MACHINEN LEARNING

PROJECT 1

Wajiha Zafar

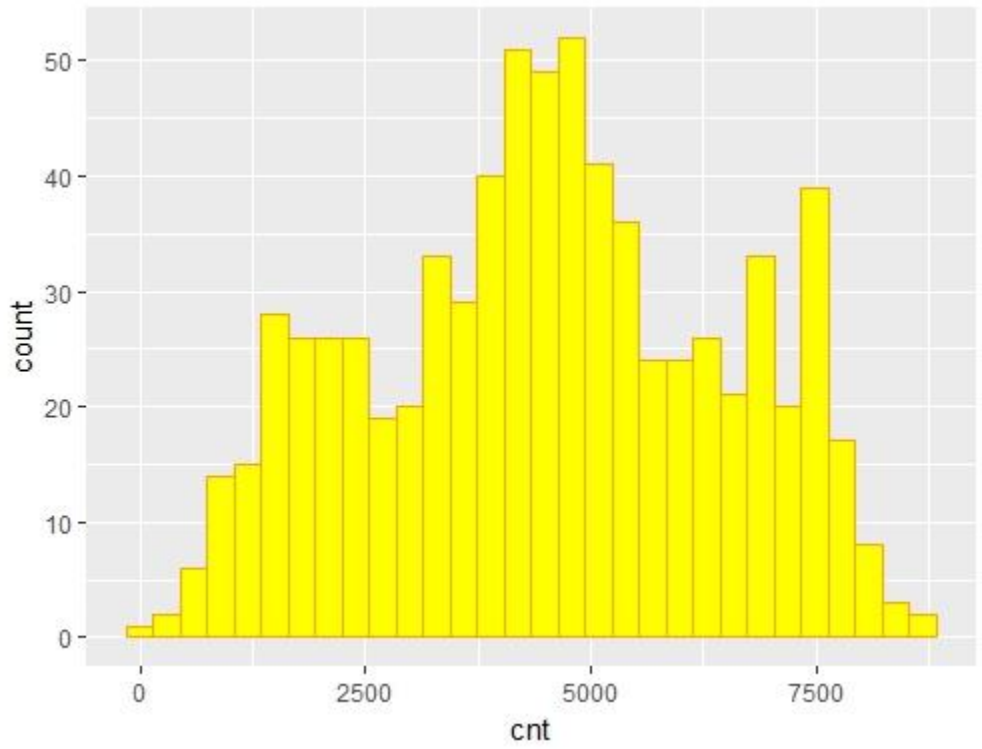
BIKE SHARING DATASET

Bike sharing is the new way of renting bike where members rent and return has become automatic. Through this system, user can easily rent bike from one point and return at other point. Right now there are 500 different bike sharing systems around the world.

I downloaded the dataset Bike Sharing Dataset from the UCI machine learning database link and performed some of the Exploratory analysis and Multivariate regression model on 4 independent variable with one dependent variable which is total bike rental.

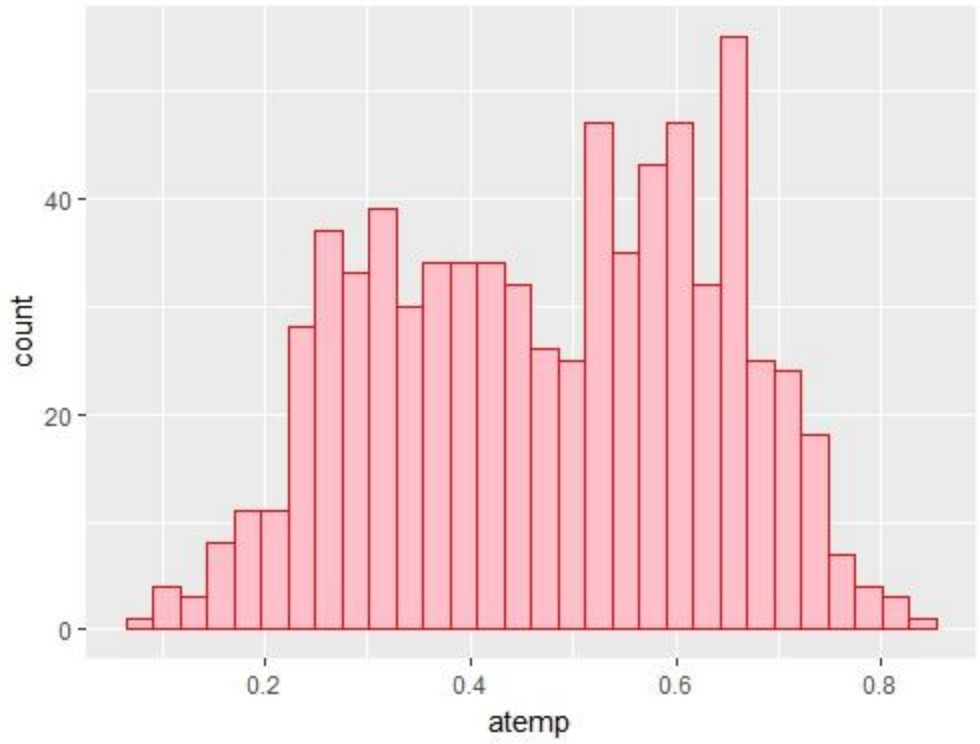
EXPLORATORY ANALYSIS

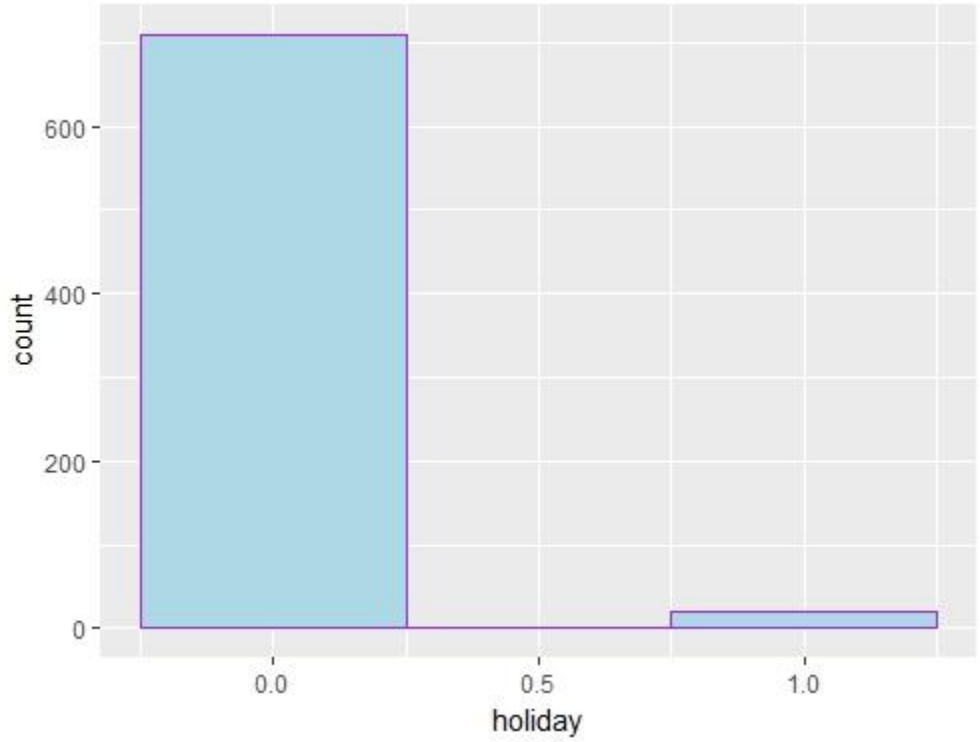
Histogram of dependent variable count which is total bike rental:



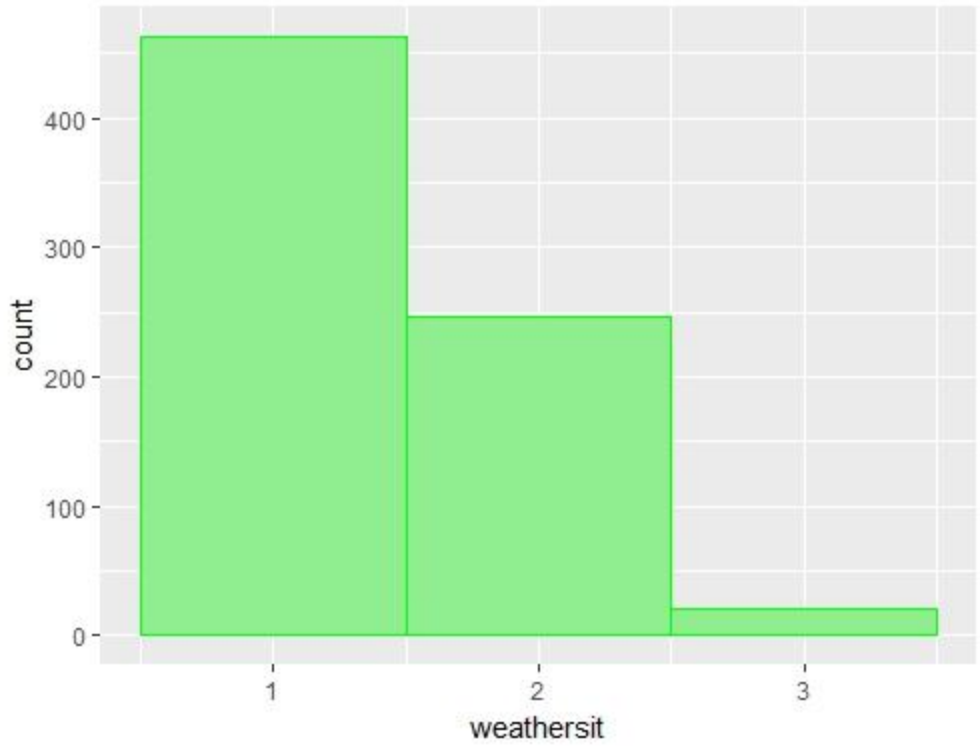
According to the plot, dependent variable is normally distributed and is also a continuous variable. This can be used as response variable in regression.

Histogram of independent variable





The holiday variable is the combination of binary values, o and 1. The o value indicates no holiday and 1 indicates about the holiday. So bike rental is increase with frequency of 0(no holiday) value.



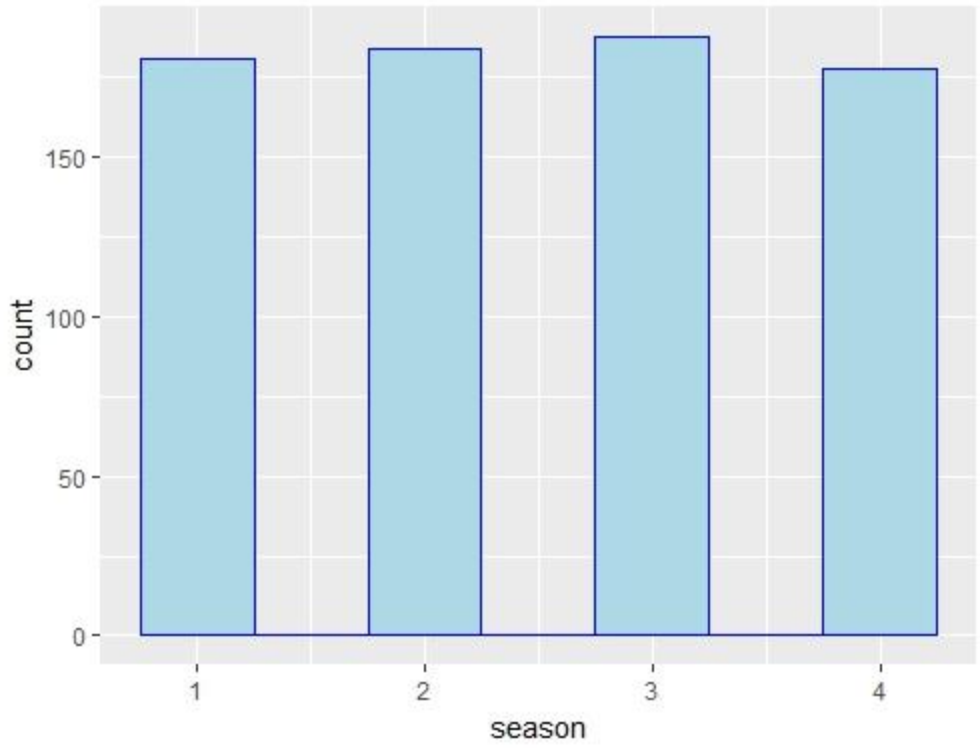
The weather situation is a combination of:

1: clear, partly cloudy, cloudy,

2: Mist and cloudy, Mist and few clouds, Mist

3: light snow, light rain and thunderstorm, lightrain and scatter clouds.

