DATA303 Interim Report - Group 5

Michael Fry 300570669 Fletcher Smith Matthew Smaill

Monday, 04 September

Contents

Background and Data

Background

The New Zealand Crash Analysis System (CAS) dataset is a comprehensive compilation of traffic crash information recorded by the Waka Kotahi, the New Zealand Transport Agency. The CAS dataset constitutes a valuable resource for gaining insights into the factors contributing to traffic crashes across New Zealand's diverse roadways and public-access areas. This report provides an overview of the CAS dataset, its significance, contents, and inherent characteristics.

Dataset (source and coverage)

The CAS data originates from the Waka Kotahi Crash Analysis System, which serves as a repository for all reported traffic crashes involving motor vehicles in New Zealand. This system is only fueled by information provided by the New Zealand Police. The scope of CAS encompasses crashes that occur on any road segment or area within the country where the public has legal access with a motor vehicle. This extensive coverage ensures that the dataset represents a wide array of scenarios, road types, and conditions.

Importance

The CAS dataset is of considerable interest due to its potential to address critical questions surrounding road safety and accident prevention. One of the central questions that this dataset can help answer is: "What statistical techniques can we use to find the relational effect that key variables have on major automotive crashes?" By analyzing the dataset, we can identify patterns, correlations, and trends that shed light on the factors contributing to major vehicle crashes.

Types of Data

The CAS dataset incorporates various types of data, each contributing to a holistic understanding of traffic crashes. This dataset comprises 12 logical variables, 2 date variables, 15 categorical variables, and 41 numeric variables. The inclusion of diverse data types allows for a multi-faceted analysis that captures both quantitative and qualitative aspects of crash incidents.

Completeness

It's important to note that the CAS dataset, while comprehensive, does contain missing values. Out of all the columns in the dataset, only X, Y, ObjectID, and crashYear are entirely devoid of missing values. However, various other variables exhibit significant instances of missing data. For instance, variables such as Bridge, debris, fence, vehicle, and waterRiver each have 488,831 missing values. This variation in missing data across variables underscores the complexity of real-world data collection and emphasizes the need for careful consideration when conducting analyses or drawing conclusions.

In conclusion, the New Zealand Crash Analysis System (CAS) dataset serves as a valuable resource for investigating the dynamics of traffic crashes in New Zealand. Its extensive coverage, diverse data types, and potential to answer crucial questions make it an essential tool for researchers, policymakers, and analysts aiming to enhance road safety and prevent major automotive crashes. However, the presence of missing data underscores the importance of thorough data preprocessing and analysis techniques to ensure accurate and meaningful insights.

Ethics and Privacy

Privacy and ethical considerations are paramount for data scientists when dealing with, and modelling data. Especially sensitive data. This section of the report details the privacy and ethical concerns that must be addressed when dealing with the crash (CAS) dataset to answer the question: Can we use statistical analysis techniques to find the relational effects that key factors have on leading to a major crash using the CAS dataset? It also discusses what steps could be taken to ensure the project data and results are secure. Privacy considerations refers to the protection of individuals' personal information and sensitive data. It involves ensuring that data is collected, stored, processed, and shared in a way that respects individuals' rights and maintains confidentiality. Ethical considerations on the other hand, is all about how the data is being used, especially when it comes to potential biases, fairness, and social implications. It is important to note; the CAS dataset is publicly available and provided by a government entity. Therefore, the following discussion includes BOTH methods that the NZTA has done to try and privatise the data, AND also the privacy and ethical considerations that still need to be taken into account within this project.

Privacy

Privacy considerations are of utmost importance when dealing with any dataset, especially one as sensitive as the Crash Analysis System (CAS). Personal identifiers, including the gender and age of the driver, and license plate numbers, have been deliberately removed from the dataset by the NZTA. These removals are critical steps taken to protect individual privacy. However, even though these direct identifiers have been removed, this alone may not suffice to guarantee privacy. The dataset still contains information about location and the financial year in which the crash occurred. While not as overt as personal identifiers, this information can still be used, in combination with other factors, to potentially identify individuals involved in crashes. Location data, in particular, can provide insights into the vicinity of the crash, which may be sensitive information. It is import for us data scientists to consider this, as it could impact peoples lives, as our analysis hasn't come with the direct permission with the participants for the use of their data (indirectly has been through the NZTA).

Ethics

Ethical considerations encompass a broad spectrum of concerns that go beyond mere data protection, extending to issues of fairness, bias mitigation, and the broader social implications of data analysis. Given that the CAS dataset is publicly accessible and provided by a government entity, the New Zealand Transport Agency (NZTA), ethical scrutiny must encompass an evaluation of the effectiveness of the NZTA's data privatization efforts. It is imperative to determine whether these efforts are robust enough to adequately safeguard individuals' privacy and whether the dataset aligns with ethical data sharing standards. Furthermore, despite the public availability of the CAS dataset, ethical concerns emerge regarding the consent of individuals whose data is included. Ensuring that the data used in research has been obtained and shared in accordance with informed consent principles is a cornerstone of ethical data analysis, particularly when dealing with sensitive crash data. This might be the case for individuals when they consent for the use of this data by NZTA, however, we have not asked for consent directly by any crash participants. Therefore, we need to do our crash analysis with respect of the people whose data we are using.

Analyzing crash data carries significant societal implications, potentially influencing policy decisions and public perceptions. Ethical considerations require data scientists to approach our analysis with a commitment to minimizing harm and promoting societal well-being. We should carefully consider how their findings may impact communities and advocate for responsible data use, recognizing the sensitivity of the information involved.

