DATA303 Interim Report - Group 5

Michael Fry 300570669 Fletcher Smith Matthew Smaill

Sunday, 03 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

Ethics

Privacy

Security

Exploratory Data Analytics

```
## `summarise()` has grouped output by 'crashSeverity'. You can override using the
## `.groups` argument.
```

Summary Tables

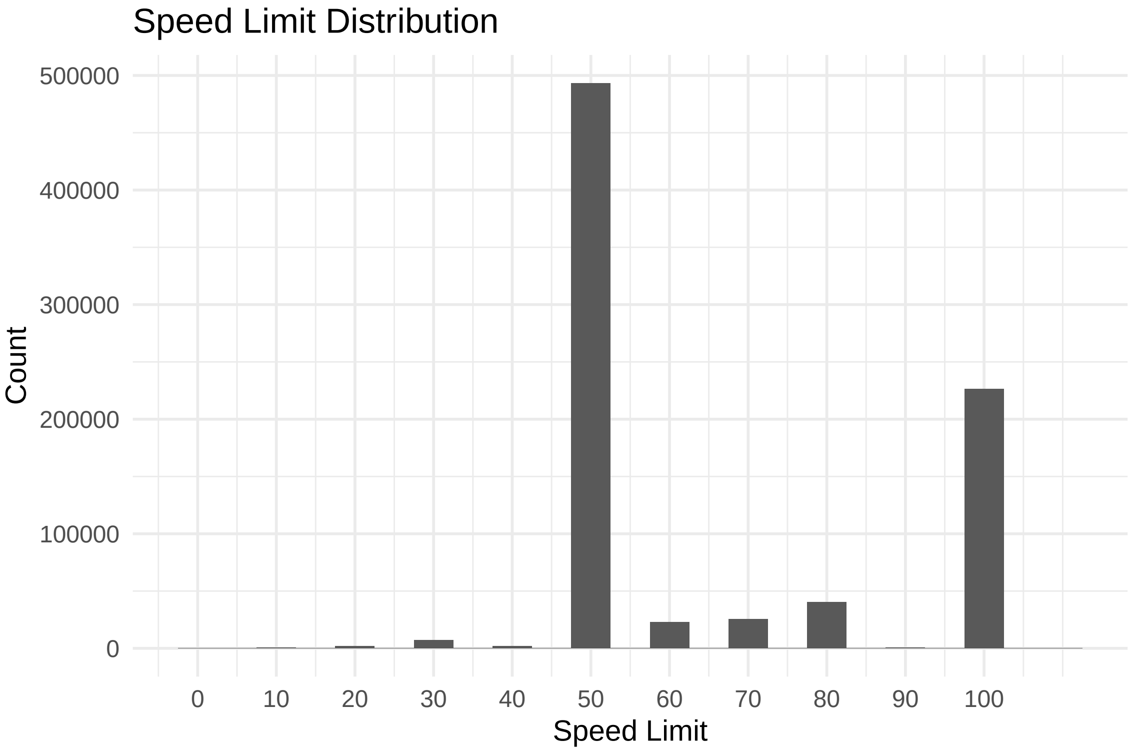
```
##                         Count Percent_Missing
## crashRoadSideRoad      821744       100.00000
## intersection           821744       100.00000
## temporarySpeedLimit    809161        98.46874
## pedestrian             795139        96.76237
## advisorySpeed          790400        96.18567
## bridge                 488831        59.48702
## cliffBank              488831        59.48702
## debris                 488831        59.48702
## ditch                  488831        59.48702
## fence                  488831        59.48702
## guardRail              488831        59.48702
## houseOrBuilding        488831        59.48702
## kerb                   488831        59.48702
## objectThrownOrDropped  488831        59.48702
## otherObject            488831        59.48702
## overBank               488831        59.48702
## parkedVehicle          488831        59.48702
## phoneBoxEtc            488831        59.48702
## postOrPole             488831        59.48702
## roadworks              488831        59.48702
## slipOrFlood            488831        59.48702
## strayAnimal            488831        59.48702
## trafficIsland          488831        59.48702
## trafficSign            488831        59.48702
## train                  488831        59.48702
## tree                   488831        59.48702
## vehicle                488831        59.48702
## waterRiver             488831        59.48702
```

This table shows the count and percentage of missing values for each variable in the dataset with more than 10% of the data missing.

Commentary: 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.

