Assignment-01

Sumanth Donthula

2023-01-28

1) Recitation Exercise

1.1

Chapter-2:

Exercise 1:

1.a) As the sample size n is extremely large and number of predictors p are less we can use a high flexible model because there will be less chances of having high variance and high bias. More over the model will get enough data and identify trend in the data.

1.b) As the sample size n is small and number of predictors are high, usually the lower flexible models give better predictions compared to higher flexible models as higher flexible models will may give high variance model(Overfitting). The higher flexible model won't be able to process and identify the trend in data as data is less.

1.c) The relationship between the predictors and response is highly non-linear, as we know that lower flexible models assumes linearity in finding patterns. SO, we the higher flexible models can process the data and give better relation model than lower flexible models.

1.d) As the variance of error term is extremely high the higher flexible models will results in higher variance(Overfitting). SO, the lower flexible models are better.

Exercise 2:

2.a) As we are making the predictions about CEO's salaries we need a regression model as we are evaluating salary range as numeric w.r.t. to predictors or features. Inferene, as we are evaluating the range of salary. The value of n is 500, p is 3.

2.b) As we are making a model which predicts whether our product is success or failure and its a classification problem. Prediction and n is 20, p is 13.

2.c) As we are predicting the USD/EURO exchange rate in relation which is numeric and its a regression problem. Prediction and n is 52, p is 3.

Exercise 4:

a) Classification:

Email Spam Recognition: Response: Email Filtering whether its spam or not.

Predictors: Content in the mail, frequency of mails, email id, subject

Prediction.

Handwritten Digit Recognition:

Response: Identify the number in Image

Predictors: Images of different digits to identify.

Prediction.

Cancer Prediction:

Response: Whether Cancer is Malign or Benign

Predictors: Tumor length, Tumor width, Area, Blood Pressure, Heart Rate

Prediction.

b) Regression:

House Price Prediction:

Response: Predict house prices

Predictors: Location, Size, Number of bed rooms, Schools nearby, Crime rate

Prediction.

Store Sales Prediction:

Response: Predict next months sales data

Predictors: Sales value, number of customers visited, Whether condition, holiday data

Prediction.

Stock value Prediction:

Response: Predict stock value

Predictors: Assets, Liabilities, Profits/Earning, Price of Stock, Economy

Prediction.

c) Clustering Analysis:

Customer Segmentation:

Response: Segmenting the customers of a company

Predictors: Type of product they buy, number of items they buy, frequency of purchase

Inference on what kind of product customer will buy.

Social Network Analysis:

Response: Segment the people in Social Network and make inferences on spending habbits

Predictors: Number of friends, Location, Content they post, Frequency of posts

Inference, as we want to analyze peoples spending habbits for making inference.

Books Clustering System in Library:

Response: Categorize books of same class in Library

Predictors: Book Title, Author, Book Field, Price of the book

Predict which area does this book falls.

Exercise 6:

Parametric Statistical Approach:

In parametric statistical approach we will assume a model function which contains parameters and models f.

Non-Parametric Statistical Approach:

A non Parametric model wont assume the model f with defined parameters and often needs more data compared to parametric models.

The advantages of parametric model is it assumes a function f, studying properties of parameters is also easy with less number of data points.

The advantages of Non Parametric model is it will be good to fit when the relation between predictors and Y is non linear, but having less number of data points will results in over fitting the model.

Exercise 7:

a) The Euclidean Distance from (0,0,0) to all the observations is

Distance d= $((X2\text{-}X1)^{2+(Y2\text{-}Y)}2+(Z2\text{-}Z1)^{2)}1/2$

Observation 1.)3 Red Observation 2.)2 Red Observation 3.)3.16 Red Observation 4.)2.23 Green Observation 5.)1.41 Green Observation 6.)1.73 Red

b)

With K=1 as our point (0,0,0) Observation 5 is nearer to the point so the model predicts Y as Green.

c)

The 3 shorter points from (0,0,0) are Observations 5,6 and 2 which are green, red and red.

The prediction we will get will be red as its probability is 2/3.

d)

If the model is non linear for lower values of k would give a better and linear decision boundaries.

Chapter-3:

Exercise 1:

The p values of BetaTV and Betaradio is < 0.0001 which is less than 0.05 and there is no significance to prove that these parameters are 0. Where as the p value of Betanewspaper is 0.8599 which is significant that this value can be 0.

Exercise 3:

a) The Model as given is

y = 50 + 20GPA + 35Level+ 0.07IQ + 0.01GPA × IQ - 10 * GPA * Level

The equation for high school graduates will be when level=0

y = 50 + 20GPA + 0.07IQ + 0.01GPA × IQ

The equation for college graduates will be when level=1

y = 50 + 20GPA + 35+ 0.07IQ + 0.01GPA × IQ - 10 X GPA = 85 + 10GPA+ 0.07IQ + 0.01GPA x IQ

It is significant that for high value of GPA high school graduates are getting more paid if GPA is greater than or equal to 3.5 from the equations. Option (iii) is valid.

b)

y=85+10*4+0.07*110+0.01*4*110 =137.1 K

c) False : We can not say that the beta of interaction term GPA/Iq is low, so it is unsignificant because we need to do hypothesis test as we can not interpret beta values by seeing the values of beta.

Exercise 4:

a)

As the true relationship between predictors and output is linear the RSS for Non Linear(Cubic Model) model should be less ideally.

b)

As we have small data n which is 100, cubic model would have over fitted the model and the RSS would be high compared to linear model. As data is less we can not infer that the RSS of Non Linear Model is high.

c) As the true relationship between predictors and output is not linear the RSS for Linear model should be less ideally.

d) With fewer data points n=100 it is inconclusive that whether the true relationship is linear or not linear. If the data is Non Linear then the RSS of non linear model would have been low or else RsS of linear model would have been low.

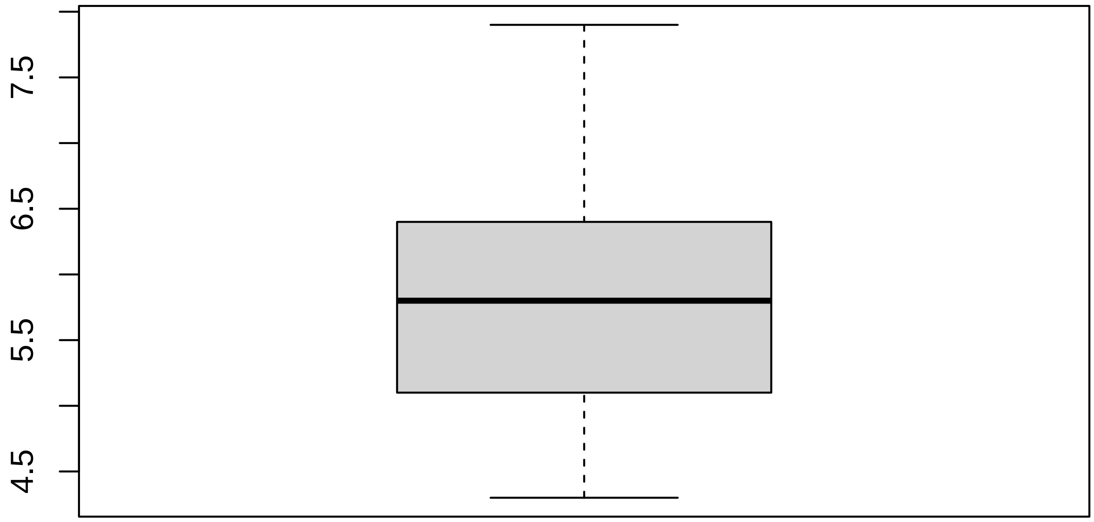
Practicum Problems:

Problem 1:

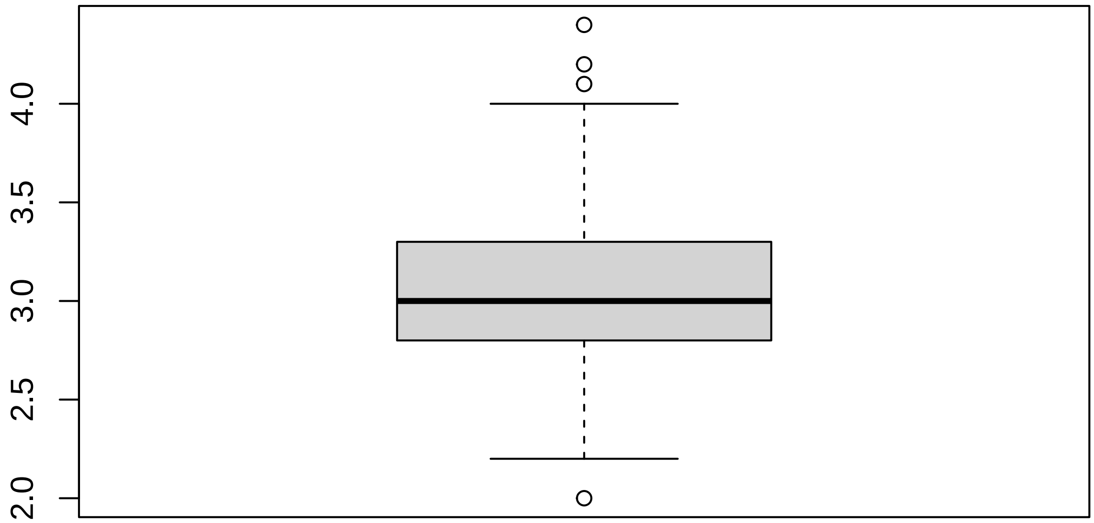
Loading the Dataset and plotting boxplot of 4 features in the dataset.

From the plot, IQR of PetalLength is quite high which is 3.5.

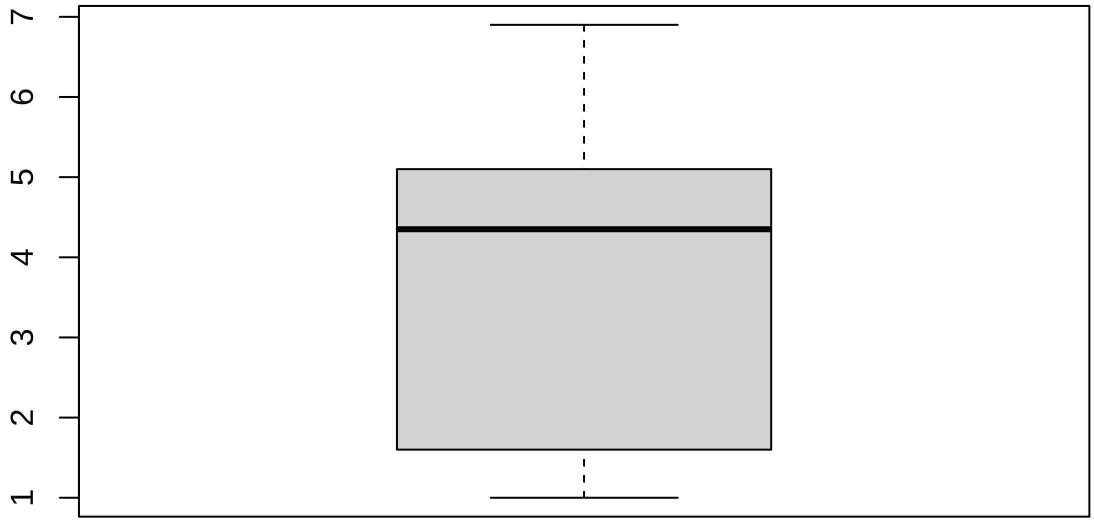
```
library(datasets)
data(iris)

sepalLength=iris[,1]
sepalWidth=iris[,2]
PetalLength=iris[,3]
petalWidth=iris[,4]
species=iris[,5]

boxplot(sepalLength)
```



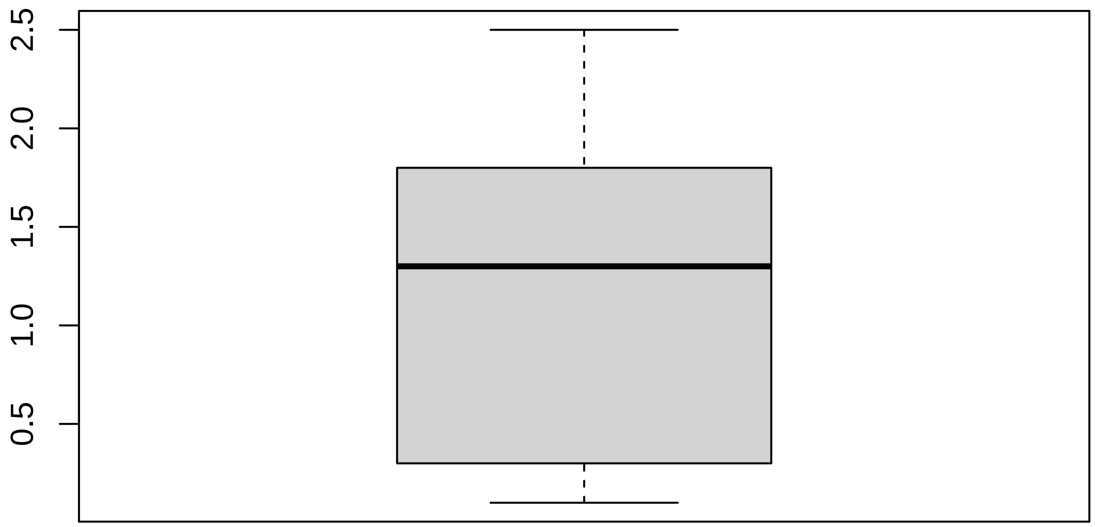
```
boxplot(sepalWidth)
```



```
boxplot(PetalLength)
```



```
boxplot(petalWidth)
```



```
print("IQR of sepalLength")
```

```
## [1] "IQR of sepalLength"
```

```
IQR(sepalLength)
```

```
## [1] 1.3
```

```
print("IQR of sepalWidth")
```

```
## [1] "IQR of sepalWidth"
```

```
IQR(sepalWidth)
```

```
## [1] 0.5
```

```
print("IQR of PetalLength")
```

```
## [1] "IQR of PetalLength"
```

```
IQR(PetalLength)
```

```
## [1] 3.5
```

```
print("IQR of petalWidth")
```

```
## [1] "IQR of petalWidth"
```

```
IQR(petalWidth)
```

```
## [1] 1.5
```

```
sd(sepalLength)
```

```
## [1] 0.8280661
```

```
sd(sepalWidth)
```

```
## [1] 0.4358663
```

```
sd(PetalLength)
```

