Title: Customer Data Analysis Report

Introduction

A key tactic for developing focused marketing campaigns and raising customer satisfaction in context of retail marketing is division of consumers into discrete groups according to their purchasing habits. To understand intricate relationship between annual income and spending score proxy for consumer purchasing habits we go into dataset from shopping center. We attempted to group consumers into useful segments by using K-Means clustering by which revealed trends that might help with marketing campaign customization and predictive association between customers income and spending propensity was investigated using linear regression.

Data Overview

Customers gender, age, yearly income and expenditure score are all included in dataset. To find patterns in spending behavior of customers data are first explored by looking at age distribution and correlation between variables and use of clustering techniques.

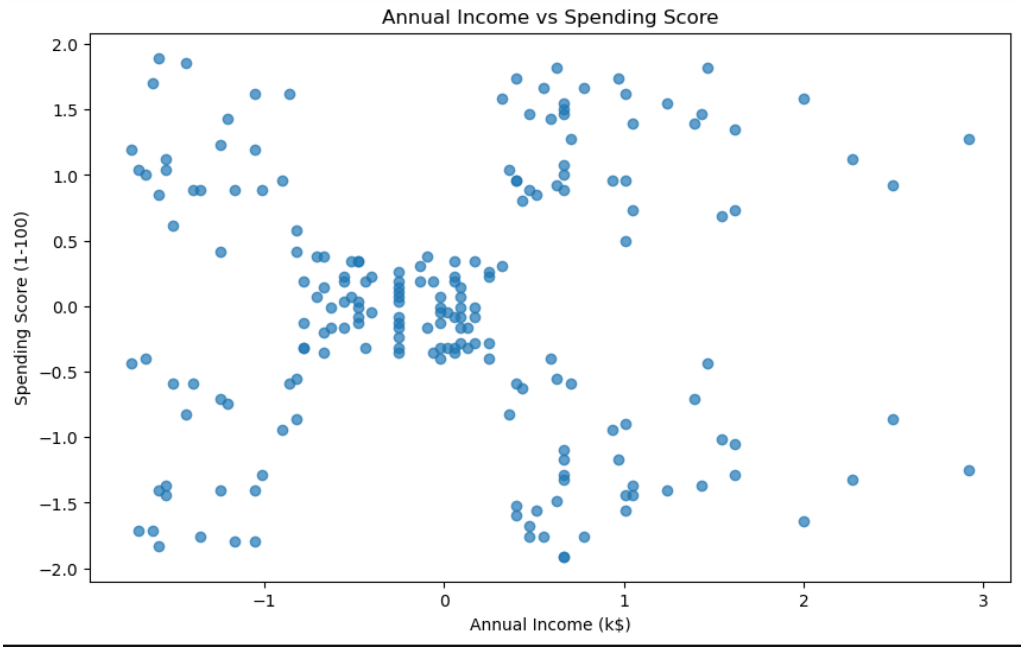
	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Data Preprocessing

Data In order to normalize size of numerical features, including age, annual income and spending score data underwent thorough standardization process. To put every feature on an even playing field conduct fair comparison and guarantee objectivity of analysis findings and standardization is essential. It reduces possibility of overemphasizing some aspects and lessons impact of outliers giving clearer and by providing more accurate picture of underlying trends.

