CAB FARE PREDICITON

Mainak Sarkar

25th May 2019

LIST OF FIGURES:

Chapter 1

INTRODUCTION

1. Problem Statement

A cab rental start-up company had successfully run the pilot project and now want to launch their cab service across the country. Historical data has been collected from the pilot project and now they want to apply analytics for fare prediction. Our task is to apply different regression models on the data and provide them with a system that could predict the fare accurately.

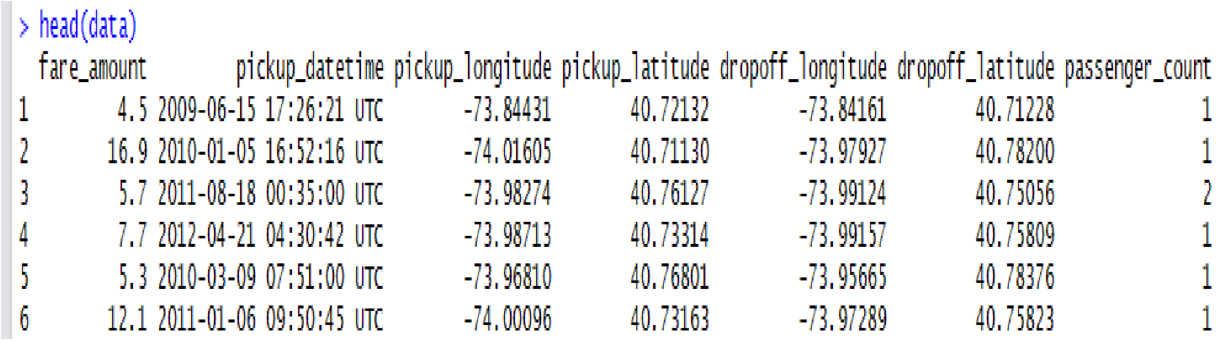
2. Data

The data that has been collected from the pilot project contains the following variables:

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

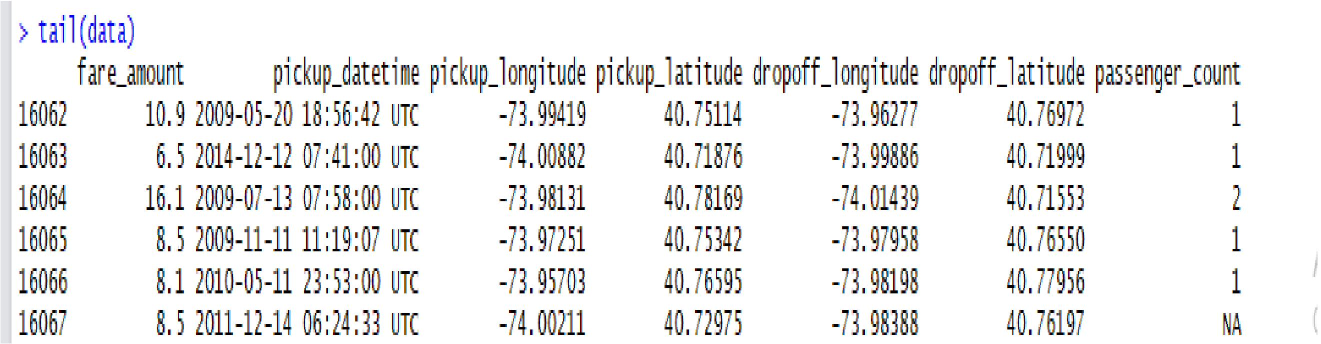
Here "fare_amount" is the dependent variable that we need to predict.

Given below is a sample of dataset from the top and bottom of the dataset respectively:



```
> head(data)
  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
```

Figure 1: Sample of data (top 6 rows)



```
> tail(data)
      fare_amount         pickup_datetime pickup_longitude pickup_latitude dropoff_longitude dropoff_latitude passenger_count
16062        10.9 2009-05-20 18:56:42 UTC        -73.99419        40.75114         -73.96277         40.76972               1
16063         6.5 2014-12-12 07:41:00 UTC        -74.00882        40.71876         -73.99886         40.71999               1
16064        16.1 2009-07-13 07:58:00 UTC        -73.98131        40.78169         -74.01439         40.71553               2
16065         8.5 2009-11-11 11:19:07 UTC        -73.97251        40.75342         -73.97958         40.76550               1
16066         8.1 2010-05-11 23:53:00 UTC        -73.95703        40.76595         -73.98198         40.77956               1
16067         8.5 2011-12-14 06:24:33 UTC        -74.00211        40.72975         -73.98388         40.76197              NA
```

Figure 2: Sample of data (bottom 6 rows)

Chapter 2

METHODOLOGY

1. Data Preprocessing

Data preprocessing or data cleaning is one of the most crucial step of building a machine learning model. Almost 80% of the time is dedicated to the data preprocessing. Because if we feed messy or uncleaned data to the model then it will generate irrelevant and wrong results. In the data mining process the data need to be pre-processed first to make them quality data to acquire the quality analysis and information to make quality decision.
The first step of data preprocessing is to check the class of each variable and then transform them as required.

Real world data are generally incomplete (lacking attribute values, lacking certain attributes of interest, or containing only aggregate data), Noisy (containing errors or outliers) and Inconsistent (containing discrepancies in codes or names), so to prepare the data for mining by using following processes is known as data preprocessing.

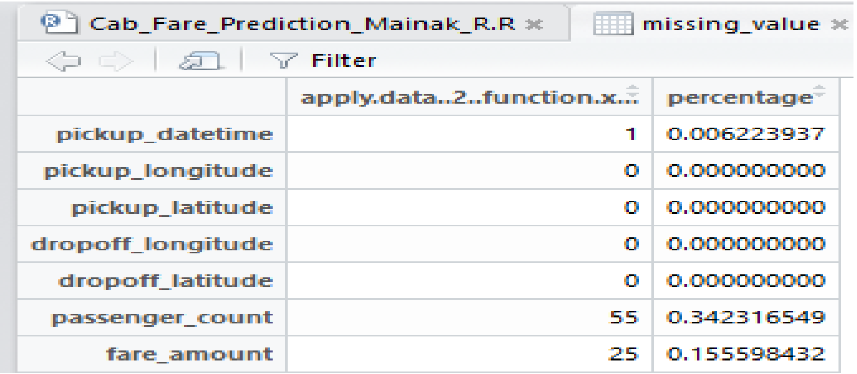
1.1. Missing Value Analysis:

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. Missing data can occur because of nonresponse or no information is provided for one or more items or for a whole unit.

Sometimes the data is found to contain a lot of missing values. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Thus, missing value analysis is a very important part of data cleaning. It can be done in two ways:

- Detecting and deletion of the rows containing missing values
- Imputing the missing values by statistical methods like- mean, median or by KNN imputation or by prediction

The figure below shows the number of missing values in our dataset and their percentages with respect to respective columns.



	apply.data..2..function.x...	percentage
pickup_datetime	1	0.006223937
pickup_longitude	0	0.000000000
pickup_latitude	0	0.000000000
dropoff_longitude	0	0.000000000
dropoff_latitude	0	0.000000000
passenger_count	55	0.342316549
fare_amount	25	0.155598432

Figure 3: Missing value count and percentage

As, we can see that the amount of missing value is too less, i.e.-less than 1% so deleting the rows won't result in any information loss.

1.2. Outlier Analysis:

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

The outliers of a dataset can be understood by checking the histogram plots of the variables.

