Group 12 Exploratory Data Analysis

DATA301

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

5 September 2023

Contents

```r
library(ggplot2)
library(dplyr)
library(readr)
library(pander)
```

```r
casdata <- read_csv("casdata.csv")
```

Background and Data

State which dataset(s) your group worked on, and their source

In our project, we focused our efforts on working with the Crash Analysis System (CAS) data sourced from Waka Kotahi, the New Zealand Transport Agency. The CAS data was accessed through the Open Data platform provided by Waka Kotahi, which offers a comprehensive dataset detailing road crashes that have occurred across New Zealand since the year 2000.

Explain briefly why the dataset is of interest, or what questions it could be used to answer; assume that the reader has never heard of your dataset

The significance of the Crash Analysis System (CAS) data lies in its potential to unveil critical insights into road safety, crash patterns, and their intricate connections with various socio-economic and geographic variables in New Zealand. This dataset provides a unique opportunity to explore and address pertinent questions, even for readers unfamiliar with the dataset itself. For instance, we 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?.

State the types of data in the dataset(s) and the structure of the dataset(s). Are the data numerical, categorical, or both? Time series? Coordinates? Diagnostic categories? This does NOT need to be an exhaustive list of every variable, just a few comments on the overall types.

```r
str(casdata)
```

```
## spc_tbl_ [821,744 x 70] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ X                          : num [1:821744] 2037858 1799424 1756461 1551129 1245391 ...
##  $ Y                          : num [1:821744] 5707835 5815528 5936053 5171320 4849172 ...
##  $ OBJECTID                   : num [1:821744] 1 2 3 4 5 6 7 8 9 10 ...
##  $ advisorySpeed              : num [1:821744] NA NA NA NA NA 30 NA NA NA NA ...
##  $ areaUnitID                 : num [1:821744] 544801 528900 507000 597513 611500 ...
##  $ bicycle                    : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ bridge                     : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ bus                        : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ carStationWagon            : num [1:821744] 2 2 0 2 1 1 1 1 2 2 ...
##  $ cliffBank                  : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ crashDirectionDescription: chr [1:821744] NA NA NA NA ...
##  $ crashFinancialYear         : chr [1:821744] "2011/2012" "2011/2012" "2012/2013" "2011/2012" ...
##  $ crashLocation1             : chr [1:821744] "SH 35 WAINUI" "HALL ST" "SHARON ROAD" "SPRINGSTON ROLL
##  $ crashLocation2             : chr [1:821744] "HIRINI ST" "LAKE ROAD" "RIDGE ROAD" "DYNES ROAD" ...
##  $ crashSeverity              : chr [1:821744] "Non-Injury Crash" "Non-Injury Crash" "Minor Crash" "Non
##  $ crashSHDescription         : logi [1:821744] NA NA NA NA NA NA ...
##  $ crashYear                  : num [1:821744] 2011 2012 2012 2011 2011 ...
##  $ debris                     : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ directionRoleDescription : chr [1:821744] "East" "South" "North" "South" ...
##  $ ditch                      : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ fatalCount               : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ fence                    : num [1:821744] NA NA NA 1 NA 0 0 NA NA 0 ...
##  $ flatHill                 : chr [1:821744] "Flat" "Flat" "Hill Road" "Flat" ...
##  $ guardRail                : num [1:821744] NA NA NA 0 NA 0 1 NA NA 0 ...
##  $ holiday                  : chr [1:821744] "Christmas New Year" NA NA NA ...
##  $ houseOrBuilding          : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ kerb                     : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ light                    : chr [1:821744] "Overcast" "Overcast" "Bright sun" "Bright sun" ...
##  $ meshblockId              : num [1:821744] 1387500 913000 162800 2719404 3139300 ...
##  $ minorInjuryCount         : num [1:821744] 0 0 1 0 0 0 0 0 0 0 ...
##  $ moped                    : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ motorcycle               : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ NumberOfLanes            : num [1:821744] 1 5 2 2 4 2 2 2 2 2 ...
##  $ objectThrownOrDropped    : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ otherObject              : num [1:821744] NA NA NA 1 NA 0 0 NA NA 0 ...
##  $ otherVehicleType         : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ overBank                 : num [1:821744] NA NA NA 0 NA 1 0 NA NA 0 ...
##  $ parkedVehicle            : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ pedestrian               : num [1:821744] NA NA 1 NA NA NA NA NA NA NA ...
##  $ phoneBoxEtc              : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ postOrPole               : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ region                   : chr [1:821744] "Gisborne Region" "Waikato Region" "Auckland Region" "Ca
##  $ roadCharacter            : chr [1:821744] "Nil" "Nil" "Nil" "Nil" ...
##  $ roadLane                 : chr [1:821744] "2-way" "2-way" "2-way" "2-way" ...
##  $ roadSurface              : chr [1:821744] "Sealed" "Sealed" "Sealed" "Sealed" ...
##  $ roadworks                : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ schoolBus                : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ seriousInjuryCount       : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ slipOrFlood              : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ speedLimit               : num [1:821744] 50 50 50 100 50 100 80 50 80 50 ...
##  $ strayAnimal              : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ streetLight              : chr [1:821744] "Unknown" "Unknown" "Off" "Unknown" ...
##  $ suv                      : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ taxi                     : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ temporarySpeedLimit      : num [1:821744] NA NA NA NA NA NA NA NA NA NA ...
##  $ tlaId                    : num [1:821744] 28 16 76 62 75 72 46 76 16 76 ...
##  $ tlaName                  : chr [1:821744] "Gisborne District" "Hamilton City" "Auckland" "Selwyn D
##  $ trafficControl           : chr [1:821744] "Give way" "Traffic Signals" "Nil" "Nil" ...
##  $ trafficIsland            : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ trafficSign              : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ train                    : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ tree                     : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ truck                    : num [1:821744] 0 0 0 0 0 0 0 1 0 0 ...
##  $ unknownVehicleType       : num [1:821744] 0 0 0 0 0 0 0 0 0 0 ...
##  $ urban                    : chr [1:821744] "Urban" "Urban" "Urban" "Open" ...
##  $ vanOrUtility             : num [1:821744] 0 0 1 0 1 0 0 0 0 0 ...
##  $ vehicle                  : num [1:821744] NA NA NA 0 NA 0 0 NA NA 1 ...
##  $ waterRiver               : num [1:821744] NA NA NA 0 NA 0 0 NA NA 0 ...
##  $ weatherA                 : chr [1:821744] "Fine" "Fine" "Fine" "Fine" ...
##  $ weatherB                 : chr [1:821744] "Unknown" "Unknown" "Unknown" "Unknown" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   X = col_double(),
##   ..   Y = col_double(),
##   ..   OBJECTID = col_double(),
##   ..   advisorySpeed = col_double(),
##   ..   areaUnitID = col_double(),
##   ..   bicycle = col_double(),
##   ..   bridge = col_double(),
##   ..   bus = col_double(),
##   ..   carStationWagon = col_double(),
##   ..   cliffBank = col_double(),
##   ..   crashDirectionDescription = col_character(),
##   ..   crashFinancialYear = col_character(),
##   ..   crashLocation1 = col_character(),
##   ..   crashLocation2 = col_character(),
##   ..   crashSeverity = col_character(),
##   ..   crashSHDescription = col_logical(),
##   ..   crashYear = col_double(),
##   ..   debris = col_double(),
##   ..   directionRoleDescription = col_character(),
##   ..   ditch = col_double(),
##   ..   fatalCount = col_double(),
##   ..   fence = col_double(),
##   ..   flatHill = col_character(),
##   ..   guardRail = col_double(),
##   ..   holiday = col_character(),
##   ..   houseOrBuilding = col_double(),
##   ..   kerb = col_double(),
##   ..   light = col_character(),
##   ..   meshblockId = col_double(),
##   ..   minorInjuryCount = col_double(),
##   ..   moped = col_double(),
##   ..   motorcycle = col_double(),
##   ..   NumberOfLanes = col_double(),
##   ..   objectThrownOrDropped = col_double(),
##   ..   otherObject = col_double(),
##   ..   otherVehicleType = col_double(),
##   ..   overBank = col_double(),
##   ..   parkedVehicle = col_double(),
##   ..   pedestrian = col_double(),
##   ..   phoneBoxEtc = col_double(),
##   ..   postOrPole = col_double(),
##   ..   region = col_character(),
##   ..   roadCharacter = col_character(),
##   ..   roadLane = col_character(),
##   ..   roadSurface = col_character(),
##   ..   roadworks = col_double(),
##   ..   schoolBus = col_double(),
##   ..   seriousInjuryCount = col_double(),
##   ..   slipOrFlood = col_double(),
##   ..   speedLimit = col_double(),
##   ..   strayAnimal = col_double(),
##   ..   streetLight = col_character(),
##   ..   suv = col_double(),
##   ..   taxi = col_double(),
##   ..   temporarySpeedLimit = col_double(),
##   ..   tlaId = col_double(),
##   ..   tlaName = col_character(),
##   ..   trafficControl = col_character(),
##   ..   trafficIsland = col_double(),
##   ..   trafficSign = col_double(),
##   ..   train = col_double(),
##   ..   tree = col_double(),
##   ..   truck = col_double(),
##   ..   unknownVehicleType = col_double(),
##   ..   urban = col_character(),
##   ..   vanOrUtility = col_double(),
##   ..   vehicle = col_double(),
##   ..   waterRiver = col_double(),
##   ..   weatherA = col_character(),
##   ..   weatherB = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

The Crash Analysis System (CAS) data is characterized by its numerical and categorical data types. Numerical attributes encompass various quantitative measures, including crash severity, the count of vehicles involved, and distances. On the categorical side, the dataset features descriptors such as crash types, road conditions, and vehicle types. Furthermore, the geographic aspect is represented through crash location one and crash location two, aswell as the region where the crash took place.

State how complete the dataset(s) are (i.e. how many missing, any structure to the missing data, whether there are errors in the data)

