New York City Taxi Fare Prediction

Data Science Capstone Project

Mehrnaz Mireslami

Overview

- Problem Statement
- Description of the Dataset
- Data Wrangling
- Driving New Features
- EDA and Storytelling
- Statistical Inference
- Machine Learning

Problem Statement

Goal:

Develop a Machine Learning based model to predict the fare amount for a taxi ride in New York City.

- Enhance customers' satisfaction, since it is given as upfront data to the customers.
- Provide better results for taxi cabs and ridesharing companies such as Uber, Lyft, etc.

Description of the Dataset

Data Source: Kaggle competition

Features:

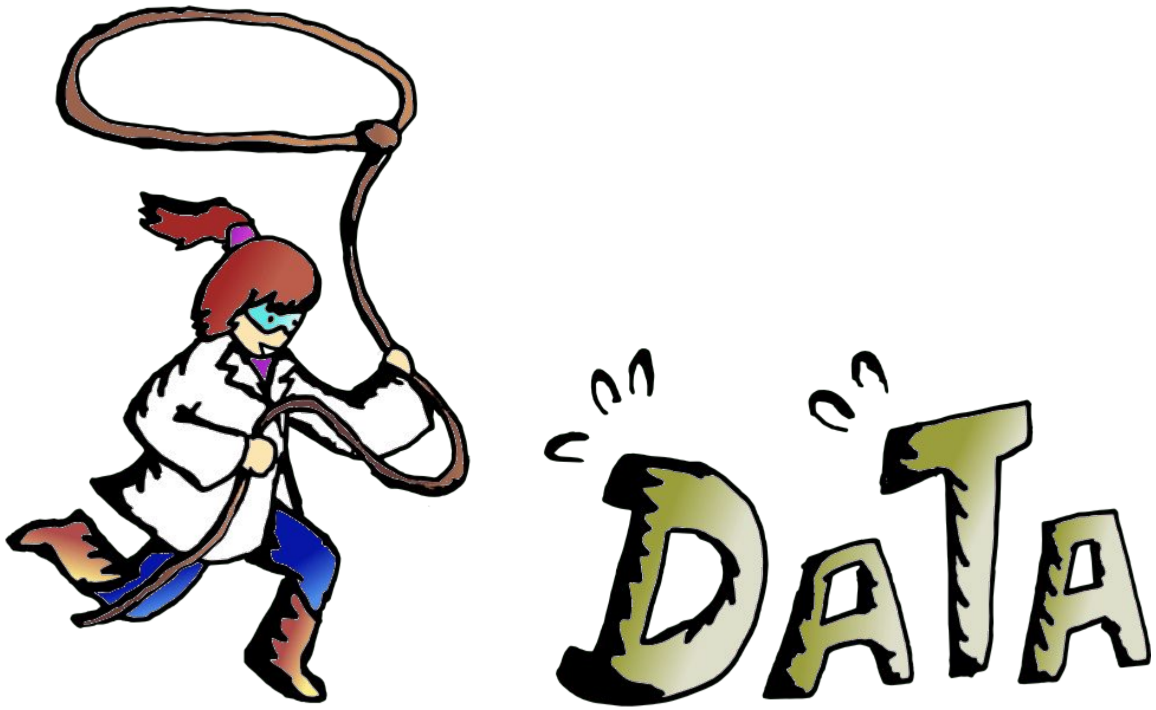
- pickup_datetime
- pickup_longitude
- pickup_latitude
- dropoff_longitude
- dropoff_latitude
- passenger_count

Target:

- fare_amount

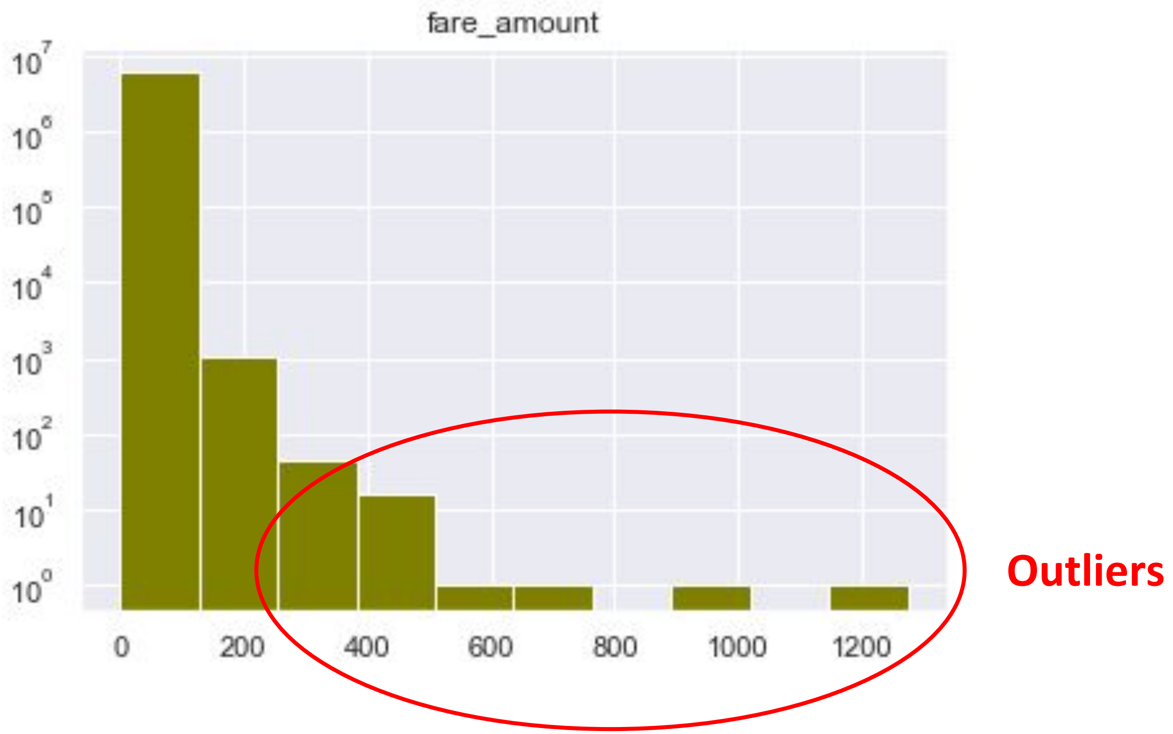
Data Wrangling

- Data was almost clean.
- Remove NaN values.
- Remove features' outliers.
- Extract new features.



Fare Amount

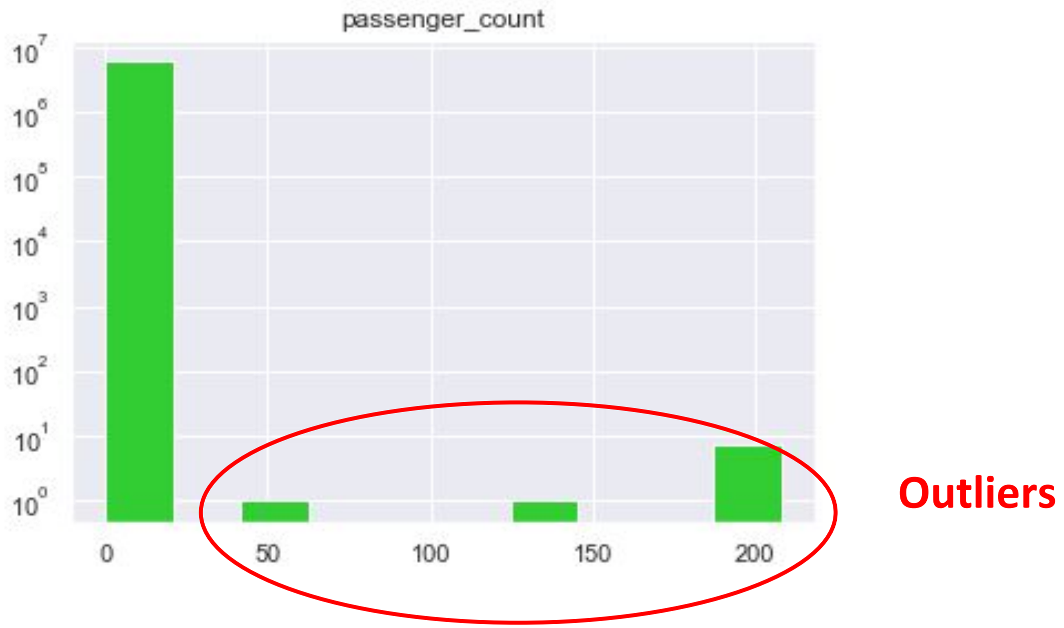
Based on the distribution of fare_amount, entries greater than $200 should be considered as outlier and dropped from the dataset.



Passenger Count

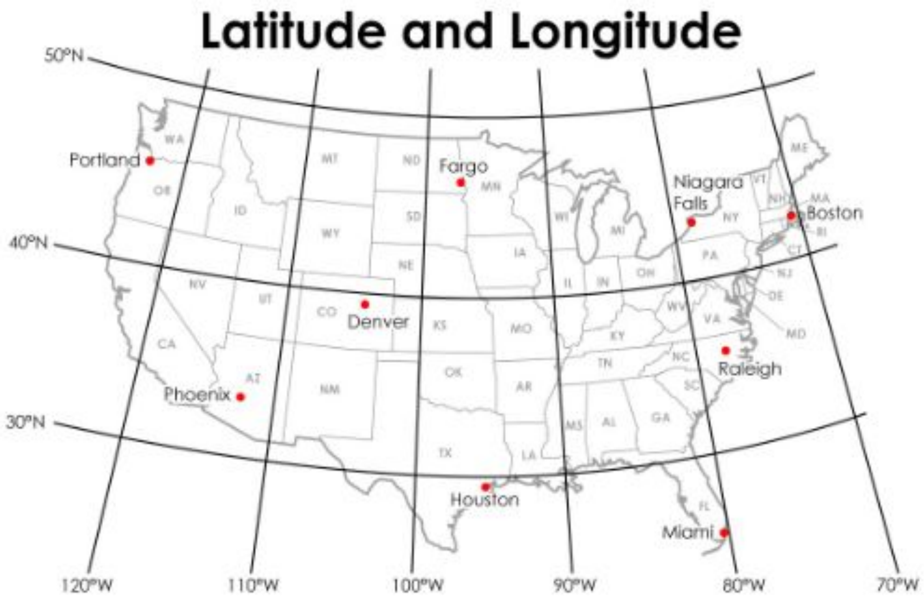
Maximum number of passengers is 208 that is not realistic for the number of seats on a taxi cab.

The outlier for number of passengers should be removed, so the number of passenger for each ride should be between 1 and 6.



