**作業1: 資料進得來,分析出得去,信用金融大數據**

1. **請依”clean required”一欄對CRX.csv進行資料之匯入與清理及轉換(transfer to)。(43%)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attribute | Defined | Transfer to | Clean required | 檢查是否完成 |
| Gender  a/b | Female/male | 0/1 | Judge yourself (5%) | V 完成 |
| Age | age |  |  | Skip |
| Debt | Outstanding debt (feature has been scaled) |  |  | Skip |
| Married | Married, u=Single/Divorced/etc, y=Married | 0/1 (5%) |  | V 完成 |
| BankCustomer  g/p | BankCustomer  Bank customer, 0=does not have a bank account, 1=has a bank account | 0/1 | Judge yourself (5%) | V 完成 |
| Industry  c, d, cc, i, j, k, m, r, q, w, x, e, aa, f |  | 1. CommunicationServices 2. ConsumerDiscretionary 3. ConsumerStaples 4. Education 5. Energy 6. Financials 7. Healthcare 8. Industrials 9. InformationTechnology 10. Materials 11. Real Estate 12. Research 13. Transport   Utilities | Judge yourself (10%) | V 完成 |
| Ethnicity  v, h, bb, j, n, z, dd, ff, o | Ethnicity | |  | | --- | | Asian | | Black | | Latino | | Other | | White | | 1. 請移除缺漏值. 2. 請移除筆數過少的值.→改用Merge合併資料 3. Judge yourself (10%) | V 完成 |
| YearsEmployed | Years employed |  |  | Skip |
| Prior default  t, f. | Prior default, 0=no prior defaults, 1=prior default | 0/1 | Judge yourself (2%) | V 完成 |
| Employed  t, f. | Employed, 0=not employed, 1=employed | 0/1 | Judge yourself (2%) | V 完成 |
| Credit score | Credit score (this feature has been scaled) |  |  | Skip |
| Drivers license,  t, f. | 0=no license, 1=has license | 0/1 | Judge yourself (2%) | V 完成 |
| Citizenship  g/p/s | Citizenship,  either ByBirth, ByOtherMeans or Temporary | ByBirth, ByOtherMeans or Temporary | Judge yourself (5%) | V 完成 |
| ZipCode |  |  | Remove | V 完成 |
| Income | Income (this feature has been scaled) |  |  | Skip |
| Approved, +/- | 0=not approved, 1=approved | 0/1 | Judge yourself (2%) | V 完成 |

1. **請貼出您的process. (12%)**

【資料轉換】

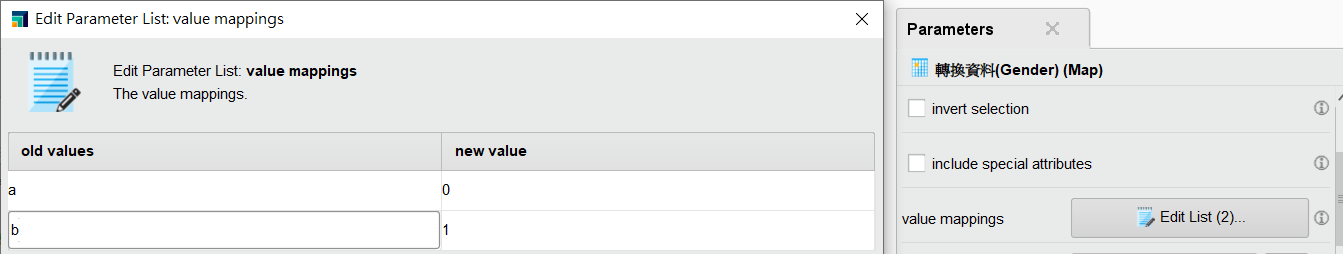
一張含有 文字 的圖片

自動產生的描述

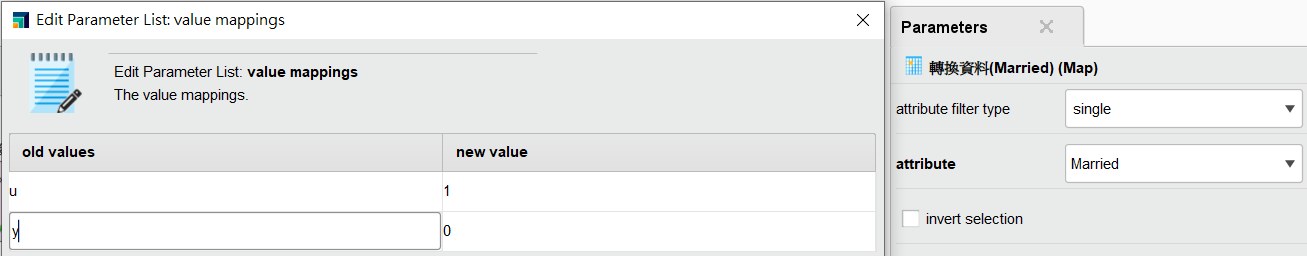
* 建議先replace，再刪除遺失值。

→這樣筆數比較能與原始資料比對確認

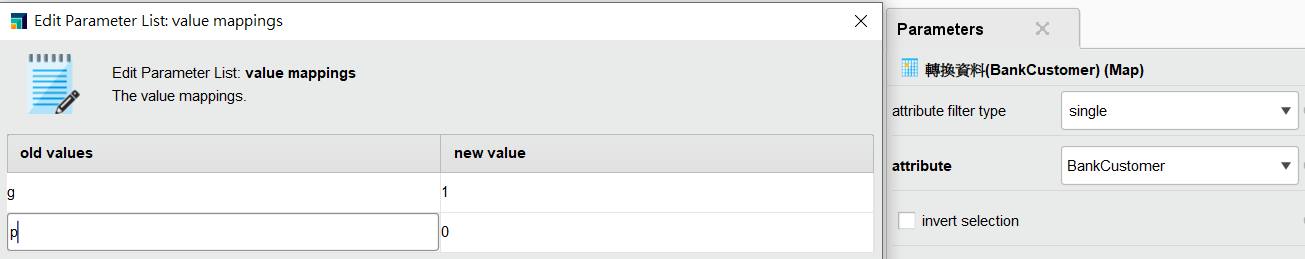
* Gender  
  a→0, b→1



* Married  
  u→1, y→0



* BankCustomer  
  g→1, p→0

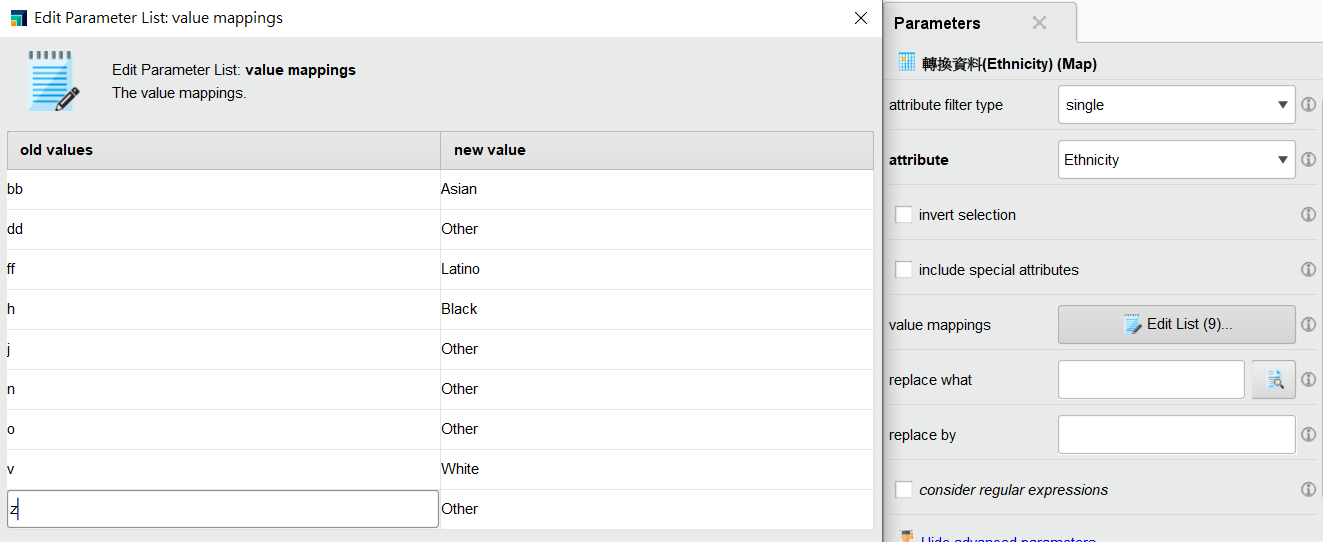


* Industry

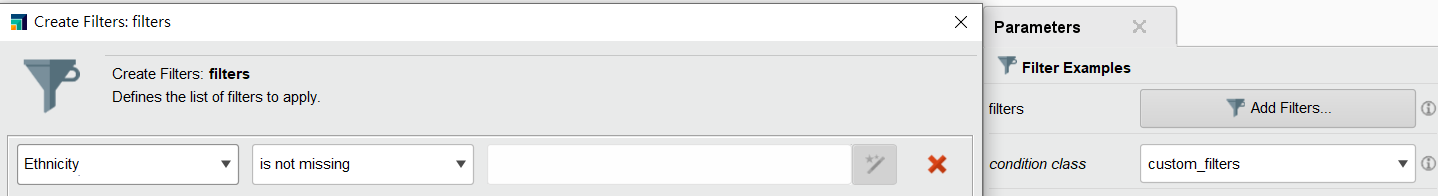
一張含有 文字, 螢幕擷取畫面, 室內 的圖片

自動產生的描述

* Ethnicity
  + Transfer轉換



* + 移除缺漏值



* + ~~移除筆數過少的值~~

|  |  |  |
| --- | --- | --- |
| Ethnicity | 次數 | 百分比 |
| White | 399 | 59% |
| Black | 138 | 20% |
| Asian | 59 | 9% |
| Latino | 57 | 8% |
| Other | 28 | 4% |

一張含有 圖表 的圖片

自動產生的描述

一張含有 圖表, 圓餅圖 的圖片

自動產生的描述

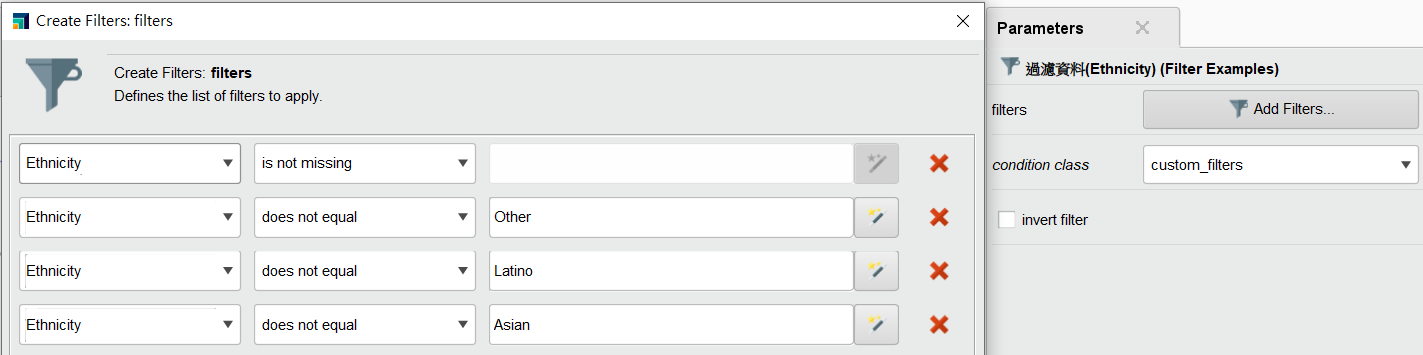
一張含有 文字 的圖片

自動產生的描述

（圓餅圖中，我在這邊輸入公式**"<b>{point.name}</b>: point.percentage:.1f}%"**來顯示百分比%）

* + - Other僅28筆資料，占4%
    - Latino僅57筆資料，占8%
    - Asian僅59筆資料，占9%

~~以上個人認為皆筆數過少，因此移除Other、Latino及Asian。~~



* + 爭議：什麼是筆數過少的欄位？要少於多少筆數才認為該資料筆數過少？

解題：與老師討論後，要藉著Domain Knowledge及Background來判斷資料是否過少以及是否應當刪除此筆資料，並且老師分享，根據經驗，通常若少於1%才算資料較少，因此在這題目我改用Merge，將Other、**Latino**及**Asian**合併為Other。

一張含有 文字 的圖片

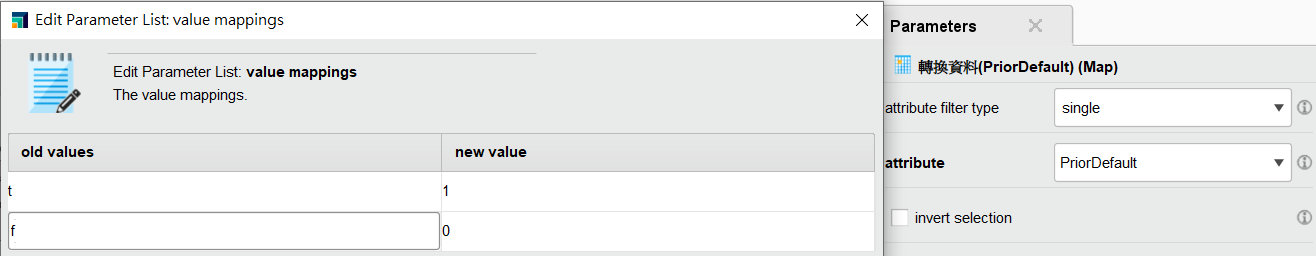
自動產生的描述

一張含有 資料表 的圖片

自動產生的描述

* PriorDefault

t→1, f→0



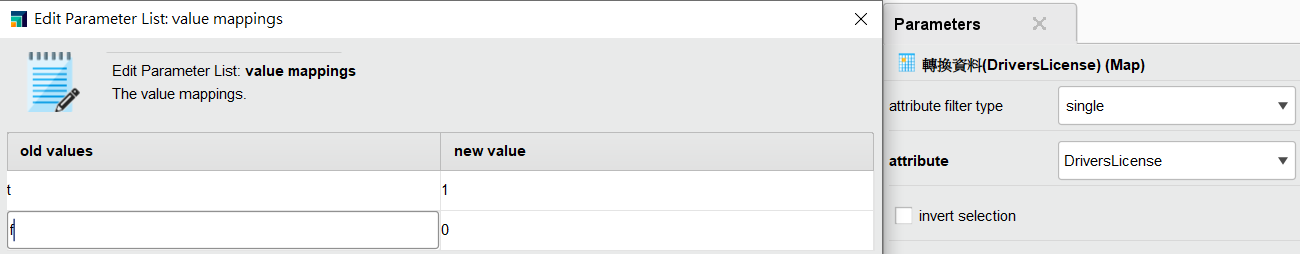
* Employed

t→1, f→0



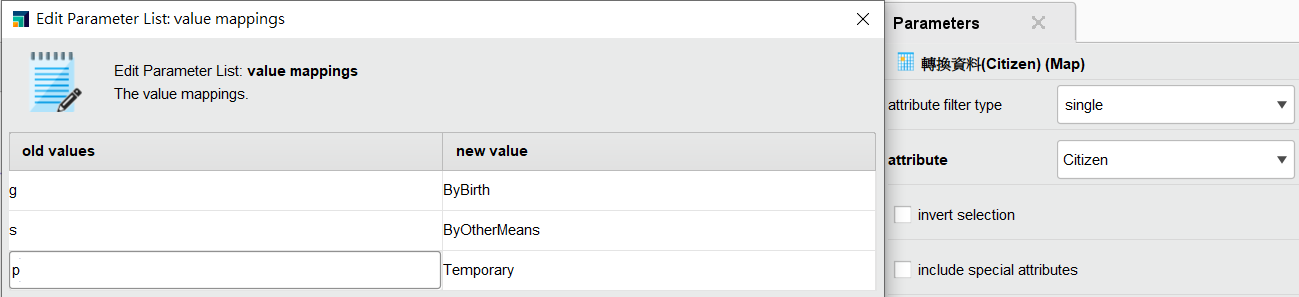
* DriversLicense

t→1, f→0

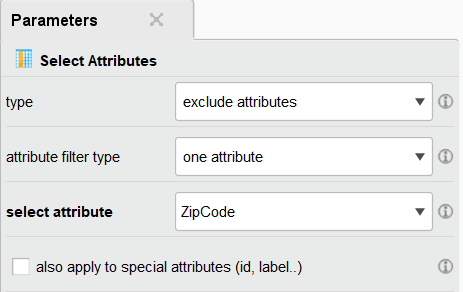


* Citizen

g→ByBirth, s→ByOtherMeans, p→Temporary

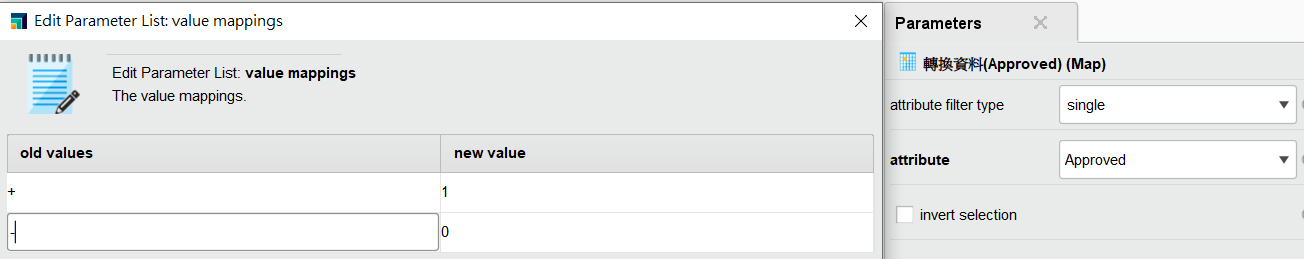


* Remove ZipCode

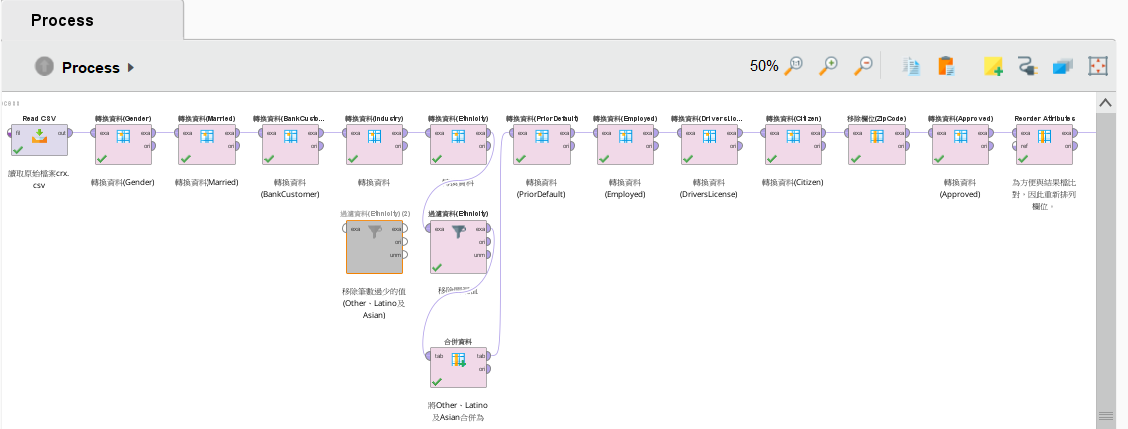


* Approved

+→1, -→0



* Process步驟流程



**2. 太棒了,現在我們終於可以開始進行一些敍述統計了,**

**2-1 (10%)請問您最後留下幾筆資料? 還有缺漏值嗎? 那些欄位?**

**2-3 (10%) 您建議怎麼處理這些缺漏值,請說明並進行必要之處置**

Ans：

* 最後留下681筆資料（移除9筆Ethnicity為缺漏值的資料），仍有缺漏值，分別為：Gender、Age、Married、BankCustomer

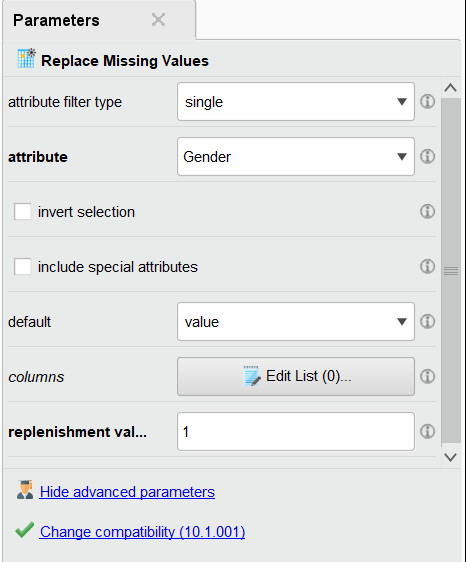
其中有些欄位疑似人為輸入錯誤，包含Married（l）、BankCustomer（gg）



* 處理缺漏值
