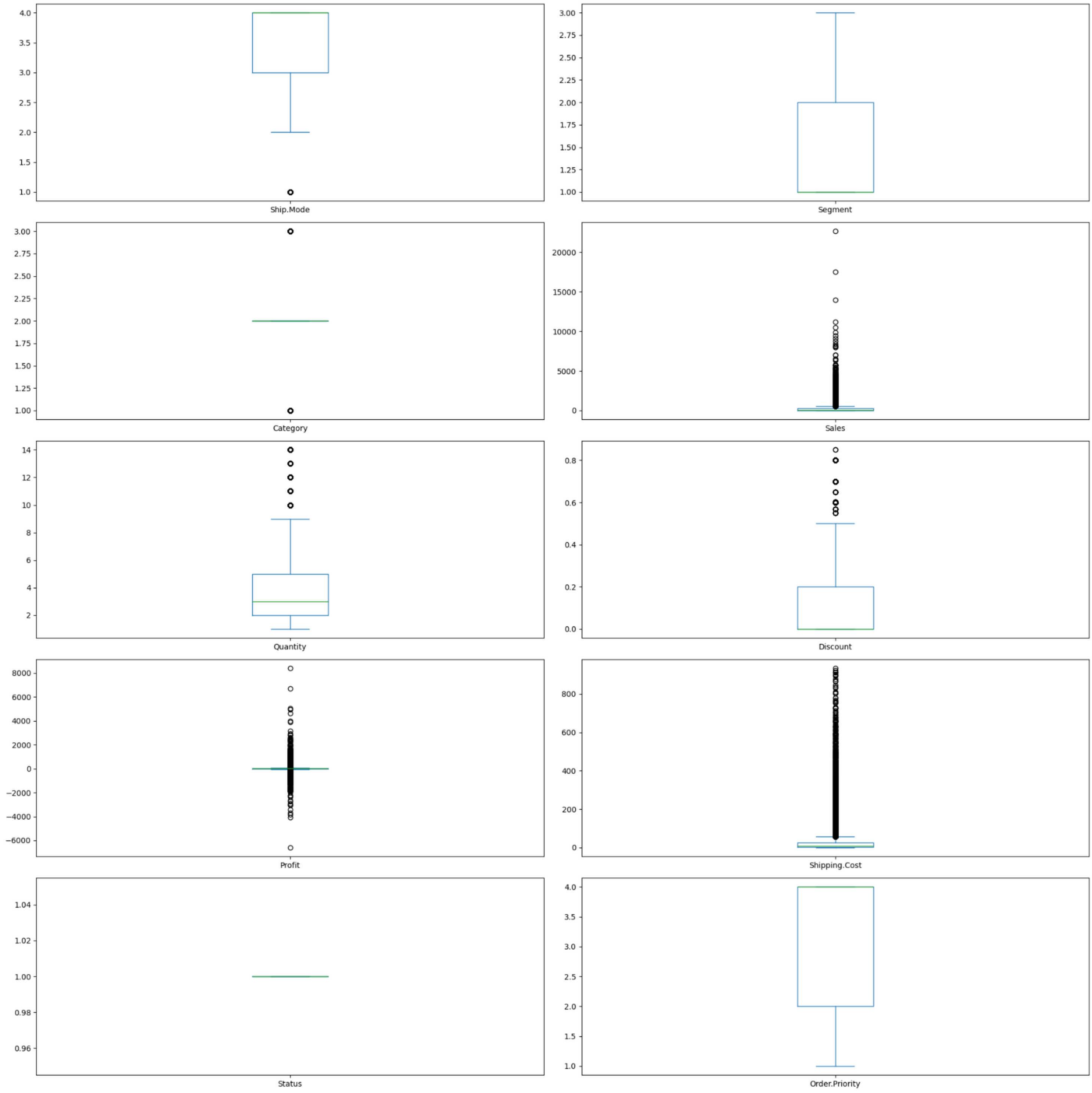
Data Mining Exercise 2

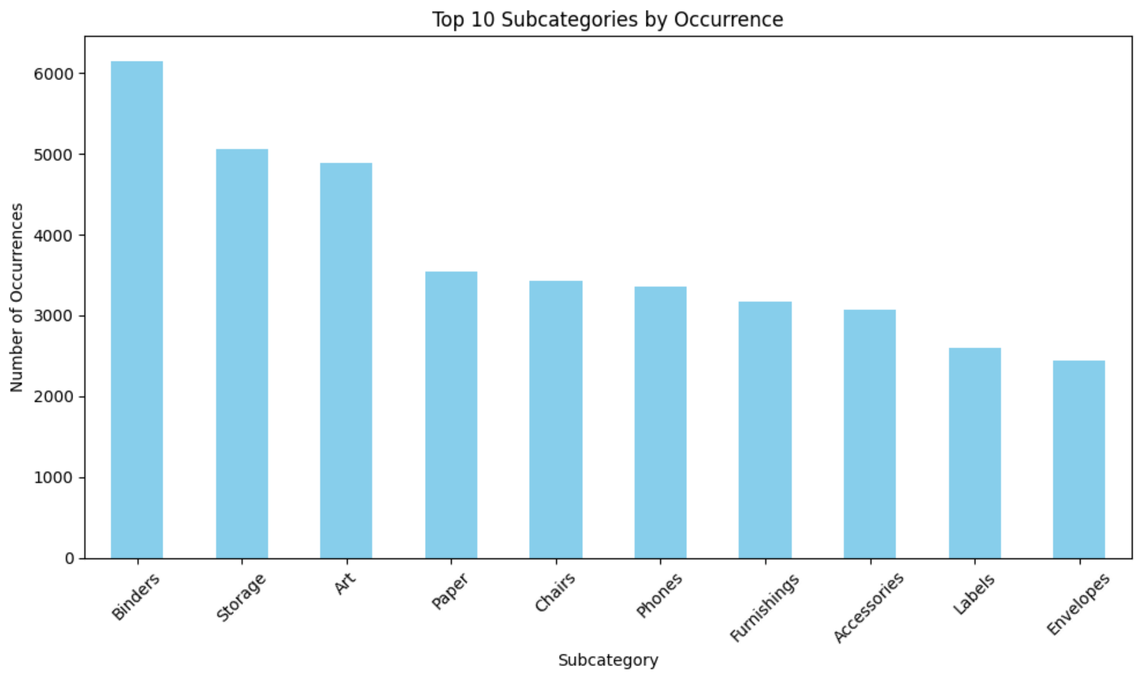
Magdalena König

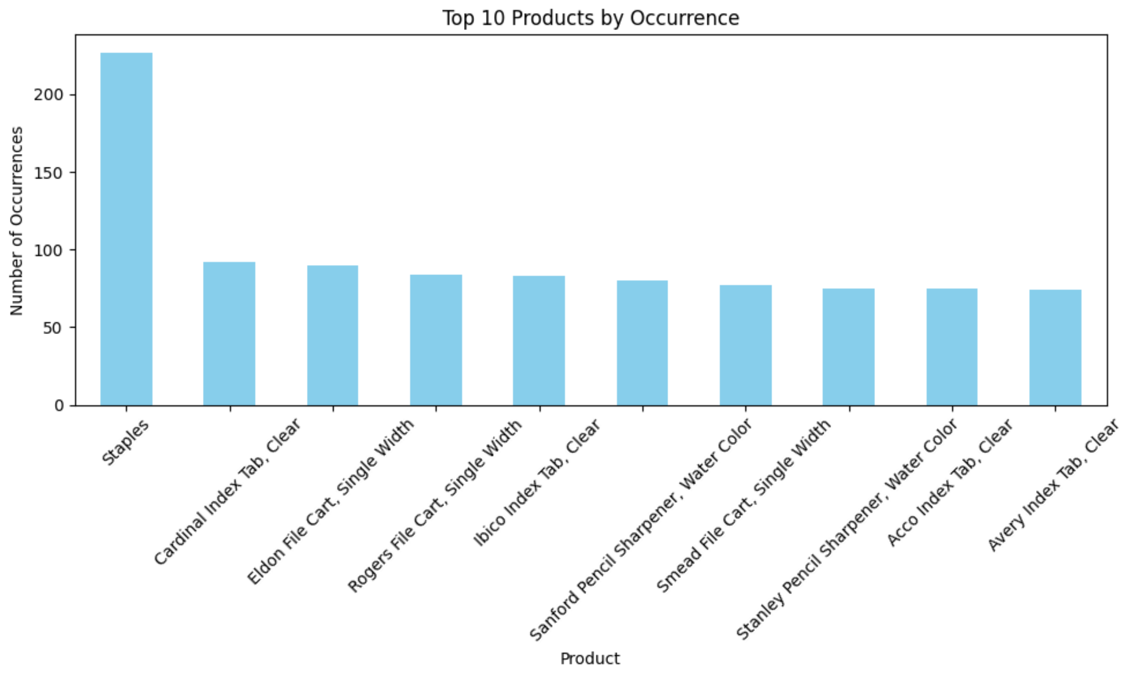
01455794

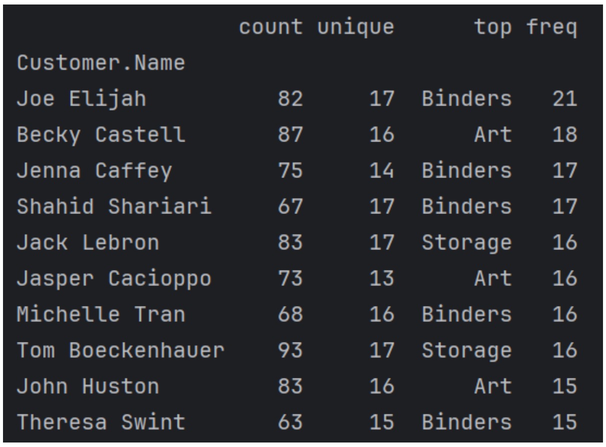
1. Product Orders

- Task 1: Data cleaning & exploration



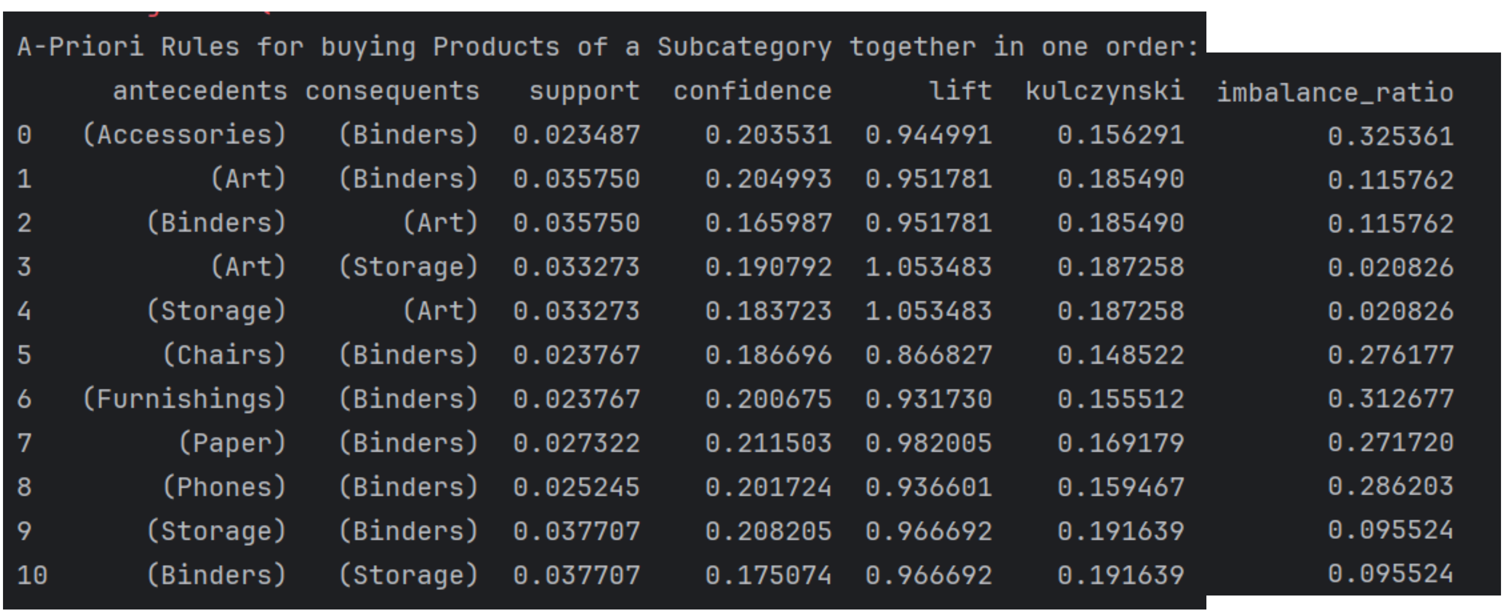






```
                  count unique     top freq
Customer.Name
Joe Elijah           82     17 Binders   21
Becky Castell        87     16     Art   18
Jenna Caffey         75     14 Binders   17
Shahid Shariari      67     17 Binders   17
Jack Lebron          83     17 Storage   16
Jasper Cacioppo      73     13     Art   16
Michelle Tran        68     16 Binders   16
Tom Boeckenhauer     93     17 Storage   16
John Huston          83     16     Art   15
Theresa Swint        63     15 Binders   15
```

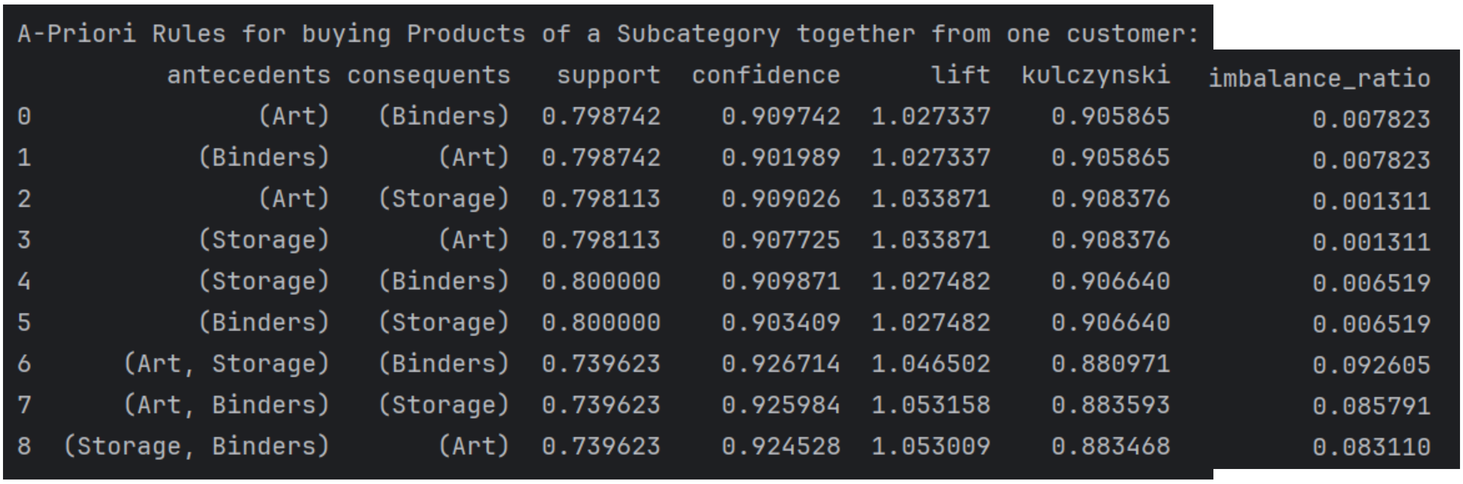
- Task 2 – Association Rules
  - Min_support and min_confidence were generally chosen so that there are 5-10 rules as a result (in a common range of support and confidence → not too many outliers)
    - Frequent purchase of products in a subcategorie per order (min_support=0.023, min_threshold=0.15)



```
A-Priori Rules for buying Products of a Subcategory together in one order:
     antecedents consequents   support  confidence      lift  kulczynski  imbalance_ratio
0  (Accessories)   (Binders)  0.023487    0.203531  0.944991    0.156291         0.325361
1          (Art)   (Binders)  0.035750    0.204993  0.951781    0.185490         0.115762
2      (Binders)       (Art)  0.035750    0.165987  0.951781    0.185490         0.115762
3          (Art)   (Storage)  0.033273    0.190792  1.053483    0.187258         0.020826
4      (Storage)       (Art)  0.033273    0.183723  1.053483    0.187258         0.020826
5       (Chairs)   (Binders)  0.023767    0.186696  0.866827    0.148522         0.276177
6  (Furnishings)   (Binders)  0.023767    0.200675  0.931730    0.155512         0.312677
7        (Paper)   (Binders)  0.027322    0.211503  0.982005    0.169179         0.271720
8       (Phones)   (Binders)  0.025245    0.201724  0.936601    0.159467         0.286203
9      (Storage)   (Binders)  0.037707    0.208205  0.966692    0.191639         0.095524
10     (Binders)   (Storage)  0.037707    0.175074  0.966692    0.191639         0.095524
```

