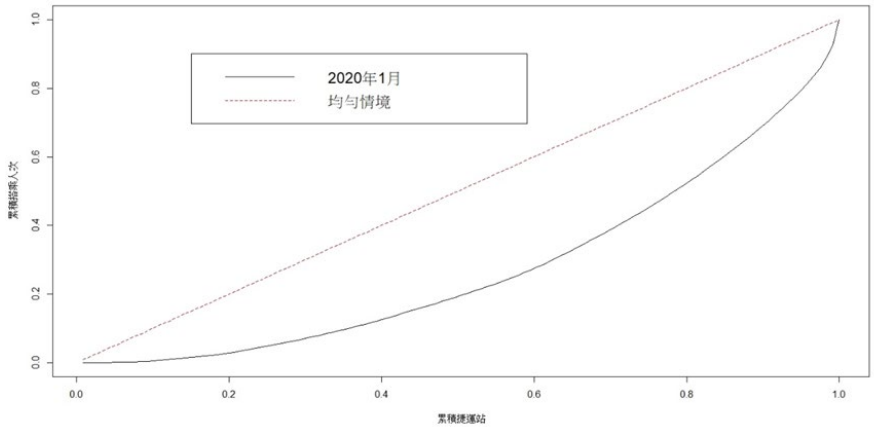
巨量資料與統計分析　　Fall 2024

第二次作業（2024/11/08 繳交）

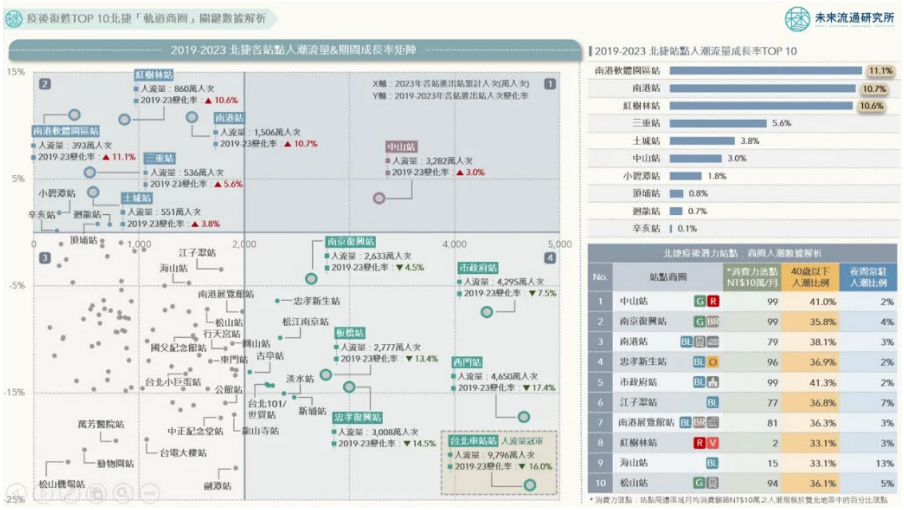
1. 臺灣政府開放許多資料可作為學術研究及大數據分析練習，以臺北各捷運站進出人數為例（https://www.metro.taipei/cp.aspx?n=FF31501BEBDD0136），可下載資料期間為 2015 年 1 月至 2024 年 9 月（117 個月），紀錄各捷運站每天進站人數、出站人數。
   (a) 以 Boxplot 繪製各捷運站的進站人數、出站人數，並說明圖形顯示的資訊及其意涵。
   (b) 以進站人數、出站人數做為解釋變數，使用群聚分析(Cluster Analysis)之類的方法，說明捷運站可分成幾群。

2. 臺北捷運自從 1996 年 3 月底開始營運，搭乘人數每月大約有 6 千萬人次，捷運站共有 117 站（2024 年 9 月底）。
   (a) 各捷運站搭乘人數差異頗大，受到武漢肺炎及新設站等因素影響，前幾名排序略有變化，搭乘人數的不均度仍然非常高，最多的 1/3 捷運站之人數高達所有人數的 2/3。而各站搭乘人數的不均度可由 Lorenz Curve 及 Gini Index 估計，下圖為 2020 年 1 月的 Lorenz Curve。請各組同學比較各年度搭乘人數的不均度（自行選擇比較的月、季、半年），並以圖形呈現其差異。



   (b) 以迴歸分析或時間數列建立捷運搭乘模型，描述台北捷運各月份（或每日、每季）進出站總人數。
   (c) （加分題：）仿造下圖探究各捷運站人數的特色及變化，並根據分析結果提出你/妳們的看法。

https://cw-image-resizer.cwg.tw/resize/uri/https%3A%2F%2Fcdn-www.cw.com.tw%2Fckeditor%2F202403%2Fckeditor-660659439dd627.30139260.png/?w=1600&format=webp



3. 臺灣擁有非常優質的人口資料時，經常會將臺灣作為比較對象。請至我網頁下載臺灣歷年死亡資料（民國 65、75、85、95 年），根據編碼簿的定義，分析及探討以下事項。

   (a) 分別計算國人在這四個年度的死亡年齡（分成男女兩性），並以圖形（像是 Boxplot）比較死亡年齡的時間變化。（註：死亡年齡也就是平均壽命，由此可推得臺灣居民的壽命趨勢。）

   (b) 除了平均壽命，請同學從編碼簿中挑出可進一步發揮的分析方向，探討與壽命（或死亡率）有關的議題。

4. 近年愈來愈多研究以房地產資料為題，套用迴歸分析估算房價。請各組同學至我的網站下載波蘭首都華沙的房地產資料（1000 筆模擬資料），其中包括一個目標變數、五個解釋變數。

   (a) 請以 EDA 工具整理所有變數的基本特性。

   (b) 以迴歸分析找出目標變數與解釋變數間的關係。

   - *m2.price*, apartments price per meter-squared (in EUR), a numerical variable range 1607 – 6595;
   - *construction.year*, the year of construction of the block of flats in which the apartment is located, a numerical variable range 1920 – 2010;
   - *surface*, apartment's total surface in square meters, a numerical variable range 20 – 150;
   - *floor*, the floor at which the apartment is located (ground floor taken to be the first floor), a numerical integer variable with values from 1 to 10;
   - *no.rooms*, the total number of rooms, a numerical variable with values from 1 to 6;
   - *district*, a factor with 10 levels indicating the district of Warsaw where the apartment is located.