DATA SCIIENCE USING PTHON

PROJECT REPORT

(Project Semester January-April 2025)

Traffic Accidents Analysis Project

Submitted by:
Priyanshu kumar
Registration no: 12306592
Section: K23EC
Course Code: INT375

Under the Guidance of

Vikas Mangotra

Discipline of CSE/IT

Lovely School of Computer Science Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Priyanshu kumar with Registration no: 12306592 has completed INT375 project titled, **"Traffic Accidents Analysis Project"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Name of the Supervisor :- Vikas Mangotra

Designation of the Supervisor-Asst. Professor

School of Computer Science

Lovely Professional University

Phagwara, Punjab.

Date: 12th April 2025

DECLARATION

I, Priyanshu kumar, student of B.tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12th April, 2025 Signature

Registration No.12306592 Priyanshu Kumar

Acknowledgment

I would like to express my deepest gratitude to Prof. **Vikas Mangotra** for his exceptional mentorship and unwavering support throughout the duration of this project. His vast knowledge in the fields of data science and machine learning, combined with his patient and thoughtful guidance, played a pivotal role in the successful completion of this work. His insightful suggestions and feedback consistently challenged me to think critically and improve the quality of my research. I am also grateful for the learning environment he fostered, which encouraged exploration and innovation.

In addition, I sincerely thank my peers and classmates for their helpful discussions, encouragement, and collaborative spirit during this project. Their input provided fresh perspectives that contributed meaningfully to the final outcome. I am also thankful to the open-source community for providing the tools, libraries, and resources that made the implementation of this project possible. Lastly, I acknowledge the dataset contributors for making this analysis feasible.

Table of Contents

1. Introduction

Road traffic accidents are one of the leading causes of injury and death worldwide, posing serious challenges to public safety and urban planning. With increasing urbanization, population growth, and vehicular density, the frequency of accidents continues to rise — making it crucial to understand the underlying factors that contribute to these incidents.

This project is an effort to analyze a real-world traffic accident dataset using Python. The objective is to extract meaningful insights from the data using Exploratory Data Analysis (EDA) techniques and visualization tools. By applying statistical methods and data visualization libraries such as **Pandas**, **Seaborn**, **Matplotlib**, and **NumPy**, the project investigates how different factors like **weather conditions**, **lighting conditions**, **crash types**, and **damage severity** influence the nature and impact of road accidents.

The dataset contains detailed records of traffic crashes, including attributes such as crash severity, vehicle damage levels, speed zones, environmental factors, and more. This analysis aims to reveal patterns such as:

- Which crash types are most frequent?
- What environmental conditions lead to more severe accidents?
- How does visibility or lighting affect crash frequency and damage levels?

The goal of the project is not only to gain insight into the causes and severity of traffic accidents but also to provide a data-driven foundation that can aid policymakers, urban developers, and traffic authorities in improving road safety strategies. Effective use of such data can potentially reduce the number of accidents and save lives.

This report outlines the steps taken to clean, explore, visualize, and interpret the dataset — transforming raw numbers into actionable knowledge.

2. Dataset Description

The dataset used in this project contains real-world data on traffic accidents. Each row represents a unique crash incident, with various attributes describing the environmental conditions, type of crash, damage levels, and other relevant factors. Key features in the dataset include:

- **Dataset Summary:**
  - **File Name:** traffic_accidents.csv
  - **Format:** CSV (Comma Separated Values)
  - **Total Rows:** 209307
  - **Total Columns:** 24

The dataset provides both categorical and numerical values, making it suitable for various forms of statistical analysis and visualization techniques. It forms a strong foundation for identifying meaningful patterns and actionable insights in traffic safety.

3. Source of Dataset

The dataset used for this project was sourced from the official **Data.gov** platform — a U.S. government open data portal that provides access to high-quality, publicly available datasets from various domains including transportation, healthcare, education, climate, and more.

Specifically, this dataset is part of a traffic crash reporting system maintained by regional or national transportation authorities. It includes detailed records of motor vehicle crashes, covering aspects such as crash severity, weather and lighting conditions, time and location, and reported damages or injuries.

🔸 **About Data.gov:**

- **Managed By:** U.S. General Services Administration (GSA)
- **Goal:** To promote transparency, innovation, and data-driven solutions by making government-collected data openly accessible.
- **URL:** https://www.data.gov
- **Licensing:** Most datasets are released under open licenses, permitting public use for analysis, research, and commercial purposes.

