Group 12 Exploratory Data Analysis

DATA301

Wian Lusse (300489294), Nicholas Gibbs (300579601), Satrio Wiradikas

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 that has the capability 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 of great interest. By using the CAS dataset we are able to analyze the data from traffic crashes offering a unique view from which we are able to generate insights towards road safety, crash patterns, and the relationship between these traffic crashes and the geographic factors specifically within the context of New Zealand. At its core the CAS data is much more than a repository of data, but a reflection of real-world events that hold the potential to generate valuable insights into traffic crashes. Every entry within the dataset encapsulates a moment of impact, actions and consequences, and the complexity of vehicular interactions. Examining this collection of data researchers 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 analyzed. 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. This extensive range of variables provides a greater opportunity for analysis and enables researchers to reveal connections that might otherwise remain concealed. Furthermore, the relevance of the CAS data and its ability to generate insights have the potential to bring tangible changes in road safety and interventions, evidence drawn researchers can help inform and design targeted measures to address the evidence. For instance, if the data reveals a connection between a road with a higher incidence of accidents during certain weather conditions, can be additional efforts put in place to improve signage, alter speed limits, and enhance road surrogacy in the specific area. Even for readers unfamiliar with the dataset itself, researchers can investigate inquiries such as: What are the primary factors contributing to road crashes in New Zealand? How do crash rates differ across distinct geographical regions and evolve over time? Are there discernible patterns connecting specific vehicle types or road conditions with heightened crash occurrence? In the context of our project, we wanted to investigate the question “What is the correlation between weather conditions and crash severity of road accidents in different regions in New Zealand?”.

The CAS data is characterized by its numerical and categorical data types, structured with 70 columns and 821744 rows. Numerical attributes encompass 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 the CAS dataset features descriptors such as the severity of a crash with possible values are ‘F’ (fatal), ‘S’ (serious), ‘M’ (minor), ‘N’ (non-injury), determined by the worst injury sustained in the crash at time of entry, the direction of the principal vehicle involved in the crash with possible values are North, South, East or West, whether the road is flat or sloped with possible values include ‘Flat or 'Hill’, and the light at the time and place of the crash with possible values: ‘Bright Sun’, ‘Overcast’, ‘Twilight, 'Dark’ or ' Unknown’. Furthermore, the geographic aspect is represented through the variables of part one of the ‘crash location’, which may be a road name, route position (RP), landmark, or other, e.g. ‘Ninety Mile Beach’ used for location descriptions in reports etc, part 2 of the ‘crash location’, maybe a side road name, landmark etc, used for location descriptions in reports etc, as well as identifies the local government (LG) region the boundaries match territorial local authority (TLA) boundaries. The data types that we have specifically chosen to answer our research question include ‘crash severity’, ‘weatherA’, ‘weatherB’, and ‘region’ which are all categorical variables.

Table 1: Types of data summary

	Type
crashSeverity	character
weatherA	character
weatherB	character
region	character

The completeness of the CAS dataset as a whole we discern a variable degree of missing information across the dataset's variables. While some variables exhibit high completeness, others display more pronounced gaps in data. This variance 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 analyze we can see that there is no missing data for crash severity, weatherA, and weatherB. However, for the region variable, there is a level of missing data with 3188 observations missing, which can be translated to a percentage of missing data of 0.0038796. In terms of errors within the specific variables that we wanted to analyze there do not seem to be any errors, the counts of the different categories are consistent and there are no counts for categories in which they appear to be errors.

Table 2: Missing data summary

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

Table 3: crashSeverity data summary

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

Table 4: weatherA data summary

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

Table 5: weatherB data summary

weatherB	Count
Frost	9254
Strong wind	14389
Unknown	798101

Table 6: region data summary

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

Ethics, Privacy and Security

Ethical considerations hold a paramount role in our project as we delve into analyzing the CAS dataset. While the dataset can be an extraordinary resource in helping analyze variables and generate insights, it presents us with ethical contemplation that we must address that are integral to the responsible execution of our research. One of the prominent ethical concerns pertains to the potential implications of publicizing the findings of our analysis. Given the sensitive nature of crash data, it is imperative that we exercise discretion in sharing insights that could inadvertently identify individuals involved in accidents. Striking a balance between informative reporting and preserving privacy is essential. Additionally, presenting the data in a way that avoids sensationalizing 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 undue distress to those affected by road accidents. Furthermore, another critical 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 imperative for road safety enhancements. This demands that the design of our communication ensures our message is focused on the overarching goal of mitigating accidents while steering clear of any inadvertent effects of distressing events. Also thinking about the broader consequences our insights can have. In this light, we 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 as an aspect of social responsibility, wherein every action we make in our analysis is aligned with the overarching goal of road safety and societal well-being.

