Module 2 Final Project

Course:ALY6110-Data Management and Big Data

Academic Term-Fall 2025

Instructor: Professor Mary Donhoffner

Group 2

Princess Chishato

Muhammad Waqas Azam

Snovy Sunil Dbritto

Junhui Hu

Ruman Liu

Shyam Sundar Nandakumar

Contents

Introduction 3

Data Description 4

Data Preprocessing 4

Dimensionality Reduction 5

Clustering Analysis 6

Implementation and Reproducibility 8

Result 9

Business view 10

Conclusion 11

Reference 12

Introduction

Our approach to this project was from a business owner's point of view. While any supermarket or online retailer is certainly interested in short-term profit, the key to success is long-term stability. To that end, having an in-depth understanding of one's customers and their purchasing habits is vital. If a retailer knows what items different types of customers are likely to buy, how often they are likely to buy them, and what is the overall structure of their shopping baskets, they can make informed business decisions to stock the right items, order the right quantities, restock at the right time, and bundle or recommend items in such a way as to increase sales. If they lack this type of data-driven understanding, then each step in the business process, from procurement to stocking to sales, may lead to avoidable losses and missed opportunities.

Instacart is a very suitable place to conduct this type of analysis, as their transactional data is both massive and very informative. Each order is a record of a customer's shopping trip and can provide insights into when, how often, and what they are buying. The only challenge to do so is that with millions of orders and thousands of unique products, clear patterns are not immediately obvious. The goal of this study was to translate raw information into business insights that can be used for more informed decision-making. Specifically, we aimed to segment Instacart customers into different behavior groups using unsupervised learning methods. Unlike predictive analytics, the focus was not on predicting but rather on discovering latent patterns that

characterize the behavior of shoppers. The results of such an analysis could be leveraged by business managers to create more targeted and effective marketing campaigns, to make inventory planning easier and more accurate, and to increase customer retention.

Data Description

The data set utilized for the project was derived from Kaggle's public Instacart Market Basket Analysis dataset. This dataset includes more than three million grocery orders from over two hundred thousand Instacart users. The database includes five tables that describe elements of the online ordering experience, including shoppers, items, aisles, and orders. The variables include ccategoricals and numeric like order sequence, weekday, and time of day orders were placed.

```
(33819106, 15)
Index(['order_id', 'product_id', 'add_to_cart_order', 'reordered',
       'product_name', 'aisle_id', 'department_id', 'aisle', 'department',
       'user_id', 'eval_set', 'order_number', 'order_dow', 'order_hour_of_day',
       'days_since_prior_order'],
      dtype='object')
```

[2]:

	order_id	product_id	add_to_cart_order	reordered	product_name	aisle_id	department_id	aisle	department	user_id	eval_set	order_number	order_dow
0	2	33120	1	1	Organic Egg Whites	86	16	eggs	dairy eggs	202279	prior	3	5
1	2	28985	2	1	Michigan Organic Kale	83	4	fresh vegetables	produce	202279	prior	3	5
2	2	9327	3	0	Garlic Powder	104	13	spices seasonings	pantry	202279	prior	3	5

The image above shows sample view of the merged Instacart dataset (33,819,106 rows × 15 columns) showing the integrated order, product, and categorical features used for analysis.

Data Preprocessing

The first step was merging all three tables into a single analysis data frame using the user ID and order ID values as keys, followed by dropping any duplicates or missing values. The image below shows missing values analysis showing that

days_since_prior_order had 6.14% missing values (2,078,868 records)

	Missing Values	Percent
days_since_prior_order	2078068	6.14

Since PCA cannot directly process NaN values, columns with a higher proportion of missing data may introduce false patterns. Therefore, we imputed missing categorical values with '-1'.

The order_id associated with products whose order number were "Unknown" or -1 were dropped from the time interval calculations since those orders do not have a valid timestamp for measuring the mean time between orders. All of the numerical columns were scaled to zero mean and unit variance in order to prevent features on larger scales from dominating features on smaller scales, in preparation for PCA and clustering.

