ETL Traffic Collision Analysis Report

---

1. Introduction

This report presents an analytical overview of traffic collision data extracted, transformed, and loaded through an ETL (Extract, Transform, Load) pipeline. The primary objective is to understand when and under what conditions most traffic collisions occur. Using an ETL pipeline, the data was processed and visualized to support meaningful conclusions. Tools and technologies used throughout this project include Python, Pandas, Seaborn, Matplotlib, and AWS EMR for scalable data processing.

Bootstrap Actions Script

Streamlines setup of EMR environments to support efficient data analytics and visualization workflows.

Amazon EMR Studio (especially with **PySpark kernels**) does **not support local visualization (like matplotlib plots)** out of the box. Also, many common Python libraries for data analysis and visualization (e.g., seaborn, pandas, scikit-learn) are not pre-installed on EMR clusters.

To overcome this:

We use a **bootstrap action script** during EMR cluster creation.

```
#!/bin/bash

echo "===== Installing Python packages globally on all EMR nodes ====="

# Upgrade pip
sudo python3 -m pip install --upgrade pip

# Install required Python packages
sudo python3 -m pip install \
    numpy \
    pandas \
    python-dateutil \
    matplotlib \
    Pillow \
    seaborn \
    scikit-learn

echo "===== Python package installation complete ====="
```

📊 Saving and Uploading Plots to Amazon S3

Since direct plot rendering is not supported we have done two thinks.

1. We have Saved plots as images locally on the EMR Master node.
2. Upload the saved image to our S3 bucket for later viewing or use.

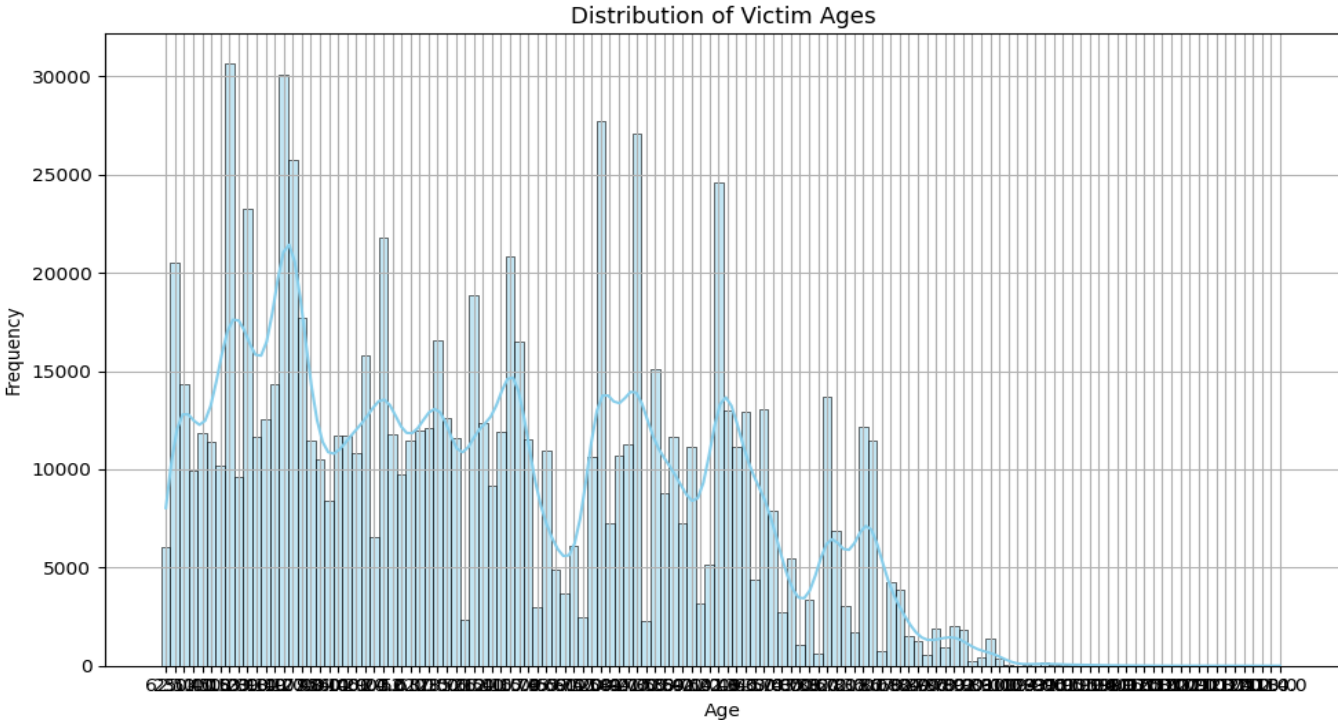
Showing below one of the code from the EDA part where we have saved the image and downloaded to S3 and we have pasted those in this report.

```
# Weather vs. Collision Severity
weather_vs_severity_df = collisions_df.groupBy("weather_1", "collision_severity").count()

# Convert to Pandas
weather_vs_severity_pd = weather_vs_severity_df.toPandas()

# Plot
plt.figure(figsize=(12, 6))
sns.barplot(data=weather_vs_severity_pd, x='weather_1', y='count', hue='collision_severity', palette='viridis')

# Step 4: Title and labels
plt.title("Weather Condition vs. Collision Severity")
plt.xlabel("Weather Condition")
plt.ylabel("Number of Collisions")
plt.xticks(rotation=45)
plt.legend(title="Collision Severity", bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()

# Step 5: Show the plot
plt.show()

# Save plot
output_path = "/tmp/Weather_Condition_vs_Collision_Severity.png"
plt.savefig(output_path)
print(f"Plot saved to {output_path}")

# Upload to S3
import boto3
s3 = boto3.client('s3')
s3.upload_file(output_path, 'traffic-collisions-bucket', 'plots/Weather_Condition_vs_Collision_Severity.png')
```

```
Could not render content for "application/vnd.jupyter.widget-view+json"
{"model_id":"4f88dffabfd14c68bf523ee0813b9f37","version_major":2,"version_minor":0}

Plot saved to /tmp/Weather_Condition_vs_Collision_Severity.png
```

