Task-3: Iris Flower Classification

Kankana Ghosh

2023-07-21

Importing all the necessary libraries required for the analysis.

```
rm(list=ls())
library(MASS)
library(klaR)
```

```
## Warning: package 'klaR' was built under R version 4.2.3
```

```
library(clusterSim)
```

```
## Warning: package 'clusterSim' was built under R version 4.2.3

## Loading required package: cluster
```

```
library(GGally)
```

```
## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggpubr)
df=iris
class(df)
```

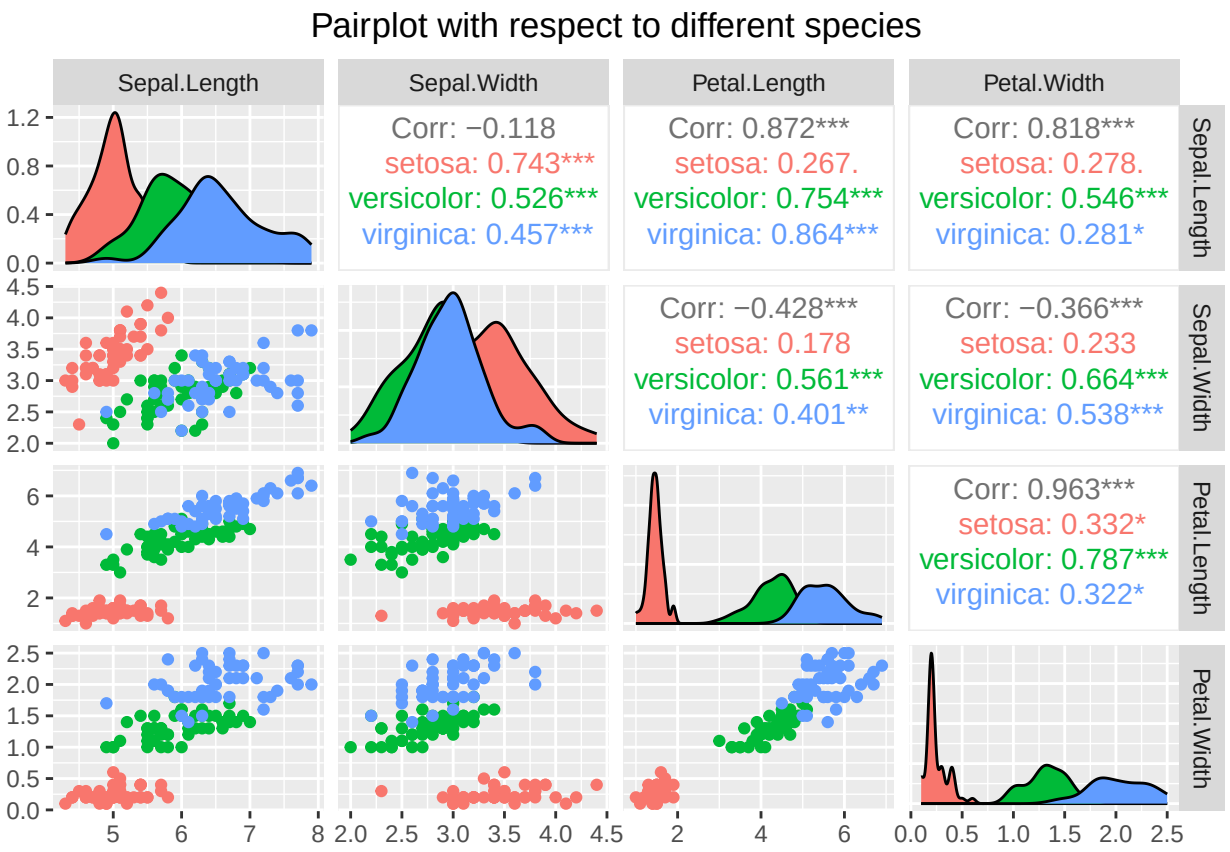
```
## [1] "data.frame"
```

```
str(df)
```

```
## '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 ...
```

