Group 12 Exploratory Data Analysis

DATA301

Wian Lusse (300489294), Nicholas Gibbs (300579601), Satrio Wiradika (300578313)

5 September 2023

Contents

Background and Data

Our project specifically focused on working with the Crash Analysis System (CAS) data sourced from Waka Kotahi, the New Zealand Transport Agency. We accessed the CAS data through the Open Data platform provided by Waka Kotahi offering a comprehensive dataset that is updated monthly containing detailed data about all traffic crashes reported by the NZ Police to Waka Kotahi. This data covers crashes on all New Zealand roadways or places where the public has legal access to a motor vehicle that has occurred across New Zealand since the 1st of January 2000. The dataset only includes crash variables that are non-personal data.

The CAS dataset is of great interest as it can be a resource to uncover insights into the domain of road safety. Specifically in our world where transportation plays a pivotal role in connecting societies, economies, and individuals, being able to comprehend the underlying dynamic of traffic crashes can be really informative. By using the CAS dataset, we are able to analyse the data from traffic crashes offering a different perspective on road safety, crash patterns, and the relationship between these traffic crashes and the geographic factors in New Zealand. At its core, the CAS data is much more than a repository of data, but a reflection of real-world events. Every entry within the dataset encapsulates a moment of impact, actions and consequences, and the complexity of vehicle interactions. By examining this collection of data, researchers will have the possibility to potentially unveil patterns within traffic crashes, and identify the factors that are contributing to their occurrences, to then create strategies to mitigate their impact. Specifically within the CAS dataset, its comprehensive nature is not being limited to a single dimension of analysis. Instead, it offers a vast range of variables that can be analysed. These variables span a wide range from the obvious elements such as weather conditions, vehicle types, and speed limit, to the more nuanced factors such as crash severity, and the region in which the traffic crash occurred. These wide ranges of variables can help researchers to inform and design targeted measures. For instance, if the data reveals a connection between a road with a higher incidence of accidents during certain weather conditions, additional efforts can be put in place to improve signage or alter the speed limits. Even for researchers who are unfamiliar with the dataset itself, they can answer inquiries such as 'What are the primary factors contributing to road crashes in New Zealand?' or 'How do crash rates differ across distinct geographical regions and evolve over time?'

The CAS data has numerical and categorical data types, structured with 70 columns and 821744 rows. Numerical attributes consist of various quantitative measures, including the unique identifier of an area unit, the number of lanes on the crash road, the unique identifier of a mesh block, and the speed limit in force at the crash site at the time of the crash may be a number, or LSZ for a limited speed zone. As to the categorical data, here are the features that we will use to answer our research question:

- 'crashSeverity' which represents The severity of a crash with possible values of 'F' (fatal), 'S' (serious), 'M' (minor), 'N' (non-injury) that are determined by the worst injury sustained in the crash at time of entry.
- 'weatherA' that indicates the weather at the crash time/place which includes 'Fine', 'Mist', 'Light Rain', 'Heavy Rain', 'Snow', and 'Unknown'.
- 'weatherB' which tells us additional information on the weather such as 'Frost', 'Strong Wind' or 'Unknown'.
- 'region' which identifies the local government (LG) region. The boundaries match territorial local authority (TLA) boundaries.

Table 1: Count of Unique Values on Each Category

crashSeverity	Count
Fatal Crash	7589
Minor Crash	191336
Non-Injury Crash	575954
Serious Crash	46865

weatherA	Count
Fine	635621
Hail or Sleet	132
Heavy rain	33153
Light rain	124210
Mist or Fog	11306
Snow	1544
Unknown	15778

weatherB	Count
Frost	9254
Strong wind	14389
Unknown	798101

Table 2: Missing and region data summary

	Count	Percentage
crashSeverity	0	0.0000000
weatherA	0	0.0000000
weatherB	0	0.0000000
region	3188	0.0038796

region	Count
Auckland Region	285346
Bay of Plenty Region	47177
Canterbury Region	82146
Gisborne Region	9784
Hawke's Bay Region	32388
Manawatū-Whanganui Region	46329
Marlborough Region	8266
Nelson Region	8076
Northland Region	33299
Otago Region	44574
Southland Region	20234
Taranaki Region	18604
Tasman Region	7541
Waikato Region	87849
Wellington Region	79725
West Coast Region	7218

