**Criteria to choose optimal number of clusters (Silhouette Score)**

1. **Cohesion**: How close the objects within the same cluster are to each other.
2. **Separation**: How well the clusters are separated from each other, or how distinct a cluster is from other clusters.

The silhouette score ranges from -1 to 1:

* A score close to **+1** indicates that the objects are well clustered, meaning that the objects are close to other objects in their own cluster and far from objects in other clusters.
* A score of **0** suggests overlapping clusters, where the distance to objects in its own cluster is about the same as the distance to objects in the nearest cluster.
* A score close to **-1** indicates that the objects have been assigned to the wrong clusters.

To calculate the silhouette score for each object in the dataset, the following steps are taken:

* **a**: The average distance between the object and all other objects in the same cluster (measuring cohesion).
* **b**: The smallest average distance between the object and all objects in any other cluster, of which the object is not a member (measuring separation).

The silhouette score for each object is then calculated as:

Silhouette score=

In clustering analysis, the silhouette score is used to determine the optimal number of clusters. A higher average silhouette score across all objects in the dataset suggests a better clustering configuration.

The cluster analysis has been conducted for both the original data and the data reduced by PCA, with each dataset divided into **two clusters**. Here are the descriptions of the characteristics of each cluster:

**Clusters Based on Original Data**

**Cluster 0:**

* Higher average values in **disc\_wage**, **disc\_jobedu**, **disc\_promotion**, and **disc\_resign**, suggesting higher perceived discrimination in these areas.
* Older age group with lower education (**edu\_cat**), and moderate income and health status.

**Cluster 1:**

* Lower average values in discrimination-related variables, suggesting less perceived discrimination.
* Younger age group with higher education and better income and health status.

**Clusters Based on PCA-Reduced Data (80% Variance)**

**Cluster 0 (PCA):**

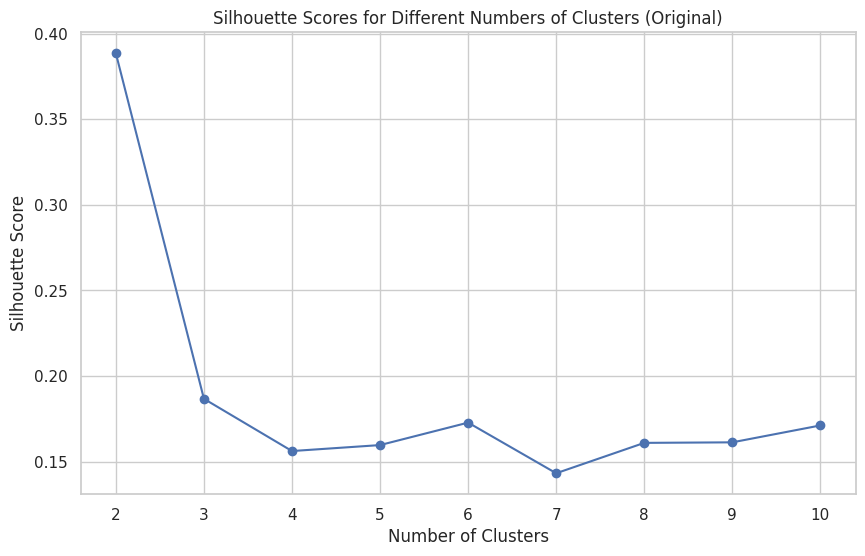
* Lower values on the principal components, indicating lower levels of the features captured by these components, such as perceived discrimination and socio-economic factors.

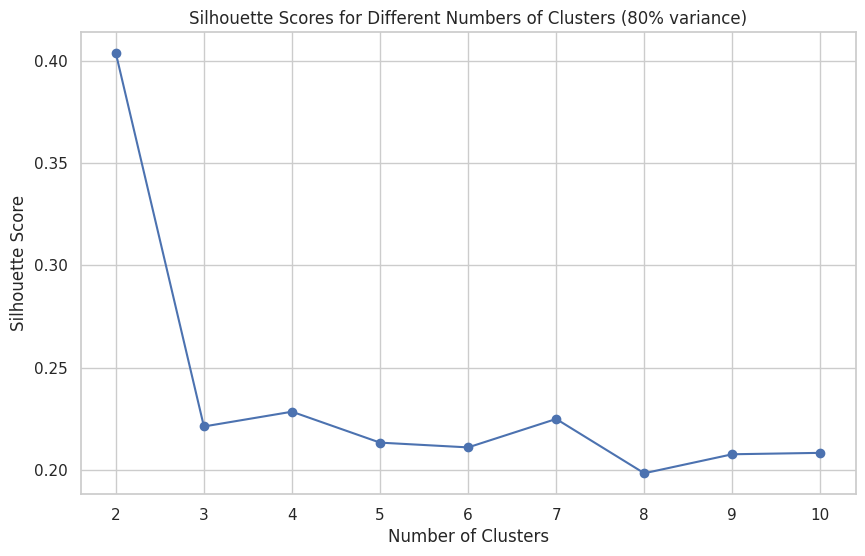
**Cluster 1 (PCA):**

* Higher values on the principal components, indicating a higher presence of the features captured by these components.

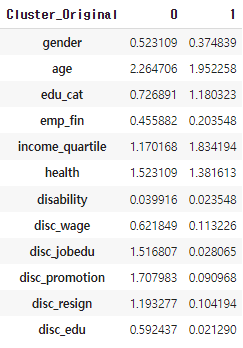
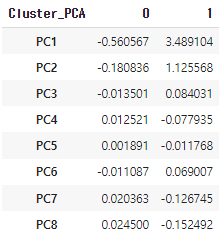
**Comparison of Clustering Results**

* The clustering based on the original data provides a detailed view of each group’s characteristics, focusing on specific variables like age, education, and discrimination experiences.
* Clustering based on PCA-reduced data offers a more abstract view, focusing on underlying patterns represented by the principal components.
* Both approaches identify distinct groups within the data. However, the PCA-based clustering is less direct in terms of original variables but captures broader patterns in the data, while the original data-based clustering provides more direct interpretability.
* In both cases, the clusters seem to differentiate between individuals with higher and lower levels of perceived discrimination and socio-economic status.





[1. Criteria to choose optimal number of clusters]

[2. Result Comparison]