FUTURE FRIDGE WITH INSTACART: PRECISION GROCERY PREDICTIONS

Team 8

Aishwarya Jayant Rauthan, Khushi Khushi, Saachi Dholakia, Tiancheng Yang, Xinyuan Hu

Executive Summary

- Data Information & Preparation
- Problem Statement
- Exploratory Data Analysis (EDA)
- Market Basket Analysis
- K-means Clustering
- Challenges
- Business Relevance(Future work/next steps)

Data Information & Preparation

Data Information

- **Database Size:** 324,345 rows, 15 columns from 30+ million records
- **Sampling:** Reduced to 1% for manageability.

Data Preparation

- **Standardization:** Used Standard Scaler for normalization.
- **KMeans Clustering:** Focused on numerical columns only.
- **Market Basket Analysis:**
  - Combined products into one column.
  - Two approaches: pre and post data sampling.

Column	Description
product_id	Unique identifier for each product.
product_name	Name or description of the product.
aisle_id	Unique identifier for each aisle where the product is located in the store.
department_id	Unique identifier for each department to which the product belongs.
aisle	Name or description of the aisle where the product is located.
department	Name or description of the department to which the product belongs.
order_id	Unique identifier for each order.
add_to_cart_order	Order in which the product was added to the cart within an order.
reordered	Indicator (0 or 1) representing whether the product has been reordered in the past.
user_id	Unique identifier for each user.
order_number	Sequential order number for each order made by a user
order_dow	Day of the week when the order was placed (0=Sunday, 1=Monday, ..., 6=Saturday).
order_hour_of_day	Hour of the day when the order was placed (0-23).
days_since_prior_order	Number of days elapsed since the user's previous order.

Problem Statement

Primary Objective: Predict which previously purchased products will be in a user's next order based on their historical order data.

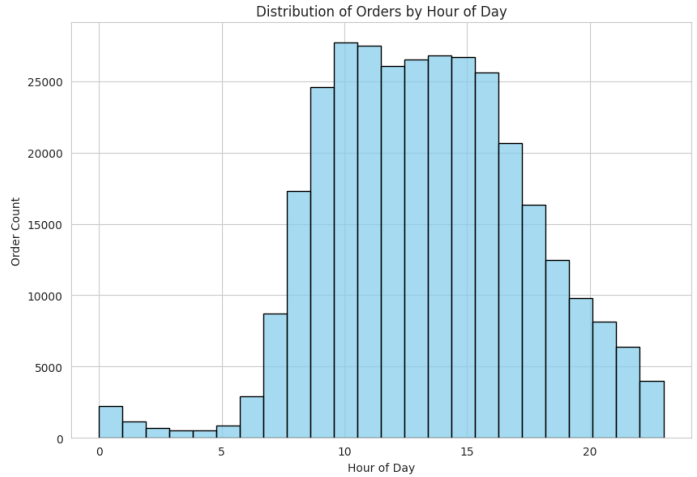
Goal: Improve the shopping experience by providing personalized product recommendations.

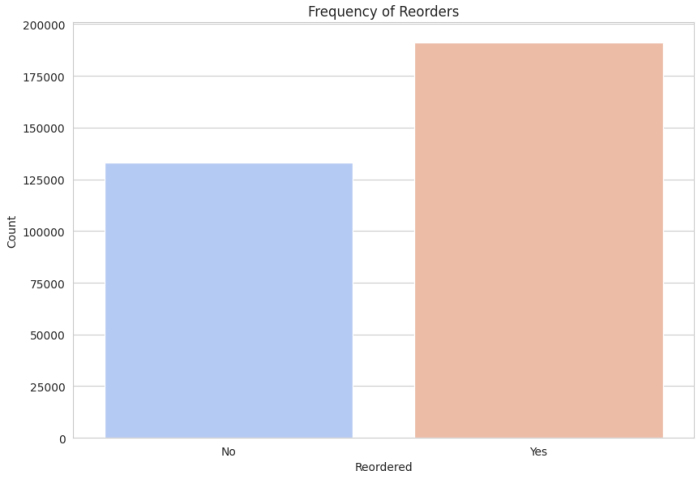
Benefits: Increase customer satisfaction and loyalty by catering to individual preferences and shopping habits.

