Capstone Project 1: Milestone Report

Introduction

The goal of this project is to be able to predict the unemployment rate for a region of the United States, given the population estimate, median household income, and education data. I want to know how unemployment rates have changed over the past 10 years, and how they vary between different states and regions of the U.S. I would also like to know how population, median household income, education level, and other factors are related to the unemployment rate. This information could be useful for agencies that are working to reduce unemployment rates, or provide services for those who are unemployed. It could provide insight on which areas should be targeted, and what strategies could be used to reduce unemployment rates.

To be able to predict unemployment rates, we need to look at past trends, and investigate the correlation between various factors that may affect unemployment rates. If we can determine which variables are strongly correlated with unemployment rates, we can use these variables to build a model to predict the unemployment rate for a region.

Data Collection and Cleaning

The data used for this project was obtained from the USDA government site and contains unemployment rates, population estimates, median household income, educational attainment, and other data for all states in the U.S. It also contains some data for all counties within each state. I was able to find additional income data from the United State Census Bureau site. The data is contained in downloadable excel files. Since the data was in several different files, I had to merge the desired data from each one. The data also required some cleaning and reformatting so that it could be plotted and analyzed. There are many variables contained in this data, so I had to determine which ones are needed to answer the questions I wanted to answer.

I began by reading the data (in csv files) into a Pandas dataframe, and looking at the overall structure, and the format of the columns. The Pandas library contains many tools that are useful for data analysis, and formats the data in a table (similar to an Excel spreadsheet) called a dataframe. In this dataframe, each column is a variable, and each row contains a value for each variable. Using some of the built-in functions in Pandas, I was able to look at the datatypes of all the columns, and find missing values. I dropped the rows containing all null values. Some of the data sets contained data from U.S. territories in addition to the state data. I wanted to use only the data from the 50 states and the District of Columbia, so I used a built-in function in Pandas to drop all rows containing data for the territories. I created a new dataframe consisting of only the desired rows and columns.

After extracting the desired data, I looked at the individual columns, and did more cleaning. Most of the columns contained numeric data, but it was formatted as 'string' type data. To be able to use this data for plotting and analysis, I needed it as integers or floating point numbers. Some of the values contained commas and/or the dollar sign, which I needed to eliminate so that the values are recognized as numbers by the computer. I defined a function that takes a column as input, then removes all non-digits from each value in the column and converts the value to a floating point number. The function returns a series containing these new values. I then replaced the column in the dataframe with this series. The data also contained a column with the 'Area', which consisted of the county name. However, some of the data files had only the county name in this column while others had the county and state. For example, some of the data had Autauga County, Alabama. Others had only Autauga County. I already had the state name in a separate column, so I did not need it again in this column. To clean this column, I used a built-in function that splits each string on the comma. I then saved only the first part, so that the county names were formatted the same throughout the column.

As part of the analysis, I wanted to compare different regions of the U.S., so I needed a column containing the region (Northeast, South, Midwest, or West). I created a list of states for each region, then used this to subset the dataframe. I created a for loop to iterate through the 'State' column and select the correct region, then added a new column containing these values. This allowed me to look at each region individually, and compare different regions of the U.S.

I also wanted a column with the year. Several of the datasets had the year in some of the column names. For example, the population data had columns with the population estimate for each year. So the column Pop_Estimate_2010 contained the population estimate for each region for 2010. I wanted one column with the year, and another column with the population estimate. I set the index as the 'State' and 'Area' columns, then used the Pandas .stack() function to combine all the remaining columns into two columns, where the first column contained the previous column names and the second column contained all the values. Then I cleaned each column using the previously mentioned function to remove all non-digits from each entry. This gave me a column containing the year, and another column containing the population estimate.

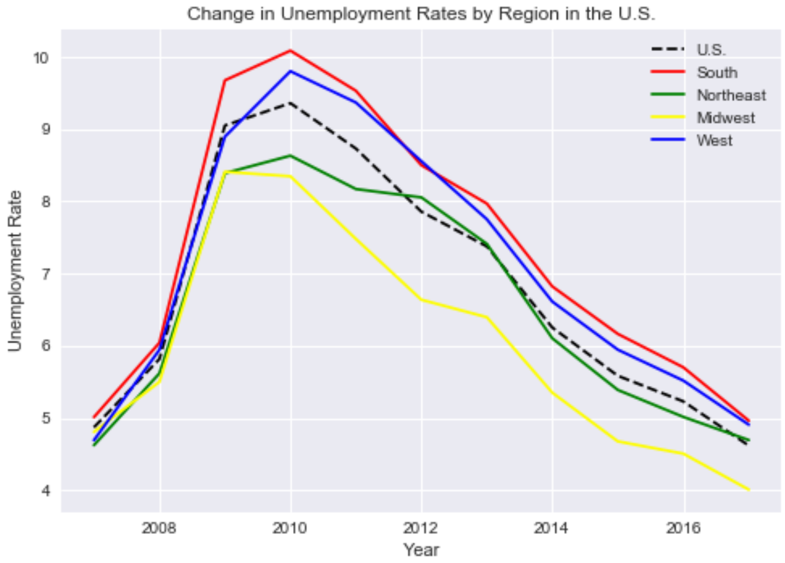
I used a similar technique for the median income data so that I had one column with the year and another column with the median income. For the education data, I was only able to find data for the years 2012-2016, so I had to fill in null values for the other years.

