BIKE RENTING

Sai Bruhath

31 MARCH 2019



**Table of Contents**

**1 Introduction**

1.1Background................................................................................................................................... 2 1.2ProblemStatement..........................................................................................................................2

**2 Exploring Data**

2.1 Data...............................................................................................................................................3

2.2 Overview of dataset……………………………………………………………………………….3

2.3 Summary of Each variable ...........................................................................................................4

2.4 Visualization of raw data……………………………………………………………………………………………………………5

2.5 Pre Processing……………………………………………………………………………………..7

2.5.1Missing Value Analysis..............................................................................................................7

2.5.2 Outlier Analysis…….................................................................................................................7

2.5.3 Feature Selection……………………………………………………………………………….....................................10

2.5.4 Variable Importance……………………………………………………………………………………………………………….12

2.5.5 Variance Inflation factor………………………………………………………………………………………………………….14

2.5.6 Feature scaling………………………………………………………………………………………………………………………..15

2.5.7 Distribution check…………………………………………………………………………………………………………………..16

**3 Sampling Techniques**......................................................................................................................17

**4 Modelling**

4.1 Random Forest ……………………………………………………………………………………18

4.2 Multiple Linear Regression……………………………………..................................................22

4.3 Decision tree ……………………………………………………………………………………...24

**5 Model Selection**……………………………………………………………………………………25

**Appendix**

RandPythonCode.................................................................................................................................26

References……………………………………………………………………………………………..27

**1 Introduction**

**1.1 Background**

A Bike rental or bike hire business is a [bicycle shop](https://en.wikipedia.org/wiki/Bicycle_shop) or other business that rents [bikes](https://en.wikipedia.org/wiki/Bicycle) for short periods of time (usually for a few hours) for a [fee](https://en.wikipedia.org/wiki/Fee). Most rentals are provided by bike shops as a sideline to their main businesses of sales and service, but some shops specialize in rentals .In this implementation the focus is to predict the count of bike rents based on environmental and sesaonal setting.

**1.2 Problem Statement**

The objective of this Case is to Predication of bike rental count on daily based on the

environmental and seasonal settings.

* instant: Record indexReason for absence (ICD).
* dteday: Date
* season: Season (1:springer, 2:summer, 3:fall, 4:winter)Seasons (summer (1), autumn (2), winter (3), spring (4))
* yr: Year (0: 2011, 1:2012)
* mnth: Month (1 to 12)
* hr: Hour (0 to 23)
* holiday: weather day is holiday or not (extracted fromHoliday Schedule)
* weekday: Day of the week
* workingday: If day is neither weekend nor holiday is 1, otherwise is 0.
* weathersit: (extracted fromFreemeteo)

1: Clear, Few clouds, Partly cloudy, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered

cloudsHit target

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

* temp: Normalized temperature in Celsius. The values are derived via
  + - (t-t\_min)/(t\_max-t\_min),
    - t\_min=-8, t\_max=+39 (only in hourly scale)
* atemp: Normalized feeling temperature in Celsius. The values are derived via
  + - (t-t\_min)/(t\_maxt\_min), t\_min=-16, t\_max=+50 (only in hourly scale)

hum: Normalized humidity. The values are divided to 100 (max)

* windspeed: Normalized wind speed. The values are divided to 67 (max)
* casual: count of casual users
* registered: count of registered users
* cnt: count of total rental bikes including both casual and registered.

**2 Exploring Data**

**2.1 - Data size and structure**

> str(Bike\_predict)

'data.frame': 731 obs. of 14 variables:

$ season : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...

$ yr : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ mnth : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...

$ holiday : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ weekday : Factor w/ 7 levels "0","1","2","3",..: 7 1 2 3 4 5 6 7 1 2 ...

$ workingday: Factor w/ 2 levels "0","1": 1 1 2 2 2 2 2 1 1 2 ...

$ weathersit: Factor w/ 3 levels "1","2","3": 2 2 1 1 1 1 2 2 1 1 ...

$ temp : num 0.344 0.363 0.196 0.2 0.227 ...

$ atemp : num 0.364 0.354 0.189 0.212 0.229 ...

$ hum : num 0.806 0.696 0.437 0.59 0.437 ...

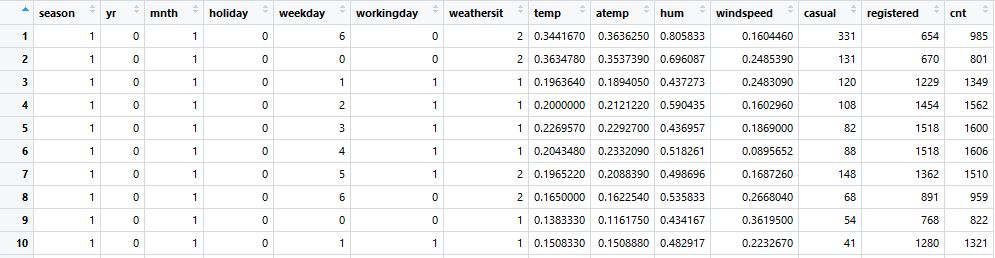
$ windspeed : num 0.16 0.249 0.248 0.16 0.187 ...

