Data Collection and Preprocessing Phase

Date	26 June 2025
Team ID	NA
Project Title	Global Energy Trends: a comprehensive analysis of key regions and generation modes using Power BI
Maximum Marks	10 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	This project uses six datasets that cover global electricity consumption and power generation patterns from 1990 to 2020. Two datasets provide information on how much electricity is consumed by countries and continents, while the other four focus on electricity generation from both renewable and non-renewable sources. The data includes regional and country-level details, mode-wise generation breakdowns (like Hydro, Solar, Wind, and Fossil Fuels), and contributions from top energy-producing nations. Together, these datasets allow for a comprehensive comparison of energy usage versus production, making it possible to explore sustainability trends and regional shifts in energy dependency.
Data Cleaning	During data cleaning in Power BI and Excel, the following steps were taken to improve data quality: **1. Handling Missing Values** Tidal energy value: in the renewable dataset was unrealistically high and likely incorrect. It was replaced with `null` or zero to avoid distortion in totals. - Blank rows and incomplete entries (if any) were removed using Power Query filtering and cleaning tools.

	2. Removing Duplicates - Duplicate entries (especially in country and continent consumption files) were removed using: - Power BI's Remove Duplicates option in Power Query. - Excel's Remove Duplicates feature for early cleaning before import. - Final datasets ensure each record is unique by year, country/region, and energy type. **3. Fixing Incorrect Values** Unit mismatches were corrected: mTOE values from Enerdata were converted to **TWh** using a standard factor (1 mTOE = 11.63 TWh). - The Tidal energy value (19,448.16 TWh), which matched a non-renewables total, was flagged as an error and corrected. - Totals that didn't match their mode-wise sums were recalculated** manually or through DAX in Power BI. **4. Standardizing Format** - Column names like "year" and "region" were made consistent across all files. - Data was reshaped into long format (one row per year-country/mode) to improve filtering and visuals in Power BI. - Unnecessary rows, decimal noise, and extra totals were removed for a neat dataset.
Data Transformation	To prepare the datasets for analysis and visualization in Power BI, I used Power Query Editor for the following transformations: **1. Filtering Data** - Removed unnecessary rows (like extra headers or totals at the bottom of raw files). - Filtered out blank or irrelevant years and entries (e.g., empty TWh rows). **2. Sorting and Arranging** - Sorted data by year and region/country for proper time-series tracking. - Sorted generation modes (Hydro, Solar, etc.) to maintain consistency across visuals. **3. Calculated Columns** - Created new columns using simple DAX or Power Query logic such as: - Total TWh by Year or Region

	- Share of Each Energy Source -YoY Growth Rate (Difference between current and previous year) **4. Merging Queries** - Combined country and continent data from separate files to enable region-wise comparisons. - Merged renewable and non-renewable sources for a full energy generation picture. **5. Column Renaming and Formatting** - Standardized column names like "TWh", "Year", "Country" to match across all tables. - Changed data types (e.g., Year to Whole Number, TWh to Decimal Number) for consistent calculations. **6. Pivoting and Unpivoting** - Used "Unpivot Columns" to convert wide-format data (columns for each year) into long-format tables . - Pivoted source columns where needed to calculate totals or percentage splits.
Data Type Conversion	**1. Converted "Year" Columns** • Original data had years sometimes stored as **text** or **decimal**. • Changed them to **Whole Number** type to enable time-series analysis and sorting. **2. Cleaned "Region" and "Country" Columns** • Ensured these were set to **Text** data type for clear filtering and label visuals. • Removed extra spaces and formatting inconsistencies. **3. Standardized "TWh" Columns** • All electricity generation and consumption values were made **Decimal Number** type. • This allowed accurate totals, averages, and percentage calculations in Power BI. **4. Replaced Errors or Nulls** • Power BI showed type errors (e.g., "Error" in place of numbers) in a few places. • Replaced these with null or fixed the format so the correct data type could be applied. **5. Applied Consistent Types Across Queries** • Made sure similar columns across different tables (like TWh, Year, and Country) had matching data types. • Helped when **merging queries** or creating relationships in the Power BI model.

Column Splitting and Merging	**Column Splitting**
	• In some original datasets, the data was packed into **wide formats** (with one year per column).
	• I used **Power Query → Unpivot Columns** to split these into:
	  ○ One column for Year
	  ○ One column for TWh (value)
	• This step was used in:
	  ○ continent table
	  ○ country table
	• It helped simplify filtering and build proper **time-series visuals** in Power BI.
	Column Merging or Combination
	• I created **Total columns** (e.g., Total TWh) by merging different source columns like Hydro, Solar, Biofuel, etc., using:
	  ○ Power Query's **"Add Column → Custom Column"**
	  ○ Or Excel's **SUM formula** across multiple columns
	• In the **Top 20 Countries** file, individual energy sources were combined to compute a final Total column.
	• In some datasets, I merged **metadata fields** (like Region + Year) into a single column to improve traceability or chart labeling.
	Why This Was Important
	• Splitting helped standardize **long format** for visuals and calculations.
	• Merging made it easier to calculate totals, percentages, and create summary visuals.
	• Both made relationships between tables stronger and visuals more flexible.

Data Modeling	In the original six datasets, there were no relationships between tables — each file stood on its own, with wide-format structures and inconsistent field names. There were no keys like Country, Continent, or Year that could be directly used to connect files. Because of this, it was not possible to filter or analyze data across different tables in a unified way. After cleaning and transforming the data into eight structured files, I created a proper data model in Power BI. I converted all

	datasets into long format, where each row represents a single year, country/region, and its corresponding energy value. This allowed me to define clear relationships: for example, I linked the country consumption table to the top 20 countries generation file using the Country field. Similarly, the continent consumption table was connected to generation totals by Continent. The Year column was unified across all datasets, enabling dynamic filtering using slicers and trend lines. These relationships allowed me to build a star-schema-style model, where central fact tables like generation and consumption were linked through shared dimension fields like Year and Country. On top of this model, I created DAX measures to calculate total generation (in TWh), percentage contributions by source or region, and year-on-year growth. These measures added flexibility, enabling dynamic insights that update instantly when filters are applied in the dashboard. This approach turned the project into a connected, interactive experience rather than just standalone files.
Save Processed Data	Each cleaned dataset was exported from Power BI or restructured in Excel and saved as **.xlsx** files.