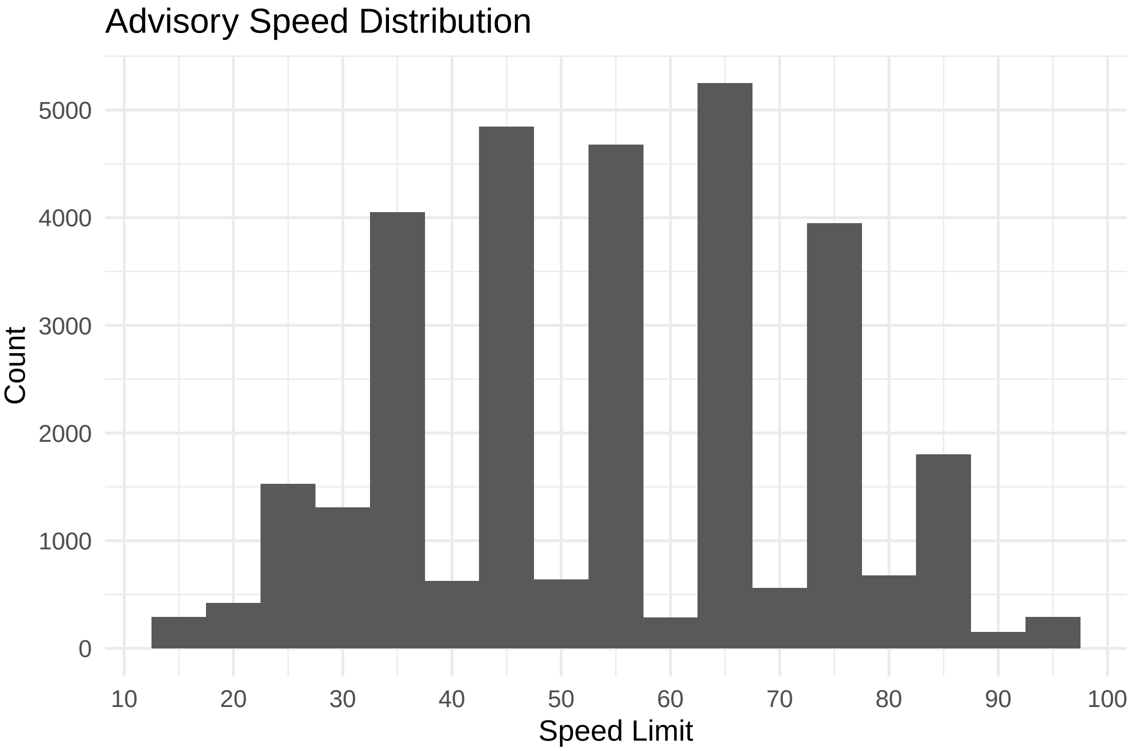
Plots

Advisory Speed Distribution



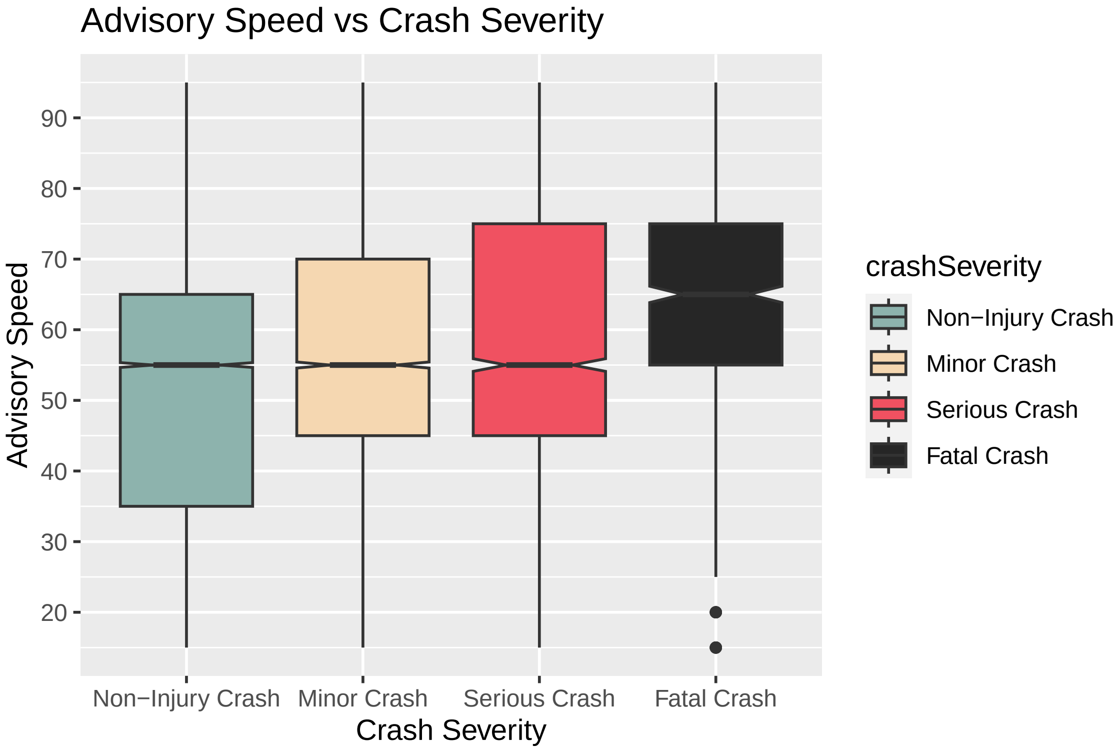
Commentary: The histogram shows the distribution of advisory speeds in the dataset. Most of the advisory speeds are clustered around 50-60 km/h, which is typical for urban and suburban areas. From advisory speeds 60 through to 100, more fatal crashes seem to occur.

