# Data-Driven Customer Segmentation: Insights from Demographics and Spending Patterns

## Introduction

Traditional market segmentation relied mainly on demographic factors such as age, gender, and income. However, the growing availability of data and advancements in analytical techniques have allowed businesses to segment customers more sophisticatedly (Wedel & Kamakura, 2012). The emergence of data-driven approaches, particularly clustering techniques like K-means and hierarchical clustering, has enabled companies to uncover hidden patterns within large datasets. These insights help businesses develop highly targeted marketing strategies that enhance customer engagement, increase loyalty, and drive sales. Typically, the features utilized for segmentation include demographic data (e.g., age, gender, income) and behavioral attributes (e.g., spending score). Both of these elements are essential for understanding purchasing behavior and customer preferences (Chawla et al., 2017). Given the complexity and diversity of customer behaviors, more advanced techniques are required to analyze data and extract meaningful insights. In this context, data science plays a vital role in customer segmentation by integrating data preprocessing, feature engineering, and machine learning algorithms. As businesses increasingly depend on data-driven decision-making, customer segmentation has evolved to be more dynamic and responsive to changes in market trends and consumer behavior (McKinsey & Company, 2020).

Despite the extensive use of customer segmentation, many businesses continue to face challenges in implementing effective strategies due to difficulties in identifying meaningful customer groups. Traditional segmentation methods frequently overlook the complexities of customer behavior and fail to account for the myriad factors influencing purchasing decisions (Hosseini & Mollah, 2020). Furthermore, modern consumers' sheer volume of data can overwhelm businesses if not analyzed appropriately. In numerous instances, companies do not segment their customers in a manner that accurately reflects their true preferences or needs. This oversight can result in inefficiencies in marketing strategies, resource allocation, and product offerings. For example, a business might deploy a single promotional campaign across its entire customer base, disregarding variations in income levels, purchasing power, or spending habits. Consequently, the campaign may fail to resonate with all customers, leading to poor returns on investment and missed opportunities for personalized engagement (Vasquez et al., 2021).

Literature Review

Early segmentation primarily relied on fundamental demographic data, such as age, gender, and income. However, this approach often overlooked the complexities inherent in consumer behavior (Kotler & Keller, 2016). With the advancement of data collection techniques, businesses began to focus more on psychographics, behaviors, and preferences, leading to a more nuanced understanding of consumer needs. In the 20th century, a pivotal development in customer segmentation was the introduction of the RFM (Recency, Frequency, and Monetary) model. This model categorizes customers based on three essential metrics: how recently they made a purchase, how often they shop, and how much they spend (Chen et al., 2018). Although RFM continues to be a widely utilized framework, it has notable limitations, particularly when managing large datasets or accounting for more intricate consumer behaviors. By being static, the RFM model emphasizes historical behaviors and does not effectively capture the progressive nature of customer preferences over time (Bonsignore et al., 2019). The rise of machine learning and big data analytics has significantly enhanced customer segmentation, fostering the development of more dynamic and adaptive methodologies in recent years. One prominent technique is cluster analysis, which organizes customers into groups based on their similarities within a multidimensional feature space. Cluster analysis has become a vital tool in customer segmentation, as it uncovers hidden patterns and relationships within the data that traditional methods might overlook (Vasquez et al., 2021). The increased popularity of clustering techniques such as k-means, DBSCAN, and hierarchical clustering can be attributed to their capacity to automatically identify meaningful groups without relying on predefined labels. Clustering techniques are particularly valuable in customer segmentation because they go beyond superficial demographic factors and enable businesses to consider a combination of features, such as purchasing behaviors, preferences, and interactions with brands. For instance, studies have shown that demographic data combined with transactional data, such as product preferences and purchase frequency, yields more accurate and actionable segments compared to using demographic data alone (Chawla et al., 2017). This is because demographic attributes only provide a high-level view of customers, while transactional and behavioral data reveal more insights into actual consumer preferences and buying habits.

Traditional methodologies often prioritize demographic data, frequently neglecting the underlying psychological and behavioral dimensions that influence purchasing decisions. Factors such as consumer preferences, motivations, and attitudes are pivotal in shaping buying behavior; however, many studies do not adequately integrate these elements (Chawla et al., 2017). The incorporation of psychographic data could yield more precise and actionable segments, thereby providing a deeper comprehension of customer needs. Moreover, although clustering techniques such as k-means and DBSCAN have been extensively implemented, there remains limited exploration of hybrid approaches that combine various algorithms or incorporate supervised learning to enhance segmentation accuracy. For instance, hybrid models that integrate clustering algorithms with classification techniques could capitalize on the strengths of both methods, enabling improved alignment between customer segments and business objectives (Xu & Wunsch, 2005). The scant rigorously tested hybrid models within the context of customer segmentation indicate a significant area for future research. Another pertinent gap in the literature is the dynamic nature of customer segments. Numerous segmentation studies treat customer segments as static and fixed; however, in practice, customer behavior evolves over time due to shifts in preferences, purchasing power, and external factors such as economic conditions or technological advancements. The temporal dimensions are often inadequately incorporated into segmentation models, limiting their effectiveness in long-term marketing strategies (Hosseini & Mollah, 2020). Future inquiries could investigate dynamic customer segmentation models that adapt and evolve as new data emerges, thereby enabling businesses to respond more adeptly to changes in customer behavior.

## ****Methodology****

### Data Collection

The dataset used in this project was sourced from Kaggle, a popular platform for data sharing and analysis, which contains 200 customer records across five attributes. These attributes provide critical insights into customer demographics and behavior. Below is an expanded description of the dataset, its relevance, and the steps taken to utilize it effectively for customer segmentation.

| **Feature Name** | **Description** | **Type** | **Value Range or Categories** |
| --- | --- | --- | --- |
| **CustomerID** | Unique identifier for each customer | Integer | N/A |
| **Gender** | Customer's gender | Categorical | Male, Female |
| **Age** | Age of the customer | Integer | [18, 70] |
| **Annual Income (k$)** | Annual income in thousands of dollars | Integer | [15, 137] |
| **Spending Score (1-100)** | Spending score indicating purchasing habits | Integer | [1, 100] |

This dataset provides a blend of categorical and numerical features, enabling a comprehensive analysis of customer segments.

### ****Data Preprocessing****

Data preprocessing is a crucial step to ensure data quality and to make it suitable for clustering algorithms. The dataset was examined for missing values, redundancies, and inconsistencies. As there were no missing values, no imputation was required. The preprocessing phase included four key steps: handling redundant columns, categorical encoding, scaling numerical data, and analyzing skewness and outliers.

### **Encoding Categorical Data**

The categorical Gender column was encoded into numerical values using binary encoding. Male was represented as 0, and Female as 1. The encoded data ensures compatibility with clustering algorithms.

| **Original Gender** | **Encoded Value** |
| --- | --- |
| Male | 0 |
| Female | 1 |

Since binary encoding creates only one column for the Gender feature, it eliminates the issue of dummy variable traps, ensuring no multicollinearity.

Numerical Scaling

First, the CustomerID column, which served as an identifier, was deemed redundant for the clustering process and removed, as it provided no meaningful information for segmentation. The remaining numerical columns (Age, Annual Income (k$), and Spending Score (1-100)) were normalized to standardize their scale, ensuring that variables with larger ranges did not disproportionately influence the clustering algorithms. The scaling process was particularly important because clustering algorithms like K-means are sensitive to scale. The standardization formula is expressed mathematically as:

Where:

* is the original value,
* is the mean of the feature, and
* is the standard deviation.

After standardization, the transformed dataset had a mean of 0 and a standard deviation of 1, ensuring that no variable dominated the clustering results.

**Outlier Analysis**

Next, skewness in numerical data was analyzed. The skewness values for Age (0.48), Annual Income (k$) (0.32), and Spending Score (1-100) (-0.05) were low, indicating that the data distribution was approximately symmetrical, and no transformation was necessary.