We found a varying degree of missing information across the variables. While some variables exhibit high degree of completeness, others display more inconsistencies in the data. This does not necessarily imply a structured pattern, such as missing data clustered in specific time periods or geographic regions, but rather the lack of information that can be collected about the specific traffic crash. When specifically looking at the variables that we want to analyse, we can see that there is no missing data for crash severity, weatherA, and weatherB.

However, for the region variable, there is a high level of missing data with 3188 observations missing. This can be translated to a percentage of missing data of 0.0038796. In terms of errors within the specific variables that we wanted to analyse, there does not seem to be any errors. The observations of the different categories are consistent and there are no observations for categories in which they appear to be errors.

Ethics, Privacy and Security

Ethical considerations hold an important role in our project as we delve into analysing the CAS dataset. While the dataset can be an extraordinary resource in helping analyse variables and generate insights, it presents us with ethical considerations that we must address. Given the sensitive nature of crash data, it is crucial that we exercise discretion in sharing insights that could inadvertently identify individuals involved in accidents. Finding a balance between informative reporting and preserving privacy is essential. Additionally, presenting the data in a way that avoids overemphasising accidents while focusing on generating useful insights is ethically crucial. Acknowledging the broader implications of our possible findings, particularly when suggesting insights, can ensure that our insights contribute positively to aspects such as road safety without causing much distress to those affected by road accidents. Another ethical responsibility lies in the manner in which we choose to convey the data to our audience. Committing to ethical practices ensures that we refrain from any form of sensationalism that can arise from the presentation of accident-related information. Instead, our approach centers on accentuating the importance for road safety improvements. This demands that the design of our communication is focused on the goal of mitigating accidents. We need to approach the task of suggesting insights with a heightened sense of responsibility. Our acknowledgement of the potential impact that our insights could have on road safety measures ensures a positive change without causing unwarranted distress to those who have been directly or indirectly affected by traffic crashes. Our project can be seen from an aspect of social responsibility, where every action we make in our analysis is aligned with the overarching goal of road safety and societal well-being.

Privacy concerns are also a consideration when dealing with the CAS dataset. Even though there is no personal information within the dataset, privacy concerns can still exist. Even if de-identified, there is still potential for the data to be linked back to the specific individual or disclose details into their personal life. The nature of the CAS dataset necessitates a comprehensive approach to ensure the mitigation of privacy risks. Anonymisation techniques such as data aggregation, suppression, and pseudonymisation, might have been employed to safeguard the identities of those mentioned in the dataset. However, the effectiveness of these techniques is not always absolute, and the complicated connections between the variables in the dataset could potentially allow for the re-identification of individuals, compromising their privacy. Furthermore, the potential for unintended consequences emerges when privacy concerns are not thoroughly addressed. The insights drawn from our analysis might inadvertently perpetuate biases. Ensuring that care is taken to ensure that the analytical process respects the privacy of the individuals involved in the dataset is a must. In light of these considerations, it is crucial to adopt a comprehensive privacy framework that encompasses the technical safeguards but also the ethical guidelines. Collaboration with ethics could help in identifying potential pitfalls, and the transparency in the methods that are used for data analysis even if indirectly are steps in ensuring privacy through the information encapsulated within the dataset.

To maintain the security of our project data and results, several steps can be taken, even though they have not been implemented for the purpose of this report. One of the foremost strategies for safeguarding sensitive information is the implementation of data encryption protocols. By encrypting data both during storage and transmission, we can prevent potential breaches and unauthorised access attempts. Encryption converts the data into an unreadable format. Without the appropriate decryption key, it is extremely challenging for malicious actors to interpret the information even if they manage to gain access. Additionally, controlling and restricting access to the data repository is crucial. This can be achieved through the deployment of secure authentication and authorisation mechanisms. By implementing strong authentication methods such as multi-factor authentication and ensuring that only authorised personnel possess the necessary credentials, we can effectively prevent unauthorised entry to the data and results. Authorisation mechanisms can be employed to establish varying levels of access privileges based on roles within the project team, thereby minimizing the risk of inadvertent data exposure. In the event that an incident occurs such as data corruption, hardware failures, or cyberattacks, regular data backups emerge as another critical line of defence. Regularly backing up project data is a reliable means of preventing data loss and facilitating swift recovery. However, it is essential to store these backups in a secure manner. Adopting a best practice approach involves storing backups in off-site locations that are well-protected and disconnected from the primary network to safeguard against potential simultaneous compromise of both primary and backup data. Furthermore, ensuring that the secure collaboration among the project stakeholders is vital. Securing the data-sharing methods plays another crucial role in maintaining data confidentiality while enabling efficient teamwork. Secure file-sharing platforms equipped with end-to-end encryption and access controls provide a secure environment for sharing sensitive documents and information. Virtual private networks can also be employed to establish encrypted communication channels, particularly useful when collaborating with remote team members or external partners.

Exploratory Data Analysis

Exploring the Difference of Weather Conditions Across All the Regions in New Zealand

Adverse weather conditions play a crucial role in shaping the dynamics of road safety, significantly amplifying the potential for accidents to escalate into fatal or critical events. These environmental factors introduce a complex set of challenges that drivers must navigate through, demanding heightened vigilance, skill, and adaptability. The impact of severe weather encompassing heavy rainfall, snow, and fog, extends beyond just the inconvenience; it directly interacts with the systems that govern road safety, creating an environment that demands caution from vehicle users.

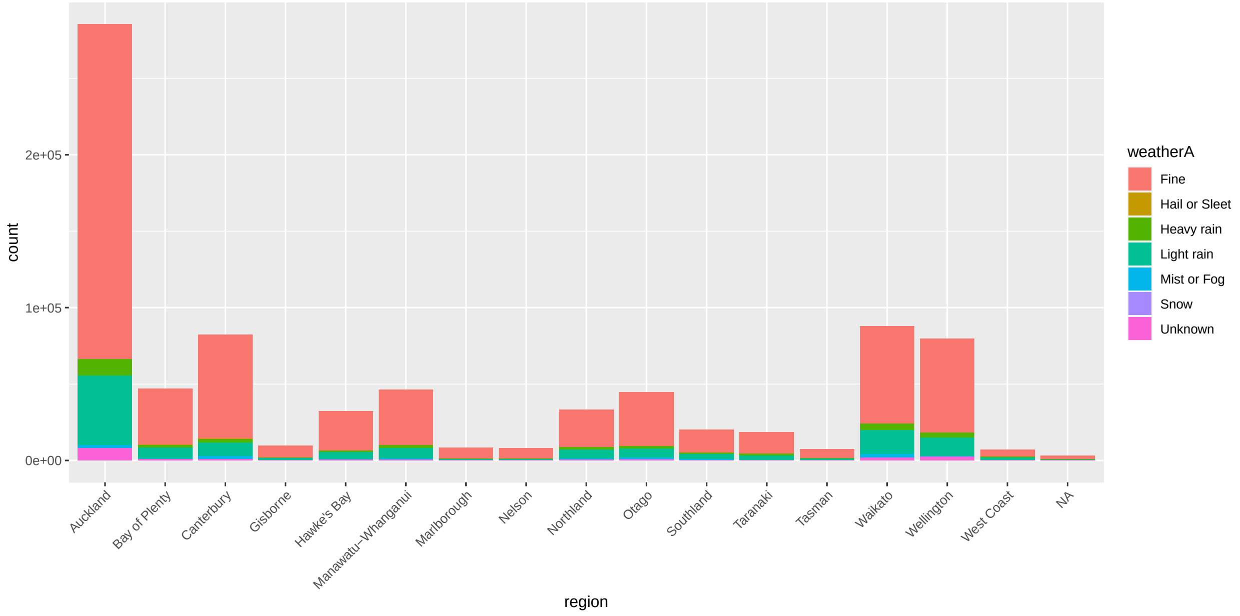


Figure 1: The Distribution of Weather Conditions in the Regions of New Zealand

Insert the key points of the plot above

Is There Any Difference on the Level of Each Crash Severity Between the Regions?

