Power BI Data Cleaning & Optimization Guide

This document outlines comprehensive data cleaning and optimization procedures performed on FEMA disaster declaration data using Power BI and Power Query Editor. The approach emphasizes performance optimization, data quality, and professional dashboard standards.

Step 1: Data Type Checks

Purpose

Ensuring correct data types is the foundation of clean data. Incorrect data types can lead to calculation errors, visual display issues, and failed relationships between tables.

Actions to Take

1.1 Review Column Data Types

- Open Power Query Editor (Home → Transform Data)
- Review the data type icon in each column header
- Identify common data types:
    - **Text (ABC):** Names, descriptions, IDs that shouldn't be calculated
    - **Whole Number (123):** Integer values, counts
    - **Decimal Number (1.2):** Prices, percentages, measurements
    - **Date:** Date values without time
    - **Date/Time:** Combined date and time values
    - **True/False:** Boolean values
    - **Currency:** Monetary values with proper formatting

1.2 Identify and Fix Data Type Mismatches

Check for common issues:

- **Dates stored as text:** Look for columns with date-like values showing ABC icon
- **Numbers stored as text:** Numeric columns that show ABC icon
- **IDs incorrectly set as numbers:** Customer IDs, Product codes should be text
- **Mixed data types:** Columns showing "Any" or errors due to inconsistent values

1.3 Change Data Types

- Right-click column header → Change Type → Select appropriate type
- Or use Transform tab → Data Type dropdown
- **Important:** Always review the "Changed Type" step in Applied Steps

1.4 Handle Conversion Errors

When changing types, watch for:

- **Error values:** Indicate data that couldn't convert (e.g., "N/A" in a number column)
- **Null values:** Check if nulls appeared after type conversion
- **Data truncation:** Decimal places lost when converting to whole numbers

Validation Checklist

- ☐ All date columns use Date or Date/Time type
- ☐ Numeric calculations use Decimal Number or Whole Number
- ☐ ID fields are set as Text (to preserve leading zeros)
- ☐ Currency columns use Currency type
- ☐ No "Any" data type columns remain
- ☐ Error values from type conversion are addressed
- ☐ Boolean/flag columns use True/False type

Data Quality Issues Identified

During the data exploration phase, the following issues were identified in the FEMA disaster declaration dataset:

1. Presence of "Not Specified" Values

The *designatedIncidentTypes* column contained a large number of "Not Specified" values, which impacted the quality of categorical analysis and aggregations.

2. Multiple Incident Types in Single Cell

Some records contained multiple incident type codes separated by commas, creating inconsistency in categorical analysis. Examples include:

- `2,M,W,F`
- `5,W,F,T`
- `R,Z`
- `2,5,M`

3. Mixed Coding Format

The incident type codes required mapping to meaningful disaster names. The following codes were identified:

Code	Disaster Type
`R`	Fire
`W`	Storm
`F`	Flood
`H`	Hurricane
`T`	Tornado
`Z`	Snow
`M`	Mudslide
`2`	Biological
`5`	Other

Cleaning Steps Performed in Power Query

Step 1: Replaced "Not Specified" with Null

Action:
- Used Replace Values in Power Query
- Converted "Not Specified" → null

Reason:
- Treated as missing data
- Improved aggregation accuracy

Step 2: Avoided Splitting Rows (Performance Optimization)

Observation: Splitting by comma significantly increased row count and file size.
Decision: Did NOT split into rows. Kept original row count: **69,089**
Reason:
- Maintain optimal model performance
- Prevent row explosion
- Reduce PBIX file size

Step 3: Created Boolean Flag Columns

Instead of splitting rows, created indicator columns using Power Query's `Text.Contains()` function:

```
Text.Contains([designatedIncidentTypes], "F")
```

New Columns Created:

Column Name	Meaning
`Is_Fire`	Fire (R)
`Is_Storm`	Storm (W)
`Is_Flood`	Flood (F)
`Is_Hurricane`	Hurricane (H)
`Is_Tornado`	Tornado (T)
`Is_Snow`	Snow (Z)
`Is_Mudslide`	Mudslide (M)
`Is_Biological`	Biological (2)
`Is_Other`	Other (5)

Values:

- **1** → Incident Type Present
- **0** → Not Present

Benefits:

- No increase in row count
- Faster dashboard performance
- Easy aggregation using SUM()

Step 4: Date Formatting

- Converted `declarationDate` to Date type
- Extracted Year and Month for trend analysis

Step 5: Removed Unnecessary Columns

Removed:

- Columns not used in visualization
- High-cardinality technical columns
- Unnecessary metadata fields

Reason:

- Reduce model size
- Improve performance
- Clean schema

Data Model Optimization

The following optimization strategies were implemented to ensure maximum performance:

- **Maintained original row count:** 69,089 rows
- Avoided row duplication
- Used numeric flag columns instead of text splitting
- Created measures instead of calculated columns where possible

Results

- Smaller file size
- Faster refresh times
- Optimized Power BI model
- Professional dashboard structure

Final Outcome

The dataset is now clean, structured, optimized, and ready for comprehensive analysis.

Prepared for Analysis

- Incident Type Distribution
- Year-wise Disaster Trends
- State-wise Disaster Analysis
- Disaster Category Comparison

Tools Used

- Power BI Desktop
- Power Query Editor
- FEMA Open Data API

Conclusion

The dataset was successfully cleaned and optimized while maintaining performance efficiency. Advanced Power BI techniques were employed to ensure professional dashboard standards, including the strategic use of boolean flag columns to avoid row explosion, proper data type assignment, and elimination of redundant columns. The resulting data model is ready for high-performance analytical visualization and reporting.