Latitude and Longitude Features

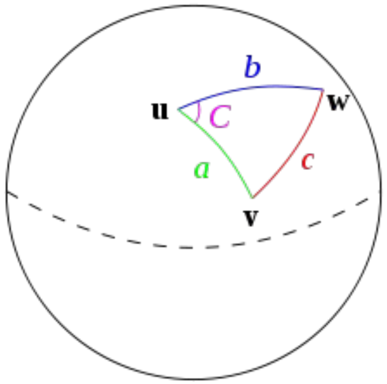
Pickup and dropoff coordinates should be in the latitude and longitude range for NYC.



Deriving New Features

- **Distance Between Pickup and Dropoff Locations**

Haversine formula is employed to calculate the distance between pickup and drop-off locations.



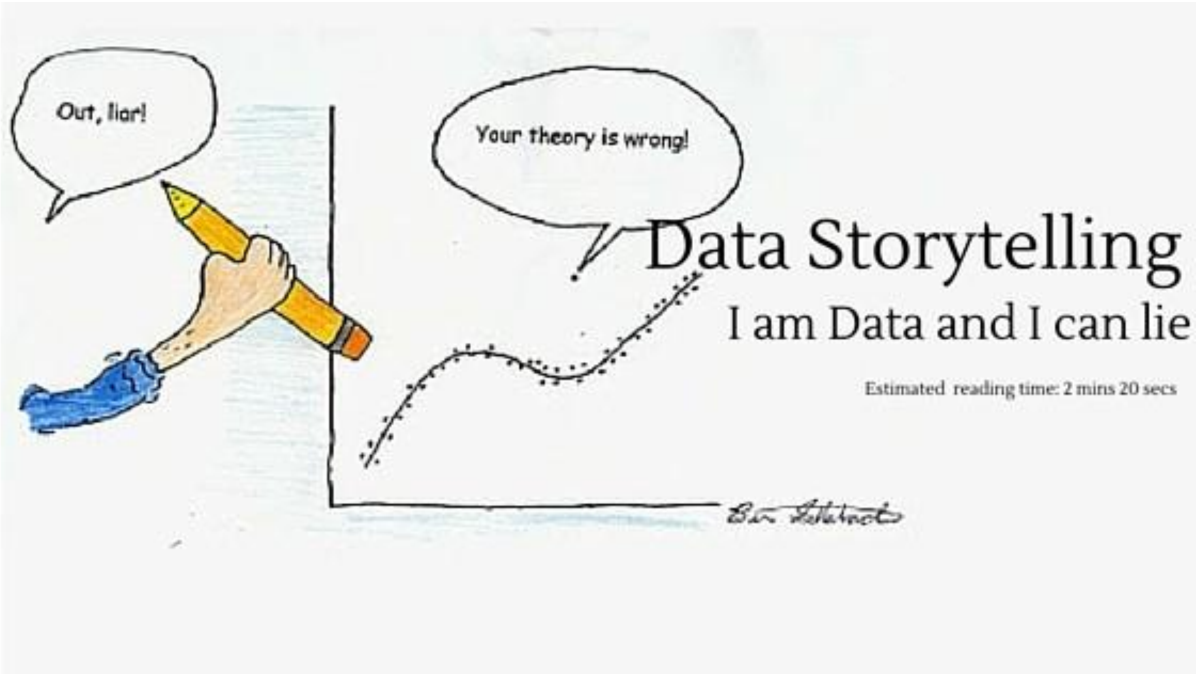
- **Time and Date Features**

The key attribute is leveraged to extract new features:

- Year
- Month
- Day
- Hour
- Day of the Week

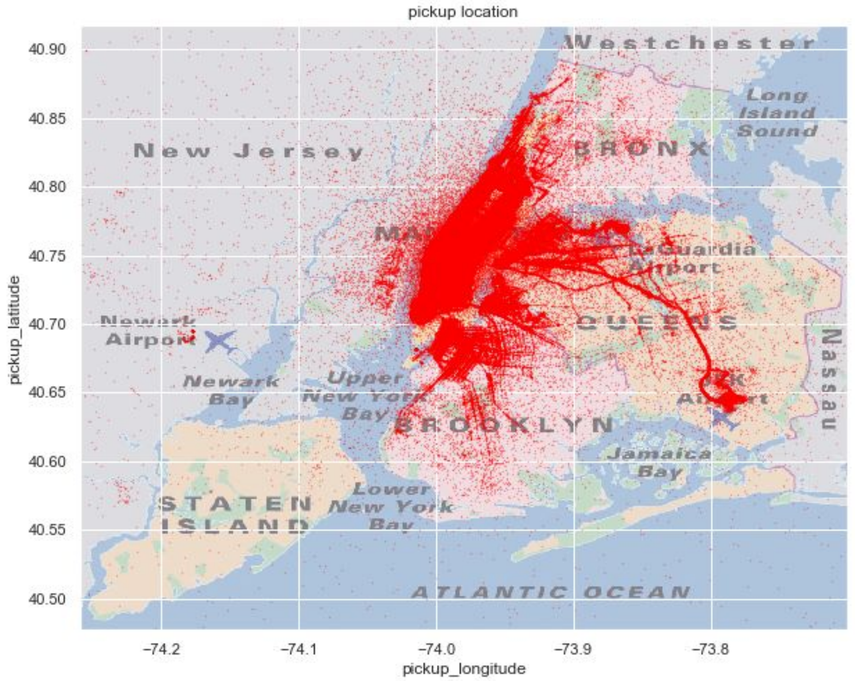
EDA and Data Storytelling

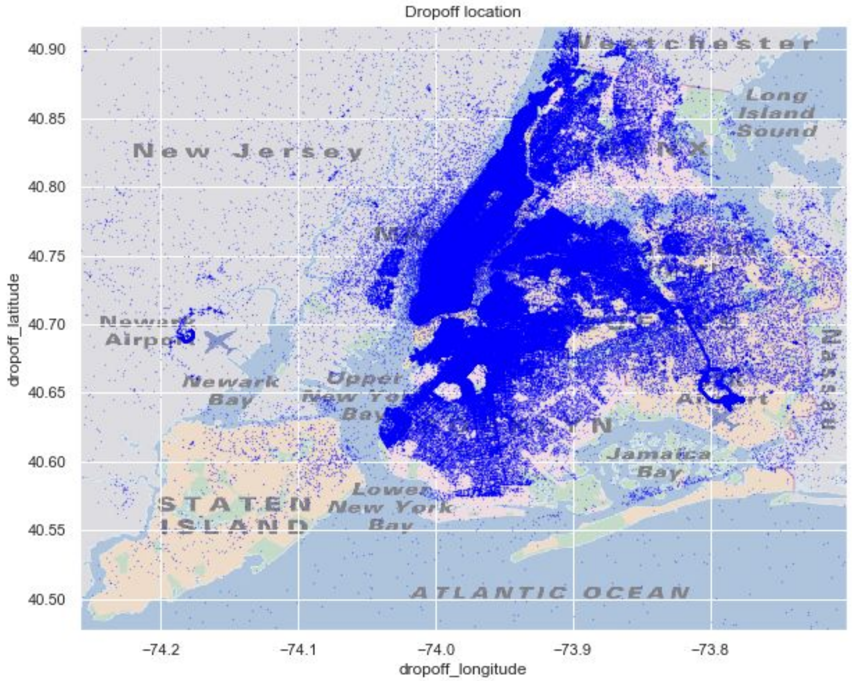
- Visualizing data to find the correlation among features.
- Inferential statistics



Pickup and drop-off visualization

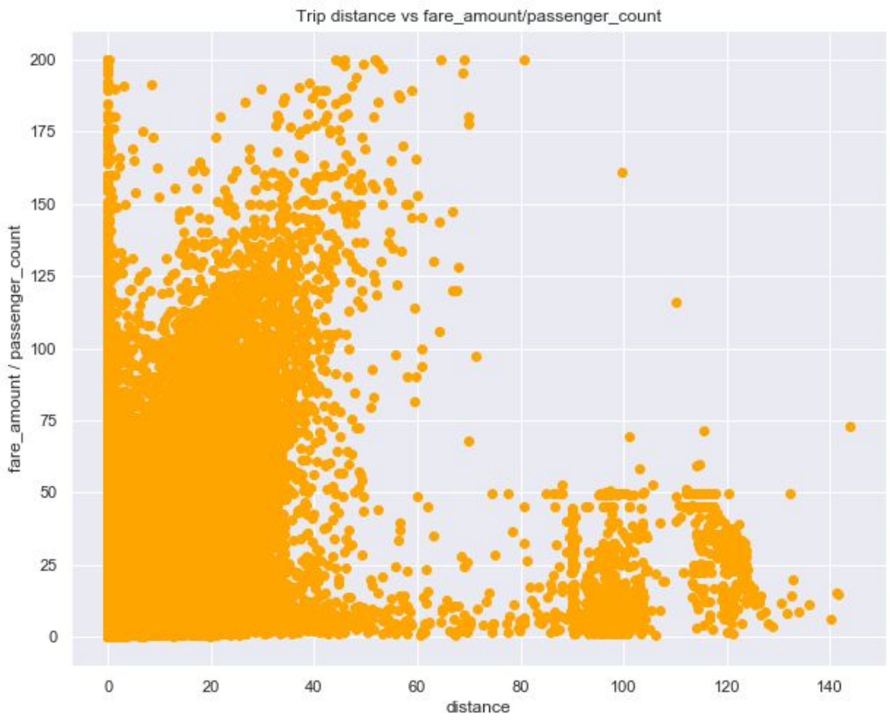
As shown in these plots, the concentration of coordinates are related to Manhattan neighborhoods Manhattan neighborhoods.





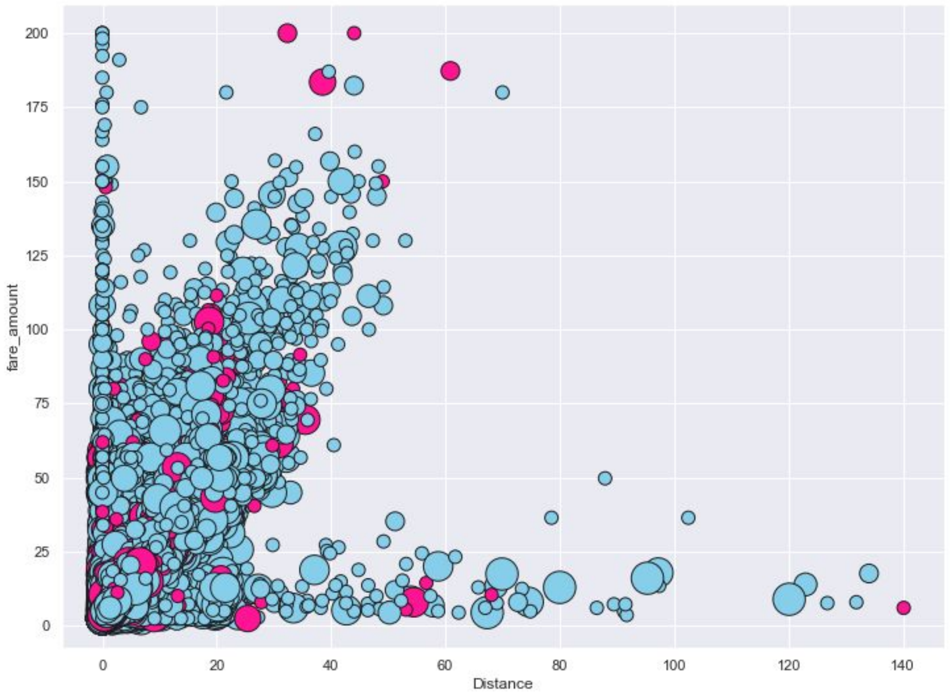
Fare_amount visualization

There is a positive correlation between distance and fare_amount per passenger, when distance < 70.

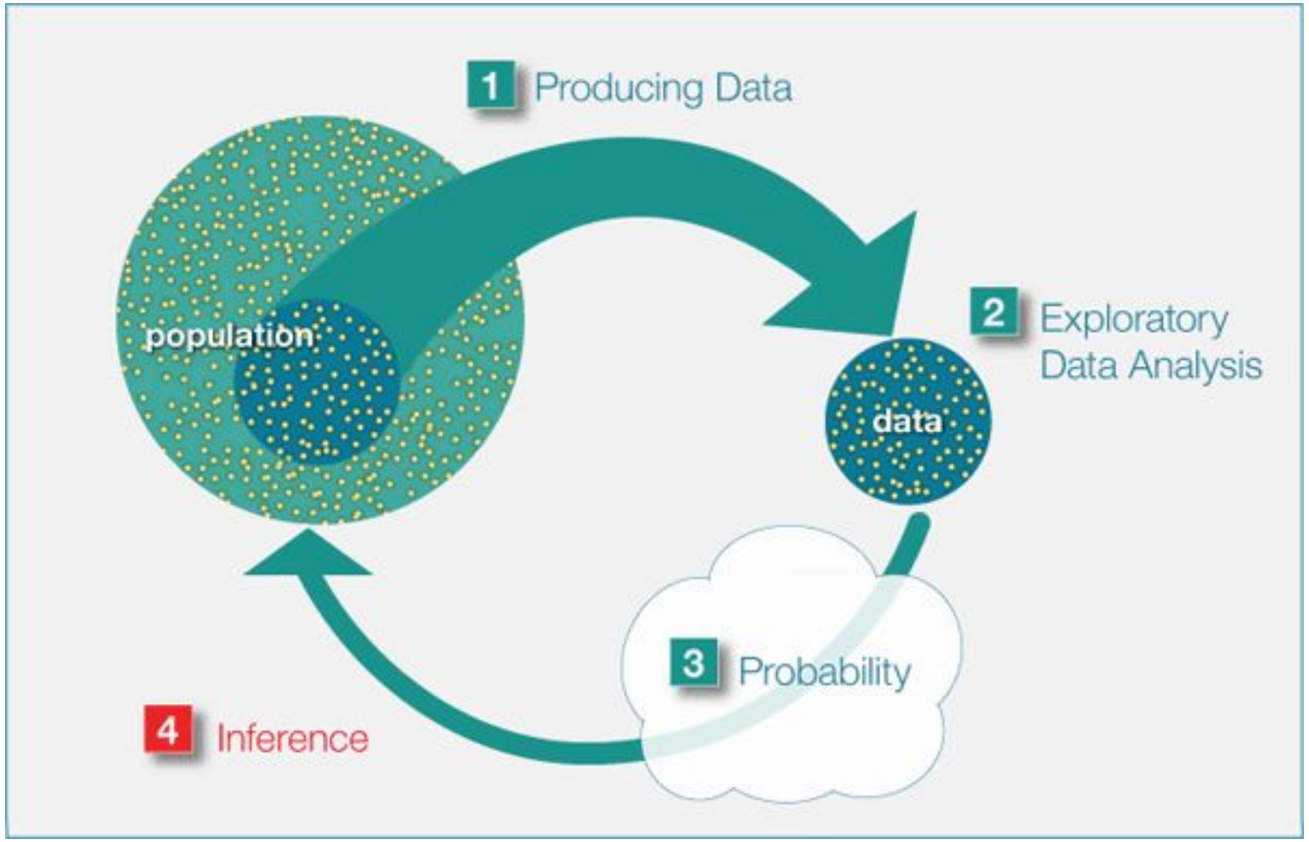


fare_amount changes over the years

This scatter plot represents that the fare_amount in the Manhattan area is lower than the other regions.



Statistical Inference

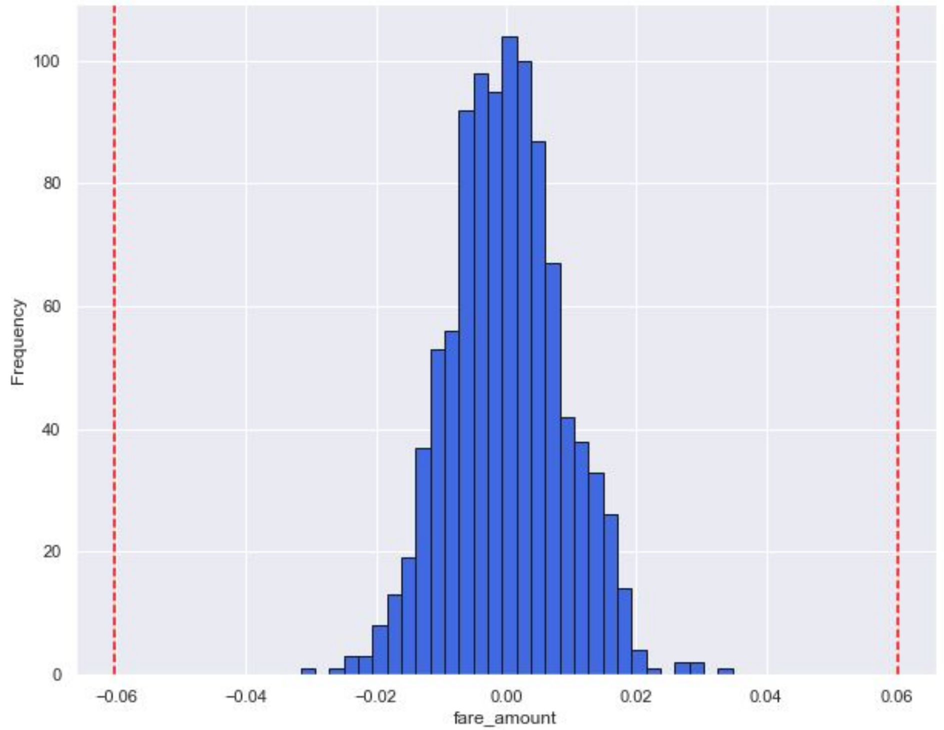


Hypothesis Testing 1

Null Hypothesis: the mean of fare_amount for weekdays and weekends are the same.

P_value is zero

the null hypothesis should be rejected .

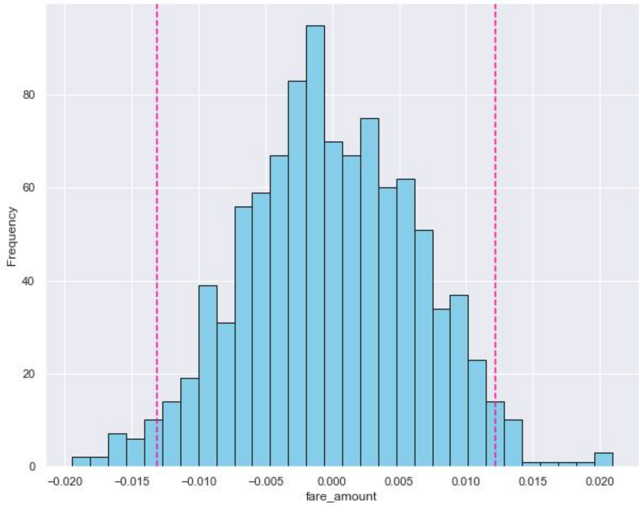


Hypothesis Testing 2

Null Hypothesis: the mean of fare_amount of Manhattan area is the same as other regions of NYC.

P_value is zero

the null hypothesis should be rejected



The mean of fare_amount of Manhattan is significantly different from the other NYC regions.

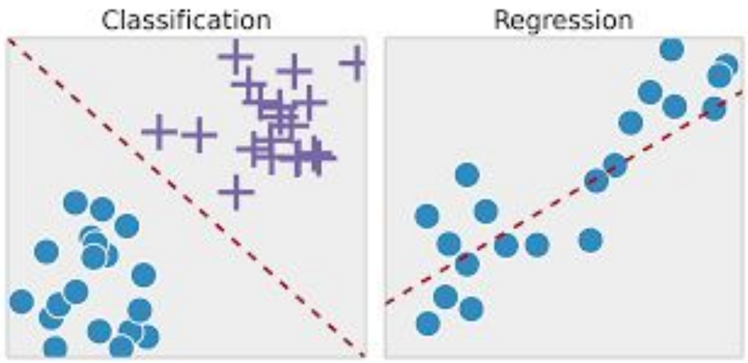
Machine Learning

- **Regression**

  Output variable is numerical (continuous)

- **Classification**

  Output variable is categorical (discrete)

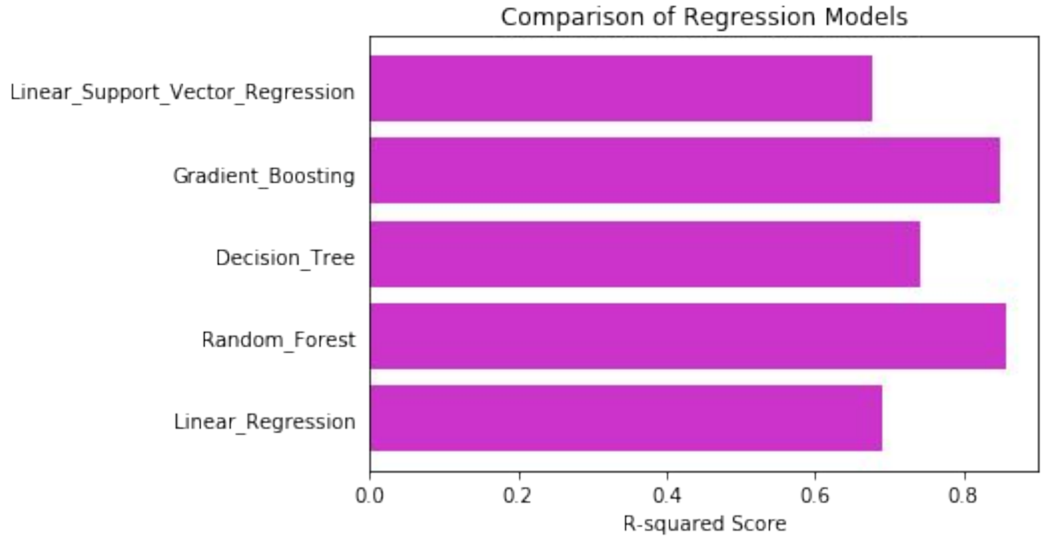


ML Regression Models Used

- Linear Regression
- Random Forest Regression
- Decision Tree Regression
- Gradient Boosting Regressor
- Linear Support Vector Regression

Evaluation: R-squared

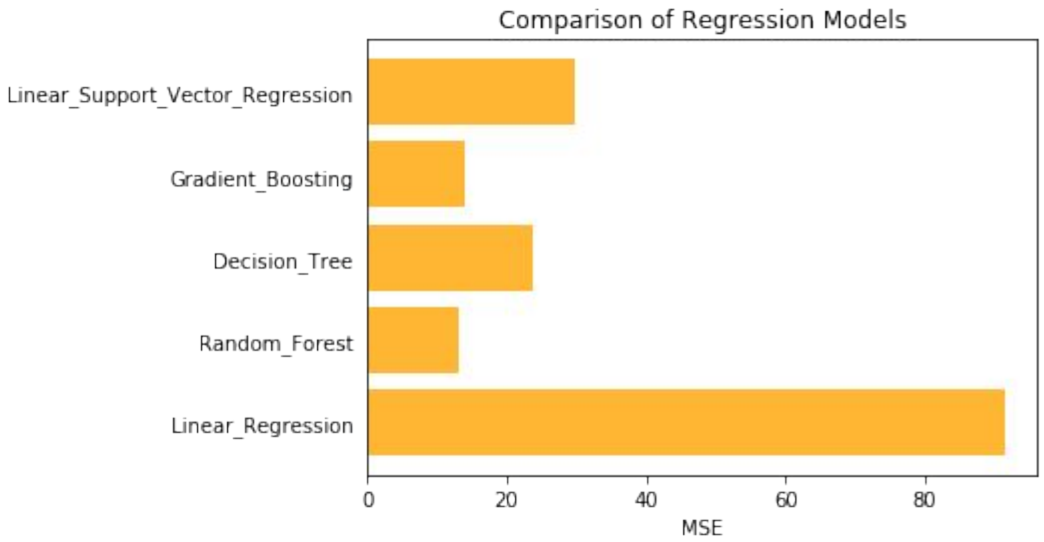
- The larger R-squared means the regression model fits the observations better.



Both Random Forest and Gradient Boosting obtain high values.

Evaluation: MSE

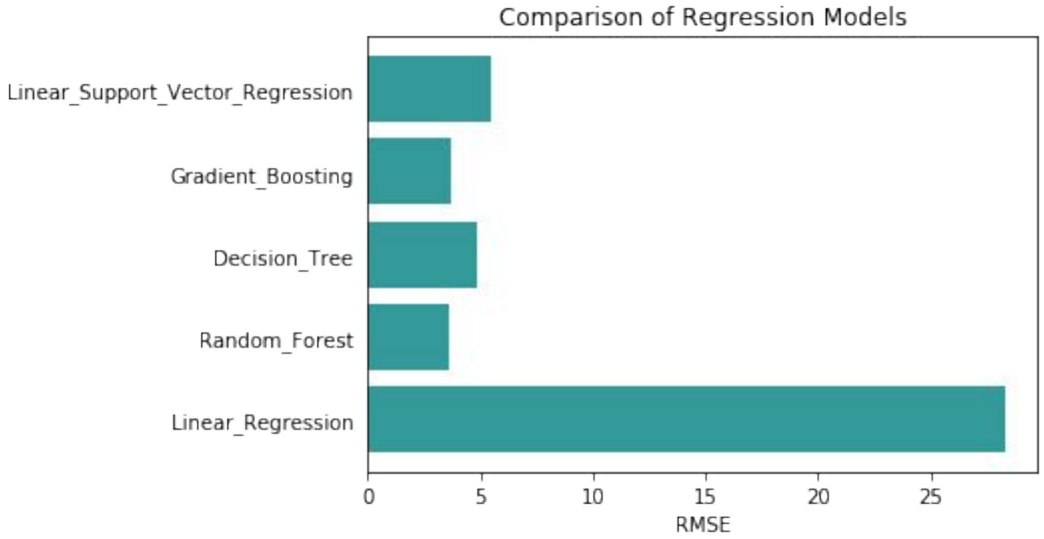
- Mean square error (MSE) : is the average of the square of the errors. The larger the number the larger the error.



Both Random Forest and Gradient Boosting have lower MSE compared to other models.

