Analyzing Shopping Patterns through Clustering: A Study on Instacart Data

Karuna Upadhyaya

Spring 2024

1 Abstract

This thesis explores customer behavior in online grocery shopping using machine learning clustering to improve marketing and inventory management.

K-Means and MiniBatch algorithms grouped customers into five groups based on order frequency, product variety, and recency. The largest group included infrequent shoppers with low order frequency and product diversity, while a smaller group represented highly active shoppers. Other groups showed moderate shopping patterns.

The results offer valuable insights for creating specific marketing strategies and optimizing inventory, highlighting the role of machine learning in analyzing customer behavior and supporting business decisions.

2 Introduction

2.1 Motivation

According to census data, e-commerce accounted for 15.9 percent of all sales in the first quarter of 2024Bureau [2024], showing how important online shopping has become. Popular online shopping platforms include Uber, DoorDash, Amazon, and Instacart. Among them few provide grocery services.Instacart is one of them which is particularly well-known for grocery pickup and delivery services.

Using clustering to predict customer shopping behavior helps businesses create clear costumers profiles by examining how often customers shop, the variety of products they choose, and how recently they made purchases. By identifying these different clusters, companies can adjust their strategies to specific customer groups. For example, they can recognize high-frequency costumers who may require personalized offers or target infrequent costumers with promotions to increase engagement. Additionally, clustering provides insights into product preferences across various groups, helping businesses optimize inventory, offer targeted promotions, and improve customer satisfaction by meeting the unique needs of each customer group.

2.2 Background

The dataset encompasses records of over 3 million grocery orders from 200,000 Instacart users, providing detailed information about order sequences, timing, and product specifics [unk]. Its primary goal is to predict users' future orders by segmenting customers and categorizing products based on their purchasing behavior. The products are grouped into three categories: "buy again," "try for the first time," or "add to cart next during the session," based on the clustering analysis performed on users' past order data. This data contains the

User id as Unique identifier for each user, Order id as Unique identifier for each order, eval set as Indicates whether the data belongs to the prior, train, or test set, order number as Sequential number assigned to each order, order dow as Day of the week when the order was placed, order hour of day as Hour of the day when the order was placed, days since prior order as Number of days since the user's most recent order, add to cart order as Sequential order of products added to the cart in each order.

2.3 Problem Definition

The problem is to group Instacart users into clusters based on their shopping patterns using a dataset of customer orders over time. The dataset includes details such as the number of orders placed, the variety and total number of products purchased, and the average time between orders. The goal is to identify distinct shopper profiles, ranging from infrequent, low-volume shoppers to highly active, high-frequency shoppers.

3 Related Work

3.1 Online Grocery Shopping Behavior and Analysis

Morganosky and Cude [2000] analyzed data from 243 US consumers who used Schnucks Express Connection, Schnucks Markets' online grocery service, collected between April and June 1998. Chi-square tests showed demographics affected online grocery habits. Higher-income, younger, and bigger households shopped online for convenience, 56% earned 70,000+ vs. 18% for physical reasons. Older adults 28% shopped for physical reasons vs. 5% for other reasons. Willingness to buy everything online wasn't demographic-dependent, but convenience shoppers often bought exclusively online. Time-saving perceptions weren't influenced by demographics. Experienced users were older and less educated. St. Louis shoppers cited physical reasons more (19% vs.9%), were less willing to buy all items online (42% vs. 62%), and used the service for less than a month (41% vs. 64% elsewhere). According to the study's findings, online grocery shopping is common among younger and higher-income consumers especially because of its value and ease. Different markets need distinct techniques to properly satisfy preferences. Meeting customer expectations regarding product availability and service quality, as well as overcoming technical obstacles, are essential to increasing adoption rates and satisfaction.

