Customer segmentation findings and business impact; ROI potential and implementation recommendation

With the hierarchical (agglomerative) clustering chosen for deployment, customers fall into one of four main groups (see appendix for a plot of feature distributions for each cluster):

- **Bulk Buyers**: These customers not only make more purchases in total than other groups, but buy more items per order. Consequently, they have by far the highest monetary value for the company. Most customers in this group have made a purchase in the last 100 days, but around a quarter have not, and re-engaging them could lead to significant profits. Since this group also saves more money through discounts than others, offering a discount on purchases with a high enough monetary value (e.g. 10% off orders of $500 or more) could be effective.
- **Engaged:** Nearly all of these customers have made a purchase in the last 100 days, and they typically purchase between 5 and 10 items per order. While not as profitable as the bulk buyers, these customers are loyal buyers.
- **At Risk:** The lowest-spending group, typically only buying a few items per order and less likely to have placed an order recently. The business should focus on retaining these customers so they do not slide into the fourth group…
- **Disengaged:** These customers' lifetime spending is on par with the Engaged group, but they have no recent purchases. Business strategy should focus on identifying why these customers "churned" and whether addressing the causes of churn could bring them back.

Key association rules and recommendation strategies

Because the dataset contained few transactions for each specific product name, I focused on association rules between types of products. Even so, the support for any given itemset tended to be quite weak. Using a very low minimum support threshold of 0.01, the rules with confidence of 0.25 or higher all involved the "Binders" subcategory. Storage, Appliances, Fasteners, and Bookcases are all associated with binder purchases. Customers who bought items in any two of the categories Phones, Paper, and Binders also tended to buy the third (all itemsets derived from this rule met the support and confidence thresholds). That is, a binder would be a good "add-on" purchase to suggest to customers buying other types of office supplies. However, ROI might be lower than ideal given that the lift for all association rules was close to 1.

Implementation methodology and results

Clustering methods tested included K-Means (K = 5 selected based on elbow and silhouette methods), hierarchical clustering, HDBSCAN (which failed to produce meaningful clusters – almost all data were assigned to the same cluster), and Gaussian Mixture (which produced three clusters, corresponding approximately to the Bulk Buyers, Disengaged, and a combination of the Engaged and At-Risk segments described above). The hierarchical clustering solution was selected for its explainability (clear distinctions between customer types).

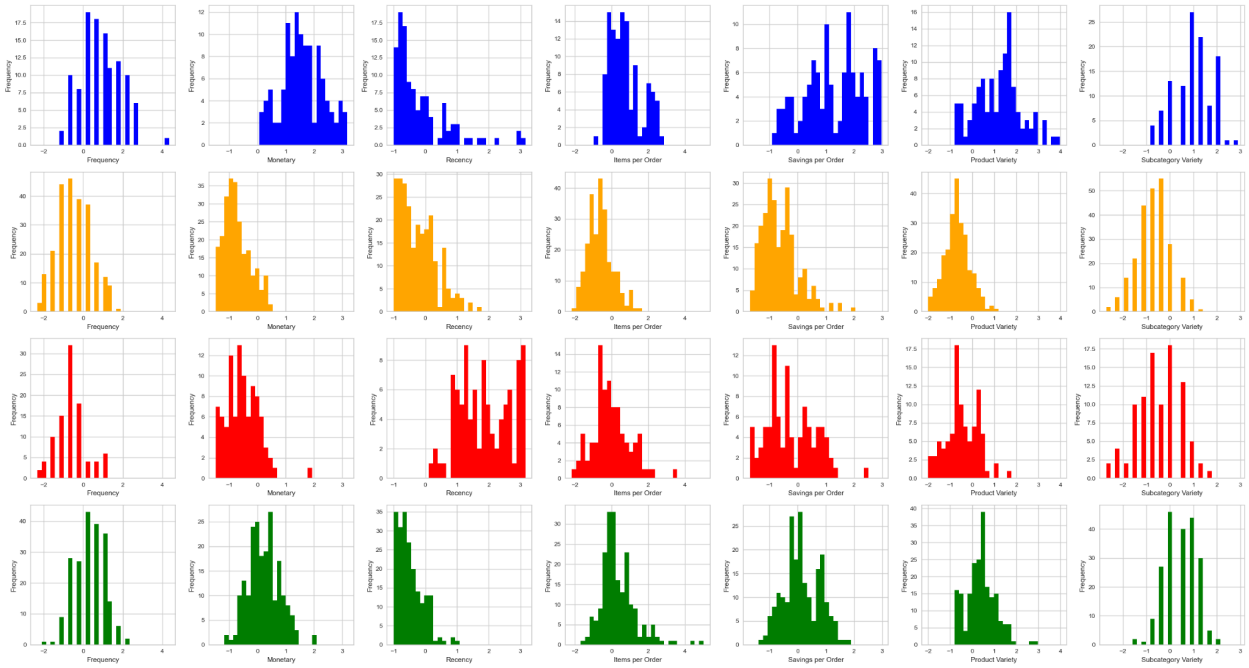
Limitations and future improvements

The dataset is small compared to the number of products, making it less than ideal for association rule mining. For clustering, considering the specific types of items purchased rather than only spending- and variety-related metrics might produce a more sophisticated model.

Appendix: Distribution of features by cluster

A larger, interactive version of this plot (with a smaller bin size) is accessible in tab 2 of the Streamlit application. Clusters are:

- Blue: Bulk Buyers
- Orange: At Risk
- Red: Disengaged
- Green: Engaged



Circular plot based on code provided in lecture: