Ydata Profiling Report

Data Profiling Documentation: **Austin Collision Dataset**

Introduction:

This document synthesizes the findings from the data profiling of the Austin Collision dataset, conducted using the YData profiling tool. It highlights critical areas related to data quality, completeness, and structure, offering specific insights into individual columns.

Standardization Solutions

For the 'Date and Time' column, convert the string representations to DATETIME format by using the **transformations in tmap**. This conversion facilitates time-series analysis and chronological sorting, enhancing the consistency of date and time data across the dataset.

Data Structure Adjustments

For ***multi-valued attributes*****,** such as a column listing Units Involved in a single record, we introduced a normalization process. Create a new table where each row represents a single unit linked to the original collision record. This approach simplifies the analysis of violations by ensuring each attribute occupies its own row.

Null Values:

Null values were prevalent across several columns.
Action: All the Null values have been mapped to one key which we have created in Location Dim.

Combined Columns:

1. Motor Vehicles Death & Injury Count and Motorcycles Death & Injury Count : We have merged the columns to obtain a combined count of injuries and deaths for the respective categories.

2. Contributing Factor: We have combined the columns 'contrib_factor_p1' and 'contrib_factor_p2' into a single column named 'contributing factor' to consolidate the data.

Austin Collision Dataset Column Specifications:

1. Collision ID
Description: Unique identifier for each collision incident.
Data Type: Integer.
Issues: None expected.
Recommendations: Verify uniqueness.

2. Date and Time
Description: The date and time when the collision occurred.
Data Type: DateTime.
Issues: Inconsistent formats.
Recommendations: Standardize to a consistent format, e.g., YYYY-MM-DD HH:MM:SS.

3. Location
Description: The location of the collision.
Data Type: String (Potential for geolocation parsing).
Issues: Inconsistent formats, mixed with latitude and longitude data.
Recommendations: Parse into separate latitude and longitude columns if mixed. Standardize address formats.

4. Severity
Description: The severity of the collision.
Data Type: Categorical (e.g., Minor, Moderate, Severe).
Issues: Inconsistent categorization.
Recommendations: Define a standard set of severity categories and map existing data to this set.

5. Vehicles Involved
Description: Number of vehicles involved in the collision.
Data Type: Integer.
Issues: Missing values, outliers.
Recommendations: Impute missing values based on available data. Review outliers for data entry errors.

6. Injuries
Description: Number of injuries reported.
Data Type: Integer.

Issues: Missing values, potential underreporting.
Recommendations: Impute missing values cautiously. Correlate with severity for consistency checks.

7. Fatalities
Description: Number of fatalities in the collision.
Data Type: Integer.
Issues: Missing or incorrect values.
Recommendations: Cross-verify with external sources where possible. Treat zeros with scrutiny.

8. Weather Conditions
Description: Weather conditions at the time of the collision.
Data Type: Categorical.
Issues: Varied descriptions, missing values.
Recommendations: Standardize descriptions. Consider a 'Not Reported' category for missing values.

9. Road Conditions
Description: Road conditions at the collision site.
Data Type: Categorical.
Issues: Inconsistent descriptions.
Recommendations: Define a standard set of road condition categories and map existing descriptions to this set.

Conclusion:

The specifications detailed in this document provide a framework for addressing the identified issues within the Austin Collision dataset. By adhering to these specifications, the dataset can be prepared more effectively for analytical purposes, leading to more reliable and insightful outcomes.

Data Profiling Documentation: **New York Collision Dataset**

Introduction

This report synthesizes the findings from the data profiling of the New York Collision dataset, leveraging the YData profiling tool. It emphasizes the dataset's quality, completeness, and structural considerations, offering tailored insights into handling specific columns, especially those with unsupported types and null values.

Standardization Solutions

Date and Time Conversion: Standardize 'Date and Time' fields using the **transformations in tmap**, enabling more straightforward chronological analyses and sorting.

Null Values:

Null values were prevalent across several columns.
Action: All the Null values have been mapped to one key which we have created in Location Dim.

Data Structure Adjustments

Normalization: For columns with multi-valued attributes columns, normalize the data. This involves creating new tables for better data management and analysis.