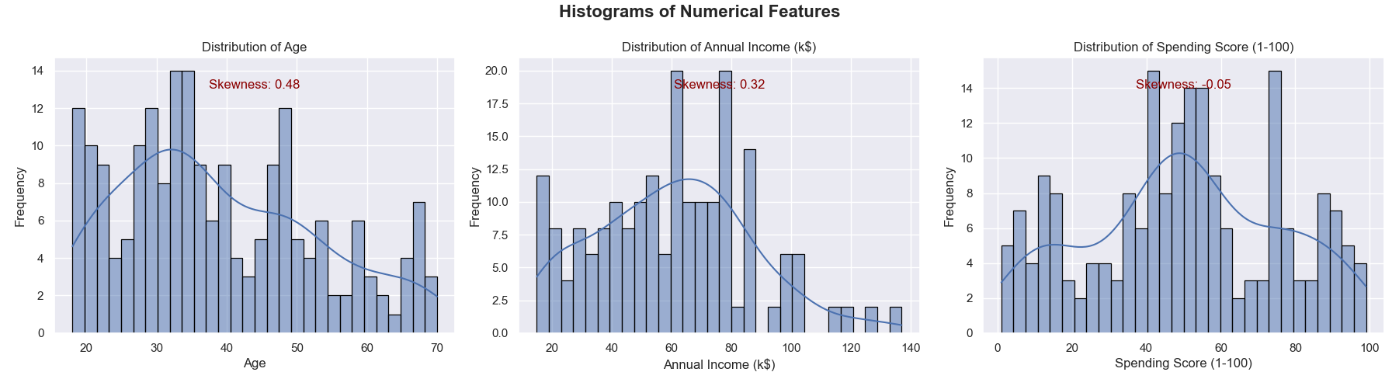


Figure 1: Distribution of the Numerical Features

Outlier detection was also conducted, revealing that Annual Income (k$) had two outliers. These outliers were retained after ensuring they represented genuine customer profiles, as their inclusion could provide meaningful insights during segmentation.

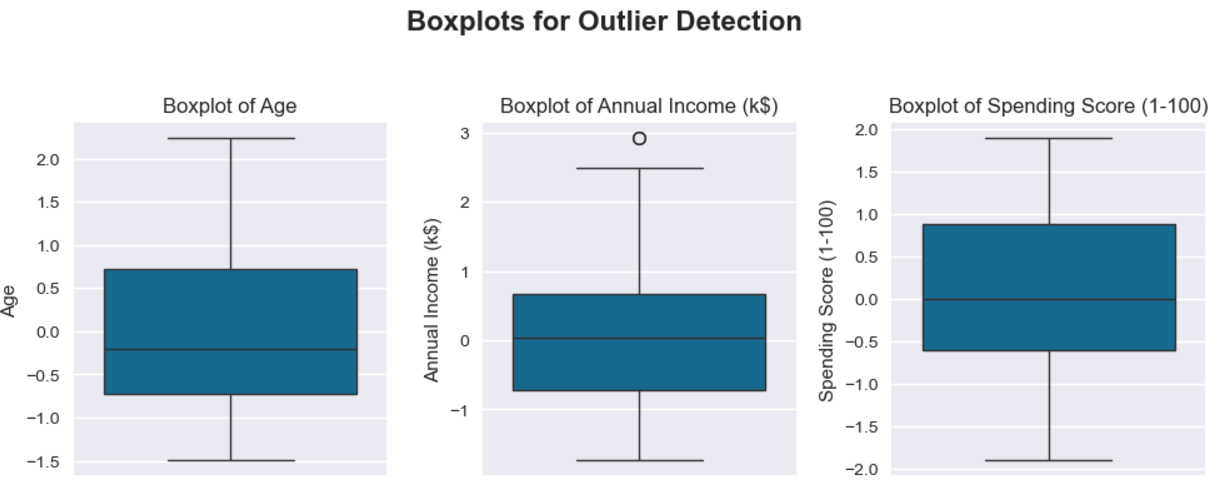


Figure 2: Outlier Detection

## ****Clustering Techniques****

To segment customers, multiple clustering algorithms were applied and evaluated to identify the most suitable approach. The clustering techniques included K-means, MiniBatch K-means, Hierarchical Clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The choice of these algorithms was motivated by their ability to group data effectively and their widespread use in customer segmentation research.

