CUSTOMER SEGMENTATION

University of Missouri – Kansas City

Lecturer: Syed Jawad Hussain Shah

24 July 2024

Group Members
Sankeerthana Challa
Vastsalya Mandagiri

INTRODUCTION:

In today's dynamic business environment, gaining profound insights into customer behavior is crucial for customizing products and services effectively. This project aims to utilize diverse clustering algorithms to categorize customers according to their characteristics, including age, gender, annual income, and spending score. By segmenting customers, businesses can refine their marketing strategies to target specific customer groups, thereby enhancing greater customer satisfaction and driving revenue growth.

RELATED WORK:

Numerous clustering techniques have been applied in the past to address challenges and offer unique advantages in the field of customer segmentation research. Because of their ease of use and scalability, traditional methods like K-Means clustering have long been preferred. They efficiently divide data into discrete clusters according to similarity metrics. Density-based clustering algorithms, like DBSCAN, are excellent at locating clusters of any size and shape while withstanding data noise and anomalies with resilience.

METHODOLOGY:

Our methodology consists of several steps. To ensure data consistency and integrity, the dataset containing customer attributes was first preprocessed to fill in missing values and encode categorical variables. After preprocessing, features were analyzed and their distribution and characteristics were visualized, through visualization. which gave important insights into the structure of the dataset. The relative importance of each feature was then determined, and possible relationships between features were revealed, by feature importance analysis.

After that, the preprocessed data was subjected to two different clustering algorithms, K-Means, DBSCAN. The goal of these algorithms was to divide the customer base into meaningful segments according to their attributes. Finally, metrics like the silhouette score were used to evaluate the quality of the resulting clusters.

Scaling:

A common preprocessing step in clustering algorithms is scaling, which makes sure that each feature contributes equally to the clustering process. It involves converting the feature values to a comparable scale, which is usually 0 to 1 or has a mean of 0 and a standard deviation of 1. Larger magnitude features are kept from influencing the distance calculations during clustering by this normalization, or standardization, of feature scales.

K-Means Clustering:

Initialization of the K-Means Clustering is done with a predetermined number of clusters (K). The elbow method was used to calculate the ideal number of clusters (K). Using this method, the within-cluster sum of squares (WCSS) is plotted against the number of clusters (K) to determine the "elbow" point, or the point at which the WCSS decreases at a significantly slower rate. This figure represents the ideal number of clusters. The final K-Means model was trained with this value of K after the ideal number of clusters was determined.

Then using the Euclidean distance as a guide, K-Means iteratively assigns each data point to the closest cluster centroid and updates the centroids based on the average of the data points within each cluster. Convergence was reached, which was demonstrated by little centroid movement or completing the maximum number of iterations. The resulting clusters formed represent the distinct groups of customers based on their attributes.

Density Based Spatial Clustering of Applications with Noise (DBSCAN):

DBSCAN stands out for its ability to automatically determine the number of clusters and handle datasets with clusters of varying shapes and sizes. Its approach revolves around grouping densely populated data points into clusters, without requiring a predefined number of clusters. This method allows for the identification of clusters that may have irregular shapes and sizes, enhancing its applicability to diverse datasets.

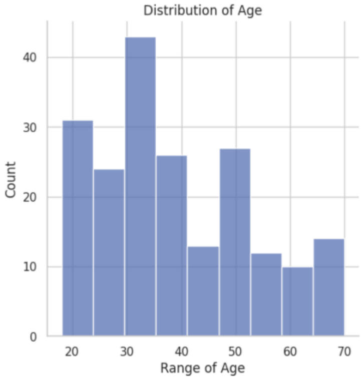
To apply the DBSCAN clustering we need to find the suitable ‘eps’ value (i.e., the maximum distance between two points for them to be considered as in the same neighborhood) and “min_samples” (i.e., The minimum number of points required in a neighborhood for a point to be considered a core point). To get “eps” value we used the K-distance plot to get idea about how data points are distributed and calculated distance matrix if there is need to perform any further analysis.

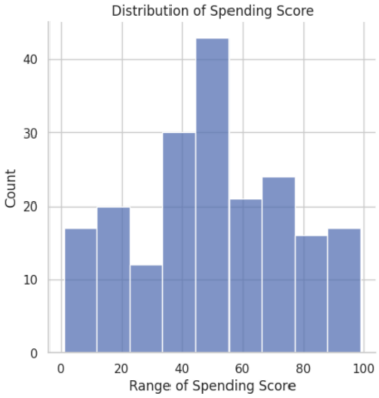
Silhouette score:

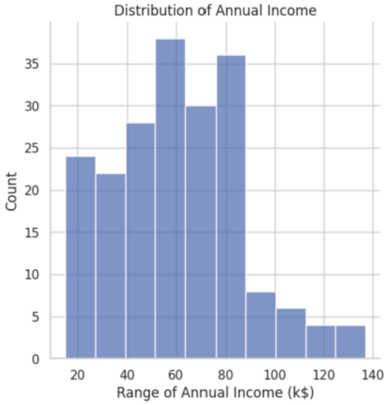
The silhouette score is a metric used to evaluate the quality of clusters produced by clustering algorithms, including K-Means, Hierarchical Clustering, DBSCAN, and others. It measures how well-separated clusters are and provides insights into the compactness and separation of clusters within the dataset.

RESULTS AND DISCUSSION:

Data Visualization:







Distribution of Age:

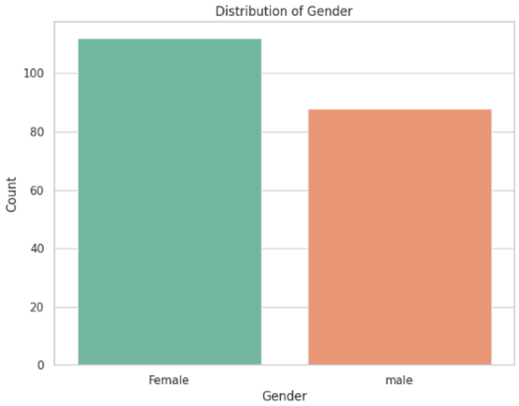
Most customers are between 30 and 40 years old. There is a smaller but significant number of customers in the 20-30 and 40-50 age ranges and very few customers are under 20 or over 70.

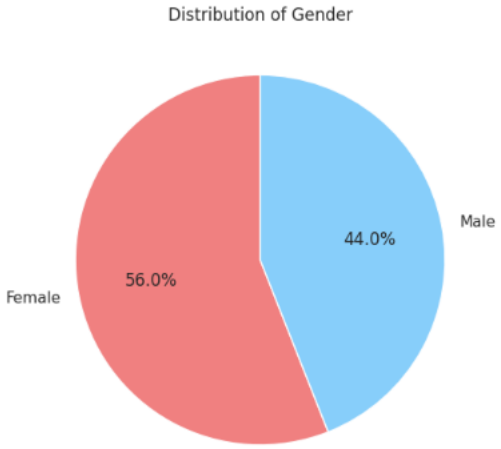
Distribution of Spending Score:

The distribution is relatively even between 0 and 80, with a slight peak around 40-60.There are fewer customers with spending scores above 80.

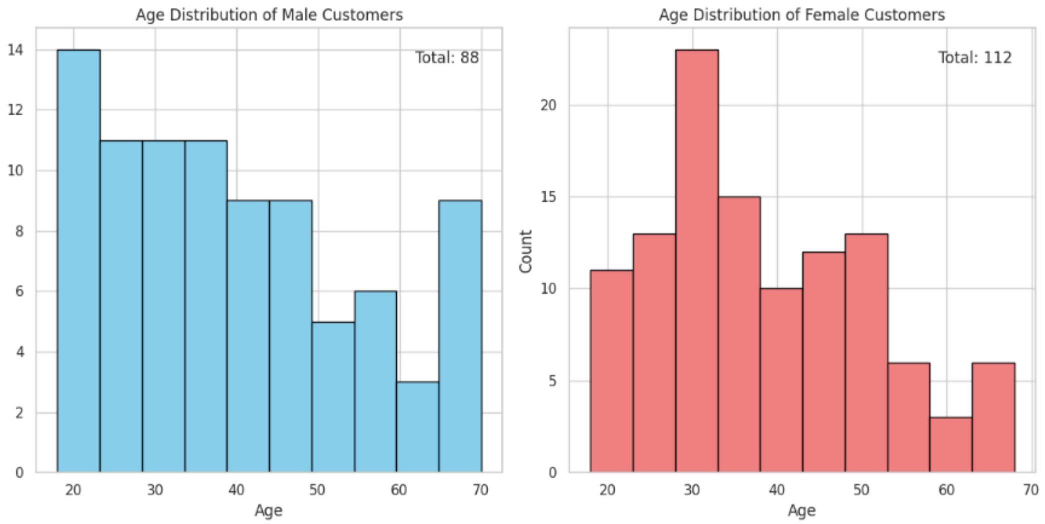
Distribution of Annual Income:

The distribution is right-skewed, meaning most customers have lower incomes. The majority of customers have annual incomes between 40,000 and 100,000 USD.There are fewer customers with annual incomes below 20,000 USD or above 120,000 USD.