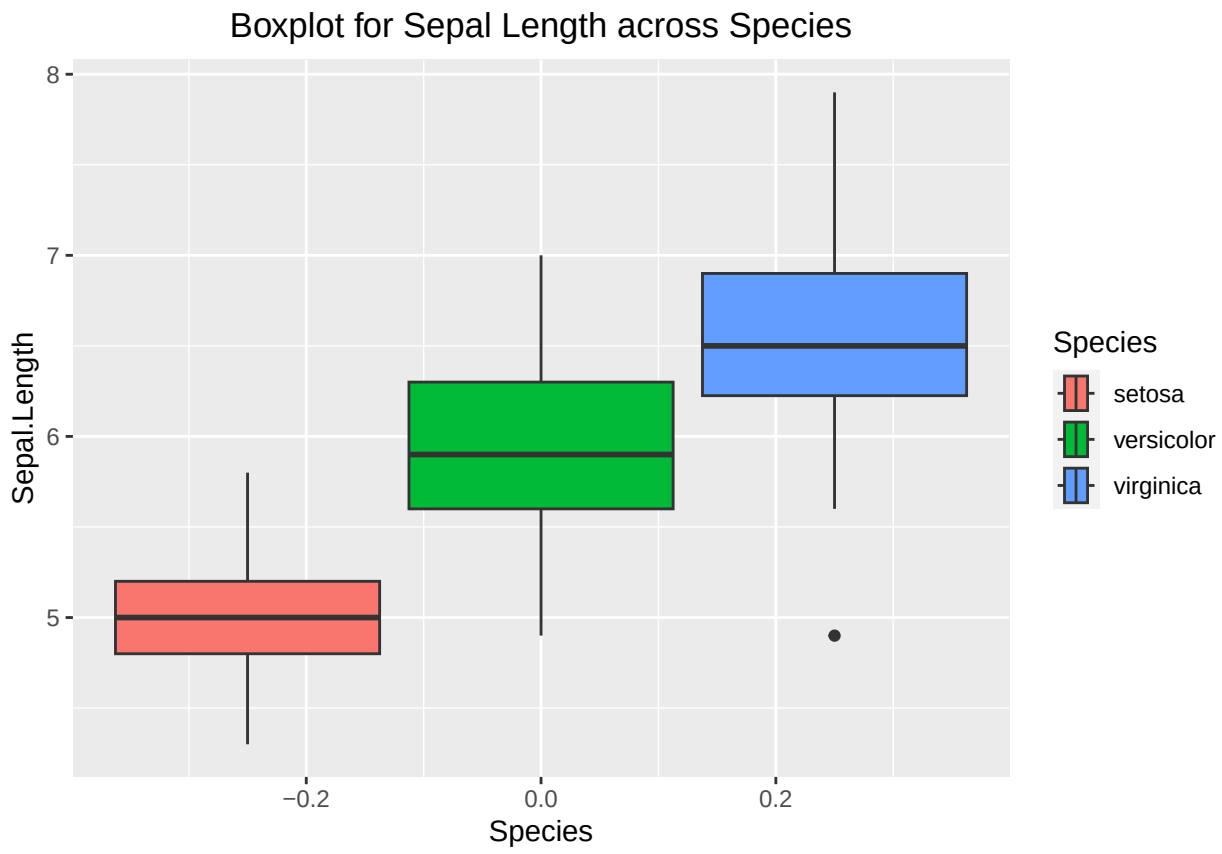
Comments: The dataset Iris is a dataframe in nature consisting 150 observations on 5 variables. The variable Species is categorical in nature having "Setosa", "versicolor" and "virginica" corresponding to which there are four variables taking measurements on it's Sepal Length, Sepal Width, Petal Length and Petal Width.

```
ggpairs(df,columns = 1:4,ggplot2::aes(colour=Species))+ggtitle("Pairplot with respect to different speci
  theme(plot.title = element_text(hjust=0.5))
```

