Predictive Analysis of Traffic Crashes in Chicago: Crash Types and Damage Estimation

1st Prathibha Vuyyala
Engineering Science - Data Science
University at Buffalo
Buffalo, US
pvuyyala@buffalo.edu

2nd Tharun Teja Mogili
Engineering Science - Data Science
University at Buffalo
Buffalo, US
tharunte@buffalo.edu

3rd Pavan Pajjuri
Engineering Science - Data Science
University at Buffalo
Buffalo, US
spajjuri@buffalo.edu

***Abstract*—This analysis of Chicago traffic accident data from June to December 2023 focuses on data cleaning, validation, and standardization. Missing values were imputed, duplicates and inconsistencies addressed, and categorical features encoded for machine learning. The cleaned dataset is prepared for further statistical analysis of traffic patterns.**

I. Introduction

An average span of four days, Chicago can record up to over a thousand car accidents. When you include drivers, passengers, pedestrians and cyclists, up to two thousand people can be effected. Forty-five percent of the people will experience a minor to fatal injury.

Traffic accidents represent a critical challenge for urban safety management, resulting in injuries, fatalities, and economic loss. This project focuses on leveraging the Chicago crashes dataset to address specific problems related to traffic crashes

II. Problem Statement

A. Problem Statement Questions

- **Predicting Crash Type:** Utilizing available data features, we aim to develop a predictive model that can accurately classify the crash type (Injury or No Injury) based on factors such as Road conditions, weather conditions, and traffic control devices. Understanding if the crash could injurious can inform targeted prevention strategies and resource allocation.
- **Estimating Damage Severity**: By analyzing contributing factors like vehicle conditions, roadway characteristics, and crash circumstances, we will build a model to estimate the potential damage severity of a crash. This can assist in anticipating resource needs for emergency services and informing policy decisions around road safety improvements.
- **Identifying Risk Factors for Severe Accidents:** We will investigate how various elements (e.g., environmental conditions, traffic volume) correlate with severe accidents. This analysis aims to uncover actionable insights that can guide interventions aimed at reducing the occurrence and impact of high-severity crashes.

B. Background of the Problem

Traffic accidents in urban environments, such as Chicago, are a persistent concern, contributing to injuries, fatalities, and substantial economic losses. With the increasing complexity of urban traffic systems and the multitude of factors influencing crash events—ranging from road conditions, weather, and traffic control devices to driver behavior—the need for predictive tools has grown. This project aims to explore these complex dynamics by utilizing the Chicago crashes dataset, which provides a wealth of structured data that can be used to uncover patterns and factors contributing to traffic accidents. By focusing on predicting crash types, estimating damage severity, and identifying risk factors for severe accidents, this project addresses key challenges in urban traffic safety management. Furthermore, the project seeks to better understand the conditions under which severe injuries and hit-and-run incidents occur, and how poor roadway surface conditions impact accident rates.

C. Significance of the Problem and Project Contribution

This project is significant because traffic accidents have far-reaching impacts on public safety, economic costs, and emergency services. The ability to predict crash types and estimate damage severity could lead to more efficient allocation of resources, such as emergency services and law enforcement. Moreover, identifying the conditions that lead to severe accidents enables targeted interventions, such as improving road infrastructure or adjusting traffic control measures in high-risk areas. Understanding trends in hit-and-run incidents and the effects of road conditions can guide policy decisions aimed at enhancing traffic safety and accountability. By providing actionable insights through data-driven models, this project has the potential to improve urban traffic management strategies, reduce accident rates, and mitigate the impacts of crashes on public health and safety.

D. Potential Contribution of the Project

This project aims to contribute significantly to the domain of urban traffic safety by developing data-driven models that can predict crash types and estimate damage severity. By leveraging predictive analytics, this project will help inform targeted strategies for accident prevention, particularly in identifying

conditions under which crashes are more likely to result in injuries. This predictive capability can enable city authorities, transportation agencies, and policymakers to allocate resources where they are most needed, reducing response times for emergency services and guiding future infrastructure investments.

Moreover, by estimating damage severity, the project will offer valuable insights for planning and improving road safety. Understanding which factors—whether environmental, vehicular, or road-related—contribute to more severe crashes can help refine traffic regulations, implement better warning systems, and design safer intersections. The analysis of hit-and-run incidents can also reveal important patterns that inform policy decisions to enhance public safety, hold offenders accountable, and provide better post-crash support for victims. Additionally, investigating how roadway conditions influence crash rates and severity offers actionable information to guide road maintenance and policy reforms, which are crucial for reducing accident rates in cities with challenging environmental conditions like Chicago. Ultimately, this project will empower stakeholders to proactively manage traffic safety, contributing to a safer, more sustainable urban transport system.

III. Data Sources

The primary dataset for this project comes from the Chicago Traffic Crashes Dataset, which is publicly available through the City of Chicago's data portal. This dataset contains detailed information about traffic accidents that have occurred in the city of Chicago, with over 870,000 records spanning multiple years, and includes variables such as crash type, weather conditions, road surface conditions, and traffic control device statuses.

The data from the Chicago Traffic Crashes dataset is comprehensive and includes numerous variables that allow for an in-depth analysis of crash incidents. With over 870,000 rows and 48 columns, this dataset is well-suited for addressing the project's core objectives of predicting crash types, estimating damage severity, and identifying risk factors for severe accidents. This data source ensures that the dataset is both large enough to yield significant insights and contains the necessary variety of features to facilitate predictive modeling and exploratory data analysis. The dataset from the Chicago site spans from 2013 to the present. For the purposes of this analysis, we will focus exclusively on a six-month period, specifically from June 2023 to December 2023

You can access the data source at Data Source.