$ casual : int 331 131 120 108 82 88 148 68 54 41 ...

$ registered: int 654 670 1229 1454 1518 1518 1362 891 768 1280 ...

$ cnt : int 985 801 1349 1562 1600 1606 1510 959 822 1321 ...

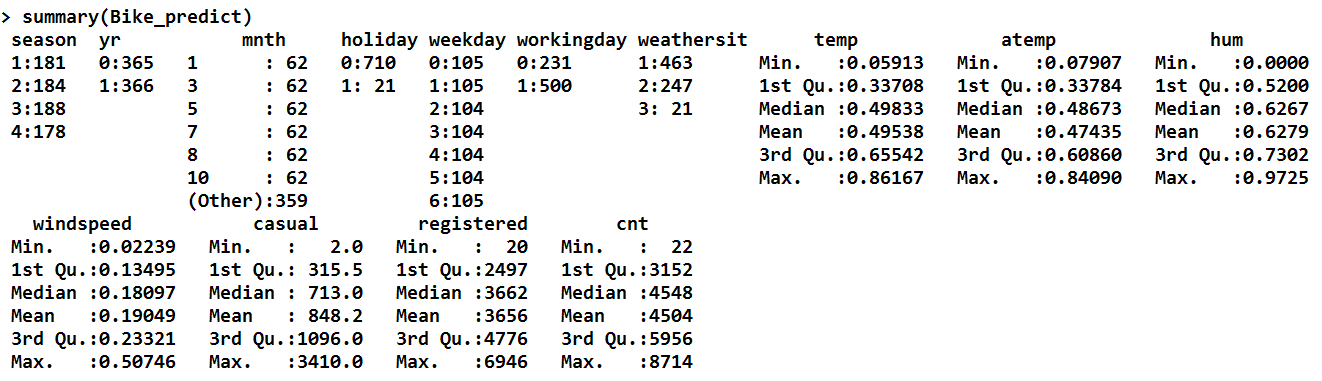
2.2 Overview of the dataset :

****

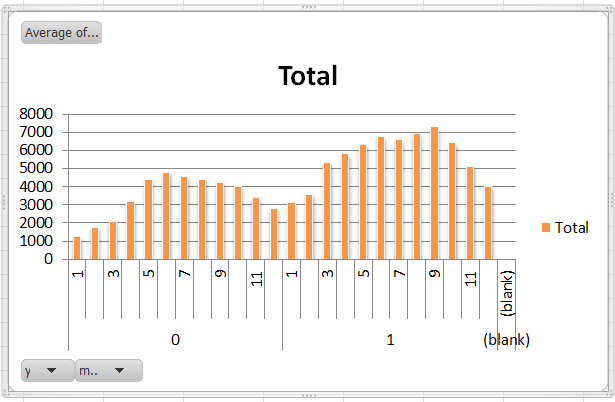
As we can see , there are seven categorical variables and 7 numeric variables .The dependent variable or the target variable is cnt, i.e. the count of people renting the bikes.

**2.3 Summary of Each Variable :**

**[1] 731 observations 12 variables**



**2.4 Visualization of raw data :**



0 : 2011

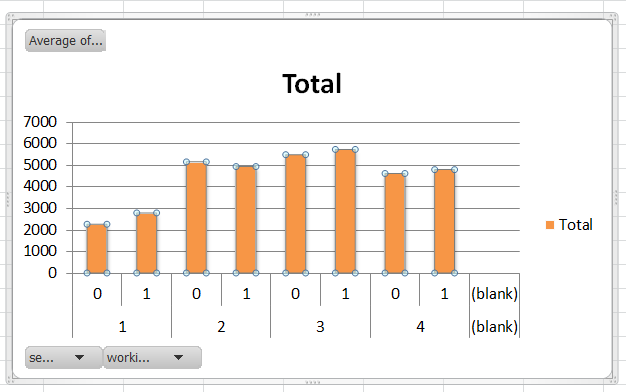
1 : 2012

The above graph shows Average count of rental bikes in each month and year.

**Insights :**

More number Of Bikes in 2012 when compared to 2011.

Less number of bikes registered in the month of January and February in both the years.

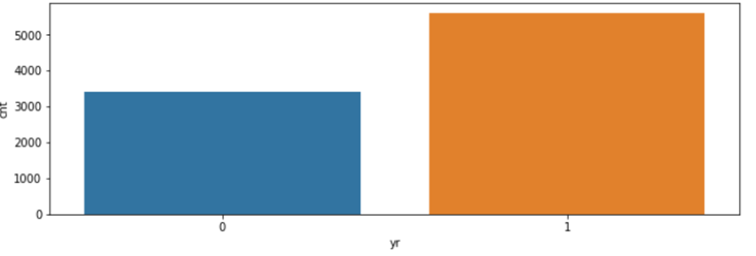


graph shows Average count of rental bikes in each on season.

**Insights from the graph :**

People rent more bikes in fall (season3) and less bikes in springer(Season1).

