ETL Traffic Collision Analysis Report

Name: Arun Santhosh, Huang-Yin Tso and Alok Kumar
Course: DS C69 Case Study Partner Details - DE track
Project Title: ETL Project – California Traffic Collisions
Date: 08/04/2025

1. Introduction

This project analyzes traffic collision data in California. The goal is to clean, transform, and explore the dataset using PySpark and generate insights to help reduce traffic-related risks. The data includes details about the collision time, weather, lighting, road conditions, victim information, and more.

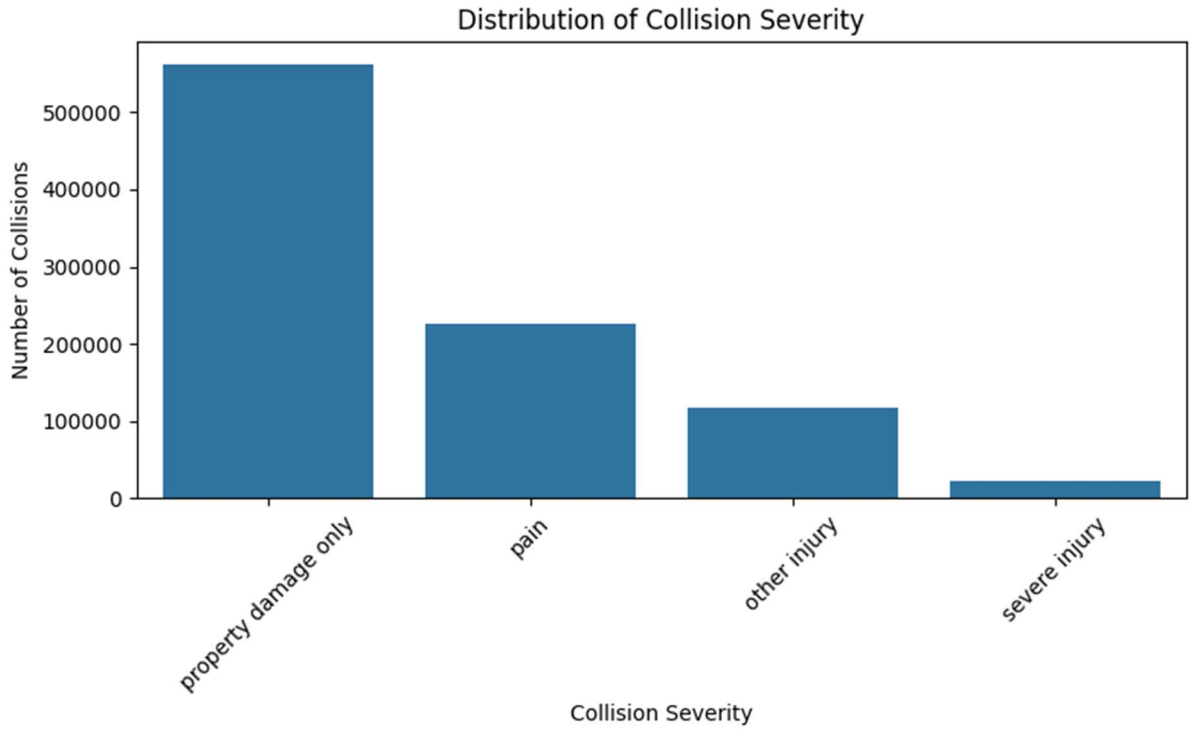
2. Data Preparation and Cleaning

- The data was extracted, loaded into dataframes, and cleaned using PySpark and Pandas.
- Missing values were handled using imputation or by removing incomplete rows.
- Outliers in numerical columns were removed using the Interquartile Range (IQR) method.
- Duplicate rows were removed.
- Categorical variables were indexed using StringIndexer.
- Data types were adjusted to enable analysis (e.g., converting time columns to string or timestamp).

