Introduction

This project aims to analyze traffic collision data using an end-to-end ETL (Extract, Transform, Load) process. The analysis identifies high-risk conditions, locations, and times that contribute to accidents, using Python (pandas, matplotlib, seaborn) and AWS S3 for storage.

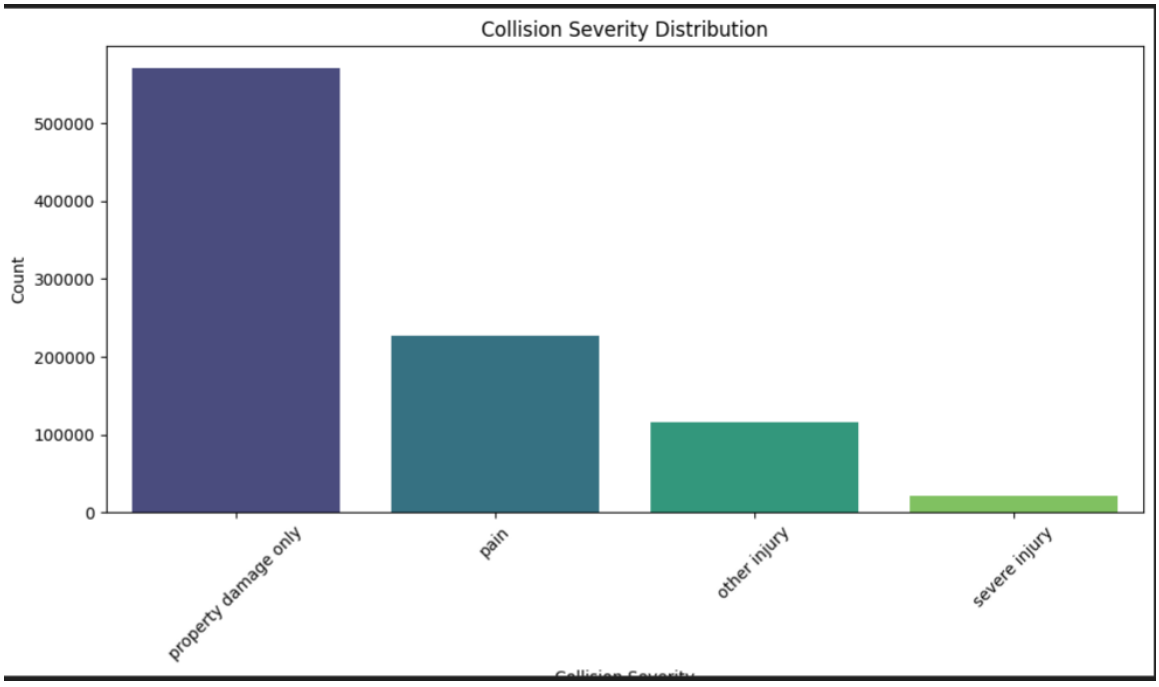
Dataset Description

The dataset includes information on collisions, victims, parties, and case metadata. Important fields include collision date, severity, victim age, weather, lighting, and geographic coordinates.

