**分散式資料庫期末報告**

**OLAP分析中風預測**

|  |  |
| --- | --- |
|  | **指導教授：蔡正發教授**  **組員：B10856012 吳明軒**  **B10856025 王郁晴**  **中華民國:111年6月6號** |

目錄

[一、 資料背景 2](#_Toc105374293)

[二、 資料集分析簡介 3](#_Toc105374294)

[三、 資料整理 4](#_Toc105374295)

[四、 OLAP分析 7](#_Toc105374296)

[1. 切片(Slice) 7](#_Toc105374297)

[2. 切丁(Dice) 11](#_Toc105374298)

[3. 向上截取(Roll-Up) 14](#_Toc105374299)

[4. 向下擷取(Drill-Down) 14](#_Toc105374300)

[5. 轉軸(Pivot) 15](#_Toc105374301)

[五、 分析結論 16](#_Toc105374302)

# 資料背景

* 來源：Kaggle資料科學平台<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
* 資料名稱：Stroke Prediction Dataset
* 分析項目：各欄位對中風影響
* 更新資料集：最後更新2021年
* 各表單的資料量：

受訪資料：4907筆

年齡：4筆

性別：2筆

高血壓：2筆

心臟病：2筆

BNI值：4筆

中風：2筆

工作類型：5筆

結婚：2筆

居住環境：2筆

抽菸：4筆

平均血糖：4筆

# 資料集分析簡介

根據世界衛生組織 (WHO)的數據，中風是全球第二大死因，約佔總死亡人數的 11%。該數據集用於根據性別、年齡、各種疾病和吸煙狀況等輸入參數預測患者是否可能中風。數據中的每一行都提供有關患者的相關信息。

1. 資料整理

**原始的資料**

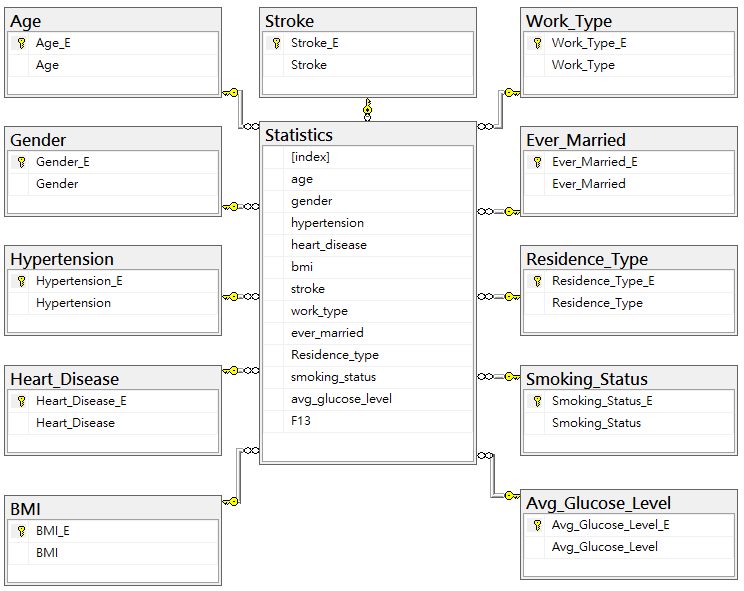


整理後

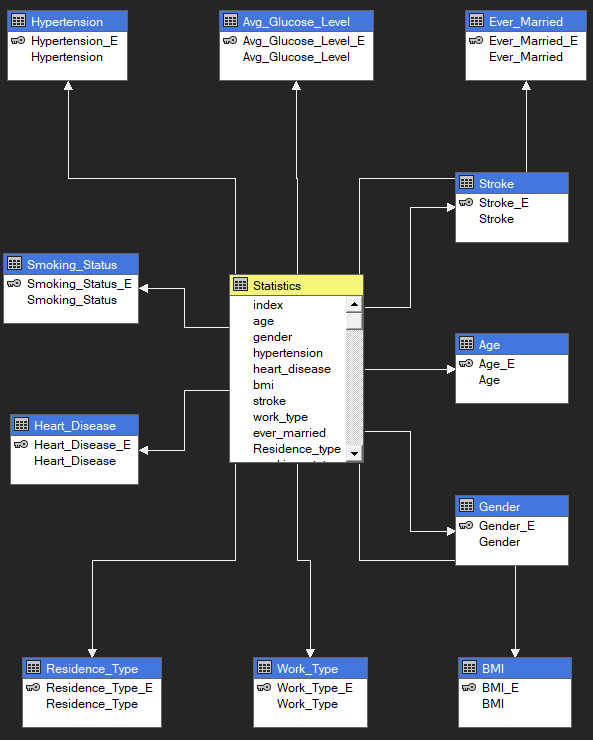
說明: 把個人id編碼刪掉，新增index重新編號從1~4907。再將所有欄位轉換為數值表示。



