# 2023 Sales Performance & Customer Insights Report

This report presents a comprehensive analysis of the 2023 retail store purchase dataset, which includes customer demographics, product categories, transaction details, and quantities purchased. The goal of this analysis is to extract meaningful insights about sales performance, customer behavior, product trends, and time-based purchasing patterns. By exploring these dimensions, we aim to identify potential opportunities for optimization, targeting, and strategic decision-making.

| | Date | Gender | Age | Product Category | Quantity | Price per Unit | Total Amount |
|---|---|---|---|---|---|---|---|
| 0 | 11/24/2023 | Male | 34 | Beauty | 3 | 50 | 150 |
| 1 | 2/27/2023 | Female | 26 | Clothing | 2 | 500 | 1000 |
| 2 | 1/13/2023 | Male | 50 | Electronics | 1 | 30 | 30 |

Example: First three rows with headers

# Sales Performance

How much total revenue did the business generate each month in 2023?


Figure 1 shows the total monthly revenue across 2023. The business experienced the highest revenue in May, making it the best performing month of the year. This peak may be influenced

by seasonal demand or external promotional events, although no such context was included in the dataset.

Which product categories contributed most to total sales?

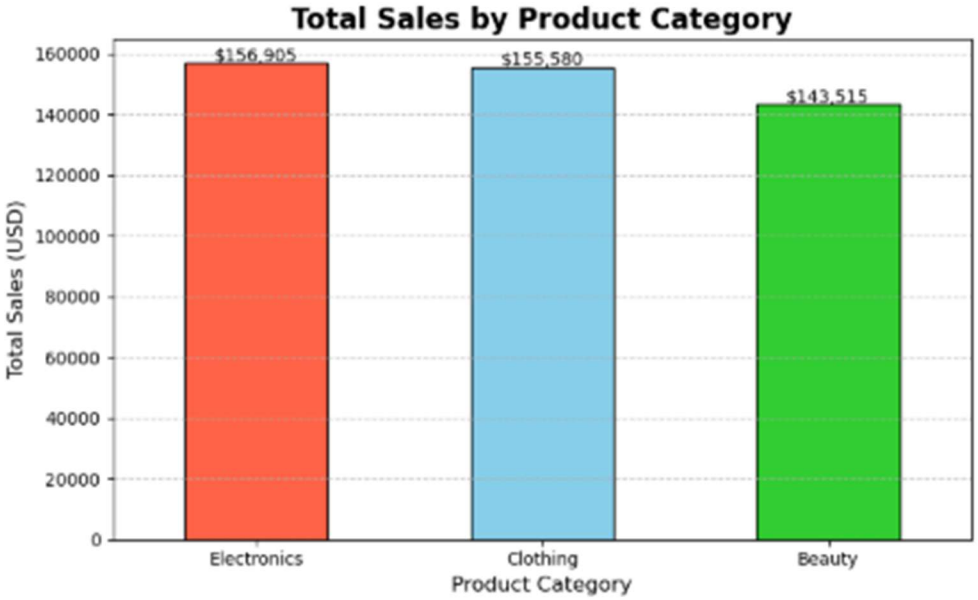



Figure 2 illustrates total revenue by product category. While Clothing, Electronics, and Beauty appear to perform similarly, an ANOVA test returned a p-value of 0.8527, indicating no statistically significant difference in sales across categories at the 95% confidence level. This suggests that no single category consistently dominates revenue, and prioritizing one over another would likely not yield a measurable advantage.

What is monthly revenue by product  category?

