Cab Fare in R

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.

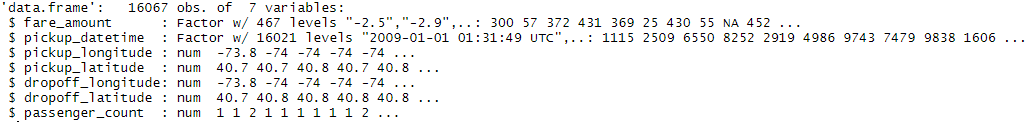
* 1. Data

Our task is to build regression model which will predict the fare of the cab depending on various others attributes. Given below is a snippet of the data set.



The data-set contains the following variable :   
 1. fare\_amount – denotes the fare of the cab.  
 2. pickup\_datetime – denotes the time when the cab was   
 taken.  
 3. pickup\_longitude – denotes the location in terms of   
 longitude.  
 4. pickup\_latitude – denotes the location in terms of latitude.

5. dropoff\_longitude- float for longitude coordinate of where   
 the ride ended.  
6. dropoff\_latitude – float for latitude coordinate of where   
 the ride ended.  
7. passenger\_count – indicates the no of passenger in the   
 cab.

The details of the variables are in the dataset as follows :  


2 METHODOLOGY

2.1Preprocessing   
 Prior to creating any machine learning models ,   
 it is imperative to have a feel of the data. Pre processing   
 involves data analysis with plots and graphs , recognizing   
 outliers , imputing missing values , feature selection and   
 feature selection etc.

2.1.2 Data Cleaning :

1.Invalid Passenger Count

Before venturing into Missing Value Analysis or Outlier Analysis , it is imperative that we get rid of invalid rows i.e. rows   
having invalid attribute column.

First step is to get rid of invalid passenger\_count - Detecting invalid passenger\_count i.e less than 1 or greater than 6 as the underlying assumption is that more than 6 passengers can't enter the car.

There are total 58 rows with passenger\_count equivalent to 0 ; which needs removal.

Next removal of passenger\_count which exceeds 6 : total 20 in number

Total 78 rows are removed and number of rows goes from 16067 to 15989

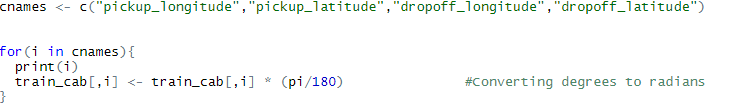
1. Latitudes and Longitudes

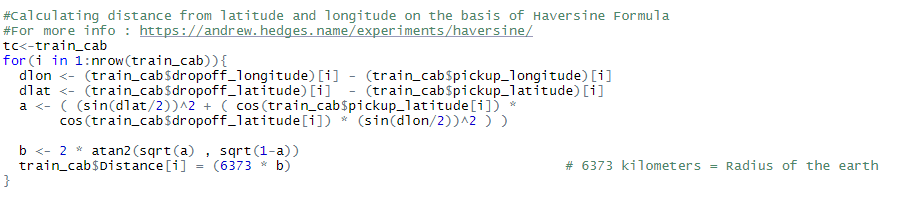
Valid values for latitudes and longitudes lies between -90 to 90 degrees and -180 t0 180 degrees respectively.

1 row exist with latitude exceeding 90 degrees which is subsequently removed upon its detection.

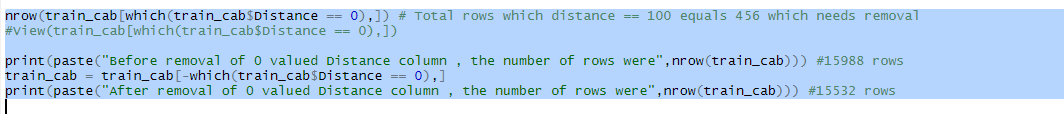
Next is feature engineering Distance from latitude and longitude  
with the help of Haversine’s Formula.

