Report on Data-Driven Review of Road Crashes in Queensland (2001–2022)

Bhavesh Hemant Vasnani

Table of Contents

Table of Figures

1. Introduction

This report analyses crash incident data in Queensland from 2001 to mid-2022. It discusses temporal, geographical and behavioural patterns that influence crash frequency and severity. Key insights were derived from the visualisations to provide data-driven recommendations aimed at improving road safety and reducing serious crash outcomes.

2. Data Cleaning and Preparation

Data preparation and cleaning are essential and ensure that the results of the subsequent analysis are accurate and reliable. For the "crash_incidents.csv" dataset, the following steps were taken to remove null values, and eliminate unnecessary columns in the dataframe:

- Imported raw data ("crash_incidents.csv") into a pandas DataFrame.

```
In [2]: df = pd.read_csv('crash_incidents.csv')
```

Figure 1 Code snippet for importing data

- Dropped non-essential columns with high null values such as Crash_Street_intersecting and State_Road_Name, with 222947 and 212150 null values, respectively.

```
In [4]: #Checking missing values

        df.isna().sum()

Out[4]: Crash_Ref_Number              0
        Crash_Severity               0
        Crash_Year                   0
        Crash_Month                  0
        Crash_Day_Of_Week            0
        Crash_Hour                   0
        Crash_Nature                 0
        Crash_Type                   0
        Crash_Longitude            598
        Crash_Latitude             598
        Crash_Street                15
        Crash_Street_Intersecting  222947
        State_Road_Name            212150
```

Figure 2 Code snippet for checking null values in columns

```
In [7]: # Drop the two unwanted columns
        df = df.drop(columns=['Crash_Street_Intersecting', 'State_Road_Name'])
```

Figure 3 Code snippet for dropping non-essential columns

- Imputed missing row values in DCA_Key_Approach_Dir using the mode (most frequent) of the column. Mode works best for categorical data, and therefore, the

measure of central tendency was used for imputing DCA_Key_Approach_Dir in this instance.

```
# Compute the mode of DCA_Key_Approach_Dir
mode_val = df['DCA_Key_Approach_Dir'].mode()[0]

# Replace NA's in DCA_Key_Approach_Dir with that mode
df['DCA_Key_Approach_Dir'] = df['DCA_Key_Approach_Dir'].fillna(mode_val)
```

Figure 4 Code snippet for imputing DCA_Key_Approach_Dir column with mode

- Dropped rows with null values in Crash_Speed_Limit, Crash_Road_Vert_Align, Crash_Longitude, Crash_Latitude, and Crash_Street.

```
In [9]: # Drop rows with nulls in either column
        df_clean = df.dropna(subset=['Crash_Speed_Limit', 'Crash_Road_Vert_Align', 'Crash_Longitude', 'Crash_Latitude', 'Crash_Street'])
```

Figure 5 Code snippet for dropping rows with null values in various columns

- Verified that no required columns contained null values.

```
In [10]: #Checking missing values

         df_clean.isna().sum()
```

```
Out[10]: Crash_Ref_Number                  0
         Crash_Severity                    0
         Crash_Year                        0
         Crash_Month                       0
         Crash_Day_Of_Week                 0
         Crash_Hour                        0
         Crash_Nature                      0
         Crash_Type                        0
         Crash_Longitude                   0
         Crash_Latitude                    0
         Crash_Street                      0
         Loc_Suburb                        0
         Loc_Local_Government_Area         0
         Loc_Post_Code                     0
         Loc_Police_Division               0
         Loc_Police_District               0
         Loc_Police_Region                 0
         Loc_Queensland_Transport_Region   0
         Loc_Main_Roads_Region             0
         Loc_ABS_Statistical_Area_2        0
         Loc_ABS_Statistical_Area_3        0
         Loc_ABS_Statistical_Area_4        0
         Loc_ABS_Remoteness                0
         Loc_State_Electorate              0
         Loc_Federal_Electorate            0
         Crash_Controlling_Authority       0
         Crash_Roadway_Feature             0
         Crash_Traffic_Control             0
         Crash_Speed_Limit                 0
         Crash_Road_Surface_Condition      0
         Crash_Atmospheric_Condition       0
         Crash_Lighting_Condition          0
         Crash_Road_Horiz_Align            0
         Crash_Road_Vert_Align             0
         Crash_DCA_Code                    0
         Crash_DCA_Description             0
         Crash_DCA_Group_Description       0
         DCA_Key_Approach_Dir              0
         Count_Casualty_Fatality           0
         Count_Casualty_Hospitalised       0
         Count_Casualty_MedicallyTreated   0
         Count_Casualty_MinorInjury        0
         Count_Casualty_Total              0
         Count_Unit_Car                    0
         Count_Unit_Motorcycle_Moped       0
         Count_Unit_Truck                  0
         Count_Unit_Bus                    0
         Count_Unit_Bicycle                0
         Count_Unit_Pedestrian             0
         Count_Unit_Other                  0
         dtype: int64
```

