第一題：

請利用 Apriori 演算法，從 Foodmart 資料庫的交易資料中，探勘符合 Minimum Support = 0.0001 且 Minimum Confidence = 0.9 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。若無法跑出結果，請簡述其原因。

先做資料預處理，分別將 sales_fact_1998.csv 和 sales_fact_dec_1998.csv 兩個檔案中 time_id 和 customer_id 皆相同者合併成一筆交易資料，再合併兩個檔案作為 transaction set ，最後藉由 apriori 演算法設定參數去尋找所求。求解過程會跑出錯誤代碼（*numpy.core._exceptions._ArrayMemoryError: Unable to allocate 92.8 GiB for an array with shape (1214461, 2, 41009) and data type bool*），即電腦內存不足，存在內存溢位問題，故無法求解。

第二題：

請利用 FP-Growth 演算法，從 Foodmart 資料庫的交易資料中，探勘符合 Minimum Support = 0.0001 且 Minimum Confidence = 0.9 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。若無法跑出結果，請簡述其原因。

先做資料預處理，分別將 sales_fact_1998.csv 和 sales_fact_dec_1998.csv 兩個檔案中 time_id 和 customer_id 皆相同者合併成一筆交易資料，再合併兩個檔案作為 transaction set ，最後藉由 FP-Growth 演算法設定參數去尋找所求，並將檔案輸出成 CSV 檔比較結果。

Confidence 考慮 antecedents 出現時，出現 antecedents 和 consequents 的機率為何，數值越大越好且最大值為一。Lift 則同時考慮 support 和 confidence。當 lift = 1 時，antecedent 和 consequent 相互獨立；當 lift < 1 antecedent 和 consequent 兩者同時存在機率極小；當 lift > 1，顧客買 antecedent 大機率也會買 consequent，故一般而言越大越好。根據此定義，可以簡單比較輸出結果的異同，我們可以發現 Confidence 高的不一定 Lift 也高，但商品的種類大致相同。

Confidence 最高的前 10 條 Rules。

antecedents	consequents	confidence
frozenset({'655', '1298'})	frozenset({'212'})	1
frozenset({'175', '564'})	frozenset({'171'})	1
frozenset({'175', '564', '968'})	frozenset({'991'})	1
frozenset({'175', '991', '968'})	frozenset({'564'})	1
frozenset({'991', '564', '968'})	frozenset({'175'})	1
frozenset({'175', '564'})	frozenset({'991', '968'})	1
frozenset({'175', '968'})	frozenset({'991', '564'})	1
frozenset({'991', '564'})	frozenset({'175', '968'})	1
frozenset({'564', '968'})	frozenset({'175', '991'})	1
frozenset({'175', '564', '991'})	frozenset({'171'})	1

Lift 最高的前 10 條 Rules。

antecedents	consequents	lift
frozenset({'171', '968'})	frozenset({'991', '564'})	8202.2
frozenset({'175', '171', '968'})	frozenset({'991', '564'})	8202.2
frozenset({'991', '564'})	frozenset({'175', '968'})	8202.2
frozenset({'175', '968'})	frozenset({'991', '564', '171'})	8202.2
frozenset({'175', '564'})	frozenset({'991', '171', '968'})	8202.2
frozenset({'991', '564'})	frozenset({'175', '171', '968'})	8202.2
frozenset({'119', '1367'})	frozenset({'951', '1161'})	8202.2
frozenset({'171', '968'})	frozenset({'991', '564', '175'})	8202.2
frozenset({'991', '564', '175'})	frozenset({'171', '968'})	8202.2
frozenset({'951', '1161'})	frozenset({'119', '1367'})	8202.2

第三題：

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{State_Province, Yearly_Income, Gender, Total_Children,Num_Children_at_Home, Education, Occupation, Houseowner, Num_cars,owned}間的 association rule。(列出 10 條)

