Project Name – Cab Fare Prediction

BY : RAHUL PRASAD SAH

1. Problem Statement :-

- You are a cab rental start-up company. You have successfully run the pilot project and now want to launch your cab service across the country. You have collected the historical data from your pilot project and now have a requirement to apply analytics for fare prediction. You need to design a system that predicts the fare amount for a cab ride in the city.

2. Data:-

- The details of data attributes in the dataset are as follows –

· **pickup_datetime** - timestamp value indicating when the cab ride started.
· **pickup_longitude** - float for longitude coordinate of where the cab ride started.
· **pickup_latitude** - float for latitude coordinate of where the cab ride started.
· **dropoff_longitude** - float for longitude coordinate of where the cab ride ended.
· **dropoff_latitude** - float for latitude coordinate of where the cab ride ended.
· **passenger_count** - an integer indicating the number of passengers in the cab ride.

3. Methodology :-

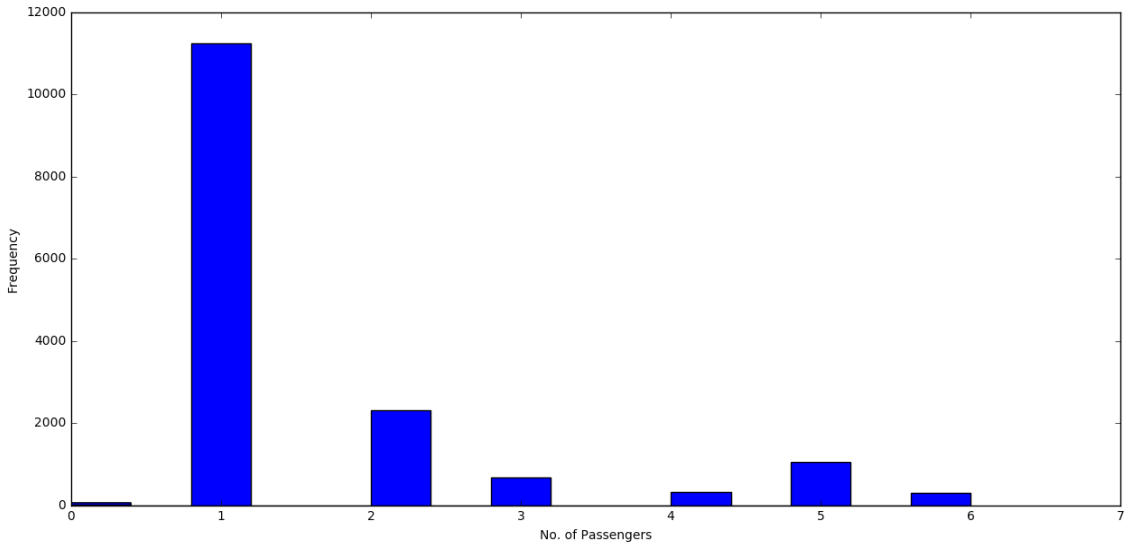
- Pre-Processing :-

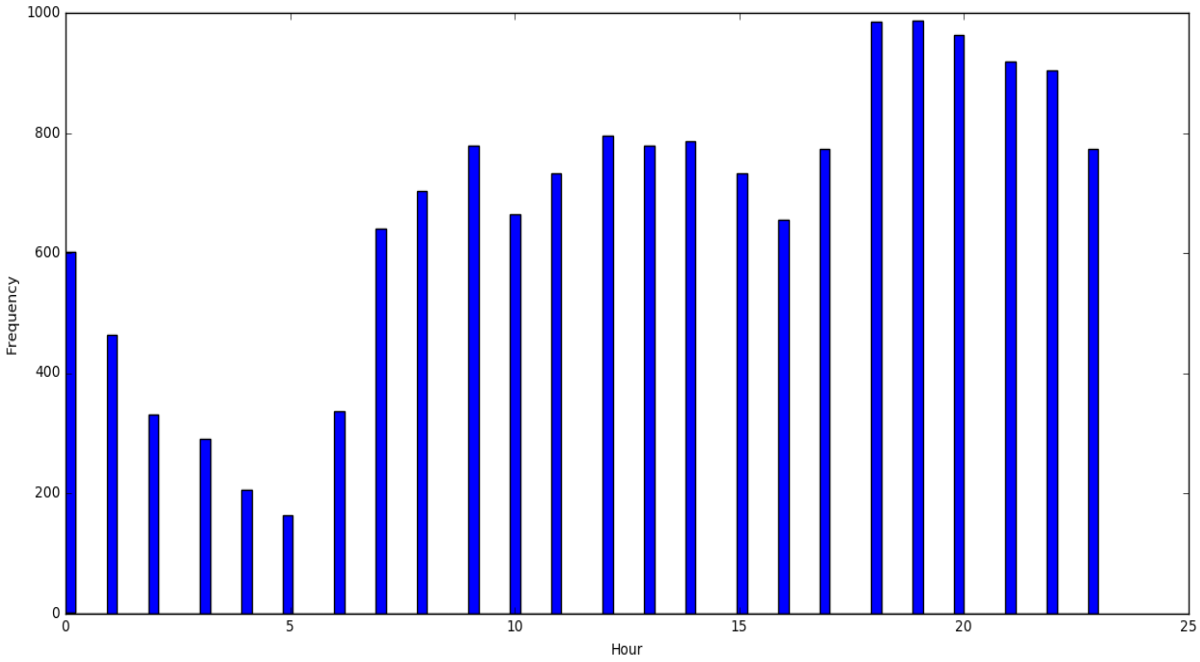
- Data pre-processing is the first stage of any type of project. In this stage we get the feel of the data. We do this by looking at plots of independent variables vs target variables. If the data is messy, we try to improve it by sorting deleting extra rows and columns. This stage is called as Exploratory Data Analysis. This stage generally involves data cleaning, merging, sorting, looking for outlier analysis, looking for missing values in the data, Imputing missing values if found by various methods such as mean, median, mode, KNN imputation, etc.

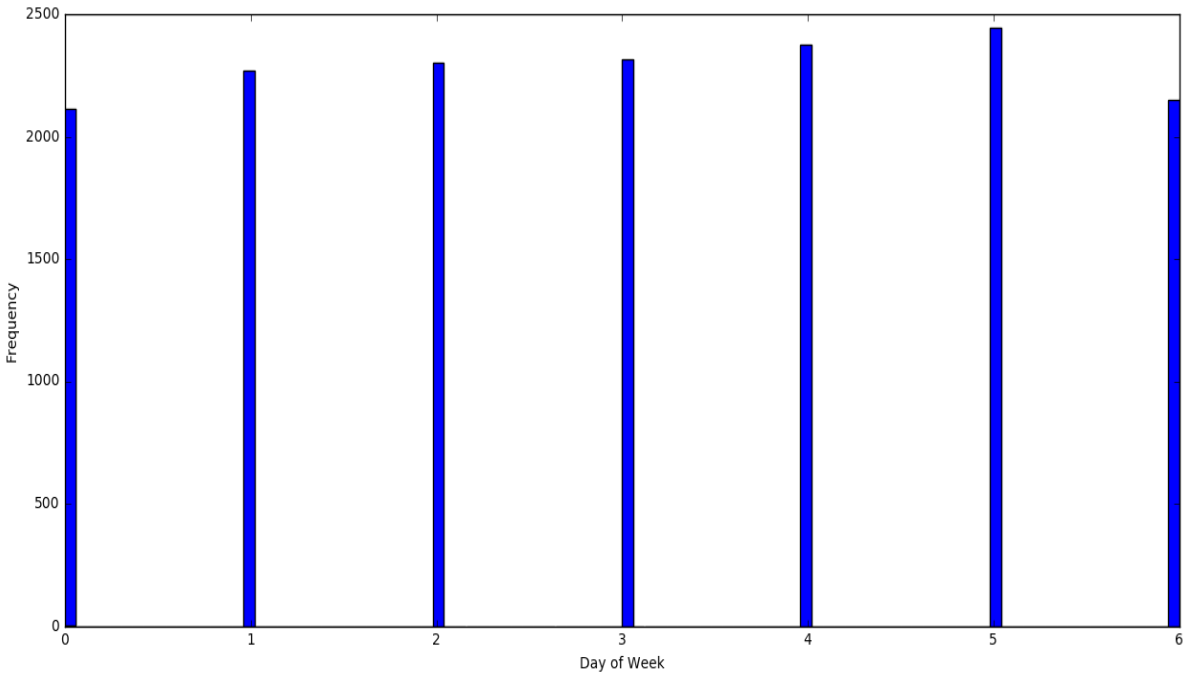
- Further we will look into what pre processing steps do this project was involved in.

