R 語言期中報告

一、匯入資料

```
Import <- function(){
  setwd(file.path("C:/Users/tingh/Desktop/mid project"))
  #更改工作目錄至期中專案中
  Data <- read.csv("pollution.csv", fileEncoding = 'UTF-8-BOM')
  #取得檔案內容，以中文編碼方式開啟
  return(Data) #回傳檔案內容
}
```

💡 **pollution 檔案**：2017 年 4 月 9 日高雄市的空氣品質即時污染指標
可用以研究高雄何處空氣污染最為嚴重，又是以何種污染源最多

二、敘述統計

觀察資料 & 資料統計特徵

```
Statistics <- function(data){
  print(head(data)) # 觀察前六筆資料
  print(tail(data)) # 觀察後六筆資料
  print(summary(data)) # 查看 data summary

  library(Hmisc) #使用 Hmisc library
  print(describe(data)) #使用 Hmisc 的內建函數來查看 data summar
}
```

觀察前後各六筆資料 ：head(data) & tail(data)

```
  SiteName Country Pollutant Status  SO2   CO  O3   NO NO2   NOx PM2.5 PM10 DataCreationDate
1     小港  高雄市             良好  3.4 0.36 4.8 2.84  16 19.21    13   25   2017/4/9 05:00
2     仁武  高雄市  懸浮微粒   普通  4.7 0.36 6.1 1.34  14 15.84    12   54   2017/4/9 05:00
3     左營  高雄市             良好  3.6 0.28 4.8 0.04  14 14.41    19   28   2017/4/9 05:00
4     林園  高雄市             良好  1.6 0.20 3.9 1.01  12 12.89    12   30   2017/4/9 05:00
5     前金  高雄市             良好 23.0 0.20 2.0 13.49 24 37.15     5   32   2017/4/9 05:00
6     前鎮  高雄市             良好  4.3 0.29 7.1 0.97  15 16.16    17   36   2017/4/9 05:00
    SiteName Country Pollutant Status SO2   CO O3   NO  NO2   NOx PM2.5 PM10 DataCreationDate
144     前鎮  高雄市             良好 3.9 0.25 17 0.96 10.0 11.07    21   35   2017/4/9 21:00
145     美濃  高雄市             良好 1.2 0.30 16 1.22  9.3 10.49    21   34   2017/4/9 21:00
146     復興  高雄市             良好 4.5 0.31 17 0.97 11.0 11.54    10   22   2017/4/9 21:00
147     楠梓  高雄市             良好 1.9 0.22 20 1.74  5.5  7.25    32   44   2017/4/9 21:00
148     鳳山  高雄市             良好 3.7 0.50 13 1.77 16.0 17.66    18   38   2017/4/9 21:00
149     橋頭  高雄市             良好 1.0 0.22 18 0.30  4.6  4.90    19   35   2017/4/9 21:00
```

1. 資料集不是按照地區名稱進行排序，而是按照時間
2. 污染源大致為氮、氧化合物

查看 data summary： `summary(data)`

```
  SiteName           Country           Pollutant           Status              SO2        
Length:149         Length:149         Length:149         Length:149         Min.   : 0.600  
Class :character   Class :character   Class :character   Class :character   1st Qu.: 1.800  
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 3.100  
                                                                            Mean   : 4.037  
                                                                            3rd Qu.: 4.900  
                                                                            Max.   :23.000  

       CO                O3               NO               NO2              NOx              PM2.5       
Min.   :0.100   Min.   : 2.00   Min.   :-0.400   Min.   : 1.700   Min.   : 2.120   Min.   : 2.00  
1st Qu.:0.190   1st Qu.:16.00   1st Qu.: 1.130   1st Qu.: 5.150   1st Qu.: 6.825   1st Qu.:12.00  
Median :0.230   Median :22.00   Median : 2.080   Median : 8.300   Median :10.420   Median :16.00  
Mean   :0.264   Mean   :20.93   Mean   : 2.832   Mean   : 8.805   Mean   :11.656   Mean   :16.13  
3rd Qu.:0.310   3rd Qu.:27.00   3rd Qu.: 3.730   3rd Qu.:12.000   3rd Qu.:15.240   3rd Qu.:20.00  
Max.   :0.830   Max.   :51.00   Max.   :13.490   Max.   :24.000   Max.   :37.150   Max.   :38.00  
                                NA's   :2        NA's   :2        NA's   :2        NA's   :1      
      PM10        DataCreationDate  
Min.   :18.00   Length:149         
1st Qu.:31.00   Class :character   
Median :38.00   Mode  :character   
Mean   :39.31                      
3rd Qu.:46.00                      
Max.   :88.00                      
```

1. 部分資料為中文編碼，無法確認其內容為何
2. 各項污染元素皆存在極端值，並非相對平均

使用 `Hmisc` 查看 data summary

```
---------------------------------------------------------------------------------------------------------
SiteName
       n  missing distinct 
     149        0       12 

lowest : 大寮 小港 仁武 左營 林園, highest: 美濃 復興 楠梓 鳳山 橋頭

Value          大寮     小港     仁武     左營     林園     前金     前鎮     美濃     復興     楠梓     鳳山
Frequency        12       12       13       12       13       12       12       13       12       12       13
Proportion    0.081    0.081    0.087    0.081    0.087    0.081    0.081    0.087    0.081    0.081    0.087

Value          橋頭
Frequency        13
Proportion    0.087
---------------------------------------------------------------------------------------------------------
Country
       n  missing distinct    value 
     149        0        1   高雄市

Value      高雄市
Frequency     149
Proportion      1
---------------------------------------------------------------------------------------------------------
Pollutant
       n  missing distinct    value 
      17      132        1 懸浮微粒

Value      懸浮微粒
Frequency        17
Proportion        1
---------------------------------------------------------------------------------------------------------

Status
       n  missing distinct 
     149        0        2 

Value          良好     普通
Frequency       132       17
Proportion    0.886    0.114
---------------------------------------------------------------------------------------------------------
SO2
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75 
     149        0       70        1    4.037    3.243     1.00     1.20     1.80     3.10     4.90 
     .90      .95 
    7.84     9.30 

lowest :  0.6  0.8  0.9  1.0  1.1, highest: 10.0 12.0 14.0 15.0 23.0
---------------------------------------------------------------------------------------------------------
```

