Problem Set 1

1. 本題練習文字資料的讀取，並探討名字「菜市場程度」與考試成績的關係。

   使用 2004 年 8 月大學入學分發榜單和 科系代碼，二個檔案的欄位定義在此，請建構錄取學生姓名與學校科系的資料 (每位錄取學生一筆資料)，並回答以下問題。

   (a) 請先將連結中的兩個檔案 (榜單以及科系代碼) 進行資料整理，並附上整理的「程式碼」以及「結果」，若是用 Stata 請附上 data browser 中的結果，若是用其他程式語言請把結果 print 出來呈現，此一資料應用於下列各問題中。

   (b) 榜單上共有幾個錄取學生？學校有幾所？錄取人數最多的前十個學校分別為何，人數多少？

   (c) 由學校名稱將錄取學校分為公立和私立，公私立學校分別錄取多少學生？

   (d) 取出考生姓名中的「主名」與「次名」，例如林怡君的主名為「怡」、次名為「君」、「名字」為「怡君」。公立大學中出現次數最多的名字為何？私立大學呢？

   (e) 如果我們將榜單中只出現一次的名字稱為「獨特名字」，公私立大學錄取學生的名字為「獨特名字」的比例各為多少? 何者較高? 是否有顯著差異? (提示: 可以使用 **ttest** 指令來檢定，或先產生「獨特名字」的虛擬變數 (dummy variable)，再對公立學校的虛擬變數作迴歸。)

   (f) 請分別列出公私立大學中出現次數最多的前十個名字、出現次數、各個名字所佔錄取學生的比例、以及常見名字的累積比例？(例如：私立學校出現最多的二個名字是「雅婷」和「怡君」，其比例分別為 0.35% 和 0.31%，那麼常見名字由出現最多的雅婷至怡君的累積比例為 0.66%)。

   (g) 如果我們將上述累積比例視為常見名字的集中度，那麼公私立學校何者的集中度較高？

2. 此部分為你可能可以用到的 Stata 程式碼，若不了解如何應用，可以試著上網搜尋或問問 ChatGPT

   (a) gen, egen, drop, replace

   (b) destring, tostring, format(), substr(), usubstr()

   (c) sort, gsort, tab, count(), merge

   (d) collapse

   (e) keep if..., count if...