Project Report on

**Cab Fare Prediction**

**By**

**PavanKumar. BL**

Contents

1. Introduction

1.1 Problem Statement

1.2 Data

1.3 Exploratory Data Analysis

2. Methodology

2.1 Pre-Processing

2.1.1 Missing Value Analysis

2.1.2 Outlier Analysis

2.1.3 Feature Selection

2.2 Modeling

2.2.1 Linear Regression

2.2.2 Random Forest

3. Conclusion

3.1 Model Evaluation

3.2 Model Selection

3.3 Answers of asked questions

Appendix

Extra Figures

Chapter 1

Introduction

* 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. Data

Our task is to Build a suitable model that will best fit for analyzing fare prediction for test data provided.

There are 07 variables in our data in which 6 are independent variables and 1 (Fare\_amount) is dependent variable. Since our target variable is continuous in nature, this is a regression problem.

Variables Information:

**1.** fare\_amount

**2.** pickup\_datetime

**3.** pickup\_longitude

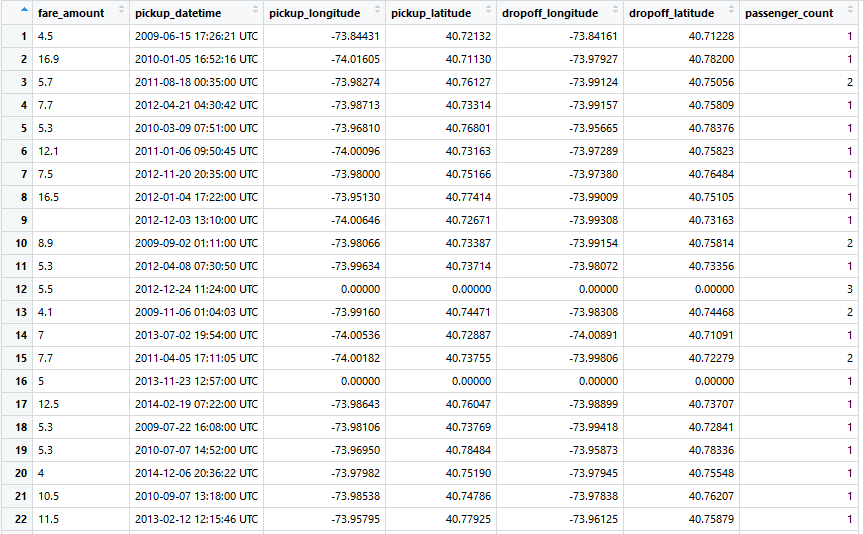
**4.** pickup\_latitude

**5.** dropoff\_longitude

**6.** dropoff\_latitude

**7.** passenger\_count

* 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.



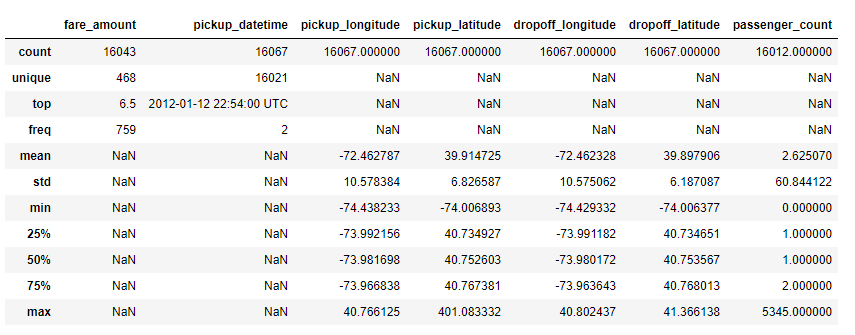
Chapter 2

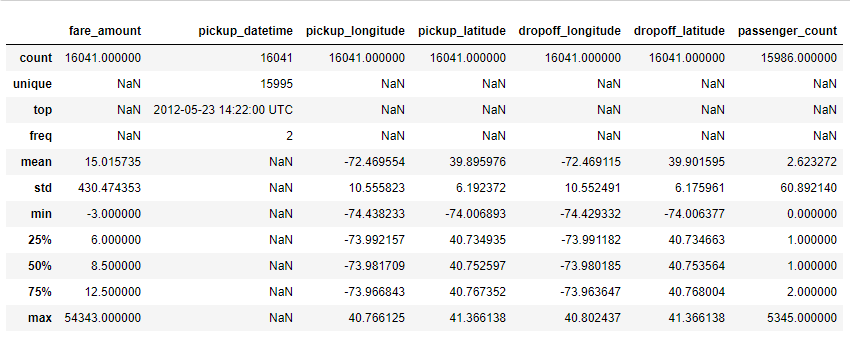
Methodology

2.1 Pre-Processing

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis.**

To start this process, we will first try and look at all the probability distributions of the variables.

