DATA303 Interim Report - Group 5

Michael Fry 300570669 Fletcher Smith Matthew Smaill

Saturday, 02 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

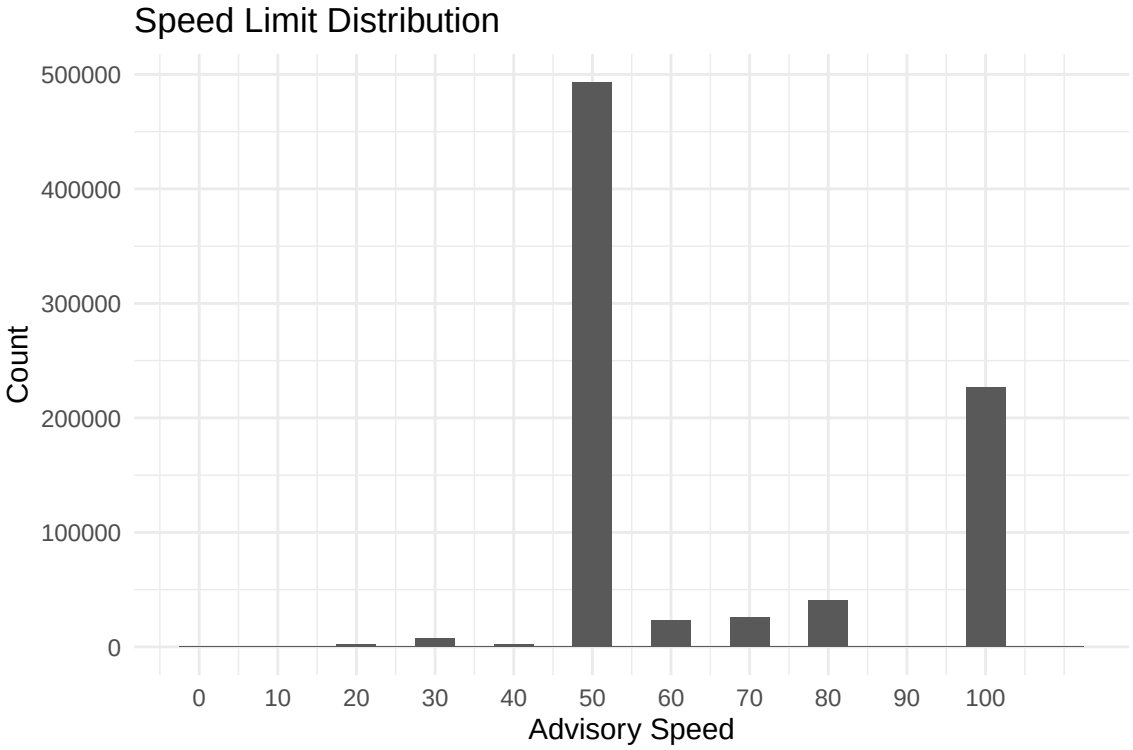
```
## `summarise()` has grouped output by 'crashSeverity'. You can override using the
## `.groups` argument.
```

Summary Tables

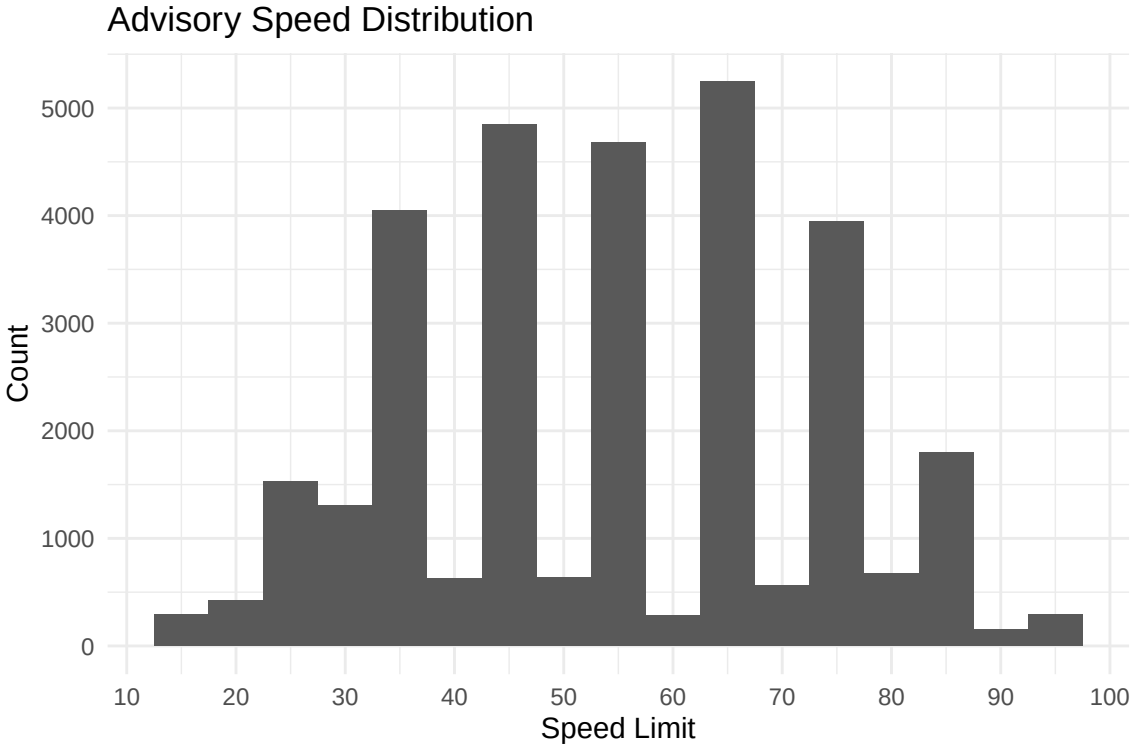
```
##                          Count Percent_Missing
## crashRoadSideRoad       821744       100.00000
## intersection            821744       100.00000
## temporarySpeedLimit     809161        98.46874
## pedestrian              795139        96.76237
## advisorySpeed           790400        96.18567
## 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
```

