1. My task is using the apriori algorithm to mine a dataset and find commonly bought items together. I chose this dataset because I was curious on what do people commonly buy together, and if I am a part of that average. I think it would be fun to use this data since it is something easy to understand like food and something that my day revolves around. One exploratory question I will want to investigate is how many combinations I will want to compare at once. Do I want to see if two types of food are bought together, three types, four types? I think I will need to do some deeper research in the dataset and an idea I have is looking into the average number of items bought on a grocery trip and use that as my combination count. Another question I will need to think about solving is what support I want to hold. I want my solution data to be something interesting and fun to look at, but not too overwhelming or underwhelming. I think it will take some playing around with seeing which support will work out the best for me.
2. The source of my data comes from <https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset> . I liked this dataset since it provided an adequate amount of data to parse through. After reading through the data, I also believe it will provide sufficient items for me to work through and compare throughout the project.
3. For this task I plan to use Apriori algorithm to solve it. I chose this algorithm because it is the one, I have enjoyed learning about most in this course. I will be using numpy and pandas and libraries to help, but I do not believe I will need any other libraries to help me solve this data set, but this may change as I work through the project. My first step for this will be creating a user profile for each customer and which items they bought. I will then compare this different customer and see if customer bough item x what is the probability that they also bought item y. This is a brief overview of my plan, but I intend to make some sort of customer-item matrix and use that to calculate my results.
4. As for a final product I am looking to return a data file that contains a combination of different food and how often they are bought together. The dataset will show the support for each combination of items. My way of measuring success in this assignment is determine if my code is scalable. My major issues I have in this class is writing efficient code for large scale datasets. I wish to improve upon my code I wrote in homework 1 that used apriori, so it is more efficient and fluid. I think this project will help me learn how to work with a real-life dataset and how to write efficient code that calculates the correct output results.