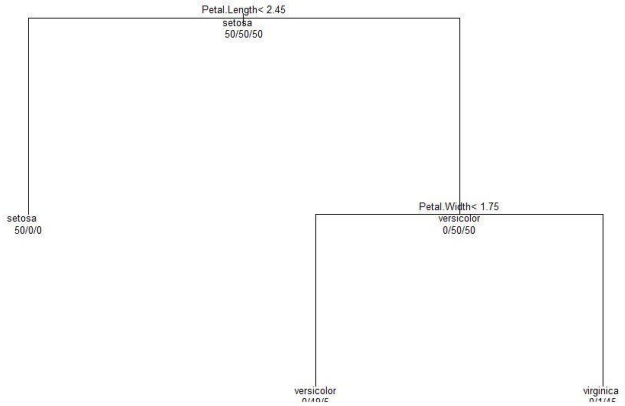
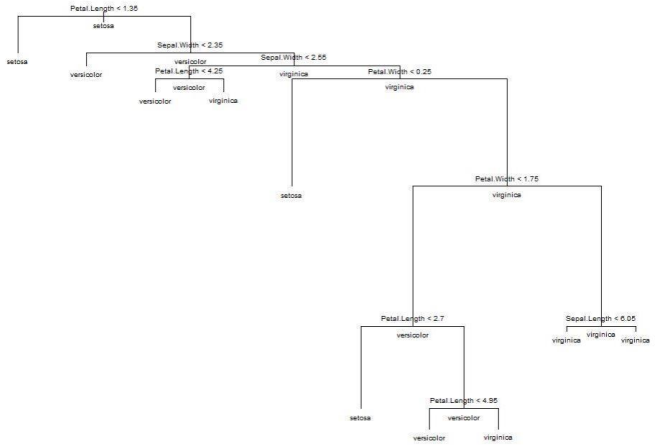
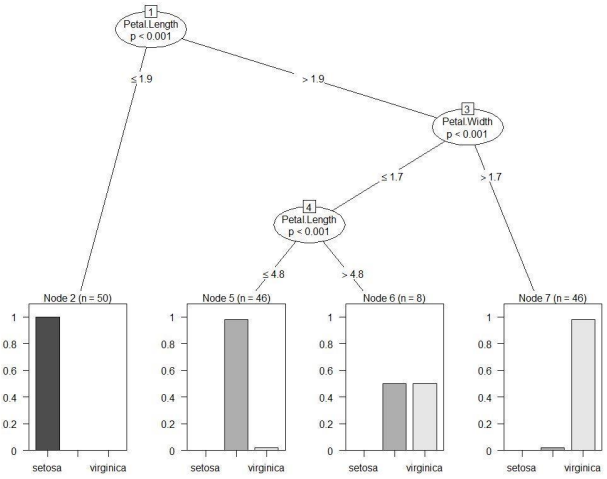
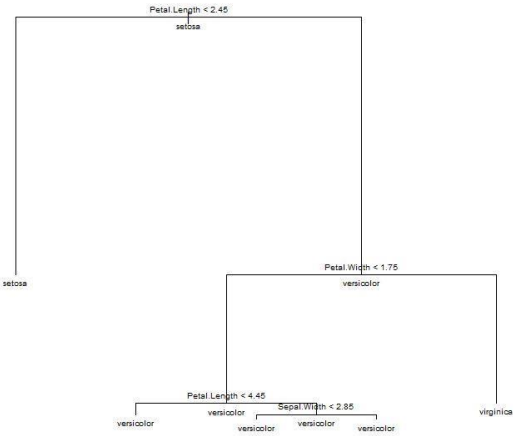
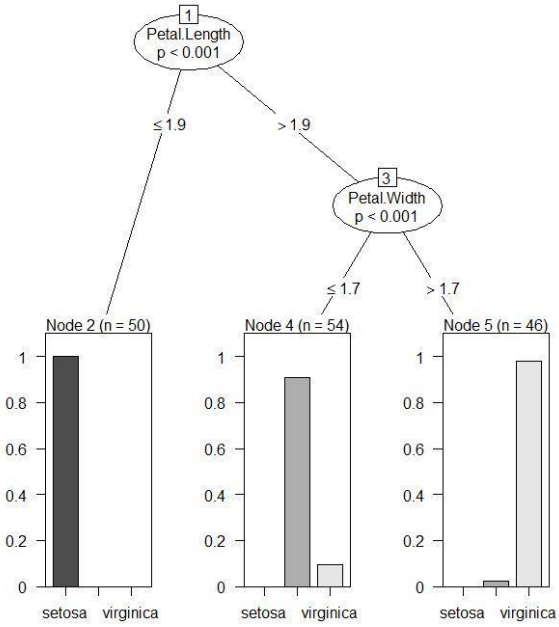
Practical 6


Petal.Length< 2.45
setosa
50/50/50
setosa
50/0/0
Petal.Width< 1.75
versicolor
0/50/50
versicolor
virginica


Petal.Length < 1.35
setosa
setosa
Sepal.Width < 2.35
versicolor
versicolor
Petal.Length < 4.25
versicolor
versicolor
virginica
Sepal.Width < 2.65
virginica
setosa
Petal.Width < 0.25
virginica
Petal.Width < 1.75
virginica
Petal.Length < 2.7
versicolor
setosa
Petal.Length < 4.95
versicolor
versicolor
virginica
Sepal.Length < 6.05
virginica
virginica
virginica

