

Credit Card Fraud Detection

Second Group Presentation

2024/05/07

小組組員

職位分工表

PM

林貫原

專案進度管理
決策管理
文件管理

PM

許政揚

協助進度管理
網頁雛形
小組報告

TS

易祐辰

資料視覺化
資料搜集

DS

楊廷紳

程式管理
資料分析

DS

周昱宏

資料清洗
資料分析

SD

留筠雅

系統架構
介面架構
使用說明書撰寫

Outline

01

刪除變數

02

調整變數

03

離群值處理

04

遺失值插補

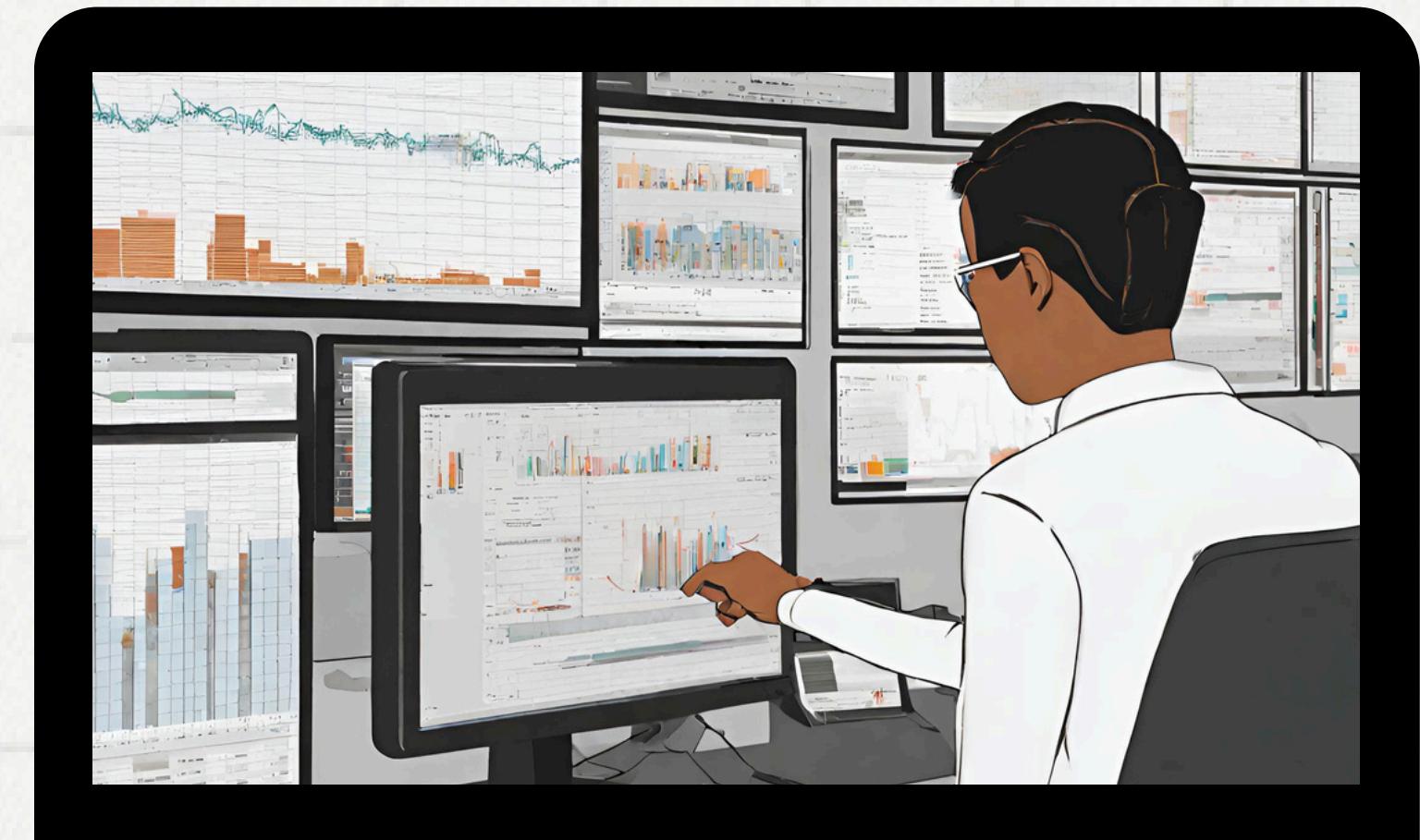
05

不平衡處理

06

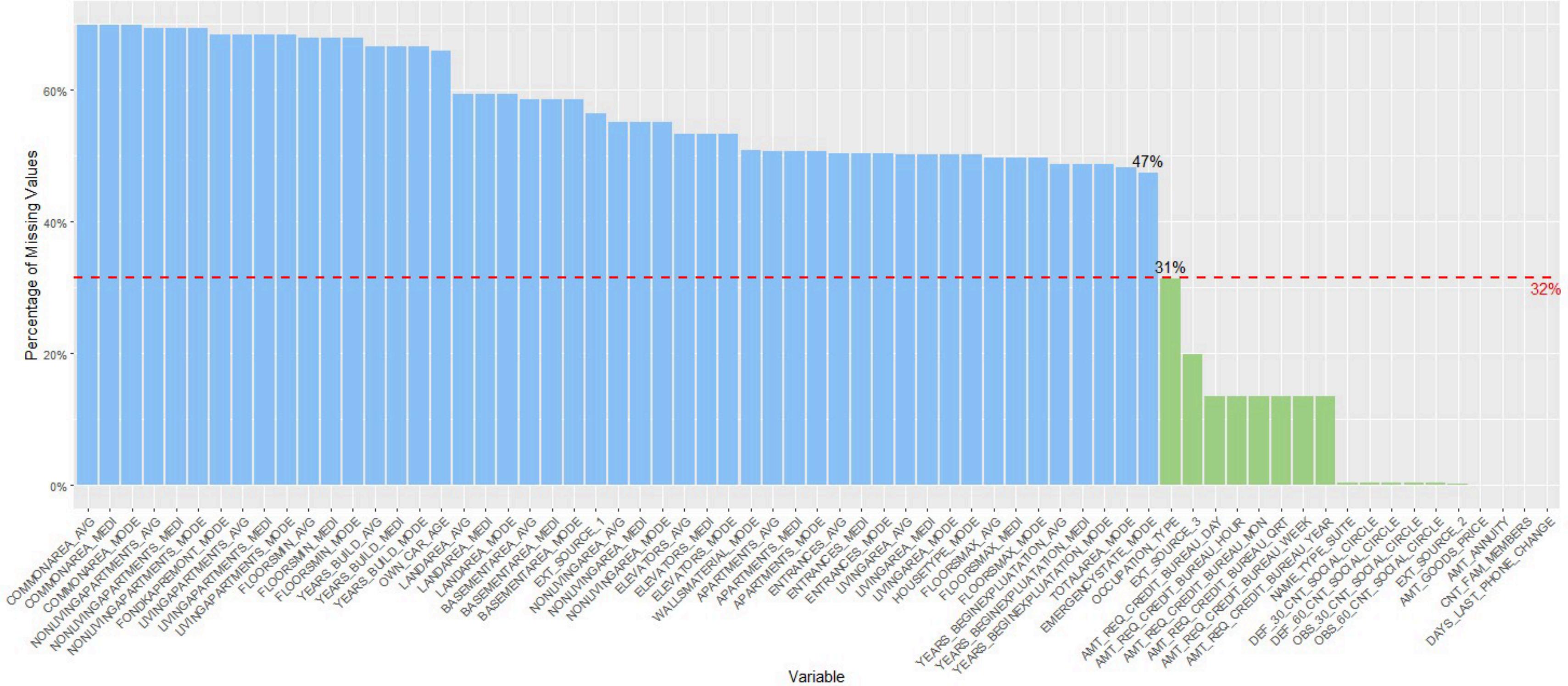
特徵選取

01. 刪除變數



去除遺失值過多變數

Percentage of Missing Values in Each Variable



02. 調整變數



新增變數

變數名稱	說明
SUM_FLAG_DOCUMENT	客戶總共簽署的文件數量

新增變數

變數名稱	說明
SUM_FLAG_DOCUMENT	客戶總共簽署的文件數量
missing_ratio	刪除了超過 32% 遺失值變數後 該資料中的遺失值個數佔全部變數的比例

