Data Preparation and Analysis(CS571)

Assignment-1

1. **Recitation Exercises**
   Chapter-2

1. (a) The sample size n is extremely large, and the number of predictors p is small.

In the above example, the performance of a flexible statistical learning method would be better because the size of the sample is extremely large. The usage of the inflexible method will result in overfitting the data.

(b)The number of predictors p is extremely large, and the number of observations n is small.

In the above example, the inflexible method would perform better than the flexible statistical method as the number of observations is small.

(c) The relationship between the predictors and response is highly non-linear.

In the above example, the performance of the flexible statistical learning method would be better because the relationship between predictors and response is given to be non-linear so there will be room for more shapes to estimate "f".

(d) The variance of the error terms, i.e. σ2 = Var(ϵ), is extremely high.

In the above example, Var(ϵ) is given to be extremely high. In this case, it would be better if we use the inflexible method as using the flexible statistical method would result in overfitting of data.

2(a) We collect a set of data on the top 500 firms in the US. For each firm, we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

The given scenario is a regression problem, and we want to know the inference in this case because we want to know the different factors that affect the salary of the CEO (our target). Here N=500 and p=3.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product, we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

The given scenario is a classification problem, and we want to know what the prediction in this case (Success or Failure) will be. In this case, we are just concerned about the outcome. Here N=20 and p=13.

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

The given scenario is a regression problem, and we want to know the prediction in this case because we want to know the change in the exchange rate i.e. one particular value. Here N=52(total no of weeks) and p=3.

4(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. Today it will rain or not. In this case, the response will be either a Yes or a No. Here, the goal will be prediction as we are just concerned with the final output. The predictors are geographical area, temperature of place, humidity and weather conditions.
2. We are considering launching a new product and wish to know whether it will be a success or a failure. Here the response will be that either the product is a success or a failure. The goal will be prediction. The predictors are product type, product brands, pricing, marketing, and budget.

3. Covid-19 vaccine trials will be successful or not. In this case, the response will be whether the trial is successful or not-did the person gets vaccinated for covid-19 or not. The goal is prediction, and the predictors will be age, general health conditions, blood sugar levels, test groups, etc.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. Pricing of houses in the next 3-5 years can be a case in which regression can prove to be useful. The response will be the price at which a house is sold in one year suppose 2019, then at what price it might be sold the next year, and so on. The predictors will be the average size of the family, location, parks, mode of transportation available nearby, family income, crime rate, etc. The goal will be inference.

2. Mileage of a car/scooter can be another case in which regression is useful. The response will be mileage and based on different parameters we will get a numerical value for it. The predictors will be fuel type, type of engine, power, and cylinders. The goal will be inference.

3. GDP increase/growth in the economy in a certain year. This can be one of the cases of regression. The response will be the GDP of countries in a given by certain year say e.g. 2035. The predictors will be Population, per capita income, and education. The goal is will be inference.

(c) Describe three real-life applications in which cluster analysis might be useful.

1. Cluster analysis can prove to be useful in earthquake studies. We can check for dangerous zones by clustering them into different earthquake-affected areas.

2. Cluster analysis can be used to divide countries of the world into developing, developed, underdeveloped and third world.

3. Cluster analysis can be used to identify shopping patterns for customers by identifying what all items are brought by customers more frequently.

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

Parametric Statistical Learning	**Non- Parametric Statistical Learning**
1. It reduces the problem of estimating f down to one of estimating set of parameters. It simplifies the problem of estimating f.	1. It does not make explicit assumptions about the functional form of f. Instead, it seeks an estimate of f that is close to data points as possible without being too rough or wiggy.
2. yi=β0+β1xi+ei	2. yi=f(xi)+ei

Advantages of Parametric approach

Its advantage to classification or regression is simplifying model f to few parameters because not many observations are required as compared to the non-parametric approach.

Disadvantages of the Parametric approach

Its disadvantage to classification or regression is an inaccurate estimate of f if the form of f is assumed to be wrong or to overfit the observations if a flexible model is chosen.

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Euclidean Distance=√($x2 - x1$) 2 + ($y2 - y1$) 2 + ($z2 - z1$) 2

Observation	X1	X2	X3	Y	Distance	Rank
1	0	3	0	Red	3	5
2	2	0	0	Red	2	3
3	0	1	3	Red	3.16	6
4	0	1	2	Green	2.23	4
5	-1	0	1	Green	1.41	1
6	1	1	1	Red	1.73	2

(a)Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0

Distance=√(X1 − 0) 2 + (X2 − 0) 2 + (X3− 0) 2

(b) What is our prediction with K = 1? Why?

Choosing the nearest option in this case which is observation 5 i.e., 1.41. Y associated with this observation is green. Hence, our prediction is green.

(c) What is our prediction with K = 3? Why?

Choosing the nearest option in this case as well. In this case we have more than one observation i.e., observation 5(distance=1.41), and observation 6(distance=1.73) and observation 2(distance=2). Y associated with these observations are Green, Red, and Red respectively. Since the Red color is in majority, our prediction is Red.

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

As K becomes larger the boundary would become linear which is inflexible. Hence, in that case, we would expect the best value of K to be smaller.

Recitation Exercises
Chapter-3

1. **Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model**.

	Coefficient	Std. error	t-statistic	p- value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
Radio	0.189	0.0086	21.89	<0.0001
Newspaper	-0.001	0.0059	-0.18	0.8599

The null hypothesis states that there is no relationship between x and y. It is represented by the formula

H0: β1 = 0

(a)The null hypothesis (TV)- TV ads have no effect on sales in the presence of radio ads and newspaper ads.

(b) The null hypothesis (Radio)- Radio ads have no effect on sales in the presence of TV ads and newspaper ads.

(c)The null hypothesis (Newspaper)- Newspaper ads have no effect on sales in the presence of radio ads and TV ads.

The Null Hypothesis is rejected for TV and radio because TV and radio have low p- values. However, the newspaper has a high p-value and in that case, the Null Hypothesis holds true for the newspaper.

3. Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get βˆ0 = 50, βˆ1 = 20, βˆ2 = 0.07, βˆ3 = 35, βˆ4 = 0.01, βˆ5 = −10.

(a) Which answer is correct, and why?

X1=GPA

X2=IQ

X3=Level 1 for college and 0 for high school

X4=Interaction between GPA and IQ

X5=Interaction between GPA and Level

$\boldsymbol{Y\hat{}} = \boldsymbol{\hat{\beta}0 + \hat{\beta}1\, X1 + \hat{\beta}2 X2 + \hat{\beta}3 X3 + \hat{\beta}4 X4 + \hat{\beta}5\, X5}$

College= 50+20(GPA)+0.07(IQ)+35(1) +0.01(GPA*IQ) +(-10(GPA*1))

50+20(GPA)+0.07(IQ)+35+0.01(GPA*IQ) +(-10(GPA)

85+10(GPA)+0.07(IQ)+0.01(GPA*IQ) (1)

High School= 50+20(GPA)+0.07(IQ)+35(0) +0.01(GPA*IQ) +(-10(GPA*0))

50+20(GPA)+0.07(IQ)+0.01(GPA*IQ) +(-10(0))

50+20(GPA)+0.07(IQ)+0.01(GPA*IQ) (2)

Equating (1) and (2) we get

85+10(GPA)+0.07(IQ)+0.01(GPA*IQ) = 50+20(GPA)+0.07(IQ)+0.01(GPA*IQ)

85-50= 20(GPA)-10(GPA)

35= 10 GPA

3.5= GPA

(iii) is the correct answer.

(b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0

Y=50+20(GPA)+0.07(IQ)+35(1) +0.01(GPA*IQ) +(-10(GPA*1))

50+20(GPA)+0.07(IQ)+35+0.01(GPA*IQ) +(-10(GPA)

85+10(GPA)+0.07(IQ)+0.01(GPA*IQ)

85+10(4.0) +0.07(110) +0.01(4.0* 110) = 137.1=$137100

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

The above statement is false. The coefficient value for GPA/IQ interaction term does not provide evidence for the effect of interaction. We need to know the p-value for a given t-statistic for checking any evidence of interaction effect.

