商業資料分析 Homework1

姓名：陳葳芃
系級：歐西四
學號：107508006

```
setwd("~/Downloads/1102 R/hw1")
sales.df <- read.csv("salesdata.csv")
prod.df <- read.csv("product_list.csv")
client.df <- read.csv("client_list.csv")
```

1. prod.df 裡將兩個變數，誤紀錄為在同一個column，其將其分開為兩個變數Product（數字部分）及Item（商品部分），取代原prod.df。

```
library(tidyverse)
prod.df<- prod.df %>%
  separate(Item,
           into=c("Product", "Item"),
           sep = "_")
```

2. 將3個報表合併為full.table。

```
sales.df$Product<- as.character(sales.df$Product)
full.table <- sales.df %>%
  left_join(client.df, by = "Client") %>%
  left_join(prod.df, by="Product")
```

說明：將Product定義為同類型資料才能進行合併

3. 在full.table. 新增一個變數「總消費」為spend = UnitPrice*Quantity。

```
full.table<- full.table %>%
  mutate( spend = UnitPrice * Quantity )
```

說明：使用mutate來新增一個名為spend的欄位

4. 在full.table將會員等級分組，其中gold和diamond的顧客為一組，其他等級的為一組，針對兩組客戶進行比較介紹（例如平均年紀、性別、國家、消費情況差異等）。

```
##資料進行分組
group1<-full.table %>%
  filter( Membership == "gold" | Membership =="diamond")
group2 <- full.table %>%
  filter( Membership != "gold" & Membership !="diamond")

#平均年紀比較
group1 %>%
  summarise(mean(Age))
group2 %>%
  summarise(mean(Age))
```

```
> #平均年紀比較
> group1 %>%
+   summarise(mean(Age))
  mean(Age)
1  27.31579
> group2 %>%
+   summarise(mean(Age))
  mean(Age)
1      32.3
```

比較分析：可以發現gold、diamond組的客戶，平均年齡（約27歲）較非gold、diamond組的客戶（約32歲）年輕。

```
#性別分布比較
group1 %>%
  group_by(Gender)%>%
    summarise(length(Gender))
group2 %>%
  group_by(Gender)%>%
  summarise(length(Gender))
```

```
> #性別分布比較
> group1 %>%
+   group_by(Gender)%>%
+     summarise(length(Gender))
# A tibble: 2 × 2
  Gender `length(Gender)`
  <chr>             <int>
1 female               13
2 male                  6
> group2 %>%
+   group_by(Gender)%>%
+   summarise(length(Gender))
# A tibble: 2 × 2
  Gender `length(Gender)`
  <chr>             <int>
1 female               10
2 male                 10
```

比較分析：gold、diamond組的客戶女性成員組成多於男性一倍，而非gold、diamond組的客戶男女分佈則相當平均，均為10人。

```r
#國家比較
group1 %>%
  group_by(Region)%>%
  summarise(length(Region))
group2 %>%
  group_by(Region)%>%
  summarise(length(Region))
```

```
> #國家比較
> group1 %>%
+   group_by(Region)%>%
+   summarise(length(Region))
# A tibble: 5 x 2
  Region   `length(Region)`
  <chr>               <int>
1 Brazil                  3
2 France                  4
3 Korea                   5
4 Spain                   4
5 Thailand                3
> group2 %>%
+   group_by(Region)%>%
+   summarise(length(Region))
# A tibble: 5 x 2
  Region   `length(Region)`
  <chr>               <int>
1 China                   2
2 Germany                 4
3 Japan                   4
4 Taiwan                  4
5 USA                     6
```

比較分析：gold、diamond組的客戶組成最多為韓國地區（5人），最少為巴西、泰國（均為3人）。非gold、diamond組的客戶佔最多的為美國（佔了6人），最少則為中國（僅有2人）。兩組共通點為均有亞洲、美洲、歐洲三個地區的客戶，但國家不同。

```r
#消費情況差異
group1 %>%
  summarise(mean(spend))
group1 %>%
  group_by(Item)%>%
  summarise(length(Item))

group2 %>%
  summarise(mean(spend))
group2 %>%
  group_by(Item)%>%
  summarise(length(Item))
```

```
> #消費情況差異
> group1 %>%
+   summarise(mean(spend))
  mean(spend)
1    241.6316
> group1 %>%
+   group_by(Item)%>%
+   summarise(length(Item))
# A tibble: 6 x 2
  Item       `length(Item)`
  <chr>               <int>
1 AirPods                 2
2 AppleWatch              2
3 iMac                    3
4 iPad                    3
5 iPhone                  5
6 MacBook                 4
> group2 %>%
+   summarise(mean(spend))
  mean(spend)
1      165.15
> group2 %>%
+   group_by(Item)%>%
+   summarise(length(Item))
# A tibble: 6 x 2
  Item       `length(Item)`
  <chr>               <int>
1 AirPods                 3
2 AppleWatch              3
3 iMac                    2
4 iPad                    5
5 iPhone                  3
6 MacBook                 4
```

比較分析：在這裡我先計算2組客戶的消費算術平均，也可發現就消費金額而言gold、diamond組的客戶確實比較高，同時，兩組顧客對於消費商品的偏好也有不同：gold、diamond組消費最多為iPhone；非gold、diamond組消費最多則是iPad。兩組共通點為第二高消費物品均為Macbook。

5. 在full.table針對女性客戶進行分析（例如平均年紀、國家、消費情況等），並對他們在不同產品的「總消費」畫圖分析。

```
group_female<-full.table %>%
  filter( Gender == "female" )
```