People rent comparatively more number of bikes in working day which is neither holiday nor weekend.



Distribution of count by year

**2.5 Pre Processing**

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

To start this process analysing the count of missing values in each variable.

**2.5.1Missing Value Analysis :**

**>sum(is.na(Bike\_predict))**

[0]

On analysing , it is seen that there are no duplicate rows or missing values present throughout the dataset. Also the numerical variables are scaled by normalizing the data ,so only two variable we need to normalize.

**2.5.2 Outlier Analysis :**

One of the other steps of pre-processingis the removal of outliers. Observations inconsistent with rest of global dataset are called as outliers.

Outlier can be caused due to several reasons :

* **Data Entry Errors:-** Human errors such as errors caused during data collection, recording, or entry can cause outliers in data. For example: Annual income of a customer is $100,000. Accidentally, the data entry operator puts an additional zero in the figure. Now the income becomes $1,000,000 which is 10 times higher. Evidently, this will be the outlier value when compared with rest of the population.
* **Measurement Error:**It is the most common source of outliers. This is caused when the measurement instrument used turns out to be faulty. For example: There are 10 weighing machines. 9 of them are correct, 1 is faulty. Weight measured by people on the faulty machine will be higher / lower than the rest of people in the group. The weights measured on faulty machine can lead to outliers.
* **Experimental Error:** Another cause of outliers is experimental error. For example: In a 100m sprint of 7 runners, one runner missed out on concentrating on the ‘Go’ call which caused him to start late. Hence, this caused the runner’s run time to be more than other runners. His total run time can be an outlier.

Impact to the model due to presence of outliers :

Outliers can drastically change the results of the data analysis and statistical modeling. It increases the error variance and reduces the power of statistical tests

* If the outliers are non-randomly distributed, they can decrease normality
* They can bias or influence estimates that may be of substantive interest
* They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

Outliers can be detected by using various methods :

* Graphical too (Box Plot)
* Stastical Technique(Grabb’s test for outliers)
* R- package outlier
* Replace with NA

In this case we use a classic approach of removing outliers, Tukey’*s method*.

We visualize the outliers using boxplot*s* and remove the outliers.

In box plot values which are falling beyond the upper and lower fence is called as outliers.

Similar way outliers has been calculated for each variable.

Outlier length :

> cnames

[1] "temp" "atemp" "hum" "windspeed" "casual" "registered" "cnt"

[1] 0

[1] 0

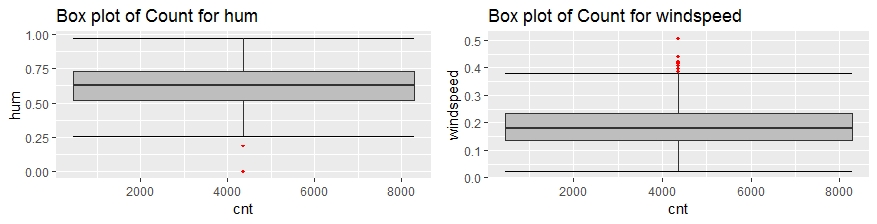
[1] 2

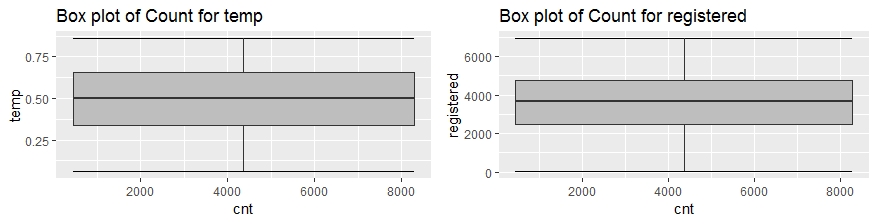
[1] 13

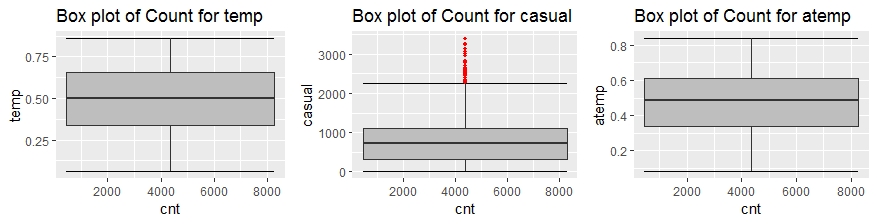
[1] 44

[1] 0

[1] 0







**2.5.3 Feature Selection**

Feature selection is a subset of relevant features (Variables,Predictors) for use in model construction.

It is also called as variable selection or attribute selection.

Feature engineering is the science (and art) of extracting more information from existing data. You are not adding any new data here, but you are actually making the data you already have more useful

Feature selection can be done in two ways ::

* Correlation Analysis
* Chi-Square test
* Analysis of variance (Anova)

**Correlation Analysis** : Correlation can be derived using following formula:

**Correlation = Covariance(X,Y) / SQRT( Var(X)\* Var(Y))**

Various tools have function or functionality to identify correlation between variables. In Excel, function CORREL() is used to return the correlation between two variables and SAS uses procedure PROC CORR to identify the correlation.