  + Gender→10筆缺漏值

|  |  |  |
| --- | --- | --- |
| 各組性別人數統計 | | |
| Ethnicity | 0.女性 | 1.男性 |
| White | 109 | **283** |
| Black | 47 | **90** |
| Other | 52 | **90** |

由上表得知男性佔多數，性別為類別資料，補最常出現的值（1.男性）。



* + Age→12筆缺漏值  
    承上題表格格式，我也會將資料依據Ethnicity種族及Gender性別分組，計算各組的中位數，條件化補中位數。

|  |  |  |
| --- | --- | --- |
| 各組性別年齡中位數 | | |
| Ethnicity | 0.女性 | 1.男性 |
| White | 24.625 | 27.58 |
| Black | 25.17 | 30.67 |
| Other | 31.835 | 34.625 |

（因為尚未找到用RapidMiner計算中位數的最佳方式，所以上圖我是用Excel的Median函數來算出中位數。）

之所以補上中位數，而非平均數，是因為平均數會受到最高值或最小值影響；再者我們無法確定是均勻的資料分布。

* + Married→2筆錯誤資料  
    該資料欄位預設僅u及y，但卻出現了I，推測是人為輸入錯誤，且無法回推正確值，因此我會刪除這兩筆資料。  
    一張含有 資料表 的圖片

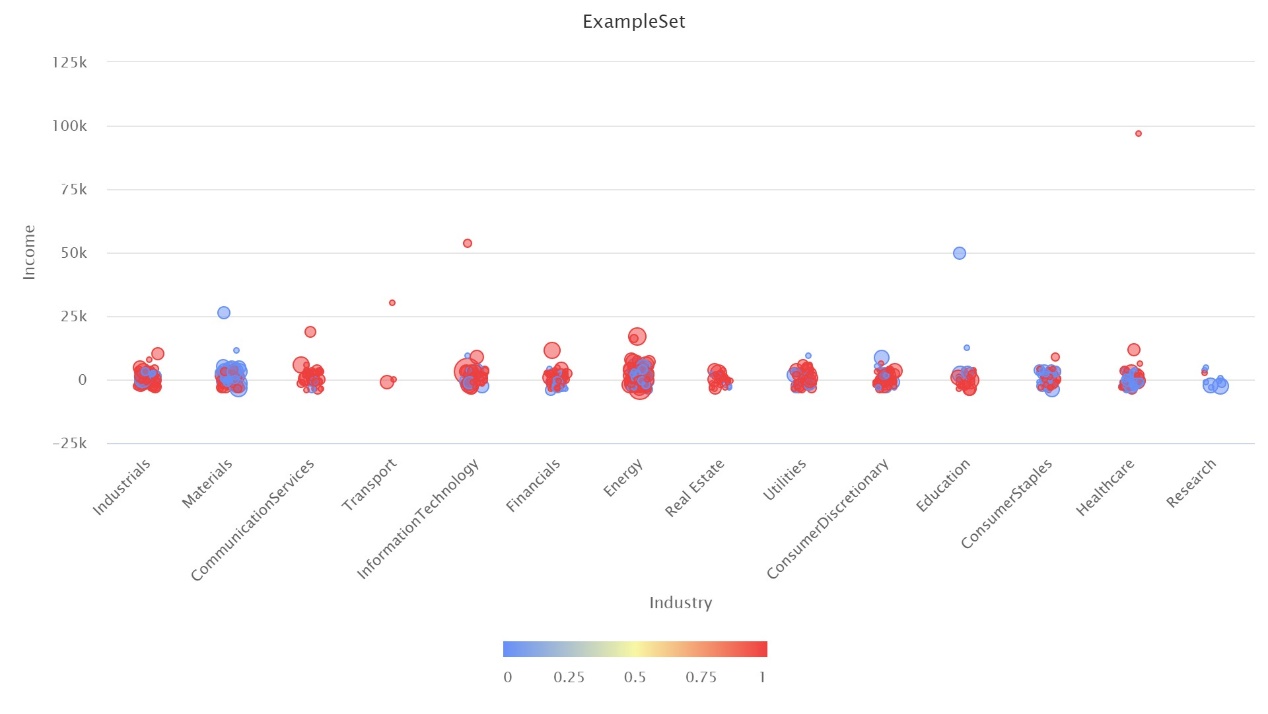
    自動產生的描述
  + BankCustomer→2筆錯誤資料  
    再刪除Married錯誤的資料後，BankCustomer也跟著一起處理好了。
* Process步驟流程

