📊 Traffic Crash Dataset – Initial Data Exploration & Understanding Report

---

1. Project Overview

Objective

The objective of this project is to explore and understand a real-world traffic crash dataset using Power BI Desktop, identify initial patterns, assess data quality, and document key findings that can guide further analysis and visualization. This phase focuses on data familiarity and readiness, which is a critical first step in any data analytics project.

---

2. Dataset Overview

Dataset Size and Structure

Total Rows: 209,306

Total Columns: 24

File Format: CSV (Comma-Separated Values)

File Size: Medium–Large (suitable for BI tools like Power BI)

Each row in the dataset represents a single traffic crash event.

Key Variables in the Dataset

Category	Variables	Description
Time-based	crash_date, crash_hour, crash_day_of_week, crash_month	When the accident occurred
Environmental	weather_condition, lighting_condition	Weather and visibility conditions
Road & Traffic	traffic_control_device, trafficway_type, alignment, roadway_surface_cond, road_defect	Road and traffic characteristics
Crash Nature	first_crash_type, crash_type	Type and manner of collision
Severity	injuries_total, injuries_fatal, injuries_incapacitating	Injury outcomes
Vehicle Info	num_units	Number of vehicles involved

---

3. Data Import & Power BI Configuration

Tools Used

Power BI Desktop (latest version)

Power Query Editor for data profiling and transformation

Import Process

Dataset was successfully imported from CSV format into Power BI.

No file corruption or delimiter issues were observed during import.

---

4. Data Quality Observations

Missing or Null Values

No missing or null values detected across the dataset.

Indicates strong data completeness and reliability.

Inconsistent Data Types

Column Name	Current Data Type	Issue Identified	Recommended Action
crash_date	Text (Object)	Cannot perform time analysis	Convert to DateTime
intersection_related_i	Text (Y/N)	Boolean stored as text	Convert to True/False
injuries_* columns	Decimal (Float)	Injury counts should be whole numbers	Convert to Integer

Duplicate Records

No exact duplicate rows were found.

Confirms uniqueness of crash records.

Irrelevant Columns

No irrelevant columns identified at this stage.

All fields provide contextual or analytical value.

---

5. Potential Data Patterns (Initial Findings)

Time-Based Trends

High crash frequency observed during:

Morning rush hours

Evening peak traffic hours

Weekday vs Weekend Behavior

Weekends show:

Lower total crash volume

Different crash type distribution compared to weekdays

Environmental Factors

A large proportion of crashes occur during:

Clear weather

Daylight conditions

➞ Indicates human behavior and traffic density are major contributors rather than weather alone.

Crash Type Distribution

Most frequent crash types:

Rear-end collisions

Turning-related crashes

Injury Distribution

Injury counts follow a long-tailed distribution:

Majority of crashes result in no or minor injuries

A small number lead to severe or fatal outcomes

6. Outliers & Anomalies

Rare cases with:
High injury counts involving multiple vehicles
Severe injuries even with moderate vehicle damage
These records require focused analysis in later stages.

---

7. Challenges & Issues Identified

Data Cleaning Requirements
Convert crash_date to DateTime format
Normalize Y/N fields to Boolean
Ensure injury counts are stored as integers
Analytical Challenges
No geographic (latitude/longitude) data available for hotspot mapping
Damage values are categorical ranges, limiting numeric cost analysis

---

8. Recommendations for Further Processing

Data Preparation
Create a Date Dimension Table for advanced time intelligence
Generate derived fields such as:
Peak vs Off-Peak hours
Weekend flag
Accident severity score
Analysis & Visualization
Hourly and daily trend analysis
Injury severity comparison by crash type
Environmental condition impact dashboards
Business Insights Focus
Identify high-risk periods
Understand conditions leading to severe injuries
Support data-driven traffic safety recommendations

---

9. First Impressions & Next Steps

Key Insights
Traffic crashes are highly influenced by time and human behavior
Environmental conditions alone do not explain accident severity
Turning and rear-end crashes should be priority areas for safety improvement
Suggested Visualizations
Line chart: Accidents by hour
Heatmap: Day vs Hour crash density
Bar chart: Crash type frequency
KPI cards: Total accidents, injuries, fatal cases

10. Power BI Dashboard Design (Visual Layout & Rationale)

Page 1: Executive Overview Dashboard

Purpose: Provide a high-level snapshot of traffic accidents and severity.

Visuals Used:

KPI Cards (Top – aligned horizontally):

Total Accidents (COUNTROWS)

Total Injuries (SUM of injuries_total)

Fatal Injuries (SUM of injuries_fatal)

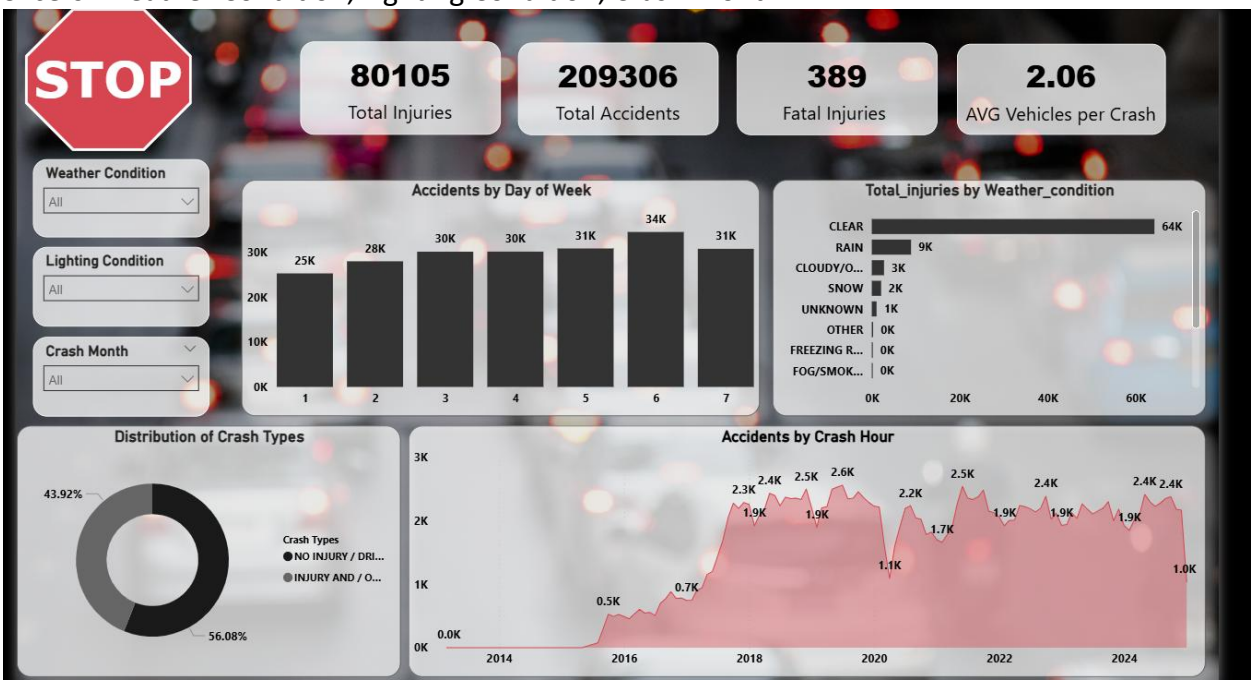
Average Vehicles per Crash (AVERAGE of num_units)

Line Chart: Accidents by Crash Hour (identifies peak traffic risk hours)

Column Chart: Accidents by Day of Week (weekday vs weekend comparison)

Donut Chart: Distribution of Crash Types

Slicers: Weather Condition, Lighting Condition, Crash Month



Page 2: Conditions & Crash Severity Analysis

Purpose: Understand environmental and crash-type impact on accidents.

Visuals Used:

Stacked Bar Chart: Weather Condition vs Accidents (segmented by lighting_condition)

Bar Chart: First Crash Type vs Total Accidents

Table/Matrix: crash_type with Total Accidents, Total Injuries, Severe Injury %

Line Chart: Monthly Accident Trend (crash_month)

11. DAX Measures Used (with Explanation)

Total Accidents

Total Accidents = COUNTROWS('Traffic Dataset')

Counts each crash event (each row represents one accident).

Total Injuries

Total Injuries = SUM('Traffic Dataset' [injuries_total])

Calculates overall injury impact across all crashes.

Fatal Injuries

Fatal Injuries = SUM('Traffic Dataset' [injuries_fatal])

Measures number of deaths caused by traffic crashes.

Average Vehicles per Crash

Avg Vehicles per Crash = AVERAGE('Traffic Dataset' [num_units])

Shows crash complexity and traffic density.

Severe Injury Rate

Severe Injury % =

DIVIDE(

SUM('Traffic Dataset'[injuries_fatal]) + SUM('Traffic Dataset'[injuries_incapacitating]),

SUM('Traffic Dataset'[injuries_total]),

0

)

Represents proportion of injuries that are fatal or incapacitating.

12. Final Insights & Recommendations

Key Insights

Accident frequency peaks during morning and evening rush hours.

Clear weather and daylight conditions still show high accident counts, indicating human behavioral factors.
Rear-end and turning-related crashes are the most frequent accident types.
Majority of crashes result in minor or no injuries; however, severe injuries occur in a small but critical percentage of cases.
Multi-vehicle crashes show higher injury severity.
Recommendations
Increase traffic enforcement and monitoring during peak hours.
Improve intersection design and turning lane safety measures.
Conduct road safety awareness campaigns focusing on driver behavior.
Use time-based crash data to plan preventive safety interventions.
Integrate geographic data in future analysis to identify accident hotspots.