Figure 6 Code snippet for verifying null data values in columns

- Exported the cleaned DataFrame into CSV to be used for analysis in R.

```
In [13]: # Export df_clean to CSV (no index column)
         df_clean.to_csv('crash_incidents_clean.csv', index=False)
```

Figure 7 Code snippet for exporting the clean dataframe

3. Visualisation, Analysis and Rationale

3.1 Time Series Chart of Monthly Crashes

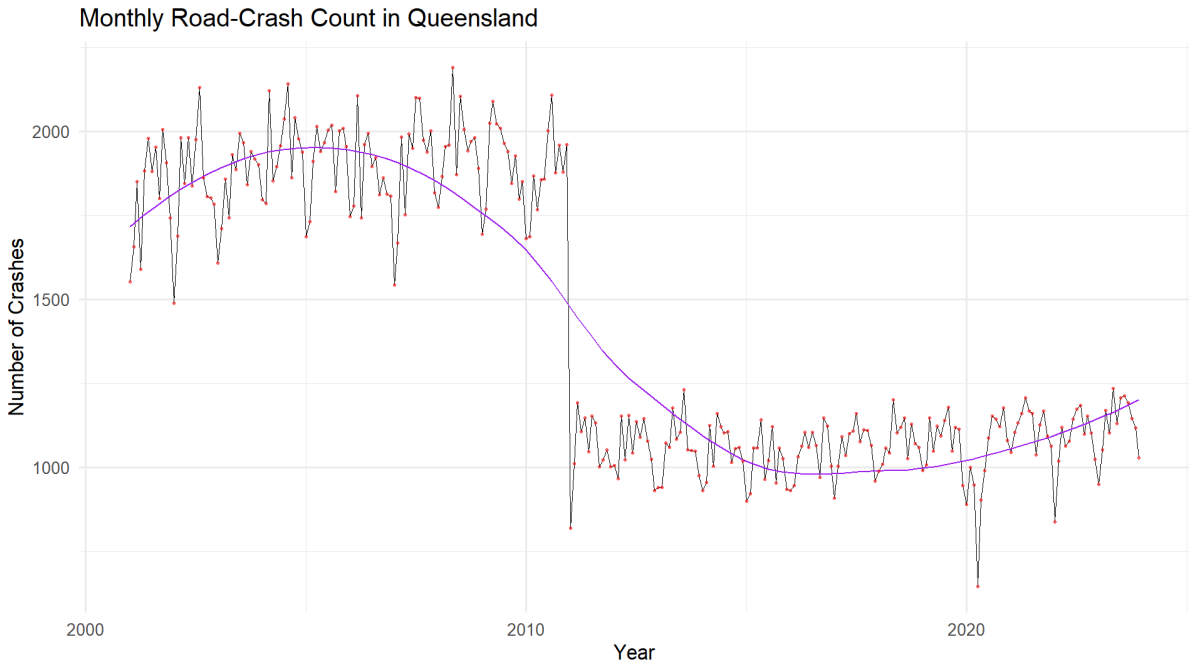

Figure 8 Monthly Road Crash Time Series

Description: A black line chart is used to demonstrate the monthly crash counts from January 2001 to mid-2022, with semi-transparent red points at each month and a purple smoothed trend line overlay.