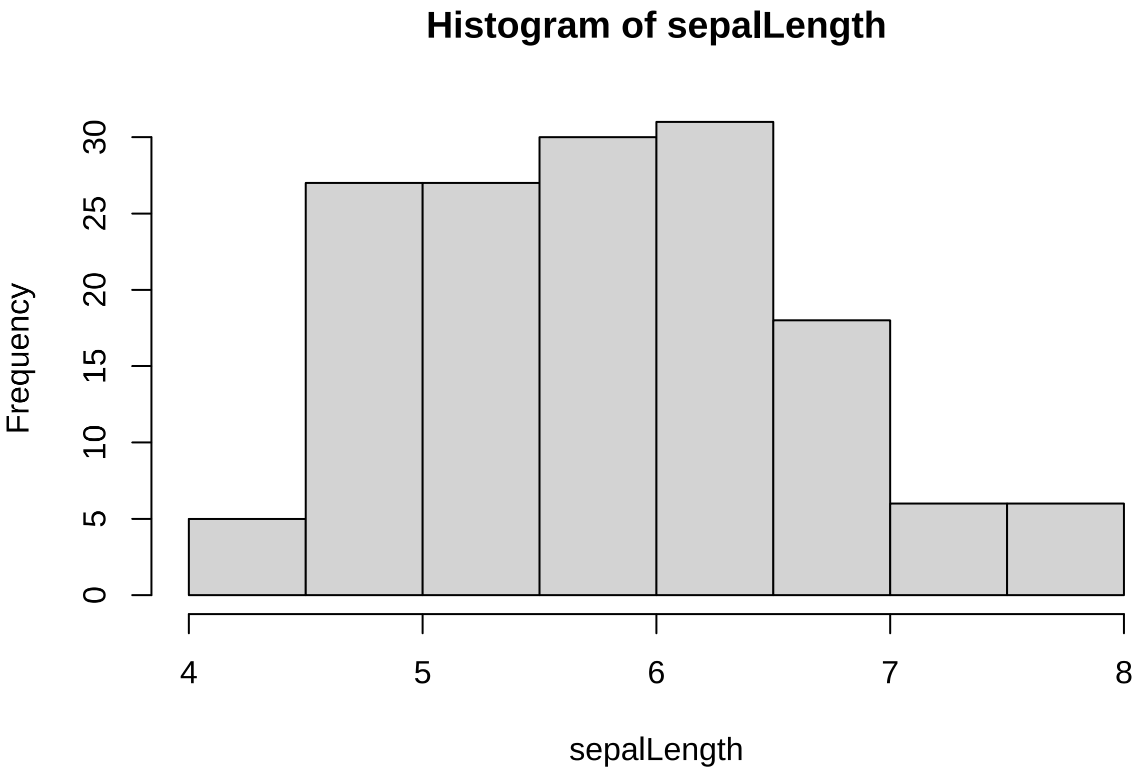
```
## [1] 1.765298
```

```
sd(petalWidth)
```

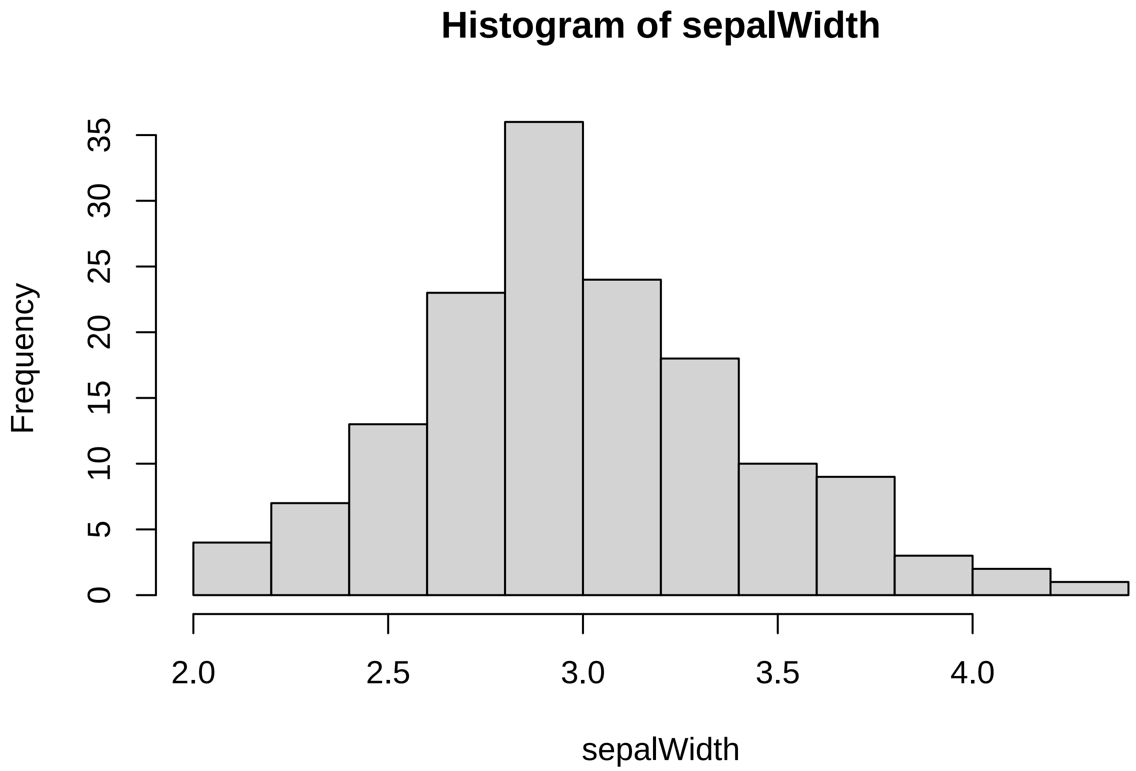
```
## [1] 0.7622377
```

From the histograms of features we can see that petalLength and petalWidth are not normally distributed and standard deviation is quite high. So, these features does not follow empirical rule.

