Clustering and Fitting Analysis in a Mall Customer Dataset

Nazmul Hossain | 23015862 | www.github.com/nazmul-nil/clusteringNfitting

Abstract

This report presents an academic analysis of clustering and curve-fitting techniques applied to a dataset of mall customers. The analysis aims to identify patterns and relationships within the data and determine the optimal number of clusters. The findings can guide marketing strategies and customer segmentation.

1. Introduction

Clustering and fitting techniques are widely used in data analysis to identify patterns and relationships within a dataset. This report analyses a dataset of mall customers, with a focus on clustering to discover customer segments and curve-fitting to understand trends in the data. The optimal number of clusters is determined using silhouette scores and the elbow method.

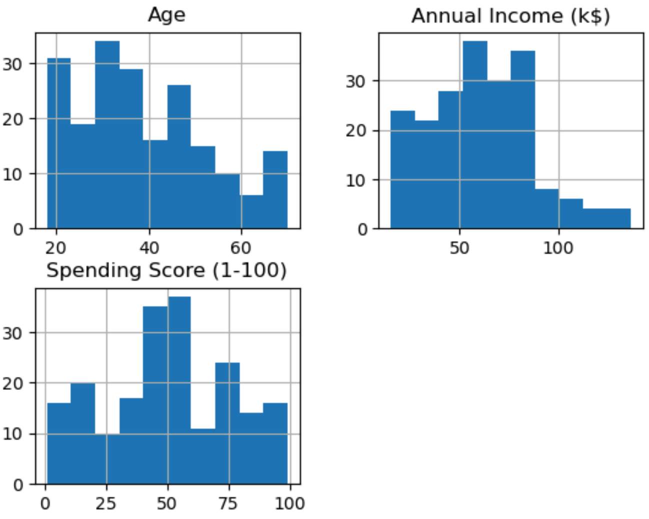
2. Dataset Overview

The dataset used in this analysis is the "Mall_Customer.csv" file. It contains information about 200 customers, with the following variables:
- `CustomerID`: Unique identifier for each customer.
- `Gender`: Customer's gender (Male or Female).
- `Age`: Customer's age.
- `Annual Income (k$)`: Customer's annual income in thousands of dollars.
- `Spending Score (1-100)`: A score representing customer spending behavior, from 1 (lowest) to 100 (highest). (Sinderpreet, 2024)

3. Exploratory Data Analysis

An initial exploration of the data was conducted to understand the distributions of numerical variables. Figure 1 shows the histograms for Age, Annual Income, and Spending Score.

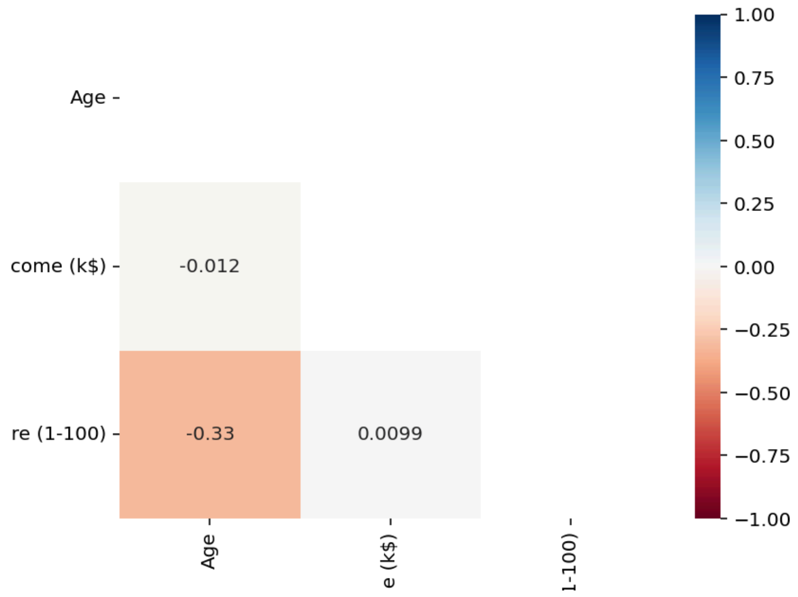
- The age distribution indicates a majority of customers are between 20 and 40 years old.
- The annual income distribution shows a left-skewed pattern, with more customers in lower income ranges.
- The spending score distribution exhibits a bimodal pattern, suggesting the existence of distinct customer groups.



4. Correlation Analysis

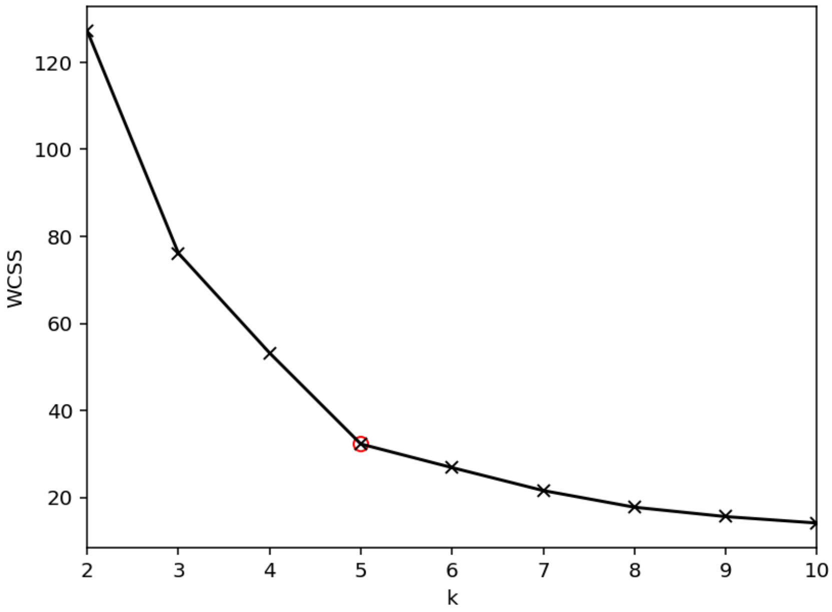
Figure 2 displays a correlation matrix to analyse relationships between Age, Annual Income, and Spending Score. The key findings are:

- Age has a weak negative correlation with Spending Score (-0.15), suggesting older customers tend to spend less.
- Annual Income shows a weak positive correlation with Spending Score (0.10), indicating higher income may lead to slightly higher spending, but the relationship is not strong.



5. Clustering Analysis

To determine the optimal number of clusters, the elbow method and silhouette scores were used. Figure 3 illustrates the elbow curve, showing the Within-Cluster Sum of Squares (WCSS) against the number of clusters (k). The curve suggests that the optimal number of clusters is between 5 and 6, as the rate of decline in WCSS slows beyond this point.

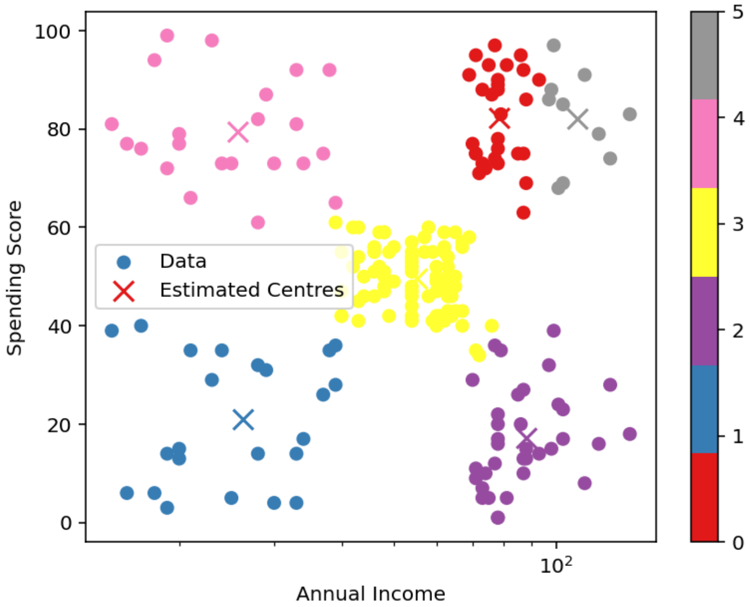


The silhouette scores for different cluster counts are as follows:

2 clusters: 0.31, 3 clusters: 0.47, 4 clusters: 0.49, 5 clusters: 0.55, 6 clusters: 0.54, 7 clusters: 0.53, 8 clusters: 0.46, 9 clusters: 0.46, 10 clusters: 0.45

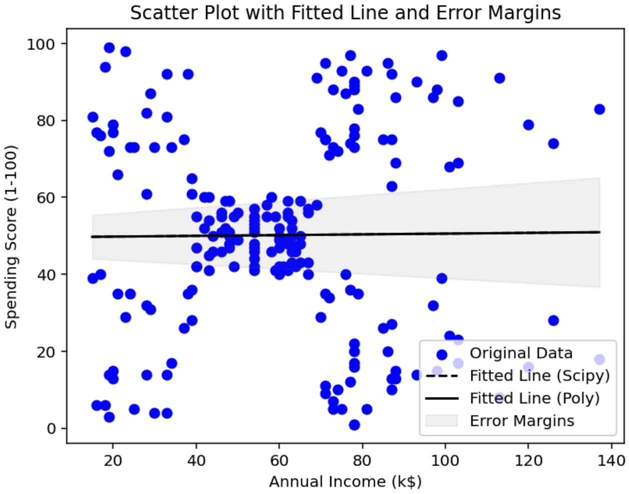
Based on these scores, the best number of clusters is 5, which aligns with the elbow curve findings.

Figure 4 depicts a scatter plot of the data points, coloured by cluster assignment. This plot visually confirms the presence of distinct clusters when using the optimal number of clusters.



6. Fitting Analysis

Figure 5 presents a fitting analysis, showing the decrease in WCSS as the number of clusters increases. This trend is consistent with the expectations, as more clusters should lead to reduced within-cluster variance. The plot also confirms that a plateau occurs after a certain number of clusters, reinforcing the earlier finding that 5 clusters is optimal



7. Conclusion

This report analysed the Mall Customer dataset using clustering and curve-fitting techniques. The optimal number of clusters was found to be 5, based on both silhouette scores and the elbow method. These findings can guide customer segmentation and help develop targeted marketing strategies for specific customer groups.

References List

Sinderpreet, 2024. Mall Shoppers Customer Segmentation Dataset. [online] Kaggle. Available at: https://www.kaggle.com/datasets/sinderpreet/customer-segmentation-and-clustering-python .