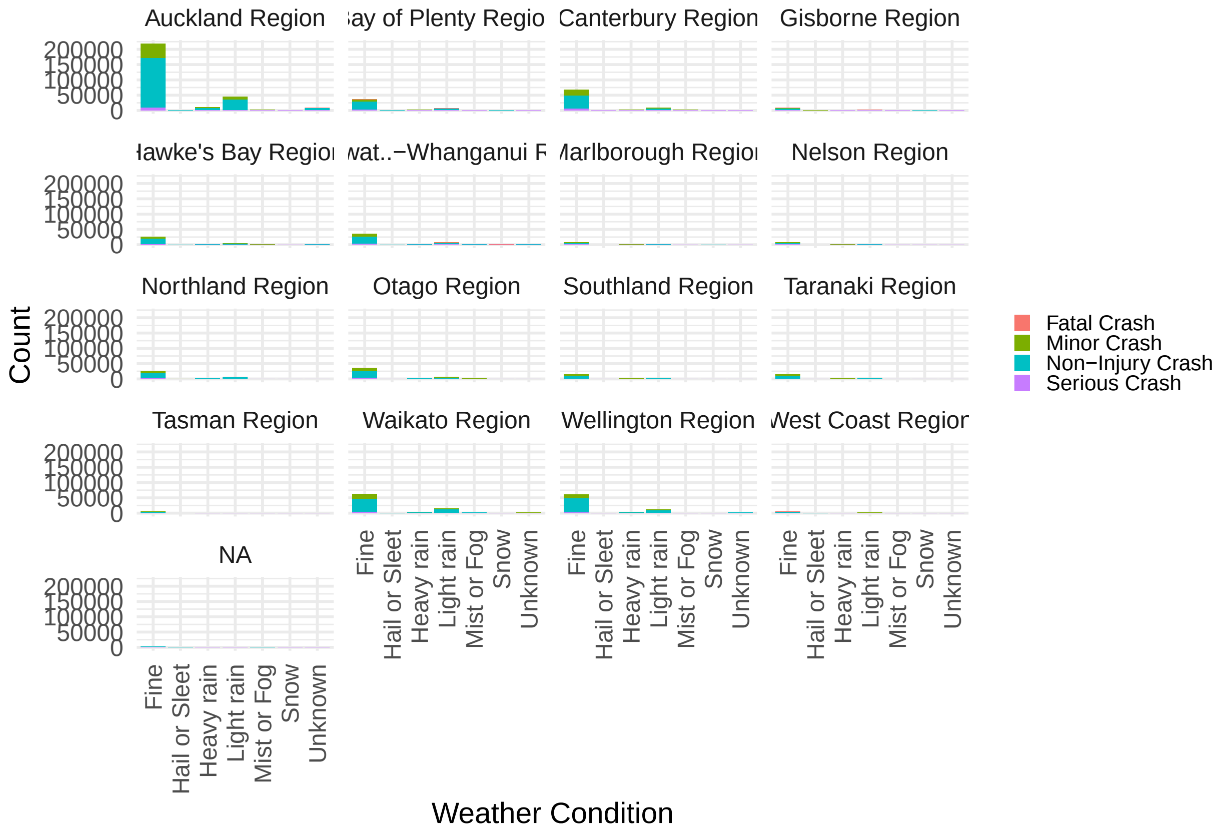
Privacy concerns are also a consideration when dealing with the CAS dataset although there is personal information within the dataset privacy concerns can still be apparent. While the dataset contains a great deal of information throughout its variables for analysis, it cannot be overlooked that there is also the potential for privacy concerns in the fact of such data, even if de-identified there is potential 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 that the mitigation of privacy risks. Anonymization techniques such as data aggregation, suppression, and pseudonymization, 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 intricate interconnections 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 and steering clear of perpetuating biases. 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 aid in identifying potential pitfalls, and the transparency in the methods that are used for data analysis and a commitment even if indirectly are steps in ensuring privacy through the information encapsulated within the dataset.

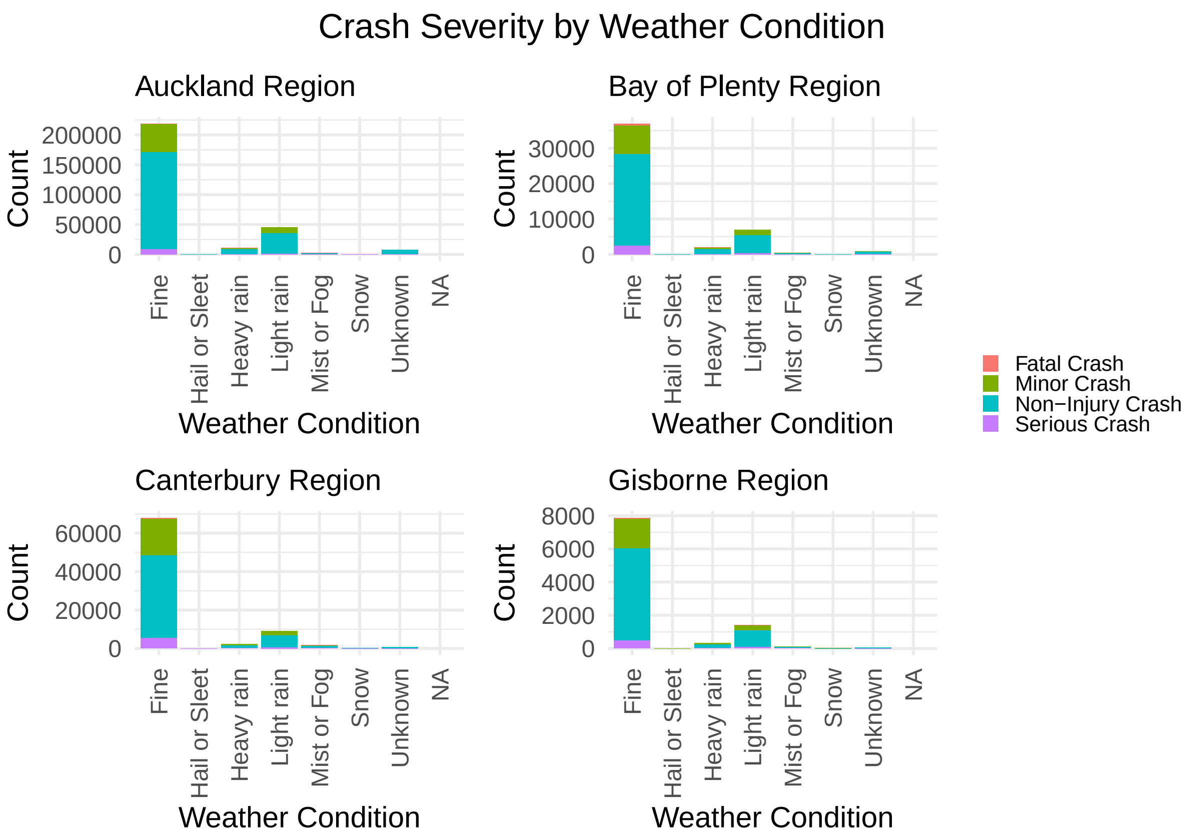
To maintain the security of our project data and results, several steps can be taken, even though they have not been implemented for the purpose of this report. One of the foremost strategies for safeguarding sensitive information is the implementation of data encryption protocols. By encrypting data both during storage and transmission, we can prevent potential breaches and unauthorized access attempts. Encryption converts the data into an unreadable formation without the appropriate decryption key, making it 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 of paramount importance. This can be achieved through the deployment of secure authentication and authorization mechanisms. By implementing strong authentication methods such as multi-factor authentication and ensuring that only authorized personnel possess the necessary credentials, we can effectively prevent unauthorized entry to the data and results. Authorization 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.

