

化整為零之集群分析

- ✓ 監督式學習／非監督式學習
- ✓ 非監督式學習－K-means

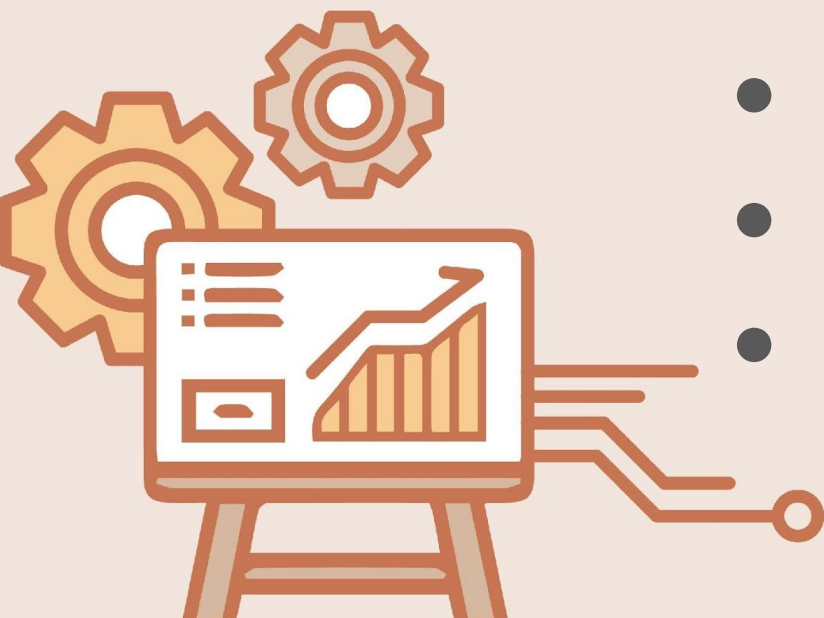


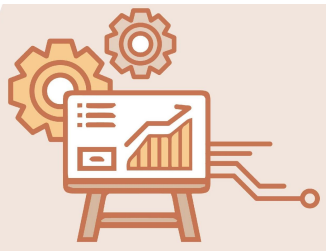
目錄

1. 什麼是非監督式學習？
2. 集群演算法介紹
3. 資料集介紹
4. RapidMiner上戰場

一、什麼是非監督學習？

- 監督式學習／非監督式學習
- 非監督學習的應用
- 非監督學習的實際案例





監督式學習／非監督式學習

監督式學習

Supervised Learning

所有資料都被「**標註**」(label)，告訴機器相對應的值，以提供機器學習在輸出時判斷誤差使用

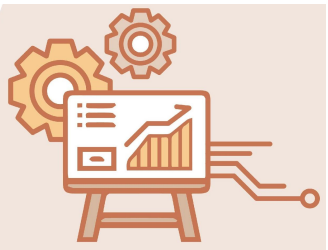
非監督式學習

Unsupervised Learning

所有資料都沒有標註，**機器透過尋找資料的特徵**，自己進行分類

類別	功能	演算法	
監督式學習 Supervised	預測 Predicting	Linear Regression	線性迴歸
		Random Forest	隨機森林
非監督式學習 Unsupervised	分類 Classification	Logistic Regression	邏輯迴歸
		Decision Tree	決策樹
	分群 Clustering	K-means	K平均法
	關聯 Association	FP Growth	關聯性分析

▲ 監督式、非監督式學習比較

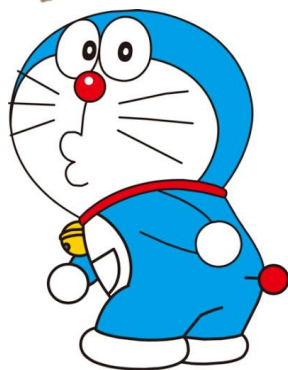


監督式學習／非監督式學習

監督式學習

Supervised Learning

有標準答案

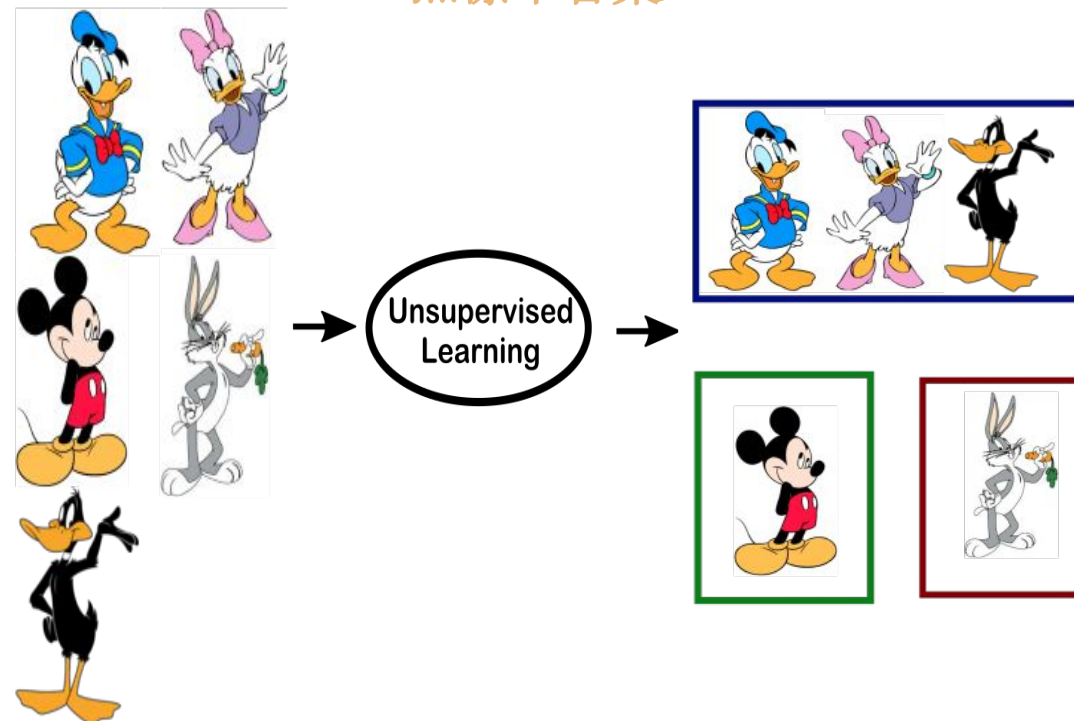


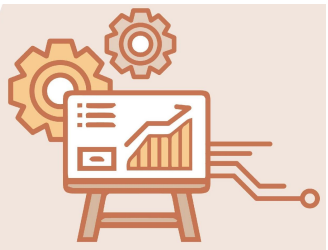
:阿我是誰咧？

非監督式學習

Unsupervised Learning

無標準答案





非監督學習的應用

銀行行銷

將客戶根據其特徵分為不同的群組，並向每個群組定制不同的行銷策略



醫學研究

根據診斷圖像、基因表達等數據分類，發現潛在的疾病，以幫助診斷和治療



社交媒體分析
將用戶行為數據進行分類和聚類，發現用戶的偏好趨勢



物流管理

路徑優化、運輸需求預測、貨物分類等進行聚類和分析

1. 提高運輸效率
2. 降低運輸成本



地震監測

根據地震測量數據進行地震預測、捕捉前兆



文檔分類
器

物品傳輸優
化

識別犯罪地
點

客戶分
類

球隊狀態分
析

保險欺詐檢
測

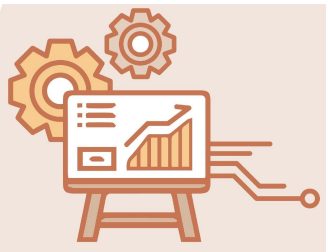
乘車數
據

網絡犯罪分
子

呼叫記錄詳
細

IT警報
自動化聚
類

▲ 其他應用

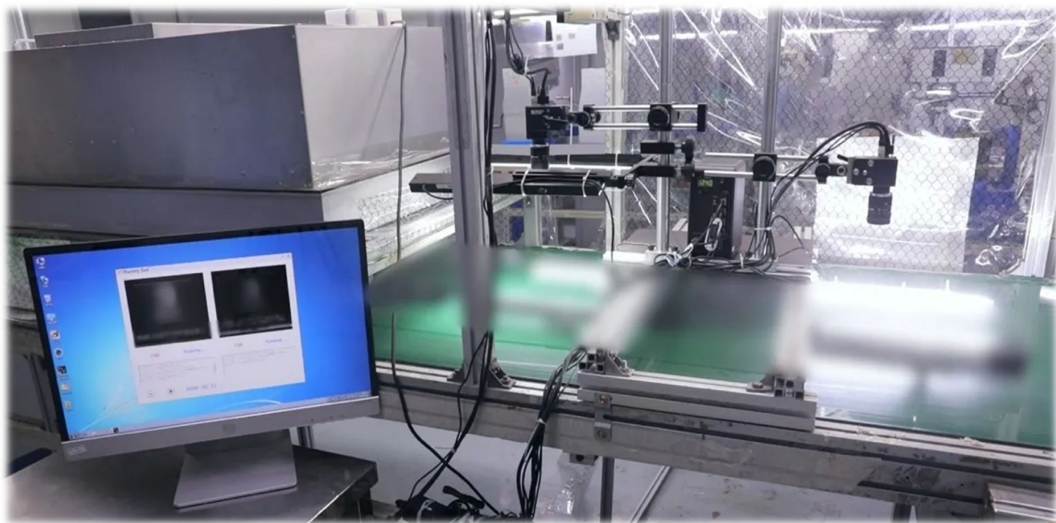


非監督學習的實際案例

鴻海研發 **非監督式學習演算法**，智慧工廠人力再砍 50%！

鴻海科技集團在 2021 年宣布推出非監督式學習人工智慧演算法「**FOXCONN NxVAE**」，運用正面表列的模型訓練方式，以 產品容易取得的正樣本進行光學檢測演算，解決 產線中瑕疵樣本取得的問題，適用於良率高的成熟 產品線，可增加 AI 模型的整體容錯能力。

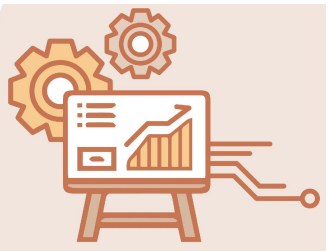
➡ 實際應用園區 內的電子 產品外觀檢測 產線上，成功降低 50% 以上的產線檢測人力



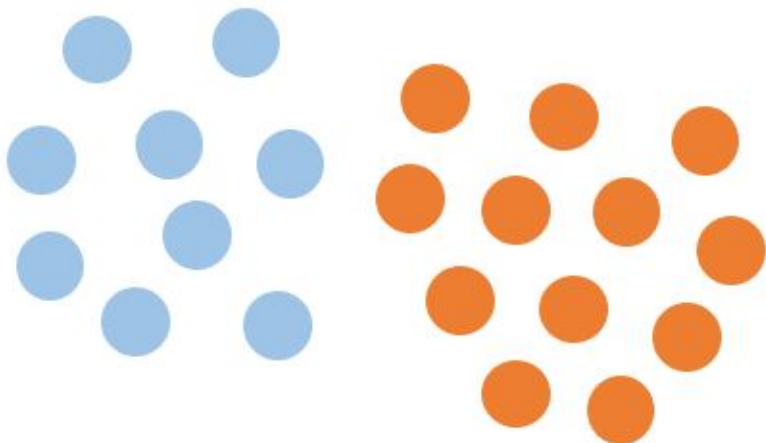
有別於監督式學習在瑕疵影像採集的困難、數據標註與分類的痛點，**Foxconn NxVAE**導入正面表列的訓練方式，沿用原本產線每日皆可取得的正樣本，解決瑕疵樣本問題，適應不同產品的智能檢測，大幅度縮短客戶導入AI檢測的時間壓力，協助定義產品檢測標準，提升生產品質、降低成本，最終提升產業價值的目標！

二、集群演算法

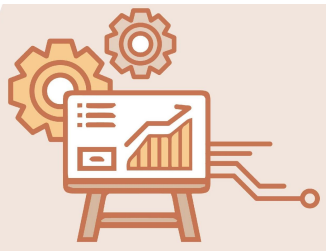




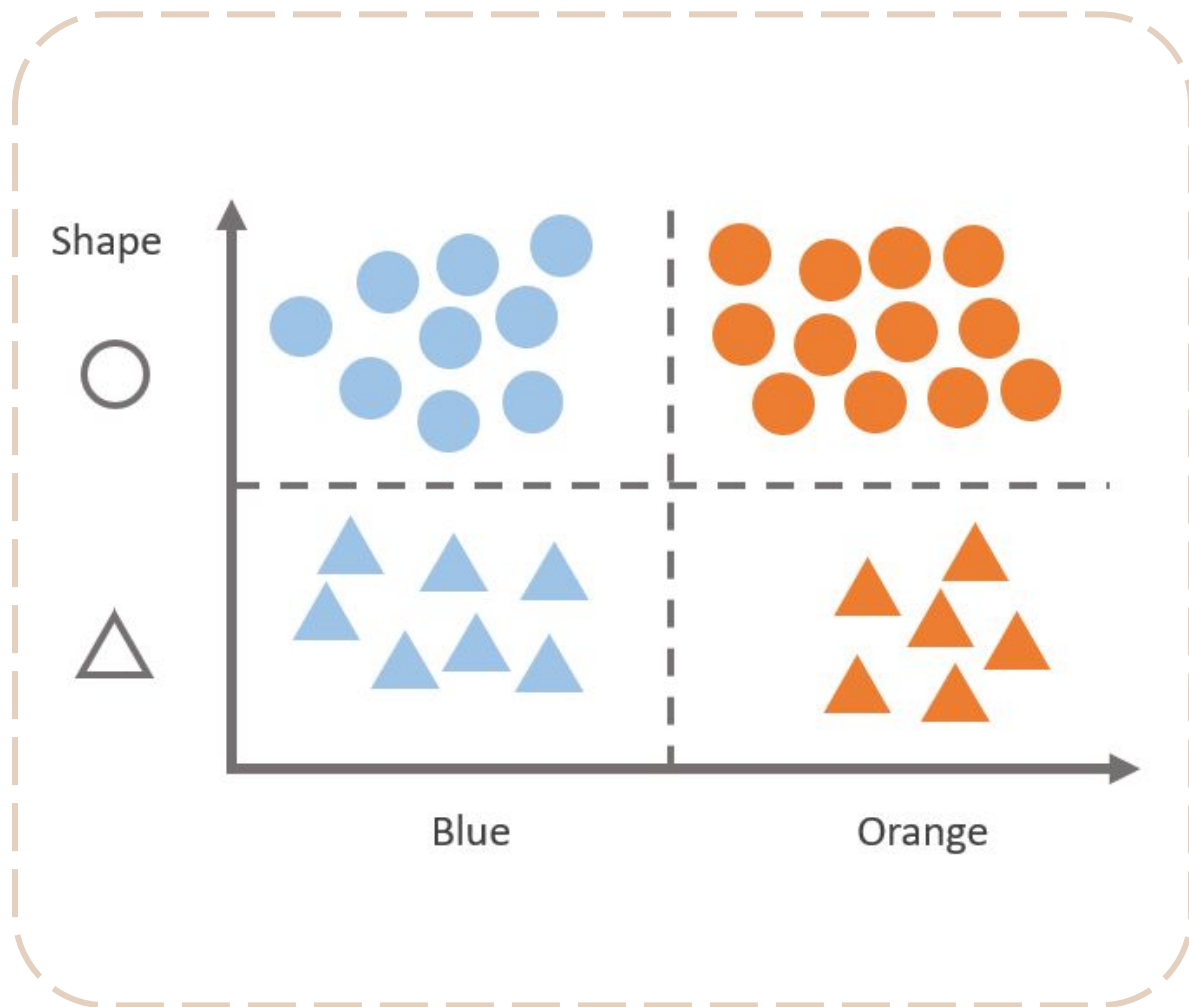
集群演算法 ---「物以類聚」。



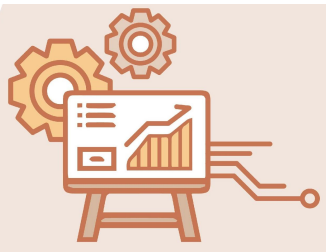
- 所有數據進行分組，相似數據歸類於同一組，一筆數據只屬於某一組，每一組稱作一個「群集 Cluster」



