機器學習觀念與應用 作業一

112703003 資訊二 黃柏淵

第一題：

Confidence 最高前十條：

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(Imagine Low Fat French Fries, CDR Hot Chocolate)	(Quick Extra Lean Hamburger)	0.000185	0.003123	0.000185	1.000000	320.220339
10	(Hilltop 200 MG Acetominifen, Just Right Canne...	(Faux Products HCL Nasal Spray)	0.000159	0.002964	0.000159	1.000000	337.375000
3	(High Top Summer Squash, Even Better Sharp Che...	(High Top New Potatos)	0.000159	0.003732	0.000159	1.000000	267.985816
14	(Cormorant Scented Toilet Tissue, Hilltop 200 ...	(Horatio No Salt Popcorn)	0.000159	0.003335	0.000159	1.000000	299.888889
6	(Booker Low Fat String Cheese, Bravo Fancy Can...	(High Top Oranges)	0.000159	0.003202	0.000159	1.000000	312.280992
11	(Plato French Roast Coffee, High Quality Sciss...	(Dollar Monthly Sports Magazine)	0.000159	0.003202	0.000159	1.000000	312.280992
1	(Quick Extra Lean Hamburger, CDR Hot Chocolate)	(Imagine Low Fat French Fries)	0.000212	0.002991	0.000185	0.875000	292.590708
15	(Cormorant Scented Toilet Tissue, Horatio No S...	(Hilltop 200 MG Acetominifen)	0.000185	0.003520	0.000159	0.857143	243.518797
13	(High Quality Scissors, Dollar Monthly Sports ...	(Plato French Roast Coffee)	0.000185	0.003282	0.000159	0.857143	261.193548
12	(Plato French Roast Coffee, Dollar Monthly Spo...	(High Quality Scissors)	0.000185	0.003229	0.000159	0.857143	265.475410

Lift 最高前十條：

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
10	(Hilltop 200 MG Acetominifen, Just Right Canne...	(Faux Products HCL Nasal Spray)	0.000159	0.002964	0.000159	1.000000	337.375000
0	(Imagine Low Fat French Fries, CDR Hot Chocolate)	(Quick Extra Lean Hamburger)	0.000185	0.003123	0.000185	1.000000	320.220339
6	(Booker Low Fat String Cheese, Bravo Fancy Can...	(High Top Oranges)	0.000159	0.003202	0.000159	1.000000	312.280992
11	(Plato French Roast Coffee, High Quality Sciss...	(Dollar Monthly Sports Magazine)	0.000159	0.003202	0.000159	1.000000	312.280992
14	(Cormorant Scented Toilet Tissue, Hilltop 200 ...	(Horatio No Salt Popcorn)	0.000159	0.003335	0.000159	1.000000	299.888889
1	(Quick Extra Lean Hamburger, CDR Hot Chocolate)	(Imagine Low Fat French Fries)	0.000212	0.002991	0.000185	0.875000	292.590708
8	(BBB Best Tomato Sauce, Imagine Frozen Cheese ...	(Best Choice Apple Fruit Roll)	0.000185	0.002938	0.000159	0.857143	291.783784
4	(High Top New Potatos, Even Better Sharp Chedd...	(High Top Summer Squash)	0.000185	0.003070	0.000159	0.857143	279.206897
3	(High Top Summer Squash, Even Better Sharp Che...	(High Top New Potatos)	0.000159	0.003732	0.000159	1.000000	267.985816
12	(Plato French Roast Coffee, Dollar Monthly Spo...	(High Quality Scissors)	0.000185	0.003229	0.000159	0.857143	265.475410

相同點：

- 規則重疊：按 Confidence 和 Lift 排序的前 10 條規則基本上相同，但排序順序不同。這表示那些置信度高的規則通常也有較高的提升度。

- 高置信度：兩組排序中的大多數規則都有非常高的置信度(1.0 或接近 1.0)，表示這些規則的前件與後件之間有強關聯性。

- 高提升度：所有規則都有極高的提升度值(>240)，表示這些規則的前件出現，顯著增加了後件出現的機率。

- 低支持度：所有規則的支持度都非常低(大約 0.00015~0.00021)，意味著這些關聯雖然強烈，但在資料集中出現的頻率較低。

相異點：

- 排序差異：依 Lift 排序時，"Hilltop 200 MG Acetominifen → Faux Products HCL Nasal Spray"的規則排名第一(Lift=337.375)。按置信度(Confidence)排序時，有 6 個規則的置信度為 1.0，它們的排名取決於資料處理順序。

