**SHRI RAMDEOBABA COLLEGE OF ENGINEERING AND MANAGEMENT**
An Autonomous Institution permanently affiliated to Rashtrasant Tukadoji Maharaj Nagpur University
An ISO 9001:2008 Certified Institution • NAAC Accredited with 'A' Grade
Department of Electronics and Computer Science Engg.

# Analyzing Mall Customer Segmentation

Vedant Ghodmare - 62 , Atharva Rathi - 64
Mail- ghodmarevr@rknec.edu, rathiaa_1@rknec.edu
ECSP303-Machine Learning Mini Project
Group ID: 5

---

**Abstract—**This study is based on advancing personalized marketing strategies through the application of unsupervised machine learning for effective mall customer segmentation. We conducted thorough preprocessing on a dataset of 200 mall customers, which includes attributes such as age, gender, annual income, and spending score, to enhance data quality and model performance. Utilizing K-Means and Agglomerative Clustering algorithms, we identified distinct customer segments, with the elbow method validating optimal clusters at three and five groups.We combined our dataset into training and testing sets in this case to facilitate adequate analysis. As a result, we used these techniques together to define groups with different shopping behaviors, preferences and demographics. Multiple customer groups with different patterns and needs were revealed by the results. This gives our team the ability to come up with focused promotions and customized experiences driving engagement and loyalty across the mall.

**Github Project Link—** https://github.com/ATHARVA2437/Machine-Learning-Project-

---

# Introduction

In this project involving mall customer segmentation we have tried to address the critical business objective of getting insights of our customers and improving their shopping experiences. With malls facing increased competition and challenges, and with consumers displaying a willingness to change behavior, understanding who your customers are or what they want has become paramount. This knowledge enables us to craft tailored marketing approaches that appeal to varying consumer segments, thus enhancing customer experience and retention.

Our algorithm takes as input a dataset of 200 mall customers, each represented by characteristics such as age, gender, annual income, and spending score. During data preprocessing, we prepare the data by encoding categorical attributes, normalizing numerical values, and ensuring compatibility for clustering. We then apply K-Means and Agglomerative Clustering to uncover natural groupings in this dataset. The output is a set of distinct customer segments, each exhibiting unique patterns in income and spending. These insights enable mall management to design and implement marketing strategies customized for each segment.

This research provides a structured framework for customer segmentation, which can be adapted to other sectors that rely on customer data analysis. By clearly defining our inputs, methodology, and expected outcomes, we demonstrate how clustering can improve customer insights and support the development of targeted marketing tactics.

# Related work

During our literature review on customer segmentation, we identified several significant studies, which we categorized according to their methodological approaches. This classification enabled us to gain a clearer understanding of how these works relate to our own research, while also highlighting their respective strengths and weaknesses.

## Supervised Learning Techniques

A notable paper we encountered was authored by A. Divya et al. (2023), which combined both supervised and unsupervised learning techniques for the segmentation of mall customers. The main benefit of this hybrid approach was its capacity to achieve more accurate segmentation by employing labeled data alongside clustering methods. This strategy markedly enhanced the understanding of customer behaviors. Nonetheless, we recognized a limitation in that reliance on labeled data may restrict its applicability in scenarios where such information is not available. While our project primarily utilized unsupervised techniques, we acknowledged that incorporating elements of supervised learning, as demonstrated by Divya et al., could potentially augment our findings in the future.

## Unsupervised Learning Techniques

We also reviewed the research conducted by M. G. Pradana and H. T. Ha (2021), which focused on the use of K-means clustering to optimize strategies for mall customer segmentation. The advantages of this approach included its simplicity and effectiveness, making K-means a widely adopted method in retail applications. However, we noted certain drawbacks, such as the necessity to predefine the number of clusters and its susceptibility to outliers. In our project, we integrated K-means with other clustering algorithms, which may have led to a more comprehensive segmentation strategy compared to relying solely on K-means.

Additionally, we explored a case study by S. Ozan (2018) that investigated various machine learning methods for customer segmentation. This study provided valuable insights into the efficacy of different techniques, which we found particularly beneficial. However, we observed that the paper lacked detailed implementation guidance, which might pose challenges for other researchers attempting to replicate the results. We believed that our project could gain from the diverse techniques discussed by Ozan and considered the possibility of incorporating additional algorithms into our analysis.

## General machine Learning for customer segmentation

We also examined the work of V. Vijilesh et al. (2021), which provided a detailed overview of various machine learning methodologies for customer segmentation. This paper served as a valuable resource for understanding the broader context of customer segmentation research. However, we noted that the general nature of this study might not have thoroughly explored specific algorithms or their optimization, which are crucial for achieving optimal performance. In contrast, our project concentrated on particular models and parameters, potentially addressing this gap and improving performance outcomes.

