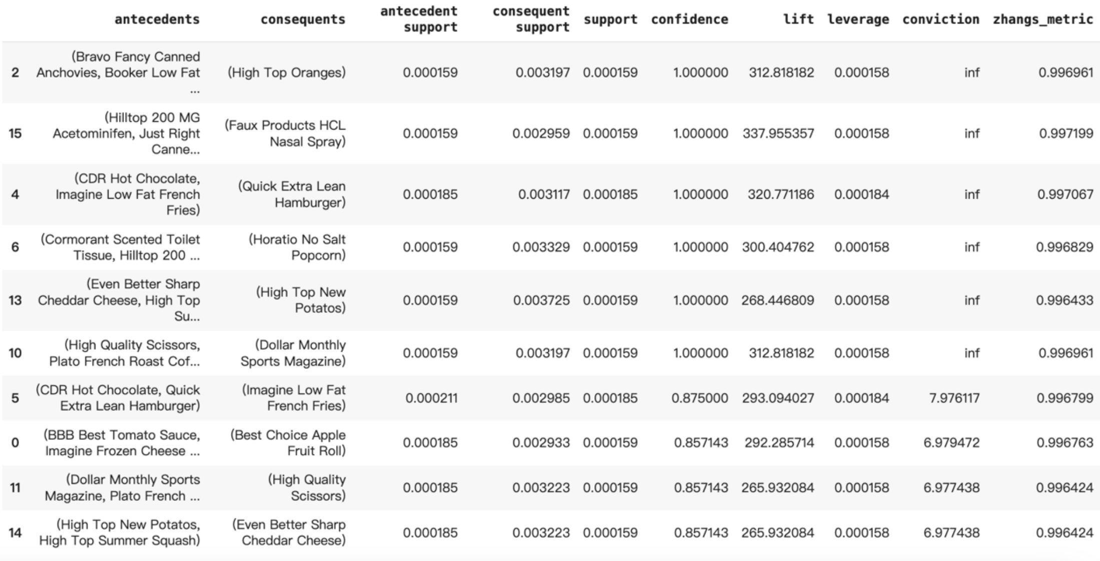
Data Mining Programming HW1　陳品伃　112753204

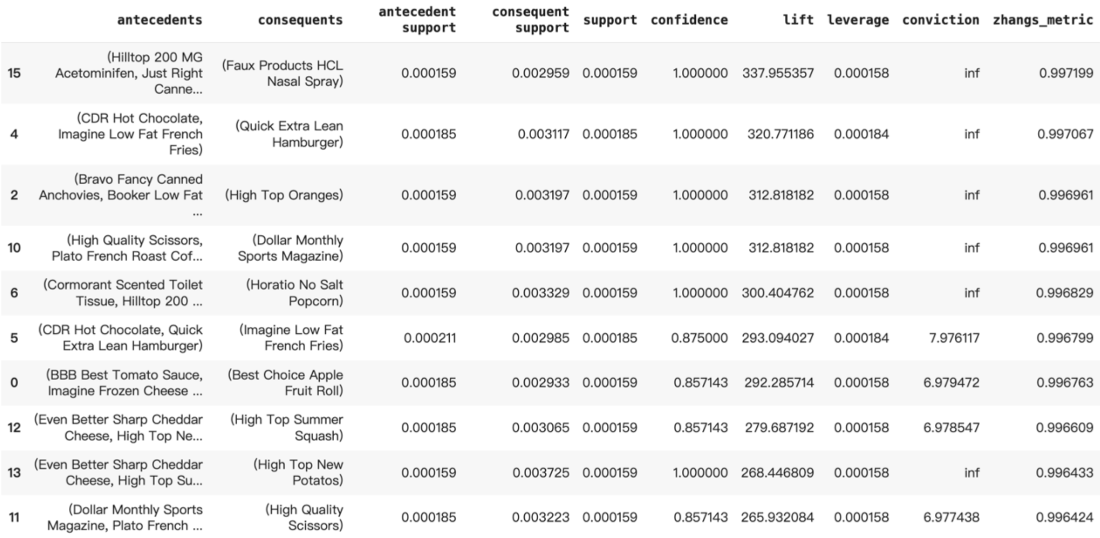
1. 從 Foodmart Data 的交易資料中，探勘符合 Minimum Support = 0.00015 且 Minimum Confidence = 0.8 的 Association Rules，並列出 Confidence 最高的前 10 條 Rules 以及 lift 最高的前 10 條，並比較這兩者的異同。

