SEGMENTATION AND CLUSTERING

THE PROBLEM

We have a dataset that contains information about customers.
We want to create clusters to categorize these customers and make assumptions about these clusters.

DATASET ATTRIBUTES

Customer ID: id of the customer
Gender: gender of the customer
Age: age of the customer
Annual Income: annual income of the customer
Spending Score: spending score of the customer

DATA EXPLORATION AND PREPROCESSING

An example of a row in the dataset

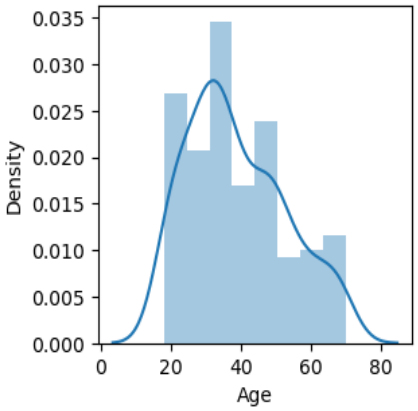
	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

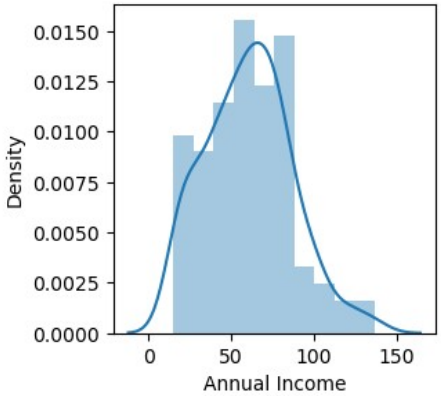
Some information about the values of the dataset

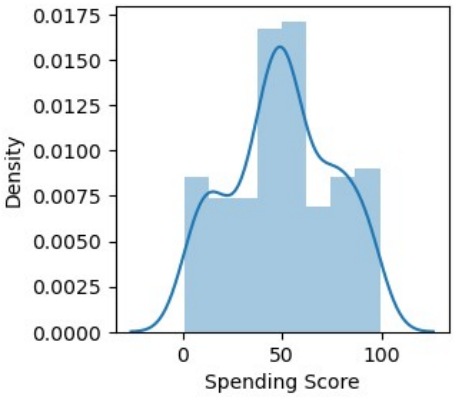
	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000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

We will remove the column Customer ID as it does not provide us with any useful information .

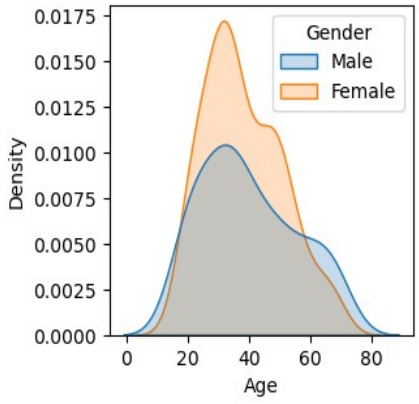
Plots of Age, Annual Income and Spending Score

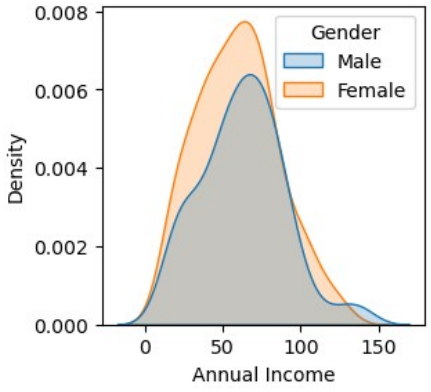


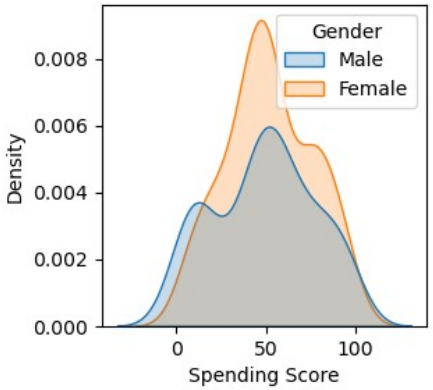




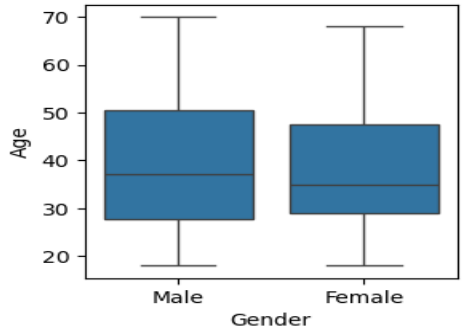
Plots of Males vs Females for Age, Annual Income and Spending Score

