R 入門實作班專題報告-日空氣品質指標 (AQI) 之研究

李明昌

2024-09-28

目录

1. 商業理解

研究目的：探討日空氣品質指標 (AQI) 在不同地區是否有差異

資料來源：政府資料開放平台-日空氣品質指標 (AQI)

資料名稱：aqx_p_434.csv

資料網址：https://data.gov.tw/dataset/40507

報告名稱：00_ 李明昌 _aqi.Rmd

2. 資料理解

資料理解包括以下主題，本研究使用免費軟體 R（R Core Team ， 2024）並參考 RWEPA 網站資料（Lee，2024）。

- 資料匯入
- 摘要
- 敘述性統計分析
- 資料視覺化
- 資料清除
- 合併
- 特徵選擇
- 資料轉換

注意

本研究為解決轉換 PDF 繪圖的標題沒有顯示中文字型問題，使用 showtext 套件。

參考：https://github.com/rwepa/ipas_bda/blob/main/ipas-r-program.R#L1348

```
library(showtext)
```

```
## 載入需要的套件: sysfonts
```

```
## 載入需要的套件: showtextdb
```

```
## Loading Google fonts (https://fonts.google.com/)
font_add_google(name = "Noto Sans TC", family = "twn")
showtext_auto()
```

資料匯入

本研究下載政府資料開放平台-日空氣品質指標（AQI），下載畫面參考下圖所示。

使用 `read.table` 匯入 aqx_p_434.csv 檔案。匯入資料名稱為 `aq`，資料筆數為 1000 筆，欄位為 11 個。

```
# 取得目前工作目錄
getwd()
```

```
## [1] "D:/00.R-Lecture-2024/2024.09.07-新明青創基地-R入門實作班/final_report_tutorial"
```

图 1: 圖 1 日空氣品質指標 (AQI) 下載圖

```
# 顯示檔案清單
dir()
```

```
## [1] "aqx_p_434.csv"             "fig_1_aqi.png"
## [3] "report_tutorial_aqi.docx" "report_tutorial_aqi.html"
## [5] "report_tutorial_aqi.pdf"  "report_tutorial_aqi.Rmd"
```

```
# 匯入資料
myfile <- "aqx_p_434.csv"
aq <- read.table(myfile, header=TRUE, sep=",")
```

