Name: Amy Nestingen  
Date: 2/22/2019

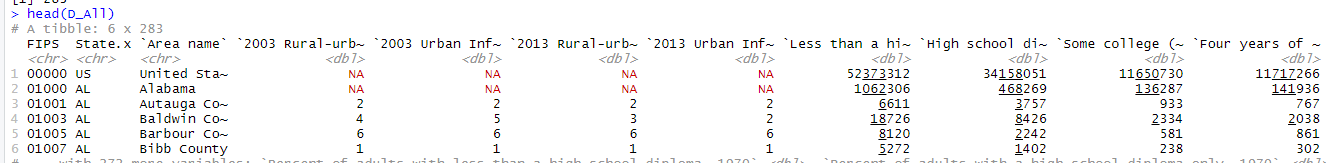
Relationship between Employment, Education, Population and Poverty

**Section 1 – Week 9**

* Intro
  + I am interested on how Poverty, unemployment, population, and education are related. If one goes, up how are the other 3 effected. How does this differ by state and region?
* Research questions
  + How can poverty rate, unemployment rate, and population predict education?
  + When unemployment goes up does education go down in subsequent years?
  + What effect does education have on poverty and unemployment rates?
  + How is the distribution of people in each educational group differ between states?
  + How are the number of children in poverty relate to the unemployment rate?
* Addressing Problem Statement
  + I plan to look at the data set in many different ways. First, I will combine the data sets into 1. Then I will predict, poverty, unemployment, population, and education in 4 different models with the other 3 variables being the predictive variable. I also want to compare states. I my break the states up in regions so that the data sets are larger.
* How your approach addresses (fully or partially) the problem.
  + My approach above will address this problem because I will be able to see the relationship between the variables clearly after modeling the 4 scenarios. Also, by grouping states together, there may be some regional trends that will be seen.
* Data
  + USDA: This link contains all 4 data sets that I will use
    - <https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/>
  + Documentation of Data Sets: This link explains how the data sets were collected and modifications made
    - <https://www.ers.usda.gov/data-products/county-level-data-sets/documentation/>
* Required Packages
  + Ggplots2
  + dplyr
* Plots and Table Needs
  + Scatter plots
  + Histograms
* Questions for future steps.
  + I need to figure out how the 4 data sets fit together. The are each broken down by state but I need go through the specifics how to connect them. I also need to make sure the data sets are using the same time frame.

**Section 2 – Week 10**

* Data importing and cleaning steps are explained in the text and in the DataCamp exercises. (Tell me why you are doing the data cleaning activities that you perform). Follow a logical process.
  + I am concatenating 4 files: Unemployment, Education, Population Estimates, and Poverty Estimates. The data is at a county level and the county ID matches throughout the 4 spreadsheets.
  + Only Education and Unemployment has Puerto Rico data so I will remove this from the data.
  + There are 3 AK counties that have spotty data so these will be thrown out too.
  + I deleted the headers so that the column header is the first row in the table.
  + I added a type column to the Education Sheet. This will say if the row is a country, state or county.
  + I used the dplyr package in R to inner join the 4 tables.
  + Thankfully, the column names are descriptive so I will be able to know what columns to use in the analysis.
  + The total state data is intermixed with the county data. I need to remove this.
* With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.
  + Below are the first 6 rows of the joined table.



* What do you not know how to do right now that you need to learn to import and cleanup your dataset?
  + I had to research how to join tables in R. Because my data sets shared a county key, this was not too difficult.
  + I also had to relearn how to subset the data so I had one data set for county, state and USA.
* Discuss how you plan to uncover new information in the data that is not self-evident.
  + I was going to look at the data on a state level but because the data is by county, I will explore the data at this level. By looking across the 4 spreadsheets, there may be other predictor variables that have not been analyzed.
* What are different ways you could look at this data to answer the questions you want to answer?
  + Scatter Plots will be helpful to compare two variables. It will be crucial to make sure I am comparing data from the same year. Some data sets group year and some have a separate column for each year. This will be a watch out.
* Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.
  + I did join the 4 tables and then separated the data sets by County, Country, and State. This is so I can compare apples to apples. A similar analysis that is done on the county data set could be applied to the state data set.
* How could you summarize your data to answer key questions?
  + A key question would be Why do some counties have higher population rates than others? What drives unemployment rates? Some predictors might be population growth, education levels, or poverty rates. I would first want to compare if the unemployment rates are statistically different. A good summary would be an ANOVA output that shares if the predicting variables are significant.
* What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate gems used, etc.).
  + Scatter Plots for an initial exploration would be valuable. Histograms will also be useful to understand the variables better. I could make a scatter plot with the x axis education levels and the y axis unemployment rate. I could then break up the counties based on the county size scale (1-12) and color the dots a different color. This combination of 3 variables can rotated across the data set to visually understand the relationship.
* What do you not know how to do right now that you need to learn to answer your questions?
  + I will have to refresh my memory on all of the ways to manipulate graphs. I will also want to continue my research on what outputs are useful for a multiple regression.
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
  + Right, now, I will not incorporate any machine learning. My data sets do not contain enough data for this.

