



## CLEANED DATASET REPORT- WEEK 2 REPORT

### Executive Summary

During Week 2, the STC Finance Industry team conducted a comprehensive data cleaning session using Power BI to prepare the Bank of America complaint dataset for analysis. The focus was on ensuring data accuracy, consistency, and readiness for visualization. The team collaboratively standardized column names, validated data types, checked for duplicates, and transformed state abbreviations into full names to enhance readability. Missing and blank values were carefully handled using Power Query tools such as Fill Down and “Not specified” placeholders to maintain data integrity. As a result, the dataset is now clean, well-structured, and ready for analytical exploration in Power BI.

#### 1. Tool Used:

PowerBI

#### 2. What issues did you find in the dataset?

During the team's data cleaning session, we identified the following issues:

- Inconsistent column names (e.g., “Submitted via” not standardized).
- Missing values in key fields such as *Sub-Product* and *Sub-Issue*.
- Abbreviated entries in the *State* column (e.g., TX, CA).

- Blank cells in multiple columns that lacked sufficient information to fill accurately.

### 3. What specific steps did you take to clean the data?

- **Checked column headers** in Power Query to ensure they aligned correctly with the data in their respective fields.
- **Renamed columns** for clarity and standardization (e.g., “Submitted via” to “Complaint\_Channel”) using the *Transform - Rename Columns* feature.
- **Validated data types** to confirm that all fields were in the correct format (e.g., date columns converted to *Date* type, text fields to *Text* type).
- **Checked for duplicates** using *Remove Duplicates* in Power Query - no duplicate records were found.
- **Transformed State abbreviations** into full state names (e.g., “TX” to “Texas”, “CA” to “California”) for better readability in visualizations.
- **Handled missing values:**
  - Used the *Fill Down* feature in Power Query to address blank cells in the *Sub-Product* column.
  - Replaced other blank or missing cells with “**Not specified**” where no logical replacement value could be determined, ensuring consistency across the dataset
  - Created new columns

## Data Cleaning & Feature Engineering

During the data preparation phase, the dataset was **cleaned, standardized, and enhanced** to support in-depth analysis of consumer complaints. This included handling missing values, correcting inconsistencies, and creating new calculated columns using **DAX functions in Power BI** to enable temporal and performance-based insights.

### New Columns Created:

**Timely\_response:** Indicates whether the company responded within the expected time frame (Yes/No).

**Positive\_Resolution\_Flag:** Flag to identify complaints resolved in a satisfactory way for the consumer.

**Timely\_Flag:** Binary indicator of timely resolution (1 = timely, 0 = delayed).

Received\_Days: Number of days between complaint submission and the company's response.

Year: Year the complaint was submitted, extracted from the date field.

Month: Month the complaint was submitted, extracted from the date field.

## **5. Is the dataset now clean and ready for analysis?**

**Yes.**

The dataset has been standardized, missing values have been handled appropriately, and column headers are now consistent and descriptive.

## **6. Additional Notes (Optional):**

- The data cleaning meeting was held on **October 20, 2025, at 7:00 PM** with all mentees of the STC Finance Industry team present.
- The main challenge faced was deciding how to handle blank fields where no logical value could be inferred; the team agreed on using “Not specified” to maintain data integrity. Also if there were any columns that would not be required for our analysis.