Customers Segmentation

Agenda

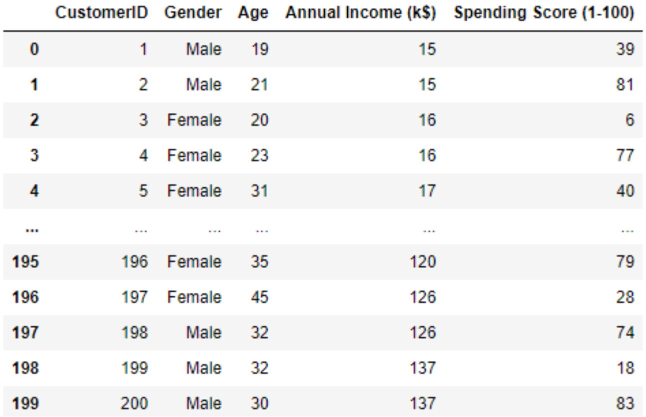
1. Introduction
2. Objectives
3. Model
4. Data preparation
5. Number of clusters and ‘silhouette score’
6. Results and discussion
7. Recommendations
8. Conclusions

1. Introduction

Customer segmentation is important for companies to get to know their customers and be able to make assertive and personalized strategies for each group. In addition, this process makes it easy to adapt and customize marketing, sales, and after-sales efforts for the needs of these specific groups. A good strategy is to separate these groups into clusters to identify the traits of each group.

For this task, let's assemble an unsupervised machine learning model called KMeans and separate the groups into clusters. In this study we will use: a little data preprocessing, KMeans algorithm, metrics to evaluate the model made, visualization of clusters made, interpretation of results and recommendations for each cluster.

Table 1 - Data



	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
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

Source: The author

This dataset contains data on age, annual gain, gender and spending score, which ranges from 0-100, and the closer to 100, the more the person spends, of customers of a shopping mall. We can see the data in table 1.

2. Objectives

The objectives of this project are: to set up a machine learning model to segment the mall's customers and propose recommendations to increase sales.

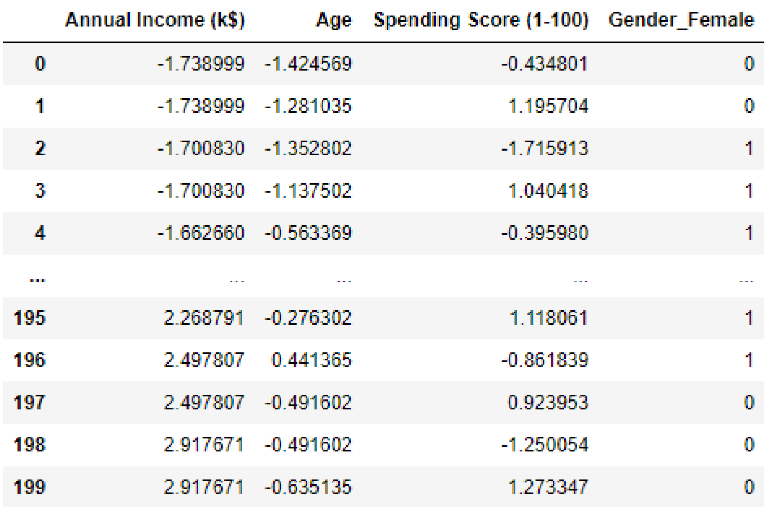
3. Model

The model that was built for this analysis was KMeans, an algorithm capable of identifying data clusters. It was made in the Python language.

4. Data preparation

The first step is to prepare the data, then the 'CustomerID' column is removed as it is not required. Then, the data was normalized, with the exception of the 'Gender' column, which will be dealt with later. This is done to improve the result of the model and get a more accurate result. Because the data in the 'Gender' column is categorical, a transformation of them to one hot encoding (0 or 1' values) was made, since the algorithm does not process categorical data. The result can be seen in table 2

