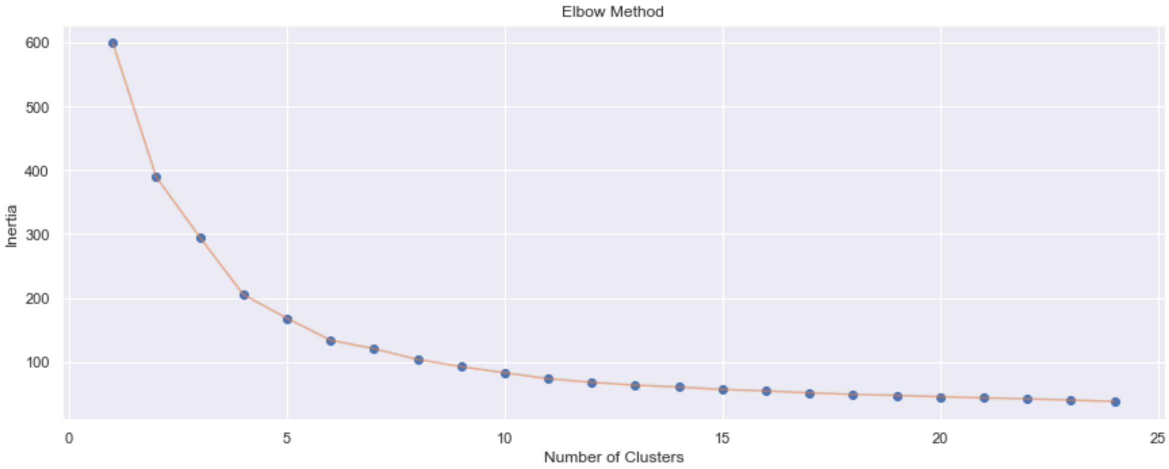
Costumer Segmentation Report

In this document will be presented the conclusions of a project where we wanted to extract the principal insights and the number of clusters (Types of customers) using the K-Means Algorithm and the Affinity Propagation Algorithm.

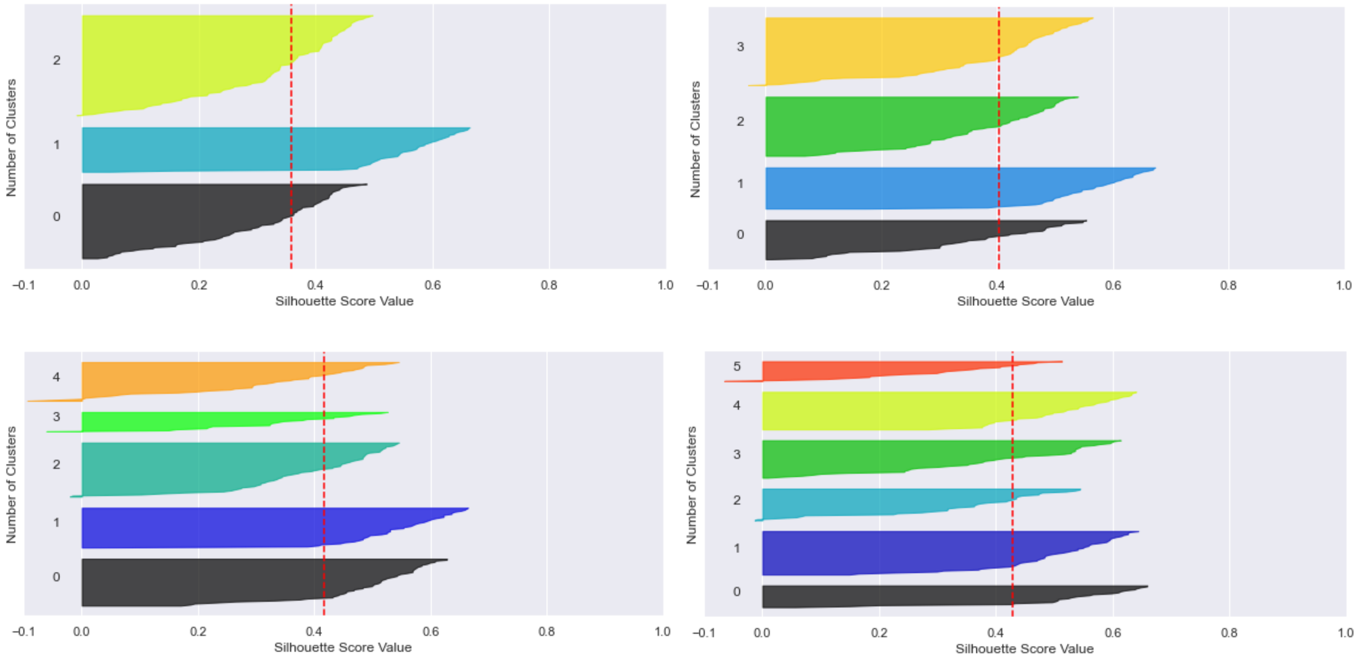
After study the quality and pre-processing the data, we started to apply machine learning to find the optimum number of clusters.

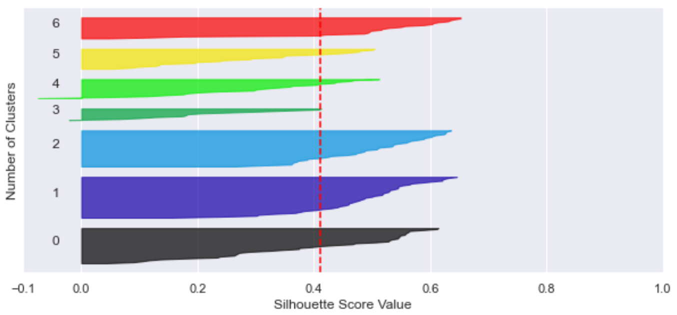
1°) The number is approached by the K-Means elbow method.



The elbow method consists in execute the k means algorithm for a range of cluster numbers (in this case from 1 to 24) and find the number where the inertia value decreases sharply. The lowest inertia value should not be used because it may cause an overfitting of the model, in other words, if I decided to select the lowest inertia it would imply many clusters and that may lead us to train a model with noise and that will lead to a bad performance.

The Elbow Method plot indicates that the number of clusters may be 5 or 6 because there is where the inertia stops to decrease sharply.
I will check this by analyzing their silhouette scores. One way to visualize the optimal number of clusters is by plotting a silhouette score plot. If a graph has all the clusters in a similar shape, we can say that the number of clusters in this plot is the optimum





.

On this occasion we have that 5 or 6 clusters may be optimal number because of their shapes. We are going to check this by analyzing the silhouettes scores:

Number of Clusters	Average Silhouette Score
3.0	0.357793
4.0	0.403958
5.0	0.416643
6.0	0.428417
7.0	0.411155

By analyzing their average silhouette score can be obtained that the optimum cluster number is 6 because is the closest to one. The silhouette score has a range from -1 to 1 and 1 mean that the clusters have an excellent cohesion and they have a marked gap between them.

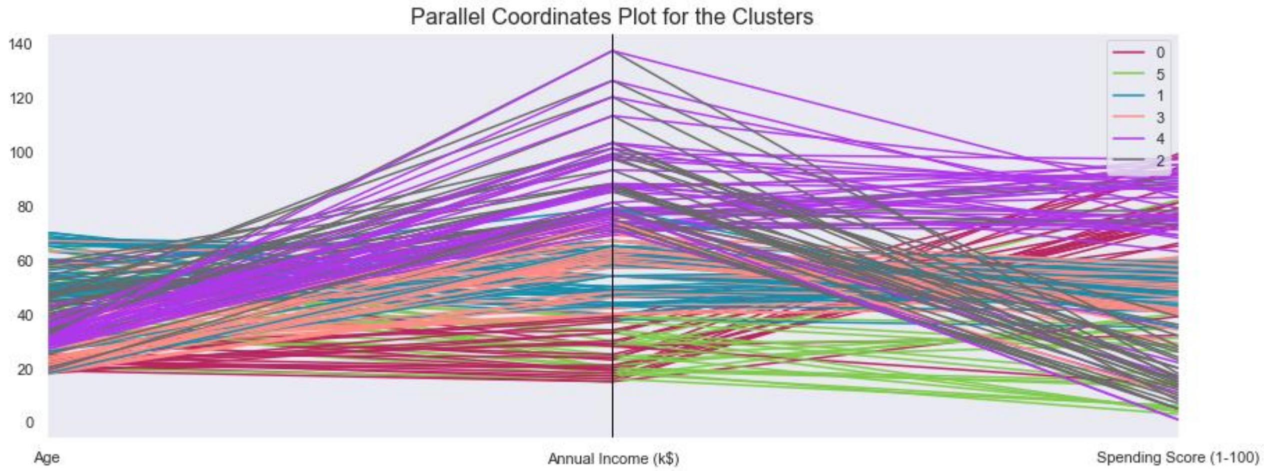
Just one algorithm isn't enough to ensure that we have 6 clusters so now we're going to use the Affinity Propagation. This algorithm gives us the number of optimal clusters by creating clusters by sending messages between pairs of samples until a coincidence appears.

Preference	Clusters	Silhouette Score
-13500.	6.0	0.451440
-6500.0	6.0	0.451440
-9000.0	6.0	0.451440
-13000.0	6.0	0.451440
-15500.0	6.0	0.450951

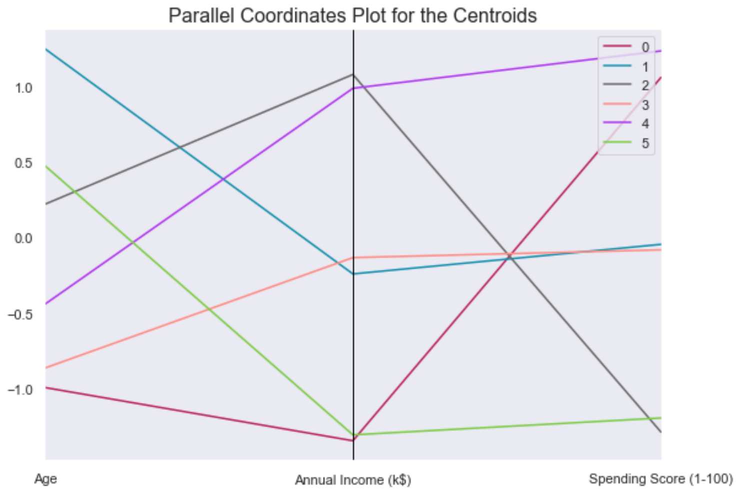
As we can see, the top five silhouette scores belongs to 6 clusters , then, we can say that we have 6 clusters

By analyzing both algorithms we can declare that this dataset has 6 clusters, i.e., it has 6 different types of customers.

Now, I'm going to train the K-Means model with 6 clusters and analyze the results by plotting each costumer and their respective segment.



As we can see this chart isn't very helpful because its full of information, so in order to visualize the customer segmentation more easily we are going to use the centroids of the clusters.



The centroids allow us to display the clusters in a very clean and helpful way, therefore, we can describe each cluster as follows:

0:

- The youngest group by age
- They have a lower annual income and the second highest spending score

1:

- The oldest group by age
- They don’t spend so much, and they have a regular spending score (approx. 50).

2:

- This group has a large variation by age.
- They have one of the highest annual incomes and one of the lowest spending scores.

3:

- The second youngest group by age.
- They have a regular annual income and spending score.

4:

- This group has people between the ages of 20 and 40 years old.
- They have one of the highest annual incomes and a higher spending score.

5:

- The second oldest group by age.
- They have the lowest income and the lowest spending score.

Assumptions

In this last step I going to share the commercial approach of the results by making assumptions. This analysis will be divided into two sections: the first is the general approach in which we make assumptions across the columns and the second is by the clusters.

1) Across the columns

1. People with high income (upper 20%) tends to spend more (high spending score).

	Annual Income (k$)	Spending Score (1-100)
Annual Income (k$)	1.000000	-0.009782
Spending Score (1-100)	-0.009782	1.000000

2. People with low income (lower 20%) tends to spend more (high spending score).

	Annual Income (k$)	Spending Score (1-100)
Annual Income (k$)	1.000000	-0.039538
Spending Score (1-100)	-0.039538	1.000000

As we can see, for both hypothesis, there is no correlation between them because it is very low. Therefore, we can't make any assumptions.

2) Clusters

As the clusters have already been presented, we will just share the commercial assumptions for each customer group. For this project we are going to simulate that a company wants to make an online marketing campaign for each type of customer.

- For the clusters with high spending score (0 and 4), regardless of their salary, we will send them much more mails than the other groups.
- For the 4th cluster we will send them more luxury goods and less sale items than the cluster 0.
- All the clusters (except 4th) will receive a lot of sale items mails.