Discovering potential relations in customer data by Apriori

組別 : Group 14
組員 :
r09942131李雲翔　　r10921099黃柏維　　r09921018周靖樺　　r10945061林宇恆

1. Dataset introduction

本次我們研究的dataset是取自於一個賣場的顧客資料集,
資料集中, 可概略分為People、Products、Promotion、Place有關的features

其中People與顧客身分有關, 如收入、學歷、婚姻狀況等;
Products與顧客消費的商品紀錄有關, 如花費在酒上的金額;
Promotion顧客接受折扣記錄相關, 如顧客有幾次是在有折扣的情況下消費;
Place為顧客消費管道有關, 如實體店面、網路下單、型錄消費次數。

我們將利用Apriori探討資料中潛在的關係。

2. General Preprocessing

此部分為通用的前處理, 除此之外我們再會針對不同的討論方向產生適合該主題的feature, 在Results的部分會有更一步的討論。

a. Nan

```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID                  2240 non-null   int64
 1   Year_Birth          2240 non-null   int64
 2   Education           2240 non-null   object
 3   Marital_Status      2240 non-null   object
 4   Income              2216 non-null   float64
 5   Kidhome             2240 non-null   int64
 6   Teenhome            2240 non-null   int64
 7   Dt_Customer         2240 non-null   object
 8   Recency             2240 non-null   int64
 9   MntWines            2240 non-null   int64
 10  MntFruits           2240 non-null   int64
```

我們發現有部分顧客缺失income的資料, 故我們直接將存有缺失值的顧客從資料中捨去。

b. Features engineering

時間相關

- 我們發現dataset的最後紀錄日期為2014-12-16, 故我們以此為依據推算顧客的年齡與會員年資。
- Year -> Age
  Dt_Customer -> Seniority

消費相關

- 'Spending'= Amount of (Wines + Fruits + Meat + Fish + Sweet + Gold)

簡化分類

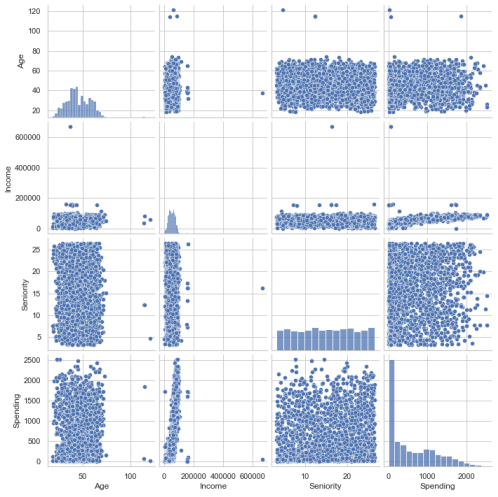
- 婚姻狀態: 'Divorced','Single','Married','Together','Absurd','Widow','YOLO'
  -> 只分類為‘'Alone’ 和 ‘In couple’
- 學歷 -> 'Basic','2n Cycle','Graduation','Master','PhD'
  -> 只分類為'Undergraduate' 和 'Postgraduate'

合併feature

- 'Kidhome' + 'Teenhome' 合併為 ‘Children'

c. Outliar

- 從下圖可以發現圖中有些零散偏離群體分布的點, 這些outliar我們將其從資料中捨去。
  - data[‘Age’] > 110
  - data[‘Income‘] > 600000



d. Data segmentstion

- 因為Apriori只處理類別型的資料(categorical data), 因此我們要對我們想分析的feature做適當的segmentation。
- Cut by value:
  - Age: [0, 30, 45, 65, 120] -> ['Young', 'Adult', 'Mature', 'Senior']
  - 'Has_child' [==0,!=0] -> ['Has child', 'No child']
- Cut by quatile:
  - Income: q=4 -> ['Low income', 'Low to medium income', 'Medium to high income', 'High income']
  - Seniority: q=4 -> ['New customers', 'Average to new customers', 'Average to old customers', 'Old customers']
  - Goods: [0, .25, .75, 1] -> ['Low consumer', 'Frequent consumer', 'Biggest consumer']

3. Apriori

- Association factors introduction:

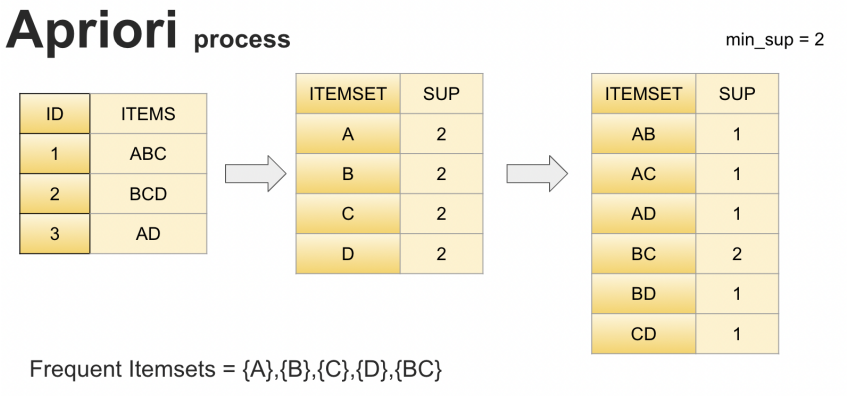
  a. 支持度(Support): support(A) = count(A) / count(All Data)

  b. 信心水準(Confidence): Confidence(A -> B) = P(B|A) = P(A ∩ B) / P(A)

  c.提升度(Lift): Lift(A -> B) = Confidence(A -> B) / P(B) = P(B|A) / P(B)

  i. 頻繁項集(Frequent Itemsets): 經常一起出現的物品集合

  ii. 關聯規則(Association Rules): 表達數據之間的可能存在很強關聯性



4. Problem Statement

a. Place-related

i. 年紀是否跟購物方式有關
ii. 消費場所和商品的關係

b. Personaliy-related

i. 分析行為與身份的關係
ii. 商品和身份的關係

c. Promotion campaign-related

i. 收入與接受促銷的關係
ii. 有無小孩與接受促銷的關係

d. Spending-related

i. 消費力和個人基本資料的關係
ii. 商品與商品之間的關係

5. Results

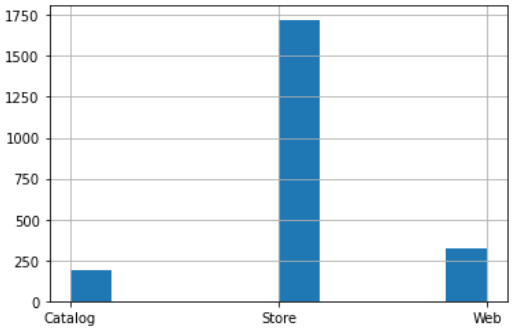
a. Place-related

前處理

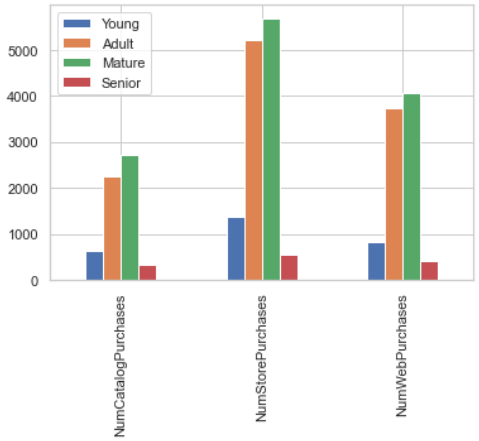
針對與購物地點有關的features: 在實體店面消費次數(NumStorePurchases)、網購消費次數(NumWebPurchases)和型錄消費次數(NumCatalogPurchases), 我們額外生成最常購賣地點:定義為最多消費次數的地方。

年紀是否跟購物方式有關

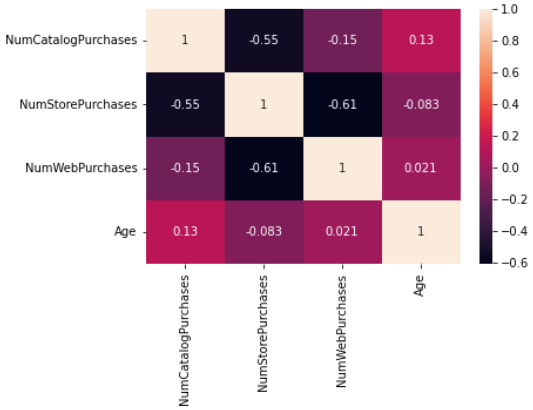
我們將每個人最常的購買地點做統計，如下圖;發現最常購入的地點是實體店面，其次是網路購物，最後才是型錄消費。



將地點的累積消費次數做統計，並依年齡層分群，如下圖;
依肉眼觀察不同購買地點的年齡分布相當相似，看不出明顯年齡上的差別



因此我們用量化的方式計算不同消費地點的消費次數和年齡之間的相關性



我們發現到年齡與通路間無明顯相關性，但我們能看見實體店面與網路消費具有負相關性。

消費場所和商品的關係

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
10195	(Education_Postgraduate, Has_child_Has child, Meat_segment_Frequent consumer)	(MostPurchasePlace_Web)	0.379	0.146	0.093	0.246	1.681	0.038	1.132

我們得出一個associstion rule為經常買肉、有小孩、大學畢業的人會比較傾向網路消費，雖然有點難以加以解釋背後的原因，但仍為。

值得一提的是，我們並不能從 dataset 中看出物品經由哪個通路購入的關聯性此表雖有前件包含 Meat_segment_Frequent Consumer，但不能指出 Meat 是在 Web 上購買。

b. Personality-related

前處理

- 收入分類:

  我們把收入分成 4 個區段，收入為前百分之 25 高的人定義成「high income」、前百分之 25 高到前百分之 50 高的人定義成「Medium to high income」、前百分之 50 高到前百分之 75 高的人定義成「low to Medium income」、後百分之 25 的人定義成「low income」。

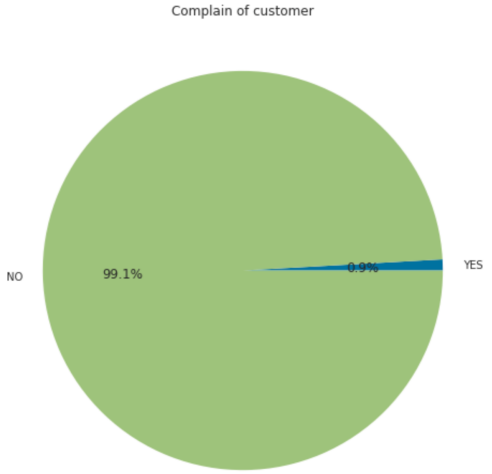
- 商品客群分類:

  我們將六個產品的客群分成三類。第一類為購買某項產品數量為前百分之 25 多的人，將此類客群定義成「Biggest consumer」。第二類為購買某項產品數量為前百分之 25 多到前百分之 75 多的人，將此類客群定義成「Frequent consumer」。第三類為購買某項產品數量為後百分之 25 多的人，將此類客群定義成「Low consumer」。

分析行為與身份的關係

- 抱怨的人通常是什麼身份的人

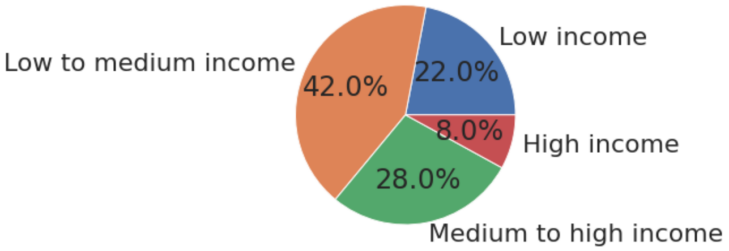
  看到資料中，含有是否曾經抱怨這項 feature，讓我們對抱怨的人通常具有什麼樣的特徵產生好奇。於是我們一開始嘗試著用 Apriori Algorithm 分析，但是沒有跑出結果。後來發現因為曾經抱怨過的 data 數目太少，所以造成 Apriori Algorithm 中的 support 太小，無法超過設定的 support 門檻 0.08。於是我們把所有曾經抱怨過的 data 展開來看。發現曾經抱怨過的 data 數目只有 21 位，佔總 data 數 2240 比例不到百分之一。雖然曾經抱怨過的 21 個資料中，「已經大學畢業」及「有小孩」的人數佔了所有曾經抱怨過的資料的 80% 以上。但是因為原始資料中，「已經大學畢業」及「有小孩」的資料本來就佔大多數，所以並無得出得「曾經抱怨過」的人有「已經大學畢業」及「有小孩」這兩項特徵。



