R：Statistical Programming Methods
R：程式、機率與統計

Data Visualization (1)

Exploring the data distribution

- A grouping of data into mutually exclusive classes showing the number of observations
  - Frequency Table
  - Percentiles and Box plot
  - Histogram
  - Density Plot
    - As opposed to the histogram, the density plot can smooth out the distribution of values and reduce the noise

Example

```
student <- read.csv("behavior.csv", header=TRUE)

#show how many male and female student
table(student$gender)
##
##  F  M
## 42 58
#in proporation
prop.table(table(student$gender))
##
##    F    M
## 0.42 0.58
```

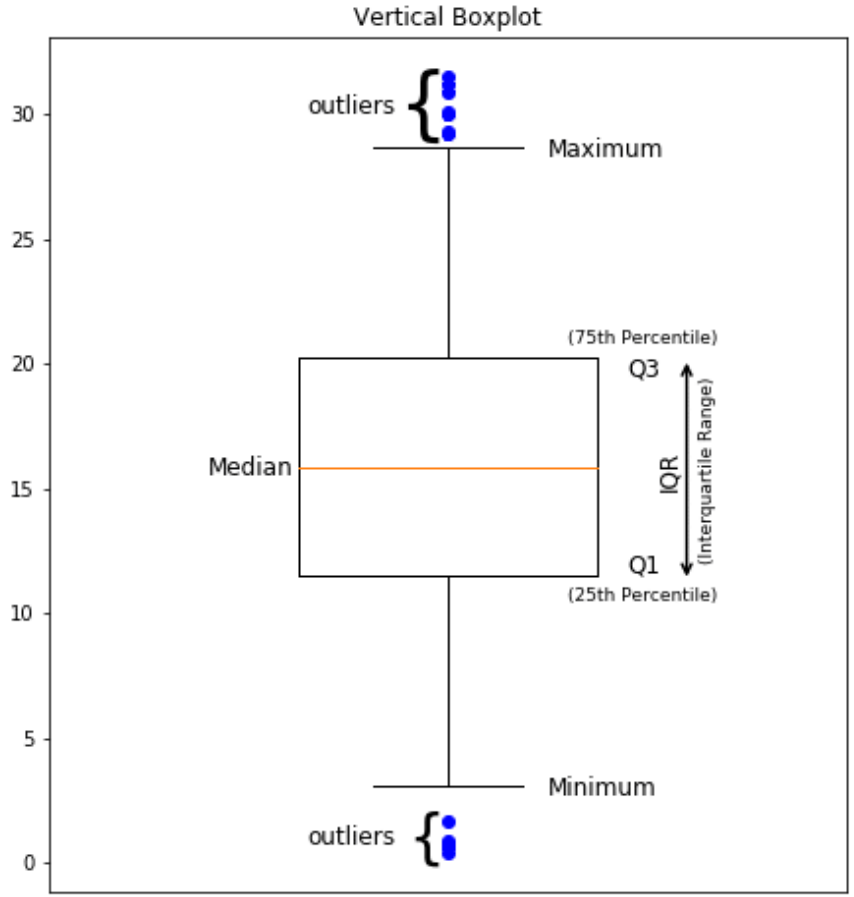
```
#in each class
table(student$class, student$gender)
```

