Chicago Crashes *Analysis*

Data-Driven Insights for Informed Decision-Making on Road Safety Policies and effect resource Allocation.

Presented by: [Peter Nyakundi]

Overview

The project is designed to reduce accidents and improve road safety by using in-depth analysis of the Chicago Car Crashes dataset. This will help identify high-risk locations, and understand contributing factors, thereby enabling more effective allocation of resources and informed decision-making to the stakeholders.

Business Understanding

The Chicago Department of Transportation, the Chicago Police Department, and Public Health and Emergency Services, as well as urban planners and community advocacy groups. Have tasked us to analyze the Chicago traffic dataset and come up with insights to help in making good decision in terms of policies, resource allocation on effective way to reduce accidents and improve road safety

The project seeks to improve road safety and reduce accidents. It categorizes crashes by severity (minor, injury-causing, fatal) to uncover conditions that lead to more severe outcomes. Time trend analysis reveals when accidents peak (e.g., rush hours, weekends), while additional evaluations focus on how weather, lighting, road conditions, and driver behaviors contribute to crash frequency and severity. Geospatial mapping of crash hotspots further pinpoints areas needing targeted safety interventions, guiding recommendations for infrastructure enhancements, traffic regulation adjustments, and public awareness campaigns.

Objectives

Goal: Provide actionable Insights for Informed Decision-Making on Road Safety Policies and effect resource Allocation.

Key Questions:

- Examine the severity of accidents and injuries?
- Analyze time trends across various scales—hours, weeks, months, and years?
- Analyze the impact of weather and lighting conditions on crash occurrences?
- Examining driver behavior?
- Analysis also extends to road conditions?
- Geospatial analysis of crash locations?

Data Understanding

- This project uses data from One primary sources:

https://www.google.com/url?q=https%3A%2F%2Fdata.cityofchicago.org%2Fd%2Faerh-rz74

- **Traffic_Crashes_-_Crashes_20250218.csv**: Contains data on Chicago trashes It has 48 columns and 109094 rows

The Dataset has **mixed data types**, including **numerical data** (e.g., injury counts, speeding), **categorical data** (e.g., crash severity levels, contributing factors), **datetime data** (e.g., crash date and time), and **geospatial data** (e.g., latitude and longitude coordinates).

Limitation

- The dataset had a lot of missing values on some columns
- The dataset was a large. Making it difficult to model using KNN which was very slow

Methodology

Steps;

- Data collection

  Data was stored and managed using csv file ensuring structured and efficient querying.

- Data cleaning and preprocessing

  Pandas was used for data manipulation, cleaning, and transformation

- Exploratory data analysis

  Seaborn, Tableau and Matplotlib were used to generate insightful visualizations to identify

  patterns and trends.

- .Modelling

  Logistic Regression, Decision Tree, Random Forest, Ensemble and Xgboost ,Lightgbm

Implementations:

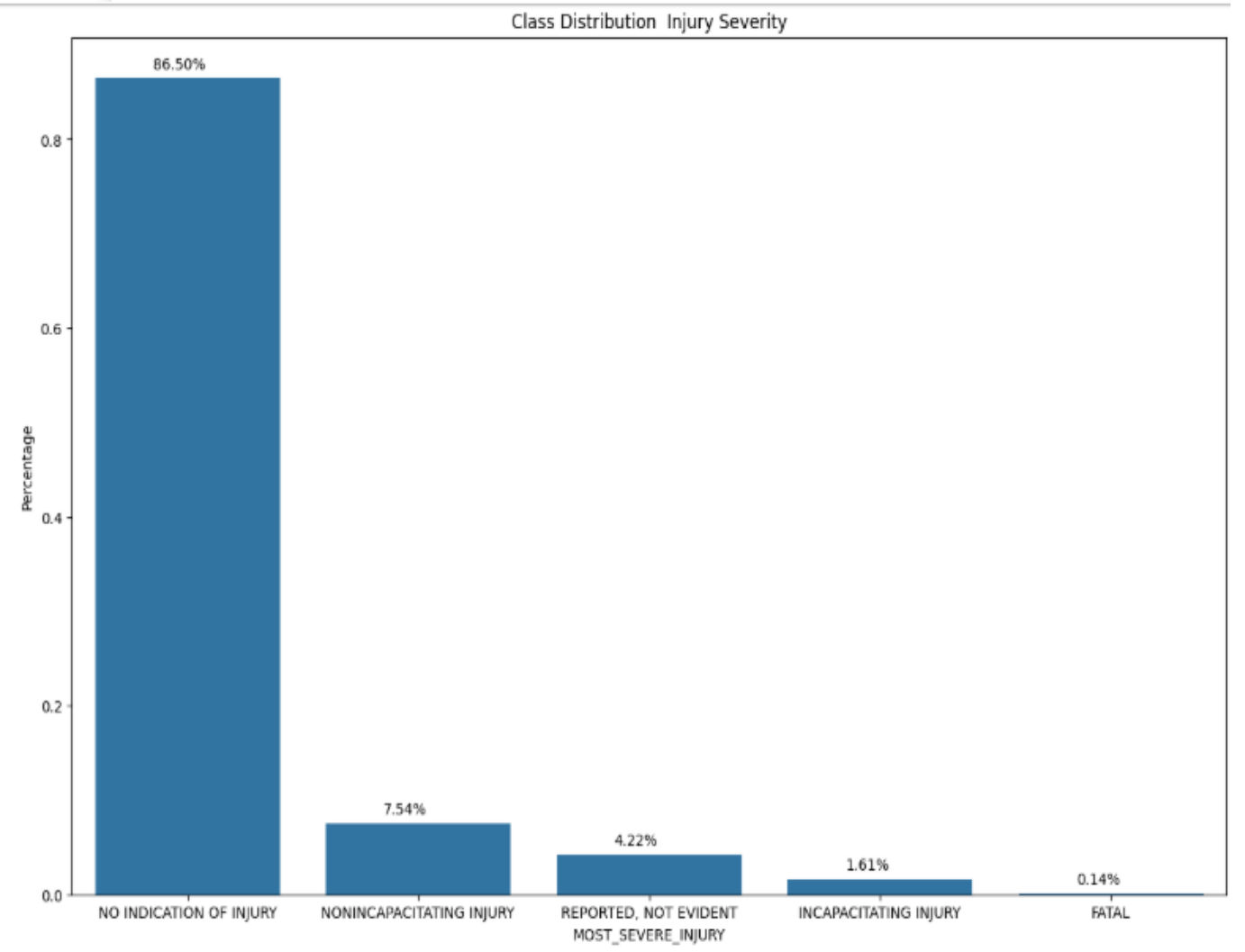
The entire workflow was implemented in Jupyter Notebook and Tableau worksheets, allowing interactive data exploration and visualization

Class Distribution of Injury Severity

- **Key Insights**

The bar chart shows the distribution of injury severity in different categories. Most cases (86.5%) have no indication of injury, while 7.54% involve non-incapacitating injuries. About 4.22% are reported injuries with no evident symptoms, 1.61% are incapacitating injuries, and only 0.14% are fatal.

This means the majority of incidents result in no injuries, while severe injuries and fatalities are rare. This means our data is highly imbalance to see the fatal accidents which we are interested.