Speed Limit Distribution



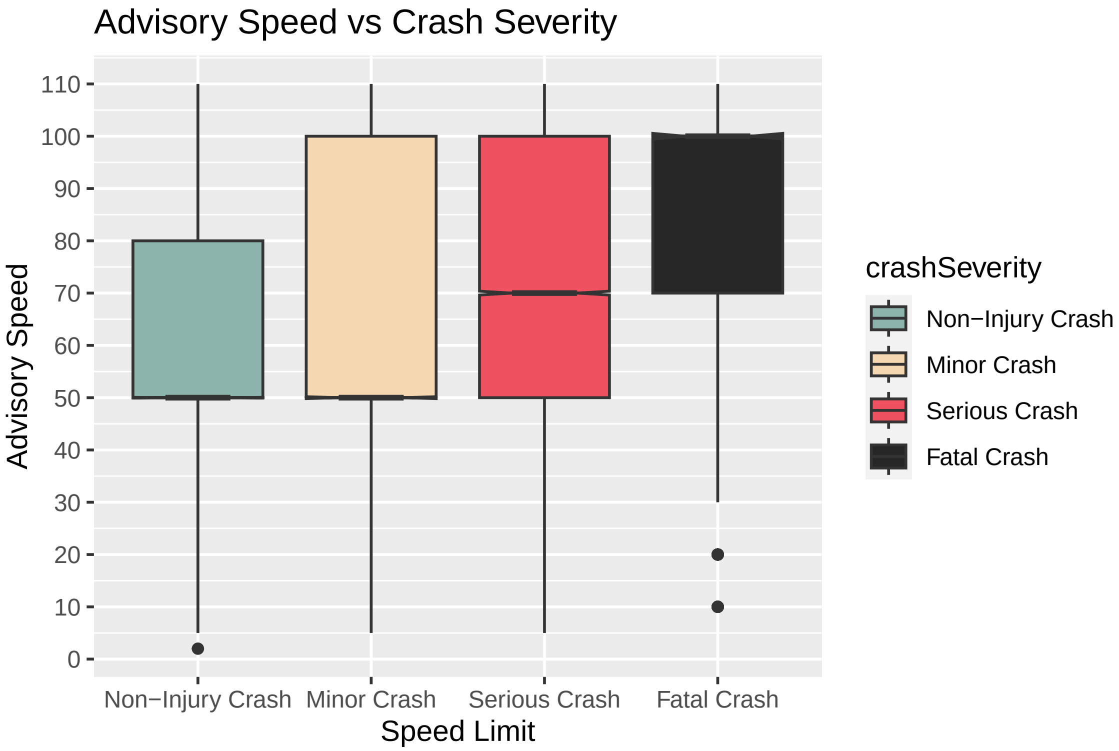
Commentary: The histogram for speed limits shows a similar trend to the advisory speed, where most of the crashes occur around the 50-60 km/h range.

Boxplot of advisorySpeed vs crashSeverity



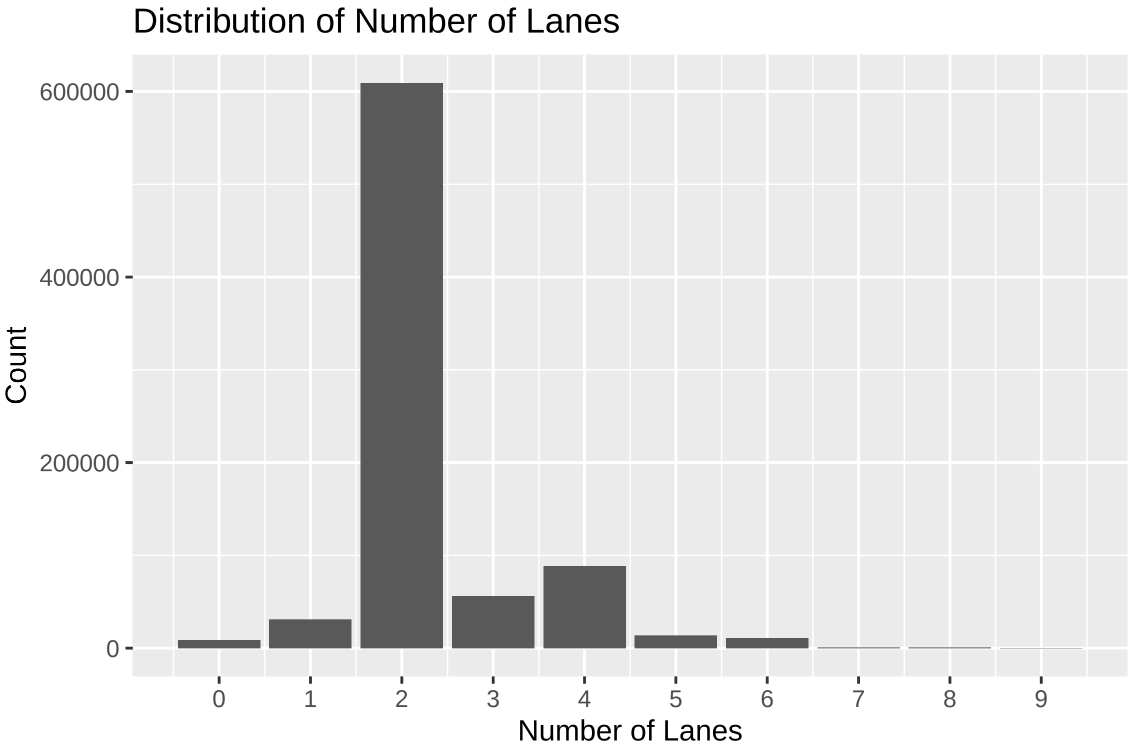
Commentary: The boxplot between advisory speed and crash severity shows that the advisory speed tends to increase slightly with the severity of the crash. The notches in the fatal crash boxplot suggests that there is a statistically significant difference between the medians of Non-injury, minor and 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.

```
## Notch went outside hinges
## i Do you want `notch = FALSE`?
## Notch went outside hinges
## i Do you want `notch = FALSE`?
## Notch went outside hinges
## i Do you want `notch = FALSE`?
```