Confidence Top 10



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
2	(Bravo Fancy Canned Anchovies, Booker Low Fat ...	(High Top Oranges)	0.000159	0.003197	0.000159	1.000000	312.818182	0.000158	inf	0.996961
15	(Hilltop 200 MG Acetominifen, Just Right Canne...	(Faux Products HCL Nasal Spray)	0.000159	0.002959	0.000159	1.000000	337.955357	0.000158	inf	0.997199
4	(CDR Hot Chocolate, Imagine Low Fat French Fries)	(Quick Extra Lean Hamburger)	0.000185	0.003117	0.000185	1.000000	320.771186	0.000184	inf	0.997067
6	(Cormorant Scented Toilet Tissue, Hilltop 200 ...	(Horatio No Salt Popcorn)	0.000159	0.003329	0.000159	1.000000	300.404762	0.000158	inf	0.996829
13	(Even Better Sharp Cheddar Cheese, High Top Su...	(High Top New Potatos)	0.000159	0.003725	0.000159	1.000000	268.446809	0.000158	inf	0.996433
10	(High Quality Scissors, Plato French Roast Cof...	(Dollar Monthly Sports Magazine)	0.000159	0.003197	0.000159	1.000000	312.818182	0.000158	inf	0.996961
5	(CDR Hot Chocolate, Quick Extra Lean Hamburger)	(Imagine Low Fat French Fries)	0.000211	0.002985	0.000185	0.875000	293.094027	0.000184	7.976117	0.996799
0	(BBB Best Tomato Sauce, Imagine Frozen Cheese ...	(Best Choice Apple Fruit Roll)	0.000185	0.002933	0.000159	0.857143	292.285714	0.000158	6.979472	0.996763
11	(Dollar Monthly Sports Magazine, Plato French ...	(High Quality Scissors)	0.000185	0.003223	0.000159	0.857143	265.932084	0.000158	6.977438	0.996424
14	(High Top New Potatos, High Top Summer Squash)	(Even Better Sharp Cheddar Cheese)	0.000185	0.003223	0.000159	0.857143	265.932084	0.000158	6.977438	0.996424