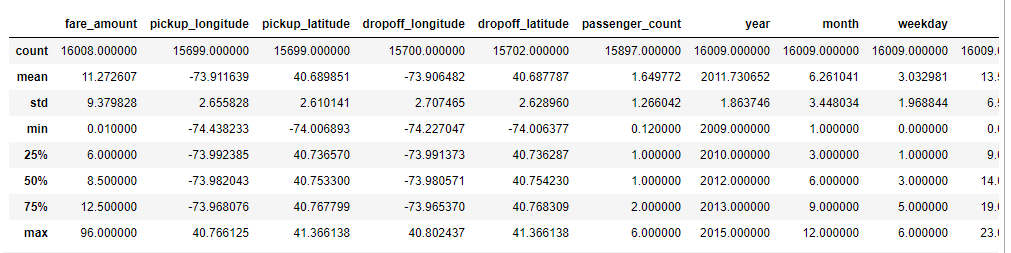




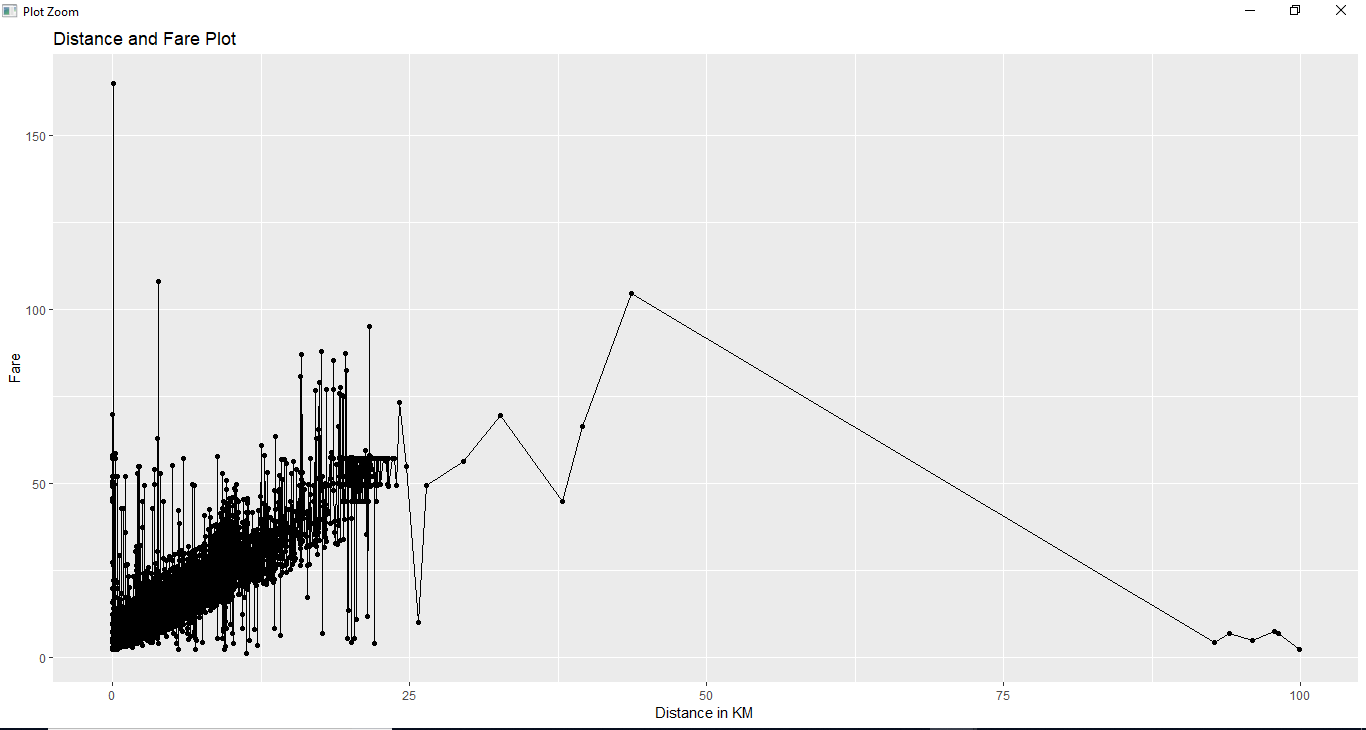
From above details we can confirm that

* there are some Outliers
* missing values
* Negative fare amount
* Max Passenger Count is High

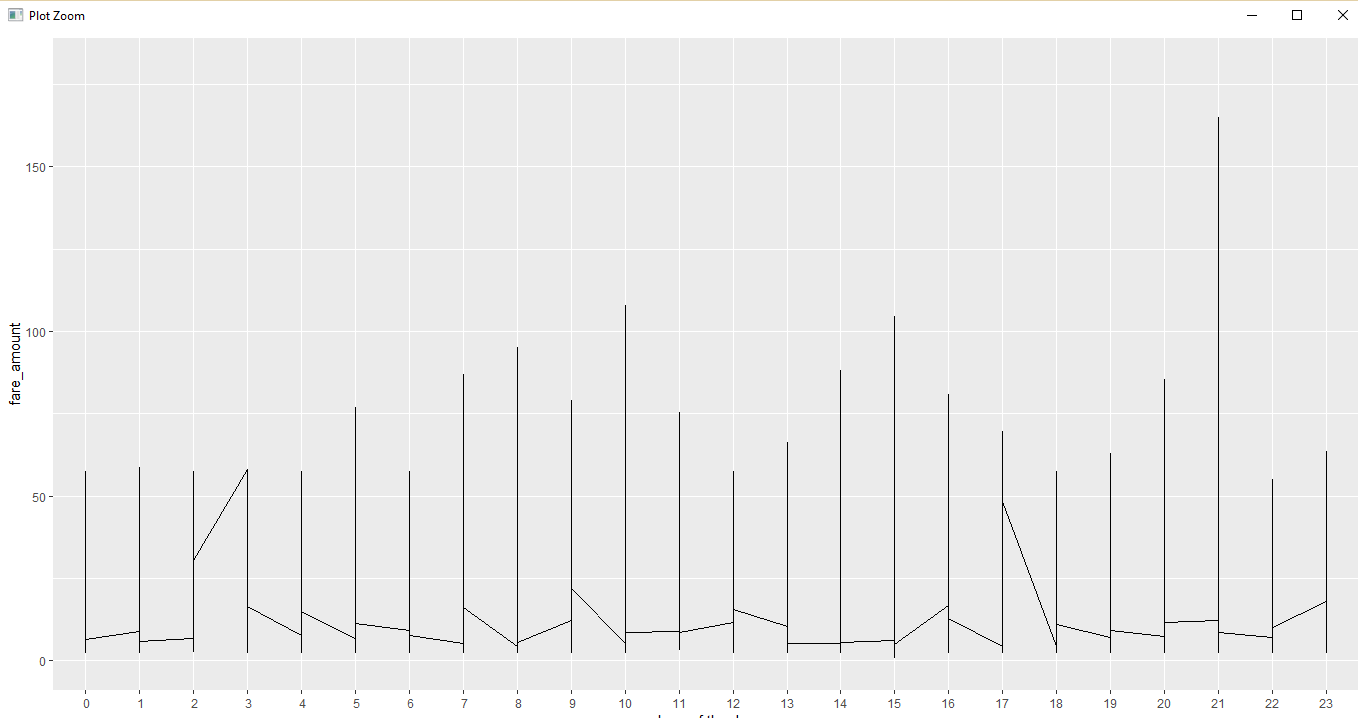
Below is the Image that explains how our data looks like after EDA



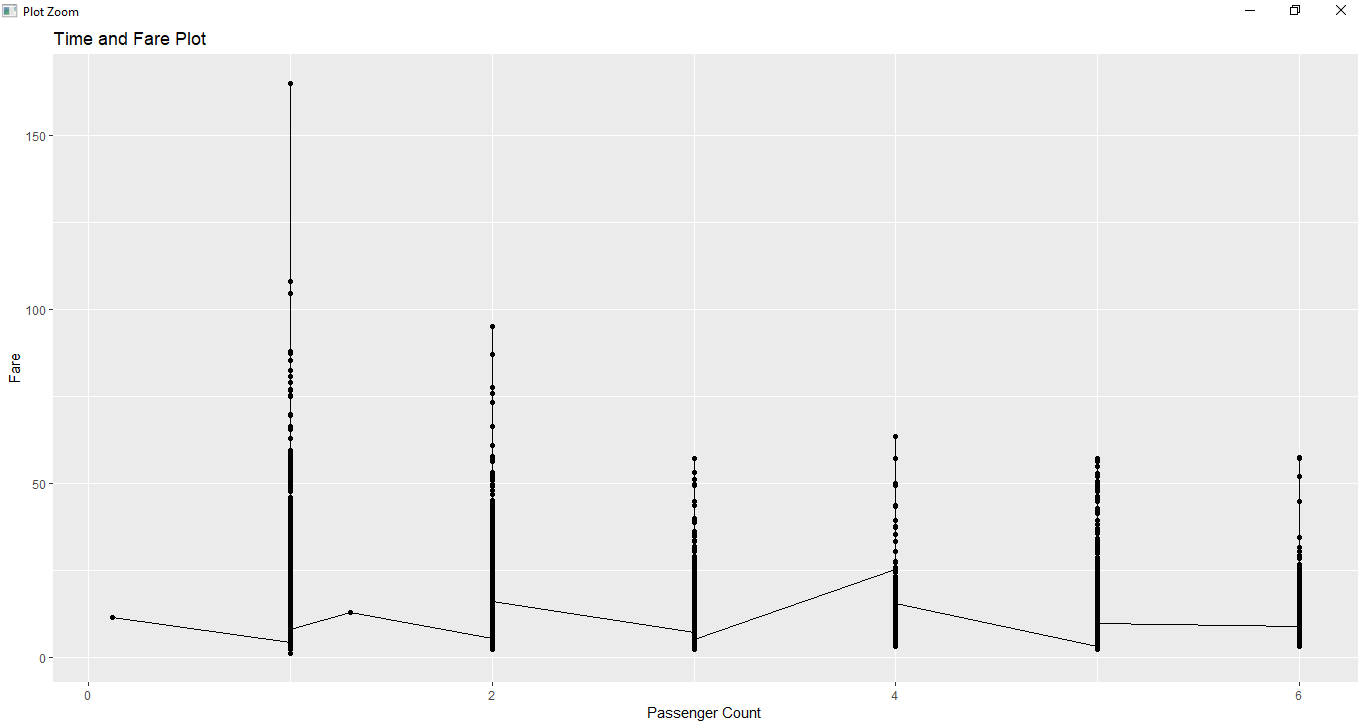
Plotting Distance Vs Fare amount



Hour Vs Fare amount

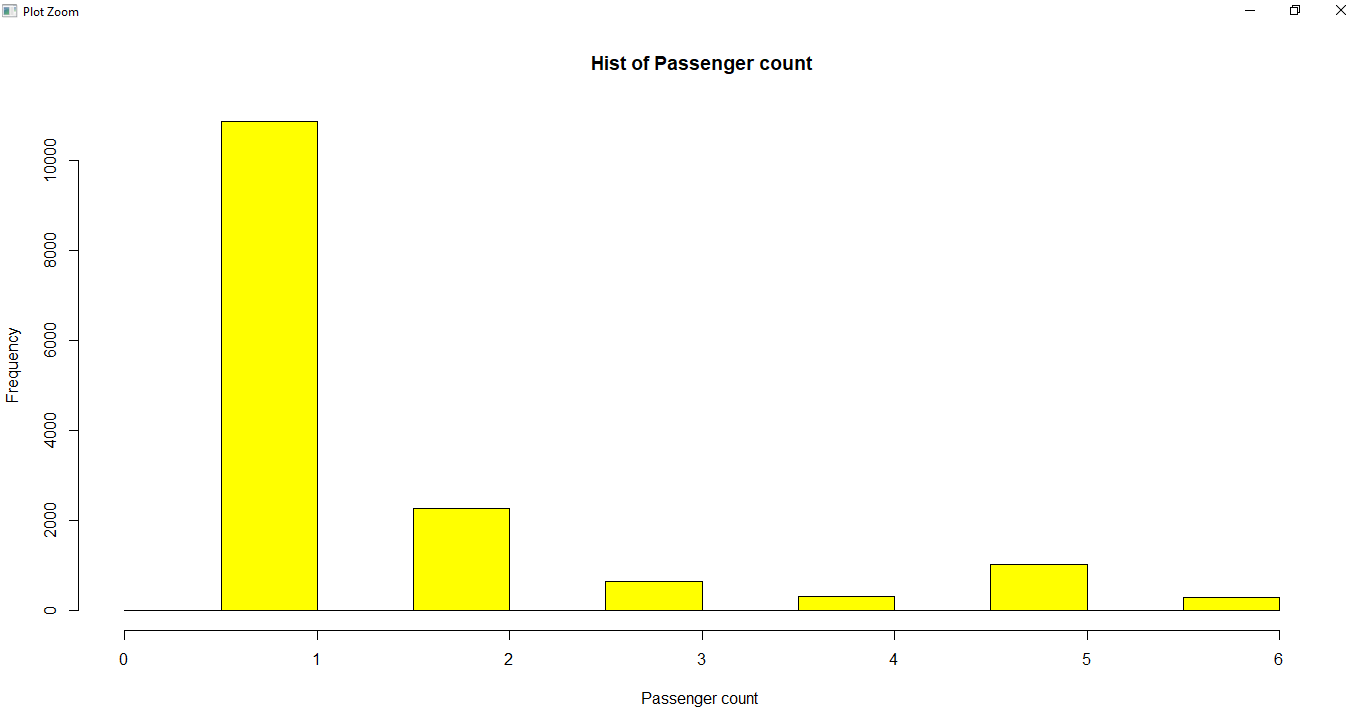


Passenger Count Vs Fare Amount

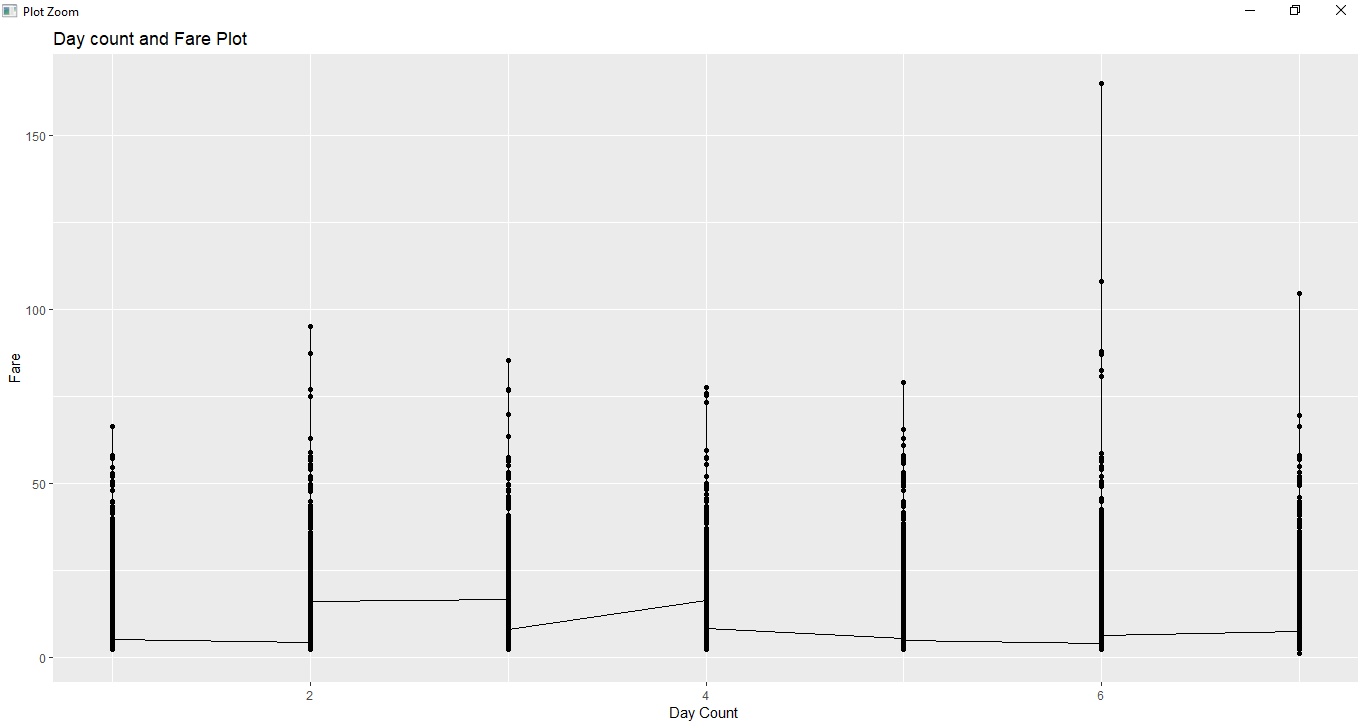


From the Graph it seems passenger count is not affecting the fare.

Frequency of 1 passenger is high.