4. Getting feel of data via visualization:-

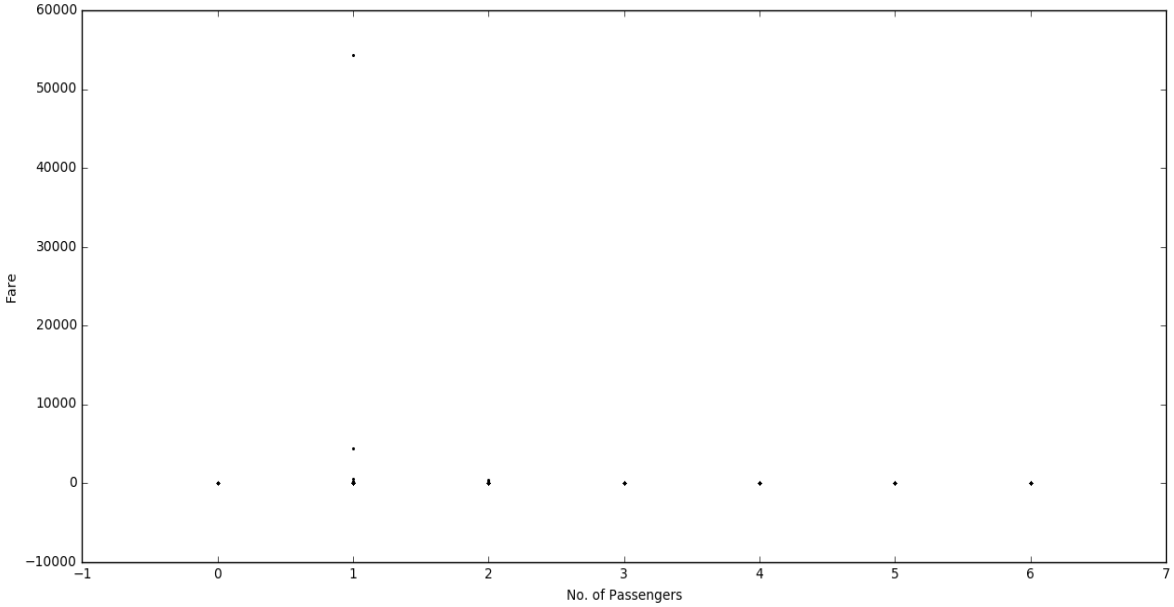
- I analysis data by bar and histogram after removing outliers :-

