**Grace Michallet**

**Project #2 – ETL**

**The questions to answer: Do access to electricity has a direct correlation to GDP and population? In other words, does higher GDP mean better access to electricity? And does a larger population mean less electricity access?**

**Extraction**

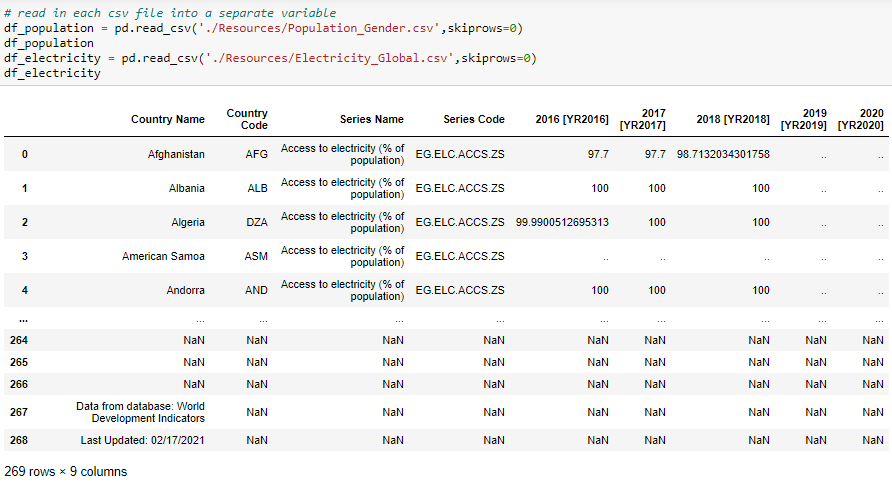
I used 3 datasets from World Bank. All of the data was based on Country from years 2016 to 2019.

The sources for my dataset are as follows:

* GDP in US$ by Country
* Population by Country
* Access to Electricity by Country (Note: It looks like data is not available for years 2019)

**Transformation**

The first steps in cleaning up the datasets was to figure out which variables are not relevant (**Figure 1**). I dropped the columns Series Name, and Series Code. I also deleted the rows that are null. I reformatted the data frame using the melt function then merged the Electricity and Population files into one data frame by Country (**Figure 2**). Then I removed values under country that is not a country (**Figure 3**). Lastly, I noticed that the year column is not formatted properly so I removed the values inside [ ] (**Figure 4**). I followed this step on all three data sets.



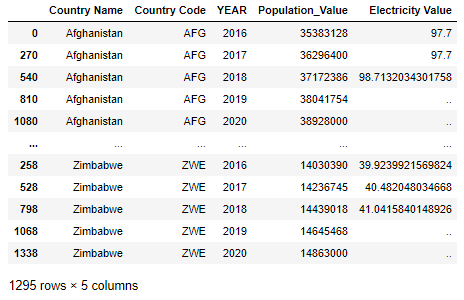
**Figure 1:** Access to Electricity Raw Data Frame



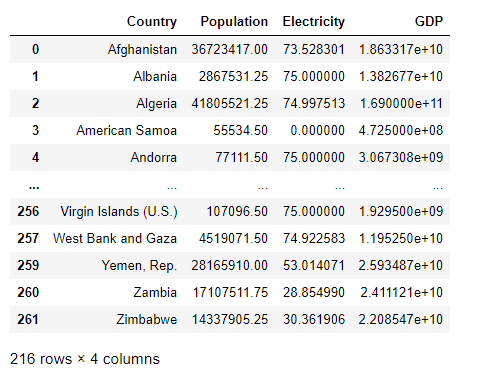
**Figure 2:** Reformatted the data sets using melt function then merged the Population and Electricity into one data frame.