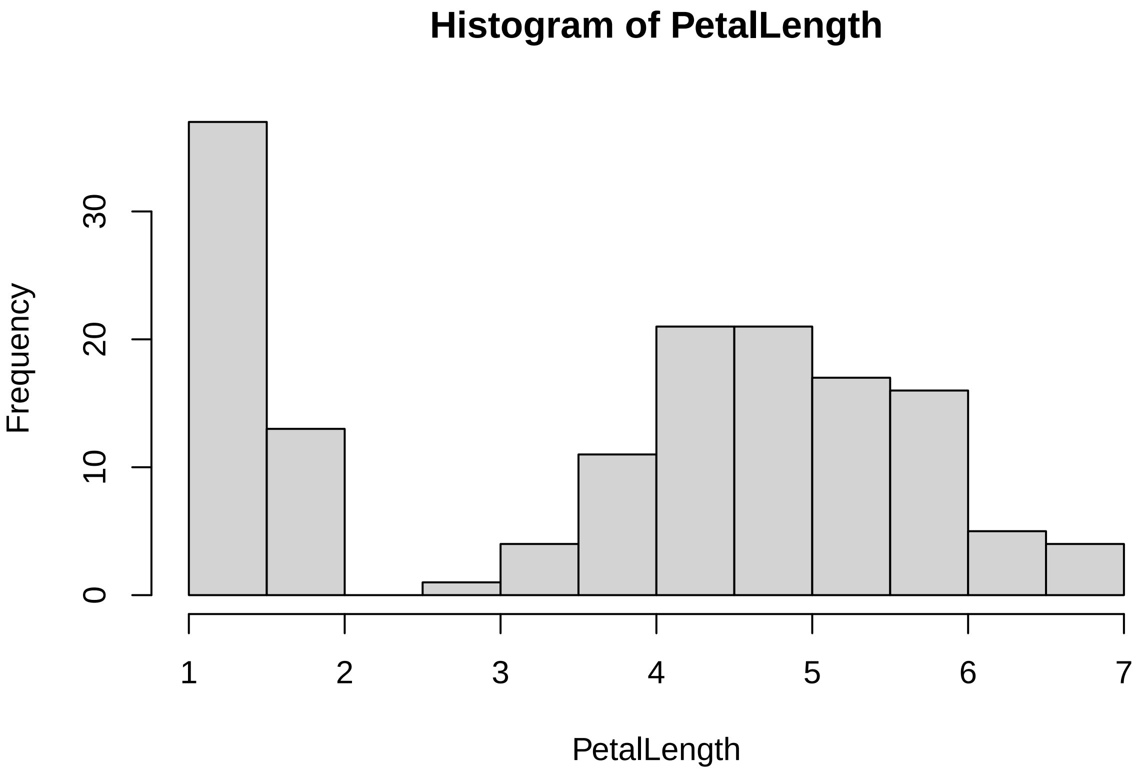
```
hist(sepalLength)
```



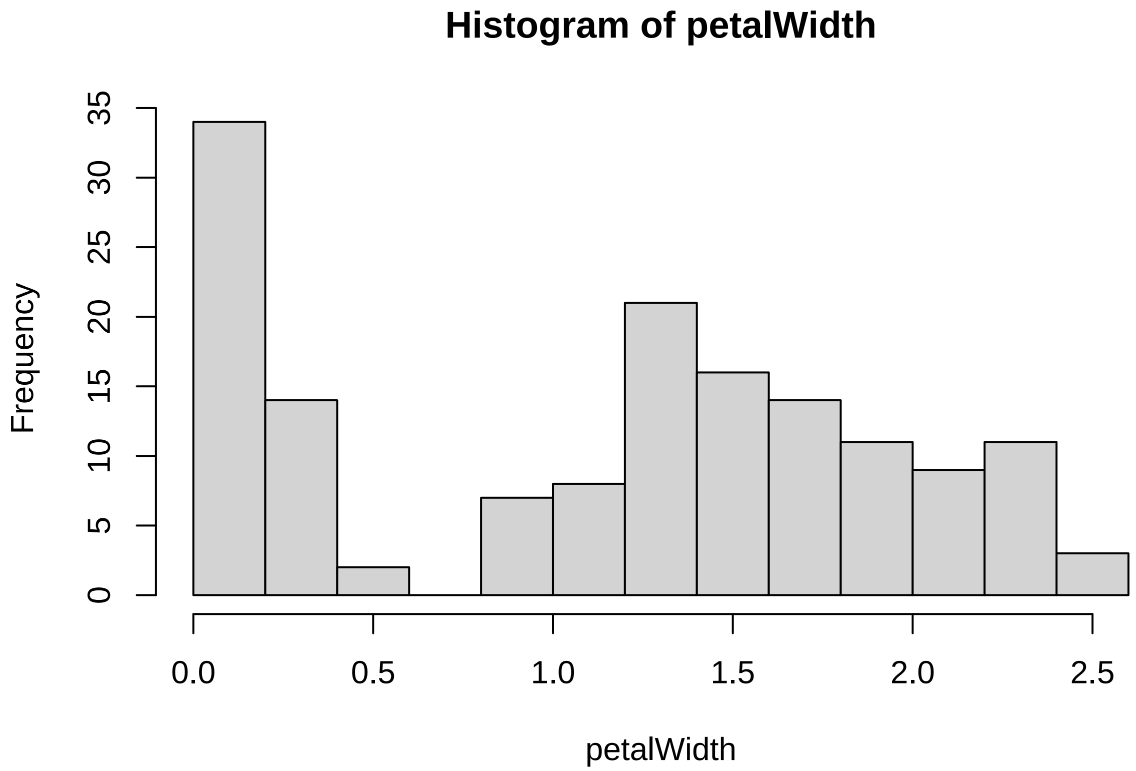
```
hist(sepalWidth)
```



```
hist(PetalLength)
```