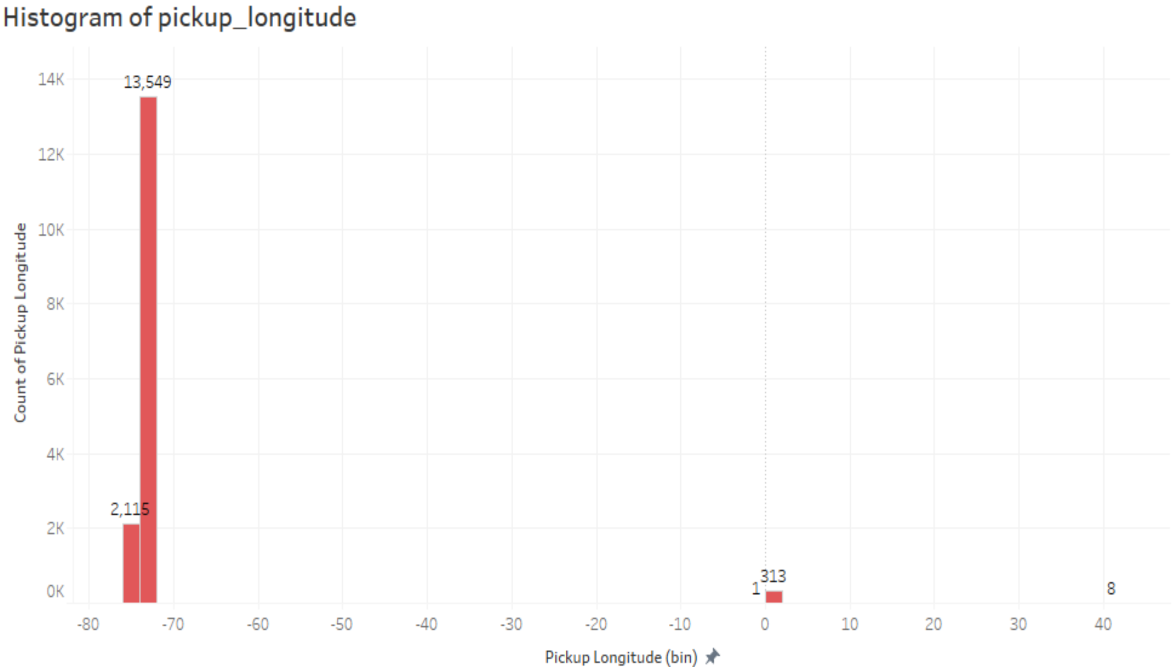


Figure 4: Histogram of pickup_longitude

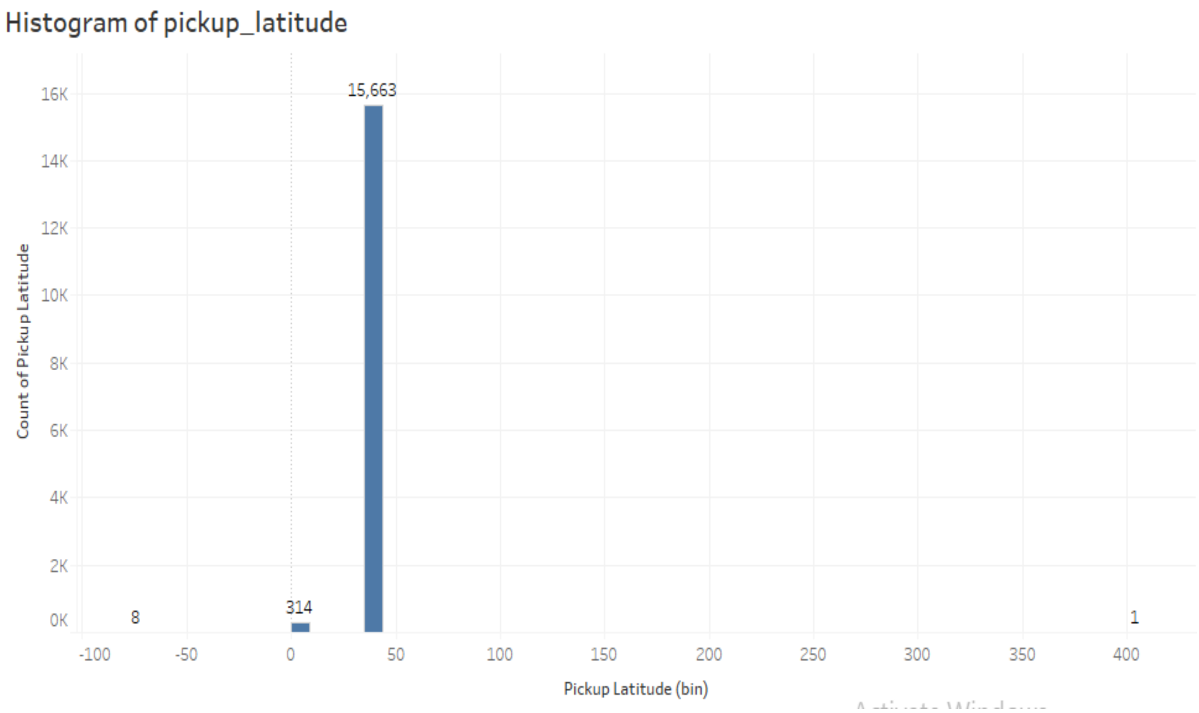


Figure 5: Histogram of pickup_latitude

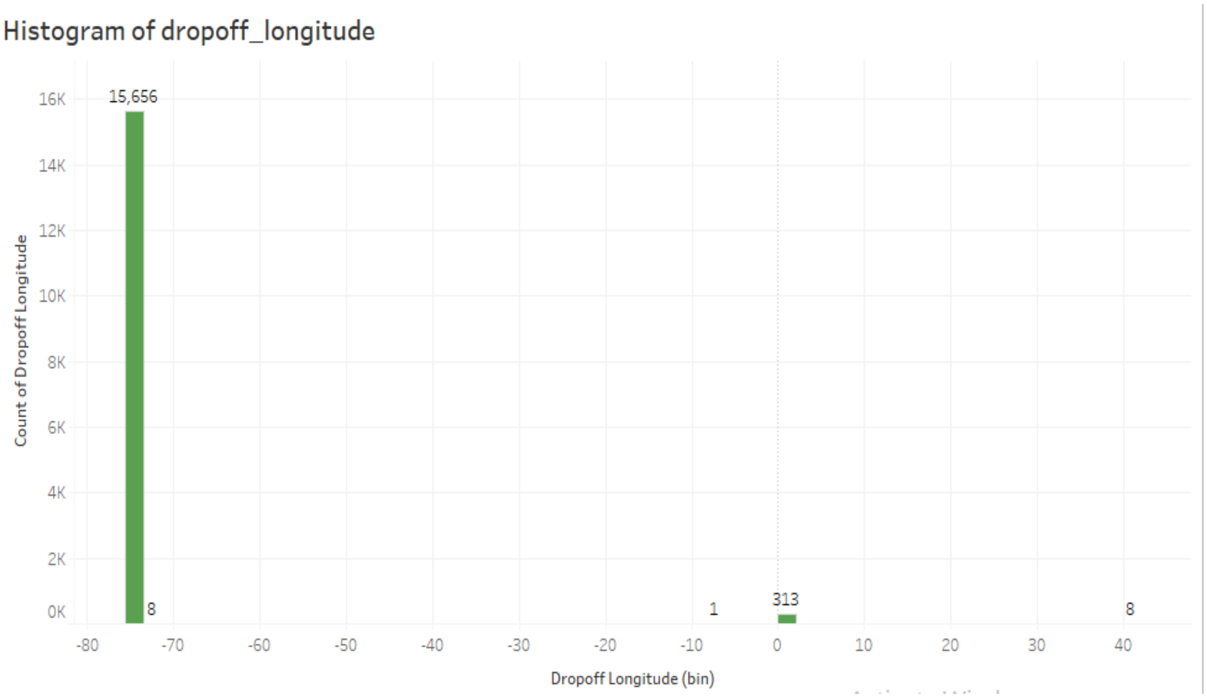


Figure 6: Histogram of dropoff_longitude

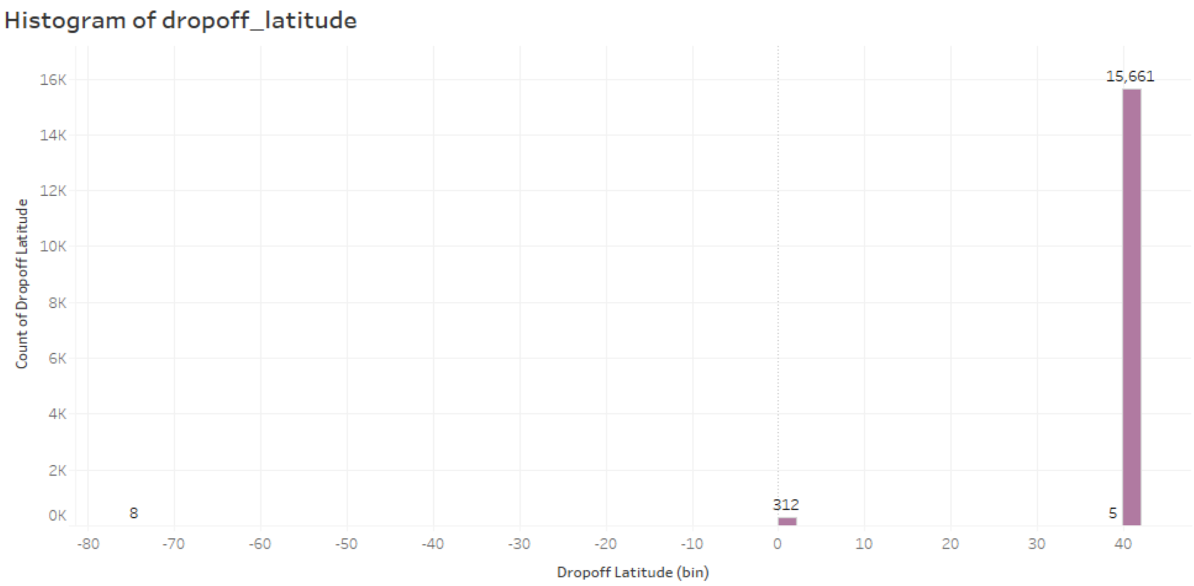


Figure 7: Histogram of dropoff_latitude

We can see from the distributions there are some bins which are away from the main bin. So, this might be due to the presence of outliers and extreme values.

Now we have carried out our investigation further by checking the box plots of the variables.

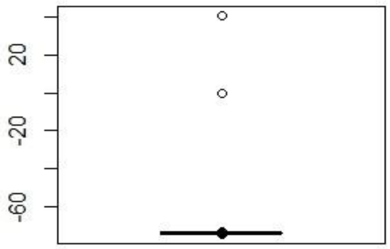


Figure 8: Boxplot of pickup_longitude

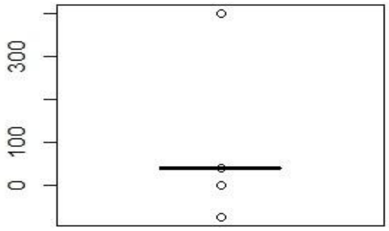


Figure 9: Boxplot of pickup_latitude

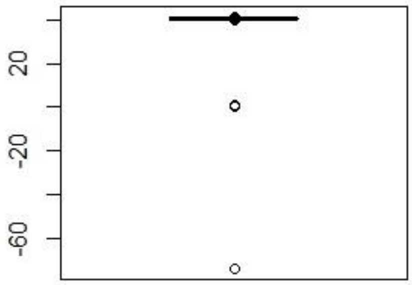


Figure 10: Boxplot of dropoff_longitude

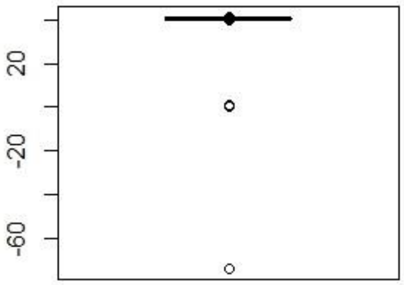


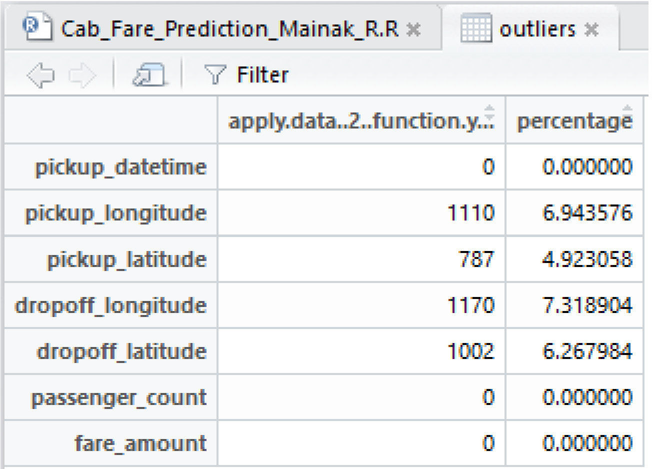
Figure 11: Boxplot of dropoff_latitude

So here the data points above the upper fence and below the lower fence (outside the box plot) show the presence of outliers.

Dealing with these outliers is a very essential part of our analysis. It can be done by following two processes:

- Deleting the outliers
- Replacing the outliers with NAs and then imputing them by statistical methods like- mean, median or by KNN imputation or by prediction

The figure below shows the number of NAs after replacing the outliers with NAs and their percentages with respect to respective columns.

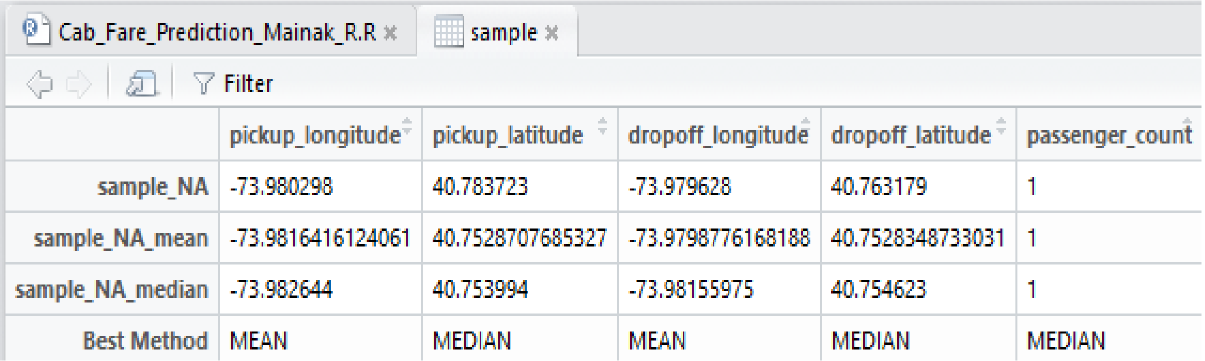


	apply.data..2..function.y...	percentage
pickup_datetime	0	0.000000
pickup_longitude	1110	6.943576
pickup_latitude	787	4.923058
dropoff_longitude	1170	7.318904
dropoff_latitude	1002	6.267984
passenger_count	0	0.000000
fare_amount	0	0.000000

Figure 12: Count of outliers and percentage

As, we can see that the number of outliers (NAs) are huge and deleting them would result in a heavy percentage of information loss. So we have to opt for imputing them.

An algorithm is designed to find the best method for imputing outliers with respect to each column. The output table shows the following:



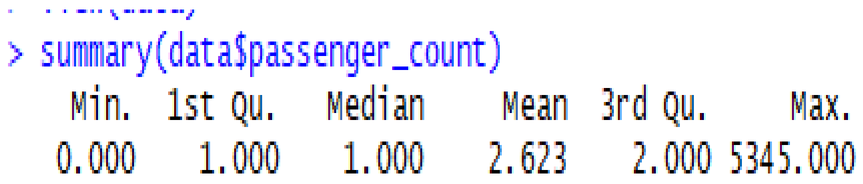
	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
sample_NA	-73.980298	40.783723	-73.979628	40.763179	1
sample_NA_mean	-73.9816416124061	40.7528707685327	-73.9798776168188	40.7528348733031	1
sample_NA_median	-73.982644	40.753994	-73.98155975	40.754623	1
Best Method	MEAN	MEDIAN	MEAN	MEDIAN	MEDIAN

Figure 13: Table for detection of Best Method for Imputation

So we have applied the above mentioned methods to the respective columns to impute in place of NAs and get rid of the outliers.

**NOTE:

- The outlier analysis technique is not applied on “passenger_count” variable as on doing so the values 4,5 and 6 are getting detected as outliers. This is because the number of such instances is very less compared to the values 1,2 and 3 and the population mean is 2.623. But those values of “passenger_count” is practically possible.



```
> summary(data$passenger_count)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
   0.000    1.000    1.000    2.623    2.000 5345.000 
```

Figure 14: Summary of passenger_count

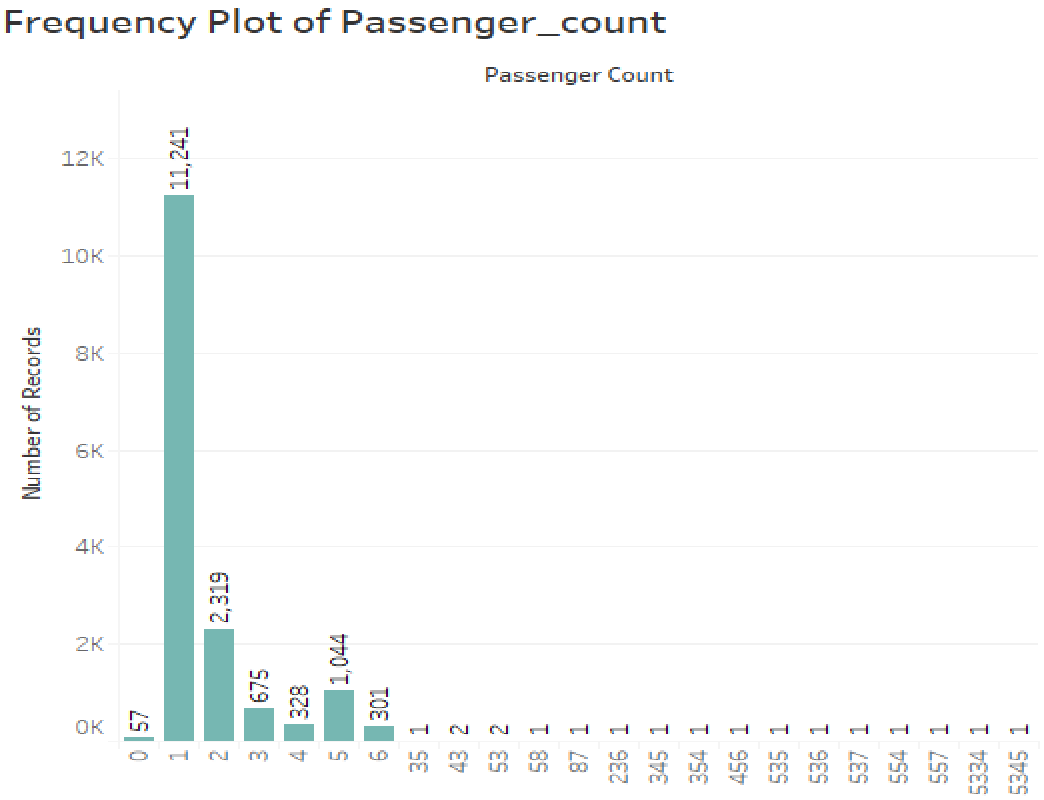


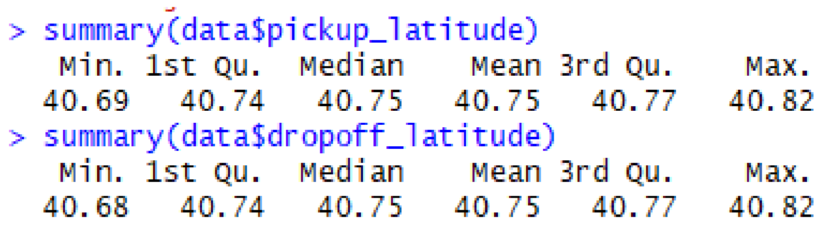
Figure 15: Frequency Plot of passenger_count

- The “fare_amount” variable being a dependent variable we are not supposed to impute it’s values based on any imputation methods. The values of “fare_amount” is based on experiments and is dependent on many variables. Thus we have not taken this variable into consideration of outlier detection.

