================= RESTART: C:/Phoebe/Phoebe Muthoni ass 2.py ================

CustomerID Gender 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

ii) Number of records : 200

iii) Features in the dataset:

Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',

'Spending Score (1-100)'],

dtype='object')

iv) The dataset does not contain missing values.

vii) Statistics by Gender:

------------------------------------

Mean Annual Income by Gender:

Gender

Female 59.250000

Male 62.227273

Name: Annual Income (k$), dtype: float64

* The average annual income for female customers is approximately $59,250, while for male customers, it's around $62,227. This suggests that, on average, male customers tend to have slightly higher annual incomes compared to female customers.

Standard Deviation of Annual Income by Gender:

Gender

Female 26.011952

Male 26.638373

Name: Annual Income (k$), dtype: float64

* The average annual income for female customers is approximately $59,250, while for male customers, it's around $62,227. This suggests that, on average, male customers tend to have slightly higher annual incomes compared to female customers.

Mean Age by Gender:

Gender

Female 38.098214

Male 39.806818

Name: Age, dtype: float64

* The average age for female customers is approximately 38.10 years, while for male customers, it's around 39.81 years. This indicates that, on average, male customers tend to be slightly older than female customers.

Standard Deviation of Age by Gender:

Gender

Female 12.644095

Male 15.514812

Name: Age, dtype: float64

* The standard deviation of age for female customers is approximately 12.64 years, and for male customers, it's around 15.51 years. This suggests that there is slightly more variability in the ages of male customers compared to female customers.

Mean Spending Score by Gender:

Gender

Female 51.526786

Male 48.511364

Name: Spending Score (1-100), dtype: float64

* The average spending score (on a scale of 1 to 100) for female customers is approximately 51.53, while for male customers, it's around 48.51. This indicates that, on average, female customers tend to have slightly higher spending scores compared to male customers.

Standard Deviation of Spending Score by Gender:

Gender

Female 24.11495

Male 27.89677

Name: Spending Score (1-100), dtype: float64

* The standard deviation of spending score for female customers is approximately 24.11, and for male customers, it's around 27.90. This suggests that there is slightly more variability in spending scores among male customers compared to female customers.



viii) Count of customers in each cluster:

Cluster

1 73

0 56

2 47

3 24

Name: count, dtype: int64

Cluster centers:

[[ 55.66071429 51.01785714]

[ 31.95890411 72.95890411]

[ 29.44680851 29.21276596]

[ 39. 106.5 ]]

The elbow plot visually demonstrates the process of determining the best value of k. From the plot, I can observe the point where the inertia starts to decrease at a slower rate. This point corresponds to the optimal value of k, which indicates the number of clusters that best represent the underlying structure of the data. I observed a distinct kink or elbow at k=2. This indicates that adding additional clusters beyond this point does not lead to significant improvement in clustering performance. Therefore, I concluded that k=2 is a suitable choice for the number of clusters to effectively capture the underlying structure of the data.