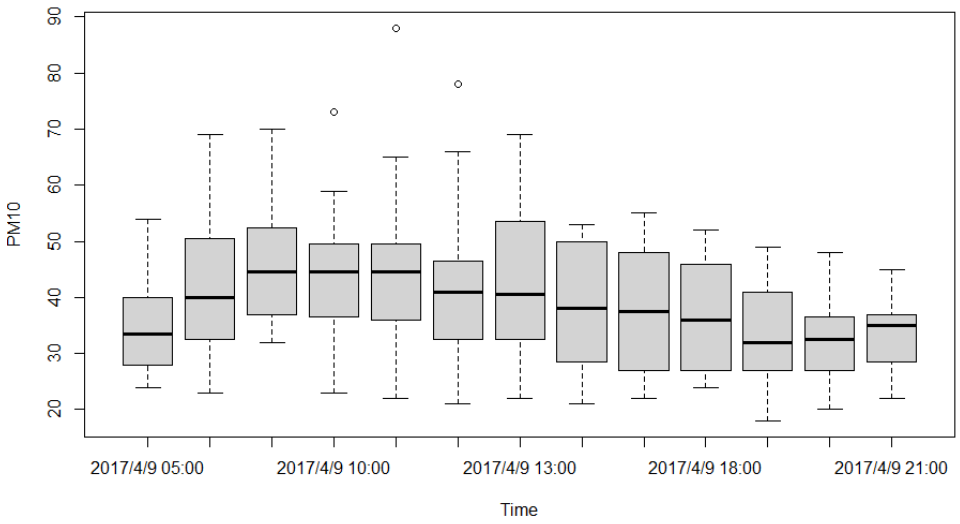
```
CO
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
     149        0       41    0.998    0.264   0.1242    0.130    0.148    0.190    0.230    0.310
     .90      .95
   0.424    0.526

lowest : 0.10 0.11 0.12 0.13 0.14, highest: 0.53 0.55 0.57 0.69 0.83
-------------------------------------------------------------------------------------------------------
O3
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
     149        0       53    0.998    20.93    10.55     4.80     6.34    16.00    22.00    27.00
     .90      .95
   32.00    33.00

lowest :  2.0  2.6  3.0  3.2  3.9, highest: 36.0 37.0 41.0 42.0 51.0
-------------------------------------------------------------------------------------------------------

NO
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
     147        2      128        1    2.832    2.467    0.547    0.860    1.130    2.080    3.730
     .90      .95
   5.974    7.868

lowest : -0.40 -0.36 -0.28  0.04  0.21, highest:  8.65  9.53 11.19 11.32 13.49
-------------------------------------------------------------------------------------------------------
NO2
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
     147        2       74    0.999    8.805    5.405     2.53     3.16     5.15     8.30    12.00
     .90      .95
   16.00    18.00

lowest :  1.7  1.8  2.1  2.4  2.5, highest: 17.0 18.0 19.0 20.0 24.0
-------------------------------------------------------------------------------------------------------
NOx
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
     147        2      142        1    11.66    7.319    3.444    4.102    6.825   10.420   15.240
     .90      .95
  20.162   24.555

lowest :  2.12  2.36  2.61  2.68  3.02, highest: 25.94 26.66 28.48 30.93 37.15
-------------------------------------------------------------------------------------------------------
PM2.5
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
     148        1       32    0.996    16.13     6.76     7.35     9.00    12.00    16.00    20.00
     .90      .95
   23.00    26.65

lowest :  2  3  4  5  6, highest: 30 32 36 37 38
-------------------------------------------------------------------------------------------------------

PM10
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
     149        0       46    0.999    39.31    13.72     22.0     24.0     31.0     38.0     46.0
     .90      .95
    54.2     62.6

lowest : 18 20 21 22 23, highest: 69 70 73 78 88
-------------------------------------------------------------------------------------------------------
DataCreationDate
       n  missing distinct
     149        0       13

lowest : 2017/4/9 05:00 2017/4/9 07:00 2017/4/9 08:00 2017/4/9 10:00 2017/4/9 11:00
highest: 2017/4/9 16:00 2017/4/9 18:00 2017/4/9 19:00 2017/4/9 20:00 2017/4/9 21:00

2017/4/9 05:00 (10, 0.067), 2017/4/9 07:00 (7, 0.047), 2017/4/9 08:00 (12, 0.081), 2017/4/9 10:00 (12,
0.081), 2017/4/9 11:00 (12, 0.081), 2017/4/9 12:00 (12, 0.081), 2017/4/9 13:00 (12, 0.081), 2017/4/9
15:00 (12, 0.081), 2017/4/9 16:00 (12, 0.081), 2017/4/9 18:00 (12, 0.081), 2017/4/9 19:00 (12, 0.081),
2017/4/9 20:00 (12, 0.081), 2017/4/9 21:00 (12, 0.081)
-------------------------------------------------------------------------------------------------------
```

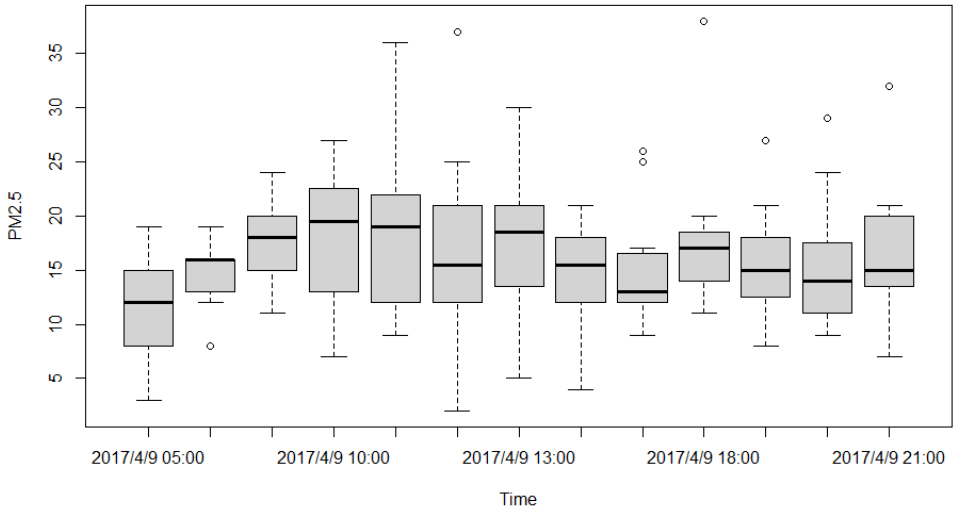
1. 合計 12 個地區，觀測時間從早上五點至晚上九點
2. 空氣品質大致上都為良好

