K-MEANS CLUSTERING

You work at a local jewelry store. You've recently been promoted, and the store owner asked you to better understand your customers. Using some sneaky magic (and the help of Environics!), you've managed to collect some useful features for a subset of your customers: age, income, spending score, and savings. Use these features to segment your customers and create customer *personas*.

1. Download the customer dataset: *jewelry_customers.csv.*
2. Perform a clustering analysis of the dataset.
    a. Try different values of parameters (e.g., K for K-means).
    b. What do you think the best parameter values are? Why?
3. Describe and interpret the clusters.
4. How good are the results?

Solution

1. The jewelry dataset contains more than 2 variables therefore, I performed PCA analysis to better visualize it.

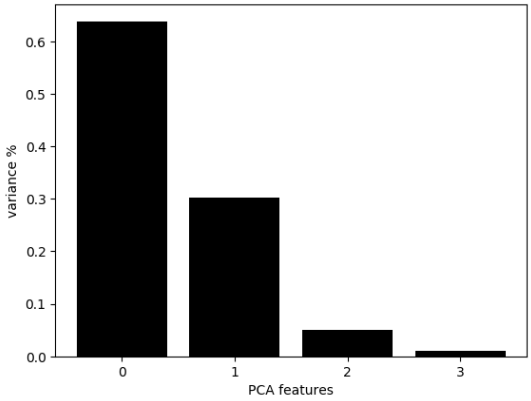

Figure 1: variance percent of components in the jewelry dataset

As seen in the figure 1, the first two components explain most of the variance in the data. First principal component contains 63.8% of the variance and the second principal component contains 30.2%, which together is 94% of the information. Because it is not 100% of the information, in this analysis, we only use PCA components for visualization purposes and not to find the actual clusters.

2. Using the top two PCA components, I performed cluster analysis on the dataset. I tried different values of K (2, 3, 4, 5, 6 & 7) for K-means, as shown in figure 2. Just from visual observation, it looks like there are 5 obvious clusters in the dataset.

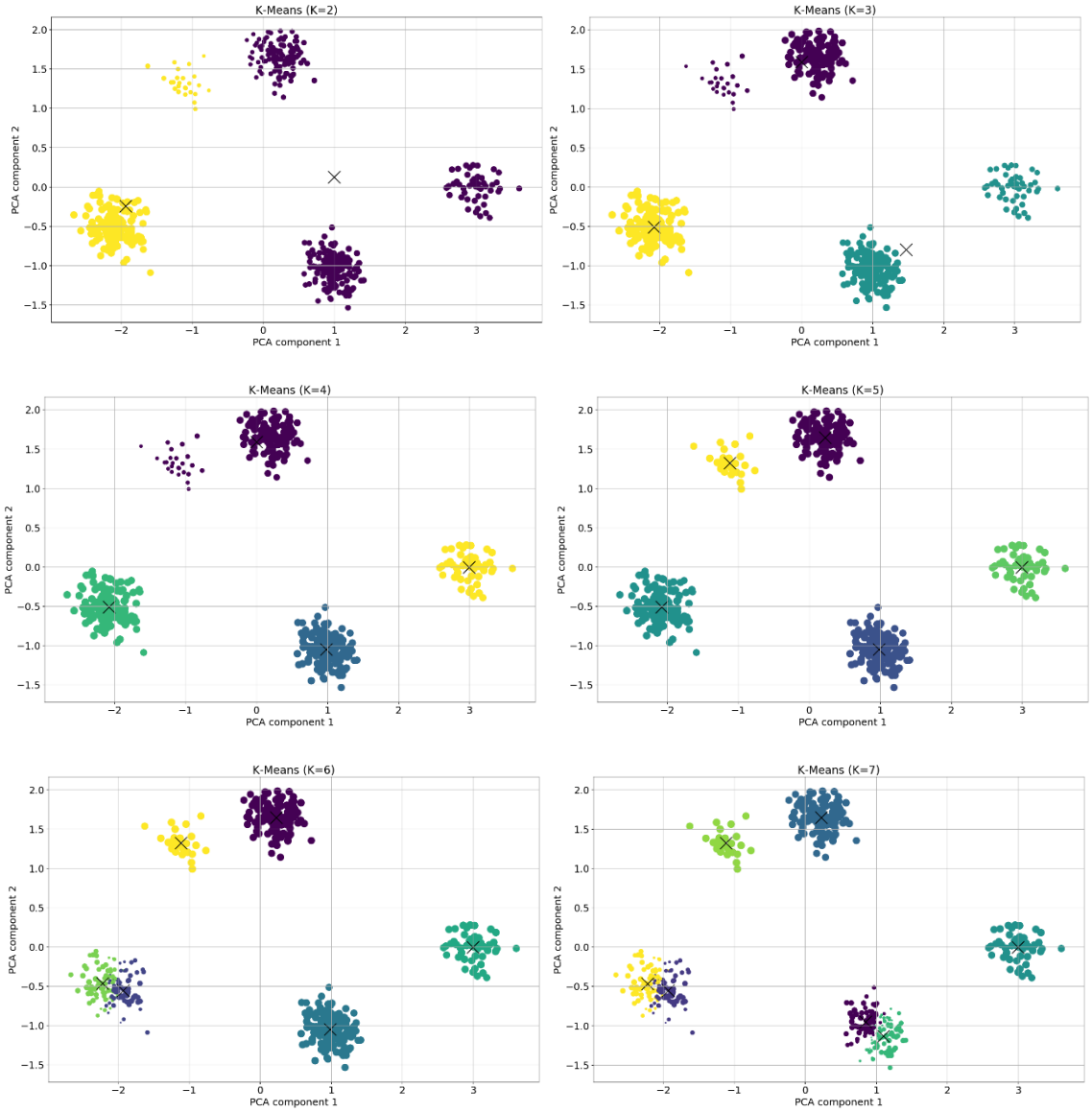

Figure 2: Experimenting with different k values for K-means clustering

3. To determine the ideal number of clusters, I used elbow method on all the features in the dataset to measure inertia and silhouette score. Below are the output figures.

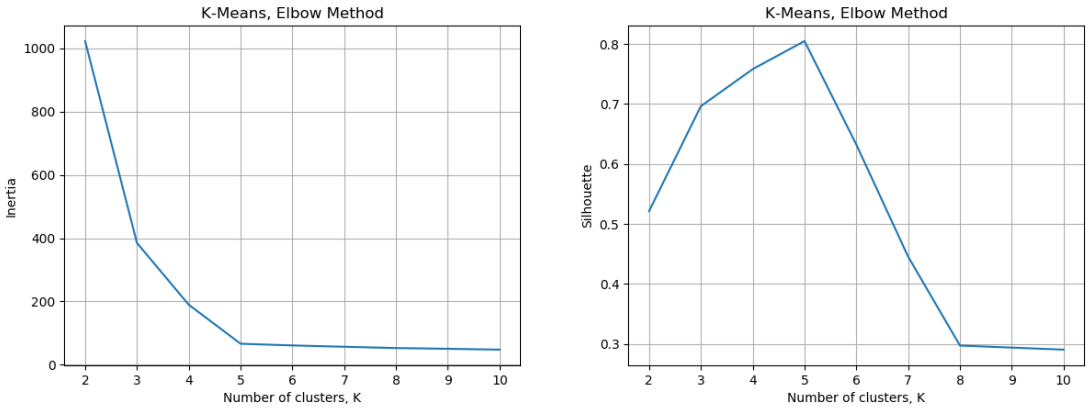

Figure 3: Elbow method to show change in inertia and silhouette for various k values

From figure 3 we can see that after 5 clusters there is no significant change in inertia. That is, the inertia is very small, and the clusters are very dense – close together. Similarly, the Silhouette score is closest to 1 when there are 5 clusters. Which indicates at 5 clusters we have minimal miss classification. Therefore, K value for K-means clustering is equal to 5.

