Tanner Sax

Capstone Project – Clustering

The value in this analysis will come from clustering the data. Through clustering, the programming languages that are often grouped together can be observed. This will help identify which programming language combinations are the most relevant in the data.

The data frame containing the job descriptions from all cities will be used. Columns with binary variables for each programming language will be created.

The specific type used will be hierarchical clustering. The process begins by identifying each observation as a separate cluster. The distances between all data points must be computed. The dist() function can be used to compute these distances using the Euclidian method. The two clusters that are closest are then merged together. This is repeated until there is only one cluster remaining. Since the occurrence of programming languages in this analysis are represented through binary variables, the clusters will mostly identify which job descriptions have zero difference between them regarding programming languages.

A dendrogram will be created to select the number of clusters. The number of clusters can be chosen by drawing a horizontal line on the dendrogram. The amount of vertical lines that are crossed is the number of clusters that would be used. Consideration should be given to the number of clusters that give the most vertical distance between horizontal lines. However, it is also important to consider the dataset being observed, and which number of clusters would be appropriate. The number of clusters selected in hierarchical clustering is not widely known as an objective approach. To select the appropriate amount, several solutions should be observed to compare the best fit.