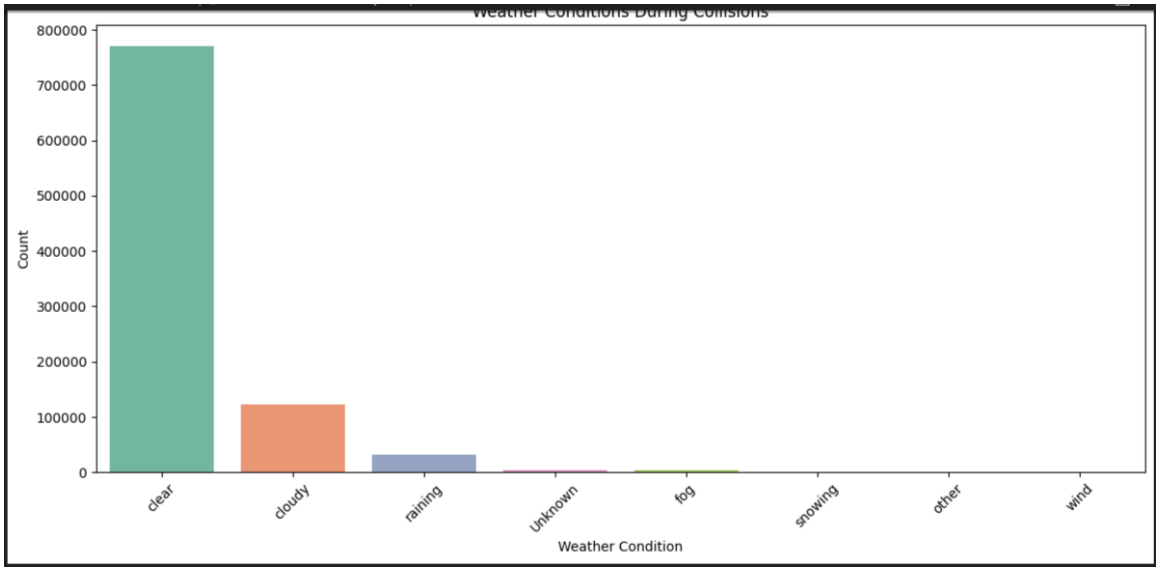
ETL Summary

Data was loaded from CSV files, cleaned by handling missing values, removing duplicates, and encoding categorical features. Sparse and invalid rows were dropped, and the resulting data was saved and uploaded to S3.

