ASSIGNMENT 1

Vaishnav Mule
A20516627

1 Recitation Exercises

1.1 Chapter 2

Exercise: 1

1.
I think a **Flexible Model** would suit better to solve this problem. As the sample size is huge the chance of overfitting even with a flexible model is very less. As the number predictors are less the flexible model will allow for a better training and understanding from the data.

2.
I think an **inflexible Model** would be a best fit. Firstly, if we were to choose a flexible model it would cause a problem of overfitting as the sample size is too small. By using a flexible model our output would result in large variance hence an inflexible model would best suit this problem.

3.
A **Flexible model** will be a perfect fit for this problem. The non-linear relationship will need more data to be understand and draw a pattern out of it, and flexible models deal will large sample data. By using a Flexible model, the value of error would be reduced to my knowledge.

4.
An **Inflexible Model** would best suit this problem. Flexible models invite a lot of noise due to the large datasets; Hence an Inflexible model would be better as compared to reduce the variance and avoid overfitting.

Exercise: 2

1.
REGRESSION : Clearly it Is a case of finding the relation between a few attributes and how they would effect the CEO salary.
INFERENCE : We are interested in inference as the question asks us to understand what factors are influencing the CEO salary.
n = 500 , p = 3

2.
CLASSIFICATION : The result we want to derive is a choice between success or failure
PREDICTION : Our main aim is to predict the success or failure and not to determine the relation between the factors effecting for that result.
n = 20, p = 13

3.
REGRESSION : we are focusing on a single output which is continuous in nature
PREDICTION : we want to predict the % change exchange rate, it is not necessary that we know what factors are effecting it .
n = 52, p = 3

Exercise: 4

1.

a) Brain Tumour detection : we can use a CNN classification Model to Classify if a given x-ray of brain, has a tumour or not. The predictors in this case will be the size/shape/density/location of the possible tumour in an x-ray. The response is a simple yes/no reflecting the presence of a tumour respectively. The goal of this problem is to predict the presence of a brain tumour.

b) Helmet Detector(or mask detector) : we can use a face Detection model to classify if a person is wearing a helmet or mask while riding a bike or in public places. This is good model used in India as a lot of motor cycles to reach form one place to another. The predictors can be featuring embeddings of the localized facial region. The output is a simple yes or no determining whether the person is wearing a helmet or not. The goal of this problem is to predict the presence of a helmet.

c) Acceptance Into a University : It is a common tool used by almost all the students who are applying for MS in US. With the help of a classification Model we can predict if the student's profile will be accepted or not. The predictors can be attribute such as UG grade/GRE score/ IELTS score/Funds available. The output is a yes/no to tell if the student is eligible for the university's benchmarks. This is a Prediction problem again, to predict acceptance into a university.

2.

a) Time Taken to Reach a Destination : We can use a regression model that will take In parameters such as the Traffic/vehicle type/speed/weather conditions and then give an estimation of time taken to reach the destination. This will be a prediction problem, predicting the time taken.

b) Stock Market Analysis : We can use the regression Model that will accept inputs as the economy/closing price/companies year performance/dividends/financial books records and then predict the stock price for the next one month or year. This is a Prediction Problem.

c) Fuel for airlines : We can use Regression model to predict the amount of fuel an airline requires for it's journey, The inputs can be Attributes such as payload, runtime, passengers onboard, the pre-flight distance etc. This is a prediction problem.

3.

a) OTT movie Suggestion: We can use a Clustering model to bring out suggestions as to which genre the account holder might prefer watching from. The Clusters can be Horror films or Adventure Films or Romantic films. It will be a prediction base algorithm.

b) Product marketing: While launching a new product into the market it is important to target the right audience, Using clustering we can study what kind of audience use that type of products in the market. The inputs can be the product, salary range, location, utility. It is an Inference based model.

c) Health Insurance : Health insurance is a sector where clustering model can be used to approach the right set of people with the correct policies. Depending on the size of families/ family Income/ age / sex The clustering model can be designed.

Exercise: 6

->Parametric Statistical Models are the models built when there are a set of know attributes form the data provided. NON Parametric on the other hand has no such defined set of parameters and we build the models based on assumptions. In parametric learning models we run a function through the model and then understand the data and their relations, in case of a non parametric learning model we have no such function to run by, the machine learns about the data and tries to build a function that suits.