Rationale: A time series chart is beneficial for observing patterns in the data. The red points indicate month-to-month volatility, while the smoothed purple line identifies the underlying trajectory.

Key Findings:

1. 2001-2010: An average of 1900 crashes per month with significant short-term fluctuations.
2. 2011-2015: There was a sharp decline in early 2011, dropping from an average of 1900 per month to 1000 per month.
3. Post-2015: The trend line remains constant till 2019, after which a COVID-19-related trough is observed in April 2020. A steady recovery is observed from 2020 to 2022, with crash counts averaging around 1200 per month.

Interpretation: The steep post-2010 decline suggests effective implementation of road safety rules and policy measures. The drop in April 2020 aligns with the lockdown-driven mobility

reductions during the pandemic. The gradual rise from 2020 signals the need for enhanced intervention by the government to increase public safety on roads.

3.2 Geographic Distribution of Crashes

The Geographic Distribution style visualisation is implemented using two approaches, namely a Point Map highlighting individual crash points by severity, and an Aggregated Bubble Map where a Bubble represents each Local Government Area.

Description: First approach, a Google basemap of Queensland displays every crash site as a point, coloured by severity category. Points are semi-transparent (alpha = 0.2) to alleviate overplotting. Second approach, the crash counts are grouped by Local Government Area (LGA) and are represented by a bubble on the same basemap.

Rationale: First approach, a point map on a basemap is an intuitive way to display individual crash locations. Clusters can be observed to determine the density of crashes in a geographical area. Second approach, aggregating by LGA, reduces overplotting and clarifies regional hotspots.

3.2.1 Geographic Distribution of Crashes by Severity

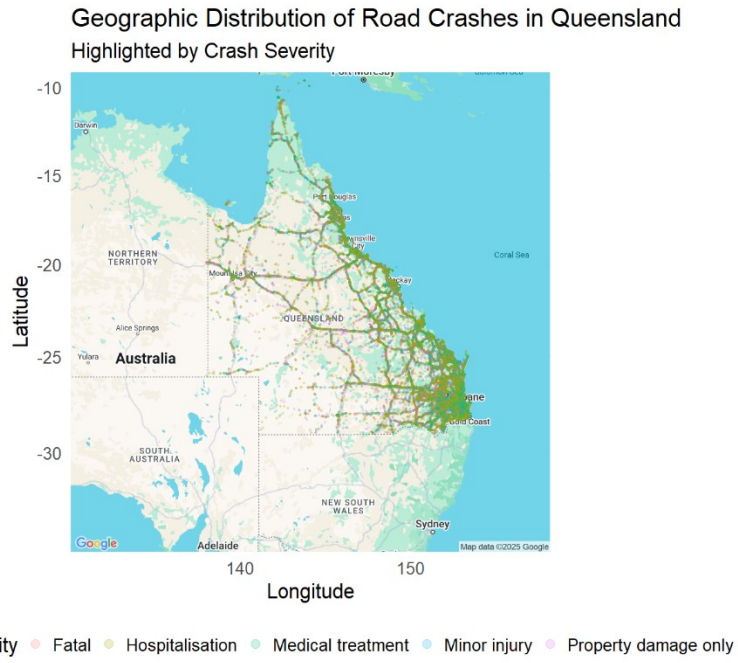

Figure 9 Geographic Distribution by Severity

Key Findings:

1. Southeast Coastal Corridor: A very high density of crashes across all severities can be observed.

2. Secondary Clusters: Townsville, Cairns and Rockhampton include medium-sized population hubs and important national highways.
3. Rural Inland: The crashes are sparsely distributed in this region, but fatal/hospitalisation points are concentrated along major highways.

Interpretations: Urban areas generate the majority of crashes due to dense populations and higher traffic volume. Fatal and hospitalisation clusters along major highways indicate the risks associated with long-distance driving and speed, and the need for better safety measures along major highways.

