Visualizing US Natural Disaster Declarations — Milestone 1 (Part-1)

Data Acquisition, Cleaning & Initial Preparation

Overview

This milestone focuses on acquiring the FEMA Disaster Declarations dataset, cleaning it, and preparing it for further exploratory data analysis. The main objective of this phase is to ensure that the dataset is consistent, reliable, and analysis-ready before moving into visualizations and trend analysis in later milestones.

The dataset contains disaster declaration records across U.S. states, including information such as disaster type, dates, program declarations, and affected regions.

Objectives of Milestone-1 (Part-1)

The following goals were completed in this phase:

- Load FEMA Disaster Declarations dataset into Pandas
- Inspect dataset structure, datatypes, and missing values
- Convert date fields to proper datetime format
- Handle missing values and duplicate records
- Standardize and normalize categorical fields (especially incidentType)
- Save the cleaned dataset for future analysis

These steps ensure data quality, consistency, and proper preprocessing before Exploratory Data Analysis.

Dataset Description

- Dataset Name: **FEMA Web Disaster Declarations**
- Format: **CSV file** loaded into Pandas dataframe
- Number of records before cleaning: **5139 rows and 26 columns**

The dataset contains real-world FEMA disaster declaration history across multiple years.

Data Loading & Initial Inspection

The dataset was imported using Pandas and previewed with head() to verify structure.

The dataframe summary (df.info()) confirmed:

- 26 total columns
- Mix of numeric and categorical fields
- Several date columns stored as string objects
- Some nullable fields (program flags, shapefile links, closeout dates)

Handling Missing Values

A missing value audit was performed using:

df.isna().sum()

This identified:

- Missing values in incidentEndDate, closeoutDate
- Missing program declaration flags
- Many nulls in geospatial URL fields

These were handled as follows:

❖ **Removed duplicate records**

df = df.drop_duplicates()

❖ **Dropped rows missing critical identifiers**

Critical analysis-dependent columns:

- disasterNumber
- state
- incidentType

❖ **Filled program declaration flags with 0 (not declared)**

Fields updated:

- ihProgramDeclared
- paProgramDeclared
- hmProgramDeclared

These columns were converted to **integer type** for consistency.

Date Standardization

The following date columns were converted to datetime format:

- declarationDate
- incidentBeginDate
- incidentEndDate

Using ***pd.to_datetime()*** with coercion mode to safely handle invalid formats.

This ensures accurate:

- time-series aggregation
- temporal analysis
- disaster timeline insights

Sample converted output confirmed correct transformation.

Standardizing *incidentType*

The ***incidentType*** field originally contained inconsistent text such as:

- Mixed casing (Fire vs FIRE)
- Variants of same category (Flood / FLOODING)
- Combined disaster type labels

To fix this:

- **Text normalized to uppercase**
- **Similar disaster types grouped using mapping**

Examples:

- HURRICANE (TYPHOON) → HURRICANE
- SEVERE STORM(S) → SEVERE STORM
- FLOODING → FLOOD

After cleaning, top categories included:

- FIRE
- SEVERE STORM
- FLOOD
- HURRICANE
- TORNADO

This improves category consistency and analysis accuracy.

Saving the Cleaned Dataset

The cleaned and standardized dataset was exported for later milestones:

df.to_csv("FEMA_cleaned_milestone1.csv", index=False)

This ensures reproducibility and separates raw vs processed data.

Conclusion

In Milestone-1 (Part-1), the FEMA Disaster Declarations dataset was successfully acquired, cleaned, and prepared for further analysis. The data preprocessing steps included inspecting the dataset, handling missing values, removing duplicate records, converting date fields to proper datetime format, and standardizing categorical fields such as incidentType. These steps improved the consistency, quality, and reliability of the dataset, ensuring that it is analysis-ready.

The cleaned dataset is now suitable for conducting Exploratory Data Analysis in the next phase, where trends, patterns, and disaster declaration insights will be examined.