Steps to Ensure Privacy and Ethical Compliance:

1. Data Anonymization: The first step involves a rigorous review of the dataset to eliminate or anonymize any remaining personally identifiable information. This process aims to further protect individuals' privacy by removing the possibility of identification.

2. Adding Noise: To bolster privacy protection, consider adding noise to sensitive attributes like crash dates. This technique obscures specific details about individuals while still allowing meaningful analysis, thereby striking a balance between utility and privacy. If the location variables, where to be used in our analysis, then we could move the area of crash within a 10km radius, which wouldn't effect the overall results of our findings.
3. Ethical Review: Conduct a comprehensive ethical review of the research project. This review should identify potential biases and ethical dilemmas, requiring consultation with experts in ethics and data privacy to ensure responsible research.
4. Fairness Assessment: Scrutinize the dataset for fairness issues, such as disparities in crash reporting or data collection. Mitigate any identified biases during data analysis and reporting to ensure fairness in the research.

Exploratory Data Analytics

Summary Tables

Aspect	Value
Number of Instances	821744
Number of Features	72
Number of Missing Values	15287398
Number of Numeric Features	50
Number of Categorical Features	22

This table shows the count of Instances, Features, Missing values, Numerical and Categorical features within the CAS Dataset. Please note, these values are different than those reported in the Background section, as it is likely that some variables will need ot be re-encoded at the data cleaning stage of the analysis.

	Count	Percent_Missing
crashRoadSideRoad	821744	100
intersection	821744	100
temporarySpeedLimit	809161	98.47
pedestrian	795139	96.76
advisorySpeed	790400	96.19
bridge	488831	59.49
cliffBank	488831	59.49
debris	488831	59.49
ditch	488831	59.49
fence	488831	59.49
guardRail	488831	59.49
houseOrBuilding	488831	59.49
kerb	488831	59.49
objectThrownOrDropped	488831	59.49
otherObject	488831	59.49
overBank	488831	59.49
parkedVehicle	488831	59.49
phoneBoxEtc	488831	59.49
postOrPole	488831	59.49
roadworks	488831	59.49
slipOrFlood	488831	59.49
strayAnimal	488831	59.49
trafficIsland	488831	59.49
trafficSign	488831	59.49
train	488831	59.49
tree	488831	59.49
vehicle	488831	59.49
waterRiver	488831	59.49

This table shows the count and percentage of missing values for each variable in the dataset with more than 10% of the data missing. The table provides insights into the completeness of the dataset. Columns with higher percentages of missing values may require further investigation or data imputation strategies.

Plots

Advisory Speed Distribution

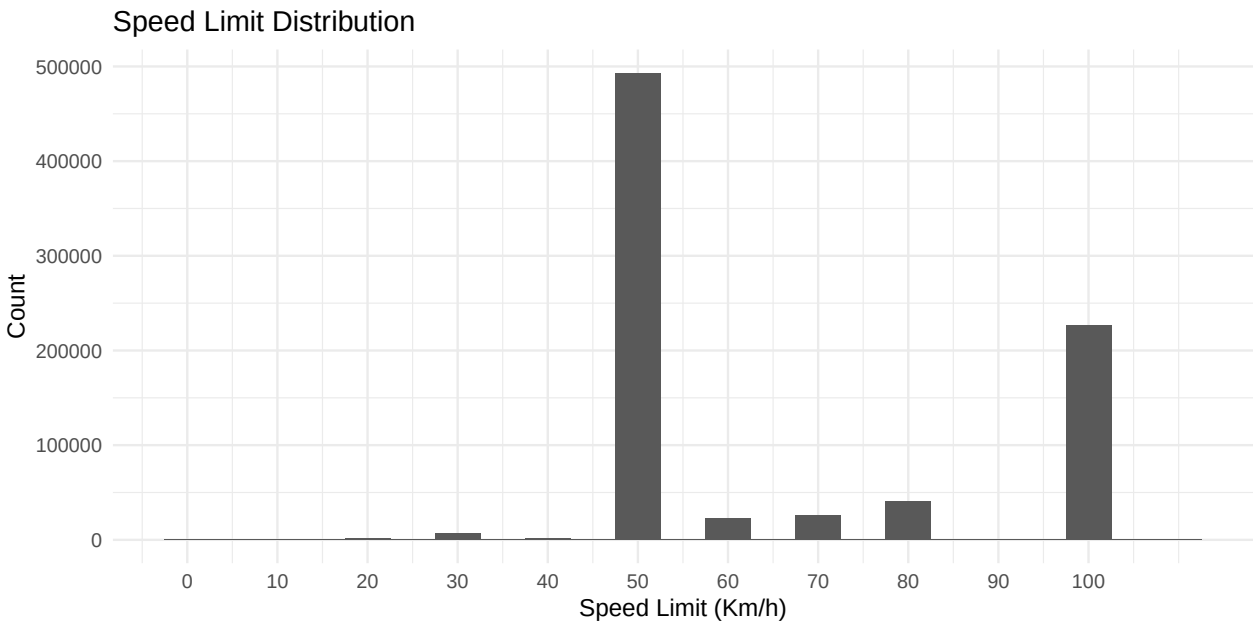

Figure 1: Histogram for distribution of advisory speeds in the dataset. Most of the advisory speeds are clustered around 50-60 km/h, which is typical for urban andsuburban areas. From advisory speeds 60 through to 100, more fatal crashes seem to occur.

Speed Limit Distribution

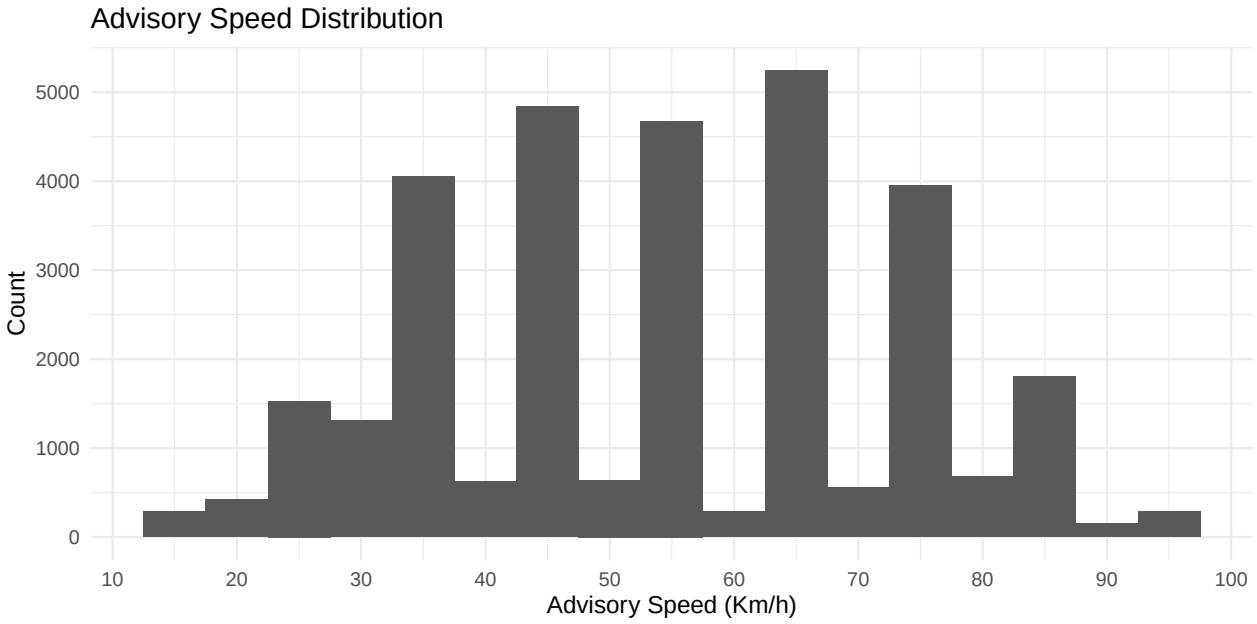

Figure 2: The histogram for speed limits shows a similar trend to the advisory speed, where most of the crashes occur around the 50-60 km/h range.

Boxplot of advisorySpeed vs crashSeverity

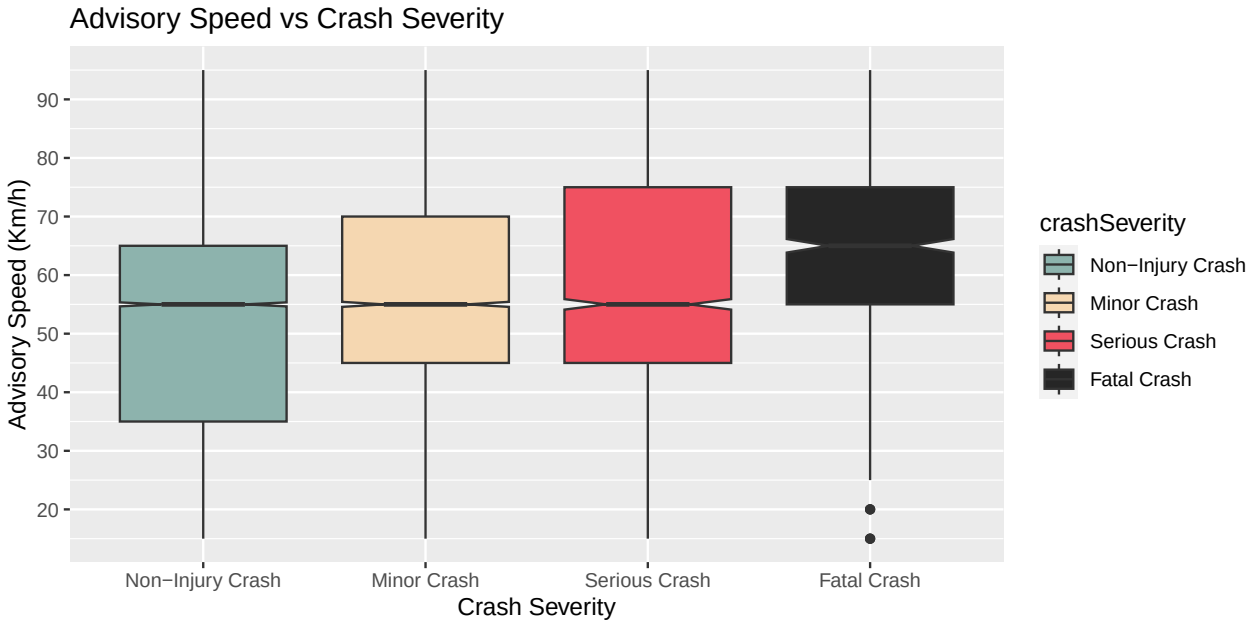

Figure 3: The boxplot between advisory speed and crash severity shows that the advisory speed tends to increase slightly with the severity of the crash. The notches in the fatal crash boxplot suggests that there is a statistically significant difference between the medians of Non-injury, minor and serious crashes with fatal crashes. This plot may imply that higher advisory speeds are associated with more severe crashes, although correlation does not imply causation.