資料摘要

使用 `head` 檢視前 6 筆資料。使用 `names` 顯示所有欄位名稱。使用 `str` 理解資料結構，其中 aq 為資料框（data.frame）物件，資料筆數有 1000 筆與 11 個欄位。最後使用 `summary` 理解資料料摘要，其中 o38subindex 與 pm25subindex 二個欄位包括遺漏值（NA）。

```
# 檢視前 6 筆資料
head(aq)
```

```
##   siteid sitename monitordate aqi so2subindex cosubindex o3subindex
```

```
## 1     85     大城  2024-09-23  29           2          0         NA
## 2     84   富貴角  2024-09-23  43           0          1         NA
## 3     83     麥寮  2024-09-23  14           2          2         NA
## 4     80     關山  2024-09-23  22           0          1         NA
## 5     78     馬公  2024-09-23  22           5          1         NA
## 6     77     金門  2024-09-23  38           2          3         NA
##   pm10subindex no2subindex o38subindex pm25subindex
## 1           16          12          29           14
## 2           43           8          41           26
## 3           11          13          NA           14
## 4           22           7          21           14
## 5            9          22          20           10
## 6            3          17          38           13
```

```r
# 欄位名稱
names(aq)
```

```
##  [1] "siteid"       "sitename"     "monitordate"  "aqi"          "so2subindex"
##  [6] "cosubindex"   "o3subindex"   "pm10subindex" "no2subindex"  "o38subindex"
## [11] "pm25subindex"
```

```r
# 資料結構
str(aq)
```

```
## 'data.frame':    1000 obs. of  11 variables:
##  $ siteid      : int  85 84 83 80 78 77 75 72 71 70 ...
##  $ sitename    : chr  "大城" "富貴角" "麥寮" "關山" ...
##  $ monitordate : chr  "2024-09-23" "2024-09-23" "2024-09-23" "2024-09-23" ...
##  $ aqi         : int  29 43 14 22 22 38 45 19 33 45 ...
##  $ so2subindex : int  2 0 2 0 5 2 0 2 12 2 ...
##  $ cosubindex  : int  0 1 2 1 1 3 2 1 5 7 ...
##  $ o3subindex  : logi  NA NA NA NA NA NA ...
##  $ pm10subindex: int  16 43 11 22 9 3 13 13 15 6 ...
##  $ no2subindex : int  12 8 13 7 22 17 7 15 30 45 ...
##  $ o38subindex : int  29 41 NA 21 20 38 45 19 NA NA ...
##  $ pm25subindex: int  14 26 14 14 10 13 18 19 33 13 ...
```

```r
# 資料摘要
summary(aq)
```

```
##      siteid        sitename         monitordate             aqi
##  Min.   : 1.00   Length:1000        Length:1000        Min.   :  6.00
##  1st Qu.:21.00   Class :character   Class :character   1st Qu.: 26.00
##  Median :40.00   Mode  :character   Mode  :character   Median : 35.00
##  Mean   :40.65                                         Mean   : 38.19
##  3rd Qu.:60.00                                         3rd Qu.: 47.00
##  Max.   :85.00                                         Max.   :150.00
##
##   so2subindex       cosubindex     o3subindex      pm10subindex
##  Min.   : 0.000   Min.   : 0.000   Mode:logical   Min.   : 1.00
##  1st Qu.: 2.000   1st Qu.: 1.000   NA's:1000      1st Qu.:12.00
##  Median : 2.000   Median : 2.000                  Median :17.00
##  Mean   : 4.716   Mean   : 2.524                  Mean   :17.91
##  3rd Qu.: 5.000   3rd Qu.: 3.000                  3rd Qu.:23.00
##  Max.   :69.000   Max.   :26.000                  Max.   :51.00
##
##   no2subindex     o38subindex      pm25subindex
##  Min.   : 0.00   Min.   :  6.00   Min.   : 0.00
##  1st Qu.:12.00   1st Qu.: 21.00   1st Qu.:14.00
##  Median :18.00   Median : 30.00   Median :23.00
##  Mean   :21.39   Mean   : 33.61   Mean   :27.63
##  3rd Qu.:28.00   3rd Qu.: 42.00   3rd Qu.:39.00
##  Max.   :64.00   Max.   :150.00   Max.   :85.00
##                  NA's   :143      NA's   :4
```

資料處理

使用 as.Date 將 monitordate 變數由 chr 資料型態轉換為 Date 資料型態。

```
# 日期: 字串 (chr) 修正為日期 (Date)
aq$monitordate <- as.Date(aq$monitordate)
str(aq)
```

```
## 'data.frame':    1000 obs. of  11 variables:
##  $ siteid      : int  85 84 83 80 78 77 75 72 71 70 ...
##  $ sitename    : chr  "大城" "富貴角" "麥寮" "關山" ...
##  $ monitordate : Date, format: "2024-09-23" "2024-09-23" ...
##  $ aqi         : int  29 43 14 22 22 38 45 19 33 45 ...
##  $ so2subindex : int  2 0 2 0 5 2 0 2 12 2 ...
##  $ cosubindex  : int  0 1 2 1 1 3 2 1 5 7 ...
##  $ o3subindex  : logi  NA NA NA NA NA NA ...
```

```
##  $ pm10subindex: int  16 43 11 22 9 3 13 13 15 6 ...
##  $ no2subindex : int  12 8 13 7 22 17 7 15 30 45 ...
##  $ o38subindex : int  29 41 NA 21 20 38 45 19 NA NA ...
##  $ pm25subindex: int  14 26 14 14 10 13 18 19 33 13 ...
```

```
# 資料處理
head(aq, n=3)
```

```
##   siteid sitename monitordate aqi so2subindex cosubindex o3subindex
## 1     85     大城  2024-09-23  29           2          0         NA
## 2     84   富貴角  2024-09-23  43           0          1         NA
## 3     83     麥寮  2024-09-23  14           2          2         NA
##   pm10subindex no2subindex o38subindex pm25subindex
## 1           16          12          29           14
## 2           43           8          41           26
## 3           11          13          NA           14
```

```
dim(aq) # 1000 列 11 行
```

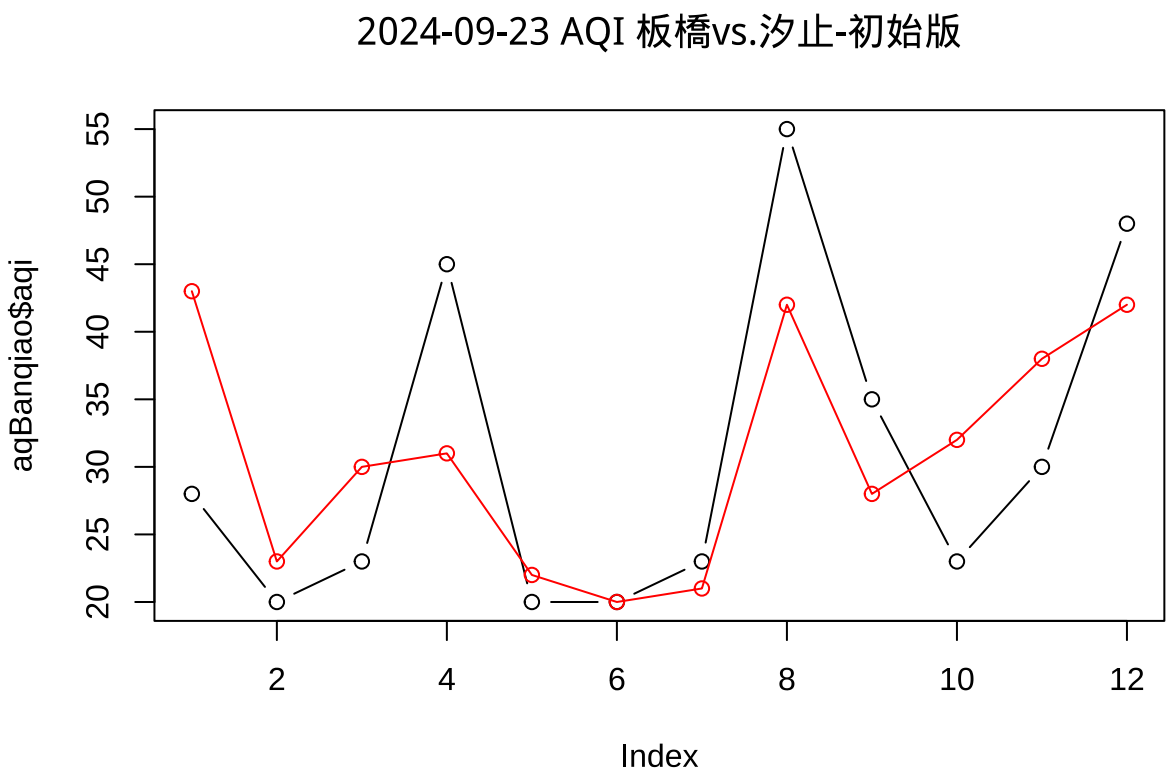
```
## [1] 1000   11
```

```
# 篩選 板橋 資料
aqBanqiao<- aq[aq$sitename == " 板橋",]

# 依照 monitordate 欄位由小至大遞增排序
aqBanqiao <- aqBanqiao[order(aqBanqiao$monitordate),]

# 篩選 汐止 資料
aqXizhi <- aq[aq$sitename == " 汐止",]

# 依照 monitordate 欄位由小至大遞增排序
aqXizhi <- aqXizhi[order(aqXizhi$monitordate),]
```