**Figure 3:** Removed values under country that is not a country



**Figure 4:** Final output after merging the Population and access to Electricity data sets

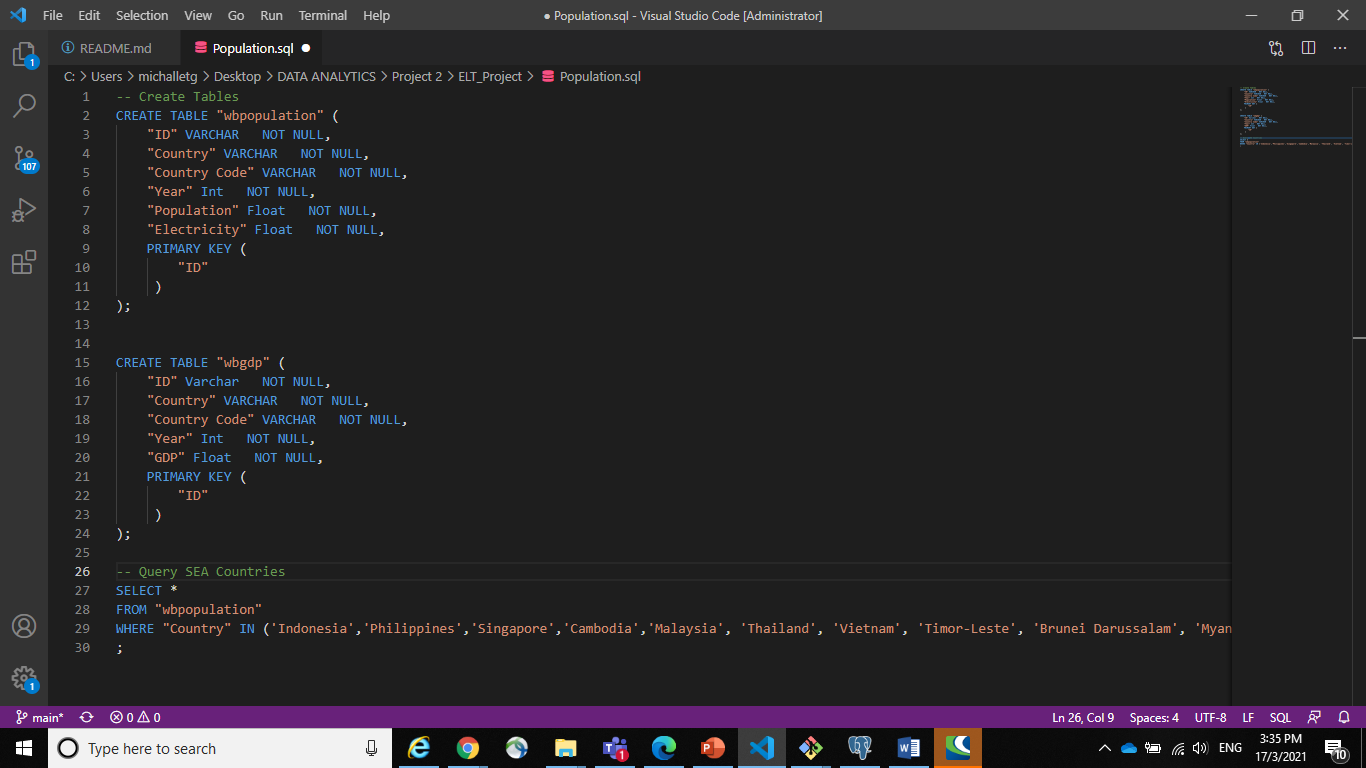


**Figure 5:** Final output of all three data sets in one data frame

**Load**

I transferred my output to a database. I created a database in PostgreSQL and created tables to match the columns from my panda’s data frame. I connected to the database using SQLAlchemy, loaded the results and performed multiple queries and analysis to answer the question.