4. I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e., Y = β0 + β1X + β2X2 + β3X3 + ϵ.

(a) It is mentioned that the relationship between X and Y is linear, but we are not given much information about the training data. Hence, we can say (or assume) that the least square fitted line would be closer to the true regression line when it is compared with the cubic regression. So, RSS for linear regression will be lower than cubic regression.

(b) It is mentioned that the relationship between X and Y is linear but again we are not given much information about test data just like we were not given much information for training data. Linear regression will have a lower test RSS given that the relationship between X and Y is linear. Here also test RSS for cubic regression would be higher.

(c) It is mentioned that the true relationship between X and Y is non-linear. Hence, we can say that the regression line that follows closely the points in the 2D area would have a lower RSS. Cubic regression may follow the points more closely, So, the training RSS for cubic regression would be lower than the training RSS for linear regression.

(d) We don't have enough information available to tell if the test RSS for linear regression will be lower or the test RSS for cubic regression will be lower. It is not given what level of flexibility will fit the data better. If it is closer to linear, then the test RSS for linear regression will be lower else, if it is closer to cubic, then the test RSS for cubic regression will be lower.

2. Practicum Problems

Problem 1

library(datasets)

data(iris)

head(iris)

```
head(iris)
Sepal.Length Sepal.Width Petal.Length Petal.Width Species
         5.1         3.5          1.4         0.2  setosa
         4.9         3.0          1.4         0.2  setosa
         4.7         3.2          1.3         0.2  setosa
         4.6         3.1          1.5         0.2  setosa
         5.0         3.6          1.4         0.2  setosa
         5.4         3.9          1.7         0.4  setosa
```

summary(iris)

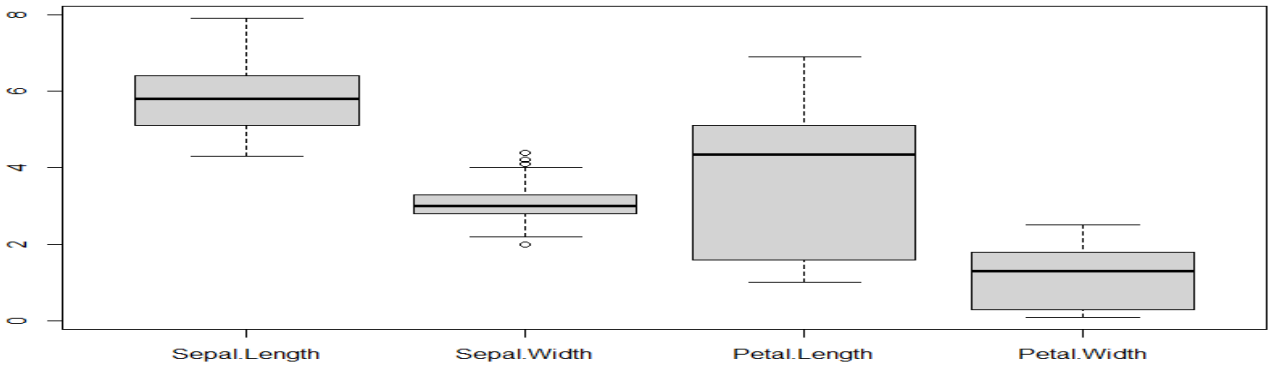
```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 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
> library(ggplot2)
```

str(iris)

```
> str(iris)
data.frame':   150 obs. of  5 variables:
$ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
$ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
$ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
$ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
$ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> boxplot(iris[, 1:4])
```

#Create a box plot of each of the four features and highlight the feature with the largest empirical IQR

boxplot (iris[, 1:4])



colnames(iris)

SLen <-iris$Sepal.Length

SWidth <-iris$Sepal.Width

PLen <-iris$Petal.Length

PWidth <-iris$Petal.Length

```
> colnames(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
> SLen <-iris$Sepal.Length
> SWidth <-iris$Sepal.Width
> PLen <-iris$Petal.Length
> PWidth <-iris$Petal.Length
```

Calculating IQR of all 4 features

IQR(SLen)

IQR(SWidth)

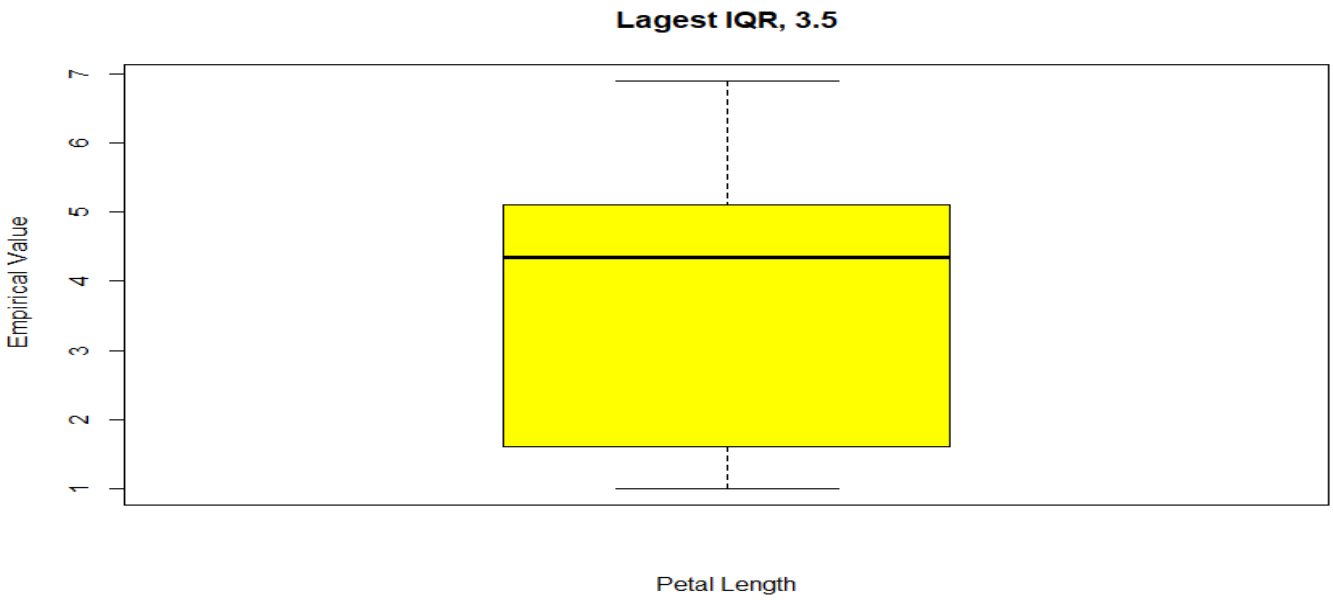
IQR(PLen)

IQR(PWidth)

```
> IQR(SLen)
[1] 1.3
>
> IQR(SWidth)
[1] 0.5
>
> IQR(PLen)
[1] 3.5
>
> IQR(PWidth)
[1] 3.5
>
```

Highlighting Petal length feature with largest empirical IQR

boxplot(PLen, main="Lagest IQR, 3.5", xlab="Petal Length", ylab=" Empirical Value", col="yellow")



Calculating the parametric standard deviation for each feature. Do your results agree with the empirical values

sd(SLen)

sd(SWidth)

sd(PLen)

sd(SWidth)