The main addition to this version of the project, which is the feature engineering component. The initial exploratory analysis was simply looking at the raw number of purchases for each aisle, but it became apparent that it is more useful to also characterize the customers by their overall purchasing behavior, rather than just their interests. The features chosen, which include number of orders, mean time between orders, mean basket size, maximum basket size, reorder ratio and mean hour of the day of orders were chosen to represent how users were using the website and to what extent, and are far more interpretable as a result.

Dimensionality Reduction

We performed Principal Component Analysis (PCA) to reduce the feature space, since the combined set had over 130 features (Scikit-learn Developers, 2024). PCA is a technique which can reduce the number of features while keeping the maximum

amount of information. It achieves this by building uncorrelated features that represent the correlations within the original variables.

The explained variance plot shows that the top 5 components have an explained variance of more than 95 percent. This validates the fact that our customer data can be modelled by only 5 different dimensions. The first component represented the overall engagement and ordering activity of the customers, the second the basket size and ordering activity, the third the ordering time preferences, the fourth the regularity in purchase intervals and the fifth the stability and consistency.

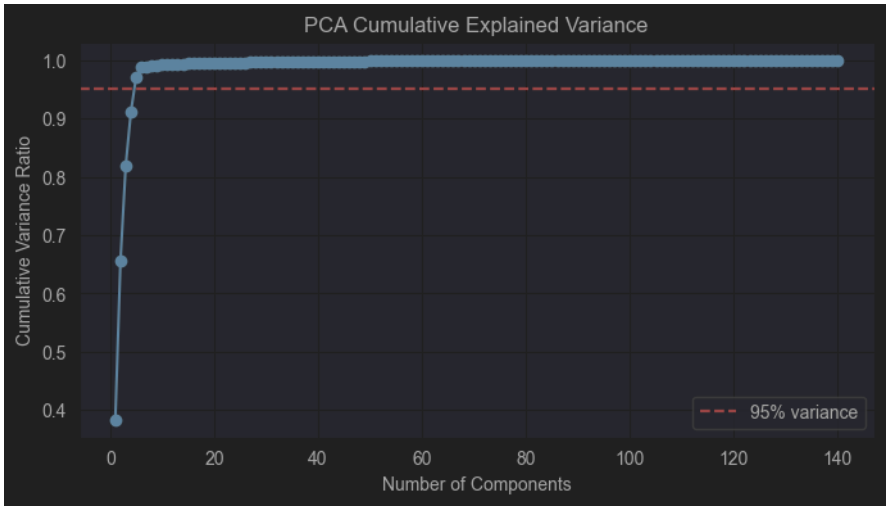


Figure 1: The cumulative explained variance with a red threshold line to 95 percent

Clustering Analysis

The preprocessed data was used for K-Means clustering. K-Means is an unsupervised machine learning algorithm that partitions the users into clusters of similar users. We decided to use K-Means since it can handle large amount of data efficiently and the cluster center is interpretable (Scikit-learn Developers, 2024; Towards Data Science,

2023). Since the K-Means model does not depend on a labelled dataset, this means that the clusters it finds are found naturally in the data.

We decided on three clusters through the Elbow method and Silhouette analysis. The Elbow plot shows an abrupt curve at 3 clusters (Figure 2(a)) and the Silhouette score was maximum when there are 3 clusters (Figure 2(b)). Since, three clusters seems to provide a good tradeoff between the cluster compactness and cluster separation we use 3 clusters.

