Project 2

Kelsey Neis - neis@umn.edu - 3942198

11/1/2021

Problem 1

The files for this problem is under Experiment 1 folder. Datasets to be used for experimentation: store transaction.csv. Jupyter notebook: apriori analysis.ipynb. In this experiment, we give a dataset of a store with thousands of transactions of customers buying several items from the store. We will use the apriori algorithm to find correlations between various items in the store.

Answer the following question :

1. How many records are there in the dataset?

7501

2. In a single transaction, what is the maximum number of items a customer has bought? We assume that each record is a separate transaction

20

3. Write any five transactions a customer has done.

{burgers,meatballs,eggs}

{chutney}

{turkey,avocado}

{mineral water,milk,energy bar,whole wheat rice,green tea}

{low fat yogurt}

4. We use the wordcloud to generate a stunning visualization format to highlight crucial textual data points and convey essential information. Generate and paste the wordcloud with max words set to 25 and 50. Briefly describe your understanding of the plot.

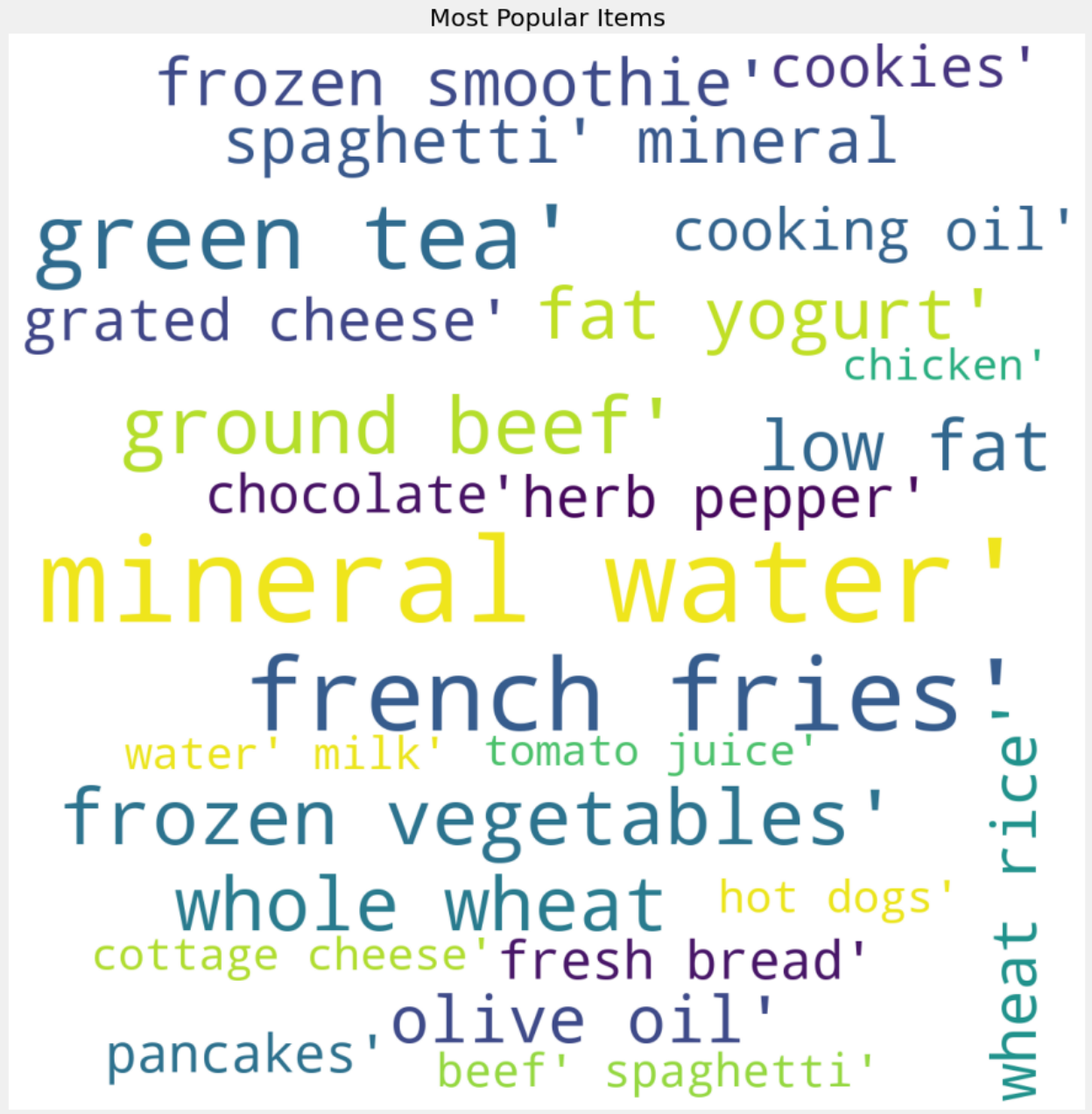
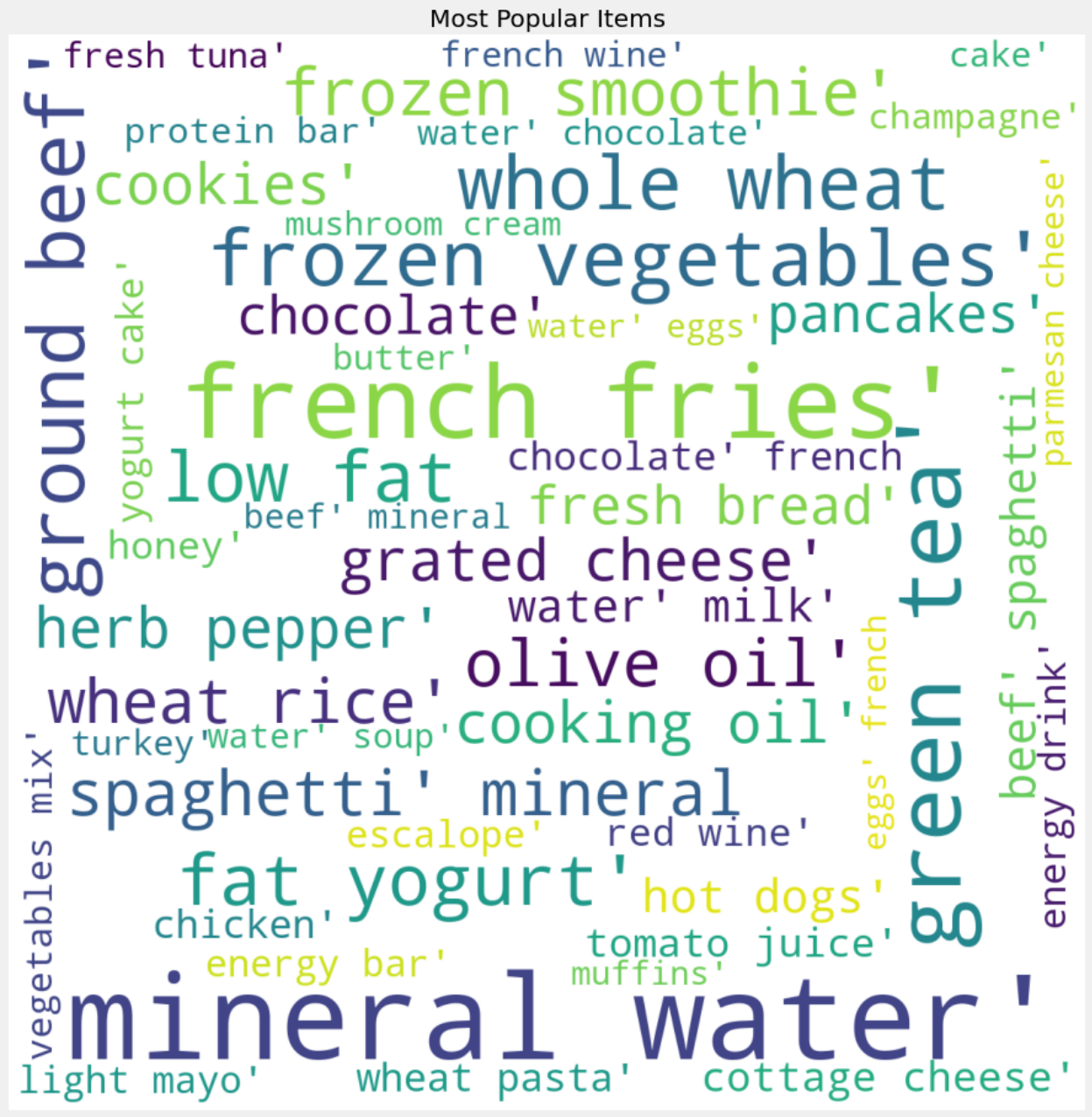


Figure 1: wordcloud, max=25



These plots show the top 25 and top 50 items bought by customers, represented by their size in the cloud.

5. What are the top 5 most frequent items in the dataset?

Mineral water, eggs, spaghetti, french fries, chocolate

6. Suppose we have the following transaction data: [['Apple', 'Beer', 'Rice', 'Chicken'], ['Apple', 'Beer', 'Rice'], ['Apple', 'Beer'], ['Apple', 'Bananas'], ['Milk', 'Beer', 'Rice', 'Chicken'], ['Milk', 'Beer', 'Rice'], ['Milk', 'Beer'], ['Apple', 'Bananas']]. Transform this input dataset into a one-hot encoded Boolean array. Hint: In the Jupyter notebook, we use TransactionEncoder to do the same.

	Apple	Bananas	Beer	Milk	Rice
0	True	False	True	False	True
1	True	False	True	False	True
2	True	False	True	False	False
3	True	True	False	False	False
4	False	False	True	True	True
5	False	False	True	True	True
6	False	False	True	True	False
7	True	True	False	False	False

7. In the input dataset, how many unique items are present?

121

8. Run Apriori to generate frequent itemsets at support thresholds of 1%, 2%, 3%, 4%, 5%, 6%, 8%,10%,12%,16% and 20%. In a single figure, for each threshold (X-axis), plot the number of itemsets (Y-axis). Comment on the general trends illustrated by the plots and the reason for the trend.

As expected, the number of itemsets decreases proportionally as the threshold increases.

9. At support threshold 1%, we see frequent itemset of size three along with size 2 and 1. However, at the support threshold of 2%, we observe itemsets of size 1 and 2 only. Why do you think this is so?

The larger the itemset, the less likely it will be frequent. 1- and 2-itemsets, in other words, 1-word and 2-word pairs are more likely to be frequent than a 3-word sequence.

10. Run Apriori to generate frequent itemsets of length 2 at support thresholds of 1%, 2%, 3%, 4% and 5%. In a single figure, for each threshold (X-axis), plot the number of itemsets of length 2 (Y-axis). Comment on the general trends illustrated by the plots and the reason for the trend.

The count of frequent 2-itemsets follows a similar trend to the plot not constrained by itemset size, except that it declines more rapidly as the support threshold increases, perhaps because of excluding the frequent 1-itemsets.

