Q1.



以此表格作為例子，{茶->咖啡}confidence為75%，喝咖啡的人為80%；

假設有一sample，他喝咖啡的機率為80%，但是「如果他喝茶」，他喝咖啡的機率卻變成75%，所以他喝咖啡的機率從80%掉到75%。

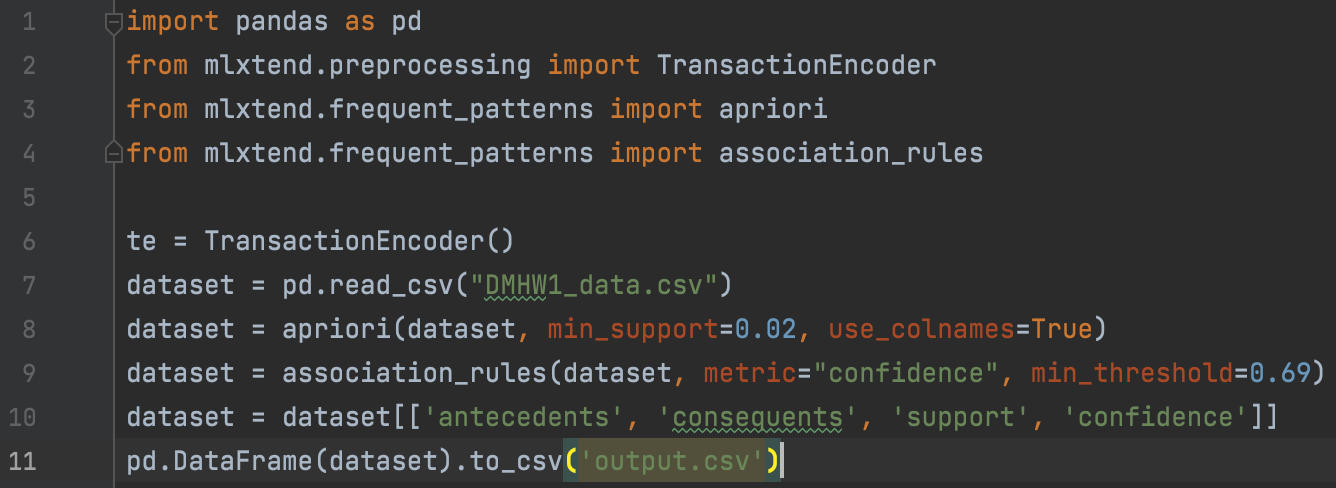
從lift的公式可以看出{茶->咖啡}confidence除以咖啡的support，咖啡的support即咖啡佔總體的比例，所以咖啡所佔比例越大對於{茶->咖啡}confidence影響越小，反之則越大，這解決了confidence沒有考慮到{咖啡}本身的機率。

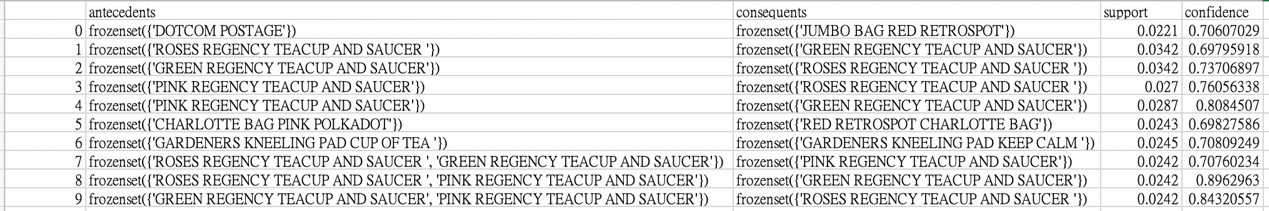
Q2.

Confidence為asymmetric以Q1為例，C{茶->咖啡}=75%、C{咖啡->茶}=18.75%，又或是『f11 / f1+ 不等於 f11 / f+1』很明顯f1+並不一定等於f+1；lift是symmetric，以Q1為例75%x1.25=18.25%x5，又或是『(f11 x N) / (f1+ x f+1) 等於 (f11 x N) / (f+1 x f1+)』從公式中不難看出兩式相等。

Q3.

計算前10筆rule

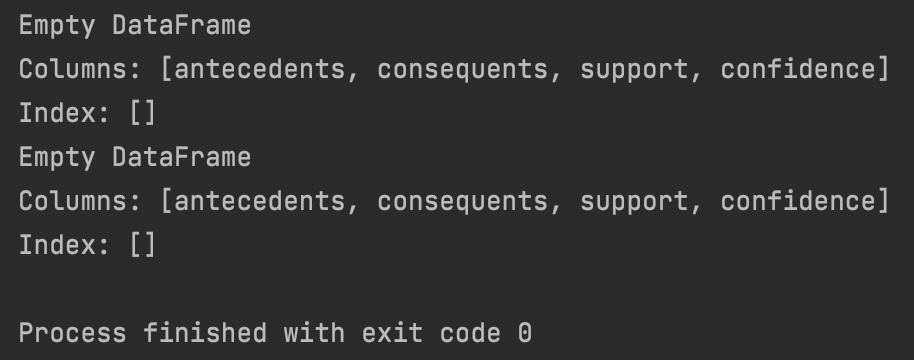
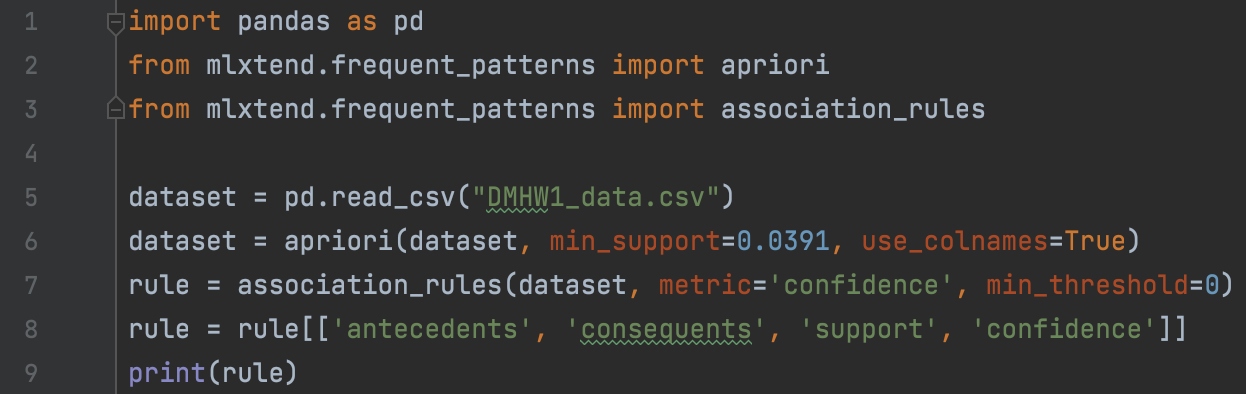




Confidence最高的前10筆中REGENCY TEACUP AND SAUCER出現的次數最多

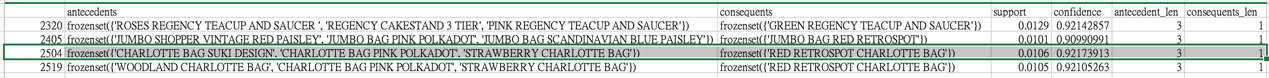
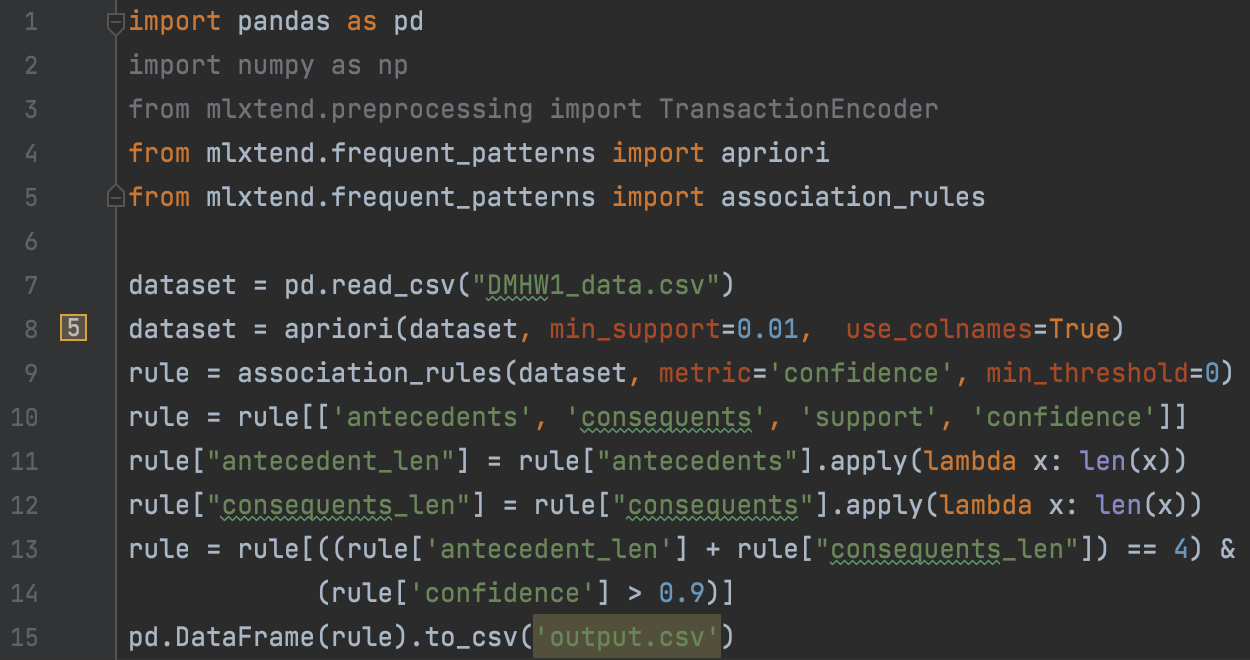
Q4.

0.39，這代表了沒有任何rule的support高於0.39



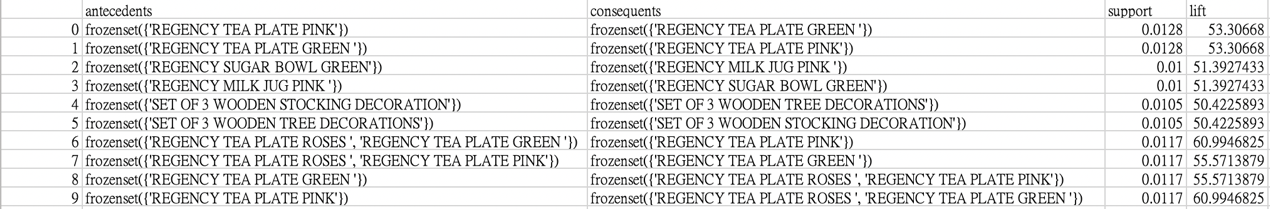
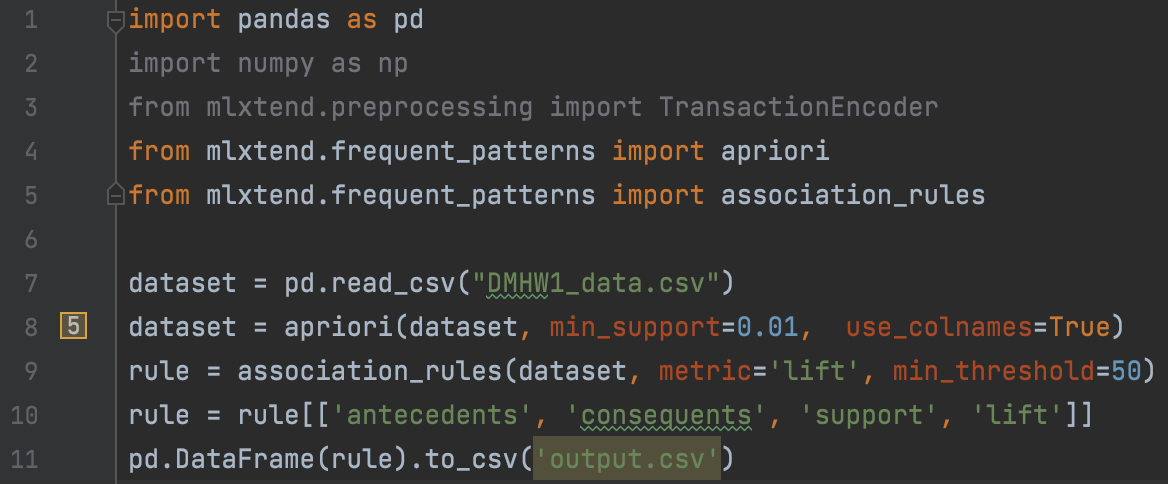
Q5.

{'CHARLOTTE BAG SUKI DESIGN', 'CHARLOTTE BAG PINK POLKADOT', 'STRAWBERRY CHARLOTTE BAG'}🡺{'RED RETROSPOT CHARLOTTE BAG'}



Q6.

Min\_support=0.01，且以lift為metric，會挑lift作為metric是因為association\_rules這個function官方API的document示範了confidence和lift兩種metric，我也懶得仔細看function的內容是否支援其他種metric所以就用lift，那在lift的metric中可以看出，最常出現的item變成了REGENCY TEA PLATE。



Q7.

使用了mlxtend套件執行apriori演算法，有用到pandas但是並不是必須要用；因為dataset的格式已經屬於Numpy\_array所以也不需要預處理；參數設定use\_colnames設true、min\_threshold要根據要篩選到剩下幾筆資料而定，剩下的都是照著題目給出的條件去設定；然後因為我的console沒辦法讓我截圖加上pandas會幫我用…省略掉中間部分所以輸出到一個新的csv檔。