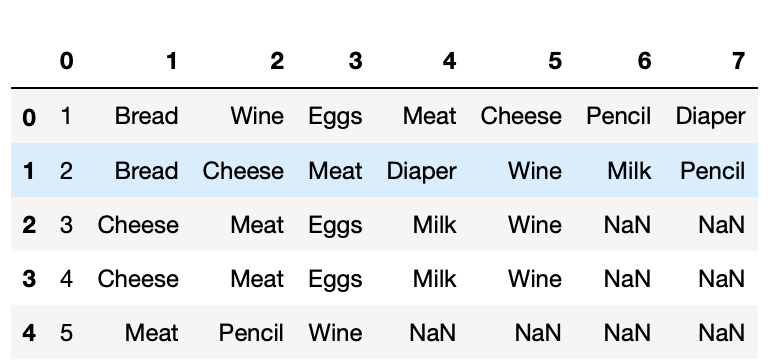
Sorry Dr. Santago, I will update my association portfolio by the end of Sunday!!!

**Data Acquisition:**

I used the dataset you shared with us in google drive titled “**Assoc\_Analysis\_Vidhya.dat.csv**”. From my understanding, it is a market dataset collected from each transaction in a grocery store. Each row represents a transaction containing all the goods bought by one person. It is a perfect data source for association analysis and we can infer the frequent purchasing list by understanding the customers’ buying patterns and preferences.



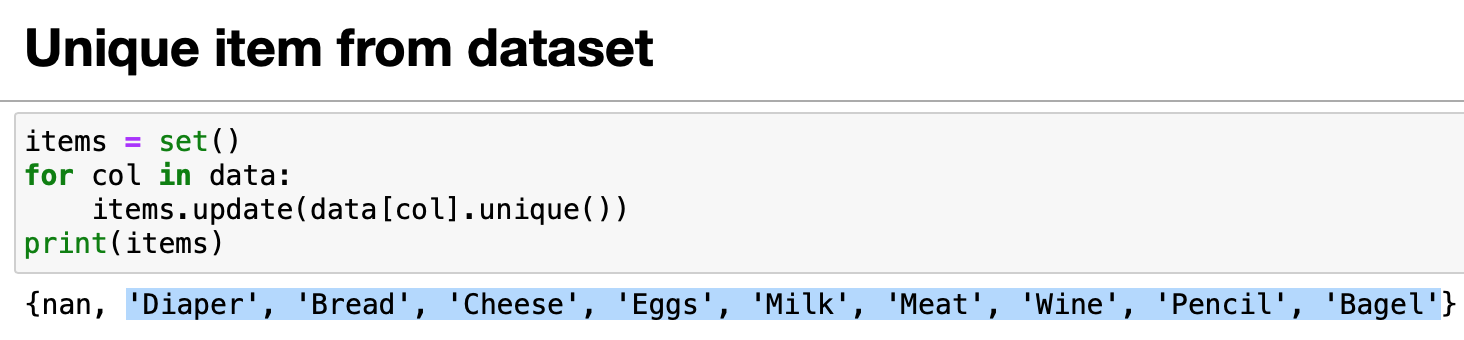
**Dimension**:

315 rows and 8 columns

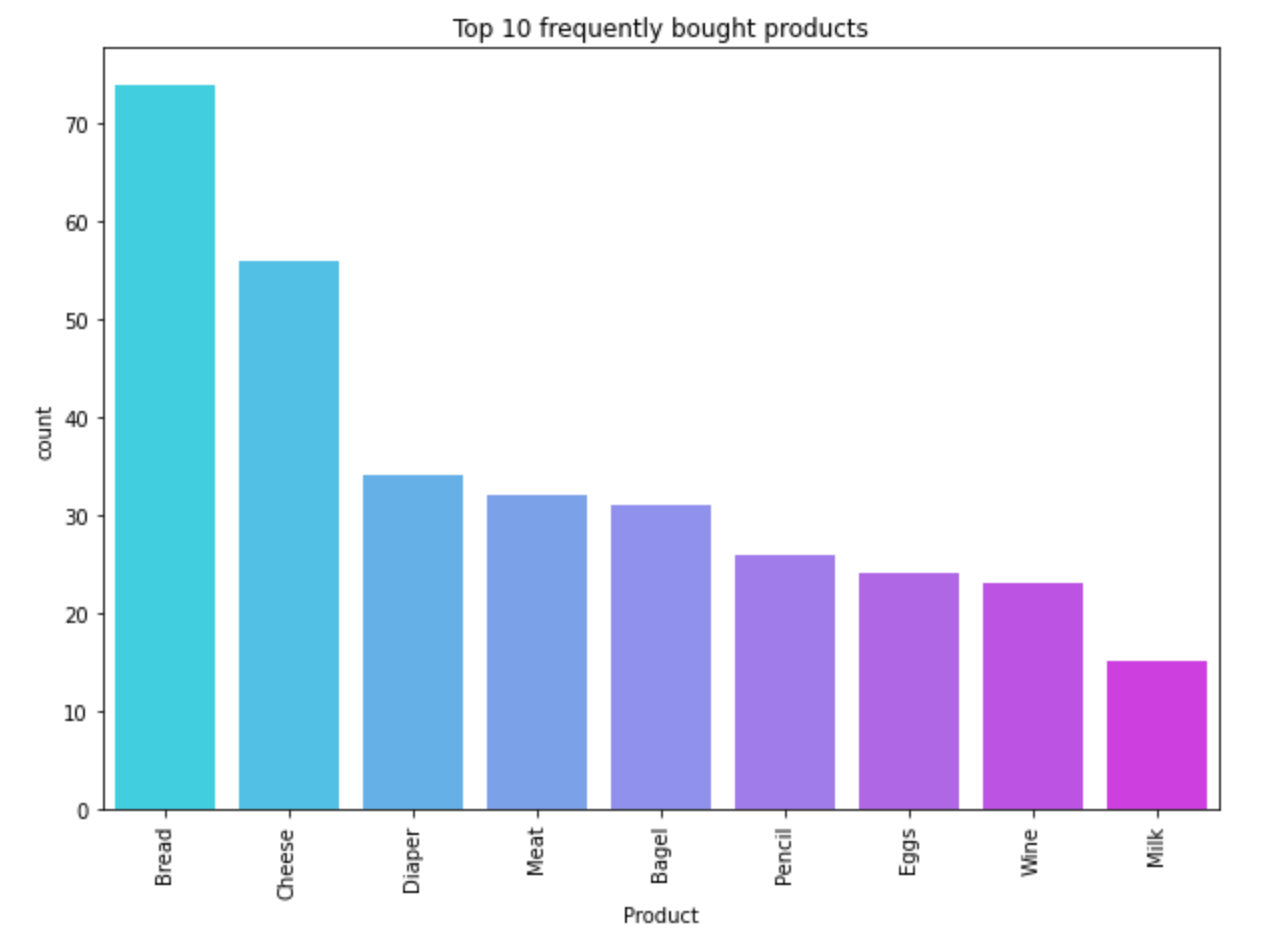
**Preprocessing and EDA:**

I delete the first column since we don’t need the index for analysis and plotting.

I first wanted to find out the unique item from the dataset and here is my result.



I was also interested in finding the most frequent item people bought and here is a visualization plot.

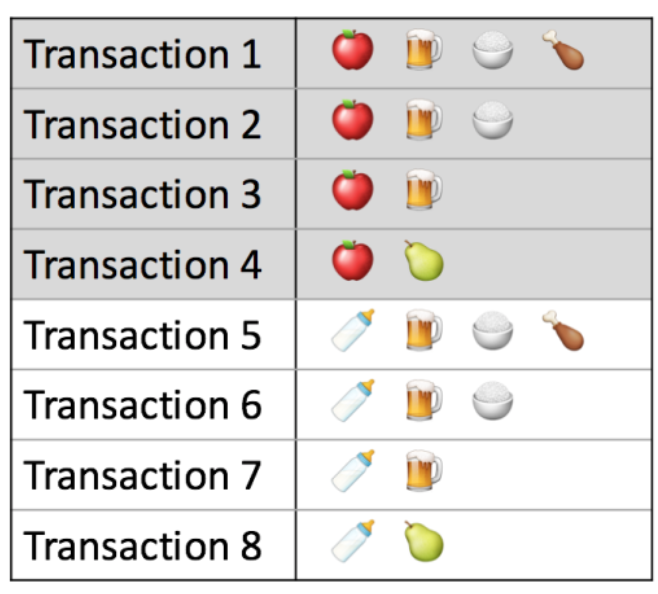


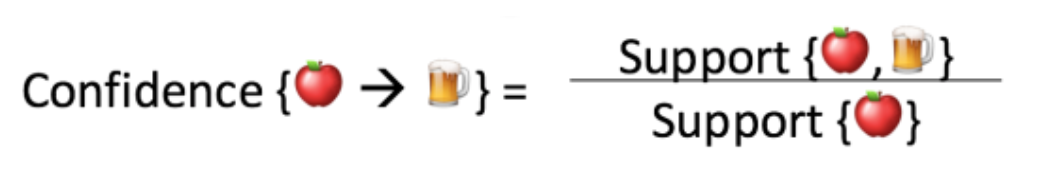
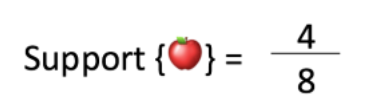
It seems that more than 25% of transactions contain bread and cheese and we believe that these are popular items.

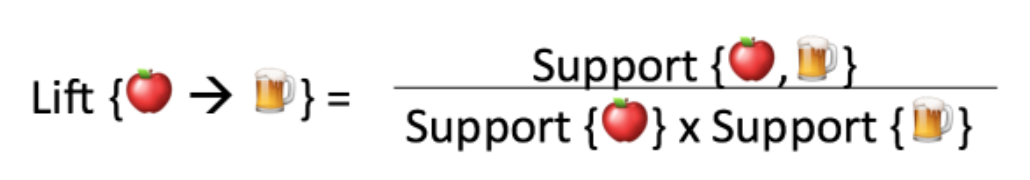
**Tasks:**

* Understand the definition of support, confidence, and lift
* Implement Apriori algorithm in python and generate frequent itemset
* Rule generation

I saw some good illustrations of understanding support, confidence, and lift. [Reference article](https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html)



