K-Means Cluster Analysis

Assignment 2  
Name: Zihan Yan

# Introduction

The objective of this assignment is to apply K-Means Cluster Analysis to the Mall Customers dataset. This analysis helps group customers into clusters based on their annual income and spending scores, providing insights into their behavior. By identifying distinct customer groups, better marketing strategies can be designed. This report outlines the steps taken to prepare the data, determine the optimal number of clusters, and interpret the results.

# Data Dimensionality

The Mall Customers dataset contains information about customer demographics and behavior. Here is a summary of the dataset’s dimensionality:

* Total number of customers: 200
* Number of attributes: 3
* Attributes and their data types:
  + Age: int64
  + Annual Income: int64
  + Spending Score: int64
* Missing values: There are no missing values in the dataset.

This clean dataset is suitable for K-Means Cluster Analysis, where we will focus on the 'Annual Income' and 'Spending Score' attributes to group customers into clusters.

A screen shot of a computer program

Description automatically generated

# Determining the Optimal K Value

To determine the optimal number of clusters, we used the elbow method. This involves plotting the inertia (sum of squared distances) for different values of K (the number of clusters) and observing the point where the rate of decline slows down, indicating the optimal K value.  
  
The elbow plot (shown below) suggests that K = 5 is the optimal number of clusters because the inertia decreases significantly up to this point, after which the reduction becomes more gradual.

A graph with a line

Description automatically generated

# K-Means Clustering Results

I applied K-Means clustering to the dataset using K = 5 clusters. The 'Annual Income' and 'Spending Score' attributes were used as the features for clustering. Each customer was assigned to one of the five clusters.  
  
The scatter plot below visualizes the clusters, with customers grouped based on their annual income and spending score.

A chart of a number of dots

Description automatically generated with medium confidence

# Is this an accurate result? Explain your reasoning.

Yes, this is an accurate result based on the data provided. K-Means clustering is effective in grouping customers based on their similarity in annual income and spending score. The elbow method was used to determine the optimal number of clusters, which showed that **K=5** was the best fit for the data. The scatter plot reveals that the customers are grouped into five distinct clusters, which are well-separated in terms of income and spending score. The clustering accurately reflects the diverse behavior of customers based on these two key factors.

# Cluster Descriptions and Reasons for Further Advertising

Each cluster represents a distinct group of customers based on their spending behavior and income. Below are the descriptions of each cluster:

* Cluster 0: Customers with low annual income and moderate spending scores.
* Cluster 1: Customers with high annual income but low spending scores.
* Cluster 2: Customers with high annual income and high spending scores.
* Cluster 3: Customers with low annual income and high spending scores.
* Cluster 4: Customers with moderate annual income and moderate spending scores.

The clustering appears accurate because the customers are well-separated based on their income and spending scores, which reflects distinct customer behaviors.

As a supervisor, these clusters can be useful for future advertising strategies. For example:

* Cluster 2 (high income, high spending) could be targeted for premium products.
* Cluster 0 (low income, moderate spending) might benefit from budget promotions.

These distinct clusters allow for more personalized marketing strategies.

# Advertising Strategy and Conclusion

The clustering results reveal distinct customer segments that can be effectively targeted for future advertising. For example, Cluster 2 represents high-income, high-spending customers who are likely to respond well to premium product offerings, while Cluster 0 represents low-income customers who might prefer budget-friendly promotions.  
  
These distinct clusters allow businesses to craft personalized advertising strategies that resonate with different customer segments, ultimately improving marketing effectiveness.