IV. Data Cleaning/ Processing

Effective data cleaning and processing is essential to ensure the accuracy and reliability of the analysis. The Chicago Traffic Crashes dataset, like most real-world datasets, contains inconsistencies, missing values, and redundant information that must be addressed. Below is an outline of the 10 distinct processing and cleaning steps applied to the dataset:

A. Dropping Null Rows and Columns

The dataset contains records with missing values, prompting the need for a structured approach to manage these gaps. A column threshold of 30% was set, leading to the removal of columns with excessive missing data. Consequently, the following columns were dropped: *WORK_ZONE_TYPE, NOT_RIGHT_OF_WAY_I, STATEMENTS_TAKEN_I, PHOTOS_TAKEN_I, WORK_ZONE_I, CRASH_DATE_EST_I, WORKERS_PRESENT_I, LANE_CNT, DOORING_I, INTERSECTION_RELATED_I*. This process reduced the dataset from 48 columns to 38 columns. Additionally, a row threshold of 20% was implemented to eliminate rows with excessive missing values, ensuring the quality of the remaining data and minimizing potential bias in subsequent analyses. Additionally, a row threshold of 20% was implemented to eliminate rows with excessive missing values, ensuring the quality of the remaining data and minimizing potential bias in subsequent analyses.

B. Handling Missing Data

The dataset was analyzed to identify missing values in both categorical and numerical columns. For categorical variables such as `REPORT_TYPE`, `HIT_AND_RUN_I`, and `MOST_SEVERE_INJURY`, missing values were imputed with the mode of each column to maintain consistency. The mode values used were:

- `REPORT_TYPE`: "NOT ON SCENE (DESK REPORT)"
- `HIT_AND_RUN_I`: "Y"
- `MOST_SEVERE_INJURY`: "NO INDICATION OF INJURY"

For numerical variables, missing values were also filled using the mode. Key columns with missing values included:

- `INJURIES_TOTAL`
- `INJURIES_FATAL`
- `INJURIES_INCAPACITATING`
- `INJURIES_NON_INCAPACITATING`
- `INJURIES_REPORTED_NOT_EVIDENT`
- `INJURIES_NO_INDICATION`
- `INJURIES_UNKNOWN`
- `LATITUDE`
- `LONGITUDE`

C. Capping Outliers

To address the issue of potential outliers in the `INJURIES_NO_INDICATION` column, a threshold was established to cap values exceeding 10. Any record with `INJURIES_NO_INDICATION` greater than 10 was adjusted to a maximum value of 10. This capping method helps mitigate the influence of extreme values on the analysis, ensuring that the dataset remains robust for further statistical evaluations.

D. Standardization of Numerical Features

The numerical features in the dataset underwent standardization to ensure that they have a mean of 0 and a standard deviation of 1. This process was applied to all numeric columns

identified in the dataset. The `StandardScaler` from scikit-learn was utilized for this purpose, transforming the data accordingly. Standardization is crucial for many machine learning algorithms as it enhances the model's performance by ensuring that each feature contributes equally to the distance calculations. A descriptive summary of the standardized features numerical features was printed for further analysis in the code.

E. Dealing with Categorical Values and Inconsistency - Mapping

The dataset contains several categorical variables that may have inconsistent entries, which could affect the analysis. A systematic mapping approach was employed to standardize the values in key categorical columns, ensuring uniformity. The following steps were taken:

- Traffic Control Device: Multiple entries were consolidated into broader categories (e.g., "TRAFFIC SIGNAL", "FLASHING CONTROL SIGNAL", and "RAILROAD CROSSING GATE" were all mapped to "SIGNAL"). This mapping reduced variability and enhanced the clarity of data, resulting in a more manageable distribution of categories.
- Weather Conditions: Various weather descriptions were similarly standardized to ensure consistency (e.g., "RAIN" includes "FREEZING RAIN/DRIZZLE"). This allows for better comparison and analysis of accident occurrences under different weather conditions.
- Primary Contributory Cause: The mapping transformed numerous specific causes of crashes into broader categories (e.g., "FAILING TO YIELD RIGHT-OF-WAY" became "YIELDING ISSUES"). This aids in summarizing the data while still retaining essential information for analysis.

The resulting distributions of the standardized categories were examined, revealing clearer patterns and relationships among the variables, which are crucial for subsequent analysis.

F. Data Type Change

The `DATE_POLICE_NOTIFIED` column was converted to a datetime format using `pd.to_datetime()`, allowing for efficient date manipulation and analysis. The format specified was `%m/%d/%Y %I:%M:%S %p`, and any errors during conversion were handled by coercing them to `NaT`.

Several columns related to the timing of crashes were transformed to the category data type to optimize memory usage and improve performance during analysis. These columns included:

- `CRASH_HOUR`: Representing the hour of the crash.
- `CRASH_DAY_OF_WEEK`: Indicating the day of the week the crash occurred.
- `CRASH_MONTH`: Denoting the month in which the crash happened.

By converting these columns to categorical types, the dataset is better structured for analysis and visualization tasks that may benefit from treating these values as distinct categories.

