2. The data is about sales of laptops at a computer chain in London in January 2008.

(1) How would you do to know the data?

Explore the data structure using str(), summary() and dim(). In addition, examine the beginning and the ending of data, head() and tail(), to make sure data is imported well, and to get sample data. (將 data frame “LaptopSalesJanuary2008Sub.csv” 重新命名成 “laptop.df”，方便後面的分析)

表一：Structure of data frame “laptop.df”.

```
> str(laptop.df)
'data.frame':    7956 obs. of    13 variables:
 $ Date                          : Factor w/ 7303 levels "1/1/2008 0:01",..: 1 2 3 3 4 5 6
7 8 9 ...
 $ Configuration                 : int    163 320 23 169 365 309 75 346 70 351 ...
 $ Customer.Postcode             : Factor w/ 834 levels "BR3 1AG","...
```

表二：Summary of data frame “laptop.df”.

```
> summary(laptop.df)
                  Date          Configuration      Customer.Postcode    Store.Postcode
 1/28/2008 16:01:     4      Min.     :    1.0      W1T 1DG :    21         SW1P 3AU:1604
 1/28/2008 23:10:     4      1st Qu.: 77.0      EC4V 2BA:    19         SE1 2BN :1232
 1/1/2008 10:06 :     3      Median :209.5      SE16 2HB:    19         SW1
```

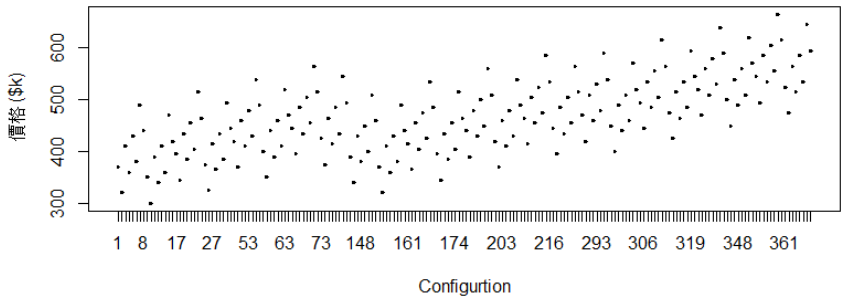
表三: dimension of data frame “laptop.df”.

```
> dim(laptop.df)
```

(2) At what price are the laptops actually selling? Hint: define “actually selling”

我們假設一般狀況下，規格越好，價格越貴。從圖 2-1 各型號和價格的 boxplot 可以看出本來的盒狀分佈圖變成只有一個點，表示同型號會有同樣的價格，而型號和價格的分佈圖隱約可以看出型號越大，價格呈現上升的趨勢。所有電腦的平均價格為 487.9349。

```
> mean(laptop.df$Retail.Price)
[1] 487.9349
```



(圖 2-1)

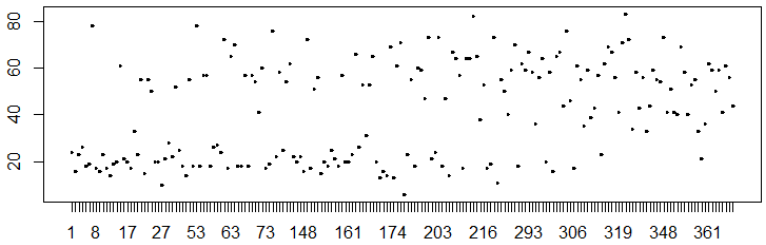
我們對於 actually selling 的想法是，雖然不同型號的電腦都有不同的價格，但是會有某個價格是在合理的規格下大部分消費者都能接受的性價比，因此我們要找出那個價格，目前有兩種方式，(2-1)最多人買的型號的價格是多少，(2-2)最多人買的、可以接受的價格是多少，而在那個價格下的電腦型號又是甚麼，進行比較。

(2-1)最多人買的電腦型號的價格是多少?

