大數據資料處理實務

李水彬Shuipin

2023/9/1

- 散佈圖是將兩個量化變數的成對觀察數值標示在二維座標系統上，點在此座標系統上的散佈可展現兩個量化變數的相關性。
- 散佈圖是以視覺化呈現變數關聯型的工具。

Chapter 07 散佈圖 scatter plot

載入套件

- 若電腦沒有以下套件, 將以下指令複製後貼在Console視窗執行
- 執行過程不用重新啟用R
- 若已經安裝, 可以略去以下指令
- copy 以下指令貼在 R script 視窗上, 將指令的前置符號 “#” 刪除

```
#install.packages("ggplot2")
#install.packages("gridExtra")
#install.packages("patternplot")
#install.packages("data.table")
#install.packages("knitr")
```

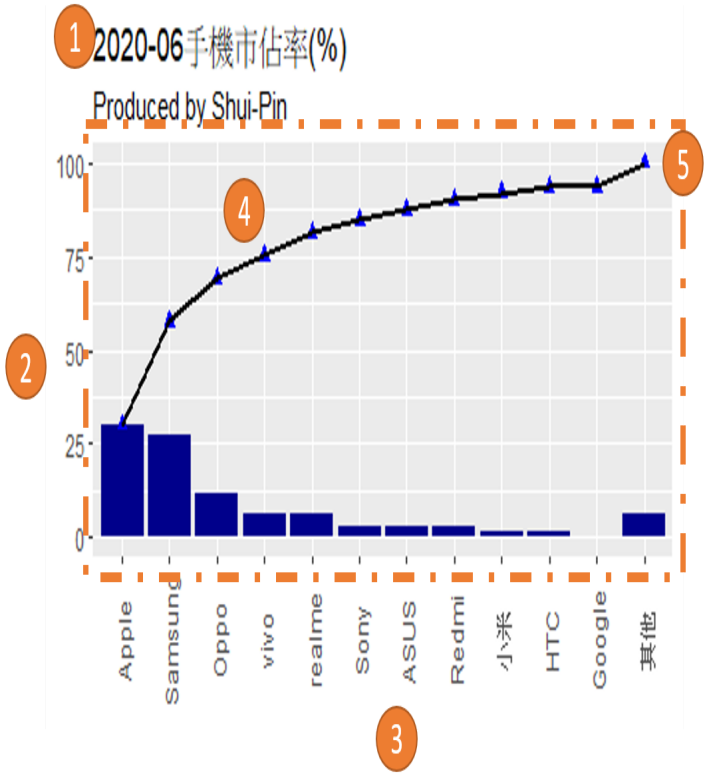
取用套件

```
library(ggplot2) #繪圖套件
library(gridExtra) #版面控制
library(patternplot)
library(data.table)
library(knitr)
```

統計圖三部曲

- 資料整理、統計量(參見之前介紹)
- 繪製統計圖
- 海報布局與說明

統計圖架構



- 標題
- Y軸座標
- X軸座標
- 統計圖
- 框架

資料整理

```
sushi<-fread("sushi.csv",header="auto") # 讀取檔案
str(sushi)
```

```
## Classes 'data.table' and 'data.frame':   48 obs. of  12 variables:
##  $ date            : int  20140301 20140302 20140303 20140304 20140305 20140306 20140307 2014
##  $ week            : int  6 7 1 2 3 4 5 6 7 1 ...
##  $ dish            : int  1511 1500 522 545 493 522 939 1094 1936 833 ...
##  $ revenue         : int  45340 45000 15660 16340 14780 15660 28160 32830 58090 25000 ...
##  $ adj-revenue     : int  45340 45000 15660 16340 14780 15660 28160 32830 58090 25000 ...
##  $ weather         : chr  "cloudy" "rainy" "rainy" "rainy" ...
##  $ promotion       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ holiday         : int  1 1 0 0 0 0 0 1 1 0 ...
##  $ temperature     : num  17.4 13.1 14.7 14.3 14.6 13.5 14.6 13.4 12.8 15.2 ...
##  $ high-temperature: num  23.5 17.5 16.9 16.6 16.2 14.7 16.1 15.5 14.3 16.4 ...
##  $ low-temperature : num  14.1 12.2 11.8 13.2 13.3 12.1 12.1 11.7 11.1 12.9 ...
##  $ humidity        : int  88 97 85 90 76 89 90 10 81 68 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

資料整理

```
Classes 'data.table' and 'data.
 $ date             : int  20140
 $ week             : int  6 7 1
 $ dish             : int  1511
 $ revenue          : int  45340
 $ adj-revenue      : int  45340
 $ weather          : chr  "clou
 $ promotion        : int  0 0 0
 $ holiday          : int  1 1 0
 $ temperature      : num  17.4
 $ high-temperature: num  23.5
 $ low-temperature : num  14.1
 $ humidity         : int  88 97
 - attr(*, ".internal.selfref")
```

- int 整數(沒有小數點)
- num 數值(有小數點)
- chr 文字(類別型資料)

數值格式轉換

```
sushi$revenue<-as.numeric(sushi$revenue)
str(sushi)
```

```
## Classes 'data.table' and 'data.frame':   48 obs. of  12 variables:
##  $ date            : int  20140301 20140302 20140303 20140304 20140305 20140306 20140307 2014
##  $ week            : int  6 7 1 2 3 4 5 6 7 1 ...
##  $ dish            : int  1511 1500 522 545 493 522 939 1094 1936 833 ...
##  $ revenue         : num  45340 45000 15660 16340 14780 ...
##  $ adj-revenue     : int  45340 45000 15660 16340 14780 15660 28160 32830 58090 25000 ...
##  $ weather         : chr  "cloudy" "rainy" "rainy" "rainy" ...
##  $ promotion       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ holiday         : int  1 1 0 0 0 0 0 1 1 0 ...
##  $ temperature     : num  17.4 13.1 14.7 14.3 14.6 13.5 14.6 13.4 12.8 15.2 ...
##  $ high-temperature: num  23.5 17.5 16.9 16.6 16.2 14.7 16.1 15.5 14.3 16.4 ...
##  $ low-temperature : num  14.1 12.2 11.8 13.2 13.3 12.1 12.1 11.7 11.1 12.9 ...
##  $ humidity        : int  88 97 85 90 76 89 90 10 81 68 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