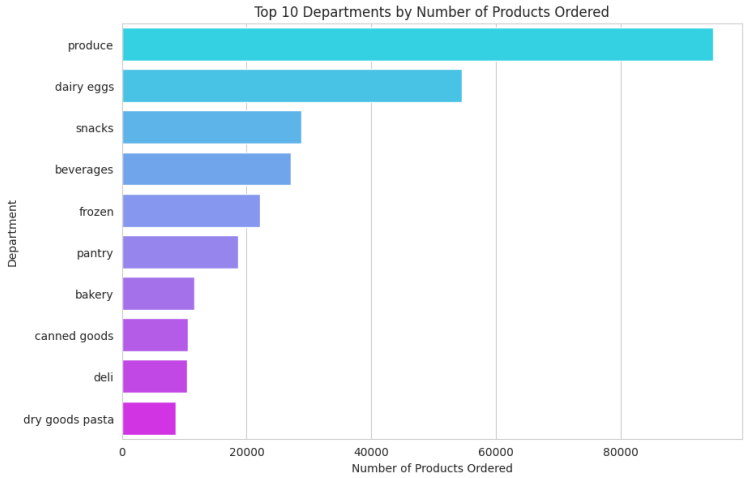
Potential methods: Enhancing the user experience by suggesting relevant products.

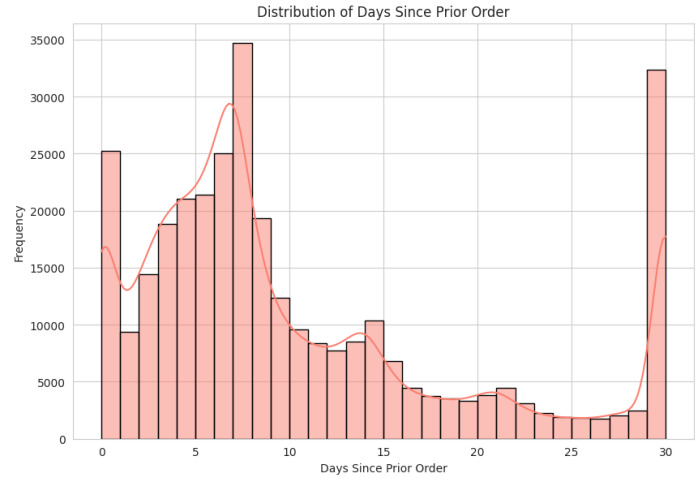
Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)









Plot 1: Order Distribution by Hour

- Morning Increase
- Lunchtime Peak
- Evening Decline.

Plot 2: Reorder Frequency

- High Reorders
- Implications

Plot 3: Department Popularity

- Produce Dominance
- Staple Departments
- Inventory Focus

Plot 4: Order Interval Distribution

- Ordering Spikes
- Routine Orders
- Subscription Potential

instacart

Market Basket Analysis

Market Basket Analysis

min_threshold=0.01

	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

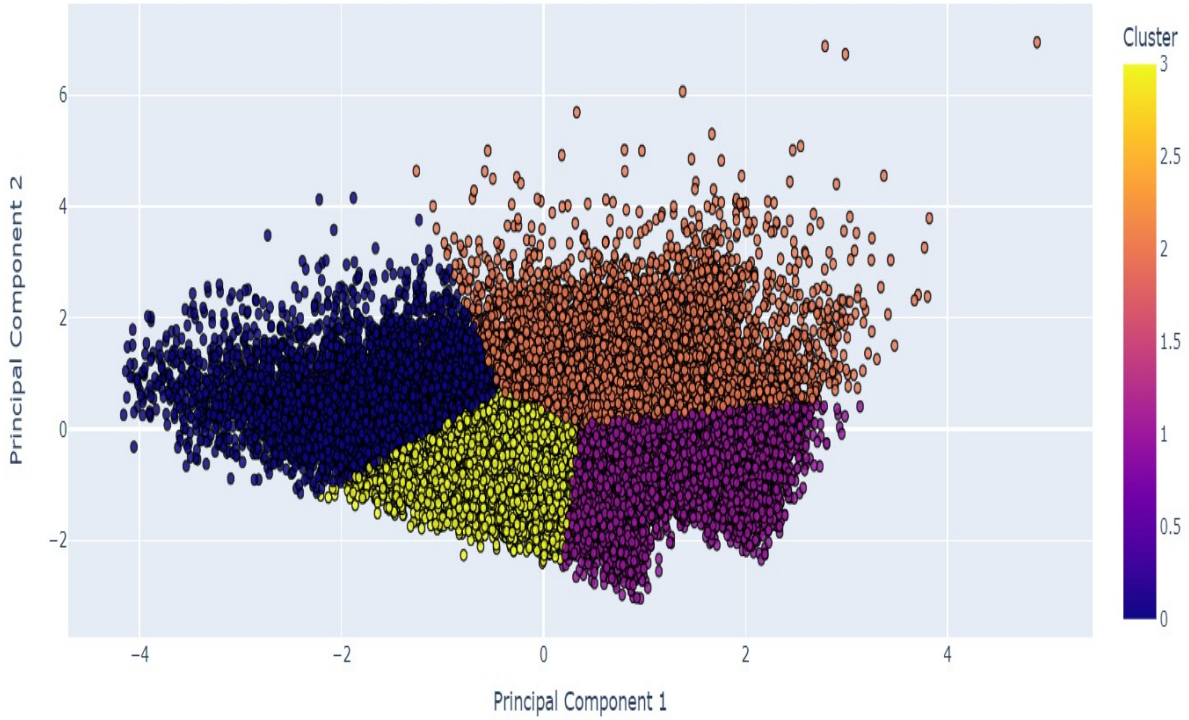
- Organic fruits form a major part of the most bought combination products
- High lift ~ High interdependence

