**Table of contents**

[**Introduction** 2](#_Toc200035621)

[**1.** **Preparing Data** 2](#_Toc200035622)

[**1.1)** **Load and explore data**: 2](#_Toc200035623)

[**1.2)** **Handling missing data:** 2](#_Toc200035624)

[**1.3)** **Type conversion and feature engineering:** 2](#_Toc200035625)

[**1.4)** **Management of outliers** 3](#_Toc200035626)

[**1.5)** **A reduction of categorical variables** 3](#_Toc200035627)

[**1.6)** **Last verification:** 3](#_Toc200035628)

[**1.7)** **Export new dataset:** 3](#_Toc200035629)

[**2. Visualization and Reason** 4](#_Toc200035630)

[**2.1)** **Visualization 1: Weekday Total Crashes** 4](#_Toc200035633)

[**Findings analysis:** 4](#_Toc200035634)

[**2.2)** **Visualization 2: Seasonal Average Monthly Crash Pattern** 5](#_Toc200035635)

[**Findings analysis:** 5](#_Toc200035636)

[**Visualization Choice Reason:** 5](#_Toc200035637)

[**2.3)** **Visualization 3: Geographic Distribution of Road Crashes** 6](#_Toc200035638)

[**Findings analysis:** 6](#_Toc200035639)

[**2.4)** **Visualisation 4: Crash Severity by Hour** 7](#_Toc200035640)

[**Findings analysis:** 7](#_Toc200035641)

[**2.5)** **Visualization 5: Crash Severity by Day of Week** 8](#_Toc200035642)

[**Findings analysis:** 8](#_Toc200035643)

[**2.6)** **Visualization 6: Proportion of Crash Severity by Roadway** 9](#_Toc200035644)

[**Analysis of Findings:** 9](#_Toc200035645)

[**Visualization Choice Reason:** 9](#_Toc200035646)

[**3.** **Pattern discussion and interpretation** 10](#_Toc200035647)

[**3.1)** **Temporal Patterns** 10](#_Toc200035648)

[**3.2)** **Geographical Pattern** 10](#_Toc200035649)

[**3.3)** **Insights on severity** 10](#_Toc200035650)

[**4.** **Data Visualization Ethics** 10](#_Toc200035651)

[**5.** **Conclusion:** 11](#_Toc200035652)

# **Introduction**

Road safety is crucial in the world or especially in Queensland by understanding road crash patterns, contributory variables, and trends is critical for effective interventions and public safety measures. This report will analyze Queensland's Government road crash data by using R for data preparation and visualization to examine crash patterns, severity, geographical distribution, and contributing factors.

After visualizations, patterns are interpreted and evidence-based recommendations are offered. This study has two primary sections: data preparation and visualization, and a detailed discussion of the findings, analytical process, and ethics.

# **Preparing Data**

Any solid evaluation is built on a foundation of reliable information. This research uses "crash\_incidents.csv," a Queensland road traffic crash dataset, which was cleaned and preprocessed in R in the following steps: Here are the steps:  
1. Load data

2. Remove missing values.  
3. Convert date column.  
4. Handle outliers.  
5. Reduce categorical variables with excessive categories.  
6. Create more features.  
7. Export the cleaned data as CSV.

## **Load and explore data**:

The raw dataset was loaded using **fread()** for efficient handling of huge data. Initial inquiry included assessing the dataset's dimensions (rows and columns), variable data types, and missing values. Understanding the data's extent and quality required this stage. The script calculated each column's missing value % to find problematic values.

## **Handling missing data:**

Missing value strategy was implemented. Columns with above 50% missing data were removed from the dataset as untrustworthy. The median value was imputed for missing numeric columns to prevent outliers. Missing categorical columns were imputed using the mode (the most frequent category) or "Unknown" if no mode could be found or all values were NA. This allowed the study of these records without bias.