->Advantages on Regression : as the parameters are already present it becomes an easy to determine what function is suitable for the problem the prediction becomes an easier task. Even with limited set

of data it becomes possible to predict required outcomes. Linear models work best with parametric models.
->Disadvantages on Regression : sometimes the huge number of parameters makes it difficult to fit into the model, Variances maybe pretty high.
->Advantages on Classification : parameters help define better boundaries and gives a clearer picture on the classification.

Exercise: 7

a)
observation 1: √ (0-0)^2 + (3-0)^2 + (0-0)^2 = 3.0

observation 2: √ (2-0)^2 + (0-0)^2 + (0-0)^2 = 2.0

observation 3: √ (0-0)^2 + (1-0)^2 + (3-0)^2 = 3.162

observation 4: √ (0-0)^2 + (1-0)^2 + (2-0)^2 = 2.236

observation 5: √ (-1-0)^2 + (0-0)^2 + (1-0)^2 = 1.414

observation 6: √ (1-0)^2 + (1-0)^2 + (1-0)^2 = 1.732

b)
k=1 :
with k=1, the nearest neighbour of our test point is obs 5 . since it is green , we say out test point is also green.

c)
k=3 :
three nearest observations are 2, 5 and 6. Since the majority is red, we say test point is also red.

d)
As we keep increasing k, the bayes decision boundary gets inflexible. So if boundary is highly non-linear, we expect k to be small.

1.2 Chapter 3

Exercise: 1

For the table 3.4, according to null hypothesis the sales are not effected by the budget allotted for TV, Radio, or Newspaper. The p value is less than 0.0001 for TV and Radio because of which we can conclude that Null Hypothesis gets rejected in this situation, TV and Radio do effect the Sales. However, for Newspaper the p value is 0.8599 so we can not reject the null hypothesis and hence it is concluded that Newspaper budget does not affect the sales.

Exercise: 3

Using the information provided :

Y’ = 50 + 20GPA + 0.07IQ + 35Gen +0.01GPA * IQ – 10GPA * Gen

Males: (0)
Y’ = 50 + 20GPA +0.07IQ +0.01GPA * IQ

Females: (1)
Y’ = 85 +10GPA + 0.07IQ + 0.01GPA * IQ

a) i) No Proof
   ii) No Proof
   iii) 50 + 20GPA >= 85 + 10GPA
      10GPA >= 35
      GPA >= 3.5
      ➔ correct
   iv) Incorrect

b) IQ = 110 and GPA = 4.0
   Y’ = 85 + 10(4) + 0.07(110) + 0.01(4)(110)
   ➔ 137.1 = $ 137,100

c) False. To evaluate the true impact , we must conduct the hypothesis test and look at the p-value to conclude.

Exercise: 4

a) The cubic model is complex and flexible so we know it will have a small RSS on the training set when compared to Linear Model. Due to the true linear nature the cubic model might have very high chances that it might overfit the training data. On the other hand, the linear model will generalize the true relation between X and Y in actual situations.
b) The True relationship between the predictors and response is linear in nature hence the linear model will have a significantly lower RSS. Since the cubic model is likely to overfit the RSS for test set has chances to be high due to memorized training test.
c) As the flexibility of model increases the train RSS reduces so if the relation is not linear flexible models will definitely reduce in train RSS. But in test we must make sure to find the perfect spot where RSS is lowest.
d) In test RSS, If the relation Is linear in nature then test RSS will be higher for cubic model. And similarly, if the nature is close to cubic in nature, then test RSS will be low for cubic models. Bottom line, it depends on the dataset’s nature.

---

2 Practicum Problems

2.1 Problem 1

Vaishnavi

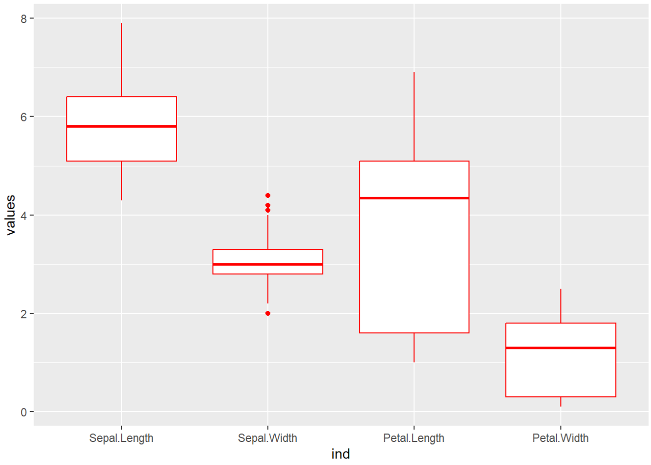
2023-09-08

➔ Loading dataset:

```
library(datasets)
data(iris)
summary(iris)
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
##        Species  
##  setosa    :50  
##  versicolor:50
##  virginica :50
##
##
##
```

