Technical Report

1. **Data & Collection**
   - List all data sources, including URLs, APIs, or repositories, with proper citations and licensing information.
     1. Energy Institute - Statistical Review of World Energy (2025); Smil (2017) – with major processing by Our World in Data. "Primary energy from other renewables" [dataset]. Energy Institute, "Statistical Review of World Energy"; Smil, "Energy Transitions: Global and National Perspectives" [original data].
     2. U.S. Energy Information Administration (2025); Energy Institute - Statistical Review of World Energy (2025); Population based on various sources (2024) – with major processing by Our World in Data
     3. World Population Prospects, United Nations (UN), uri: population.un.org/wpp, publisher: UN Population Division; Statistical databases and publications from national statistical offices, National Statistical Offices, uri: unstats.un.org/home/nso_sites, publisher: National Statistical Offices; Eurostat: Demographic Statistics, Eurostat (ESTAT), uri: ec.europa.eu/eurostat/data/database?node_code=earn_ses_monthly, publisher: Eurostat; Population and Vital Statistics Report (various years), United Nations (UN), uri: unstats.un.org, publisher: UN Statistics Division
   - Describe the data structure (number of records, key variables, data types) and explain why the data was chosen for your topic.
     1. Three datasets were used in this project, merged into 2 datasets total.
        1. Per Country Dataset
           1. 41,655 records
           2. Entity (Country) - object, Code - object, Year - int64, Primary energy consumption per capita - float, Population (total) - int64
        2. World Energy Types Dataset
           1. 1,463 records
           2. Entity - object, Code - object, Year - int64, Other renewables (TWh, substituted energy), Biofuels (TWh, substituted energy), Solar (TWh, substituted energy), Wind (TWh, substituted energy), Hydropower (TWh, substituted energy), Nuclear (TWh, substituted energy), Gas (TWh, substituted energy), Oil (TWh, substituted energy), Coal (TWh, substituted energy), Traditional biomass (TWh, substituted energy)

        ii. These data were chosen for our topic because they give us the best quantitative data over a long period of time that is best for building an analysis.

- Note any known limitations or biases in the dataset and how they may affect interpretation.

        i. One limitation was a lack of per-country data for different energy types, which forced our group to make broad generalizations about worldly awareness and use of different types of energy. It also forced our project to use multiple datasets rather than simply one.

2. **Data Cleaning & Preparation**

- Explain how you cleaned, merged, or transformed the data.
- Include information on handling missing values, filtering, feature selection, and derived variables.

1. **Importing the Datasets-** We used three datasets: per-capita energy consumption (1964-2024), population by country (1960-2024, World Bank), and global energy by type (1800-2024, aggregated). The population dataset contained extra metadata rows, so we used skiprows=4 when loading it:

```
capita_data = pd.read_csv(capita_file_path)

population_data = pd.read_csv(population_file_path, skiprows=4)

types_data = pd.read_csv(types_file_path)
```

2. **Cleaning the Population Dataset-** We removed extra unnamed columns and stripped whitespace from column names. The population data, originally wide format, was reshaped into long format so each row represented a single country-year. Non-numberic years were removed, the year column was converted to integers, and missing values were filled using linear interpolation within each country:

```
population_data.columns = population_data.columns.str.strip()
```

```
print(population_data.columns.tolist())
population_long = population_data.melt(
    id_vars=['Country Name','Country Code','Indicator Name','Indicator Code'],
    var_name='Year',
    value_name='Population'
)
population_long = population_long[population_long['Year'].str.isnumeric()]
population_long['Year'] = population_long['Year'].astype(int)
print(population_long.head())
```

3. **Merging Per-Capita Energy with Population-** The per-capita energy dataset was merged with the cleaned population dataset using country and year as keys. We used an inner join to include only countries with both population and energy data:

```
merged = pd.merge(
    capita_data,
    population_long,
    left_on=['Entity','Year'],
    right_on=['Country Name','Year'],
    how='inner'
)
```

4. **Derived variables-** We created a new column Total_Energy_Demand in the merged dataset using:

```
merged['Total_Energy_Demand'] = merged['Population'] * merged['Primary energy consumption per capita (kWh/person)']
```

This variable calculates the total energy demand of each country in a given year, combining its population with per-capita energy consumption. It allows analysis of absolute energy use rather than only per-capita values.

5. **Handling Missing Values and Filtering-** Removed extra unnamed columns that contained no useful data. Stripped whitespace from column names to ensure consistency. Reshaped the population dataset from wide to long format, so each row represents a single (country, year) observation. Filtered only numeric years and converted them to integers. Merged per-capita energy and population datasets with an inner join, automatically removing rows with missing country or year matches:

```
# Load datasets
capita_data = pd.read_csv(capita_file_path)
population_data = pd.read_csv(population_file_path, skiprows=4)
types_data = pd.read_csv(types_file_path)
# Clean population dataset, Remove extra unnamed columns
population_data.columns = population_data.columns.str.strip()  # Strip whitespace
# Reshape from wide to long format
population_long = population_data.melt(
    id_vars=['Country Name','Country Code','Indicator Name','Indicator Code'],
    var_name='Year',
    value_name='Population'
)
# Keep only numeric years and convert to integers
```

```
population_long = population_long[population_long['Year'].str.isnumeric()]
population_long['Year'] = population_long['Year'].astype(int)
# Merge per-capita energy and population datasets
merged = pd.merge(
    capita_data,
    population_long,
    left_on=['Entity','Year'],
    right_on=['Country Name','Year'],
    how='inner'
)
```

6. **Feature Selection and Standardization-** Selected columns necessary for analysis: country/entity, year, population, and per-capita energy. Year values converted to integers to ensure proper numeric handling, No additional continent mapping or unit conversions were applied beyond what was already in the datasets:

```
population_data.columns = population_data.columns.str.strip()
population_long = population_data.melt(
    id_vars=['Country Name','Country Code','Indicator Name','Indicator Code'],
    var_name='Year',
    value_name='Population'
)
population_long = population_long[population_long['Year'].str.isnumeric()]
population_long['Year'] = population_long['Year'].astype(int)
merged = pd.merge(
    capita_data,
    population_long,
    left_on=['Entity','Year'],
    right_on=['Country Name','Year'],
    how='inner'
)
```

3. **Processing & Methods**
   - Report relevant parameters, tools, and library functions (with version numbers when possible).
     - i. Pandas, Numpy, Matplotlib, Seaborn all used within a Google Colab Notebook
   - **Visual Mapping Choices:** explain how analytical results were translated into visual form — why specific chart types, encodings, or visual structures were selected for the data and task.
     - i. Bar Charts
       1. These were chosen because of the quantitative data surrounding our kWh. Comparing different countries' energy use is easier to see with a bar chart, especially when comparing the top and bottom 10 countries for energy consumption.
     - ii. Line Graphs
       1. These were chosen because of the time series feature in our other dataset. Data ranging from 1800-2024 gives us a clear look into the intense growth of different energy types over 2 centuries.
   - If interactivity or animation is involved, describe the logic behind it (e.g., filtering functions, hover effects, transitions).
     - i. Interactivity in our dashboard is helpful for seeing countries' energy consumption in relation to others. Filtering by a country on one page of the dashboard drills down the map to the country. Another page of the dashboard allows the user to see in real-time when different energy types were introduced with manual filtering.
4. **Validation & Quality Checks**
   - Explain how you verified data accuracy and analysis correctness (e.g., sanity checks, comparison against source statistics, outlier detection).
   - Discuss alternative analyses or views you explored and why you kept or discarded them.
   - Summarize what measures you took to ensure reliability and integrity of results.

Validation & Quality Checks

Sanity Checks:

- Ensured all datasets were properly loaded and columns were correctly interpreted.
- Stripped whitespace from column names to maintain consistency across datasets.
- Filtered population data to include only numeric years and converted them to integers.
- Performed reshaping of the population dataset from wide to long format to ensure one row per country-year.
- Verified that all derived values, such as Total_Energy_Demand (Population × Per-Capita Energy), were calculated correctly for each country-year combination.
- Ensured merged datasets included only rows with matching country and year values, preventing inconsistencies.

Dataset Merging and Integrity:

- Merged the per-capita energy and population datasets using country and year as keys, keeping only rows that had data in both datasets.
- Verified that column names were consistent across datasets to avoid errors during analysis.

Outlier Detection:

- Filtered the population dataset to include only numeric years.
- Removed extra unnamed or invalid columns that contained no useful data.
- Used linear interpolation to fill missing values with each country.

Documentation and Reproducibility:

- All cleaning, merging, and transformation steps were documented in version-controlled notebooks.
- Derived variables and processing steps were clearly recorded to allow reproducibility and verification.

5. **Reproducibility & Environment**
   - ○ Provide clear instructions to reproduce the results (e.g., repository link, required software/libraries, environment setup).
     i. This technical report is placed within our Github Repository, where our datasets, Design Report, dashboard, and other materials for this project are housed. To reproduce our results, do the following:
        1. Download datasets in CSV format, upload to Google Drive, and start a Google Colab notebook
        2. Import libraries (pandas, numpy, matplotlib, seaborn) and sync Google Drive with Google Colab
        3. Preprocessing to combine datasets and reformat them
        4. Build time series charts, bar charts, line graphs using the new datasets
   - ○ If your artifact uses an online tool (e.g., Tableau, Power BI, Observable, Streamlit), explain where and how the data is connected.
     i. Our Power BI data is connected directly to our datasets in CSV format and are uploaded directly to the Power BI cloud service.
6. **Ethics & Transparency**
   - ○ Address any ethical considerations in data use, privacy, or potential misrepresentation of results.
   - ○ Cite all datasets, libraries, templates, and AI tools used. Include a brief transparency statement describing how any AI assistance was incorporated.

Ethics Considerations:

- All datasets in this project are publicly available and were used solely for research, educational, and analytical purposes. No private or sensitive information was accessed or included.
- We ensured that results were represented accurately. Historical and projected data sources, such as UN projections and History Database of the Global Environment data, were clearly identified, and any known limitations or biases were noted in our analysis.
- Visualizations and interpretations were created to reflect trends truthfully, without exaggeration or misleading representations.

Citations of Tools and Data:

Energy Institute - Statistical Review of World Energy (2025); Smil (2017) – with major processing by Our World in Data. “Primary energy from other renewables” [dataset]. Energy Institute, “Statistical Review of World Energy”; Smil, “Energy Transitions: Global and National Perspectives” [original data].

U.S. Energy Information Administration (2025); Energy Institute - Statistical Review of World Energy (2025); Population based on various sources (2024) – with major processing by Our World in Data

“World Population Prospects.” *United Nations*, United Nations, population.un.org/wpp/. Accessed 5 Dec. 2025.

“UNSD - Partners.” *United Nations*, United Nations, unstats.un.org/home/nso_sites/. Accessed 5 Dec. 2025.

“Navigation.” *Database - Eurostat*, ec.europa.eu/eurostat/data/database?node_code=earn_ses_monthly. Accessed 5 Dec. 2025.

“UNSD - Welcome to UNSD.” *United Nations*, United Nations, unstats.un.org/. Accessed 5 Dec. 2025.

The analysis and visualizations in this project were performed using Python libraries, including pandas, numpy, and matplotlib. Power BI was used for creating the interactive dashboards. All libraries are open-source and publicly available.

Transparency Statement:

AI assistance was incorporated solely for improving clarity, grammar, and formatting in written sections of the report. All data processing, cleaning, analysis, and visualizations were conducted manually by the team.