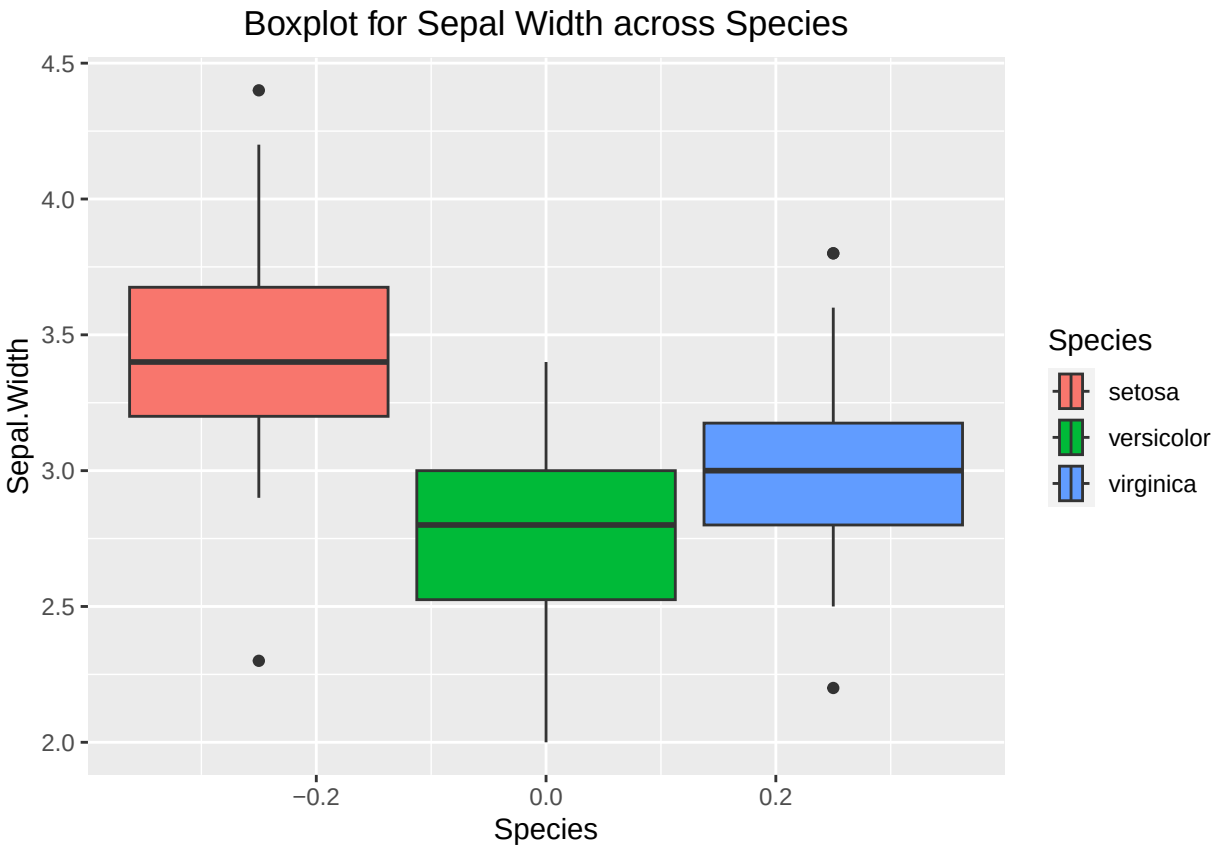


Comments: From the pairplot we get the idea of the correlation between different measurements of the 3 species.