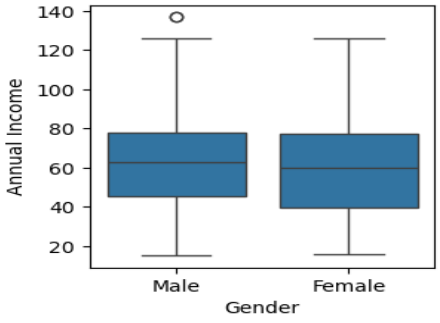


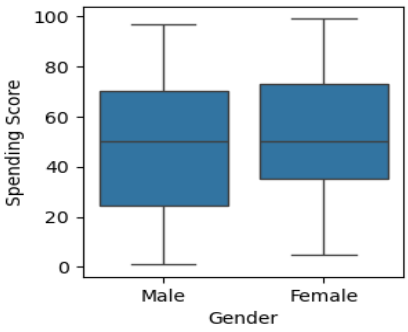




Box plot of Males vs Females for Age, Annual Income and Spending Score



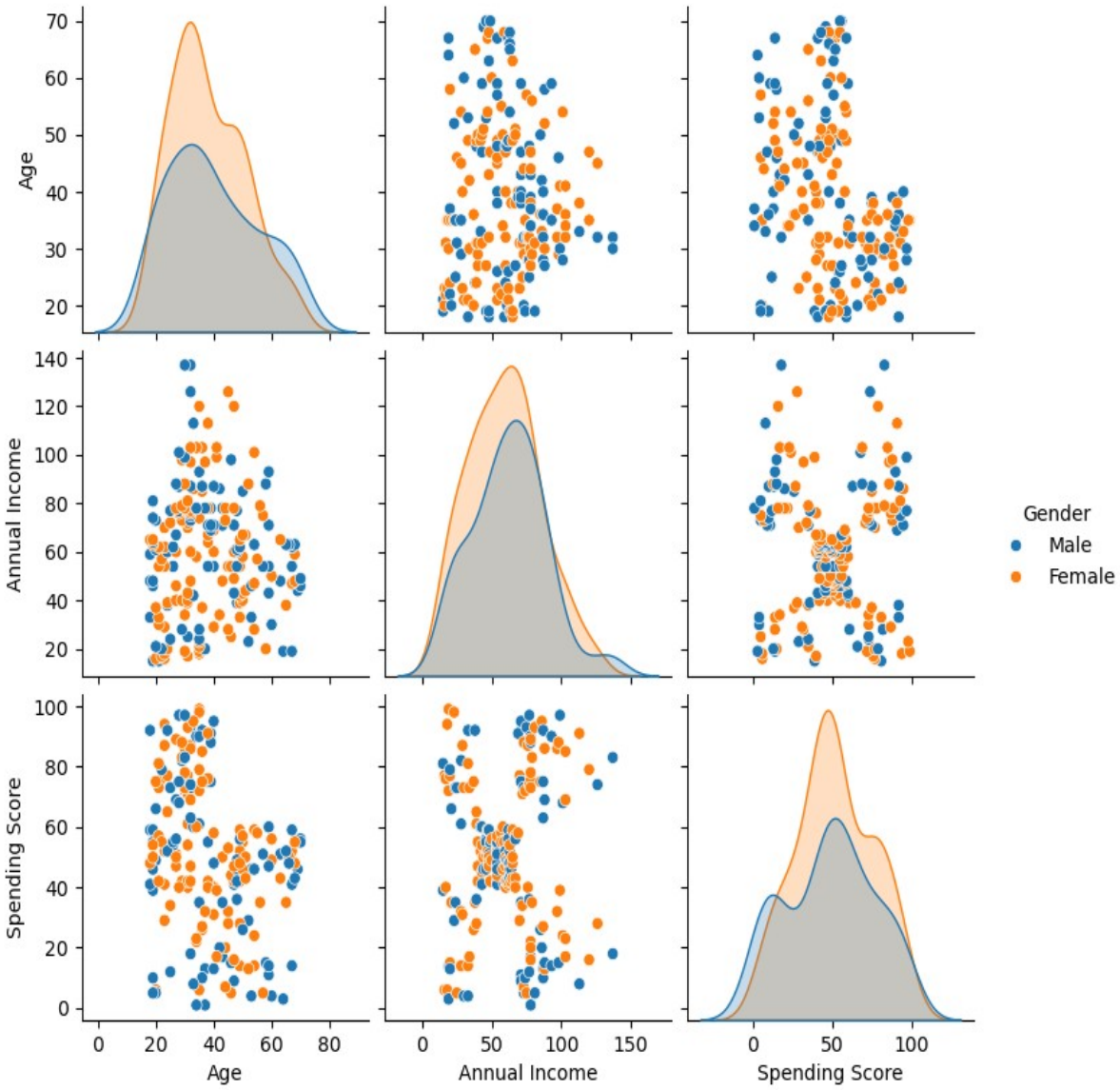




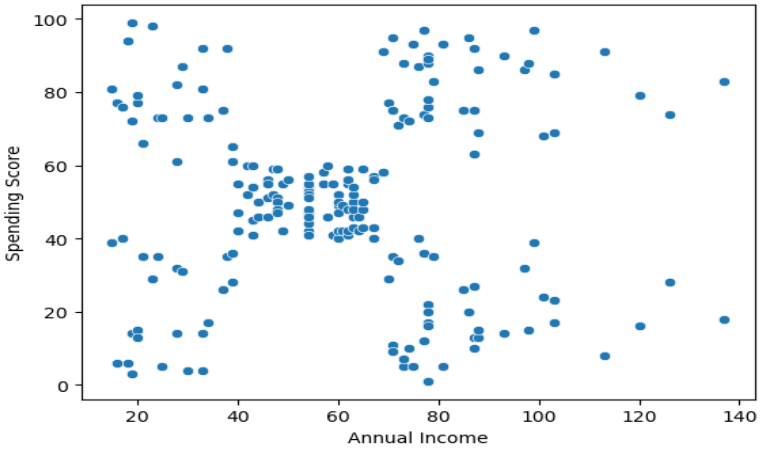
Percentage of Males vs Females

```
Gender
Female    0.56
Male      0.44
Name: proportion, dtype: float64
```

Some plots about Male vs Female



