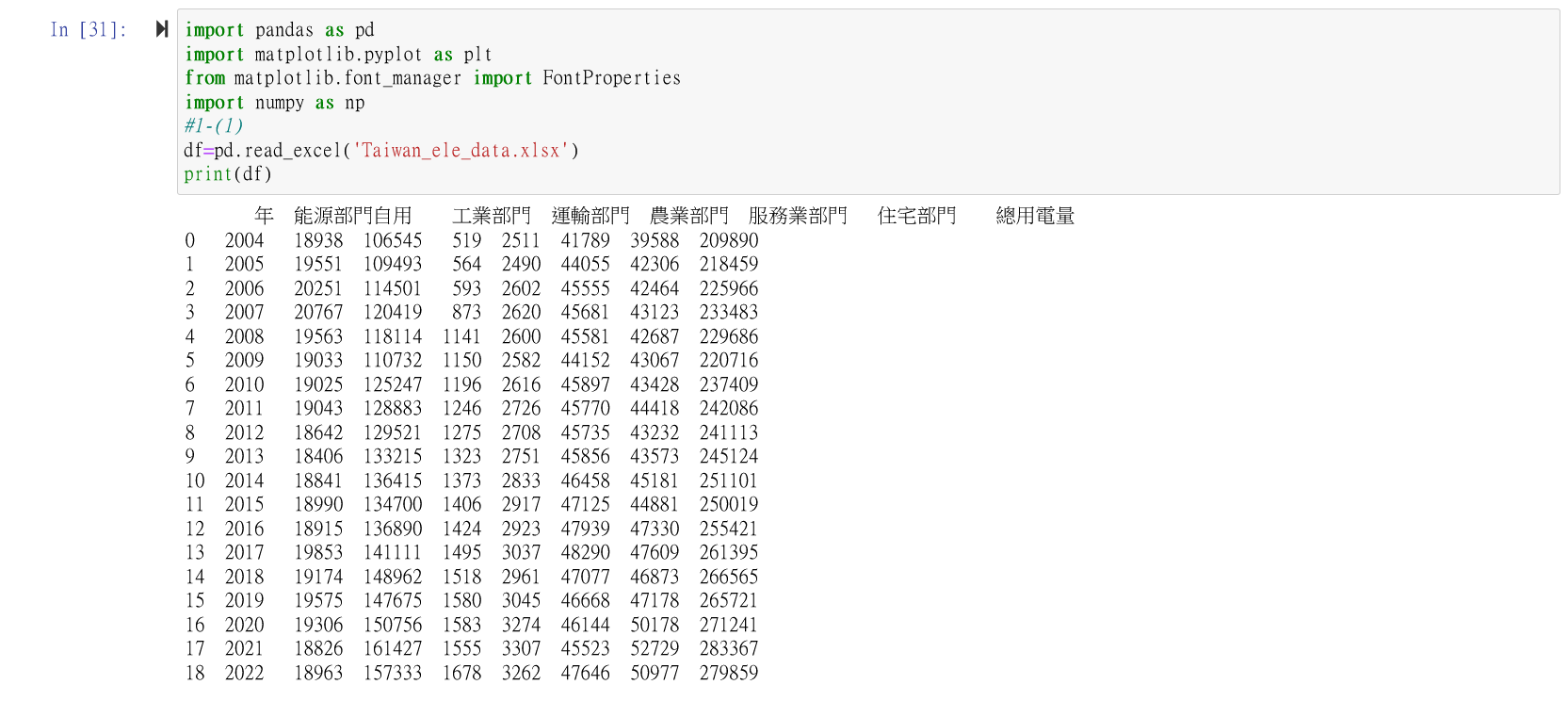
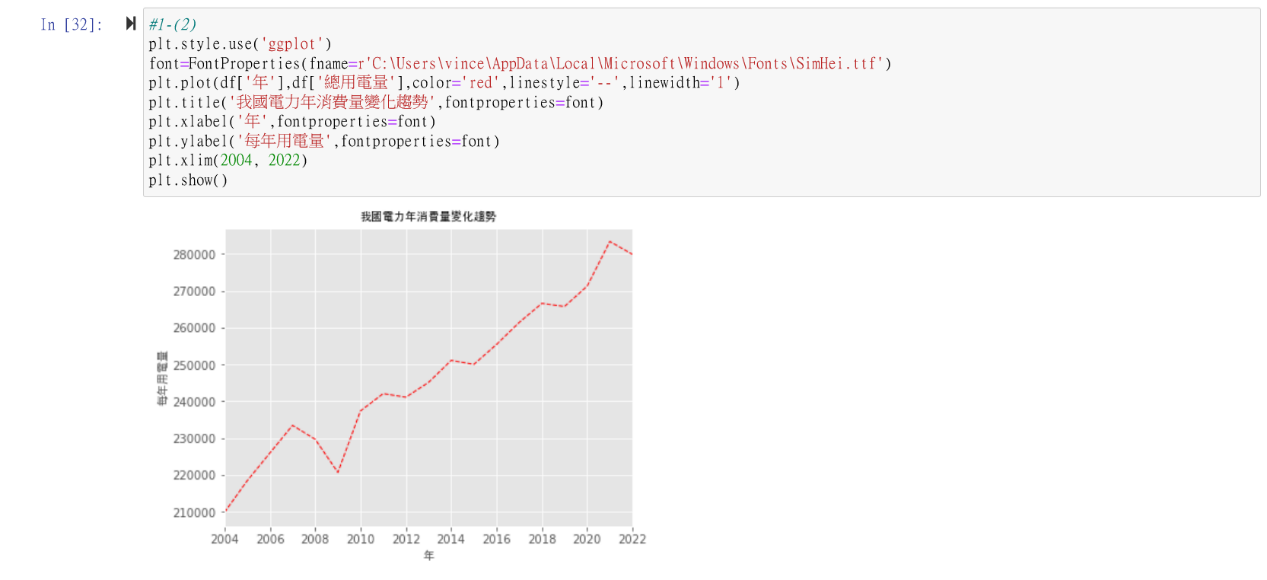
大數據分析與資料探勘-第二次作業

繳交日期: 2022/4/12

使用Python程式進行分析並回答以下所有問題。請直接在word檔中每個問題的下面，貼上python程式與執行結果，最後將檔案上傳到MOODLE，包含此word檔及python程式檔。兩個檔案，麻煩請壓縮成一個檔在上傳，檔名請命名為Homework2\_姓名.zip，另程式碼中請詳細標明對應的題數，若看不懂的話，一律不給分；此外，若檔案有缺少的話，亦會扣分。上傳期限為4/12晚上11點59分前，超過繳交期限就無法上傳。

1. 請匯入Taiwan\_ele\_data資料檔(台灣各部門電力消費資料，電力消費單位為百萬度)至DataFrame，並加以顯示。
2. 請以年為橫軸、總用電量為縱軸，繪製折線圖。此外，在此折線圖中必須包含：X軸標籤名稱為年，X軸座標起始值為2004、終止值為2022，Y軸標籤名稱為每年用電量，圖標題為我國電力年消費量變化趨勢(以上中文字體請使用SimHei字體)，折線圖線條樣式為虛線，線條顏色為紅色，線條寬度為1。
3. 一張含有 圖表 的圖片

   自動產生的描述請以年為橫軸、各部門用電量為縱軸，以折線圖繪製六大部門的每年用電量(六條折線圖)。此外，折線圖中必須包含：X軸標籤名稱為年，X軸座標起始值為2004、終止值為2022，Y軸標籤名稱為每年用電量，標題為我國各部門電力年消費量變化趨勢。另請在左上角顯示圖例(標示那一條線為哪個部門的用電量)，figsize設定為[9,7] (以上中文字體請使用SimHei字體)。
4. 請以年為橫軸、住宅部門用電量為縱軸，繪製散佈圖。此外，散佈圖中必須包含：X軸標籤名稱為年，X軸座標起始值為2004、終止值為2022，Y軸標籤名稱為每年用電量，標題為住宅部門年用電資料。另請在繪圖區中添加格線，網格線樣式為“--”，並在左上角顯示圖例，圖例名稱為住宅部門用電量(以上中文字體請使用SimHei字體)。

