**CUSTOMER SEGMENTATION**

University of Missouri – Kansas City

# Lecturer: Syed Jawad HussaiShah

MAY09 2025

**Group Members**

**Chandrahaas Paladugula**

**Vamsi Krishna Meka**

**Srinivas Galla**

**Nikhilesh Chintala**

**INTRODUCTION:**

Customer segmentation is a data-driven approach used by businesses to categorize customers into distinct groups based on shared characteristics, behaviors, or preferences. This project focuses on using clustering techniques, specifically K-Means , Hierarchical Clustering and DBscan, to effectively segment customers. The goal is to identify customer segments that can help businesses optimize their marketing strategies, enhance customer satisfaction, and drive revenue growth.

**RELATED WORK:**

Customer segmentation is a widely studied topic in data science, with numerous techniques proposed to classify customers into distinct groups. Traditional approaches primarily use K-Means clustering due to its simplicity, speed, and scalability, making it a popular choice for customer segmentation in retail and marketing industries. However, K-Means is sensitive to the initial centroids and may not perform well on non-spherical clusters. Hierarchical Clustering is another popular technique that builds a tree-like structure (dendrogram) representing customer relationships. It provides a detailed view of how clusters are formed but suffers from high computational complexity, making it less suitable for large datasets. Variants of hierarchical clustering, such as Agglomerative and Divisive Clustering, offer flexibility in how clusters are formed. Recent research has explored advanced methods, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), which is effective for detecting clusters of varying shapes and identifying outliers. Hybrid models that combine multiple clustering methods, such as ensemble clustering, have demonstrated superior performance by leveraging the strengths of different algorithms.  
Additional Dataset tested: *https://www.kaggle.com/datasets/zubairmustafa/shopping-mall-customer-segmentation-data?resource=download*

**METHODOLOGY:**

This section describes the step-by-step approach used for customer segmentation in this project.

**1. Data Preparation**

* The customer dataset was loaded, containing features such as **Annual Income** and **Spending Score**.
* The dataset was preprocessed to ensure data quality:
  + Handle any missing values by either removing them or filling them with appropriate values.
  + Ensured that all data was numerical for clustering analysis.
* The data was standardized using **Standard Scaling** to bring all feature values within a consistent range (0 to 1), preventing any single feature from dominating the clustering process.

**2. K-Means Clustering**

* The optimal number of clusters (K) was determined using the **Elbow Method**, which involves plotting the within-cluster sum of squares (WCSS) for a range of K values and identifying the point of diminishing returns (elbow point).
* K-Means clustering was applied using the optimal K value(‘5’ from the Elbow method)
* Each customer was assigned a cluster label representing their segment.
* A scatter plot was generated to visualize the K-Means clustering results, with each cluster color-coded for clarity.

**3. Hierarchical Clustering**

* Hierarchical clustering was implemented to explore the hierarchical relationships among customers:
  + Four linkage methods were tested: **Ward, Average, Complete, and Single**.
  + Two distance metrics were explored: **Euclidean** and **Manhattan**.
* Dendrograms were generated for each linkage method to visualize the hierarchical structure of customer clusters:
  + The height of the vertical lines in the dendrogram represents the distance (or dissimilarity) between merged clusters.
  + A horizontal line was used to identify the optimal number of clusters.
* The hierarchical clustering results were compared with K-Means to understand the differences in cluster formation.

**4. DBSCAN Clustering**

* **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** was applied as an alternative clustering technique:

Two main hyperparameters were configured:

* + - **eps (Epsilon)**: Defined the maximum distance between two points for them to be considered as neighbors.
    - **min\_samples**: Specified the minimum number of points required to form a dense region (cluster).
* The plots were compared with changing the hyper parameters chosen above individually for the model and the visuals were compared. The observation is the ideal value for eps and min\_samples was 0.5 and 10 respectively.
* The resulting DBSCAN clusters were visualized using a scatter plot, where outliers were clearly marked.
* The DBSCAN results were compared with K-Means and Hierarchical Clustering to assess their effectiveness in customer segmentation.