1.3. Data Manipulation:

In this section we have applied different real-life constraint check to our variables:

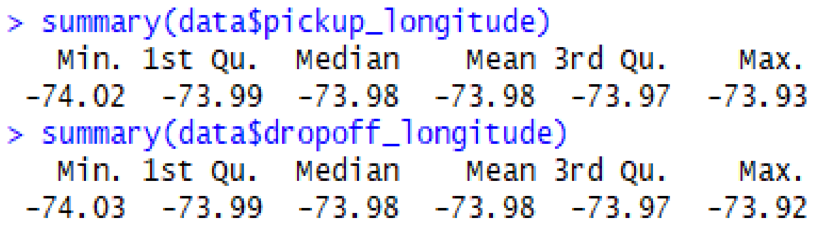
i) **Pickup_latitude:** The latitude value should be between +90 to -90



```
> summary(data$pickup_latitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  40.69   40.74   40.75   40.75   40.77   40.82 
> summary(data$dropoff_latitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  40.68   40.74   40.75   40.75   40.77   40.82 
```

Figure 16: Summary of latitudes

ii) **Pickup_longitude:** The longitude value should be between +180 to -180



```
> summary(data$pickup_longitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -74.02  -73.99  -73.98  -73.98  -73.97  -73.93 
> summary(data$dropoff_longitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -74.03  -73.99  -73.98  -73.98  -73.97  -73.92 
```

Figure 17: Summary of longitudes

iii) **Pickup and Drop off location:** The pickup and drop off location can't be same in a cab ride as that would indicate 0 distance travelled, which is practically not possible.

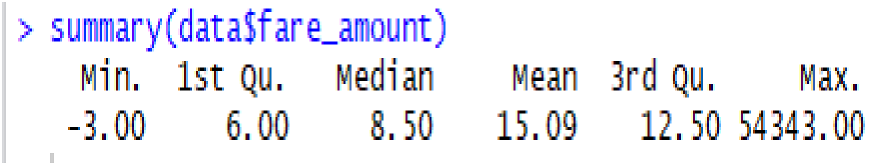
We have counted that there are 118 cases where the distance travelled is 0, so we have deleted such instances.

iv) **Passenger_count:**

- In a car the number of passengers can't be greater than 6. We have found 19 rows where the number of passengers is more than 6 and thus we have deleted such cases.
- The 'passenger_count" can't be less than 1 as well. So we have imputed the minimum "passenger_count' to 1.
- As the "passenger_count" value can't be a fractional so we have rounded off those values to the nearest integer.

v) Fare_amount:

We have checked the summary of the “fare_amount” and found that the minimum value is -3, maximum value is 54343 and the population mean and median are at 15.09 and at 8.50 respectively.



```
> summary(data$fare_amount)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  -3.00     6.00     8.50    15.09    12.50 54343.00
```

Figure 18: Summary of fare_amount

Moreover from the frequency plot we can see that there are very less number of instances of fare greater than 58 and less than 2.5.

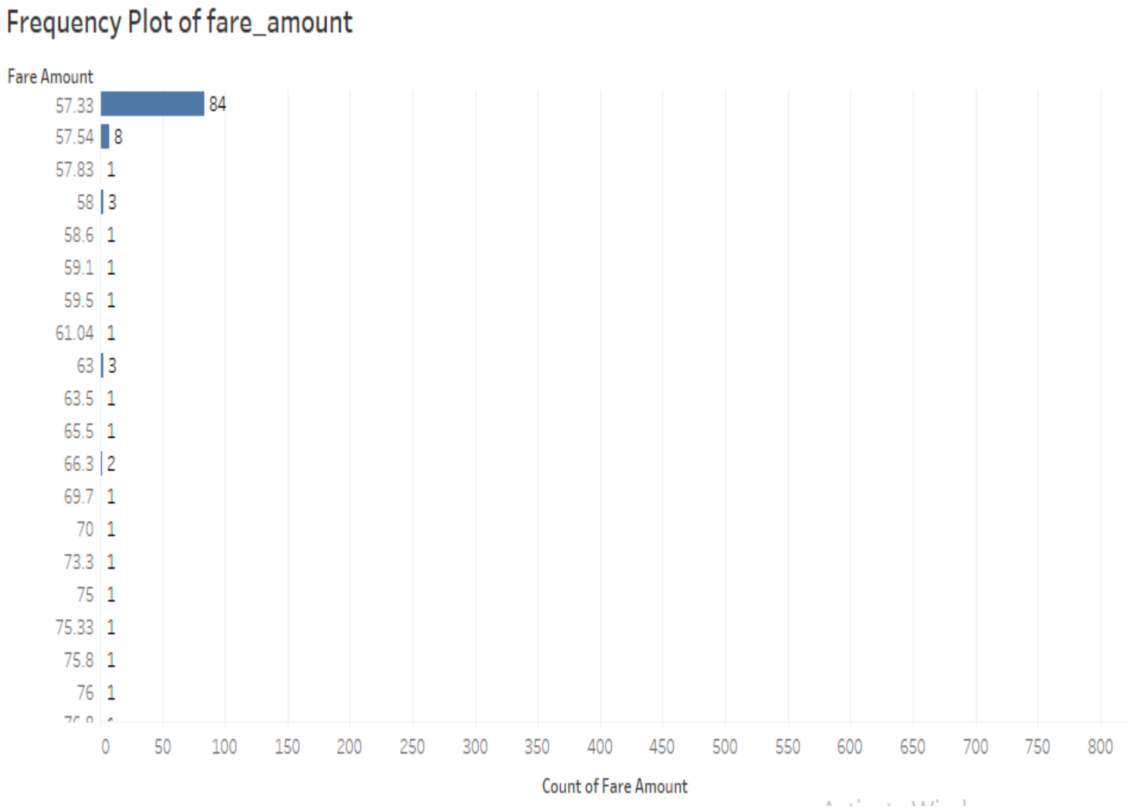


Figure 19: Frequency Plot of fare greater than 58

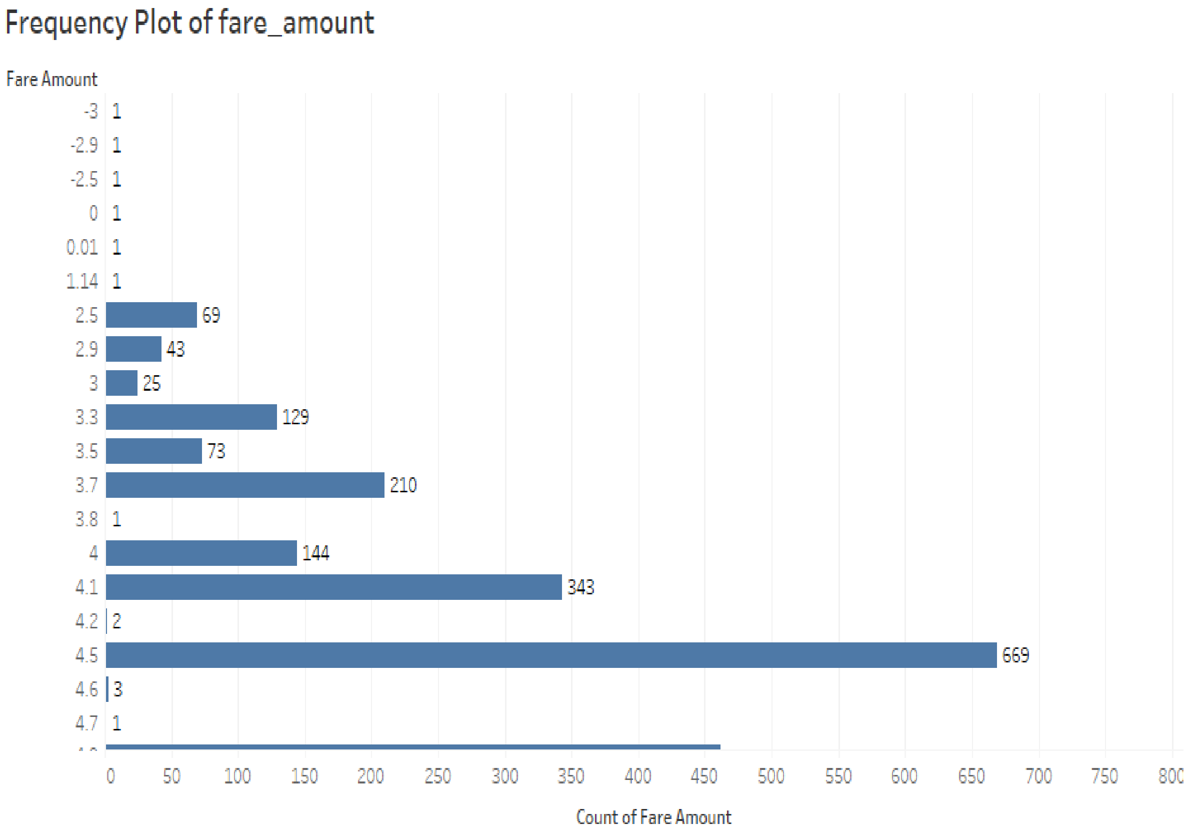


Figure 20: Frequency Plot of fare less than 2.5

- We have got 38 rows where the fare amount is greater than 58 and we have deleted such instances as those fares appear only few times so the actual existence of those values is highly suspicious.
- We have found 6 cases where the fare amount is less than 6 and even 0 or negative in some cases. So these cases might be due to wrong input of data. So we would delete those instances as well.

1.4. Feature Creation:

In the problem we have to predict the cab fare and our independent variables give us the pickup and drop off locations and date-time of ride of the user. So we have created new variables to find out the distance travelled by the user based on pickup and drop off locations. We have used **Vincenty's formula** to calculate the elliptical distance accurately.

Moreover the "pickup_datetime" variable is separated to find "year", "month", weekday" and "hour_bin" variables.

1.5. Removing Unnecessary Variables:

We have omited the variables: “pickup_latitude”, “pickup_longitude”, “dropoff_latitude”, “dropoff_longitude”, “pickup_datetime”.

1.6. Exploratory Data Analysis:

It involves visualizing the data for understanding and analyzing purpose and finding different insights.

- **Year:**

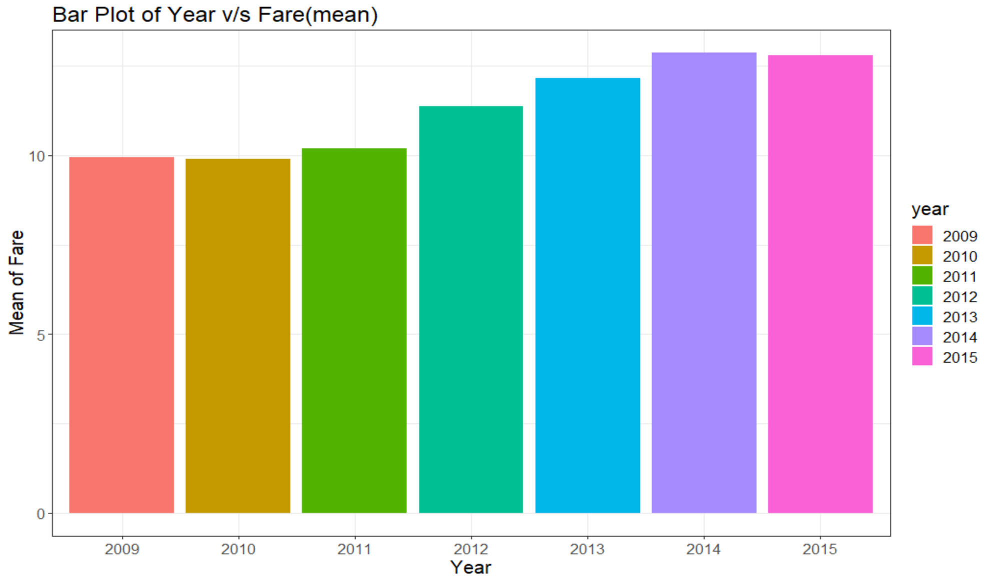


Figure 21: Bar plot of Year v/s Fare (mean)

So we can see that the mean fare has increased from 2012 onwards and a slight decrease in 2015.

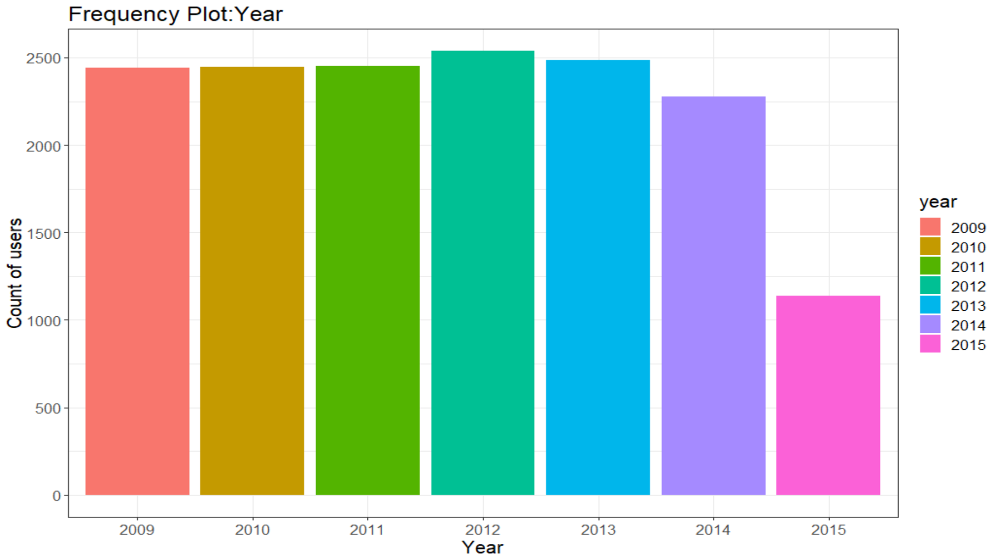


Figure 22: Frequency Plot of Year

Number of users also has a harsh decrease in 2015(or there is less number of observations). So we can say that actual fare has some how increased from 2012 onwards as number of users was almost constant for those years. But due to the increase in fare the number of users is decreasing from 2014 onwards and it has a harsh decrease in 2015.