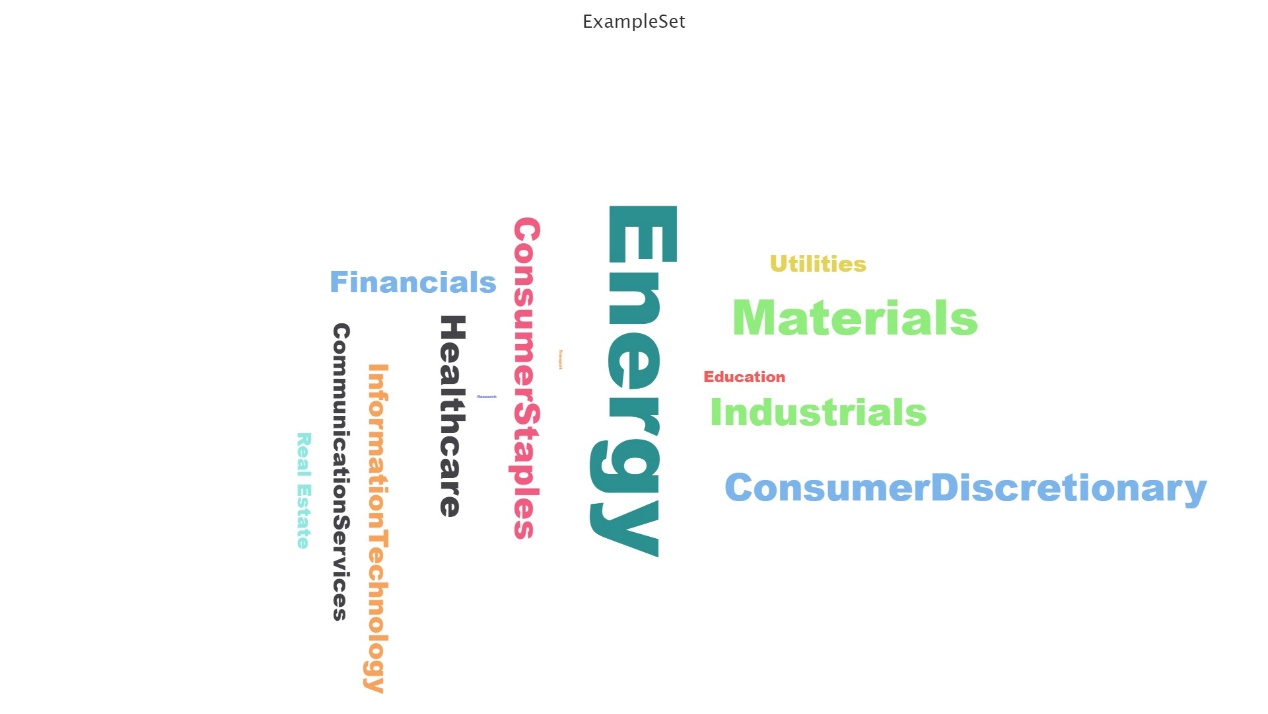
一張含有 圖表 的圖片

自動產生的描述

**3. 有圖有真相,請畫出2張圖,並加以說明。(25%)**

範例：





* 第一個議題：男女在各產業的分佈。
  + 1.男性→總共472筆

一張含有 圖表 的圖片

自動產生的描述

一張含有 圖表, 圓餅圖 的圖片

自動產生的描述

當中以Energy佔比22.2%最多，共有105位男性從事Energy能源產業相關工作。

* + 1.女性→總共207筆

一張含有 圖表 的圖片

自動產生的描述

一張含有 圖表, 圓餅圖 的圖片

自動產生的描述

當中以Materials佔比22.2%，共有46位女性從事Materials材料產業相關工作。

* + Step1：合併長條圖  
    為了方便檢視男性與女性在各產業的人數，將長條圖合併為一張圖表。

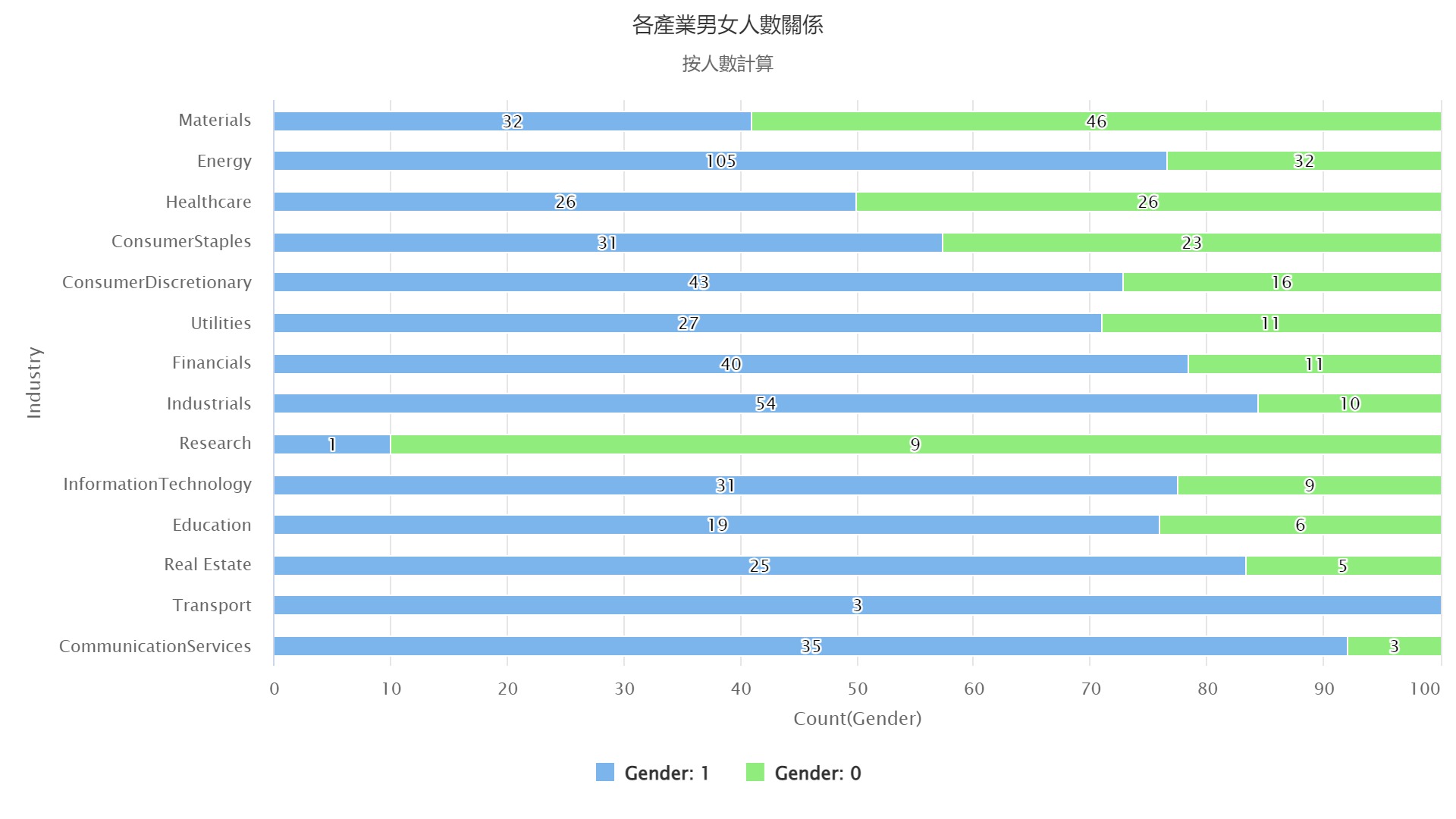
一張含有 圖表 的圖片

自動產生的描述

（圖中藍色代表男性；綠色代表女性）

* + Step2：Stack to 100%

更能看出各產業的男女人數大小（男多於女或男少於女）



（圖中藍色代表男性；綠色代表女性）

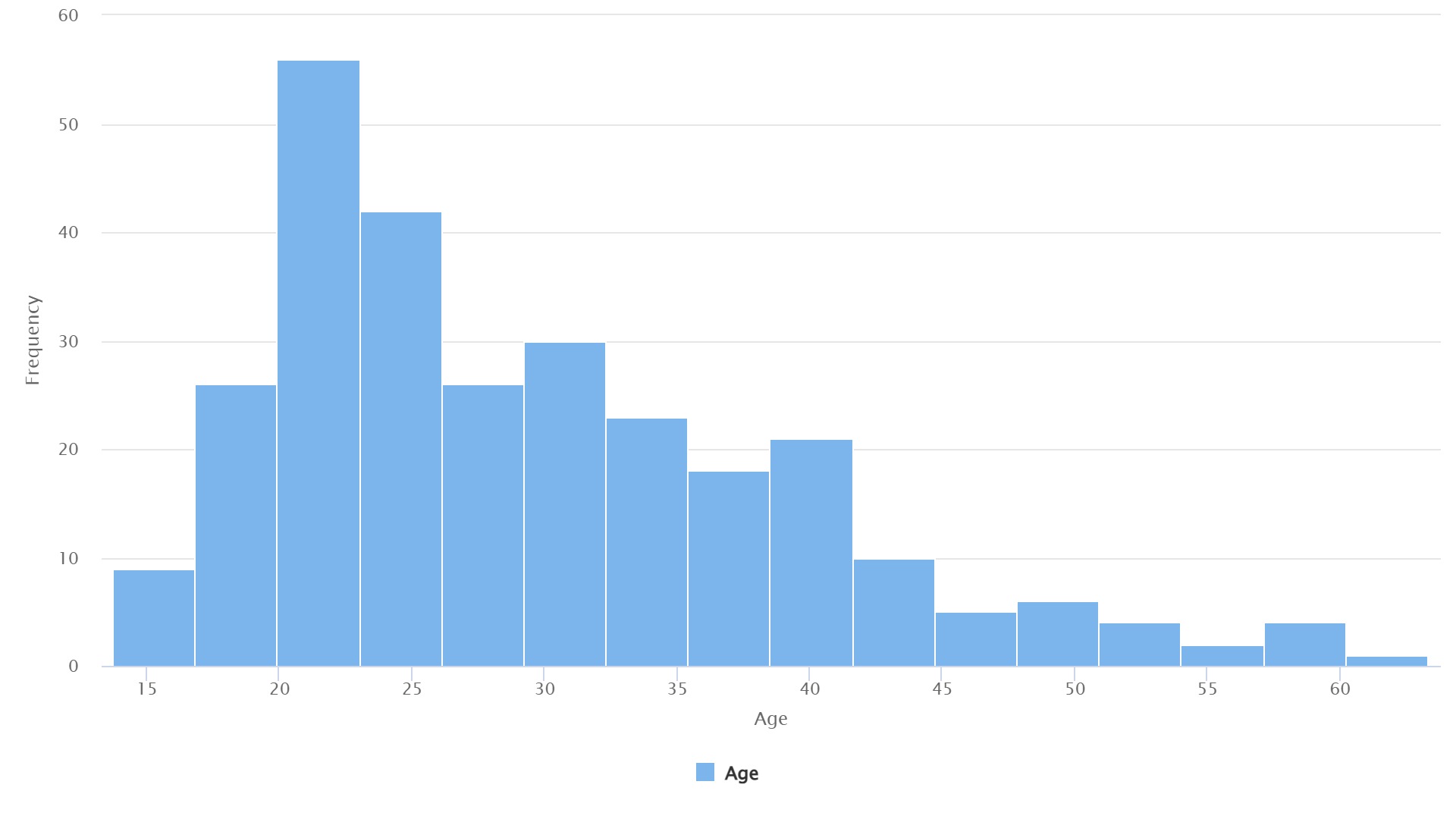
* + - Energy能源產業當然是男性人數完全輾壓女性人數
    - 然而Research卻呈現女性多於男性，這點讓我有點意外，因為我所處在的研究室環境或研發產業，可說是男性占居絕對多數。不過仍得再細看這份數據是如何調查而得到的為主。
