Customer Segmentation using Clustering Techniques

A Machine Learning Approach to Identify and Analyze Key Customer Groups

Felipe Chen, Mykyta Shutov, Misha Semenov, Amareswar Doddi

FORDHAM UNIVERSITY

Introduction & Objective

Objective : Explore customer purchasing behavior and identify key characteristics of the best buyers.

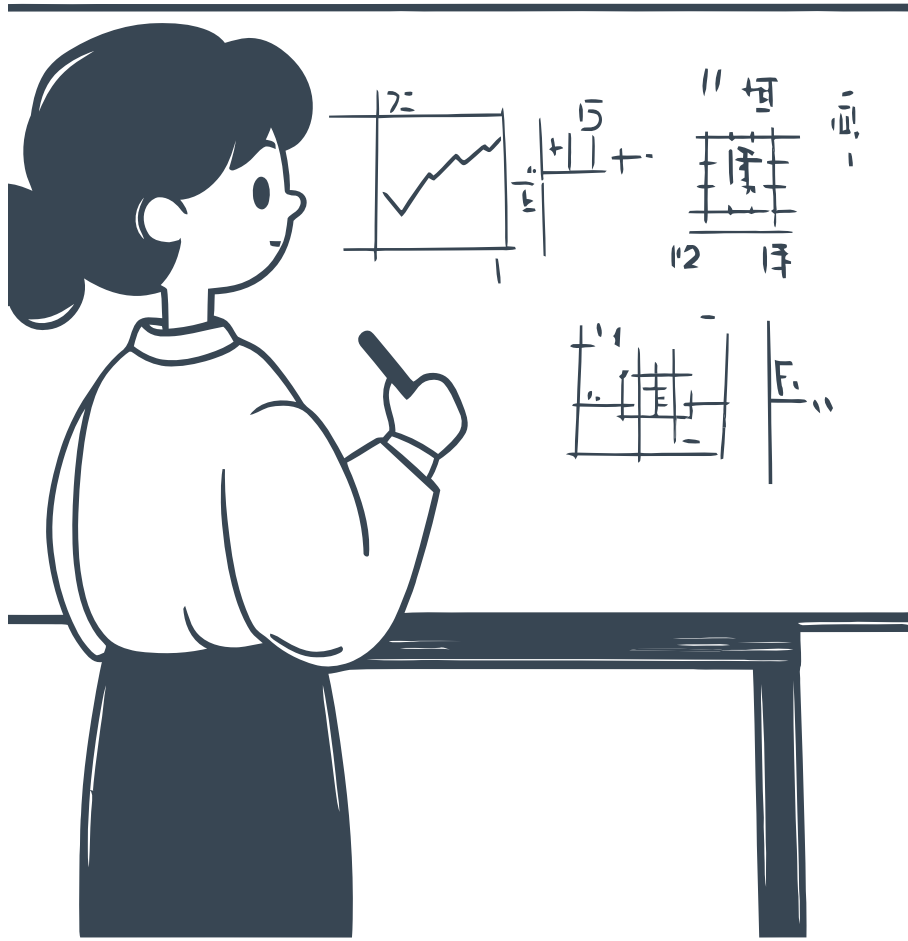
Importance : Understanding which customers contribute most to sales can help businesses develop targeted marketing strategies.

Method : Applied Unsupervised learning (K-Means and Hierarchical Clustering)

Problem-Solving Statement

Businesses struggle to identify which customers generate the most value and how to target them effectively.

- The goal is to use unsupervised ML (clustering) to group customers based on their spending patterns, income, and preferences.
- These insights can help identify high-value segments and guide marketing and retention strategies.

