R 套件教學簡介

***在正式進行 R 實作課程之前，以下幾點建議提供學員們前置作業時參考：

1. 建議攜帶個人筆電，事先安裝完成 R 及 R studio，安裝細節可參考以下網頁
http://www.cc.ntu.edu.tw/chinese/epaper/0030/20140920_3006.html
http://cran.csie.ntu.edu.tw/bin/windows/base/
http://joe11051105.gitbooks.io/r_basic/content/environment_settings/RStudio_introduction.html
https://www.rstudio.com/products/rstudio/download/
http://www.r-tutor.com/elementary-statistics/numerical-measures/skewness

2. 以下指令設定工作路徑，在讀取或寫入檔案時，可以省掉很多麻煩。
```
setwd("C:\\Users\\Chih-Hsuan  Wang\\Desktop\\R 檔案\\")
setwd("C:/Users/chtti/Desktop/R 檔案")
```

3. 請先將個人檔案存成 DAT、CSV 或 TXT 檔，在 R 環境下可用以下指令抓取
```
heart<-  read.table("heart.dat",  header=T,  sep="  ")
glass<-  read.table("glass.txt",  header=T,  sep=",")
bank<-  read.table("bank.csv",  header=T,  sep=";")
rm(list=ls())  #清除記憶體指令，避免先前的殘留資料
```

4. 上課時不會對外部資料庫進行直接連結，也不會進行資料清理或格式不一致的處理；同時建議欄位名稱盡量用英文命名，避免許多原因未明的操作干擾

5. 若有缺失值欄位(屬性)的資料請事先補齊(平均數、中位數、眾數)或直接刪除，避免屆時 R 套件處理發生問題(查詢缺失值之指令如下)，亦可用迴歸方式進行差補

6. R 操作的範例檔案已壓縮存在資料夾中，有興趣的學員可以事先感覺一下

***2006 年在香港舉辦的 IEEE 的 ICDM 國際會議選出了 10 大最具影響力的資料採礦演算法，列舉如下：C4.5, K-means, SVM, Apriori, EM, PageRank, AdaBoost, KNN, Naïve Bayes, CART，除 PageRank 外，其他皆在影音及實體課程授課範圍

第一週

```
data(iris)
attributes(iris)
summary(iris)
plot(iris)
plot(Species, Sepal.Length, main="Distribution of Sepal.Length")

attach(iris)
var=c(1:4)
colMeans(iris[,-5])
cor(Sepal.Length,  Sepal.Width)
corr=cor(iris[var],  use="pairwise")
corr

cov(Petal.Length,  Petal.Width)
covv=cov(iris[var],  use="pairwise")
covv

install.packages("timeDate")
library(timeDate)
skewness(iris[,1:4])  #  https://en.wikipedia.org/wiki/Skewness
kurtosis(iris[,1:4])   #  https://en.wikipedia.org/wiki/Kurtosis
```

偏態值 ＞0，為正偏態，分配集中在平均數以下，低分群的個體較多。
偏態值 ＜0，為負偏態，分配集中在平均數以上，高分群的個體較多。
峰度值 ＞3，為高狹峰，較常態分配來得高瘦。
峰度值 ＜3，為低闊峰，較常態分配來得矮胖。

常態性檢定

```
hist(Sepal.Length, breaks=seq(4.0, 8.0, 0.25))
hist(Sepal.Length, breaks=seq(4.0, 8.0, 0.25), prob=TRUE)
qqnorm(Sepal.Length, xlab= "Z-score", ylab="Sepal.Length")
qqline(Sepal.Length, col="red")
curve(dnorm(x, mean(Sepal.Length), sd(Sepal.Length)), 4.0, 8.0, col="red")
shapiro.test(Sepal.Length)

qqplot(Sepal.Length, Sepal.Width) #判定兩者的機率分配是否相似
install.packages("nortest")
library(nortest)
```

ad.test(Sepal.Width)
sf.test(Sepal.Width)
cvm.test(Sepal.Width)
lillie.test(Sepal.Width)
pearson.test(Sepal.Width)
shapiro.test(Sepal.Width)

卡方檢定/比例檢定
male= c(Bush=315, Perot=152, Clinton=337)
female= c(Bush=346, Perot=126, Clinton=571)
rbind(male, female)
chisq.test(rbind(male, female))

citizen= c(sum(male), sum(female))
bush= c(male[1], female[1])
prop.test(bush, citizen, alternative="greater")
perot= c(male[2], female[2])
prop.test(perot, citizen, alternative="two.sided")
prop.test(perot, citizen, alternative="greater")
clinton= c(male[3], female[3])
prop.test(clinton, citizen, alternative="less")

雙群樣本平均值檢定 (F test/T test)
##先以 var.test 函數進行變異數相同與否的 F 檢定：若變異數相同，則執行 t.test 時設定 var.equal=TRUE，若變異數不相同，則設定 var.equal= FALSE 或省略。
data(iris)
setosa=subset(iris, Species=="setosa")
versicolor=subset(iris, Species=="versicolor")
var.test(setosa$Petal.Width, versicolor$Petal.Width)
t.test(setosa$Petal.Width, versicolor$Petal.Width, var.equal=FALSE)
t.test(setosa$Petal.Width, versicolor$Petal.Width, alternative ="less", var.equal=FALSE)