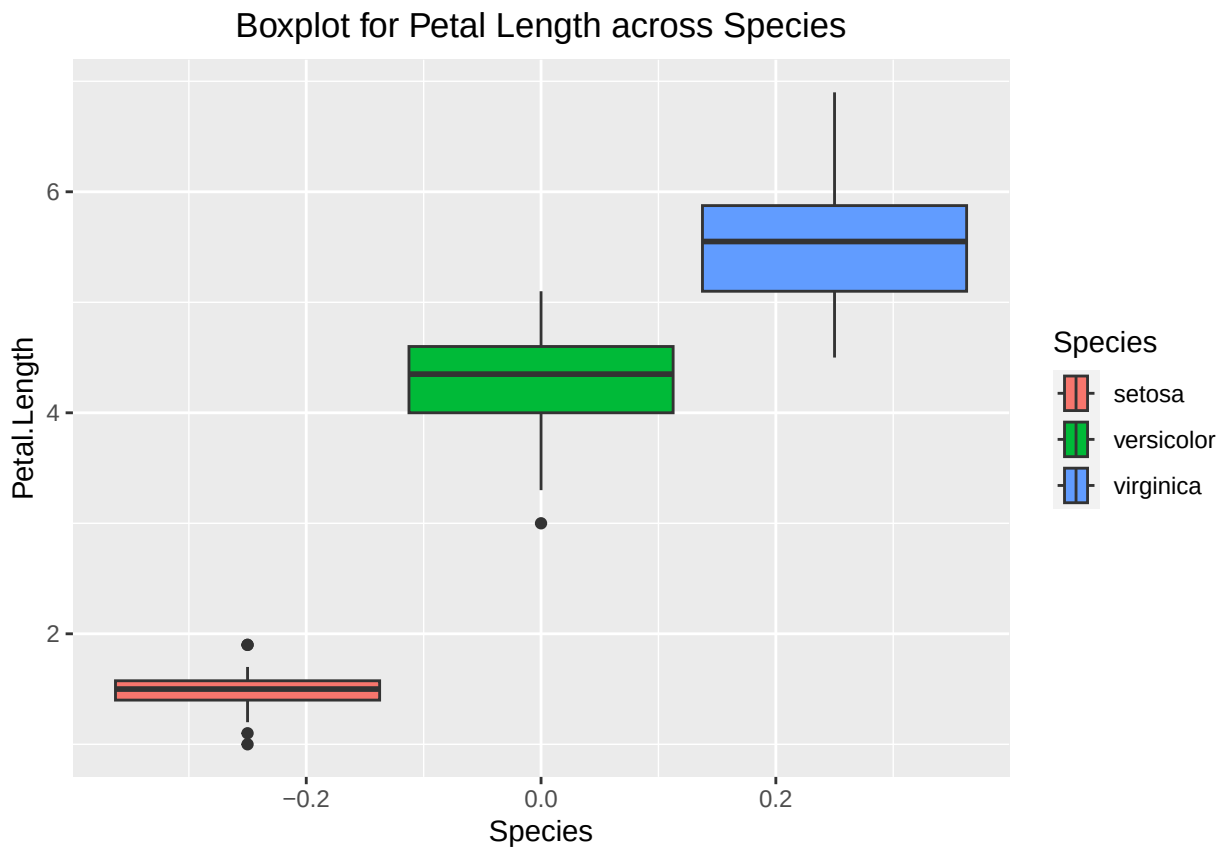
```
ggplot(data=df,aes(x=Sepal.Length,fill=factor(Species)))+geom_boxplot()+scale_fill_discrete(name="Speci
  coord_flip()+ylab("Species")+ggtitle("Boxplot for Sepal Length across Species")+
  theme(plot.title = element_text(hjust=0.5))
```



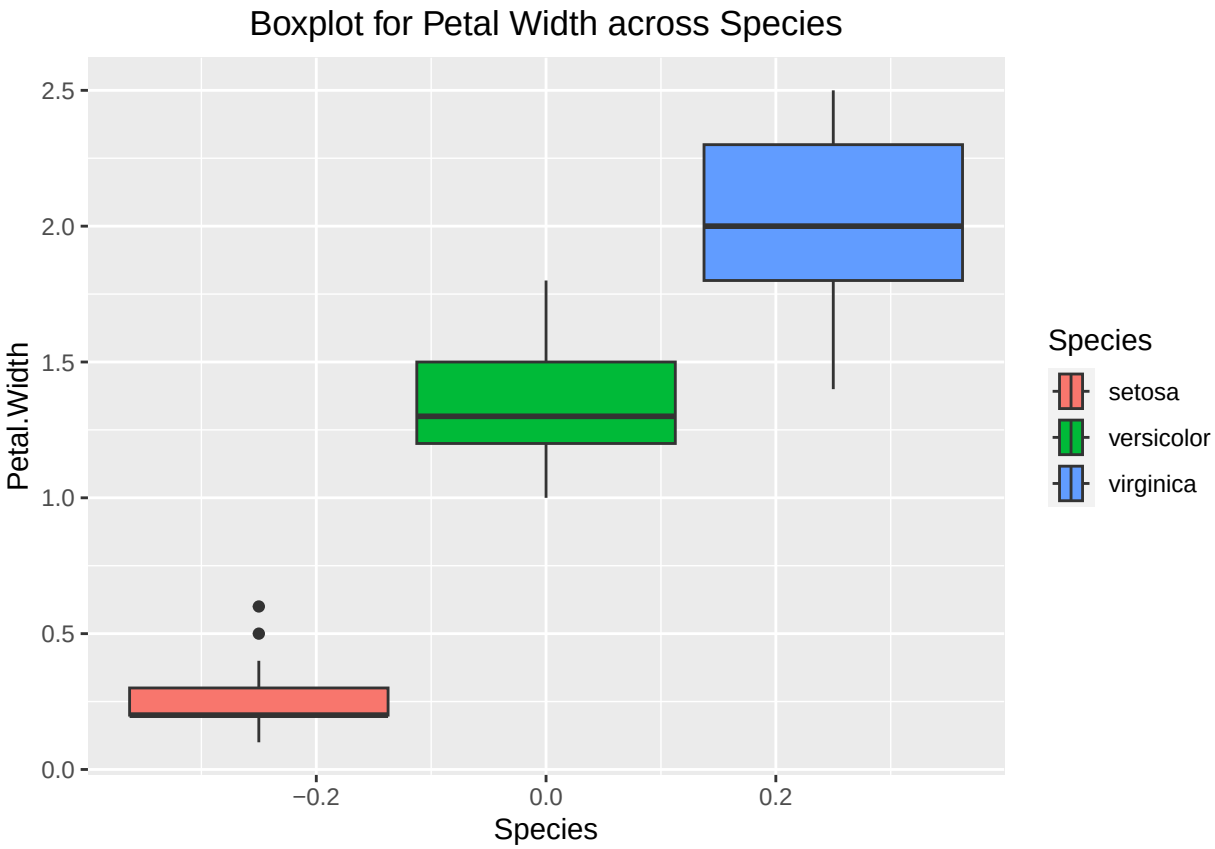

```
ggplot(data=df,aes(x=Sepal.Width,fill=factor(Species)))+geom_boxplot()+scale_fill_discrete(name="Species")+
  coord_flip()+ylab("Species")+ggtitle("Boxplot for Sepal Width across Species")+
  theme(plot.title = element_text(hjust=0.5))
```




```
ggplot(data=df,aes(x=Petal.Length,fill=factor(Species)))+geom_boxplot()+scale_fill_discrete(name="Speci
  coord_flip()+ylab("Species")+ggtitle("Boxplot for Petal Length across Species")+
  theme(plot.title = element_text(hjust=0.5))
```




```
ggplot(data=df,aes(x=Petal.Width,fill=factor(Species)))+geom_boxplot()+scale_fill_discrete(name="Species
  coord_flip()+ylab("Species")+ggtitle("Boxplot for Petal Width across Species")+
  theme(plot.title = element_text(hjust=0.5))
```