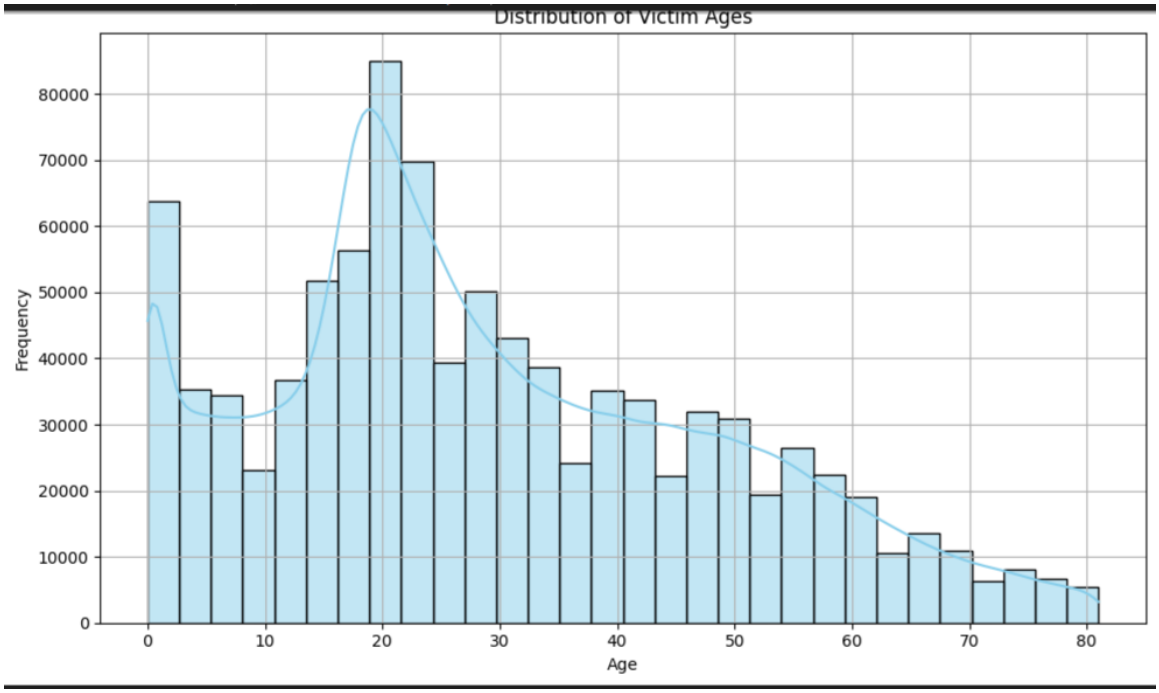
Collision Severity Distribution



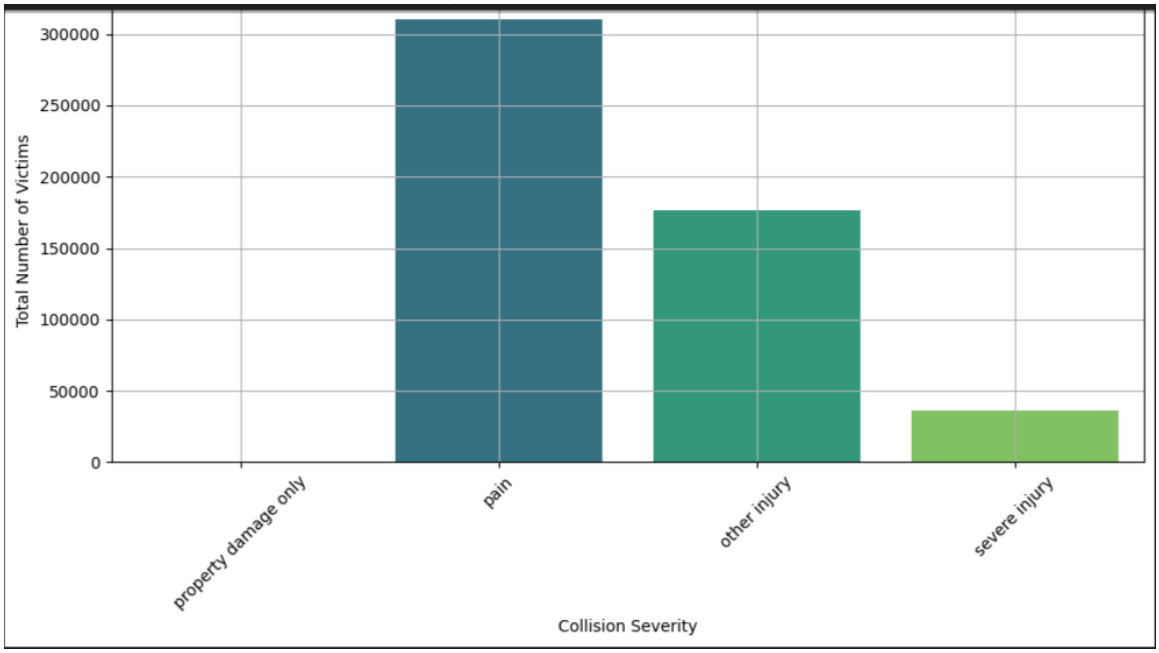
Weather Conditions During Collisions



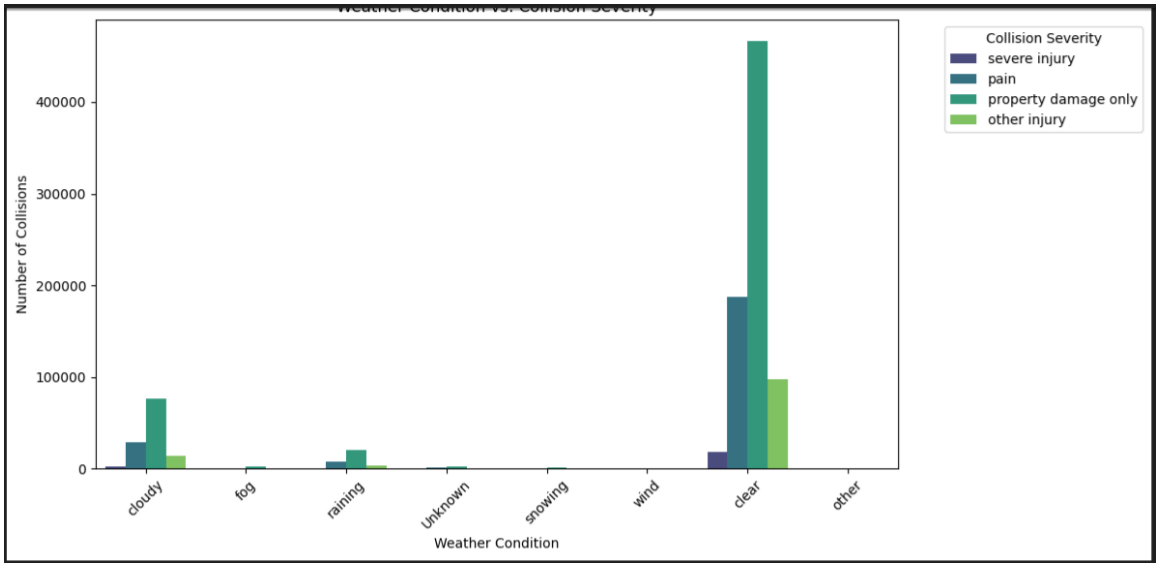
