資料科學導論期中考

1. Let X be a random variable with c.d.f. $F_X(x)$, and $X_1, X_2, \cdots, X_n \overset{iid}{\sim} F_X$ be a random sample. Please show that:

   (a) (15 分) $\mathrm{E}[X]$ (the expectation of X) is the solution of the

   $$\arg\min_a \mathrm{E}\left[(X-a)^2\right].$$

   In other words, $\mathrm{E}[X]$ is the minimizer of the risk function defined by the quadratic loss function.

   (b) (15 分) $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is the solution of

   $$\arg\min_a \sum_{i=1}^n (X_i - a)^2.$$

   That is, the sample mean $\bar{X}$ is the minimizer of the empirical risk function defined by the quadratic loss function.

2. 我們在 Homework 3 中利用探索式資料分析來比較台北市與臺中市空氣污染程度的差異。這邊我們要藉由比較 Airbox 物聯網裝置與環保署官方觀測站的 PM 2.5 讀數來探討物聯網裝置的量測準確度：

   (a) (25 分) 請從「民生公共物聯網」資料集下載 2020 年 9 月份的「校園空品微型感測器」及「環保署國家空品測站」資料，並分別從中取出「臺中市立惠文高中」及「忠明測站」的 PM 2.5 歷史讀數資料。(Hint: 善用 os 與 zipfile 套件)

   (b) (20 分) 將校園空品微型感測器資料轉為逐時樣本平均值。利用探索式資料分析觀察校園空品微型感測器資料與環保署忠明測站資料讀數是否一致（或至少相關）。請注意兩種來源資料在時間上必須一致方能公平比較。

   (c) (35 分) 由於校園空品微型感測器資料中存在許多離群值，而樣本平均數 $\bar{X}$ 又很容易受離群值影響，因此 (b) 的結果可能容易受離群值干擾。令 $\bar{T}$ 為下列最佳化問題的解：

   $$\arg\min_a \sum_{i=1}^n L_5(X_i - a),$$

   其中 L_5 為 $\delta = 5$ 的 huber loss function。請利用 $\bar{T}$ 取代上題中的樣本平均數，觀察校園空品微型感測器資料與環保署忠明測站資料讀數是否一致或至少相關。