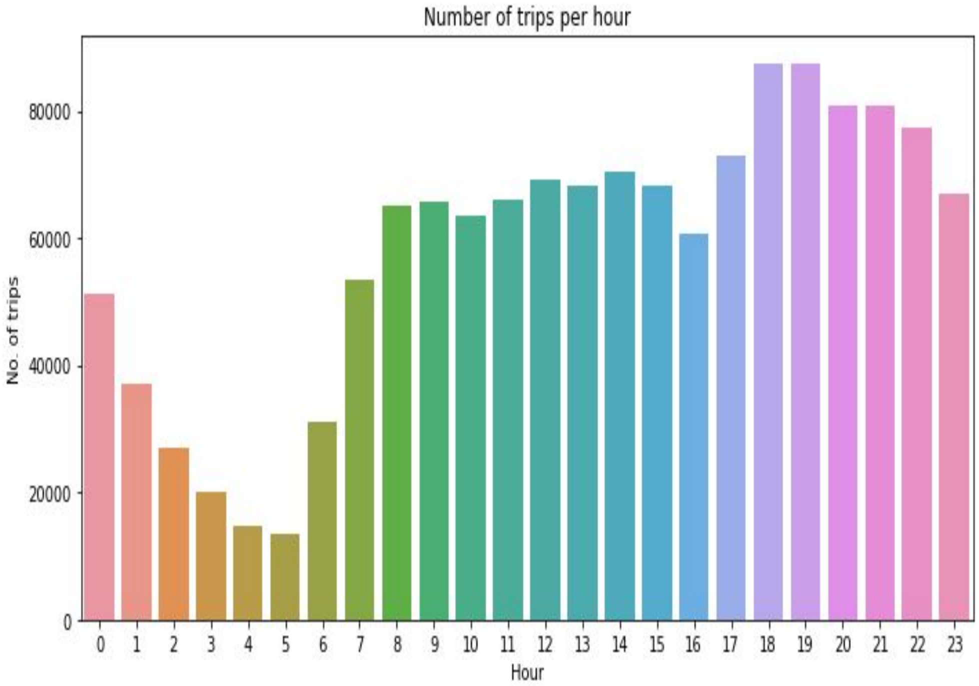
Outliers detection (Box Plots)

- Every Parameter has extreme values
- We can remove most extreme value and later remove remaining anomalies



Avg trip per hour

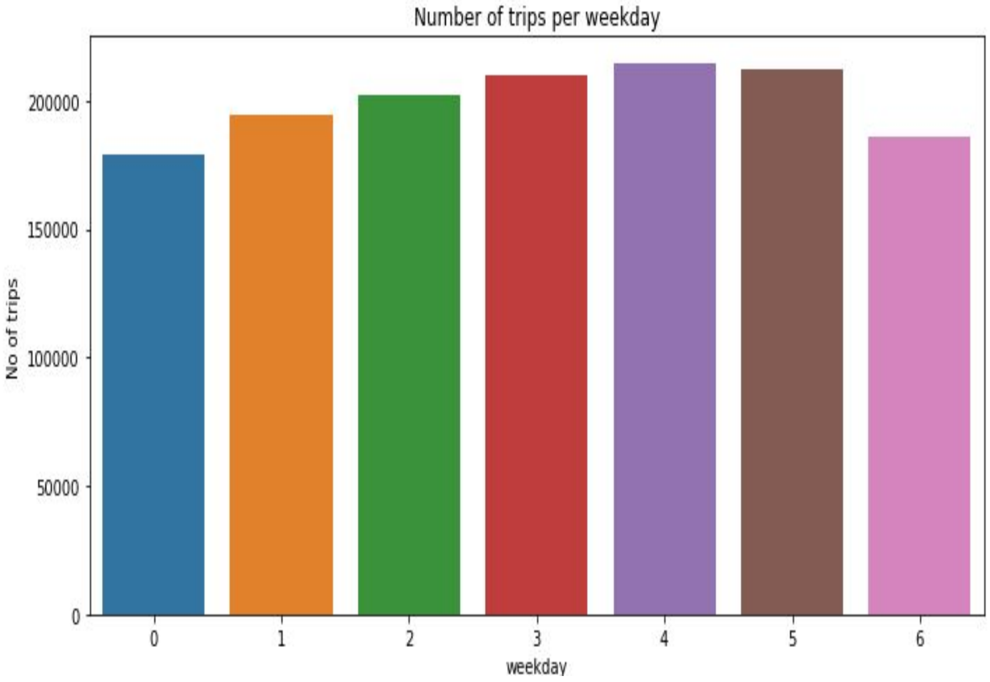
- Number of trips start increasing at 6 early morning
- Then morning to late afternoon trips remain almost same
- Evening has highest number of trip per hour
- Late night trip are least