散佈圖(雙變數)

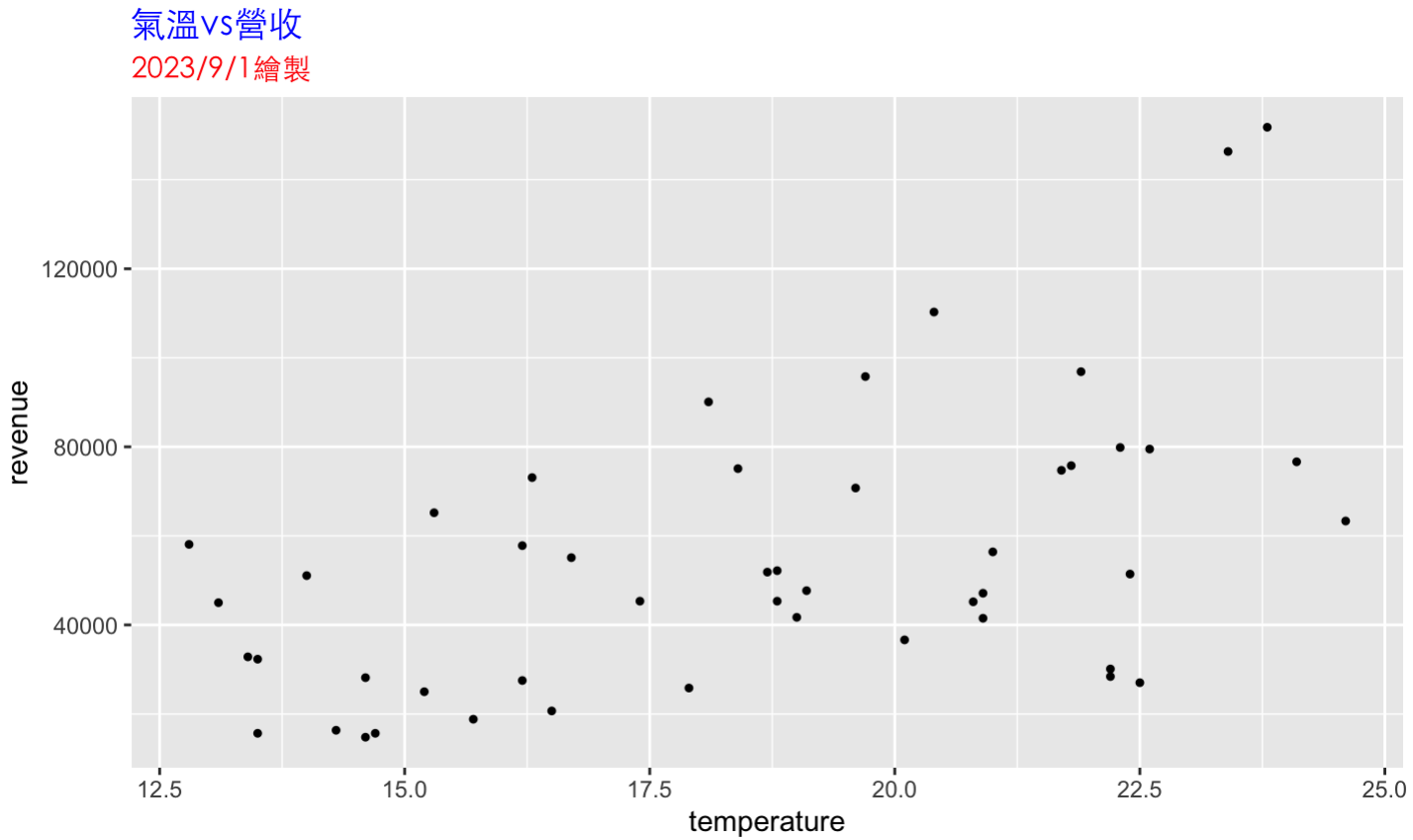
```
g<-ggplot(data=sushi,aes(x=temperature, y=revenue))
#啟動繪圖程式, 以溫度 temperature 為x軸
g<-g+geom_point(shape=20) # 散佈圖,y 軸為營收
g<-g+labs(title="氣溫vs營收",subtitle="2023/9/1繪製") #標題
```

- ggplot() 啟動繪圖
  - data=sushi 繪圖資料來源(資料集)
  - aes() 定義X 軸與Y 軸的變數
- geom_point() 描點
  - shape=20 點的圖示編號

- labs() 標題名稱

圖片風格設定

```
g<-g+theme(legend.position = "none", plot.title=element_text(family="STHeitiTC-Light",face=1,col
           plot.subtitle=element_text(family="STHeitiTC-Light",face="bold",color = "red")) #風格
g
```