We will use the combination Annual Income/Spending Score
Scatter plot of Annual Income/Spending Score

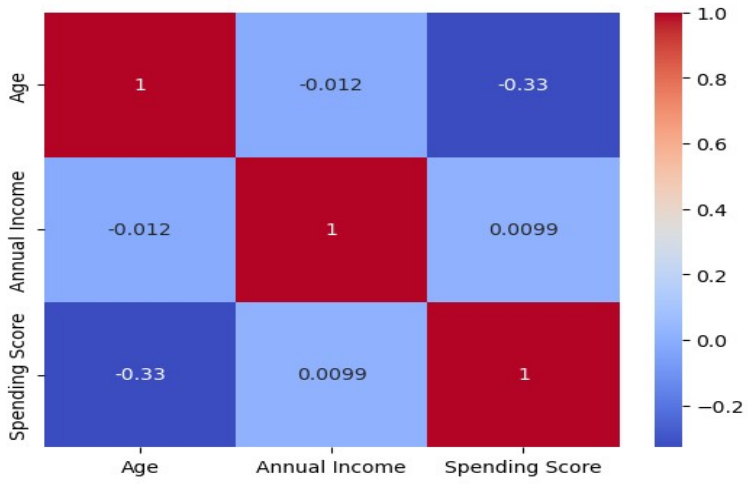


We observe that we may need 5 clusters.

We group the data by Gender and find the mean Annual Income/Spending Score

Gender	Age	Annual Income	Spending Score
Female	38.098214	59.250000	51.526786
Male	39.806818	62.227273	48.511364

We find if our attributes have any correlation between them .



They do not have any correlation

CLUSTERING

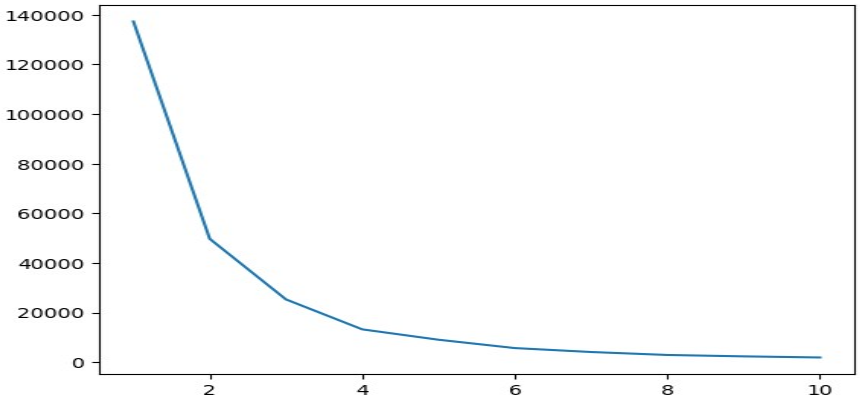
We will use the K means algorithm

UNIVARIATE CLUSTERING

Create clusters based on Annual Income

First lets find the inertia scores and plot them, inertia scores show how well the dataset was clustered by K means.

```
[137277.28,
 49761.737012987,
 25341.285871863223,
 13278.112713472487,
 9100.830157289776,
 5742.224880382775,
 4156.433857808857,
 2964.763636363636,
 2404.4269230769232,
 1955.2266067266069]
```



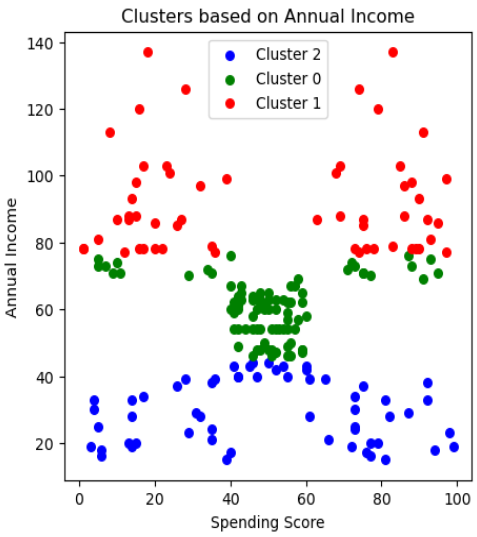
We observe that the ‘elbow’ in the graph starts at Number of Cluster = 3. So 3 clusters may be the most optimal.
Lets see how many values its Cluster has

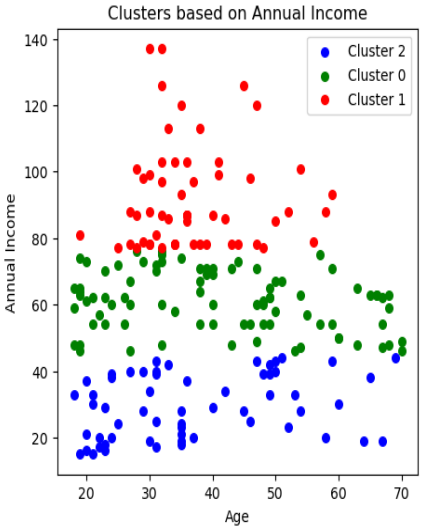
```
Income Cluster
0    86
2    58
1    56
Name: count, dtype: int64
```

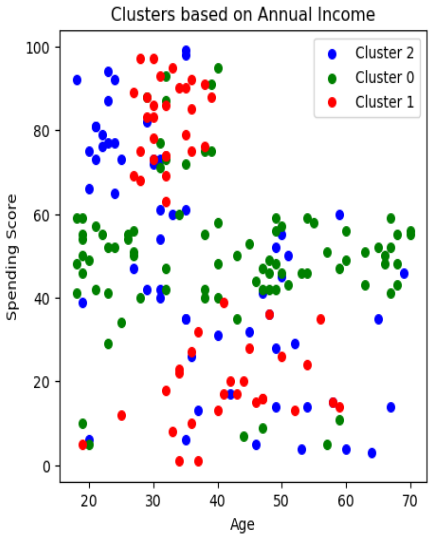
Lets group the data by the Annual Income Clusters and see the mean Age, Annual Income, Spending Score

Income Cluster	Age	Annual Income	Spending Score
0	41.279070	60.906977	50.337209
1	36.910714	92.142857	50.517857
2	37.120690	29.551724	49.689655