Avg trip per day

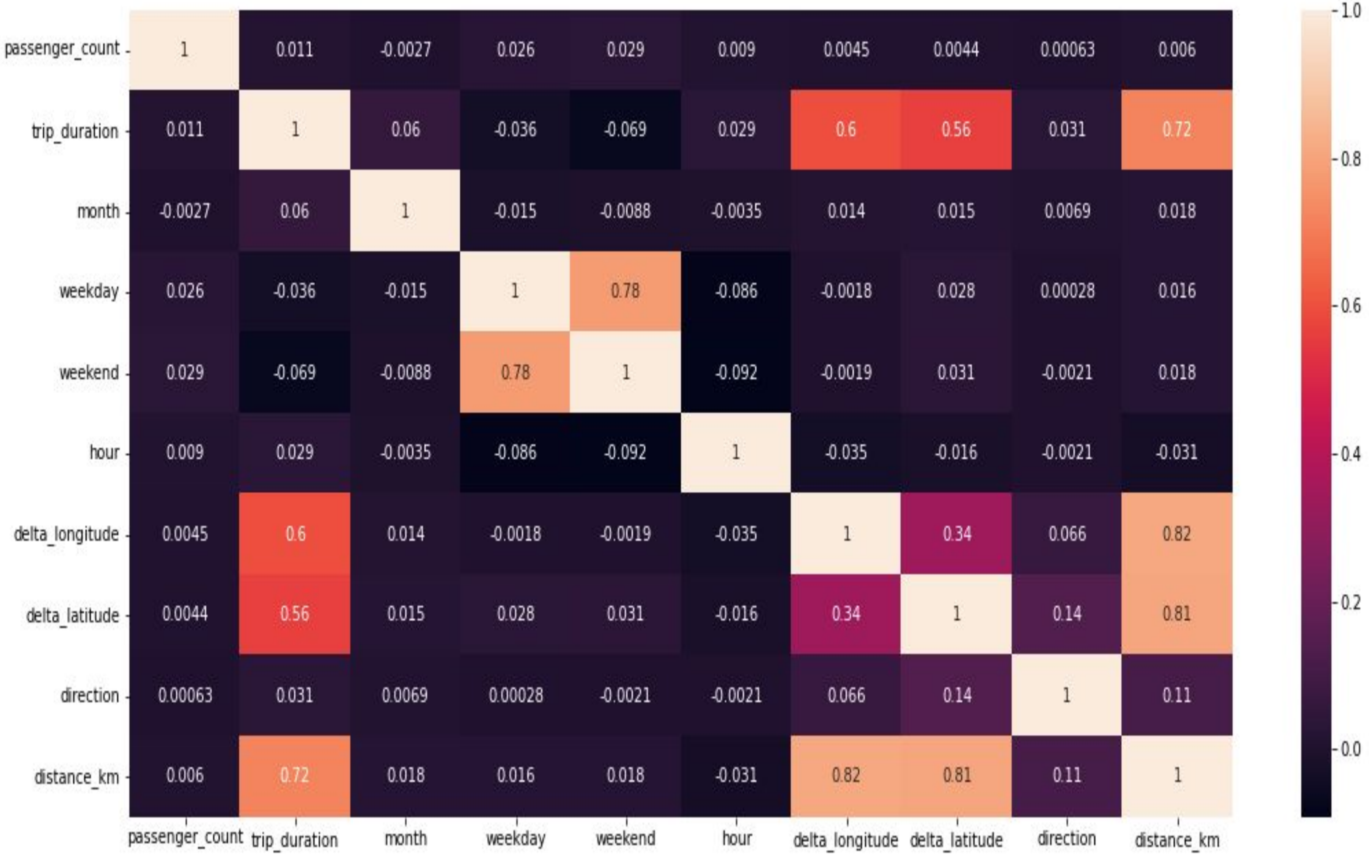
- There is no significant change in trips per day

