

PROJECT REPORT

Cab Fare Prediction Model

By:-

Shashwat Pandey

JIIT, Noida

TABLE OF CONTENTS

1. Introduction

- 1.1. Problem Statement
- 1.2. Data

2. Methodology

- 2.1. Pre-Processing
- 2.2. Model Selection
- 2.3. Model Deployment

3. Pre-Processing

- 3.1. Missing Value Analysis
- 3.2. Outlier Analysis
- 3.3. Exploratory Data Analysis
- 3.4. Feature Creation
- 3.5. Visualizations
- 3.6. Feature Selection
- 3.7. Feature Scaling

4. Model Selection

- 4.1. Modelling
- 4.2. Linear Regression
- 4.3. Decision Tree
- 4.4. Random Forest
- 4.5. Support Vector Regression

5. Conclusion

- 5.1. Model Evaluation
- 5.2. Error Metrics
- 5.3. Prediction

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

To start with the analysis, we will first observe the dataset and the attributes governing the dataset. We will then find independent and target variables (which are already given to us here), and then continue with the procedure of data cleaning, processing, modelling etc.

	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

Training Dataset

	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
1	2015-01-27 13:08:24 UTC	-73.97332	40.76381	-73.98143	40.74384	1
2	2015-01-27 13:08:24 UTC	-73.98686	40.71938	-73.99889	40.73920	1
3	2011-10-08 11:53:44 UTC	-73.98252	40.75126	-73.97965	40.74614	1
4	2012-12-01 21:12:12 UTC	-73.98116	40.76781	-73.99045	40.75164	1
5	2012-12-01 21:12:12 UTC	-73.96605	40.78977	-73.98856	40.74443	1
6	2012-12-01 21:12:12 UTC	-73.96098	40.76555	-73.97918	40.74005	1
7	2011-10-06 12:10:20 UTC	-73.94901	40.77320	-73.95962	40.77089	1
8	2011-10-06 12:10:20 UTC	-73.77728	40.64664	-73.98508	40.75937	1
9	2011-10-06 12:10:20 UTC	-74.01410	40.70964	-73.99511	40.74137	1
10	2014-02-18 15:22:20 UTC	-73.96958	40.76552	-73.98069	40.77072	1

Testing Dataset

Number of attributes:

- 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.

METHODOLOGY

2.1 Pre-Processing

When we are required to build a predictive model, we need to look and manipulate the data before we start modelling which includes multiple preprocessing steps such as exploring the data, cleaning the data as well as visualizing the data through graph and plots, all these steps are combined under one shed which is Exploratory Data Analysis, which includes following steps:

- Data exploration and Cleaning
- Missing values treatment
- Outlier Analysis
- Feature Selection
- Features Scaling
 - Skewness and Log transformation
- Visualization

2.2 Model Selection

Once all the Pre-Processing steps have been done on our data set, we will now further move to our next step which is modelling. Modelling plays an important role to find out the good inferences from the data. Choice of models depends upon the problem statement and data set. As per our problem statement and dataset, we will try some models on our preprocessed data and post comparing the output results we will select the best suitable model for our problem. As per our data set following models need to be tested:

- Linear regression
- Decision Tree
- Random forest

2.3 Model Deployment

The final step of our methodology will be the selection of the model based on the different output and results shown by different models. We have multiple parameters which we will study further in our report to test whether the model is suitable for our problem statement or not.

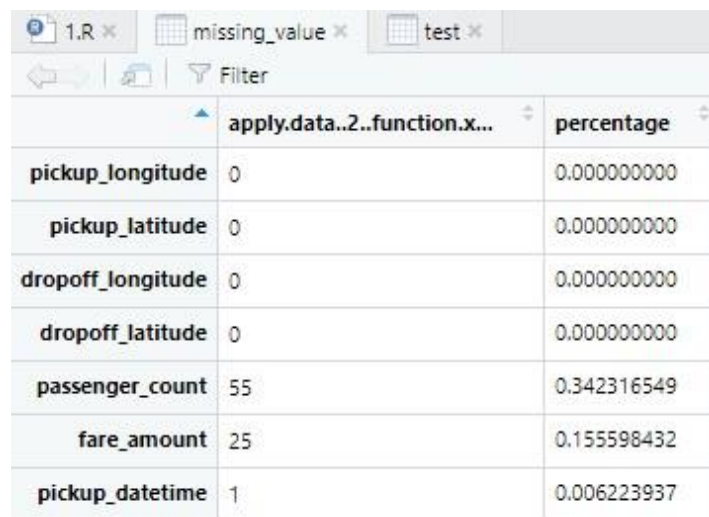
PRE-PROCESSING

3.1 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_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
pickup_datetime	1	0.006223937

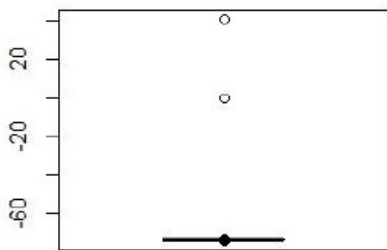
As, we can see that the amount of missing value is too less, i.e.-less than 1% so deleting the rows won't result in any information loss.

3.2 Outlier Analysis

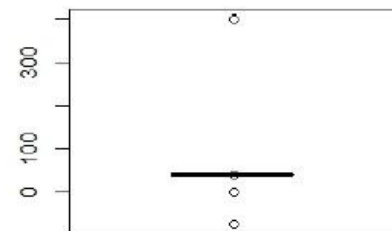
In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

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.

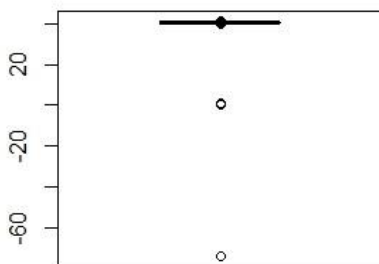
Now we have carried out our investigation further by checking the box plots of the variables.



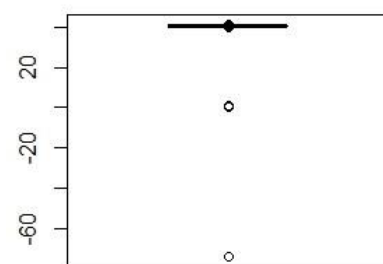
Box-Plot for pickup-longitude



Box-Plot for pickup-latitude



Box-Plot of dropoff-longitude



Box-Plot of dropoff-latitude

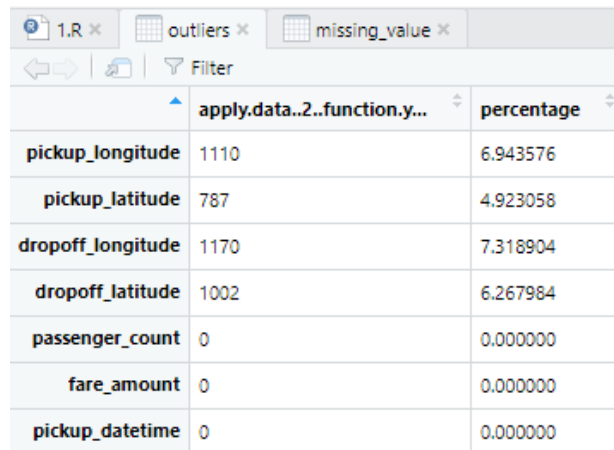
So here the data points above the upper fence and below the lower fence (outside the box plot) show the presence of outliers.