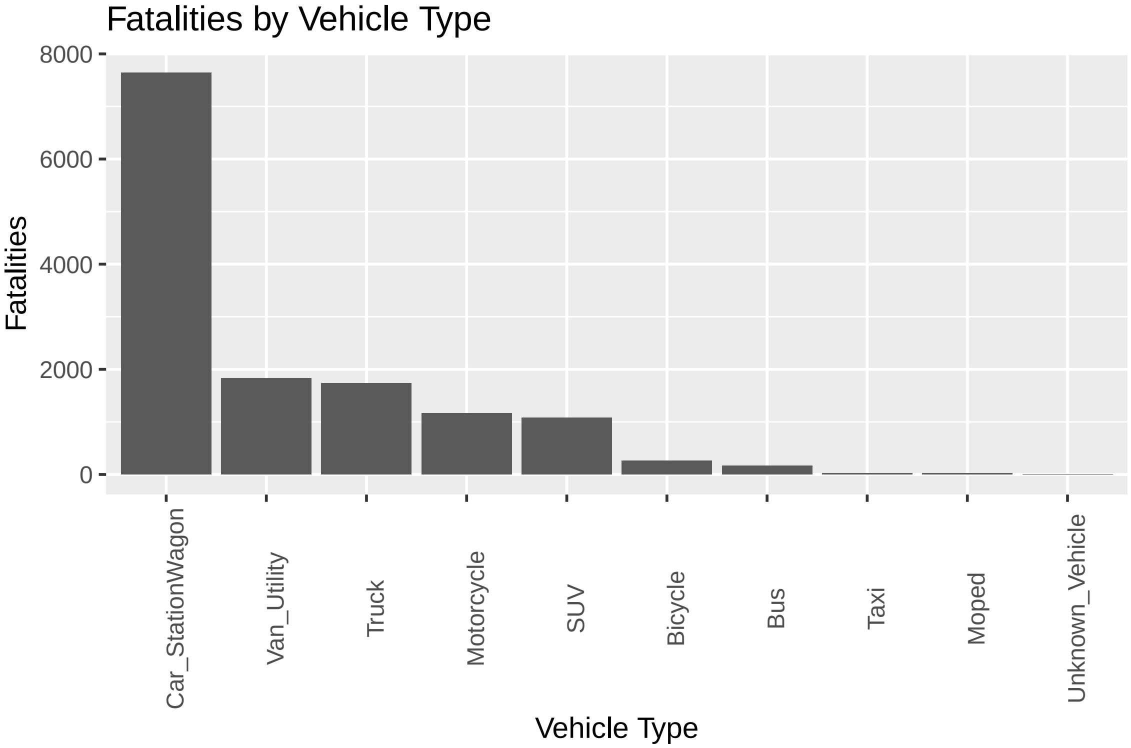
Commentary: The boxplot between Speed Limit and crash severity shows some of the same relationship as the advisory speed graph above.

