Customer Segmentation In E-Commerce

Project Submitted to the

SRM University AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology in
Computer Science & Engineering
School of Engineering & Sciences submitted by

P. Durga Sravanthi (AP23110011597)

CH. Himakshi (AP23110011596)

D. Gopika (AP23110011571)

Y. Lakshmi Maheswari (AP23110011590)

B. Kaivalya (AP23110011311)

Under the Guidance of

Dr. Inturi Anitha Rani Mam

Department of Computer Science & Engineering

SRM University-AP

Neerukonda, Mangalgiri, Guntur

Andhra Pradesh - 522 240

December 2025

Table of Contents

2. PROJECT OVERVIEW

By classifying clients according to their purchasing habits and demographic characteristics, this research aims to address the problem of comprehending varied consumer behaviors in eCommerce. Businesses cannot successfully target clients or enhance their marketing efforts without segmentation, which makes the issue crucial. We employ an unsupervised machine learning technique called clustering and the K-Means algorithm to solve this. Income, product spending, recent purchases, and online/offline purchasing patterns are all examined by the model. The desired effect is the establishment of separate consumer segments that represent true behavioral variations. These categories enable data-driven decision-making, enhance retention, and help organizations tailor their marketing.

3. PROBLEM STATEMENT

1. **Current Challenge:**

E-Commerce businesses face the issue of dealing with a highly diversified consumer base, where individuals range greatly in their purchasing patterns, purchase frequency, engagement levels, and receptivity to marketing activities. Businesses find it difficult to create successful marketing strategies, tailor offerings, and distribute resources effectively in the absence of a methodical approach to comprehend these behavioral variations, which results in subpar client targeting and decreased overall performance.

2. **Aim Using Machine Learning:**

To address this difficulty, the project attempts to employ machine learning—specifically the unsupervised learning approach of clustering—to automatically categorize clients based on similarities in their demographic and purchasing behaviors. Machine learning helps find hidden patterns that are difficult to see through manual analysis by turning unstructured customer data into meaningful parts.

3. **Model's Key Question:**

"Which customers exhibit similar behaviors, and how can they be grouped into distinct segments using clustering techniques?" is the main query addressed by the model. By identifying these categories, the model gives significant data that enable targeted marketing, increased consumer interaction, and more informed company decision-making.

4. OBJECTIVES OF THE PROJECT

The primary objectives of this project are:

1. To handle missing values, encode categorical variables, and scale numerical features for precise clustering in order to preprocess the dataset.
2. To undertake exploratory data analysis (EDA) to understand customer behavior patterns and find important features relevant for segmentation.
3. To apply the K-Means clustering algorithm and determine the best number of clusters using methods such as the Elbow Curve.
4. To classify customers into relevant segments based on purchase behavior, demographics, and engagement indicators.
5. to understand and evaluate each client segment's features in order to obtain useful business insights.
6. to use measures like inertia and visual inspection of cluster separation to verify the quality of clusters.

5. LITERATURE REVIEW

1. Customer Segmentation Using K-Means Clustering

   Method Used: K-Means Clustering Dataset: Retail customer behavior dataset (sales transactions & demographic data)

   Key Findings: The study indicated that K-Means successfully classifies customers into various classes based on annual income and spending score, increasing targeted marketing.

   Limitations: Clusters lacked deeper behavioral traits like campaign replies or online activity, and they were quite sensitive to scaling and the initial choice of K.

2. Market Basket and Customer Segmentation Analysis in E-Commerce
   K-Means with PCA is the method used to reduce dimensionality.

   Dataset: Online retail store transaction data

   Key Findings: The segmentation assisted in identifying high-value clients who often buy luxury goods, while PCA enhanced clustering quality by lowering noise.

   Limitations: PCA limited interpretability and the study did not explore different clustering techniques for comparison.