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

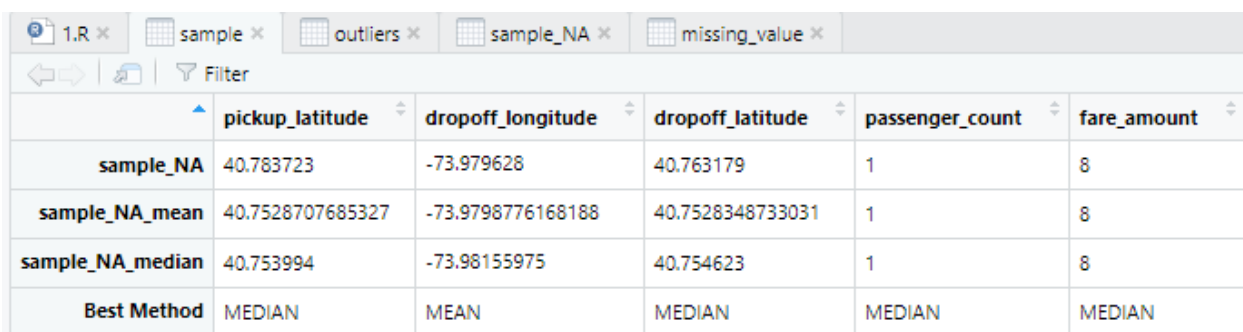
The figure below shows the number of NAs after replacing the outliers with NAs and their percentages with respect to respective columns.



	apply.data..2..function.y...	percentage
pickup_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
pickup_datetime	0	0.000000

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_latitude	dropoff_longitude	dropoff_latitude	passenger_count	fare_amount
sample_NA	40.783723	-73.979628	40.763179	1	8
sample_NA_mean	40.7528707685327	-73.9798776168188	40.7528348733031	1	8
sample_NA_median	40.753994	-73.98155975	40.754623	1	8
Best Method	MEDIAN	MEAN	MEDIAN	MEDIAN	MEDIAN

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.
- 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.

3.3 Exploratory Analysis

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

- Pickup_latitude:** The latitude value should be between +90 to -90
- Pickup_longitude:** The longitude value should be between +180 to -180
- 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. Hence, deleted the 118 observations with distance as 0.
- 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.
- 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.

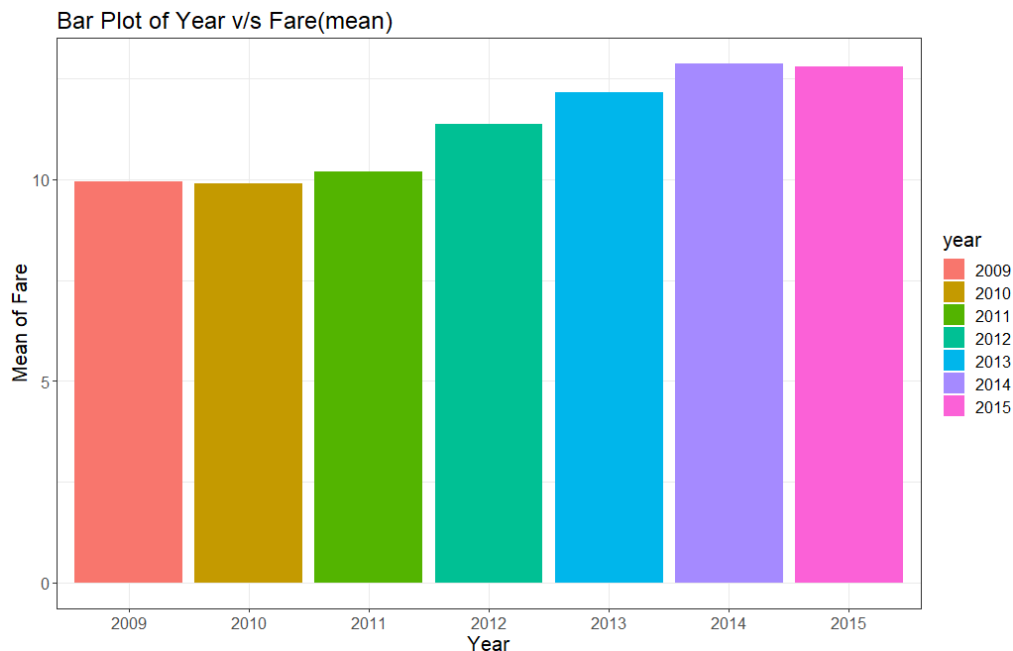
- a. We have got 38 rows where the fare amount is greater than 58 and we have deleted such instances as those fares appear only few times so the actual existence of those values is highly suspicious.
- b. We have found 6 cases where the fare amount is less than 6 and even 0 or negative in some cases.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

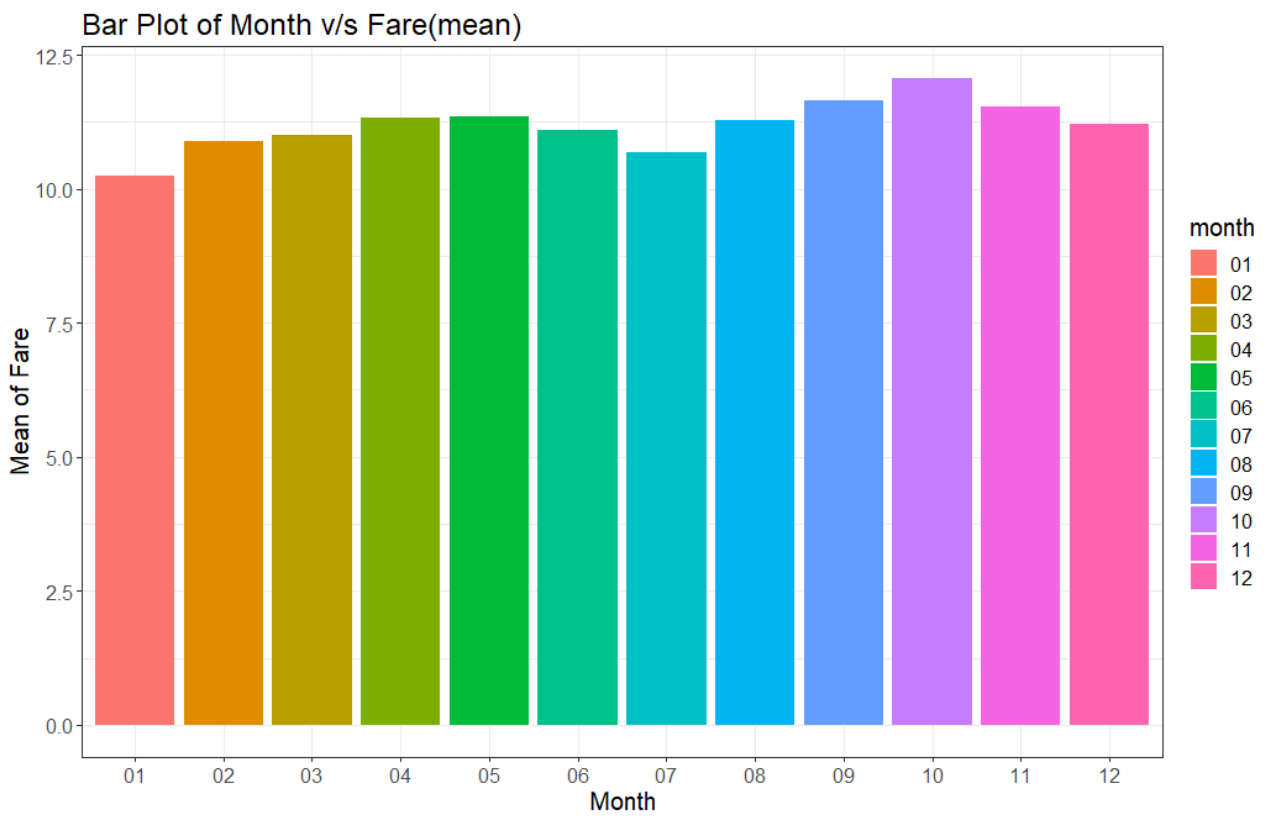
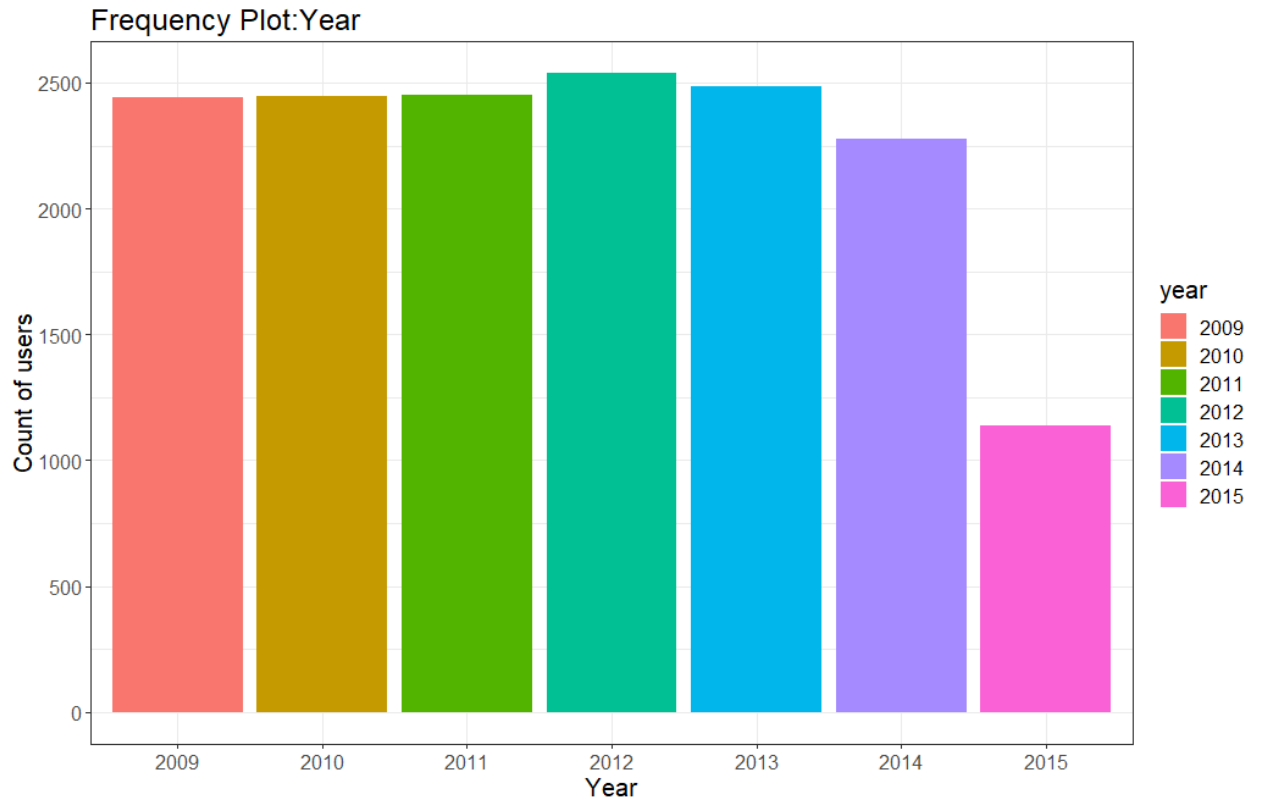
3.4 Feature Creation

