Data Sources

* % population access to electricity
  + Source
    - <https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?locations=1W&start=1998&view=chart>
    - This is an external source as it is collected by a third party via industry, national surveys, and international sources.
    - It is collected by the International Energy Agency (IEA), International Renewable Energy Agency (IRENA), United Nations Statistics Division (UNSD), World Bank and the World Health Organization (WHO). The World Bank is a reliable source as it works with countries and national organizations to collect the most accurate data and follows internationally accepted standards which credits it as a reliable source of information.
  + Collection
    - This data is compiled from national surveys, census data and other sources going back to 1990. It also incorporates data from other databases like Socio-Economic Database for Latin America and the Caribbean. Overall, this data is a culmination of 1,375 surveys from 149 countries from the years 1990-2021
    - A limitation to this data is that is collected via survey and from multiple groups which means that is a chance for human error within this data.
  + Content
    - This data provides a year-by-year report from 1990 – 2021 on the percent of the population for each country that has access to electricity.
  + Relevance
    - This data shows information for many of the countries around the world and will provide good insights to the different levels of energy access from region to region.
* % of energy from oil, gas and coal
  + Source
    - <https://data.worldbank.org/indicator/EG.ELC.FOSL.ZS?locations=1W&start=1998&view=chart>
    - This is an external source as this data comes from the International Energy Agency via the World Bank. These are two well-known and trusted agencies so while the data may not be perfect, it is accepted as reliable.
  + Collection
    - This data was collected by the IEA. It used surveys and other measuring tools to account for energy produced by each country from oil, gas and coal.
  + Content
    - “sources of electricity refer to inputs used to generate electricity.” For this data that will include oil, coal and gas along with their derivatives.
  + Relevance
    - This will allow us to analyze changes in fossil fuels compare to changes in renewable energy, along with changes in overall GDP.
* % of energy from renewables and nuclear sources
  + Source
    - (renewables)<https://data.worldbank.org/indicator/EG.ELC.RNWX.KH?locations=1W&start=1998&view=chart>
    - (nuclear)<https://data.worldbank.org/indicator/EG.ELC.NUCL.ZS?locations=1W&start=1998&view=chart>
    - This is an external source as this data comes from the International Energy Agency via the World Bank. These are two well-known and trusted agencies so while the data may not be perfect, it is accepted as reliable.
  + Collection
    - This data was collected by the IEA. It used surveys and other measuring tools to account for energy produced by each country from renewable and nuclear energy sources. The IEA did provide the following limitations to this data:  
      * “IEA occasionally revises its time series to reflect political changes. For example, the IEA has constructed historical energy statistics for countries of the former Soviet Union. In addition, energy statistics for other countries have undergone continuous changes in coverage or methodology in recent years as more detailed energy accounts have become available. Breaks in series are therefore unavoidable.”
  + Content
    - This data contains a year-by-year percentage of total energy produced that came from renewable sources. These renewable sources include geothermal, solar, tides, wind, biomass, and biofuels. They do not include hydroelectric sources. This also includes nuclear energy, but will be provided in a separate column.
  + Relevance
    - This data will contribute to the understanding of each country’s energy make up and impact of overall GDP. This will also give a view into the non-fossil fuel distribution of energy production.
* GDP growth factors (country resources)
  + Source
    - <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2022&locations=1W&start=1961>
    - <https://wdi.worldbank.org/table/4.2>
    - This is an external source that comes from the World Bank national accounts and The Organization for Economic Cooperation and Development (OECD) National Accounts data files. These are both nationally and internationally recognized organizations that can be trusted for their data collection methods.
  + Collection
    - This data was collected by the World Bank and OECD.
    - “World Bank staff review the quality of national accounts data and sometimes make adjustments to improve consistency with international guidelines. Nevertheless, significant discrepancies remain between international and standards and actual practice…Among difficulties faced by compilers of national accounts is the extent of unreported economic activity I the informal or secondary economy. In developing countries, a large share of agricultural output is either not exchanged (because it is consumed within the household) or not exchanged for money.”
  + Content
    - That information used from this data will reflect the changes in different factors that contribute to GDP by country, comparing the year 2015 to 2022. The factors include Agriculture, Industry, Manufacturing, and Servies. It looks at the percent of GDP generated from each factor in 2015 and compares in toe the percent of GDP generated from each factor in 2022.
  + Relevance
    - This will enable a more indepth look on how GDP changes and the different factors that can affect it. Along with
* GDP per capita
  + Source
    - <https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2022&locations=1W&start=1990>
    - This is an external source that comes from the World Bank national accounts and The Organization for Economic Cooperation and Development (OECD) National Accounts data files. These are both nationally and internationally recognized organizations that can be trusted for their data collection methods.
  + Collection
    - This data was collected by using other data sources that reflect current U.S. Dollar GDP and total population.
    - “GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus and subsidies not included in the value of all products. It is calculated without making any deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. Dollars.
  + Content
    - This data is a year-by-year report on each country’s GDP according to the guidelines in the collection section above.
  + Relevance
    - This will enable the analysis of GDP compared to access to electricity as well as energy production. It will also work in the reverse to see if increased energy production helps drive GDP upwards.