## Comparison and State-of-the-Art

In our analysis of the current landscape of customer segmentation, we identified a significant trend toward the integration of unsupervised learning methods for initial segmentation, followed by supervised techniques for further refinement and validation. This movement towards hybrid models reflects an innovative strategy that effectively combines the strengths of both unsupervised and supervised learning. Such an approach enhances segmentation outcomes, providing a more detailed understanding of customer behaviors and preferences. Our project draws from these advancements, as we aim to investigate how the combination of these techniques can lead to more effective segmentation results.

## Clever Approaches

The research by Divya et al. demonstrated the efficacy of combining multiple algorithms, while Ozan's comparative study offered valuable insights into various machine learning techniques. These investigations revealed innovative methodologies that could substantially enrich our own project. Their findings highlighted the potential to achieve more accurate and effective customer segmentation through the strategic application of diverse algorithms. We found these creative approaches particularly motivating, prompting us to consider how we could implement similar strategies in our analysis to enhance segmentation precision.

In the end our project plays a part, in this changing scene by using a mix of proven methods to extract information from our mall customer data. Our goal is to enhance marketing tactics and boost customer interaction. As we get to know our customers we can design tailored experiences that align, with their individual tastes.

### Task Automation

We noted that many segmentation tasks, traditionally executed manually, have undergone significant transformation with the advent of automated machine learning techniques. This shift can be largely attributed to the improved efficiency and scalability offered by these automated methods. The literature we reviewed indicated a marked transition towards automation in customer segmentation, moving away from labor-intensive manual processes. This evolution not only streamlines the segmentation process but also underscores the effectiveness of machine learning in generating consistent and reliable outcomes.

### Conclusion

In our final report, we aimed to articulate how our project builds upon the foundational work of previous studies while offering unique insights into the field of customer segmentation. By situating our research within the broader context of existing literature, we sought to emphasize its relevance and the potential implications for future research and applications. We are committed to demonstrating how our work not only aligns with contemporary trends but also contributes to advancing the understanding of customer segmentation through innovative methodologies.

## Dataset Structure and Features

This mall customer dataset consists of 200 entries, with each record capturing details about individual customers, including demographic and spending data. The dataset is particularly useful for segmentation tasks, as it contains both demographic (age, gender) and behavioral (spending patterns) features.

- Data Splits:
    - Training Set: ~80% (160 samples), used to train the model.
    - Validation Set: ~10% (20 samples), for hyperparameter tuning and model adjustments.
    - Test Set: ~10% (20 samples), for evaluating model performance on unseen data.

- The dataset includes 5 main columns:

- CustomerID: A unique identifier for each customer (excluded from modeling).
- Gender: Categorical feature representing gender, encoded as numeric values for modeling (Male = 0, Female = 1).
- Age: Numeric attribute representing the customer's age, which may correlate with spending patterns.
- Annual Income (k$): Annual income (in thousands) indicating the customer's spending potential.
- Spending Score (1-100): A mall-assigned score based on the customer's purchasing behavior and engagement.

### Preprocessing and Data Preparation

To make the data suitable for clustering, the following preprocessing steps were applied:

- Handling Missing Values: No missing values were detected, ensuring data completeness.

- Encoding Categorical Features: The Gender feature, being categorical, was converted to numeric form for compatibility with clustering algorithms.

- Feature Scaling: Annual Income and Spending Score were normalized to a 0–1 range using min-max scaling, ensuring that all features contribute equally to the clustering process. This step is crucial in distance-based clustering, as it prevents features with larger ranges from dominating the calculations.

