資料視覺化

Sung-Chien Lin

2018 年 9 月 15 日

課程簡介

課程簡介

- 本次課程的目的為介紹 R 語言上主要的資料視覺化工具– ggplot2
- 包括以下的內容：
  - 從 csv 檔讀入資料，儲存為 data frame 資料形態
  - 利用 tibble 整理資料成適合分析的形式
  - 進行簡單分析
- 本次課程將以空氣品質指標資料為例，進行資料視覺化

學習目標

- 能夠說明在資訊視覺化中各種圖表類型適合的應用問題
- 能夠將輸入的資料檔案轉換成可以進行視覺化的形式
- 能夠利用 ggplot2 套件進行資料視覺化

資料視覺化的基本概念

資料處理與分析的步驟

1. 問題擬定
2. 資料取得
3. 資料清理
4. 資料分析
5. 結果解釋

- 其中的第 4 和第 5 步驟都可以進行資料視覺化
- 第 4 步驟：進行資料的探索分析，概觀資料的樣貌，擬定相關資料模型
- 第 5 步驟：以圖形增強研究結果說明的印象和效用

視覺化常見的應用問題

- 分布(distribution)：某一數值資料的散佈情形
  - 觀察重點： 最大值、最小值、集中化、異常
- 比較(comparison)：比較某一類別資料在其對應數值上的差異
  - 觀察重點： 大小、順序、分布範圍
- 組成(compositon)：某一類別資料其對應數值的比例
  - 觀察重點： 比例大小
- 關係(relationship)：某兩個數值資料之間的關係
  - 觀察重點： 模式(直線、週期)、分布範圍
- 這些問題可以組合，例如比較類別的分布

資料視覺化

- 運用各種類型的圖表，以圖表上的位置、顏色、形狀、大小等視覺線索表現資料的分布、比較、組成和關係
- 分布：以直方圖表示某一數值資料的散佈情形
- 比較：以長條圖比較某一類別資料在其對應數值上的差異
- 組成：以圓餅圖表示某一類別資料其對應數值的比例
- 關係：以散佈圖或折線圖觀察某兩個數值資料之間的關係
  - 如果兩個數值資料中，有一個是時間資料，特別適合使用折線圖觀察
- 比較+分布：以盒狀圖或小提琴圖比較某一類別資料其對應數值的散佈情形

R 語言的資料視覺化工具 | ggplot2

- ggplot2 提供一套標準圖表與資訊視覺化的語法和套件
  - Data 資料來源
  - Aesthetics 視覺線索：位置、顏色、形狀、大小
  - Geometrics 圖表類型
  - Scales 資料編碼的表現形式
  - Coordinates 座標類型
  - Facets 圖表層面
  - Themes 圖表外觀

ggplot2 的運作概念

- 將做為資料來源的 data frame 上的每一筆紀錄，視為圖形上要呈現的一個圖形，
- 先將紀錄上的欄位對應到位置、顏色、形狀或大小等資料編碼方式，
- 然後選擇圖表類型，
- 設定各種資料編碼方式的樣式，例如座標的尺度、調色板等，
- 必要時調整圖形的座標類型、圖表層面和圖表外觀。

幾種常用的 ggplot2 資料編碼

- x：x 軸位置
- y：y 軸位置
- color：點、線的顏色
- fill：填入的顏色
- size：大小
- linetype：線的樣式
- shape：點的樣式

幾種常用的 ggplot2 圖表類型

- geom_histogram()：直方圖
- geom_col()：長條圖
- geom_line()：折線圖
- geom_point(), geom_jitter(), geom_count()：散佈圖、點狀圖
- geom_boxplot()：盒狀圖
- geom_violin()：小提琴圖
- geom_area()：區域圖
- geom_tile()：方塊圖

預備工作

準備工作目錄與檔案

- 在 rCourse 下，建立工作目錄 04
- 將 02 的空氣品質指標資料檔案複製到 04 下

設定工作目錄

- 首先開啟新的 Script
- 在 Script 上，設定工作目錄

```
setwd("rCourse/04")
```

使用 ggplot2 套件

- ggplot2 是 tidyverse 中包含的一個套件
- 可以使用 library(tidyverse)載入
- 在 Script 上輸入

```
library(tidyverse)
```

- 或單獨使用 library(ggplot2)載入

```
library(ggplot2)
```

- 建議採用 library(tidyverse)，可以一次載入多個資料處理套件

讀取粉絲專頁發文資料 csv 檔案

- 在 Script 上輸入

```
aqi_data <- read.csv(file="AQI_20180128061645.csv",
                     fileEncoding="UTF-8-BOM", stringsAsFactors=FALSE)
```

讀取縣市地區對照資料

- 在 Script 上輸入

```
ca <- read.csv(file="county_area.csv",
                     fileEncoding="UTF-8", stringsAsFactors=FALSE)
```

- 注意：county_area.csv 與 AQI_20180128061645.csv 的編碼方式不同，前者是 UTF-8 without BOM，後者則是 UTF-8 with BOM

將地區資料加入空氣品質指標資料

```
aqi_data <- aqi_data %>%
  left_join(ca)
```

將部分 Character 型態的資料欄位改成 Factor

```
aqi_data <- aqi_data %>%
  mutate(County=factor(County)) %>%
  mutate(Status=factor(Status, levels=c("良好", "普通", "對敏感族群不健康
","對所有族群不健康"), ordered=TRUE)) %>%
  mutate(Area=factor(Area))
```

本次課程範例

以資料視覺化的方式回答以下的問題

- 全台各地空氣品質指標的分布情形為何？
- 各地區(北、中、南、東、離島)的觀測站數量與比例為何？
- 各地區觀測站測得的空氣品質分布情形為何？
- 各地區觀測站測得的 PM10 與 PM2.5 分布
- PM2.5 指標與 AQI 之間的關係為何？

全台各地空氣品質指標分布情形的視覺化

先了解各地空氣品質指標的統計摘要

- 利用統計摘要 summary()描述資料分布
- 在 Script 上輸入

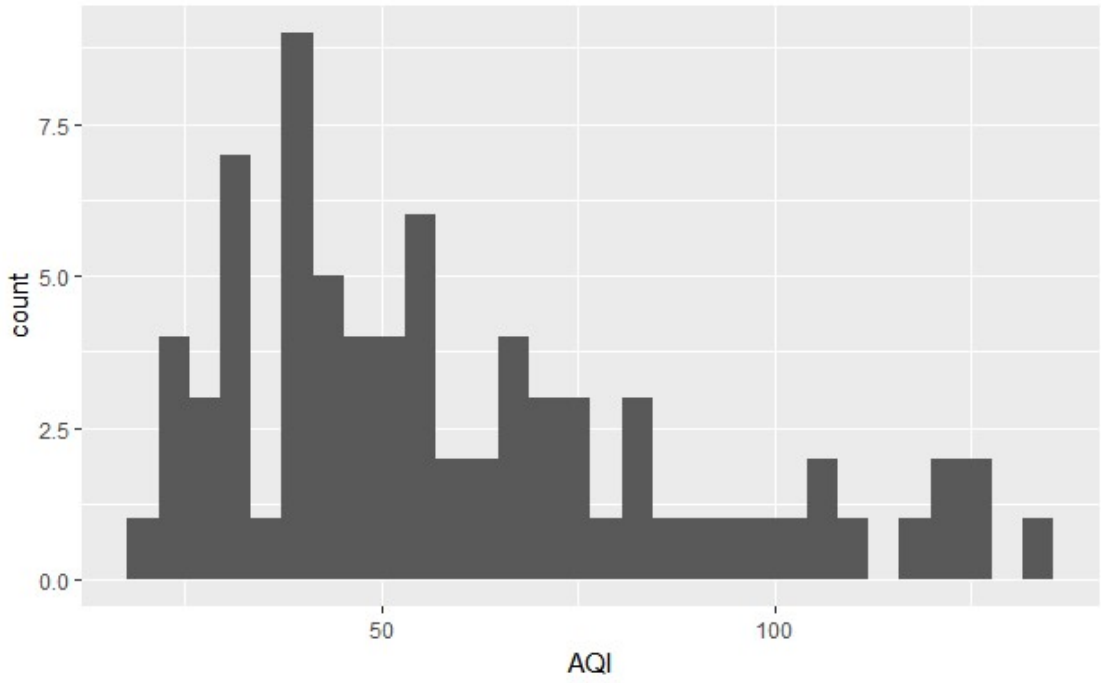
```
aqi_data %>%

  select(AQI) %>%          # 選擇AQI 值欄位

  summary()                # 對AQI 值欄位資料統計摘要
```

```
      AQI
 Min.   : 19.00
 1st Qu.: 39.75
 Median : 53.50
 Mean   : 60.11
 3rd Qu.: 73.50
 Max.   :133.00
 NA's   :1
```

資料分布一般採用直方圖

- 在篩選非 NA 的 AQI 資料(刪除沒有內容的 AQI 資料)之後，以直方圖繪製 AQI 分布
- 在 Script 上輸入
- ggplot()：呼叫 ggplot2 繪圖程式
- geom_histogram()：產生直方圖

```
aqi_data %>%
  filter(!is.na(AQI)) %>%     # 將沒有內容的AQI 資料刪除
  ggplot(aes(AQI)) +          #以直方圖繪製AQI 分布
  geom_histogram()
```