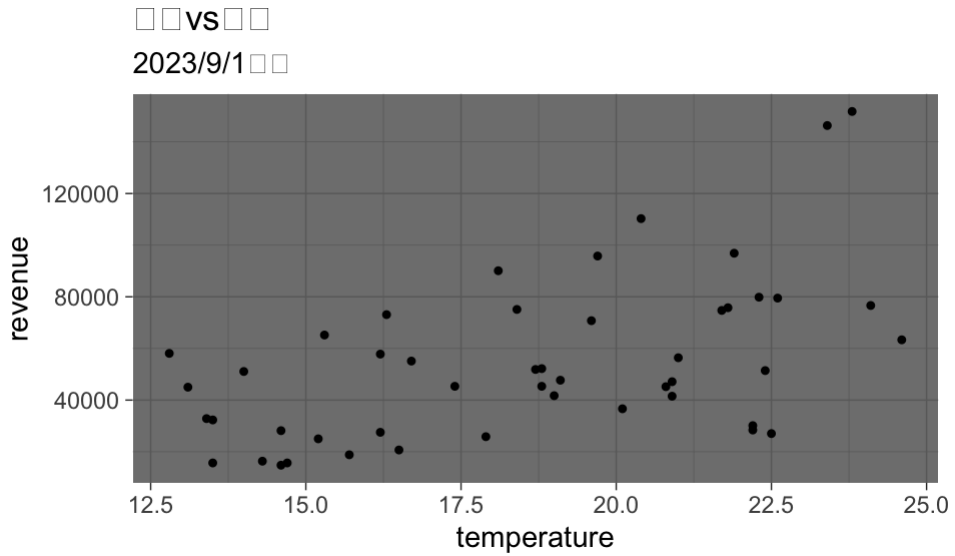
- theme() 圖片物件風格

內建風格

內建風格

- theme_dark() 黑色風格

```
gdark<-g+theme_dark() #黑色風格
gdark
```

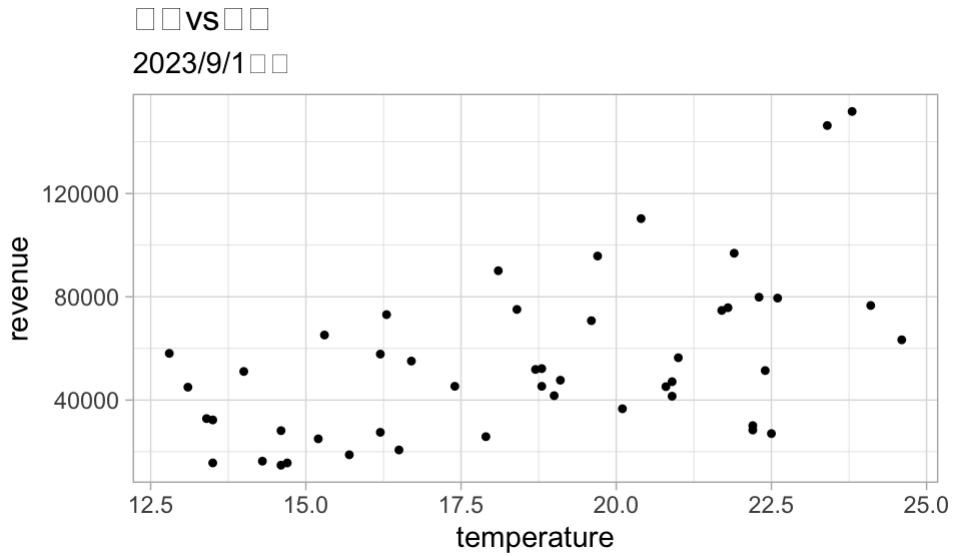


內建風格

內建風格

- theme_light() 明亮風格

```
glight<-g+theme_light() #明亮風格
glight
```

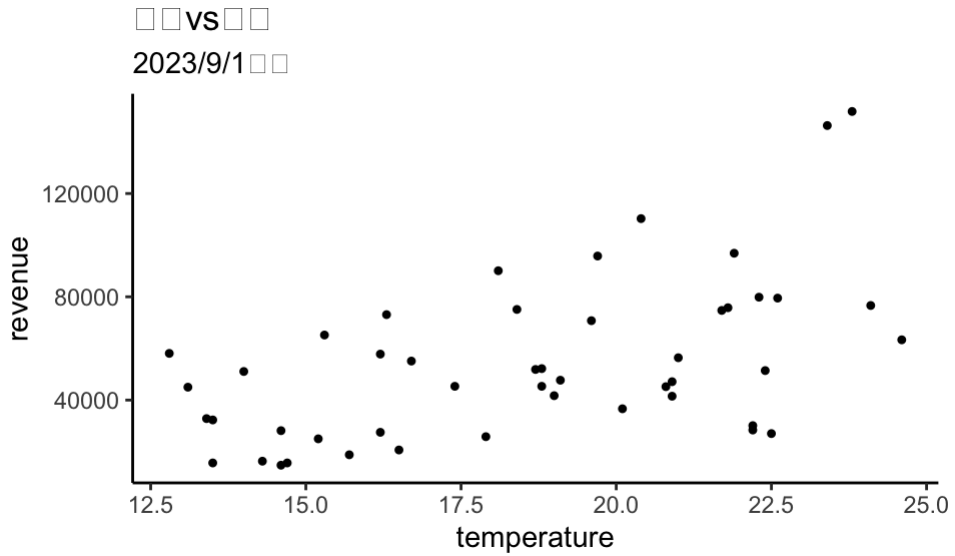