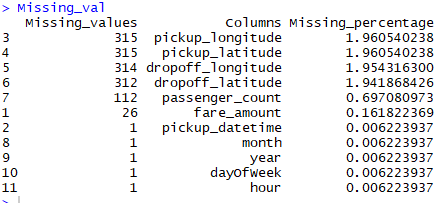


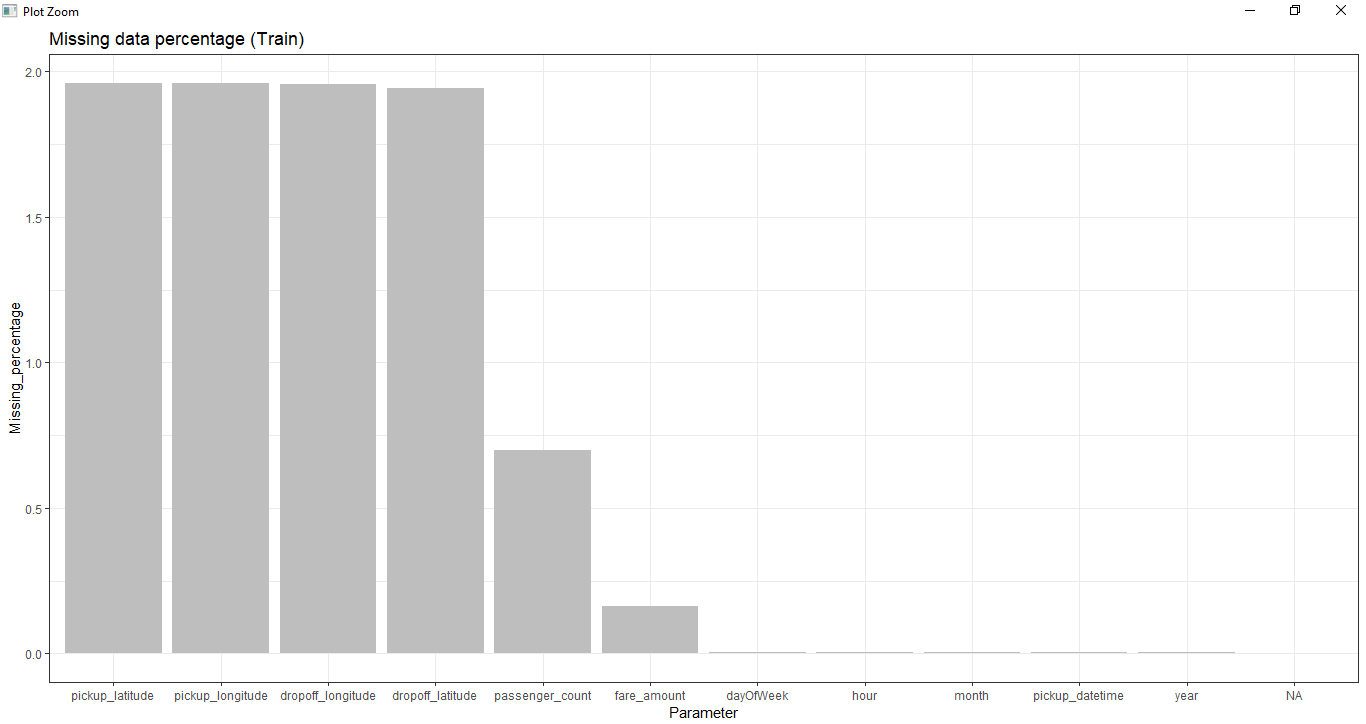
Day Vs Fare Amount



2.2.1 Missing Value Analysis

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a columns has more than 30% of data as missing value either we ignore the entire column or we ignore those observations. In the given data the maximum percentage of missing value is





Missing Values have been ignored in the R coding since distribution of missing values are same accros the different variable .

As a another try Missing values are being imputed using KNN in Python.

2.1.2 Outlier Analysis \ Cleaning

We can clearly observe from Summary in R and Describe Function in Python that Passenger counts of maximum values is very high and Pickup\drop off longitude and latitude has been kept under 90 and 180 as per geographical information. Passenger count is limited to 8 since only cab can accommodate only 8 if consider its SUV. Distance is also Minimized to 100km.

**2.1.3** Feature Engineering

We have converted Pickup \ drop off latitude and longitude as absolute location points and from these variables we have extracted the total distance travelled. From Pick date and Time extracted Year, Month, day, Hours. Here is some graphical representation of the same.A picture containing writing implement, pencil, stationary

Description generated with very high confidence

**2.2** Modeling

**2.2.1** Linear Regression

Regression is a parametric technique used to predict continuous (dependent) variable given a set of independent variables. It is parametric in nature because it makes certain assumptions (discussed next) based on the data set. If the data set follows those assumptions, regression gives incredible results. Otherwise, it struggles to provide convincing accuracy.

Mathematically, regression uses a linear function to approximate (predict) the dependent variable given as:

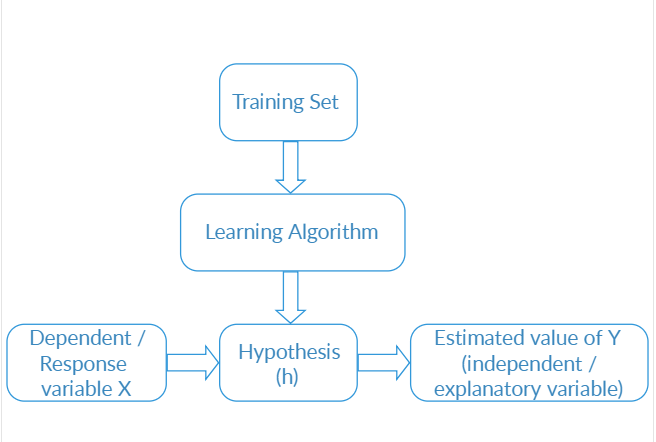
Y = ?o + ?1X + ?

where, Y – Dependent variable  
X – Independent variable  
?o – Intercept  
?1 – Slope  
? – Error

?o and ?1 are known as coefficients. This is the equation of simple linear regression. It’s called ‘linear’ because there is just one independent variable (X) involved. In multiple regression, we have many independent variables (Xs).  If we recall, the equation above is nothing but a line equation (y = mx + c)

**Simple Linear Regression**

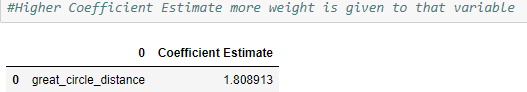
**Predicting Fare amount based on the Distance**



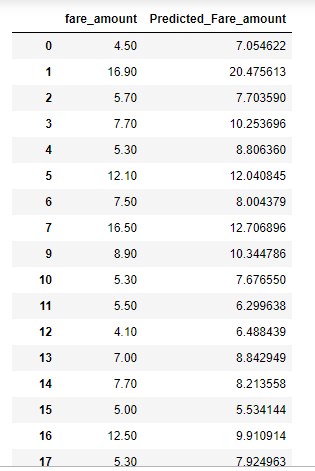
Model Details

|  |  |
| --- | --- |
| Score of the model (R2 Score | 0.610013367 |
| Model Coefficients | 1.80891257 |
| Model Intercept | 5.190060444 |

Contribution of Distance variable to predict Fare amount.