Lets plot the Clusters



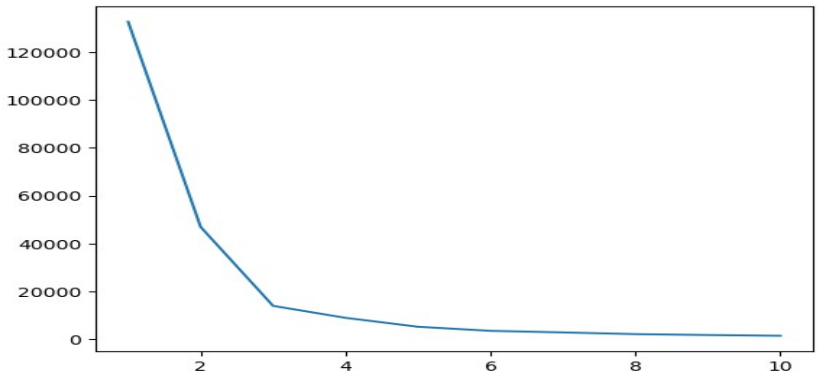




Create clusters based on Spending Score

First lets find the inertia scores and plot them, inertia scores show how well the dataset was clustered by K means.

```
[132704.0,
 46936.47916666667,
 13901.380274071398,
 8899.422693960478,
 5161.663758460416,
 3466.772391393773,
 2802.5106747412224,
 2085.4320661230618,
 1713.6779338334081,
 1397.7471675544596]
```



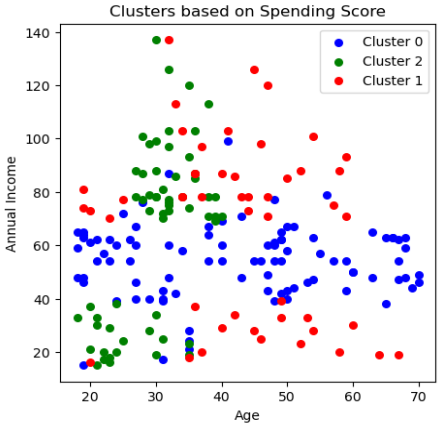
We observe that the‘elbow’ in the graph starts at Number of Cluster = 3. So 3 clusters may be the most optimal.
Lets see how many values its Cluster has

```
Spending Cluster
0    93
1    58
2    49
Name: count, dtype: int64
```

Lets group the data by the Spending Score Clusters and see the mean Age, Annual Income, Spending Score

Spending Cluster	Age	Annual Income	Spending Score
0	42.247312	54.215054	48.709677
1	42.877551	67.000000	15.306122
2	30.000000	65.293103	82.068966

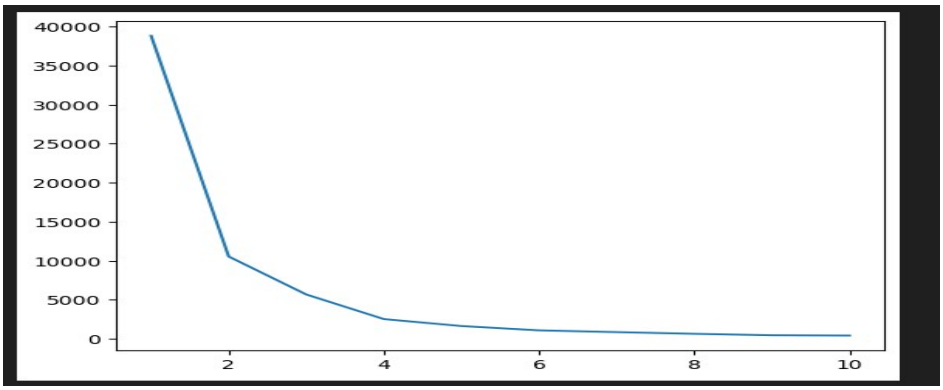
Lets plot the Clusters



Create clusters based on Age

First lets find the inertia scores and plot them, inertia scores show how well the dataset was clustered by K means.

```
[38831.50000000001,
 10549.679966044141,
 5685.520687645688,
 2534.1262305206155,
 1643.843142405674,
 1091.5176193608406,
 866.1143966547193,
 649.2761580526541,
 463.5767325684055,
 424.844529105768]
```