**RESULTS AND DISCUSSION:**

This section presents the results of the clustering methods used in this project, along with a comparative analysis of their performance.

**1.Data Visualizations**: To understand the distribution of customer data, the following visualizations were generated:

Figure 1: Distribution of Customer Ages. Figure 2: Distribution of Customer Annual Income. Figure 3: Distribution of Customer Spending Score.

A graph of age distribution

AI-generated content may be incorrect.A graph of a distribution of income

AI-generated content may be incorrect.A graph of a distribution of spending

AI-generated content may be incorrect.

**2. K-Means Clustering Results**

* The optimal number of clusters (K) was determined using the Elbow Method. The Elbow Method helps identify the point where adding more clusters does not significantly reduce the within-cluster variance.

Figure 4: Elbow Method Plot for K-Means Clustering. Figure 5: K-Means Clustering Scatter Plot.

A graph with a line and a point

AI-generated content may be incorrect.A graph with many colored dots

AI-generated content may be incorrect.

**3. Hierarchical Clustering Results**

* Dendrograms were generated using various linkage methods.

Figure 6: Dendrogram (Ward Linkage, Euclidean Distance). Figure 7: Dendrogram (Average Linkage, Manhattan Distance).

A diagram of a customer dendrogram

AI-generated content may be incorrect.A graph of a diagram

AI-generated content may be incorrect.

**4.DBSCAN Clustering Results**

* DBSCAN detected clusters of varying shapes and identified outliers.
* Figure 8: DBSCAN Scatter Plot with Clusters and Outliers. A chart of different colored dots

  AI-generated content may be incorrect. A chart of different colored dots

  AI-generated content may be incorrect.

**LIMITATIONS:**

* K-Means is sensitive to initial centroid selection and assumes spherical clusters.
* Hierarchical Clustering is computationally expensive for large datasets.
* DBSCAN is sensitive to the choice of eps and min\_samples and may not perform well on high-dimensional data.

**Justification of Methodology:**

* K-Means was selected for its efficiency in large datasets and ease of interpretation.
* Hierarchical Clustering was chosen to explore hierarchical relationships between customers.
* DBSCAN was included to identify non-linear clusters and detect outliers.

**Recommendations:**

* For larger datasets, consider using Mini-Batch K-Means for faster clustering.
* Explore advanced clustering methods such as Gaussian Mixture Models (GMM) for probabilistic clustering.
* Fine-tune DBSCAN hyperparameters use automated optimization techniques for better results.

**CONCLUSION:**

This project successfully implemented and compared three clustering techniques—K-Means, Hierarchical Clustering, and DBSCAN—for customer segmentation. K-Means provided well-separated clusters with the highest silhouette score, making it the most effective method for this dataset. Hierarchical Clustering offered a detailed hierarchical view of customer relationships, but its computational complexity limited its scalability. DBSCAN effectively identified non-linear clusters and detected outliers, making it valuable for datasets with varied cluster shapes.

**Key Findings**

* K-Means clustering achieved the highest silhouette score, indicating well-defined clusters.
* Hierarchical Clustering provided a detailed hierarchical structure but struggled with larger datasets.
* DBSCAN detected clusters of varying shapes and identified outliers but was sensitive to hyperparameter selection (eps and min\_samples).

**FUTURE WORK**

* Implement advanced clustering methods such as Gaussian Mixture Models (GMM) or Deep Embedded Clustering (DEC) for more accurate segmentation.
* Explore hybrid clustering techniques that combine K-Means, Hierarchical, and DBSCAN for improved performance.
* Fine-tune DBSCAN hyperparameters using automated optimization methods.
* Integrate additional customer features (e.g., age, gender, purchase history) to enhance segmentation quality.