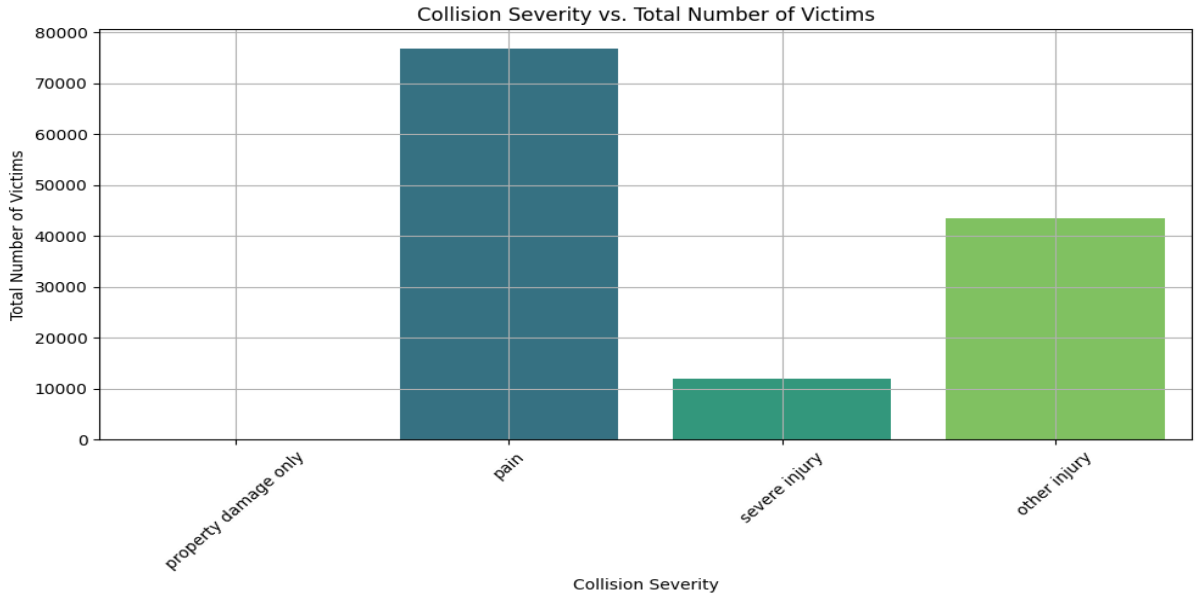
1.Distribution of Victim Ages

age distribution of collision victims, showing that most victims are between 20 and 40 years old.



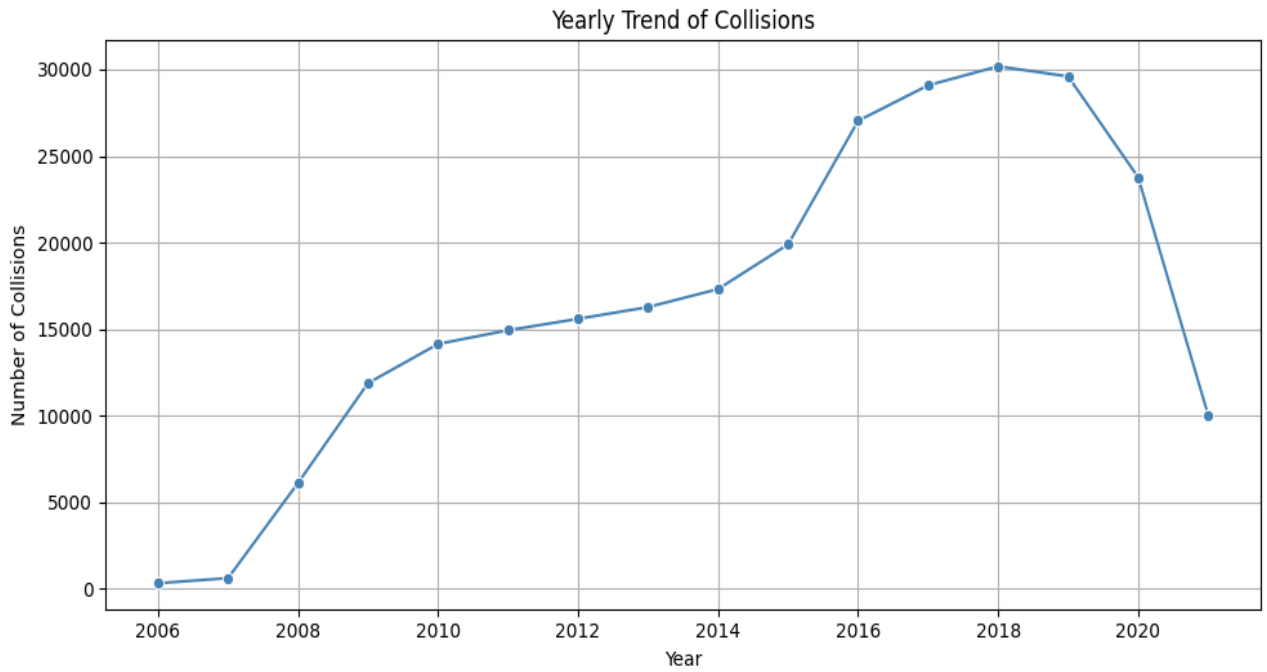
2. Collision Severity vs. Total Number of Victims

