CUSTOMER SEGMENTATION USING MACHINE LEARNING

Archana Akella
akellaarchanas@gmail.com

Abstract: Nowadays Customer segmentation became very popular method for dividing company's customers for retaining customers and making profit out of them, in the following study customers of different organizations are classified on the basis of their behavioral characteristics such as spending and income, by taking behavioral aspects into consideration makes these methods an efficient one as compares to others. For this classification a machine learning algorithm named as k means clustering algorithm is used and based on the behavioral characteristic's customers are classified. Formed clusters help the company to target individual customer and advertise the content to them through marketing campaigns and social media sites which they are really interested in.

Keywords: Machine learning, Customer segmentation, K-means algorithm

1. Introduction

Today many of the businesses are going online and, in this case, online marketing is becoming essential to hold customers, but during this, considering all customers as same and targeting all of them with similar marketing strategy is not very efficient way rather it's also annoys the customers by neglecting his or her individuality, so customer segmentation is becoming very popular and also became the efficient solution for this existing problem. Customer segmentation is defined as dividing company's customers on the basis of demographic (age, gender) and behavioral (annual income, spending score) aspects. Since demographic characteristics do not emphasize the individuality of customers because the same age groups may have different interests, behavioral aspects are a better approach for customer segmentation as it focuses on individuality and we can do proper segmentation with the help of it.

2. Problem Setting:

A mall wants to segment its customer base in order to better target marketing efforts and improve overall revenue. The mall has collected data on customer demographics, purchasing behavior, and spending patterns.
The goal of the customer segmentation project is to use machine learning

techniques to group customers based on their income levels, spending scores, and purchasing habits, hoping to tailor their marketing strategies to the ideal customer segments and enhance the shopping experience. One of the challenges in this project will be handling high-dimensional data, as well as addressing any outliers or inconsistencies in the dataset. Additionally, selecting an appropriate number of clusters and determining the optimal parameters for the clustering algorithm will also be critical.

3. Problem Definition:

By understanding the purchasing behavior and demographics of mall customers, the mall management can optimize marketing campaigns and provide a more personalized shopping experience. The insights derived from this segmentation will help the mall better serve its customers, improve customer loyalty, and increase profitability.

Questions addressed in the project:

1. What are the ideal customer segments within the mall's customer base?
2. Demographics: What is the age, gender, and income distribution of mall customers?
3. Can the identified customer segments improve targeted marketing and boost sales?
4. What are the purchasing habits and spending patterns of customers?
5. What are the probable spending behaviors based on income and previous shopping habits?
6. What products or services should the mall prioritize for each customer segment?

4. Data Source:

https://www.kaggle.com/datasets/nelakurthisudheer/mall-customer-segmentation

5. Data Description

The dataset used in this project contains 200 rows and 5 columns, with detailed information about customers. The columns include the following:

1. Demographics:
    - Customer ID
    - Gender

    - Age
    - Annual Income (k$)
2. Behavioral Attributes:
    - Spending Score (1–100): A measure of customer spending behavior based on factors such as income and purchasing habits.

We selected Kaggle's 'Mall Customers' dataset for our customer segmentation project due to its concise structure, variety of customer-related attributes, and direct applicability to real-world scenarios. The dataset provides valuable insights into a retail mall's customer demographics and spending patterns, making it ideal for segmentation analysis.

Purpose and Approach:

Our goal is to analyze this data to identify patterns and segment customers into different groups based on their spending behavior and income levels. By doing so, we aim to help mall management better understand their customers and design tailored marketing strategies that helps in unique needs of each segment. Overall, we believe that the 'Mall Customers' dataset is a powerful resource for conducting segmentation analysis, and this project will provide actionable insights for improving customer experience and driving sales.

Table 1. Description of Variables

SNo:	Variable	Description
1	Customer ID	Customer's unique identifier
2	Gender	Gender of the customer
3	Age	Age of the customer, derived from their year of birth
4	Annual Income (k$)	Annual income of the customer in thousands of dollars
5	Spending Score (1-100)	Customer purchasing habits

6. Data Processing

A. Data Cleaning

1. **Changing the Data Type**
    - The Gender column was transformed using LabelEncoder, where 'Male' was encoded as 0 and 'Female' as 1.

2. **Generating New Features**
    - **Age of Customers:**
      Understanding age distribution help in identifying customer preferences and shopping habits among different age groups.
    - **Spending Behavior Categories:**
      Customers were grouped into spending behavior categories (e.g., low, medium, high spenders) based on their Spending Scores.
3. **Confirming Data Completeness**
    - No missing values were found in the dataset during data validation.
    - Data consistency across all features was confirmed, ensuring the dataset was ready for clustering and analysis without requiring imputation or handling of missing values.