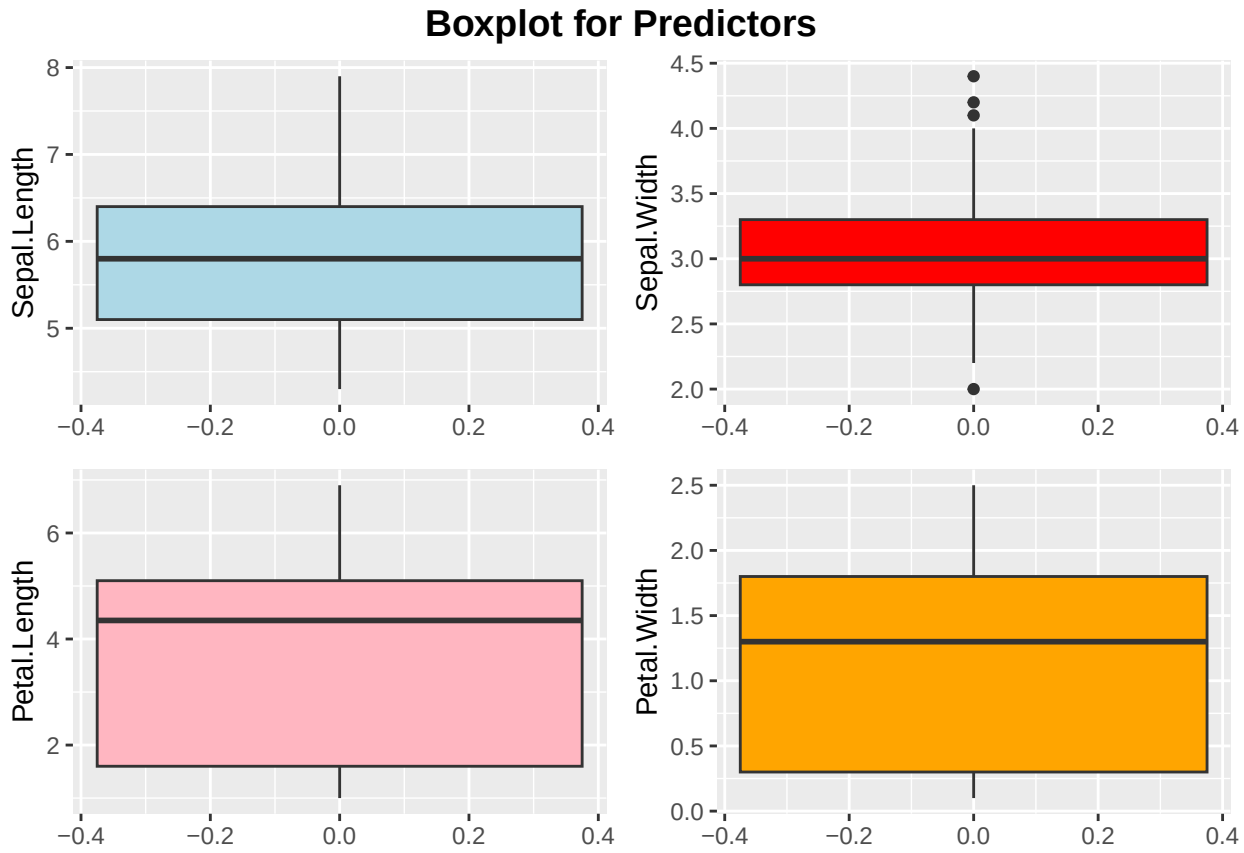
Comments: From the boxplot for Sepal Length across species it is clear that the Sepal Length of the specie virginica is more compared to the other two species.Virginica has outlier present in them.

