Problem 1

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 questions:

1. How many records are there in the dataset?

   There are 7501 transactions in the dataset.

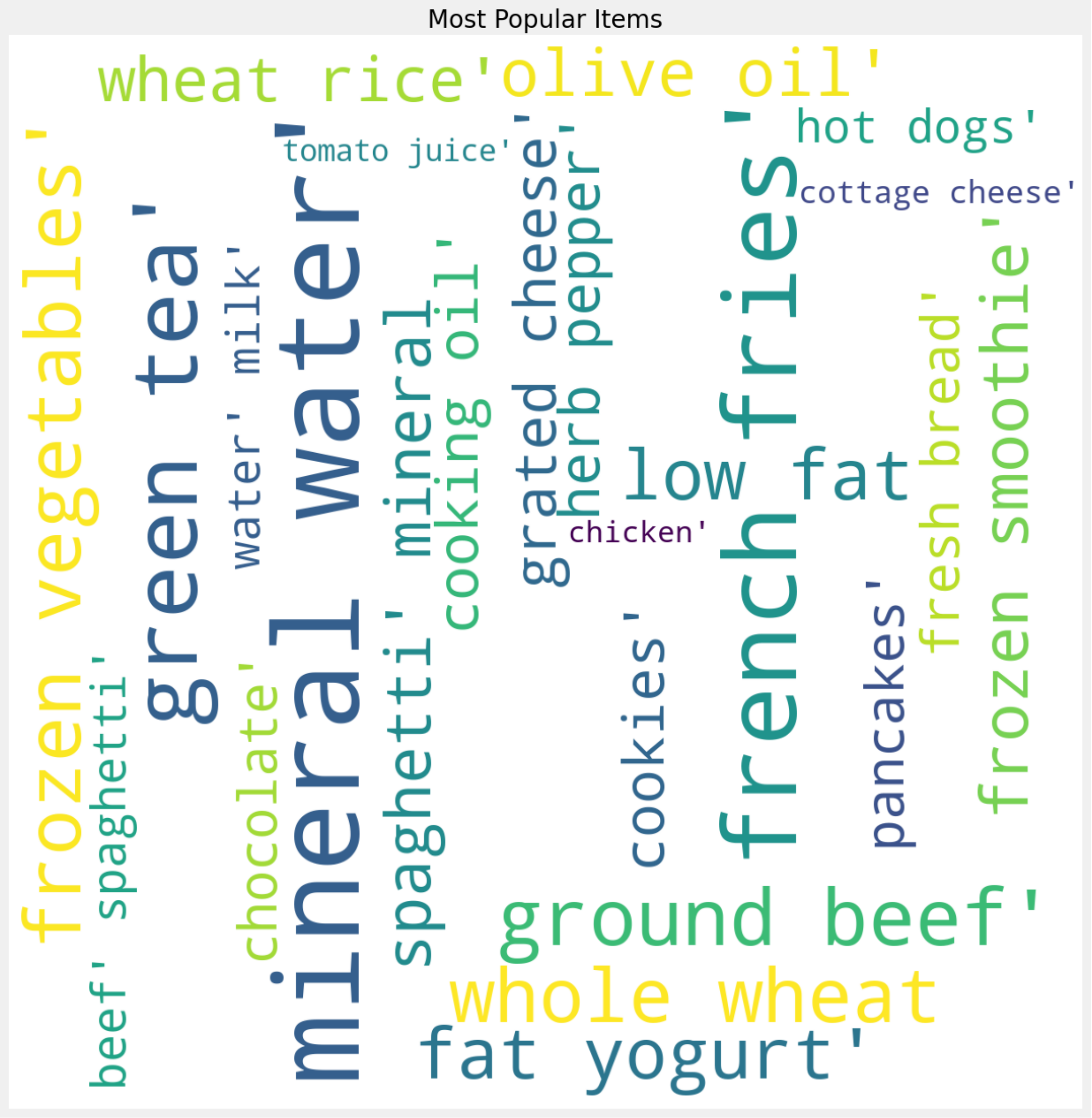
2. In a single transaction, what is the maximum number of items a customer has bought? We assume that each record is a separate transaction