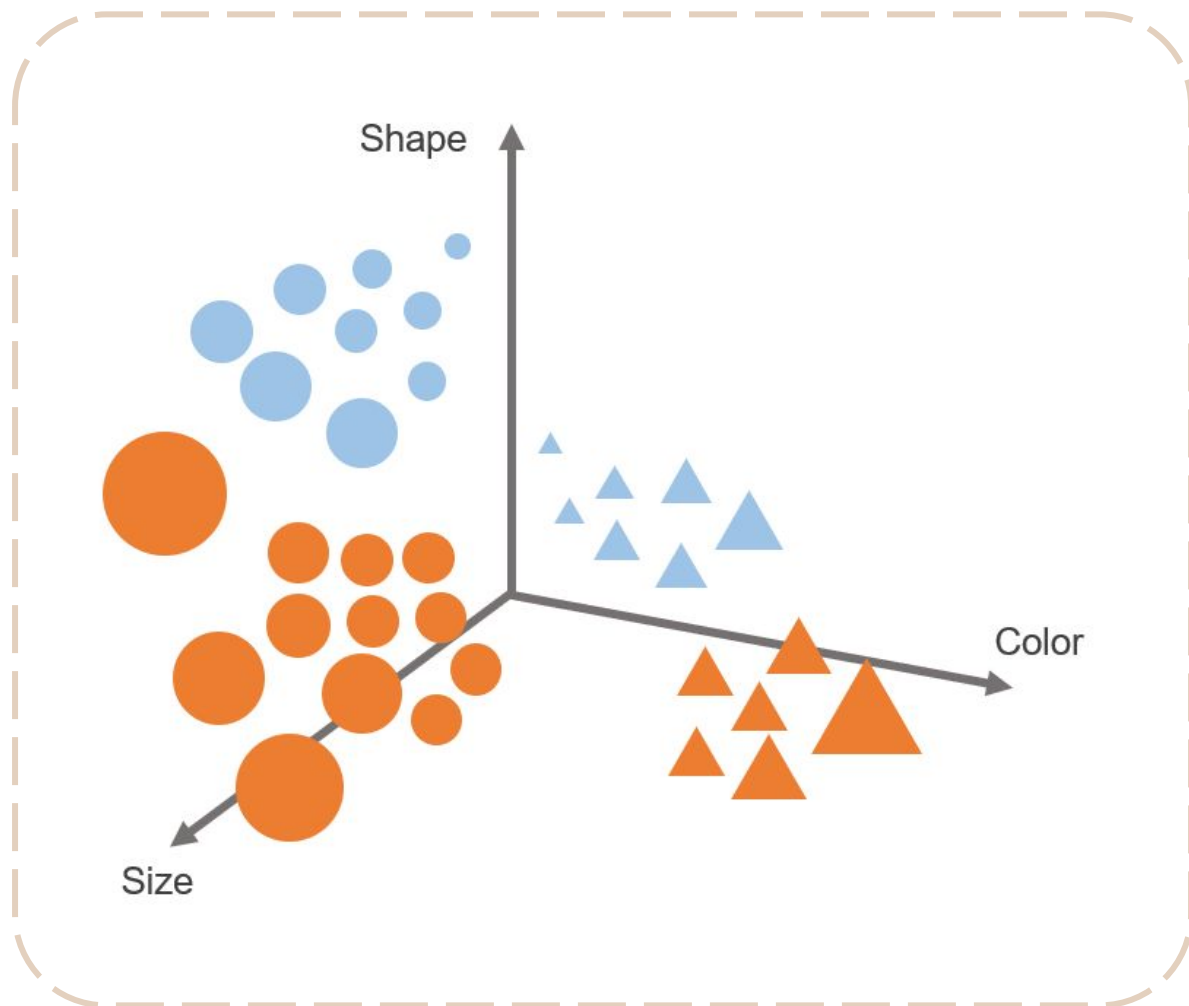
集群演算法 --- 近朱者赤、近墨者黑



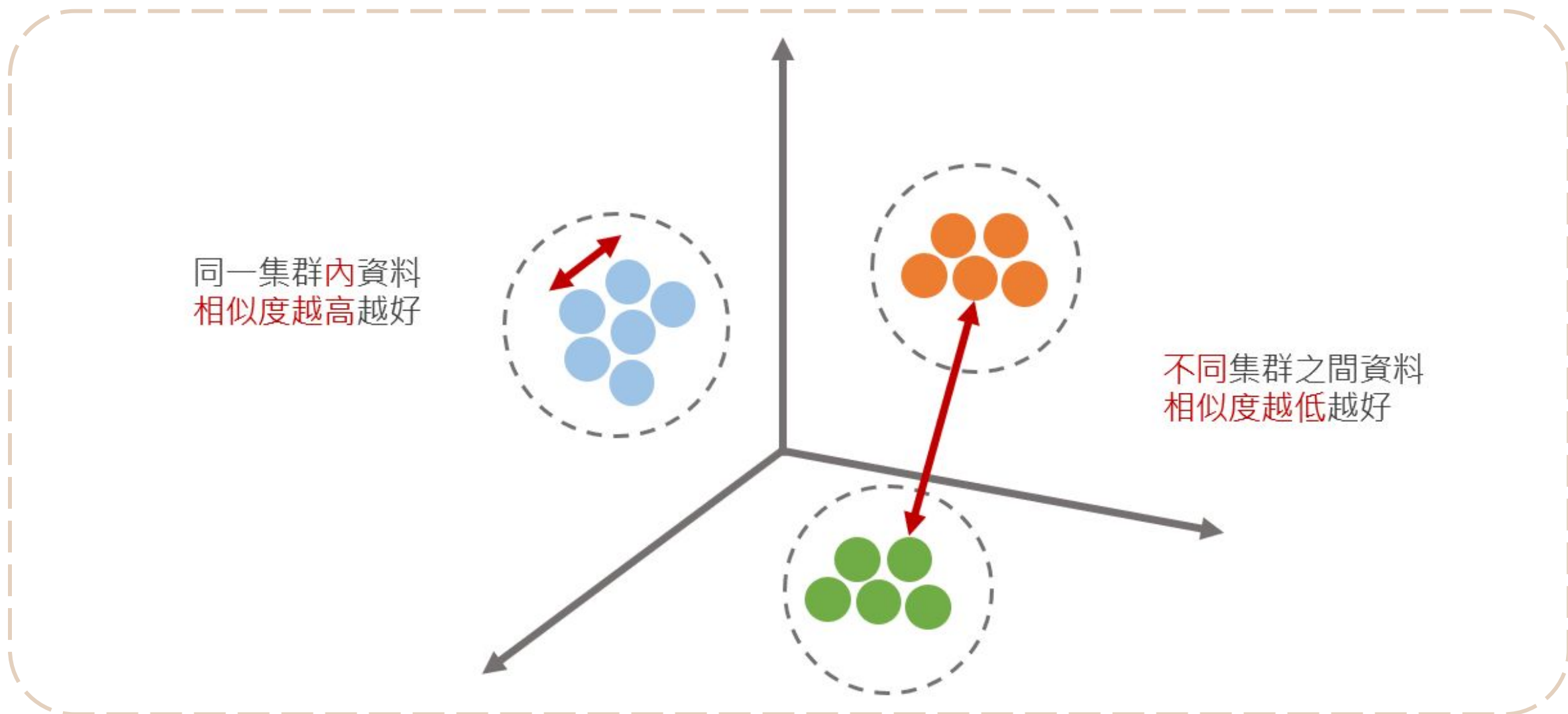
- 依屬性做區分
- 距離越近，推定為越相似



集群分析



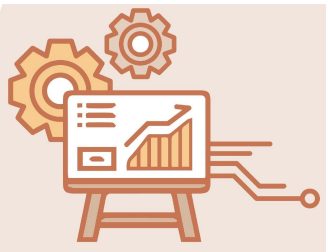
- 適用於多維特徵空間中的分群，並使得分群結果在相似度測度上最佳化
- 可以透過其他套件降維