K-means Clustering

K-means was chosen for its simplicity and effectiveness in dividing data into distinct clusters. The Elbow Method was employed to determine the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) against the number of clusters. Additionally, silhouette scores and the Davies-Bouldin index were used to evaluate clustering performance. The final results indicated that five clusters provided a good balance between interpretability and cluster cohesion. K-means minimizes the within-cluster sum of squares (WCSS). Mathematically, the objective function is:

Where is the number of clusters, ​ is a data point, and ​ is the centroid of the cluster .

The Elbow Method was used to determine the optimal number of clusters. A plot of WCSS against the number of clusters revealed a kink at k=5.

MiniBatch K-means

MiniBatch K-means, a faster variant of K-means, was used to verify the consistency of the clustering results on the same dataset. This approach processes smaller random subsets of data, making it suitable for large-scale datasets or faster computations.

**Hierarchical Clustering**

Agglomerative hierarchical clustering was implemented to explore how customers form nested groupings. This algorithm builds clusters iteratively by merging the closest data points or clusters. Dendrograms were used to visualize the clustering process and determine an appropriate number of clusters. This method offered valuable insights into the hierarchical relationships among customer profiles. Agglomerative clustering was implemented, starting with each data point as its own cluster and successively merging the closest clusters. The linkage between clusters was based on Ward’s method, minimizing the variance within clusters.

DBSCAN  
DBSCAN was employed to identify clusters of arbitrary shape and to detect noise or outliers in the dataset. It uses density-based spatial clustering to group data points close to each other, ignoring points in low-density regions. However, the algorithm struggled to form meaningful clusters in this dataset due to its reliance on density parameters.

## ****Evaluation Metrics****

To assess the performance of clustering algorithms, three metrics were utilized:

**Silhouette Score**: The silhouette score measures the cohesion and separation of clusters. It is calculated as:

Where is the mean intra-cluster distance, and is the mean nearest-cluster distance. Scores range from -1 (poor clustering) to 1 (perfect clustering).

**Davies-Bouldin Index**: This index evaluates clustering quality by considering the ratio of within-cluster scatter to between-cluster separation. A lower index indicates better clustering.

**Elbow Method**: The WCSS plot revealed that adding more clusters beyond 5 provided diminishing returns, confirming k=5k = 5k=5 as the optimal number of clusters.

## Results and Discussion

### K-Means Clustering Analysis

The optimal number of clusters for the K-Means algorithm was determined by evaluating the Elbow method, Silhouette score, and Davies-Bouldin index. A comparative analysis of these metrics strongly supports the selection of **5 clusters** as the best option.

The Elbow method helps us understand how the Within-Cluster Sum of Squares (WCSS) changes as we increase the number of clusters. For example, with 4, 5, and 6 clusters, the WCSS values are 254.36, 217.29, and 207.18, respectively. The drop in WCSS from 4 to 5 clusters is significant, showing that the clusters are tighter and more cohesive. However, the small decrease in WCSS when going from 5 to 6 clusters suggests that adding more clusters can make the model more complex, but does not greatly improve cluster compactness.

The Silhouette score is another helpful measure for cluster quality. It looks at how well the clusters stick together and how distinct they are from each other. The Silhouette scores for 4, 5, and 6 clusters are 0.295, 0.355, and 0.318, respectively. The highest score at 5 clusters shows that this setup offers the best balance between each cluster being close together and being separate from other clusters. The slight drop in the Silhouette score at 6 clusters may mean that the clusters are starting to overlap, which could be worth exploring further.

The Davies-Bouldin index is another way to measure clustering quality. Lower scores indicate clusters that are more compact and well-separated. The index values for 4, 5, and 6 clusters are 1.069, 1.034, and 1.149, respectively. The best score again occurs at 5 clusters, emphasizing that this is a strong choice for clustering. The higher index at 6 clusters suggests some fragmentation, pointing to an area to consider for improving the clustering approach.