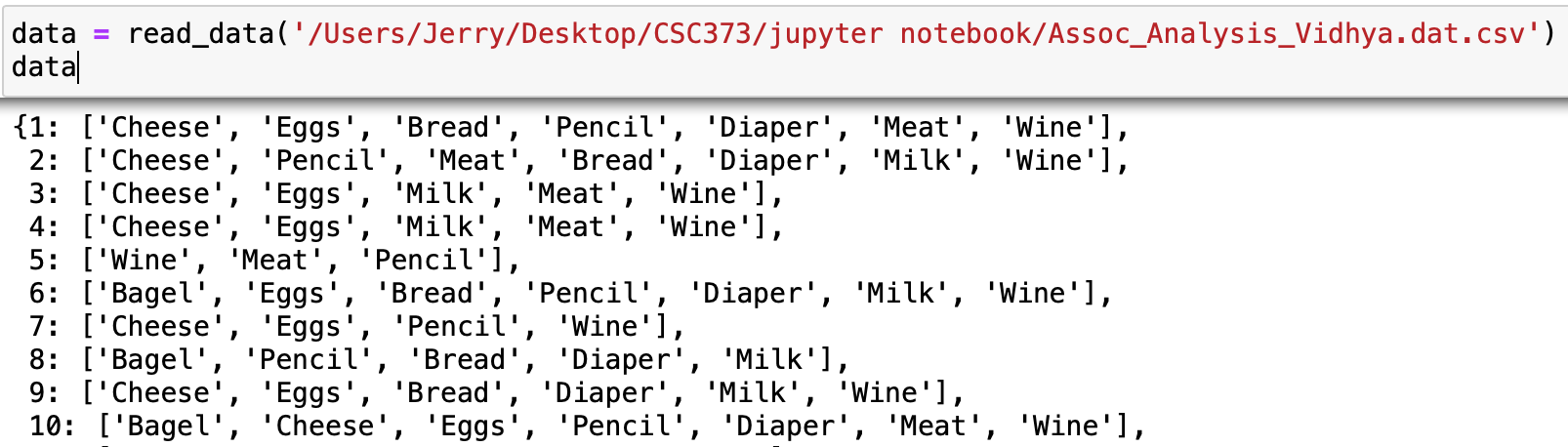




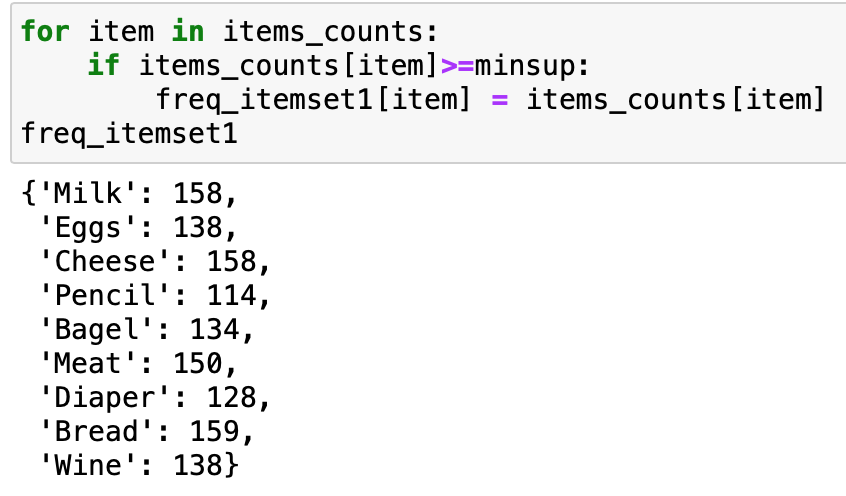
**Code Development:**

[**Link to Jupyter Notebook**](https://colab.research.google.com/drive/1MCuh3PiinQpoYdL1YtnIyJMH4CZ9NMVV)

I created two functions, one is for reading in data and the other is counting the frequency of each input itemset. I read data from .csv file and stored it in a dictionary.

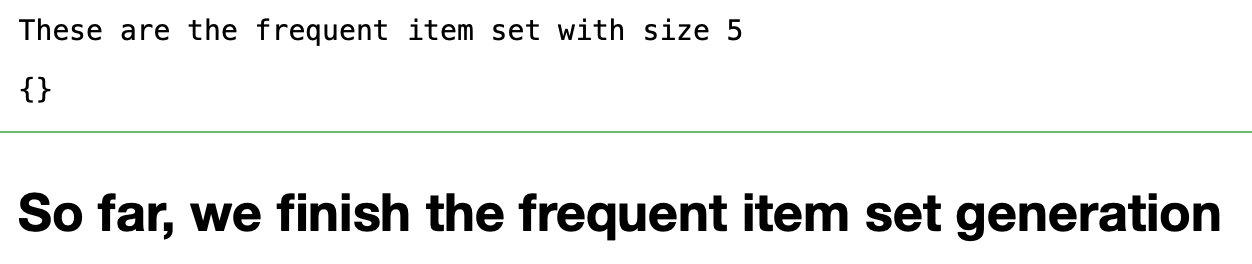


Firstly, I wanted to get all 1-itemset. In order to do this, I created an empty set and add each item from the dataset to the set. The property of the Python set is that it does not allow repeated items, so the set will only contain the unique items, which is our 1-itemset. I then pass the 1-itemset as input to the frequency count function and get their occurrence. I set minsup=30 and filter out some items below this threshold.



Then, I created a While loop to repeat this process: while freq\_itemset3 != {}:

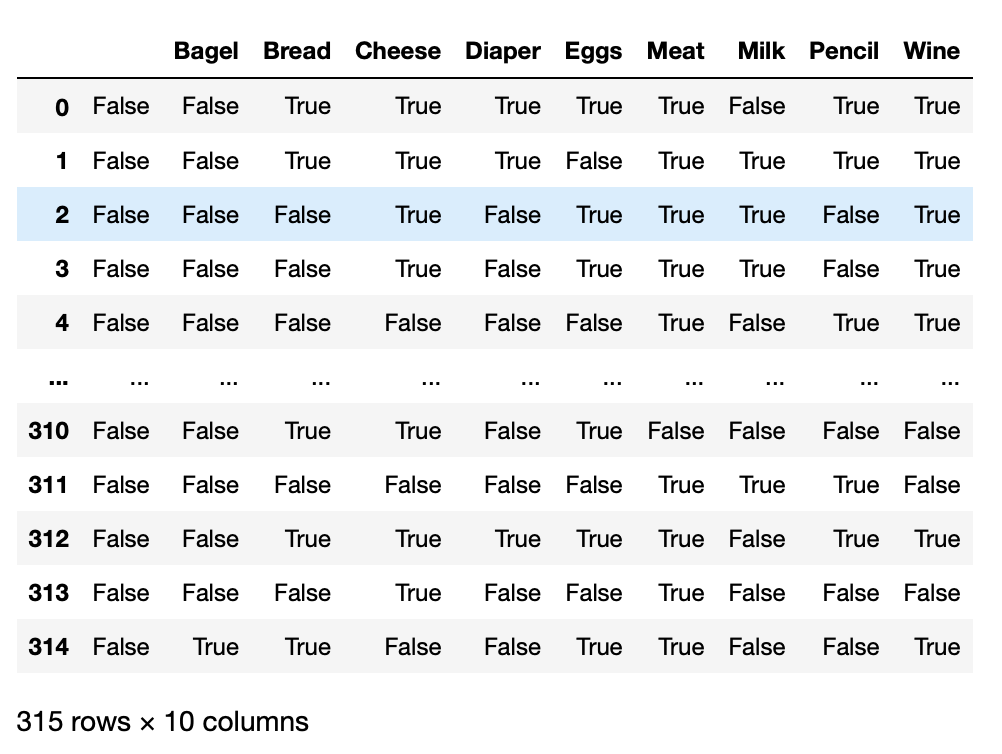
The stopping condition is when the k=itemset is empty. Finally, I got all the k-itemset and this concludes our first step of association analysis.