3. Exploratory Data Analysis (EDA)

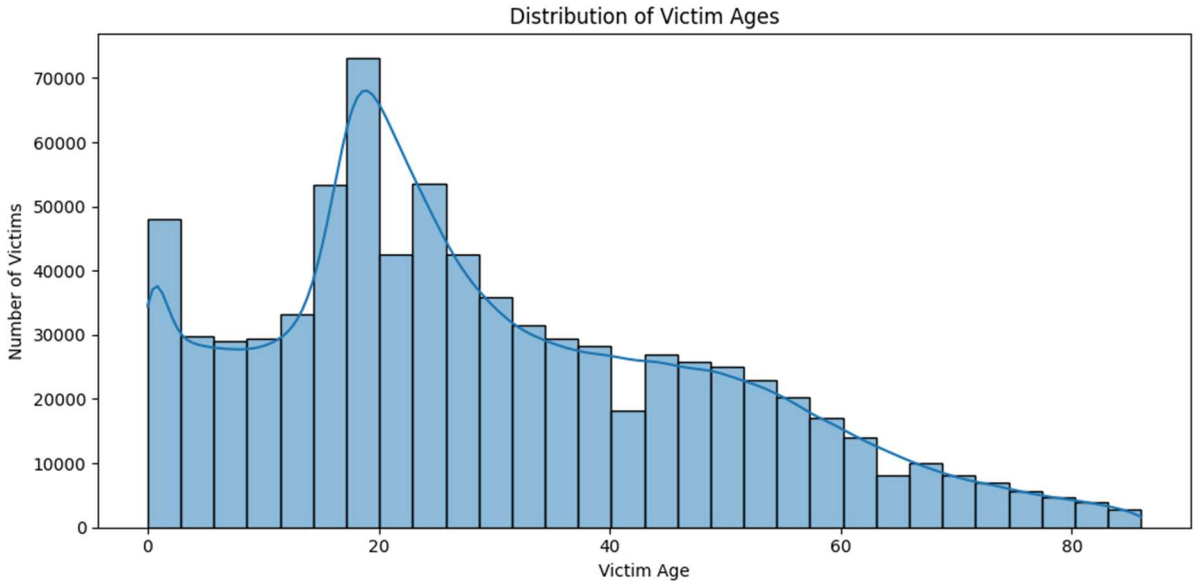
3.1 Collision Severity

- Most collisions were categorized as **property damage only** or **complaints of pain**.
- Very few resulted in fatalities.



