Clustering Analysis and Evaluation

Your Name

May 21, 2024

1 Introduction

This document presents the clustering analysis performed on the dataset. The analysis includes using the elbow method to determine the optimal number of clusters, finding the optimal epsilon for DBSCAN, and evaluating clustering performance using various metrics.

2 KMeans Clustering

2.1 Age Clustering

We created a subset containing columns just for age

2.1.1 Elbow Method for Optimal Number of Clusters

The elbow method was used to determine the optimal number of clusters. The plot below shows the within-cluster sum of squares (WCSS) as a function of the number of clusters. The result is in Figure 1.

From here, we go with 4 clusters.

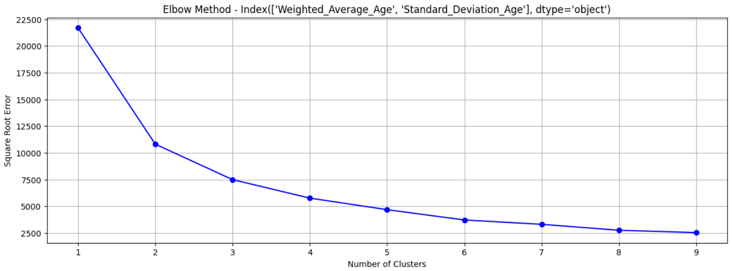

Figure 1: Elbow Method for Determining Optimal Number of Clusters

2.1.2 Clustering Evaluation Score

```
  Silhouette Score: 0.34670912923552283,
Davies-Bouldin Index: 0.8967056702996291,
Calinski-Harabasz Index: 870.8409692455747
```

2.2 Social Clustering

We created a subset containing columns for the social dimension

2.2.1 Elbow Method for Optimal Number of Clusters

The elbow method was used to determine the optimal number of clusters. The plot below shows the within-cluster sum of squares (WCSS) as a function of the number of clusters. The result is in Figure 2.

From here, we go with 4 clusters.

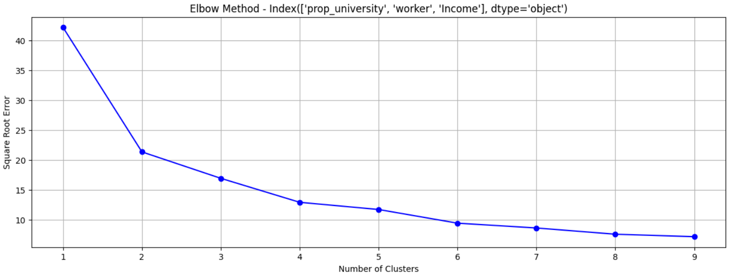

Figure 2: Elbow Method for Determining Optimal Number of Clusters

2.2.2 Clustering Evaluation Score

```
  Silhouette Score: 0.35776906830109717,
Davies-Bouldin Index: 0.9363643893605348,
Calinski-Harabasz Index: 723.2620432968293
```

2.3 Demograhic Clustering

We created a subset containing columns for the demographic dimension

2.3.1 Elbow Method for Optimal Number of Clusters

The elbow method was used to determine the optimal number of clusters. The plot below shows the within-cluster sum of squares (WCSS) as a function of the number of clusters. The result is in Figure 3. From here, we go with 2 clusters.

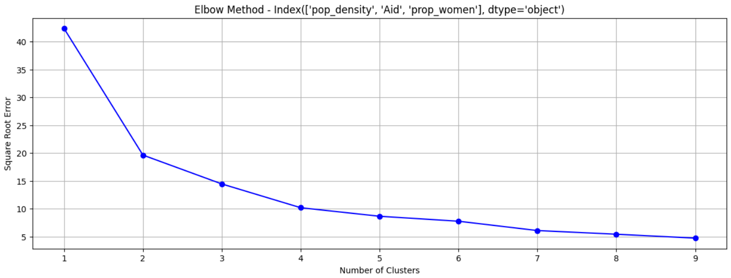

Figure 3: Elbow Method for Determining Optimal Number of Clusters

2.3.2 Clustering Evaluation Score

```
  Silhouette Score: 0.5431516201635206,
Davies-Bouldin Index: 0.704558772933317,
Calinski-Harabasz Index: 1109.8168696956805
```

3 DBSCAN Clustering

3.1 Age Clustering

3.1.1 Finding Epsilon for DBSCAN

The graph below shows the k-distance plot used to determine the optimal epsilon value for the DBSCAN algorithm. The optimal epsilon in this case is 0.6