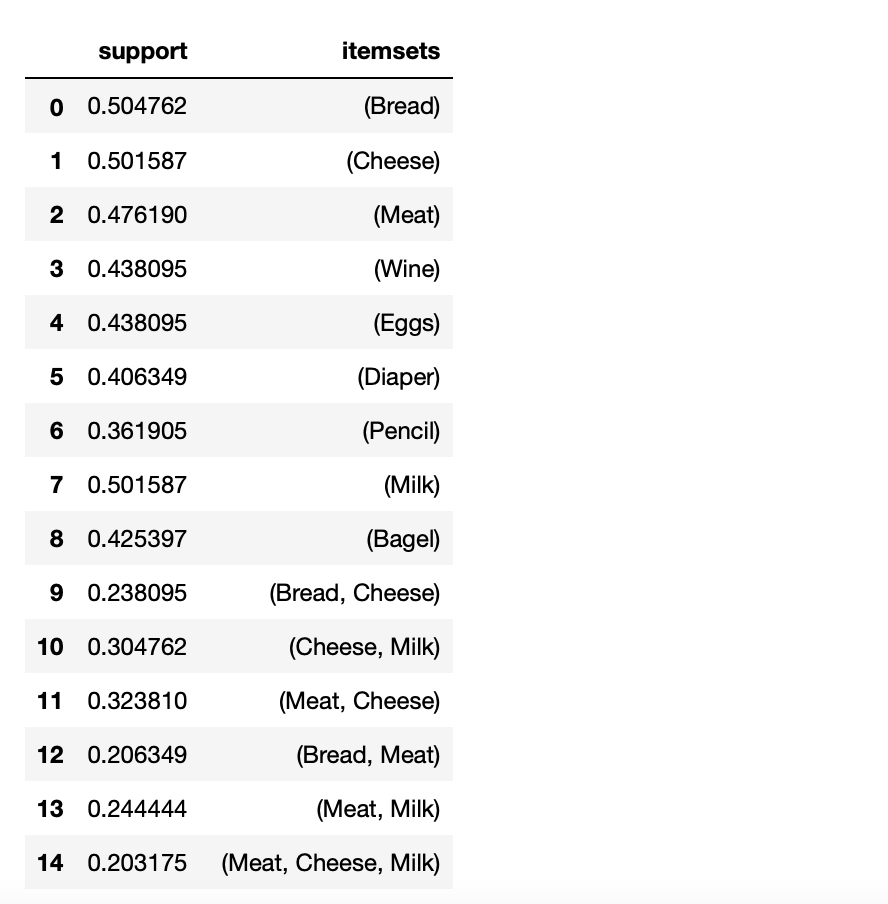
**Package Used:**

[**Link**](https://colab.research.google.com/drive/1sI3cH7_eoR0M5vh0y1eyHeWufDGlgh__)

In order to find the frequent itemset using mlxtend.preprocessing and mlxtend.frequent\_patterns, the dataset should be in a 2D list format. Then, we transformed our dataset to this format. See the screenshot below. We put each unique item as a column and each transaction as a row.