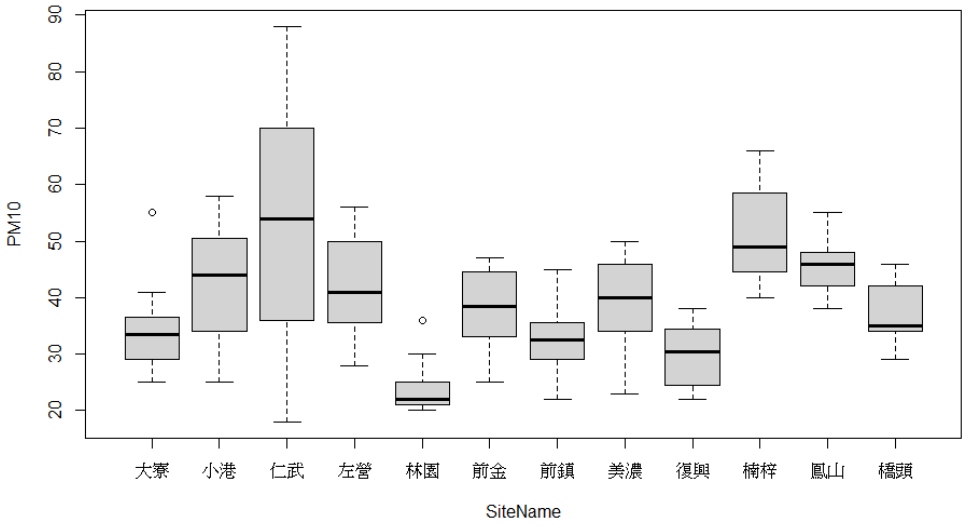
資料比較

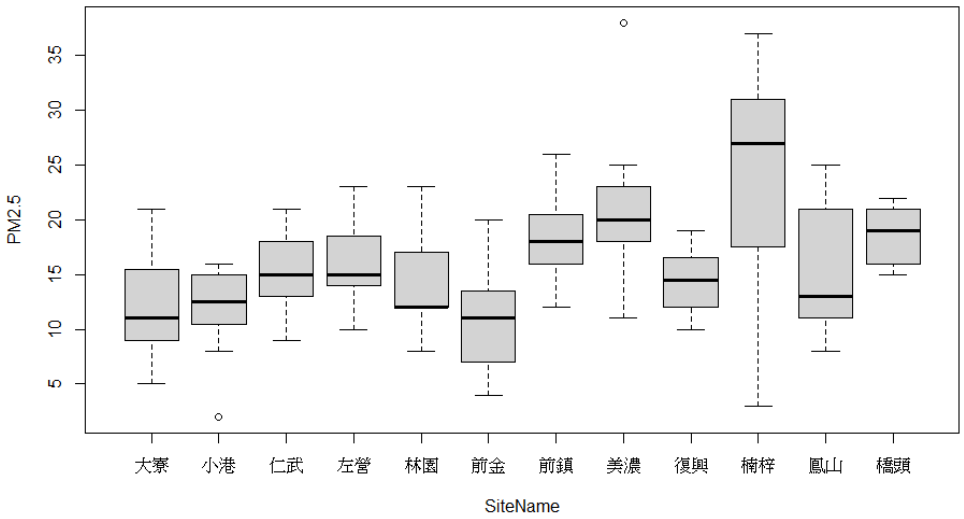
```
Histogram <- function(data){
  #觀察各地區的空汙含量
  boxplot(data[,11]~data[,1], data, xlab = "SiteName", ylab = 'PM2.5')
  boxplot(data[,12]~data[,1], data, xlab = "SiteName", ylab = 'PM10')

  #觀察各時間的空汙含量
  boxplot(data[,11]~data[,13], data, xlab = "Time", ylab = 'PM2.5')
  boxplot(data[,12]~data[,13], data, xlab = "Time", ylab = 'PM10')

  #觀察不同時間下，PM2.5 和 PM10 之間的關係
  cor.all <- by(data[,c(11,12)],INDICES = data$DataCreationDate,cor)
  print(cor.all)
}
```

