Project Report for Clustering on Mall Customer Segmentation Dataset

Sharon Saronian

The goal of this project is to do clustering using different methods on the Mall Customer Segmentation dataset. For this specific supermarket mall dataset, we have some basic data about it's customers which are gathered through membership cards. The goal of our project is to use clustering methods on the data in order to gain insight and identify different types of customers that shop from the mall, so that different marketing strategies can later be used for different types. In other words, we try to locate the customers which will benefit the mall the most so that marketing strategies and methods can be used for them to improve the profiting and sales of the mall through them.

As mentioned above, different clustering methods will be used. For each method, we will try to obtain the optimal number of clusters and then do the clustering algorithms based on that. Then we will plot the clusters and the centroids and see what information can be interpreted.

The methods used will be the Kmeans algorithm, and the Agglomerative Clustering methods from Hierarchial Clustering algorithms.

I. Dataset Information

The dataset for this project is not from UCI, but from Udemy's Machine Learning course(https://github.com/SteffiPeTaffy/machineLearningAZ/blob/master/Machine%20Learning%20A-Z%20Template%20Folder/Part%204%20-%20Clustering/Section%2025%20-%20Hierarchical%20Clustering/Mall_Customers.csv).

There are 5 attributes, the first 3 being nominal, and the later 2 being numerical in the following form:

1. CustomerID: Unique ID assigned to the customer
2. Gender: Gender of the customer
3. Age: Age of the customer

4. Annual Income(k$): Annual income of the customer
5. Spending Score (1-100): Score assigned by the mall based on customer behavior and spending

There are no label attributes, as this is a clustering unsupervised dataset.

According to the metadata, there are 200 instances and no missing values.

II. Clustering

Before we apply any algorithms, let's see what the data looks like and what statistical information it has:

	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

	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

False

From all the attributes we should decide which ones to select for the clustering task. CustomerID won't do us any help so we'll put that aside. The same can be said for Gender. The three features left could be the most useful for the clustering. In other words, we can cluster the customers according to their age and spending score, or their income and spending score. If we use the first option, we will know how much different age groups of customers spend in the mall, and for the second option, we will identify customers with different incomes and how much they spend. Both of the options give us valuable information which can later be used to plan different strategies, whereas knowing how much each gender spends doesn't give us much:

	Annual Income (k$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

(200, 2)

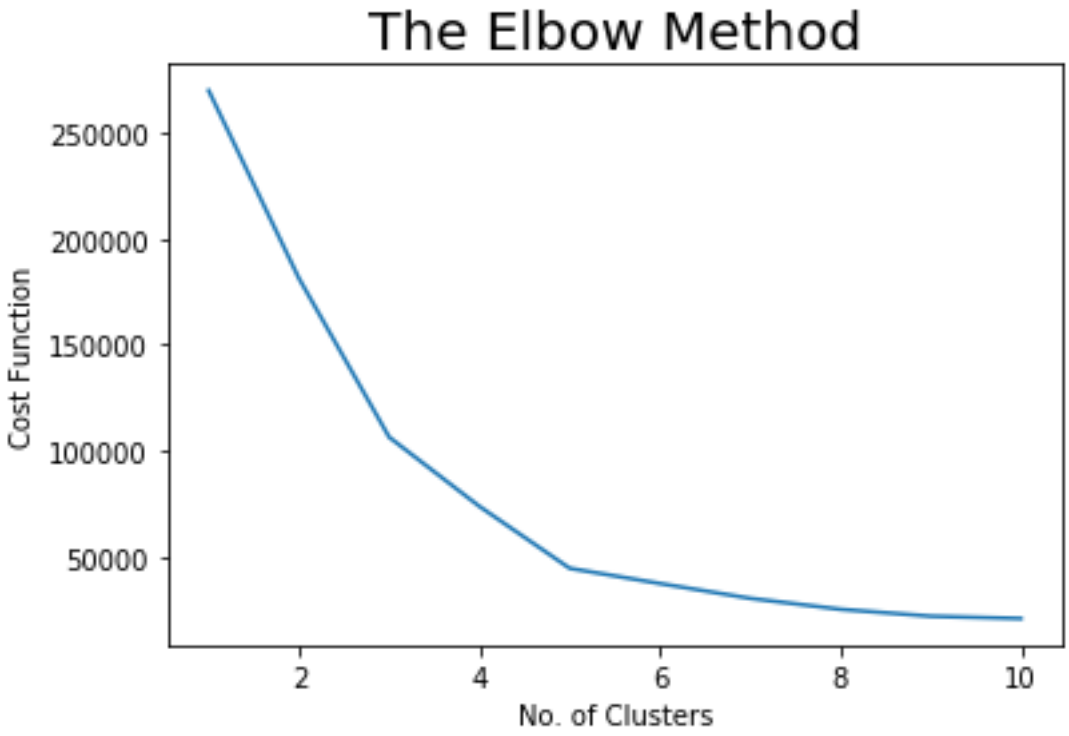
	Age	Spending Score (1-100)
0	19	39
1	21	81
2	20	6
3	23	77
4	31	40

(200, 2)

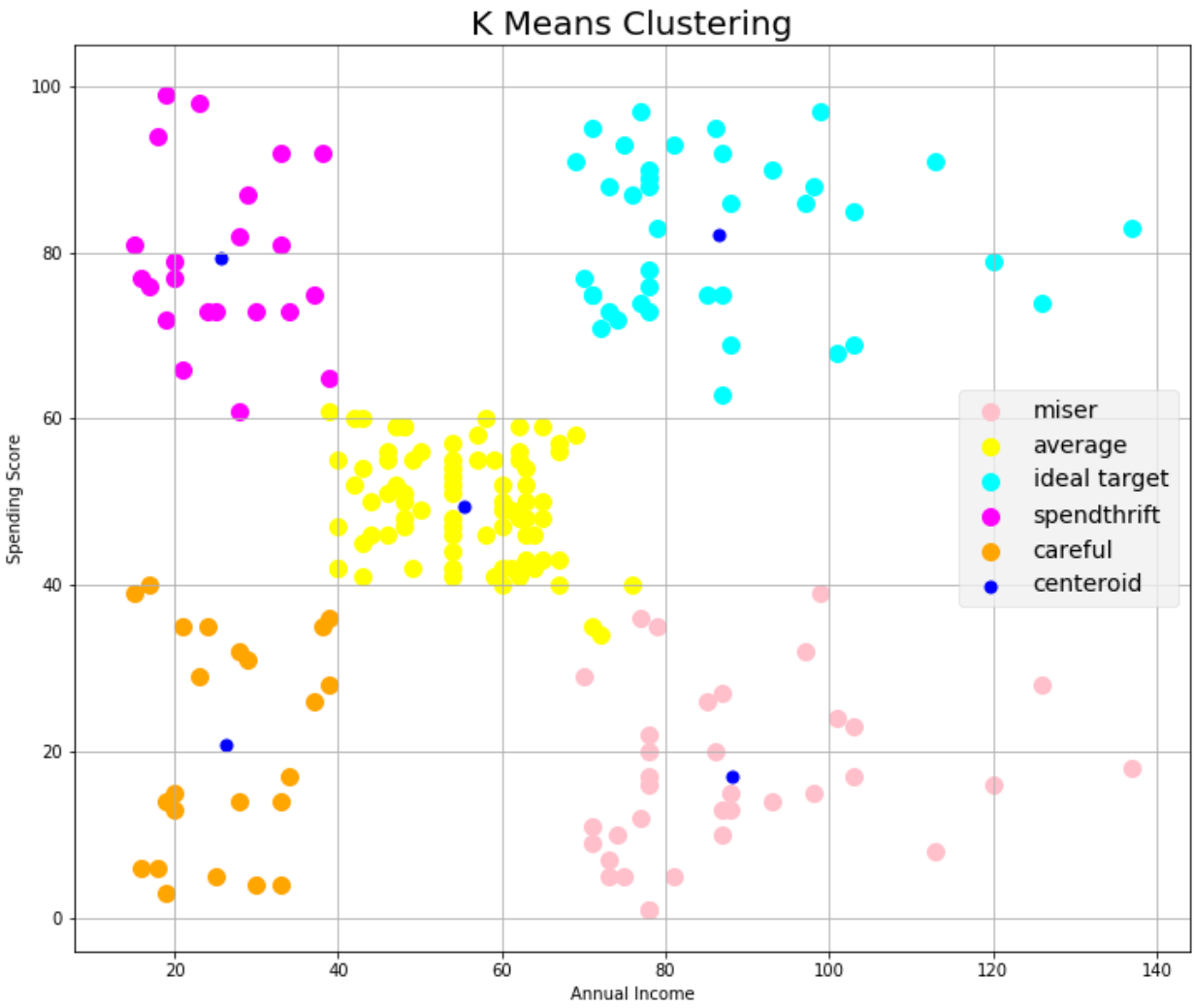
First, we will implement Kmeans and Hierarchial clustering for Annual Income and Spending Score.

The first thing to do in Kmeans is to determine the optimal number of clusters. For that matter, the algorithm calculates the loss function for different K values and plots them. A sharp change in the plot means that by increasing the K, the difference between points increases significantly, which means that the K before the sharp change is the optimal K. This method is known as the elbow method.

Below is the result from implementing the Kmeans with the elbow method on Annual Income and Spending Score:



The figure suggests that the optimal number of clusters is 5. So, we'll apply 5-means clustering on income and spending score and plot the clusters and centroids:



The clustering gives us a very valuable insight about the diverse segments of the customers. There are five segments or groups which I've named miser(خسيس), average, ideal target, spendthrift(ولخرج), and careful customers. Let's see what interpretation we can make of these clusters.

The customers in the careful cluster have low income and they don't spend much in the mall. So, they're careful customers, meaning they spend very carefully and proportional to their income.

The customers in the spendthrift cluster have low incomes as well, but they spend very much. That's what makes them spendthrifts. They know they have a low income but spend a lot regardless.

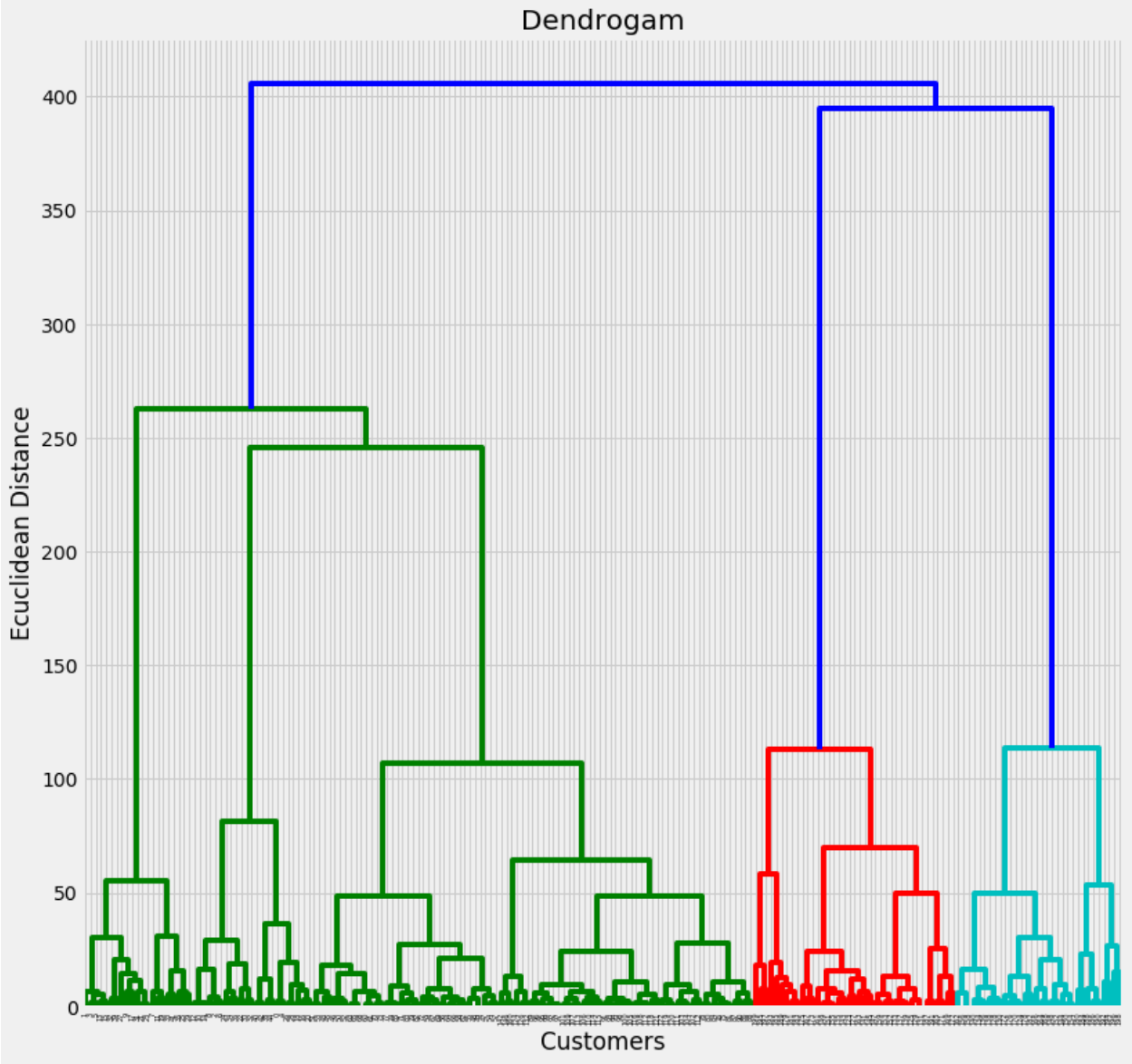
The average cluster are the normal customers you can find in any mall, the have average income and they're spending score is average as well. So, they're just regular customers.

The miser cluster have very high incomes but they don't spend much. That's why they're miser.

The most desirable and interesting cluster are the ideal targets. The have a very high income and they also spend a lot. So, this could be the most profitable group of customers any mall can get and identifying them is crucial.

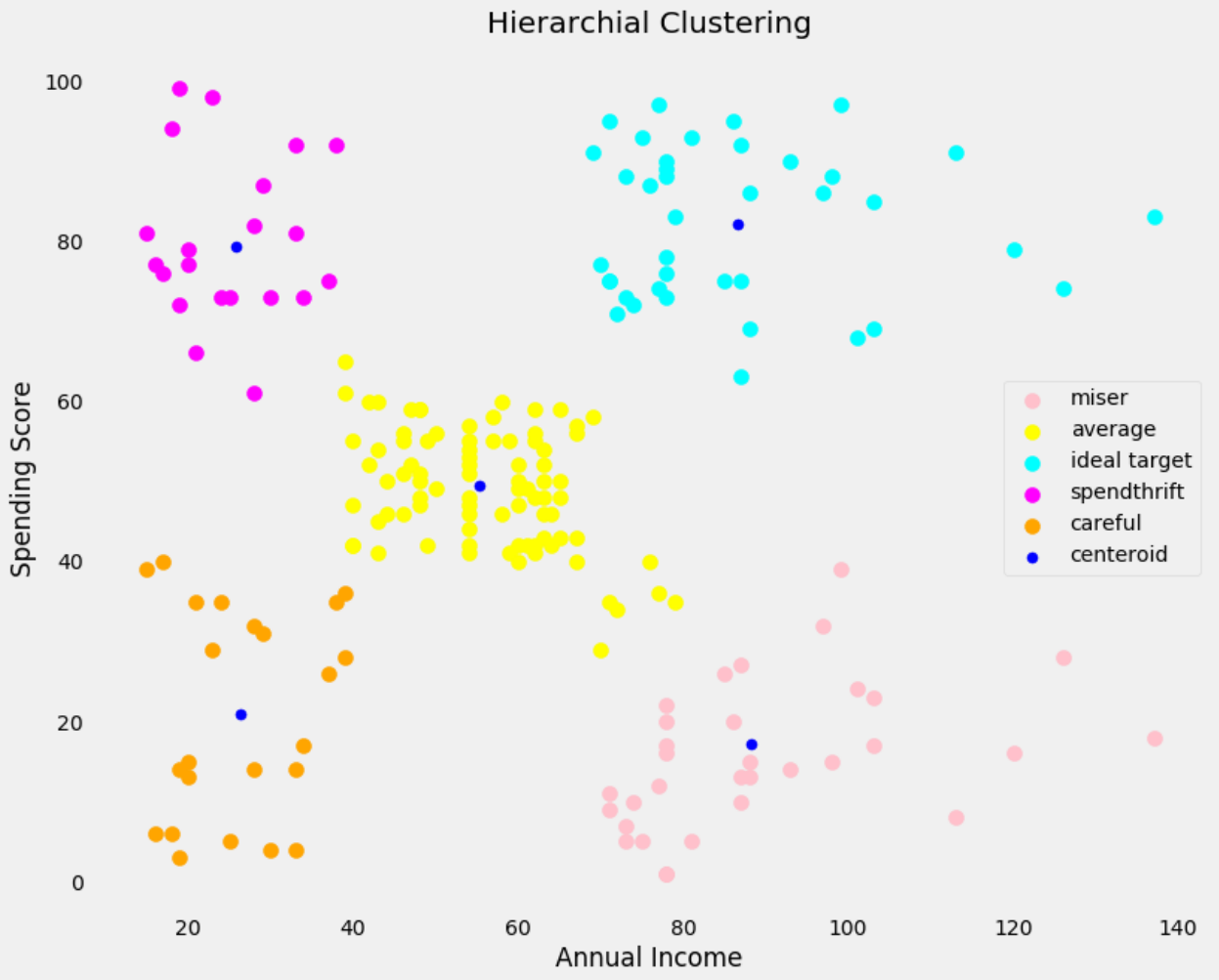
For the next part, we'll try hierarchial clustering on income and spending score and see if it gives the same results of the Kmeans algorithm.

This time we'll use Dendograms to find the number of clusters:



A sharp change in the vertical line of the dendogram is where the distance suddenly increases. Similar to the Elbow method, this is where we should stop. In our figure, the stopping point is just a little above the 100. Similar to the previous part, the optimal number of clusters are 5.

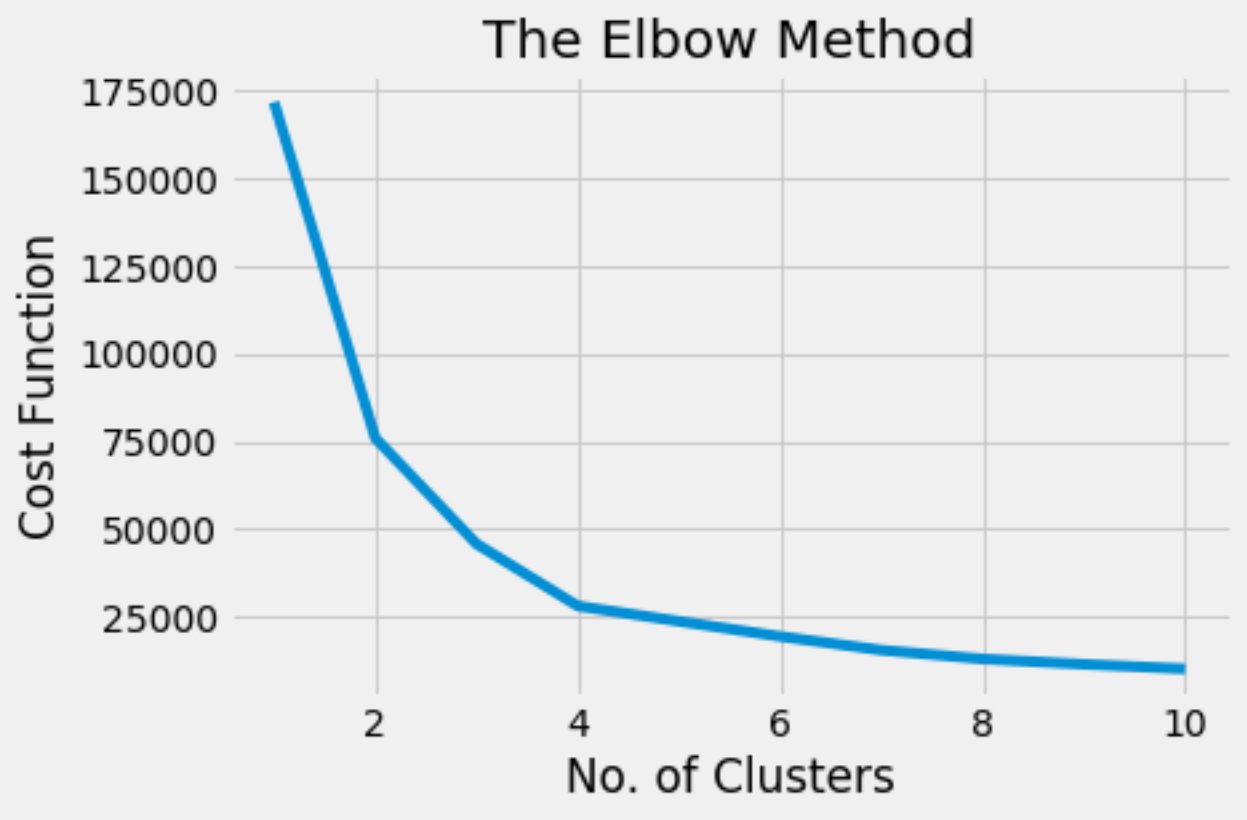
Now, we'll do Agglomerative clustering with 5 clusters and visualize it's clusters:



We can see that the clusters are the same as the Kmeans method and they both suggest that we've done the right thing and have found our information correctly.

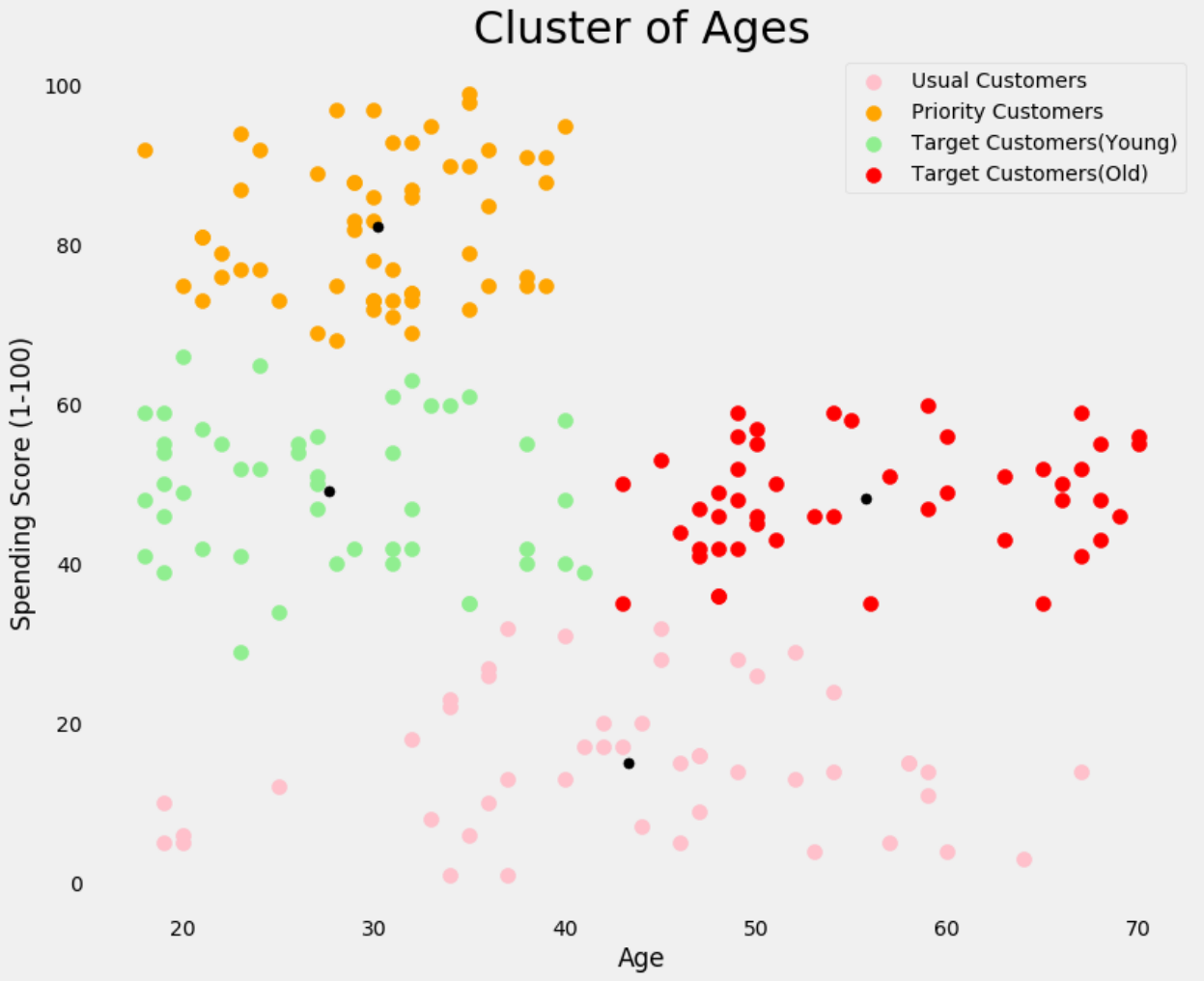
For the last part we'll apply Kmeans to age and spending score and see what results we get.

Again, we should use the elbow method for the number of customers:



The figure suggests that we have 4 clusters according to age and spending score.

It’s time to apply 4-means clustering and visualize the clusters:



This figure shows us the 4 clusters based on Age and spending score. I've named the clusters Usual customers, Priority customers, young target customers, and old target customers.

The customers in the usual cluster are from ages 20-70 (all ranges) and their spending score is low. So, this cluster doesn't give us much. It just shows that we have people from different ages who don't spend that much.

The target customer clusters are young and old groups of customers who have average spending scores. That's why they're targets. They can have some profit for the mall and different marketing strategies can be done on them based on their ages.

The Priority cluster is the most profitable group for the mall. It shows that we have a group of young people between ages 20-40 who spend a great deal in the market. So, they require very different strategies compared with other clusters.

Finally, let's visualize the whole dataset by using 5-means clustering on all three attributes of age, income, and spending score. The plot it gives will be a summary of all the discussions we made above: