# 1.Data Cleaning & Preprocessing Report

Dataset: Traffic Report Dataset
Initial Size: 209,306 rows × 24 columns
Final Clean Size: 135,864 rows × 25 columns

---

## Missing Values Report

**Identification**

Checked all columns for null values.
Missing values were found mainly in:
Injury-related columns
Unit count fields

**Missing Value Summary**

| Category | Details |
|---|---|
| Columns with Missing Data | Injury counts, unit counts |
| Percentage | Varied by column (low to moderate) |
| Critical Fields | No date/time or categorical dimension was fully missing |

## Handling Techniques Used

Categorical columns: Mode imputation
(e.g., weather_condition, traffic_control_device)
Numerical columns: Median imputation
(e.g., injuries_total, num_units)
No columns removed to preserve analytical depth

**Result**

100% missing values eliminated
Dataset is now complete with no NULL values

---

# 2.Standardized Date & Time Format

**Fields Standardized**

crash_date → Converted to YYYY-MM-DD
Created a new column:
crash_datetime → YYYY-MM-DD HH:MM:SS
by combining:

- crash_date
- crash_hour

**Transformations Applied**

Used datetime parsing with error handling
Converted crash hour into time delta for consistency

**Analytical Benefits**

Now you can easily:

Filter crashes by specific date ranges

Group data by:

- Day
- Week
- Month

Perform time-series traffic pattern analysis

**Example:**

Identify peak crash hours

Analyze monthly accident trends

---

## 3. Outlier Detection & Removal Summary

**Method Used**

Interquartile Range (IQR) Method

Q1 (25th percentile)

Q3 (75th percentile)

Outliers detected beyond:

Q1 − 1.5 × IQR and Q3 + 1.5 × IQR

**Outliers Identified**

| Column | Outlier Count |
|---|---|
| num_units | 19,940 |
| injuries_total | 46,494 |
| injuries_no_indication | 7,008 |

## 4. Refined, Clean Dataset:

- The final, cleaned dataset ready for analysis, with the following improvements:
  - No missing values.
  - Standardized date and time fields.
  - Outliers removed or adjusted.

**A brief summary of how the dataset is now better suited for traffic pattern analysis.**

“The dataset was cleaned by imputing missing values using median and mode, standardizing date-time formats for time-series analysis, and removing statistical outliers using the IQR method. The refined dataset is now reliable, consistent, and suitable for accurate traffic pattern and accident severity analysis.”