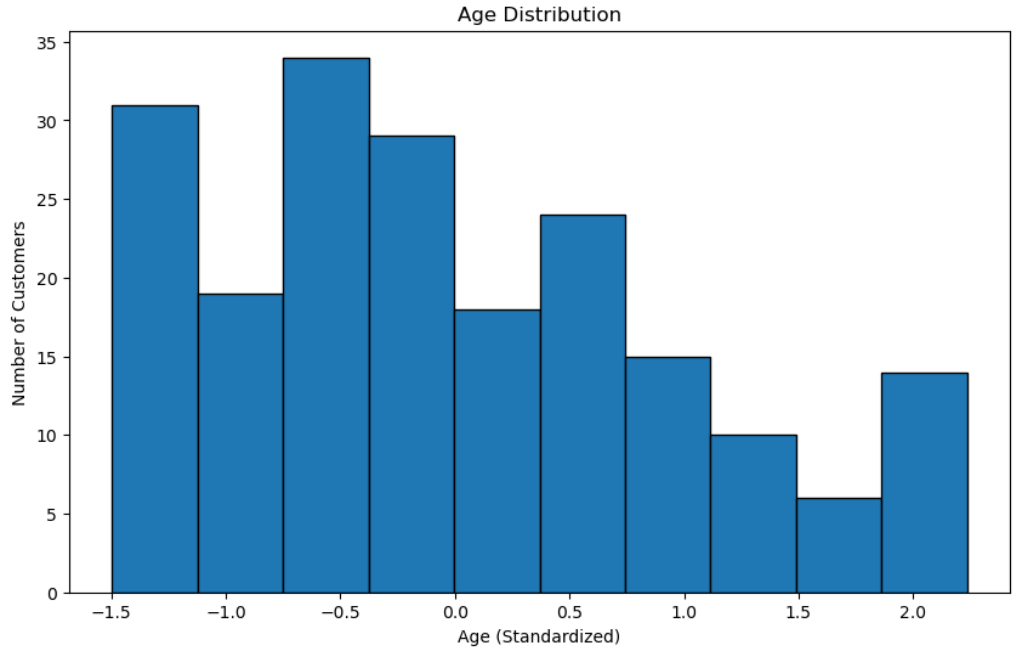
Exploratory Data Analysis

Annual Income vs. Spending Score



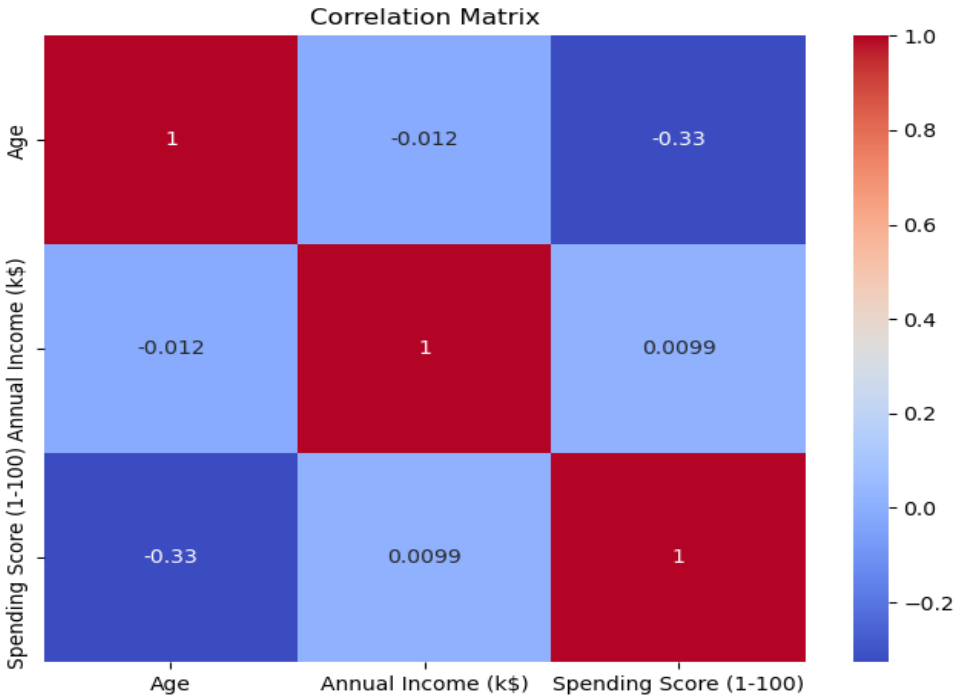
The annual income and spending score do not exhibit an obvious linear relationship when plotted against each other. Within data possible groupings or clusters are shown by point distribution.

Age Distribution



The customer age distribution histogram shows nearly normal distribution suggesting that customer base is made up of variety of age groups.

Correlation Matrix



The correlation matrix shows that there is almost no association between annual income and spending score and very minor negative correlation between age and spending score. This shows that spending habits may be slightly influenced by age.