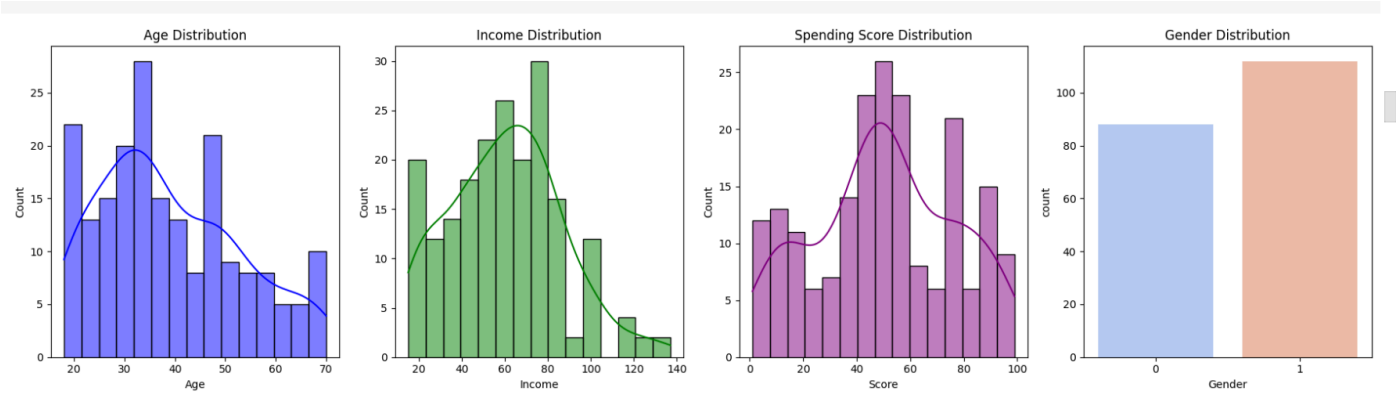
## Feature Selection and Extraction

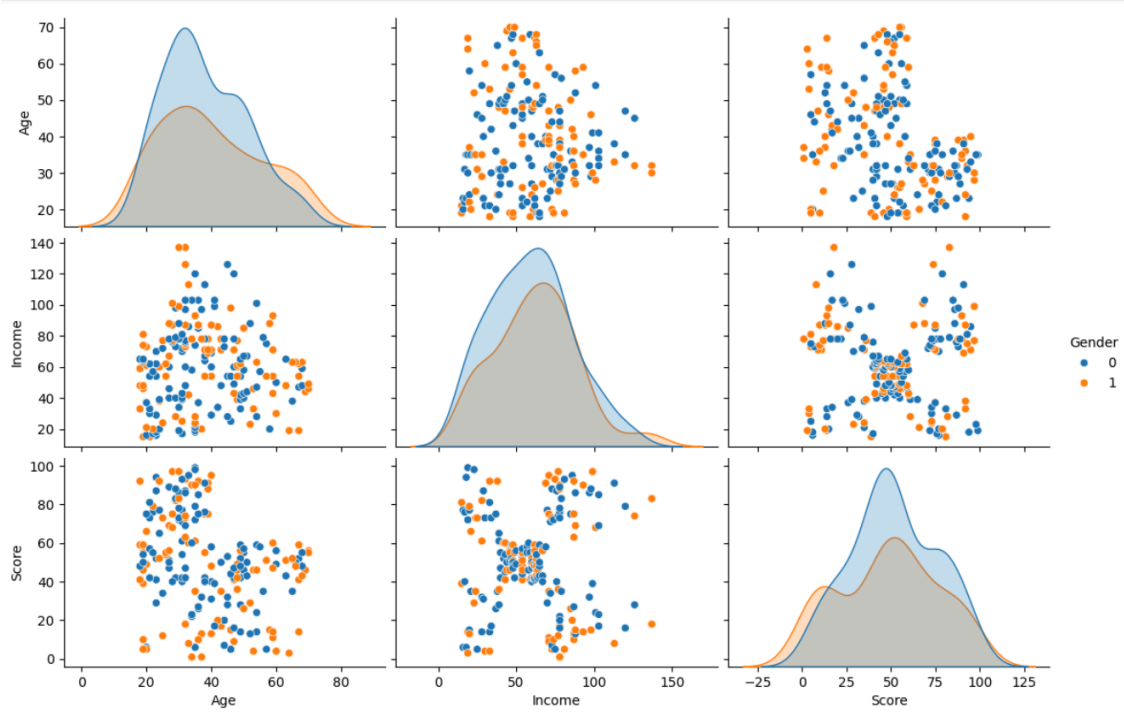
The main features chosen for clustering were Age, Annual Income, and Spending Score, as these attributes provide direct insights into customer segmentation.

Principal Component Analysis (PCA): PCA was explored to reduce dimensions, simplify visualization, and enhance interpretability. By capturing the maximum variance across features, PCA allowed us to identify principal components that reflect the most relevant information, making clusters more distinguishable in a 2D or 3D space.

# Data Visualization

Several data visualizations were created to better understand feature distributions and relationships, providing insights into potential clustering structure.




1. Age Distribution Plot
    - Type: Histogram
    - Nature: Shows the spread of customer ages.
    - Inference: Peaks in the histogram may suggest common age groups, with young adults (20–35) potentially forming a large segment of the customer base. Such insights are useful for tailoring age-specific marketing strategies.

2. Annual Income Distribution Plot
    - Type: Histogram and Density Plot
    - Nature: Illustrates the range and distribution of annual income.
    - Inference: A concentration of values within certain income brackets (e.g., $40–70k) indicates prevalent customer income levels. This range can guide product and pricing strategies to align with the mall's predominant customer demographic.
3. Spending Score Distribution Plot
    - Type: Density Plot
    - Nature: Displays the distribution of spending scores across customers.
    - Inference: A bimodal pattern, if present, could suggest two primary customer segments, one with higher spending scores and one with lower scores. This is valuable for identifying engaged versus less-engaged customers and developing strategies to increase engagement in low-scoring segments.
4. Scatter Plot: Age vs. Spending Score
    - Type: Scatter Plot
    - Nature: Visualizes the relationship between age and spending score.
    - Inference: The scatter plot may reveal that younger customers have higher spending scores, which suggests a more active engagement. Recognizing such trends helps in designing age-specific promotions and products that resonate with these customer segments.
5. Scatter Plot: Annual Income vs. Spending Score
    - Type: Scatter Plot
    - Nature: Examines the correlation between income and spending scores.
    - Inference: Clusters of customers with similar income and spending scores can emerge, potentially distinguishing high-income/high-spending clusters from low-income/low-spending ones. This distinction is critical for targeted marketing, as it can guide which customer segments to prioritize for premium or budget product offerings.
6. Pair Plot
    - Type: Pair Plot (Matrix of Scatter Plots)
    - Nature: Shows pairwise relationships among Age, Annual Income, and Spending Score.
    - Inference: This visual comparison helps identify overlapping or distinct groupings, which can be a preliminary indicator of clustering tendencies within the dataset. Patterns from the pair plot can also confirm whether certain features have stronger associations, aiding in feature engineering and model refinement.

