PARTH RATHOD

A20458817

CSP - 571 DPA

FALL 2021

Recitation Exercises

Chapter 1

1) For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

Answer:-

a) Since the sample size (n) is high, the performance of the model will be high and have low variance and hence the flexible method will perform better than inflexible method.

b) Since the sample size (n) is low, the performance of the model will be low and will result in overfitting, hence the flexible method will perform worse than inflexible method.

c) Since the flexible method are good at developing models which have a non-linear relationship. Hence flexible methods will perform better than inflexible method. Also inflexible methods may develop models which result in underfitting due to non linear relationship.

d) Since high variance of error terms means a lot of noise is present in dataset which may result in overfitting of model in case of flexible method. Hence flexible method will perform worse than inflexible method.

2) Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

Answer:-

a) Since CEO salary will be continuous variable hence this is a regression problem. Also, this is an inference problem because here we are interested in finding how independent variables(profit, number of employees, industry) impact the salary of CEO. Here n=500 and p=3

b) Since we want to know whether the product will be success or failure so this is a classification problem. Also, this is a prediction problem because we are mainly interested in the success or failure of the company and not on how vastly different parameters impact the firm. Here n=20 and p=13

c) Since we want to know % change which will be a continuous value and hence this is a regression problem. Also, this is a prediction problem because here we are mainly interested in the % change in dollar not on the % impact of other parameters on % dollar change. Here n=52 and p=3

4) You will now think of some real-life applications for statistical learning.

Answer:-

a) i)Classification model will be useful to classify whether a person is eligible to take loan or not, based on different parameters like marital status,monthly expenses, gender, age, salary.

```
Response: eligible/not eligible
Predictors: Gender, Age, Salary, Marital status, Monthly expenses
Goal: Prediction because here we are interested in finding whether a
person is eligible for loan or not.

ii)Classification model will be useful to classify whether a person
should buy a specific product, based on parameters
   like a requirement, quality, price, offers

Response: Buy/Dont Buy
Predictors: requirement, quality, price, offers
Goal: Prediction because here we are interested in finding whether he
should buy the product or not.

iii)Classification model will be useful to classify whether a person
should purchase a vehicle to travel to his office,       based on
following parameters like distance to the office, travel time, regularly
spent on public service transportation,
    salary.

Response: Buy/Dont Buy
Predictors: distance to the office, travel time, regularly spent on
public service transportation,salary
Goal: Prediction because here we are interested in finding whether a
person should purchase a vehicle or not.
```

b) i)% of precipitation

```
Response: Percentage of rainfall (Precipitation).
Predictors: Temperature, Air Pressure, Moisture, Humidity.
Goal: Inference.

ii)Predict insurance premium

Response: PInsurance premium amount.
Predictors: age, deductibles applied, accident history, salary.
Goal: Prediction because here we are interested in finding insurance
premium for different people.

iii)Per capita income of an economy

Response: Per capita income of an economy.
Predictors: GDP, Population, Literacy Rate, Tax Revenue,Inflation rate.
Goal: Inference.
```

c) i) People can be clustered into different groups based on their liking about different genres of movies (like action, horror, comedy, etc.) and accordingly new upcoming movies can be promoted by recommending them in a group.

```
ii) Cluster analysis can be used to identify customer buying patterns.
Frequent items bought together can be identified.

iii)Recognizing communities within large groups of people. Predictors:
Country, 1st Language, 2nd Language.
```

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 nonparametric approach)? What are its disadvantages?

Answer:-

```
Parametric Learning:- i) Reduces the problem of estimating f down to one
of estimating set of parameters because it assumes a form for f.
                      ii) yi=β0+β1xi+ei
                      iii) In parametric model, the model to fit the data
is known in prior.

NonParametric Learning:-i) Does not assume a particular form of f, so
requires a very large sample to accurately estimate f.
                        ii) yi=f(xi)+ei
                        iii) In non-parametric model, the data tells,
what the 'regression' should look like.
```

Advantage: The main advantage of a parametric model to either a classifier or a regressor is the simplification of model f to a few parameters as not many observations are required as compared to the non-parametric model.

Disadvantage: The disadvantages of a parametric model to either a classifier or a regressor are potentially inaccurate estimate f, if the form of f assumed is wrong or to overfit the observations if more flexible models are used.

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

Answer:-

a) $Euclidian\ Distance$ = √($x2 - x1$)^2 + ($y2 - y1$)^2 + ($z2 - z1$)^2

```
Obs.  Euclidean_distance
1           3
```

```
2           2
3           3.16
4           2.24
5           1.41
6           1.73
```

b) For K = 1, we choose single nearest point, which in this case is the 5th observation (Distance 1.41).Hence class associated with that observation is Green. Hence for K = 1 our prediction is Green.

c) For K = 3, we choose three nearest points, which in this case are Obs. 5 (Distance 1.41), Obs.6 (Distance 1.73) and Obs. 2 (Distance 2). Majority of these observation is class Red. Hence class associated with that observation is Red. Hence for K = 3 our prediction is Red.

d)The Bayes boundary will become almost linear for higher values of K. However, in this case Bayes boundary is highly noncolinear which suggest that the value for K should be small.

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.