    - Transport是唯一沒有任何女性的產業。
* 第二個議題：如下表格，依據Ethnicity種族及Gender性別分組，觀察在填補缺漏值之前，各年齡層的關係。

（個人作法：會先將年齡的缺漏值移除後，再做計算。）

|  |  |  |
| --- | --- | --- |
| 各組性別年齡中位數 | | |
| Ethnicity | 0.女性 | 1.男性 |
| White | 24.625 | 27.58 |
| Other | 31.835 | 34.625 |

|  |  |  |
| --- | --- | --- |
| 各組性別年齡平均數 | | |
| Ethnicity | 0.女性 | 1.男性 |
| White | 28.44509259 | 29.52572438 |
| Other | 36.7466 | 37.36933333 |

* WhiteMale白人男性→總共283人



以19.94- 23.04這年齡層人數最多，約有56人，大約佔兩成。

* + 中位數：27.58
  + 平均數：29.52572438

→若以平均數來填補缺漏值，會有些被拉高。（平均數>中位數）

* OtherMale其他人種的男性→總共90人

一張含有 圖表 的圖片

自動產生的描述

以30.66- 37.25這年齡層人數最多，約有21人，大約佔兩成三。

* + 中位數：34.625
  + 平均數：37.36933333

→若以平均數來填補缺漏值，會有些被拉高。（平均數>中位數）

* WhiteFemale白人女性→總共108人