3.2.2 Geographic Distribution by Local Government Area

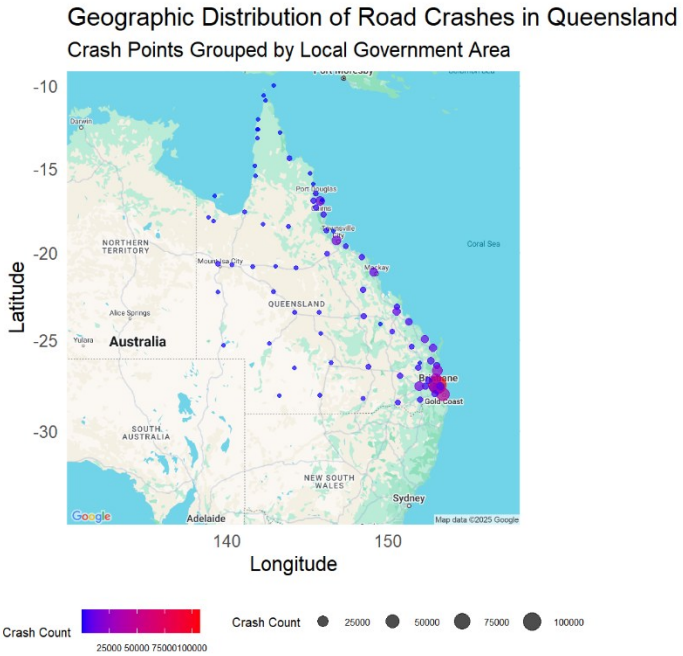

Figure 10 Geographic Distribution by LGA

Key Findings:

1. Brisbane City LGA: Has the largest bubble (approximately 100,000 crashes), followed by Logan, Gold Coast, and Ipswich.
2. Northern Queensland: Cairns and Townsville have medium-sized bubbles, indicating a moderate number of crashes.
3. Remote LGAs: Small Bubbles are observed, but high per-capita fatalities due to sparse population

Interpretation: The concentration in metropolitan LGAs, such as Brisbane, suggests the need for stricter enforcement in populous regions. Regional LGAs require improvements in safety on key highways and intersections.

3.3 Seasonal Boxplot of Monthly Crash Counts

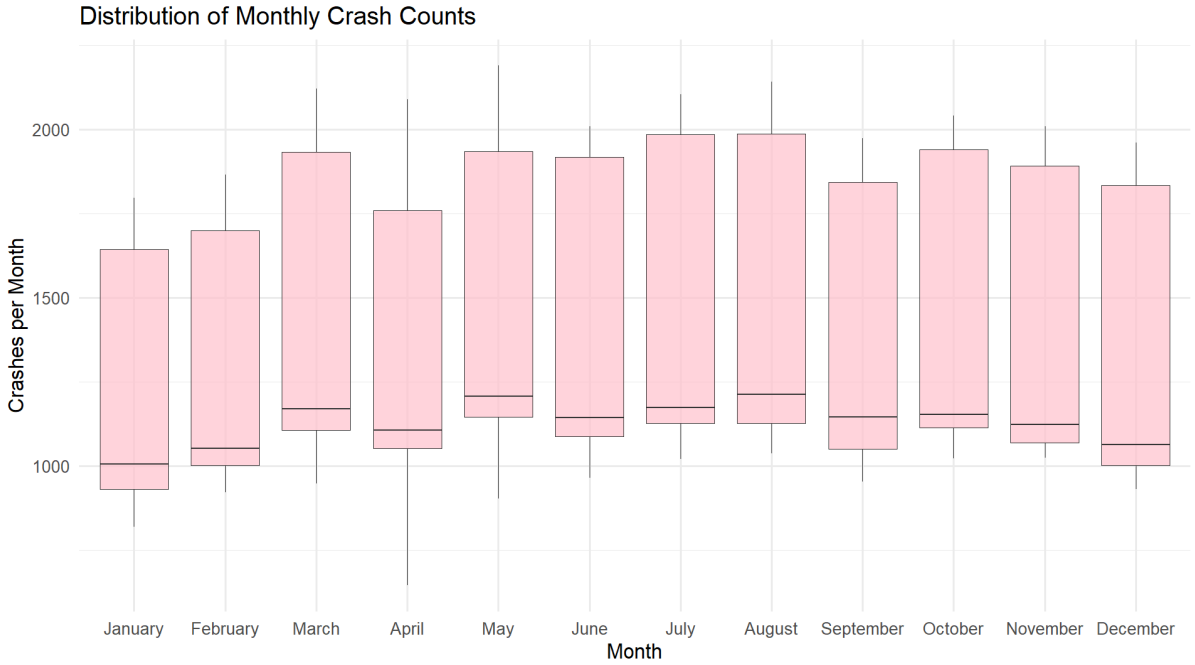

Figure 11 Monthly Crash Counts by Month (Boxplot)

