# Crash Fatalities

JAlexandra

2025-05

## Purpose

I am looking at understanding the relationship between fatalCount (the count of fatalities associated with a crash) and the rest of the variables in the Crash_Analysis_System_(CAS)_data data set.

## Looking at the data descriptions

On https://opendata-nzta.opendata.arcgis.com/pages/cas-data-field-descriptions there are the descriptions of the variables along with their variable names in the Crash data set.

I found that the variables crashDistance, easting, northing, and roadMarkings are listed as in the data set on that page, but are not in this csv.

The variables X, Y, objectID, and crashRoadSideRoad are in the data set but are not listed on this page.

The X-coordinate is often referred to as the "Easting", it seems like a reasonable assumption to say that the easting variable is the X variable and the northing variable is the Y variable in this dataset.

However there is no obvious link between the variables crashDistance, roadMarkings, objectID and crashRoadSideRoad.

Because I cannot determine the exact meanings of the variables objectID, and crashRoadSideRoad I will be removing them from the data set.

Small discrepencies in variable names when comparing the field descriptions variable names to the data sets variable names:

The variable intersectionMidblock mentioned in the field descriptions appears to be the variable intersection in the dataset.

Likewise with roadCharacter1 and roadCharacter.

The variables fatalCount, crashSeverity, seriousInjuryCount and minorInjuryCount contain similar information and I'm choosing to drop crashSeverity, seriousInjuryCount and minorInjuryCount

There are a significant amount variables that give location data, such as X, Y, and region. For the purposes of this research much of this information is redundant, as a result I will be removing the following variables crashLocation1, crashLocation2, directionRoleDescription and tlaName.

The variables crashYear and crashFinancialYear contain similar information, and I am choosing to drop crashFinancialYear.

The variable urban is derived from the variable speedLimit, since it contains the same information I will be removing it from the data set.

# Libraries

```
library(readxl)
library(finalfit)
library(naniar)
library(scales)
library(psych)
library(ggcorrplot)
library(caret)
library(moments)
library(MVN)
library(reshape2)
library(ggplot2)
library(pander)
library(car)
library(MASS)
library(dplyr)
library(AER)
library(performance)
library(DHARMa)
library(tidyverse)
library(broom)
library(knitr)
library(Metrics)
```

# Loading Data

```
Crash <- read.csv("Crash_Analysis_System_(CAS)_data.csv",
                  na.strings = c("", "Unknown", "Null", "Nil"))
```

Dropping unknown variables:

```
Crash <- subset(Crash, select = -c(OBJECTID, crashRoadSideRoad))
```

Dropping unnecessary variables:

```
Crash <- subset(Crash, select = c(-crashSeverity, -crashLocation1, -crashLocation2,
        -crashFinancialYear, -tlaName, -tlaId, -directionRoleDescription,
        -seriousInjuryCount, -minorInjuryCount, -urban))
```

# EDA

## Quantitative analysis

```
pander(data.frame(
  mean = sapply(select_if(Crash, is.numeric), mean, na.rm = TRUE),
  median = sapply(select_if(Crash, is.numeric), median, na.rm = TRUE),
  iqr = sapply(select_if(Crash, is.numeric), IQR, na.rm = TRUE)
))
```

| | mean | median | iqr |
|---|---|---|---|
| **X** | 1721397 | 1757428 | 89494 |
| **Y** | 5644547 | 5802445 | 479769 |
| **advisorySpeed** | 54.21 | 55 | 25 |
| **areaUnitID** | 546282 | 536651 | 53813 |
| **bicycle** | 0.02873 | 0 | 0 |
| **bridge** | 0.0137 | 0 | 0 |
| **bus** | 0.01593 | 0 | 0 |
| **carStationWagon** | 1.302 | 1 | 1 |
| **cliffBank** | 0.1063 | 0 | 0 |
| **crashYear** | 2012 | 2011 | 13 |
| **debris** | 0.00844 | 0 | 0 |
| **ditch** | 0.09428 | 0 | 0 |
| **fatalCount** | 0.01045 | 0 | 0 |
| **fence** | 0.2104 | 0 | 0 |
| **guardRail** | 0.08129 | 0 | 0 |
| **houseOrBuilding** | 0.02353 | 0 | 0 |
| **kerb** | 0.03547 | 0 | 0 |
| **meshblockId** | 1351845 | 1177900 | 1525399 |
| **moped** | 0.007022 | 0 | 0 |
| **motorcycle** | 0.03711 | 0 | 0 |
| **NumberOfLanes** | 2.332 | 2 | 0 |
| **objectThrownOrDropped** | 0.002169 | 0 | 0 |
| **otherObject** | 0.02378 | 0 | 0 |
| **otherVehicleType** | 0.005422 | 0 | 0 |
| **overBank** | 0.0422 | 0 | 0 |
| **parkedVehicle** | 0.2575 | 0 | 0 |
| **pedestrian** | 1.044 | 1 | 0 |
| **phoneBoxEtc** | 0.01256 | 0 | 0 |
| **postOrPole** | 0.1221 | 0 | 0 |
| **roadworks** | 0.003147 | 0 | 0 |
| **schoolBus** | 0.0008131 | 0 | 0 |
| **slipOrFlood** | 0.002536 | 0 | 0 |
| **speedLimit** | 65.94 | 50 | 50 |
| **strayAnimal** | 0.004231 | 0 | 0 |
| **suv** | 0.107 | 0 | 0 |
| **taxi** | 0.01064 | 0 | 0 |
| **temporarySpeedLimit** | 45.33 | 40 | 20 |
| **trafficIsland** | 0.0288 | 0 | 0 |
| **trafficSign** | 0.04928 | 0 | 0 |
| **train** | 0.001449 | 0 | 0 |
| **tree** | 0.1029 | 0 | 0 |
| **truck** | 0.0799 | 0 | 0 |
| **unknownVehicleType** | 0.003839 | 0 | 0 |
| **vanOrUtility** | 0.1783 | 0 | 0 |
| **vehicle** | 0.02407 | 0 | 0 |

| | mean | median | iqr |
|---|---|---|---|
| **waterRiver** | 0.009777 | 0 | 0 |

The variables are on extremely different scales, which could affect the values of regression coefficients, but will not affect the statistical significance or interpretation of the coefficients for the later regression.

```
pander(summary(Crash))
```

Table 2: Table continues below

| X | Y | advisorySpeed | areaUnitID |
|---|---|---|---|
| Min. :1150346 | Min. :4721798 | Min. :15.00 | Min. :500100 |
| 1st Qu.:1704319 | 1st Qu.:5434056 | 1st Qu.:40.00 | 1st Qu.:519710 |
| Median :1757428 | Median :5802445 | Median :55.00 | Median :536651 |
| Mean :1721397 | Mean :5644547 | Mean :54.21 | Mean :546282 |
| 3rd Qu.:1793813 | 3rd Qu.:5913825 | 3rd Qu.:65.00 | 3rd Qu.:573523 |
| Max. :2465388 | Max. :6190095 | Max. :95.00 | Max. :626801 |
| NA | NA | NA's :836776 | NA's :4 |

Table 3: Table continues below

| bicycle | bridge | bus | carStationWagon |
|---|---|---|---|
| Min. :0.00000 | Min. :0.00 | Min. :0.00000 | Min. : 0.000 |
| 1st Qu.:0.00000 | 1st Qu.:0.00 | 1st Qu.:0.00000 | 1st Qu.: 1.000 |
| Median :0.00000 | Median :0.00 | Median :0.00000 | Median : 1.000 |
| Mean :0.02873 | Mean :0.01 | Mean :0.01593 | Mean : 1.302 |
| 3rd Qu.:0.00000 | 3rd Qu.:0.00 | 3rd Qu.:0.00000 | 3rd Qu.: 2.000 |
| Max. :5.00000 | Max. :4.00 | Max. :3.00000 | Max. :11.000 |
| NA's :5 | NA's :513897 | NA's :5 | NA's :5 |

Table 4: Table continues below

| cliffBank | crashDirectionDescription | crashSHDescription | crashYear |
|---|---|---|---|
| Min. :0.00 | Length:870753 | Length:870753 | Min. :2000 |
| 1st Qu.:0.00 | Class :character | Class :character | 1st Qu.:2005 |
| Median :0.00 | Mode :character | Mode :character | Median :2011 |
| Mean :0.11 | NA | NA | Mean :2012 |
| 3rd Qu.:0.00 | NA | NA | 3rd Qu.:2018 |
| Max. :3.00 | NA | NA | Max. :2025 |
| NA's :513897 | NA | NA | NA |

Table 5: Table continues below

| debris | ditch | fatalCount | fence |
|---|---|---|---|
| Min. :0.00 | Min. :0.00 | Min. :0.00000 | Min. :0.00 |
| 1st Qu.:0.00 | 1st Qu.:0.00 | 1st Qu.:0.00000 | 1st Qu.:0.00 |

| debris | ditch | fatalCount | fence |
|---|---|---|---|
| Median :0.00 | Median :0.00 | Median :0.00000 | Median :0.00 |
| Mean :0.01 | Mean :0.09 | Mean :0.01045 | Mean :0.21 |
| 3rd Qu.:0.00 | 3rd Qu.:0.00 | 3rd Qu.:0.00000 | 3rd Qu.:0.00 |
| Max. :7.00 | Max. :3.00 | Max. :9.00000 | Max. :4.00 |
| NA's :513897 | NA's :513897 | NA's :1 | NA's :513897 |

Table 6: Table continues below

| flatHill | guardRail | holiday | houseOrBuilding |
|---|---|---|---|
| Length:870753 | Min. :0.00 | Length:870753 | Min. :0.00 |
| Class :character | 1st Qu.:0.00 | Class :character | 1st Qu.:0.00 |
| Mode :character | Median :0.00 | Mode :character | Median :0.00 |
| NA | Mean :0.08 | NA | Mean :0.02 |
| NA | 3rd Qu.:0.00 | NA | 3rd Qu.:0.00 |
| NA | Max. :4.00 | NA | Max. :2.00 |
| NA | NA's :513897 | NA | NA's :513897 |

Table 7: Table continues below

| intersection | kerb | light | meshblockId |
|---|---|---|---|
| Mode:logical | Min. :0.00 | Length:870753 | Min. : 100 |
| NA's:870753 | 1st Qu.:0.00 | Class :character | 1st Qu.: 602401 |
| NA | Median :0.00 | Mode :character | Median :1177900 |
| NA | Mean :0.04 | NA | Mean :1351845 |
| NA | 3rd Qu.:0.00 | NA | 3rd Qu.:2127800 |
| NA | Max. :3.00 | NA | Max. :3209003 |
| NA | NA's :513897 | NA | NA's :4 |

Table 8: Table continues below

| moped | motorcycle | NumberOfLanes | objectThrownOrDropped |
|---|---|---|---|
| Min. :0.000000 | Min. :0.00000 | Min. :0.000 | Min. :0 |
| 1st Qu.:0.000000 | 1st Qu.:0.00000 | 1st Qu.:2.000 | 1st Qu.:0 |
| Median :0.000000 | Median :0.00000 | Median :2.000 | Median :0 |
| Mean :0.007021 | Mean :0.03711 | Mean :2.332 | Mean :0 |
| 3rd Qu.:0.000000 | 3rd Qu.:0.00000 | 3rd Qu.:2.000 | 3rd Qu.:0 |
| Max. :4.000000 | Max. :8.00000 | Max. :9.000 | Max. :4 |
| NA's :5 | NA's :5 | NA's :2094 | NA's :513897 |

Table 9: Table continues below

| otherObject | otherVehicleType | overBank | parkedVehicle | pedestrian |
|---|---|---|---|---|
| Min. :0.00 | Min. :0.000000 | Min. :0.00 | Min. :0.00 | Min. :1.00 |
| 1st Qu.:0.00 | 1st Qu.:0.000000 | 1st Qu.:0.00 | 1st Qu.:0.00 | 1st Qu.:1.00 |
| Median :0.00 | Median :0.000000 | Median :0.00 | Median :0.00 | Median :1.00 |
| Mean :0.02 | Mean :0.005422 | Mean :0.04 | Mean :0.26 | Mean :1.04 |

| otherObject | otherVehicleType | overBank | parkedVehicle | pedestrian |
|---|---|---|---|---|
| 3rd Qu.:0.00 | 3rd Qu.:0.000000 | 3rd Qu.:0.00 | 3rd Qu.:0.00 | 3rd Qu.:1.00 |
| Max. :5.00 | Max. :3.000000 | Max. :4.00 | Max. :8.00 | Max. :7.00 |
| NA's :513897 | NA's :5 | NA's :513897 | NA's :513897 | NA's :842125 |

Table 10: Table continues below

| phoneBoxEtc | postOrPole | region | roadCharacter |
|---|---|---|---|
| Min. :0.00 | Min. :0.00 | Length:870753 | Length:870753 |
| 1st Qu.:0.00 | 1st Qu.:0.00 | Class :character | Class :character |
| Median :0.00 | Median :0.00 | Mode :character | Mode :character |
| Mean :0.01 | Mean :0.12 | NA | NA |
| 3rd Qu.:0.00 | 3rd Qu.:0.00 | NA | NA |
| Max. :3.00 | Max. :4.00 | NA | NA |
| NA's :513897 | NA's :513897 | NA | NA |