**Star Schema**

****

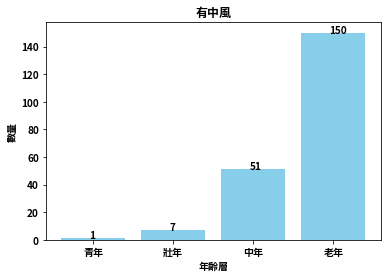
**Cube**

****

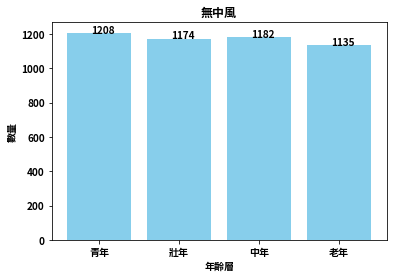
# OLAP分析

## 切片(Slice)

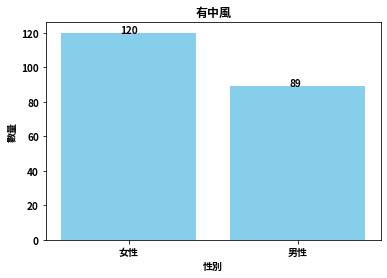
中風者的年齡

****

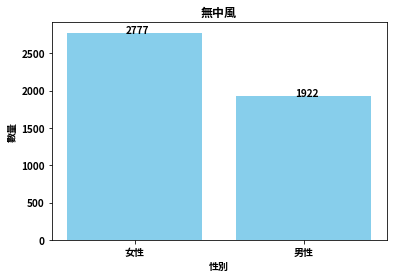
無中風者的年齡

****

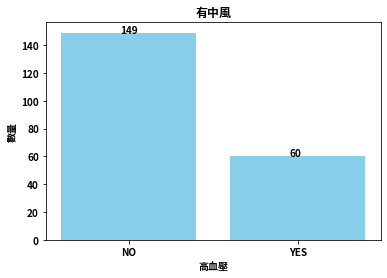
中風者的性別

****

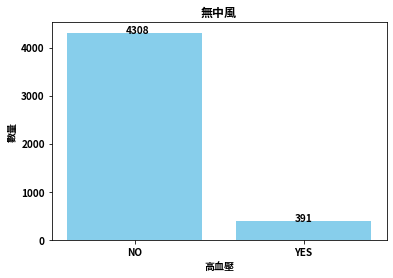
無中風者的性別

****

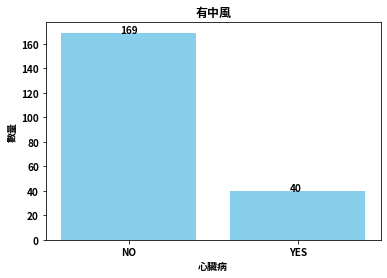
中風者是否患有高血壓

****

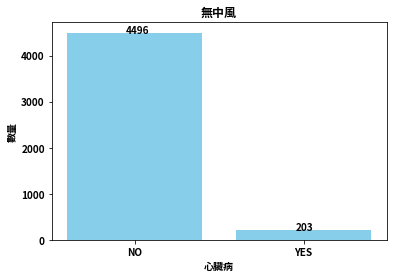
無中風者是否患有高血壓

****

中風者是否患有心臟病

****

無中風者是否患有心臟病

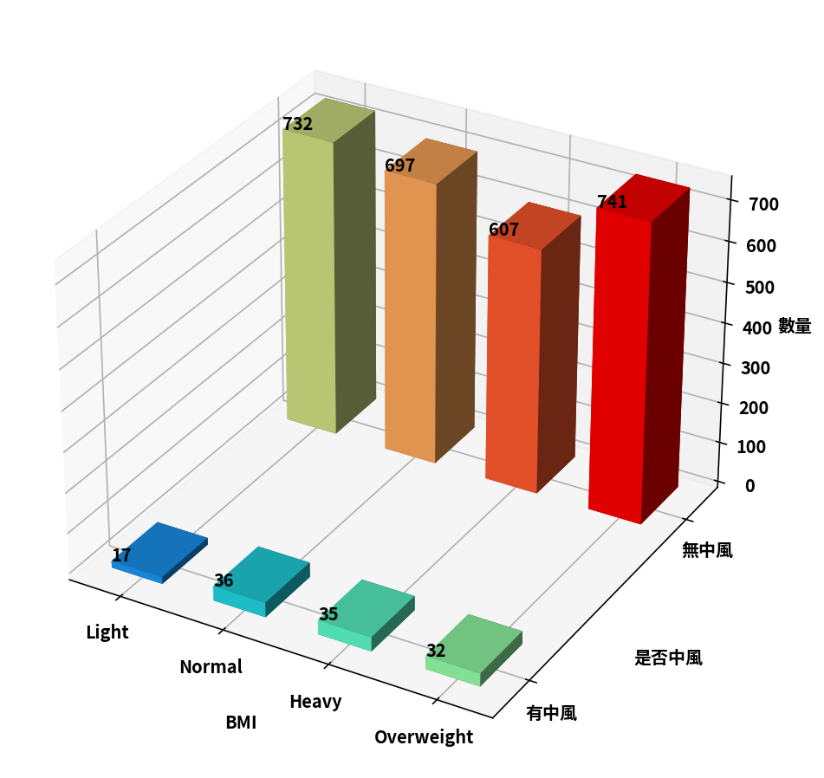
****

## 切丁(Dice)

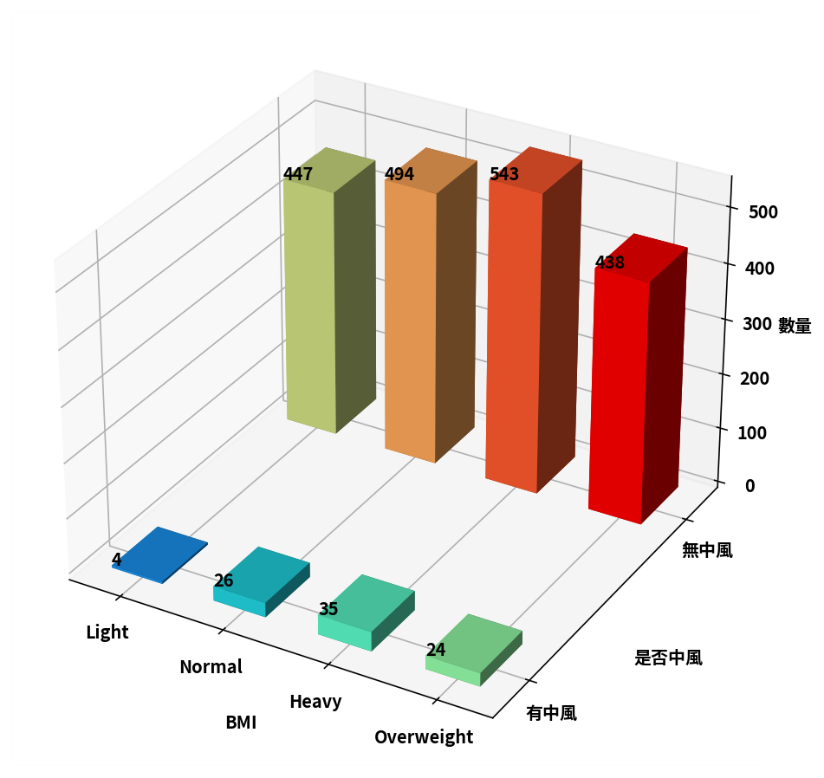
* 劃分男女性別分析對中風的影響。

主要分析男女的生活狀況是否與中風有關。

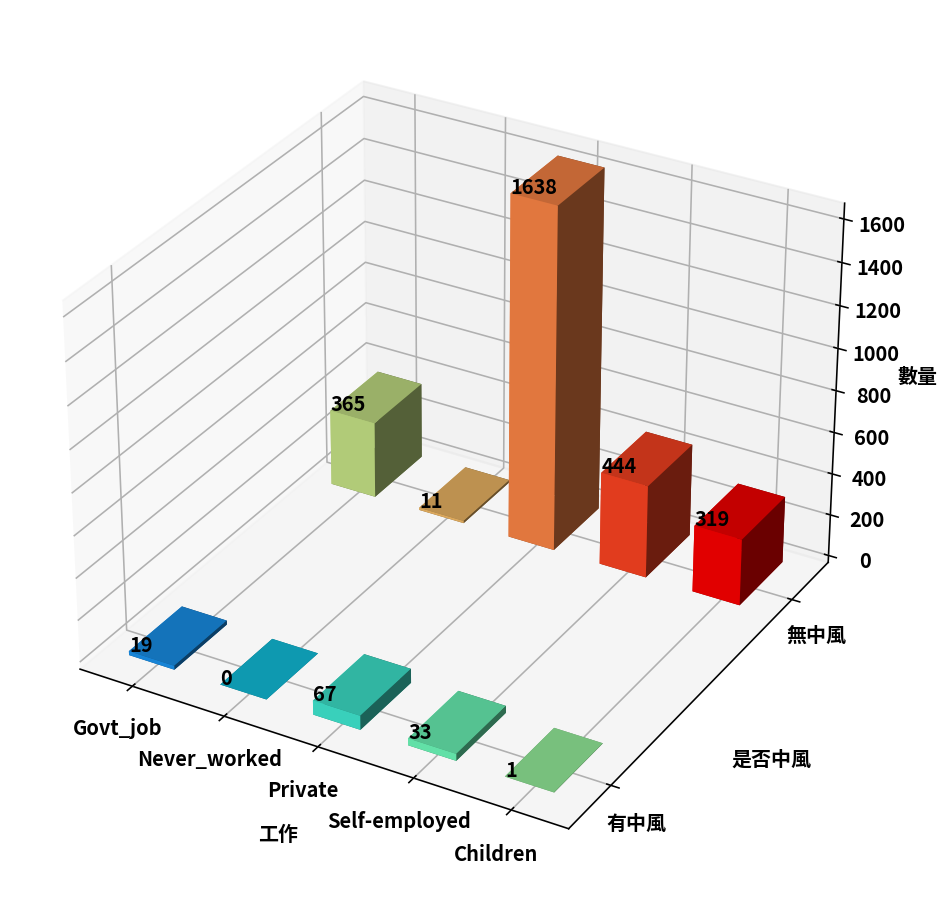
女性BMI指數對中風的影響

******

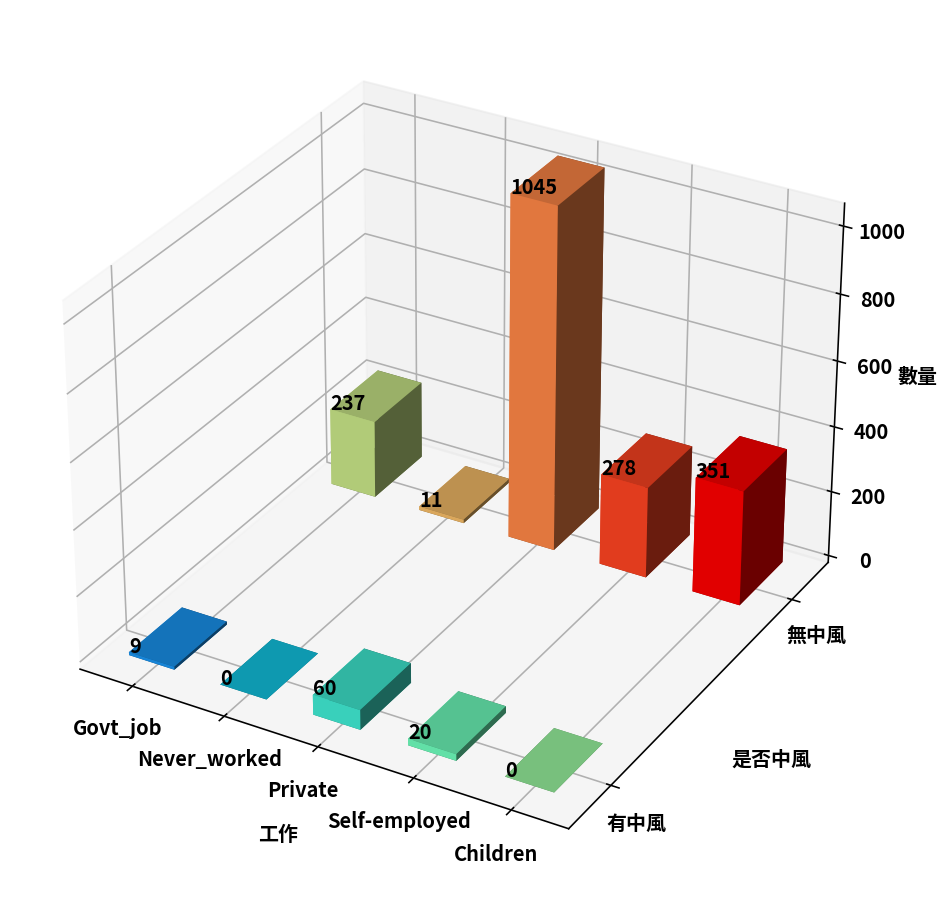