Boxplot of speedLimit vs crashSeverity

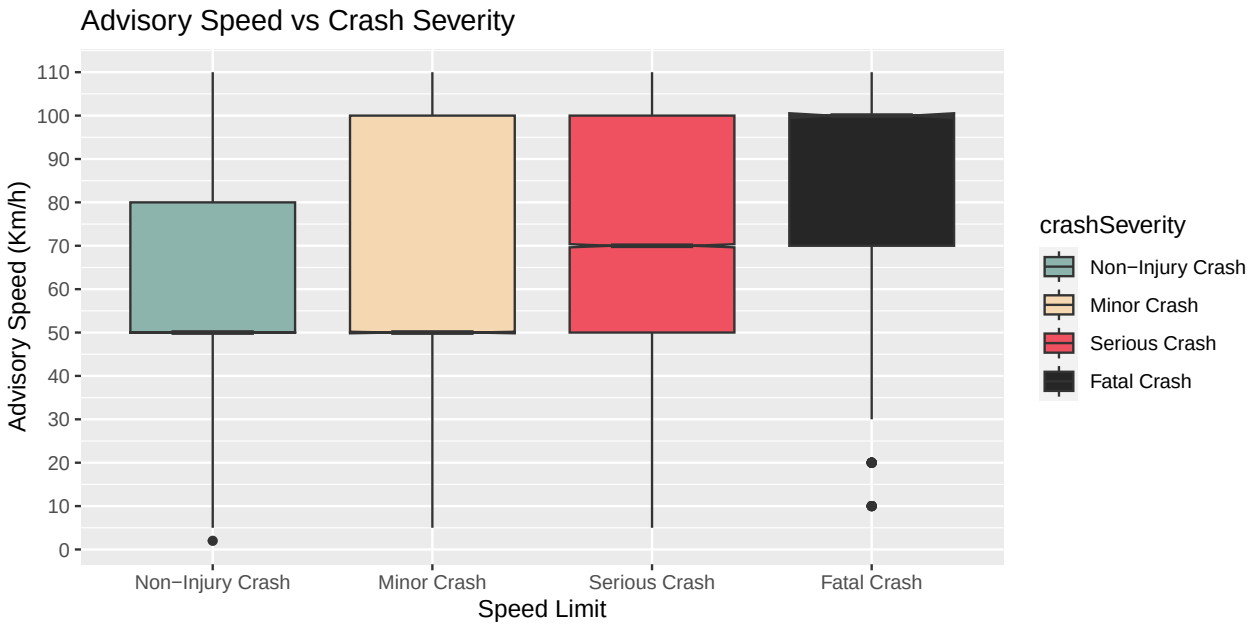

Figure 4: The boxplot between Speed Limit and crash severity shows some of the same relationship as the advisory speed graph above.

Number of lanes

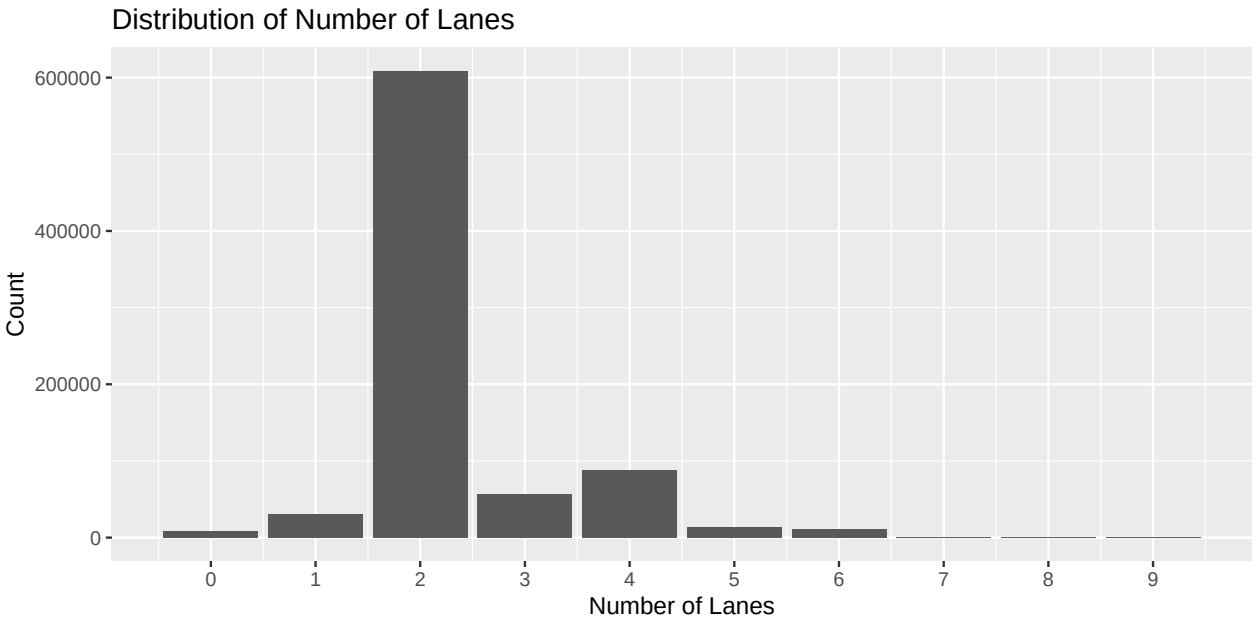

Figure 5: The bar plot shows that most crashes occur on roads with 1 or 2 lanes.