By performing these data cleaning steps and feature engineering, the dataset was optimized for segmentation analysis, ensuring high-quality input for machine learning algorithms.

7. Data Exploration

- Gender Encoding
- Income and Spending Analysis
- Correlation matrix
- Feature scaling

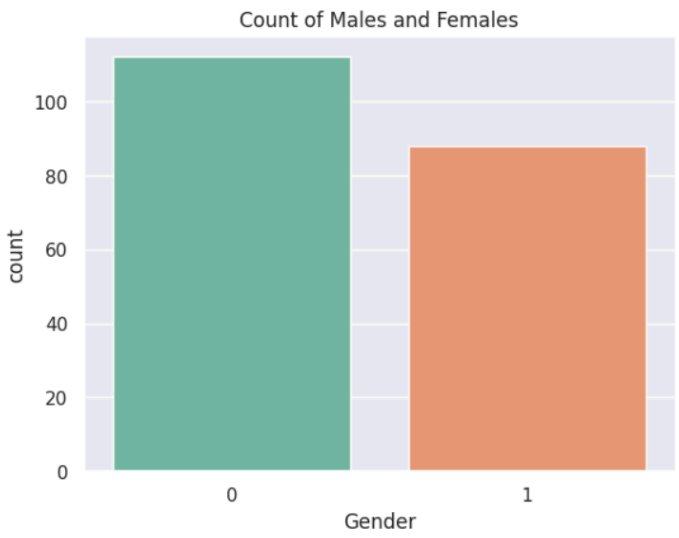
8.Gender-Based Analysis

1. Encoding Gender Data

The `LabelEncoder` is used to convert the categorical "Gender" column into numerical values: 0 - Female, 1 - Male.

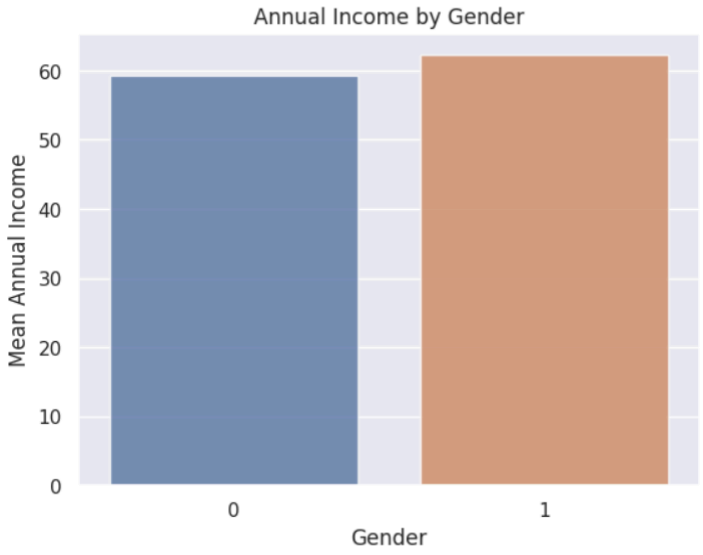
2. Gender Count

The distribution of male and female customers is calculated using `value_counts()`



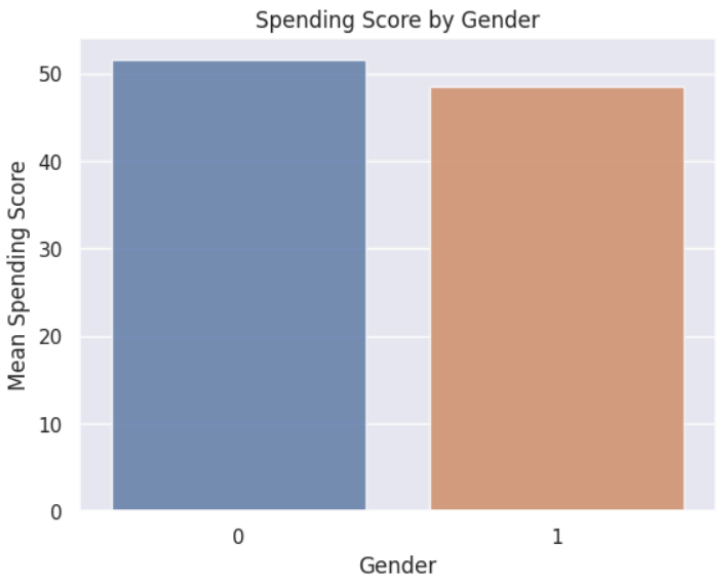
3. Annual Income by Gender

The average annual income for males and females is calculated and visualized



4. Spending Score by Gender

The average spending score for males and females is calculated and visualized.



