Predicting Cab Rentals

(Chandradeep Pokhariya)

CONTENTS

INTRODUCTION

1.1 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.

1.2 Data

Data includes two datasets, one for training and another for testing. Data contains 7 variables and around 16067 observations. The variables are:

fare_amount - fare in a single cab ride.
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.

In these variables, **fare_amount** is independent variable while all other 6 variables are dependent.
The dataset also contains missing value, so we will have to first do the **missing value analysis.**

fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
16.900	2010-01-05 16:52:16 UTC	-74.01605	40.71130	-73.97927	40.78200	1
5.700	2011-08-18 00:35:00 UTC	-73.98274	40.76127	-73.99124	40.75056	2
7.700	2012-04-21 04:30:42 UTC	-73.98713	40.73314	-73.99157	40.75809	1
5.300	2010-03-09 07:51:00 UTC	-73.96810	40.76801	-73.95665	40.78376	1
12.100	2011-01-06 09:50:45 UTC	-74.00096	40.73163	-73.97289	40.75823	1
7.500	2012-11-20 20:35:00 UTC	-73.98000	40.75166	-73.97380	40.76484	1
16.500	2012-01-04 17:22:00 UTC	-73.95130	40.77414	-73.99009	40.75105	1
15.015	2012-12-03 13:10:00 UTC	-74.00646	40.72671	-73.99308	40.73163	1
8.900	2009-09-02 01:11:00 UTC	-73.98066	40.73387	-73.99154	40.75814	2
5.300	2012-04-08 07:30:50 UTC	-73.99634	40.73714	-73.98072	40.73356	1
4.100	2009-11-06 01:04:03 UTC	-73.99160	40.74471	-73.98308	40.74468	2
7.000	2013-07-02 19:54:00 UTC	-74.00536	40.72887	-74.00891	40.71091	1
7.700	2011-04-05 17:11:05 UTC	-74.00182	40.73755	-73.99806	40.72279	2

METHODOLOGY

2.1 Data Pre Processing

As the dataset may contain missing values and outliers, we need to first clean the data before feeding it to the algorithm. So at first we will proceed with missing value analysis and then outlier analysis and followed by feature selection.

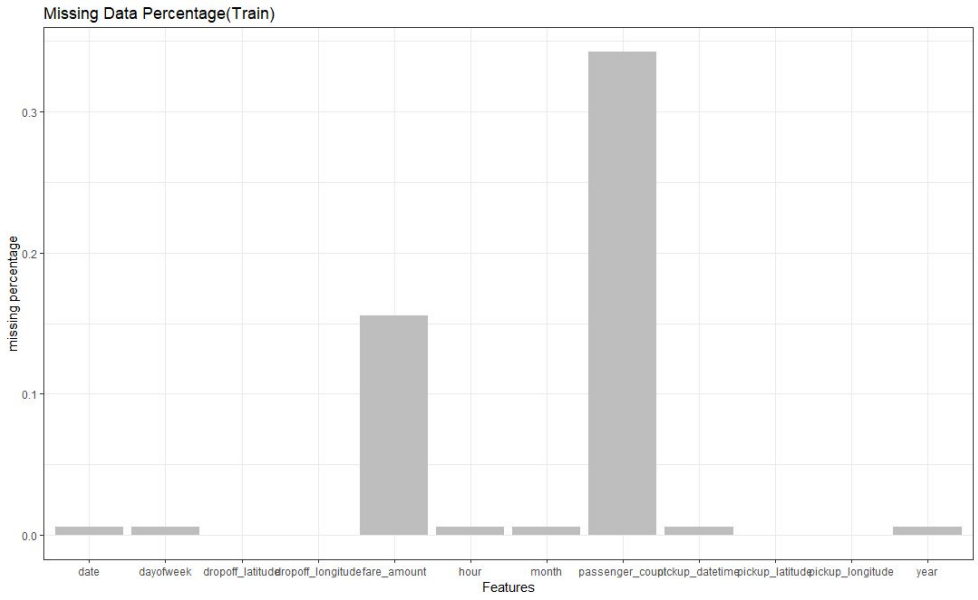
We have ‘pickup_datetime’ feature which is not clean, so we have to extract features from it. At first with the help of POSIXCt, we will extract date, time, month, year etc.

Also through haversine formula we calculate the distance feature which calculate the distance between given longitudes and latitudes.