Answer:-

```
i) The null hypothesis for "TV”: In the presence of radio ads and
newspaper ads, TV ads have no
   effect on sales.
ii) The null hypothesis for "radio": In the presence of TV and newspaper
ads, radio ads have no
    effect on sales.
iii)The null hypothesis for "newspaper": In the presence of TV and radio
ads, newspaper ads have
    no effect on sales.
iv) However, because of the low p-values for TV and radio, the null
hypotheses are rejected.
v) On the other hand, the high p-value of newspaper suggests that the
null hypothesis holds true
   for 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 β0 = 50, β1 = 20, β2 = 0.07, β3 = 35, β4 = 0.01, β5 = −10.

Answer:-

a) Y = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5 Y = 50 + 20(GPA) + 0.07(IQ) + 35($Gender$) + 0.01($GPA \times IQ$) − 10($GPA \times Gender$) Also, we know in Gender, Male = 0 and Female = 1,

Males: Y = 50 + 20(GPA) + 0.07(IQ) + 35(0) + 0.01(GPA × IQ) − 10(GPA × 0)

Y= 50+20(GPA)+0.07(IQ)+0.01(GPA × IQ)

Females: Y = 50 + 20(GPA) + 0.07(IQ) + 35(1) + 0.01(GPA × IQ) − 10(GPA × 1)

Y= 85+10(GPA)+0.07(IQ)+0.01(GPA × IQ)

On solving above two equations we get GPA = 3.5,

Therefore, Males earn more than Females if and only if GPA is greater than 3.5 (High GPA)

Therefore, iii is correct.

b) From the equation of females derived from the above question we have, Females: Y= 85+10(GPA)+0.07(IQ)+0.01(GPA × IQ) Therefore, Y = 85+10(4.0) + 0.07(110) + 0.01(4.0 × 110) Salary of female = 137.1 = $137100/-

c) False. To examine if the GPA/IQ has an impact on the quality of the model we need to test the hypothesis H0: β ₪ = 0 and look at the p-value associated with the t-statistic to draw a conclusion.

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.

Answer:-

a) Without knowing more details about the training data, it is difficult to know which training RSS is lower between linear or cubic. However, as the true relationship between X and Y is linear, it can be assumed that the least squares line to be close to the true regression line, and consequently the RSS for the linear regression may be lower than for the cubic regression.

b) The test RSS depends on test data, thus to draw any conclusion, enough information isn't available. However, it can be assumed that polynomial regression will have a higher test RSS as the overfit from training would have more error than the linear regression.

c)Due to the high flexibility, Polynomial regression has lower train RSS as compared to the linear fit. So, the more flexible model will closer follow points and reduce train RSS.

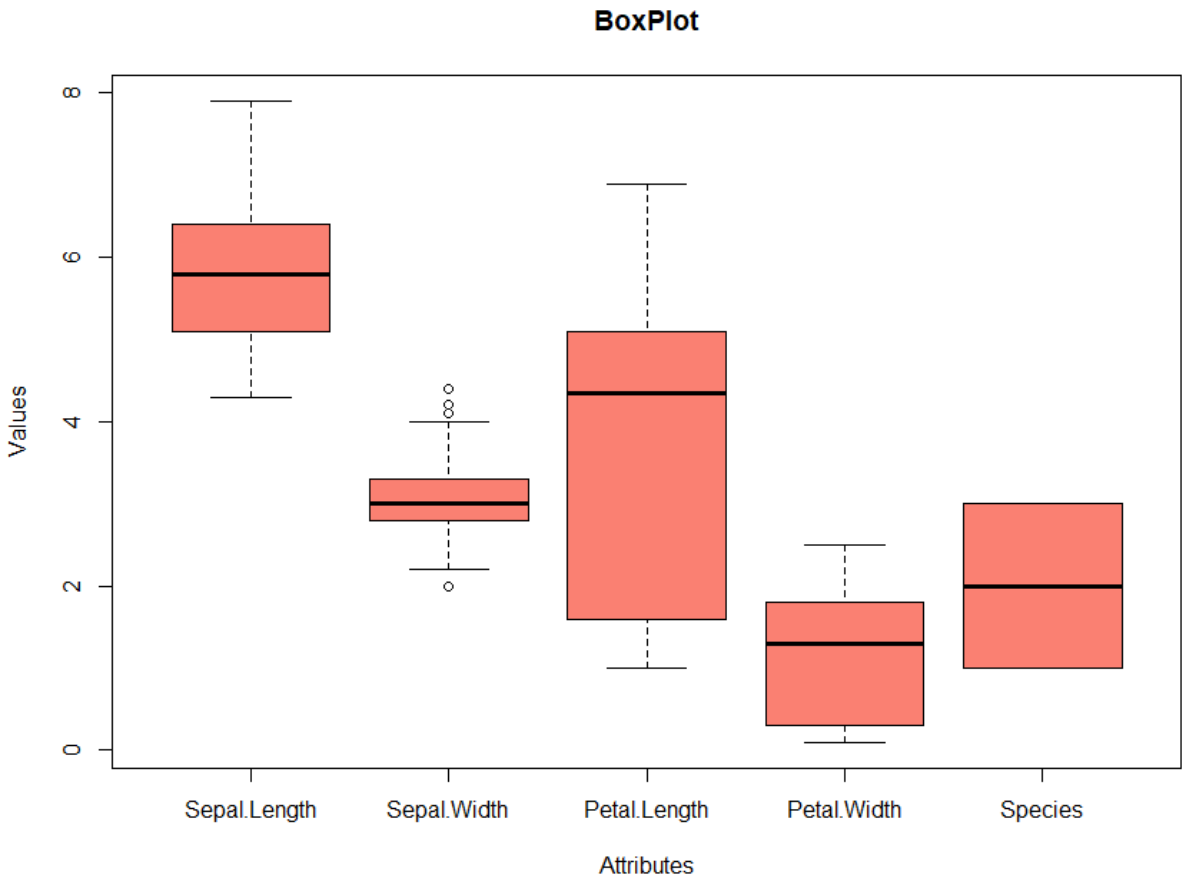
d)The information available is not sufficient enough to tell which test RSS would be lower for either regression as it is not clear that what level of flexibility will fit data better. If it is closer to linear,the linear regression test RSS would be lower and if it closer to cubic then the cubic regression test RSS could be lower.

Coding Problems

Problem 1

In [ ]:
```
#Load Libraries
```

```
library(gridExtra)
library(ggplot2)

#Load Iris Dataset
iris <-data.frame(iris)

#BoxPlot
boxplot(iris,main="BoxPlot",xlab="Attributes",ylab="Values", col = "salmon")
```



In [ ]:

```
> #Inter Quartile Range Calculation:
> print("IQR of SepalLength")
 "IQR of SepalLength"
> IQR(iris$Sepal.Length)
 1.3
> print("IQR of SepalWidth")
 "IQR of SepalWidth"
> IQR(iris$Sepal.Width)
 0.5
> print("IQR of PetalLength")
 "IQR of PetalLength"
> IQR(iris$Petal.Length)
 3.5
> print("IQR of PetalWidth")
 "IQR of PetalWidth"
> IQR(iris$Petal.Width)
 1.5
```

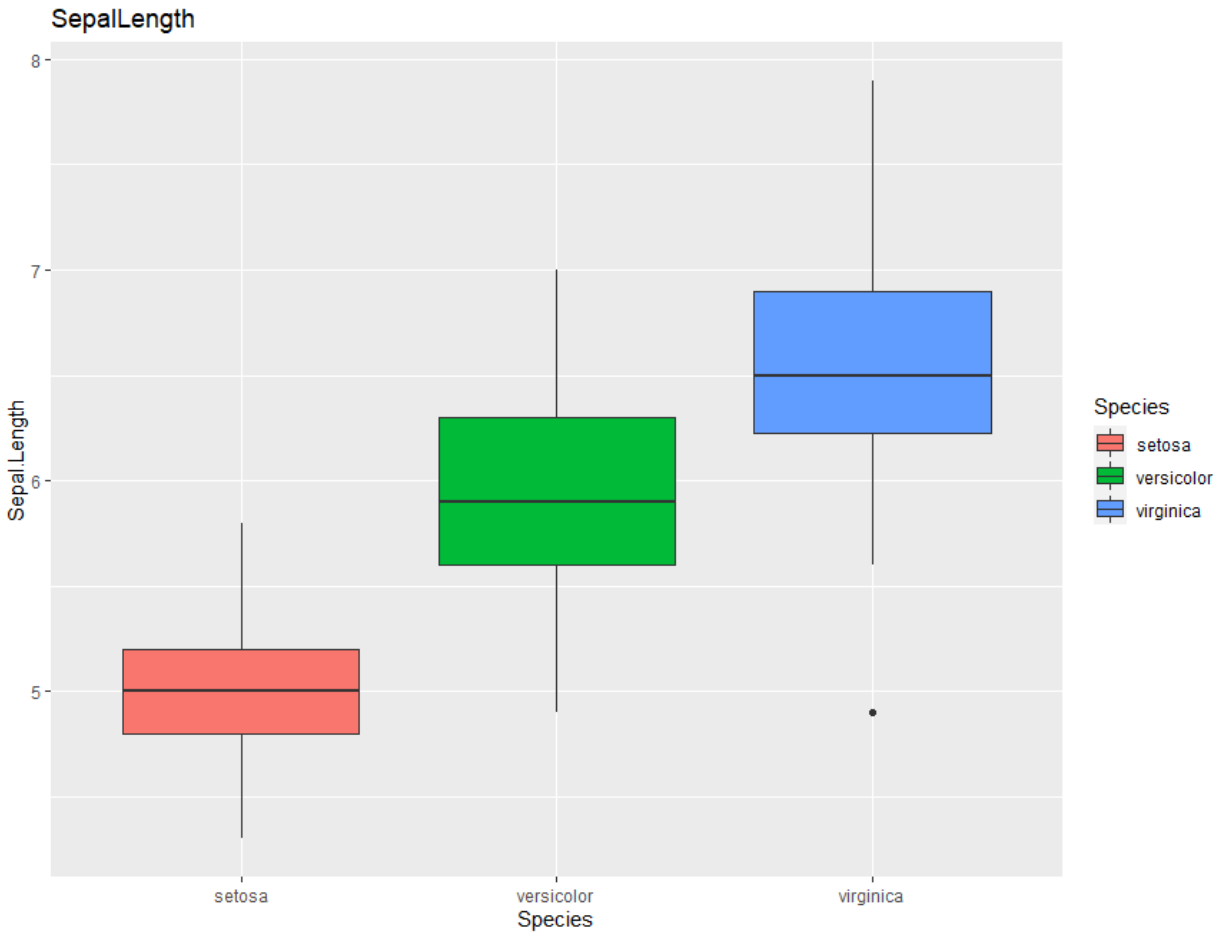
Petal Length has maximum length boxplot. So Petal Length will have maximum Inter Quartile Range.

In [ ]:
```
> #SD
> print("Standard Deviation of SepalLength")
 "Standard Deviation of SepalLength"
> sd(iris$Sepal.Length)
 0.8280661
> print("Standard Deviation of SepalWidth")
 "Standard Deviation of SepalWidth"
> sd(iris$Sepal.Width)
 0.4358663
> print("Standard Deviation of PetalLength")
 "Standard Deviation of PetalLength"
> sd(iris$Petal.Length)
 1.765298
> print("Standard Deviation of PetalWidth")
