Customer segmentation

A K means clustering project

Mijail Dragowski

2023

Introduccion

The objective of this project is to carry out a segmentation of clients in different clusters in order to identify different consumer groups within the total group of clients. To carry out this objective, the most popular machine learning model to identify clusters, k means clustering, will be developed.

In **section 1** (Data) we will explain the data used in this project, in **section 2** (methodology) we will give a brief theoretical explanation of how the k means clustering model works, in **section 3** (results) we will show the main results of the model.

This project was programmed and developed in the python language, but for simplicity purposes, this pdf was written and its purpose is to give a summary illustration of what the project is. Those people who are interested in the technical details of programming and coding of data processing, data analysis and machine learning, You can see it in the notebook with the code saved in this same repository

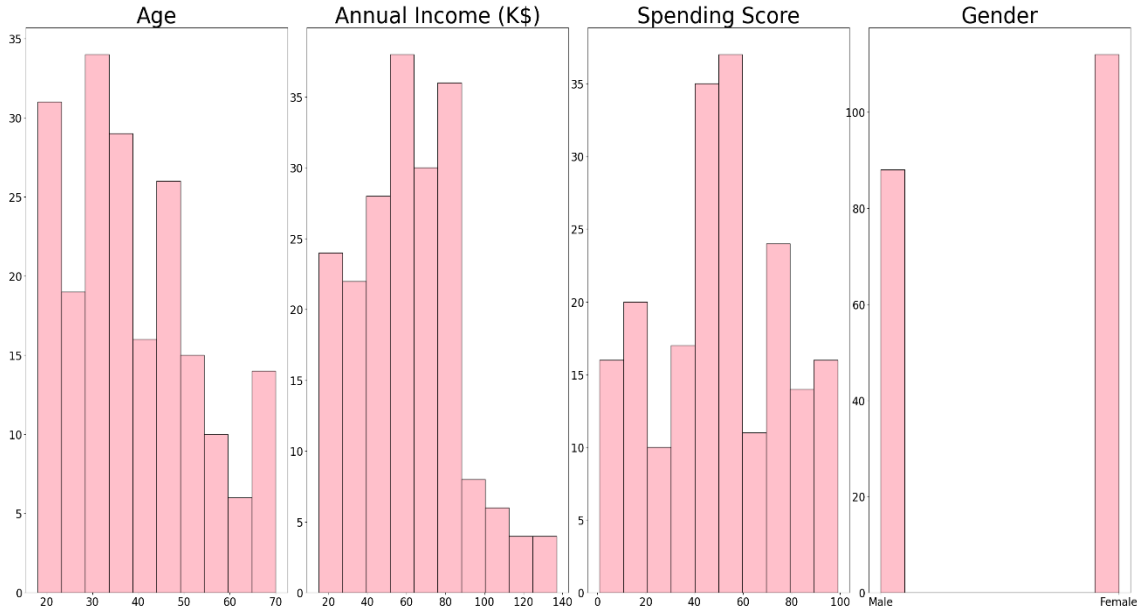
The data set observations represent each client, and the variables that we will use to segment them are their characteristics. The variables are:

Gender: Customer gender

Age: Age of the costumer

Annual income: Annual income of the costumer (in k$)

Spending score (1-100): Score assigned by the mall based on customer behavior and spending nature
We plot histograms of the variables to visualize their respective distributions**:**



2. Metodology

In this section, we will give a brief summary of the k means clustering model in order for the reader to obtain a more intuitive understanding of the procedure performed to segment customers into clusters.

k-means clustering is a unsupervised machine learning model that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.k means clustering minimizes the distance of data within the cluster and maximizes the distance between clusters. The stages of the model are as follows:

1. The optimal number of clusters is chosen.

The optimal number of clusters (k) for this model is popularly defined by the elbow method. The optimal cluster will be the one that minimizes the sum square error (SSE) of the model. The higher k is, the lower the sse will be, with which the optimal k will be the one that most minimizes the sse in marginal terms.