- 規則組成：Lift 排序強調了前件和後件之間關聯性的規則(相對於它們的獨立出現機率)。Confidence 排序強調了前件出現時後件也出現的機率高的規則，不管它們的獨立出現頻率如何。

第二題：

按 Lift 排序：

前 10 條關聯規則(按 Lift 排序)：

antecedents	consequents	support	confidence	lift
yearly_income=$50K - $70K, occupation=Professional	education=Bachelors Degree	0.097	0.95	3.73
homeowner=Y, yearly_income=$50K - $70K, occupation=Professional	education=Bachelors Degree	0.054	0.95	3.71
yearly_income=$50K - $70K, occupation=Professional, num_children_at_home=0	education=Bachelors Degree	0.063	0.94	3.70
occupation=Manual, yearly_income=$10K - $30K, num_children_at_home=0	education=Partial High School	0.065	0.97	3.22
homeowner=Y, occupation=Manual, yearly_income=$10K - $30K	education=Partial High School	0.060	0.97	3.22
occupation=Manual, yearly_income=$10K - $30K, gender=F	education=Partial High School	0.050	0.97	3.22
occupation=Manual, yearly_income=$10K - $30K	education=Partial High School	0.101	0.96	3.21
homeowner=Y, yearly_income=$10K - $30K, occupation=Skilled Manual	education=Partial High School	0.056	0.96	3.21
gender=M, occupation=Manual, yearly_income=$10K - $30K	education=Partial High School	0.051	0.96	3.20
occupation=Skilled Manual, yearly_income=$10K - $30K, num_children_at_home=0	education=Partial High School	0.060	0.96	3.19

透過上圖可以得知，高收入（$50K-$70K）、專業職業、無子女的顧客多為學士學位的 lift ≈ 3.7；低收入（$10K-$30K）、體力勞動職業的顧客多為部分高中教育的 lift ≈ 3.2。支持度：5.1%-10.1%，表示規則覆蓋少數但顯著的顧客群。置信度：0.94-0.97，顯示規則高度可靠。房屋擁有（homeowner=Y）和高收入相關，性別影響較小。因此，可以推論出教育程度與收入、職業強相關。

第三題：

1. **12 月平均交易數量高於前 11 個月的平均：**
    - 12 月交易數量為 18325，前 11 個月平均交易為 14960，顯示在 12 月期間，顧客更為願意進行消費。

```
12 月交易數量: 18325
1-11 月平均交易數量: 14959.82
12 月交易占比: 10.02%
```

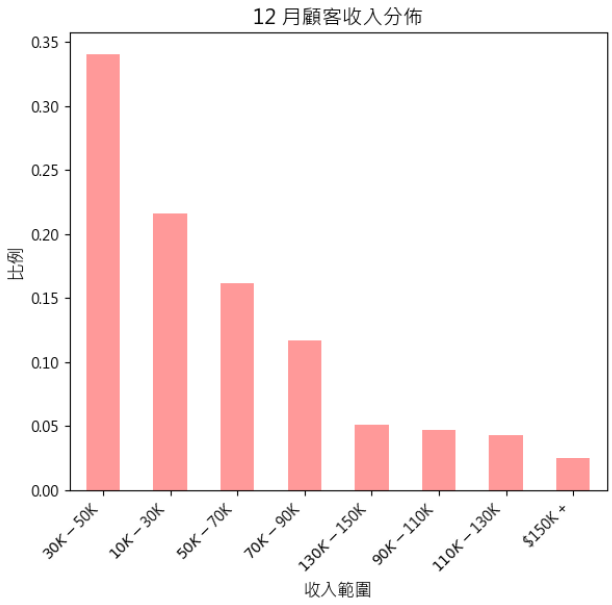
2. **熱門產品類型差異：**
    - 12 月偏向節日相關商品：熱門產品包括 "American Sliced Ham"（火腿）、"Urban Large Eggs"（雞蛋）、"Super Grape Jam"（果醬），這些與節日烹飪或烘焙密切相關。此外，"Hilltop 200 MG Ibuprofen"（止痛藥）進入前 10，可能反映節日壓力或活動增加的需求。