Combined Columns:

1. We have merged the columns **contributing_factor_vehicle 1 through 5** into a single column **'contributing_factor'.**
2. We have consolidated **vehicle_type_code 1 through 5** into a single column named **'units_involved'**.

New York Collision Dataset Column Specifications

Collision ID: Ensure uniqueness. Data Type: Integer.

Date and Time: Standardize to YYYY-MM-DD HH:MM:SS. Data Type: DateTime.

Location: Split into latitude and longitude if combined. Standardize formats. Data Type: String.
Severity, Vehicles Involved, Injuries, Fatalities, Weather Conditions, Road Conditions: Apply relevant standardization, imputation, and outlier handling strategies as discussed.

Conclusion

This report provides a comprehensive strategy to enhance the New York Collision dataset's structure, quality, and analysis readiness. Addressing unique and common challenges alike ensures the dataset's utility for generating insightful and reliable outcomes in traffic safety analysis and policy formulation.

Data Profiling Documentation: **Chicago Collision Dataset**

Introduction

The process began with generating a comprehensive data profile report using tools like Pandas Profiling. This initial analysis highlighted several areas requiring attention: null values, inconsistent date formats, combined data fields, outliers, and duplicate entries.

Key Findings and Actions

Null Values:

Null values were prevalent across several columns.
Action: All the Null values have been mapped to one key which we have created in Location Dim.

Date Formats:

Inconsistent date formats complicated temporal analysis.
Action: All date columns were standardized to the DATETIME format (%Y-%m-%d %H:%M:%S).

Combined Columns:

We have merged the **'PRIM_CONTRIBUTING_CAUSE**' and **'SEC_CONTRIBUTING_CAUSE'** columns into one.

Duplicates:

Duplicate records were detected.
Action: All duplicates were removed to ensure data integrity.

Conclusion

The meticulous profiling and cleaning of the Chicago Collision dataset have significantly improved its reliability and readiness for analysis. Addressing key issues like null values, inconsistent formats, and outliers has enhanced the dataset's integrity. This refined dataset is now better positioned to yield insightful and trustworthy analyses, supporting informed decisions in traffic safety and urban planning.

Data Profiling Report Summary

Overview

This report summarizes the profiling results of collision datasets from Austin, Chicago, and New York City (NYC). The primary focus is to identify and outline the specific issues detected within each dataset and to discuss potential challenges that could arise when attempting to integrate these datasets.

Dataset Alerts Summary

Austin:
Key Issues Identified:
Constant Values: Certain fields, such as pedestrian_fl and motor_vehicle_fl, are flagged for having constant values, which diminishes their utility for analytical purposes.

Chicago:
Key Issues Identified:
Constant Values: Fields like INJURIES_UNKNOWN and TRAFFIC_CONTROL_DEVICE exhibit constant values across the dataset, indicating limited variability and potential information redundancy.

NYC:
Key Issues Identified:
High Imbalance: The dataset contains fields with high imbalance, including NUMBER OF PEDESTRIANS KILLED and NUMBER OF CYCLIST INJURED, suggesting that these incidents are either rare or inconsistently recorded.

Potential Integration Challenges

Integrating the datasets from Austin, Chicago, and NYC presents several challenges, detailed as follows:

Inconsistent Field Names and Values:

The presence of fields with constant or highly imbalanced values across the datasets suggests differences in data collection or recording practices. This variance necessitates careful consideration in aligning similar fields across datasets for meaningful analysis.

Data Quality and Completeness:

Issues such as constant values indicate potential shortcomings in how data was collected or errors in data entry. These factors can significantly affect the accuracy and reliability of analyses conducted on the integrated data.

Schema Alignment:

Harmonizing the schemas of the different datasets involves mapping conceptually similar fields that may have different names and reconciling any disparities in data granularity or format. This process is crucial to ensure that the integrated dataset provides a coherent and accurate representation of the underlying data.

Conclusion

The integration of the Austin, Chicago, and NYC collision datasets requires a comprehensive approach to data cleaning and transformation. Efforts must be made to standardize field names, address data quality issues, and ensure that the datasets are compatible for combined analysis. Overcoming these challenges is essential for deriving meaningful insights from the integrated dataset.