```
missing_counts <- colSums(is.na(casdata))
missing_counts
```

```
##                         X                         Y                  OBJECTID
##                         0                         0                         0
##             advisorySpeed                areaUnitID                   bicycle
##                    790400                        97                         5
##                    bridge                       bus           carStationWagon
##                    488831                         5                         5
##                 cliffBank crashDirectionDescription        crashFinancialYear
##                    488831                    309268                         0
##            crashLocation1            crashLocation2             crashSeverity
##                         0                      1273                         0
##        crashSHDescription                 crashYear                    debris
##                    821744                         0                    488831
##  directionRoleDescription                     ditch                fatalCount
##                        72                    488831                         1
##                     fence                  flatHill                 guardRail
##                    488831                      6317                    488831
##                   holiday           houseOrBuilding                      kerb
##                    776922                    488831                    488831
##                     light               meshblockId          minorInjuryCount
##                         0                        97                         1
##                     moped                motorcycle             NumberOfLanes
##                         5                         5                      1813
##     objectThrownOrDropped               otherObject          otherVehicleType
##                    488831                    488831                         5
##                  overBank             parkedVehicle                pedestrian
##                    488831                    488831                    795139
##               phoneBoxEtc                postOrPole                    region
##                    488831                    488831                      3188
##             roadCharacter                  roadLane               roadSurface
##                         0                         0                       977
##                 roadworks                 schoolBus        seriousInjuryCount
##                    488831                         5                         1
##               slipOrFlood                speedLimit               strayAnimal
##                    488831                       838                    488831
##               streetLight                       suv                      taxi
##                         0                         5                         5
##       temporarySpeedLimit                     tlaId                   tlaName
##                    809161                      3188                      3188
##            trafficControl             trafficIsland               trafficSign
##                         0                    488831                    488831
##                     train                      tree                     truck
##                    488831                    488831                         5
##        unknownVehicleType                     urban              vanOrUtility
##                         5                         0                         5
##                   vehicle                waterRiver                  weatherA
##                    488831                    488831                         0
##                  weatherB
##                         0
```

Data completeness within the Crash Analysis System (CAS) data, we discern a variable degree of missing information across the dataset's attributes. 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.

Ethics, Privacy and Security

Brief discussion of any ethical considerations that apply to your project

Ethical considerations hold a paramount role in our project as we delve into analyzing crash data from Waka Kotahi's Crash Analysis System (CAS). One prominent ethical concern 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 safety improvements is ethically crucial. Acknowledging the broader implications of our findings, particularly when suggesting insights, ensures that our project contributes positively to road safety without causing undue distress to those affected by road accidents.

Brief discussion of any privacy concerns that might arise connected to your project

Privacy concerns are an inherent consideration when dealing with datasets that involve personal information, such as the Crash Analysis System (CAS) data. CAS data contains information about individuals involved in accidents, and demographic datasets might include personal identifiers. Therefore, our project must adhere to stringent privacy regulations to safeguard individual privacy rights. One of the major concerns is ensuring that any data shared or published is anonymized to prevent the identification of specific individuals or vehicles. This entails removing or aggregating any identifying attributes. By diligently addressing these concerns, we can prevent potential breaches of privacy that could arise from the publication of sensitive data.

Brief discussion of what steps you could take to keep your project data and results secure (you do NOT need to carry this out, you just need to talk about it in the report)

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. First, data encryption protocols can be implemented to safeguard data during storage and transmission. Encryption ensures that unauthorized access to sensitive information is significantly mitigated. Moreover, restricting access to the data is pivotal. Utilizing secure authentication and authorization mechanisms ensures that only authorized personnel can access the project data and results. Regular data backups would help prevent data loss due to unforeseen circumstances. Storing backups securely, ideally in an off-site location, is a best practice. Lastly, utilizing secure data sharing methods, such as secure file sharing platforms or virtual private networks (VPNs), can facilitate collaboration while maintaining data security. By integrating these security measures, we can uphold the integrity and confidentiality of our project data and outcomes.

Exporatory Data Analysis

```
selected_data <- casdata[, c("crashSeverity", "weatherA", "region")]
summary(selected_data)
```

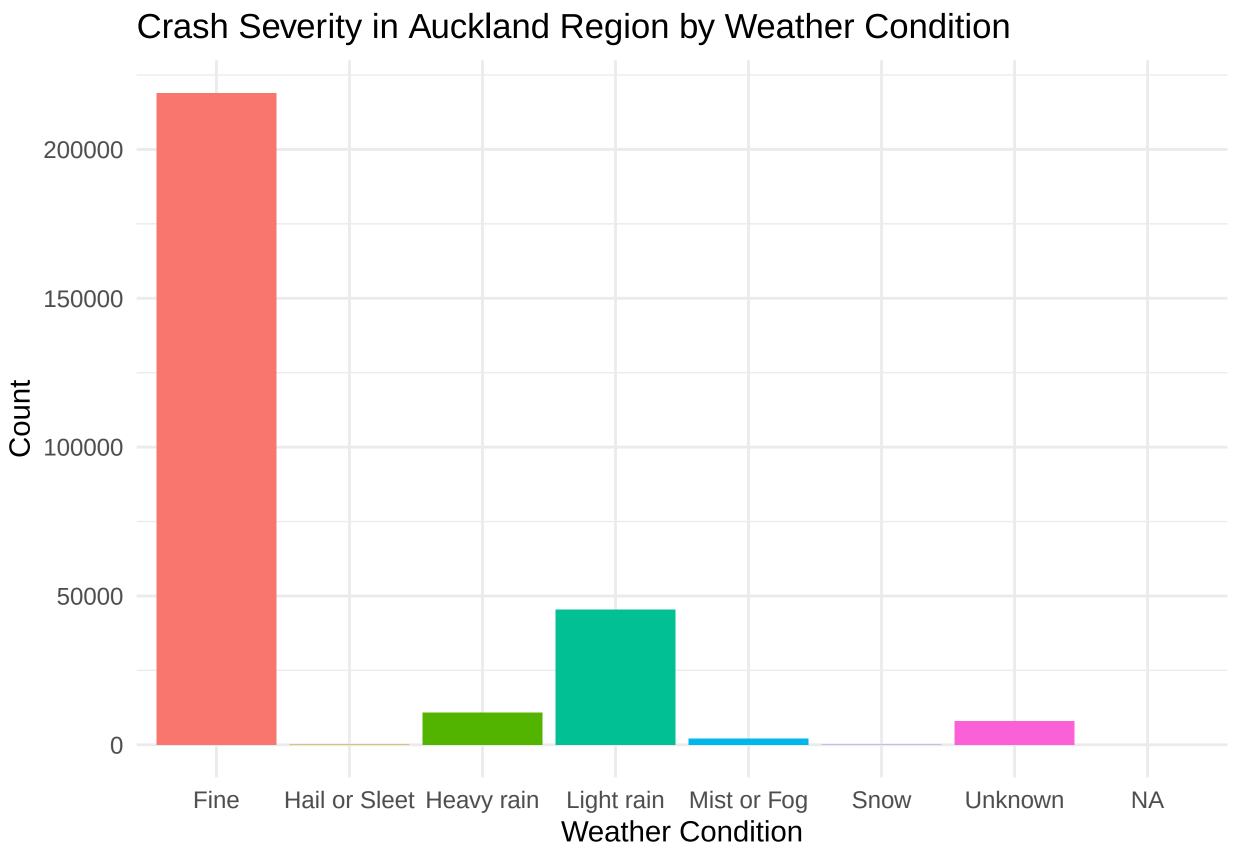
```
##  crashSeverity        weatherA            region         
##  Length:821744      Length:821744      Length:821744     
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
```

