Name: Ojo Covenant Oludapomola

Student id: 23013682

Course name and number: Applied Data Science 1, 7PAM2000

Github: https://github.com/Covpet/Clustering-and-Fitting

INTRODUCTION

The dataset used in this analysis is derived from a study conducted at a mall to understand the shopping behaviors and preferences of its customers. The dataset contains information about various demographic and behavioral attributes of customers. These attributes include:

Customer ID: A unique identifier assigned to each customer.

Gender: The gender of the customer categorized as male or female.

Age: The age of the customer, indicating their age at the time of data collection.

Annual Income (k$): The annual income of the customer, measured in thousands of dollars.

Spending Score: A score assigned to each customer based on their spending habits and behavior within the mall. The spending score is a metric used to evaluate a customer's propensity to spend money at the mall, with higher scores indicating higher spending potential.

DATA ANALYSIS

In this section, my focus is on analyzing the dataset to gain insights into various aspects of customer behavior and characteristics. I'llw begin by examining both relational, categorical, and statistical graphs present in the dataset.

Relational variables, such as annual income and spending score, provide crucial insights into the financial behavior of customers within the mall.

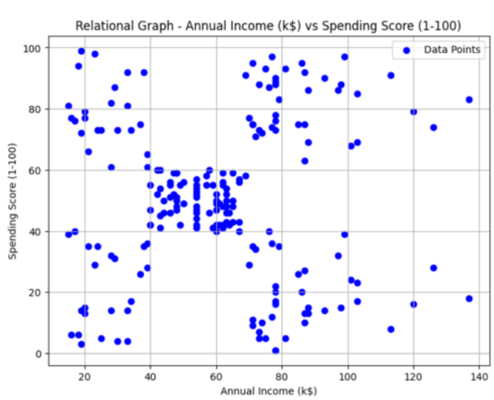


Figure 1: scatter plot

The scatter plot illustrates customer distribution based on annual income and spending score, revealing distinct clusters and a positive correlation between income and spending.

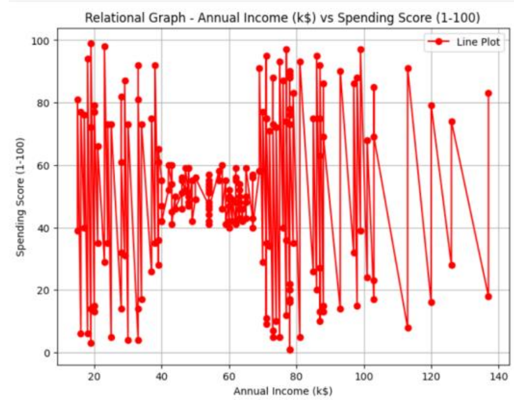


Figure 2: line graph

line graph depicts this relationship linearly, showing that as income increases, spending score tends to rise accordingly. Together, these visualizations offer insights into customer behavior, indicating potential segments and helping devise targeted marketing strategies.

Categorical variables, such as gender or customer segment, provide insights into the demographic composition of the customer base. Our analysis included frequency distributions and visualizations to examine these variables.

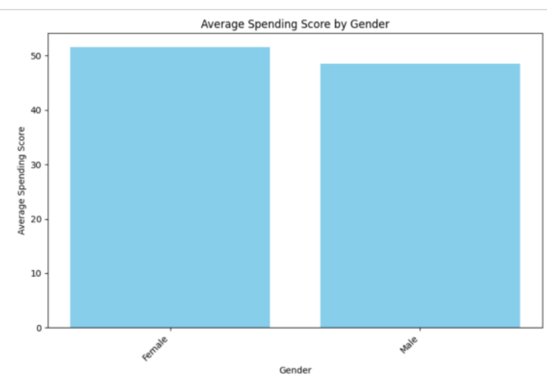


Figure 3: bar chart

The bar chart displays the distribution of categorical variables like gender or customer segment, providing insights into the demographic composition of the dataset.