2.1.1 Missing Value Analysis

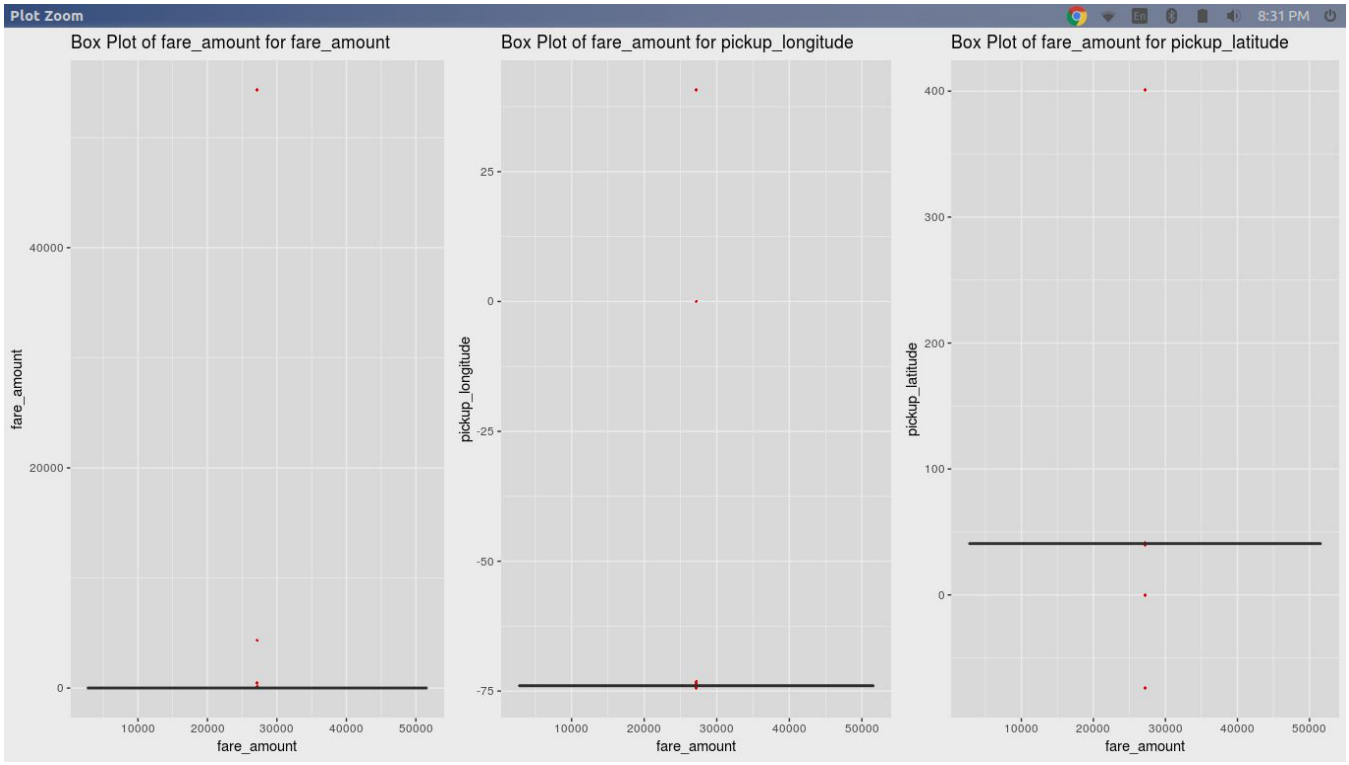
Missing values can be due to various reasons. It may be due to some human error, refuse to answer while surveying or optional box in questionnaire. We will first find the missing values in the dataset and then impute them with different methods of imputation.

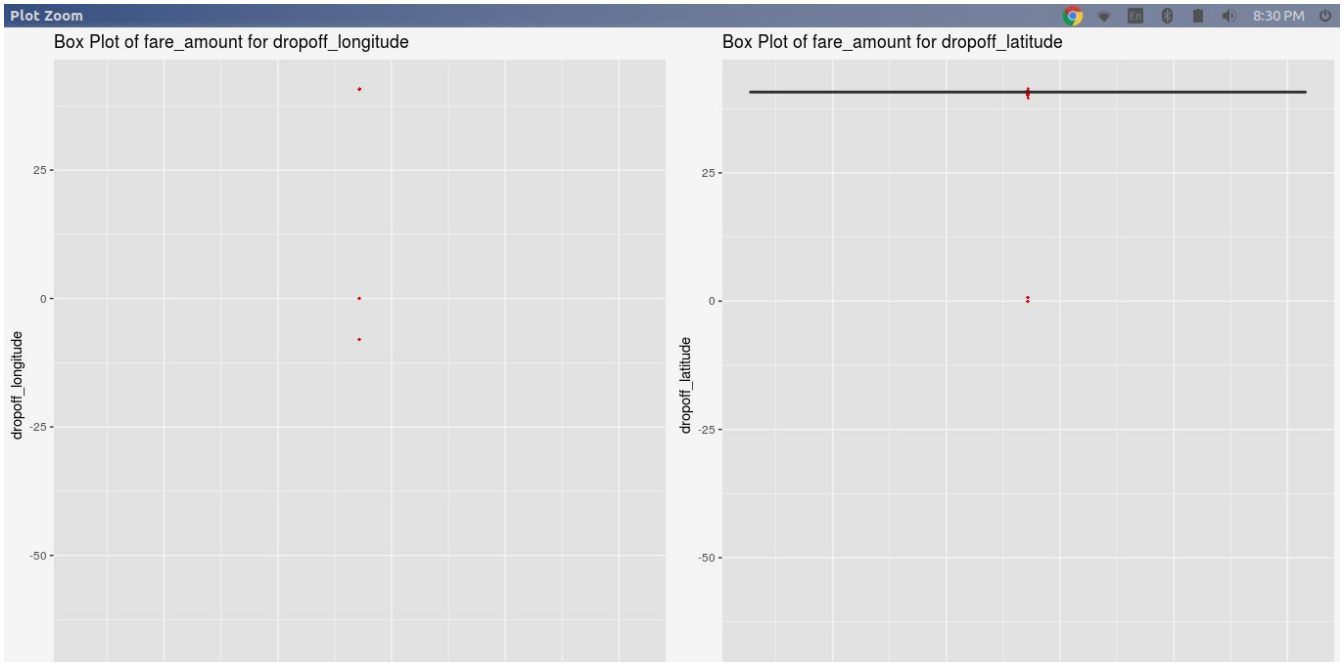
But here we are directly removing NA value, as these incorrect values are very less and we can ignore them..

