K-Means Clustering Report

# 1. Introduction

This project performs unsupervised machine learning using the K-Means clustering algorithm on the Mall\_Customers dataset. The objective is to segment customers based on their annual income and spending score to uncover natural groupings that could guide targeted marketing strategies.

# 2. Tools and Libraries Used

- Python  
- Pandas  
- Matplotlib  
- Seaborn  
- Scikit-learn  
- PCA (Principal Component Analysis) for visualization  
- Silhouette Score for evaluation

# 3. Dataset Description

Dataset: Mall\_Customers.csv  
Attributes Used:  
- Age  
- Annual Income (k$)  
- Spending Score (1–100)  
  
Non-numeric attributes like 'Gender' and identifiers like 'CustomerID' were excluded.

# 4. Steps Performed

## 4.1 Data Preprocessing

Dropped non-numeric columns and retained only numerical features for clustering.

## 4.2 Elbow Method

The Elbow Method was used to determine the optimal number of clusters (K). The inertia was plotted against K from 1 to 10 to identify the 'elbow point', where adding more clusters results in diminishing returns. Based on the elbow plot, K=5 was selected.

## 4.3 K-Means Clustering

K-Means clustering was applied using the optimal number of clusters (K=5). Each data point was assigned a cluster label based on proximity to cluster centroids.

## 4.4 PCA Visualization

Principal Component Analysis (PCA) was used to reduce dimensionality to 2D for visualizing clusters in a scatter plot. Clusters were color-coded for better distinction.

## 4.5 Evaluation

The Silhouette Score was calculated to assess clustering performance. A score of 0.444 indicates moderately well-separated clusters.

# 5. Conclusion

K-Means clustering successfully segmented mall customers into five distinct groups based on income and spending behavior. These insights can help businesses develop personalized marketing strategies. The use of the elbow method, PCA, and silhouette score collectively validated the clustering approach.