instacart

K-means Clustering

K-means Clustering— PCA

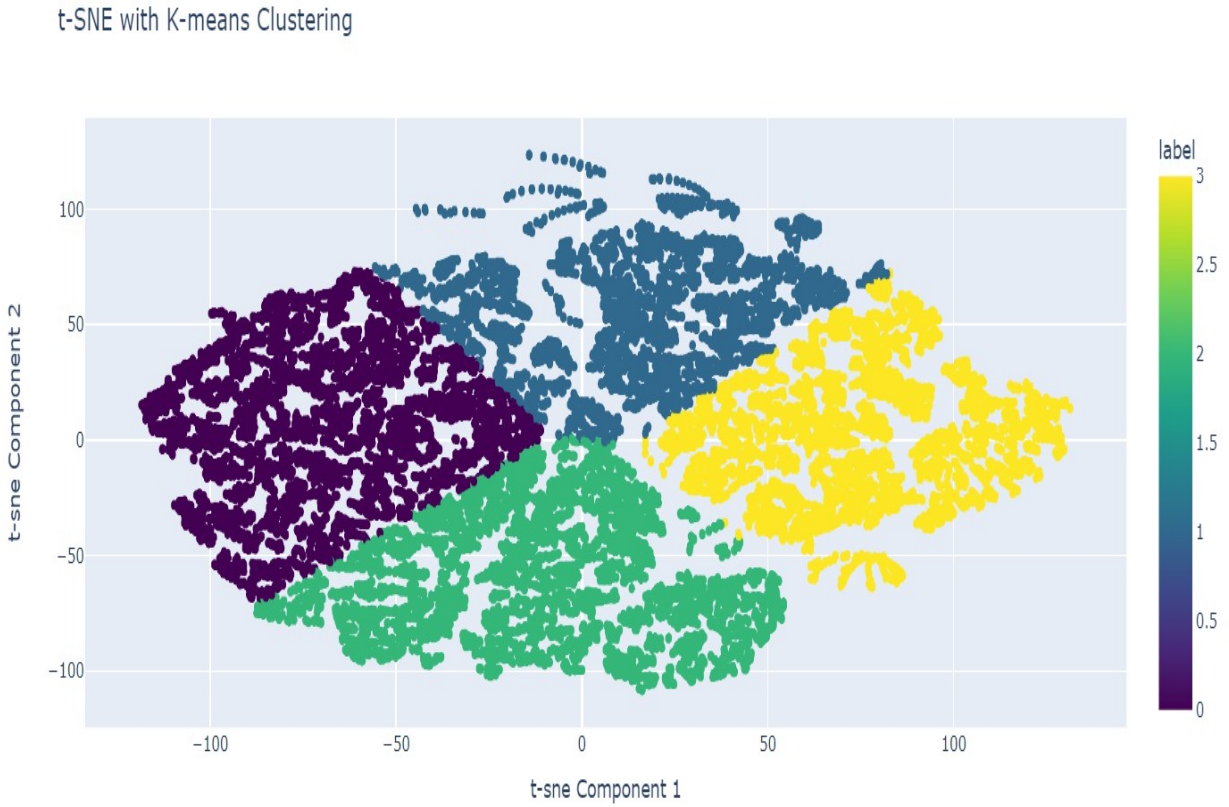
Before PCA



After PCA

K-means Clustering — T-SNE

Before T-SNE

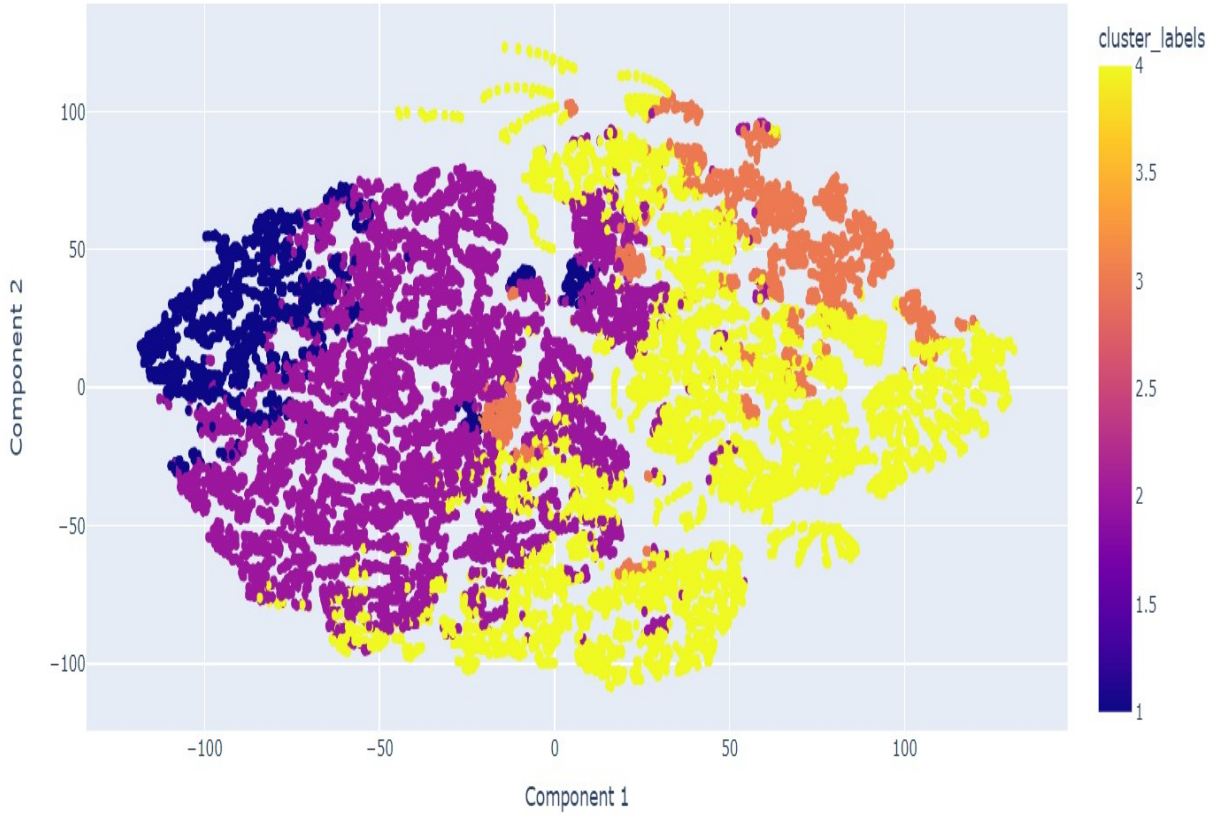


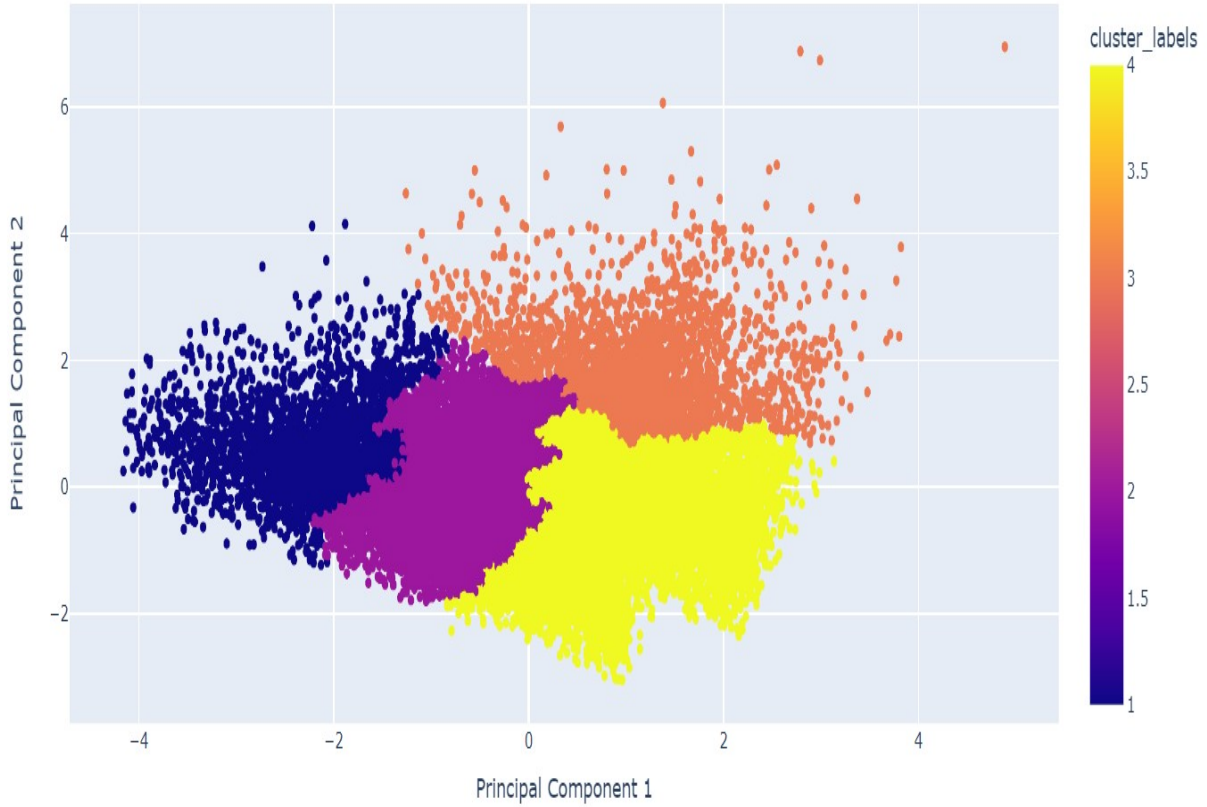
After T-SNE

instacart

Hierarchical Clustering

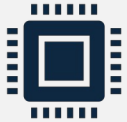
Hierarchical Clustering





Challenges

Challenges



Challenges with Large Data

1.Extensive sampling and resampling required.

2.Frequent computational crashes due to memory limits.

Market Basket Analysis Approaches

1.Initial insights from sampled data, limited by partial data coverage.

2. Later, a comprehensive method using product combinations for broader analysis, facing computational challenges.

Dimensionality Reduction Techniques

1.t-SNE and Hierarchical clustering applied, necessitating data resampling for computational feasibility.

2. Successful visualization and interpretation of high-dimensional data.

Business Relevance(Future work/next steps)

Business Relevance(Future work/next steps)

Optimization Approach and Business Strategy	Emphasize that precise clustering avoids over-segmentation
Personalized marketing campaigns	Increasing customer engagement and sales
Enhancing the user experience by suggesting relevant products	Leading to higher conversion rates and customer retention

Future work/next steps

Maintaining Model Relevance	Continuously update and fine-tune the model with new data, and possibly integrate adaptive learning mechanisms to evolve with changing patterns.	**Adapting to New Data**	Continuously update and fine-tune the model with new data, and possibly integrate adaptive learning mechanisms to evolve with changing patterns.
Cost Efficiency	Continuously update and fine-tune the model with new data, and possibly integrate adaptive learning mechanisms to evolve with changing patterns.	**Data Privacy and Ethics**	Continuously update and fine-tune the model with new data, and possibly integrate adaptive learning mechanisms to evolve with changing patterns.

Thanks for Your listening

Questions!