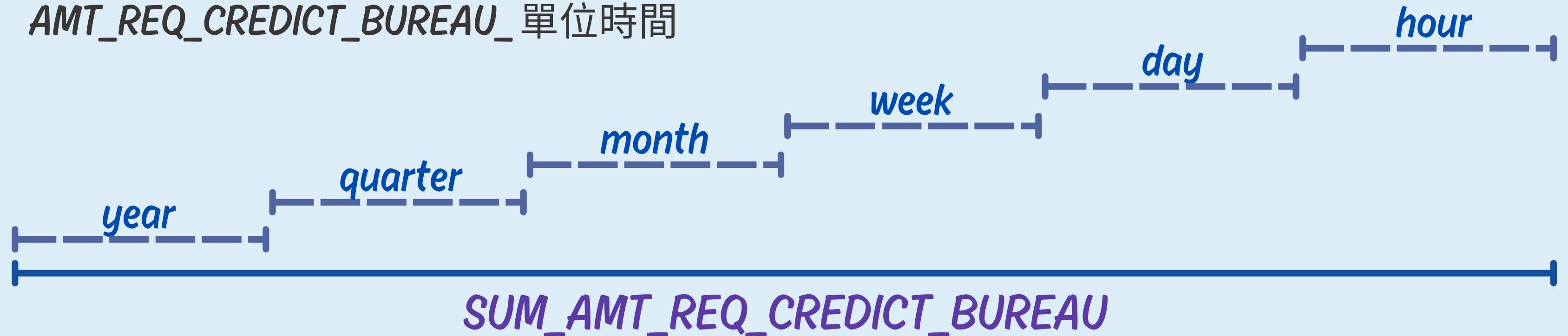
新增變數

變數名稱	SUM_AMT_REQ_CREDIT_BUREAU
說明	在申請信用卡之前一年內向信用局查詢客戶信用報告的總次數

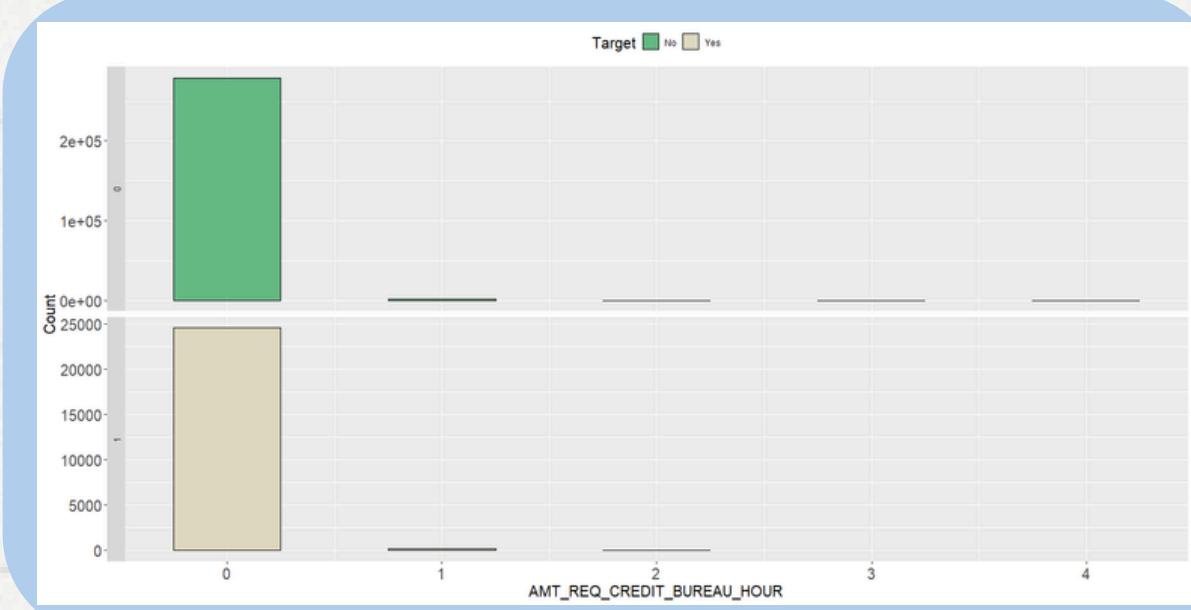
新增變數

變數名稱	SUM_AMT_REQ_CREDIT_BUREAU
說明	在申請信用卡之前一年內向信用局查詢客戶信用報告的總次數

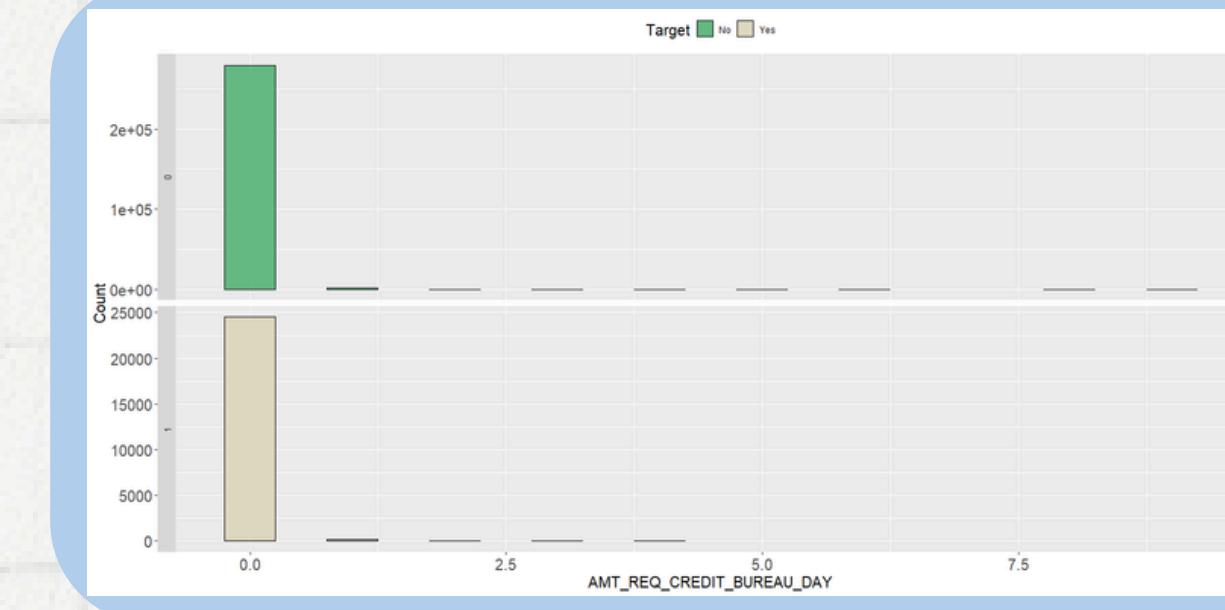
AMT_REQ_CREDIT_BUREAU 單位時間



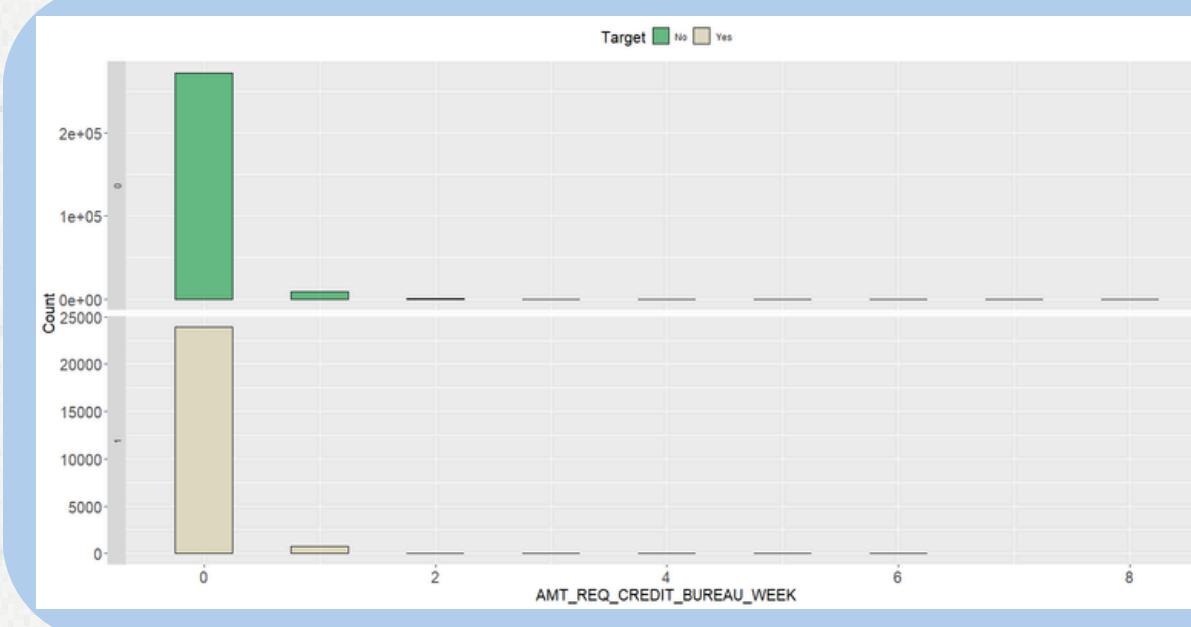
02調整變數



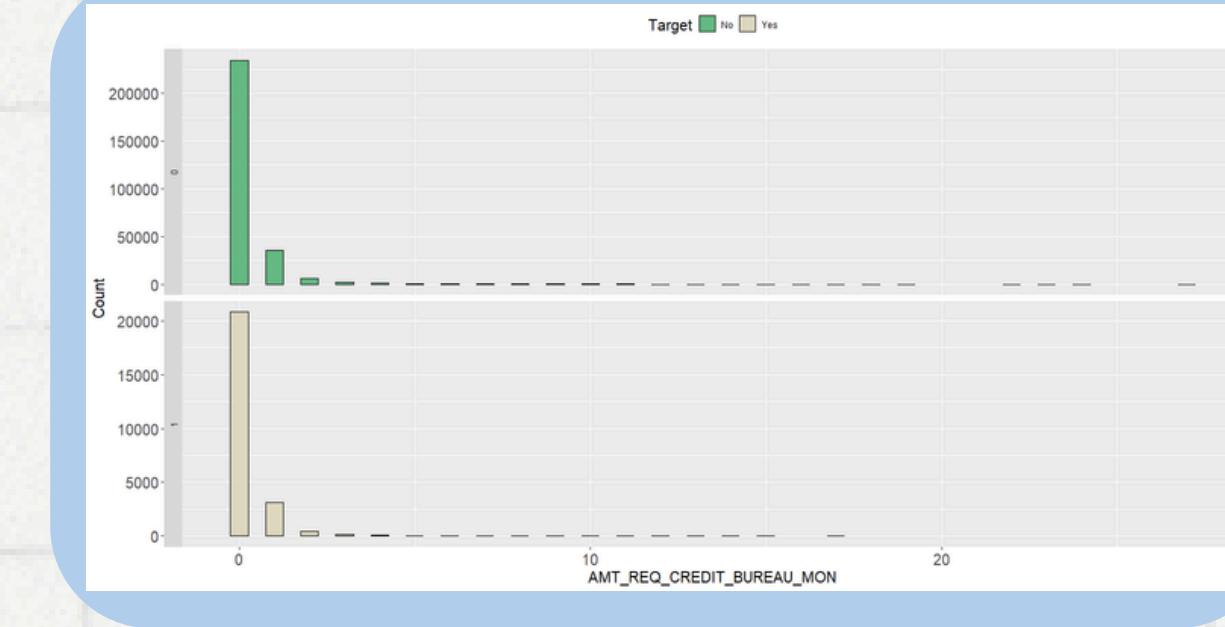
HOUR



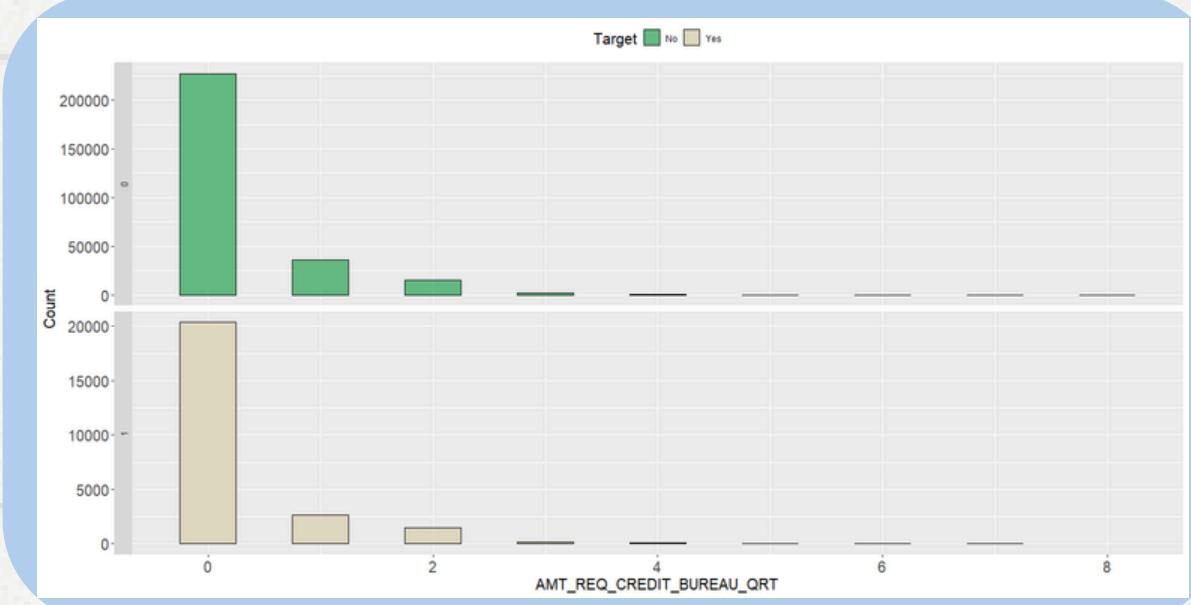
DAY



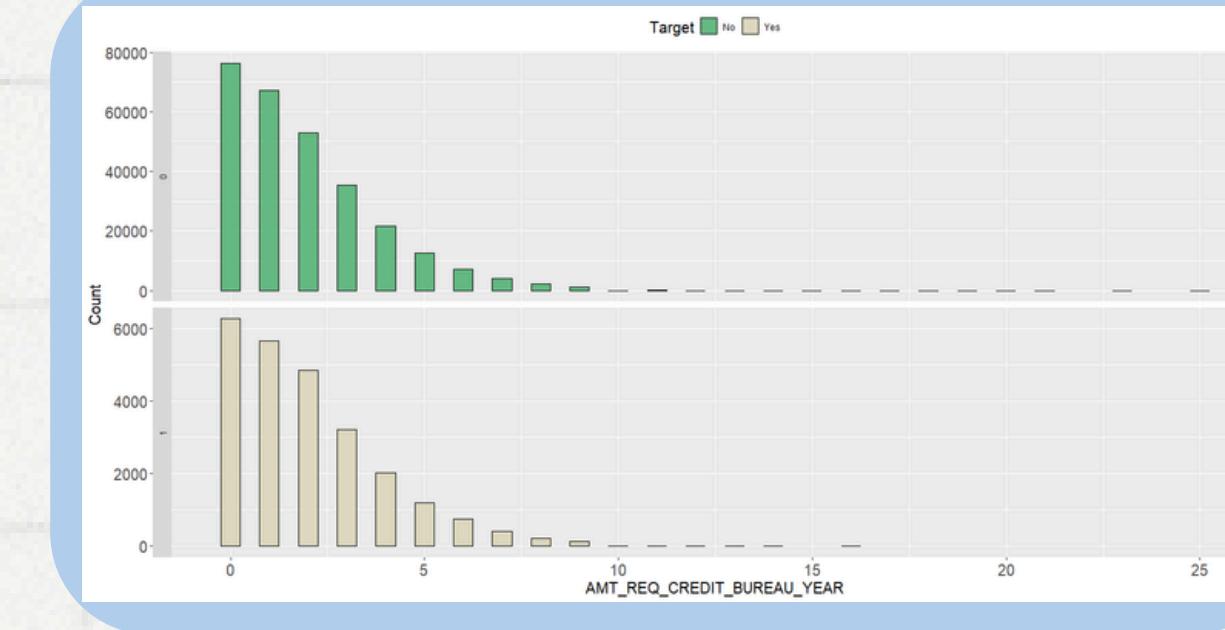
WEEK



MONTH



QUARTER



YEAR

資料修改 -- 數值型

變數名稱	說明	變動內容
DAY_BIRTH		單位皆為天數且值皆為負數
DAY_EMPLOYED		將資料除以 365.25
DAY_REGISTRATION	變數中的所有資料	讓單位由天改為年，並取絕對值
DAY_ID_PUBLISH		變數名稱： DAY 改為 YEAR
DAY_LAST_PHONE_CHANGE		

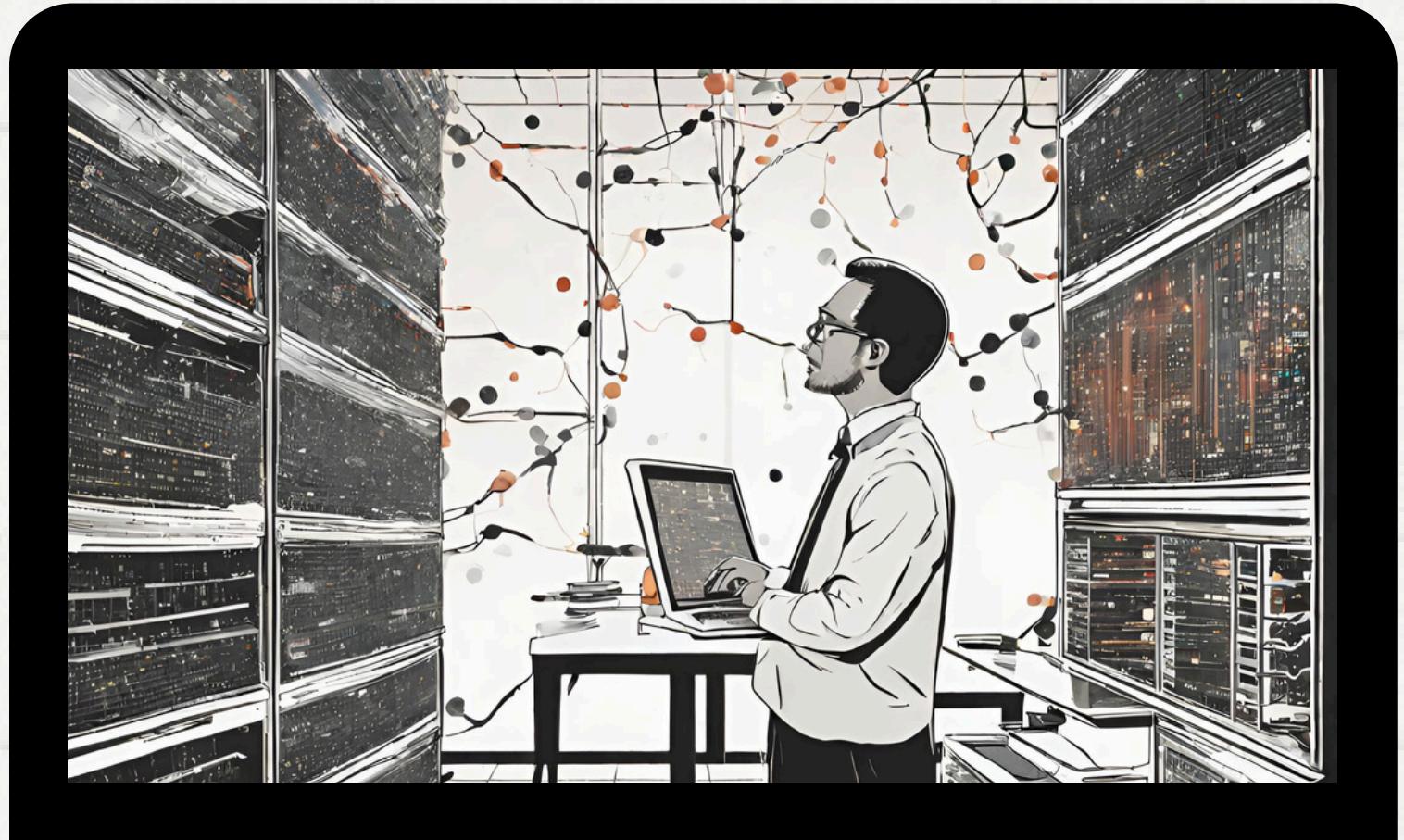
資料修改 -- 類別型

變數名稱	說明	變動內容
GENDER	變數中的 XNA	不進行插補 將其歸類為一個新的類別 others

資料修改 --類別型

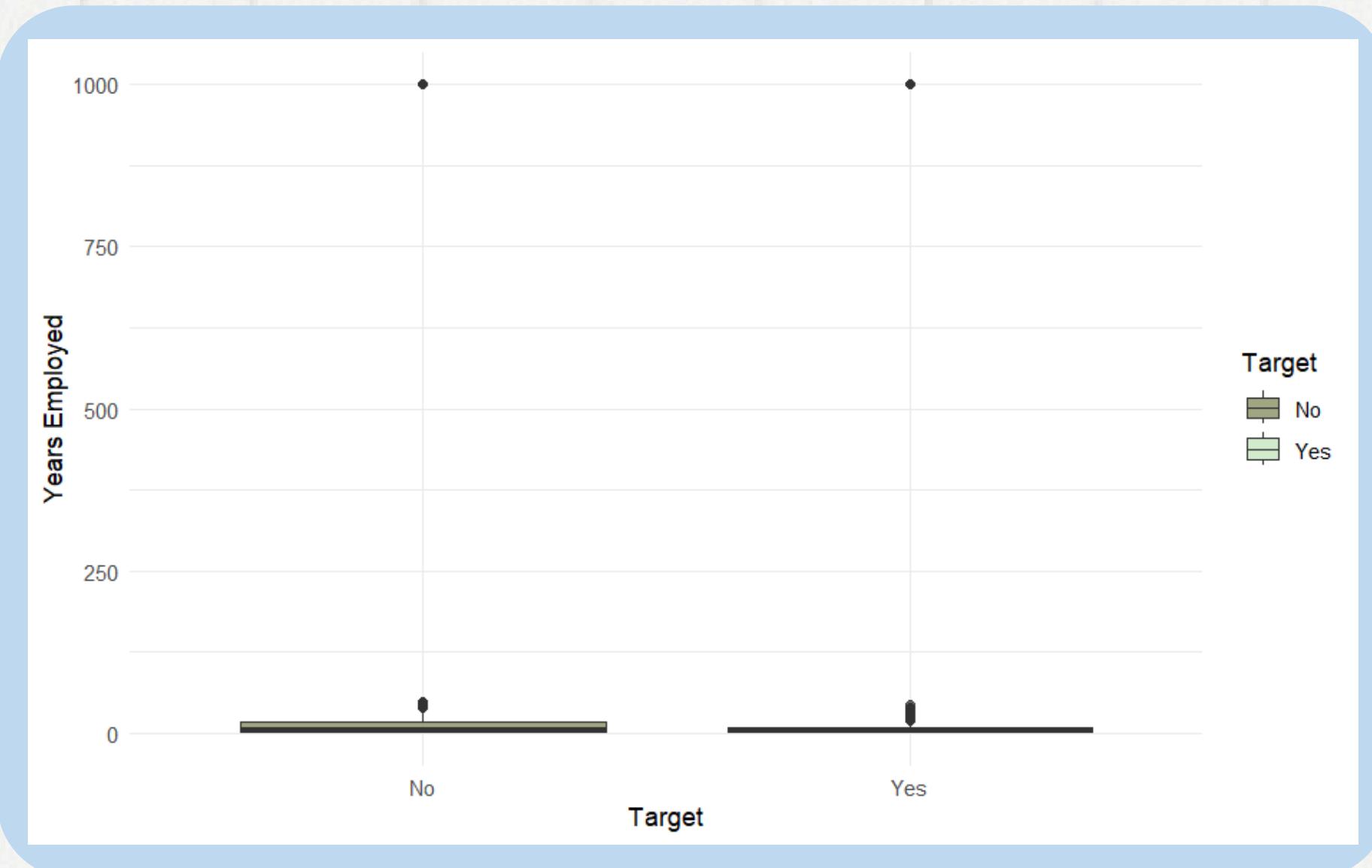
變數名稱	說明	變動內容
GENDER	變數中的 XNA	不進行插補 將其歸類為一個新的類別 others
ORGANIZATION_TYPE		歸類為新類別 Pensioner ，表示其為退休的人