→ Lifts are gernally all under 1 – so purchasing a certain subcategory doesn't increase to purchase the other (except storage and art). The confidence and support is rather small. So is the Kulszynski coefficent – so there is only a very low connection appliceable. The imbalance ratio shows a quite balanced occurance of the subcategories.

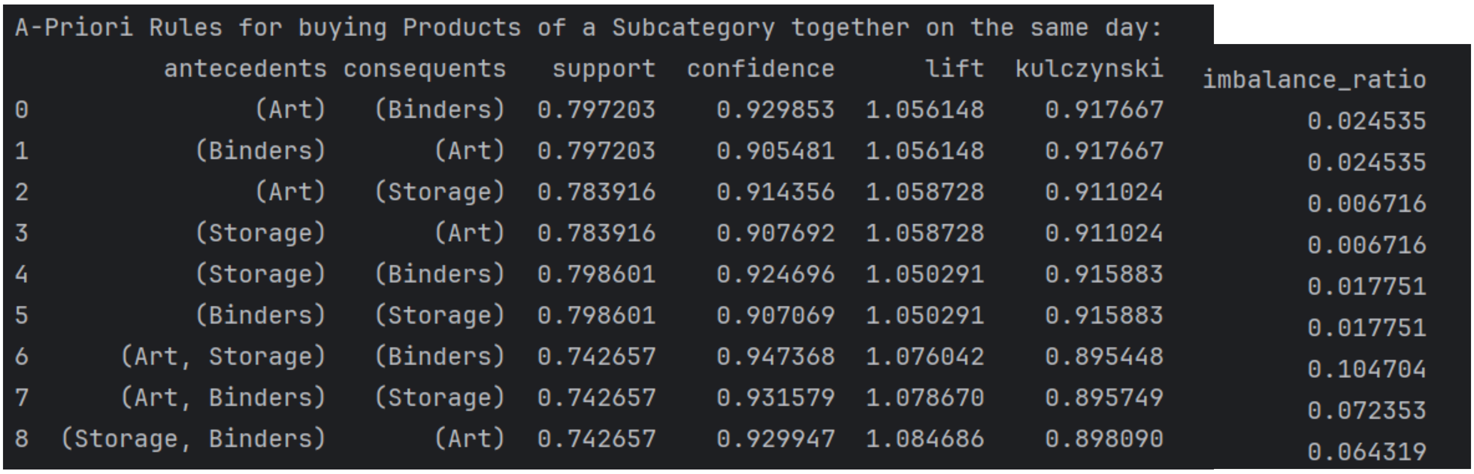
- Frequent purchase of products in a subcategorie per customer (0.7;0.9)



```
A-Priori Rules for buying Products of a Subcategory together from one customer:
          antecedents consequents   support  confidence      lift  kulczynski  imbalance_ratio
0               (Art)   (Binders)  0.798742    0.909742  1.027337    0.905865         0.007823
1           (Binders)       (Art)  0.798742    0.901989  1.027337    0.905865         0.007823
2               (Art)   (Storage)  0.798113    0.909026  1.033871    0.908376         0.001311
3           (Storage)       (Art)  0.798113    0.907725  1.033871    0.908376         0.001311
4           (Storage)   (Binders)  0.800000    0.909871  1.027482    0.906640         0.006519
5           (Binders)   (Storage)  0.800000    0.903409  1.027482    0.906640         0.006519
6      (Art, Storage)   (Binders)  0.739623    0.926714  1.046502    0.880971         0.092605
7      (Art, Binders)   (Storage)  0.739623    0.925984  1.053158    0.883593         0.085791
8  (Storage, Binders)       (Art)  0.739623    0.924528  1.053009    0.883468         0.083110
```

→ Lifts are gernally all over 1 – so purchasing a certain subcategory from one customer does increase the purchase the other. The confidence is high. So is the Kulszynski coefficent – so there is a high connection appliceable. The imbalance ratio shows a very balanced occurance of the subcategories.

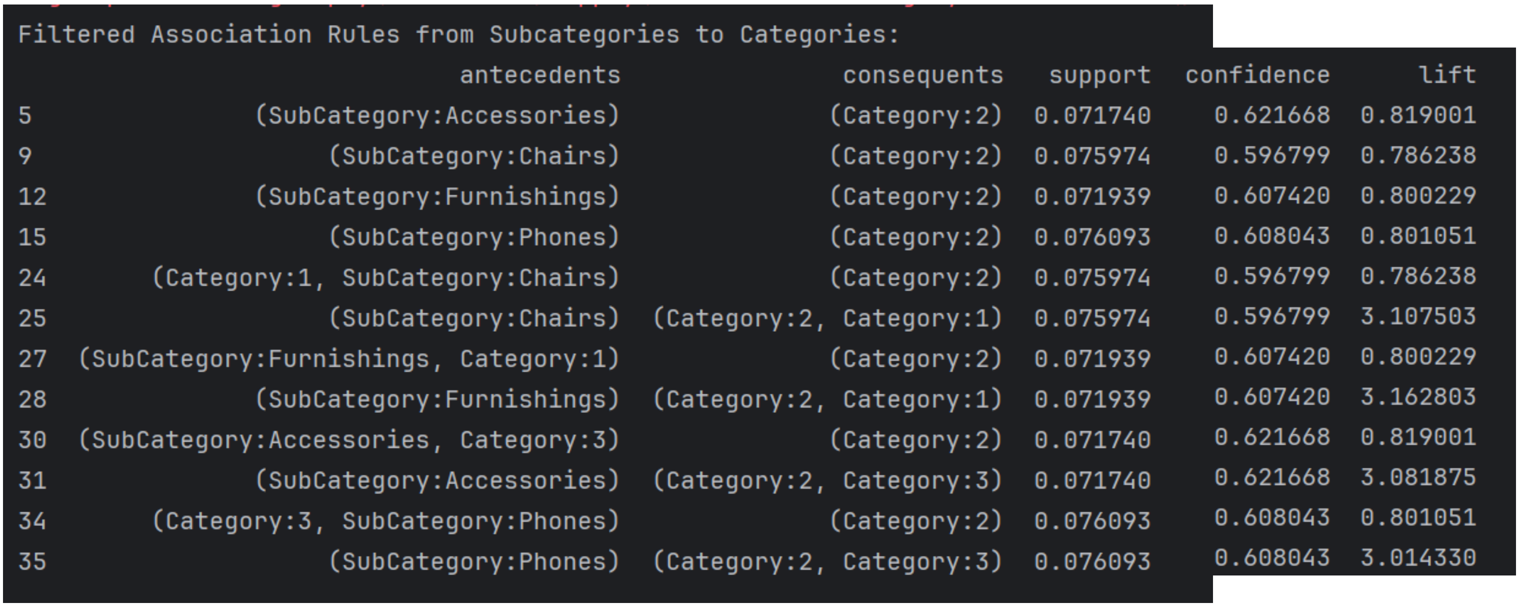
- Products bought together on the same day (0.74;0.9)



```
A-Priori Rules for buying Products of a Subcategory together on the same day:
          antecedents consequents   support  confidence      lift  kulczynski  imbalance_ratio
0               (Art)   (Binders)  0.797203    0.929853  1.056148    0.917667         0.024535
1           (Binders)       (Art)  0.797203    0.905481  1.056148    0.917667         0.024535
2               (Art)   (Storage)  0.783916    0.914356  1.058728    0.911024         0.006716
3           (Storage)       (Art)  0.783916    0.907692  1.058728    0.911024         0.006716
4           (Storage)   (Binders)  0.798601    0.924696  1.050291    0.915883         0.017751
5           (Binders)   (Storage)  0.798601    0.907069  1.050291    0.915883         0.017751
6      (Art, Storage)   (Binders)  0.742657    0.947368  1.076042    0.895448         0.104704
7      (Art, Binders)   (Storage)  0.742657    0.931579  1.078670    0.895749         0.072353
8  (Storage, Binders)       (Art)  0.742657    0.929947  1.084686    0.898090         0.064319
```