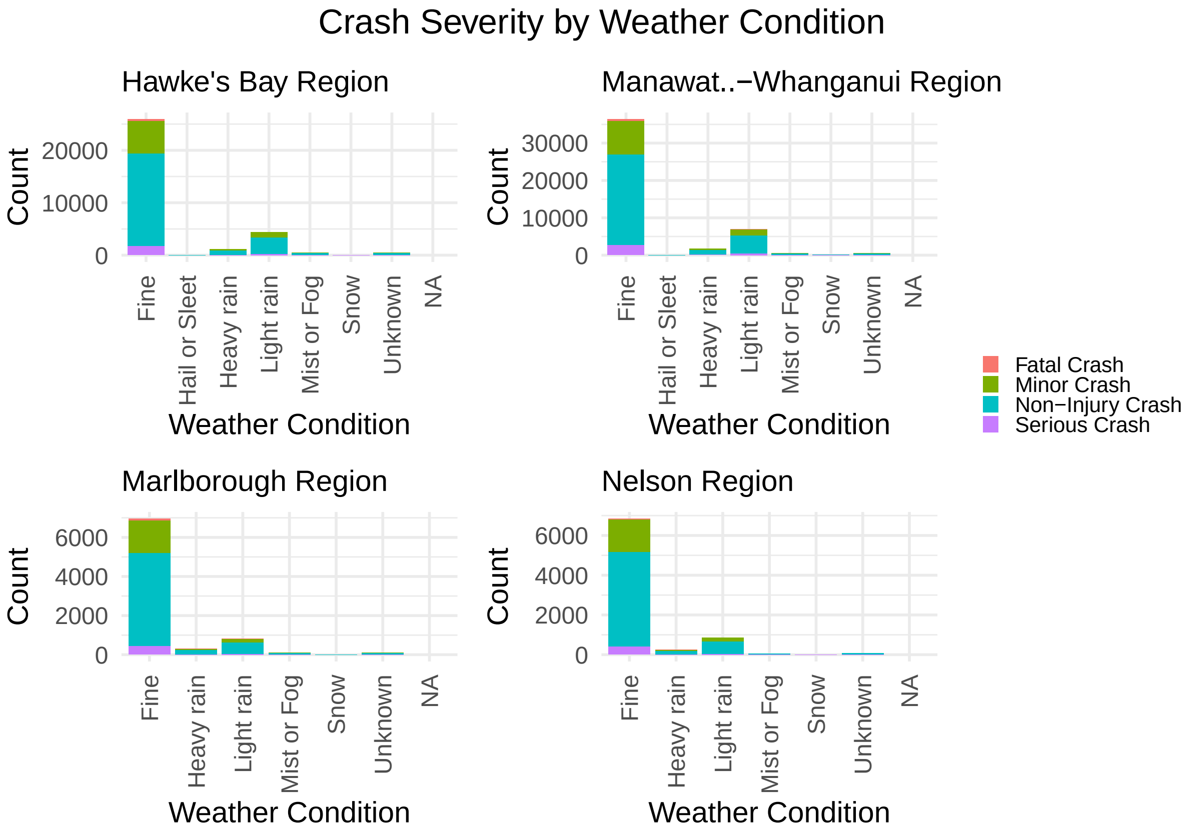
Exploratory Data Analysis

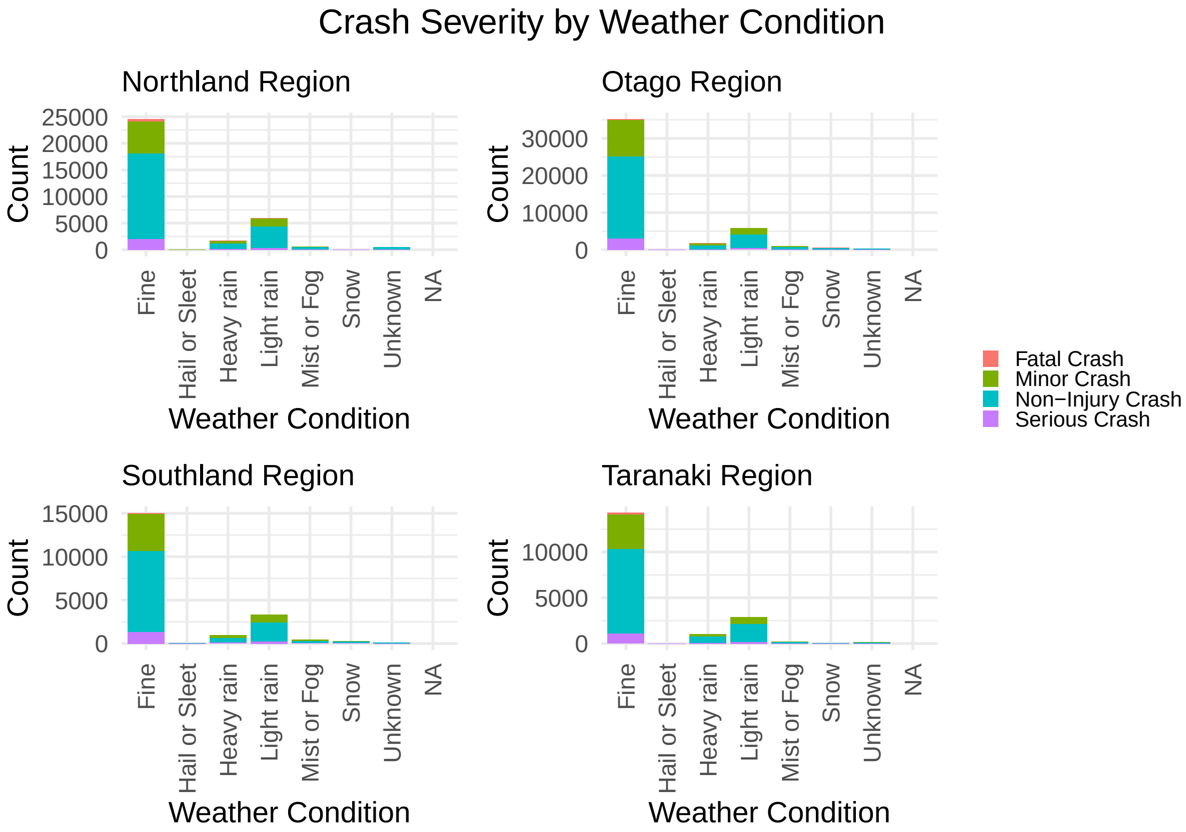
Exploring the effect of weather conditions on the severity of vehicle crashes

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. The relationship between road users and adverse weather begins with heavy rainfall. The raindrops can obscure the visions to an extent where recognizing other vehicles, pedestrians or road signs becomes a difficult task. The rain drops on the windshield refract light, which leads to distorted perceptions of distance and size. This compromised visibility heightens the risk of collisions, particularly due to the reduced reaction time available to drivers when unexpected hazards emerge. Moreover, the road surface undergoes a transformation, transitioning from a stable grip to slippery, the combination of water and residue on the road can significantly diminish the coefficient of friction between the tires and the pavement, leading to an increased stopping distance and a heightened potential for skidding, and loss of control. The entrance of snow into the equation brings a new set of challenges. Snow-covered roads not only impair the driver's ability to see road markings but also introduce an element of uncertainty. The subtle layers of snow can conceal underlying ice patches, rendering roads unpredictably slick. Maneuvering becomes a difficult act, steering or sudden braking can result in a loss of control, causing vehicles to slide uncontrollably. Additionally, the accumulation of snow can narrow roadways, reducing the space and further elevating the risk of unintended contact between vehicles. Fog serves as the disruption of the visibility is dramatically reduced, often extending to a few meters ahead. This drastically impairs the driver's ability to anticipate the road's trajectory and the presence of obstacles in their path. The limited field of vision forces drivers to slow down significantly often well below the speed limit. However, not all drivers might respond to the altered conditions in a timely and appropriate manner, potentially leading to abrupt braking and inadequate braking distances. According to a study conducted in 2019 (Fanny Malin 2019), the relative accident risks are increased for poor road weather conditions; however, they are highest for icy rain and slippery and very slippery road conditions. Especially those stemming from extreme weather, are more likely to result in grave consequences compared to accidents unaffected by adverse weather.

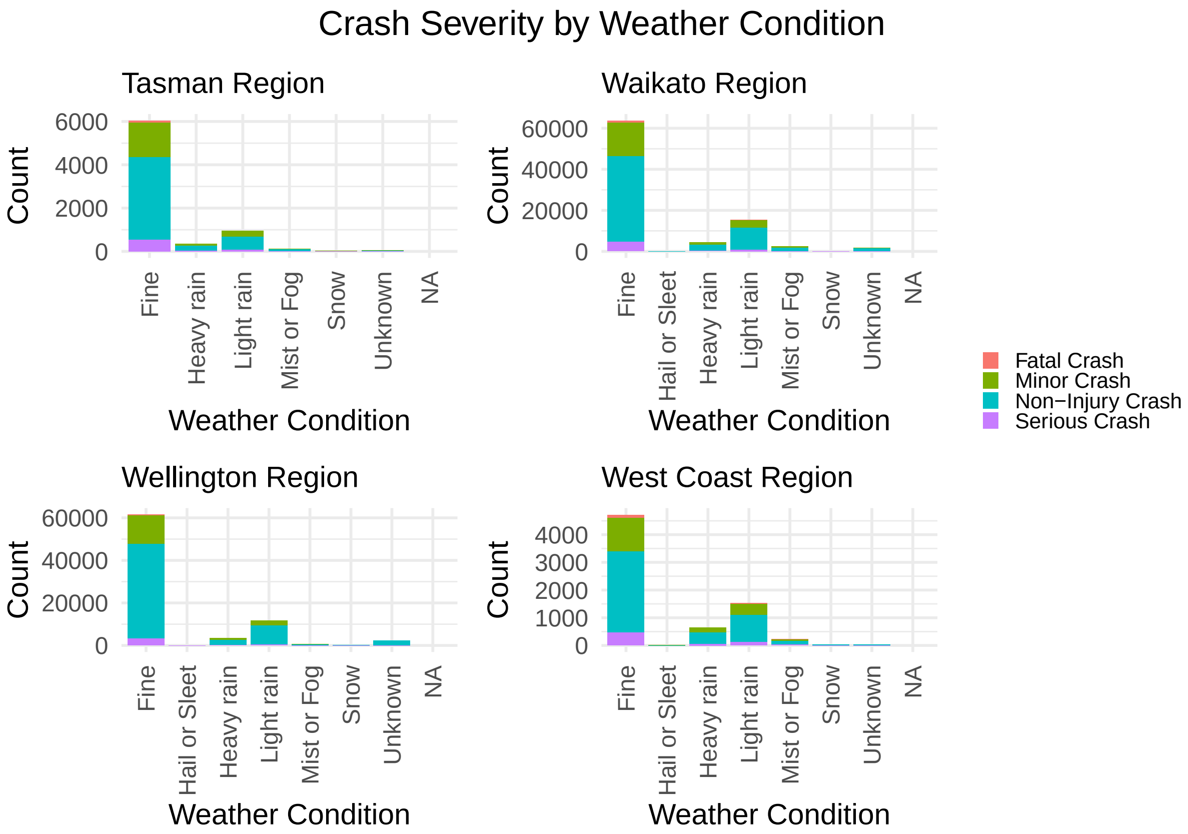



Crash Severity by Weather Condition
Auckland Region
Bay of Plenty Region
Canterbury Region
Gisborne Region
Count
200000
150000
100000
50000
0
30000
20000
10000
60000
40000
20000
8000
6000
4000
2000
Fine
Hail or Sleet
Heavy rain
Light rain
Mist or Fog
Snow
Unknown
NA
Weather Condition
Fatal Crash
Minor Crash
Non−Injury Crash
Serious Crash

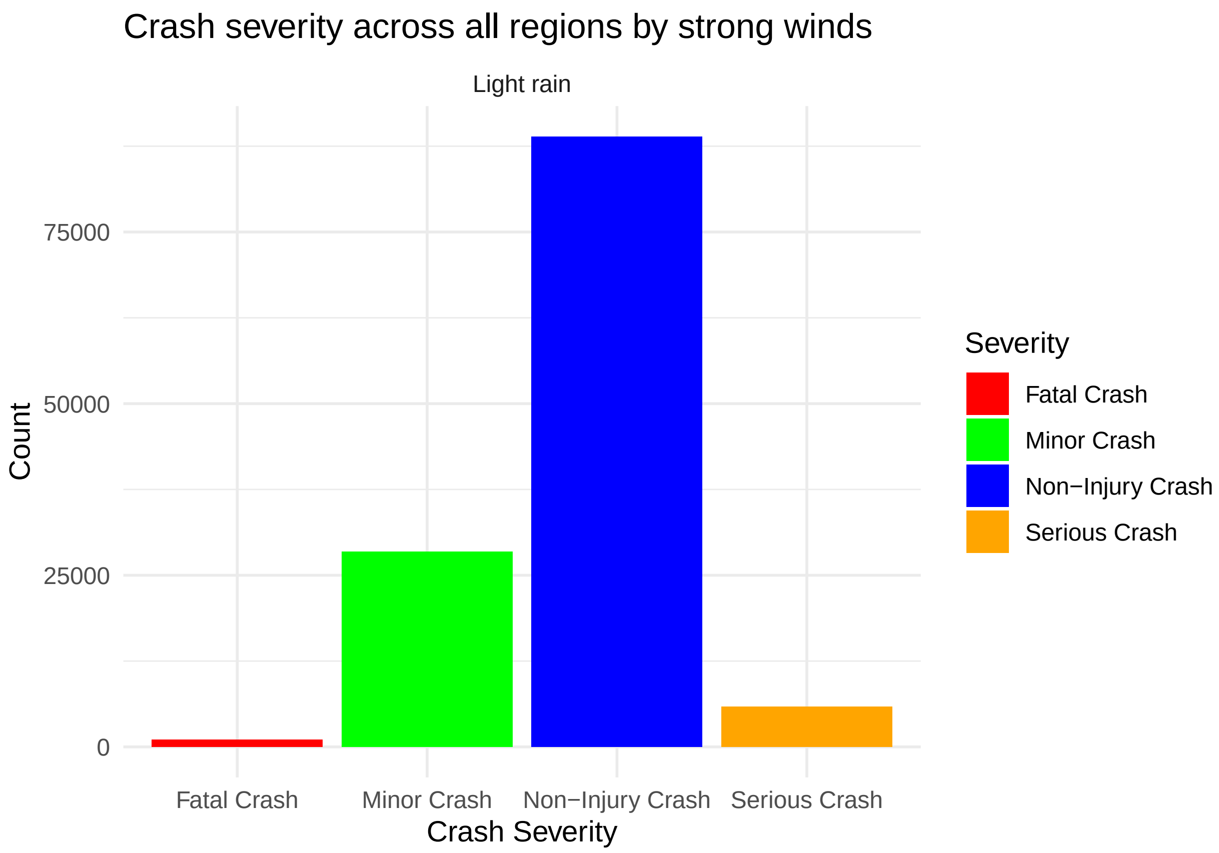

Crash Severity by Weather Condition
Hawke's Bay Region
Count
0
10000
20000
Fine
Hail or Sleet
Heavy rain
Light rain
Mist or Fog
Snow
Unknown
NA
Weather Condition
Manawat..−Whanganui Region
Count
0
10000
20000
30000
Fine
Hail or Sleet
Heavy rain
Light rain
Mist or Fog
Snow
Unknown
NA
Weather Condition
Fatal Crash
Minor Crash
Non−Injury Crash
Serious Crash
Marlborough Region
Count
0
2000
4000
6000
Fine
Heavy rain
Light rain
Mist or Fog
Snow
Unknown
NA
Weather Condition
Nelson Region
Count
0
2000
4000
6000
Fine
Heavy rain
Light rain
Mist or Fog
Snow
Unknown
NA
Weather Condition



Crash Severity by Weather Condition
Northland Region
Otago Region
Southland Region
Taranaki Region
Count
25000
20000
15000
10000
5000
0
30000
20000
10000
0
15000
10000
5000
0
10000
5000
0
Fine
Hail or Sleet
Heavy rain
Light rain
Mist or Fog
Snow
Unknown
NA
Weather Condition
Fatal Crash
Minor Crash
Non−Injury Crash
Serious Crash



The above graphs shows that fine weather is present at most accidents with light rain being present at the second most accidents.



Exploring the effect of wind on the severity of vehicle crashes

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.

