**Project 2 - Report**

**ETL Project – Consumption of US Energy Sources**

**Group 4 Members:** Robin Khan, Ted Phillips, John Klinges

**Project Summary:**

1. **Data Scope and Sources:** See EIA government website cited below.
2. **Uploading Data:** Data was read from excel files.
3. **Transform**: Data was cleaned and structured to suit project needs.
4. **Create a Database:** A local database was created.
5. **Load**: loaded the data into a database for storage that can be used for future analysis.
6. **Data scope and sources:**

**a) Data Scope:** We focused on the

1) annual consumption and production data, 2) for the U.S. only 3) for the following energy sources:

* Natural Gas
* Petroleum Energy
* Coal Energy
* Renewable (Solar energy, Electric energy, nuclear, wind etc.)

**b) Data Sources:** We used the following website (<https://www.eia.gov>) which was well designed for our project needs and because it provided both energy production and consumption data.

**c) Time Period:** We choose the time range between 1980 to 2019 to give us plenty of data in order

to do year over year comparisons.

1. **Uploaded Data**:
2. We uploaded original data sources from excel files.

**b)** We loaded the data into excel data files sorted by energy type:

natural\_gas\_df = pd.read\_excel('Resources/Natural\_Gas.xlsx')

petroleum\_df = pd.read\_excel('Resources/Petroleum.xlsx')

coal\_df = pd.read\_excel('Resources/coal.xlsx')

renewables\_df = pd.read\_excel('Resources/Renewable\_Energy.xlsx')

1. **Transformed The Data**:Data from the website was relatively easy to clean and unnecessary columns were dropped/removed. We cleaned out data from the data frame for columns that were not consistent with all the chosen energy sources including:
2. **For Natural Gas:**

"Natural Gas Gross Withdrawals,", "NGPL Production", "Natural Gas Net Imports”, "Supplemental Gaseous Fuels", "Natural Gas Balancing Item”.

1. **For Petroleum:**

"Crude Oil Production, Alaska (Thousand Barrels per Day, "Natural Gas Liquids Production”,

" Renewable Fuels and Oxygenate Plant Net Production”,

“Petroleum Processing Gain,", "Petroleum Adjustments"

1. **For Coal:**

"Waste Coal Supplied (Thousand Short Tons)" “Coal Imports”, "Coal Losses and Unaccounted for (Thousand Short Tons)"

‘

1. **For Renewable Energy**: We kept and merged renewable data (including clean coal ) into a data frame.
2. **Other Cleaning**:

* We dropped year 2020 due to lack of consistent data (after August).
* We merged / combined all data by month for the four energy sources reviewed.

1. **Create a data base(s):**
2. Once the clean data frame was set, we used quick DB to create our source table and categories **b**) The data was then saved in a SQL based data frame. We used sql. lite to upload the data into a data frame, which allowed us to use Panda data functions.

**c**) All of the data that we extracted were placed in CSV file, with the column heads and data populated and uploaded.

**d) Entity Relationship Diagram** – We used a ERD to show a snapshot of data structures and visualize the consumption and production relationships between tables within that database. For example, imports, exports, and stock were used with data serving as the primary key as all.

|  |
| --- |
|  |

1. **Loaded Data**:
2. We loaded the final database, tables/collections, showing both production and consumption data: In order to combine the data, we wrote a function to get/combine data for the four energy sectors for the US. (There were no unique IDS used for joining purposes).

drop\_final.to\_csv('file1.csv')

**b**) Final Information was cleaned up and saved in Junypter Notebook (using Python/Pandas)

**c)** Information was copied, pasted, and downloaded into GitHub (see <https://github.com/robinkhan1984/Project-2.git>).