內建風格

內建風格

- theme_classic() 經典風格

```
gcls<-g+theme_classic() #經典風格
gcls
```



儲存統計圖

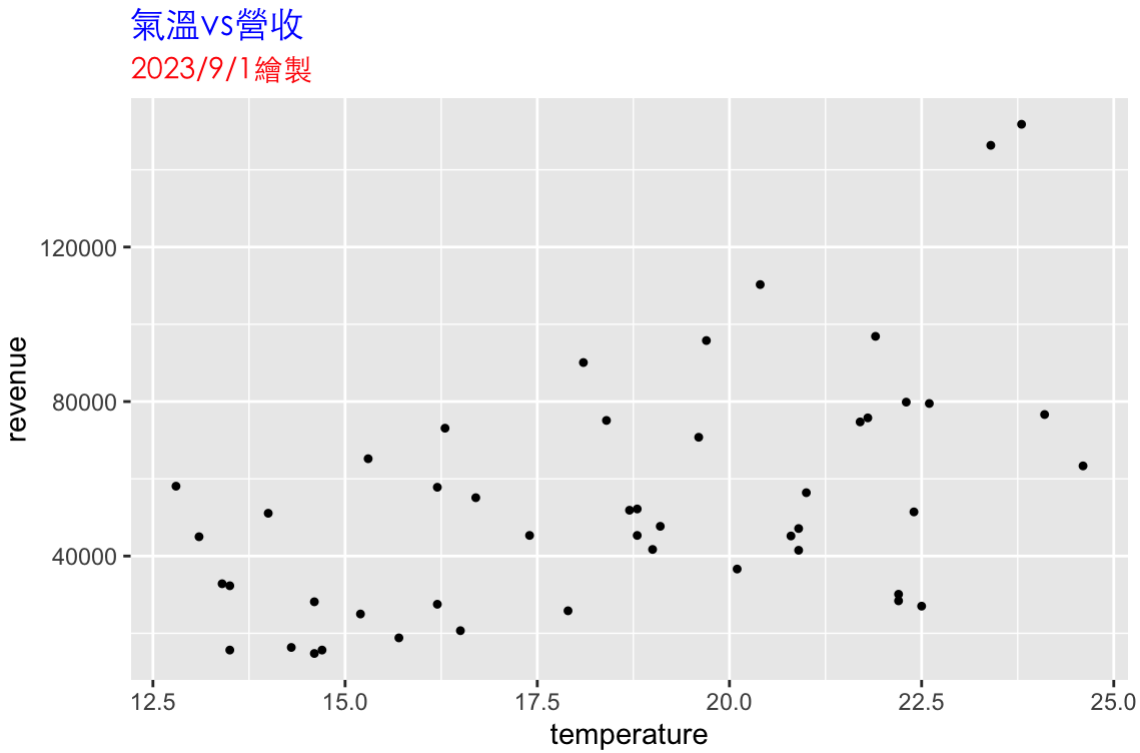
```
png("sushi_scatter_v01.png", width=640,height=480)#檔案名稱，圖片大小設定
print(g)
dev.off()
```

```
## quartz_off_screen
##                 2
```