03. 離群值處理



申請信用卡前多少天開始目前的工作 **DAYs_EMPLOYED**

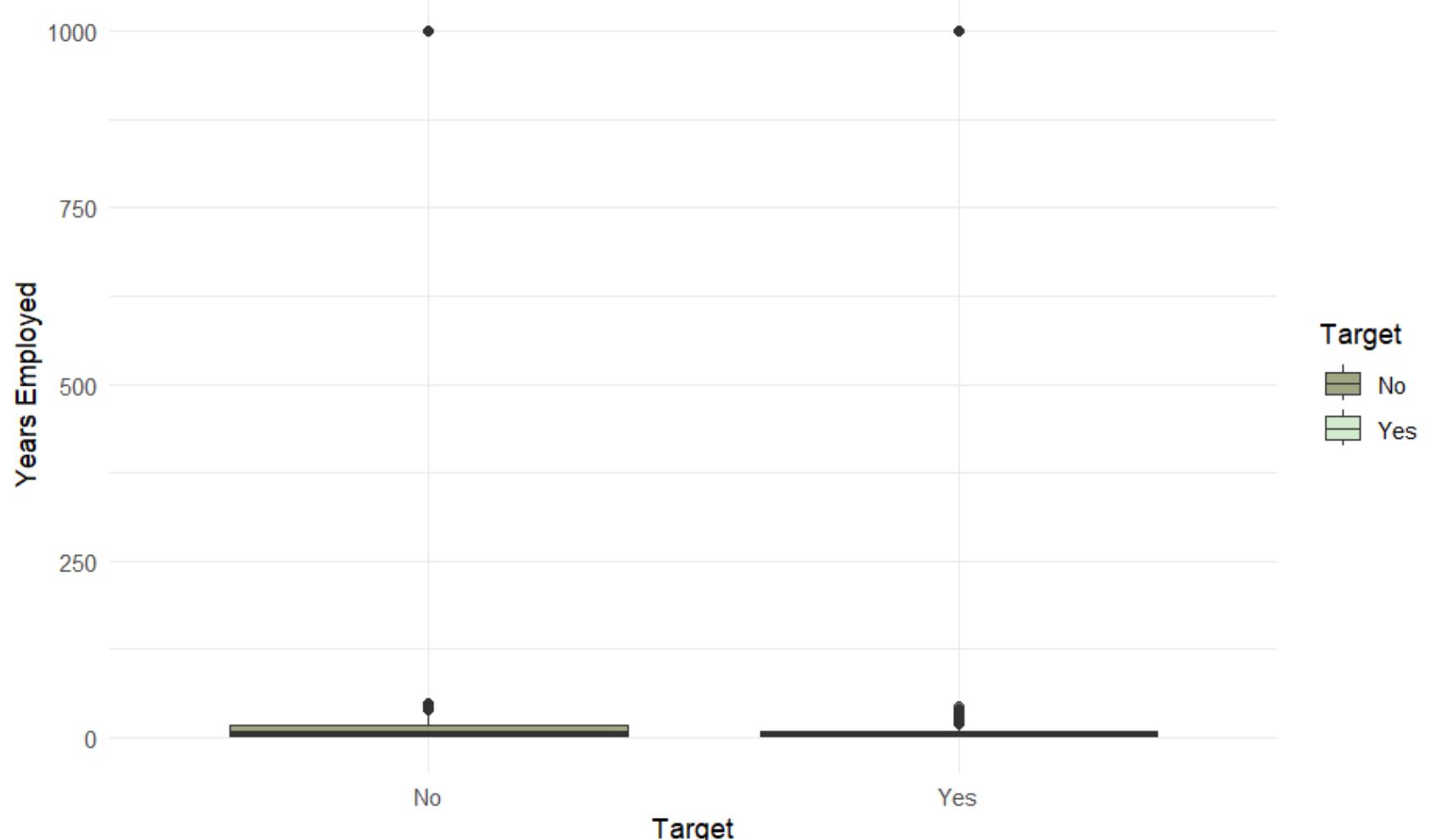
有很多資料顯示天數為-365243，經過轉換過後大約是 1000 年



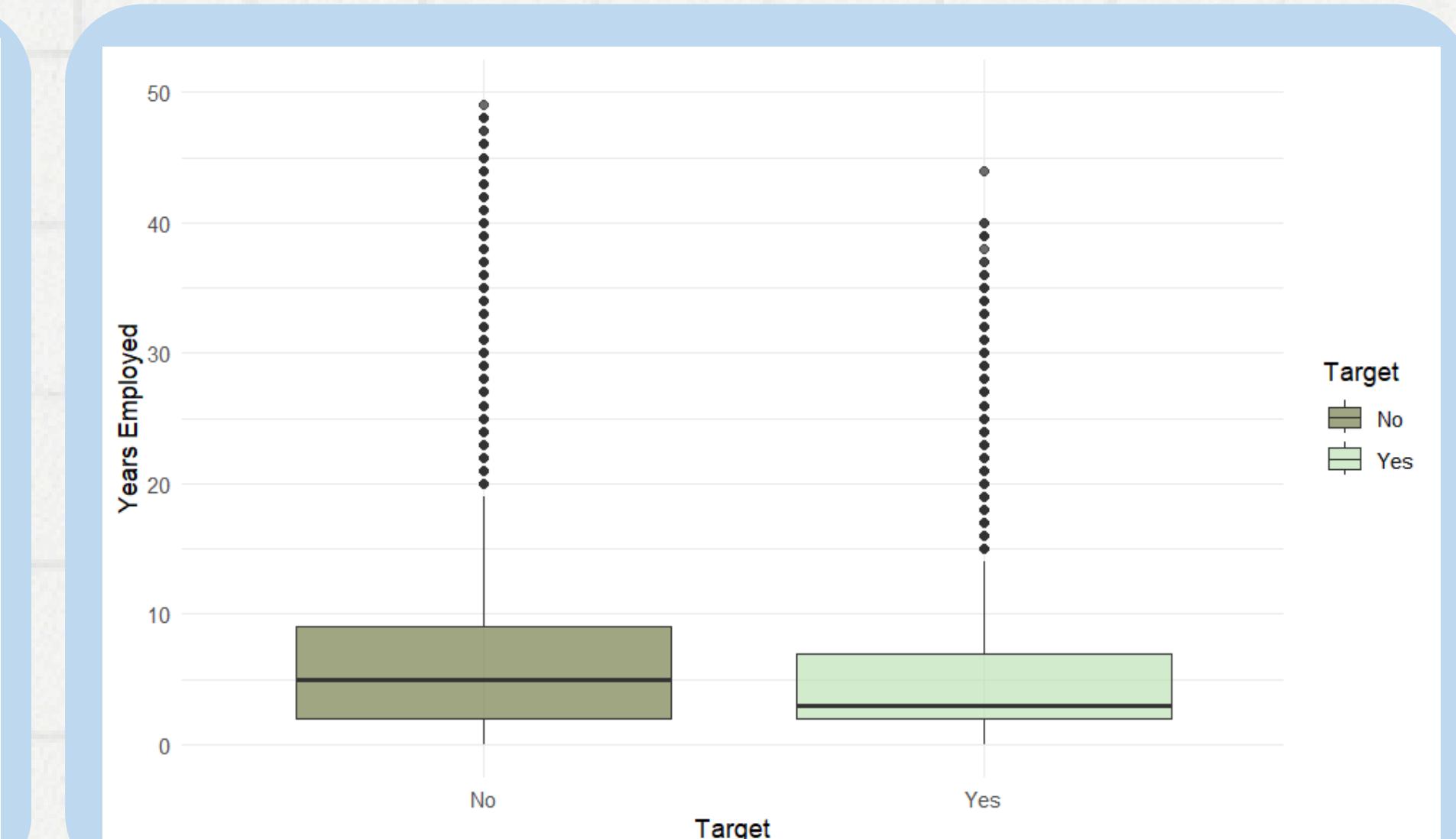
保留1000年資料

申請信用卡前多少天開始目前的工作 **DAYS_EMPLOYED**

有很多資料顯示天數為-365243，經過轉換過後大約是 1000 年

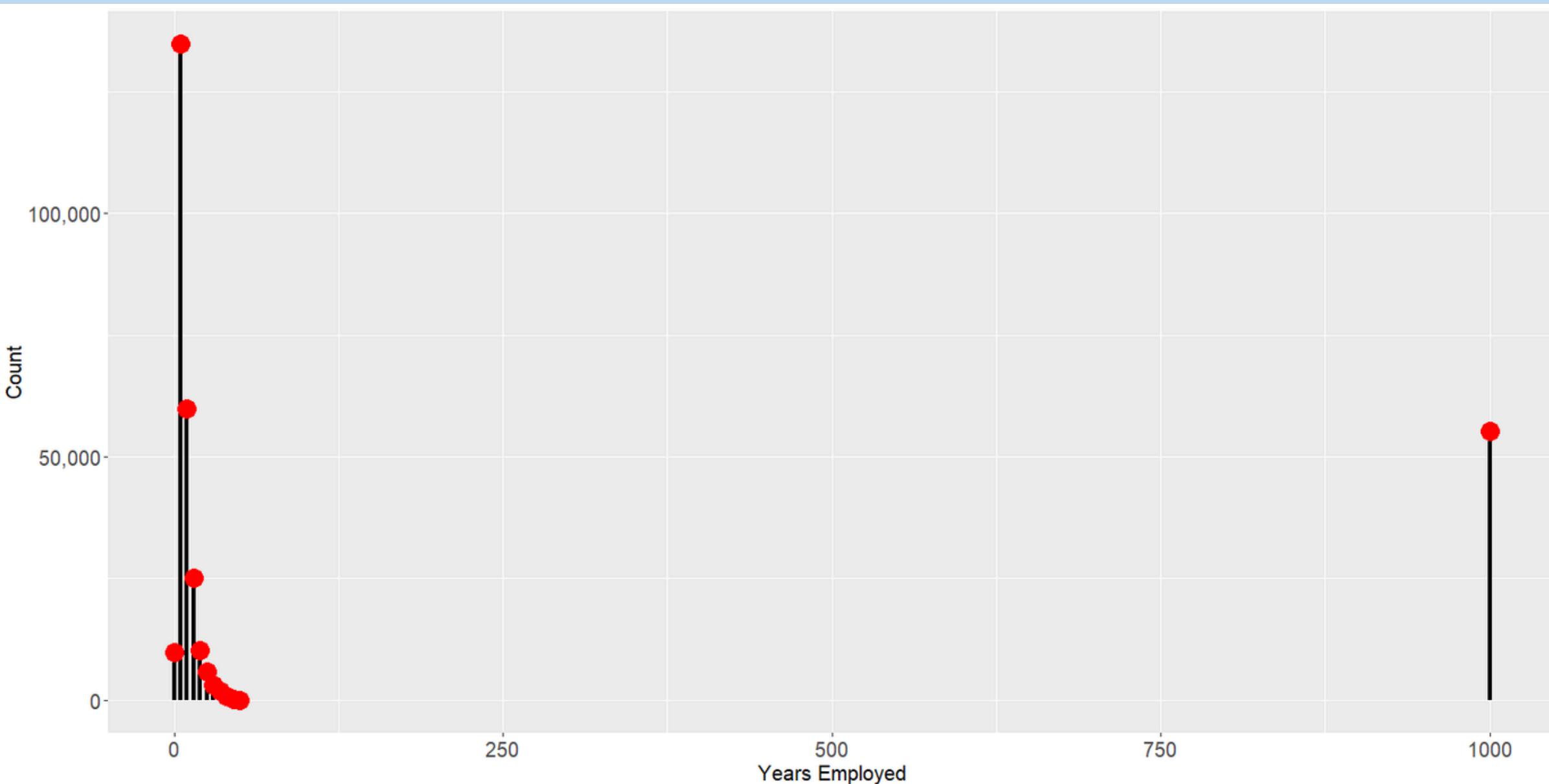


保留1000年資料



去除1000年資料

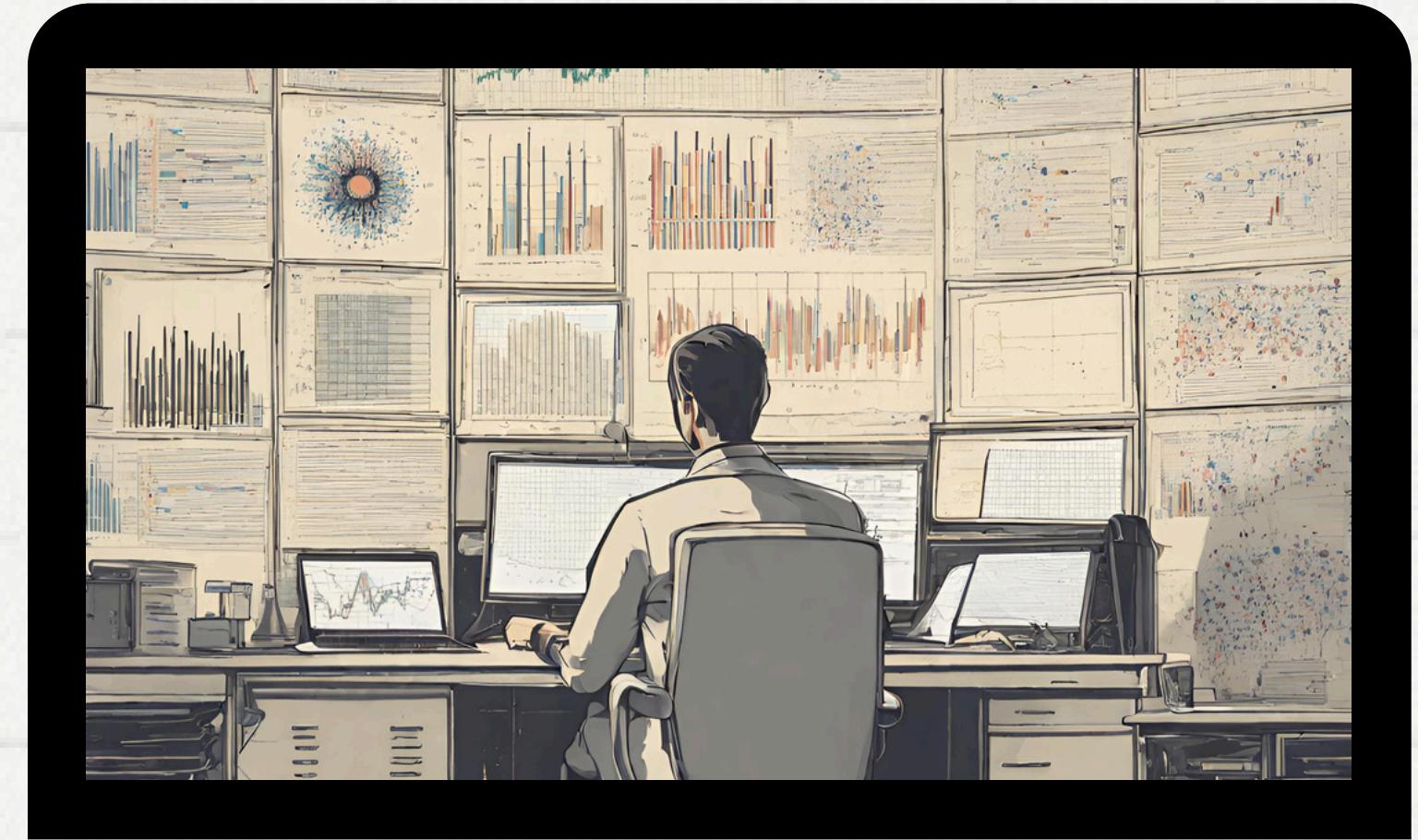
整體以五年為一個組距之莖葉圖：



資料修改 -- 類別型

變數名稱	說明	變動內容
GENDER	變數中的 XNA	不進行插補 將其歸類為一個新的類別 others
ORGANIZATION_TYPE		歸類為新類別 Pensioner ，表示其為退休的人

04. 遺失值插補



平均數插補

- **AMT_ANNUITY** (12 筆)

這個變數代表客戶的貸款年金

- **AMT_GOODS_PRICE**

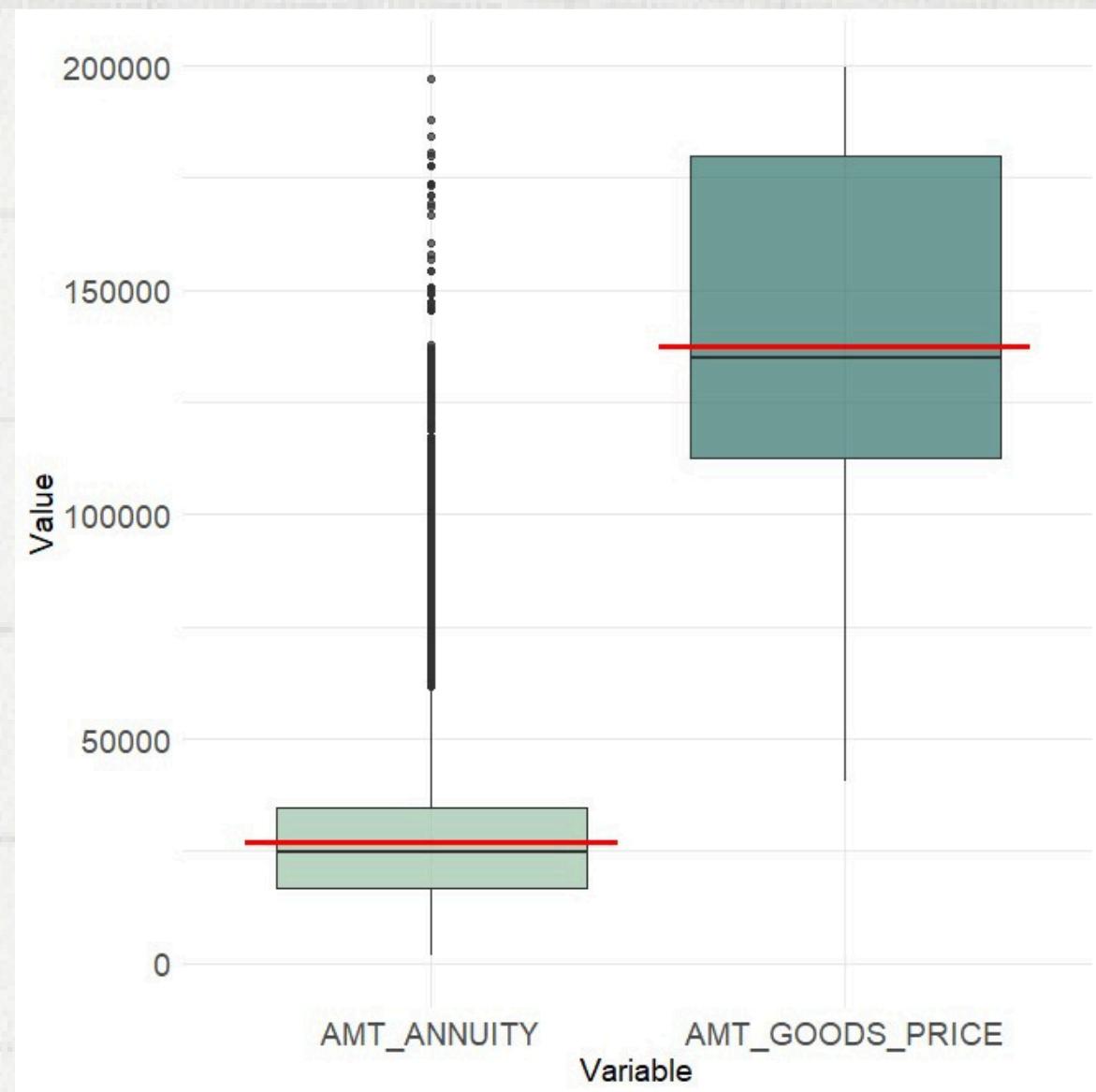
(276 筆，佔全部資料的 0.09%)

這個變數代表客戶欲購買商品的價格

平均數插補

- **AMT_ANNUITY** (12 筆)
這個變數代表客戶的貸款年金

- **AMT_GOODS_PRICE**
(276 筆，佔全部資料的 0.09%)
這個變數代表客戶欲購買商品的價格



平均數插補

- **AMT_ANNUITY** (12 筆)
這個變數代表客戶的貸款年金

- **AMT_GOODS_PRICE**
(276 筆，佔全部資料的 0.09%)
這個變數代表客戶欲購買商品的價格



變數 統計量	AMT_ANNUITY	AMT_GOODS_PRICE
平均數	27123.36	538694.10
中位數	24930	450000
標準差	14475.81	369455.07
q1	16564.50	238500.00
q3	34596.00	679500.00

平均數插補

- **EXT_SOURCE_2**

(658 筆，佔全部變數 0.2%)

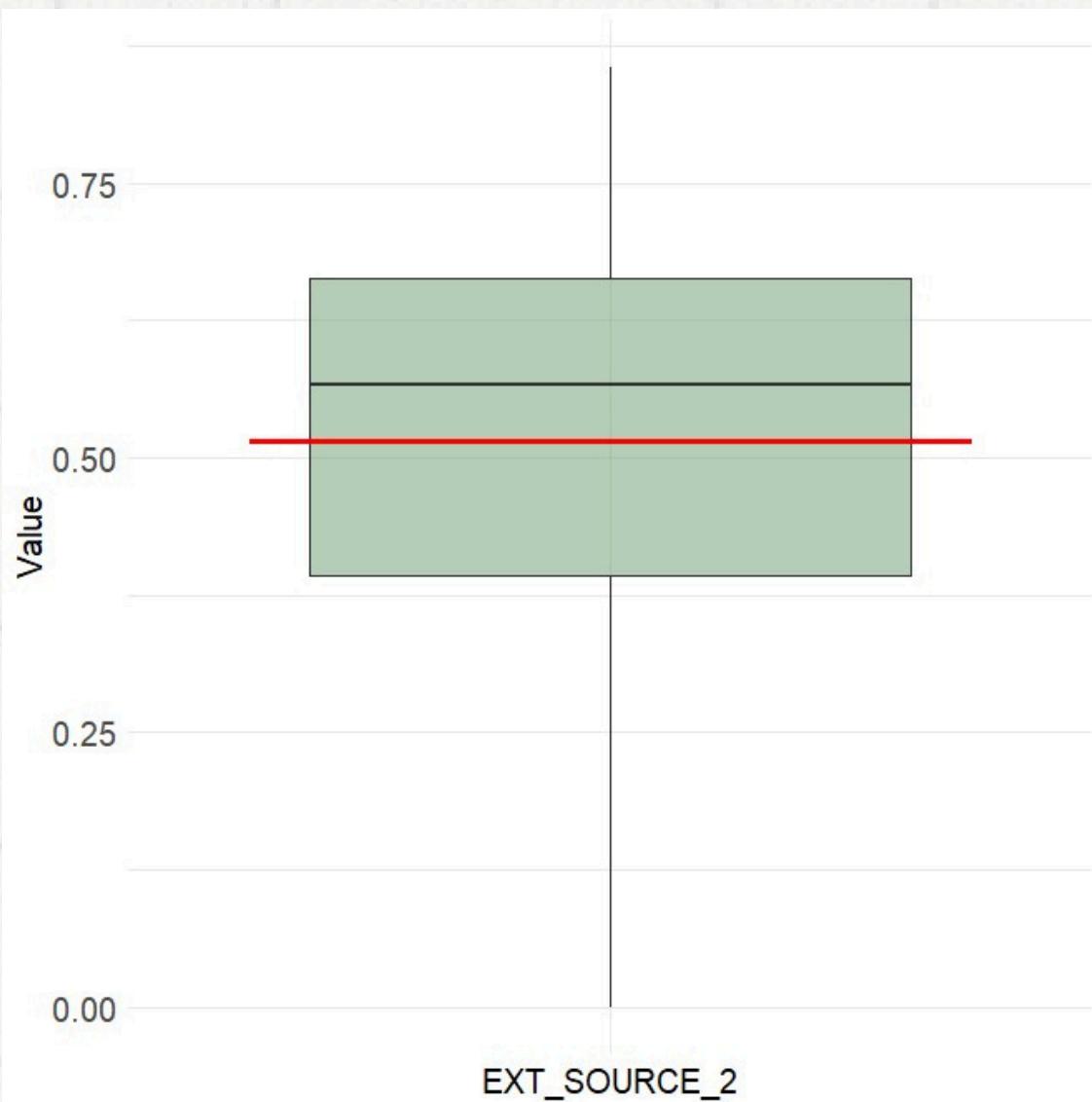
這個變數代表外部數據來源 2 的正規化分數

平均數插補

- **EXT_SOURCE_2**

(658 筆，佔全部變數 0.2%)

