Data Report for Data Exploration and Risk Assessment of Vehicle Collisions in NYC

Question

What are the primary factors contributing to motor vehicle collisions in New York City?

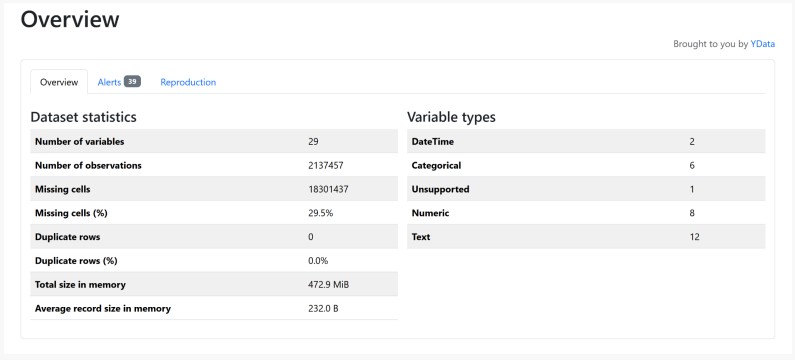
Data Sources

Dataset Description:

The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police-reported motor vehicle collisions in NYC. The police report (MV104-AN) must be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage.

Dataset Details:

- **Source**: NYC Open Data platform, Motor Vehicle Collisions - Crashes | NYC Open Data
- **Dataset Name**: "Motor Vehicle Collisions - Crashes"
- **Metadata URL**: Metadata
- **Data URL**: Data CSV
- **Data Type**: CSV
- **Agency**: Police Department (NYPD)
- **Rows:** 2.14M
- **Columns:** 29



Reason for Choosing:

This dataset provides comprehensive information on motor vehicle collisions across NYC, making it suitable for understanding collision patterns and high-risk factors.

Structure and Quality:

- **Structure**: Tabular, with many valuable columns for timestamps, geospatial data (latitude, longitude), and categorical/contributing factors.
- **Quality**:
  - **Completeness**: Some critical columns contain very few missing values (e.g., geospatial data and contributing factors).
  - **Volume**: Large dataset covering several years, ideal for identifying trends but requires efficient handling.

Licenses:

- **License**: NYC OpenData. [Motor Vehicle Collisions - Crashes | NYC Open Data]
- **Why Allowed**: This license permits unrestricted use, provided attribution is given.
- **Plan to Fulfill Obligations**:
  - Include attribution and visualizations in the final report.
  - Provide proper references to NYC Open Data in all outputs.

Data Pipeline

Technology

- **Language**: Python
- **Tools**: Pandas, NumPy, GeoPandas for geospatial analysis.

Steps

1. **Import Data**:
   - Load CSV into a Pandas DataFrame.
2. **Cleaning**:
   - Handle missing values by:
     - Removing rows where critical data (e.g., geospatial coordinates) is missing.
     - Replacing missing values with 0 or NULL
   - Removing duplicate rows if any.
3. **Transformation**:
   - Convert timestamps to datetime objects for time-series analysis.
4. **Quality Assurance**:
   - Ensure no duplicate entries.
5. **Error Handling**:
   - Log rows with errors (e.g., missing data) for future review.

Challenges and Solutions

- **Handling Missing Data**: Developed a strategy to retain as much data as possible while ensuring analysis accuracy.

Meta-Quality Measures

- Implemented logging for each pipeline stage.
- Integrated checks to validate new data against existing schema (e.g., geospatial ranges).

Result and Limitations

Output Data

- **Data Structure**: Cleaned tabular data with consistent datetime, geospatial, and contributing factor information.
- **Quality**:
    - High-quality geospatial data for hotspot analysis.
    - Reduced ambiguity in categorical variables.

Format

- **Chosen Format**: CSV for compatibility with other tools and dashboards.
- **Reason**: Easy to integrate with visualization libraries and share with stakeholders.

Reflection and Limitations

- **Strengths**:
    - Clear insights into spatial and temporal collision trends.
    - Robust error handling and logging.
    - With over 2 million observations, the dataset provides a rich source for uncovering trends and patterns in motor vehicle collisions across NYC.
- **Limitations**:
    - Potential underreporting or misclassification in source data.
    - High Proportion of Missing Values: 29.5% missing cells may lead to incomplete or biased analysis, especially if critical information
    - With 6 categorical variables, inconsistent or ambiguous categories (e.g., "Unspecified") may require significant cleaning and standardization efforts.
    - Text variables may require additional preprocessing (e.g., tokenization or normalization) for effective analysis, especially if they contain inconsistent formatting.