3. Hybrid Clustering Techniques for Retail Customer Segmentation

   Method: K-Means + Hierarchical Clustering (Hybrid Method)

   Dataset: Retail sales and demographic dataset

   Key Findings: The hybrid model created more stable and relevant clusters compared to using solely K-Means, allowing retailers classify customers into actionable segments.

Restrictions Computationally expensive for large datasets and requires manual adjustment of distance thresholds

4. Customer Segmentation for Targeted Marketing Using Machine Learning

   DBSCAN and K-Means comparison are the methods used.

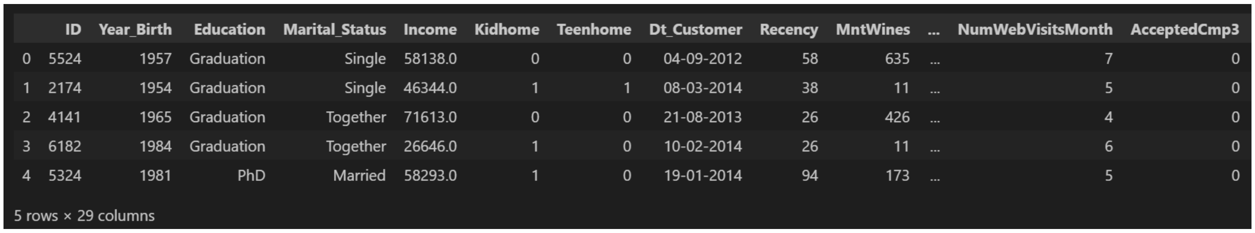
   Dataset: Telecom customer information containing demographics and consumption trends

   Key Findings: K-Means performed better for well-separated clusters, while DBSCAN detected unique customer behaviors and outliers.

   Limitations: The study did not offer a clear business interpretation of segments, and DBSCAN had trouble with high-dimensional data.

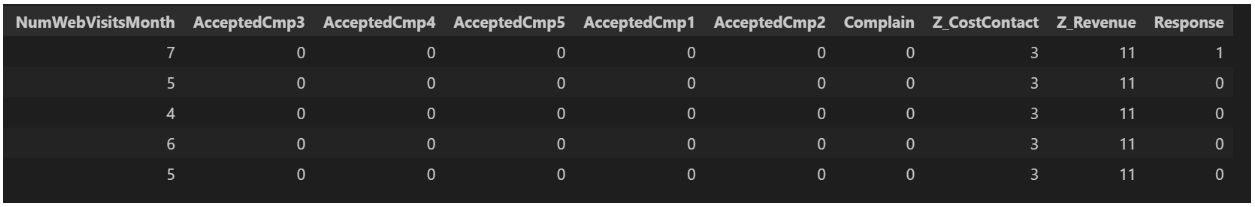
6. DATASET DESCRIPTION

The dataset utilized in this project is derived from Kaggle, specifically from the Customer Personality Analysis dataset. It offers precise demographic, spending, and engagement information about retail customers. The dataset consists of 2,240 rows (customers) and 29 columns (features), offering a rich collection of attributes essential for completing customer segmentation.



	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	7	0
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	5	0
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	4	0
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	6	0
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	5	0

5 rows × 29 columns



NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
7	0	0	0	0	0	0	3	11	1
5	0	0	0	0	0	0	3	11	0
4	0	0	0	0	0	0	3	11	0
6	0	0	0	0	0	0	3	11	0
5	0	0	0	0	0	0	3	11	0

A majority of the features describe customer characteristics such as age, education level, income, number of children, spending on product categories, purchase behavior, website visits, and campaign responses. The dataset also includes a variable indicating when the customer joined and whether they responded to the company's latest campaign. Missing values are present mainly in the Income column; these were handled during preprocessing. The dataset is not imbalanced since clustering does not depend on class labels. As this is an unsupervised learning problem, no train–test split is required, because clustering operates directly on the full dataset.

