Project Report on

Cab Fare Prediction

By

PavanKumar. BL

Contents

Chapter 1

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

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.

	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
1	4.5	2009-06-15 17:26:21 UTC	-73.84431	40.72132	-73.84161	40.71228	1
2	16.9	2010-01-05 16:52:16 UTC	-74.01605	40.71130	-73.97927	40.78200	1
3	5.7	2011-08-18 00:35:00 UTC	-73.98274	40.76127	-73.99124	40.75056	2
4	7.7	2012-04-21 04:30:42 UTC	-73.98713	40.73314	-73.99157	40.75809	1
5	5.3	2010-03-09 07:51:00 UTC	-73.96810	40.76801	-73.95665	40.78376	1
6	12.1	2011-01-06 09:50:45 UTC	-74.00096	40.73163	-73.97289	40.75823	1
7	7.5	2012-11-20 20:35:00 UTC	-73.98000	40.75166	-73.97380	40.76484	1
8	16.5	2012-01-04 17:22:00 UTC	-73.95130	40.77414	-73.99009	40.75105	1
9		2012-12-03 13:10:00 UTC	-74.00646	40.72671	-73.99308	40.73163	1
10	8.9	2009-09-02 01:11:00 UTC	-73.98066	40.73387	-73.99154	40.75814	2
11	5.3	2012-04-08 07:30:50 UTC	-73.99634	40.73714	-73.98072	40.73356	1
12	5.5	2012-12-24 11:24:00 UTC	0.00000	0.00000	0.00000	0.00000	3
13	4.1	2009-11-06 01:04:03 UTC	-73.99160	40.74471	-73.98308	40.74468	2
14	7	2013-07-02 19:54:00 UTC	-74.00536	40.72887	-74.00891	40.71091	1
15	7.7	2011-04-05 17:11:05 UTC	-74.00182	40.73755	-73.99806	40.72279	2
16	5	2013-11-23 12:57:00 UTC	0.00000	0.00000	0.00000	0.00000	1
17	12.5	2014-02-19 07:22:00 UTC	-73.98643	40.76047	-73.98899	40.73707	1
18	5.3	2009-07-22 16:08:00 UTC	-73.98106	40.73769	-73.99418	40.72841	1
19	5.3	2010-07-07 14:52:00 UTC	-73.96950	40.78484	-73.95873	40.78336	1
20	4	2014-12-06 20:36:22 UTC	-73.97982	40.75190	-73.97945	40.75548	1
21	10.5	2010-09-07 13:18:00 UTC	-73.98538	40.74786	-73.97838	40.76207	1
22	11.5	2013-02-12 12:15:46 UTC	-73.95795	40.77925	-73.96125	40.75879	1

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.

	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	16043	16067	16067.000000	16067.000000	16067.000000	16067.000000	16012.000000
unique	468	16021	NaN	NaN	NaN	NaN	NaN
top	6.5	2012-01-12 22:54:00 UTC	NaN	NaN	NaN	NaN	NaN
freq	759	2	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	-72.462787	39.914725	-72.462328	39.897906	2.625070
std	NaN	NaN	10.578384	6.826587	10.575062	6.187087	60.844122
min	NaN	NaN	-74.438233	-74.006893	-74.429332	-74.006377	0.000000
25%	NaN	NaN	-73.992156	40.734927	-73.991182	40.734651	1.000000
50%	NaN	NaN	-73.981698	40.752603	-73.980172	40.753567	1.000000
75%	NaN	NaN	-73.966838	40.767381	-73.963643	40.768013	2.000000
max	NaN	NaN	40.766125	401.083332	40.802437	41.366138	5345.000000

	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	16041.000000	16041	16041.000000	16041.000000	16041.000000	16041.000000	15986.000000
unique	NaN	15995	NaN	NaN	NaN	NaN	NaN
top	NaN	2012-05-23 14:22:00 UTC	NaN	NaN	NaN	NaN	NaN
freq	NaN	2	NaN	NaN	NaN	NaN	NaN
mean	15.015735	NaN	-72.469554	39.895976	-72.469115	39.901595	2.623272
std	430.474353	NaN	10.555823	6.192372	10.552491	6.175961	60.892140
min	-3.000000	NaN	-74.438233	-74.006893	-74.429332	-74.006377	0.000000
25%	6.000000	NaN	-73.992157	40.734935	-73.991182	40.734663	1.000000
50%	8.500000	NaN	-73.981709	40.752597	-73.980185	40.753564	1.000000
75%	12.500000	NaN	-73.966843	40.767352	-73.963647	40.768004	2.000000
max	54343.000000	NaN	40.766125	41.366138	40.802437	41.366138	5345.000000

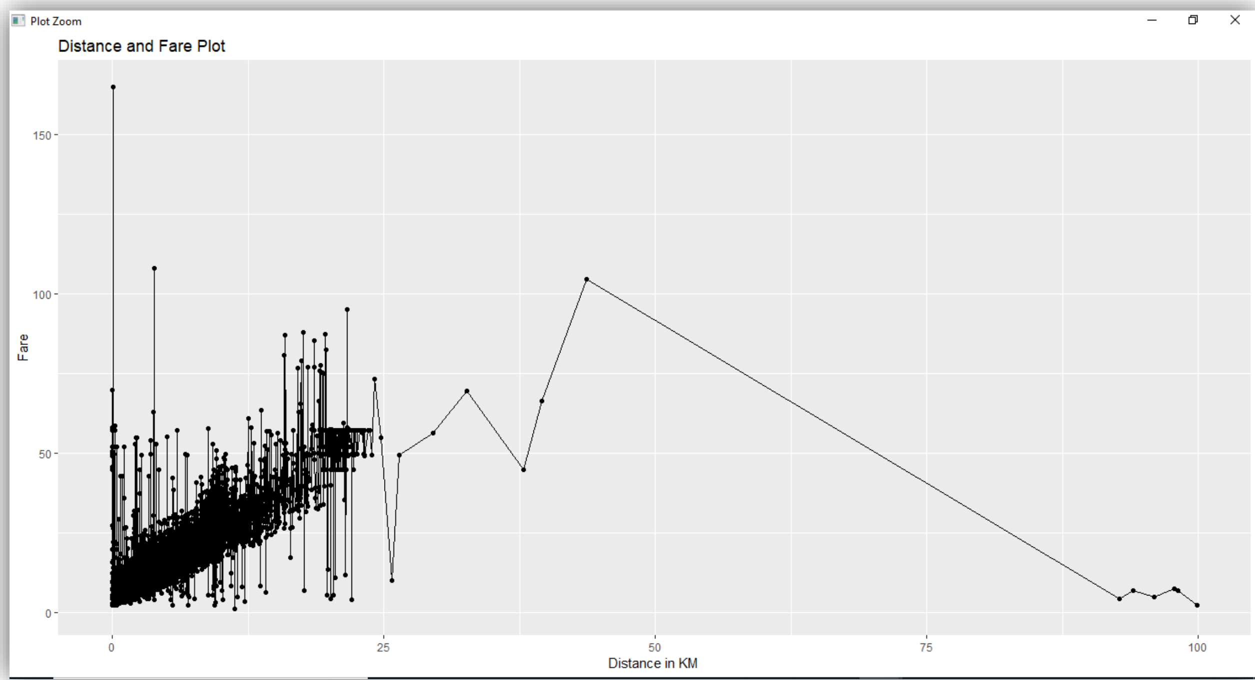
From above details we can confirm that

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

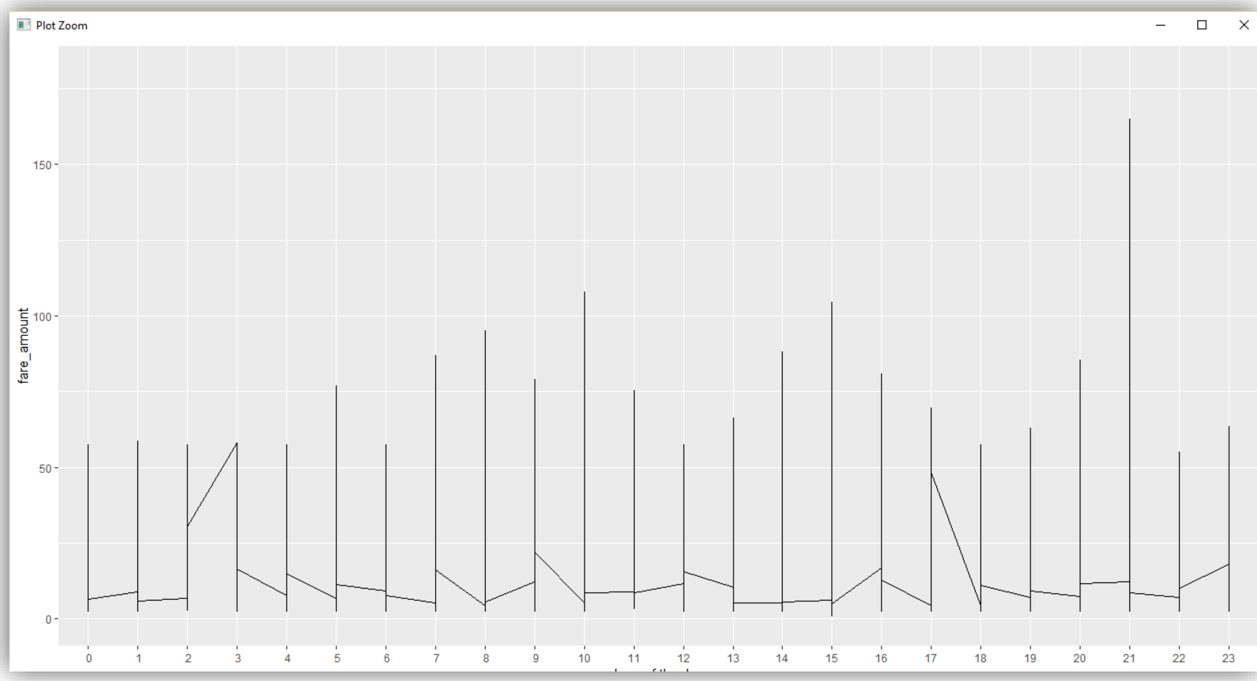
Below is the Image that explains how our data looks like after EDA

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	year	month	weekday	
count	16008.000000	15699.000000	15699.000000	15700.000000	15702.000000	15897.000000	16009.000000	16009.000000	16009.000000	16009.
mean	11.272607	-73.911639	40.689851	-73.906482	40.687787	1.649772	2011.730652	6.261041	3.032981	13.
std	9.379828	2.655828	2.610141	2.707465	2.628960	1.266042	1.863746	3.448034	1.968844	6.
min	0.010000	-74.438233	-74.006893	-74.227047	-74.006377	0.120000	2009.000000	1.000000	0.000000	0.
25%	6.000000	-73.992385	40.736570	-73.991373	40.736287	1.000000	2010.000000	3.000000	1.000000	9.
50%	8.500000	-73.982043	40.753300	-73.980571	40.754230	1.000000	2012.000000	6.000000	3.000000	14.
75%	12.500000	-73.968076	40.767799	-73.965370	40.768309	2.000000	2013.000000	9.000000	5.000000	19.
max	96.000000	40.766125	41.366138	40.802437	41.366138	6.000000	2015.000000	12.000000	6.000000	23.

