Digital Egypt Pioneers Initiative - DEPI

Electricity Group 2 (Eng. Sherihan Ali)

Project Administrators

Abdelrahman Adel Ahmed
Nada Mahmoud Hamed
Ahmad Frhat Mohamad
Hady Mohamed Kamel
Islam Mohamed Sayed
Bassem Amr Mohamed
Aya Mohamed Khamis

Assumptions and Remarks

Date and Time

The time **5:78 PM** in the 2005 sheet was identified as a typographical error. This occurred due to the proximity of the numbers 7 and 4 on the keyboard, leading to an incorrect entry.

Converted the day's events:

- Evening: 6:00 PM
- Noon: 12:00 PM
- Midnight: 12:00 AM

In cases where the date and time of restoration are unknown and the Demand Loss (MW) is recorded as zero, we assumed that the electric current was restored simultaneously. Therefore, the date and time of restoration will be considered the same as the recorded time and date of return.

Data Status

The input is considered faulty if the power outage duration exceeds one hour and the Demand Loss (MW) is recorded as zero.

If the Number of Customers Affected and Demand Loss (MW), the event start date, and the restoration date are NULLs, the data is considered anonymous.

If the Number of Customers Affected, Demand Loss (MW), Duration Time, and NERC Region are NULLs, the data is considered Full Data.

Otherwise, It is considered Fair Data.

Demand Loss (MW)

We assumed that **none** is equivalent to **zero** during data processing.

For (PG&E) values: Assumptions used in calculating the impact of power outages:

1. Average household consumption: Assumes each household consumes 1.2 kWh per hour according to official data of NERC.

2. Outage duration: Calculated as the difference (in hours) between the Time Event Began and the Time of Restoration.
3. Number of individuals per household: Assumed to be 3 individuals per household.
4. Equal energy distribution: The lost energy is assumed to be distributed equally among the affected households.

We used AI to help us get the Demand Loss value for these two records:

1. 133 on 5/21/04 between 3:00 a.m. and 4:00 a.m., 392 on 5/21/04 between 4:00 p.m. and 5:00 p.m.
2. 177 on 5/21/04 between 3:00 p.m. and 5:00 p.m.

Using the following assumptions that it gave to us:

- Energy loss during the specified time
  total_loss = loss_3_to_4_am + loss_4_to_5_pm
- Assume the remaining hours have a similar loss (average of the known periods)
  remaining_hours = duration_hours – 2
  - ➢ Subtract the two known periods
    average_loss = total_loss / 2
  - ➢ The average loss of the two known periods
  - ➢ Total energy lost total_energy_lost = total_loss + (remaining_hours * average_loss)

So the results are:

1st record's total_energy_lost = 18,375 MW
2nd record's total_energy_lost = 6,195 MW

If there was a peak value and an accumulated value, we took the peak value.

Cleaning Steps

We divided the sheet into groups: **2002-2010**, **2011-2014**, **2015-2022**, and **2023**, ensuring that each group maintains the same structure. After cleaning the date and time columns, we will append these groups into a single query. Subsequent cleaning steps will then be applied.

We added this record manually to the 2003 sheet as it was grouped to another record and deleted while cleaning
27| 2/5/2003 SERC 12,897 (Alabama) 8:00 p.m. Alabama Severe Thunderstorms 130 12,897 (Alabama) 5/03/03, 8:00 a.m.

We didn't delete any records except the Blanks Rows and the Duplicated ones which are 15 records so the final total records is **3936** after it was **3951**.

Basic Cleaning Steps for Each Group

1. Remove Blank Rows
2. Remove Top Rows
3. Promote Headers

We utilized functions and custom columns to optimize performance and reduce file load during cleaning. We focused on applying **dynamic** changes rather than manual adjustments to streamline the workflow.

Now, let's examine what has occurred in each column of the dataset. We will focus on identifying trends, inconsistencies, and any necessary adjustments made during the cleaning process.

Date and Time

Replaced wrong data like

- 7/01//05 in the 2005 sheet with 1/7/2005
- Text values like "Unknown", "Ongoing", "NA", and "(Trans. Only)" with NULL
- Wrong input data like the year 2024 in the Date of Restoration in the beginning data 2006-2010
- 18/3/2001 and 29/8/2077 in the 2011 sheet with 18/3/2011

Area Affected

When we placed the **Area** column on a map visualization, we observed that some entries, such as "Eastern Montana" and "Vallee, California," were incorrectly located

outside the USA, specifically in Asia and Australia. As a result, we corrected these entries to ensure accurate geographic representation.

NERC Region

The **North American Electric Reliability Corporation (NERC)** divides the U.S. and parts of Canada into six major regional entities, each responsible for overseeing the reliability of the power grid in their area.

Below are the main NERC regions along with their abbreviations and the territories they cover:

1. **Midwest Reliability Organization (MRO)**: Covers parts of the U.S. Midwest and Canada, including Minnesota, Wisconsin, Iowa, and Manitoba.
2. **Northeast Power Coordinating Council (NPCC)**: Includes New York, New England, Ontario, Quebec, and the Maritime provinces of Canada.
3. **ReliabilityFirst Corporation (RFC)**: Covers the Great Lakes region, including parts of Ohio, Pennsylvania, Maryland, New Jersey, and Virginia.
4. **Southeastern Electric Reliability Council (SERC)**: Covers the southeastern U.S., including the Carolinas, Georgia, and parts of Mississippi, Alabama, and Tennessee.
5. **Texas Reliability Entity (TRE)**: Focuses on the state of Texas, which operates mostly independently from the national grid.
6. **Western Electricity Coordinating Council (WECC)**: Covers the western U.S., including California, Arizona, Nevada, and portions of Canada and Mexico.

These two regions aren't set under the control of NERC but they were in the data:

7. **Hawaii (HI)** has its own utility and grid operators, like Hawaiian Electric Company (HECO), which are not part of the continental NERC regions
8. **Puerto Rico (PREPA)** has its own utility and grid operators and doesn't fall within the main NERC regions.

Event Type

Multiple event types could be grouped under broader categories. After classification, the final categories are:

- **Natural Disaster**
- **Fire**
- **Vandalism**
- **Shedding Load**
- **Operational Malfunction**

Natural Disaster Type

We observed that "**Natural Disaster**" was the most common category. To gain more insights, we added this new column to classify specific types of natural disasters, such as winds, floods, and earthquakes.

Demand Loss (MW)

We replaced wrong data like

- Text values like "Unknown", "All", and "NA" with NULL

There were numbers separated by **"to"** and **"-"** so we took the average, like "65 to 100" and "8000-10000"

Number of Customers Affected

We replaced wrong data like

- Text values like "Unknown", "utilities", "industrial", and "Interruptible" with NULL