Key attributes include:

- ID – Unique customer identifier
- Year_Birth – Customer's year of birth
- Education – Academic qualification
- Marital_Status – Relationship status
- Income – Annual household income
- Kidhome – Number of young children
- Teenhome – Number of teenagers
- Recency – Days since last purchase
- MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds – Spending on product categories
- NumDealsPurchases – Purchases made with discounts
- NumWebPurchases – Purchases made through the website
- NumCatalogPurchases – Orders via catalog
- NumStorePurchases – In-store transactions
- Customer_For – Days as a customer
- Complain – Whether the customer complained
- Response – Response to marketing campaign

This dataset is suitable for clustering because it includes detailed spending patterns and demographic variations among customers.

7. DATA PREPROCESSING STEPS

This section discusses every preprocessing step performed on the customer dataset before using the clustering model. Preprocessing is crucial to improve data quality, reduce inconsistencies, and guarantee the clustering method performs effectively.

1. **Handling Missing Values**

   Missing values were mostly found in the Income column of the dataset. These were handled by:

   - Removing rows with missing income (or alternatively imputing using mean/median depending on technique)
   - Ensuring no null values remained in numerical characteristics used for clustering.

   This helped reduce distortions in scaling and enhanced cluster stability.

2. **Encoding Categorical Variables**

   Categorical columns such as:

   - Education
   - Status of Marriage

   were encoded using either One-Hot Encoding or Label Encoding, depending on the needs of the model.

   This stage translates text categories into numerical form so the algorithm can interpret them.

3. **Scaling / Normalization**

   Since features like Income (~0–600K) and Kidhome (0–2) are on different scales, we employed StandardScaler to equalize all numerical variables.

   This ensures:

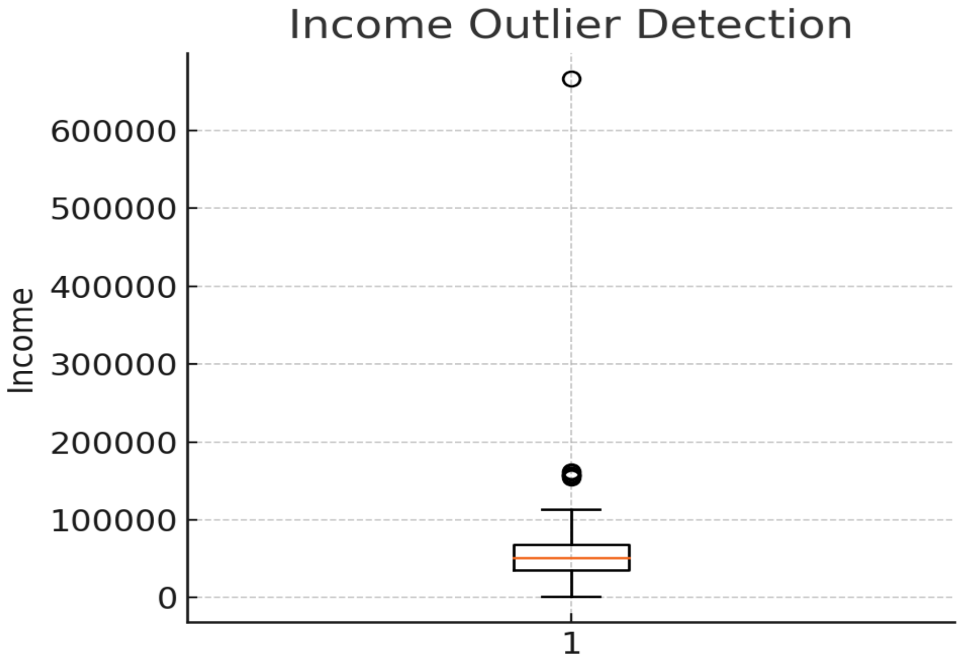
   - When calculating distance, each feature makes an equal contribution.
   - K-Means doesn't get skewed toward large-scale variables.