Number of lanes



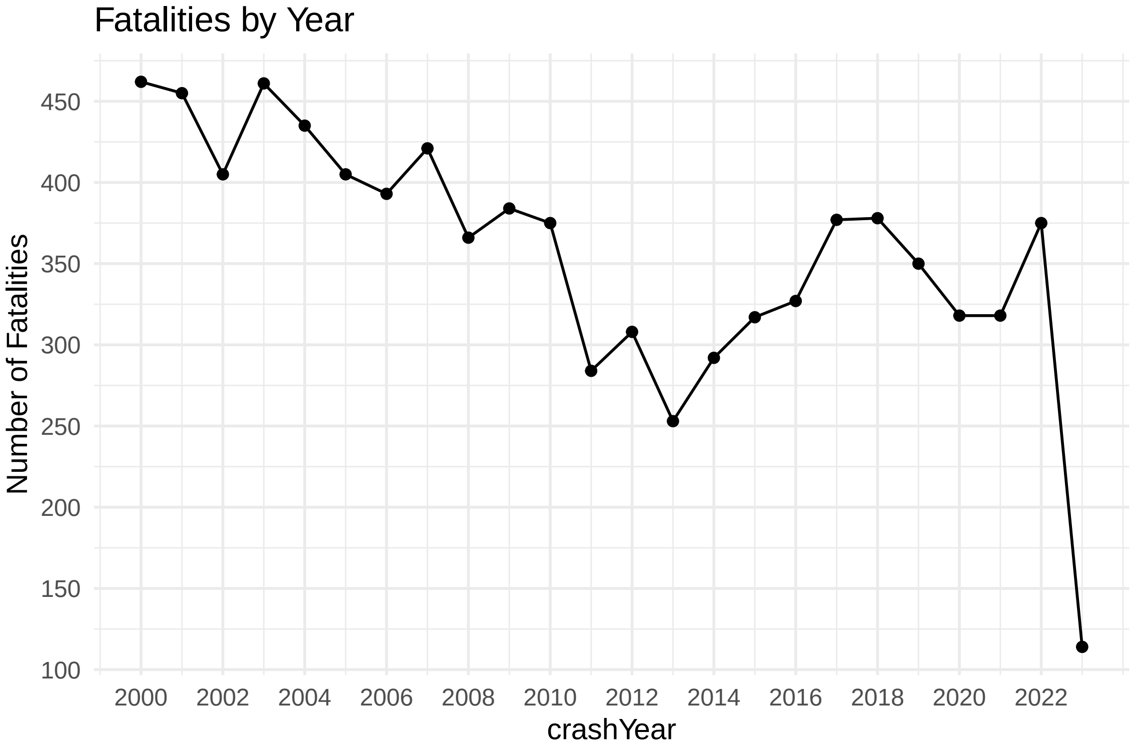
Commentary: The bar plot shows that most crashes occur on roads with 1 or 2 lanes.

Barplot of Fatalities by Vehicle Type



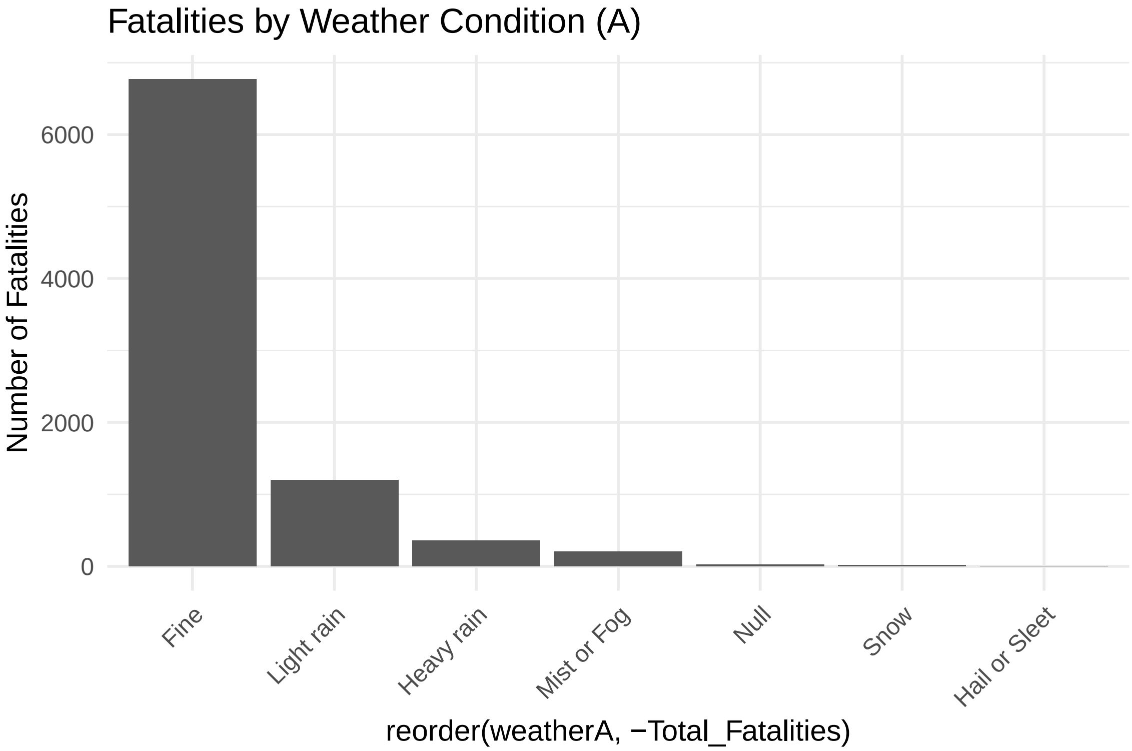
Commentary: The bar plot shows the number of fatalities by vehicle type. Cars and station wagons have the highest number of fatalities, though this is likely simply because they are more common on the roads.