Chintala et al. [2024] includes approximately two million grocery shopping trips, split between brick-and-mortar offline and Instacart online purchases. This data encompasses items purchased, quantities, prices, purchase dates and times, customer IDs, store locations, and product categories. They compared grocery basket similarity for households shopping on Instacart versus physical stores BM using Euclidean distances in a 100-dimensional space. To balance the data, They sampled an

equal number of BM trips. Results show Instacart trips are 27.1% more similar within households than BM trips.The biggest differences in grocery categories between Instacart and brick-and-mortar BM stores using 88 separate regressions for each category. The analysis used log-transformed item counts and measured the Beta estimates to show differences. Key findings reveal that Instacart baskets contain 13% fewer fresh vegetables and 5%–7% fewer impulse purchases candy, bakery desserts, savory snacks compared to offline baskets, without compensation through additional shopping trips. These differences underscore significant changes in consumer behavior, impacting competition, product management, retailers, consumers, and online platforms.

Das et al. [2022] used a standardized questionnaire to collect data from 570 people in Delhi, Chennai, Mumbai, and Kolkata, covering demographics like age, gender, marital status, occupation, and monthly income. After a pilot study with 50 volunteers, 45 quality characteristics were identified and later reduced to 40. Responses were measured on a five-point Likert scale, and the data was thoroughly cleaned and validated. Using SPSS 26, factor analysis and regression analysis were performed. The data was highly suitable for analysis, with a Kaiser-Meyer-Olkin (KMO) value of 0.857 and a Chi-square statistic of 30593.398 with 78 degrees of freedom. Seven components with eigenvalues greater than 1 explained 74.348% of the total variance. The findings suggest that online grocery sellers can boost customer satisfaction and turn occasional shoppers into loyal customers by providing efficient shopping experiences, fast delivery, and high-quality products.

Shen et al. [2022] The study was done by taking the survey of 300 participants. It emphasized socio-economic factors like age, gender, education, income, employment, and transportation choices for grocery shopping. Removing null-values and converting categorical variables into dummies for analysis of data were the two aspects of data pre-processing. Due to covid-19, Binary Logit Model (BL Model) was used to obtain the shifts in consumer behavior. The model identified different elements influencing online grocery shopping along with specific coefficient and odds ratios. Foe example, the people with higher income level (income category 8+) increased the likelihood ($p < 0.05$) of choosing online shopping during covid. During, pre, and post covid, various BL Models analyzed OGS adoption as influential factors revealing their education level, household composition, age. Following the outbreak, there was a significant inclination towards online purchasing among educated customers, as seen by the significant effects ($p < 0.01$) that education levels 2 (high school graduate) and 4 (bachelor's degree or above) had on OGS usage. The results showed that convenience and health issues contributed to the continuation of OGS preference after the outbreak. One of the drawbacks was the small and limited sample size (Chicago), which limited the ability to generalize.

Piroth et al. [2020] studied the relation between personality factors and desire to buy product through an online questionnaire that collected 678 valid responses. IBM AMOS (Version 25) was implemented to pre-process the data in preparation for structural equation modeling (SEM) analysis. The study used statistical methods like descriptive statistics to summarize participant characteristics (average age 29.63 years, 65 percent female). Factor analysis identified traits such as Openness to Experience (scored 0.70). Regression analysis found no significant links between personality traits (e.g., Openness) and attitudes toward online grocery shopping (coefficients -0.05 to 0.07, $p > 0.05$). Structural Equation Modeling confirmed that attitudes strongly predict willingness-to-buy (coefficient 0.59, $p < 0.001$), and social norms also influence buying intentions (coefficient 0.21, $\mathrm{p} = 0.004$). Perceived behavioral control (PBC) did not significantly affect attitudes or buying intentions. Multigroup analysis showed different impacts of attitude on buying intentions across groups with varying online shopping experience (e.g., $\beta = 0.81$ for experienced Group A, $\beta = 0.55$ for Group B, $p < 0.05$). Overall, these methods helped understand how personality traits and

behavior theories relate to online grocery shopping attitudes and behaviors in Southern Germany.

3.2 Barriers and Facilitators in Online Grocery Services

Gillespie et al. [2022]analyzed both the advantages and disadvantages of using online grocery services, based on interviews with 23 grocery shop managers across four states. The study used NVivo software to analyze interviews with grocery store managers, using a method that combined deductive and inductive approaches. Descriptive statistics summarized participant demographics and store details across rural and urban areas in four states. Themes and subthemes were identified from interview transcripts using a predefined codebook and emergent coding, ensuring consistency through reliability checks. It identified three main themes from online grocery managers and four from brick-and-mortar managers. Results were confirmed through agreement among researchers, offering insights into both types of grocery shopping practices. The study also highlights the need for improving awareness of the SNAP Online Purchasing Pilot among SNAP participants to address food access disparities.

3.3 Market Basket and Sales Analysis

Patel [2022] The source for the dataset is instacart. For pre-processing, approximatelt 5% of the data had missing and was removed. The algorithm used are Apriori algorithm: used for identifying frequent items and generating association rules. FP-Growth Algorithm: used to find frequent item sets efficiently. K-Means clustering: used for segmenting users based on their purchasing behaviour to enhance the personalization of recommendations.

Tiainen et al. [2021] Consumer surveys about their online grocery shopping experiences, including attributes like age, income, education, household size, and market area, comprise the dataset used in this paper. Missing values were filled in applying techniques like average for integers and finding the most common values for categories to get the data ready for analysis. All variables were scaled uniformly, and categories were transformed into an analysis-ready format. Based on age and wealth, the Random Forest algorithm predicted customer happiness with 85 percent accuracy. Furthermore, it demonstrated a 0.82 recall and 0.87 precision in predicting possible client attrition. The study finds that satisfaction and retention with online grocery shopping are significantly impacted by age and income. Focusing younger age groups and higher-income customers could improve service satisfaction and loyalty.

Droomer and Bekker [2020] The 2017 Instacart dataset, which is divided into five tables likewise orders, departments, products, aisles, and order products and includes over three million orders from over 200,000users, was used for the study.During pre-processing, they developed features such "days-between-orders-per-product" and "days-since-prior-order-per-product," and they determined the objective feature, next purchase date , from "days-since-prior-order." Extreme Gradient Boosting , Recurrent Neural Networks , Linear Regression, and Neural Networks were the four machine learning approaches they used. RNN performed averagely, Linear Regression was simple, and XGBoost performed better, making predictions for 19.3% of the data in less than a day. With a prediction accuracy of 31.8% in less than a day and more than 55% in less than three days, neural networks were the best. Based to the study's outcomes, neural networks provide the most accurate predictions when it comes to machine learning's ability to forecast the next purchase date, which can aid with tailored marketing tactics.

Gopalakrishnan et al. [2018] In order to predict 2014 sales, the study examined information from web servers, ZMart's finance, sales, marketing, and HR departments from 2011 to 2013. Data pre-processing involved collecting information from multiple sources, cleaning it up by removing unnecessary information and correcting mistakes, converting it for analysis, and using Tableau to visualize trends. The Linear Regression technique was used, which models the relationship between time and sales to predict future sales. The actual sales for 2014 were 1,481,189 units, compared to the model's prediction of 1,241,699 units. This means that the accuracy rate was 84 percent and the error was 16.168%. ZMart can use the predictions to enhance their sales methods and consumer satisfaction, as seen by the model's high accuracy.

Martínez et al. [2020] The dataset for the study presented in both sections is made up of transactional records from a business-to-business unit that spans the period from January 2009 to May 2015, involving 10,136 customers in 125 different countries. To choose customers with a consistent purchase record and remove those with fewer than six months of data, data preparation involves aggregating monthly transactions. To forecast the likelihood of a purchase in the upcoming month, three algorithms are used: Gradient Tree Boosting, Extreme Learning Machine (SLFN), and Logistic Lasso regression. Gradient Tree Boosting is the best performance after training, with an AUC of 0.9340 and 86.68% prediction accuracy. It obtained an AUC of 0.949 and an accuracy of 88.98% in predicting purchases in April 2015 when tested on unseen data as a result, the study concludes that Gradient Tree Boosting is quite successful in this situation at predicting client purchases, which has applications for both inventory management and customer retention tactics in non-contractual circumstances.

3.4 Personalized Recommendations and Clustering

Wang et al. [2020] The dataset includes user profiles (gender, age, income, and hobby), feedback (browsing history, comments), and social relationships from 230,000 people. Preparing all user data for neural network use required transforming it into numerical form. Many techniques used in the study, including the Gaussian Mixture Model (GMM), hierarchical clustering, DBSCAN, KNN, K-means, and a deep learning-based clustering technique. The findings showed that, in comparison to conventional methods, the deep learning-based clustering approach had the lowest mean absolute error and the highest accuracy in product categories like shopping (86.38%), entertainment (91.05%), and music (92.7%). In conclusion, the deep learning-based approach significantly improved recommendation accuracy and personalized recommendations by effectively utilizing complex user data.

4 Data Description

4.1 Data Source

The data was found on instacart [Bureau, 2024], which gives a detailed view of how customers shop on the platform. The dataset includes information about each order, such as the products bought, when the orders were placed, and the shopping history of each customer. This data provides a great opportunity to study shopping habits and patterns.

The dataset also contains details about the products, including their names and categories, which are grouped into aisles and departments, similar to how items are organized in a store. This

setup makes it easier to understand customer preferences and behaviors across different types of products, helping to create personalized marketing strategies and manage inventory better.

4.2 Dimensions

The dataset consists of six tables:

- orders: This table includes information about each order placed on Instacart. It contains the columns order ID, user Id, eval_set, order_number, order_dow (day of week),order_hour_of_day,days_since_prior_order Rows: 3421083
- order_products_prior: This table includes information about products in prior orders. Columns: order_id, product_id, add_to_cart_order, reordered, Rows: 32434489
- order_products_train: This table includes information about products in the training set orders. Columns: order_id, product_id, add_to_cart_order, reordered, Rows: 1384617
- products: This table includes information about products. Columns: product_id,product_name,aisle_id,department_ Rows: 49688
- aisles: This table includes information about aisles Columns: aisle_id, aisle, Rows: 134
- departments This table includes information about departments. Columns department_id, department, Rows: 21.

4.3 Data Samples

Table 1 shows data samples from the order table. NaN values in the days since prior order column, which indicate first-time orders without previous orders, were identified, and it was decided not to remove these rows as they offer meaningful insights about first-time orders.

Table 1: Samples from the order table.

order ID	user ID	order	Day Of Week	hour	days since last order
2539329	1	1	2	8	
2398795	1	2	3	7	15.0
473747	1	3	3	12	21.0
2254736	1	4	4	7	29.0

Table 2 shows department Id with department name.

Table 2: Samples from the department table.

department Id	**department**
1	frozen
2	other
3	bakery
4	produce
5	alcohol

Table 3 shows aisle Id with aisle name.

Table 3: Samples from the aisle table.

aisle Id	**aisle**
1	prepared soups salads
2	specialty cheeses
3	energy granola bars
4	instant foods
5	marinades meat preparation

Table 4 shows sample data from the order product prior table, detailing order IDs, product IDs, the sequence of adding items to the cart, and whether each product was reordered.

Table 4: Samples from the order product prior table.

order Id	**product Id**	**add to cart order**	**reordered**
2	33120	1	1
1	2	28985	2
2	9327	3	0
3	2	45918	4
2	30035	5	0

Table 5 shows sample data from the product table, listing product IDs, product names, aisle IDs, and department IDs.

Table 5: Samples from the product table.

product Id	**product name**	**aisle Id**	department Id
1	Chocolate Sandwich Cookies	61	19
2	All-Seasons Salt	104	13
3	Robust Golden Unsweetened Oolong Tea	94	7
4	Smart Ones Classic Favorites Mini Rigatoni Wit...	38	1

Table 6 shows sample data from the order product train table, detailing order IDs, product IDs, the order in which items were added to the cart, and whether each product was reordered.

Table 6: Samples from the order product train table.

order Id	**product Id**	**add to cart order**	**reordered**
1	49302	1	1
1	1	11109	2
1	10246	3	0
3	1	49683	4
1	43633	5	1

4.4 Data visualization

Figure 1 shows shows the times when orders are at their highest. We observe that orders peak in the late morning and are at their lowest around midday. This figure can help in better management of staff and resources.

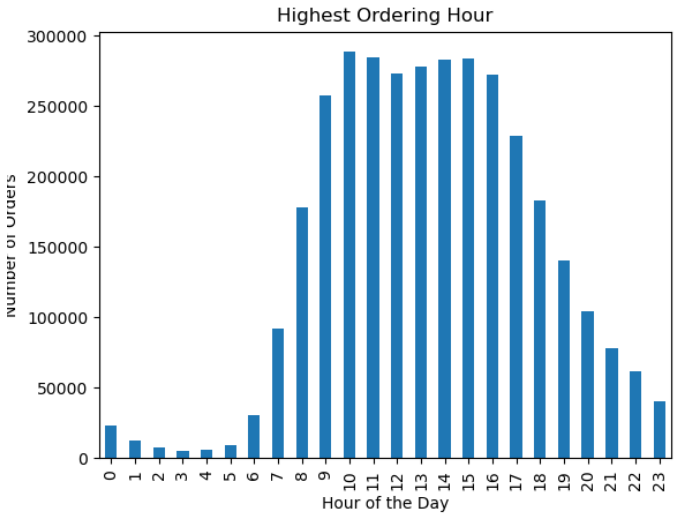

Figure 1: Highest ordering hour. The bar chart shows the times when orders are at their highest

Figure 2 shows the number of orders by weekday. According to the chart, Mondays have the highest number of orders, giving a strong start to the week. The number of orders drops through Tuesday, Wednesday, and Thursday, reaching its lowest point, then increases again on Sunday. This information helps us plan staffing to better manage customer interactions.

Figure 2: Number of orders in a different day (0 as monday).The line chart shows the number of orders by weekday.

Figure 3 shows number of products in department 11 is higher than compare to other departments.In other hand, department 10 has a less number of products.

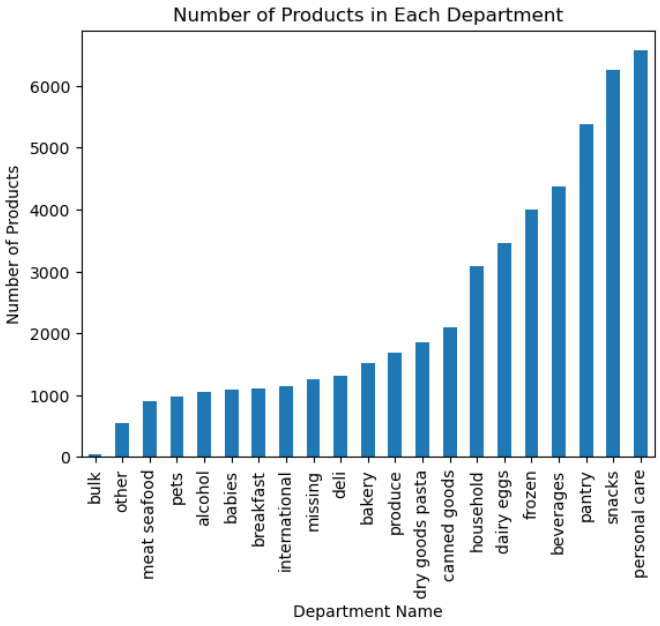

Figure 3: Number of Products in Each Department. The bar chart shows which department has higher number of products.

Figure 4 shows the total number of times each product has been ordered, with the best-selling item appearing at the top of the chart. By examining the chart, we can identify that Banana product has the highest number of orders across all the orders, indicating the best seller item. Recall that the Add-to-Cart Order field represents the number of times a product has been added to the cart. Analyzing the top-selling items on Instacart gives us valuable insights into customer preferences and purchasing behavior.