➔ -- Lifts are gernally all over 1 – so purchasing a certain subcategory on one day does increase the purchase the other. The confidence is high. So is the Kulszynski coefficent – so there is a high connection appliceable. The imbalance ratio shows a very balanced occurance of the subcategories.

- Task 3: Multiassociation-Rules
  - Association Rules from Subcategories to Categories (0.06, 0.5)

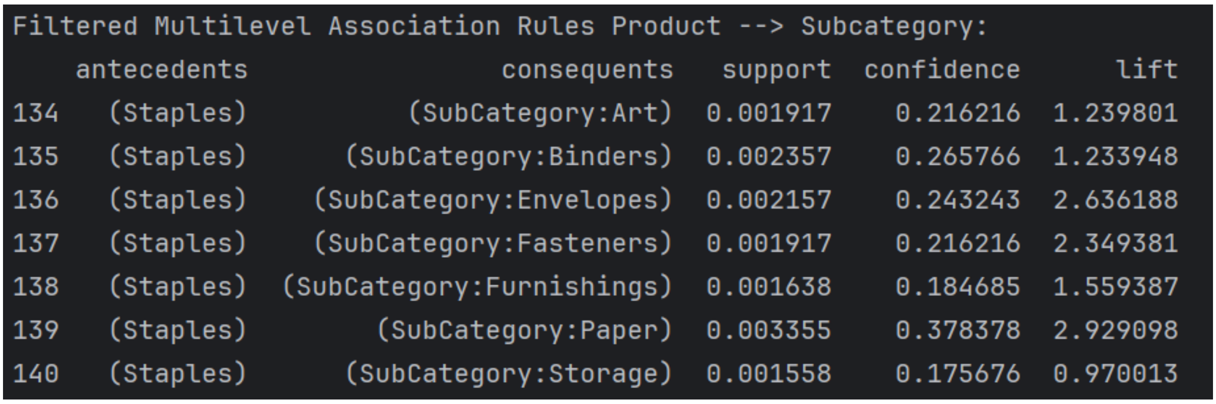


```
Filtered Association Rules from Subcategories to Categories:
                             antecedents                  consequents   support  confidence      lift
5              (SubCategory:Accessories)               (Category:2)  0.071740    0.621668  0.819001
9                   (SubCategory:Chairs)               (Category:2)  0.075974    0.596799  0.786238
12             (SubCategory:Furnishings)               (Category:2)  0.071939    0.607420  0.800229
15                  (SubCategory:Phones)               (Category:2)  0.076093    0.608043  0.801051
24      (Category:1, SubCategory:Chairs)               (Category:2)  0.075974    0.596799  0.786238
25                  (SubCategory:Chairs)  (Category:2, Category:1)  0.075974    0.596799  3.107503
27 (SubCategory:Furnishings, Category:1)               (Category:2)  0.071939    0.607420  0.800229
28             (SubCategory:Furnishings)  (Category:2, Category:1)  0.071939    0.607420  3.162803
30 (SubCategory:Accessories, Category:3)               (Category:2)  0.071740    0.621668  0.819001
31             (SubCategory:Accessories)  (Category:2, Category:3)  0.071740    0.621668  3.081875
34      (Category:3, SubCategory:Phones)               (Category:2)  0.076093    0.608043  0.801051
35                  (SubCategory:Phones)  (Category:2, Category:3)  0.076093    0.608043  3.014330
```

➔ The lift varies greatly, sometimes a subcategory increases the likehood of purchasing a product in a specific category, sometimes it decreases. The confidence is medium.

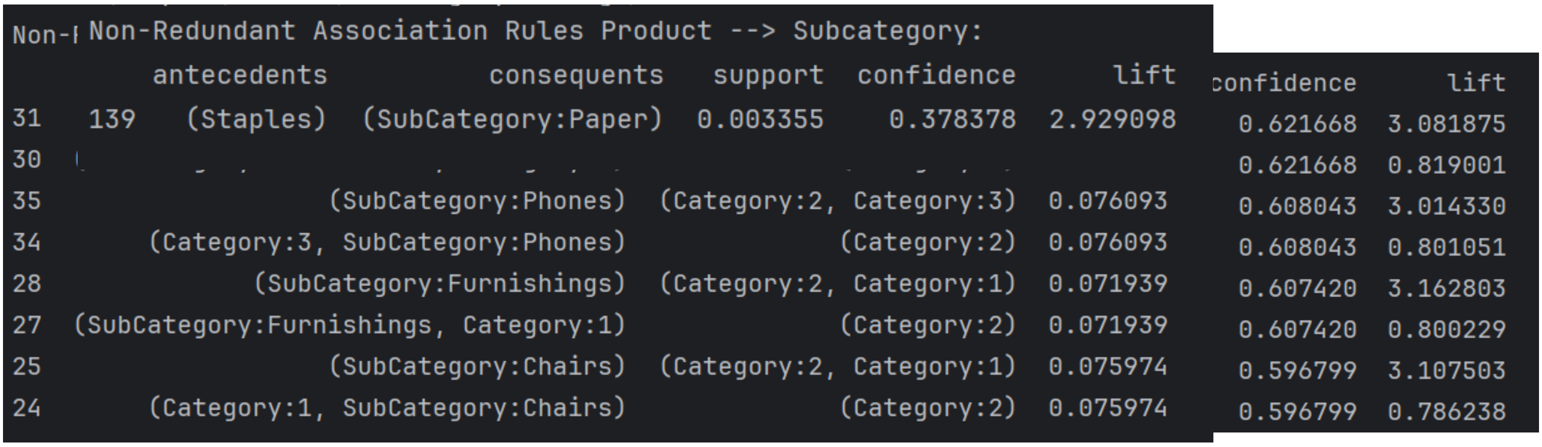
➔ Discovered rules are interesting to see the purchase of which subcategories lead to the purchase of which categegories and how theres two purchases are connected.

  - Association Rules from Products to Subcategories (0.0015, 0.1)



```
Filtered Multilevel Association Rules Product --> Subcategory:
    antecedents                  consequents   support  confidence      lift
134   (Staples)          (SubCategory:Art)  0.001917    0.216216  1.239801
135   (Staples)      (SubCategory:Binders)  0.002357    0.265766  1.233948
136   (Staples)    (SubCategory:Envelopes)  0.002157    0.243243  2.636188
137   (Staples)    (SubCategory:Fasteners)  0.001917    0.216216  2.349381
138   (Staples)  (SubCategory:Furnishings)  0.001638    0.184685  1.559387
139   (Staples)        (SubCategory:Paper)  0.003355    0.378378  2.929098
140   (Staples)      (SubCategory:Storage)  0.001558    0.175676  0.970013
```

➔ Positive Lift (except the last one) – so the product Staples increases the likehood of a purchase in the subcategory art, binders,…

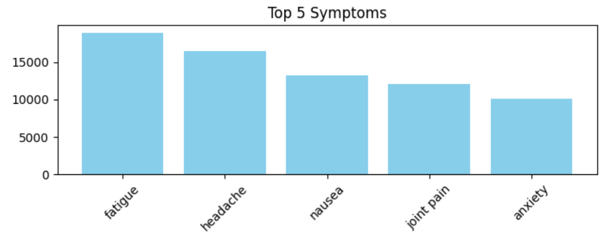


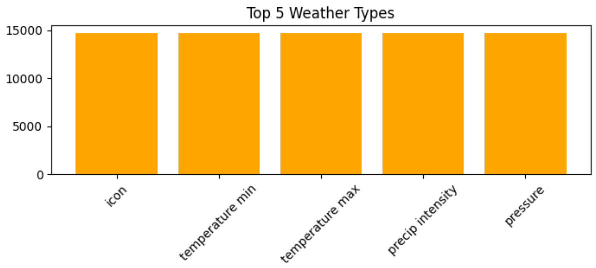
```
Non-I Non-Redundant Association Rules Product --> Subcategory:
        antecedents          consequents   support  confidence      lift
31  139   (Staples)  (SubCategory:Paper)  0.003355    0.378378  2.929098
30
35                 (SubCategory:Phones)  (Category:2, Category:3)  0.076093
34      (Category:3, SubCategory:Phones)              (Category:2)  0.076093
28            (SubCategory:Furnishings)  (Category:2, Category:1)  0.071939
27  (SubCategory:Furnishings, Category:1)              (Category:2)  0.071939
25                 (SubCategory:Chairs)  (Category:2, Category:1)  0.075974
24      (Category:1, SubCategory:Chairs)              (Category:2)  0.075974
```

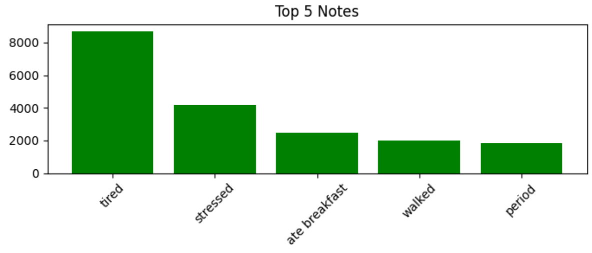
```
confidence      lift
  0.621668  3.081875
  0.621668  0.819001
  0.608043  3.014330
  0.608043  0.801051
  0.607420  3.162803
  0.607420  0.800229
  0.596799  3.107503
  0.596799  0.786238
```

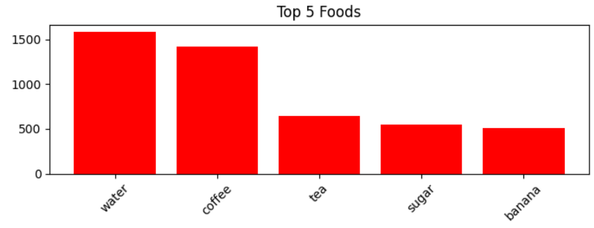
- Discovered rules are interesting to see the purchase of which products lead to the purchase of which subcategories and how theres two purchases are connected. However as seen here, neither the support not the confidence are high. And after the redandency check only one rule remains (all the other consequents were already covered by the SubCategory:Paper and therefore redandent)

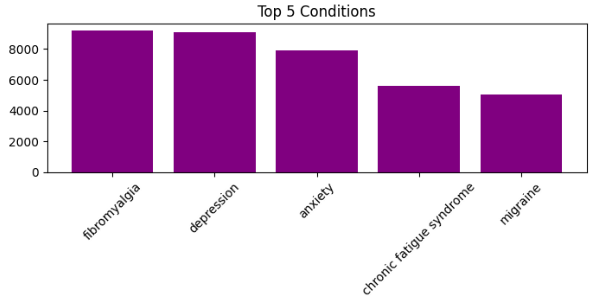
  - Exercise 2 – Health data
    - Task 1: Data cleaning and exploration

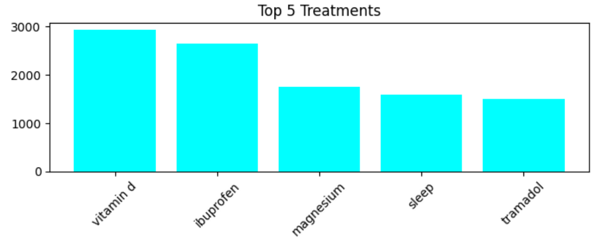


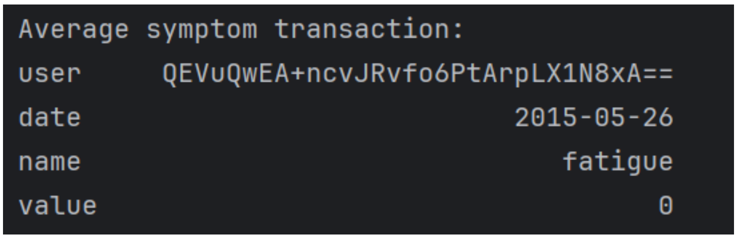




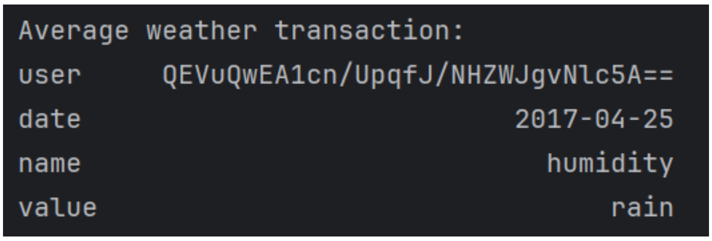




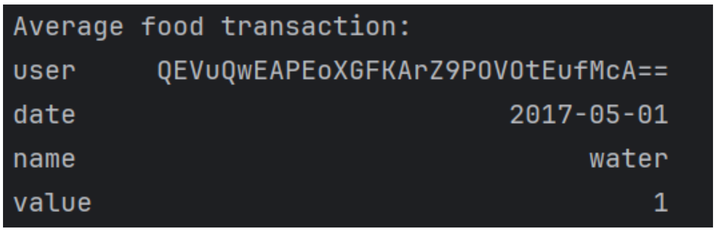




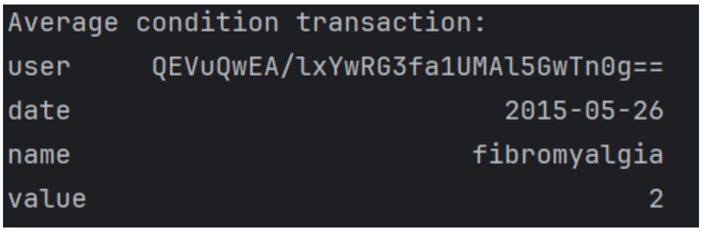
```
Average symptom transaction:
user      QEVuQwEA+ncvJRvfo6PtArpLX1N8xA==
date                            2015-05-26
name                               fatigue
value                                    0
```