G. Redundant Columns Removal

Certain columns identified as redundant were removed from the dataset to streamline the data and enhance its utility for analysis. These columns were deemed of little use either due to their limited analytical value or because their information was already captured in other columns. The columns removed include:

- `LOCATION`: Contains information that is duplicative or less relevant.
- `SEC_CONTRIBUTORY_CAUSE`: Provides minimal additional insights beyond the primary cause.
- `STREET_DIRECTION`: Does not significantly contribute to the analysis.
- `STREET_NAME`: Offers little additional context for the analysis being conducted.
- `STREET_NO`: Similar to `STREET_NAME`, it does not add meaningful information.

This cleaning step helps focus on the most relevant data for further analysis and model building.

H. Temporal Consistency Checks

- The check for logical consistency revealed that all accident records are correctly timestamped, with no instances where the `CRASH_DATE` is later than the `DATE_POLICE_NOTIFIED`.
- This indicates that all reported accidents were notified appropriately without any discrepancies in the temporal order of events.
- The analysis found a significant number of duplicate crash records, totaling 34,419 occurrences across various `CRASH_DATE` entries.
- Specifically, there are 11,996 unique dates with multiple records, indicating that numerous accidents were reported on the same date and time.
- This could be due to multiple accidents occurring simultaneously or errors in data entry.

I. Cross-Validation of Related Columns

- The validation check for the relationship between `INJURIES_TOTAL` and `MOST_SEVERE_INJURY` found 0 invalid records. This means that all records align logically; when there are injuries reported (`INJURIES_TOTAL` ¿ 0), the severity classification does not mistakenly indicate "NO INDICATION OF INJURY."
- The cross-validation of the relationship between `WEATHER_CONDITION` and `LIGHTING_CONDITION` identified 230 records where the weather was reported as "SNOW" while the lighting condition was classified as "DAYLIGHT." This inconsistency suggests a potential data quality issue, as snowy conditions typically occur when there is limited daylight, indicating possible errors in reporting or categorization.

J. Categorical Encoding

- Ordinal Encoding: Specific columns with a natural order were transformed using ordinal encoding to convert categorical values into numerical representations. The mappings for each column were defined as follows:
  - `Lighting Condition`: Values were assigned from 1 (*DAYLIGHT*) to 6 (*UNKNOWN*).
  - `Most Severe Injury`: Injury severity was mapped from 1 (*NO INDICATION OF INJURY*) to 5 (*FATAL*).
  - `Report Type`: Categorical values were encoded to 1 (*NOT ON SCENE*) and 2 (*ON SCENE*).
  - `Crash Type`: Encoded as 1 (*NO INJURY*) and 2 (*INJURY/TOW*).
  - `Hit and Run Indicator`: Mapped to binary values, where 'Y' is 1 and 'N' is 0.
  - `Damage`: Categorized into three levels, from 1 (*OVER $1,500*) to 3 (*$500 OR LESS*).
- Nominal Encoding: For columns that do not have a natural order, one-hot encoding was applied. This method created binary columns for each category within the specified nominal columns:
  - `Traffic Control Device`
  - `Device Condition`
  - `Weather Condition`
  - `First Crash Type`
  - `Trafficway Type`
  - `Alignment`
  - `Roadway Surface Condition`
  - `Road Defect`
  - `Primary Contributory Cause`

V. Exploratory Data Analysis

1) Summary Statistics: The histograms and boxplots provide valuable insights into the distribution of various numerical features in the dataset, as well as potential outliers.

From the histograms, it's clear that variables such as posted speed limit, lane count, and number of units involved in crashes have skewed distributions, with most crashes occurring at lower speed limits (under 40 mph), involving fewer lanes and fewer vehicles. The histogram for crash hour shows that crashes are more frequent during the middle of the day, with a peak between 12 PM and 6 PM. Other variables, like injury-related counts, display a concentration of data around lower values, indicating that the majority of crashes result in minimal or no injuries.

The boxplots highlight outliers in some features, particularly in variables such as street number and number of units, where extreme values extend beyond the whiskers. These outliers could represent crashes in unique locations or incidents involving multiple vehicles. For instance, the beat of occurrence shows variability with a noticeable range of values, but extreme cases still occur beyond the normal distribution. The presence of outliers in injury-related variables is less prominent, but outliers in date differences suggest

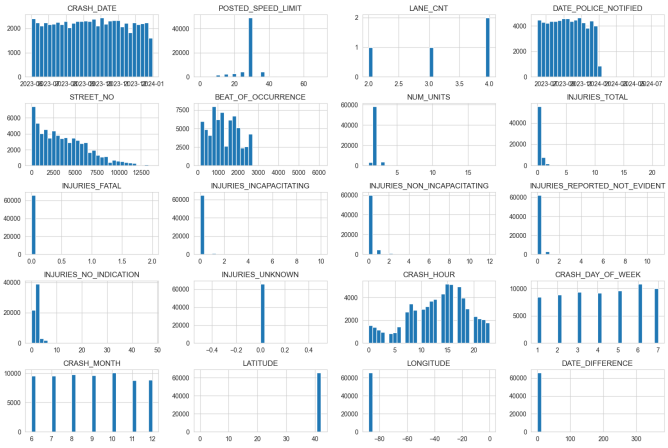

Fig. 1. Data Distribution of Features

that some incidents take significantly longer to be reported than others. These insights suggest that while the majority of crashes follow expected patterns, there are exceptions that may warrant further investigation, particularly those involving delayed reporting or unusual locations.