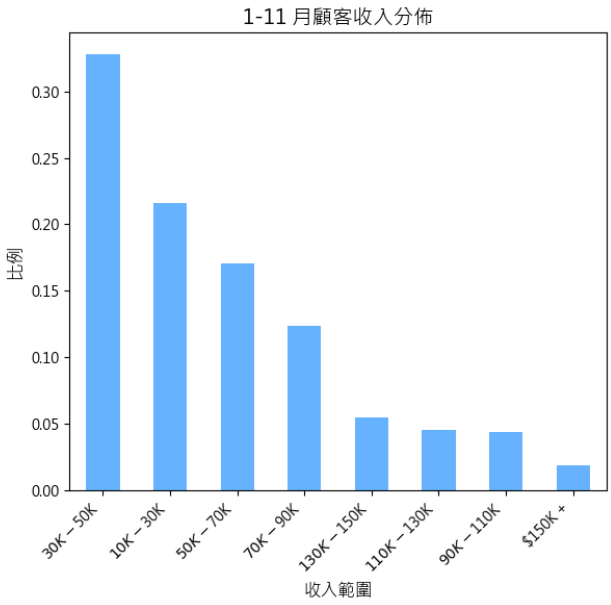
    - 1-11 月偏向日常零食與便利品：熱門產品以 "Great English Muffins"（鬆餅）、"Carrington Ice Cream"（冰淇淋）、"Nationeel Fudge Brownies"（布朗尼）為主，屬於日常零食或即食產品。

```
12 月熱門產品 (前 10):
product_name
Hilltop 200 MG Ibuprofen            25
American Sliced Ham                 23
Booker Low Fat Cottage Cheese       23
Super Grape Jam                     23
Moms Roasted Chicken                22
Landslide Vegetable Oil             22
Urban Large Eggs                    21
Sunset 75 Watt Lightbulb            21
Hermanos Limes                      21
Top Measure White Zinfandel Wine    21
Name: count, dtype: int64

1-11 月熱門產品 (前 10):
product_name
Great English Muffins           13.00
Carrington Ice Cream            12.73
Nationeel Dried Apples          12.55
Nationeel Fudge Brownies        12.55
Ebony Mixed Nuts                12.45
Excellent Orange Juice          12.45
Booker String Cheese            12.45
Steady Childrens Cold Remedy    12.36
Moms Roasted Chicken            12.27
Super Chunky Peanut Butter      12.18
Name: count, dtype: float64
```

3. **顧客收入分佈相似：**
   - 主要顧客群：
     i. "$30K - $50K" 在 12 月（34.02%）和 1-11 月（32.79%）均為最大群體，顯示中收入顧客是全年消費的主力。"$10K - $30K" 緊隨其後，比例幾乎相同（12 月 21.60%，1-11 月 21.59%），表明低收入顧客也是穩定的消費基礎。
     ii. 12 月與 1-11 月的顧客收入分佈整體相似，中低收入群體（$10K - $50K）全年穩定佔據約 55%，顯示 FoodMart 的顧客基礎不受節日影響而大幅變動。

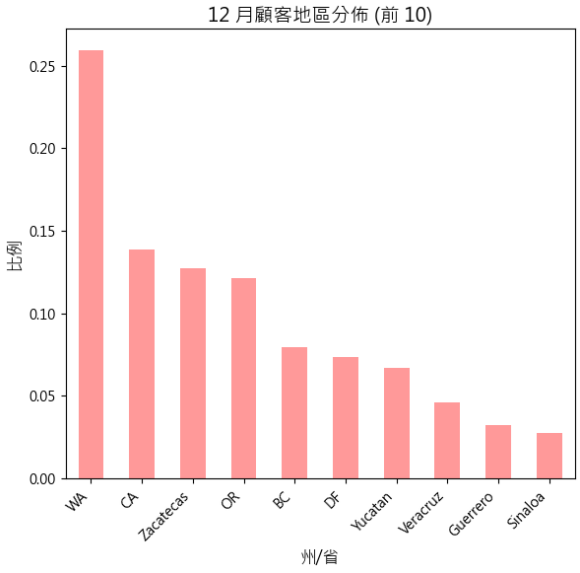
```
12 月顧客收入分佈:
yearly_income
$30K - $50K      0.34
$10K - $30K      0.22
$50K - $70K      0.16
$70K - $90K      0.12
$130K - $150K    0.05
$90K - $110K     0.05
$110K - $130K    0.04
$150K +          0.02
Name: proportion, dtype: float64

1-11 月顧客收入分佈:
yearly_income
$30K - $50K      0.33
$10K - $30K      0.22
$50K - $70K      0.17
$70K - $90K      0.12
$130K - $150K    0.05
$110K - $130K    0.05
$90K - $110K     0.04
$150K +          0.02
Name: proportion, dtype: float64
```