11. For the following itemset, write down its corresponding support value:

- Mineral Water

0.238

- Chocolate

0.164

- Eggs

0.18

- Eggs, Mineral Water

0.051

- Chocolate, Mineral Water

0.053

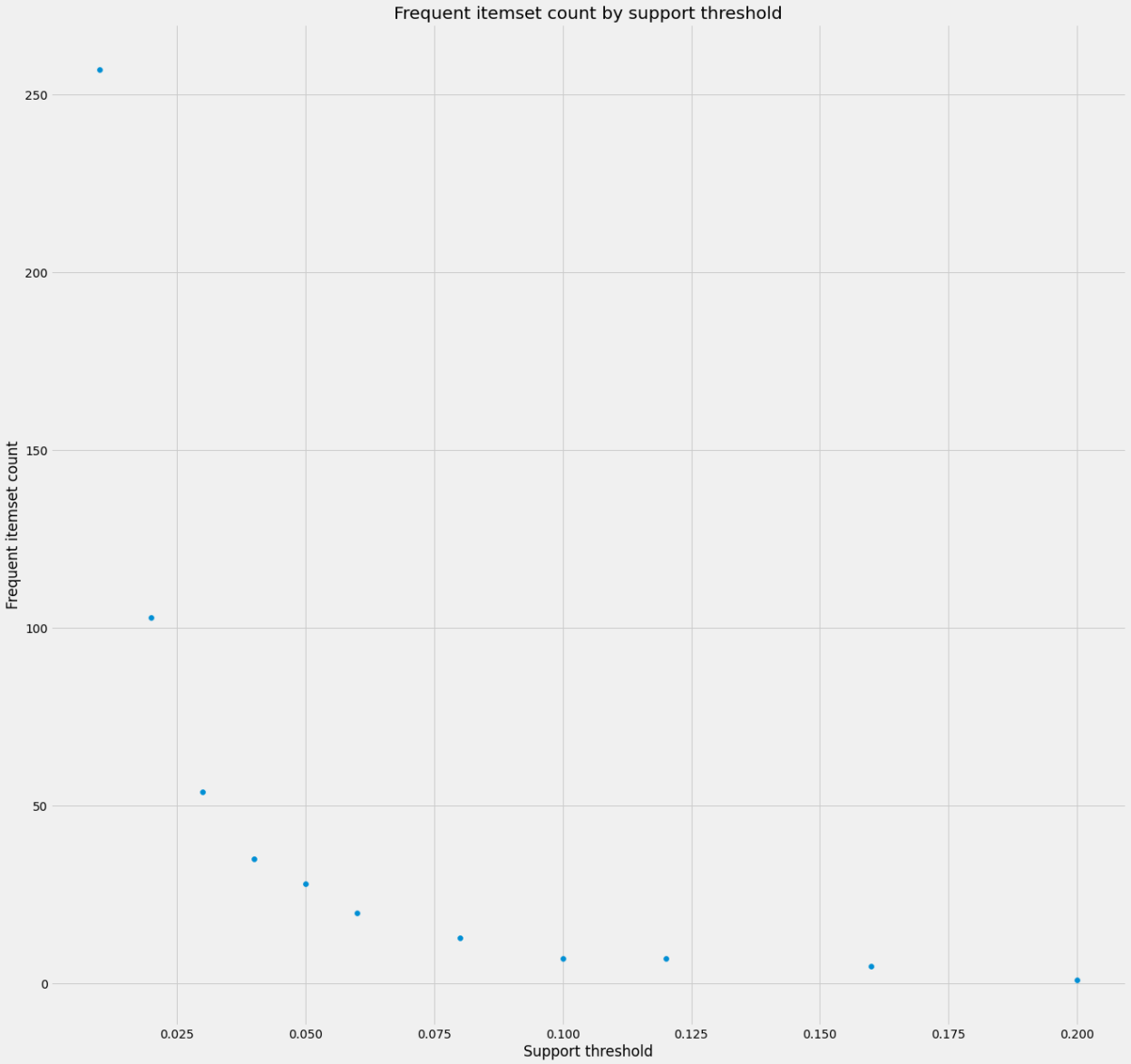


Figure 2: Apriori for different thresholds

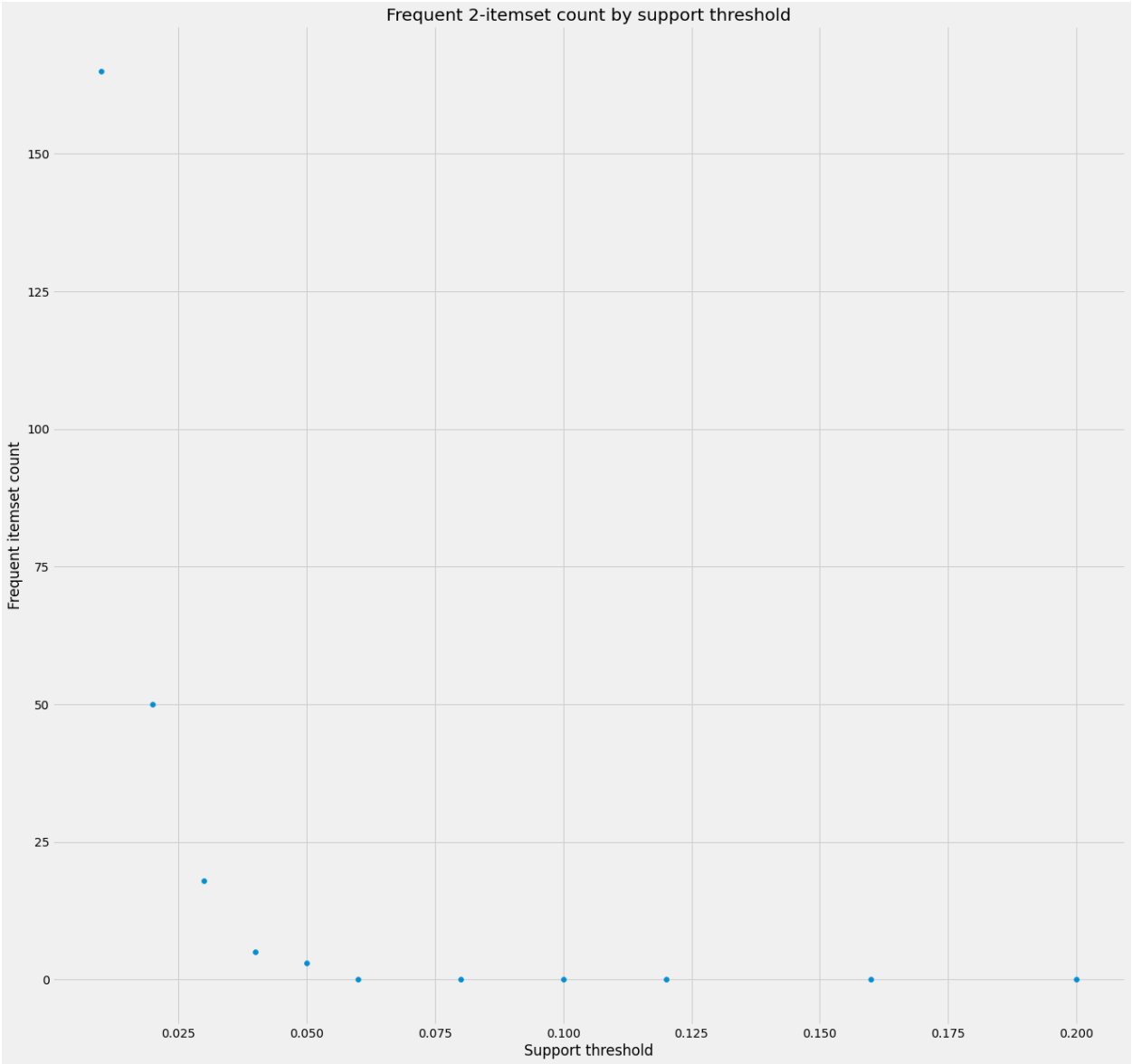


Figure 3: 2-itemsets

Problem 2

The files for this problem is under Experiment 2 folder. Datasets to be used for experimentation: instacart transaction.csv. Jupyter notebook: Instacart association.ipynb. Instacart, an online grocer, has graciously made some of their datasets accessible to the public (https://www.instacart.com/datasets/grocery-shopping-2017). In this experiment, we will use apriori algorithm to find correlations between the different items in the store. Answer the following question:

1. Given following transactions:

- order 1: apple, egg, milk
- order 2: carrot, milk
- order 3: apple, egg, carrot
- order 4: apple, egg
- order 5: apple, carrot

Using the apriori algorithm, write down the pair of items having a minimum threshold of 3. Briefly describe your steps.

In the first round, gather 1-itemsets: {apple}, {egg}, {milk}, {carrot}. Eliminate {carrot} and {milk} since they don't meet the minimum threshold of 3.

In the second round, gather 2-itemsets from the 1-itemsets that were not pruned: $\boxed{\{\text{apple, egg}\}}$ is the only one, and it meets the threshold of 3.