The use of this dataset ensures that the project is based on real-world, authentic data collected through standardized reporting methods. This enhances the reliability of insights derived from the analysis and supports evidence-based recommendations in traffic safety and urban planning.

🔍 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in any data science or analytical project. It involves summarizing the main characteristics of the dataset, often with the help of visual methods, to uncover hidden patterns, detect anomalies, test hypotheses, and check assumptions.

In this project, EDA was performed to:

- Understand the structure and distribution of the dataset
- Identify missing or inconsistent values
- Analyze the relationships between various features (e.g., crash severity and weather)
- Discover trends and patterns in traffic accidents under different conditions

🛠 Techniques Used:

- Descriptive Statistics: To get an overview of data types, value counts, means, and distributions.
- Missing Value Analysis: Checked for nulls or inconsistent entries and handled them accordingly.
- Univariate Analysis: Examined single features using bar plots, histograms, and KDE plots to understand distributions.
- Bivariate Analysis: Explored relationships between variables (e.g., Crash Type vs. Severity) using scatter plots and box plots.
- Categorical Analysis: Analyzed frequency of crash types, weather conditions, damage levels, etc. using countplots and pie charts.
- Correlation Matrix: To identify numeric feature correlations for potential multivariate insights.

📊 Tools & Libraries Used:

- Pandas for data manipulation and cleaning
- Matplotlib and Seaborn for data visualization
- NumPy for numerical operations

The EDA phase helped in building an intuition about the data and guided the direction for further analysis, visualization, and conclusions

4. Analysis on Dataset

i. Introduction

After performing Exploratory Data Analysis (EDA), several important observations and patterns were identified in the traffic accident dataset:

- **Key Insights:**

1. **Crash Severity Trends:**
    - Minor crashes were the most frequent, followed by major and fatal crashes.
    - A noticeable portion of severe crashes occurred during night or poor lighting conditions.
2. **Weather Conditions:**
    - Most crashes occurred under **clear weather**, suggesting human error plays a major role even in ideal conditions.
    - Rainy and foggy weather also contributed significantly to crash counts, often with higher severity.
3. **Lighting Conditions:**
    - **Daylight** was the most common lighting condition for crashes, likely due to high traffic volume.
    - However, **crashes during darkness** (especially without street lights) tended to have **higher injury and fatality rates**.
4. **Crash Types:**
    - **Rear-end and side-impact crashes** were the most common.
    - Head-on collisions, though less frequent, showed a higher fatality rate.
5. **Damage Levels:**
    - Most vehicles suffered **moderate to low damage**, but severe damages correlated strongly with high-speed zones and nighttime crashes.
6. **Temporal Patterns:**
    - A spike in crash frequency was observed during **peak traffic hours**, particularly **morning and evening**.
    - **Fridays and weekends** showed a higher number of incidents, indicating lifestyle and travel patterns.
7. **Speed Zones:**
    - Higher speed zones saw more severe crashes, emphasizing the need for speed regulation and enforcement in such areas.

1. **Data Visualization Techniques**
    - **Histograms** for distribution analysis
    - **Scatter plots** to identify relationships between two numerical variables
    - **Heatmaps** for correlation analysis
    - **Boxplots** to compare popularity distributions across genres
    - **Pie charts** for categorical analysis

5. Conclusion

The analysis of the traffic accident dataset revealed several valuable insights into the factors contributing to road crashes. Through systematic Exploratory Data Analysis (EDA), we uncovered patterns related to crash severity, weather conditions, lighting, and vehicle damage levels.

Key findings indicate that even under ideal weather and lighting, human behavior and traffic density play a major role in accidents. Moreover, crashes during nighttime and in high-speed zones tend to be more severe, highlighting the importance of infrastructure improvements and enforcement of speed regulations.

This project not only helped in understanding the underlying trends in traffic incidents but also demonstrated the power of data analysis and visualization in solving real-world problems. The insights derived from this dataset can be used by authorities to design smarter road safety measures, optimize traffic management systems, and ultimately save lives.

Moving forward, integrating more datasets such as driver behavior, vehicle type, or real-time traffic data could enhance the depth of analysis and lead to more comprehensive solutions.