Data Profiles

* Electricity Access
  + Shape
    - Started with 7,348 rows and 4 columns
    - No missing values and no duplicates
    - Ended with the same shape, 7,348 rows and 4 columns
  + Descriptive Statistics
    - Year
      * Minimum: 1990
      * Maximum: 2021
    - Percent of Population with Access to Electricity
      * Minimum: 0.53%
      * Maximum: 100.0%
      * Mean: 80.76%
      * Standard Deviation: 28.75%
  + Changes Made
    - Renmaed column ‘electricity\_access\_%\_population’ to ‘ea\_%\_population’ for simplicity.
    - Exported file name: ‘electricity\_access\_wrangled.csv’
  + Limitations
    - As said above, this data is collected from multiple sources and third parties. Because it is cumulative of so many different countries with different levels of development there may be errors within the data. This is the most up to and accurate data that we can have access to.
* Energy from Nuclear Sources
  + Shape
    - Started with 6916 columns and 4 columns
    - No duplicate values found
    - 2282 missing values in the percent of energy from nuclear sources
      * 198 of these countries have no data collected and should have zero imputed.
        + 225 total missing values
      * 68 of these countries are missing values for one or more their years and should have the weighted average imputed.
        + 27 total missing values
    - After imputing zero for the 198 and weighted averages by country for the 68 there are no missing values
    - Ended with 6916 rows and 4 columns
  + Descriptive Statistics
    - Year
      * Minimum: 1990
      * Maximum: 2015
    - Percent Energy from Nuclear
      * Minimum: 0.00%
      * Maximum: 87.99%
      * Mean: 4.17%
      * Standard Deviation: 11.57%
  + Changes Made
    - Renamed column ‘%\_energy\_from\_nuclear’ to ‘%\_energy\_nuclear’ to make less wordy
    - Exported file name ‘energy\_from\_nuclear\_wrangled.csv’
  + Limitations
    - As said above, this data is collected from multiple sources and third parties. Because it is cumulative of so many different countries with different levels of development there may be errors within the data. This is the most up to and accurate data that we can have access to.
* Energy from Oil, Gas, and Coal
  + Shape
    - Started with 6914 rows and 4 columns
    - No duplicates were found
    - 2155 missing values in the percent of energy from oil, gas, and coal sources
      * Similar to energy from nuclear some missing values require a imputed value of zero while others require a weighted average.
      * This time I imputed the weighted average values first by making a new weighted average column that filled in averages for missing values, which decreased the number of missing values to 2080.
      * Then I filled the weighted average column with 0 to replace all NaN points to finish imputing all missing values.
    - No more missing values after imputations
    - Ended with 6916 rows and 4 columns
  + Descriptive Statistics
    - Year
      * Minimum: 1990
      * Maximum: 2015
    - Energy from Fossil Fuels (oil, gas, coal)
      * Minimum: 0.00%
      * Maximum: 100.00%
      * Mean: 42.29%
      * Standard Deviation: 37.89%
  + Changes Made
    - Made a new column ‘weighted avg’
    - Removed original ‘%\_energy\_fossil\_fuels’
    - Renamed ‘weighted\_avg’ to new ‘%\_energy\_fossil\_fuels’ column
    - Exported file name: ‘energy\_from\_oil\_gas\_coal\_wrangled.csv’
  + Limitations
    - As said above, this data is collected from multiple sources and third parties. Because it is cumulative of so many different countries with different levels of development there may be errors within the data. This is the most up to and accurate data that we can have access to.
* Energy from Renewables
  + Shape
    - Started with 6916 rows and 4 columns
    - No duplicates were found