2. How many unique orders and unique items are there in the dataset? Are unique order same as number of records in the dataset? On average, per order, how many items does a customer order?

Unique orders: 3214874, Unique items: 49677

3. Run apriori to generate pairs of itemset at support thresholds of 1%, 2%, 4%, 6%, 8%, and 10%. In a single figure, for each threshold (X-axis), plot the number of association rules (Y-axis). Comment on the trend of algorithm runtime at different thresholds.

The runtime was quite long for the first two, but decreased after that. The total runtime was 17 minutes.

4. Run apriori at support thresholds of 1%, 2%, 4%, 6%, 8% and 10%. For each threshold, write a pair of association rules (you can choose any) along with its key metrics (i.e. freqAB, supportAB, freqA,supportA, freqB, support, confidenceAtoB, confidenceBtoA, lift). As a data scientist for the retailer giant, after observing the association rules, what would you do to increase the sales.

threshold	itemA	itemB	freqAB	supportAB	freq
.01	Grain Free Chicken Formula Cat Food	Grain Free Turkey Formula Cat Food	318	0.010553	1809
.02	Banana	Bag of Organic Bananas	654	0.021704	4700
.03	Original Hummus	Organic Blackberries	1163	0.040018	7111
.04	Cucumber Kirby	Organic Large Extra Fancy Fuji Apple	1163	0.040018	9697
.06	Carrots	Michigan Organic Kale	2229	0.080007	7248
.08	Bag of Organic Bananas	Organic Orange Bell Pepper	2233	0.080150	3734
.10	Green Bell Pepper	Red Peppers	2733	0.100014	5527

I would suggest placing popular itemset pairs near each other, and move all organic produce to be close together. I'd also make sure that similar items with different color/flavor are placed close together and place healthy snacks together as well. For example, the hummus -> organic blackberry pair would inspire customers to try something new if they were in the mood for a healthy snack.

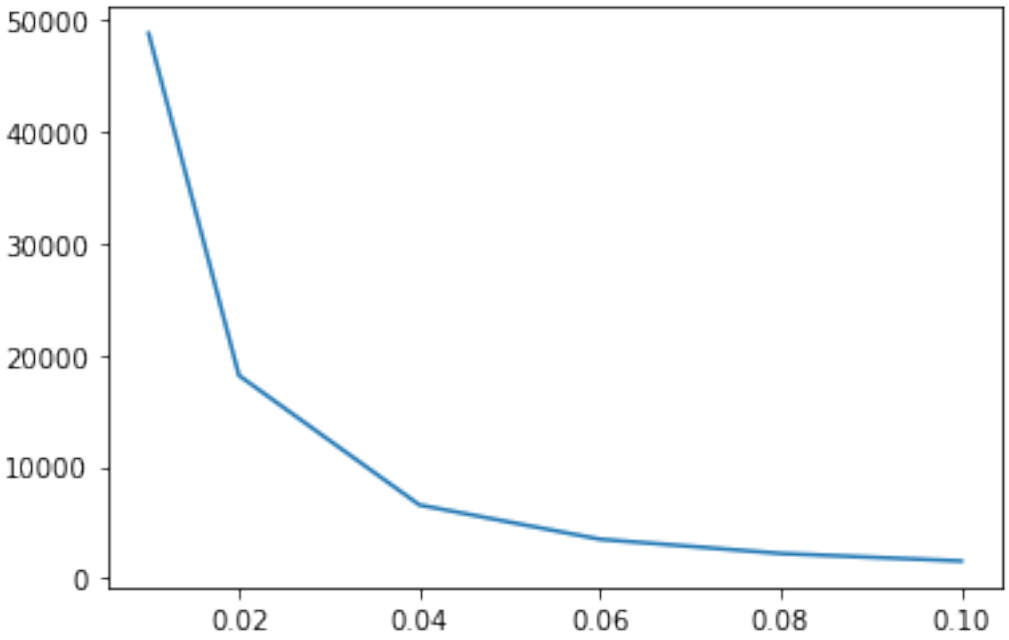


Figure 4: itemset pairs