- dev.off() 完成圖片以檔案輸出，改回螢幕輸出。

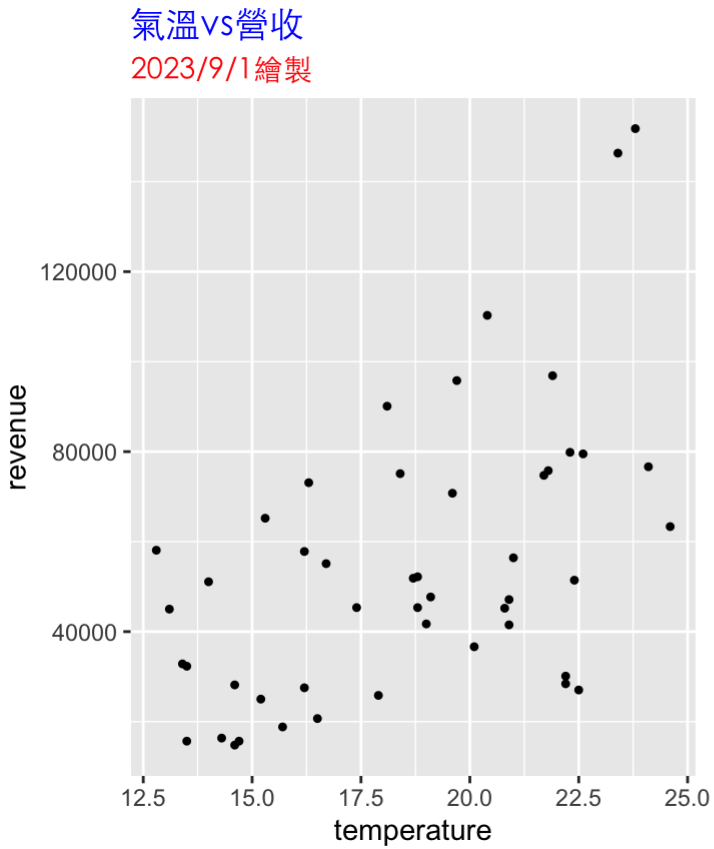
螢幕展示

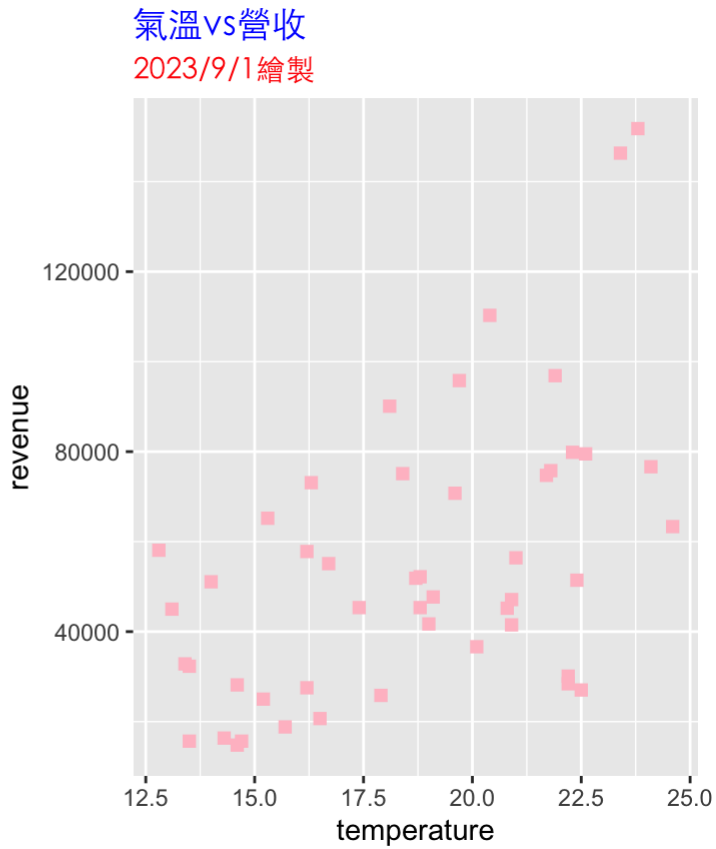
g # *打g可以在螢幕顯示散佈圖*



主體參數調整

```
g1<-g+geom_point(shape=15,colour="pink",size=2)
```





主體參數調整

- shape 點的圖示

- color 顏色
- size 大小調整

x 軸的控制-範圍與刻度控制

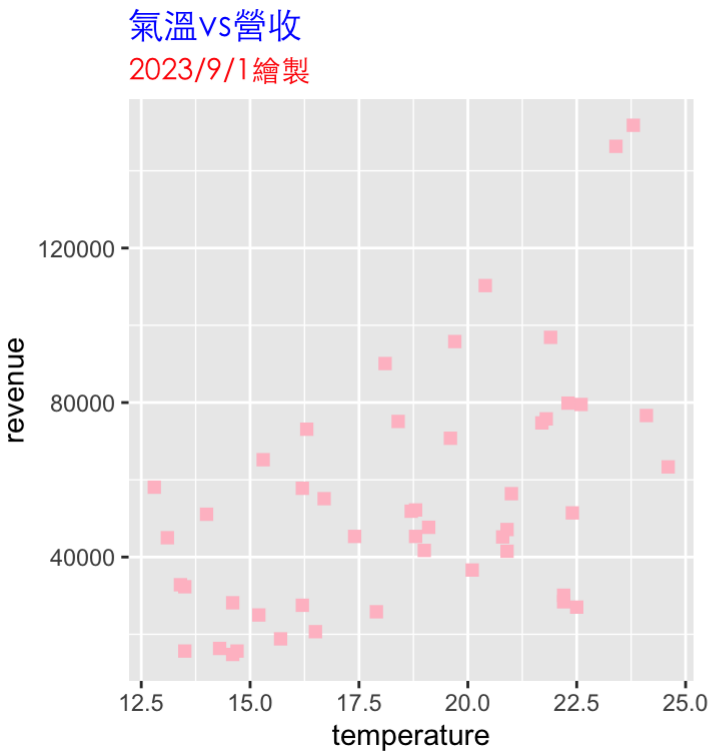
```
g2<-g1+scale_x_continuous(breaks=seq(12,25,by=1),limits=c(12,25))
```

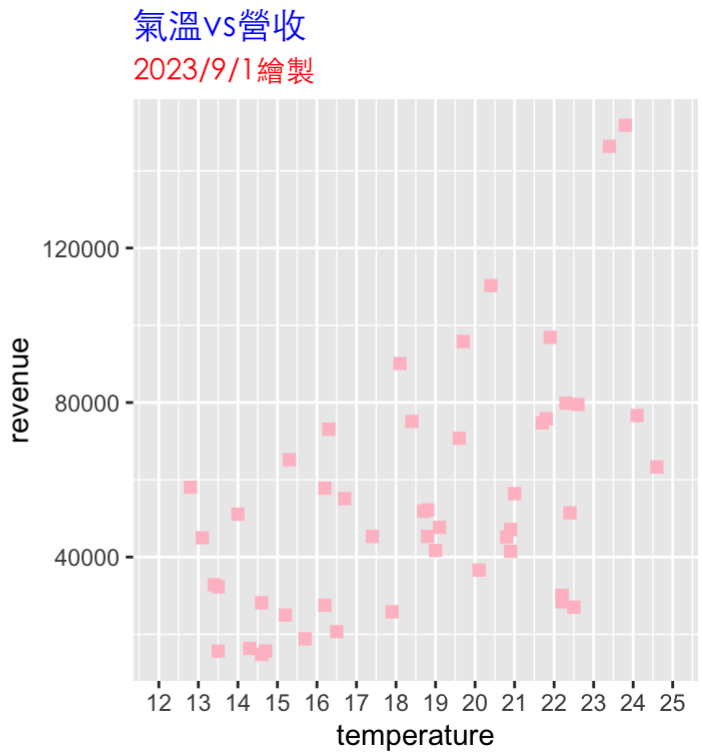
- 溫度是連續型資料 (continuous data), 所以使用scale_x_continuous()函數設定軸的參數。
- breaks (斷點): 設定斷點的引數 (argument)。
- limits (界線): 設定邊界的引數。
- seq(,,) 為生成等差序列函數。甚麼是等差序列呢?
    - seq(起始, 結束, 間隔)

```
seq(5, 10, 2) #分割點不超過10
```

```
## [1] 5 7 9
```