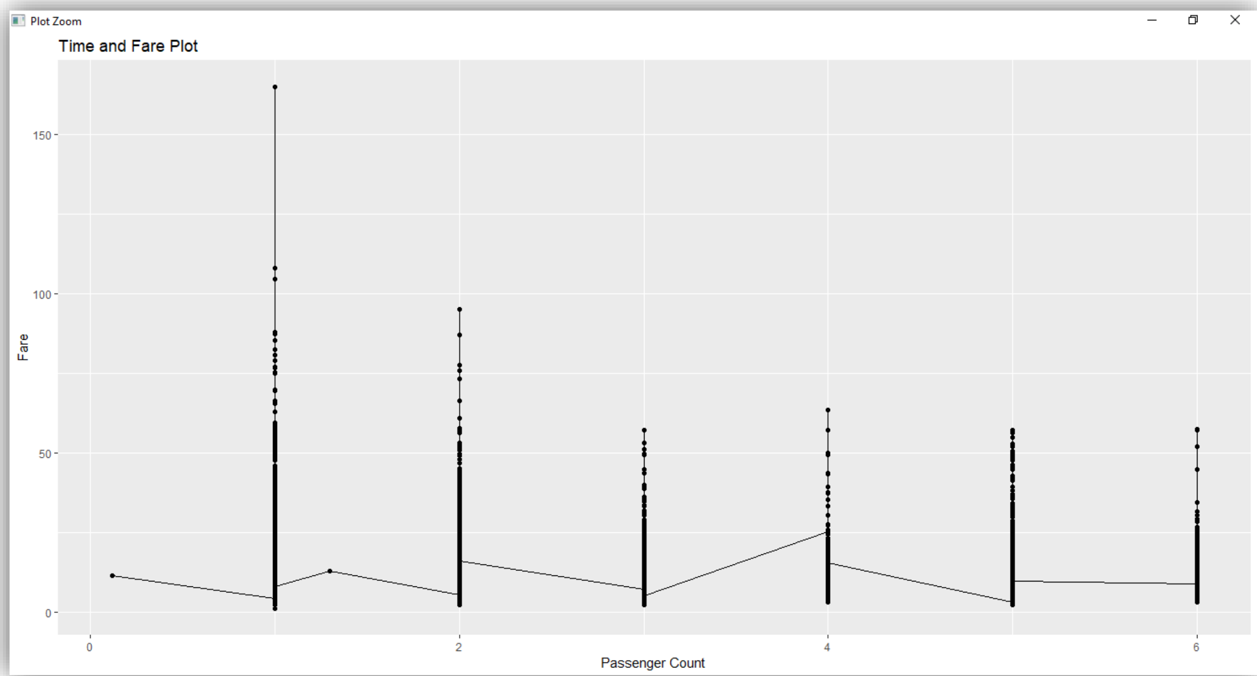
Plotting Distance Vs Fare amount



Hour Vs Fare amount

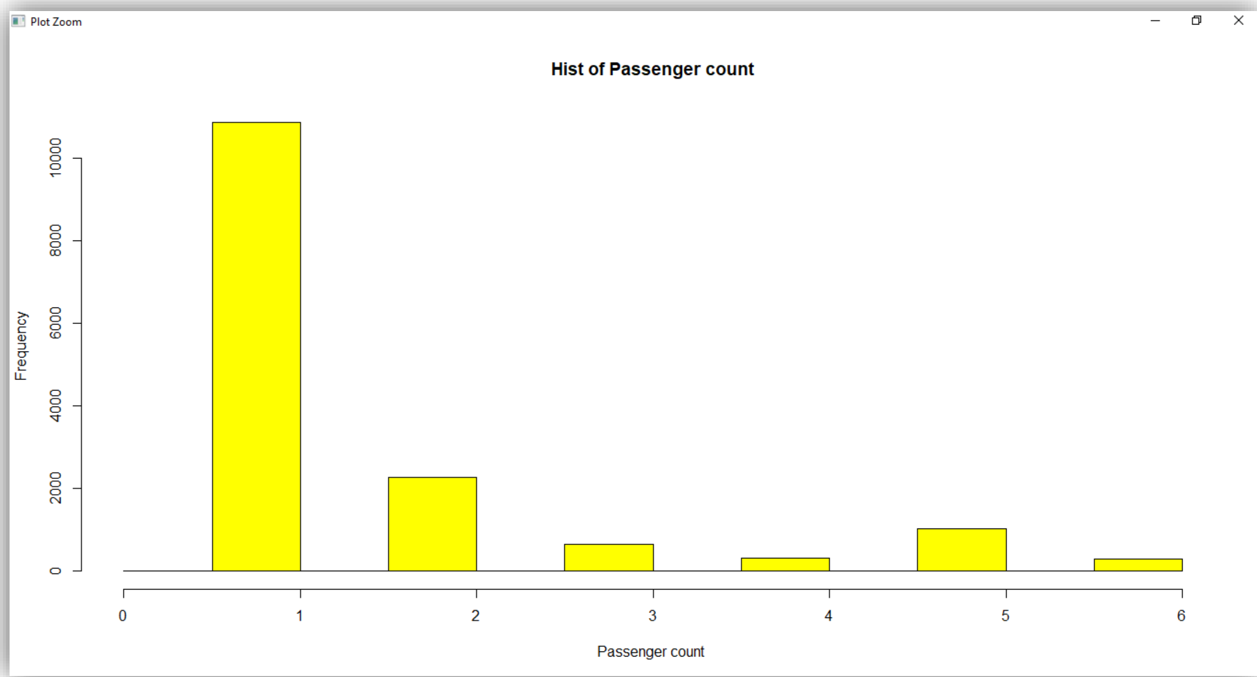


Passenger Count Vs Fare Amount

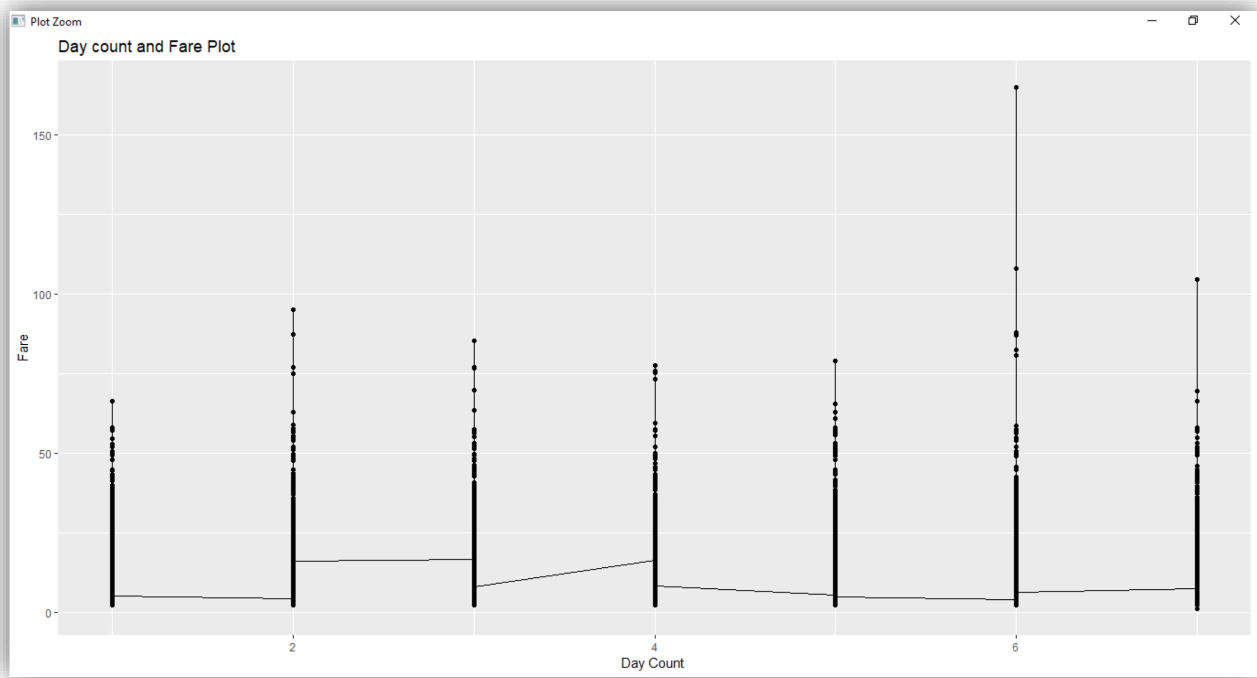


From the Graph it seems passenger count is not affecting the fare.

Frequency of 1 passenger is high.