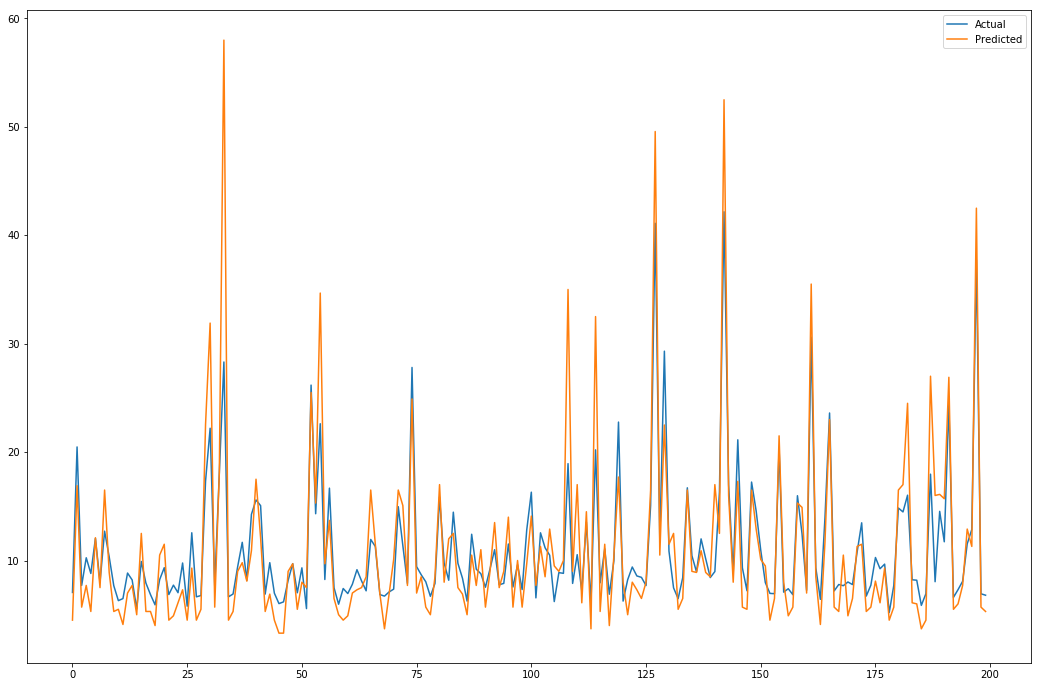
Comparison of Actual Fare amount vs Predicted fare amount.



Error Metrics

|  |  |
| --- | --- |
| *Mean Squared Error* | *33.90217895* |
| *Root Mean squared error* | *5.822557767* |

**Visualization Actual Vs Predicated fare amount.**



A close up of a map

Description generated with very high confidence

Linear Regression Model 2

In this Model we have used Passenger count and absolute Location to find the fare amount

Passenger count  
Abs Longitude  
Abs latitude

dependent Variable X [Fare amount]

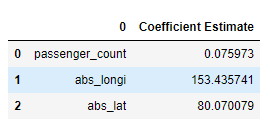
Predicted Output  
 Y

**LM**

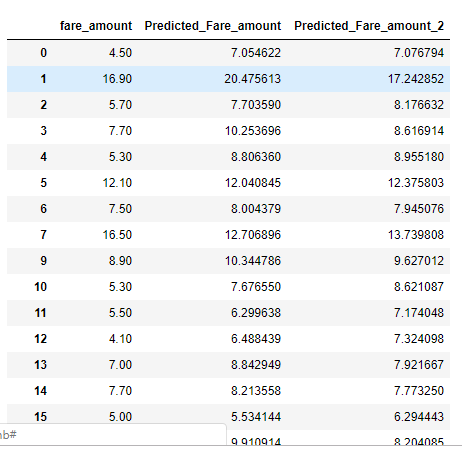
Model Details

|  |  |
| --- | --- |
| Score of the model (R2 Score | 0.599491739 |
| Model Coefficients | array([7.59731353e-02, 1.53435741e+02, 8.00700792e+01]) |
| Model Intercept | 5.8624776 |

Coefficient estimate



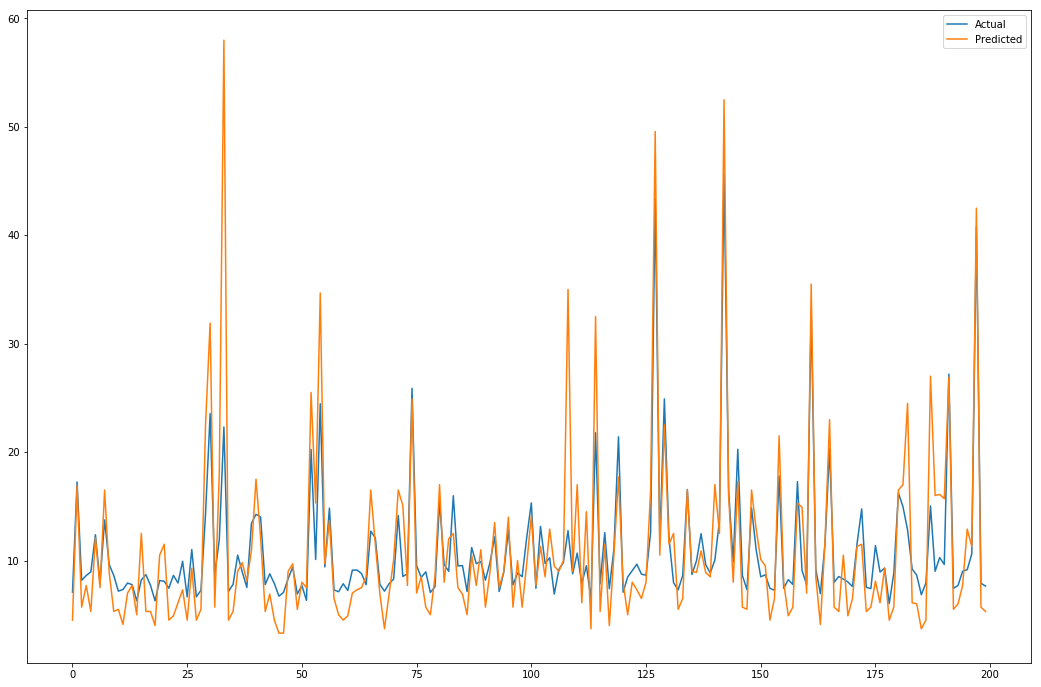
Predicted Fare amount vs Actual

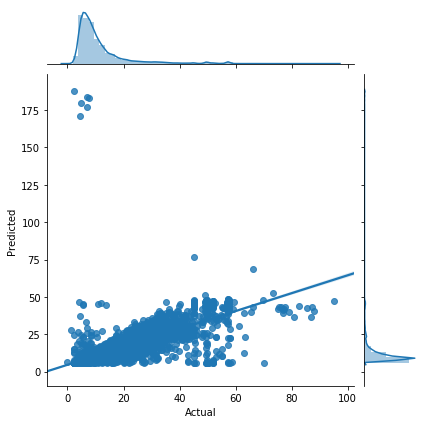


Error Metrics

|  |  |
| --- | --- |
| Mean Squared Error | 34.81684133 |
| Root Mean squared error | 5.900579745 |

Visualization

****

****

# Linear Regression Model 3

Passenger count, Abs Longitude, Abs latitude, year, Month, week day, Hour Distance

dependent Variable X [Fare amount]

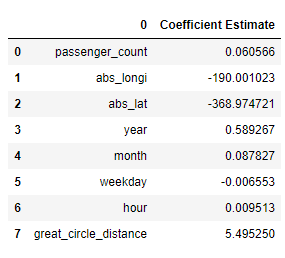
Predicted Output  
 Y

**LM**

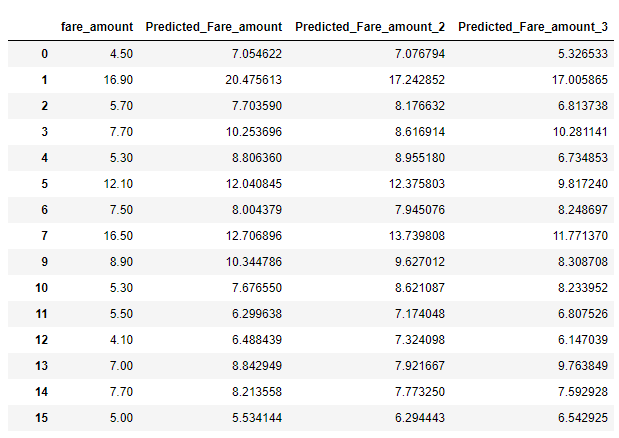
Model Details

|  |  |
| --- | --- |
| Score of the model(R2 Score | 0.67335737 |
| Model Coefficients | array([ 6.05664608e-02, -1.90001023e+02, -3.68974721e+02, 5.89267345e-01,  8.78271733e-02, -6.55347232e-03, 9.51270932e-03, 5.49525025e+00]) |
| Model Intercept | -1181.076022 |

Coefficient estimate



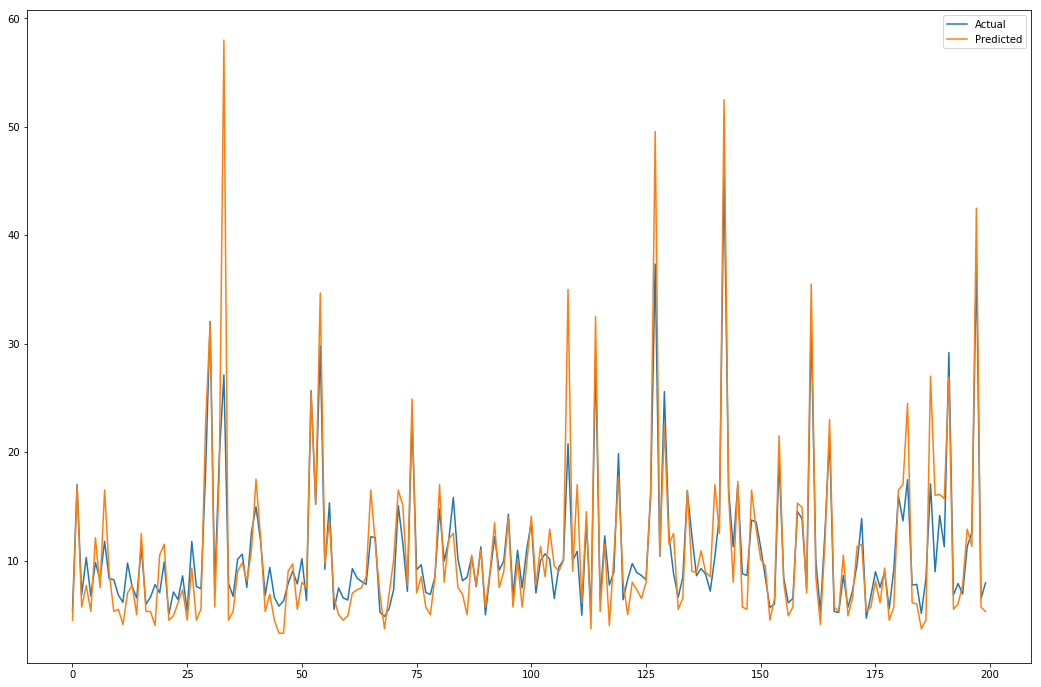
Predicted Vs Actual Fare amount

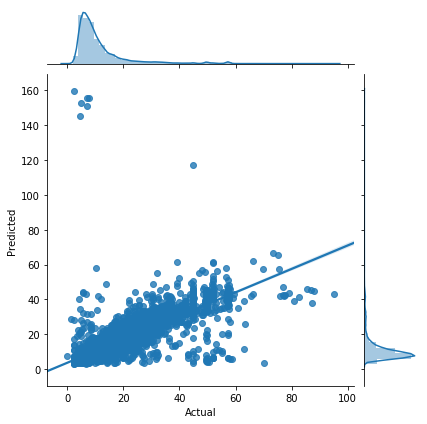


Error Metrics

|  |  |
| --- | --- |
| Mean Squared Error | 28.39558067 |
| Root Mean squared error | 5.328750386 |

Visualization

****

****

# Linear Regression Model Using Ridge Method

Passenger count, Abs Longitude, Abs latitude, year, Month, week day, Hour Distance

dependent Variable X [Fare amount]

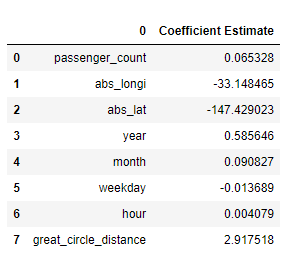
**LM-Ridge**

Predicted Output  
 Y

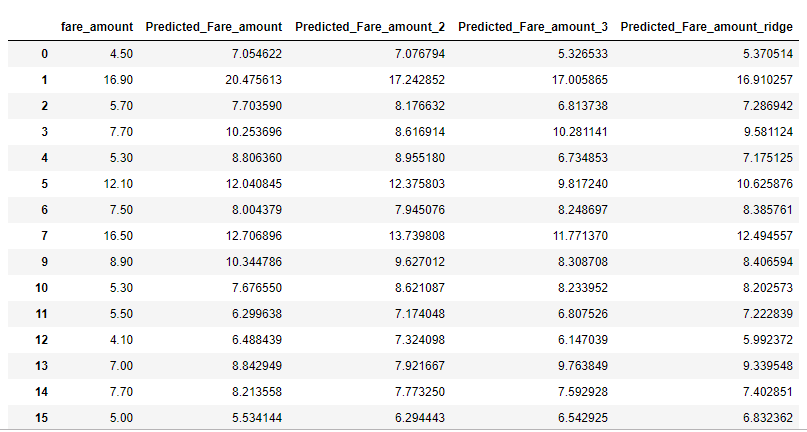
Model Details

|  |  |
| --- | --- |
| Score of the model(R2 Score | 0.659612795 |
| Model Coefficients | array([ 6.53277609e-02, -3.31484653e+01, -1.47429023e+02, 5.85646007e-01,  9.08269852e-02, -1.36890879e-02, 4.07918207e-03, 2.91751751e+00]) |
| Model Intercept | -1173.456782 |