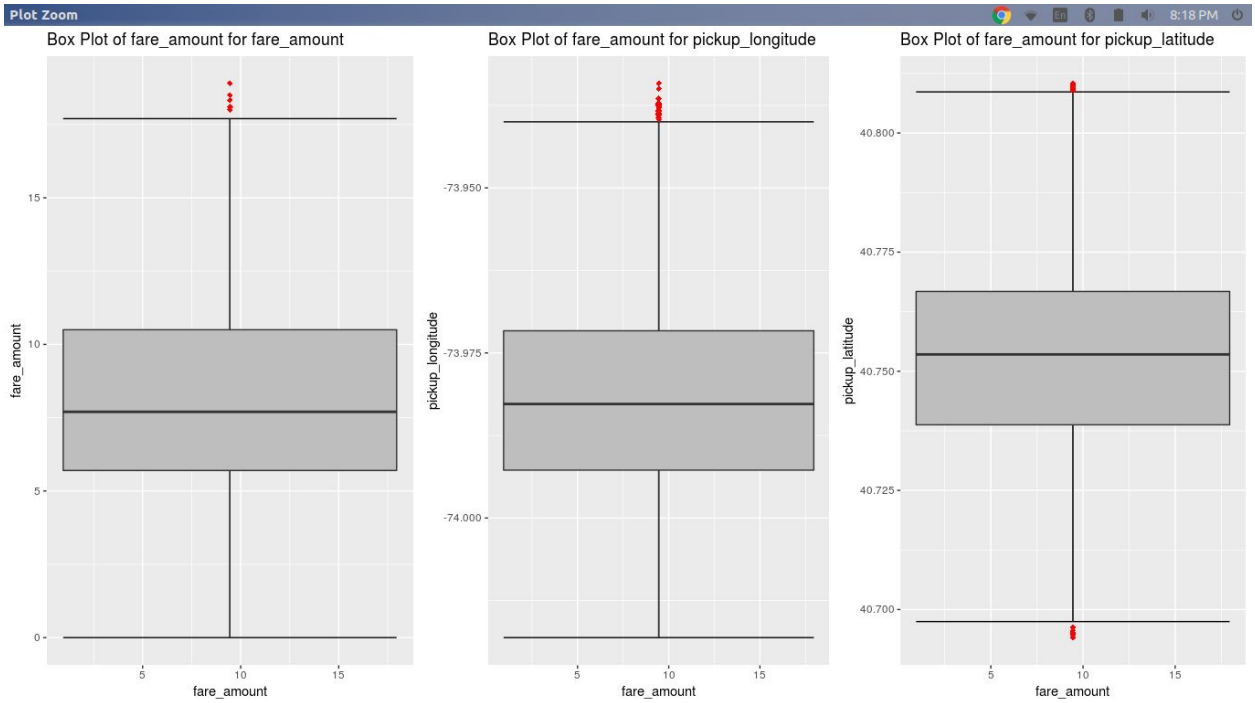


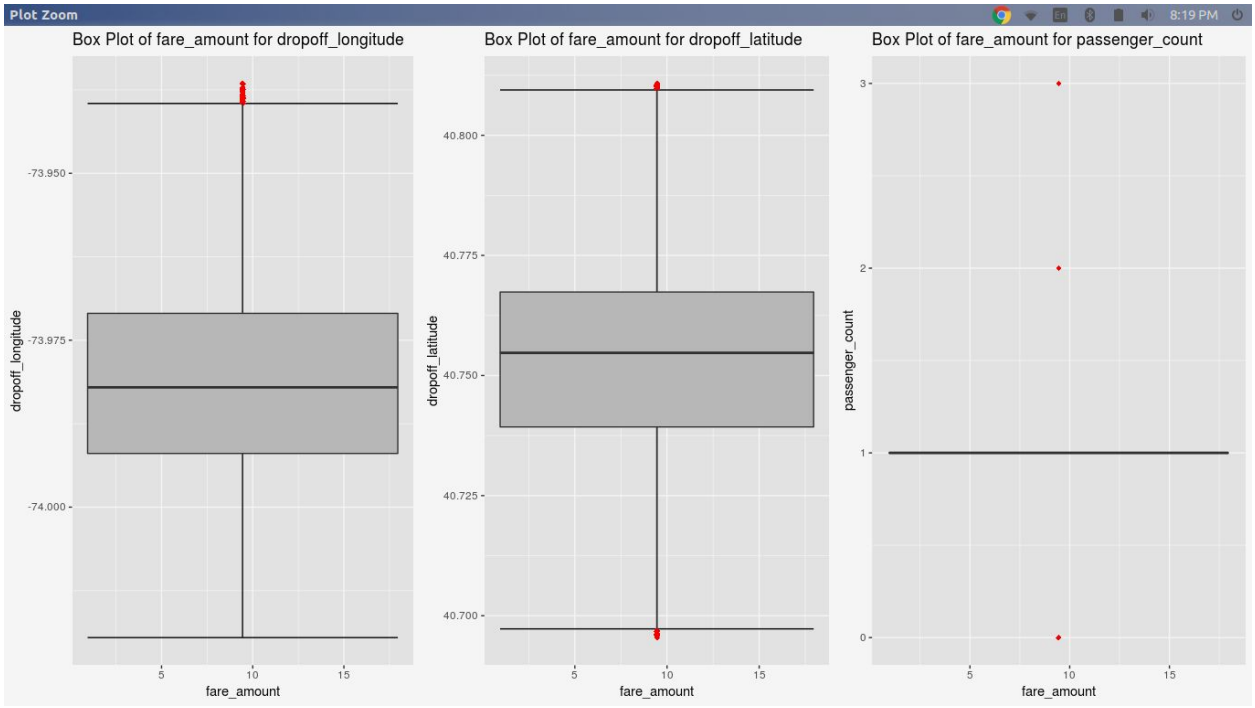
2.1.2 Outlier Analysis

In outlier analysis we check the data which does not lie in the normal range of data. Either it is too more or too less than the data. Outlier are the observations which are inconsistent with the rest of the data.
Outliers can be caused due to the poor quality, low quality measurements manufacturing equipment, manual error etc.
There are different techniques to detect outliers. Among them box plot is the best method to detect and remove the outlier.