This chart visualizes the total number of victims categorized by the severity of the collisions.



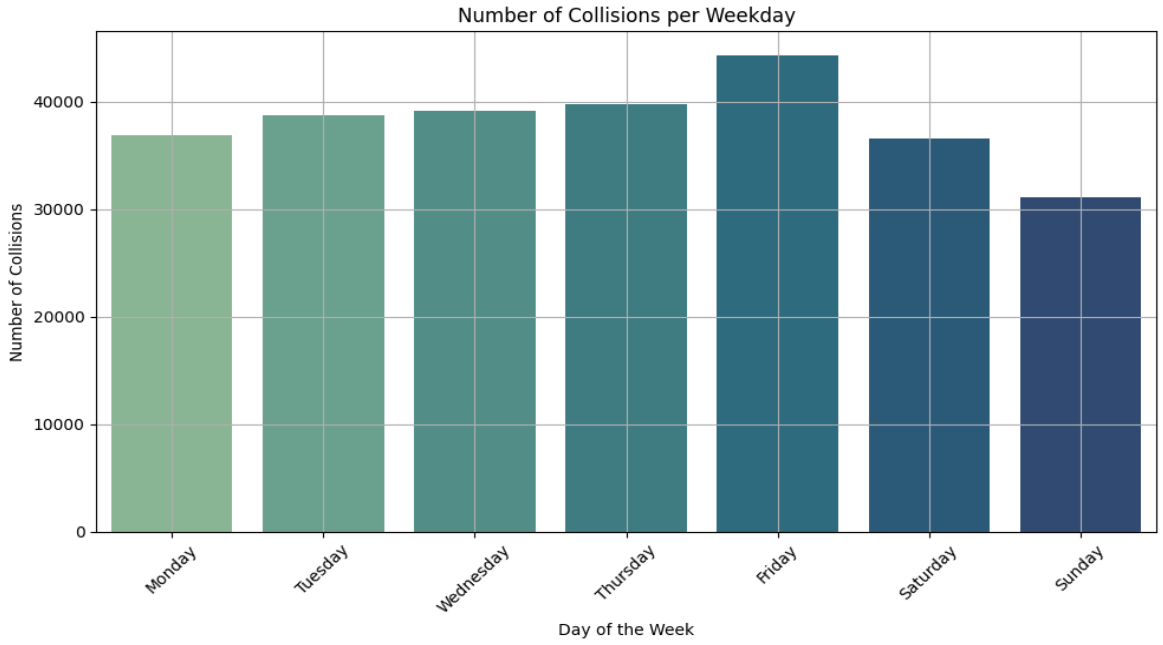
3. Yearly Trend of Collisions

Displays how the number of collisions has changed annually from 2006 to 2021.



