UNIVERSITI KUALA LUMPUR

BKB 40603: ARTIFICIAL INTELLIGENCE

K-MEANS CLUSTERING: MALL CUSTOMERS CLUSTERING ANALYSIS

NAZHAN HAKEEM BIN AHMAD ZAKI (51218118204)

MOHAMAD DANIAL HEZRI BIN KHAIRI (51218118193)

MUHAMMAD HANIF BIN SALLEHUDIN (51218118208)

MUHAMAD HUSAINI BIN YUSRI (51218118195)

WHAT IS CLUSTERING?

- Classify each data point into a specific group.
- Data points in the same group should have comparable qualities and features, whereas data points other groups should have distinct properties and features.
- Unsupervised learning is a widely utilized statistical data analysis approach in a variety of fields.

TYPES OF CLUSTERING

1. Hard Clustering
    - Each data point is either totally or partially associated with a Cluster.
    - This means that every data point will belong to one and only one cluster at a time.
2. Soft Clustering
    - Each data point can belong to Multiple Clusters at the same time.
    - This Means data points can be associated to Multiple Clusters at the same time.

CENTROID BASED CLUSTERING

- These models are based on the idea that similarity is determined by a data point's closeness to the cluster's centroid.
- The Points will be divided into different Clusters based on the Closeness to a Centroid.
- The K Means Algorithm is one of the most well-known algorithms in this category.

POPULAR APPLICATIONS OF CLUSTERING

IN THE FIELD OF SALES AND MARKETING

- Optimization of Marketing Ad Campaigns for Best Returns.
- Tracking Down Target Customers for better Revenue Generation.
- Grouping Customer with Similar Characteristics to Increase Sales.

PROBLEM STATEMENT

Imagine you own a mall and want to learn more about the consumers, such as who can easily converge [Target Customers], so that the marketing staff may design their approach properly to learn about customer segmentation and market basket analysis.

OBJECTIVES

01 To elaborate in details regarding K-means clustering with practical example.

02 To classify the data of mall customers into specific groups and assign label to each group.

03 To implement K-means clustering method using JUPYTER Notebook application.

SOFTWARE USED - JUPYTER NOTEBOOK

- Jupyter supports over 40 programming languages, including Python, R, Julia, and Scala.
- Notebooks can be shared with others using email, Dropbox, GitHub and the Jupyter Notebook Viewer.
- Data Visualisation: As a component, the shared notebook Jupyter supports visualisations and includes rendering some of the data sets.
- Your code can produce rich, interactive output: HTML, images, videos, LaTeX, and custom MIME types.
- Leverage big data tools, such as Apache Spark, from Python, R and Scala.

PROGRAMMING LANGUAGE USED- PYTHON

Library used

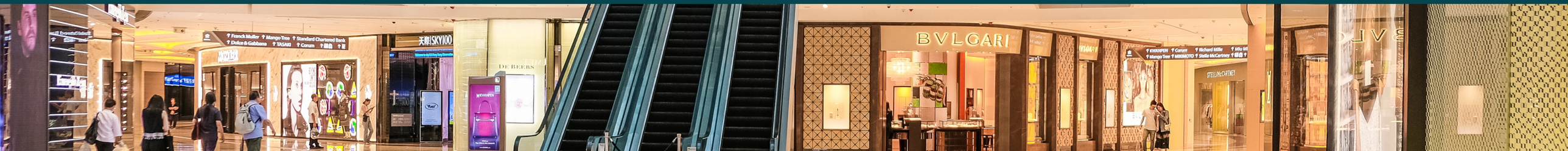
Numpy (basic mathematic operations.)

Pandas (dataframe manipulations)

Matplotlib.pyplot (data visualizations)

