NAME: AGBELEYE VICTOR

USERNAME: VA23ABB

COURSE TITLE: APPLIED DATA SCIENCE

COURSE CODE: 7PAM2000-0105-2023

STUDENT NUMBER: 22079387

GITHUB LINK: https://github.com/AgVicCodes/clustering_analysis

MALL CUSTOMERS DATASET ANALYSIS REPORT

INTRODUCTION

The dataset comprises information stored on customer membership cards. The data comprises the customer Id, age, gender, annual income and spending score - based on purchasing data.

DATA EXPLORATION

The cleaned and prepared dataframe has 200 rows and 5 columns with the following data types:

customer_id:	int64
age:	int64
annual_income:	int64
spending_score:	int64
total_amount:	float64
gender_class:	category
gender:	object
age_group:	category

The dataframe has no duplicate or non-available (nan) values. Some columns were added to the dataframe. The total amount signifies the estimated sum of money spent by each customer as their spending score is based on purchasing data and spending habits. Gender class refers to numbers assigned to different genders – 1 for males and 2 for females. Age group is an interval that classifies customers by their age brackets, from 10 – 19, 20 – 29,… to 70+.

MOMENTS

The average age in the dataframe is 39, and the standard deviation for age is 16.07, inferring that the distribution has a large width and the ages are distributed largely around the mean. The distribution is positively skewed as the skewness is 0.31, and the Kurtosis of the age distribution is -1.13 – the age distribution is platykurtic (Broad peak).

The average income is $45172, the standard deviation for income is $15721, and the salary distribution also has a large width. The distribution is negatively skewed as its skewness is -0.39. The kurtosis of the income distribution is -1.07 therefore, the distribution is also platykurtic.

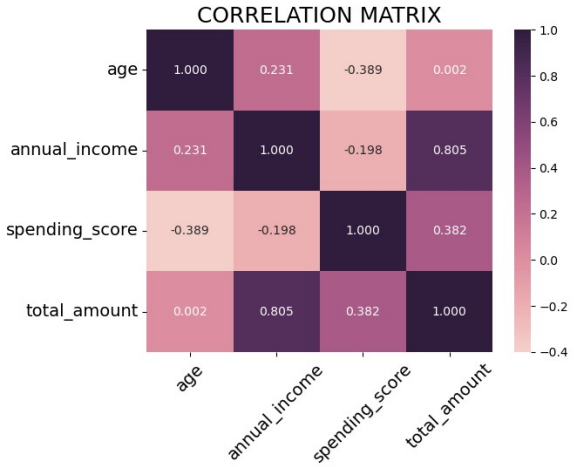
The average spending score is 52, and the standard deviation for the spending score is 17 which also infers a wide distribution. The skewness of the spending score distribution is 0.36 inferring that this distribution is almost symmetric. The Kurtosis of the spending score distribution is 0.73 which signifies a leptokurtic distribution (Sharp peak).

VISUALISATION

I created a correlation matrix using the default method (Pearson's) to show the relationships between all the numerical columns.

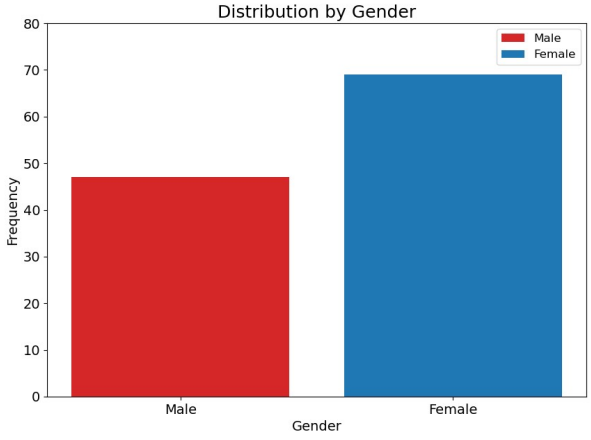
From the pairplot below, we can confirm the structure of the distributions as highlighted in the moments. We immediately notice some correlations and dissimilarities. I would like to perform cluster analysis on the annual income against spending score scatter plot to segregate our customers into different classes or categories (clusters) and fit the annual income

against the total amount spent per annum to be able to predict the amount a person will likely spend based on his income. The correlation plot is provided here:



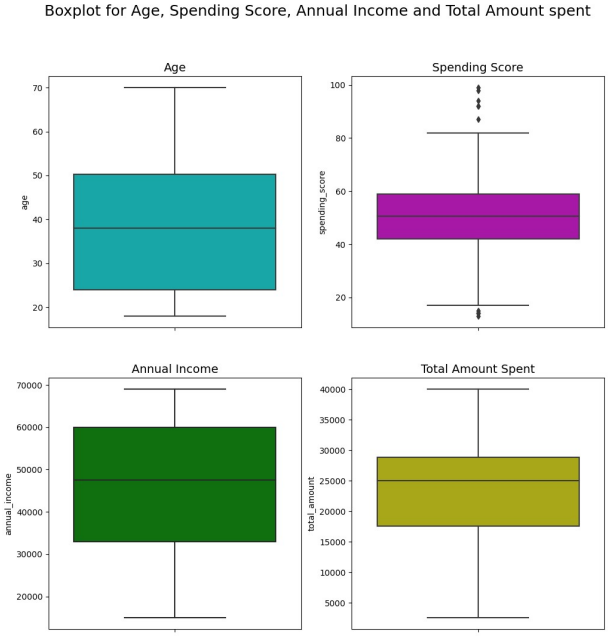
And the corresponding pairgrid below:

Before going into the data clustering and fitting, a glance at our data is necessary.



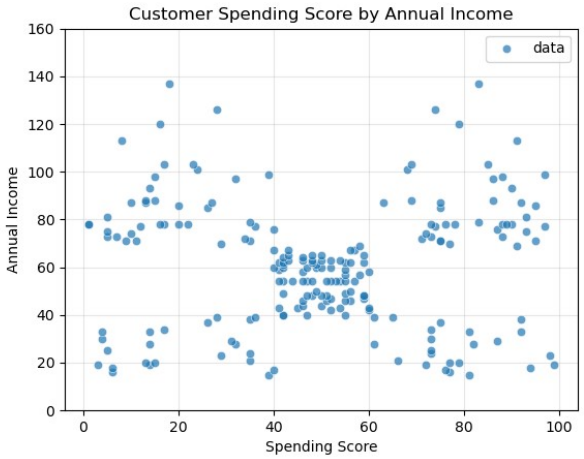
As we see from above there are more women (69) than men (47) in our distribution.

Now let's detect potential outliers in our dataframe.



From the above boxplots, we notice some outliers among the spending score, but they are not necessarily bad as the spending score can only range from 0 to 100.

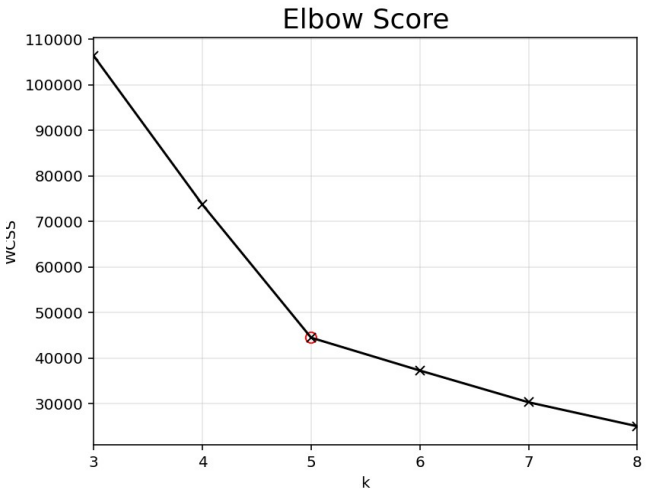
To start clustering, we will need to take an initial guess for the number of clusters.



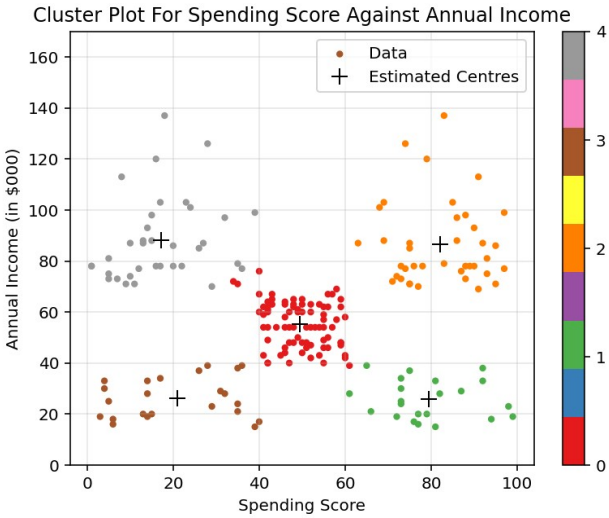
From the above plot, by eye, I would guess 5 clusters. Let's check the result of the silhouette score.

3 clusters silhouette score = 0.47
4 clusters silhouette score = 0.49
5 clusters silhouette score = 0.55

6 clusters silhouette score = 0.54
7 clusters silhouette score = 0.53
8 clusters silhouette score = 0.45



The elbow score also shows that 5 is the ideal number of clusters.



From the diagram above, we have 5 separate data clusters and their respective midpoints are shown by “o” and “+” respectively.

The **brown** cluster indicates our **economical customers.** They have low incomes and spend their money accordingly.

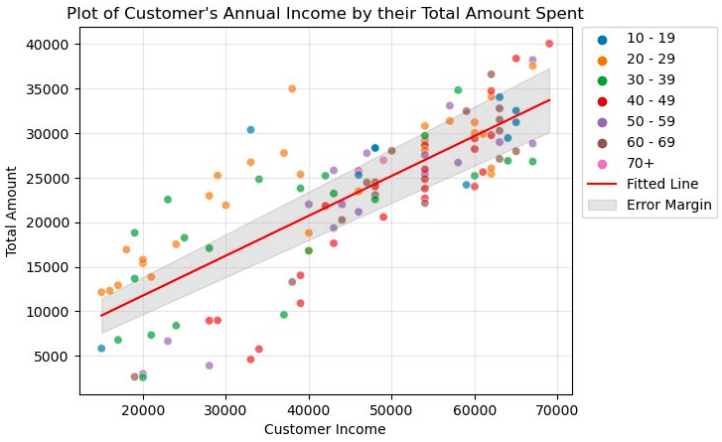
The **green** cluster indicates our **extravagant customers**. Although they have little income also, they spend a lot. These people most likely have shopping addictions.

The **red** cluster indicates our **normal customers** with average income and average spending scores.

The **grey** cluster indicates the **roamers.** Although they have very high incomes, they seem to spend less due to factors like customer satisfaction or something else.

And lastly, the **orange** cluster indicates our **loyal customers** who have high incomes and also high spending scores.

For the fitting, after testing multiple fitting functions, the best fit was a linear fit: y = ax + b. Anything else was overfitting.



References

CHOUDHARY, VIJAY. 2018. *Mall Customer Segmentation Data.* March 16. Accessed March 22, 2024. https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python.