Project I: Descriptive Analytics Data Mining Project

Victoria Roberts

ECON 494 (1:00PM)

Oct. 25, 2020

As COVID is transitioning into our ‘new normal’, unemployment has become an underlying cause of graduating students and individuals' constant stress in joining the workforce. Given that I’m currently a senior in college, joining the workforce has been of utmost importance these past few months. The data collected is about unemployment across countries and the variables that might cause or correlate the effects of unemployment. The dataset is focused particularly on 2017 and the fluctuations there might have been in the unemployment rate. This cross-sectional dataset helps understand the relationship between unemployment and other variables. Variables such as: Inflation Rate, Population, Public Spending on Education, and Nominal GDP. The data was sourced from publicly available information on gapminder and the global economy. The variables were chosen because of their relevance to each other. The tradeoff between inflation rate and unemployment rate have been the most monitored economic indicators. If unemployment were to decrease, workers would push for higher wages, resulting in the increase of inflation. Next, given the size of a country's population could help analyze the unemployment rate. I decided to incorporate an educational variable for this specific dataset. The education a person has received or the availability of education can help understand unemployment. Public spending on education is important to understand when viewing unemployment because if there are no resources for individuals to study they most likely will not be able to get a job. Lastly, changes in GDP can substantially burden the unemployment rate.

Once the data was collected from the public sources, each data column was inserted into excel. Given that the data was gathered through different resources, there were two ‘Country’ columns. Before downloading the dataset into RStudio I went through both country columns and deleted the countries that did not align with each other. For example, Bermuda wasn't available in the GDP data and was therefore eliminated. After deleting the specific countries that did not align as well as their data, I deleted one of the country columns so there was only one as reference for the whole dataset. Next, I imported the data as a CSV file in RStudio to further clean and analyze the cross-sectional dataset. Once in RStudio, I began a preliminary exploratory analysis to get situated with the data. I used functions such as dim, head, tail, summary, and summary specific to a variable to get the overall descriptive statistics. These functions are useful to see the dimensions of the dataset as well as its observations. Through this, I was able to see that there were 175 observations and 11 variables in the dataset. In addition, there were added columns that were not relevant to the dataset. From these preprocessing and preliminary tools, I

was able to proceed with cleaning the data. First, downloading the dplyr package was necessary to easily delete columns and rows. After downloading, I changed the datasets name to dataset2 so the cleaned data can be visible on a new RScript tab. This will also make it easier to compare datasets.

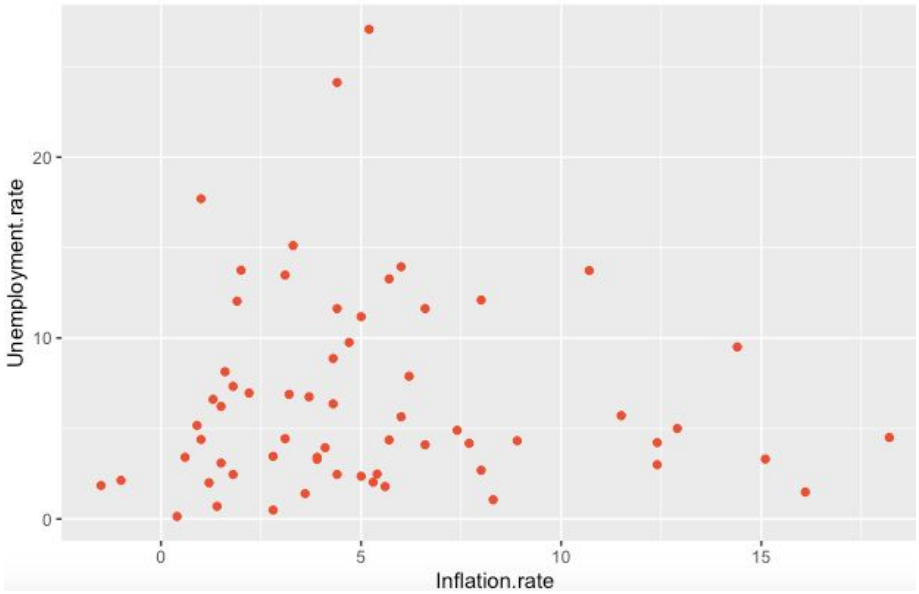
Next, I deleted the last three columns (X,X.1, and X.2) that were irrelevant to the dataset and gave the list a null value. To have a clean and tidy dataset, columns and rows that do not help explain or analyze the variables should be deleted. For this reason, I decided to eliminate columns 'Code' and 'Year.' After deleting these 5 variables, using the View() tool, I opened dataset2 on a new R Script tab to see if the variables were erased. The new dataset now portrayed the correct columns needed to identify correlations and relationships between variables, but the cleaning process was not done. The data was not complete and there was an array of missing and null values. The easiest way to remove rows and NA values is through a function called na.omit. The na.omit function returns a list of rows without NA values. Conversely, the complete.cases function returns a vector of rows with NA values. After using these two tools to see how many missing values were available in the dataset, I used the na.omit function to erase these values and assigned the new data to dataset2. Using the view function, I was able to see that the function worked and the new dataset had no missing values. Lastly, using the rename function I changed the column header names so they were more visually appealing and easily understood.

Now that the dataset is cleaned, an exploratory analysis can be used to further analyze the data and develop unique visualizations. Before starting the analysis, I used the preliminary tools such as head, tail, and summary to see the changes between the tidy and untidy datasets. After preprocessing and cleaning the data, dataset2 now contains 63 observations and 6 variables. Before downloading the ggplot2, I decided to explore the data with simple visualizations. Using the basic hist() function, RStudio generated histograms for variables such as unemployment, public spending on education, and inflation rate. To further analyze these variables I checked if the variables were normally distributed by using the hist() function with a prob= TRUE. Fig. 1 shows the first histogram of the unemployment variable. Fig. 2 shows the histogram with an added calibrated normal density curve. The histogram shows how the unemployment variable is skewed to the left. In addition, public spending on education is not normally distributed and is skewed to the left. The third variable, inflation, is somewhat normally distributed with a bell shaped curve (Fig.7). After reviewing the data through various histograms and scatter plots

(Fig.3), I downloaded the ggplot2 package to develop more unique visualizations to narrate a story with the dataset.

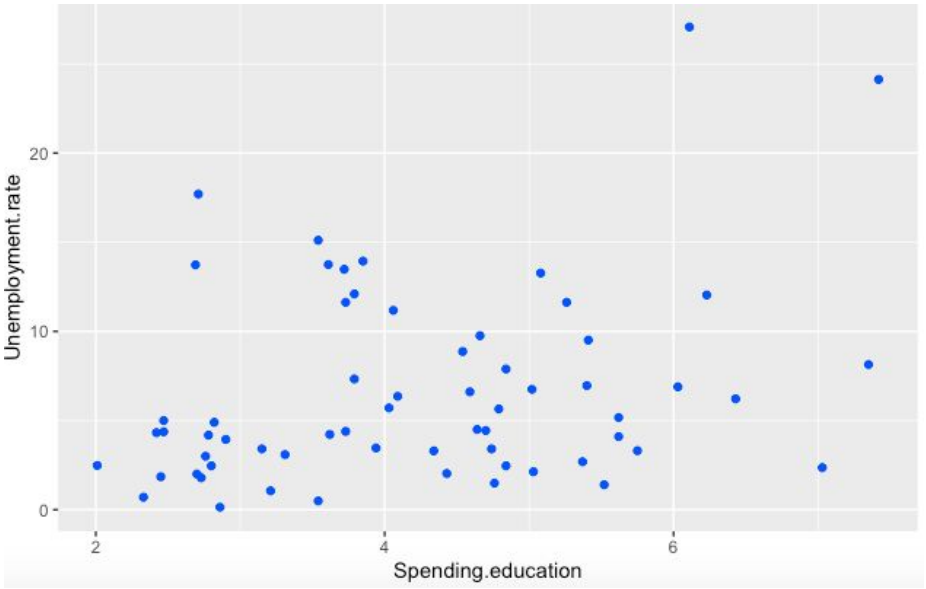
Using the ggplot, I created a scatter plot between inflation and unemployment rate to see if there was a relationship between these two variables. The scatter plot was generated by color rendering and using the point geom.

Figure 4: Scatter Plot



The scatterplot shown above, between the variables inflation and unemployment, highlights an inverse correlation. This means, as inflation increases unemployment decreases and vice versa. This makes sense because if unemployment falls, workers would want higher payments, resulting in higher prices and increased inflation rate.

Figure 5: Scatter plot (Spending on education vs Unemployment rate)



Next, Figure 5 scatter plot shows a little to no relationship between public spending on education and unemployment rate. There are some outlier points shown on the right corner that might affect the data. This shows that public spending on education does not correlate with unemployment rate and therefore does not affect its increase or decrease. Lastly, Figure 6 shows the scatter plot between unemployment and population size. The graph does not show a strong relationship between the two variables. The increase in population usually tends to increase the unemployment rate because there is such a high demand to join the workforce.

Overall, the results show the strong and weak relationships between what might affect unemployment rate. There is an inverse relationship between unemployment and inflation rate. This suggests that in 2017, there was a slight increase in inflation and decrease in unemployment. In addition, there was a weak relationship between public spending on education and unemployment. There is no correlation between these variables. I chose education as one of the variables in the dataset because I thought it would be interesting to see the effect of the unemployment rate, if there was more education available. The more education someone would have the less unemployment rate. Lastly, as GDP decreases the unemployment rate will increase given that there are less resources available.

Appendix

Figure 1: Histogram with Unemployment variable

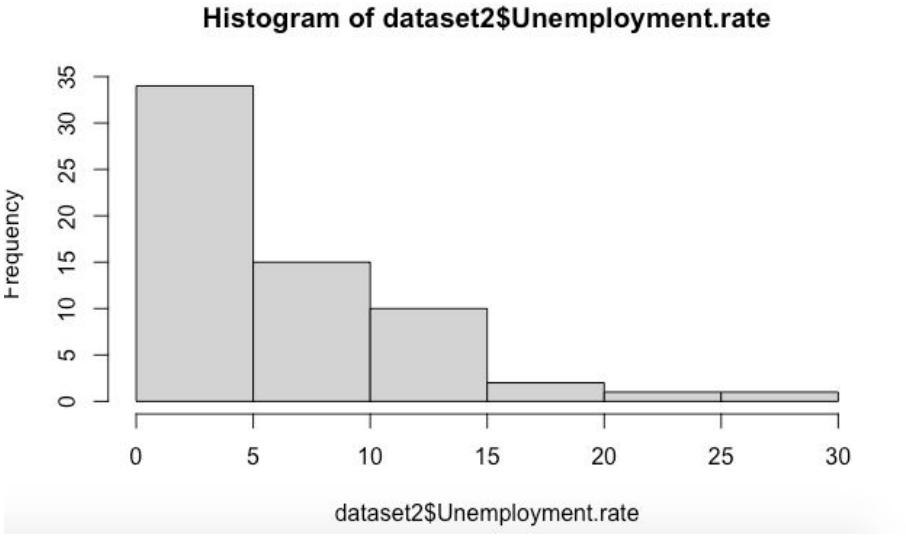


Figure 2: Histogram of Unemployment variable with calibrated normal density curve

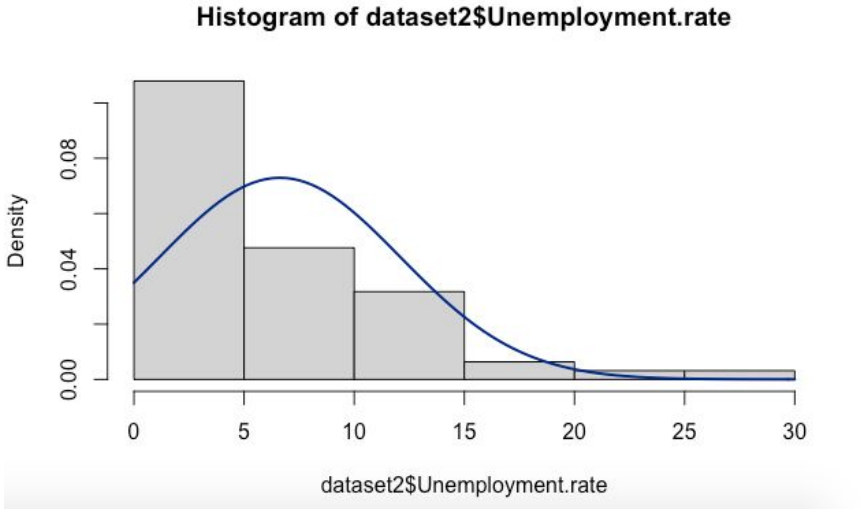


Figure 3: Scatter plot (Inflation vs Unemployment)