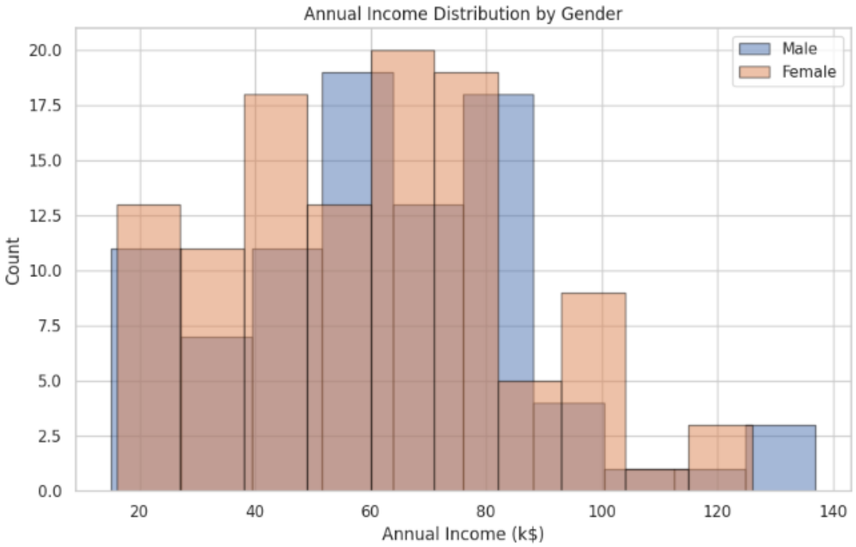


The data indicates a higher proportion of females in the dataset. Approximately 56% of the individuals represented are female, while 44% are male.



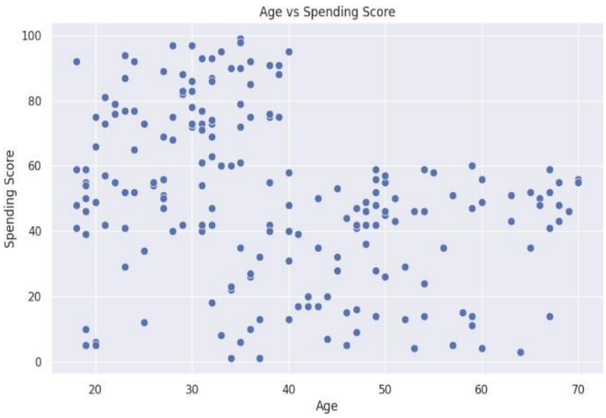
Through this visualization we can say that both male and female customers exhibit a primary age range between 20 and 40 years old. The total count of female customers (112) is higher than that of male customers (88). The highest frequency for both genders falls within the 20-30 age bracket. Both distributions exhibit a right-skewed pattern, indicating a larger proportion of younger customers.

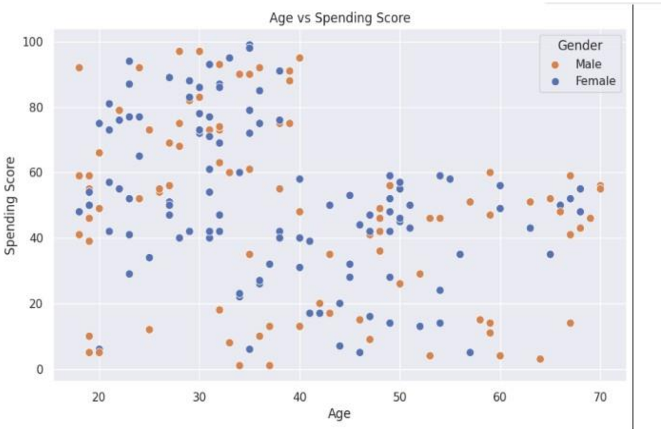
The data suggests a potential focus on marketing and product offerings tailored to the 20-40 age demographic. The higher number of female customers may warrant gender-specific marketing strategies or product lines.