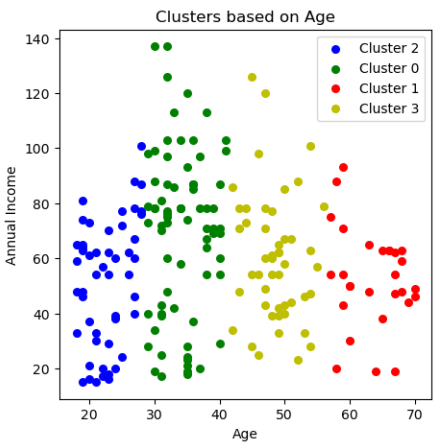
We observe that the‘elbow’ in the graph starts at Number of Cluster = 4. So 4 clusters may be the most optimal.
Lets see how many values its Cluster has

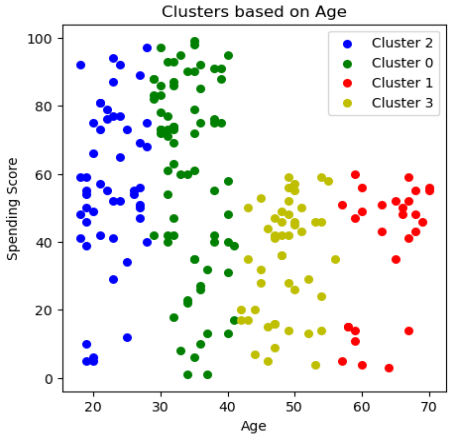
```
Age Cluster
0    74
2    50
3    48
1    28
Name: count, dtype: int64
```

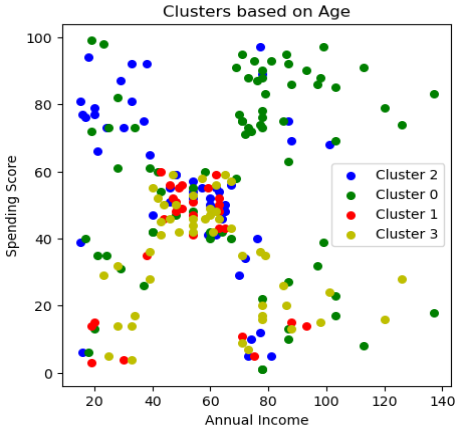
Lets group the data by the Spending Score Clusters and see the mean Age, Annual Income, Spending Score

Age Cluster	Age	Annual Income	Spending Score
0	34.229730	69.878378	60.148649
1	63.535714	52.785714	38.500000
2	22.600000	51.440000	56.780000
3	48.500000	60.229167	34.833333

Lets plot the Clusters





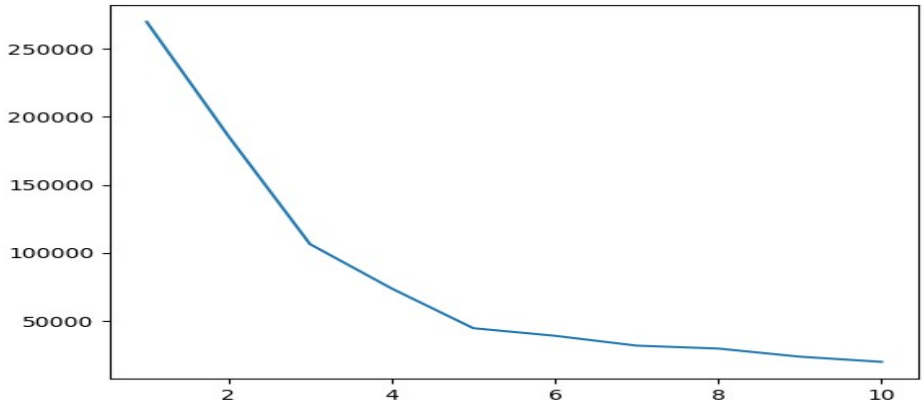


BIVARIATE CLUSTERING

Create clusters based on Annual Income + Spending Score

First lets find the inertia scores and plot them, inertia scores show how well the dataset was clustered by K means.

```
[269981.27999999997,
185917.14253928518,
106348.37306211119,
73679.78903948834,
44448.4554479337,
38858.95997514391,
31605.86838023088,
29459.12901314177,
23466.20585925702,
19657.78360870396]
```