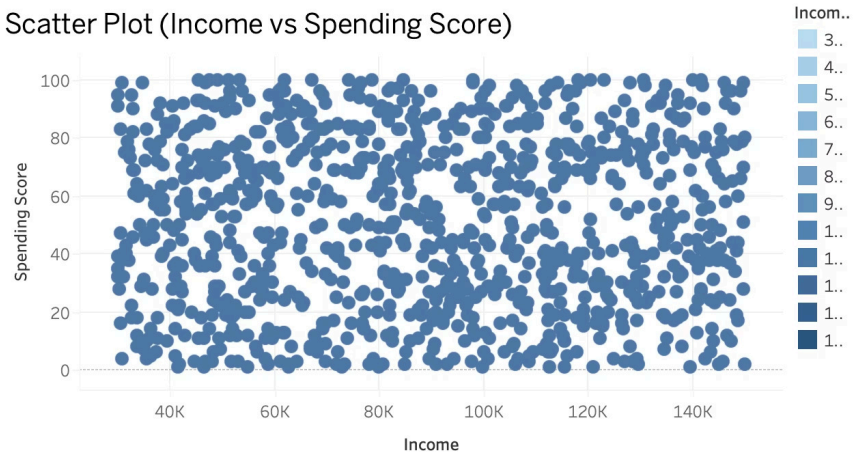


Dataset Description

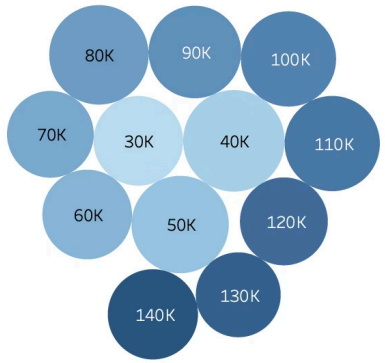
Feature	Type	Description
id	Integer	Unique customer identifier
age	Integer	Customer's age
gender	Categorical	Gender of the customer (Male, Female, Other)
income	Integer	Annual income of the customer
spending_score	Integer	Score representing customer's spending habits (1–100)
membership_years	Integer	Years of membership with the company
purchase_frequency	Integer	Number of purchases per year
preferred_category	Categorical	Product category most frequently purchased
last_purchase_amount	Float	Amount spent in the last purchase

Exploratory Data Analysis (Before Clustering)

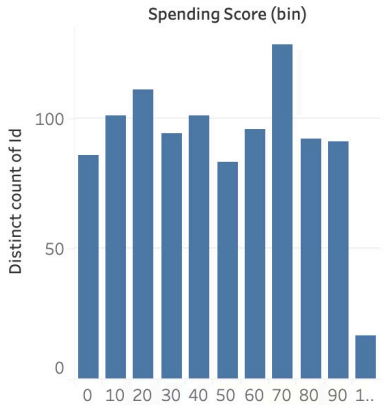
Scatter Plot (Income vs Spending Score)



Bubble Chart (Income Distribution by Bins)



Histogram of Spending Score



- **Scatter Plot (Income vs Spending Score) :** Spending behavior varies widely across all income levels, indicating no simple linear relationship between income and spending score.
- **Bubble Chart (Income Distribution by Bins) :** Most customers fall into mid-to-high income ranges (50K–110K), with fewer customers at extreme low- or high-income levels.
- **Histogram of Spending Score (Binned) :** Spending scores are spread evenly across the 0–100 scale, showing multiple customer behavior types.

Methodology (*Project Workflow*)

1

Data Collection

Downloaded dataset from Kaggle.

2

Data Preprocessing

Encoded categorical data, normalized numeric fields.

3

Feature Selection

Chose *Income* and *Spending Score* for clustering.

4

Model Building

Applied K-Means and Hierarchical clustering.

5

Evaluation

Used Elbow Method, Inertia, and Silhouette Coefficient.

6

Visualization

Used Tableau for Charts

Model Implementation

K-Means:

Identified optimal cluster count using the Elbow Method.

Hierarchical Clustering:

Validated with Dendrogram.

How K-Means Clustering Works:

01

K-Means groups customers into four segments based on income and spending behavior.

02

Each cluster represents a unique customer persona with distinct purchasing patterns.

03

Cluster centroids highlight the average profile of each group, improving interpretation.

04

K = 4 was selected because it clearly separates high-value, medium-value, and low-value customers.

05

This segmentation supports targeted marketing, cross-selling, retention, and customer value analysis.

06

Clustering reveals hidden patterns in customer data that are not visible in simple EDA charts.

How K-Means Clustering Works:

01

```
# Clusters seems to be the best (k=4)

kmeans = KMeans(n_clusters=4, init='k-means++', max_iter=500, random_state=42)
y = kmeans.fit_predict(X_scaled)
labels_kmeans = kmeans.labels_
labels_kmeans
```

```
# Silhouette score is reasonable

score = silhouette_score(X_scaled, labels_kmeans)
print(score)
```

```
0.4189735063734962
```

Visualization : K-Means Clustering

Cluster 1: High spending score - low income.

Cluster 2: High spending score - high income.

Cluster 3: Low spending score - low income.