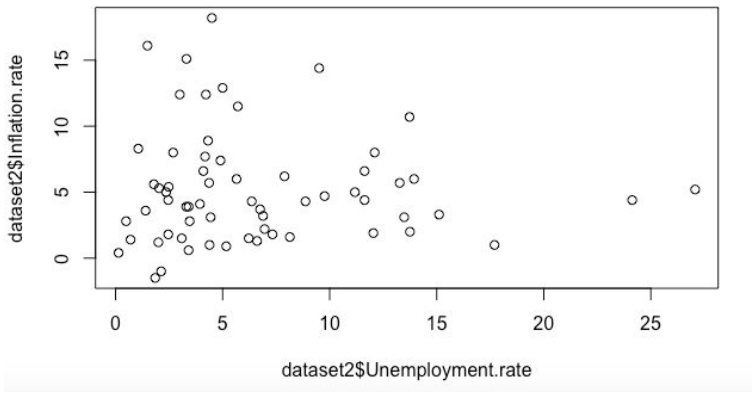


Figure 4: Scatter plot

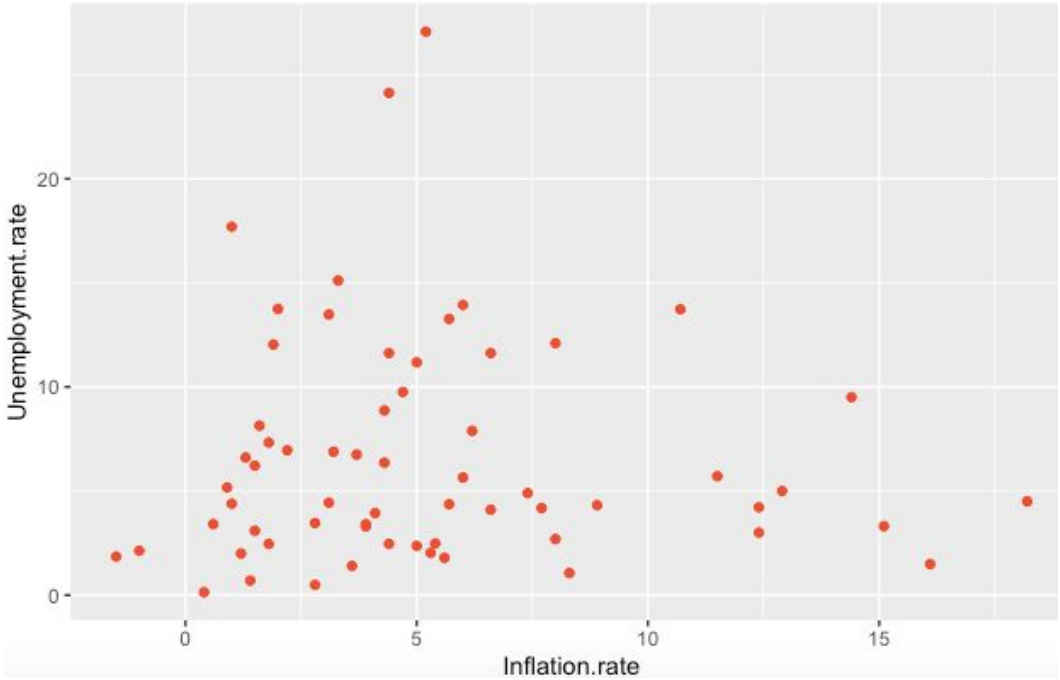


Figure 5: Scatter Plot

