QLD Crash Data 2018 - Python Analysis

Tristan Hayes, 46969842

Feburary 2025

Contents

1 Abstract

This report presents an analysis of crash data with a focus on identifying key relationships between crash severity and variables such as speed limits, crash types, time of occurrence, and geographic location. The findings indicate that crashes in higher-speed zones tend to result in more severe outcomes, with fatality rates significantly increasing at speeds above 80 km/h. Single-vehicle and pedestrian crashes were observed to have the highest fatality rates, highlighting the need for targeted safety measures. A seasonal trend analysis suggested a possible anomaly in April's data, warranting further investigation into data integrity. Additionally, geographic analysis revealed that high-crash suburbs, particularly Southport and Brisbane City, are likely influenced by traffic density, road infrastructure, and urban design. These insights provide valuable guidance for road safety interventions and urban planning improvements.

2 Introduction

Understanding traffic crash data is crucial for improving road safety and identifying high-risk factors associated with severe accidents. This report analyzes crash data from 2018, focusing on relationships between crash severity and various factors, including speed limits, crash types, time trends, and geographical distribution.

The primary objectives of this analysis are:

1. Examine the impact of speed limits on crash severity and identify which speed ranges contribute to the most severe accidents.
2. To investigate the relationship between crash types and severity, determining which crash categories result in the most dangerous outcomes.
3. Analyze monthly crash trends in 2018 and assess time-based variations in crash frequency and severity.
4. Identify the top ten suburbs with the highest number of crashes and assess their severity distributions.
5. Explore any additional patterns or insights that emerge from the dataset, including geospatial trends.

Python was used for data processing and visualization, with libraries such as pandas, matplotlib, seaborn, and Plotly Express to perform statistical and graphical analysis. Data quality checks were conducted to ensure consistency and accuracy before deriving insights. This report presents a structured breakdown of the findings, supported by visualizations and interpretations.

3 Data Integrity & Cleaning

Ensuring data integrity is a critical step before conducting analysis, as inconsistencies or errors can lead to misleading conclusions. To maintain data accuracy, duplicate entries were identified and removed from the Crash Facts, Date Dimensions, and Location Dimensions datasets. Additionally, categorical values were checked for inconsistencies, revealing a minor formatting issue in the Crash Severity column where ”Medical treatment” was incorrectly recorded as ”Medical Treatement.”. This was corrected.

Furthermore, the Crash Speed Limit field contained speed ranges in string format that were not plotted in the natural order as expected. To address this, the values were converted into a categorical data type with a predefined order of increasing speed limit to ensure that they are correctly interpreted in visualizations and statistical comparisons.
These integrity checks enhance the reliability of the dataset, allowing for accurate insights in the subsequent analysis.

4 Analysis & Visualisations

4.1 Crash Frequency

Initially, we start with a basic analysis of the frequency of crashes in each speed limit zone. Note that these are raw numbers, and therefore suffer bias from certain traffic zones having a higher influx of traffic and therefore more crashes as a result. From Figure 1, we can immediately see that the number of crashes is significantly higher in zones with a speed limit of 60km/h. While this does not account for the fact that the volume of cars driving in 60km/h zones is much higher than others, it does evidently show that the most amount of crashes do occur there. Notably as well, the lowest number of crashes occurs in 70km/h zones.