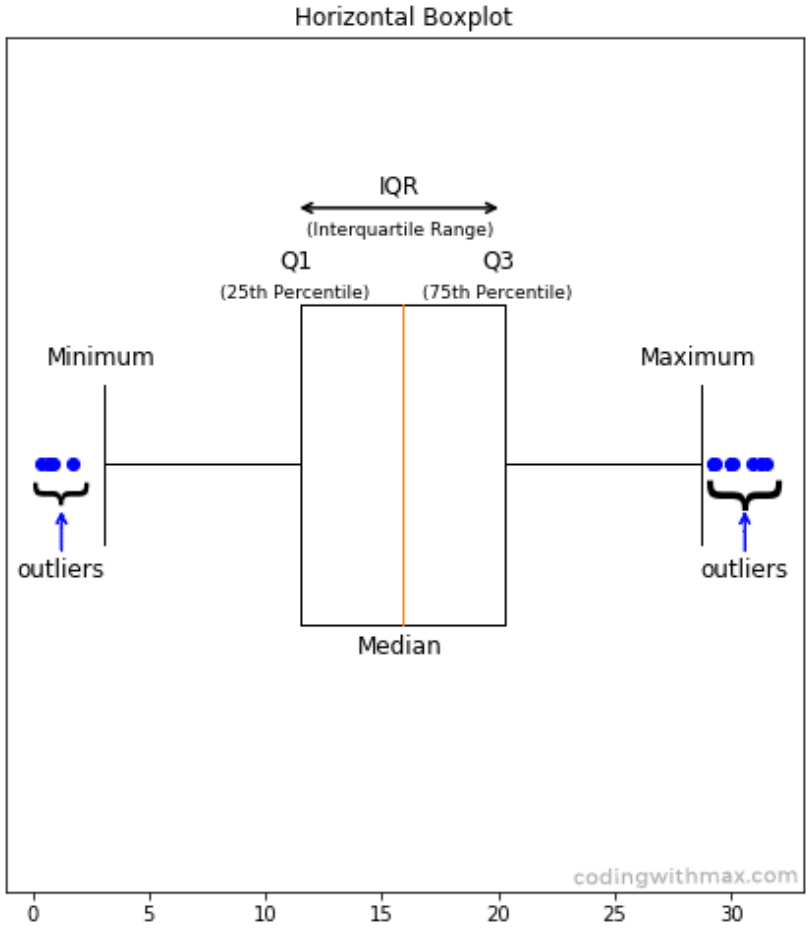
```
##
##       F  M
##    A  7 10
##    B  8  8
##    C  5 13
##    D  6  7
##    E 11  9
##    F  5 11
```

```
prop.table(table(student$class, student$gender))
```

```
##
##          F    M
##    A 0.07 0.10
##    B 0.08 0.08
##    C 0.05 0.13
##    D 0.06 0.07
##    E 0.11 0.09
##    F 0.05 0.11
```

