Digital Egypt Pioneers

Cleaning Steps

U.S. ELECTRIC GRID OUTAGE ANALYSIS

DEPI-GIZ1_DAT2_G1e
-Group1-

Presented to: Eng. Sherihan Ali

30/09/2024

Project Cleaning Steps: -

General Cleaning Steps for all Years 2002-2023 (each separated query)

1. Remove Rows:

- Remove Top rows to promote headers, empty Rows, and any unnecessary “null” valued or out of context bottom rows.

2. Promote Headers:

- Promote the proper rows to headers.

3. Add Custom Columns:

- Extract the month name and year from the date column.

4. Filter Columns & Replace Values:

- Filter null values in key columns like NERC Region, Date, and Time Event Began.

5. Split Columns:

- Split the Restoration column by delimiter (space or period), creating separate columns for Restoration Time and Restoration Date based on delimiters like space or period.

6. Change Data Types:

- Update the data types for columns like Time Event Began, Time of Restoration, and Date Event Began to their appropriate types.

7. Rename Columns:

- Rename columns for clarity and consistency.

8. Trim Text Columns & Remove Empty Columns:

- Trim extra spaces from Text columns like (e.g., Restoration or Type of Disturbance).
- Remove columns doesn’t needed

7. Replace Values:

- Replace errors at Time of Restoration ad Date of Restoration columns to **null.**

Project Cleaning Steps: -

Unique Cleaning Steps on each year

- **2002**:
  - Replace “**Evening**” in the Time column with **“6:00:00 PM”.**
  - Replace “**Noon**” in the Date of Restoration column with “**11/10/2002**”.
  - Replace “**November 10**” in the Time column with “**12:00:00 PM**”.
- **2003:**
  - Transpose
  - Merge columns To solve the Date of Restoration (merged in Excel) problem.
  - Replace “**Approximately** “ and “**Approximately “** in Restoration column to “”.
  - Replace “**14/8/2003**“ in Date Event Began column to “**8/14/2003**”.
- **2004:**
  - Remove duplicates.
- **2005:**
  - Replace **“,”** in Restoration column to **“, ”**.
  - Replace **“//”** in Date Event Began column to **“/”.**
  - Replace errors in Event Month column to **July**.
  - Replace errors in Date of Restoration column to **null**.
  - Replace “**5:78 p.m**” in Time Event Began column to “**5:48 p.m**”.
- **2006:**
  - Add Custom column to add (**2006**) to the Date of Restore column.
  - Replace “**1/6/2006**” in Date of Restoration column to “**1/6/2007**”
- **2007**:
  - Add Custom column to add (**2007**) to the Date of Restore column.

Project Cleaning Steps: -

- **2008:**
  - Replace “**1/1/2008**” in Date of Restoration column to “**1/1/2009**”.
  - Add Custom column to add (**2008**) to the Date of Restore column.
  - Replace “**.**” in Time of Restoration column to “**.m**”
  - Replace “**.** ” in Date of Restoration column to “”
- **2009:**
  - Add Custom column to add (**2009**) to the Date of Restore column.
- **2010:**
  - Replace “**1/12/2010**” in Date of Restoration column to “**1/12/2011**”.
  - Add Custom column to add (**2010**) to the Date of Restore column.
- **2011:**
  - Replace “**3/18/2001**” in Date of Restoration column to “**3/18/2011**”.
  - Replace “**8/29/2077**” in Date of Restoration column to “**8/29/2011**”.
- **2015:**
  - Replace “**null**” at NERC Region column to **“Not_NERC”.**
- **2016:**
  - Replace “**null**” at NERC Region column to **“Not_NERC”.**
- **2019:**
  - Replace “**8/18/2018**” in Date of Restoration column to “**8/18/2019**”.
- **2023:**
  - Replace “Unknown” in Number of Customers Affected and Demand Loss (MW) columns to “null”.
  - Extract the Time from Time Event Began column.

Project Cleaning Steps: -

1. Change Data Types
2. Use “**IF else**” function to clean the following columns “**Time Event Began - Time of Restoration** - **Date of Restoration-Demand Loss (MW)**- **Number of Customers Affected**- **NERC Region**).
3. Categorize event types.
4. **Extract** Text before delimiter (-) in **Demand Loss (MW)** column to remove (-, --)
5. **Trim** Columns Like (Demand Loss (MW) column-Number of Customers Affected column) to remove text within those columns, then change the column type to Decimal Number for the demand loss & Whole Number for the Customers affected.
6. **Categorize Alert Criteria**
7. **Add** new custom column to calculate the **Duration of the event**, and checking the column type for **“Duration”.**
8. **Add** new custom column to Display the duration in “**hours”**
9. **Filter** Total Event Hours column to remove negative values
10. **Trim** NERC Region column
11. **Add** custom column to clean the NERC Region column according to the latest updates in NERC Regions
12. **Remove unused** columns old NERC Region, Event type and alert criteria.