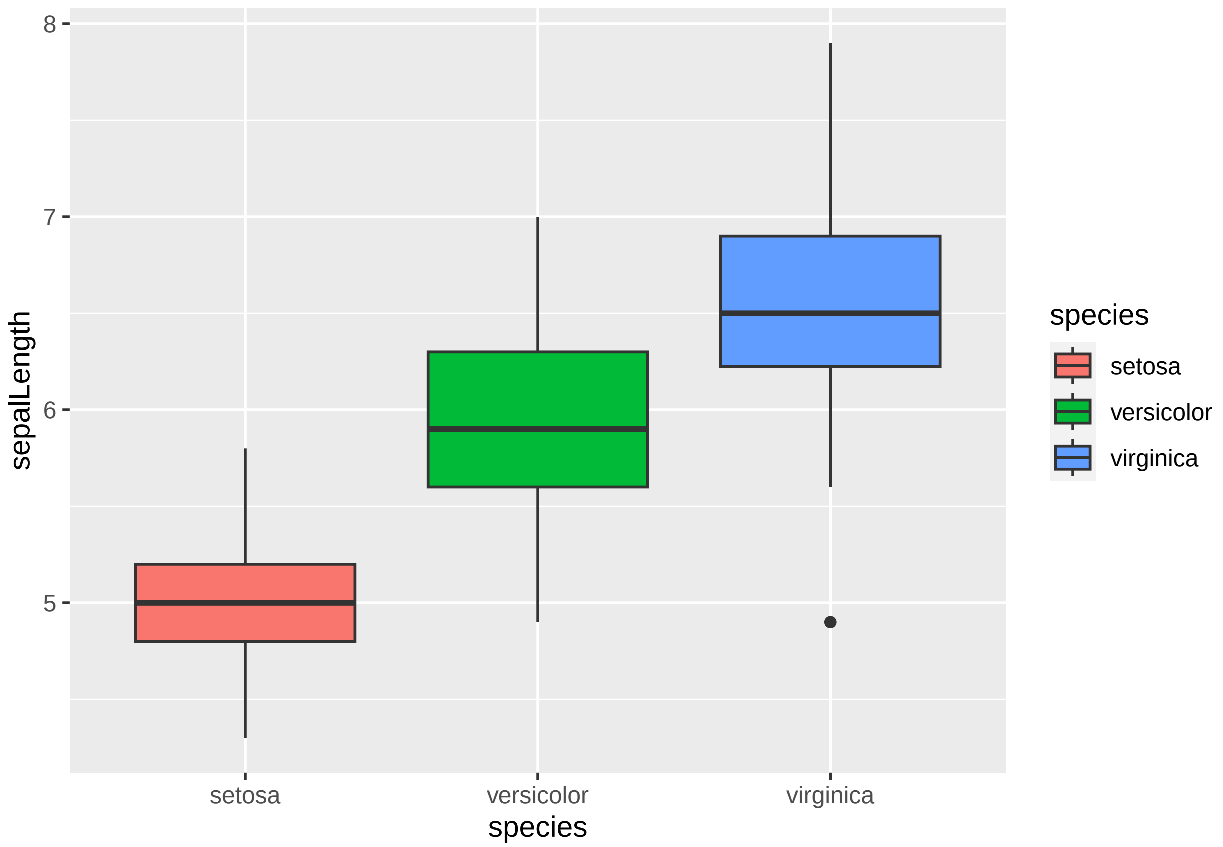


```
hist(petalWidth)
```

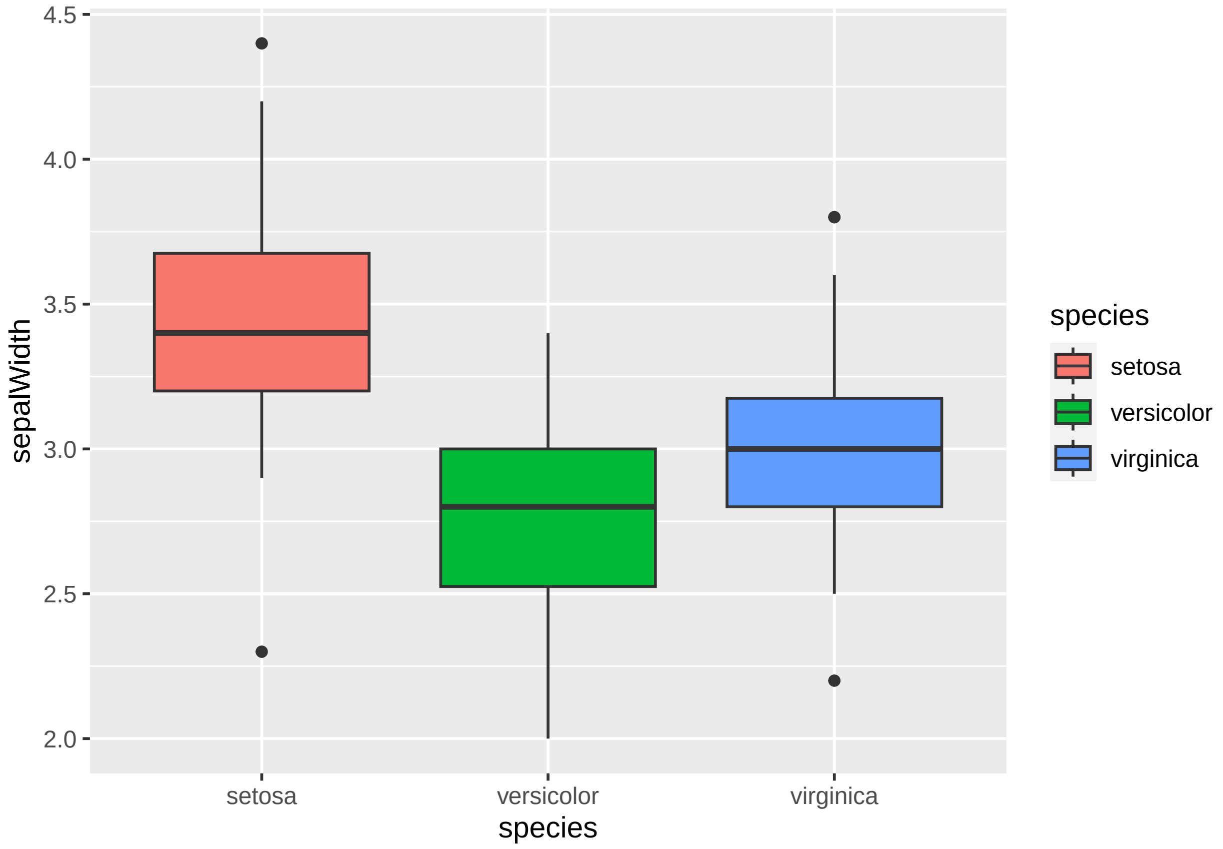


Plotting boxplots of features with respect to species class. Verginica, because its IQR is high compared to other classes.

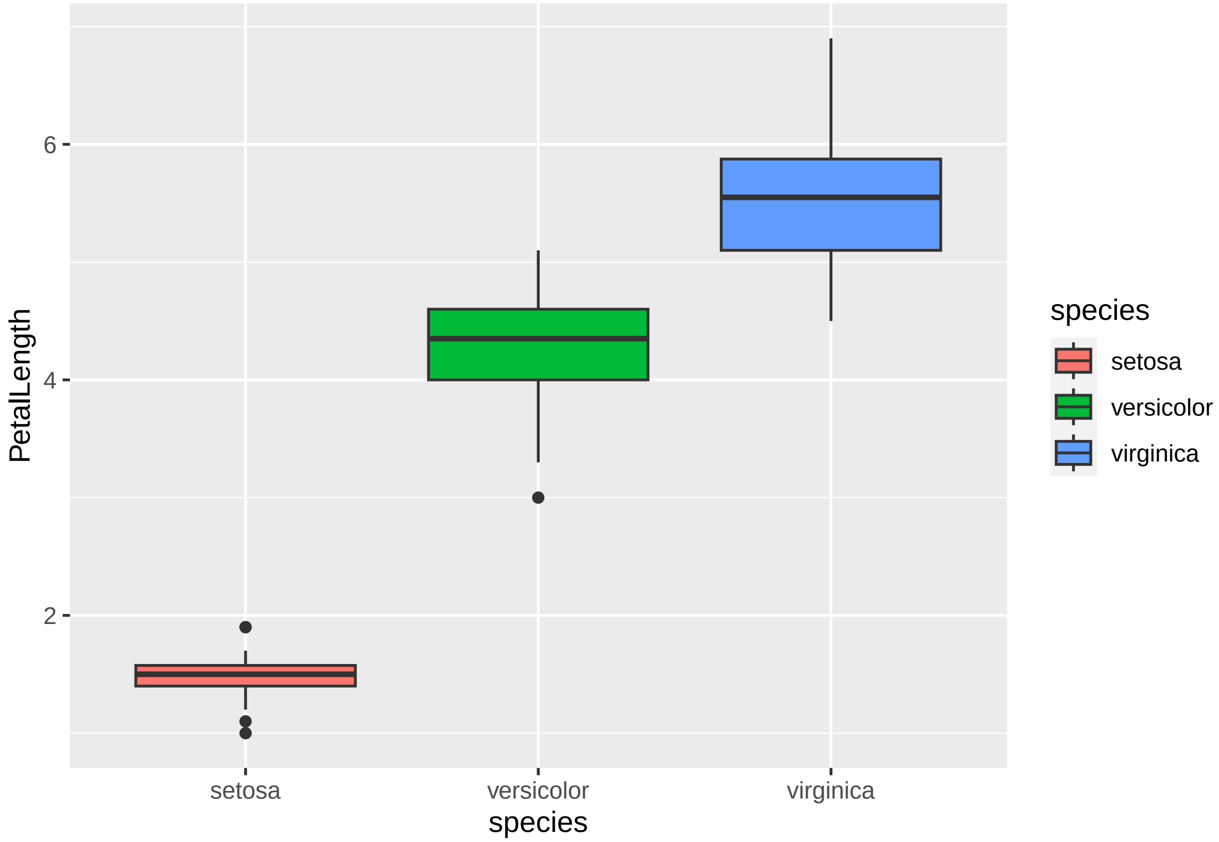
```
library(ggplot2)
ggplot(iris, aes(y = sepalLength, x = species,fill=species)) +
  geom_boxplot()
```



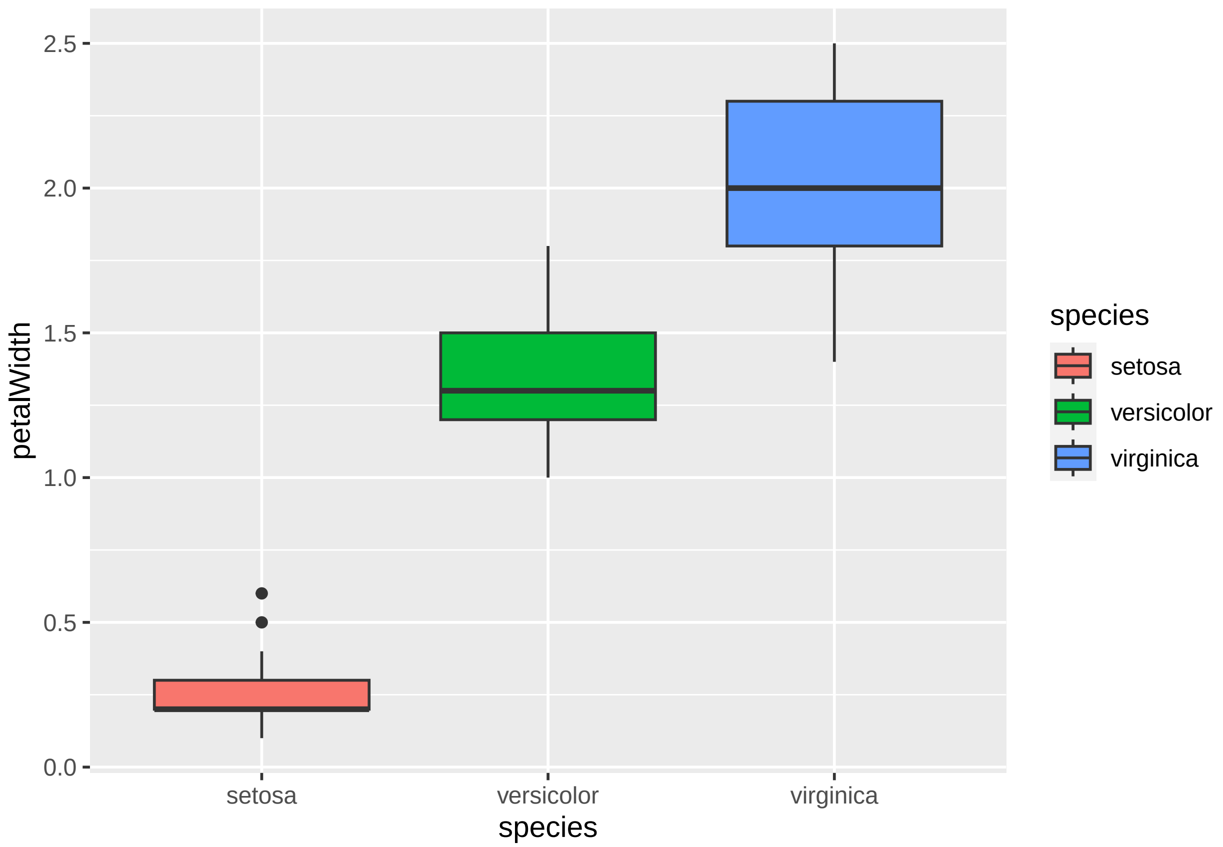
```
ggplot(iris, aes(y = sepalWidth, x = species,fill=species)) +
  geom_boxplot()
```



```
ggplot(iris, aes(y = PetalLength, x = species,fill=species)) +
  geom_boxplot()
```



```
ggplot(iris, aes(y = petalWidth, x = species,fill=species)) +
  geom_boxplot()
```