```
grouped_data <- selected_data %>%
  group_by(crashSeverity, weatherA, region) %>%
  summarise(count = n()) %>%
  arrange(crashSeverity, weatherA, region)
```

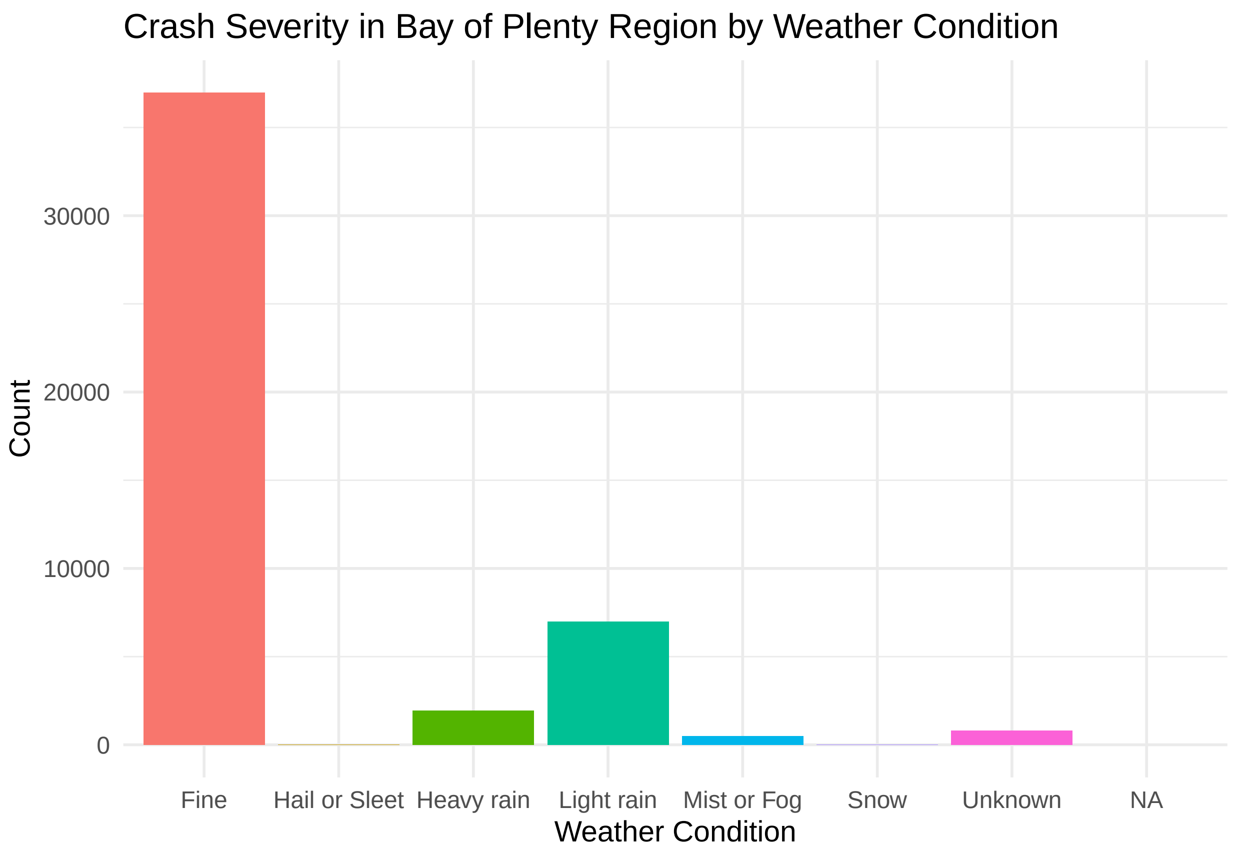
```
## ‘summarise()‘ has grouped output by ’crashSeverity’, ’weatherA’. You can
## override using the ‘.groups‘ argument.
```

```
regions <- unique(grouped_data$region)

plots <- list()

for (region in regions) {
  region_data <- grouped_data[grouped_data$region == region, ]
  plot <- ggplot(region_data, aes(x = weatherA, y = count, fill = weatherA)) +
    geom_bar(stat = "identity") +
    labs(x = "Weather Condition", y = "Count", fill = "Weather Condition", title = paste("Crash Severit
    theme_minimal() +
    theme(legend.position = "none")