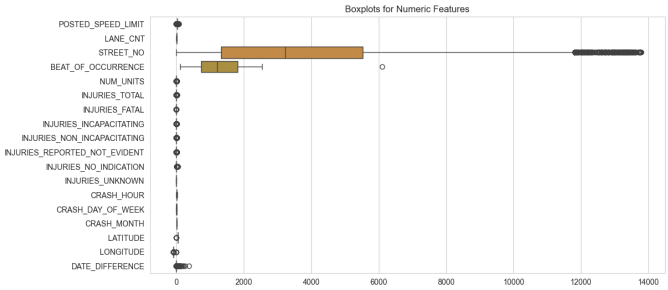

Fig. 2. Box plot for numeric features

2) Binning Hours to Analyze Distribution of Accidents: The distribution of crashes by hour bins reveals several important patterns. The data shows that the highest number of crashes occurs between 12 PM and 6 PM, likely corresponding to peak traffic periods such as lunch breaks and afternoon commutes. This suggests that increased traffic volume, coupled with potentially higher stress levels during these hours, plays a role in the elevated crash rates. The 6 AM to 12 PM time bin also shows a considerable number of crashes, which could be attributed to morning rush hours as people head to work or school, indicating another peak period for traffic accidents.

In contrast, significantly fewer crashes are observed between midnight and 6 AM, which is expected due to reduced traffic volume during late-night hours. The crash rate between 6 PM and midnight is lower than in the afternoon but remains substantial, likely due to evening activities and post-work commutes. These findings suggest that targeted interventions, such as increased traffic monitoring or public safety campaigns, may be most effective when implemented during afternoon and morning rush hours, where the risk of crashes appears to be the highest.

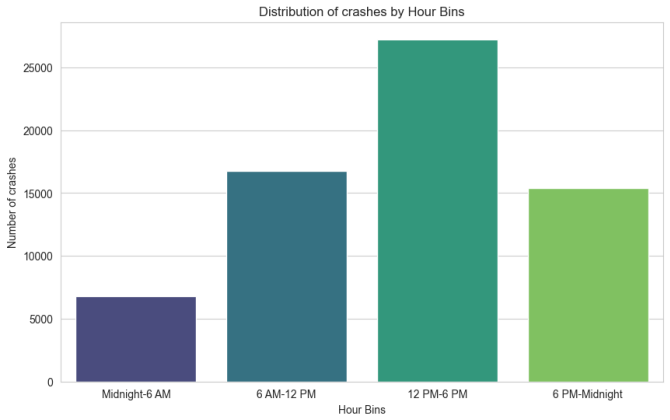

Fig. 3. Distribution of crashes by Hour Bins

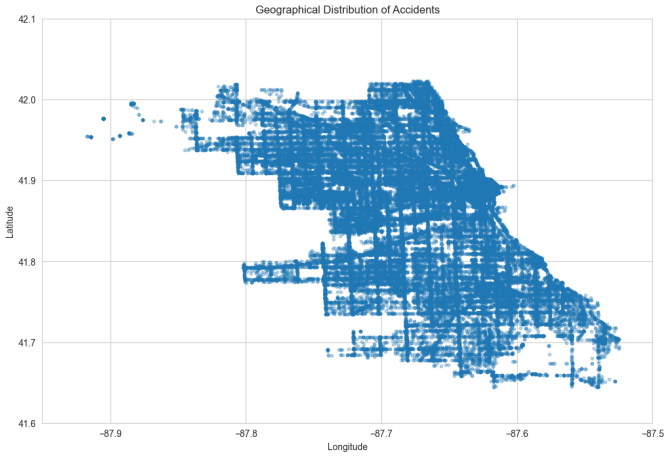

Fig. 4. Geographical Distribution of Accidents

3) Geographical Distribution of Accidents: The scatter plot visualizing the geographical distribution of accidents provides significant insights into the spatial concentration of crashes. The dense clustering of points indicates that accidents are highly concentrated in specific regions, which may correspond to urban areas with high traffic volume. These clusters can help identify potential accident-prone zones, possibly due to a combination of factors such as road design, traffic congestion, and the presence of key intersections or landmarks. Urban planners and local authorities can use this data to focus on improving road safety in these regions, perhaps by installing better traffic control devices or redesigning intersections. Furthermore, the spread of points outside the dense areas suggests that accidents also occur in less populated or suburban areas, though at a much lower frequency. By analyzing these patterns further, transportation officials can determine whether these accidents are due to factors like road conditions, inadequate lighting, or speeding. This geographical analysis is a crucial component in identifying high-risk areas and implementing targeted interventions to reduce the occurrence of accidents in both densely populated urban areas and less traveled suburban regions.

4) Traffic Accidents by Months: The line plot showing traffic accidents by month for the year 2023 highlights noticeable fluctuations in the number of accidents throughout the year. A peak in accidents can be observed in October, with a slight decline following it. This suggests that there could be seasonal factors, such as changes in weather or traffic volume, contributing to a higher accident rate during this period. In contrast, there is a sharp drop in accidents in December, which may be due to holidays or different traffic patterns during the winter months, resulting in reduced road activity or heightened caution. Interestingly, August and November also demonstrate higher accident rates, possibly due to transitional periods between summer and fall, where weather conditions might vary, leading to unpredictable road safety. This analysis can help city planners and authorities identify specific months when increased road safety measures should be implemented to mitigate the risk of traffic accidents, particularly during peak months like October and transitional months like August and November.