5. Age Distribution by Gender

The average age of customers is calculated for each gender.

	Gender	Age
0	0	38.098214
1	1	39.806818

Looking at the gender distribution in our dataset, we observe that the number of female customers is somewhat more than male customers. This highlights an opportunity to tailor marketing efforts to female customers, as female customers are significant portion of the customer base.

From the Annual Income analysis by gender, we find that male customers tend to have a higher average annual income compared to female customers.

When it comes to spending scores, female customers exhibit a higher average spending score than male customers, suggesting that female customers might be more active shoppers

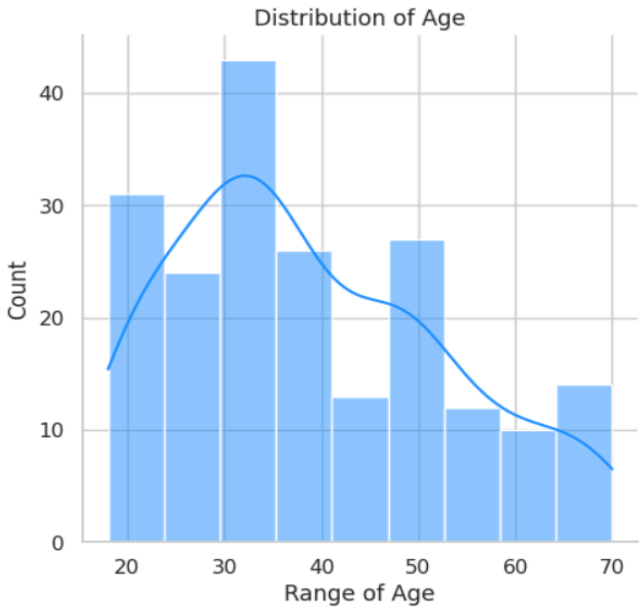
analyzing the age distribution by gender, we see a fairly even distribution of ages across genders, but specific trends of certain age groups could help refine targeted strategies for each group.

These insights suggest the need for customized marketing approaches for each gender to enhance engagement and maximize sales opportunities.

9.Distribution of Age, Annual Income (k$), Spending Score (1-100)

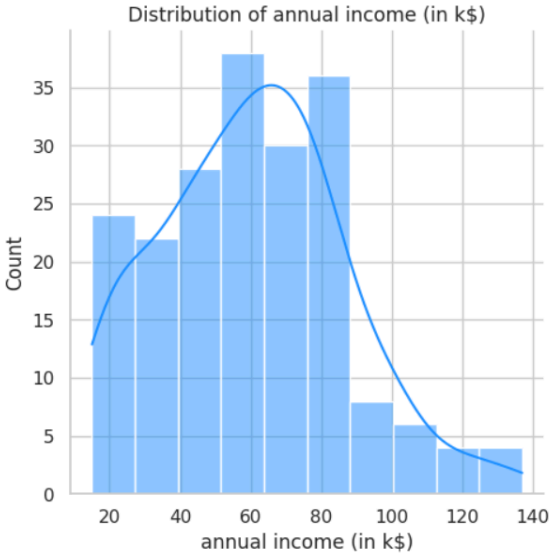
Age Distribution: Dataset shows majority of the customers are within 26-40 age range compared to the 41-60 age range. It indicates that customers are mostly young to middle aged individuals. The smallest group of customers are below 25 and above 60.

- Marketing operations should prioritise more young and middle age individuals and should find opportunities to engage younger and older age group



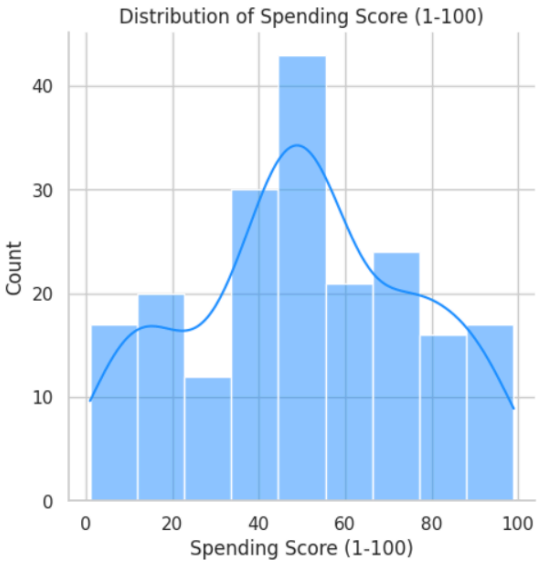
Annual Income (k$): Most of the customers fall within the income range of $15000 to $75000. Only small group of customers are having higher income range crossing $100000, so majority of the customer are middle income individuals

