EXPOSYS DATA LABS

Mall Customer Segmentation

Unlocking Insights, Enhancing Experiences

DATA SCIENCE

(One Month)

SMIT RANA

Abstract

This project explores the application of unsupervised learning, specifically K-means clustering, for customer segmentation. The primary objective is to identify distinct customer groups based on common characteristics such as gender, age, annual income, and spending habits. By categorizing customers into these segments, businesses can tailor their marketing strategies more effectively to meet the needs and preferences of each group.

The analysis begins with the application of K-means clustering to group customers into distinct segments. The clustering process utilizes key features such as age, gender, annual income, and spending scores to form meaningful clusters. The results are visualized to illustrate the distribution of customers across these segments.

Additionally, the project includes a detailed visualization and analysis of the gender and age distributions within each cluster. This helps to further understand the demographic makeup of each customer segment. The annual income and spending scores of customers are also analyzed, providing insights into the financial behaviors and purchasing power within each group. Through this approach, the project demonstrates the effectiveness of K-means clustering in uncovering actionable insights from customer data. The findings can assist businesses in devising targeted marketing strategies, improving customer satisfaction, and ultimately driving better business outcomes.

Acknowledgments

I would like to extend my sincere gratitude to Exposys Data Labs for providing me with the opportunity to work on this project as part of my Data Analyst internship. The guidance, resources, and support provided by the Exposys Data Labs team were invaluable in the successful completion of this project.

I am grateful for this internship because project has been instrumental in enhancing my understanding of customer segmentation and the application of K-means clustering in real-world scenarios.

Additionally, I would like to thank entire Exposys Data Labs team for creating a collaborative and enriching work environment. The experience and knowledge gained during this internship have significantly contributed to my professional growth and development as a data analyst.

Finally, I would like to acknowledge my family and friends for their unwavering support and encouragement throughout this internship period.

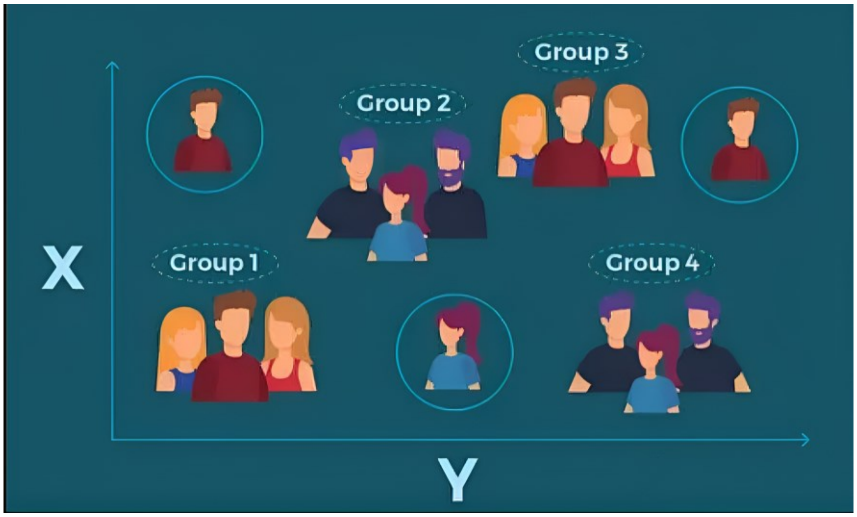
Thank you.

Table of Contents

1 Introduction

In the bustling world of retail, understanding customer behavior and preferences is essential for optimizing marketing strategies and enhancing the shopping experience. Mall operators and retailers seek to leverage data-driven insights to segment their customer base effectively, tailoring offerings and experiences to different demographic and psychographic profiles.

To address this challenge, this project aims to utilize K-means clustering, a powerful unsupervised machine learning technique, to segment mall customers based on various characteristics such as age, gender, income, and spending behavior. By employing K-means clustering, the goal is to identify distinct customer segments within the mall's diverse clientele. These segments will enable targeted marketing campaigns, personalized promotions, and optimized tenant mix, ultimately leading to a more personalized and engaging shopping experience for customers and improved business outcomes for retailers.