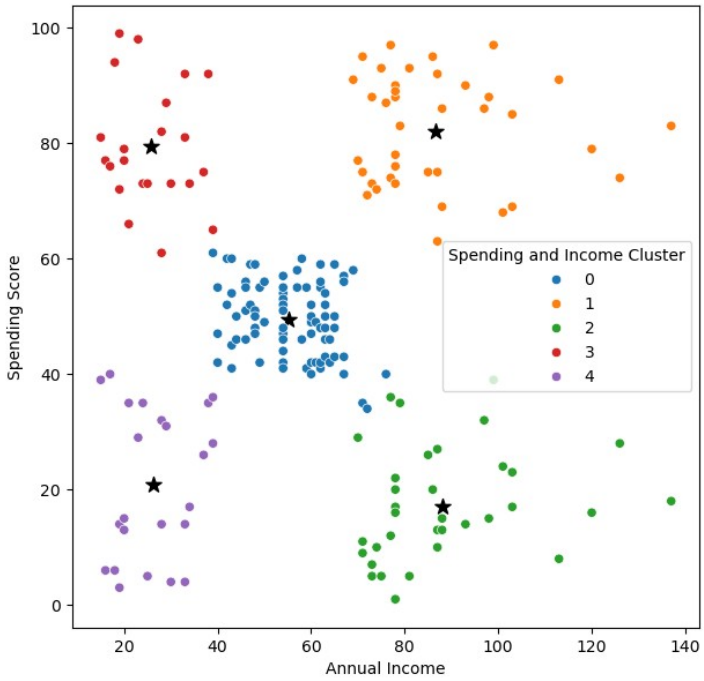
We observe that the‘elbow’ in the graph starts at Number of Cluster = 5. So 5 clusters may be the most optimal.
Lets see how many values its Cluster has

```
Spending and Income Cluster
3    81
0    39
2    35
1    23
4    22
Name: count, dtype: int64
```

Lets group the data by the Spending Score Clusters and see the mean Age, Annual Income, Spending Score

Spending and Income Cluster	Age	Annual Income	Spending Score
0	32.692308	86.538462	82.128205
1	45.217391	26.304348	20.913043
2	41.114286	88.200000	17.114286
3	42.716049	55.296296	49.518519
4	25.272727	25.727273	79.363636

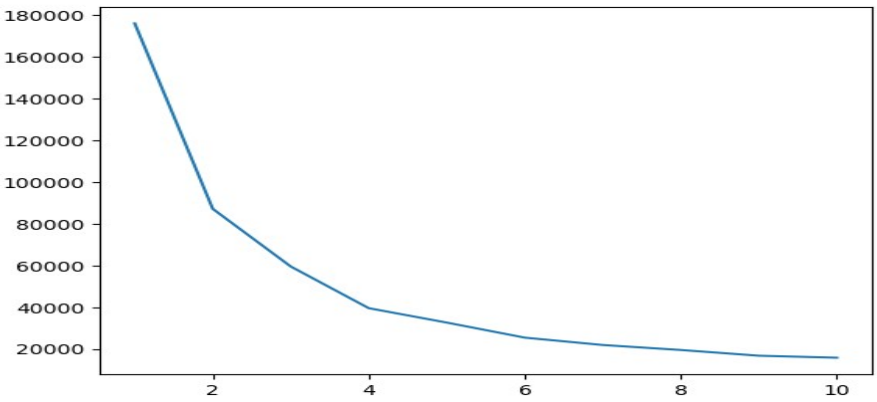
Lets plot the Clusters



Create clusters based on Annual Income + Age

First lets find the inertia scores and plot them, inertia scores show how well the dataset was clustered by K means.

```
[176108.77999999997,
 87179.72383252818,
 59525.85757379874,
 39529.88796736351,
 32620.945198551453,
 25399.02200441699,
 21873.211057947894,
 19480.630799755796,
 16737.771403145547,
 15774.944696599221]
```