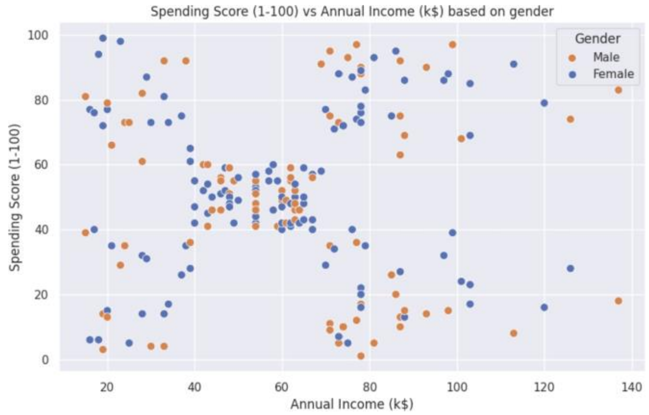
The income ranges for both genders significantly overlap, indicating a shared income spectrum. While there's overlap, the male distribution shows a slight tendency towards higher income brackets, with a more pronounced tail in the upper income ranges. The female distribution appears to be concentrated in the lower to middle income brackets, with a steeper decline in frequency as income increases Both genders exhibit income peaks around the 60-80k annual income range.

The data suggests potential gender-based income disparities, with men tending to have higher earning potential. The overlapping distributions highlight the presence of both high-earning women and lower-income men.

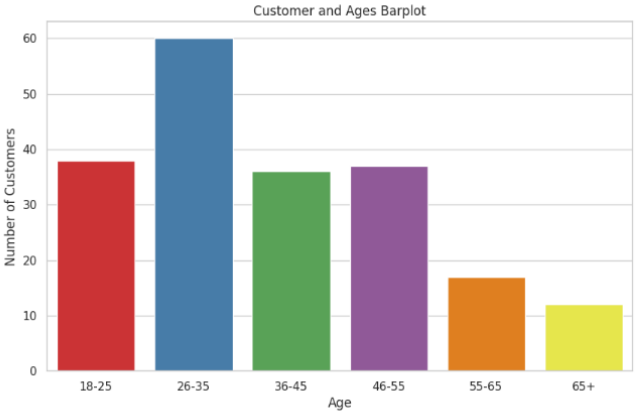


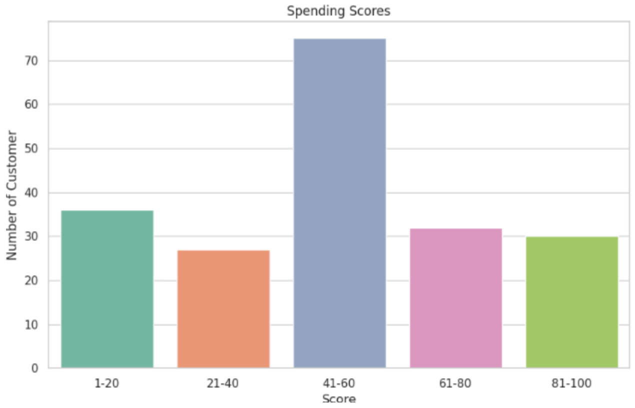


This visualization illustrates the relationship between customers' ages and their spending scores, segmented by gender. It reveals that customers aged 20-40 tend to have higher spending scores across both genders, indicating a more active spending behavior in this age group.



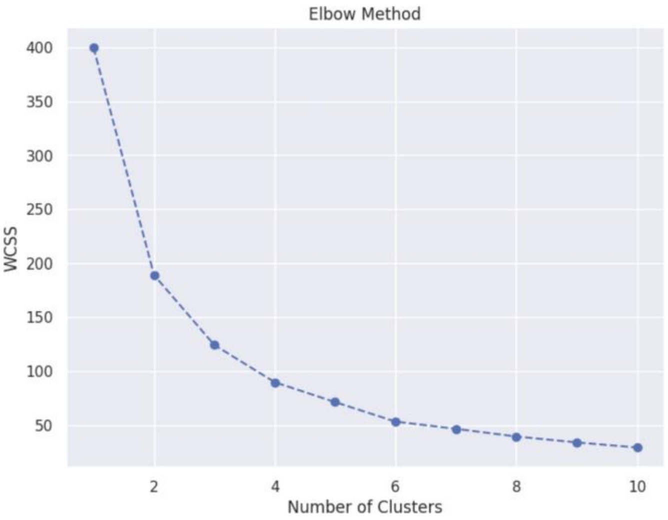
This plot shows the correlation between annual income and spending score, categorized by gender. It shows that customers with an annual income between $40,000 and $60,000 generally have a moderate spending score (40-60). This suggests a threshold where income levels stabilize spending behavior.