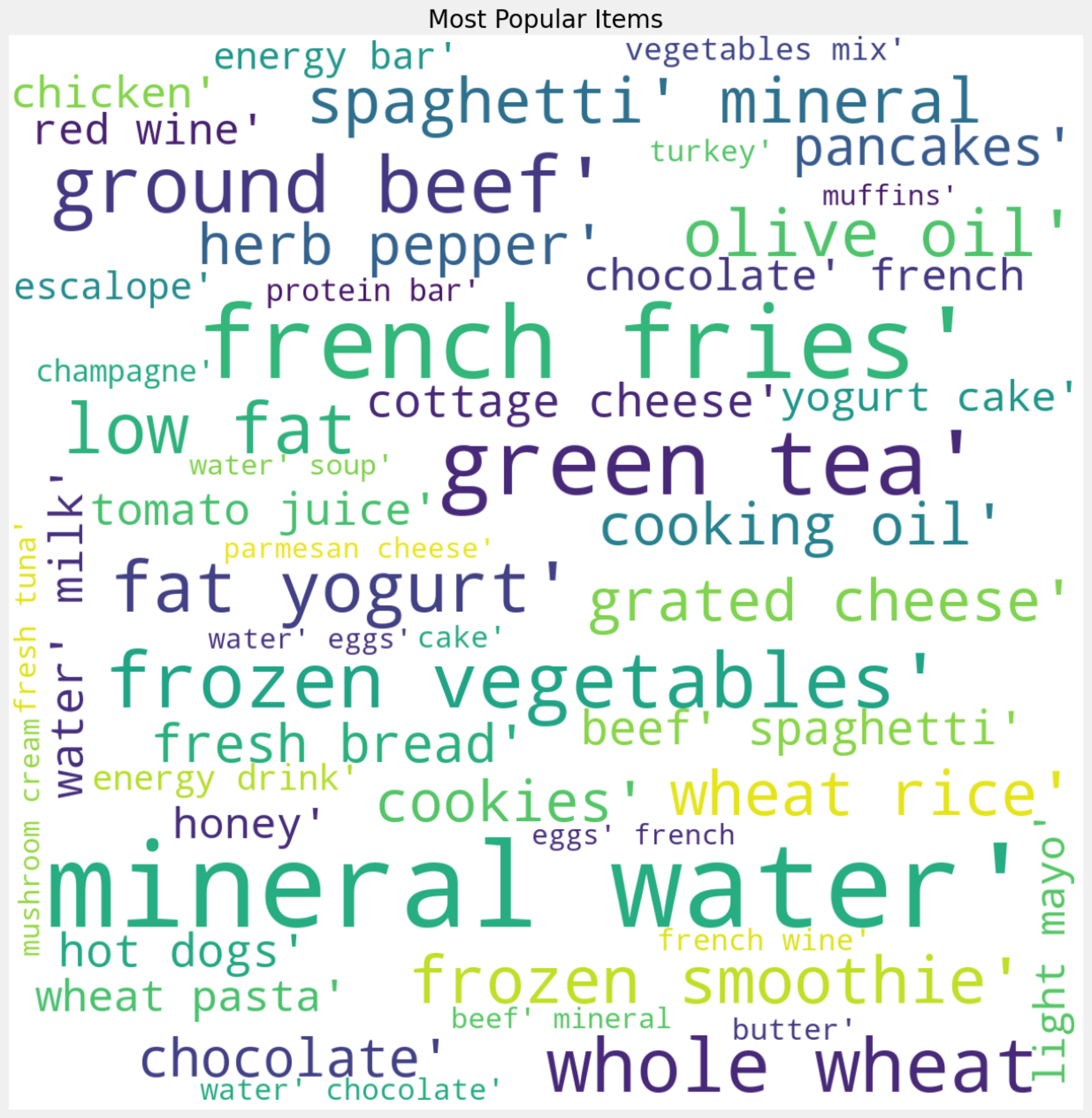
   20

3. Write any five transactions a customer has done.

   1. shrimp,almonds,avocado,vegetables mix,green grapes,whole weat flour,yams,cottage cheese,energy drink,tomato juice,low fat yogurt,green tea,honey,salad,mineral water,salmon,antioxydant juice,frozen smoothie,spinach,olive oil
   2. burgers,meatballs,eggs
   3. chutney
   4. turkey,avocado
   5. mineral water,milk,energy bar,whole wheat rice,green tea

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.


Most Popular Items
wheat rice'
olive oil'
hot dogs'
tomato juice'
cottage cheese'
frozen vegetables'
green tea'
water' milk'
mineral water'
mineral
cooking oil'
grated cheese'
herb pepper'
french fries'
low fat
chicken'
fresh bread'
frozen smoothie'
spaghetti'
beef' spaghetti'
chocolate'
cookies'
pancakes'
ground beef'
whole wheat
fat yogurt'



The wordcloud with max_words=25 will print top 25 frequency items from the transcations. The wordcloud with max_words=50 will print top 50 frequency items from the transcations.

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

   mineral water, eggs, spaghtti, 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.

