Title: The Effect of Renewable Energy Statistics on Greenhouse Gas Emissions **Name: Payton Bartow**

Greenhouse gasses are accredited with climate change. Thus, this project's purpose is to find which countries contribute the most to emissions and what energy measurements can forecast emissions. This will allow leaders to identify how they contribute to emissions and how they can alter behavior to reach net zero emissions.

The data used for this project came from two sources, I treated each pairing of country and year as an observation. Data that contains greenhouse gas emissions in carbon dioxide equivalents was obtained from Our World in Data, https://github.com/owid/co2-data. I used this dataset to compute the total emissions of each observation by adding together the emissions of carbon dioxide, methane, and nitrous oxide since they were in equivalent units (1). I also used this data to find the share of world emissions (2) and the share of the world population (3) by dividing the observations values by the world values (a share is a percentage). I combined the data on emissions with data from Energy Data Info/World Bank, https://energydata.info/dataset/world-global-tracking-framework-2017. This dataset contained values related to renewable energy. I used the individual renewable energy output percent (4), renewable energy output (5), population share with electricity access (6), population share with access to clean cooking tech (7), urban population share with electricity access (8), rural population share with electricity access (9), and energy intensity (inefficiency of the energy economy) (10) and computed the non-renewable energy output (11) by subtracting the renewable output from total output. Next, I combined data yielding 7226 observations; 68.4% were missing one or more values. Therefore, I only included observations that contained values for every variable in the year 2014, the most recent year. This reduced the number of observations to 154. Some countries names did not match between datasets so 41 countries of 195 in the world were not included.

After cleaning the data, I compared each country's emissions share (2), their share of the world population (3), and individual renewable energy percentage (4), as seen in **Figure 1**. Only countries greater than or equal to 1% of emissions were individually represented, all other countries were placed in the "Other" category, their emissions percent and population share were summed, however, their renewable energy percentage was averaged. Furthermore, the top 13 emission producers accounted for 61.92% of the emissions, held 55% of the world's population, and had an average renewable energy percent of 17.46%, while 141 countries created 21.38% of emissions, held 25.57% of the population, and averaged 36% renewable energy output. The top 13 emission producers have the most work to do to ensure the global ecosystem does not collapse, especially the US and China who account for 38.43% of emissions and only 23.55% of the population.

Then I completed a principal component analysis (PCA) with variables 5 through 11. I used two iterations of a pipeline, with and without a standard scalar. **Figure 2** shows the cumulative explained variance ratio of both pipelines as the number of components increases. The standard scalar is the best fit as some values are percentages while other values are energy outputs. For example, the average values for renewable and nonrenewable energy output were $4.62 * 10^5$ and $1.93 * 10^6$ Terajoules. Without standardizing values, they are not comparable, yielding a high explained variance ratio for the first component of the non-scaled pipeline, .99. The explained variance ratio of the first component for the scaled pipeline was .53. Moreover, 5 components explain over 97% of the variance, so the dimensions could be reduced to 5 without losing information. However, the 5 components have limited predictive power. When the 5 components are used to predict total emissions via linear regression (linear regression was added to the end of the pipeline) the model scores .55/1 on 25% of the emissions data. When all components are used it scores a .88/1, therefore, all components prove necessary for regression.

For my last analysis, I investigated linear regression. I used variables 5 through 11 to predict the emissions (1) of an observation. I split the data into 75% training data and 25% testing data (the same 25% used to score the PCA regression). Then I fitted the training data using three different iterations of a pipeline to find the most accurate model. The first iteration of the pipeline fit the linear regression alone yielding a score of .84/1. The second iteration used a standard scalar transformation prior to fitting and had a score of .84/1. The third iteration used a standard scalar and 2 polynomial features and scored .39/1. The standard scalar and the untransformed regression had the same score; however, I chose to use the standard scalar pipeline as my model since its coefficients were more comparable. Moreover, **Figure 3** shows that the most important variables in the final regression model were nonrenewable and renewable energy output, with weights of 908 and 377, respectively (in the figure a negative value of weight means that the value is subtracted while a positive value is added). This makes sense since energy output has a direct relationship to emissions and even renewable energy produces emissions. All other variables were smaller in magnitude and unimportant.

Based on my findings the world has a long way to go before it reaches net-zero emissions as the top 13 emission producing countries in 2014 produce an average of 17.46% of their energy through renewables. Policy makers should try to optimize the best balance of total nonrenewable and renewable energy output since renewable output has a weight which is 41.5% of the weight of nonrenewable energy output.

Special thanks to my dad, brother, Sarah Eckhardt, and Professor Tyler Caraza-Harter for their guidance.

Figure 1: Comparison of Carbon Emissions, Renewable Energy Percentage, and Population Share

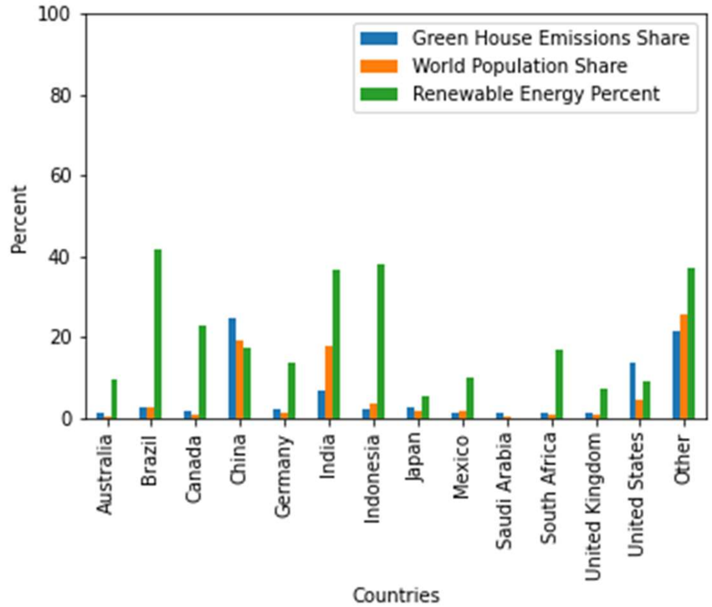

Figure 2: Cumilitave Principal Components

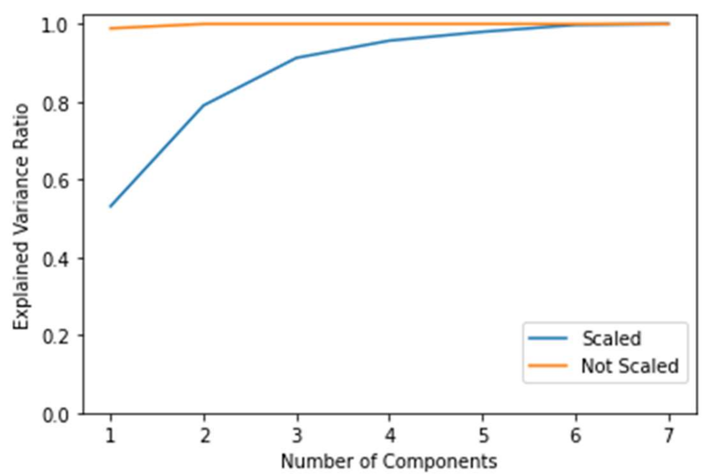

Figure 3: Linear Regression Coeffecients