- Product Offers and Pricing strategies are suitable to middle income individuals and premium offers targeting smaller high income group



Spending Score (1-100): Distribution shows two groups, one with low spending scores(1-40) and another high spending score(60-100)

- Analyzes one spends minimally and another that spends highly, Helps businesses in targeted strategies



10.Extracting relevant features

- Extracting relevant features from dataset for further analysis and in this case of customer segmentation, might focus on features like Age, Annual Income (k$) and Spending Score(1-100)
- we use `.values` to extract the values as a numpy array, which is typically the format required for many machine learning algorithms
- Used `df.iloc[:,3:5].values` and `df.iloc[:,2:4].values` to extract annual income and spending score , age and annual income

11.Standardize the Features:

- It's important to scale and standardise the features if they are on different ranges like annual income in thousands and spending score in 1-100
- StandardScaler helps to ensure both features contribute equally to analysis

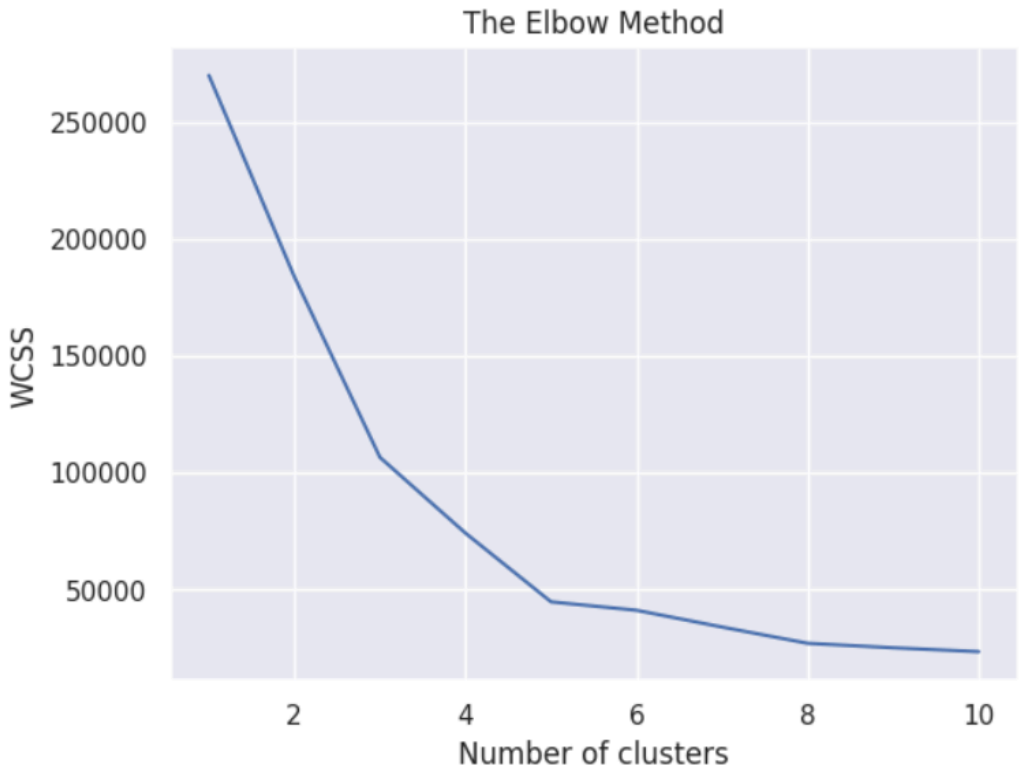
12.WCSS(Within Cluster Sum of Square) & Elbow Method

- The WCSS is a metric used in clustering algorithms, particularly in K-Means clustering, to evaluate the compactness of the clusters formed.
  measures how far each data point is from the center of its cluster, if value is low then the data points are closer

- By using the **elbow method**, we plot the WCSS values for different numbers of clusters. The point where the WCSS starts decreasing(forming an "elbow")

helps us determine the optimal number of clusters for segmenting the customers.