## **Type conversion and feature engineering:**

Several columns needed data type conversion. Lubridate functions converted date and time columns into datetime objects. The Crash\_Year, Crash\_Month, Crash\_Day, and Crash\_Day\_Of\_Week characteristics, needed for time-series analysis and temporal patterns, were derived from these.

To maintain chart logic, the visualization script ordered Crash\_Severity into Property Damage Only, Minor Injury, Medical care, Hospitalisation, and Fatal. Similarly, Crash\_Day\_Of\_Week was chronological.

Vehicle\_age\_at\_crash, obtained by subtracting the vehicle\_year from the crash\_date\_year and ensuring both were numeric, was engineered. Negative numbers indicating data discrepancies were NA.

## **Management of outliers**

The 1.5 \* IQR rule was used to identify outliers in numeric features omitting newly created date components. To avoid affecting statistical summaries and visualizations, outliers were capped at the lower and upper boundaries while maintaining the information that an extreme number occurred.

## **A reduction of categorical variables**

Highly cardinality categorical variables (>100 unique values) can clutter visualizations and make interpretation difficult. For such columns, the script kept the top 50 most frequent categories and sorted the rest into "Other". The technique balances detail and interpretability.

## **Last verification:**

Critical columns with keywords like 'crash', 'fatal','severity', and 'injury' were identified and rows with missing data were removed as a final cleaning step. This ensured complete core data for studying crucial outcomes like severity.

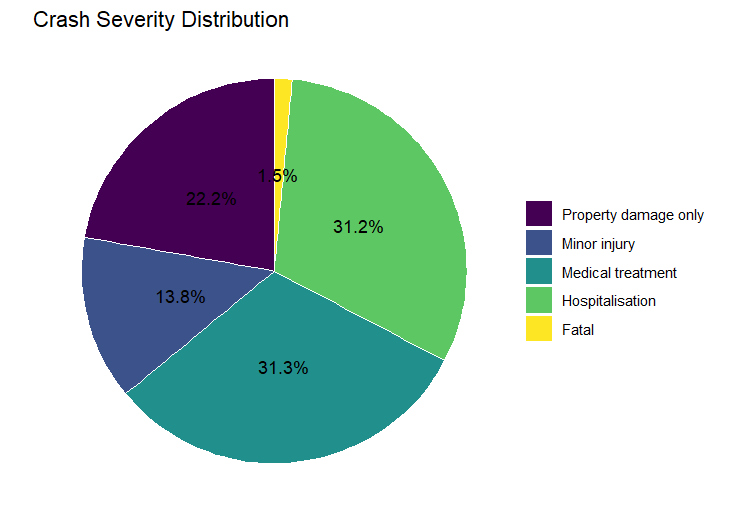
## **Export new dataset:**

Final step is export it into new dataset and use that new dataset for further analysis. New data file name “crash\_incidents\_cleaned.csv”

# **2. Visualization and Reason**



## **Visualization 1: Crash Severity Distribution**



**Figure 1: Crash Severity Distribution**

**Findings analysis:**  
In Figure 1, we can observe a general insight from the dataset though out many kinds of severities. Looking at the pie chart we can see the the distribution of “minor injury” and “medical treatment” have a close percentage of distribution (31.2% and 31.3% respectively). After that is the “property damage only” with 22.2% and minor injury with 13.8%. Fatal still spot at a very low value about 1.5% but still need to consider for further investigation.  
  
**Visualization Choice Reason:**

This pie chart act as a general openning for further deep dive into the dataset later on. This is ideal for highlighting which categories place at the high proportion which is low. A fundamental chart for further analysis later on

## **Visualization 2: Seasonal Average Monthly Crash Pattern**

A graph with green lines

AI-generated content may be incorrect.

**Figure 2: Average Monthly Crash Pattern in Queensland**

**Findings analysis:**  
Figure 1 shows a seasonal pattern in Queensland road crashes. The average number of crashes is lowest in January (around 1250 crashes) and peaks in March (around 1480 crashes). May, July, and August also peak, with August having the highest average (over 1500 crashes). September dips before a smaller peak in October, followed by a decline at the end of the year.