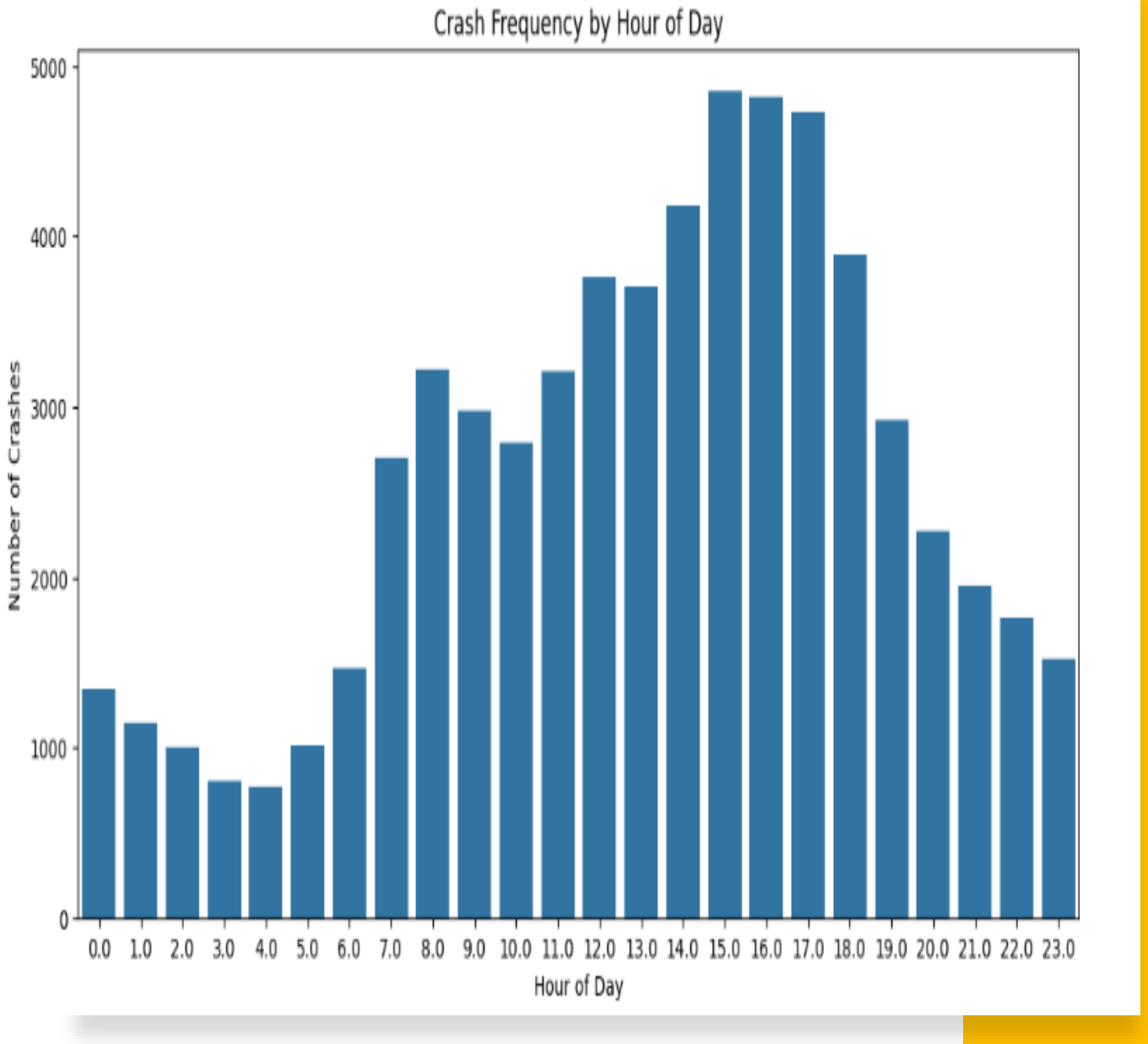


Crash Frequency by Hour of Day

Insights

15:00 – 17:00 had the highest Crashes

- **hours** (e.g., 8 AM, 5 PM) This makes sense because there's more traffic during these times .Because this times are Rush Hours were people go to Work place or it is time they are coming out of work place.

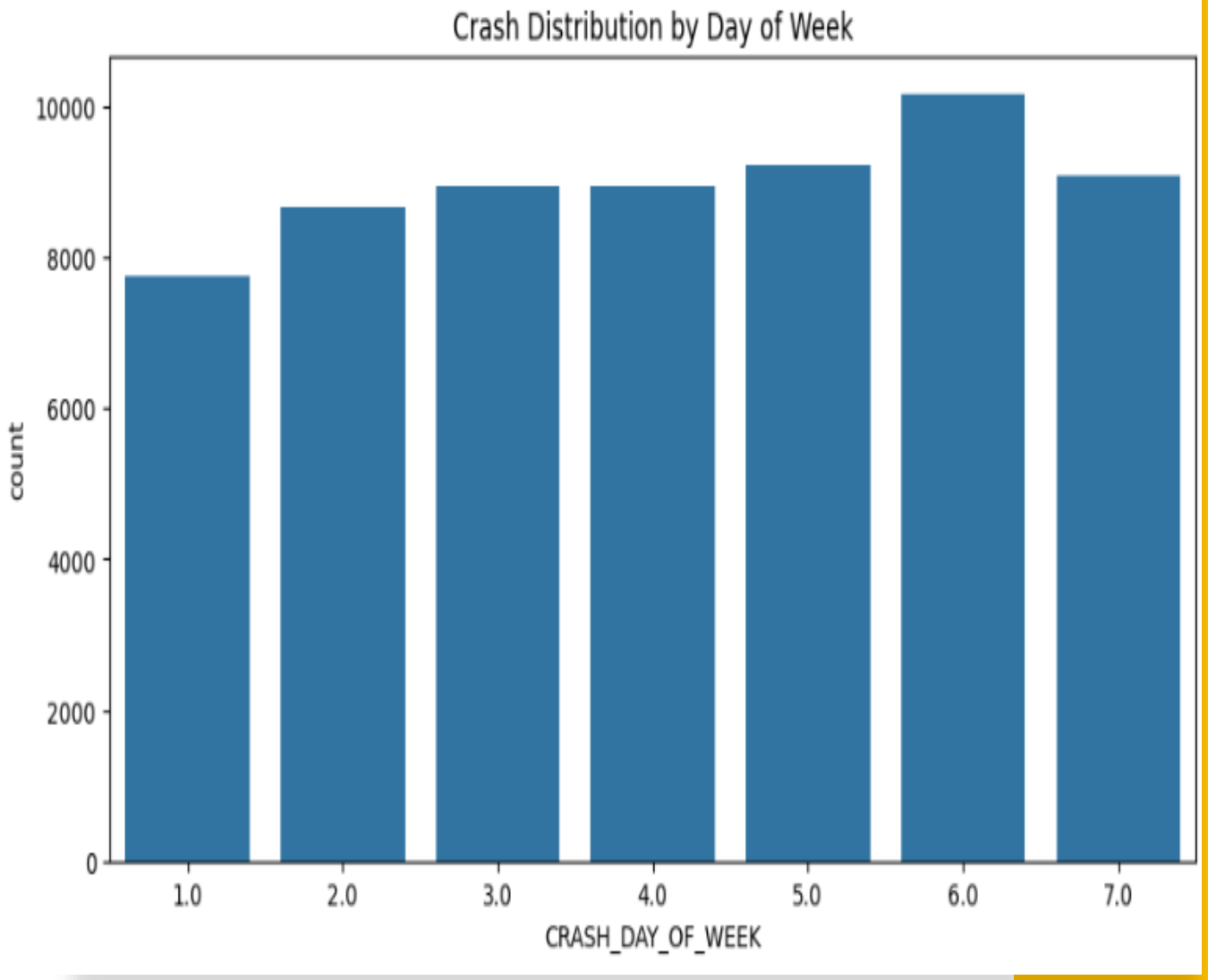


Crash Frequency by Day of Week

Insights

Friday had the highest Crashes

- **Daily**: The highest number of crashes occurs on Fridays and Saturdays, possibly due to increased social activities or traffic. Mostly going to weekend.