Cluster 4: Low spending score - high income.

```
# Getting numerical summary of clusters
df_scaled.astype(float)
df_scaled.groupby('Clusters').mean().round(2)
```

Clusters	age	income	spending_score	membership_years
0	43.32	-0.93	-0.92	5.45
1	43.59	0.82	-0.83	5.42
2	44.04	-0.79	0.83	5.61
3	44.19	0.94	0.93	5.36

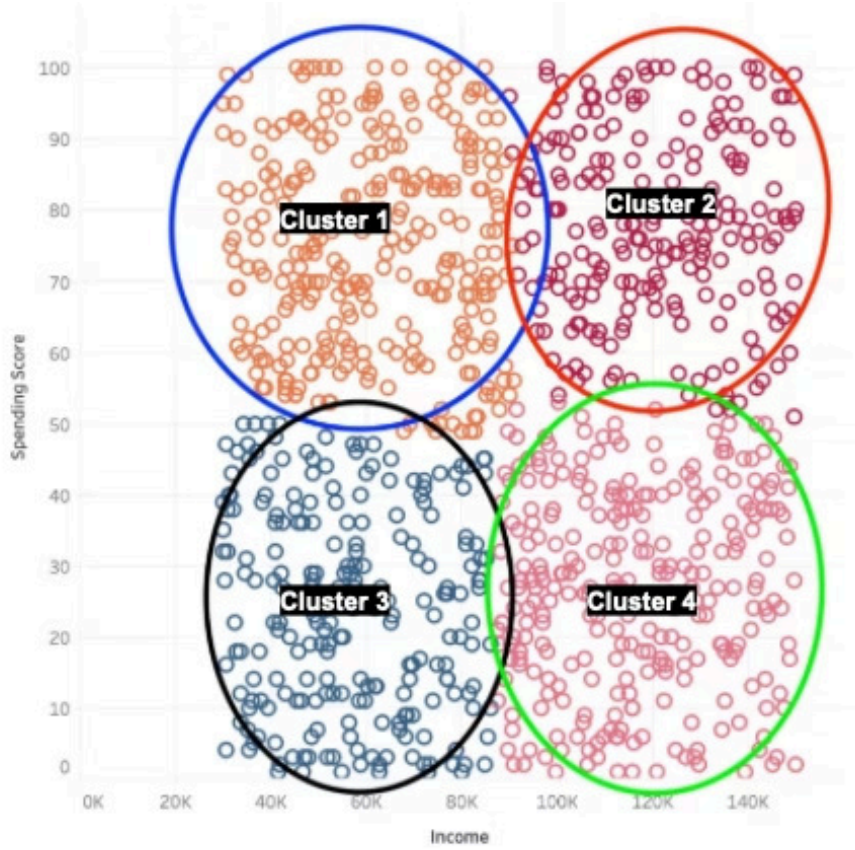


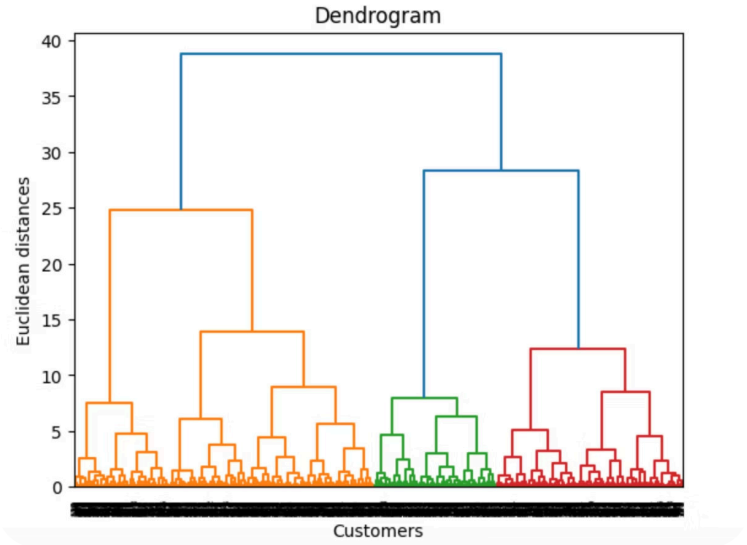
Chart 1: K-Means – Income vs Spending Score

How Hierarchical Clustering Works:

01

```
import scipy.cluster.hierarchy as sch

dendrogram = sch.dendrogram(sch.linkage(X_scaled, method = 'ward'))
plt.title('Dendrogram')
plt.xlabel('Customers')
plt.ylabel('Euclidean distances')
plt.show()
```




```
from sklearn.cluster import AgglomerativeClustering

hc = AgglomerativeClustering(n_clusters = 5, metric = 'euclidean', linkage = 'ward')
y_hc = hc.fit_predict(X_scaled)
```

```
# Silhouette Score is lower compared to K-means

print('Silhouette Score HC:', silhouette_score(X_scaled, y_hc))
```

```
Silhouette Score HC: 0.37411356233203535
```

Visualization : Hierarchical Clustering (HC)

Cluster 1: High spending score - low income.

