

顧客行為分析

專案01

目錄

- 資料集概要

- 1-1 資料選擇

- 1-2 特徵涵義

- 1-3 資料觀察

- 1-4 鎖定分析方向

- 分析方法

- 3-1 PCA降維處理

- 3-2 KMeans模型

- 資料預處理

- 2-1 填補缺失值

- 2-2 創建特徵

- 2-3 類別型特徵整理

- 2-4 偏離值處理

- 2-5 特徵圖表說明

- 2-6 相關性

- 結果呈現

- 4-1 說明模型後續調整方向

- 4-2 總結

- 4-3 專案相關資料

資料集概要

1-1 資料選擇

1-2 特徵涵義

1-3 資料觀察

1-4 鎖定分析方向

為什麼選擇此 資料集

用意是建立顧客行為相關的分析專案，以證明有相關的分析經驗能參考。

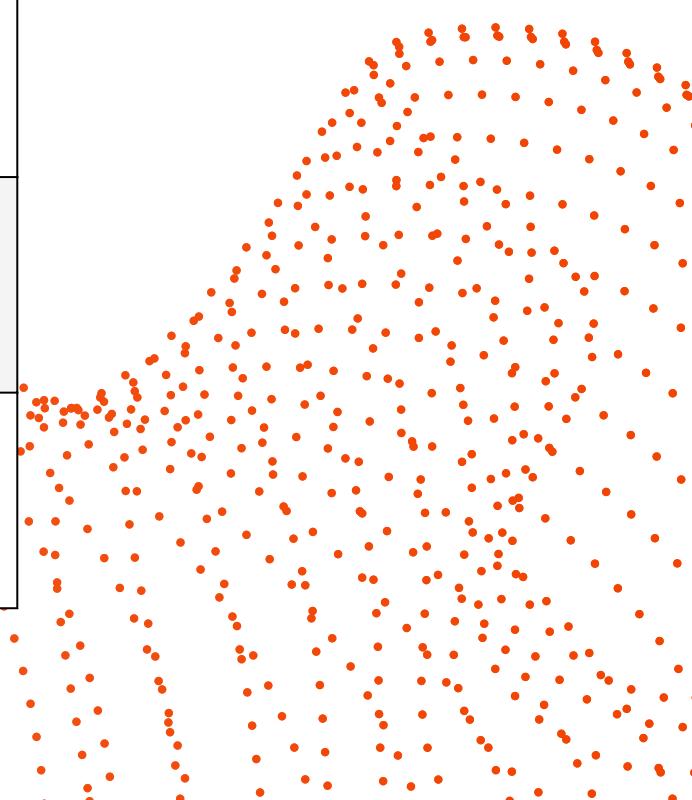
資料集來源

這副資料集是取自Kaggle平台上的資料。

1-2特徵涵義

- ID：顧客編號(數值)
- Year_Birth：顧客生日(數值)
- Education：顧客教育程度(類別)
- Marital_Status：顧客配偶(類別)
- Income：顧客配偶(類別)
- Kidhome：家中有小孩為1，反之為0(類別)
- Teenhome：家中有青少年為1，反之為0(類別)
- Dt_Customer：會員註冊日期(日期)
- Recency：累計未消費天數(數值)
- Complain：顧客兩年內有抱怨為1，反之為0(類別)

特徵	ID	Year_Birth	Education	Marital_Status	Income
範例	9360	1967	Graduation	Married	59354
特徵	Kidhome	Teenhome	Dt_Customer	Recency	Complain
範例	1	0	15-11-2013	23	1



1-2特徵涵義

- MntWines : 顧客在酒類上的消費額(數值)
- MntFruits : 顧客在水果類上的消費額(數值)
- MntMeatProducts : 顧客在肉類上的消費額(數值)
- MntFishProducts : 顧客在魚類上的消費額(數值)
- MntSweetProducts : 顧客在甜品類上的消費額(數值)
- MntGoldProds : 顧客在精品類上的消費額(數值)
- NumDealsPurchases : 顧客在結帳時打折的次數(數值)
- AcceptedCmp1 : 活動1中是否接受了優惠有為1，反之為0
- AcceptedCmp2 : 活動2中是否接受了優惠有為1，反之為0
- AcceptedCmp3 : 活動3中是否接受了優惠有為1，反之為0

特徵	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts
範例	520	42	118	24	546
特徵	MntGoldProds	NumDealsPurchases	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3
範例	236	5	1	0	1

1-2特徵涵義

- AcceptedCmp4 : 活動4中是否接受了優惠有為1，反之為0
- AcceptedCmp5 : 活動5中是否接受了優惠有為1，反之為0
- Response : 最新活動中是否接受了優惠有為1，反之為0
- NumWebPurchases : 透過網站購買次數(數值)
- NumCatalogPurchases: 透過書刊購買次數(數值)
- NumStorePurchases : 透過商店購買次數(數值)
- NumWebVisitsMonth : 顧客上個月瀏覽網站次數(數值)
- Z_CostContact : 沒有解釋含意
- Z_Revenue : 沒有解釋含意

特徵	AcceptedCmp4	AcceptedCmp5	Response	NumWebPurchases	NumCatalogPurchases
範例	1	0	1	20	12
特徵	NumStorePurchases	NumWebVisitsMonth	Z_CostContact	Z_Revenue	
範例	6	32	3	11	

1-3資料觀察

- Z_CostContact及Z_Revenue找不到意義，且所有值都相同，所以不需要。
- Income有少量缺失值。
- Education要標籤化。
- Marital_Status要標籤化。
- ID可以拿掉。
- Year_Birth要改成長年紀。
- Dt_Customer可做成註冊多久。
- 新增商品總花費。
- 用Teenhome,Kidhome新增是否為家庭。

Z_CostContact	1
Z_Revenue	1

Marital_Status	2240 non-null
Income	2216 non-null
Kidhome	2240 non-null

ID	Year_Birth	Education	Marital_Status
5524	1957	Graduation	Single
2174	1954	Graduation	Single
4141	1965	Graduation	Together
6182	1984	Graduation	Together
5324	1981	PhD	Married

	MntWines	MntFruits	MntMeatProducts	MntFishProducts
635	88	546	172	
11	1	6	2	

Dt_Customer	04-09-2012
	08-03-2014
	21-08-2013
	10-02-2014

Kidhome	Teenhome
0	0
1	1
0	0
1	0
1	0

以了解顧客組成主要分析方向

- 因為沒有特徵中無標籤紀錄顧客的組成，所以打算透過分群的方式，來了解那些特徵會明顯影響顧客組成。
- 透過了解顧客接受活動優惠的比例，降低不必要的顧客行銷成本。
- 了解顧客購買的產品之間是否有相關性。
- 了解顧客的主要消費管道。

資料預處理

2-1 填補缺失值

2-2 創建特徵

2-3 類別型特徵整理

2-4 偏離值處理

2-5 特徵圖表說明

2-6 相關性

2-1 填補缺失值

- 因為不想將缺值移除，所以採取補入Income中位數的方法。

count	2216.000000
mean	52247.251354
std	25173.076661
min	1730.000000
25%	35303.000000
50%	51381.500000
75%	68522.000000
max	666666.000000
Name:	Income, dtype: float64

2-2創建特徵

- 透過當天日期與會員註冊日期的相減，創建用於表示顧客累計創建天數的新特徵，並取名為 Be_Customer。
- 先把 Kidhome 和 Teenhome 加起來，然後創建一個叫做 Children的新欄位用於紀錄未成年著的數量。
- 再根據 Children的值來判斷是否為父母，並創建另一個叫做 Parent的類別欄位是父母則表示為1，反之0。
- 創建Spent特徵用於代表購買種類的加總
- 創建Age特徵用於代表顧客的年紀

	Be_Customer	Spent	Children	Parent	Age
4014		1617	0	0	65
3168		27	2	1	68
3515		776	0	0	57
3108		53	1	1	38
3364		422	1	1	41
...	
3584		1341	1	1	55
3104		444	3	1	76
3358		1241	0	0	41
3359		843	1	1	66
3825		172	2	1	68

2-3類別型特徵整理

- 由於模型對於文字型態的處理不佳，所以通常回將類別型的特徵轉為數值型態。
- 將Alone, Absurd, YOLO, Widow, Divorced 納入Single，並將Married 納入Together。
- 0為Together，1為Single。

```
Married      864
Together    580
Single       480
Divorced    232
Widow        77
Alone         3
Absurd        2
YOLO          2
Name: Marital_Status, dtype: int64
```

```
0      1444
1      796
Name: Marital_Status, dtype: int64
```

- 將Education的中的教育程度，轉成有序的數字型態。
- 0為Basic
- 1為2n Cycle
- 2為Graduation
- 3為Master
- 4為PnD

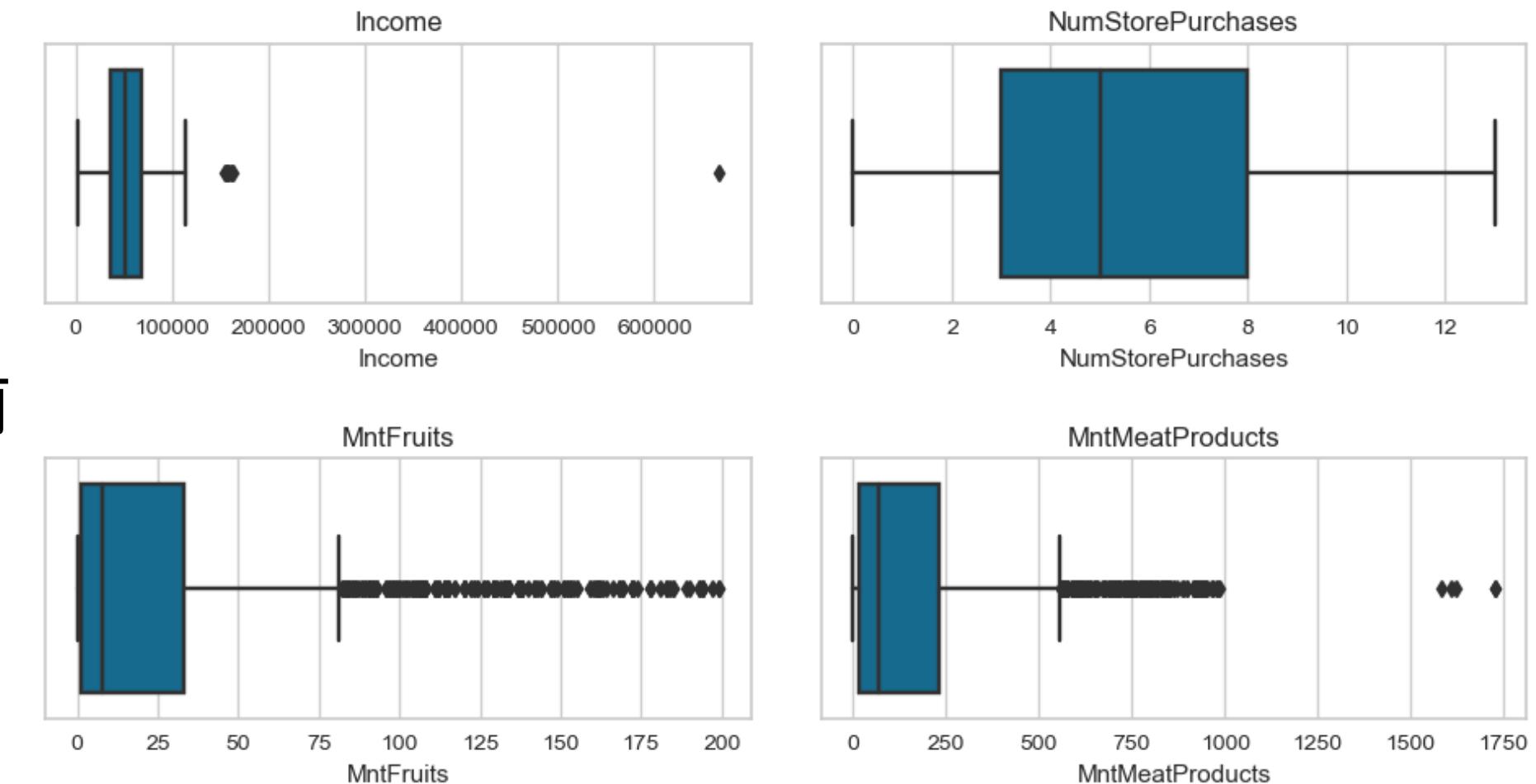
```
Graduation   1127
PhD          486
Master        370
2n Cycle     203
Basic         54
Name: Education, dtype: int64
```

```
2      1127
4      486
3      370
1      203
0      54
Name: Education, dtype: int64
```

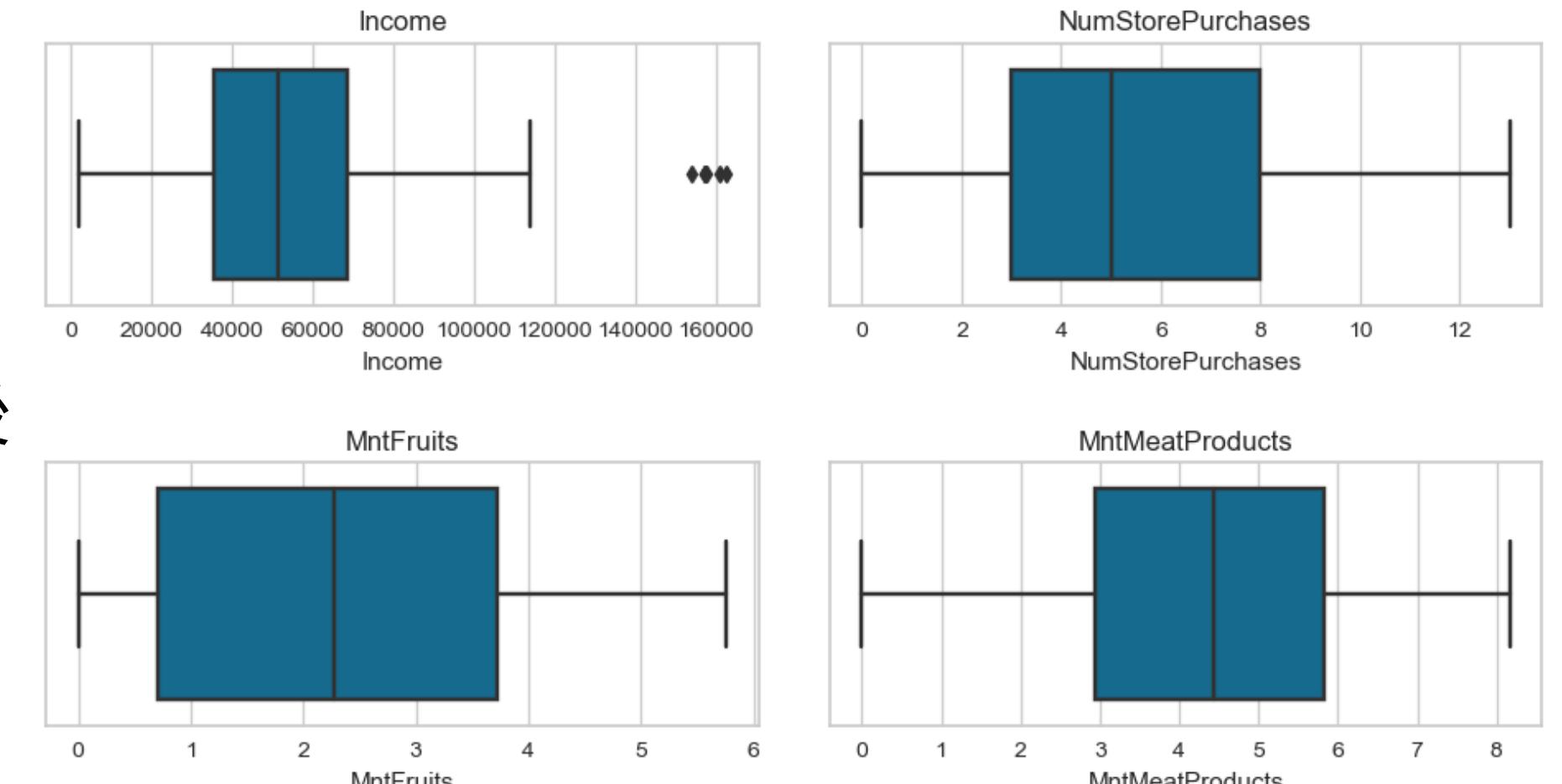
2-4 偏離值處理

- 大多數數值特徵的偏離值會使用 boxcox 這個轉常態分佈的分法處理，特殊處理會特別說明。
- Income 再轉常態分佈前，注意到只有一個值的極端化最明顯，所以我會先將大於 600000 以上的值排除掉。
- 轉化完後我會再將 Income 圖表中四分位數之外的值排除掉，調整後的圖片將在 2-4 最後展示。

調整前



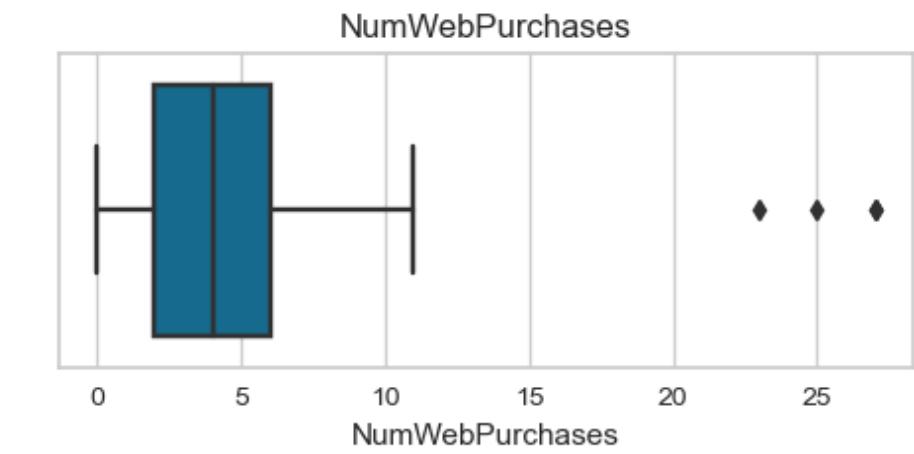
調整後



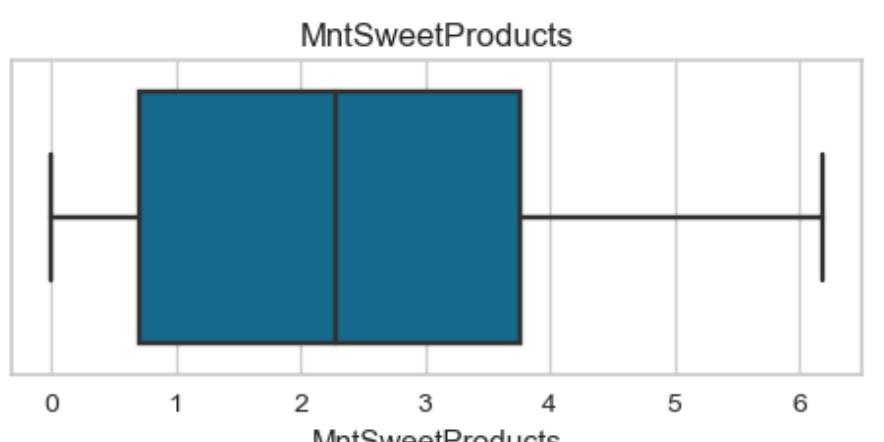
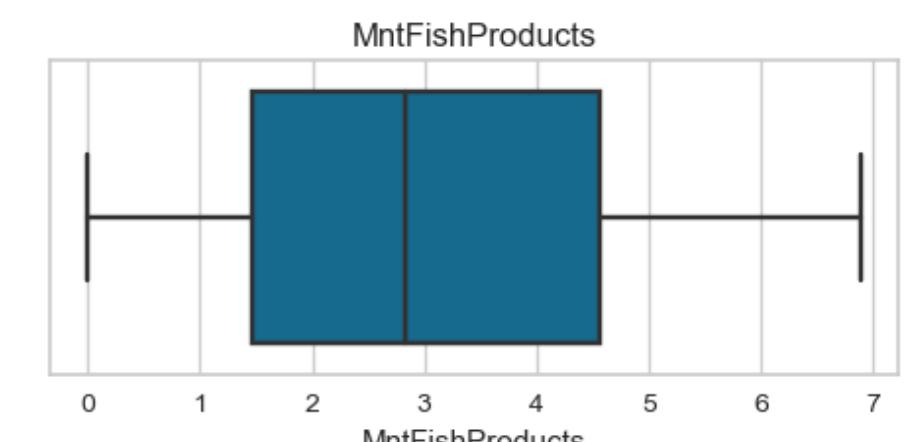
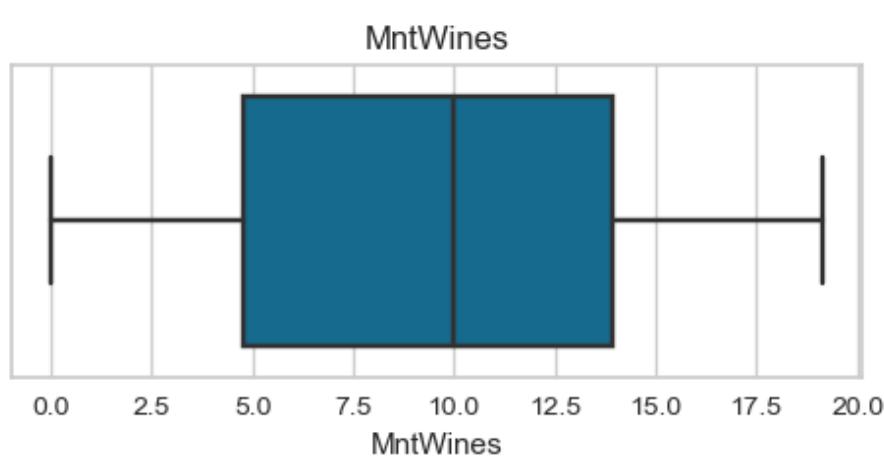
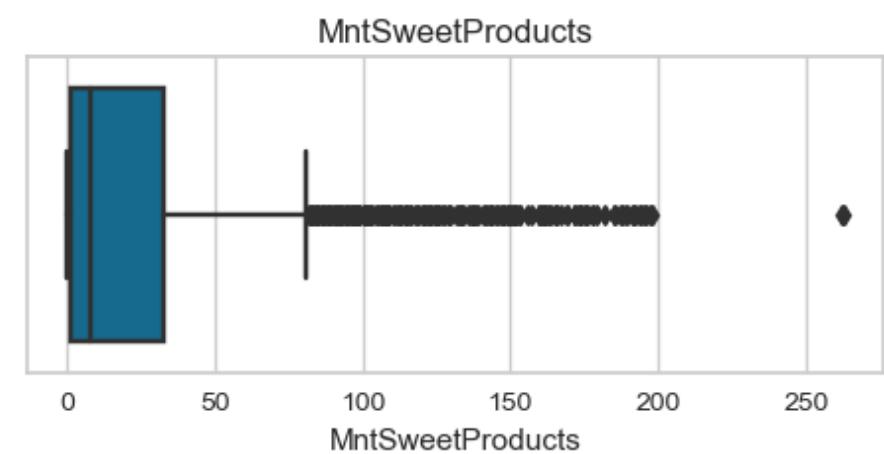
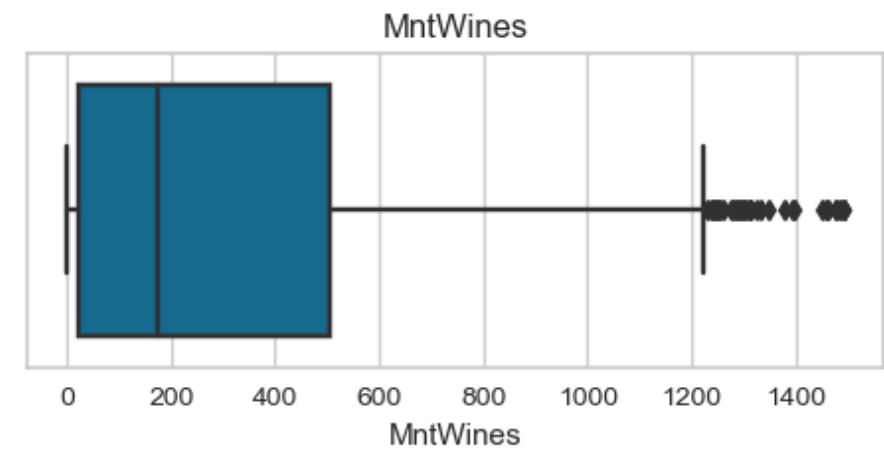
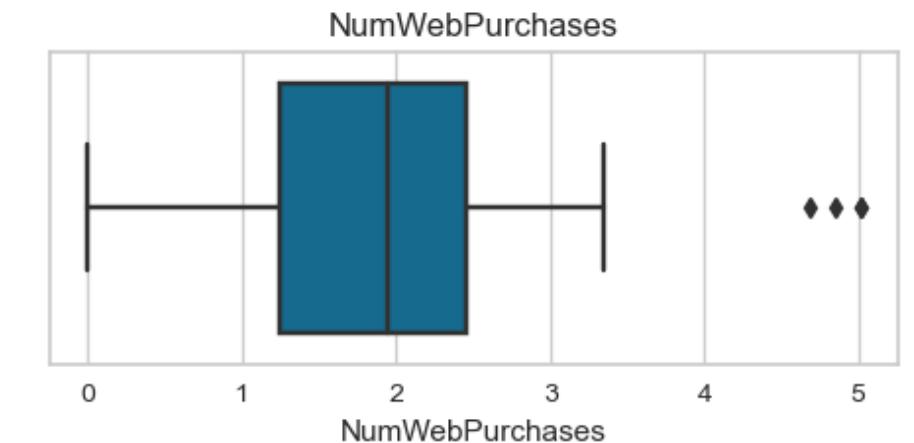
2-4 偏離值處理

- 轉化完後我會再將 NumWebPurchases 圖表中四分位數之外的值排除掉，調整後的圖片將在 2-4 最後展示。

調整前



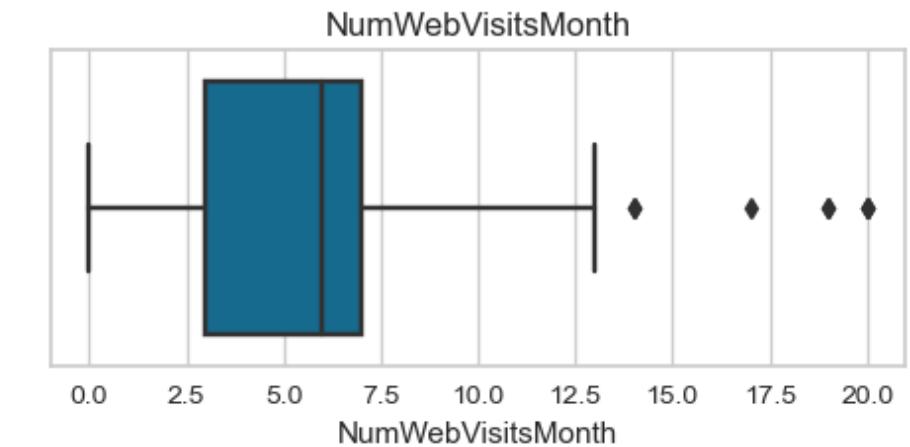
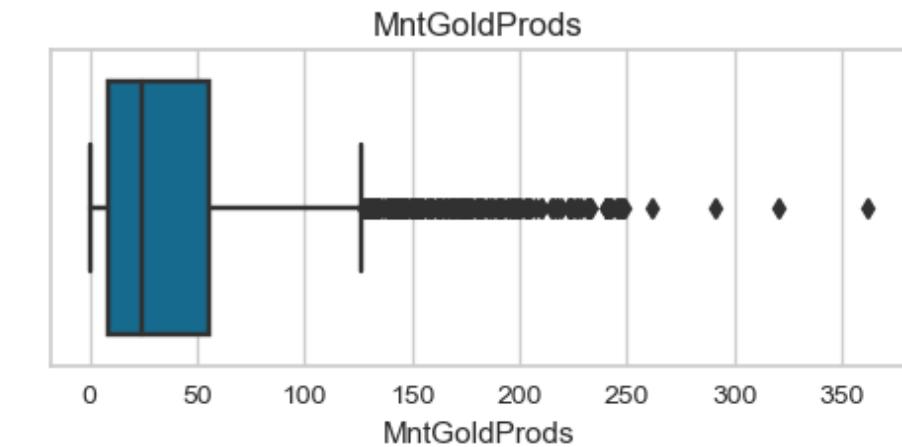
調整後



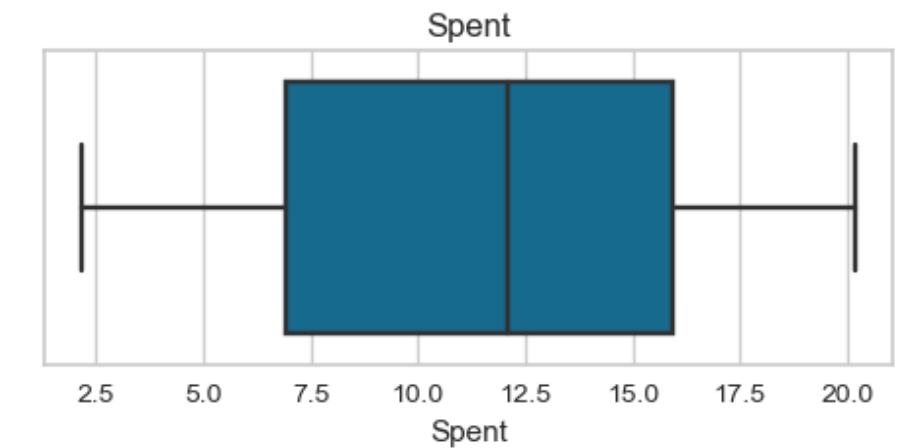
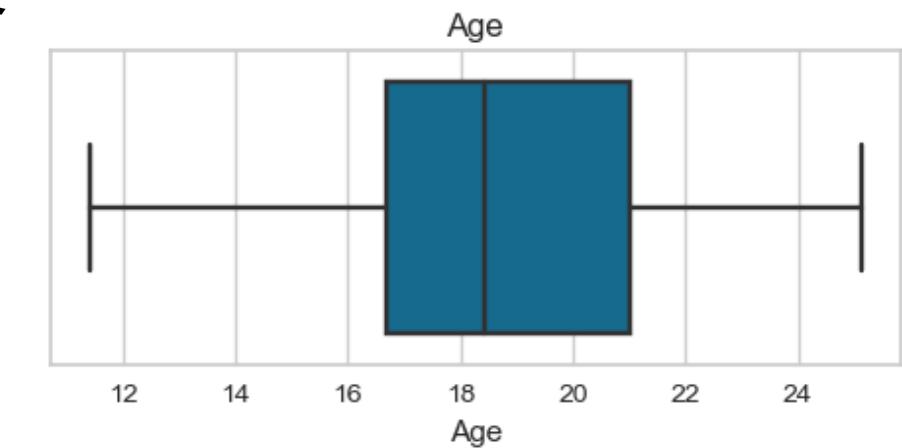
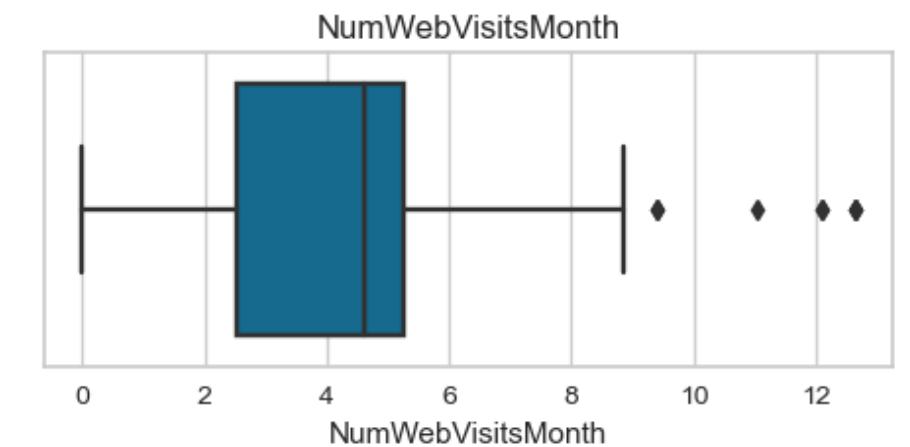
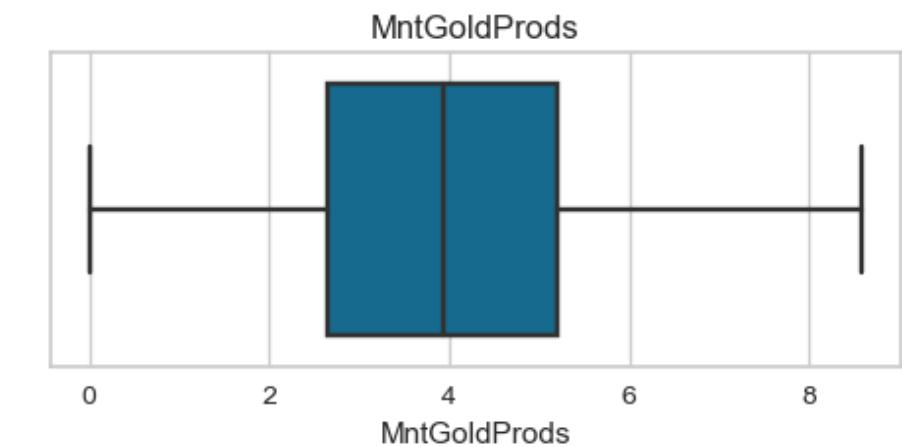
2-4 偏離值處理

