Unsupervised Learning (Customer Segmentation)

Done By: Toh Kien Yu (P2222291)

Background:
Customer Segmentation is the process to group customers who share the same characteristics in order to understand the customer's spending behavior and is an important to effectively market to a specific target audience . This will allow the shopping mall to personalize marketing to specific groups and boost profability.

We can group customers based on factors such as gender, age, income, amount spent and their interests

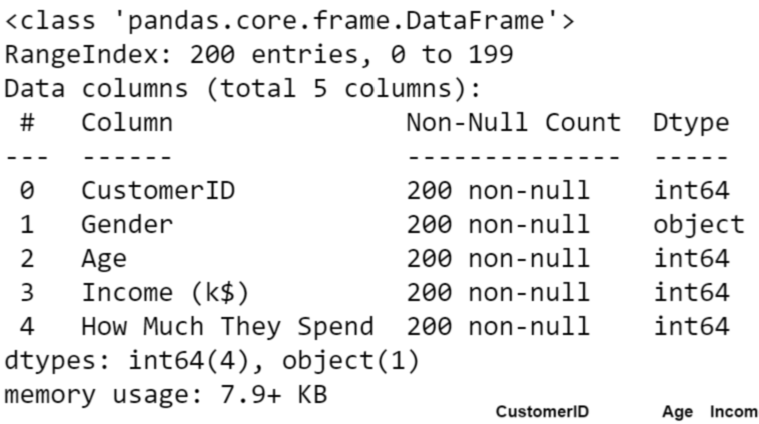
Nature of Dataset

<u>Categorical Data</u>

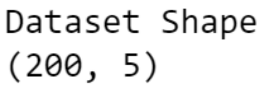
1. Gender

<u>Numeric Data</u>

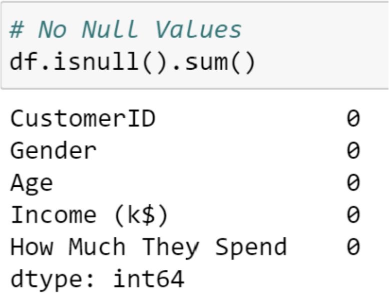
1. CustomerID
2. Age
3. Income (k$)
4. How Much They Spend



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   CustomerID          200 non-null    int64
 1   Gender              200 non-null    object
 2   Age                 200 non-null    int64
 3   Income (k$)         200 non-null    int64
 4   How Much They Spend 200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

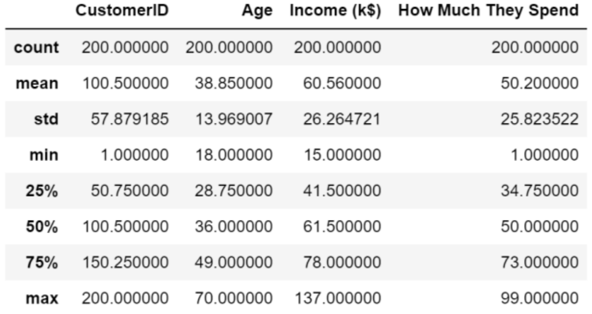


```
Dataset Shape
(200, 5)
```



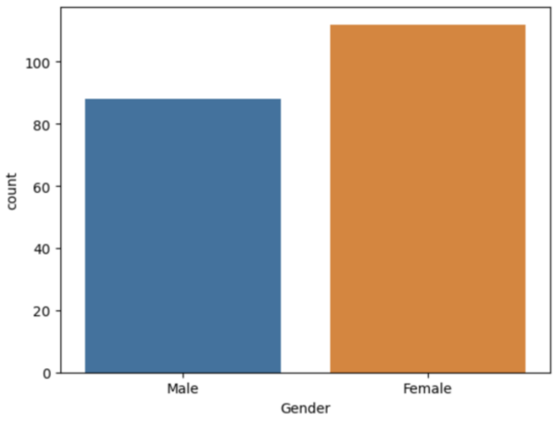
```
# No Null Values
df.isnull().sum()
```

```
CustomerID             0
Gender                 0
Age                    0
Income (k$)            0
How Much They Spend    0
dtype: int64
```

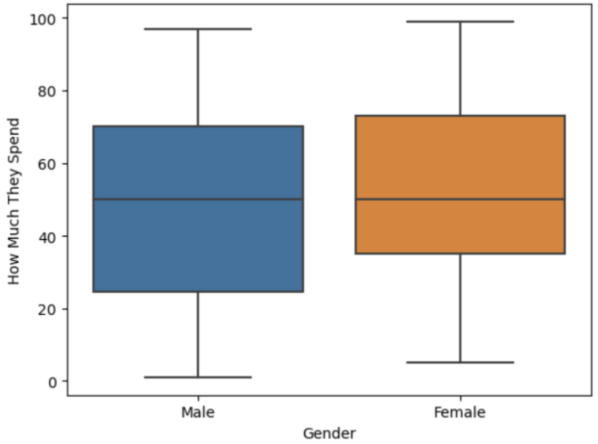


	CustomerID	Age	Income (k$)	How Much They Spend
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000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