Cluster 2: High spending score - high income.

Cluster 3: Low spending score - low income.

Cluster 4: Low spending score - middle income.

Cluster 5: Low spending score - high income.

HC_Cluster	age	income	spending_score	membership_years
0	43.68	-0.82	0.80	5.56
1	43.06	0.22	-0.69	5.39
2	44.97	0.99	0.92	5.38
3	43.56	-1.01	-1.15	5.78
4	43.61	1.32	-0.82	5.13

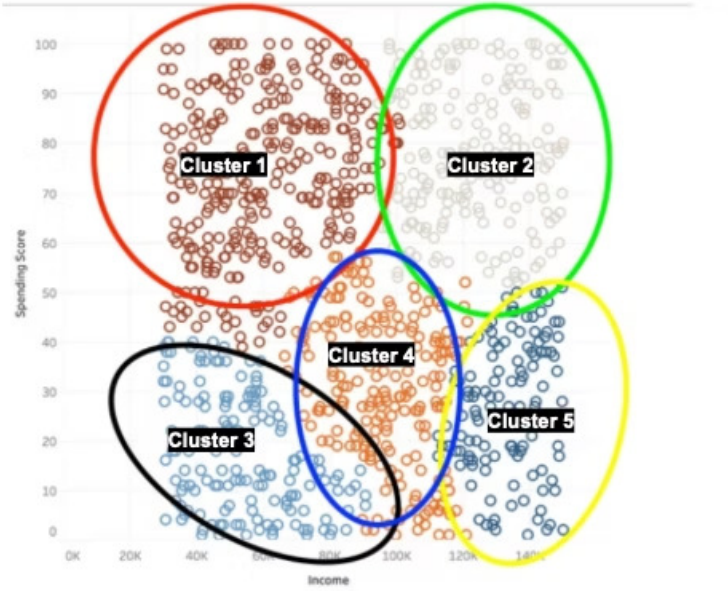


Chart 2: HC – Income VS Spending Score

K-Means Cluster Profiles & Key Segment Insights

- **Cluster Size Chart** : The largest group is Low Income – High Spend, showing strong spending despite lower income levels.
- **Average Age** : Average age is similar across all clusters (early 40s), meaning age is not a major segmentation factor.
- **Average Annual Income** : High Income – High Spend has the highest income, aligning with premium customer potential.
- **Average Spending Behavior Score** : High Income – Low Spend customers represent an opportunity for targeted promotions to increase spending.

Insight from clusters

The data clusters clearly show that **Spending Score** is the most influential and actionable factor in customer segmentation.

It creates two distinct and consistent groups - high spenders and low spenders - more reliably than income does.

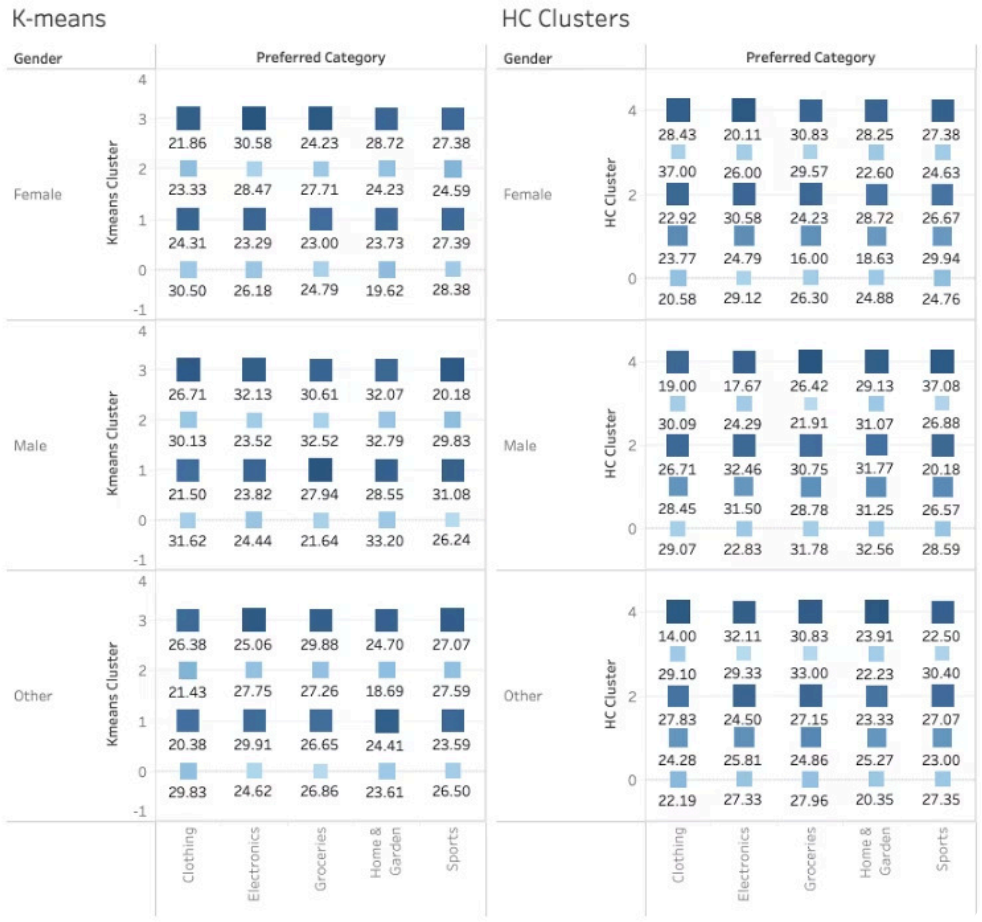
From a business perspective, the strategic priority should be to strengthen purchasing behavior within the large low-spending segment. Building a strong brand presence and fostering customer loyalty would be effective approaches to achieving this goal as well as keeping high spending customers.

Clustering Visualization Insights

This visualization allows us to examine how each cluster behaves in terms of purchasing preferences. By breaking the clusters down by gender and preferred product category, we can see differences in buying patterns even among customers with similar income and spending profiles.

Across K-means and hierarchical clusters, females, males, and customers identifying as "other" show distinct category preferences, and their purchase frequencies vary noticeably within the same cluster level.

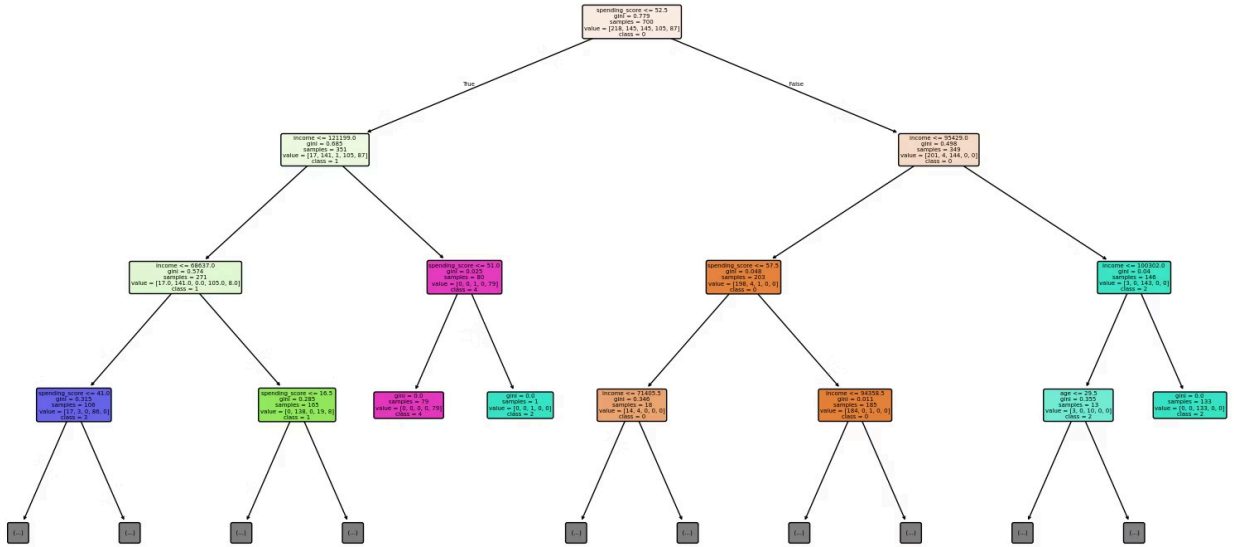
This suggests that demographic factors, especially gender, continue to influence buying behavior independent of income-based segmentation.