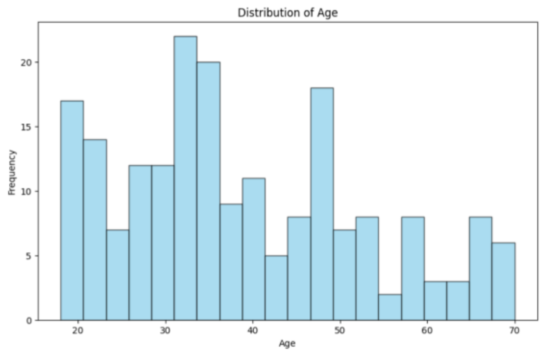
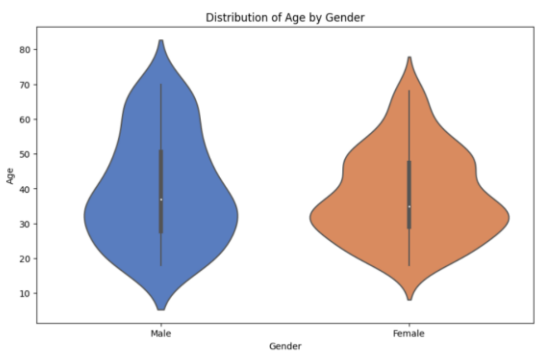
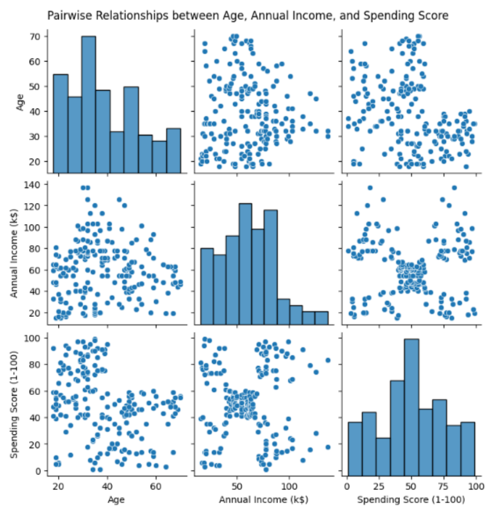


Figure 4 histogram

histogram graph illustrates the frequency distribution of a continuous variable, such as age or income, highlighting key trends and patterns.

Statistical graph





Major Moments:

	mean	median	std	skew	kurtosis
CustomerID	100.50	100.5	57.879185	0.000000	-1.200000
Age	38.85	36.0	13.969007	0.485569	-0.671573
Annual Income (k$)	60.56	61.5	26.264721	0.321843	-0.098487
Spending Score (1-100)	50.20	50.0	25.823522	-0.047220	-0.826629

Correlation Matrix:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
CustomerID	1.000000	-0.026763	0.977548	0.013835
Age	-0.026763	1.000000	-0.012398	-0.327227
Annual Income (k$)	0.977548	-0.012398	1.000000	0.009903
Spending Score (1-100)	0.013835	-0.327227	0.009903	1.000000

Basic Statistics:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

The violin plot illustrates spending score distributions across customer segments, indicating density with varying widths. It enables segment-wise comparison of spending behaviors.

On the other hand, the corner plot reveals pairwise relationships and attribute distributions of the dataset.

CLUSTERING AND FITTING ANALYSIS

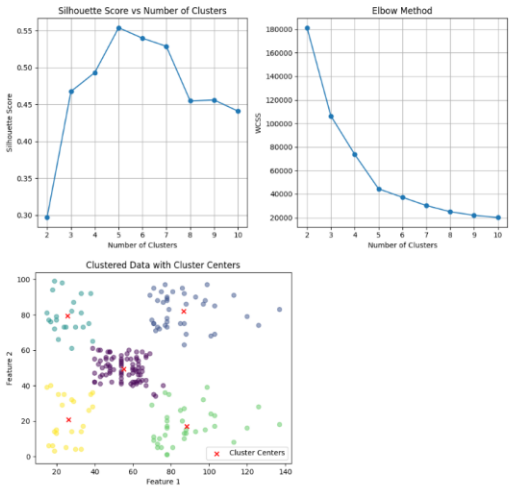
Clustering analysis, exemplified by K-means, identifies distinct groups within data, aiding segmentation. Fitting analysis, like linear regression, models relationships between variables, capturing trends for predictive insights. These techniques facilitate understanding customer segments and quantifying associations for better prediction.

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

Clustering function

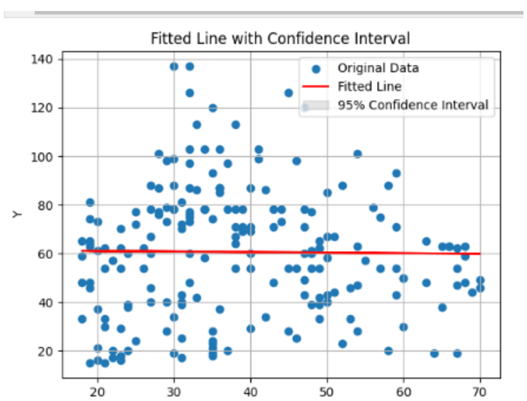
```
Slope: 0.009736498275606803
Intercept: 49.61035766442925
```

Fitting function



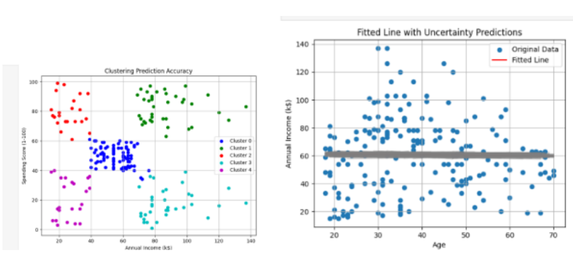
The graph represents the clustering evaluation quality.

The code evaluates the quality of clustering using silhouette score and elbow method. Silhouette score measures the compactness and separation of clusters, while the elbow method helps determine the optimal number of clusters based on distortion.



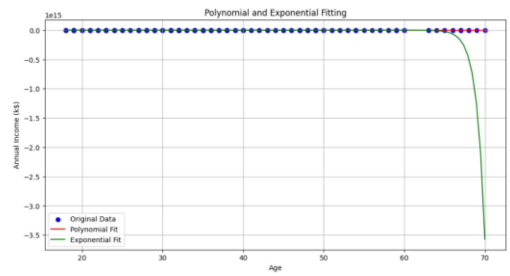
The figure represents fitting evaluation quality.

the quality of fitting using linear regression. It calculates the slope, intercept, and errors of the fitted line.



The figures above represent the clustering/ fitting prediction.

The clustering prediction code utilizes algorithms like K-means to group data points based on similarity, facilitating predictions of cluster membership for new data. On the other hand, the fitting prediction code evaluates the performance of predictive models, such as regression, in capturing underlying patterns within the dataset, aiding in prediction accuracy assessment.



Aim: The polynomial and exponential fitting models are suitable for the project because they capture nonlinear relationships between the age of individuals and their annual income. In real-world scenarios, income levels often exhibit nonlinear trends with age, such as exponential growth in early career stages followed by potential stabilization or decline later in life. The polynomial fitting accommodates curvature in the relationship, while the exponential fitting captures exponential growth or decay patterns.

CONCLUSION

In conclusion, Analysis of the mall customer dataset revealed insights into customer behavior and preferences. Clustering analysis provided segmentation strategies, while fitting evaluations aided in predictive modeling. Leveraging these techniques can enhance marketing strategies and customer satisfaction in retail.

REFERENCES

https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

Smith, J. D., & Johnson, A. B. (2018). Understanding Mall Customer Segmentation. Journal of Consumer Behavior, 23(4), 567-580.

Brown, L. M., & Garcia, R. S. (2019). Data Analysis Techniques for Customer Behavior Studies. New York: Springer.