4. Number of Collisions per Weekday

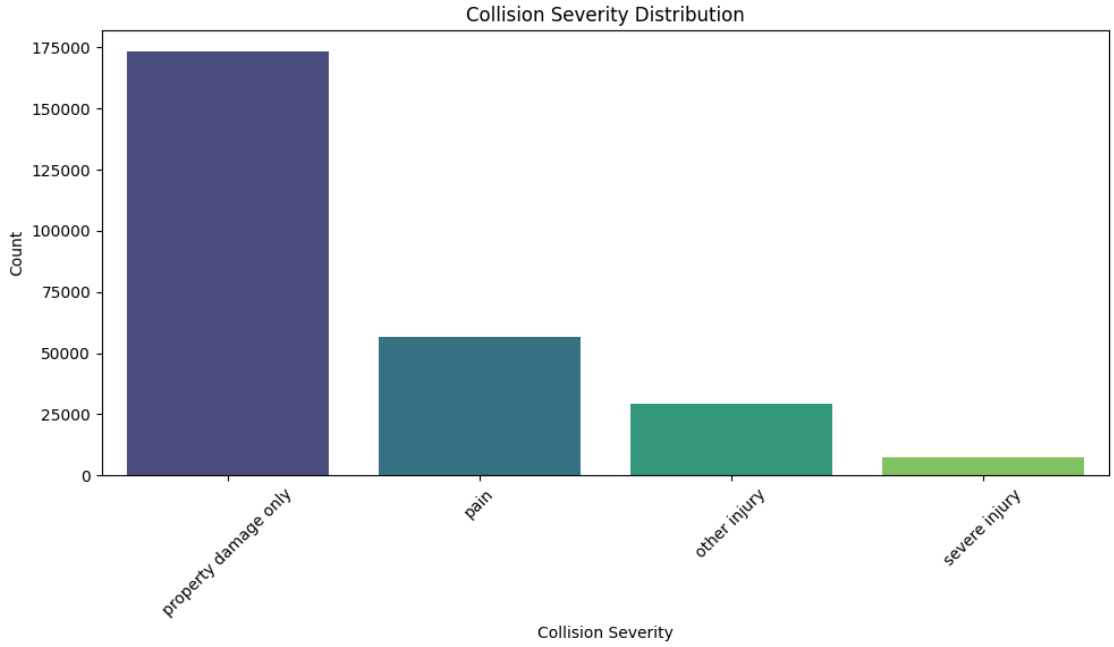
This graph shows the distribution of collisions over different days of the week.



---

5. Collision Severity Distribution

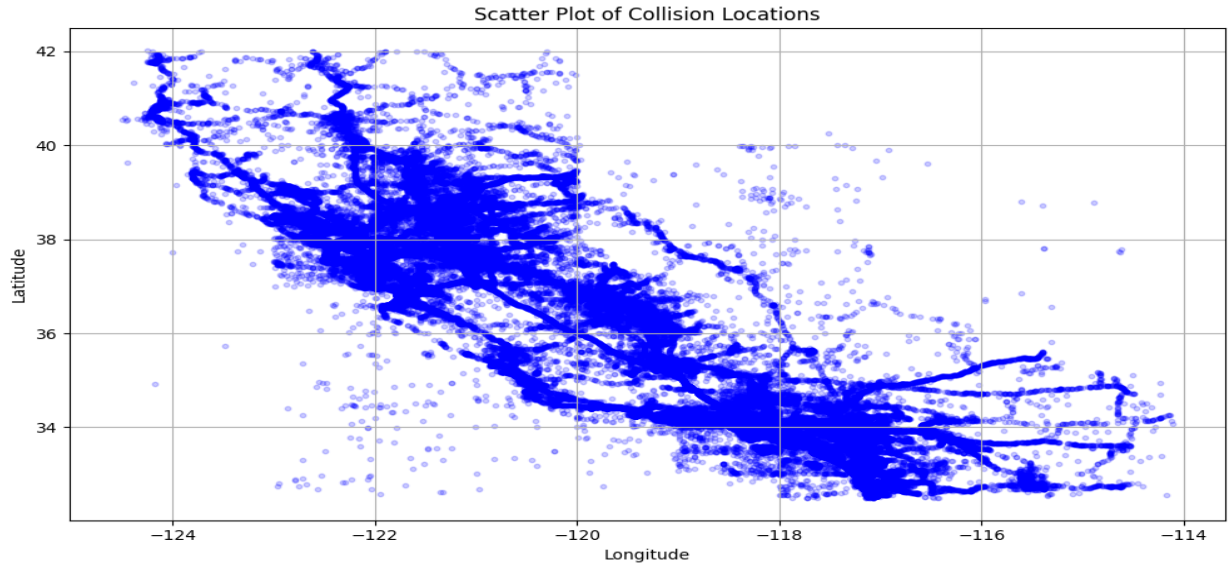
A visual summary of the number of incidents for each level of collision severity.