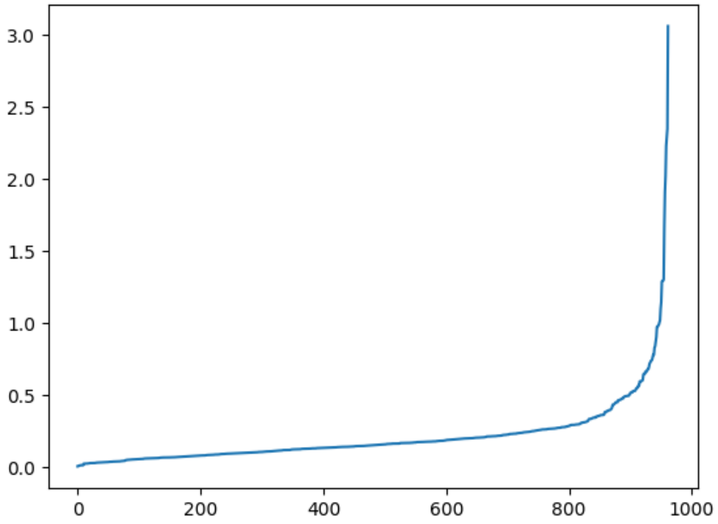

Figure 4: K-Distance Graph for Finding Optimal Epsilon

3.1.2 Clustering Evaluation Scores

```
  Silhouette Score: 0.10905903800976988,
Davies-Bouldin Index: 3.1030322814328968,
Calinski-Harabasz Index: 36.32797925043744
```

3.2 Social Clustering

3.2.1 Finding Epsilon for DBSCAN

The graph below shows the k-distance plot used to determine the optimal epsilon value for the DBSCAN algorithm. The optimal epsilon in this case is 0.1.

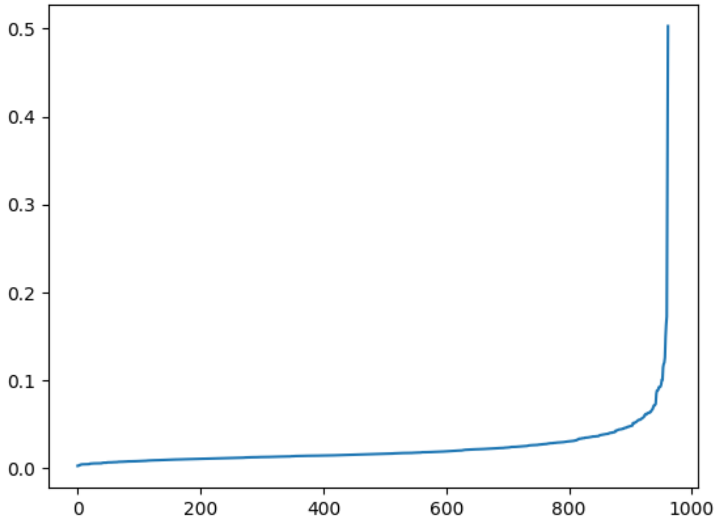

Figure 5: K-Distance Graph for Finding Optimal Epsilon

3.2.2 Clustering Evaluation Scores

```
  Silhouette Score: 0.10905903800976988,
Davies-Bouldin Index: 3.1030322814328968,
Calinski-Harabasz Index: 36.32797925043744
```

3.3 Demographic Clustering

3.3.1 Finding Epsilon for DBSCAN

The graph below shows the k-distance plot used to determine the optimal epsilon value for the DBSCAN algorithm. The optimal value in this case is 0.06.

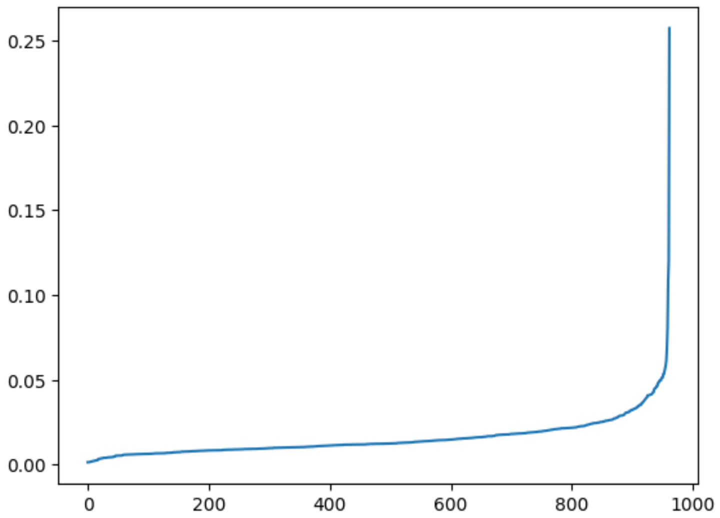

Figure 6: K-Distance Graph for Finding Optimal Epsilon

3.3.2 Clustering Evaluation Scores

```
 Silhouette Score: 0.5125078223003526,
Davies-Bouldin Index: 1.2146959623375966,
Calinski-Harabasz Index: 76.59937735480838
```