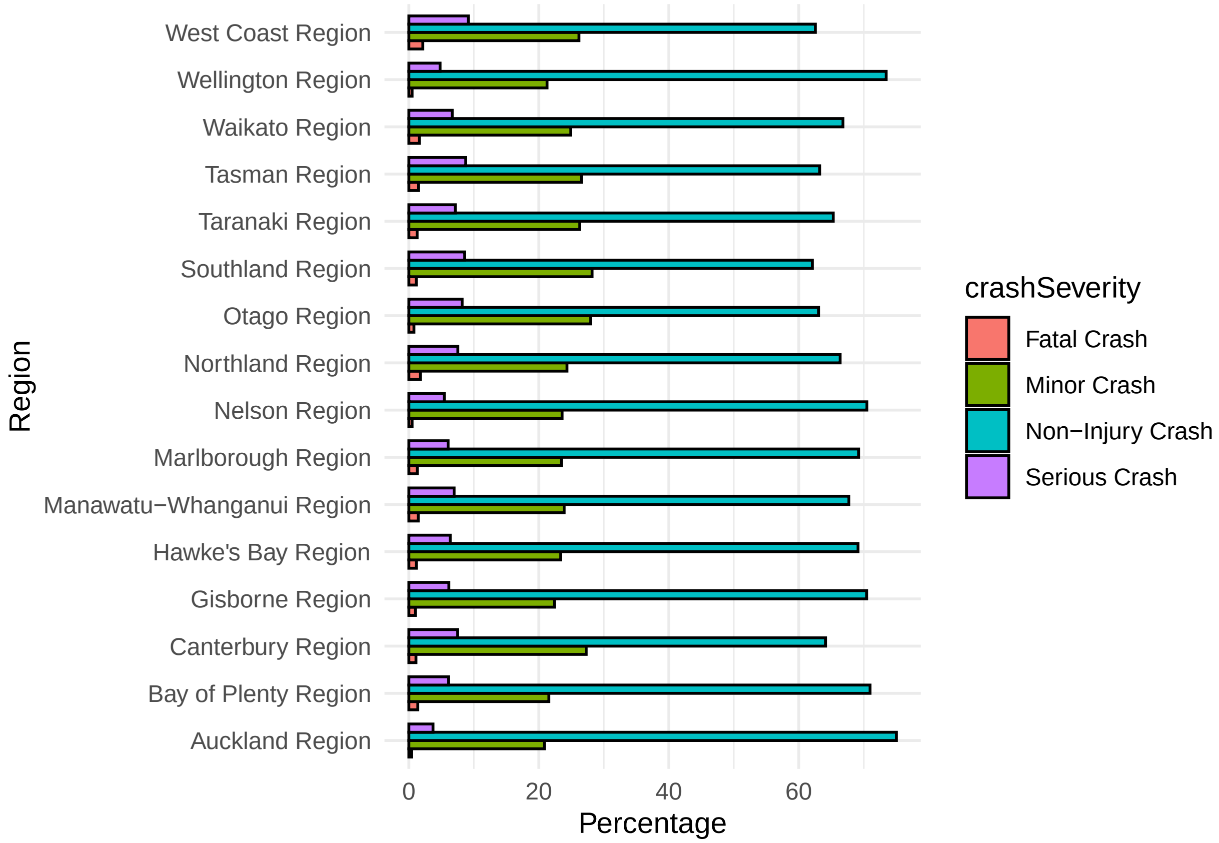


Figure 2: Percentage Grouped Bar Chart of Crash Severity by Region

Exploring the effect of wind on the severity of vehicle crashes

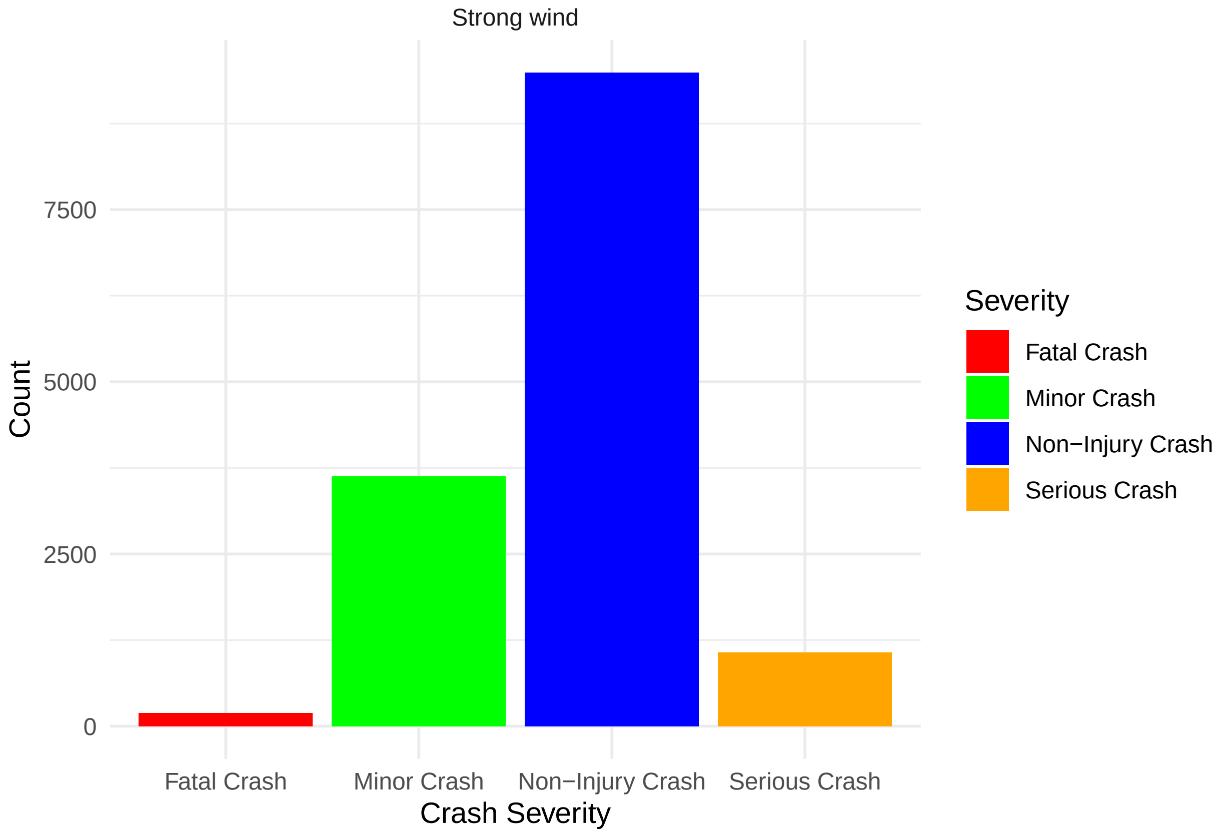


Figure 3: Crash Severity Across All Regions by Strong Winds

Strong winds can play a role in increasing the risk of accidents and potentially be a cause of fatal or serious vehicle crashes. Strong winds can pose multiple hazards on the road. High winds can lead to sudden obstacles on the road and can also turn loose debris like rock into projectiles. High winds can also reduce vehicle control making it harder to stay in lanes safely or even be able to push large vehicles off the side of the road. According to this study in 2019 (Bhattachan et al. 2019), the fatality rate related to winds is almost twice as high as the rate in accident caused my weather conditions other than winds. This might suggest that wind-related accidents tend to be more lethal compared to accidents caused my other type of weather.