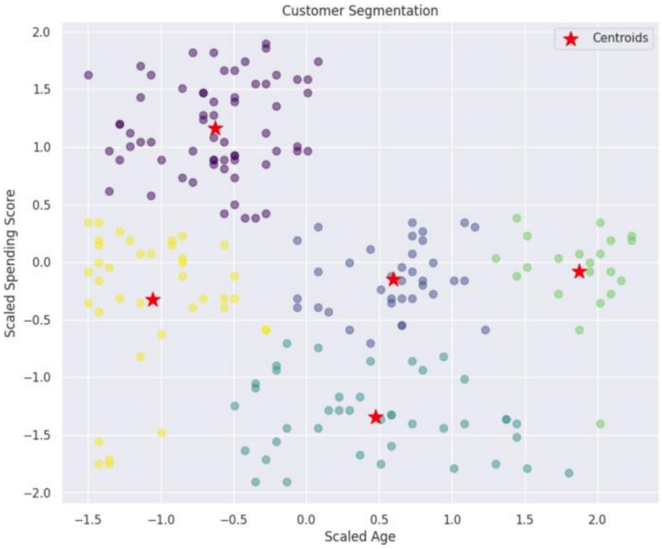




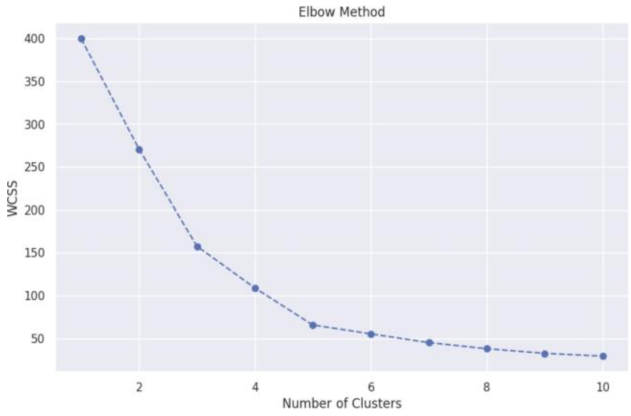
K-Means Clustering:

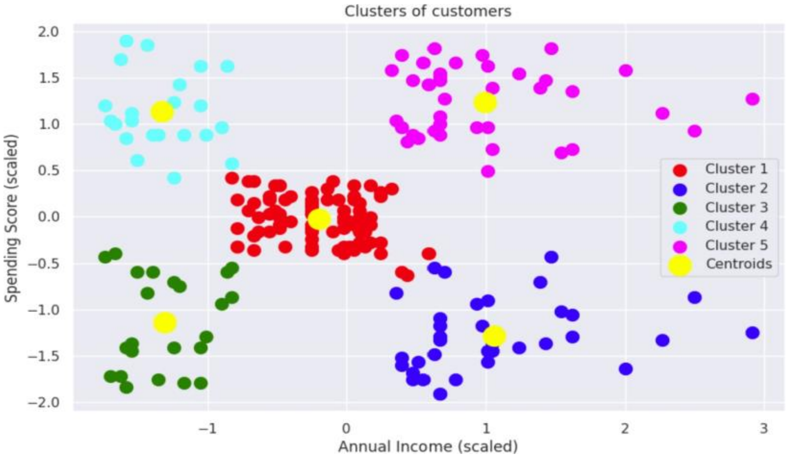
The elbow curve helps determine the optimal number of clusters for the K-Means algorithm by plotting the within-cluster sum of squares (WCSS) against the number of clusters (K).
The curve indicates an "elbow" at K=5, suggesting that five clusters best represent the data with diminishing returns in reducing WCSS beyond this point.





