Adelyn Yeoh – Data Analysis Project
ajyeoh@tepper.cmu.edu

Introduction

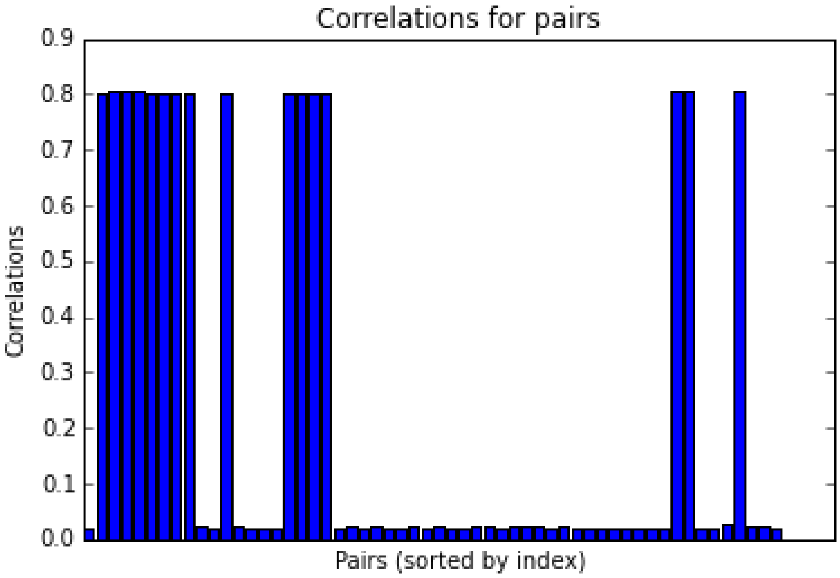
I took two approaches to analyzing the dataset. The first approach I used was to look at the correlation between each item. The second approach I took was to look at the frequency of each possible combination of orders for different group sizes. I chose the second approach as it is easier to implement for larger group sizes. The items that I found that are most commonly purchased together are:

- (item_2, item_7, item_29)
- (item_3, item_5, item_22)
- (item_1, item_9, item_35, item_39, item_42)

I conducted most of my analysis in Python. I used Pandas and Numpy for data analysis and statistical tools. I used Matplotlib to create charts. See corr.py to view code for the first approach, and graphprob.py to view code for the second approach.

Approach

I used two approaches to analyzing this dataset. For first approach I computed correlation between each column and searched for pairs that had a positive correlation of above 0.05. I chose this level because I noticed that the pairs which have positive correlations tended to be either above 0.5 or less than 0.05 (See graph below). So I wanted to analyze the subset of data with stronger correlation.



The second approach I decided to compute the frequency of pairs of orders. I did this in the following manner:

- Computing all possible pairs of orders
- Computing the frequency of each of the order

To implement this approach in Python, I chose the dictionary data structure as a way to store and read through data.

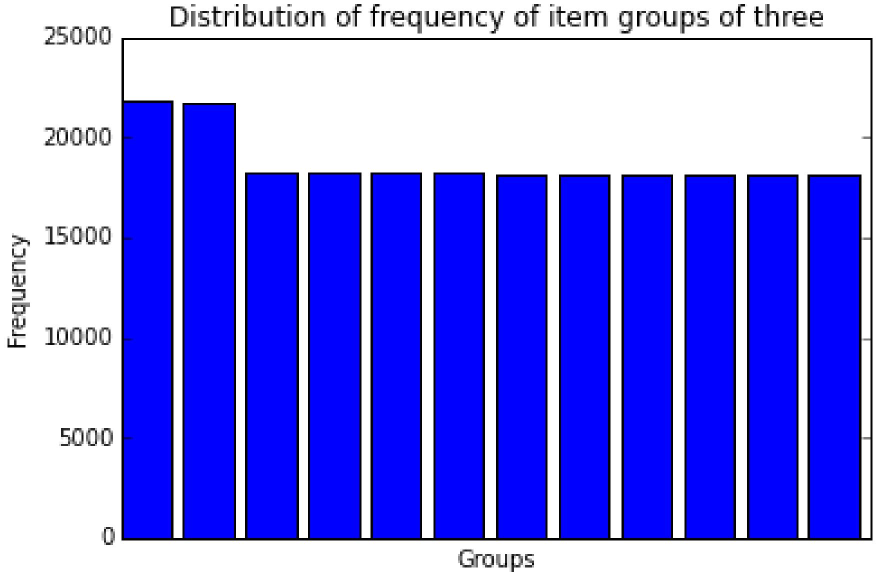
Both approaches yielded the same results:

Pair	Correlation	Weight
(item_1, item_9)	0.802991872	20224
(item_1, item_35)	0.804620425	20245
(item_1, item_39)	0.805169815	20281
(item_1, item_42)	0.806342152	20287
(item_2, item_7)	0.803573041	24328
(item_2, item_29)	0.800563524	24266
(item_3, item_5)	0.802109673	24173
(item_3, item_22)	0.802757187	24151
(item_5, item_22)	0.802248486	24201
(item_7, item_29)	0.803029725	24306
(item_9, item_35)	0.799761815	20152
(item_9, item_39)	0.802520759	20228
(item_9, item_42)	0.800879037	20183
(item_35, item_39)	0.80403863	20247
(item_35, item_42)	0.803996542	20231
(item_39, item_42)	0.805704261	20288

However, I noticed that each transaction tended to contain orders for more than two items. Thus, I chose the second approach to determine items that are most frequently purchased together. This is because the second approach tends to scale a lot easier.

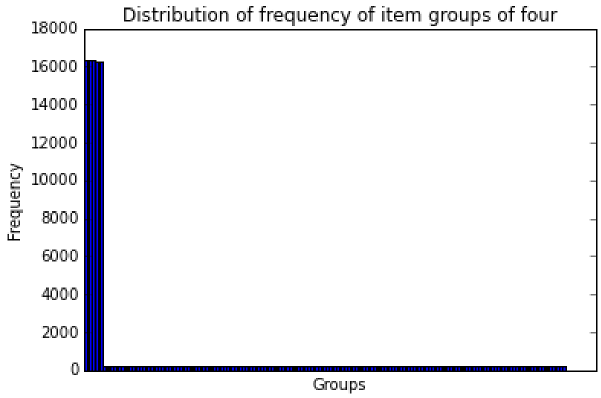
Results

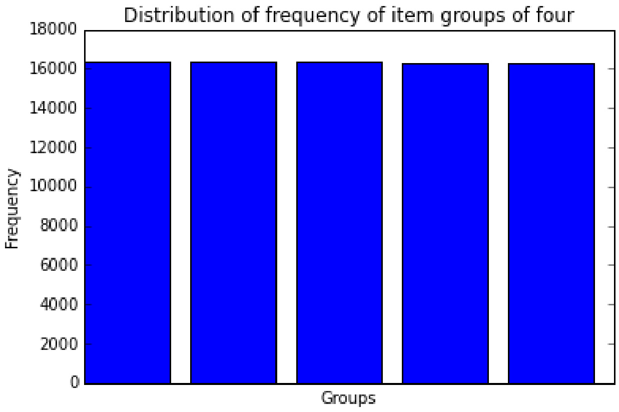
Groups of three



- (item_2, item_7, item_29): 21767
- (item_3, item_5, item_22): 21670
- (item_1, item_39, item_42): 18235
- (item_35, item_39, item_42): 18207
- (item_1, item_35, item_39): 18186
- (item_1, item_35, item_42): 18179
- (item_1, item_9, item_39): 18177
- (item_9, item_39, item_42): 18173
- (item_1, item_9, item_42): 18163
- (item_9, item_35, item_39): 18138
- (item_1, item_9, item_35): 18108
- (item_9, item_35, item_42): 18099