- Sunday has least tris per day while friday and saturday highest



Correlation matrix

- Distance has the highest correlation with trip duration



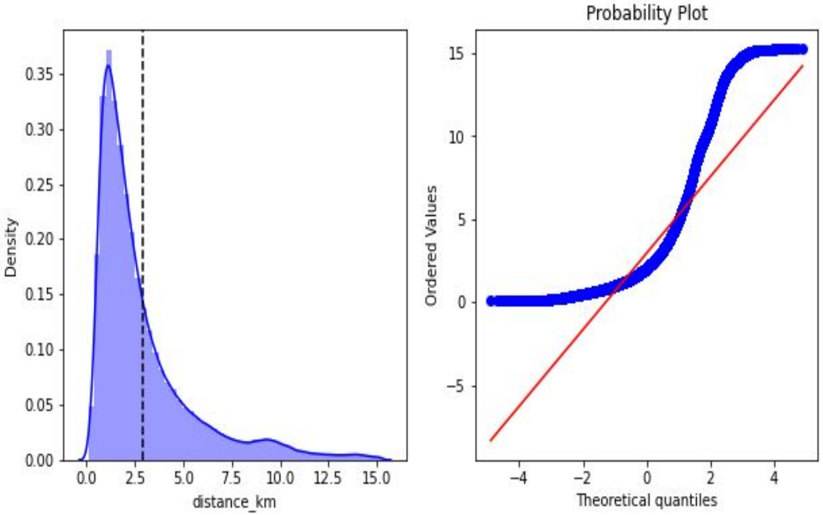
	passenger_count	trip_duration	month	weekday	weekend	hour	delta_longitude	delta_latitude	direction	distance_km
passenger_count	1	0.011	-0.0027	0.026	0.029	0.009	0.0045	0.0044	0.00063	0.006
trip_duration	0.011	1	0.06	-0.036	-0.069	0.029	0.6	0.56	0.031	0.72
month	-0.0027	0.06	1	-0.015	-0.0088	-0.0035	0.014	0.015	0.0069	0.018
weekday	0.026	-0.036	-0.015	1	0.78	-0.086	-0.0018	0.028	0.00028	0.016
weekend	0.029	-0.069	-0.0088	0.78	1	-0.092	-0.0019	0.031	-0.0021	0.018
hour	0.009	0.029	-0.0035	-0.086	-0.092	1	-0.035	-0.016	-0.0021	-0.031
delta_longitude	0.0045	0.6	0.014	-0.0018	-0.0019	-0.035	1	0.34	0.066	0.82
delta_latitude	0.0044	0.56	0.015	0.028	0.031	-0.016	0.34	1	0.14	0.81
direction	0.00063	0.031	0.0069	0.00028	-0.0021	-0.0021	0.066	0.14	1	0.11
distance_km	0.006	0.72	0.018	0.016	0.018	-0.031	0.82	0.81	0.11	1