6. Future Scope

This project lays the groundwork for deeper and more actionable traffic accident analysis. However, there are several directions in which this work can be extended to gain even more meaningful insights:

- **1. Integration with Real-Time Data**
  - Incorporating real-time traffic, weather, and GPS data can enable dynamic crash prediction and live monitoring systems.
- **2. Advanced Predictive Modeling**
  - Using machine learning models such as decision trees, random forests, or neural networks to predict crash severity or probability based on environmental and situational factors.
- **3. Geospatial Analysis**
  - Mapping accident locations using GIS tools to identify high-risk zones or accident hotspots and support urban planning.
- **4. Driver Behavior Analysis**
  - Including factors such as driver age, experience, intoxication, or distraction levels (e.g., mobile phone usage) for behavior-based crash analysis.
- **5. Policy Impact Evaluation**
  - Analyzing the effectiveness of traffic policies (e.g., new speed limits, stricter fines) over time using before-and-after datasets.
- **6. Public Awareness Dashboards**
  - Creating interactive dashboards to present accident trends to the public and policymakers in an intuitive format using tools like Plotly, Dash, or Power BI.

[11]:
```
import pandas as pd
df = pd.read_csv('traffic_accidents.csv')
df.head(10)
```

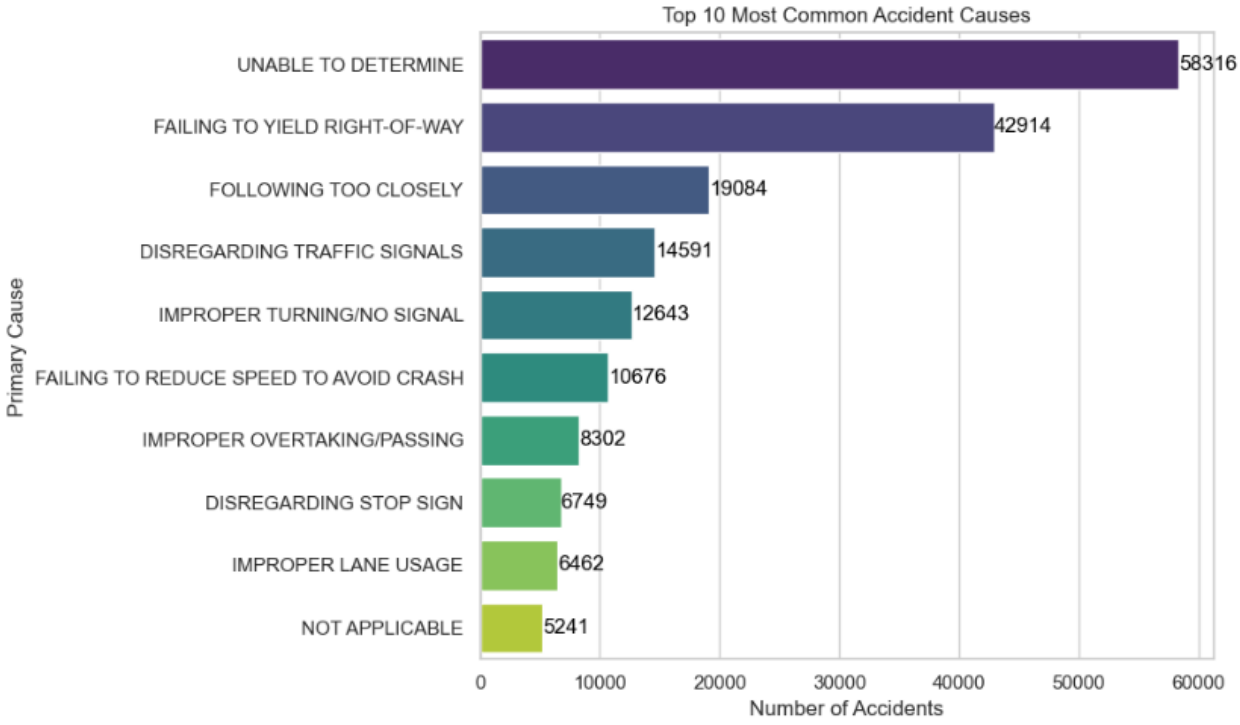
[11]:

	crash_date	traffic_control_device	weather_condition	lighting_condition	first_crash_type	trafficway_type	alignment	roadway_surface_cond	road_defect	crash_type
0	07/29/2023 01:00:00 PM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	TURNING	NOT DIVIDED	STRAIGHT AND LEVEL	UNKNOWN	UNKNOWN	NO INJURY / DRIVE AWAY
1	08/13/2023 12:11:00 AM	TRAFFIC SIGNAL	CLEAR	DARKNESS, LIGHTED ROAD	TURNING	FOUR WAY	STRAIGHT AND LEVEL	DRY	NO DEFECTS	NO INJURY / DRIVE AWAY
2	12/09/2021 10:30:00 AM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	REAR END	T-INTERSECTION	STRAIGHT AND LEVEL	DRY	NO DEFECTS	NO INJURY / DRIVE AWAY
3	08/09/2023 07:55:00 PM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	ANGLE	FOUR WAY	STRAIGHT AND LEVEL	DRY	NO DEFECTS	INJURY AND / OR TOW DUE TO CRASH
4	08/19/2023 02:55:00 PM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	REAR END	T-INTERSECTION	STRAIGHT AND LEVEL	UNKNOWN	UNKNOWN	NO INJURY / DRIVE AWAY
5	09/06/2023 12:59:00 AM	NO CONTROLS	RAIN	DARKNESS, LIGHTED ROAD	FIXED OBJECT	NOT DIVIDED	STRAIGHT AND LEVEL	WET	UNKNOWN	INJURY AND / OR TOW DUE TO CRASH
6	12/20/2022 11:45:00 AM	TRAFFIC SIGNAL	CLEAR	DAYLIGHT	REAR TO FRONT	FOUR WAY	STRAIGHT AND LEVEL	DRY	NO DEFECTS	NO INJURY / DRIVE AWAY
7	09/20/2023 02:38:00 PM	NO CONTROLS	CLEAR	DAYLIGHT	ANGLE	DIVIDED - W/MEDIAN (NOT RAISED)	CURVE, LEVEL	DRY	NO DEFECTS	INJURY AND / OR TOW DUE TO CRASH

[7]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Setting seaborn style
sns.set(style="whitegrid")
plt.rcParams['figure.figsize'] = (10, 6)