  plots[[region]] <- plot
}

for (region in regions) {
  print(plots[[region]])
}
```

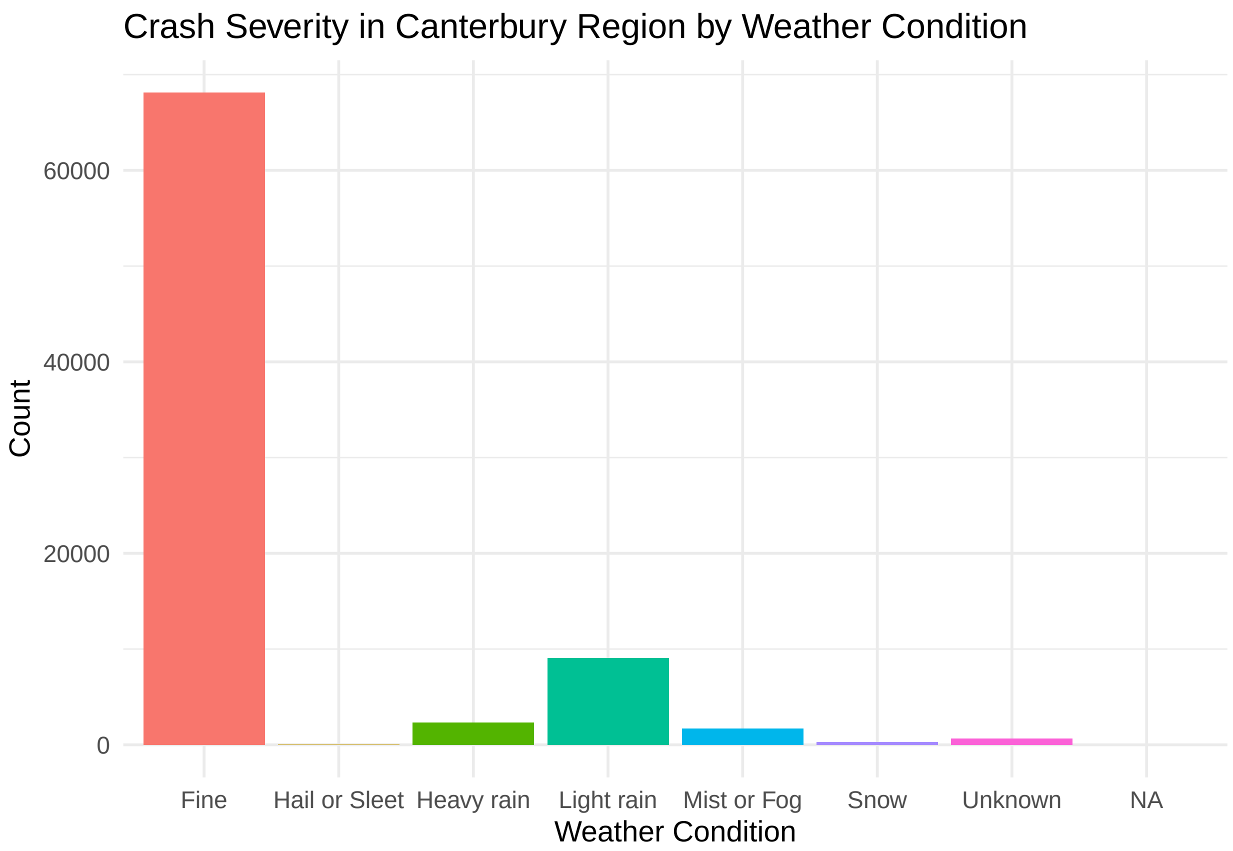
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



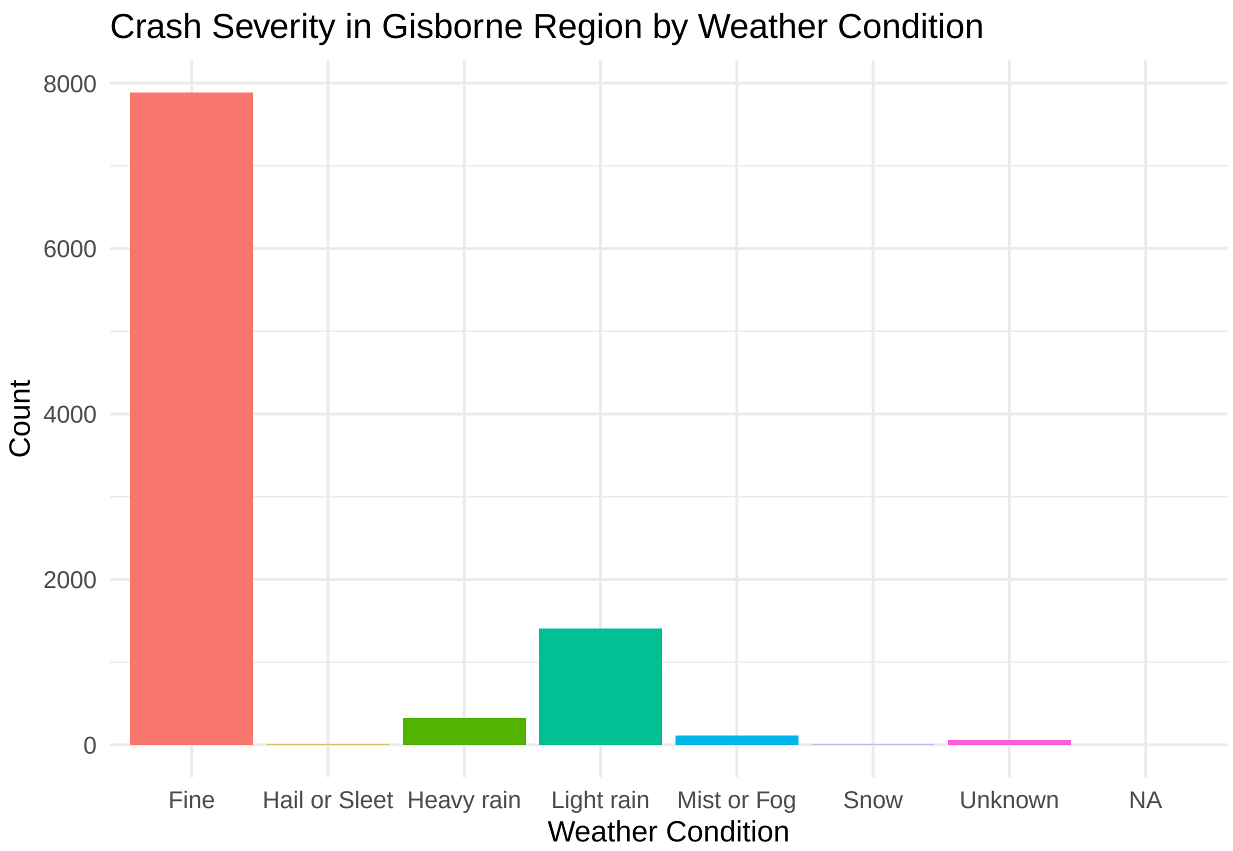
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



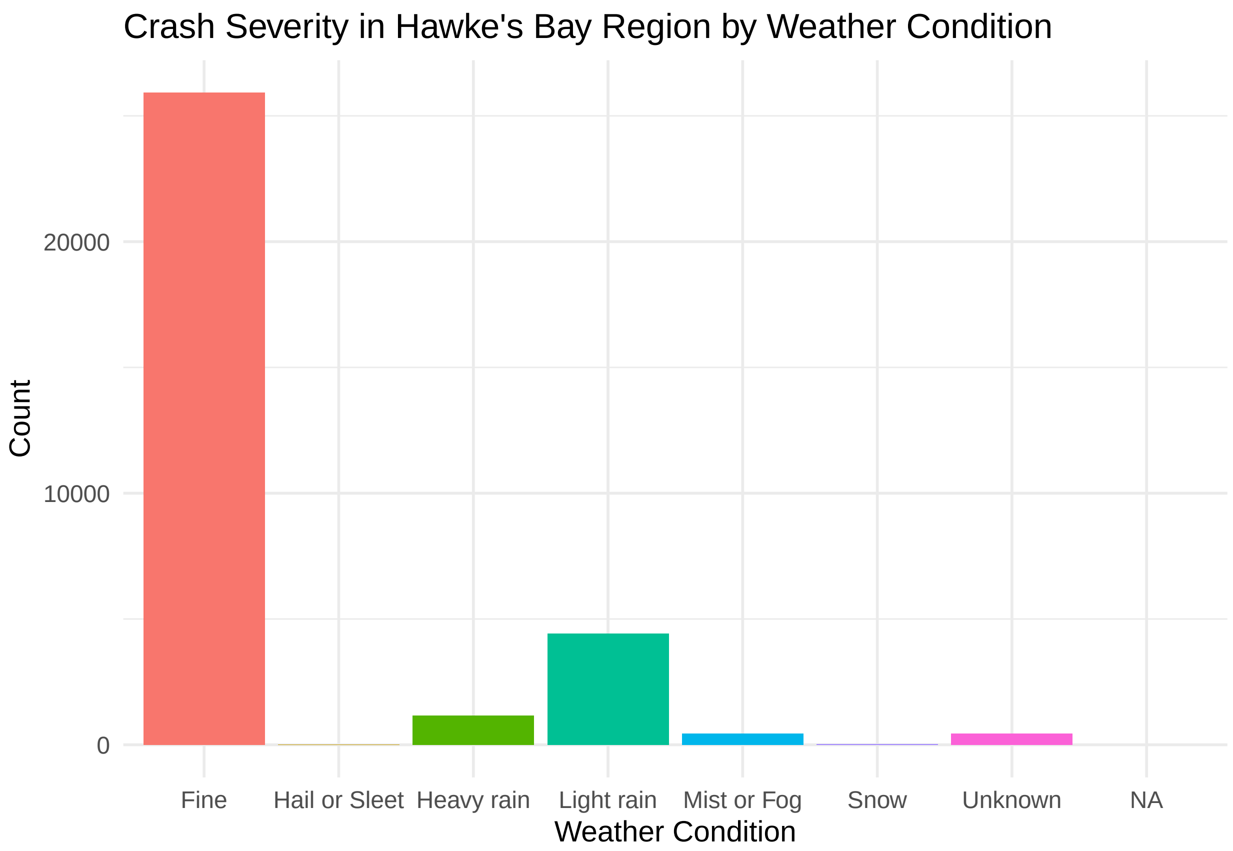
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



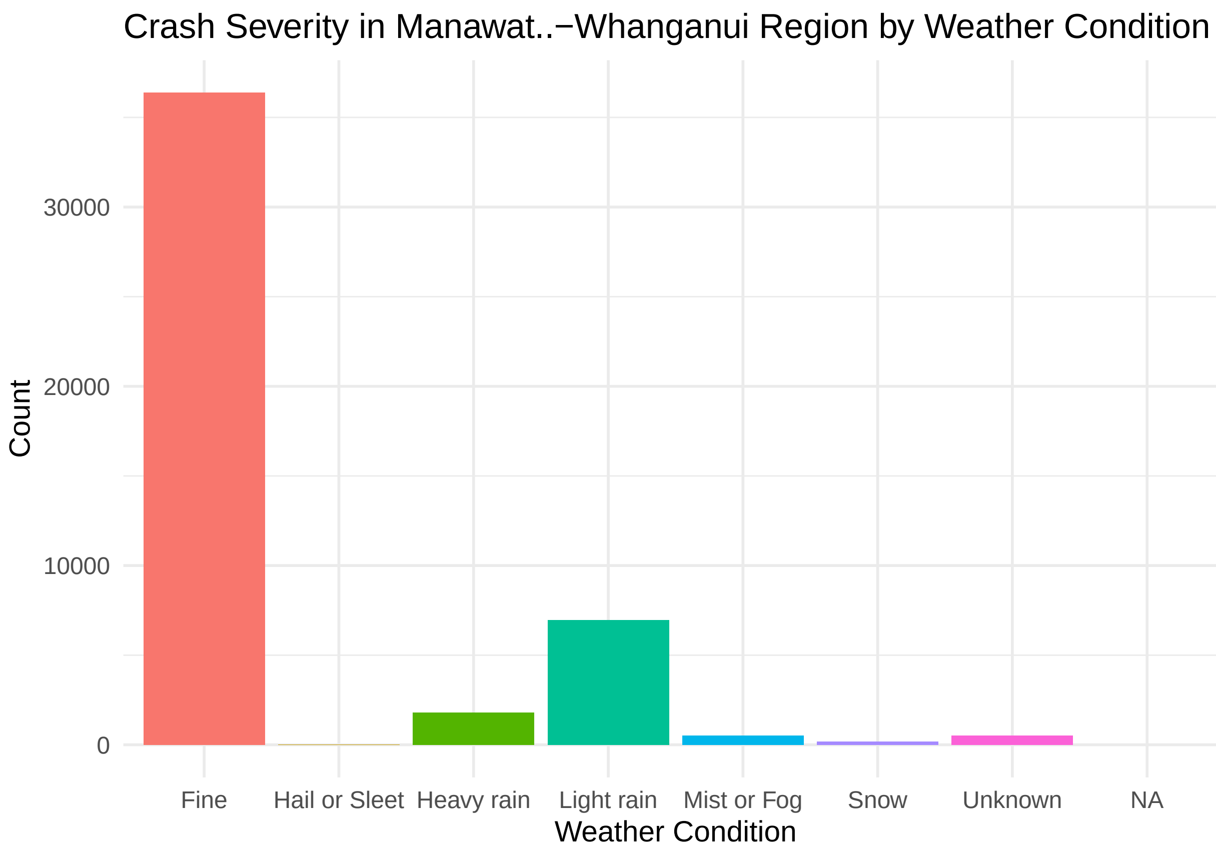
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



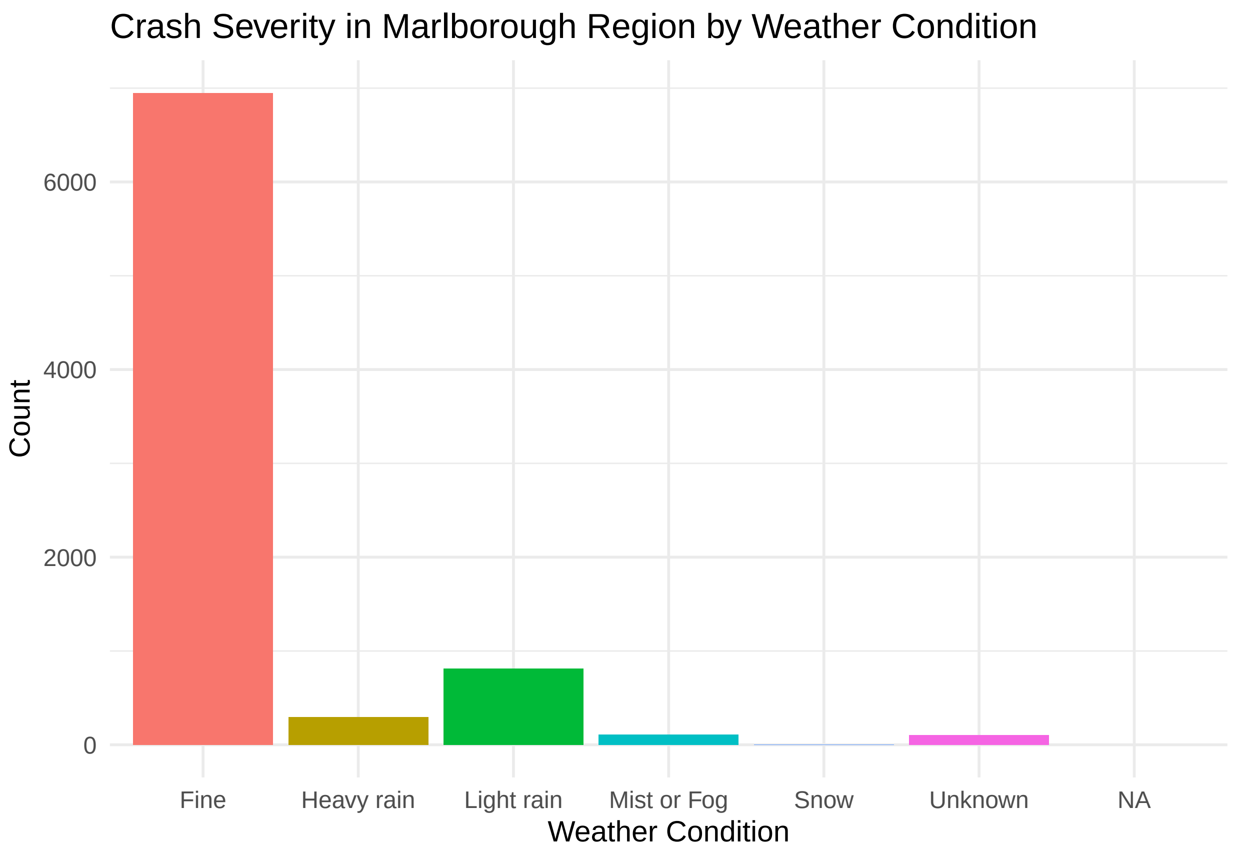
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



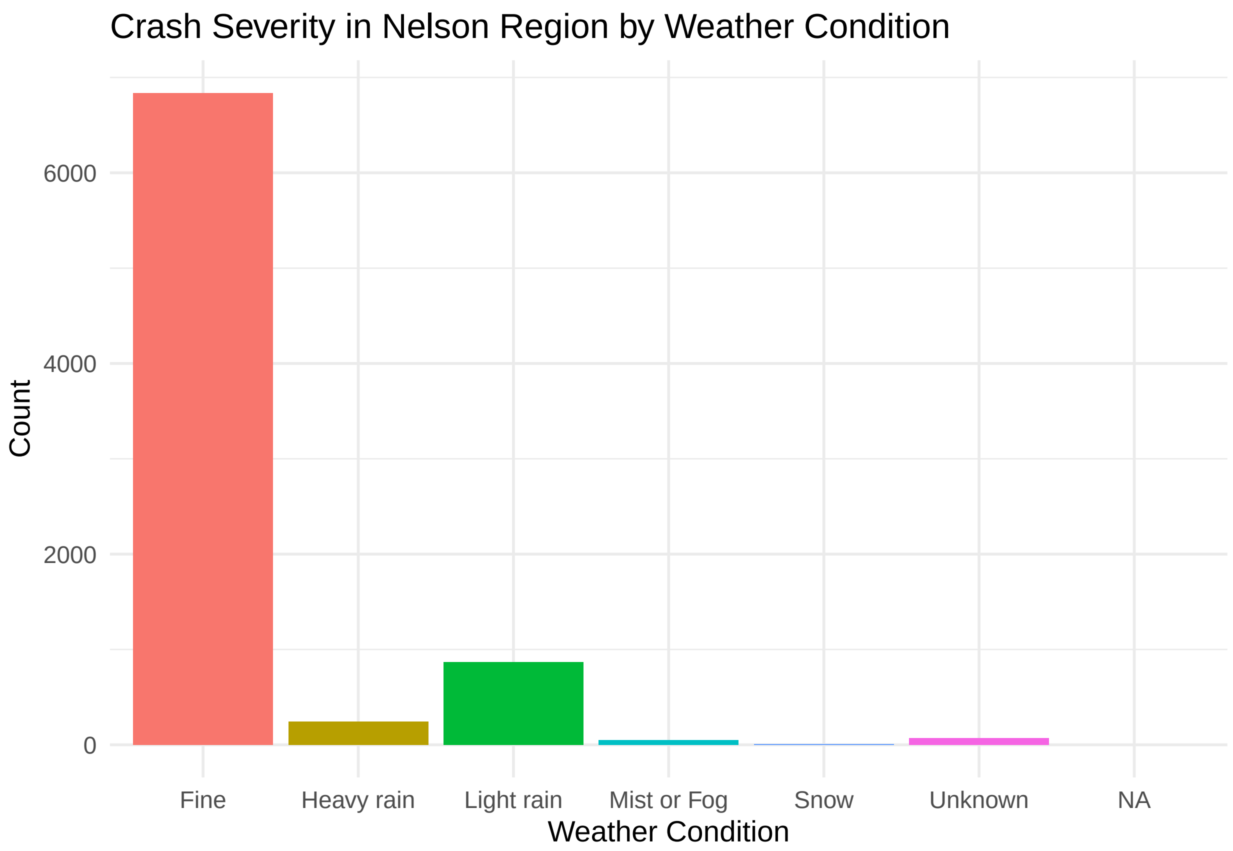
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <ab>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <ab>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <ab>
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <ab>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <ab>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <ab>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <ab>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <c5>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Crash Severity in Manawatū-Whanganui Region by Weather
## Condition' in 'mbcsToSbcs': dot substituted for <ab>
```