各地區 & 各時間的空汙含量









1. 楠梓觀測站檢測到的 PM2.5 含量最高
2. 仁武、楠梓和鳳山觀測站檢測到的 PM10 含量最高
3. 早上 10 點左右的空汙含量最高

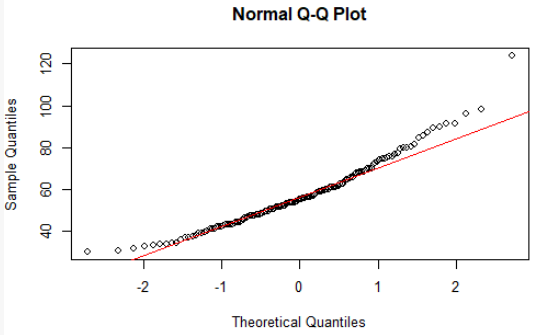
不同時間下，PM2.5 和 PM10 之間的關係

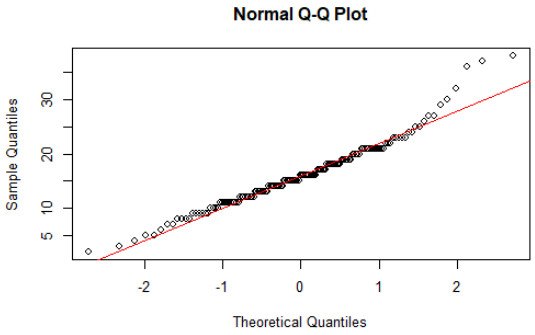
```
data$DataCreationDate: 2017/4/9 05:00
           PM2.5       PM10
PM2.5  1.0000000 -0.2576925
PM10  -0.2576925  1.0000000
----------------------------------------
data$DataCreationDate: 2017/4/9 07:00
         PM2.5     PM10
PM2.5 1.000000 0.657128
PM10  0.657128 1.000000
----------------------------------------
data$DataCreationDate: 2017/4/9 08:00
           PM2.5      PM10
PM2.5 1.0000000 0.3221448
PM10  0.3221448 1.0000000
----------------------------------------
data$DataCreationDate: 2017/4/9 10:00
           PM2.5       PM10
PM2.5 1.00000000 0.05288103
PM10  0.05288103 1.00000000
----------------------------------------
data$DataCreationDate: 2017/4/9 11:00
           PM2.5      PM10
PM2.5  1.000000 -0.039059
PM10  -0.039059  1.000000
----------------------------------------
data$DataCreationDate: 2017/4/9 12:00
           PM2.5      PM10
PM2.5 1.0000000 0.4246804
PM10  0.4246804 1.0000000
----------------------------------------
data$DataCreationDate: 2017/4/9 13:00
         PM2.5     PM10
PM2.5 1.000000 0.386751
PM10  0.386751 1.000000
----------------------------------------
data$DataCreationDate: 2017/4/9 15:00
            PM2.5        PM10
PM2.5  1.00000000 -0.08877629
PM10  -0.08877629  1.00000000
----------------------------------------
data$DataCreationDate: 2017/4/9 16:00
           PM2.5      PM10
PM2.5 1.0000000 0.1692629
PM10  0.1692629 1.0000000
----------------------------------------
data$DataCreationDate: 2017/4/9 18:00
      PM2.5 PM10
PM2.5     1   NA
PM10     NA    1
----------------------------------------
data$DataCreationDate: 2017/4/9 19:00
           PM2.5      PM10
PM2.5 1.0000000 0.3384636
PM10  0.3384636 1.0000000
----------------------------------------
data$DataCreationDate: 2017/4/9 20:00
           PM2.5      PM10
PM2.5 1.0000000 0.5356689
PM10  0.5356689 1.0000000
----------------------------------------
data$DataCreationDate: 2017/4/9 21:00
           PM2.5      PM10
PM2.5 1.0000000 0.4933335
PM10  0.4933335 1.0000000
```