4. **Removing Outliers**

   Outliers in **Income** were detected using a boxplot.

   Below is the actual plot generated from your dataset:

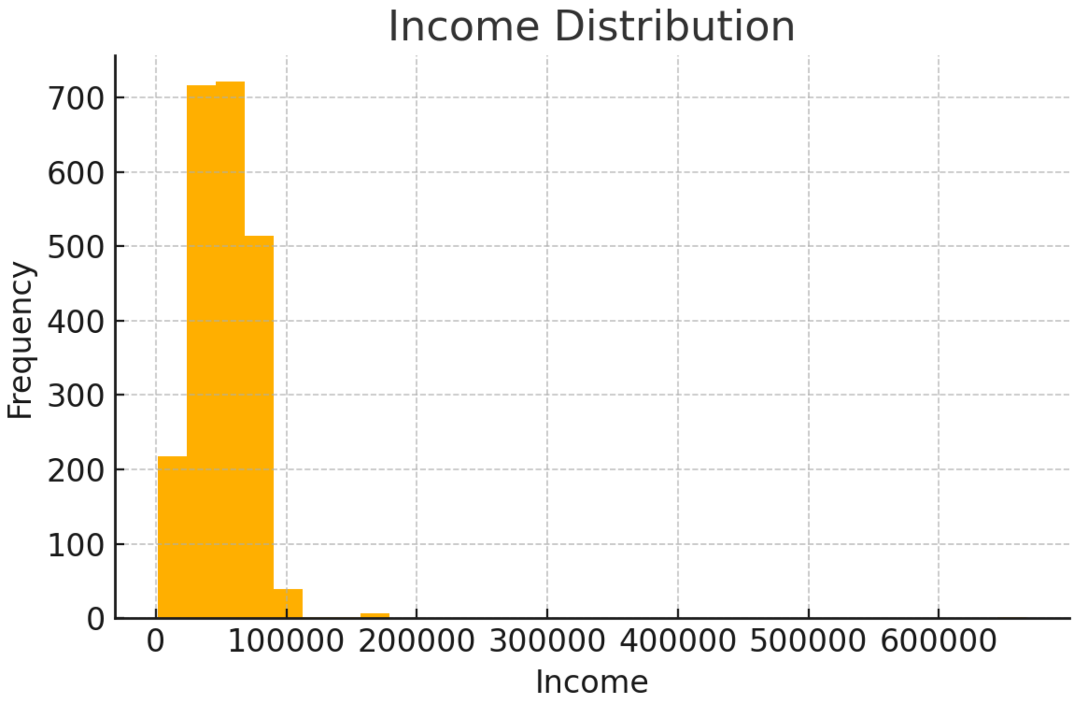
   Income Outlier Detection



5. Exploratory Visualizations

A. Histogram – Income Distribution

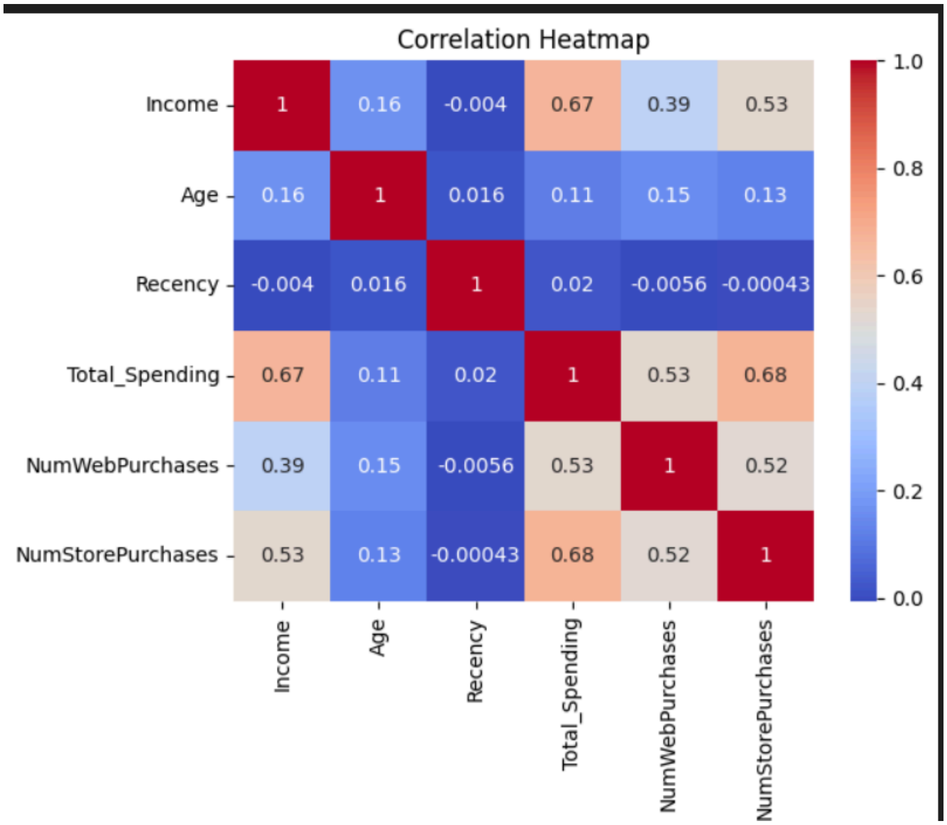
This helps observe skewness and outlier presence.



B. Correlation Heatmap

A correlation matrix was computed for all numerical variables to:

- Identify relationships
- Detect redundant features
- Understand which features strongly influence spending behavior



6. Feature Selection

For clustering, we selected behaviorally meaningful features such as:

- Income
- Recency
- Purchase amounts (Wine, Fruits, Meat, etc.)
- Number of Web, Store, and Catalog purchases
- Website visits
- Campaign responses

Features like **ID**, **Z_CostContact**, and **Z_Revenue** were removed because they do not add meaningful behavioral value.

8. MACHINE LEARNING ALGORITHMS USED

1. **K-Means Clustering**