Lift Top 10

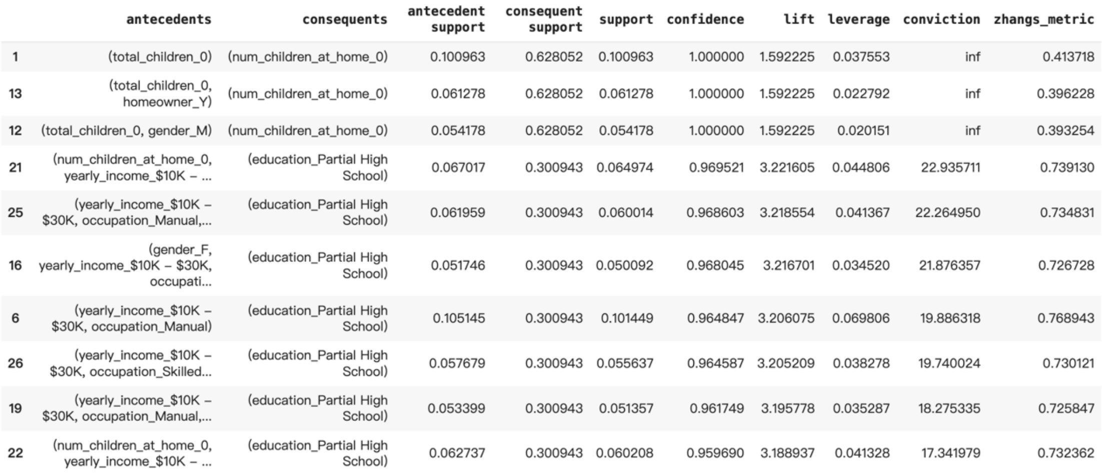


	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
15	(Hilltop 200 MG Acetominifen, Just Right Canne...	(Faux Products HCL Nasal Spray)	0.000159	0.002959	0.000159	1.000000	337.955357	0.000158	inf	0.997199
4	(CDR Hot Chocolate, Imagine Low Fat French Fries)	(Quick Extra Lean Hamburger)	0.000185	0.003117	0.000185	1.000000	320.771186	0.000184	inf	0.997067
2	(Bravo Fancy Canned Anchovies, Booker Low Fat ...	(High Top Oranges)	0.000159	0.003197	0.000159	1.000000	312.818182	0.000158	inf	0.996961
10	(High Quality Scissors, Plato French Roast Cof...	(Dollar Monthly Sports Magazine)	0.000159	0.003197	0.000159	1.000000	312.818182	0.000158	inf	0.996961
6	(Cormorant Scented Toilet Tissue, Hilltop 200 ...	(Horatio No Salt Popcorn)	0.000159	0.003329	0.000159	1.000000	300.404762	0.000158	inf	0.996829
5	(CDR Hot Chocolate, Quick Extra Lean Hamburger)	(Imagine Low Fat French Fries)	0.000211	0.002985	0.000185	0.875000	293.094027	0.000184	7.976117	0.996799
0	(BBB Best Tomato Sauce, Imagine Frozen Cheese ...	(Best Choice Apple Fruit Roll)	0.000185	0.002933	0.000159	0.857143	292.285714	0.000158	6.979472	0.996763
12	(Even Better Sharp Cheddar Cheese, High Top Ne...	(High Top Summer Squash)	0.000185	0.003065	0.000159	0.857143	279.687192	0.000158	6.978547	0.996609
13	(Even Better Sharp Cheddar Cheese, High Top Su...	(High Top New Potatos)	0.000159	0.003725	0.000159	1.000000	268.446809	0.000158	inf	0.996433
11	(Dollar Monthly Sports Magazine, Plato French ...	(High Quality Scissors)	0.000185	0.003223	0.000159	0.857143	265.932084	0.000158	6.977438	0.996424

從結果看來，兩者差異並不大，大部分confidence排序的前幾名也都與lift排序的前幾名相同。

2. 有時候我們有興趣的資料不只有產品間的資訊，也會想要由 User Profile 探勘顧客的基本資料。在給定 Minimum Support = 0.05 且 Minimum Confidence = 0.9 的條件下，探勘 Foodmart 顧客基本資料的屬性 {customer_state_province, yearly_income, gender, total_children, num_children_at_home, education, occupation, homeowner} 間的 Association Rules，並列出10條。

Sorted by Support & Confidence Top10



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
1	(total_children_0)	(num_children_at_home_0)	0.100963	0.628052	0.100963	1.000000	1.592225	0.037553	inf	0.413718
13	(total_children_0, homeowner_Y)	(num_children_at_home_0)	0.061278	0.628052	0.061278	1.000000	1.592225	0.022792	inf	0.396228
12	(total_children_0, gender_M)	(num_children_at_home_0)	0.054178	0.628052	0.054178	1.000000	1.592225	0.020151	inf	0.393254
21	(num_children_at_home_0, yearly_income_$10K – ...	(education_Partial High School)	0.067017	0.300943	0.064974	0.969521	3.221605	0.044806	22.935711	0.739130
25	(yearly_income_$10K – $30K, occupation_Manual,...	(education_Partial High School)	0.061959	0.300943	0.060014	0.968603	3.218554	0.041367	22.264950	0.734831
16	(gender_F, yearly_income_$10K – $30K, occupati...	(education_Partial High School)	0.051746	0.300943	0.050092	0.968045	3.216701	0.034520	21.876357	0.726728
6	(yearly_income_$10K – $30K, occupation_Manual)	(education_Partial High School)	0.105145	0.300943	0.101449	0.964847	3.206075	0.069806	19.886318	0.768943
26	(yearly_income_$10K – $30K, occupation_Skilled...	(education_Partial High School)	0.057679	0.300943	0.055637	0.964587	3.205209	0.038278	19.740024	0.730121
19	(yearly_income_$10K – $30K, occupation_Manual,...	(education_Partial High School)	0.053399	0.300943	0.051357	0.961749	3.195778	0.035287	18.275335	0.725847
22	(num_children_at_home_0, yearly_income_$10K – ...	(education_Partial High School)	0.062737	0.300943	0.060208	0.959690	3.188937	0.041328	17.341979	0.732362