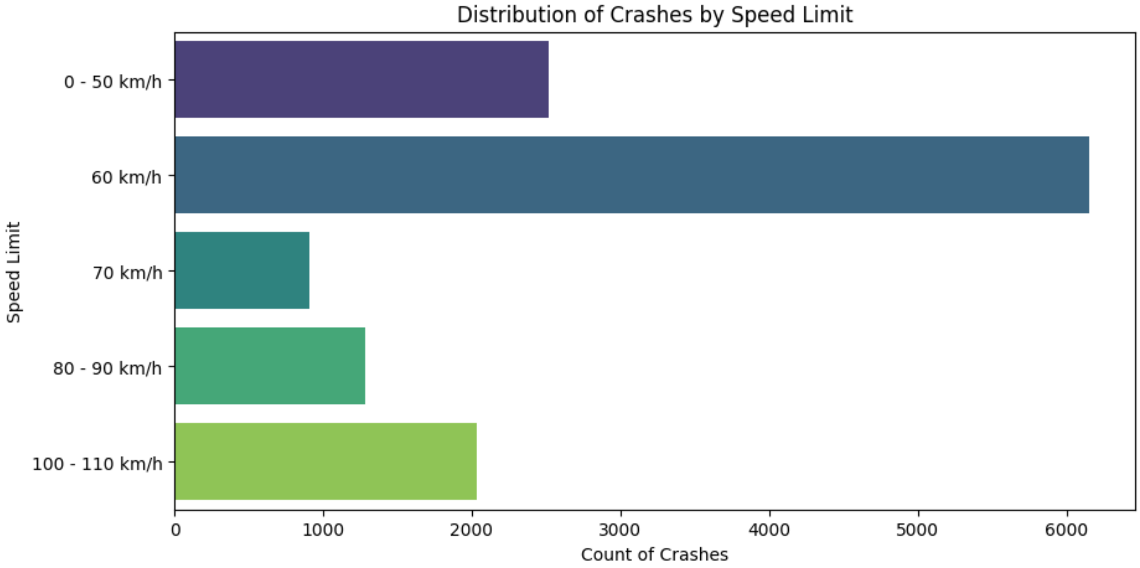

Figure 1: Crash Frequency for each Speed Limit Zone

4.2 Crash Speed Limit vs Severity

A big factor to investigate is how the the speed zone affects the severity of the incident, as this information will provide extremely insightful data into what needs to be considered most in regards to safety precautions for each zone. We can use a stacked bar graph firstly to gauge the basic trends in each zone, At a first glance at Figure 2, we can noticeably see that the proportion of fatal crashes is high in 100-110km/hr zones, as well as 80km/h, and lowers in speed zones with

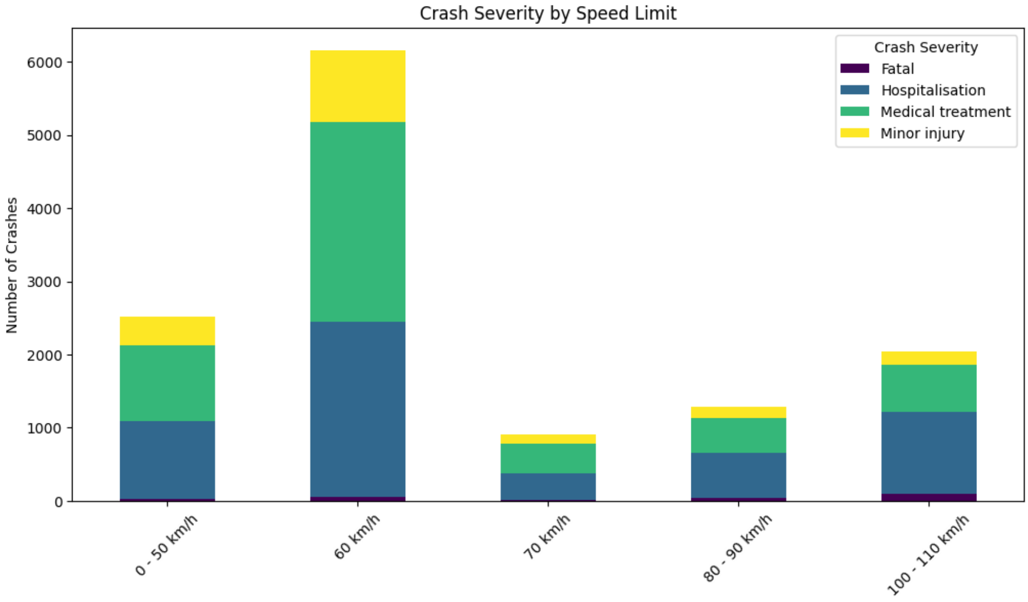

Figure 2: Severity vs Speed Limit Stacked Bar Plot

decreased speed limits. This behaviour is expected, as higher speed crashes are expected to be more lethal. However, this visualisation is hard to interpret still due to the inconsistency in number of crashes in each zone. To account for this, the data will be normalised and visualised on a heat map to investigate the proportions in each speed zone relative to the number of crashes in them respectively. Consulting Figure 3, it is immediately evident that the proportion of fatal crashes in 100-110 km/h zones is significantly higher than lower speed zones, sitting at 5/06% of all crashes in that zone. Minor injuries are also the least common, and most injuries require hospitalization or medical treatment. It is much more evident to see now that even though 60 km/h zones have significantly higher total crashes, they involve less hospitalization or fatalities than all other speed zones, which can be influenced by a variety of factors. Notably, the fatalities in the 0-50km/h zones are of the same proportion as the 60km/h zones which can be due to factors such as an increase in pedestrian-vehicle collisions, etc.