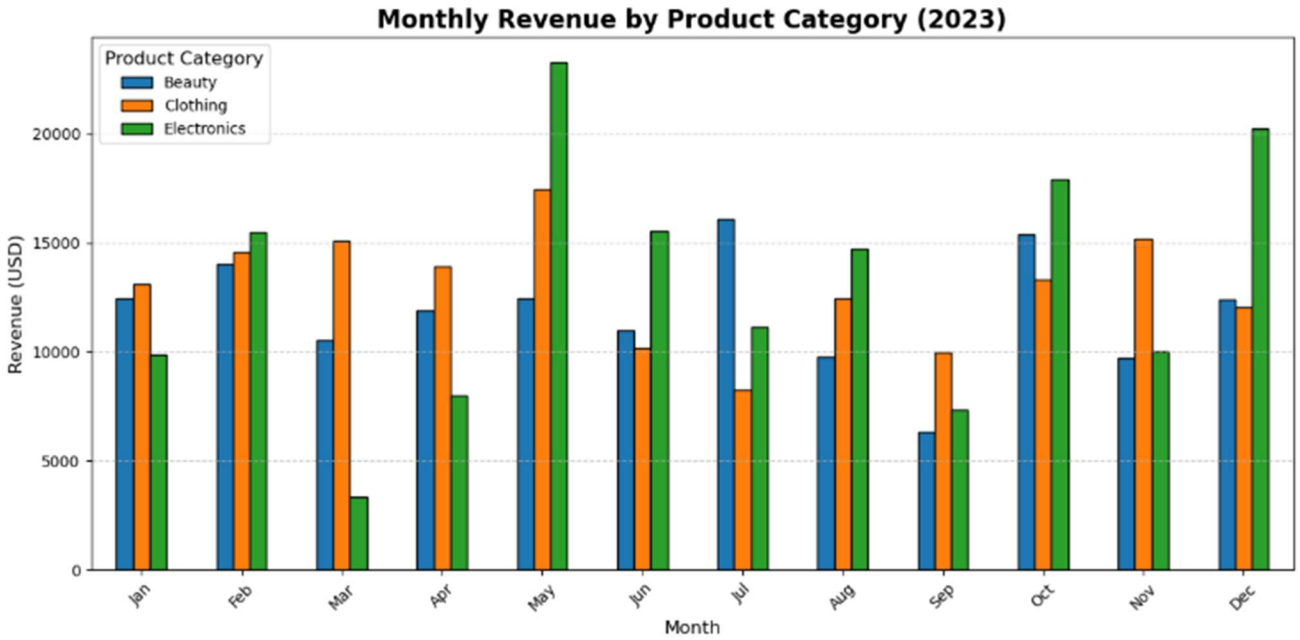



Figure 3 displays monthly sales trends by product category. While the dataset does not include marketing or promotional campaign data, the observed variations may reflect customer responses to external stimuli such as advertisements or seasonal promotions. For example, certain months show category-specific spikes, suggesting that timing and external factors likely influenced customer behavior — though this cannot be confirmed without additional context.

What is the average order value per customer or transaction?

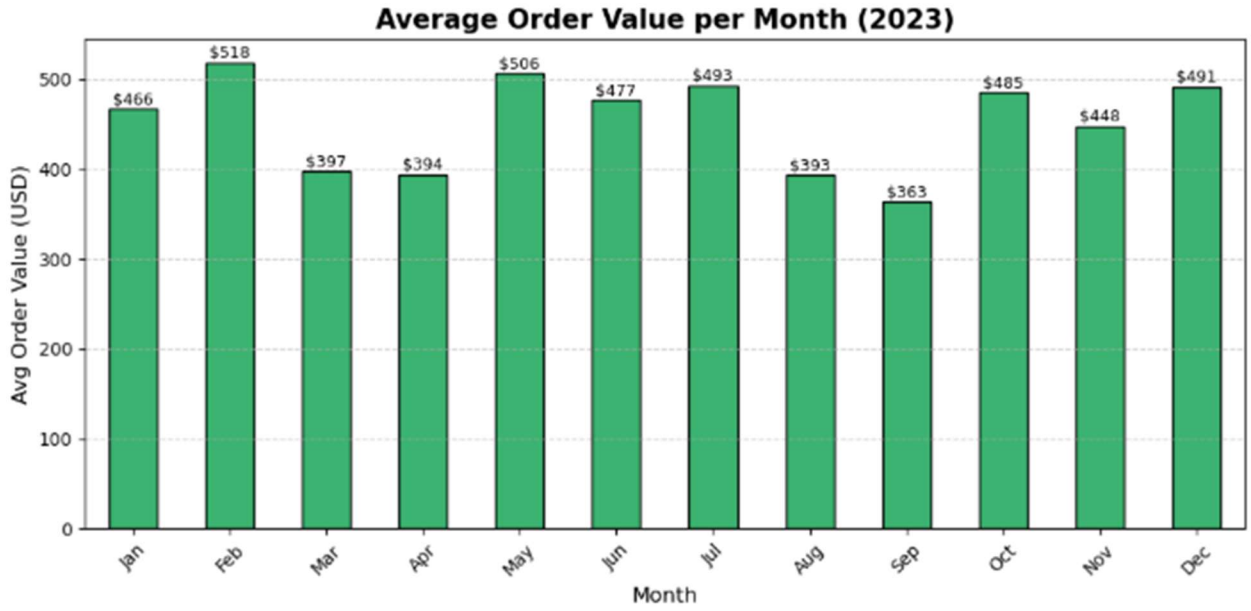


Figure 4 shows the Average Order Value (AOV) for each month in 2023. There is no observable upward or seasonal trend, and the AOV remains relatively stable throughout the year. Since AOV is a key performance indicator for retail growth, the lack of change suggests that customer spending per transaction did not increase over time. Further analysis by product category is needed to explore if any subgroup exhibits a different pattern.