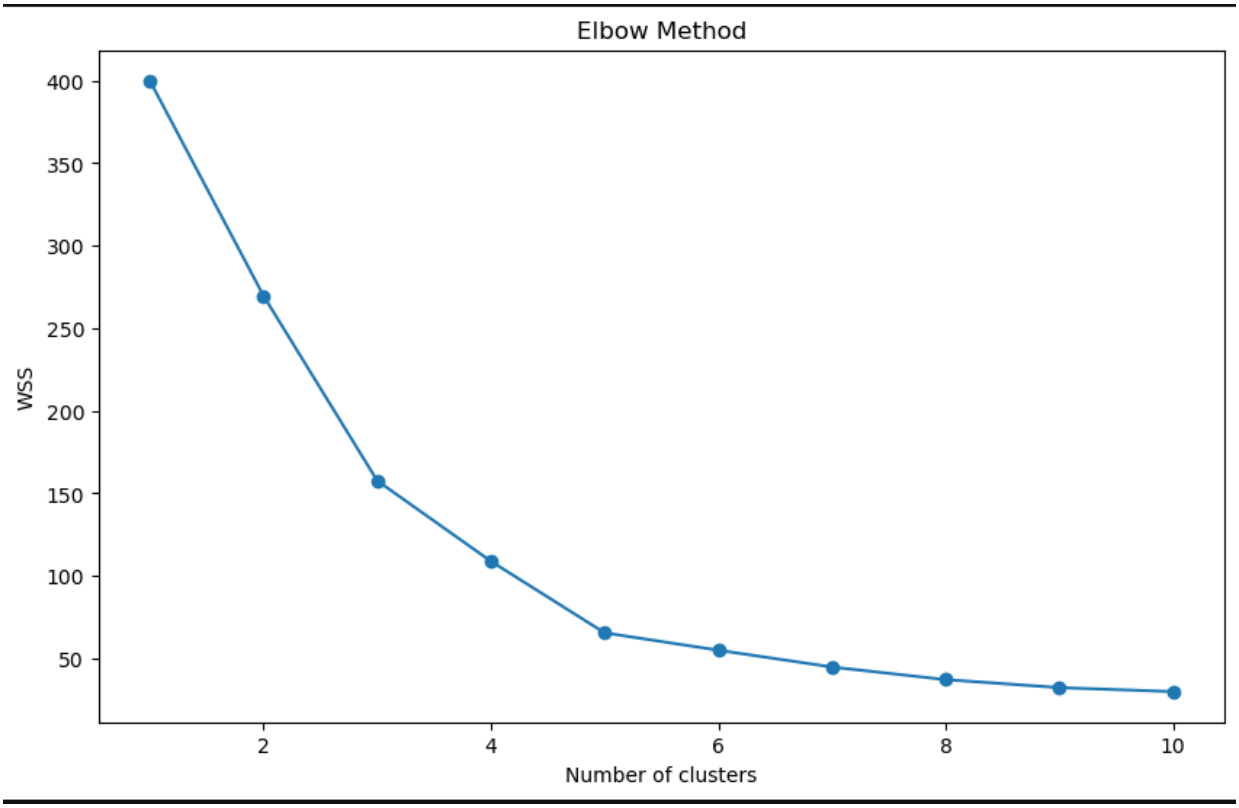
Statistical Analysis and Correlation

```
Descriptive Statistics:
                  Age  Annual Income (k$)  Spending Score (1-100)
count     2.000000e+02        2.000000e+02            2.000000e+02
mean     -1.021405e-16       -2.131628e-16           -1.465494e-16
std       1.002509e+00        1.002509e+00            1.002509e+00
min      -1.496335e+00       -1.738999e+00           -1.910021e+00
25%      -7.248436e-01       -7.275093e-01           -5.997931e-01
50%      -2.045351e-01        3.587926e-02           -7.764312e-03
75%       7.284319e-01        6.656748e-01            8.851316e-01
max       2.235532e+00        2.917671e+00            1.894492e+00
skewness  4.855689e-01        3.218425e-01           -4.722020e-02
kurtosis -6.715729e-01       -9.848709e-02           -8.266291e-01

Correlation Matrix:
                           Age  Annual Income (k$)  Spending Score (1-100)
Age                   1.000000           -0.012398               -0.327227
Annual Income (k$)   -0.012398            1.000000                0.009903
Spending Score (1-100) -0.327227           0.009903                1.000000
```

Important details about client profile are revealed by statistical analysis of dataset. Upon standardizing data we discover that every variable has an effective mean of zero guaranteeing cross category comparability. Standard deviations close to one show that variability is constant. The dataset appears to contain higher proportion of younger customers based on skewness and kurtosis metrics. And distributions of expenditure and income scores show no evidence of large tails. While there is no significant linear association between income and spending score correlation analysis indicates minor negative relationship between age and expenditure score suggesting that spending declines with age. These results demonstrate need for more complicated models to effectively represent distinctions of consumer behavior.