**Section 3 – Week 11**

* What information is not self-evident?
  + I will start with Census\_2010 Population as the dependent variable. I will set up the code so that the population estimates from 2010 through 2018 can also be used as dependent variables. I will choose independent variables from 2010 to predict the 2010 population for each county. I will look at scatter plots with each independent variable and match it up with the dependent variable. Then I will test several regression options to decide how to best predict the population of a county.
* What are different ways you could look at this data?
  + I think I will start looking at the data one year at a time. I should also look at the data year over year to understand if there are any variables that should be weighted heavier in later years. There are rural and urban tags that could also be applied. I can look at the more urban counties as one group and the rural counties as another group.
* How do you plan to slice and dice the data?
  + I have combined 4 data frames to create one master table. The independent variable (population) was stored in one data frame and the other 3 dependent variables (poverty, education, and unemployment) were in 3 other data frames. The data frames came from the same website so they were relatively easy to join. They contained generally the same counties with just a few exceptions. Counties that were not consistent across the data frames were removed.
* How could you summarize your data to answer key questions?
  + I will be heavily relying on ANOVA test to summarize the significance of the dependent variables. This will help me determine what variables to keep in or take out. It will also be important to check for correlation between the independent variables. I will use a correlation matrix to observe this.
* What types of plots and tables will help you to illustrate the findings to your questions?
  + Scatter plots of the most significant variables will be helpful but this is only a 2D graph. I don't find 3D graphs helpful so I won't be using them. I could add color to a scatter plot to show a third variable. Creating a map that indicates the main driver of population for each county would be interesting.
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
  + I do not know how to create a map chart in r. This would be something I need to research. I also know that there are many ways to choose the best dependent variables. I will continue to research what ways I want to try and will execute a few of them.
* Questions for future steps.
  + As of right now, I do not plane to execute any machine learning. I do not believe my data set is large enough to do this.

**Section 4 – Week 12**

* A story / narrative that emerged from your data. Follow this structure.
  + Introduction.
    - It is important to understand what drives poverty. When the society understands what the main drivers of poverty are, they can be a more directed approach to combat it.
  + The problem statement you addressed.
    - My problem statement is "Predict poverty rate in a US County". By understanding what factors contribute most to unemployment rate, there could be more focused help on providing assistance to struggling counties.
    - I have combined 4 data sets together to create a master file. There is a data set for poverty, unemployment, population, and education. In my final model, I used data from poverty, unemployment, and education.
    - I started by creating scatterplots with poverty rate in 2018 as the dependent variable versus each of the main independent variables. The independent variables I was focused on included Population Estimate in 2018, Rural Urban Continuum Code 2013, 2013 Urban Influence Code, Percent of adults with less than a high school diploma 2014-18, Percent of adults with a high school diploma only, 2014-18, Percent of adults completing some college or associate's degree, 2014-18, and Percent in Unemployment 2018. From this investigation, I found that population had little to do with unemployment rate in 2018.
    - The scatter plots below gave me a visual representation of how the variables correlate to each other.
  + How you addressed this problem statement
    - I looked at different models to determine which one was the best. I also tried a log transformation. I determined the outliers and made a decision on which ones to remove.
  + Analysis.
    - The model that I thought was the best is log mod\_3 with the subset with the outliers removed. All of the coefficients are significant and there aren't any unnecessary variables included. The adjusted R squared value is .53. I thought it was interesting that the poverty rate could be predicted by the unemployment rate and education. These factors could be used to focus on decreasing the poverty rate.
  + Implications.
    - Because I removed outliers, this model would not do a good job predicting the counties that I removed. A deeper investigation would have to be done to determine the cause of these outliers.
  + Limitations.
    - One limitation is that the time span was not taken into account. It is possible that the unemployment in years prior could predict the poverty rate more effectively than the unemployment rate of the same year.
  + Concluding Remarks
    - The main drivers of poverty that I found is unemployment and education. There could be many other drivers that were not explored in this analysis. This is an important topic to continue to investigate.