- Elbow Graph is a visualization that helps in determining the optimal number of clusters, in this case mall customers based on annual income and spending score



- The elbow point in the graph occurs at **k=5**, indicating that the optimal number of clusters for segmenting the mall customers based on their Annual Income and Spending Score is 5
- we can conclude that the dataset can be effectively divided into 5 different customer segments.

13. K Means Clustering

K-means Clustering is a clustering Algorithm in which we are given data points with its data set and features and the mechanism is to categorise those data points into clusters as per their similarities. The algorithm forms K clusters based on its similarity. To calculate the similarity Kmeans uses Euclidean distance measurement method.

Steps

i. In the first step, we randomly initialize k points.

ii. K-means classifier categorizes each data point to its nearest mean and rewrites the mean's coordinates.

iii. Iteration is continuing up till all data points are classified

- KMeans clustering algorithm to segment the mall customers into distinct groups based on their Annual Income and Spending Score.
- `n_clusters=5`: This specifies that we want to divide the dataset into 5 clusters, which is based on the result from the elbow method.

```
kmeans = KMeans(n_clusters=5, init='k-means++', random_state=0 )
Y = kmeans.fit_predict(X)
Y
```

```
array([3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4,
       3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 0,
       3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 1, 0, 1, 2, 1, 2, 1,
       0, 1, 2, 1, 2, 1, 2, 1, 2, 1, 0, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1,
       2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1,
       2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1,
       2, 1], dtype=int32)
```

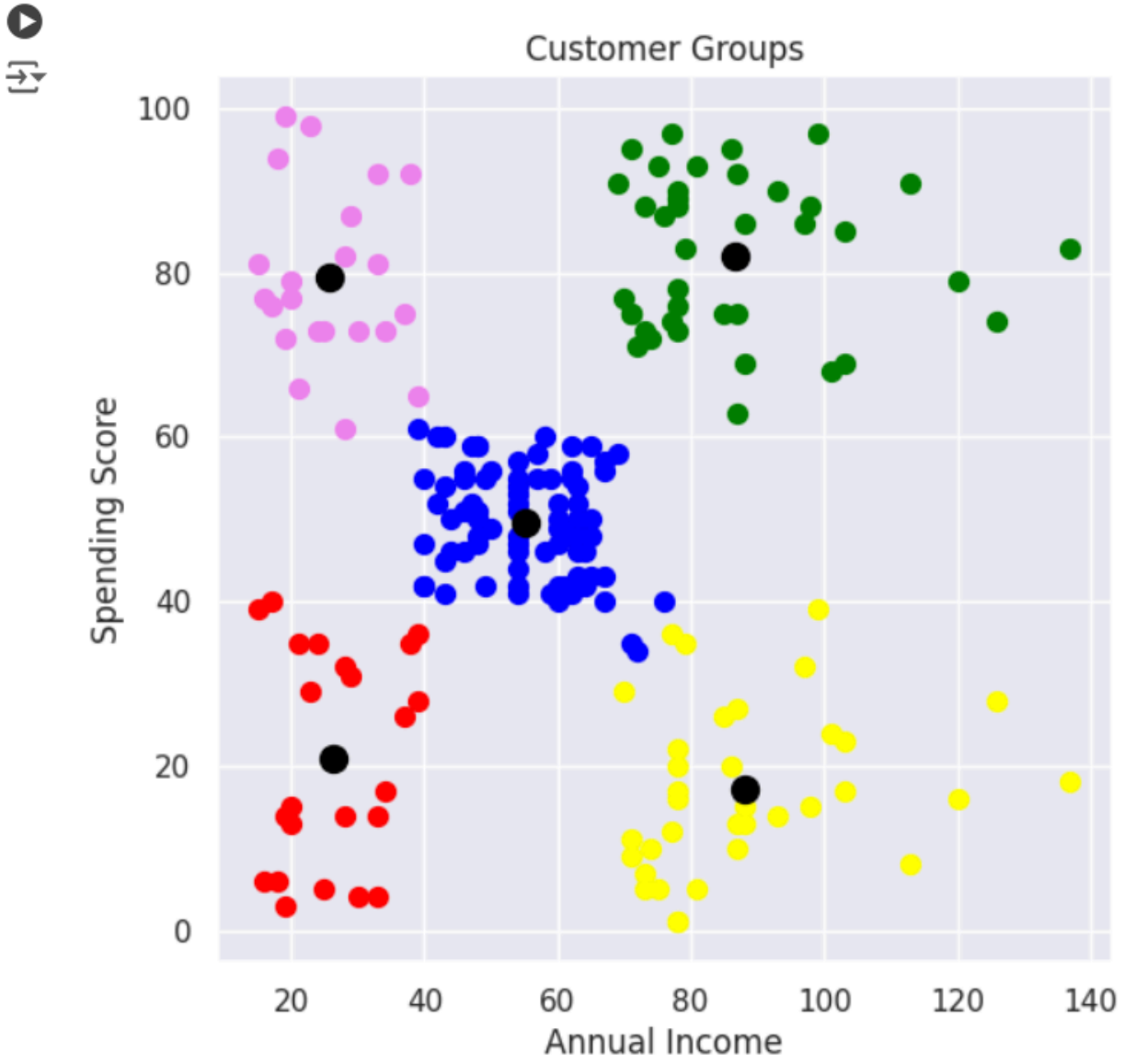
- `init='k-means++'`: This is the initialization method.
- `Y = kmeans.fit_predict(X)`: fits the KMeans model to the input data `X` (which contains the Annual Income and Spending Score values) and then predicts the cluster labels.