The following line of code is used:





Upon inspecting further , its found that 456 rows exists with Distance valued 0 – which needs removal as 0 Distance route seems impossible. The following snippet depicts the removal of invalid Distance row:

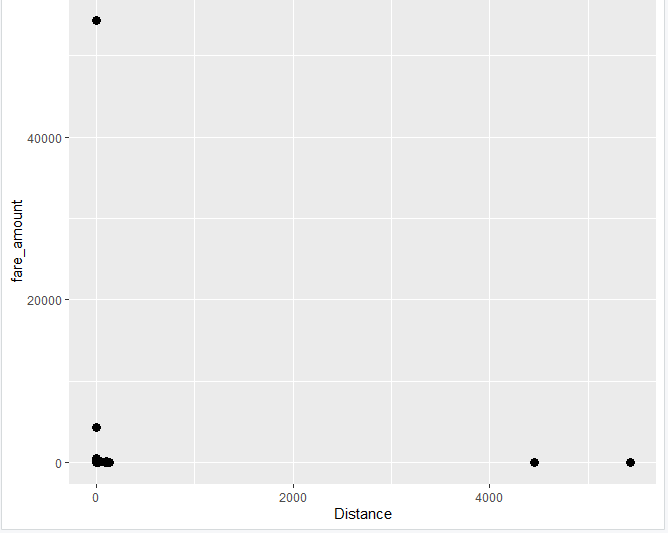


Total number of rows goes down from 15988 to 15532

Also, latitude and longitude valued 0 needs removal as its invalid.

The number of rows decreases further to 15511.

1. Invalid fare\_amount :



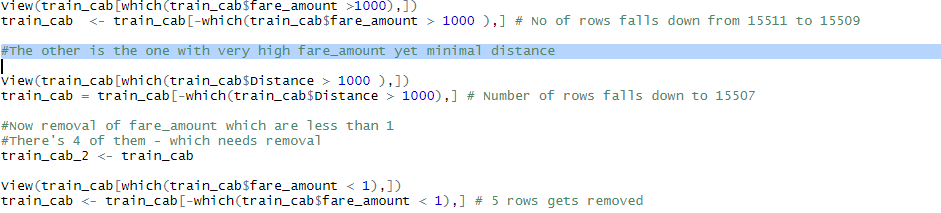
#From the above scatterplot we can observe some peculiar observations :

2 Observations where the distance exceeds 1000 km yet the fare remains very bare minimum.Upon further inspection , it is revealed that the passenger\_count is 1.

The other is the one with very high fare\_amount yet minimal distance

Thus ,it's better to remove these observations as its meaningless  
Also removal of rows with fare\_amount less than 1 gets removed as well.

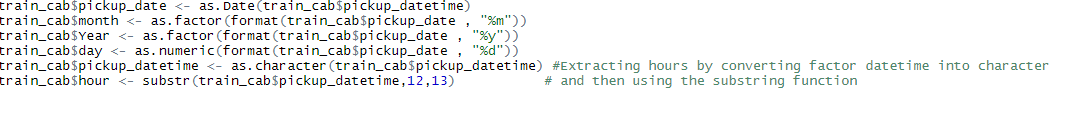
Following is the snippet of the code :



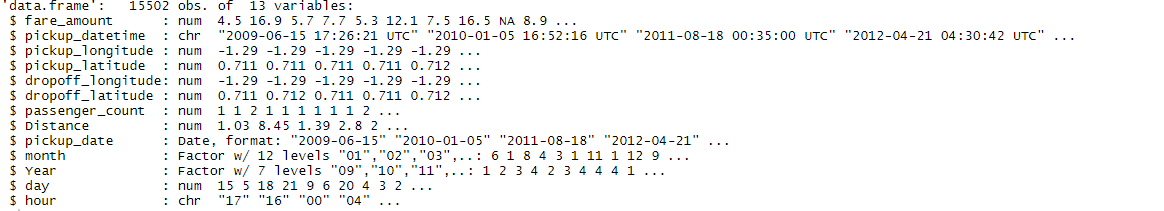
The numbers we have left after data cleaning is 15502 i.e. 565 rows gets removed.