▲ 集群演算法目標

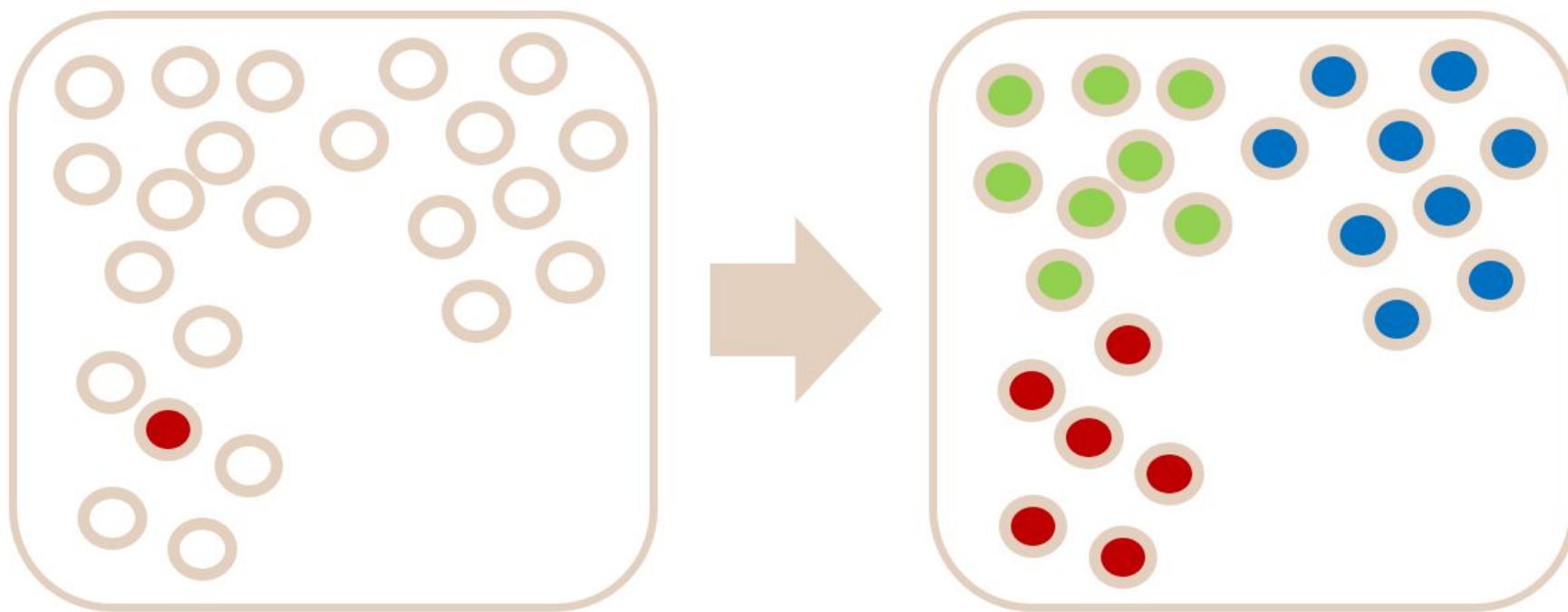
K-means演算法

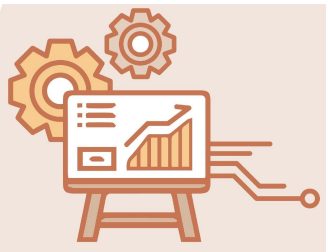




K-means演算法

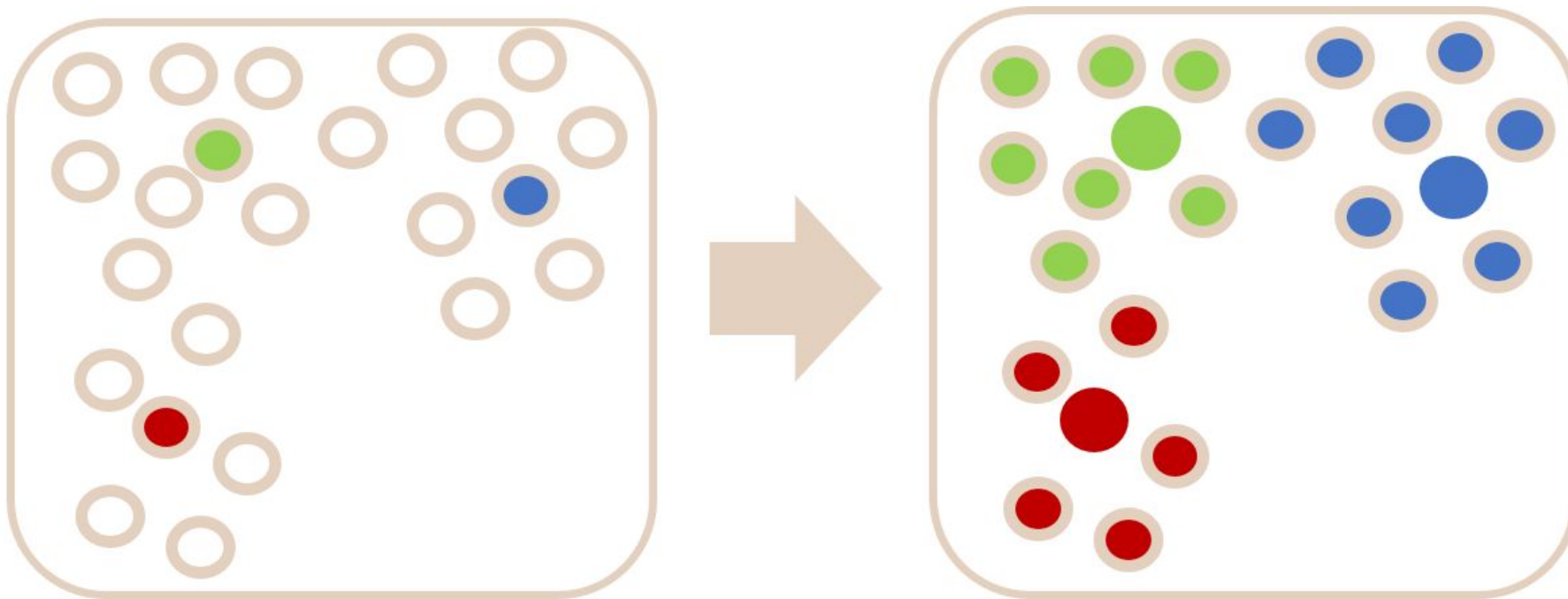
- 第一步 □ 隨機選取資料組中的n筆資料，先決定好k值，要把它們分成幾群

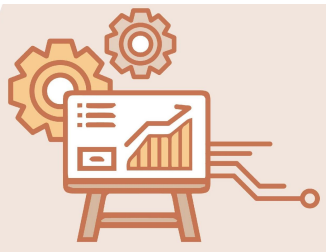




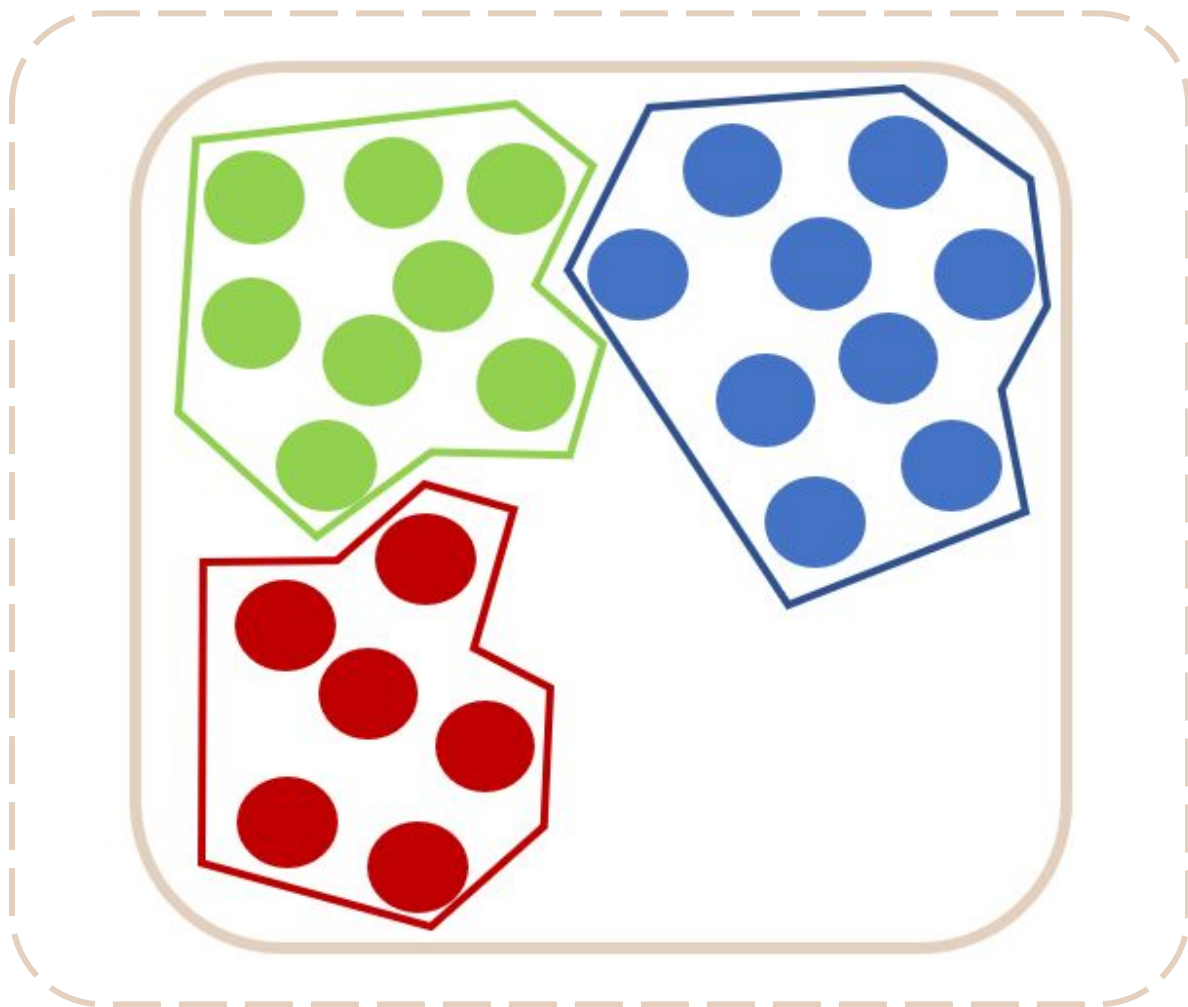
K-means演算法

- 第二步 □ 決定中心點

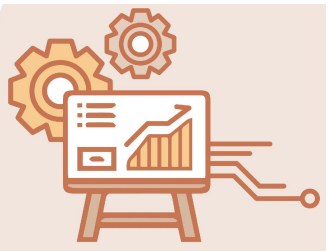




K-means演算法



- 第三步 □ 每一個人去計算跟中心點的位置
- 第四步 □ 分好組後，重新計算中心點
- 第五步 □ 重複以上步驟，直到中心點不再改變



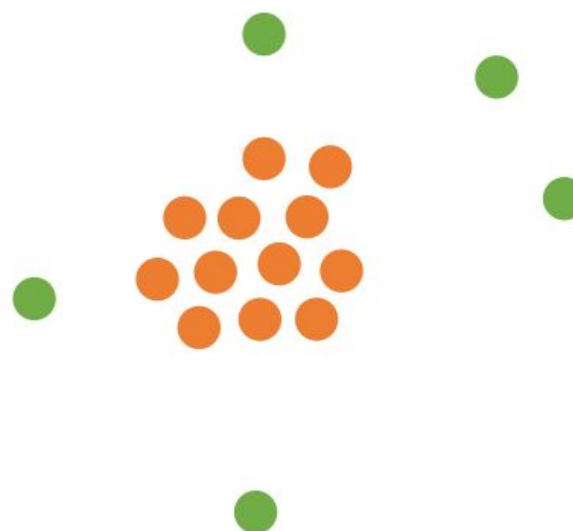
K-means演算法

• K-means演算法有一些限制，當...

1. 大小不一

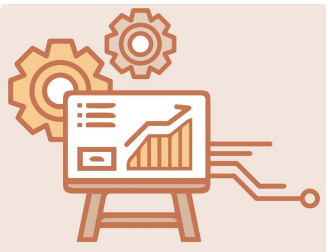


2. 密度(距離)不同



3. 數據分布呈甜甜圈





K-means怎麼選K？

- 計算每筆資料和中心點的距離，用來比較群集誤差的結果
- 手肘法 (Elbow Method)

誤差平方和(SSE): 計算各點到cluster中心的距離的平方的和

- 輪廓係數法 (Silhouette Coefficient)

「找出相同群凝聚度越小、不同群分離度越高」的 值

將所有的點都計算 S 後再總和。S 值越大，表示效果越好，適合作為 K

$$S = \frac{b - a}{\max(a, b)}$$

a: 與相同群內的其他點的平均距離

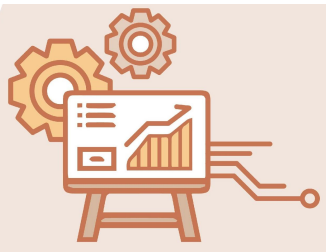
b: 與不同群的其他點的平均距離

S: 以一個點作為計算的 值



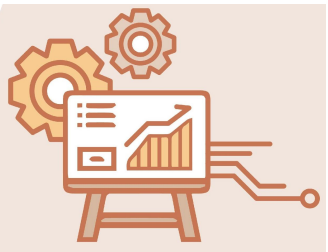
三、資料集介紹





資料集介紹

欄位名稱	資料型態	分布	欄位名稱(中文)
recency	Date-time	2018/12/31~2019/10/23	客戶最近一次的消費日期
frequency	Integer	01~246	客戶消費的頻率
monetary	Integer	約13萬~百億美元	客戶消費的總金額
company_id	Nominal	類別、屬性無大小之分	客戶的公司id
province	Nominal		運送地址的州/省
district	Nominal		運送地址的行政區
village	Nominal		運送地址的村(里)
address	Nominal		客戶的運送地址
phone_number	Nominal		客戶的聯絡電話



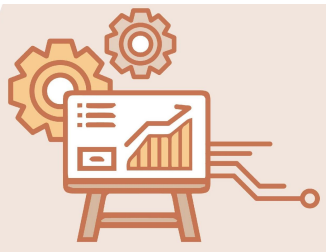
資料集的應用

• 企業如何將顧客價值進行分群？

• 顧客關係管理中的 RFM 模型

隨著顧客的購買通路越來越多樣，企業可以透過觀察顧客行為，將顧客分群針對不同類型的顧客提供差異化行銷，來優化顧客體驗。





資料集的應用

• 為什麼要使用 RFM模型？

• 協助認識顧客

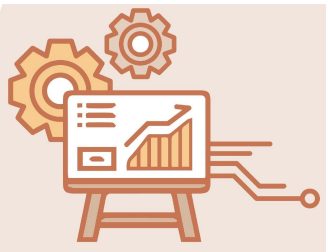
可以從分類中找出顧客的特性，並進一步提出個人化行銷，將行銷預算用在刀口上。

舉個例子：RFM模型可以帶我們了解哪一群會員是品牌忠誠客戶？哪些會員即將流失，而需要怎麼盡快挽救的？




透過了解會員行為，可以擬定出更適當的會員策略。

• 直觀好操作

RFM模型使用的門檻較低，僅需要公司本來就有存取的顧客資料、簡單的統計工具，以及對RFM 模型架構的基本認識，便可以執行。透過幾個簡單的分析步驟，就能更深入了解顧客在品牌中的行為。



資料集的應用

 Recency 最近消費日	 Frequency 消費頻率	 Monetary 消費金額
最近有互動（指消費、點擊、轉換）的顧客活躍度較高	互動頻率越高，對公司的黏著度越高	消費金額高，表示消費能力高
適用於 CRM、EDM 或 Web 的分析指標		
<ul style="list-style-type: none">• 上次的消費日期• 上次打開電子郵件的日期• 上次潛在顧客轉換的日期	<ul style="list-style-type: none">• 一段時間內購買的次數• 一段時間內開信的次數• 一段時間內的潛在顧客轉換次數	<ul style="list-style-type: none">• 一段時間內顧客花費的總金額• 一段時間內，根據每個顧客的獲得成本及利潤等因素估算的價值• 一段時間內從不同指標得出的顧客參與度分數

RFM模型的三大指標

- (x軸) frequency (F)

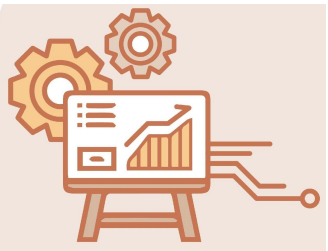
: 一段時間內訂購商品的次數

- (y軸) recency (R)

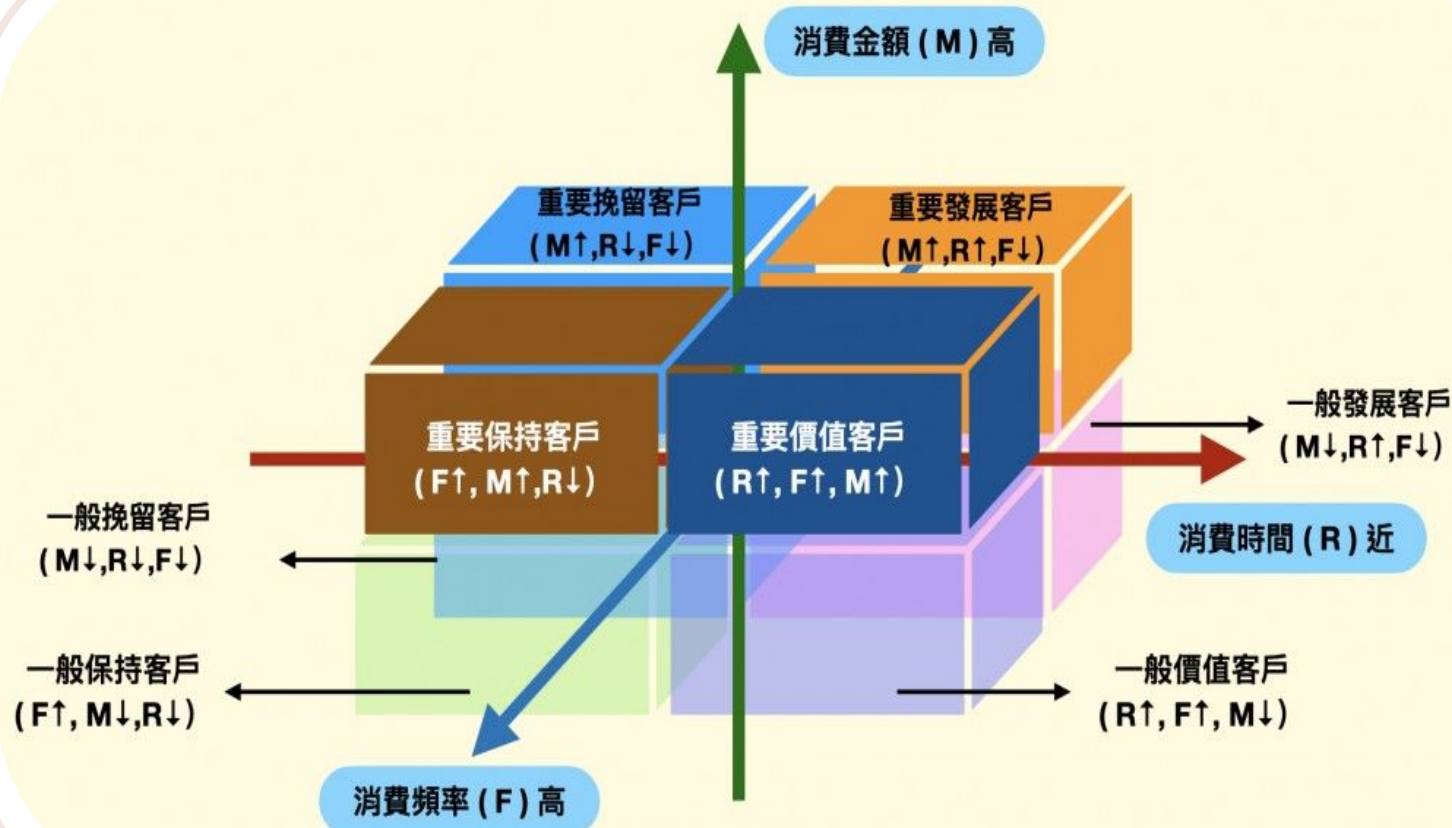
: 最近一次訂購商品的時間

- (z軸) monetary (M)