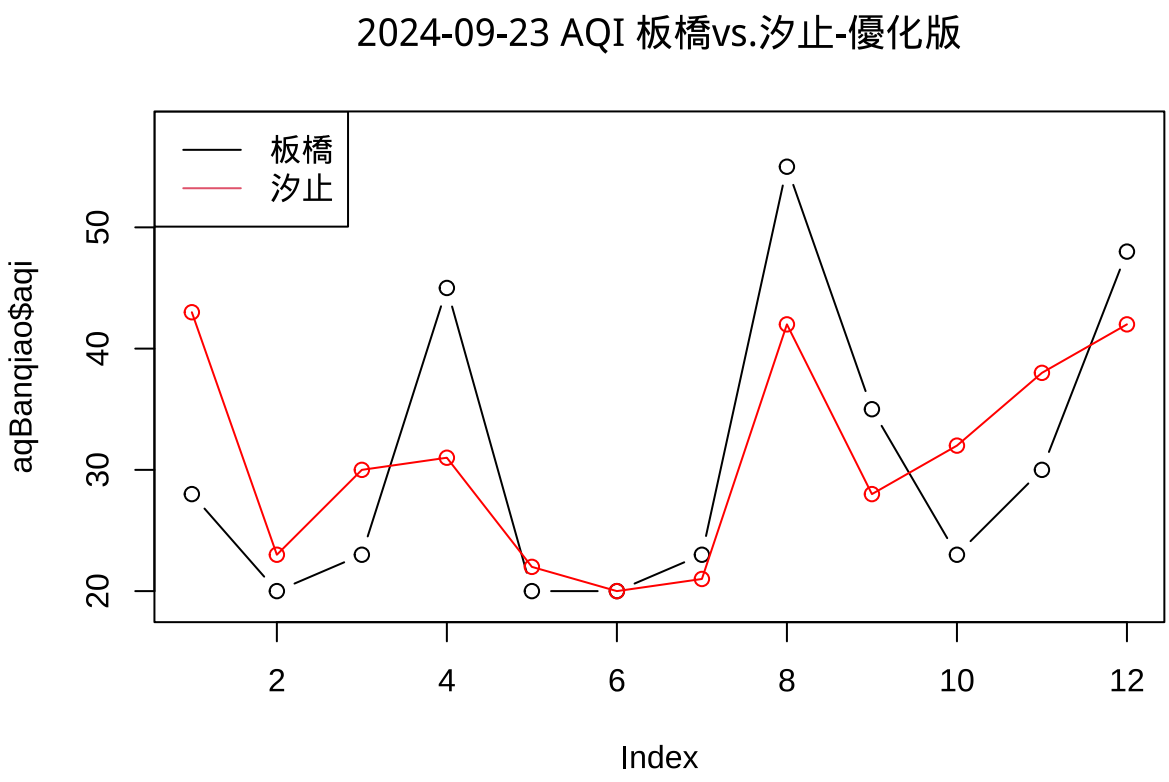
資料視覺化

```
# 板橋暨汐止 AQI 趨勢圖
plot(aqBanqiao$aqi,
     type="b",
     main = paste0(aq$monitordate[1], " AQI 板橋 vs. 汐止-初始版"))
```

```
lines(aqXizhi$aqi, col="red")
points(aqXizhi$aqi, col="red")
```

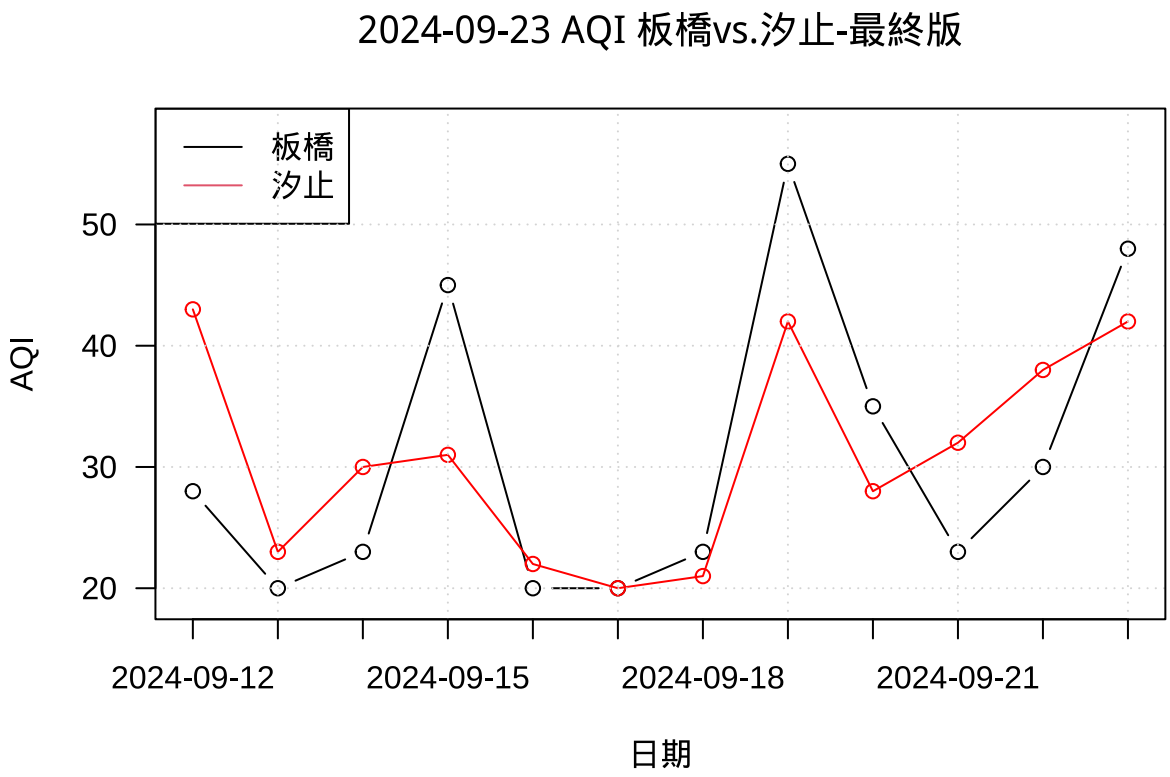



```
# 優化趨勢圖
# 客製化 Y 軸最小值, 最大值
ymin <- min(aqBanqiao$aqi, aqXizhi$aqi) - 1
ymax <- max(aqBanqiao$aqi, aqXizhi$aqi) + 3

plot(aqBanqiao$aqi,
     type = "b",
     ylim = c(ymin, ymax),
     main = paste0(aq$monitordate[1], " AQI 板橋 vs. 汐止-優化版"))
lines(aqXizhi$aqi, col="red")
points(aqXizhi$aqi, col="red")
legend("topleft", legend=c(" 板橋", " 汐止"), col=c(1,2), lty=1)
```




```r
# 優化趨勢圖-revised
plot(aqBanqiao$aqi,
     type = "b",
     ylim = c(ymin, ymax),
     axes=FALSE,
     xlab = " 日期",
     ylab = "AQI",
     main = paste0(aq$monitordate[1], " AQI 板橋 vs. 汐止-最終版"))
lines(aqXizhi$aqi, col="red")
points(aqXizhi$aqi, col="red")

# Add axis
# 1=below, 2=left, 3=above and 4=right
axis(side=1, at = 1:12, labels = aqBanqiao$monitordate)
axis(side=2, las = 2)

# 圖例
legend("topleft", legend=c(" 板橋", " 汐止"), col=c(1,2), lty=1)
```

```r
# 網格線
grid()

# 外框線
box()
```