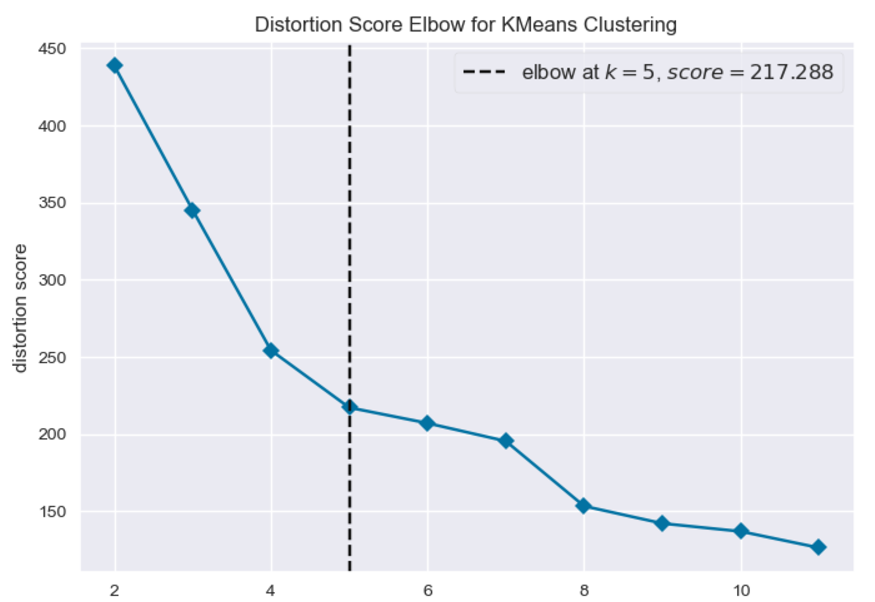
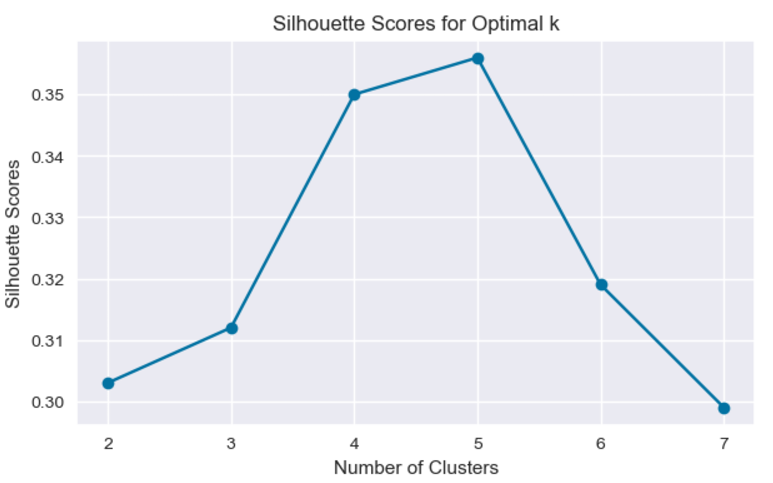
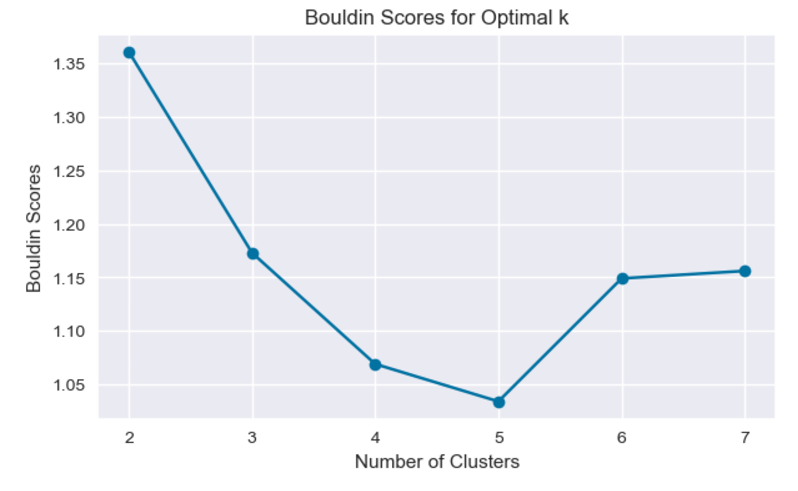