x 軸的控制-範圍與刻度控制



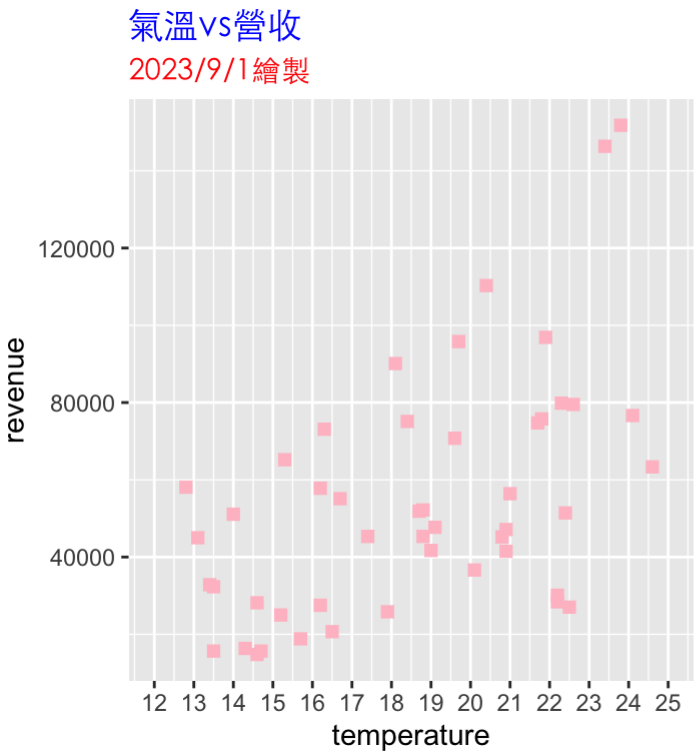


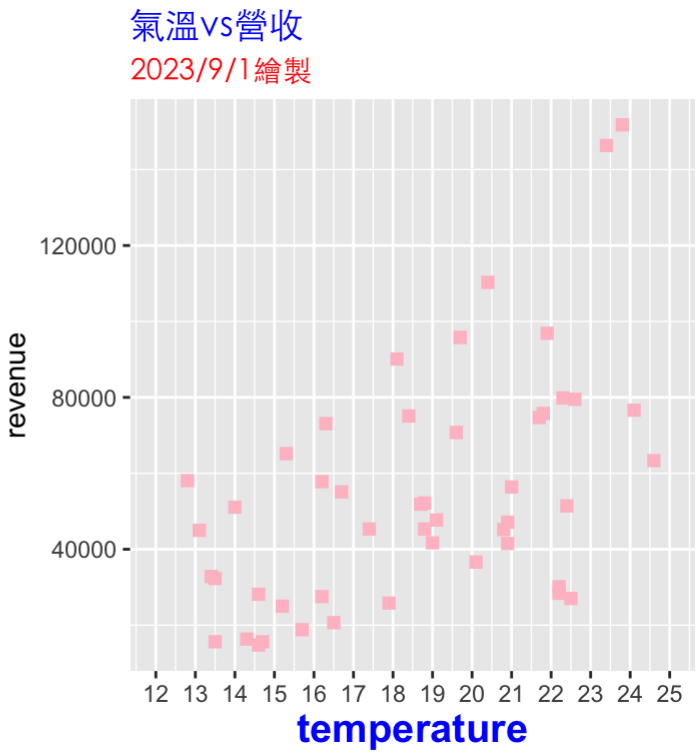
x 軸的控制-字體與顏色控制

```
g3<-g2+theme(axis.title.x=element_text(face=2,color="blue",size=15))
```

- axis.title.x 指定x軸標題
- element_text() 文字物件
  - face 字體
  - color 顏色
  - size 大小

x 軸的控制-字體與顏色控制



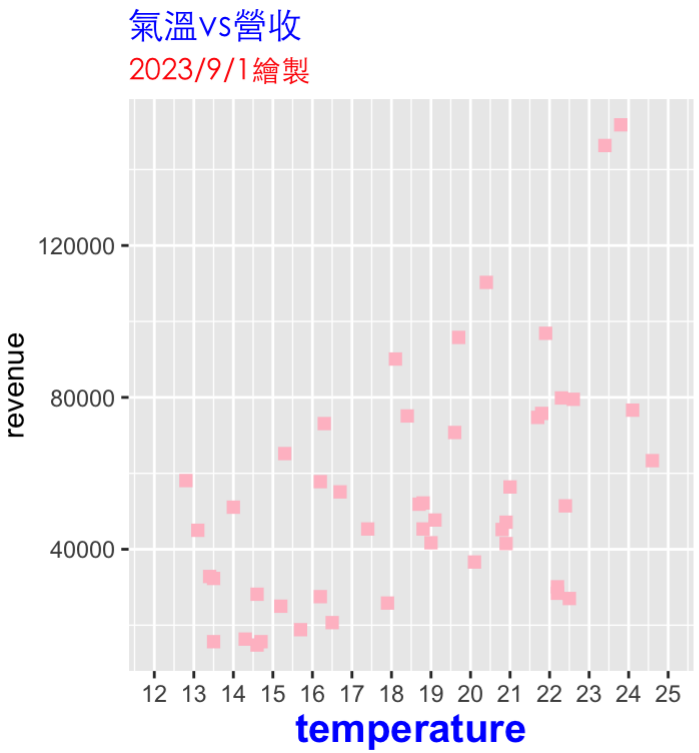


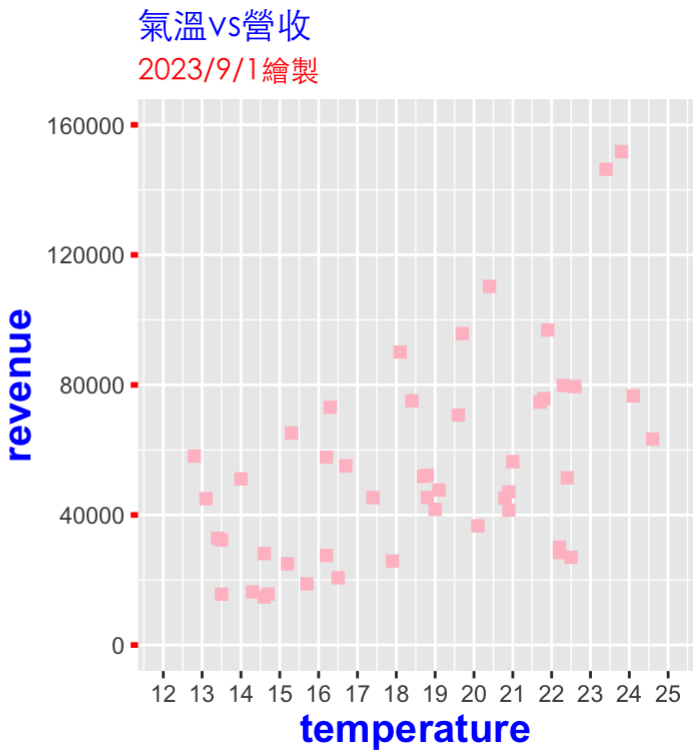
y 軸的控制-與X軸相同

```
ybreaks<-seq(0,16,by=4)*10000 #設定y軸斷點
g4<-g3+scale_y_continuous(breaks=ybreaks,limits=c(0,16)*10000)
g4<-g4+theme(axis.ticks.y =element_line(linewidth=1,color="red"),
             axis.title.y=element_text(face=2,color="blue",size=15))
```