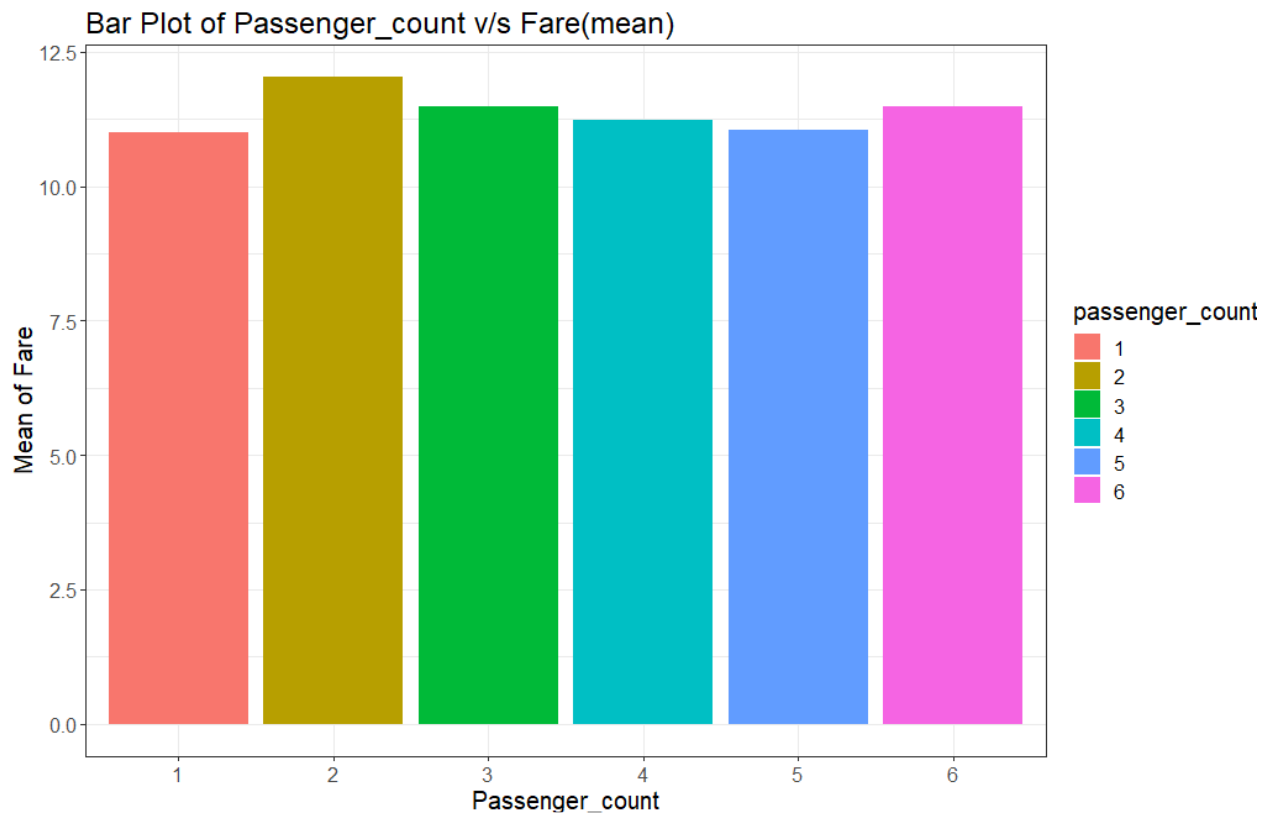
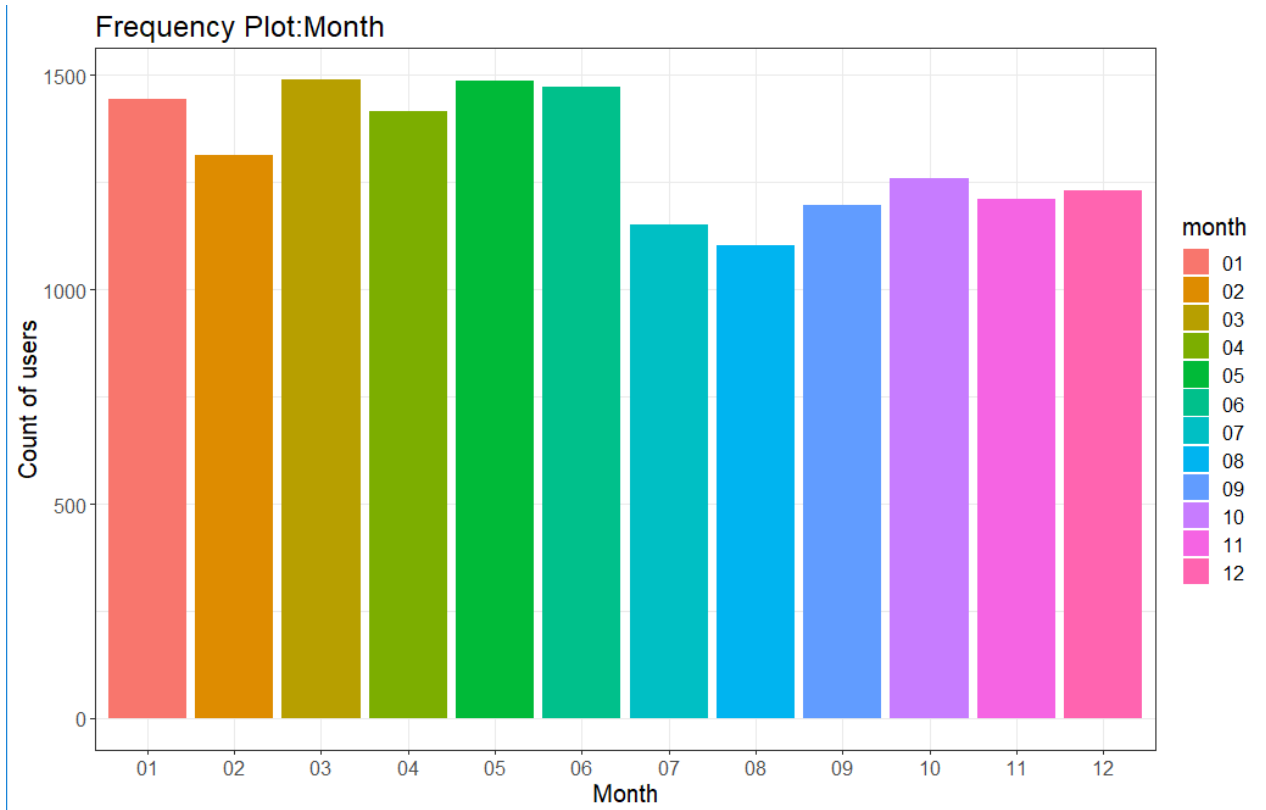
In the problem we have to predict the cab fare and our independent variables give us the pickup and drop off locations and date-time of ride of the user. So we have created new variables to find out the distance travelled by the user based on pickup and drop off locations. We have used Vincenty's formula to calculate the elliptical distance accurately. Moreover the "pickup_datetime" variable is separated to find "year", "month", weekday" and "hour_bin" variables.