2. The centroids of each cluster are assigned.

The centroids of each cluster are assigned randomly, so a high error is likely to be obtained at the beginning. The objective of the model is to minimize this error, which is the distance of each point with respect to its centroid.

3. The error is calculated.

Once a cluster has been assigned to each observation, its distance is calculated with respect to its respective centroids to calculate the total error of the model. To calculate the distance we use the Euclidean distances.

4. Compute the new centroids

In this stage, the centroids of each cluster are updated to reduce the error obtained in the previous stage.

5. Repeat the process until there are no more changes.

It is an iterative process, in each iteration the centers are moved, calculating the distances between the new centers and the observations, assigning a cetroid to each observation again. This is done until the total distance of the observations with their respective centroids can no longer be minimized.

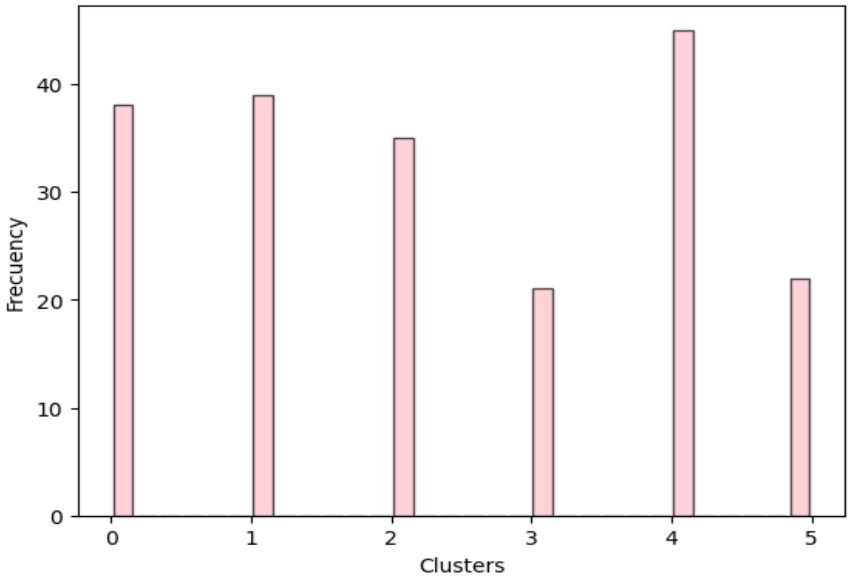
3. Results

Once the model is developed (for more details see the code) we can see its main results. The model divided the customer database into 6 different clusters.

Now we can see for each observation to which cluster it belongs, for example the first 5 observations:

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

We see how many clients there are for each cluster:

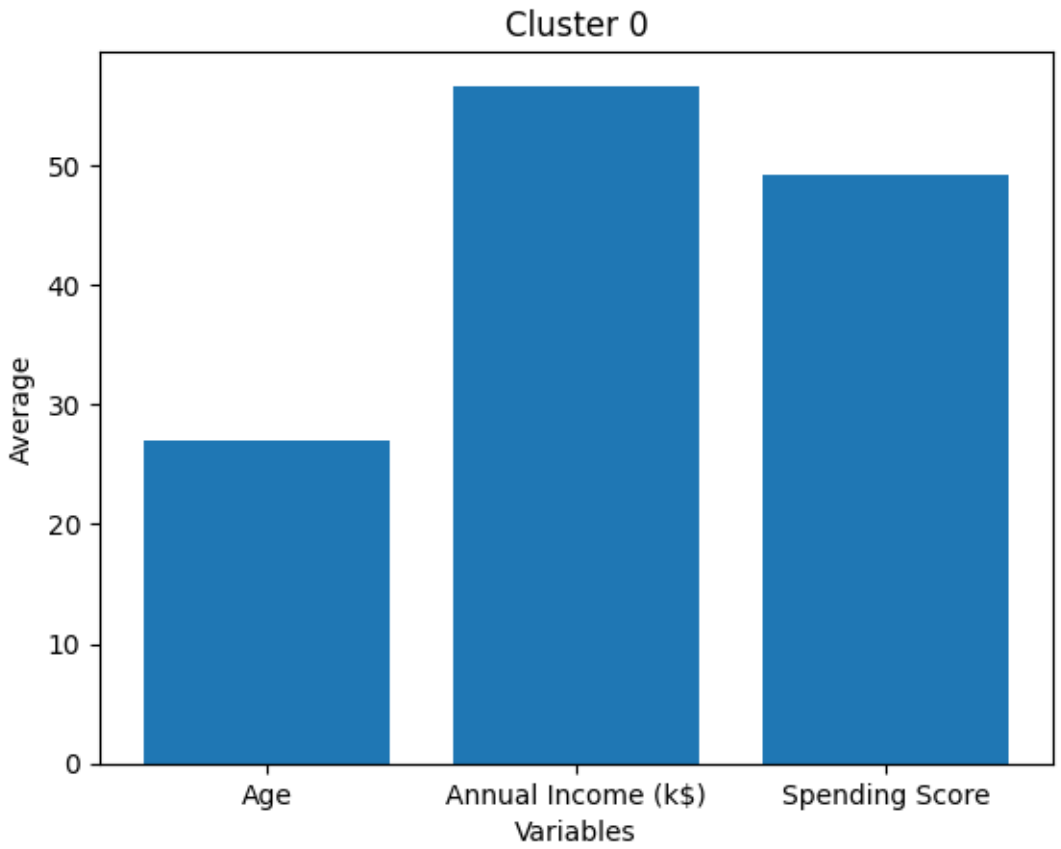


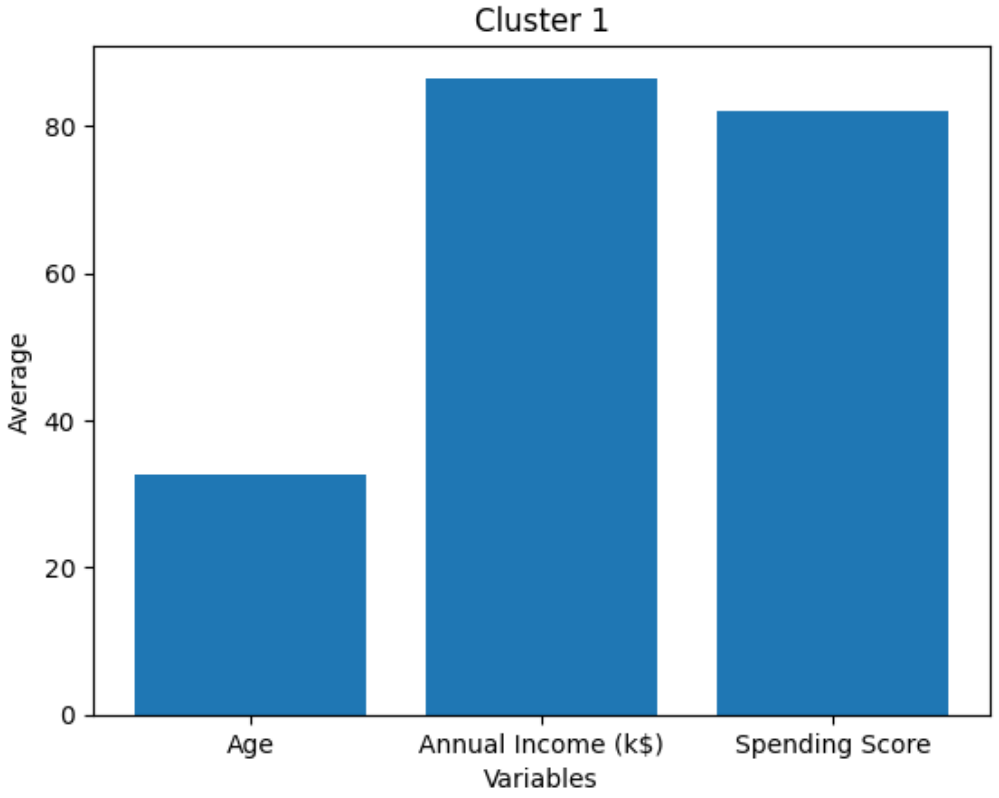
For each cluster we can see its average characteristics.

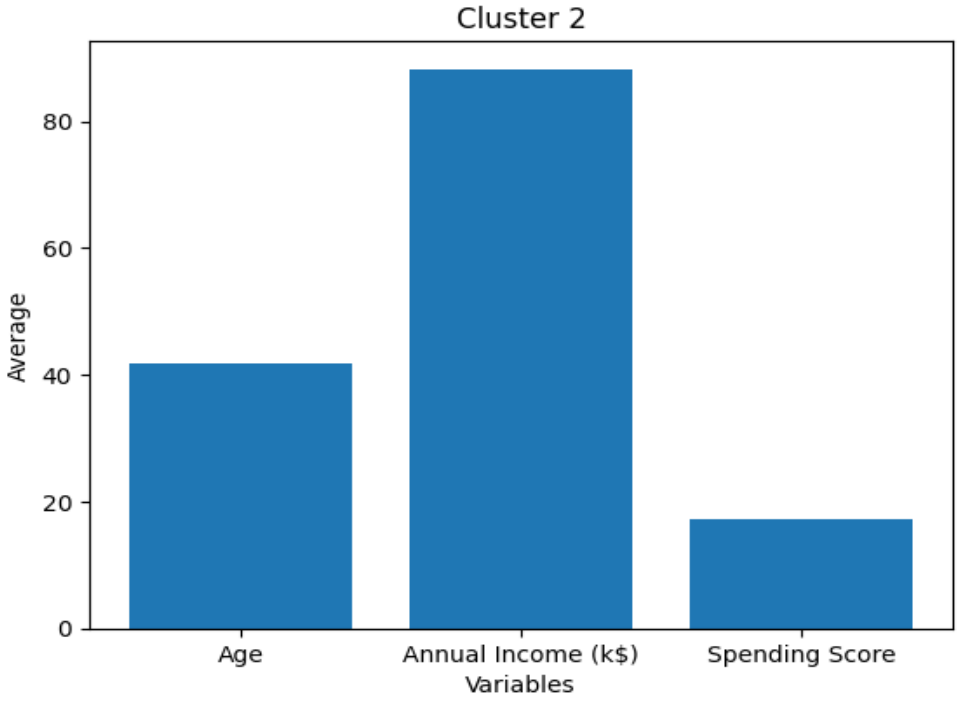
Note that the gender variable takes 2 values, 1 when the customer is a man and 0 when the customer is a woman. With which, if its average is below 0.5, it means that for that cluster there are more women than men.

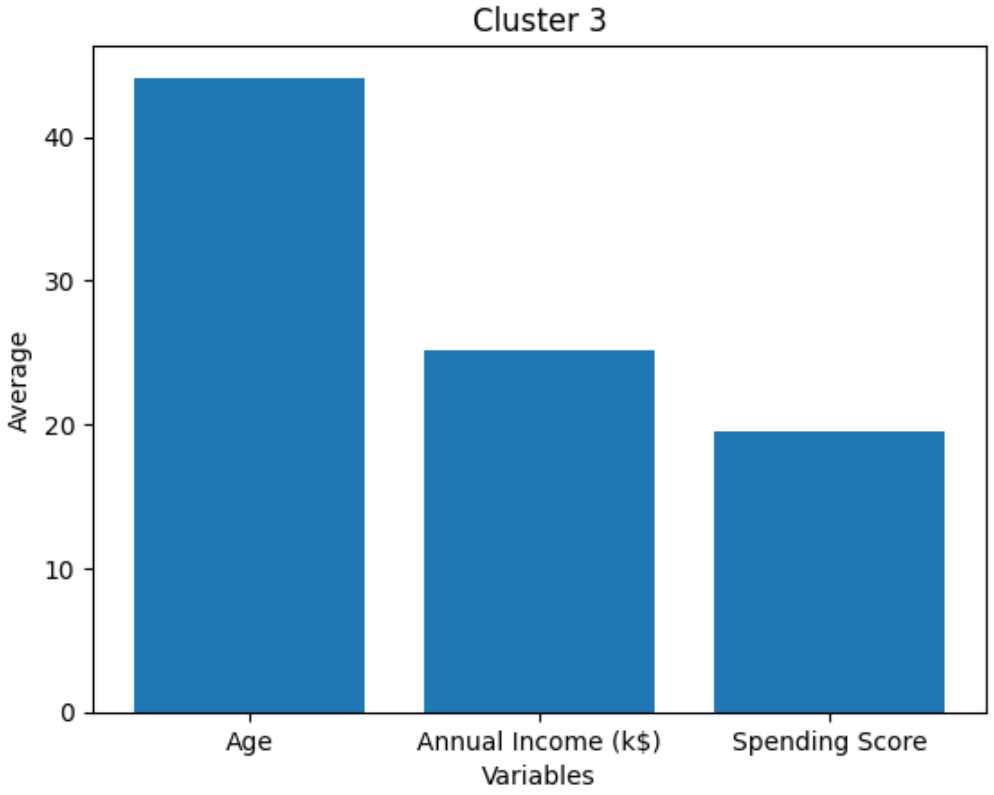
Clusters	Age	Annual Income (k$)	Spending Score	Gender
0	27.000000	56.657895	49.131579	0.342105
1	32.692308	86.538462	82.128205	0.461538
2	41.685714	88.228571	17.285714	0.571429
3	44.142857	25.142857	19.523810	0.380952
4	56.155556	53.377778	49.088889	0.444444
5	25.272727	25.727273	79.363636	0.409091

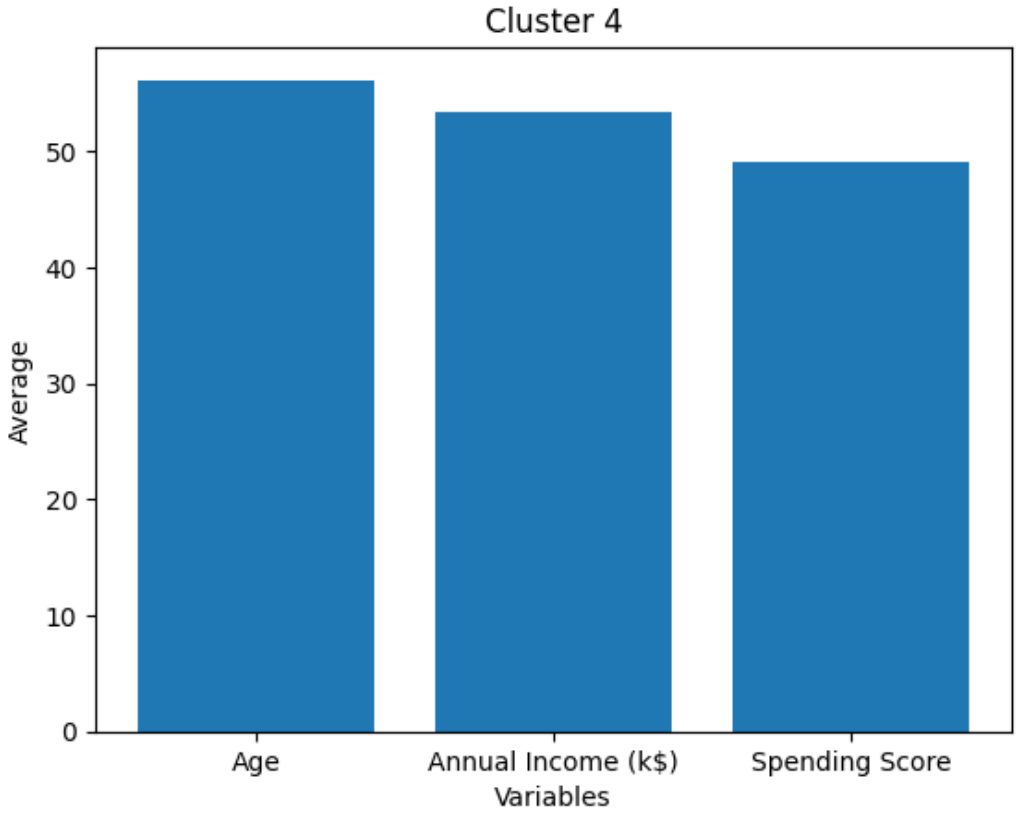
We can plot the average values of each cluster:

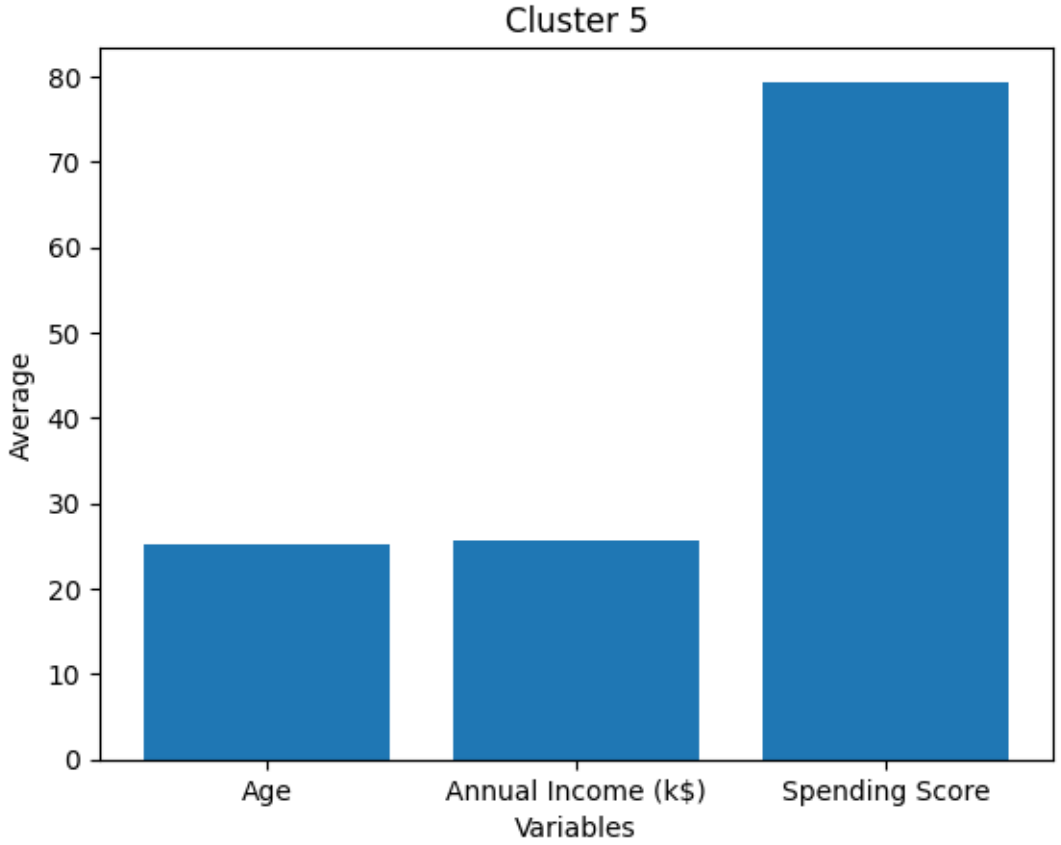

Cluster 0
Average
0
10
20
30
40
50
Age
Annual Income (k$)
Spending Score
Variables


Cluster 1
Average
0
20
40
60
80
Age
Annual Income (k$)
Spending Score
Variables


Cluster 2
Average
0
20
40
60
80
Age
Annual Income (k$)
Spending Score
Variables


Cluster 3
Average
0
10
20
30
40
Age
Annual Income (k$)
Spending Score
Variables


Cluster 4
Average
0
10
20
30
40
50
Age
Annual Income (k$)
Spending Score
Variables


Cluster 5
Average
0
10
20
30
40
50
60
70
80
Age
Annual Income (k$)
Spending Score
Variables