1. 目的、研究動機

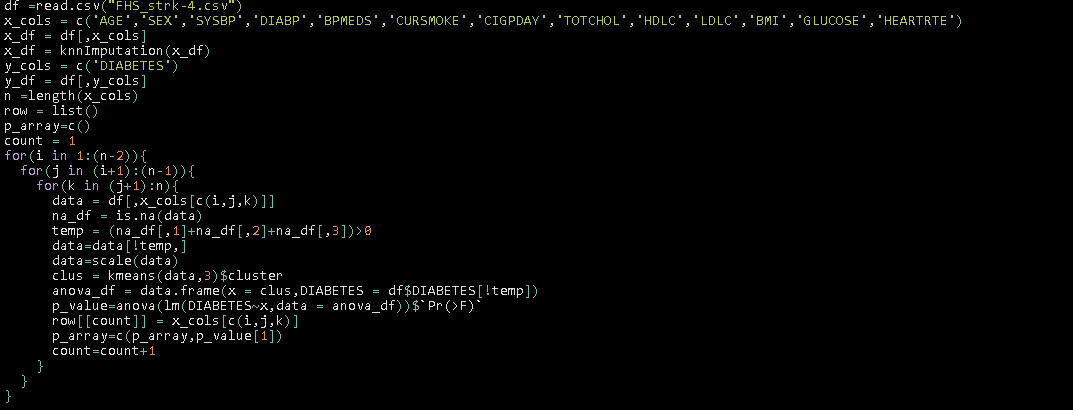
分析此份資料，試圖找出對DIABETES關鍵影響因子，若有複數，則將複數因子轉化成綜合評估指標，試圖以此指標評估得到人群是否為高風險群。

1. 手法: KMEANS、PCA、ANOVA
2. 資料分析
3. 群聚分析找出關鍵因子

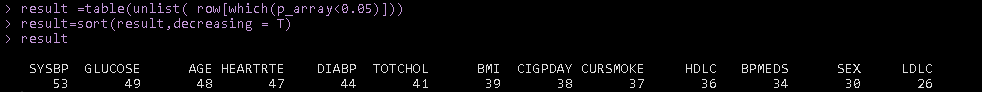
X資料集 : AGE,SEX,SYSBP,DIABP,BPMEDS,CURSMOKE,CIGPDAY,TOTCHOL,HDLC,LDLC,BMI,GLUCOSE,HEARTRTE

-每次以3個因子，做為1組，固有13\*12\*11/3!=286組組合

每組組合，以kmeans方式進行分3組，驗證此分組，是否有對糖尿病有無顯著影響。紀錄每組p-value值



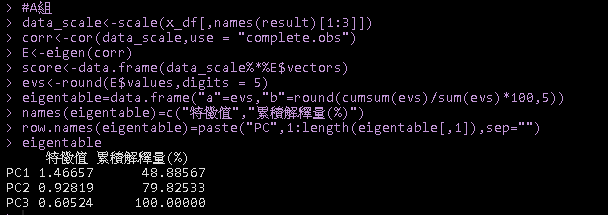
挑選p-value小於0.05組合，並進行統計因子出現次數

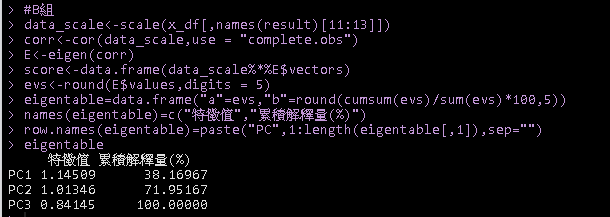


挑選次數出現最多前3名，作為A組

挑選次數出現最少前3名，作為B組

將A、B組以PCA方式進行資料降維(目標1維)。





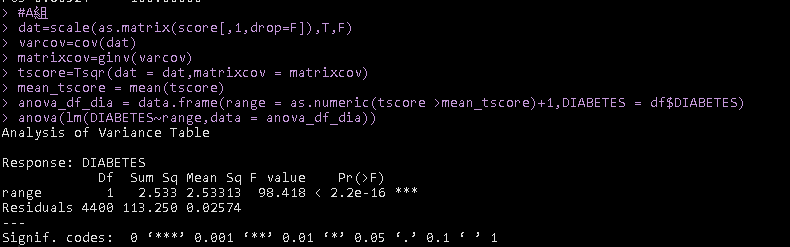
PC數挑選，以特徵值大於1作為原則

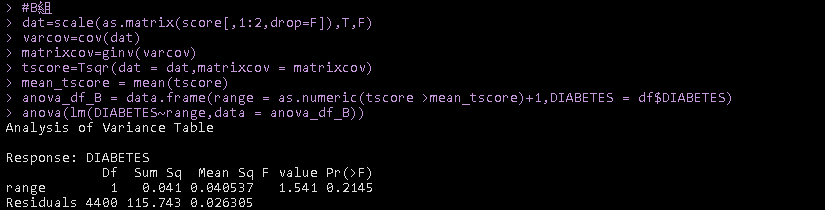
將資料降成一維，並將其所得到數值，分成2組，1-大於平均 2-小於平均，再次以ANOVA方式進行驗證。

虛無假設:

H0: 1組得到糖尿病人數與2組得到糖尿病人數相當

H1: 1組得到糖尿病人數與2組得到糖尿病人數不同





由數據可以看出，

A組在經過資料降維後，仍對糖尿病有顯著影響(拒絕H0)

B組在經過資料降維後，對糖尿病無顯著影響

四、結論

糖尿病患病風險，與血糖、收縮壓、年齡有一定正相關性。

在糖尿病確診條件中，與血糖有直接的條件定義，即空腹8小時以上血糖  126 mg/dl，故在此次分析中，得到糖尿病關鍵因子有此因子，是在預期中的。而收縮壓、舒張壓此三條件，也有一定的相關性，可以提供後續相關研究繼續分析。