Table 11: Table continues below

| roadLane | roadSurface | roadworks | schoolBus |
|---|---|---|---|
| Length:870753 | Length:870753 | Min. :0 | Min. :0.0000000 |
| Class :character | Class :character | 1st Qu.:0 | 1st Qu.:0.0000000 |
| Mode :character | Mode :character | Median :0 | Median :0.0000000 |
| NA | NA | Mean :0 | Mean :0.0008131 |
| NA | NA | 3rd Qu.:0 | 3rd Qu.:0.0000000 |
| NA | NA | Max. :3 | Max. :3.0000000 |
| NA | NA | NA's :513897 | NA's :5 |

Table 12: Table continues below

| slipOrFlood | speedLimit | strayAnimal | streetLight |
|---|---|---|---|
| Min. :0 | Min. : 2.00 | Min. :0 | Length:870753 |
| 1st Qu.:0 | 1st Qu.: 50.00 | 1st Qu.:0 | Class :character |
| Median :0 | Median : 50.00 | Median :0 | Mode :character |
| Mean :0 | Mean : 65.94 | Mean :0 | NA |
| 3rd Qu.:0 | 3rd Qu.:100.00 | 3rd Qu.:0 | NA |
| Max. :4 | Max. :110.00 | Max. :3 | NA |
| NA's :513897 | NA's :1148 | NA's :513897 | NA |

Table 13: Table continues below

| suv | taxi | temporarySpeedLimit | trafficControl |
|---|---|---|---|
| Min. :0.000 | Min. :0.00000 | Min. : 10.00 | Length:870753 |
| 1st Qu.:0.000 | 1st Qu.:0.00000 | 1st Qu.: 30.00 | Class :character |
| Median :0.000 | Median :0.00000 | Median : 40.00 | Mode :character |
| Mean :0.107 | Mean :0.01064 | Mean : 45.33 | NA |
| 3rd Qu.:0.000 | 3rd Qu.:0.00000 | 3rd Qu.: 50.00 | NA |
| Max. :6.000 | Max. :5.00000 | Max. :100.00 | NA |

| suv | taxi | temporarySpeedLimit | trafficControl |
|---|---|---|---|
| NA's :5 | NA's :5 | NA's :856353 | NA |

Table 14: Table continues below

| trafficIsland | trafficSign | train | tree | truck |
|---|---|---|---|---|
| Min. :0.00 | Min. :0.00 | Min. :0 | Min. :0.0 | Min. :0.0000 |
| 1st Qu.:0.00 | 1st Qu.:0.00 | 1st Qu.:0 | 1st Qu.:0.0 | 1st Qu.:0.0000 |
| Median :0.00 | Median :0.00 | Median :0 | Median :0.0 | Median :0.0000 |
| Mean :0.03 | Mean :0.05 | Mean :0 | Mean :0.1 | Mean :0.0799 |
| 3rd Qu.:0.00 | 3rd Qu.:0.00 | 3rd Qu.:0 | 3rd Qu.:0.0 | 3rd Qu.:0.0000 |
| Max. :4.00 | Max. :4.00 | Max. :1 | Max. :4.0 | Max. :5.0000 |
| NA's :513897 | NA's :513897 | NA's :513897 | NA's :513897 | NA's :5 |

Table 15: Table continues below

| unknownVehicleType | vanOrUtility | vehicle | waterRiver |
|---|---|---|---|
| Min. :0.000000 | Min. :0.0000 | Min. :0.00 | Min. :0.00 |
| 1st Qu.:0.000000 | 1st Qu.:0.0000 | 1st Qu.:0.00 | 1st Qu.:0.00 |
| Median :0.000000 | Median :0.0000 | Median :0.00 | Median :0.00 |
| Mean :0.003839 | Mean :0.1783 | Mean :0.02 | Mean :0.01 |
| 3rd Qu.:0.000000 | 3rd Qu.:0.0000 | 3rd Qu.:0.00 | 3rd Qu.:0.00 |
| Max. :3.000000 | Max. :6.0000 | Max. :4.00 | Max. :2.00 |
| NA's :5 | NA's :5 | NA's :513897 | NA's :513897 |

| weatherA | weatherB |
|---|---|
| Length:870753 | Length:870753 |
| Class :character | Class :character |
| Mode :character | Mode :character |
| NA | NA |
| NA | NA |
| NA | NA |
| NA | NA |

Count data: fatalCount, bicycle, bridge, bus, carStationWagon, cliffBank, debris, ditch, fence, guardRail, houseOrBuilding, kerb, moped, motorcycle, NumberOfLanes, objectThrownOrDropped, otherObject, otherVehicleType, overBank, parkedVehicle, pedestrian, phoneBoxEtc, postOrPole, roadworks, schoolBus, slipOrFlood, strayAnimal, suv, taxi, trafficIsland, trafficSign, train, tree, truck, unknownVehicleType, vanOrUtility, vehicle, waterRiver.

A large amount of the count data variables are derived variables.

Discrete variables (not including count data): crashYear

Continuous variables: X, Y, advisorySpeed, areaUnitID, meshblockId, speedLimit, temporarySpeedLimit

All of the categorical variables are nominal (lack an inherent order).

The data set is majority made up of count data and categorical variables.

There is one logical variable in the data set named intersection it has 870753 NA's which is equal to the total amount of observations in the data set. This means that this variable has no data for any of the observations. As a result I will be removing it

```
Crash <- subset(Crash, select = -c(intersection))
```

```
pander(head(Crash))
```

Table 17: Table continues below

| X | Y | advisorySpeed | areaUnitID | bicycle | bridge | bus |
|---|---|---|---|---|---|---|
| 1243177 | 4849584 | NA | 611210 | 0 | NA | 0 |
| 1832358 | 5584384 | NA | 559220 | 0 | 0 | 0 |
| 1749496 | 5918077 | NA | 514801 | 0 | NA | 0 |
| 2038048 | 5708026 | NA | 544701 | 0 | NA | 0 |
| 1834941 | 5642955 | NA | 554900 | 0 | 0 | 0 |
| 1693702 | 5676441 | NA | 551800 | 0 | NA | 0 |

Table 18: Table continues below

| carStationWagon | cliffBank | crashDirectionDescription | crashSHDescription |
|---|---|---|---|
| 2 | NA | NA | Yes |
| 0 | 0 | South | Yes |
| 1 | NA | NA | Yes |
| 2 | NA | NA | No |
| 0 | 1 | South | Yes |
| 2 | NA | West | Yes |

Table 19: Table continues below

| crashYear | debris | ditch | fatalCount | fence | flatHill | guardRail |
|---|---|---|---|---|---|---|
| 2003 | NA | NA | 0 | NA | Flat | NA |
| 2003 | 0 | 1 | 0 | 1 | Flat | 0 |
| 2003 | NA | NA | 0 | NA | Flat | NA |
| 2004 | NA | NA | 0 | NA | Flat | NA |
| 2005 | 0 | 0 | 0 | 0 | Flat | 0 |
| 2003 | NA | NA | 0 | NA | Flat | NA |

Table 20: Table continues below

| holiday | houseOrBuilding | kerb | light | meshblockId | moped |
|---|---|---|---|---|---|
| NA | NA | NA | Overcast | 3109000 | 0 |
| NA | 0 | 0 | Dark | 1748400 | 0 |
| NA | NA | NA | Dark | 390401 | 0 |
| NA | NA | NA | Overcast | 1382300 | 0 |
| NA | 0 | 0 | Overcast | 1673301 | 0 |
| NA | NA | NA | Bright sun | 1594000 | 0 |

Table 21: Table continues below

| motorcycle | NumberOfLanes | objectThrownOrDropped | otherObject |
|---|---|---|---|
| 0 | 4 | NA | NA |
| 0 | 2 | 0 | 0 |
| 0 | 3 | NA | NA |
| 0 | 2 | NA | NA |
| 0 | 2 | 0 | 0 |
| 0 | 4 | NA | NA |

Table 22: Table continues below

| otherVehicleType | overBank | parkedVehicle | pedestrian | phoneBoxEtc |
|---|---|---|---|---|
| 0 | NA | NA | NA | NA |
| 0 | 0 | 0 | NA | 0 |
| 0 | NA | NA | NA | NA |
| 0 | NA | NA | NA | NA |
| 0 | 0 | 0 | NA | 0 |
| 0 | NA | NA | NA | NA |

Table 23: Table continues below

| postOrPole | region | roadCharacter | roadLane |
|---|---|---|---|
| NA | Southland Region | NA | 2-way |
| 0 | Manawatū-Whanganui Region | NA | 2-way |
| NA | Auckland Region | Motorway ramp | 1-way |
| NA | Gisborne Region | NA | 2-way |
| 0 | Manawatū-Whanganui Region | NA | 2-way |
| NA | Taranaki Region | NA | 1-way |

Table 24: Table continues below

| roadSurface | roadworks | schoolBus | slipOrFlood | speedLimit | strayAnimal |
|---|---|---|---|---|---|
| Sealed | NA | 0 | NA | 50 | NA |
| Sealed | 0 | 0 | 0 | 100 | 0 |
| Sealed | NA | 0 | NA | 100 | NA |
| Sealed | NA | 0 | NA | 50 | NA |
| Sealed | 0 | 0 | 0 | 100 | 0 |
| Sealed | NA | 0 | NA | 50 | NA |

Table 25: Table continues below

| streetLight | suv | taxi | temporarySpeedLimit | trafficControl |
|---|---|---|---|---|
| NA | 0 | 0 | NA | Traffic Signals |
| None | 0 | 0 | NA | NA |
| On | 0 | 0 | NA | NA |
| NA | 0 | 0 | NA | Give way |

| streetLight | suv | taxi | temporarySpeedLimit | trafficControl |
|---|---|---|---|---|
| NA | 0 | 0 | NA | NA |
| NA | 0 | 0 | NA | NA |

Table 26: Table continues below

| trafficIsland | trafficSign | train | tree | truck | unknownVehicleType |
|---|---|---|---|---|---|
| NA | NA | NA | NA | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| NA | NA | NA | NA | 1 | 0 |
| NA | NA | NA | NA | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 |
| NA | NA | NA | NA | 0 | 0 |

| vanOrUtility | vehicle | waterRiver | weatherA | weatherB |
|---|---|---|---|---|
| 0 | NA | NA | Fine | NA |
| 1 | 0 | 0 | Fine | NA |
| 0 | NA | NA | Heavy rain | NA |
| 0 | NA | NA | Fine | NA |
| 0 | 0 | 0 | Heavy rain | NA |
| 0 | NA | NA | Fine | NA |

I can see that some variables have lots of NA's such as advisorySpeed which has 836776 NA's - almost the entire variable's data is missing data.

## Data cleaning

### Converting Categorical data for regression

All categorical variables in the data set:

```
colnames(select_if(Crash, is.character))
```

```
##  [1] "crashDirectionDescription" "crashSHDescription"
##  [3] "flatHill"                  "holiday"
##  [5] "light"                     "region"
##  [7] "roadCharacter"             "roadLane"
##  [9] "roadSurface"               "streetLight"
## [11] "trafficControl"            "weatherA"
## [13] "weatherB"
```

```
sort(unique(Crash[["crashSHDescription"]]))
```

```
## [1] "No"  "Yes"
```

Note: crashSHDescription "Indicates where a crash is reported to have occurred on a State Highway (SH) marked '1', or on another road type marked '2' " according to the field descriptions, but in this data set it is coded with "No" and "Yes".