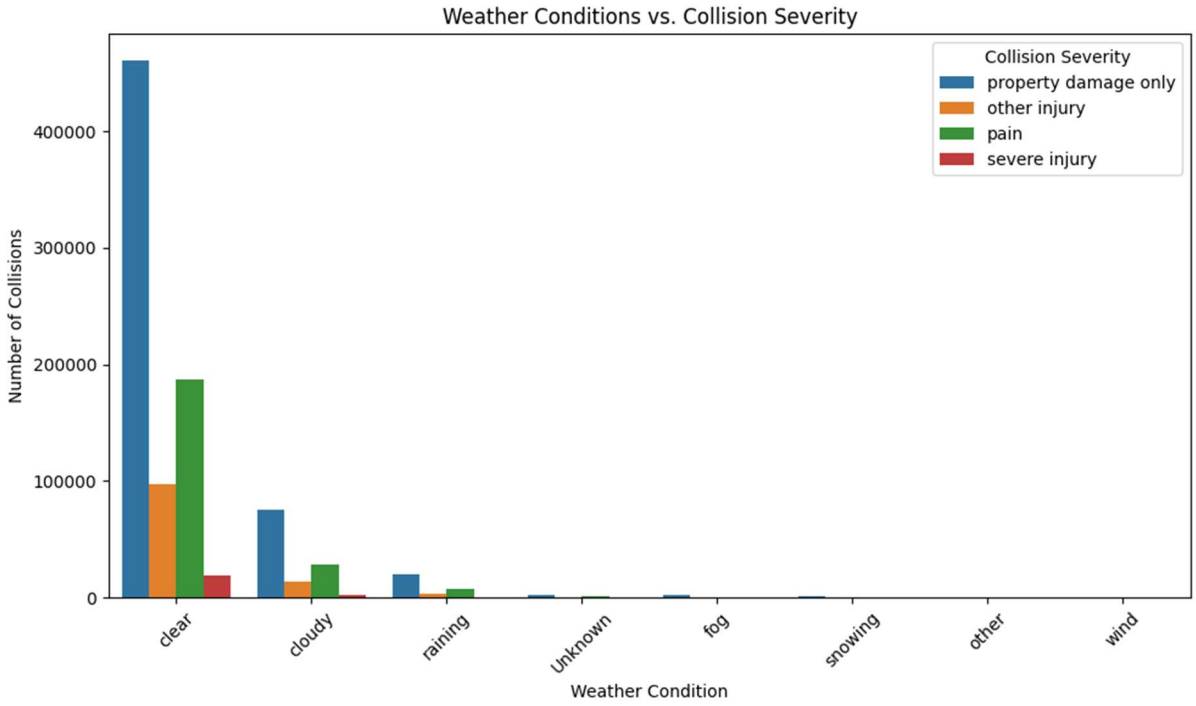
3.2 Victim Demographics

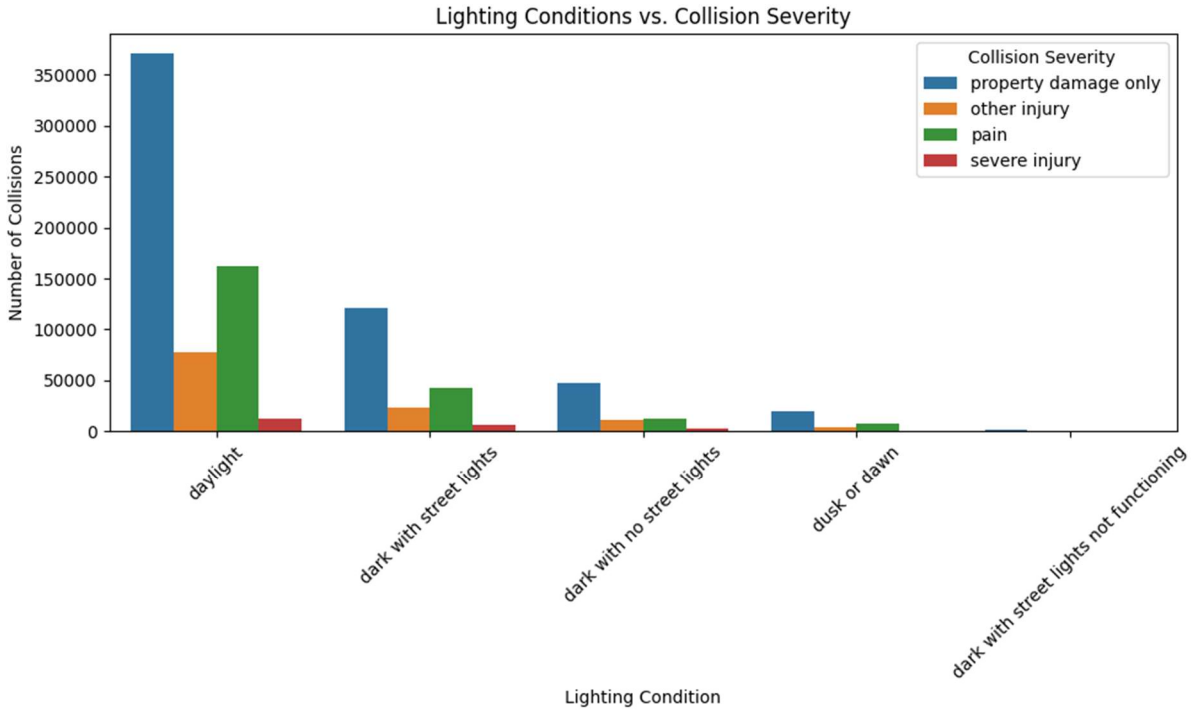
- Most victims were aged between **20 to 40 years old**.
- Male victims were slightly more common than female victims.



3.3 Weather and Lighting Conditions

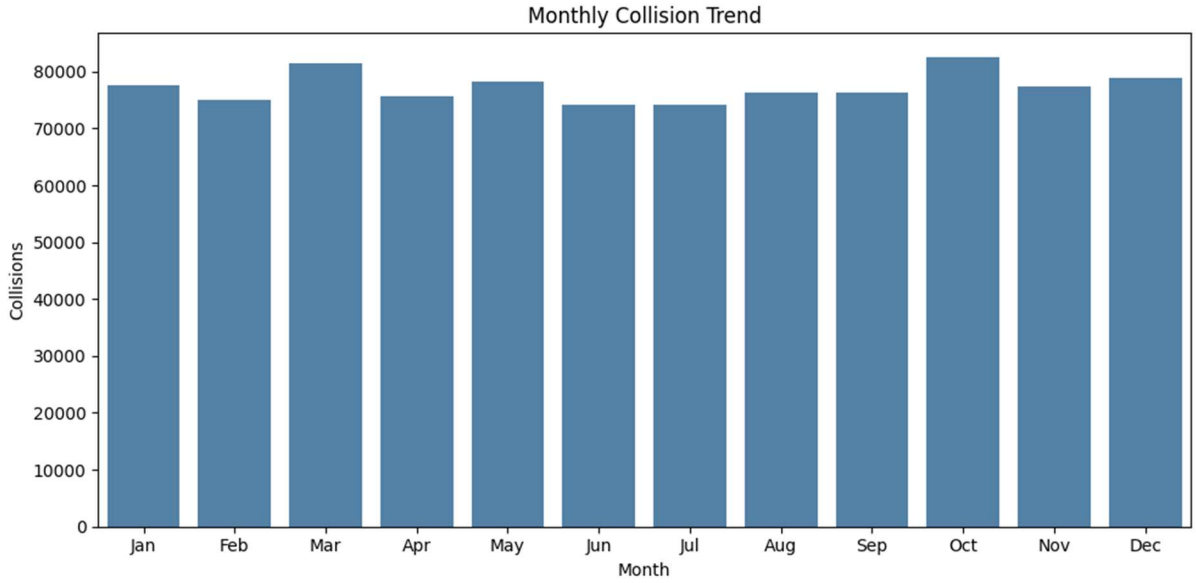
- Most collisions occurred in **clear weather** and **daylight**.
- This suggests high activity during normal driving conditions.