From the boxplot for Sepal Width across species it is clear that the Sepal Width of the specie setosa is more compared to the other two species.Virginica and setosa has outlier present in them.

From the boxplot for Petal Length across species it is clear that the Petal Length of the specie virginica is more compared to the other two species.Versicolor and setosa has outlier present in them.

From the boxplot for Petal Width across species it is clear that the Petal Width of the specie virginica is more compared to the other two species.Setosa has outlier present in them.

```
a=ggplot(data=df,aes(x=Sepal.Length))+geom_boxplot(aes(Sepal.Length),fill="lightblue")+coord_flip()
b=ggplot(data=df,aes(x=Sepal.Width))+geom_boxplot(aes(Sepal.Width),fill="red")+coord_flip()
d=ggplot(data=df,aes(x=Petal.Length))+geom_boxplot(aes(Petal.Length),fill="lightpink")+coord_flip()
e=ggplot(data=df,aes(x=Petal.Width))+geom_boxplot(aes(Petal.Width),fill="orange")+coord_flip()
plot=ggarrange(a,b,d,e)
annotate_figure(plot,top=text_grob("Boxplot for Predictors",face="bold",size=14))
```




```
apply(df[,-5], 2,mean)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##     5.843333     3.057333     3.758000     1.199333
```

```
apply(df[,-5], 2,var)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.6856935    0.1899794    3.1162779    0.5810063
```

```
apply(df[,-5], 2,median)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##         5.80         3.00         4.35         1.30
```

Splitting the dataset into train and test data in the ratio of 80:20.

```
s=sample(c(TRUE,FALSE),nrow(df),replace=TRUE,c(0.8,0.2))
train=df[s,]
test=df[!s,]
model=lda(Species~.,data=train)
model
```

```
## Call:
## lda(Species ~ ., data = train)
##
## Prior probabilities of groups:
##     setosa versicolor  virginica
##  0.3362069  0.3362069  0.3275862
##
## Group means:
##            Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa         5.048718    3.494872     1.482051   0.2461538
## versicolor     5.825641    2.712821     4.207692   1.3051282
## virginica      6.589474    2.976316     5.518421   1.9894737
##
## Coefficients of linear discriminants:
##                     LD1        LD2
## Sepal.Length  0.5993432  0.4858063
## Sepal.Width   1.8912842  2.1512012
## Petal.Length -2.0885186 -0.7682193
## Petal.Width  -2.9318113  2.2088190
##
## Proportion of trace:
##    LD1    LD2
## 0.9888 0.0112
```

```
pred=predict(model,newdata =  test)
pred
```

```
## $class
##  [1] setosa     setosa     setosa     setosa     setosa     setosa
##  [7] setosa     setosa     setosa     setosa     setosa     versicolor
## [13] versicolor versicolor versicolor versicolor virginica  versicolor
## [19] versicolor versicolor versicolor versicolor virginica  virginica
## [25] virginica  virginica  virginica  virginica  virginica  virginica
## [31] virginica  virginica  virginica  virginica
## Levels: setosa versicolor virginica
##
## $posterior
##           setosa   versicolor    virginica
## 4   1.000000e+00 2.047868e-17 1.064074e-35
## 6   1.000000e+00 9.915839e-23 3.874540e-41
## 9   1.000000e+00 5.638856e-16 4.575321e-34
## 10  1.000000e+00 2.141038e-19 1.664058e-38
## 23  1.000000e+00 1.448959e-26 3.295230e-47
## 29  1.000000e+00 1.012175e-22 3.072896e-42
## 32  1.000000e+00 4.677872e-20 3.570922e-38
## 36  1.000000e+00 3.310750e-22 7.949707e-42
## 42  1.000000e+00 1.332743e-10 2.234045e-27
## 46  1.000000e+00 7.638000e-17 9.202614e-35
## 48  1.000000e+00 3.028676e-19 5.310734e-38
## 51  7.611917e-20 9.993618e-01 6.382191e-04
## 57  3.451182e-23 9.782351e-01 2.176491e-02
## 59  1.345560e-21 9.995096e-01 4.904378e-04
## 66  8.305014e-19 9.998006e-01 1.993792e-04
## 76  5.454078e-20 9.996500e-01 3.500330e-04
```

