**Data Cleaning and Preparation**

This dataset required extensive cleaning and it involved a series of steps.

**Note**: This documentation is for my reference and also will help me optimize the steps in future. I have used Power Query for cleaning, this time. The initial steps, I was doing certain things manually, as in clicking on few values and replacing them individually, but once I got the hang of it, I was able to make certain changes from the Advanced editor, which reduced the time taken to clean. This is the Major learning here!

**The Cleaned Dataset:**

The data with the null date events, and empty rows were removed. I assumed and mapped the data for the null NERC Regions by Areas Affected column and hence they are taken into consideration. The Counties and Cities are considered at a State Level. All the fields in the sheet for 2002-2022 are prepped to be consistent with the 2023 sheet. The NERC regions, Alert Criteria, Events are grouped into broader categories and the Areas affected column is cleaned and updated with the respective State Names. The date fields are cleaned, and a standard time format is achieved.

**Assumptions made (on a higher level, more granular details, in the later section):**

**NERC Region**: The NERC Regions are segregated, and the typo or half written Regions are identified (based on the Areas affected and checking them online) and they are replaced accordingly. If a value has got 2 or more regions, they are split accordingly. **Note**: Few of the NERC Region mentioned were either dissolved or merged but I have kept them not to lose the essence of information, if I find anything interesting with knowing about the previous existent region. I do not want to assume them with any other or completely ignore them as Unknown / Other, as the other details along with them may be affected. Hence, I chose to keep them.

* **Alert Criteria**: There were many incomplete Criteria listed, they are segregated, and I got a clean list of alert criteria. Then I ensured they map with the details in the PDF provided. Also, I have categorized them into broad categories like (Alert Reporting System) - **Emergency alert, Normal Report, System report** based on the reporting time as mentioned in the PDF.
* **Event Type**: The Event Type are grouped into broader categories like - **Cybersecurity Issues**, **Environmental factors, Operational failures, Public appeal**, **Security Incidents, System Operations, Others / Unknown**.
* **Areas Affected:** If the values contain multiple states, they are split and kept, if they are split into cities / counties they are considered only at a state level and ignored at county or city level. The split ups have unique ID and hence not considered repetitive. The data is standardized / Normalized with the correct State list. Also non US states were ignored but PR was considered. This is again the challenging part where I took a lot of time deciding how to proceed.

**The Detailed Process:**

**Splitting into groups:**

The cleaning is completely done using Power Query. The data was presented for each year in a separate sheet. Out of this, 2002-2010 data had 8 columns (but this was split into 2 groups for easy cleaning up of restoration dates as there were 2 major format variations), 2011-2014 had 9 columns, 2015-2022 had 11 columns and 2023 had 12 columns. **First** **task** was to **group** them as per the number of columns. Then worked on each group.

The common tasks done across all the files were:

1. Remove blank row.
2. Made the first row as header.
3. Changed the column names to be consistent throughout.

This was then grouped as mentioned earlier. So, I had 5 groups in total. I started working on each group and did the following across each group. Once they are done, they are then grouped into a single table.

**Date Event Began:**

The date format was checked to ensure it is as per the Date type. There were no major cleaning required.

**Assumptions / Note:**

1. The null values are removed.

**Time Event Began:**

The time were displayed in different formats like a.m, p.m, midnight, evening, noon, etc. They were changed to be in standard format.

1. Assumptions / Note:
2. The a.m and p.m were replaced with AM and PM.
3. The midnight, noon, evening was considered as 12:00AM, PM, and 6:00PM respectively.
4. If it is “Approximately 5 PM”, it was considered as 5 PM.
5. N/A is considered null

**Date and Time of Restoration:**

For few sheets, both the date and time were in the same column. First Split them into 2, then worked on dates.

**My Experience:** I will share the experience first, I initially worked on the 2002-2010 data together and tried several ways to format the restoration time. They were displayed in many different formats. Few things were ok to clean but then when the time was broadly divided into 2 categories like time, date and date, time. It was difficult to split and work on. Maybe there would be a proper or easy way to handle this in Advanced M code or other options, but as ‘m not well versed in M coding and new., it was difficult to split them, when they are grouped together. I then identified that these formats each are same across 2002-2005 and 2006-2010 files. Hence, I did go back and split the data further into 2 groups and worked on the same. This required re-work but worth it and a learning to be documented. Ok, on to the process!

**Assumptions / Note:**

1. The N/A, null, Ongoing, approximately and Unknown were changed to null.
2. Few of the data that has misplaced data, as in if there is date mentioned in time or region mentioned in time field, such data was removed.
3. Then there was mismatch with formats dd/mm/yyyy and mm/dd/yyyy, this couldn’t be achieved as there was some issues with the format of number of digit mentioned for month. Hence used Custom function to handle it and parse date.

For the restoration time, the values such as Noon, Midnight, Noon, Evening, were replaced as mentioned earlier in the time event began section.

As mentioned above, once the restoration date and time were worked on, all the groups were appended as a single table. Then I worked on all the remaining 6 columns.

**NERC Region:**

Few of the NERC Region mentioned were either dissolved or merged but I have kept them not to lose the essence of information, if I find anything interesting with knowing about the previous existent region.

**Assumptions / Note:**

1. Few of the Regions has typo, few were incorrect, hence made changes.
2. Some of the rows had multiple regions, have split them, and had a unique identifier. The splitting was done using a custom function.

**Demand Loss:**

There were some replacements to do,

**Assumptions / Note:**

1. The NA, None, All, were changed to 0.
2. The Unknown was changed to null.
3. Peak, greater than etc was removed.
4. If there is a value that says greater than 500, it was considered 500.
5. If there is an approximation, it was considered the same value Ex: Approx 300 as 300.
6. For ranges like 500-800, average was taken.

**Number of Customer Affected:**

The changes were similar to the ones mentioned for Demand loss. Apart from that, I had made few other changes that includes,

1. The words like Cumulative, utilities, million, etc was removed.
2. The unknown was changed to null.

**Alert Criteria:**

The Alert criteria had different values (half entered, etc). They were first grouped into the 24 items listed in the provided PDF.

Then they were further divided into 3 broad categories depending on the reporting - Emergency Aler (which required filing within an hour), Normal report(filing within 6 hours) and System report (Filing within 1 business day).

**Event Type:**

There were numerous Event Types.

**Assumptions / Note:**

I have split them into the following categories,

1. Environmental Factors (Includes weather, natural disaster, etc)
2. Security Incidents (Includes vandalism, physical attack, threat, sabotage, etc)
3. Cybersecurity Issues (Cyber related)
4. Operational Failures (Grid, Generator, Transmission, failure at high voltage, etc)
5. Public Appeal
6. System Operations (System Operations, Interruption of electrical systems etc)
7. Others / Unknown.

**Areas Affected:**

This is again the challenging part where I took a lot of time deciding how to proceed, what to keep for analysis and how granular I’m going to represent the data.

**Assumptions / Note:**

The Areas were split into separate rows where there were multiple States, but the granularity with regards to City / County was not taken into account and the state level was considered. Only US states were considered and others were ignored. Then I used conditional column to have consistent naming of the 52 States.

**Extra Fields for Analysis:**

1. Event Year
2. Event Month
3. Alert Reporting System
4. Event Type Grouping
5. Index
6. Quarter

**Conclusion:**

This is how I approached the data cleaning process, I have spent countless hours working on restoration date and time as well as NERC and Areas affected on deciding how to split values, what to consider for analysis, etc. There might for sure be a way to get through these challenges a bit more easier, but with the skillset I have, I decided to take up this challenge and survive! With further projects, I’m sure I will find ways to optimize the process!