- axis.ticks.y 選定y軸的刻度
- elment_line() 線屬性控制
  - linewidth 寬
  - color 顏色

y 軸的控制-與X軸相同

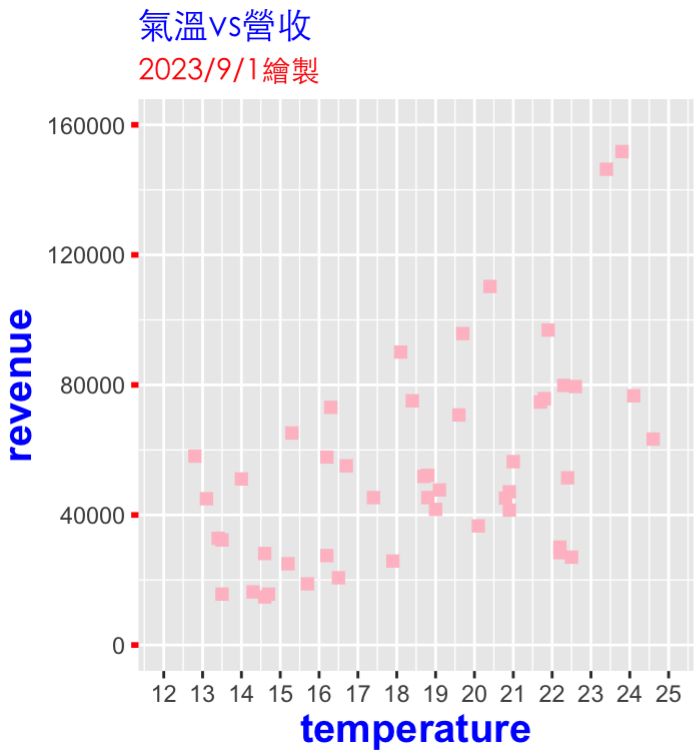

氣溫vs營收
2023/9/1繪製
120000
80000
40000
revenue
12 13 14 15 16 17 18 19 20 21 22 23 24 25
temperature

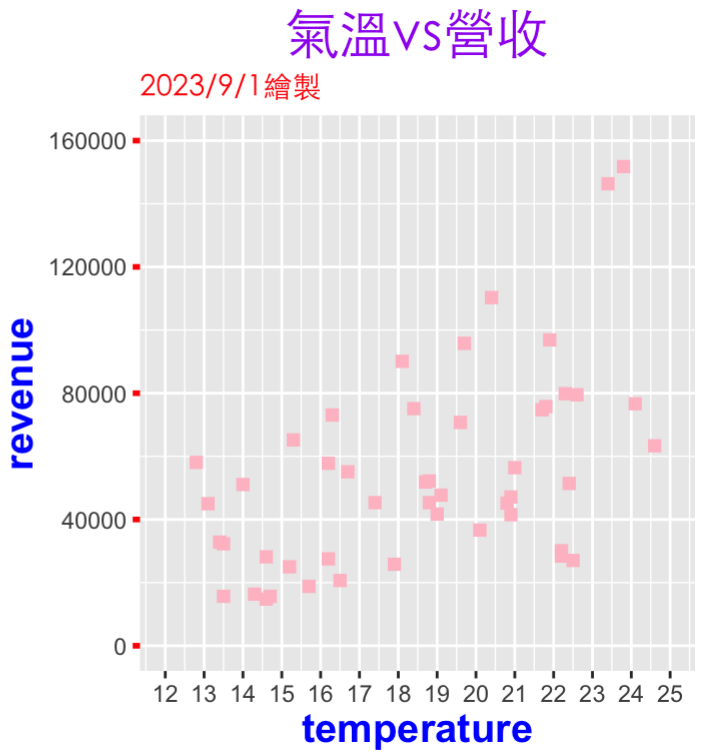

氣溫vs營收
2023/9/1繪製
160000
120000
80000
40000
0
revenue
12 13 14 15 16 17 18 19 20 21 22 23 24 25
temperature

標題控制

```
g5<-g4+theme(plot.title=element_text(face=3,
    color="purple",size=20,hjust=0.5))
```

標題控制


氣溫vs營收
2023/9/1繪製
160000
120000
80000
40000
0
revenue
12 13 14 15 16 17 18 19 20 21 22 23 24 25
temperature


氣溫vs營收
2023/9/1繪製
160000
120000
80000
40000
0
revenue
12 13 14 15 16 17 18 19 20 21 22 23 24 25
temperature

關聯性分析

1. 計算相關係數 (correlation)

```
r<-cor(sushi$temperature, sushi$revenue)
r
```

```
## [1] 0.5327026
```

- |r| < 0.4 低度線性相關
- 0.4 ≤ |r| < 0.7 中度線性相關
- |r| ≥ 0.7 高度線性相關

關聯性分析

1. 加入迴歸直線

假設營收(y) 與溫度(x)的模型為

$$y = \alpha + \beta x$$

α 為截距、β為斜率。

關聯性分析

・ 模型估計(revenue = α + β × temperature)

```
aa<-lm(sushi$revenue~sushi$temperature)
summary(aa)
```

```
##
## Call:
## lm(formula = sushi$revenue ~ sushi$temperature)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46413 -20018  -3983  15409  72050
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -35462      21409  -1.656    0.104
## sushi$temperature     4840       1134   4.269 9.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26570 on 46 degrees of freedom
## Multiple R-squared:  0.2838, Adjusted R-squared:  0.2682
## F-statistic: 18.23 on 1 and 46 DF,  p-value: 9.714e-05
```

