Traffic Crashes EDA

Group 12: Divya Sharma (ds655), Dhaval Potdar(dsp50), Jiayi Zhou (jz456), Jiechen Li (jl1254)

2023-10-20

1. Data Overview

Dataset: The data-set has data on Traffic Crashes in Chicago from 2015 onwards, and contains details of each crash, including:

- `Location`: The latitude and longitude of the crash.
- `Date and time`: The date and time of the crash.
- `Injuries`: The number and type of injuries that occurred.
- `Damage`: The estimated damage cost of the crash.
- `Crash type`: The details about the type of crash - hit and run, no right of way, intersection related, causes etc.
- `Conditions`: The details about the weather, lighting, traffic, traffic control devices, and roads etc.

This data can be used to identify the probabilities of crashes happening, and their eventual consequences, given the surrounding situations such as time of day, weather, road conditions, traffic conditions, lighting, and roadway surface etc.

Source: The dataset was obtained from an online repository, though the specific origin or institution responsible for its collection and maintenance was not specified. The dataset is selected from City of Chicago's official page.

Sample Size: The dataset contains a total of 769,100 observations.

Number of Variables: There are 49 variables in the dataset.

Collection Method: The data is collected from traffic crash reports in Chicago. These reports can be sourced from various entities such as traffic police, surveillance systems, or reporting by involved parties. The exact method of data collection, however, has not been provided.

Unit of Analysis: Each row in the dataset represents a unique traffic crash incident in Chicago. The details of the crash, including specifics like location, time, type of crash, conditions at the time of the crash, and the aftermath (like injuries and damages), are captured across the 49 variables.

Research Questions:

1. Given that there is a crash, based on the location, time, crash type, and surrounding road and traffic conditions, how much damage is to be expected (<=1.5K$ or >1.5K$)?

- **Outcome Variable** - Damage (<=1.5K$ or >1.5K$) - categorical variable
- **Input/Explanatory Variables** - Location, date, time of the day, crash type, and surrounding conditions

2. How long it takes in minutes for police to be notified upon accident, given the nature of the accident?

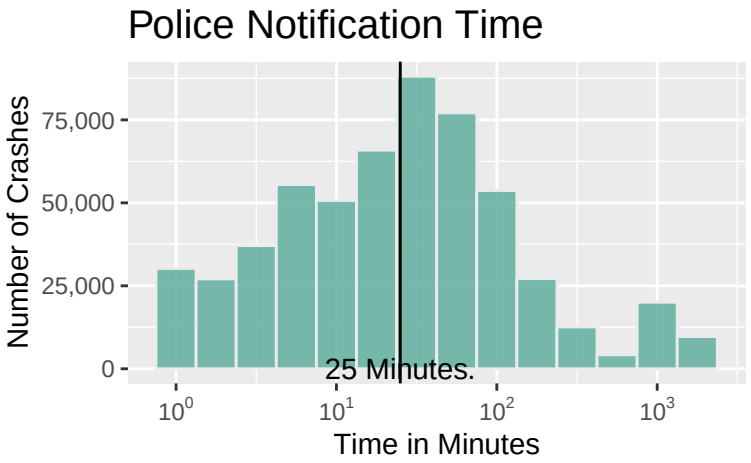
- **Outcome Variable** - Police Notification Time (in minutes) - continuous variable
- **Input/Explanatory Variables** - Location, date, time of the day, crash type, and surrounding conditions.

2. EDA for Outcome Variables

2.1 Police Notification Time

- The distribution is heavily right skewed with a majority of accidents being reported to the police in less than 25 minutes. Thus, the outcome is better visualized on a log scale.
- The median notification time is 25 minutes, the mean is 110 minutes (1 hour and 50 minutes) and the maximum is 1440 minutes (1 day).
- Given the heavy skewing of the outcome variable, we will need to do appropriate transformations on the outcome variables in the modeling step.