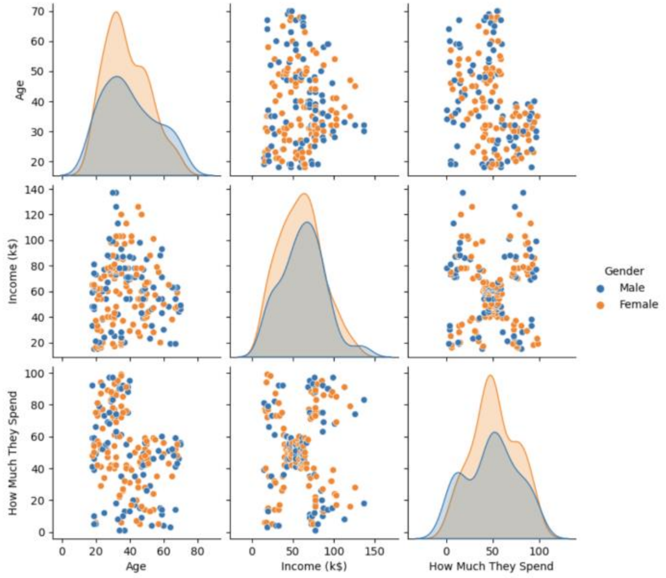
Data Exploration



There are more females than males in the dataset

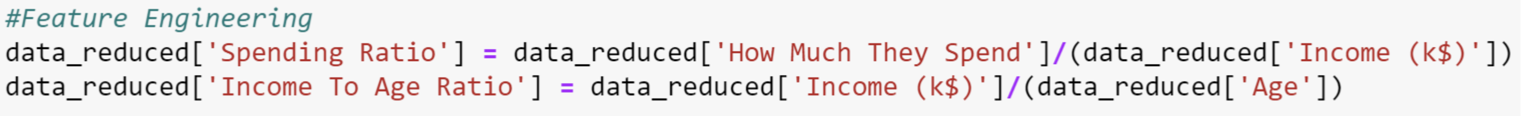


Median of how much a male and female spent is about the same.



From the pair plot, gender does not seem to have any correlation to age, income and how much they spend.

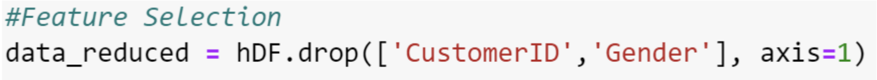
Data Preprocessing



```
#Feature Engineering
data_reduced['Spending Ratio'] = data_reduced['How Much They Spend']/(data_reduced['Income (k$)'])
data_reduced['Income To Age Ratio'] = data_reduced['Income (k$)']/(data_reduced['Age'])
```

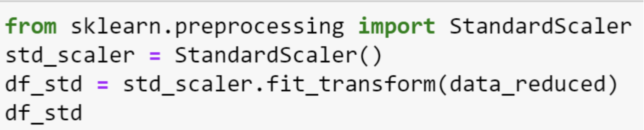
Feature engineer is the process to create additional relevant features from the existing raw features of the data to improve the learning algorithm.

1. ‘Spending Ratio’ is created to find out how much an individual is willing to allocate their income to spending.
2. ‘Income To Age Ratio’ is created to find out the relationship between one’s different age group’s income.



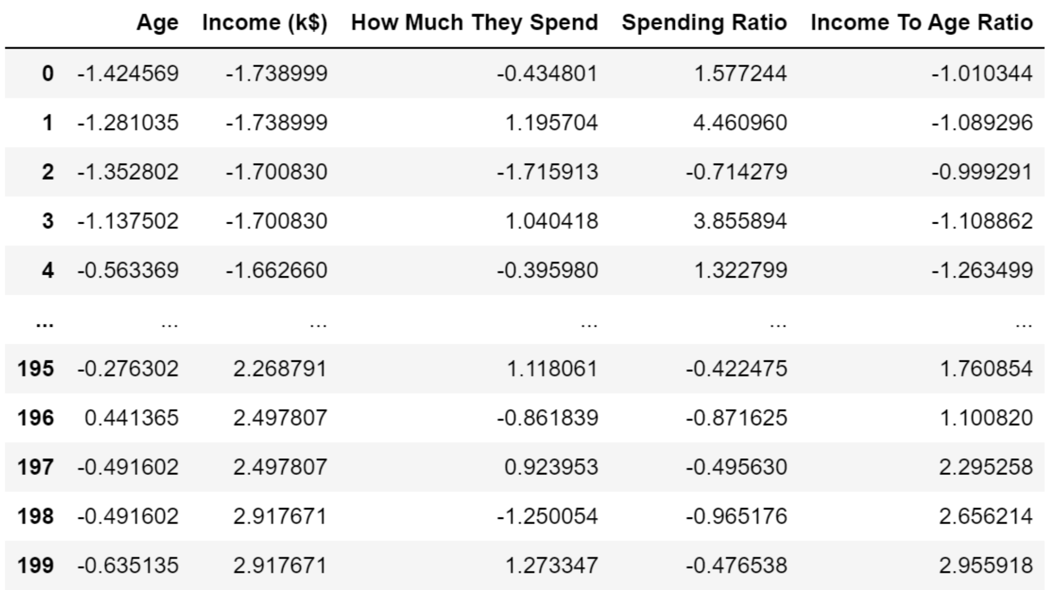
```
#Feature Selection
data_reduced = hDF.drop(['CustomerID','Gender'], axis=1)
```

- Feature Selection is done to get rid of features that are redundant by taking the best subset of the dataset.
- ‘CustomerID’ is dropped as it is an identifying column and ‘Gender’ is dropped as there is no correlation to other variables.



```
from sklearn.preprocessing import StandardScaler
std_scaler = StandardScaler()
df_std = std_scaler.fit_transform(data_reduced)
df_std
```