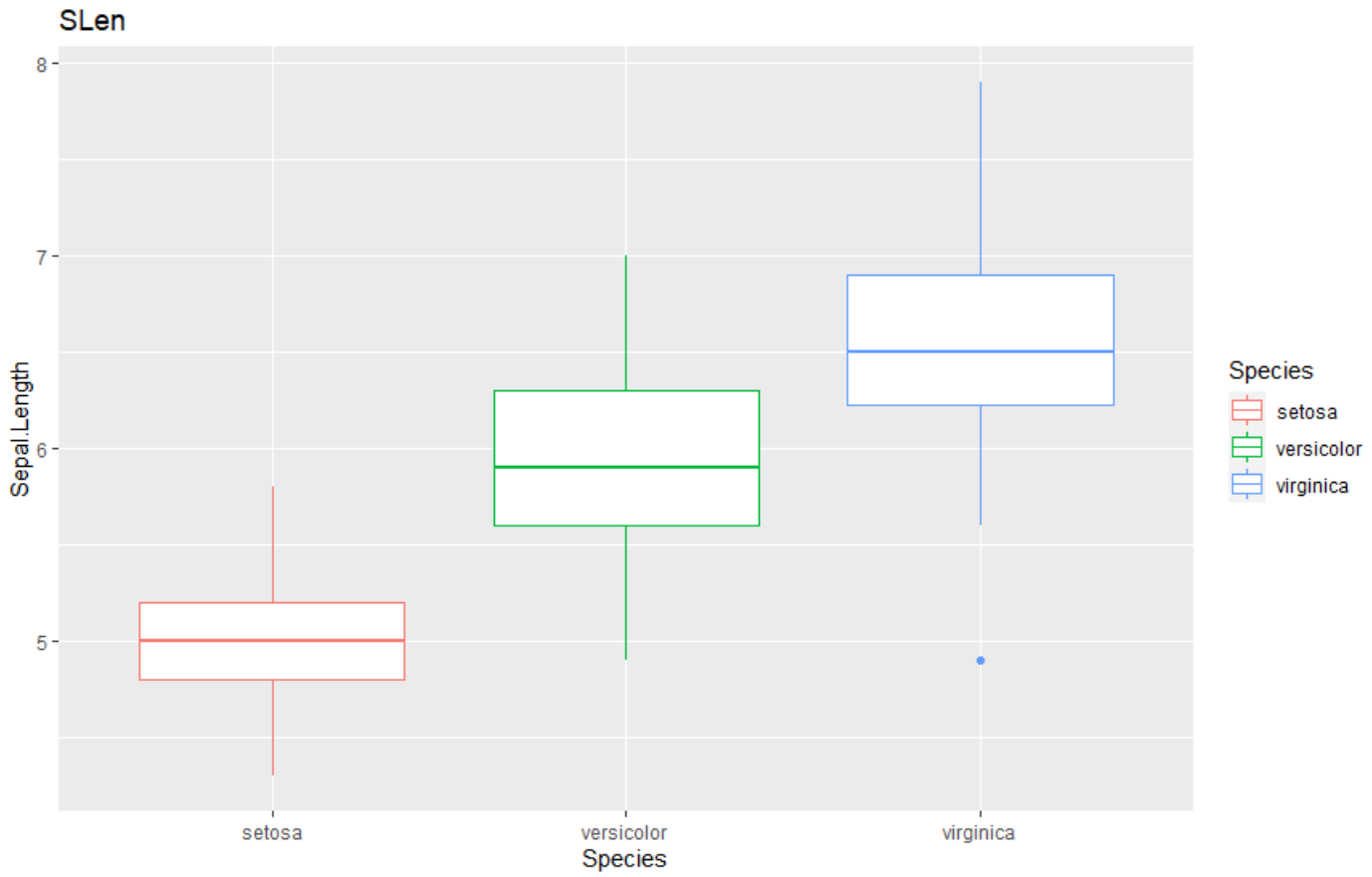
```
> #Calculating the parametric standard deviation for each feature. Do your results agree with the empirical values
>
> sd(SLen)
[1] 0.8280661
>
> sd(SWidth)
[1] 0.4358663
>
> sd(PLen)
[1] 1.765298
>
> sd(PWidth)
[1] 1.765298
>
```

Ans- As seen above, the **Standard Deviation(sd)** results do not agree with empirical values. We were asked to compare the parametric function of standard deviation with values and parametric functions are assumed to follow a normal distribution, we can see by the result that Petal Length and Petal width do not follow the normal distribution and hence normal distribution cannot be assumed for two different features of a Petal.

Use the ggplot2 library from CRAN to create a colored boxplot for each feature, with a box-whisker per flower species. Which flower type exhibits a significantly different Petal Length/Width once it is separated from the other classes

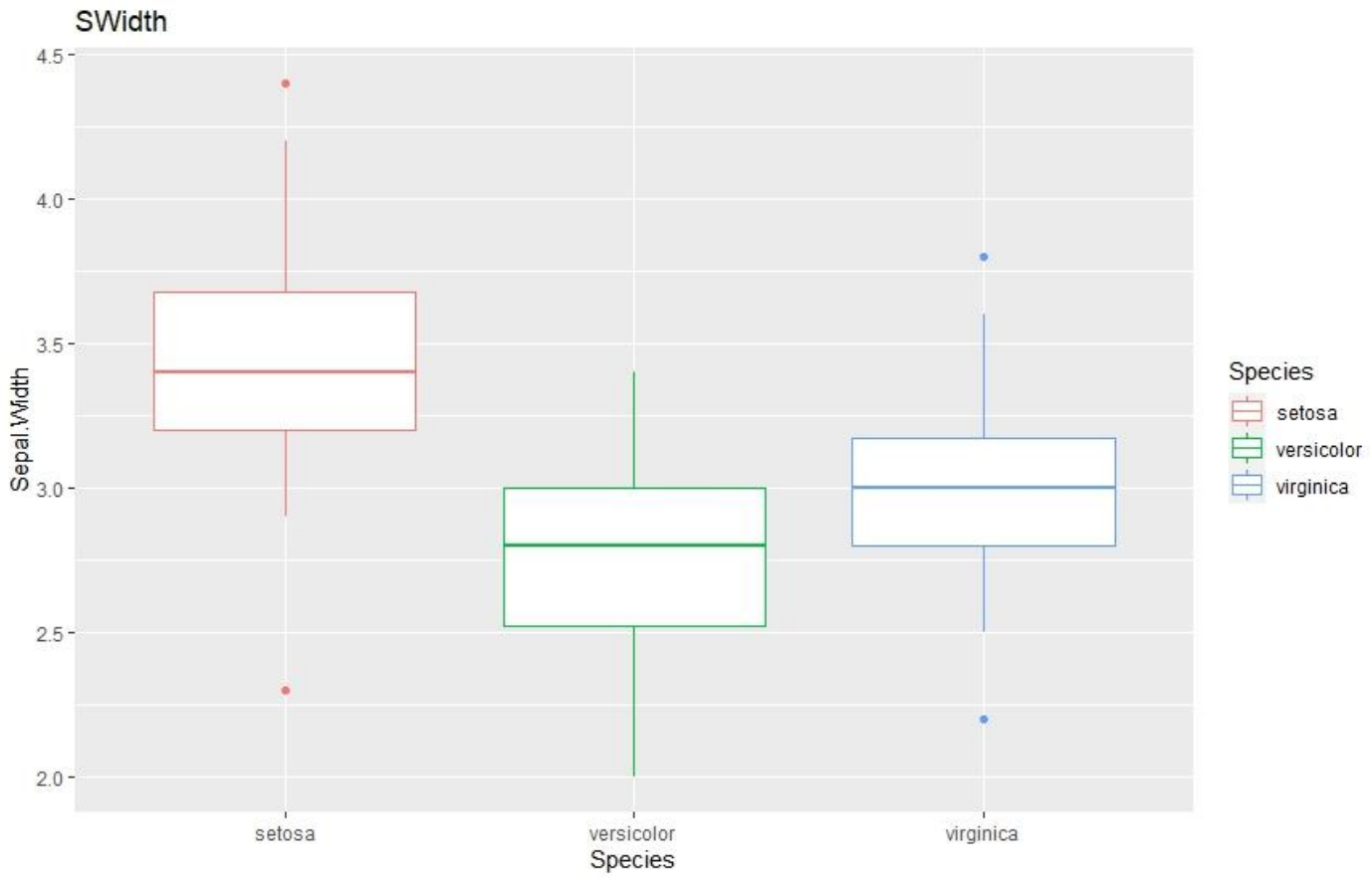
library(ggplot2)

ggplot(data = iris, aes(x = Species, y = Sepal.Length, color = Species)) + geom_boxplot() + ggtitle("SLen")