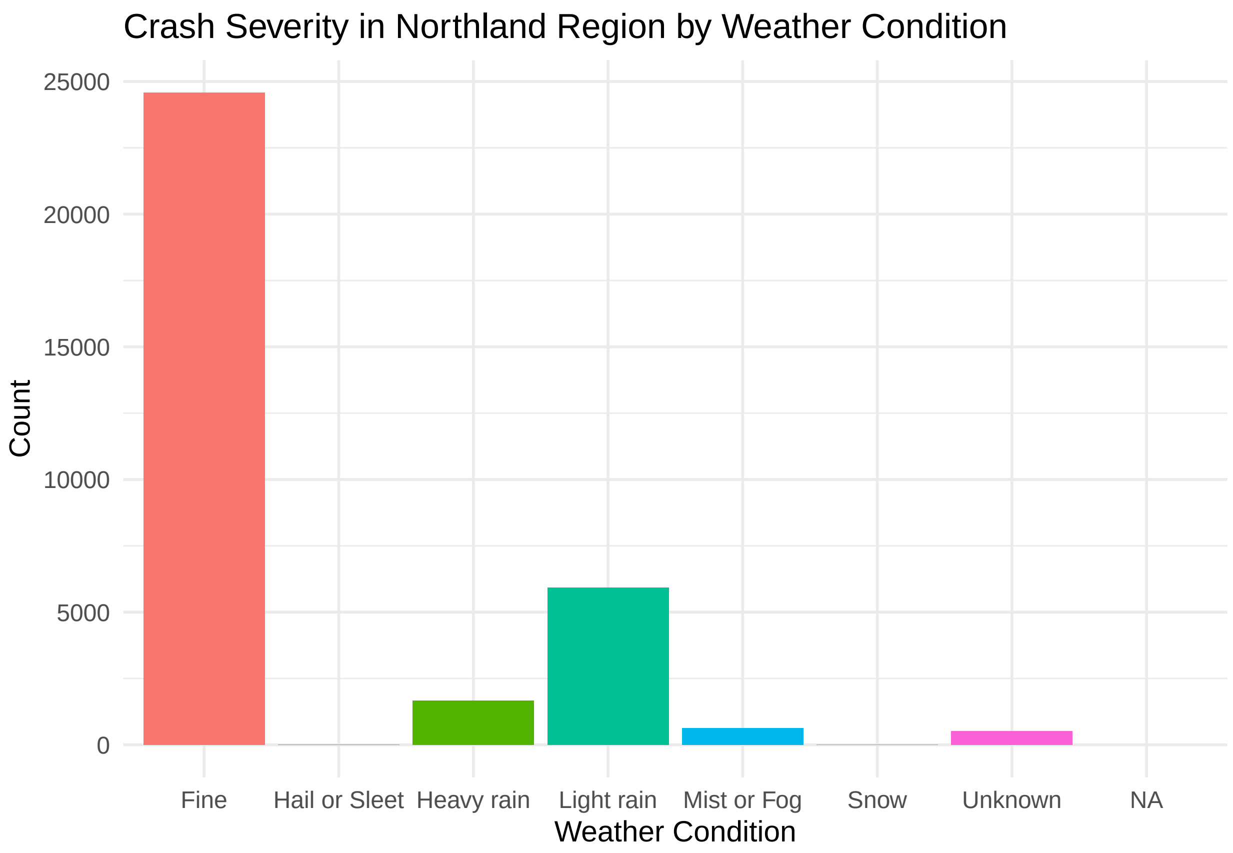
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



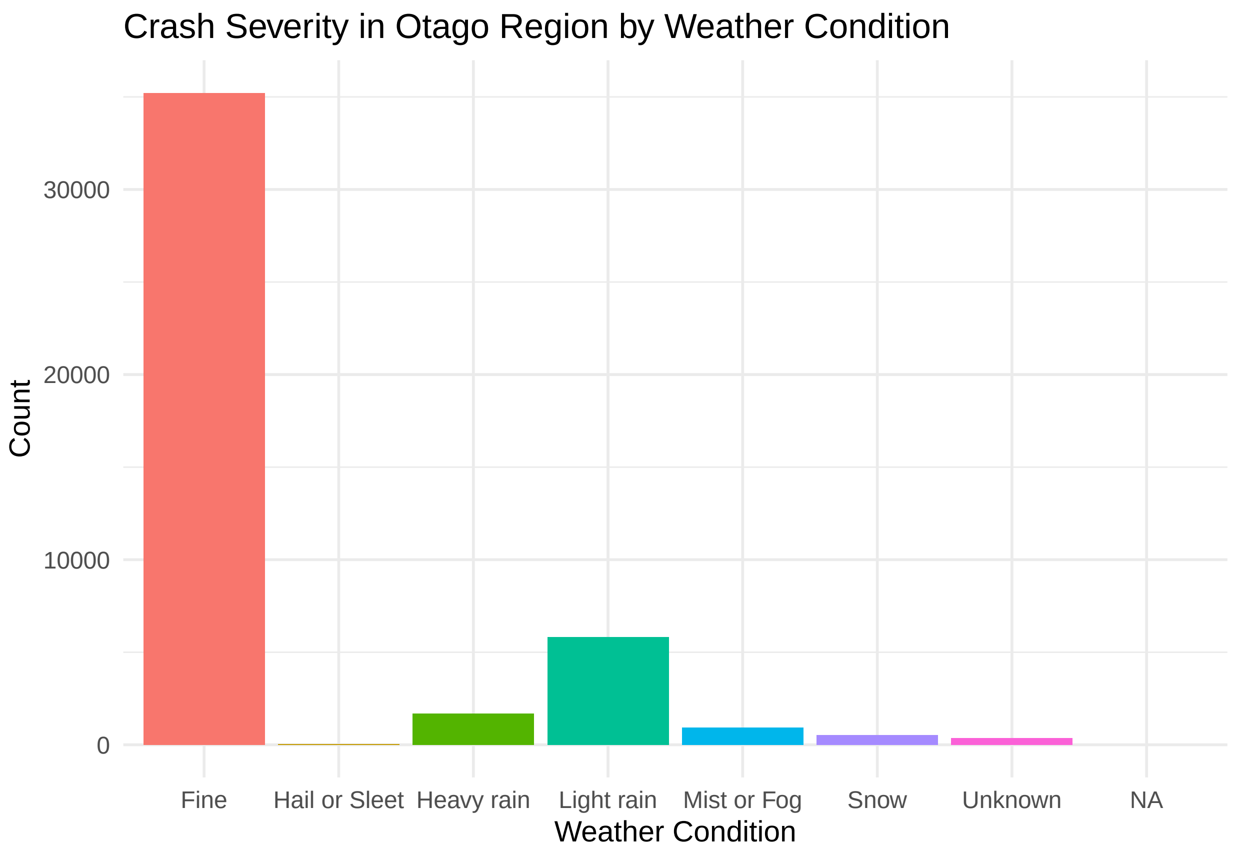
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



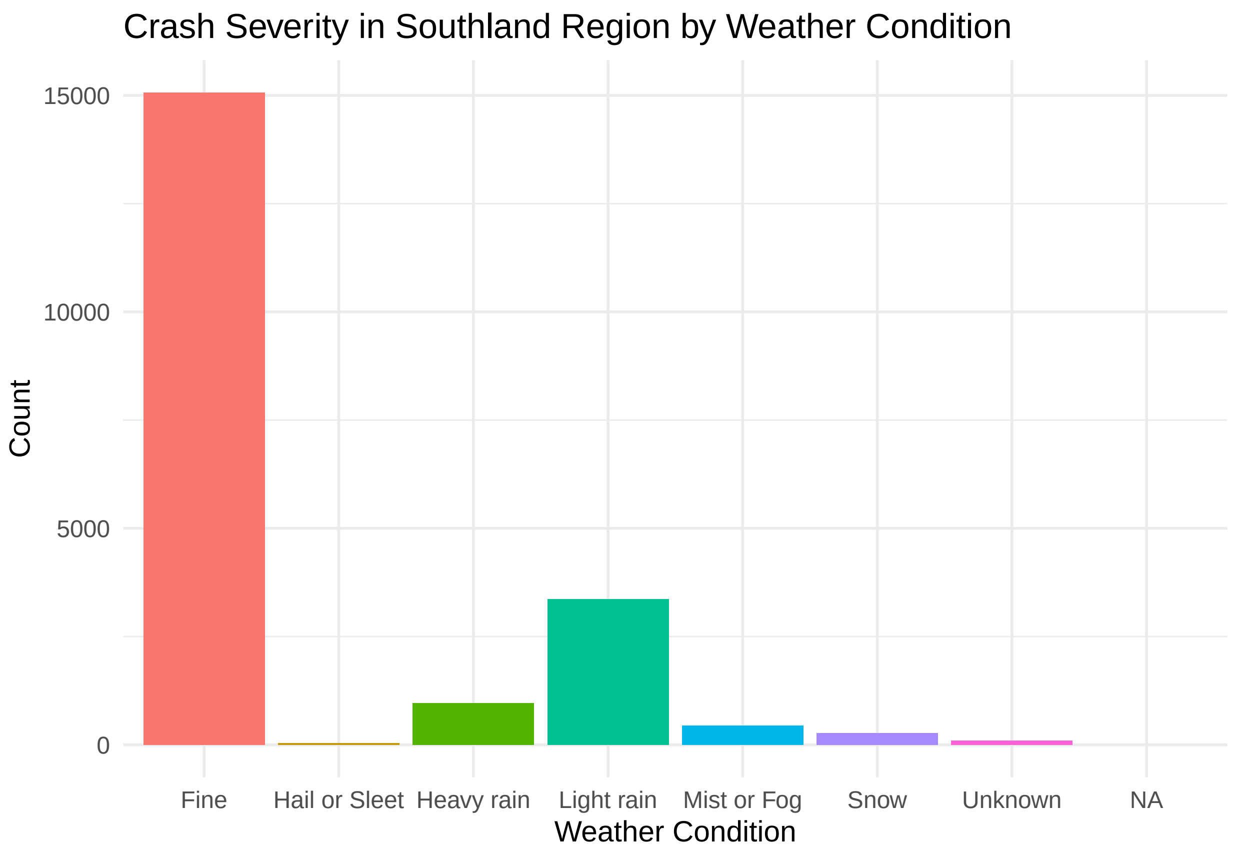
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



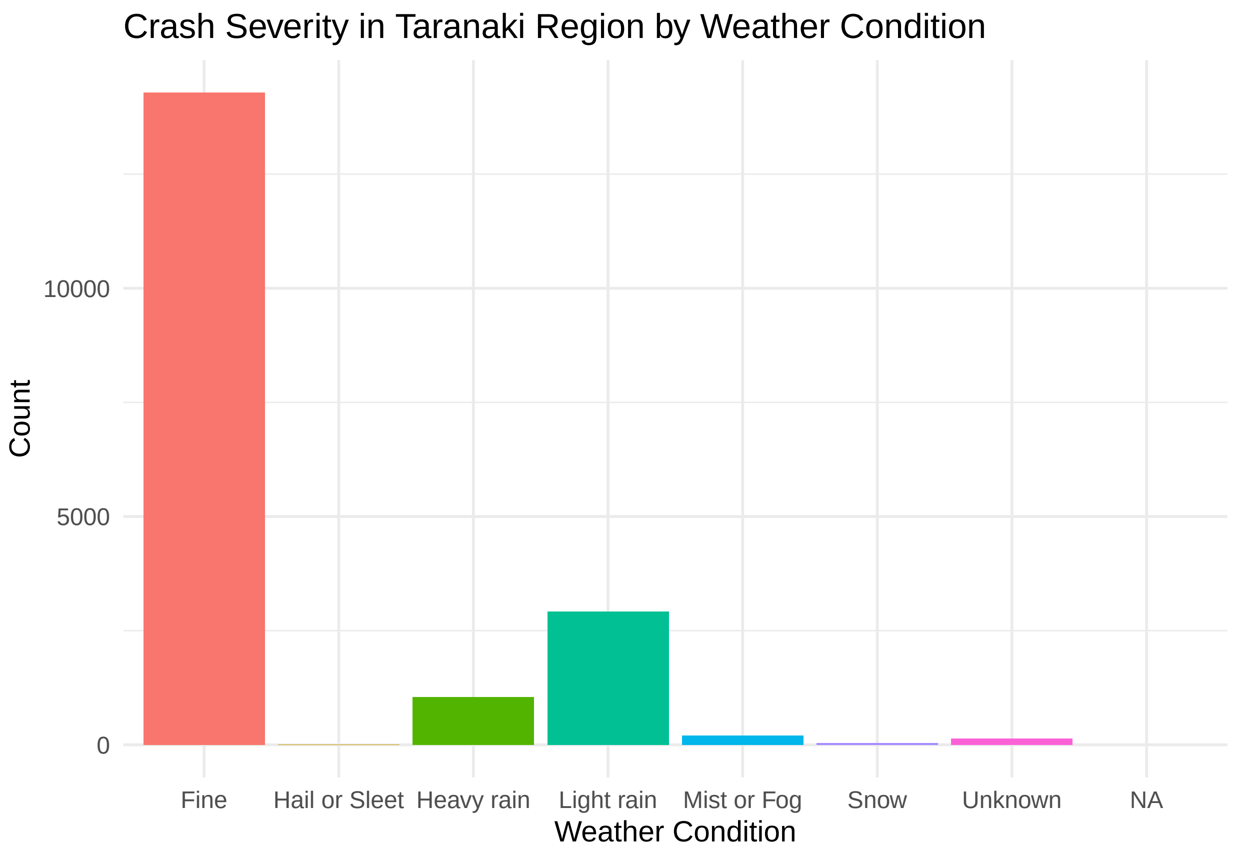
```
## Warning: Removed 18 rows containing missing values ('position_stack()').
```



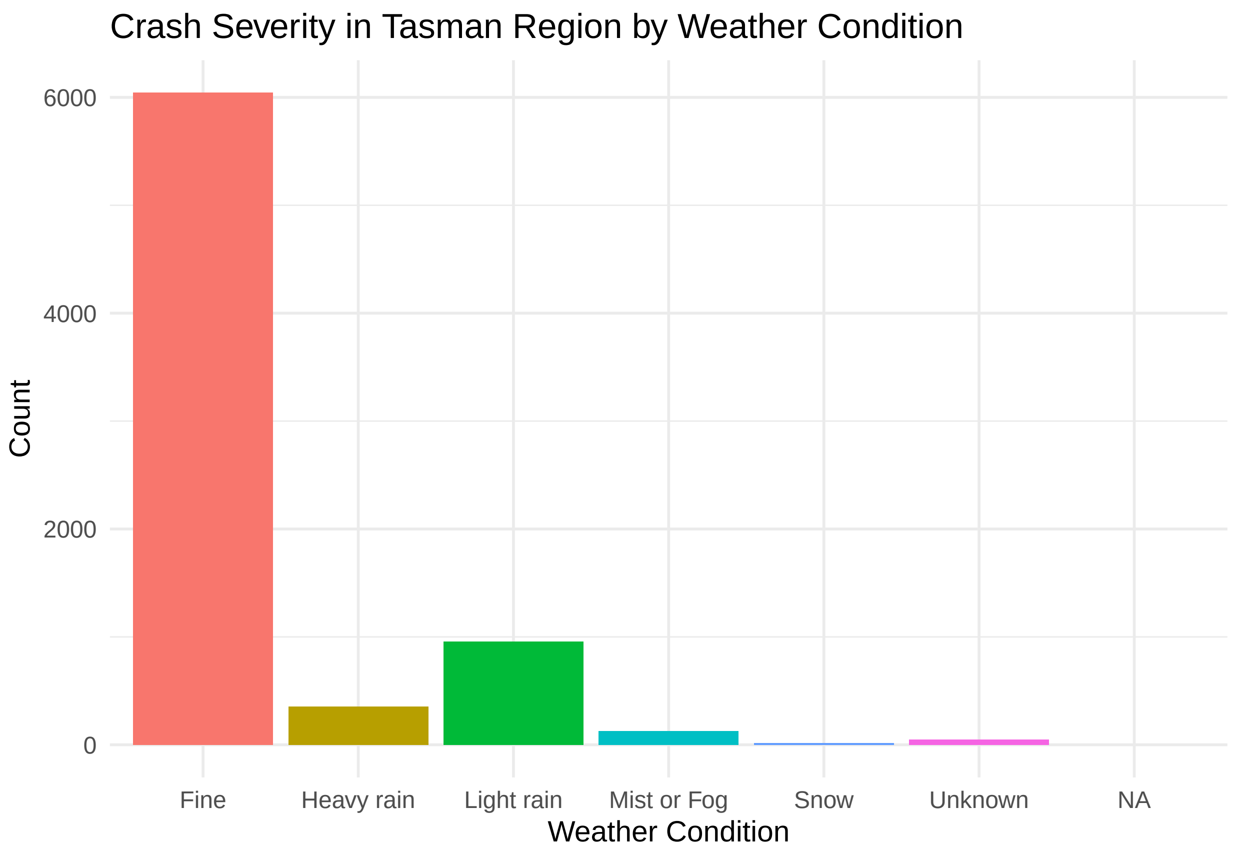
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



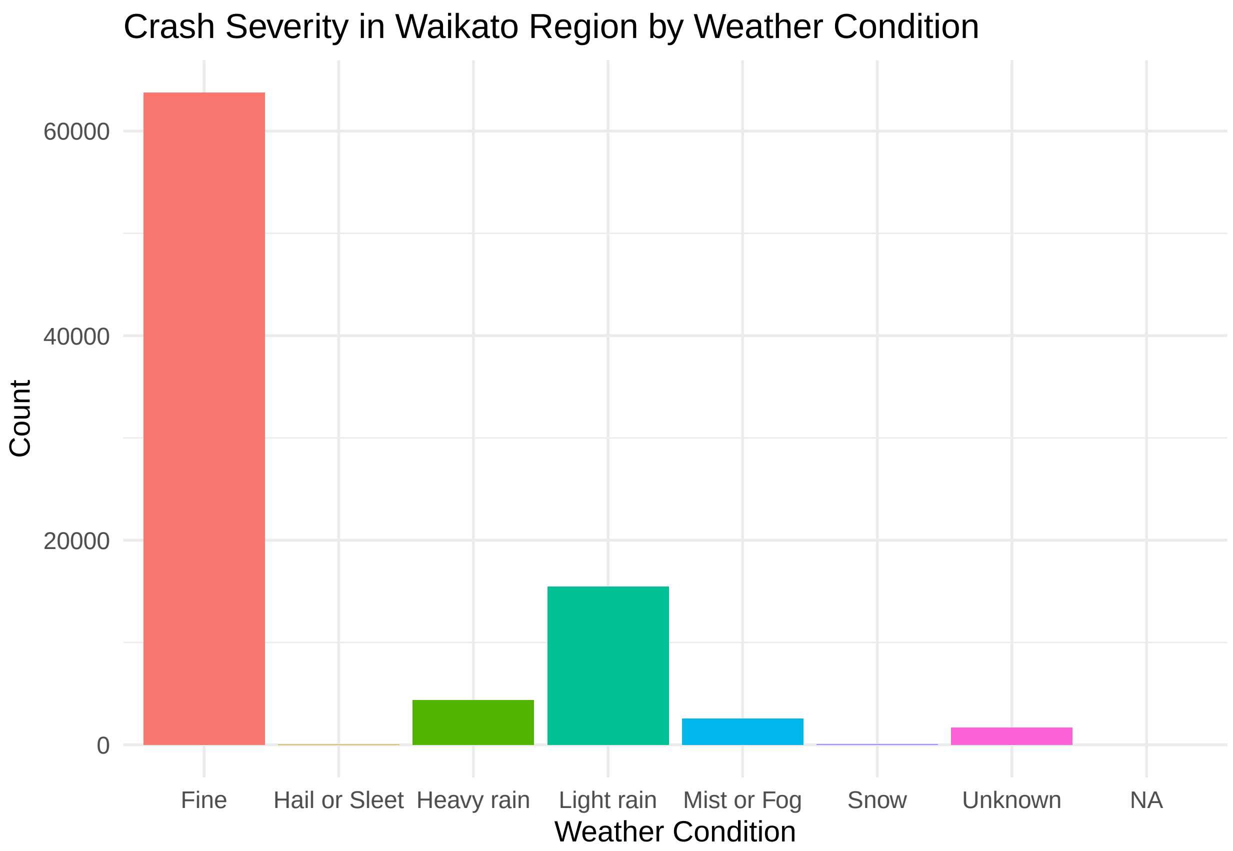
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



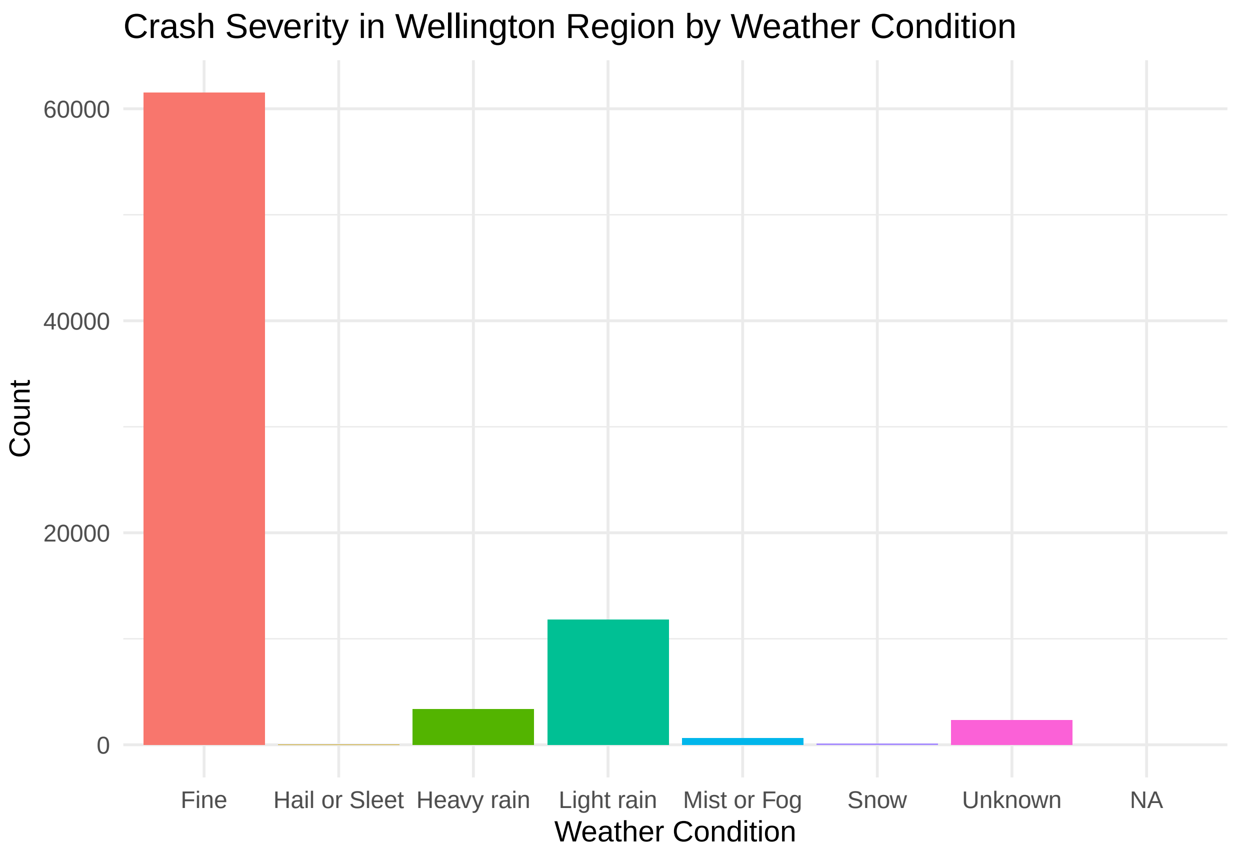
```
## Warning: Removed 18 rows containing missing values (`position_stack()`).
```



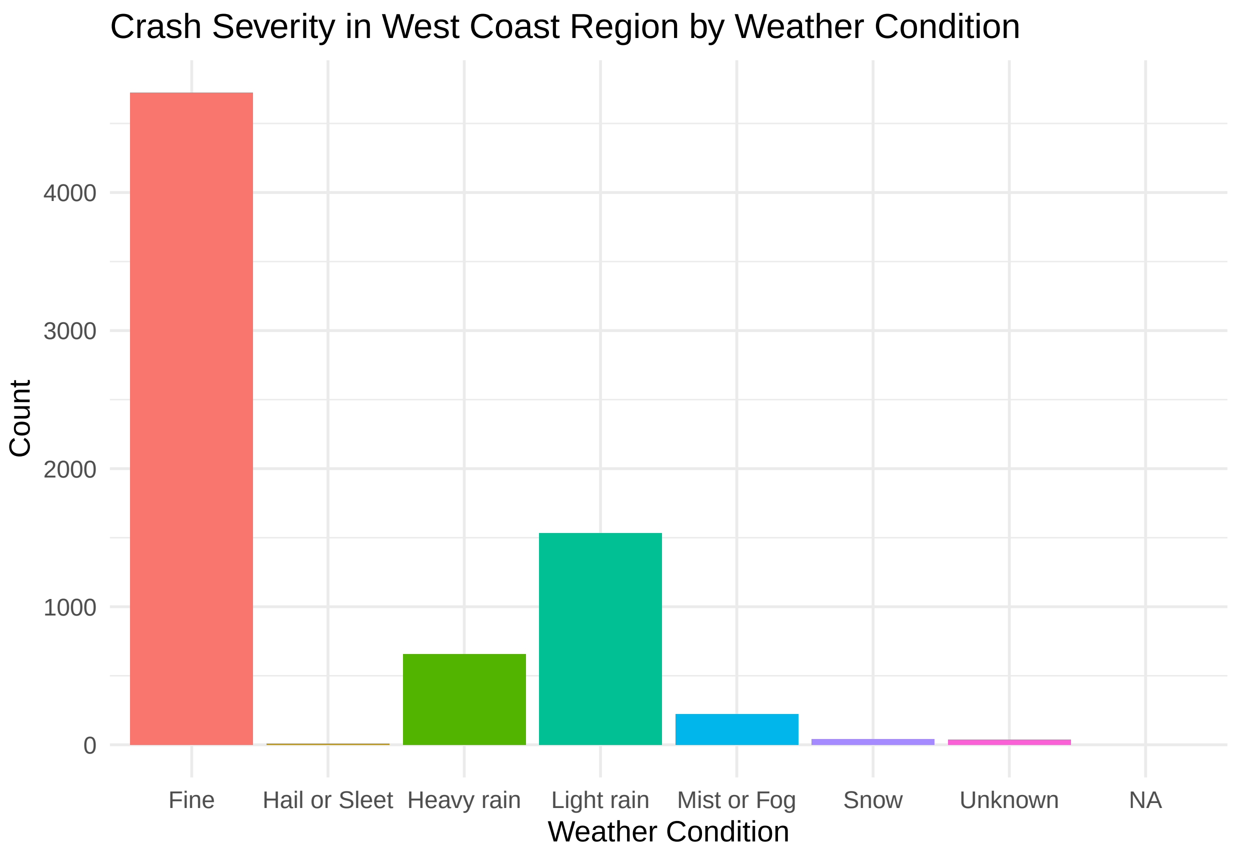
```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



```
## Warning: Removed 18 rows containing missing values (‘position_stack()‘).
```



```
## NULL
```

Individual Contributions

References