2.1.3 Feature Engineering

It’s imperative to convert pickup\_datetime from factor to datetime and extract month,Year ,Hour and other valuable derivatives from it. Following is the snippet of the code :

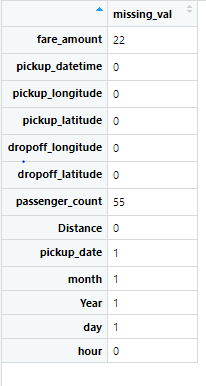


This is now the structure of the data-frame:

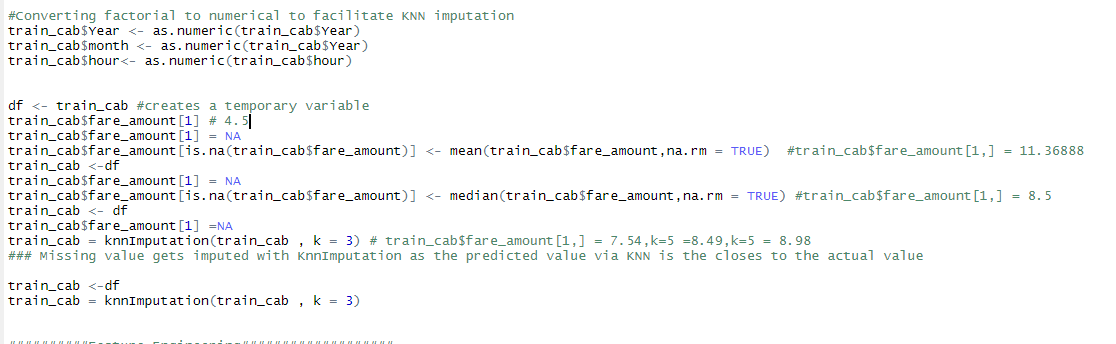


2.2 Missing value Analysis   
 Missing value involves imputing the missing values with   
 three methods : mean , median or KNN Imputation.

Before imputing missing values , it is important to discover   
 the number of missing values per column.



The missing value is imputed with trial and error method ,  
 first the value is imputed with , then median and finally with   
 KNN imputation method and the method producing the   
 closest prediction to the actual value gets the nod.Following   
 is the snippet of the code :



As the value imputed is the closest to the actual value which is 4.5  
 ,KNN Imputation method is selected.

Also , before performing KNN Imputation Method , we convert the   
 all the factorial to numerical in order to facilitate KNN Imputation.

All the missing values gets imputed.

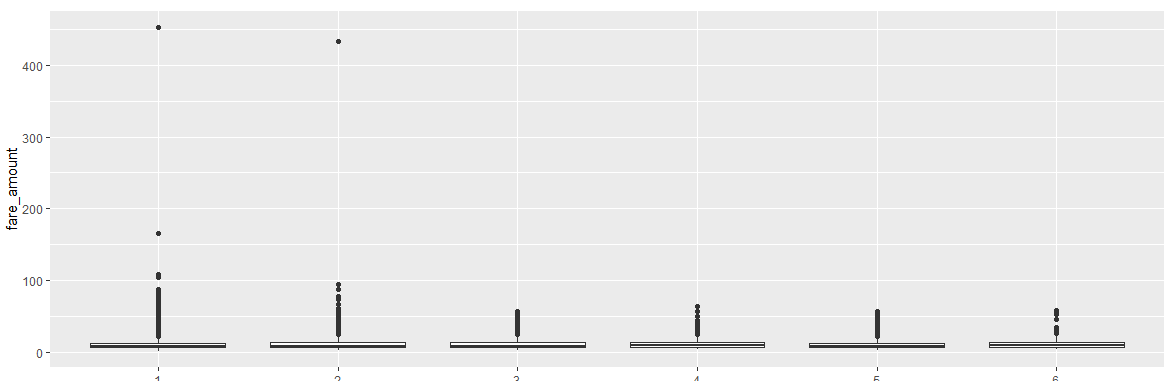
.

