**作業5: 誰和誰一國**

接下來，我們要開始進行一個「相對」而言模糊的分析，也就是集群分析，讓我們開始吧!

沒有最佳解，只要組內相似性越高，組間相異性越高，就會是相對較好的分群

用未知的資料作集群，用已知的資料作分析

演算法：k-means(常用歐基里得距離計算)－只能處理數值型資料

1. 開始之前，讓我們了解一下，集群分析是一項　監督式　非監督式　學習 （請圈選）　(5%)

2.請開啟onlineRetail.csv這是一個交易記錄的資料集。請進行必要之前處理前，並請存為ETL，也請稍加說明您做了那些處理(10%)

Step 1

使用excel做資料前處理，將原始檔案onlineRetail.csv，改為以CustomerID為主體，儲存為新檔案onlineRetail\_new.csv，留下的欄位如下：

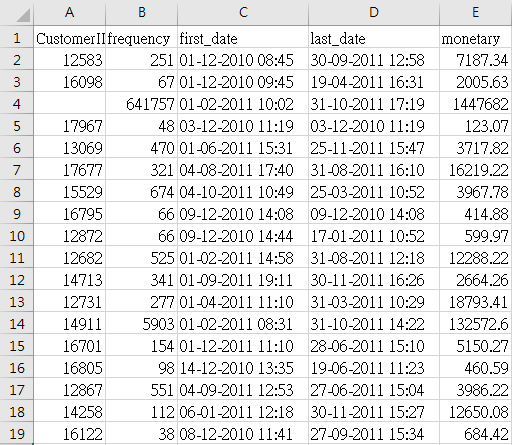
CustomerID

frequency(計算個別CustomerID的列數→即是每一位顧客消費次數)

first\_date(同一CustomerID中InvoiceDate最早的日期→即是每一位顧客第一次購買日期)

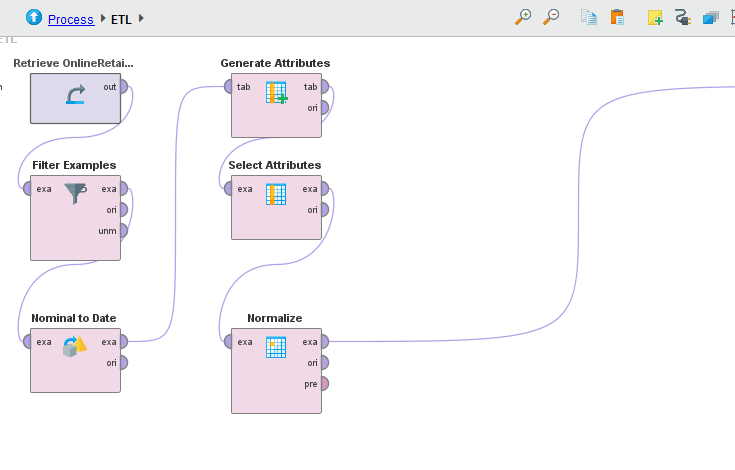
last\_date(同一CustomerID中InvoiceDate最晚的日期→即是每一位顧客最後一次購買日期)

monetary(同一CustomerID中Quantity\*UnitPrice的加總→即是每一位顧客的消費總金額)



Step 2

* 將新檔案匯入RapidMiner
* 過濾值 (將缺失值刪除)
* 轉換格式(將last\_date的格式從nomial轉換為date)
* 新增欄位(新增recency欄位，計算顧客最後一次購買日與今天日期差幾天)
* 篩選欄位(將CustomerID篩選掉，僅留下後續會用到的欄位)
* 正規化(將數值正規化到0-1的區間，讓資料較符合常態分佈，並減少離群值對的影響)



3. 接下來，我們要使用傳說中的RFM模型。但等等，目前的資料裡並沒有這3個欄位呀，那怎麼辦呢? 請產生這3個欄位唷，說明如下:

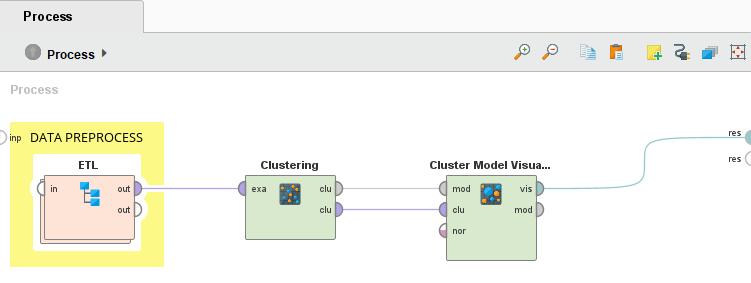
R: 距離今天的消費日期（recency）

F: 總共消費的次數（frequency）

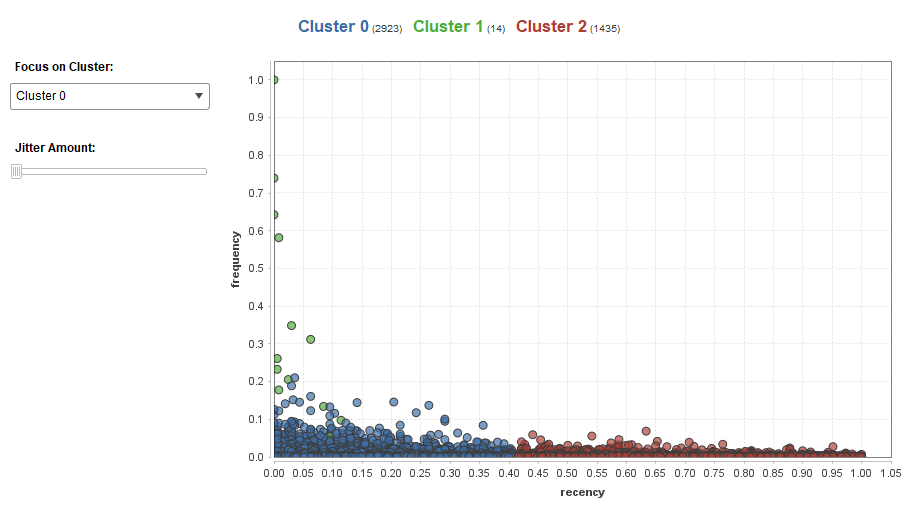
M: 總共消費的金額（monetary）

第二題中已提到過

第3題，我們是從消費者的角度出發唷! 請分別將R/F/M之資料產制方式存入Subprocess. (20%) 如下圖所示



4. 最後，請對上述整併後的資料進行Cluster分析，並請說明相關結果，請將結果存為學號\_1.rmp(15%)。並請匯製一張您覺得最能說明相關分析結果之分群說明圖，加以說明(5%)。



從圖中可以看到：

藍色群(R低F低)屬於最近有購買，且消費次數很少的客群，可能是賣場的新客戶，因此賣方應該為客戶提供良好的消費體驗，給他好的印象，以留住這類型的客戶

紅色群(R高F低)屬於最近沒有購買，且消費次數也很少的客群，可能是賣場過去的未留住的新客戶，可以透過分析此類型的客戶來去探討客戶流失的原因

綠色群(R低F高)屬於最近有購買，且消費次數也很多的客群，屬於賣場應重點關注的客群，是賣場的潛在VIP顧客，應想辦法留住這類顧客

5. 接下來，我們要改變分析角度，變成是對產品面項使用傳說中的RFM模型，但等等，目前的資料裡還是沒有這3個欄位呀，所以請產生這3個欄位唷，說明如下:

R: 距離今天的最近一次被購買的時間

F: 總共被購買的次數

M: 總共被購買的金額

資料前處理

Step 1

使用excel做資料前處理，將原始檔案onlineRetail.csv，改為以CustomerID為主體，儲存為新檔案onlineRetail\_new\_stock.csv，留下的欄位如下：

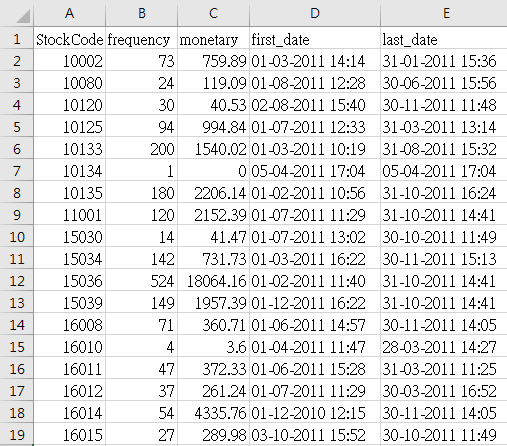
CustomerID

frequency(計算個別StockCode的列數→即是每一個商品的被購買次數)

first\_date(同一StockCode中InvoiceDate最早的日期→即是每一個商品第一次被購買的日期)

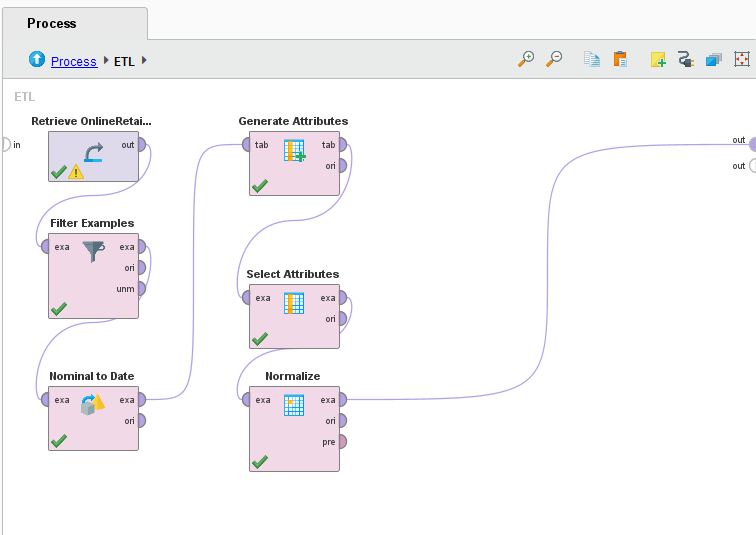
last\_date(同一StockCode中InvoiceDate最早的日期→即是每一個商品最後一次被購買的日期)

monetary(同一StockCode中Quantity\*UnitPrice的加總→即是每一個商品被購買總金額)

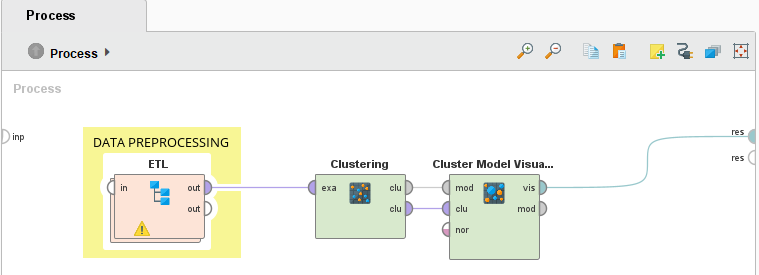


Step 2

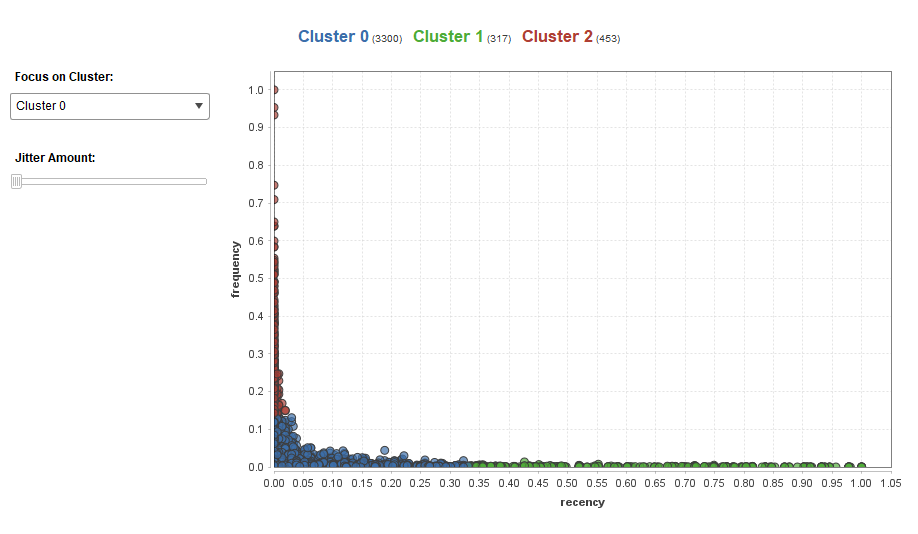
* 將新檔案匯入RapidMiner
* 過濾值 (將缺失值刪除)
* 轉換格式(將last\_date的格式從nomial轉換為date)
* 新增欄位(新增recency欄位，計算商品最後一次被購買日與今天差幾天)
* 篩選欄位(將StockCode篩選掉，僅留下後續會用到的欄位)
* 正規化(將數值正規化到0-1的區間)



第5題，我們是以產品的角度出發唷! 請分別將R/F/M之資料產制方式存入Subprocess. (20%)



6. 最後，請對上述整併後的資料進行Cluster分析，並請說明相關結果，請將結果存為學號\_2.rmp(15%)。並請匯製一張您覺得最能說明相關分析結果之分群說明圖，加以說明(10%)。



從圖中可以看到：

藍色群(R低F低)屬於最近有被購買，且被購買次數很少的商品，可能屬於新進商品，可以將這類商品放置於宣傳物顯眼處，使顧客能一眼看到這類商品，提升這類商品的被購買率

綠色群(R高F低)屬於最近沒有被購買，且被購買次數也很少的商品，可能屬於被淘汰的商品，可以降低此類商品的進貨量，以降低整體進貨成本

紅色群(R低F高)屬於最近有被購買，且被購買次數也很多的商品，屬於賣場應重點關注的商品，可能可以為賣場提升營收，可將這類商品擺放在商場中顯眼的位置，做為商場標誌商品