Turning all categorical variables into factors for regression:

```
Crash$crashSHDescription <- as.factor(Crash$crashSHDescription)
Crash$flatHill <- as.factor(Crash$flatHill)
Crash$holiday <- as.factor(Crash$holiday)
Crash$light <- as.factor(Crash$light)
Crash$region <- as.factor(Crash$region)
Crash$roadCharacter <- as.factor(Crash$roadCharacter)
Crash$roadLane <- as.factor(Crash$roadLane)
Crash$roadSurface <- as.factor(Crash$roadSurface)
Crash$streetLight <- as.factor(Crash$streetLight)
Crash$trafficControl <- as.factor(Crash$trafficControl)
Crash$weatherA <- as.factor(Crash$weatherA)
Crash$weatherB <- as.factor(Crash$weatherB)
Crash$crashDirectionDescription <- as.factor(Crash$crashDirectionDescription)
```

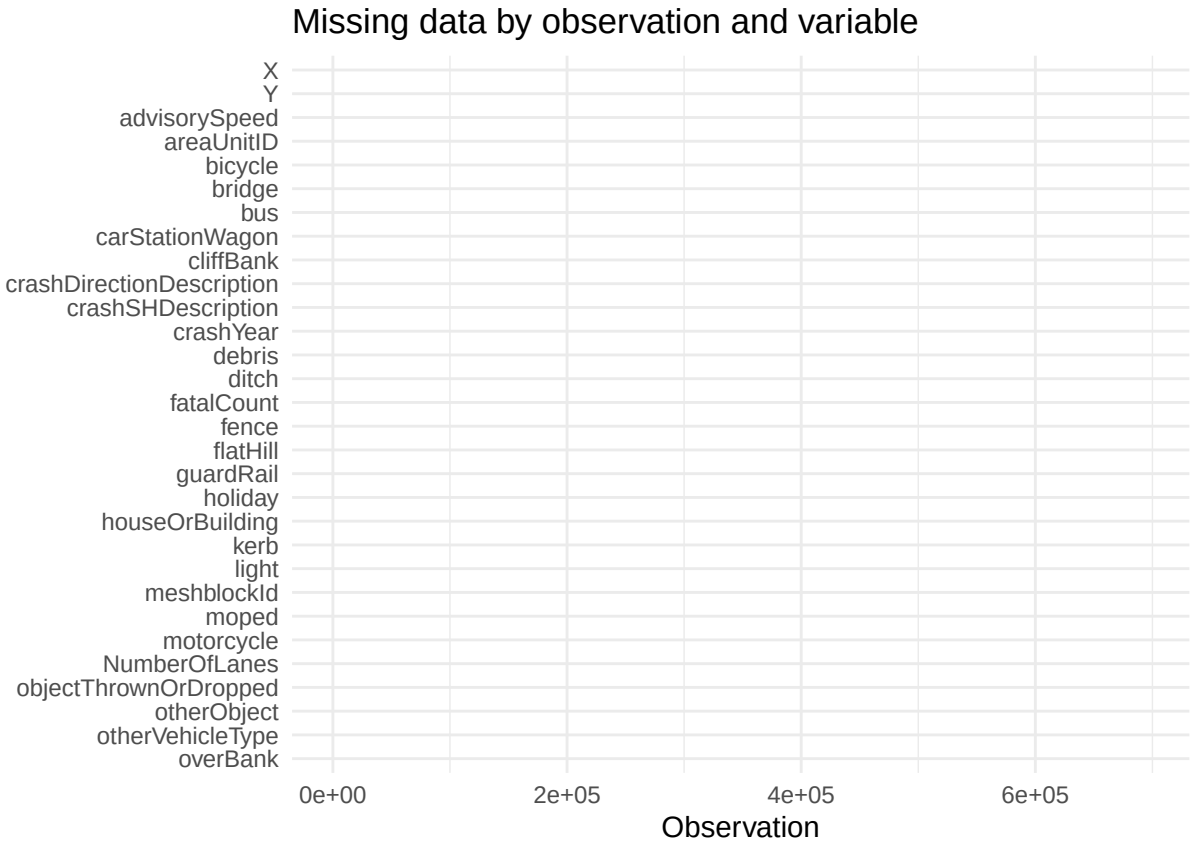
### Train, test split

```
set.seed(1)
sample <- sample(c(TRUE, FALSE), nrow(Crash), replace = TRUE, prob = c(0.8, 0.2))
Train_set  <- Crash[sample, ]
Test_set   <- Crash[!sample, ]
```

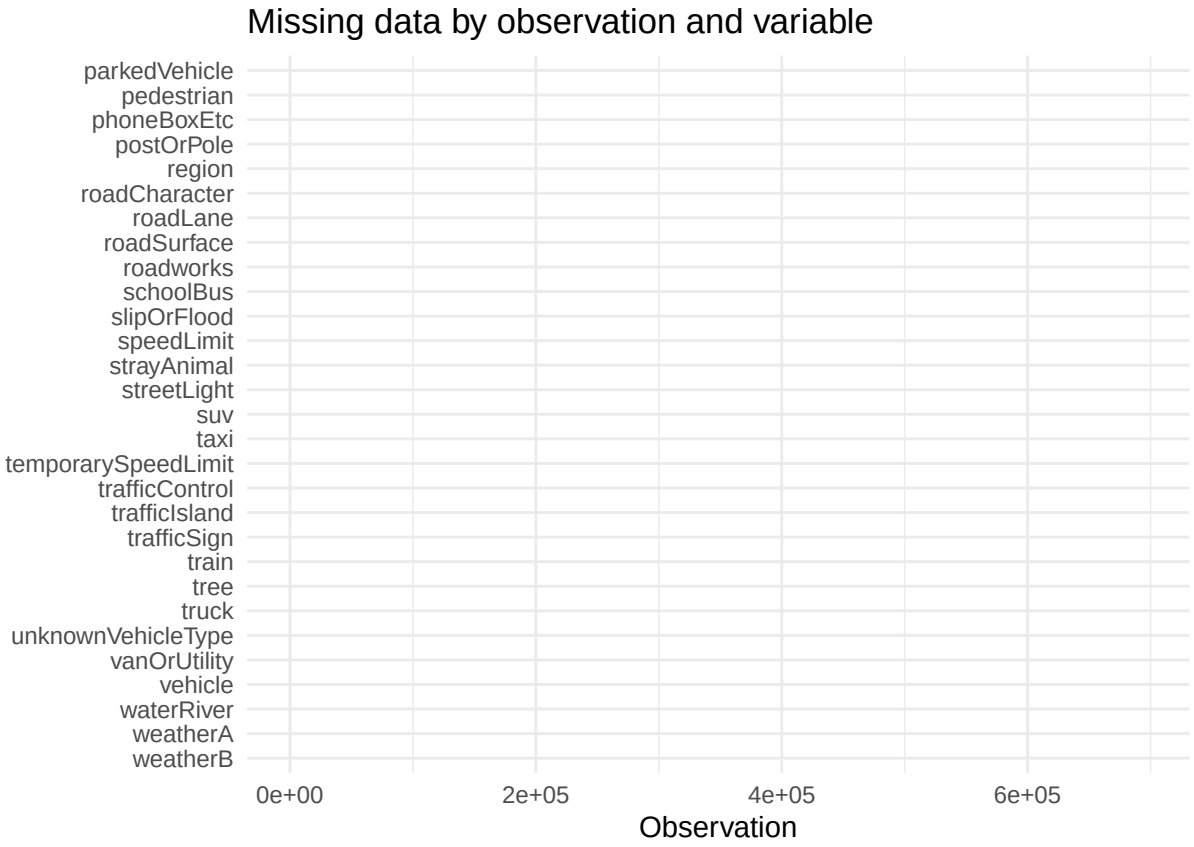
### Missing data

There are 59 variables and trying to place them onto one plot caused it to be unreadable. To resolve this it has been split into separate plots.

```
# Producing a missing data plot for the data frame.
missing_plot(Train_set[,1:30], title = "Missing data by observation and variable")
```




```
missing_plot(Train_set[,31:59], title = "Missing data by observation and variable")
```




```
# Saving plots:
# plotA <- missing_plot(Train_set[,1:30], title = "Missing data by observation and variable")
# plotB <- missing_plot(Train_set[,31:59], title = "Missing data by observation and variable")
# ggsave(plotA,
#       filename = "Missing data by observation and variable (variables 1 to 31).png",
#       device = "png")

#ggsave(plotB,
#       filename = "Missing data by observation and variable (variables 31 to 59).png",
#       device = "png")
```

It seems that most variables in the Crash data set have large amounts of missing data. Particularly temporarySpeedLimit, pedestrian, the encoded WeatherB variables, the encoded holiday variables, and adviserySpeed which appear to be majority missing data.

Calculating how many variables have missing data:

```
percentages <- c()
missing_percents_colnames <- c()
x = 0
for (i in colnames(Train_set)){
  if (sum(is.na(Train_set[[i]])) > 0){
    percentages <- append(percentages, prop_miss(Train_set[[i]]))
    missing_percents_colnames <- append(missing_percents_colnames, colnames(Train_set[i]))
    x = x + 1
  }
```

```
}
print(x)
```

```
## [1] 56
```

```
# Turning proportion to percentage
percentages <- percent(percentages, accuracy = 0.01)
```

There are 55 variables with missing data out of 59.

Finding out what proportion of each variables data is missing:

```
missing.table <- do.call(rbind, Map(data.frame, variable = missing_percents_colnames,
                                    percentage = percentages))
missing.table <- missing.table[rev(order(missing.table$percentage)), ]
row.names(missing.table) <- c(1:nrow(missing.table))
missing.table
```

```
##                      variable percentage
## 1         temporarySpeedLimit     98.34%
## 2                    weatherB     96.79%
## 3                  pedestrian     96.70%
## 4               advisorySpeed     96.08%
## 5               roadCharacter     95.98%
## 6                     holiday     94.48%
## 7              trafficControl     66.34%
## 8                  waterRiver     58.99%
## 9                     vehicle     58.99%
## 10                       tree     58.99%
## 11                      train     58.99%
## 12                trafficSign     58.99%
## 13              trafficIsland     58.99%
## 14                strayAnimal     58.99%
## 15                slipOrFlood     58.99%
## 16                  roadworks     58.99%
## 17                 postOrPole     58.99%
## 18                phoneBoxEtc     58.99%
## 19              parkedVehicle     58.99%
## 20                   overBank     58.99%
## 21                otherObject     58.99%
## 22     objectThrownOrDropped     58.99%
## 23                       kerb     58.99%
## 24           houseOrBuilding     58.99%
## 25                 guardRail     58.99%
## 26                     fence     58.99%
## 27                     ditch     58.99%
## 28                    debris     58.99%
## 29                 cliffBank     58.99%
## 30                    bridge     58.99%
## 31 crashDirectionDescription     37.57%
## 32                streetLight     34.47%
## 33                   weatherA      1.84%
```

```
## 34                    light      0.96%
## 35                  flatHill      0.74%
## 36                    region      0.37%
## 37             NumberOfLanes      0.23%
## 38                speedLimit      0.13%
## 39               roadSurface      0.12%
## 40                  roadLane      0.06%
## 41         crashSHDescription      0.02%
## 42               vanOrUtility      0.00%
## 43         unknownVehicleType      0.00%
## 44                     truck      0.00%
## 45                      taxi      0.00%
## 46                       suv      0.00%
## 47                 schoolBus      0.00%
## 48           otherVehicleType      0.00%
## 49                 motorcycle      0.00%
## 50                     moped      0.00%
## 51               meshblockId      0.00%
## 52                fatalCount      0.00%
## 53            carStationWagon      0.00%
## 54                       bus      0.00%
## 55                   bicycle      0.00%
## 56                areaUnitID      0.00%
```

30 variables have more than 59% of their data missing. Another 2 variables have around 30% of their data missing.

I think it would be best to drop these variables as other methods such as imputation for the missing data would be too computationally expensive given the scale of the data.

```
# Dropping variables
Train_set <- Train_set %>% select(-missing.table$variable[1:32])
```

There are now have 27 variables in the data set.

Reducing the data down to complete cases only:

```
Train_set <- Train_set[complete.cases(Train_set), ]
# making the dependent variable the first column
Train_set <- Train_set %>% relocate(fatalCount)
```

This causes the amount of observations I have to go from 870,753 to 844,965, a reduction of 2.96% (to 2 d.p).

# Plotting data

## Boxplots of categorical data

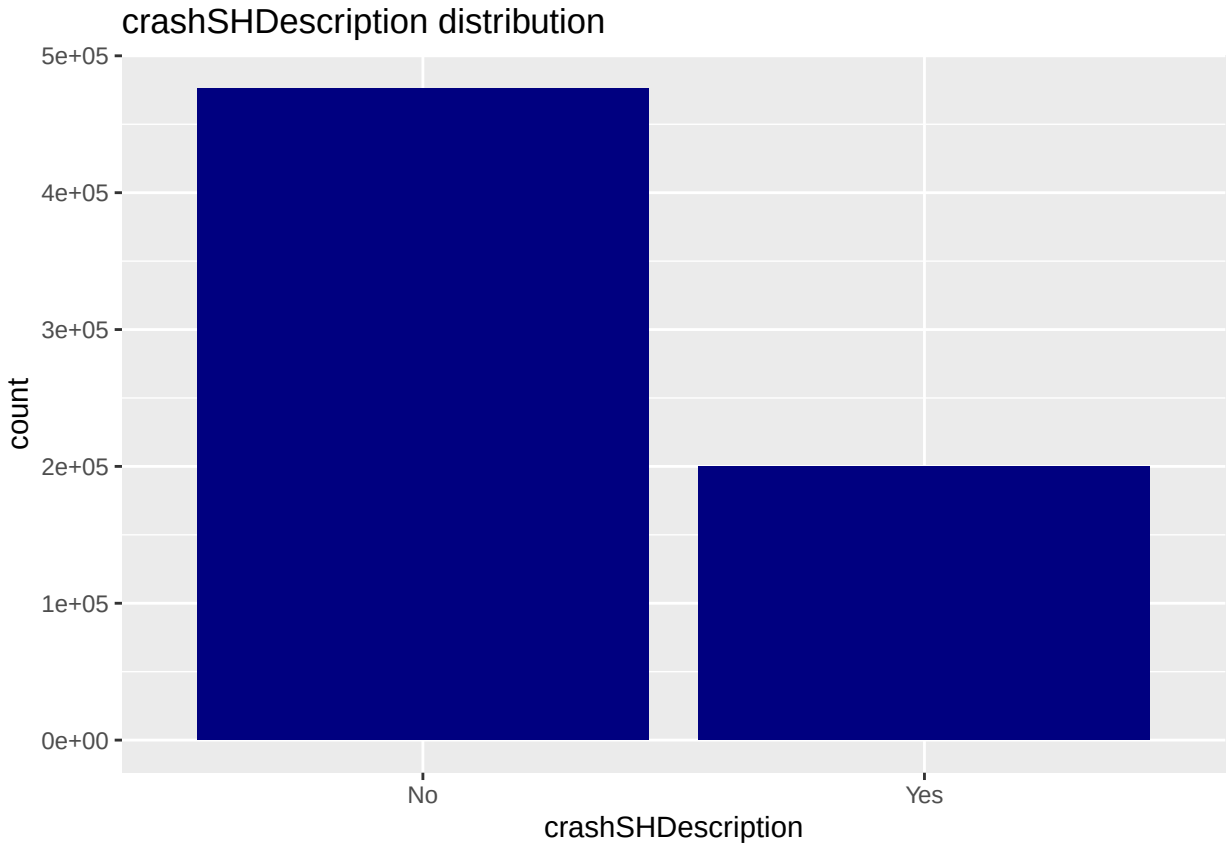
All categorical variables:

```
colnames(select_if(Train_set, is.factor))
```

```
## [1] "crashSHDescription" "flatHill"           "light"
## [4] "region"             "roadLane"           "roadSurface"
## [7] "weatherA"
```