3. 請探勘 Foodmart Data 中，顧客背景資料與其交易資料之間的關係 (Quantitative Association Rules)。例如 80% 女性顧客常買保養品。請自行設定 Minimum Support、Minimum Confidence， 找出 10 條你覺得有意義的 Rules。請說明你的作法及相關參數設定。

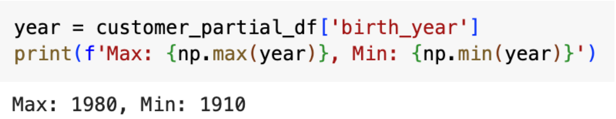
欄位選取

- Transaction：costumer_id、product_id
- Product：product_id、product_name
- Customer：
    - customer_id
    - customer_state_province(省份)
    - yearly_income(年薪)
    - gender(性別)
    - total_children(總共有幾個小孩)
    - birthdate(生日)
    - occupation(職業)

資料處理

- 將birthdate(出生年月日)只提取年份出來，並已10年為一個間距，重新定義。

  - ■ 查看資料分布，顧客出生年介於1910至1980年間

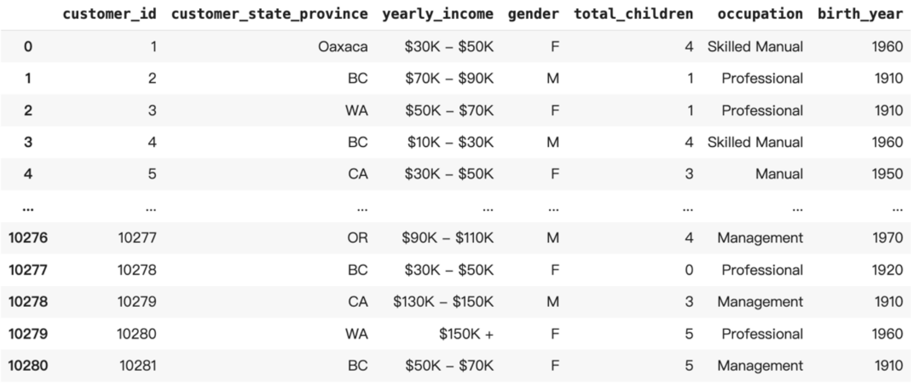


```
year = customer_partial_df['birth_year']
print(f'Max: {np.max(year)}, Min: {np.min(year)}')
```

```
Max: 1980, Min: 1910
```

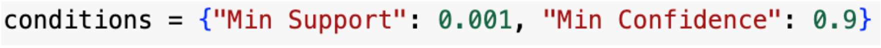
  - ■ 將資料以10年為一個區間來整理

    例：年份為1910~1919的都統一改寫成1910



	customer_id	customer_state_province	yearly_income	gender	total_children	occupation	birth_year
0	1	Oaxaca	$30K – $50K	F	4	Skilled Manual	1960
1	2	BC	$70K – $90K	M	1	Professional	1910
2	3	WA	$50K – $70K	F	1	Professional	1910
3	4	BC	$10K – $30K	M	4	Skilled Manual	1960
4	5	CA	$30K – $50K	F	3	Manual	1950
...	...	...	...	...	...	...	...
10276	10277	OR	$90K – $110K	M	4	Management	1970
10277	10278	BC	$30K – $50K	F	0	Professional	1920
10278	10279	CA	$130K – $150K	M	3	Management	1910
10279	10280	WA	$150K +	F	5	Professional	1960
10280	10281	BC	$50K – $70K	F	5	Management	1910