4. Figure 4 shows the 5 clusters in the dataset. For visualization purposes, I used only the top two PCA components to plot these clusters.

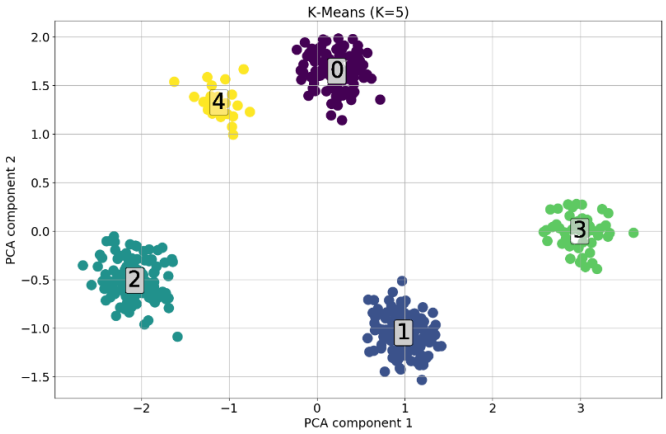

Figure 4: k = 5 clusters using top two PCA components from customer jewelry dataset

Figure 5, the silhouette plot, shows the distribution of clusters. The silhouette score 0.805 tells us that the clusters are dense, and the objects are well matched within each cluster. For this plot I used all the features because I wanted to show how well matched the objects within each cluster are and how poorly matched they are to the neighboring clusters.

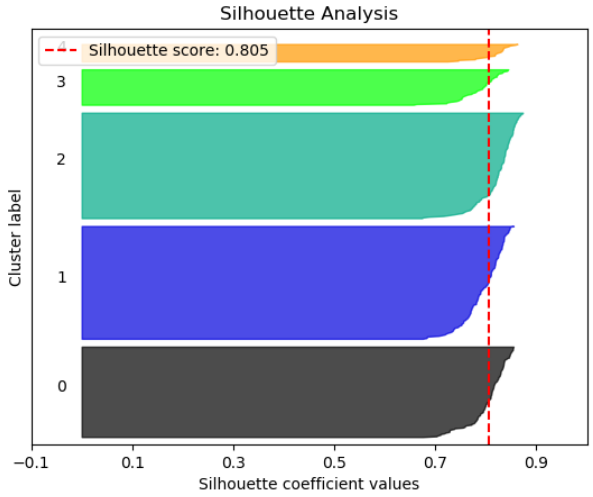

Figure 5: Silhouette plot showing the density of the 5 clusters

5. To interpret the clusters, I created a relative importance plot and compared clusters with each other.

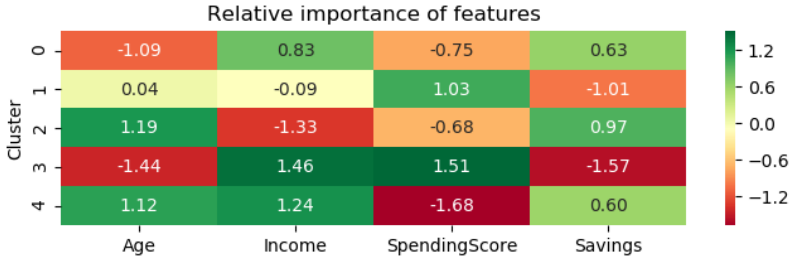

Figure 6: Relative importance plot to interpret the clusters

   a. Cluster 0: includes customers that are young with medium income, low spending score and high savings.
   b. Cluster 1: includes customers that are medium age with medium income, high spending score and low savings.
   c. Cluster 2: includes customers that are old with very low income, low spending score and high savings.
   d. Cluster 3: includes customers that are very young with very high income, very high spending score and very low savings.
   e. Cluster 4: includes customers that are old with high income, very low spending score and medium savings.

5. I believe the results from the above analysis are good because the clusters are very clean and dense as seen in figures 4 and 5. They use various features to create customer personas as explained using relative importance plot in figure 6. This model can be used to identify different customer segments and make business decisions accordingly.