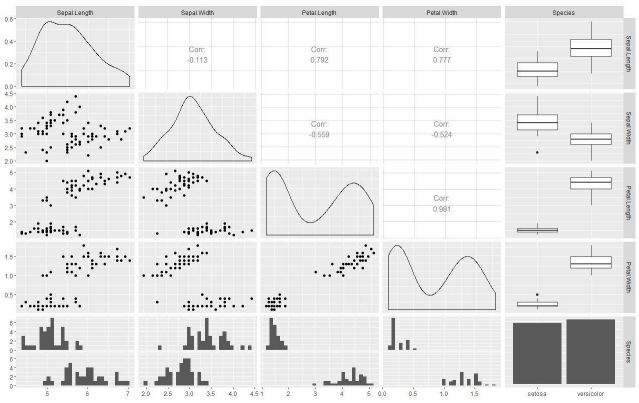

1
Petal.Length
p < 0.001
≤ 1.9
> 1.9
3
Petal.Width
p < 0.001
≤ 1.7
> 1.7
4
Petal.Length
p < 0.001
≤ 4.8
> 4.8
Node 2 (n = 50)
Node 5 (n = 46)
Node 6 (n = 8)
Node 7 (n = 46)
setosa
virginica


Petal.Length < 2.45
setosa
setosa
Petal.Width < 1.75
versicolor
Petal.Length < 4.45
versicolor
versicolor
Sepal.Width < 2.85
versicolor
versicolor
versicolor
virginica


1
Petal.Length
p < 0.001
≤ 1.9
> 1.9
3
Petal.Width
p < 0.001
≤ 1.7
> 1.7
Node 2 (n = 50)
Node 4 (n = 54)
Node 5 (n = 46)
setosa
virginica

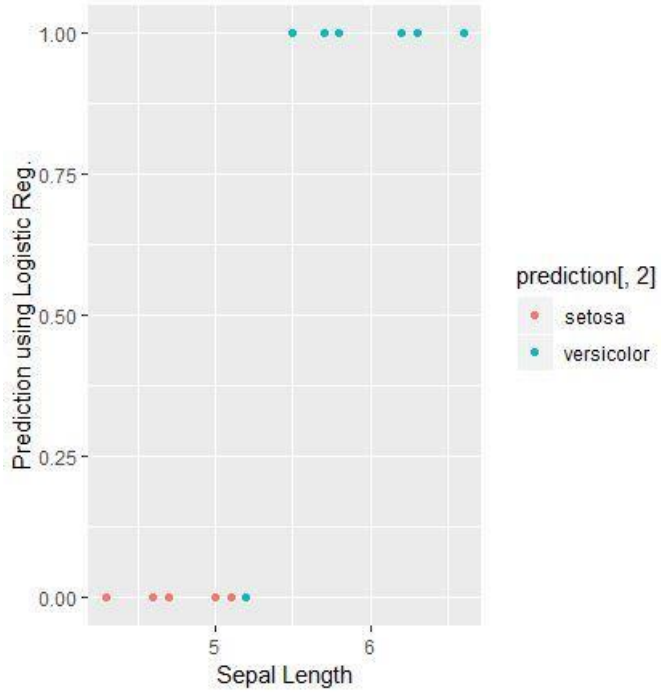
Pracrtical 5

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> str(ir_data)
'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 ...
```



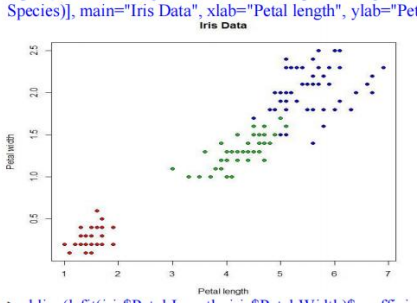

```
> summary(glfit)

Call:
glm(formula = y ~ x, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.94538  -0.50121   0.04079   0.45923   2.26238

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -25.386      5.517  -4.601 4.20e-06 ***
x              4.675      1.017   4.596 4.31e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 110.854  on 79  degrees of freedom
Residual deviance:  56.716  on 78  degrees of freedom
AIC: 60.716