14. Results:

After performing analysis on selected features(Annual Income and Spending Score) using K means clustering, so based on two selected features the algorithm successfully identified 5 clusters

These 5 clusters represent different customer segments with varying characteristics and behaviour

These clusters can be used by the marketing team to optimize the customer engagement



Cluster 1:- Middle Income, Medium Spending Score
Also can be classified as “Balanced Shoppers”
Customers with average income and spending habits
Prefers seasonal sales and offers, flexible payment methods

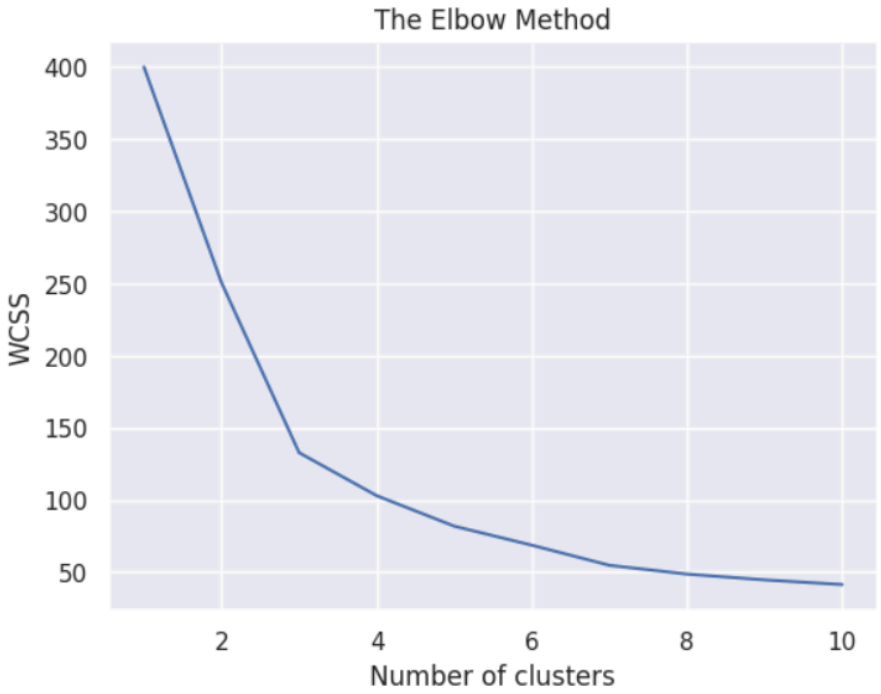
Cluster 2:- High Income, High Spending Score
Also can be classified as “Premium Shoppers”
Customers with high annual income and high spending scores.
Likely prefers premium products.

Cluster 3:- High Income, Low Spending Score
Can classify as “Budget-conscious individuals”
Customers with high income but low spending scores.
Prefer quality over quantity and need offers & sales to spend more

Cluster 4:- Low Income, Low Spending Score
Can classify as “price-sensitive customers”
Customers with low income and low spending scores.
Sensitive to discounts and promotional offers.