Figure 3: Elbow Curve Showing the Optimal Number of Clusters

The table below summarizes the evaluation metrics for clusters 4, 5, and 6:

| **Number of Clusters** | **WCSS** | **Silhouette Score** | **Davies-Bouldin Index** |
| --- | --- | --- | --- |
| 4 | 254.36 | 0.295 | 1.069 |
| 5 | 217.29 | 0.355 | 1.034 |
| 6 | 207.18 | 0.318 | 1.149 |

From the table, **5 clusters** are consistently superior across all metrics:

* The reduction in WCSS justifies tighter groupings without unnecessary complexity.
* The highest Silhouette score demonstrates optimal separation and compactness.
* The lowest Davies-Bouldin index confirms the best clustering quality.

### MiniBatch K-Means Clustering

| **Number of Clusters** | **Silhouette Score** | **Davies-Bouldin Index** |
| --- | --- | --- |
| 3 | 0.366 | 0.915 |
| 4 | 0.404 | 0.854 |
| 5 | 0.407 | 0.803 |
| 6 | 0.378 | 0.829 |

Both metrics suggest that **5 clusters** offer the best clustering solution. This conclusion aligns with the findings from the standard K-Means method, confirming that **5 clusters** is the optimal configuration for this dataset.

### Hierarchical Clustering Results

In the analysis of hierarchical clustering, we evaluate two key metrics: **Silhouette Score** and **Davies-Bouldin Index** for different numbers of clusters (k = 4, 5, and 6). These metrics help us assess the quality of the clustering solution based on the compactness and separation of the clusters.

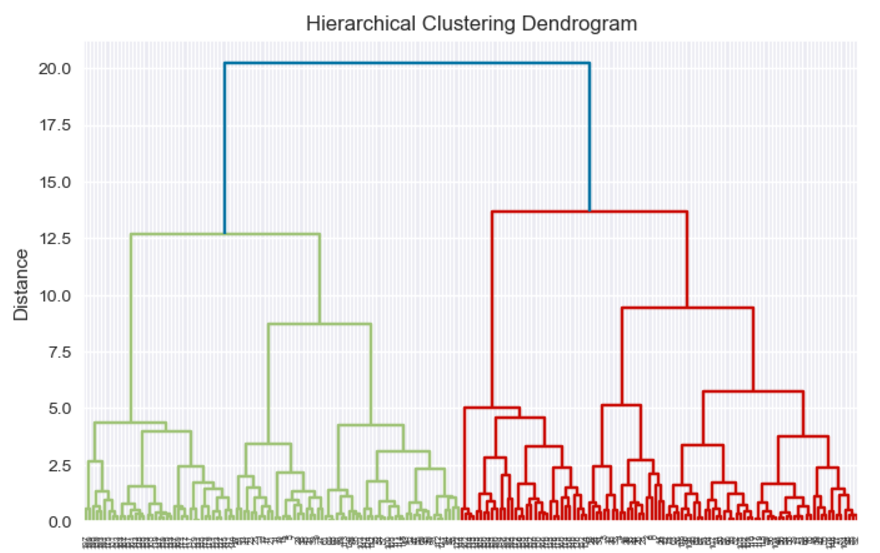


Figure 4: Dendogram Plot

| **Number of Clusters (k)** | **Silhouette Score** | **Davies-Bouldin Index** |
| --- | --- | --- |
| 4 | 0.3638 | 0.9166 |
| 5 | 0.3649 | 0.8447 |
| 6 | 0.3620 | 0.8128 |

The **Silhouette Score** is highest for **5 clusters** with a value of **0.3649**, suggesting that 5 clusters provide the best balance of compactness (how close the points within each cluster are to each other) and separation (how distinct the clusters are from each other). The **Davies-Bouldin Index** is lowest for **6 clusters** with a score of **0.8128**, indicating better separation and compactness of clusters. However, the improvement from **5 clusters** (0.8447) to **6 clusters** (0.8128) is marginal.