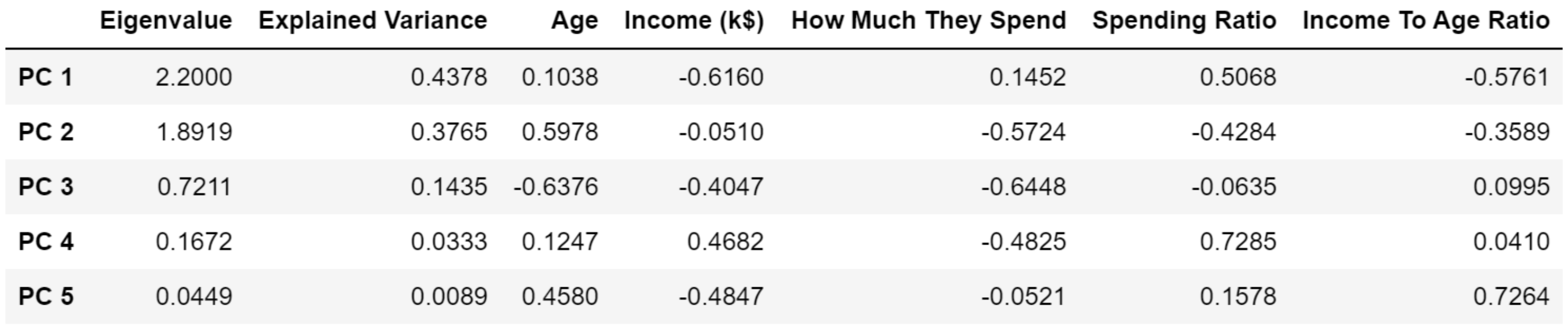
- Data is then standardized to so that the features are centred around 0 with a standard deviation of 1.



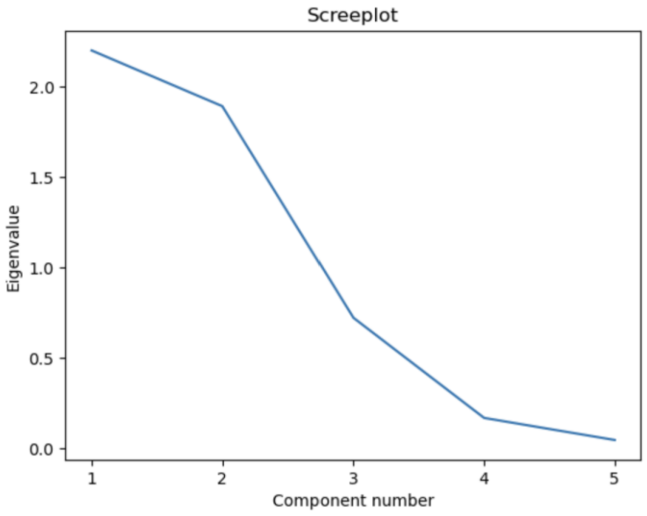
	Age	Income (k$)	How Much They Spend	Spending Ratio	Income To Age Ratio
0	-1.424569	-1.738999	-0.434801	1.577244	-1.010344
1	-1.281035	-1.738999	1.195704	4.460960	-1.089296
2	-1.352802	-1.700830	-1.715913	-0.714279	-0.999291
3	-1.137502	-1.700830	1.040418	3.855894	-1.108862
4	-0.563369	-1.662660	-0.395980	1.322799	-1.263499
...	...	...	...	...	...
195	-0.276302	2.268791	1.118061	-0.422475	1.760854
196	0.441365	2.497807	-0.861839	-0.871625	1.100820
197	-0.491602	2.497807	0.923953	-0.495630	2.295258
198	-0.491602	2.917671	-1.250054	-0.965176	2.656214
199	-0.635135	2.917671	1.273347	-0.476538	2.955918

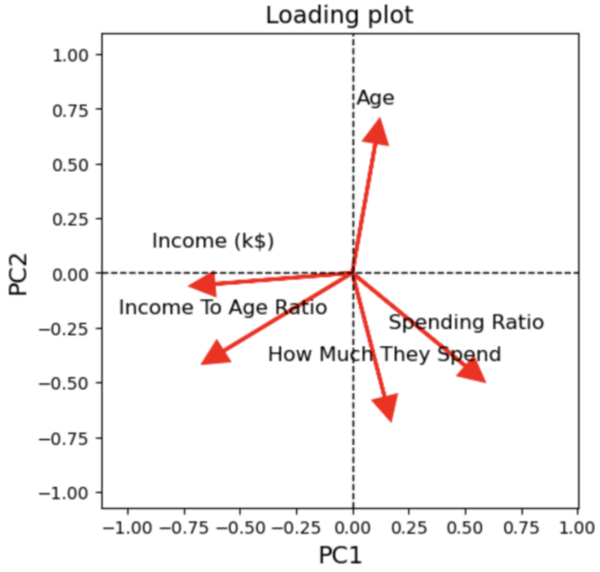
Dimension Reduction

Principal Component Analysis (PCA) is then performed into a lower dimensional space while retaining as much information as possible



	Eigenvalue	Explained Variance	Age	Income (k$)	How Much They Spend	Spending Ratio	Income To Age Ratio
PC 1	2.2000	0.4378	0.1038	-0.6160	0.1452	0.5068	-0.5761
PC 2	1.8919	0.3765	0.5978	-0.0510	-0.5724	-0.4284	-0.3589
PC 3	0.7211	0.1435	-0.6376	-0.4047	-0.6448	-0.0635	0.0995
PC 4	0.1672	0.0333	0.1247	0.4682	-0.4825	0.7285	0.0410
PC 5	0.0449	0.0089	0.4580	-0.4847	-0.0521	0.1578	0.7264