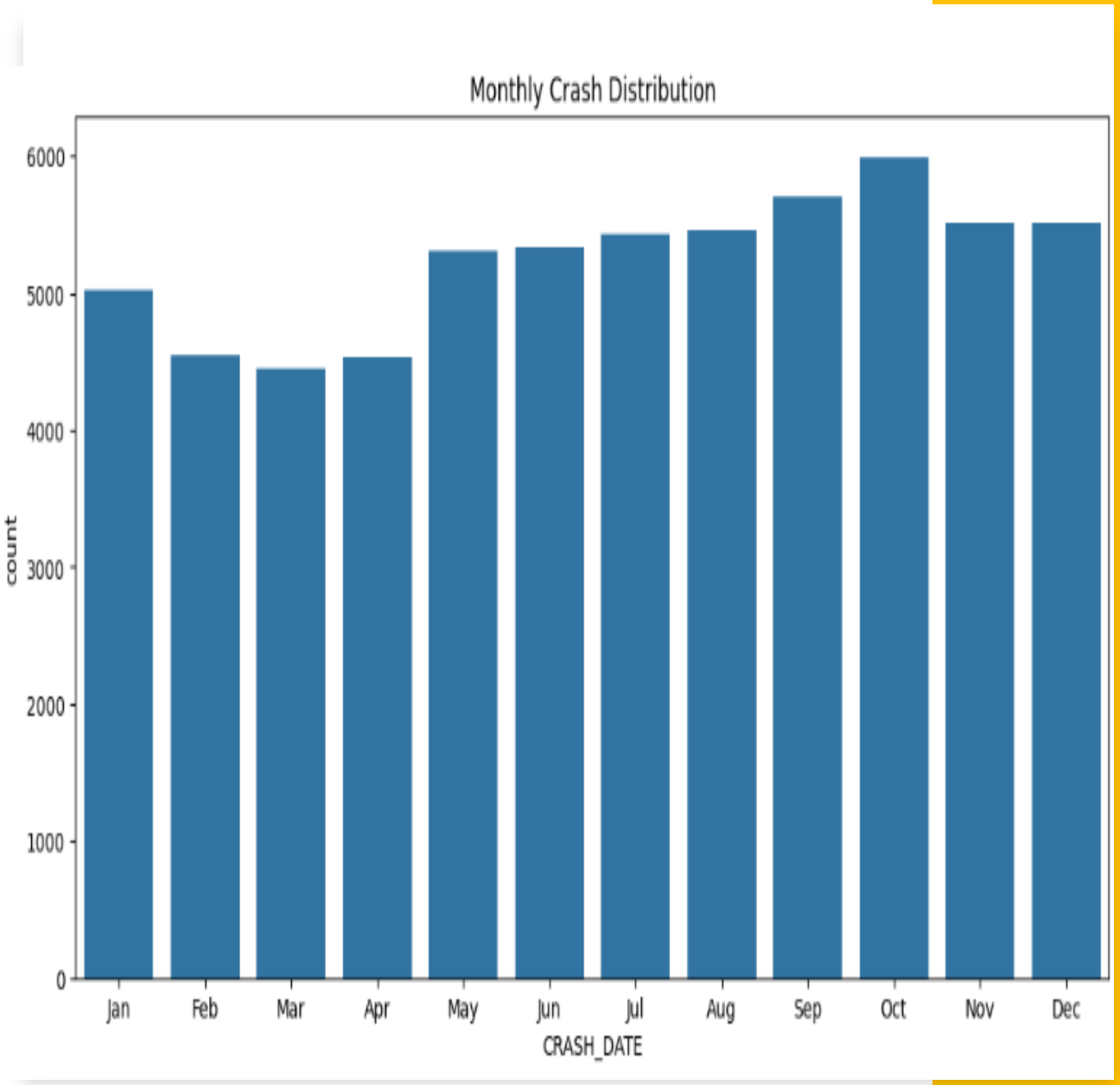


Crash Frequency by Month

Key Insights

October Had the Highest Crashes

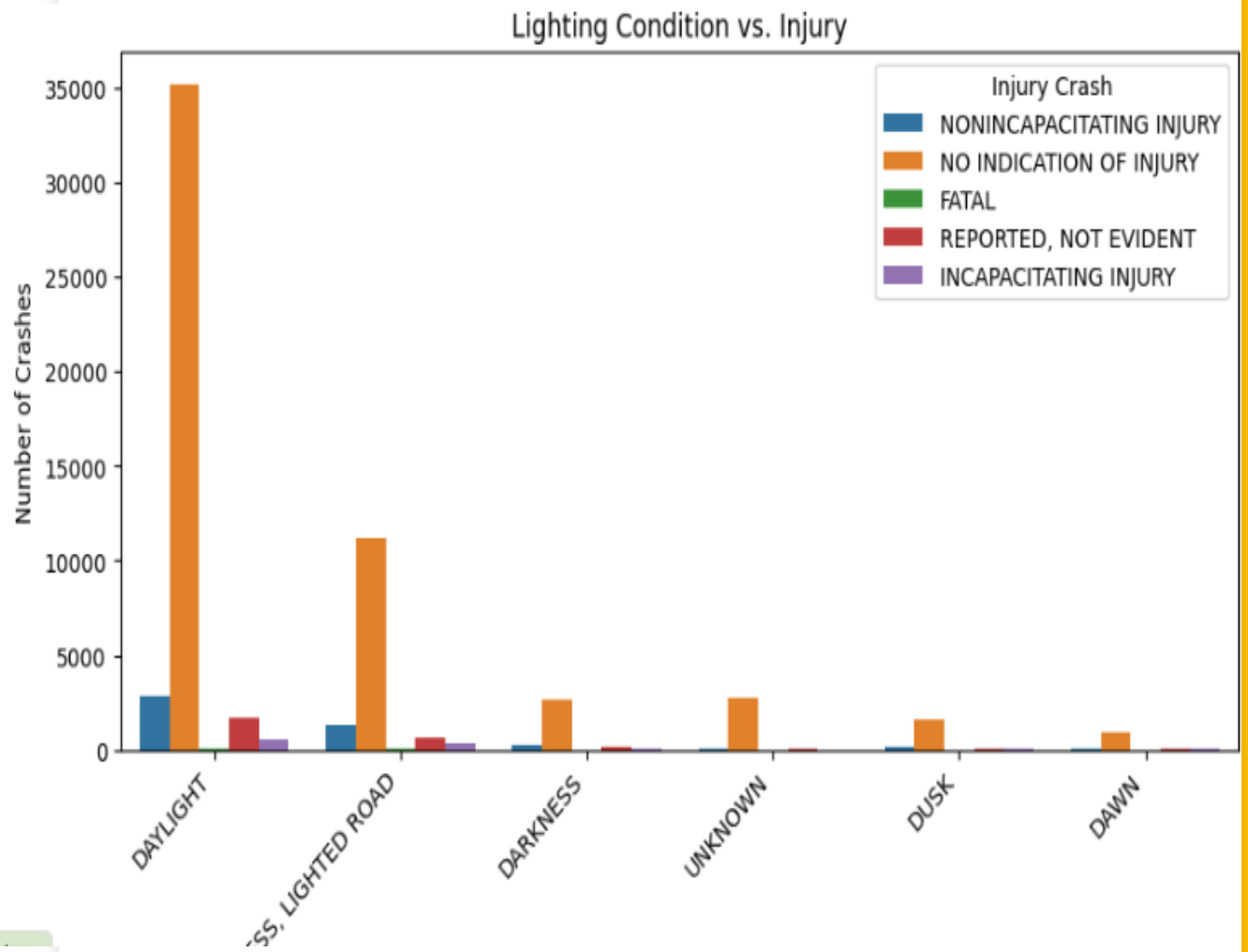
Crashes might be higher in winter months (September to February) due to snow and ice.