關聯性分析

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -35462      21409  -1.656    0.104
sushi$temperature   4840       1134   4.269 9.71e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26570 on 46 degrees of freedom
Multiple R-squared:  0.2838,    Adjusted R-squared:  0.2682
F-statistic: 18.23 on 1 and 46 DF,  p-value: 9.714e-05
```

1. 變數temperature 對營收revenue 是否有影響的判定，此值(p_value)<0.05 判有影響。

2. 此模型的解釋變數為temperature, 反應變數為revenue。整個線性模型是否有解釋力，此值(p_value)<0.05 判有解釋力。因為這個模型只有temperature一個變數，所以1和2的值是相同的。不過，通常一個線性模型可能有很多解釋變數。

3. α, β 的估計值

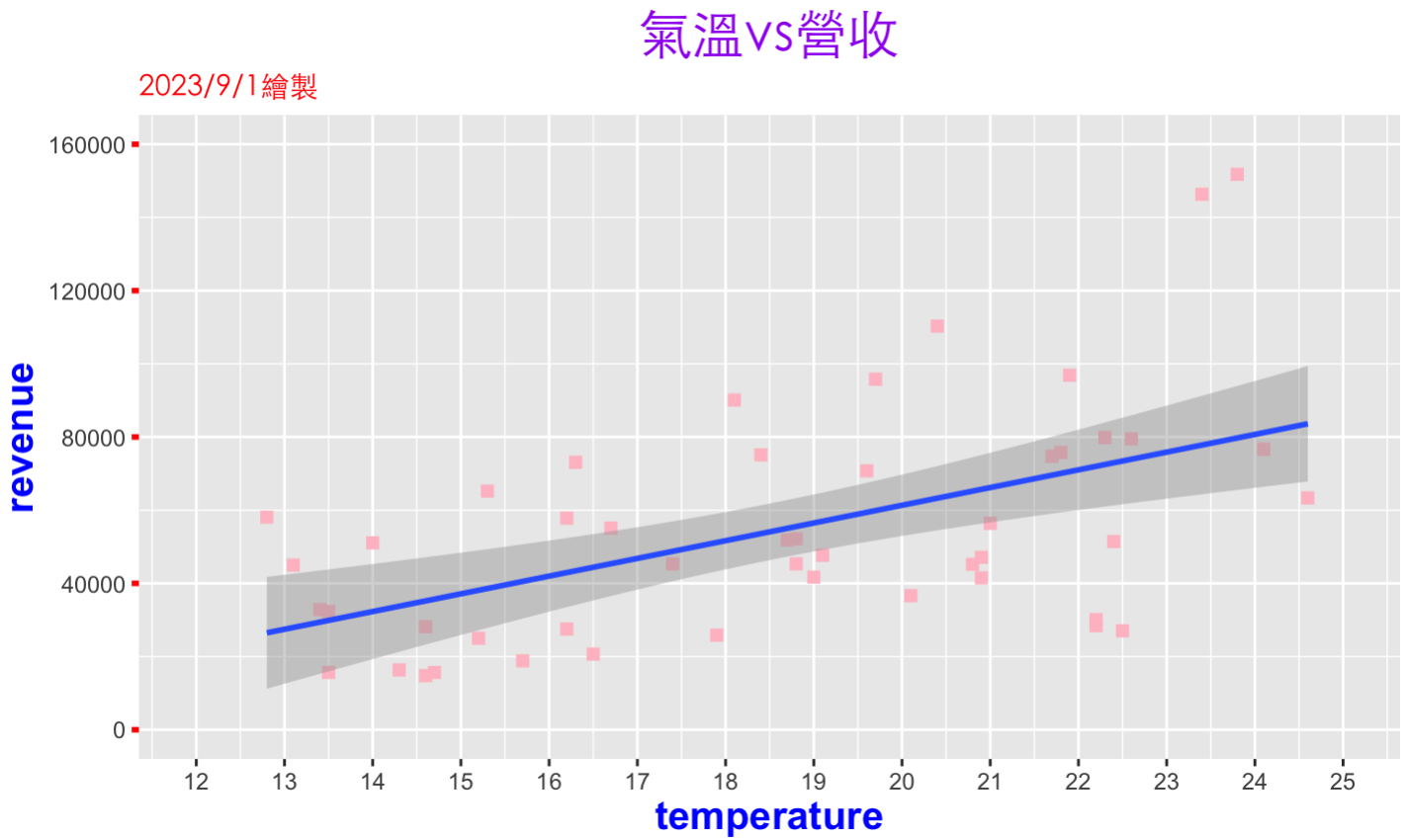
模型解釋

$$\text{revenue} = -35462 + 4840\text{temperature}$$

- 截距 α的估計值為-35462不具實質意義，可以不用特別說明。
- 斜率 β的估計值為4840代表氣溫增加一度C，預期營收(revenue)的**增幅**

加入迴歸直線

```
g5<-g5+geom_smooth(method='lm',formula=y~x,show.legend=TRUE) #加入平滑曲線, 表現整體趨勢
g5
```



散佈圖說明範本

- 主題說明

蒐集桃園區某珍先壽司店從3月到4月中旬氣溫與營收的資料共48筆, 這張散佈圖是用來說明該店營收與氣溫的相關性。

- 資料範圍
    - 解釋變數- 範圍、平均值、標準差 此圖顯示氣溫變化的範圍約在12和25之間，該店每日營收約在16萬元以內。
    - 反應變數– 範圍、平均值、標準差

散佈圖說明範本

- 趨勢(整體現象)
  整體看來，該店營收與氣溫有關，氣溫越高營收有越好的現象，呈現出一種線性相關。

- 異常情況(偏離整體現象)

  - 說明有無與趨勢不同的現象
  - 個別變數有無異常值(通常需要用盒型圖表現)

散佈圖說明-解釋變數(temperature)

```
a1<-max(sushi$temperature) #最大值
a2<-min(sushi$temperature) #最小值
a3<-round(mean(sushi$temperature),1) #平均值，四捨五入小數第一位
a4<-round(sd(sushi$temperature),1) #標準差，四捨五入小數第一位
print(c(a1,a2,a3,a4))
```

```
## [1] 24.6 12.8 18.6  3.4
```

2014年三月至四月中旬，桃園地區氣溫介於12.8到 24.6之間，平均氣溫約為 18.6，標準差為3.4，因為在春天氣溫變動大，日漸上升。(可以加入氣溫的趨勢圖)

散佈圖說明-反應變數(temperature)

```
a1<-max(sushi$revenue) #最大值
a2<-min(sushi$revenue) #最小值
a3<-round(mean(sushi$revenue),1) #平均值，四捨五入小數第一位
a4<-round(sd(sushi$revenue),1) #標準差，四捨五入小數第一位
a5<-round(sd(sushi$revenue)/mean(sushi$revenue)*100,1) #變異係數，四捨五入小數第一位
print(c(a1,a2,a3,a4,a5))
```

```
## [1] 151785.0  14780.0  54455.3  31065.4     57.0
```

2014年三月至四月中旬，本店每日營業額介於14780到 151785之間，平均每日營業額約為 54455.3，平均每日營業額標準差為31065.4，營業額的變異係數為57% 顯示每日營業額波動很大。(可以加入營業額的趨勢圖和直方圖)

散佈圖說明- 整體趨勢

- 氣溫高低會影響到營收，相關係數0.5327026 為中度線性相關。
- 根據營收與氣溫的線性模型，得知氣溫提高一度c，預期營收會增加 4840.216。