Customer segmentation is the process of dividing a customer base into groups of individuals who are similar in specific ways relevant to marketing, such as demographics, behaviors, preferences, or needs. The goal of segmentation is to better tailor marketing efforts and product offerings to the distinct characteristics and needs of different customer groups, ultimately improving customer satisfaction and profitability.

Customer Segmentation Applications to a Mall Setting

1. **Demographic Segmentation**: Customers are divided based on factors like age, gender, income, occupation, and family size. For example, promotions or events might target families with young children, affluent shoppers, or young professionals.

2. **Psychographic Segmentation:** Understanding shoppers' lifestyles, values, interests, and attitudes helps tailor offerings. Some may prefer upscale brands and luxury experiences, while others prioritize convenience and value.

3. **Behavioral Segmentation:** Analyzing shoppers' behavior in the mall provides insights like visit frequency, time spent, store preferences, and purchasing habits. Segments might include regulars, occasional visitors, bargain hunters, or luxury shoppers.

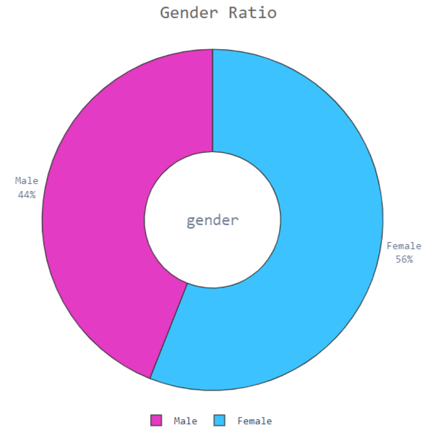
4. **Geographic Segmentation:** Location influences the customer base, considering nearby neighborhoods, cities, or regions. Understanding local demographics helps tailor marketing and store offerings.

5. **Loyalty Segmentation:** Loyalty programs segment customers based on engagement and spending. This allows for targeted promotions, rewards, and personalized communication to boost repeat visits and retention.

---

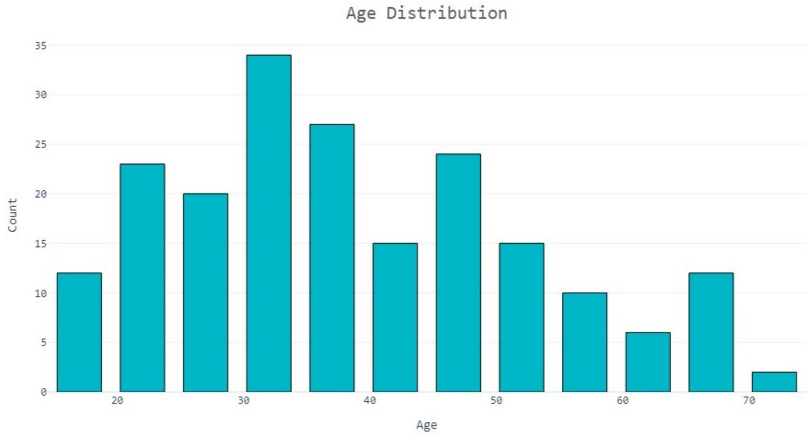
2 Data Exploration and EDAs

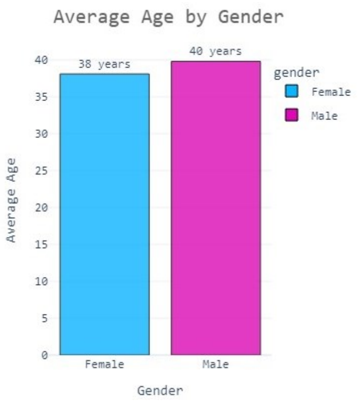
Gender Distribution



There is a slight predominance of female customers compared to male customers, with 112 females as opposed to 88 males. Females comprise 56% of the total customer base.

