Customer Shopping Behavior Analysis Report

1. Project Overview

This project focuses on analyzing customer shopping behavior using a dataset comprising 3,900 purchases across various product categories. The primary goal is to derive actionable insights into customer spending patterns, segmentation, product preferences, and subscription trends. These findings are intended to guide strategic business decisions and optimize marketing efforts.

2. Dataset Summary

The analysis was performed on a transactional dataset summarized as follows:

Metric	Detail
Rows (Purchases)	3,328 *
Columns	18
Key Features	Customer Demographics (Age, Gender, Location, Subscription Status), Purchase details (Category, Purchase Amount, Color, Size, Item Purchased), Shopping behavior (Discount Used, Previous Purchases, Review Rating, Shipping Type)

3. Methodology and Exploratory Data Analysis (EDA)

The project began with data preparation and cleaning using Python (as confirmed by the uploaded Jupiter Notebook).

- **Data Preparation:** Initial steps involved data loading using the Pandas library and basic exploration (df.info (), df.describe()).
- **Feature Engineering:** Columns were processed for consistency, and potentially new features like age_group were created to facilitate deeper analysis.
- **Clustering:** The "Coustmer_Trend_Analysis Project_file(K-Means clustering).ipynb" confirms the application of advanced segmentation techniques, specifically **K-Means Clustering**, to group customers based on their purchase behavior.

4.Exploratory Data Analysis

A rigorous exploratory analysis was conducted to examine distribution patterns, detect anomalies, and understand relationships between variables.

The **boxplots generated in the notebook** indicate that customer age is well distributed, with most customers falling between 25 and 60 years. Purchase amounts vary between 20 and 100 units, with consistent distribution and no extreme behavior. Previous purchases show a skewed distribution, which is typical in retail where most customers shop only a few times while a smaller segment contributes disproportionately to repeat purchases. Purchase frequency in days exhibited a small number of unusually large values, revealing periods of long inactivity, which were useful for retention analysis.

The **histograms** reinforced these distributions and highlighted natural segmentation tendencies among age groups, purchase frequency, and review ratings. Additionally, the **correlation heatmap** confirmed that no strong linear relationships existed between most behavioral variables, which justified the need for machine learning clustering to uncover underlying groupings that are not visible through simple correlations.

These early findings established a clear understanding of the dataset and shaped the direction of the clustering strategy.

5. Feature Engineering

To enrich the dataset and improve clustering performance, several engineered features were added. Age groups were created to categorize customers into meaningful demographic segments such as young adults, adult, middle-aged, and seniors. Purchase frequency in days provided a more intuitive measure of how often a customer engaged with the platform. Categorical variables such as gender, discount usage, and subscription status were encoded numerically for compatibility with machine learning algorithms. These engineered variables helped the model identify deeper behavioral differences and enhance the interpretability of the final segmentation.

6. Key Findings and Analysis

The Exploratory Data Analysis (EDA) yielded several critical insights into customer demographics, loyalty, and revenue contributions:

Metric	**Value**
Number of Customers	3.9K (3,900 transactions) *
Average Purchase Amount	$59.76
Average Review Rating	3.75

6.2 Subscription Status and Loyalty

A significant portion of the customer base is not subscribed, but purchase frequency correlates strongly with subscription likelihood:

- **Non-Subscribers:** 73%
- **Subscribers:** 27%
- **Loyalty Trend:** Buyers with five or more previous purchases are more likely to subscribe to the service.

6.3 Revenue Breakdown by Age Group

Revenue analysis shows that younger customer segments contribute the highest total revenue.

Age Group	Total Revenue (USD)
Young Adult	53,657
Middle-aged	51,450
Adult	46,996
Senior	46,559

The Young Adult segment generates the most revenue, suggesting targeted marketing efforts should maintain engagement with this demographic while seeking to maximize lifetime value from the Middle-aged and Senior segments.

6.4 Revenue by Product Category

Analysis indicates varying revenue contribution across categories such as Accessories, Clothing, Footwear, and Outerwear.

7. Machine Learning — Customer Segmentation

7.1 K-Means Clustering

K-Means clustering was selected as the primary segmentation algorithm due to its efficiency and interpretability. The **Elbow Method**, which compares within-cluster distortion across different values of k, indicated that **three clusters (k=3)** provided an optimal balance between cohesion and separation.

Cluster analysis revealed three distinct customer personas:

Cluster 0 – Occasional Middle-Age Buyers
This group purchases sporadically, maintains moderate spending levels, and displays average review ratings. Their purchasing cycles suggest they may need targeted engagement strategies such as seasonal offers or interest-based recommendations.

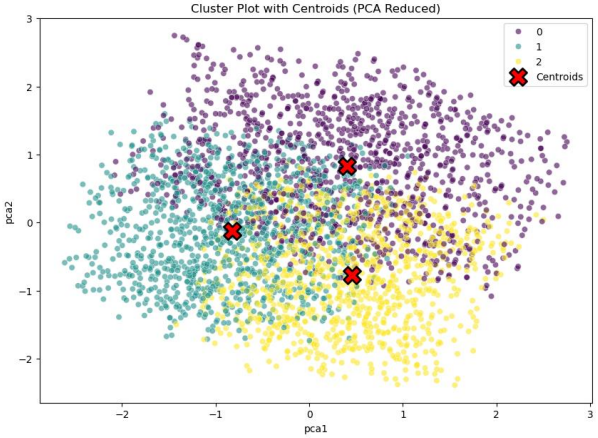
Cluster 1 – Younger Frequent Buyers
These customers shop more frequently, respond strongly to discounts, and tend to make value-driven purchases. Their behavior suggests high engagement potential, making them suitable for promotional campaigns and upsell strategies.

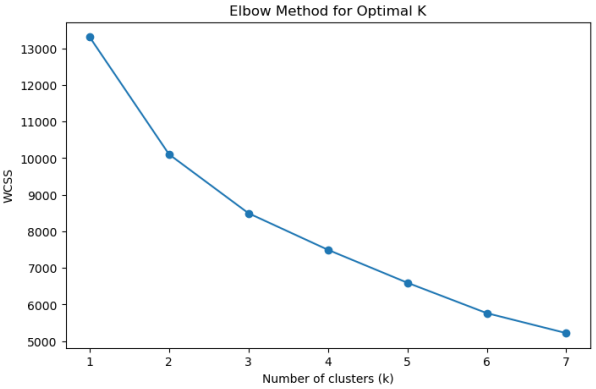
Cluster 2 – High-Value Loyal Buyers
This segment exhibits the highest purchase amounts, strong repeat behavior, and higher review ratings. They

represent the most profitable customer group and would benefit from premium loyalty programs, exclusive access offers, and personalized communication.

The **PCA visualization** clearly showed separation between these groups, reinforcing the segmentation's reliability.

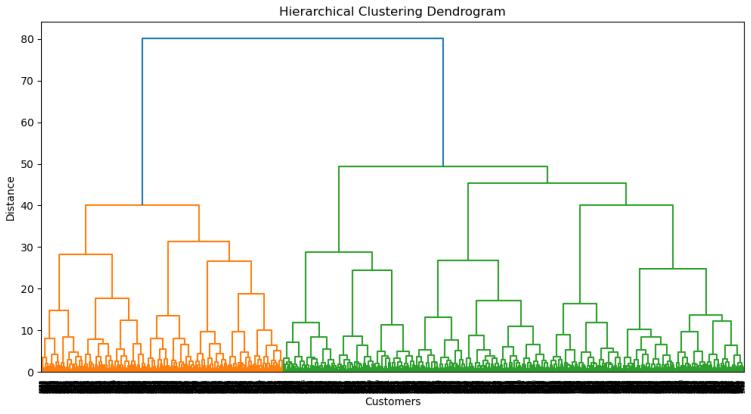
Although Cluster 0 contributes the highest total revenue due to a larger customer base, Cluster 2 represents high-value customers with higher average spending and stronger loyalty. This makes Cluster 2 strategically important despite generating slightly lower overall revenue.





7.2 Hierarchical Clustering

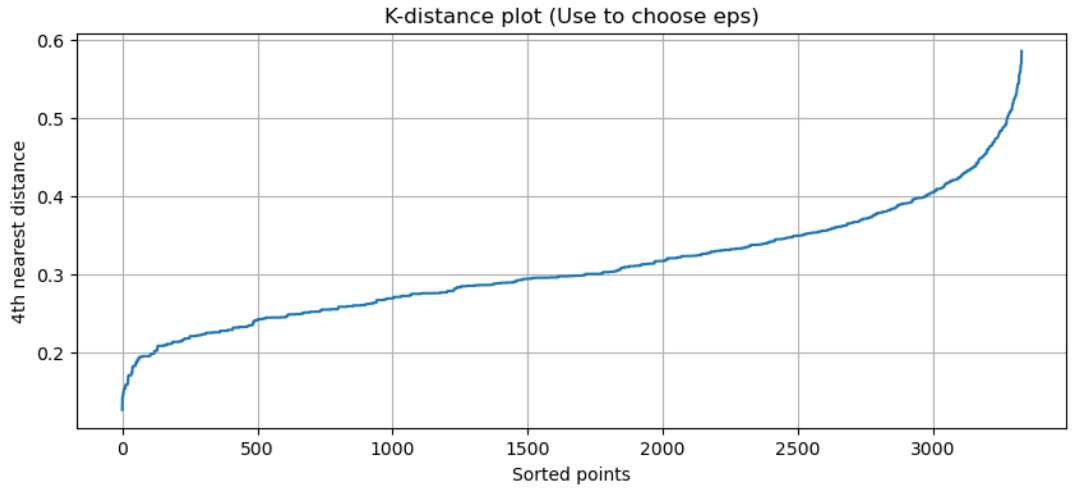
Hierarchical clustering was applied to validate the K-Means segments and visualize natural customer groupings. The **dendrogram** demonstrated clear boundaries that closely aligned with the three-cluster structure found in K-Means. This cross-validation showed that the segmentation is not arbitrary but rooted in strong natural divisions in customer behavior.



7.3 DBSCAN Clustering

DBSCAN, a density-based clustering method, was used to detect outliers and secondary cluster formations. This analysis uncovered two main behavioral groups and several outlier points—customers whose purchasing patterns were atypical or significantly divergent. In a business context, such customers may represent either

high-value VIPs, churn risks, or fraudulent behavior depending on their characteristics. DBSCAN complemented the main clustering model by revealing edge case customers who might require special handling.



8. Visualizations and Power BI Dashboard

All derived insights have been compiled into a comprehensive and interactive dashboard to facilitate visual analysis and decision-making. The report leverages the provided Power BI file (Customer_trend_analysis.pbix) for interactive data exploration.

The dashboard includes key visuals such as:

- Donut charts display the percentage breakdown of customers by Gender and Subscription Status.
- Bar charts illustrate Revenue by Category and Revenue by Age Group.
- KPI cards highlight the Average Purchase Amount and Average Review Rating.

A separate segmentation page highlights the machine-learning clusters, comparing their revenue contribution, purchase frequency, and customer distribution. The design is simple, well-structured, and supports dynamic filtering, enabling the user to explore trends and customer segments efficiently. Overall, the dashboard effectively translates analytical findings into an intuitive business-ready format.

9. Business Recommendations

Based on the observed customer behavior, the following recommendations are proposed:

1. **Boost Subscriptions:** Implement promotions that specifically target users with a high number of previous purchases (e.g., >5 purchases), as this group demonstrates high loyalty and conversion potential for subscription status. Promote exclusive subscriber benefits vigorously.
2. **Targeted Age Campaigns:** Maintain strong engagement with the Young Adult segment (highest revenue) while developing specific campaigns to increase spending among the Middle-aged and Senior demographics.
3. **Inventory Optimization:** Use the Revenue by Category breakdown to prioritize stocking and marketing of high-performing product categories.

4. **Data Utilization:** Leverage the customer clusters identified via K-Means to personalize marketing messages and product recommendations for distinct customer personas.

Conclusion

This project successfully demonstrates an end-to-end analytical approach for understanding and segmenting customer behavior using modern data science tools. Through detailed exploration, robust clustering techniques, SQL-driven business analytics, and visual storytelling in Power BI, the project uncovers meaningful customer personas and provides strategic business insights. The segmentation identifies three distinct customer groups with unique behaviors and revenue patterns, guiding targeted marketing, loyalty initiatives, and personalized communication strategies.

The multi-layered analysis combining machine learning, statistical understanding, and business intelligence creates a comprehensive foundation for customer-centric decision-making. This project represents a real-world analytical workflow and serves as a strong demonstration of integrated Python, ML, SQL, and BI skills.