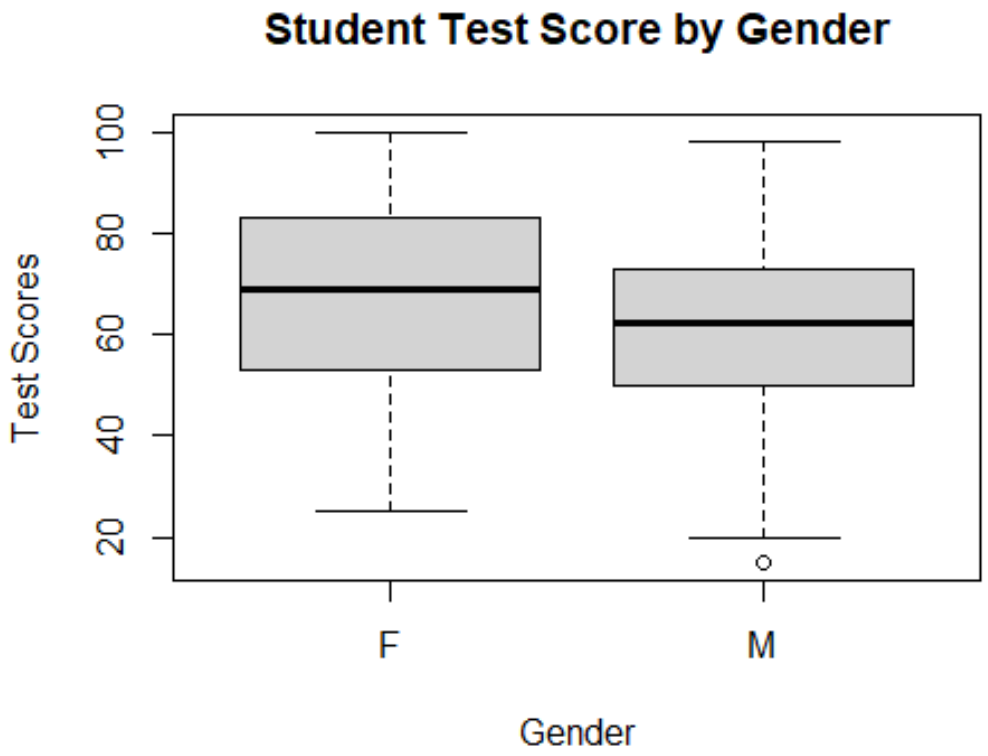
Boxplot 箱型圖/盒鬚圖





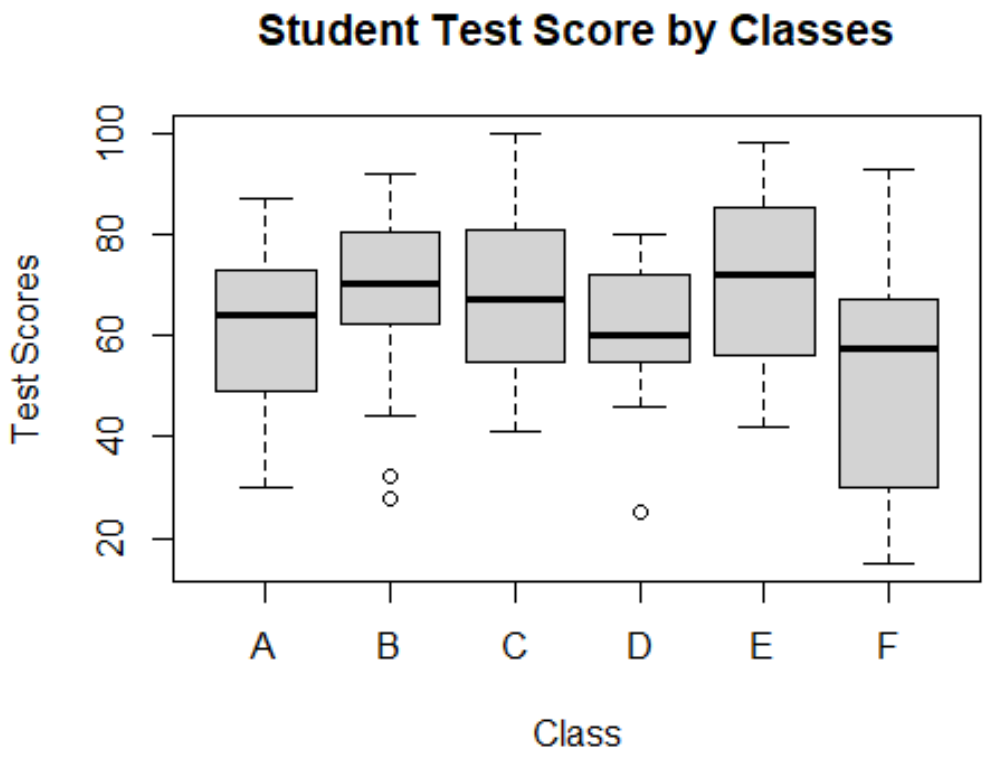
Boxplot (by gender)

```
boxplot(score~gender,
data=student,
        xlab="Gender",
        ylab="Test Scores",
        main="Student Test
Score by Gender")
```