我們發現從 192 個電腦型號中，最多數量的電腦型號是 337，總共有 83 個人買此型號，規格為 15 吋、電池續航力 6 小時、RAM 2GB、HD size 40GB、執行速度 1.5GHz、具有整合無線網路還有一些附帶套件。這是最多人買的型號價格為 520，略高於平均價格，根據規格越高價格越貴，表示此規格高於平均，我們可以認為在意規格的人，也可能是跟電腦工作較相關的人，價格不是他們最在意的點，而在此規格下，高於平均的價格可能是電腦工作相關的人會購買的型號。而此筆資料總共有 7000 多筆資料，但最多人買的電腦型號只有 83 個人購買，並不能代表整體消費者的購買意願。再從圖 2-2 各型號和分別購買數量的 scatter plot 來看，最高的 83 比購買數量和其他型號相比都不是高太多，表示各型號各有愛好者。

此　方　法　不　夠　具　有　代　表　性　。

```
> laptop.df[maxConfigPrice[1],c(2,5:12)]#最多人買的電腦型號規格
   Configuration Retail.Price Screen.Size..Inches. Battery.Life..Hours. RAM..GB. Processor.Speeds..GHz.
33           337          520                   15                    6        2                    1.5
   Integrated.Wireless. HD.Size..GB. Bundled.Applications.
33                  Yes          40                   Yes
>
```

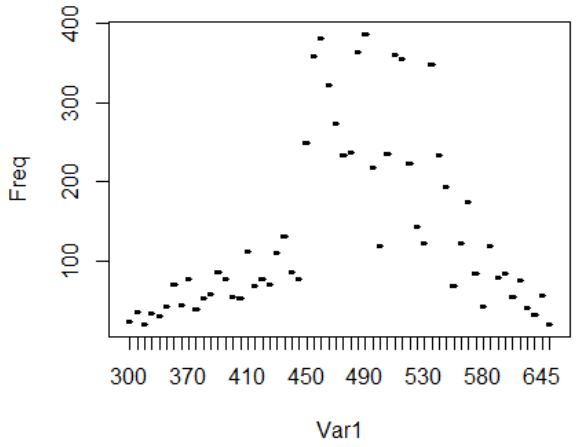


(圖 2-2)

(2-2) 最多人買的價格是多少，而在那個價格下的電腦型號又是甚麼?

我們發現總共有 58 種價格，其中最多人購買的價格為 490，總共有 387 筆購買數，490 略高於平均，且此價格下總共有六種規格，分別為 7,56,159,208,299,348，其中又屬規格 7 最多購買筆數，規格為 15 吋、電池續航力為型號中最弱的 4 小時、RAM 1GB、HD size 300GB、執行速度 1.5GHz、具有整合無線網路還有一些附帶套件，7 號以最大的 HD size 為優勢，是許多人考量的點，次高為 348，以 2GB RAM 和 6 小時電池續航力勝出。這六個型號都是 15 吋螢幕、執行速度 1.5GHz，其他略有差異。

(圖 2-3)



各價格的數量分佈，其中 490 為最多人的購買價格。

```
> MaxPriceconfig
  ConfigmaxPrice Freq
1              7   78
2             56   57
3            159   57
4            208   64
5            299   58
6            348   73
>
```

(圖 2-4)

此為不同電腦型號在 490 元分別的購買數量。

```
> laptop.df[CC[c(1:4,9:10)],c(2,5:12)]#查看最大購買價格筆數的configuraiton的規格
    Configuration Retail.Price Screen.Size..Inches. Battery.Life..Hours. RAM..GB.
20            299          490                    15                    6        1
57              7          490                    15                    4        1
59            159          490                    15                    5        1
75            348          490                    15                    6        2
107            56          490                    15                    4        2
171           208          490                    15                    5        2
    Processor.Speeds..GHz. Integrated.Wireless. HD.Size..GB. Bundled.Applications.
20                     1.5                   No          80                    Yes
57                     1.5                  Yes         300                    Yes
59                     1.5                   No         300                    Yes
75                     1.5                   No          80                     No
107                    1.5                  Yes         300                     No
171                    1.5                   No         300                     No
```

我們認為在 490 的價格下為 actually selling，這些規格都各有愛好者也能符合大眾所需，可以知道一般人的規格需求是甚麼。在規格制訂時可以推出略高於平均

且價格相同，而功能規格上略微差異可以符合更多不同消費者的需求，才能得到更多獲利。

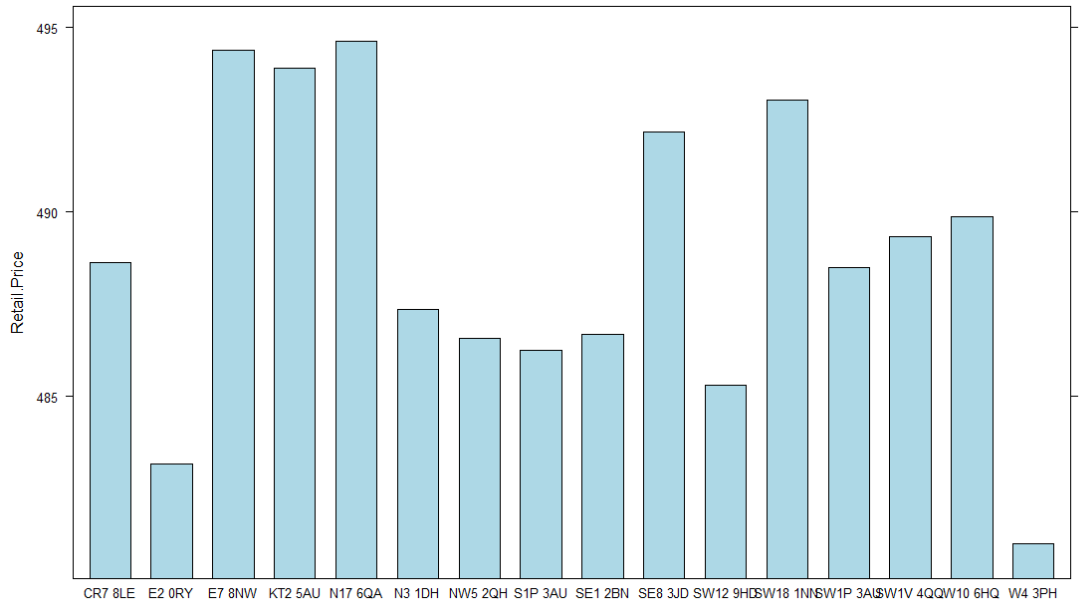
(3) Create a bar chart, showing the average retail price by store. Which store has the highest average? Which has the lowest?

表一：Aggregate function and bar chart codes.

```
>price.mean<-aggregate(Retail.Price~Store.Postcode, data=laptop.df, mean)
>library(lattice)
>barchart(Retail.Price~Store.Postcode,data=price.mean,col="lightblue")
```

表一為程式碼，先使用 aggragate function 將每家店的零售價根據店家郵遞區號做整理，再呼叫 lattice 功能將每家店的零售價根據店家郵遞區號做出 bar chart。而從下方表二的 bar chart 圖可看出 Store Postcode. N17 6QA 的店家有最高的平均零售價；而 Store Postcode. W4 3PH 則有最低的平均零售價。

表二: bar chart 圖



(4) Are average prices consistent across retail outlets in general?

從第三題我們算出一月各店家的平均銷售金額，但是店家的實際銷售金額是否一樣，我們可以用 ANOVA 進行檢定。

我們將以各店家進行分組，和銷售額做 ANOVA 分析:

H0:16 個店家母體銷售額平均均相同

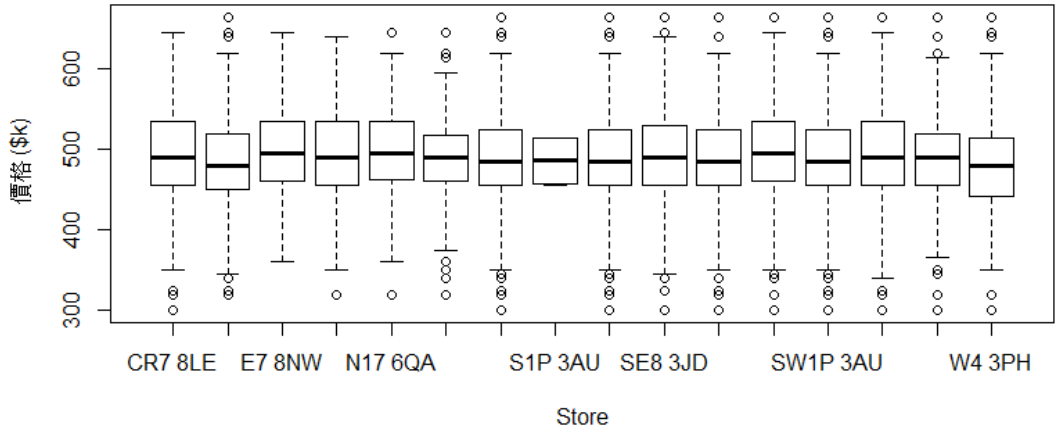
H1:至少一家店的母體銷售額平均與其他店不相同

Store.Postcode 為組間資料，Residuals 為組內資料，Mean Squre 為變異數，F value 為組間變異數除以組內變異數。從圖 4-1 中可以看到 F value>1，表示組間差距大於組內差距， p value 為 0.2737>0.05，表示沒有顯著差異，不能否定 H0，所以我們可以說 16 家店的平均銷售額沒有顯著差異。

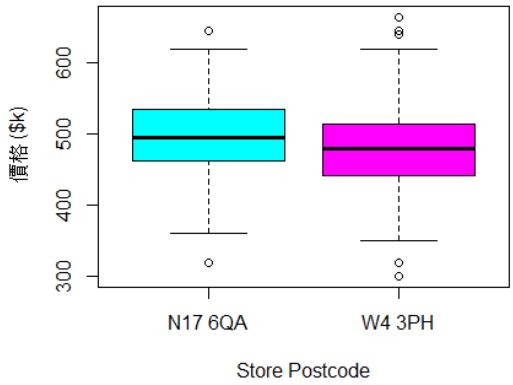
```
> anova(seg.aov.own)
Analysis of Variance Table

Response: Retail.Price
                Df   Sum Sq Mean Sq F value Pr(>F)
Store.Postcode  15    67337  4489.1  1.1866 0.2737
Residuals     7940 30038584  3783.2
```

(圖 4-1)

(5) To better compare retail prices across stores, create boxplots of retail price by store. Now compare the prices in the two stores from (3). Does there seem to be a difference between their price distributions?



(圖 5-1)此為各店家銷售金額盒狀分佈圖



(圖 5-2)此為平均最大(N17 6QA)和最小銷售額店家(W4 3PH)銷售金額分佈盒狀圖比較

表一：Basic statistics codes such as quartiles and range of price.

```
minPrice<-laptop.df$Retail.Price[laptop.df$Customer.Postcode=='W4 3PH']
maxPrice<-laptop.df$Retail.Price[laptop.df$Customer.Postcode=='N17 6QA']
IQR(minPrice)
IQR(maxPrice)
range(maxPrice)[2] - range(maxPrice)[1]
range(minPrice)[2] - range(minPrice)[1]
```

表一分別呈現出店家 W4 3PH 和 N17 6QA 各自的四分位差(IQR)以及全距(range)。店家 W4 3PH 的四分位差為 70，全距為 285，四分位差和全距的比例為 0.6087。店家 N17 6QA 的四分位差為 83.75，全距為 285，四分位差和全距的比例為 0.2939。兩者的比例相差 2.07 倍(0.6087/0.2939)，因此兩者的分布情況以店家 W4 3PH 較為分散，店家 N17 6QA 較為集中。

除了用 boxplot 觀察最大平均銷售金額店家和最小平均銷售金額店家的銷售額分佈狀況，我們也可以用 t-test 和 f-test 來估計實際母體分佈狀況是否一樣。

圖 5-3 的 t-test:

H0:最大平均銷售金額店家的母體銷售額平均(x1)=最小平均銷售金額店家的母體銷售額平均(x2)

H1: 最大平均銷售金額店家的母體銷售額平均(x1)不等於最小平均銷售金額店家的母體銷售額平均(x2)

從圖中可以看到 p value>0.05，且信心區間包含 0，表示 x1-x2 有可能等於 0，沒有顯著差異，所以 HO 成立，接著我們再檢查變異數是否有差異。

圖 5-4 的 f-test:

H0:最大平均銷售金額店家的母體銷售額變異數(v1)=最小平均銷售金額店家的母體銷售額變異數(v2)

H1: 最大平均銷售金額店家的母體銷售額變異數(v1)不等於最小平均銷售金額店家的母體銷售額變異數(v2)