Table 3: Counts of Crash Severity Levels for Strong Wind Cases

crashSeverity	count
Non-Injury Crash	9489
Minor Crash	3632
Serious Crash	1074
Fatal Crash	194

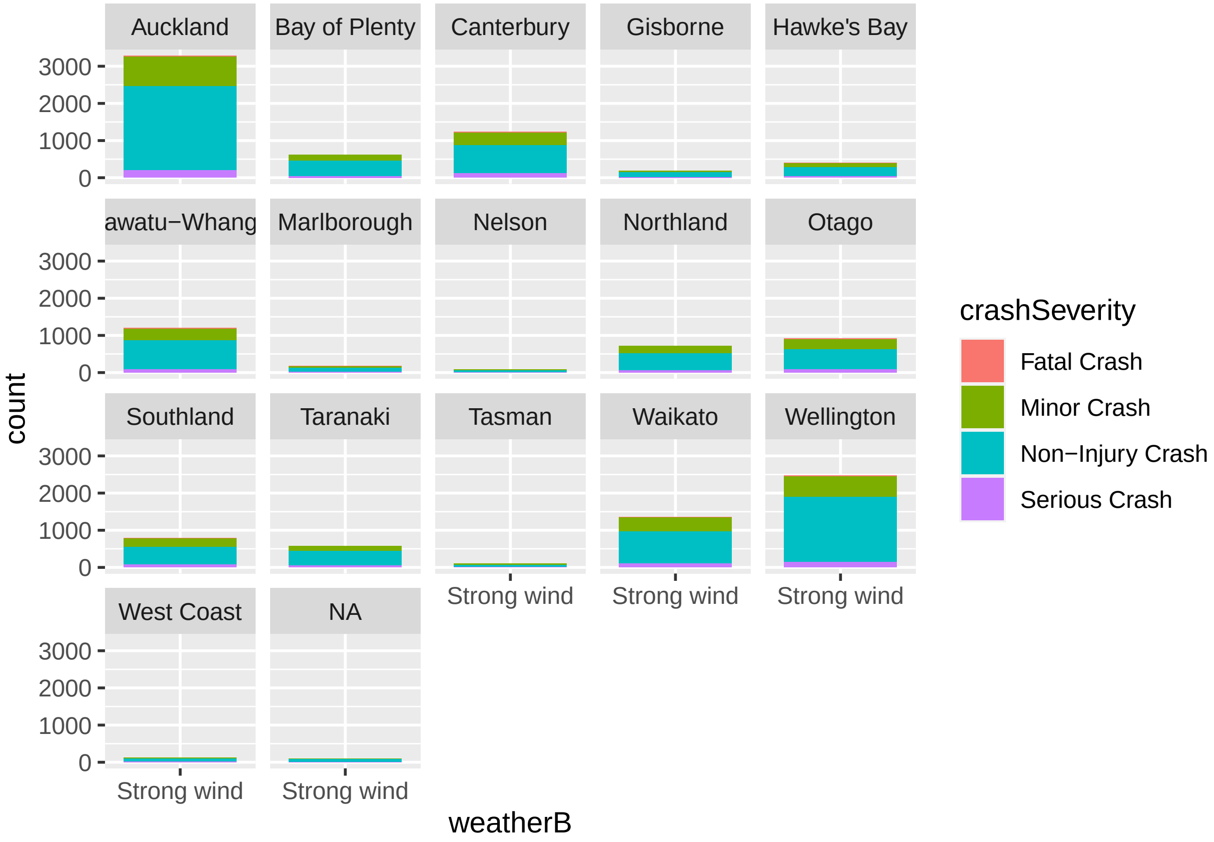
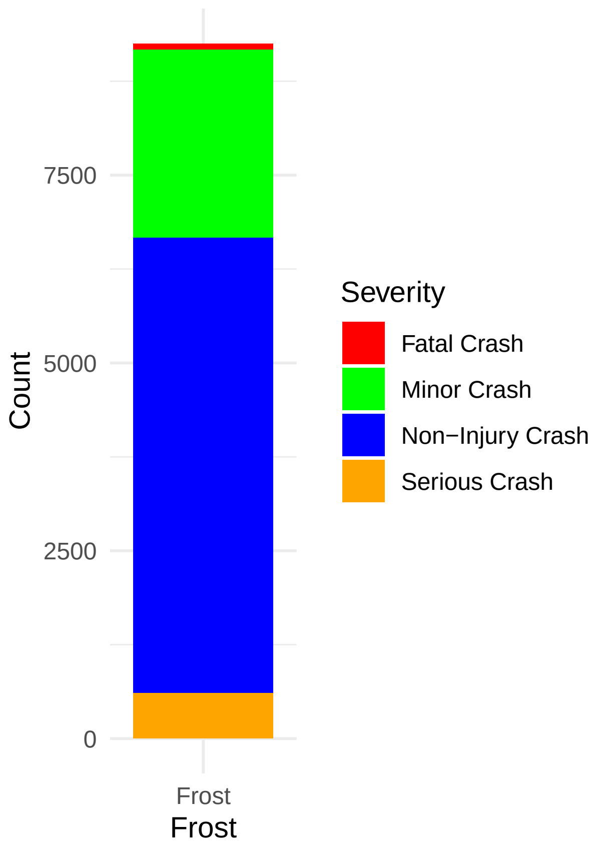
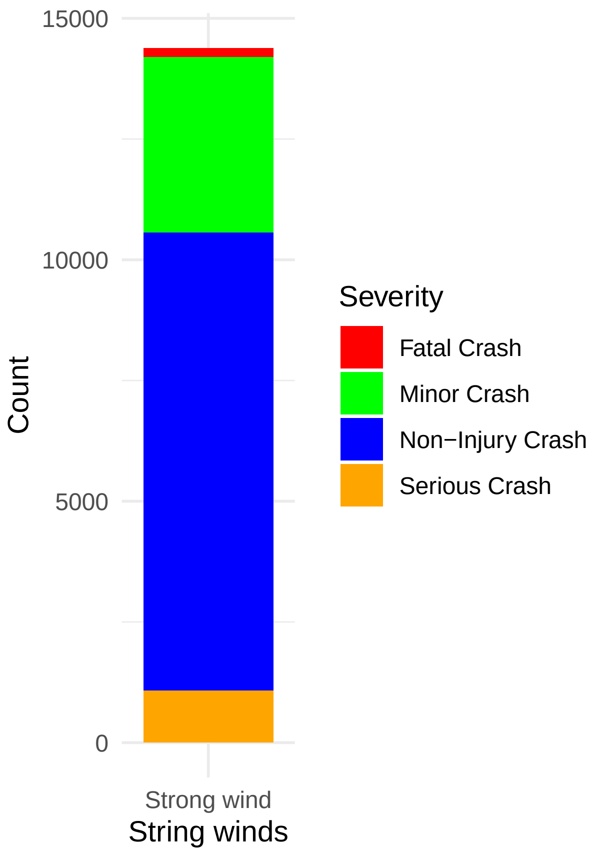


Figure 4: Crash Severity in Each Region for Strong Wind Condition

Comparing light rain to strong winds





Light rain is present in 5 times more fatal accidents than where strong winds were present.

Individual Contributions

Wian Lusse (300489294)

Nicholas Gibbs (300579601)

Satrio Wiradikas

References

Bhattachan, Abinash, Gregory S Okin, Junzhe Zhang, Solomon Vimal, and Dennis P Lettenmaier. 2019. “Characterizing the Role of Wind and Dust in Traffic Accidents in California.” *GeoHealth* 3 (10): 328–36.