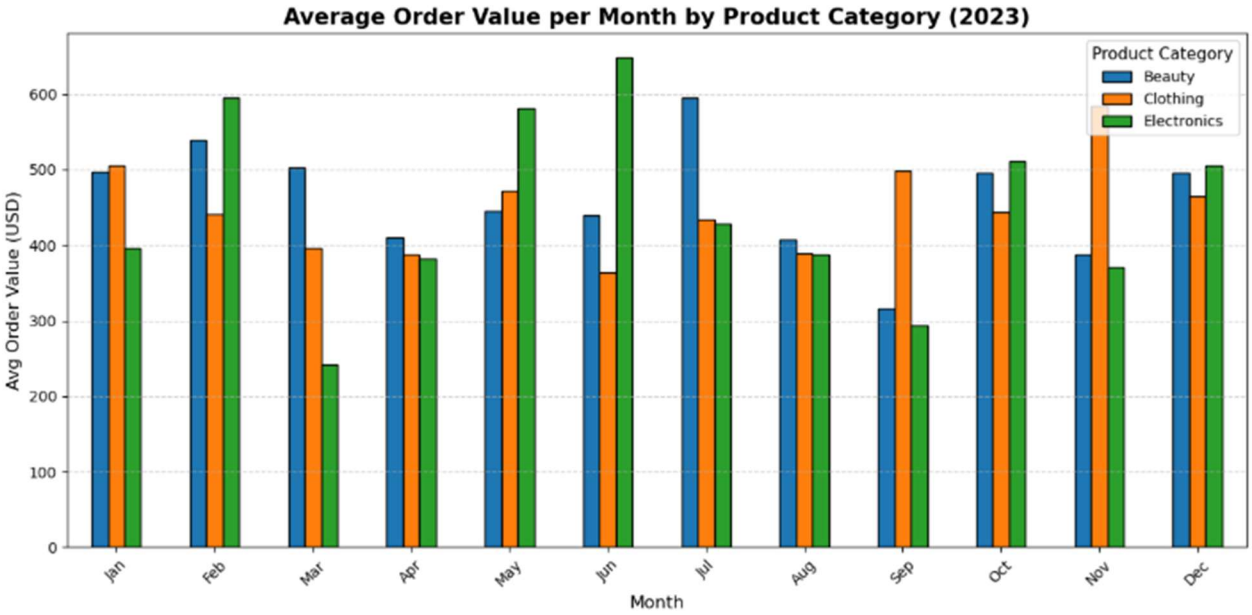



Figure 5 shows the AOV trend for each product category throughout 2023. Similar to the overall AOV, there is no clear pattern or seasonal fluctuation across categories. This suggests that factors like holidays or seasonal promotions (e.g., Christmas) did not significantly influence customer spending behavior. As a result, it may be more valuable to explore other variables — such as customer segmentation, discount strategies, or marketing campaigns — to identify potential growth levers.

# Customer Demographics:

Understanding who the customers are is just as important as knowing what they purchase. Demographic insights such as gender and age can help businesses tailor marketing strategies, adjust product offerings, and identify underserved segments. In this section, we explore how sales differ between male and female customers, how age influences buying behavior, and whether certain demographic groups show clear preferences for specific product categories. Identifying these tendencies is key to building more targeted and effective sales strategies

How do sales differ between male and female customers?

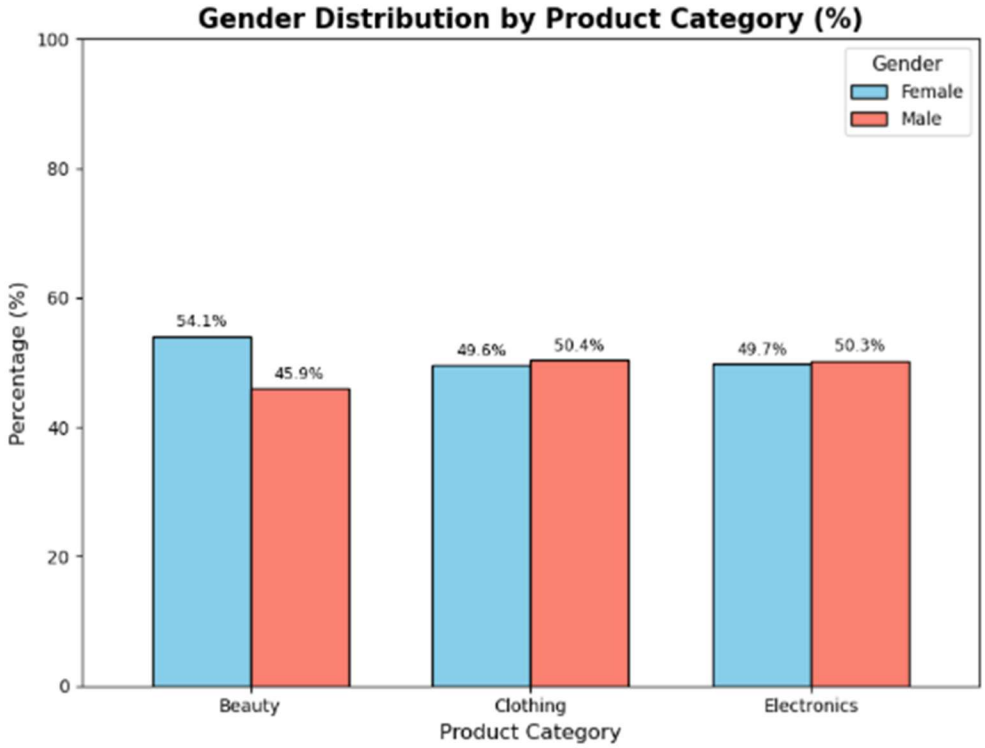



Fig6: shows gender preference toward specific items, where we see that females and males tend to have almost same preference in clothing and electronics, but in beauty category females tend to be dominant, but upon doing a chi-test for all three groups we end up with P-value of 0.433 which suggests that the difference in preference between genders is not statistically significant,

also doing a separate test for beauty category only, we get P-value of 0.2207, which suggests that the difference is NOT significant with 95% confidence level.
All of these findings mean that gender has no preference.