3. 資料準備

資料準備主要工作是將資料隨機區分為二大類：訓練集 (train dataset), 測試集 (test dataset), 有的模型會加上驗證集.

本例將進行 T-檢定，因此暫無須區分訓練集與測試集。

4. 模式建立 (使用訓練集)

模式建立包括推論統計, 機器學習, 深度學習, 生成式學習等方法.

本研究採用 T-檢定方式進行，目的是比較板橋與汐止平均日 AQI 是否相等。相關假設條件如下所示：

- p-value = 0.05
- H0: 平均 AQI_ 板橋等於平均 AQI_ 汐止
- H1: 平均 AQI_ 板橋不等於平均 AQI_ 汐止

```
# 使用雙尾 T檢定
aqi_ttest <- t.test(x = aqBanqiao$aqi, y = aqXizhi$aqi)
print(aqi_ttest)
```

```
##
##  Welch Two Sample t-test
##
## data:  aqBanqiao$aqi and aqXizhi$aqi
## t = -0.038657, df = 19.743, p-value = 0.9696
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.167660  8.834327
## sample estimates:
## mean of x mean of y
##  30.83333  31.00000
```

```
print(aqi_ttest$p.value)
```

```
## [1] 0.9695521
```

5. 評估與測試 (使用測試集)

```r
# p 值 > 0.05, 接受 H0
ifelse(aqi_ttest$p.value > 0.05, " 接受 H0", " 接受 H1")
```

```
## [1] "接受H0"
```

6. 佈署應用與結論

本研究顯示板橋的平均 AQI 與平均汐止的 AQI 沒有顯著差異，未來研究亦可考慮其他觀測站的 T 檢定或變異數分析（Analysis of variance，ANOVA）。

本研究使用套件與函數與功能參考下表所示。

表 1: 本研究使用套件與函數表

套件	函數	功能
utils	read.table	匯入文字檔
showtext	font_add_google, showtext_auto	處理中文字型問題
getwd	base	顯示檔案清單
utils	head	顯示前 6 筆資料
base	names	欄位名稱
utils	str	資料結構
base	summary	資料摘要
base	as.Date	轉換為日期資料
graphics	plot	繪圖
graphics	lines	加入線
graphics	points	加入點
stats	t.test	T 檢定

參考文獻

1. Lee, Ming-Chang. (2024, September 26). RWEPA. https://rwepa.blogspot.com/

2. R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.