Lighting Condition VS Injuries

Key Insights

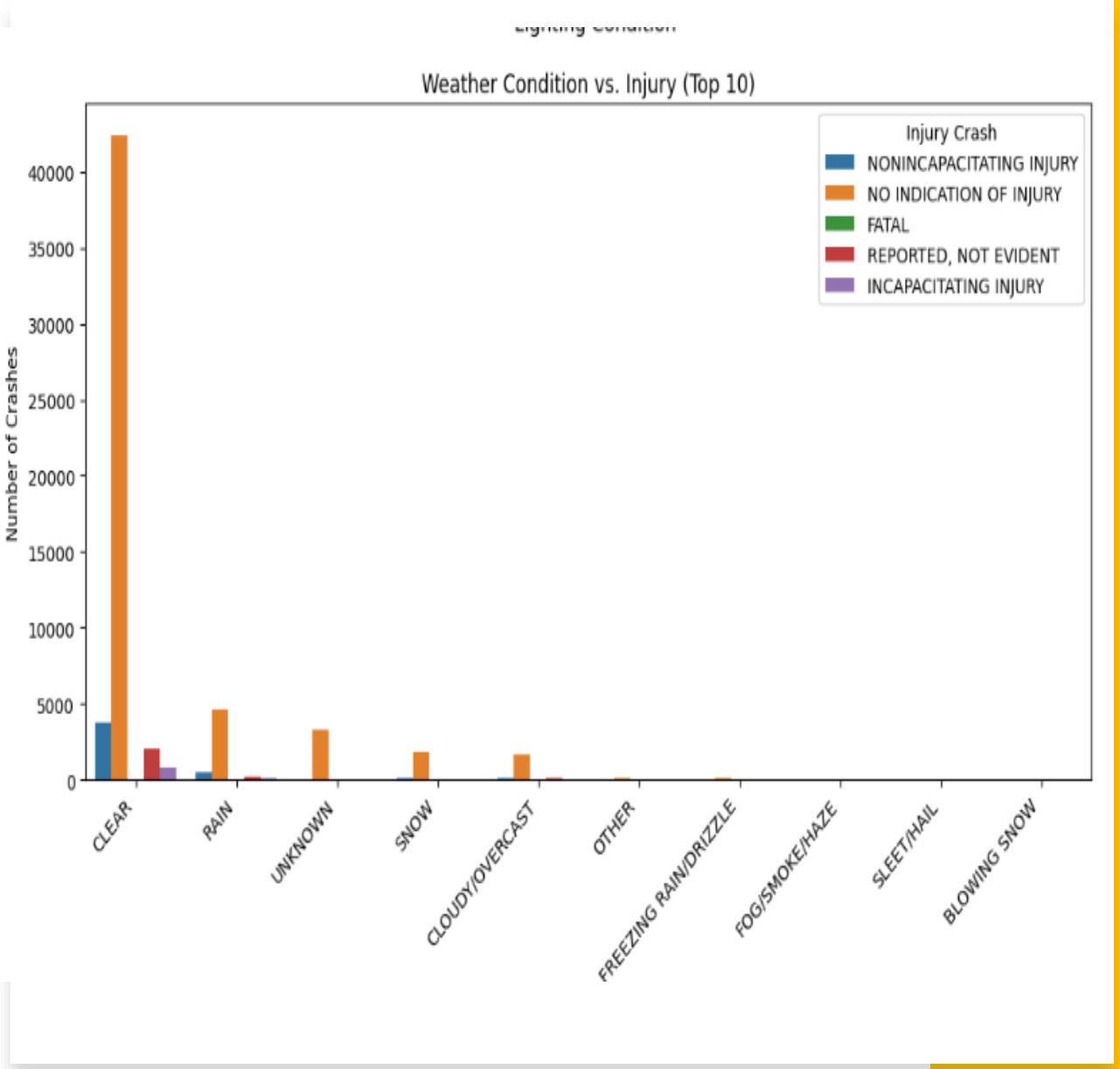
Lighting: Most crashes happen in daylight, which could be due to higher traffic volumes. However, darkness with lighted roads also has a significant number, suggesting visibility isn't the only factor.



Weather Condition VS Injuries

Key Insights

Weather: Clear weather has the most crashes, probably because people drive more. Adverse weather like rain or snow shows fewer crashes but possibly higher severity.

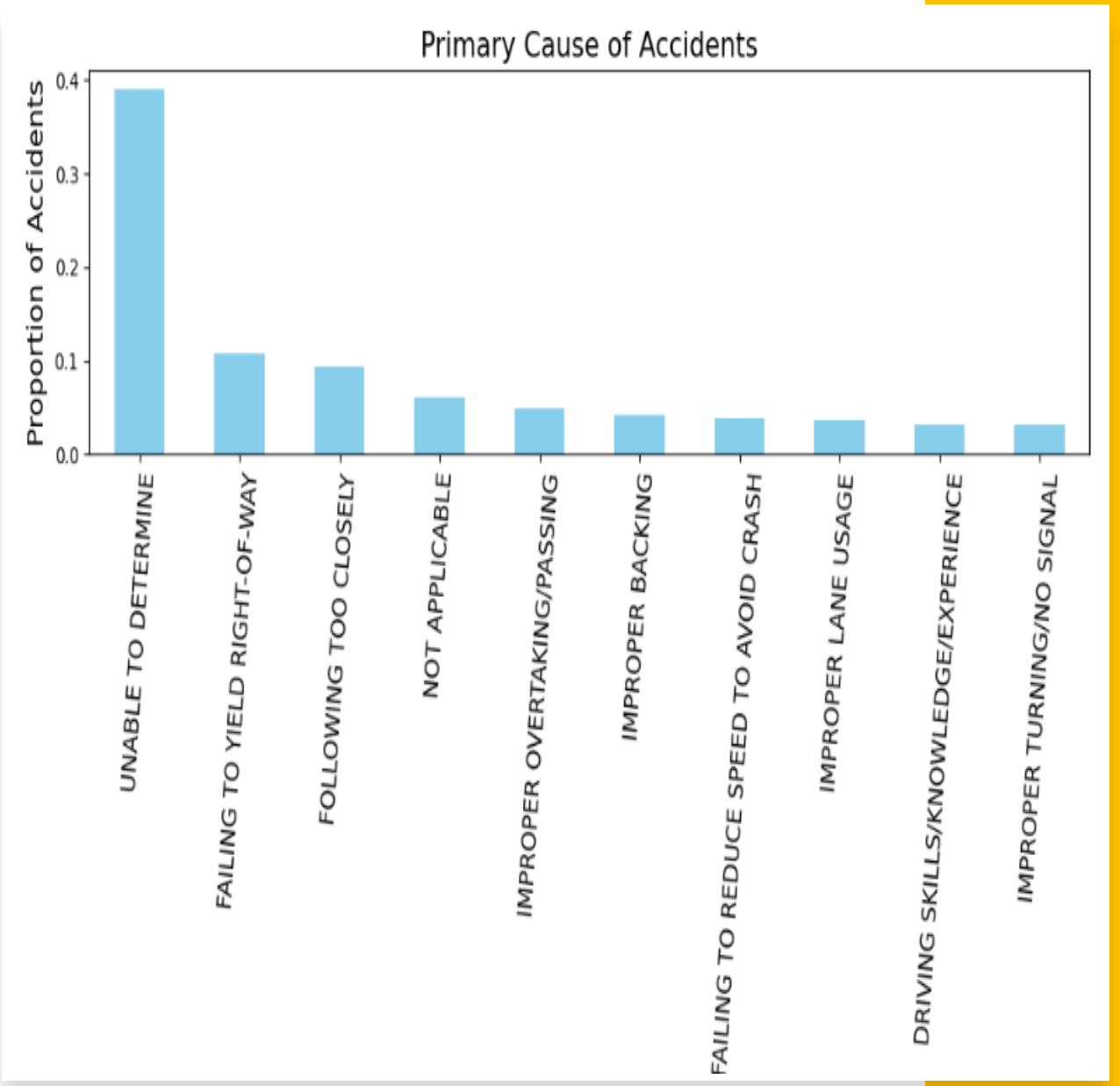


Primary Causes for Crashes

Key Insights

The primary causes include following unable to determine , fail to yield way, too closely, failing to reduce speed, and improper backing.