Description: Boxplots for each month visualise the distribution of monthly crash counts across all years. The median is marked by a central black line, and the whiskers extend to 1.5 x IQR.

Rationale: Boxplots concisely depict the distribution, variability, and outliers per month. Ordering months chronologically can also reveal seasonal patterns.

Key Findings:

1. March, May, July and August have the highest medians due to weather conditions such as rain. Whereas December, January and February have the lowest medians, likely due to summer holidays.

Interpretation: Peak crash volumes are observed in cooler and work-dominated months when commuting is highest and weather conditions are not favourable. Lower crash volumes during the peak holiday season indicate lower commuter traffic.

3.4 Crash Severity and Hospitalisation by Hour of Day

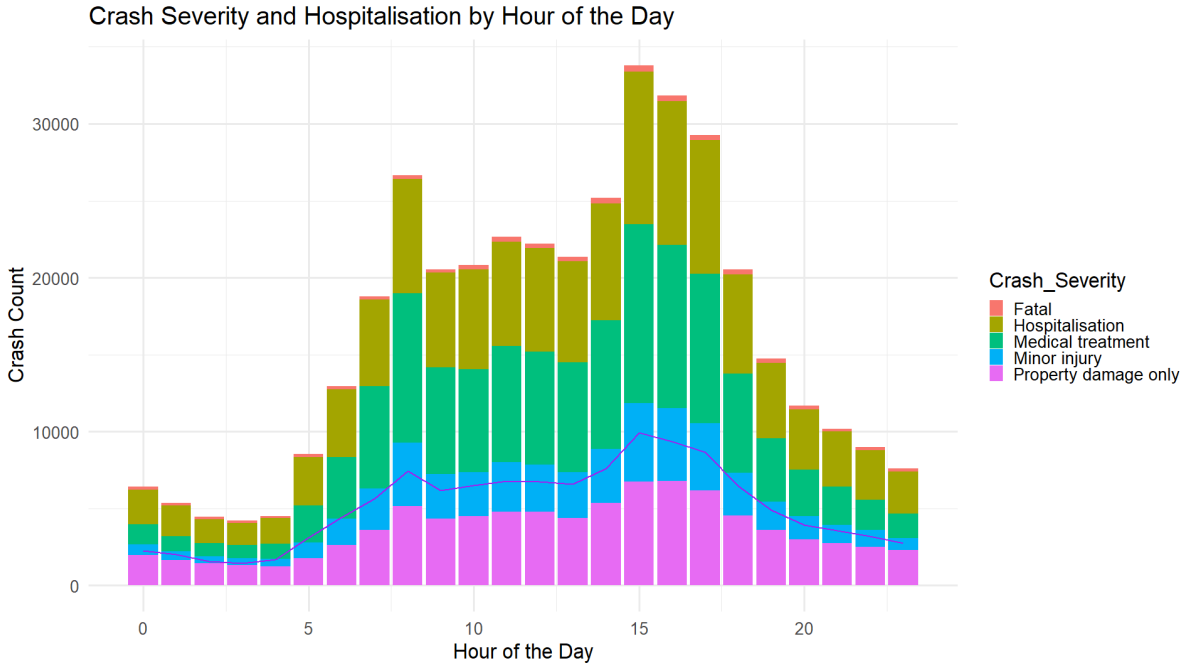

Figure 12 Crash Severity by Hour of the Day

Description: A 24-hour bar plot (hours 0-23) stacked by severity category. A purple smoothed line overlays the counts of hospitalisation by hour.

Rationale: The stacked bar reveals both total hourly volume and relative severity breakdown during the hour. The overlaid line focuses on hospitalisation counts specifically to focus on high-risk incidences.

Key Findings:

1. 00:00–04:00 observe the lowest volume, and they begin rising at 05:00. On the other hand, peak volumes can be observed from 07:00-08:00 and 15:00-16:00, coinciding with morning commute, afternoon commute, and school runs.
2. Peak hospitalisations between 14:00-17:00 indicate increased severity in late afternoons.
3. 18:00-06:00 have higher proportions of severe crashes, reflecting the risk of fatigue and low visibility during the night.

Interpretation: Late-afternoon peaks suggest the need for increased enforcement during commuter and school times. Increased night-time severity highlights the need for improved road lighting and impaired-driving countermeasures.

3.5 Trend of Hospitalisation Crashes

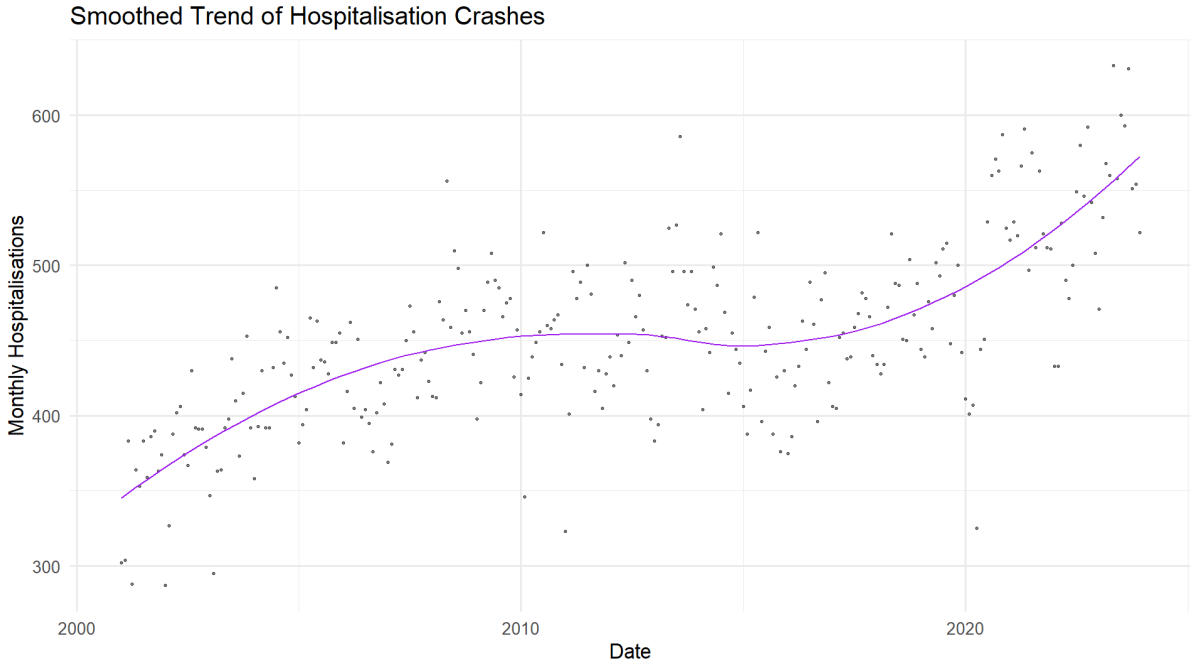

Figure 13 Trend of Hospitalisation Crashes

Description: Monthly hospitalisation crash counts from 2001 to mid-2022 with a purple smoothed line.

Rationale: A scatterplot with smoothing focusing on the hospitalisation subset removes noise and aids in focusing on severe crashes. Neutral grey points keep focus on the purple trend line.

Key Findings:

1. 2001-2010: Hospitalisations rose from 300 to 500 per month.
2. 2010-2016: They plateau around 400 – 500 per month.
3. 2017-Mid-2020: Declined to around 380 per month, then a COVID-related trough at around 330 per month.
4. Post-2020: A sharp rise to 650 by mid-2022.