Bar plot of crashSHDescription:

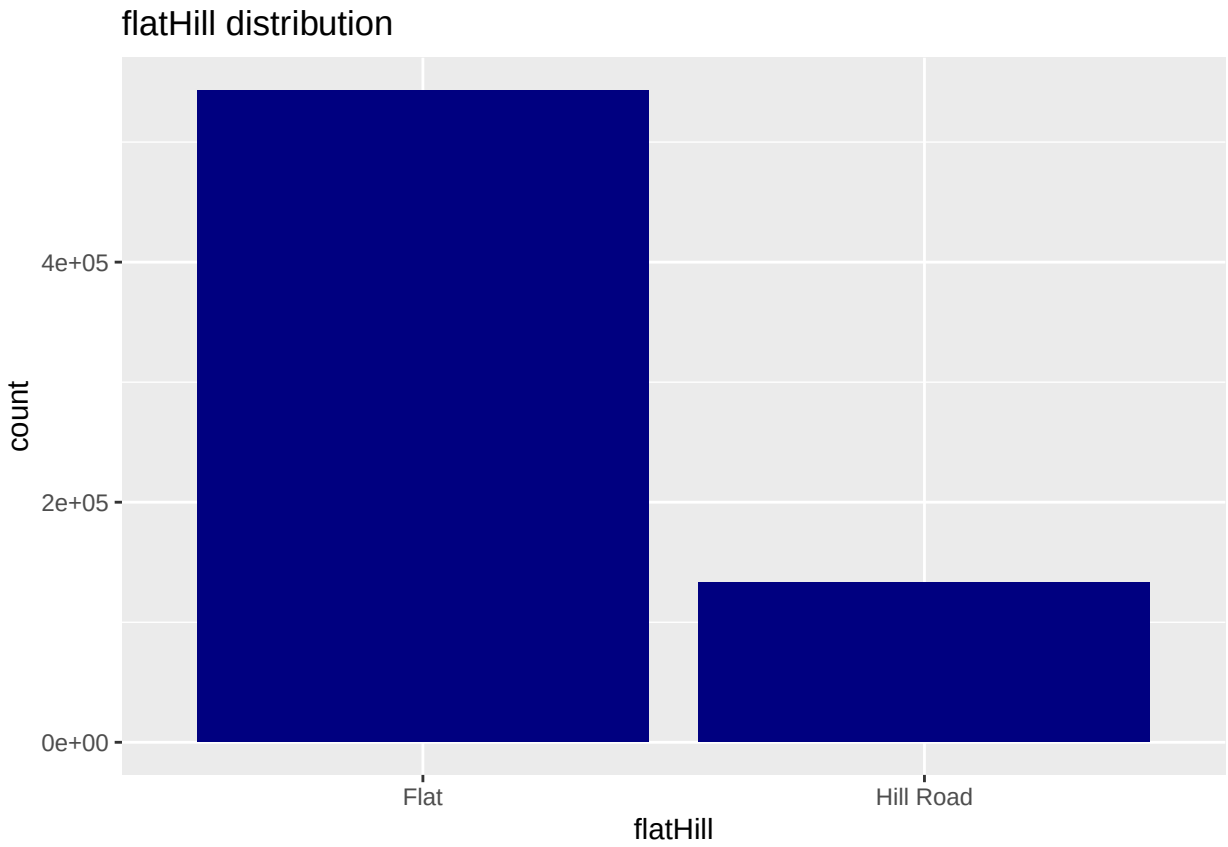
```
ggplot(Train_set, aes(x = crashSHDescription)) +
  geom_bar(fill = "navy") +
  ggtitle("crashSHDescription distribution")
```




Less crashes occurred on a state highway, around 200,000 of the observations occurred on a state highway.

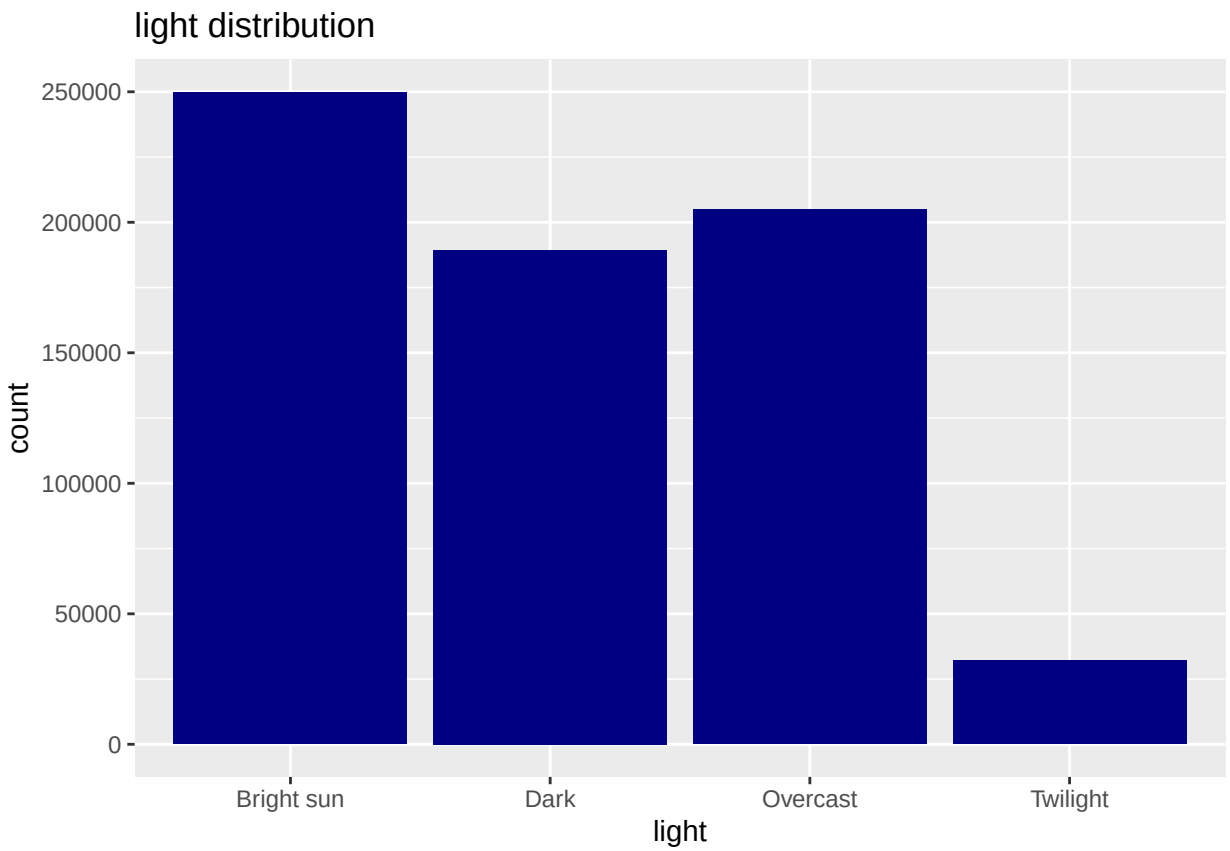
Bar plot of flatHill:

```
ggplot(Train_set, aes(x = flatHill)) +
  geom_bar(fill = "navy") +
  ggtitle("flatHill distribution")
```




Bar plot of light:

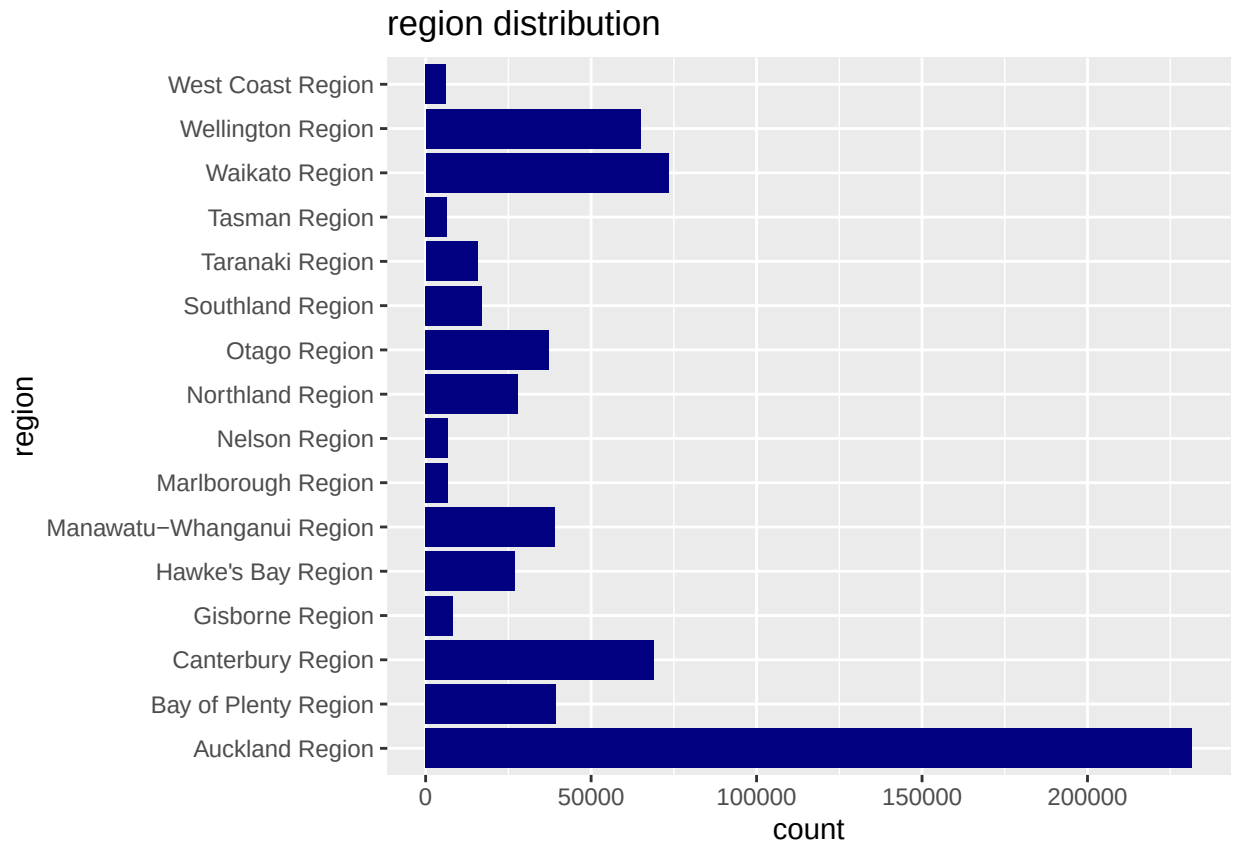
```
ggplot(Train_set, aes(x = light)) +
  geom_bar(fill = "navy") +
  ggtitle("light distribution")
```




There are less twilight observations, less than 50,000 of the observations occurred during twilight.