This informs the choice of K for clustering, ensuring that the number of customer segments is both meaningful and interpretable. It prevents overfitting and ensures that clusters are distinct enough for actionable insights.
Based on the Elbow Method, the optimal number of clusters for this dataset appears to be 5. The scatter plot reveals distinct clusters of customers based on their age and spending score. Analyzing the distribution of points within each cluster can provide insights into customer segments, such as "young spendthrifts," "middle-aged savers," or "older high spenders."
The location of the centroids can help characterize the typical customer profile for each cluster.





Like the previous visualization, this plot shows clusters based on annual income and spending score. The Elbow Method strongly suggests that 5 clusters are optimal for this too.

Cluster Characteristics:

- **Cluster 1 (Red):** Likely represents customers with high spending scores but relatively low annual income.
- **Cluster 2 (Green):** Might represent customers with both high annual income and high spending scores.
- **Cluster 3 (Blue):** Could represent customers with average annual income and spending scores.
- **Cluster 4 (Cyan):** May represent customers with low annual income and low spending scores.
- **Cluster 5 (Magenta):** Could represent customers with high annual income but low spending scores.

This segmentation is crucial for financial planning, product positioning, and understanding the spending capacity of different income groups, allowing for more effective marketing and product development strategies.

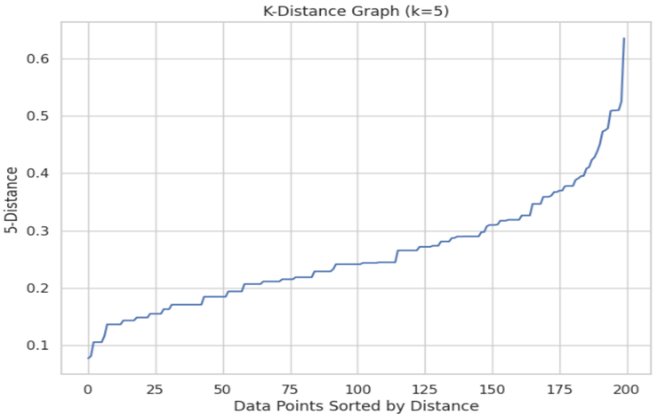
Density Based Spatial Clustering of Application with Noise (DBSCAN):

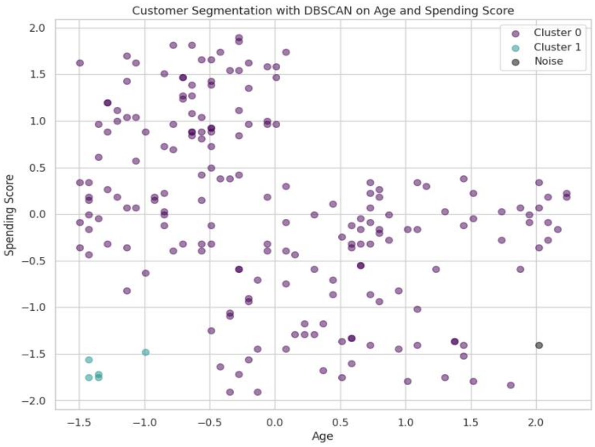
The K-Distance plot helps determine the appropriate 'eps' parameter for DBSCAN by showing the distance to the k-th nearest neighbor. It indicates that a suitable 'eps' value lies between 5 and 10, guiding the DBSCAN algorithm in identifying dense regions or clusters.

Proper selection of the 'eps' value ensures that DBSCAN accurately captures clusters of varying shapes and sizes, including detecting outliers (noise), which are crucial for understanding anomalies or niche customer segments.

Here we have taken k=5

With the application of Model DBSCAN to data given we got the results and visualized on graph