Feature Engineering

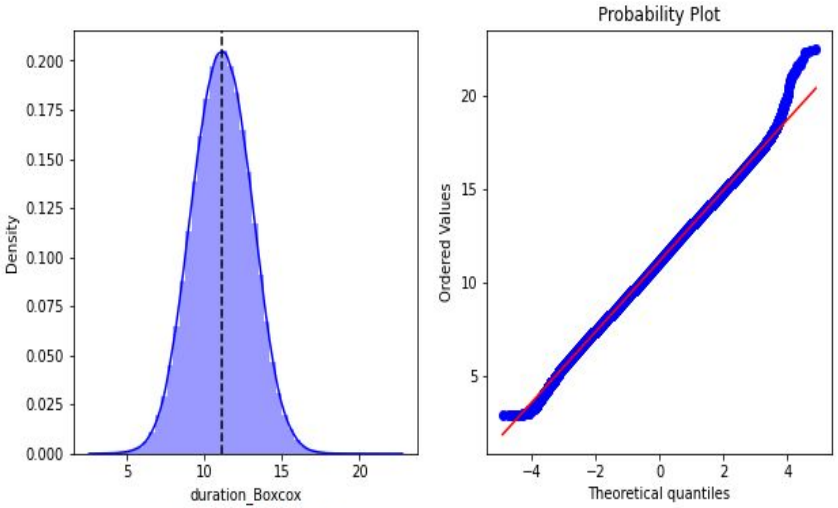
- Features extractions from “Pickup Datetime”
  - Month
  - day
  - is_Weekend
  - Hour
  - Peak Hour

- New Derived features from Latitudes and Longitudes
  - Delta longitude / latitude
  - Direction
  - Distance_km

Gaussian Distribution - distance

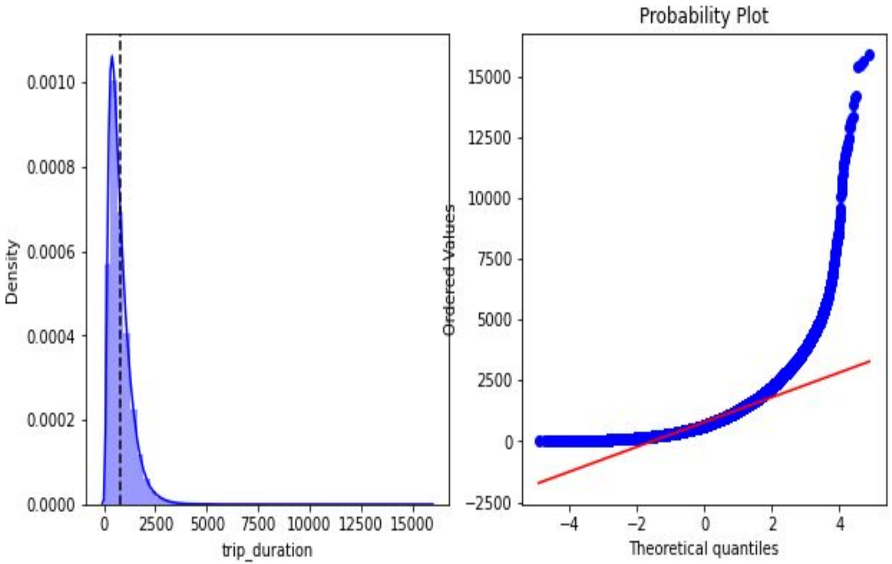


```
skewness  :  1.9517243542550862
```