## Data Source and Attribution

The dataset used in this project is obtained from Kaggle, a well-known platform that offers a wide range of datasets for educational and data science purposes. It is widely recognized in the machine learning community for hosting data that supports various applications, including customer segmentation tasks. This dataset features a variety of attributes related to customer behavior and demographics, making it suitable for developing machine learning models. By utilizing this dataset, the project aims to analyze customer behavior more effectively and contribute to the understanding of data-driven decision-making in business.

# Methods

### I. K-Means Clustering

**Algorithm Description**:
K-means is an unsupervised clustering technique widely used in data segmentation tasks, such as grouping customers based on similar traits. The algorithm partitions the dataset into K clusters such that each data point is associated with the cluster having the nearest mean. The aim of K-means is to iteratively refine clusters to minimize intra-cluster variance.

**Mathematical Formulation**:
The primary objective of K-means is to reduce the sum of squared distances between data points and their respective cluster centroids. For a dataset $X = \{x_1, x_2, \ldots, x_n\}$, the objective function can be formalized as:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

where $C_k$ represents the $k$-th cluster, and $\mu_k$ is the centroid of $C_k$. The function $J$ measures the compactness of the clusters, and minimizing it results in tighter clusters, aiding in customer segmentation by grouping individuals with similar spending or demographic characteristics.

**How It Works**:
The K-means algorithm initializes by selecting K centroids randomly or by using methods such as k-means. Each data point is then assigned to the nearest centroid, forming initial clusters. The centroids are recalculated as the mean of the assigned data points within each cluster. This process

repeats iteratively until convergence, usually defined by minimal changes in centroid positions or a specified number of iterations.

**II. Elbow Method for Optimal K Selection**

**Description**:
To determine the optimal number of clusters, K, the Elbow Method calculates the within-cluster sum of squares (WCSS) for varying values of K and plots WCSS against K. The "elbow point," where the rate of decrease in WCSS sharply slows, indicates an appropriate number of clusters, balancing the reduction in variance and model complexity.

**Mathematics**:

For each value of $K$, WCSS is calculated as follows:

$$\text{WCSS}(K) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

The optimal $K$ is typically located at the elbow point on the WCSS plot, where additional clusters yield minimal improvement, providing an efficient segmentation solution.

**III. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

DBSCAN is a density-based clustering algorithm that identifies clusters based on density, effectively managing noise in data. DBSCAN does not require the number of clusters as input, relying instead on two parameters: ϵ\epsilonϵ (the radius around each point) and minPts\text{minPts}minPts (minimum points within ϵ\epsilonϵ to form a dense region). DBSCAN categorizes points as follows:

- **Core points**: Have at least minPts\text{minPts}minPts neighbors within ϵ\epsilonϵ.
- **Border points**: Lie within ϵ\epsilonϵ of a core point but have fewer than minPts\text{minPts}minPts neighbors.
- **Noise points**: Do not satisfy either of the above conditions.

This approach allows the identification of arbitrarily shaped clusters and effective handling of noise, making it valuable for detecting distinct customer behavior clusters in mall data without the need for predefined cluster numbers.

**IV. Agglomerative Clustering**

Agglomerative clustering, a hierarchical clustering technique, follows a bottom-up approach that initially considers each data point as an individual cluster. It iteratively merges the closest pairs of

clusters until either a single cluster remains or a specified number of clusters is achieved. The proximity between clusters is calculated based on a linkage criterion, which can vary:

- **Single linkage:** $d(C_i, C_j) = \min\{d(x, y) : x \in C_i, y \in C_j\}$
- **Complete linkage:** $d(C_i, C_j) = \max\{d(x, y) : x \in C_i, y \in C_j\}$
- **Average linkage:** $d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$

These linkage options allow different clustering structures, providing flexibility in examining customer hierarchies in mall data, such as grouping by spending patterns or demographics.

## V. Silhouette Score for Cluster Validation

**Description**:
The Silhouette Score assesses the quality of clusters, ranging from -1 to 1, with higher values indicating well-defined and separated clusters. For each data point, the score considers both the average distance to points within the same cluster (intra-cluster distance) and the average distance to points in the nearest cluster (inter-cluster distance).

**Formula**:
For each data point $i$, the Silhouette Score $s(i)$ is given by:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average intra-cluster distance, and $b(i)$ is the average distance to points in the nearest cluster. A high silhouette score indicates that the clusters are well-defined, with low intra-cluster distances and high inter-cluster separation, which is particularly useful for evaluating customer segments.