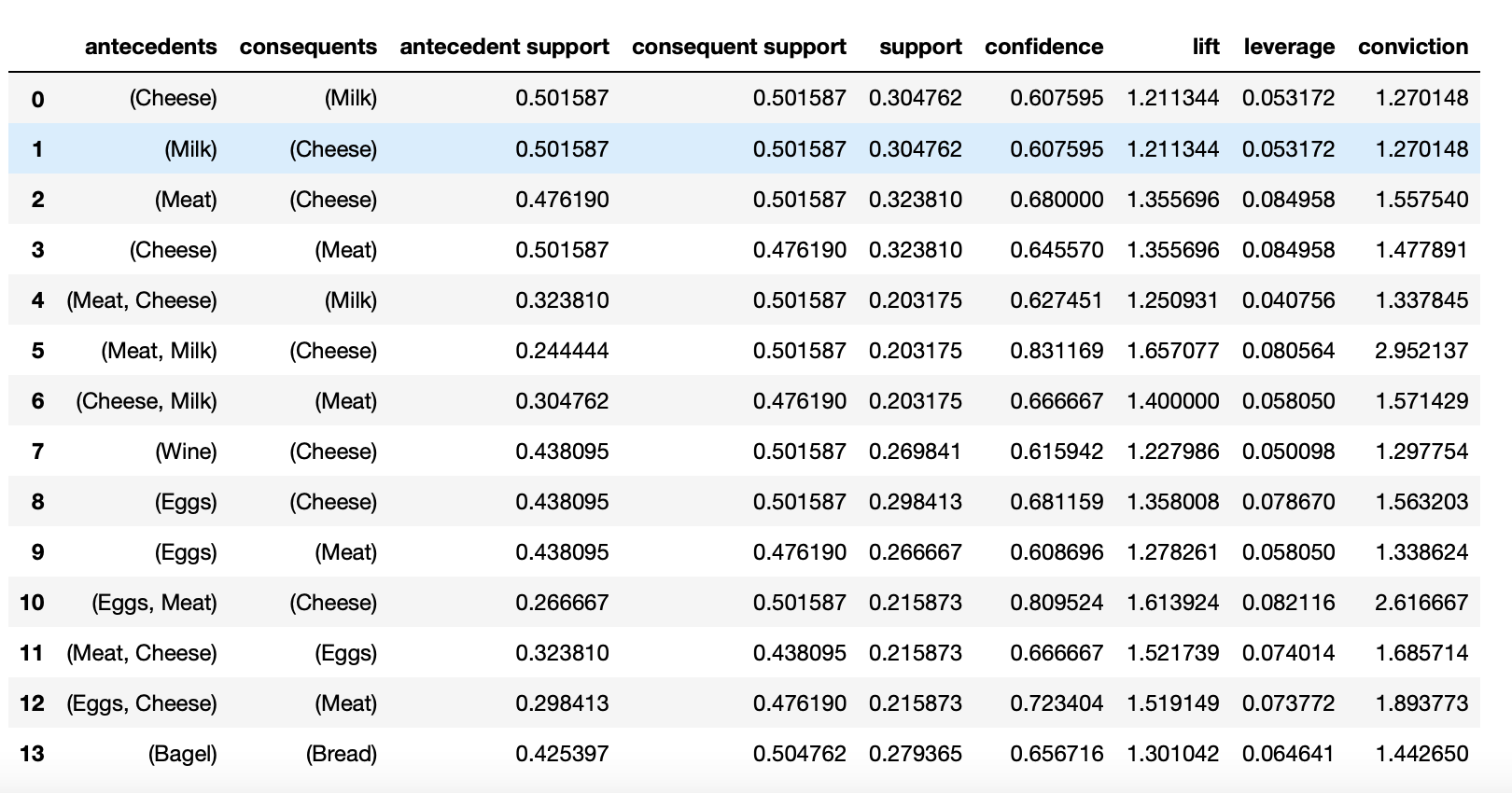


Then, I compute the support of frequent itemset = 1 using minsup = 0.2.