Bar plot of region:

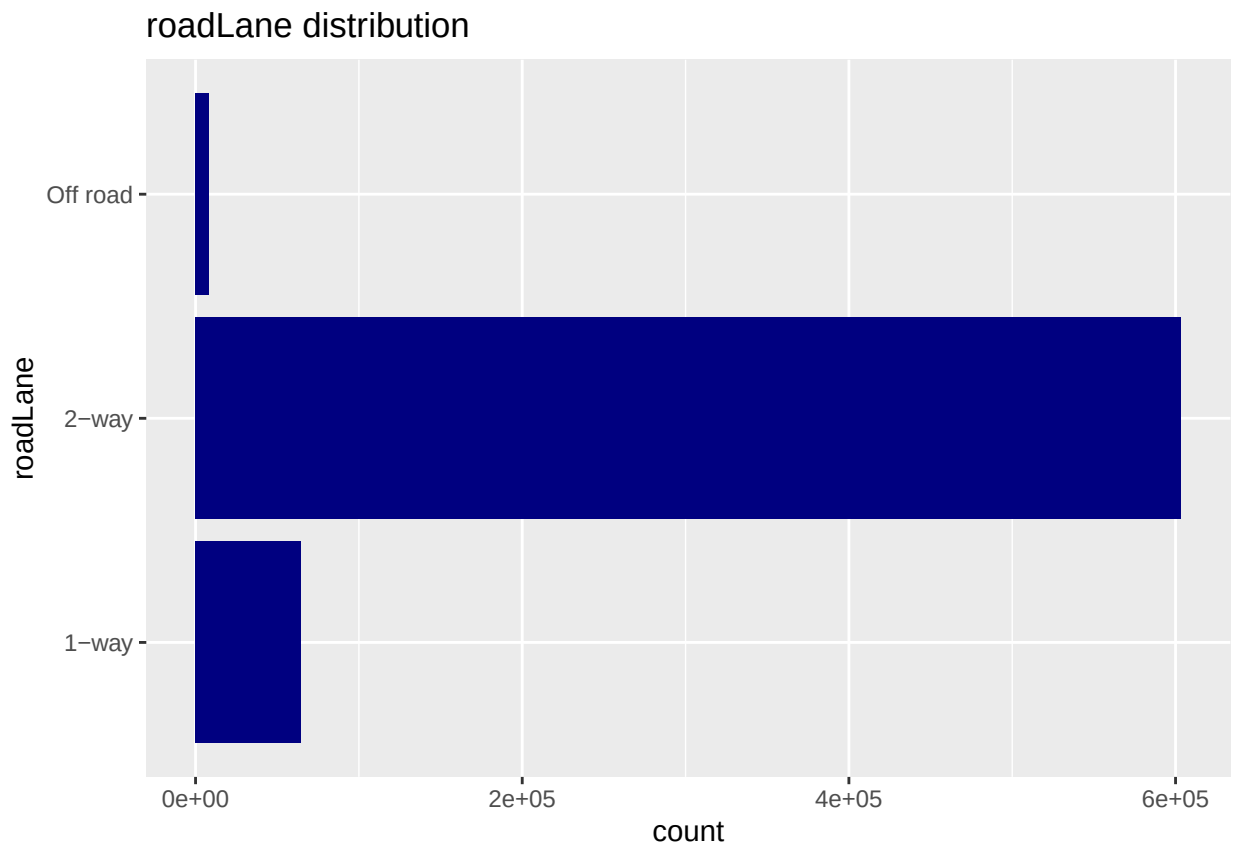
```
ggplot(Train_set, aes(x = region)) +
  geom_bar(fill = "navy") +
  coord_flip() +
  ggtitle("region distribution")
```




The majority of car crashes occur in Auckland.

Bar plot of roadLane:

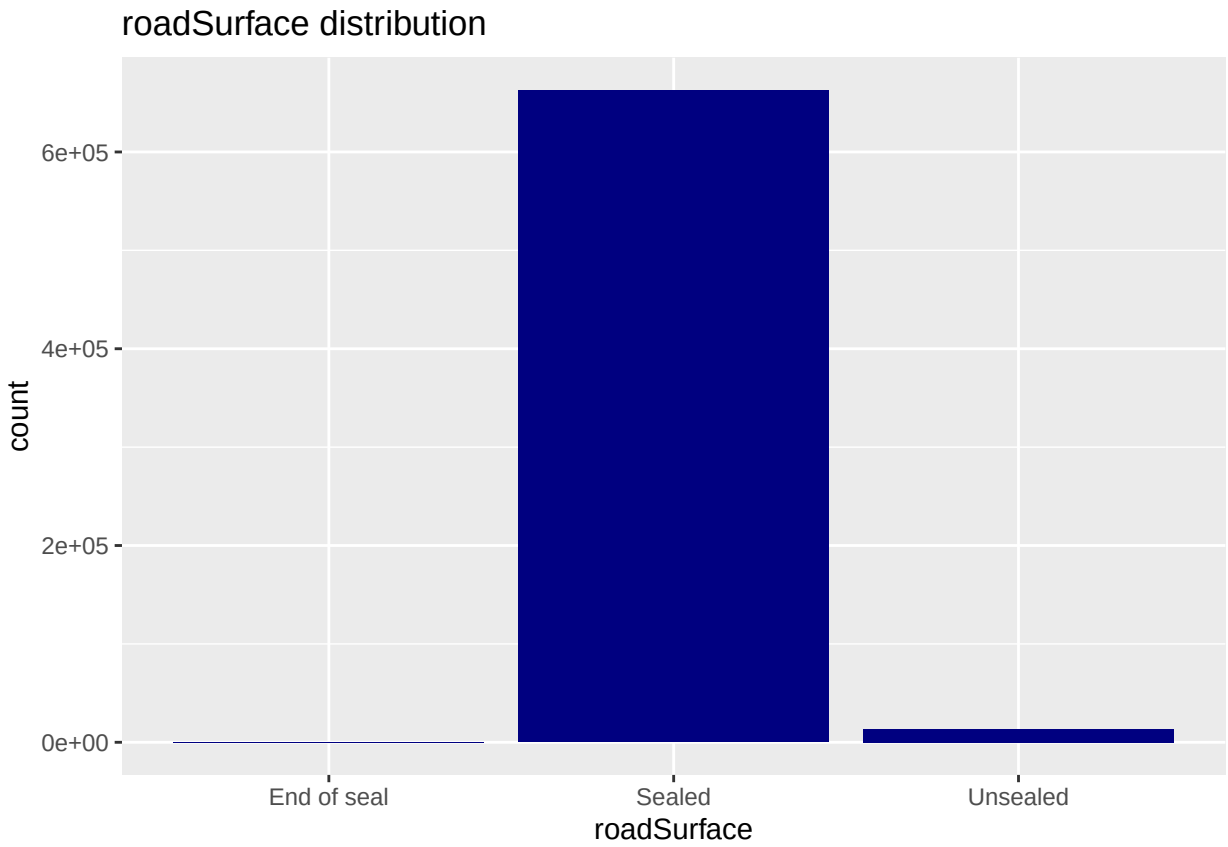
```
ggplot(Train_set, aes(x = roadLane)) +
  geom_bar(fill = "navy") +
  coord_flip() +
  ggtitle("roadLane distribution")
```




Very few off road observations

Bar plot of roadSurface:

```
ggplot(Train_set, aes(x = roadSurface)) +
  geom_bar(fill = "navy") +
  ggtitle("roadSurface distribution")
```



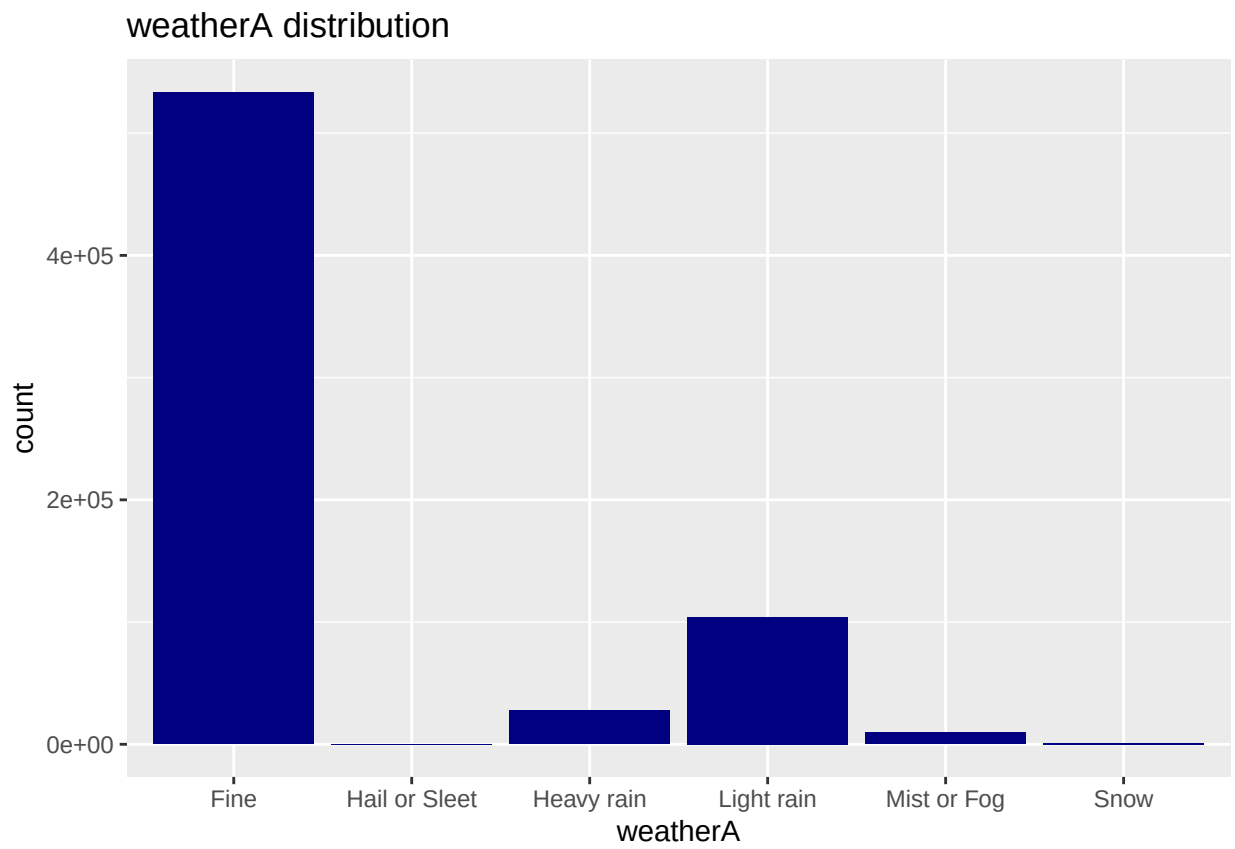

There are very few End of seal observations

```
sum(Train_set$roadSurface == "End of seal")
```

```
## [1] 95
```

Bar plot of weatherA:

```
ggplot(Train_set, aes(x = weatherA)) +
  geom_bar(fill = "navy") +
  ggtitle("weatherA distribution")
```




```
sum(Train_set$weatherA == "Hail or Sleet")
```

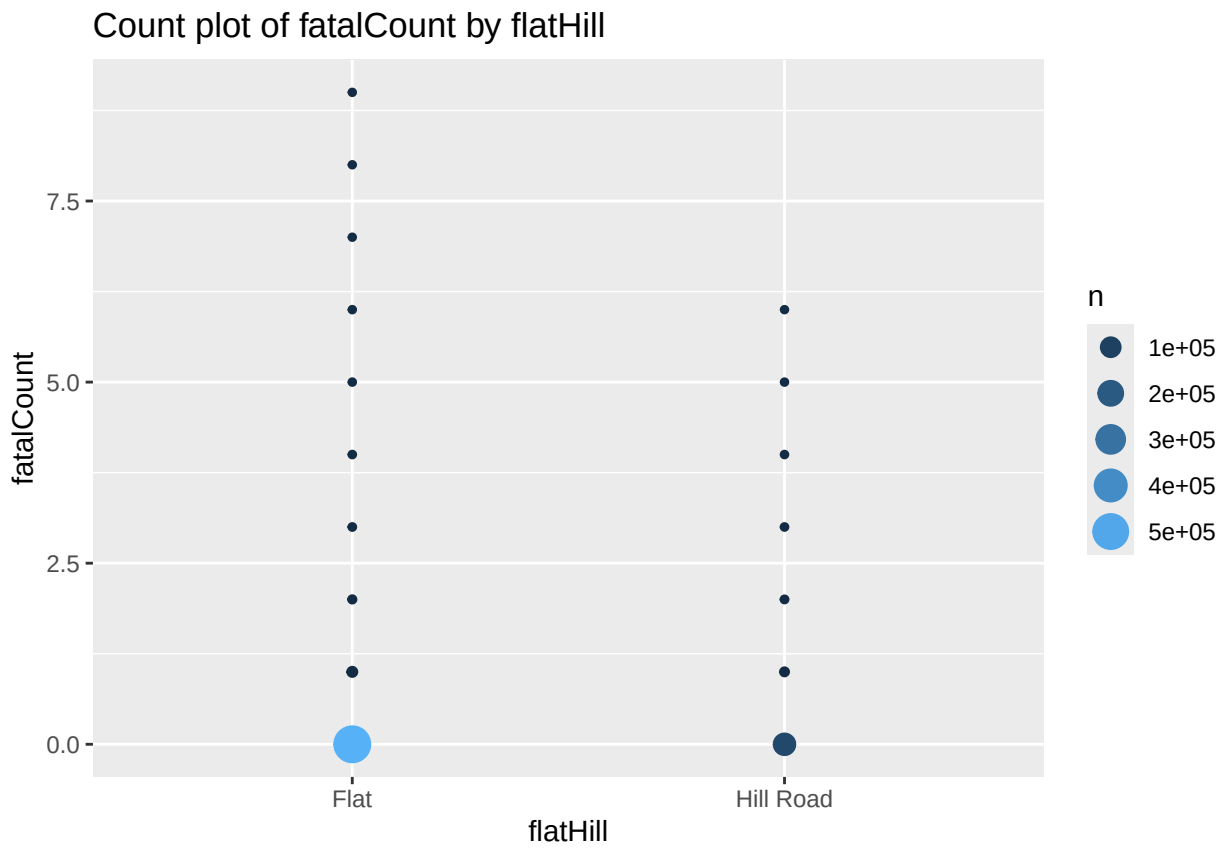
```
## [1] 141
```

There are only 141 Hail or Sleet observations Many of the categorical variables are imbalanced

## Numerical data plots

Count plot of response variable fatalCount:

```
ggplot(Train_set, aes(x = flatHill, y = fatalCount)) +
  geom_count(aes(color = after_stat(n), size = after_stat(n))) +
  guides(color = 'legend') + ggtitle("Count plot of fatalCount by flatHill")
```




There are few observations with at least 1 death.

```
sum(Train_set$fatalCount == 0)
```

```
## [1] 669796
```

```
sum(Train_set$fatalCount >= 1)
```

```
## [1] 6392
```