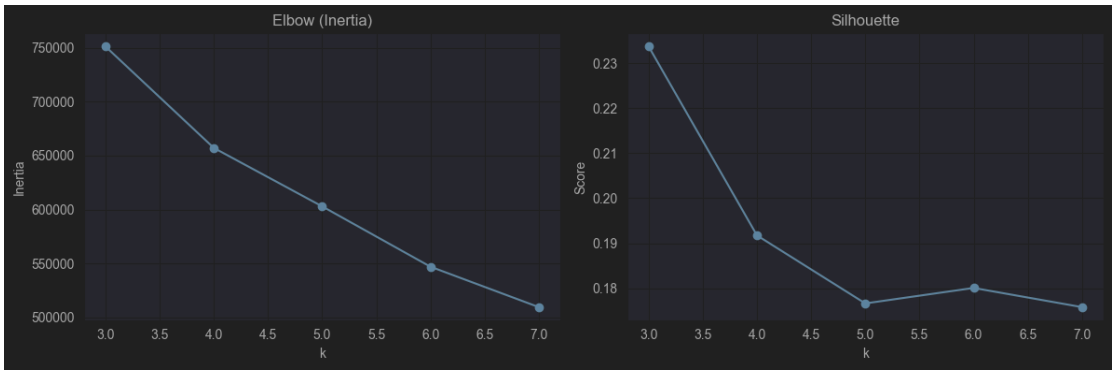
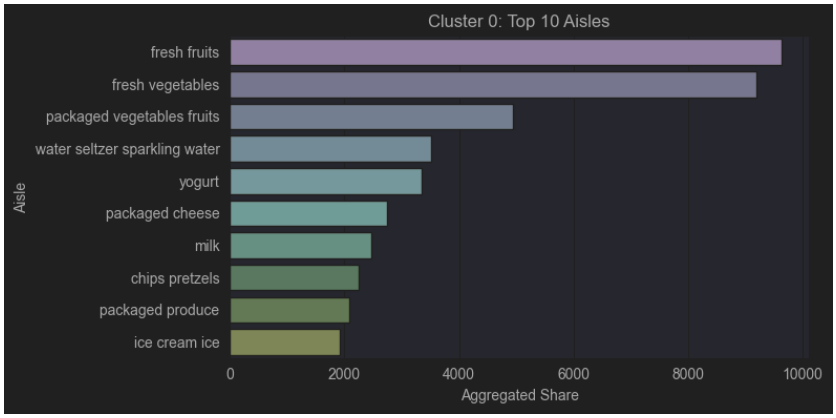
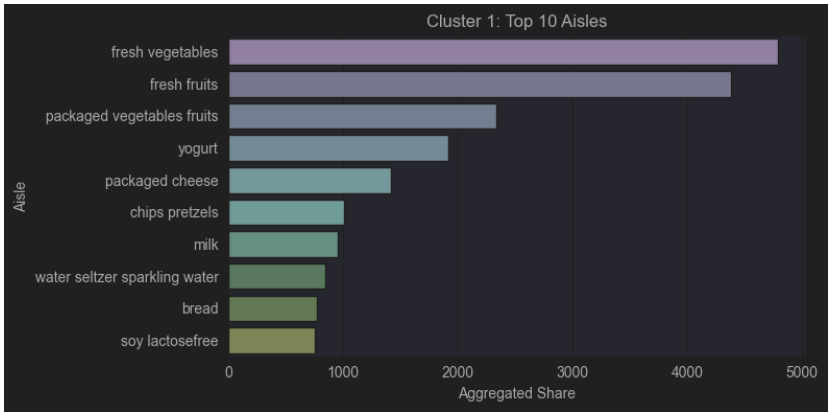
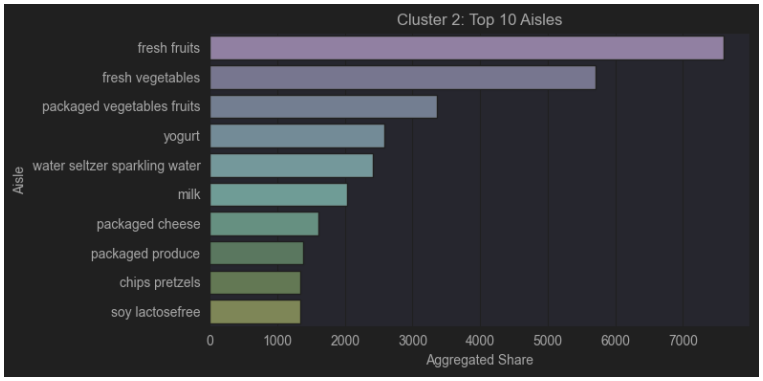


Figure 2. Elbow and Silhouette plots showing that three clusters yield optimal performance.

Each user was assigned to one of the three clusters based on their behavioral profile. We then examined both behavioral and product-level differences to interpret the meaning of each group. Mean feature comparisons helped us identify differences in order frequency, basket size, and reorder rates, while top-aisle and top-product visualizations revealed distinct shopping preferences.