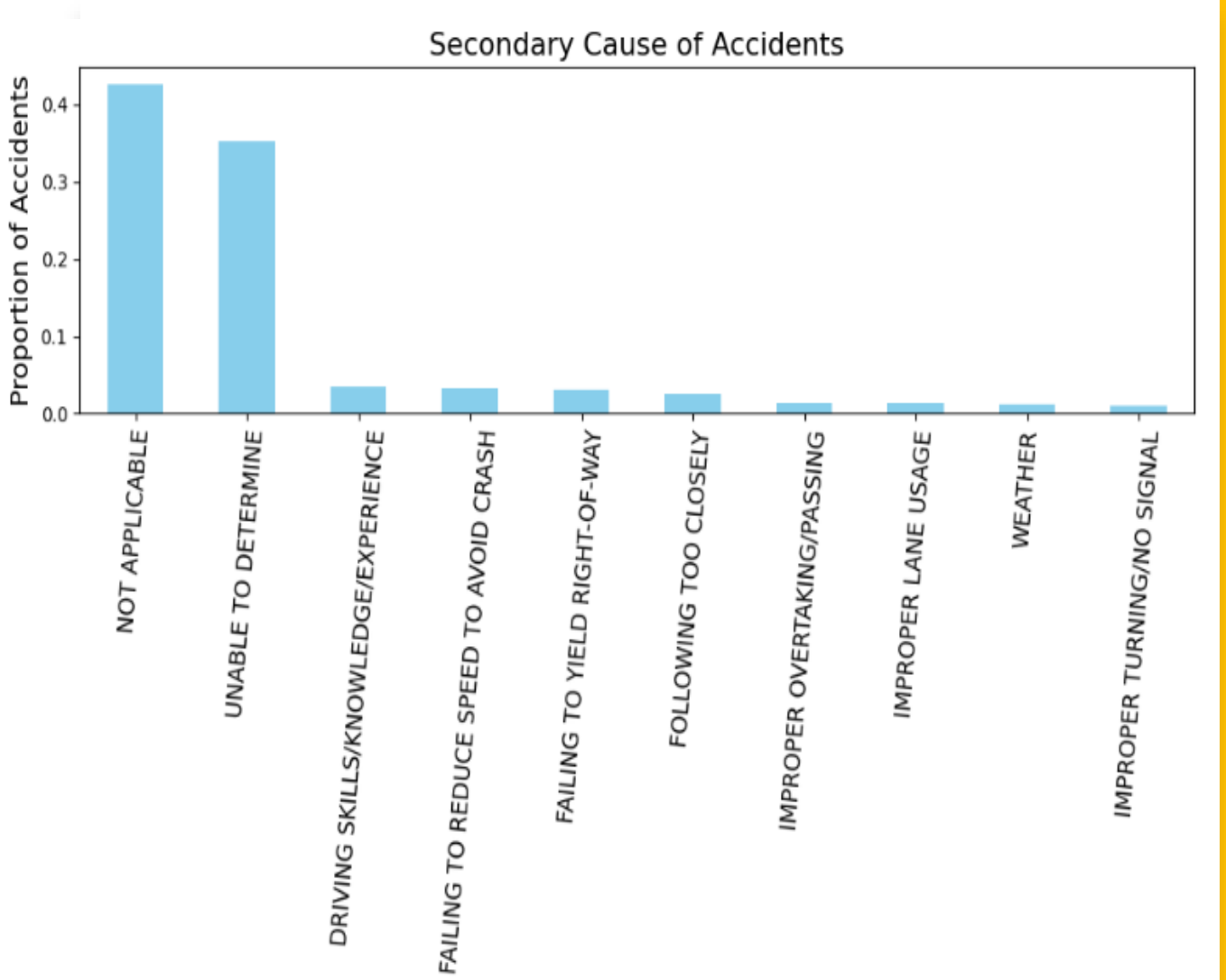
This highlights human error as a major factor.



Secondary Causes for Crashes

Insights

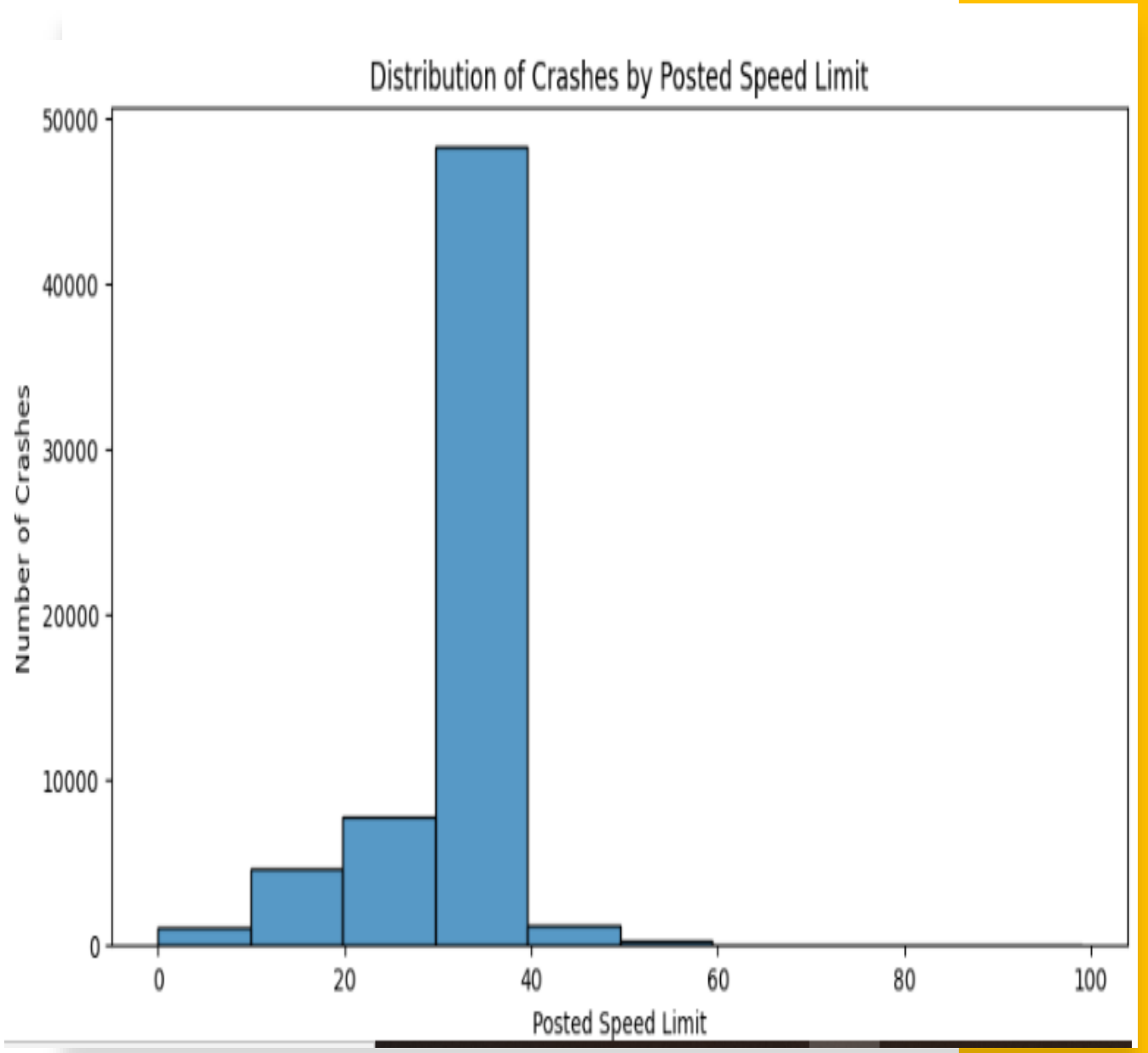
Secondary causes involve distractions and reckless driving. This highlights human error as a major factor.



Posted Speed Limit Distribution

Insights

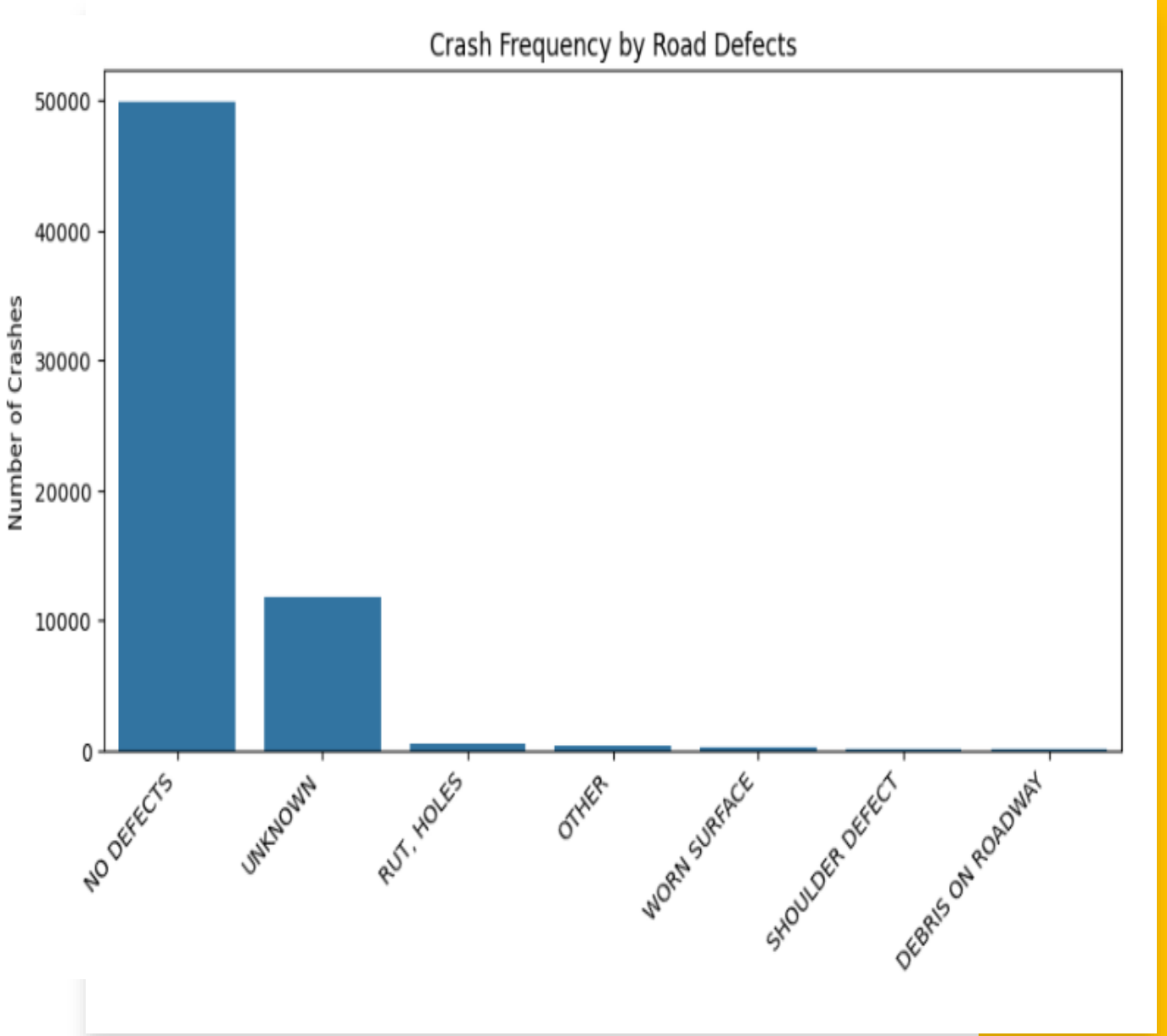
Speed posted is average in 30's this is low but it can indicate more accident occur when the road population is densely populated may be due to Rush hour or it occur in high densely populated like Western Avenue where they is high accidents.



Crash Frequency by Road Defects

Key Insights

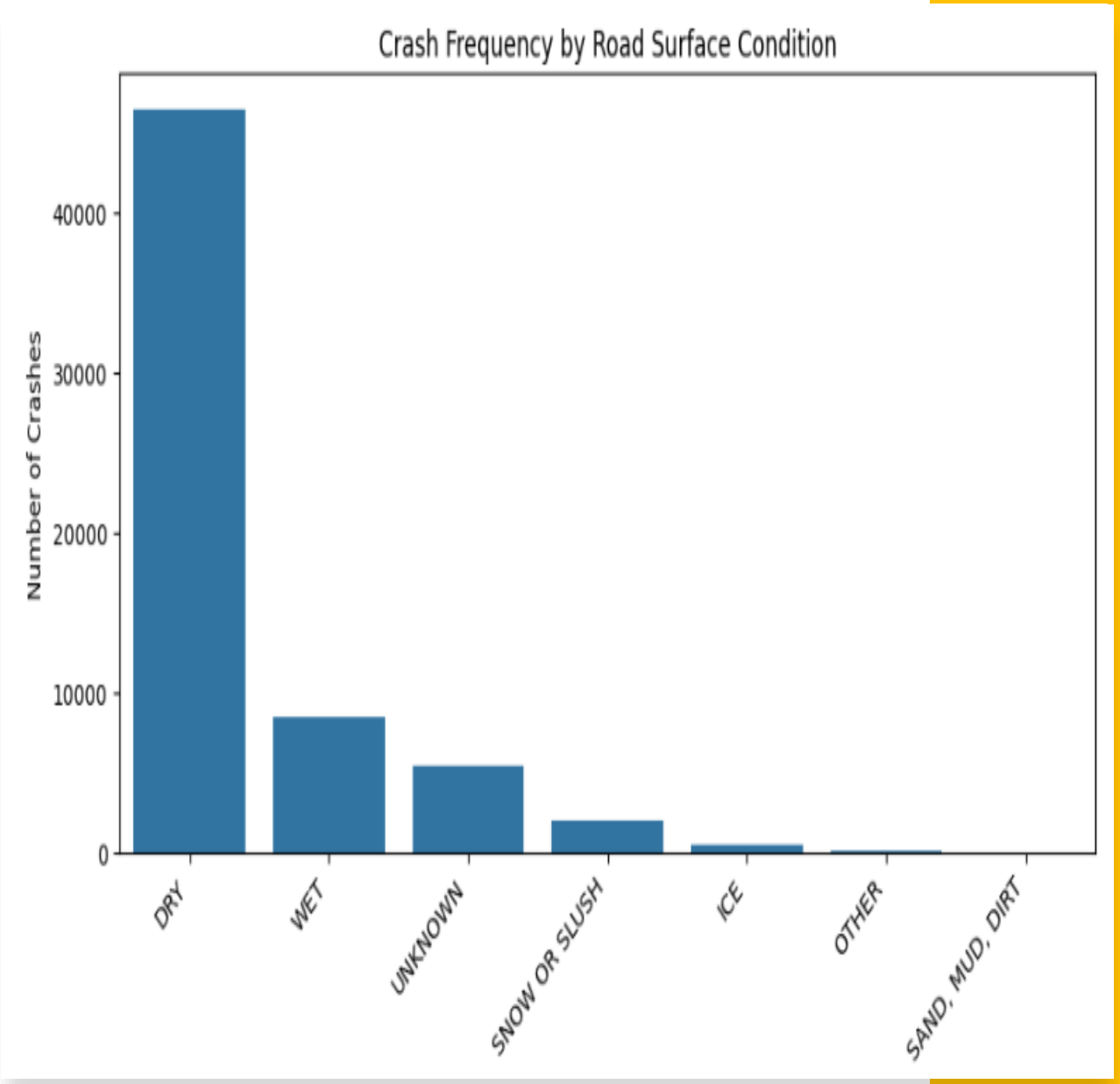
Road defects like potholes contribute but aren't the main issues. There may be other factors. **which again might be due to higher traffic.**