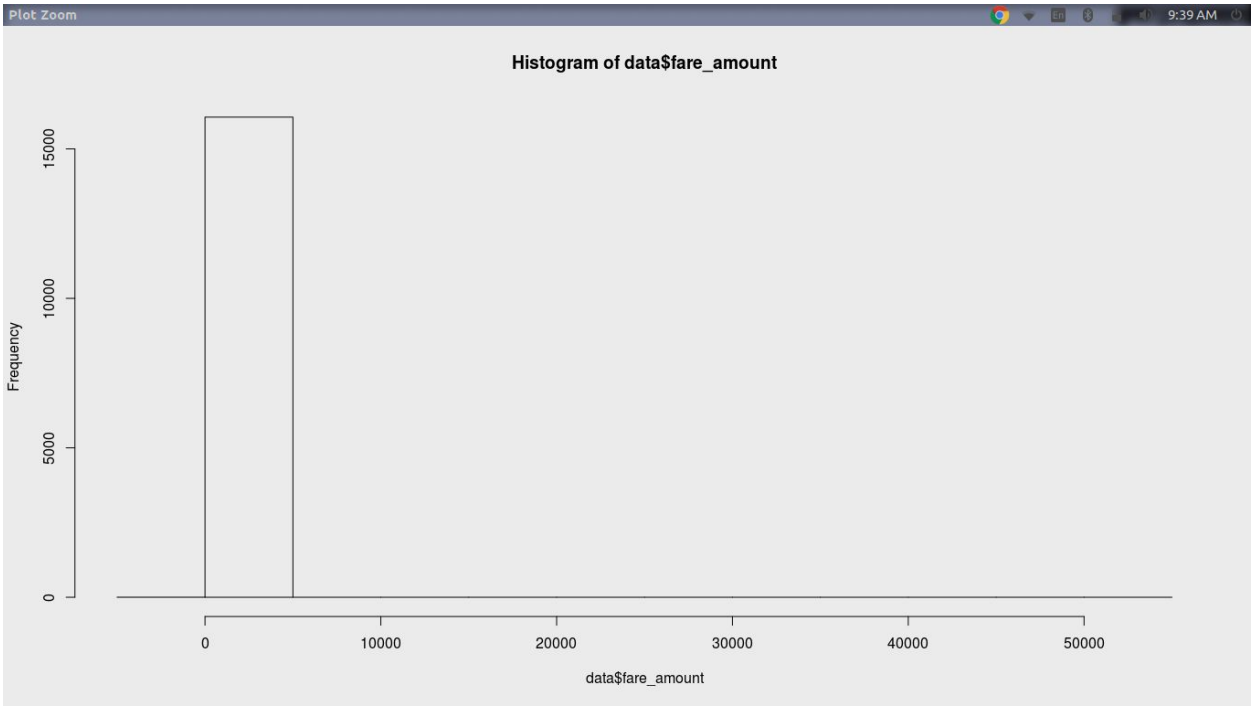


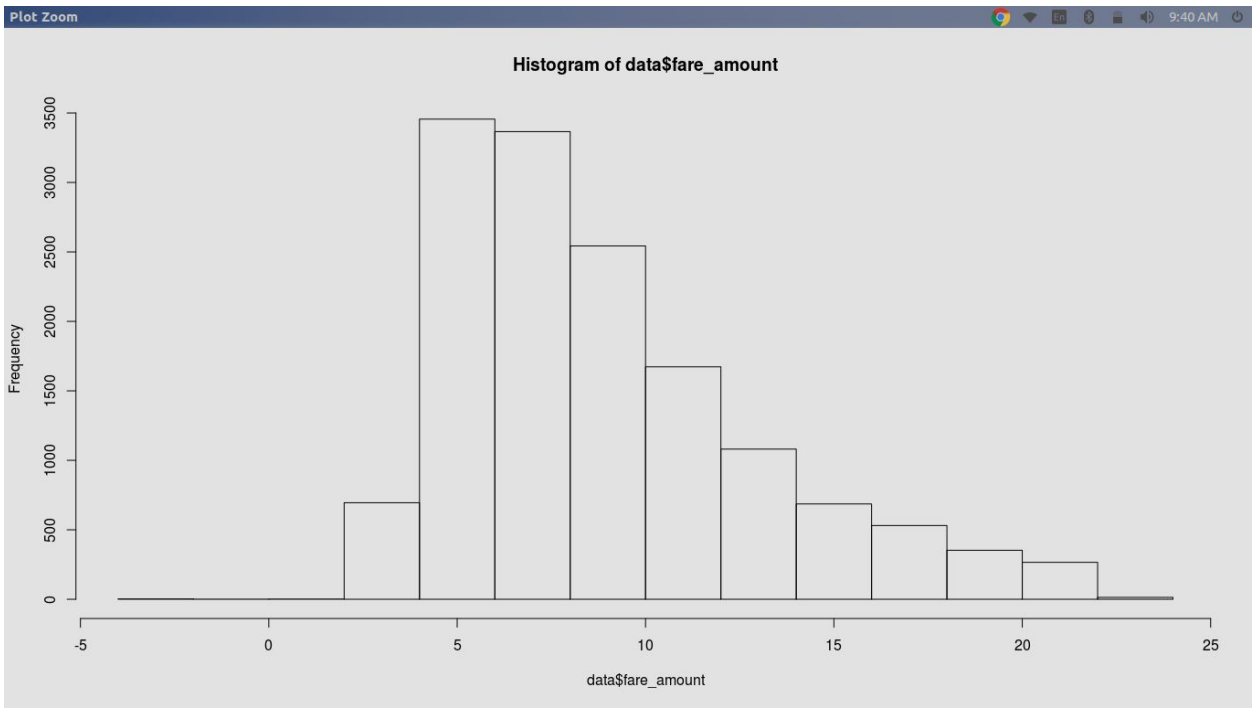



Plot Zoom
8:18 PM
Box Plot of fare_amount for fare_amount
fare_amount
fare_amount
Box Plot of fare_amount for pickup_longitude
pickup_longitude
-73.950
-73.975
-74.000
fare_amount
Box Plot of fare_amount for pickup_latitude
pickup_latitude
40.800
40.775
40.750
40.725
40.700
fare_amount


Plot Zoom
8:19 PM
Box Plot of fare_amount for dropoff_longitude
dropoff_longitude
-73.950
-73.975
-74.000
fare_amount
Box Plot of fare_amount for dropoff_latitude
dropoff_latitude
40.800
40.775
40.750
40.725
40.700
fare_amount
Box Plot of fare_amount for passenger_count
passenger_count
fare_amount