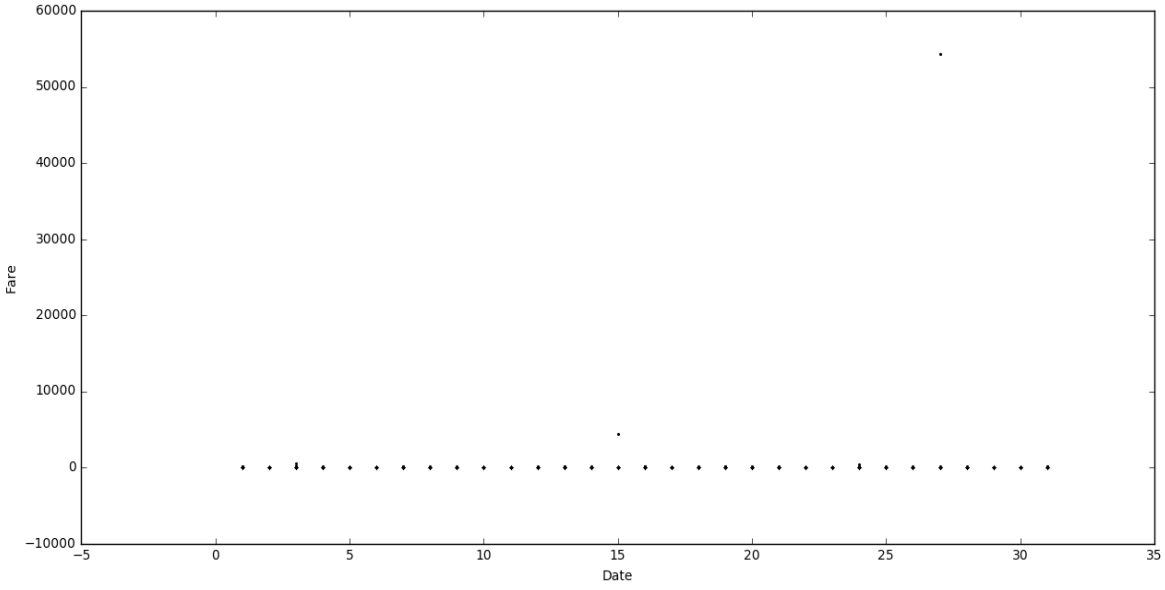


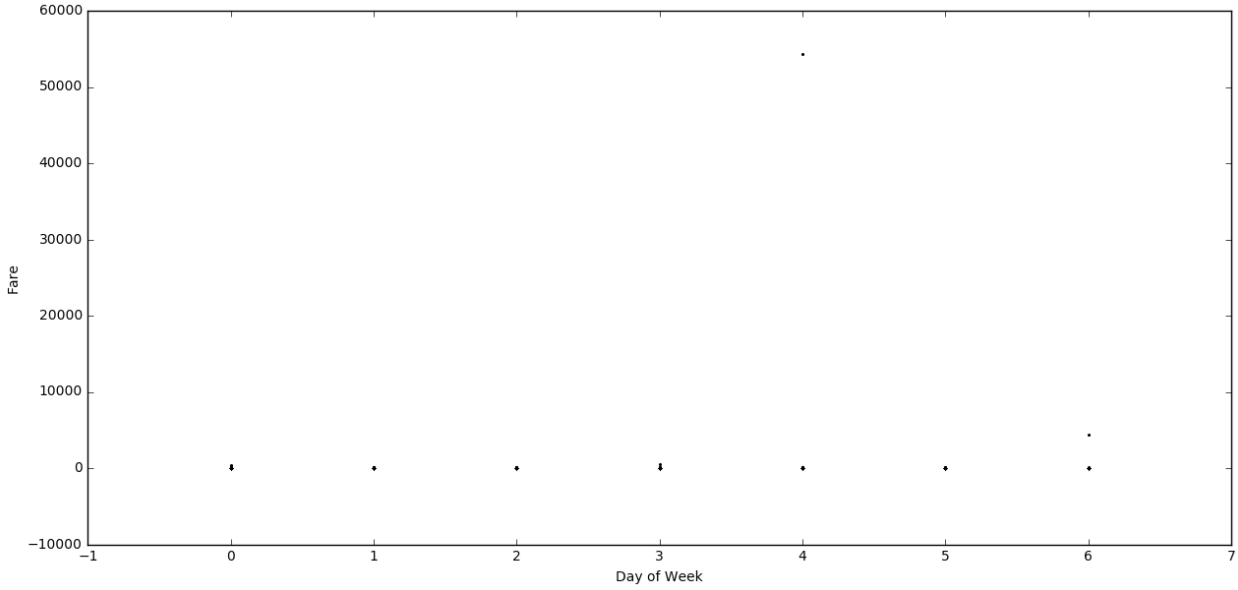

1000
800
600
400
200
0
Frequency
0
5
10
15
20
25
Hour


2500
2000
1500
1000
500
0
Frequency
0
1
2
3
4
5
6
Day of Week