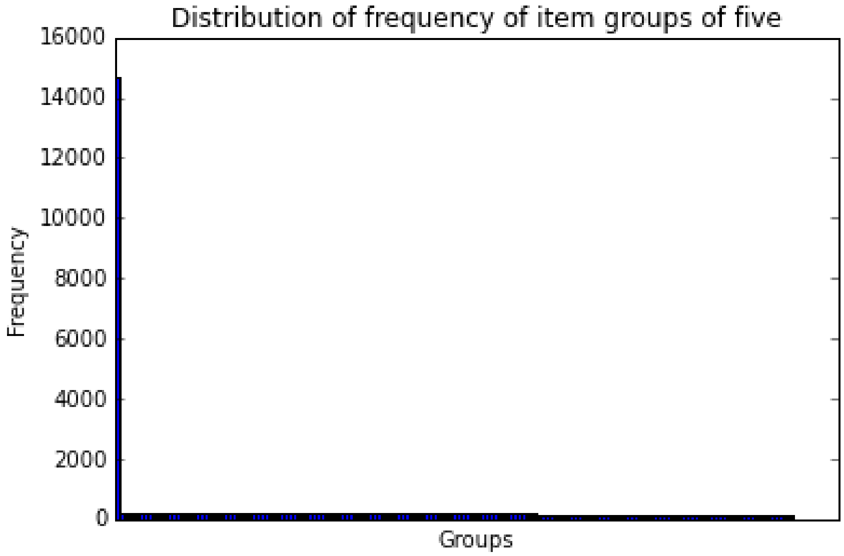
Groups of four





- (item_1, item_35, item_39, item_42): 16374
- (item_1, item_9, item_39, item_42): 16374
- (item_9, item_35, item_39, item_42): 16332
- (item_1, item_9, item_35, item_39): 16314
- (item_1, item_9, item_35, item_42): 16293

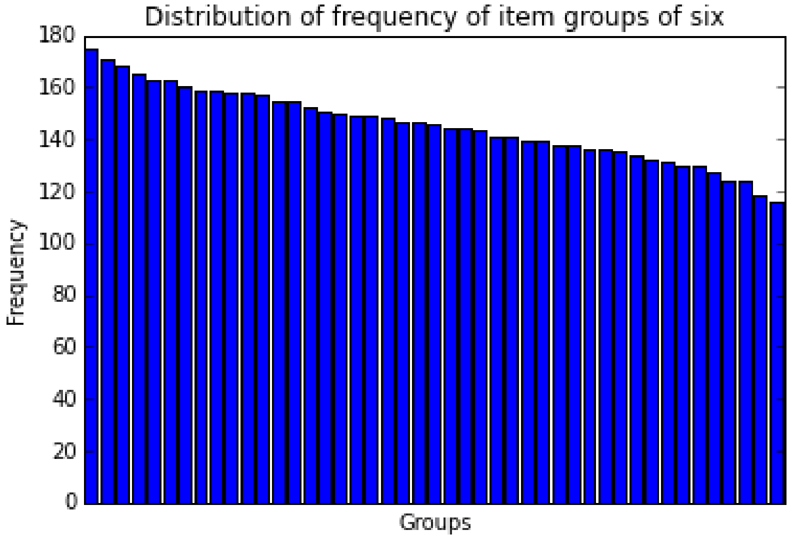
Groups of five



For groups of five, there seemed to be only one group that stood out:

- (item_1, item_9, item_35, item_39, item_42): 14692

Groups of six



For groups of six, there seemed to be a lot more variation to the groups. Frequency of occurrences for groups of six are significantly smaller compared to other grouping sizes. So I chose to stop analysis at this point.

Adelyn Yeoh – Data Analysis Project
ajyeoh@tepper.cmu.edu

Summary

After identifying the groups that were most significant for each group size order, I removed groups that were combinations of larger groups particularly if the frequency of orders are about the same. For example, notice how the group of four is simply a combination of the group of five.

Group of Four	Group of Five
(item_1, item_35, item_39, item_42) (item_1, item_9, item_39, item_42) (item_9, item_35, item_39, item_42) (item_1, item_9, item_35, item_39) (item_1, item_9, item_35, item_42)	(item_1, item_9, item_35, item_39, item_42)

After conducting that analysis, these groups of items are most frequently purchased together is as follows:

- (item_2, item_7, item_29)
- (item_3, item_5, item_22)
- (item_1, item_9, item_35, item_39, item_42)

Discussion

While I analyzed the data this way, one thing I did not take into consideration is the average order size for 100k transactions. Perhaps one way to extend this analysis is to consider groups of items ordered, given a particular order size.