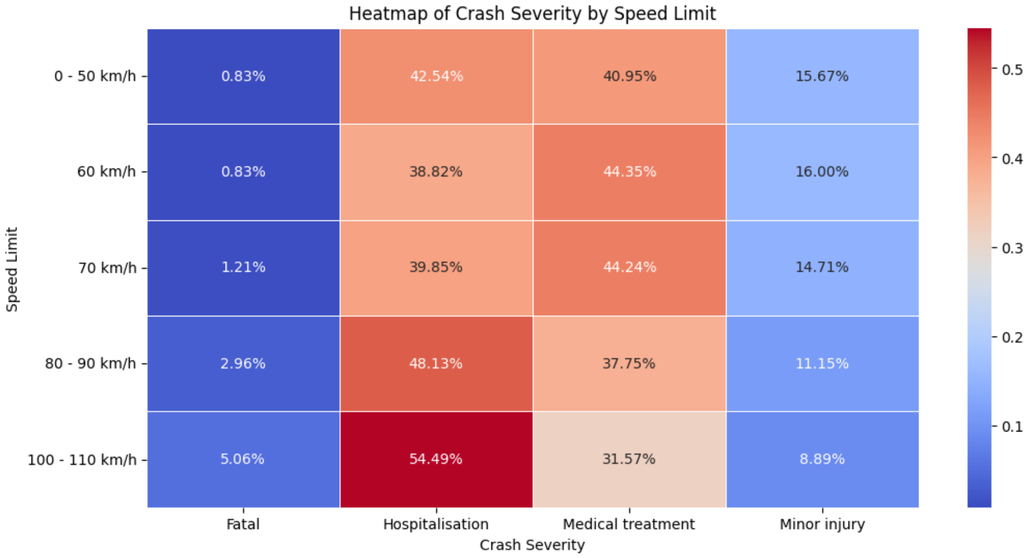

Figure 3: Severity vs Speed Limit Heat Map

4.3 Crash Type vs Severity

Another important factor to consider is the type of crash vs the severity. While understanding how the speed zone affects the severity is crucial, understanding what the specific accidents are that cause the most severe outcomes more easily allows for action to be taken. Consulting Figures 4 and 5, we can uncover some interesting trends. Noticeably at a first glance of Figure 4, we can notice that the proportion of fatal incidents in single vehicle accidents is large compared to the other types, and so we look to Figure 5 to confirm this. Notably, incidents involving a single vehicle crash or pedestrians are the most fatal accidents. This is likely due to a range of factors, some easy to describe and others not. Pedestrian incidents are more simple to justify, as pedestrians have no safety features when walking and coming into contact with a vehicle unlike in other incidents such as a multi-vehicle crash, and therefore consequently would suffer a higher hospitalisation and fatality rate. However, single vehicle crashes are a bit harder to explain, as there is a much broader range of factors that can be considered, such as hitting poles, getting stuck without help, and others that could contribute to a high death and hospitalization rate. Multi-vehicle crashes are statistically the safest of the three (excluding the Other category), and can be justified by modern safety features such as seatbelts, airbags, as well as nearby human support.