➔ box plots:

```
library('ggplot2')
ggplot(data = stack(iris), mapping = aes(x=ind, y = values)) + geom_boxplot(color = 'red')
## Warning in stack.data.frame(iris): non-vector columns will be ignored
```



➔ IQR for all 4 attributes:

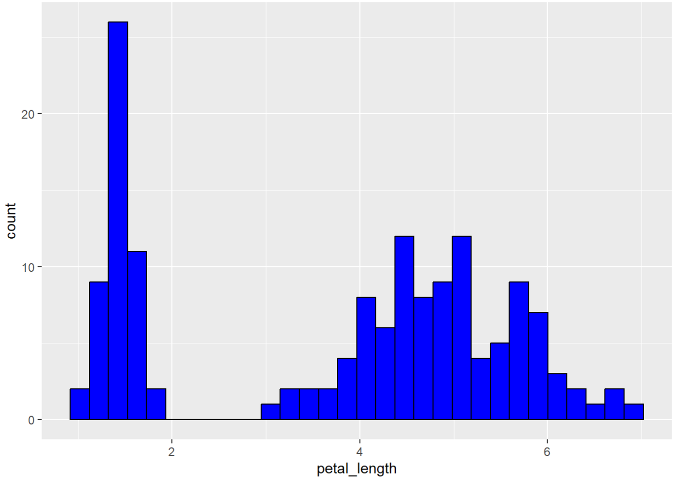
```
sepal_width = iris$Sepal.Width
sepal_length = iris$Sepal.Length
petal_width = iris$Petal.Width
petal_length = iris$Petal.Length
sprintf("IQR of Sepal Width: %.3f", IQR(sepal_width))
## [1] "IQR of Sepal Width: 0.500"
sprintf("IQR of Sepal Length: %.3f", IQR(sepal_length))
## [1] "IQR of Sepal Length: 1.300"
sprintf("IQR of Petal Width: %.3f", IQR(petal_width))
## [1] "IQR of Petal Width: 1.500"
sprintf("IQR of Petal Length: %.3f", IQR(petal_length))
## [1] "IQR of Petal Length: 3.500"
```

Answer: Petal length has highest IQR = 3.500

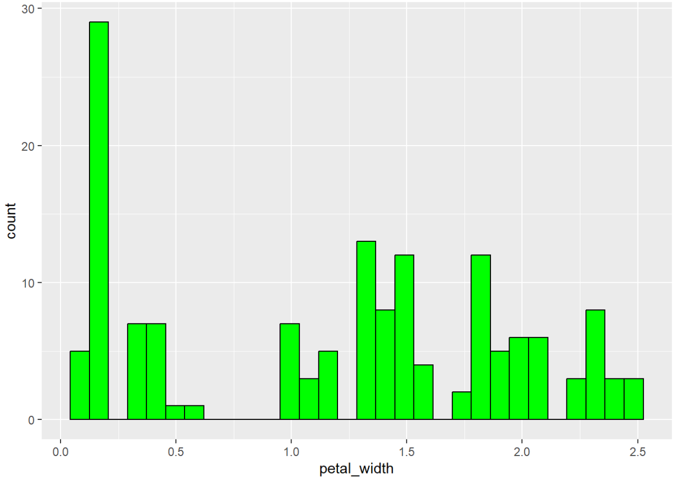
➔ Standard Deviation for all features:

```
sprintf("SD of Sepal Width: %.3f", sd(sepal_width))
## [1] "SD of Sepal Width: 0.436"
sprintf("SD of Sepal Length: %.3f", sd(sepal_length))
## [1] "SD of Sepal Length: 0.828"
sprintf("SD of Petal Width: %.3f", sd(petal_width))
## [1] "SD of Petal Width: 0.762"
sprintf("SD of Petal Length: %.3f", sd(petal_length))
## [1] "SD of Petal Length: 1.765"
ggplot(iris, aes(x=petal_length)) +geom_histogram(color="black",fill="blue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



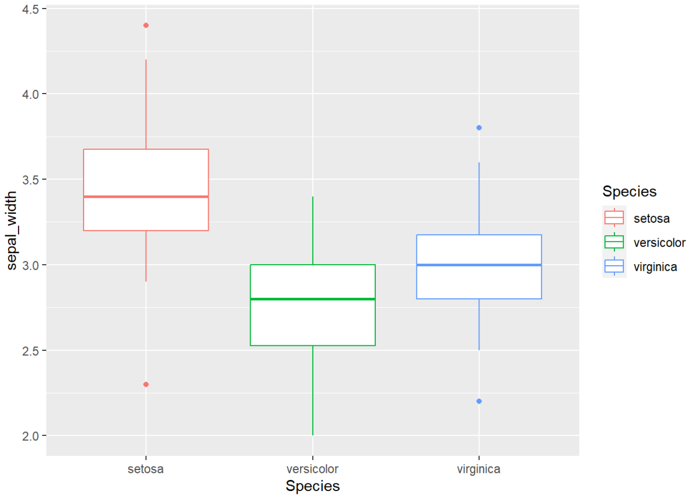
```
ggplot(iris, aes(x=petal_width)) +geom_histogram(color="black",fill="green")
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



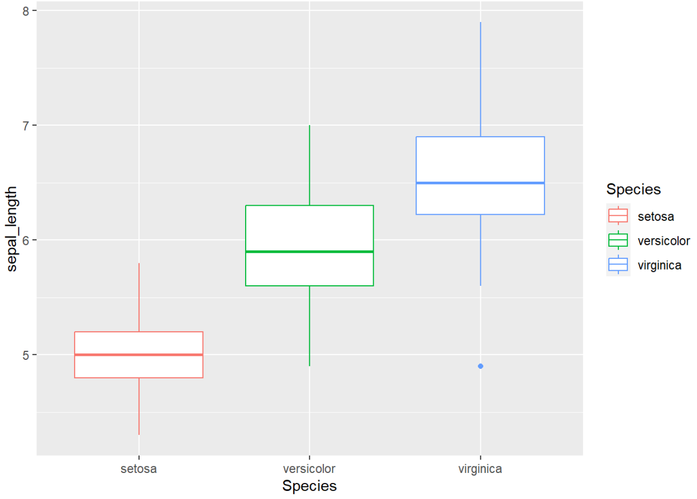
answer : Both Petal length and petal width follow different distributions as observed from the standard deviation values obtained. Hence the findings do not match the empirical values.

➔ Boxplots for each feature in the iris dataset, with box-whisker per flower spicies.

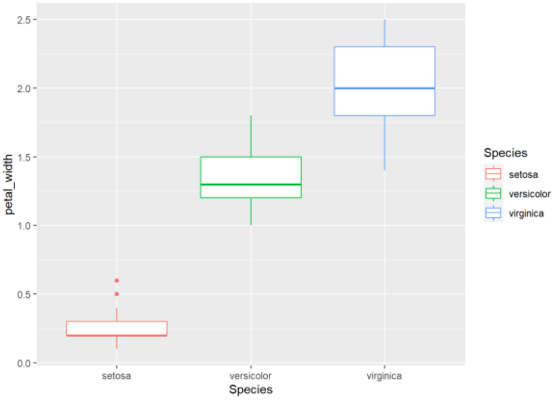
```
Ggplot(data=iris) + geom_boxplot(mapping = aes(x=Species,y=sepal_width, color = Species))
```



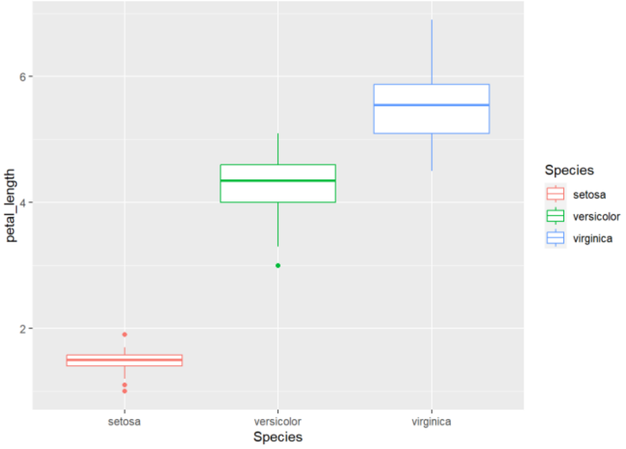
```
ggplot(data=iris) + geom_boxplot(mapping = aes(x=Species,y=sepal_length, color = Species))
```



```
ggplot(data=iris) + geom_boxplot(mapping = aes(x=Species,y=petal_width, color = Species))
```



```
ggplot(data=iris) + geom_boxplot(mapping = aes(x=Species,y=petal_length, color = Species))
```



Answer = Virginica (Blue)

2.2 Problem 2

Vaishnavi

2023-09-08

➔ Loading Dataset :

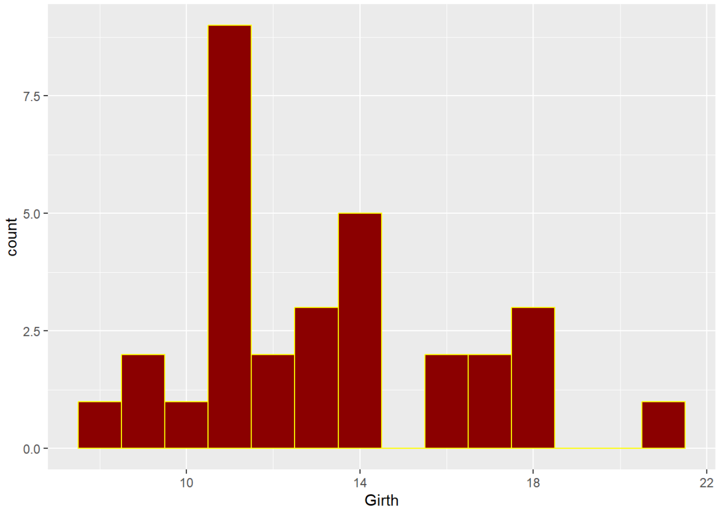
```
library(datasets)
library('ggplot2')
data(trees)
summary(trees)
##      Girth          Height       Volume
##  Min.   : 8.30   Min.   :63   Min.   :10.20
##  1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
##  Median :12.90   Median :76   Median :24.20
##  Mean   :13.25   Mean   :76   Mean   :30.17
##  3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
##  Max.   :20.60   Max.   :87   Max.   :77.00
```

➔ 5-number summary of all features :

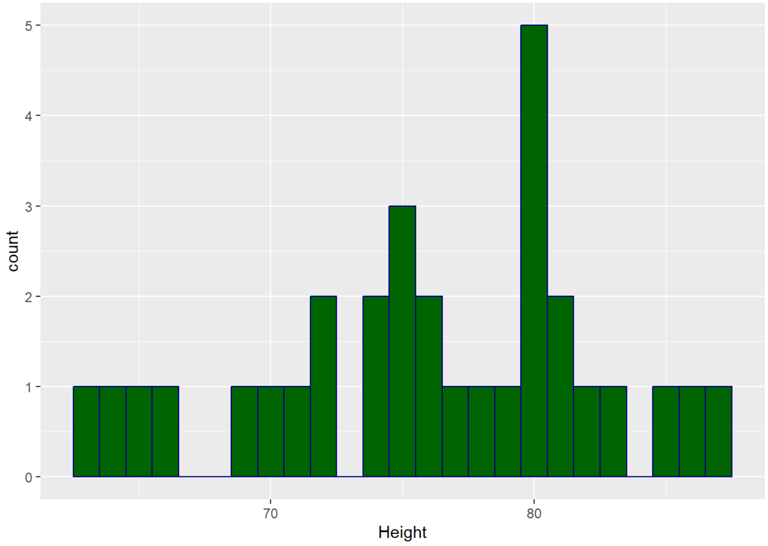
```
print("5-number summary of Girth")
## [1] "5-number summary of Girth"
fivenum(trees$Girth)
## [1]  8.30 11.05 12.90 15.25 20.60
print("5-number summary of Height")
## [1] "5-number summary of Height"
fivenum(trees$Height)
## [1] 63 72 76 80 87
print("5-number summary of Volume")
## [1] "5-number summary of Volume"
fivenum(trees$Volume)
## [1] 10.2 19.4 24.2 37.3 77.0
```

Histrogram for each feature :