It is applicable only for Numerical Variables.

Assumptions :

1. There should be no (or) less correlation exist between two independent variables.
2. High correlation should exist between independent and dependent variable.

Correlation value ranges from -1 to +1.

**Chi-Square test :** This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

It is applicable for categorical variables.

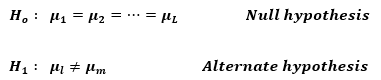
Assuumptions :

Null Hypothesis : Two variables are independent

Alternative hypothesis : Two variables are not independent.

**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.

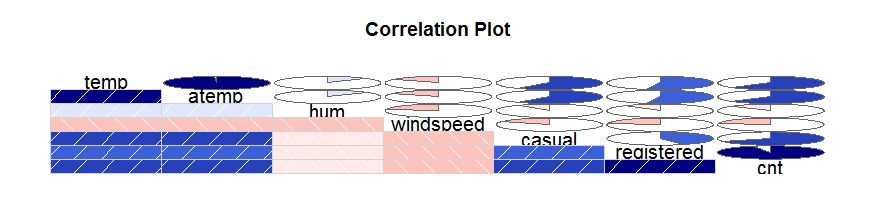
ANOVA also uses a Null hypothesis and an Alternate hypothesis. The Null hypothesis in ANOVA is valid when all the sample means are equal, or they don’t have any significant difference. Thus, they can be considered as a part of a larger set of the population. On the other hand, the alternate hypothesis is valid when at least one of the sample means is different from the rest of the sample means. In mathematical form, they can be represented as:



where https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/12/image0111.pngbelong to any two sample means out of all the samples considered for the test. In other words, the null hypothesis states that all the sample means are equal or the factor did not have any significant effect on the results. Whereas, the alternate hypothesis states that at least one of the sample means is different from another. But we still can’t tell which one specifically.

In this project performed correlation and ANOVA , Correlation is applied for numerical variables and Anova is for Categorical variables as Anova is used to compare one numerical and group of categorical variables.(Target variables is continuous) compared each categorical variable with Target variable.

**2.5.4 Variable Importance:**



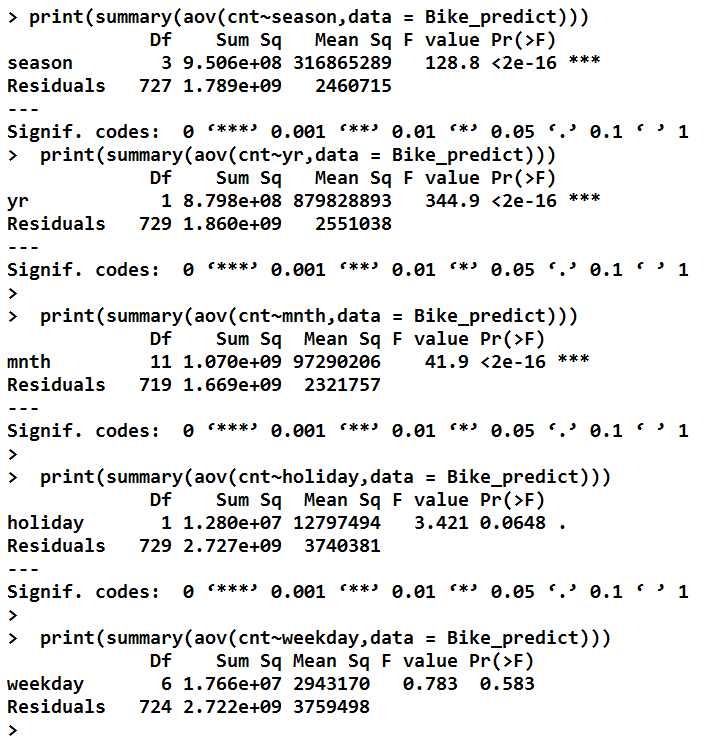
Above figure shows the correlation analysis table

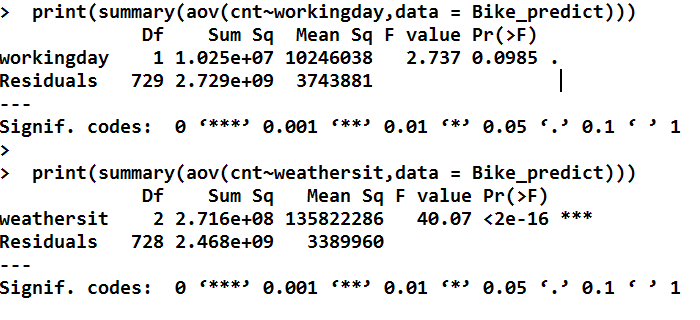
Insights from Correlation Plot:

1. Temp and atemp are highly correlated each other.
2. Registered and our target variable cnt are highly correlated.

As we know correlation between two independent variables should be low and correlation between target variable and independent variable should be high. In temp and atemp one of the variable has to be dropped.

**ANOVA – Categorical Variables:**





**2.5.5 Variance Inflation Factor(VIF) :** We have performed VIF test using function VIF which is used to check whether variables have multicollinearity .