Table 2 - Clean data



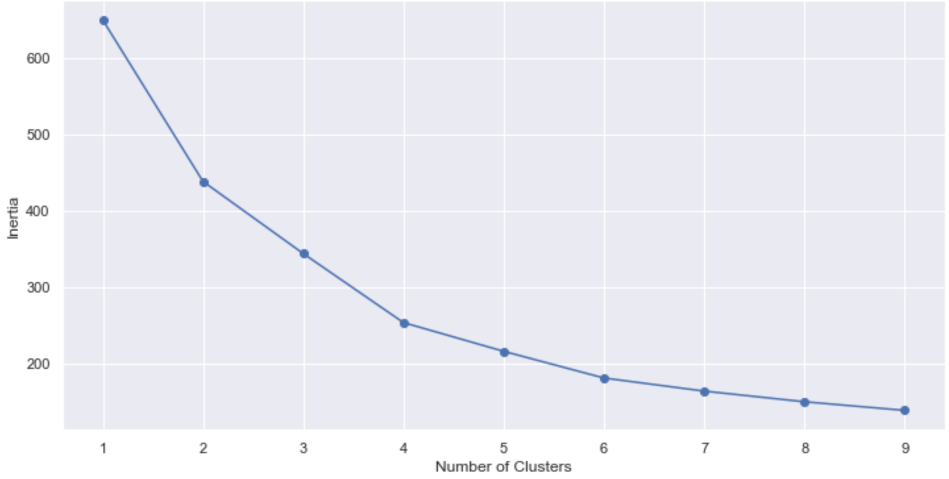
	Annual Income (k$)	Age	Spending Score (1-100)	Gender_Female
0	-1.738999	-1.424569	-0.434801	0
1	-1.738999	-1.281035	1.195704	0
2	-1.700830	-1.352802	-1.715913	1
3	-1.700830	-1.137502	1.040418	1
4	-1.662660	-0.563369	-0.395980	1
...	...	...	...	...
195	2.268791	-0.276302	1.118061	1
196	2.497807	0.441365	-0.861839	1
197	2.497807	-0.491602	0.923953	0
198	2.917671	-0.491602	-1.250054	0
199	2.917671	-0.635135	1.273347	0

Source: The Author

5. Number of clusters and 'silhouette score'

Another crucial part to achieve the best result is to determine the optimal number of clusters. So, the elbow method was used, in which, through a graphical representation, it was determined the ideal number of clusters was four, as seen in figure 1

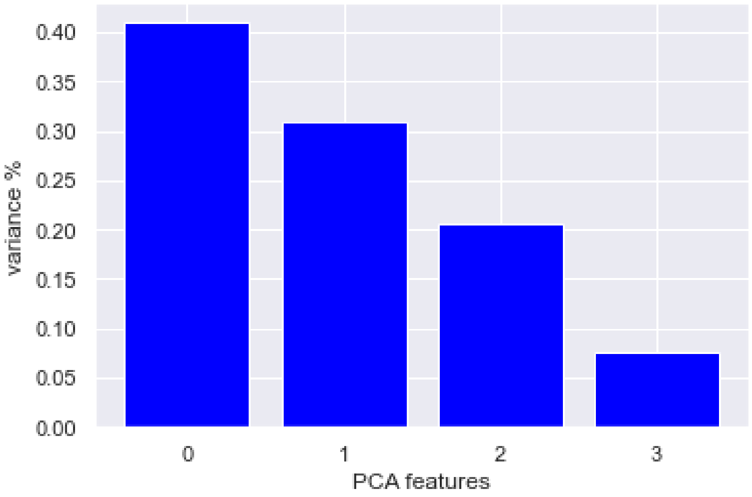
Figure 1 – Elbow technique



Source: The author

To evaluate the quality of clusters made by algorithm, the silhouette score was used. It ranges the distance between all points within the same cluster, and the smaller the distance, the better the score. It ranges from -1 to 1 and the closer to 1, the better the model. For these settings, we have that the 'silhouette score' was 0.35.

Figure 2 – PCA chart



Source: The author

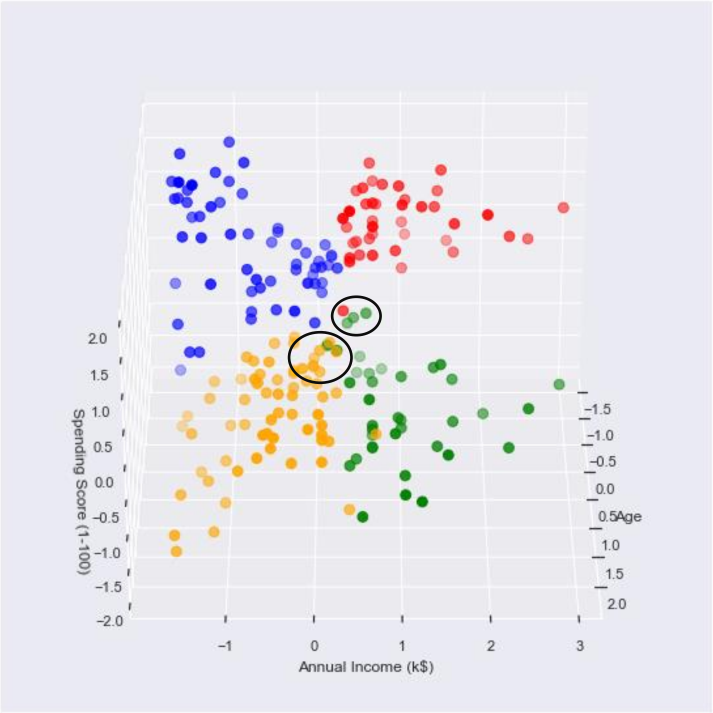
In an attempt to improve this score, we will use a technique called ‘Principal Component Analysis (PCA)'. It is a feature selection technique, which will help us reduce the dimension of the dataset. The chart generated is seen in Figure 2. The chart shows each component of the PCA and its variation. With this, we have that the first two components of the PCA represent basically 70 % of the variation, so they are the main components and were used to make a new model.

Using the result given by the PCA, we have that the number of clusters remains, four, but the silhouette score increased to 0.42, which is means an improvement in the model.

6. Results and discussion

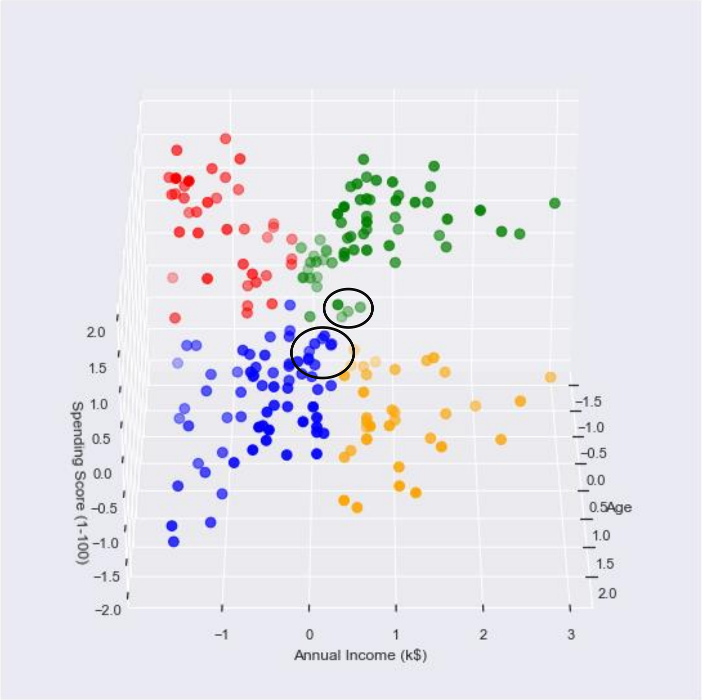
The first result we have is the graphical representation of the clusters.

Figure 3 – Clusters



Source: The author

Figure 4 – Clusters using PCA components



Source: The author

In Figure 3, we can see that there is a misclassification of points within the clusters, circled in black. Therefore, this is not the ideal model. In Figure 4, the model that presented the highest silhouette score and made with the main PCA components, we have an improvement in the classification of points in clusters. There is no longer a misclassification of points circled in black. We also have in table 3, the average of the features to help us in the analysis and in the table 4 we have the count of males and females in each cluster.

Table 3 – Average of all the features

Cluster	Age	Annual Income (k$)	Spending Score (1-100)
0	52.14	43.33	40.07
1	25.80	32.63	67.50
2	30.0	79.08	70.77
3	41.68	88.22	17.28

Source: The author

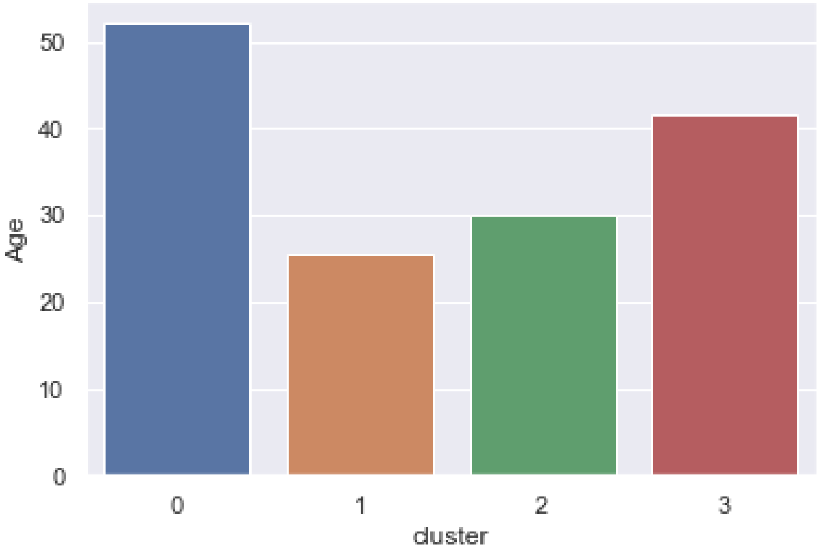
Table 4 – Count of males and females in each cluster

Cluster	Gender	Count
0	Female	34
	Male	24
1	Female	40
	Male	29
2	Female	23
	Male	15
3	Female	15
	Male	20

Source: The author

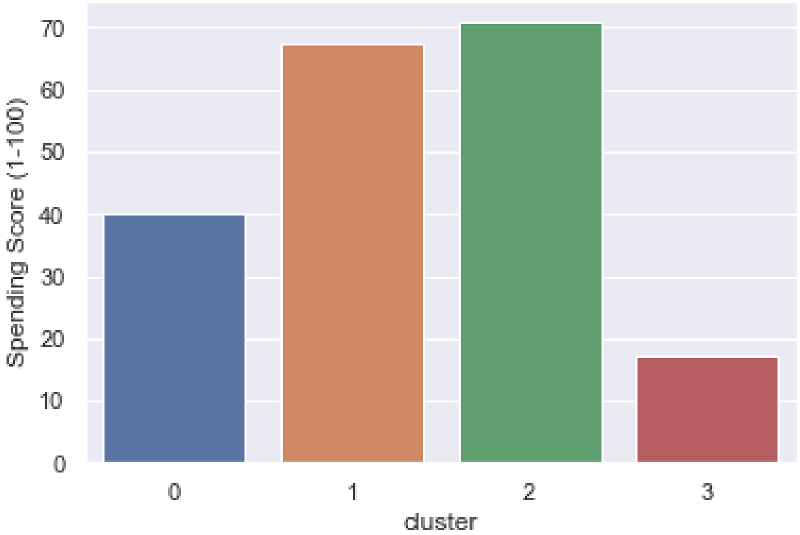
To better analyze the segmentation of customers in each cluster, the average of 'Age', 'Annual Income (k$)' and 'Spending Score (1-100) were calculated, seen in table X. With these results, we can obtain a visual representation for better analysis, seen in the figures 5, 6 and 7

Figure 5 – cluster x age



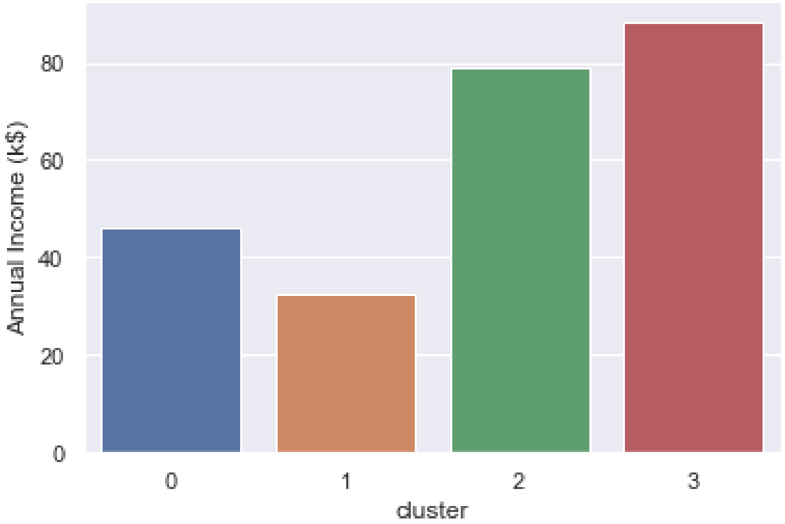
Source: The author

Figure 6 –cluster x spending score



Source: The author

Figure 7 – cluster x annual income



Source: The author

In table 4 we see that there are more women than men in the dataset, so this has to be considered when making recommendations.

We can do the following analysis of the clusters:

- Cluster 0: We have the oldest group of the 4 who earn just over $40,000 a year and don't spend as much.
- Cluster 1: We have the youngest group that spends a lot, but have the lowest annual income of all.
- Cluster 2: We have the group with the average age of 30 years, who earn well and who spend a lot.
- Cluster 3: We have the group of people just over 40 years old, who have the highest income of all and who are not spenders.

7. Recommendations

With the results of the clusters, we can make the following recommendations to try to increase the number of sales in this context.

For cluster 0 the social network Meta (Facebook) would be more indica to perform marketing, since currently this network is more used by older people. Health-related products may be a good idea.

In cluster 1 we have a younger group that spends a lot, but earns little, so discount coupons in the areas of clothing and beauty would be ideal for this segment of customers.

Cluster 2 is a good people who earn money and spend a lot, so customer loyalty programs would encourage them to keep buying products.

Finally, in cluster 3, they are the richest group, but they are careful about how much they spend, so discount codes, coupons are the most recommended for that group.

8. Conclusions

The KMeans clustering model was successfully made for customer segmentation. An analysis of the results generated by the model helped make recommendations to increase sales