**Data Cleaning and Documentation**

**Data Sources description and licensing**

All our data sources come from the US Energy Information Administration:

* U.S. Energy Information Administration. (Oct 2008). International Energy Statistics / Total Primary Energy Consumption [Data file]. Retrieved Oct 30, 2016. Available from<http://www.eia.gov/beta/international/data/browser/index.cfm#?ord=CR&cy=2013&v=H&vo=0&so=0&io=0&start=1980&end=2013&vs=INTL.44-2-AFRC-QBTU.A~INTL.44-2-ASOC-QBTU.A~INTL.44-2-CSAM-QBTU.A~INTL.44-2-EURA-QBTU.A~INTL.44-2-EURO-QBTU.A~INTL.44-2-MIDE-QBTU.A~INTL.44-2-NOAM-QBTU.A&c=4100000002000060000000000000g0002&pa=000000001s&f=A&ug=4&ct=1&tl_type=p&tl_id=44-A&s>=
* U.S. Energy Information Administration. (Oct 2008). International Energy Statistics / Total Primary Energy Production [Data file]. Retrieved Oct 30, 2016. Available from<http://www.eia.gov/beta/international/data/browser/index.cfm#?ord=CR&cy=2013&v=H&vo=0&so=0&io=0&start=1980&end=2013&vs=INTL.44-1-AFRC-QBTU.A~INTL.44-1-ASOC-QBTU.A~INTL.44-1-CSAM-QBTU.A~INTL.44-1-EURA-QBTU.A~INTL.44-1-EURO-QBTU.A~INTL.44-1-MIDE-QBTU.A~INTL.44-1-NOAM-QBTU.A&c=4100000002000060000000000000g0002&pa=004&f=A&ug=4&ct=1&tl_type=p&tl_id=44-A&s>=

Our original data source from the Info Seeking assignment was Max Roser (2016). This source is comprised of many datasets and the intent was to use one or many of them in addition to the UN Statistics database for our analysis. We noticed that the dataset we wanted to use from this source referenced the US Energy Information Administration (EIA) database as its original source. So instead of using Max Roser (2016) dataset’s data, we decided to go directly to the source of its data and get perhaps more current and updated data.

We also found that the UN Statistics database data was too granular and too complex to navigate in a way that it did not make sense to our analysis; at times we were lost in trying to figure out how to group the various data categories into a single one; hence the use of the EIA database for our analysis.

We encountered no licensing constraints with using EIA’s data as it is public domain and not subject to copyright protection. More details available at:<http://www.eia.gov/about/copyrights_reuse.php>.

**Metadata and data issues**

EIA website’s presentation of the data is appealing to us because it gives you an interactive way of formatting the data in different ways. The metadata is structured along different metrics and dimensions which makes the data easy to understand and interpret. Using the interactive interface for example, we could compile energy data, grouped in categories such as total production or go even granular by splitting it up by energy source such as total coal production or total petroleum production as you see fit. It also gives us the ability to categorize our metrics into different geographical regions, again with the option to go granular to individual countries if needed.

Missing values, or Not Applicable/Available values were one of the main issues we encountered in dealing with the data. A lot of countries did not have any data available up until about two decades ago and some of them had data points or figures that were too small for the number of decimal places shown in the data sheets. We also encountered some issues like data recorded for countries that do not exist anymore in today’s world geography. Examples include records of energy production data for West Germany in 1990.

**Cleaning process**

To remedy the issues with the data, we made a couple of decisions. The very first one is to remove all countries that do not have any energy production or consumption data available from the data source. As our analysis focuses on looking at energy production and consumption around the world, we felt that keeping these countries around would not be useful. Second, we reconciled country names with current nomenclatures. Some of the names used for some of the countries are no longer current as they have changed. Countries that have split over the years have been removed and their data has been merged into their remnants if possible.

Since our dataset files are in csv format, countries that have two word names like “South Korea” were put in between two commas. They were represented as “,Korea,North,” instead of “,Korea North,”. This caused some rows to overflow one column over into the next record when we attempted to read the files. We had to locate and readjust these records so everything is uniform. As far as the content goes, the data points in the datasets were recorded with the precision of 3 decimal places (0.001). Those that were under or too small for three decimal places were not shown and just marked with “(s)”. We assigned to these “(s)” values, a value of 0.0001 (lowest value of 4 decimal places notation).  Then, all empty and not available values have been replaced by zero. We made the assumption that for our particular analysis for example, the empty value would be interpreted on a graph or a plot somewhat the same way as zero; meaning that there would be either no energy production or no energy consumption for at that particular point for that time period for that country/region.

Following the above algorithm, the specific steps taken for our datasets are the following:

* Remove all empty rows
* Rename countries:
  + Congo (Kinshasa) was renamed to Republic of The Congo
  + Congo (Brazzaville) was renamed to Democratic Republic of The Congo
  + Korea, North was renamed North Korea
  + Korea, South was renamed South Korea
  + Virgin Islands, US was renamed US Virgin Islands
  + Virgin Islands, British was renamed British Virgin Islands
* Remove countries:
  + Former Serbia and Montenegro
  + Former Yugoslavia
  + Germany West
  + Germany East
  + U.S.S.R.
  + Former Czechoslovakia
* Merge countries:
  + Data for Germany West and Germany East before 1991(data was empty after 1991) were added together into data for Germany (data was empty before 1991). Both rows for West and East Germany were removed.
  + Data for USSR before 1991(data was empty after 1991) was merged into data for Russia (data was empty before 1991). Row for USSR was removed.
  + Data for Former Czechoslovakia before 1992 (data was empty after 1992) was merged into Slovakia (data was empty before 1992). Row for Czechoslovakia was removed.
* Values marked (“s”) are replaced by 0.0001.
* Empty values, N/A (not available and not applicable values) were replaced by zero.

**Citations**

1. Max Roser (2016) – ‘Energy Production & Changing Energy Sources’. *Published online at OurWorldInData.org.* Retrieved Sept 12, 2016. Retrieved from: https://ourworldindata.org/energy-production-and-changing-energy-sources/ [Online Resource]

Word Count: 976