Crash Frequency by Road Surface Condition

Key insights

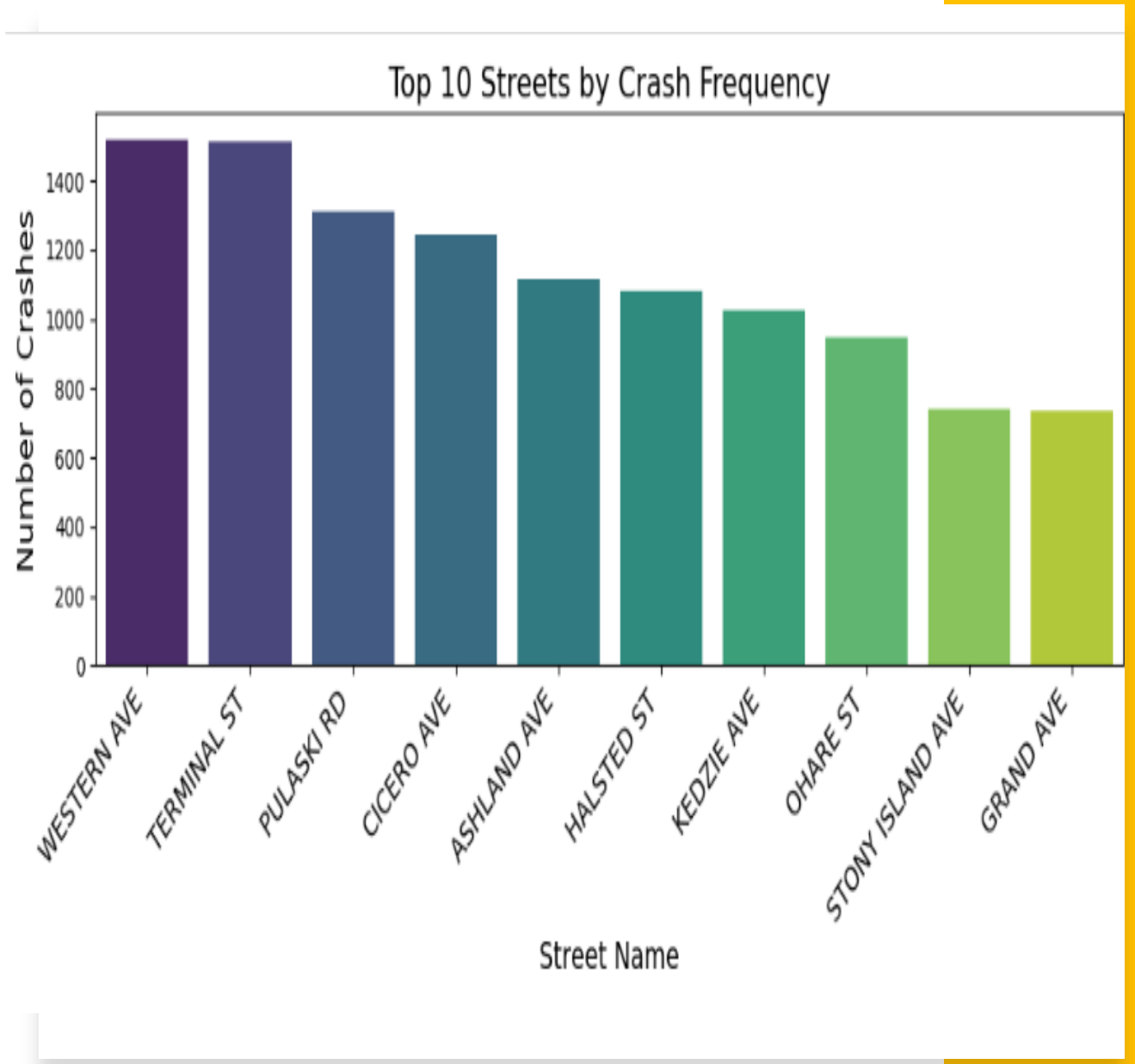
Most crashes occur on dry roads, which again might be due to higher traffic. Road defects like potholes contribute but aren't the main issue. Trafficway types like parking lots and divided roads have higher crash frequencies.



Top 10 Streets by Crash Frequency

Key Insights

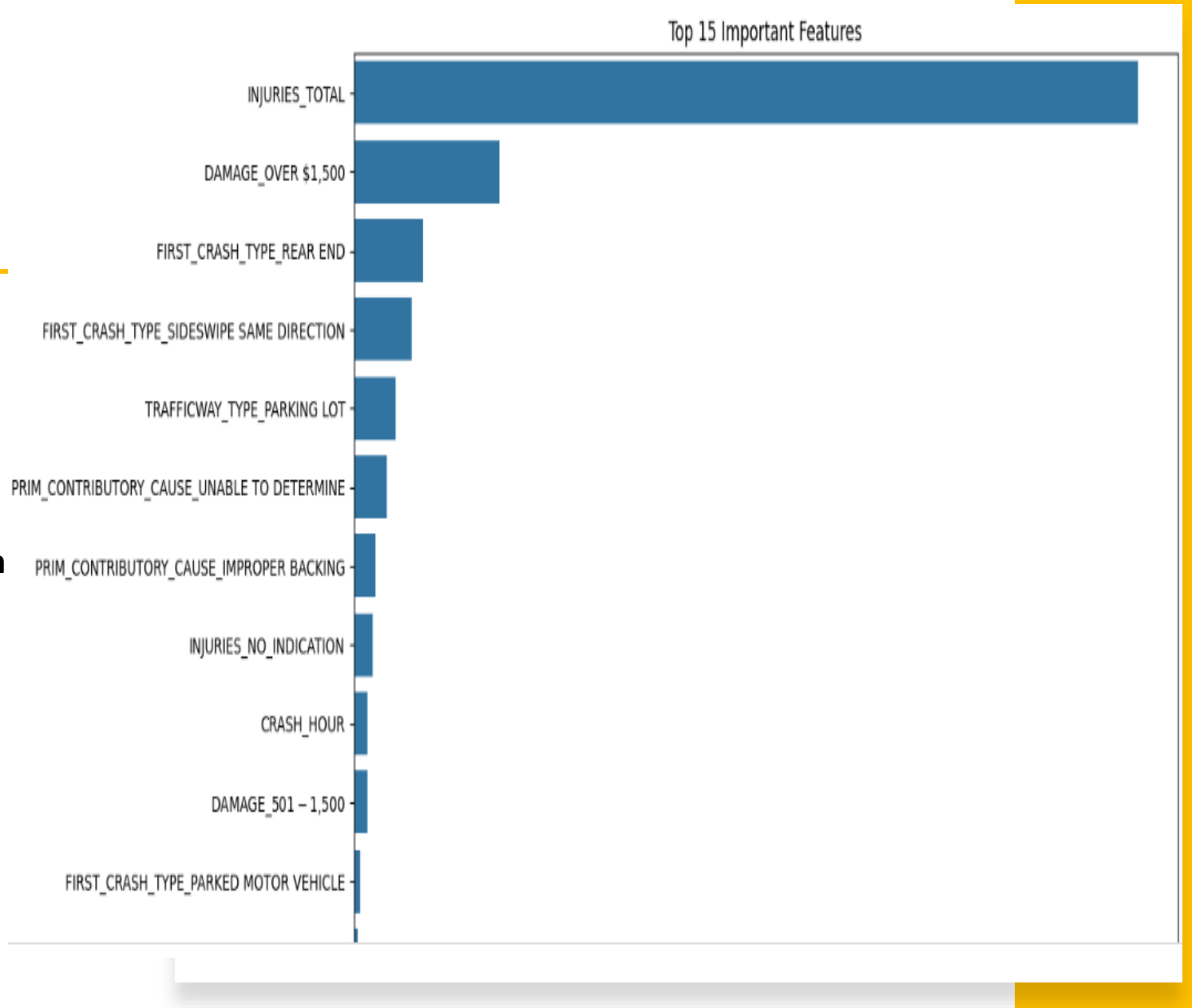
Top streets with high crashes are major roads like Western Ave. This suggests intersections or high-traffic areas need attention.