: 一段時間內訂購商品的總金額



資料集的應用



RFM指標的高低

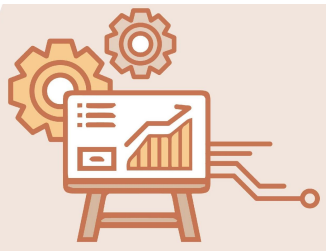
客戶根據RFM指標的高低可以區分出八種價值的客戶，每一類型的客戶都有相對應的行銷策略。

重要發展客戶 (M↑, R↑, F↓)

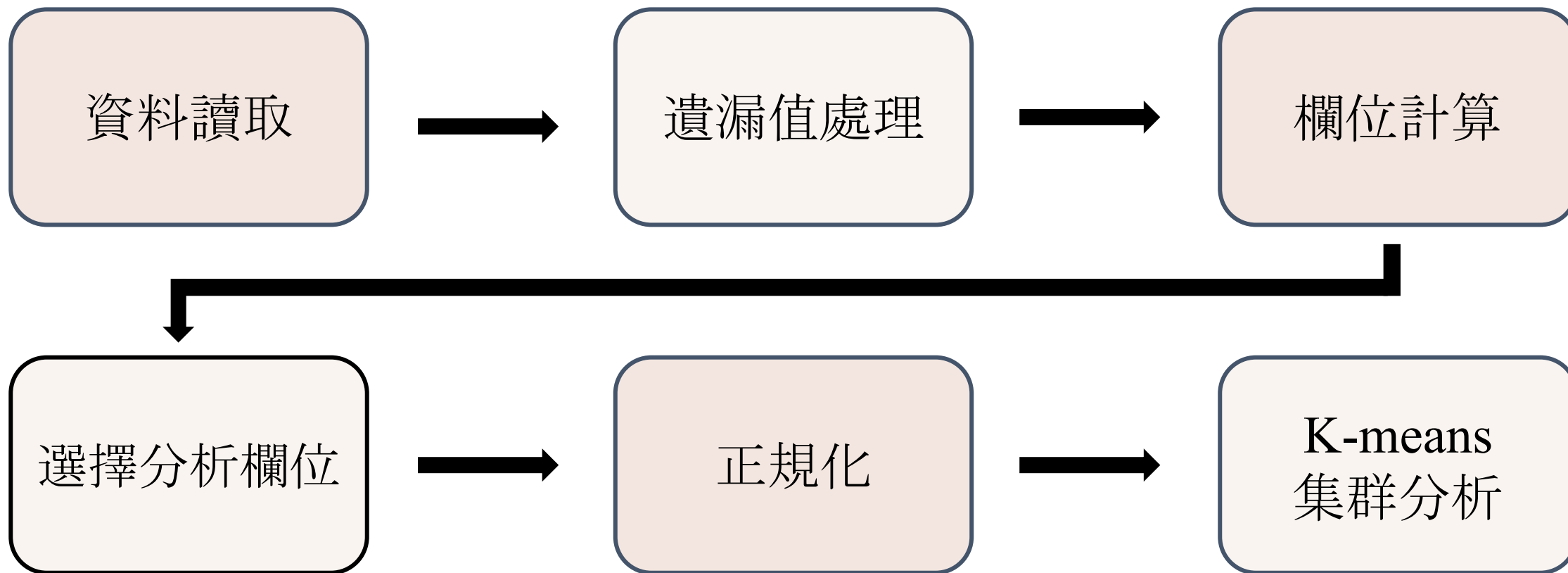
這類型的客戶已經有了第一次的消費，且消費金額偏高，是不可或缺的客戶類型。為了提升該類型客戶的消費頻率，可以透過信件提醒或是消息推播提醒，並應盡可能培養成重要價值客戶。

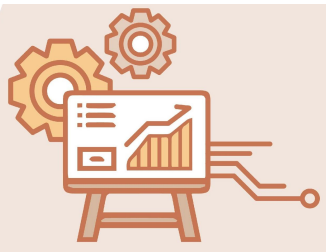
四、RapidMiner上戰場





集群分析流程



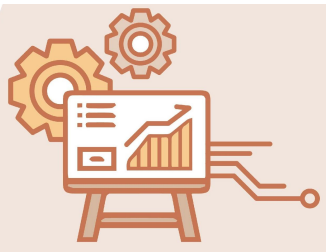


Step 1. 資料讀取

1. 左下角Operators搜尋框中輸入「read」
2. 在Read目錄中找到「Read CSV」
3. 將Read CSV拖曳至右側Process，並點擊兩下
4. 選擇放置於桌面的「cluster_practice.csv」

The screenshot illustrates the steps to read a CSV file in a data processing tool. The interface is divided into three main panels: Repository, Process, and Operators. The Operators panel shows a search for 'read' and a list of operators, with 'Read CSV' selected. The Process panel shows the 'Read CSV' operator being added to the workflow. The 'Import Data' dialog shows the selection of 'cluster_practice.csv' from the desktop.

File Name	Size	Type	Last Modified
--- Last Directory			
cluster_practice.csv	61 KB	Microsoft Excel 逗點分隔值檔...	May 6, 2023
金融大數據			



Step 1. 資料讀取

Import Data - Format your columns.

Format your columns.

Date format: yyyy/MM/dd 95%

2. ☒ Replace errors with missing values

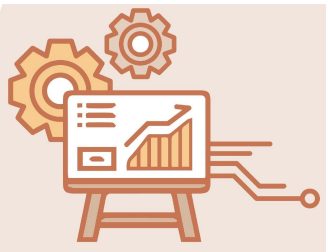
	recency date	frequency integer	monetary integer	company_id integer	address polynomial	province polynomial	district polynomial	village polynomial
1	Dec 31, 2018	2	1382000	544	Toko Susu Cariss...	Jawa Barat	Kab. Bekasi	Karanganyar
2	Dec 31, 2018	1	7200000	268	Perumahan Lemb...	Jawa Barat	Kota Depok	Cipayung Jaya
3	Jan 2, 2019	1	5242000	210	Jalan Pemuda No...	Jawa Bara	Bekasi	Kranji
4	Jan 4, 2019	1	1918200	223	Jalan LAMPIRI RT ...	Jawa Barat	Kota Bekasi	Jatibening Bar
5	Jan 5, 2019	1	4684800	237	JLN PORIS RAW...	Jawa Barat	PONDOK MELATI	Kota Bekasi
6	Jan 5, 2019	2	5273800	429	Jl.swatantra II no. 1...	Jawa Barat	Kota Bekasi	Jatiasih
7	Jan 7, 2019	1	5453200	711	Jalan Lampiri rt 4 r...	Jawa Barat	Pondok Gede	Bekasi
8	Jan 8, 2019	1	20665000	5	/1 kelurahan karan...	Jawa Barat	Kab. Bekasi	Karangasih
9	Jan 8, 2019	1	20320000	202	Jalan Tipar Halim 4...	Jawa Barat	Depok	Mekarsari
10	Jan 8, 2019	1	1333133	333	Rempah Speed ...	DKI Jakarta	Kota Jakarta Pusat	Gunung Sahar
11	?	?	?	?	?	?	?	?
12	Jan 10, 2019	1	1516800	722	Perumahan Perm...	Banten	Kota Tangerang S...	Bakti Jaya
13	Jan 10, 2019	1	4398400	867	perumahan Green ...	Banten	Cinodoh	Kota Tangerang
14	Jan 10, 2019	2	19509200	427	Toko Susu Bila Kid...	Jawa Barat	Kab. Bogor	Pakansari
15	Jan 11, 2019	1	1993000	872	Jl. Bandengan Utar...	DKI Jakarta	Kota Jakarta Utara	Penjaringan
16	Jan 11, 2019	1	4165600	718	Jalan Rawajati Bar...	DKI Jakarta	Kota Jakarta Selat...	Rawa Jati
17	Jan 12, 2019	1	1771600	887	Jl. Duri Raya No. 3...	DKI Jakarta	Gambir	Jakarta Pusat

1. ?

3. 5 warnings View Details

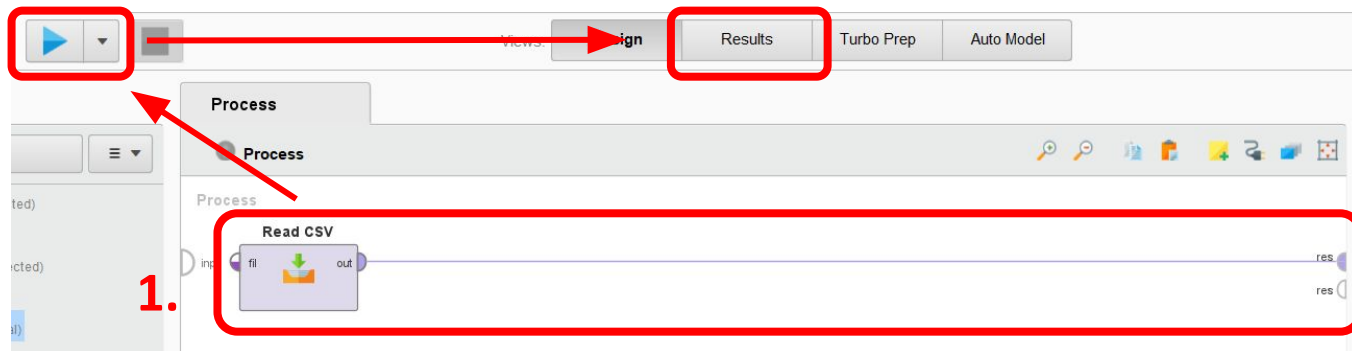
Previous Finish Cancel

1. 點擊Next後，進入最後格式設定的畫面，可以發現有異常資料
2. 勾選「Replace errors with missing values」，以遺漏值方式取代異常資料
3. 點擊「Finish」，完成資料讀取



Step 2. 遺漏值處理

2.

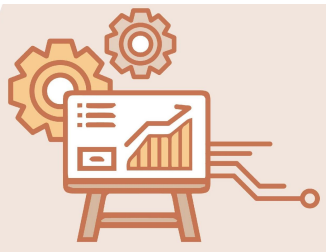


1. 將Read CSV的out連線至res
2. 點擊左上角執行按鈕，再點擊Results，以查看執行後結果

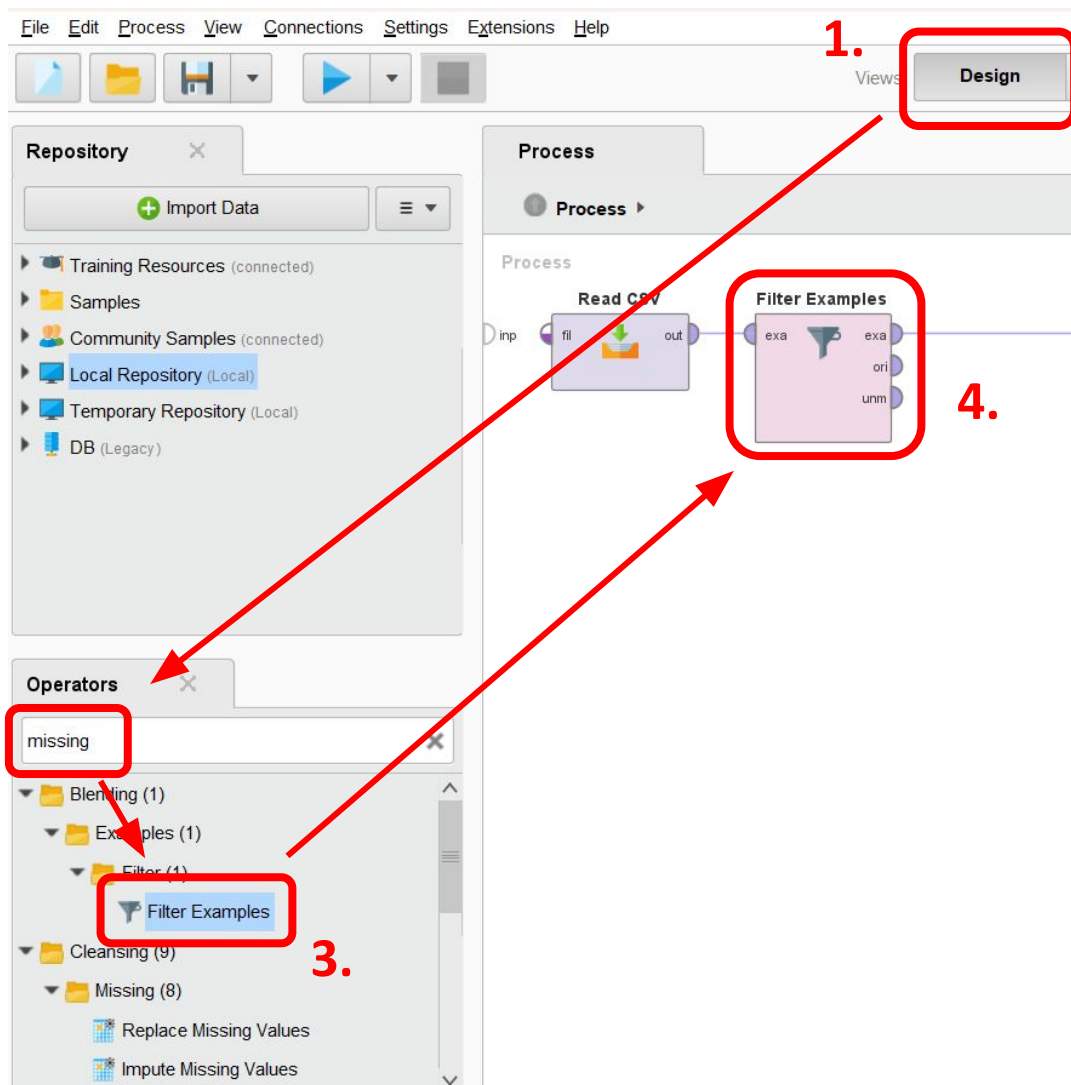
3.

Name	Type	Missing	Statistics
recency	Date-time	20	Earliest date: Dec 31, 2018; Latest date: Oct 23, 2019; Duration: 296 days
frequency	Integer	20	Min: 1; Max: 246; Average: 9.157
monetary	Integer	20	Min: 133200; Max: 17120020000; Average: 370122719.764
company_id	Integer	20	Min: 36; Max: 7981; Average: 2448.815
address	Nominal	20	Least: taman pu [...] no.21 (1); Most: Jl pondo [...] no 14 (2); Values: Jl pondo [...] h1 no 14 (2), Perumaha [...] Pertanian (2), ...[428 more]
province	Nominal	20	Least: Yogyakarta (1); Most: DKI Jakarta (210); Values: DKI Jakarta (210), Jawa Barat (147), ...[5 more]
district	Nominal	20	Least: sukasari (1); Most: Kota Jakarta Barat (47); Values: Kota Jakarta Barat (47), Kota Jakarta Timur (39), ...[85 more]
village	Nominal	20	Least: karangasih (1); Most: Depok (8); Values: Depok (8), Cengkareng Timur (7), ...[276 more]

3. 執行結果中，點擊Statistics，閱覽描述性統計資料，可以發現到幾乎每一個欄位都存在著遺漏值



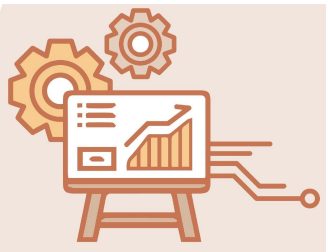
Step 2. 遺漏值處理



1. 點選上方「Design」，回到作業區
2. 於左下角搜尋「missing」
3. 在Filter目錄中，找到「Filter Examples」
4. 將Filter Examples拖曳至線上，並點擊兩下，以打開Create Fliter頁面

※ 學生筆記－為什麼是選擇過濾，而非遺漏 值填補？

遺漏值可能存在於同一筆資料中的多個欄位，進行過濾能夠先將這些幾乎不具代表性/影響力、過多遺漏值的資料優先排除



Step 2. 遺漏值處理

1. **recency**

2. **is not missing**

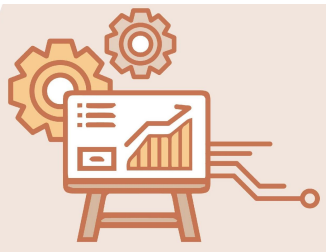
3. **OK**

Name	Type	Missing	Statistics
recency	Date	0	Earliest Date: Dec 31, 2018; Latest Date: Oct 23, 2019; Duration: 296 days
frequency	Integer	0	Min: 1; Max: 246; Average: 9.157
monetary	Integer	0	Min: 133200; Max: 17120020000; Average: 370122719.764
company_id	Integer	0	Min: 36; Max: 7981; Average: 2448.815
address	Polynome	0	Latest: laman pu [..] no 21 (1); Oldest: Jl pondo [..] no 14 (2); Values: Jl pondo [..] h1 no 14 (2), Perumaha [..] Pertanian (2), [426 more]
province	Polynome	0	Latest: Yogyakarta (1); Oldest: DKI Jakarta (210); Values: DKI Jakarta (210), Jawa Barat (147), [5 more]
district	Polynome	0	Latest: sokasari (1); Oldest: Kota Jakarta Barat (47); Values: Kota Jakarta Barat (47), Kota Jakarta Timur (39), [85 more]
village	Polynome	0	Latest: karangash (1); Oldest: Depok (8); Values: Depok (8), Cengkareng Timur (7), [276 more]

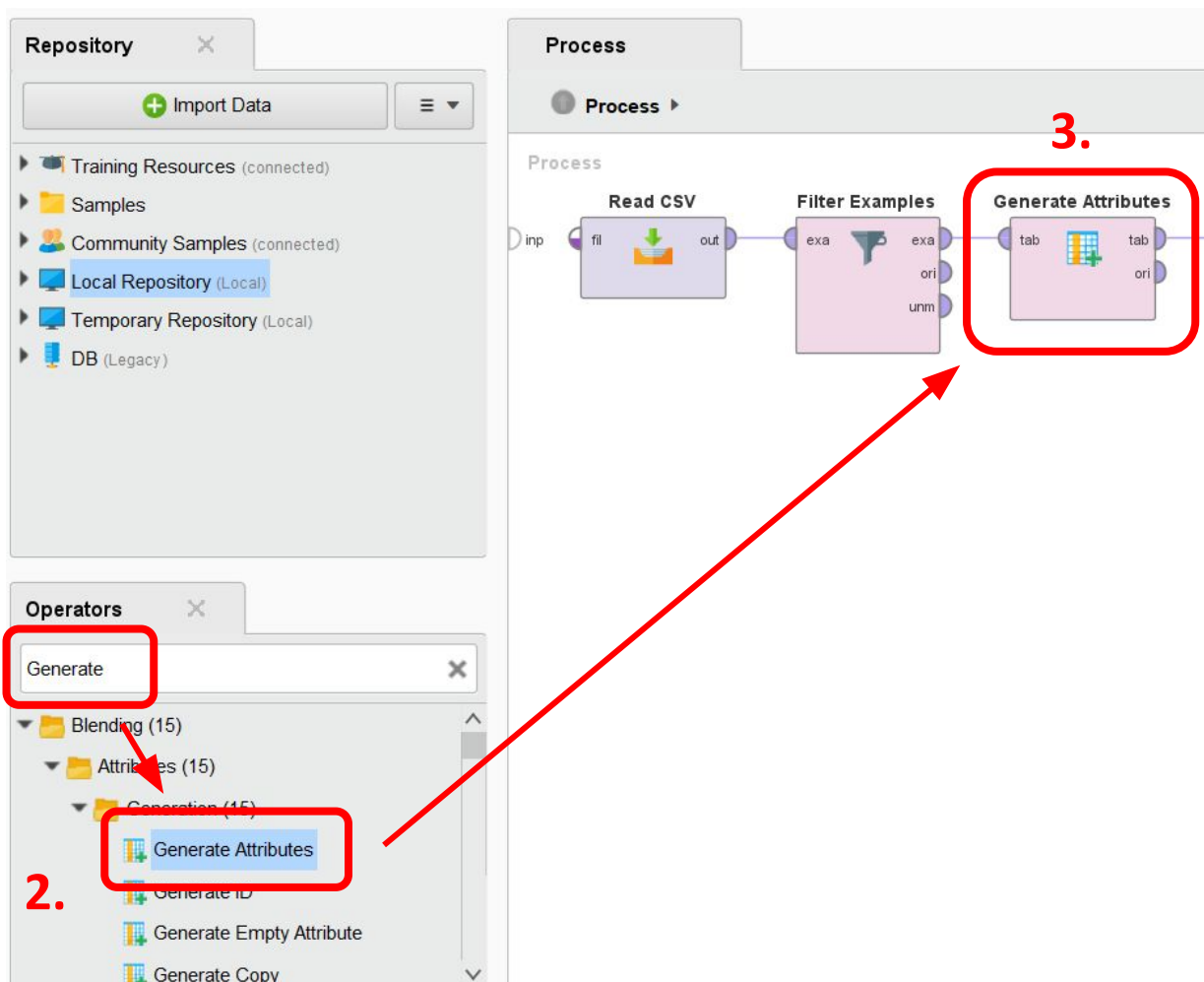
1. 選擇曾出現遺漏值的欄位「**recency**」
2. 對該欄位的過濾選擇「**is not missing**」，確保欄位必須有值
3. 點選右下角「**OK**」，並**重複**前面檢查遺漏值的動作，可以發現到**資料中已無遺漏值**，完成遺漏值處理

※ 學生筆記－為什麼是選擇 **recency**？

recency是出現過遺漏值的欄位，但不一定非得選擇 **recency**進行過濾，也可以選擇其他欄位，但最終的目標是「減少遺漏 值」的出現



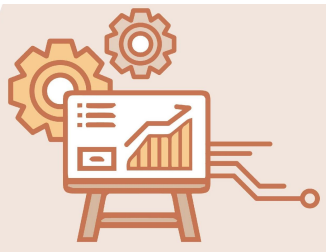
Step 3. 欄位計算



1. 於左下角搜尋欄搜尋「Generate」
2. 在Generation目錄中，找到「Generate Attributes」
3. 將Generate Attributes拖曳至線上，並點擊兩下，以打開Edit Parameter List頁面

※ 學生筆記－為什麼要欄位計算？

RFM中的R是以與最近一次消費的「間隔時間」，作為考量。因此，需要計算最近一次消費時間與現在的間隔



Step 3. 欄位計算

1. 於column name中新增/輸入「r」的欄位

2. 點擊右側計算機，進入Edit Expression頁面

3. 於Expression輸入日期計算公式，如下
`date_diff(recency,date_now(),DATE_UNIT_DAY,"America/Los_Angeles")`

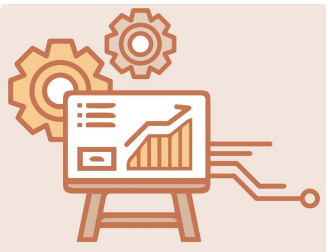
4. 點選兩次「Apply」，完成欄位計算

1. 於column name中新增/輸入「r」的欄位
2. 點擊右側計算機，進入Edit Expression頁面
3. 於Expression輸入日期計算公式，如下
`date_diff(recency,date_now(),DATE_UNIT_DAY,"America/Los_Angeles")`
4. 點選兩次「Apply」，完成欄位計算

※ 學生筆記－公式講解

`date_diff (recency, date_now(), DATE_UNIT_DAY, "America/Los_Angeles")`

日期加減 (開始日, 結束日, 計算後單位(天), 時區)



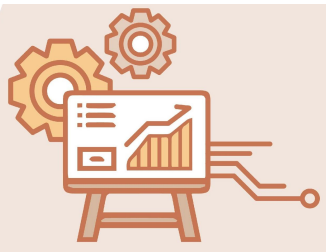
Step 4. 選擇分析欄位

1. 於左側搜尋框中搜尋「select」

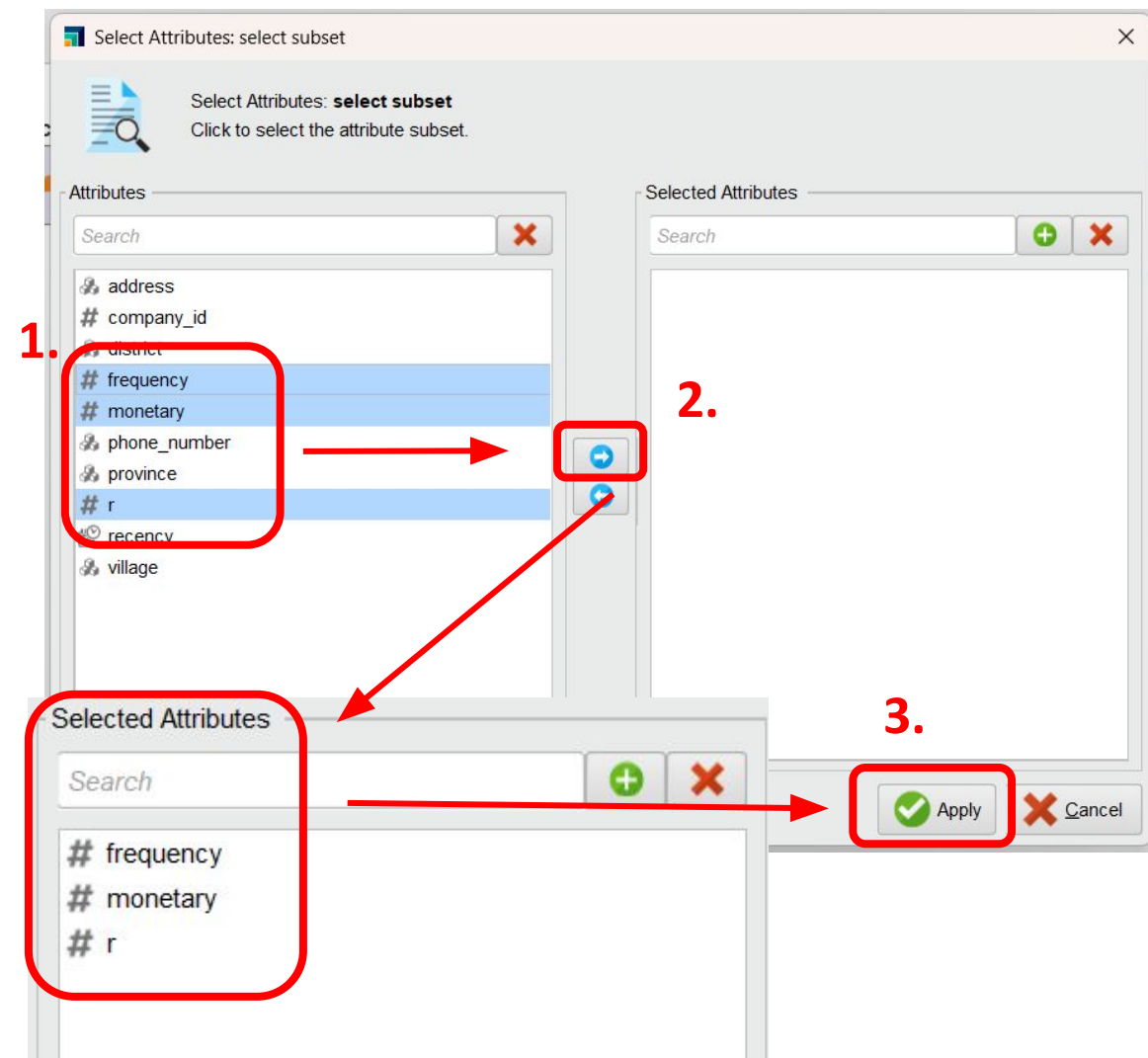
2. 在Selection目錄中，找到「Select Attributes」

3. 將Select Attributes拖曳至線上

4. 選取Select Attributes時，右側Parameters會出現相關設定，於attribute filter type中選擇「a subset」，並點開下方「Select Attributes...」，進入select subset



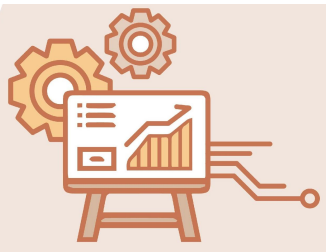
Step 4. 選擇分析欄位



1. 於Attributes中，**按住Ctrl點選**「frequency、monetary、r」三個欄位
2. 點選**右向鍵**，將三個欄位新增於Selected Attributes中
3. 點選「Apply」，完成欄位選擇

※ 學生筆記－為什麼是選擇 r，而不是recency？

欄位r是前面欄位計算所產生的欄位，目的在記錄現在時間與最近一次消費時間的間距，而recency僅記錄最近一次消費時間的日期，並不符需求



Step 5. 正規化

The screenshot illustrates the process of adding and configuring the 'Normalize' operator in Orange3. It includes the following elements:

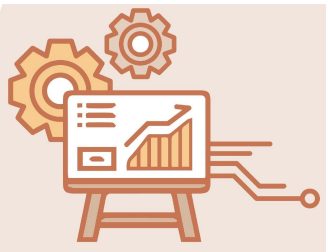
- Repository:** A list of data sources on the left, including 'Training Resources', 'Samples', 'Community Samples', 'Local Repository (Local)', 'Temporary Repository (Local)', and 'DB (Legacy)'.
- Operators:** A panel on the bottom left showing a search for 'Normalize' under the 'Normalization (2)' category. A red box highlights the 'Normalize' operator, with a red arrow pointing to the 'Process' area.
- Process:** A central workspace showing a workflow with operators: 'Read CSV', 'Filter Examples', 'Generate Attributes', and 'Select Attributes'. A red box highlights the 'Normalize' operator being added to the workflow.
- Parameters:** A panel on the bottom right showing the configuration for the 'Normalize' operator. It includes fields for 'attribute filter type' (set to 'all'), 'invert selection' (unchecked), 'include special attributes' (unchecked), 'method' (set to 'range transformation'), 'min' (0.0), and 'max' (1.0). A red box highlights the 'method' dropdown, with a red arrow pointing to it from the 'Show advanced parameters' link.
- Show advanced parameters:** A link at the bottom left, highlighted with a red box, that allows users to view more configuration options for the operator.

Red numbers 1, 2, 3, and 4 are placed next to the corresponding steps in the workflow.

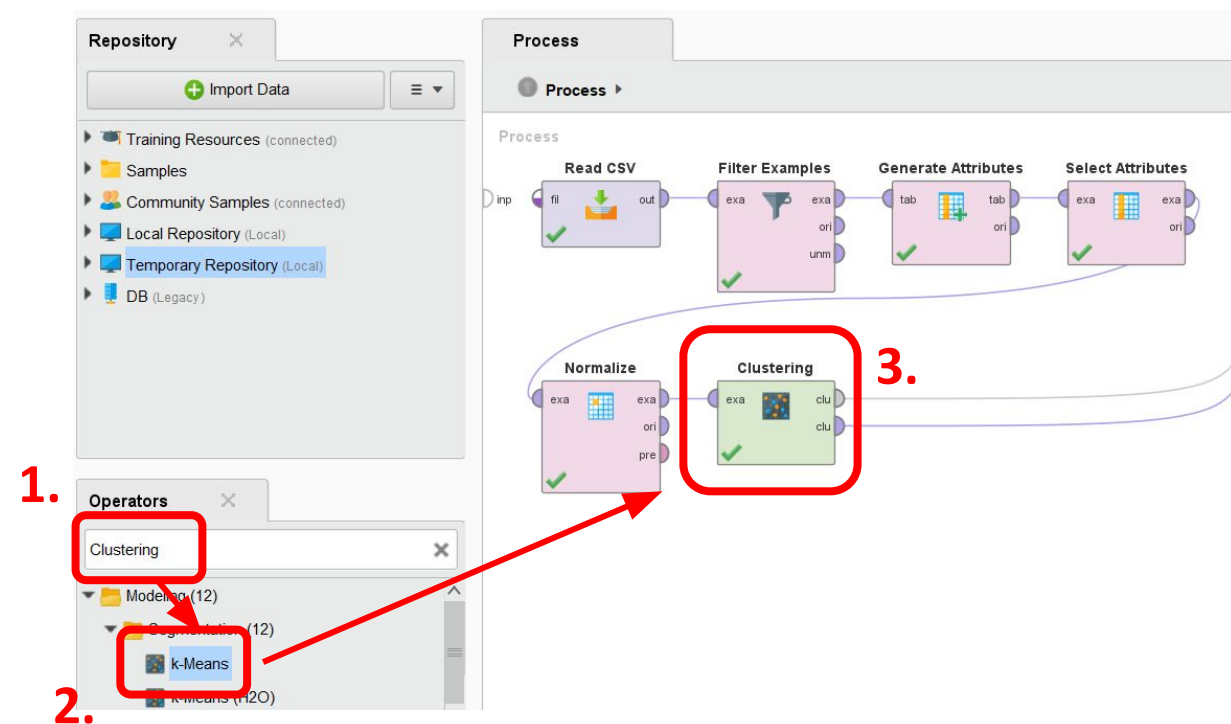
1. 於左側搜尋框中搜尋「Normalize」
2. 在Normalization目錄中，找到「Normalize」
3. 將Normalize拖曳至線上
4. Normalize**選取狀態**時，點選右側Parameters下方「Show advanced parameters」，並於method中選擇「range transformation」