- 轉化完後我會再將 NumWebVisitsMonth 圖表中四分位數之外的值排除掉，調整後的圖片將在 2-4 最後展示。

調整前



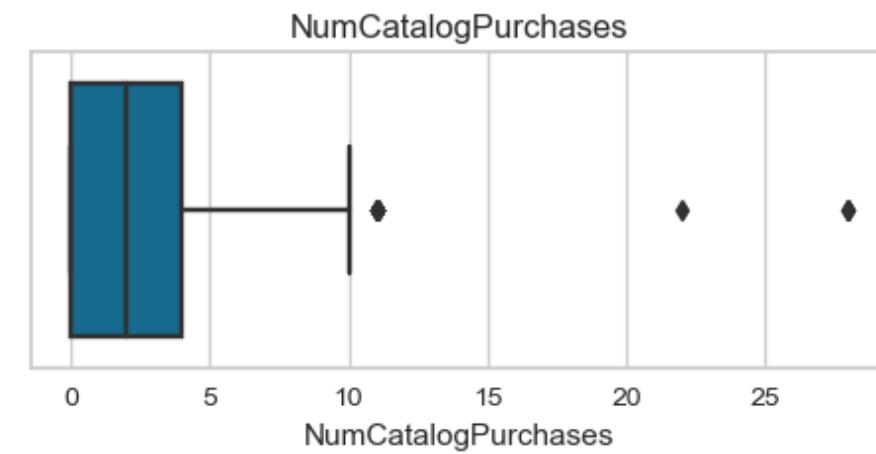
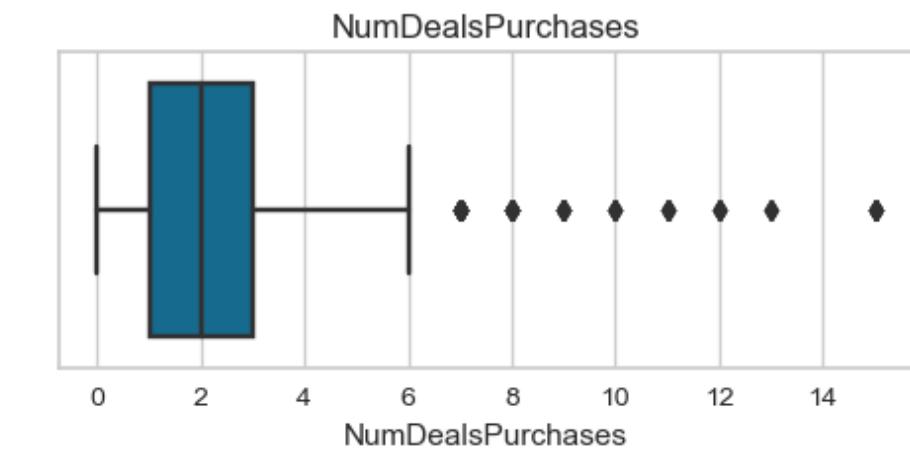
調整後



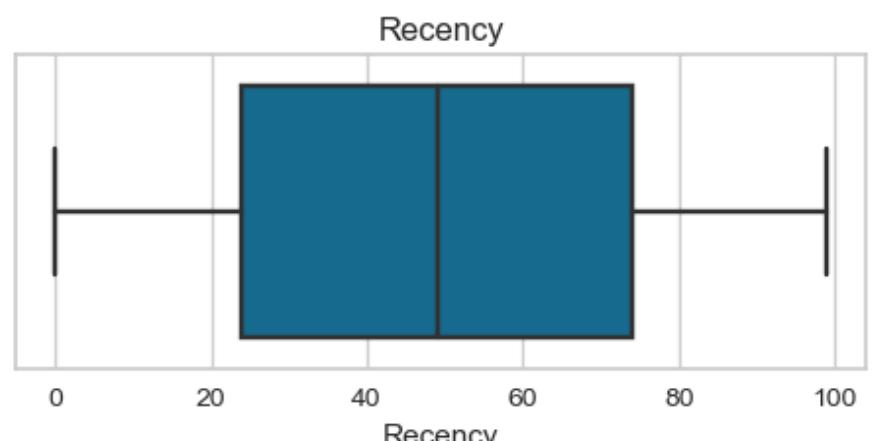
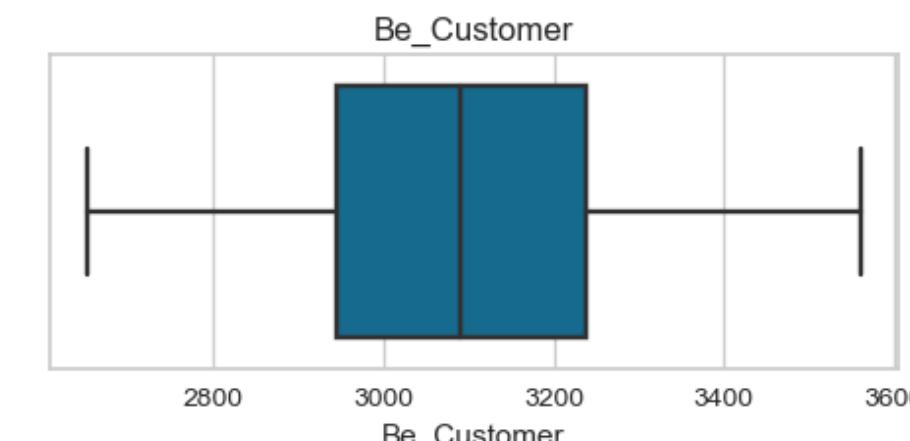
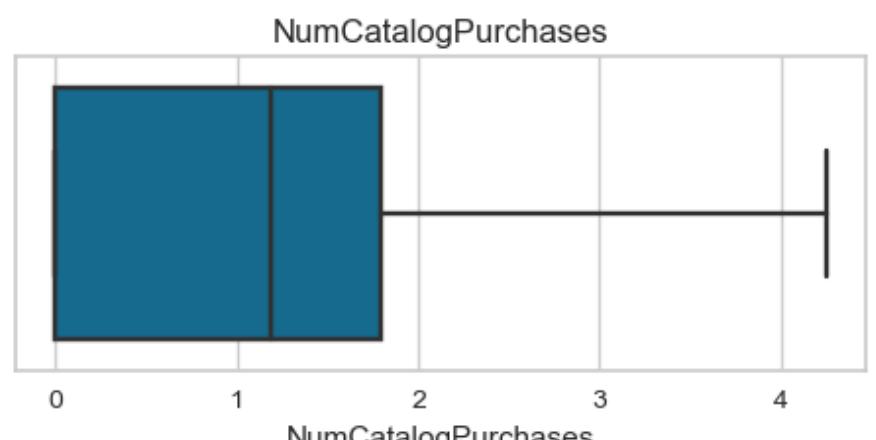
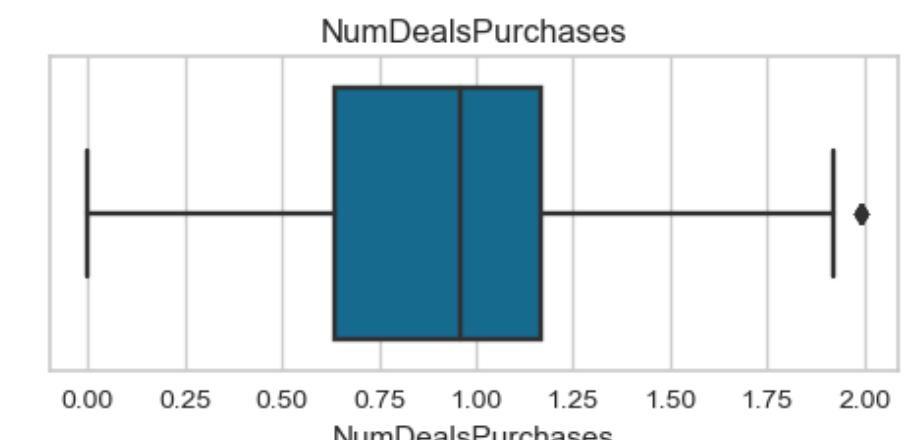
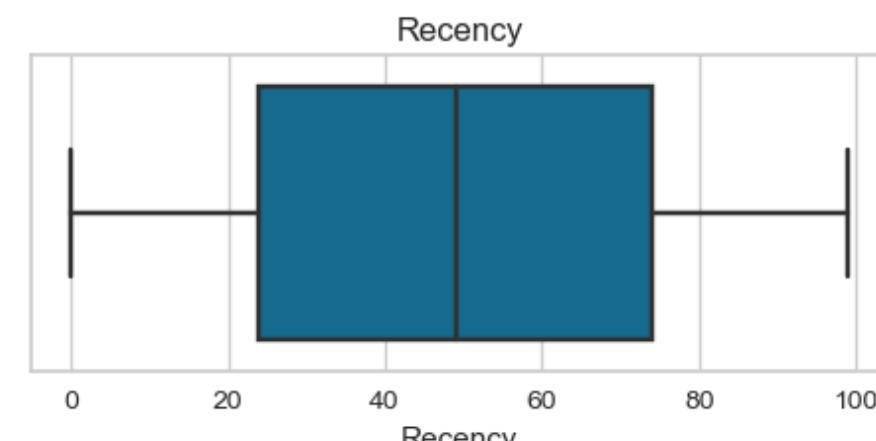
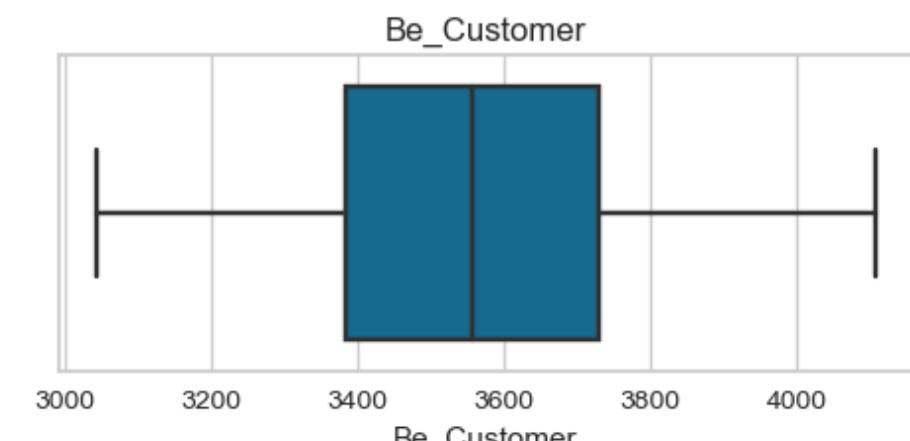
2-4 偏離值處理

- 轉化完後我會再將Be_Customer 圖表中四分位數之外的值排除掉，調整後的圖片將在2-4最後展示。

調整前

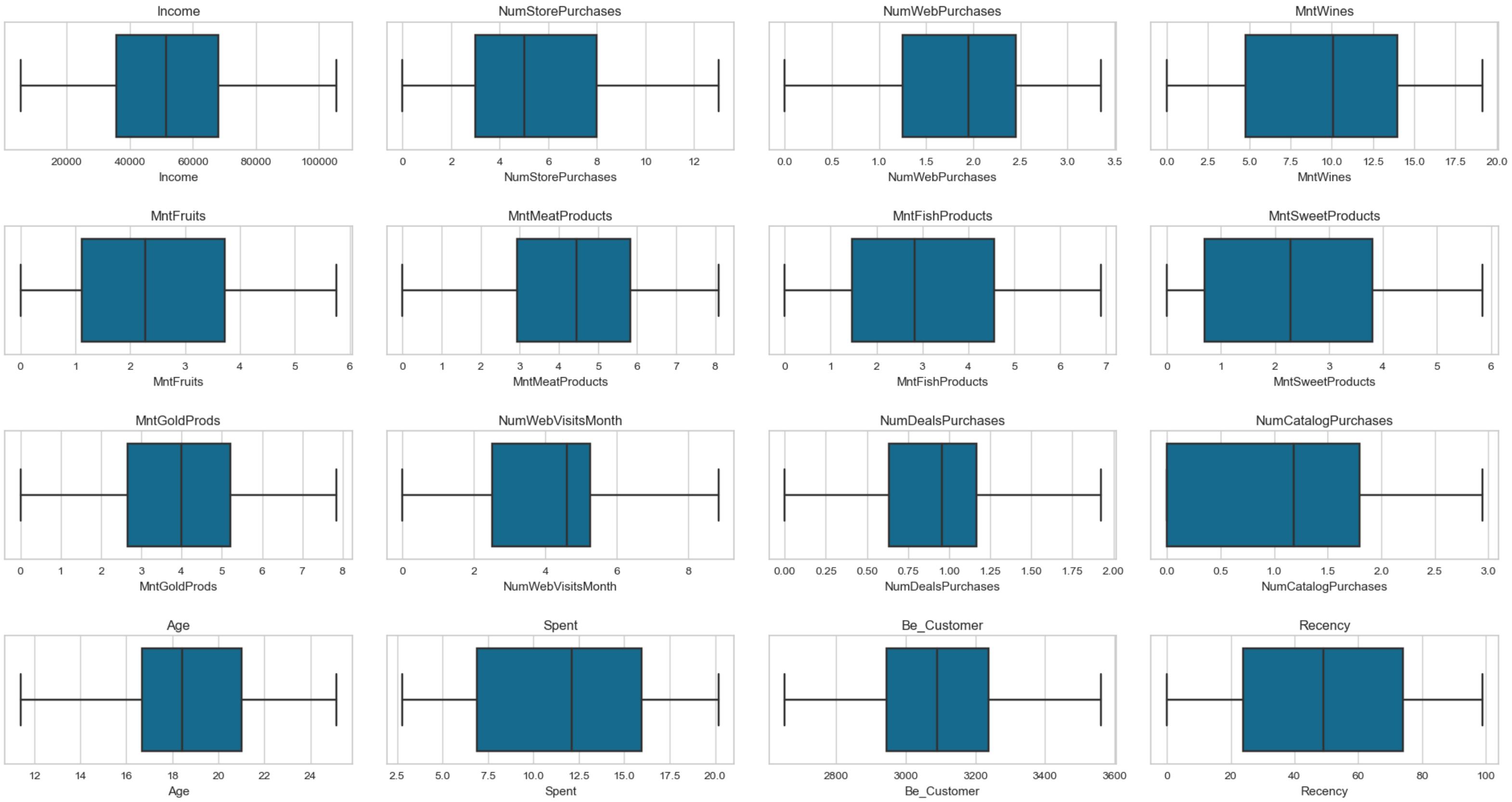


調整後



2-4 偏離值處理

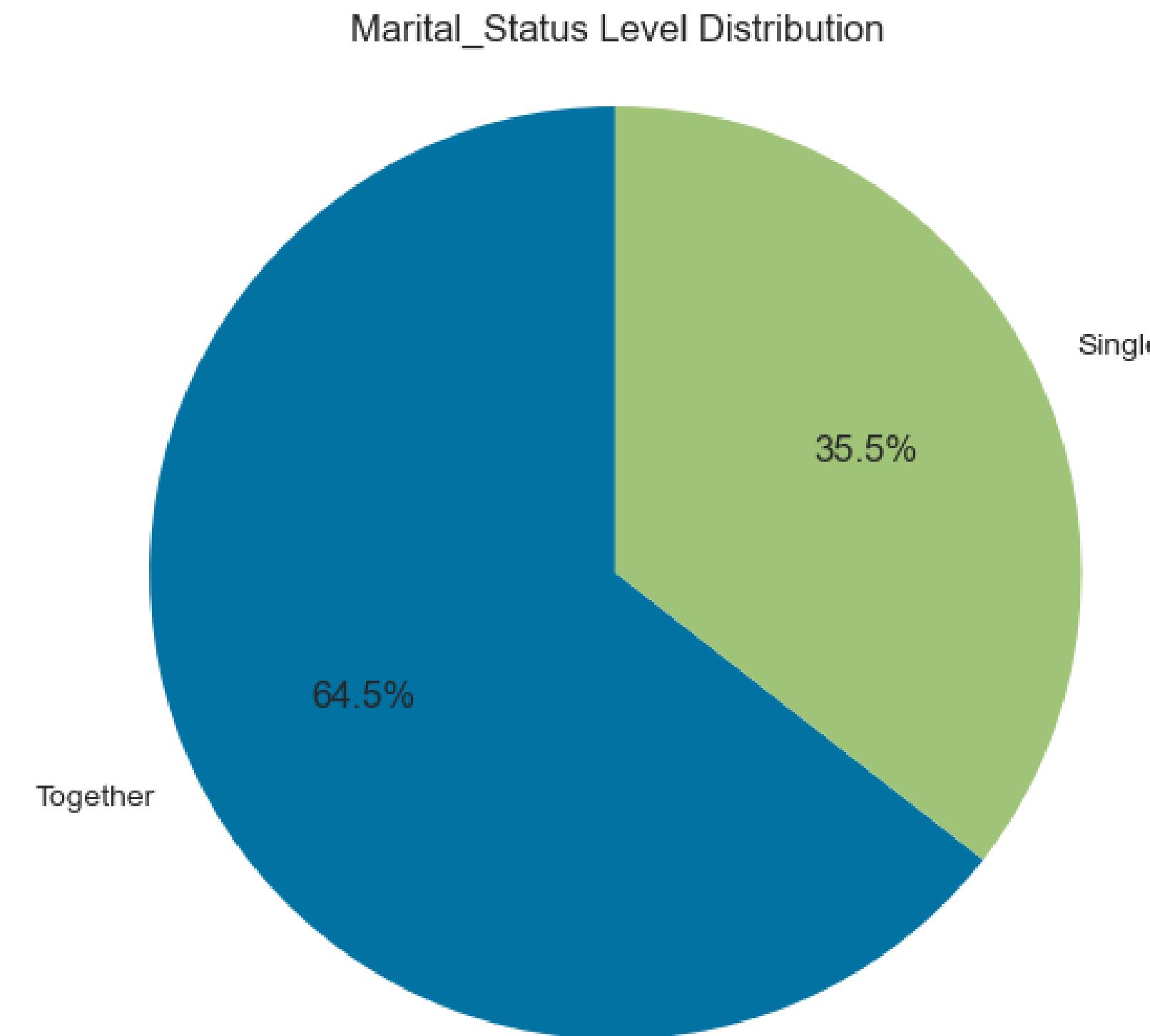
各特徵的四分位距圖



2-5 特徵圖表說明

顧客關係狀態圓餅圖

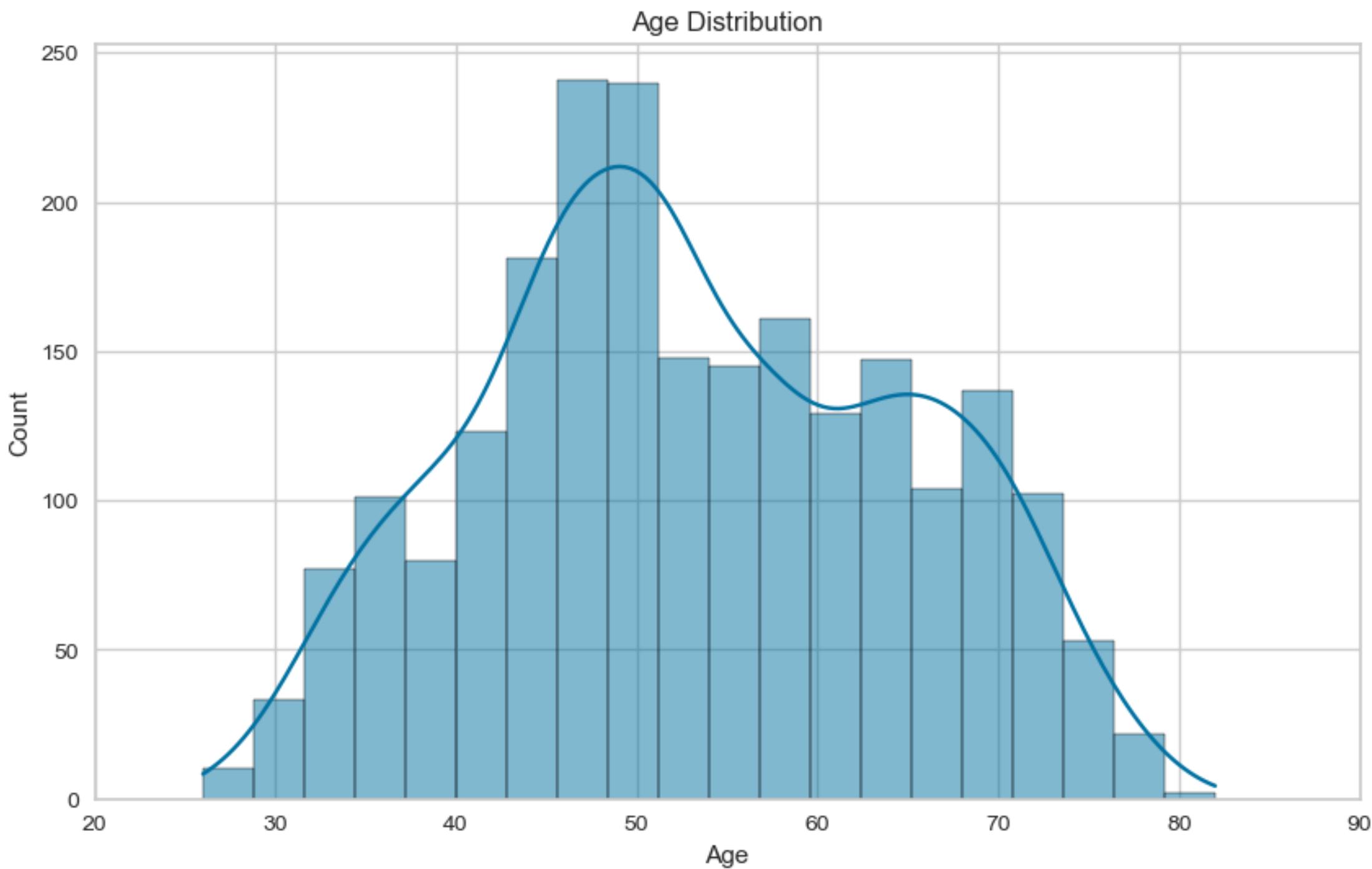
- 有伴侶狀態的佔顧客整體裡的6成。



2-5 特徵圖表說明

- 40~50歲之見的顧客佔多數。
- 60歲以上的族群多於40以下的族群。

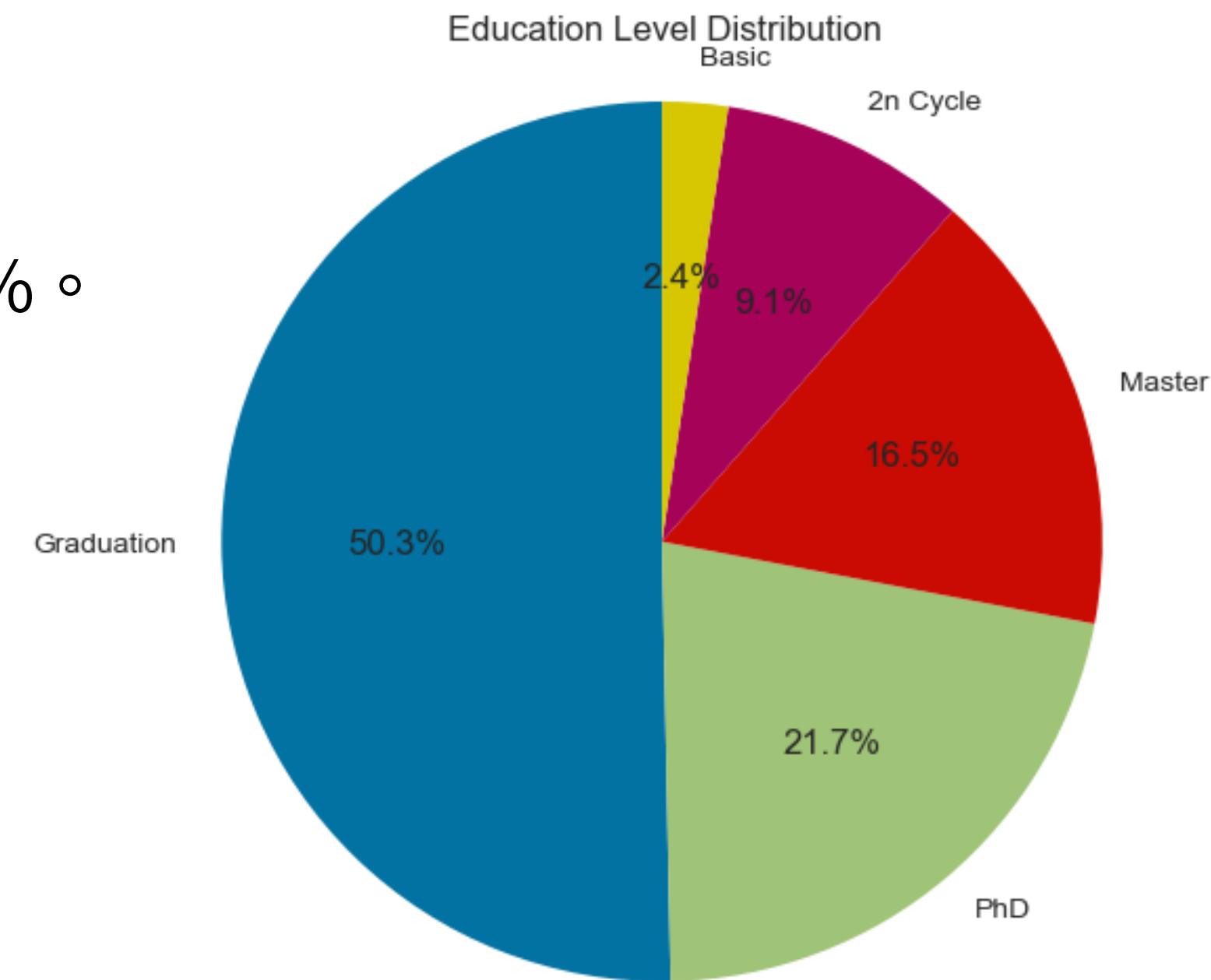
顧客年齡分布直條圖



2-5 特徵圖表說明

顧客教育程度圓餅圖

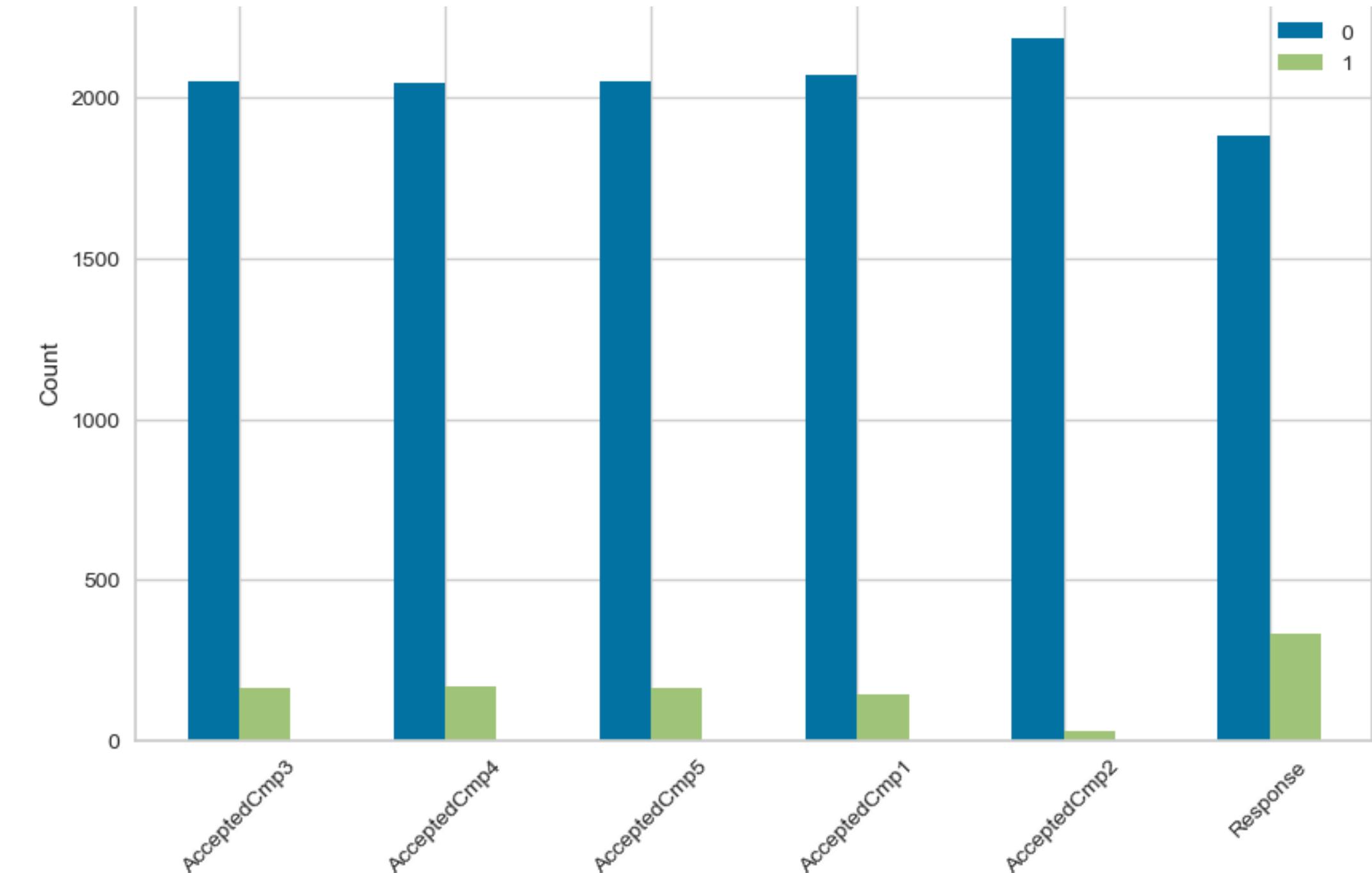
- 學士學歷顧客佔整體的一半。
- 學士學歷以下的顧客佔比約整體的10%。
- 顧客整體擁有高度的教育程度。



2-5 特徵圖表說明

顧客接受活動優惠值條圖

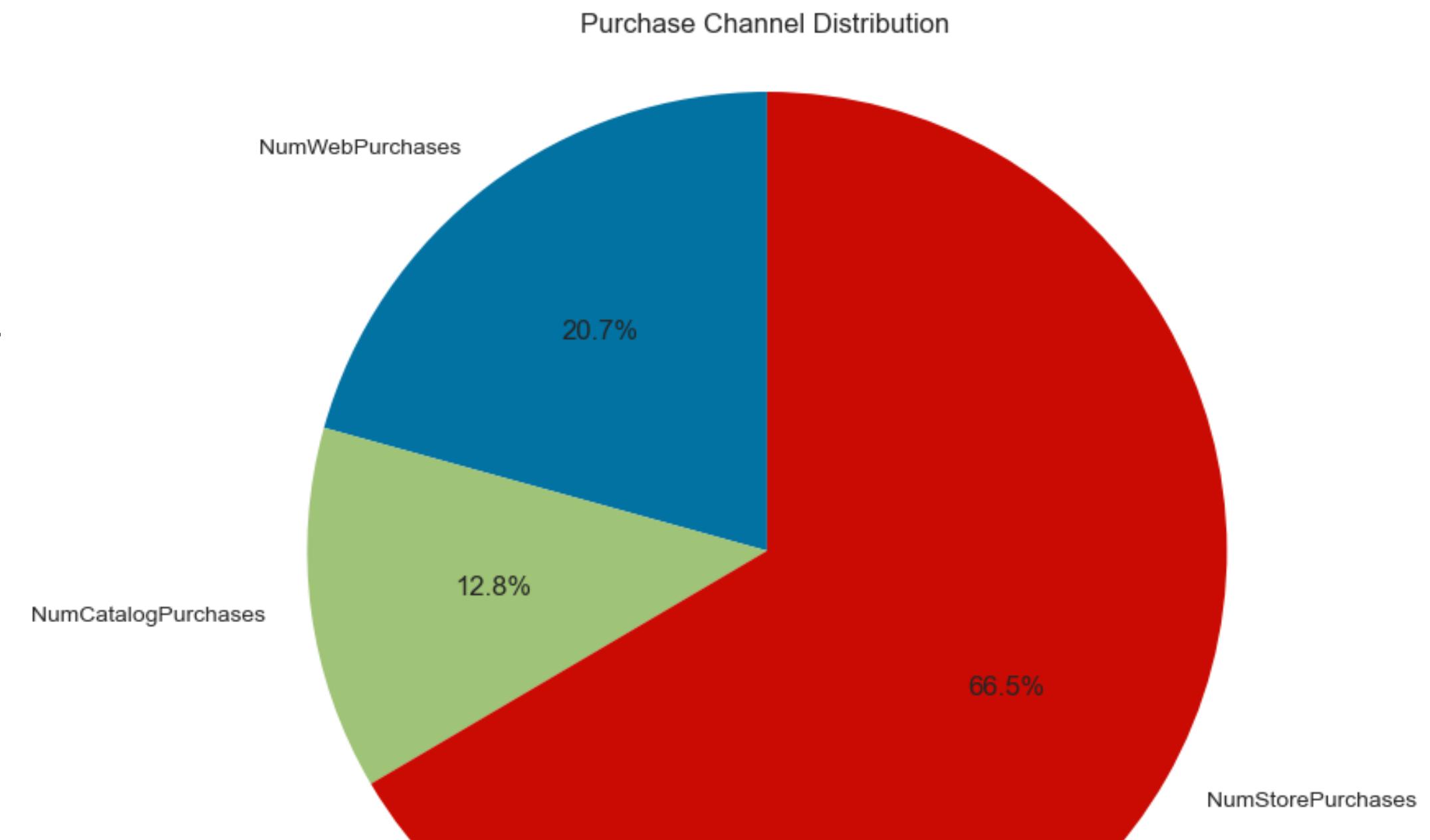
- 在最新的活動上接受優惠得顧客最多。
- 在活動2上接受優惠的顧客最少。
- 普遍顧客在活動上接受優惠的佔比不大。



2-5 特徵圖表說明

顧客購買途徑圓餅圖

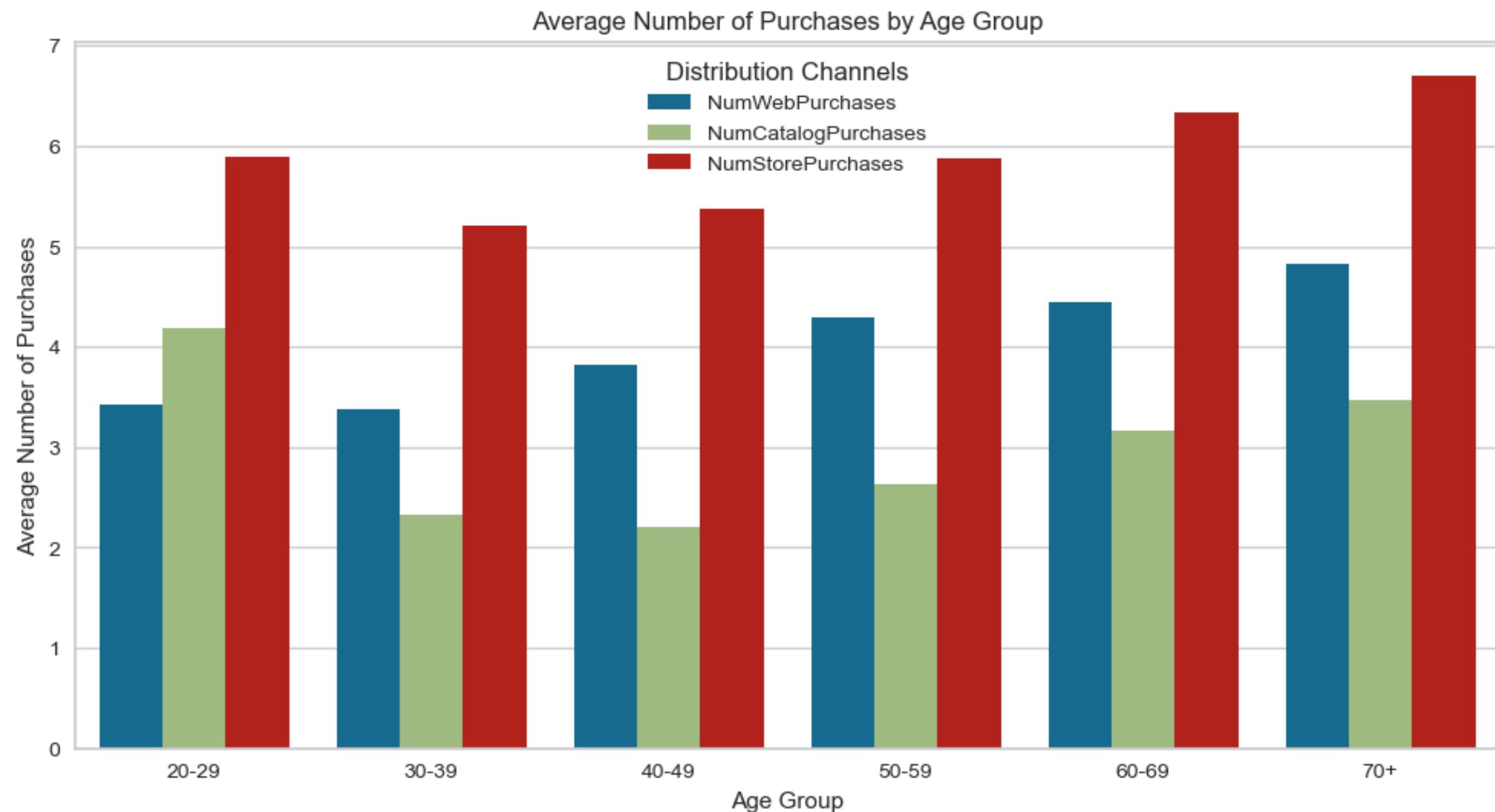
- 透過店面購買產品的顧客佔整體的一半以上。



2-5 特徵圖表說明

顧客購買途徑與年齡關係線直條圖

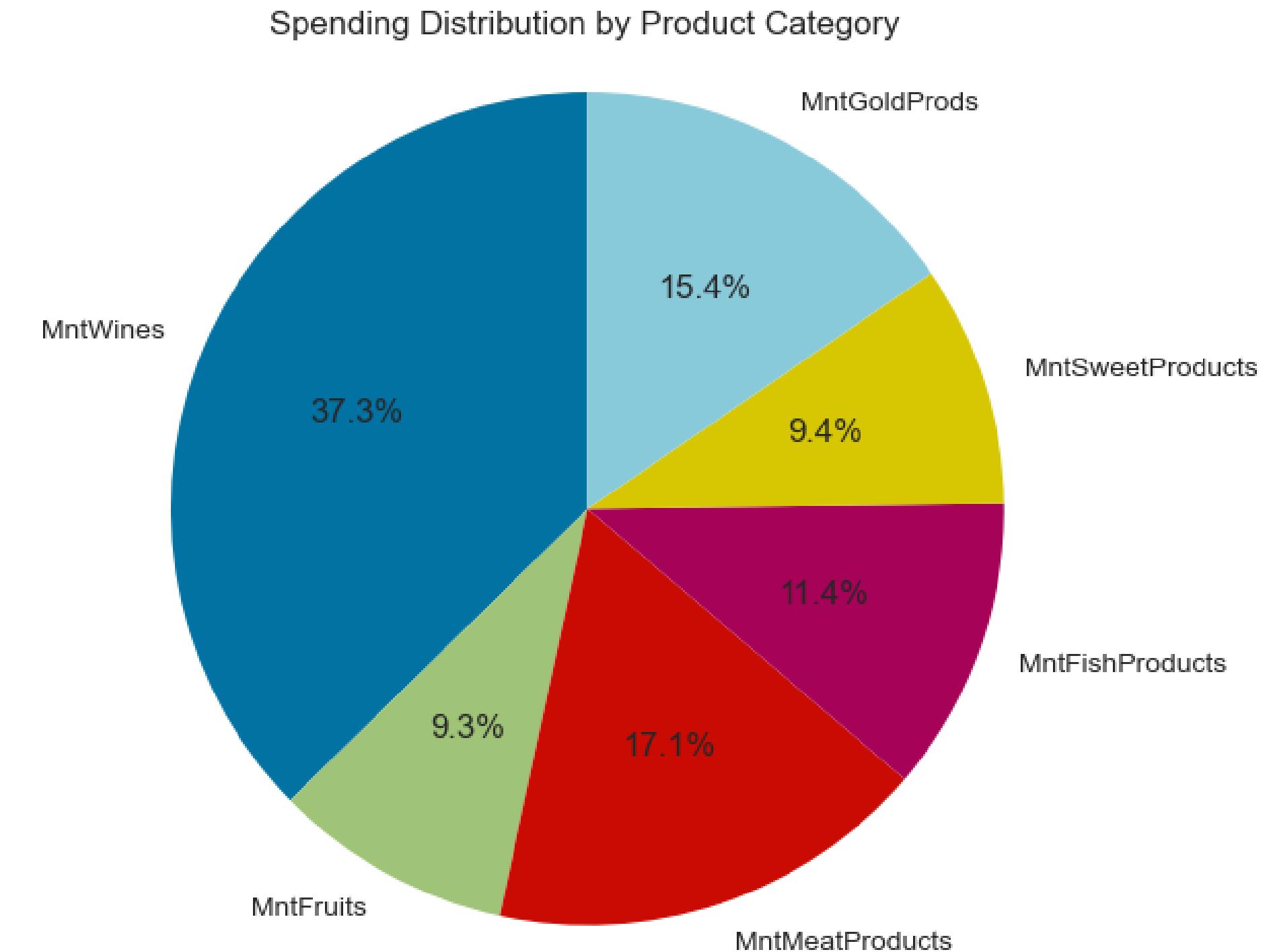
- 店面購買在所有年齡層都是最高的。
- 在透過書刊購買管道中 20~29歲的比例最高。



2-5 特徵圖表說明

顧客購買產品圓餅圖

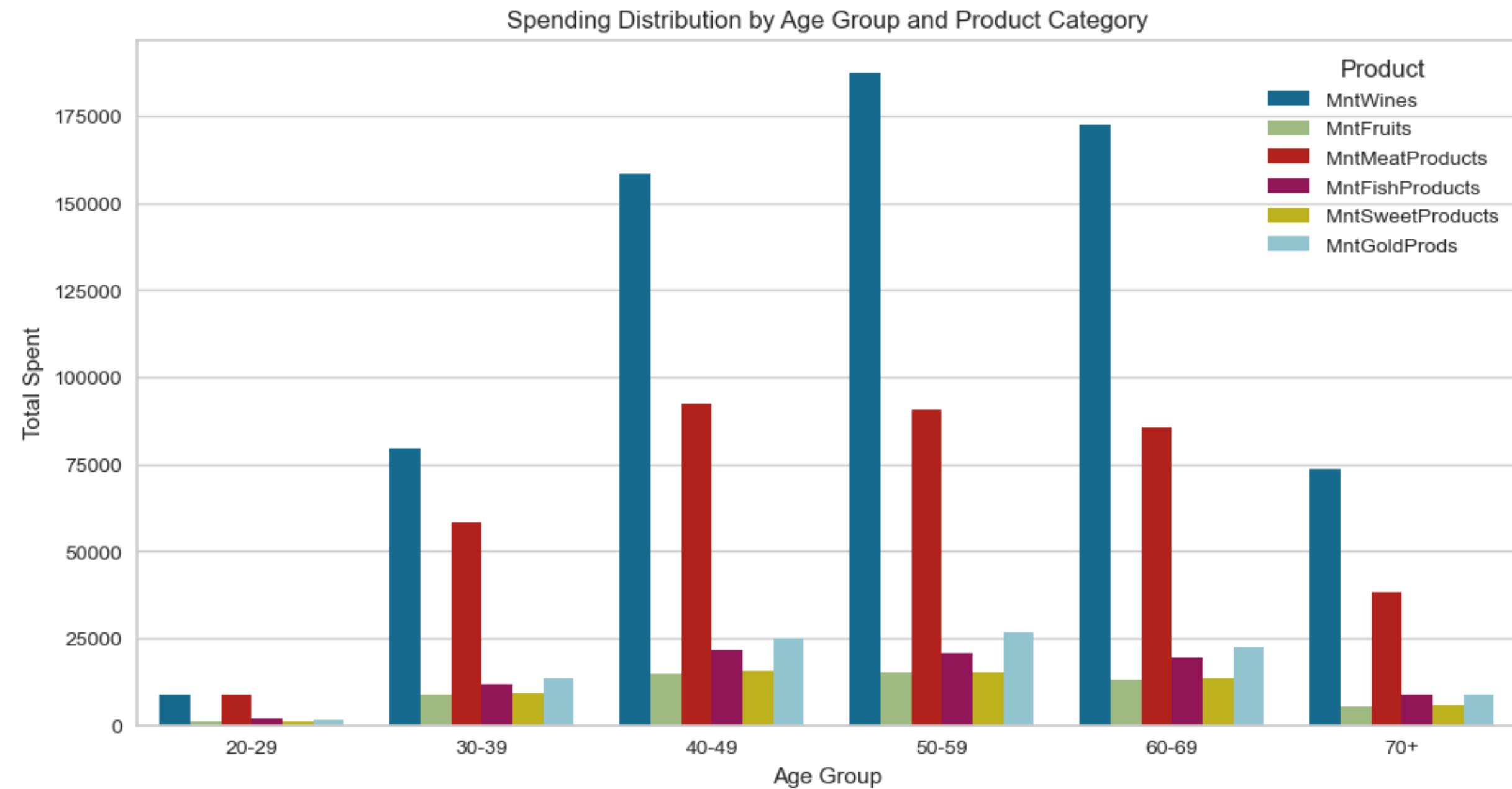
- 酒類產品的銷售量占總銷售的近四成。
- 金飾與肉類的銷售率相近



2-5 特徵圖表說明

顧客購買產品與年齡關係線直條圖

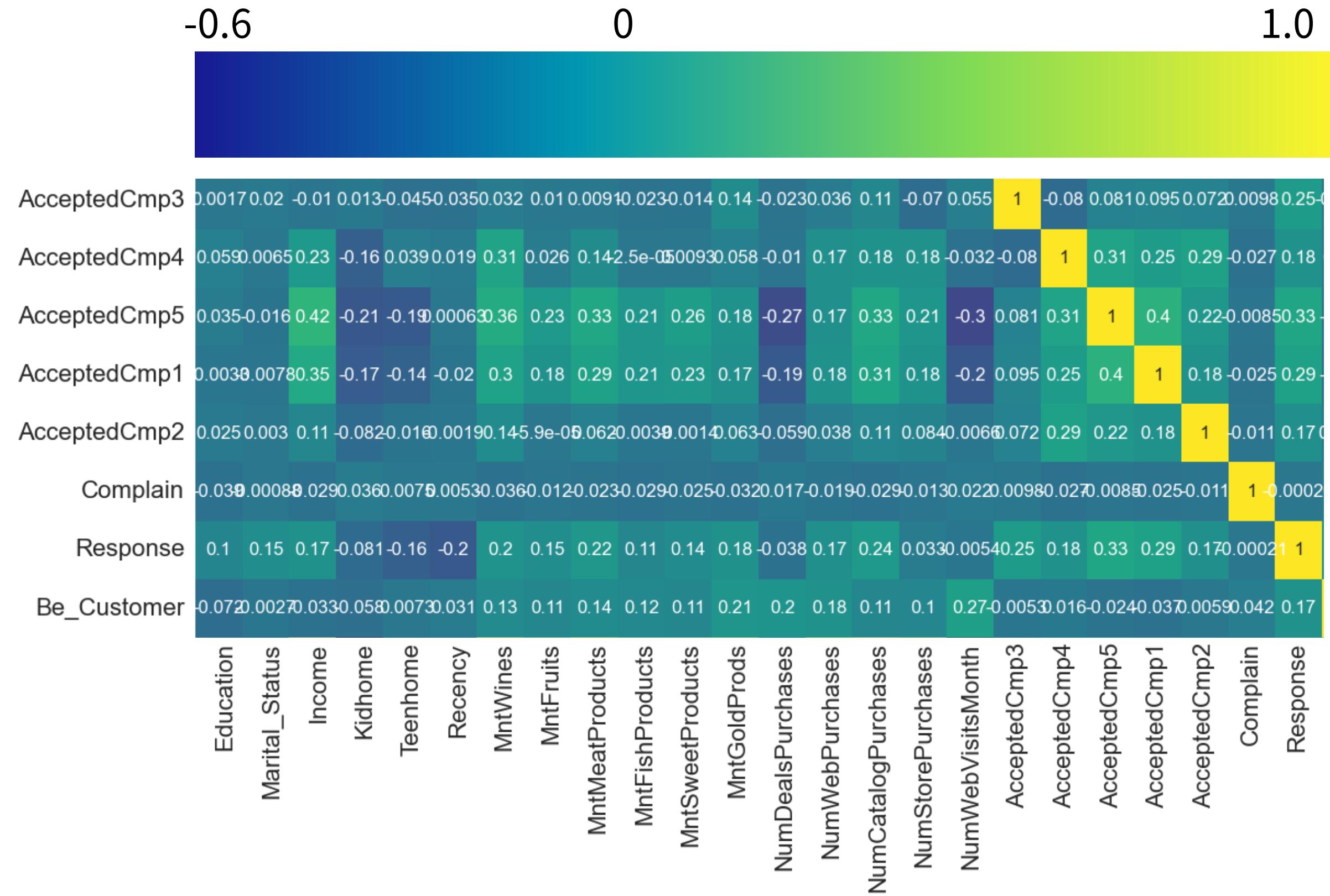
- 酒類產品在各個年齡層的銷售都是最高的。
- 肉類產品為第二高。
- 20~29歲的消費力最低。



2-6相關性

無相關性特徵

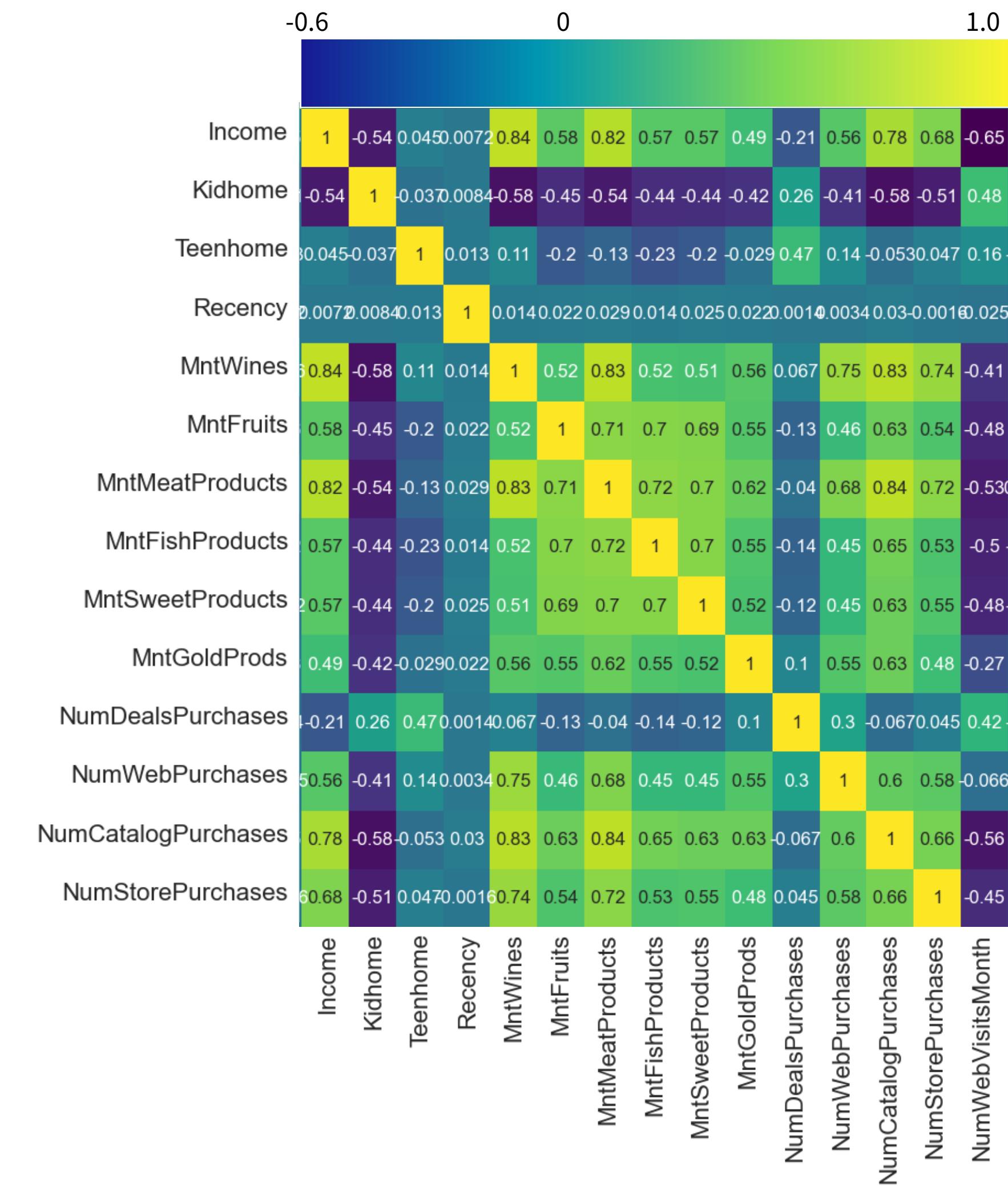
- 由於特徵過多因此只會挑有明顯相關性狀態的特徵做展示。
- 從熱力圖中可看出於y軸上的特徵與其他特徵的相關係數趨近於零，因此能得知這些特徵與其他特徵之間較無關係性。



2-6相關性

有明顯相關性特徵

- 從熱力圖中可看出於y軸上除了Recency, Teenhome與NumDealsPurchases的特徵與其他特徵無明顯差異以外其他y軸的特徵都有明顯的正負特徵。



分析方法

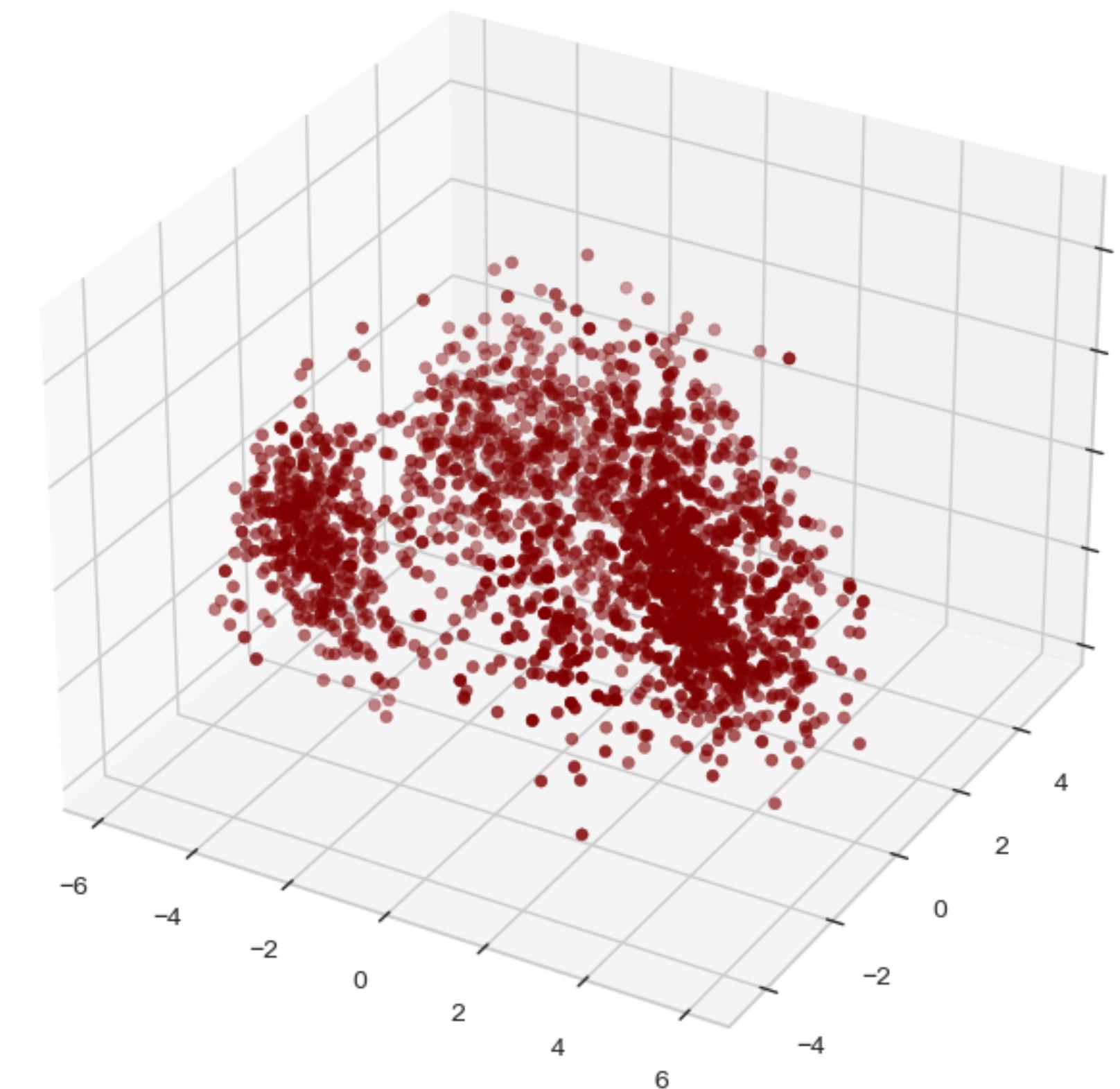
3-1 PCA降維處理

3-2 KMeans模型

3-1 PCA降維處理

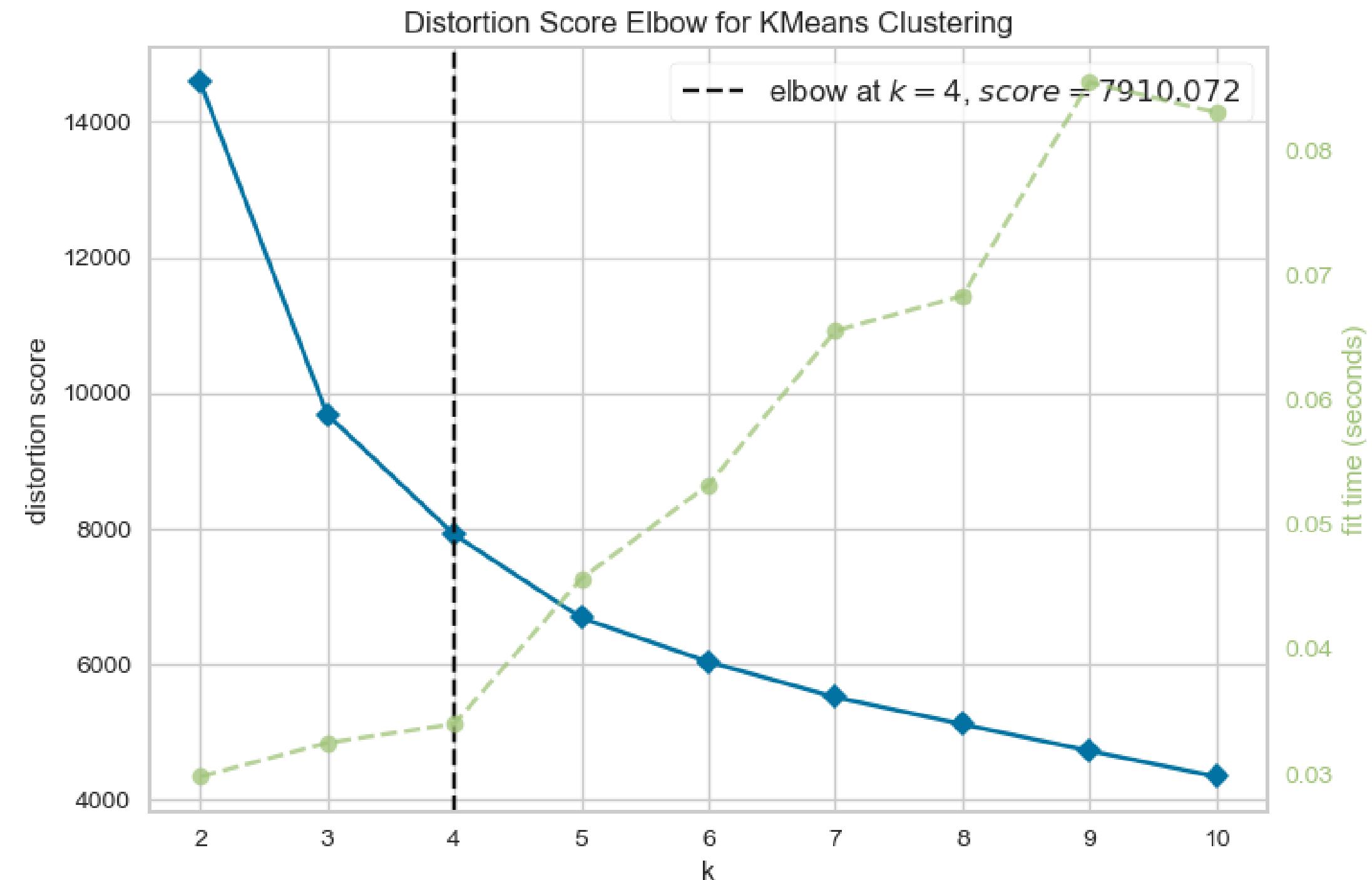
- 由於特徵的數量過多，所以在進行分群之前要先做降為處理。
- 右圖為降維成3個特徵的散點資料集的3D化視覺圖。

3D Scatter Plot Of Data Distribution



3-2 KMeans模型

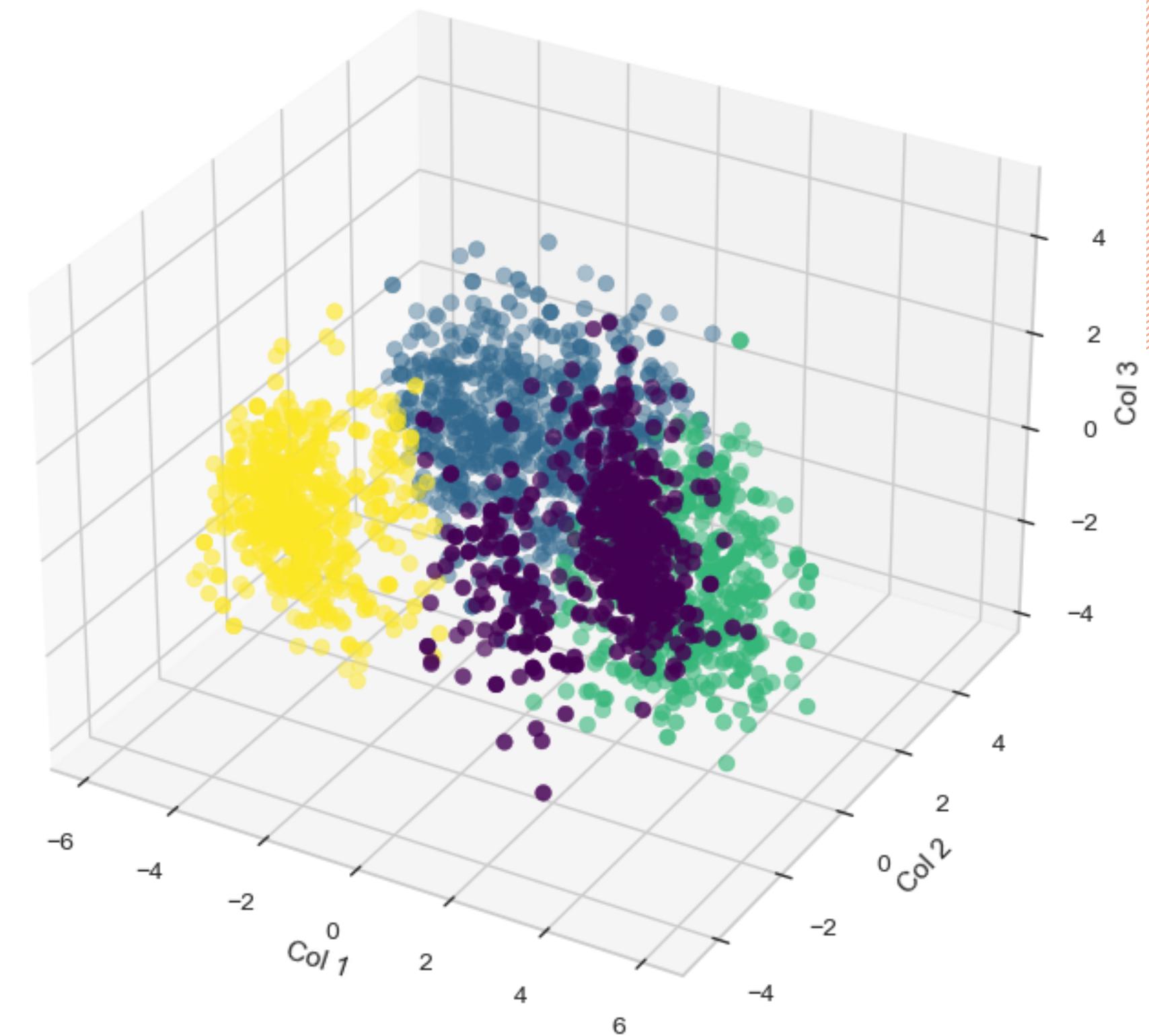
- 在進行分群的機械學習前，先透過聚類可視化(ElbowVisualizer)了解此資料集適合分群的數量。
- 從圖表中可看出此資料集適合4個分類的標籤。



3-2 KMeans模型

- 經過聚類可視化，我們得到了 4 個分類標籤，這些標籤將作為輸入，進入我們的模型中進行分析和預測。
- 可看出分群後的分布過於集中，沒有明顯得分群效果。
- Silhouette Score: 0.17(分數越趨近於1越好)
- Inertia: 7910.63(分數越低越好)
- 從兩個參考指標的分數中也能看出分群的表現不佳。
- 雖然聚類分析結果不佳，但我們可以針對每個群集提出一些建議來改善這個數據集，以提升聚類模型的精確度。