After I finished cleaning each data set individually, I exported each one as a csv file. I now had files containing unemployment rates, education data, population data, and median incomes. Each one was formatted in a way that would make it easy to plot the data and conduct analysis. The next step was to combine all this data into one dataframe. I read in each csv file as a dataframe, then used the Pandas df.merge() function to combine the dataframes into one. After merging the education data, I used numpy.nan to fill in the values for the years for which I did not have education data. I also added a column with the change in unemployment rate from the previous year. I did this by using the 'Year' and 'Unemployment Rate' columns. Using a for loop, I subtracted the previous year unemployment rate for each one. If the year was 2007, I filled in 'nan', because I did not have data for the previous year.

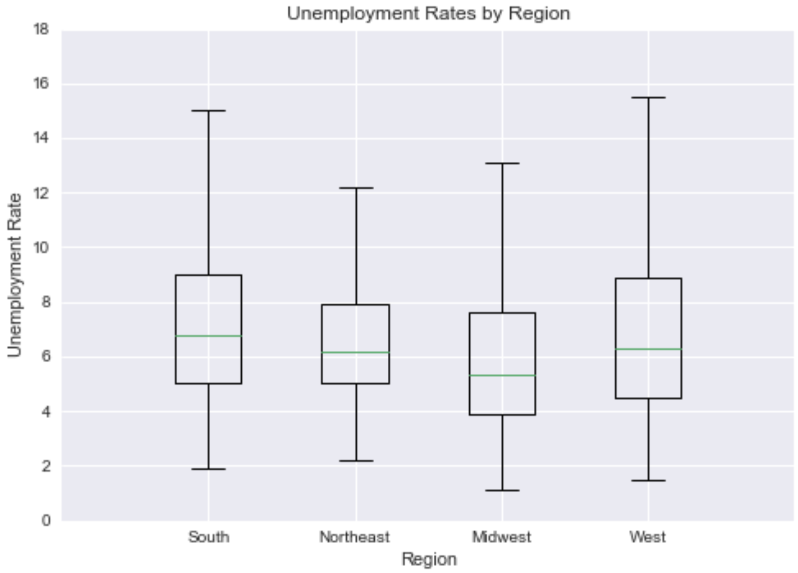
After merging all the data, I had a dataframe with 14 columns and 35,121 rows. I had a column for the state, one for the area, and one for the year. The rest of the columns contained unemployment rates, income data, population estimates, and education information for each state and area. Using this data, I was able to construct some plots and gain more insight about the dataset.

Exploratory Data Analysis

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(shown below) for the U.S. data, and 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.



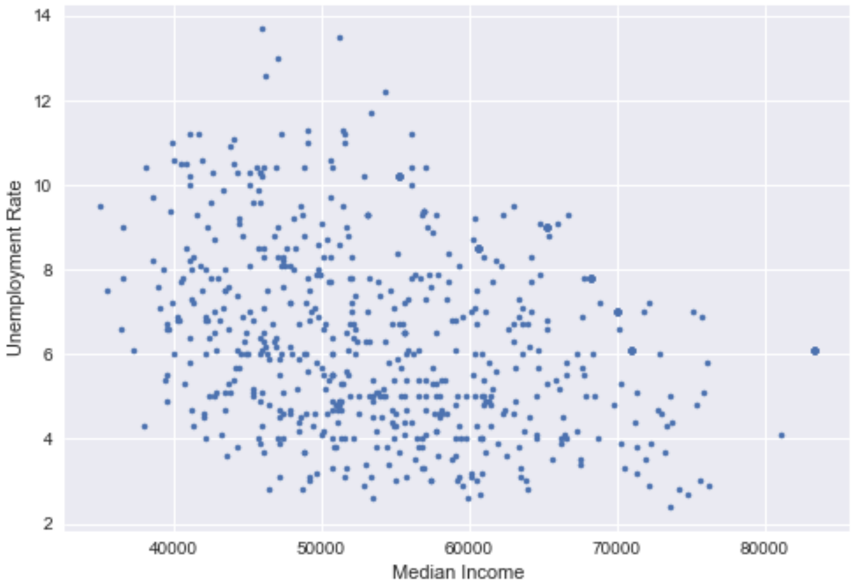
In this plot, we can also see the difference between regions of the U.S. It is obvious that unemployment rates in the South are typically higher than average, and in the Midwest they are lower than average. This difference is shown again in the boxplot below, where we can also compare the median and the range for the regions.