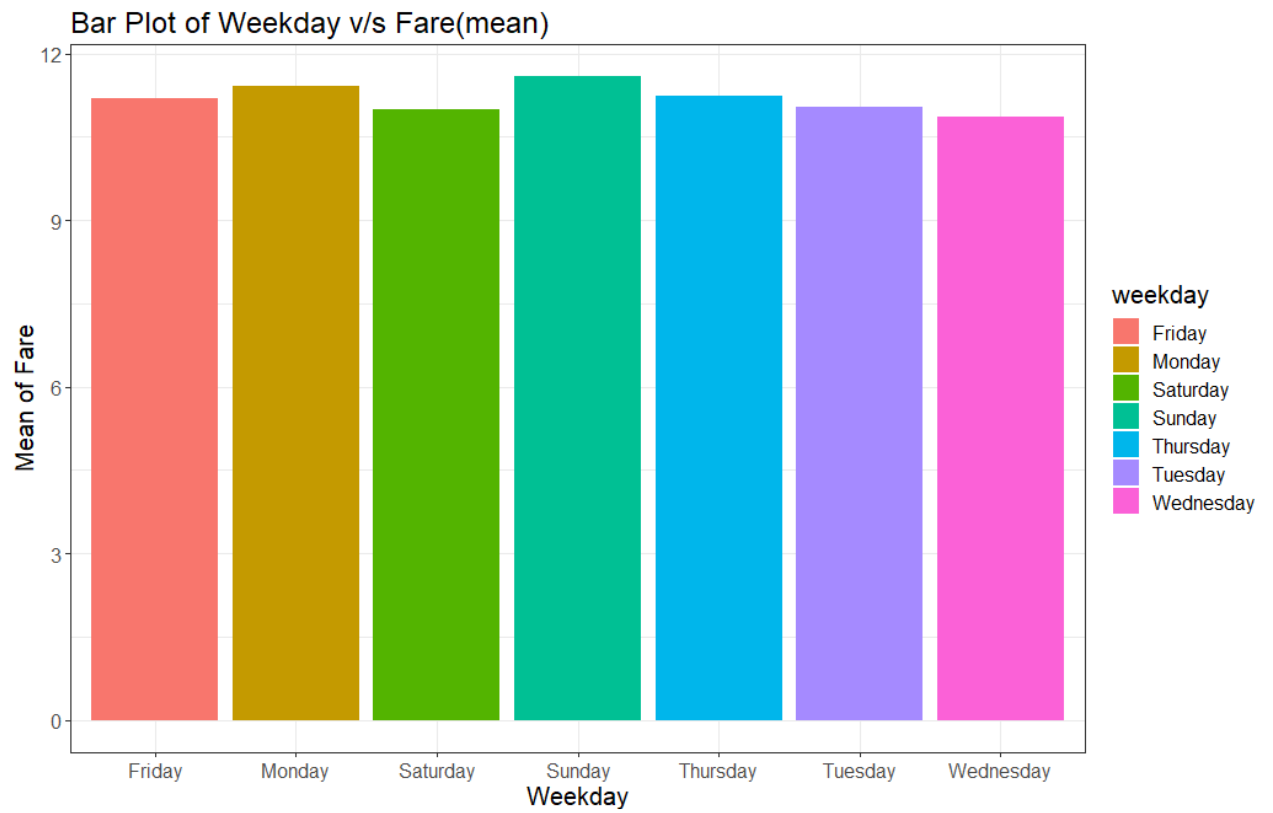
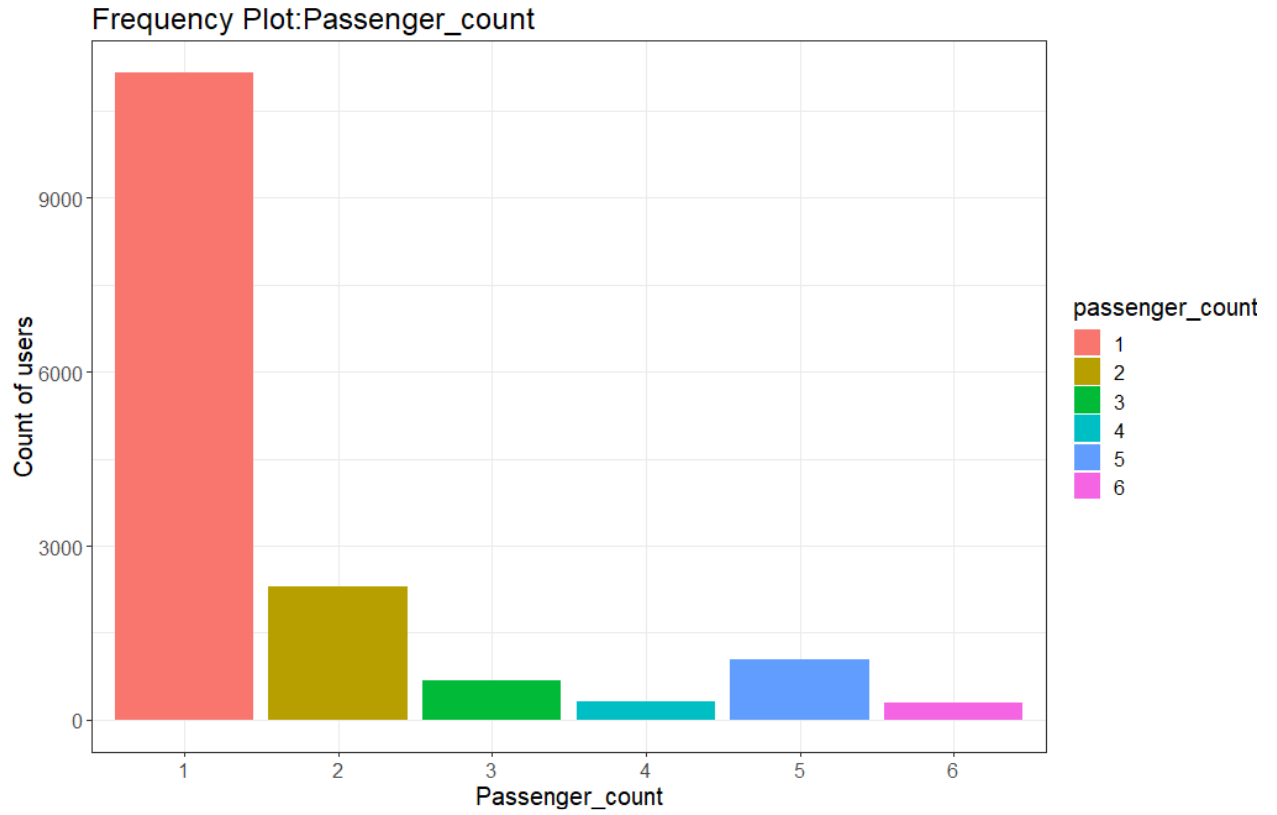
3.5 Visualizations