從圖中可以看到 p value>0.05，且信心區間包含 1，表示 v1、v2 沒有顯著差異，所以 HO 成立。

以上檢定可以證明最大平均銷售金額店家和最小平均銷售金額店家的母體銷售額分佈沒有顯著差異。

```
> t.test(Price ~ Store, data=minmax)

        Welch Two Sample t-test

data:  Price by Store
t = 1.8015, df = 276.88, p-value = 0.07271
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.263606 28.519320
sample estimates:
mean in group N17 6QA  mean in group W4 3PH
             494.6341              481.0063
```

(圖 5-3)

```
> var.test(Price ~ Store, data=minmax)

        F test to compare two variances

data:  Price by Store
F = 0.73824, num df = 122, denom df = 158, p-value = 0.0792
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5297771 1.0362857
sample estimates:
ratio of variances
         0.7382369
```

(圖 5-4)

(6) Which stores are selling the most in terms of quantities? Provide a visual presentation, and a description about which store it is. Do store quantities vary in general?

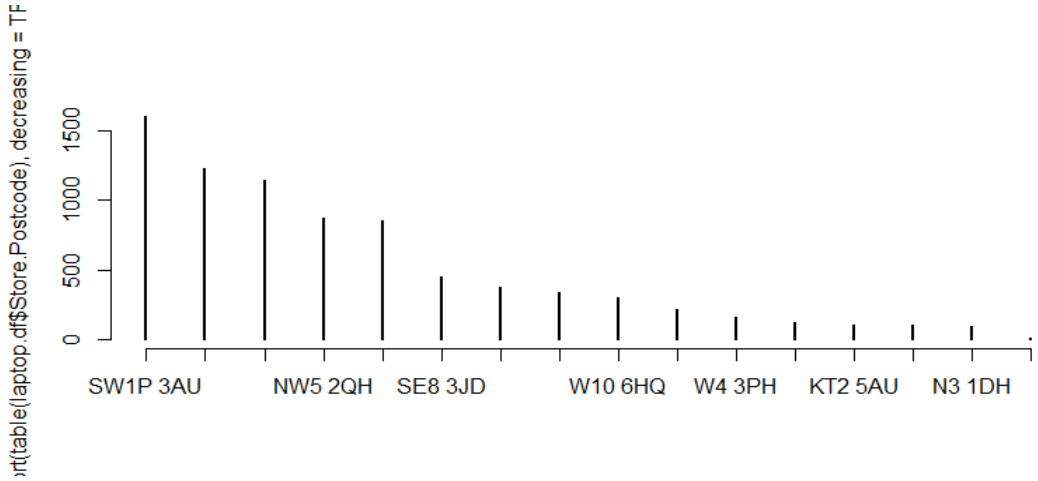
圖 6-1 為各店家賣出的商品總數，從圖中可以看出賣出最多的店家為 SW1P 3AU。而我們也發現賣出最多數量的商品的店家同時也是顧客跟店家平均距離最近的店家，可以證明若顧客離店家距離越近，銷售量也會提高，其他店家附近的住戶沒有前來有可能是其他店家附近的住戶量不夠高或是跟店家賣的產品種類有關。

根據圖 6-3，觀察各店家所賣產品數量整體上來說是否分布不均，我們以 chi-square test 檢定:

H0:各店家賣的母體商品數量佔母體整體數量比例均相同

H1:各店家賣的母體商品數量佔母體整體數量比例不完全相同

各店家所賣的產品數量整體上來說是有顯著差異的，因為 p value<<0.05，表示 H0 不成立，所以可以得知各店家所賣的產品數量有顯著差異。

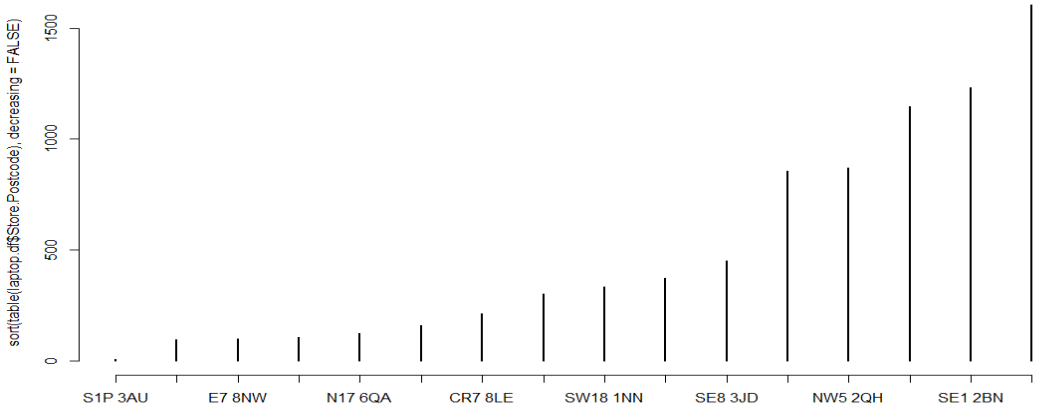


(圖 6-1)

(圖 6-2)

```
> chisq.test(table(laptop.df$Store.Postcode))

        Chi-squared test for given probabilities

data:  table(laptop.df$Store.Postcode)
X-squared = 7222.9, df = 15, p-value < 2.2e-16
```

(圖 6-3)



(7) In general, are stores' sales quantities depend on computer configurations?

我們要依資料分析所賣的商品數量是否會根據不同型號的電腦而有不同，我們用 chi square test 檢定不同型號對應賣出的數量:

H0:不同型號賣出的母體數量占母體整體賣出數量的比例均相同

H1:不同型號賣出的母體數量占母體整體賣出數量的比例不完全相同

根據圖 7-1，我們發現 p value<<0.05，所以 H0 不成立，表示不同型號對應賣出的數量有顯著差異，也就是說根據不同型號的電腦會賣出不同的數量，所以賣出數量跟電腦型號具有相關性。

```
> chisq.test(table(laptop.df$Configuration))

        Chi-squared test for given probabilities

data:  table(laptop.df$Configuration)
X-squared = 1983.3, df = 191, p-value < 2.2e-16
```

圖(7-1)

(8) What are revenues of stores?

表一: Filtering Laptop.df to find configuration type with highest sales.

```
>table(laptop.df$Retail.Price)
>laptop.df %>%
+    group_by(Store.Postcode, Configuration) %>%
+    filter(Retail.Price == 460) ->K
>table(K$Configuration)
```

(8-1)表一的程式碼介紹: 首先我們想知道哪一款型號的 laptop 銷量最高，因此我們用 table function，進而發現價格為 460 元的 laptop 最暢銷。而接下來我們用 group function 讓 data frame laptop.df 根據店家郵遞區號和 laptop 型號進行分類，再用 fiter function 篩選出價格為 460 元並命名為 K。最後再用 table function 呈現出 data frame K 中出現最多筆的型號。

表二: Laptop configuration type with highest sales.

```
> table(K$Configuration)
 51  61 152 203 292 302
 55  72  56  73  59  67
```

從表二可看出型號為 203 的 Laptop 賣的最好，共出現 73 筆銷量。

表三: 各店家一月銷售額

```
> Storesum
   Postcode      x
1   CR7 8LE 102610
2    E2 0RY 413595
3    E7 8NW  47955
4   KT2 5AU  51860
5   N17 6QA  60840
6    N3 1DH  46300
7   NW5 2QH 423325
8   S1P 3AU   1945
9   SE1 2BN 599590
10  SE8 3JD 221480
11 SW12 9HD 180530
12 SW18 1NN 164675
```

```
13 SW1P 3AU 783565
14 SW1V 4QQ 560300
15  W10 6HQ 146960
16   W4 3PH  76480
```

(8-2)我們想用 bar chart 做出各店家的銷售額圖，首先用 sum function 計算出一月所有店家的總銷售額為 3,882,010(sum(laptop.df$Retail.Price))，另外用 aggregate function 算出各店家的總銷售額並將其命名為 Storesum.(Storesum = aggregate(laptop.df$Retail.Price, by=list(Postcode=laptop.df$Store.Postcode), sum))。最後再用 bar chart function 以店家郵遞區號為橫軸，各店家總銷售額為縱軸呈現(如圖一)

圖一: Bar chart of each store's sales.