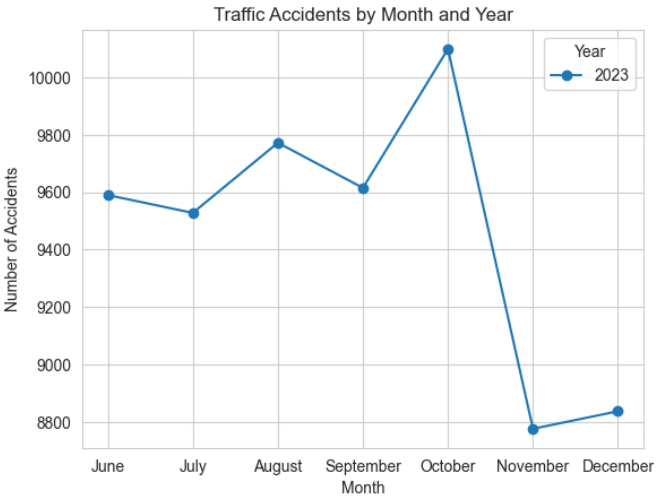

Fig. 5. Traffic Accidents by Months

5) Total Accidents by Primary Cause of Traffic Accidents: The bar plot visualizing the primary causes of traffic accidents reveals some significant insights. The leading cause of accidents is failing to yield the right-of-way, accounting for 7,578 incidents. This suggests that right-of-way violations remain a persistent issue on the roads, possibly due to lack of awareness or disregard for traffic rules. Another major contributor is following too closely, with 5,469 incidents, which may indicate a prevalent issue with tailgating behavior, especially in high-traffic areas. These two causes highlight the need for stricter enforcement and education around defensive driving techniques. An interesting observation from the chart is that a substantial number of accidents (27,734 cases) are marked as unable to determine the cause. This ambiguity suggests either inadequate data collection or challenges in pinpointing the specific cause of accidents, signaling an opportunity

for better reporting mechanisms. Additionally, other causes, such as improper lane usage and disregarding traffic signals, also contribute significantly, underscoring the complexity of traffic safety issues that need to be addressed through both infrastructure improvements and driver education programs

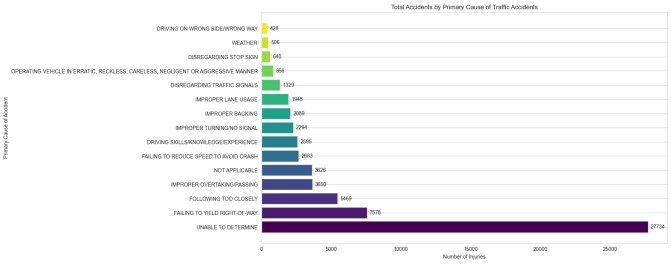

Fig. 6. Total Accidents by Primary Cause of Crash

6) Weather and Lighting Conditions: The heatmap displaying accidents based on weather and lighting conditions reveals some clear trends. Clear weather and daylight conditions show the highest number of accidents, with 35,206 accidents recorded under these conditions. This is likely due to the fact that most road activity happens during clear and daylight conditions, resulting in more opportunities for crashes despite the favorable weather and visibility. Similarly, clear weather combined with lighted roads also has a significant number of accidents, suggesting that despite better road conditions, other factors such as traffic density might play a role in accident occurrences. On the other hand, severe weather conditions like fog, snow, and rain under low visibility conditions (such as darkness or dawn) show relatively fewer accidents. This could imply that drivers are more cautious during these conditions or that there are fewer vehicles on the road. However, the risks posed by such weather should not be underestimated, as the combination of rain and darkness still results in a notable number of accidents (2,434 cases). This suggests that while extreme weather reduces the number of vehicles on the road, it increases the danger for those who are driving, necessitating targeted interventions like weather advisories or road safety campaigns.

7) Distribution of Crash Types Across Different Device Conditions Using a Violin Plot: The violin plot illustrates the distribution of crash types across various device conditions, providing insights into how traffic control devices impact the severity of crashes. Functioning properly devices have the most variation in crash types, with a significant proportion resulting in injury and/or tow due to crash, suggesting that even when devices are operational, other factors like human error or road conditions contribute to severe crashes. This highlights the need for further investigation into human behavior, even in areas with functioning traffic control systems. On the other hand, device conditions such as functioning improperly or no controls show a relatively lower variance in crash types, but with more extreme outcomes leaning towards severe crashes. This could imply that malfunctioning or absent traffic devices can lead to more severe accidents, emphasizing the importance

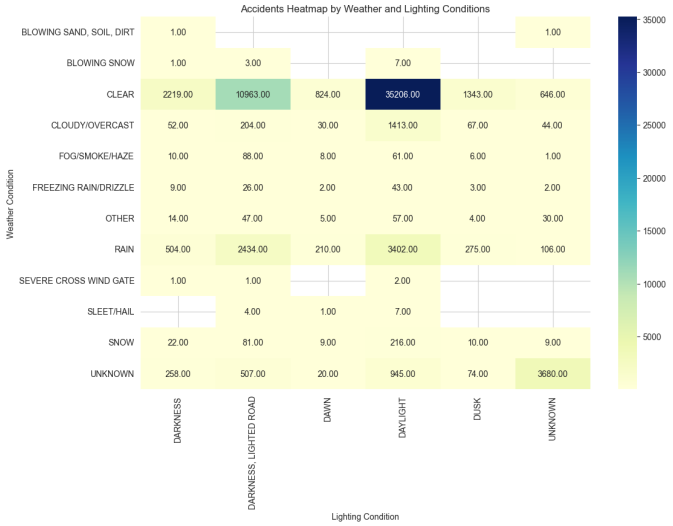

Fig. 7. Crashes Heatmap by Weather and Lighting Conditions

of maintaining and installing adequate traffic control systems to mitigate crash severity. The presence of worn reflective material and missing devices further supports the need for effective traffic control maintenance to prevent more severe outcomes in accidents.