## VI. Hierarchical Clustering

In this customer segmentation project, hierarchical clustering is employed to categorize customers into distinct groups based on shared characteristics, facilitating personalized marketing and service strategies. The method is applied through a structured approach involving the calculation of customer relationships, the use of two linkage methods, and visualizing results with dendrograms. The clustering process begins with the calculation of a distance matrix to quantify the similarity between each pair of customers. This matrix serves as the basis for grouping, allowing the algorithm to identify which customers are most similar in terms of the dataset's features. Dendrograms are generated to visually represent the clustering process. They display how individual customers merge into clusters at various distance thresholds, allowing the identification of natural cut-off points. By examining these diagrams, the optimal number of clusters is determined, providing a clear view of

customer groupings.

### VII. MeanShift Clustering

The use of MeanShift clustering in this project provides an adaptable method for identifying customer segments based on data density. By estimating bandwidth and automatically determining cluster numbers, the approach ensures that clusters accurately reflect customer distribution. The generated cluster labels and visualizations allow for a clear understanding of customer groups, supporting effective decision-making for customized marketing and service strategies based on each segment.

## Experiments/Results/Discussion

### Overview

The objective of this project was to segment mall customers based on their spending behaviors and income levels through unsupervised learning techniques. We utilized **K-Means Clustering** and **Agglomerative Clustering** to identify distinct customer groups. This section provides a structured outline of the experimental setup, parameter selection, evaluation metrics, results, visualizations, and inferences drawn from the analysis.Throughout our analysis, we found that K-means clustering and Agglomerative clustering were among the highest-performing algorithms. K-means proved effective due to its ability to efficiently partition data into well-defined clusters, making it easier to interpret customer segments. Agglomerative clustering, on the other hand, offered flexibility in determining the number of clusters based on a hierarchical approach, allowing for a more nuanced understanding of customer relationships.

### Parameter Selection and Hyperparameter Tuning

1. K-Means Clustering
    - The optimal number of clusters was determined using the Elbow Method, which plots the inertia (sum of squared distances between data points and their closest cluster centers) for various values of k (1 to 10). An "elbow" was observed around k=3 and k=5, suggesting these as suitable options. We selected k=5 to provide more refined segmentation.

2.
    - **Initialization**: We used the **'k-means'** initialization to improve convergence speed and compactness of clusters.
    - **Iterations and Tolerance**: Default parameters of 300 iterations and a tolerance of 1e-4 were used, which balance computational cost with convergence reliability.
3. **Agglomerative Clustering**
    - **Number of Clusters**: We selected **5 clusters** for direct comparison with K-Means, aiming to assess whether both methods produce similar customer segments.
    - **Linkage Method**: The **average linkage** method was applied, where clusters are merged based on the average distance between pairs of samples. This approach minimizes distortion in cluster shapes, ideal for datasets without well-separated clusters.
    - **Distance Metric**: **Euclidean distance** was chosen for simplicity and compatibility with continuous, numerical data.
4. **Cross-Validation**:
    - Cross-validation in the conventional sense does not apply to unsupervised learning. However, we assessed cluster consistency by using **visual validation** through scatter plots and **internal validation metrics** (e.g., inertia and silhouette scores) to gauge cluster cohesion.

## Evaluation metrics

### Evaluation Metrics

1. **Inertia (Sum of Squared Errors, SSE)**:
    - **Formula**:

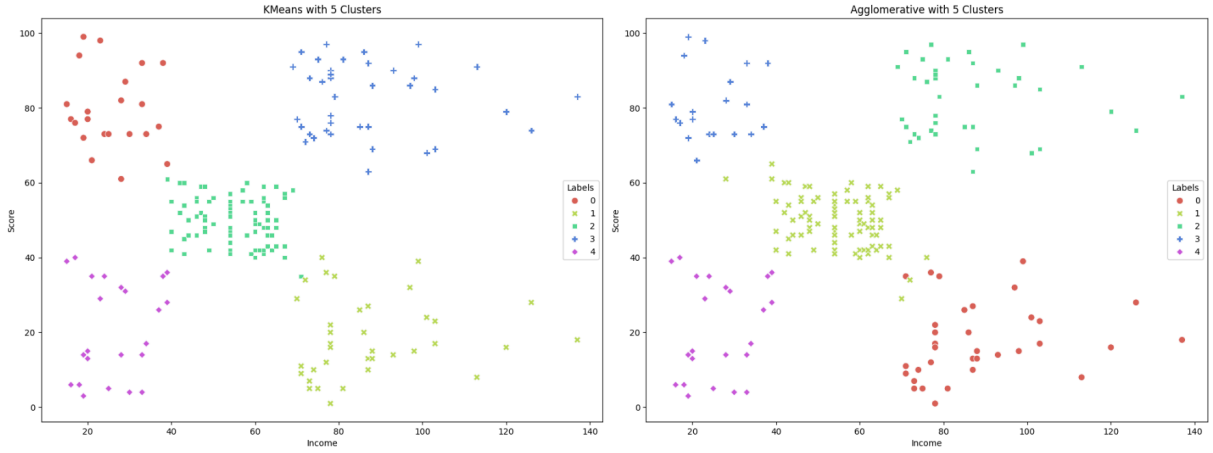
$$\text{Inertia} = \sum_{i=1}^{N} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