Age Distribution

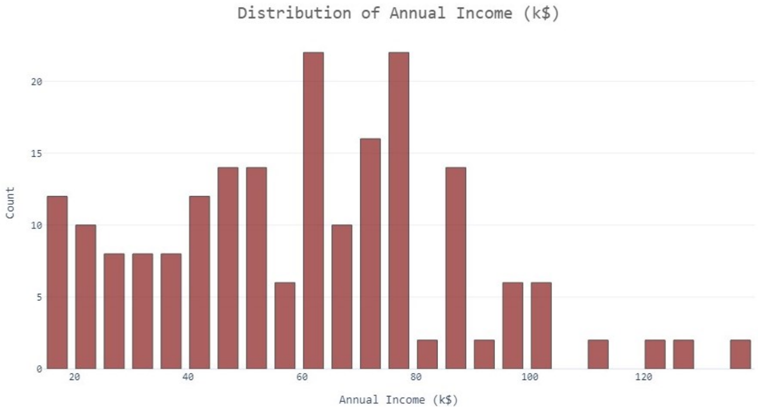




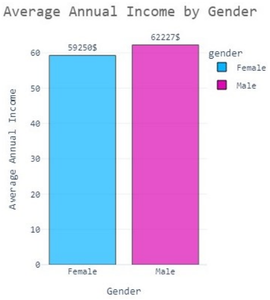
On average, male customers tend to be slightly older than female customers, with an average age of 40 years compared to 38 years for females.

The distribution of ages among male customers appears to be more evenly spread out compared to females, among whom the largest age group falls within the range of 30-35 years old.

Distribution of Annul Income



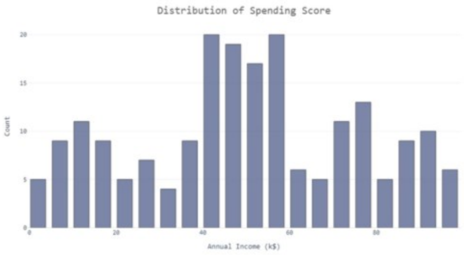
The average income for male customers is higher compared to female customers, with males having a mean income of $62,227 compared to $59,250 for females.

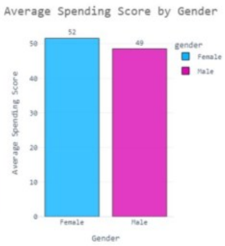


Additionally, the median income for male customers, at $62,500 is also higher than that of females, which is $60,000.

Despite these differences, the standard deviation of incomes is almost similar ($26,638 and $26,012) for both groups. Notably, there is one outlier in the male group with an annual income of approximately $137,000.

Distribution of Spending Score





The average spending score for women, at 52, appears to be higher than that of men, which stands at 49.

3 Proposed Method for Segmentation

Selected relevant features for segmentation based on three combinations of variables:

a. Age and Spending Score
b. Spending Score and Annual Income
c. Age, Spending Score, and Annual Income

4 Methodology

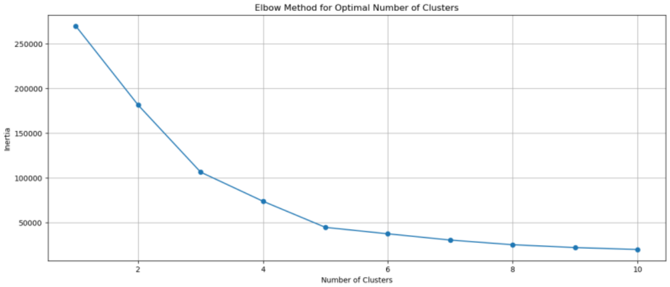
1. Data Preparation
   Extract the relevant features from the dataset
   - X1 = df[['age', 'spending_score']].values
   - X2 = df[['annual_income','spending_score']].values
   - X3 = df[['age', 'annual_income', 'spending_score']].values

2. Used Elbow method to find optimal Numbers of Clusters
3. K-means Clustering:
   - Define the K-means algorithm with specified parameters (e.g., number of clusters, initialization method, maximum iterations).
   - Fit the algorithm to the data to obtain cluster labels and centroids.
4. Visualization:
   - Create a meshgrid to visualize the decision boundaries of the clusters.
   - Predict cluster labels for each point in the meshgrid.
   - Plot the clusters using imshow for the decision boundaries and scatter plot for the data points colored by cluster labels.
   - Display centroids as red points.

