Within the next phase of the exploratory analysis section, the unsupervised learning method of ‘Clustering’ was chosen to determine potential groupings / subsets within the ‘Food Desert’ data set. Ideally, various states would be grouped together to determine potential similarities, meaningful structure, generative features, etc.

To begin the clustering process, the data set needed to be transformed in terms of aggregating observations by the ‘State’ variable. State data was then converted from a basic column/variable to a row name of the data frame. Non-numeric variables were then omitted from the analysis (i.e., ‘State,’ ‘CensusTract’, and ‘County’).

An additional analysis step was conducted to determine the optimal amount of clusters to use for the data. All in all, the ‘Average Silhouette Method’ and its respective R/RStudio functions determined the ‘best’ number of clusters to choose was 3 via the ‘fviz\_nbclust’ function from the ‘factoextra’ package.

Chart, line chart

Description automatically generated

Following this munging/preliminary step in the process, various models and functions were prepped for use within the analysis section of the overall ‘Clustering’ topic; for the hierarchical aspect, ‘agnes’ and ‘hclust’ algorithms were chosen. Before running the respective algorithms, an overall method was determined before use. Available options for these methods in determining distance are listed as follows:

* Average
* Single
* Complete
* Ward

In order to determine which method would serve best for the analysis, a custom function was developed to generate, compute, and compare agglomerative coefficients (which conveys overall quality and fit for the clustering algorithm). In ranking from best to worst, the ‘ward’ method indicated ~0.88, followed by ‘complete,’ ‘average,’ and ‘single which generated results of ~0.85, ~0.73, and ~0.60, respectively.

Since ‘ward’ yielded the highest agglomerative coefficient value, this method was chosen for the ‘agnes’ and ‘hclust’ algorithms. With keeping 3 clusters in mind, the following dendrogram was generated (with state information shown near the bottom of the figure):

A picture containing diagram

Description automatically generated

Lastly, the ‘K-Means’ algorithm was also chosen as a viable algorithm for the ‘Clustering’ topic. The three (3) overall clusters consisted of 16, 3, and 31 observations (which equate to 50 – the total number of US states). The ‘Sum of Squares’ values by cluster equate to the following:

* ~75.7
* ~11.5
* ~108.9

The algorithm is also accompanied by a visualization via the ‘fviz\_cluster’ function of the same ‘factoextra’ package that was introduced earlier in this section.

Chart

Description automatically generated

As shown in the Cluster Plot, three (3) distinctive clusters are formed, containing US state information. The larger the geometric point, the higher the mean poverty rate for the state. These subsets of states may lead to valuable classification in latter stages of the report when supervised learning methods are introduced to the audience. We may be able to preliminarily state that clusters with higher poverty rates may be correlated with food deserts.