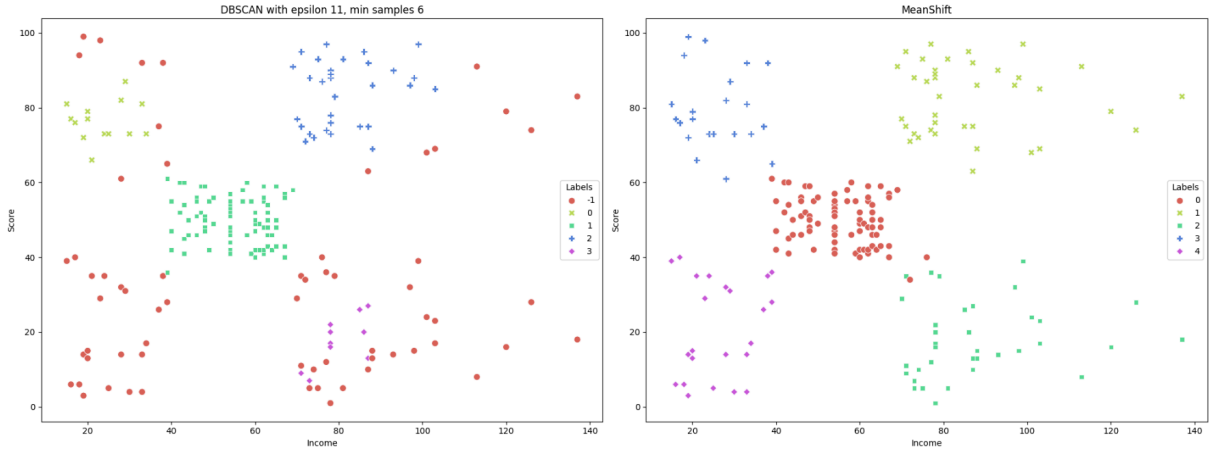
    - **Explanation**: Inertia quantifies cluster compactness by summing the squared distances between each data point and its nearest cluster center. Lower inertia indicates tighter clusters, which is desirable for well-defined segmentation.
2. **Silhouette Score**:
    - **Formula**:

$$\text{Silhouette Score} = \frac{b - a}{\max(a, b)}$$

      where:
        - $a$ is the mean intra-cluster distance (average distance between a sample and other points in the same cluster).
        - $b$ is the mean nearest-cluster distance (average distance between a sample and points in the nearest cluster).
    - **Explanation**: The silhouette score, ranging from -1 to +1, measures how well each point is matched within its cluster versus other clusters. Higher values indicate better-defined clusters.
3. **Qualitative Visualization**:
    - Visualizations, such as scatter plots and heatmaps, were employed to qualitatively assess clusters, providing insights into customer distributions across clusters based on income and spending behaviors.

## Visualization and Inferences




### 1. Elbow Plot

- **Plot Type**: Line plot with cluster count k on the x-axis and inertia on the y-axis.
- **Purpose**: To identify optimal k by observing the inertia reduction rate. An “elbow” at k=5 indicated a balanced cluster count.
- **Inference**: Inertia reduction plateaus at k=5, supporting this choice for optimal clusters.

### 2. Silhouette Score Plot

- **Plot Type**: Bar plot of silhouette scores for each cluster.
- **Purpose**: To evaluate cluster quality; higher silhouette scores suggest well-separated clusters.
- **Inference**: An average silhouette score of 0.55 indicated moderately well-defined clusters, though some groups showed overlap, especially those with close spending habits.