Number of Fisher Scoring iterations: 6
```

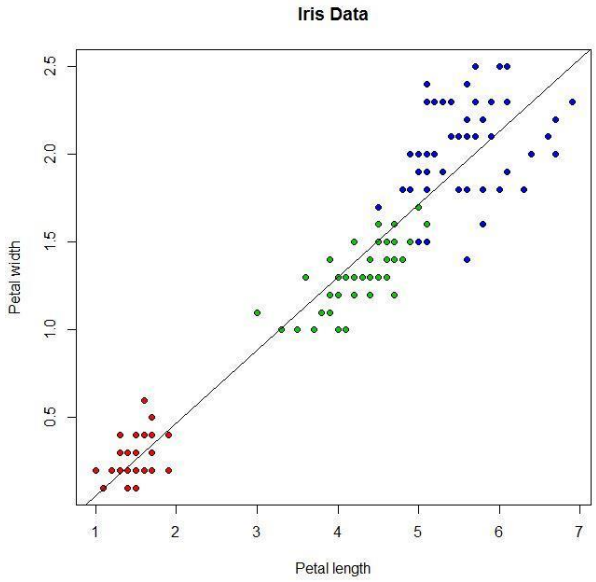
```
> prediction
  ir_ctrl.Sepal.Length ir_ctrl.Species predicted_val
1                  5.1         setosa   0.176005274
2                  4.7         setosa   0.031871367
3                  4.6         setosa   0.020210042
4                  5.0         setosa   0.118037011
5                  4.6         setosa   0.020210042
6                  4.3         setosa   0.005048194
7                  4.6         setosa   0.020210042
```



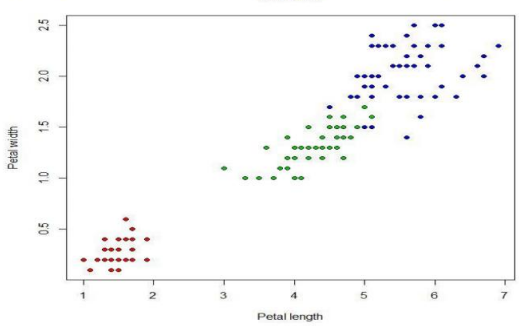
Practical 4

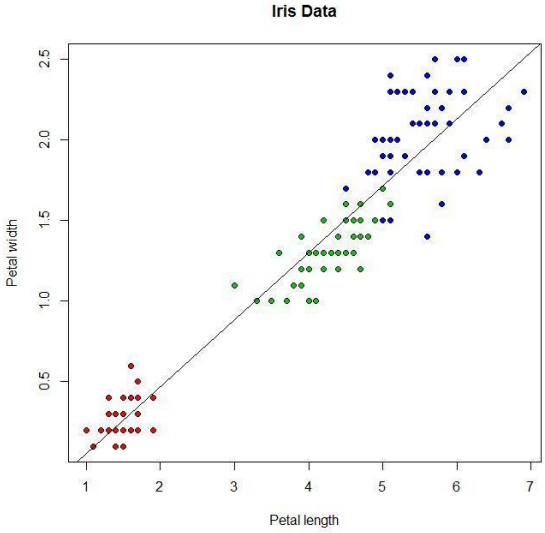
```
> lsfit(iris$Petal.Length, iris$Petal.Width)$coefficients
 Intercept          X
-0.3630755  0.4157554

> plot(iris$Petal.Length, iris$Petal.Width, pch=21, bg=c("red","green3","blue")[unclass(iris$
Species)], main="Iris Data", xlab="Petal length", ylab="Petal width")
```





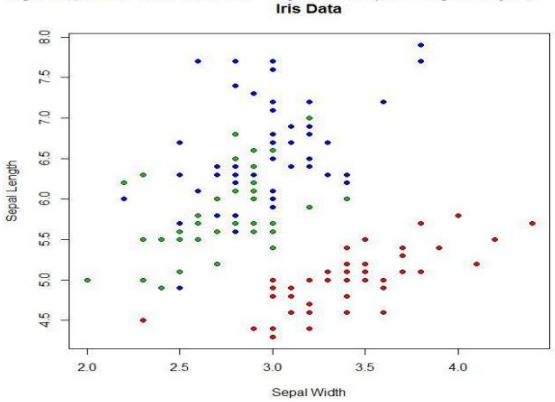

```
> lm(Petal.Width ~ Petal.Length, data=iris)$coefficients
 (Intercept) Petal.Length
  -0.3630755    0.4157554

> plot(iris$Petal.Length, iris$Petal.Width, pch=21, bg=c("red","green3","blue")[unclass(iris$
Species)], main="Iris Data", xlab="Petal length", ylab="Petal width")
```



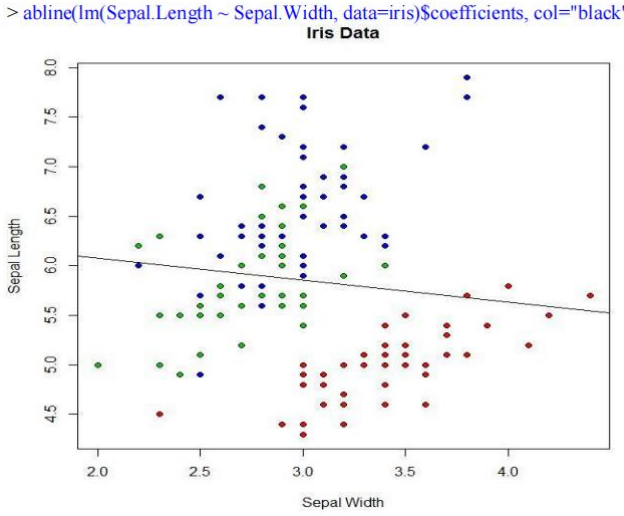



```
> summary(lm(Petal.Width ~ Petal.Length, data=iris))

Call:
lm(formula = Petal.Width ~ Petal.Length, data = iris)
Residuals:
     Min       1Q   Median       3Q      Max
-0.56515 -0.12358 -0.01898  0.13288  0.64272
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
---Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

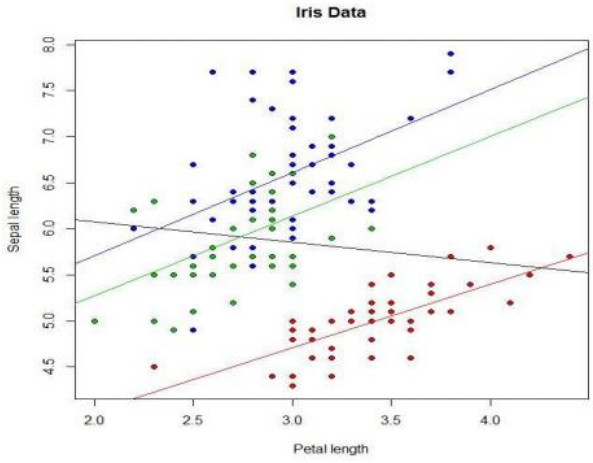
```
> plot(iris$Sepal.Width, iris$Sepal.Length, pch=21, bg=c("red","green3","blue")[unclass(iris
$Species)], main="Iris Data", xlab="Sepal Width", ylab="Sepal Length")
```



```
> abline(lm(Sepal.Length ~ Sepal.Width, data=iris)$coefficients, col="black")
```




```
> summary(lm(Sepal.Length ~ Sepal.Width, data=iris))

Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5561 -0.6333 -0.1120  0.5579  2.2226

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5262     0.4789   13.63   <2e-16 ***
Sepal.Width  -0.2234     0.1551   -1.44    0.152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8251 on 148 degrees of freedom
Multiple R-squared:  0.01382,    Adjusted R-squared:  0.007159
F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519
```



The coefficients doing separate per species regressions of Sepal.Length ~ Sepal.Width are:

```
> lm(Sepal.Length ~ Sepal.Width, data=iris[which(iris$Species=="setosa"),])$coefficients
(Intercept) Sepal.Width
  2.6390012   0.6904897
> lm(Sepal.Length ~ Sepal.Width, data=iris[which(iris$Species=="versicolor"),])$coefficient
s
(Intercept) Sepal.Width
  3.5397347   0.8650777
> lm(Sepal.Length ~ Sepal.Width, data=iris[which(iris$Species=="virginica"),])$coefficients
(Intercept) Sepal.Width
  3.9068365   0.9015345
```

The equivalent linear model would be something like Sepal.Length ~ Petal.Length:Species + Species- 1, which gives identical coef□cients (see later for why I did this):

```
> lm(Sepal.Length ~ Sepal.Width:Species + Species - 1, data=iris)$coefficients
          Speciessetosa      Speciesversicolor       Speciesvirginica
              2.6390012               3.5397347               3.9068365
Sepal.Width:Speciessetosa Sepal.Width:Speciesversicolor  Sepal.Width:Speciesvirginica
              0.6904897               0.8650777               0.9015345
```

```
> summary(lm(Sepal.Length ~ Sepal.Width:Species + Species - 1, data=iris))

Call:
lm(formula = Sepal.Length ~ Sepal.Width:Species + Species - 1,
    data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26067 -0.25861 -0.03305  0.18929  1.44917

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
Speciessetosa                 2.6390     0.5715   4.618 8.53e-06 ***
Speciesversicolor             3.5397     0.5580   6.343 2.74e-09 ***
Speciesvirginica              3.9068     0.5827   6.705 4.25e-10 ***
Sepal.Width:Speciessetosa     0.6905     0.1657   4.166 5.31e-05 ***
Sepal.Width:Speciesversicolor 0.8651     0.2002   4.321 2.88e-05 ***
Sepal.Width:Speciesvirginica  0.9015     0.1948   4.628 8.16e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4397 on 144 degrees of freedom
Multiple R-squared:  0.9947,    Adjusted R-squared:  0.9944
F-statistic:  4478 on 6 and 144 DF,  p-value: < 2.2e-16
```

```
> summary(step(lm(Sepal.Length ~ Sepal.Width * Species, data=iris)))
Start:  AIC=-240.59
Sepal.Length ~ Sepal.Width * Species

                     Df Sum of Sq    RSS     AIC
- Sepal.Width:Species  2   0.15719 28.004 -243.75
<none>                             27.846 -240.59

Step:  AIC=-243.74
Sepal.Length ~ Sepal.Width + Species
```

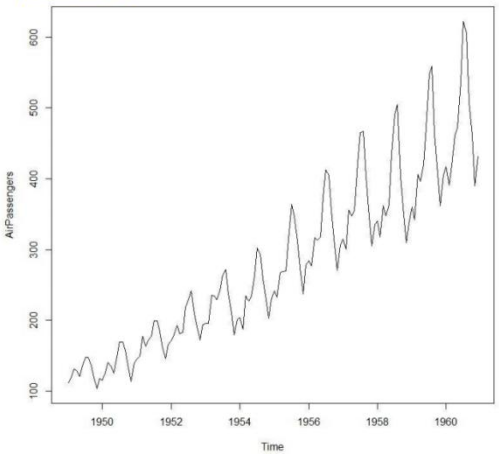
```
> lm(Sepal.Length ~ Sepal.Width:Species + Species - 1, data=iris)$coefficients
          Speciessetosa      Speciesversicolor       Speciesvirginica
              2.6390012               3.5397347               3.9068365
Sepal.Width:Speciessetosa Sepal.Width:Speciesversicolor  Sepal.Width:Speciesvirginica
              0.6904897               0.8650777               0.9015345
```

The use of the "- 1" in the model above told R not to automatically include a default intercept term. The alternative is the following:

```
> lm(Sepal.Length ~ Sepal.Width:Species + Species, data=iris)$coefficients
          (Intercept)      Speciesversicolor       Speciesvirginica
              2.6390012               0.9007335               1.2678352
Sepal.Width:Speciessetosa Sepal.Width:Speciesversicolor  Sepal.Width:Speciesvirginica
              0.6904897               0.8650777               0.9015345
```

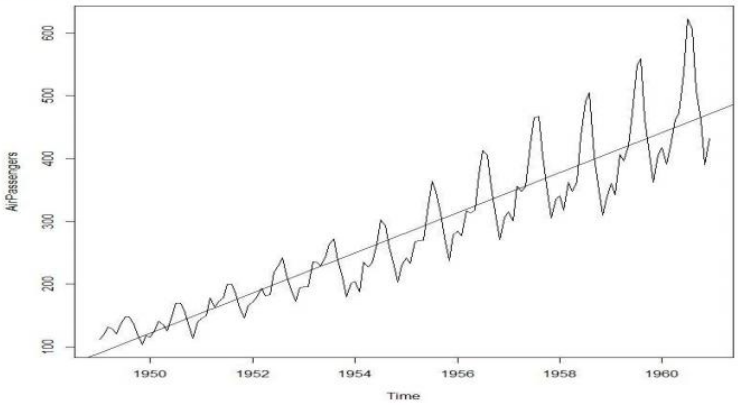
Practical 3

```
> plot(AirPassengers)
```



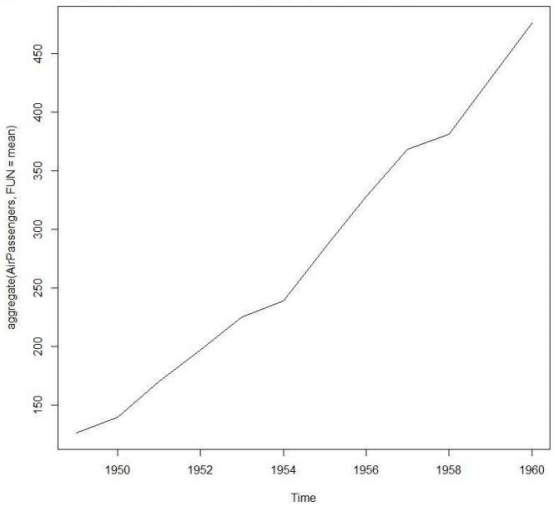

```
> abline(reg=lm(AirPassengers~time(AirPassengers)))
```

This will fit in a line




```
> plot(aggregate(AirPassengers,FUN=mean))
```

#This will aggregate the cycles and display a year on year trend

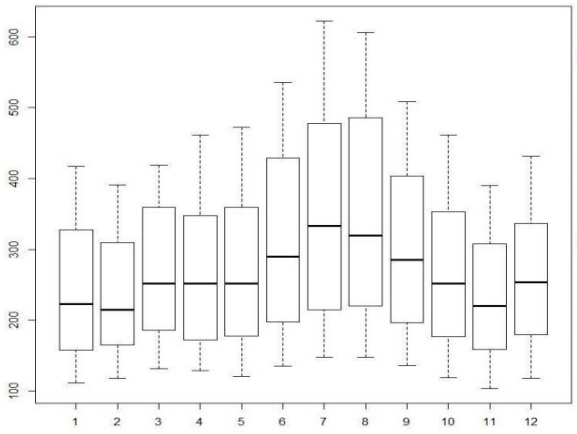


	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	1	2	3	4	5	6	7	8	9	10	11	12
1950	1	2	3	4	5	6	7	8	9	10	11	12
1951	1	2	3	4	5	6	7	8	9	10	11	12
1952	1	2	3	4	5	6	7	8	9	10	11	12
1953	1	2	3	4	5	6	7	8	9	10	11	12
1954	1	2	3	4	5	6	7	8	9	10	11	12
1955	1	2	3	4	5	6	7	8	9	10	11	12
1956	1	2	3	4	5	6	7	8	9	10	11	12
1957	1	2	3	4	5	6	7	8	9	10	11	12
1958	1	2	3	4	5	6	7	8	9	10	11	12
1959	1	2	3	4	5	6	7	8	9	10	11	12
1960	1	2	3	4	5	6	7	8	9	10	11	12

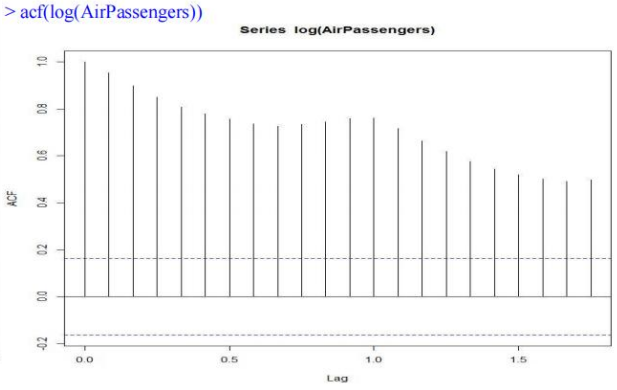
#This will print the cycle across years.

```
> boxplot(AirPassengers~cycle(AirPassengers))
```

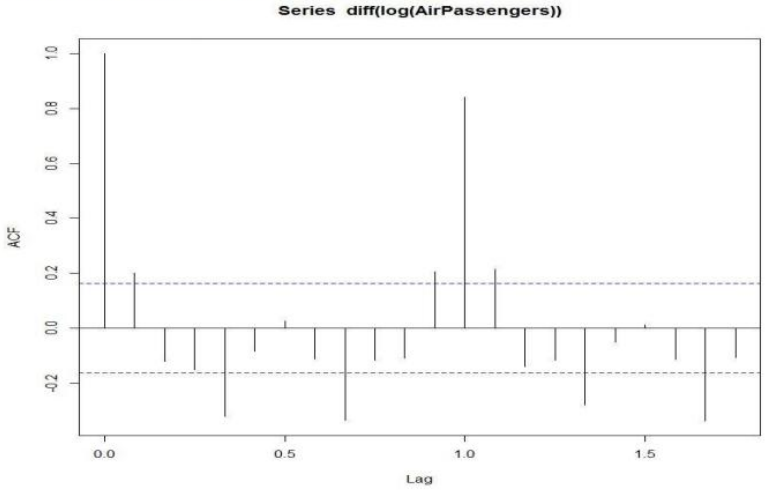
#Box plot across months will give us a sense on seasonal effect




```
> acf(log(AirPassengers))
```