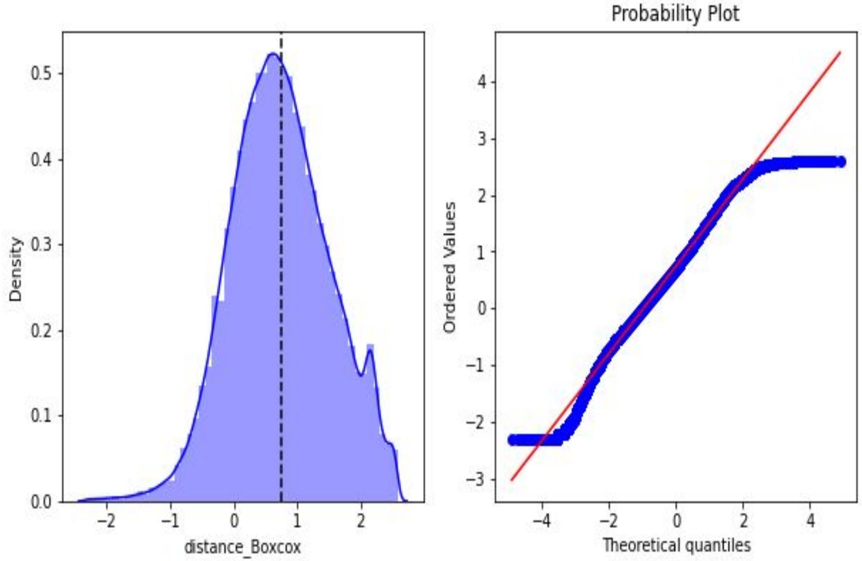


```
skewness  :  -0.0021837506737579933
```

Gaussian Distribution - Trip duration



```
skewness  :  2.08118809062355
```



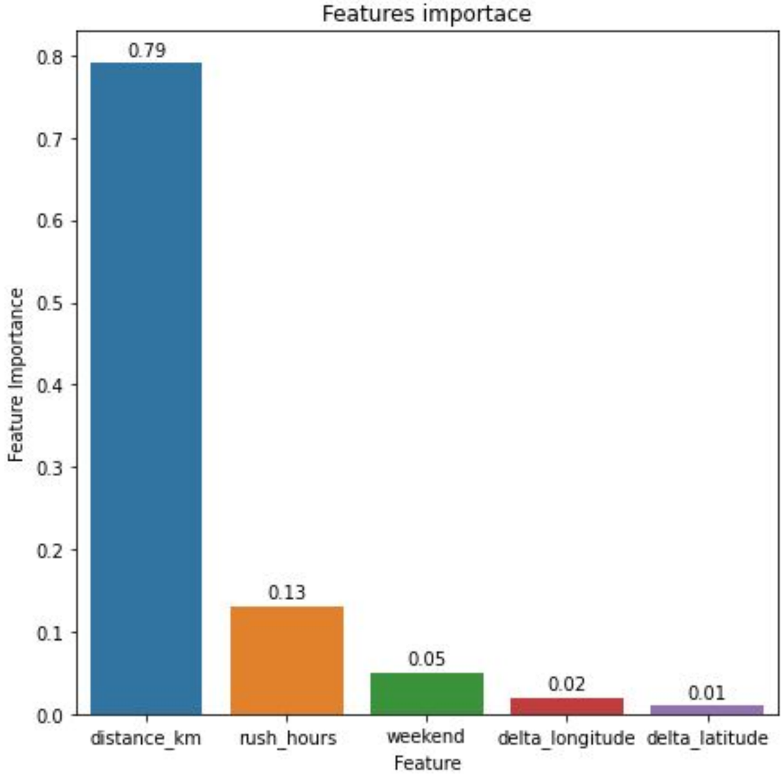
```
skewness  :  0.0012122340625846646
```

Models

Model Name	Evaluation Metrics		
	R-Squared	RMSE	MAE
Linear Regression	0.55	372	241
Lasso Regression	0.55	372	241
Ridge Regression	0.55	372	241
Decision Tree	0.61	346	230
Random Forest	0.6	348	232
Xgboost	0.62	341	224

Parameters of importance

- Distance is most important parameter
- Rush hour is second most important parameter
- Delta latitude and Delta longitude is not that important as distance already captures the information from these parameters



Conclusion

- Linear, Lasso, and Ridge regressions are giving similar results.
- In the Decision Tree regressor, the results are slightly improved.
- Out of all tried models, Random Xgboost Regressor is giving the best result.

Future Challenges

- Distance and time of day are not sufficient to predict the time of ride it requires a thorough understanding of traffic and that the actual path of the ride is important.
- Deep Neural Networks can improve this performance.