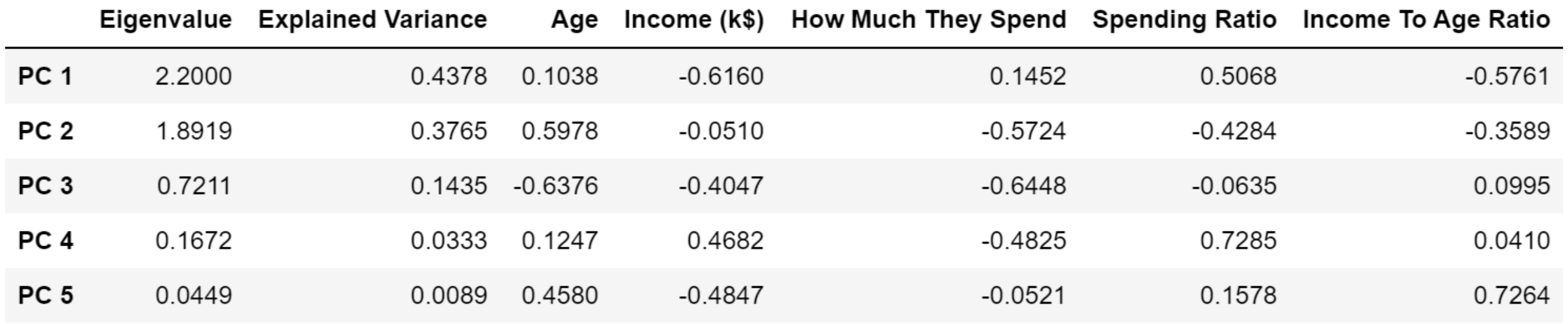


Loading Plot allows us to visualize how each variable influences PC1 and PC2

1. By Kaiser's rule, extract the first 2 PCs where eigenvalues (2.2000, 1.8919) are > 1.
2. First 2 PCs accounted for 81.43% of the total variance.
3. Scree plot shows elbow at PC3 suggesting 1st 2 PCs to extract

Hence, lets extract 2 PCs.

Dimension Reduction



	Eigenvalue	Explained Variance	Age	Income (k$)	How Much They Spend	Spending Ratio	Income To Age Ratio
PC 1	2.2000	0.4378	0.1038	-0.6160	0.1452	0.5068	-0.5761
PC 2	1.8919	0.3765	0.5978	-0.0510	-0.5724	-0.4284	-0.3589
PC 3	0.7211	0.1435	-0.6376	-0.4047	-0.6448	-0.0635	0.0995
PC 4	0.1672	0.0333	0.1247	0.4682	-0.4825	0.7285	0.0410
PC 5	0.0449	0.0089	0.4580	-0.4847	-0.0521	0.1578	0.7264

x_1: **Age**, x_2: **Income**, x_3: **How Much They Spend**, x_4: **Spending Ratio**, x_5: **Income To Age Ratio**

PC1: $0.1038x_1 - 0.6160x_2 + 0.1452x_3 + 0.5068x_4 - 0.5761x_5$

The loading on Age, How Much They Spend and Spending Ratio is opposite in sign to the other loading This PC seems to measure a contrast of Age, How Much They Spend and Spending Ratio against the remaining variables

1. Income has the highest negative loading of -0.6160 which tells us that low-income customers will score high on PC1, and high-income customers will score low in PC1.
2. Income To Age Ratio has a second highest negative loading of -0.5761 which tells us that customers with a low Income To Age Ratio score high in PC1.
3. Spending Ratio has the highest positive loading of 0.5068 which tells us that customers with high Spending Ratio will score high on PC1

PC1 mainly captures customer income levels and spending. A customer with low income and high spending ratio will score high on PC1.

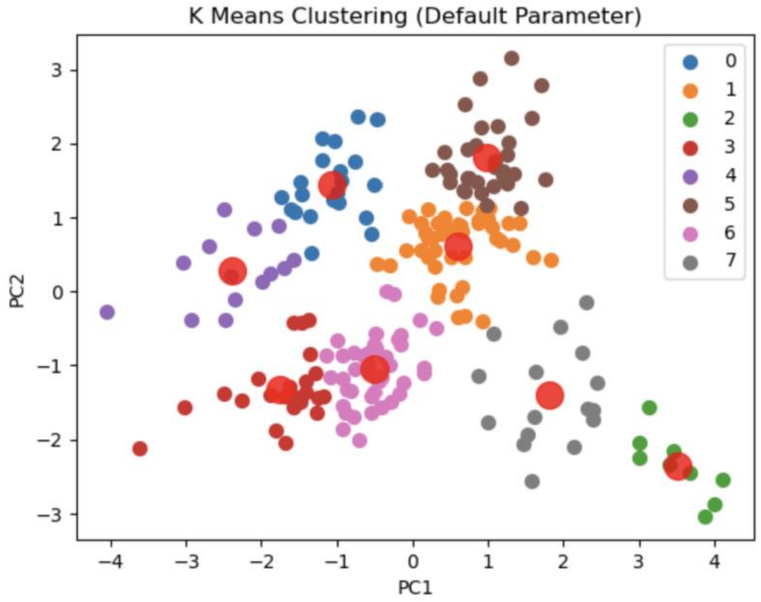
PC2: $0.5978x_1 - 0.0510x_2 - 0.5724x_3 - 0.4284x_4 - 0.3589x_5$

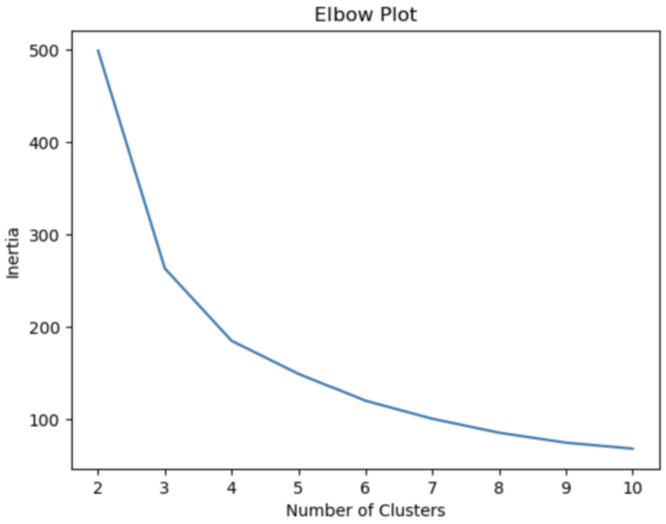
The loading on Age is opposite in sign to the other loading This PC seems to measure a contrast of Age against the remaining variables.