Fatalities by year



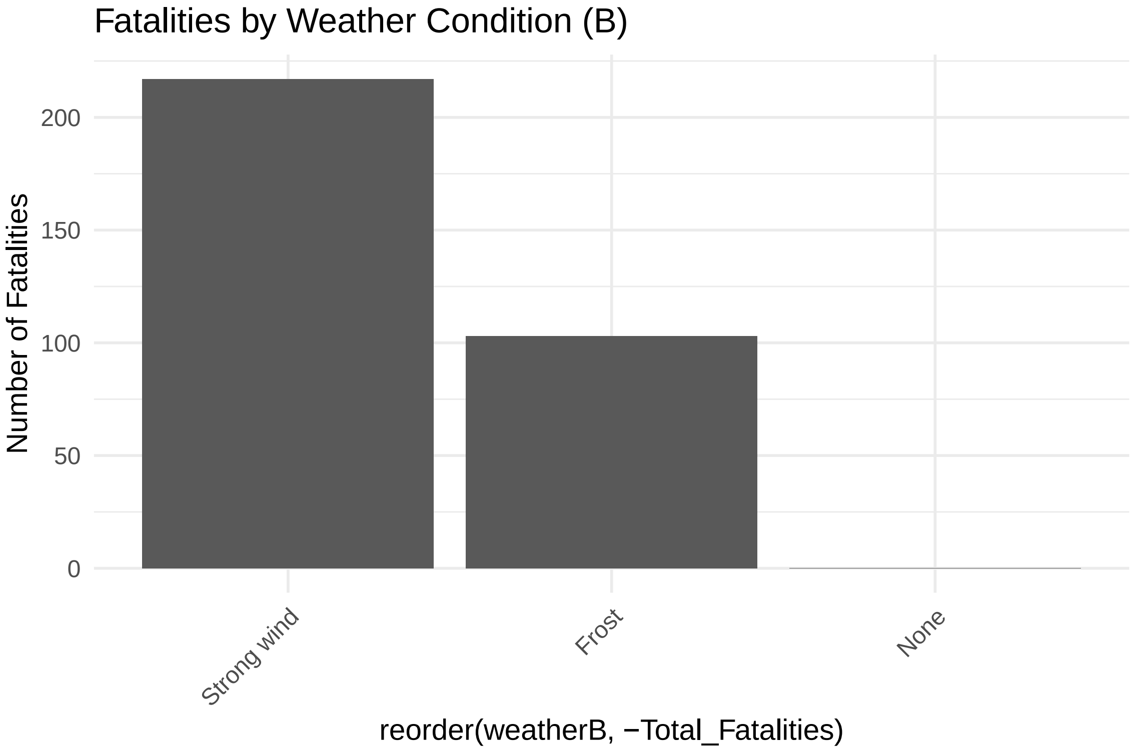
Commentary: The line plot shows a clear trend in the number of fatalities. The period between 2000 and 2013 showes a clear downward trend of the number of fatalities. The period between 2013 and 2022 (Excluding 2023 as there is not data for the whole year) there is a clear upward trend in number of fatalities.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

Fatalities by weather A



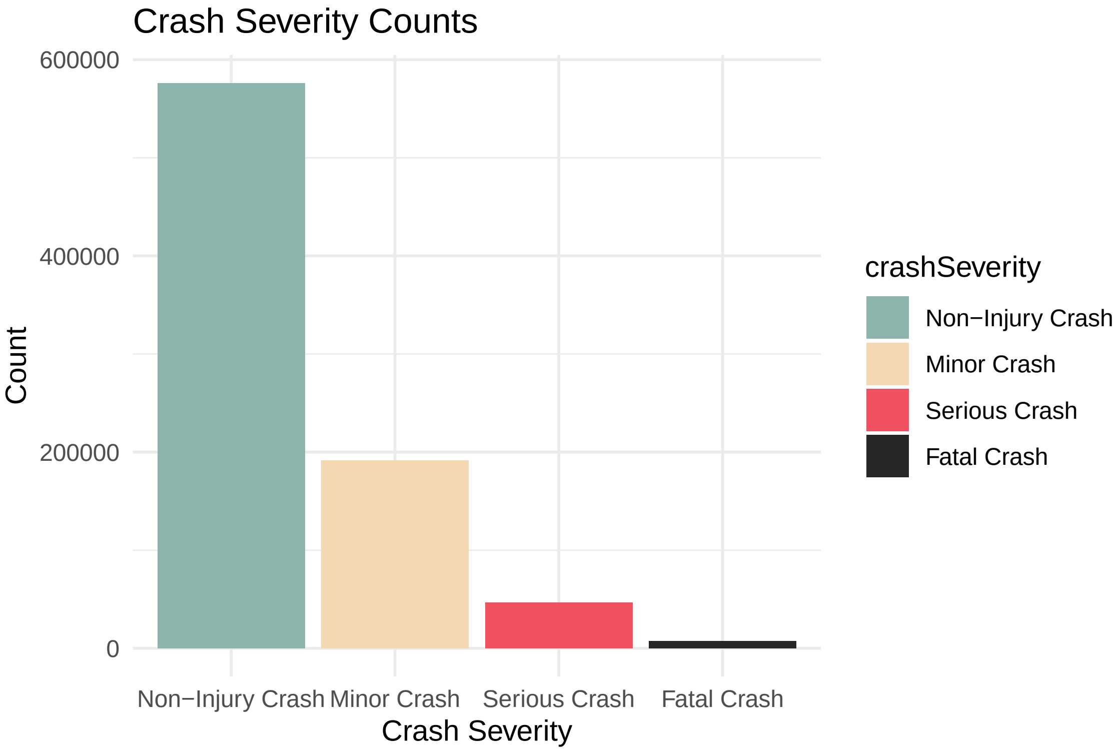
Commentary: The bar plot shows the number of fatalities categorized by primary weather conditions.