Decision Tree Model Insights

A Decision Tree splits the data based on the most important features to predict whether a customer is a **high spender or low spender**.

The tree identifies **Income, Spending Score,** and **Age** as key predictors.



Customer Segments Overview

Devoted Fans

High engagement despite lower income

Power Customers

Premium buyers with high value

Emerging Relationships

Early-stage customers with growth potential

Opportunity Accounts

Untapped high-income prospects

Devoted Fans

High Spend / Low Income

Profile

- **Characteristics:** High spending score despite lower income levels
- **Behavior:** Loyal customers who prioritize value and quality over price
- **Size:** Largest customer segment in our data

Marketing Actions

Loyalty Programs

Reward frequent purchases with points, discounts, and exclusive access

Value-Based Messaging

Emphasize quality, durability, and long-term value in communications

Community Building

Create brand communities and user-generated content campaigns

Flexible Payment Options

Offer installment plans and seasonal promotions to support purchasing power

Power Customers

High Spend / High Income

Profile

- **Characteristics:** High income with high spending scores - premium customer segment
- **Behavior:** Quality-focused buyers willing to pay premium prices for superior products
- **Value:** Highest revenue per customer with strong profit margins

Premium Strategies

VIP Treatment

Exclusive access to new products, private sales, and premium customer service

Premium Product Lines

Curate high-end collections and luxury offerings specifically for this segment

Personalized Experiences

One-on-one consultations, custom solutions, and white-glove service

Strategic Partnerships

Collaborate with luxury brands and offer exclusive co-branded experiences

Emerging Relationships

Low Spend / Low Income

Profile

- **Characteristics:** Lower income customers with currently low spending patterns
- **Behavior:** Price-sensitive buyers who need encouragement to increase purchase frequency
- **Potential:** Early-stage relationships with significant growth opportunity

Cost-Efficient Strategies

Entry-Level Products

Offer affordable starter products to build initial purchasing habits

Educational Content

Provide value through tutorials, tips, and product education to build trust

Gradual Engagement

Use email marketing and social media to maintain regular, low-cost touchpoints

Referral Incentives

Encourage word-of-mouth growth through friend referral programs and social sharing

Opportunity Accounts

Low Spend / High Income

Profile

- **Characteristics:** High income customers with surprisingly low spending scores
- **Behavior:** Have purchasing power but aren't currently engaged with our brand
- **Opportunity:** Significant untapped revenue potential from affluent but inactive customers

Reactivation Strategies

Targeted Promotions

Create compelling offers specifically designed to re-engage dormant high-value prospects

Premium Positioning

Showcase high-quality products and services that match their income level and lifestyle

Competitive Analysis

Research what competitors are offering and develop superior value propositions

Direct Outreach

Personal sales calls and account management to understand barriers and rebuild relationships

Strategic Takeaways

Our clustering analysis reveals four distinct customer segments that require tailored marketing approaches.

- Devoted Fans represent our most loyal base and should receive loyalty-focused retention strategies.
- Power Customers demand premium experiences and represent our highest-value opportunities for revenue growth.
- Emerging Relationships need cost-effective nurturing to develop their purchasing potential over time.
- Opportunity Accounts present the greatest untapped revenue potential, requiring targeted reactivation campaigns to convert their high income into high spending.

By implementing segment-specific strategies, we can optimize marketing spend, improve customer lifetime value, and drive sustainable revenue growth across all customer groups.

Expected Results & Conclusion

- The model successfully identified **key customer segments** using both **Hierarchical Clustering (5 groups)** and **K-Means (4 groups)**.
- The **K-Means** formed compact, centroid-based clusters for efficient segmentation, while **Hierarchical Clustering** revealed broader, distance-based linkages for deeper analysis.
- These insights enable **targeted marketing**, **optimized product placement**, and **customer retention** strategies.
- Provides a **scalable foundation** for future customer segmentation using Machine Learning.

Future Scope & References

Future Work:

- A/B Test
- Add behavioral features like purchase history, location, or sentiment.
- Build a real-time dashboard in Tableau or Power BI.

References:

- Dataset: Kaggle
- Tools: Python (scikit-learn, pandas, seaborn), Tableau
- University resources and academic materials.