STEVENS
INSTITUTE of TECHNOLOGY
School of Business

STEVENS
INSTITUTE of TECHNOLOGY
School of Business

Revision: 1.0

Table of Figures

COURSE PROJECT PART 3

ASSIGNMENT M11.E1.1

ASHLEY LAU
BIA-678 BIG DATA TECHNOLOGIES
Stevens Institute of Technology – Summer 2023

1 EXECUTIVE SUMMARY

Instacart is an American online service via website and mobile app that allows customers to stop from a network of groceries at the ease of their fingertips. A customer selects the grocery store they'd like to shop from as well as a list of items for purchase. After an order is processed, a personal shopper will retrieve the items and deliver at a pre-selected time frame. Customers can choose from a set of delivery options such as contactless delivery, pickup and club store deliveries. Additional fees are included like delivery fees ($3.99 for orders $35+ or $7.99 under that amount), pickup fees, service feeds (dependent on grocery location, alcohol purchases, etc.) There is an additional option of Instacart+ membership for other benefits. Instacart has become a popular service due to its broad network of access to grocery chains, like Kroger, Safeway, Costco, and Target, convenience, and was an immense help for families at home during the pandemic.

To get a better understanding of Instacart user's client base and shopping patterns, customer segmentation was performed amongst product selection and user id. The summary will highlight the technical approach to calculate the data segments and point out key insights from the analysis.

Before conducting customer segmentation, reference master data was designed to further analyze product orders into separate clusters. By utilizing the product data, Instacart order, department/aisle referential data, and prior order information, a master data frame was built capturing product features. Clustering segmentation was done with a combination of principal component analysis (PCA) and KMeans clustering. Principal component analysis (PCA) is a statistical procedure that is used to reduce the dimensionality of a dataset while preserving as much information as possible. This is done by transforming the dataset into a new set of variables called principal components, which are ordered by their ability to explain the variance in the original dataset. K-means clustering is an unsupervised learning algorithm that groups data points into k clusters, where k is a user-specified number. K-means clustering is a popular clustering algorithm because it is simple to implement and understand. This method outputs 5 distinct clusters.

At a quick glance, we initially identified clusters by calculating the count of order amounts by department and sum of average reorders. The range of order department amounts varied from 3,021 to little less than 17,000. Of this cluster 1 and 3 were the most significant with order department amounts greater than then thousand respectively. From the average reorder perspective, the range was between 643 – 7,635; surprisingly the most significant clusters were cluster 1 and 4 in this comparison. Visualizations for both segments are highlighted in Figure 1 and Figure 2. Furthermore, to get a better sense of what items are bought by customers in these clusters, product details were checked. In cluster 4, the top 3 products ordered aligned with popularity rankings mentioned previously – bananas were number 1 product in this cluster followed by bag of organic bananas and organic strawberries. In cluster 1, the top 3 products were organic garbanzo beans, organic diced tomatoes, and organic frozen peas.

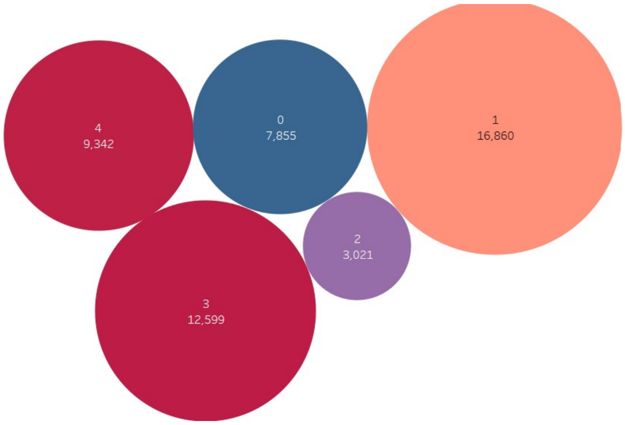

Figure 1 Cluster by product: Order Amount Department

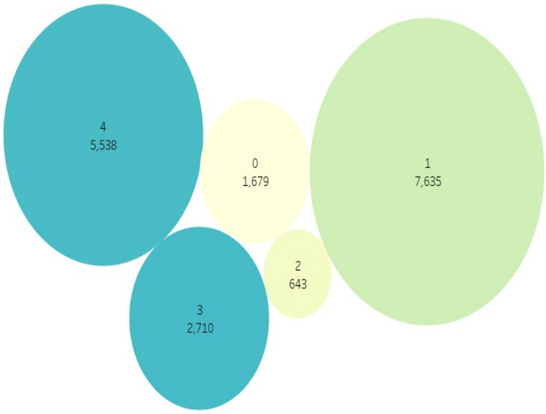

Figure 2 Cluster by product: Sum(Average Reordered)

A second method, to conduct customer segmentation, was done with the help of Tableau data visualization tool. Tableau is a business intelligence (BI) and analytics platform that helps people see and understand data. Tableau is easy-to-use and has capabilities to create collaborative, scalable dashboards. From the list of all Instacart user ids in the database Tableau auto identified 8 clusters; to narrow scope an additional filter was added to force Tableau to arrange the data into 4 clusters. .

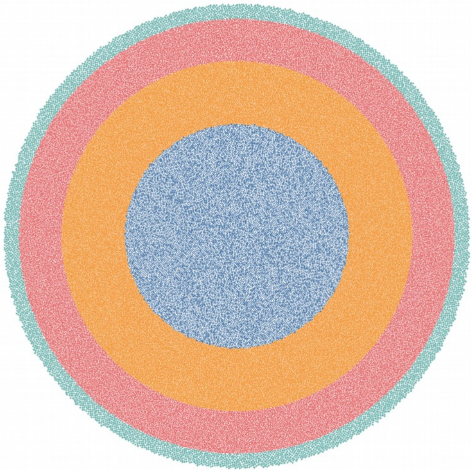

Figure 3 Cluster by UserID: 4 Populated Clusters

Another aspect of customer segmentation entailed looking at the hour of day orders were placed. From this we found 4 segments of clientele: the first group of customers places orders throughout a 24-hour cycle with a count of orders under 40K. The second group of customers places orders between 7am-10pm with order count from 90K-140K. The third group of customers places orders between 8am-6pm with order count from 180K-228K. Lastly the fourth group of customers places orders between 9am-4pm with order count from 257K-288K. This is shown in Figure 4 below.

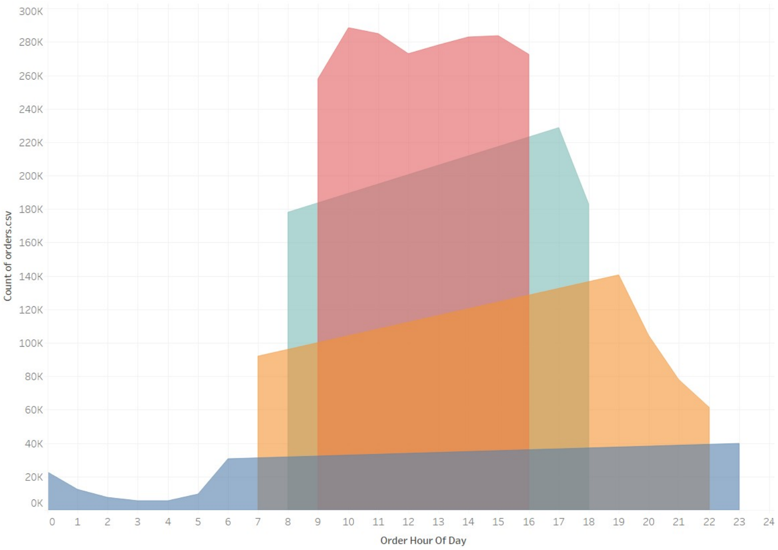

Figure 4 Cluster by UserID: Order Hour of Day

Two more views of customer segmentation were done. In terms of days since prior order, 5 clusters were found. The biggest cluster of Instacart users made an Instacart order 3-6 days ago; second largest cluster of Instacart users were customers who placed an order a week or exactly 1 month (30 days) ago. In terms of day of week, there were 4 main clusters discovered. The 4 clusters were people who ordered on Sundays/Mondays, Tuesdays only, Fridays/Saturdays, and Wednesdays/Thursdays respectively. Of this, the first cluster had the greatest number of orders, once more in line with analysis done in the second part of the project.

Finally, the last version of customer segmentation analyzed the average number of days since prior order, with setting arranged to create 10 individual clusters; top 3 dominant clusters were cluster 9, 8, and 5, in that order. Common reordered item categories in cluster 9 were baked goods, beverages, canned goods, dairy eggs, frozen foods, pantry items, produce and snacks. Similar trends can be found in cluster 8 and 5 as well; this is not as surprising of a finding since these department items are essential food items for households.

Overall, this analysis has shown interesting slices of Instacart data to portion customers and Instacart orders into separate groups. While being able to understand clients better and target marketing more effectively to make better pricing, product development, or marketing decisions, there are some downsides to the analysis as well. If instructions on how to segment customers are not clear or properly analyzed the results for customer segmentation can be inaccurate or lead to endless rabbit holes. The insights could also hold bias against certain groups of customers. It is also time-consuming to do as the process is highly manual in nature. Analysts require the right data and demographic details to get clearer pictures of the customer base and forge an understandable profile of the user’s wants and needs.