Implementation and Reproducibility

The analysis was conducted in Python using pandas for data manipulation, scikit-learn for modeling, and matplotlib and seaborn for visualization (Matplotlib Development Team, 2024). The workflow was organized into modular steps of data cleaning, feature engineering, PCA transformation, and clustering. Random seeds were set for reproducibility, and each step was documented to allow for replication of the analysis

with the same settings.

Result

PCA showed that customer behavior can be modeled using five principal components that summarize engagement, frequency, recency and stability. The three clusters from K-Means showed three well-differentiated customer groups.

Cluster 0 had low-frequency shoppers with small baskets and low reorder rate. They are likely infrequent or new users that place occasional orders. Cluster 1 shoppers were of medium frequency; they placed large baskets and reordered at a moderate rate. These users had more balanced behavior that points to planned shopping such as family needs or bulk buying. Cluster 2 had highly engaged users that reordered often with short recency periods. Although they shopped with smaller baskets, they had consistent behavior with higher rates of reordering and loyalty.

Behavioral feature means by cluster:

3 rows × 6 cols

cluster	total_orders	avg_days_between_orders	avg_basket_size	max_basket_size	reorder_rate	avg_order_hour
0	7.42	19.51	7.68	12.98	0.31	13.81
1	15.83	15.22	18.42	32.06	0.50	13.55
2	30.71	9.66	7.94	16.39	0.62	13.13

The most frequent aisles and products were similar across all three groups, with fruits, vegetables, yogurt, and milk being the most popular. This is in line with the general Instacart shopper that has a tendency to buy fresh and healthy foods. The clusters were well-differentiated by behavior metrics such as frequency and reorder rate. For instance, Cluster 2 had more than twice the reorder rate as Cluster 0, which would show habit formation and long-term engagement.

The analysis revealed that customer segmentation should consider not only what people buy but how they buy. The combination of behavioral and product-level analysis helped to identify actionable differences to help personalize marketing and operations.

Business view

On a business level, we can take actionable insights from the analysis. Cluster 0 users can be targeted to increase frequency, Cluster 1 shoppers may be enticed with family packs, loyalty points, or subscription delivery models for bulk purchases, Cluster 2 users are the most valuable in terms of engagement and retention and can be incentivized with early access offers, referral bonuses, or automated reorder alerts.

Segmentation also enables data-driven decision making across business units. Inventory managers can anticipate demand for different product categories based on customer group size, marketing teams can craft messages that align with lifestyles and purchase cadence, operationally, understanding behavioral clusters enables more efficient allocation of resources and minimizes waste in restocking or promotional efforts.

This study has some limitations as well. The analysis was done on a subset of behavioral and product features, price and promotion influences were not factored in. Customers may also be driven by income, regional availability, seasonal factors in their choice behavior. Incorporating these aspects can refine segmentation models. Other clustering techniques such as hierarchical clustering or Gaussian Mixture

Models can uncover additional structure in the data.

Conclusion

In this project, we applied unsupervised learning to solve the business problem of identifying customer segments in Instacart shopping data. By doing feature engineering, PCA, and K-means clustering, we were able to detect three separate segments of Instacart shoppers. Each group has their own distinct behaviors and personality in shopping. We may use these insights to make marketing and operations more efficient.

In business terms, these results have clear implications: it's vital to know not only what people buy, but also how they buy. Segmenting customers by data can allow for personalization, efficient inventory control, and a more durable long-term relationship with the consumer.

Reference

Scikit-learn Developers. (2024). Principal component analysis (PCA) — scikit-learn documentation. Scikit-learn. https://scikit-learn.org/stable/modules/decomposition.html#pca

GeeksforGeeks. (2025, July 11). Determine the optimal value of K in K-Means clustering – ML. GeeksforGeeks. https://www.geeksforgeeks.org/ml-determine-the-optimal-value-of-k-in-k-means-clustering/

Scikit-learn Developers. (2024). *K-means clustering — scikit-learn documentation.* Scikit-learn. https://scikit-learn.org/stable/modules/clustering.html#k-means

Scikit-learn Developers. (2024). *Silhouette coefficient — scikit-learn documentation.* Scikit-learn. https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient

Matplotlib Development Team. (2024). *Matplotlib: Visualization with Python.* https://matplotlib.org/stable/index.html

Waskom, M. L. (2024). *Seaborn: Statistical data visualization.* Seaborn. https://seaborn.pydata.org/