CODE CLAUSE

Chris Aryan

Data Science Intern

TASK

AIM :

To apply K-Means clustering to segment customers based on their purchase behavior.

DESCRIPTION :

Using a customer purchase dataset to identify distinct segments using the K-Means clustering algorithm.

TASK

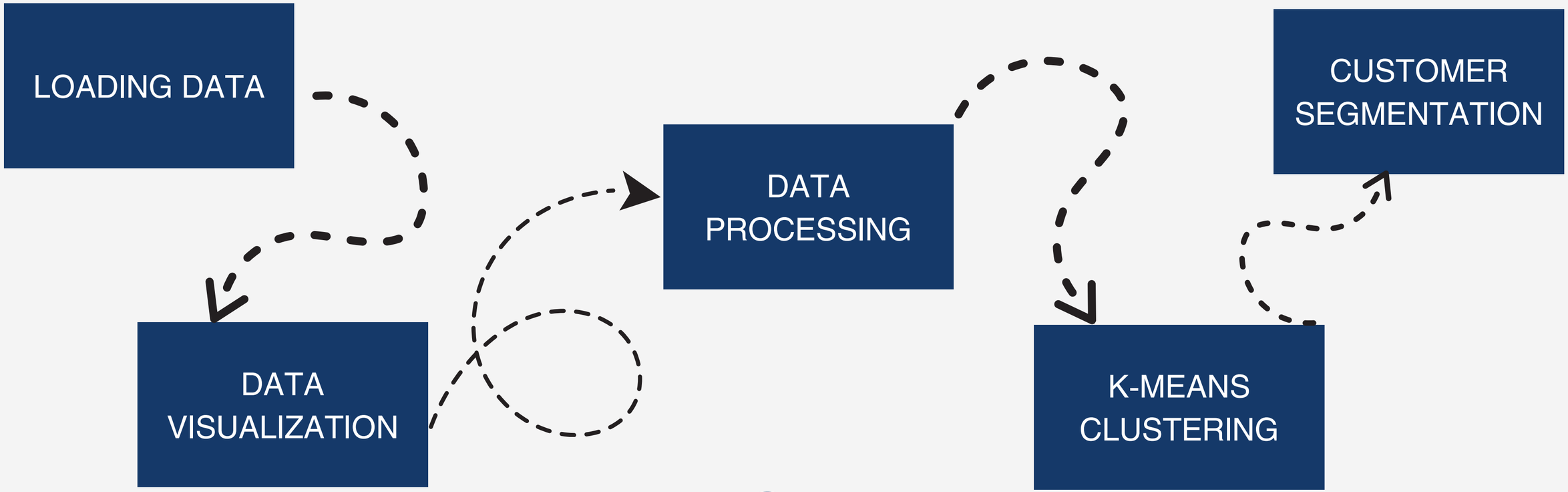
TECHNOLOGIES USED :

Python, Pandas, Numpy, Matplotlib, Seaborn, Sklearn.cluster

DATASET :