Coefficient Estimate



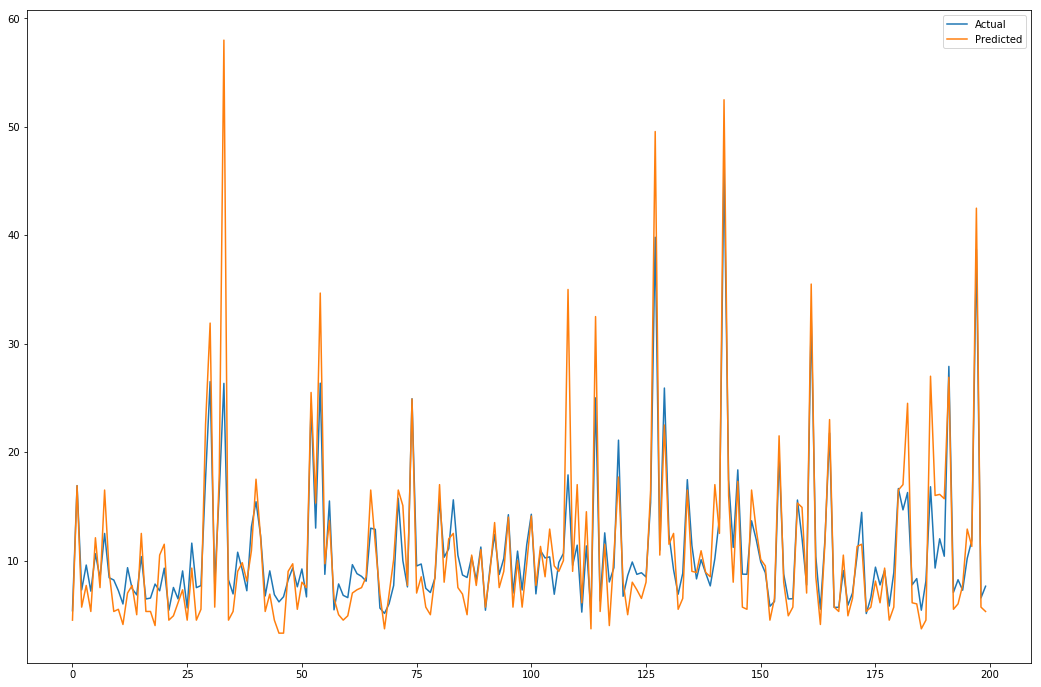
Predicted Vs Actual Fare

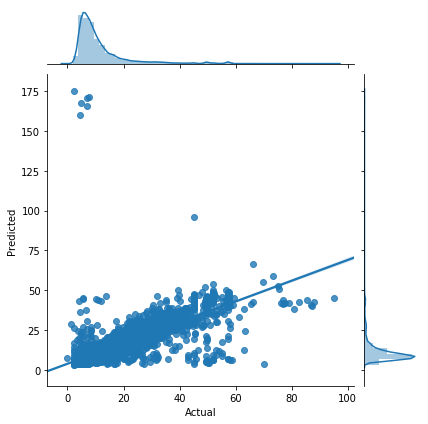


Error Metrics

|  |  |
| --- | --- |
| Mean Squared Error | 29.59041911 |
| Root Mean squared error | 5.439707631 |

Visualization

****

****

**2.2.2 Random Forest**

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data.

Passenger count, Abs Longitude, Abs latitude, year, Month, week day, Hour Distance

dependent Variable X [Fare amount]

**Random Forest**

Predicted Output  
 Y

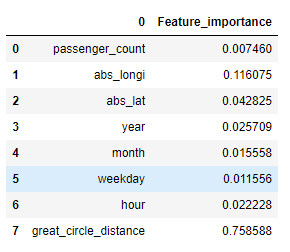
Model Score

|  |  |
| --- | --- |
| Score of the model(R2 Score | 0.964387689 |

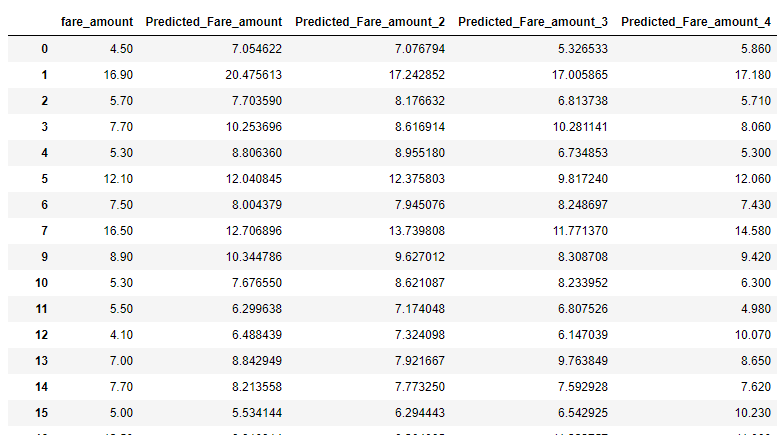
Model Parameters

{'bootstrap': True, 'criterion': 'mse', 'max\_depth': None, 'max\_features': 'auto', 'max\_leaf\_nodes': None, 'min\_impurity\_decrease': 0.0, 'min\_impurity\_split': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'min\_weight\_fraction\_leaf': 0.0, 'n\_estimators': 10, 'n\_jobs': None, 'oob\_score': False, 'random\_state': None, 'verbose': 0, 'warm\_start': False}

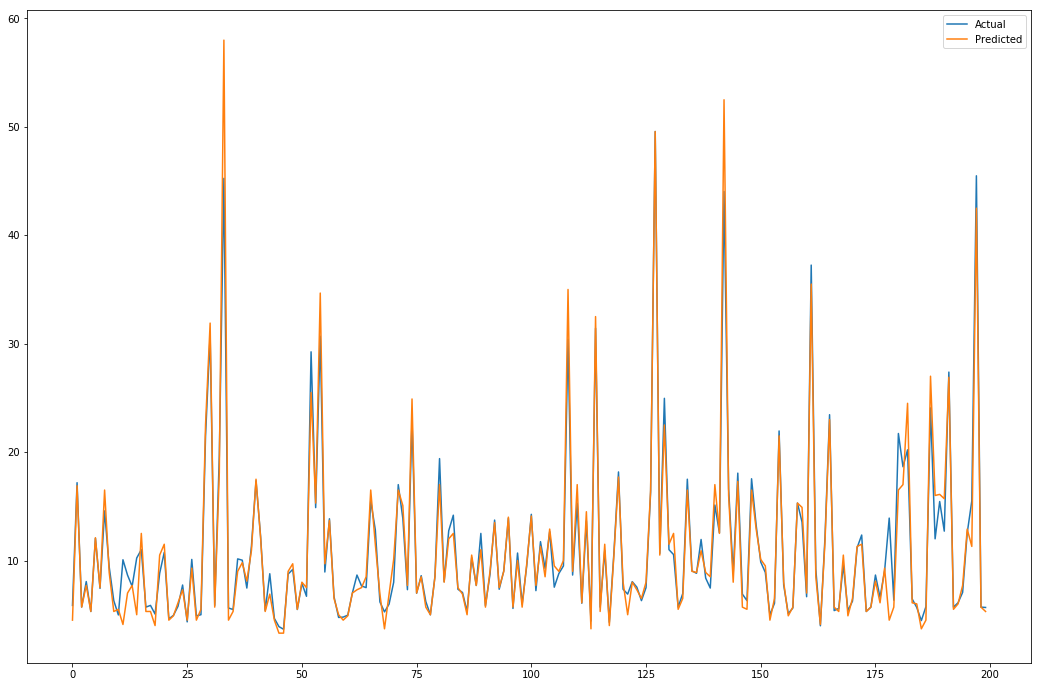
Feature Importance

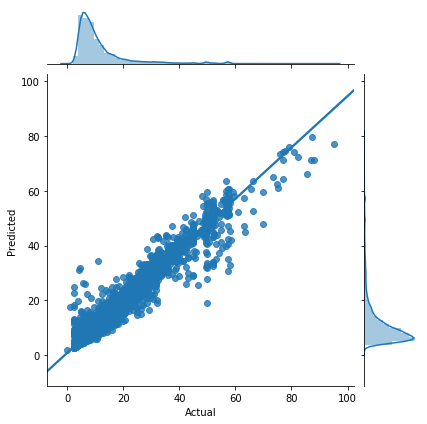


Predicted Vs Actual Amount



Visualizing

****

****

Chapter 3

Conclusion

In this chapter we are going to evaluate our models, select the best model for our dataset and try to get answers of the asked questions.

