Caleb Warwick and Grace Okamoto
CPT_S 315
Professor Jana Doppa
Course Project Proposal

Our data mining task will be to predict a user's next Instacart purchase, using data on between 4-100 of their previous purchases obtained through Kaggle: Instacart Market Basket Analysis | Kaggle. We chose this data set as it is closely related to the work we have done in class and in our homework assignments, but the nature of our research question is intricate enough that we will be required to combine a variety of the skills we have developed in this course thus far in order to obtain the most accurate results.

For our software tools, we intend on using Python and Jupyter Notebooks in Visual Studio Code, as those are the tools we have found most helpful for our previous homework assignments. Importing the Numpy and Pandas libraries to read and manipulate our data into the appropriate arrays and tables will reduce the amount of programming needed and allow our program to run more efficiently.

Our methodology will most likely adapt once we have started on the project and gain a better understanding of which algorithms will contribute to the most accurate results. However, some initial ideas include using the A Priori algorithm for frequent items as a way of predicting which items are most frequently appearing for a particular user. In addition, we could use a form of collaborative filtering to predict any new items that a user may not have purchased previously.

A cursory look at the data reveals that it is split into three subsets: prior, train, and test. The prior data consists of the historical orders of a particular user and the training data is used to develop our algorithm. We will compare the results of our program to the test data in order to determine how accurate our algorithm is, and to explore and analyze what additional changes we may need to make to our program in order to get the most accurate results.