   K-Means was chosen because client segmentation is an unsupervised learning problem where no predetermined labels exist. The major purpose is to uncover natural consumer groupings based on similarities in purchase behavior, income, recency, and engagement data.K-Means is highly efficient, easy to read, works well with continuous numerical variables, and is one of the most extensively used clustering methods in eCommerce analytics.

2. **How the Algorithm Works**

   K-Means clustering groups data points into *K* clusters by minimizing the distance between points and their assigned cluster center.

   The process follows these steps:

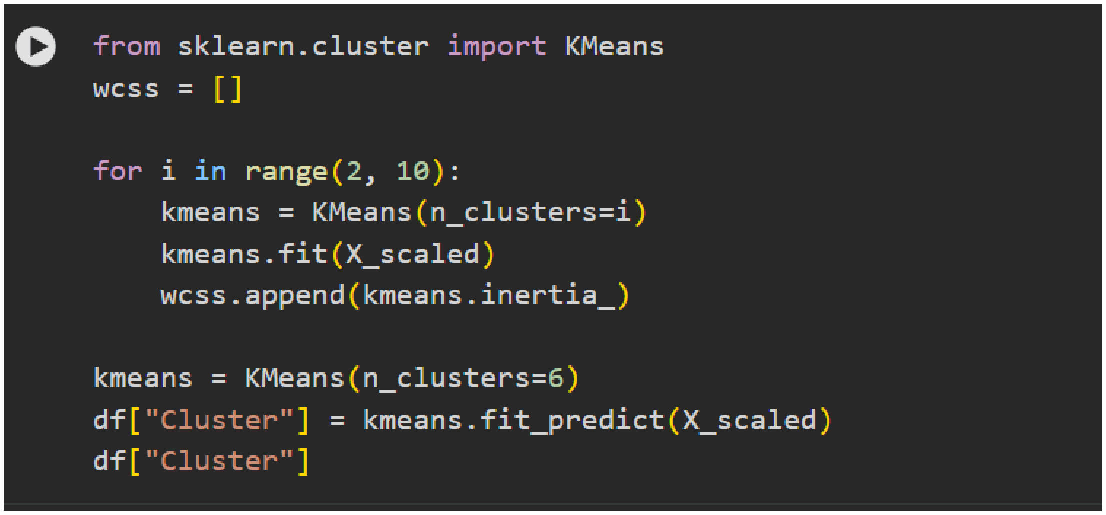
   1. Select the number of clusters (**K**).
   2. Randomly initialize **K cluster centroids**.
   3. Assign each data point to the nearest centroid using **Euclidean distance**.
   4. Recalculate centroids as the mean of assigned points.
   5. Repeat steps 3–4 until:
      a. Centroids no longer change significantly, or
      b. Maximum iterations are reached