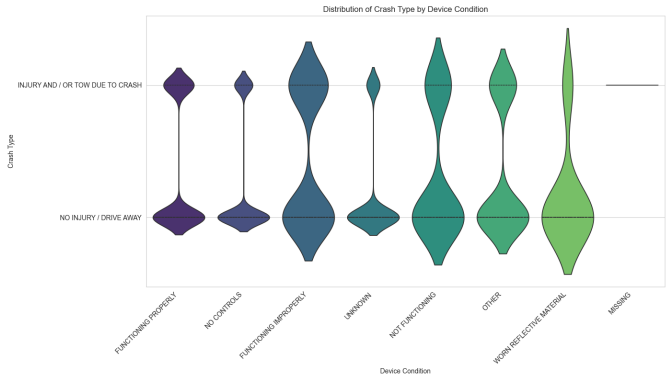

Fig. 8. Distribution of Crash Type by Device Condition

8) Distribution of Traffic Control Devices in Traffic Accidents: The pie chart displaying the distribution of traffic control devices involved in accidents provides a clear breakdown of how different traffic controls are linked to crash occurrences. The most significant portion of crashes involves no controls, indicating that areas without traffic regulation devices are particularly prone to accidents. This finding suggests a need for installing more traffic signals or control mechanisms in such locations to mitigate accident risks. Another notable observation is that a large proportion of accidents occur at traffic signals and stop signs, which may indicate issues related to driver behavior, such as ignoring signals or misjudging their timing. This emphasizes the importance of enforcing traffic rules more strictly at these intersections and potentially re-evaluating signal timings to ensure they are optimally config-

ured for safety. Additionally, the relatively smaller segments involving devices like yield signs and pedestrian crossing signs highlight specific areas where further safety measures could be introduced to enhance protection for vulnerable road users like pedestrians.

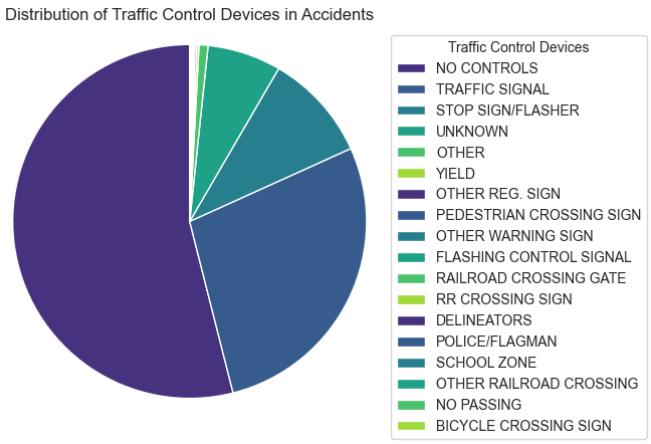

Fig. 9. Distribution of Traffic Control Devices in Accidents

9) Distribution of Damage Categories Across Different First Crash Types: The stacked bar plot depicting the damage categories by first crash type reveals key insights into the extent of financial damage across different types of crashes. For most crash types, particularly angle collisions and rear-end collisions, the majority of crashes result in damages exceeding $1,500. This suggests that these types of accidents are likely to involve significant property damage, possibly due to the nature of the collisions where vehicles impact each other directly and with force.

In contrast, less severe collisions, such as pedestrian-related and fixed object crashes, display a more even distribution across the three damage categories, with some resulting in minimal damage (under $500). Non-collision incidents, including crashes with animals or other non-vehicle objects, tend to fall under the lower damage categories, which could indicate less severe impacts and minor vehicle repairs. These insights suggest that angle and rear-end collisions require more attention from insurance providers and vehicle safety experts, as they often result in costly repairs.

10) Heatmap of Crashes by Road Conditions: The heatmap showing accidents by roadway alignment, surface condition, and road defects highlights several important patterns. The majority of accidents occurred on straight and level roads with dry surfaces (40,027 accidents), suggesting that even under favorable conditions, human factors like speeding, distraction, or following too closely may play a significant role in accidents. This finding underscores the need for continuous driver awareness and safety campaigns, even on seemingly safe roads. Interestingly, wet conditions on straight and level roads also show a relatively high number of accidents (6,394 incidents), indicating that adverse weather, such as rain, greatly increases the risk of accidents. Moreover, road defects such as debris

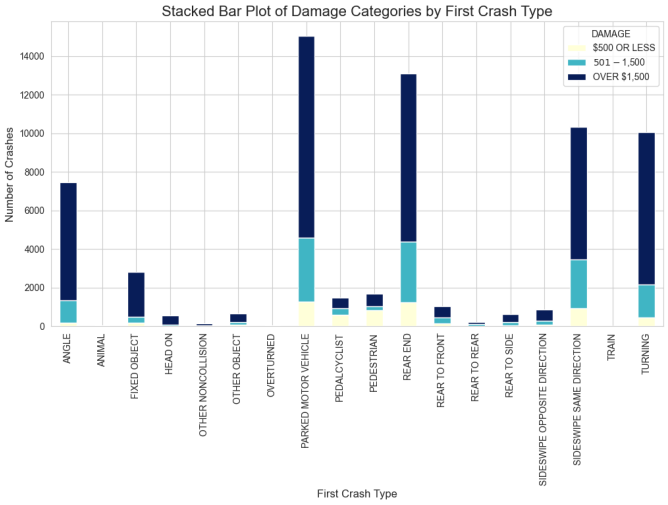

Fig. 10. Stacked Bar plot of Damage Categories by First Crash Type

on the roadway and worn surfaces appear to exacerbate these risks, particularly on straight and level roads, as evidenced by the significant accident counts in these categories. These findings suggest that maintaining road surfaces and clearing debris can have a substantial impact on reducing accident occurrences, particularly during adverse weather conditions.

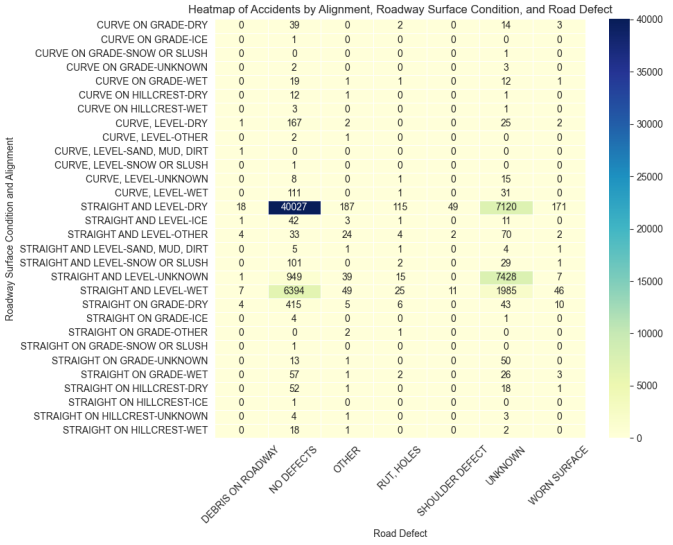

Fig. 11. Heatmap of Crashes by Alignment, Road way surface condition and Road defect

11) Distribution of Most Severe Injuries by First Crash Type: The bar chart illustrating the distribution of the most severe injuries by first crash type reveals several key insights. The majority of accidents, particularly those involving rear-end and angle collisions, tend to result in non-incapacitating injuries or no indication of injury, which implies that while these types of crashes are common, they are less likely to lead to severe or fatal injuries. This suggests that rear-end and angle collisions may occur more frequently due to

traffic congestion or tailgating but tend to happen at lower speeds, reducing the severity of injuries. On the other hand, more severe outcomes, such as incapacitating injuries and fatalities, are associated with less frequent crash types, such as pedestrian collisions and fixed object impacts. These crash types, while less common, tend to be more dangerous, likely due to higher speeds or the vulnerability of pedestrians. The data underscores the need for targeted safety measures, such as pedestrian crossings and better vehicle control near high-risk zones like intersections, to reduce the likelihood of severe or fatal injuries in these types of accidents.

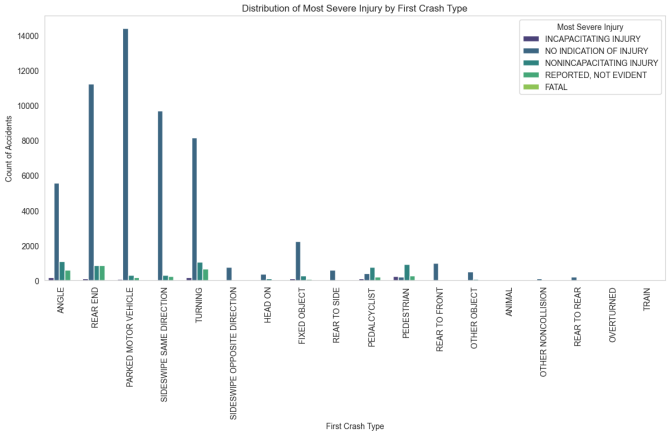

Fig. 12. Distribution of Most Severe Injuries by First Crash Type

12) Correlation Matrix of Traffic Accident Variables: The correlation matrix of traffic accident variables provides key insights into how different factors relate to one another. Notably, there is a strong positive correlation between the total number of injuries and the number of severe injuries (0.74), indicating that accidents involving a high number of injuries tend to also have more severe injuries. Additionally, the correlation between most severe injury and crash type (0.61) suggests that the nature of the crash plays a significant role in determining the injury severity. Another important observation is the negative correlation between posted speed limit and various injury variables. This indicates that higher posted speed limits may not directly correlate with more injuries, possibly due to higher speed limits being implemented in less congested areas with lower crash risk. Finally, the strong negative correlation between latitude and longitude (-0.98) is expected, as these are geographical variables that naturally move in opposite directions. These correlations can help in further refining traffic safety measures by focusing on the factors that most strongly influence crash severity and injury outcomes