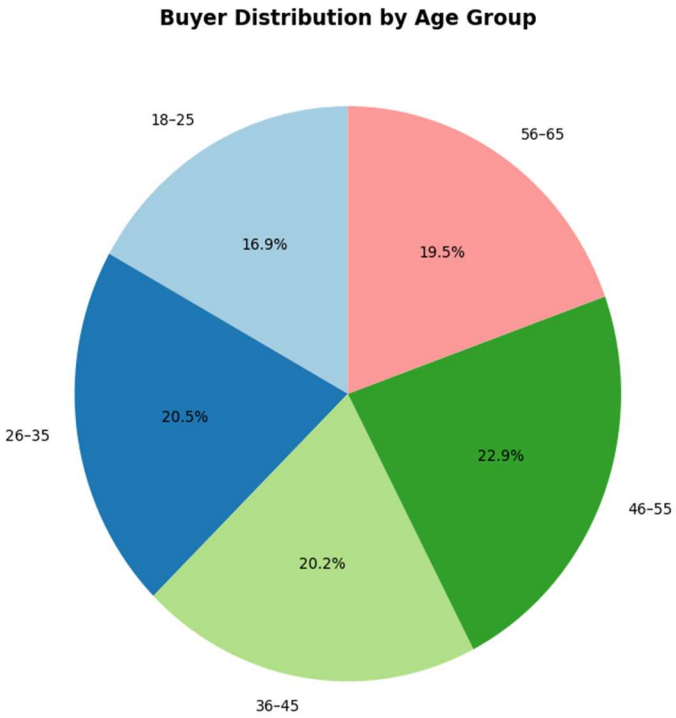



Figure 7 displays the age distribution of all customers in the dataset. The distribution appears balanced and evenly spread across age groups, with no single age segment dominating the purchasing activity. This suggests that the store attracts a broad age range of customers, and that age alone may not be a strong differentiator in buyer behavior

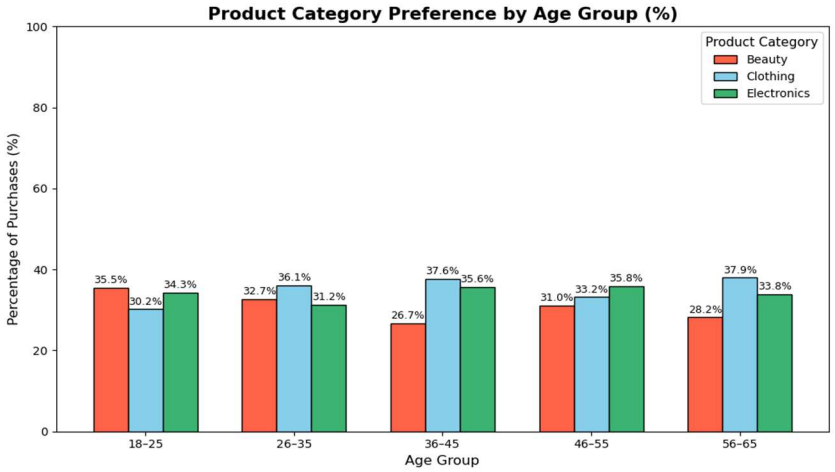



Figure 8 explores whether product category preferences vary across age groups. Visually, there are no strong patterns indicating age-based preferences. To confirm this, a statistical test was performed, which showed no significant relationship between age and product category selection. This reinforces the conclusion that age is not a major driver of product preference in this dataset.

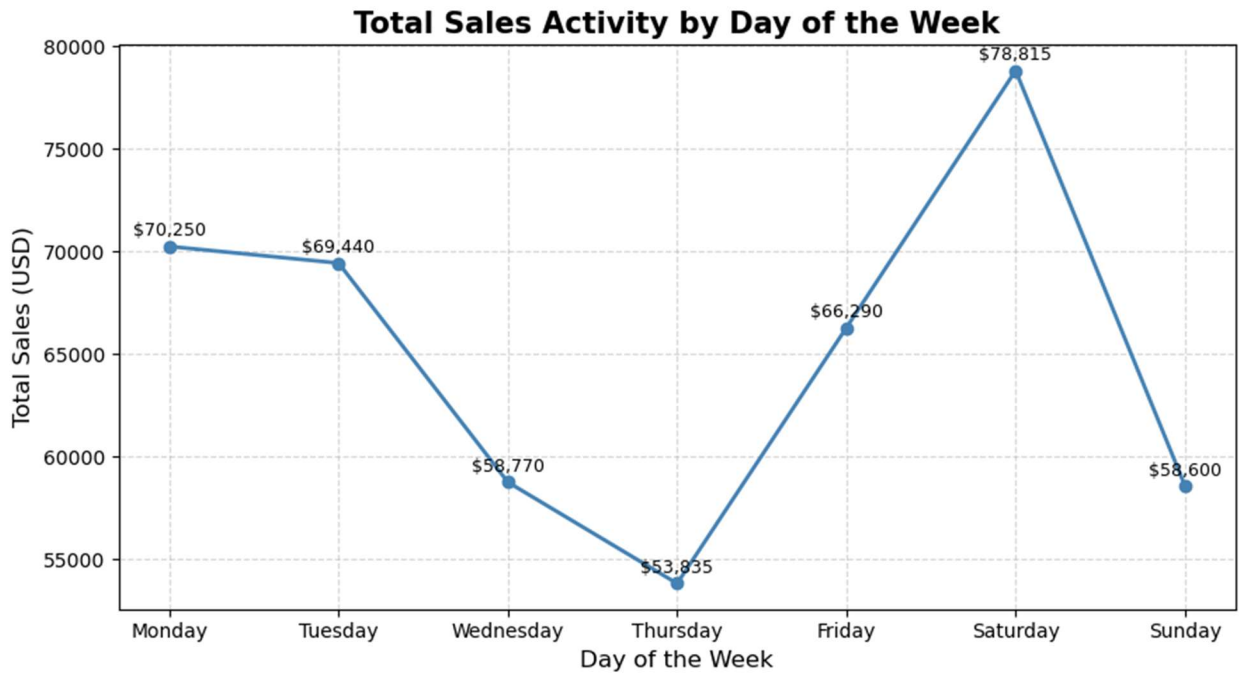
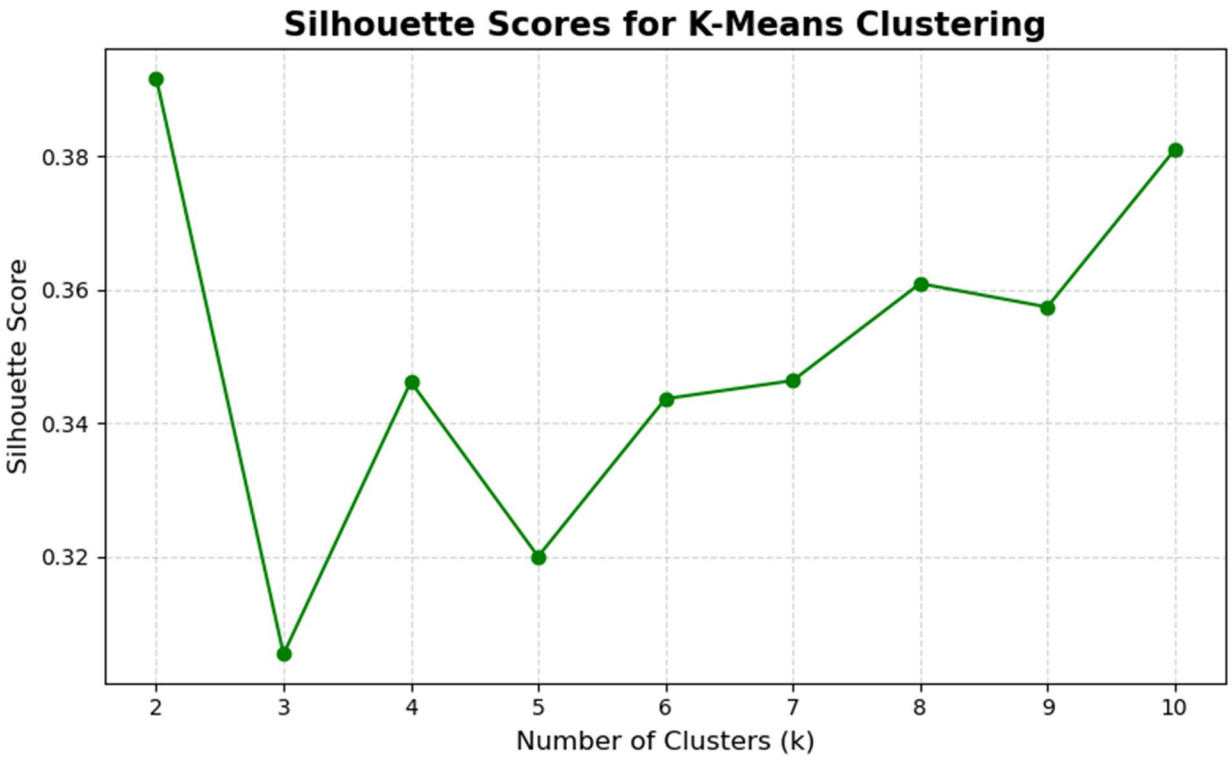


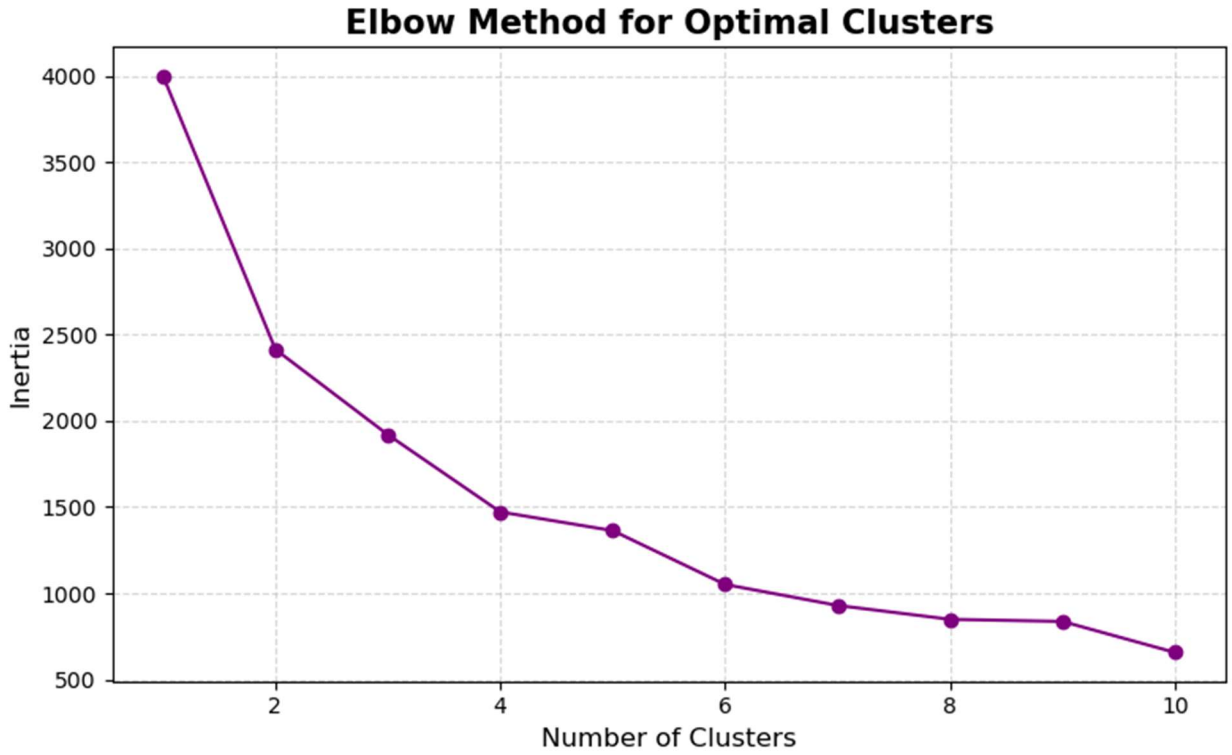

Figure 8 illustrates total sales by day of the week. While visual inspection may suggest that some days perform better than others, an ANOVA test returned a p-value of 0.6787, indicating that

these differences are not statistically significant. Therefore, day of the week does not appear to meaningfully influence sales performance based on the data available.