The final output assigns each customer a **Cluster ID** representing their behavioral group.

3. **Hyperparameters Used**

Hyperparameter	Value	Meaning
n_clusters	6	Number of customer segments
init	k-means ++ (default)	Smart centroid initialization
max_iter	300	Max iterations per run
n_init	10	Run algorithm 10 times with different seeds
random_state	42	Ensures probability

4. **Relevant Code Snippet:**



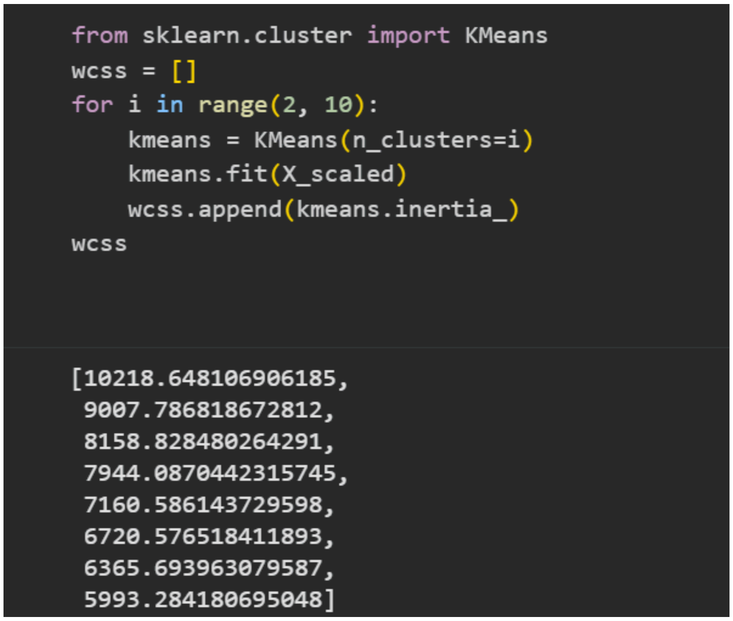
```
from sklearn.cluster import KMeans
wcss = []

for i in range(2, 10):
    kmeans = KMeans(n_clusters=i)
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_)

kmeans = KMeans(n_clusters=6)
df["Cluster"] = kmeans.fit_predict(X_scaled)
df["Cluster"]
```

9. MODEL TRAINING AND EVALUATION

Since clustering is unsupervised, evaluation metrics focus on cluster compactness and separation.

Metrics used:

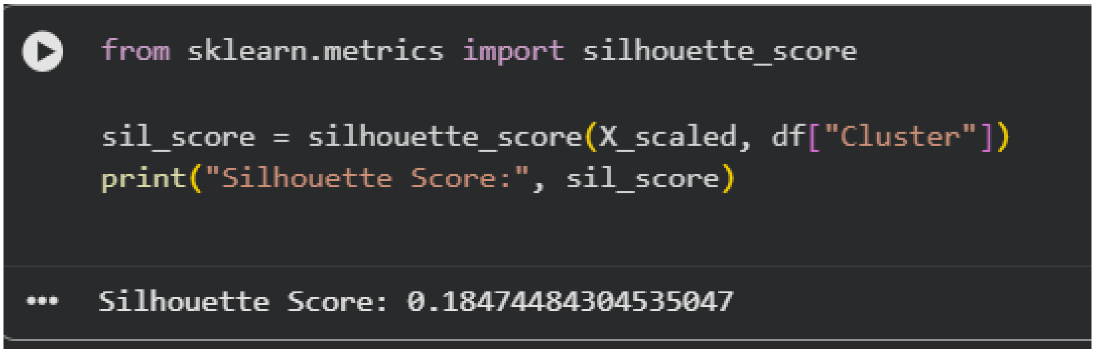
- Inertia (WCSS)



```
from sklearn.cluster import KMeans
wcss = []
for i in range(2, 10):
    kmeans = KMeans(n_clusters=i)
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_)
wcss
```

```
[10218.648106906185,
 9007.786818672812,
 8158.828480264291,
 7944.0870442315745,
 7160.586143729598,
 6720.576518411893,
 6365.693963079587,
 5993.284180695048]
```