Barplot of Fatalities by Vehicle Type

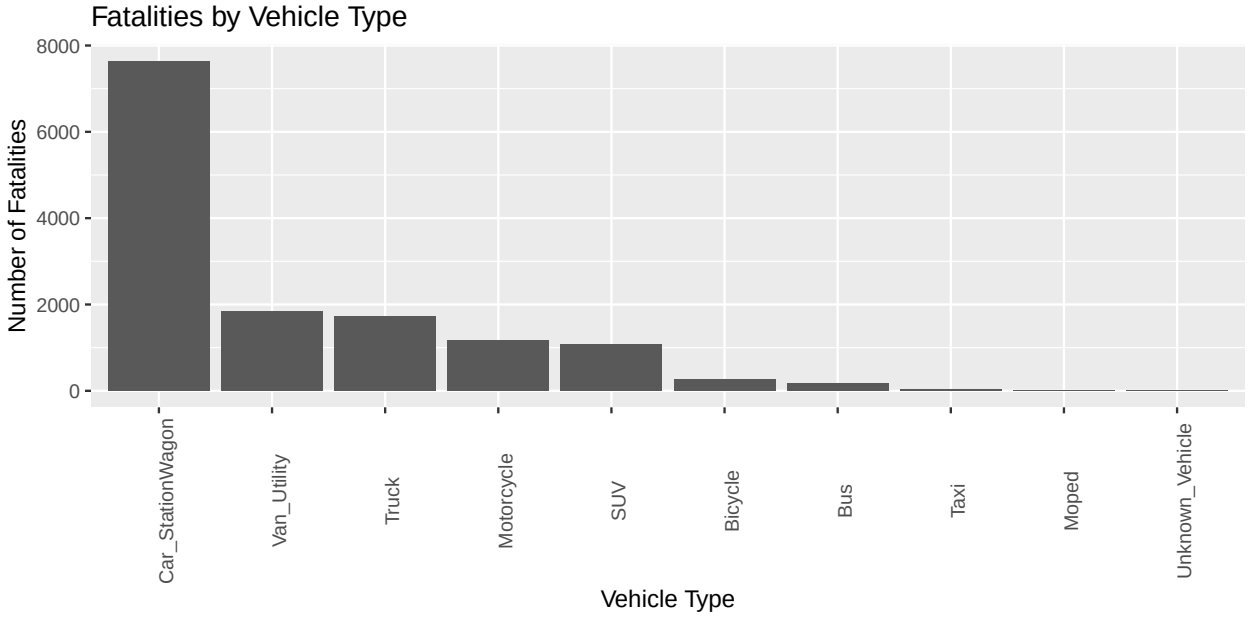

Figure 6: The bar plot shows the number of fatalities by vehicle type. Cars and station wagons have the highest number of fatalities, though this is likely simply because they are more common on the roads.

Fatalities by year

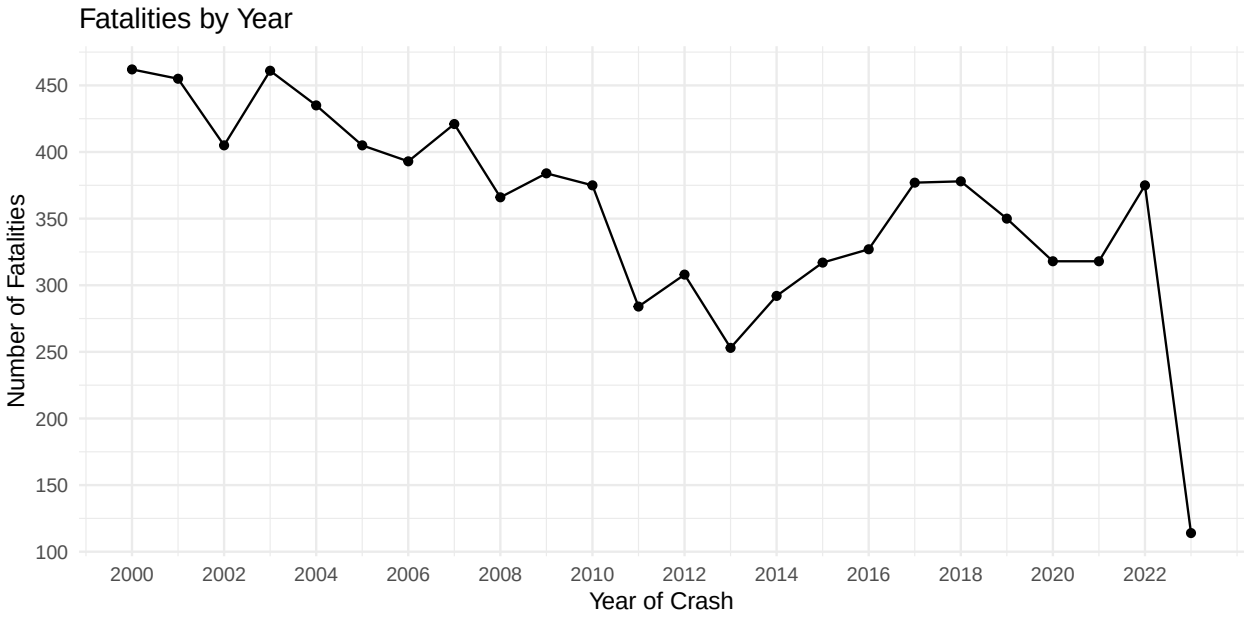

Figure 7: The line plot shows a clear trend in the number of fatalities. The period between 2000 and 2013 showes a clear downward trend of the number of fatalities. The period between 2013 and 2022 (Excluding 2023 as there is not data for the whole year) there is a clear upward trend in number of fatalities. This could be due to numerious things, population growth, increased reporting or other factors.

Fatalities by weather A

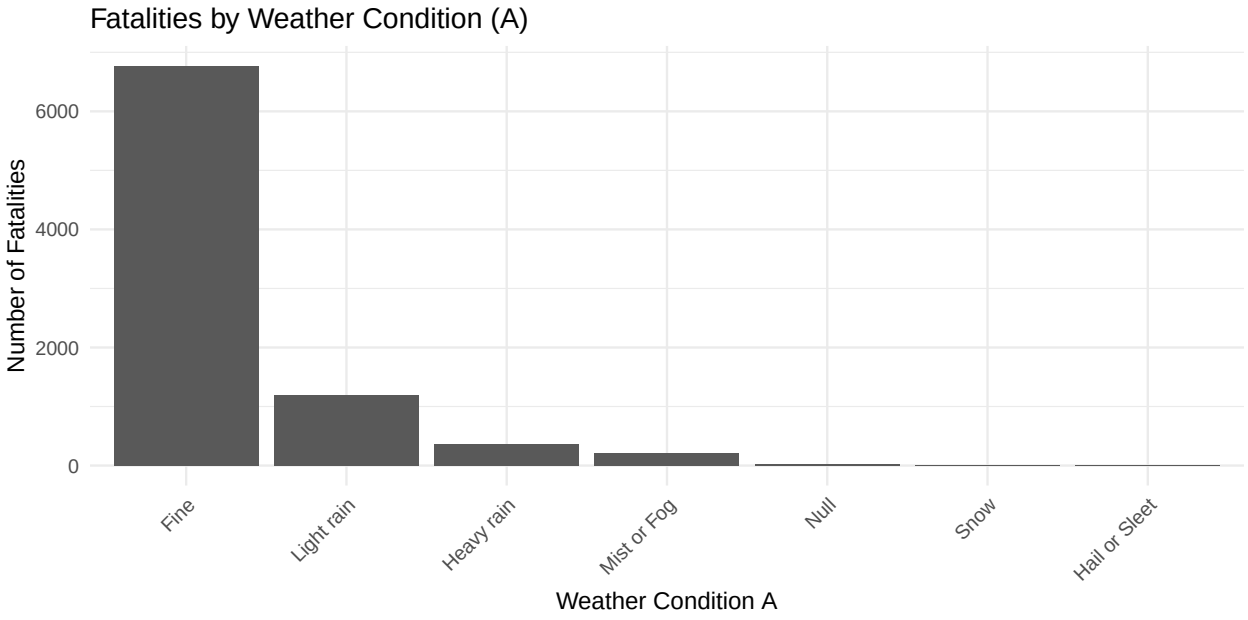

Figure 8: The bar plot shows the number of fatalities categorized by primary weather conditions.

Fatalities by weather B

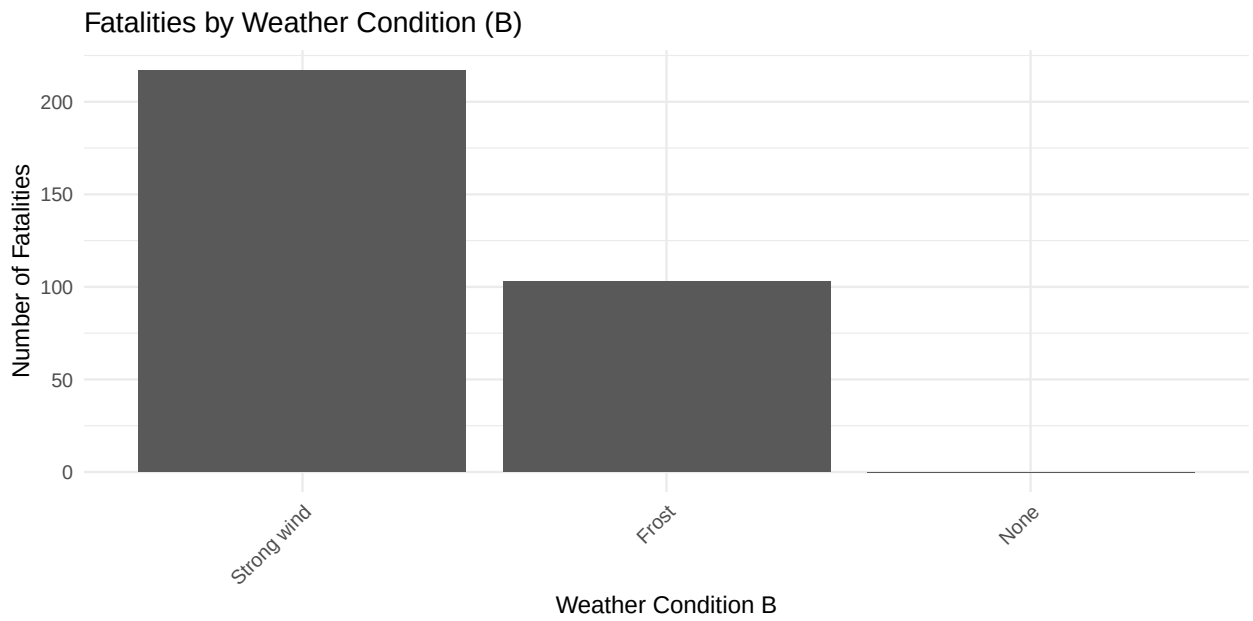

Figure 9: Similar to weatherA, this bar plot shows fatalities by secondary weather conditions.

Crash Severity

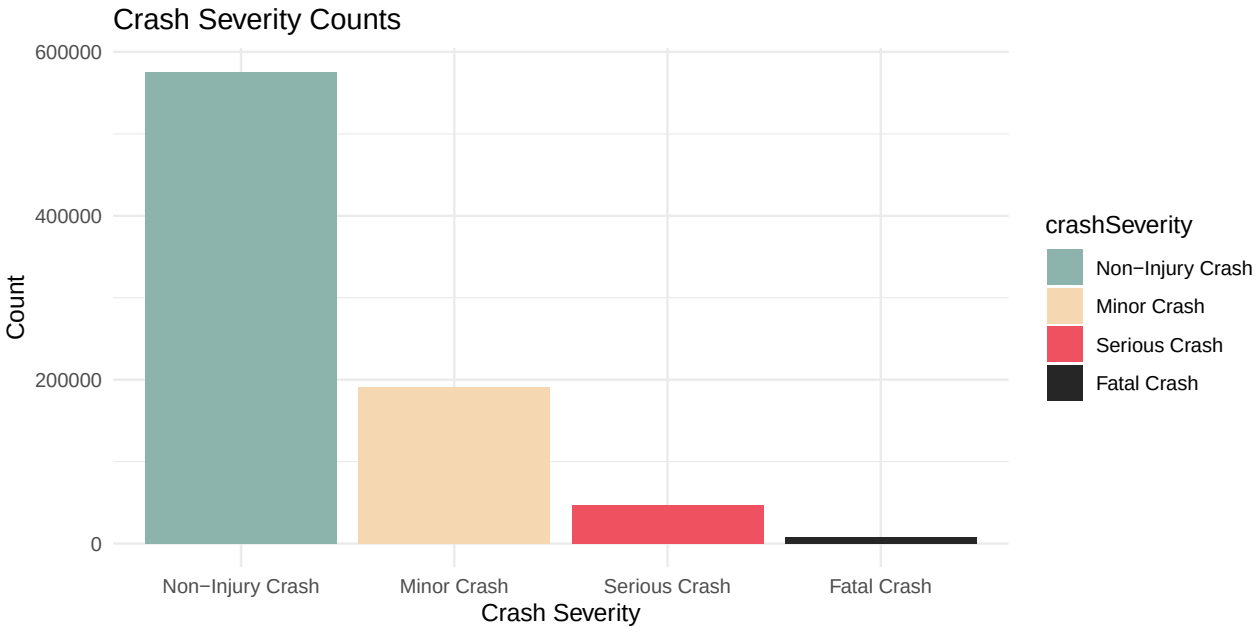

Figure 10: The bar plot gives an overview of the crash severity distribution in the dataset. Most crashes are non-injury or minor, but there is a significant number of serious and fatal crashes as well.

Crash Severity by Region

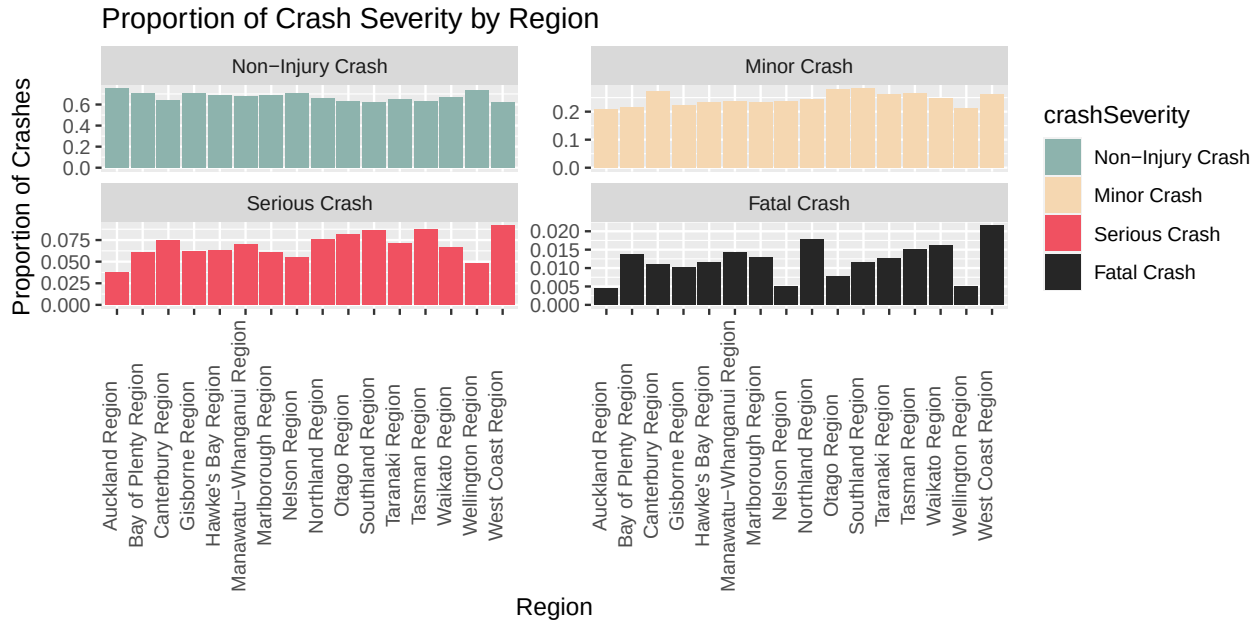

Figure 11: The bar plot shows the regional distribution of proportion of crashes at each crash severity. This can give insights into which regions have a higher proportion of more severe crashes.