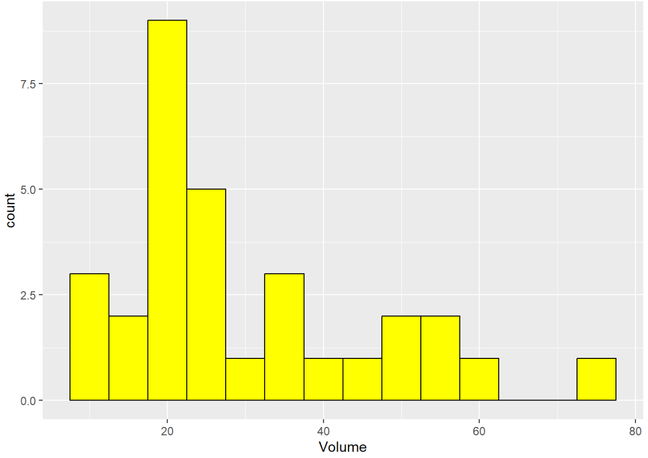
```
ggplot(trees, aes(x=Girth)) + geom_histogram(color="yellow", fill="darkred", binwidth = 1)
```



```
ggplot(trees, aes(x=Height)) +
  geom_histogram(color=”darkblue”, fill=”darkgreen”, binwidth = 1)
```



```
ggplot(trees, aes(x=Volume)) +
  geom_histogram(color=”black”, fill=”yellow”, binwidth = 5)
```



Answer : Height has a normal distribution. Girth and Volume exhibit positive skewness.

- ➔ Skewness of each feature :

```
library(“moments”)

10print(“Skewness of Girth: %.3f”, skewness(trees$Girth))
## [1] “Skewness of Girth: 0.526”
10print(“Skewness of Height: %.3f”, skewness(trees$Height))
## [1] “Skewness of Height: -0.375”
10print(“Skewness of Volume: %.3f”, skewness(trees$Volume))
## [1] “Skewness of Volume: 1.064”
```

Answer : The values match with those of the plots (Histograms)

2.3 Problem 3

Vaishnavi

2023-09-08

- ➔ Load auto-mpg dataset using UCI data repo :

```
library('ggplot2')

url="https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data"


auto_data <- read.csv(file=url, header=FALSE, sep="", as.is =4&9,

   col.names = c("mpg","cylinders","displacement","horsepower","weight","acceleration", "model_year", "origin", "car name"))
```

- ➔ Converting horsepower column from string to numeric data type. And the replacing all NaN values with the median :

```
auto_data$horsepower = as.numeric(auto_data$horsepower)

## Warning: NAs introduced by coercion

na_replaced_hp = auto_data$horsepower


na_replaced_hp[which(is.na(na_replaced_hp))] = median(na_replaced_hp, na.rm =TRUE)
```

- ➔ Comparing old and new means :

```
mean(auto_data$horsepower, na.rm=TRUE)

## [1] 104.4694

mean(na_replaced_hp)

## [1] 104.304
```

Answer : The old mean is calculated ignoring NA values in the dataset, While the new Mean is calculated filling up the NA values with median hence the change in the mean values.