一張含有 圖表 的圖片

自動產生的描述

以20.267- 24.784這年齡層人數最多，約有31人，大約佔兩成九。

* + 中位數：24.625
  + 平均數：28.44509259

→若以平均數來填補缺漏值，會有些被拉高。（平均數>中位數）

* OtherFemale其他人種的女性→總共50人

一張含有 圖表 的圖片

自動產生的描述

以16.08- 31.34這年齡層人數最多，約有24人，佔四成八。

* + 中位數：24.625
  + 平均數：36.7466

→若以平均數來填補缺漏值，會有些被拉高。（平均數>中位數）

* 若以整體來看，先移除缺漏值的年齡後，會有669筆資料。

一張含有 圖表 的圖片

自動產生的描述

以21.31-23.83這年齡層人數最多，約有94人，約佔一成四。

* + 中位數：28.42
  + 平均數：31.45188341

→仍是平均數>中位數，所以我會用中位數來填補缺漏值，以免整體年齡被拉高。

* Age年齡的分組是以RapidMiner當中的Turbo Prep來自動劃分。

Reference:

* 如何重新命名欄名稱  
  Select Attributes Operator in RapidMiner - Data Mining  
  <https://www.youtube.com/watch?v=tQ7oDnQXhmQ>
* Replace Missing Values

<https://academy.rapidminer.com/learn/video/replace-missing-values>