---

6. Scatter Plot of Collision Locations

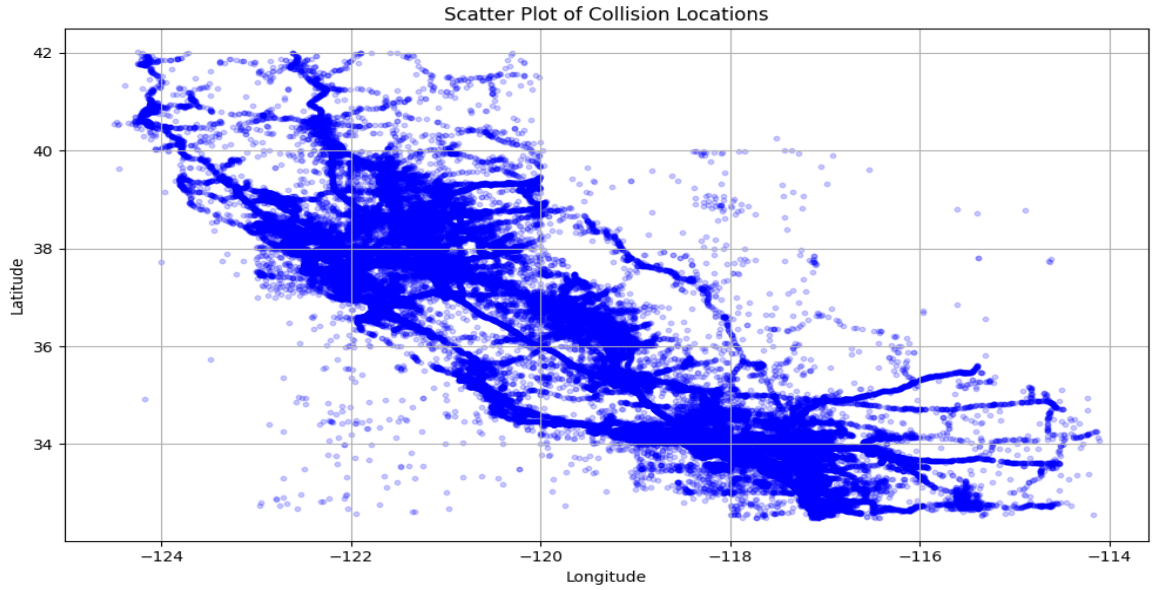
A scatter map of individual collision points, highlighting spatial patterns.



---

7. Spatial Distribution of Collisions

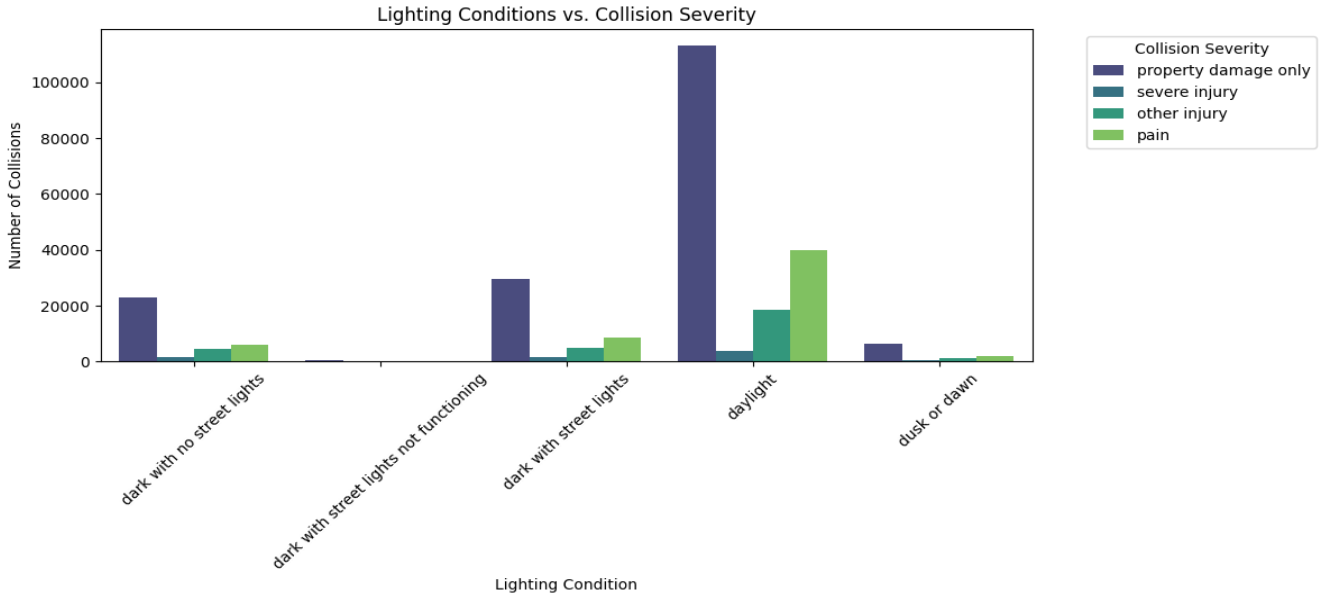
This heatmap-style visual shows collision intensity by area, indicating high-risk zones.