Top 15 Importance Features

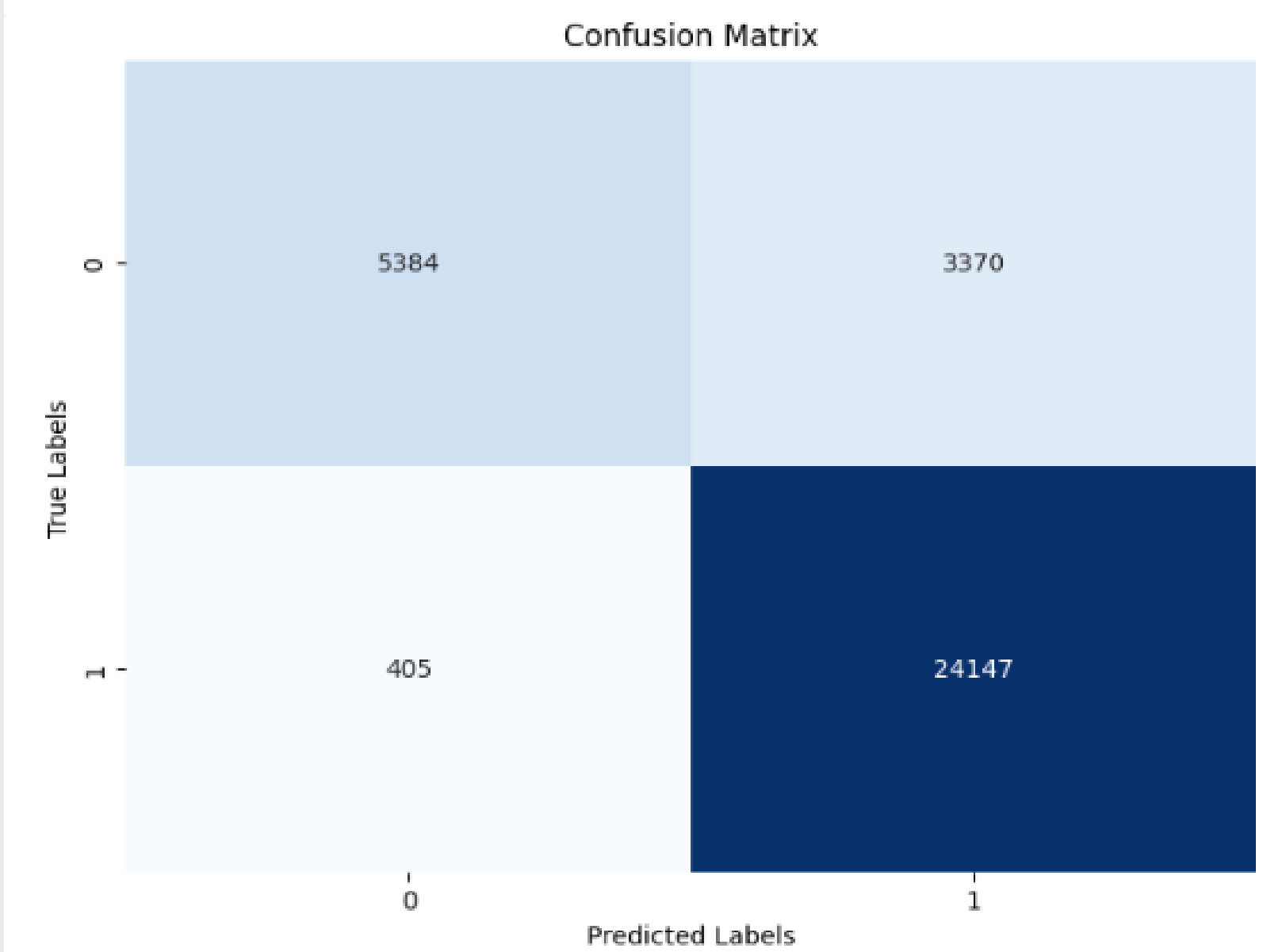
Insights

Injuries _total and Damage we the leading in matters pertaining predicting.



Best Performing Model

- **Insights**
- It correctly predicts nearly 89% of all cases. It catches almost all critical cases (class 1) with a 98% recall. Its high F1-score (93% for class 1) shows a good balance between precision and recall.
- The confusion matrix confirms very few critical cases are missed. In short, the model is very reliable at detecting the most important cases while maintaining strong overall performance.



Recommendations

01

target improvements in infrastructure and lighting could reduce the most dangerous incidents

02

Accidents tend to peak during specific times like rush hours and weekends. This insight can help optimize the deployment of law enforcement and emergency services, ensuring a stronger presence when and where accidents are more likely.

03

Analysis of driver-related factors, including speeding, distracted, or impaired driving, indicates that public awareness campaigns and stricter traffic enforcement could play a significant role in reducing accident rates.

04

Mapping crash locations has highlighted specific areas with higher accident frequencies. These hotspots are prime candidates for immediate safety interventions, such as enhanced signage, redesigned intersections, or increased monitoring.

Conclusion

This project demonstrates that machine learning models can play a vital role in predicting the severity of Chicago car crashes—information that is crucial for improving road safety and informing policy decisions. By starting with simple baseline models (such as Logistic Regression, Decision Tree, Random Forest, and KNN) and then addressing a key data challenge (class imbalance) with SMOTE, we observed significant improvements in performance. In particular, the Random Forest classifier achieved an accuracy of approximately 91.7% and an F1 score of 0.92 on resampled test data, making it a strong candidate for deployment. the final model had the highest balance

For stakeholders like the Chicago Department of Transportation, the Police Department, and urban planners, these results offer actionable insights. For example, the model can help pinpoint high-risk locations and conditions, enabling targeted safety interventions and optimized resource allocation. However, the analysis also highlights limitations—such as missing values and the inherent imbalance in the crash severity data—that need to be continuously monitored. Future work should include further feature selection, real-time data integration, and ongoing validation to ensure that the model remains effective in evolving traffic conditions and supports long-term road safety improvements.

QUESTION AND ANSWER?

THANK YOU