**Visualization Choice Reason:**  
A line chart was chosen for this time-series data because it clearly shows the rise and fall in average crash numbers, making it easy to identify peaks, troughs, and seasonality. The group = 1 aesthetic in ggplot2 ensures a single line connects the monthly averages, emphasizing the cyclical trend.

## **Visualization 3: Geographic Distribution of Road Crashes**



**Figure 3: Geographic Distribution of Sampled Road Crashes in Queensland**

In this section we will select about 5000 sample point to get a general observation from the dataset, due to when set all the point we can see the plotted-data point clearly, selection of 5000 points giving us a overal assumption for the data.

**Findings analysis:**  
In Figure 3, road crashes are widespread in Queensland, but clusters are visible in and around major urban centers like Brisbane, Gold Coast, Sunshine Coast, Townsville, and Cairns (inferred from typical Queensland geography, as city labels are not clear on the image). Rural and remote areas have more dispersed incidents.  
  
**Visualization Choice Reason:**  
An interactive point map using leaflet is the best way to show the geographic distribution of incidents. Users can zoom and pan to explore specific areas in a live environment. Coloring points by Crash\_Severity adds another layer of information, helping to determine if high-risk areas correlate with higher severity outcomes.

## **Visualisation 4: Crash Severity by Hour**

A graph of a graph showing the amount of time

AI-generated content may be incorrect.

**Figure 4: Crash Severity by Hour of Day**

**Findings analysis:**  
Figure 4 shows that crash frequency and severity vary throughout the day. "Minor injury" (light green) and "Medical treatment" (orange) crashes account for most crashes, but "Hospital" crashes account for a small percentage. To be more specific, in late-night hours, despite having fewer crashes overall, might show a relatively higher proportion of severe crashes, which would be a critical insight for targeted enforcement or safety campaigns. The plot shows "Fatal" crashes as a very small proportion across all hours, but consistently present.  
  
**Visualization Choice Reason:**  
The position = "stack" argument in ggplot2 allows a stacked bar chart to display the total number of crashes per hour and the severity composition within each hour. A distinct color palette (like "Spectral" in the code or the one shown in the plot) helps differentiate severity levels.

## **Visualization 5: Crash Severity by Day of Week**

A graph of a graph showing the results of a crash

AI-generated content may be incorrect.

**Figure 5: Crash Severity Distribution by Day of Week**

**Findings analysis:**  
Figure 5 shows how crash severity distributes across the week. Friday has the highest total number of crashes (Figure 2), but this chart shows the severity breakdown. Weekdays (Monday-Thursday) have a large number of "Minor injury" (light green) and "Medical treatment" (orange) crashes, as does Friday. "Hospitalisation" (red) crashes peak on Friday and remain high on Saturday. Fatal crashes (blue) seem relatively consistent but might show slight increases on weekends (Friday/Saturday), although their overall volume is much lower than other categories.  
  
**Visualization Choice Reason:**  
Like severity by hour, a stacked bar chart shows both the total crashes per day and their severity composition, helping to determine if the risk profile (in severity) changes with the week. Position = "stack" and a clear color scheme for severity levels are essential for interpretability.

## **Visualization 6: Proportion of Crash Severity by Roadway**

A graph of crash on a road

AI-generated content may be incorrect.

**Figure 6: Proportion of Crash Severity by Roadway**

**Analysis of Findings:**

Figure 6 provides insights into how crash severity profiles differ across various roadway features. For instance, crashes at "No Roadway Feature" (which might indicate mid-block sections or straight roads) appear to have a certain proportion of "Minor injury", "Medical treatment", and "Hospitalisation". Intersections like "T-Junction" and "Cross Intersection" also show their respective severity distributions. It's important to look for features that have a disproportionately higher percentage of severe outcomes (Hospitalisation or Fatal). The plot suggests that "Minor injury" (light green) is the predominant outcome across most listed features. The "Fatal" (dark blue/purple) proportion appears small across all features shown but is still present.

**Visualization Choice Reason:**  
The proportion of crash severities across different roadway features was compared using a 100% stacked bar chart to normalize the data and compare the severity mix. Filtering for roadway features with a large number of crashes (e.g., >1000 as in the code) helps focus the comparison.

# **Pattern discussion and interpretation**

## **Temporal Patterns**

* Crash rates peak in Autumn (March/May) and late Winter/early Spring (August/October), then drop in January. Holidays, weather (e.g., Queensland wet/dry seasons), and school terms might alter traffic.
* Weekly cycle: End-of-week commuter traffic, social travel, and tiredness may cause the most collisions on Fridays, Sunday has the fewest.
* Daily cycle: Morning and afternoon commuter hours are crash-heavy (8 AM and 3-5 PM). Late-night crashes are rarer, but the fraction of severe crashes is worth investigating (Figure 4 displays absolute numbers, a common pattern). Work/school scheduling, traffic congestion, and driver attention are factors.

## **Geographical Pattern**

* Due to higher traffic volumes and population density, crashes are concentrated in urbanized areas and along major transport corridors. However, crashes in regional or remote areas may be more severe due to higher speeds, longer emergency response times, or different road characteristics.

## **Insights on severity**

* Most crashes result in "minor injury" and "medical treatment" at varied times and places.  
  While intersections are common crash locations, the proportion of severe crashes at "Bridge/Causeway" or "Median Opening" should be examined (Figure 6 suggests minor injuries are dominant, but specific high-severity proportions are key).
* Fridays and Saturdays may have high crash numbers and hospitalization-level injuries.  
  Possible Contributors (General Discussion):
* Several factors can affect observed patterns, including:
  + Drunk/drug driving, speeding, distraction, weariness.
  + Road design (intersections, curves, lighting), surface, and safety obstacles.
  + Vehicle factors: Age, safety, upkeep.
  + Traffic: Peak-hour congestion, mixed vehicles.
  + Environmental factors: Rain, fog, daylight/darkness.
  + Demographic factors: Density, travel (commuting, leisure).
  + Further research linking crash data to these external factors may reveal causative links.

# **Data Visualization Ethics**

Several industry-standard ethical guidelines were followed when evaluating and presenting road crash data:

**Truthfully in representation:**

* **Application:** Bar chart axes start at zero to avoid emphasizing differences, and outlier capping (unless seriously lacking) protected data integrity. The map's sampling and performance were noted.
* **Relevance:** Misrepresenting data in road safety analysis can lead to poor policy decisions, resource allocation, and public mistrust. Honesty helps identify problems.

**Being objective and bias-free:**

* **Application:** Colors, chart types, and aggregations were chosen to present findings objectively, not to sensationalize or downplay specific aspects. For example, using normalized percentages for comparing roadway feature severity (Figure 6) avoids bias caused by extremely different total crash numbers per feature type.
* **Relevance:** Analysts must present data neutrally so stakeholders can draw their own conclusions. Even unintentional bias can skew perceptions and lead to flawed interventions..

**Clarity, Transparency, Interpretability:**

* **Application:** Clear titles, axis labels, and legends helped understand visualizations. The rationale for choosing specific chart types was to best convey the intended insight..
* **Relevance:** Data visualizations should clearly communicate complex information to a diverse audience. Confusing, misleading, or lack-of-context visualizations are less useful and can even be harmful if misinterpreted. Transparency in methods builds trust and allows scrutiny and replication.

# **Conclusion:**

Seasonal and weekly changes in crash situations, urban event focus, and severity profiles linked with different times of day and roadway features are seen in Queensland road crash data. Fridays and peak commuter hours are crash-prone, and roadway designs may contribute.

General recommendations from these studies: Friday road safety activities should increase. Safety auditing highway characteristics having a higher proportion of severe crashes, even if crash data are low. Seasonality and higher crash rates on certain days/times can be determined using weather, traffic flow, and local events data.