The dataset is taken from Kaggle. Please find the link below.
DATASET

FLOWCHART

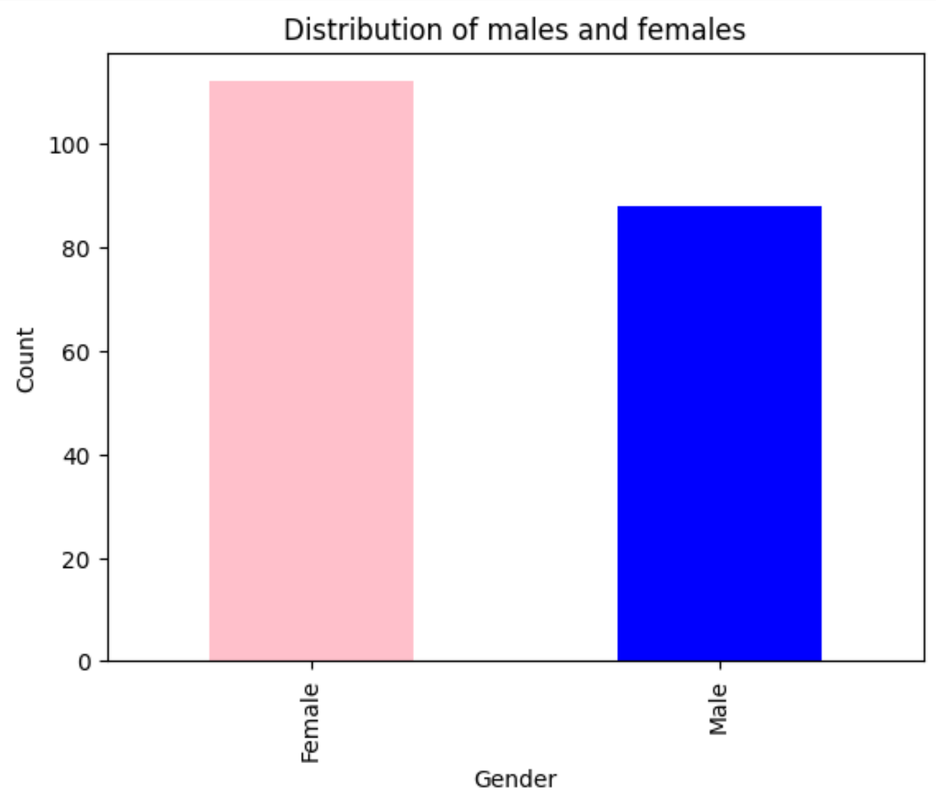

LOADING DATA
DATA VISUALIZATION
DATA PROCESSING
K-MEANS CLUSTERING
CUSTOMER SEGMENTATION

DATA VISUALIZATION

	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

- We observe that our dataset has five attributes.

DATA VISUALIZATION



- We can see that there are significant numbers of male and female customers visiting the mall.
- So, we can apply clustering for males and females separately and based on their annual income we can perform customer segmentation for both the genders.

DATA PROCESSING

```
null_count= df.isnull().sum()
```

```
CustomerID                0
Gender                    0
Age                       0
AnnualIncome(k$)          0
SpendingScore(1-100)      0
dtype: int64
```

- We can see that there are no null values in any columns.
- We are going to perform segmentation of both the genders based on their annual income and spending score.

DATA PROCESSING

```
males_data = df[df['Gender'] == 'Male'][['AnnualIncome(k$)', 'SpendingScore(1-100)']]
```

```
males_data
```

	AnnualIncome(k$)	SpendingScore(1-100)
0	15	39
1	15	81
8	19	3
10	19	14
14	20	13
...	...	...
187	101	68
192	113	8
197	126	74
198	137	18
199	137	83

```
females_data = df[df['Gender'] == 'Female'][['AnnualIncome(k$)', 'SpendingScore(1-100)']]
```

```
females_data
```

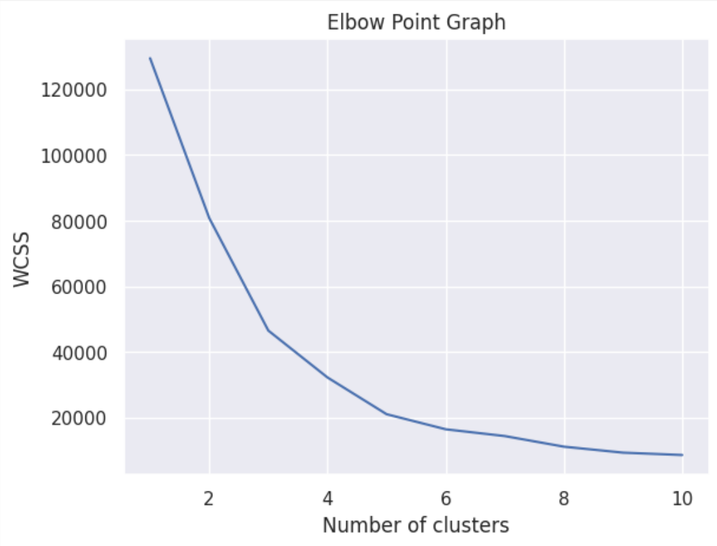
	AnnualIncome(k$)	SpendingScore(1-100)
2	16	6
3	16	77
4	17	40
5	17	76
6	18	6
...	...	...
191	103	69
193	113	91
194	120	16
195	120	79
196	126	28

112 rows × 2 columns

- Extracting males and females column from our dataset with their respective annual income and spending score.