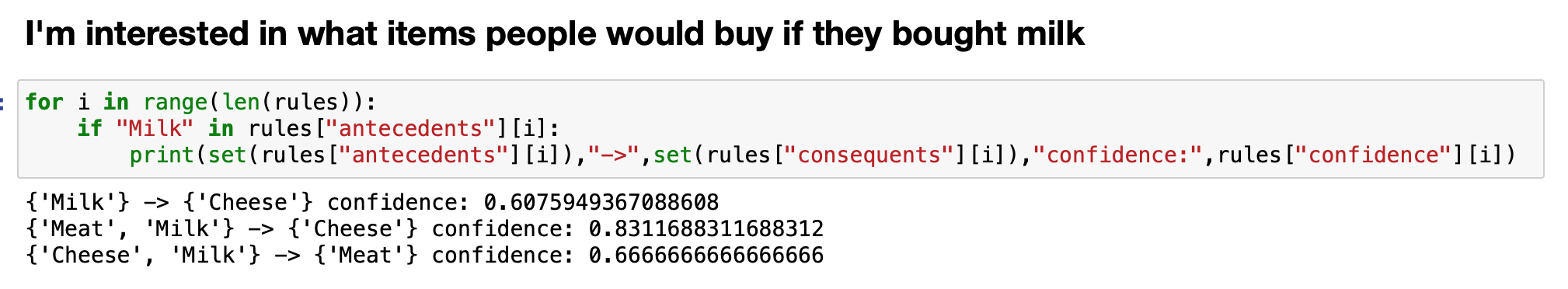


After getting the frequent itemset, I want to find the associated rule generation and their corresponding confidence.

This is the entire association table with min\_confidence = 0.4

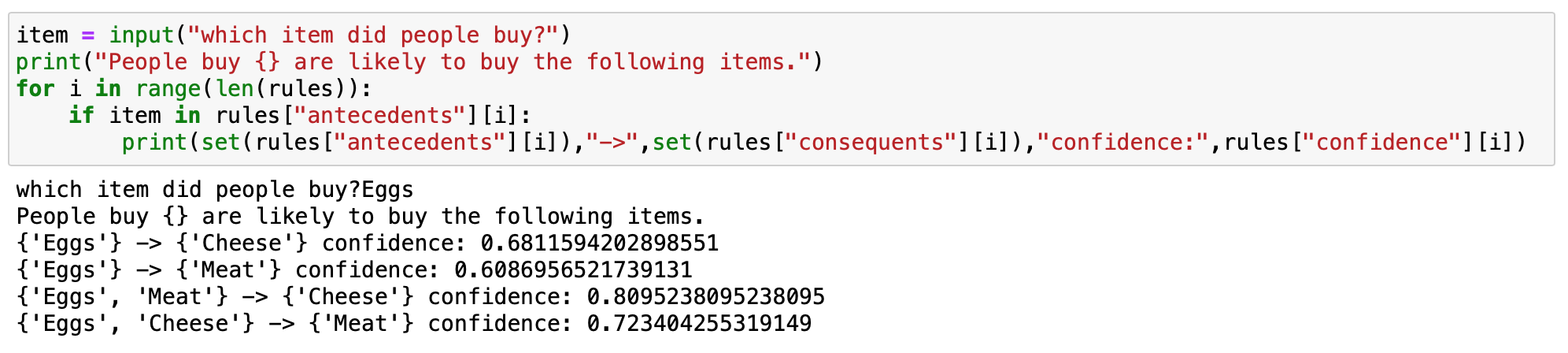


I’m also trying to find the association rule for a specific item so I tried this:



The output shows that people who purchased milk indicates that they are also likely to buy cheese and meat. This association rule is important to the grocery stores’ managers because they can put associate items on nearby shelves so that people are likely to spend more money.

I built an interactive query for search association rule for one item:



I used Eggs as input and the result shows that people who bought eggs are likely to buy cheese and meat.

This interactive tool is helpful for grocery stores to arrange items.

**Theory exercise**

**Under the same folder**

**Self-assessment**

I think I learned a lot through building this portfolio. Association analysis is not something new in daily life because almost all online shopping websites utilize this as part of their recommendation system. However, it is my first time conducting market basket analysis and generating association rules by data mining tools. At first, I was so confused about how the Apriori algorithm and FP-growth work, so I watched these videos for detailed explanation and these help me a lot in understanding the computation details and developing my code.

<https://www.youtube.com/watch?v=VB8KWm8MXss>

<https://www.youtube.com/watch?v=h_l3b2CIQ_o&t=807s>

Overall, I think I have a solid understanding right now but I will keep exploring more about association analysis.