13) Analysis of Average Reporting Delay by First Crash Type: The bar chart showing the average date difference by first crash type highlights several interesting trends. Crashes involving non-collision events, such as collisions with animals or other objects, exhibit the longest delay between the crash and police notification, with an average date difference of over 5 days. This could indicate that non-collision events are often

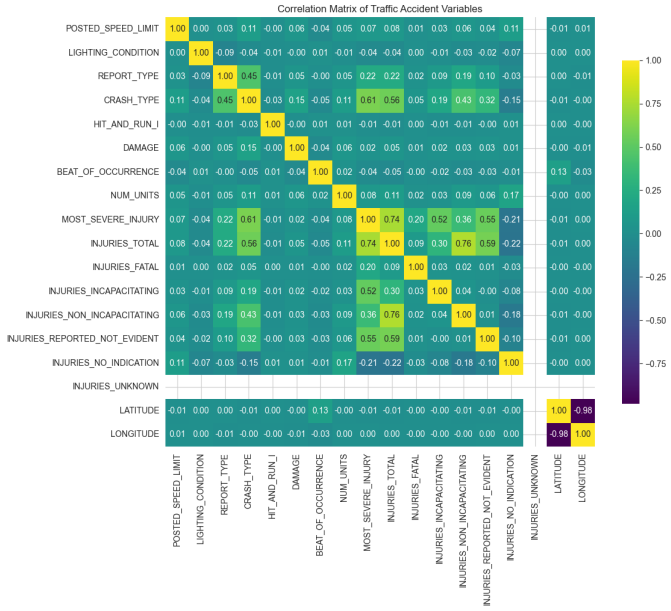

Fig. 13. Correlation Matrix of Traffic Accident Variables

considered less urgent or are less likely to result in immediate reporting, potentially due to lower perceived severity or logistical challenges in rural or remote areas. In contrast, crashes involving more severe or direct interactions, such as train accidents, turning collisions, and angle crashes, show a much shorter average date difference, often within a day or less. These crash types likely demand quicker intervention due to the higher risk of severe injuries and property damage. These insights suggest a potential gap in timely reporting for less severe or non-collision accidents, which could hinder data collection accuracy and delay appropriate responses in certain cases.

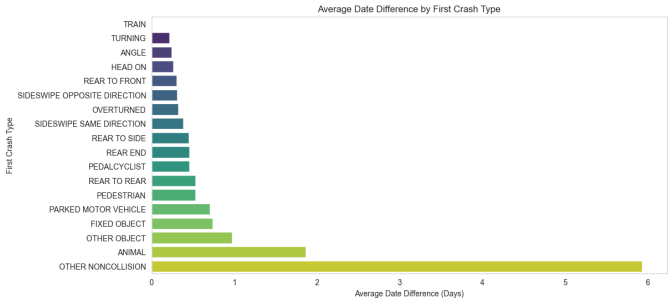

Fig. 14. Average Date Difference by First Crash type

14) Top 10 Streets with the Highest Number of Traffic Accidents in Chicago: The bar chart illustrating the 10 most dangerous streets in Chicago, based on the number of traffic accidents, reveals that Pulaski Road (1,806 accidents) and Western Avenue (1,747 accidents) are the two streets with the highest number of crashes. These findings suggest that these roads may have higher traffic volumes or potentially problematic road conditions, making them prone to accidents.

Pulaski Road, in particular, stands out as the most accident-prone, indicating a need for targeted safety interventions such as traffic calming measures or better traffic signal management.

Additionally, other streets like Cicero Avenue (1,520 accidents) and Ashland Avenue (1,401 accidents) also rank high on the list, highlighting the importance of traffic safety improvements in these areas. These streets are known for heavy traffic flow, potentially explaining the elevated accident counts. The analysis underscores the need for authorities to focus on these high-risk roads to reduce traffic accidents and improve overall road safety in Chicago.

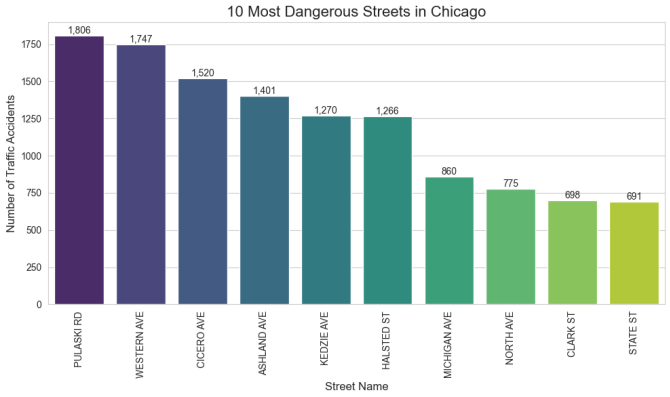

Fig. 15. 10 Most Dangerous Streets in Chicago

REFERENCES

[1] U.S. Department of Transportation, "Traffic Crashes Dataset," data.gov, 2023. [Online]. Available: https://catalog.data.gov/dataset/traffic-crashes-crashes/resource/858674f2-8acc-4803-ba50-91c7faf54030. [Accessed: Month Day, Year].

[2] C. O'Neill and R. Schutt, *Doing Data Science*, O'Reilly, 2013.

[3] National Institute of Standards and Technology (NIST), "Exploratory Data Analysis," 2021. [Online]. Available: https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm. [Accessed: February 2021].

[4] John Tukey Biography, [Online]. Available: https://mathshistory.st-andrews.ac.uk/Biographies/Tukey/. [Accessed: 2021].