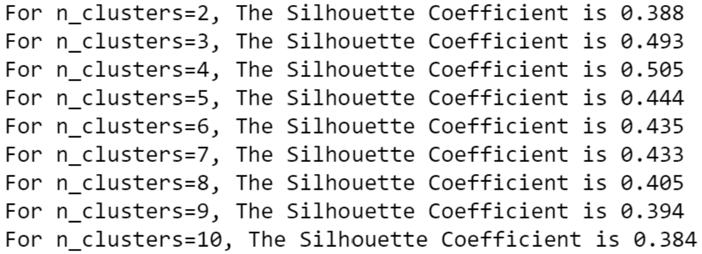
1. Age has the highest positive loading of 0.5978 which tells us that older customers will score high on PC2.
2. How Much They Spend has the highest negative loading of -0.5724 which tells us that customers with a low spending will score high in PC2.
3. Customers with a high spending ratio will score low on PC2

PC2 mainly captures customer age-related spending patterns. A customer who is older and has low spending will score high in PC2.

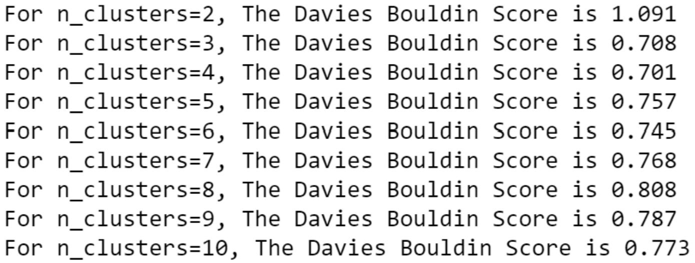
Modelling: K Means Clustering (Before Tuning)





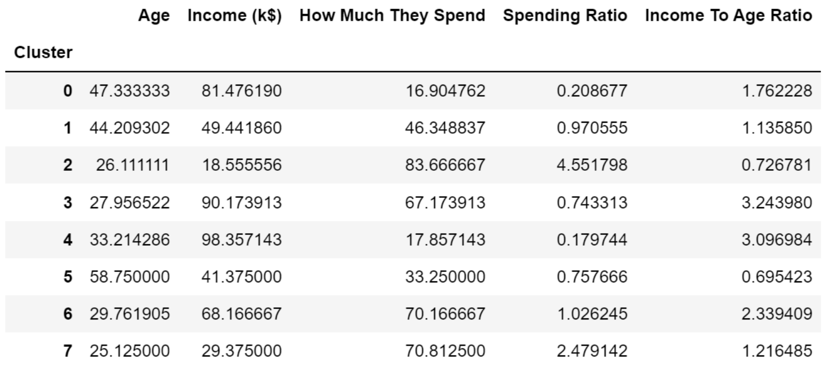


```
For n_clusters=2, The Silhouette Coefficient is 0.388
For n_clusters=3, The Silhouette Coefficient is 0.493
For n_clusters=4, The Silhouette Coefficient is 0.505
For n_clusters=5, The Silhouette Coefficient is 0.444
For n_clusters=6, The Silhouette Coefficient is 0.435
For n_clusters=7, The Silhouette Coefficient is 0.433
For n_clusters=8, The Silhouette Coefficient is 0.405
For n_clusters=9, The Silhouette Coefficient is 0.394
For n_clusters=10, The Silhouette Coefficient is 0.384
```



```
For n_clusters=2, The Davies Bouldin Score is 1.091
For n_clusters=3, The Davies Bouldin Score is 0.708
For n_clusters=4, The Davies Bouldin Score is 0.701
For n_clusters=5, The Davies Bouldin Score is 0.757
For n_clusters=6, The Davies Bouldin Score is 0.745
For n_clusters=7, The Davies Bouldin Score is 0.768
For n_clusters=8, The Davies Bouldin Score is 0.808
For n_clusters=9, The Davies Bouldin Score is 0.787
For n_clusters=10, The Davies Bouldin Score is 0.773
```

Customer’s Profile for different clusters



	Age	Income (k$)	How Much They Spend	Spending Ratio	Income To Age Ratio
Cluster					
0	47.333333	81.476190	16.904762	0.208677	1.762228
1	44.209302	49.441860	46.348837	0.970555	1.135850
2	26.111111	18.555556	83.666667	4.551798	0.726781
3	27.956522	90.173913	67.173913	0.743313	3.243980
4	33.214286	98.357143	17.857143	0.179744	3.096984
5	58.750000	41.375000	33.250000	0.757666	0.695423
6	29.761905	68.166667	70.166667	1.026245	2.339409
7	25.125000	29.375000	70.812500	2.479142	1.216485

The default parameters for K Means Clustering groups customers into 8 clusters.

Cluster 2 and 7 depicts young customers who spends a lot but have low income. Although Cluster 2 has a higher spending ratio than Cluster 7, Cluster 7 has a higher income than Cluster 2

Cluster 3 and 6 shows young customers who spends a lot but have a high income. Cluster 3 has a higher income than Cluster 6

Cluster 1 and 5 represents older customers with average income and moderate spending. Cluster 5 made up of senior citizens and Cluster 1 is made up of working adults

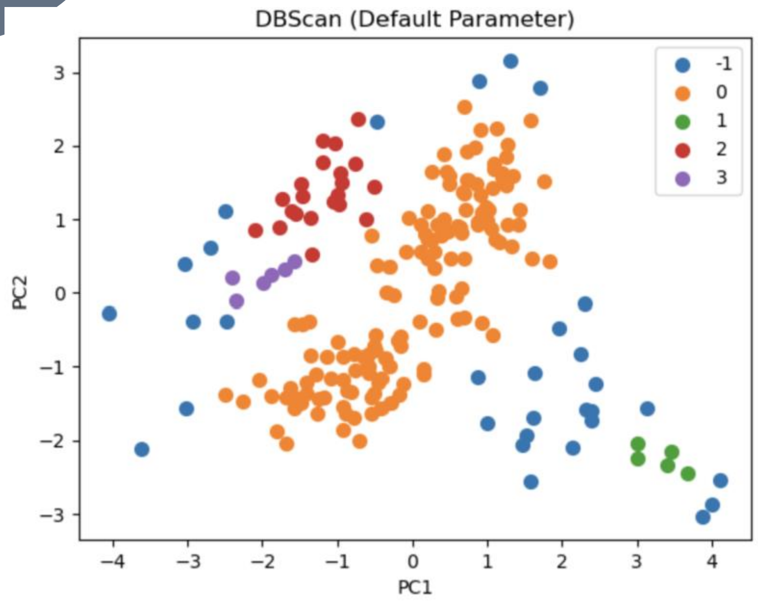
Cluster 4 and 0 are customers with very high income and low spending. Cluster 0 is made up of older individuals compared to Cluster 4