Interpretation: The post-2020 surge in hospitalisations despite moderate increases in total crashes is suggestive of riskier driving behaviour and a lack of safety innovation in modern cars.

4. Patterns, Trends, and Factors Influencing Crash Incidence and Severity

Patterns and trends observed in the crash data suggest a direct relationship between temporal, geographic, and behavioural factors. A significant decline in incidents after 2010 is

directly related to improved road safety policies by the government. The COVID-19-related trough in April 2020 is a result of reduced mobility due to lockdown restrictions. The increase in crashes after 2020 indicates the rising number of cars on the road, and possibly due to increased reckless driving.

Geographically, population-dense urban areas like Brisbane City have a higher frequency of crashes due to higher traffic levels. Fatal and hospitalisation incidents along major highways highlight the risks associated with speed and fatigue, and long-distance driving. Seasonally, higher crashes in cooler and wetter months indicate risks related to weather conditions and inadequate infrastructure. On the other hand, the holiday season exhibited a lower number of instances, indicating lower commuting and stricter law enforcement during the period.

Hourly analysis reveals higher incidents during commuting hours and increased severity during late afternoon and nighttime. The post-2020 surge in hospitalisations despite a moderate increase in crashes suggests severe incidents due to factors such as riskier driving behaviours, distracted driving, or lagging vehicle safety technology. Thus, initiatives like targeted enforcement, improved road infrastructure, and behavioural interventions are essential for effective crash prevention.

5. Ethical principles applied

Adherence to ethical principles is crucial in the visualisation process to ensure accuracy, honesty, clarity, simplicity, privacy and trust (Shahazad, 2024).

First, accuracy and honesty guided my rationale for developing monthly crash counts and severity. The visualisations represented true proportions without any exaggerations, aligning with the ethical recommendations by Webber and Morn (2019), who advocate that visuals should not have misleading scales or data distortions.

Secondly, clarity and simplicity ensured visuals clearly expressed complex patterns, such as time series and geographical crash clusters, without overwhelming detail. This aligns with Tufte's (2001) principles of precise yet straightforward representations to aid the viewer's understanding.

Lastly, privacy and trust were maintained by using tokenised crash data, which prevents mapping the data to an individual, thus adhering to ethical privacy standards. Webber and Morn (2019) highlight the importance of safeguarding personal information to uphold trust and ethics.

These principles are guidelines for ethical data practices and are essential for accurate decision-making in an organisation.

6. Conclusion and Recommendations

The analysis highlights significant patterns related to crash incidents and severity across Queensland. Policy improvements after 2010 saw a significant reduction in crashes, but recent data highlights rising hospitalisation rates, indicating reckless behaviour and inadequate vehicle safety technology. The highest number of crashes is observed during commute hours and in densely populated areas such as Brisbane, highlighting high-risk zones.

Recommendations include:

1. Enhanced Enforcement: Increase targeted police monitoring during commuter and school-run times, especially late afternoons, to mitigate the risk of crashes.
2. Infrastructure Improvement: Upgrading road lighting and safety barriers should be prioritised, particularly along major highways and roads, to reduce nighttime crashes.
3. Public Awareness Campaigns: Public awareness campaigns should be implemented to address risky behaviour (speeding, distracted driving) to reduce the rise of severe crashes after 2020.
4. Vehicle Safety Incentives: Encourage adoption of advanced safety features by automakers through policy incentives or regulation.

Adopting these strategies can significantly reduce hospitalisations, improve road safety and enhance public trust.

References

Shahazad, M. (2024, January 23). *Ethics of Data Visualization: Avoiding Deceptive Practices*. Www.analyticodigital.com. https://www.analyticodigital.com/blog/ethics-of-data-visualization-avoiding-deceptive-practices

Tufte, E. (2001). Graphical excellence. In *The visual display of quantitative information* (pp. 13–52). Graphics Press Cheshire, CT.

Webber, K. L., Morn, J., & Webber, K. (2019). Limitations in data analytics: Considerations related to ethics, security, and possible misrepresentation in data reports and visualizations. *IHE Research Projects Series*, *3*, 2019.