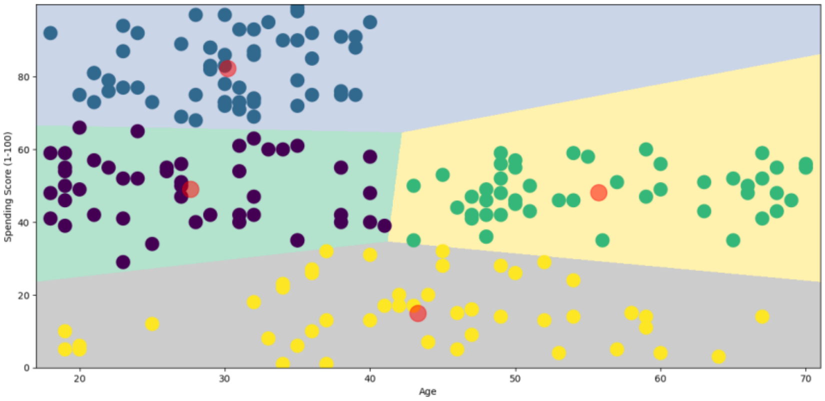
Overall, this method helps visualize the segmentation of mall customers based on feature selection, identifying distinct clusters and centroids to understand customer behavior patterns.

5 Implementation

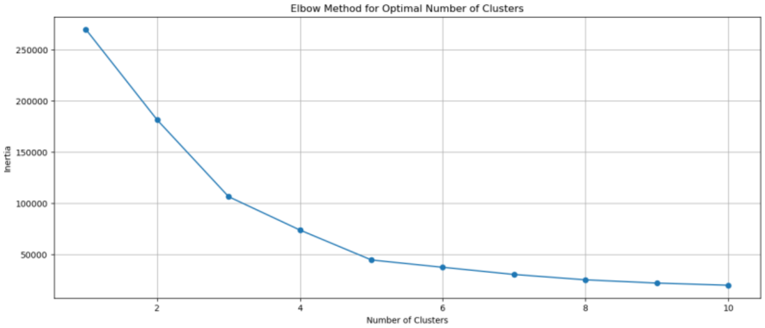
Age and Spending Score: Utilized a segmentation approach based on age and spending score to identify clusters of customers with similar age profiles and spending behavior.



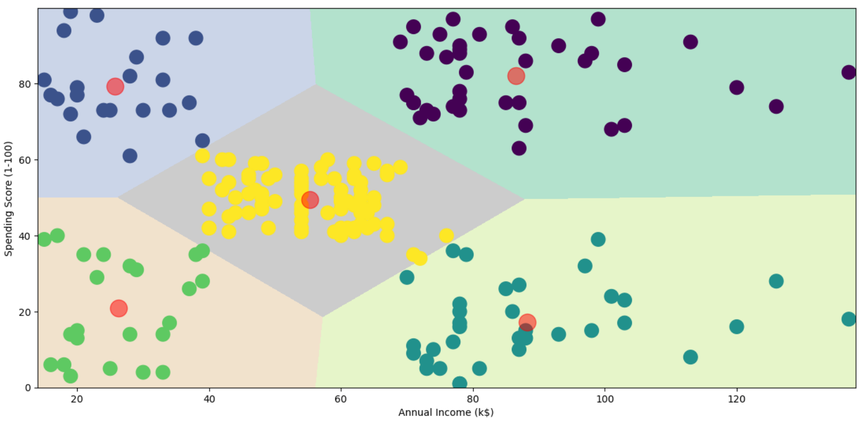
When observing the plot, look for a juncture where the decrease in inertia slows down notably, typically forming a bend known as the "elbow". The rationale is that after this bend, introducing additional clusters doesn't substantially reduce the inertia. Hence, in this case, **cluster=4** would be the optimal choice.