VIF outpuut :

2 variables from the 7 input variables have collinearity problem:

atemp cnt

After excluding the collinear variables, the linear correlation coefficients ranges between:

min correlation ( casual ~ hum ): -0.07700788

max correlation ( casual ~ temp ): 0.5432847

---------- VIFs of the remained variables --------

Variables VIF

1 temp 1.852244

2 hum 1.177663

3 windspeed 1.154336

4 casual 1.506753

5 registered 1.546630

Conclusions : From the various test has been drawn that temp and atemp are highly correlated so need to drop atleast one variable .From the ANOVA test we have found that variable ‘weekday’ have p-value greater than 0.05(significance value) so need to drop variable weekday.

**2.5.6 Feature Scaling :**

When we want to change the scale of a variable or standardize the values of a variable for better understanding. While this transformation is a must if you have data in different scales, this transformation does not change the shape of the variable distribution.

Feature scaling is applicable only for Continuous variables.

Feature scaling has two methods to transform the data :

* Standardization
* Normalization

**Standardization /Z –score** :

Basically it converts each data point unique of standard deviation.

Z=

Z -represents the difference between raw score and population mean in the units of standard deviation.

If the data is normally distributed (or) uniformly distributed then standardization is used.

**Normalization :**

Normalization to bring all the variables into proportion with one another variable.

New value =(Value-Min value)/ (max value – minvalue).

Range is from 0 to 1.

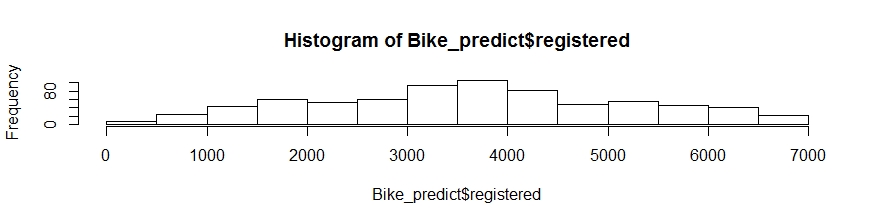
If the data is left skewed or right skewed other than unformally distrubuted then Nrmalization is used.

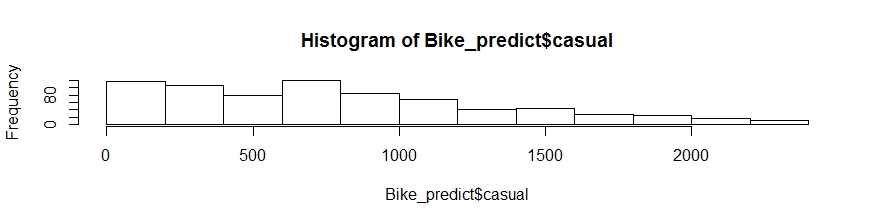
As our data is normally distributed standardization technique is used for most of the variables.

**2.5.7 Density Plot: Distribution of data points**

Checking the distribution of values for each variable. All the variable seems to be not normally distributed Hence Normaliation technique is used to change the scale of variables.

The below plots shows the variables pattern in the dataset.





Now we have completed all the pre-processing techniques that has to be applied on the data.

Data available after pre-processing : 731 observations 12 variables

**3. Sampling Techniques :**

Selection of Subset from the whole population is called sampling.

Sampling can be categorized into two categories

* Probability samples : Probability sampling is a sampling technique, in which the subjects of the population get an equal opportunity to be selected as a representative sample.

It is selected Randomly.

* Non – Probability samples: Nonprobability sampling is a method of sampling wherein, it is not known that which individual from the population will be selected as a sample.

It is selected Arbitarly.

Different types of Probability sampling are as follows :

1. Simple Random Sampling.
2. Systematic Radom Sampling.
3. Stratified Sampling.
4. Multi-stage cluster Sampling.

1. **Simple Random Sampling** : Random Sampling is purest form of probability sampling.

Selected by using chance or random numbers.

Each individual subject has an equal chance of being selected.

EX: Random Numbers, Drawing names from hat.

1. **Systematic Sampling** : It is also called kth name selection technique.

After required sample size has been calculated every kth record is selected from list of population numbers.

K=N/n ,where N= no of observations , n= Desired shape.

EX: If K=3

P1,P2,P3,P4,P5,P6,P7,P8,P9

P3,P6,P9 is selected.

This process is selected if the data desn’t contain an hidden order.

1. **Stratfied Sampling** : In stratified sampling we have to select categorical variables on equal propotions.

In this project simple random sampling is used.

After sampling distribution of the data are as follows :

Test : 147 Observations ,12Variables

Train : 584 Observations,12 Variables.

**4.Modeling :**

To make predictions, I have used Random Forest and Multiple linear regression,

* 1. **Random Forest :** Random forest is popular Regression method. random forest is better at fitting non-linear data. It can also work well even if there are correlated features, which can be a problem for interpreting Linear regression.

