

Customer Segmentation Analysis: Clustering and Fitting

Student Name: Mohammed Abdul Jilani

Student Number: 24168848

Course: Applied Data Science 1

Instructor: Dr. William Cooper

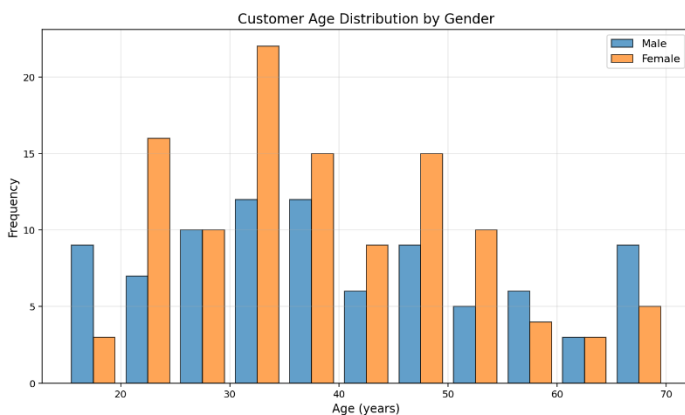
Date: November 2025

I. INTRODUCTION

Customer segmentation is a basic operation of retail analytics that would assist businesses determine the behaviour of their customers and use the knowledge to develop marketing strategies. This paper provides an analysis of Mall Customer Segmentation Dataset, which has 200 customers and the feature of age, gender, annual income, and spending score. The aim is to determine the different customer groups with the use of K-mean clustering and find out the dependency between income and spending behaviour using the use of a poly-curve fitting. The importance of this analysis is to design specific marketing campaigns and efficient allocation of resources.

II. AGE DISTRIBUTION ANALYSIS

Age distribution of the customers is also analysed using a histogram as divided by gender which gives demographic trends that are used in marketing strategy. Female customers represent slightly higher percentage of the population (56%), and are concentrated in younger age groups (20-35 years), whereas male customers (44) are more evenly divided in the age groups. The data has the means of age as 38.85 years and standard deviation of 13.97 years. The skewness of 0.49 shows that the distribution is moderately skewed to the right with kurtosis of -0.10, which shows that the distribution is quite flat. Such indicators imply that there are numerous age group categories that have to be approached differently.



III. INCOME AND SPENDING RELATIONSHIP

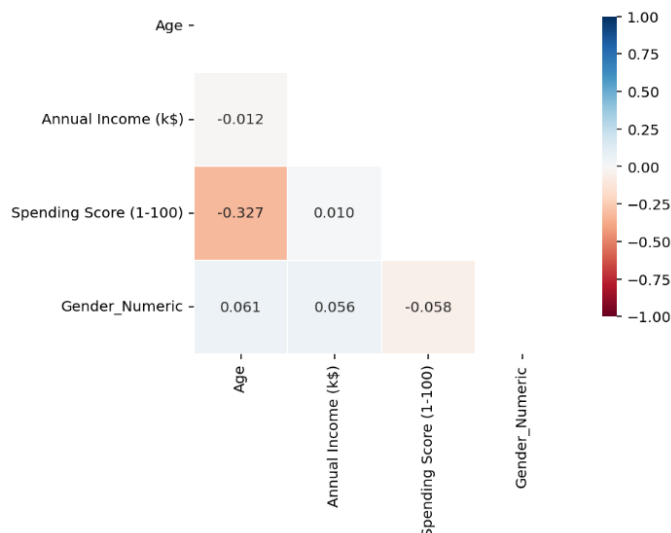
A correlation of annual income and spending scores revealed that there is almost no correlation between the two ($r = 0.009$), that is income is not a strong predictor of customer spending. Four natural quadrants of the scatter plot depict four types of behaviours: low-income/ low-spending, low-income/ high-spending, high-income/ low-spending and high-income/ high-spending. The customers who are females are rated slightly higher on the spending score than the male customers though both gender is spread out in all the quadrants. This complexity underscores the fact that multi-dimensional segmentation should be adopted as opposed to income only.



IV. STATISTICAL RELATIONSHIPS

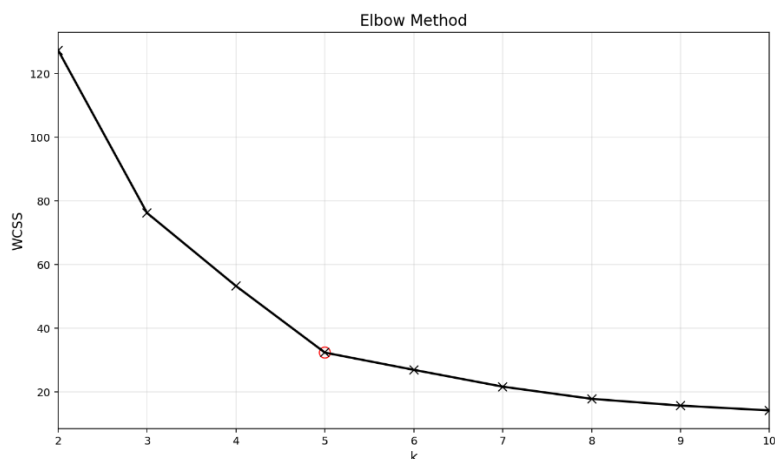
A correlation heatmap helps to give information about the correlations between variables. Age and spending score have a weak negative correlation (-0.327) and income and spending score are effectively unrelated (0.009). Gender only has a slight relationship with other characteristics. These poor bivariate associations warrant the adoption of clustering methods which are able to represent more intricate patterns.

Correlation Matrix of Customer Features



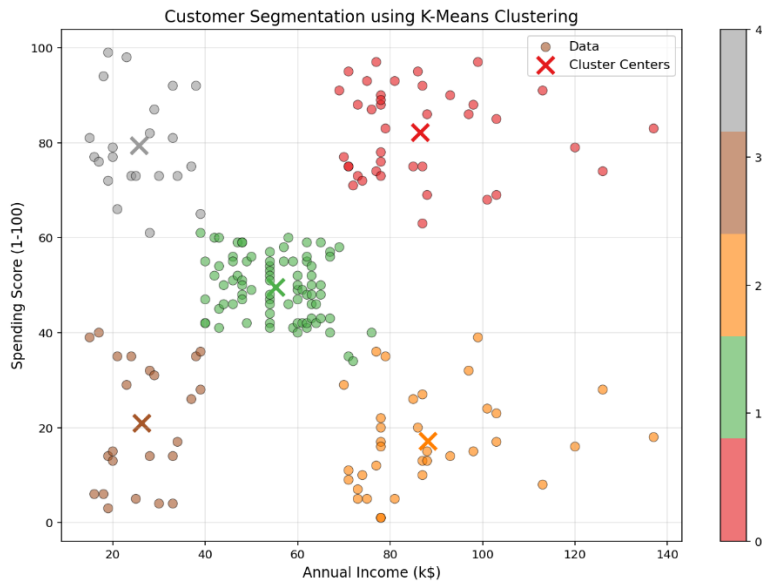
V. OPTIMAL CLUSTER DETERMINATION

The number of best clusters to use in K-means was identified using the elbow technique and silhouette analysis. Elbow plot shows that the within-cluster variance reduction is diminishing after the 5th cluster, and the silhouette score is highest at the 5th cluster, which means that the clusters are well-separated and internally cohesive. These findings offer strong reasons as to why five clusters were chosen.



VI. CUSTOMER SEGMENTATION RESULTS

K-means clustering with five clusters reveals distinct customer groups:

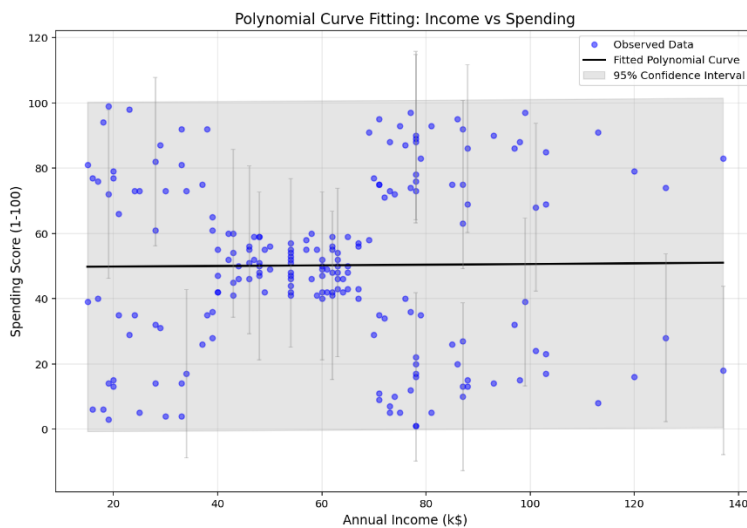


VII. POLYNOMIAL CURVE FITTING ANALYSIS

A second-degree regression was estimated to determine the relationship between income (x) and spending score (y):

$$[y = 0.0007x^2 - 0.1072x + 54.87]$$

The quadratic term implies a U-shaped change that there is initially a negative relation between spending and income, and the further the income the positive the relation. Standard deviation (residual) = 25.32, and standard deviation confidence interval values lie within a wide range of about ± 49.6 indicating a high degree of variability in individual customers.



VIII. DISCUSSION

As it is shown by the analysis, income alone fails to predict spending behaviour and five clusters are effective in grouping the customer base into actionable groups. The broad trends of incomes-spending are captured in the model of the polygons yet there is high variability in each individual, which supports the use of multi-dimensional clustering as an efficient method of strategic decision-making. It is possible to attach predicted points to clusters to show predictive applicability of the clusters to unseen customers.

IX. CONCLUSIONS

This report illustrates the application of K-means clustering alongside regression to conduct customer segmentation and expenditure analysis. Main findings comprise:

- It identified five distinctive customer groups that were verified using silhouette and elbow analyses.
- Income and spending score exhibit a linear relationship highlighting the impact of multiple factors.
- A pattern shows complex connections between income and spending, revealed using the polynomial fitting method.