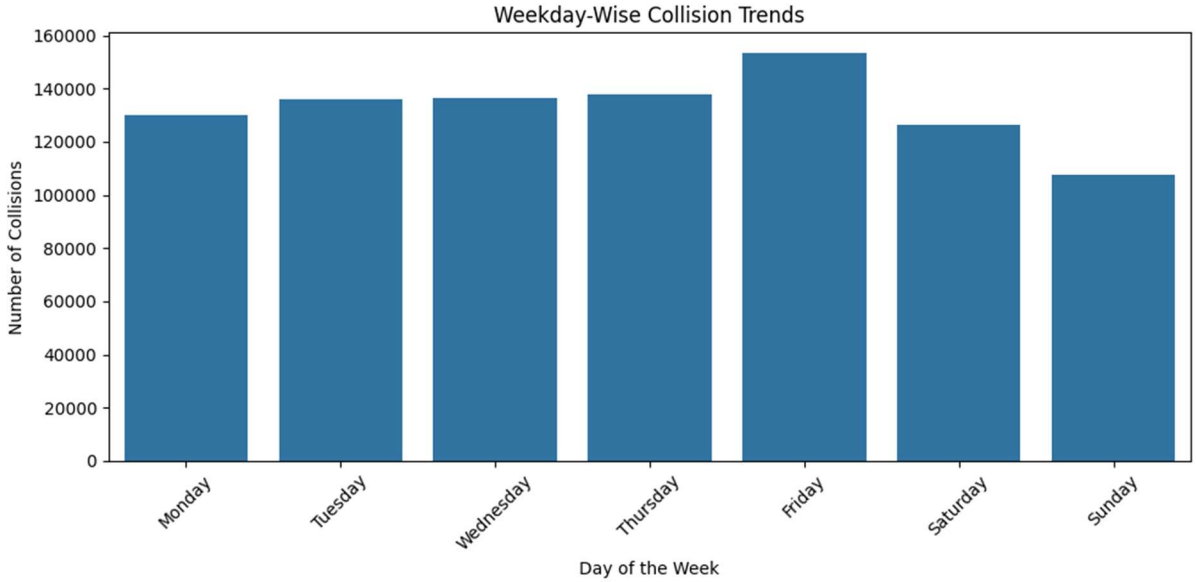


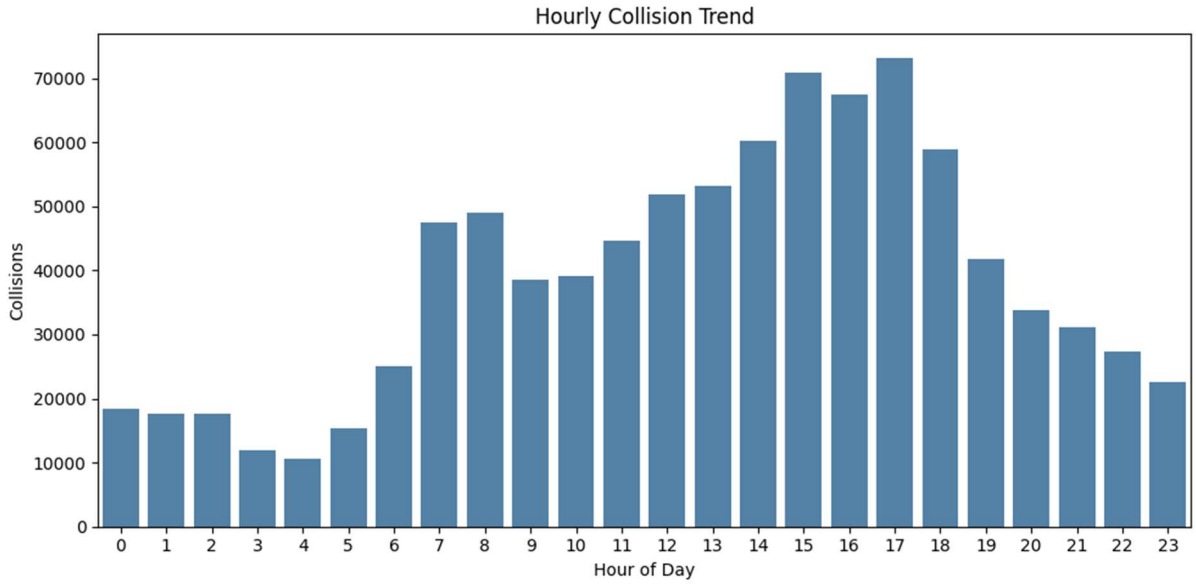


3.4 Temporal Trends

- Collisions were most frequent around **3 PM**, during afternoon traffic.
- **Fridays** had the highest collision count.
- **Monthly trends** showed peaks during summer months.

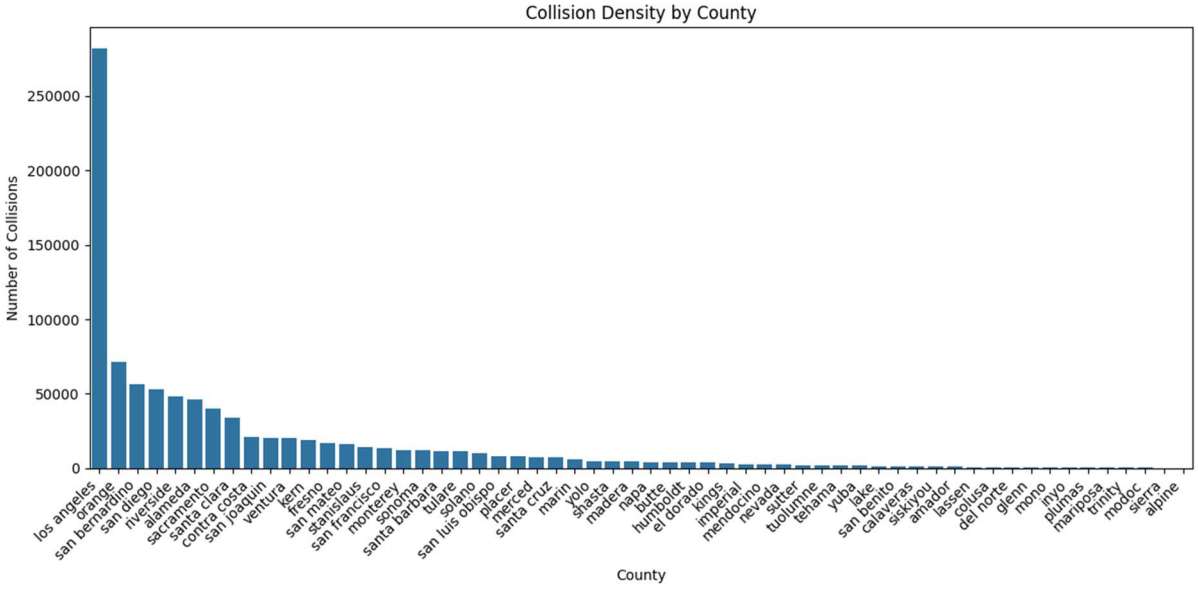






3.5 Spatial Distribution

- The highest number of collisions occurred in **Los Angeles County**, followed by San Bernardino and San Diego.



4. Key Findings and Insights

- **3 PM and Fridays** are the most dangerous times to drive.
- Most collisions happen in normal conditions, not extreme weather — this indicates that driver behavior is more important than the environment.
- Large, populated counties like **Los Angeles** require special attention.

- **Dry roads and daylight** are not indicators of safety — collisions are still frequent.

5. Recommendations

- Increase **patrol and traffic enforcement** during peak hours and in top 5 counties.
- Launch public **awareness campaigns** targeting afternoon and Friday drivers.
- Encourage **safe driving in normal conditions**, as most collisions happen when drivers may feel relaxed or distracted.
- Use collision patterns to improve **road safety infrastructure** in high-risk zones.

6. Assumptions

- It was assumed that killed_victims > 0 implies a fatal collision.
- Time values were assumed to follow the "HH:mm:ss" format for conversion.
- If location fields were missing (county_location), they were not included in the spatial analysis.

7. Conclusion

- The analysis shows that collisions are more affected by human behavior than by weather or road conditions. Data-driven strategies such as targeted policing, driver education, and urban planning can help reduce the risk of traffic accidents.