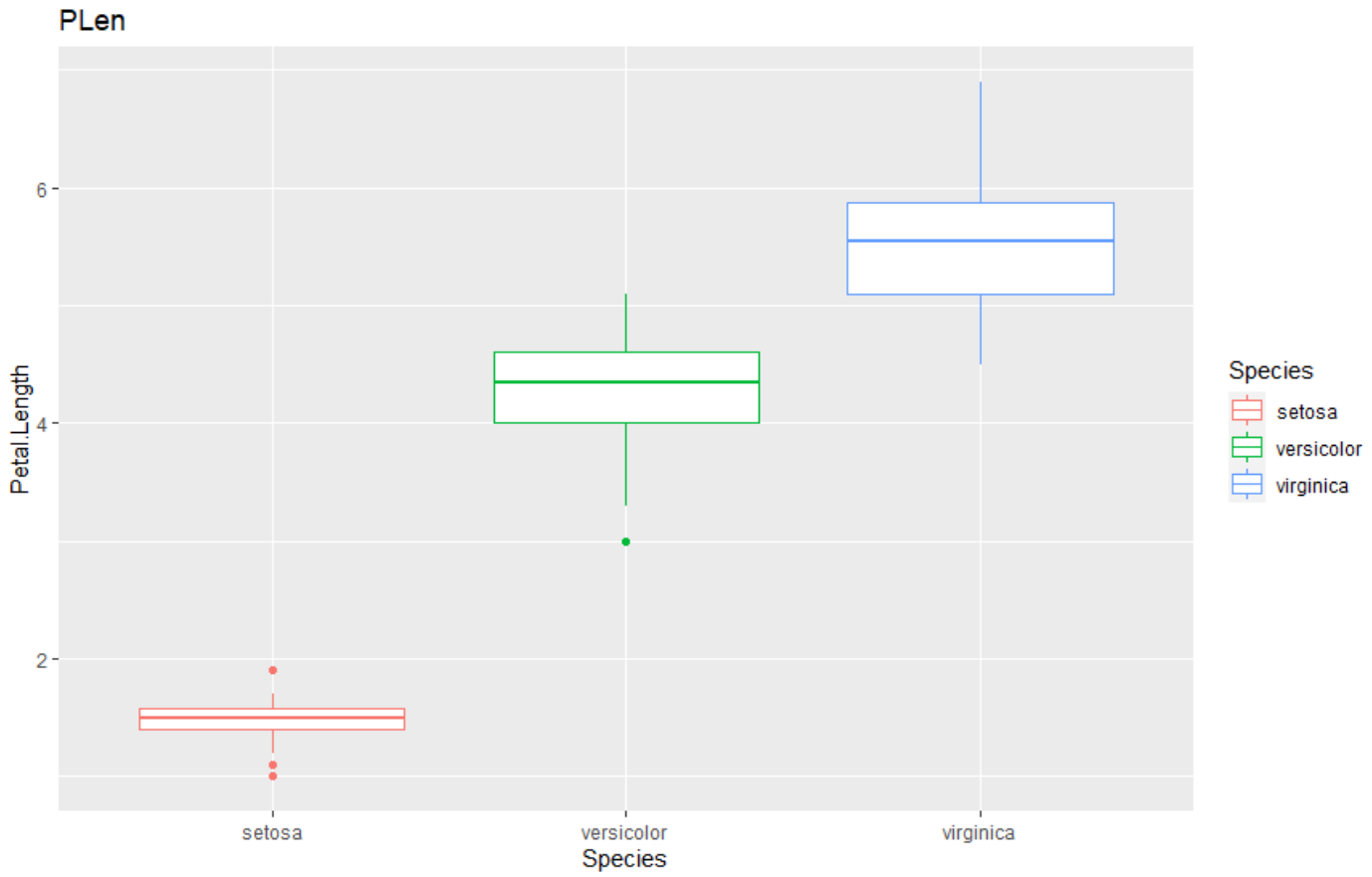
ggplot(data = iris, aes(x = Species, y = Sepal.Width, color = Species)) +

geom_boxplot() + ggtitle("SWidth")



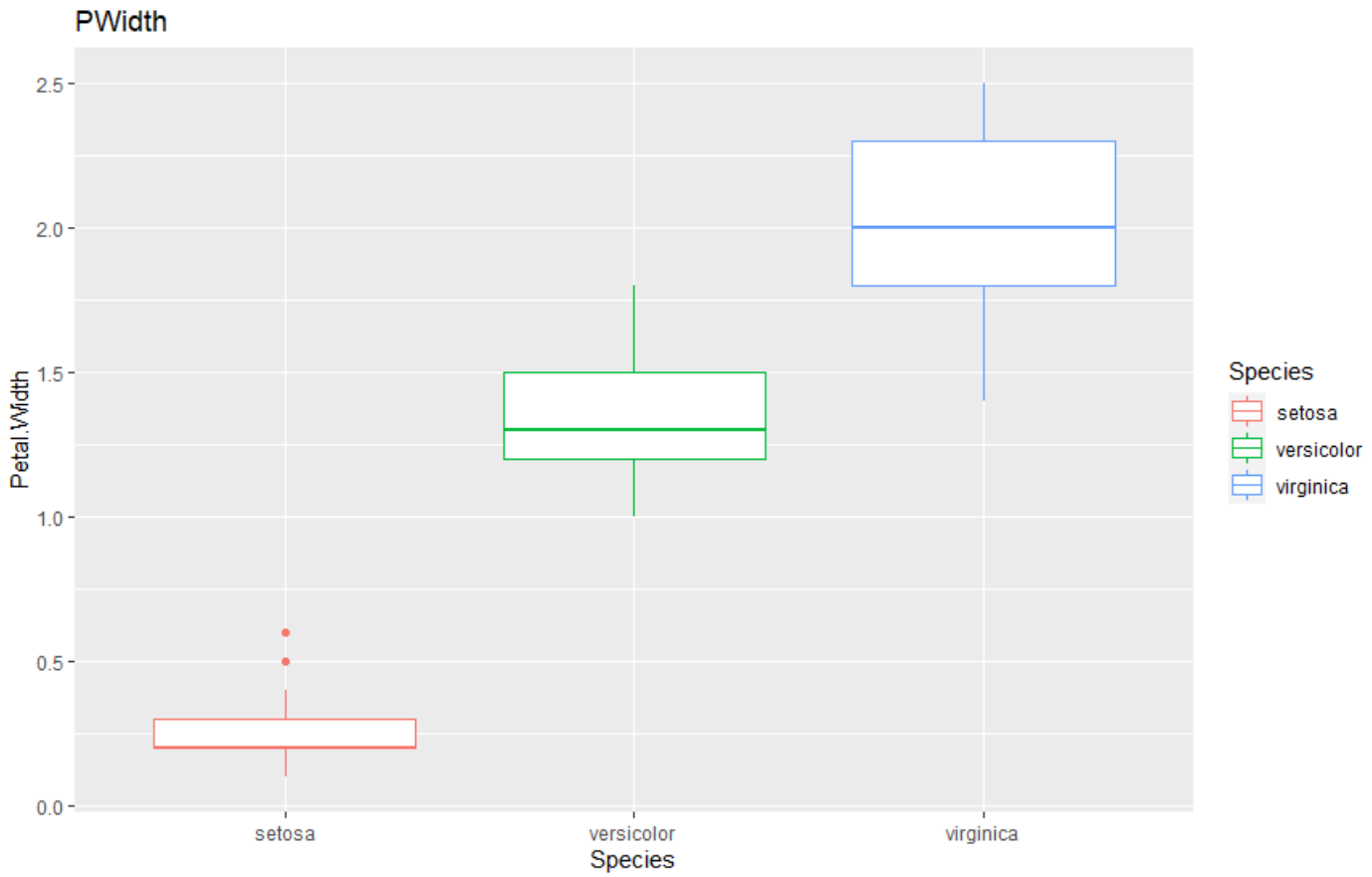
ggplot(data = iris, aes(x = Species, y = Petal.Length, color = Species)) +

geom_boxplot() + ggtitle("PLen")



