The overall purpose we have chosen for this project is to explore the data set [1] using multiple data mining techniques studied in this course in order to determine strategies for marketing to Instacart shoppers in ways that will generate increased overall revenue. At this stage we have several ideas for what conclusions will be able to be drawn from the data. The first is making general staffing suggestions based on order volume at particular times of day and days of the week, which can also inform times for delivering marketing materials to customers. The second is exploring purchases as organic or non-organic products, and then determining if customers fall into the category of organic shoppers to target marketing of organic items to those customers, which could also lead to a possible need for an expansion of organic products available on Instacart. The third idea is to explore the purchase volume and frequency of customers to determine what a high value customer “looks” like. For example, a high value customer might be one that makes large orders on a regular basis or a customer that makes smaller frequent purchases. The goal being to create a recommendation for a rewards program that will target the high value customers and not only reward their loyalty but encourage them to increase their purchase volume.

The data we are exploring is a set of over three million Instacart grocery shopping transactions that have been made publicly available for learning purposes such as this project. The data includes orders with many attributes including; day of week, hour of the day, days since last order, and an evaluation identifier. Since the data set is designed for mining education and training, each order has been designated to the purpose of training, testing, and prior orders. There are nearly two-hundred thousand training records in the set which is a very large amount to process with a personal computer, so we plan to do random samplings of one-thousand records between ten and twenty times to train the models we use to make predictions.

There will need to be a substantial amount of pre-processing to the data before we can begin modeling. For example, we plan to make recommendations about staffing based on purchase time and day of the week so in order to use decision trees or random forest the data will need to be categorized to the time of day; day (7a - 3p), evening (3p - 11p), overnight (11p - 7a). The data will also need to be aggregated to determine customer value, by selecting a specific period of time for example 90 days, so that the highest dollar value ordering individuals can be identified to draw conclusions about shopping patterns.

At this point, we know that our primary strategies will be to use decision trees, random forests, and other techniques learned in class to make grouping or classification decisions from the training data to make predictions about the test data about those classifications. Each group member has begun exploring the data to identify any categories that stand out as classification determinants. As we explore the data and make models to predict future purchases there will likely be changes to the specific models we use, but these are our initial thoughts.

Data Set Reference

“The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on May 1, 2019.