Project Report: Customer Segmentation Task

1. The Task I Was Assigned

I was given the task of building a model to segment customers based on their annual income and spending score, using the provided Mall Customer dataset. The initial requirement was to apply K-Means clustering, determine the optimal number of clusters, and visualize the results. The assignment also included bonus objectives to experiment with different clustering algorithms and analyze average spending per cluster.

2. How I Approached the Task

To complete the assignment, I broke down the problem into a clear, multi-stage process, focusing on a logical progression of data science steps.

Data Acquisition and Initial Exploration

- **My Goal:** To load the dataset and gain a preliminary understanding of its structure and content.
- **What I Did:** I loaded the `Mall_Customers.csv` dataset into a Pandas DataFrame. I performed initial data inspections using functions such as `df.head()`, `df.info()`, `df.describe()`, and `df.isnull().sum()` to check for missing values and understand data types.
- **The Result:** The data was successfully loaded, and I confirmed it was clean with no missing values. I also identified 'Annual Income (k$)' and 'Spending Score (1-100)' as the key features relevant for the clustering task.
- **My Conclusion:** This initial step provided a solid foundation, confirming data readiness for subsequent processing.

Data Preprocessing and Scaling

- **My Goal:** To prepare the selected features for clustering by normalizing their scales.
- **What I Did:** I isolated the 'Annual Income (k$)' and 'Spending Score (1-100)' features. Recognizing that K-Means is sensitive to feature scales, I applied `StandardScaler` from `sklearn.preprocessing` to transform these numerical features. This process centers the data around zero with a unit standard deviation.
- **The Result:** The features were successfully scaled, making them suitable for the K-Means clustering algorithm. Visualizations of the unscaled and scaled data confirmed the effectiveness of this transformation.
- **My Conclusion:** Data preprocessing was crucial for ensuring the clustering algorithm would perform optimally and yield meaningful results.

K-Means Clustering: Optimal K Determination

- **My Goal:** To determine the most appropriate number of clusters (K) for the dataset.
- **What I Did:** I employed the Elbow Method. This involved iterating K-Means clustering for a range of K values (from 1 to 10) and calculating the Sum of Squared Errors (SSE) or inertia for each iteration. I then plotted the SSE values against the number of clusters.
- **The Result:** The plot clearly showed a distinct "elbow point" at **K=5**. This indicated that increasing the number of clusters beyond 5 yielded diminishing returns in terms of reducing the SSE.
- **My Conclusion:** K=5 was chosen as the optimal number of clusters, balancing model complexity with the clarity of segmentation.

K-Means Model Training and Cluster Assignment

- **My Goal:** To train the K-Means model with the optimal number of clusters and assign each customer to a segment.
- **What I Did:** I initialized and trained the KMeans model using n_clusters=5 on the scaled data. After fitting, I extracted the cluster labels for each data point and added them as a new 'Cluster' column to the original DataFrame. I also inspected the coordinates of the cluster centers in both scaled and original data scales.
- **The Result:** The K-Means model successfully segmented the 200 customers into 5 distinct clusters. The cluster assignments and center coordinates provided the basis for detailed segment interpretation.
- **My Conclusion:** The core clustering task was successfully completed, providing a foundational segmentation of the customer base.

Cluster Visualization and Interpretation

- **My Goal:** To visually represent the customer segments and interpret their characteristics.
- **What I Did:** I created a 2D scatter plot using 'Annual Income (k$)' and 'Spending Score (1-100)' as the axes. Data points were color-coded according to their assigned K-Means cluster, and the cluster centroids were overlaid on the plot as prominent markers.
- **The Result:** The visualization clearly depicted 5 well-separated customer segments. This allowed for a qualitative interpretation of each group's income and spending behavior, identifying segments such as "high-income, high-spending" or "low-income, low-spending" customers.
- **My Conclusion:** The visual interpretation confirmed the effectiveness of the clustering and provided actionable insights into customer behaviors.

Bonus Tasks

Exploring Different Clustering Algorithms (DBSCAN)

- **My Goal:** To explore an alternative clustering approach and compare its segmentation with K-Means.
- **What I Did:** I applied DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to the scaled data, using `eps=0.5` and `min_samples=5`. I then visualized the DBSCAN clusters alongside the K-Means results.
- **The Result:** DBSCAN identified 2 primary clusters and a set of noise points (assigned cluster label -1). This demonstrated DBSCAN's ability to find density-based clusters of arbitrary shapes and explicitly handle outliers, a key difference from K-Means.
- **My Conclusion:** DBSCAN offered a valuable alternative perspective, highlighting the presence of noise and different cluster structures, though K-Means provided a more straightforward segmentation for this specific business context.

Analyzing Average Spending per Cluster

- **My Goal:** To quantitatively characterize each K-Means cluster based on its average spending.
- **What I Did:** I calculated the mean 'Spending Score (1-100)' for each of the 5 K-Means clusters using the `groupby()` function on the DataFrame.
- **The Result:** This analysis provided precise average spending scores for each segment, confirming the qualitative interpretations from the visualization. For instance, clusters identified visually as "high-spending" indeed showed significantly higher average spending scores.
- **My Conclusion:** This quantitative analysis further solidified the understanding of each customer segment, providing concrete metrics for targeted strategies.

3. Final Result of My Work

After completing all assigned tasks and bonus objectives, my final conclusion is that **K-Means clustering with 5 clusters** provides a clear, actionable, and interpretable segmentation of mall customers based on their income and spending habits. The distinct clusters identified (e.g., high-income high-spenders, low-income high-spenders, average-income average-spenders, high-income low-spenders, and low-income low-spenders) offer valuable insights for developing targeted marketing strategies. While DBSCAN provided an interesting alternative perspective by identifying density-based clusters and noise, K-Means offered a more straightforward and actionable segmentation for this particular dataset's structure. This project successfully demonstrated the application of unsupervised learning techniques for customer segmentation, providing a robust foundation for business insights.