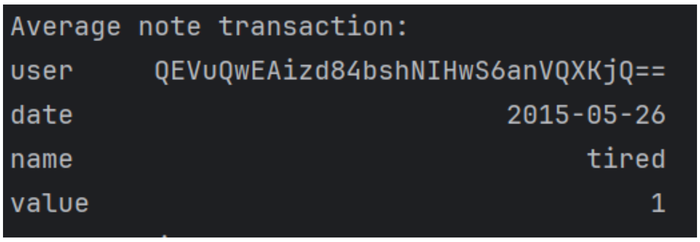
```
Average weather transaction:
user      QEVuQwEA1cn/UpqfJ/NHZWJgvNlc5A==
date                            2017-04-25
name                              humidity
value                                 rain
```



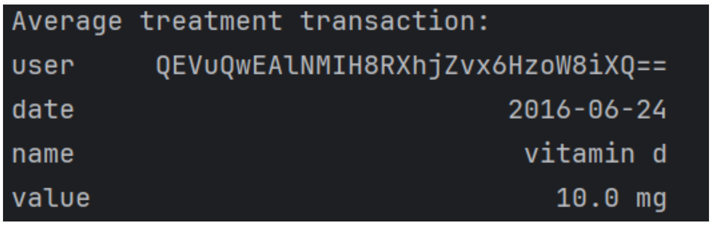
```
Average food transaction:
user      QEVuQwEAPEoXGFKArZ9POVOtEufMcA==
date                            2017-05-01
name                                 water
value                                    1
```



```
Average condition transaction:
user      QEVuQwEA/lxYwRG3fa1UMAl5GwTn0g==
date                            2015-05-26
name                           fibromyalgia
value                                    2
```

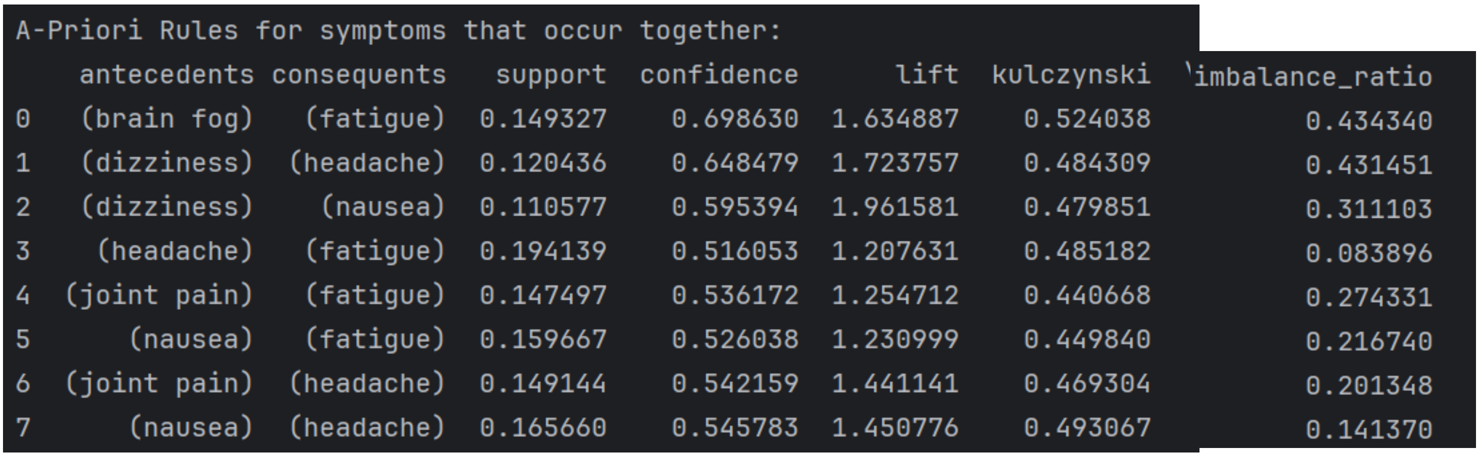


```
Average note transaction:
user      QEVuQwEAizd84bshNIHwS6anVQXKjQ==
date                            2015-05-26
name                                 tired
value                                    1
```



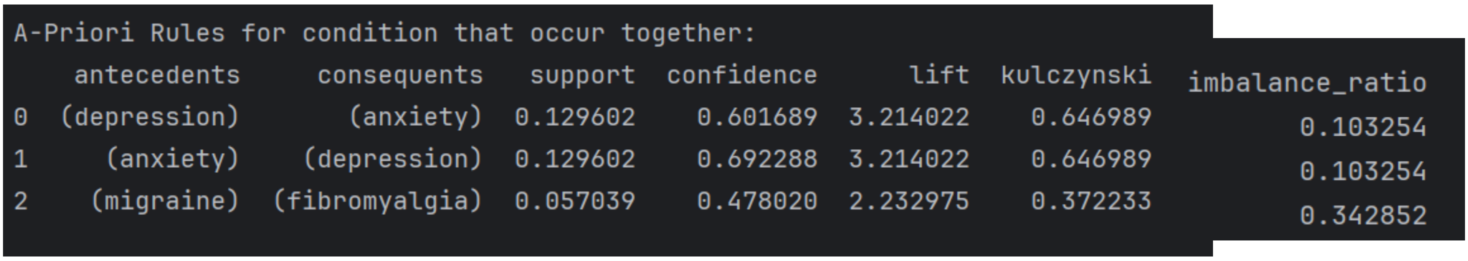
```
Average treatment transaction:
user      QEVuQwEAlNMIH8RXhjZvx6HzoW8iXQ==
date                            2016-06-24
name                             vitamin d
value                              10.0 mg
```

- Task 2 - Association Rules
  - → Min_support and confidence was generally choose to get between 3 and 10 rules in a common range of support and confidence (not too many outliers)
  - Symptoms that occur together (0.1; 0.5)



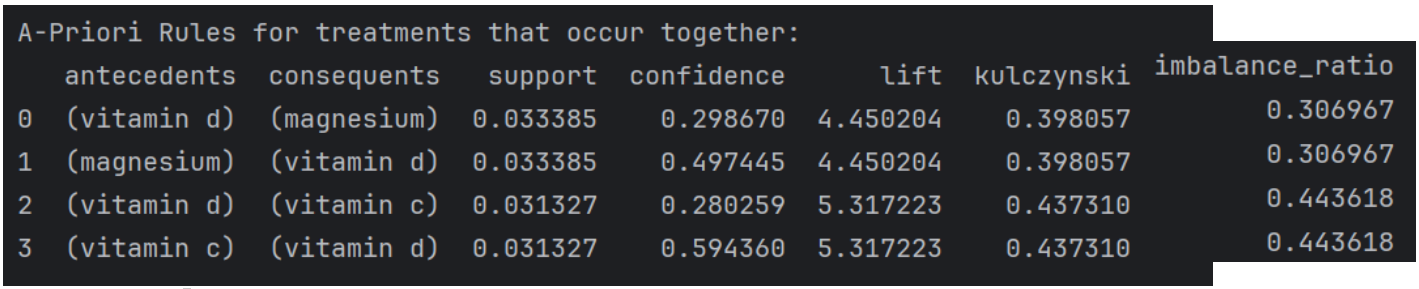
```
A-Priori Rules for symptoms that occur together:
   antecedents consequents   support  confidence      lift  kulczynski  \imbalance_ratio
0   (brain fog)   (fatigue)  0.149327    0.698630  1.634887    0.524038          0.434340
1   (dizziness)  (headache)  0.120436    0.648479  1.723757    0.484309          0.431451
2   (dizziness)    (nausea)  0.110577    0.595394  1.961581    0.479851          0.311103
3    (headache)   (fatigue)  0.194139    0.516053  1.207631    0.485182          0.083896
4  (joint pain)   (fatigue)  0.147497    0.536172  1.254712    0.440668          0.274331
5      (nausea)   (fatigue)  0.159667    0.526038  1.230999    0.449840          0.216740
6  (joint pain)  (headache)  0.149144    0.542159  1.441141    0.469304          0.201348
7      (nausea)  (headache)  0.165660    0.545783  1.450776    0.493067          0.141370
```

    - Condition that occur together (0.05; 0.4)



```
A-Priori Rules for condition that occur together:
    antecedents     consequents   support  confidence      lift  kulczynski  imbalance_ratio
0  (depression)       (anxiety)  0.129602    0.601689  3.214022    0.646989         0.103254
1     (anxiety)    (depression)  0.129602    0.692288  3.214022    0.646989         0.103254
2     (migraine)  (fibromyalgia)  0.057039    0.478020  2.232975    0.372233         0.342852
```

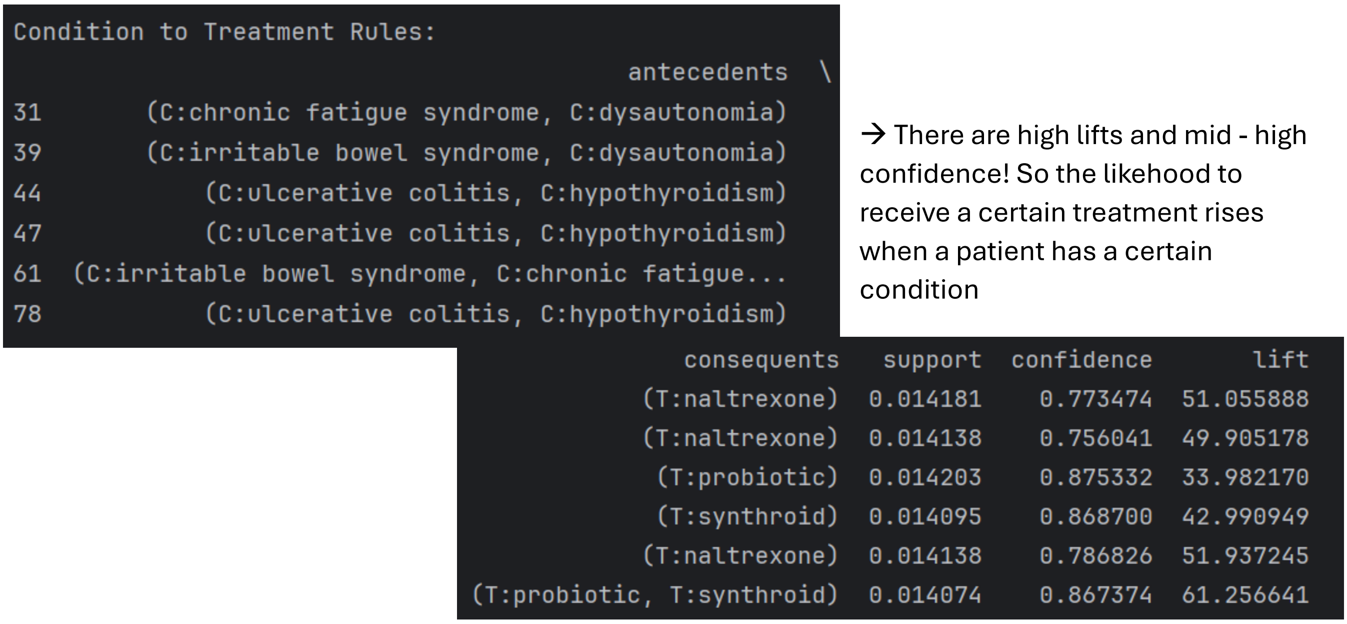
    - Treatments that occur together (0.03; 0.05)



```
A-Priori Rules for treatments that occur together:
   antecedents  consequents   support  confidence      lift  kulczynski  imbalance_ratio
0  (vitamin d)  (magnesium)  0.033385    0.298670  4.450204    0.398057         0.306967
1  (magnesium)  (vitamin d)  0.033385    0.497445  4.450204    0.398057         0.306967
2  (vitamin d)  (vitamin c)  0.031327    0.280259  5.317223    0.437310         0.443618
3  (vitamin c)  (vitamin d)  0.031327    0.594360  5.317223    0.437310         0.443618
```

➔ Positive lift – so the likehood of one condition rises when the other is here. However Kulzynski signalises a low – mid connection, so ist he imbalance ratio (number of overall occurance)

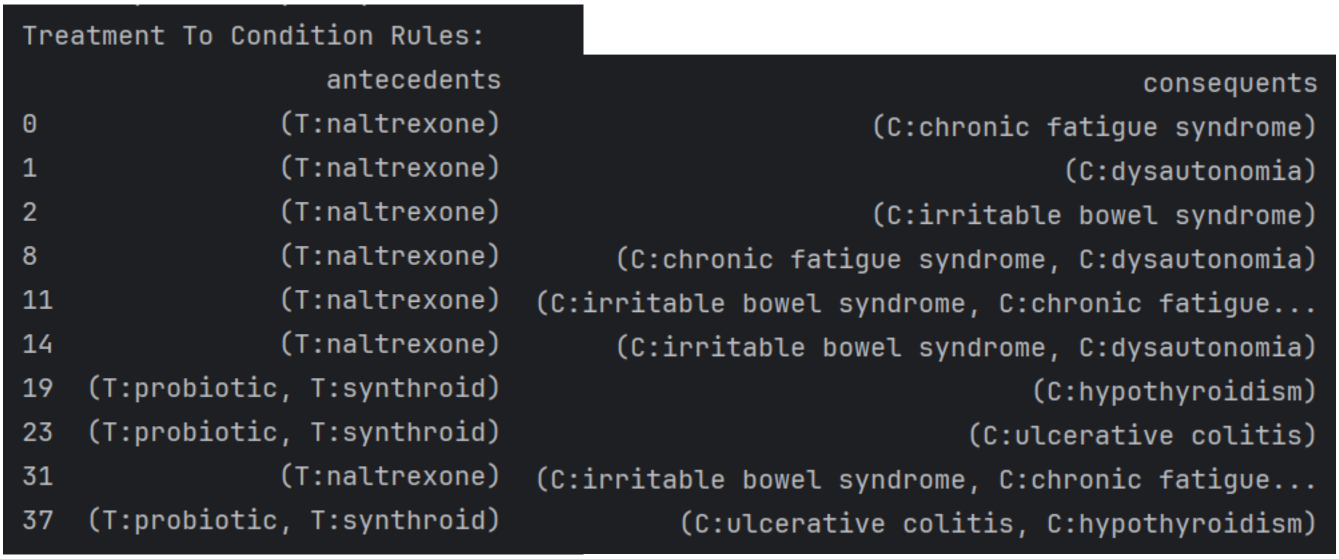
- Multi-Association Rules
    - Condition → Treatment (0.014; 0.5)



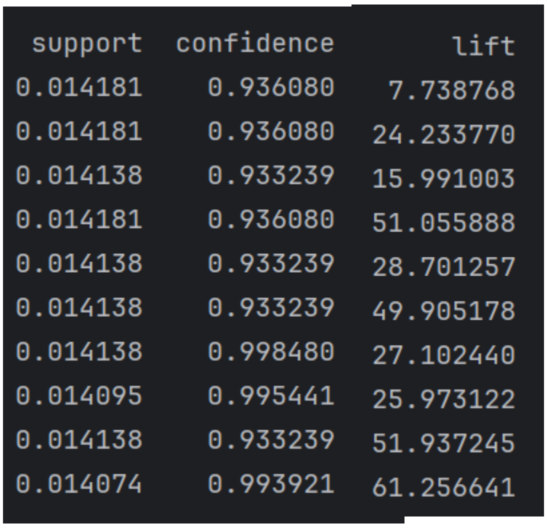
```
Condition to Treatment Rules:
                                         antecedents  \
31       (C:chronic fatigue syndrome, C:dysautonomia)
39       (C:irritable bowel syndrome, C:dysautonomia)
44           (C:ulcerative colitis, C:hypothyroidism)
47           (C:ulcerative colitis, C:hypothyroidism)
61   (C:irritable bowel syndrome, C:chronic fatigue...
78           (C:ulcerative colitis, C:hypothyroidism)

             consequents   support  confidence       lift
          (T:naltrexone)  0.014181    0.773474  51.055888
          (T:naltrexone)  0.014138    0.756041  49.905178
           (T:probiotic)  0.014203    0.875332  33.982170
           (T:synthroid)  0.014095    0.868700  42.990949
          (T:naltrexone)  0.014138    0.786826  51.937245
(T:probiotic, T:synthroid)  0.014074    0.867374  61.256641
```