参數設定

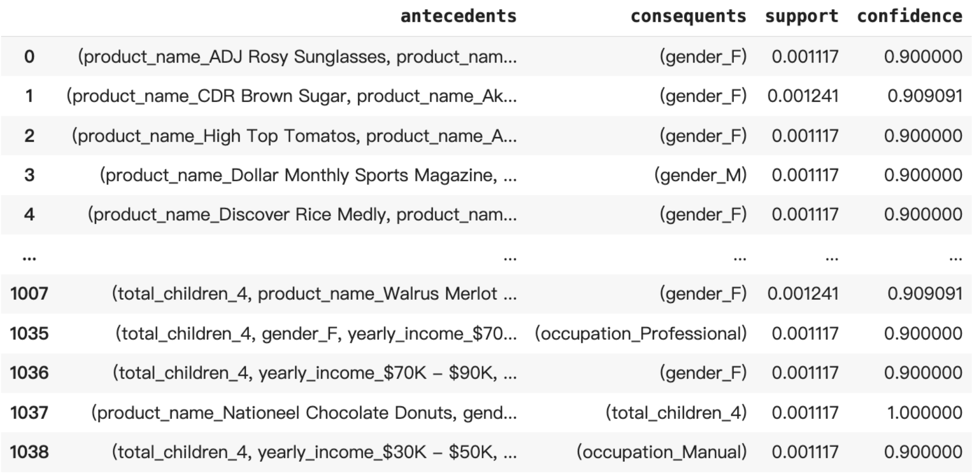


```
conditions = {"Min Support": 0.001, "Min Confidence": 0.9}
```

執行結果

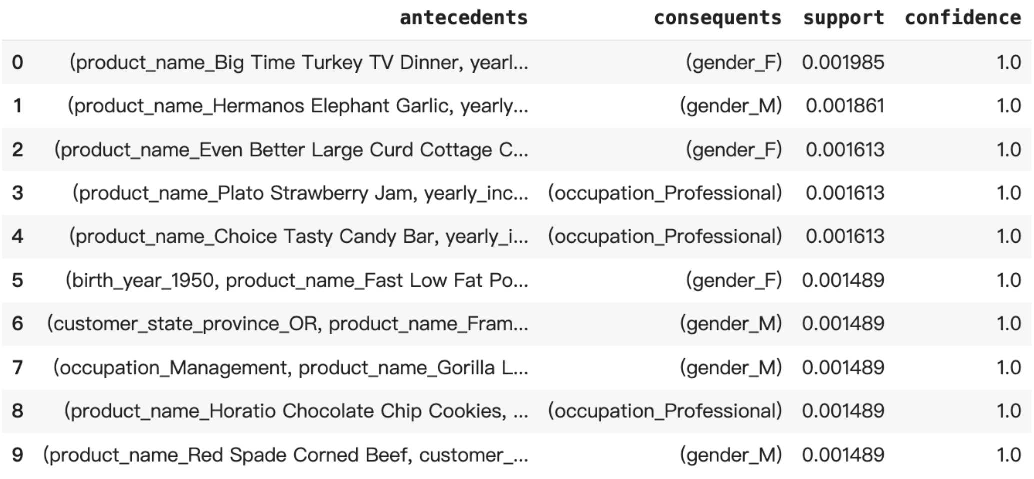
- 將antecedents或consequents中有包含product的留下來