Problem 2

Loading data set and producing a 5-number summary of each feature.

```
library(datasets)
data(trees)


girth=trees[,1]
height=trees[,2]
volume=trees[,3]




summary(fivenum(girth))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.30   11.05   12.90   13.62   15.25   20.60
```

```
summary(fivenum(height))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    63.0    72.0    76.0    75.6    80.0    87.0
```

```
summary(fivenum(volume))
```

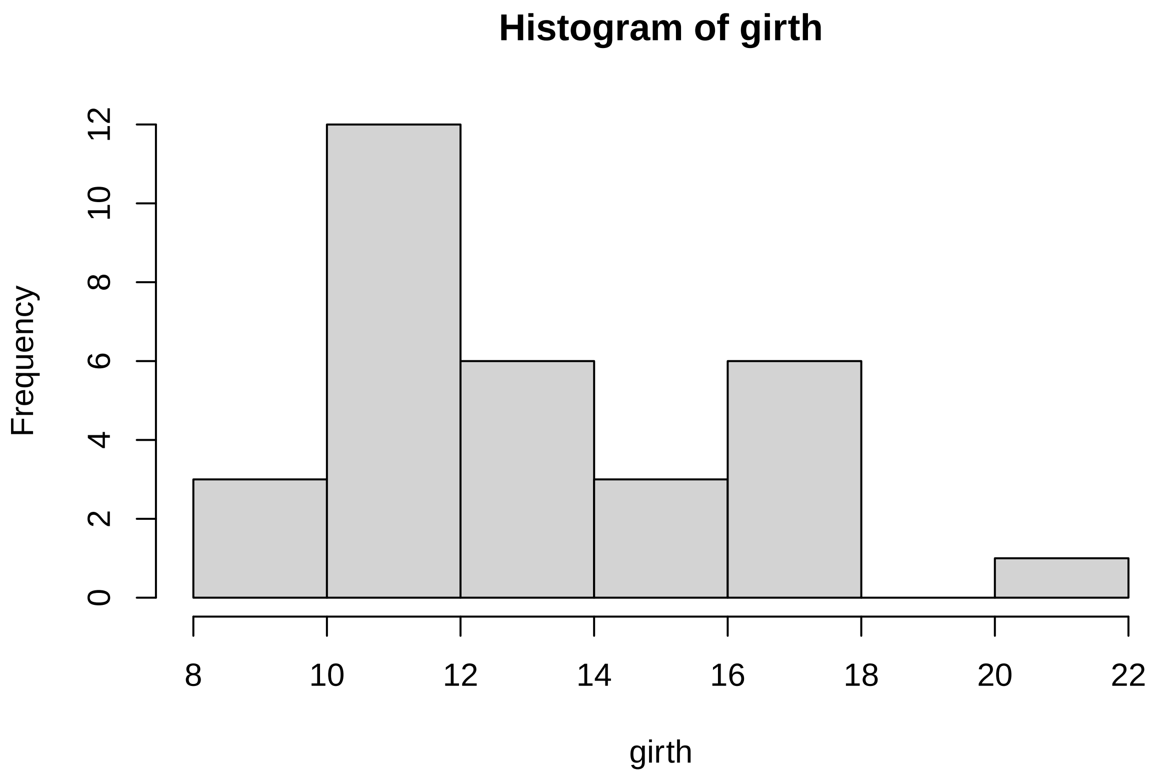
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.20   19.40   24.20   33.62   37.30   77.00
```

Creating histogram of each feature.

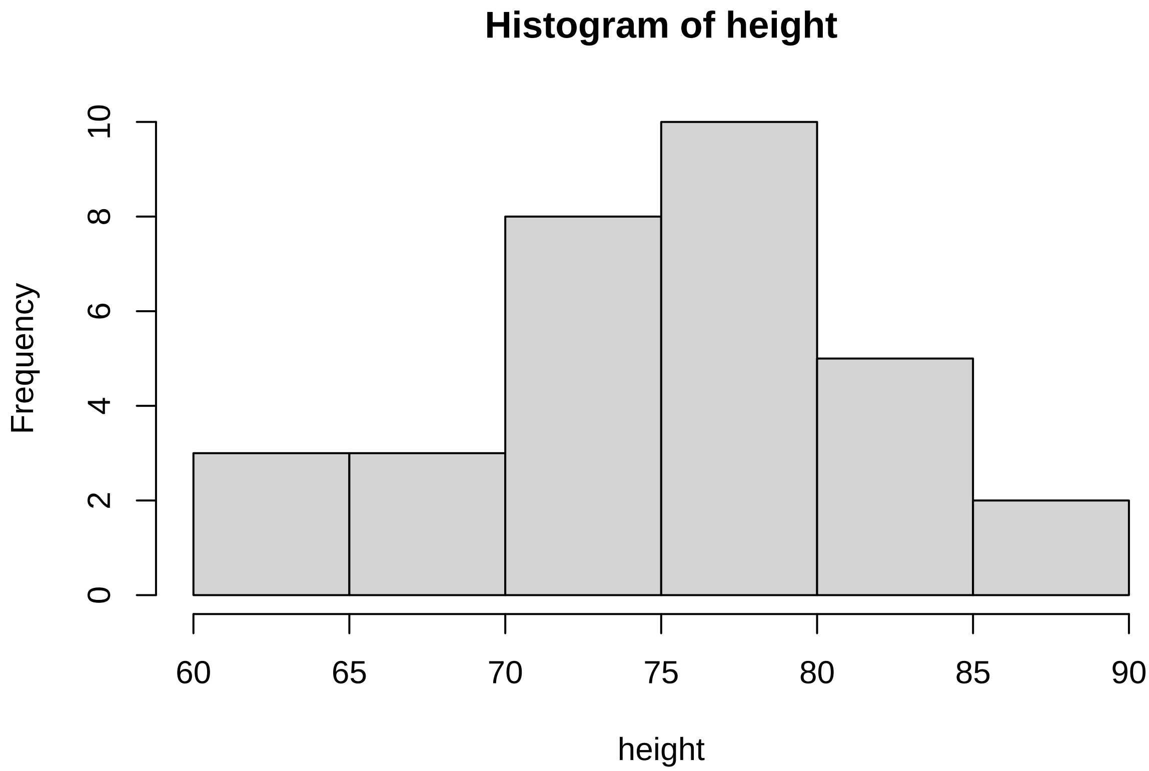
The features girth and height appeared to be normally distributed.

The feature volume seems to be Right skewed.

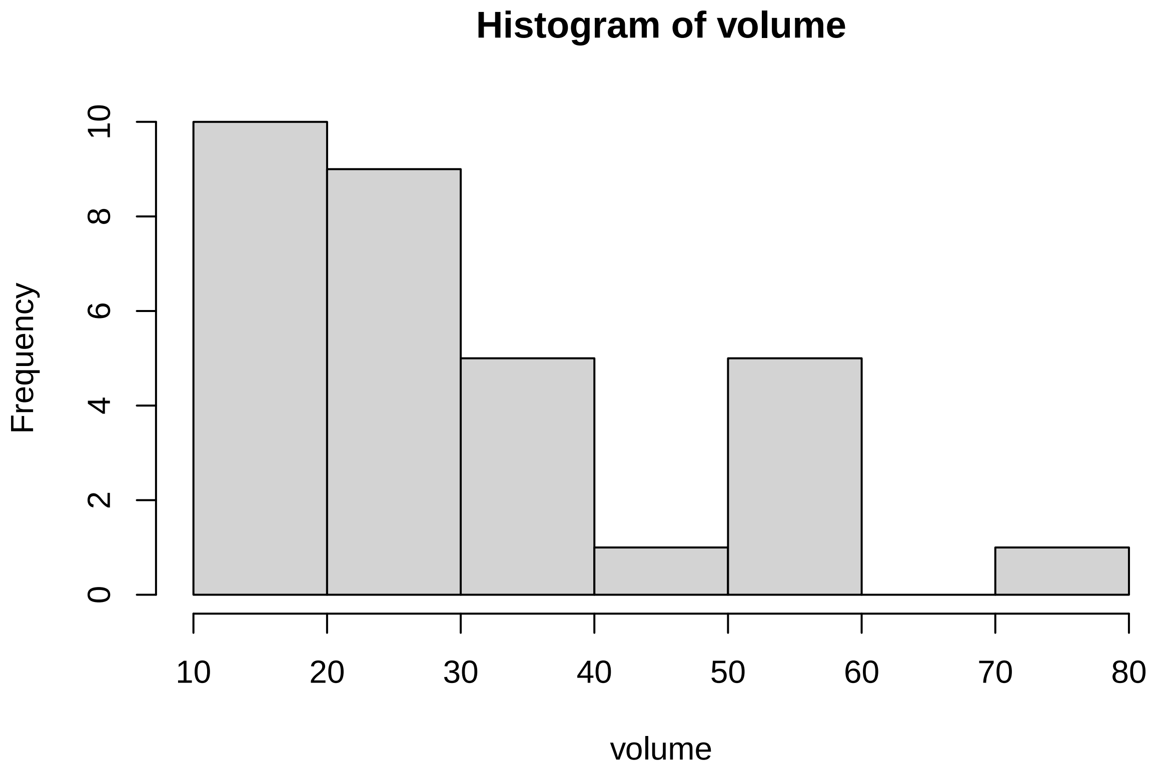
```
hist(girth)
```