這個變數代表外部數據來源 2 的正規化分數

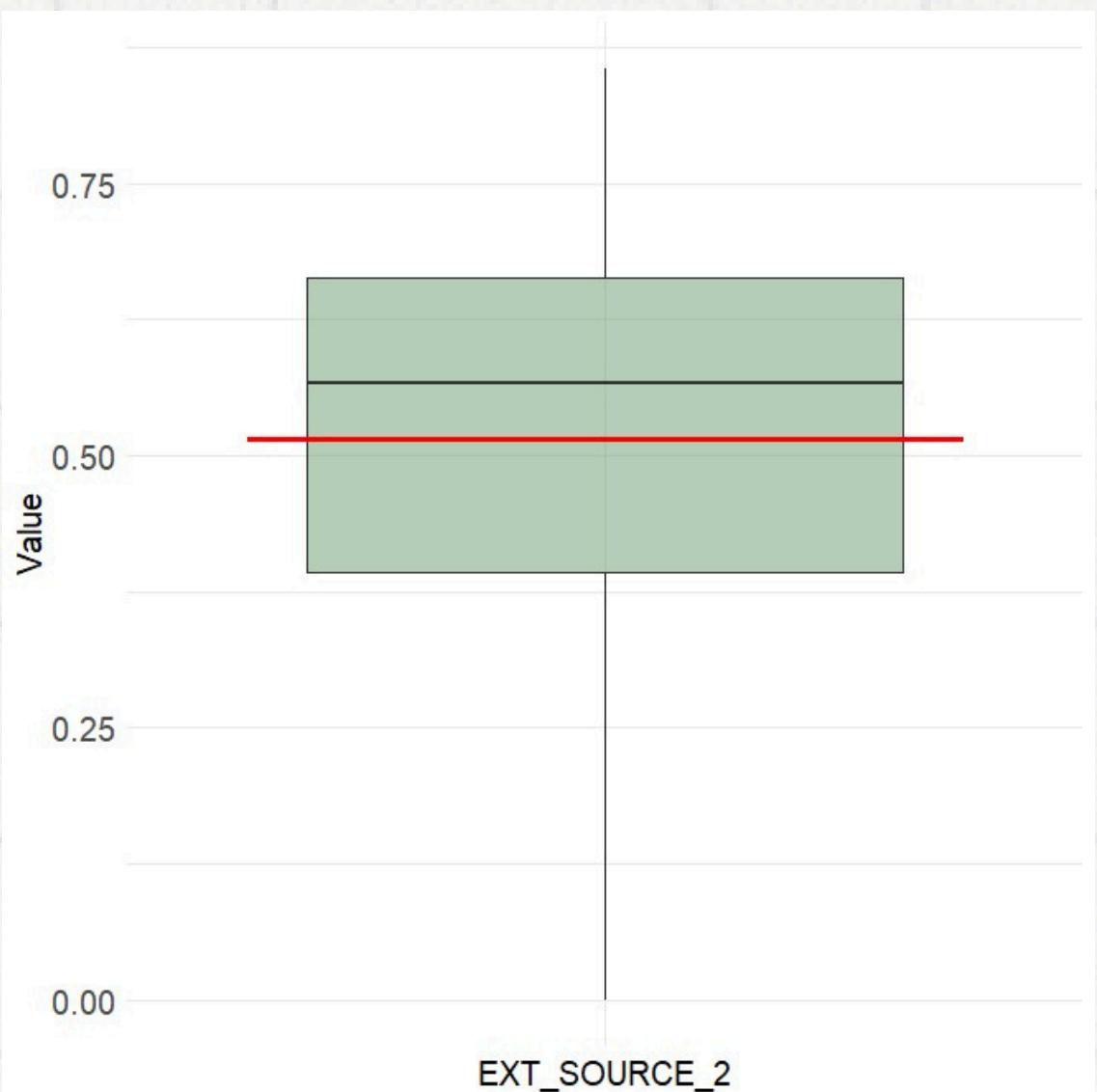


平均數插補

- **EXT_SOURCE_2**

(658 筆，佔全部變數 0.2%)

這個變數代表外部數據來源 2 的正規化分數

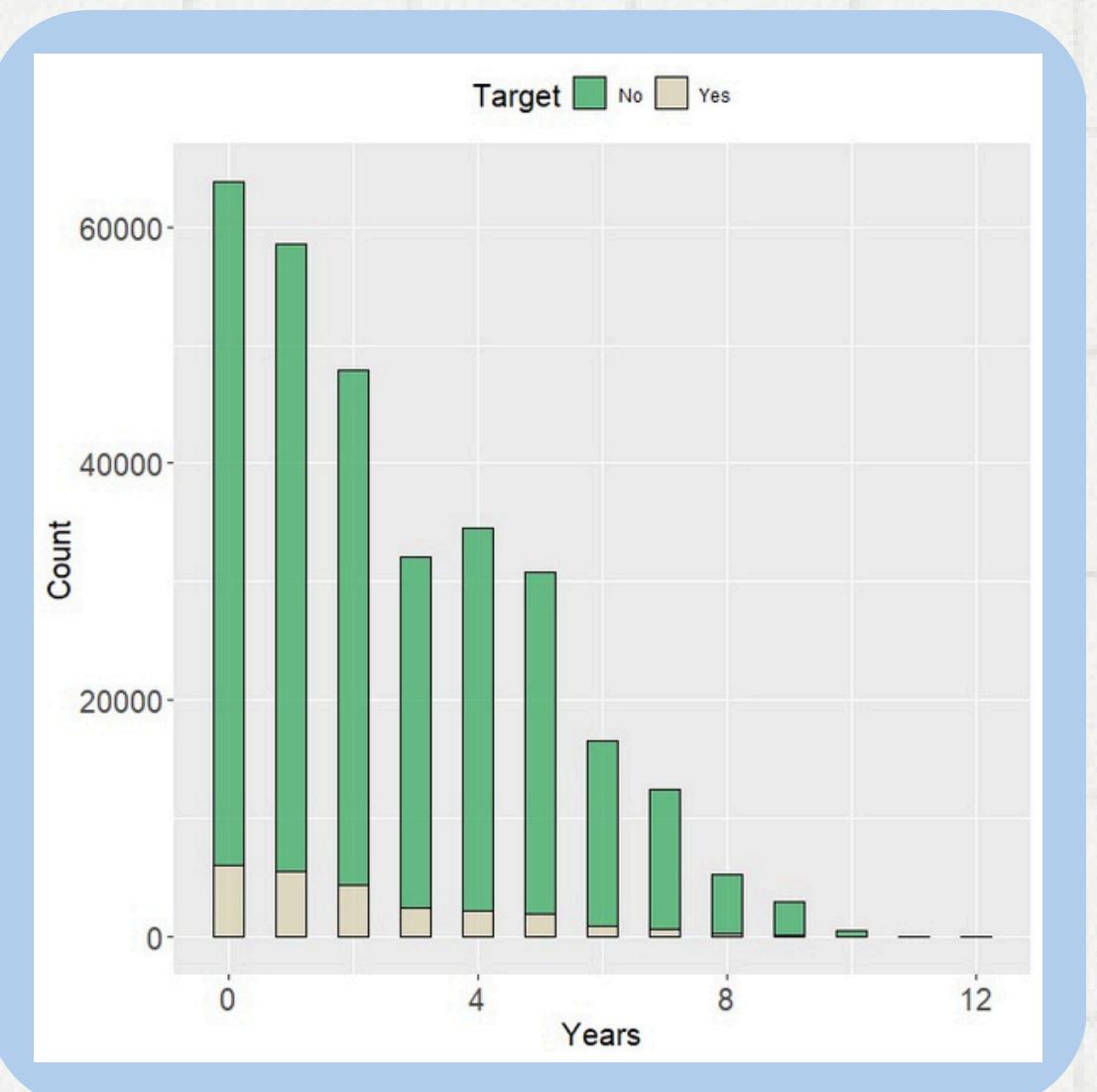


統計量	變數
平均數	EXT_SOURCE_2
中位數	0.5143
標準差	0.5659
q1	0.1911
q3	0.3924

眾數插補

- **YEARS_LAST_PHONE_CHANGE** (1 筆):

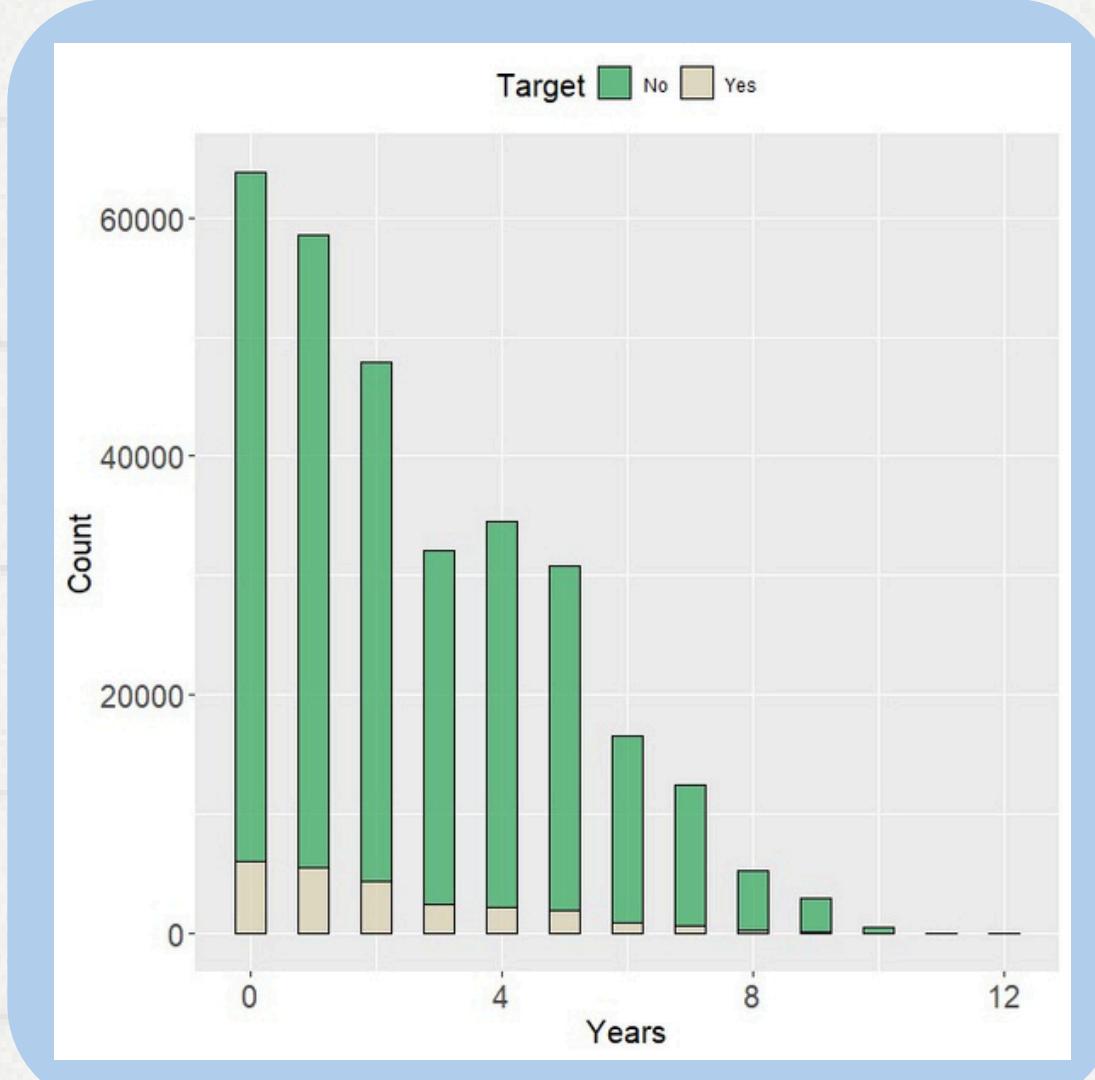
該類別用於記錄客戶在申請貸款前多少年換過手機
將遺失值替換為眾數值 0



眾數插補

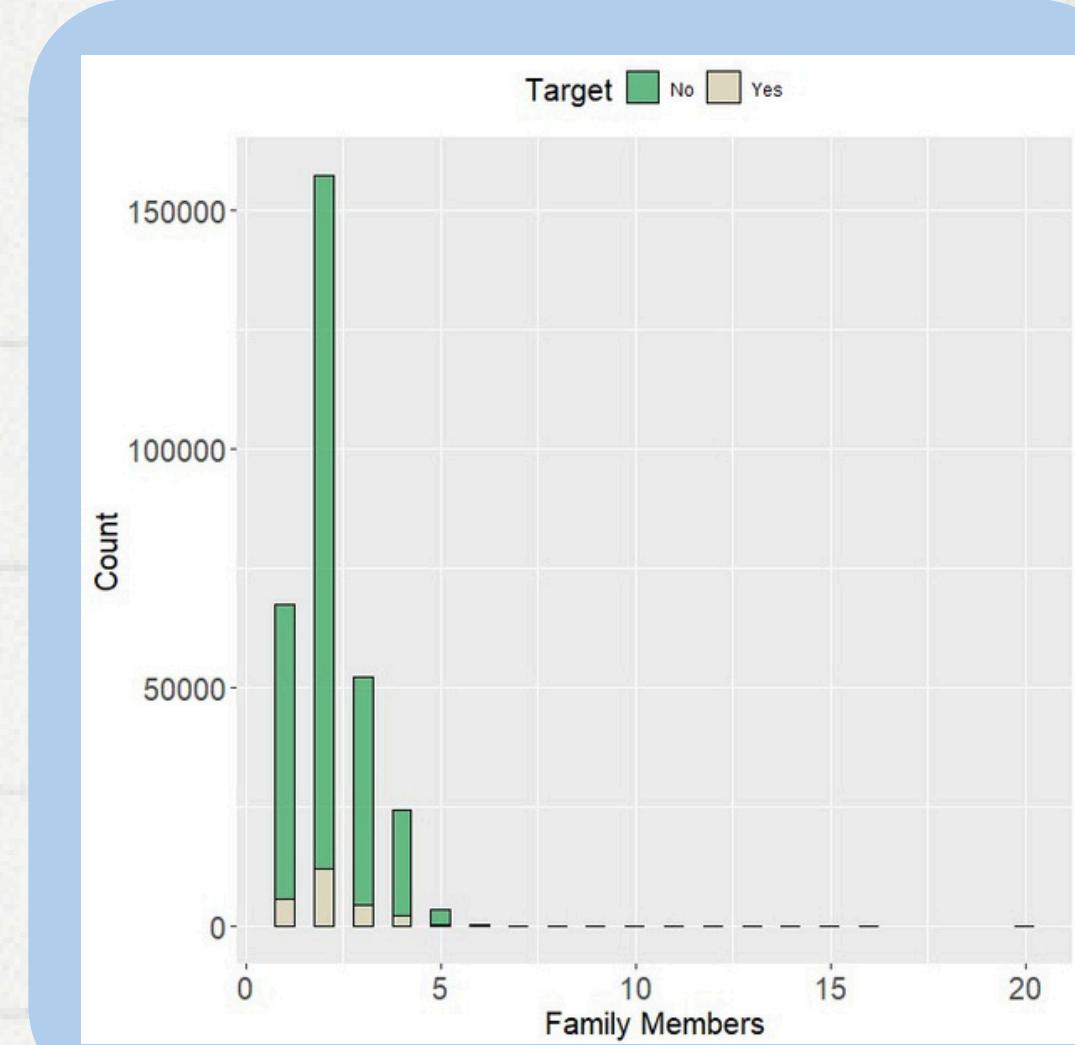
- **YEARS_LAST_PHONE_CHANGE** (1 筆):

該類別用於記錄客戶在申請貸款前多少年換過手機
將遺失值替換為眾數值 0



- **CNT_FAM_MEMBERS** (2 筆):

該類別用於記錄客戶家庭成員數量
將遺失值替換為眾數值 2



新增新類別

變數名稱	資料量	說明
OCCUPATION_TYPE	96109 筆 佔全部資料 31%	用於記錄客戶的職業 將所有遺失值歸類為一個新類別「others」

新增新類別

變數名稱	資料量	說明
OCCUPATION_TYPE	96109 筆 佔全部資料 31%	用於記錄客戶的職業 將所有遺失值歸類為一個新類別「others」
NAME_TYPE_SUITE	1289 筆 佔全部資料 0.4%	用於記錄客戶在申請貸款時的陪同人員 歸類為一個新的類別「Non collected」 表示陪同人員的資訊是未知的或未提供的

刪除資料

變數名稱

在客戶的社交環境中的貸款情況：

OBS_30_CNT_SOCIAL_CIRCLE --- 30 天過期

DEF_30_CNT_SOCIAL_CIRCLE --- 30 天內未按時還款

OBS_60_CNT_SOCIAL_CIRCLE --- 60 天過期

DEF_60_CNT_SOCIAL_CIRCLE --- 60 天內未按時還款

刪除資料

變數名稱

在客戶的社交環境中的貸款情況：

OBS_30_CNT_SOCIAL_CIRCLE --- 30 天過期

DEF_30_CNT_SOCIAL_CIRCLE --- 30 天內未按時還款

OBS_60_CNT_SOCIAL_CIRCLE --- 60 天過期

DEF_60_CNT_SOCIAL_CIRCLE --- 60 天內未按時還款

1020 筆

佔全部資料的 0.3%

PMM

使用 R 程式語言的 MICE 套件(Multivariate Imputation by Chained Equations)
PMM (Predictive Mean Matching) 是一種基於模型的資料插補方法

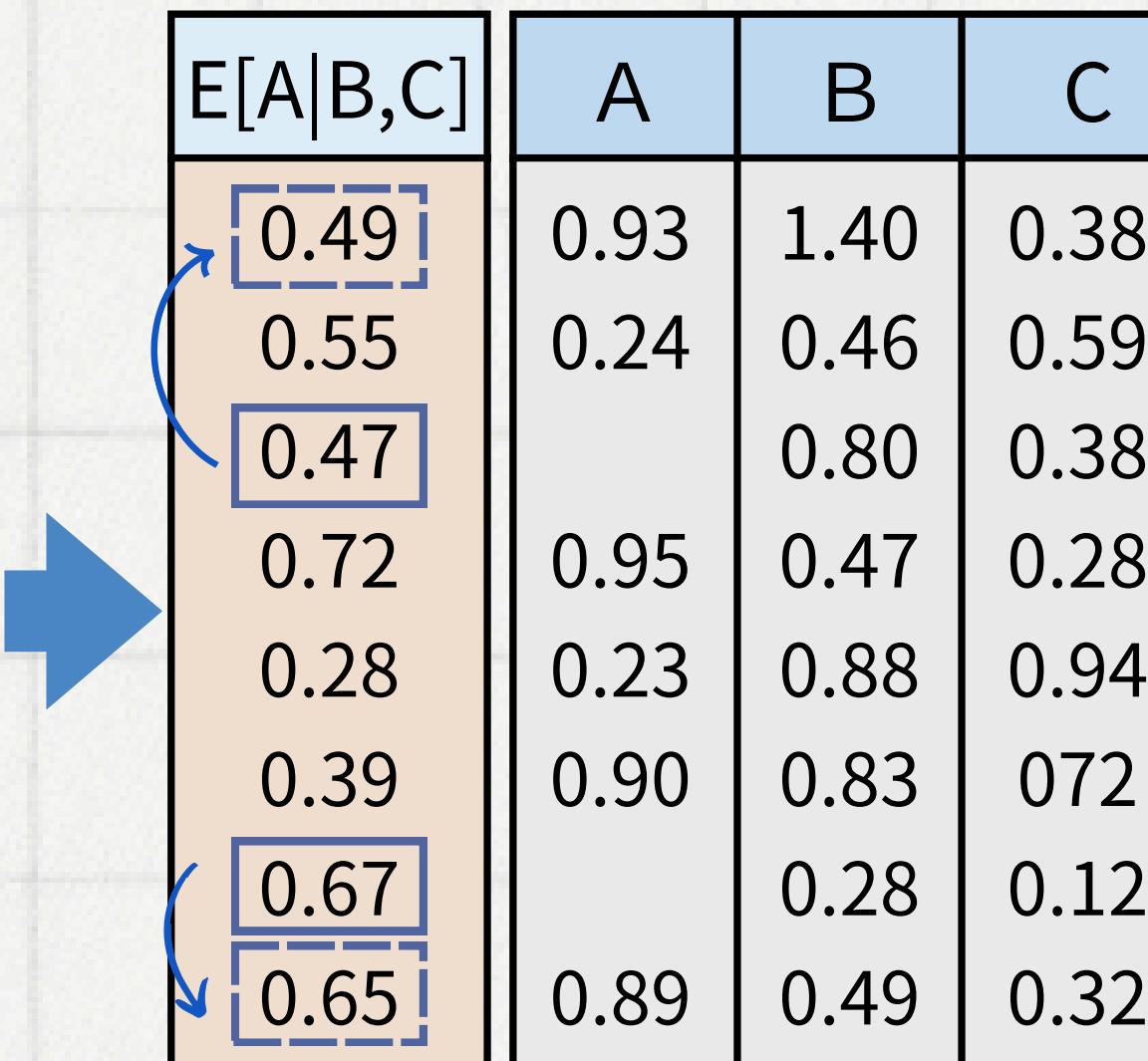
PMM

使用 R 程式語言的 MICE 套件(Multivariate Imputation by Chained Equations)
PMM (Predictive Mean Matching) 是一種基於模型的資料插補方法

$E[A B,C]$	A	B	C
0.49	0.93	1.40	0.38
0.55	0.24	0.46	0.59
0.47	0.00	0.80	0.38
0.72	0.95	0.47	0.28
0.28	0.23	0.88	0.94
0.39	0.90	0.83	0.72
0.67	0.00	0.28	0.12
0.65	0.89	0.49	0.32

PMM

使用 R 程式語言的 MICE 套件(Multivariate Imputation by Chained Equations)
 PMM (Predictive Mean Matching) 是一種基於模型的資料插補方法

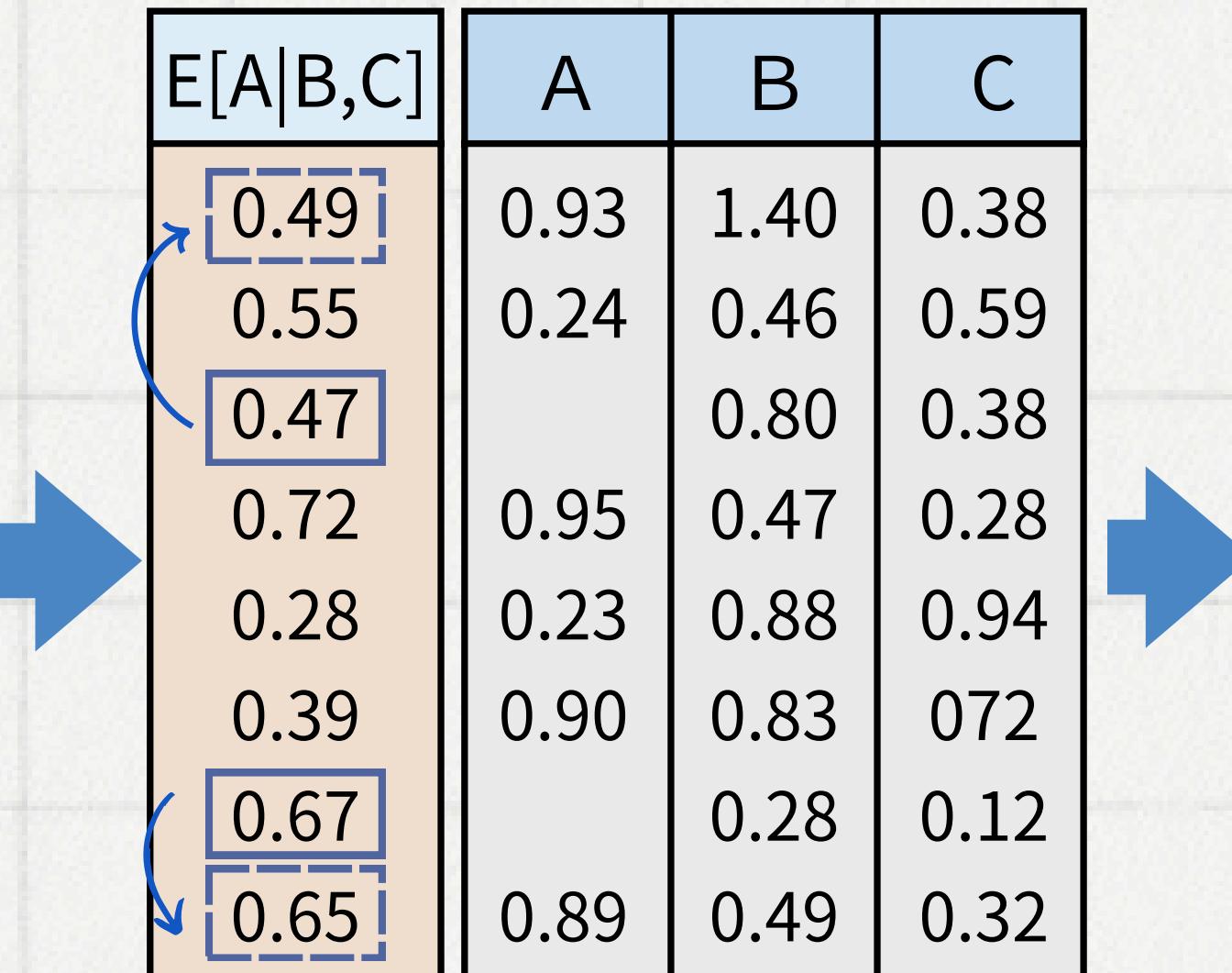


$E[A B,C]$	A	B	C
0.49	0.93	1.40	0.38
0.55	0.24	0.46	0.59
0.47			0.38
0.72	0.95	0.47	0.28
0.28	0.23	0.88	0.94
0.39	0.90	0.83	0.72
0.67			0.12
0.65	0.89	0.49	0.32

$E[A B,C]$	A	B	C
0.49	0.93	1.40	0.38
0.55	0.24	0.46	0.59
0.47	0.80	0.38	0.38
0.72	0.95	0.47	0.28
0.28	0.23	0.88	0.94
0.39	0.90	0.83	0.72
0.67	0.28	0.28	0.12
0.65	0.89	0.49	0.32

PMM

使用 R 程式語言的 MICE 套件(Multivariate Imputation by Chained Equations)
 PMM (Predictive Mean Matching) 是一種基於模型的資料插補方法



$E[A B,C]$	A	B	C
0.49	0.93	1.40	0.38
0.55	0.24	0.46	0.59
0.47	0.80	0.38	
0.72	0.95	0.47	0.28
0.28	0.23	0.88	0.94
0.39	0.90	0.83	0.72
0.67	0.28	0.12	
0.65	0.89	0.49	0.32

$E[A B,C]$	A	B	C
0.49	0.93	1.40	0.38
0.55	0.24	0.46	0.59
0.47	0.80	0.38	
0.72	0.95	0.47	0.28
0.28	0.23	0.88	0.94
0.39	0.90	0.83	0.72
0.67	0.28	0.12	
0.65	0.89	0.49	0.32

$E[A B,C]$	A	B	C
0.49	0.93	1.40	0.38
0.55	0.24	0.46	0.59
0.47	0.80	0.38	
0.72	0.95	0.47	0.28
0.28	0.23	0.88	0.94
0.39	0.90	0.83	0.72
0.67	0.28	0.12	
0.65	0.89	0.49	0.32

PMM插補的步驟：

PMM插補的步驟：

Step 1. 使用 PMM 插補五次

PMM插補的步驟：

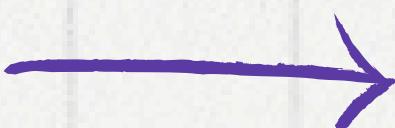
Step 1. 使用 PMM 插補五次



Step 2. 繪畫五次插補的密度分配圖

PMM插補的步驟：

Step 1. 使用 PMM 插補五次



Step 2. 繪畫五次插補的密度分配圖



Step 3. 目測觀察後挑選與原始資料最像的

PMM插補的步驟：

Step 1. 使用 PMM 插補五次



Step 2. 繪畫五次插補的密度分配圖



Step 4. Kolmogorov-Smirnov test



Step 3. 目測觀察後挑選與原始資料最像的

PMM插補的步驟：

Step 1. 使用 PMM 插補五次



Step 2. 繪畫五次插補的密度分配圖



Step 4. Kolmogorov-Smirnov test



Step 5. 看統計量 D 比較五次插補

PMM插補的步驟：

Step 1. 使用 PMM 插補五次



Step 2. 繪畫五次插補的密度分配圖



Step 4. Kolmogorov-Smirnov test



Step 5. 看統計量 D 比較五次插補



Step 6. 確定最終選擇

第一次PMM插補：

插補之變數名稱

AMT_REQ_CREDIT_BUREAU_HOUR

AMT_REQ_CREDIT_BUREAU_DAY

AMT_REQ_CREDIT_BUREAU_WEEK

AMT_REQ_CREDIT_BUREAU_MON

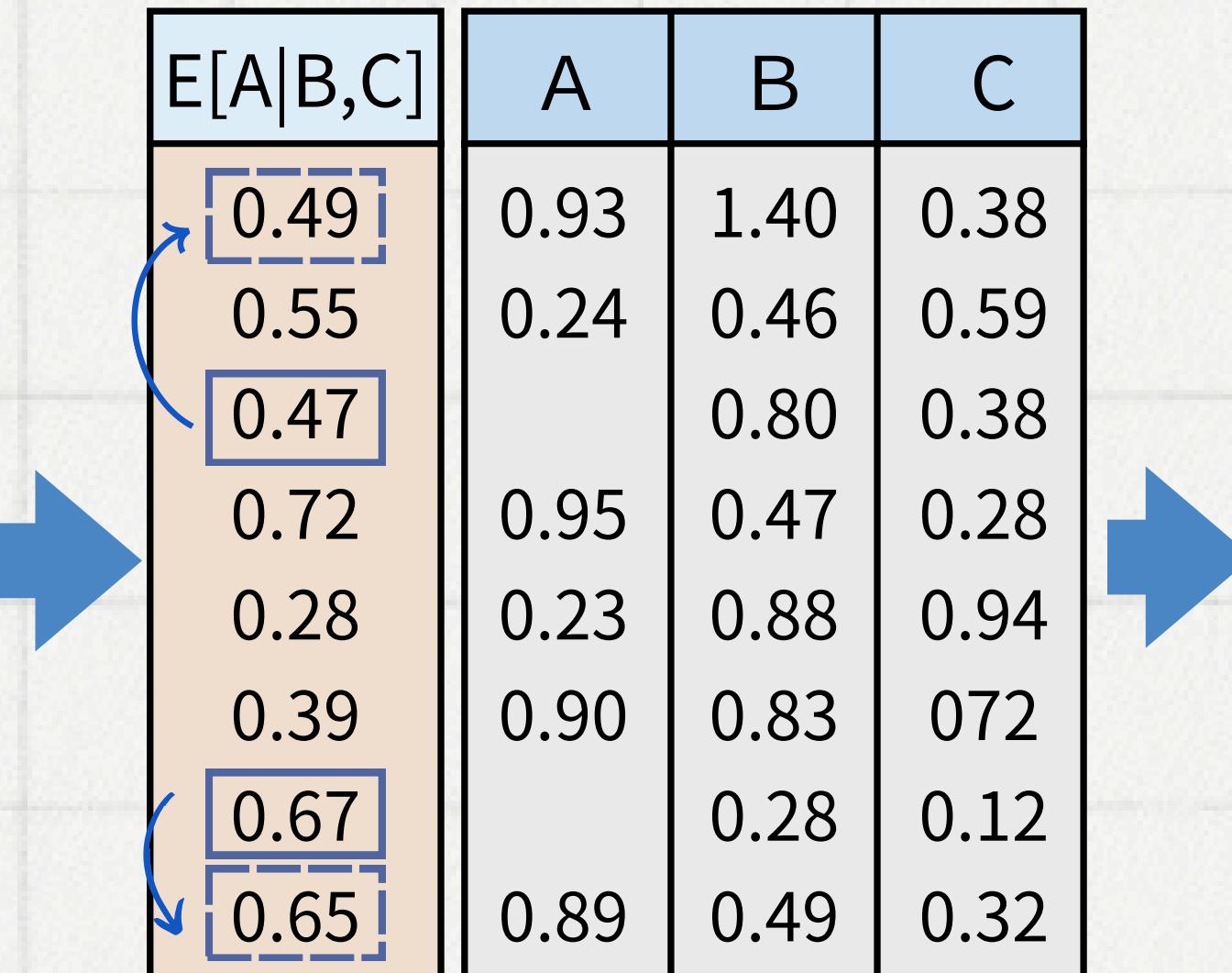
AMT_REQ_CREDIT_BUREAU_QRT

AMT_REQ_CREDIT_BUREAU_YEAR

AMT 系列的六個變數遺失值皆為 41376 筆
且為同一群資料，佔全部變數 13.5%

PMM

使用 R 程式語言的 MICE 套件(Multivariate Imputation by Chained Equations)
 PMM (Predictive Mean Matching) 是一種基於模型的資料插補方法



$E[A B,C]$	A	B	C
0.49	0.93	1.40	0.38
0.55	0.24	0.46	0.59
0.47	0.80	0.38	0.38
0.72	0.95	0.47	0.28
0.28	0.23	0.88	0.94
0.39	0.90	0.83	0.72
0.67	0.28	0.12	0.12
0.65	0.89	0.49	0.32

$E[A B,C]$	A	B	C
0.49	0.93	1.40	0.38
0.55	0.24	0.46	0.59
0.47	0.80	0.38	0.38
0.72	0.95	0.47	0.28
0.28	0.23	0.88	0.94
0.39	0.90	0.83	0.72
0.67	0.28	0.12	0.12
0.65	0.89	0.49	0.32

$E[A B,C]$	A	B	C
0.49	0.93	1.40	0.38
0.55	0.24	0.46	0.59
0.47	0.80	0.38	0.38
0.72	0.95	0.47	0.28
0.28	0.23	0.88	0.94
0.39	0.90	0.83	0.72
0.67	0.28	0.12	0.12
0.65	0.89	0.49	0.32

第一次PMM插補：

插補之變數名稱

AMT_REQ_CREDIT_BUREAU_HOUR

AMT_REQ_CREDIT_BUREAU_DAY

AMT_REQ_CREDIT_BUREAU_WEEK

AMT_REQ_CREDIT_BUREAU_MON

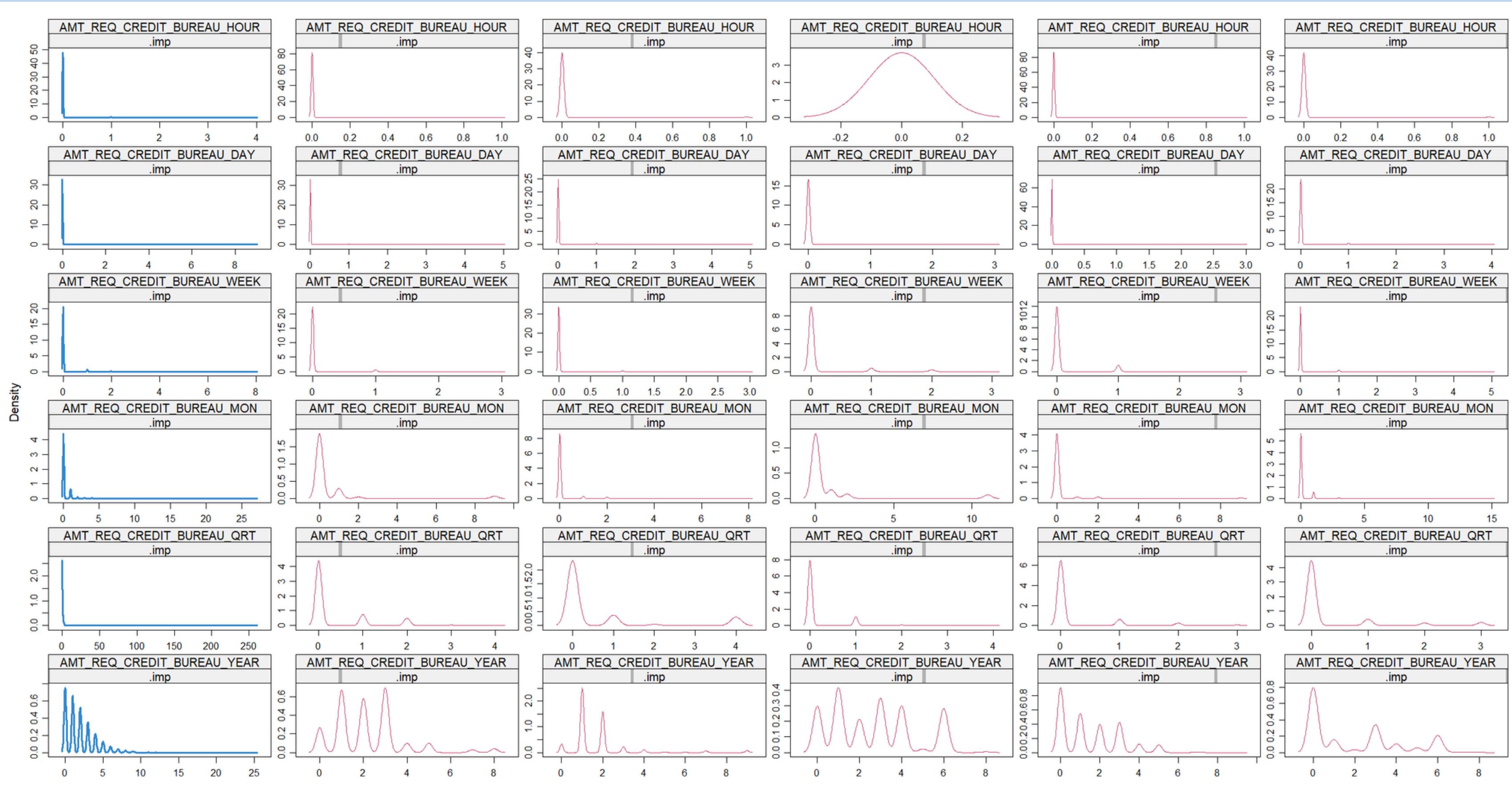
AMT_REQ_CREDIT_BUREAU_QRT

AMT_REQ_CREDIT_BUREAU_YEAR

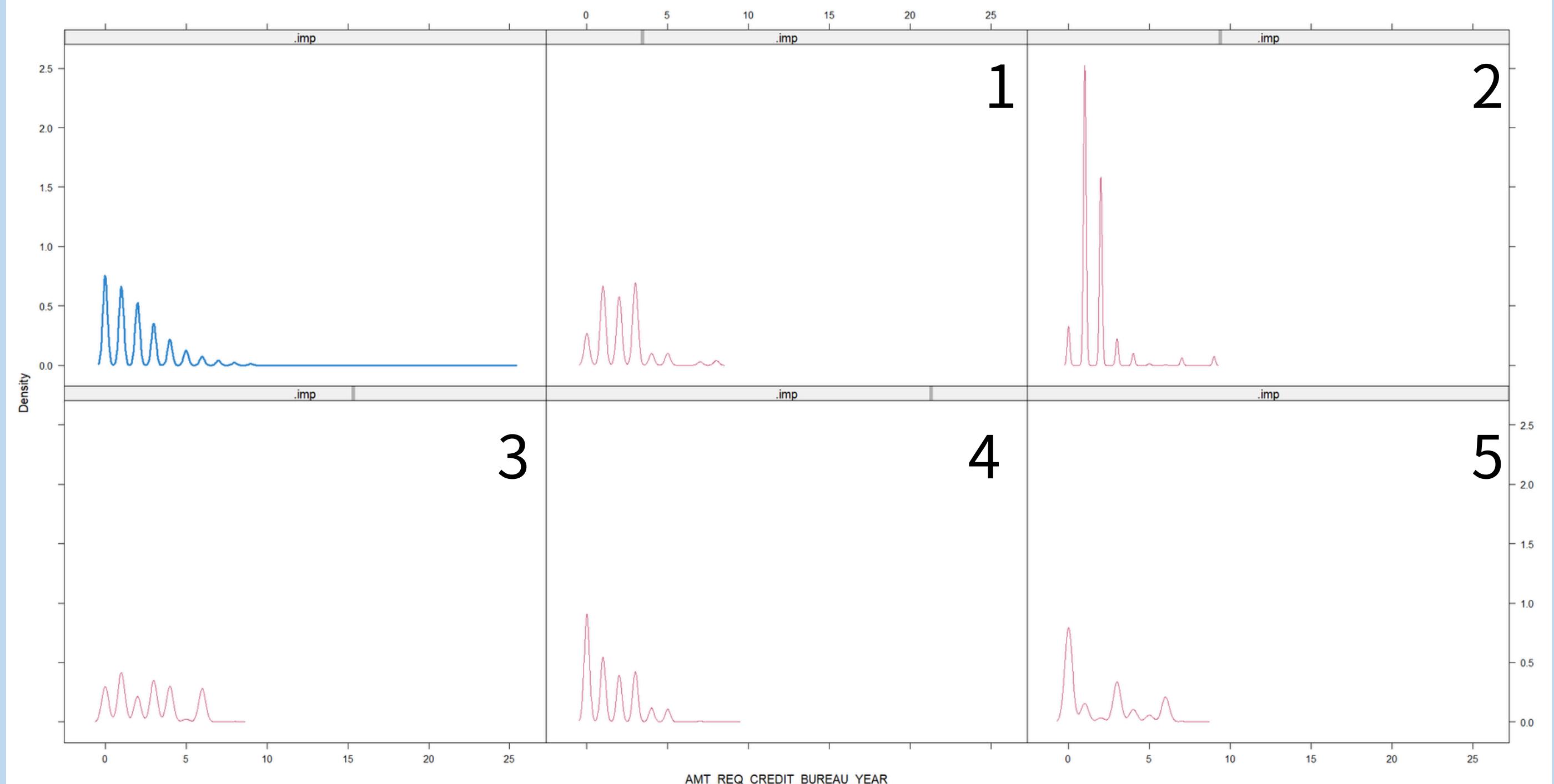
AMT 系列的六個變數遺失值皆為 41376 筆
且為同一群資料，佔全部變數 13.5%

密度分配圖：

04遺失值插補

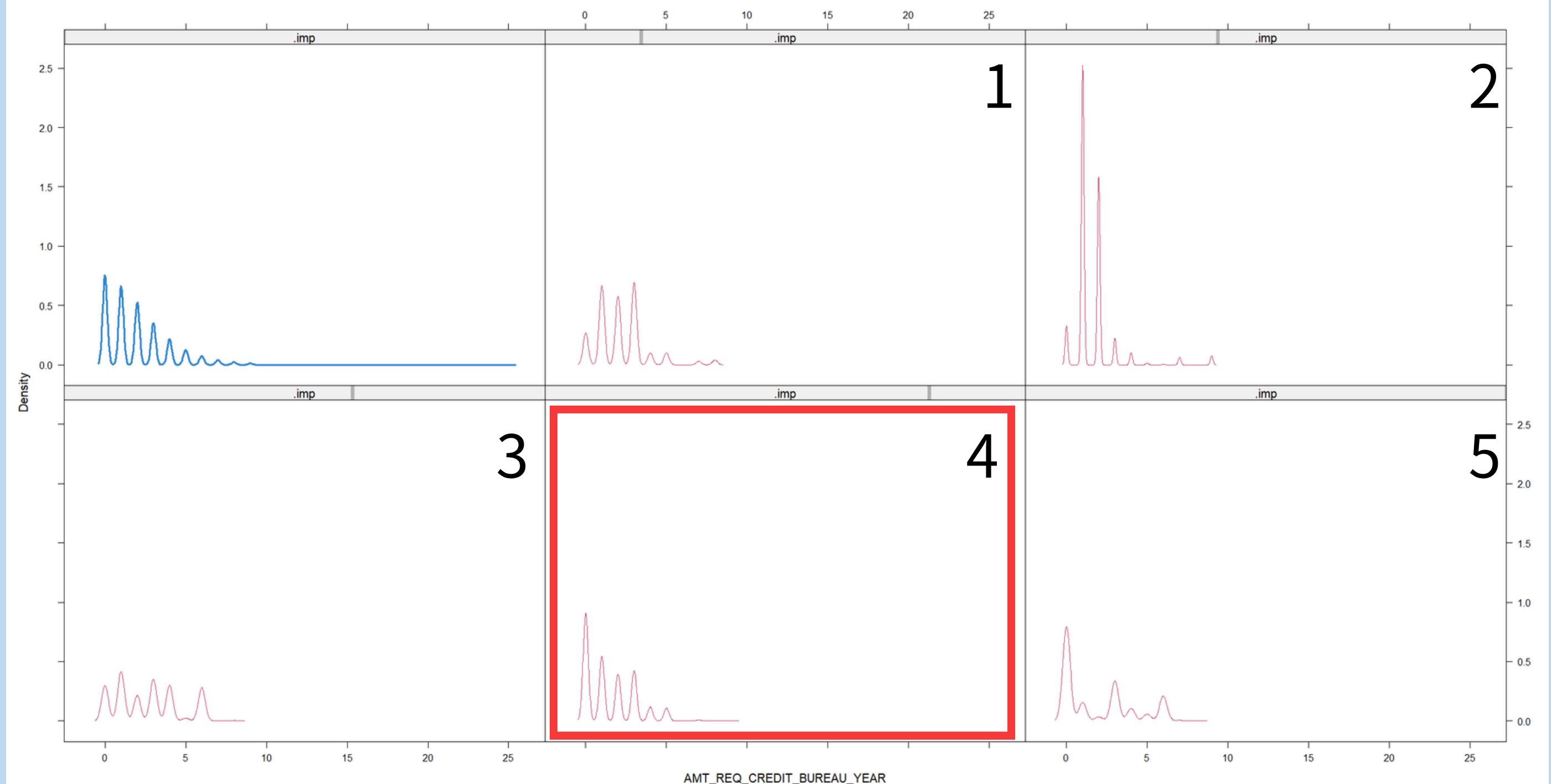


AMT_REQ_CREDIT_BUREAU_YEAR插補結果：



AMT_REQ_CREDIT_BUREAU_YEAR插補結果：

第四次迭代的表現較為出色



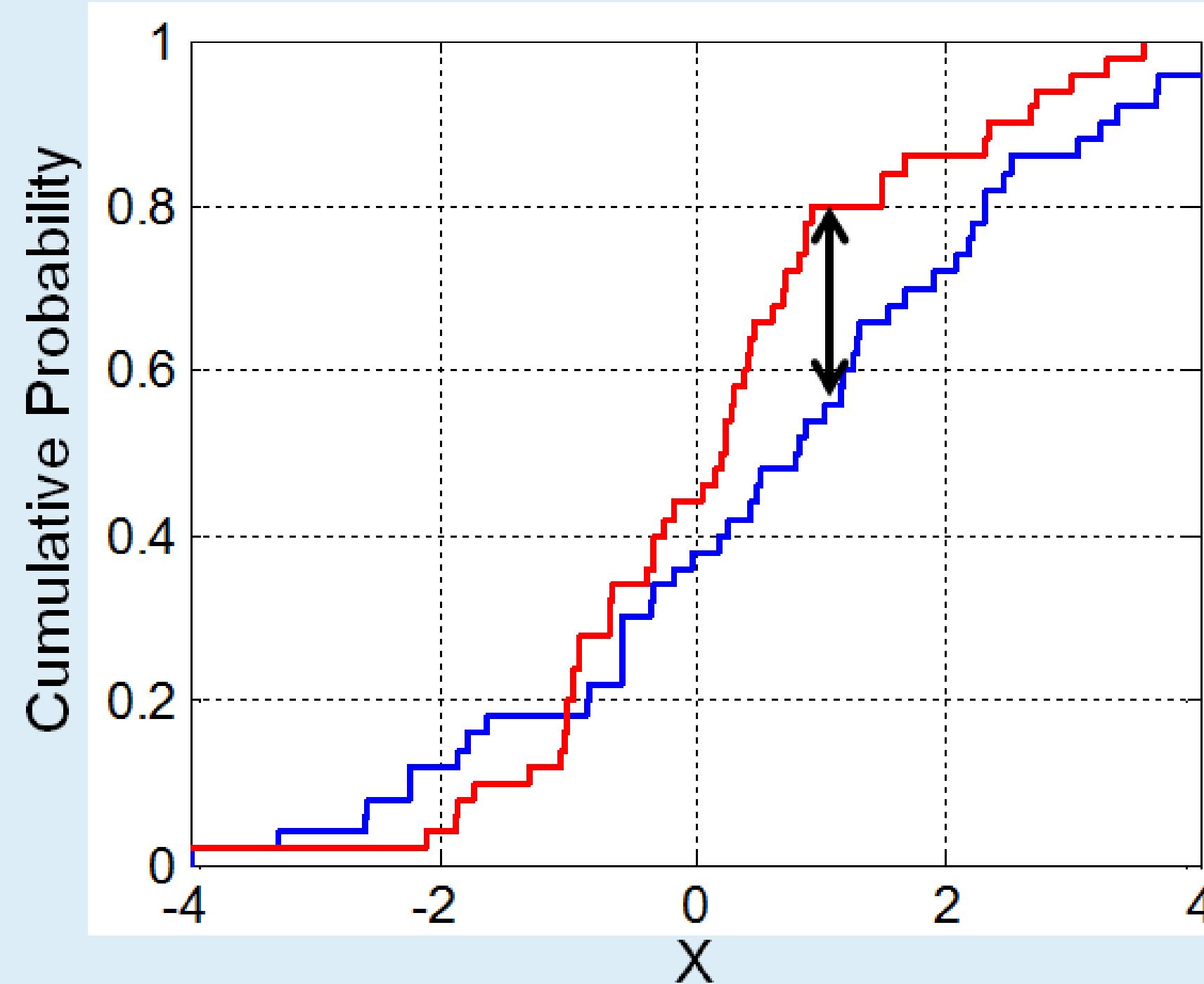
H₀ :The ith distribution is not significantly different from the original distribution
 $i = 1, 2, 3, 4, 5.$

H₁ :The ith distribution is significantly different from the original distribution
 $i = 1, 2, 3, 4, 5.$

D 統計量:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

K-S Test 統計量示意圖：



H₀ :The ith distribution is not significantly different from the original distribution
 $i = 1, 2, 3, 4, 5.$

H₁ :The ith distribution is significantly different from the original distribution
 $i = 1, 2, 3, 4, 5.$

D 統計量:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

統計量 第 i 次 迭代	1	2	3	4	5
D	0.0030	0.0020	0.0028	0.0014	0.0025

H₀ :The ith distribution is not significantly different from the original distribution
 $i = 1, 2, 3, 4, 5.$

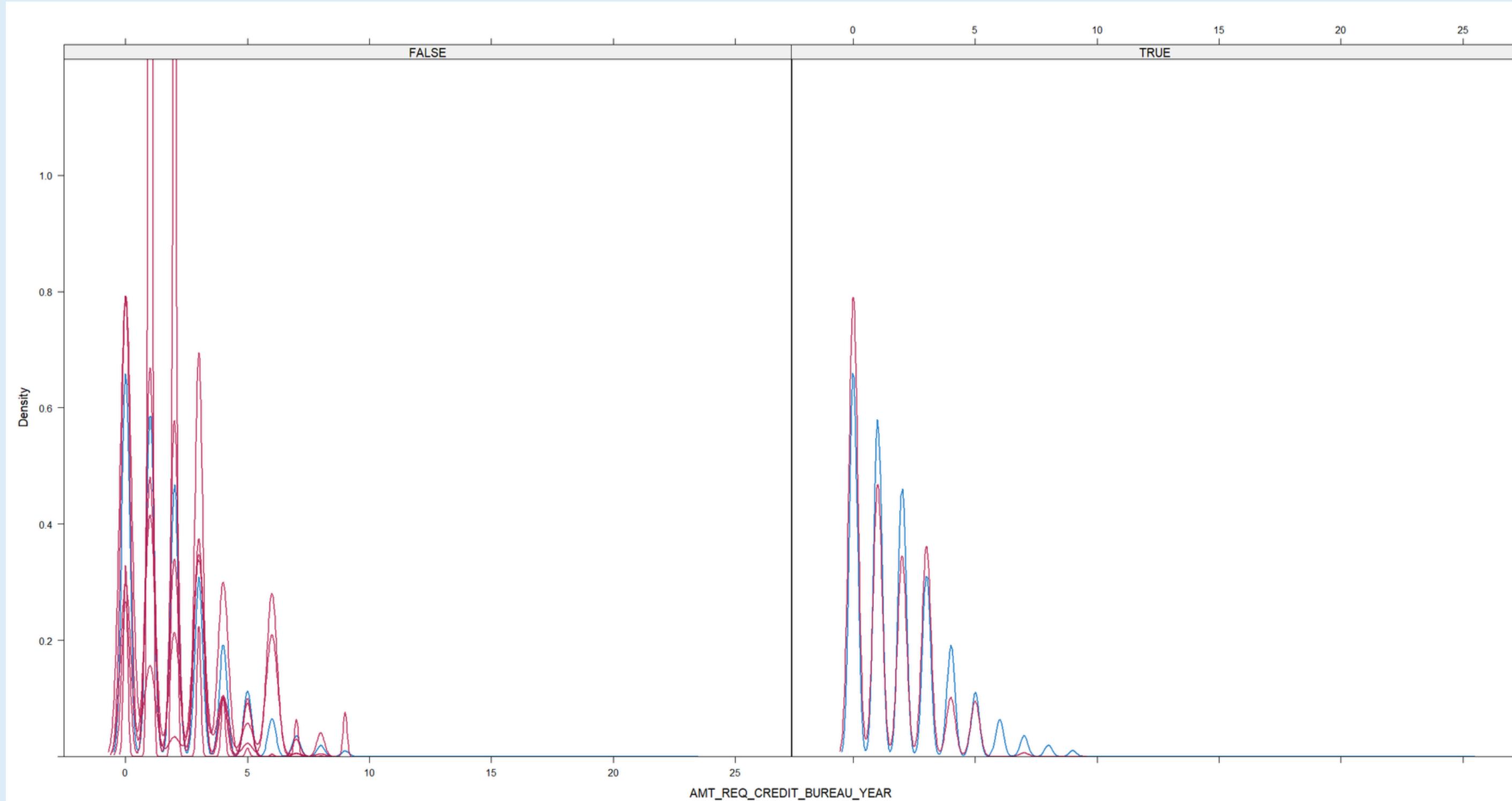
H₁ :The ith distribution is significantly different from the original distribution
 $i = 1, 2, 3, 4, 5.$

D 統計量:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

統計量 第 i 次 迭代	1	2	3	4	5
D	0.0030	0.0020	0.0028	0.0014	0.0025

AMT_YEAR的PMM最終插補密度分配圖：

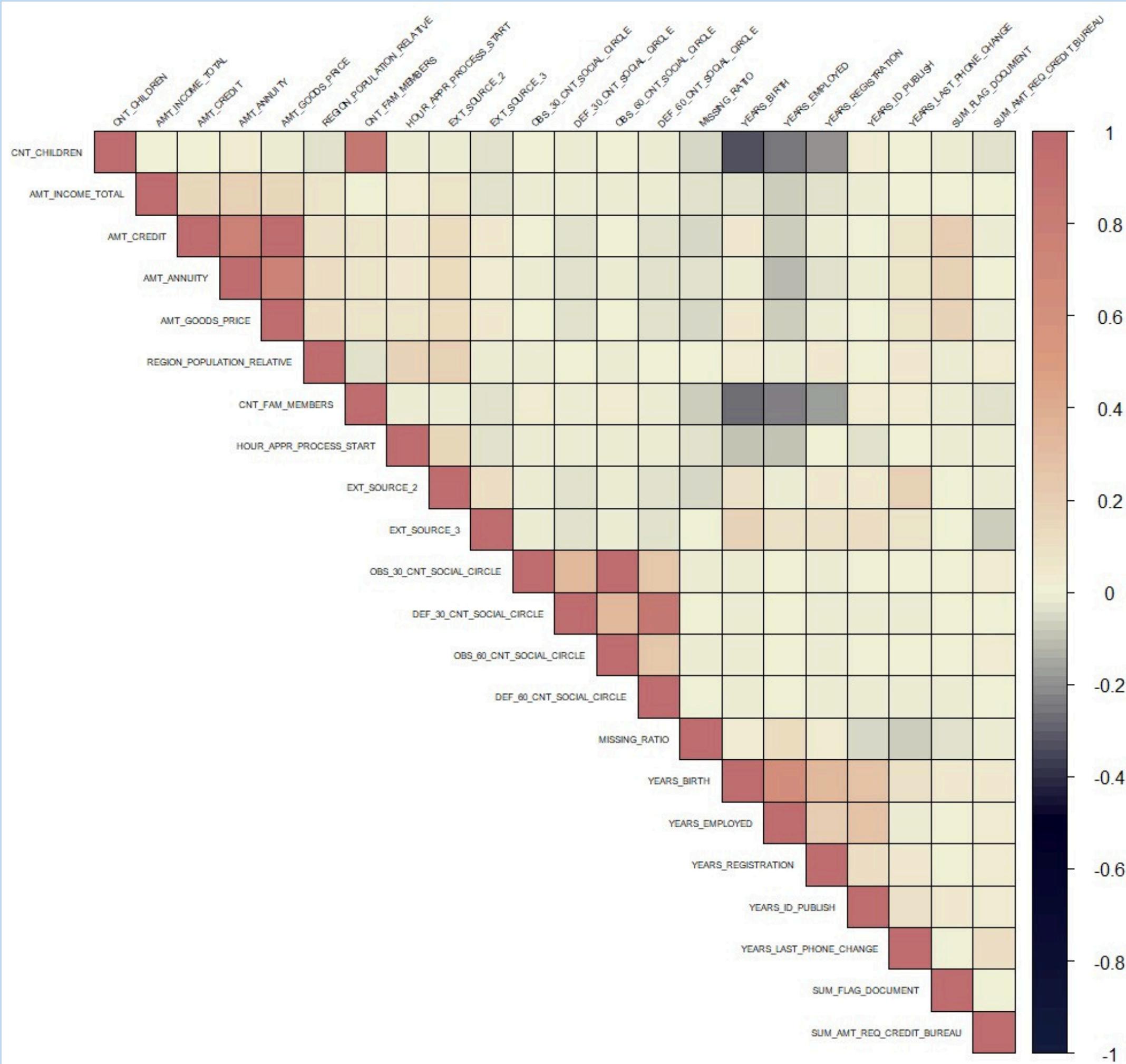


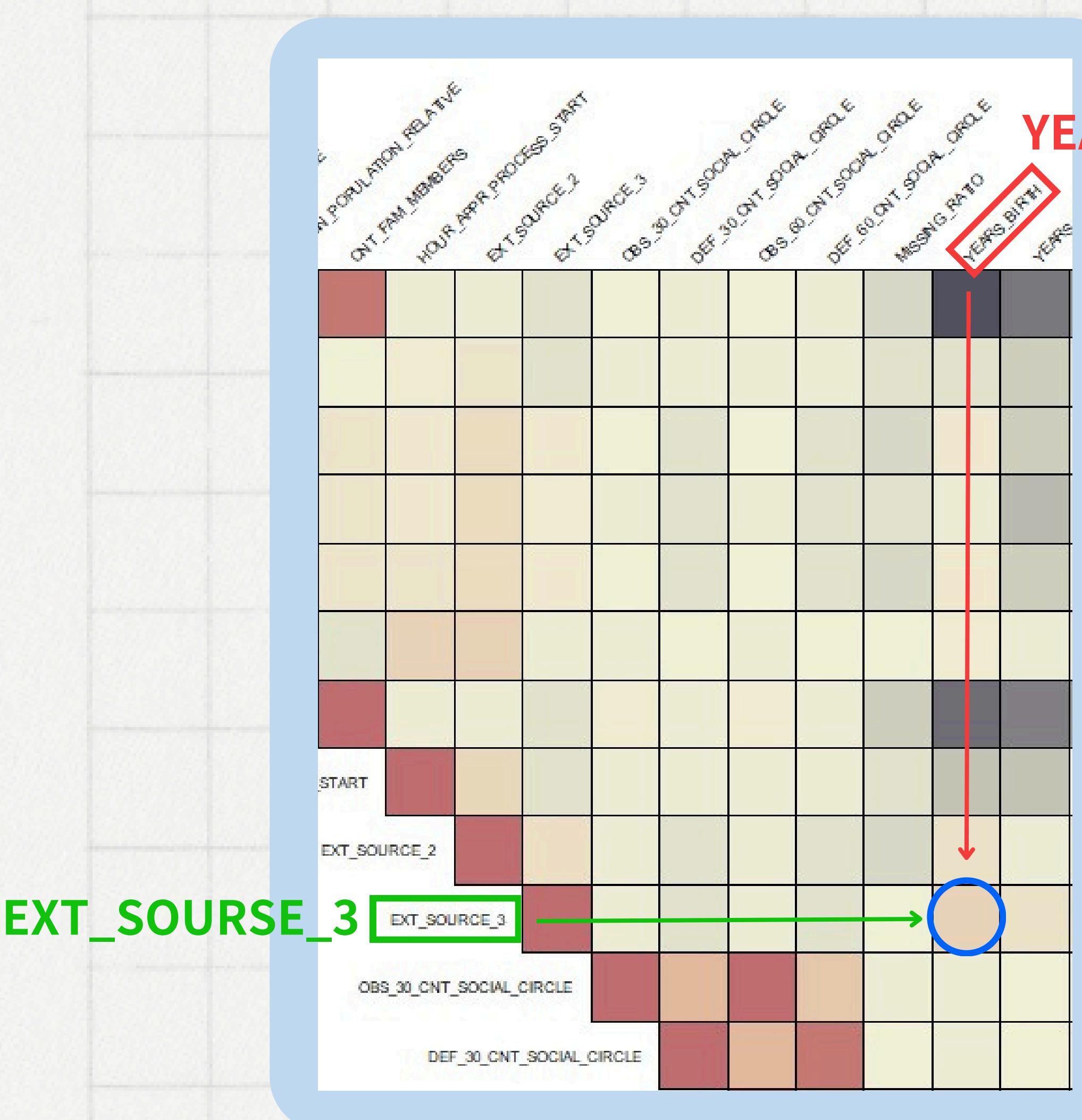
第二次PMM插補：

PMM 使用現有資料對遺失值插補且需要至少兩個有遺失值的變數

- 僅剩EXT_SOURCE_3尚未插補
- 為了更好去預測EXT_SOURCE_3的遺失值，因此觀察相關係數圖

04 遺失值插補





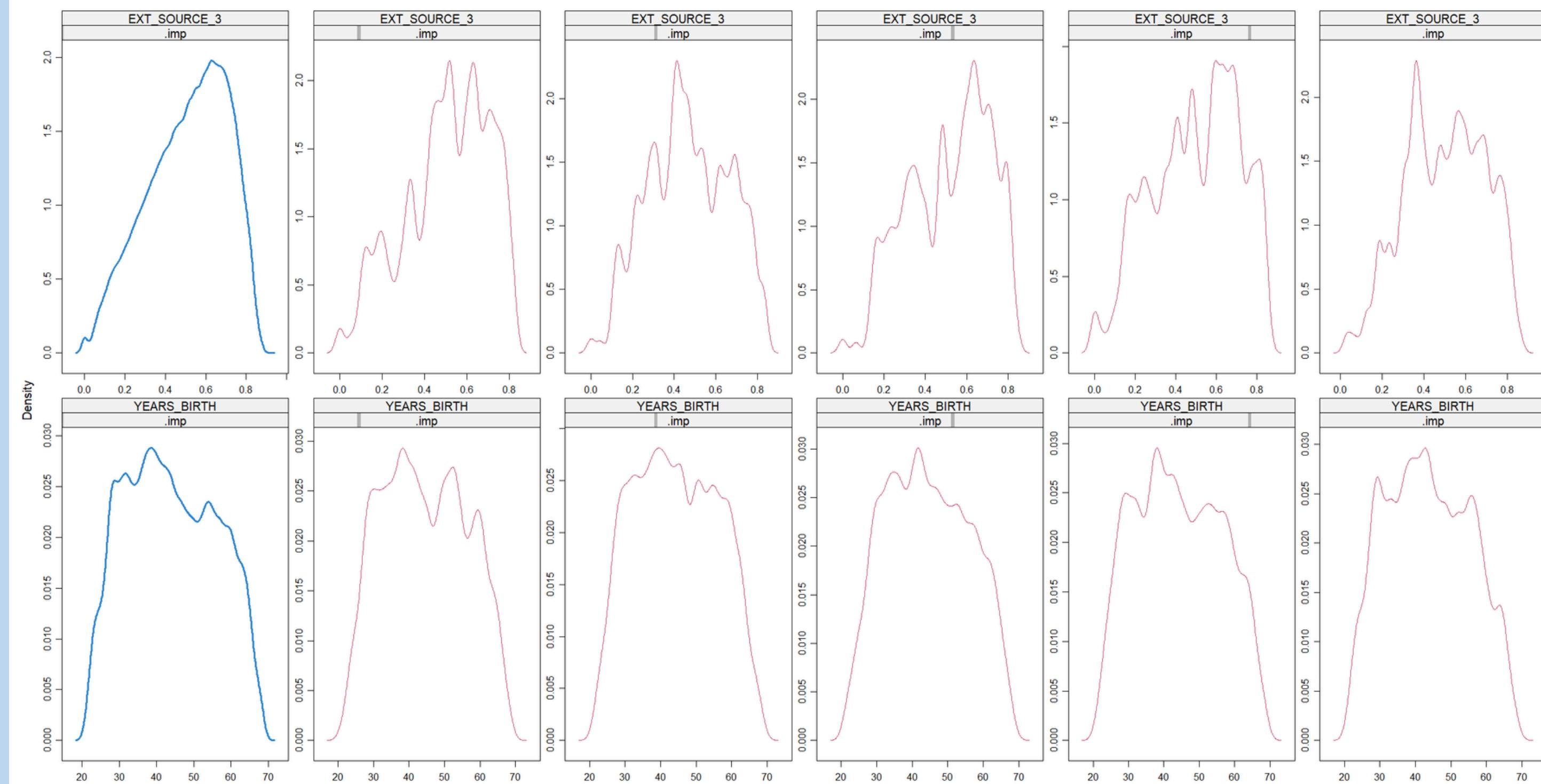
- **EXT_SOURCE_3**

(60771 筆，佔全部變數 19.8%)

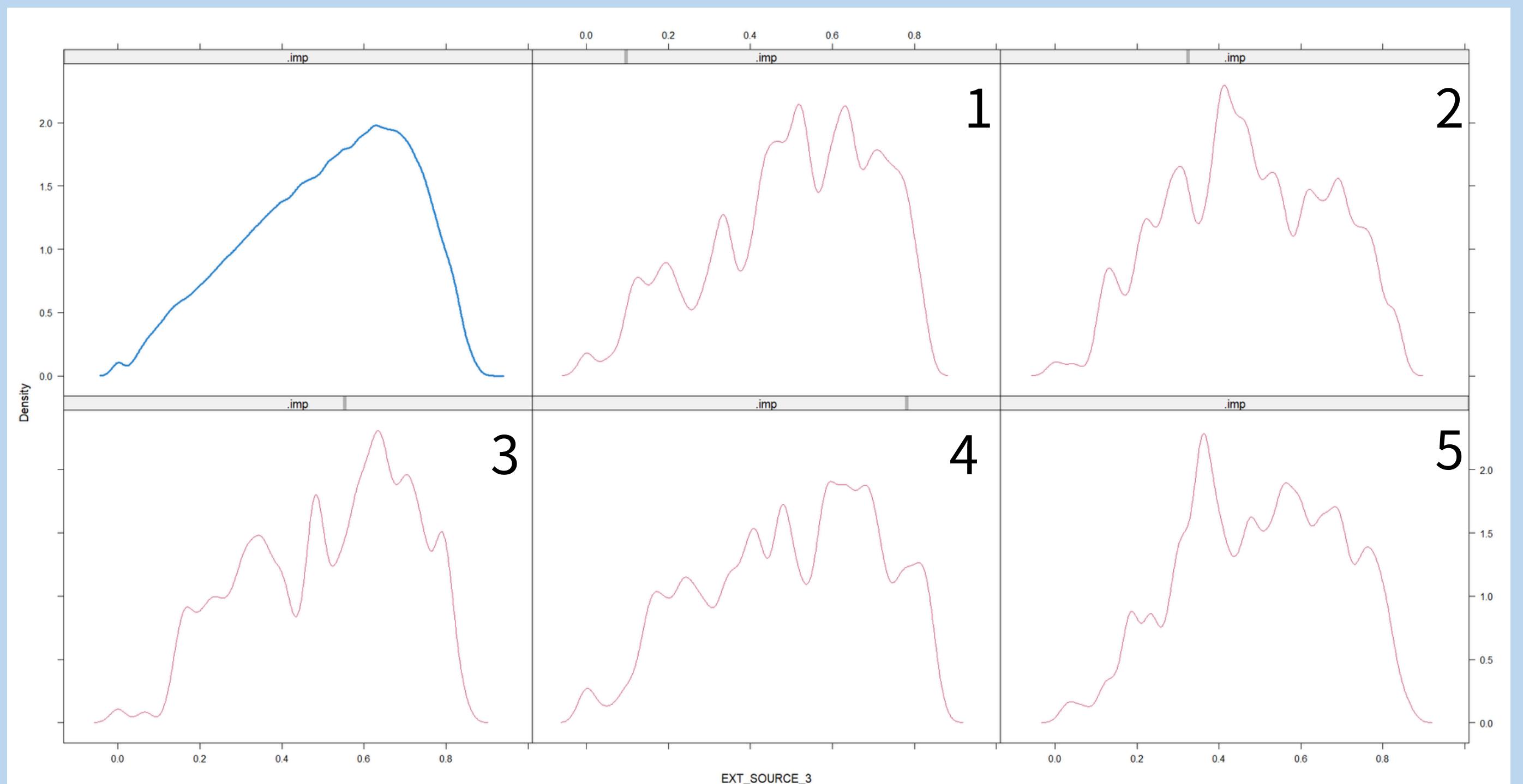
- **YEARS_BIRTH**

(自行生成 10% 遺失值)

兩個變數PMM插補密度分配圖：

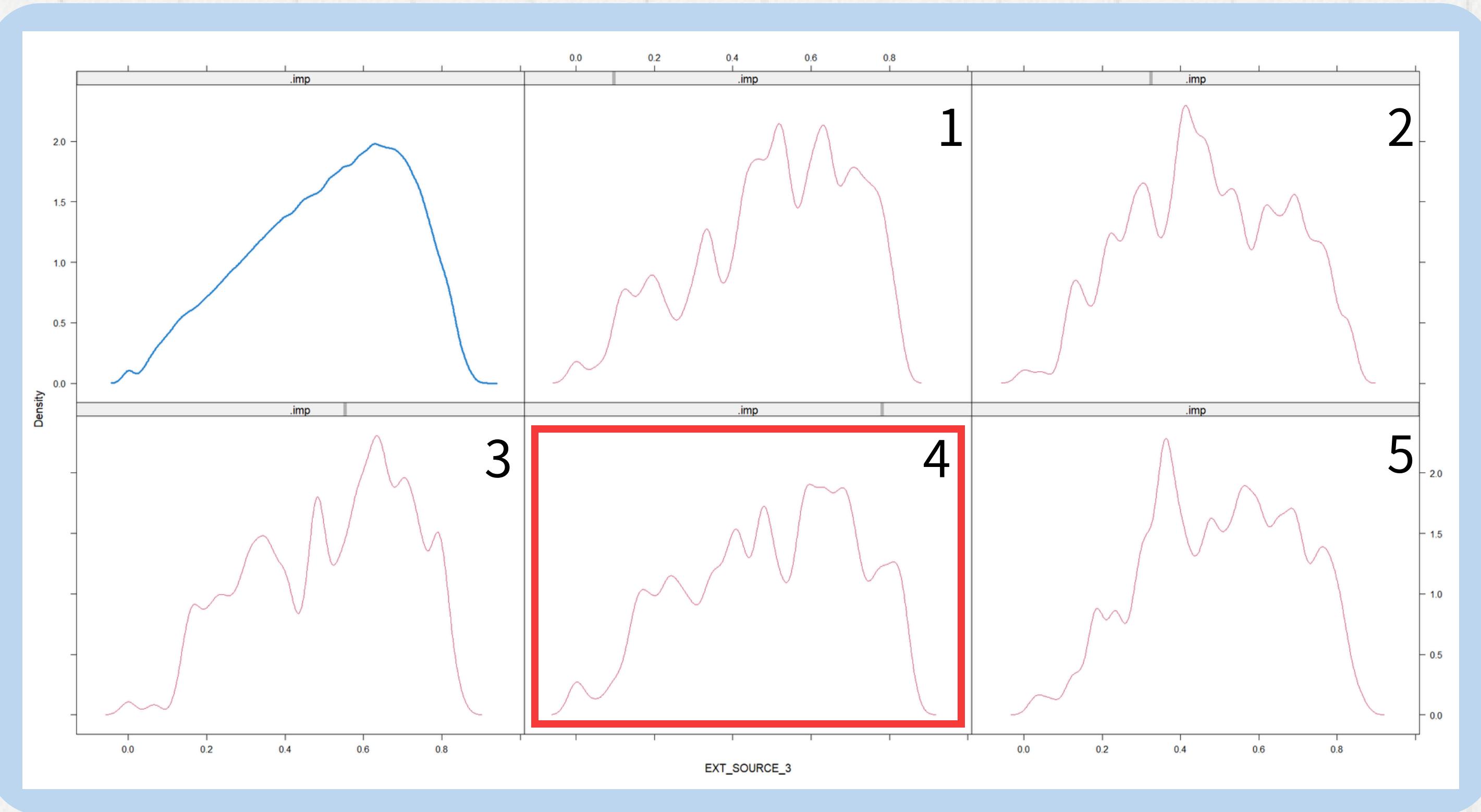


EXT-3的PMM插補密度分配圖：



EXT-3的PMM插補密度分配圖：

第四次迭代的表現較為出色



H₀ :The ith distribution is not significantly different from the original distribution
 $i = 1, 2, 3, 4, 5.$

H₁ :The ith distribution is significantly different from the original distribution
 $i = 1, 2, 3, 4, 5.$

D 統計量:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

統計量 第 i 次 迭代	1	2	3	4	5
D	0.0219	0.0275	0.0277	0.0126	0.0265

H₀ :The ith distribution is not significantly different from the original distribution
 $i = 1, 2, 3, 4, 5.$

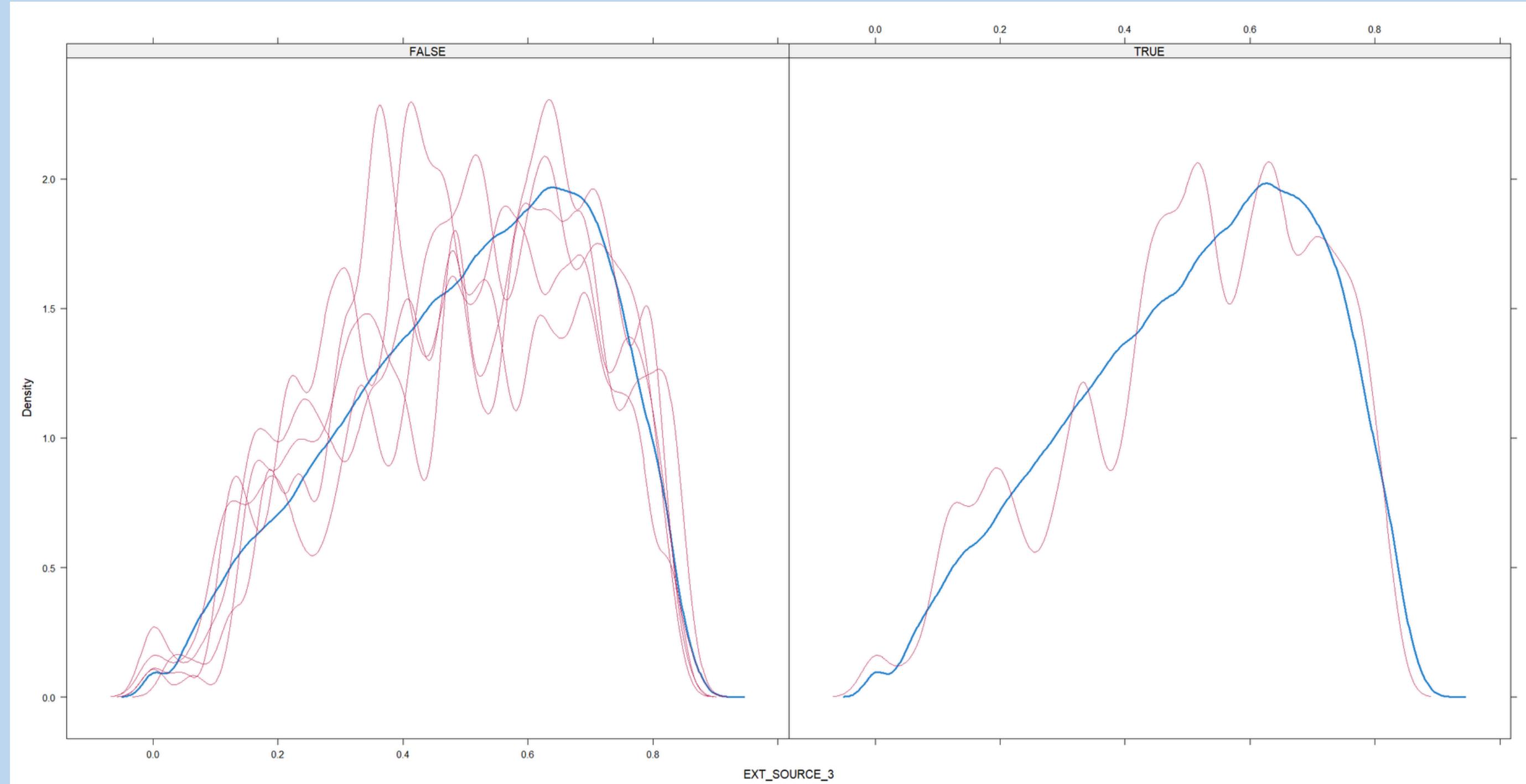
H₁ :The ith distribution is significantly different from the original distribution
 $i = 1, 2, 3, 4, 5.$