從資料庫中讀取所需的資料並尋找 association rule。

	antecedents	consequents	support	confidence	lift
0	frozenset({'houseowner:Y', 'eduation:High School Degree', 'occupation:Skilled Manual'})	frozenset({'yearly_income:$30K - $50K'})	0.06	0.90	2.79
1	frozenset({'gen:M', 'eduation:High School Degree', 'occupation:Skilled Manual'})	frozenset({'yearly_income:$30K - $50K'})	0.05	0.90	2.79
2	frozenset({'occupation:Professional', 'yearly_income:$50K - $70K'})	frozenset({'eduation:Bachelors Degree'})	0.10	0.95	3.73
3	frozenset({'num_childen_at_home:0', 'occupation:Professional', 'yearly_income:$50K - $70K'})	frozenset({'eduation:Bachelors Degree'})	0.06	0.94	3.70
4	frozenset({'houseowner:Y', 'yearly_income:$50K - $70K', 'occupation:Professional'})	frozenset({'eduation:Bachelors Degree'})	0.05	0.95	3.71
5	frozenset({'yearly_income:$10K - $30K'})	frozenset({'eduation:Partial High School'})	0.20	0.93	3.08
6	frozenset({'gen:M', 'yearly_income:$10K - $30K'})	frozenset({'eduation:Partial High School'})	0.10	0.93	3.10
7	frozenset({'gen:M', 'yearly_income:$10K - $30K', 'houseowner:Y'})	frozenset({'eduation:Partial High School'})	0.06	0.94	3.11
8	frozenset({'gen:M', 'yearly_income:$10K - $30K', 'num_childen_at_home:0'})	frozenset({'eduation:Partial High School'})	0.06	0.93	3.10
9	frozenset({'yearly_income:$10K - $30K', 'occupation:Skilled Manual'})	frozenset({'eduation:Partial High School'})	0.10	0.96	3.19

第四題：

請探勘 Foodmart 資料庫中，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。請自行嘗試設定 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。

Step1: 用 pd.read_csv 讀 sales_fact_1998、sales_fact_dec_1998 和 product。

Step2: 用 pd.merge 合併 sales_fact_1998 和 product 中 product_id 相同者。
用 pd.merge 合併 sales_fact_dec_1998 和 product 中 product_id 相同。

Step3: 用 pd.concat 合併 step2 的兩個檔案。

Step4: 讀取 customer.csv 並依 customer_id 和 Step3 的檔案合併。

Step5: 利用 fpgrowth 尋找 association_rule 再匯成 CSV 檔輸出。

Case1: 設 support=0.01 和 confidence = 0.3。從下圖可以發現顧客買新鮮的蔬菜（product_class_id_61）的客群相當多元，不論男女生或是否為家庭主婦（父）皆可能購買，若以學歷來看則以 Partial_High_School 為大宗，年收入約落在 30K-50K。

antecedents	consequents	support	confidence	lift
frozenset({'product_class_id_61'})	frozenset({'houseowner_Y'})	0.046866029	0.607527644	1.004951909
frozenset({'product_class_id_61'})	frozenset({'gender_F'})	0.039090566	0.506733768	0.992747635
frozenset({'product_class_id_61'})	frozenset({'occupation_Professional'})	0.024518408	0.317833853	0.981666023
frozenset({'product_class_id_61'})	frozenset({'num_cars_owned_2'})	0.023364665	0.3028778	1.000726286
frozenset({'product_class_id_61'})	frozenset({'education_Partial High School'})	0.023599788	0.305925716	1.005094992
frozenset({'product_class_id_61'})	frozenset({'gender_M'})	0.038051651	0.493266232	1.00756155
frozenset({'product_class_id_61'})	frozenset({'houseowner_N'})	0.030276188	0.392472356	0.992430202
frozenset({'product_class_id_61'})	frozenset({'yearly_income_$30K - $50K'})	0.024923038	0.323079104	0.981489631

NOTE:上述參數設定下，我們得出的關係大多與蔬菜有關，因此我們放寬參數繼續討論。

Case2: 設 support=0.01 和 confidence = 0.1。放寬條件（下圖）之後可以發現 product_class_id_99（水果）的關係度被列出來，不論性別或是否為家庭主婦（父）皆會購買水果。

antecedents	consequents	support	confidence	lift
frozenset({'product_class_id_99'})	frozenset({'houseowner_Y'})	0.026771214	0.601769912	0.995427661
frozenset({'product_class_id_99'})	frozenset({'gender_F'})	0.022604616	0.508112094	0.995447929
frozenset({'product_class_id_99'})	frozenset({'gender_M'})	0.021882843	0.491887906	1.004746137
frozenset({'product_class_id_99'})	frozenset({'occupation_Professional'})	0.01456669	0.327433628	1.011316021
frozenset({'product_class_id_99'})	frozenset({'houseowner_N'})	0.017716245	0.398230088	1.006989565
frozenset({'product_class_id_99'})	frozenset({'yearly_income_$30K - $50K'})	0.014561222	0.327310718	0.99434495
frozenset({'gender_F', 'product_class_id_99'})	frozenset({'houseowner_Y'})	0.013560582	0.599903241	0.992339877
frozenset({'product_class_id_99', 'houseowner_Y'})	frozenset({'gender_F'})	0.013560582	0.506535948	0.992360083
frozenset({'product_class_id_99'})	frozenset({'gender_F', 'houseowner_Y'})	0.013560582	0.304818092	0.996835778
frozenset({'gender_M', 'product_class_id_99'})	frozenset({'houseowner_Y'})	0.013210632	0.603698151	0.99861729
frozenset({'product_class_id_99', 'houseowner_Y'})	frozenset({'gender_M'})	0.013210632	0.493464052	1.007965625

結論：根據 Case1 和 Case2，我們可以發現這些資料的客群相當注重健康，經過 FP-Tree 得出的結果多和蔬菜和水果有關，其中又以新鮮蔬果（Product_class_id_61）為主。

第五題：

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？

在 support = 0.0001 和 confidence = 0.9 情況下比較。

整體而言，相同處和相異處列舉如下：

	1 到 11 月	**12 月**
相異	Lift 排序前 10 名商品相關度高	Lift 排序前 10 名商品相關度低
	Confidence 排序下香料佔比高	Confidence 排序下香料佔比低
	Confidence 下{1236,414}→{271}佔比高	Confidence 下{1236,414}→{271}佔比低
相同	Confidence 最高值是一	Confidence 最高值是一

以 Lift 前 10 名為例，可以發現顧客購買商品皆不同，1 到 11 月以商品編號{1236,1436,414,1512}→{271,287,244}為主，12 月則無固定之關係。

antecedents	consequents	support	confidence	lift
frozenset({'787', '26'})	frozenset({'49', '1180', '138'})	0.000241429	1	4142
frozenset({'1236', '1436', '414', '271', '287', '1340'})	frozenset({'1512', '244', '1024', '237', '404', '1186', '1446'})	0.000241429	1	4142
frozenset({'1236', '1436', '414', '1512', '244', '404'})	frozenset({'271', '287', '1024', '237', '1340', '1186', '1446'})	0.000241429	1	4142
frozenset({'1236', '1436', '414', '1512', '244', '1186'})	frozenset({'271', '287', '1024', '1446', '237', '1340', '404'})	0.000241429	1	4142
frozenset({'1236', '1436', '414', '1512', '244', '1446'})	frozenset({'271', '287', '1024', '237', '1340', '1186', '404'})	0.000241429	1	4142
frozenset({'1236', '1436', '414', '1512', '1024', '1340'})	frozenset({'271', '287', '244', '237', '404', '1186', '1446'})	0.000241429	1	4142
frozenset({'1236', '1436', '414', '1512', '1024', '237'})	frozenset({'271', '287', '244', '1446', '1340', '1186', '404'})	0.000241429	1	4142
frozenset({'1236', '1436', '414', '1512', '1024', '404'})	frozenset({'271', '287', '244', '237', '1340', '1186', '1446'})	0.000241429	1	4142
frozenset({'1236', '1436', '414', '1512', '1024', '1186'})	frozenset({'271', '287', '244', '1446', '237', '1340', '404'})	0.000241429	1	4142
frozenset({'1236', '1436', '414', '1512', '1024', '1446'})	frozenset({'271', '287', '244', '237', '1340', '1186', '404'})	0.000241429	1	4142