男性BMI指數對中風的影響



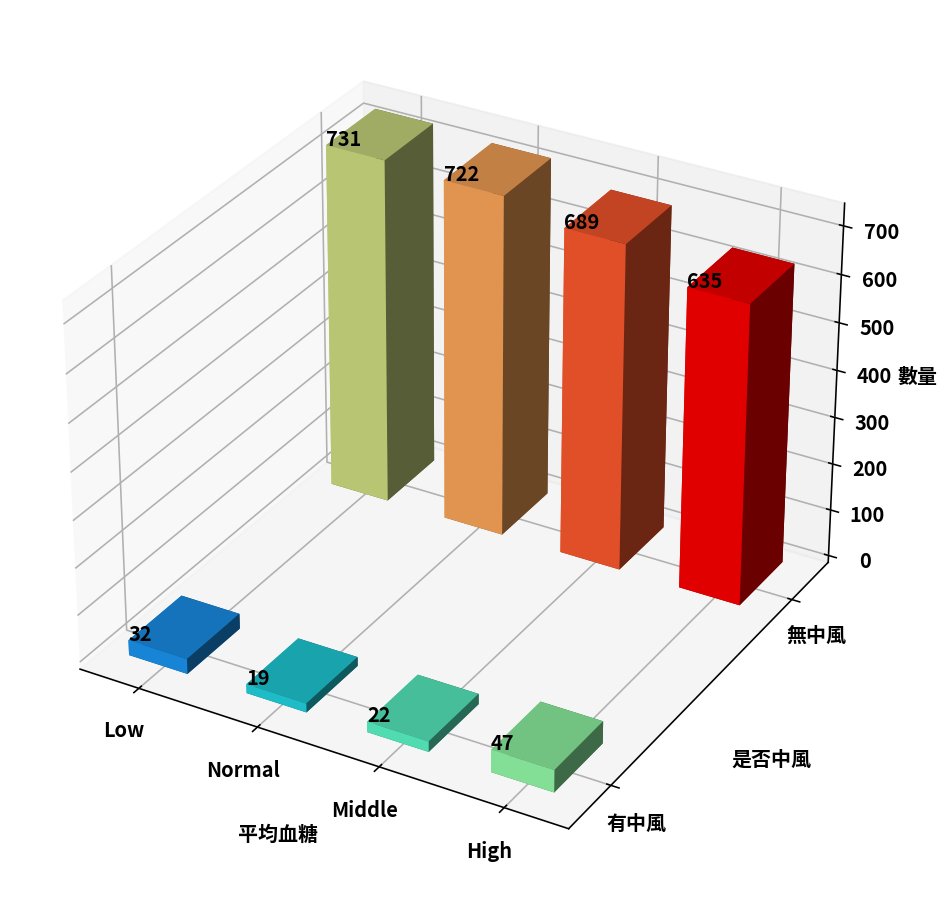
女性職業對中風的影響



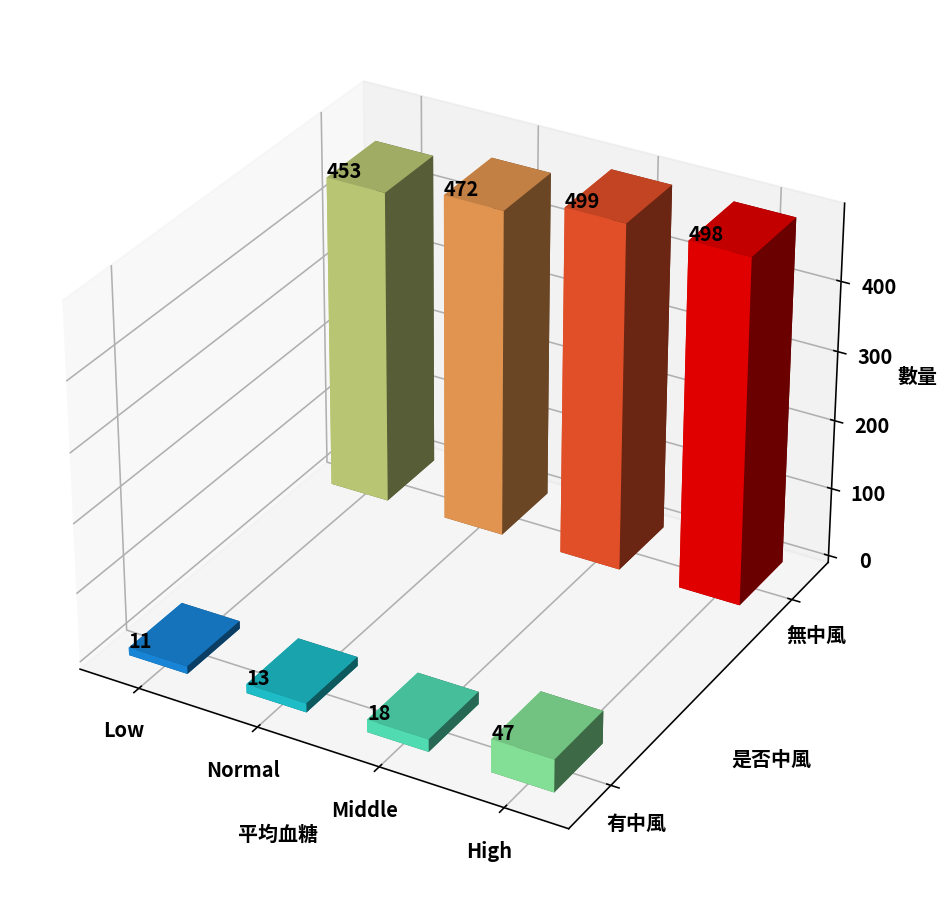
男性職業對中風的影響

****

女性平均血糖對中風的影響

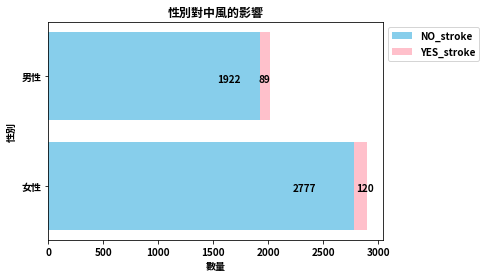


男性平均血糖對中風的影響



## 向上截取(Roll-Up)

男性與女性的中風比率分布



## 向下擷取(Drill-Down)

男性與女性的抽菸中風比率分布



[14, 29, 25, 21]

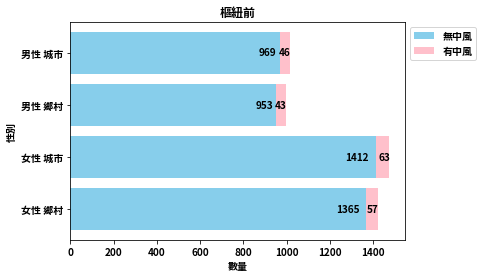
[15, 28, 59, 18]

796 430 1144 407

658 349 624 291