    - 2155 missing values in the percent of energy from renewable sources
      * Like energy from oil, gas and coal, some missing values require an imputed value of zero while others require a weighted average.
      * This time I imputed the weighted average values first by making a new weighted average column that filled in averages for missing values, which decreased the number of missing values to 2080.
      * Then I filled the weighted average column with 0 to replace all NaN points to finish imputing all missing values.
    - After imputations, no missing values
    - Ended with 6916 rows and 4 columns
  + Descriptive Statistics
    - Year
      * Minimum: 1990
      * Maximum: 2015
    - Energy from Renewables
      * Minimum: 0.00%
      * Maximum: 65.44%
      * Mean: 1.71%
      * Standard Deviation: 4.63%
  + Changes Made
    - Made a new column ‘weighted avg’
    - Removed original ‘%\_energy\_renewables’
    - Renamed ‘weighted\_avg’ to new ‘%\_energy\_renewables’ column
    - Exported file name: ‘energy\_from\_renewables\_wrangled.csv’
  + Limitations
    - As said above, this data is collected from multiple sources and third parties. Because it is cumulative of so many different countries with different levels of development there may be errors within the data. This is the most up to and accurate data that we can have access to.
* GDP per Capita
  + Shape
    - Started with 8778 rows and 4 columns
    - No duplicates were found
    - 572 missing gdp values
      * Most of the missing values are older values, but there are a handful of more recent values that are missing.
      * Imputed future values using a rolling or weighted average, which look at the past 3 values and imputed the average of those values for the missing value.
        + This account for 98 missing values
      * Imputed later values using a rolling average and bfill() function to earliest year within the data frame. Some of these values remained constant, but this will work for the data frame because most of the missing values are in the 1990s, and data improves beginning in the 2000s.
    - There were 4 countries that had no data, so rather than imputing zero they were removed from the dataframe. The number of rows removed was 132
    - Ended with 8646 rows and 4 columns
  + Descriptive Statistics
    - Year
      * Minimum: 1990
      * Maximum: 2022
    - GDP
      * Minimum: 22.85
      * Maximum: 234317.08
      * Mean: 12366.15
      * Standard Deviation: 20484.29
  + Changes Made
    - Created a new column ‘gdp\_copy’ to cross reference changes made to data
    - Created a new column ‘trend’ that had the weighted averages of the data
    - Dropped both columns when data was done being checked for consistency
    - Exported file name: ‘gdp\_per\_capita\_wrangled.csv’
  + Limitations
    - As said above, this data is collected from multiple sources and third parties. Because it is cumulative of so many different countries with different levels of development there may be errors within the data. This is the most up to and accurate data that we can have access to.

Questions

* Is there a significant correlation between GDP and access to electricity?
* Is there a threshold of access to electricity that significantly increases a country’s GDP?
* Are there significant correlations between GDP and energy production from different sources?
* Which, if any, energy production type improves a country’s GDP the most from year to year.
* What are the trends of energy production from year to year, for each country?
  + Do all countries follow the same trend?
* What countries are the top for energy production in fossil fuels, renewables and nuclear?
* What is the World’s GDP and how has it changed over time?
* Are there continents and regions that have a higher GDP than others?
  + What role does electricity access and energy production play in that?