3D Scatter Plot of Clusters



結果呈現

4-1 說明模型後續調整方向

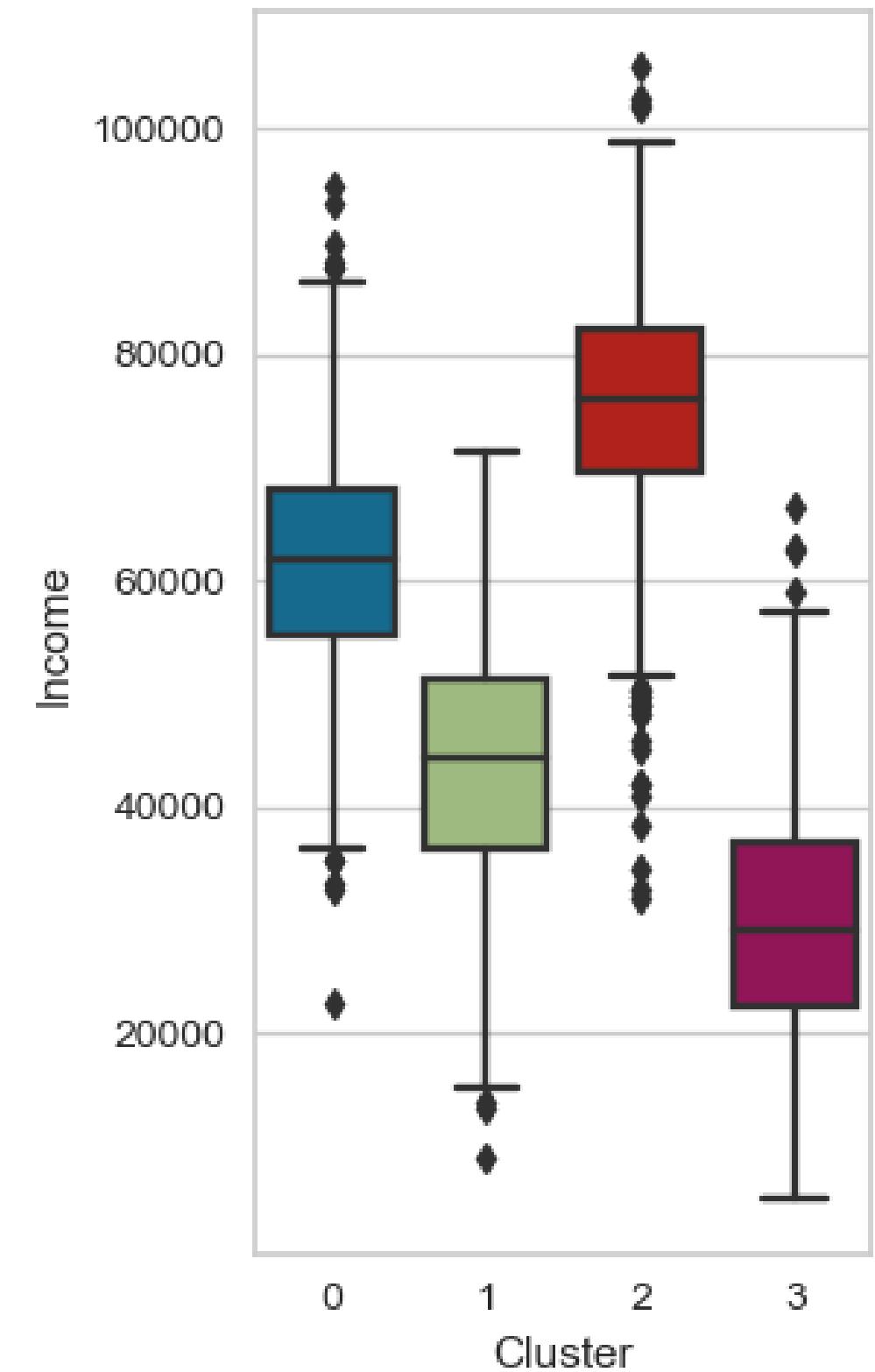
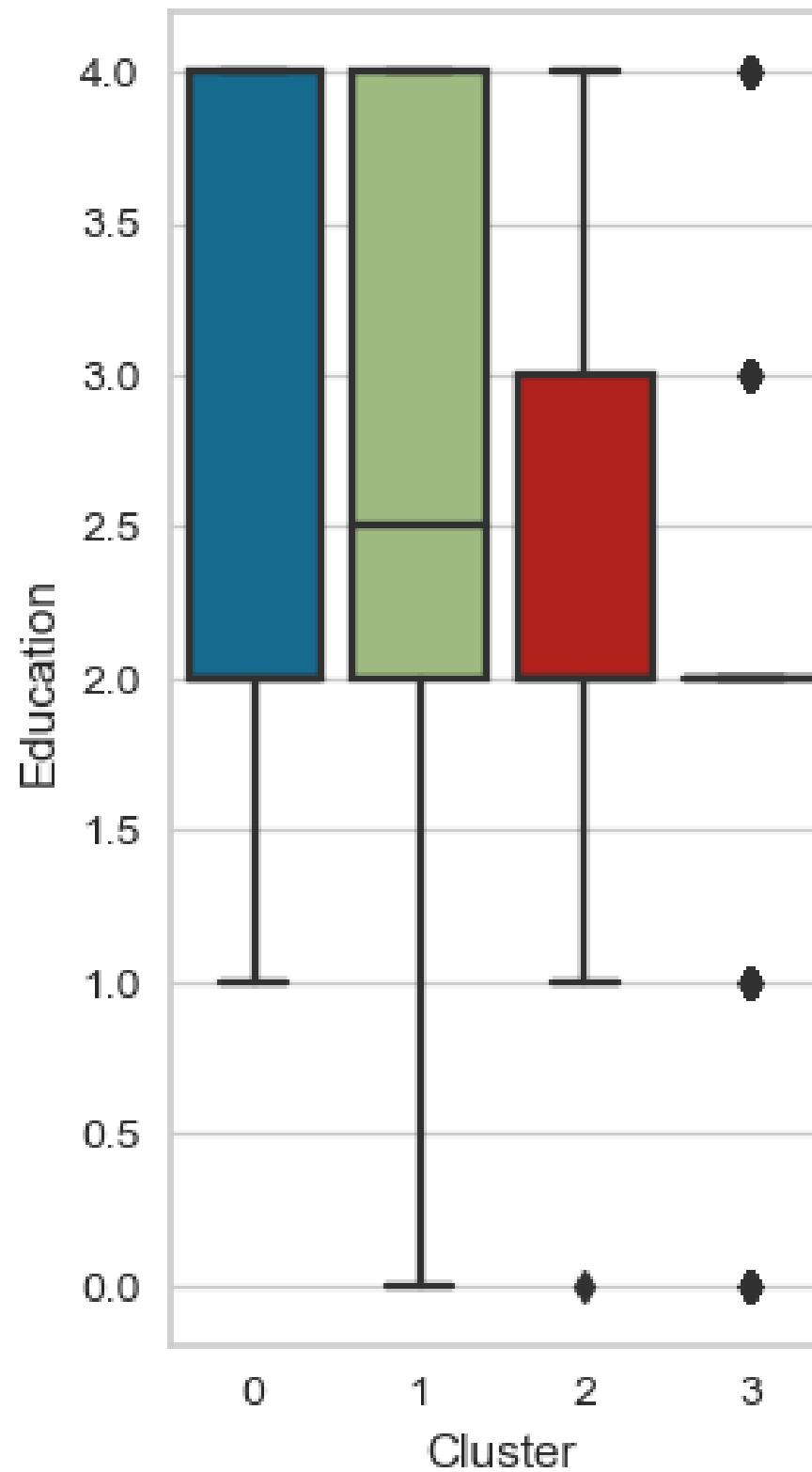
4-2 總結

4-3 專案相關資料

4-1 說明模型後續調整方向

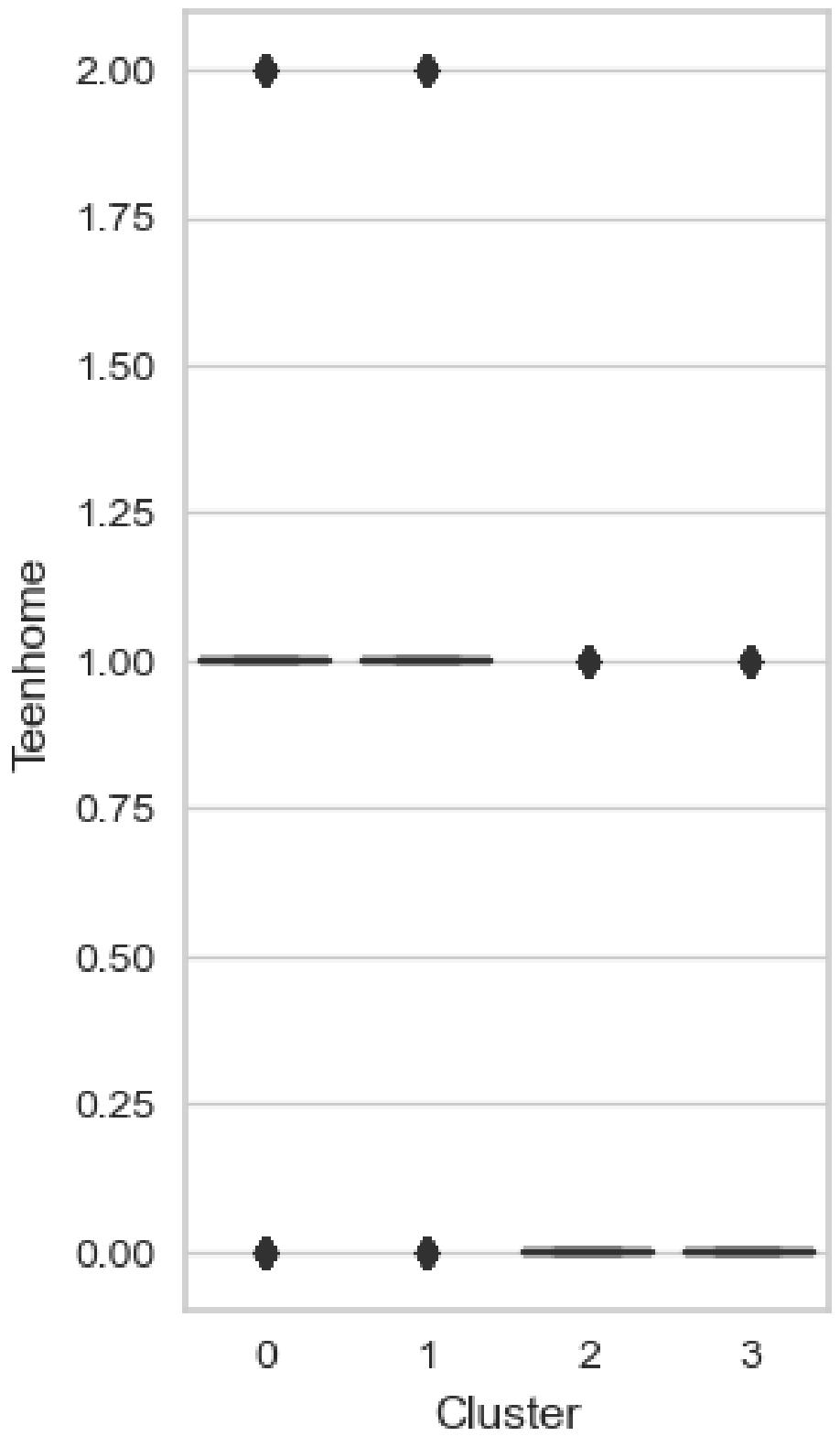
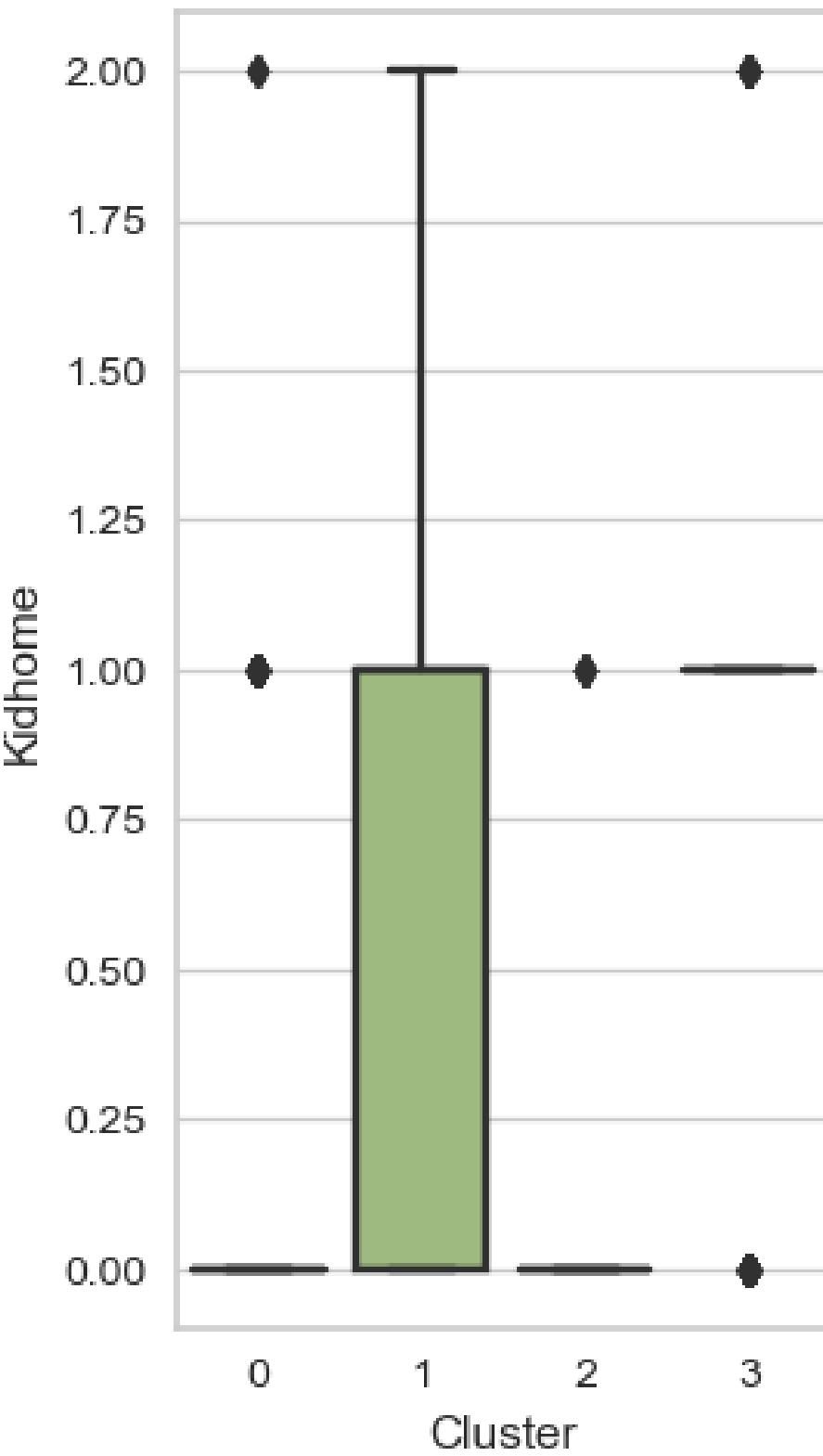
從聚類結果可以看到，顧客被分為4個不同的群組（Cluster 0、1、2 和 3）。每個群組的中位數值顯示了該群組客戶的特徵。

- 由於特徵過多，這邊之列出就有區分的特徵。
- Cluster 0: 教育程度和收入中位數較高。
- Cluster 1: 教育程度中位數略高，收入中位數較低。
- Cluster 2: 教育程度中位數較高，收入中位數最高。
- Cluster 3: 教育程度中位數較低，收入中位數最低。



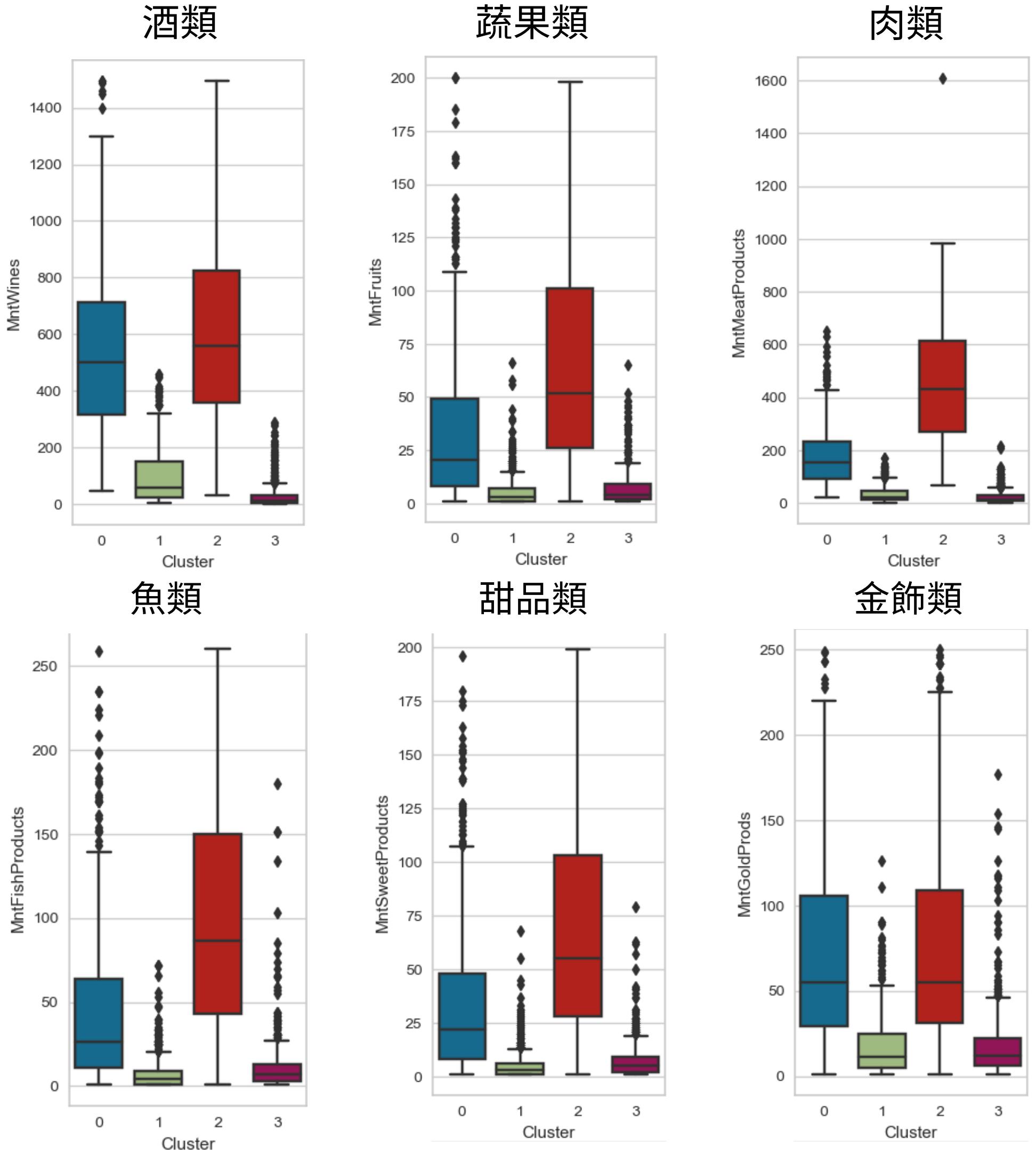
4-1 說明模型後續調整方向

- Cluster 0:沒有孩子，擁有1位青少年。
- Cluster 1:擁有1個孩子和1個青少年。
- Cluster 2:沒有孩子和青少年。
- Cluster 3:擁有1個孩子，沒有青少年。



4-1 說明模型後續調整方向

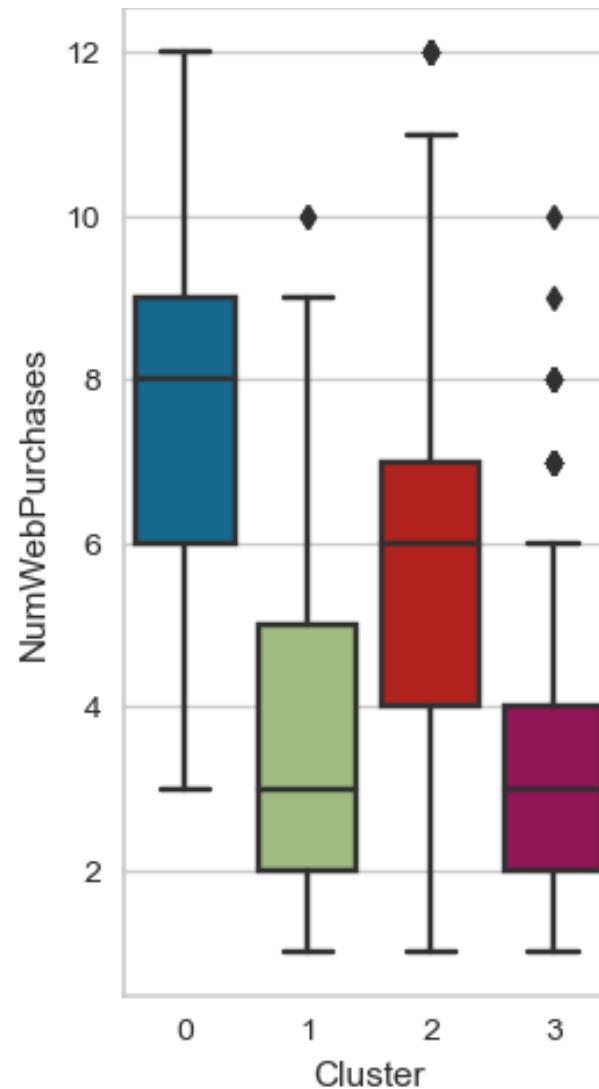
- Cluster 0:消費金額主要集中在酒類、肉類產品和金飾產品上。
- Cluster 1:消費金額主要集中在酒類、金飾產品上，其他消費較低。
- Cluster 2:所以品項消費最多的。
- Cluster 3:所以品項消費最低的。



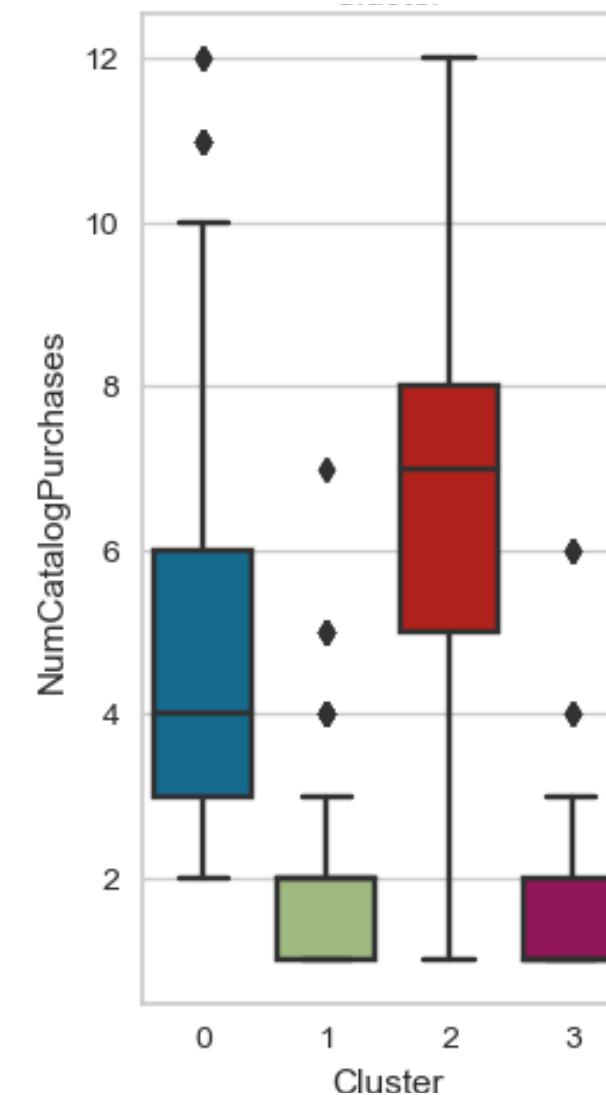
4-1 說明模型後續調整方向

- Cluster 0:較高的網絡購物次數和店面購物次數。
- Cluster 1:網絡和店面購物次數和目錄購物次數比Cluster 3高。
- Cluster 2:最高的書刊購物次數，中等的網絡購物次數和實體店購物次數。
- Cluster 3:三種購物管道都最低。

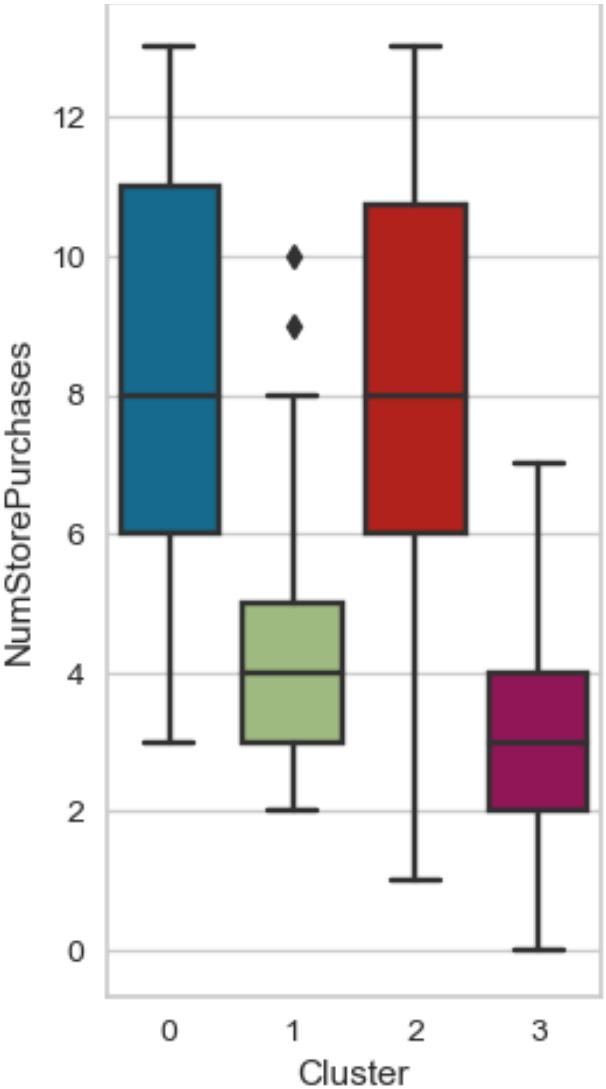
網絡



書刊



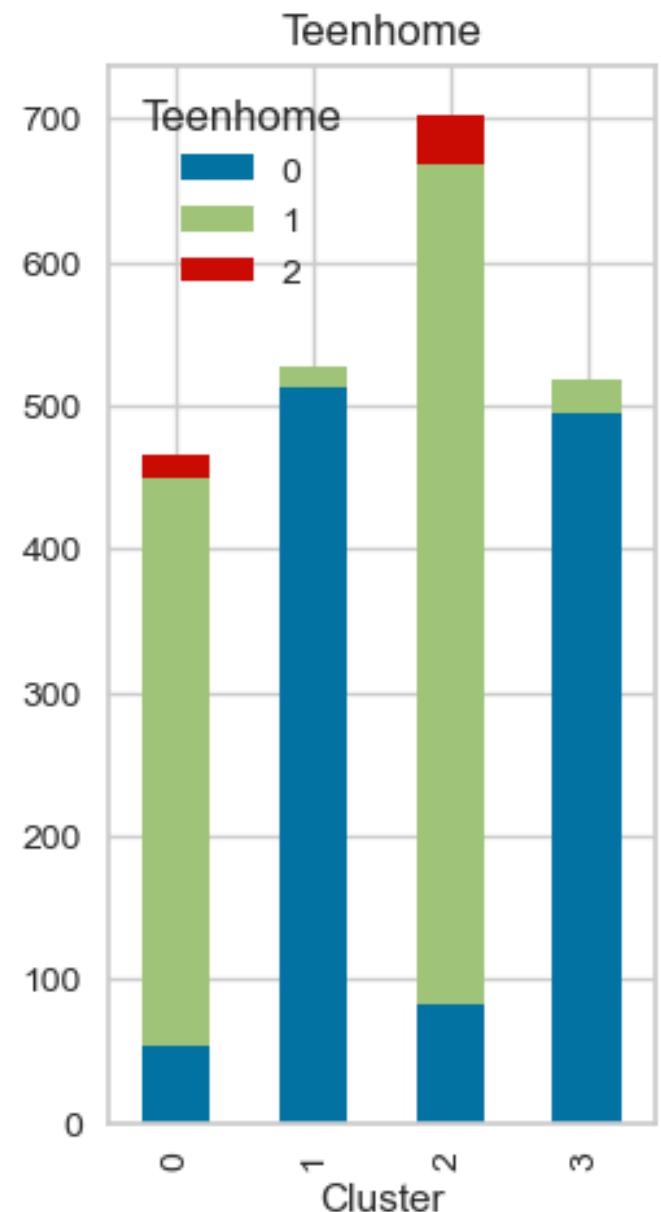
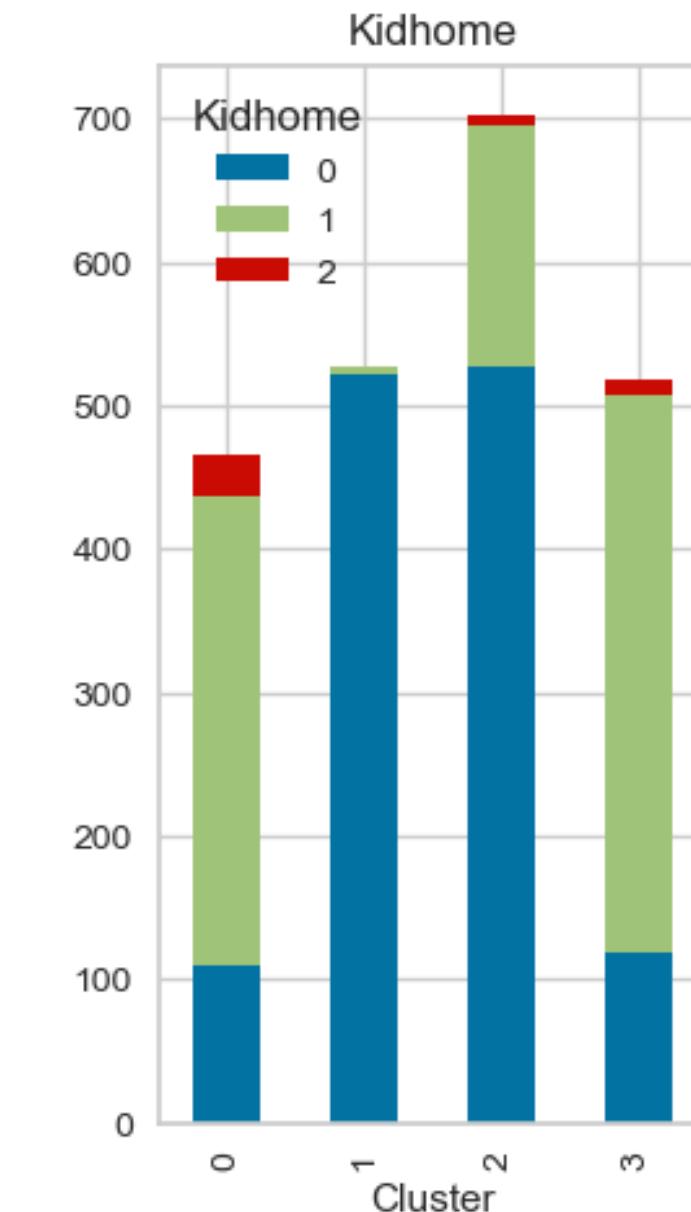
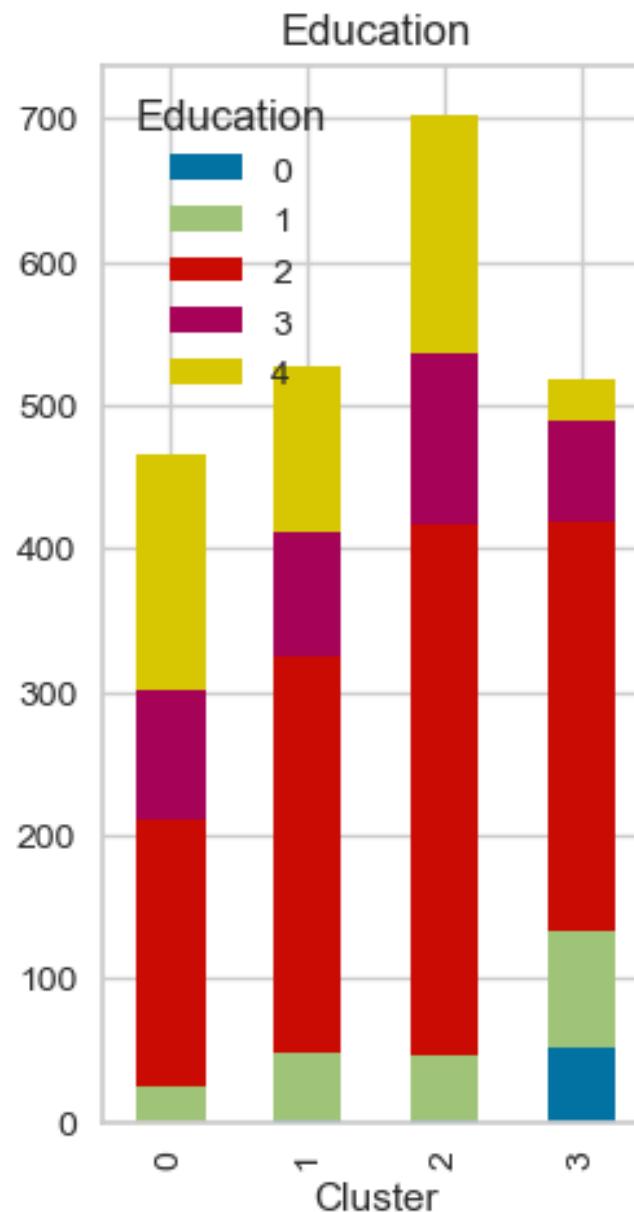
店面



4-1 說明模型後續調整方向

透過堆積直條圖能看出各類別中特徵內容組成

- Cluster 0: 顧客人數占比最少。
- Cluster 1: 小孩和青少年的人數占比都很少。
- Cluster 2: 顧客人數和有青少年人數占比最多。
- Cluster 3: 教育程度不及其他三者，小孩占比多，青少年人數占比少。



4-2 總結

透過前面的圖表關係我這邊提出各分類的改善建議用於增進這份數據集的完整性及聚類模型的精確度。

- Cluster 0：此群集的客戶具有較高的教育程度和收入，因此可以嘗試提供更多高品質或高價位的產品，以吸引此類客戶。由於他們的消費主要集中在酒類、肉類產品和金飾產品上，可以考慮增加這些類別的產品種類和優惠活動，以便進一步提高客戶的消費額。
- Cluster 1：這個群集的客戶在教育程度上表現較好，但收入較低。考慮到他們的消費主要集中在酒類和金飾產品上，可以嘗試提供更多價格合理的產品，以吸引這一群體。另外，可以嘗試通過增加與孩子和青少年相關的產品種類和促銷活動，來提高此類客戶在其他消費品項上的消費。

4-2 總結

- Cluster 2：這個群集的客戶在所有品項上都有很高的消費。為了維持這一群體的高消費水平，可以通過提供更多的優惠、積分獎勵和會員活動等，來進一步增加他們在各個消費品項上的消費。此外，可以針對這一群體的客戶提供更多定制化的產品和服務，以提高客戶滿意度和忠誠度。
- Cluster 3：此群集的客戶在教育程度和收入方面相對較低，因此可以考慮提供更多針對這一群體的實惠產品和促銷活動。由於他們在三種購物渠道上的購物次數都相對較低，可以嘗試通過提高線上購物平台的易用性、優化實體店的購物體驗以及提高目錄購物的吸引力，來提升這一群體的消費頻率。

4-3 專案相關資料

整個分析專案的程式碼、資料集和圖表都儲存在Github裡面，
如果想了解詳細分析方法，還請您複製網址查看，謝謝。

Github:https://github.com/Zerotoen/marketing_campaign

感謝您的閱讀