Crash Severity by Region - Stacked

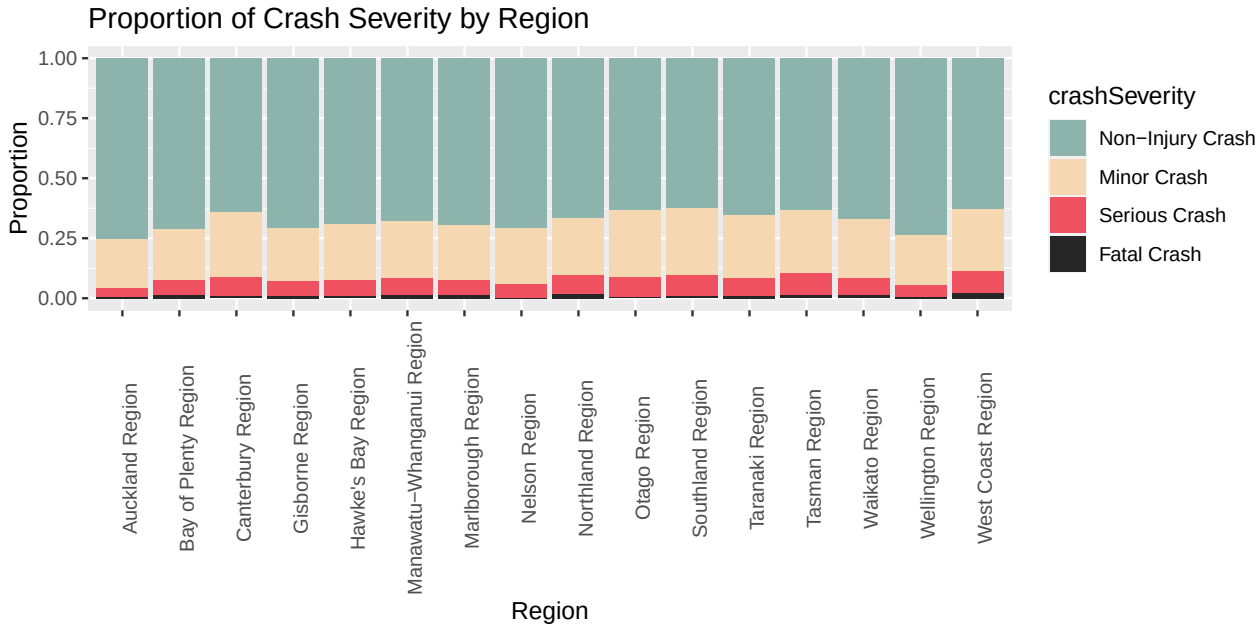

Figure 12: The stacked bar plot shows the same regional distribution of the proportion of crashes at each severity as above. The stacked nature of the bars allows for a different view of the data, providing an easier way to compare regions with one plot.

Heatmap of Fatal Crashes by year and region.

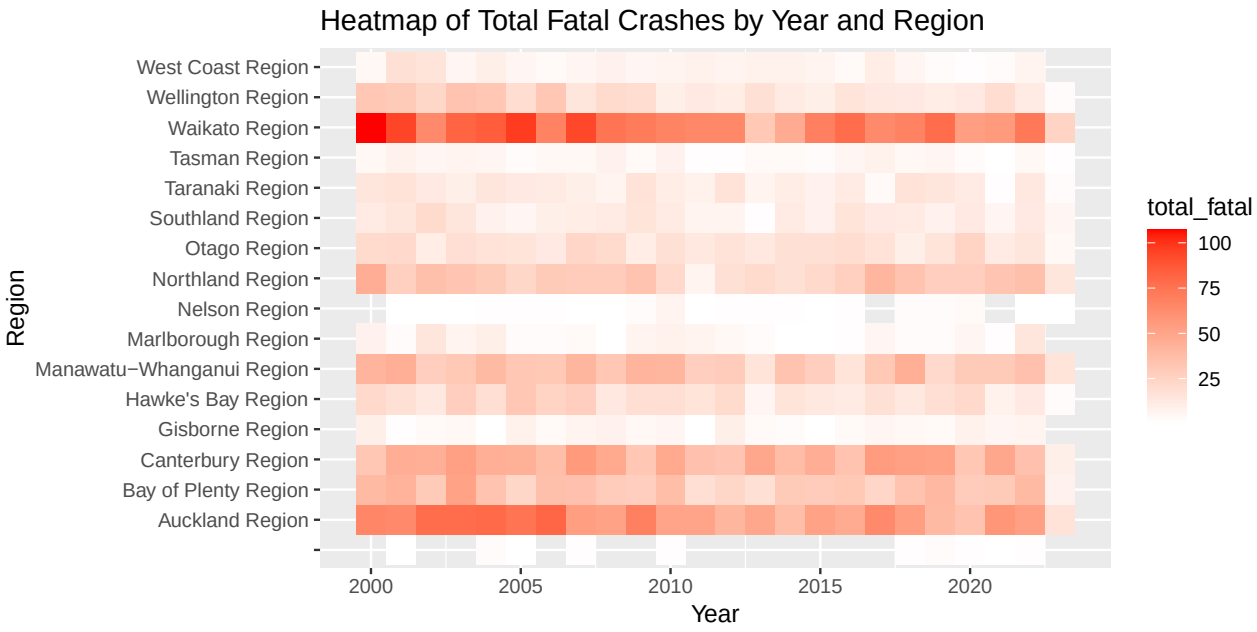

Figure 13: The heatmap shows the distribution of total fatal crashes by year and region. Darker colors represent higher numbers of fatal crashes.

Things to note going forward

There are a significant number of variables with missing data. Each will need to be handled differently. Significant pre-processing will be needed before the fitting of any models.

It is likely, due to the significant number of features present in the dataset, that significant feature selection will be required to reduce the dimensionality of the dataset to a usable level.

Individual Contributions

Michael Fry

- Created and formatted the interm report document
  - Knittr options, layout, formatting etc.
- Exploratory Data Analytics Section
  - Summary
  - Grouping etc
  - Plots
  - Captions

Fletcher Smith

Matthew Smaill

References