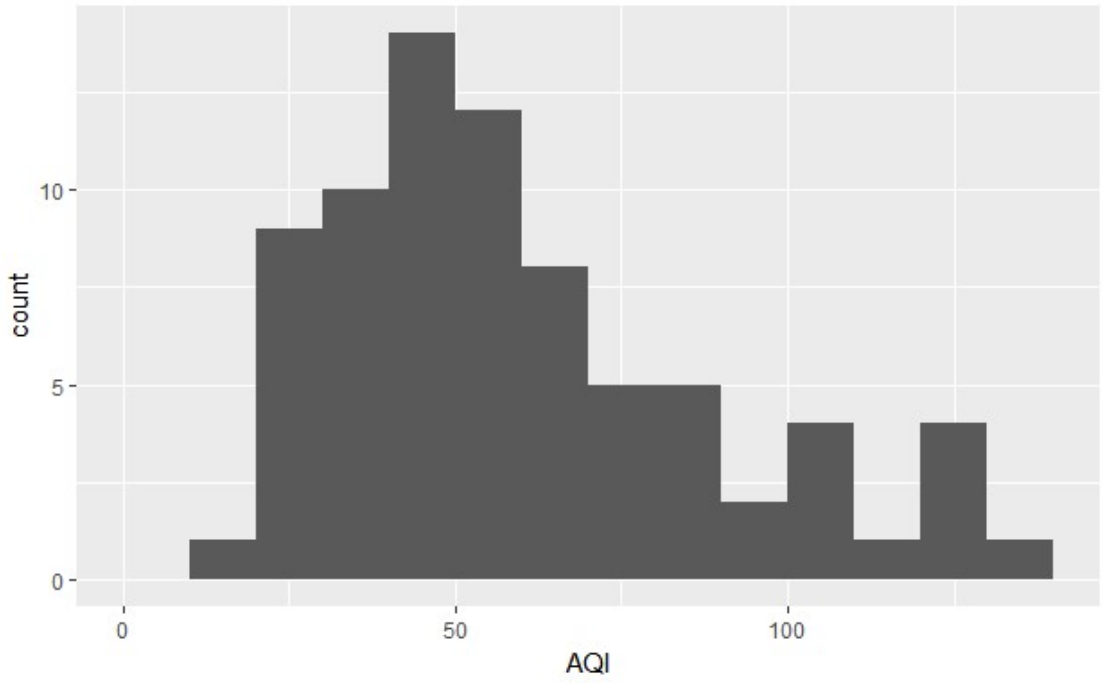


- ggplot2 的直方圖 geom_histogram()
  - 將 AQI 值分為若干區間(例如：0~5, 6~10, 11~15, ...)
    - 預設為 30 個區間
  - 統計 AQI 值在各區間內的觀測站數

設置區間範圍

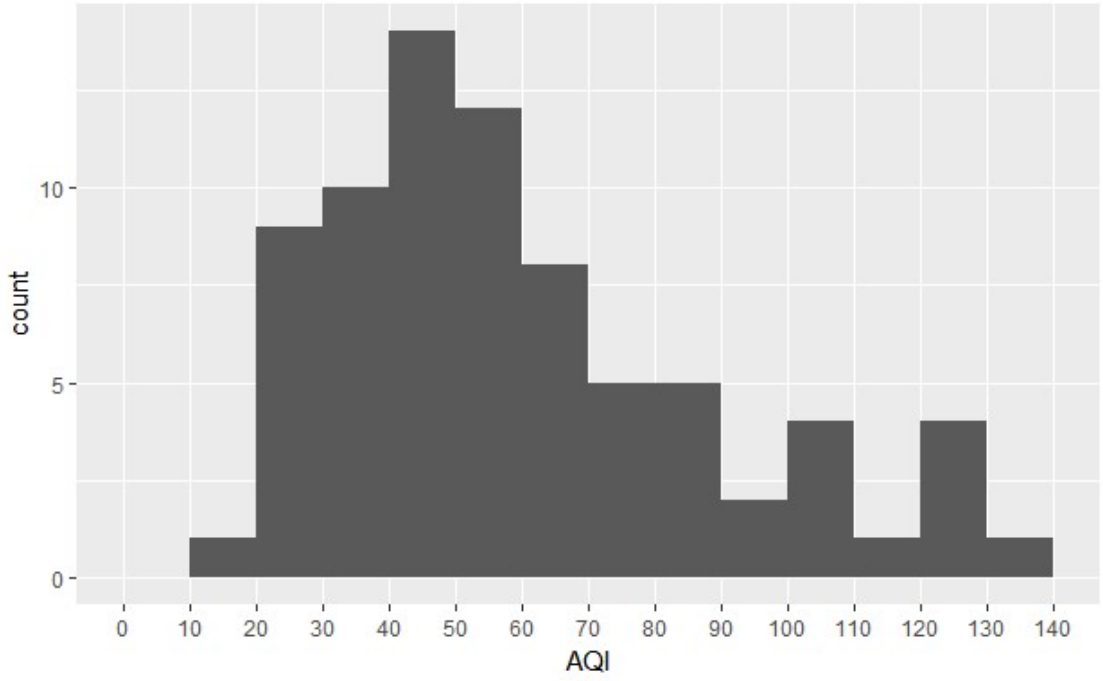
- 將區間的寬度設為 10
- 改寫前段的程式

```
x_breaks <- seq(0, max(aqi_data$AQI, na.rm=TRUE)+10, 10) # 產生劃分資料
的區間


aqi_data %>%
  filter(!is.na(AQI)) %>%     # 將沒有內容的 AQI 資料刪除
  ggplot(aes(AQI)) +
  geom_histogram(breaks=x_breaks)
```

- 

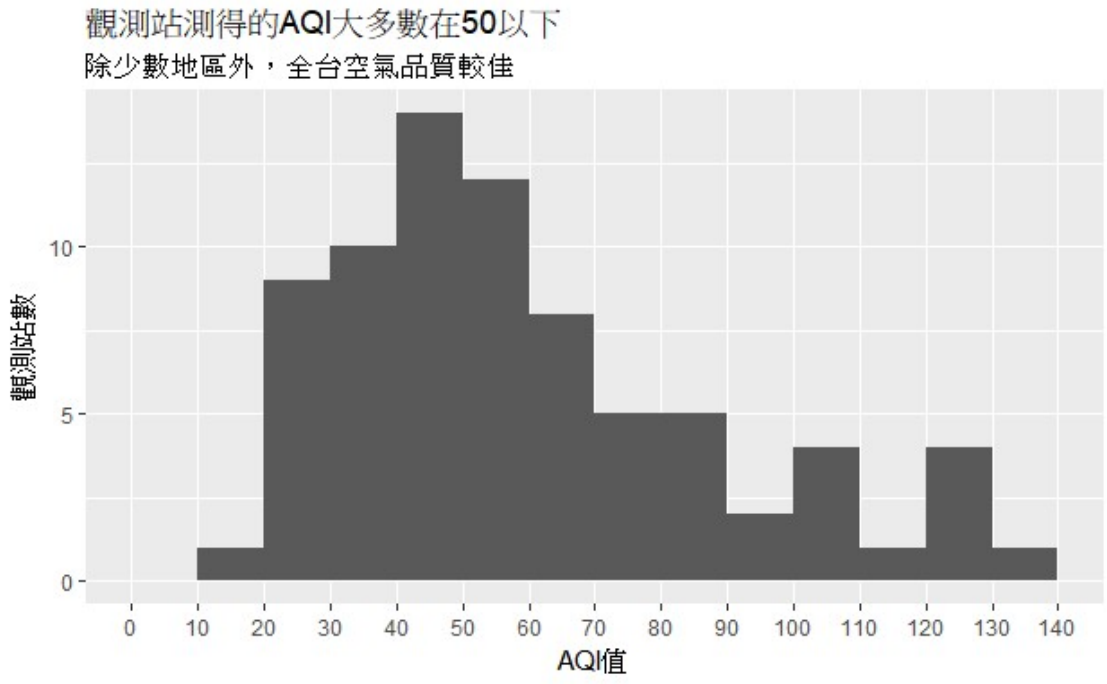
加入 X 軸座標

```
aqi_data %>%
  filter(!is.na(AQI)) %>%     # 將沒有內容的 AQI 資料刪除
  ggplot(aes(AQI)) +
  geom_histogram(breaks=x_breaks) +
  scale_x_continuous(breaks = x_breaks, minor_breaks=NULL)
```

- 

修改與加入標題

```
aqi_data %>%
  filter(!is.na(AQI)) %>%     # 將沒有內容的 AQI 資料刪除
  ggplot(aes(AQI)) +
  geom_histogram(breaks=x_breaks) +
  scale_x_continuous(breaks = x_breaks, minor_breaks=NULL) +
  labs(x="AQI 值", y="觀測站數",
       title="觀測站測得的 AQI 大多數在 50 以下",
       subtitle="除少數地區外，全台空氣品質較佳")
```



練習

- 將直方圖的 y 軸座標(觀測站數)主要間隔設為 1

練習

- 所有的 PM2.5 資料分布情形

簡要複習

- ggplot2 如何畫出 data frame 上的每一筆紀錄：
  - ggplot(aqi_data, aes(AQI))：使用資料來源 aqi_data。根據 AQI 的值，決定記錄在 x 軸上的位置。
  - geom_histogram(breaks=x_breaks)以直方圖做為圖表。根據 AQI 的值，將所有紀錄指派到 x_breaks 表示的各個區間上，以每一個區間上的記錄數量為這個區間的高度。
- scale_x_continuous(breaks=x_breaks, minor_breaks=NULL)將圖表的主要刻度設為 x_breaks 上的各個區間，並且不要次要刻度。

各地區的觀測站數量與比例為何？

統計各地區的觀測站數量

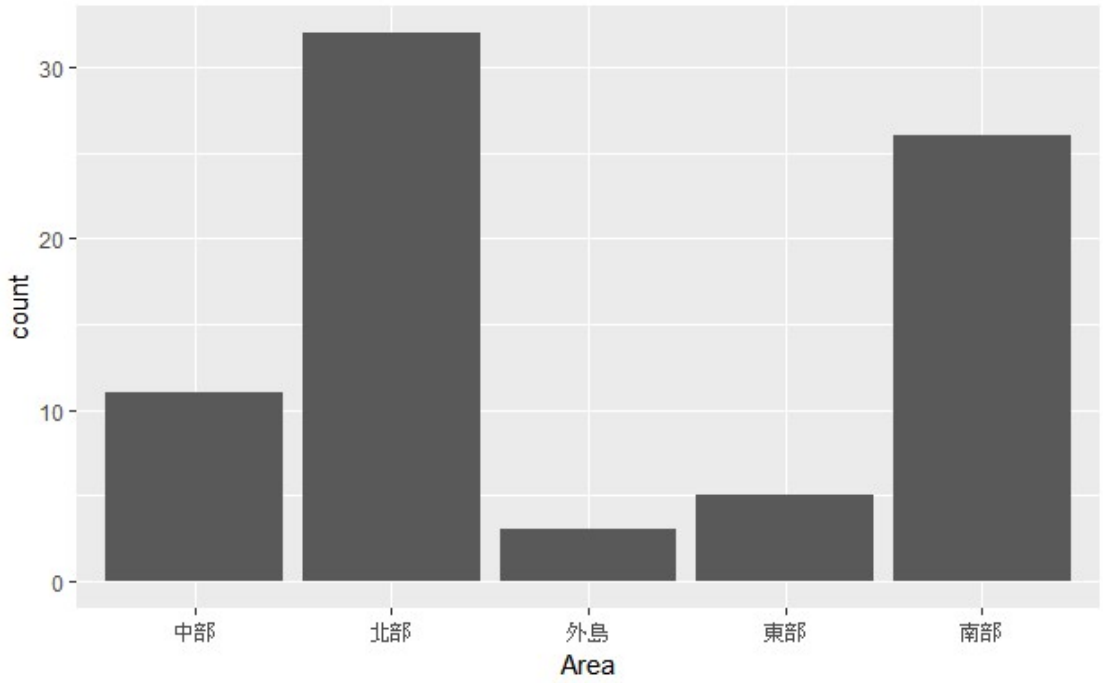
- 將紀錄依據地區分類，統計各地區上的觀測站數量，形成新的 data frame。
- 新的 data frame 上每一筆紀錄對應到一個地區以及它的觀測站數量。

```
area_data <- aqi_data %>%
  group_by(Area) %>%
  summarise(count=n())
```

運用 ggplot 的長條圖繪製各地區的觀測站數量

- ggplot(area_data, aes(x=Area, y=count))
  - 以 area_data 做為資料來源，比較每一個地區的觀測站數量。
  - 地區放置在圖形的 x 軸上
  - 每一筆紀錄的發文數量放置在圖形的 y 軸上
- geom_col()長條圖

```
area_data %>%
  ggplot(aes(x=Area, y=count)) +
  geom_col()
```

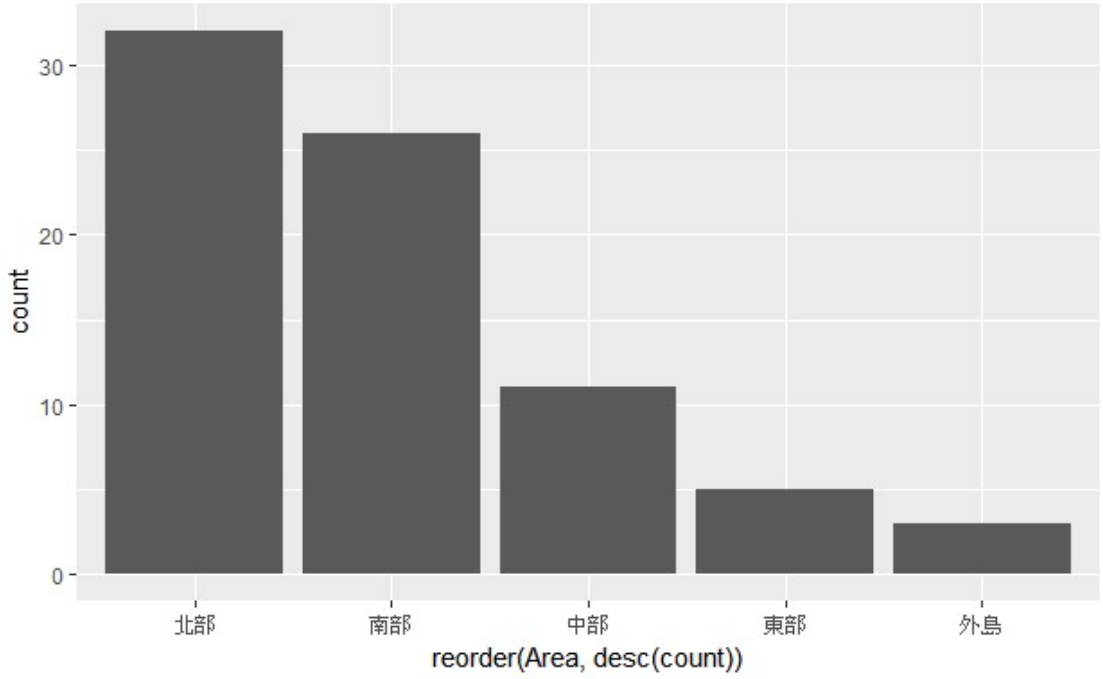


-
- 最好將長條圖進行排序

加上排序的長條圖

- 利用 ggplot2 繪製長條圖時，X 軸改為依據 count 降冪排序後的地區。
- 在 Script 上加入下面指令

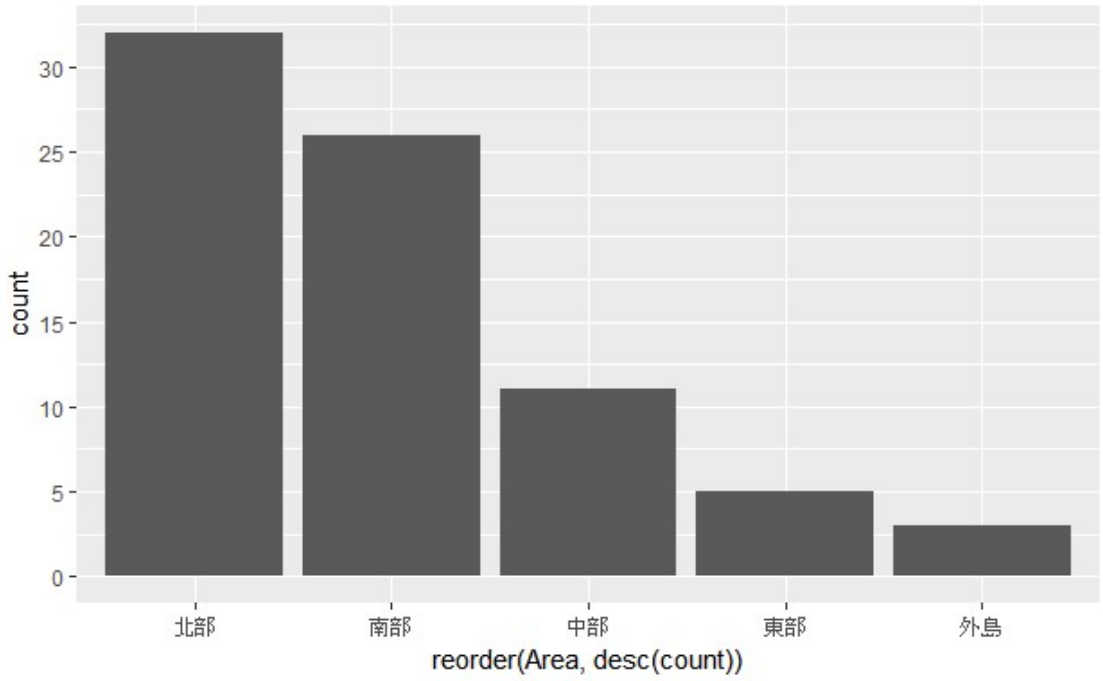
```
area_data %>%

  ggplot(aes(x=reorder(Area, desc(count)), y=count)) +

  geom_col()
```

- 

修改 y 軸上的標示

- 修改 Script，加上 scale_y_continuous()修改 y 軸上的標示

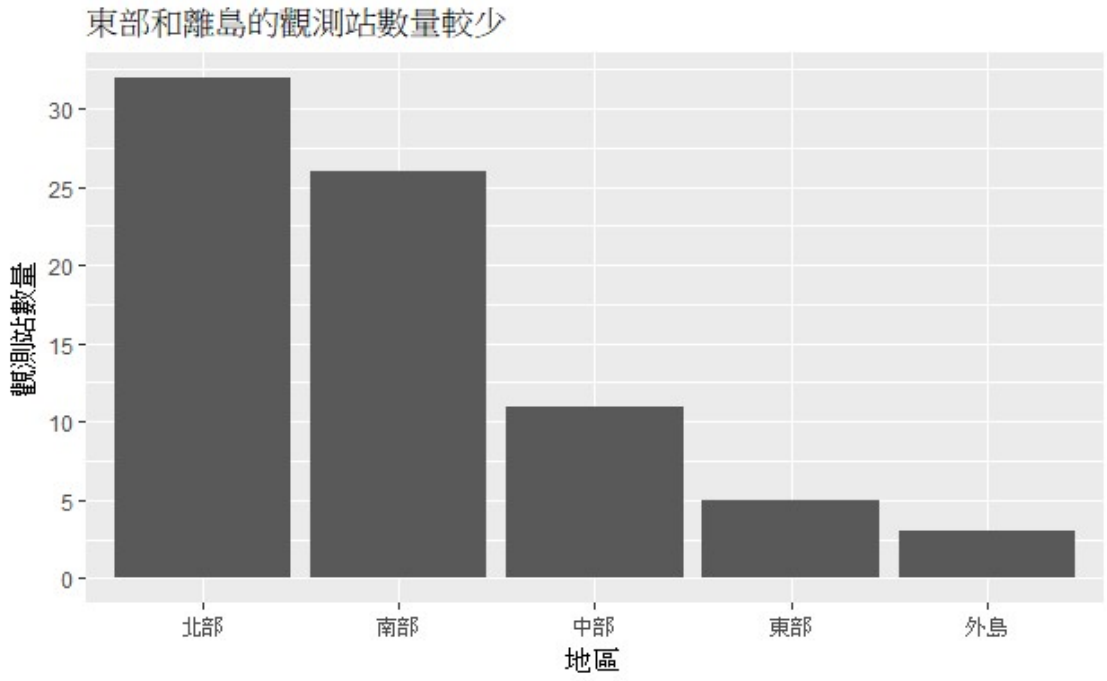
```
area_data %>%
  ggplot(aes(x=reorder(Area, desc(count)), y=count)) +
  geom_col() +
  scale_y_continuous(breaks=seq(0, max(area_data$count)+5, 5))
```

- 

加上標題

- 修改 Script，利用 labs()加上標題

```
area_data %>%
  ggplot(aes(x=reorder(Area, desc(count)), y=count)) +
  geom_col() +
  scale_y_continuous(breaks=seq(0, max(area_data$count)+5, 5)) +
  labs(x="地區", y="觀測站數量", title="東部和離島的觀測站數量較少")
```