2.3 Outlier Analysis And Data Visualization

Outlier Analysis is performed to detect the outliers present in   
the data set in order to weed out the unnecessary variables. Graphical methods can be used, in this case Boxplot is used to visualize the presence of outliers.

It becomes imperative to get rid of outliers as their presence can generate ambiguity.

Boxplot is generated which looks like this :

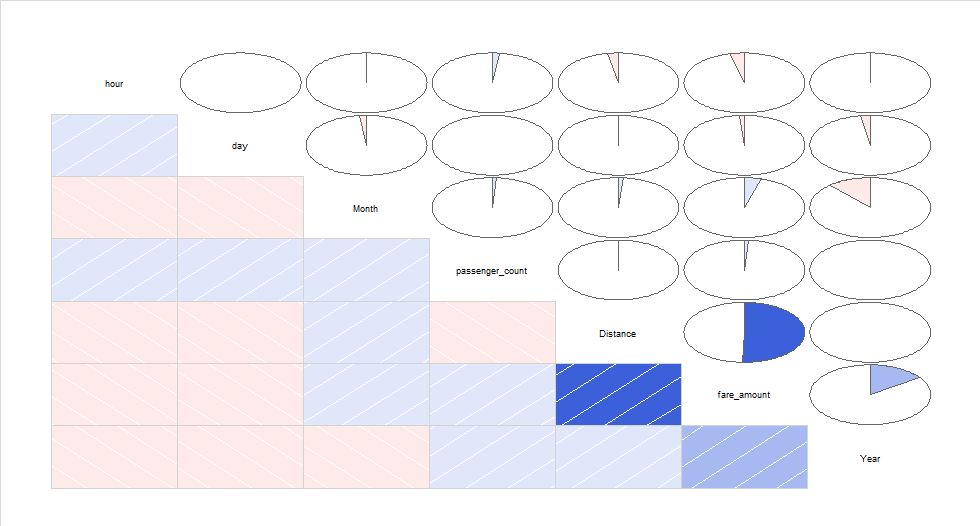


As it can be seen , we have high number of outliers present in  
the data-frame. Total number of fare\_amount exceeding 20 is  
1624 in number while the mean of fare\_amount is 11.37.

This can cause heavy deflection in our prediction

Thus we need to remove them.

Co-relation Analysis:



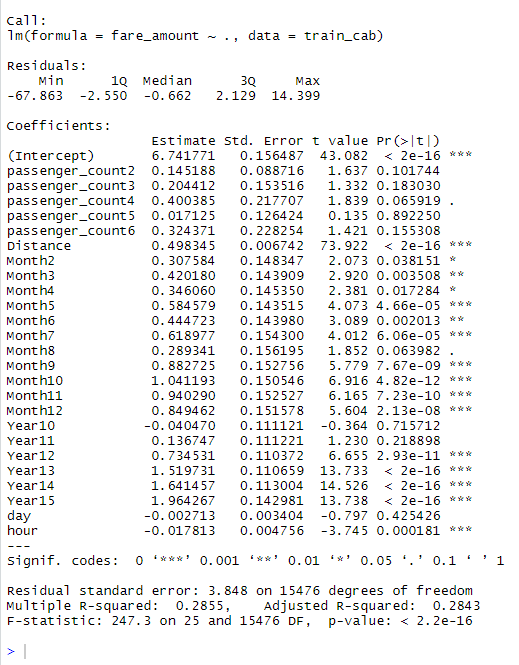
Here we can see that the fare\_amount is dependent on distance  
(almost 50%) and very low co-relation with passenger\_amount.

There’s a 25% co-relation with Year which probably indicates that average  
fare\_amount increases as the year goes by.

