Luke Watson

Exercise 8 Write up

11-4-24

**Exercise 8**

*Ward’s Hierarchical Cluster Analysis*

Using the occupation dataset provided, I conducted two cluster analyses. The first of which was a hierarchical cluster analysis. Using the hclust() command from the stat package in R, I ran a hierarchical cluster analysis using a Ward’s D method on a scaled version of the occupation data. I decided to cut the resulting dendrogram into three clusters. I leveraged the consensus of the NbClust() command from the NbClust package. This command consults several cluster indices and summarizes what they deem to be the optimal number of clusters for my analysis. Commonly, researchers will choose whichever number of clusters is preferred by most of the indices, and as Figure 1 shows, there was a tie between 3 and 15. Specifically, the CCC index returned 15 as the optimal number of clusters, but it was only one of six opting for that many clusters. I believe there is more evidence supporting parsimonious solutions for this cluster analysis. The second most frequent optimal number of clusters was 4, so I opted to go with 3 clusters.

With my optimal number of clusters, I cut the dendrogram and split the occupations along their hierarchical categories. The result is depicted in Figure 2, where I plotted a radial dendrogram of the occupations and colored them by their cluster assignment.

**Figure 1**

A graph of a number of clusters

Description automatically generated

*Figure 1* – Frequency plot generated by the fviz\_nbclust() command showing how many indices indicated that number of clusters (k) as the most optimal partitioning of the data.

**Figure 2**

A circular pattern of text and numbers

Description automatically generated with medium confidence

*Figure 2* – Dendrogram of the results of the hierarchical cluster analysis. Each color represents a unique cluster.

*K-means cluster analysis*

I also ran a k-means cluster analysis on the same dataset to compare the occupations. Like the hierarchical analysis, I consulted NbClust() to see what the majority thought about the optimal number of clusters. Again, the optimal number of clusters determined by the majority was 3, and this time it was a much easier decision (Figure 3).

**Figure 3**

A graph of a number of clusters

Description automatically generated

*Figure 3* – Majority of the indices support k = 3 for the k-means analysis.

With my optimal number of clusters, I conducted the analysis using the kmeans() command from the stats library in R. The results are plotted in Figure 4 with the fviz\_cluster() command which uses a principal components analysis to reduce the data to two dimensions for visual analysis. Looking at the results of this cluster analysis, I’ve decided on classifications for each of the clusters.

Cluster 1 (green ellipse in Figure 4) is defined by above average IQ as well as above average complexity in the data and people dimensions while the thing complexity dimension was below average. I classified this cluster as the “generalists.” These types of occupations likely require knowledge of a lot of different modalities and fields to be integrated into one to support the overall functioning of a particular system with a lot of moving parts. Some exemplars like librarians, teachers’ aides, and executives have to keep track of all these moving parts, whether that is employees, students, or books. They leverage their intelligence to track lots of data and/or people on a daily basis.

Cluster 2 (blue ellipse in Figure 4) is defined by below average IQ as well as below average complexity on data and people but average thing complexity. I classified this cluster as the “connectors.” These occupations seem to involve very formulaic, but crucial tasks that help run the logistical systems we all rely on. Some exemplars include drivers, clerks, utilities, and dockworkers.

Cluster 3 (red ellipse in Figure 4) is defined by average IQ, data complexity, people complexity, but high thing complexity. I see these occupations as “specialists” that require the worker to become skilled and very knowledgeable about a narrow range of things to help troubleshoot issues when our complex machinery or systems fail to work as intended. Some exemplars include computer programmers, electricians, news writers, and nurses.

**Figure 4**

A screen shot of a computer

Description automatically generated

*Figure 4* – Results for the k-means cluster analysis for the occupation dataset plotted along dimensionally reduced data. The ellipses were generated using a multidimensional t distribution from the stat.ellipse() command in the ggplot2 library. Colors of the points and ellipses represent cluster relationships determined by the analysis and are labelled accordingly.