TP 1: Les règles d'associations

Emna Belhaj - 3IDL2

1) Créer Un DataFrame en utilisant les données de fichier ‘’market_basket.txt’’ qui contient les identifiants des caddies et les produits associées.
2) Afficher les 10 premières lignes du DataFrame.

	ID	Product
0	1	Peaches
1	2	Vegetable_Oil
2	2	Frozen_Corn
3	3	Plums
4	4	Pancake_Mix
5	5	Cheese
6	6	Cauliflower
7	7	2pct_Milk
8	8	98pct_Fat_Free_Hamburger
9	8	Potato_Chips

3) Afficher les dimensions du dataframe. `(12935, 2)`
4) Écrire un script python qui permet de Construire un table binaire indiquant la présence de chaque produit au niveau des caddies (True:1 si le produit est présent dans le caddie et 0 dans le cas réciproque).
5) Tester la bibliothèque pandas.crosstab pour construire la table binaire et vérifier que vous avez les mêmes résultats de votre script
6) Afficher les 30 premières transactions et les 3 premiers produits

Product	100_Watt_Lightbulb	2pct_Milk	40_Watt_Lightbulb
ID			
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	1	0
8	0	0	0
9	0	0	0
10	0	0	0
11	0	0	0
12	0	0	0
13	0	0	0
14	0	0	0
15	0	0	0
16	0	0	0
17	0	0	0
18	0	0	0
19	0	0	0
20	0	0	0
21	0	0	0
22	0	0	0
23	0	0	0
24	0	0	0
25	0	1	0
26	1	0	0
27	0	0	0
28	0	1	0
29	0	1	0
30	0	0	0

7) Écrire un script python de la fonction a_priori() qui permet l'extraction des itemsets les plus fréquents. (on définit un min_supp=0.025 et un longueur maximum de 4 produits)
8) Afficher les 15 premiers itemsets

	support	itemsets
0	0.030147	(100_Watt_Lightbulb)
1	0.109559	(2pct_Milk)
2	0.037500	(60_Watt_Lightbulb)
3	0.031618	(75_Watt_Lightbulb)
4	0.093382	(98pct_Fat_Free_Hamburger)
5	0.031618	(AA_Cell_Batteries)
6	0.025735	(Apple_Cinnamon_Waffles)
7	0.026471	(Apple_Drink)
8	0.031618	(Apple_Fruit_Roll)
9	0.032353	(Apple_Jam)
10	0.033088	(Apple_Jelly)
11	0.032353	(Apple_Sauce)
12	0.053676	(Apples)
13	0.066912	(Aspirin)
14	0.027941	(Avocado_Dip)

9) Ecrire une fonction is_inclus() qui permet de vérifier si un sous-ensemble items est inclus dans l'ensemble x
10) Afficher les itemsets comprenant le produit 'Aspirin'

	support	itemsets
13	0.066912	(Aspirin)
208	0.034559	(2pct_Milk, Aspirin)
249	0.027941	(Aspirin, 98pct_Fat_Free_Hamburger)
272	0.027206	(Cola, Aspirin)
273	0.025735	(Domestic_Beer, Aspirin)
274	0.038235	(Eggs, Aspirin)
275	0.027206	(Hot_Dogs, Aspirin)
276	0.027206	(Onions, Aspirin)
277	0.027206	(Aspirin, Pepperoni_Pizza_-_Frozen)
278	0.025000	(Popcorn_Salt, Aspirin)
279	0.036029	(Potato_Chips, Aspirin)
280	0.030147	(Potatoes, Aspirin)
281	0.030147	(Sweet_Relish, Aspirin)
282	0.028676	(Toilet_Paper, Aspirin)
283	0.025000	(Tomatoes, Aspirin)
284	0.030882	(Toothpaste, Aspirin)
285	0.025000	(Wheat_Bread, Aspirin)
286	0.041912	(White_Bread, Aspirin)
509	0.025735	(2pct_Milk, Eggs, Aspirin)
510	0.025000	(2pct_Milk, Potato_Chips, Aspirin)
511	0.027206	(2pct_Milk, White_Bread, Aspirin)
552	0.025000	(Eggs, Potato_Chips, Aspirin)
553	0.029412	(Eggs, White_Bread, Aspirin)
554	0.027206	(Potato_Chips, White_Bread, Aspirin)
555	0.025000	(White_Bread, Potatoes, Aspirin)
556	0.025735	(Toothpaste, White_Bread, Aspirin)

11) Afficher les itemsets contenant Aspirin et Eggs

	support	itemsets
274	0.038235	(Eggs, Aspirin)
509	0.025735	(2pct_Milk, Eggs, Aspirin)
552	0.025000	(Eggs, Potato_Chips, Aspirin)
553	0.029412	(Eggs, White_Bread, Aspirin)

12) Nous produisons les règles à partir des itemsets fréquents. Elles peuvent être très nombreuses, nous en limitons la prolifération en définissant un seuil minimal (min_threshold = 0.75) sur une mesure d'intérêt, en l'occurrence la confiance dans notre exemple (metric = ‘’confidence’’).
13) Afficher les 5 premières règles

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Aspirin, 2pct_Milk)	(White_Bread)	0.034559	0.119118	0.027206	0.787234	6.608878	0.023089	4.140147
1	(Bananas, White_Bread)	(2pct_Milk)	0.032353	0.109559	0.025735	0.795455	7.260525	0.022191	4.353268
2	(Bananas, 2pct_Milk)	(White_Bread)	0.031618	0.119118	0.025735	0.813953	6.833190	0.021969	4.734743
3	(Wheat_Bread, Cola)	(2pct_Milk)	0.032353	0.109559	0.025735	0.795455	7.260525	0.022191	4.353268
4	(Popcorn_Salt, 2pct_Milk)	(Eggs)	0.033088	0.122794	0.027206	0.822222	6.695941	0.023143	4.934283

14) Filtrer les règles en affichant celles qui présentent un LIFT supérieur ou égal à 7

	antecedents	consequents	antecedent support	consequent support	...	confidence	lift	leverage	conviction
1	(Bananas, White_Bread)	(2pct_Milk)	0.032353	0.109559	...	0.795455	7.260525	0.022191	4.353268
3	(Wheat_Bread, Cola)	(2pct_Milk)	0.032353	0.109559	...	0.795455	7.260525	0.022191	4.353268
8	(Onions, Wheat_Bread)	(2pct_Milk)	0.034559	0.109559	...	0.829787	7.573897	0.024890	5.231342
10	(Potatoes, Wheat_Bread)	(2pct_Milk)	0.032353	0.109559	...	0.772727	7.053081	0.021455	3.917941
13	(Toothpaste, Wheat_Bread)	(2pct_Milk)	0.034559	0.109559	...	0.808511	7.379694	0.024155	4.650082
16	(Hamburger_Buns, White_Bread)	(98pct_Fat_Free_Hamburger)	0.034559	0.093382	...	0.765957	8.202379	0.023243	3.873730
17	(Wheat_Bread, 98pct_Fat_Free_Hamburger)	(White_Bread)	0.029412	0.119118	...	0.900000	7.555556	0.022967	8.808824
29	(Hot_Dog_Buns, Sweet_Relish)	(Hot_Dogs)	0.036029	0.092647	...	0.836735	9.031422	0.026809	5.557537
35	(Potatoes, Toilet_Paper)	(White_Bread)	0.028676	0.119118	...	0.871795	7.318772	0.021584	6.870882
37	(Toothpaste, Toilet_Paper)	(White_Bread)	0.029412	0.119118	...	0.875000	7.345679	0.022232	7.047059
40	(Eggs, Potato_Chips, White_Bread)	(2pct_Milk)	0.033824	0.109559	...	0.782609	7.143274	0.022765	4.096029
43	(Eggs, White_Bread, Toothpaste)	(2pct_Milk)	0.032353	0.109559	...	0.795455	7.260525	0.022191	4.353268
46	(Potato_Chips, 2pct_Milk, White_Bread)	(Toothpaste)	0.033088	0.079412	...	0.755556	9.514403	0.022372	3.766043
47	(Toothpaste, Potato_Chips, White_Bread)	(2pct_Milk)	0.030147	0.109559	...	0.829268	7.569160	0.021697	5.215441
48	(Toothpaste, Potato_Chips, 2pct_Milk)	(White_Bread)	0.028676	0.119118	...	0.871795	7.318772	0.021584	6.870882
49	(Toothpaste, White_Bread, 2pct_Milk)	(Potato_Chips)	0.033088	0.097794	...	0.755556	7.725982	0.021764	3.690842

[16 rows x 9 columns]

15) Filtrer les règles en affichant celles menant au conséquent {‘2pct_milk’}

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1	(Bananas, White_Bread)	(2pct_Milk)	0.032353	0.109559	0.025735	0.795455	7.260525	0.022191	4.353268
3	(Wheat_Bread, Cola)	(2pct_Milk)	0.032353	0.109559	0.025735	0.795455	7.260525	0.022191	4.353268
7	(Eggs, Wheat_Bread)	(2pct_Milk)	0.040441	0.109559	0.030882	0.763636	6.970104	0.026452	3.767251
8	(Onions, Wheat_Bread)	(2pct_Milk)	0.034559	0.109559	0.028676	0.829787	7.573897	0.024890	5.231342
9	(Toothpaste, Potato_Chips)	(2pct_Milk)	0.037500	0.109559	0.028676	0.764706	6.979866	0.024568	3.784375
10	(Potatoes, Wheat_Bread)	(2pct_Milk)	0.032353	0.109559	0.025000	0.772727	7.053081	0.021455	3.917941
13	(Toothpaste, Wheat_Bread)	(2pct_Milk)	0.034559	0.109559	0.027941	0.808511	7.379694	0.024155	4.650082
40	(Eggs, Potato_Chips, White_Bread)	(2pct_Milk)	0.033824	0.109559	0.026471	0.782609	7.143274	0.022765	4.096029
43	(Eggs, White_Bread, Toothpaste)	(2pct_Milk)	0.032353	0.109559	0.025735	0.795455	7.260525	0.022191	4.353268
47	(Toothpaste, Potato_Chips, White_Bread)	(2pct_Milk)	0.030147	0.109559	0.025000	0.829268	7.569160	0.021697	5.215441