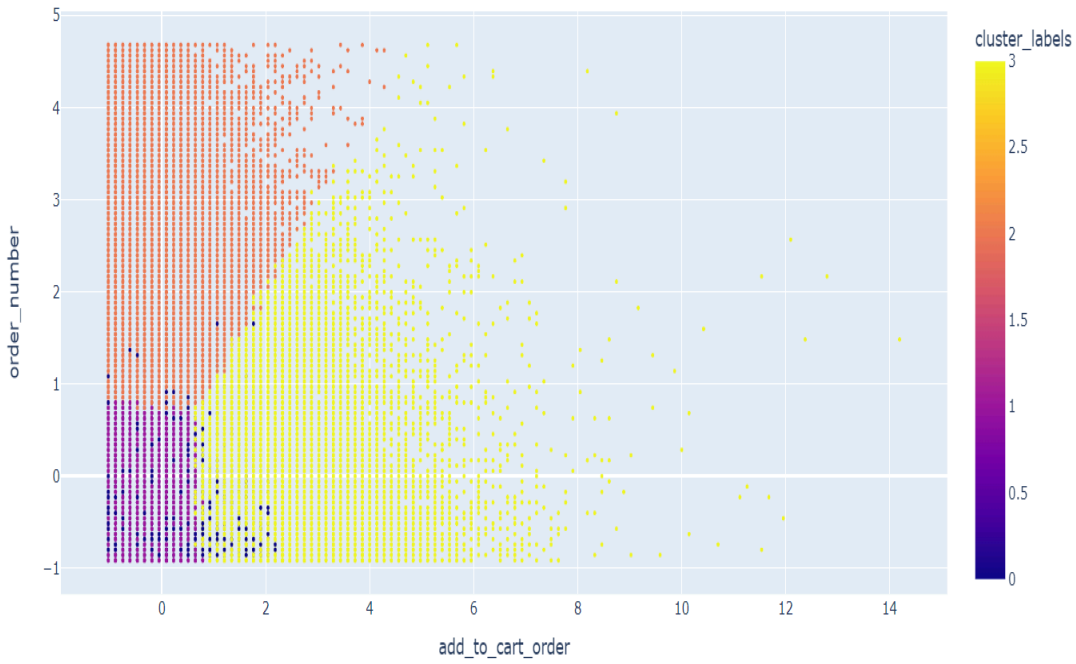
Final Project Deliverable (BA 820_T8)

FUTUREFRIDGE WITH INSTACART: PRECISION GROCERY PREDICTIONS

Appendix:

1. Methodologies

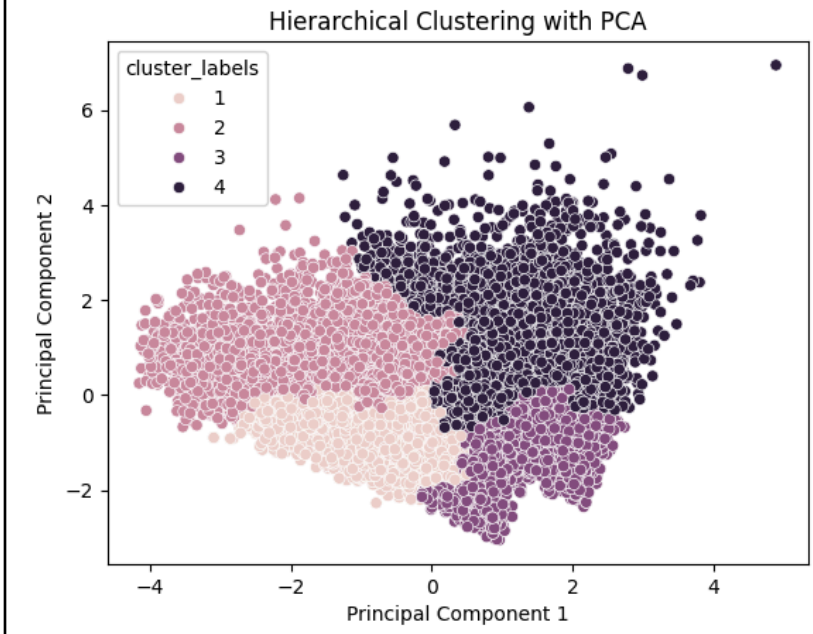
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
5	(Bag of Organic Bananas)	(Organic Hass Avocado)	0.121466	0.067249	0.019503	0.160563	2.387582	0.011334	1.111163	0.661518
4	(Organic Hass Avocado)	(Bag of Organic Bananas)	0.067249	0.121466	0.019503	0.290009	2.387582	0.011334	1.237388	0.623067
9	(Bag of Organic Bananas)	(Organic Strawberries)	0.121466	0.082615	0.018134	0.149296	1.807120	0.008099	1.078383	0.508384
8	(Organic Strawberries)	(Bag of Organic Bananas)	0.082615	0.121466	0.018134	0.219503	1.807120	0.008099	1.125609	0.486855
18	(Banana)	(Organic Strawberries)	0.144732	0.082615	0.017326	0.119708	1.448977	0.005368	1.042136	0.362294



