元　　智　　大　　學

資 訊 工 程 學 系

學　士　論　文

Data Processing

研 究 生：李明昕

指導教授：簡廷因教授

中 華 民 國　一一一　年　四　月

資料處理
DATA PROCESSING

研　究　生：李明昕　Student : Ming-Shin Lee
指導教授：簡廷因　Advisor : Ting-Ting Chien

元智大學
資訊工程學系
_學_ 士論文

A Thesis

Submitted to Department of Computer Science and Engineering
College of Imformatics
Yuan Ze University
in Partial Fulfillment of the Requirements
for the Degree of
Bachelor of Informatics
in
Computer Science and Enginerring
April2022
Chungli, Taiwan, Republic of China.

中華民國　111　年　4　月

口試委員會審定書(我不知道怎麼放，他的邊界大小與其他頁不相同)

資料處理

學生：李明昕　　指導教授：簡廷因教授

元　智　大　學資訊工程學系

摘　　要

本論文研製之內容為資料處理，其目標乃為將一未知之資料集整理成合理的數據，首要項目便是觀察資料的數據有何種特徵並依據該特徵做最適當之處理以求清洗出最適當之資料集，在此目的下進行此論文研究報告。

Data Processing

Student：Ming-Shin Lee Advisors：Dr. Ting-Ying Chien

Submitted to Department of Computer Science and Engineering
College of Imformatics
Yuan Ze University

ABSTRACT

A procedure is for the sake of data processing. I want to convert an Unknown data into reasonable data. First, I need to observe the data whether the chracteristic they have, so I can use the best way to clean the dataset.

誌　　謝

非常感謝簡廷因教授的指導，讓我能在大學生活中，一路從 web 程式語言、線性代數、精準醫療、直到如今的大數據資料分析，使我獲益良多，在求學的路上能更有自信地踏出腳步向前邁進，使自己能更在如今這個時刻寫下這份致謝詞，雖然只是作業而已。

符　號　說　明

A1-A23 ：皆為直行名稱

第一章、問 題 說 明

第一節　處理動機

壹　處理方法

一、 Excel 觀察法

1、column 名稱

觀察名稱是否有特殊符號如百分比(%)

2、數值的眾數

①逗號是小數點類型的

②逗號是千位分隔符

3、實施方式

打開 excel

二、 R code 呈現

1、使用 dataframe

能夠清楚地呈現直行(column)名稱

2、使用 dplyr

可以利用此方法便利的完成資料觀察

3、實施方式

需使用 Rstudio 指令執行。

貳　研究限制

一、　時間及算力限制

1、時間限制

二、　作業的繳交期限

2、算力限制

電腦跑不夠快，請求更新電腦

第二節 處理目的

壹　短期目標

完成作業一，並獲得高分

貳　長期目標

獲得較高的學期成績

第二章、處 理 方 法

第一節 Excel 觀察法

壹、 行列編排

一、　該資料表直行(column)的第一格皆為欄位名稱，並且第一欄的名稱為 Date，故推估該資料是依照時間作排列，且個欄位擁有著不同的資料。

二、 該資料表的欄位名稱中有某些有百分比(%)符號，推估該行(column)的資料為百分比數據，並且百分比數據不會大於 100。

貳、 欄位資料

一、 逗號(,)

1、 百分比的逗號大多是呈現在最後兩位前，推估其資料皆取到小數點後第二位。

2、 A3 的資料也大多呈現在最後兩位前，推估其資料亦取到小數點後第二位。

3、 其餘資料大多為千位分隔符。

二、 極大值

許多欄位有極大值，並且觀察極大值大多在前面幾位與眾數相近。

三、 極小值

較少欄位有極小值，但觀察期數值大多與眾樹的前面幾位相近。

第二節 R code 呈現

壹、 使用 dataframe 能幫我們抓出直行名稱，無須藉由數值的指定便可以直接抓取直航全部的值。

貳、 使用 dplyr，可將資料表運用近似資料庫的指令呈現，方便觀察資料中的特殊值，並加以處理。其中更可以利用篩選器(filter)將特殊值和一般值做分割，讓不同的極端值經由不同的方式進行處理，以達成最好的補值效果。

第三章、程 式 解 析

第一節 A1 和 A2

A1 和 A2 皆無極大值和極小值，全字串長度皆不大於 5，且多可視為有兩位小數，因此無三位數的存在，故我們只需藉由 gsub 這種函式將欄位中的逗號(,)轉成小數點(.)，便能夠讓資料變成合法的資料。

第二節 A22 和 A23

有極大值的出現並且依據各行(column)不同的特徵，可以發現 A23 的字串多只有三個數字，故此我們先將逗號(,)轉成空字符()，再將長度超過數字量(A22 為 4、A23 為 3)的與沒有超過的做分割，此時超過數字量的我們命名為 tmp1，我們要將 tmp1 藉由 substr 做字串剪接成符合數字量(A22 為 4、A23 為 3)的長度，最後將 tmp1 與沒有超過的合併，便可得到純數字組成的字串，藉由 paste 和

substr 兩函式合作可以將小數點插入其中，最後將其轉成數字(numeric)，便能使資料成為無極大值的資料。

第三節 A3

觀察 A3 可以發現大多數逗號(,)後皆為兩個數字且前面為四個數字，可以推測其為千位數字且有兩位小數點，在極大值的作法與第二節並無二致，但因為其數字較大，若是出現極小值時，勢必要將其補成合理的數值，於是我們多分出 tmp2 來收納長度小於 6(數字量)的值，此時原本的資料只儲存長大恰好等於 6 的，於是我們藉由 str_pad 在長度不足的資料後面補 0 使得長度符合 6(數字量)，最後將分割出來的全部合併，便可得到由純數字組成程度為 6(數字量)的字串了。

其餘作法與第二節並無二致，藉由 paste 和 substr 將小數點插入其中並轉成數字，便可得到乾淨的資料。

第四節 其餘

其餘的大多為千位分隔符且沒有小數點，因此作法相近，我們藉由 paste0 將所有相似的製成矩陣(matrix)，作為迴圈(for)的參數運行，此時的做法和第三節幾乎相同，只有兩點不一樣。

第一點是我們並不清楚每行相對應的數字量是多少，所以我們在每次開始前，先計算我們當次所要執行的行，每一項的字串長度為多少，找出字串長度的眾數作為我們的數字量依據。

第二點是因為沒有小數點只有千位分隔符，所以在最後我們無須新增小數點，僅需直接將以數字量為長度的字串直接轉成數字就好。

第四章、討 論 與 結 論

第一節 討論

在 A3 時我們發現有些數值的長度在接受補 0 後仍舊發生長度不對等的事件，經由後續觀察發現該數字在補 0 後的字串樣式呈現首項為 0 的事件，我們推估他可能本來就只是百位數字。

有許多步驟大量的重複，應藉由迴圈來使得程式碼更為簡便。

最後的輸出並不清楚是要由命令列顯示還是輸出成 csv 檔，故本人選擇使用命令列輸出。

我沒想到我不會用頁首頁尾，搜尋了好一段時間卻無法解決輸入其中一頁時其他頁的值也會跟著改變，不知道怎麼樣同時進行兩種編號(羅馬數字和阿拉伯數字)。

我不知道格式是純粹本文的部分還是怎樣，所以我就虎頭蛇尾的將前面本文的前面認真地完成，本文的後面無所適從，並直接使用投影片提供之網址的附件作為書寫依據，特此說明。

第二節 結論

最後的結論由於先前有對資料新增名稱為 long 的直行，故需在程式輸出時藉由 grep 將 long 刪除。