Distribution Of Victim Ages



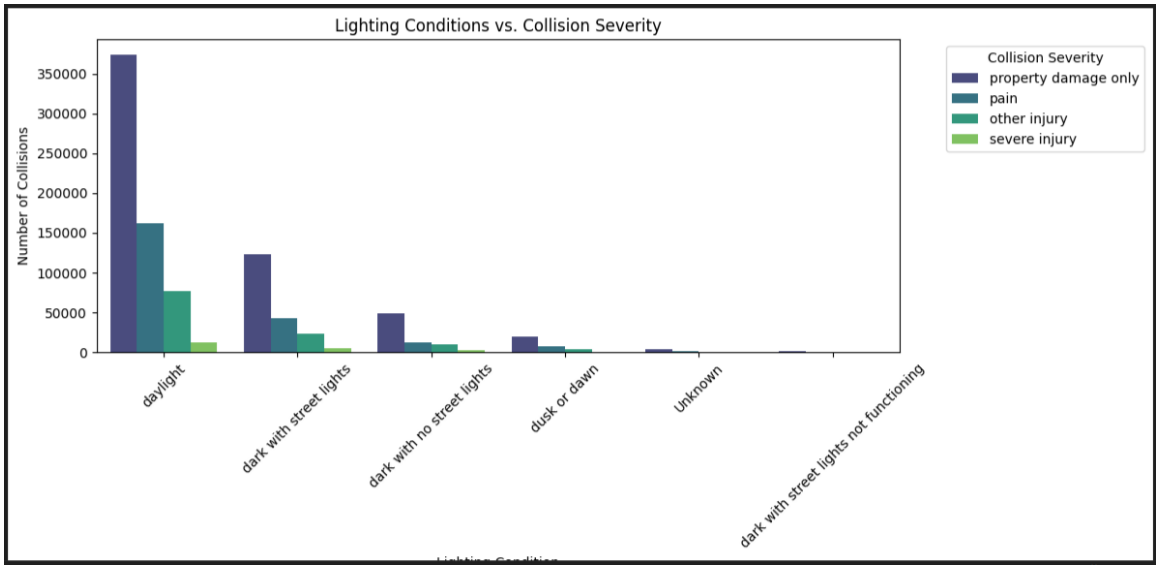
Collisions Severity Vs Total No. Of Victims



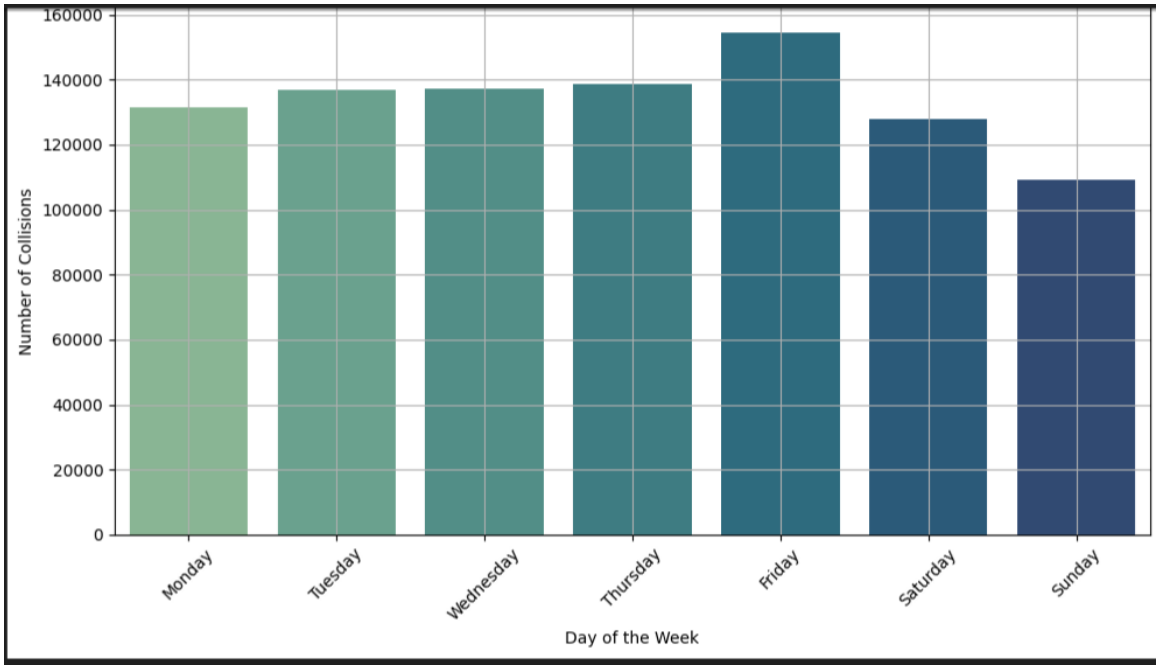
Weather Conditions Vs Collisions Severity