Cluster 5:- Low Income, High Spending Score
Customers that can be classified as “young shoppers”
Customers with lower income but high spending scores.
Customers who focus more on trendy and aesthetic items

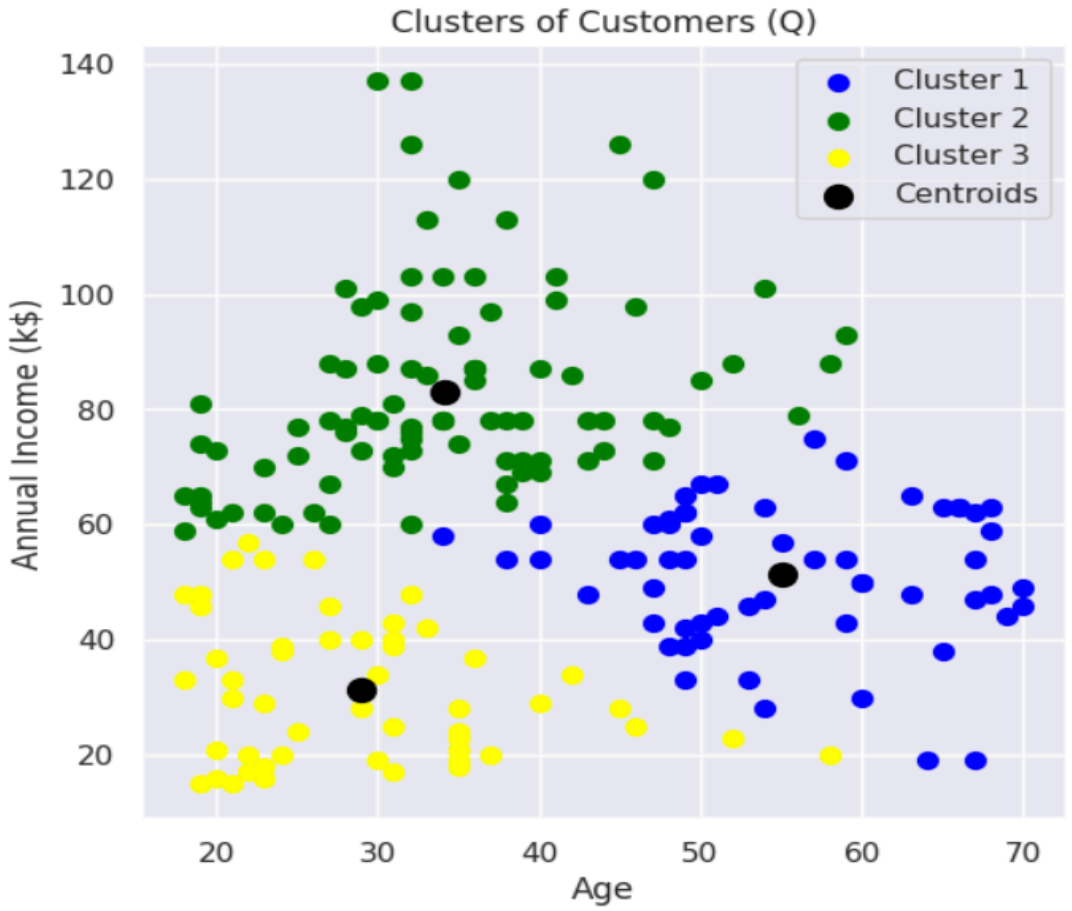
Elbow Graph of another two selected features (Age, Annual Income)



- The elbow point in the graph occurs at **k=3**, indicating that the optimal number of clusters for segmenting the mall customers based on their Age and Annual Income is 3
- we can conclude that the dataset can be effectively divided into 3 different customer segments.

Results:
After performing analysis on selected features(Age and Annual Income) using K means clustering, so based on two selected features the algorithm successfully identified 3 clusters

These 3 clusters represent different customer segments with varying characteristics and behaviour



Cluster 1: Low Income, Younger Customers
younger individuals, often students, recent graduates, etc
relatively low annual incomes

Cluster 2: Low Income, Older Customers
Individuals in later stages of their careers or retired.
with fixed incomes or those in low-wage jobs.

Cluster 3: High Income, Younger Professionals
Individuals with successful careers or entrepreneurs
more likely to spend on premium or luxury products.

Age and Spending Score (1-100):
The relation between age and spending score is a measure of purchasing behaviour, often showing the trends that provide overview on customer preferences.