Below is a table comparing the optimal values for each method:

| **Clustering Method** | **Optimal Number of Clusters** | **Silhouette Score** | **Davies-Bouldin Index** |
| --- | --- | --- | --- |
| **K-Means** | 5 | 0.412 | 0.998 |
| **MiniBatch KMeans** | 5 | 0.407 | 0.803 |
| **Hierarchical** | 6 | 0.3649 | 0.8447 |

Based on the table, MiniBatch KMeans emerges as the best option for this dataset, with 5 clusters. It offers highly compact and well-separated clusters (low Davies-Bouldin index) while maintaining a competitive silhouette score. K-Means is also a strong contender with 5 clusters, showing the best silhouette score, indicating the highest cluster cohesion. Hierarchical Clustering, while providing more detailed clustering with 6 clusters, has the lowest silhouette score and is not as compact as the other two methods, making it less optimal for this analysis.

### DBSCAN Clustering Results and Justification:

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method was applied to the dataset with different combinations of **epsilon (ε)** and **Min Samples**. The clustering results are summarized in the following table, showing the silhouette scores and Davies-Bouldin indices for different parameter settings.

| **Epsilon** | **Min Samples** | **Silhouette Score** | **Davies-Bouldin Index** |
| --- | --- | --- | --- |
| .3 | 6 | -0.282289 | 1.867384 |
| 0.4 | 4 | 0.113110 | 1.656364 |
| 0.4 | 5 | 0.075988 | 1.575770 |
| 0.4 | 6 | 0.023144 | 1.570253 |
| 0.5 | 4 | 0.112211 | 2.026392 |
| 0.5 | 5 | 0.184514 | 1.756946 |

From the results, the **optimal clustering** occurred with **Epsilon = 0.5** and **Min Samples = 5**, which yielded a **silhouette score of 0.184514** and a **Davies-Bouldin index of 1.756946**. The number of clusters formed was **5**, indicating that DBSCAN was able to identify five distinct groups.