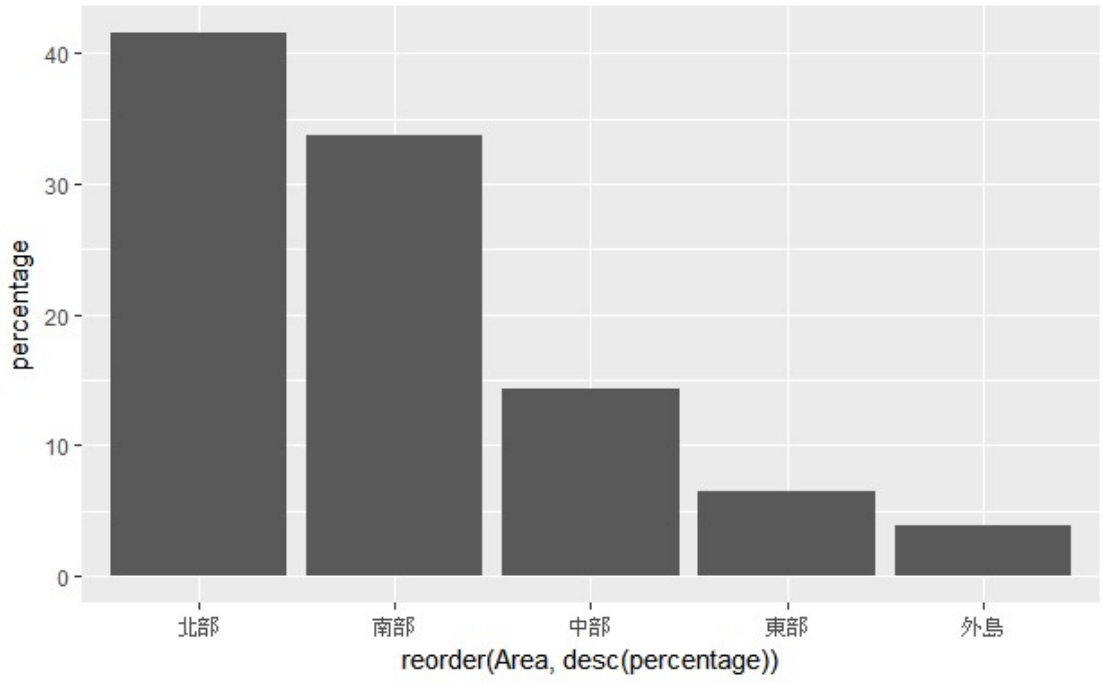
- 

統計各類型的發文比例

```
area_data <- area_data %>%
  mutate(percentage=round(count/sum(count)*100, 2))
```

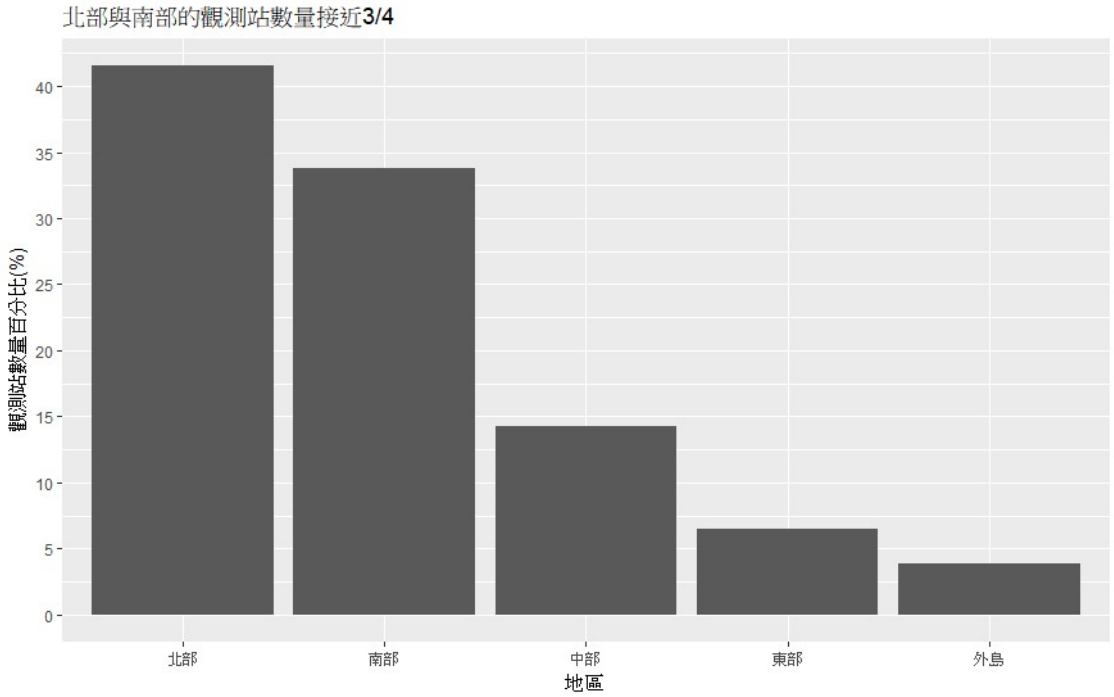
畫出長條圖

```
area_data %>%
  ggplot(aes(x=reorder(Area, desc(percentage)), y=percentage)) +
  geom_col()
```



練習

- 修改 y 軸標示與加上標題



簡要複習

- 長條圖需要指定 data frame 上的兩個欄位
  - x 軸為名目尺度的資料型態 Area
  - y 軸為連續尺度的資料型態 count 或 percentage
- 因為長條圖用於比較記錄之間在某一個欄位的大小以及表現它們的次序，所以可以加上大小排序

各地區觀測站測得的空氣品質分布情形為何？

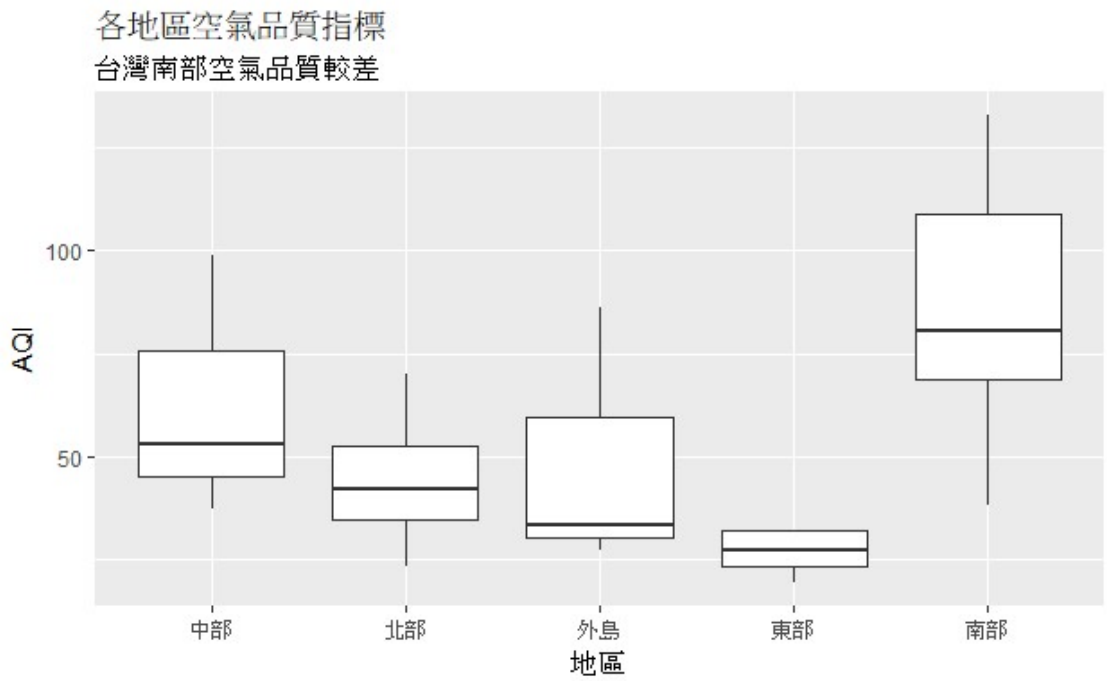
比較各地區觀測站測得的空氣品質分布

- 比較+分布
- 在一張圖上，同時顯示多個分布圖
  - 將不同地區放置於橫軸上的各個部分
  - 縱軸表示分布情形

使用盒狀圖畫出各地區觀測站測得的空氣品質分布

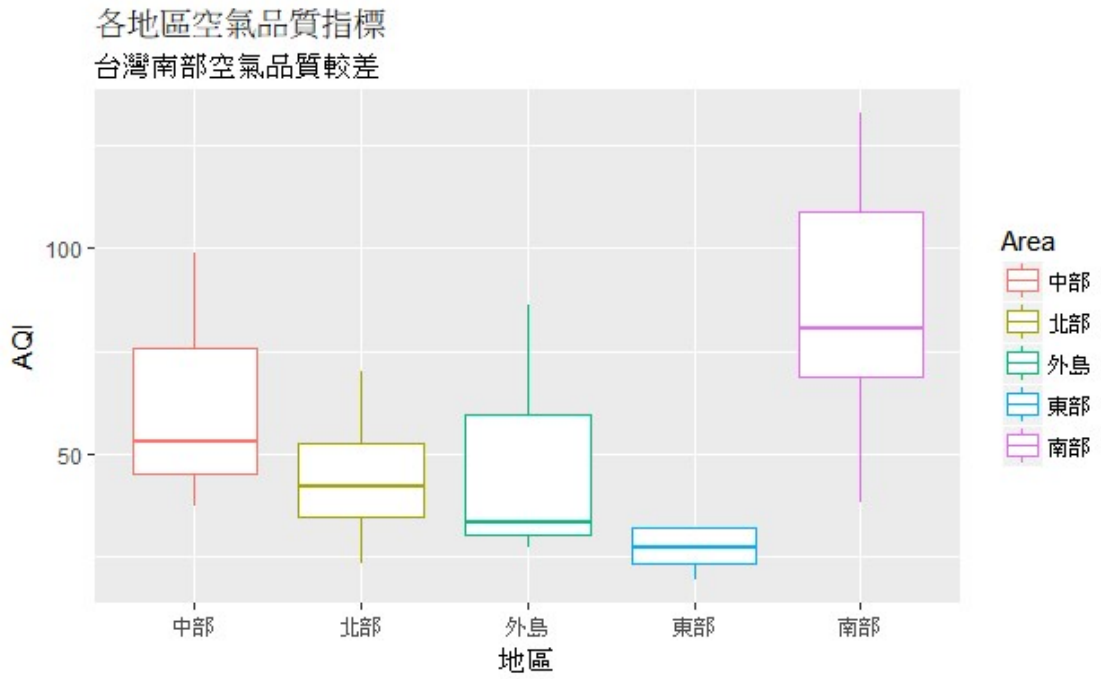
- geom_boxplot()盒狀圖
- 盒子中的橫線表示中位數，上下緣分別代表第三和第一四分位數
- 盒子的高是四分位距，代表資料的分散程度，盒子愈高表示資料愈分散
- 盒子上下的直線外的點表示極端值或離群值

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  ggplot(aes(x=Area, y=AQI)) +
  geom_boxplot() +
  labs(x="地區", y="AQI", title="各地區空氣品質指標", subtitle="台灣南部空氣品質較差")
```



以不同的顏色呈現各地區的 AQI 分布

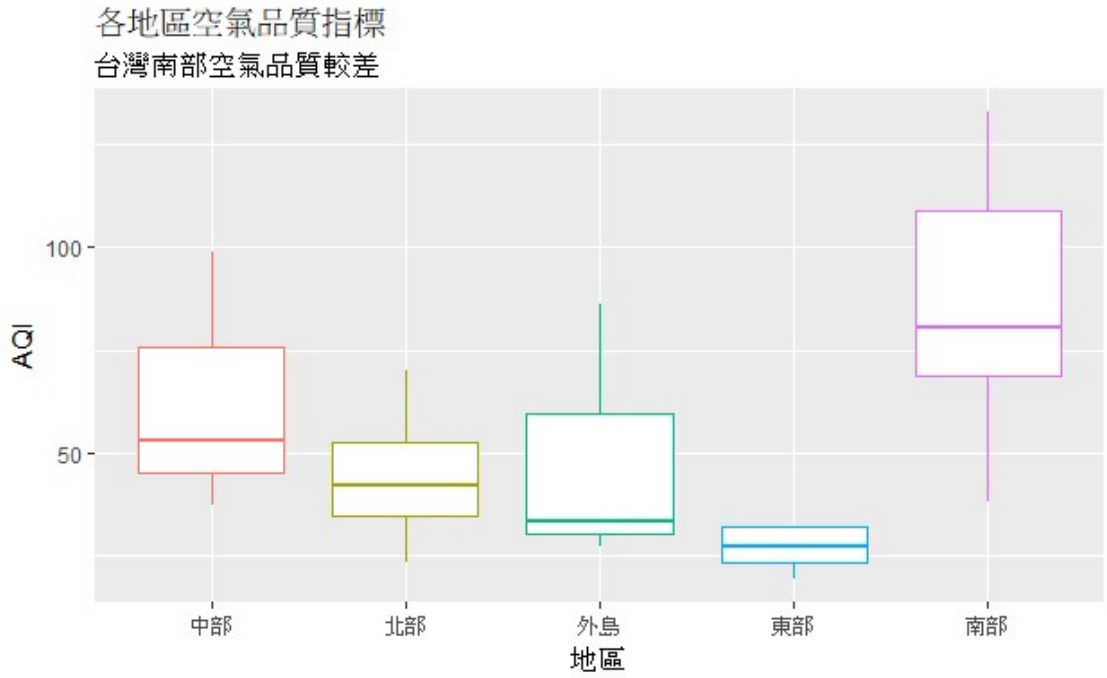
```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  ggplot(aes(x=Area, y=AQI, color=Area)) +
  geom_boxplot() +
  labs(x="地區", y="AQI", title="各地區空氣品質指標", subtitle="台灣南部空氣品質較差")
```



取消圖例(legend)

- theme(legend.position="none")取消圖例

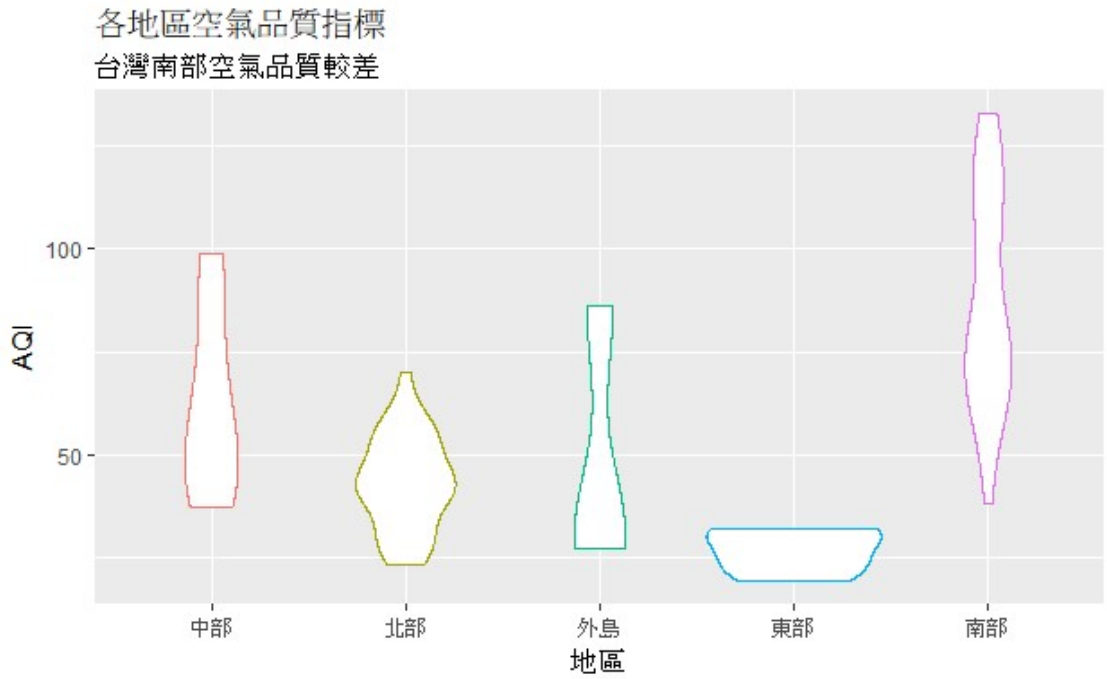
```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  ggplot(aes(x=Area, y=AQI, color=Area)) +
  geom_boxplot() +
  labs(x="地區", y="AQI", title="各地區空氣品質指標", subtitle="台灣南部空氣品質較差") +
  theme(legend.position="none")
```

- 

使用小提琴圖畫出各地區觀測站測得的空氣品質分布

- geom_violin()小提琴圖
- 小提琴的寬度表示資料在這個數值的分布情形，愈高的話分布的資料愈多

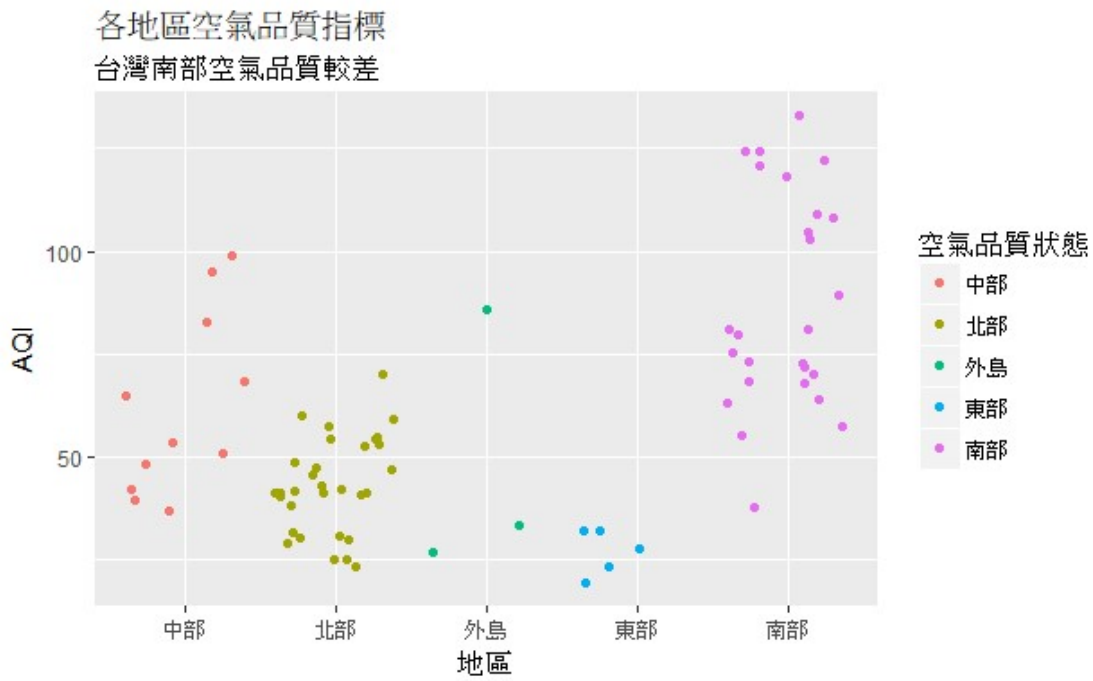
```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  ggplot(aes(x=Area, y=AQI, color=Area)) +
  geom_violin() +
  labs(x="地區", y="AQI", title="各地區空氣品質指標", subtitle="台灣南部空氣品質較差") +
  theme(legend.position="none")
```

- 

使用點狀圖畫出各地區觀測站測得的空氣品質分布

*以各發文類型為 x 軸，將每一筆紀錄依據它們的按讚次數在相對應的發文類型上畫出一個點

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  ggplot(aes(x=Area, y=AQI, color=Area)) +
  geom_jitter() +
  labs(x="地區", y="AQI", color="空氣品質狀態", title="各地區空氣品質指標
", subtitle="台灣南部空氣品質較差")
```