D 統計量:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

統計量 第 i 次 迭代	1	2	3	4	5
D	0.0219	0.0275	0.0277	0.0126	0.0265

EXT_SOURCE_3 的 PMM 最終插補密度分配圖：



05. 不平衡處理



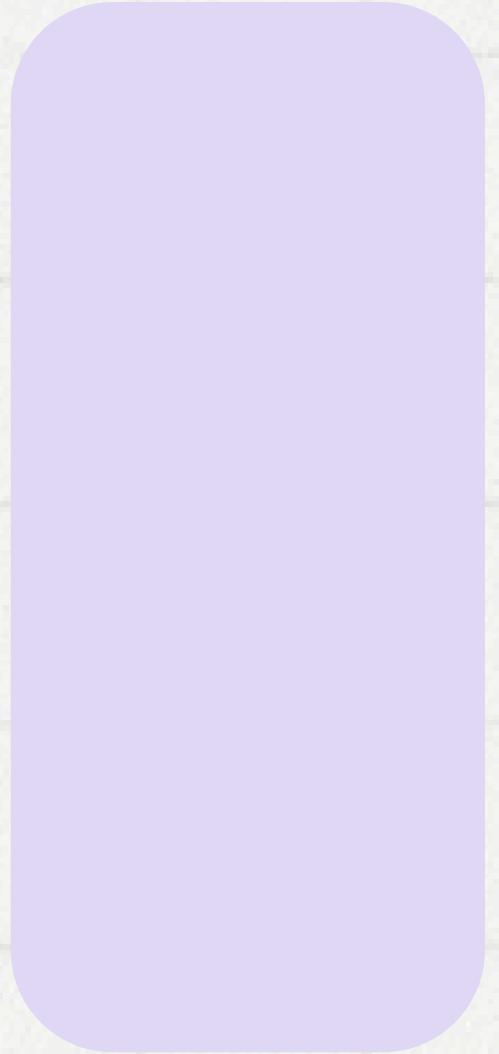
常見處理方式

名稱	作法說明
Undersampling	刪除多數類別的資料

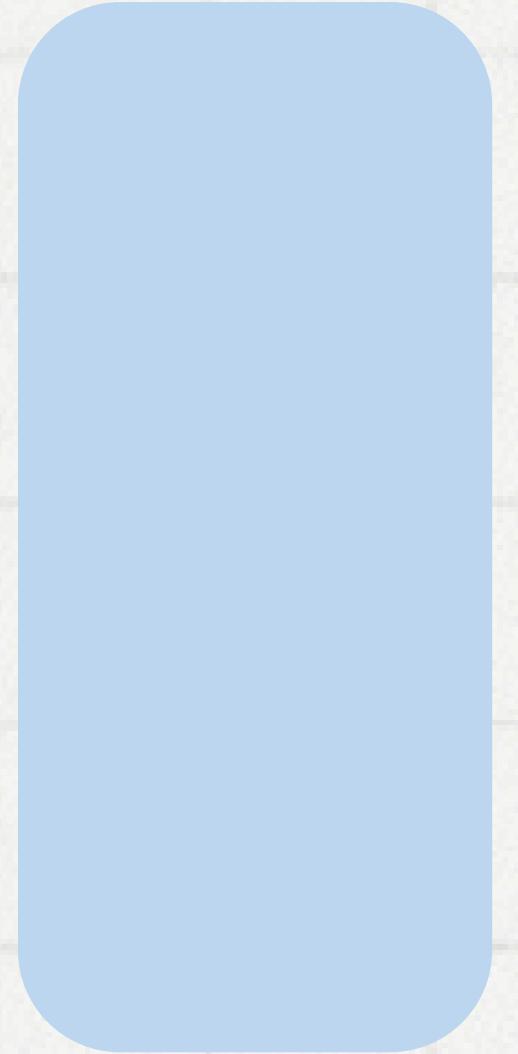
常見處理方式

名稱	作法說明
Undersampling	刪除多數類別的資料
Oversampling	增加少數類別的資料

Raw Data



Over-Sampling



Under-Sampling



Raw Data

Over-Sampling

Under-Sampling

K-Fold Validation
切分四層

Raw Data

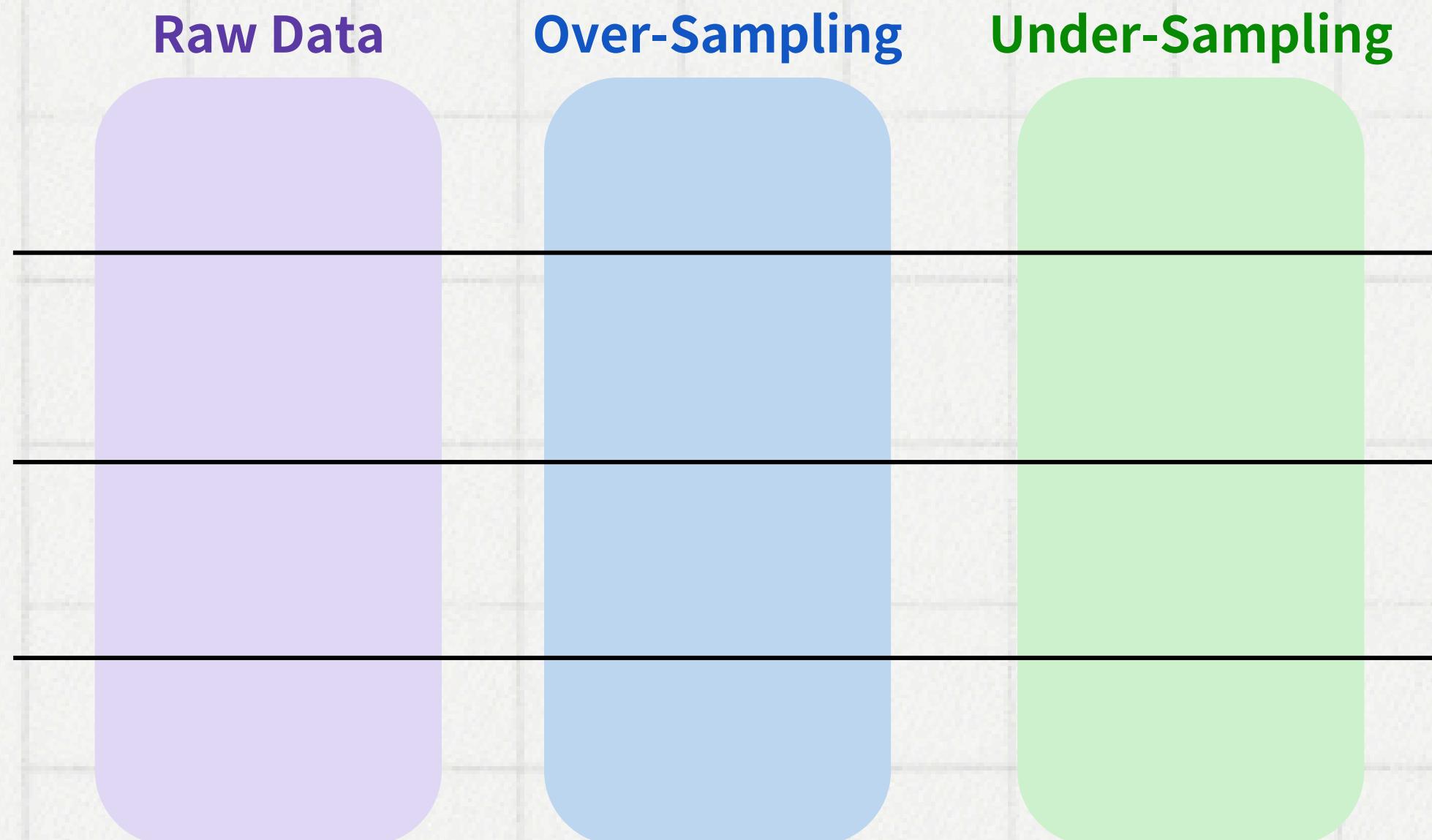
Over-Sampling

Under-Sampling

去除不平衡變數再執行

K-Fold Validation
切分四層

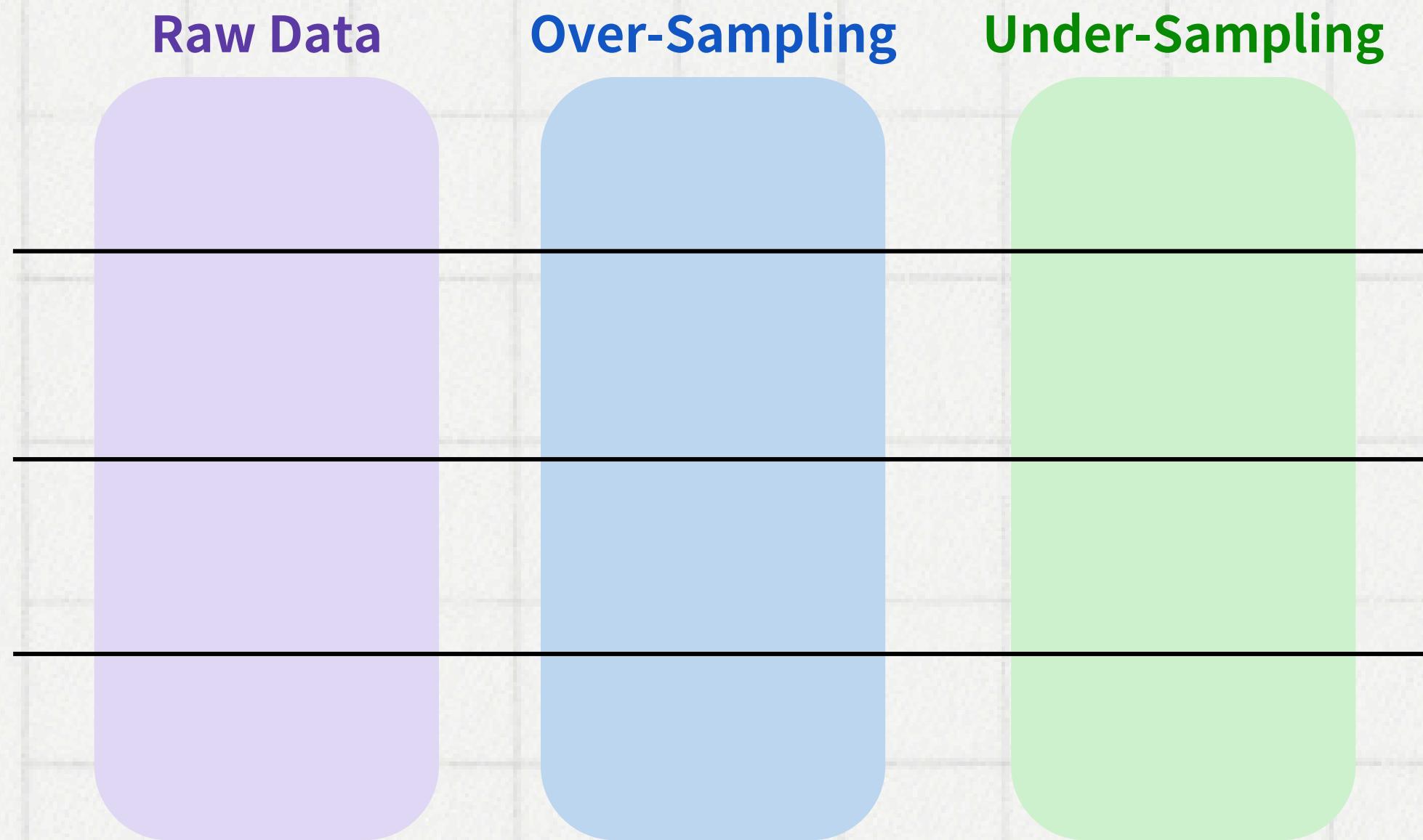




去除不平衡變數再執行

K-Fold Validation
切分四層

- 羅吉斯迴歸 (Logistic Regression)
- 決策樹 (Decision Tree)
- 隨機森林 (Random Forest)



去除不平衡變數再執行

K-Fold Validation
切分四層

- 羅吉斯迴歸 (Logistic Regression)
- 決策樹 (Decision Tree)
- 隨機森林 (Random Forest)

AUC

K-Fold Validation 之 AUC 值：

資料集 \ 模型方法	Logistic Regression	Decision Tree	Random Forest
Raw Data	0.73118	0.5	0.708897

K-Fold Validation 之 AUC 值：

資料集 \ 模型方法	Logistic Regression	Decision Tree	Random Forest
Raw Data	0.73118	0.5	0.708897
Undersampling Data	0.730358	0.645275	0.7068887

K-Fold Validation 之 AUC 值：

資料集 \ 模型方法	Logistic Regression	Decision Tree	Random Forest
Raw Data	0.73118	0.5	0.708897
Undersampling Data	0.730358	0.645275	0.7068887
Oversampling Data	0.731453	0.646103	0.7112601

Mutual Information :

- 定義 :

衡量兩個隨機變數之間相依性的指標

- 目的 :

查看變數與目標變數 (TARGET) 之間的相關性

Mutual Information :

- 定義 :

衡量兩個隨機變數之間相依性的指標

- 目的 :

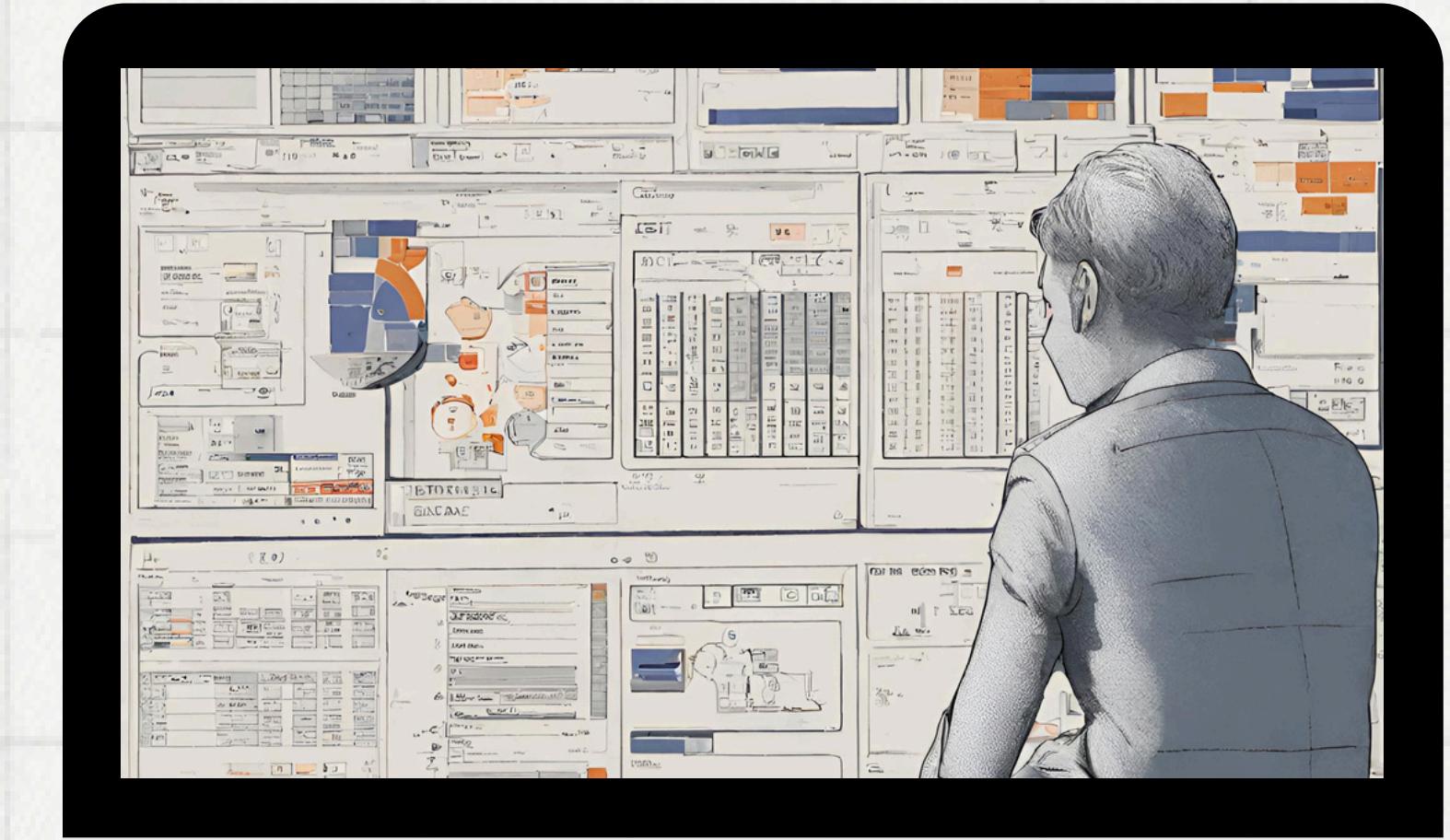
查看變數與目標變數 (TARGET) 之間的相關性

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

交叉驗證前被去除變數與目標變數的 Mutual Information :

變數名稱	Raw Data	Undersampling	Oversampling
CODE_GENDER	0.001455	0.004762	0.004725
NAME_INCOME_TYPE	0.002115	0.007182	0.007373
FLAG_MOBIL	0	0	0.000001
NAME_FAMILY_STATUS	0.000809	0.002415	0.002642
FLAG_DOCUMENT_2	0.000009	0.000020	0.000032
FLAG_DOCUMENT_3	0.001005	0.003616	0.003407
FLAG_DOCUMENT_20	0	0	0
FLAG_DOCUMENT_21	0.000006	0.000033	0.000011

06. 特徵選取



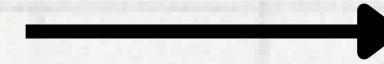
變數篩選：

連續型變數

類別型變數

變數篩選：

連續型變數



Point-Biserial Correlation

類別型變數

變數篩選：

連續型變數



Point-Biserial Correlation

類別型變數



Mutual Information

Point-Biserial Correlation :

定義：衡量一個二元類別變數和一個連續變數之關係

對象：類別型變數 (TARGET) VS 數值型變數

Point-Biserial Correlation :

定義：衡量一個二元類別變數和一個連續變數之關係

對象：類別型變數 (TARGET) VS 數值型變數

公式計算Point-Biserial Correlation

$$r_{pb} = \frac{M_1 - M_0}{s_y} \sqrt{\frac{N_0 N_1}{N(N-1)}}$$

類別型變數 (TARGET) VS 數值型變數：

Pearson's Product-Moment
Correlation Test

set alpha = 0.05

H₀:

True correlation is equal to 0

H₁:

True correlation is not equal to 0

類別型變數 (TARGET) VS 數值型變數：

Pearson's Product-Moment Correlation Test

`set alpha = 0.05`

$H_0:$

True correlation is equal to 0

$H_1:$

True correlation is not equal to 0

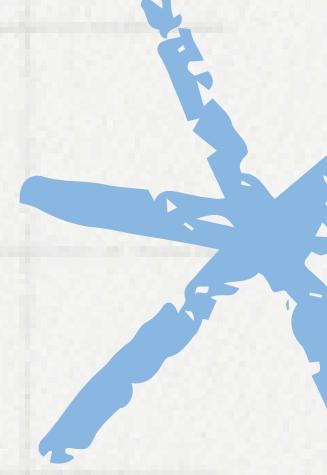
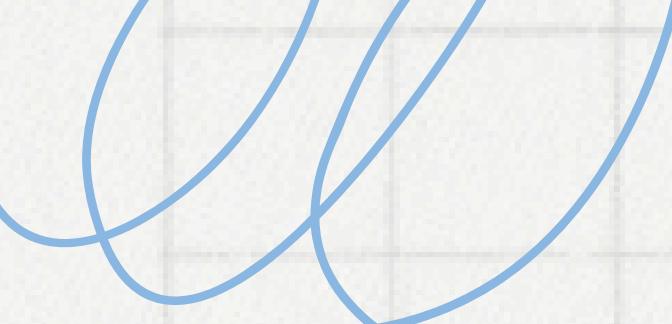
數值型變數	P-value	Correlation	名次
EXT_SOURCE_2	0	-0.268909	1
EXT_SOURCE_3	0	-0.242487	2
YEARS_BIRTH	0	-0.137025	3
YEARS_LAST_PHONE_CHANGE	0	-0.102210	4
YEARS_ID_PUBLISH	0	-0.093831	5
YEARS_EMPLOYED	0	-0.093730	6
YEARS_REGISTRATION	0	-0.079577	7
AMT_GOODS_PRICE	0	-0.077160	8
REGION_POPULATION_RELATIVE	0	-0.071720	9
AMT_CREDIT	0	-0.058562	10

類別型變數 (TARGET) VS 類別型變數 :

類別型變數	$I(X; Y)$	名次
OCCUPATION_TYPE	0.010219	1
ORGANIZATION-TYPE	0.008290	2
NAME_INCOME_TYPE	0.007373	3
REGION_RATING_CLIENT_W_CITY	0.006174	4
NAME_EDUCATION_TYPE	0.006086	5
REGION_RATING_CLIENT	0.005799	6
CODE_GENDER	0.004725	7
FLAG_EMP_PHONE	0.003968	8
REG_CITY_NOT_WORK_CITY	0.003497	9
FLAG_DOCUMENT_3	0.002722	10

最終選取變數：

數值型變數	類別型變數
EXT_SOURCE_2	OCCUPATION_TYPE
EXT_SOURCE_3	ORGANIZATION-TYPE
YEARS_BIRTH	NAME_INCOME_TYPE
YEARS_LAST_PHONE_CHANGE	REGION_RATING_CLIENT_W_CITY
YEARS_ID_PUBLISH	NAME_EDUCATION_TYPE



Thank you