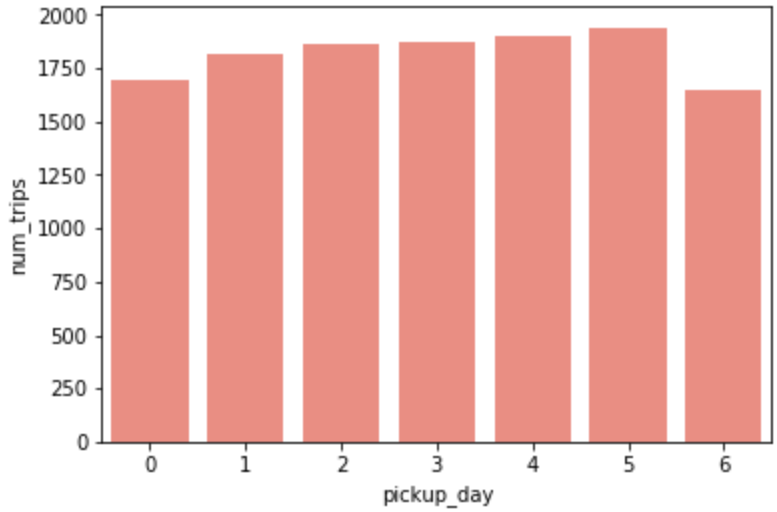
Histogram of fare_amount before and after the outlier analysis



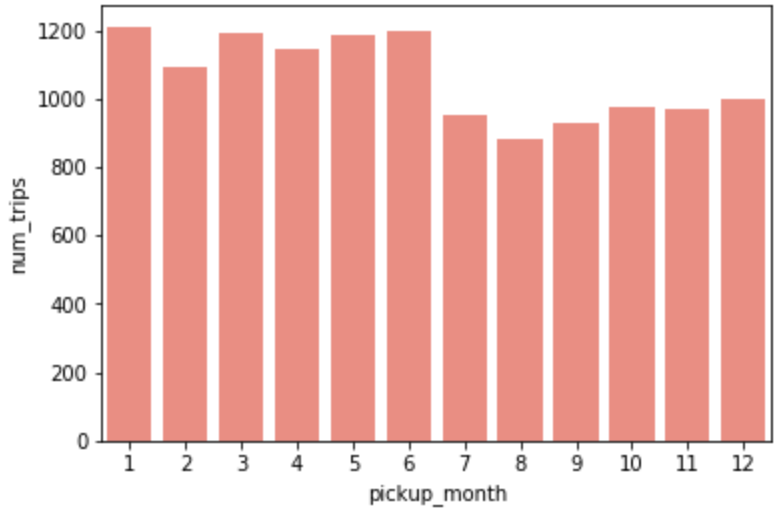


Distribution of the features with the target variable:

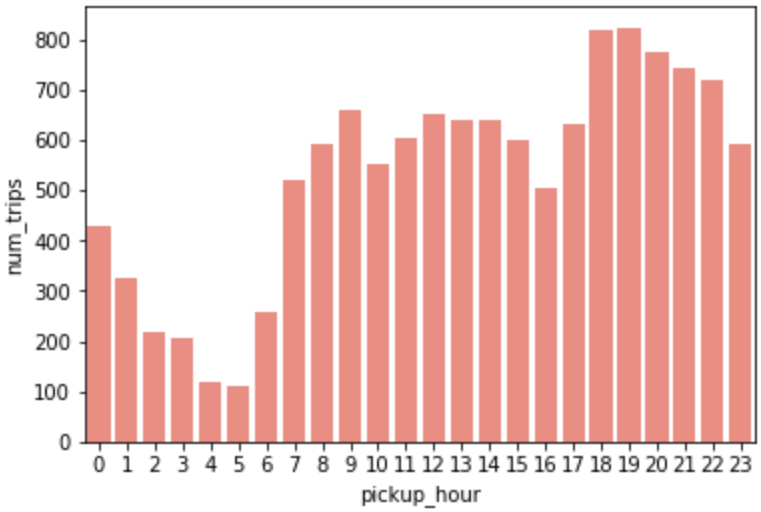
Pickup day & number of trips:



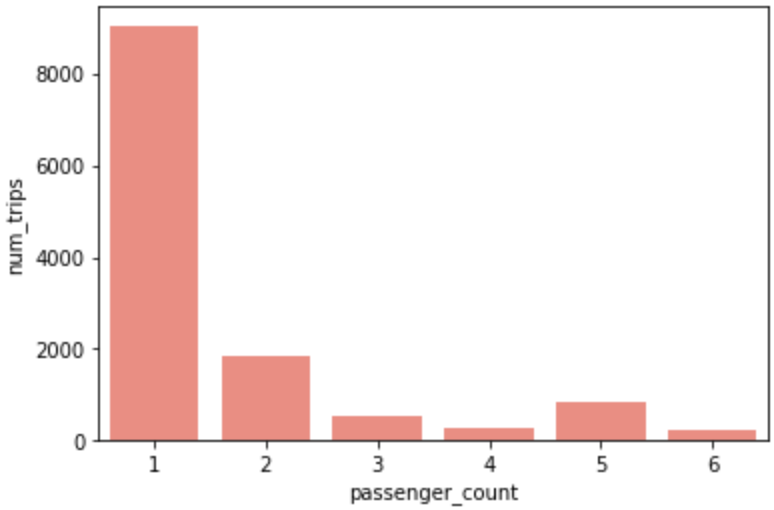
Pickup month & number of trips



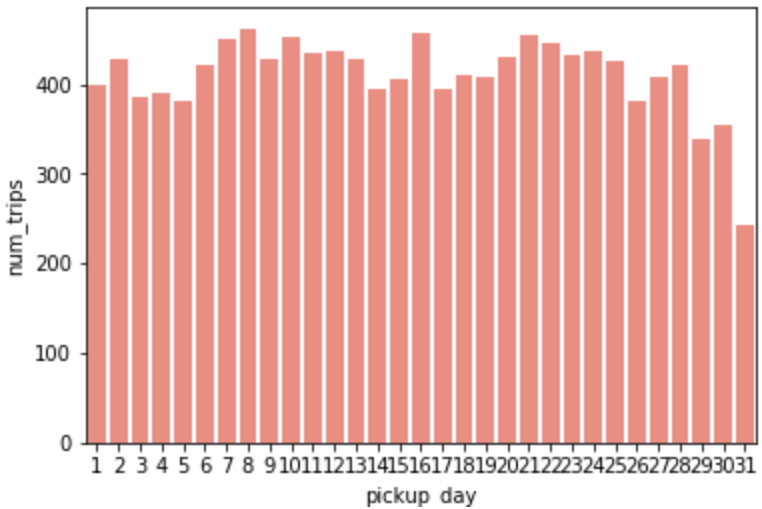
Pickup hour & number of trips:



Passenger count & number of trips

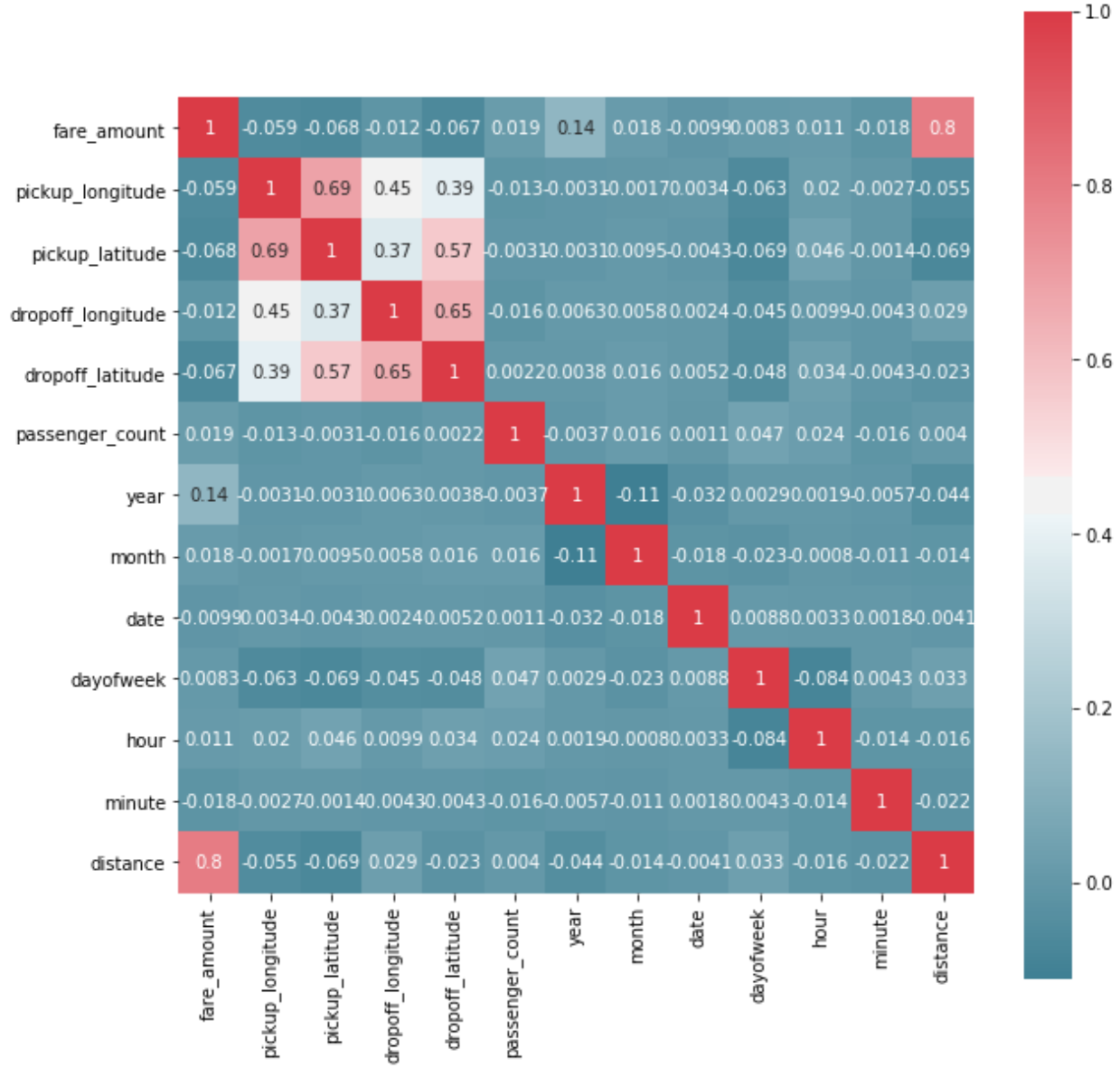


Pickup day & number of trips

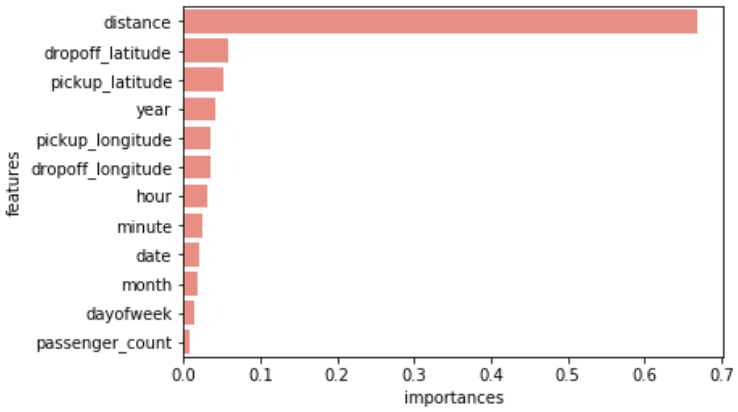


2.1.3 Feature Selection

In feature selection we select a subset of relevant features or variables for use in model construction. Basically we check which features of the dataset do not contain the correlation with the dependent variable (fare_amount).
So at first i made the correlation graph and checked the correlation between different features with the target variable. The feature which has the least relation with the target variable is dropped out.
This is the correlation matrix i got:



I further checked the importance of the features through random forest.
Here is the result:



2.2 Modeling

At first we started with splitting the data, putting 75% in the training set and 25% in test set.
Then we checked for the multicollinearity with the help of 'vif' function.
After checking for the multicollinearity we found no multicollinearity in the dataset.

Random Forest----------

Random forest is very popular model in predicting the data. I started with random forest model because of its 'importance' function which helps in calculating the importance of the features in the model.
Here is the result of the model :
Accuracy: 82.08912
MAPE:17.916
r2_score : 0.691

Linear Regression----------

After Random Forest I used the linear regression model, very popular model in predicting regression datas.

For Linear regression model, there should be no multicollinearity. And our dataset passes this multicollinearity test.

Here are the result of the Linear regression model:
Accuracy: 82.062
MAPE:17.9376
r2_score : 0.687

Decision Tree----------

Decision Tree model can be used for both classification and regression model. Here we use it for this regression model to find the fare amount.