Plots

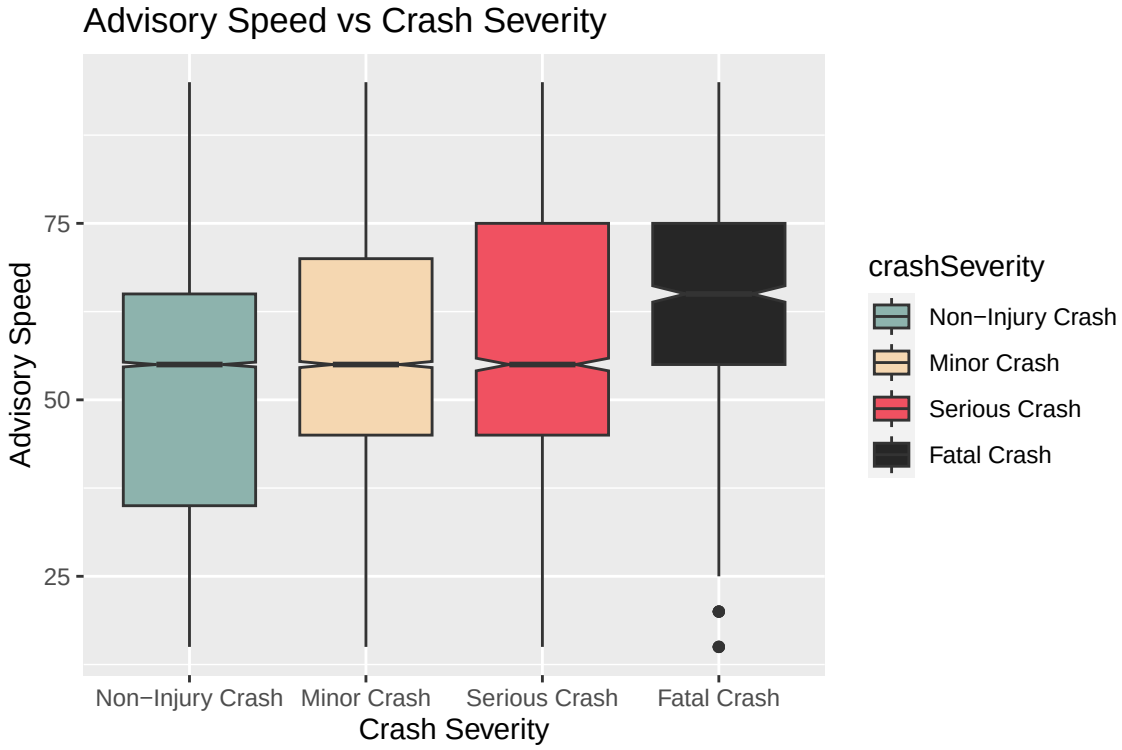
Advisory Speed Distribution



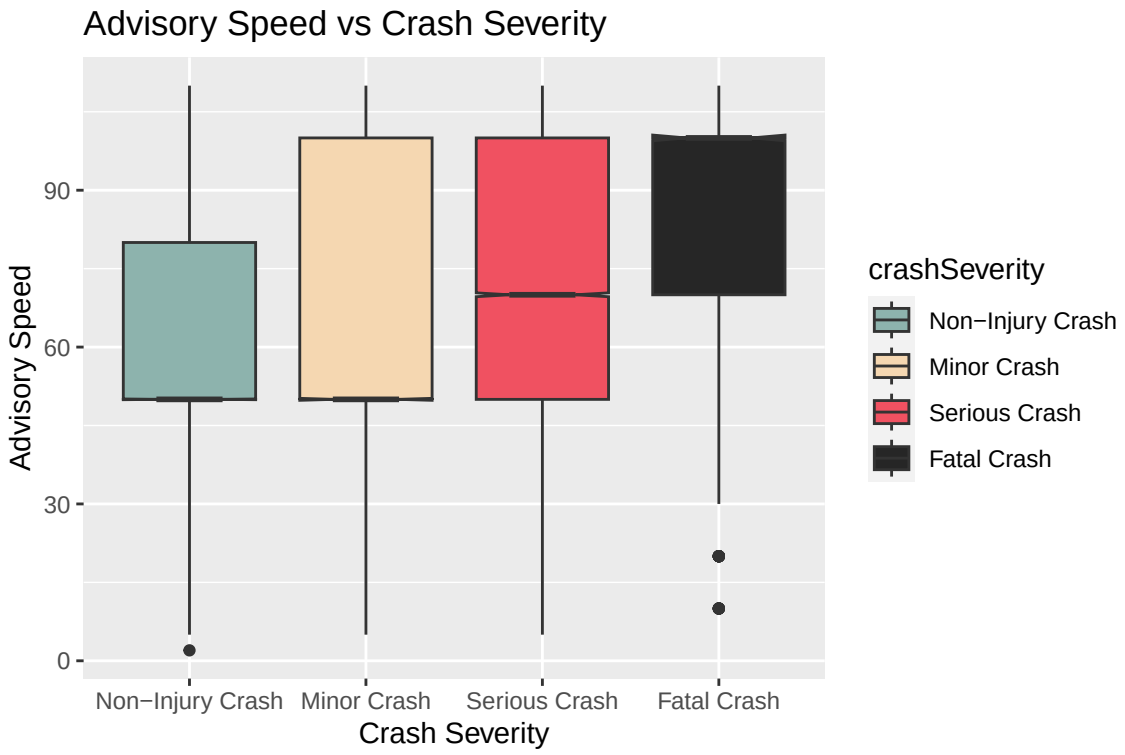
