1. 資料表格處理

(合併不同類型資料並利用 key 串聯資料)

甲、**讀取 SAS 資料集**，P.18。*副檔名為.sas7bdat*

- ✓ 檔案(左上角)=>開啟=>資料

乙、**匯入 Excel 檔**，P.20。*副檔名為.xlsx*

- ✓ 檔案(左上角)=>匯入資料

丙、**匯入 csv 檔**，P.24。*副檔名為.csv*

- ✓ 檔案(左上角)=>匯入資料

丁、**附加表格**，P.28。*合併相同欄為名稱、相同資料型態表格*

- ✓ 先確定要合的同構資料表，雙極其中一張，資料(中上)=>附加表格=>增加表格

戊、**堆疊欄**，P.33。*變成 transaction 形式，把原本多個展開的 sparse 欄位，變成都在同一欄方便計算(P.37 結果比較圖)，堆疊欄位要是相同資料型態。分割欄位將單一欄未分割為複數《*

- ✓ 輸出資料表格=>資料=>堆疊欄=>不要動要保留(當成 key 值的)放到下面的分析群組依據；要進行合併的欄位放到要堆疊的欄(變成衍生欄)，衍生欄原本對應到的值叫做新值欄

己、**聯結表格**，P.38。*Merge 串聯不同資料表，文氏圖 inner(兩邊都要有才留)/outer(一邊有就留,最大)/left(以左邊資料表為主，左邊沒有的資料就不要了)/right*

- ✓ 跟其他資料表要透過某個 key join，資料表右鍵=>查詢產生器=>聯結表格=>增加表格(要跟誰合在一起)=>確認關係=>把要留著的資料拖曳到選取資料去

2. 資料彙總與衍生欄位

(選取部分的變數、產生新的變數，資料彙總分析)

甲、**選取變數**，P.49。*搭配聯結表格為以下乙丙丁的前置動作*

- ✓ 聯結表格，把變數拖到右邊

乙、**遺失值補值**，P.50。*將遺漏不具值的 nan 取代為其他值*

- ✓ 聯結完表格後，查詢產生器右邊的計算機=>重編碼的欄=>增加=>取代條件(運算子未知、取代值)=>完成後拖曳到想要的位子

丙、**新增變數**。

A. 為原本變數設置範圍區間分組(P.54)

✓ 查詢產生器右邊計算機=>重編碼欄=>增加=>取代範圍(0~10、10.01~具有此值)=>(此值置於引號)欄類型:字元=>格式刪掉留白

B. 取代原本的值變為其他字串、數值類別(P.59)

✓ 重編碼欄=>取代值=>置於引號=>字元=>格式刪除

C. 利用運算式對原本時間資料進行取代處理(P.65)

✓ 查詢產生器右邊計算機=>進階運算式=>選函數、選變數

丁、**篩選、排序資料**，P.67。*以某個欄位當成限制條件，大於、等於、小於；遞增、遞減；尋找符合某條件的人*

✓ (拖曳「and」篩選)查詢產生器=>篩選資料(中間偏上)=>運算子(未遺漏)、值

✓ (優先順位排序)查詢產生器=>排序資料

戊、**移除重複值**，P.71。*僅保留 unique 值，拿掉不看的欄位*

✓ 雙擊資料表=>資料(中間偏上)=>排序資料=>拖曳排序 key 值、不要的欄位=>選項(左邊)=>如果排序依據的 key 值有重複很多個只保留第一個做為代表

己、**彙總資料**，P.73。*各種簡單統計變量 Avg、Count、Std、Max/Min、Range、Sum、Var…，對某欄位的計算與想要一起看得欄位*

✓ 對輸出的資料表(紅色)右鍵查詢產生器=>把要看的欄位跟要統計的欄位，拖曳到選取欄位(右邊)(又要計算又要看的欄位要拖兩次)=>要計算的欄位於摘要選擇下拉選單(統計變量)=>執行

3. 初步統計分析

甲、**次數表**，P.79。*(各分群)計算所選欄位的(累積)次數、(累積)百分比*

✓ 雙擊輸出資料表(紅色)=>(工作)描述=>單因子次數=>拖曳想要看次數的欄位=>結果(左邊)=>輸出資料排序依據(次數遞減)=>執行

乙、**摘要統計** P.82 匯出資料 P.89。*基本統計(平均值、標準差、全距、總和、最大小值、觀測數目、遺漏數目、四分位數)；匯出 excel、html；摘要表、階層式分群*

