Tim Lefebvre & Mari Sisco

November 16th, 2024

Group Assignment 3: Clustering U.S. Household Income Data

The missing data in this dataset were only a little over 300 records which had no information in the important columns. We dropped these entries because the data set is large enough for it to make little difference.

We do not use the non-numerical variables for clustering or analysis because they are either not diverse enough or would not be helpful to cluster. In order to get a grouping based more on the character of each location rather than its proximity to others, we opted not to use the Zipcodes, Area-codes, or latitude and longitude together. Our clusters are made using a K-means algorithm with the variables mean, median, and standard deviation, with mean weighted heavier. Our K-Means was done on a standardized measure of those variables.

These cluster variables allow us to see differences between the clusters on the area of land and number of households. Based on the residual measurements and silhouette score for the different numbers of k-means clusters, the best number of clusters for these variables was three. The Silhouette score was 0.5167, meaning that the clusters are somewhat fitting with moderate difference between and within clusters. Additionally, when there were 3 clusters they were at the 'elbow' of the SSE and number of clusters graph. While this level is not ideal, and being able to group with a larger number of variables might get better insights, adding variables makes it more difficult to get good fitting models, as the silhouette scores were lower as more variables were added.

From the 3 clusters produced, the following summary statistics were derived:

Cluster 0: average mean income of $44968, average land area of 1.414166e+08 and an average 617 houses in that area.

Cluster 1: average mean income of $77506, average land area of 1.186226e+08 and an average 625 houses in that area.

Cluster 2: average mean income of $126148, average land area of 1.959726e+07 and an average 322 houses in that area.

This means that Cluster 0 and 1 contain a larger number of houses with a smaller land area and a smaller average income compared to Cluster 2. Cluster 2 describes a set of locations that contain a larger land space, with smaller houses and a larger income. Ultimately, richer communities live in spacious areas with a lower number of surrounding houses.