Speed Limit Distribution



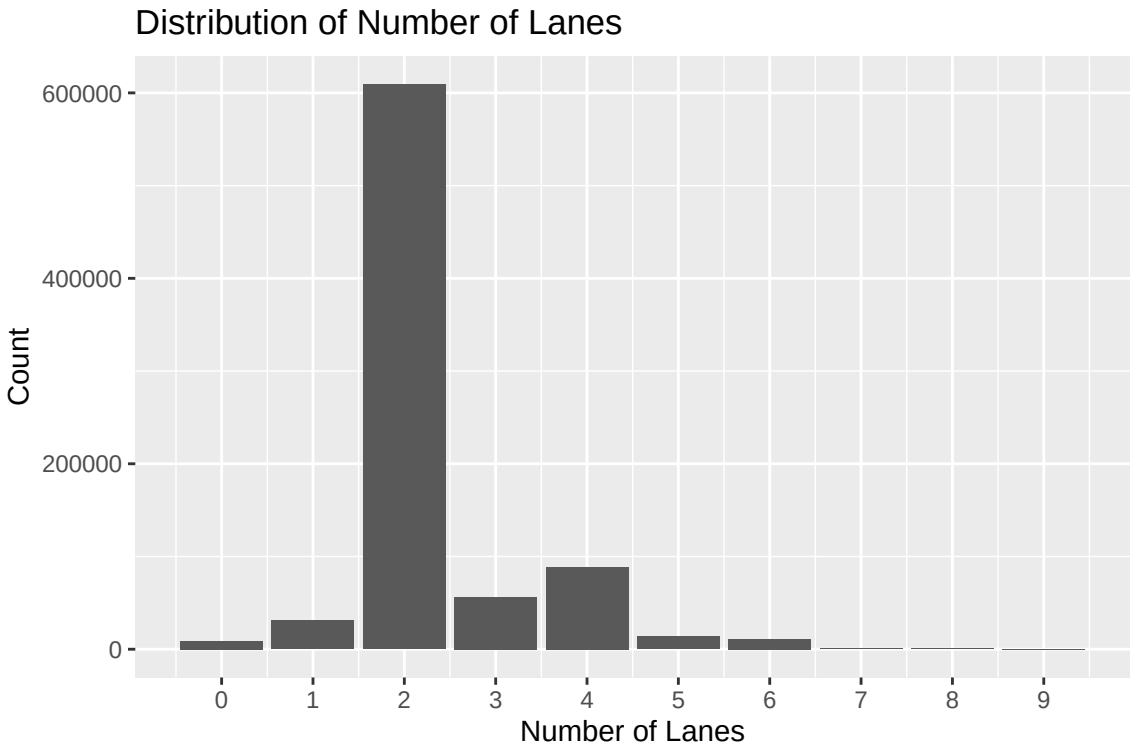
Boxplot of advisorySpeed vs crashSeverity



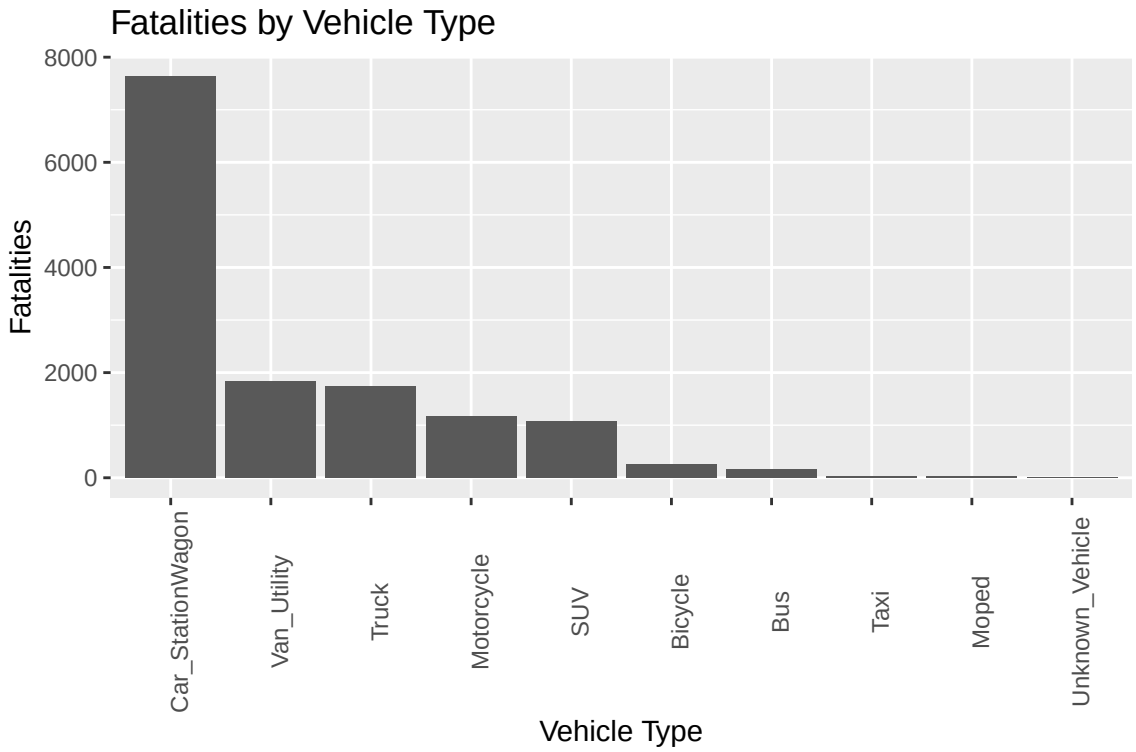

```
## Notch went outside hinges
## i Do you want `notch = FALSE`?
## Notch went outside hinges
## i Do you want `notch = FALSE`?
## Notch went outside hinges
## i Do you want `notch = FALSE`?
```