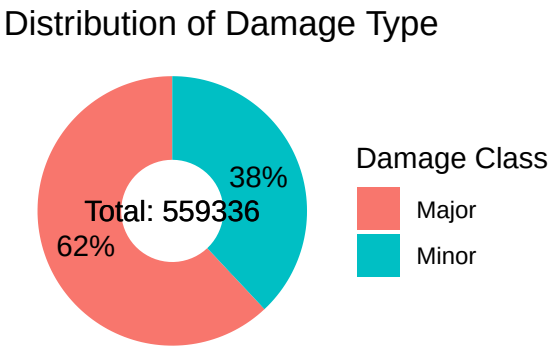
Fig 1: Distribution of Police Notified Time (Continuous Outcome)



2.2 Expected Damage

- The categorical variable has two outcomes - Major Damage (>$1,500) and Minor Damage (<$1,500)
- 62% of the damages come under the “Major” category

Fig 2: Distribution of Damage Type (Categorical Outcome)



3. Primary Variables of Interest

3.1 For Time Taken to Notify Police (Continuous Outcome)

While the average time taken to notify police in case of a crash is 25 mins, we can look at this distribution across different cuts such as day of the week or weather -

- When we look at the distribution of time taken to notify police during different times of the day, the median time taken to notify is highest between 12PM to 4PM and lowest between 12AM to 4AM
- For damages above $1,500, time taken to notify police (103 minutes) is lesser than time taken for damages between $501-$1,500 (133 minutes)
- We observe that mondays have a much higher average time to notify police as compared to Sundays
- For Crash Type "no Injury", the average time taken is 134 minutes, which is almost 3x the time taken for crash type "Injury", which is 46 minutes

Fig 3: Distribution of Time Taken to Notify Across Different Times of the Day

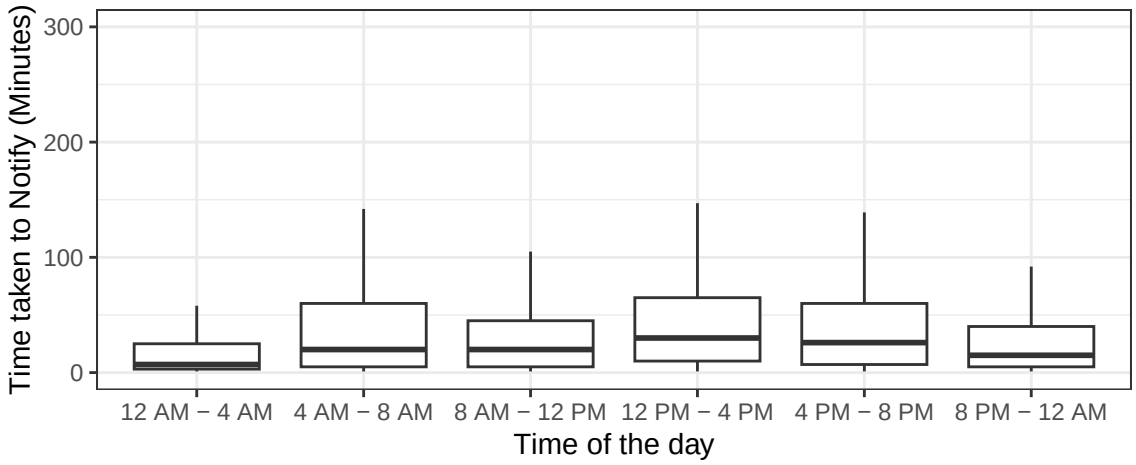


Fig 4: Distribution of Average Time Taken to Notify Across Different Damage Types

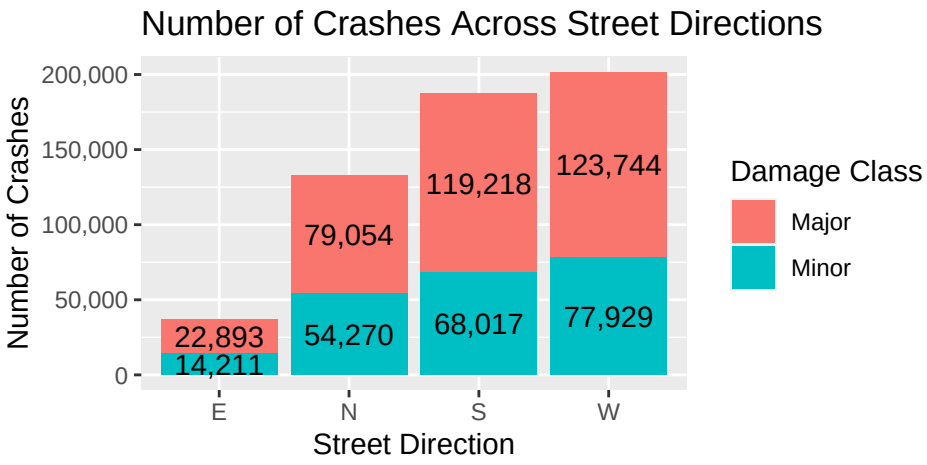
Damage ($)	Average time taken to notify Police (Minutes)
$500 OR LESS	97
$501 - $1,500	133
OVER $1,500	103

3.2 For Damage Class (Categorical Outcome)

Major Damage (>$1500) is much more common than Minor Damage (<$1500) in the Chicago Traffic Crashes data. We can view this distribution across different dependent variables such as street direction, or posted speed limit.

- Street directions South and West have a higher number of crashes, both major as well as minor
- Posted speed limit 30 has the highest number of crashes, with 258,700 major damage and 155,725 minor damage
- Intersection related crashes have a much higher share of major damages than minor damages

Fig 5: Number of Crashes across Street Directions

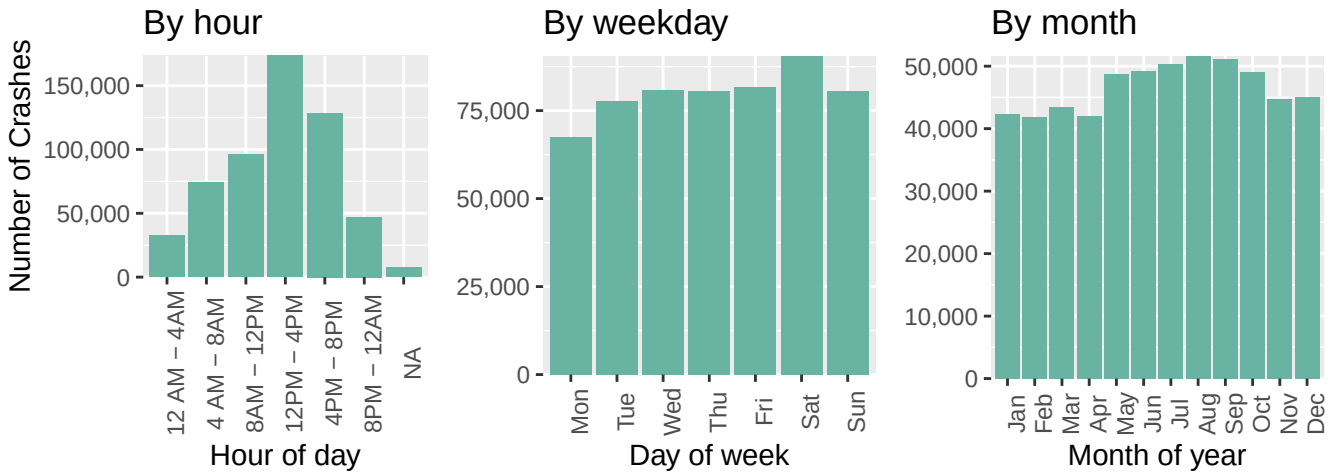


4. Other Characteristics

4.1 Date and Time Data

Information like crash hour (between 0 to 23 based on hour of the day), crash day of week (between 0 to 7 depending on day of the week), and crash month (January to December depending on the month) can be used to ascertain the times when the highest amount of crashes occurred.

Fig 6: Number of Crashes across the Year, Week, or Day



- Most common time of the day - 12PM to 4PM
- Most common days of the week - Fridays and Saturdays
- Most common crash hours - July, August, and September

4.2 Weather and Lighting Data

Similar analysis can be done for weather and lighting details, and road conditions. We observed that **65.6%** crashes occur during daylight, **79.8%** occur with clear weather, and **44%** occur in non-divided roads.

5. Potential Challenges

- Code is missing in the original source, so we need to interpret variables based solely on variable names.
- Missing values or unknown values in columns such as device condition, road surface condition, and road defects need to be addressed or removed, depending on different circumstances.
- Outliers and unreasonable values in variables, such as police response times exceeding 1000 minutes in the dataset, need to be cleaned and adjusted.
- For categorical variables, some, like primary contributory cause, posted speed limit, and weather conditions, exhibit imbalances with one category dominating.
- If we manually clean and fill in the dataset, many assumptions need to be made, which might not accurately reflect real-world situations.
- The predictors in the dataset are categorical variables. It may be necessary to merge this dataset with another one to include additional continuous variables, such as daily temperature or distance to downtown.