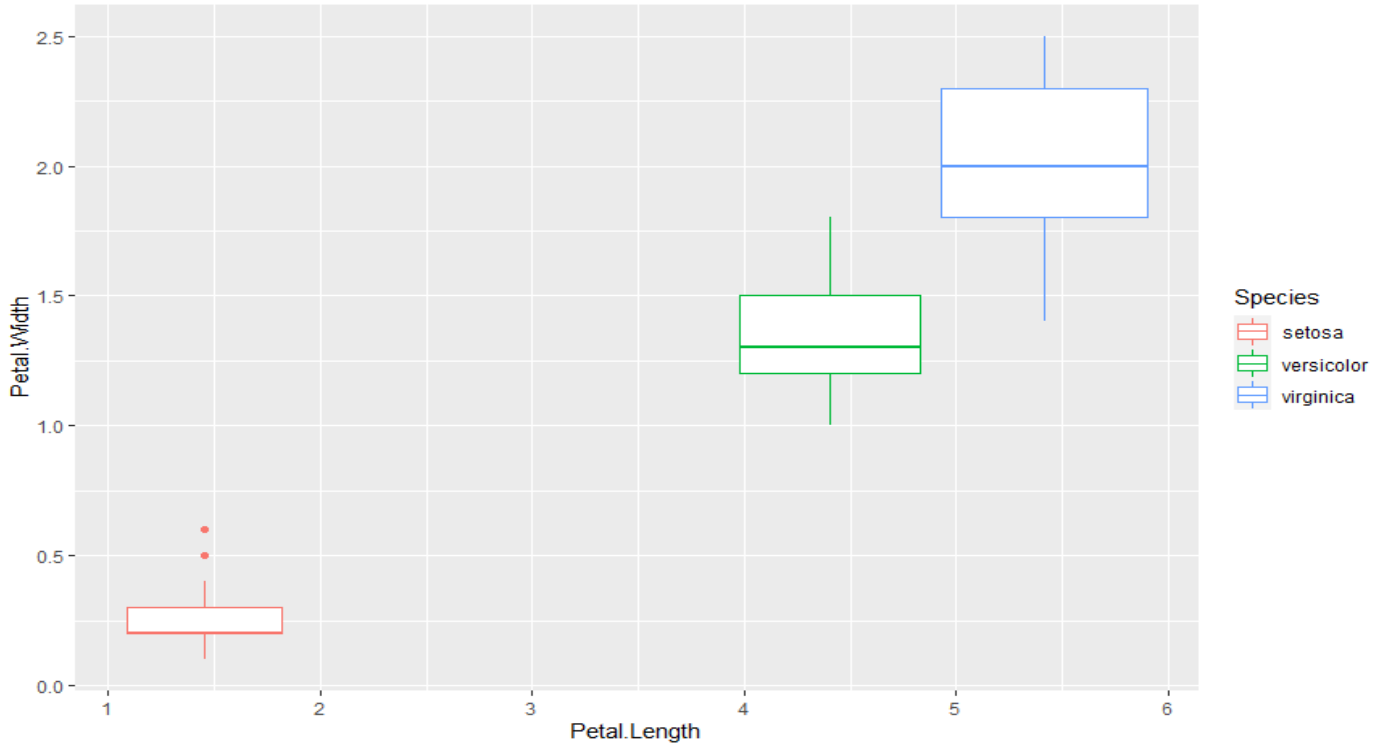
ggplot(data = iris, aes(x = Species, y = Petal.Width, color = Species)) +

geom_boxplot() + ggtitle("PWidth")



Which flower type exhibits a significantly different Petal Length/Width once it is separated from the other classes.

ggplot(data = iris, aes(x = Petal.Length, y = Petal.Width, color = Species)) + geom_boxplot()



Ans- **Setosa flower** exhibits different **Petal Length/Width** once it is separated from other classes.

Problem 2

Load the trees sample data set into R using a data frame (it is a built-in dataset) and produce a 5-number summary of each feature.

data(trees)

head(trees)

summary(trees)

```
> data(trees)
> head(trees)
  Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
5  10.7     81   18.8
6  10.8     83   19.7
> 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
>
```

Girth <- trees$Girth

Height <- trees$Height

Volume <- trees$Volume

```
> Girth <- trees$Girth
> Height <- trees$Height
> Volume <- trees$Volume
>
```

fivenum(Girth)

fivenum(Height)

fivenum(Volume)

```
> fivenum(Girth)
[1]  8.30 11.05 12.90 15.25 20.60
>
> fivenum(Height)
[1] 63 72 76 80 87
>
> fivenum(Volume)
[1] 10.2 19.4 24.2 37.3 77.0
>
```

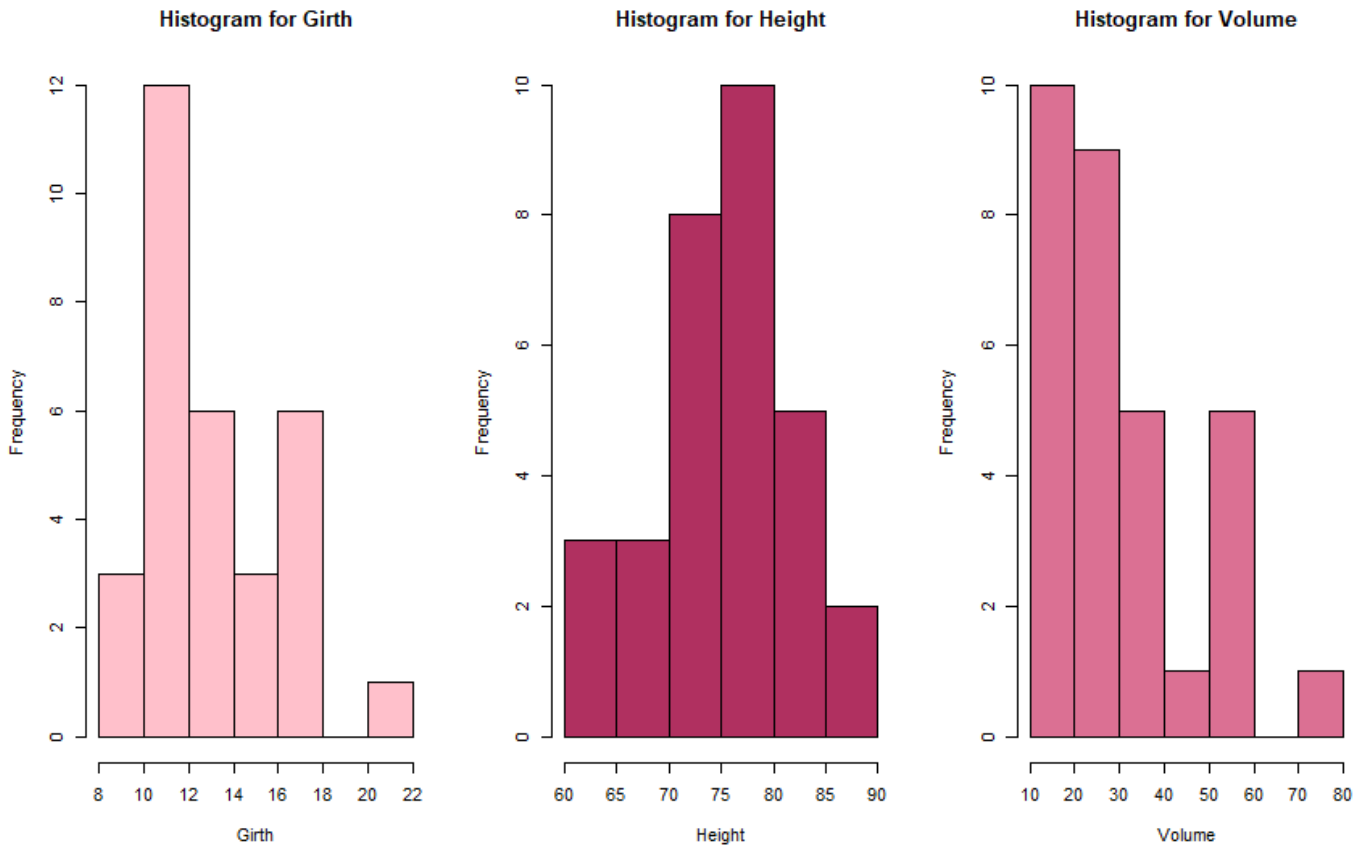
Create a histogram of each variable - which variables appear to be normally distributed based on visual inspection?

par(mfrow=c(1,3)) **#arranging all the 3 plots in 1 row and 3 different columns**

hist(Girth, xlab ="Girth", main="Histogram for Girth", col="pink")

hist(Height, xlab ="Height", main="Histogram for Height", col="maroon")

hist(Volume, xlab ="Volume", main="Histogram for Volume", col= "pale violet red")



Do any variables exhibit positive or negative skewness? Install the moments library from CRAN use the skewness function to calculate the skewness of each variable. Do the values agree with the visual inspection?