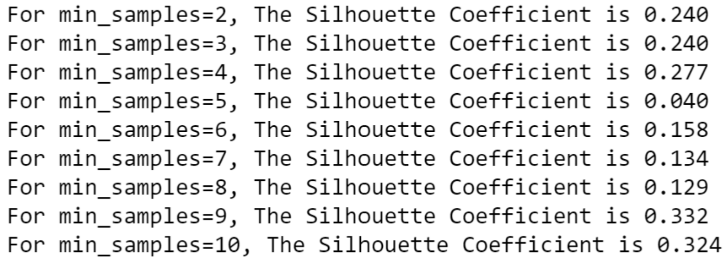
We will choose N Cluster = 4 as,

1. It has the highest silhouette coefficient of 0.505.
2. The elbow plots show an elbow at 5, which suggests to choose N Cluster = 4
3. Davies Bound-in Score is the lowest at N Cluster = 4

Modelling: DBScan and Hierarchical Clustering (Before Tuning)

DBScan

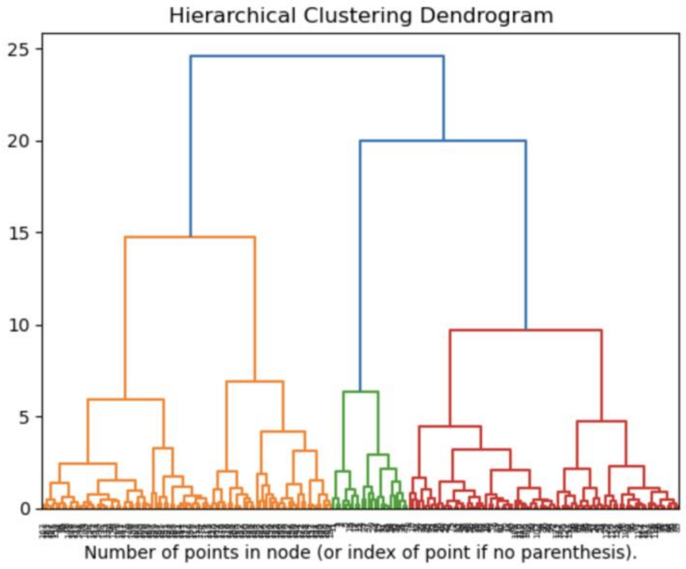


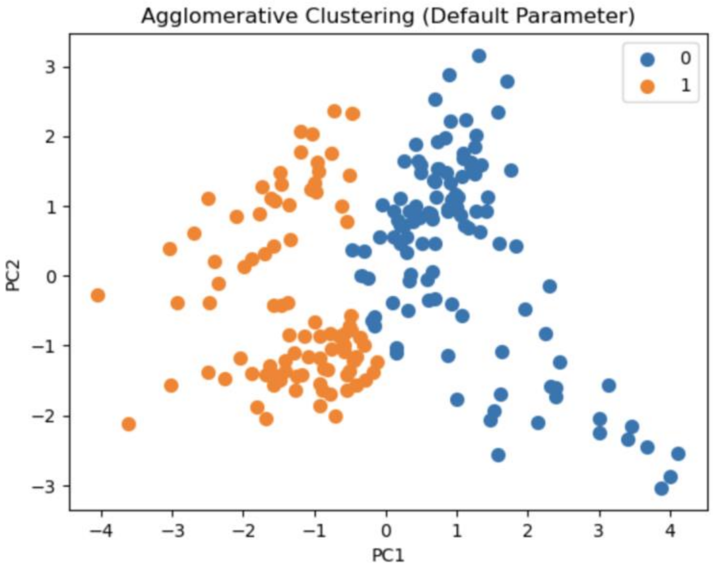


```
For min_samples=2, The Silhouette Coefficient is 0.240
For min_samples=3, The Silhouette Coefficient is 0.240
For min_samples=4, The Silhouette Coefficient is 0.277
For min_samples=5, The Silhouette Coefficient is 0.040
For min_samples=6, The Silhouette Coefficient is 0.158
For min_samples=7, The Silhouette Coefficient is 0.134
For min_samples=8, The Silhouette Coefficient is 0.129
For min_samples=9, The Silhouette Coefficient is 0.332
For min_samples=10, The Silhouette Coefficient is 0.324
```

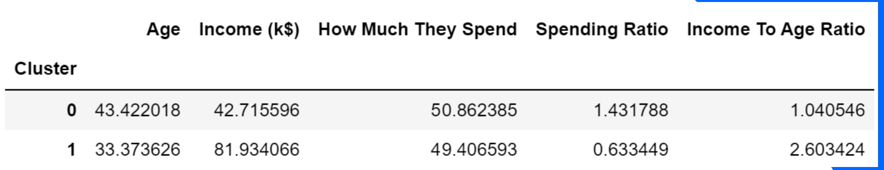
Highest silhouette score is only 0.332 which is not very high.
Hence, DBScan is not effective for customer segmentation due to the dataset having large difference of density

Hierarchical Clustering (Agglomerative Clustering)