antecedents	consequents	support	confidence	lift
frozenset({'1284', '1534'})	frozenset({'232', '475', '708'})	0.000108492	1	9217.25
frozenset({'1161', '1151'})	frozenset({'768', '1367', '951'})	0.000108492	1	9217.25
frozenset({'1284', '1534', '708'})	frozenset({'232', '475'})	0.000108492	1	9217.25
frozenset({'232', '475'})	frozenset({'1284', '1534', '708'})	0.000108492	1	9217.25
frozenset({'232', '708'})	frozenset({'1284', '475', '1534'})	0.000108492	1	9217.25
frozenset({'768', '1161'})	frozenset({'1151', '951'})	0.000108492	1	9217.25
frozenset({'1151', '951'})	frozenset({'768', '1161'})	0.000108492	1	9217.25
frozenset({'768', '1161'})	frozenset({'1151', '1367'})	0.000108492	1	9217.25
frozenset({'1151', '1367'})	frozenset({'768', '1161'})	0.000108492	1	9217.25
frozenset({'768', '1161', '951'})	frozenset({'1151', '1367'})	0.000108492	1	9217.25

以 confidence 前 10 名為例，12 月跟 1 到 11 月商品項目雖然不同但至少不會跟 lift 12 月數據一樣較無規則。

antecedents	consequents	support	confidence	lift
frozenset({'324', '1290'})	frozenset({'214'})	0.000241429	1	345.1667
frozenset({'1186', '244'})	frozenset({'287', '1024', '1436', '1236', '94'})	0.000241429	1	4142
frozenset({'1186', '1024', '94'})	frozenset({'1236', '244', '1436', '287'})	0.000241429	1	4142
frozenset({'1236', '1186', '1436'})	frozenset({'1024', '244', '94', '287'})	0.000241429	1	4142
frozenset({'1236', '1436', '94'})	frozenset({'1024', '1186', '244', '287'})	0.000241429	1	4142
frozenset({'1186', '1436', '94'})	frozenset({'1024', '244', '1236', '287'})	0.000241429	1	4142
frozenset({'1236', '1186', '94'})	frozenset({'1024', '244', '1436', '287'})	0.000241429	1	4142
frozenset({'244', '287'})	frozenset({'1024', '1436', '1236', '1186', '94'})	0.000241429	1	4142
frozenset({'1024', '244'})	frozenset({'287', '1436', '1236', '1186', '94'})	0.000241429	1	4142
frozenset({'244', '1436'})	frozenset({'287', '1024', '1236', '1186', '94'})	0.000241429	1	4142

antecedents	consequents	support	confidence	lift
frozenset({'173', '1222'})	frozenset({'872'})	0.000108492	1	279.31061
frozenset({'951', '1151', '119'})	frozenset({'1161', '768'})	0.000108492	1	9217.25
frozenset({'951', '1151', '768'})	frozenset({'1161', '119'})	0.000108492	1	6144.8333
frozenset({'119', '1151', '768'})	frozenset({'1161', '951'})	0.000108492	1	7373.8
frozenset({'1161', '951', '1151'})	frozenset({'119', '768'})	0.000108492	1	7373.8
frozenset({'1161', '1151', '119'})	frozenset({'951', '768'})	0.000108492	1	7373.8
frozenset({'1161', '951', '768'})	frozenset({'1151', '119'})	0.000108492	1	9217.25
frozenset({'1161', '119', '768'})	frozenset({'951', '1151'})	0.000108492	1	9217.25
frozenset({'1161', '1151', '768'})	frozenset({'951', '119'})	0.000108492	1	6144.8333
frozenset({'119', '951', '1151', '768'})	frozenset({'1161'})	0.000108492	1	320.6