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	3 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle	1. Tidal Energy Mismatch: In the renewables dataset, the Tidal energy value (19,448.16 TWh) is unrealistically high and also matches the total value in the non-renewables file. This likely happened due to a copy-paste mistake or a misclassified value. 2. Missing or Expired File Content:A few datasets had to be reuploaded because their content had expired. This delayed analysis and shows the importance of maintaining organized file backups.	Moderate	**1. Tidal energy number is too high**: Replace it with a blank or zero, and double-check the original source. **2. Files lost their content after upload:** Re-uploaded them and kept a backup saved properly. **3. Totals didn't match the individual values:** Recalculated totals yourself instead of trusting the one already there. **4. Years and regions were written in different ways:** Cleaned and renamed columns so everything lines up.

	3. Inconsistent Totals: In the file showing total renewable generation, the total is shown as 6,384.25 TWh, but the sum of individual sources is much higher. This signals a possible calculation error or outdated total. 4. Format Variations:Some files used different formats — for example, year columns were not consistent across all datasets. This required extra cleaning to ensure timelines align correctly. 5. Units Conversion Needed: Consumption data from Enerdata was originally in mTOE (million tonnes of oil equivalent), and had to be converted to TWh to match the generation datasets. 6. No Metadata or Source Columns: Most files didn't mention where the data was sourced from or what units were used. Without clear labels or descriptions, verifying and interpreting the data took more time. 7. Region vs. Country Confusion: Consumption was separated by country in one file and by continent or group (like BRICS or OECD) in another. While this adds richness, it also made it tricky to match both directly in one visual.		**5. Units didn't match:** Converted all mTOE values into TWh to keep it the same. **6. Files didn't say where the data came from:** Added a "source" column or made a note about where each file came from. **7. Mixing country and continent data:** Created groups or mapped countries into continents so they can be analyzed together.
Kaggle	Missing country codes for some regions	moderate	To address missing country codes, use a standard ISO lookup (via Python, Excel, or online mappers) to fill them in.

	Duplicate region entries (like “Asia” and “Asia Pacific”) Inconsistent naming		For duplicate region names, decide on a consistent label and filter or merge using data cleaning tools like Power Query. Clean inconsistent names by applying text functions (like remove or replace) to standardize entities.