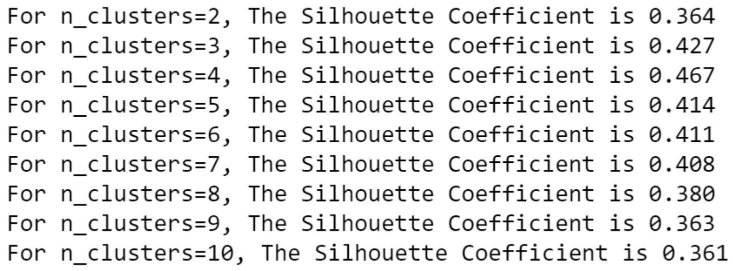




Agglomerative Clustering's Customer Profile



	Age	Income (k$)	How Much They Spend	Spending Ratio	Income To Age Ratio
Cluster					
0	43.422018	42.715596	50.862385	1.431788	1.040546
1	33.373626	81.934066	49.406593	0.633449	2.603424



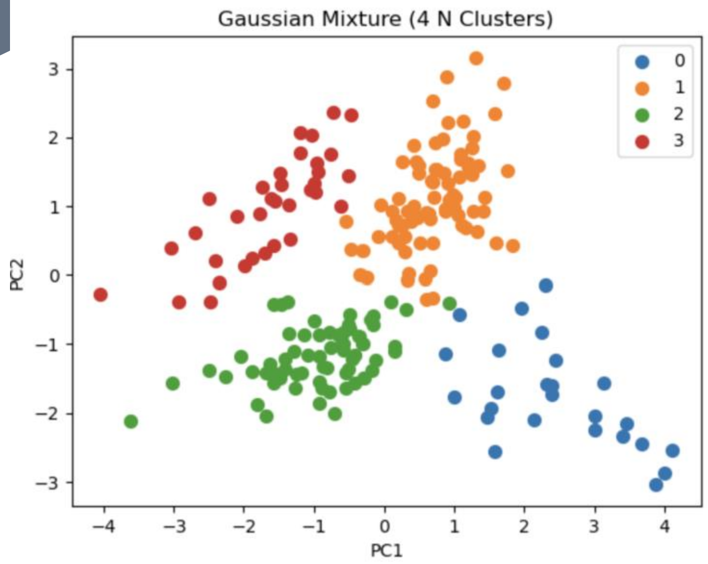
```
For n_clusters=2, The Silhouette Coefficient is 0.364
For n_clusters=3, The Silhouette Coefficient is 0.427
For n_clusters=4, The Silhouette Coefficient is 0.467
For n_clusters=5, The Silhouette Coefficient is 0.414
For n_clusters=6, The Silhouette Coefficient is 0.411
For n_clusters=7, The Silhouette Coefficient is 0.408
For n_clusters=8, The Silhouette Coefficient is 0.380
For n_clusters=9, The Silhouette Coefficient is 0.363
For n_clusters=10, The Silhouette Coefficient is 0.361
```

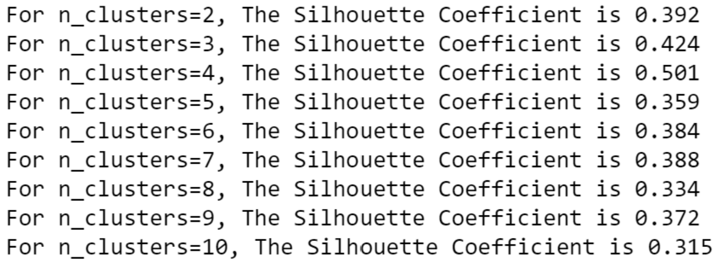
- Customers in Cluster 0 have low income and high spending ratio
- Customers in Cluster 1 have high income and low spending ratio

4 Clusters are produced, upon cutting a straight line at 10 on the Dendogram.

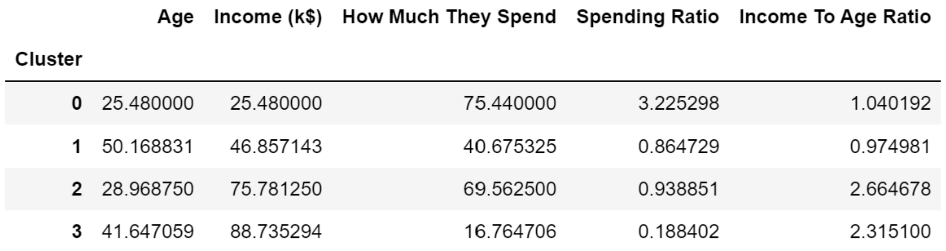
We will choose N Cluster = 4 as it has the highest silhouette coefficient of 0.467

Modelling: Gaussian Mixture Model (After Tuned)





```
For n_clusters=2, The Silhouette Coefficient is 0.392
For n_clusters=3, The Silhouette Coefficient is 0.424
For n_clusters=4, The Silhouette Coefficient is 0.501
For n_clusters=5, The Silhouette Coefficient is 0.359
For n_clusters=6, The Silhouette Coefficient is 0.384
For n_clusters=7, The Silhouette Coefficient is 0.388
For n_clusters=8, The Silhouette Coefficient is 0.334
For n_clusters=9, The Silhouette Coefficient is 0.372
For n_clusters=10, The Silhouette Coefficient is 0.315
```



	Age	Income (k$)	How Much They Spend	Spending Ratio	Income To Age Ratio
Cluster					
0	25.480000	25.480000	75.440000	3.225298	1.040192
1	50.168831	46.857143	40.675325	0.864729	0.974981
2	28.968750	75.781250	69.562500	0.938851	2.664678
3	41.647059	88.735294	16.764706	0.188402	2.315100