```
hist(height)
```


Histogram of height
Frequency
0
2
4
6
8
10
60
65
70
75
80
85
90
height

```
hist(volume)
```



The value of skewness should be Symmetric if Values between -0.5 to 0.5 Modarately skewed : -1 and -.5 or 0.5 and 1 Highly skewed : <-1 and >1

The skewness values for height and girth is in range -0.5 to 0.5 which is symmetric and the value of volume is >1 which is skewed interms of distribution. This values agree with visual representation.

```
library(moments)

skewness(girth)
```

```
## [1] 0.5263163
```

```
skewness(height)
```

```
## [1] -0.374869
```

```
skewness(volume)
```

```
## [1] 1.064357
```

Problem 3:

Loading the dataset and filling the values '?' in horsepower with the median of the horsepower.

There is no much change in mean after filling the missing values with median.

```r
autoMpg = read.table(url("https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.da

summary(as.numeric(autoMpg$horsepower))
```

```
## Warning in summary(as.numeric(autoMpg$horsepower)): NAs introduced by coercion
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    46.0    75.0    93.5   104.5   126.0   230.0       6
```

```r
print("Mean before filling NA's or ? with Median")
```

```
## [1] "Mean before filling NA's or ? with Median"
```

```r
mean(as.numeric(autoMpg$horsepower),na.rm=TRUE)
```

```
## Warning in mean(as.numeric(autoMpg$horsepower), na.rm = TRUE): NAs introduced by
## coercion
```

```
## [1] 104.4694
```

```r
autoMpg$horsepower[is.na(as.numeric(autoMpg$horsepower))] = median(as.numeric(autoMpg$horsepower),na.rm=
```

```
## Warning in median(as.numeric(autoMpg$horsepower), na.rm = TRUE): NAs introduced
## by coercion
```

```
## Warning in autoMpg$horsepower[is.na(as.numeric(autoMpg$horsepower))] =
## median(as.numeric(autoMpg$horsepower), : NAs introduced by coercion
```

```r
mean(as.numeric(autoMpg$horsepower),na.rm=TRUE)
```

```
## [1] 104.304
```

```r
print("Mean after filling NA's or ? with Median ")
```

```
## [1] "Mean after filling NA's or ? with Median "
```

```r
mean(as.numeric(autoMpg$horsepower),na.rm=TRUE)
```

```
## [1] 104.304
```

Problem 4:

Loading the dataset adn fitting the linear model.

```
library(MASS)
Boston=Boston


Y=Boston$medv
X=Boston$lstat

LinearModel=lm(Y~X)
```

Predictions and confidence intervals for the data points 5,10 and 15 are as follows.

The confidence intervals and prediction intervals are different as the confidence intervals implies the assumptions based on statistical population parameters it is less compared to prediction interval which uses the estimated function to find the intervals.

```
print("Prediction for 5")
```

```
## [1] "Prediction for 5"
```

```
predict(LinearModel,data.frame(X=5))
```

```
##        1
## 29.80359
```

```
print("Prediction for 10")
```

```
## [1] "Prediction for 10"
```

```
predict(LinearModel,data.frame(X=10))
```

```
##        1
## 25.05335
```

```
print("Prediction for 15")
```

```
## [1] "Prediction for 15"
```

```
predict(LinearModel,data.frame(X=15))
```

```
##       1
## 20.3031
```

```
print("Prediction and confidence interval for 5")
```

```
## [1] "Prediction and confidence interval for 5"
```

```
predInt = predict(lm(Y ~ X), data.frame(X=5), interval = "prediction")
predInt
```

```
##        fit      lwr      upr
## 1 29.80359 17.56567 42.04151
```

```
confInt = predict(lm(Y ~ X), data.frame(X=5), interval = "confidence")
confInt
```

```
##        fit      lwr      upr
## 1 29.80359 29.00741 30.59978
```

```
print("Prediction and confidence interval for 10")
```

```
## [1] "Prediction and confidence interval for 10"
```

```
predInt = predict(lm(Y ~ X), data.frame(X=10), interval = "prediction")
predInt
```

```
##        fit      lwr      upr
## 1 25.05335 12.82763 37.27907
```

```
confInt = predict(lm(Y ~ X), data.frame(X=10), interval = "confidence")
confInt
```

```
##        fit      lwr      upr
## 1 25.05335 24.47413 25.63256
```

```
print("Prediction and confidence interval for 15")
```

```
## [1] "Prediction and confidence interval for 15"
```

```
predInt = predict(lm(Y ~ X), data.frame(X=15), interval = "prediction")
predInt
```

```
##       fit      lwr      upr
## 1 20.3031 8.077742 32.52846
```

```
confInt = predict(lm(Y ~ X), data.frame(X=10), interval = "confidence")
confInt
```

```
##        fit      lwr      upr
## 1 25.05335 24.47413 25.63256
```

The R2 value of Non Linear model is high than Linear model which says that it explains Y better than Linear Model does.