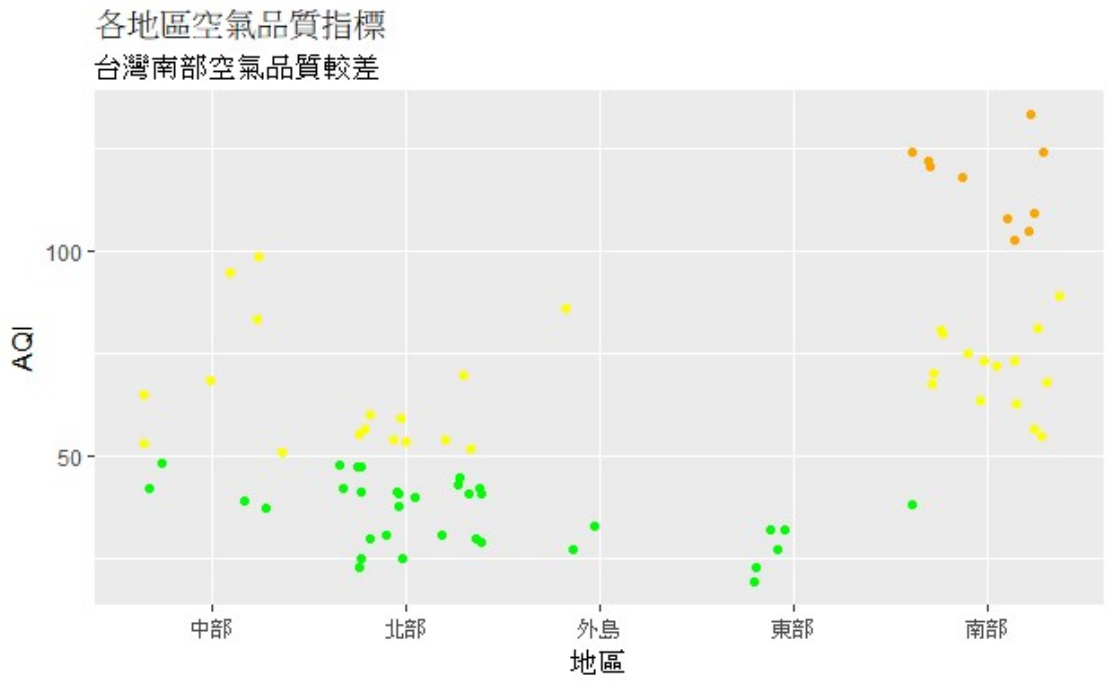


-
- 不用 geom_point()的原因是圖形上的點若是 x 和 y 軸座標相同的話，會重疊在一起
- geom_jitter()當 x 和 y 軸座標相同的話，會盡量繪製在附近，避免重疊。

修改顏色

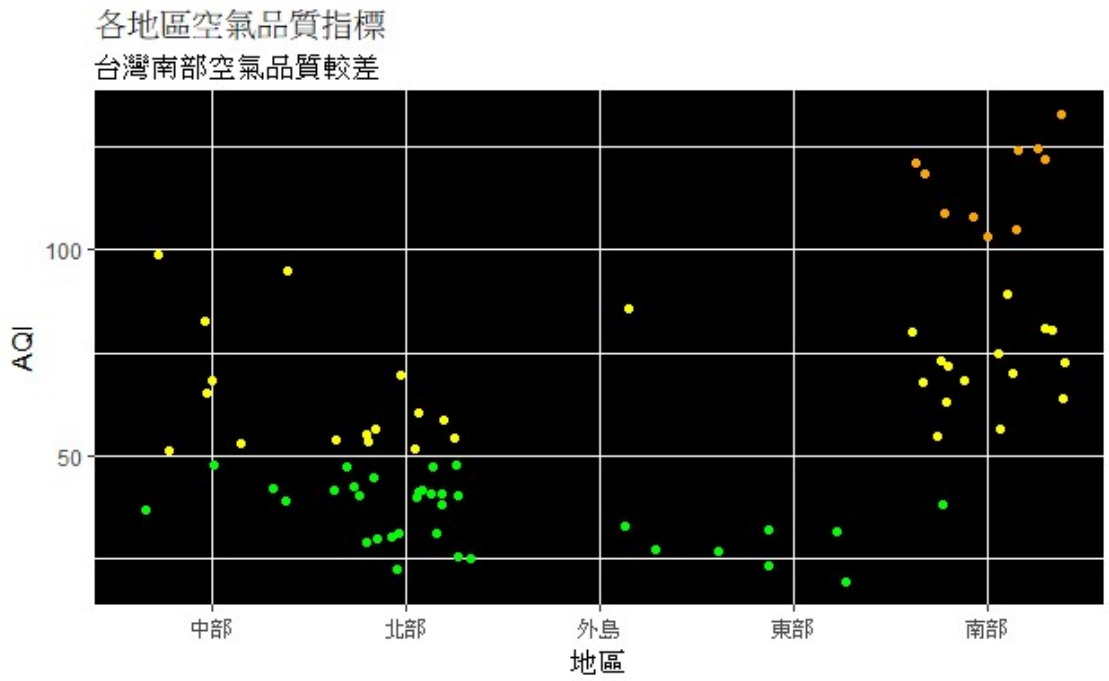
- 將每個地區觀測站的 AQI，依照不同狀態呈現
- 良好：綠色，普通：黃色，對敏感族群不健康：橙色

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  mutate(StatusColor=case_when(
    Status=="良好"~ "green",
    Status=="普通"~ "yellow",
    Status=="對敏感族群不健康"~ "orange"
  )) %>%
  ggplot(aes(x=Area, y=AQI, color=StatusColor)) +
  geom_jitter() +
  labs(x="地區", y="AQI", title="各地區空氣品質指標", subtitle="台灣南部空氣品質較差") +
  scale_color_identity()
```

- 
各地區空氣品質指標
台灣南部空氣品質較差
AQI
100
50
中部
北部
外島
東部
南部
地區

修改背景顏色

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  mutate(StatusColor=case_when(
    Status=="良好"~ "green",
    Status=="普通"~ "yellow",
    Status=="對敏感族群不健康"~ "orange"
  )) %>%
  ggplot(aes(x=Area, y=AQI, color=StatusColor)) +
  geom_jitter() +
  labs(x="地區", y="AQI", color="空氣品質狀態", title="各地區空氣品質指標
", subtitle="台灣南部空氣品質較差") +
  scale_color_identity()+
  theme(panel.background=element_rect(fill="black"))
```

- 
各地區空氣品質指標
台灣南部空氣品質較差
AQI
100
50
中部
北部
外島
東部
南部
地區

各地區觀測站測得的 PM10 與 PM2.5 分布

並列的圖表

- 將多個相同類型的圖表並列
- 以下以各地區觀測站的 PM10 與 PM2.5 資料為例，在並列的圖表上畫出表示兩個指標分布的盒狀圖

取出各地區觀測站測得的 PM10 與 PM2.5

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  select(Area, SiteName, PM10, PM2.5)
```

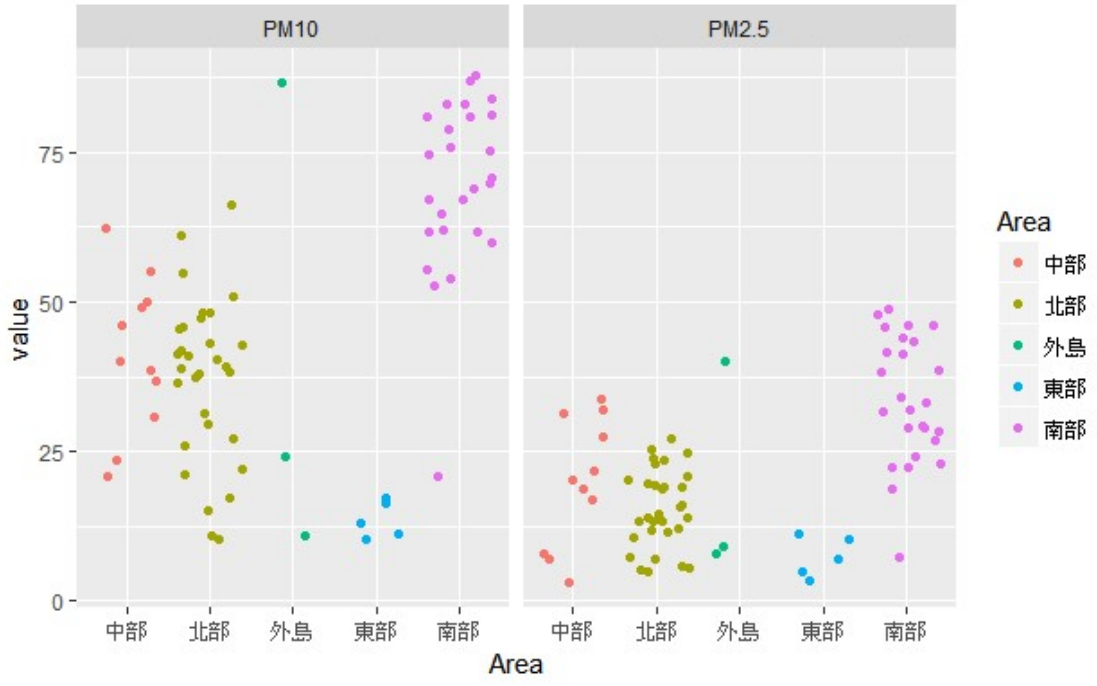
從 wide data format 改為 long data format

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  select(Area, SiteName, PM10, PM2.5) %>%
  gather(key=index, value=value, PM10, PM2.5)
```

利用圖表層面將各地區觀測站測得的 PM10 與 PM2.5 分布畫在並列的圖表上

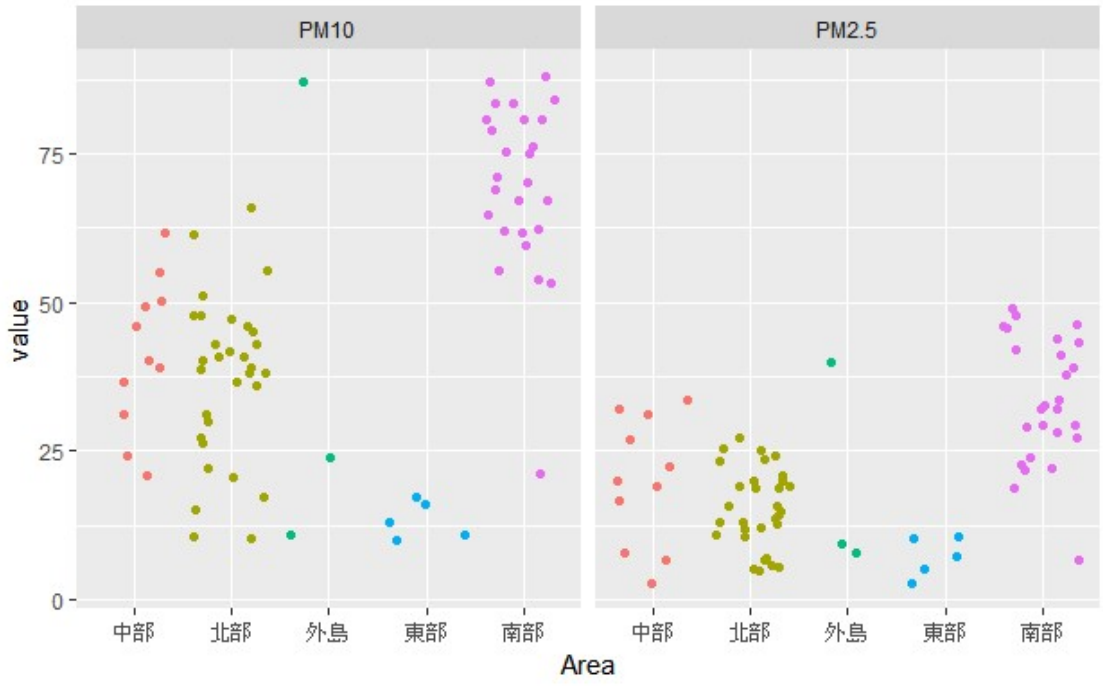
- 盒狀圖類似上面的做法，但利用 facet_wrap(~index)，將兩種指標的盒狀圖並列

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  select(Area, SiteName, PM10, PM2.5) %>%
  gather(key=index, value=value, PM10, PM2.5) %>%
  ggplot(aes(x=Area, y=value, color=Area)) +
  geom_jitter() +
  facet_wrap(~index)
```

- 

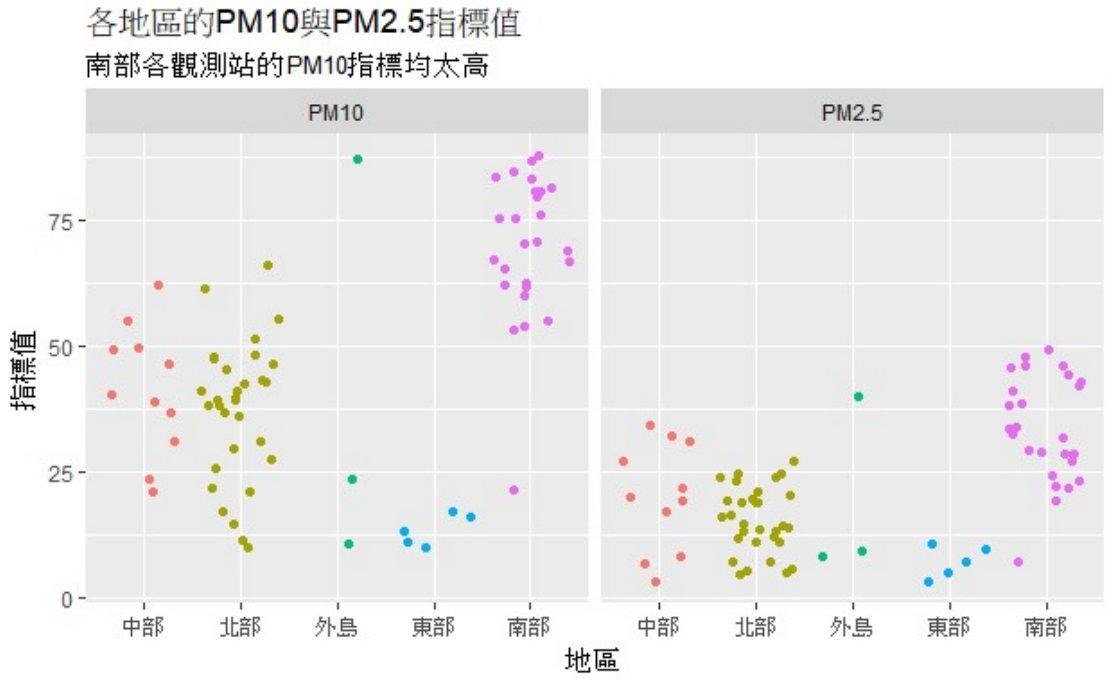
取消圖例

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  select(Area, SiteName, PM10, PM2.5) %>%
  gather(key=index, value=value, PM10, PM2.5) %>%
  ggplot(aes(x=Area, y=value, color=Area)) +
  geom_jitter() +
  facet_wrap(~index) +
  theme(legend.position="none")
```



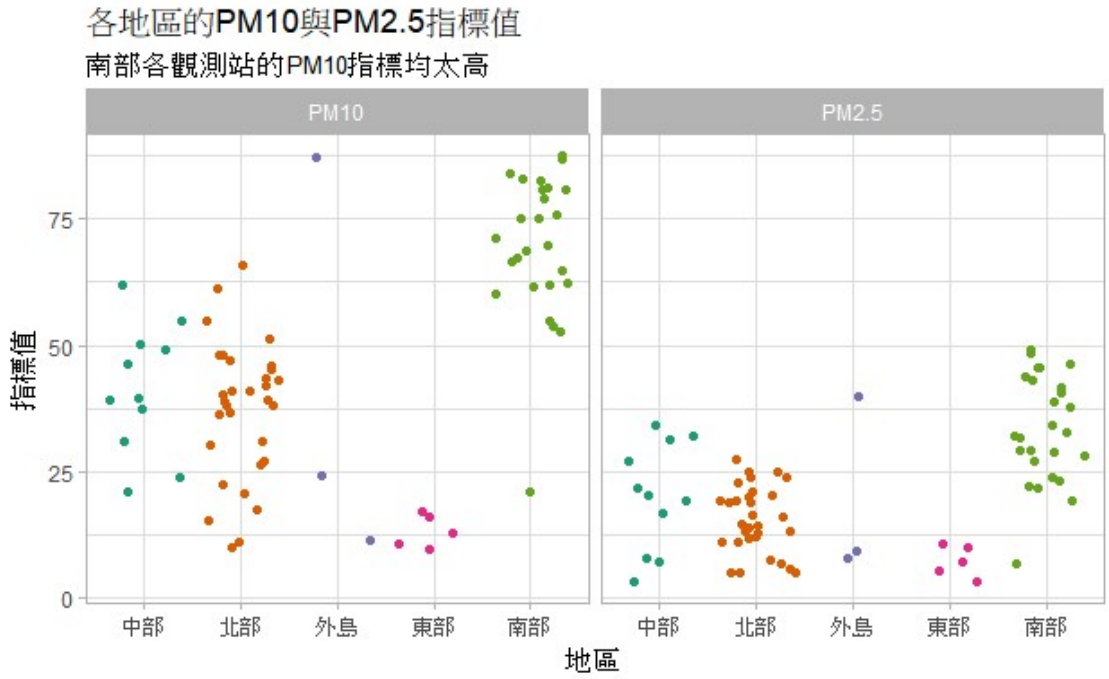
練習

- 將 X 軸與 Y 軸的標題改為中文，並加上圖表標題



練習

- 利用 scale_color_brewer()修改圖形顏色



-

簡要複習

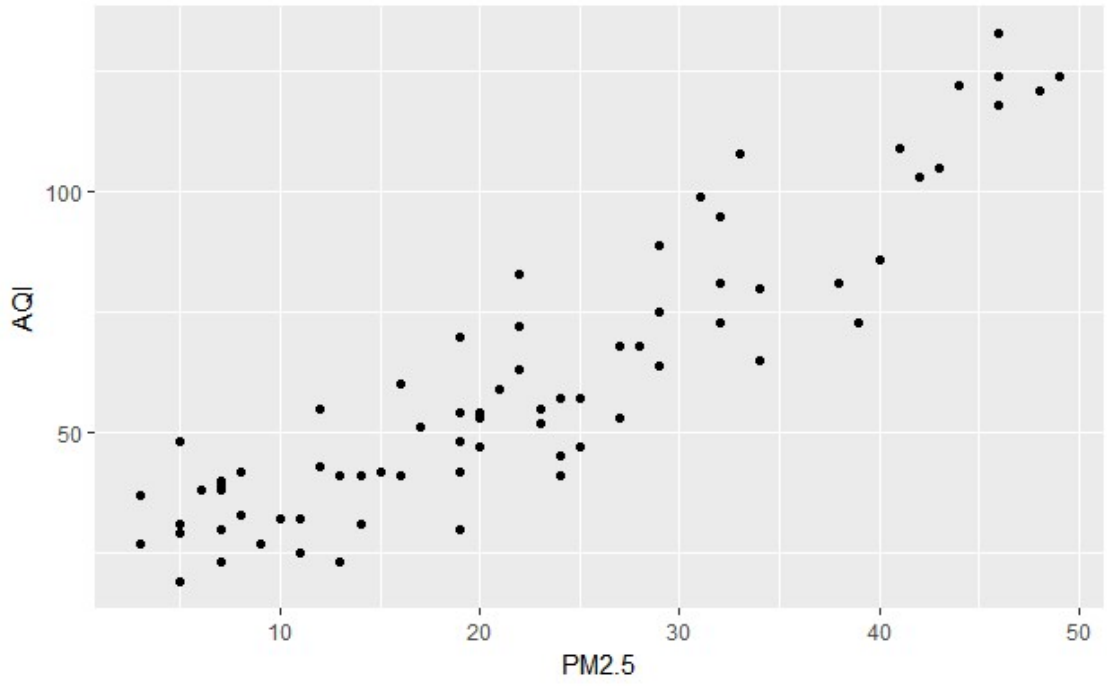
- 比較資料分布可利用點狀圖、盒狀圖和小提琴圖，三種圖形各有擅長表現的訊息。
  - 點狀圖藉由點的密度表現資料分布情形，也能夠比較不同種類資料的數量，但無法表現重要的統計訊息，如中位數等。
  - 盒狀圖節由關鍵位置的視覺圖示(如盒子的上下緣和中間的橫線等)表現出資料分布的重要統計訊息，但缺乏較細緻的資料分布情形，並且無法比較不同種類資料的數量。
  - 小提琴圖由圖形寬度變化表現資料分布情形，可以大致看出資料的集中與分散，但無法表現重要的統計訊息，也無法比較不同種類資料的數量。
- facet_wrap 可以產生並列的圖形，但需要注意座標是否需要綁定。
- 圖形的外觀可由 theme()進行設定。

PM2.5 指標與 AQI 之間的關係為何？

散佈圖可以用來兩個數值之間的關係

- 每一個觀測站在圖形上為一個點
- PM2.5 指標為點的 x 軸座標，AQI 為點的 y 軸座標

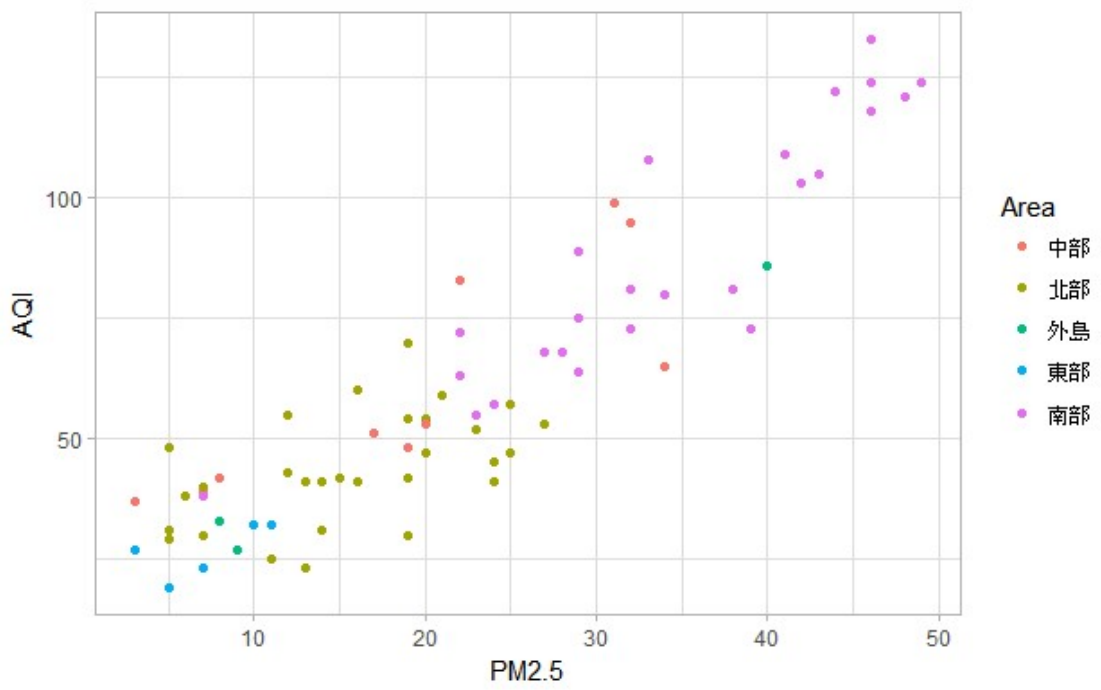
```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  ggplot(aes(x=PM2.5, y=AQI)) +
  geom_point()
```



各地區的 PM2.5 與 AQI 分布情形為何？

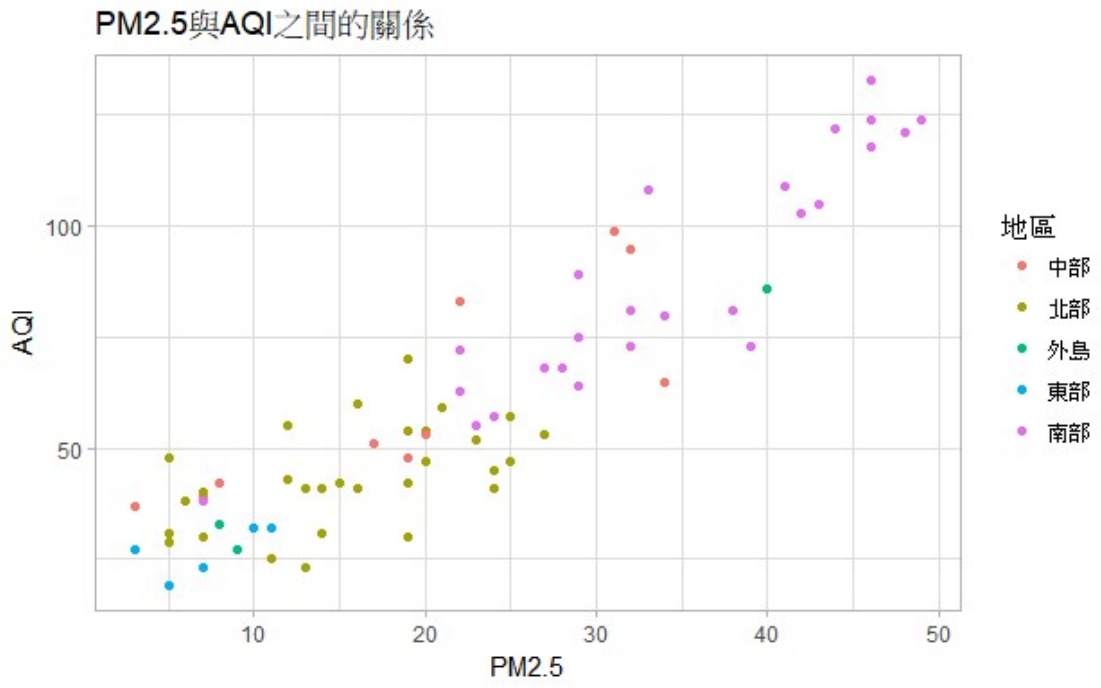
- 將地區的資訊加入散佈圖
- 使不同地區的觀測站在散佈圖上有不同顏色

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  ggplot(aes(x=PM2.5, y=AQI, color=Area)) +
  geom_point() +
  scale_color_discrete() +
  theme_light()
```

- 

修改標題

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  ggplot(aes(x=PM2.5, y=AQI, color=Area)) +
  geom_point() +
  scale_color_discrete() +
  labs(title="PM2.5 與 AQI 之間的關係", color="地區") +
  theme_light()
```



- 是否觀察到什麼樣的樣式？PM2.5 與 AQI 之間有什麼現象？

進行線性迴歸分析

```
line_model <- lm(AQI~PM2.5, aqi_data, na.action = "na.omit")
```

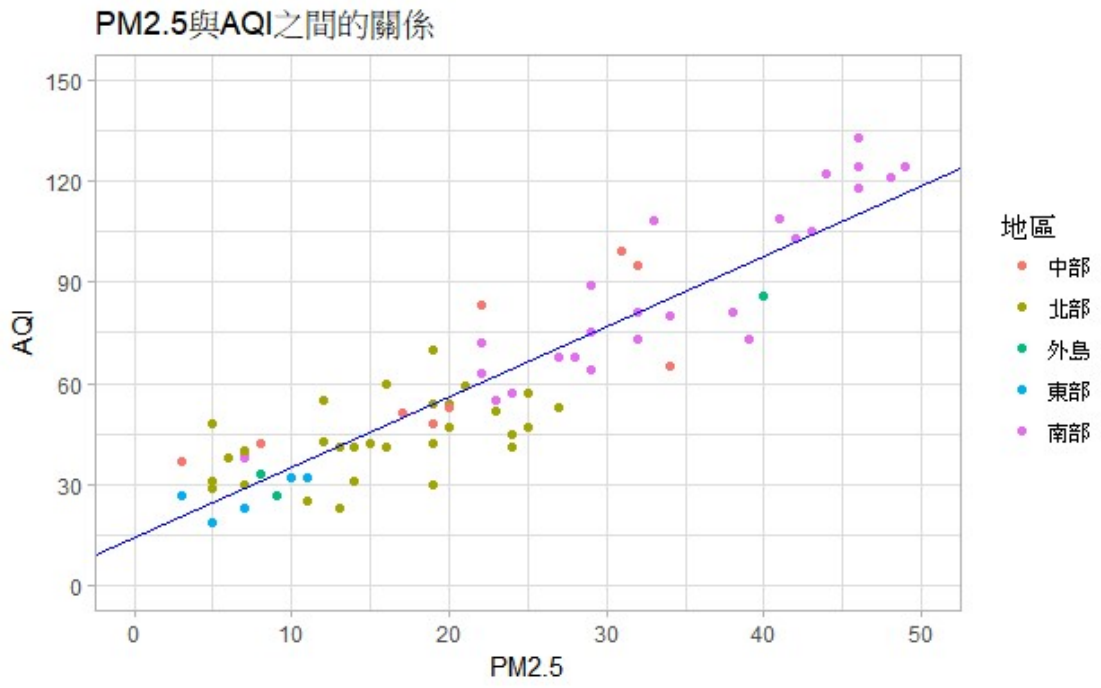
- 利用 summary(line_model)查看此次線性迴歸分析的結果
- Intercept：線的 y 軸截距
- PM2.5：線的斜率
- R-squared：模型的符合情形(在 0 與 1 之間，R-squared 愈大，模型愈符合)

```
Call:
lm(formula = AQI ~ PM2.5, data = aqi_data, na.action = "na.
omit")

Residuals:
     Min       1Q   Median       3Q      Max
-24.0346  -7.6481  -0.3414   9.2822  24.7384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.3695     2.8100   5.114  2.4e-06 ***
PM2.5         2.0876     0.1114  18.740  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 12.14 on 74 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.826,     Adjusted R-squared:  0.8236
F-statistic: 351.2 on 1 and 74 DF,  p-value: < 2.2e-16
```

-

畫出直線

```
aqi_data %>%
  filter(!is.na(AQI)) %>%
  ggplot(aes(x=PM2.5, y=AQI, color=Area)) +
  geom_point() +
  scale_color_discrete() +

  scale_x_continuous(limits=c(0, 50), breaks=seq(0, 50, 10)) +
  scale_y_continuous(limits=c(0, 150), breaks=seq(0, 150, 30)) +
  labs(title="PM2.5 與 AQI 之間的關係", color="地區") +
  theme_light() +
  geom_abline(intercept=14.3695, slope=2.0876, color="blue")
```



本次課程小結

小結

- 資料數值的分布情形： 直方圖
- x 軸：數值資料
- 同種資料數值之間的比較： 長條圖
- x 軸：類型資料
- y 軸：數值資料
- 資料數值隨時間變化的關係： 折線圖
- x 軸：時間資料
- y 軸：數值資料
- 兩種資料數值之間的關係： 散佈圖
- x 軸：數值資料
- y 軸：數值資料
- 若有第三種以上資料
  - 類型資料： 點的顏色
  - 數值資料： 點的大小
- 資料分布的比較：盒狀圖、小提琴圖
- x 軸：類型資料
- y 軸：數值資料

小結

- ggplot2 畫圖步驟

1. data frame –> ggplot()的第一項
2. aes(x, y, color, fill, size) –> ggplot()的第二項
3. geom_ –> 圖表樣式
4. scal_ –> 資料樣式
5. labs –> 標題
6. theme –> 圖表外觀

延伸思考

1. 在 Excel 上畫圖時使用的方式為 wide data format，但在其他繪圖工具時則是使用 long data format，從你自己的觀察，就一般人的思考角度來說，兩者在使用上各有何優缺點？
2. 就你自己的體會，當進行資訊視覺化的問題時，怎麼樣的進行步驟比較合適？