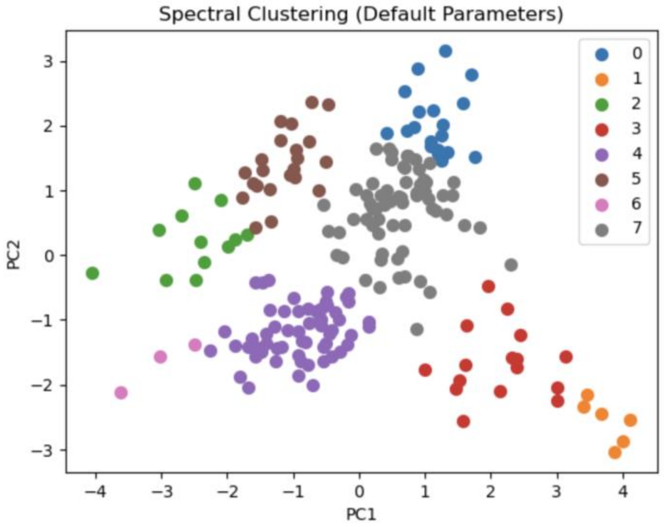
As the default parameter for N Cluster for Gaussian Mixture is 1. There is not much insights for default parameters. We will look at Gaussian Mixture with 4 N Clusters since it has the highest silhouette coefficient of 0.501.

Cluster 0: Young And Undisciplined Customers. Despite, these customers having a low income, they are big spenders and have the highest spending ratio.

Cluster 1: Senior Citizens. These customer have moderate spending and have average income.

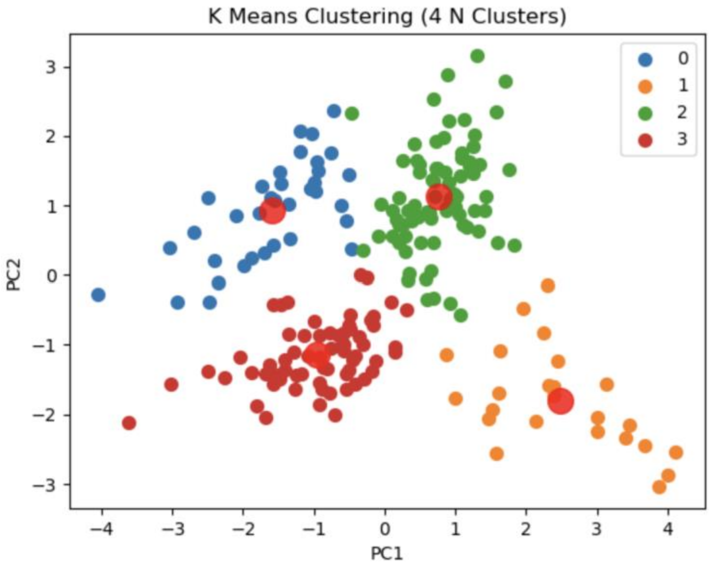
Cluster 2: Young and High-Income Customers. These customers are from the above average income group and have a high spending.

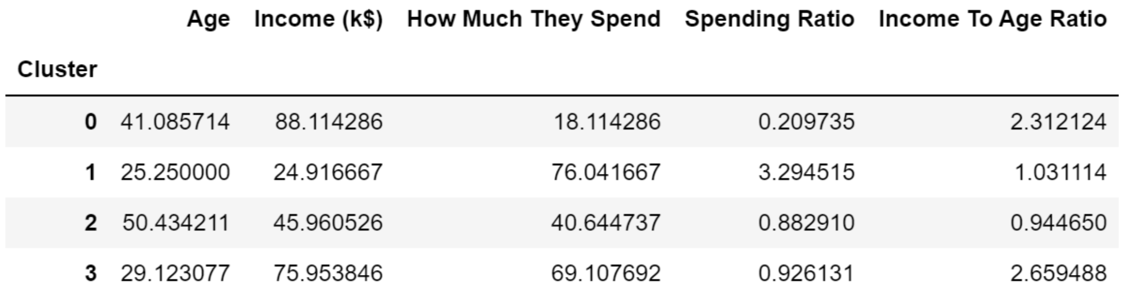
Cluster 3: Middle-Aged Citizens and High Income. These customers have a low spending ratio despite having a high income. They are financially stable and knows how to budget well.



I have also looked at Spectral Clustering but decided not to use it as it is quite sensitive to noise and outliers.

Final Model: K Means Clustering (After Hyperparameter Tuning)





Cluster	Age	Income (k$)	How Much They Spend	Spending Ratio	Income To Age Ratio
0	41.085714	88.114286	18.114286	0.209735	2.312124
1	25.250000	24.916667	76.041667	3.294515	1.031114
2	50.434211	45.960526	40.644737	0.882910	0.944650
3	29.123077	75.953846	69.107692	0.926131	2.659488

For my final model, I decided to choose K-Means Clustering as

1. It has the highest silhouette coefficient of 0.505
2. It has the lowest Davies Bouldin Score of 0.701
3. K-Means is easy to implement and interpret.

Interpretation of Clusters:

Cluster 0: Middle-Aged Citizens and High Income. These customers have a low spending ratio despite having a high income. They are financially stable and knows how to budget well.

Cluster 1: Young And Undisciplined Customers. Despite, these customers having a low income, they are big spenders and have the highest spending ratio.

Cluster 2: Senior Citizens. These customer have moderate spending and have average income.

Cluster 3: Young and High-Income Customers. These customers are from the above average income group and have a high spending

Conclusion

Cluster 0 and 3 from the K-Means Algorithm are the most valuable segment to the shopping mall.
Cluster 0 holds a group of middle-aged customers with high income but not spending much.
Cluster 3 holds customers with high income and have high spending.

Recommendations:

1. The shopping mall can induce Cluster 0 to spend more by targeting advertisements and loyalty programs that offer discounts to attract them to frequently shop in the mall.

2. As for Cluster 3, these customers have high spending habits. The mall can gather customer feedbacks to enhance the shopping experience and create a special Membership for these customers in Cluster 3 for benefits such as discounts

This customer segmentation journey has been very fruitful, I have learned more in depth about how I can apply unsupervised learning to real world applications.