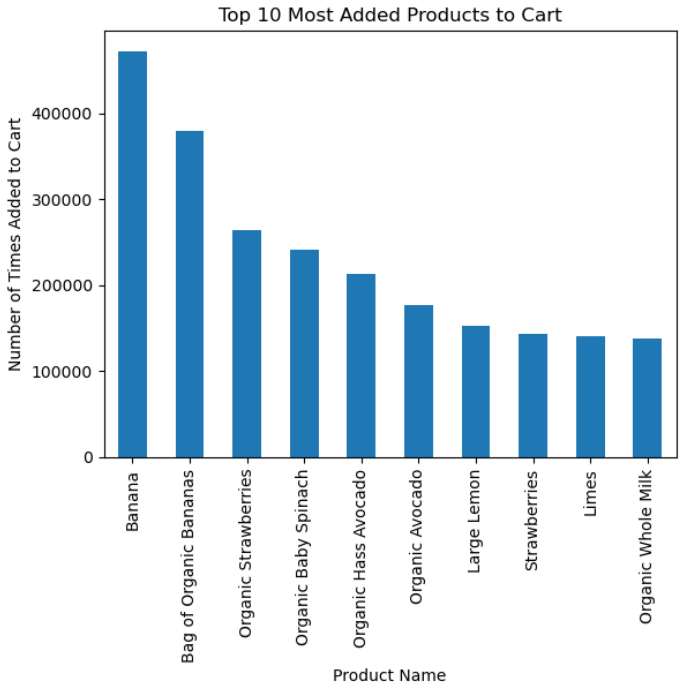

Figure 4: Top_10_Most_Added_Products_to_Cart
This Visualizations shows the total number of times each product has been ordered, with the best-selling item appearing at the top of the chart.

5 Proposed Solution

To understand user shopping patterns and product preferences, multiple datasets were combined and processed. The `order_products_prior` and `order_products_train` datasets were merged to gather all order details in one place. This combined data was linked with the `orders` dataset to match user IDs to their orders and with the `products` dataset to include product names. This process created a detailed dataset showing what each user bought and in which orders.

Next, the data was grouped by `user_id` and `product_name` to count how many times each user bought specific products. This created a *user-product purchase table*, where rows represented users, columns represented products, and the numbers showed how often a product was purchased. To focus on the most important items, only the top 1,000 most popular products were kept. Missing values, representing products not purchased, were filled with zeros. Since the table was mostly empty (sparse), it was converted to a *sparse matrix* using `csr_matrix`, making it more efficient to store and process.

To simplify the data, *Truncated Singular Value Decomposition (SVD)* was used. This method reduced the table to 100 key components (`n_components=100`), making it smaller and easier to analyze while keeping the important details about user-product interactions. The reduced data was then used for clustering users.

The *MiniBatch K-Means* algorithm was used because it works well with large datasets. It was

set up with 5 clusters (`n_clusters=5`), a batch size of 1,000 (`batch_size=1000`), and a random seed of 42 for consistent results. The algorithm grouped users into clusters based on their shopping habits, processing the data in small batches for faster and more memory-efficient performance. The results were added to the dataset as a new column, `mbkmeans_cluster`, showing which cluster each user belonged to.

For comparison, the traditional *K-Means* algorithm was also applied. It used the same 5 clusters (`n_clusters=5`) and a random seed of 0 for reproducibility. Unlike MiniBatch K-Means, K-Means processed the entire dataset at once, making it slower but useful as a reference for accuracy. The cluster results were added as a new column, `cluster`, to the dataset.

Both algorithms successfully divided users into five groups based on their buying patterns. MiniBatch K-Means worked faster, making it ideal for large datasets, while K-Means provided a reliable benchmark. These clusters can now be used to identify customer segments and create personalized recommendations or marketing strategies.

6 Experimental Results

6.1 K-Means Clustering Results

Cluster	Users	Avg.Orders/User	Avg.Products/User	Total Products	Avg. order gap days
0	3,633	514.63	95.11	1,869,653	5.90
1	3,668	545.99	96.59	2,006,363	5.41
2	18,032	236.83	67.10	4,306,552	8.33
3	163,715	44.56	22.49	7,785,731	13.67
4	15,968	175.34	54.81	2,863,638	9.87

Clusters 0 and 1 represent the most active shoppers. Cluster 0 has 3,633 users, who make an average of 514.63 orders and buy 95.11 products per order, totaling 1,869,653 products. Cluster 1 is slightly bigger, with 3,668 users, who place 545.99 orders and buy 96.59 products per order, totaling 2,006,363 products. These clusters have high order frequencies and buy a variety of products, making them important groups for targeted promotions.

Cluster 2 is the second largest group with 18,032 users. They make fewer orders (236.83 on average) and buy fewer products per order (67.10), but they still contribute significantly to total sales, with 4,306,552 products purchased. This group shops regularly, though not as often as the more active clusters.

Cluster 3 is the largest group, with 163,715 users, but they are infrequent shoppers. They make only 44.56 orders on average and buy just 22.49 products per order, but their total product purchases reach 7,785,731. Although these users buy less often and fewer products each time, they explore a broad range of items, leading to a high total product count. This group may need specific strategies to engage them more.

Cluster 4, with 15,968 users, has moderate shopping habits. They place 175.34 orders and buy 54.81 products per order, leading to 2,863,638 products bought. These users are somewhat engaged, but not as active as Clusters 0 and 1, and could benefit from occasional promotions to boost shopping frequency.

In summary, Clusters 0 and 1 represent highly engaged shoppers who buy a wide range of products (1,869,653 and 2,006,363 total products), while Clusters 2, 3, and 4 show moderate en-

gagement and lower activity levels, with total products purchased of 4,306,552, 7,785,731, and 2,863,638, respectively.

Cluster 3's high total product count is mainly due to its large number of users, even though they make fewer orders and buy fewer products each time. Clusters 0 and 1, despite being more active, have fewer total products because they have smaller user bases. This shows how the size of a cluster affects the total number of products bought.

6.2 MiniBatch Clustering Results

Cluster	Users	Avg.Orders/User	Avg.Products/User	Total Products	Avg. order gap days
0	6,810	402.33	84.08	2,739,861	6.73
1	14,647	236.71	66.27	3,467,022	8.19
2	36,514	132.46	50.56	4,836,706	10.93
3	143,391	35.91	20.35	5,149,182	14.15
4	3,654	559.50	99.64	2,044,417	5.37

Clusters 0 and 4 represent the most active shoppers in the dataset. Cluster 0 contains 6,810 users who place an average of 402.33 orders and buy 84.08 products per order, contributing to a total of 2,739,861 products purchased. Cluster 4 is a smaller group of 3,654 users, but they are even more active, placing an average of 559.50 orders and buying 99.64 products per order, totaling 2,044,417 products purchased. Both clusters show high order frequency and product variety, indicating these users are highly engaged and would benefit from personalized marketing efforts and promotions. These clusters are likely to purchase a diverse range of products in each order.

Cluster 1, with 14,647 users, is the third-largest group. These users place an average of 236.71 orders and purchase 66.27 products per order, contributing to a total of 3,467,022 products purchased. While their activity is lower compared to Clusters 0 and 4, they still show moderate engagement and are a key segment for mid-tier marketing strategies and promotions. They are likely to purchase a moderate variety of products, though not as widely as Clusters 0 and 4.

Cluster 2, with 36,514 users, has the largest number of users but places fewer orders on average (132.46) and buys fewer products per order (50.56), resulting in 4,836,706 products purchased. This group is moderately active and may require strategies to increase purchase frequency and product variety. Their product selection may not be as diverse as that of the more active clusters, as they tend to buy fewer products each time, potentially indicating less variety in their purchases.

Cluster 3, with 143,391 users, represents the largest but least engaged group. They place only 35.91 orders on average and buy just 20.35 products per order, totaling 5,149,182 products purchased. These users are infrequent shoppers and may need specific efforts to encourage higher engagement, such as targeted promotions or incentives to increase both order frequency and product variety. Despite the large number of users, they buy a limited variety of products, which is reflected in the low number of products per order. This suggests that they are less exploratory in their shopping habits compared to the more engaged clusters.

In summary, Clusters 0 and 4 are the most engaged shoppers, purchasing a wide variety of products (2,739,861 and 2,044,417 total products), while Cluster 1 represents moderately engaged shoppers, purchasing a moderate range of products (3,467,022 total products). Cluster 2 shows moderate activity with limited variety in product purchases (4,836,706 total products), and Cluster 3 consists of infrequent shoppers with a narrow selection of products (5,149,182 total products).