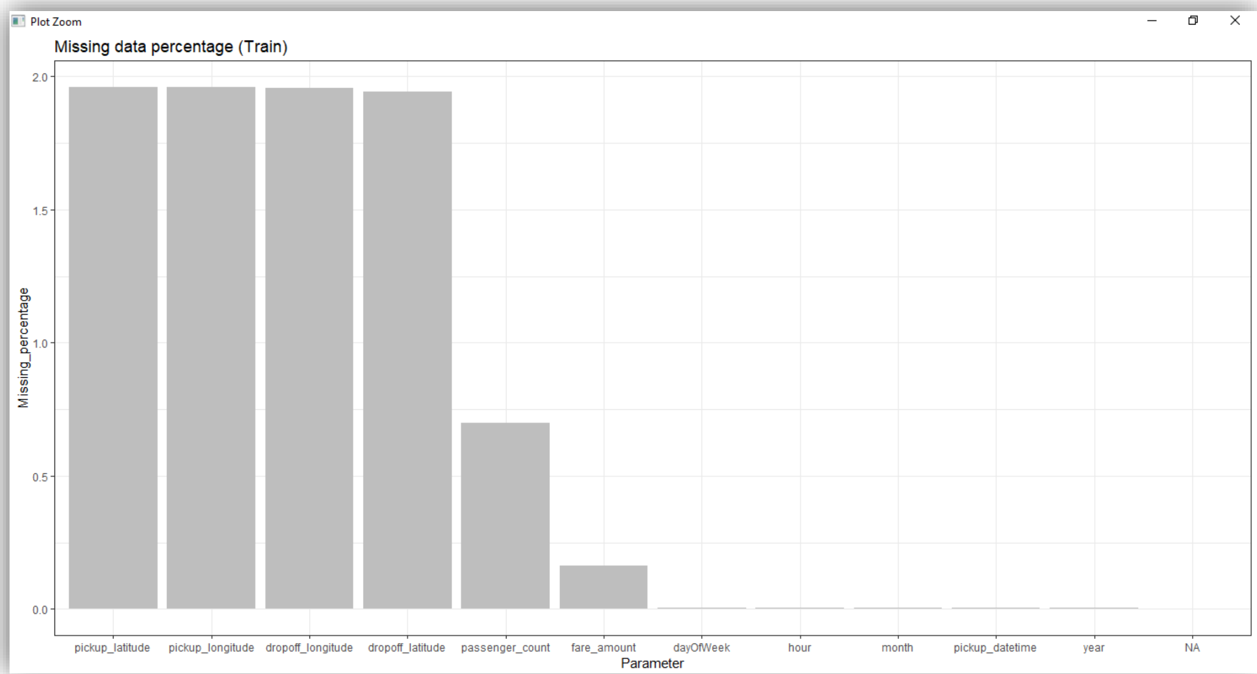
Day Vs Fare Amount



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_val
   Missing_values           Columns Missing_percentage
3             315  pickup_longitude        1.960540238
4             315   pickup_latitude        1.960540238
5             314 dropoff_longitude        1.954316300
6             312  dropoff_latitude        1.941868426
7             112   passenger_count        0.697080973
1              26       fare_amount        0.161822369
2               1   pickup_datetime        0.006223937
8               1             month        0.006223937
9               1              year        0.006223937
10              1         dayOfWeek        0.006223937
11              1              hour        0.006223937
```



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

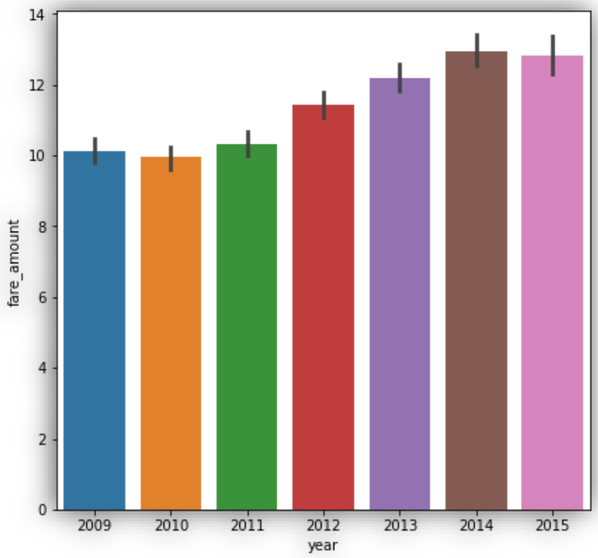
As a another try Missing values are being imputed using KNN in Python.

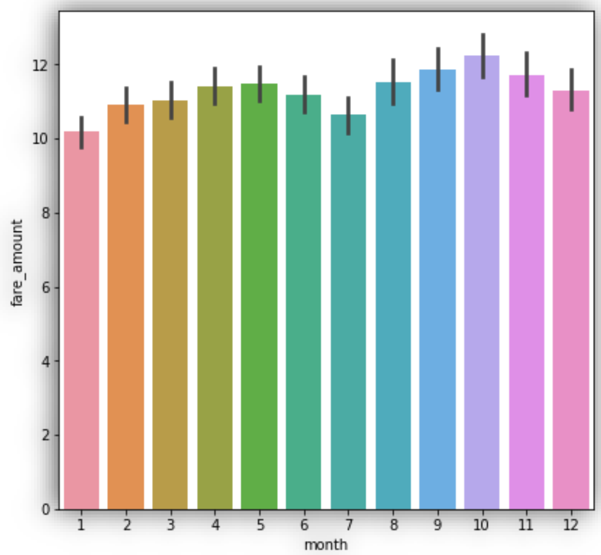
2.1.2 Outlier_Analysis \ Cleaning

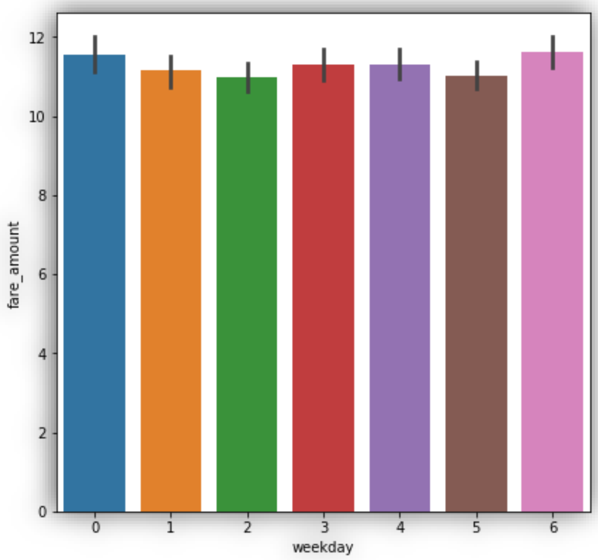
We can clearly observe from Summary in R and Describe Function in Python that Passenger counts of maximum values is very high and Pickup\drop off longitude and latitude has been kept under 90 and 180 as per geographical information. Passenger count is limited to 8 since only cab can accommodate only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

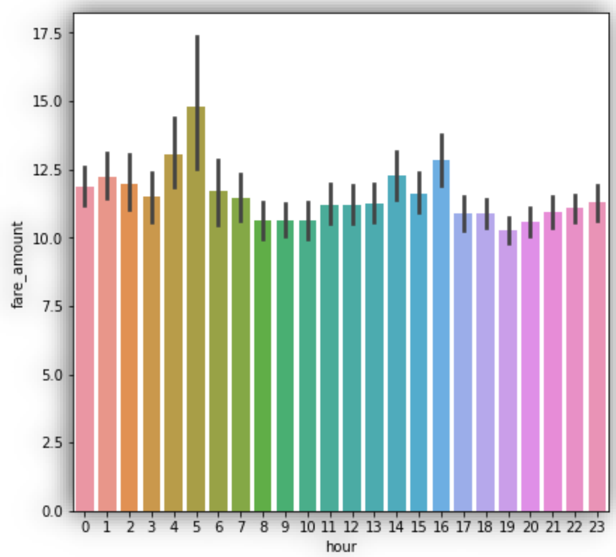
2.1.3 Feature Engineering

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.









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:

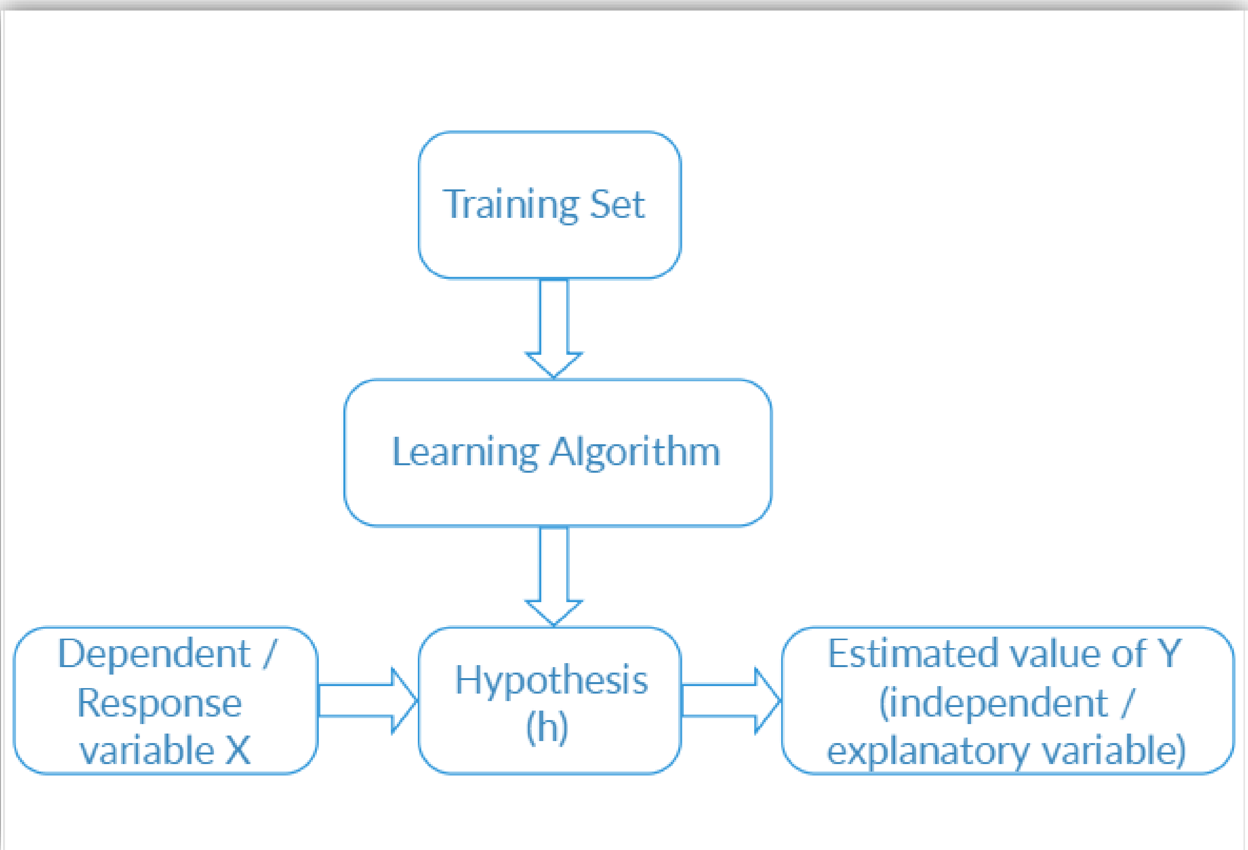
Y = ?o + ?1X + ?

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.

```
#Higher Coefficient Estimate more weight is given to that variable
```

	0	Coefficient Estimate
0	great_circle_distance	1.808913

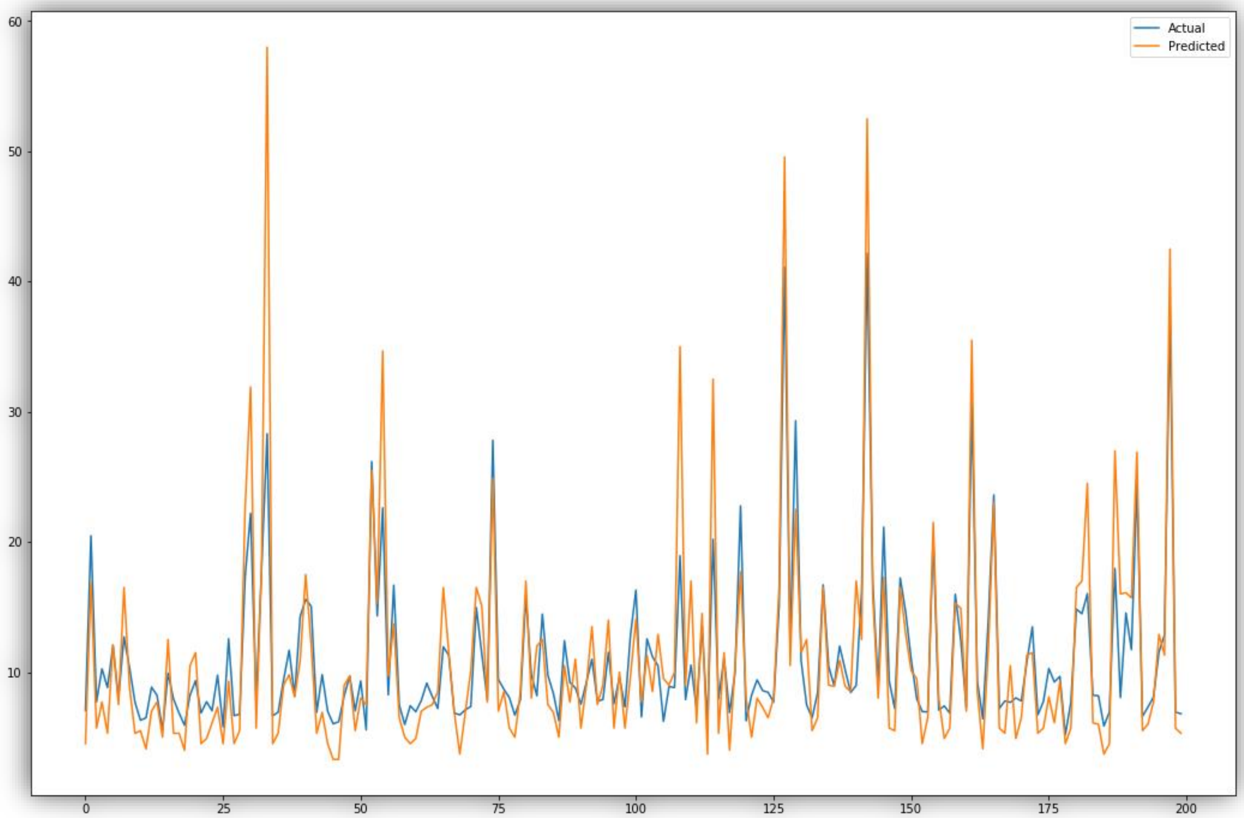
Comparison of Actual Fare amount vs Predicted fare amount.

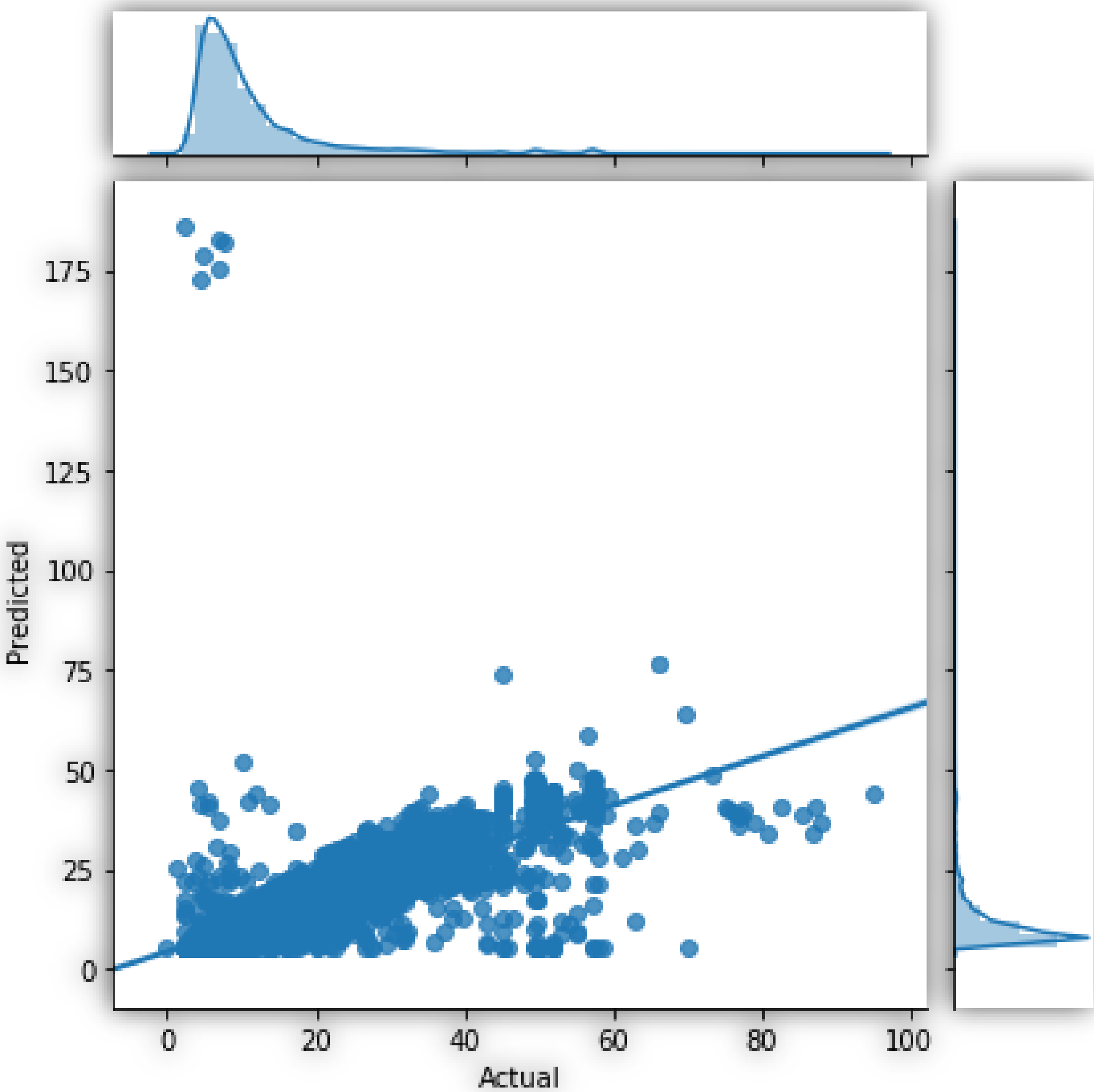
	fare_amount	Predicted_Fare_amount
0	4.50	7.054622
1	16.90	20.475613
2	5.70	7.703590
3	7.70	10.253696
4	5.30	8.806360
5	12.10	12.040845
6	7.50	8.004379
7	16.50	12.706896
9	8.90	10.344786
10	5.30	7.676550
11	5.50	6.299638
12	4.10	6.488439
13	7.00	8.842949
14	7.70	8.213558
15	5.00	5.534144
16	12.50	9.910914
17	5.30	7.924963

Error Metrics

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

Visualization Actual Vs Predicated fare amount.

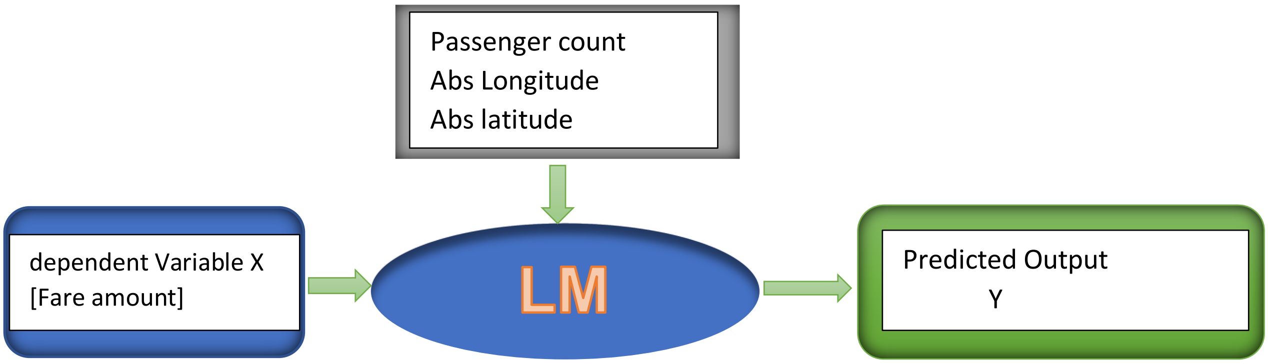




175
150
125
100
75
50
25
0
Predicted
0
20
40
60
80
100
Actual

Linear Regression Model 2

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



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

	0	Coefficient Estimate
0	passenger_count	0.075973
1	abs_longi	153.435741
2	abs_lat	80.070079

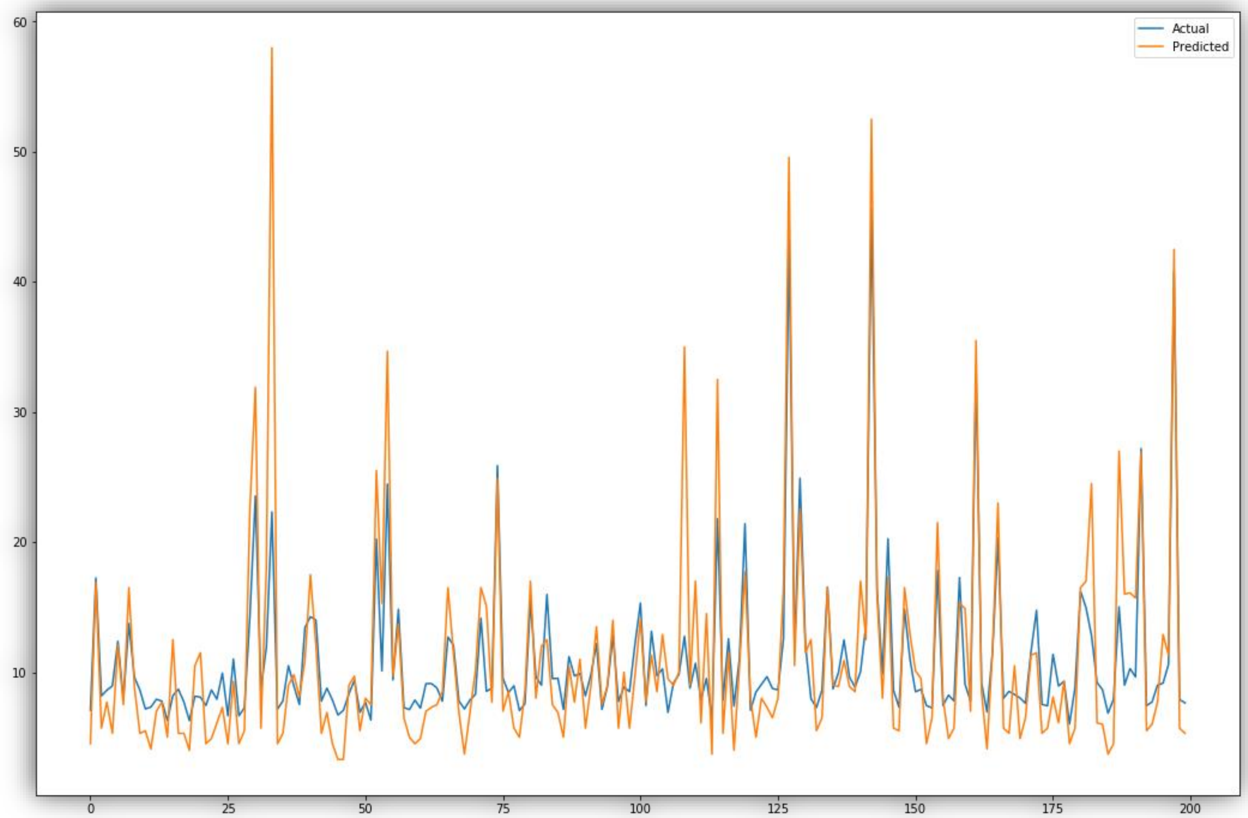
Predicted Fare amount vs Actual

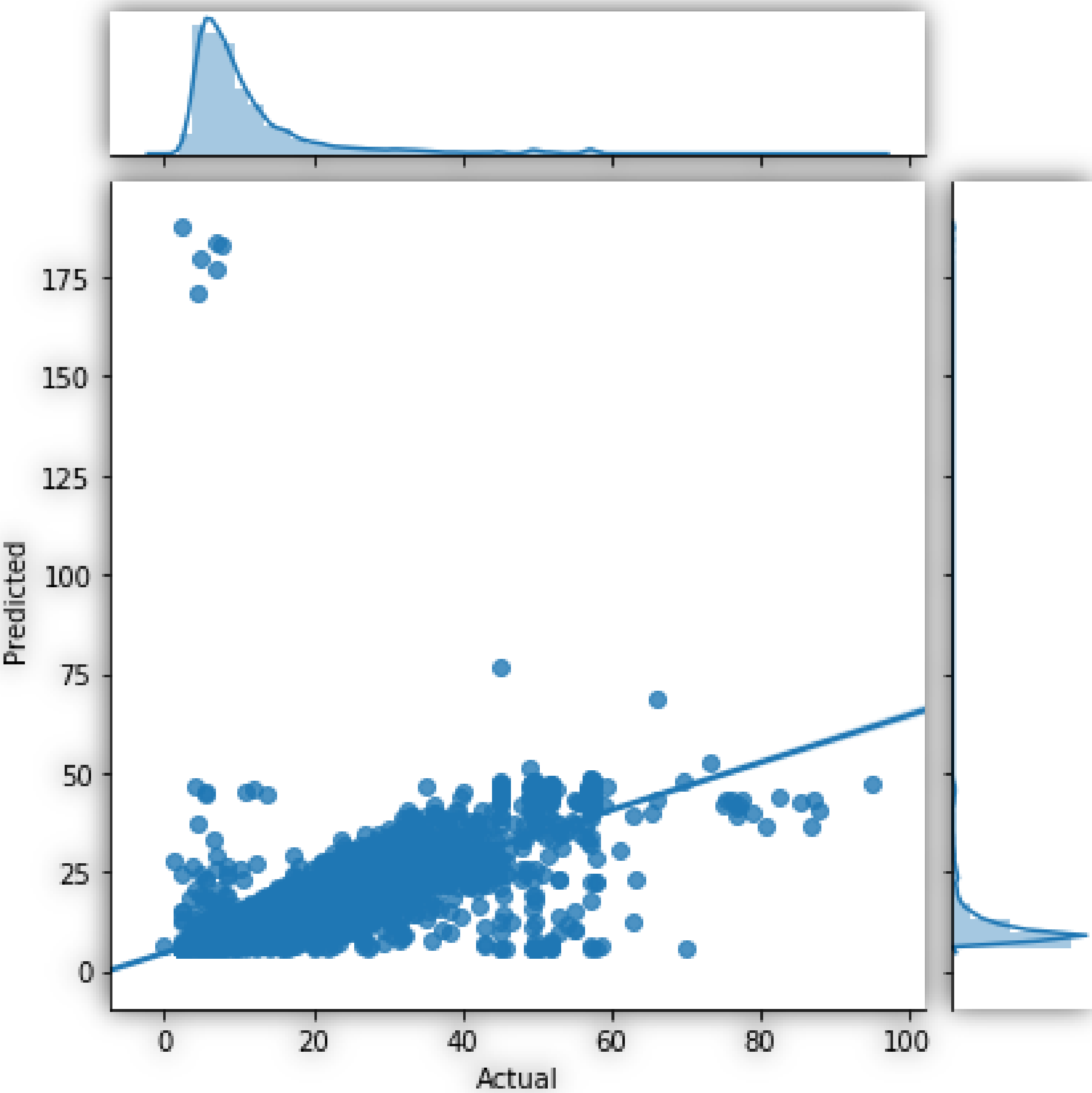
	fare_amount	Predicted_Fare_amount	Predicted_Fare_amount_2
0	4.50	7.054622	7.076794
1	16.90	20.475613	17.242852
2	5.70	7.703590	8.176632
3	7.70	10.253696	8.616914
4	5.30	8.806360	8.955180
5	12.10	12.040845	12.375803
6	7.50	8.004379	7.945076
7	16.50	12.706896	13.739808
9	8.90	10.344786	9.627012
10	5.30	7.676550	8.621087
11	5.50	6.299638	7.174048