※ 學生筆記－什麼是 range transformation？

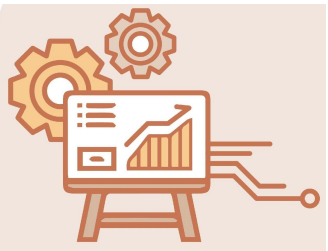
一種常見的資料正規化方法，用於將數值資料轉換到特定的範圍內，目的是將不同範圍的資料映射到統一的標準範圍內，以便進行比較和分析



Step 6. K-means 集群分析



1. 於左側搜尋框中搜尋「Clustering」
2. 在Segmentation目錄中，找到「k-Means」
3. 將k-Means拖曳至線上，記得將下方clu也一併連線至res，才能產生圖表



Step 6. K-means 集群分析

Parameters

Clustering (k-Means)

☒ add cluster attribute

☐ add as label

☐ remove unlabeled

k: 5

max runs: 10

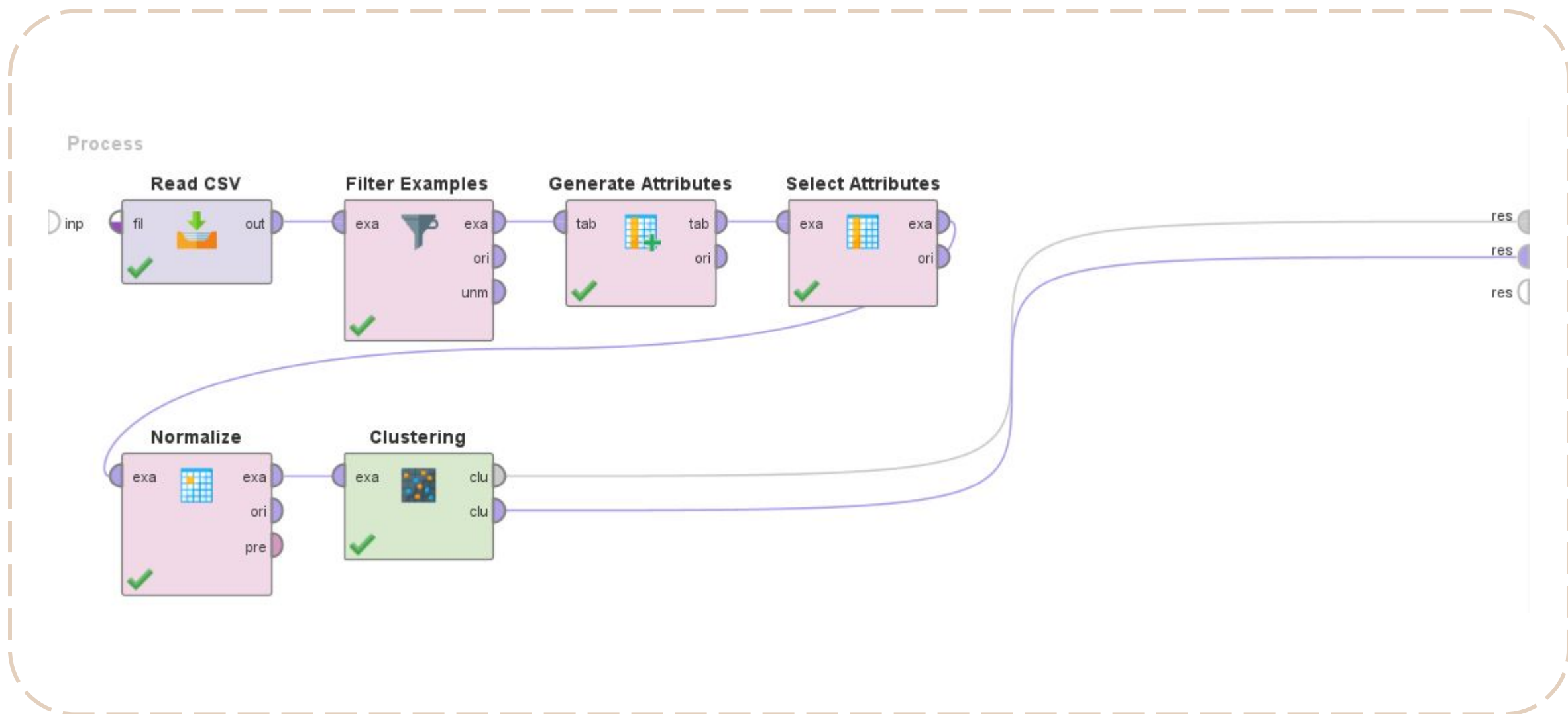
☒ determine good start values

《自行決定參數》



於Clustering選取時，右側Parameters能夠進行最後的參數設置，**參數不同，結果也會不同**

k—**分群數量**（想將資料分為幾個群）

max runs—**最大執行次數**（可能影響精確性，同時也會拉長執行時間）



▲ 分群分析流程圖。[練習專案下載](#)

Open in  Turbo Prep  Auto Model

Row No.	id	cluster	frequency	monetary	r
1	1	cluster_1	0.004	0.000	1
2	2	cluster_1	0	0.000	1
3	3	cluster_1	0	0.000	0.993
4	4	cluster_1	0	0.000	0.986
5	5	cluster_1	0	0.000	0.983
6	6	cluster_1	0.004	0.000	0.983
7	7	cluster_1	0	0.000	0.976
8	8	cluster_1	0	0.001	0.973
9	9	cluster_1	0	0.001	0.973
10	10	cluster_1	0	0.000	0.970
11	11	cluster_1	0	0.000	0.966
12	12	cluster_1	0	0.000	0.966
13	13	cluster_1	0.004	0.001	0.966
14	14	cluster_1	0	0.000	0.963
15	15	cluster_1	0	0.000	0.963
16	16	cluster_1	0	0.000	0.959
17	17	cluster_1	0	0.006	0.953
18	18	cluster_1	0	0.000	0.949

ExampleSet (433 examples, 2 special attributes, 3 regular attributes)

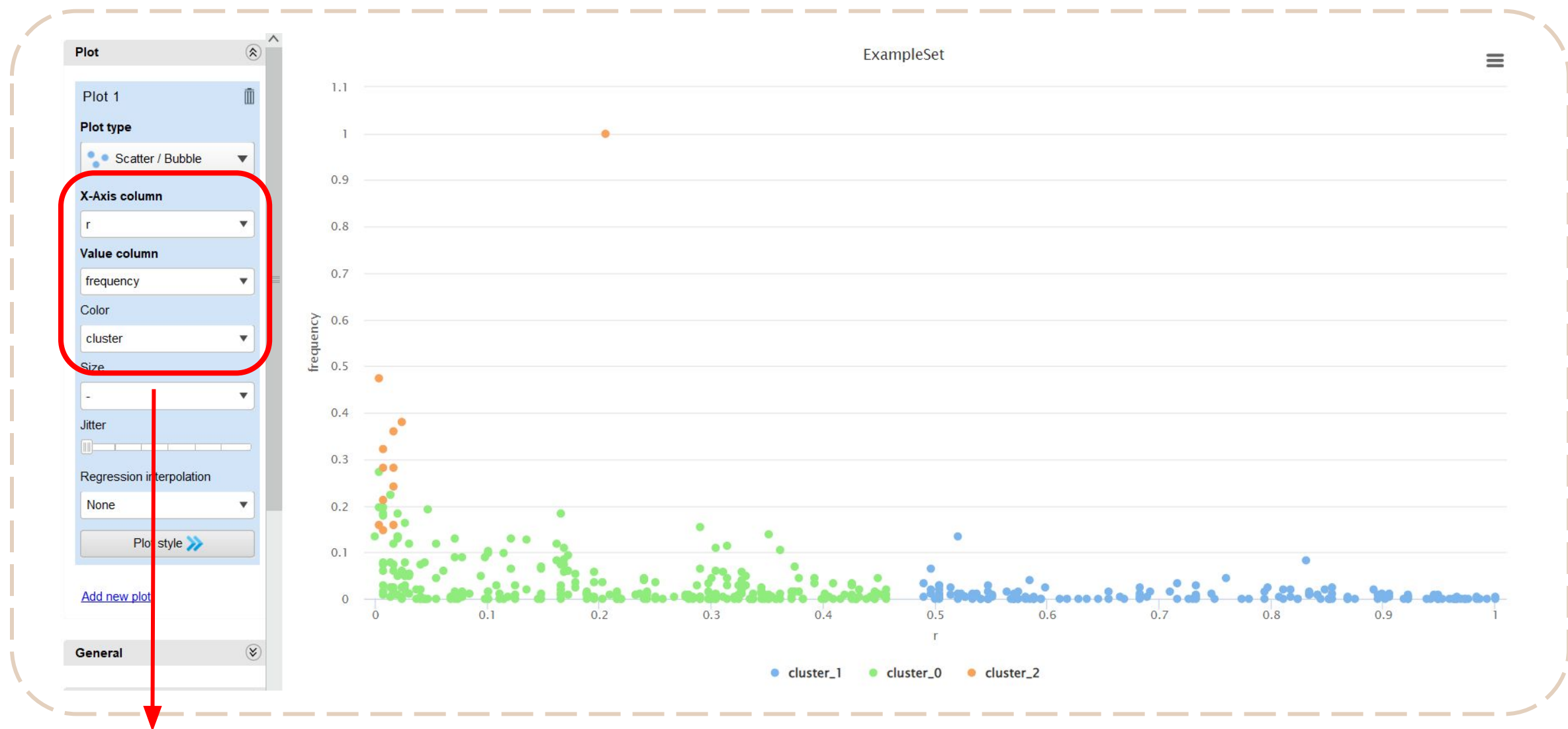
執行後，點選上方Results可以看到...

Data－每筆資料分群後的個別結果

Statistics－分群後的描述性統計

Visualizations－視覺化圖表

▲ 執行結果



可自行調整座標軸

▲ 執行結果視覺化

Q & A 時間