  共1011 rows



	antecedents	consequents	support	confidence
0	(product_name_ADJ Rosy Sunglasses, product_nam...	(gender_F)	0.001117	0.900000
1	(product_name_CDR Brown Sugar, product_name_Ak...	(gender_F)	0.001241	0.909091
2	(product_name_High Top Tomatos, product_name_A...	(gender_F)	0.001117	0.900000
3	(product_name_Dollar Monthly Sports Magazine, ...	(gender_M)	0.001117	0.900000
4	(product_name_Discover Rice Medly, product_nam...	(gender_F)	0.001117	0.900000
...	...	...	...	...
1007	(total_children_4, product_name_Walrus Merlot ...	(gender_F)	0.001241	0.909091
1035	(total_children_4, gender_F, yearly_income_$70...	(occupation_Professional)	0.001117	0.900000
1036	(total_children_4, yearly_income_$70K – $90K, ...	(gender_F)	0.001117	0.900000
1037	(product_name_Nationeel Chocolate Donuts, gend...	(total_children_4)	0.001117	1.000000
1038	(total_children_4, yearly_income_$30K – $50K, ...	(occupation_Manual)	0.001117	0.900000

Sorted by Confidence & Support Top 10



	antecedents	consequents	support	confidence
0	(product_name_Big Time Turkey TV Dinner, yearl...	(gender_F)	0.001985	1.0
1	(product_name_Hermanos Elephant Garlic, yearly...	(gender_M)	0.001861	1.0
2	(product_name_Even Better Large Curd Cottage C...	(gender_F)	0.001613	1.0
3	(product_name_Plato Strawberry Jam, yearly_inc...	(occupation_Professional)	0.001613	1.0
4	(product_name_Choice Tasty Candy Bar, yearly_i...	(occupation_Professional)	0.001613	1.0
5	(birth_year_1950, product_name_Fast Low Fat Po...	(gender_F)	0.001489	1.0
6	(customer_state_province_OR, product_name_Fram...	(gender_M)	0.001489	1.0
7	(occupation_Management, product_name_Gorilla L...	(gender_M)	0.001489	1.0
8	(product_name_Horatio Chocolate Chip Cookies, ...	(occupation_Professional)	0.001489	1.0
9	(product_name_Red Spade Corned Beef, customer_...	(gender_M)	0.001489	1.0

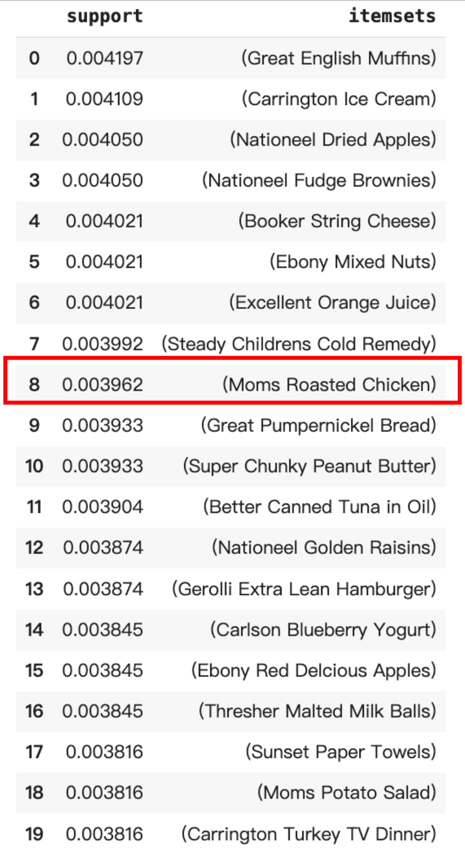
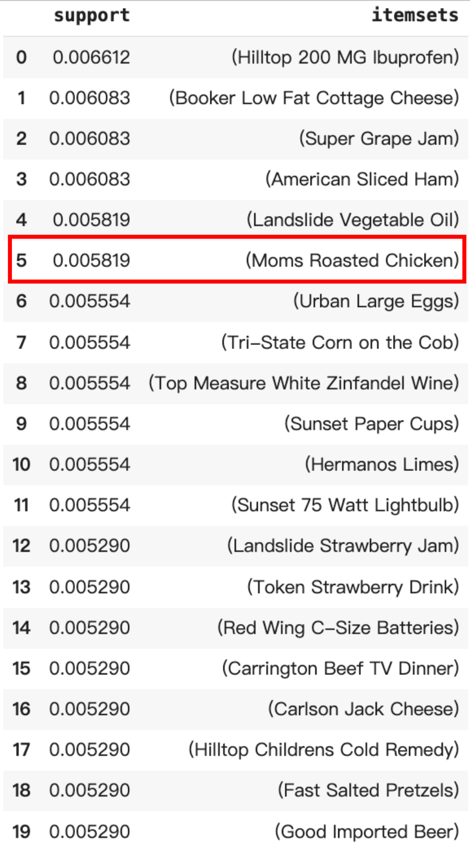
```
[Detail]
Big Time Turkey TV Dinner -> gender_F
Hermanos Elephant Garlic -> gender_M
Even Better Large Curd Cottage Cheese -> gender_F
Plato Strawberry Jam -> occupation_Professional
Choice Tasty Candy Bar -> occupation_Professional
Fast Low Fat Popcorn -> gender_F
Framton City Map -> gender_M
Gorilla Low Fat Cottage Cheese -> gender_M
Horatio Chocolate Chip Cookies -> occupation_Professional
Red Spade Corned Beef -> gender_M
```

4. 在美國由於聖誕節，12月是購物的旺季。請探勘分析比較 12 月與 1~11月的顧客購物行為。 有哪些相似的地方，有哪些差異的地方？

參數設定

```
conditions = {"Min Support": 0.0003, "Min Confidence": 1e-10}
```

執行結果

1-11月	12月
	

```
    support                       itemsets
0   0.004197        (Great English Muffins)
1   0.004109        (Carrington Ice Cream)
2   0.004050        (Nationeel Dried Apples)
3   0.004050     (Nationeel Fudge Brownies)
4   0.004021        (Booker String Cheese)
5   0.004021            (Ebony Mixed Nuts)
6   0.004021       (Excellent Orange Juice)
7   0.003992 (Steady Childrens Cold Remedy)
8   0.003962        (Moms Roasted Chicken)
9   0.003933      (Great Pumpernickel Bread)
10  0.003933     (Super Chunky Peanut Butter)
11  0.003904     (Better Canned Tuna in Oil)
12  0.003874      (Nationeel Golden Raisins)
13  0.003874  (Gerolli Extra Lean Hamburger)
14  0.003845       (Carlson Blueberry Yogurt)
15  0.003845     (Ebony Red Delcious Apples)
16  0.003845     (Thresher Malted Milk Balls)
17  0.003816          (Sunset Paper Towels)
18  0.003816            (Moms Potato Salad)
19  0.003816     (Carrington Turkey TV Dinner)
```

```
    support                            itemsets
0   0.006612             (Hilltop 200 MG Ibuprofen)
1   0.006083       (Booker Low Fat Cottage Cheese)
2   0.006083                    (Super Grape Jam)
3   0.006083                (American Sliced Ham)
4   0.005819             (Landslide Vegetable Oil)
5   0.005819               (Moms Roasted Chicken)
6   0.005554                   (Urban Large Eggs)
7   0.005554            (Tri-State Corn on the Cob)
8   0.005554   (Top Measure White Zinfandel Wine)
9   0.005554                  (Sunset Paper Cups)
10  0.005554                    (Hermanos Limes)
11  0.005554           (Sunset 75 Watt Lightbulb)
12  0.005290           (Landslide Strawberry Jam)
13  0.005290             (Token Strawberry Drink)
14  0.005290          (Red Wing C-Size Batteries)
15  0.005290          (Carrington Beef TV Dinner)
16  0.005290               (Carlson Jack Cheese)
17  0.005290       (Hilltop Childrens Cold Remedy)
18  0.005290               (Fast Salted Pretzels)
19  0.005290                (Good Imported Beer)
```

相同之處只有Moms Roasted Chicken且它的support在12月中有提升一些，其餘皆不同。 但可以注意到的是，12月出現的商品有較多是酒類或季節性飲品，如：第8筆的 Top Measure White Zinfandel Wine、第13筆的Token Strawberry Drink、第19筆的Good Imported Beer…等。