Fatalities by weather B

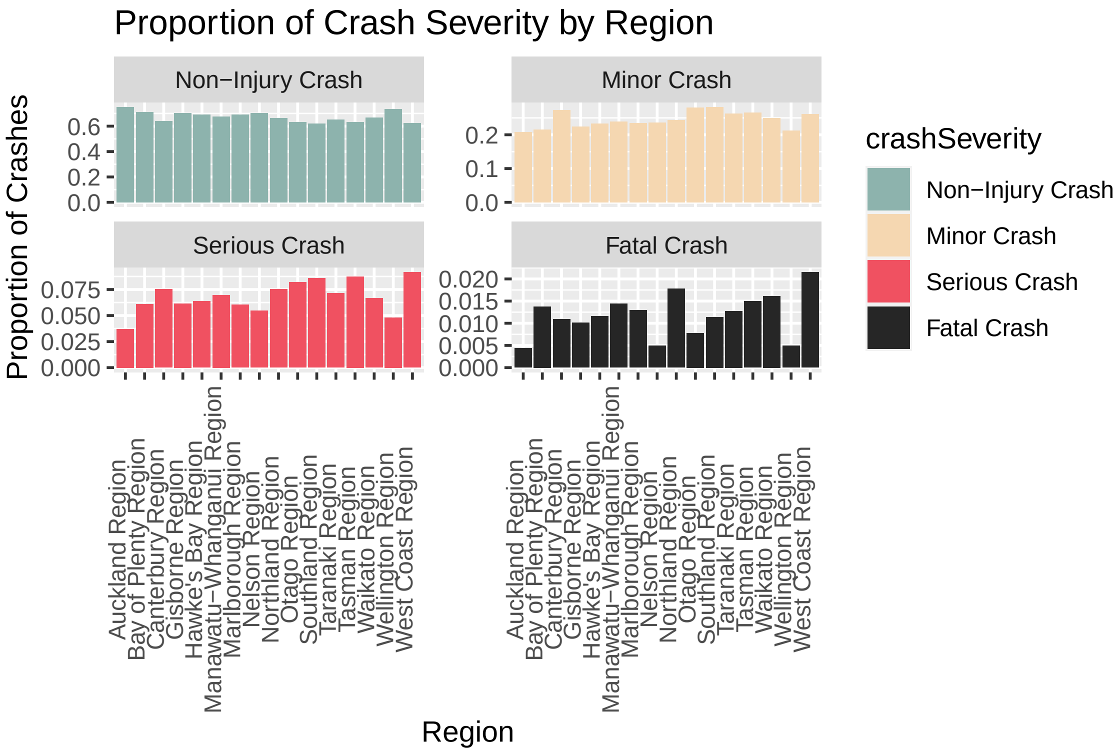


Commentary: Similar to weatherA, this bar plot shows fatalities by secondary weather conditions.