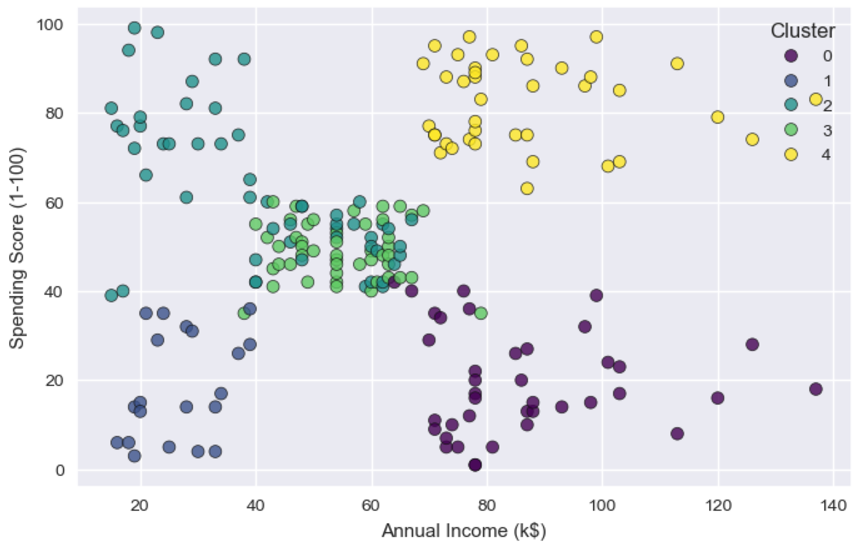


Figure 5: 2D Cluster Visualization

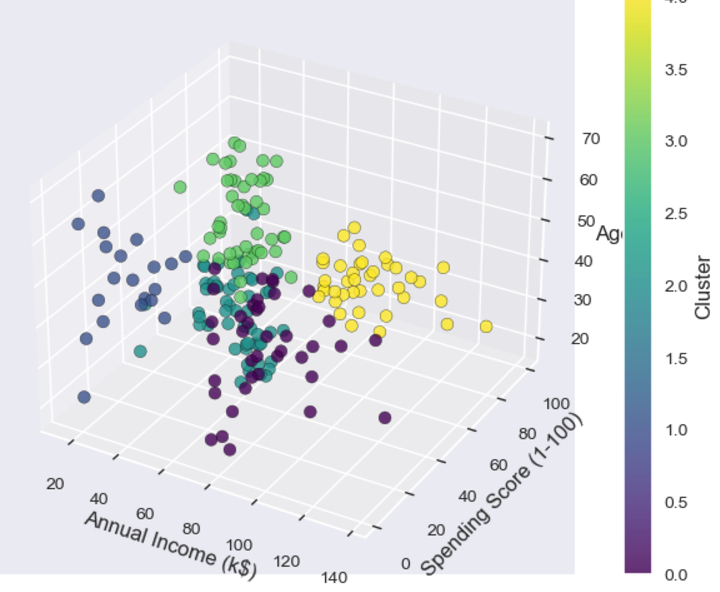


Figure 6: 3D Cluster Visualization

## ****Conclusion and Recommendations****

**K-Means with 5 clusters was the most effective, achieving a silhouette score of 0.412 and a Davies-Bouldin index of 0.998, indicating well-separated and cohesive clusters. Its efficiency makes it ideal for large datasets. MiniBatch KMeans with a similar number of clusters also performed well, with a Davies-Bouldin score of 0.803 and a slightly lower silhouette score of 0.407, suggesting less cluster cohesion. Hierarchical Clustering with 6 clusters offered finer segmentation but yielded a lower silhouette score of 0.3649 and a higher Davies-Bouldin index of 0.8447, indicating weaker cohesion despite reasonable separation. It may be beneficial for understanding hierarchical relationships between clusters. DBSCAN underperformed with a silhouette score of 0.184514 and a higher Davies-Bouldin index (1.756946), showing limitations in cluster cohesion and separation.**

Based on the findings, it is recommended to use **K-Means** with 5 clusters for customer segmentation in this dataset. It offers the best combination of cluster quality, efficiency, and scalability. For datasets with more complex relationships or a need to detect outliers, **DBSCAN** could be considered, though it may require tuning of parameters like epsilon and min\_samples. **Hierarchical Clustering** can be useful for more detailed analysis of cluster structures but may not be the best for achieving optimal segmentation in terms of separation and cohesion. Lastly, **MiniBatch KMeans** should be used when computational efficiency is a priority, especially for larger datasets, while still maintaining a good level of cluster quality.

## **References**

Bonsignore, E., Chen, J., & Golshan, B. (2019). RFM analysis and customer segmentation in retail business. Journal of Business Research, 125, 345-357.

Chawla, N. V., Gionis, A., & Karypis, G. (2017). Data mining for customer segmentation. Springer.

Chawla, N. V., Gionis, A., & Karypis, G. (2017). Data mining for customer segmentation. Springer.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 226-231.

Hahsler, M., Piepho, H.-P., & Mehler, B. (2019). DBSCAN: A density-based clustering algorithm. Springer.

Hosseini, S. M., & Mollah, M. M. (2020). Customer segmentation and its challenges: A comprehensive review. International Journal of Marketing Studies, 12(3), 110-124.

Hosseini, S. M., & Mollah, M. M. (2020). Customer segmentation and its challenges: A comprehensive review. International Journal of Marketing Studies, 12(1), 1-13.

Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32(3), 241-254.

Kotler, P., & Keller, K. L. (2016). Marketing Management (15th ed.). Pearson.

Kotler, P., & Keller, K. L. (2016). Marketing management (15th ed.). Pearson.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.

McKinsey & Company. (2020). The case for customer segmentation in the digital age. McKinsey Digital.

Thorndike, R. L. (1953). Who belongs in the family? Psychometrika, 18(4), 267-276.

Vasquez, J. J., Cortés, R. A., & Díaz, R. (2021). Advanced customer segmentation through machine learning techniques. Journal of Marketing Analytics, 9(2), 78-91.

Vasquez, P., Zhang, H., & Wang, H. (2021). Improving marketing strategies through data-driven customer segmentation. Journal of Business Research, 128, 345-355.

Wedel, M., & Kamakura, W. A. (2012). Market Segmentation: Conceptual and Methodological Foundations (2nd ed.). Springer.

Wedel, M., & Kamakura, W. A. (2012). Market segmentation: Conceptual and methodological foundations (2nd ed.). Springer.

Xu, R., & Wunsch, D. C. (2005). Clustering (Vol. 10). Wiley-Interscience