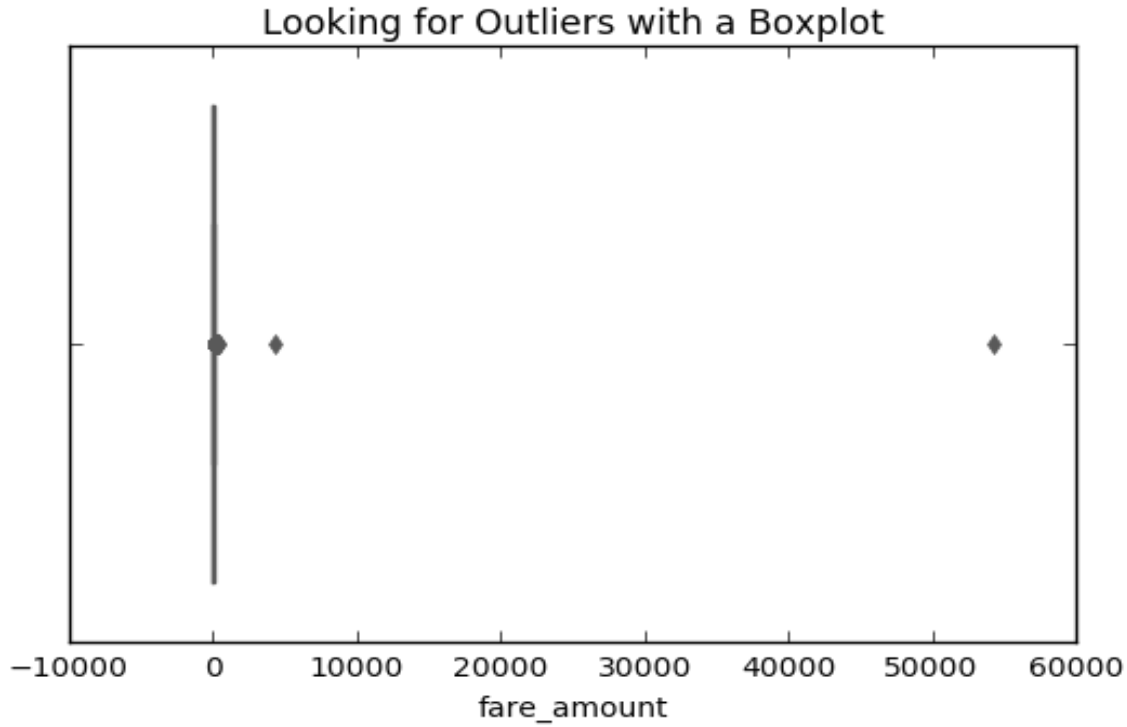
- Analysis data by Scatter plot after removing outliers :-

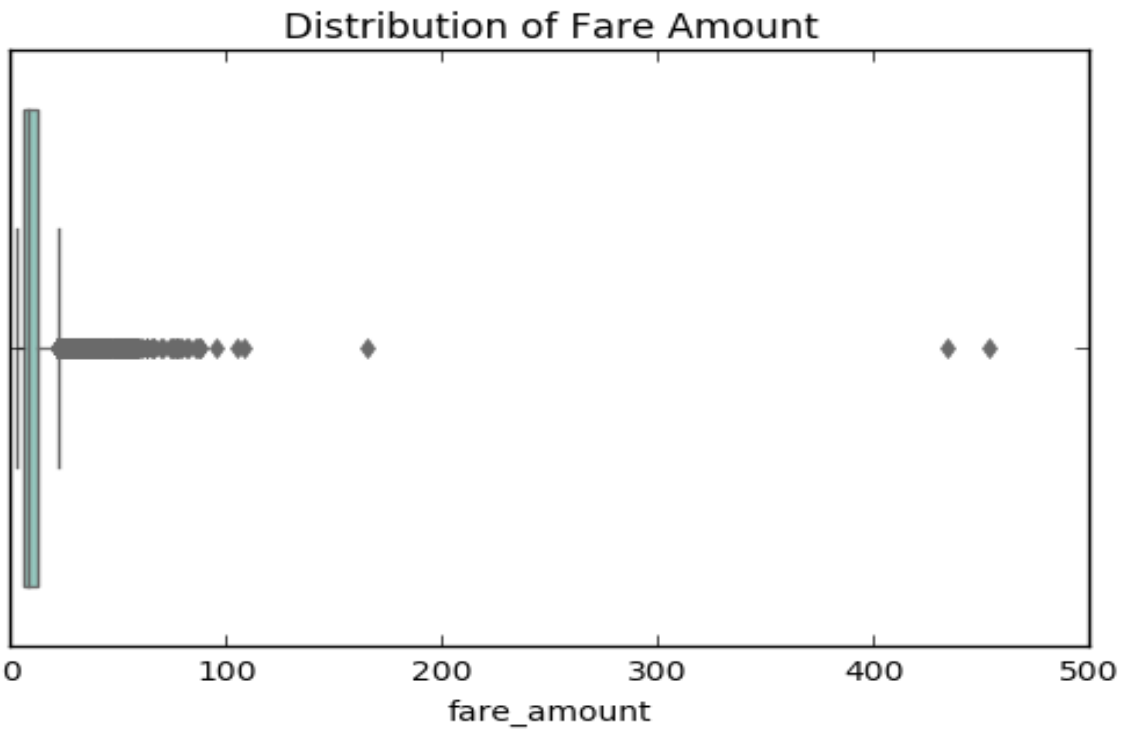






- Looking for outliers by Using Box Plot Method :-





- Visualize data by Distance Plot :-