Clustering Analysis



Silhouette Score

- Silhouette Score for 2 clusters: 0.3212707813918878
- Silhouette Score for 3 clusters: 0.46658474419000145
- Silhouette Score for 4 clusters: 0.4939069237513199
- Silhouette Score for 5 clusters: 0.5546571631111091
- Silhouette Score for 6 clusters: 0.5398800926790663

- Silhouette Score for 7 clusters: 0.5281492781108291
- Silhouette Score for 8 clusters: 0.4552147906587443
- Silhouette Score for 9 clusters: 0.4570853966942764
- Silhouette Score for 10 clusters: 0.4431713026508046

The customer data was segmented using K-Means technique. The elbow technique indicates bend at five clusters suggesting ideal ratio of within cluster sum of squares WSS to number of clusters. Five clusters were selected based on silhouette analysis with greatest silhouette score for these clusters being roughly 0.5546.

Clear segmentation is evident in visualization of clustered data indicating different customer groups according to their spending and annual income scores.

Regression Analysis

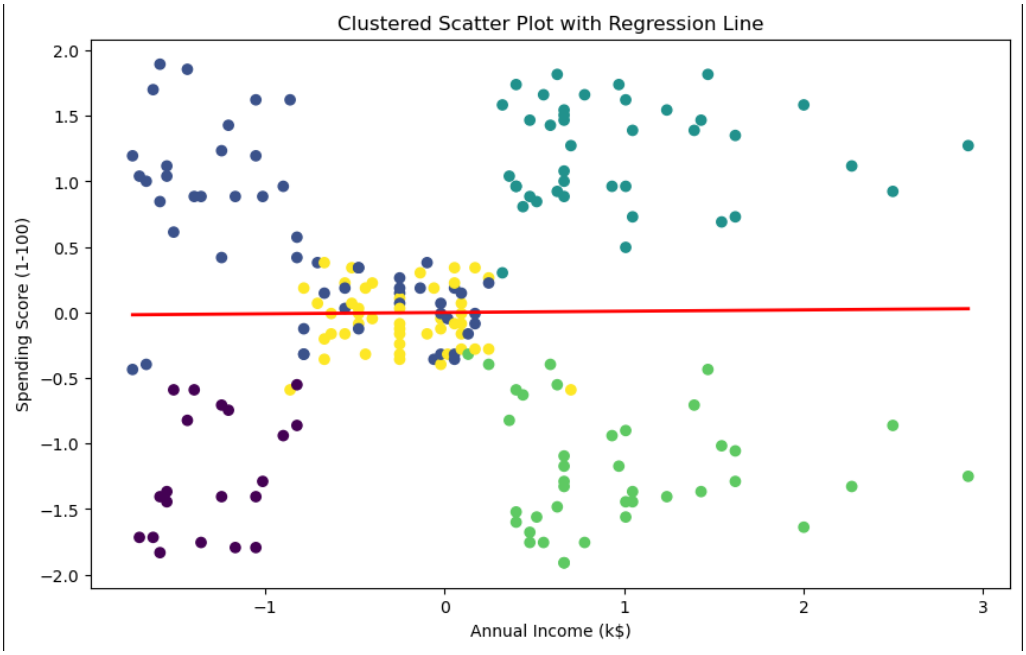
Evaluation Metrics

Mean Absolute Error (MAE): 0.81

Mean Squared Error (MSE): 1.00

Root Mean Squared Error (RMSE): 1.00

These numbers suggest that there is an average difference of 0.81 standard units between models predictions and actual spending scores. An RMSE of 1.00 is relatively high indicating that model's predictions are not very accurate given standardized nature of dataset.



To investigate association between annual income and spending score a linear regression analysis was conducted. On other hand model does not explain any variability in spending score from annual income as indicated by R-squared value of 0.00. The correlation matrix which likewise showed nonlinear connection that is consistent with this conclusion. The poor prediction ability of annual income on expenditure score is further illustrated by regression line superimposed on clustered scatter plot.

Conclusions

The smart clustering analysis identified discrete consumer base groupings that can be targeted using various tactics. Regression analysis however was unable to produce solid model for forecasting consumer spending scores solely based on yearly income. This suggests that spending behavior may be influenced by other variables not included in model.

Incorporating other data like client demographics or purchase history could enhance future analysis and provide more reliable model of customer spending. Customizing marketing and service offerings to identified client categories and their preferences is goal.

References

Subramanian, R. & Dhandayudam, Prabha & Maheswari, B. & Aswini, J.. (2021). Customer Analysis Using Machine Learning Algorithms: A Case Study Using Banking Consumer Dataset. 10.3233/APC210263.

Sun, Y. & Liu, H. & Gao, Y.. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. 10.1016/j.heliyon.2023.e13384.

Choudhury, Adil & Nur, Kamruddin. (2019). A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior. 10.1109/ICREST.2019.8644458.

Gupta, Er & Gupta, Er & Mishra, Amit. (2012). RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS.