Capstone Project 1: Exploratory Data Analysis

The purpose of this data analysis was to identify trends and relationships between variables in the data set that may be useful in building a model to predict unemployment rates. To be able to build a model, I need to know which variables are useful in predicting unemployment rates, and which ones are not useful. Including the correct variables, and eliminating those that are not useful will ensure that the model is more accurate.

As I explored the data, and applied inferential statistics techniques, I noticed several interesting patterns and relationships. Some of them were similar to what I expected to find, but there were others that were surprising. I used various types of graphs and plots to investigate trends in the data, and then used linear regression, confidence intervals, and hypothesis tests to confirm the findings.

I was interested in how the unemployment rates in the U.S. have changed over the period from 2007 to 2017. I constructed a time series plot for the U.S. data, and also for each region in the U.S. (Northeast, South, Midwest, and West). I noticed a very similar trend for each graph, an increase from 2007 to 2010, and then a gradual decrease since then. Further analysis indicated that overall, there has been a slightly negative trend in unemployment rates.

I wanted to know which other variables might be correlated to Unemployment Rates. To look at the correlation between various variables, I used the pairplots function from Seaborn to construct scatterplots of all possible pairs of variables. Before constructing these plots, I had to deal with missing values in some of the columns. Since I had median income data for the states, but not for all the counties within the state, I filled in the county values with the median income for the state for that year. For the education data, I had estimates for the period from 2012 to 2016, so I used this same value for the other years. I had population data for 2011 through 2017. To estimate the population for the other years, I calculated the rate of change from 2011 to 2012, then used that rate to estimate the value for the previous years.

After constructing the scatterplots, I looked for patterns. I was especially interested in the correlation between education level and unemployment rates. I have four variables for the education level: percent of population with less than a high school education, percent of population with a high school education, percent of population with some college education, and percent of population with a bachelor's degree. For the first two, the scatterplots seemed to indicate a positive correlation, and for the last two, a slight negative correlation. For percent with less than high school, the correlation is much stronger than for percent with a high school education. For percent with some college and percent with a bachelor's degree, the strength of the correlation was about the same for both. It appears that education level may be a good predictor of unemployment rates.

Another variable that I was interested in was median income. Although I was only able to access median income for the states, and not for individual counties, it appeared to be correlated to unemployment rates. The scatterplot shows an obvious negative trend, and the correlation coefficient was – 0.36. This means states that have a higher median income tend to have lower unemployment rates.

After visualizing the data, and looking at the correlation between the variables, I conducted several hypothesis tests. I noticed that there was a negative correlation between percent of population with a bachelor's degree and unemployment rate. I wanted to know if the difference in the unemployment rate between areas where 25% or more have a bachelor's degree, and areas where the percent is less than 25%, is significant. My null hypothesis was that the mean unemployment rate was the same for these two areas, and the alternative hypothesis was that the mean is higher in areas where the percentage is less than 25%.

After dividing the data into two groups based on the percentage with a bachelor's degree, I looked at the distributions. Both distributions were positively skewed, and the scipy normaltest indicated that the data was not from a normal distribution. Also, there was a large difference in the sample sizes, and also the variances. I used the Mann-Whitney test, since it is more robust, and does not require that the data comes from a normal distribution. This test gave me a p-value of almost zero, so we have evidence to support the alternative hypothesis. This indicates that the mean unemployment rate is higher in areas where less than 25% of the population has a bachelor's degree. I then tested the same hypothesis using the bootstrap approach to simulate the data. I generated 100,000 bootstrap replicates, and got a p-value of zero. So the bootstrap approach results in the same conclusion.

I also have several categorical variables, State, Area, and Region, that I will consider as possible predictor variables for my model. I converted these to numerical values, so that each value of the variable is represented by a numerical code. This will allow me to include these in the model, and look at how it affects the accuracy of my prediction. Now that I have some information about the relationships between the variables, I am ready to start building a model to predict the unemployment rate for an area.