1. 轉軸(Pivot)

男性與女性的居住環境對是否中風的影響





1365 1412 953 969

[57, 63, 43, 46]

# 分析結論

1. 在有中風的年齡層從中年到老年有往上增長的趨勢。
2. 資料集中女性資料多於男性資料，所以在有中風的女性高於男性，屬正常範圍。
3. 有無高血壓在中風的占比，有高血壓和沒有高血壓在有中風的占比為60:149，是將近三成的比例。所以高血壓對中風的影響有少許比例。
4. 在中風者是否患有心臟病中，無心臟病高出有心臟病的人，所以心臟病也跟中風比較無關。
5. 女性和男性的BMI指數對中風的影響，可以看出，在過輕BMI值得中風人數遠低於其他範圍BMI值得中風人數。
6. 女性職業對中風的影響中，可以感覺到在有中風中，自僱人士比私人的的比例多一些。而男性職業對中風的影響，在有中風中，政府工作比平均的比例低了一些。
7. 女性平均血糖對中風的影響，在低血糖與高血糖有中風的人數高出其他範圍的人數。而男性平均血糖對中風的影響，在高血糖影響較多。
8. 從男性與女性的抽菸中風比率分布來看，曾經抽過菸及抽菸的人有中風的比例都比沒抽菸的人有種風的比例高一些。
9. 最後的轉軸部分，男性與女性的居住環境對是否中風的影響不大，分別占的比例都差不多。
10. 結合上述的分析資料，我們發現有些欄位像是心臟病和居住環境對重風的影響的關係比較小，而其他欄位像是年齡、高血壓、BMI值、職業、平均血糖和抽菸，就可以比較明顯看出影響的大小。

經事實證明:中風常發生於中年以後的男性、肥胖、高血壓或糖尿病控制不良、工作忙碌、心情緊張、盛怒爭吵、用力解便者。在做完以上分析以及看完這些數據後，發現不能單看單看一個決定，各個因數都會在有中風的影響占一點比例，更包括一些生活上的因數。