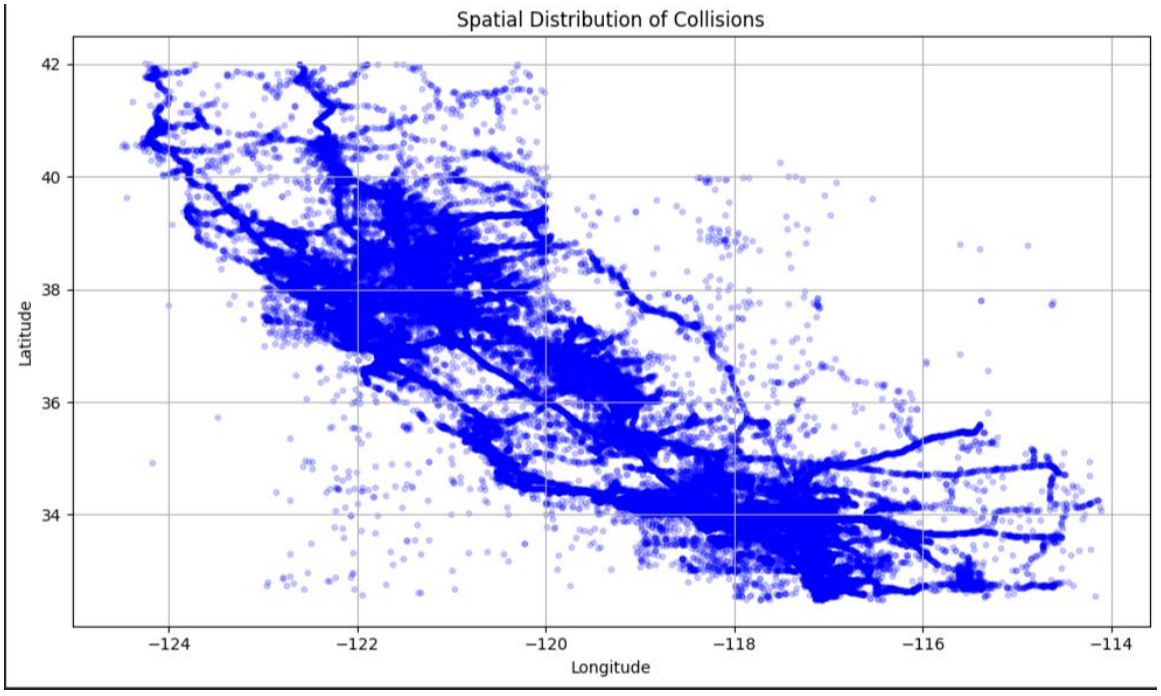
Lightning Conditions Vs Collision Severity