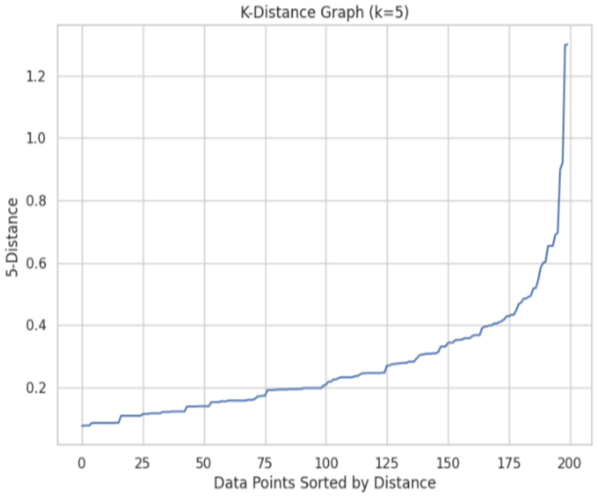


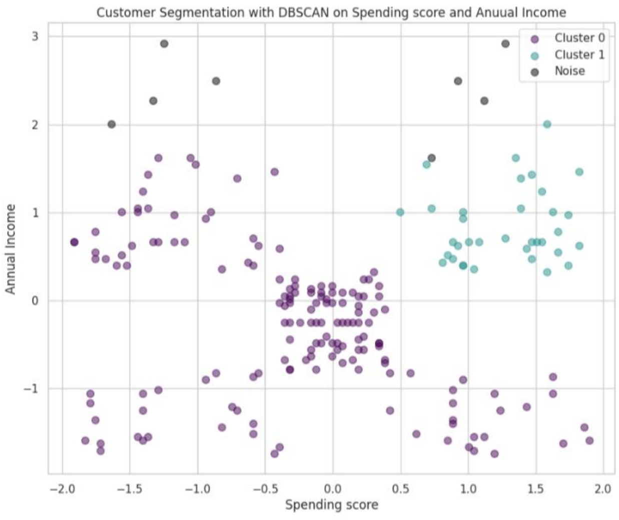
The K-Distance graph (k=5) shows the distance to the 5th nearest neighbor for each data point.The graph has a general upward trend, indicating that as you consider points further from the origin, their distance to the 5th nearest neighbor tends to increase.There's a noticeable elbow or bend in the curve around a distance of 0.4. This suggests that a suitable value for epsilon in the DBSCAN algorithm might be around 0.4.

Customer Segmentation with DBSCAN on Age and Spending Score

The scatter plot visualizes the results of applying DBSCAN clustering to the dataset, considering age and spending score. Three distinct clusters are evident, labeled Cluster 0, Cluster 1, and Noise. Cluster 0 appears to be concentrated in a specific region of the age-spending score space. Cluster 1 is more spread out but still forms a discernible group. Noise points are scattered across the plot, representing data points that didn't fit into any defined cluster.

By identifying distinct customer clusters based on age and spending behavior, businesses can tailor marketing campaigns and messages to specific segments. For instance, cluster 1, which might represent high-spending customers, could be targeted with exclusive offers or loyalty programs.





The second graph visualizes the results of DBSCAN clustering based on spending score and annual income. Identifies three clusters (Cluster 0, Cluster 1) and noise points. Cluster 0 appears denser and more compact than Cluster 1. Noise points are scattered across the plot.

The DBSCAN algorithm has identified two distinct clusters in the data, suggesting heterogeneity among customers based on spending score and annual income. Further analysis would be required to understand the specific characteristics of each cluster, such as average income, spending patterns, and demographics.

Silhouette score:

	K-Means	DBSCAN
Annual Income vs Spending Score	**0.55465**	**0.35044**
Age vs Spending Score	**0.44754**	**0.08622**

K-Means generally outperforms DBSCAN For both feature combinations, K-Means achieved higher Silhouette scores, suggesting it produced better-defined clusters. The combination of Annual Income and Spending Score resulted in higher Silhouette scores for both algorithms compared to Age and Spending Score. This indicates that these features might be better suited for clustering in this dataset.

CONCLUSION AND FUTURE WORK:

In conclusion We have explored and applied four different clustering techniques like “**k-means**” “**DBSCAN**” to the problem of Customer Segmentation based on demographic and behavioral attributes. Each technique offers unique advantages and considerations, contributing valuable insights to our analysis.

In Future work We Aim to enhance Customer Segmentation techniques by exploring additional features and transformations to capture nuanced customer behavior better. Additionally, we plan to investigate ensemble clustering methods to improve the robustness and stability of segmentation solutions. By integrating clustering with predictive modeling techniques will enable us to leverage customer segments for personalized marketing campaigns and recommendations.

REFERENCES

[1]. Jomark Pablo Noriega; Luis Antonio Rivera , Jose Alfredo Herrera. Machine Learning for Credit Risk Prediction: A Systematic Literature Review, 2023.

[2]. Norshakirah Aziz; Emelia Akashah Patah Akhir; Izzatdin Abdul Aziz. A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems, 2020.

[3]. Aized Amin Soofi; Classification Techniques in Machine Learning: Applications and Issues, 2017.

[4]. Swastik Satpathy; SMOTE for Imbalanced Classification with Python, 2023