install.packages("moments") #installing pakages

```
> install.packages("moments")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate v
ersion of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/aasth/AppData/Local/R/win-library/4.2'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/moments_0.14.1.zip'
Content type 'application/zip' length 56264 bytes (54 KB)
downloaded 54 KB

package 'moments' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\aasth\AppData\Local\Temp\RtmpCwynOW\downloaded_packages
```

library(moments)

skewness(Girth)

skewness(Height)

skewness(Volume)

```
> library(moments)
> skewness(Girth)
[1] 0.5263163
> skewness(Height)
[1] -0.374869
> skewness(Volume)
[1] 1.064357
>
```

Yes, **Height** exhibits **negative skewness**, and **Volume and Girth** represent **positive skewness**. The value of Height is -0.374869 which is quite close to 0 and is in line with the visual inspection. On visually inspecting **Height** shows **Normal Distribution.**

Problem 3

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

df <- 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"))

summary(df)

str(df)

```
R 4.2.1 · ~/
> url="https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data"
>
> df <- 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"))
>
> summary(df)
      mpg          cylinders      displacement    horsepower           weight      acceleration     model.year
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Length:398         Min.   :1613   Min.   : 8.00   Min.   :70.00
 1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.2   Class :character   1st Qu.:2224   1st Qu.:13.82   1st Qu.:73.00
 Median :23.00   Median :4.000   Median :148.5   Mode  :character   Median :2804   Median :15.50   Median :76.00
 Mean   :23.51   Mean   :5.455   Mean   :193.4                      Mean   :2970   Mean   :15.57   Mean   :76.01
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0                      3rd Qu.:3608   3rd Qu.:17.18   3rd Qu.:79.00
 Max.   :46.60   Max.   :8.000   Max.   :455.0                      Max.   :5140   Max.   :24.80   Max.   :82.00
     origin        car.name
 Min.   :1.000   Length:398
 1st Qu.:1.000   Class :character
 Median :1.000   Mode  :character
 Mean   :1.573
 3rd Qu.:2.000
 Max.   :3.000
```

```
> str(df)
'data.frame':   398 obs. of  9 variables:
 $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower  : chr  "130.0" "165.0" "150.0" "150.0" ...
 $ weight      : num  3504 3693 3436 3433 3449 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ model.year  : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ car.name    : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel sst" ...
```

HorsePower <-df$horsepower

HorsePower = as.numeric(HorsePower) **#converting factors to numeric types**

```
> HorsePower <-df$horsepower
> HorsePower = as.numeric(HorsePower)
Warning message:
NAs introduced by coercion
```

mean(HorsePower, na.rm = TRUE) **#original Mean**

median(HorsePower, na.rm = TRUE)

HorsePower[is.na(HorsePower)] <-median(HorsePower, na.rm = TRUE) **# replacing NA's with median**

mean(HorsePower, na.rm = TRUE)

```
> mean(HorsePower, na.rm = TRUE)
[1] 104.4694
> median(HorsePower, na.rm = TRUE)
[1] 93.5
> HorsePower[is.na(HorsePower)] <-median(HorsePower, na.rm = TRUE)
> mean(HorsePower, na.rm = TRUE)
[1] 104.304
>
```

How does this affect the value obtained for the mean vs the original mean when the records were ignored?

Ans- We can see above that after replacing NA's with median value, the **mean** has decreased a little bit from 104.464 to 104.304 respectively. The original mean of **104.464** calculated for horsepower ignored 7 values in the NA column. So, when a median of 93.5 had to fill up the 7 values, the **median** was slightly changed **from 104.464 to 104.304.**

Problem 4

library(MASS)

library(ggplot2)

data(Boston)

attach(Boston) **#database is searched by R when checking/evaluating variables using attach**

bostondata= data.frame(Boston)

names(bostondata)

head(bostondata)

```
[1] 93.5
> library(MASS)
> library(ggplot2)
> data(Boston)
> attach(Boston)     #database is searched by R when checking/evaluating variables using attach
> bostondata= data.frame(Boston)
>
> names(bostondata)
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"     "rad"     "tax"     "ptratio"
[12] "black"   "lstat"   "medv"
> head(bostondata)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

Use lm to fit a regression between medv and lstat - plot the resulting fit and show a plot of fitted values vs. residuals. Is there a possible non-linear relationship between the predictor and response?

Linearmodel1 = lm(medv~lstat, data = bostondata)

Linearmodel1

summary(Linearmodel1)

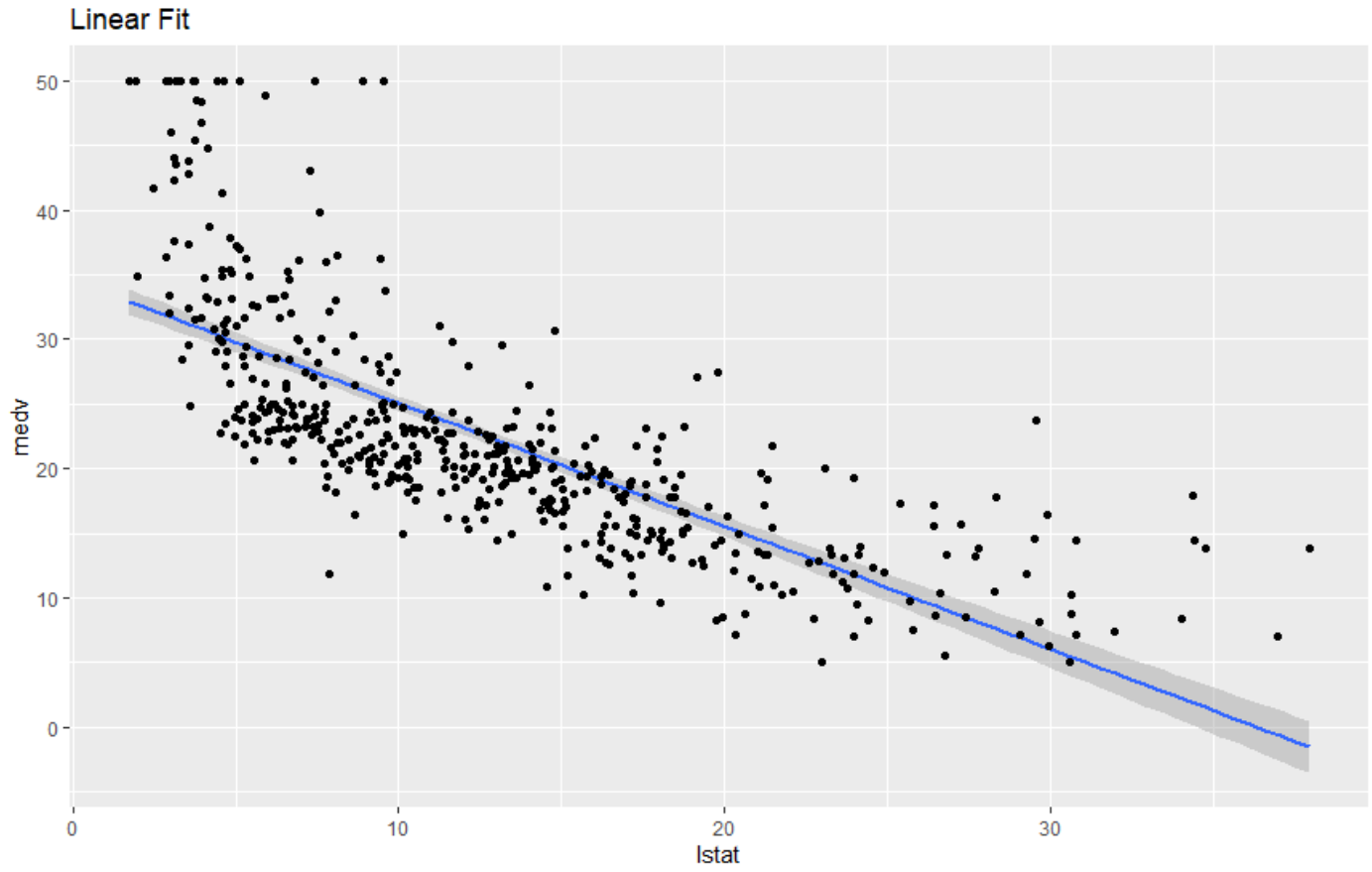
R^2 for Linear fit(Linearmodel1) =0.5441

```
> Linearmodel1 = lm(medv~lstat, data = bostondata)
> Linearmodel1 = lm(medv~lstat, data = bostondata)
> Linearmodel1

Call:
lm(formula = medv ~ lstat, data = bostondata)

Coefficients:
(Intercept)        lstat
      34.55        -0.95
```

```
R 4.2.1 · ~/

> summary(Linearmodel1)

Call:
lm(formula = medv ~ lstat, data = bostondata)

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
```

ggplot(bostondata, aes(lstat, medv)) + stat_smooth(method = lm, se = TRUE) + geom_point() +

ggtitle("Linear Fit")

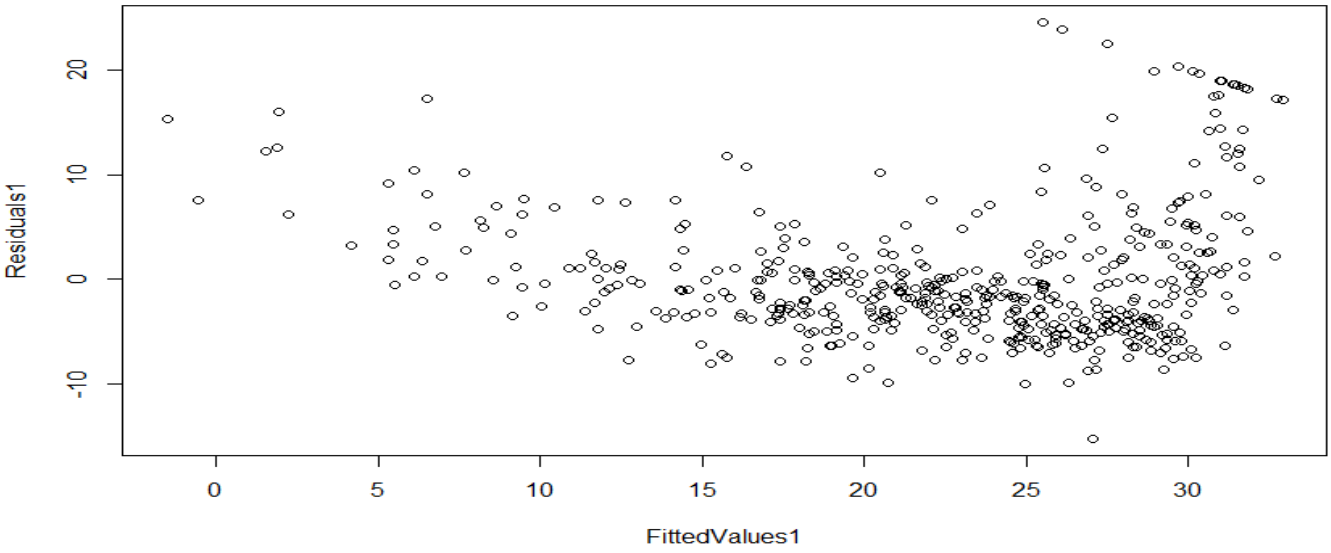


Ans- There appears to be some non-linear relationship between predictor and response.

FittedValues1 <-Linearmodel1$fitted.values

Residuals1 <-Linearmodel1$residuals

plot(FittedValues1, Residuals1)



Use the predict function to calculate values response values for lstat of 5, 10, and 15 – obtain confidence intervals as well as prediction intervals for the results - are they the same? Why or why not?

predict(Linearmodel1, data.frame(lstat=(c(5,10,15))), interval = "confidence")

```
> predict(Linearmodel1, 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(Linearmodel1, data.frame(lstat=(c(5,10,15))), interval = "prediction")

```
> predict(Linearmodel1, 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
> |
```

Ans- As seen the prediction and the confidence intervals are not the same. The prediction interval results are a little wider than confidence intervals. We can see that for a value of 25.05 the prediction intervals are wider than confidence intervals.

Modify the regression to include lstat2 (as well lstat itself) and compare the R2 between the linear and non-linear fit - use ggplot2 and stat smooth to plot the relationship.

Linearmodel2 = lm(medv~lstat +I(lstat^2), data = bostondata)

Linearmodel2

```
> Linearmodel2 = lm(medv~lstat +I(lstat^2), data = bostondata)
> Linearmodel2

Call:
lm(formula = medv ~ lstat + I(lstat^2), data = bostondata)

Coefficients:
(Intercept)        lstat   I(lstat^2)
   42.86201     -2.33282      0.04355

> |
```

summary(Linearmodel2)

```
> summary(Linearmodel2)

Call:
lm(formula = medv ~ lstat + I(lstat^2), data = bostondata)

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

> |
```

Rsquare1 = summary(Linearmodel1)$r.sq

Rsquare1

```
> Rsquare1 = summary(Linearmodel1)$r.sq
> Rsquare1
[1] 0.5441463
> |
```

Rsquare2 = summary(Linearmodel2)$r.sq

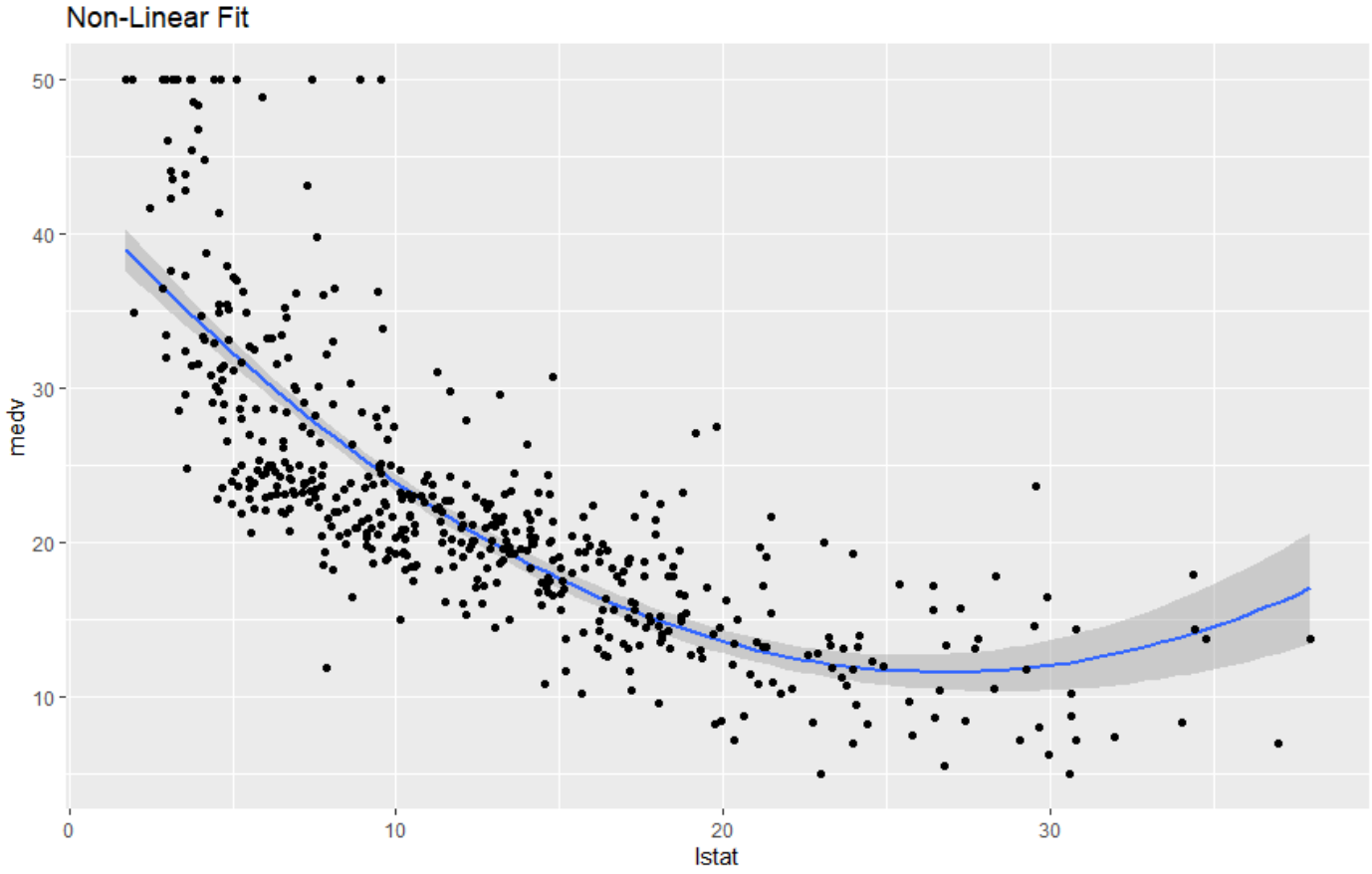
Rsquare2

```
> Rsquare2 = summary(Linearmodel2)$r.sq
> Rsquare2
[1] 0.6407169
> |
```