說明：將所需資料存成group_female

```
#平均年紀
group_female%>%
  summary()
```

```
> #平均年紀
> group_female%>%
+   summary()
    salesID        Store             Product             Client         UnitPrice
 Min.   : 5.0   Length:23          Length:23          Min.   : 2.000   Min.   :4.000
 1st Qu.:12.0   Class :character   Class :character   1st Qu.: 3.000   1st Qu.:5.000
 Median :21.0   Mode  :character   Mode  :character   Median : 6.000   Median :7.000
 Mean   :21.0                                        Mean   : 5.565   Mean   :6.522
 3rd Qu.:29.5                                        3rd Qu.: 8.000   3rd Qu.:8.000
 Max.   :39.0                                        Max.   :10.000   Max.   :9.000
    Quantity        Region              Age         Membership          Gender
 Min.   : 3.00   Length:23          Min.   :19.00   Length:23          Length:23
 1st Qu.: 7.50   Class :character   1st Qu.:21.00   Class :character   Class :character
 Median :34.00   Mode  :character   Median :26.00   Mode  :character   Mode  :character
 Mean   :30.65                      Mean   :27.78
 3rd Qu.:45.00                      3rd Qu.:34.00
 Max.   :64.00                      Max.   :35.00
     Item               spend
 Length:23          Min.   : 12.0
 Class :character   1st Qu.: 48.0
 Mode  :character   Median :216.0
                    Mean   :208.3
                    3rd Qu.:313.5
                    Max.   :512.0
```

分析：由summary函式可以看見女性客戶平均年紀落在27.78歲（四捨五入約28歲）。

```
#國家
group_female %>%
  group_by(Region)%>%
  summarise(length(Region))
```

分析：可以發現女性客戶最多落在美國；次之則是韓國。

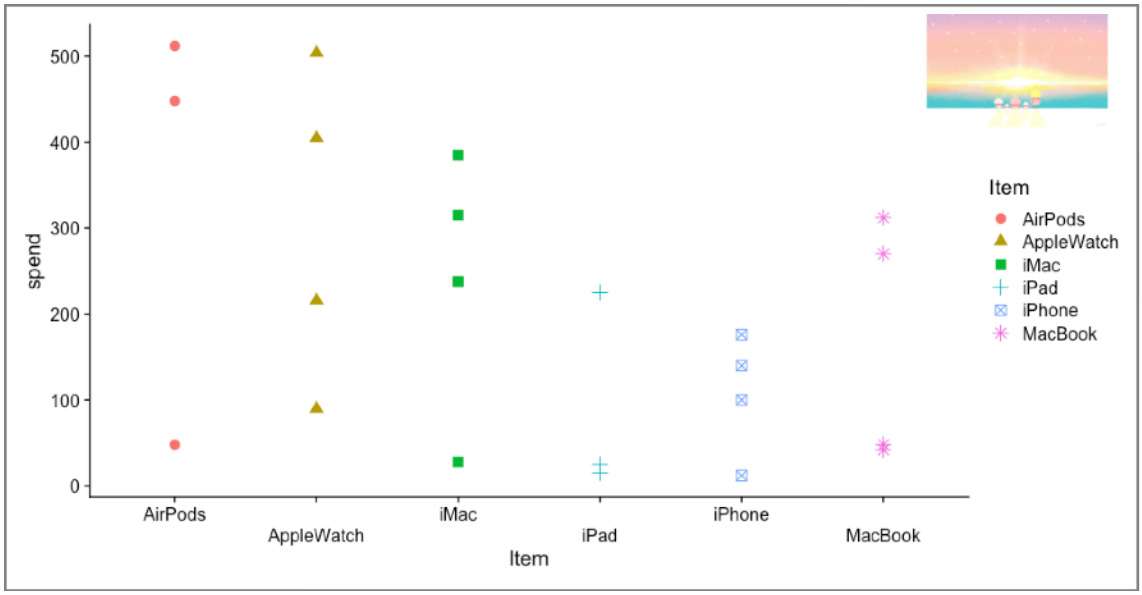
```
> #國家
> group_female %>%
+   group_by(Region)%>%
+   summarise(length(Region))
# A tibble: 5 x 2
  Region `length(Region)`
  <chr>             <int>
1 France                4
2 Japan                 4
3 Korea                 5
4 Spain                 4
5 USA                   6
>
```

```
#消費情形
group_female %>%
  group_by(Membership)%>%
  count(Region)
```

```
> #消費情形
> group_female %>%
+   group_by(Membership)%>%
+   count(Region)
# A tibble: 5 x 3
# Groups:   Membership [3]
  Membership Region     n
  <chr>      <chr>  <int>
1 diamond    France     4
2 diamond    Korea      5
3 gold       Spain      4
4 silver     Japan      4
5 silver     USA        6
>
```

分析：在此我先用會員等級進行分組再看看各國家在不同等級的分佈情形，很平均的是每個國家基本上只對照一種會員等級（意即：此資料集中單一國家只有一種會員等級），而在女性客戶中完全沒有一般等級的會員，推測女性會員消費都有抵達一定水準。

```
##在不同產品的「總消費」畫圖分析
library(cowplot)
ii <- group_female %>%
  ggplot(aes(x=Item, y=spend ,color=Item,shape=Item)) +
  scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  geom_point(size=3) +
  theme_cowplot()
ggdraw(ii) +
  draw_image("foto.jpg", x = 1, y = 1, width = 0.2, height = 0.2,hjust = 1, vjust = 1)
```



分析說明：在不同產品分類下（X軸），女性客戶每筆訂單的總消費畫出的cowplot，可以透過不同icon來清楚分辨不同產品，並可對照該筆訂單總消費的金額來看出在Y軸的大約數值。從圖表可以觀察出單筆訂單最大值出現在Airpods，最小則可能出現在iPhone或iPad，且iPhone訂單消費金額的離散程度較小（較集中）。