**Model summary :**

> print(RF\_model)

Call:

randomForest(x = X\_train, y = Y\_train, ntree = 100)

Type of random forest: regression

Number of trees: 100

No. of variables tried at each split: 3

Mean of squared residuals: 104755.7

% Var explained: 97.24

> summary(RF\_model)

Length Class Mode

call 4 -none- call

type 1 -none- character

predicted 584 -none- numeric

mse 100 -none- numeric

rsq 100 -none- numeric

oob.times 584 -none- numeric

importance 11 -none- numeric

importanceSD 0 -none- NULL

localImportance 0 -none- NULL

proximity 0 -none- NULL

ntree 1 -none- numeric

mtry 1 -none- numeric

forest 11 -none- list

coefs 0 -none- NULL

y 584 -none- numeric

test 0 -none- NULL

inbag 0 -none- NULL

> getTree(RF\_model,1,labelVar = TRUE)

left daughter right daughter split var split point status prediction

1 2 3 yr 1.000000e+00 -3 4452.384

2 4 5 casual 1.722074e-01 -3 3423.010

3 6 7 registered 5.639619e-01 -3 5569.989

4 8 9 workingday 1.000000e+00 -3 2061.333

5 10 11 season 1.000000e+00 -3 4206.150

6 12 13 mnth 2.051000e+03 -3 3749.835

7 14 15 temp 4.550000e-01 -3 6534.770

8 16 17 season 5.000000e+00 -3 1449.846

9 18 19 season 7.000000e+00 -3 2248.376

10 20 21 temp 3.720835e-01 -3 2762.857

11 22 23 temp 4.695835e-01 -3 4260.468

12 24 25 registered 3.483973e-01 -3 3031.857

13 26 27 registered 3.154057e-01 -3 4730.488

14 28 29 mnth 2.551000e+03 -3 5345.469

15 30 31 casual 2.943262e-01 -3 6786.808

16 32 33 mnth 2.052000e+03 -3 1022.895

17 34 35 registered 3.275339e-01 -3 2608.714

18 36 37 hum 7.122915e-01 -3 1783.236

19 38 39 weathersit 6.000000e+00 -3 3101.133

20 0 0 <NA> 0.000000e+00 -1 2132.000

21 0 0 <NA> 0.000000e+00 -1 3015.200

22 40 41 weathersit 2.000000e+00 -3 3482.316

23 42 43 registered 4.114929e-01 -3 4460.264

24 44 45 hum 7.330165e-01 -3 2265.083

25 46 47 temp 3.233335e-01 -3 3606.937

26 0 0 <NA> 0.000000e+00 -1 2425.000

27 48 49 mnth 3.951000e+03 -3 4848.718

28 50 51 windspeed 2.960270e-01 -3 4798.308

29 52 53 windspeed 1.308320e-01 -3 5719.842

30 54 55 registered 7.513716e-01 -3 5144.647

31 56 57 temp 7.995830e-01 -3 6995.142

32 0 0 <NA> 0.000000e+00 -1 775.600

33 58 59 casual 4.853723e-02 -3 1111.214

34 0 0 <NA> 0.000000e+00 -1 2431.000

35 0 0 <NA> 0.000000e+00 -1 2742.000

36 0 0 <NA> 0.000000e+00 -1 1694.278

37 60 61 registered 3.072481e-01 -3 1951.789

38 62 63 registered 4.413803e-01 -3 2848.762

39 64 65 temp 3.512500e-01 -3 3690.000

40 66 67 mnth 4.000000e+00 -3 2890.091

41 68 69 casual 2.912234e-01 -3 3723.593

42 70 71 hum 6.314585e-01 -3 3799.619

43 72 73 workingday 2.000000e+00 -3 4569.504

44 74 75 registered 2.246607e-01 -3 2379.316

45 0 0 <NA> 0.000000e+00 -1 1831.000

46 76 77 registered 4.956685e-01 -3 3408.059

47 78 79 casual 3.690160e-01 -3 3832.333

48 80 81 casual 5.135195e-01 -3 4517.286

49 82 83 weathersit 6.000000e+00 -3 5692.364

50 84 85 season 5.000000e+00 -3 4720.750

51 0 0 <NA> 0.000000e+00 -1 5729.000

52 0 0 <NA> 0.000000e+00 -1 5356.500

53 86 87 mnth 3.583000e+03 -3 5816.733

54 88 89 casual 1.812943e-01 -3 4977.133

55 0 0 <NA> 0.000000e+00 -1 6401.000

56 90 91 casual 8.523310e-01 -3 7086.669

57 92 93 windspeed 1.166080e-01 -3 5860.200

58 0 0 <NA> 0.000000e+00 -1 913.500

59 94 95 temp 2.887500e-01 -3 1190.300

60 96 97 windspeed 2.843645e-01 -3 1640.857

61 0 0 <NA> 0.000000e+00 -1 2822.400

62 98 99 registered 2.624170e-01 -3 2664.471

63 0 0 <NA> 0.000000e+00 -1 3632.000

64 100 101 temp 2.866665e-01 -3 3559.286

65 0 0 <NA> 0.000000e+00 -1 4147.500

66 0 0 <NA> 0.000000e+00 -1 1693.000

67 102 103 registered 2.975744e-01 -3 3156.111

68 104 105 mnth 2.056000e+03 -3 3482.455

69 106 107 mnth 1.032000e+03 -3 3889.375

70 108 109 registered 3.535230e-01 -3 3434.556

71 110 111 casual 6.288011e-01 -3 4073.417

72 112 113 weathersit 2.000000e+00 -3 4442.919

73 114 115 casual 7.728280e-01 -3 5017.071

74 0 0 <NA> 0.000000e+00 -1 1777.333

75 116 117 mnth 1.000000e+00 -3 2492.187

76 118 119 temp 2.995835e-01 -3 3186.727

77 120 121 windspeed 1.998770e-01 -3 3813.833

78 122 123 registered 5.047647e-01 -3 3726.385

79 0 0 <NA> 0.000000e+00 -1 4521.000

80 124 125 weathersit 2.000000e+00 -3 4133.625

81 126 127 hum 8.402085e-01 -3 5028.833

82 0 0 <NA> 0.000000e+00 -1 5075.800

83 128 129 season 1.300000e+01 -3 6206.167

84 130 131 registered 5.884349e-01 -3 4595.375

85 0 0 <NA> 0.000000e+00 -1 4971.500

86 132 133 mnth 4.087000e+03 -3 5627.636

87 0 0 <NA> 0.000000e+00 -1 6336.750

88 0 0 <NA> 0.000000e+00 -1 4588.400

89 134 135 hum 6.633335e-01 -3 5171.500

90 136 137 windspeed 2.848210e-01 -3 7177.676

91 138 139 casual 9.605496e-01 -3 6472.375

92 0 0 <NA> 0.000000e+00 -1 6786.000

93 140 141 temp 8.200000e-01 -3 5628.750

94 142 143 registered 1.566561e-01 -3 1241.143

95 0 0 <NA> 0.000000e+00 -1 1071.667

96 144 145 windspeed 2.306265e-01 -3 1728.154

97 0 0 <NA> 0.000000e+00 -1 506.000

98 0 0 <NA> 0.000000e+00 -1 1607.000

99 146 147 registered 3.579988e-01 -3 2730.562

100 0 0 <NA> 0.000000e+00 -1 3483.000

101 0 0 <NA> 0.000000e+00 -1 3616.500

102 0 0 <NA> 0.000000e+00 -1 2437.667

103 148 149 registered 3.491193e-01 -3 3515.333

104 150 151 windspeed 2.481420e-01 -3 3167.000

105 0 0 <NA> 0.000000e+00 -1 3861.000

106 152 153 casual 3.752216e-01 -3 3699.818

107 0 0 <NA> 0.000000e+00 -1 4306.400

108 0 0 <NA> 0.000000e+00 -1 3200.000

109 0 0 <NA> 0.000000e+00 -1 3622.200

110 154 155 casual 5.498670e-01 -3 3773.286

111 0 0 <NA> 0.000000e+00 -1 4493.600

112 156 157 registered 5.368178e-01 -3 4118.314

113 158 159 registered 5.216575e-01 -3 4620.438

114 160 161 temp 6.729165e-01 -3 4808.615

115 162 163 mnth 3.807000e+03 -3 5197.733

116 164 165 windspeed 1.896750e-01 -3 2339.625

117 166 167 casual 1.119238e-01 -3 2644.750

118 168 169 hum 4.306250e-01 -3 3254.875

119 0 0 <NA> 0.000000e+00 -1 3005.000

120 0 0 <NA> 0.000000e+00 -1 3821.200

121 0 0 <NA> 0.000000e+00 -1 3777.000

122 170 171 season 8.000000e+00 -3 3390.857

123 172 173 temp 3.425000e-01 -3 4117.833

124 174 175 mnth 1.316000e+03 -3 4090.636

125 0 0 <NA> 0.000000e+00 -1 4228.200

126 176 177 casual 6.482228e-01 -3 5089.100

127 0 0 <NA> 0.000000e+00 -1 4727.500

128 0 0 <NA> 0.000000e+00 -1 5464.000

129 0 0 <NA> 0.000000e+00 -1 6577.250

130 0 0 <NA> 0.000000e+00 -1 4271.000

131 0 0 <NA> 0.000000e+00 -1 4790.000

132 178 179 hum 5.291665e-01 -3 5459.000

133 0 0 <NA> 0.000000e+00 -1 5830.000

134 0 0 <NA> 0.000000e+00 -1 5445.333

135 180 181 windspeed 1.756855e-01 -3 5054.143

136 182 183 registered 8.439215e-01 -3 7280.711

137 184 185 mnth 4.063000e+03 -3 6269.091

138 186 187 temp 6.500000e-01 -3 6670.571

139 188 189 season 1.300000e+01 -3 6318.222

140 0 0 <NA> 0.000000e+00 -1 5905.000

141 0 0 <NA> 0.000000e+00 -1 5463.000

142 0 0 <NA> 0.000000e+00 -1 1148.800

143 0 0 <NA> 0.000000e+00 -1 1472.000

144 190 191 weathersit 3.000000e+00 -3 1664.636

145 0 0 <NA> 0.000000e+00 -1 2077.500

146 192 193 hum 8.540670e-01 -3 2531.375

147 194 195 temp 3.945835e-01 -3 2929.750

148 0 0 <NA> 0.000000e+00 -1 3134.500