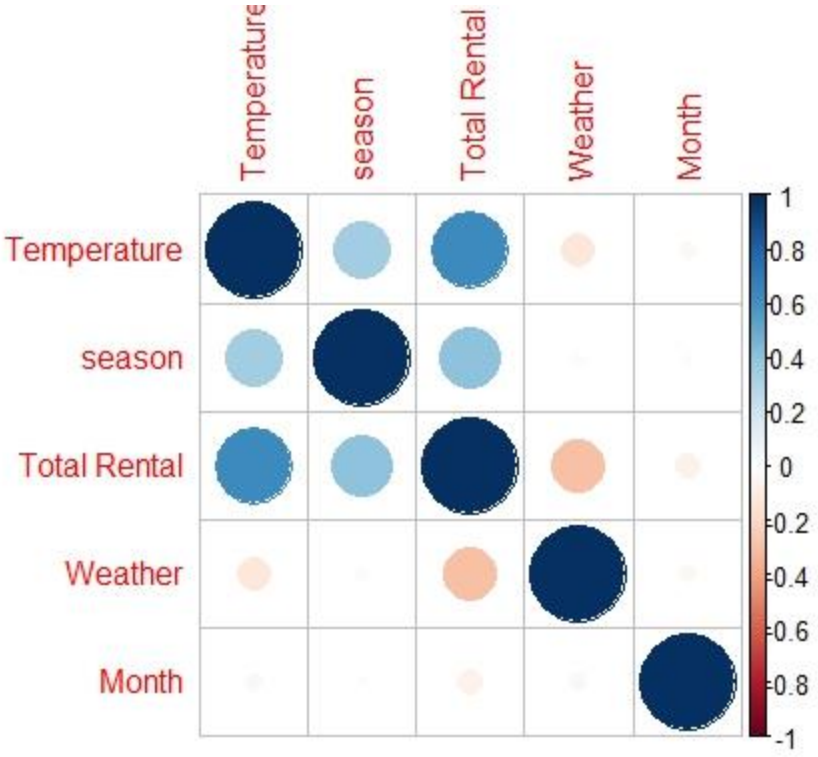
The frequency of bike rental is more in clear and partly cloudy weather as compared to the cloudy and snow and rainy weather. The bike rental is showing negative correlation with snow and rainy weather.



SUMMARY STATS:

Attributes	Mean	Median	Variance
Count	4504.35	4548	3752788
Temperature	0.4744	0.486	0.0266

CORRELATION MATRIX




```
> cor(df1)    #Correlation of Data
             Temperature      season Total Rental     Weather     Holiday
Temperature   1.00000000  0.34287561   0.63106570 -0.12158335 -0.03250669
season        0.34287561  1.00000000   0.40610037  0.01921103 -0.01053666
Total Rental  0.63106570  0.40610037   1.00000000 -0.29739124 -0.06834772
Weather      -0.12158335  0.01921103  -0.29739124  1.00000000 -0.03462684
Holiday      -0.03250669 -0.01053666  -0.06834772 -0.03462684  1.00000000
> corplot_data <- cor(df1)
```

According to the correlation plot and matrix, there is slightly positive correlation between the dependent variable and actual temperature of 0.63. There is also some correlation between the season and dependent variable of around 0.40. Where as weather situation and dependent variable showed negative correlation of -0.297 and also correlation between holiday and total rental is negative. The highest correlation is seen between the season and temperature. This shows that the change in temperature and season effect the bike rental frequency.

ANALYSIS:

MULTIVARIANT REGRESSION ANALYSIS:

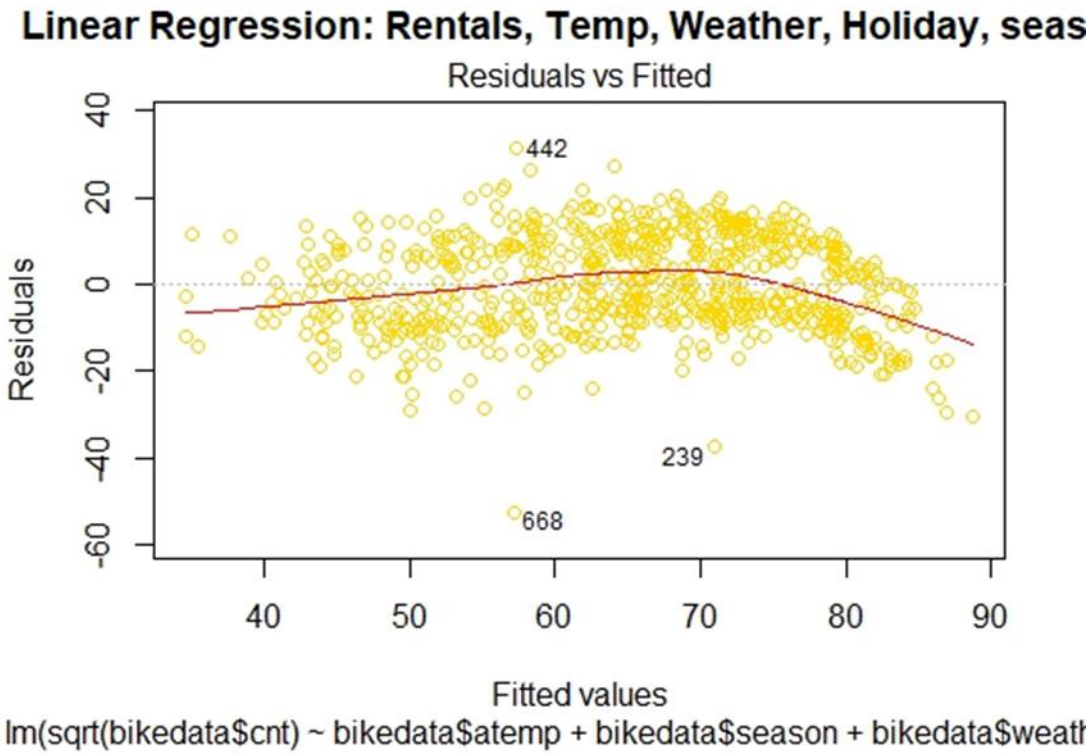
The choice of algorithm was to calculate/compute the multivariate regression analysis, because we have to predict the numerical outcome of the bike rental. The response variable in this case is Bike rental count. The regression is performed well in numeric and normal distributed variables. Linear regressions are often used to predict the influence of variables to the values. They have ability to easily identify outliers in data. The linear regression is also dependent on the quality of data points.

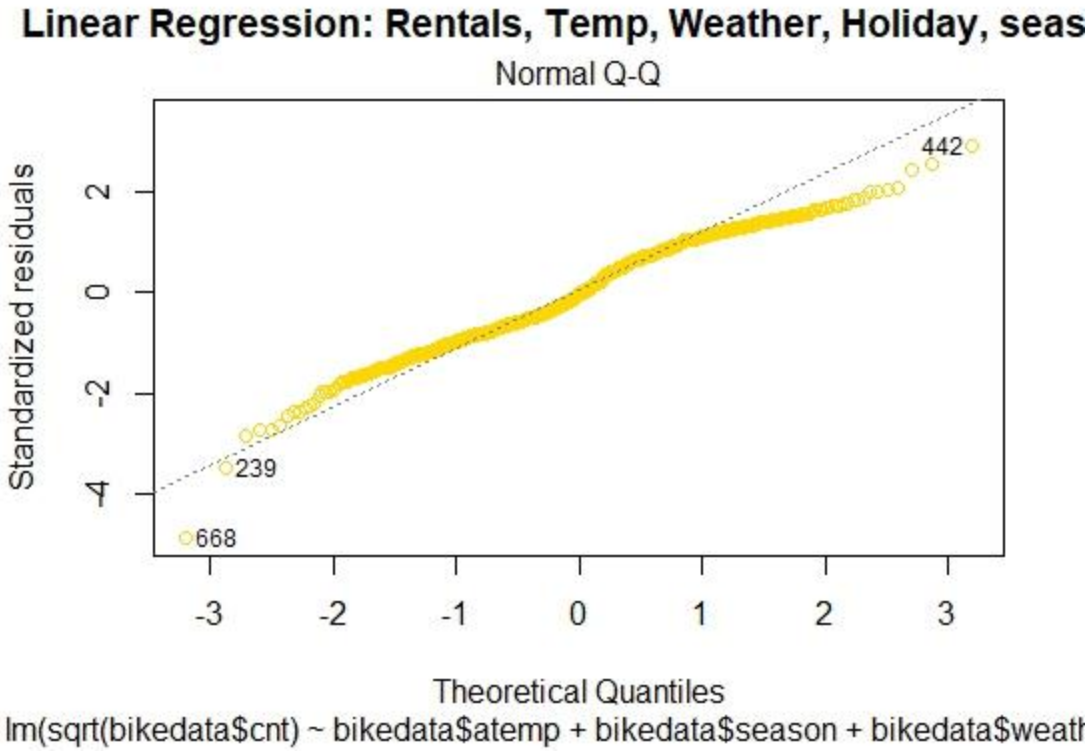
The Rsquare and RMSE from the model is : 0.5724 and 1194.066, which are not too bad because the 25th percentile of the variable is around 2500.

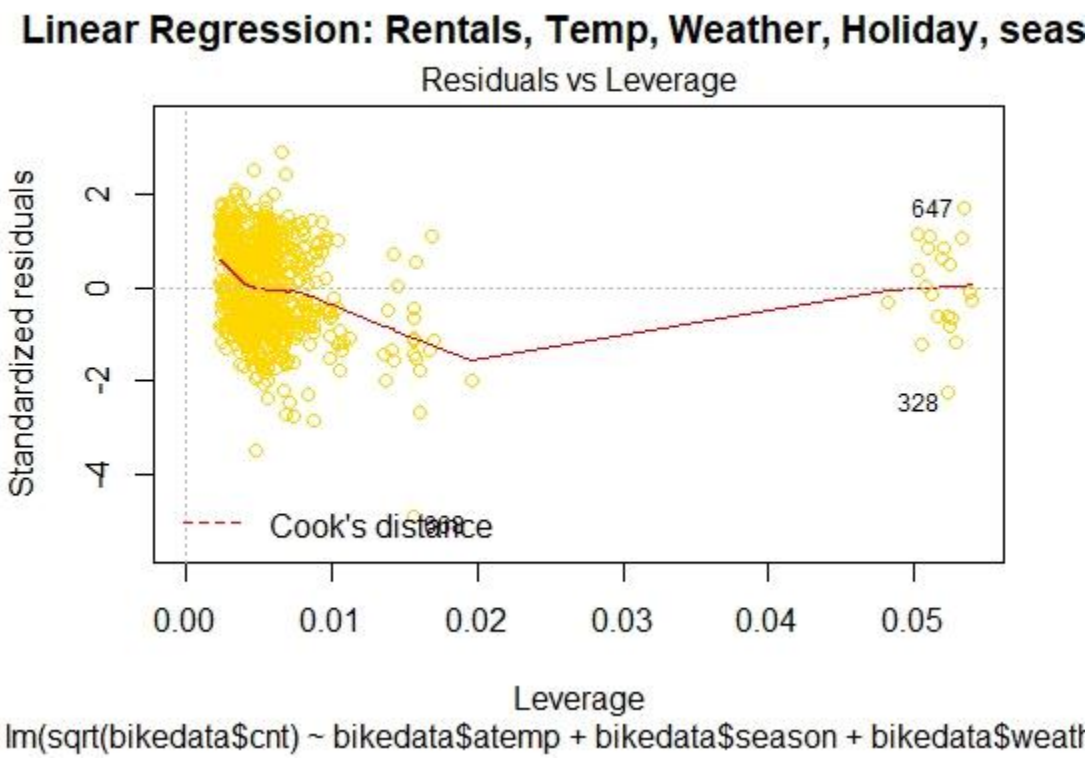
The overall dataset was not that massive have only 731 observations and after picking 4 independent variables against a response variable its not very bad.

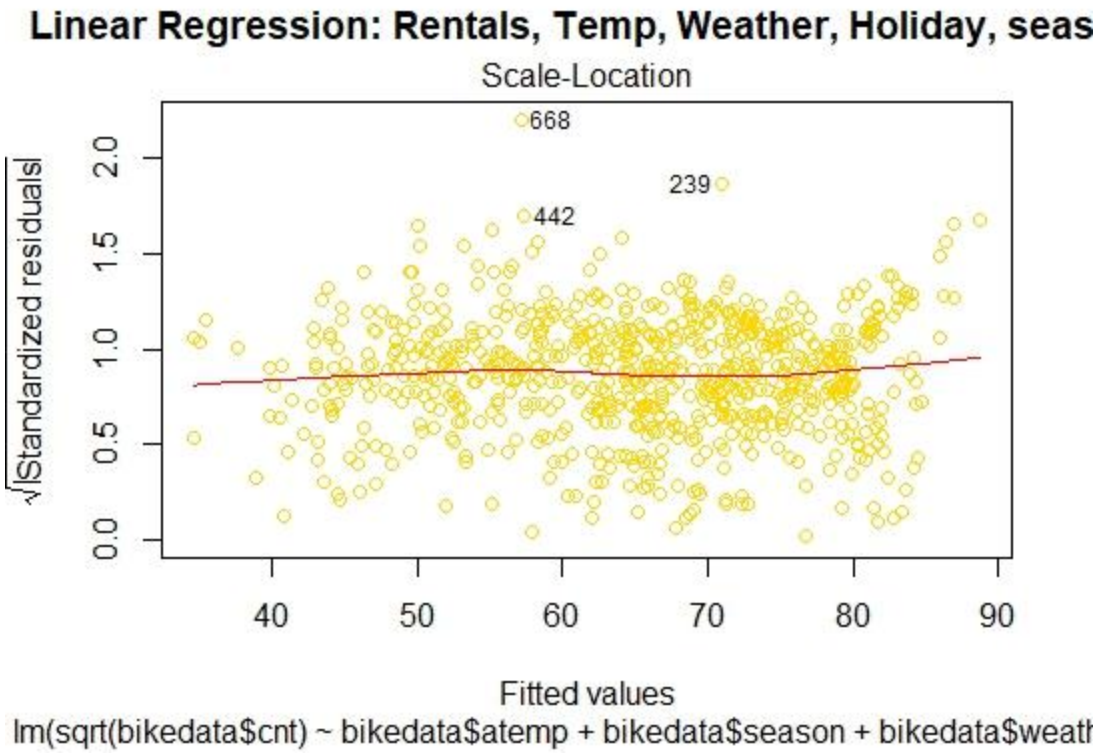
```
Call:
lm(formula = sqrt(bikedata$cnt) ~ bikedata$atemp + bikedata$season +
    bikedata$weathersit + bikedata$holiday)

Coefficients:
        (Intercept)       bikedata$atemp      bikedata$season  bikedata$weathersit
             42.713               50.968                3.504               -7.329
   bikedata$holiday
             -5.905
```




Linear Regression: Rentals, Temp, Weather, Holiday, seas
Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(sqrt(bikedata$cnt) ~ bikedata$atemp + bikedata$season + bikedata$weath
442
239
668


Linear Regression: Rentals, Temp, Weather, Holiday, seas
Residuals vs Leverage
Standardized residuals
Leverage
lm(sqrt(bikedata$cnt) ~ bikedata$atemp + bikedata$season + bikedata$weath
647
328
Cook's distance