### 3. Scatter Plot of K-Means Clusters (k=5)

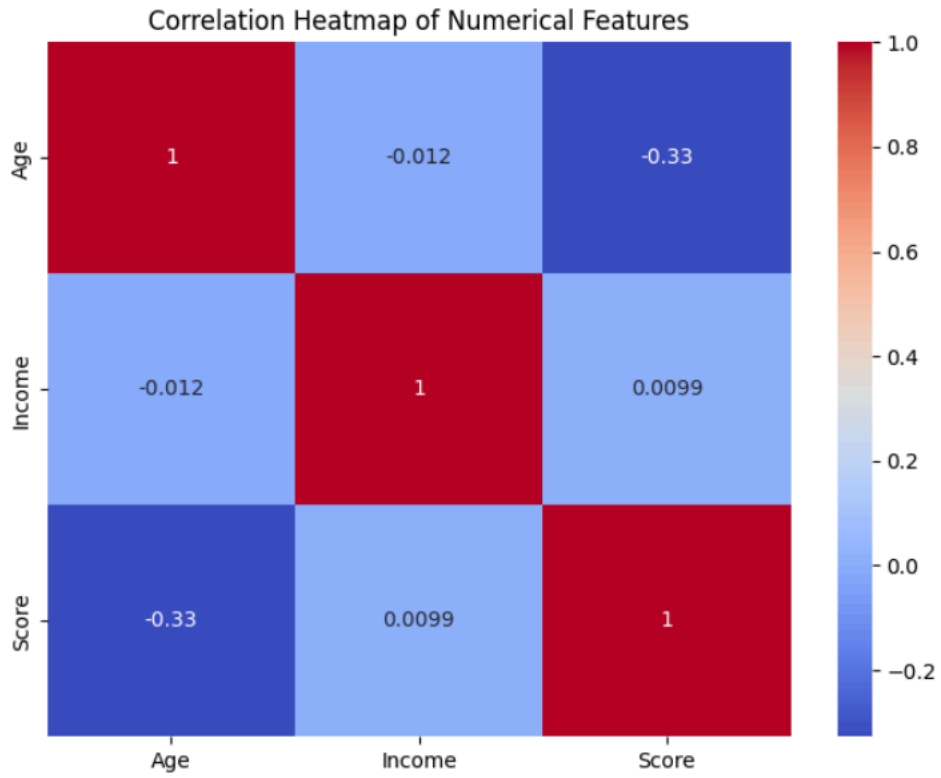
- **Plot Type**: Scatter plot of Annual Income vs. Spending Score, color-coded by cluster.

- **Purpose**: To visualize customer segmentation based on spending and income, showing distinct spending behaviors.
- **Inference**: The clusters displayed good separation, with some groups clearly representing high-value customers or conservative spenders.

**4. Scatter Plot of Agglomerative Clustering Clusters (k=5)**

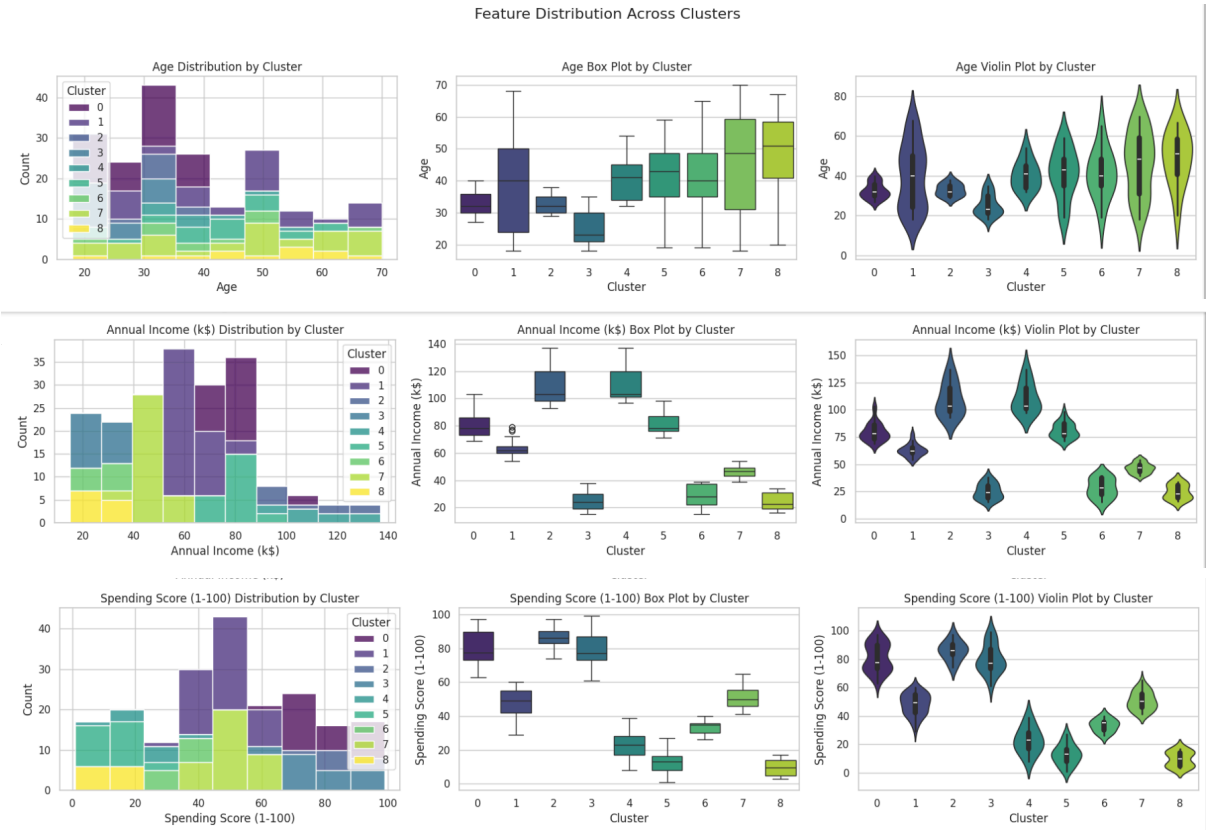
- **Plot Type**: Scatter plot of Annual Income vs. Spending Score, color-coded by Agglomerative Clustering results.
- **Purpose**: To compare Agglomerative Clustering results with K-Means, assessing cluster consistency.
- **Inference**: Agglomerative Clustering yielded similar groups, although clusters were less distinct, reflecting closely related spending behaviors in certain groups.