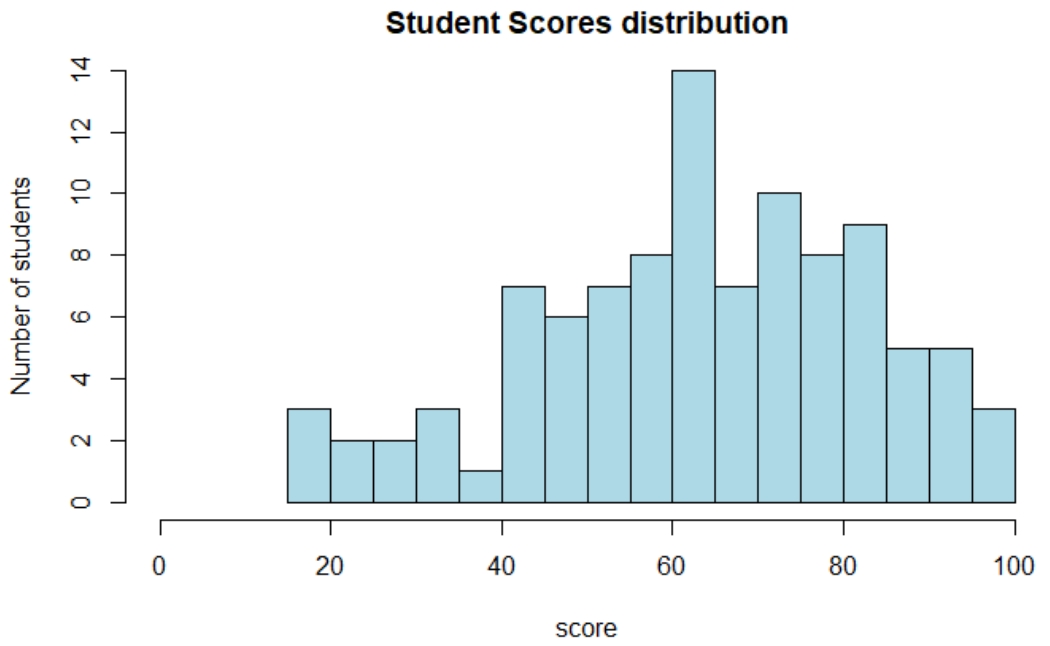
Boxplot (by class)

```
#how to create boxplot for score by class
boxplot(score~class,
data=student,
        xlab="Class",
        ylab="Test Scores",
        main="Student Test Score by Classes")
```



Histogram 長條圖

```
graph_h2 <- hist(student$score,
                 main="Student Scores
distribution",
                 xlab="score",
                 ylab="Number of students",
                 col="lightblue",
                 xlim=c(0,100),
                 breaks=20)
```



Correlation

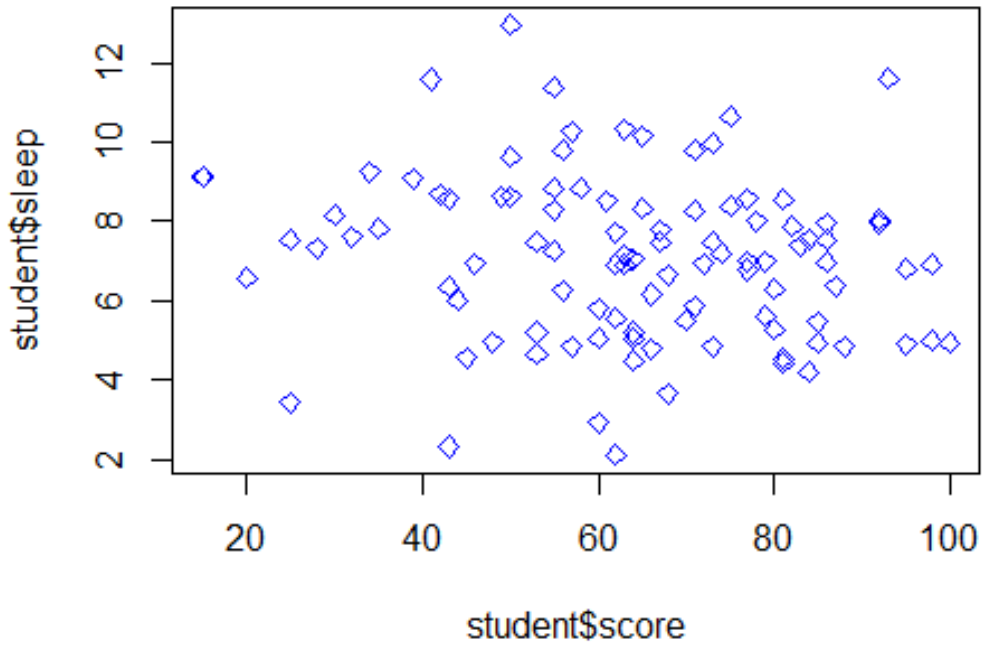
- Scatterplot 散布圖
  - A plot in which the x-axis is the value of one variable, and the y-axis the value of the other variable
- Correlation coefficients 相關係數
  - A metric that measures the extent to which numeric variables are associated with one another (ranges from –1 to +1).
  - Pearson's correlation coefficient