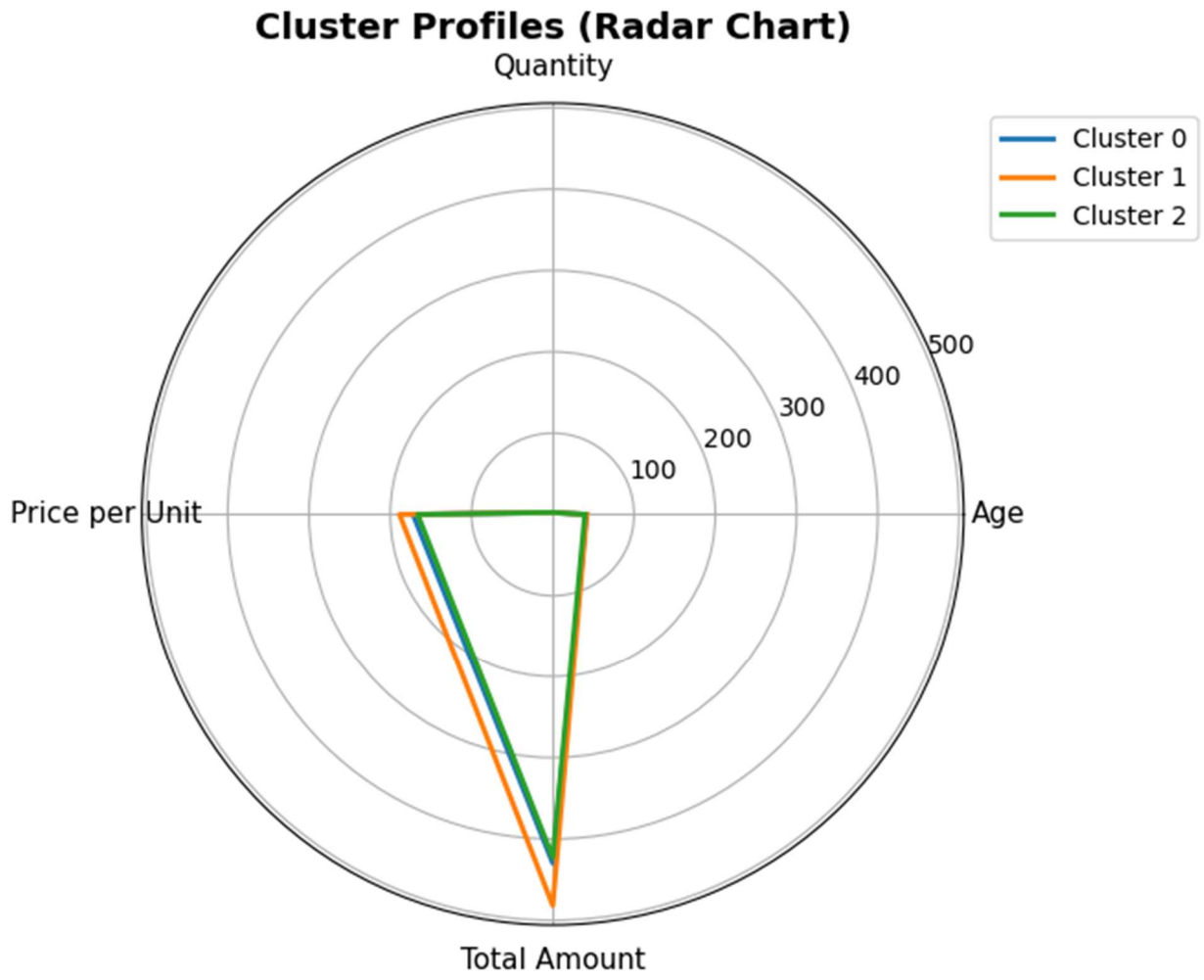
# Exploring Customer Segmentation with Clustering

After addressing key business questions, it became clear that the dataset lacks a strong, singular driver of sales — no individual factor like age, gender, day of the week, or product category significantly influenced purchasing behavior. However, rather than stop there, we advanced into more complex analysis to investigate whether hidden patterns or customer personas could still be discovered. We began with K-means clustering to explore the possibility of identifying natural customer segments based on their purchase behavior. As a first step, we used the elbow method and silhouette scores to determine the most suitable number of clusters. These tools help assess how well the data can be grouped and whether clear, separable clusters exist.


Silhouette Scores for K-Means Clustering
Silhouette Score
0.38
0.36
0.34
0.32
2
3
4
5
6
7
8
9
10
Number of Clusters (k)