```
NonLinearModel=lm(Y~X+I(X^2))
summary(LinearModel)
```

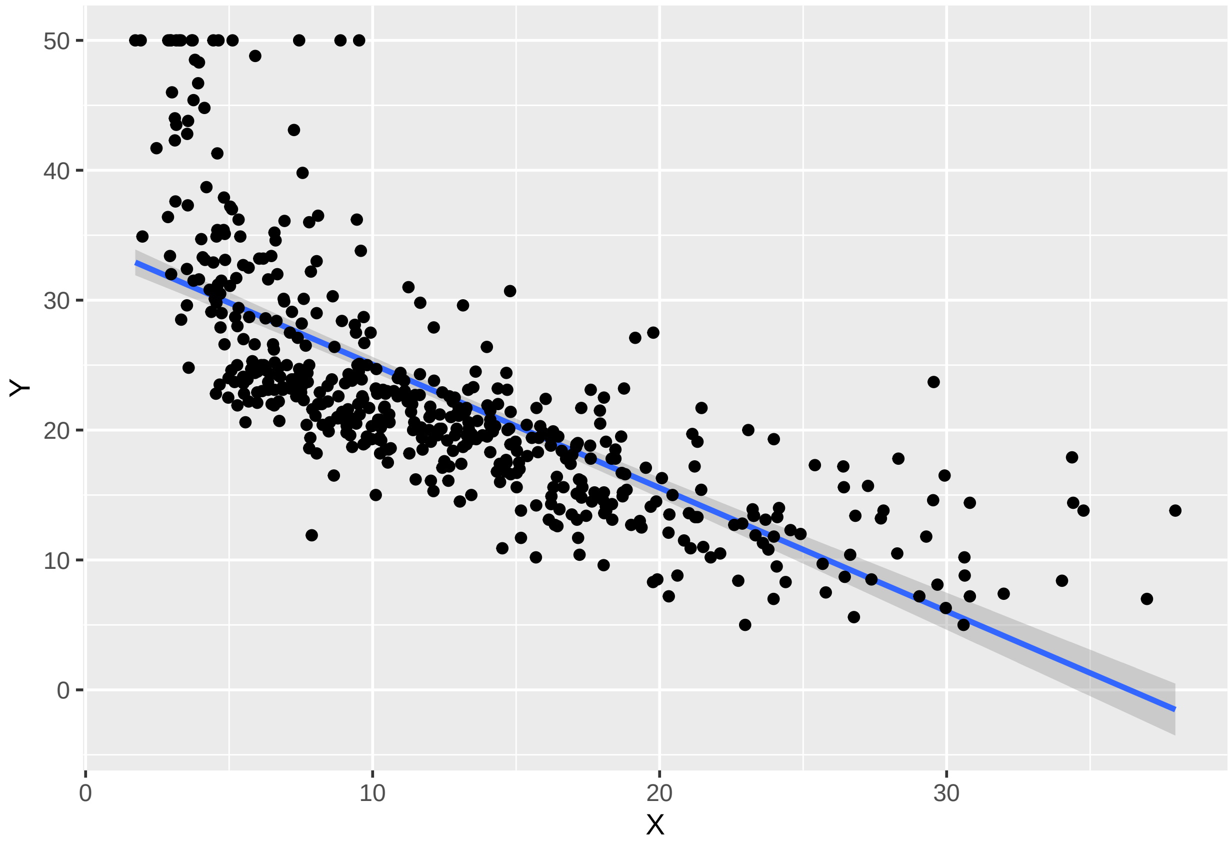
```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## X           -0.95005    0.03873  -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
```

```
summary(NonLinearModel)
```

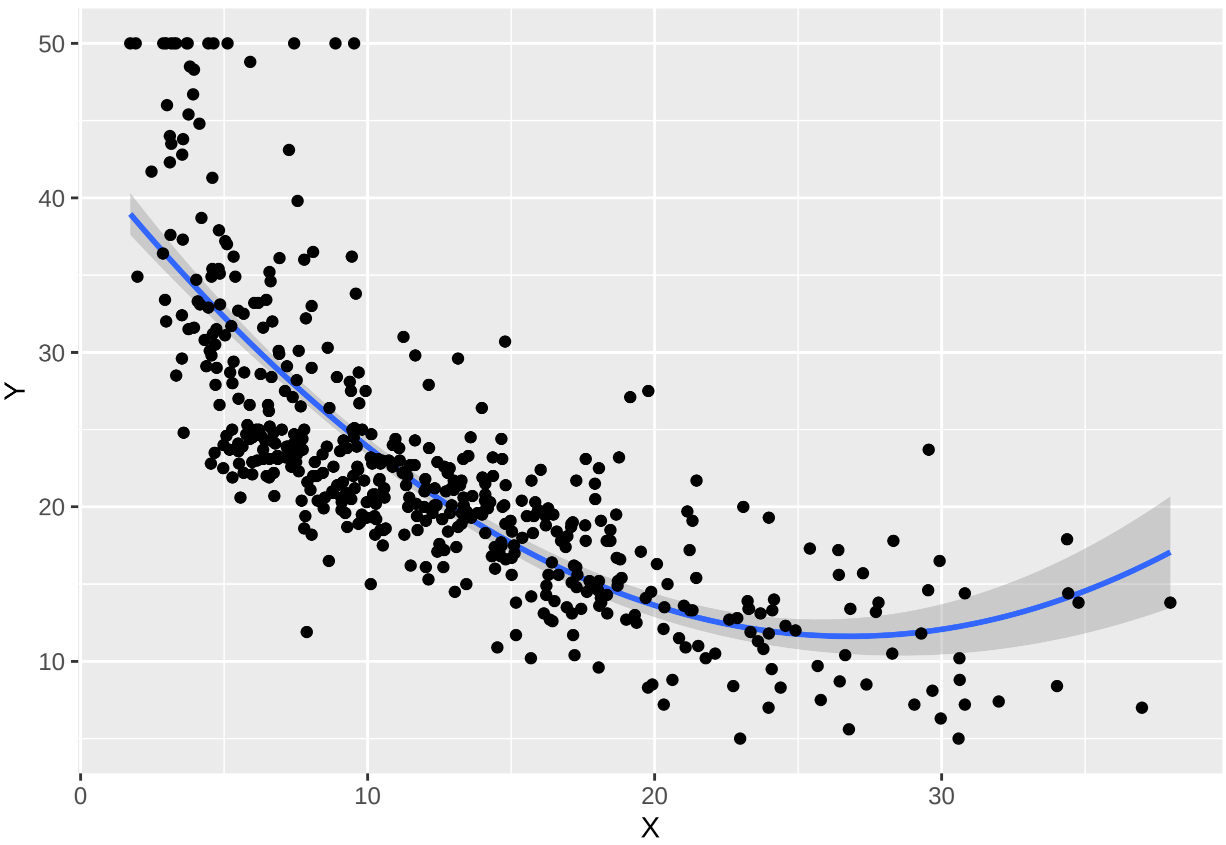
```
##
## Call:
## lm(formula = Y ~ X + I(X^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2834  -3.8313  -0.5295   2.3095  25.4148
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.862007   0.872084   49.15   <2e-16 ***
## X           -2.332821   0.123803  -18.84   <2e-16 ***
## I(X^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
```

Using GPlot to plot the linear and non linear models.

```
ggplot(data = Boston, aes(x = X, y = Y)) +
stat_smooth(method = "lm", aes(x = X, y = Y),formula=y~x,fullrange = TRUE) +
geom_point()
```



```
ggplot(data = Boston, aes(x = X, y = Y)) +
stat_smooth(method = "lm", aes(x = X, y = Y),formula=y~x+I(x^2),fullrange = TRUE) +
geom_point()
```


50
40
30
20
10
Y
0
10
20
30
X