ECON 860 Final

Ross Bowen

Step 1

**final\_factor\_analysis.py**

Using the provided dataset, I perform factor analysis using factor\_analyzer.

Graph “plot\_factor\_analysis.png” leads to the conclusion that there are three groups.

I export “result.csv” for use in various clustering exercises that follow.

Step 2 – Kmeans

**final\_run\_kmc.py**

K-means is used to analyze “result.csv”.

The resulting silhouette scores are shown in graph “silhouette\_kmeans.png”.

K-means results suggest that there are 2 groupings. The silhouette score is near .80 for two groups, then drops quickly to near .50 for groups 3,4 and so on.

Step 3 – Kmedoids

**final\_kmedoid.py**

K-medoid analysis produces results similar to k-means.

Graphs “silhouette\_kmedoid.png” and “kmedoid\_ssd.png” both suggest that there are two groups.

Step 4 – Gaussian Mixture Model

**final\_run\_gmm.py**

GMM model returns very similar silhouette scores. However, examining the scatter plots shows a plausible grouping of three. The larger group of data is broken into two groups, with seemingly coming under the remaining data.

Step 5 – Agglomerative Hierarchical Clustering

**final\_ahc.py**

AHC results in a interesting dendrogram. Most of the benefit comes from the first division into two groups. The large group is divided into two groups near the next step down. This is consistent with the idea from the GMM model.

The model was unable to run until I added these two lines

import sys

sys.setrecursionlimit(5000)

The model wasn’t able to converge at 1,000 or 2,000.

**Optimal Number of Clusters**

For all models, silhouette scores strongly suggest two groups. A closer examination of GMM makes it seem there could be three groups. In the GMM model, the two group division is different than the other clustering algorithms. It shows the points “to the right and above” are in one group and the points “to the left and bottom” are a different group. This implies that the bunch on the bottom left is really a different group.

All the numerical evaluation suggests two groups. I prefer the three groups as determined by GMM.

**final\_run\_kmc\_usonly.py - Country**

I first approached the country problem by sorting by country. The US was by far the biggest, followed by Australia, Great Britain, and Canada. I created a subset containing only the US records, about half of the total. If the result of the k-means clustering analysis on US only is materially different from the clustering results of all other countries records, then country warrants further investigation.

The silhouette score was lower than the previous analysis (.61 vs .80). The scatterplot looks completely different. It appears that the smaller blob belonged to non-US records, and it is gone almost completely. The remaining graph looks like the elongated “cigar” shape from the original scatterplot.

I conclude that we need to do further analysis on each country, and use those results by country if we implement the recommendations in real life. The US looks quite different than all other countries combined.