▲抱怨的人佔所有資料的比例圓餅圖

- 小孩多的人通常是怎樣的人的

在原始資料中，小孩人數這個 feature 總共有四種。分別是「沒有小孩」、「有一個小孩」、「有兩個小孩」、「有三個小孩」。我們想要分析顧客資料中，小孩多的人通常具有什麼樣的特徵。於是我們一開始先分析「有三個小孩」的人通常具有什麼樣的特徵，我們使用如上述一樣的方法，先試著使用 Apriori Algorithm 去分析，結果遇到與上述同樣的問題，就是「有三個小孩」的資料數太少（只有 53 筆資料，約佔原資料的 2%），無法超過設定的 support 門檻。於是我們退而求其次，分析「有兩個小孩」的人。我們使用 Apriori Algorithm 分析之後，發現 lift 為前兩高的兩項特徵都有「收入為 low to Medium income」這項特徵。於是我們猜測「有三個小孩」的人可能也會有這項特徵。因此我們把「有三個小孩」的 53 筆資料的收入情況使用圓餅圖分析。分析結過如下圖所示，發現「有三個小孩」的資料中，收入為 low to Medium income 的比例佔了 42%。於是我們整合上面的發現，得到小孩多的人的收入狀況通常位於 low to Medium income 這個區間。

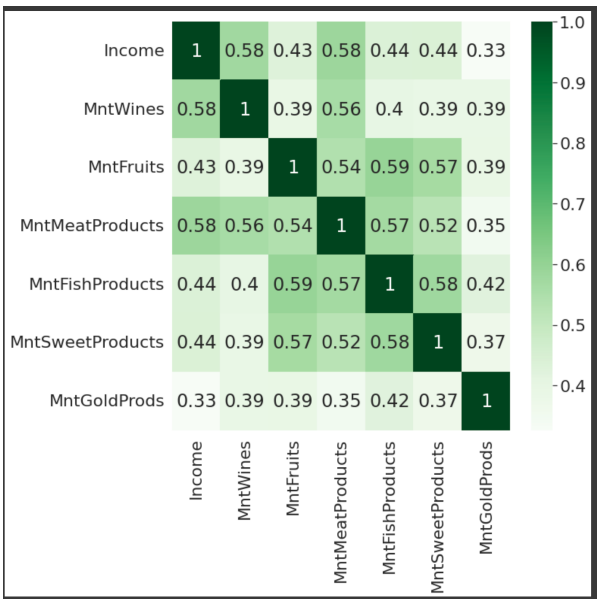


▲ 有三個小孩的人的收入情況圓餅圖

商品和身份的關係

- 高收入的人購買哪種產品的比例較其他產品高

我們把收入及商品客群用上面的定義進行分類，並一樣使用 Apriori Algorithm 下去跑，最小的 support 為 0.08。分析出來的結果發現 lift 前三高（分別為 3.489, 3.489, 3.473）的特徵在「酒」和「肉」這兩項產品都是屬於 Biggest consumer（也就是購買「酒」和「肉」的數量為前百分之 25 多的人），但是在其他項產品並沒有這項關係。此外，我們另外用 correlation 分析高收入與購買產品的相關程度，並做成以下的圖。發現收入越高，在酒和肉所購買的數量和其他商品相比，相關程度明顯較高。所以我們得到收入高的人購買「酒」和「肉」產品的比例較其他產品高。



▲ 收入多寡與購買產品數量相關程度

c. Promotion campaign-related

前處理

●新增feature: "AcceptCmpornot" (曾接受過促銷:Accept ;不曾接受過:No)

○data['AcceptedCmp']=data['AcceptedCmp1']+data['AcceptedCmp2']+data['AcceptedCmp3']+data['AcceptedCmp4']+data['AcceptedCmp5']

○data['AcceptCmpornot'] = np.where(data.AcceptedCmp> 0, 'Accept', 'No')

AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	→	AcceptedCmp	AcceptCmpornot
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	0	0		0	No
1	0	0	0	0		1	Accept
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	1	1	0		2	Accept
0	0	0	0	0		0	No
0	0	0	0	0		0	No
0	0	0	1	0		1	Accept
0	0	0	0	0		0	No

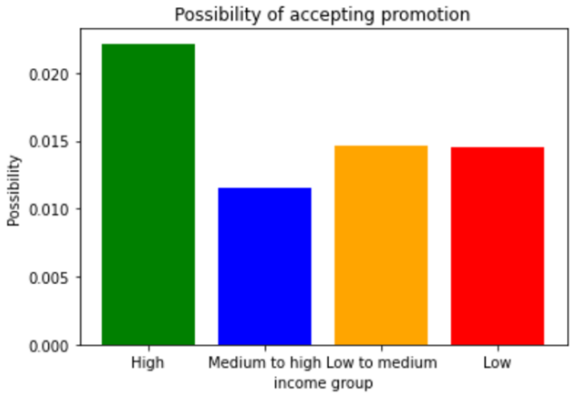
●接受促銷的人的特徵--- 運用Apriori進行關聯性的分析

●接受促銷(AcceptCmpornot_Accept)的rules:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
748	(Income_group_High income, Wines_segment_Biggest consumer)	(AcceptCmpornot_Accept)	0.149	0.207	0.085	0.571	2.758	0.054	1.850
736	(Marital_Status_In couple, Wines_segment_Biggest consumer)	(AcceptCmpornot_Accept)	0.158	0.207	0.080	0.510	2.461	0.048	1.618
743	(Income_group_High income, Has_child_No child)	(AcceptCmpornot_Accept)	0.174	0.207	0.087	0.499	2.407	0.051	1.581
6	(Wines_segment_Biggest consumer)	(AcceptCmpornot_Accept)	0.249	0.207	0.121	0.486	2.347	0.069	1.544
712	(Education_Postgraduate, Wines_segment_Biggest consumer)	(AcceptCmpornot_Accept)	0.235	0.207	0.112	0.479	2.311	0.064	1.521
754	(Income_group_High income, Meat_segment_Biggest consumer)	(AcceptCmpornot_Accept)	0.190	0.207	0.086	0.452	2.183	0.046	1.448
4	(Income_group_High income)	(AcceptCmpornot_Accept)	0.250	0.207	0.107	0.429	2.068	0.055	1.387
706	(Income_group_High income, Education_Postgraduate)	(AcceptCmpornot_Accept)	0.231	0.207	0.098	0.422	2.036	0.050	1.371
10	(Meat_segment_Biggest consumer)	(AcceptCmpornot_Accept)	0.250	0.207	0.094	0.377	1.821	0.043	1.273

收入與接受促銷的關係

- 收入族群以四分位距分成四個族群，由高到低分別為"High","Medium to high",”Low to medium”,”Low”.
- 計算各族群接受促銷的機率: (該族群接受促銷的人/該族群的總消費次數)
- High: 0.02215, Medium to high:0.01158,
  Low to medium:0.01461, Low:0.01457

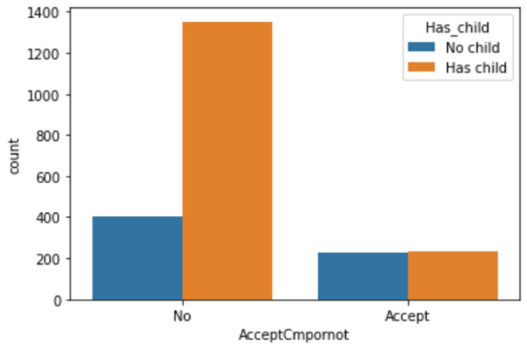


觀察: 收入高的人接受促銷的機率最高，是其他收入族群的1.5倍以上

有無小孩與接受促銷的關係

- 分別計算有無小孩的人接受與不接受促銷的人數，進而分別計算有小孩跟沒有小孩這兩個族群接收促銷的機率

```
Possibility of accepting promotion:
Has child: 0.15
No child: 0.36
```



觀察: 沒有小孩的人接受促銷的機率較高，是有小孩的人的2.4倍

d. Spending-related

前處理

- 婚姻:

  Divorced（離婚）,Single（單身包含Absurd,YOLO and Widow）,In Couple（有伴侶的，包含情侶以及夫妻）

- 學歷:

  分別為Basic(基礎學歷), 2n cycle（二技）, Graduation（學士）, Master（碩士）, PhD（博士）

- 子女數量:

  分別為No child, 1 child, 2 children, 3 children

- 年齡:

  我們把年齡分為4個區段，四等分，分別是 Young(0-18),Adult(18-45),Mature(45-65),Senior(65-120)

- 收入分類:

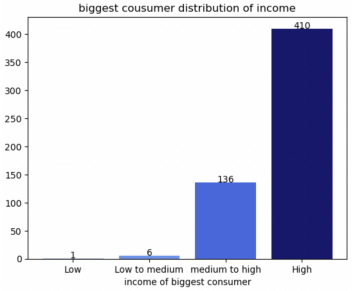
  我們把收入分成 4 個區段，收入為前百分之 25 高的人定義成「high income」、前百分之 25 高到前百分之 50 高的人定義成「Medium to high income」、前百分之 50 高到前百分之 75 高的人定義成「low to Medium income」、後百分之 25 的人定義成「low income」。

消費力和個人基本資料的關係

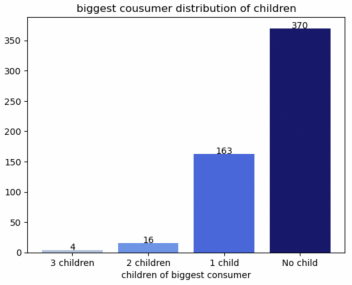
- 依序分析了消費力和婚姻、學歷、子女數量、年齡的關係得到的結果:

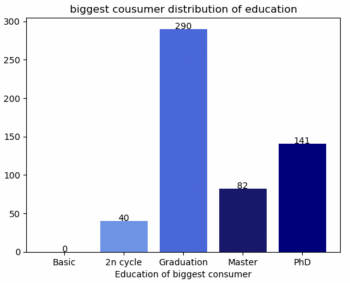
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
298	(Children_segment_No child, Income_segment_High income)	(Spending_segment_Biggest consumer)	0.174	0.250	0.141	0.813	3.258	0.098	4.013
384	(Children_segment_No child, Marital_Status_segment_In couple, Income_segment_High income)	(Spending_segment_Biggest consumer)	0.105	0.250	0.084	0.803	3.216	0.058	3.801
244	(Income_segment_High income, Education_segment_Medium education)	(Spending_segment_Biggest consumer)	0.137	0.250	0.102	0.749	3.002	0.068	2.992
67	(Income_segment_High income)	(Spending_segment_Biggest consumer)	0.250	0.250	0.185	0.740	2.966	0.123	2.887
293	(Marital_Status_segment_In couple, Income_segment_High income)	(Spending_segment_Biggest consumer)	0.159	0.250	0.116	0.728	2.917	0.076	2.759
284	(Children_segment_No child, Education_segment_Medium education)	(Spending_segment_Biggest consumer)	0.144	0.250	0.086	0.594	2.379	0.050	1.847

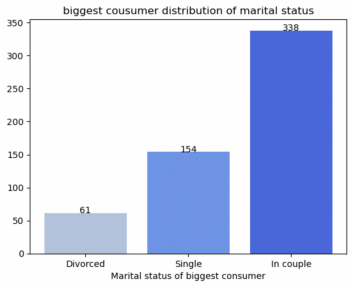
- 消費力和收入呈正相關

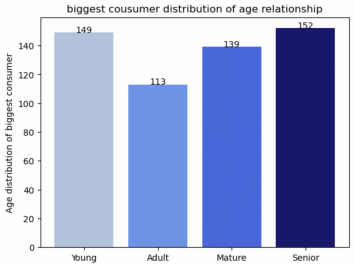


- 消費力和子女數量呈負相關
- 消費力和學歷無特別關係
- 有伴侶的消費者通常花費較多
- 年齡分佈和消費力沒有特別的關係





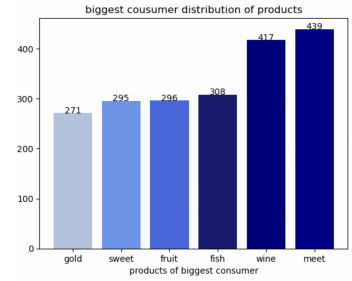


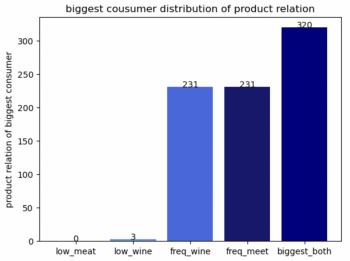


商品與商品之間的關係

我們分析了消費力高的人與商品種類的關係，得到的結論為他們花費最多的商品種類為肉和酒，因此進而分析酒類與肉類和其他商品的關係

- 酒類和肉類為花費比重最高之商品種類





- 同時為酒類以及肉類的最大消費者接近6成，而其中很少買酒或肉的，也不會買另一樣商品
- 肉類的最大消費者，購買菜和魚商品的比例差不多