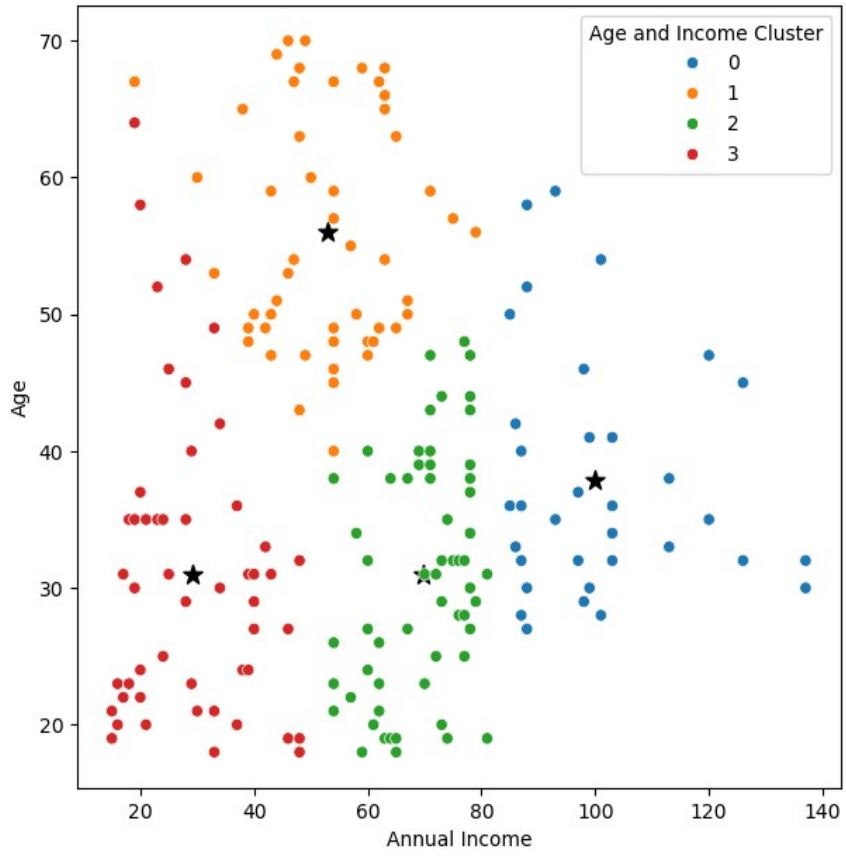
We observe that the‘elbow’ in the graph starts at Number of Cluster = 4. So 4 clusters may be the most optimal.
Lets see how many values its Cluster has

```
Age and Income Cluster
2    61
1    53
3    50
0    36
Name: count, dtype: int64
```

Lets group the data by the Spending Score Clusters and see the mean Age, Annual Income, Spending Score

Age and Income Cluster	Age	Annual Income	Spending Score
0	37.833333	99.888889	50.638889
1	56.000000	52.867925	44.490566
2	30.967213	69.852459	52.065574
3	31.020000	29.060000	53.660000

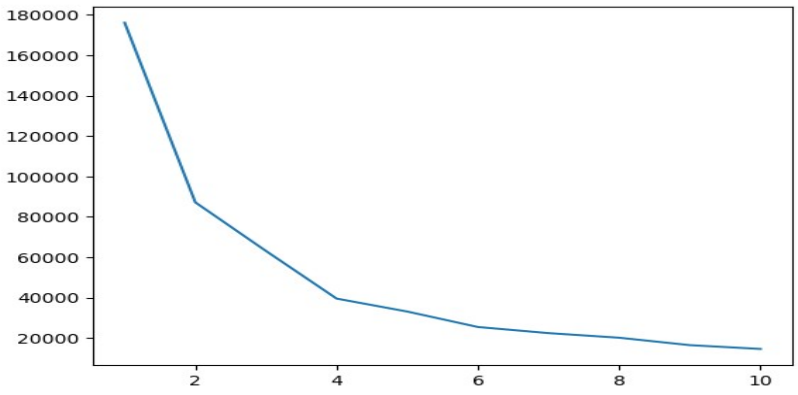
Lets plot the Clusters



Create clusters based on Age + Spending Score

First lets find the inertia scores and plot them, inertia scores show how well the dataset was clustered by K means.

```
[176108.77999999997,
 87189.04625974289,
 63120.43491305989,
 39529.88796736351,
 33115.9440925088,
 25460.547371910277,
 22434.36902025804,
 20158.357328492282,
 16486.89254721027,
 14602.440701759468]
```



We observe that the‘elbow’ in the graph starts at Number of Cluster = 5. So 5 clusters may be the most optimal.
Lets see how many values its Cluster has

```
Age and Spending Cluster
3    57
2    49
0    48
1    46
Name: count, dtype: int64
```

Lets group the data by the Spending Score Clusters and see the mean Age, Annual Income, Spending Score

Age and Spending Cluster	Age	Annual Income	Spending Score
0	43.291667	66.937500	15.020833
1	27.326087	52.282609	49.369565
2	55.408163	55.673469	48.040816
3	30.175439	66.070175	82.350877

Lets plot the Clusters