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.

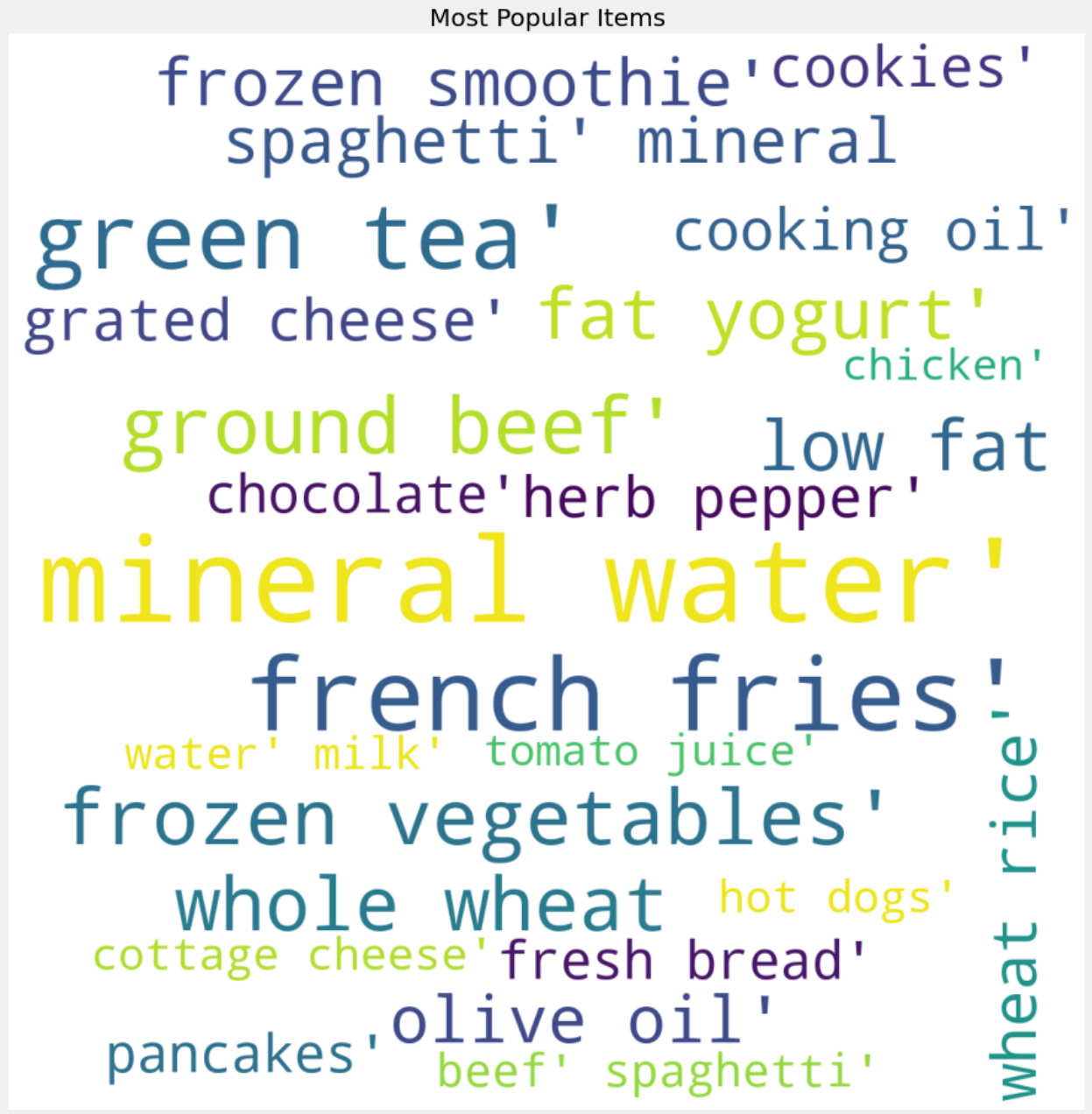
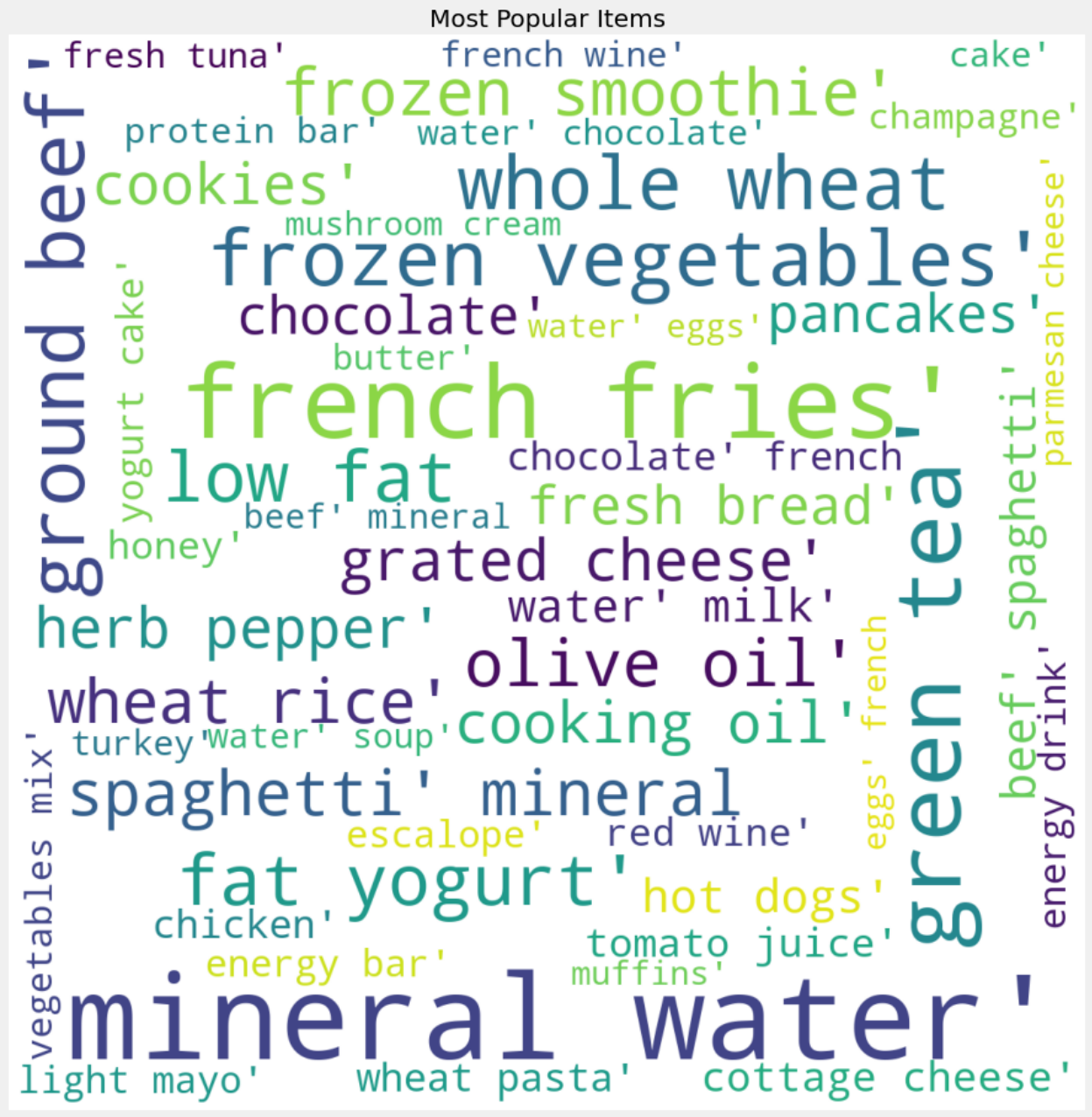


Figure 1: wordcloud, max=25



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

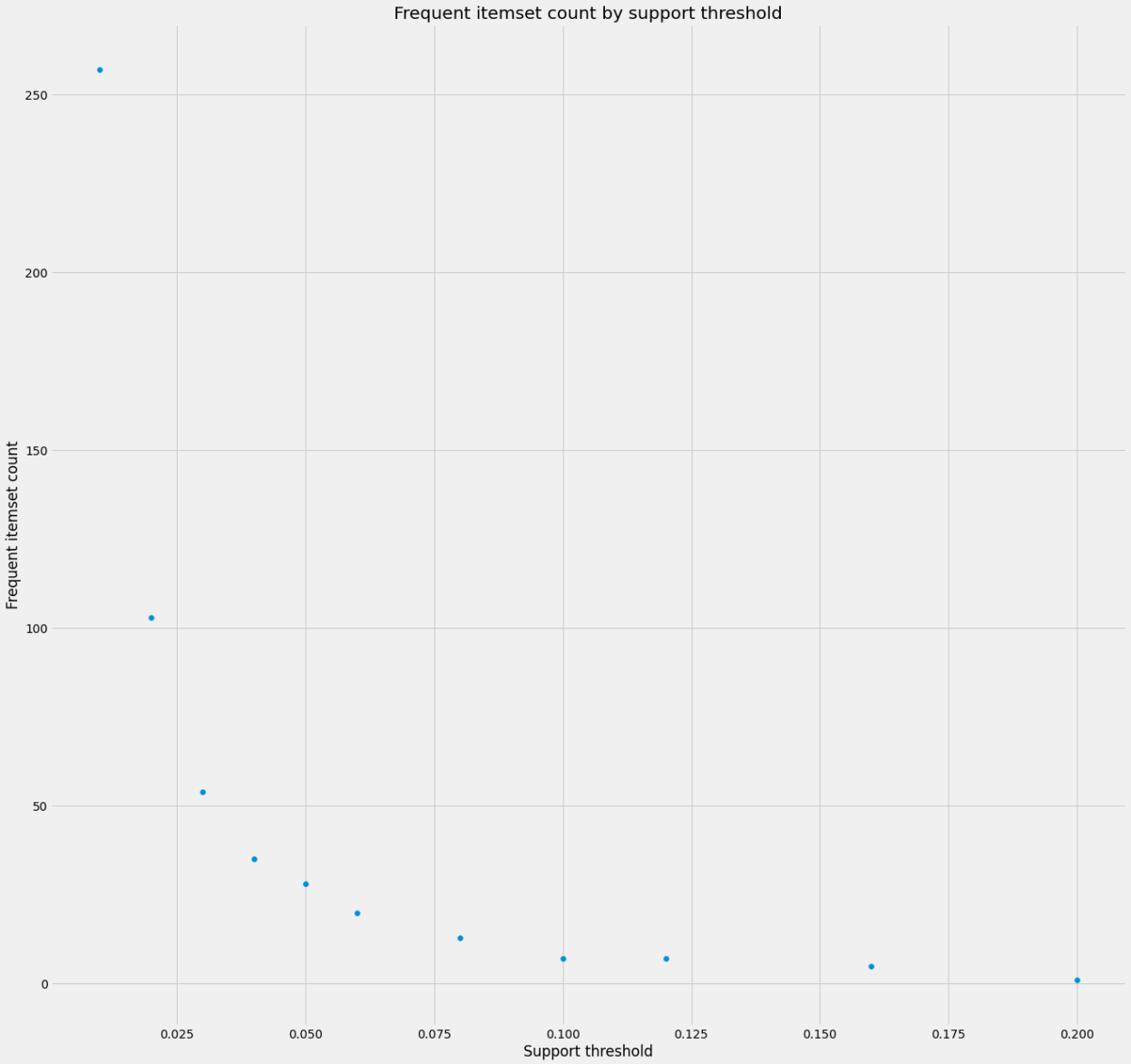


Figure 2: Apriori for different thresholds

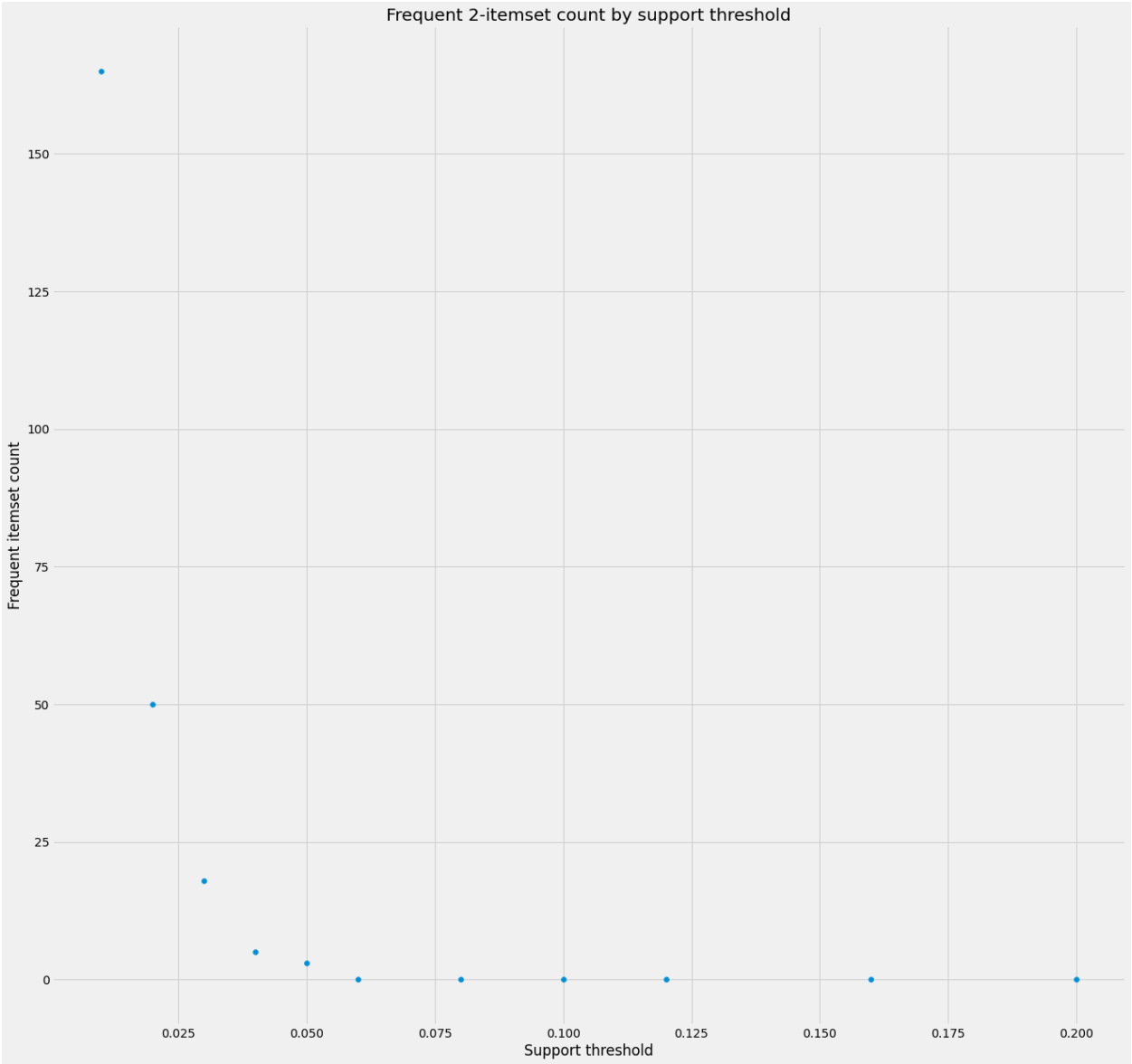


Figure 3: 2-itemsets