- **Month:**

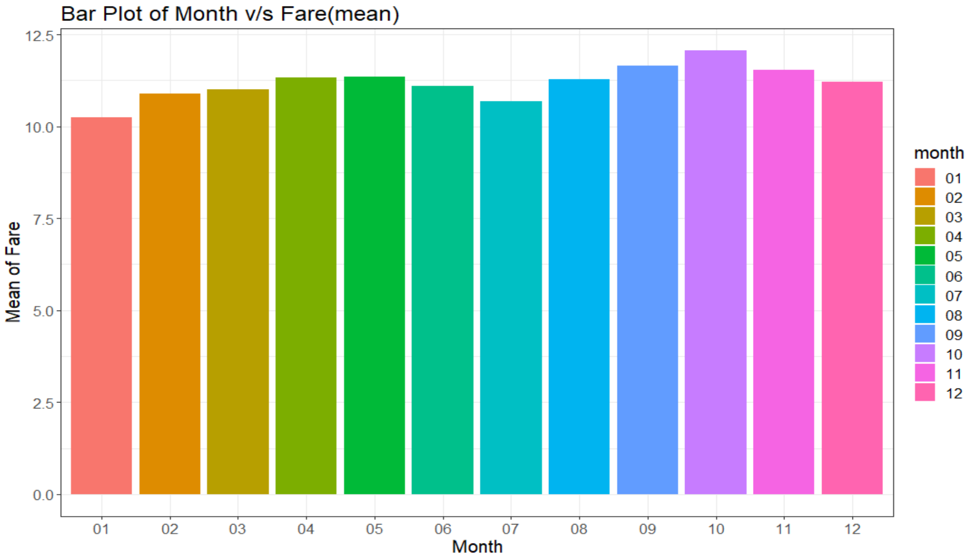


Figure 23: Bar plot of Month v/s Fare (Mean)

So we can see that mean fare is almost same for all months ranging between 10 - 12.5 . The fare is just slightly high from 7th month and highest at 10th.

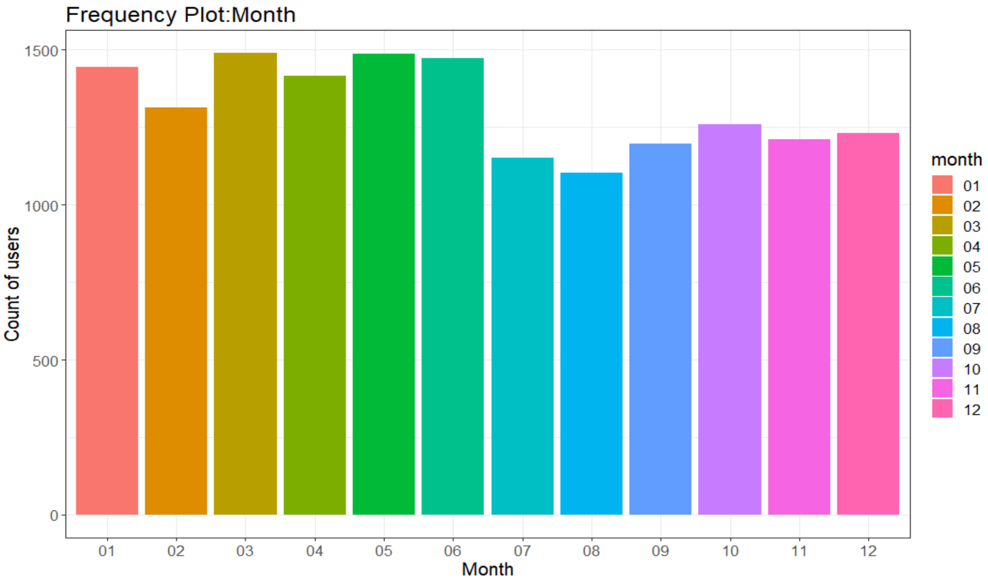


Figure 24: Frequency Plot of Month

So here from the distribution we can see that the no. of users is high for first 6 months and low for next 6 months. Thus, the actual fare is more in the last 6 months compared to the first 6 months.

- **Passenger Count:**

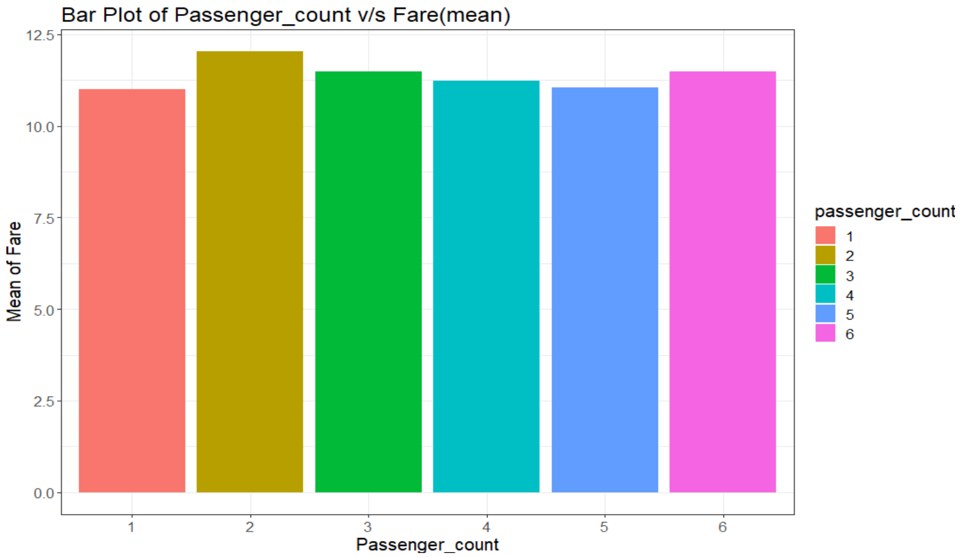


Figure 25: Bar Plot of Passenger Count v/s Fare (Mean)

So mean of fare is almost similar for any passenger group ranging from 11-12 approx.

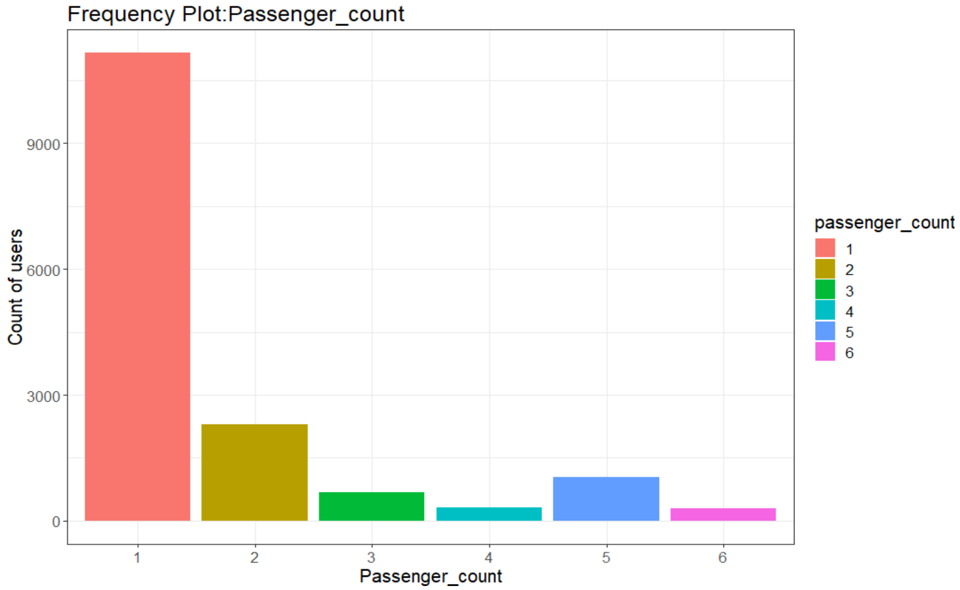


Figure 26: Frequency Plot of Passenger Count

So number of users travelling single is very much higher compared to others.

- **Weekday:**

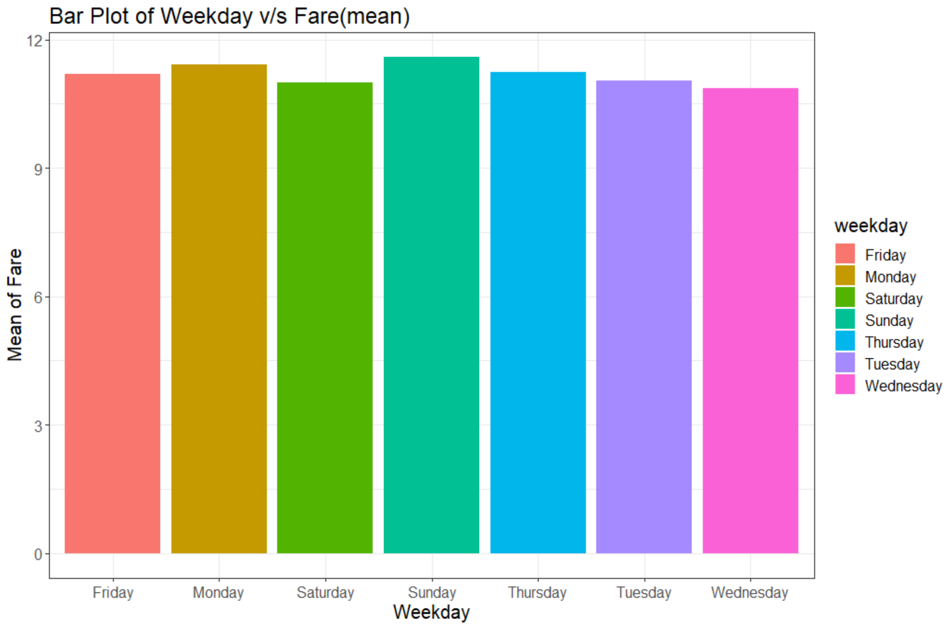


Figure 27: Bar Plot of Weekday v/s Fare (Mean)

So mean of fare is almost equal irrespective of weekday.

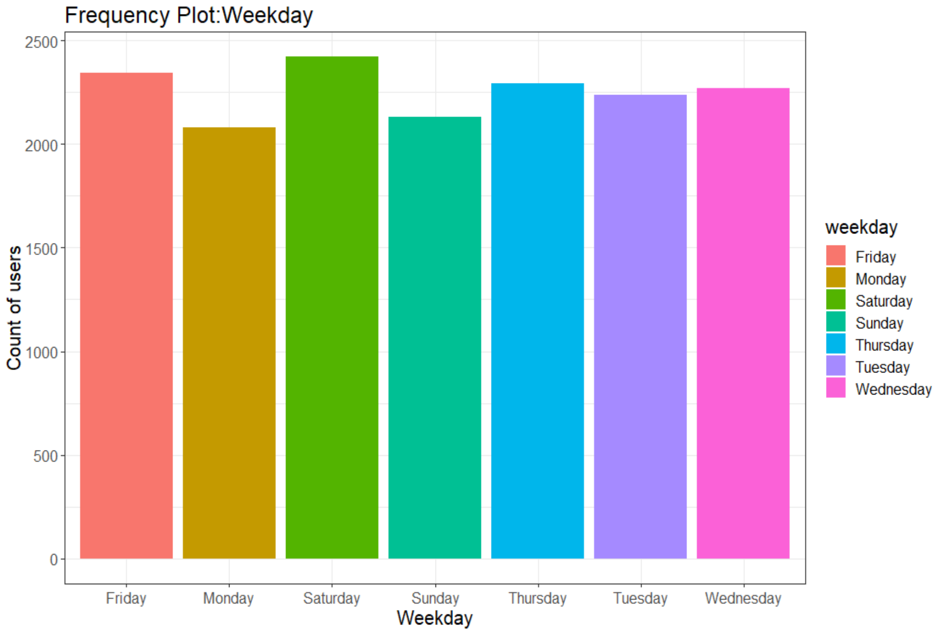


Figure 28: Frequency Plot of Weekday

So number of users is almost similar for everyday and a bit high on Friday and Saturday while lower on Sunday and Monday. So this variable shouldn't have much effect on fare amount.

- **Hour Bin:**

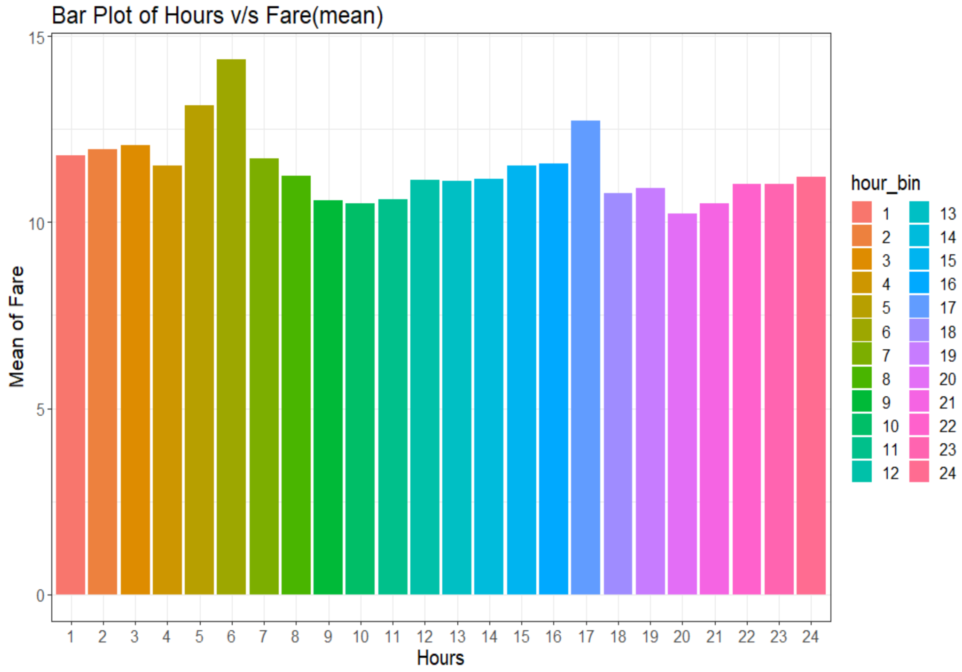


Figure 29: Bar Plot of Hour v/s Fare (Mean)

Mean of fare is highest at 6am and other notable higher fares are at 5am, 3pm and 5pm.

