Project Report

For this project, my focus in on extracting and organizing data about renewable energy in China, Germany and the United States. To elaborate, the data are on electricity installed capacity (MW) for both non and renewable energy, total energy consumption, renewable energy type consumption, CO2 emission, and 2030 and 2050 agricultural GHG emission goals.

Extract:

I extracted my dataset from four different sources: irena.org, bp.com, wesr.unep.org, and eia.gov. The dataset from irena.org was formatted as Excel (XLSB) file, the ones from bp.com were formatted as Excel (XLXS) file, the one from wesr.unep.org was formatted as CSV, and the one from eia.gov was in the form of JSON from API calling.

For easier read:

1. irena.org, bp.com
   1. Excel (XLSB) file
2. bp.com
   1. Excel (XLXS) file
3. wesr.unep.org
   1. CSV
4. eia.gov
   1. API

Transform:

I performed cleaning, filtering, joining, changing dataframe layout, aggregating and performing normalization for the datasets. I will list what I did for each dataset in details below:

1. irena.org, bp.com,
   1. Cleaned the data to get the dataset display in the proper way for filtering, renaming, aggregating and performing normalization
   2. Filtered out unneeded columns and rows, and then extracted data only from my three selected countries
   3. Aggregated the three countries’ datasets together as one
   4. Performed normalization of the dataset then renamed the columns
2. bp.com
   1. Cleaned the data to get the dataset display in the proper way for filtering, renaming, aggregating and performing normalization
   2. Filtered out unneeded columns and rows, and then extracted data only from my three selected countries
   3. Changed the dataframe layout, so the year values are under the year column
   4. Renamed the columns
   5. Aggregated the three countries’ datasets together as one
   6. Performed normalization of the dataset
3. wesr.unep.org
   1. Cleaned the data to get the dataset display in the proper way for filtering, renaming, aggregating and performing normalization
   2. Filtered to get only my selected years 2030 and 2050
   3. Aggregated the datatsets
   4. Performed normalization of the dataset then renamed the columns
4. eia.gov
   1. Renamed the columns and changed the order of the years from descending to ascending
   2. Filtered out some years for the US one
   3. Performed normalization of the dataset

Load:

I loaded the transformed dataframes to postgres in the form of tables. I chose to do that because the datasets are related to each other either by year or by country.