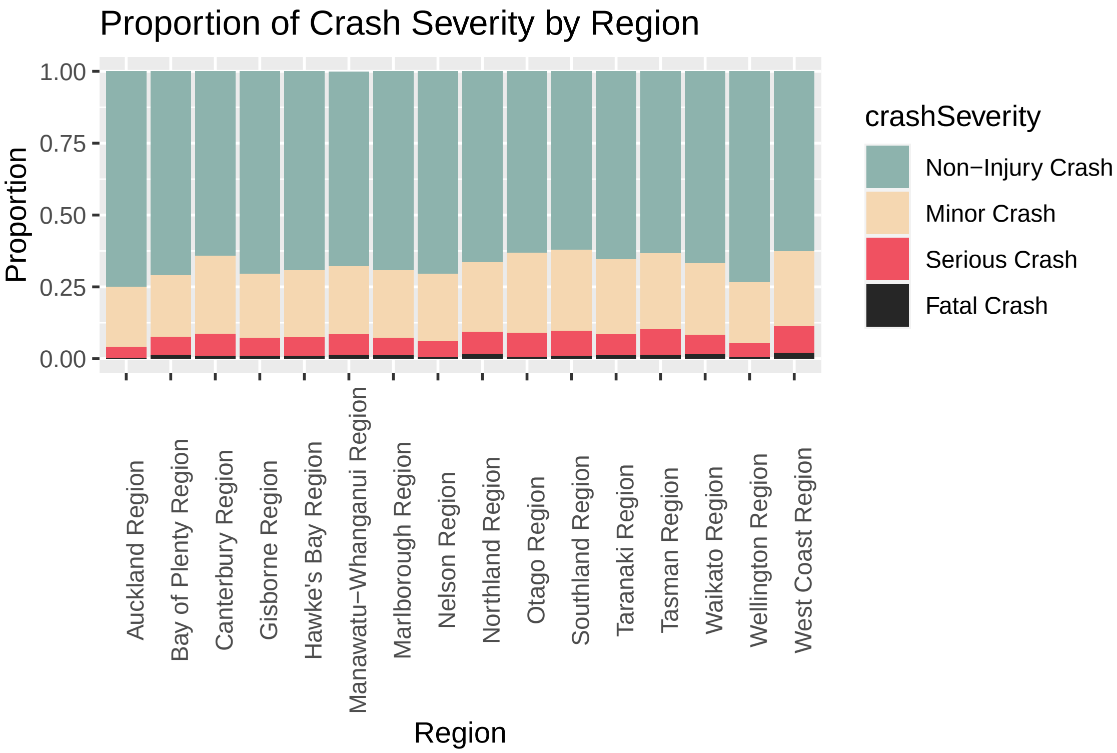
Crash Severity



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

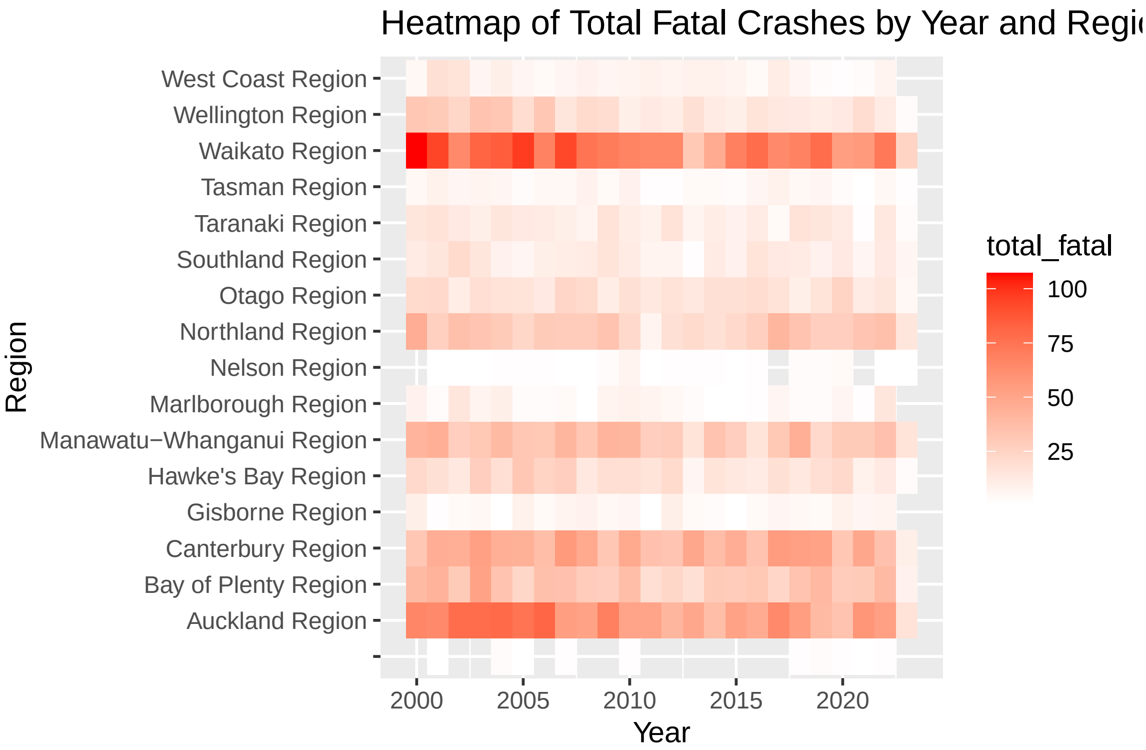


Commentary: The bar plot shows the regional distribution of proportion of crashes at each crash severity. This can give insights into which regions have a higher proportion of more severe crashes.



Commentary: 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.

```
## `summarise()` has grouped output by 'crashYear'. You can override using the
## `.groups` argument.
```



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

Individual Contributions

Michael Fry

Fletcher Smith

Matthew Smaill

References