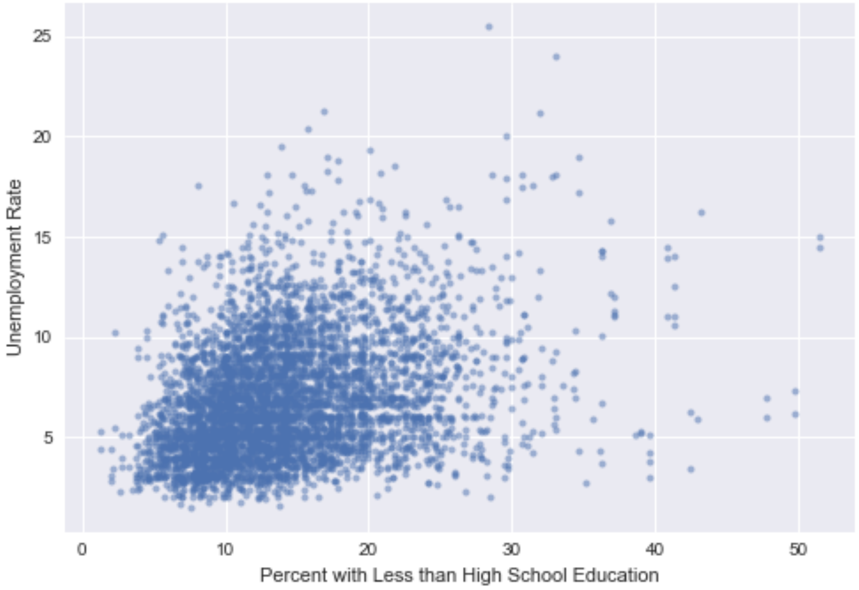
I also wanted to know how the individual variables might be correlated with Unemployment Rates. I used the pairplots function in the Seaborn library to construct scatterplots of all possible pairs of variables in the dataset. This would allow me to see possible correlations between any of the variables, so that I can determine which ones will be most useful in predicting Unemployment Rates. Before constructing the plots, I had to deal with the missing values in the dataset, because the pairplots function will only work if there is no missing data. 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 population for each year prior to 2011.

There are also several categorical variables in the data set that might be useful for predicting unemployment rates. 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.

After looking at the pairplots for all the variables, I constructed a few individual plots to investigate the relationships further. The plot showing the relationship between Median Income and Unemployment Rates is shown below. I used only the data from the states for this plot. There is a noticeable downward trend, which indicates that states with higher median incomes generally have lower unemployment rates. So it appears that median income might be a good predictor to include in the model.



I was also interested in the correlation between education level and unemployment rates. The education level variables appeared to be more strongly correlated with unemployment rates than most of the other variables. One example is shown in the plot below, which shows the relationship between 'Percentage with Less Than a High School Education' and 'Unemployment Rate'. I used a random sample of 5000 of the data points to construct this plot, so that the plot is easier to interpret. There is an obvious upward trend, which indicates that areas where a higher percentage of the population has less than a high school education tend to have higher unemployment rates.



After looking at the correlations between variables, I conducted several hypothesis tests to investigate these relationships further. For example, I wanted to know if there was a significant difference in unemployment rates between areas where less than 25% of the population have a bachelor's degree, and areas where the percentage is at least 25. My null hypothesis was that there is no difference, and the alternative hypothesis was that the mean unemployment rate 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 to determine what type of test to use. Both distributions were positively skewed, and there was a large difference in the sample size and the variance. I used the Mann-Whitney test, since it does not require that the data is from a normal distribution. This test resulted in a p-value of almost zero, which means that we have evidence to support the alternative hypothesis. I also tested the hypothesis using the bootstrap approach to simulate the data, and got similar results. So we can conclude that the mean unemployment rate is significantly higher in areas where less than 25% of the population has a bachelor's degree.

As I begin building the model, it is likely that I will find other interesting patterns and relationships in this data. I may need to conduct other tests to investigate these, and to determine how to construct the model so that I can accurately predict unemployment rates for any region.