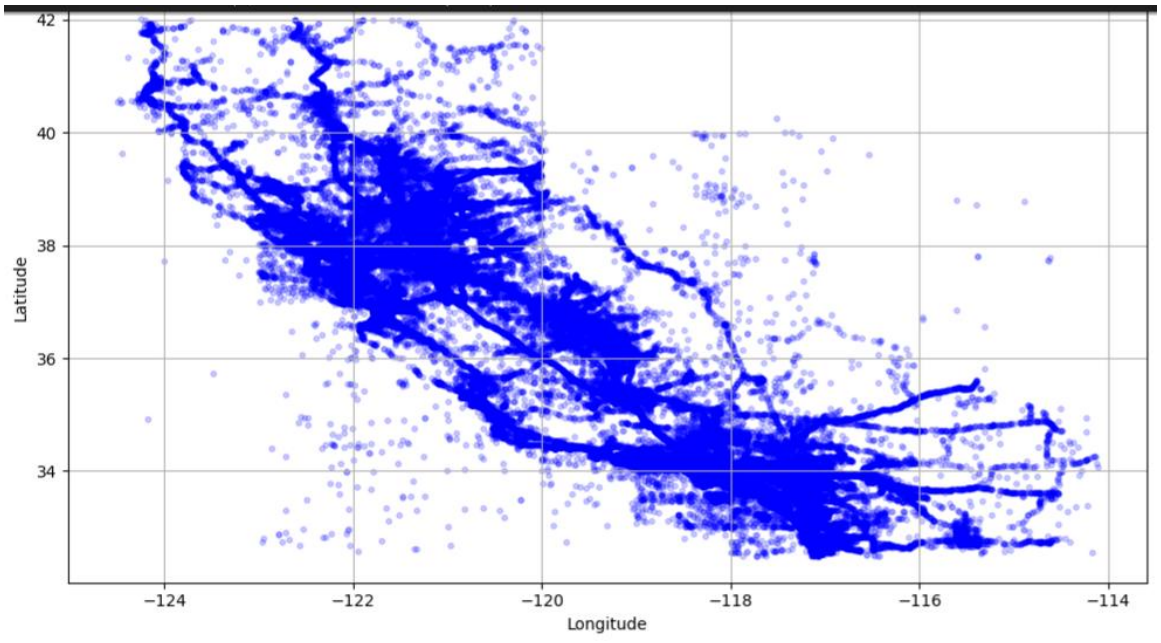
Number Of Collisions Per Week Day



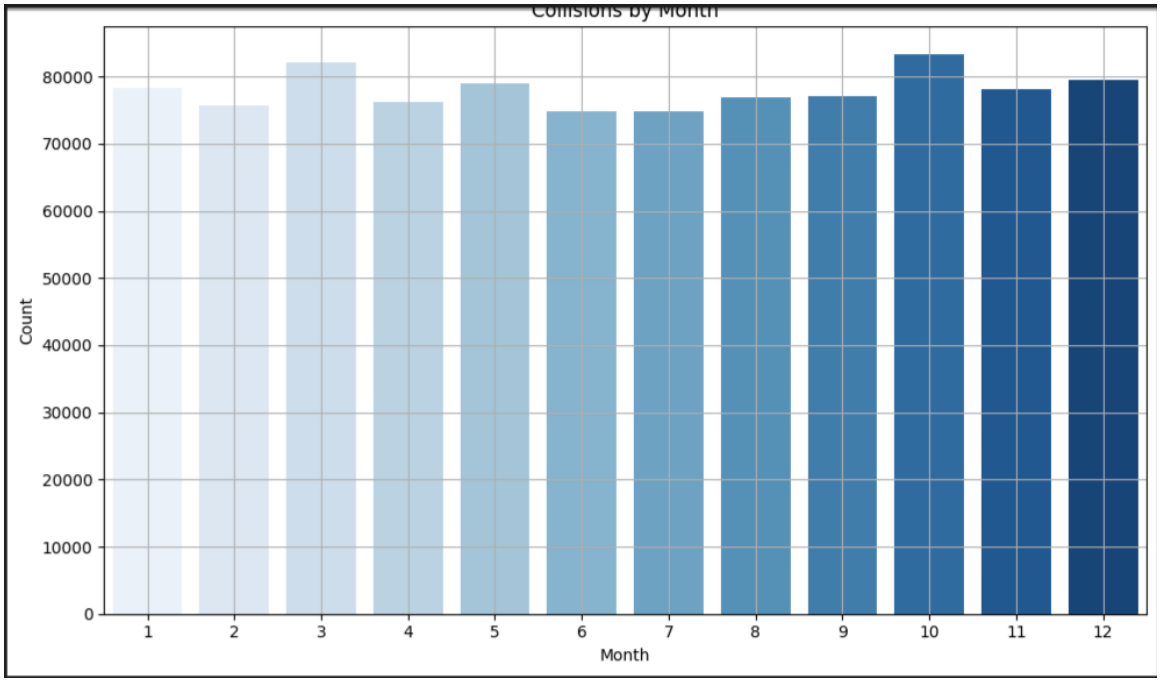
Spatial Distribution Of Collisions



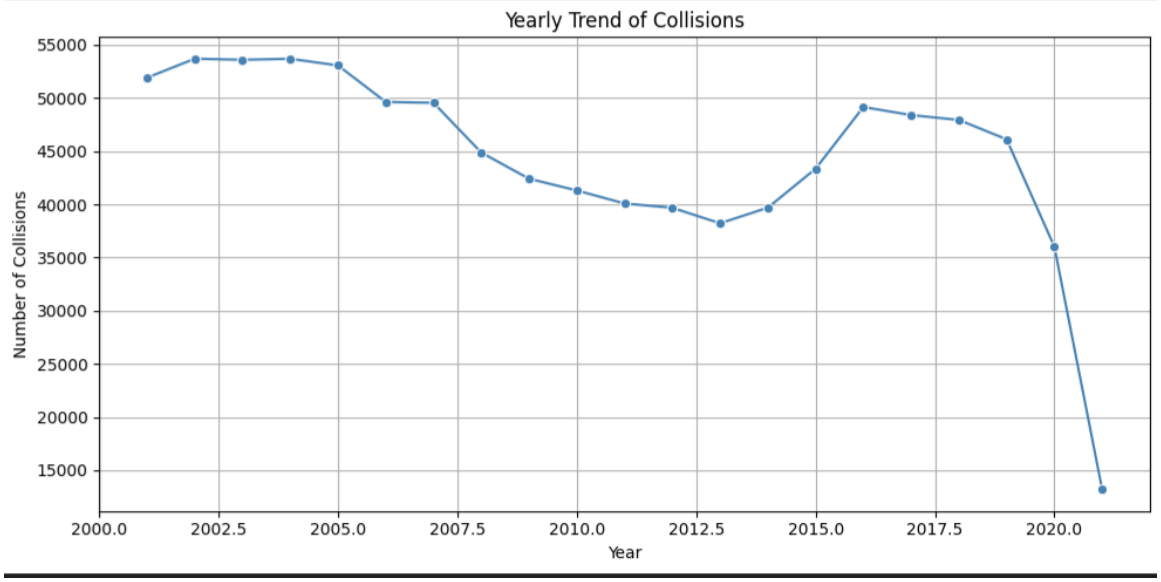
Scatter Plot Of Collisions Locations



Collisions By Month

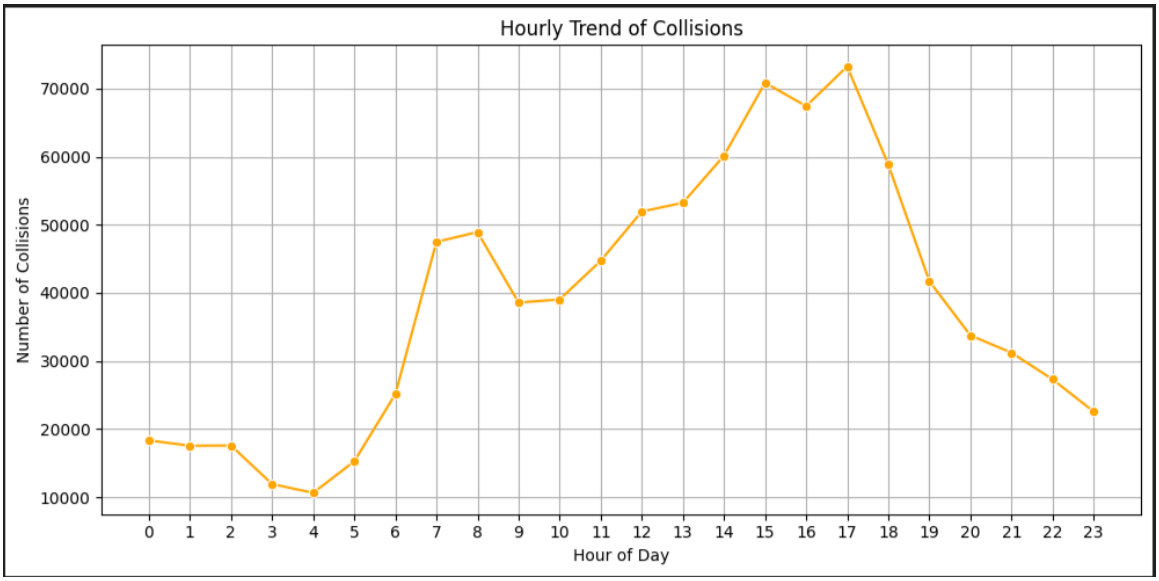


Yearly Trend Of Collisions



Monthly Trend Of Collisions

Hourly Trend Of Collisions



Key Insights

Collision Severity Distribution

- The majority of reported collisions were minor or involved only property damage.
- A small but noteworthy percentage resulted in serious injuries or fatalities.

Temporal Trends

- Collisions peaked during rush hours, especially between **4 PM and 6 PM**.
- **Fridays** saw the highest number of collisions during the week, while **Sundays** had the fewest.

- **October** experienced the highest monthly collision count.

Weather & Road Conditions

- Most collisions occurred during **clear weather**, likely due to increased traffic volumes.
- **Wet or slippery road surfaces** also contributed significantly to collision counts.
- Lighting conditions such as **"Daylight"** and **"Dark – Street Lights On"** were associated with a large number of incidents.

Geographic Trends