12	4.10	6.488439	7.324098
13	7.00	8.842949	7.921667
14	7.70	8.213558	7.773250
15	5.00	5.534144	6.294443
		9.910914	8.204085

Error Metrics

Mean Squared Error	**34.81684133**
Root Mean squared error	**5.900579745**

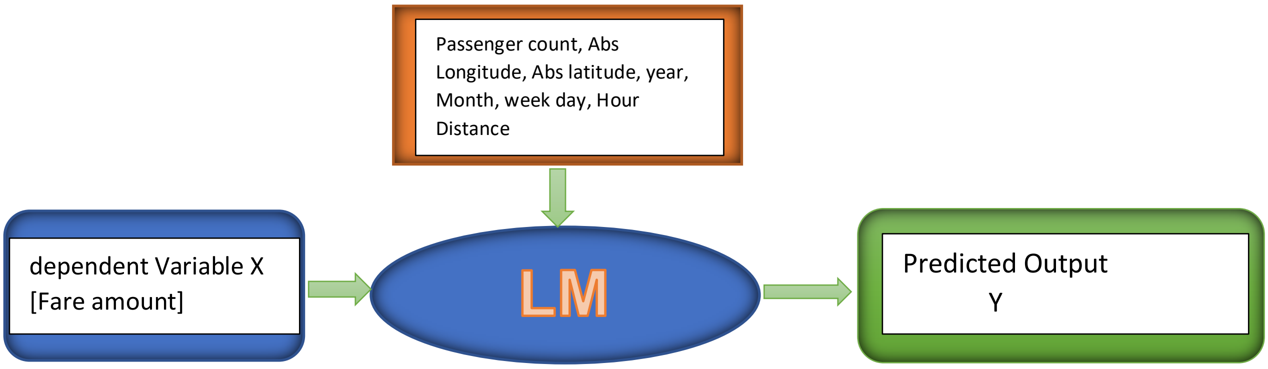
Visualization





175
150
125
100
75
50
25
0
Predicted
0
20
40
60
80
100
Actual

Linear Regression Model 3



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

	0	Coefficient Estimate
0	passenger_count	0.060566
1	abs_longi	-190.001023
2	abs_lat	-368.974721
3	year	0.589267
4	month	0.087827
5	weekday	-0.006553
6	hour	0.009513
7	great_circle_distance	5.495250

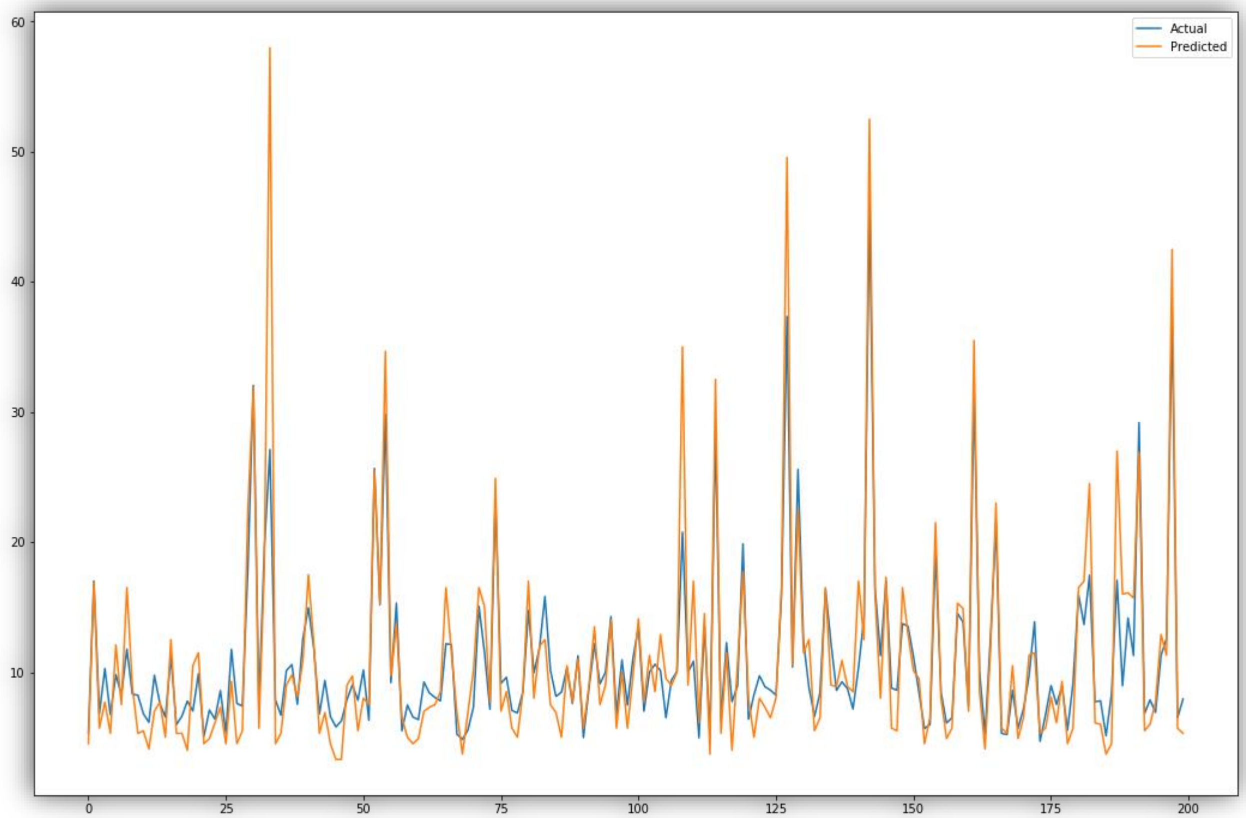
Predicted Vs Actual Fare amount

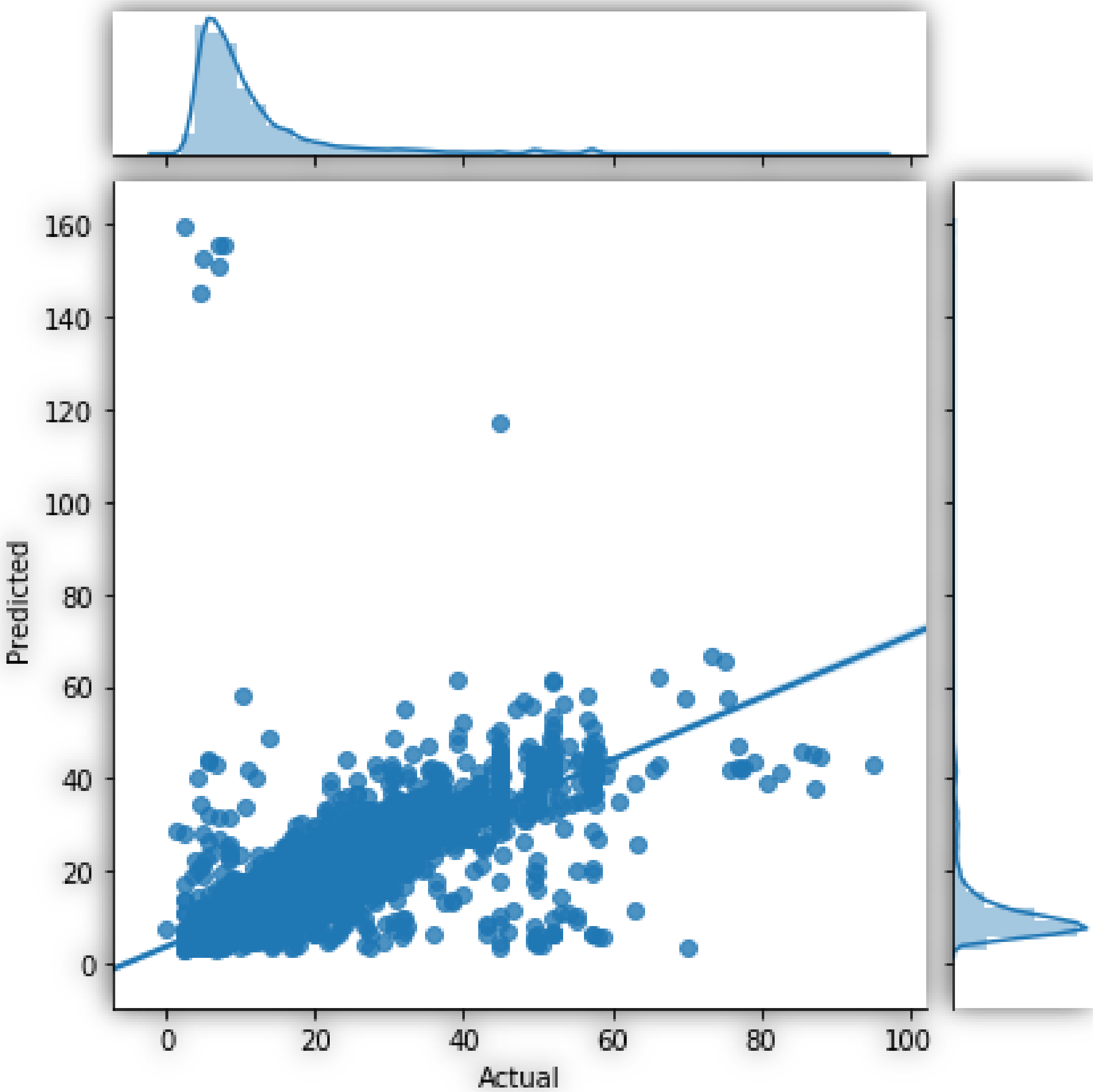
	fare_amount	Predicted_Fare_amount	Predicted_Fare_amount_2	Predicted_Fare_amount_3
0	4.50	7.054622	7.076794	5.326533
1	16.90	20.475613	17.242852	17.005865
2	5.70	7.703590	8.176632	6.813738
3	7.70	10.253696	8.616914	10.281141
4	5.30	8.806360	8.955180	6.734853
5	12.10	12.040845	12.375803	9.817240
6	7.50	8.004379	7.945076	8.248697
7	16.50	12.706896	13.739808	11.771370
9	8.90	10.344786	9.627012	8.308708
10	5.30	7.676550	8.621087	8.233952
11	5.50	6.299638	7.174048	6.807526
12	4.10	6.488439	7.324098	6.147039
13	7.00	8.842949	7.921667	9.763849
14	7.70	8.213558	7.773250	7.592928
15	5.00	5.534144	6.294443	6.542925

Error Metrics

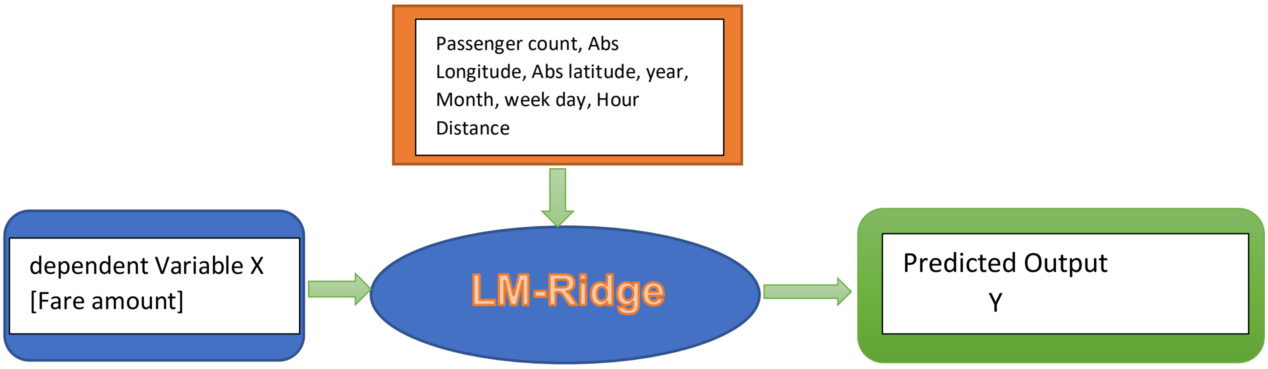
Mean Squared Error	**28.39558067**
Root Mean squared error	**5.328750386**

Visualization




160
140
120
100
80
60
40
20
0
Predicted
0
20
40
60
80
100
Actual

Linear Regression Model Using Ridge Method



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

	0	Coefficient Estimate
0	passenger_count	0.065328
1	abs_longi	-33.148465
2	abs_lat	-147.429023
3	year	0.585646
4	month	0.090827
5	weekday	-0.013689
6	hour	0.004079
7	great_circle_distance	2.917518

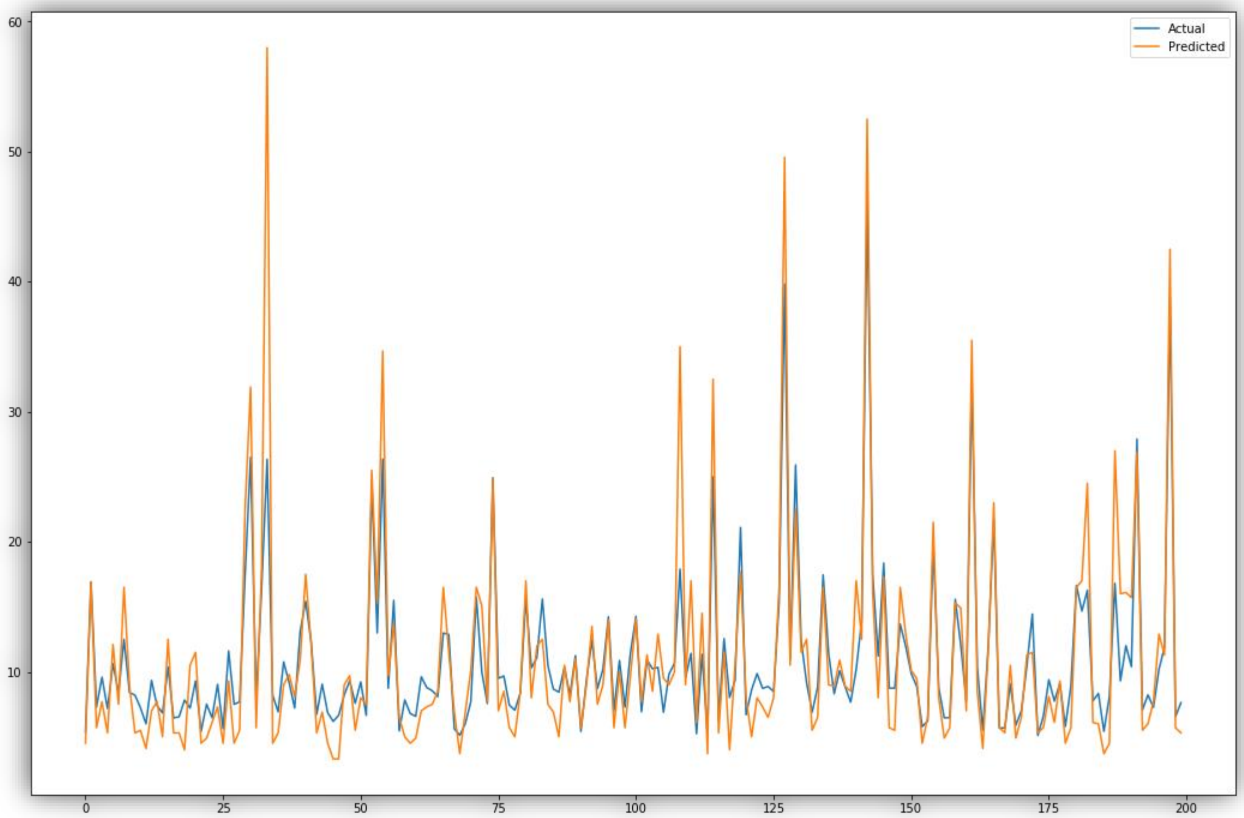
Predicted Vs Actual Fare

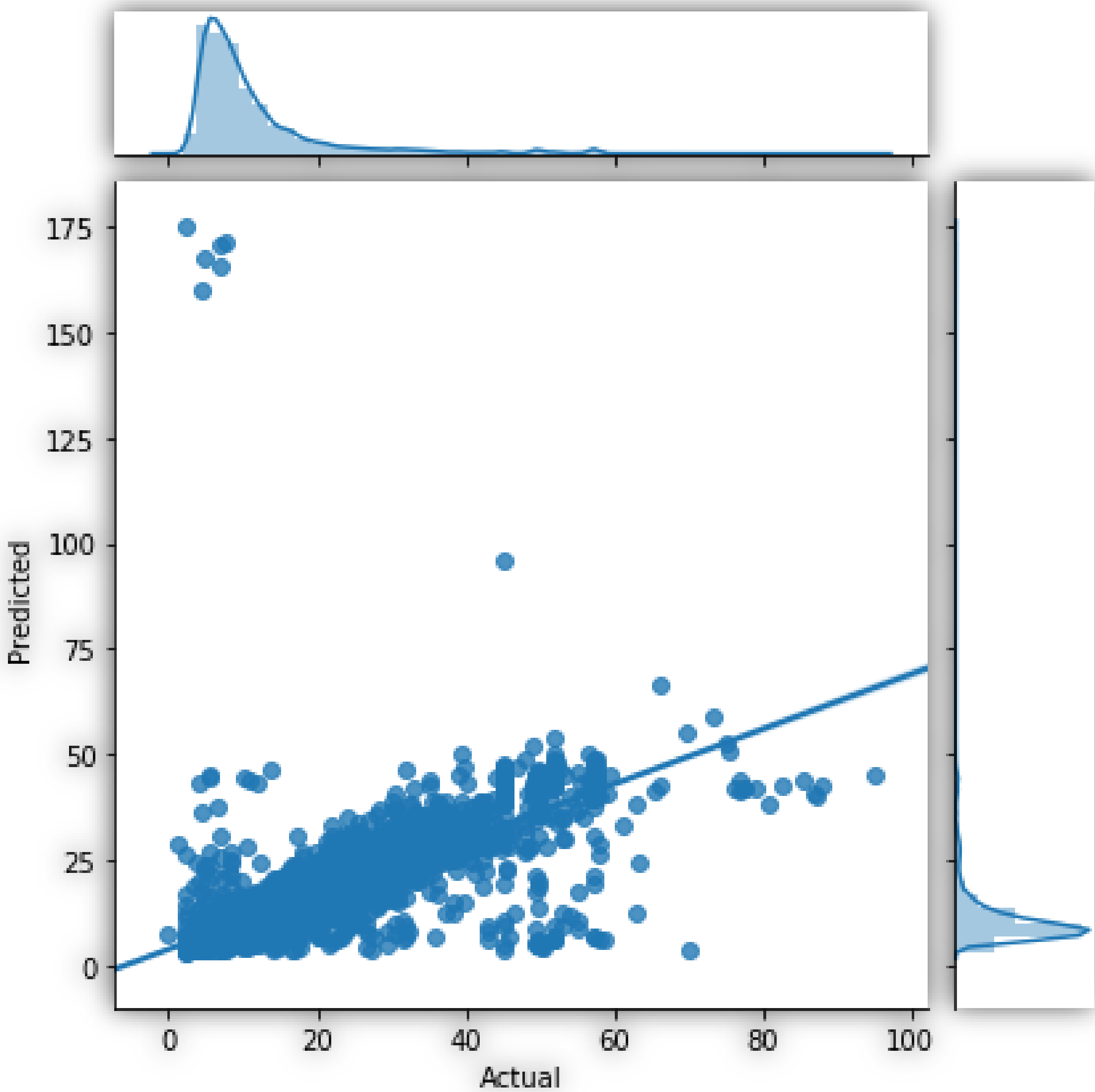
	fare_amount	Predicted_Fare_amount	Predicted_Fare_amount_2	Predicted_Fare_amount_3	Predicted_Fare_amount_ridge
0	4.50	7.054622	7.076794	5.326533	5.370514
1	16.90	20.475613	17.242852	17.005865	16.910257
2	5.70	7.703590	8.176632	6.813738	7.286942
3	7.70	10.253696	8.616914	10.281141	9.581124
4	5.30	8.806360	8.955180	6.734853	7.175125
5	12.10	12.040845	12.375803	9.817240	10.625876
6	7.50	8.004379	7.945076	8.248697	8.385761
7	16.50	12.706896	13.739808	11.771370	12.494557
9	8.90	10.344786	9.627012	8.308708	8.406594
10	5.30	7.676550	8.621087	8.233952	8.202573
11	5.50	6.299638	7.174048	6.807526	7.222839
12	4.10	6.488439	7.324098	6.147039	5.992372
13	7.00	8.842949	7.921667	9.763849	9.339548
14	7.70	8.213558	7.773250	7.592928	7.402851
15	5.00	5.534144	6.294443	6.542925	6.832362

Error Metrics

Mean Squared Error	**29.59041911**
Root Mean squared error	**5.439707631**

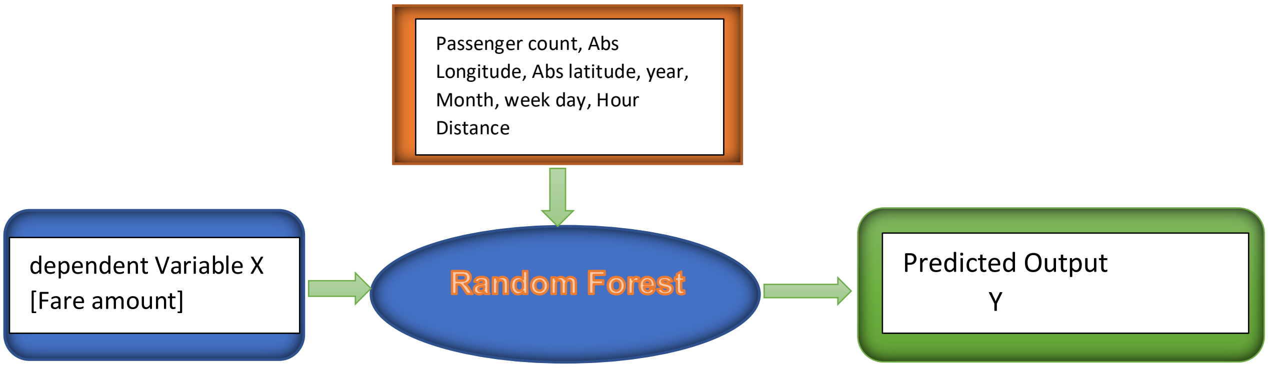
Visualization




175
150
125
100
75
50
25
0
Predicted
0
20
40
60
80
100
Actual

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.



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}
```