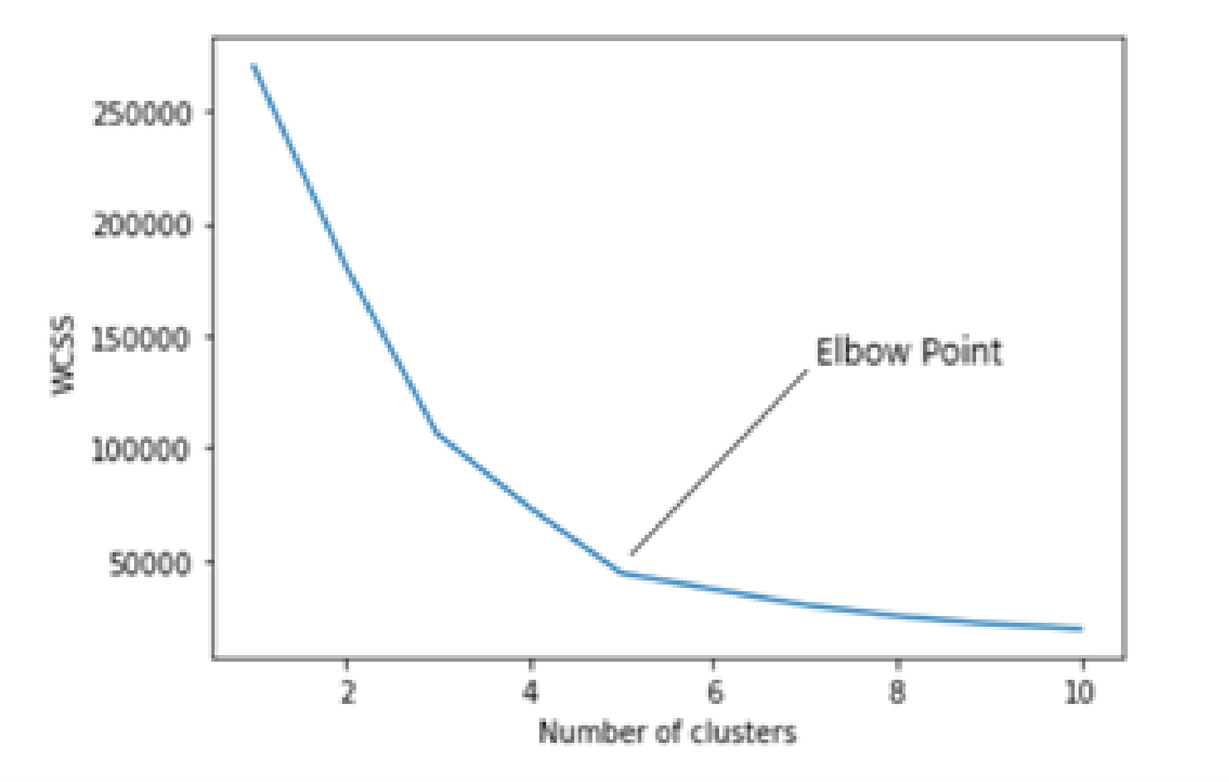
Seaborn (data visualizations)

Dabl. (data analysis)

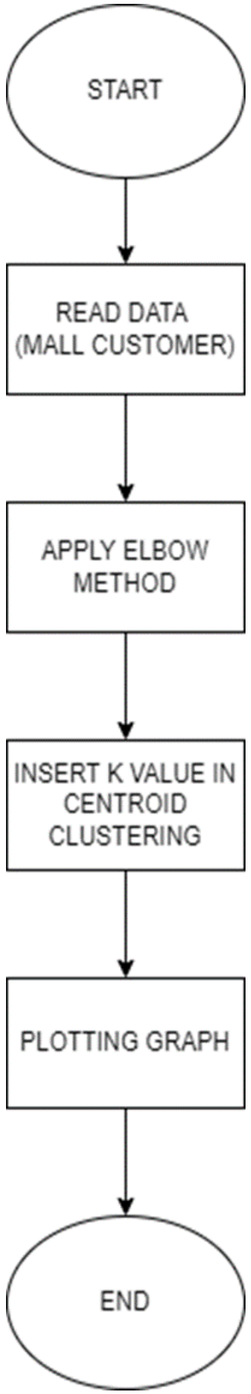


ELBOW METHOD

1. To find and determine the best number of clusters by fitting the model with a range of values for K
2. For each value of K, it will be computing WCSS (Within-Cluster Sum of Square)
3. WCSS value is greatest when K = 1