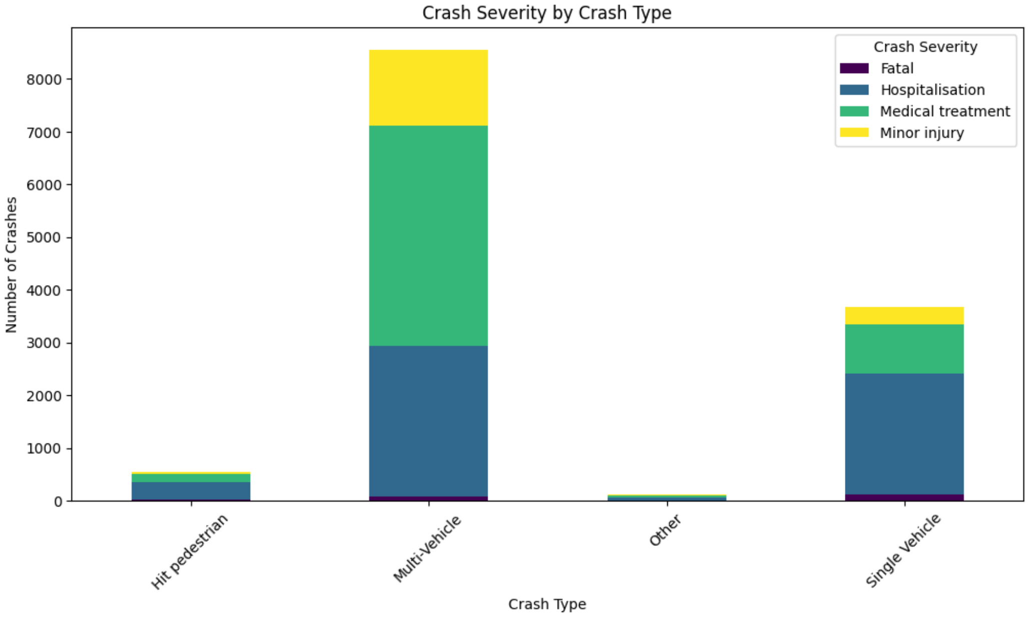

Figure 4: Severity vs Crash Type Stacked Bar Plot

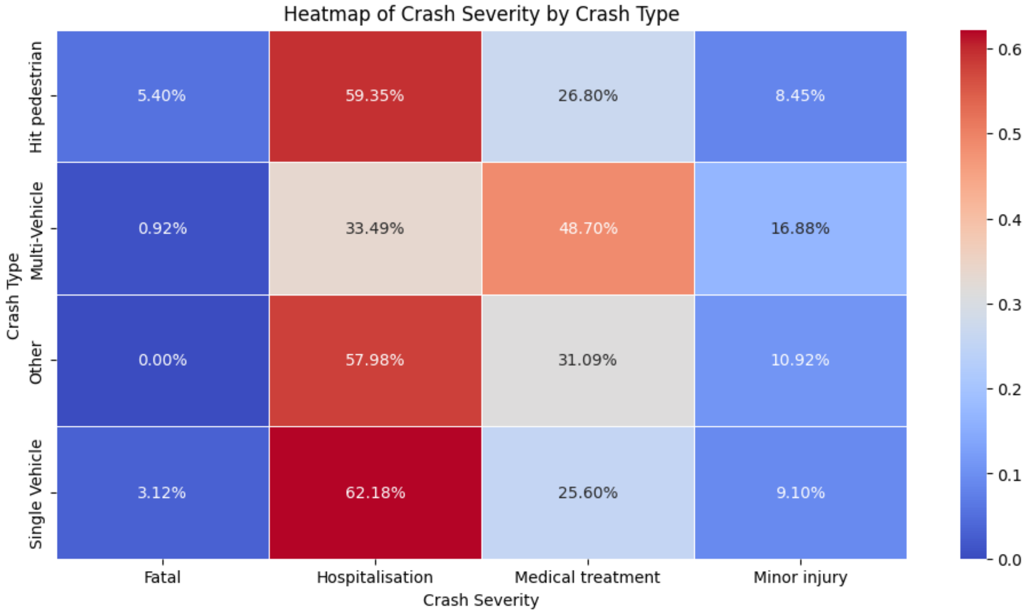

Figure 5: Severity vs Crash Type Stacked Bar Plot

4.4 Crashes vs Severity for each Month of 2018

Another statistic to consider is the distribution of crashes throughout the year to consider factors such as seasonal changes. From Figure 6, immediately, there is a noticeable drop in the number of