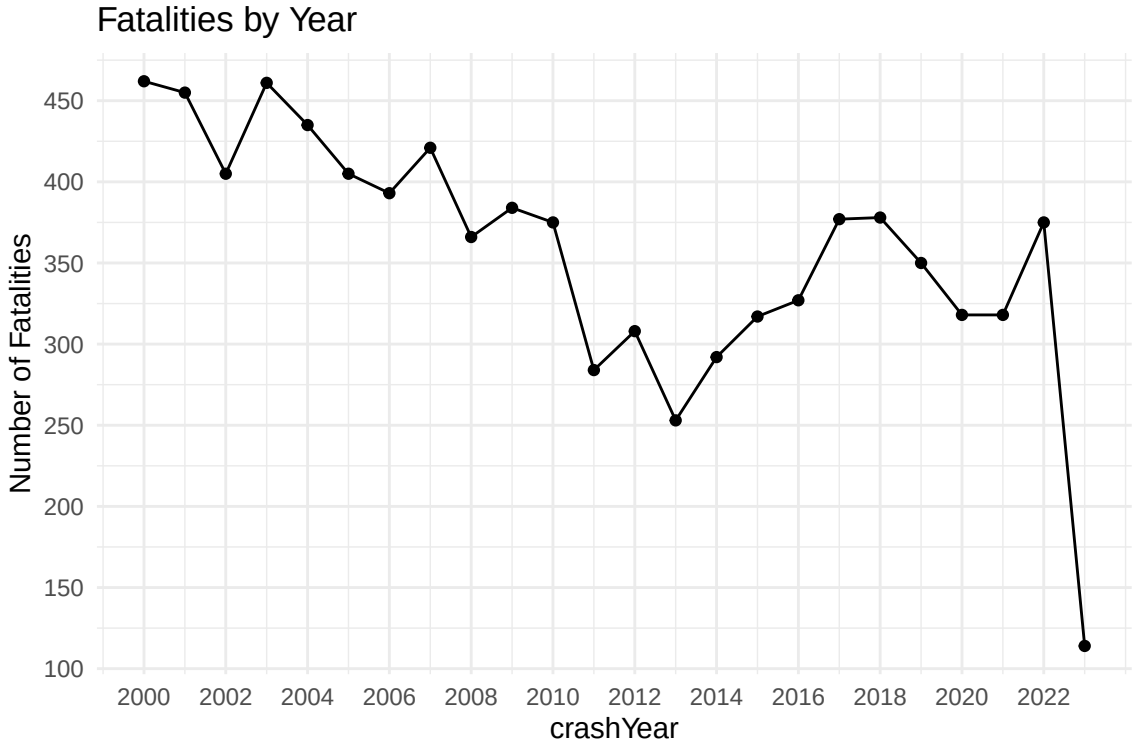
Number of lanes



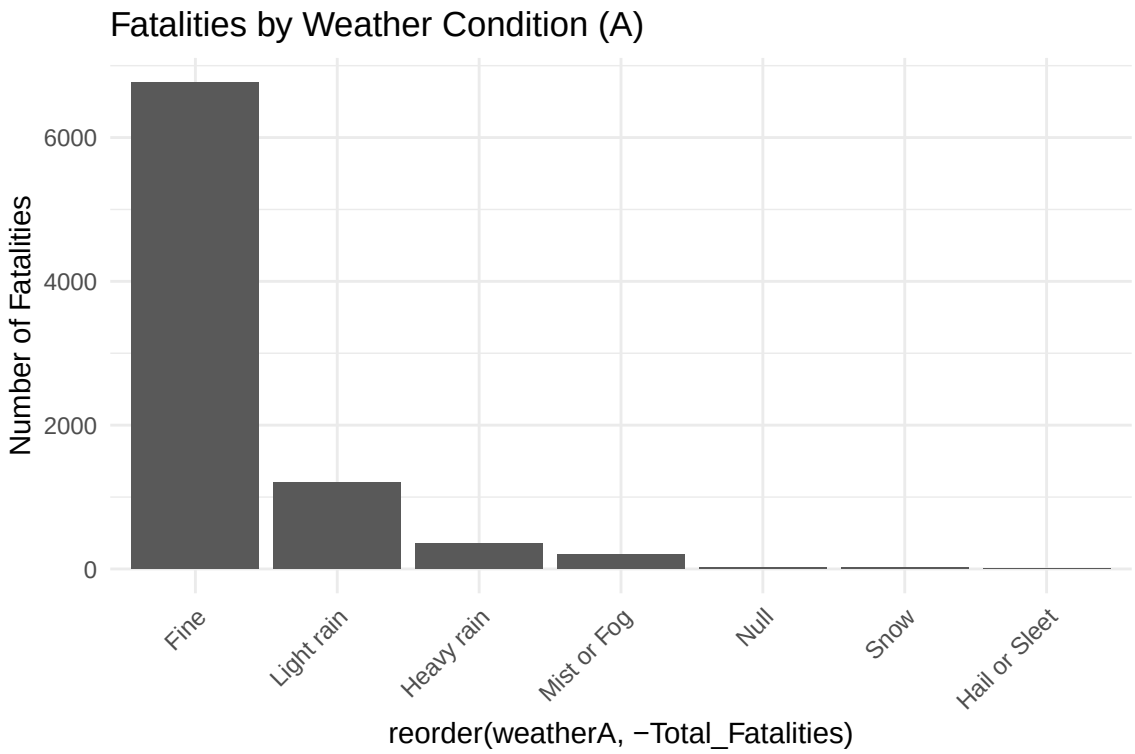
Barplot of Fatalities by Vehicle Type



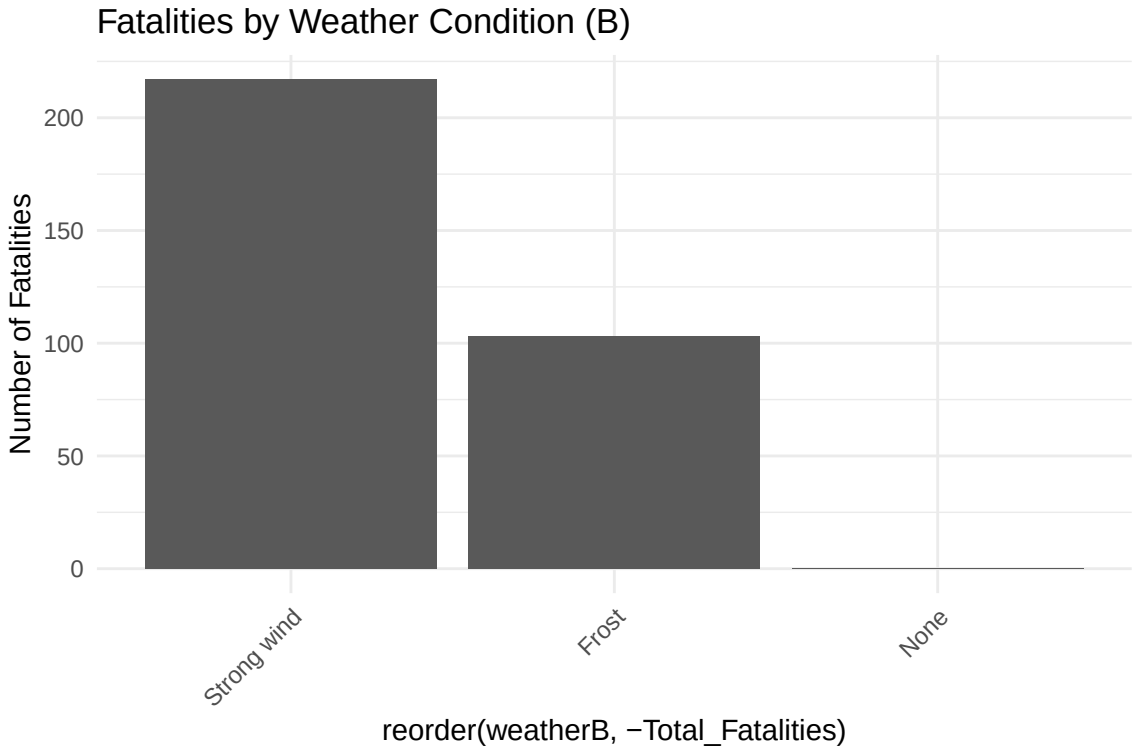
Fatalities by year



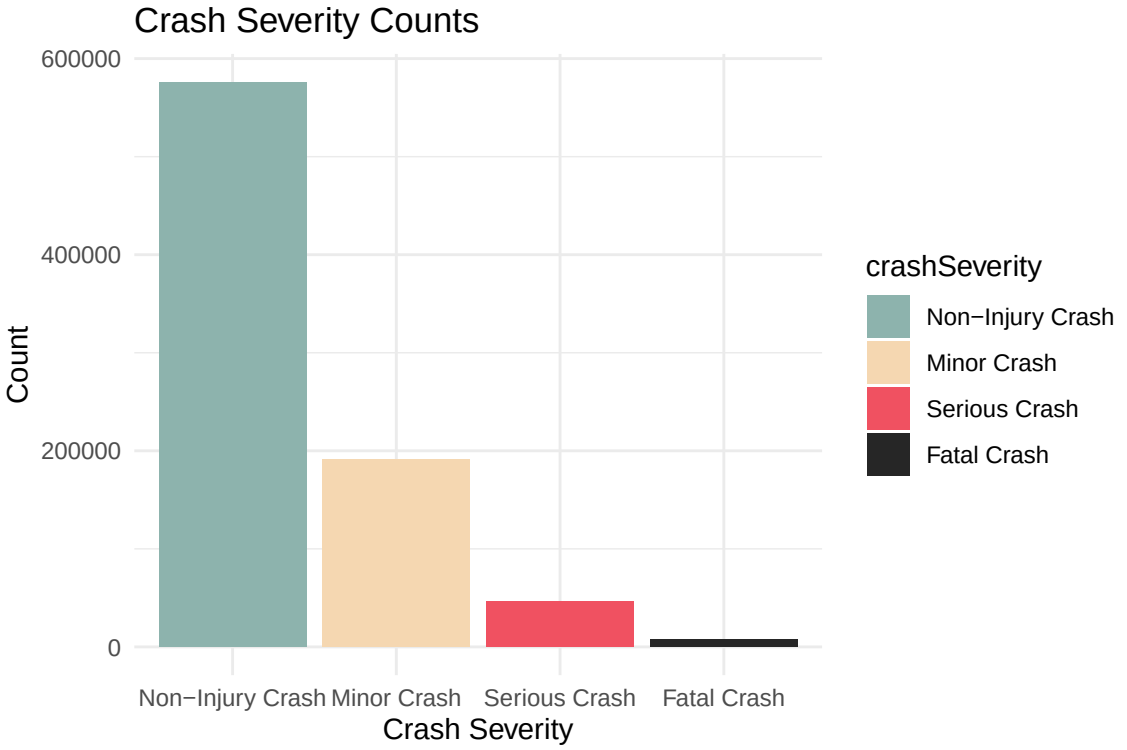
Fatalities by weather A