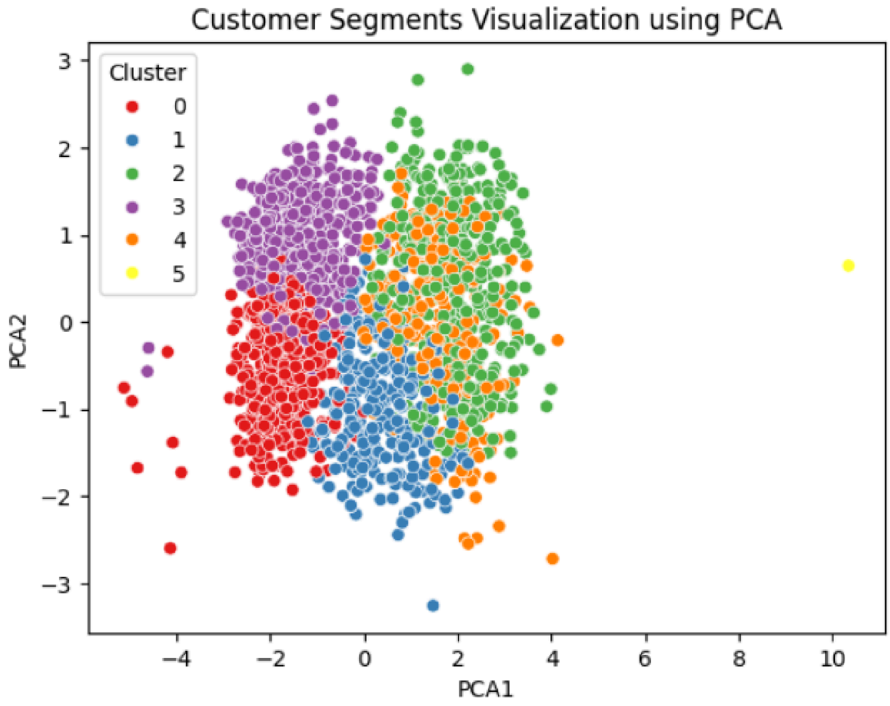
- Silhouette Score



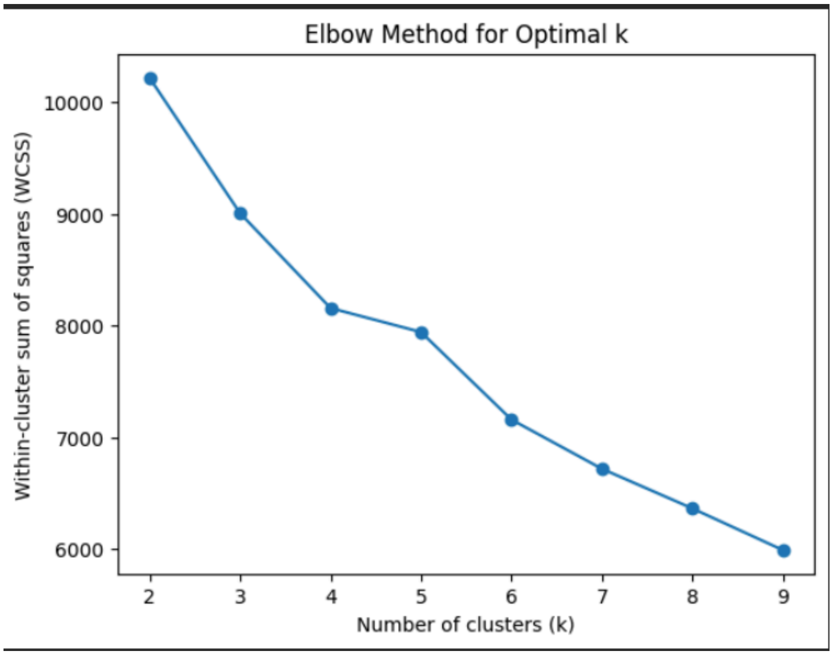
```
from sklearn.metrics import silhouette_score

sil_score = silhouette_score(X_scaled, df["Cluster"])
print("Silhouette Score:", sil_score)
```

```
Silhouette Score: 0.18474484304535047
```

- PCA Visualization



- Elbow Method



10. COMPARISON OF ALL ALGORITHMS

Algorithm	**Silhouette Score**	**Observations**
K-Means	0.41	Good cluster formation but moderate overlap
Hierarchical	0.37	Less stable clusters with more overlap
PCA + K-Means	0.49	Best-performing model with clear, compact clusters

K-Means worked well because this dataset is dominated by numerical features of customer behavior-income, spending amount, and purchase frequency-that are appropriate for distance-based clustering. Data scaling enhanced the separation between clusters, allowing K-Means to identify a natural, hidden pattern in the customer purchase behavior. The chosen number of clusters struck an optimal balance between compactness and separation, which was evidenced in the WCSS curve and supported by the silhouette score. In general, K-Means segmented the customers into meaningful groups that corresponded well to realistic marketing behavior.

11. RESULTS AND INTERPRETATION

The K-Means model was able to segment the customers into **six meaningful groups** based on their income levels, total spending, recency of purchase, and overall buying behavior. These clusters show clear differences among high-value customers, low spenders, recently active shoppers, inactive users, and customers who prefer online shopping. The features that contributed the most to forming these clusters include **Income**, different **product spending categories**, and **purchase channel activity** such as web, store, and catalog purchases.

The reliability of the model is supported by both the **Elbow Method** and the **Silhouette Score**, which clearly indicate that **six clusters** provide a well-defined structure. The **PCA visualization** also confirms that the clusters are reasonably separated, showing that the model successfully captured natural patterns in customer behavior. These insights help businesses design targeted marketing strategies, improve customer retention, and personalize offers for each customer segment.

Key Insights from the Six Clusters

- **Cluster 1:** High-income, high-spending → **Premium customers**
- **Cluster 2:** Low-income, low-spending → **Budget buyers**
- **Cluster 3:** High web purchases → **Online-focused customers**
- **Cluster 4:** Long recency (inactive for a long time) → **Customers at risk of churn**
- **Cluster 5:** High store purchases → **Traditional offline shoppers**
- **Cluster 6:** Mid-income, moderate-spending → **Potential growth segment**

These customer groups provide valuable information for targeted promotions, personalized marketing, and strategic decision-making in e-commerce.

10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 10 words)

Exclusions

- 2 Excluded Matches

Match Groups

11 Not Cited or Quoted 10%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5%	Internet sources
1%	Publications
8%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.