Exploratory Data Analysis (EDA) Report

Dataset: Traffic Crashes – Crashes (Chicago)
Shape: 920,282 rows × 48 columns
Tools Used: Python Libraries - Pandas, Matplotlib, Seaborn
Objective: To perform statistical and graphical exploration of the dataset to recognize significant patterns, relationships, and anomalies concerning Chicago traffic accidents

1. Dataset Overview

This dataset records detailed information about reported traffic crashes in Chicago. It includes features such as crash types, dates, times, road conditions, weather, lighting, injuries, speed limits, and geographical locations.

Sample Columns:

- POSTED_SPEED_LIMIT, WEATHER_CONDITION, LIGHTING_CONDITION
- INJURIES_TOTAL, INJURIES_FATAL, HIT_AND_RUN_I
- CRASH_HOUR, CRASH_DAY_OF_WEEK, CRASH_MONTH
- LATITUDE, LONGITUDE (used for mapping hotspots)

2. Data Preprocessing

Key steps taken:

- **Renamed columns** for better readability (e.g., HIT_AND_RUN_I → HIT_AND_RUN)
- **Dropped unnecessary columns** such as DOORING_I, WORK_ZONE_TYPE, PHOTOS_TAKEN_I, etc.
- **Removed duplicates** using .duplicated() and .drop_duplicates()
- **Handled missing values**:

- Categorical columns filled with 'Unknown' or 'No'
- Numerical/null-heavy rows dropped selectively (e.g., LATITUDE, LONGITUDE)
- Used .mode() to fill dominant values (e.g., BEAT_OF_OCCURRENCE)

- Final shape after cleaning: **(depends on your final df.shape)**

3. Statistical Overview

Used:

- .describe(), .info(), .value_counts()

Observations:

- Majority of crashes occurred under **clear or cloudy** weather.
- Most entries for MOST_SEVERE_INJURY were **non-severe** or **not indicated**.
- Many crashes occurred with **no fatal injuries**.
- POSTED_SPEED_LIMIT ranged from **5 to 70 mph**, with most common being **30 mph**.

4. Visual Analysis & Insights

Crash Timing & Frequency

- **Insight**: Most crashes occur during **rush hours (7–9 AM & 5–7 PM)**.
- **Density Plot & Heatmap**: Show crash **peaks by hour and day of the week**, especially on weekdays.

Speed Limit vs. Injuries

- **Insight**: Higher **speed limits** generally correlate with **higher injury counts**.

Weather Impact

- **Grouped Bar Chart**: Compared **average injuries** across different **weather conditions**.
- **Insight**: Hazardous weather types (e.g., rain, snow) show **higher average injuries** despite fewer incidents.

Monthly & Seasonal Trends

- **Line Plot**: Visualized **crash trends per month**, showing **seasonal peaks**.
- **Insight**: Certain months (e.g., **OCTOBER**) had more reported crashes, potentially due to holidays and weather.

Injury Severity

- **Pie Chart**: Showed proportions of **fatal, incapacitating, non-incapacitating**, and **unknown** injuries.
- **Insight**: Majority of injuries were **non-severe**, but **fatal injuries**, though rare, were non-negligible.

Environmental Conditions

- **Box Plot**: Displayed **Crash Hour by Day of the Week**.
- **Insight**: Weekends show a wider spread of crash times, including late nights. Weekdays have crashes mostly during commute hours.
-

Multivariate Relationships

- **Pairplot** & **Heatmap of correlations**: Showed that:
    - INJURIES_TOTAL is highly correlated with INJURIES_NON_INCAPACITATING
    - Many injury types are positively correlated with one another

5. Key Findings

- **Peak crash times** align with **morning and evening rush hours**.
- **Speed and weather** significantly influence the **severity of injuries**.
- **Winter and summer** months see increased crash rates, possibly due to holidays and road/weather conditions.

- **Lighting and road surface** play a role in crash severity.
- **Geospatial hotspots** can inform better traffic planning and enforcement.

6. Conclusion

This EDA uncovers vital trends in Chicago traffic crashes. Knowing when, where, and under what conditions crashes happen inform city planners, traffic engineers, and policymakers about how to improve safety interventions.

Further steps could include:

- Building predictive models for crash severity
- Mapping high-risk intersections for preventive infrastructure
- Temporal analysis for policy-making during seasonal peaks