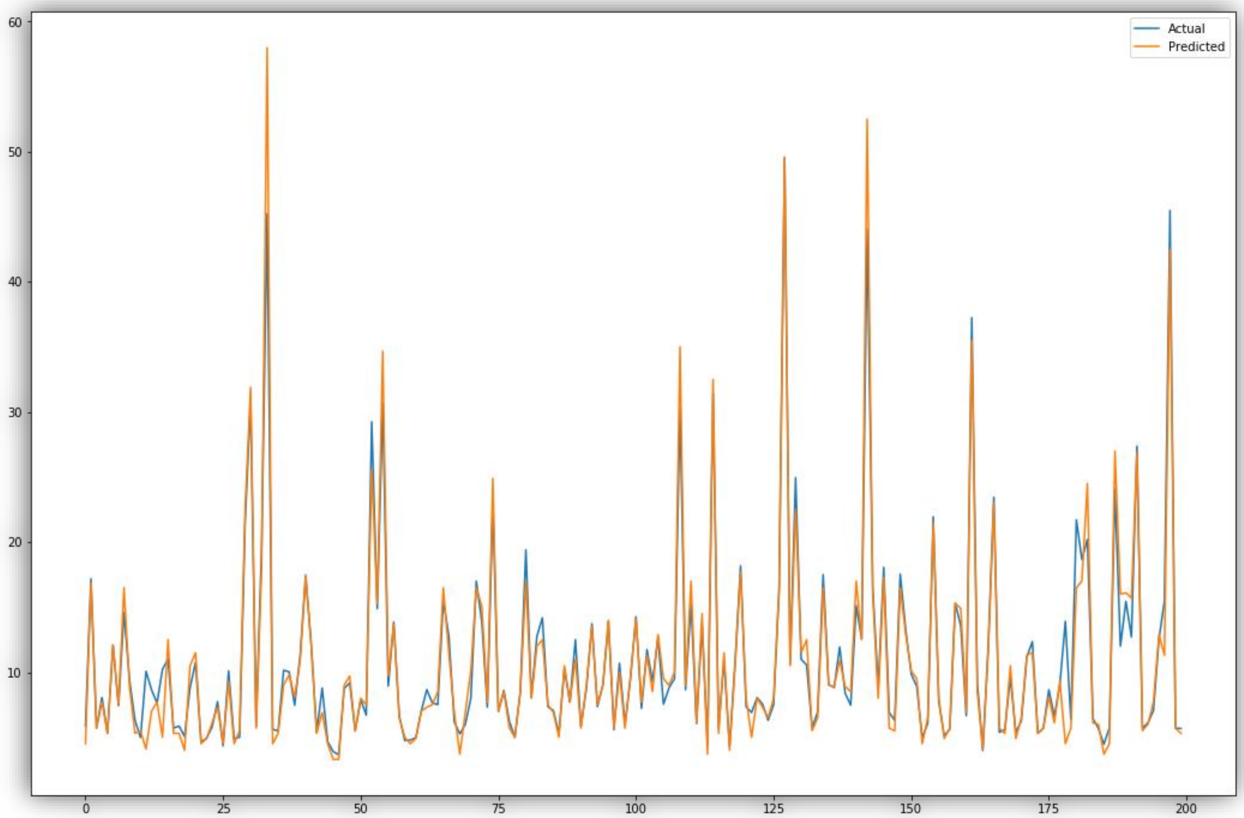
Feature Importance

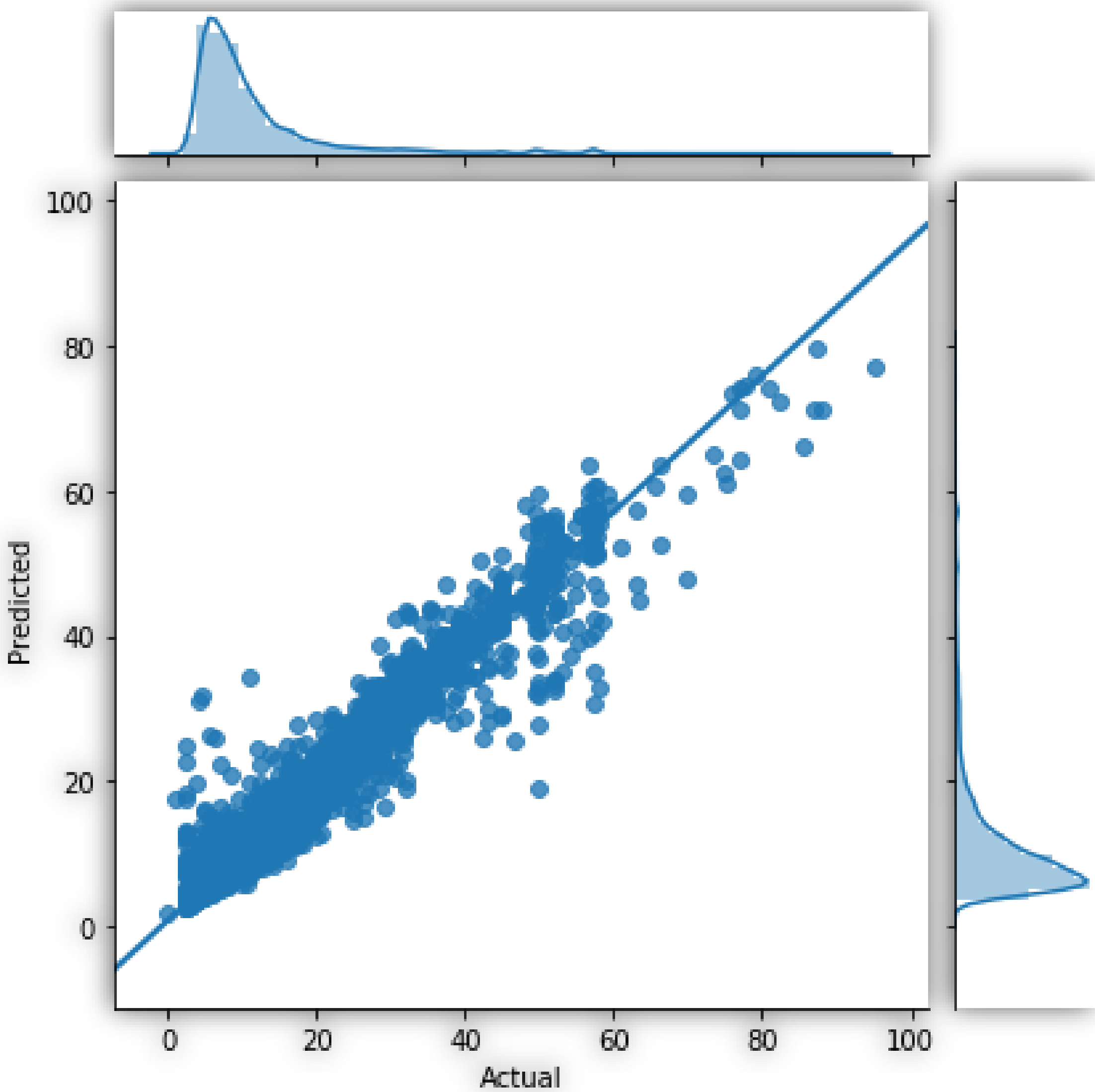
	0	Feature_importance
0	passenger_count	0.007460
1	abs_longi	0.116075
2	abs_lat	0.042825
3	year	0.025709
4	month	0.015558
5	weekday	0.011556
6	hour	0.022228
7	great_circle_distance	0.758588

