**Unsupervised Learning/Clustering Analysis**

**Overview**

In contrast to the linear regression and random forest models, unsupervised models do not have an accuracy score. Rather, unsupervised machine learning looks for clusters within the data that may not be obvious otherwise. In an unsupervised analysis, the first step is to determine how many clusters to instruct the algorithm to look for. Next, the model is fit to the data and predictions are made. Then comes the fun part, which is exploring how the algorithm grouped the data.

**Methodology**

To create my unsupervised learning model, I first imported my preprocessed data from my random forest model. Then, as in the random forest model, I scaled the data. Next, I reduced the dimensions in the data using principal component analysis, which reduced the number of features being input into the model to 30, which accounted for about 80% of the information in the data. The next step was an elbow curve analysis which suggested that 7 was a good number of clusters to look for in the data. Finally, I fit and ran K-Means clustering model which predicted a class for each county between 0-6.

**Results**

The clusters that emerged from the analysis can be seen in the diagram below, or for a better (and bigger) look at the same chart, click this link to view it on Tableau Public: <https://public.tableau.com/app/profile/martha.richardson/viz/JobGrowth_16536764067130/Clusters?publish=yes>.

A picture containing timeline

Description automatically generated

The clusters here can be used in a variety of ways. They can help you understand trends you already know about. They can be used to inform how you split data into subsets for more accurate models. I reran the multiple regression for each cluster individually, and the accuracy of the results was an improvement over the whole dataset regression for every group. Just looking at these clusters suggested to me that location is an important part of the underlying factors that influence rates of job growth. Location is very closely related to % rural, and so I decided to rerun the multiple regression with subsets based on how rural or urban a county is, and those results were the most accurate ones I got.

Below are three scatter charts that show the average job growth rate for each class against three different measurements: average % rural, average median income, and average rate of high school graduation. The tables can help to identify some of the characteristics of the clusters, though there are dozens of other charts we could make also.

**Average Employment Growth Rate Vs. Average % Rural**

Chart, scatter chart

Description automatically generated

**Average Employment Growth Rate Vs. Average Median Household Income**

Table

Description automatically generated

**Average Employment Growth Rate Vs. Average High School Graduation Rate**

Table, scatter chart

Description automatically generated