For Non-Linear Fit(Linearmodel2) the value of R^2 is 0.6407

ggplot(bostondata,aes(x= lstat,y = medv)) + stat_smooth(method = "lm",

formula = y ~ x + I(x^2), se = TRUE) + geom_point() + ggtitle("Non-Linear Fit")



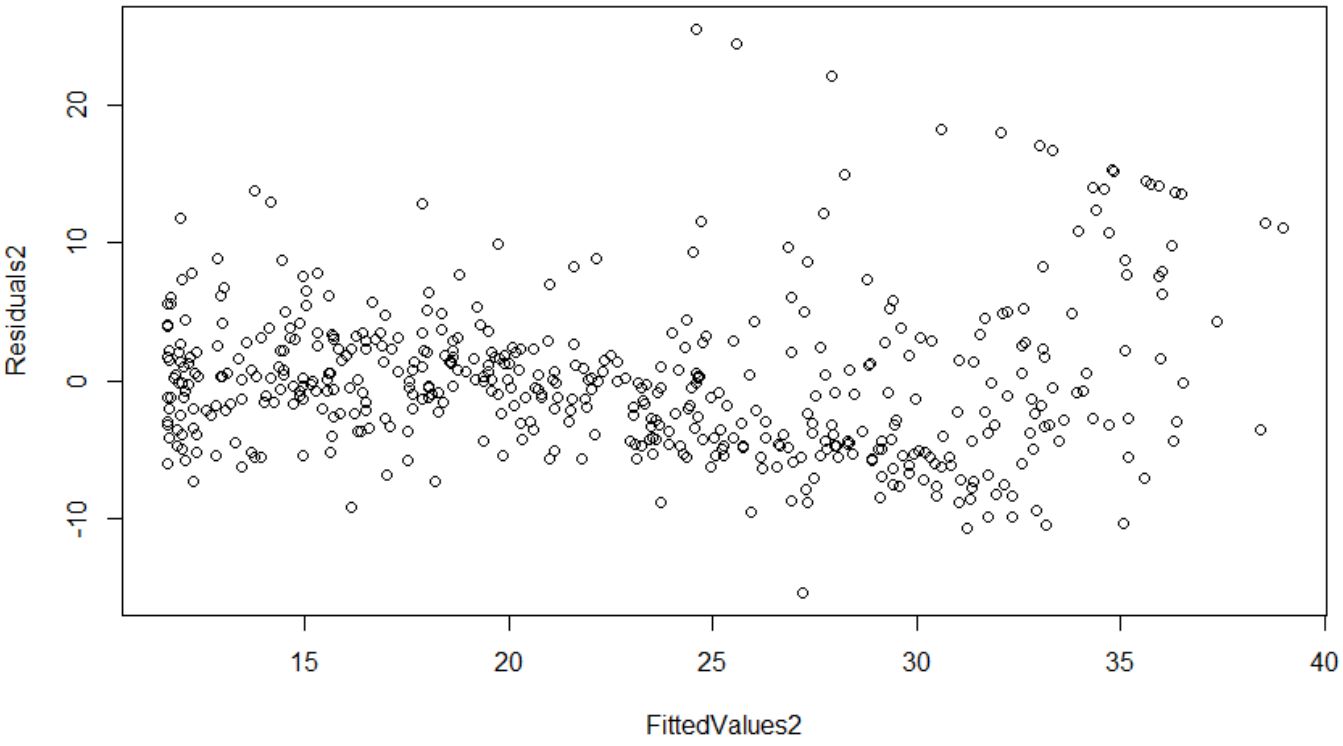
ggplot(bostondata, aes(x =FittedValues1, Residuals1)) + stat_smooth(se = TRUE) +

geom_point() + labs(x = "FittedValues", y = "Residuals") + ggtitle("Fitted Values VS Residuals for Linear Model")

FittedValues2 <-Linearmodel2$fitted.values

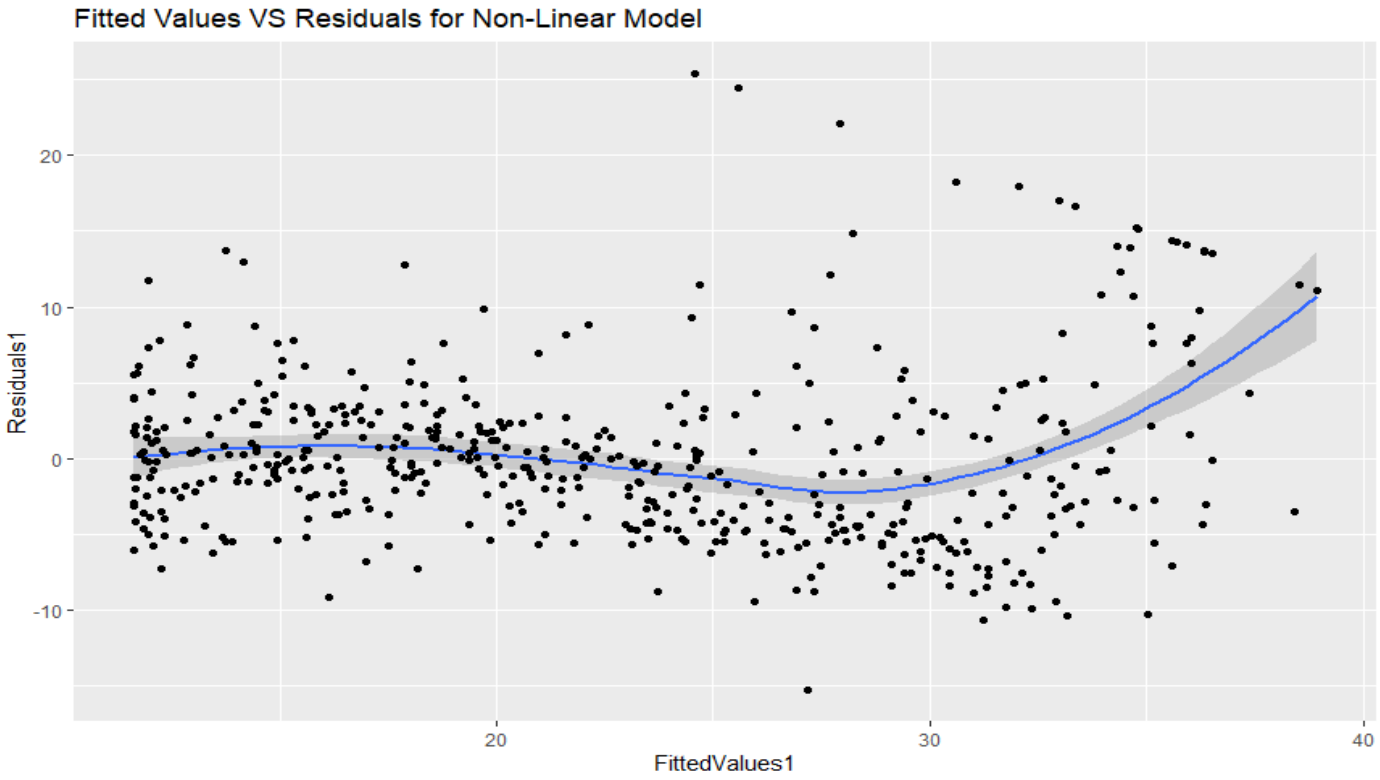
Residuals2 <-Linearmodel2$residuals

plot(FittedValues2, Residuals2)



ggplot(bostondata, aes(x =FittedValues2, Residuals2)) + stat_smooth(se = TRUE) +

geom_point() + labs(x = "FittedValues1", y = "Residuals1") + ggtitle("Fitted Values VS Residuals for Non-Linear Model")



Submitted By: -

Aastha Dhir

CWID-A20468022

adhir2@hawk.iit.edu