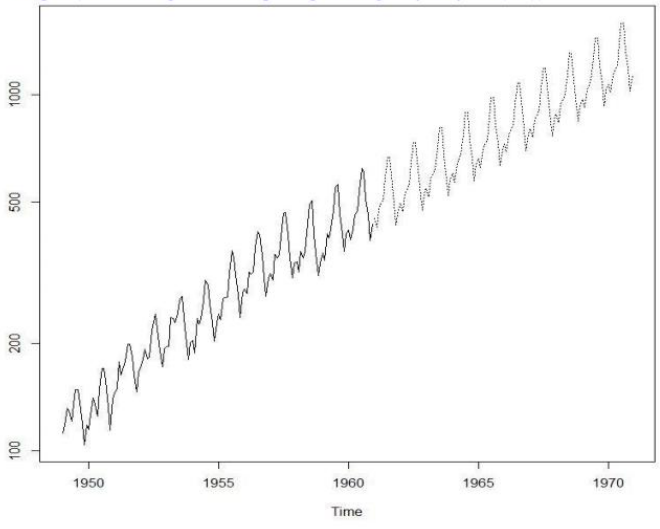


```
> acf(diff(log(AirPassengers)))
```



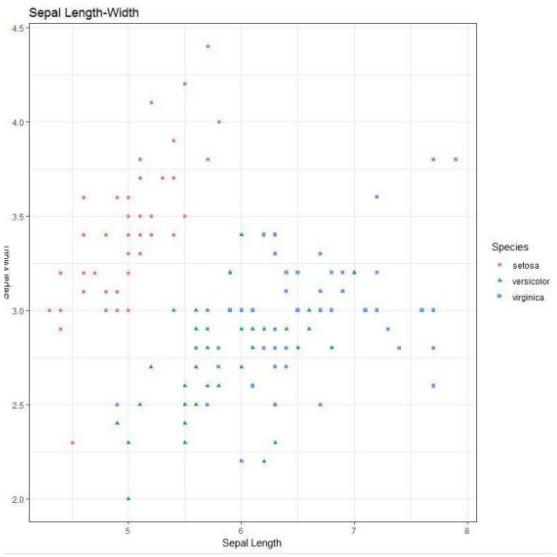

```
> (fit <- arima(log(AirPassengers), c(0, 1, 1),seasonal = list(order = c(0, 1, 1), period = 12)))

Call:
arima(x = log(AirPassengers), order = c(0, 1, 1), seasonal = list(order = c(0,
    1, 1), period = 12))

Coefficients:
          ma1     sma1
      -0.4018  -0.5569
s.e.   0.0896   0.0731

sigma^2 estimated as 0.001348:  log likelihood = 244.7,  aic = -483.4
```

```
> pred <- predict(fit, n.ahead = 10*12)
> ts.plot(AirPassengers,2.718^pred$pred, log = "y", lty = c(1,3))
```

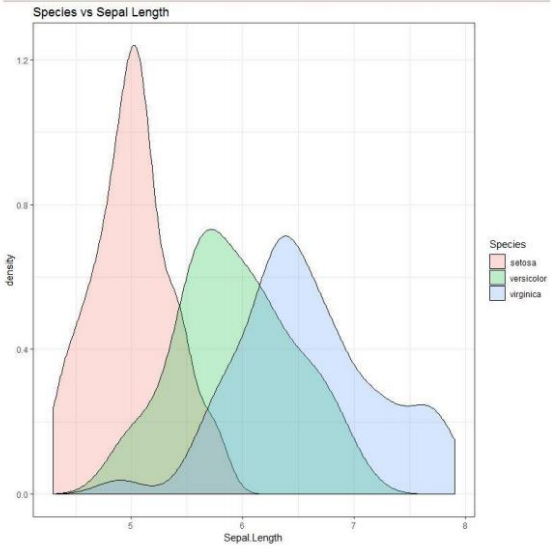


Practical 2

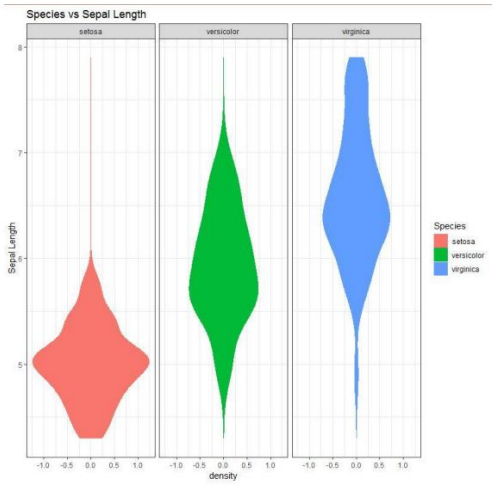
```
> scatter + geom_point(aes(color=Species, shape=Species)) +
+   theme_bw()+
+   xlab("Sepal Length") +  ylab("Sepal Width") +
+   ggtitle("Sepal Length-Width")
```



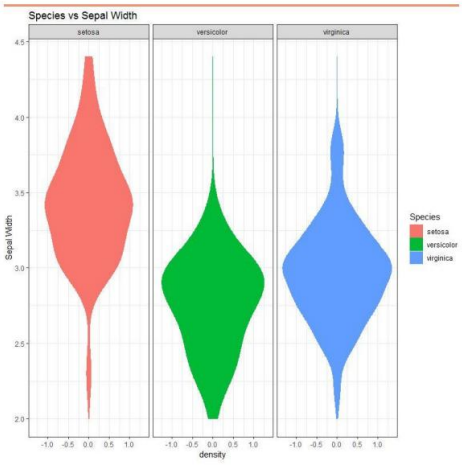

```
> ggplot(data=iris, aes(Sepal.Length, fill = Species))+
+   theme_bw()+
+   geom_density(alpha=0.25)+
+   labs(x = "Sepal.Length", title="Species vs Sepal Length")
```



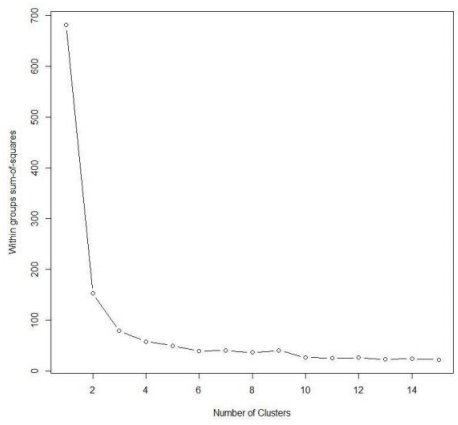

```
> vol + stat_density(aes(ymax = ..density.., ymin = -..density..,
+                        fill = Species, color = Species),
+                    geom = "ribbon", position = "identity") +
+   facet_grid(. ~ Species) + coord_flip() + theme_bw()+labs(x = "Sepal Length", title="Spec
ies vs Sepal Length")
```

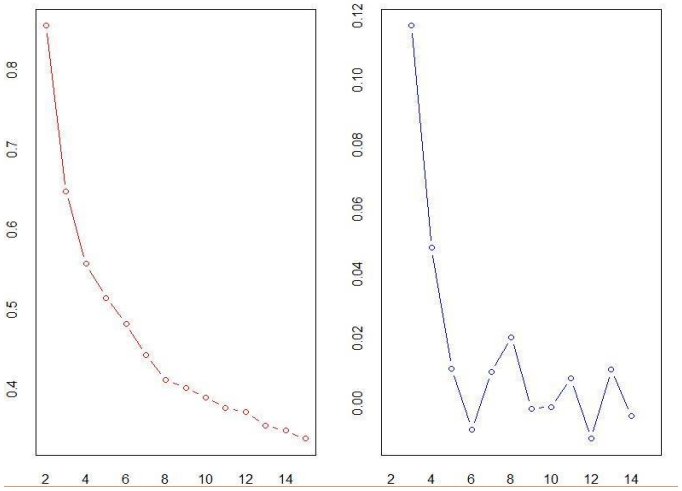



```
> vol <- ggplot(data=iris, aes(x = Sepal.Width))
> vol + stat_density(aes(ymax = ..density.., ymin = -..density..,
+                        fill = Species, color = Species),
+                    geom = "ribbon", position = "identity") +
+   facet_grid(. ~ Species) + coord_flip() + theme_bw()+labs(x = "Sepal Width", title="Speci
es vs Sepal Width")
```

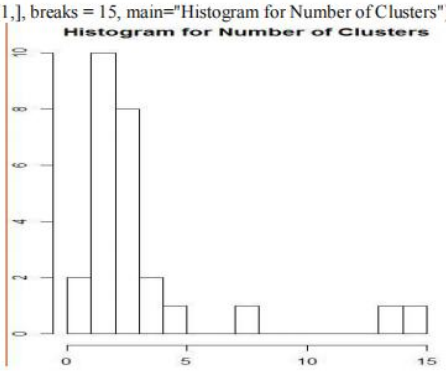



```
# Scree plot - Use plot function to plot values of tot_wss against no-of-clusters
> plot(x=1:15,                    # x= No of clusters, 1 to 15
+      y=totalwSS,                # tot_wss for each
+      type="b",                  # Draw both points as also connect them
+      xlab="Number of Clusters",
+      ylab="Within groups sum-of-squares")
```



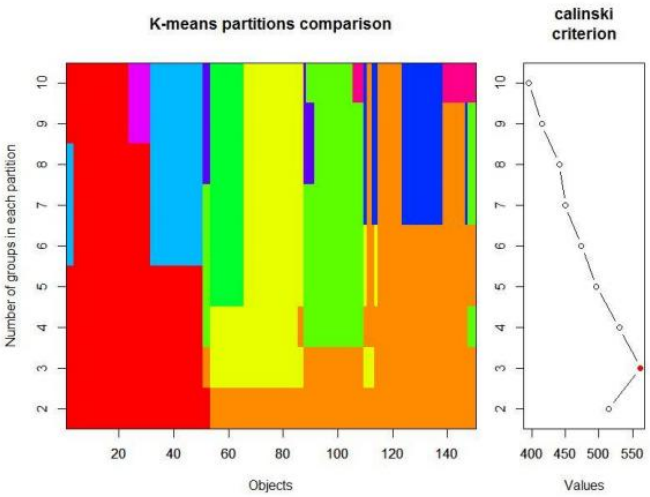


> hist(nb$Best.nc[1,], breaks = 15, main="Histogram for Number of Clusters")