✓ 雙擊輸出資料表=>描述(中偏上)=>摘要統計=>摘要統計要分析的欄位放在分析變數(摘要統計的值)、階層式分類欄位放在分類變數(依變數排序(上而下外而內))、要依照啥欄

位做成不同份表(分析群組依據)=>基本(左邊)=>選擇想要的統計變量=>百分位數=>結果(左邊)=>將統計值儲存至資料集=>執行

- ✓ 輸出資料(上面)=>匯出(中上偏右)=>匯出 XXX 做為專案中的步驟=>選剛剛儲存的資料集=>xlsx=>不要勾使用標籤做為欄名稱

丙、**摘要表** P.93 匯出表格 P.101。*階層式分群、篩選特定資料；匯出 html*

- ✓ 輸出資料表雙擊=>工作(左上)=>描述=>摘要表=>編輯(右邊)=>下拉要篩選的欄位、篩選條件值=>每個 cell 要出現的值拖曳到右邊的分析變數、想要分析的階層欄位拖曳到分類變數=>先把分析變數拖到直的最上面格子=>看直的第一階層、第二階層分類變數=>橫的第一階層、第二階層分類變數=>拖曳要看的統計變數到直的格子下方去=>執行
- ✓ 匯出=>XXX 專案中的步驟=>html

丁、**群集分析**，P.105。*利用變數分群、K-means 、R-square*

- ✓ 雙擊輸出資料表=>工作(左上)=>多變量=>群集分析=>把要分析的哪幾個維度(欄位)拖曳到分析變數=>群集(左邊)=>K 平均值演算法(最大群集數)=>執行

戊、**相關性分析**，P.108。*變數(欄位)間的線性關係程度、相關係數(信心水準) 、p 值(顯著性)*

- ✓ 雙擊輸出資料表=>工作(左上)=>多變量=>相關...=>拖曳要分析的變數到分析變數=>執行

己、**分配分析**，P.110。*檢定資料是否為常態分布、偏態峰度分配曲線*

- ✓ 雙擊輸出資料表=>工作=>描述=>分配分析=>拖曳一個要檢定的變數到分析變數=>分配摘要(左邊)=>常態、核=>標繪圖外觀(左邊)=>直方圖=>表格(左邊)=>多勾動差、常態性檢定=>執行

庚、**變異數分析**，P.113。*ANOVA 、F 值、Pr 值(>0.05 則表示不顯著、沒差異)。群之間的哪個東西有否差異*

- ✓ 雙擊輸出表格=>工作=>ANOVA=>單向 ANOVA=>要分析哪個欄位中各群拖曳到自變數=>要分析各群的哪個值拖曳到應變數(=>檢定=>比較)=>執行

4. 統計圖表製作

甲、**折線圖**，P.118。*欄位於時間的連續變化趨勢*

- ✓ 雙擊輸出資料表=>工作(左上)=>圖形=>折線圖=>資料(左邊)=>編輯(右邊)=>篩選欄位條件(SAS 格式)=>出現在 X 軸的欄位拖曳到水平、出現在 Y 軸的欄位拖曳到垂直=>打勾為

每個相異水平直做成摘要=>執行

乙、**長條圖**，P.122。*不連續資料、沒順序性，XYZ 軸*

- ✓ 3D 群組/堆疊垂直長條圖=>資料(左邊)=>X 橫軸欄位拖曳到要繪製的欄=>Y 軸各項加總和的欄位拖曳到總和=>Z 軸欄位中值分群的拖曳到長條圖群組依據=>長條圖分層種類拖曳到堆疊=>執行

丙、**盒鬚圖**(五數彙整盒形圖)，P.125。*百分位數、四分位數、離群值、IQR(匯總資料)*

- ✓ 盒形圖=>資料(左邊)=>編輯(設定欲分析值的上界/下界)=>要依據哪個欄位中的值分群組做盒形圖拖曳到水平=>要分析哪個欄位的數值資料拖曳到垂直=>執行

※僅有直方圖在工作=>描述=>分配分析，其他各種圖都在工作=>圖形

日月年: '13SEP1994'd

5. 探索性資料分析