K-MEANS CLUSTERING

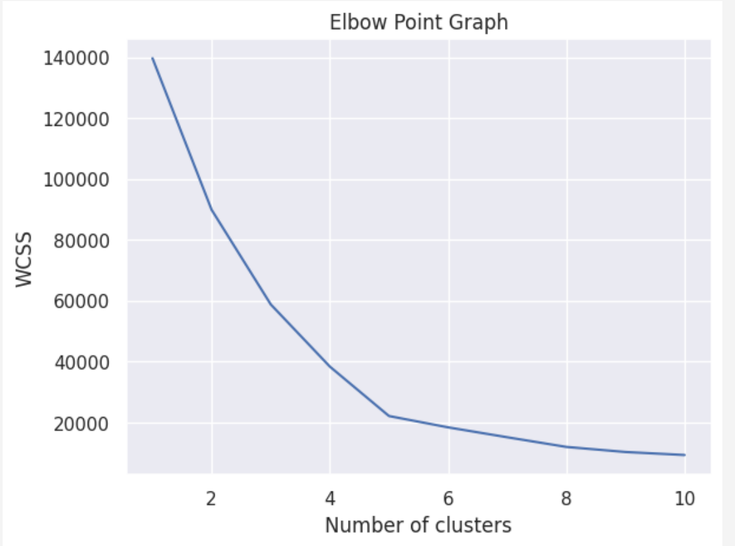
- **Now, applying elbow method to find the optimal number of clusters for both the genders.**



FOR MALES

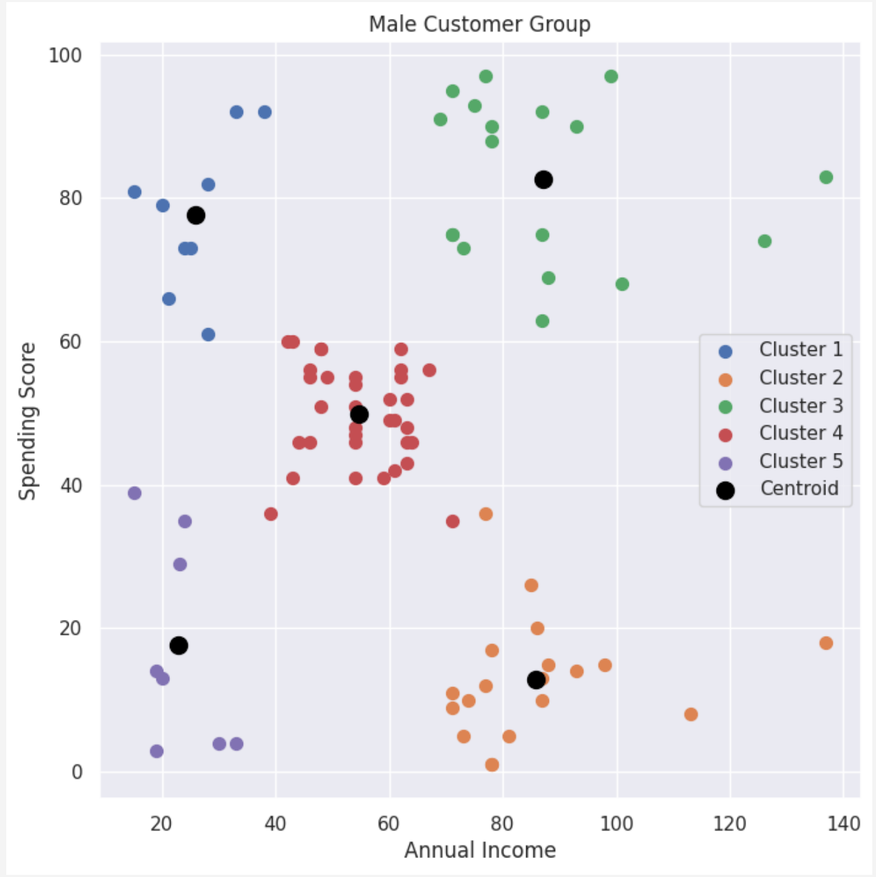
- We observe that for both males and females, the elbow graph shows a steep fall when the number of clusters equals 5.