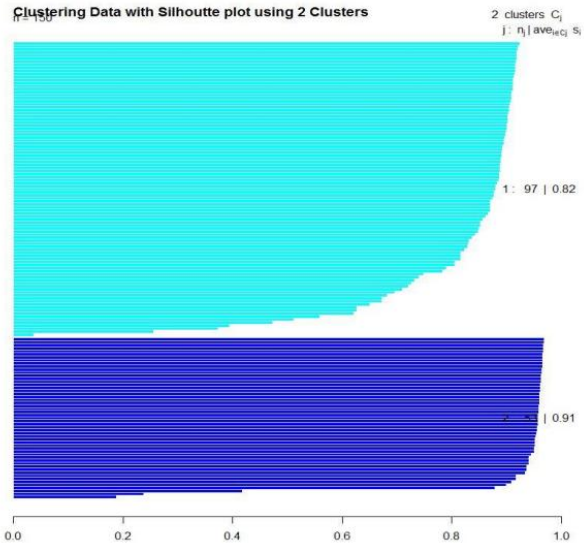


Assistant Professor-Sumit R. Mishra

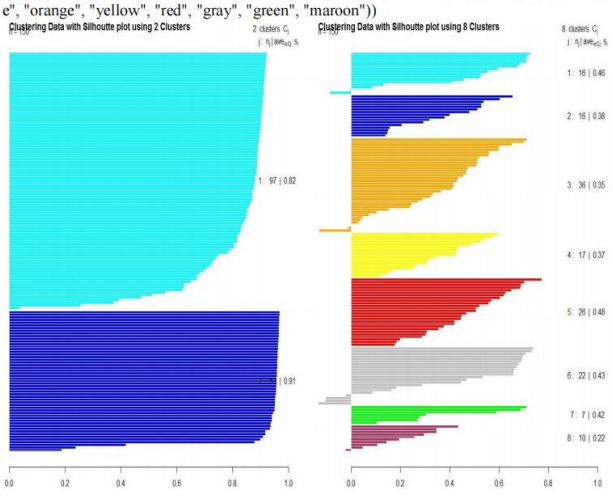
> plot(modelData, sortg = TRUE)



> plot(sil, main = "Clustering Data with Silhoutte plot using 2 Clusters", col = c("cyan", "blue"))

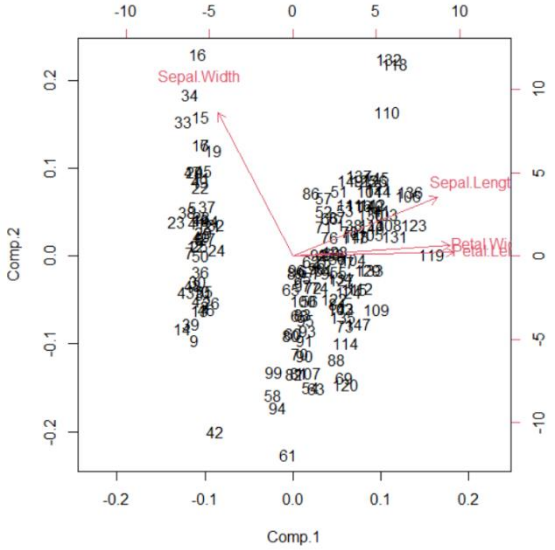


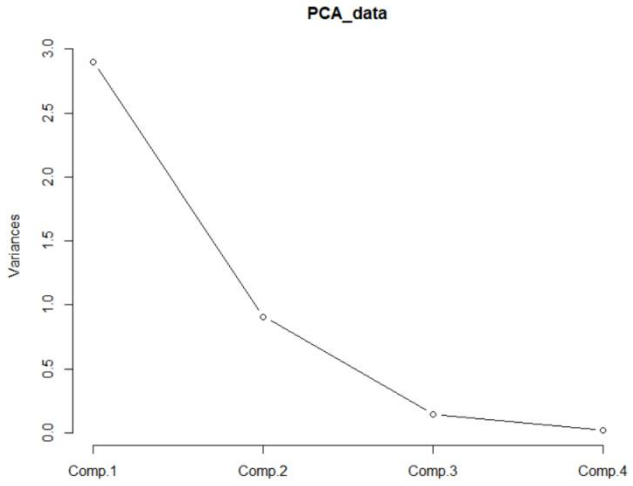
> plot(sil, main = "Clustering Data with Silhoutte plot using 8 Clusters", col = c("cyan", "blue", "orange", "yellow", "red", "gray", "green", "maroon"))



Practical 1

```
> Eigen_data$values
[1] 2.91849782 0.91403047 0.14675688 0.02071484
> PCA_data$sdev^2
    Comp.1     Comp.2     Comp.3     Comp.4
2.89904116 0.90793693 0.14577850 0.02057674
> PCA_data$loadings[,1:4]
                Comp.1     Comp.2     Comp.3     Comp.4
Sepal.Length  0.5210659 0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 0.06694199 -0.6342727  0.5235971
> Eigen_data$vectors
          [,1]        [,2]       [,3]       [,4]
[1,]  0.5210659 -0.37741762  0.7195664  0.2612863
[2,] -0.2693474 -0.92329566 -0.2443818 -0.1235096
[3,]  0.5804131 -0.02449161 -0.1421264 -0.8014492
[4,]  0.5648565 -0.06694199 -0.6342727  0.5235971
> summary(PCA_data)
Importance of components:
                          Comp.1    Comp.2     Comp.3      Comp.4
Standard deviation     1.7026571 0.9528572 0.38180950 0.143445939
Proportion of Variance 0.7296245 0.2285076 0.03668922 0.005178709
Cumulative Proportion  0.7296245 0.9581321 0.99482129 1.000000000
```






```
> mod2<-naiveBayes(model2_scores, iris[,5])
> # Accuracy for the first model
> table(predict(mod1, iris[,1:4]), iris[,5])

            setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         47         3
  virginica       0          3        47
> # Accuracy for the second model
> table(predict(mod2, model2_scores), iris[,5])

            setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         44         5
  virginica       0          6        45
```