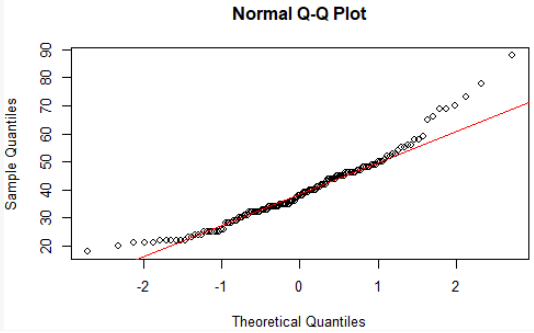
1. 一整天下來，兩者間只有微弱的正相關

三、常態檢定

```
Normal_test <- function(data){
  par(mfrow = c(2,2))

  total <- data$SO2 + data$CO + data$O3+
    data$NO + data$NO2 + data$NO2 + data$NOx
  qqnorm(total); # 所有汙染物的常態機率圖
  qqline(total,col='red') # 最佳斜線
  print(shapiro.test(total)) # shapiro-wilk 檢定

  qqnorm(data$PM2.5); # PM2.5 的常態機率圖
  qqline(data$PM2.5,col="red") #最佳斜線
  print(shapiro.test(data$PM2.5)) # shapiro-wilk 檢定

  qqnorm(data$PM10); # PM10 的常態機率圖
  qqline(data$PM10,col="red") # 最佳斜線
  print(shapiro.test(data$PM10)) # shapiro-wilk 檢定}
```







```
        Shapiro-Wilk normality test

data:  total
W = 0.955, p-value = 0.0001035


        Shapiro-Wilk normality test

data:  data$PM2.5
W = 0.96194, p-value = 0.000411


        Shapiro-Wilk normality test

data:  data$PM10
W = 0.9483, p-value = 2.525e-05
```

檢測 PM2.5, PM10 是否為常態分布

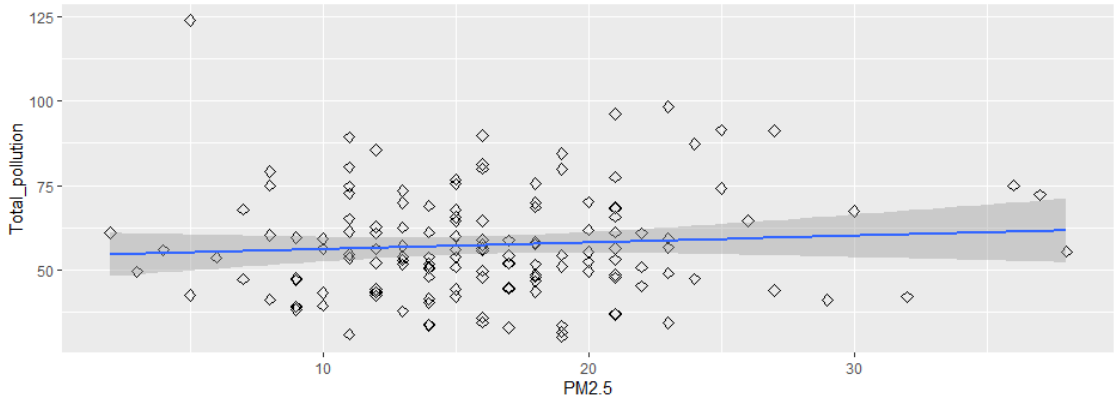
1. 三者 Shapiro-Wilk 檢定中得出的 p-value 皆小於 0.05
2. 最佳斜線與分布圖差異小
3. 三者皆為常態性分布