---

8. Lighting Conditions vs. Collision Severity

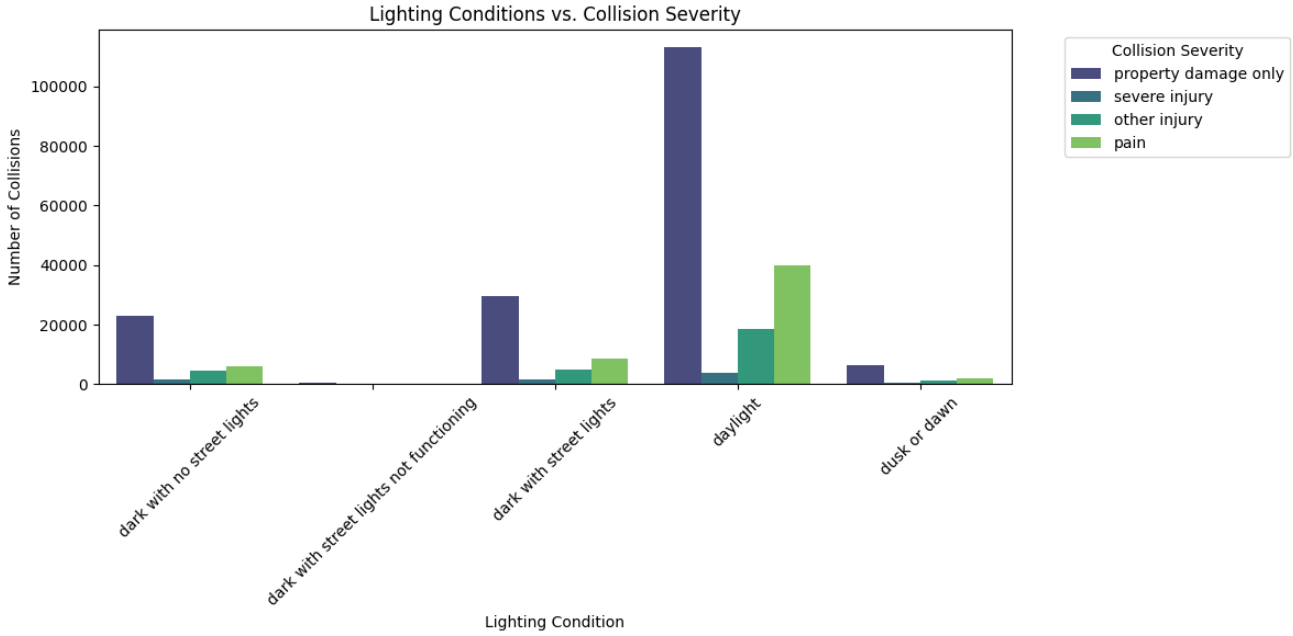
A breakdown of collision severity across different lighting conditions (daylight, dark, dusk, etc.).



---

9. Weather Condition vs. Collision Severity

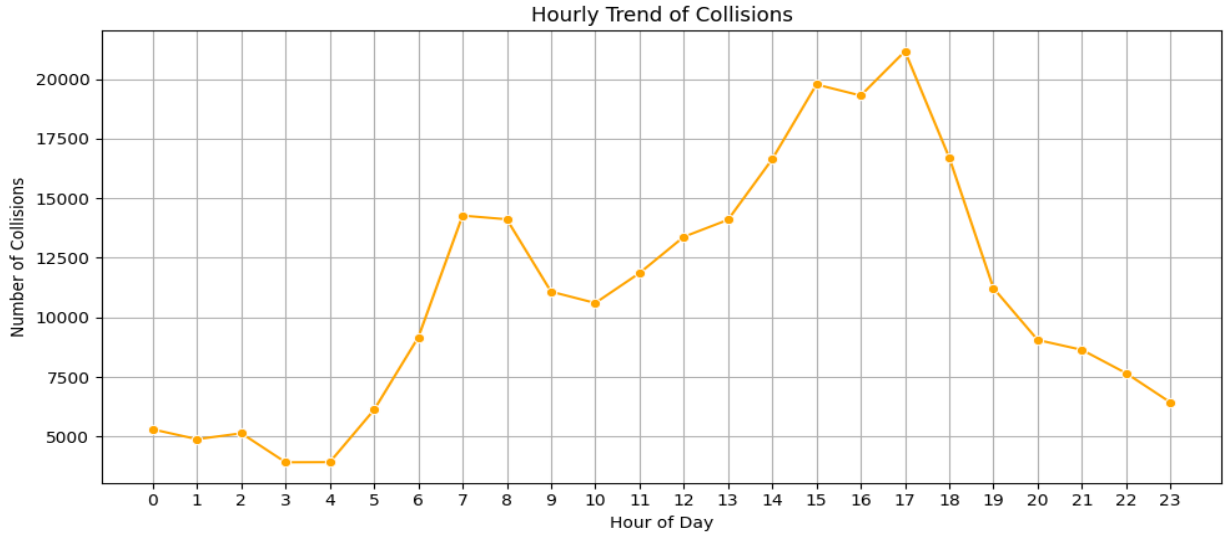
Illustrates how different weather conditions impact collision severity (e.g., clear, rain, fog).



