Lecture 2: Exploratory Data Analysis

Xiao Guo

2023/2/21

## 2.1. Exploratory Data Analysis

### Roles of Data Visualization

* Role 1: Exploratory data analysis (pre stage);
* Role 2: Visual presentation of results (after stage).
* John W. Tukey (1977; Exploratory Data Analysis): “The greatest value of a picture is when it forces us to notice what we never expected to see.”

knitr::include\_graphics("tukey.png")



### John Tukey (1915-2000)

Proposed “Exploratory Data Analysis”

* Coined terms: Boxplot, Stem-and-Leaf plot, ANOVA (Analysis of Variance)
* Coined terms “Bit” and “Software”
* Co-Developed Fast Fourier Transform algorithm, Projection Pursuit, Jackknife estimation
* Famous quote: “The best thing about being a statistician is that you get to play in everyone’s backyard.”
* <https://en.wikipedia.org/wiki/John_Tukey>

### John Tukey: Exploratory Data Analysis (1977)

* Five-number summary
* Stem-and-Leaf plot
* Scatter plot
* Box-plot, Outliers
* Residual plot
* Smoother
* Bag plot (two or three dimensional ‘box’ plot)

## 2.2. Simple Base Graphics

### Iris Dataset

DataX = iris   
str(DataX)

## '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 ...

#DataX$Species  
dim(DataX)

## [1] 150 5

head(DataX)

## 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

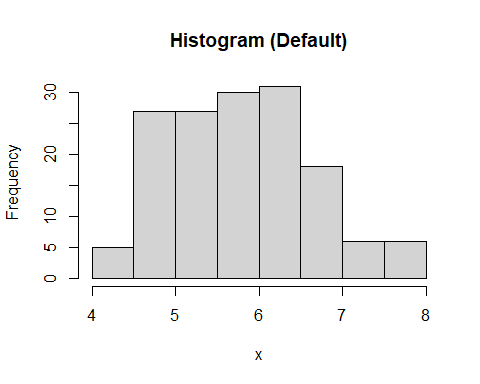
summary(DataX)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## 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   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

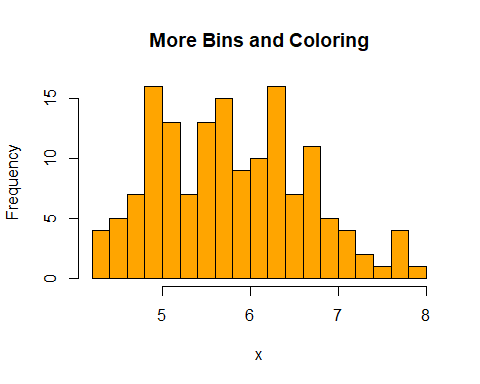
### Basic R Plots

#### Histogram

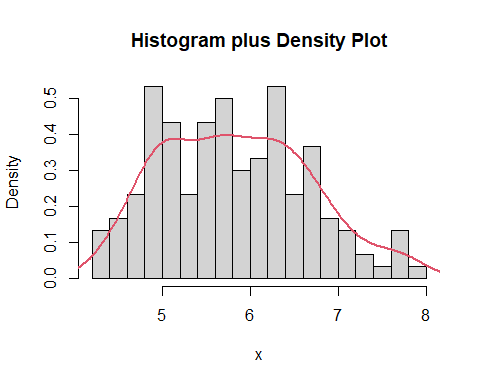
x = DataX[,1]  
hist(x, main='Histogram (Default)')



hist(x, breaks=20, col="orange", main='More Bins and Coloring')

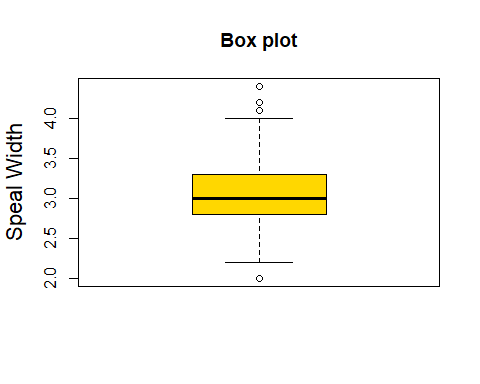


hist(x, breaks=20, freq=F, main='Histogram plus Density Plot') # using freq=FALSE  
lines(density(x), col=2, lty=1, lwd=2) #add the density curve

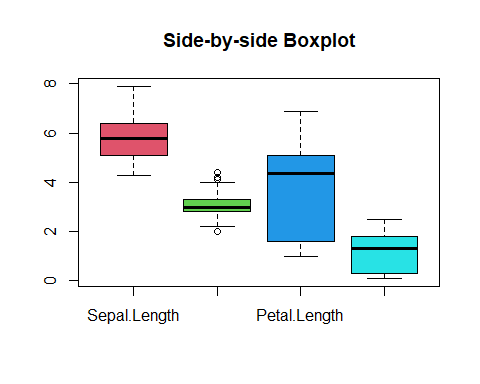


#### Box plot

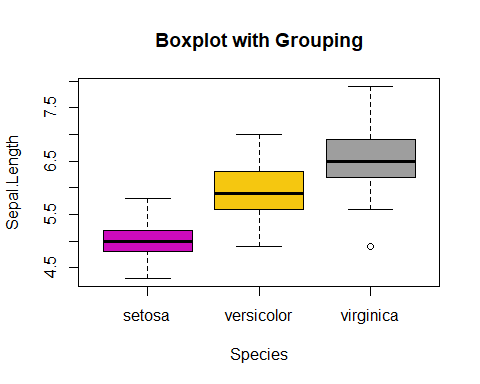
x = DataX[,2]  
boxplot(x,main="Box plot",col="gold")  
mtext("Speal Width", side = 2, line = 2.8,cex=1.4)



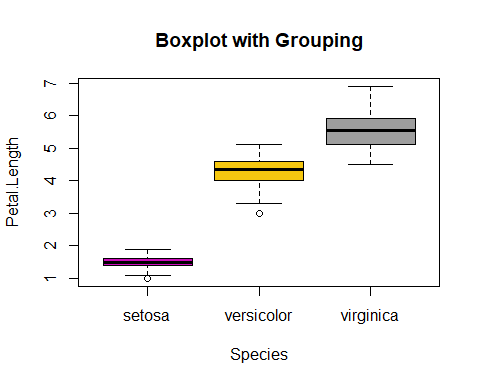
boxplot(DataX[,1:4], col=c(2,3,4,5), main='Side-by-side Boxplot')



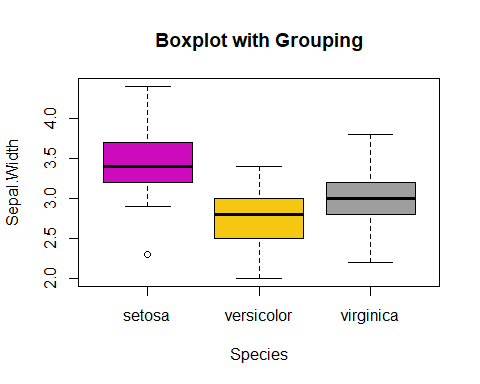
boxplot(Sepal.Length~Species, DataX, col=c(6,7,8), main="Boxplot with Grouping")



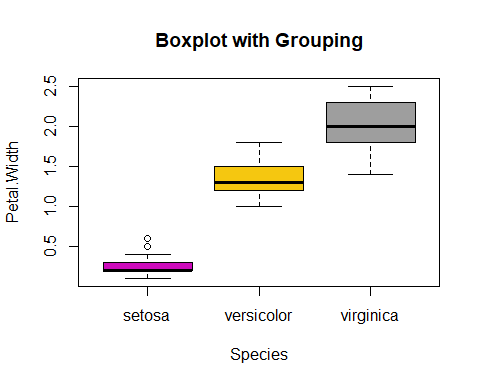
boxplot(Petal.Length~Species, DataX, col=c(6,7,8), main="Boxplot with Grouping")



boxplot(Sepal.Width~Species, DataX, col=c(6,7,8), main="Boxplot with Grouping")

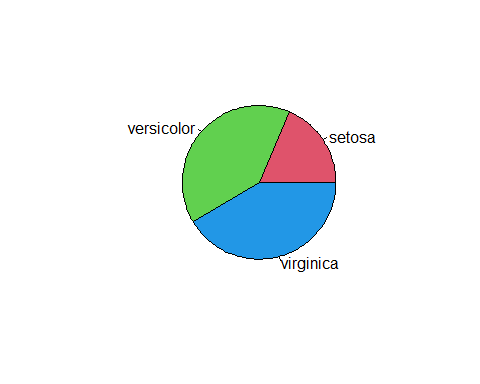


boxplot(Petal.Width~Species, DataX, col=c(6,7,8), main="Boxplot with Grouping")

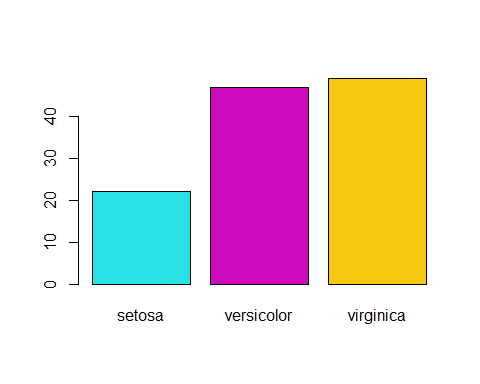


#### Pie and Bar Charts

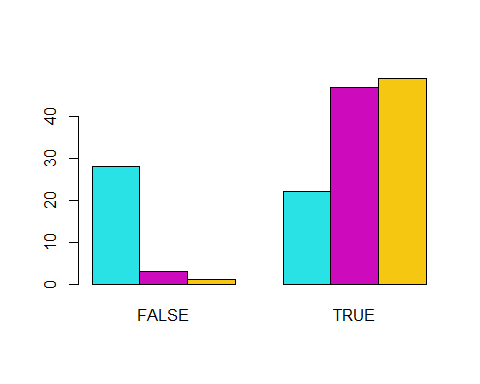
DataX$Flag = DataX$Sepal.Length>5 # Create a binary flag  
pie(table(DataX$Species[DataX$Flag]), col=c(2,3,4))



barplot(table(DataX$Species[DataX$Flag]), col=c(5,6,7))

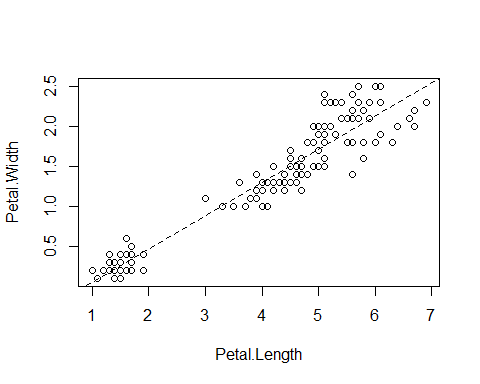


barplot(table(DataX$Species, DataX$Flag), col=c(5,6,7), beside=T)

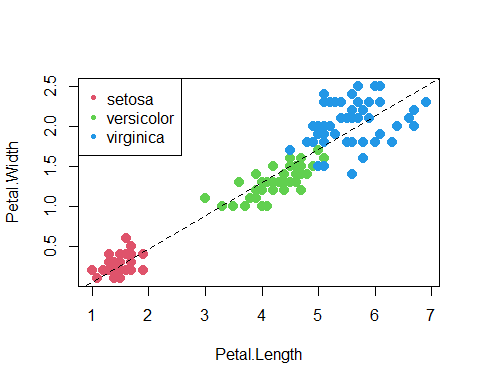


#### Relationship Between Variables

x = DataX$Petal.Length; y = DataX$Petal.Width; z = DataX$Species  
plot(x, y, xlab="Petal.Length", ylab="Petal.Width")   
abline(coef(lm(y~x)), col=1, lty=2)

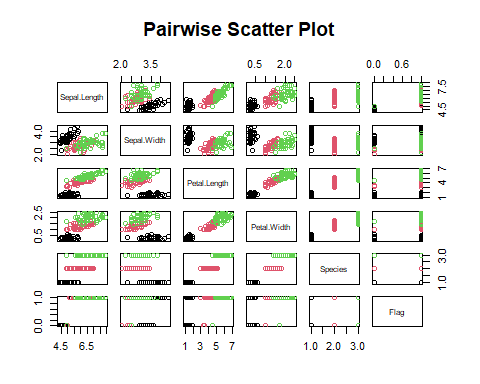


plot(x, y, col=c(2,3,4)[z], pch=20, cex=2.0, xlab="Petal.Length", ylab="Petal.Width")   
abline(lm(y~x), col=1, lty=2)   
legend("topleft", levels(z), pch=20, col=c(2,3,4))

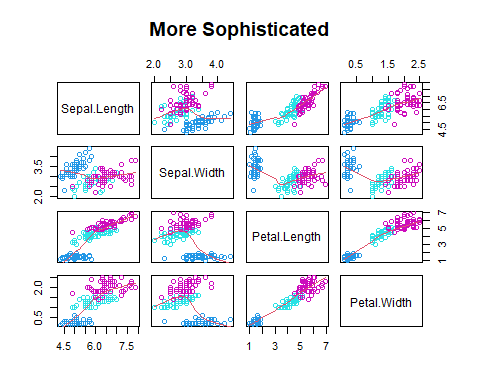


#### Pairwise Scatter Plot

plot(DataX, col=DataX$Species, main="Pairwise Scatter Plot")



pairs(DataX[,1:4], panel = panel.smooth, col = c(4,5,6)[DataX$Species], main="More Sophisticated")



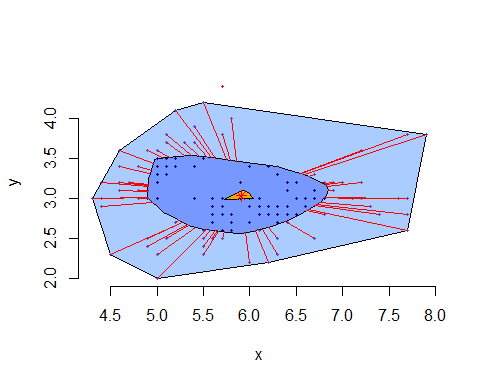
#### Bag Plot

* The bagplot consists of three nested polygons, called the “bag”, the “fence”, and the “loop”.
* The inner polygon, called the bag, is constructed on the basis of Tukey depth, the smallest number of observations that can be contained by a half-plane that also contains a given point.[4] It contains at most 50% of the data points.
* The outermost of the three polygons, called the fence is not drawn as part of the bagplot, but is used to construct it. It is formed by inflating the bag by a certain factor (usually 3). Observations outside the fence are flagged as outliers.
* The observations that are not marked as outliers are surrounded by a loop, the convex hull of the observations within the fence.

library(aplpack)

## Warning: package 'aplpack' was built under R version 4.0.5

bagplot(DataX[, 1:2])



#### Violin Plot

A violin plot is a hybrid of a box plot and a kernel density plot, which shows peaks in the data. It is used to visualize the distribution of numerical data. Unlike a box plot that can only show summary statistics, violin plots depict summary statistics and the density of each variable.

library(vioplot)

## Loading required package: sm

## Package 'sm', version 2.2-5.7: type help(sm) for summary information

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

vioplot(DataX[,1], DataX[,2],col=4)