149 0 0 <NA> 0.000000e+00 -1 3705.750

150 0 0 <NA> 0.000000e+00 -1 3238.800

151 0 0 <NA> 0.000000e+00 -1 2808.000

152 0 0 <NA> 0.000000e+00 -1 4189.000

153 196 197 temp 3.658335e-01 -3 3591.111

154 0 0 <NA> 0.000000e+00 -1 3351.000

155 198 199 temp 7.266665e-01 -3 3843.667

156 200 201 mnth 8.720000e+02 -3 3996.241

157 202 203 windspeed 1.890585e-01 -3 4708.333

158 204 205 hum 5.466665e-01 -3 4106.312

159 206 207 casual 5.115248e-01 -3 4791.812

160 0 0 <NA> 0.000000e+00 -1 5054.750

161 208 209 hum 7.072920e-01 -3 4699.222

162 210 211 registered 4.639763e-01 -3 5161.083

163 0 0 <NA> 0.000000e+00 -1 5344.333

164 0 0 <NA> 0.000000e+00 -1 2493.000

165 212 213 registered 3.023390e-01 -3 2288.500

166 0 0 <NA> 0.000000e+00 -1 2169.000

[ reached 'max' / getOption("max.print") -- omitted 181 rows ]

> MAPE(RF\_predict,Y\_test)

[1] 0.05220401 -----5.2 %

**4.2 Linear Regression** :

> summary(lm\_model)

Call:

lm(formula = cnt ~ ., data = train)

Residuals:

Min 1Q Median 3Q Max

-387.98 -132.20 -28.85 57.03 1439.80

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 137.8701 91.6486 1.504 0.1331

season2 46.0595 66.3074 0.695 0.4876

season3 -15.6680 78.7644 -0.199 0.8424

season4 -40.5977 72.7956 -0.558 0.5773

yr1 50.3559 38.2823 1.315 0.1889

mnth2 -8.3310 54.6969 -0.152 0.8790

mnth3 7.9148 63.3454 0.125 0.9006

mnth4 -48.8858 93.2608 -0.524 0.6004

mnth5 -21.9637 100.7996 -0.218 0.8276

mnth6 -94.0688 105.0086 -0.896 0.3707

mnth7 -89.1360 115.2883 -0.773 0.4398

mnth8 -133.8276 112.4253 -1.190 0.2344

mnth9 35.8566 98.2565 0.365 0.7153

mnth10 15.1926 90.9716 0.167 0.8674

mnth11 -6.5852 85.5279 -0.077 0.9387

mnth12 -12.2583 66.2418 -0.185 0.8533

holiday1 -97.3597 65.2701 -1.492 0.1364

workingday1 -256.5734 42.7173 -6.006 3.42e-09 \*\*\*

weathersit2 33.1922 29.7221 1.117 0.2646

weathersit3 80.4441 81.5706 0.986 0.3245

temp 319.4232 164.7684 1.939 0.0531 .

hum -192.7500 112.5384 -1.713 0.0873 .

windspeed 0.8308 170.0572 0.005 0.9961

casual 2248.8383 95.7152 23.495 < 2e-16 \*\*\*

registered 7104.9131 125.3966 56.660 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 255.8 on 559 degrees of freedom

Multiple R-squared: 0.9835, Adjusted R-squared: 0.9828

F-statistic: 1390 on 24 and 559 DF, p-value: < 2.2e-16

> MAPE(lm\_predict,Y\_test)

[1] 0.03545968----3.5%

**Decision Tree :**

**> summary(DT\_model)**

**Call:**

**rpart(formula = cnt ~ ., data = train, method = "anova")**

**n= 584**

**CP nsplit rel error xerror xstd**

**1 0.59618462 0 1.0000000 1.0084478 0.045229526**

**2 0.20293369 1 0.4038154 0.4114661 0.020145372**

**3 0.05409479 2 0.2008817 0.2108550 0.011433640**

**4 0.02693918 3 0.1467869 0.1593397 0.010483785**

**5 0.01609060 4 0.1198477 0.1371291 0.008764885**

**6 0.01316393 5 0.1037571 0.1215173 0.008029645**

**7 0.01000000 6 0.0905932 0.1088963 0.007350386**

Variable importance

registered yr casual temp mnth season weathersit workingday

44 16 16 10 8 5 1 1

**> MAPE(DT\_predict,Y\_test)**

**[1] 0.1310449-----13.1%**

**5.Model Selection :**

By comparing all the models we could see the MAPE is least for linear regression model (3.5%).

Conclusion :

Forecasting for the future data should be done with linear regression model as this model gives higher accuracy.

.

**Appendix:**

R and Python code can be found in the submission folder.

**References :**

Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE.

*Wickham, Hadley. 2009. Ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business Media.g*