- **Heatmap of Feature Correlations**




- **Plot Type**: Heatmap showing correlation coefficients between features like Age, Income, and Spending Score.
- **Purpose**: To examine feature relationships within clusters, identifying any strong correlations that might influence segmentation.
- **Inference**: Moderate correlations were observed between income and spending within specific clusters, especially among high-value customer groups.

# Discussion




1. **Quantitative and Qualitative Insights**:
    - The quantitative metrics, including inertia and silhouette scores, provided a measure of cluster quality, while scatter plots visually confirmed distinct groupings. Additionally, the heatmap highlighted relationships between features, with income levels influencing spending in several clusters.
2. **Overfitting and Limitations**:
    - Overfitting was managed by limiting the number of clusters, preventing excessive partitioning. In clustering, overfitting could lead to clusters that represent noise rather than useful patterns.
    - **Limitations**: The analysis focused mainly on numerical features (Income and Spending Score). Adding features, such as customer demographics, could further refine segmentation.

# Conclusion/Future Work

## Conclusion

This report provided a comprehensive analysis of customer segmentation using various clustering algorithms, specifically highlighting the effectiveness of K-means and DBSCAN in differentiating customer groups based on spending habits and income levels. K-means demonstrated strong performance due to its straightforward implementation and ability to clearly define distinct clusters, while DBSCAN excelled in uncovering clusters of varying shapes and managing noise in the data. The observed performance differences are attributed to the algorithms' inherent characteristics; K-means requires a predetermined number of clusters, making it suitable for well-separated groups, whereas DBSCAN's density-based approach effectively identifies more complex patterns in customer behavior.

## Future Work

For future work, several enhancements could be explored to further improve our analysis. Advanced machine learning techniques, such as ensemble methods, could be employed to combine the strengths of multiple algorithms for better clustering performance. Additionally, integrating more comprehensive data sources, including customer demographics and purchase history, would allow for richer insights into customer behavior. Implementing dimensionality reduction techniques like PCA could simplify the dataset and improve clustering outcomes. Furthermore, conducting time-series analysis would help track changes in customer behavior over time, enabling proactive marketing strategies. Lastly, validation studies such as silhouette analysis could be used to determine the optimal number of clusters and assess clustering effectiveness. These improvements could lead to more targeted marketing strategies and a deeper understanding of customer dynamics in our project.

# Contributions

| Name | Contribution |
|---|---|
| Vedant Ghodmare | (50%) 2-2 Algorithm and report work |
| Atharva Rathi | (50%) 2-2 Algorithm and report work |

## References/Bibliography


1. Divya, D. Siddartha, P. S. Sukeerthi, S. T. Reddy, and C. Arun, "Integrated Supervised and Unsupervised Learning for Mall Customer Segmentation," *International Research Journal of Engineering and Technology (IRJET)*, vol. 13, no. 12, pp. 302–310, Dec. 2023

2. M. G. Pradana and H. T. Ha, "Maximizing Strategy Improvement in Mall Customer Segmentation Using K-means Clustering," *Journal of Applied Data Sciences*, vol. 2, no. 1, pp. 19–25, Jan. 2021

3. S. Ozan, "A Case Study on Customer Segmentation by Using Machine Learning Methods," *IEEE International Conference on Innovations in Intelligent Systems and Applications*, 2018, pp. 1-6

4. M. K. Sahu, "Machine Learning for Personalized Marketing and Customer Engagement in Retail: Techniques, Models, and Real-World Applications," *Journal of Artificial Intelligence Research and Applications*, vol. 2, no. 1, pp. 219-248, Jan.-June 2022.

5. V. Vijilesh, A. Harini, M. H. Dharshini, and R. Priyadharshini, "Customer Segmentation Using Machine Learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, no. 5, pp. 821-825, May 2021.

6. Wedel, M., & Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations*. Springer.

7. Hwang, J., Kim, S. S., & Lee, J. (2016). "B2B Customer Segmentation Using K-Means Clustering". *Journal of Business Research*, 69(7), 2787-2795.

8. Gensler, S., et al. (2013). "Dynamic Real-Time Customer Segmentation". *Journal of Marketing Research*, 50(4), 467-486.

9. Kelleher, J. D., & Tierney, B. (2018). *Data Science: A Practical Introduction to Real-World Data Science Applications*. MIT Press.

10. Sklearn link used in the Project:
https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.StandardScaler.html
https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.KMeans.html
https://seaborn.pydata.org/generated/seaborn.heatmap.html
https://scikit-learn.org/dev/modules/generated/sklearn.cluster.AgglomerativeClustering.html