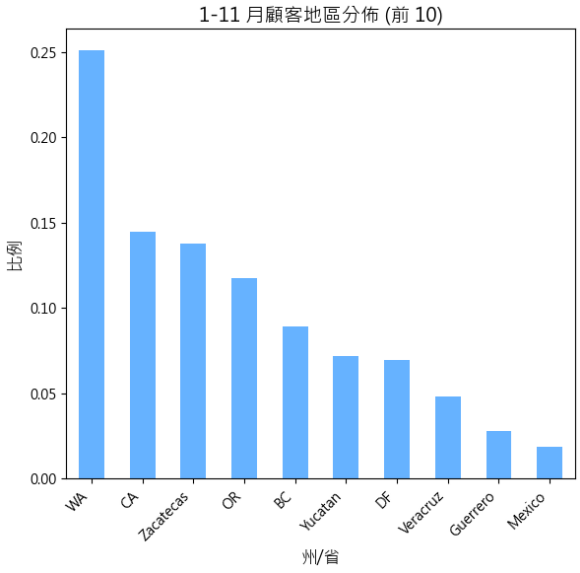




4. **地區分佈穩定：**
   - 分析顯示，業務主要集中在四個關鍵地理區域，這些區域構成了顧客分佈的核心：華盛頓州（WA）、加利福尼亞州（CA）、墨西哥薩卡特卡斯州（Zacatecas）、俄勒岡州（OR)這四個戰略市場共計佔據公司整體客戶地理分佈的約 65%，且該比例在全年各季度中展現出顯著的一致性與穩定性。

```
12 月顧客地區分佈 (前10):
customer_state_province
WA            0.26
CA            0.14
Zacatecas     0.13
OR            0.12
BC            0.08
DF            0.07
Yucatan       0.07
Veracruz      0.05
Guerrero      0.03
Sinaloa       0.03
Name: proportion, dtype: float64

1-11 月顧客地區分佈 (前10):
customer_state_province
WA            0.25
CA            0.14
Zacatecas     0.14
OR            0.12
BC            0.09
Yucatan       0.07
DF            0.07
Veracruz      0.05
Guerrero      0.03
Mexico        0.02
Name: proportion, dtype: float64
```





5. **復購比例的顯著差異**

- 12 月的復購率明顯較高，顯示顧客在聖誕節期間的購物頻次增加。這可能與節日準備有關，例如顧客多次購買以滿足節日需求（如食材、禮物）。1-11 月平均復購率低（7.22%），表明在非節日期間，顧客的購物行為更分散，平均每月交易次數很少超過 1 次，反映日常消費的低頻性。

- 12 月有近 72% 的顧客只交易一次，比例雖高，但遠低於 1-11 月的 93%。這顯示 12 月仍有較多一次性購物者，但相對全年平均，更多顧客選擇多次購買。1-11 月平均的 93% 一次性顧客比例，表明大多數顧客在全年平均每月交易不超過 1 次，顯示日常購物的低忠誠度。

```
12 月顧客數量: 2689
  一次性顧客 (1 次): 1929, 比例: 71.74%
  復購顧客 (>1 次): 760, 比例: 28.26%
1-11 月顧客數量: 7824
  一次性顧客 (平均 ≤ 1 次): 7259, 比例: 92.78%
  復購顧客 (平均 > 1 次): 565, 比例: 7.22%
```