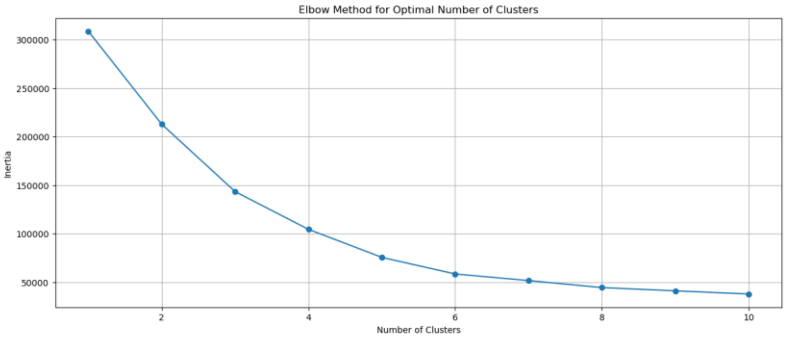
Spending Score and Annual Income: Employed segmentation techniques using spending score and annual income to categorize customers based on their purchasing power and spending habits.



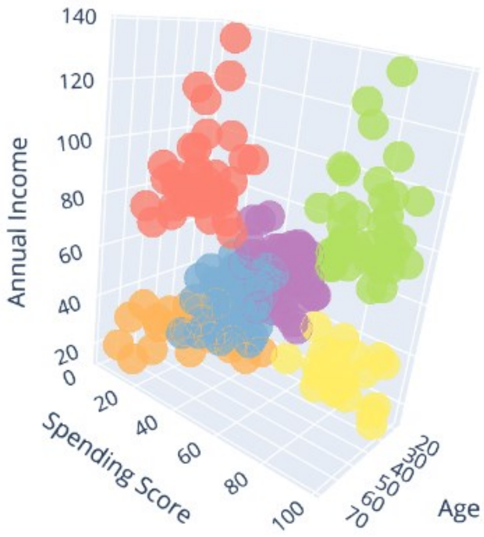
In the second segmentation, pinpoint the spot where the inertia's descent becomes less pronounced is 5, Therefore, selecting **cluster=5** would be most appropriate.



Age, Spending Score, and Annual Income: Integrated all three variables (age, spending score, and annual income) to perform comprehensive customer segmentation, considering both demographic and behavioral **aspects.**



In this third plot, point 6 is the juncture where the inertia reduction rate becomes less steep, Consequently, for this scenario, selecting cluster=6 would be ideal.



6 Conclusions

In this analysis, segmentation was performed on three combinations of variables:

1. age and spending_score

For the combination of spending_score and annual_income, the K-Means algorithm identified five distinct clusters:

- Customers with lower annual incomes (20-40) but higher spending scores (60 or more).
- Customers with moderate annual incomes (40-60) and moderate spending scores (40-60).
- Customers with higher annual incomes (80 or more) but lower spending scores (less than 40).
- Customers with higher annual incomes (80 or more) and higher spending scores (60 or more).
- Customers with lower annual incomes (80 or more) and lower spending scores (less than 40).

2. spending_score and annual_income

However, no clear distinctions based on age were observed among these groups (for the age and spending_score combination).

3. age, spending_score, and annual_income

Considering customer spending habits, the company could target potential members by focusing on the clusters with higher spending scores (above 60) identified by the K-Means model. These customers are likely more receptive to membership programs.

Notably, these clusters predominantly include customers under the age of 40, suggesting that the marketing strategy for the membership program could be tailored toward younger customers with higher spending scores.

For customers with lower spending scores and aged above 40, strategies involving popular products and promotions could be effective in engaging them further.

To gain deeper insights into customer preferences, future analyses could incorporate additional data such as purchase frequency and types of purchases, allowing for more customized product offerings tailored to each segment.

References

[1] https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

[2] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[3] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[4] https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/