$$r = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

- Correlation matrix
  - A table where the variables are shown on both rows and columns, and the cell

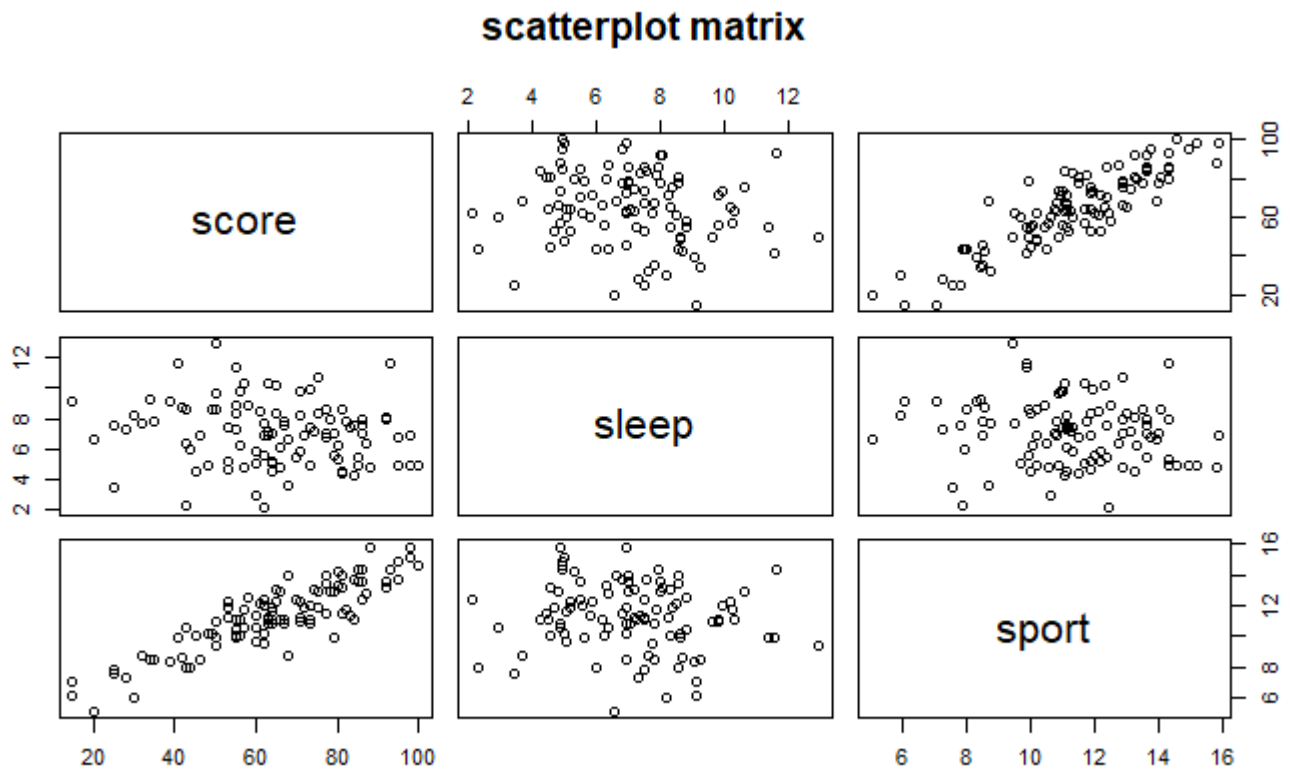
Scatterplot

```
#scatterplot
plot(student$score,
student$sleep,
     col="blue",
     pch=23)
```