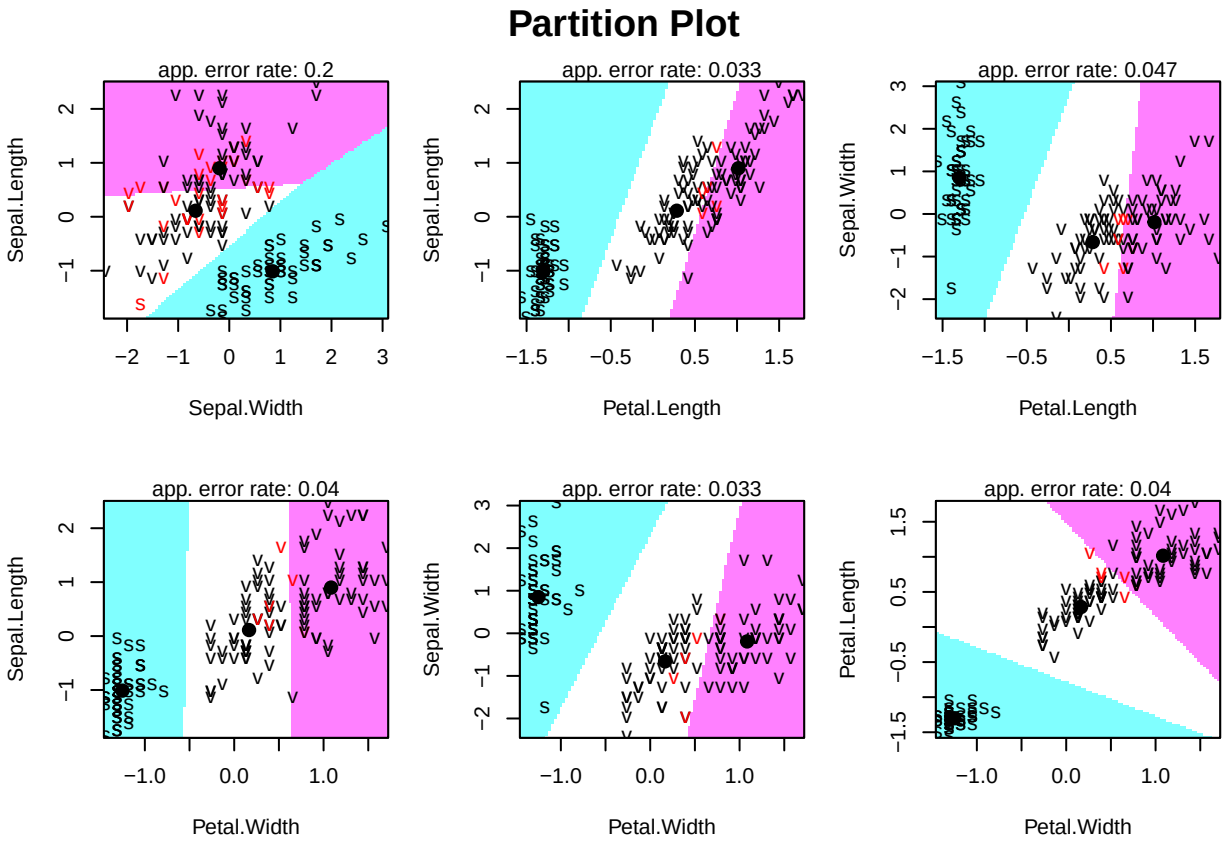
```
## 78  4.580187e-29 4.366423e-01 5.633577e-01
## 79  1.225128e-24 9.890144e-01 1.098557e-02
## 80  3.428066e-13 9.999999e-01 5.266430e-08
## 87  9.851189e-23 9.936329e-01 6.367051e-03
## 95  1.749642e-22 9.996507e-01 3.492697e-04
## 97  2.140070e-20 9.998674e-01 1.325625e-04
## 103 1.718318e-45 1.048235e-05 9.999895e-01
## 104 2.253639e-40 1.113440e-03 9.988866e-01
## 106 2.630086e-52 1.964442e-07 9.999998e-01
## 107 2.038843e-34 1.533546e-01 8.466454e-01
## 110 3.581246e-48 1.183468e-07 9.999999e-01
## 112 2.692839e-40 1.066075e-03 9.989339e-01
## 115 1.088976e-47 1.919613e-06 9.999981e-01
## 116 5.805412e-42 2.674824e-05 9.999733e-01
## 119 1.749380e-63 3.376208e-10 1.000000e+00
## 129 2.058535e-46 1.193930e-05 9.999881e-01
## 137 1.413180e-45 1.602819e-06 9.999984e-01
## 140 5.340804e-39 3.447349e-04 9.996553e-01
##
## $x
##             LD1         LD2
## 4    6.83515585 -0.95230333
## 6    7.82359179  1.44542263
## 9    6.54588223 -1.40288290
## 10   7.30813995 -1.02744334
## 23   8.82505724  0.50740696
## 29   7.97099890  0.06136275
## 32   7.29565342  0.52346588
## 36   7.89057714 -0.31239488
## 42   5.38671678 -2.34731917
## 46   6.68156681 -0.77255835
## 48   7.23313613 -0.66036127
## 51  -1.73872510  0.62103283
## 57  -2.55549919  0.93785235
## 59  -2.04381465 -0.36271003
## 66  -1.48110090  0.49063662
## 76  -1.73016364  0.22693587
## 78  -3.80288387  0.47723060
## 79  -2.78093098 -0.13560807
## 80   0.02630506 -1.26290049
## 87  -2.40083761  0.48105272
## 95  -2.18600726 -0.97146883
## 97  -1.74781610 -0.49264796
## 103 -6.61553785  0.86368331
## 104 -5.77804186 -0.17226175
## 106 -7.77782926  0.56883292
## 107 -4.78308445 -1.08871168
## 110 -7.01126126  2.93286842
## 112 -5.76298993 -0.10257366
## 115 -6.98166938  1.07911606
## 116 -5.99007236  1.85655456
## 119 -9.68732645 -0.03176894
## 129 -6.78677935  0.32384446
## 137 -6.59148656  2.22863028
```

```
## 140 -5.50201877  1.36575185
```

```
mean(pred$class==test$Species)
```

```
## [1] 0.9705882
```

```
df.standard=data.Normalization(df[,-5],type="n1")
Species=df[,5]
df1=cbind(Species,df.standard)
partimat(Species~.,data = df1,method="lda")
```



Comments: Now to classify iris flowers into different species based on their sepal and petal measurements I have used the LDA model. Fitting the model on the train dataset I got the prior probabilities as 0.35,0.34,0.30 for setosa, versicolor and virginica respectively.It means that the chance to classify a specie into setosa, versicolor and virginica is the above mentioned probabilities.Here, we get two discriminants hence the linear combinations of the predictors (i.e, the four measurements) that are used to form the LDA decision rule are:

LD1=0.70* Sepal.Length + 1.40* Sepal.Width - 2.25* Petal.Length - 2.39* Petal.Width LD2=-0.54* Sepal.Length + 2.19* Sepal.Width - 0.72* Petal.Length + 2.88* Petal.Width

The accuracy of this model is coming out to be 1, hence the above model is a perfect model, i.e., the above decision rules can correctly classify a specie based on the different measurements.

From the partition plot we get an idea of the classification of the species into three regions. In each case we get the approximate error rate which indicates the incorrect classification.