四、線性迴歸

因懸浮微粒並非單一物質所構成，所以將所有化合物的數量加總計算

```
regression_pm2.5 <- function(data){
  library(ggplot2) # 使用 ggplot2 套件

  # 建立模型
  total_pollution <- data$SO2 + data$CO + data$O3+
                     data$NO + data$NO2 + data$NO2 + data$NOx
  pm2.5 <- data$PM2.5
  LM2.5 <- lm(pm2.5~total_pollution, data=data)

  # 分布&預測圖
  ggplot(data, aes(x=pm2.5, y=total_pollution)) +
    geom_point(shape=5, size=2) + geom_smooth(method=lm) +
    labs(x="PM2.5", y="Total_pollution")

  summary(LM2.5) # 取得方程式參數

  new <- data.frame(total_pollution = 48) # 假定總汙染值為 48
  result <- predict(LM2.5, newdata=new) # 進行預測
  print(result)
}
```

建立總汙染物質&PM2.5 的模型及畫出分布圖



取得方程式參數 ：summary(LM)

```
Call:
lm(formula = pm2.5 ~ total_pollution, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-14.2780  -3.7560  -0.8117   3.6644  21.8943

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     14.39139    1.95160   7.374  1.2e-11 ***
total_pollution  0.03088    0.03278   0.942    0.348
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.234 on 144 degrees of freedom
  (因為不存在，3 個觀察量被刪除了)
Multiple R-squared:  0.006124,  Adjusted R-squared:  -0.0007775
F-statistic: 0.8874 on 1 and 144 DF,  p-value: 0.3478
```

1. 回歸模型公式可寫成
   PM2.5 = (14.39139) + (0.03088)×total_pollution
2. Adjusted R-squared 非常小，表示此模型的預測能力極低

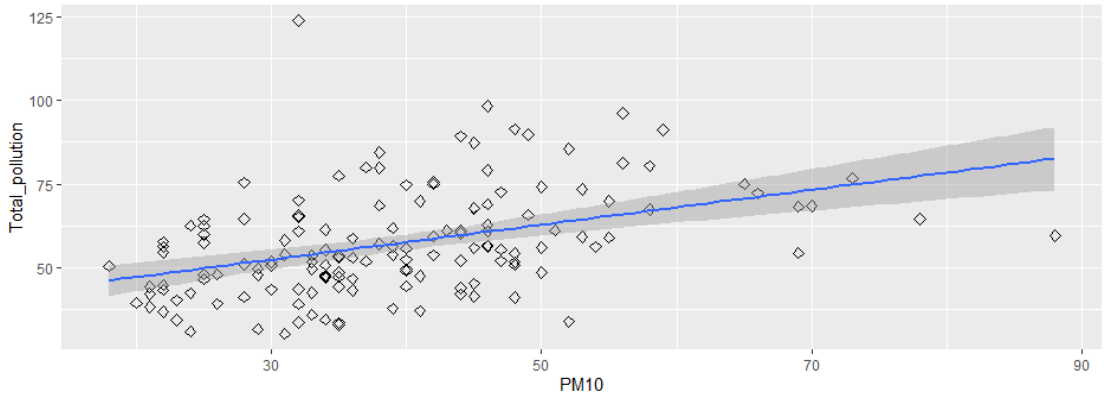
進行預測

```
> regression_pm2.5(data)
       1
15.87378
```

輸入 total_pollution = 48, 模型預測出的 PM2.5 為 15.87378

```r
regression_pm10 <- function(data){
  library(ggplot2) # 使用 ggplot2 套件

  # 建立模型
  total_pollution <- data$SO2 + data$CO + data$O3+
                       data$NO + data$NO2 + data$NO2 + data$NOx
  pm10 <- data$PM10
  LM10 <- lm(pm10~total_pollution, data=data)

  # 分布&預測圖
  ggplot(data, aes(x=pm10, y=total_pollution)) +
    geom_point(shape=5, size=2) + geom_smooth(method=lm) +
    labs(x="PM10", y="Total_pollution")

  summary(LM10) # 取得方程式參數

  new <- data.frame(total_pollution = 48) # 假定總汙染值為 48
  result <- predict(LM10, newdata=new) # 進行預測
  print(result)
}
```