一張含有 圖表 的圖片

自動產生的描述

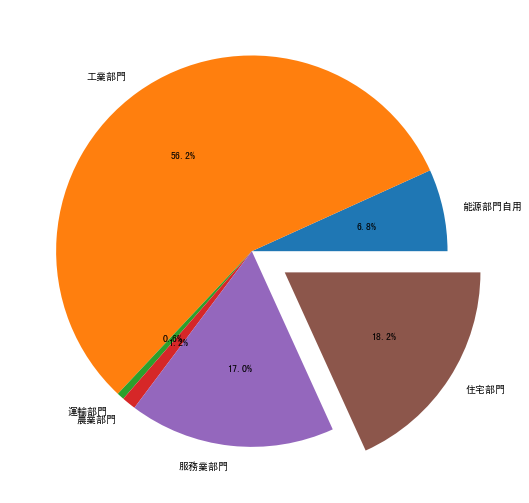
1. 一張含有 圖表 的圖片

   自動產生的描述請以年為橫軸、每年用電量為縱軸，以長條圖繪製服務業部門與住宅部門這兩個部門的每年用電量比較(亦即繪製兩條長條圖)。此外，長條圖中必須包含：X軸標籤名稱為年，Y軸標籤名稱為每年用電量，標題為服務業部門與住宅部門年用電量比較，並在左上角顯示圖例，圖例名稱分別為服務業部門用電量與住宅部門用電量(以上中文字體請使用SimHei字體)。
2. 假設現在想知道6大部門在最新年度的用電量占比，請使用6大部門在2022年的年用電量，然後計算每個部門年用電量占我國年總用電量的比例，繪製成圓餅圖，請把每個部門名稱及占比顯示在圖上 (以上中文字體請使用SimHei字體、占比請顯示至小數點第1位)(若部門名稱或占比都擠在一起的話，可把圖形尺寸放大)。

一張含有 圖表, 圓餅圖 的圖片

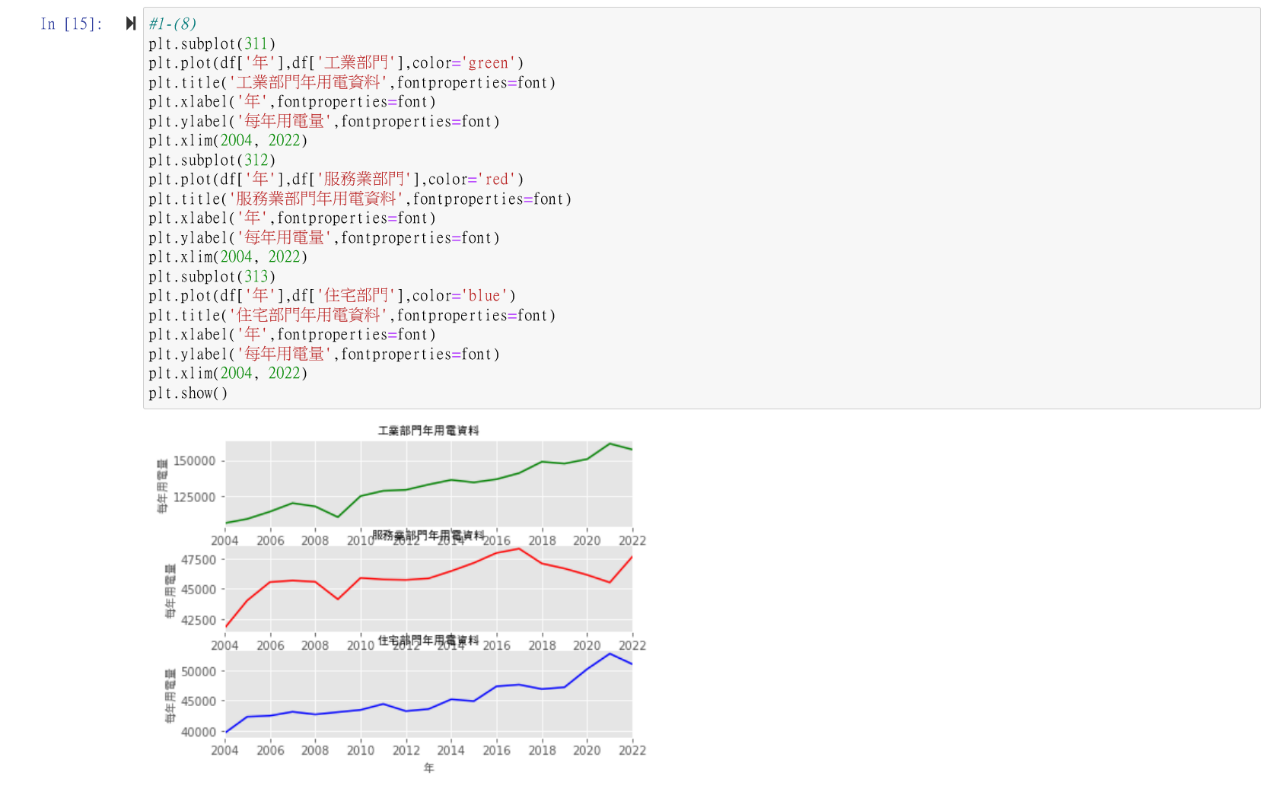
自動產生的描述

1. 承上題，分析人員發現住宅部門的用電量占比雖然並非最大，但是未來此部門極具節電潛力，想要在圓餅圖中特別凸出來看，如下圖所示，請繪製如以下之圖形。



一張含有 圖表, 圓餅圖 的圖片

自動產生的描述

1. 請用多圖合一方式(3x1)，最上面圖形畫工業部門、中間圖形畫服務業部門、最下面圖形畫住宅部門每年用電量的折線圖。此外，圖形中必須包含：X軸標籤名稱為年，X軸座標起始值為2004、終止值為2022，Y軸標籤名稱為每年用電量，三個圖形的標題分別為工業部門年用電資料、服務業部門年用電資料與住宅部門年用電資料。最上面圖形的線條請用綠色，中間圖形的線條用紅色、最下面圖形的線條用藍色(以上中文字體請使用SimHei字體)。
2. 針對這6大部門的每年用電量，繪製堆疊面積圖。此外，堆疊圖中必須包含：X軸標籤名稱為年，X軸座標起始值為2004、終止值為2022，Y軸標籤名稱為每年用電量，標題為各部門年用電量之堆疊圖，並在左上角顯示圖例(標示那一塊面積為哪個部門的用電量)。(以上中文字體請使用SimHei字體)

一張含有 資料表 的圖片

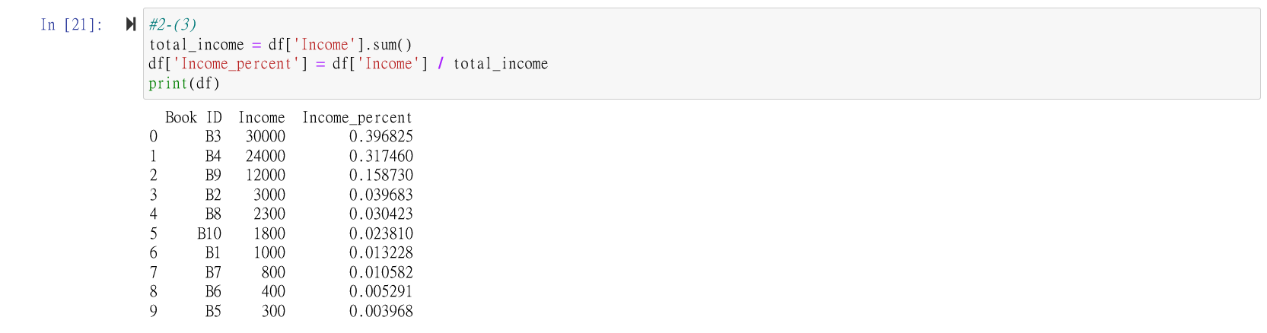
自動產生的描述

Note：繪圖問題中有多項要求，請一一滿足，若僅完成一部分，則僅會得到部分分數。

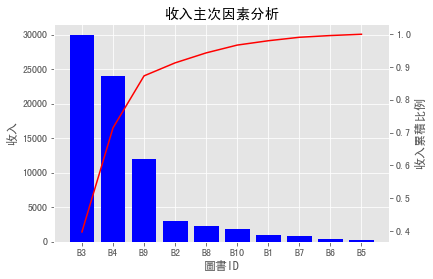
2.

1. 請用手動建立一個DataFrame如下表，並加以顯示。

|  |  |
| --- | --- |
| Book ID | Income |
| B1 | 1000 |
| B2 | 3000 |
| B3 | 30000 |
| B4 | 24000 |
| B5 | 300 |
| B6 | 400 |
| B7 | 800 |
| B8 | 2300 |
| B9 | 12000 |
| B10 | 1800 |
|  |  |

1. 請以收入(Income)做遞減排序，並reset index後，重新顯示資料表。
2. 承上題，請將排序後的資料表每本書的收入(Income)除以全部書本加總後的總收入，請在上題的DataFrame最後再增加一個新的column，顯示所計算出的收入比例，並將此新的column名稱命名為「Income\_percent」。
3. 一張含有 資料表 的圖片