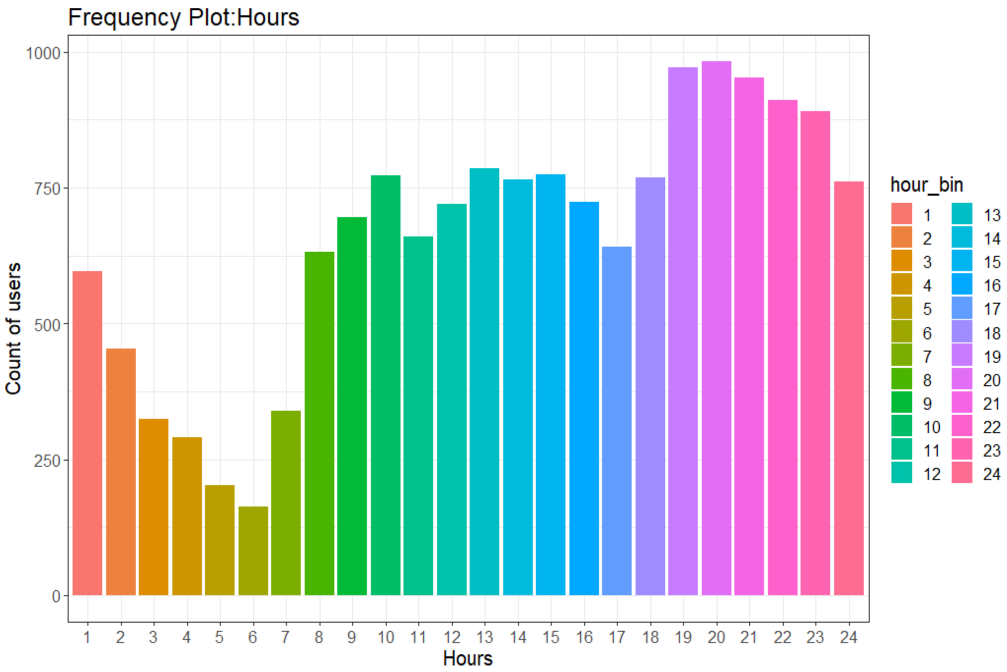


Figure 30: Frequency Plot of Hour Bin

The number of users is increasing from 6pm and highest during 8pm and then has a decreasing slope having lowest value during 6am. This increase in number of users after 6pm might be due office hours. So we can say that the actual fare remains high after midnight till 8am and becomes very high at 6am.

- **Distance:**

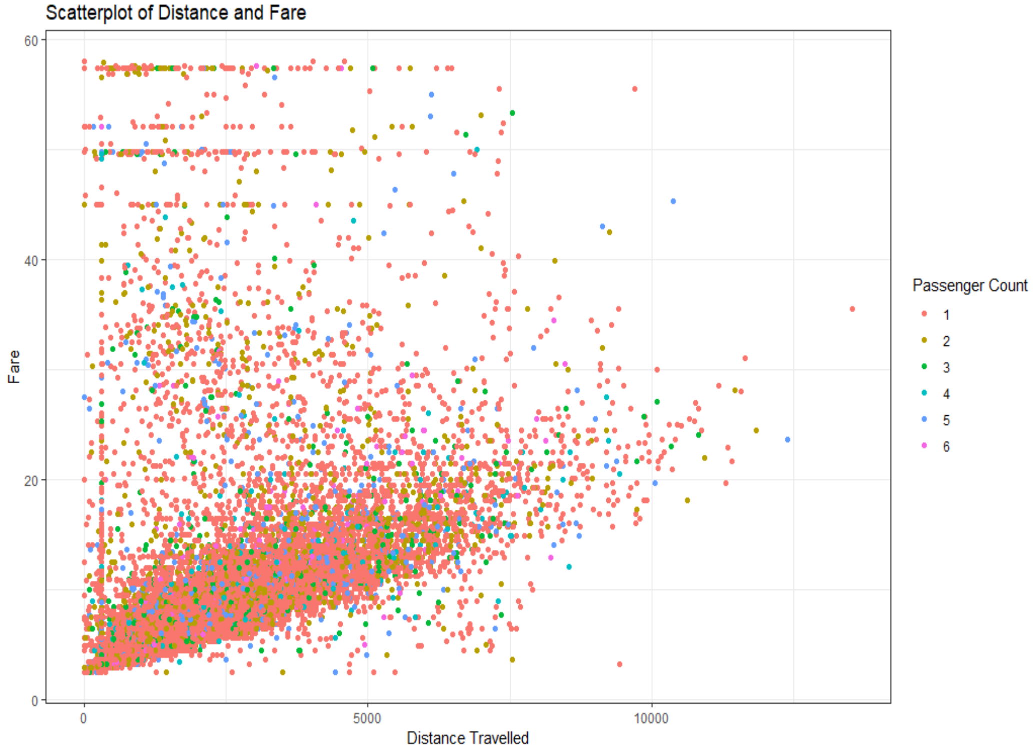


Figure 31: Scatter Plot of Distance v/s Fare

From the scatter plot we can as that on increasing distance the fare increases so they have a linear relationship with each other.

- **Fare Amount:**

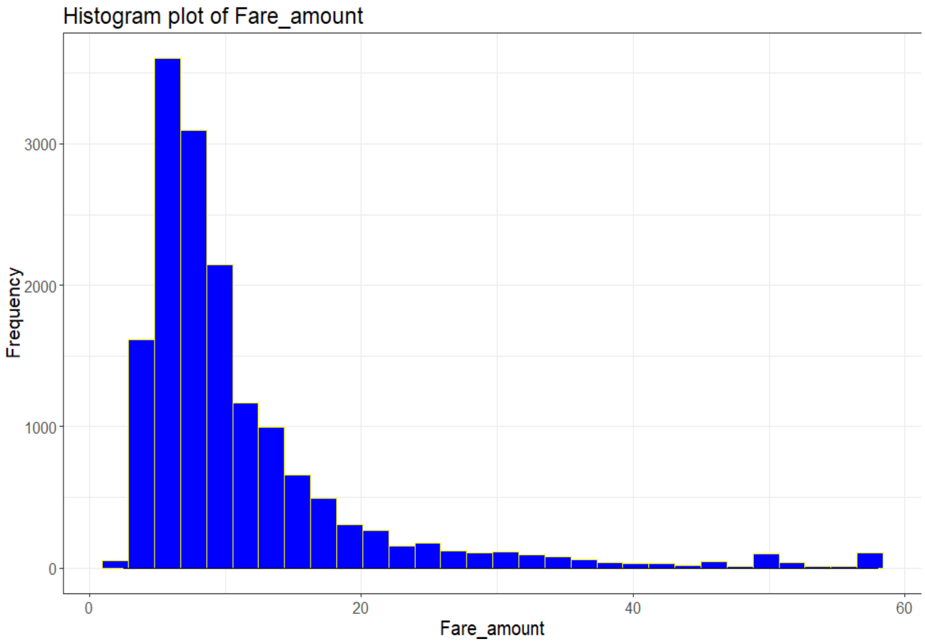


Figure 32: Histogram Plot of Fare Amount

We can clearly see that the distribution is right skewed.

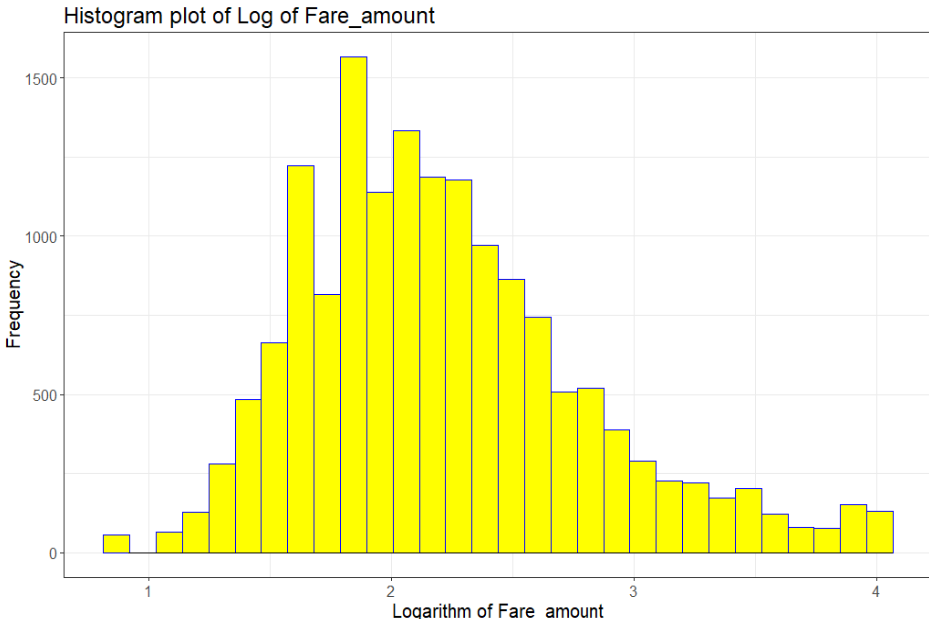


Figure 33: Histogram Plot of logarithm of Fare Amount

Taking the natural log value gives us a better distribution very close to bell shape. So we have considered logarithm of fare amount.

1.7. Feature Selection:

1.7.1. Correlation Analysis:

It is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables. This type of analysis is done to check the multi collinearity effect. If two or more independent variables are strongly correlated then only one of them is enough to predict the dependent variable so, others need to be removed. While a strong correlation between a dependent and independent variable is highly appreciable.

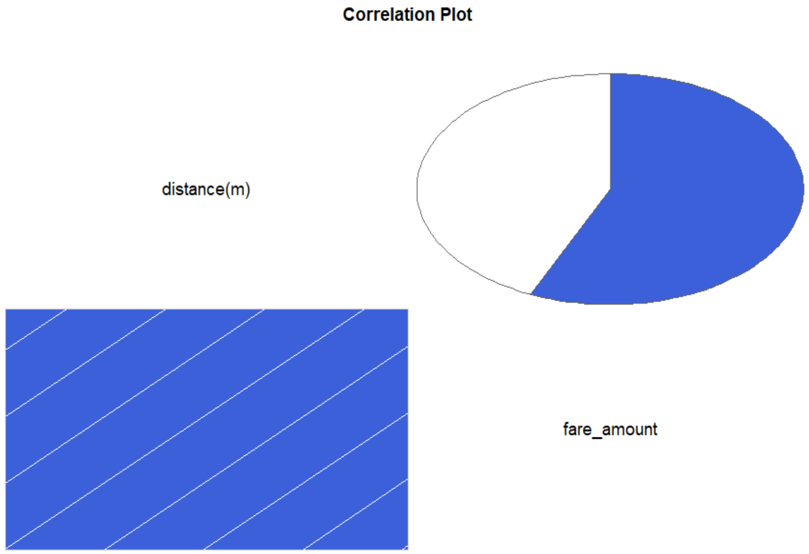
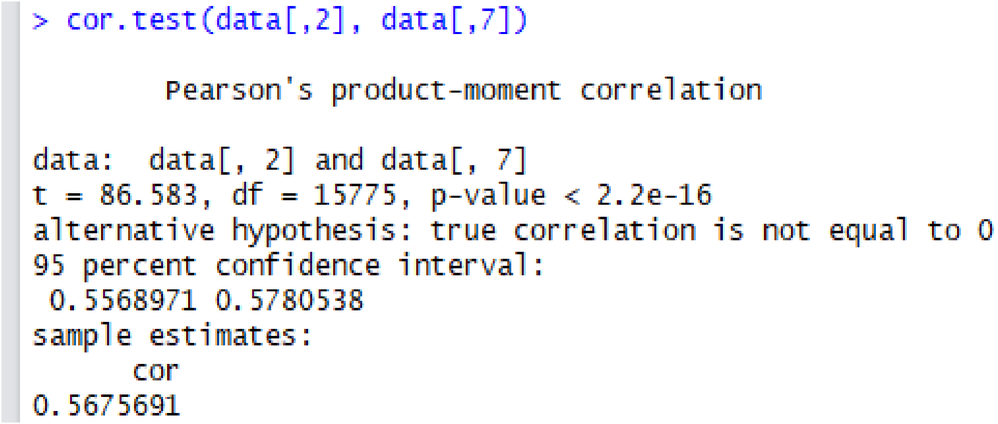


Figure 34: Correlation Plot



```
> cor.test(data[,2], data[,7])

        Pearson's product-moment correlation

data:  data[, 2] and data[, 7]
t = 86.583, df = 15775, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5568971 0.5780538
sample estimates:
      cor
0.5675691
```

Figure 35: Correlation value of Fare and Distance

Here the correlation coefficient of “distance(m)” and “fare_amount” is 0.57. So we can say that there is a moderate positive correlation between them.

And the p-value being less than 0.05 we can say that “distance(m)” is a significant predictor of “fare_amount”.

1.7.2. ANOVA (Analysis of Variance) Test:

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. We can use ANOVA to prove/disprove if all the medication treatments were equally effective or not.

We have applied ANOVA test for our variables and found the following insights:

- **Population Count:**

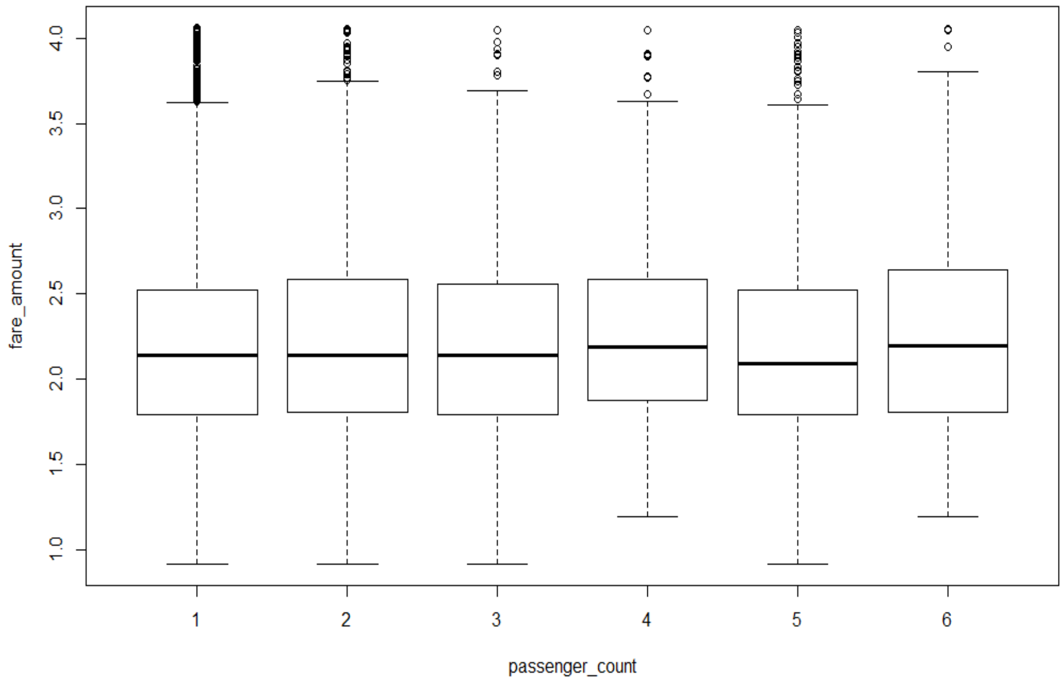
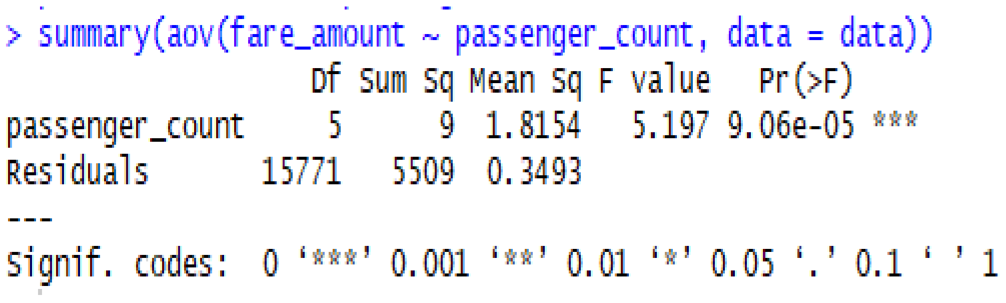


Figure 36: Box Plot of Fare v/s Population Count

From the box plot it is evident that not all population means, across the groups, are equal. So we can reject the Null Hypothesis (H0). Thus the variable is significant in explaining the variance of dependent variable.



```
> summary(aov(fare_amount ~ passenger_count, data = data))
                   Df Sum Sq Mean Sq F value   Pr(>F)    
passenger_count     5      9  1.8154   5.197 9.06e-05 ***
Residuals       15771   5509  0.3493                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 37: ANOVA Test of Passenger Count

p-value less than 0.05 indicates the importance of the variable.

- **Year:**

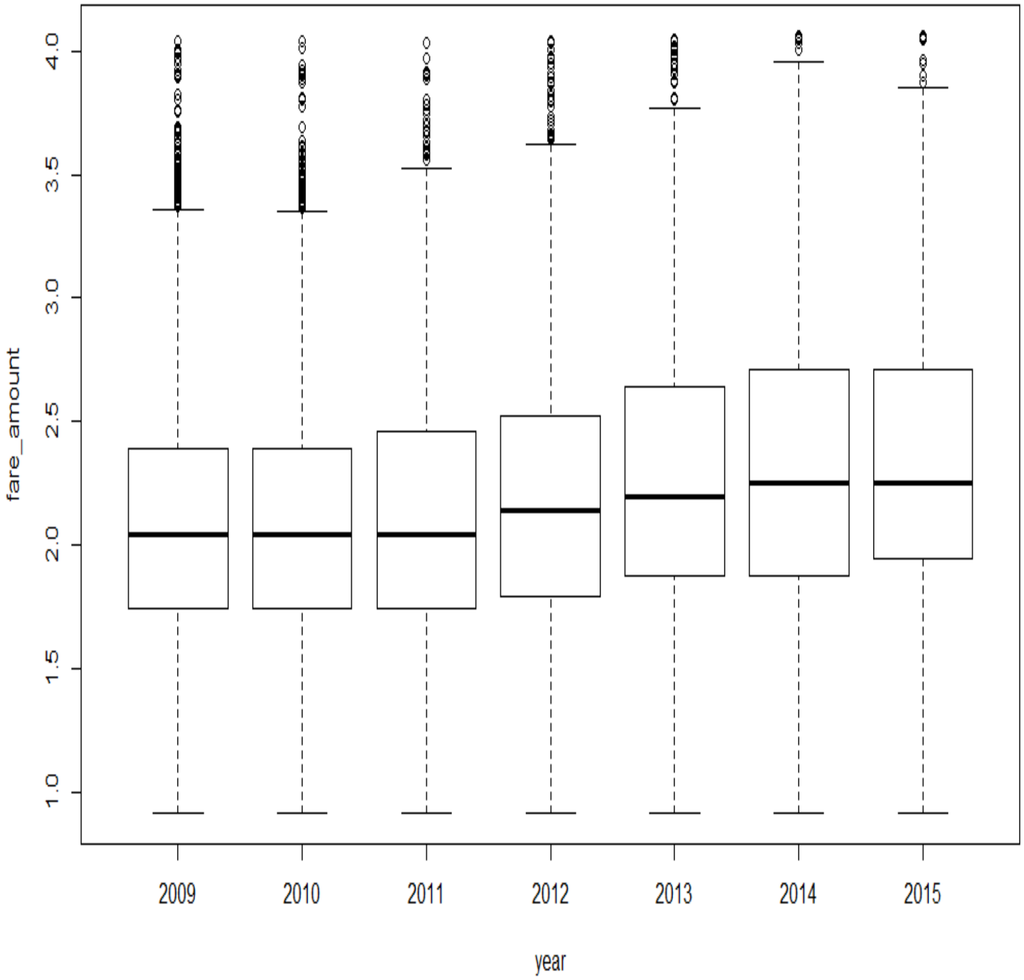
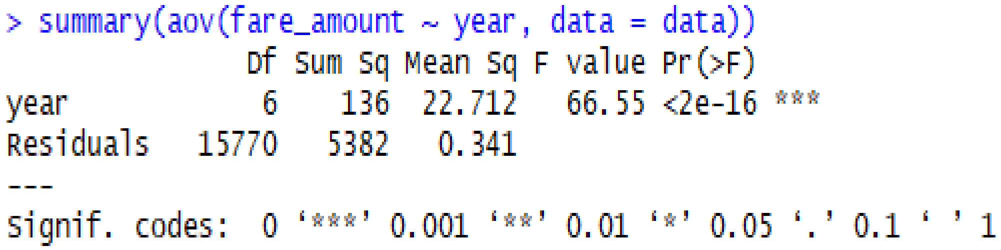


Figure 38: Box Plot of Fare v/s Year

From the box plot it is evident that not all population means, across the groups, are equal. So we can reject the Null Hypothesis (H0). Thus the variable is significant in explaining the variance of dependent variable.



```
> summary(aov(fare_amount ~ year, data = data))
                Df Sum Sq Mean Sq F value Pr(>F)    
year             6    136  22.712   66.55 <2e-16 ***
Residuals    15770   5382   0.341                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 39: ANOVA Test of Year

p-value less than 0.05 indicates the importance of the variable.

- **Month:**

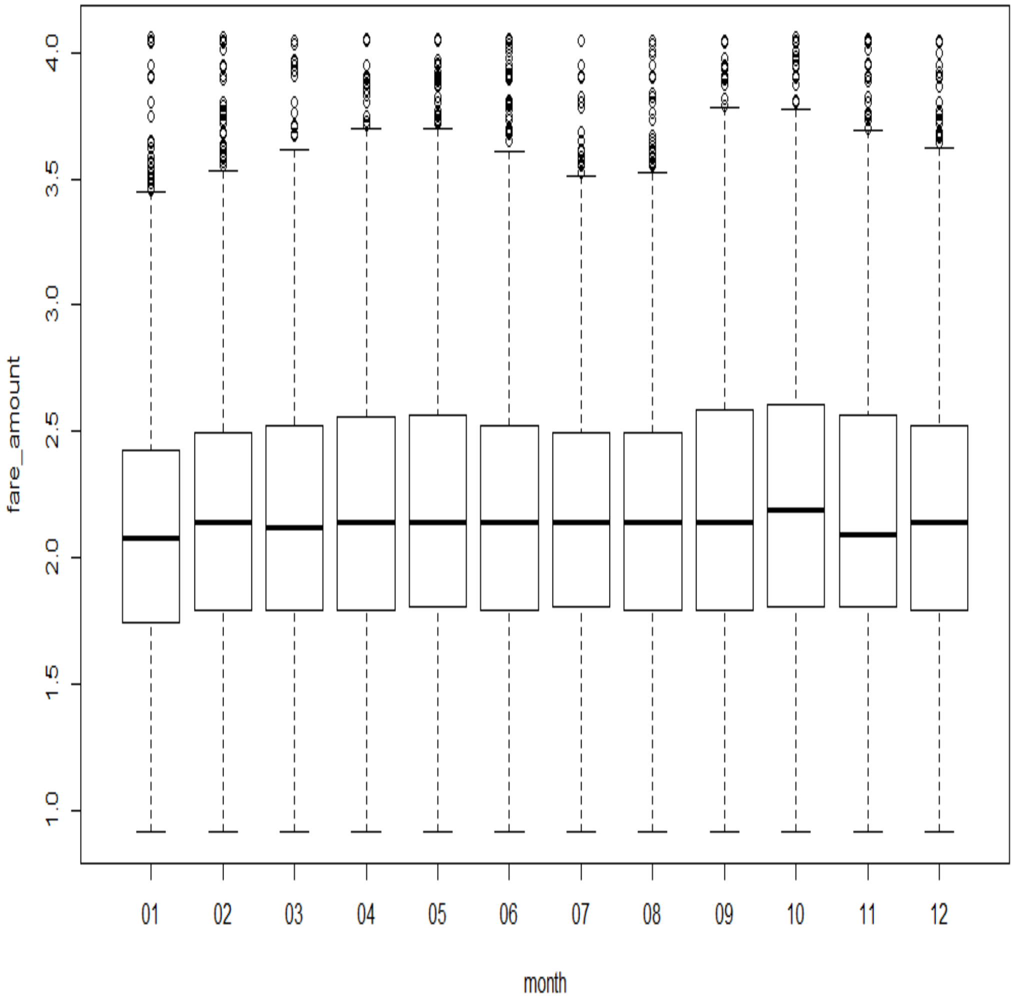
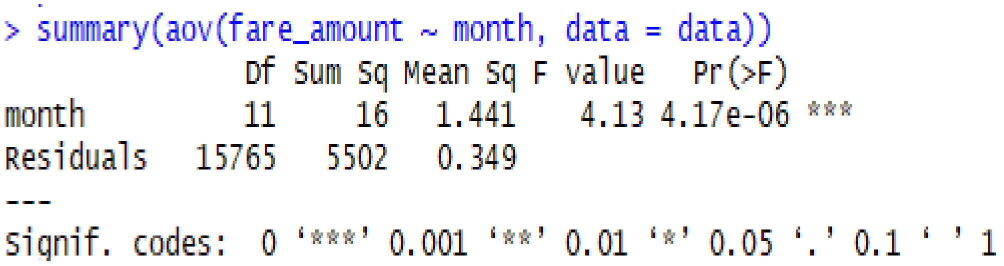


Figure 40: Box Plot of Fare v/s Month

From the box plot it is evident that not all population means, across the groups, are equal. So we can reject the Null Hypothesis (H0). Thus the variable is significant in explaining the variance of dependent variable.



```
> summary(aov(fare_amount ~ month, data = data))
               Df Sum Sq Mean Sq F value   Pr(>F)    
month          11     16   1.441    4.13 4.17e-06 ***
Residuals   15765   5502   0.349                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 41: ANOVA Test of Year

p-value less than 0.05 indicates the importance of the variable.

- **Weekday:**

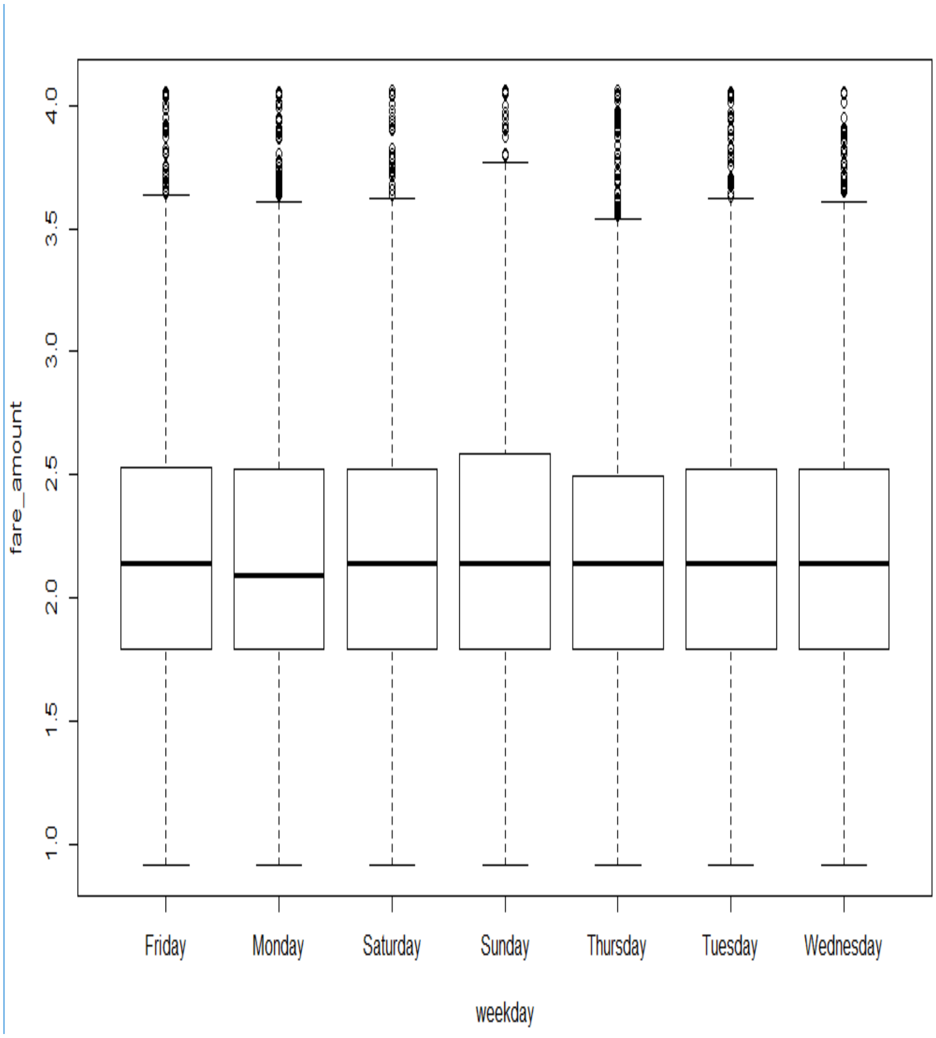
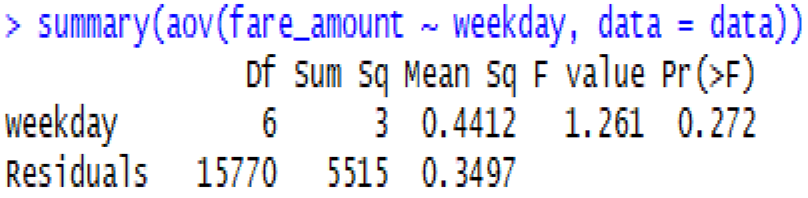


Figure 42: Box Plot of Fare v/s Weekday

In this case we can see that all the population means across the groups are almost equal. So we do not have enough evidence to reject the Null Hypothesis (H0). Thus this variable can’t explain the variance of the dependent variable significantly.