- The top five counties—**Los Angeles, Orange, Riverside, San Diego, and San Bernardino**—consistently reported the highest collision numbers.
- Collision hotspots were predominantly located in densely populated **urban areas**, indicating spatial clustering.

Victim Demographics

- Victims spanned a broad age range, but individuals aged **18–35** were most commonly involved.
- Most reported injuries were classified as **complaint of pain**, followed by **visible but not severe injuries**.

Fatal Collisions

- Although fatal collisions represented only **[0]%** of the total, they were disproportionately concentrated during specific **hours of the day** and on certain **road types**.

Recommendations

Enhanced Lighting & Infrastructure

- Improve roadway lighting in high-risk areas with poor visibility, particularly those identified as **"Dark – No Street Lights"**, to reduce nighttime collision risk.

Targeted Law Enforcement During Peak Hours

- Increase law enforcement presence during high-risk periods, especially **late afternoons and weekends**, to deter reckless driving and manage traffic flow.

Public Awareness Campaigns

- Launch educational initiatives aimed at **younger drivers and daily commuters**, emphasizing the dangers of **distracted driving**, especially during **clear weather conditions** when perceived risk is often underestimated.

Prioritize High-Risk Counties

- Allocate safety resources and infrastructure investments to counties with consistently high collision rates, such as **Los Angeles, Orange, Riverside, San Diego, and San Bernardino**.

Predictive Analytics for Proactive Safety

- Leverage the cleaned dataset to develop **machine learning models** capable of identifying emerging high-risk areas, enabling **data-driven interventions** before accidents occur.

Conclusion

Improved road lighting, targeted enforcement during peak hours, and public awareness campaigns can help reduce risk. This cleaned and structured data can also support machine learning models for future risk prediction.

In this ETL project, we effectively **ingested, cleaned, transformed, and analyzed** traffic collision data across multiple dimensions—**temporal patterns, geographic distribution, weather conditions, lighting scenarios**, and **injury severity**. The primary objective was to uncover key risk factors and recurring patterns that contribute to collisions, ultimately generating **actionable insights** to support **data-driven road safety improvements**.