**Final Report**

ETL Project – Consumption of Energy Sources by Countries

Members: Michael Kung, Mary Gong, Kristen Cannon

***Extract: your original data sources and how the data was formatted (CSV, JSON, MySQL, etc).***

All of our original data sources were extracted from:

* <https://www.eia.gov>
* <http://data.un.org/Default.aspx>
* <https://yearbook.enerdata.net/>

We looked into multiple other website but majority of them only focused on US data, we wanted to find energy consumption data for all countries instead and the above two websites contained the most information.

We focused on the following energy sources and the annual consumption data across all countries:

* Solar Energy consumption
* Electric Energy consumption
* Natural Gas consumption
* Oil Energy consumption
* Coal Energy consumption

All of the data that we extracted were in CSV file.

***Transform: what data cleaning or transformation was required.***

Natural gas & Solar energy data from UN website was pretty clean and easy to use, not much data cleaning was needed. We had to remove a few unnecessary columns using the drop method.

For electric energy data from UN website, we had to rename all columns in order to separate the data between resident, industrial, commercial, and all sector. We also had to remove the commas in the data and convert it to numeric. A few rows with ‘NA’ data were dropped.

Natural gas data from EIA required a bit of cleaning and renaming, we had to rename all of the columns to start with a letter in order to call it in SQL (ex: 1990 cannot be used as column header on its own), so we ended up adding ‘Year\_’ In front of columns and dropped it for the first two column which did not start with an integer (Country, Unit).

We dropped unnecessary Rows that had ‘NA’ data, and had to figure out what type of data the table contained for SQL purpose, tried out both Float and Decimal.

***Load: the final database, tables/collections, and why this was chosen.***

In order to combine the data, we wrote a function to get the list of countries from UN database, it creates a country ID for unique countries, and this is used for joining purpose.

We chose to grab data for all countries that had available energy consumption data instead of our original idea of focusing on US only, this allows us to expand on research & analysis in the future if we want to look into specific reason why certain countries may use more energy compare to others (ex: population, household size, salary income, etc.) We also chose the time range between 1990 to 2017 to give us plenty of data in order to do comparison year over year.

We ended up with six different tables:

* Solar energy - UN
* Electric energy - EIA
* Natural Gas energy - UN
* Natural Gas energy – EIA
* Oil Consumption – Enerdata
* Coal Consumption - EIA