Course Final Project on Unsupervised Learning

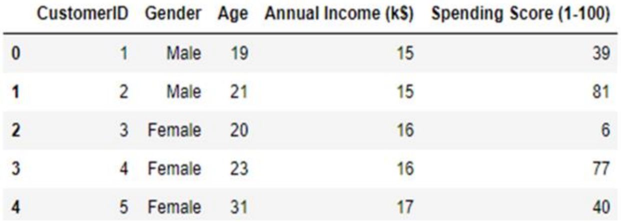
Main Objectives:

1. To Cluster the Customers based on their behavior
2. To draw insights from "Mall Customer" Data to increase the sales
3. Applying the clustering algos & choosing the best one among them

About The Data:

It is "Mall Customer Data" that contains the customers details like Customer ID, age, gender, annual income and spending score.

1. **Customer ID**: Unique ID assigned to the customer.
2. **Age**: Age of the customer
3. **Gender**: Gender of the customer
4. **Annual Income**: Annual Income of the customer in "Thousands Dollar"
5. **Spending Score**(1-100): Score assigned by the mall based on customer behavior and spending nature.

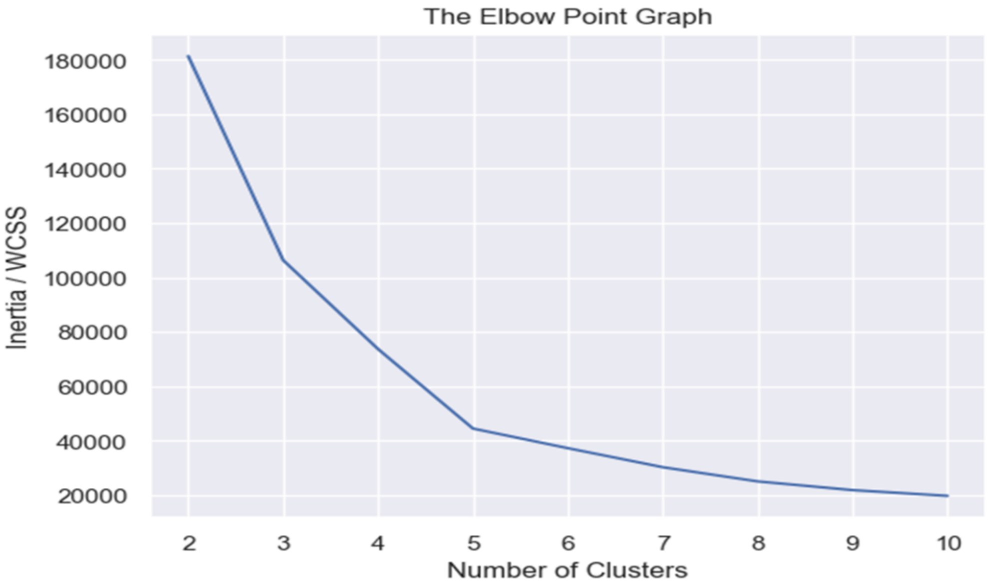


	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

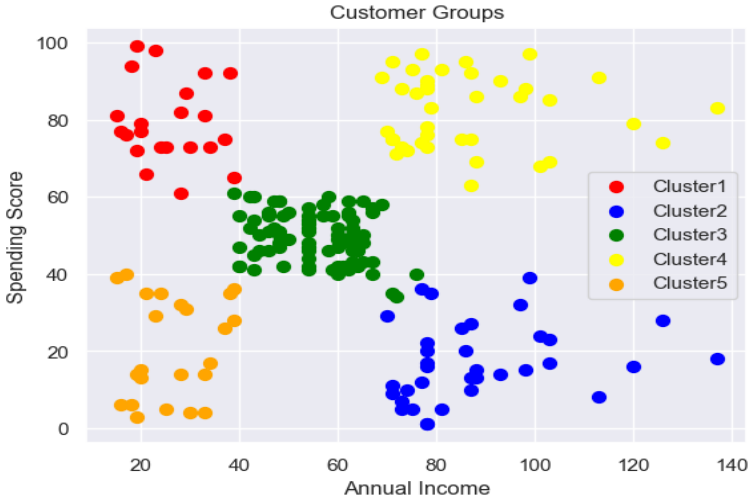
But here, we will be using "Annual Income" and "Spending Score" columns to form the clusters of customers.

Applying the Clustering Algorithms:

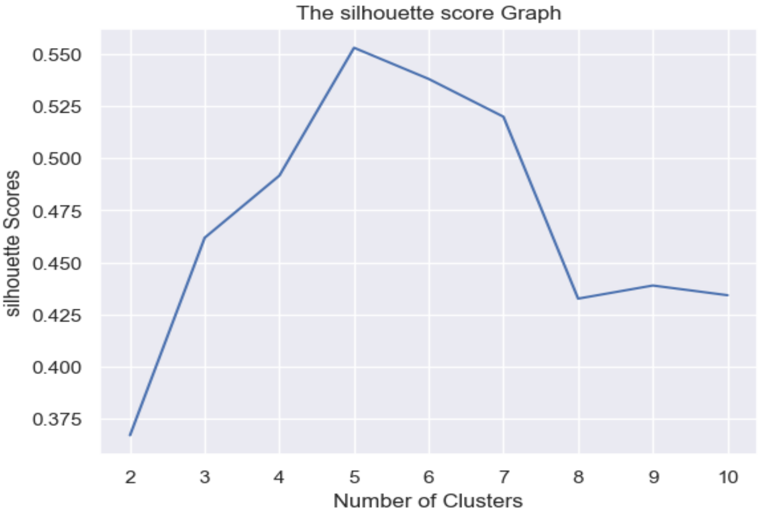
1. **K-Means**: K-Means algorithm wants the optimum value of "k" a.k.a number of clusters. So, The "Elbow Method" is used for the same.



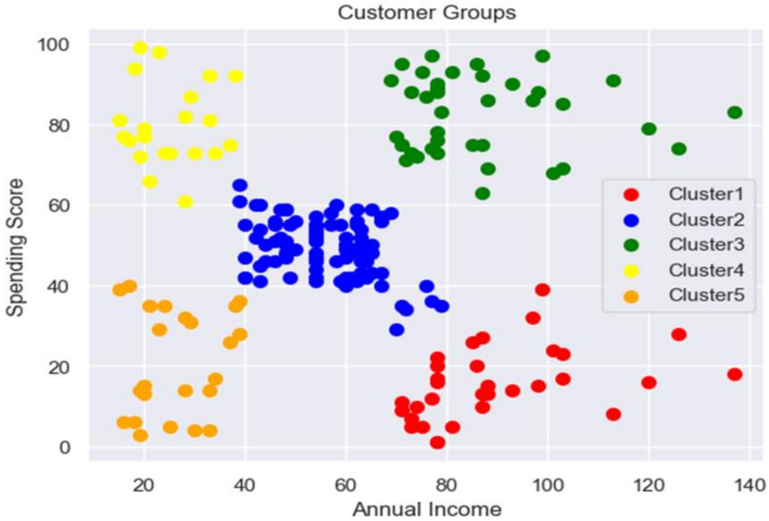
Upon observing the above graph, I have used "k=5" in K-Means algorithm which gives me the clusters as shown below:



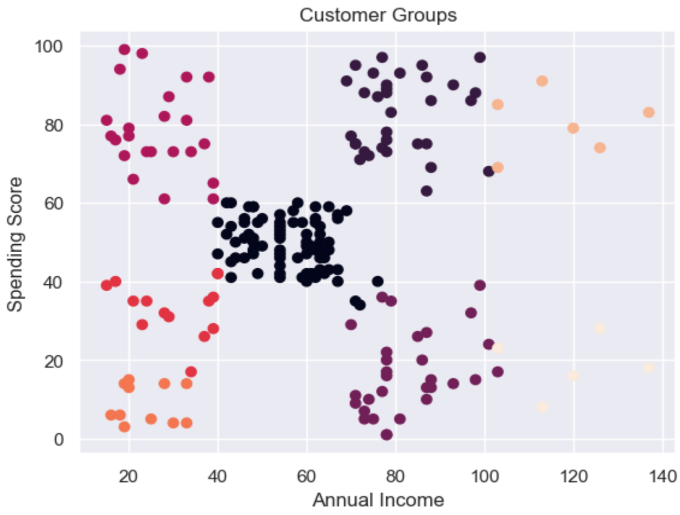
2. **Agglomerative Clustering**: It also demands the number of clusters. So, silhouette score is used to find the optimum value of "k". Upon exploration, I have found k=5 as best value.