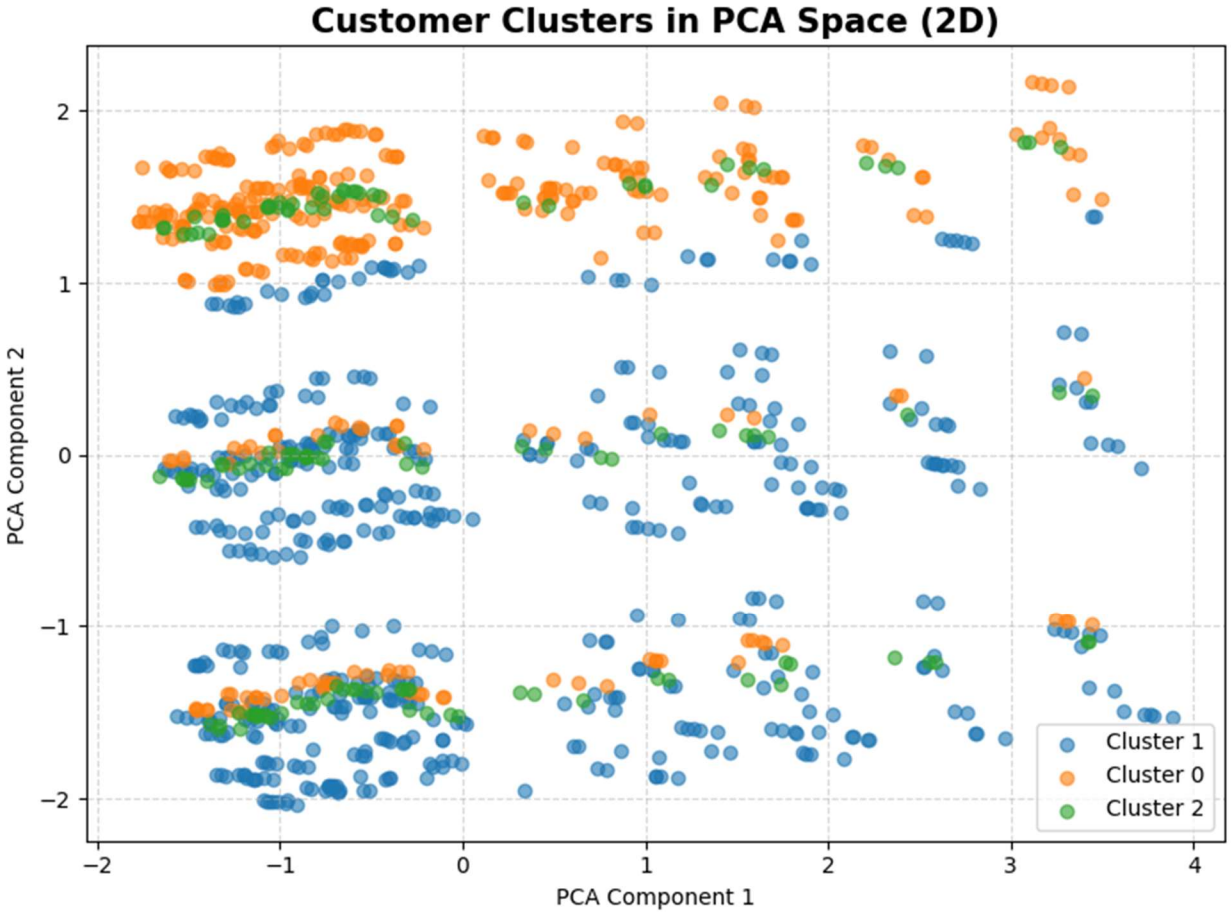

Elbow Method for Optimal Clusters
Inertia
4000
3500
3000
2500
2000
1500
1000
500
2
4
6
8
10
Number of Clusters

Both the elbow method and silhouette score suggested that the data does not naturally form well-defined groups. While the metrics pointed toward 2 clusters as the most technically appropriate choice, we still proceeded with 3 clusters to explore whether a slightly more granular segmentation might reveal meaningful customer patterns. This allowed us to test for any subtle differences across potential customer personas, despite the overall low separability in the data.




The radar chart visualization shows that the clusters heavily overlap across most features, indicating that there are no clear, distinct customer personas within the data. This suggests that

no specific combination of factors consistently defines one group versus another, and the clustering fails to reveal any actionable segmentation.




After reducing the dataset to two principal components using PCA (Principal Component Analysis), we visualized the 3-cluster solution in a 2D space. The resulting scatter plot showed significant overlap between clusters, with no clear separation between groups. This further confirms that the data does not naturally support meaningful segmentation, and that clustering does not provide actionable insights with the current feature set

## Predictive Modeling: Regression and Behavior-Based Approaches.

We analyzed a dataset containing 999 sales transactions, aiming to identify patterns and build predictive models for total purchase amount. After preprocessing and encoding categorical variables, we applied multiple regression techniques:

| Model | **Train $R^2$** | **Test $R^2$** | **RMSE** | MAE |
|---|---|---|---|---|
| Linear Regression | 0.85 | 0.85 | ~214 | |
| Ridge Regression | 0.85 | 0.85 | ~206 | |
| Random Forest Regressor | 1.00 | 1.00 | ~0 | 0.00 |

All models showed that over 99% of the variance in total sales could be explained by just two features: Quantity and Price per Unit. This confirms that the total purchase amount is almost entirely determined by a simple arithmetic relationship (Total ≈ Quantity × Price), leaving no room for deeper behavioral insights. Other variables such as Age, Gender, Product Category, and Weekday had negligible predictive impact.

# Predicting Quantity with Interaction Effects

To move beyond formula-driven patterns, we shifted the prediction target from Total to Quantity purchased, aiming to model underlying behavioral influences. We introduced interaction terms using a second-degree polynomial model, allowing the model to learn how variables might influence each other (e.g., how price affects purchases differently by category or age).

# Model Performance (Target: Quantity)

| Model | $R^2$ (Test) | MAE |
|---|---|---|
| **Ridge Regression (with Interactions)** | **0.39** | **0.84** |

Although this method introduced more nuance, the model still failed to explain most of the variation in quantity. These results further confirmed that the current dataset lacks the behavioral complexity and feature diversity required to support effective predictive modeling beyond basic arithmetic relationship

# Conclusion

This analysis explored the 2023 retail purchase dataset from multiple angles—sales performance, customer demographics, product trends, time patterns, clustering, and predictive modeling. While the dataset was well-structured, it lacked the complexity and behavioral richness needed to produce deep, actionable insights. Statistical tests revealed no significant differences across gender, age, or weekday behavior, and clustering failed to identify distinct customer segments. Predictive modeling showed that total sales were almost entirely determined by quantity and price, and even advanced methods like interaction terms and classification models produced limited improvement. To unlock more meaningful patterns in the future, we recommend expanding the dataset to include behavioral, contextual, and transactional features—such as discounts, repeat visits, marketing exposure, time of purchase, or geographic segmentation. With richer data, more targeted strategies and customer-driven decisions will become possible