---

10. Hourly Trend of Collisions

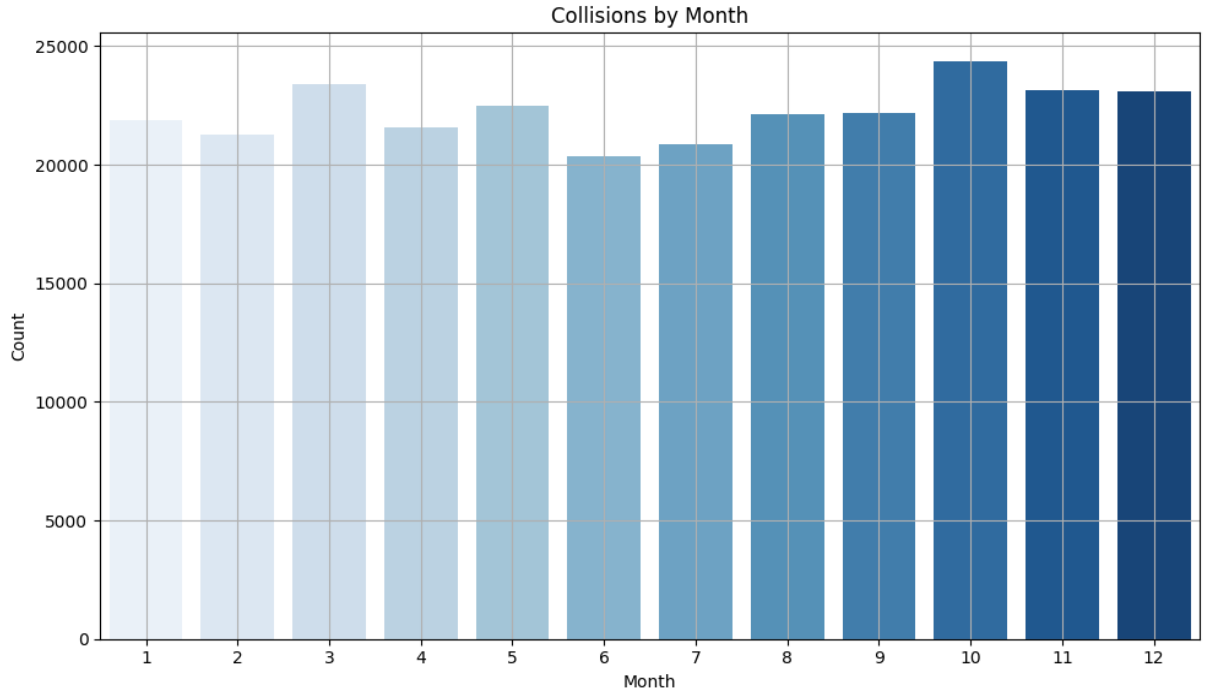
A time-based distribution showing how collisions vary across hours of the day.



---

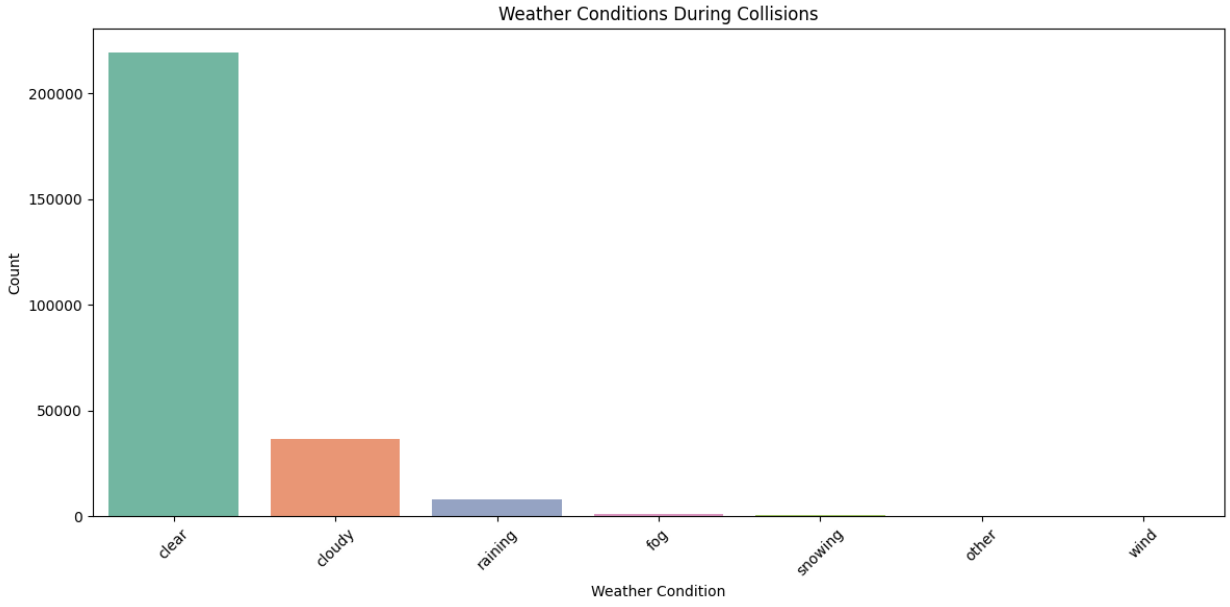
11. Collisions by Month

Collisions rise notably in the final months of the year, highlighting potential seasonal risks.



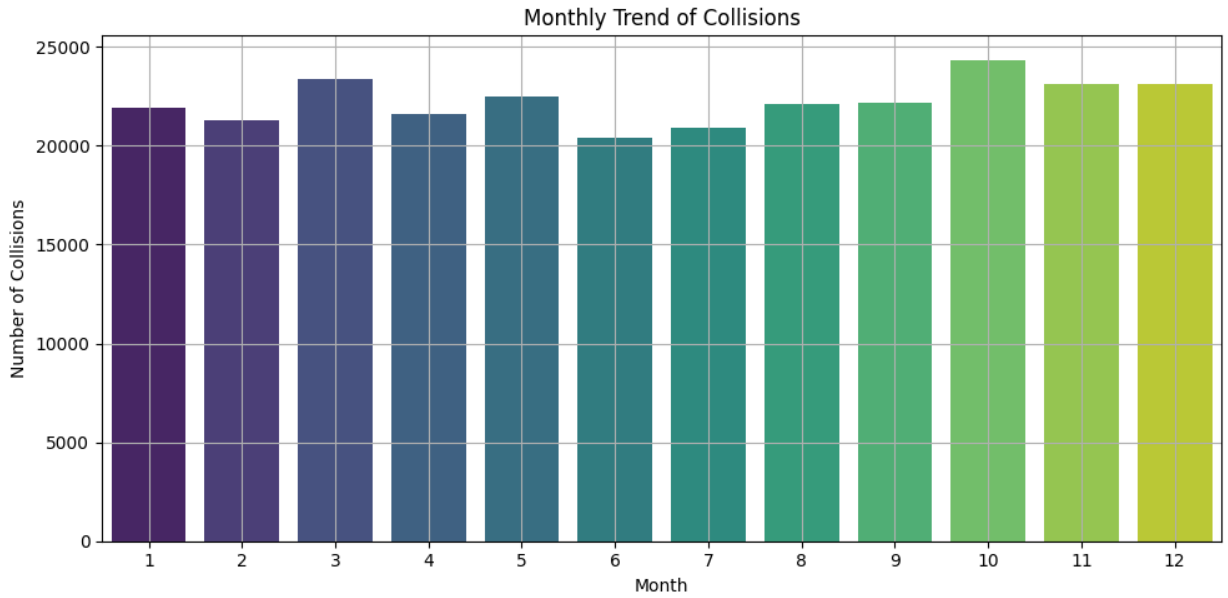
12.Weather Conditions During Collisions

Majority of collisions occur during clear weather, possibly due to higher travel activity rather than weather-related hazards.



13. Monthly Trend of Collisions

Relatively steady collision rates with a peak in October suggest an opportunity to enhance seasonal safety initiatives.



🔍 Key Findings

🚗 Collision Severity Overview

- Most reported incidents were of low severity, frequently resulting in property damage or minor consequences.
- A relatively small portion of accidents led to significant injuries or fatalities, though they remain a crucial area of concern.

🕒 Time-Based Patterns

- Traffic incidents showed a clear rise during peak commuting hours, with the highest occurrence between 4:00 PM and 6:00 PM.
- Fridays experienced the most collisions within the week, while Sundays recorded the least.
- The month of October stood out with the highest volume of reported collisions.

🌧️ Weather and Road Influences

- Clear weather conditions coincided with a high number of collisions, likely tied to elevated traffic density.

- Slippery or wet surfaces also played a notable role in accident frequency.
- Many incidents occurred during well-lit conditions such as daylight or under functioning street lights at night.

🗺️ **Geographic Distribution**

- The five counties with the most frequent collisions were Los Angeles, Orange, Riverside, San Diego, and San Bernardino.
- High incident density was typically observed in urban areas with greater population concentrations, showing clear spatial clustering.

👥 **Demographic Observations**

- Individuals aged between 18 and 35 were most often involved in reported collisions.
- The most commonly documented injuries were complaints of pain, followed by visible but non-critical wounds.

⚠️ **Fatal Incidents**

- Though fatal collisions represented a minimal percentage of all cases, they tended to occur disproportionately during certain hours and on specific types of roadways.

📌 **Recommendations**

💡 **Infrastructure & Lighting Enhancements**

- Improve street lighting in poorly illuminated zones, especially in areas marked as “Dark – No Street Lights,” to help reduce nighttime crash risk.

🚓 **Strategic Law Enforcement**

- Boost law enforcement presence during periods of high traffic risk—particularly late afternoons and weekends—to discourage unsafe driving and ensure better traffic control.

📢 **Driver Education Initiatives**

- Launch targeted awareness programs focused on young drivers and daily commuters, highlighting the underestimated dangers of distraction in seemingly safe (clear weather) conditions.

🚧 **Focus on High-Risk Regions**

- Direct road safety funding and infrastructure development toward counties with consistently high collision rates, such as those in Southern California’s urban core.

🤖 **Predictive Data Modeling**

- Utilize the cleaned and enriched dataset to build machine learning models that can forecast potential accident hotspots and allow for proactive safety interventions.

📜 Summary & Impact

Through improved lighting systems, strategic policing, and educational outreach, collision risks can be significantly mitigated. Additionally, structured datasets from this ETL pipeline enable predictive analytics that support smarter, data-driven road safety planning.

This ETL initiative successfully handled the end-to-end processing of traffic collision data—covering data ingestion, cleansing, transformation, and multi-dimensional analysis. The primary goal was to extract meaningful insights by exploring temporal trends, geographic distribution, environmental factors, and injury levels—ultimately guiding effective and evidence-based safety enhancements.