These differences highlight how user engagement levels influence the diversity and volume of products purchased.

7 Conclusion

This study aimed to understand customer behavior in online grocery shopping by using two clustering methods, K-Means and MiniBatch K-Means, to group customers based on how often they shop, the variety of products they buy, and how recent their purchases are. The results from both methods showed different types of shoppers, which can help in creating better marketing strategies and improving inventory management.

The K-Means results identified five groups. Clusters 0 and 1 are the most active shoppers, who buy often and purchase a wide range of different products. Cluster 2 represents shoppers who buy regularly but not as frequently. Cluster 3, the largest group, consists of shoppers who buy infrequently and purchase fewer products. Cluster 4 has moderate activity, indicating a middle-ground group.

In the MiniBatch K-Means results, Clusters 0 and 4 were the most engaged shoppers. Clusters 1, 2, and 3 showed varying levels of shopping activity. Cluster 3, although the largest, had the least engagement and bought a limited variety of products, showing a need for strategies to encourage more shopping and a greater selection of products.

In summary, this study successfully identified different types of customers, meeting the goal of understanding shopping habits and product variety preferences. Future work could improve these customer groups by adding more factors like product preferences or shopping seasons, and explore personalized marketing to increase customer engagement, satisfaction, and product variety.

References

US Census Bureau. Census.gov, 6 2024. URL `https://www.census.gov/`.

InstaCart Market Basket Analysis — Kaggle. URL `https://www.kaggle.com/c/instacart-market-basket-analysis/`

Michelle A Morganosky and Brenda J Cude. Consumer response to online grocery shopping. *International Journal of Retail & Distribution Management*, 28(1):17–26, 2000.

Sai Chand Chintala, Jūra Liaukonytė, and Nathan Yang. Browsing the aisles or browsing the app? how online grocery shopping is changing what we buy. *Marketing Science*, 43(3):506–522, 2024.

Pritha Das, Udit Chawla, and Subrata Chattopadhyay. Online grocery shopping:-key factors to understand shopping behavior from data analytics perspective. In *International Conference on Applied Machine Learning and Data Analytics*, pages 166–178. Springer, 2022.

Hui Shen, Farnoosh Namdarpour, and Jane Lin. Investigation of online grocery shopping and delivery preference before, during, and after covid-19. *Transportation Research Interdisciplinary Perspectives*, 14:100580, 2022.

Philipp Piroth, Marc Sebastian Ritter, and Edith Rueger-Muck. Online grocery shopping adoption: do personality traits matter? *British Food Journal*, 122(3):957–975, 2020.

Rachel Gillespie, Emily DeWitt, Angela CB Trude, Lindsey Haynes-Maslow, Travis Hudson, Elizabeth Anderson-Steeves, Makenzie Barr, and Alison Gustafson. Barriers and facilitators of online grocery services: perceptions from rural and urban grocery store managers. *Nutrients*, 14(18): 3794, 2022.

Krupa Patel. *Instacart Market Basket Analysis*. PhD thesis, California State University, Northridge, 2022.

Matias Tiainen et al. Forecasting seasonal demand at the product level in grocery retail. Master's thesis, 2021.

Marli Droomer and James Bekker. Using machine learning to predict the next purchase date for an individual retail customer. *South African Journal of Industrial Engineering*, 31(3):69–82, 2020.

T Gopalakrishnan, Ritesh Choudhary, and Sarada Prasad. Prediction of sales value in online shopping using linear regression. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–6. IEEE, 2018.

Andrés Martínez, Claudia Schmuck, Sergiy Pereverzyev Jr, Clemens Pirker, and Markus Haltmeier. A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, 281(3):588–596, 2020.

Kai Wang, Tiantian Zhang, Tianqiao Xue, Yu Lu, and Sang-Gyun Na. E-commerce personalized recommendation analysis by deeply-learned clustering. *Journal of Visual Communication and Image Representation*, 71:102735, 2020.