3.1 Model Evaluation

In the previous chapter we have seen the **Root Mean Square Error** (RMSE) and **R-Squared** Value of different models. **Root Mean Square Error** (RMSE) is the standard deviation of the residuals (prediction **errors**). Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Whereas **R**-**squared** is a relative measure of fit, **RMSE** is an absolute measure of fit. As the square root of a variance, **RMSE** can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of **RMSE** and higher value of **R-Squared Value** indicate better fit.

3.2 Model Selection

From the observation of all **RMSE Value** and **R-Squared** Value we have concluded that **Random Forest Model**  has minimum value of RMSE and it’s **R-Squared** Value is also maximum (i.e. 0.96).

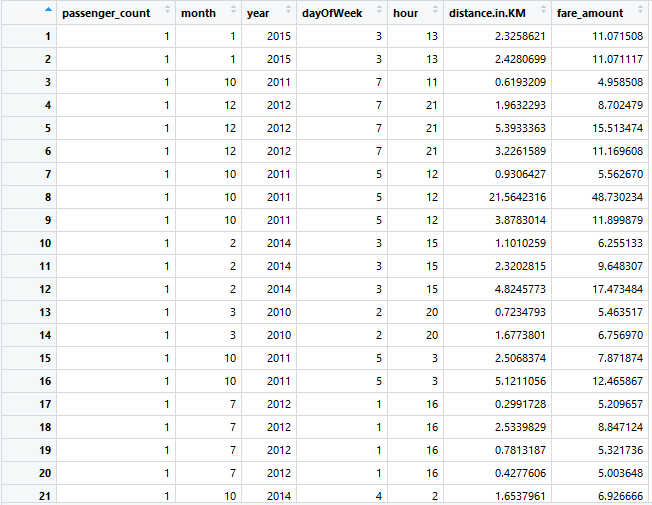
The RMSE value of Testing data and Training does not differs a lot this implies that it is not the case of overfitting.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model 1 Simple Linear Regression | Model 2 Multiple Linear Regression | Model 3 Multiple Linear Regression | Model 4 Linear Regression with Ridge | Model 5 Random Forest |
| Score of the model(R2 Score | 0.610013367 | 0.599491739 | 0.67335737 | 0.659612795 | 0.964387689 |
| Model Coefficients | 1.80891257 | array([7.59731353e-02, 1.53435741e+02, 8.00700792e+01]) | array([ 6.05664608e-02, -1.90001023e+02, -3.68974721e+02, 5.89267345e-01,  8.78271733e-02, -6.55347232e-03, 9.51270932e-03, 5.49525025e+00]) | array([ 6.53277609e-02, -3.31484653e+01, -1.47429023e+02, 5.85646007e-01,  9.08269852e-02, -1.36890879e-02, 4.07918207e-03, 2.91751751e+00]) |  |
| Model Intercept | 5.190060444 | 5.8624776 | -1181.076022 | -1173.456782 |  |

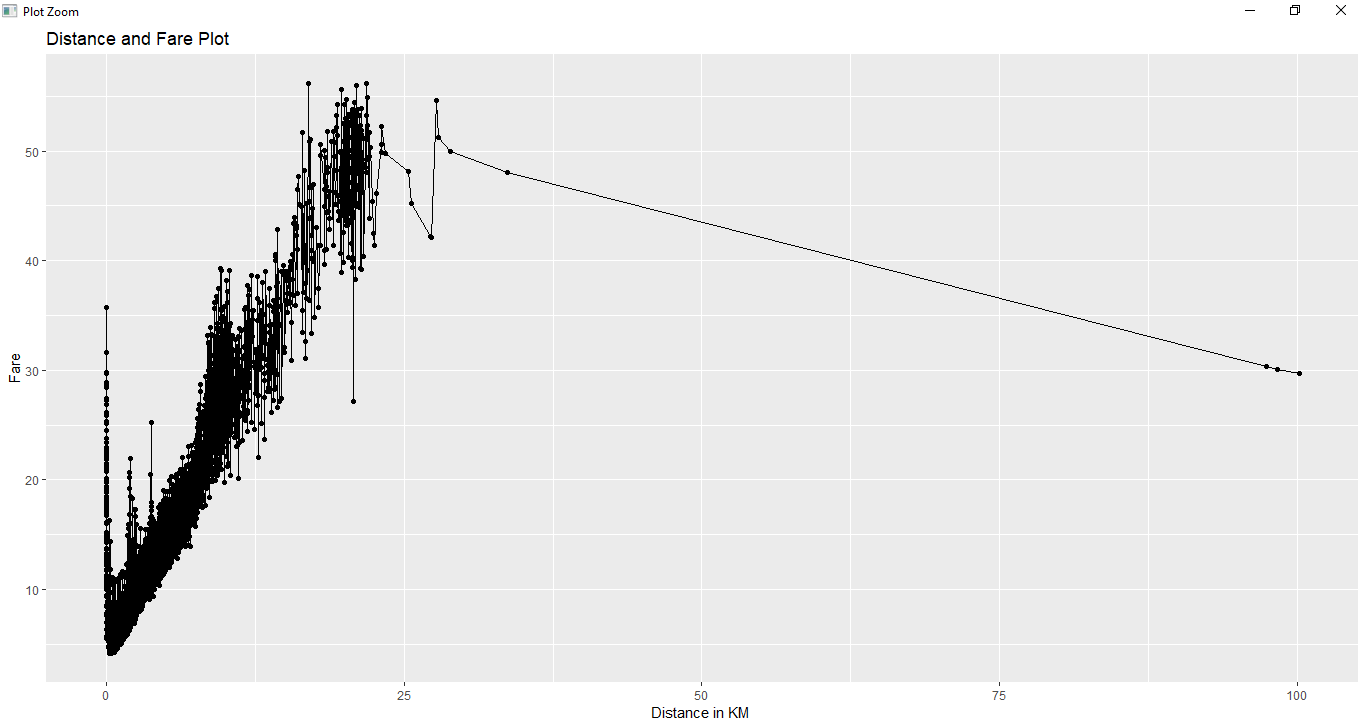
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean Squared Error | 33.90217895 | 34.81684133 | 28.39558067 | 29.59041911 | 3.0958367 |
| Root Mean squared error | 5.822557767 | 5.900579745 | 5.328750386 | 5.439707631 | 1.759499 |

**3.2 Answers of asked questions**

Predicted Fare amount for a Test Cab Data



Predicted Fare amount for Distance



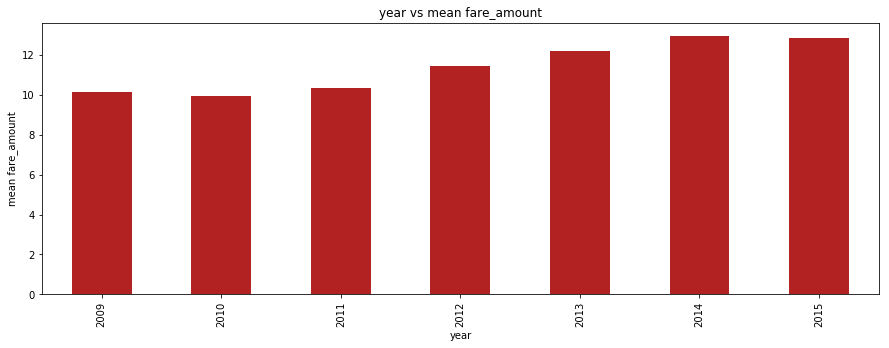
Appendix

Extra Figures

Correlation Plot



Year Vs Fare amount



Distance Frequency