FLOWCHART


START
READ DATA (MALL CUSTOMER)
APPLY ELBOW METHOD
INSERT K VALUE IN CENTROID CLUSTERING
PLOTTING GRAPH
END

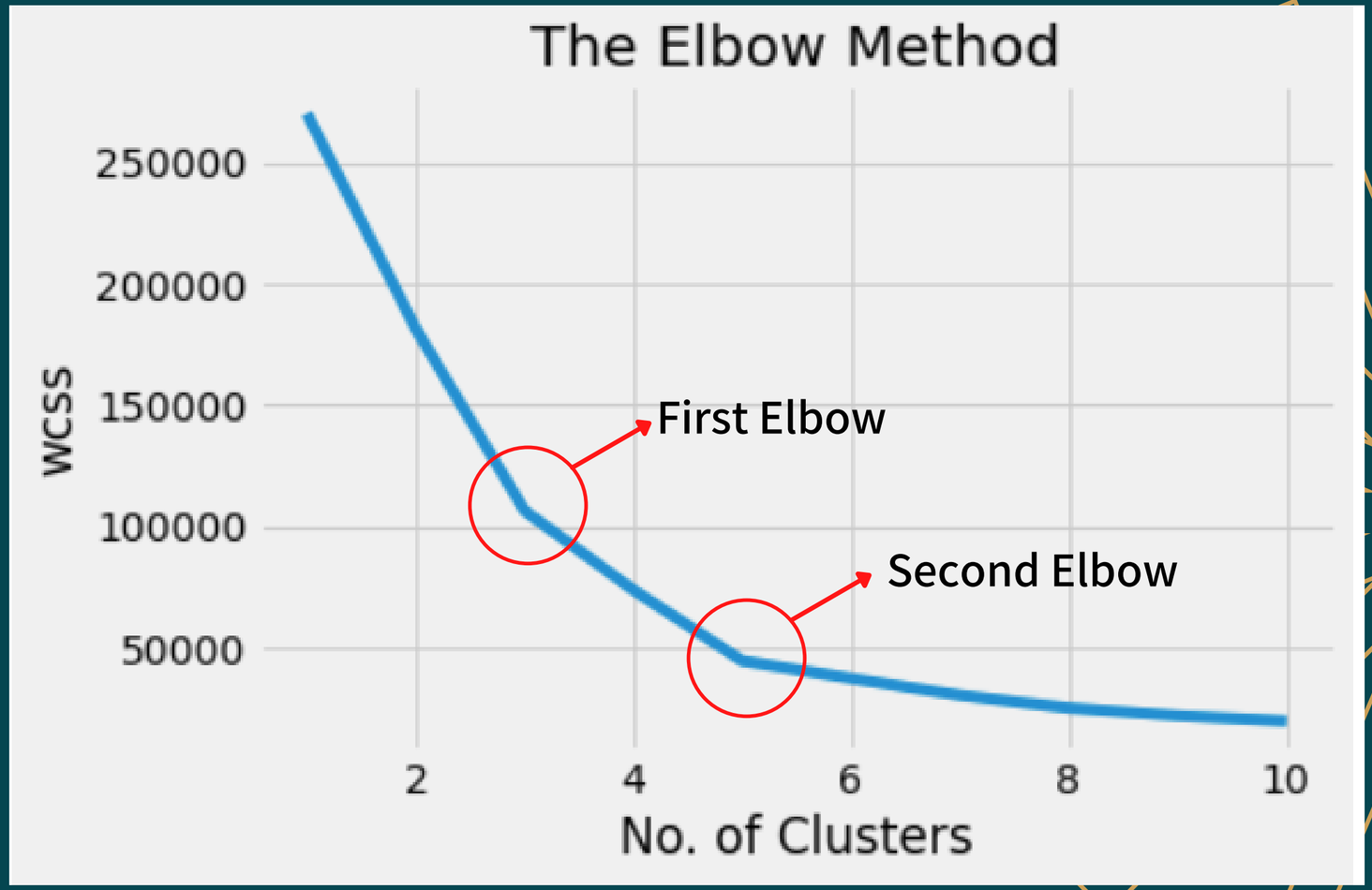
RESULTS

Descriptive 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

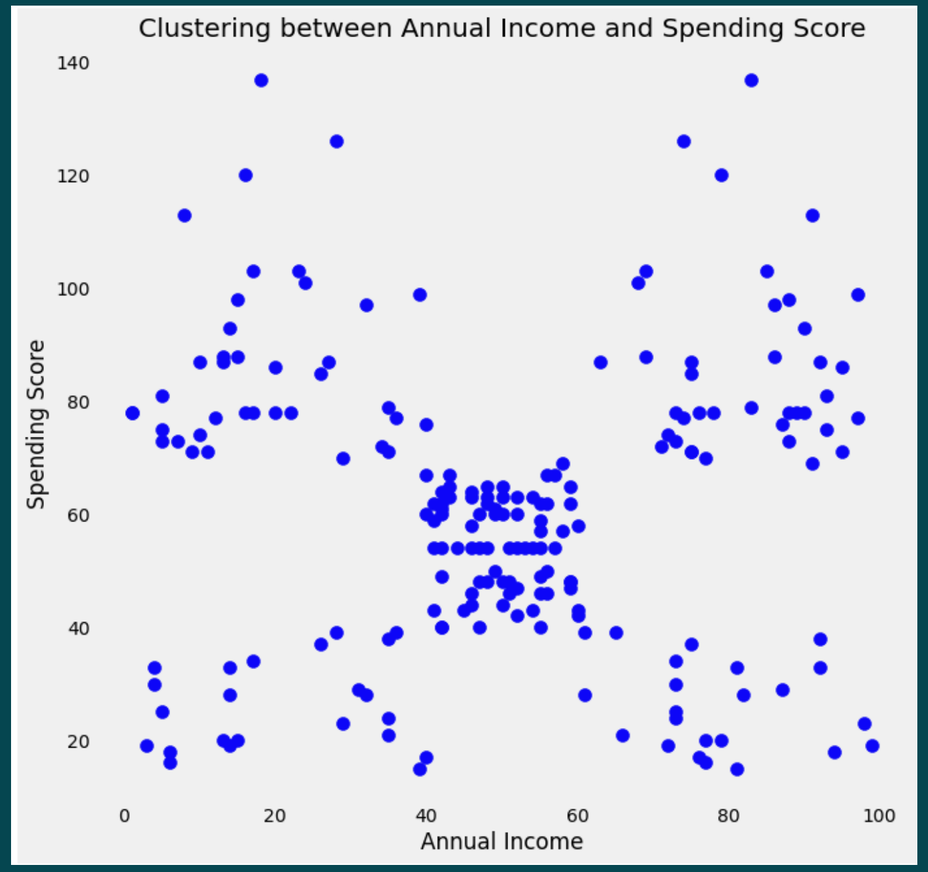
ELBOW METHOD

To determine the optimal number of clustes

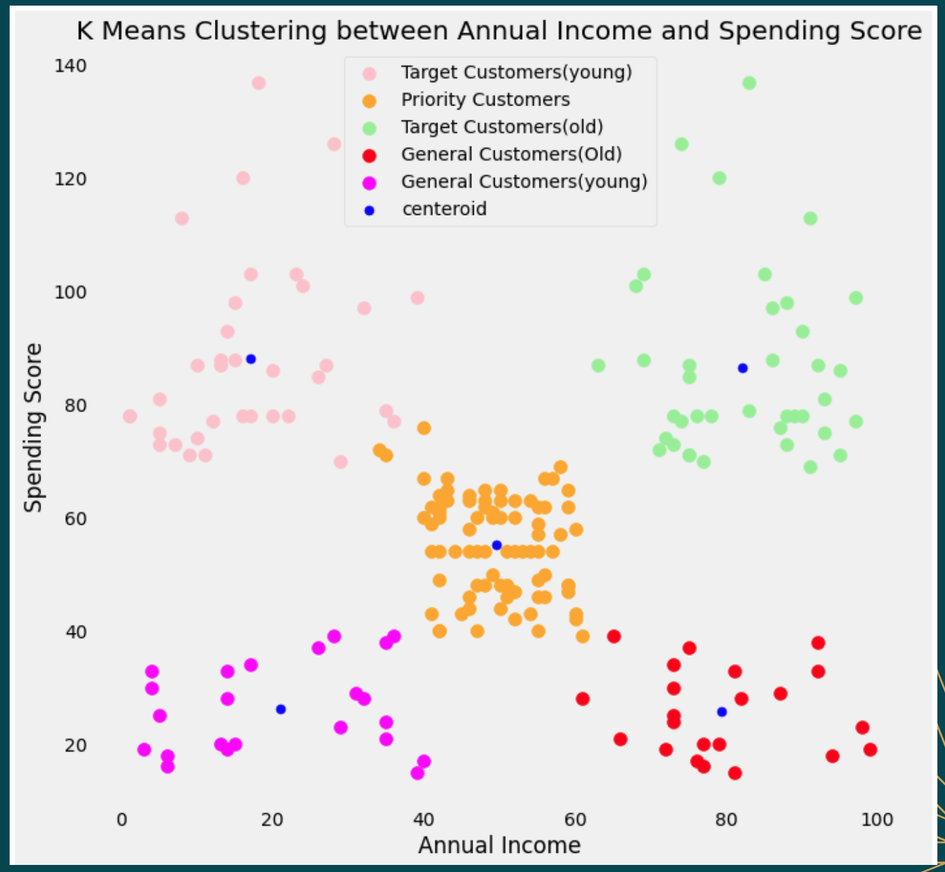


K-MEANS CLUSTERING

Before apply KMC



After apply KMC



SILHOUETTE SCORE

From coding:

```
from sklearn.metrics import silhouette_score

#calculate the score
score=silhouette_score(x, km.labels_, metric='euclidean')

#print the score
print('Silhouette Score: %.3f' % score)
```