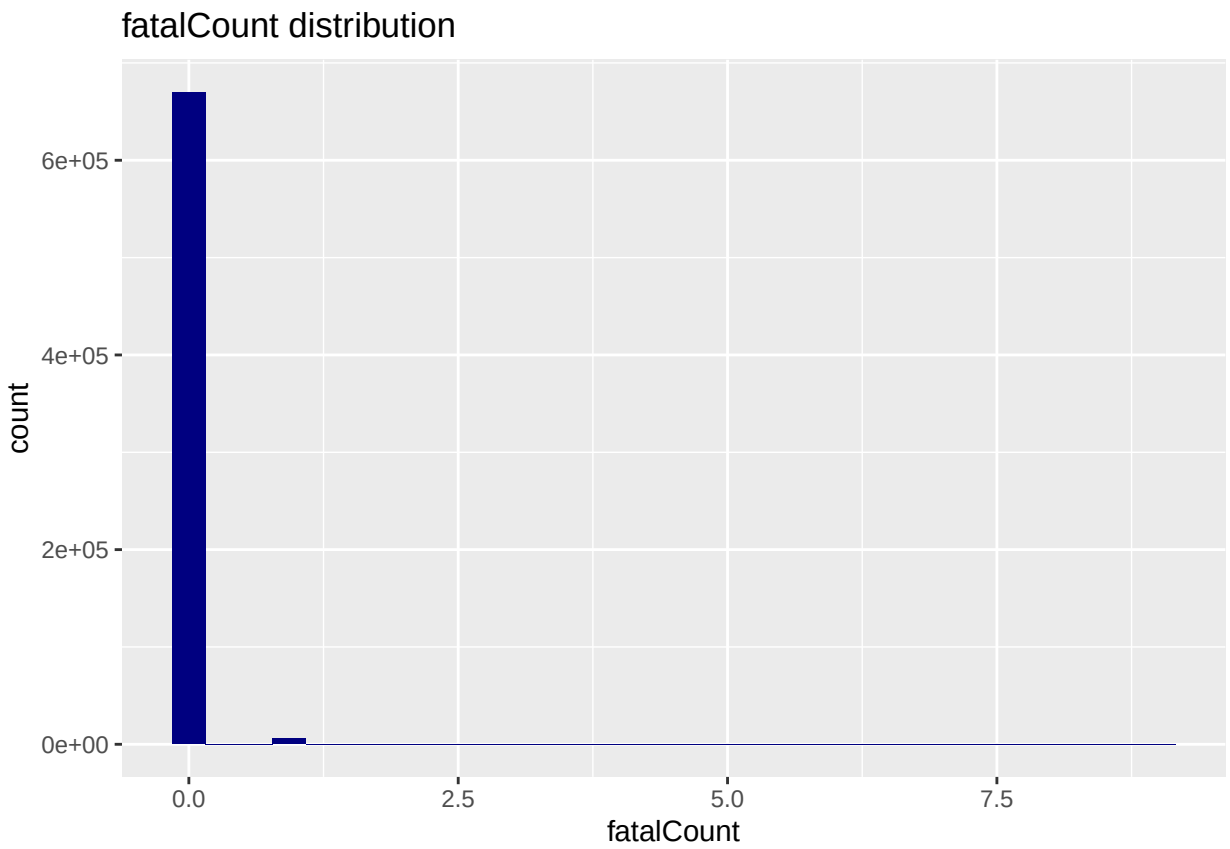
```
round((7968/836997)*100, 2)
```

```
## [1] 0.95
```

Observations with 1 or more deaths make up only 0.95% of observations.

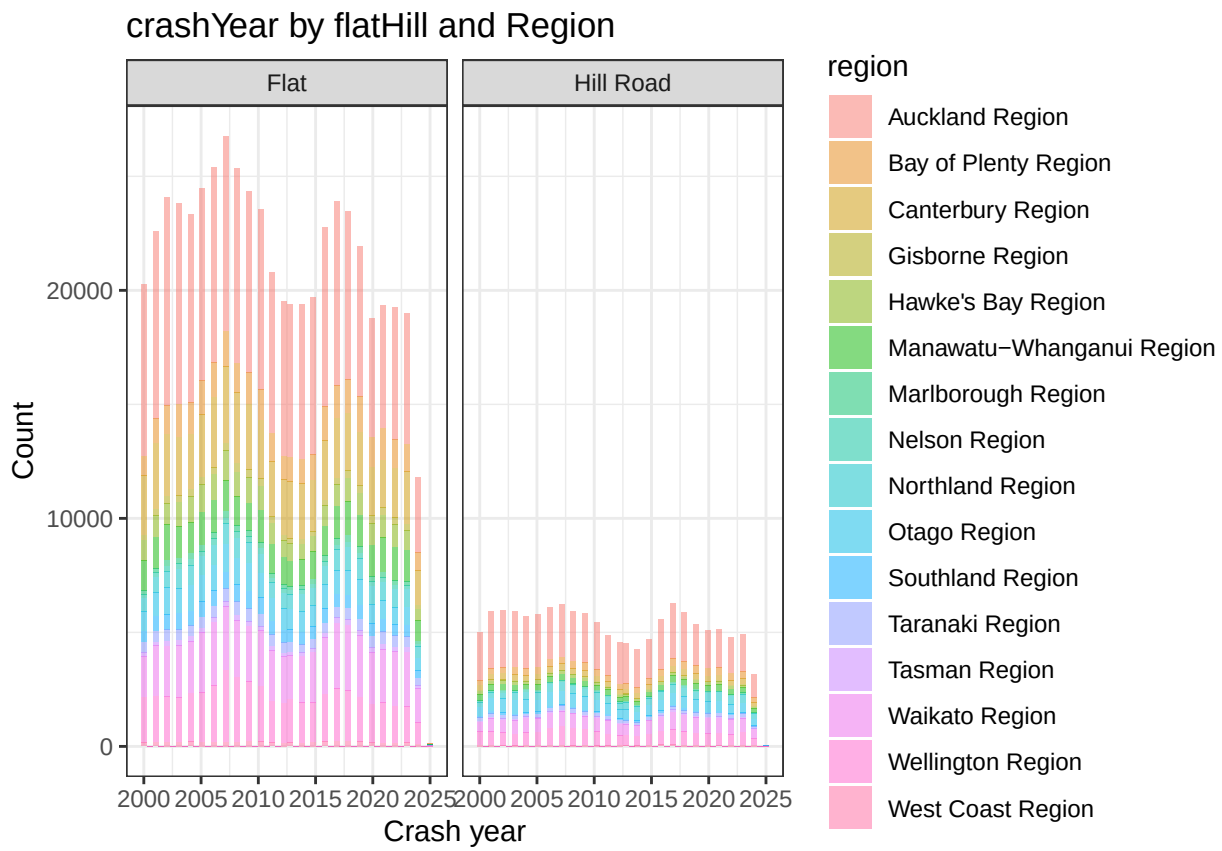
```
ggplot(Train_set, aes(x = fatalCount)) +
  geom_histogram(fill = "navy")  + ggtitle("fatalCount distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



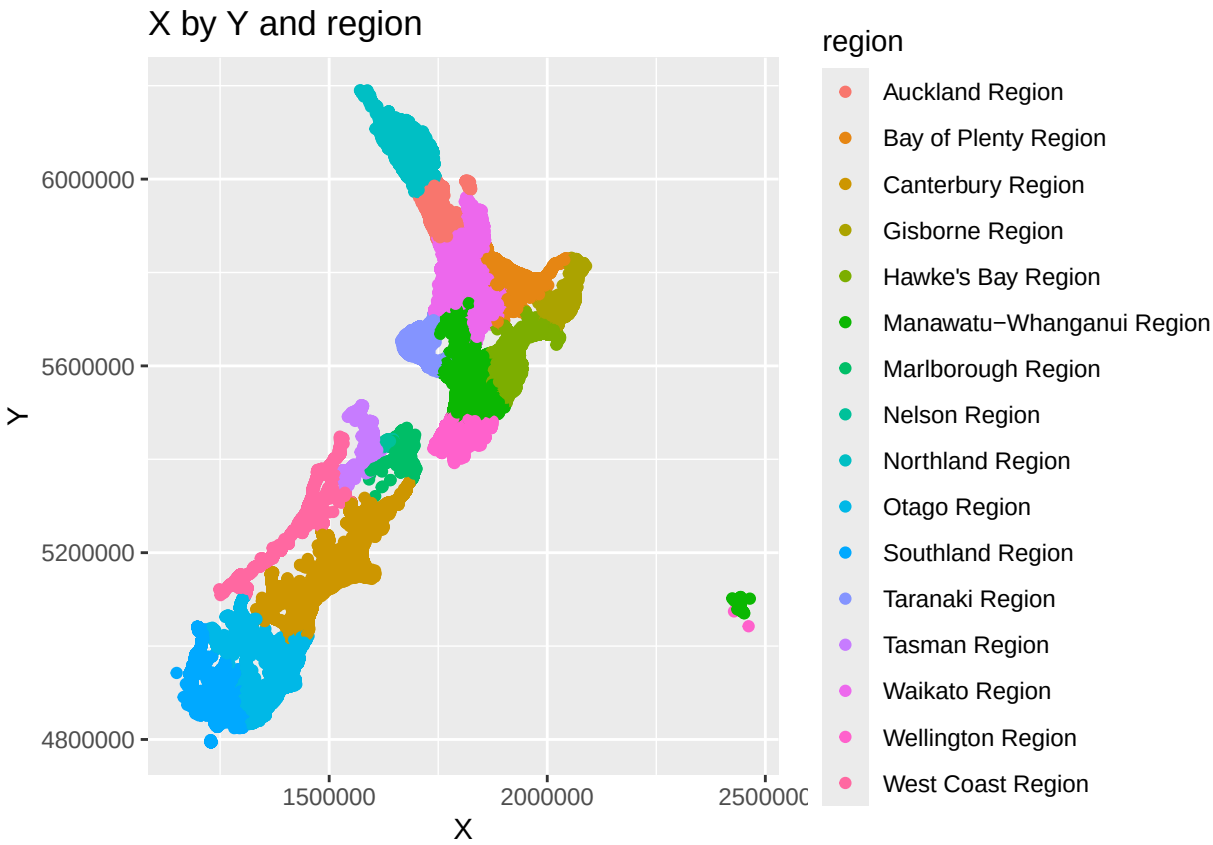

Zero deaths appear to be the vast majority of car crashes, this could potentially indicate zero inflation.