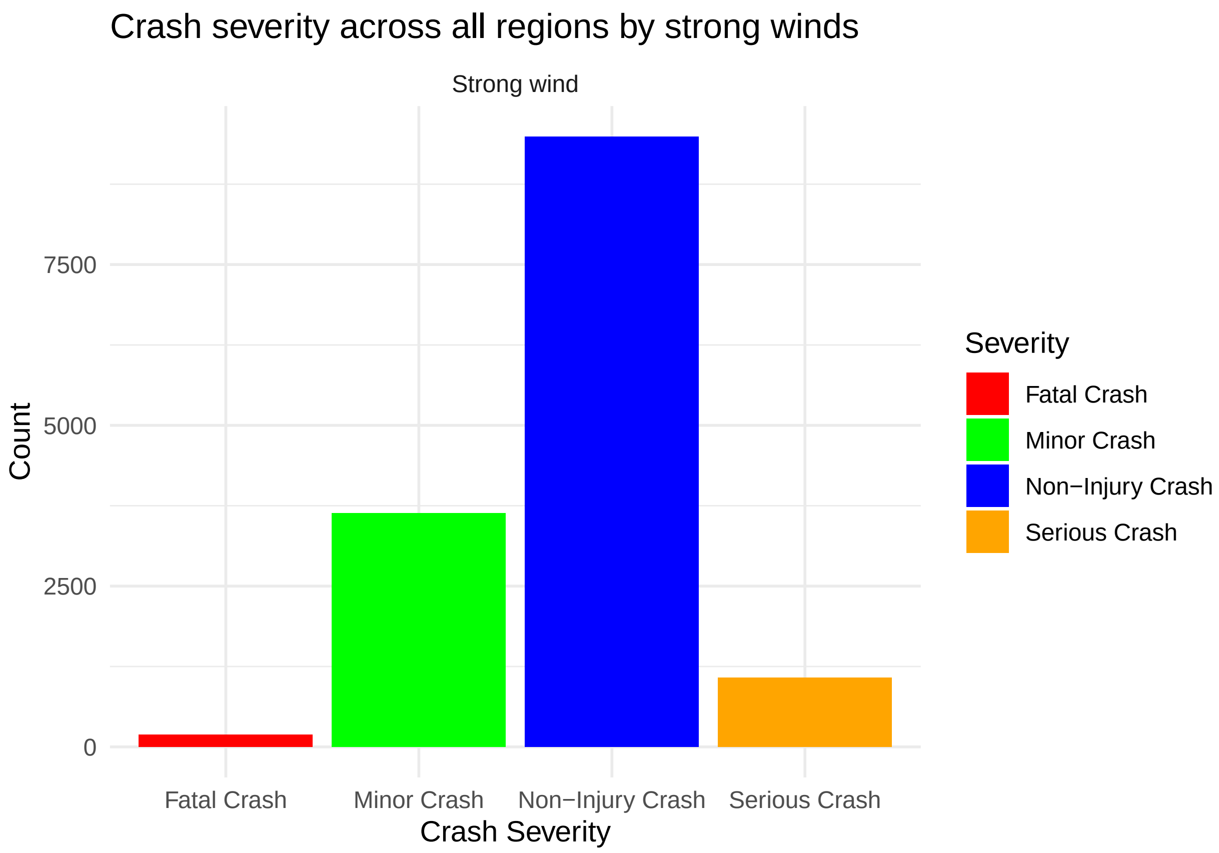
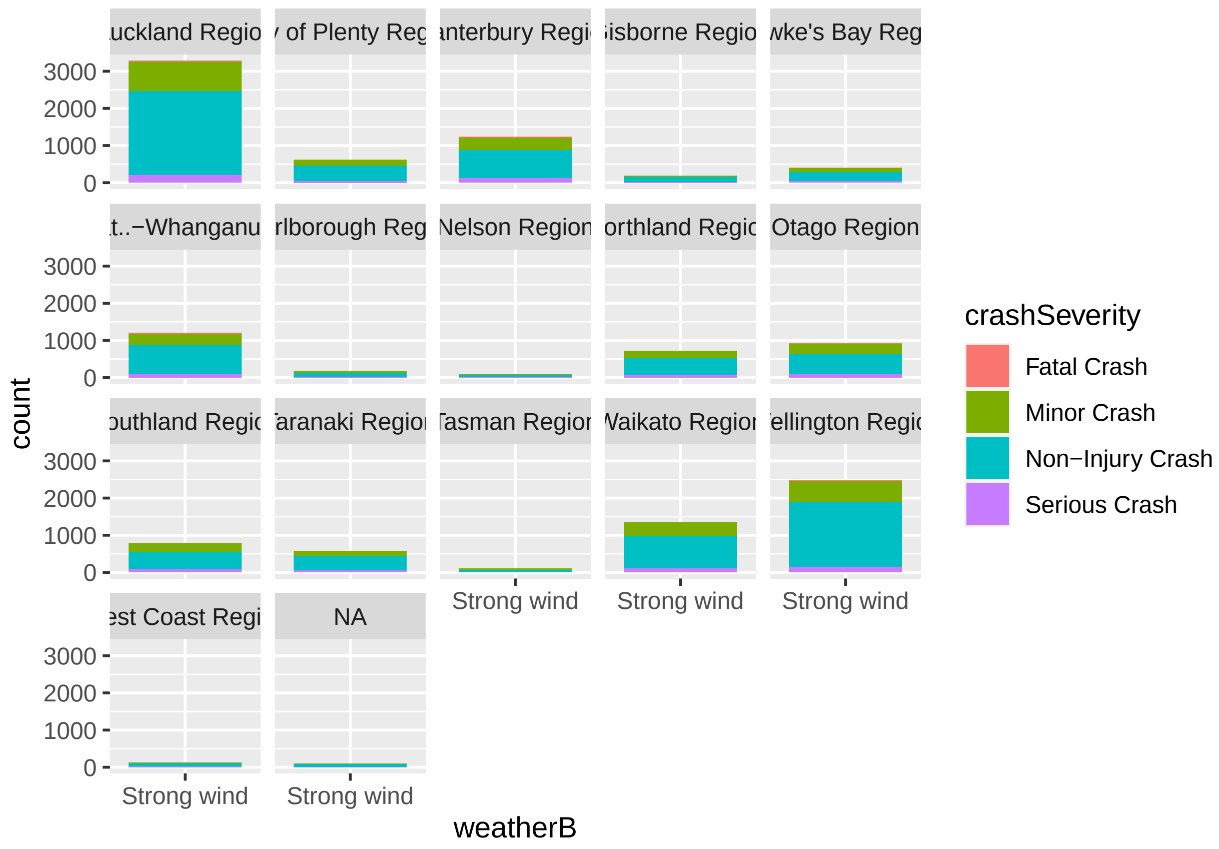