Evaluation: RMSE

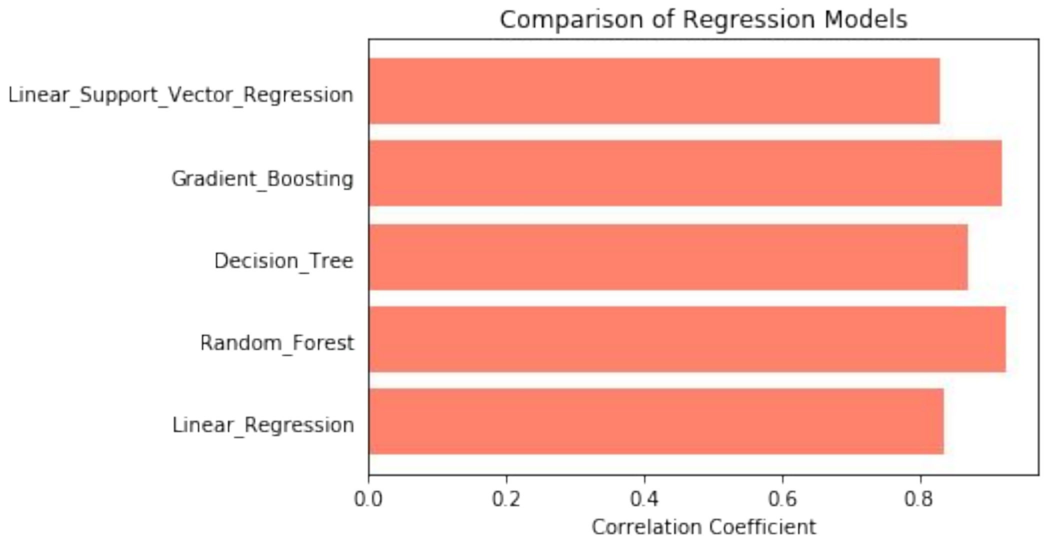
- Root Mean Square Error (RMSE) is a method of measuring the difference between values predicted by a model and their actual values.



Both Random Forest and Gradient Boosting obtain low values.

Evaluation: Correlation Coefficient

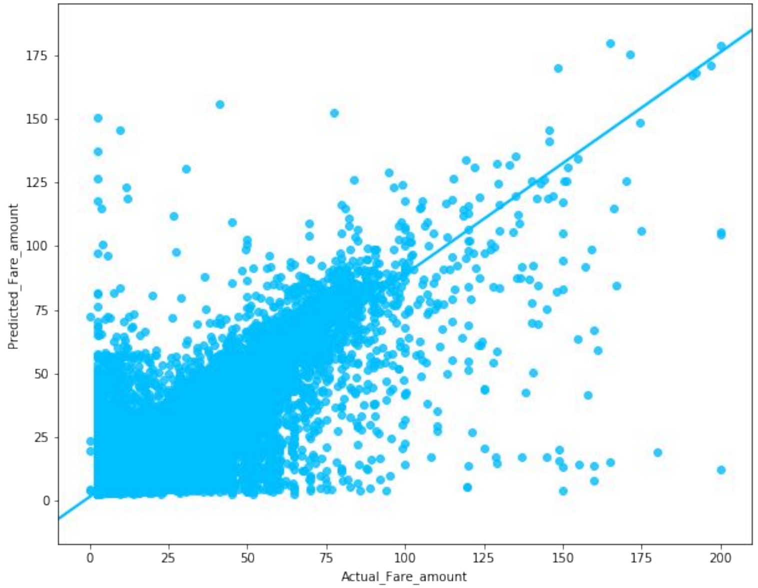
- Correlation Coefficient returns a value between -1 (full negative correlation) and 1 (full positive correlation).



Both Random Forest and Gradient Boosting indicate a notable correlation between target value and predicted values.

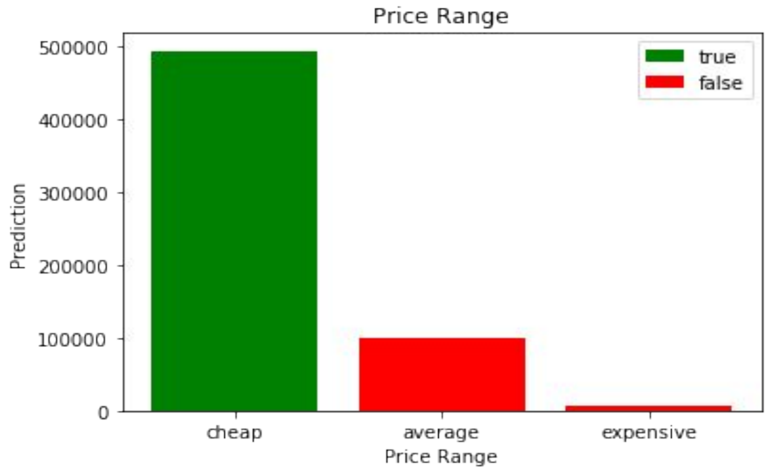
The chosen Regression Model

Random Forest outperforms all other models.



Classification Model

Naive Bayes Classifier: New York city taxi fare prediction is a regression problem. The continuous target is changed to labeled target by considering 3 different fare_amount categories.



Since most of the training dataset are in the cheap category, the prediction accuracy of this category is higher.

Conclusion

- For app-based hiring vehicles companies accurate prediction of taxi fare is important.
- To find the most appropriate prediction model, many factors are considered as problem features such as pickup or dropoff locations.
- Various regression and classification models are used to find the best prediction model.
- The results show that Random Forest outperforms all other models and obtains an accurate prediction model.