- Therefore, assuming the optimal number of clusters to be 5 for both the genders.



FOR FEMALES

Customer Segmentation

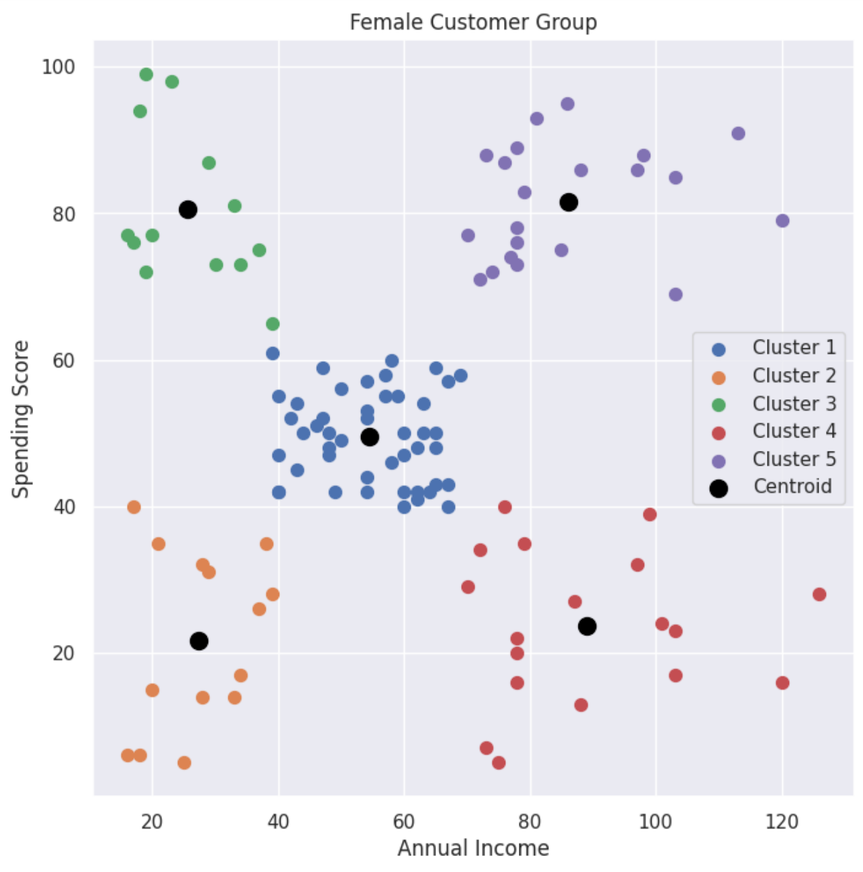


- The graph shows the male customers segmented in clusters in accordance with their spending score and annual income.

OBSERVATIONS:

- The number of customers with low annual income and low spending score is less, which makes sense.
- There are significant number of customers who have high annual income but they have low spending score. To such customers, price is not an issue, so the mall needs to update their collection.

Customer Segmentation



- The graph shows the female customers segmented in clusters in accordance with their spending score and annual income.

OBSERVATIONS:

- Most of the female customers are having mid annual income and mid spending score. The females in this cluster can be given special offers to increase their spending score.
- We also see that the number of females with high annual income and high spending score are comparatively more. This shows that the mall offers better quality at good rates to the females as compared to men.

THANKYOU

Chris Aryan

Data Science Intern