Milestone 1 Report: Data Cleaning, Integration, and Exploratory Data Analysis

1. Introduction

This report documents the entire Milestone 1 process for the NYC Collision Dataset, including data cleaning, integration of crash-level and person-level datasets, and exploratory data analysis. The goal is to prepare a clean, reliable dataset and extract meaningful preliminary insights.

2. Dataset Overview

Two primary datasets were used: the crashes dataset and the persons dataset. The crashes dataset contains crash-level information such as date, time, location, and contributing factors. The persons dataset contains person-level data such as age, injury status, and position in vehicle. Both datasets are related through the field COLLISION_ID.

3. Data Cleaning

Cleaning steps included: standardizing column names, converting CRASH_DATE and CRASH_TIME to proper datetime formats, removing duplicates, addressing invalid geographic coordinates using NYC latitude/longitude boundaries, dropping columns with excessive missing values, and handling missing entries using appropriate imputation techniques. CRASH_TIME was fully standardized into a datetime format to support time-of-day analysis.

4. Persons Dataset Cleaning

The persons dataset was cleaned by converting person age into numeric form, removing duplicates, dropping rows missing essential identifiers like COLLISION_ID, and imputing missing ages using the median. Injury descriptions were standardized by replacing missing entries with 'Unknown'.

5. Data Integration

Integration was performed by aggregating the persons dataset at the COLLISION_ID level, producing: TOTAL_PERSONS (count of individuals per crash), TOTAL_INJURED (individuals with non-zero injury), and AVG_PERSON_AGE (average age per crash). These aggregated values were merged into the crash dataset using a left join.

6. Post■Integration Cleaning

After merging, new missing values arising from unmatched COLLISION_ID entries were addressed. Aggregated metrics were filled with zeros or median values where appropriate. Final checks ensured that all merged fields contained correct and consistent data types.

7. Exploratory Data Analysis

EDA included missing value analysis, boxplots for numerical distributions, summary statistics for both numerical and categorical features, and examination of injury distributions. Crash trends were analyzed across dimensions such as hour of day, weekday, and month. The top contributing factors leading to crashes were identified, as well as correlations between key numeric variables. Visualizations show that traffic congestion and human behavior strongly influence collision frequency.

8. Key Insights

• Crashes peak during evening rush hour (3–6 PM). • Brooklyn and Queens contribute the highest number of collisions. • Driver inattention is the most common contributing factor. • Most crashes involve zero or minor injuries, but severe outliers exist. • Average age distributions reflect typical driver demographics. • Weekdays see significantly more crashes than weekends.

9. Conclusion

The dataset is now fully cleaned, validated, and integrated. EDA has revealed strong temporal and geographic patterns, as well as the influence of human factors. The processed dataset is ready for subsequent analysis, modeling, or dashboard development in later milestones.