Data Processing Documentation: Crime Motives Dashboard

Overview

This document explains the approach, logic, and challenges encountered during the preprocessing of monthly crime data used for the Crime Motives Dashboard. The raw data provided was highly inconsistent, with various formatting issues, ambiguous column names, and inconsistent categorization. The goal was to create a unified, analyzable dataset from unstructured CSV files representing monthly crime reports.

Key Challenges

1. **Inconsistent Column Names**\ Each monthly CSV file had slightly different column names due to formatting errors, manual entry issues, or naming convention inconsistencies (e.g., "S. No", "SLNO", "Sl. No.", etc.). Some files even contained unnamed columns or whitespace-filled headers.

2. **Non-standard Formatting**\ Files included irregular spacing, hyphens, periods, or mixed casing. For example:

3. `Heads of Crime`, `heads_of_crime`, `HEADS-OF-CRIME`, etc.

4. Column names with trailing or leading whitespace.

5. **Missing or Misplaced Data**\ Some files had missing key columns, or columns appeared in unexpected positions. In certain cases, essential fields like `current_year_upto` were labeled differently such as "month under review" or "current year (upto)".

6. **Ambiguous Motive Classifications**\ The dataset included a vast list of `minor_head` values which represent specific motives. These needed to be grouped into broader, generalized categories (`super_motive_category`) for dashboard-level filtering.

7. **Encoding Errors**\ Some CSVs could not be read with UTF-8 encoding due to special characters, requiring fallback to Latin-1.

Processing Strategy

1. **Dynamic Column Renaming with Heuristics**\ Used custom logic to scan each column name and map it to a standardized set using pattern matching. For example:

2. Columns containing both `sl` and `no` → mapped to `sl_no`

3. Columns referring to crime descriptions → mapped to `heads_of_crime`

4. Applied consistent casing, removed extra spaces, punctuation, and replaced symbols with underscores.

5. **Column Fallback Filling**\ Ensured all dataframes had a consistent set of columns by adding missing ones with `pd.NA` if not found. This ensured compatibility for concatenation.

6. **Standard Column Set**\ All files were transformed to conform to:

```
standard_columns = [
    "sl_no", "heads_of_crime", "major_head", "minor_head",
    "current_year_upto", "prev_year_same_month",
    "prev_month", "current_month"
]
```

1. **Motive Grouping Logic**\ Built a mapping from `minor_head` to `motive` and from `motive` to `super_motive_category`. This mapping grouped specific crimes into logical higher-level categories like:

2. `Sexual Motives` (e.g., `rape`, `molestation`, `sexual_harassment`)

3. `Financial Motives` (e.g., `robbery`, `cheating`, `embezzlement`)

4. `Domestic/Relationship Motives`, `Property Disputes`, etc.

5. **Month Tagging**\ Each dataframe was tagged with the month it represented by extracting the file name, enabling trend tracking over time.

6. **Concatenation & Export**\ All cleaned monthly dataframes were combined using `pd.concat()` into a master file: `all_months_concatenated.csv`.

Dashboard Application (dashboard.py)

To make this data explorable and insightful for analysts and stakeholders, a Streamlit dashboard was built that enables intuitive filtering and visual exploration:

1. **Filter by Super Motive Category**\ Users can filter the dataset by high-level motive categories like `Financial`, `Sexual`, `Domestic/Relationship`, `Property Disputes`, etc.

2. **Nested Display of Specific Motives**\ Under each super-category, the app displays stats for the specific motives (e.g., within "Financial", show `theft`, `fraud`, `cheating`, etc.).

3. **Aggregated Statistics**\ Bar plots and value counts were shown for the distribution of cases across different motives and categories.

4. **Time Trends**\ Optionally, month-wise trends are plotted to show how particular motives evolve over time.

Final Output

- **Individual Cleaned Files**: Saved into `/cleaned_csvs/`
- **Final Aggregated File**: `all_months_concatenated.csv`
- **Enhanced File for Dashboard**: `enhanced_crime_data.csv` (includes motive and category mappings)
- **Streamlit App Script**: `dashboard.py`

Reflection

This preprocessing workflow wasn't straightforward data cleaning-it involved robust logic for ambiguous column standardization, handling missing or corrupted data, dynamically restructuring dataframes, and carefully designing a mapping hierarchy that would power intuitive filtering on the dashboard. The success of the visualization layer relied heavily on getting this foundational processing right.

Building the dashboard added an extra layer of complexity and creativity, ensuring that real-world, messy data could be made actionable through interactive visuals. The work highlights a strong problem-solving ability, end-to-end ownership, and thoughtful design in both data wrangling and interface development.