   [[True False True True False True False] [True False True False False True True] [True False True False False False True] [True True False False False False True] [False False True True True True False] [False False True False True True True] [False False True False True False True] [True True False False False False True]]

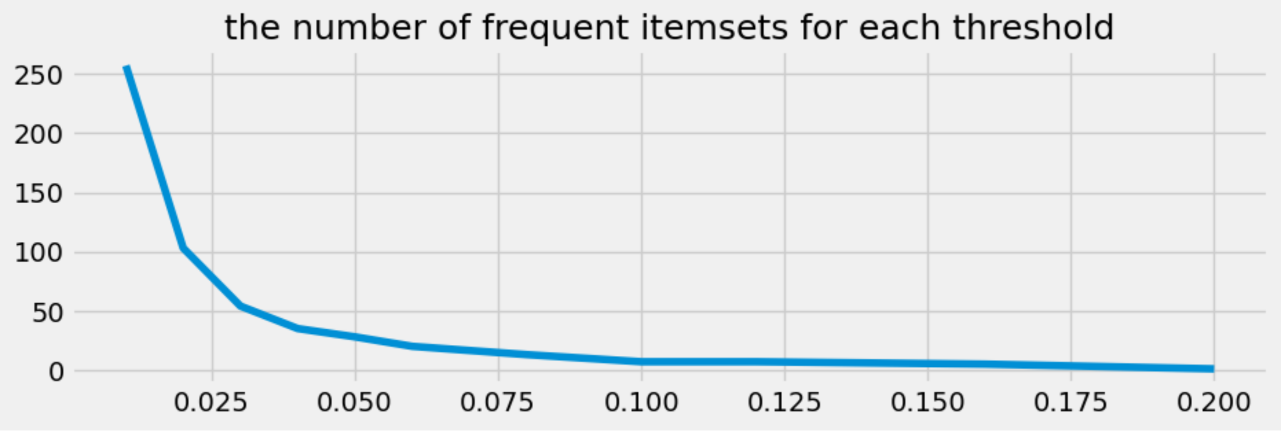
	Apple	Bananas	Beer	Chicken	Milk	Rice	nan

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

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

   There are 120 unique items present.

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.

   The number of frequent itemsets decreases as the support threshold increases. With higher support threshold, there must be fewer itemsets that satisfy the requirements.

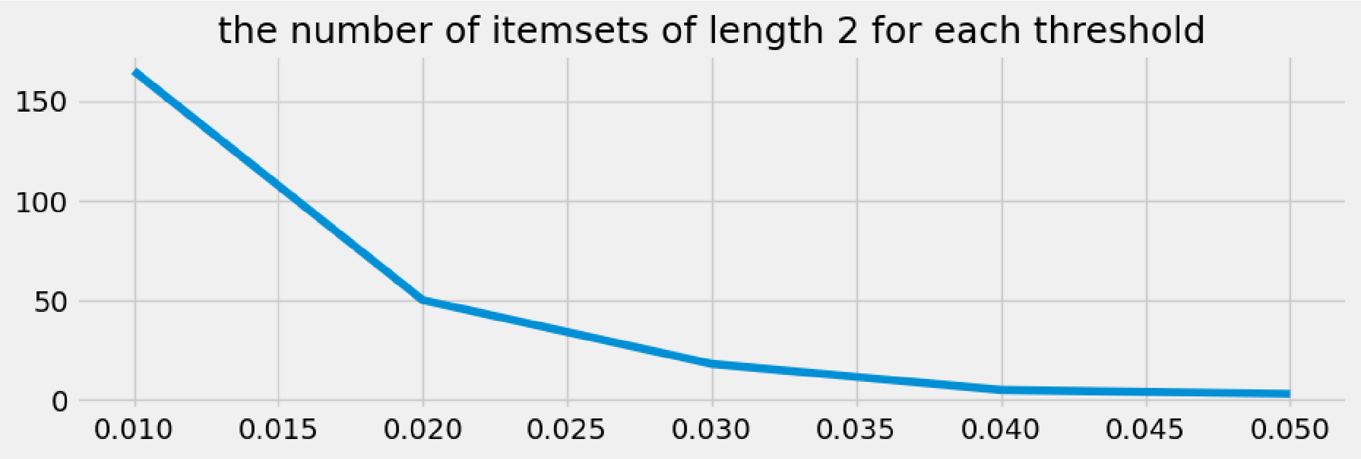
9. At support threshold 1%, we see frequent itemset of size three along with size 2 and 1. However, at the support threshold of 2%, we observe itemsets of size 1 and 2 only. Why do you think this is so?

   If an itemset meet the requirement of min_support > 0.02, it must has a support greater than 1%.

   The set of frequent itemsets at the support threshold of 2% is a subset of the set of frequent itemsets at the support threshold of 1%.

   There are 3-itemsets has a support > 0.01, but their supports are all less than 2%.

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 number of frequent itemsets of length 2 decreases as the support threshold increases. With higher support threshold, there must be fewer itemsets that satisfy the requirements.

11. For the following itemset, write down its corresponding support value:
    - sup(Mineral Water) = 0.238368
    - sup(Chocolate) = 0.163845
    - sup(Eggs) = 0.179709
    - sup(Eggs, Mineral Water) = 0.050927
    - sup(Chocolate, Mineral Water) = 0.05266