Predicted Vs Actual Amount

	fare_amount	Predicted_Fare_amount	Predicted_Fare_amount_2	Predicted_Fare_amount_3	Predicted_Fare_amount_4
0	4.50	7.054622	7.076794	5.326533	5.860
1	16.90	20.475613	17.242852	17.005865	17.180
2	5.70	7.703590	8.176632	6.813738	5.710
3	7.70	10.253696	8.616914	10.281141	8.060
4	5.30	8.806360	8.955180	6.734853	5.300
5	12.10	12.040845	12.375803	9.817240	12.060
6	7.50	8.004379	7.945076	8.248697	7.430
7	16.50	12.706896	13.739808	11.771370	14.580
9	8.90	10.344786	9.627012	8.308708	9.420
10	5.30	7.676550	8.621087	8.233952	6.300
11	5.50	6.299638	7.174048	6.807526	4.980
12	4.10	6.488439	7.324098	6.147039	10.070
13	7.00	8.842949	7.921667	9.763849	8.650
14	7.70	8.213558	7.773250	7.592928	7.620
15	5.00	5.534144	6.294443	6.542925	10.230

Visualizing




100
80
60
40
20
0
Predicted
0
20
40
60
80
100
Actual

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	

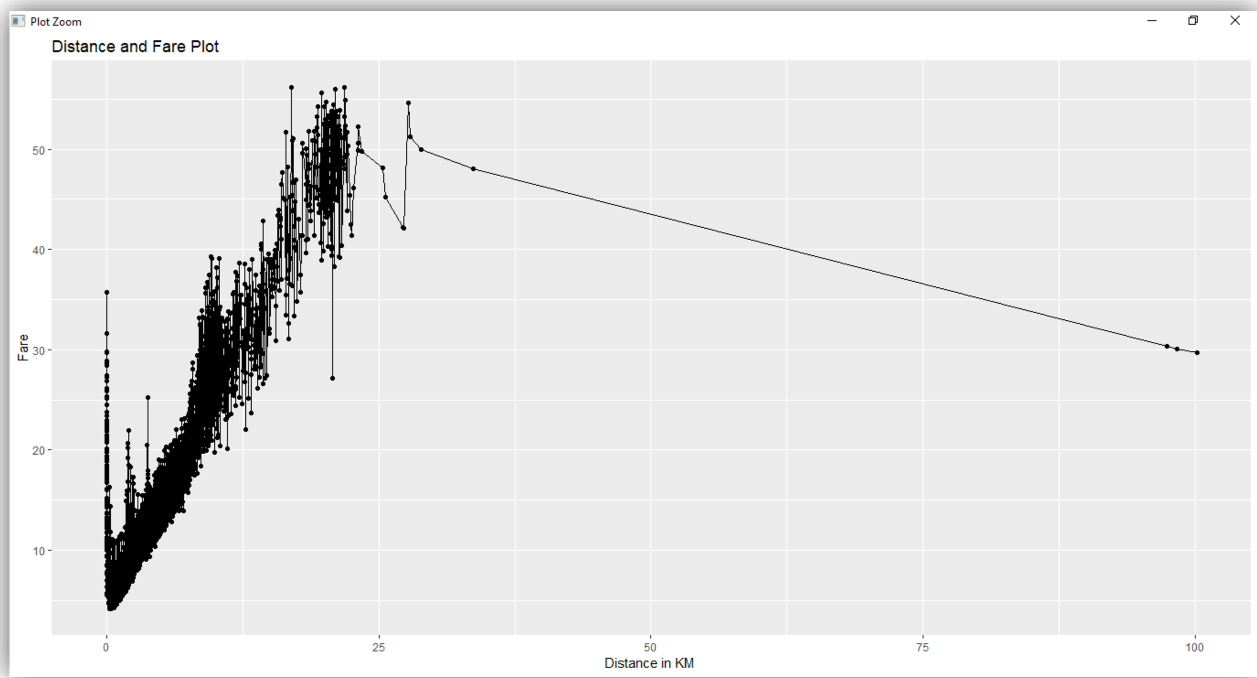
Mean Squared Error	33.90217895	34.81684133	28.39558067	29.59041911	3.0958367
Root Mean squared error	5.822557767	5.900579745	5.328750386	5.439707631	1.759499

3.2 Answers of asked questions

Predicted Fare amount for a Test Cab Data

	passenger_count	month	year	dayOfWeek	hour	distance.in.KM	fare_amount
1	1	1	2015	3	13	2.3258621	11.071508
2	1	1	2015	3	13	2.4280699	11.071117
3	1	10	2011	7	11	0.6193209	4.958508
4	1	12	2012	7	21	1.9632293	8.702479
5	1	12	2012	7	21	5.3933363	15.513474
6	1	12	2012	7	21	3.2261589	11.169608
7	1	10	2011	5	12	0.9306427	5.562670
8	1	10	2011	5	12	21.5642316	48.730234
9	1	10	2011	5	12	3.8783014	11.899879
10	1	2	2014	3	15	1.1010259	6.255133
11	1	2	2014	3	15	2.3202815	9.648307
12	1	2	2014	3	15	4.8245773	17.473484
13	1	3	2010	2	20	0.7234793	5.463517
14	1	3	2010	2	20	1.6773801	6.756970
15	1	10	2011	5	3	2.5068374	7.871874
16	1	10	2011	5	3	5.1211056	12.465867
17	1	7	2012	1	16	0.2991728	5.209657
18	1	7	2012	1	16	2.5339829	8.847124
19	1	7	2012	1	16	0.7813187	5.321736
20	1	7	2012	1	16	0.4277606	5.003648
21	1	10	2014	4	2	1.6537961	6.926666

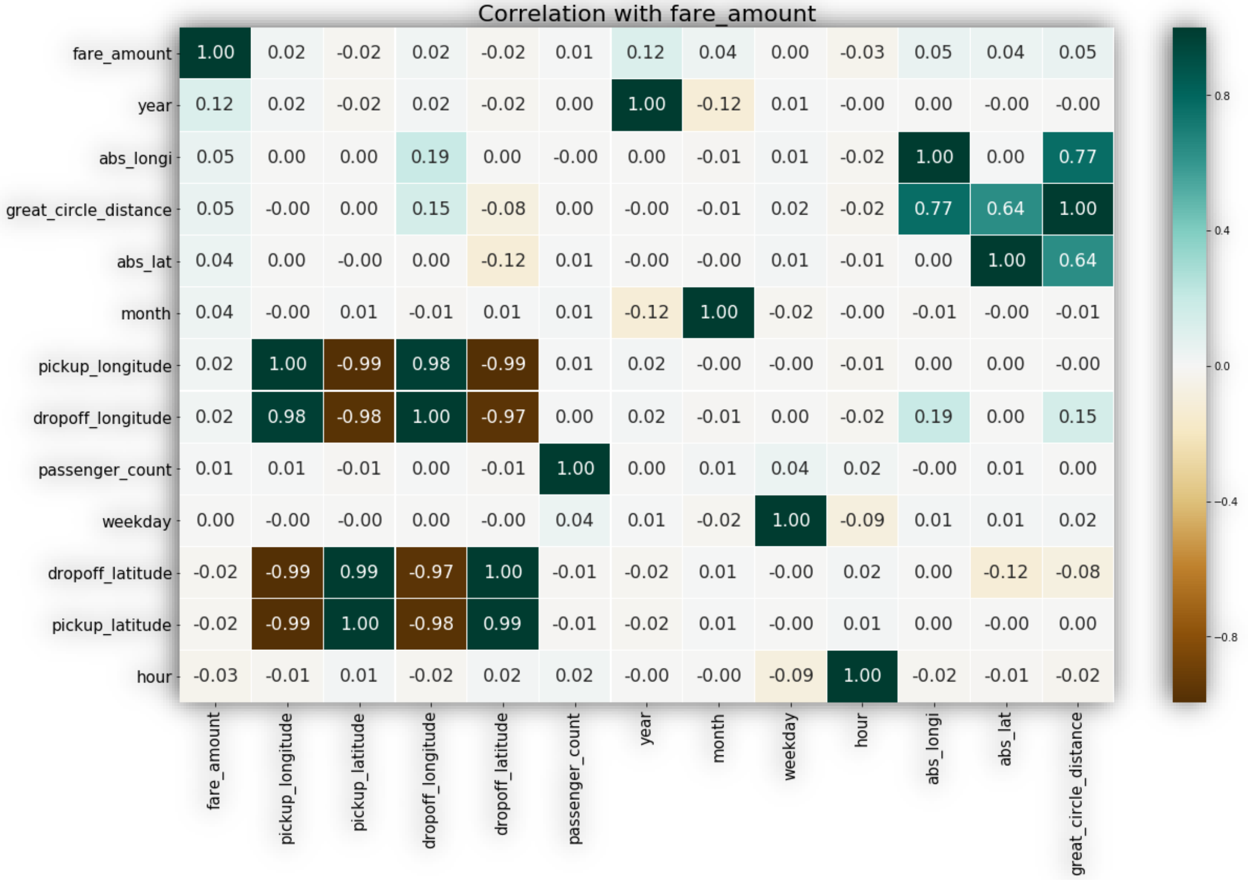
Predicted Fare amount for Distance



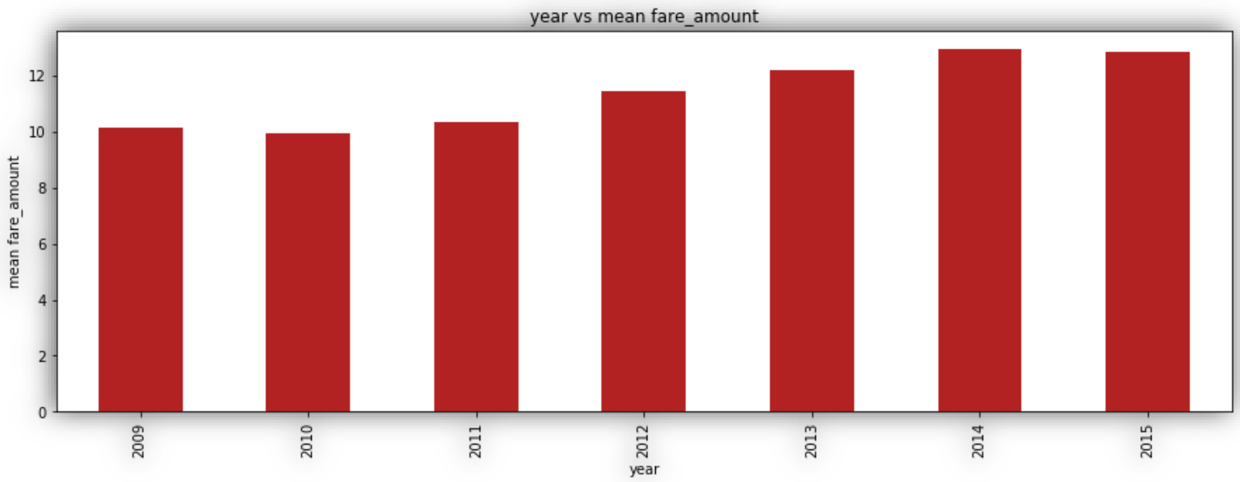
Appendix

Extra Figures

Correlation Plot



Year Vs Fare amount



Distance Frequency