**THE ALGORITHM**

One of the algorithms that we chose to run our data through was the Apriority algorithm that was discussed in class. As we learned in class, the Apriority algorithm identifies frequent individual item sets in a set of data and extends that to larger and larger item sets as long as they appear above a given threshold. The item sets then found can be used to determine some association rules that hopefully, will highlight general trends in a given set of data.

**THE DATA & THE MOTIVATION**

Our lending data had a lot of different dimension associated with it so when we ran the Apriority algorithm we extracted a few specific columns we wanted to test. We generated frequent item sets for the following four columns: Fund Requested, Employee Title, Annual Income and Reason for Requesting a Loan. The reason we chose to look at these four columns specifically was to answer some specific queries. We were interested in seeing possible correlation between the funds requested versus the amount of money the individual made. The hope was to identity some kind of connection between how much money one made and if there was some kind of trend in how much money they may need. Another trend we could identify using these columns was the possibility of correlating Annual Income and Employee title. There were a lot of records in the data and we could potentially make a link between amount made and job title. Lastly, we were hoping to see a connection between the amount of funds requested and the reason for requesting a loan. This was probably the most promising connection we could make. Some of the reasons for requesting a loan included buying a car, debt consolidation and a wedding. It would have been very interesting to see if there was a consistent range for people requesting a certain amount and what they were putting the money toward.

**PROCESSING THE DATA**

Part of the code we used was the Apriori algorithm provided to us in class. Alongside this we had to clean up the data and extract the columns we were interested in using. Since there were a lot of different values for loans and annual income, we put these values into ‘buckets’ to better cluster our data. For every integer or float value, we rounded the value down to the closest 1000 value. This allowed us to better cluster our numbers since many were unique and produce better results. Since we had a lot of records in our data, it took a number of hours to run each time.

**RESULTS**

**FUTURE WORK**

One of the tests we wanted to run was amalgamating the Accepted and Rejected data into one trial. It would be interested to see if different types of transactions are affected by being accepted or rejected. One of the problems with doing this with our current data was that the rejected data sample did not have all the necessary dimension. What could be done in the future is a modification of the analysis we did using different dimensions. Because we had so much data with so many different dimensions there are a lot of interesting questions we could be asking. Having so many records means we can look for many different rules in many different ways but unfortunately, there isn’t enough time to answer all those questions right now.