```
> summary(aov(fare_amount ~ weekday, data = data))
               Df Sum Sq Mean Sq F value Pr(>F)
weekday         6      3  0.4412   1.261  0.272
Residuals   15770   5515  0.3497
```

Figure 43: ANOVA Test of Weekday

P-value being greater than 0.05 indicates that the variable is not much important to our analysis.

- **Hour Bin:**

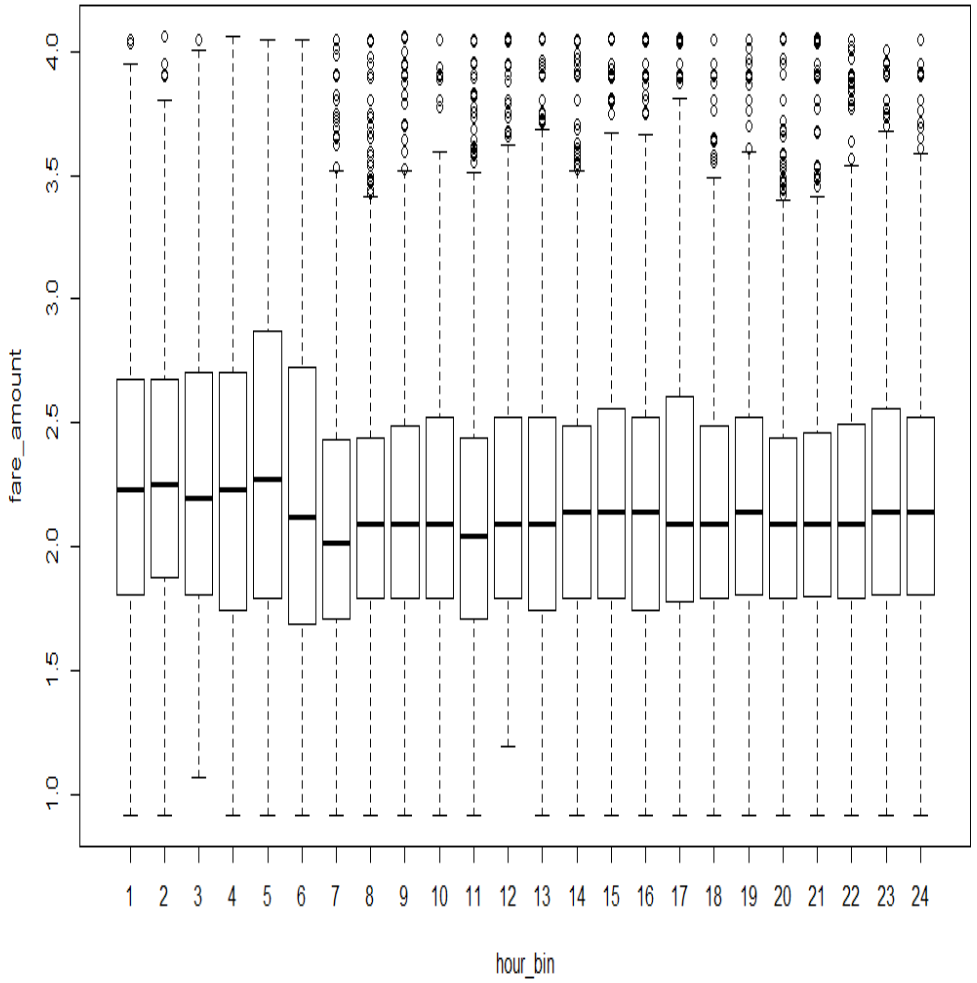
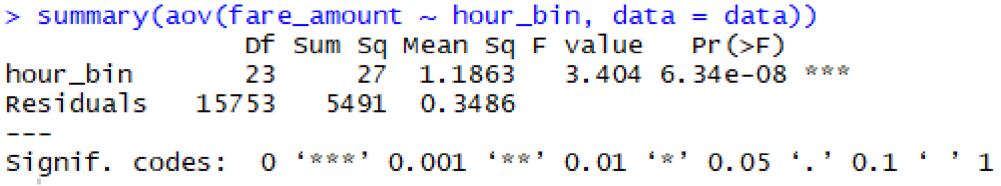


Figure 44: Box Plot of Fare v/s Hour Bin

From the box plot it is evident that not all population means, across the groups, are equal. So we can reject the Null Hypothesis (H0). Thus the variable is significant in explaining the variance of dependent variable.



```
> summary(aov(fare_amount ~ hour_bin, data = data))
               Df Sum Sq Mean Sq F value   Pr(>F)
hour_bin       23     27  1.1863   3.404 6.34e-08 ***
Residuals   15753   5491  0.3486
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 45: ANOVA Test of Hour Bin

p-value less than 0.05 indicates the importance of the variable.

From the ANOVA test we can see that only the “weekday” variable is having p-value less than 0.05. So we need to delete that variable for further consideration.

1.8. Feature Scaling:

Feature scaling is a method used to standardize the range of independent variables or features of data.

But here as we have one numerical independent variable so feature scaling is irrelevant.

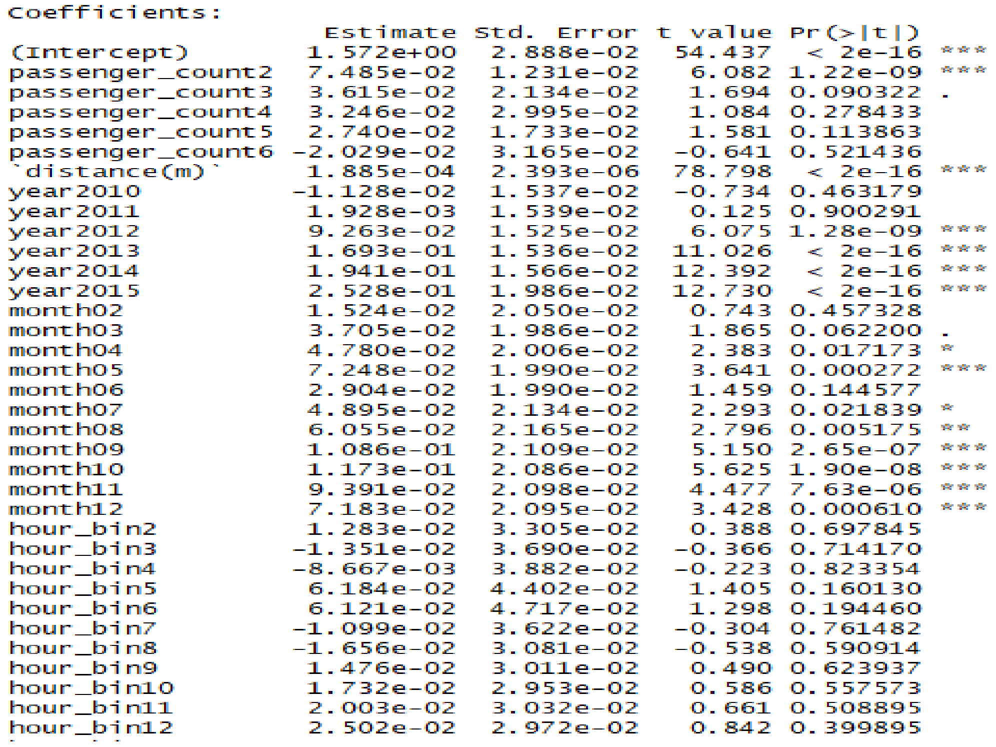
2. Model Building

As we have prepared and cleaned the data, now the next step involves feeding the data to model and train it. But to check the error metrics we have divided the dataset into training and test sets. But dividing randomly may generate low instances of records for some categorical variables. So to fix that we have applied Stratified Sampling with “passenger_count” as the categorical variable. The training set consisting of 80percent of the data is used to train the model and predict on the independent variables of the test set. Then the predicted values are compared to the dependent variable of the test set to calculate the error metrics.

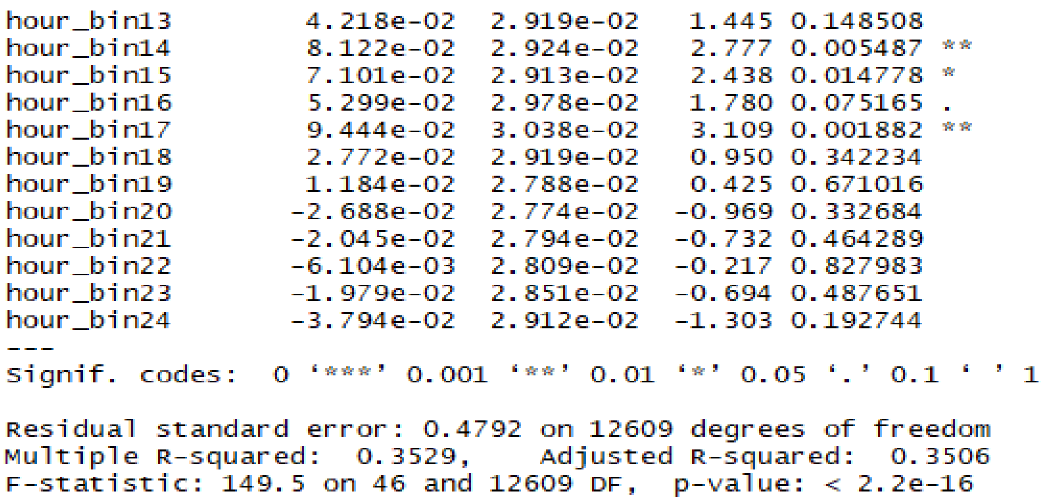
Here we have to predict the fare of the cab ride. Fare being a continuous numerical variable we have to apply different regression models on the data. We have applied several regression models from simple to complex one by one and then at last we have compared the results.

2.1. Multiple Linear Regression:

At first multiple linear regression model is applied to the training set and we got the following result.



```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         1.572e+00  2.888e-02  54.437  < 2e-16 ***
passenger_count2    7.485e-02  1.231e-02   6.082 1.22e-09 ***
passenger_count3    3.615e-02  2.134e-02   1.694 0.090322 .  
passenger_count4    3.246e-02  2.995e-02   1.084 0.278433    
passenger_count5    2.740e-02  1.733e-02   1.581 0.113863    
passenger_count6   -2.029e-02  3.165e-02  -0.641 0.521436    
`distance(m)`       1.885e-04  2.393e-06  78.798  < 2e-16 ***
year2010           -1.128e-02  1.537e-02  -0.734 0.463179    
year2011            1.928e-03  1.539e-02   0.125 0.900291    
year2012            9.263e-02  1.525e-02   6.075 1.28e-09 ***
year2013            1.693e-01  1.536e-02  11.026  < 2e-16 ***
year2014            1.941e-01  1.566e-02  12.392  < 2e-16 ***
year2015            2.528e-01  1.986e-02  12.730  < 2e-16 ***
month02             1.524e-02  2.050e-02   0.743 0.457328    
month03             3.705e-02  1.986e-02   1.865 0.062200 .  
month04             4.780e-02  2.006e-02   2.383 0.017173 *  
month05             7.248e-02  1.990e-02   3.641 0.000272 ***
month06             2.904e-02  1.990e-02   1.459 0.144577    
month07             4.895e-02  2.134e-02   2.293 0.021839 *  
month08             6.055e-02  2.165e-02   2.796 0.005175 ** 
month09             1.086e-01  2.109e-02   5.150 2.65e-07 ***
month10             1.173e-01  2.086e-02   5.625 1.90e-08 ***
month11             9.391e-02  2.098e-02   4.477 7.63e-06 ***
month12             7.183e-02  2.095e-02   3.428 0.000610 ***
hour_bin2           1.283e-02  3.305e-02   0.388 0.697845    
hour_bin3          -1.351e-02  3.690e-02  -0.366 0.714170    
hour_bin4          -8.667e-03  3.882e-02  -0.223 0.823354    
hour_bin5           6.184e-02  4.402e-02   1.405 0.160130    
hour_bin6           6.121e-02  4.717e-02   1.298 0.194460    
hour_bin7          -1.099e-02  3.622e-02  -0.304 0.761482    
hour_bin8          -1.656e-02  3.081e-02  -0.538 0.590914    
hour_bin9           1.476e-02  3.011e-02   0.490 0.623937    
hour_bin10          1.732e-02  2.953e-02   0.586 0.557573    
hour_bin11          2.003e-02  3.032e-02   0.661 0.508895    
hour_bin12          2.502e-02  2.972e-02   0.842 0.399895    
```



```
hour_bin13        4.218e-02  2.919e-02   1.445 0.148508    
hour_bin14        8.122e-02  2.924e-02   2.777 0.005487 ** 
hour_bin15        7.101e-02  2.913e-02   2.438 0.014778 *  
hour_bin16        5.299e-02  2.978e-02   1.780 0.075165 .  
hour_bin17        9.444e-02  3.038e-02   3.109 0.001882 ** 
hour_bin18        2.772e-02  2.919e-02   0.950 0.342234    
hour_bin19        1.184e-02  2.788e-02   0.425 0.671016    
hour_bin20       -2.688e-02  2.774e-02  -0.969 0.332684    
hour_bin21       -2.045e-02  2.794e-02  -0.732 0.464289    
hour_bin22       -6.104e-03  2.809e-02  -0.217 0.827983    
hour_bin23       -1.979e-02  2.851e-02  -0.694 0.487651    
hour_bin24       -3.794e-02  2.912e-02  -1.303 0.192744    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4792 on 12609 degrees of freedom
Multiple R-squared:  0.3529,	Adjusted R-squared:  0.3506 
F-statistic: 149.5 on 46 and 12609 DF,  p-value: < 2.2e-16
```

Figure 46: Multiple Linear Regression

As you can see the Adjusted R-squared value, we can explain only about 35% of the data using our multiple linear regression model. This is not very impressive, but at least looking at the F-statistic and combined p-value we can reject the Null Hypothesis that target variable does not depend on any of the predictor variables.