# Load the dataset
df = pd.read_csv('traffic_accidents.csv')

# Display basic information
print("Dataset Info:\n")
print(df.info())
print("\nFirst 5 rows:\n")
print(df.head())

# Check for missing values
print("\nMissing values:\n")
print(df.isnull().sum())
```

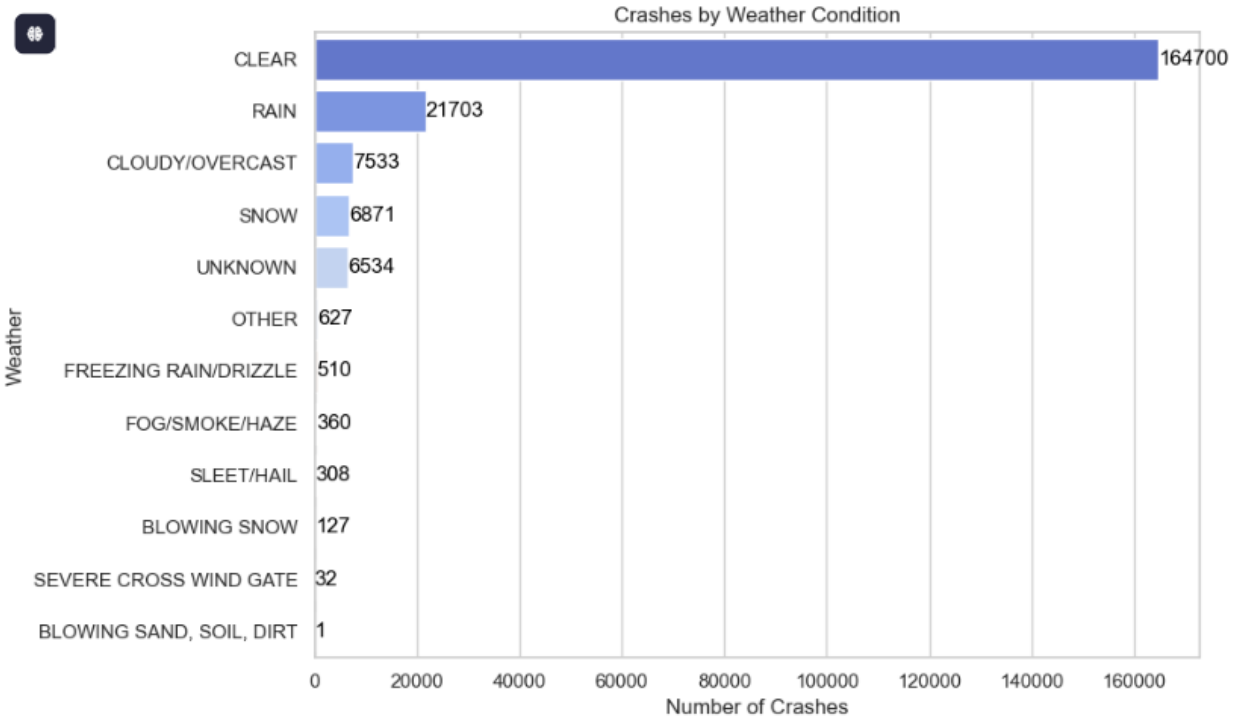
```
 1   traffic_control_device         209306 non-null  object
 2   weather_condition              209306 non-null  object
 3   lighting_condition             209306 non-null  object
 4   first_crash_type               209306 non-null  object
 5   trafficway_type                209306 non-null  object
 6   alignment                      209306 non-null  object
 7   roadway_surface_cond           209306 non-null  object
 8   road_defect                    209306 non-null  object
 9   crash_type                     209306 non-null  object
 10  intersection_related_i         209306 non-null  object
 11  damage                         209306 non-null  object
 12  prim_contributory_cause        209306 non-null  object
 13  num_units                      209306 non-null  int64
 14  most_severe_injury             209306 non-null  object
 15  injuries_total                 209306 non-null  float64
 16  injuries_fatal                 209306 non-null  float64
 17  injuries_incapacitating        209306 non-null  float64
 18  injuries_non_incapacitating    209306 non-null  float64
 19  injuries_reported_not_evident  209306 non-null  float64
```

[9]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# ------------------------------
# Objective 1: Most Common Accident Cause
# ------------------------------
cause_counts = df['prim_contributory_cause'].value_counts().head(10)
sns.barplot(x=cause_counts.values, y=cause_counts.index, hue=cause_counts.index, dodge=False, palette='viridis', legend=False)

plt.title('Top 10 Most Common Accident Causes')
plt.xlabel('Number of Accidents')
plt.ylabel('Primary Cause')
for i, v in enumerate(cause_counts.values):
    plt.text(v + 10, i, str(v), color='black', va='center')
plt.tight_layout()
plt.show()
```

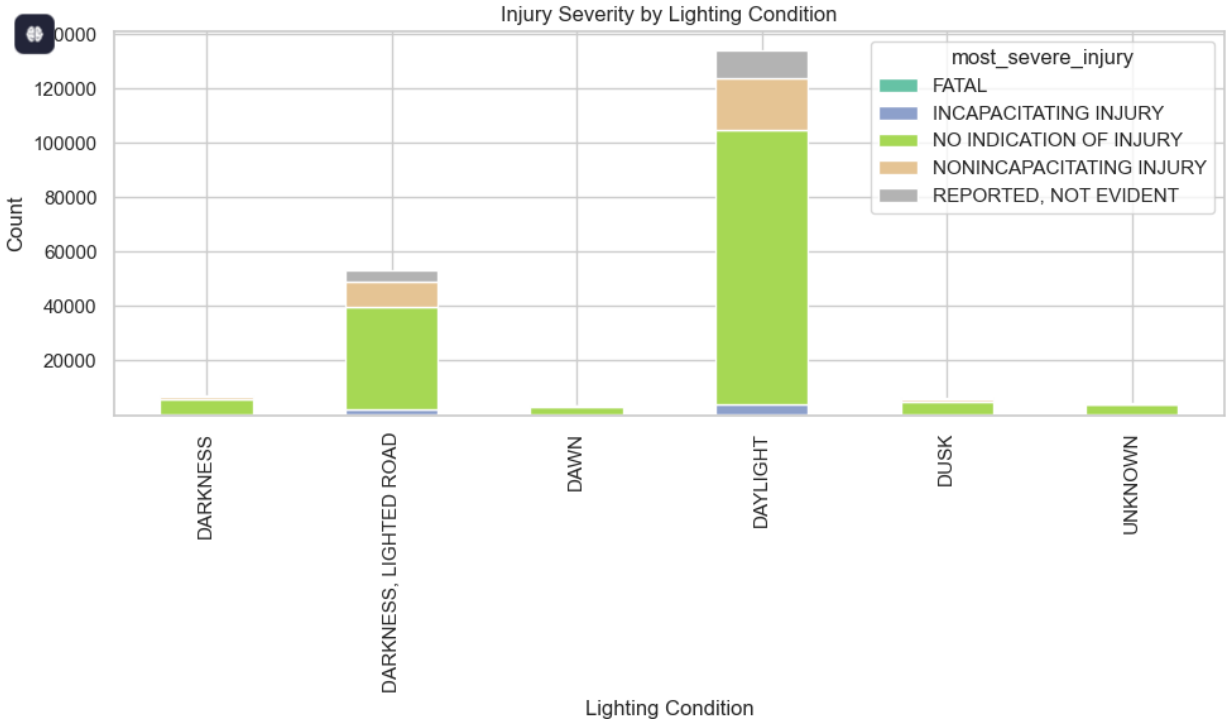


Number of Accidents

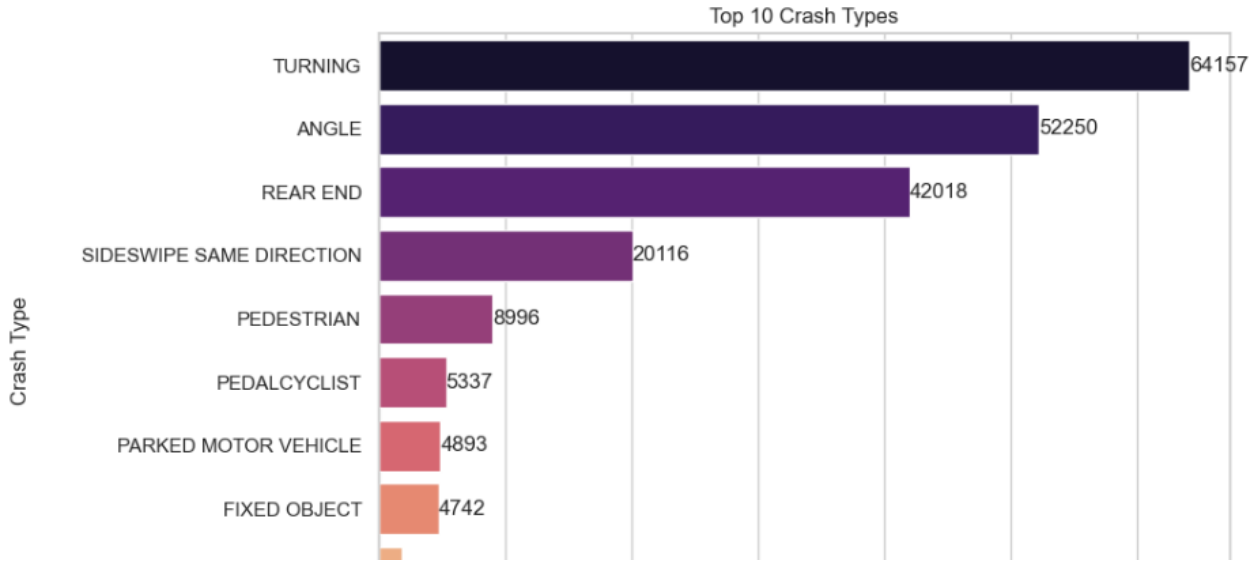
[4]:
```
# Objective 2: Crashes in Different Weather
# ------------------------------
weather_crashes = df['weather_condition'].value_counts()
sns.barplot(x=weather_crashes.values, y=weather_crashes.index, hue=weather_crashes.index, dodge=False, palette='coolwarm', legend=False)

plt.title('Crashes by Weather Condition')
plt.xlabel('Number of Crashes')
plt.ylabel('Weather')
for i, v in enumerate(weather_crashes.values):
    plt.text(v + 10, i, str(v), color='black', va='center')
plt.tight_layout()
plt.show()
```