Possible Clusters:

Cluster 1:- Young, High Spenders

Cluster 2:- Middle-Aged, Moderate to Low Spenders

Cluster 3:- Older, Low Spenders

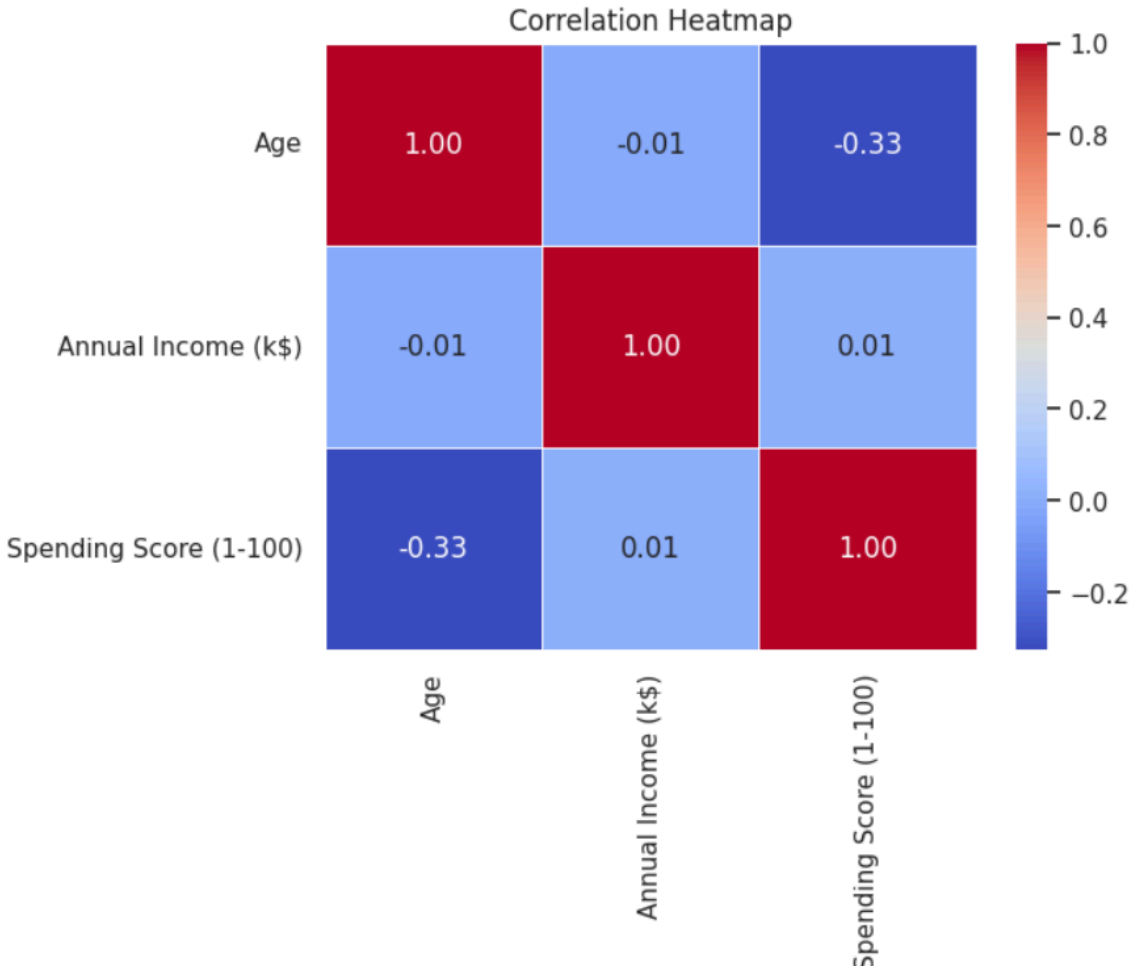
Correlation heatmap of Age, Annual Income (k$) and Spending Scores (1-100)

After performing a correlation heatmap analysis on the Mall Customer Segmentation dataset, we observed the relationships between key numerical features: Age, Annual Income (k$), and Spending Score (1-100**)**

- The heatmap visually represented these correlations, with colors ranging from blue to red. Blue indicated a negative correlation, red represented a positive correlation, and white indicated no correlation.
- And value close to 1 is positive correlation, value close to -1 is negative correlation and zero means no correlation or zero correlation.

A high correlation between Annual Income and Spending Score suggests that customers with higher incomes tend to spend more or less, depending on the value.

A negative correlation between Age and Spending Score would suggest that older customers spend less than younger ones.



Conclusion :

The Mall Customer Segmentation project successfully analyzed customer data to derive meaningful insights and classify customers in different groups based on Age, Annual Income (k$) and Spending Score (1-100). By Data preprocessing, exploratory data analytics, clustering using K Means clustering. The elbow method determined optimal number of clusters and data exploration revealed key patterns such as different age groups and gender based analysis.

These insights helps in creating targeted market strategies like offering premium products and rewards to high income customers and providing discounts to low income customers.

This project shows how data can help understand the customer behaviour better and improve business results.