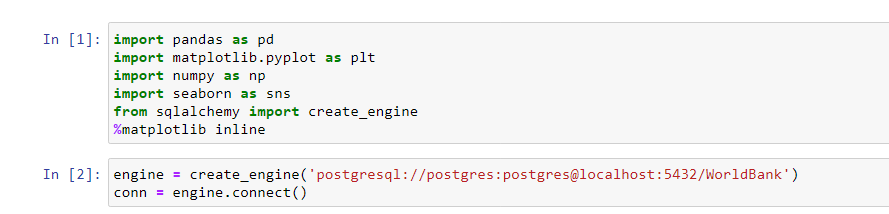
Sample Query (**Figure 6**) using PostgreSQL that returns all the variables for the South East Asian Countries.



**Figure 4:** Sample Query

**Summary**

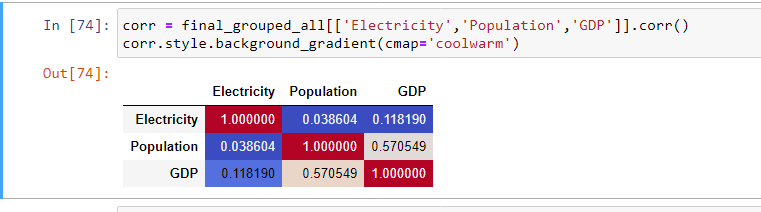
I performed an analysis using SQL Alchemy to determine whether GDP and population directly correlate to access to electricity.



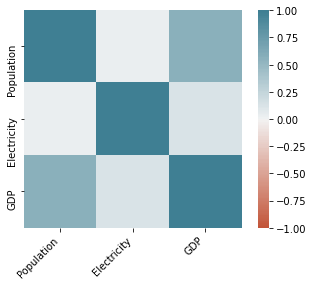
**Figure 5:** SQL Alchemy in Pandas

**Analysis 1:** **Do access to electricity has a direct correlation to GDP and population?**

Based on the correlation matrix below, the answer is there is no strong correction between electricity access, GDP and population.

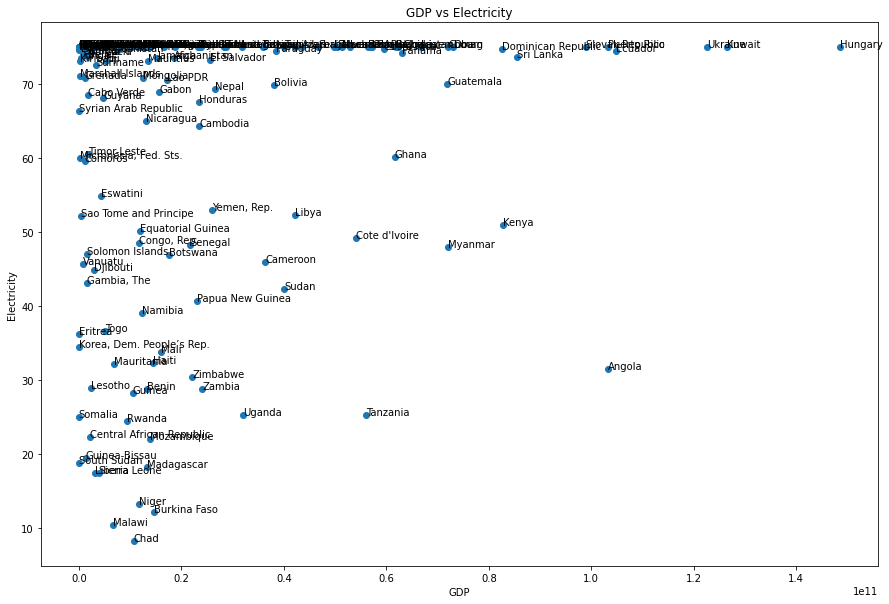


Green means positive, red means negative. The stronger the color, the larger the correlation magnitude.

****

**Analysis 2: Does higher GDP mean better access to electricity?**

Based on the below plot, it doesn’t appear that higher GDP necessarily mean better access to electricity.



**Analysis 3: Does a larger population mean lesser access to electricity?**

Based on the plot below, it doesn’t appear that a larger population mean lesser access to electricity.



The final output also will help recognize which Country has the following:

* Most or least population
* Most or least GDP
* Most or least access to electricity