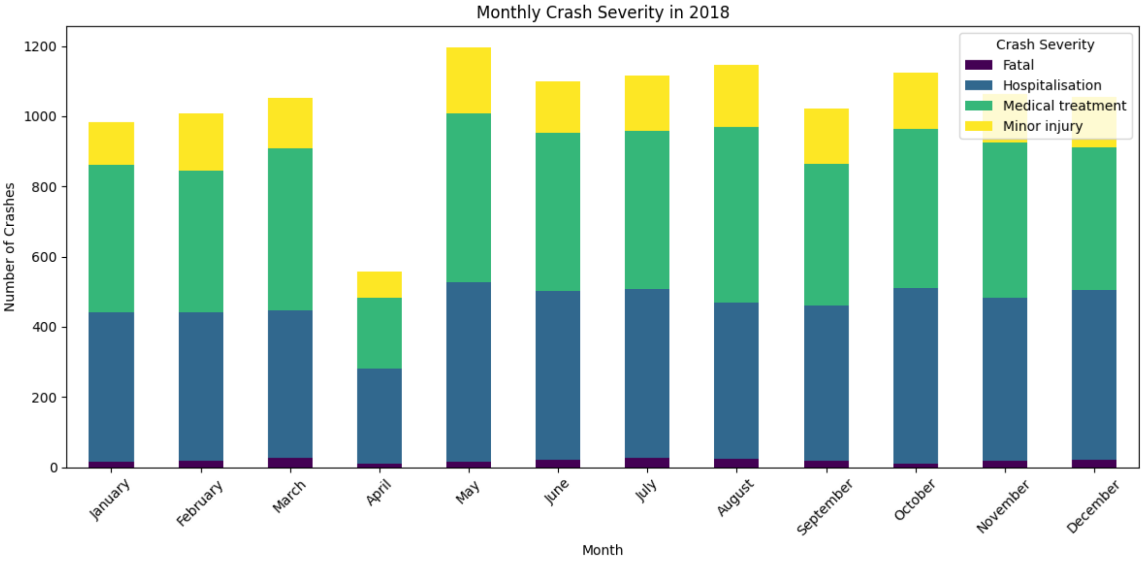

Figure 6: Crash Frequency vs Severity for each Month

crashes in April. Considering the rest of the months, there does not appear to be any evident trend or change in the number and severity of crashes over the months that cannot be explained with random deviations. However, the month of April has significantly reduced total crashes. While this could potentially be explained/justified through reasons, the absence of any evident trend and the sudden drop could be an indication that data has been lost for this month.

4.5 Most Dangerous Suburbs

Another variable to consider is the locations at which the higher frequency of crashes occur. This information can provide insights into which areas need to be considered high priority investigation. This information is most easily presentable on a stacked bar chart in decreasing order of the number of crashes. Considering Figure 7, we immediately see that the most dangerous suburbs are as follows

1. Southport
2. Brisbane City
3. Caboolture
4. Woolloongabba
5. Slacks Creek
6. Fortitude Valley
7. Eight Miles Plains
8. Surfers Paradise
9. Upper Mount Gravatt

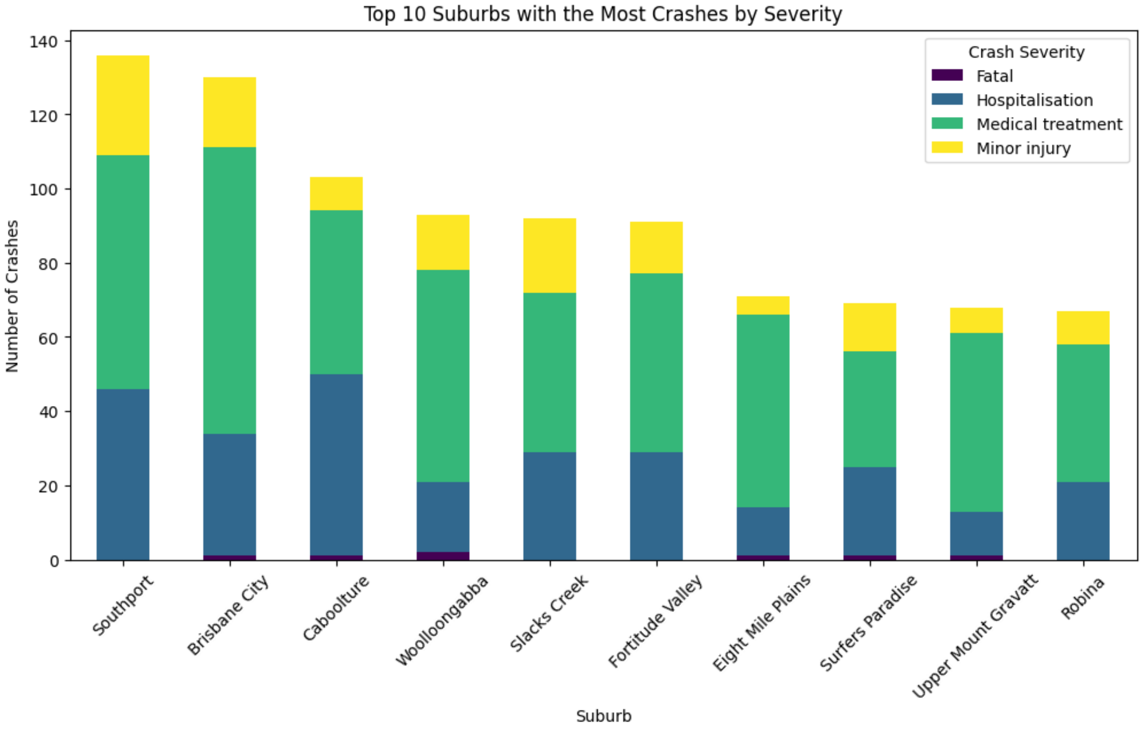

Figure 7: Top 10 Suburbs with the Highest Crash Frequency

10. Robina

There are numerous factors that can lead to these suburbs having the highest frequency of crashes. The most explainable is the volume of traffic flow. Areas such as Brisbane City have much larger traffic flow than areas that are more suburbia away from the city, and would obviously have a larger amount of crashes. Other factors could be considered as well however, such as road design, lane width, driver training etc. Other noticeable features from Figure 7 include the high proportion of fatalities in Woolloongabba, which could be cause for concern regarding these factors that needs to be monitored.

4.6 Geographic Heatmap of Crashes

A very simple way to view the density of crashes in areas is through the use of a geographical heatmap. The Plotly Express library provides a simple and convinent way to create a powerful visualisation for this, in the form of an interactable geographical heatmap of crashes. Figure 8 provides a static image snapshot of this. As evident immediately, the density of crashes increases drastically as the density of housing and consequently traffic flow would increase. However, we can zoom in to specific regions to gain more meaningful insights. Figure 9 provides a more zoomed in analysis of Brisbane City, which was noted earlier as the region with the 2nd highest crash frequency. From this information, we can now more easily identify specific regions of Brisbane City that have a high frequency of crashes, and investigate features such as road design that could be contributing

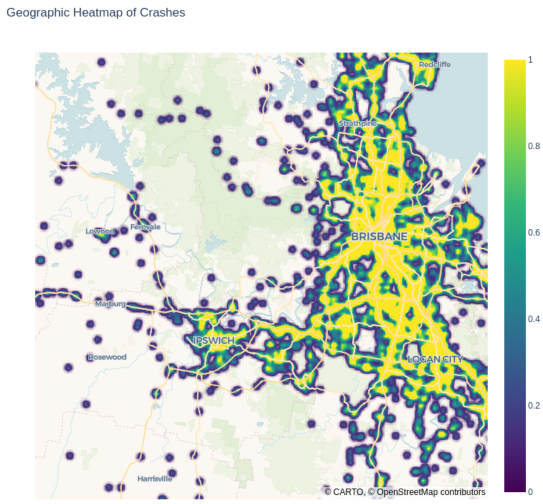

Figure 8: Geographical Heatmap Image 1

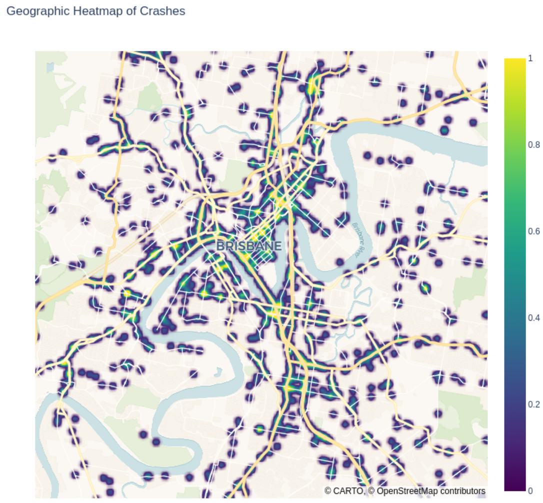
Geographic Heatmap of Crashes
BRISBANE
© CARTO, © OpenStreetMap contributors
1
0.8
0.6
0.4
0.2
0

Figure 9: Geographical Heatmap Image 2

to a larger frequency of crashes. Notably, Central Brisbane appears to have the highest frequency of crashes, due to factors such as higher traffic flow, lane width, etc.

5 Conclusion

The analysis of crash data has provided key insights into factors influencing crash severity and frequency. The results confirm that higher-speed zones correlate with increased crash severity, with fatal crashes being most prevalent in 100–110 km/h zones. Single-vehicle and pedestrian crashes emerged as the most hazardous crash types, likely due to the absence of protective measures compared to multi-vehicle collisions. Monthly trends did not indicate significant variations, except for a notable drop in April, which may indicate missing or incomplete data. Geographical heat map analysis identified high-crash areas, particularly within dense urban regions, suggesting a need for targeted safety measures such as improved road infrastructure, speed regulations, and pedestrian safety enhancements. Future research should explore additional variables such as driver behavior and environmental conditions to further refine risk assessments and mitigation strategies.