建立總汙染物質&PM10 的模型及畫出分布圖



取得方程式參數 ：summary(LM)

```
Call:
lm(formula = pm10 ~ total_pollution, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-29.231  -7.629  -1.227   6.161  47.958

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     20.35282    3.58082   5.684 6.99e-08 ***
total_pollution  0.33009    0.06023   5.480 1.83e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.47 on 145 degrees of freedom
  (因為不存在，2 個觀察量被刪除了)
Multiple R-squared:  0.1716,    Adjusted R-squared:  0.1659
F-statistic: 30.03 on 1 and 145 DF,  p-value: 1.833e-07
```

1. 回歸模型公式可寫成

   PM10 = (20.35282) + (0.33009)×total_pollution
2. Adjusted R-squared 非常小，表示此模型的預測能力極低

進行預測

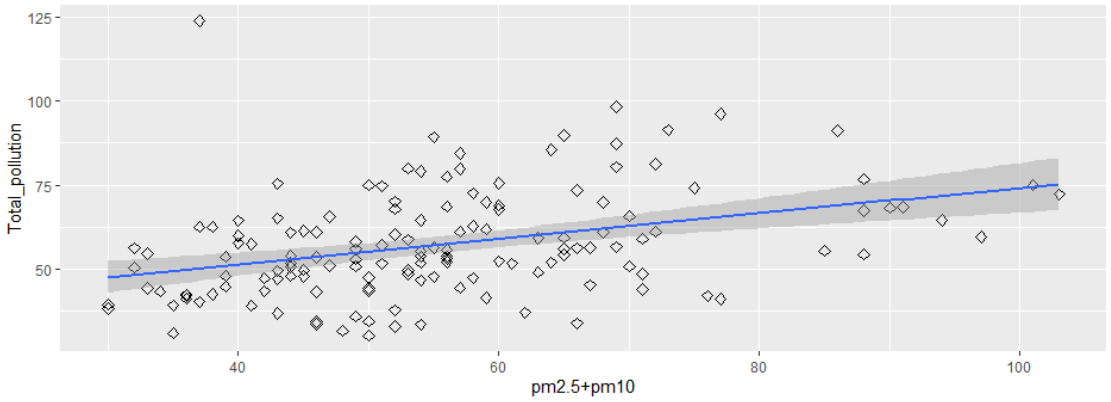
```
> regression_pm10(data)
       1
36.1969
```

輸入 total_pollution=48, 模型預測出的 PM10=36.1969

五、複雜性回歸 & 預測

```
regression_all <- function(data){
  library(ggplot2) # 使用 ggplot2 套件

  # 建立模型
  total_pollution <- data$SO2 + data$CO + data$O3+
                       data$NO + data$NO2 + data$NO2 + data$NOx
  pm2.5 <- data$PM2.5
  pm10 <- data$PM10
  LM <- lm(pm2.5+pm10 ~ total_pollution,  data= data)

  # 分布&預測圖
  ggplot(data, aes(x=pm2.5+pm10, y=total_pollution)) +
    geom_point(shape = 5, size = 2) + geom_smooth(method = lm) +
     labs(x = "pm2.5+pm10", y = "Total_pollution")

  summary(LM) # 取得方程式參數

  new <- data.frame(total_pollution = 48) # 假定總汙染值為 48
  result <- predict(LM, newdata=new) # 進行預測
  print(result)
}
```

建立總汙染物質&(PM10+PM2.5)的模型及畫出分布圖



取得方程式參數 ：summary(LM)

```
Call:
lm(formula = total_pollution ~ pm2.5 + pm10, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-30.187  -9.051  -2.153   7.625  69.536

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.80342    4.55027   8.308 6.78e-14 ***
pm2.5       -0.06972    0.19954  -0.349    0.727
pm10         0.52654    0.09910   5.313 4.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.49 on 143 degrees of freedom
  (因為不存在，3 個觀察量被刪除了)
Multiple R-squared:   0.17,     Adjusted R-squared:  0.1584
F-statistic: 14.64 on 2 and 143 DF,  p-value: 1.641e-06
```

1. 回歸模型公式可寫成
   total_pollution = (-0.06972)×PM2.5 + (0.52654) ×PM10+37.80342
2. Adjusted R-squared 非常小，表示此模型的預測能力極低

進行預測

```
> regression_all(data)
       1
52.17446
```

輸入 total_pollution=48, 模型預測出的 PM10+PM2.5 = 52.17446