[6]:
```
# Objective 3: Injuries by Lighting Condition
# ------------------------------
injury_lighting = df.groupby('lighting_condition')['most_severe_injury'].value_counts().unstack().fillna(0)
injury_lighting.plot(kind='bar', stacked=True, colormap='Set2')
plt.title('Injury Severity by Lighting Condition')
plt.xlabel('Lighting Condition')
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```

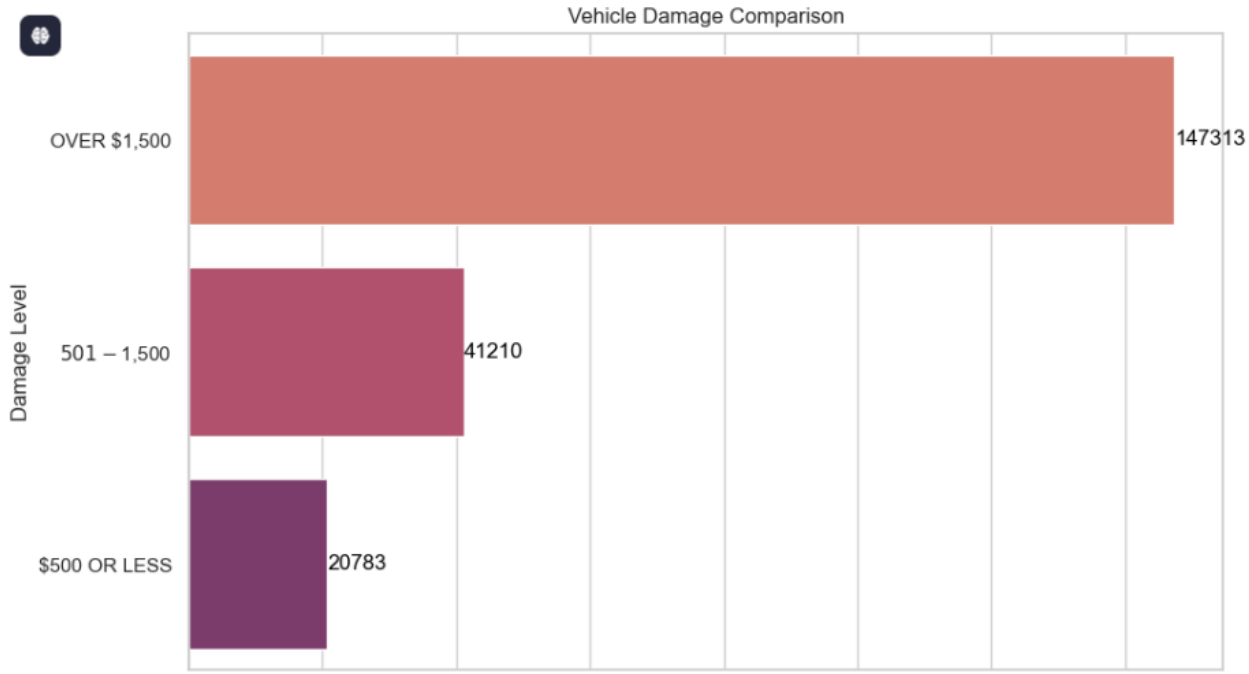


```
crash_type_counts = df['first_crash_type'].value_counts().head(10)

sns.barplot(
    x=crash_type_counts.values,
    y=crash_type_counts.index,
    hue=crash_type_counts.index,
    dodge=False,
    palette='magma',
    legend=False
)
plt.title('Top 10 Crash Types')
plt.xlabel('Frequency')
plt.ylabel('Crash Type')

# Add value labels
for i, v in enumerate(crash_type_counts.values):
    plt.text(v + 10, i, str(v), va='center')

plt.tight_layout()
plt.show()
```



Frequency

[15]:
```
# Objective 5: Damage Comparison
# ------------------------------
damage_counts = df['damage'].value_counts()
sns.barplot(x=damage_counts.values, y=damage_counts.index, hue=damage_counts.index, palette='flare', dodge=False, legend=False)

plt.title('Vehicle Damage Comparison')
plt.xlabel('Number of Crashes')
plt.ylabel('Damage Level')

for i, v in enumerate(damage_counts.values):
    plt.text(v + 10, i, str(v), color='black', va='center')
plt.tight_layout()
plt.show()
```



```python
# Select only numerical columns
numeric_df = df.select_dtypes(include=['float64', 'int64'])

# Calculate the correlation matrix
correlation_matrix = numeric_df.corr()

# Set the figure size and plot the heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Heatmap of Numerical Features Correlation")
plt.show()
```