→ There are high lifts and mid - high confidence! So the likehood to receive a certain treatment rises when a patient has a certain condition

- Treatment → Condition (0.014; 0.9)



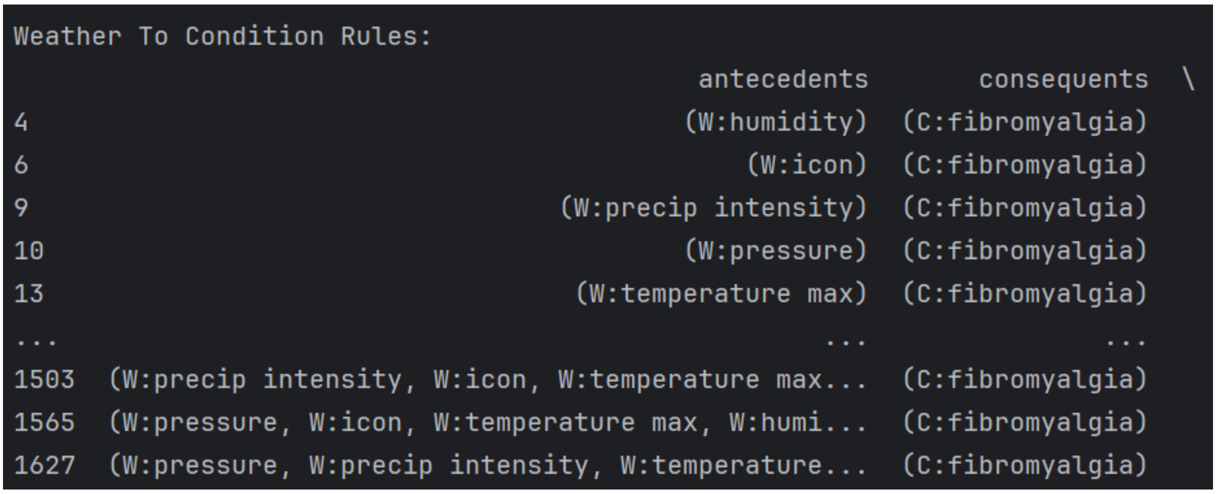
```
Treatment To Condition Rules:
                   antecedents                                          consequents
0               (T:naltrexone)                   (C:chronic fatigue syndrome)
1               (T:naltrexone)                                (C:dysautonomia)
2               (T:naltrexone)                   (C:irritable bowel syndrome)
8               (T:naltrexone)      (C:chronic fatigue syndrome, C:dysautonomia)
11              (T:naltrexone)  (C:irritable bowel syndrome, C:chronic fatigue...
14              (T:naltrexone)      (C:irritable bowel syndrome, C:dysautonomia)
19  (T:probiotic, T:synthroid)                              (C:hypothyroidism)
23  (T:probiotic, T:synthroid)                          (C:ulcerative colitis)
31              (T:naltrexone)  (C:irritable bowel syndrome, C:chronic fatigue...
37  (T:probiotic, T:synthroid)         (C:ulcerative colitis, C:hypothyroidism)
```



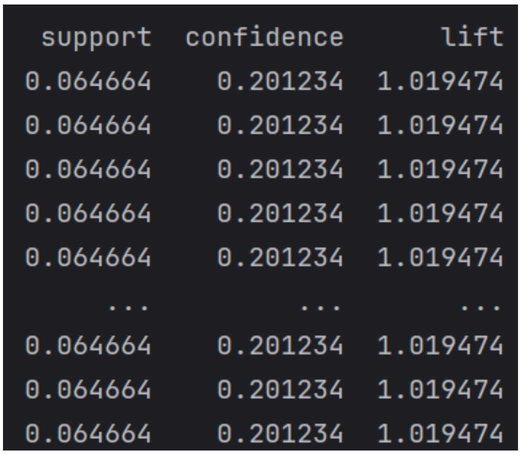
```
 support  confidence       lift
0.014181    0.936080   7.738768
0.014181    0.936080  24.233770
0.014138    0.933239  15.991003
0.014181    0.936080  51.055888
0.014138    0.933239  28.701257
0.014138    0.933239  49.905178
0.014138    0.998480  27.102440
0.014095    0.995441  25.973122
0.014138    0.933239  51.937245
0.014074    0.993921  61.256641
```

→ There are high lifts and high confidence! So the likehood of a certain condition rises when a patient receives a certain treatment.

- Weather → Condition (0.06; 0.2)



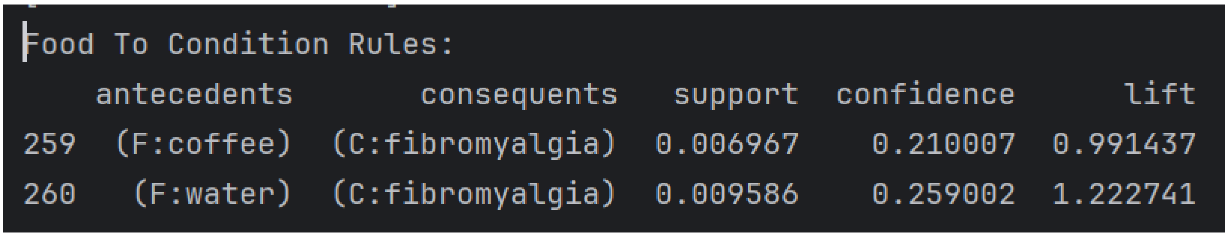
```
Weather To Condition Rules:
                                        antecedents       consequents  \
4                                     (W:humidity)  (C:fibromyalgia)
6                                         (W:icon)  (C:fibromyalgia)
9                               (W:precip intensity)  (C:fibromyalgia)
10                                    (W:pressure)  (C:fibromyalgia)
13                             (W:temperature max)  (C:fibromyalgia)
...                                            ...               ...
1503  (W:precip intensity, W:icon, W:temperature max...  (C:fibromyalgia)
1565  (W:pressure, W:icon, W:temperature max, W:humi...  (C:fibromyalgia)
1627  (W:pressure, W:precip intensity, W:temperature...  (C:fibromyalgia)
```



```
 support  confidence      lift
0.064664    0.201234  1.019474
0.064664    0.201234  1.019474
0.064664    0.201234  1.019474
0.064664    0.201234  1.019474
0.064664    0.201234  1.019474
     ...         ...       ...
0.064664    0.201234  1.019474
0.064664    0.201234  1.019474
0.064664    0.201234  1.019474
```

➔ Every rule (63) has theh same support, confidence and lift. Positive Lift → so it seems like many weather condition increse the likehood of fibromyalgia

- Food → Condition (0.005; 0.2)



```
Food To Condition Rules:
    antecedents       consequents   support  confidence      lift
259  (F:coffee)  (C:fibromyalgia)  0.006967    0.210007  0.991437
260   (F:water)  (C:fibromyalgia)  0.009586    0.259002  1.222741
```

➔ The likehoof of the appearance of fibromyalgia decreases, than when a patient drinks water.

- Condition → Food (-)

Couldn’t be computed!