```
ggplot(data = Train_set,
       mapping = aes(x = crashYear, fill = region)) +
  geom_histogram(alpha = 0.5, bins = 50) +
  labs(x = "Crash year", y = "Count",
       title = "crashYear by flatHill and Region") +
  facet_grid(. ~ flatHill) +
  theme_bw()
```




More crashes occur on flat roads than hill roads for every year recorded. But that could be due to there being more flat roads overall in New Zealand than there are hill roads.

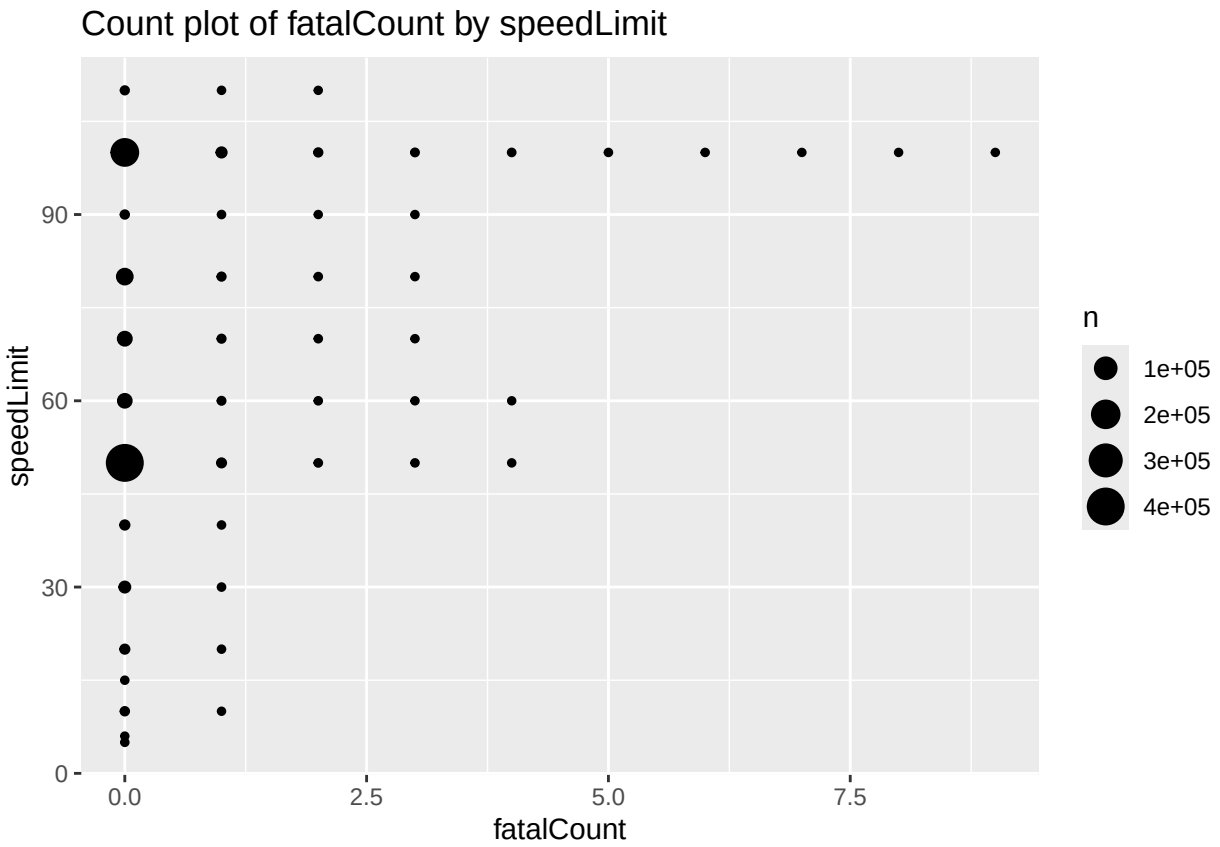
```
ggplot(Train_set, aes(x = X, y = Y, color = region)) +
  geom_point() + ggtitle("X by Y and region")
```




It appears that X and Y contain redundant information that can be modeled by region.

Count plot of response variable fatalCount:

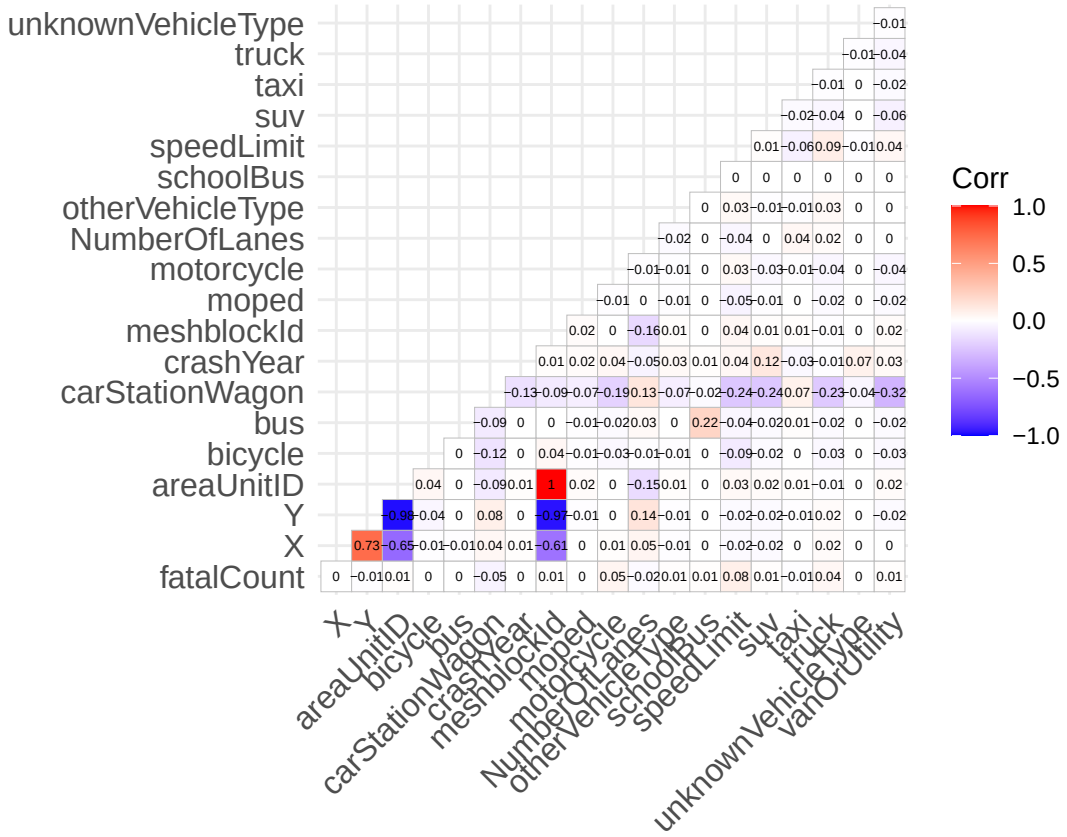
```
ggplot(Train_set, aes(x = fatalCount, y = speedLimit)) +
  geom_count() + ggtitle("Count plot of fatalCount by speedLimit")
```




fatalCount's of 5 and above appear to only occur when speedLimit is above 90

### Correlation plot

```
ggcorrplot(cor(select_if(Train_set, is.numeric)),
method = "square",
lab = TRUE,
lab_size = 1.9,
type = "lower")
```




```
findCorrelation(cor(select_if(Train_set, is.numeric)), cutoff = 0.7, names = TRUE)
```

```
## [1] "areaUnitID" "Y"
```

Y is strongly correlated with areaUnitID and meshblockID with a correlation of -0.98 and -0.97 respectively.

The variable areaUnitID is strongly correlated with meshblockID with a correlation of 1.

Due to X and Y being strongly correlated with a correlation of 0.73 and that they contain redundant information already contained in region, I will be dropping X and Y.

Due to the strong correlation between these variables I will be removing the following variables: areaUnitID and meshblockId

```
Train_set <- Train_set %>% select(-areaUnitID, -meshblockId, -Y, -X)
```

I now have 23 variables in the Crash data set.

## Skewness

```
pander(skewness(select_if(Train_set, is.numeric)))
```

Table 28: Table continues below

| fatalCount | bicycle | bus | carStationWagon | crashYear | moped |
|---|---|---|---|---|---|
| 15.22 | 6.164 | 8.068 | 0.4384 | 0.09674 | 11.91 |

Table 29: Table continues below

| motorcycle | NumberOfLanes | otherVehicleType | schoolBus | speedLimit | suv |
|---|---|---|---|---|---|
| 5.685 | 1.801 | 14.25 | 36.41 | 0.6734 | 3.156 |

| taxi | truck | unknownVehicleType | vanOrUtility |
|---|---|---|---|
| 10.53 | 3.555 | 19.53 | 2.231 |

The response variable fatalCount has a skewness of 15.22 this is significant departure from the normal distribution which has a skewness of 0.

The majority of the numerical variables have a skewness value above 3. Only vanOrUtility, speedLimit, NumberOfLanes, crashYear and carStationWagon aren't skewed.

## Kurtosis

```
pander(kurtosis(select_if(Train_set, is.numeric)))
```

Table 31: Table continues below

| fatalCount | bicycle | bus | carStationWagon | crashYear | moped |
|---|---|---|---|---|---|
| 373.7 | 46.77 | 69.79 | 4.322 | 1.805 | 145.4 |

Table 32: Table continues below

| motorcycle | NumberOfLanes | otherVehicleType | schoolBus | speedLimit | suv |
|---|---|---|---|---|---|
| 45.9 | 7.593 | 216.1 | 1435 | 1.726 | 13.3 |

| taxi | truck | unknownVehicleType | vanOrUtility |
|---|---|---|---|
| 131.4 | 15.89 | 409.3 | 7.696 |

The response variable fatalCount has a kurtosis of 373.7, which means it has a leptokurtic distribution (high peak) this is significant departure from the normal distribution where the absolute kurtosis value should not exceed 7.1.

The variables taxi, motorcycle, bus, moped, otherVehicleType, schoolBus and unknownVehicleType all have extremely high kurtosis values, indicating leptokurtic distributions.

The distributions of most of the variables seem highly skewed with high peaks.

The kurtosis and skewness of fatalCount (the response variable) indicates that the variables distribution deviates significantly from a normal distribution meaning that the assumption of normality has been violated.

Due to the non-normality of the data I will fit a generalized linear model to the data due to its robustness to non-normality (particularly when the data set is large).

Given that the response variable is discrete count data (counting the number of deaths per car crash) I will attempt to fit a poisson model to the data first.

## Feature selection:

Feature importance:

```
roc_imp <- filterVarImp(x = Train_set[,-1], y = Train_set$fatalCount, nonpara = TRUE)
roc_imp <- data.frame(cbind(variable = rownames(roc_imp), score = roc_imp[,1]))
roc_imp$score <- as.double(roc_imp$score)
roc_imp <- roc_imp[order(roc_imp$score,decreasing = TRUE),]
pander(roc_imp)
```

| | variable | score |
|---|---|---|
| **16** | speedLimit | 0.006882 |
| **3** | carStationWagon | 0.002259 |
| **9** | motorcycle | 0.002153 |
| **4** | crashSHDescription | 0.002027 |
| **19** | truck | 0.001489 |
| **12** | region | 0.0005106 |
| **10** | NumberOfLanes | 0.0003955 |
| **13** | roadLane | 0.0003136 |
| **6** | flatHill | 0.0001683 |
| **11** | otherVehicleType | 0.0001206 |
| **21** | vanOrUtility | 6.217e-05 |
| **17** | suv | 5.246e-05 |
| **14** | roadSurface | 4.155e-05 |
| **18** | taxi | 4.123e-05 |
| **15** | schoolBus | 2.595e-05 |
| **8** | moped | 2.271e-05 |
| **20** | unknownVehicleType | 1.684e-05 |
| **5** | crashYear | 8.203e-06 |
| **2** | bus | 6.606e-06 |
| **1** | bicycle | 3.218e-06 |
| **7** | light | 3.751e-07 |
| **22** | weatherA | 1.78e-08 |

Reducing the data set to the ten most important variables:

```
Train_set2 <- Train_set %>% select(fatalCount, roc_imp$variable[1:10])
```

## Fitting Poisson Regression Model

Fitting a poisson regression model:

```
model_glm_P <- glm(fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +
                     NumberOfLanes + otherVehicleType + crashSHDescription +
                     carStationWagon + flatHill, data = Train_set2, family = poisson)
pander(summary(model_glm_P))
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | -8.292 | 0.1086 | -76.37 | 0 |
| **speedLimit** | 0.03301 | 0.0006757 | 48.85 | 0 |
| **motorcycle** | 0.8387 | 0.02356 | 35.6 | 1.497e-277 |
| **roadLane2-way** | 1.458 | 0.07408 | 19.68 | 3.016e-86 |
| **roadLaneOff road** | 1.598 | 0.1603 | 9.971 | 2.038e-23 |
| **truck** | 0.6495 | 0.02798 | 23.22 | 3.132e-119 |
| **regionBay of Plenty Region** | 0.4213 | 0.05298 | 7.953 | 1.822e-15 |
| **regionCanterbury Region** | 0.3667 | 0.048 | 7.639 | 2.186e-14 |
| **regionGisborne Region** | 0.08144 | 0.1046 | 0.7789 | 0.436 |
| **regionHawke's Bay Region** | 0.2337 | 0.06226 | 3.753 | 0.0001745 |
| **regionManawatū-Whanganui Region** | 0.2869 | 0.05307 | 5.407 | 6.404e-08 |
| **regionMarlborough Region** | 0.2424 | 0.1021 | 2.373 | 0.01763 |
| **regionNelson Region** | -0.2285 | 0.1724 | -1.326 | 0.1849 |
| **regionNorthland Region** | 0.3692 | 0.05506 | 6.706 | 2e-11 |
| **regionOtago Region** | -0.1397 | 0.06532 | -2.139 | 0.03248 |
| **regionSouthland Region** | 0.05677 | 0.07678 | 0.7393 | 0.4597 |
| **regionTaranaki Region** | 0.2307 | 0.07406 | 3.114 | 0.001844 |
| **regionTasman Region** | -0.03117 | 0.1052 | -0.2964 | 0.7669 |
| **regionWaikato Region** | 0.3624 | 0.04414 | 8.21 | 2.221e-16 |
| **regionWellington Region** | -0.1169 | 0.06075 | -1.924 | 0.0544 |
| **regionWest Coast Region** | 0.2046 | 0.09325 | 2.195 | 0.0282 |
| **NumberOfLanes** | -0.1654 | 0.0198 | -8.354 | 6.601e-17 |
| **otherVehicleType** | 0.3159 | 0.0949 | 3.328 | 0.0008732 |
| **crashSHDescriptionYes** | 0.1969 | 0.02712 | 7.259 | 3.89e-13 |
| **carStationWagon** | -0.1614 | 0.0183 | -8.82 | 1.145e-18 |
| **flatHillHill Road** | -0.03563 | 0.02762 | -1.29 | 0.1971 |

(Dispersion parameter for poisson family taken to be 1 )