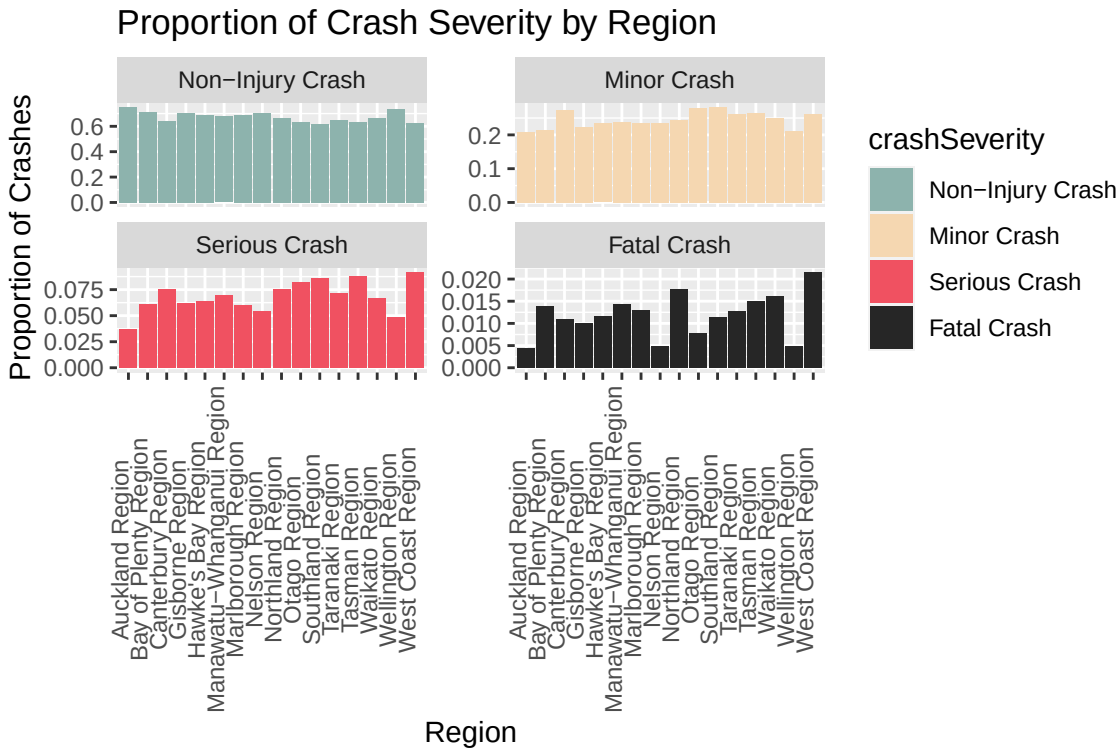


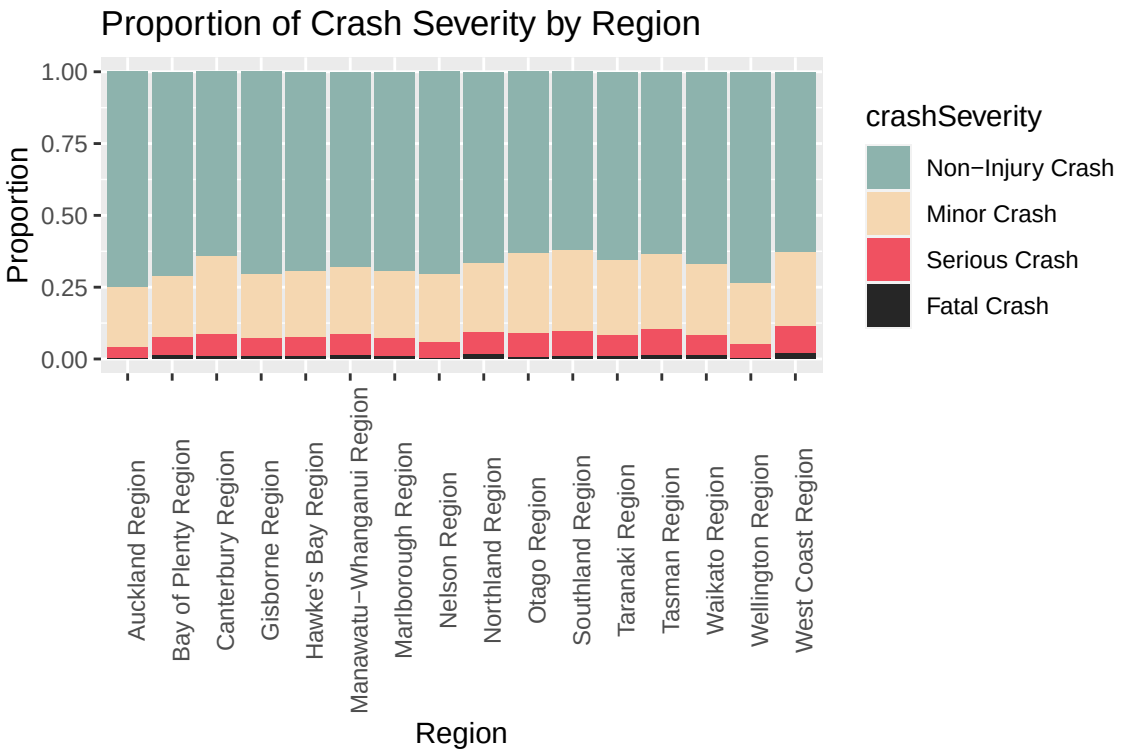
Fatalities by weather B



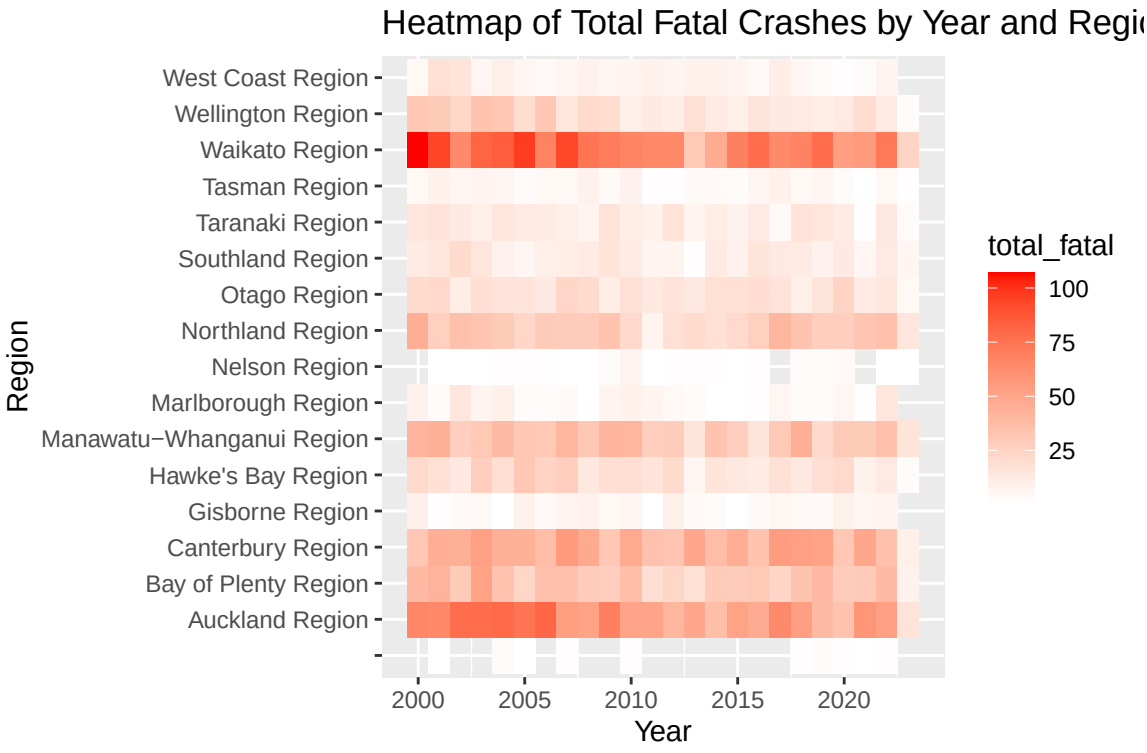
Crash Severity








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



Individual Contributions

Michael Fry

Fletcher Smith

Matthew Smaill

References