- **VIF (Variance Inflation Factor):**

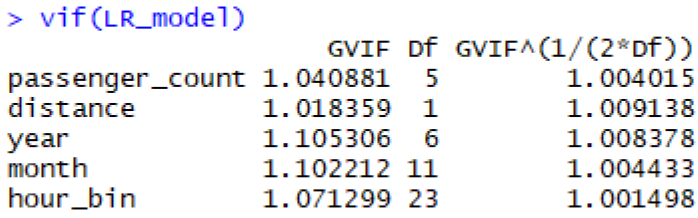
In statistics, the variance inflation factor (VIF) is the ratio of variance in a model with multiple terms, divided by the variance of a model with one term alone. It is a measure of multi-collinearity in a regression design matrix. The formula for for determining VIF is:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

Where r^2 is the coefficient of determination

It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity. The ideal value of VIF should be 1. If the value is between 5 to 10 then there is presence of multi-collinearity among the variables. But if the value exceeds 10 then there is very high multi-collinearity and it should be taken care of.

The values of VIF for our model is shown below:



```
> vif(LR_model)
                    GVIF Df GVIF^(1/(2*Df))
passenger_count 1.040881  5        1.004015
distance        1.018359  1        1.009138
year            1.105306  6        1.008378
month           1.102212 11        1.004433
hour_bin        1.071299 23        1.001498
```

Figure 47: VIF

From the above table we can see that the VIF values are close to one which ensures that multi-collinearity doesn't exists among the variables. Thus we have considered all the variables for our further models.

2.2. Decision Tree Regression:

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. The decision tree algorithm is mainly based on Information Gain. It will select that parameter as the parent node which will have more information gain value.

Information gain is the difference between information entropy of the system before splitting and information entropy of the system after splitting. Information entropy is the average rate at which information is produced.

$$S = -\sum_i P_i \log P_i$$

Where P_i is the probability of occurrence of dependent variable.

Now we have applied decision tree regression model to predict our fare_amount target variable. The decision tree is shown below:

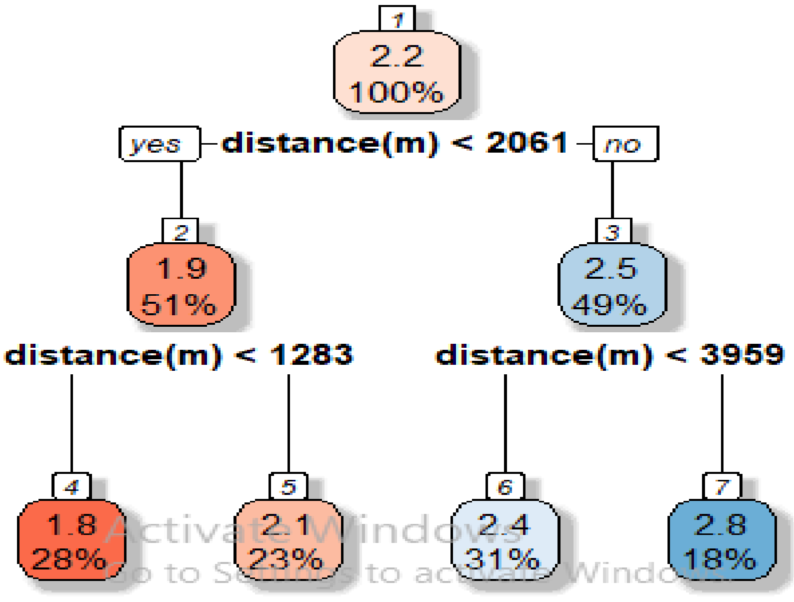


Figure 48: Decision Tree

2.3. Random Forest Regression:

Random Forest is an ensemble that consists of many decision trees. To build each decision tree we use the different portion of the whole data. This reduces error and increases accuracy. The idea behind the Random Forest is that a single decision tree may not be able to explain the variance of the whole data set, so, we use many trees to extract as much variance as possible. The Random Forest algorithm uses the Gini Index to select the parent nodes.

$$\textbf{Gini} = 1 - \sum(Pi)^2$$

Gini Index measures the amount of impurity of the data. It selects that variable whose Gini Index is lowest. Then for each node it will randomly select few variables(m) to build the first tree and this must be very much less than the number of variables(M) and may be based on the formula (m = sqrt.(M)).

Then the tree takes the bootstrap sample, i.e.- it randomly selects 67% of the observation for training and the remaining 33% for testing. This is called ‘Out of Bag’ sample method. Then it applies the CART algorithm on the training data, to predict the class of the test data and thus the error of the tree is estimated comparing the actual and the predicted values. Then whatever observation is misclassified is fed to the next decision tree. Then it will keep on splitting until it finds the leaf node based on the error rate. It will build trees until the error no longer decreases. When the same error value will repeat it will stop growing the trees.

We have applied the random forest algorithm to the model without mentioning the number of tree so that it can grow until it finds the lowest error value.

2.4. SVR (Support Vector Regression):

SVR is a type of model in which we try to set the error within a certain threshold while in linear regression we try to minimise the error rate. Support Vector Machine can be applied not only to classification problems but also to the case of regression. Still it contains all the main features that characterize maximum margin algorithm: a non-linear function is leaned by linear learning machine mapping into high dimensional kernel induced feature space. The capacity of the system is controlled by parameters that do not depend on the dimensionality of feature space. In the same way as with classification approach there is motivation to seek and optimize the generalization bounds given for regression. They relied on defining the loss function that ignores errors, which are situated within the certain distance of the true value. This type of function is often called – epsilon intensive – loss function.

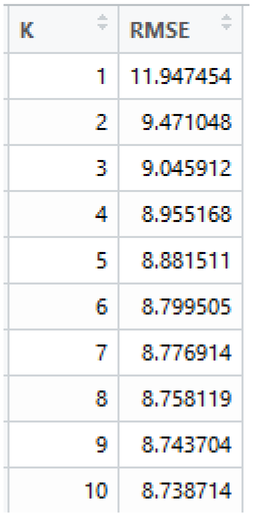
We have applied SVR on our training set and predicted the values on test set.

2.5. KNN (K-Nearest Neighbours):

The algorithm uses **‘**feature similarity**’** to predict values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. The first step is to calculate the distance between the new point and each training point. There are various methods for calculating this distance, of which the most commonly known methods are – Euclidian, Manhattan and Hamming distance. Then the value of k is user defined which denotes the number of nearest neighbours that should be considered when a new point comes.

It is a lazy learning method, where the function is only approximated locally; i.e.- the model doesn’t saves the value of training data like other algorithms instead the value is determined instantly when a new point comes.

One of the challenging task for KNN algorithm is to find the value of K. Here we have imputed different values of ‘K’ for the algorithm and computed their respective rmse values.



K	RMSE
1	11.947454
2	9.471048
3	9.045912
4	8.955168
5	8.881511
6	8.799505
7	8.776914
8	8.758119
9	8.743704
10	8.738714

Figure 49: K_values

From the above table we can see that the value of rmse is decreasing on increasing K-value. But if we take a very low k value then the model over fits the training data, which results in high error rate on the validation side. On the other hand for a high value of K the model performs poorly on both train and test data sets. One way to find a descent value of K is to plot the rmse values with their respective K values.

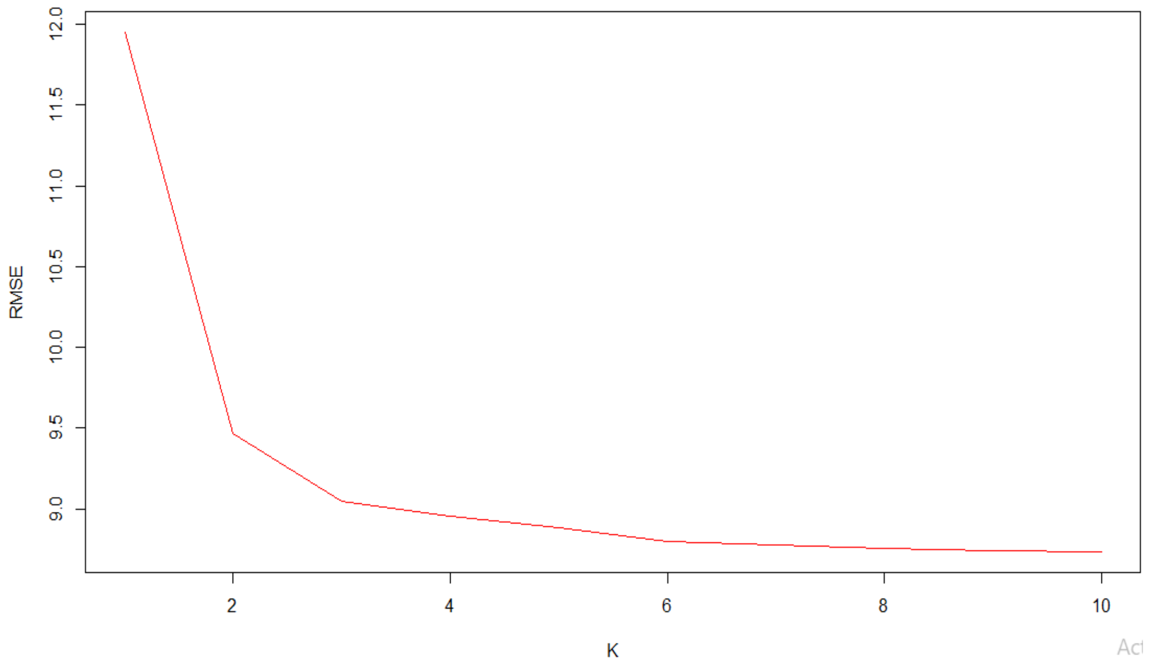


Figure 50: Elbow Curve

The above curve is also known as Elbow Curve. From the plot we can see that the elbow appears at a value of k = 3. But here as we want lower error for our model so we have imputed k = 6 for our model to predict the test cases.

Now of the different regression models we will choose the best model by comparing error metrics.

Chapter 3

CONCLUSION

1. Model Evaluation:

Now that we have a few models for predicting the target variable and we need to decide which one to choose. There are several methods by which we can compare the models. As the dependent variable is a continuous regression model so we have compared the models based on different error metrics.

a) Mean Absolute Percentage Error (MAPE):

It 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 expresses accuracy as a percentage, and is defined by the formula:

$$\mathrm{M} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|,$$

Where, A_t is the actual value and F_t is the predicted value. The difference between A_t and F_t is divided by the actual value A_t again. The absolute value in this calculation is summed for every predicted point and is divided by the number of total points n. Multiplying by 100% makes it a percentage error.
Lower MAPE indicates better model.

b) Root Mean Squared Error (RMSE):

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. 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.

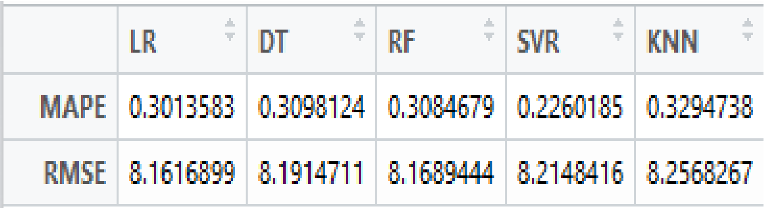
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{model,i})^2}{n}}$$

Where, X_{obs} is observed values and X_{model} is modelled values at time/place i.

Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.

2. Model Selection:

Thus we have calculated the MAPE and RMSE values for prediction of different models using the predicted values of respective models and the dependent variable of the test set. Then we have saved the data in the following table for comparison:



	LR	DT	RF	SVR	KNN
MAPE	0.3013583	0.3098124	0.3084679	0.2260185	0.3294738
RMSE	8.1616899	8.1914711	8.1689444	8.2148416	8.2568267

Figure 51: Error Metrics

From the above table we can clearly see that there is no major difference between the RMSE values of different models. Thus we can take this metrics out of our comparison and select a model based on MAPE. The MAPE value of the SVR model is significantly lower than the other models.

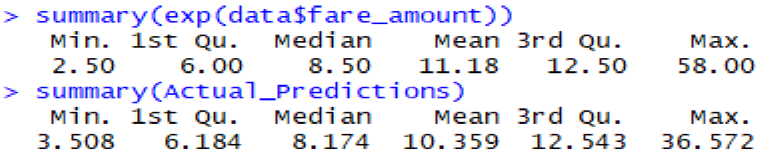
So we have chosen the SVR model as the best model for our dataset.

3. Prediction of Results:

We have applied the required data preprocessing steps for the test data as well, because the model can only predict if the training and test set data have similar variables. Then we have trained our model on the whole data set – "train_cab.csv". Then we have predicted the result on "test.csv" data. This result gives us the predicted values in the form of natural logarithm of original values. So we have applied the exponential operation to convert them to actual fare amount.

4. Conclusion:

At last we have compared the statistical metrics of the predicted results with that of the dependent variable from the training dataset and got the following results:



```
> summary(exp(data$fare_amount))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.50    6.00    8.50   11.18   12.50   58.00 
> summary(Actual_Predictions)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.508   6.184   8.174  10.359  12.543  36.572 
```

Figure 52: Statistical Comparison

As almost all the statistical metrics are close to each other except the maximum value which might be due to values of the dependent variable present in the original data set, so we can conclude that the cab fare is dependent on distance of travel, year, month, hour of travel and no. of passengers in the cab.

5. References:

- https://en.wikipedia.org/
- https://learning.edwisor.com/
- https://medium.com/
- https://www.statisticshowto.datasciencecentral.com/
- https://www.theanalysisfactor.com/