Now, let's see the clusters made by the "Agglomerative clustering" algorithm.

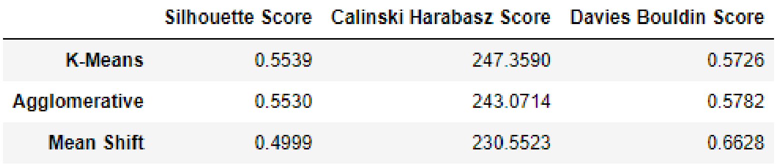


3. **Mean Shift Clustering Algorithm**: This algorithm doesn't have any "k" parameter. But, there is one parameter "bandwidth" that I have set to "15.21". It has formed 8 clusters. The Clusters looked like below:



Choosing the Best Model:

For choosing the best among three models, three metrics are used namely: "Silhouette Score", "Calinski Harabasz Score" and "Davies Bouldin Score". The Score Data Frame is given below:



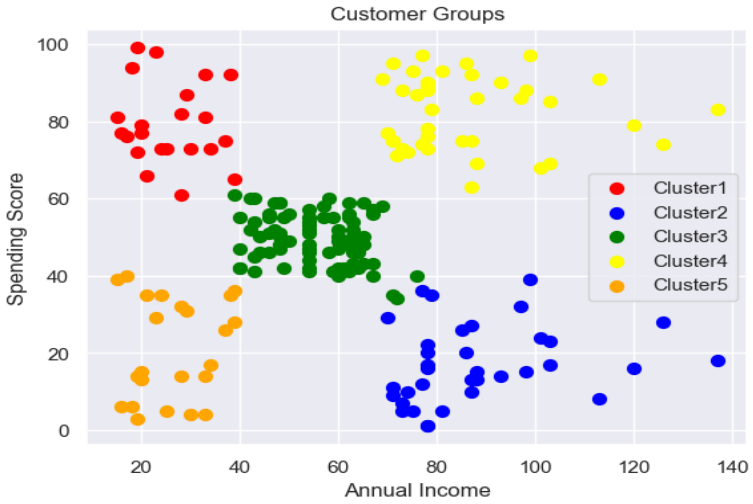
	Silhouette Score	Calinski Harabasz Score	Davies Bouldin Score
K-Means	0.5539	247.3590	0.5726
Agglomerative	0.5530	243.0714	0.5782
Mean Shift	0.4999	230.5523	0.6628

This Data Frame shows the values of three scores for three different model (i.e. K-Means, Agglomerative Clustering, Mean Shift Algorithm).

From this data frame, it is clear that the K-Means algorithm performs the best in comparison to other two algorithm. **So, The winner is K-Means algorithm**.

Key Findings:

Since, the winner is “K-Means” algorithm, let’s look again the clusters formed by K-Means.



It has formed 5 clusters. But our focus should be on “Cluster1” and “Cluster4”. These two clusters are showing those group of customers that are having the high “Spending Score” i.e. spending more money in the Mall.

So, If we target those customers of “Cluster1” and “Cluster4” by attract them via giving gifts or providing other benefits(like discounts) then it can surely boost the sales of that particular Mall.

Further Improvements:

Since two columns named “Annual Income” and “Spending Score” are used for clustering, we can also include “Age” and “Gender” for clustering to gain further insights on the behavior of customers.

So, we can probably be able to target the potential customer of a particular age group and gender type to increase the sales.