2017/04/28(五) 9:10~12:00, 巨量資料分析 期中考

==========

注意事項:

1. 自己做，不要討論、不要參看別人的答案。可上網查詢資料。(若要上廁所，請安靜去回)
2. 不要貼一堆圖或是報表，然後沒有任何解釋!
3. 作答完上傳(檔名: 學號-姓名-BigData-MidtermExam.docx): 帳號/密碼 (bigdata105/1fxx) 於 http://www.hmwu.idv.tw 選按【作業考試上傳區】
4. 資料
   - 行政院環境保護署 空氣品質監測網
   - http://taqm.epa.gov.tw/taqm/tw/YearlyDataDownload.aspx
   - 北部空品區 105 年監測資料檔 (105_HOUR_01_20170301.zip)
   - 空氣品質監測 105 年年報: 105_YEAR_00.pdf
   - (註: 此監測網資料包含全台灣地區，因時間關係，僅取北部空品區練習)
5. [重要] 105 年年報已有完整的問題、分析及圖表，你可以參考裡面一些背景知識，發掘一些問題，也可以使用 R 將它裡面的圖表重覆再做一次，驗証看看。
6. 老師評分標準: 回答題目時，是否有真的在探索資料、有想去了解資料本質?
7. 請按照平時上課之座位入座，但左右兩邊不要有人，至少隔一個空位。
8. 全程可上網查詢，但不能用通訊軟體討論。
9. 可使用計算機、自己的筆記型電腦及平板電腦，不可使用手機。
10. 有問題者，請在 FB 群組發問，老師會利用時間回覆，但也有可能不會回覆。
11. 請全部使用 R 處理及分析，不可用其它軟體，例如: Excel。
12. 作答請依序「執行後的 R 程式碼 => 執行結果及圖表 => 簡述結果」，例如:
    (不要直接貼「未執行」的程式碼!!)

```
> attach(iris)
> iris.lm <- lm(Sepal.Length ~ Sepal.Width)
> summary(iris.lm)
Call:
lm(formula = Sepal.Length ~ Sepal.Width)
... (篇幅關係，這裡有一些省略了)
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8251 on 148 degrees of freedom
Multiple R-squared:  0.01382,   Adjusted R-squared:  0.007159
F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519
# 簡述
利用 Sepal.Length 預測 Sepal.Width，配適簡單線性迴歸，得到之 R-squared 為 0.01382.
F 檢定並不顯著。
```

題目： 請運用目前課程所學，探索「空氣品質」資料:

1. 讀取全部檔案，並利用一些圖形或統計量檢查資料之正確性。(資料有沒有問題?你可能要知道每一變數之觀察值合理的範圍是什麼。)
2. 資料的基本統計量為何? 觀察值的範圍為何？分佈為何?
3. 條列出你可能想了解的問題(列出問題即可，不要管能不能解決或不切實際; 例如：群組比較、變數間的相關或影響等等)。
4. 以上的問題，我期待(猜測)的答案(結果)是什麼？需要什麼額外的資料來輔助分析嗎?(可能經過分析之後，答案不一定是正確的，沒關係)
5. [因時間關係這題不用做]空氣品質指標(AQI)的定義: http://taqm.epa.gov.tw/taqm/tw/b0201.aspx
   維基百科、搜尋「空氣品質指數」，查詢公式。https://zh.wikipedia.org
   為簡化起見，假設「污染物項目濃度」在計算 AQI 過程中是以 24 小時平均值為標準。
   請計算「板橋」站在 105 年度 12 個月份之空氣品質指標(AQI)。
6. 將資料整理成以下兩個 data.frame，(取名 **airdata.mean**，**airdata.var**)，表格中的 **value** 是一整個月(一天有 24 個紀錄值)的平均數(遺失值不列入計算)或變異數。格式如下:

Area	Year	Month	AMB.TEMP	CH4	…
三重站	2016	1	value		
	2016	2			
	2016	⋮			
	2016	12			
土城站	2016	1			
⋮	⋮	2			
		⋮			
⋮	⋮	12			
		⋮			
⋮	⋮				
觀音站		1			
⋮		⋮			
	2016	12			

7. 利用 **airdata.mean (airdata.var)**畫出每一監測站 PM2.5 之時間序列圖:橫軸為(1~12 月)，縱軸是 PM2.5 月平均值(變異數)。圖中每一條線代表一個監測站。兩張圖你有什麼發現?
8. 以 **airdata.mean** 為例，CO，SO2 兩污染物(變數)的分佈為何? 請做 QQplot 及常態分佈檢定(參考老師講義)。需要考慮做資料轉換嗎? 試著使用三種不同資料轉換(其中一個是 Cox-Box)，並解釋為何要採用所選的轉換方式。轉換前後有什麼差別?
9. 依照「B01-2: 遺失值、離群值處理，76/84」之準則，自選 4 種遺失值補值方法，評估哪一個是最佳的。

```
tmp <- airdata.meam[, -(1:3)]
np <- nrow(tmp) * ncol(tmp)
id <- sample(1:np, floor(np* 0.1))
tmp[id] <- NA
airdata.mean.miss <- cbind(airdata.mean[,1:3], tmp)
```

10. 答案卷最後可列出參考的網站、書本、或參考資料。