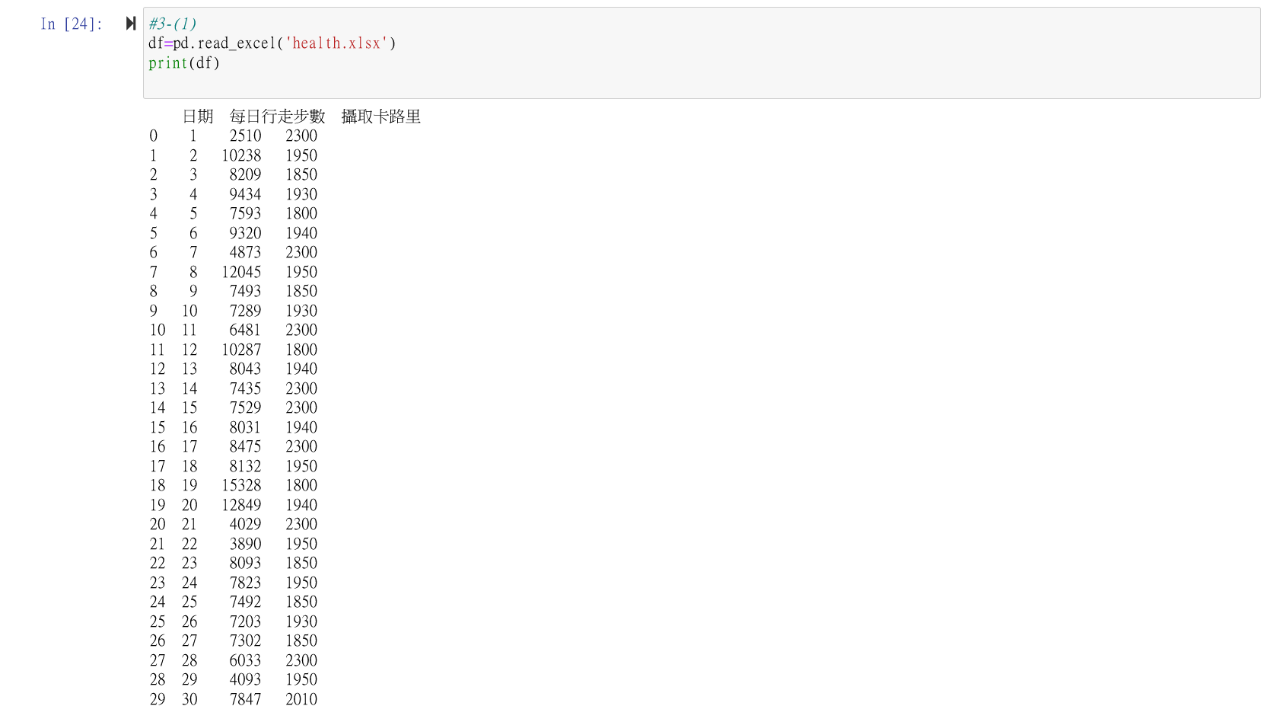
   自動產生的描述承上題，請將收入比例由上至下依序累加為收入累積比例，並在上題的DataFrame最後再增加一個新的column，顯示所計算出的收入累積比例，並將此新的column名稱命名為「Income\_cum\_percent」。
4. 請以Book ID為橫軸、收入(Income)為左邊縱軸，繪製長條圖，顏色設定為藍色。此外，在此圖形下同樣以Book ID為橫軸、收入累積比例為右邊縱軸，繪製折線圖，顏色設定為紅色。其中， X軸標籤名稱為圖書ID，左邊Y軸標籤名稱為收入，右邊Y軸標籤名稱為收入累積比例，圖標題為收入主次因素分析 (以上中文字體請使用SimHei字體)，繪製出如以下的圖形



一張含有 圖表 的圖片

自動產生的描述

3.

1. 請匯入health資料檔(此資料檔為小明在某個月的每天行走步數與攝取卡路里)至DataFrame，並加以顯示。
2. 在此DataFrame最後增加一個新的column，計算「每日行走步數」除上「攝取卡路里」，將此新的column名稱命名為「每單位卡路里之行走步數」，並加以顯示此新的資料表。

一張含有 資料表 的圖片

自動產生的描述

1. 一張含有 資料表 的圖片

   自動產生的描述承上題，依據「每單位卡路里之行走步數」可判斷出每天的健康指數，若值小於等於2則顯示「低」，若值大於2、小於等於4.5則顯示「中」，若值大於4.5則顯示「高」，請在上題的DataFrame最後再增加一個新的column，顯示這些健康指數，並將此新的column名稱命名為「健康指數」。
2. 一張含有 資料表 的圖片

   自動產生的描述承上題，由於這些健康指數為類別型資料，使用One-hot Encoding方法將其擴展為三個columns (亦即對應於該列的column為1，其餘columns為0)，將這三個新columns增加到上題DataFrame的最後，column名稱分別為「健康指數\_高」，「健康指數\_中」和「健康指數\_低」。