Code:

```
install.packages("ggplot2")

install.packages("GGally")

install.packages("corrplot")

install.packages("MLmetrics")

library(ggplot2)

library(GGally)

library(corrplot)

library(scales)

library(grid)

library(RColorBrewer)

library(KernSmooth)

library(MLmetrics)
```

```
day <- read.csv("C:/Users/Wajiha/Desktop/day.csv")
bikedata <- subset(day, select = c("cnt","season","holiday","weathersit","atemp"), stringsAsFactors = F)
head(bikedata)
ggplot(bikedata, aes(x=atemp)) + geom_histogram(color = "red", fill = "pink")
ggplot(bikedata, aes(x=weathersit)) + geom_histogram(binwidth = 1, color = "green", fill = "lightgreen")
ggplot(bikedata, aes(x=season)) + geom_histogram(binwidth = 0.5, color = "blue", fill = "lightblue")
ggplot(bikedata, aes(x=cnt)) + geom_histogram(color = "orange", fill = "yellow")
ggplot(bikedata, aes(x=holiday)) + geom_histogram(binwidth = 0.5, color = "purple", fill = "lightblue")
```

#Mean and Median and Variance:

```
mean(bikedata$atemp)
mean(bikedata$cnt)
mean(bikedata$weathersit)
mean(bikedata$season)
mean(bikedata$holiday)
median(bikedata$atemp)
median(bikedata$holiday)
median(bikedata$season)
median(bikedata$cnt)
median(bikedata$weathersit)
var(bikedata$cnt)
var(bikedata$holiday)
var(bikedata$season)
var(bikedata$atemp)
var(bikedata$weathersit)
```

#Correlation matrix and plot

```
df1 <- data.frame(bikedata$atemp, bikedata$season, bikedata$cnt, bikedata$weathersit,
bikedata$holiday)
colnames(df1)[1] <- "Temperature"
```

```
colnames(df1)[2] <- "season"
colnames(df1)[3] <- "Total Rental"
colnames(df1)[4] <- "Weather"
colnames(df1)[5] <- "Holiday"
cor(df1)   #Correlation of Data
corplot_data <- cor(df1)
corrplot(corplot_data, method = "circle")  #Correlation plot
```

#Multivariant Regression

```
set.seed(20)
sample <- sample.int(n = nrow(bikedata), size = floor(0.85*nrow(bikedata)), replace = F)
trainset = bikedata[sample, ]
testset = bikedata[-sample, ]
```

#Multiple Regression function

```
regressionModel <- function(trainset,testset){
  y <- trainset$cnt
  # matrix of feature variables from dataset
  x <- as.matrix(trainset[,1:5])
  int <- rep(1, length(y))
  # add an intercept to the variables
  #x <- cbind(int, x)
  # calculate the beta / Coefficent
  beta <- solve(t(x) %*% x) %*% t(x) %*% y
  intercept<-beta[1]
  seasoncoef<-beta[2]
  holidaycoef<-beta[3]
  weathercoef<-beta[4]
  atempcoef<-beta[5]
  model <- intercept + seasoncoef * testset$season + holidaycoef * testset$holiday + weathercoef *
testset$weathersit + atempcoef * testset$atemp
```

```
  return(model)
}
Rsquare <- function(actual,predicted){
  return(1 - (sum((actual-predicted)^2)/sum((actual-mean(actual))^2)))
}
# Mean Squared erorr:
rmse <- function(error){
  return(sqrt(mean(error^2)))
}
evalMetrix<-function(test_data){
  R2 <- Rsquare(testset$cnt,testset$pred)
  RMSE<-rmse(test_data$error)
  evaldf<-data.frame(Rsquare = R2, RMSE = RMSE)
  return(evaldf)
}
# Feature Scaling
trainset["atemp"] = scale(trainset["atemp"])
testset["atemp"] = scale(testset["atemp"])
predict <- regressionModel(trainset, testset)
testset$predict <- predict
# error calculation:
testset$error <- testset$cnt - testset$predict
evaldf<-evalMetrix(test_data = testset)
evaldf
#Linear Regression Model ploting
lmtest<-
lm(sqrt(bikedata$cnt)~bikedata$atemp+bikedata$season+bikedata$weathersit+bikedata$holiday)
lmtest
plot(lmtest,col = "gold", main = "Linear Regression: Rentals, Temp, Weather, Holiday, season")
```

References:

1. http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization
2. http://www.theanalysisfactor.com/r-tutorial-recoding-values/
3. http://pingax.com/logistic-regression-r-step-step-implementation-part-2/
4. **http://a-little-book-of-r-for-bayesian-statistics.readthedocs.io/en/latest/src/bayesianstats.html#bayesian-statistics**