2.4 Problem 4

Vaishnavi

2023-09-08

➔ Loading dataset-Boston :

```
library(MASS)
library('ggplot2')
data(Boston)
attach(Boston)
summary(Boston)
```

```
##      crim                zn             indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##      nox               rm             age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##      rad               tax           ptratio          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##     lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

➔ Fitting regression between medv and istat :

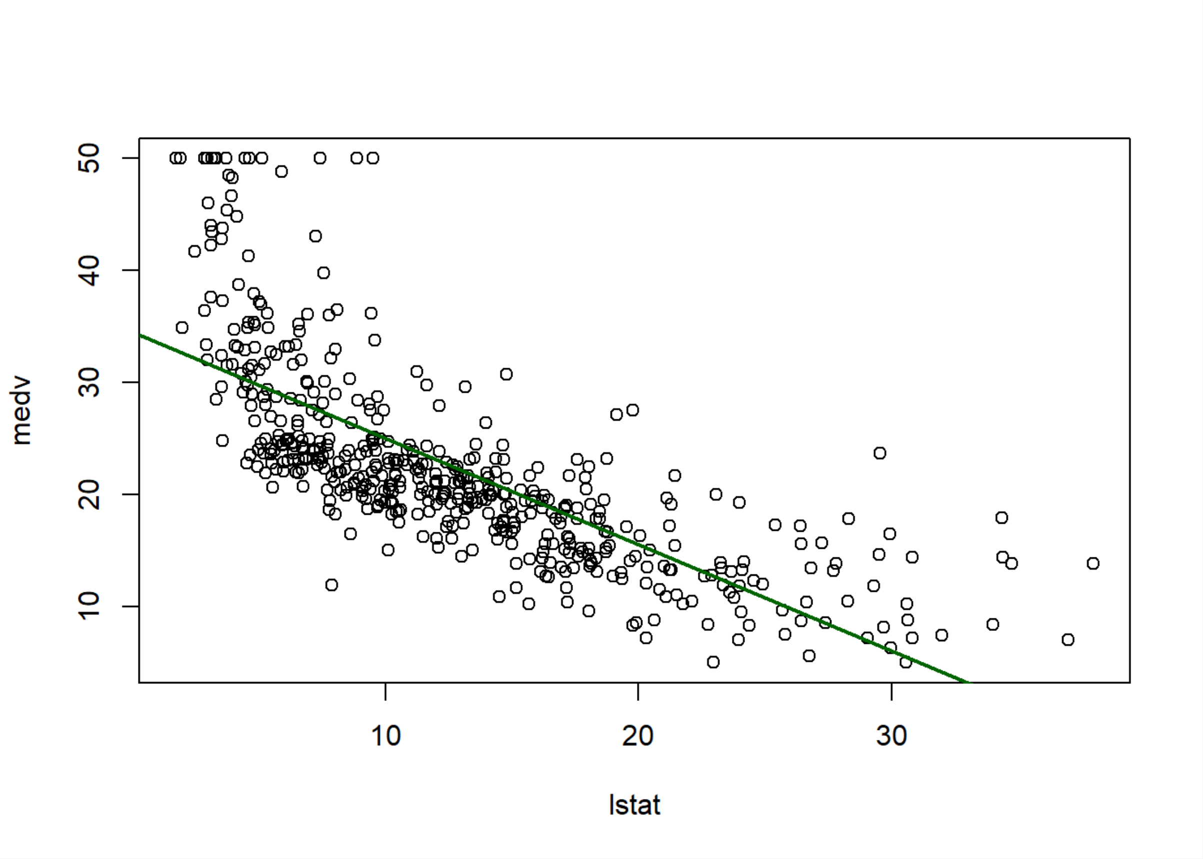
```
lm.fit=lm(medv~lstat)
summary(lm.fit)
##
## Call:
## lm(formula = medv ~ lstat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

➔ Plot the fitted values vs residual :

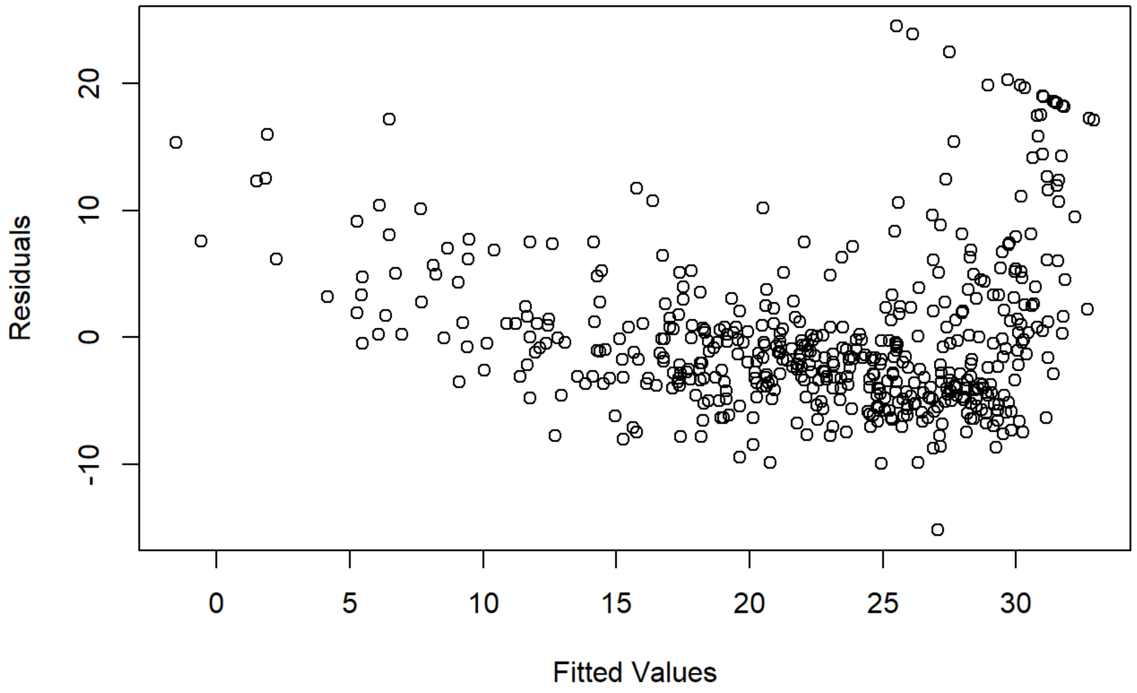
```
plot(lstat ,medv)

abline (lm.fit)


abline (lm.fit ,lwd = 2, col = 'darkgreen')
```



```
plot(predict (lm.fit), residuals (lm.fit), xlab="Fitted Values", ylab="Residuals")
```



Answer: Looking at the plot above, the relationship between medv and lstat does depict some non-linearity.

- ➔ Predicting reponse values for lstst = 5, 10, and 15. And finding confidence and prediction intervals.

```
predict(lm.fit ,data.frame(lstat=c(5,10,15) ),interval="confidence")
##        fit      lwr      upr
## 1 29.80359 29.00741 30.59978
## 2 25.05335 24.47413 25.63256
## 3 20.30310 19.73159 20.87461
predict(lm.fit ,data.frame(lstat=c(5,10,15) ),interval="prediction")
##        fit       lwr      upr
## 1 29.80359 17.565675 42.04151
## 2 25.05335 12.827626 37.27907
## 3 20.30310  8.077742 32.52846
```

Answer: Confidence and Prediction interval essentially predicts the response so they do the same thing, but the interpretation is different. The prediction interval will be wider than the confidence interval since it must account for the variability in the mean and standard deviation estimators.

- ➔ Including lstat^2

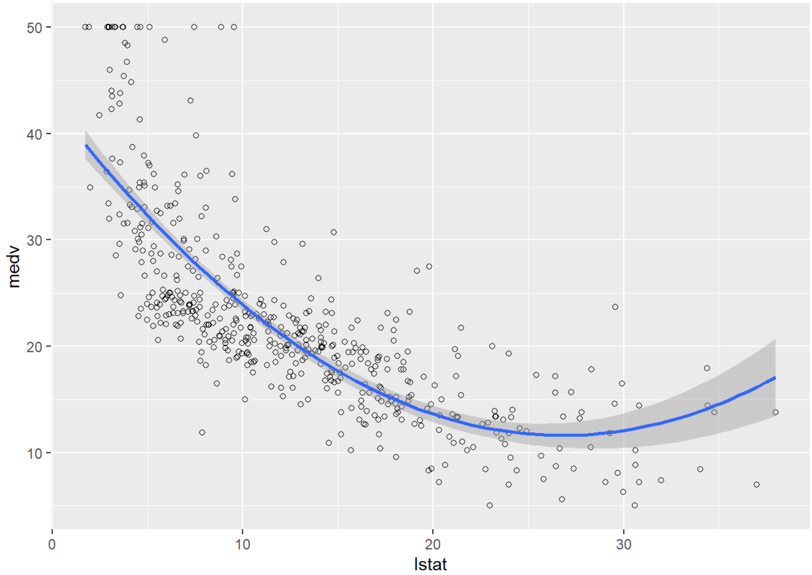
```
flexible_lm.fit=lm(medv~lstat+I(lstat^2),data=Boston)
summary(flexible_lm.fit)
##
## Call:
```

```
## lm(formula = medv ~ lstat + I(lstat^2), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2834  -3.8313  -0.5295   2.3095  25.4148
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.862007   0.872084   49.15   <2e-16 ***
## lstat       -2.332821   0.123803  -18.84   <2e-16 ***
## I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.524 on 503 degrees of freedom
## Multiple R-squared:  0.6407, Adjusted R-squared:  0.6393
## F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

Answer : Linear Model adjusted r^2 = 0.5432 non-linear Model adjusted r^2 = 0.6393

➔ Plotting relation between predictor and response :

```
ggplot(Boston, aes(x = lstat, y = medv)) + geom_point(shape=1) +
stat_smooth(method = 'lm', formula = y~x+poly(x,2))
## Warning in predict.lm(model, newdata = data_frame0(x = xseq), se.fit = se, :
## prediction from a rank-deficient fit may be misleading
```


medv
lstat