2. Challenges

3. Practical applications of our analysis

4. Result



Coding Contribution:

Name of the Contributor	Topics	Coding Contribution
Aishwarya Jayant Rauthan	Data Cleaning, Market Basket Analysis Approach- 1, Tsne	21%
Khushi Khushi	Market Basket Analysis Approach- 2, Hierarchical Clustering	20%
Saachi Dholakia	EDA, Inferences, PCA	21%
Xinyuan Hu	EDA & Kmeans	19%
Tiancheng Yang	Plt.px, EDA & Kmeans	19%

Links

1. **Colab:**https://colab.research.google.com/drive/1Bg1ETFVKgLqUW6TPAEGXoj1kCkPKpUhO?usp=sharing
2. **Github:** https://github.com/lilchengzi/BA820-project.git
3. **Presentation:**https://docs.google.com/presentation/d/1V087R3D1SnE8L3T9xM9nsu26hRQIazkMRBUrDk3xkL4/edit?usp=sharing
4. **Kanban:**https://cyclic-columnist-bd0.notion.site/b46e409361db46c79ef785d2467e5ba4?v=7d13fa5575ff44d48a357b3870210e5a&pvs=4

Methodologies:

1. **Market Basket Analysis:** Our market basket analysis has yielded crucial insights for retailers. We've pinpointed strong associations like 'Organic Hass Avocado' and 'Bag of Organic Bananas', suggesting lucrative opportunities for strategic product placement. By understanding common buying patterns such as the affinity between 'Organic Strawberries' and 'Bag of Organic Bananas', personalized recommendations can be tailored, enhancing customer satisfaction. Optimizing inventory management to ensure consistent availability of these popular items can streamline operations and drive revenue growth. Leveraging metrics like support, confidence, and lift, our analysis provides a robust foundation for informed decision-making in retail strategies.
2. **K-means:** Our k-means clustering exploration focused on determining the optimal number of clusters. We found a notable shift in the within-cluster sum of squares decline rate after identifying six clusters, supported by similar silhouette scores for two or three clusters. With our market segmented into six distinct clusters, we're ready to execute targeted marketing campaigns tailored to each segment's preferences. Understanding customers' needs within each cluster enables personalized marketing initiatives, boosting engagement and conversion rates. Additionally, our clustering results form a basis for developing recommendation systems, enhancing the shopping experience and fostering customer loyalty.
3. **TSNE:** In our approach to dimensionality reduction using t-SNE, we first sampled the data to address computational complexities. Following this, after applying t-SNE and clustering the data based on the two most significant dimensions, we observed a distinct division into four quadrants. Cluster 0 predominantly comprises lower values of component 2, whereas cluster 1 is characterized by higher values of component 2. Similarly, cluster 2 exhibits elevated values of component 1, while cluster 3 predominantly features lower values of component 1. Notably, there is no overlap between clusters, indicating the efficiency of the clustering process. This delineation provides valuable insights into the underlying structure of the data, aiding in further analysis and interpretation.
4. **PCA :** In our analysis, we employed PCA as a robust dimensionality reduction technique, particularly suitable for handling sparse datasets commonly encountered in text processing and recommendation systems. By sidestepping the necessity for centering, PCA facilitated efficient computation while preserving essential information. Utilizing PCA allowed us to extract meaningful components capturing variance in the data, aiding in exploratory data analysis and feature extraction. This methodological approach enabled us to uncover significant structures within the dataset, enhancing our understanding of underlying patterns. Subsequently, we applied K-means clustering post-PCA to identify distinct clusters within the reduced feature space. Our observations revealed well-separated clusters along the first principal component, indicative of its role

in delineating cluster boundaries. Additionally, we identified outliers, suggesting potential anomalies or unclustered data points warranting further investigation. Overall, our utilization of PCA and subsequent clustering techniques yielded valuable insights into the inherent structure of the data, facilitating informed decision-making across various applications.

5. **Hierarchical clustering** : In our exploration of hierarchical clustering using t-SNE and PCA, we employed silhouette scores to determine the optimal number of clusters. For t-SNE, the silhouette scores indicated that the optimal number of clusters is likely 4, as it displayed a high score comparable to 2 and 3, yet with the advantage of providing greater separation between the clusters. Similarly, when utilizing PCA, silhouette scores suggested that clusters of 2, 3, and 4 have the highest scores. Despite this, we opted to proceed with 4 clusters to maintain uniformity and consistency across our analysis. This decision was informed by the high score associated with 4 clusters, ensuring robust cluster separation and facilitating clearer delineation of distinct groupings within the data. By leveraging these techniques and informed decisions on cluster numbers, we aim to uncover meaningful patterns and structures within the dataset, aiding in deeper insights and actionable outcomes for various applications.

Challenges:

In our journey with the Instacart dataset, we faced some serious hurdles due to its enormous size. Wrangling this data required extensive sampling and resampling efforts to make it workable for analysis. However, these processes often push the limits of our computers, leading to frequent crashes in our collaborative environment due to memory constraints.

To execute market basket analysis, we had to employ two distinct approaches. Initially, we utilized sampled data to derive insights. However, this method had inherent limitations as it only provided a snapshot of the data. Since we were only working with a fraction of the data, there was a risk of missing out on subtle patterns present in the complete dataset. In our second approach, we opted for a more exhaustive method by generating combinations of products and then sampling these combinations for analysis. This promised a broader understanding of associations within the data but also presented computational challenges.

In addition to market basket analysis, we also encountered challenges when executing t-SNE and Hierarchical clustering for dimensionality reduction. Due to the dataset's size, we had to resample the data to make these techniques computationally feasible. Despite these challenges, we managed to implement both t-SNE and Hierarchical clustering successfully, enabling us to visualize and understand the high-dimensional data in a more interpretable manner.

In conclusion, working with the Instacart dataset wasn't without its challenges. However, through perseverance and creative problem-solving, we were able to overcome computational limitations and extract meaningful insights that could have real-world implications.

Practical applications of our analysis：

Our analysis provides practical applications that span the breadth of the grocery industry, enabling businesses to make data-driven decisions that drive growth and enhance customer satisfaction. By utilizing insights derived from market basket analysis and customer segmentation, businesses can offer personalized product recommendations and targeted marketing campaigns tailored to individual preferences and behaviors. This not only improves the shopping experience for users but also increases the likelihood of repeat purchases and customer loyalty. Moreover, our analysis facilitates inventory management optimization by identifying popular products and predicting demand patterns, allowing businesses to maintain optimal stock levels while minimizing excess inventory and stockouts. Additionally, our insights enable businesses to optimize operational processes, such as delivery route planning and supply chain management, resulting in cost savings and improved operational efficiency. Overall, these practical applications empower grocery retailers and delivery apps to adapt to changing market dynamics and deliver exceptional value to their customers.

Result: Our project aimed to revolutionize the grocery ordering and delivery experience through data-driven analysis and machine learning techniques. Through comprehensive exploration and analysis of the Instacart dataset, we have achieved several key outcomes:

1. **Enhanced Customer Experience**: By leveraging market basket analysis, we uncovered valuable insights into product associations and common purchasing behaviors. This enabled us to make personalized product recommendations and offer targeted promotions, ultimately improving customer satisfaction and loyalty.
2. **Increased Sales and Retention**: Through customer segmentation using clustering techniques like k-means and hierarchical clustering, we identified distinct customer segments with unique preferences and behaviors. This allowed us to execute tailored marketing initiatives, leading to higher engagement, conversion rates, and customer retention.
3. **Practical Applications**: Our findings have practical implications for grocery retailers and delivery apps, offering strategies for optimizing inventory management, strategic product placement, and personalized recommendations. These strategies can drive sales growth, streamline operations, and enhance the overall shopping experience for users.