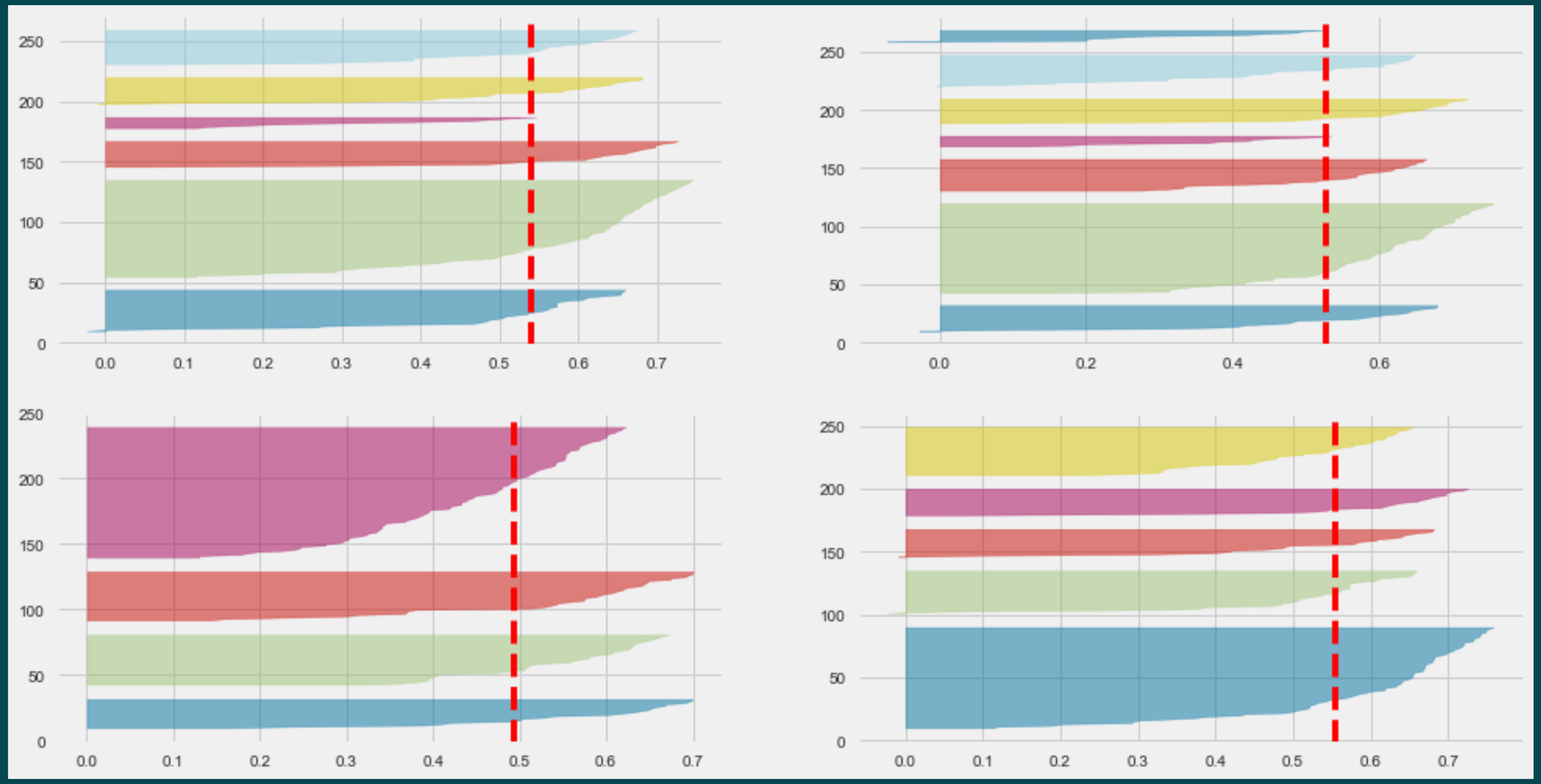
```
Silhouette Score: 0.554
```

- Value +1 indicates the sample is very far away from neighbouring cluster and very close to its cluster.
- Value 0 indicates the sample is at the boundary of the distance between two clusters.
- Value -1 indicates the sample is closer to the neighbouring cluster than its cluster.

COMPARISON OF SILHOUETTE SCORE

K value=6, SS=0.538

K value=7, SS=0.526



K value=4, SS=0.493

K value=5, SS=0.554

CONCLUSION

The mall customers were able to be classify and grouped into five different categories based on their annual income and spending score using K-means clustering. This method helps mall owners to properly identify their consumers and accordingly strategize the marketing techniques used to suit their target customers.