Scatterplot Matrix

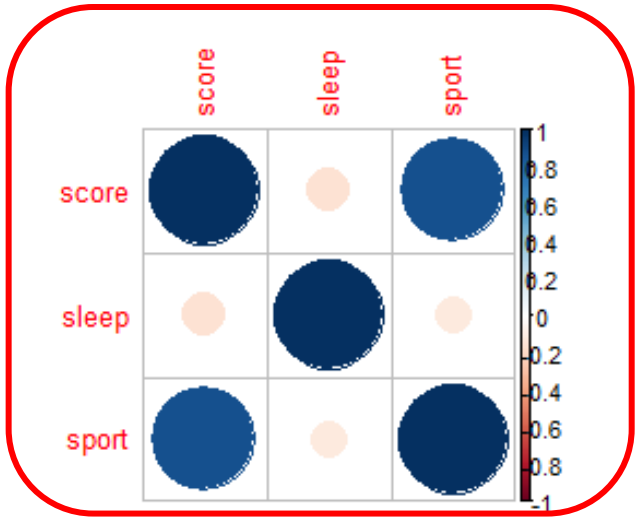
```
pairs(~score+sleep+sport,
data=student,
      main="scatterplot
matrix")
```

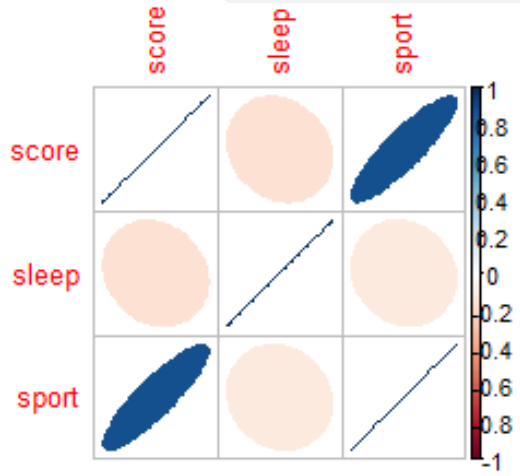


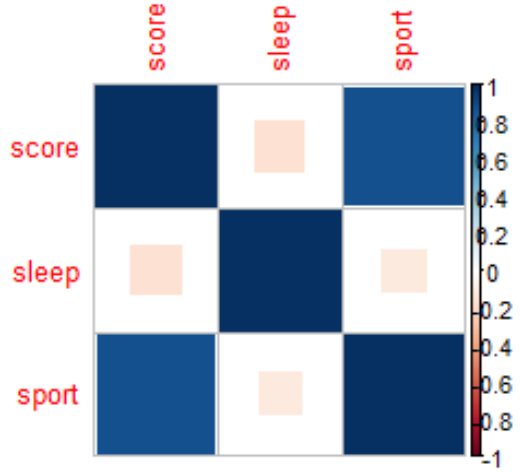
Correlation Matrix

```
library(corrplot)
corrplot(cor(performanc
e), method="circle")

#circle, square,
ellipse, number, color,
pie can be used
```







Practice

- Create a Table to show the count of gender and profession
  - Both number counts and proportion
- Calculate the quantile of 20%, 40%, 60% and 80% of the spending score
- Draw the boxplot of spending score by profession
  - X-axis shown “Working Profession”
  - Y-axis shown “Spending Score”
  - Title “Spending Score by Profession”
  - Use a **comment in R file to specify who profession has a highest score**

Practice

- Draw a histogram to show the distribution of spending score
  - Identify the number of customers whose score fall between 50 and 60.
  - Use comment to specify

```
df <- customer[,c("Age", "Score", "Work", "Family", "Income")]
df$Income <- log(df$Income)
```

- Draw the correlation matrix among
  - Score
  - Income
  - Work
  - Family
  - Age