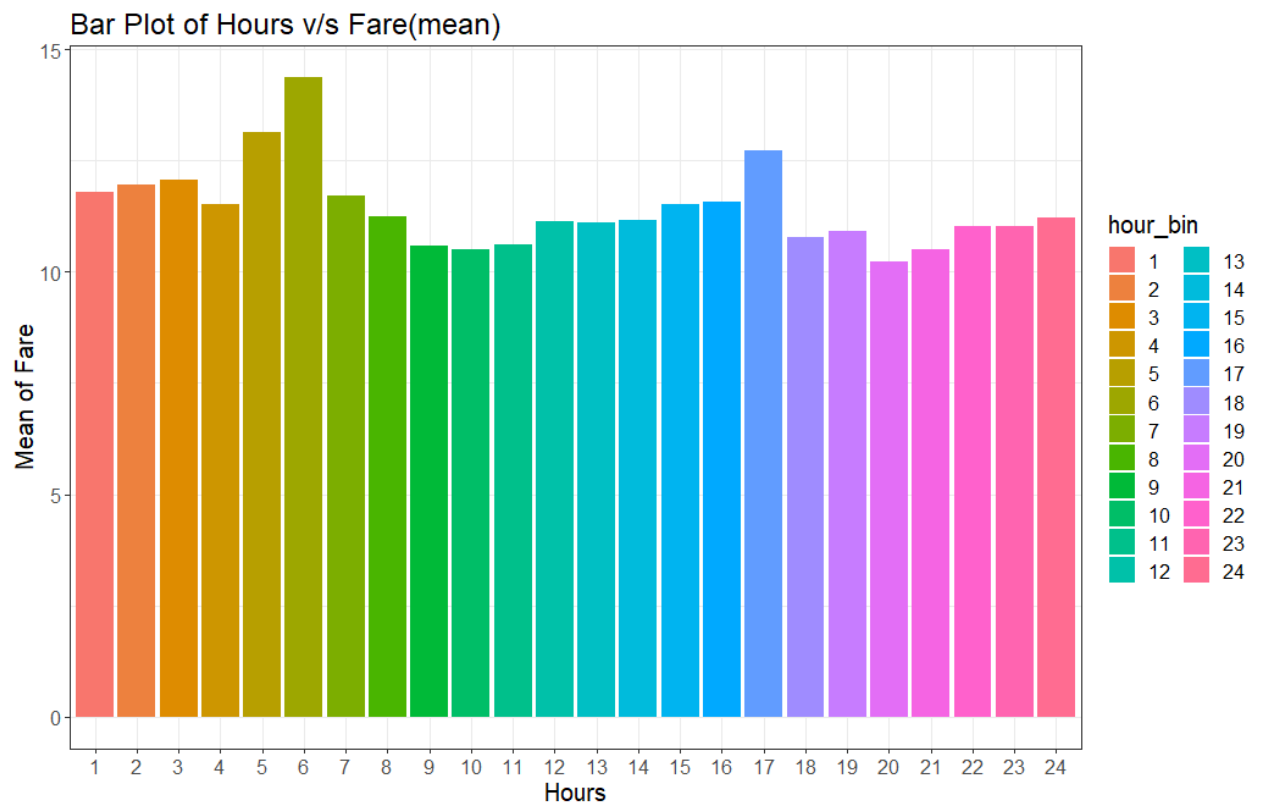
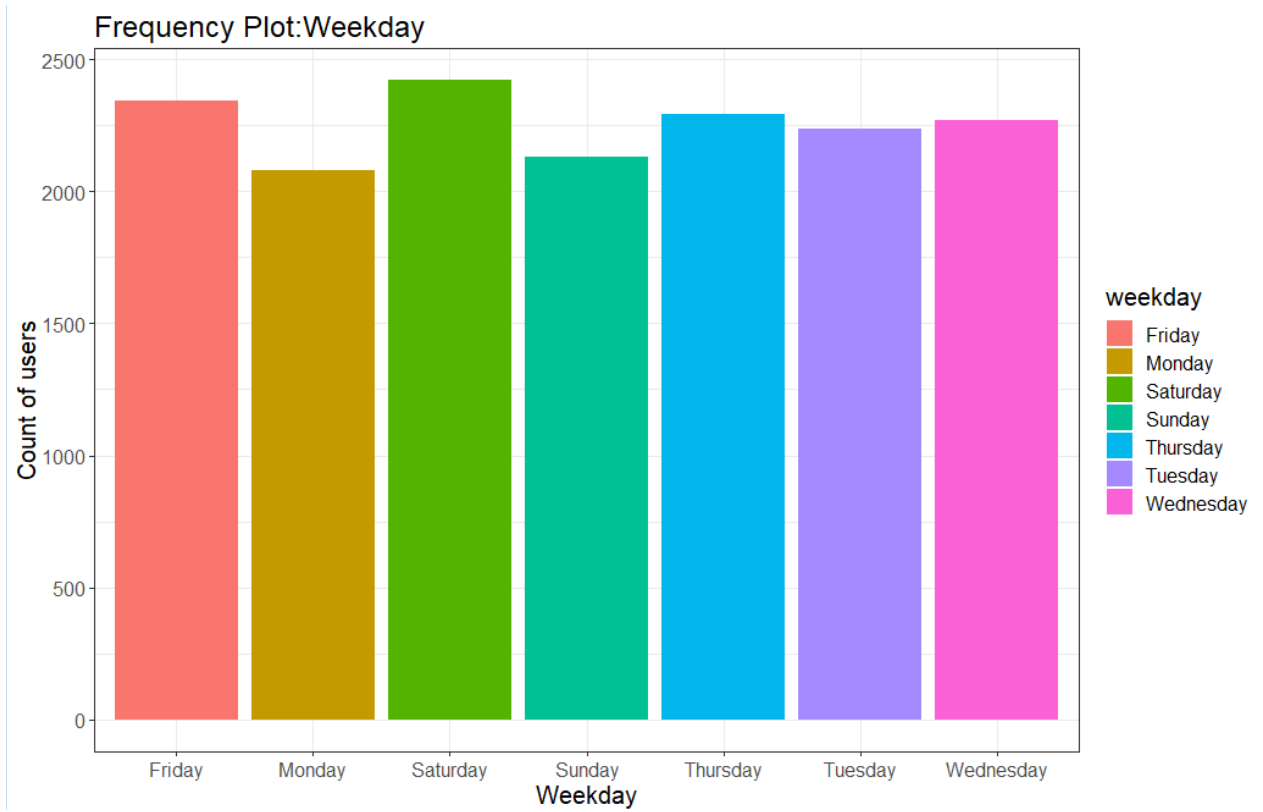
Trying to understand the dataset better with visual representation.

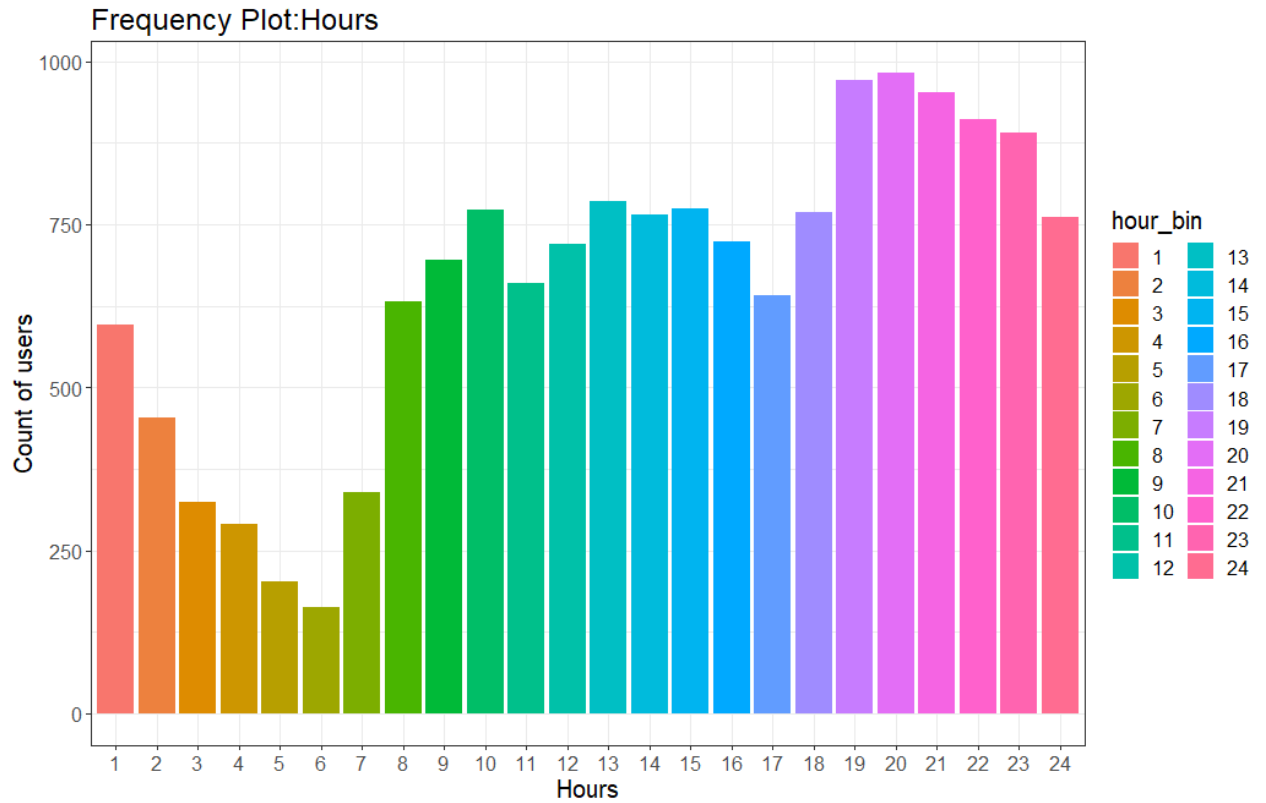




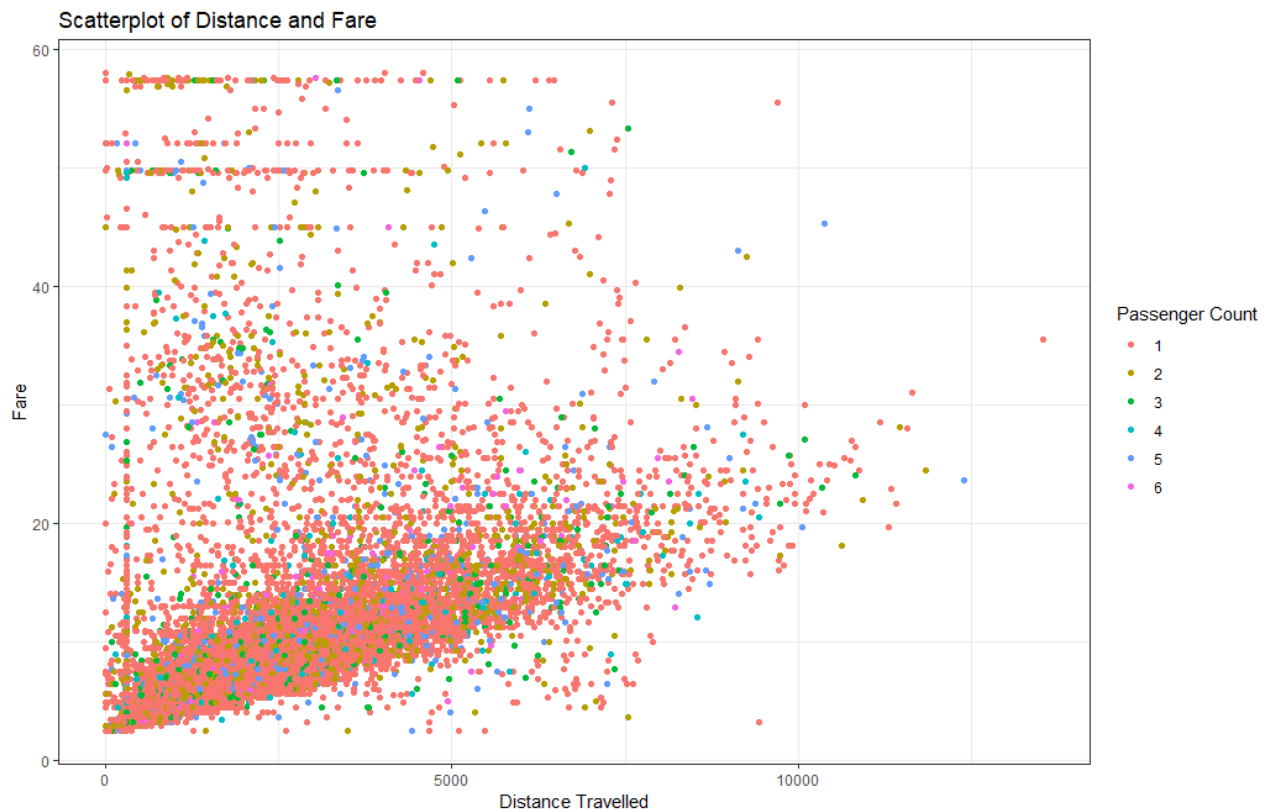




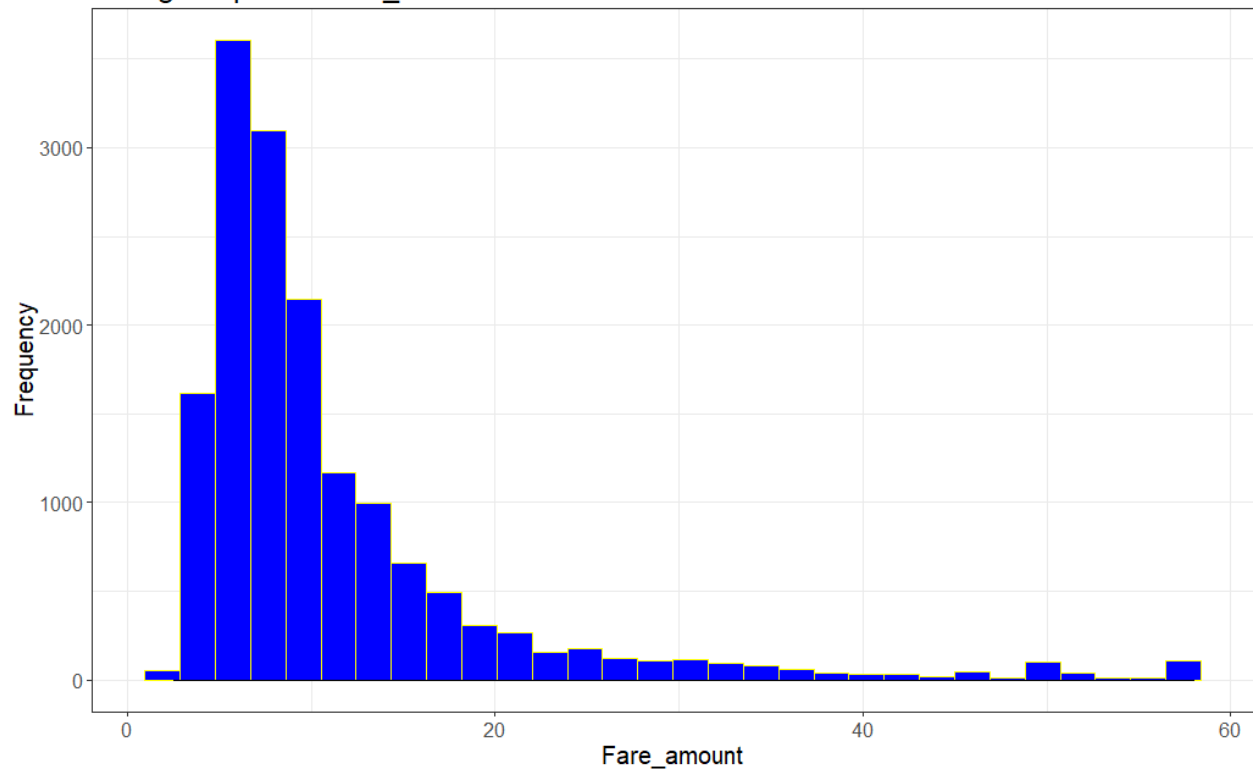




As we can see, the frequency plot of hours, unlike previous plots is right skewed.

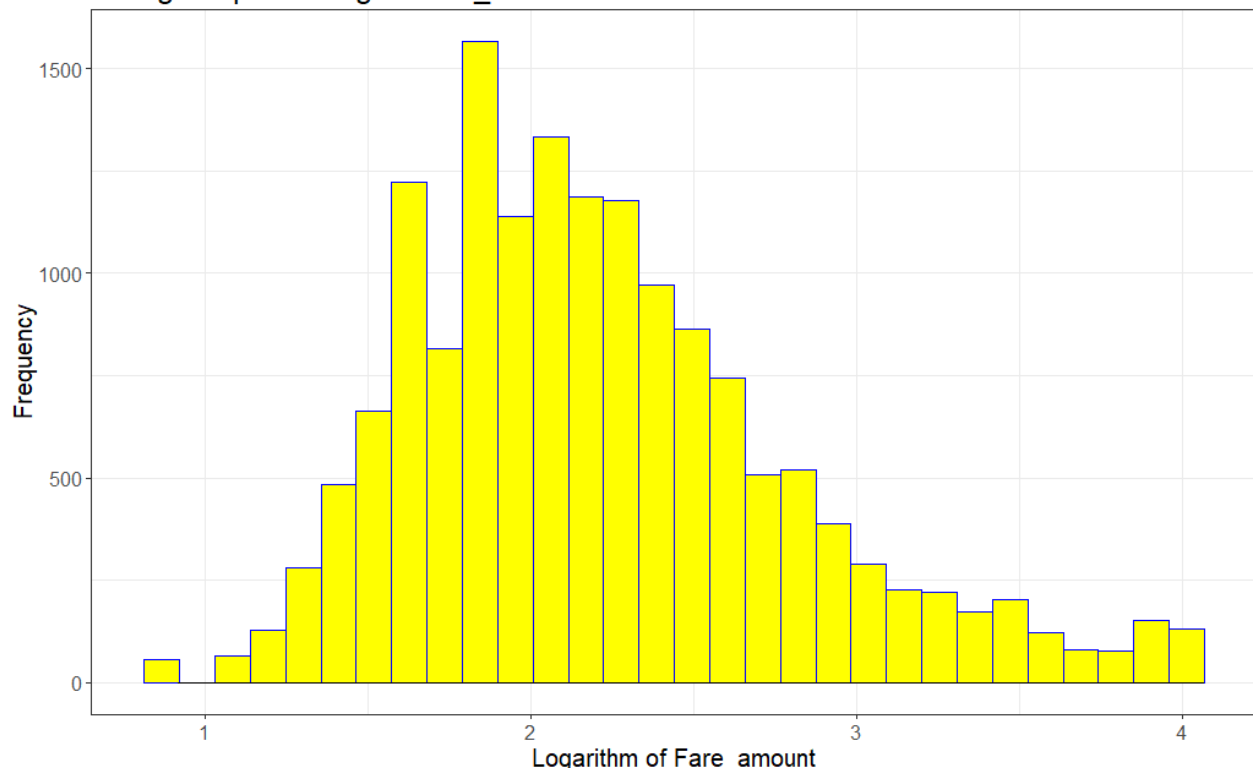


Histogram plot of Fare_amount



The above histogram is left skewed.

Histogram plot of Log of Fare_amount



We have considered the natural logarithmic values in order to give a better distribution of the data.

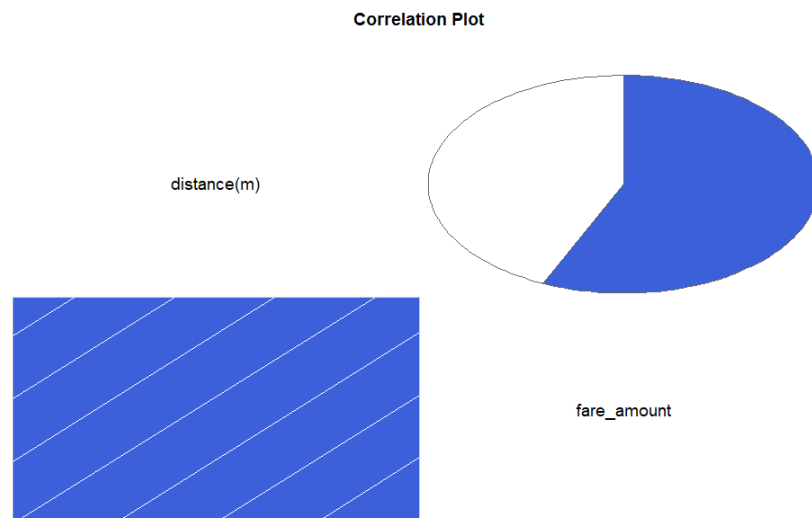
3.6 Feature Selection

In this step we would allow only to pass relevant features to further steps. We remove irrelevant features from the dataset. We do this by some statistical techniques, like we look for features which will not be helpful in predicting the target variables. In this dataset we have to predict the fare_amount.

Further below are some types of test involved for feature selection:

i. Correlation Plot:

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.



```
> 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
```

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”.

ii. Analysis of Variance(Anova) Test

It is carried out to compare between each group in a categorical variable. ANOVA only lets us know the means for different groups are same or not. It doesn't help us identify which mean is different.

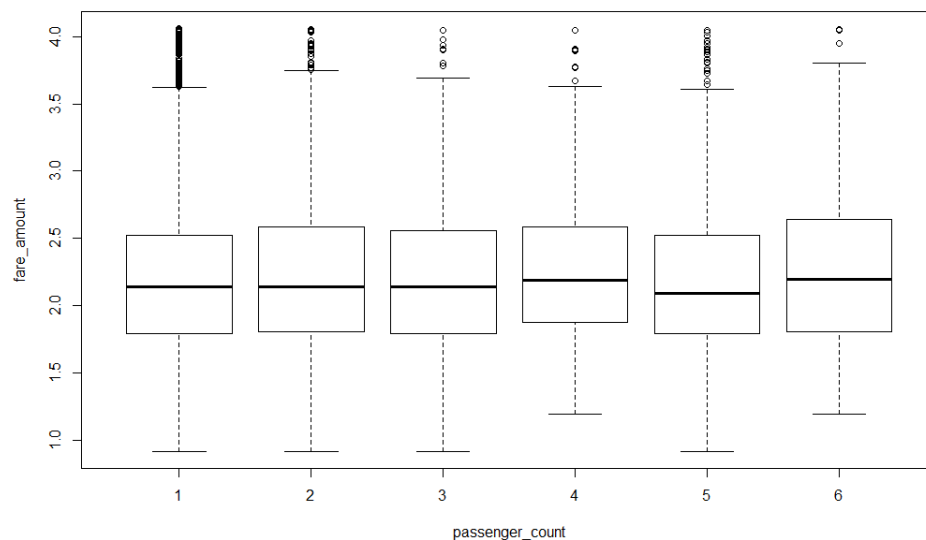
Hypothesis testing:

- Null Hypothesis: mean of all categories in a variable are same.
- Alternate Hypothesis: mean of at least one category in a variable is different.
- If p-value is less than 0.05 then we reject the null hypothesis.
- And if p-value is greater than 0.05 then we accept the null hypothesis.

Below is the anova analysis table for each categorical variable:

a. Population Count:

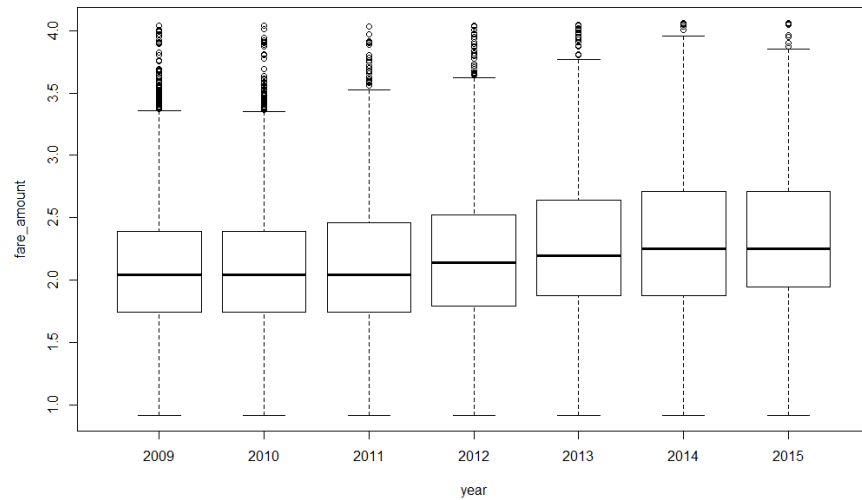
- b. From the box plot it is evident that not all population means, across the groups, are equal. So we can reject the Null Hypothesis (H0). Thus the variable is significant in explaining the variance of dependent variable. P-value was observed to be <0.05 .



c. Year:

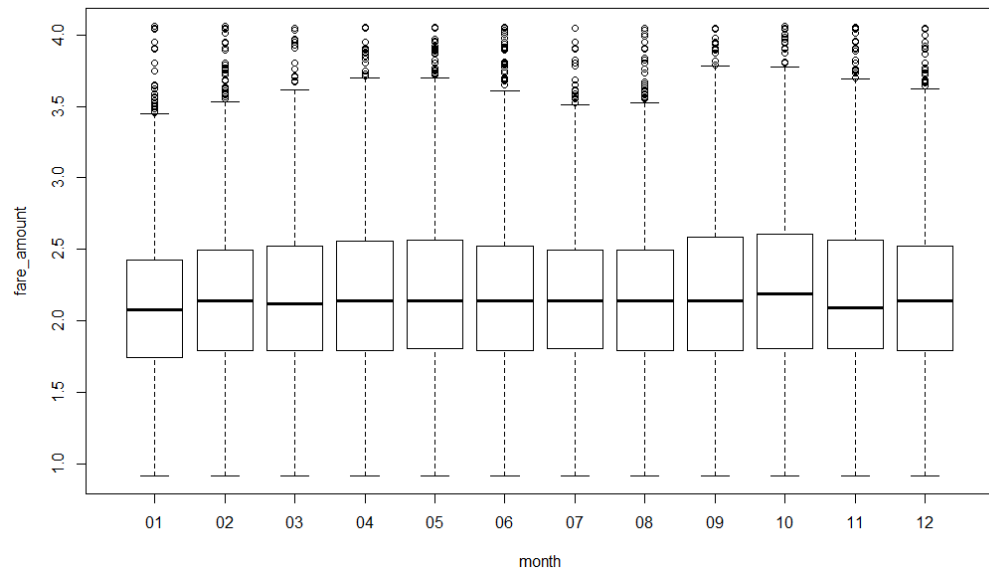
From the box plot it is evident that not all population means, across the groups, are equal. So we can reject the Null Hypothesis (H0).

Thus the variable is significant in explaining the variance of dependent variable. P-value was observed to be <0.05 .



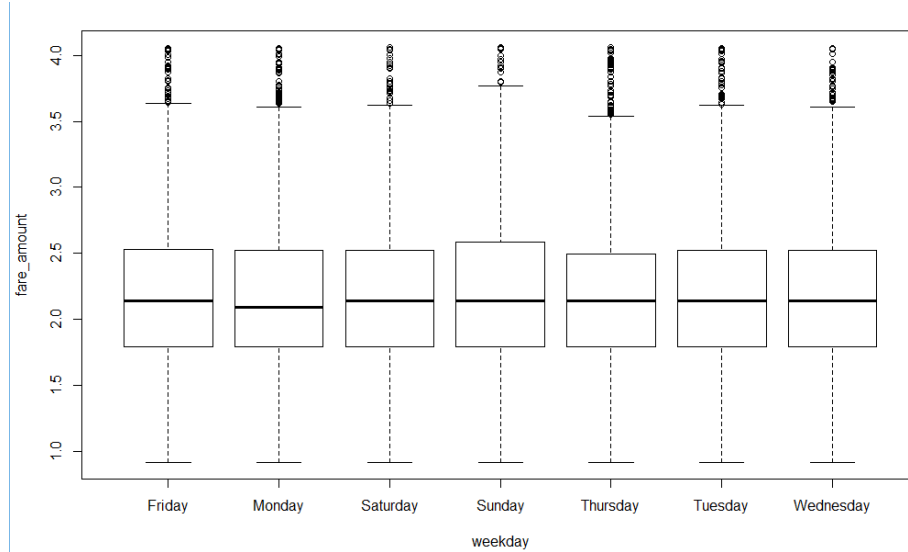
d. Month

From the box plot it is evident that not all population means, across the groups, are equal. So we can reject the Null Hypothesis (H_0). Thus the variable is significant in explaining the variance of dependent variable. P-value was observed to be <0.05 .



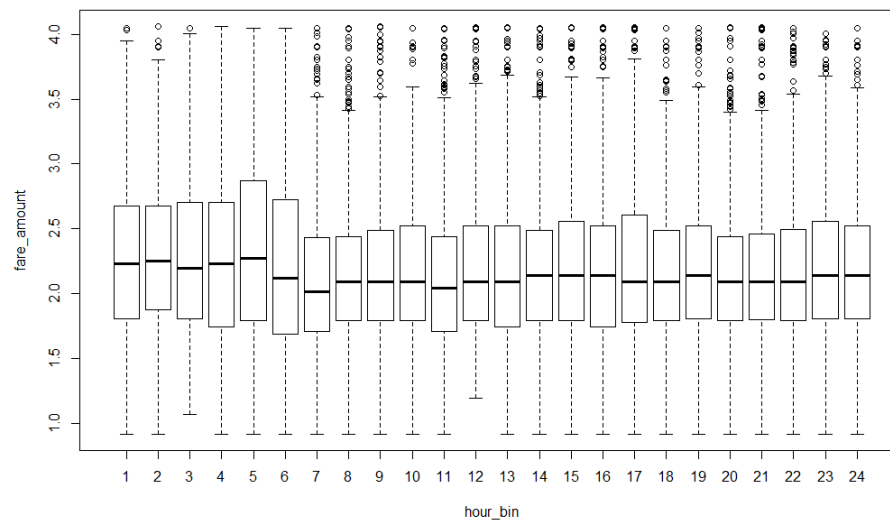
e. Weekday

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 (H_0). Thus this variable can't explain the variance of the dependent variable significantly. P-value was also observed to be >0.05 .



f. Hour

From the box plot it is evident that not all population means, across the groups, are equal. So we can reject the Null Hypothesis (H_0). Thus the variable is significant in explaining the variance of dependent variable. P-value was observed to be <0.05 .



3.7 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.

MODEL SELECTION

4.1 Modelling

Once completing data cleaned next process is model selection it is based on problem statement. In car fare prediction problem statement understood that it comes under supervised machine learning because it has both input and output variables and its regression problem as our target variable is fare_amount which is of numeric / continuous type. So, we can consider linear regression, Decision Tree, Random Forest etc. In our project used three models viz., linear regression, Decision Tree, Random Forest.

Error matrix chosen for the given problem statement is Root Mean Squared Error (RMSE) and R2(R-Squared). Before building any model we divided the preprocessed data set in to train and test set. Data was divided into 80:20 ratio, 80% of data was used as 'train' set and rest of the 20% was used as 'test' set. The training set is used to fit the model and the test set is used to estimate the model prediction accuracy.

4.2 Linear Regression

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

Linear Regression, unlike other algorithms, stores information in terms of coefficients. It is a statistical model. We cannot use this for classification. It describes relationship among variables.

Our aim is – we always want a model with low RMSE value i.e. minimum calculated errors and high R square value i.e. the independent variables should have maximum potential to explain about the target variable.

```

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

```

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. It's formulated as:

$$VIF_i = \frac{1}{1 - R_i^2}$$

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

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

4.3 Decision Trees

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.

4.4 Random Forest

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.

$$Gini = 1 - \sum (P_i)^2$$

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.

4.5 Support Vector Regression (SVR)

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 minimize 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 learned 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.

CONCLUSION

5.1 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.

5.2 Error Metrics

- 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:

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

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.

- 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.

5.3 Prediction

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
MAPE	0.3013583	0.3098124	0.3084679	0.2260185
RMSE	8.1616899	8.1914711	8.1689444	8.2148416

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.

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.

Finally, we compare 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
```

Since, most of the statistical metrics are more or less similar to each other except the max_value which could be due to values of the dependent variable present in the original data set, 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.

REFERENCES:

- <https://en.wikipedia.org/>
- <https://learning.edvisor.com/>
- <https://medium.com/>
- <https://www.statisticshowto.datasciencecentral.com/>
- <https://www.theanalysisfactor.com/>