Table 7: Counts of Crash Severity Levels for Strong Wind Cases

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



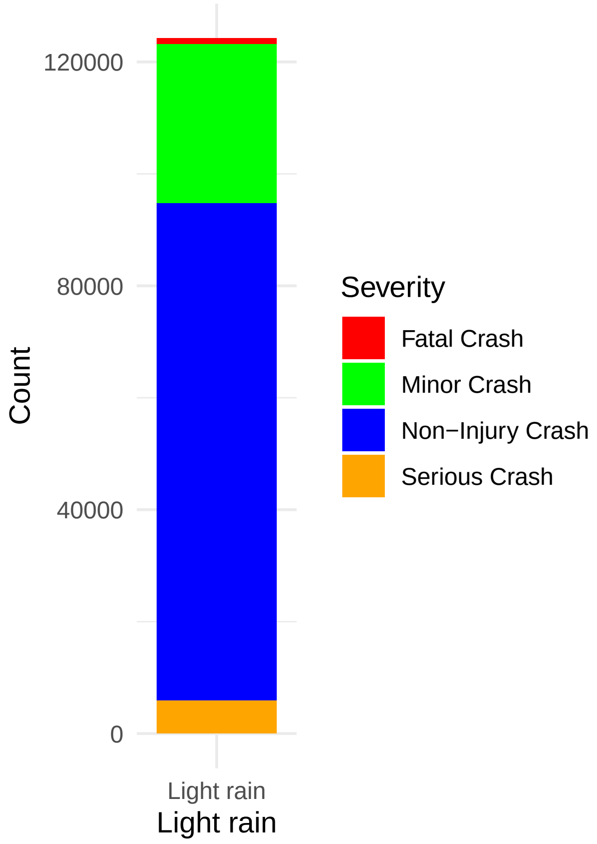
```
## NULL
```

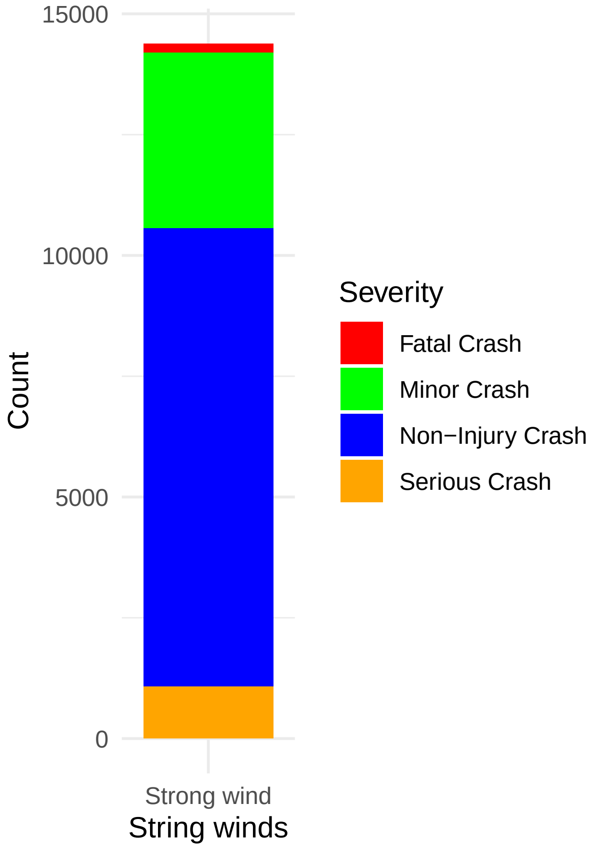
Table 8: Counts of Crash Severity Levels for Strong Wind Cases by Region

crashSeverity	region	count
Non-Injury Crash	Auckland Region	2271
Minor Crash	Auckland Region	789
Serious Crash	Auckland Region	198
Fatal Crash	Auckland Region	27
Non-Injury Crash	Bay of Plenty Region	418
Minor Crash	Bay of Plenty Region	152
Serious Crash	Bay of Plenty Region	43
Fatal Crash	Bay of Plenty Region	5
Non-Injury Crash	Canterbury Region	760
Minor Crash	Canterbury Region	340
Serious Crash	Canterbury Region	116
Fatal Crash	Canterbury Region	24
Non-Injury Crash	Gisborne Region	130
Minor Crash	Gisborne Region	42
Serious Crash	Gisborne Region	18
Fatal Crash	Gisborne Region	2
Non-Injury Crash	Hawke’s Bay Region	247
Minor Crash	Hawke’s Bay Region	109

crashSeverity	region	count
Serious Crash	Hawke’s Bay Region	37
Fatal Crash	Hawke’s Bay Region	12
Non-Injury Crash	Manawatū-Whanganui Region	780
Minor Crash	Manawatū-Whanganui Region	317
Serious Crash	Manawatū-Whanganui Region	87
Fatal Crash	Manawatū-Whanganui Region	26
Non-Injury Crash	Marlborough Region	110
Minor Crash	Marlborough Region	39
Serious Crash	Marlborough Region	21
Fatal Crash	Marlborough Region	5
Non-Injury Crash	Nelson Region	55
Minor Crash	Nelson Region	22
Serious Crash	Nelson Region	6
Non-Injury Crash	Northland Region	468
Minor Crash	Northland Region	190
Serious Crash	Northland Region	55
Fatal Crash	Northland Region	8
Non-Injury Crash	Otago Region	531
Minor Crash	Otago Region	275
Serious Crash	Otago Region	93
Fatal Crash	Otago Region	18
Non-Injury Crash	Southland Region	473
Minor Crash	Southland Region	227
Serious Crash	Southland Region	78
Fatal Crash	Southland Region	14
Non-Injury Crash	Taranaki Region	399
Minor Crash	Taranaki Region	127
Serious Crash	Taranaki Region	45
Fatal Crash	Taranaki Region	11
Non-Injury Crash	Tasman Region	67
Minor Crash	Tasman Region	29
Serious Crash	Tasman Region	4
Non-Injury Crash	Waikato Region	873
Minor Crash	Waikato Region	368
Serious Crash	Waikato Region	100
Fatal Crash	Waikato Region	25
Non-Injury Crash	Wellington Region	1754
Minor Crash	Wellington Region	550
Serious Crash	Wellington Region	146
Fatal Crash	Wellington Region	17
Non-Injury Crash	West Coast Region	81
Minor Crash	West Coast Region	30
Serious Crash	West Coast Region	17

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.

Fanny Malin, Satu Innamaa, Ilkka Norros. 2019. "Accident Risk of Road and Weather Conditions on Different Road Types." *ScienceDirect* 122 (1): 181–88.