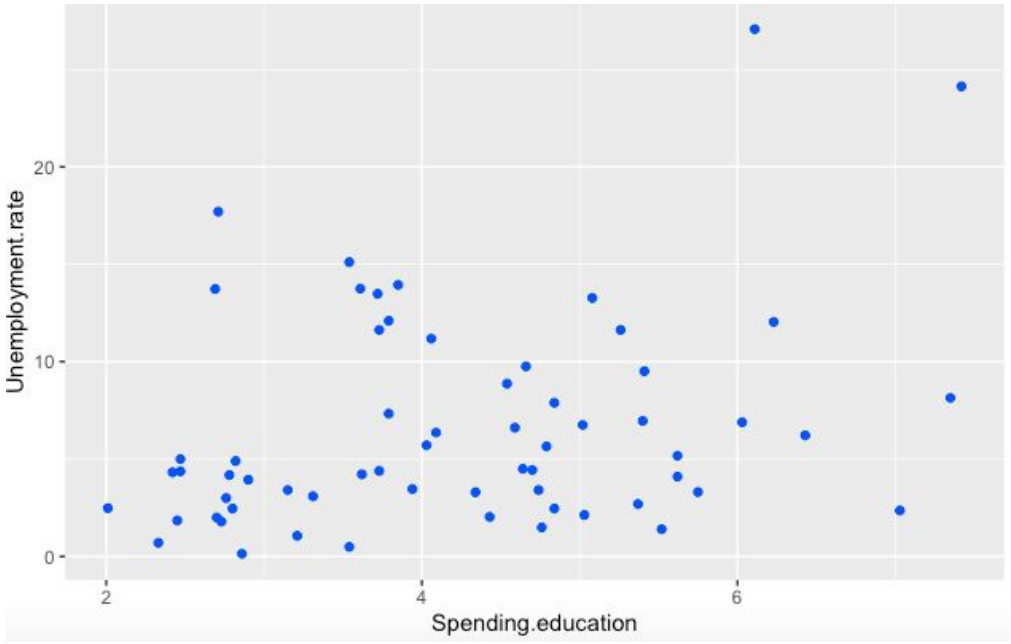


Figure 6: Scatter Plot

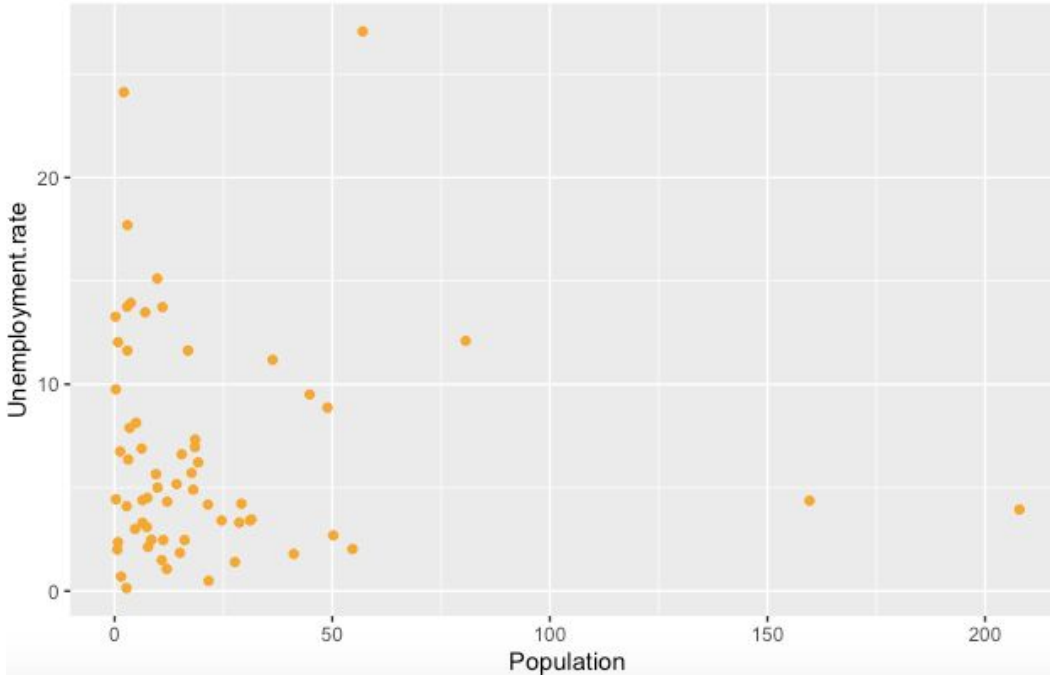


Figure 7: Histogram with calibrated normal density curve