MACHINE LEARNING

Once the pre-processing steps are completed , it is the next step to build   
 model. Since it’s a regression model , we will select from one of the   
 three – Linear Regression , Decision Tree Model or Random Forest   
 Model.

1. Linear Regression Model : Following is the summary of the model developed.



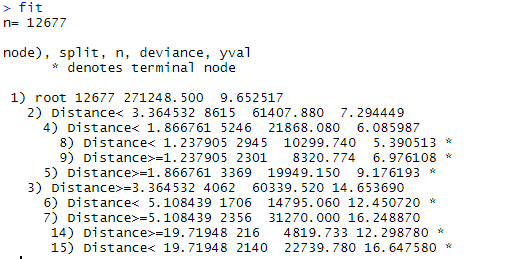
Following inferences can be made :

a.Fare\_amount is independent of passenger\_count

b.Fare\_amount is heavily dependent on Year , hour of the day cab is taken and also months

c.Adjusted R squared value is 0.2853 which is not acceptable i.e. independent variable can only explain 28% variation of the target variable

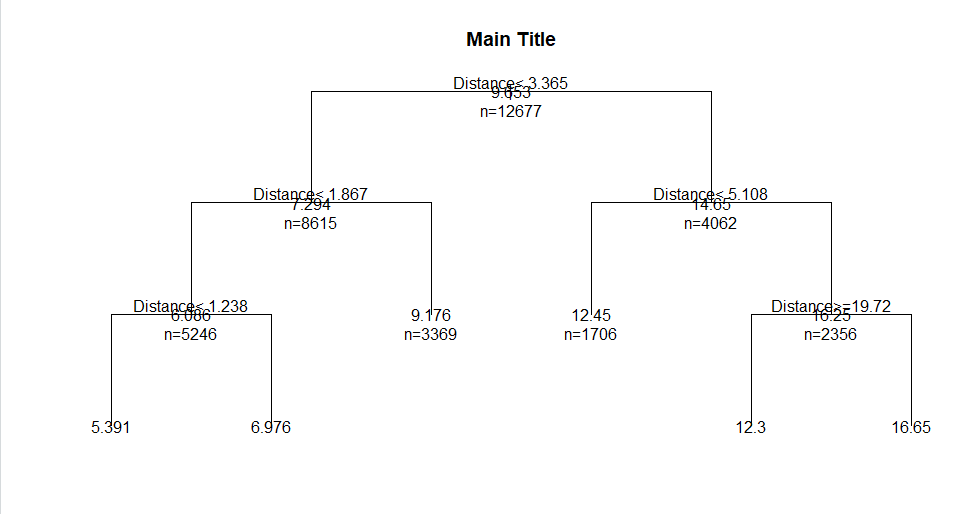
d.Also the error rate came out to be 36.36% which means that the model is only 67% accurate(MAPE Based)

2.Decision Tree Model : With Decision Tree , higher accuracy was obtained i.e. approximately 80% (MAPE based),which is not commendable but definitely an improvement over the Regression Model.Here’s a snippet of the model : 

Here n = 12677 indicates no of observations used

The above picture indicates that the fare\_amount is predicted soley on the basis of Distance travelled.

A better picture would like look this :

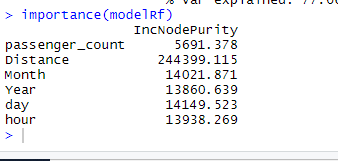


The picture speaks volumes in regards to the dependency of fare\_amount with Distance travelled.

Random Forest:

Random Forest is an extension of Decision Tree Algorithm. The decision is made by combining various decision trees.

In our case , it gave a staggering accuracy of 99%.(MAPE based)



The IncNodePurity indicates the importance of every predictor variable in regards to the target variable.

Random Forest , again reaffirms the fact that fare\_amount is heavily relied on Distance travelled.

CONCLUSION

Due to High Accuracy , Random Forest is selected as the algorithm to predict cab fare.