 "Standard Deviation of PetalWidth"
> sd(iris$Petal.Width)
 0.7622377
```

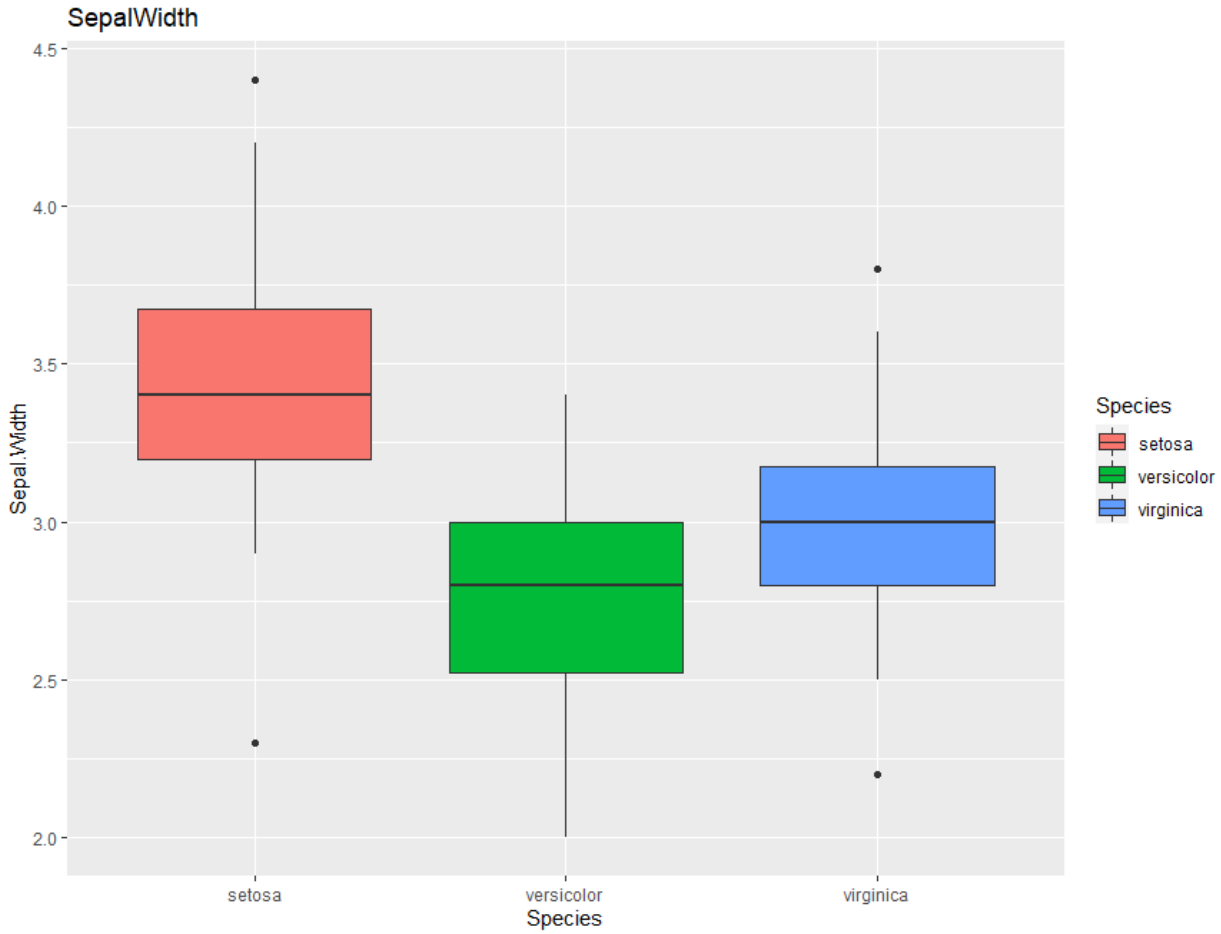
Feature with largest Standard Deviation: Petallength

Box Plot of attributes with class species

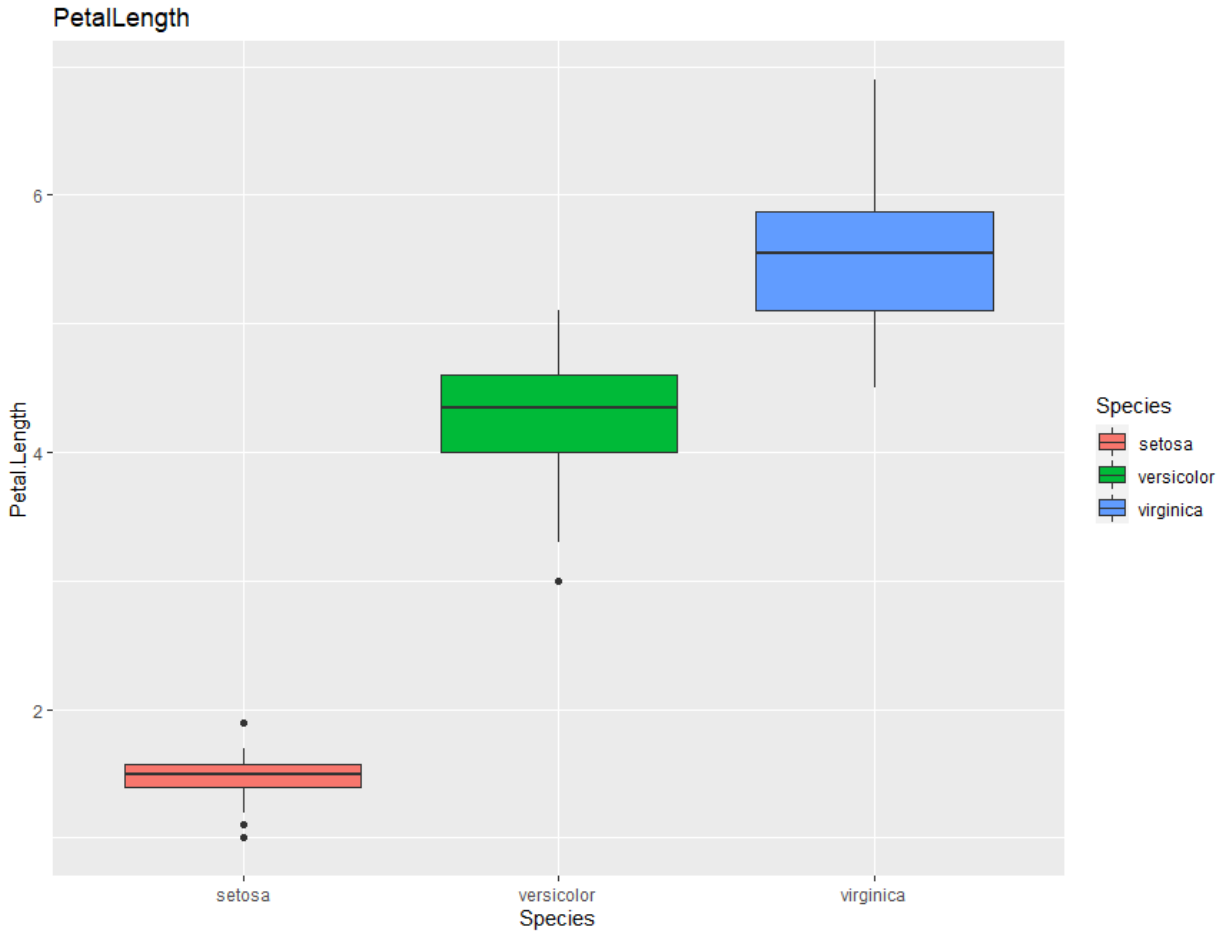
In [ ]:
```
ggplot(data = iris, aes(x = Species, y = Sepal.Length, fill = Species)) + geom_boxplot(
```



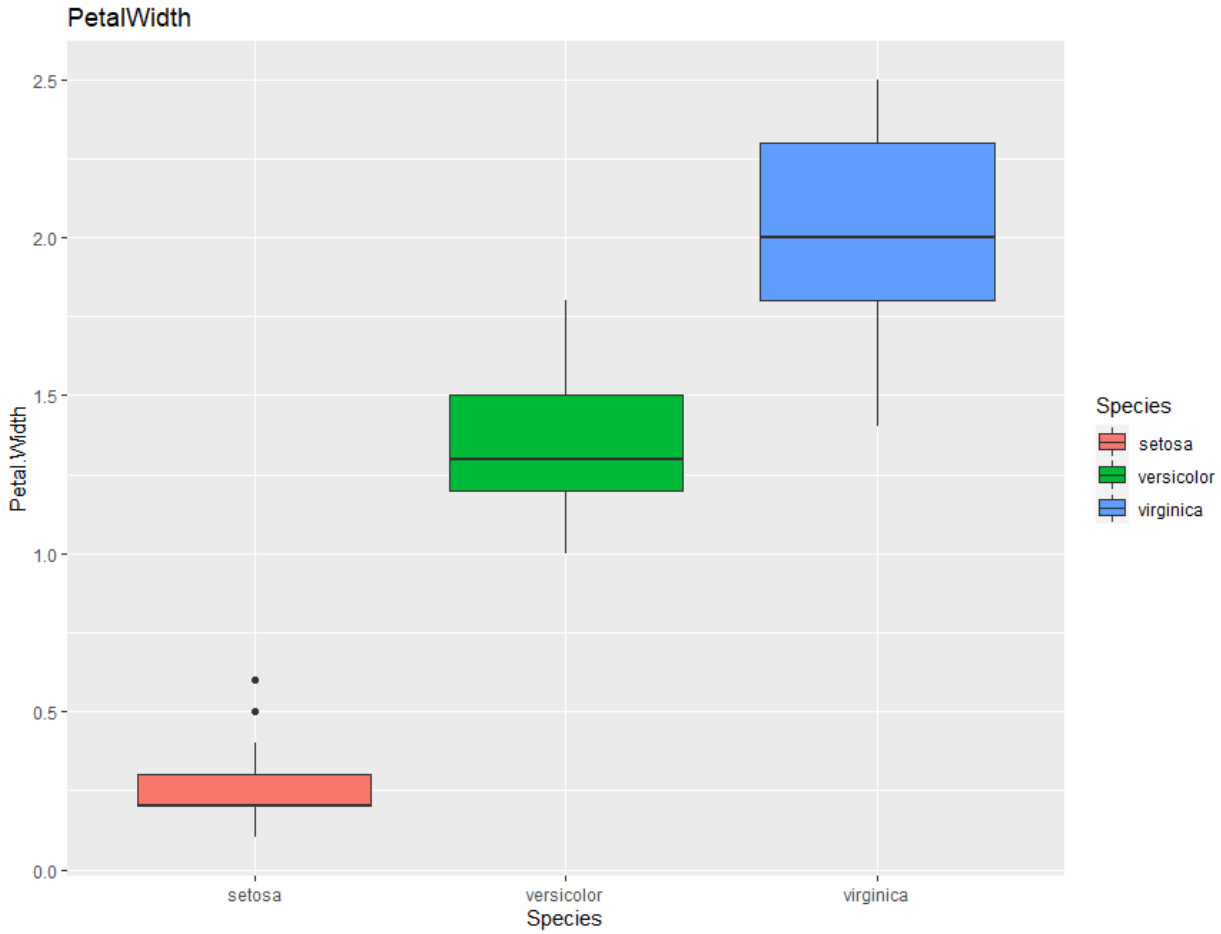
In [ ]:
```
ggplot(data = iris, aes(x = Species, y = Sepal.Width, fill = Species)) + geom_boxplot()
```



In [ ]:
```
ggplot(data = iris, aes(x = Species, y = Petal.Length, fill = Species)) + geom_boxplot(
```



In [ ]:
```
ggplot(data = iris, aes(x = Species, y = Petal.Width, fill = Species)) + geom_boxplot()
```



From aboce BoxPlot, we can conclude that Setosa Species has comparatively different values of petal length and width.

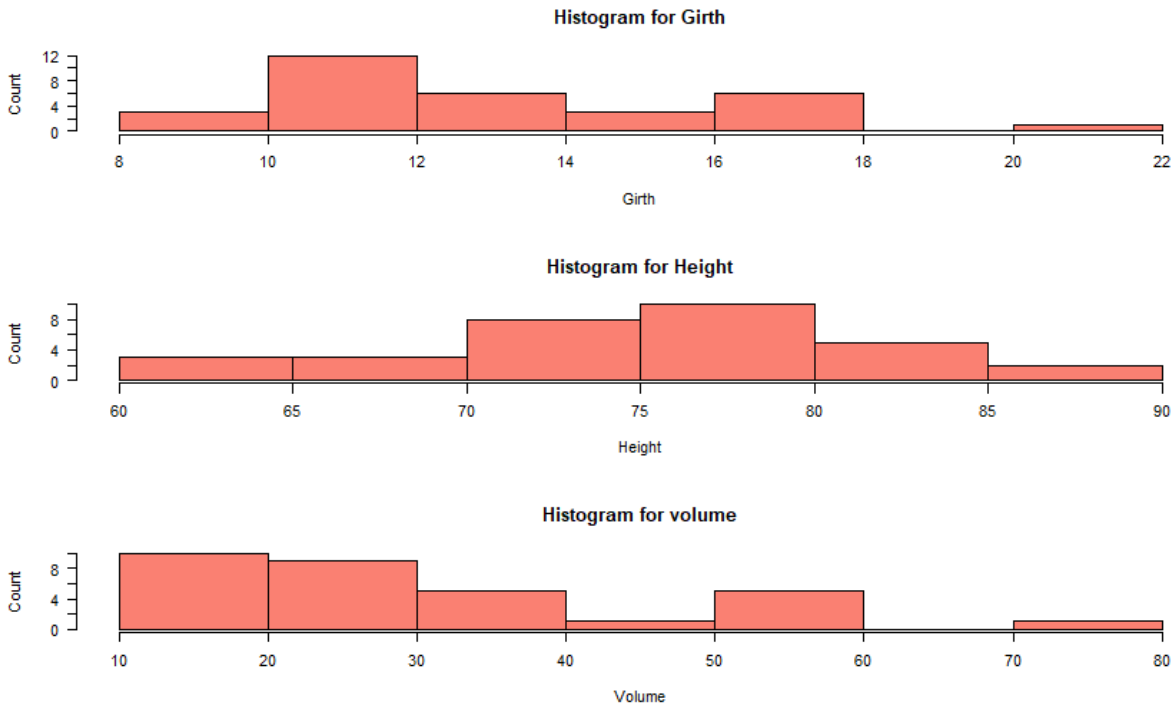
Problem 2

In [ ]:
```
#Load Library
library(moments)

#Load Tree Sample Dataset
trees <-data.frame(trees)

#summary of Attributes
summary(trees)
```

In [ ]:
```
     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  
```

In [ ]:
```
#Histogram of Attributes
par(mfrow=c(3,1))
hist(trees$Girth,main="Histogram for Girth", xlab = "Girth", ylab = "Count",border="bla
```

```
hist(trees$Height,main="Histogram for Height", xlab = "Height", ylab = "Count",border="
hist(trees$Volume,main="Histogram for volume", xlab = "Volume", ylab = "Count",border="
```



From above graph, we can conclude:-

Height is somewhat close to Normal Distributed and has Negative Skewness.

Girth has Positve Skewness.

Volume has Positive Skewness.

In [ ]:
```
print("Skewness of Girth:")
skewness(trees$Girth)
print("Skewness of Height:")
skewness(trees$Height)
print("Skewness of Volume:")
skewness(trees$Volume)
```

In [ ]:
```
> print("Skewness of Girth:")
 "Skewness of Girth:"
> skewness(trees$Girth)
 0.5263163
> print("Skewness of Height:")
 "Skewness of Height:"
> skewness(trees$Height)
 -0.374869
> print("Skewness of Volume:")
 "Skewness of Volume:"
> skewness(trees$Volume)
 1.064357
```

We can see that visual Inspection and Calculation of Skewness leads to same.

The Skewness of height attribute is in range from -0.5 to +0.5 and hence it is close to Normal Distribution and has Negative Skewness value.

Skewness of Girth is in range from 0.5 to 1 hence it is somewhat Positively Skewed.

Skewness of Volume is greater than 1 and hence it is Positively Skewed.

Problem 3

In [ ]:
```
mpgData = read.table(url("https://archive.ics.uci.edu/ml/machine-learning-databases/aut
mpgHeader = c("mpg","cylinders","displacement","horsepower","weight","acceleration","mo
colnames(mpgData) = mpgHeader                          # setting up the headers
mpgData$horsepower = as.numeric(mpgData$horsepower)    # converting factors to numeric
print("Mean before replacment:")
mean(mpgData$horsepower, na.rm = T)
mpg_median = median(mpgData$horsepower, na.rm = T)
mpgData$horsepower[is.na(mpgData$horsepower)] = mpg_median
print("Mean after replacment:")
mean(mpgData$horsepower, na.rm = T)
```

In [ ]:
```
> print("Mean before replacment:")
[1] 104.4694
> print("Mean after replacment:")
[1] "Mean after replacment:"
 104.304
```

From above we can conclude that , On replacing all NA values of Horsepower attribute with Median value, mean of the attribute will slightly change from 104.469 to 104.304

Problem 4

In [ ]:
```
#Load Libraries
library(MASS)
library(ggplot2)
bostonData = data.frame(Boston)
attach(Boston)

#Display Summary
summary(bostonData)
```

```
     crim                zn              indus            chas
   nox               rm             age
```

Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000 Min. :0.3850 Min. :3.561 Min. : 2.90
1st Qu.: 0.08205 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000 1st Qu.:0.4490 1st Qu.:5.886 1st Qu.: 45.02
Median : 0.25651 Median : 0.00 Median : 9.69 Median :0.00000 Median :0.5380 Median :6.208 Median : 77.50
Mean : 3.61352 Mean : 11.36 Mean :11.14 Mean :0.06917 Mean :0.5547 Mean :6.285 Mean : 68.57
3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10 3rd Qu.:0.00000 3rd Qu.:0.6240 3rd Qu.:6.623 3rd Qu.: 94.08
Max. :88.97620 Max. :100.00 Max. :27.74 Max. :1.00000 Max. :0.8710 Max. :8.780 Max. :100.00
dis rad tax ptratio black lstat medv
Min. : 1.130 Min. : 1.000 Min. :187.0 Min. :12.60 Min. : 0.32 Min. : 1.73 Min. : 5.00

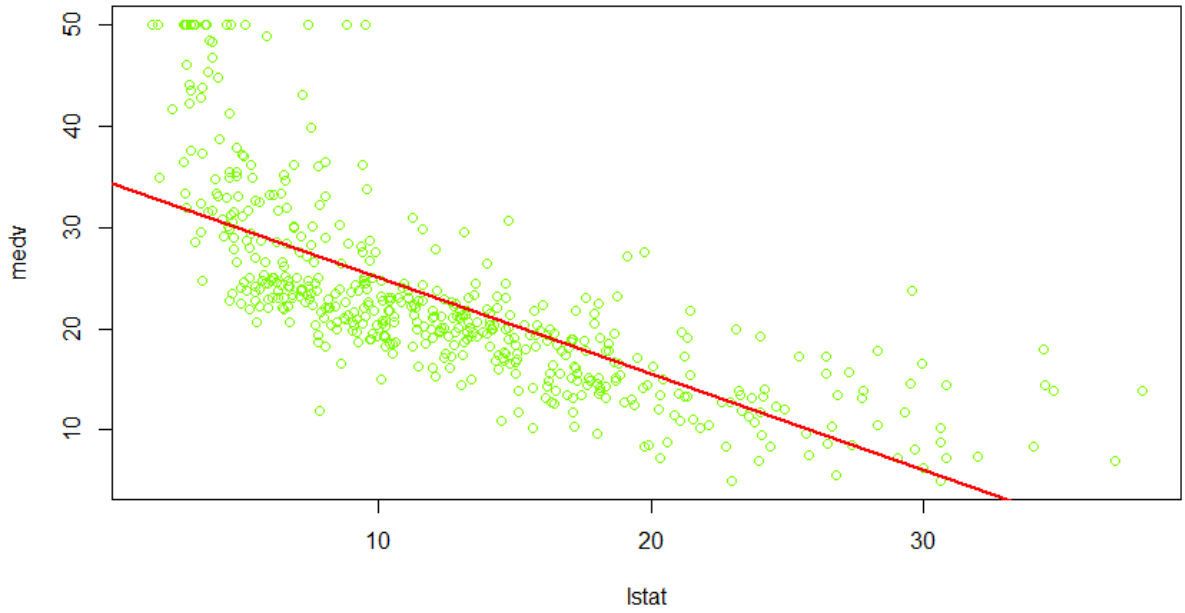
```
1st Qu.: 2.100 1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.:375.38 1st Qu.: 6.95 1st Qu.:17.02
Median : 3.207 Median : 5.000 Median :330.0 Median :19.05 Median :391.44 Median :11.36 Median
:21.20
Mean : 3.795 Mean : 9.549 Mean :408.2 Mean :18.46 Mean :356.67 Mean :12.65 Mean :22.53
3rd Qu.: 5.188 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:396.23 3rd Qu.:16.95 3rd Qu.:25.00
Max. :12.127 Max. :24.000 Max. :711.0 Max. :22.00 Max. :396.90 Max. :37.97 Max. :50.00
```

In [ ]:
```
#Fit a Regression Model
linearModel1 = lm(medv~lstat, data = bostonData)
summary(linearModel1)
coef(linearModel1)
cat("R-squared for Linear Model: ",summary(linearModel1)$r.sq)
```

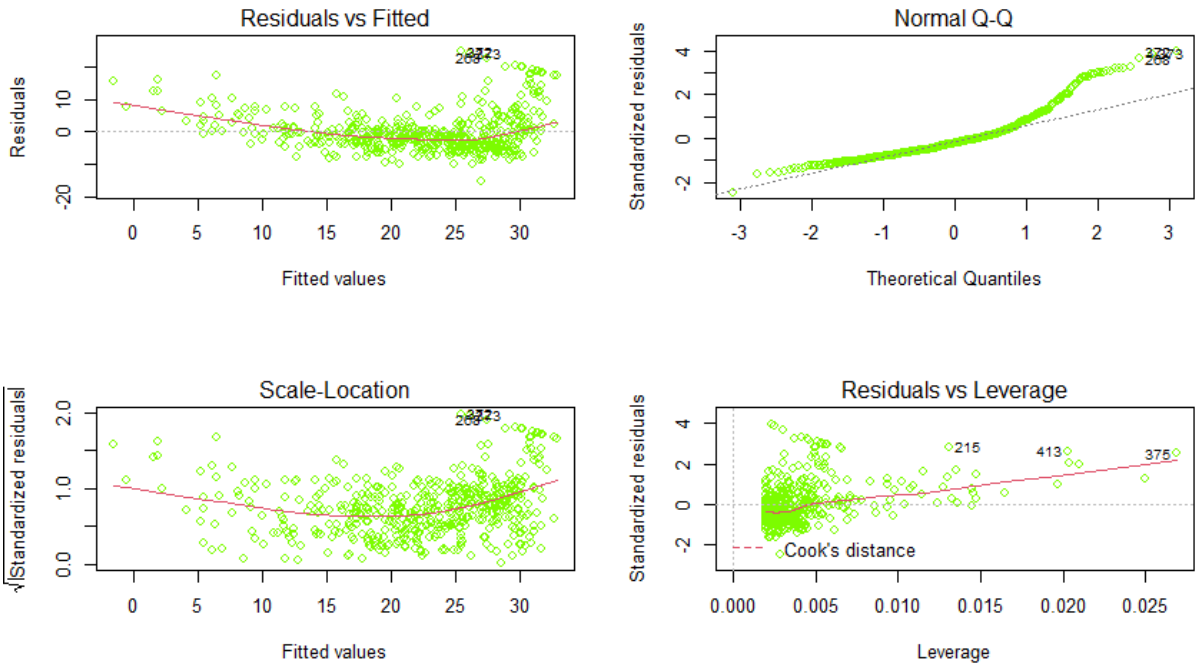
In [ ]:
```
Call:
lm(formula = medv ~ lstat, data = bostonData)

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

> coef(linearModel1)
(Intercept)       lstat
 34.5538409  -0.9500494
> cat("R-squared for Linear Model: ",summary(linearModel1)$r.sq)
R-squared for Linear Model:  0.5441463
```

For Linear Model we have R2 = 0.5441

In [ ]:
```
#Plot Resulting Fit
plot(lstat,medv,col="lawngreen")
abline(linearModel1, col="red",lwd=2)
```

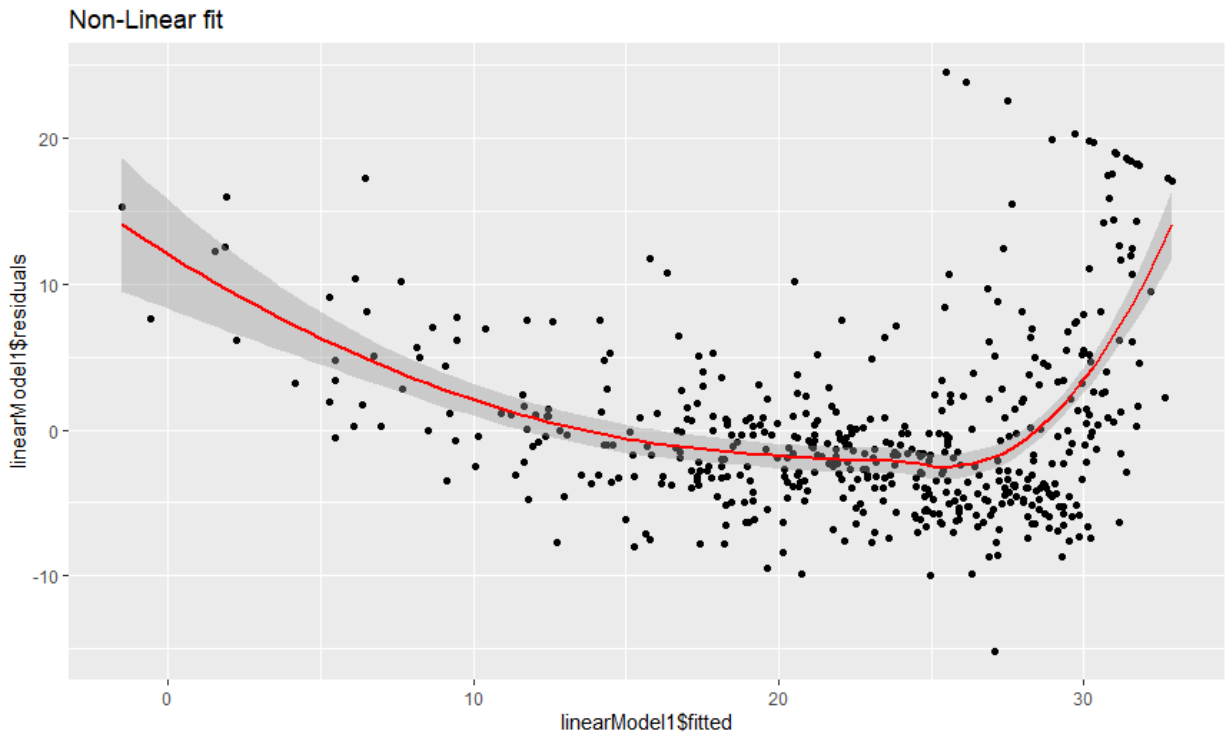


In [ ]:
```
par(mfrow=c(2,2))
plot(linearModel1, col="lawngreen")
```



In [ ]:
```
#Plot for Non-Linear Fit:
p1 <- ggplot(data=bostonData,aes(x=linearModel1$fitted,y=linearModel1$residuals)) +
  geom_point()+
  stat_smooth(col="red")+
  ggtitle("Non-Linear fit")
p1
coef(linearModel1)
```

In [ ]:
```
> coef(linearModel1)
(Intercept)       lstat
 34.5538409  -0.9500494
```



From the Graph we can see that there is Non-Linear Relationship between Predictor and Response variables.

In [ ]:
```
#predictions for lstat values of 5, 10 and 15
test=data.frame(lstat=c(5,10,15))
predict(linearModel1,test,interval = "confidence")
predict(linearModel1,test,interval = "predict")
```

In [ ]:
```
> predict(linearModel1,test,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,test,interval = "predict")
       fit       lwr      upr
1 29.80359 17.565675 42.04151
2 25.05335 12.827626 37.27907
3 20.30310  8.077742 32.52846
```

The Confidence and Prediction interval are not same for Response Values.

The Fitted Values are same for both the intervals, just that the range is wider in case of Prediction Interval.This is due to Standard Error in Prediction Interval.

Prediction interval shows the uncertainty around a single value, while confidence interval shows the uncertainty around the mean prediction, that is why prediction interval has wider range.

In [ ]:
```
#Modified Plot
cat("Modifying the model to include lstat^2")
linearModel2=lm(medv~lstat + I(lstat^2))
```

```
summary(linearModel2)
coef(linearModel2)
cat("R-squared for Non-Linear Model: ",summary(linearModel2)$r.sq)
par(mfrow=c(2,2))
plot(linearModel2,col="lawngreen")
```

In [ ]:

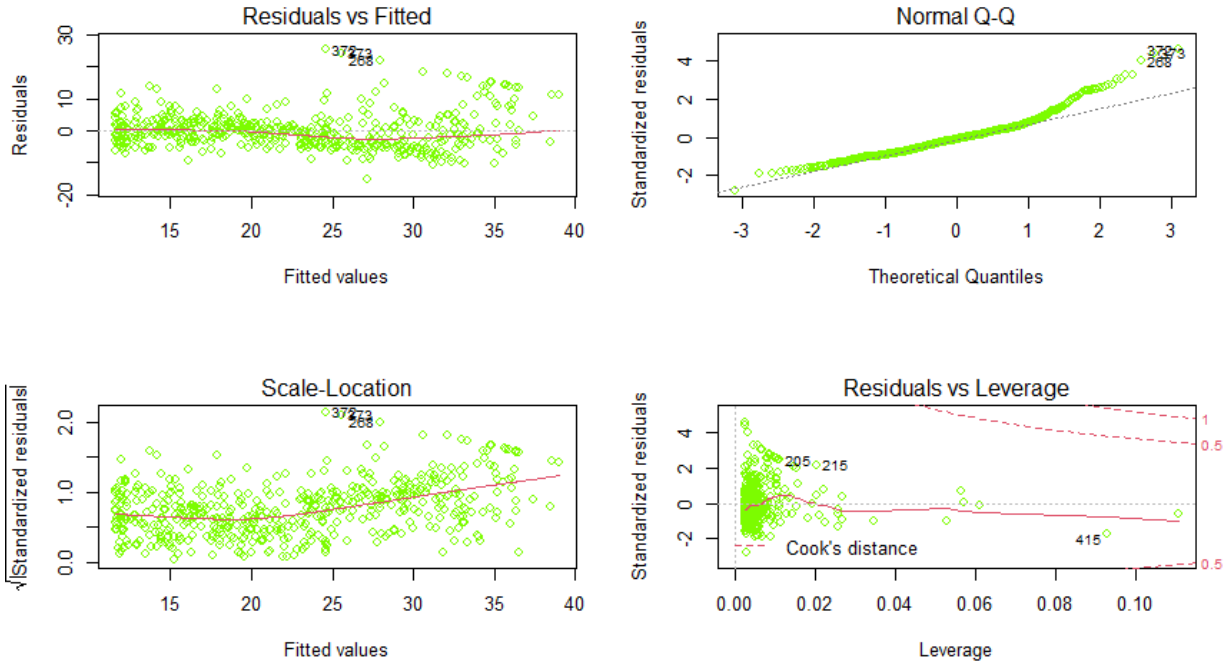
```
Call:
lm(formula = medv ~ lstat + I(lstat^2))

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

> coef(linearModel2)
(Intercept)       lstat  I(lstat^2)
42.86200733 -2.33282110  0.04354689
> cat("R-squared for Non-Linear Model: ",summary(linearModel2)$r.sq)
R-squared for Non-Linear Model:  0.6407169
```

The Value of R-Squared has increased from 54% to 64%. Meaning 10% more variance can be explained by the model.

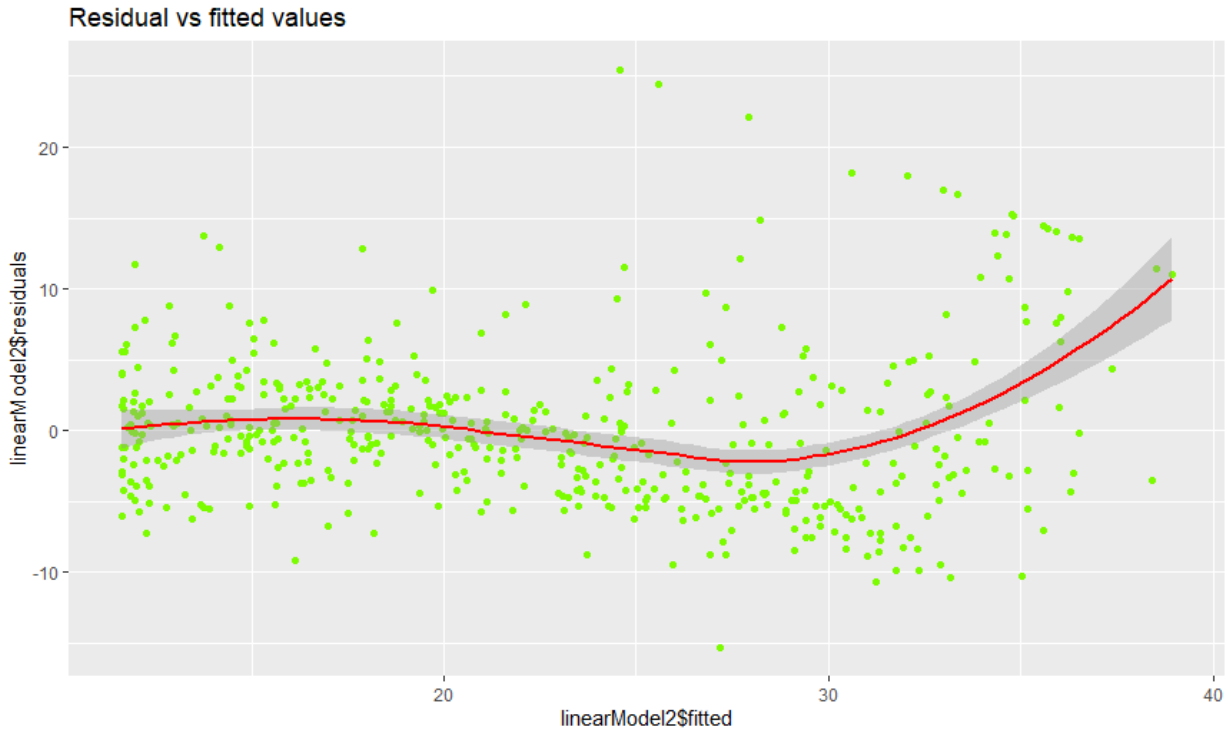
Hence Performance of Model has improved as we moved to Higher Degree of Polynomial.

In [ ]:

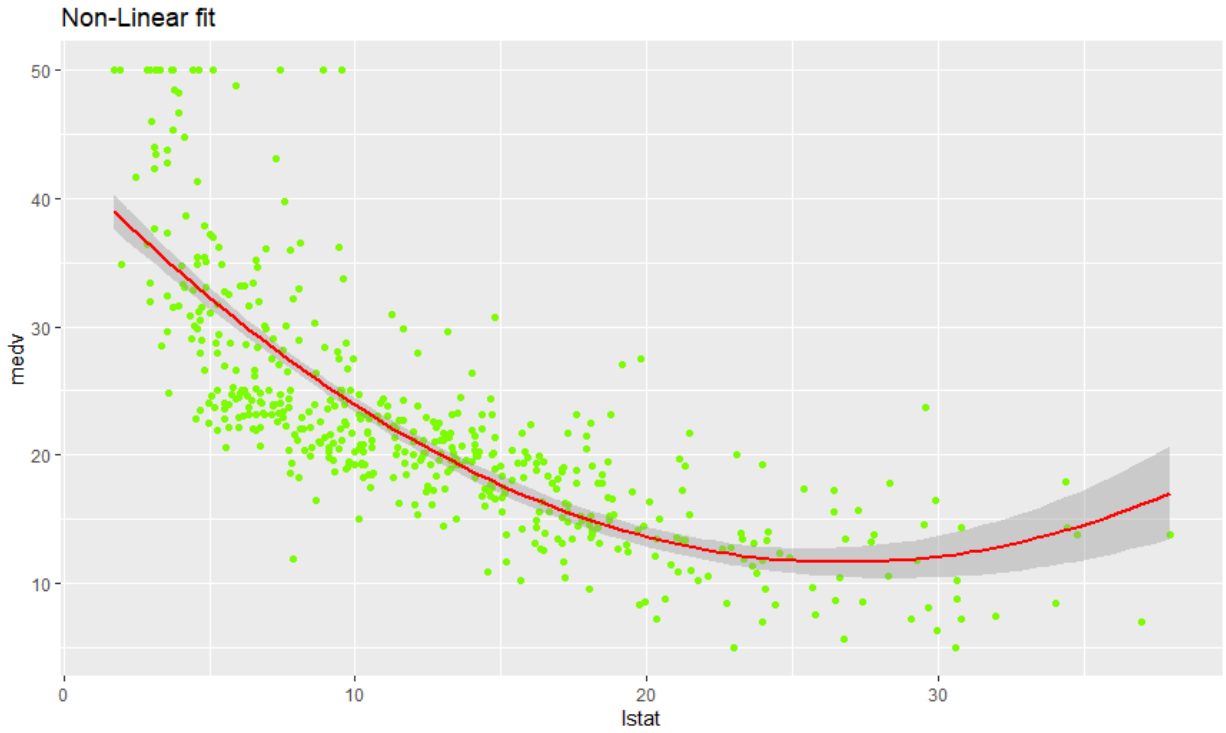
```
par(mfrow=c(2,2))
plot(linearModel2,col="lawngreen")
```



In [ ]:
```
#plot for fitted values vs residual:
ggplot(data=bostonData,aes(x=linearModel2$fitted,y=linearModel2$residuals)) +
  geom_point(col="lawngreen")+
  stat_smooth(col="red")+
  ggtitle("Residual vs fitted values")
```



In [ ]:
```
#plot for non-linear fit:
ggplot(data=bostonData,aes(x=lstat,y=medv)) +
  geom_point(col="lawngreen")+
  stat_smooth(formula = y ~ x + I(x^2),method="lm",col="red")+
  ggtitle("Non-Linear fit")
```



In [ ]:
```
anova(linearModel1,linearModel2)
```

In [ ]:
```
Analysis of Variance Table

Model 1: medv ~ lstat
Model 2: medv ~ lstat + I(lstat^2)
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    504 19472
2    503 15347  1    4125.1 135.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```