Decision Tree is a flow chart type structure, where each terminal denotes a test set on an attribute, each branch denotes the outcome of a test and each leaf holds a class level.

The topmost node in a tree is a root node.

Here are the results of the decision tree model:
MAPE 23.55109570556832
Accuracy 76.44890429443169
r2_score 0.4247394322767284

2.3 Model Evaluation:

Since in our dataset the data was continuous, so we consider basically two type of error matrices here.

1. MAPE error matrix
2. R squared matrix

1. MAPE error matrix:
MAPE is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a loss function for regression problems in machine learning.It usually express accuracy as a percentage.

2. R Squared Error:
R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model.

R-squared = Explained variation / Total variation

2.4 Model selection:

Accuracy of different models:

1. Random Forest: Accuracy (**82.08912**) / MAPE (**17.916**) / r-squared (**0.691**)
2. Linear Regression: Accuracy (**82.062**) / MAPE (**17.9376**) / r-squared (**0.687**)
3. Decision Tree Regressor: Accuracy (**76.448**) / MAPE (**23.5510**) / r-squared (**0.424**)

Since the accuracy of the Random Forest is highest, we will freeze this model for the test dataset.

Cleaning the test data:

After we freeze the model, we clean the test data. We create the features by extracting the date, time, month etc from the 'pickup_datetime' feature.
After that we drop the variable.
We further check for the missing value, and in our case there is none.

After doing these sort of operations we put our model in test data and predict the fare amount for the test dataset, then we include the obtained values in a new column in the test data and export it as csv file.