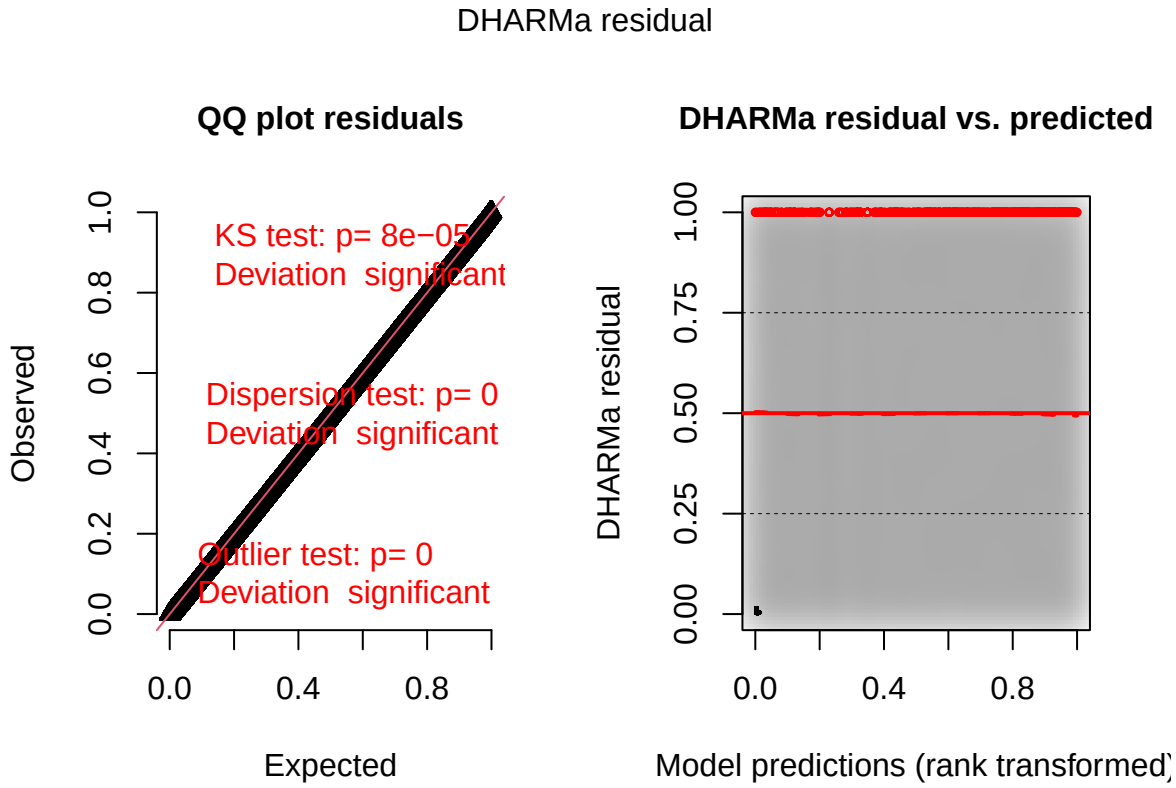
| | |
|---|---|
| Null deviance: | 68460 on 676187 degrees of freedom |
| Residual deviance: | 59759 on 676162 degrees of freedom |

## Checking model assumptions:

I need to check if the response variable fatalCount follows a poisson distribution:

```
simulationOutput <- simulateResiduals(fittedModel = model_glm_P)
plot(simulationOutput)
```

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```




The p-values of the KS test, dispersion test and outlier test are very small, I conclude that the data is not sampled from a poisson distribution.

The dispersion test below also adds evidence to my conclusion:

```
dispersiontest(model_glm_P, alternative = "greater")
```

```
##
##  Overdispersion test
##
## data:  model_glm_P
## z = 11.542, p-value < 2.2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   1.200603
```

The small p-value indicates that the data does not fit a poisson distribution. Overdispersion means the assumptions of the model are not met.

To handle the overdispersion I could fit a quasipoission distribution or a negative binomial distribution to the data instead of a poisson distribution.

Because I want to use AIC or BIC for model selection I will fit a negative binomial model, as quasi-poisson models cannot use AIC or BIC for model selection. This is due to quasi-Poisson models using quasi-likelihood rather than true likelihood.

```
model_nb <- glm.nb(fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +
                     NumberOfLanes + otherVehicleType + crashSHDescription +
                     carStationWagon + flatHill, data = Train_set2)
```

```
BIC(model_glm_P, model_nb)
```

```
##             df      BIC
## model_glm_P 26 73358.36
## model_nb    27 70686.34
```

I can see that the negative binomial model is a better fit to the data than the poisson model according to BIC. Note the degrees of freedom increases because the negative binomial model has a dispersion parameter.

Previously I found that only 0.95% of fatalCount data (the response variable) had a count different from 0.

I am going to test for zero inflation due to this:

```
check_zeroinflation(model_nb)
```

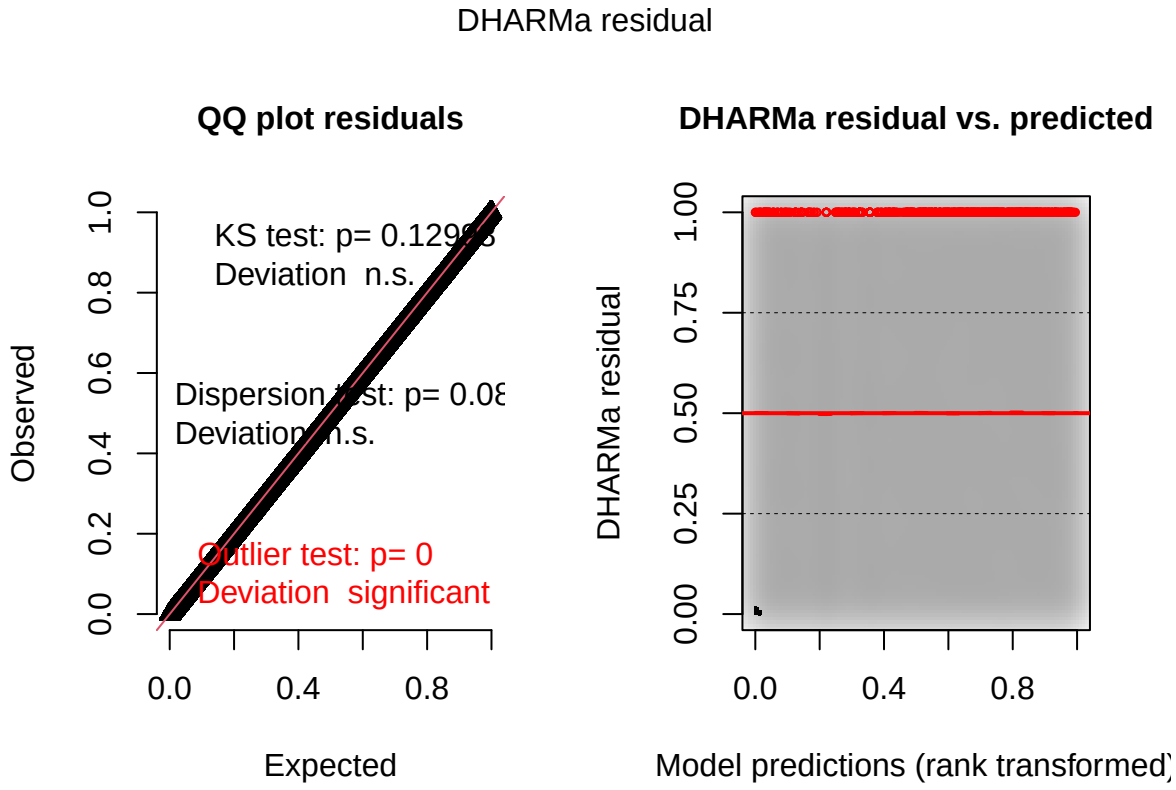
```
## # Check for zero-inflation
##
##    Observed zeros: 669796
##   Predicted zeros: 669804
##             Ratio: 1.00

## Model seems ok, ratio of observed and predicted zeros is within the
##   tolerance range (p = 0.904).
```

The ratio of observed and predicted zeros is within the tolerance range which means that zero inflation is not an issue for the negative binomial model, so there is no need to fit a zero-inflated negative binomial model.

## Residual plots

```
simulationOutput2 <- simulateResiduals(fittedModel = model_nb)
plot(simulationOutput2)
```

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```




The KS test and the dispersion test are not statistically significant at a significance level of 0.05

The outlier test has a p-value of 0 which is concerning, this indicates that there could potentially be outliers in the data which could negatively affect the model by leading to overfitting.

## Diagnostic measures

Is there severe multicollinearity in the data set:

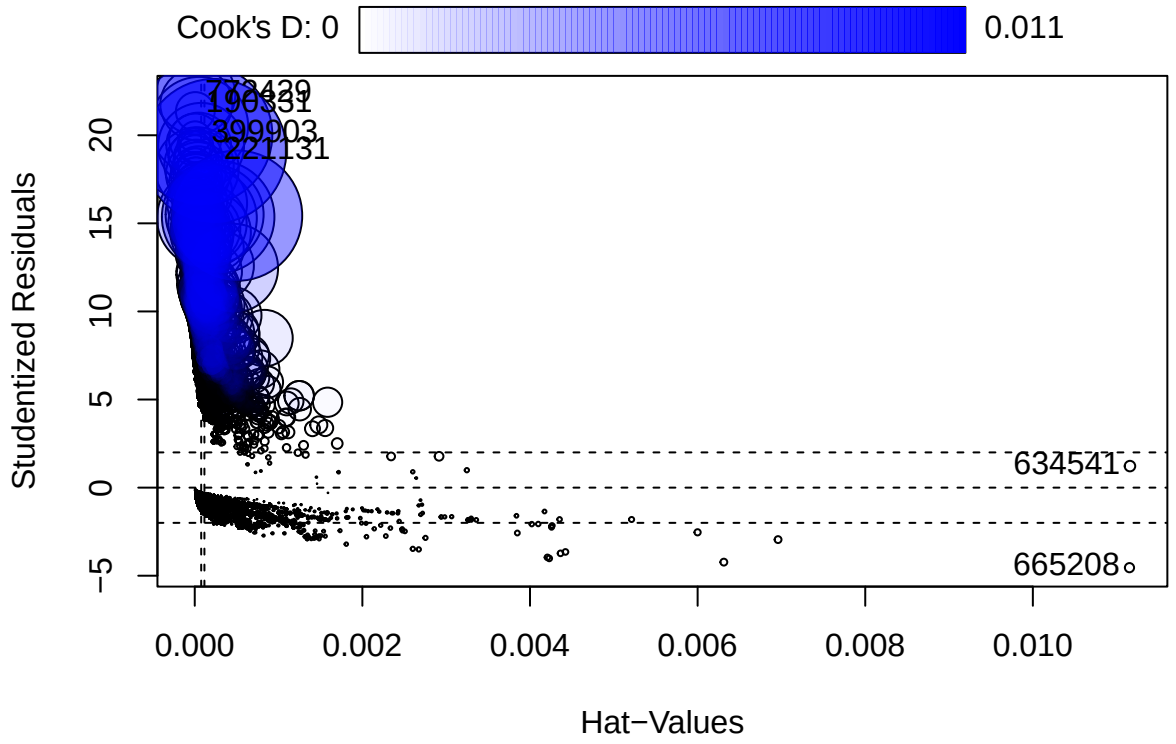
```
pander(vif(model_nb))
```

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| **speedLimit** | 1.466 | 1 | 1.211 |
| **motorcycle** | 1.109 | 1 | 1.053 |
| **roadLane** | 1.225 | 2 | 1.052 |
| **truck** | 1.141 | 1 | 1.068 |
| **region** | 1.396 | 15 | 1.011 |
| **NumberOfLanes** | 1.191 | 1 | 1.091 |
| **otherVehicleType** | 1.009 | 1 | 1.004 |
| **crashSHDescription** | 1.356 | 1 | 1.164 |
| **carStationWagon** | 1.323 | 1 | 1.15 |
| **flatHill** | 1.054 | 1 | 1.027 |

The variance inflation factors (VIF's) are all below 10, meaning there is no evidence of severe multicollinearity of the predictors.

Cook's distance:

```
influencePlot(model_nb)
```




```
##          StudRes          Hat        CookD
## 190331 21.760441 1.934052e-05 2.746457e-03
## 221131 19.106698 2.333606e-04 1.096035e-02
## 399903 20.036303 8.526701e-05 1.047698e-02
## 634541  1.221948 1.115846e-02 5.850745e-05
## 665208 -4.543209 1.115276e-02 4.488761e-05
## 772429 22.296721 1.090614e-05 3.323258e-03
```

There are six oberservations that are influential points according to cook's distance, I will remove them from the data set.

```
Train_set2 <- Train_set2[-c(190331, 221131, 399903, 634541, 665208, 772429),]
```

Finding outliers using the standardized residuals:

```
suppressWarnings({model.diag.metrics <- augment(model_nb)})
model.diag.metrics <- model.diag.metrics %>%
mutate(index = 1:nrow(model.diag.metrics)) %>%
select(index, everything())

large_std_resid <- model.diag.metrics[model.diag.metrics$.std.resid > 3 | model.diag.metrics$.std.resid
nrow(large_std_resid)
```

```
## [1] 158
```

There are 133 observations that have standardized residuals larger than 3 in absolute value.

I am going to remove these from the data set:

```
Train_set2 <- Train_set2[-c(large_std_resid$index), ]
```

## Variable selection

Refitting model using data excluding influential points and outliers:

```
model_nb <- glm.nb(fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +
                     NumberOfLanes + otherVehicleType + crashSHDescription +
                     carStationWagon + flatHill, data = Train_set2)
```

Stepwise BIC:

```
step(model_nb, direction = "both", k = log(nrow(Train_set2)))
```

```
## Start:  AIC=69481.1
## fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +
##     NumberOfLanes + otherVehicleType + crashSHDescription + carStationWagon +
##     flatHill
##
##                       Df Deviance   AIC
## - flatHill             1    36593 69469
## - otherVehicleType     1    36600 69476
## <none>                      36592 69481
## - region              15    36795 69483
## - crashSHDescription   1    36631 69507
## - NumberOfLanes        1    36659 69535
## - carStationWagon      1    36705 69581
## - truck                1    36953 69829
## - roadLane             2    37077 69940
## - motorcycle           1    37147 70022
## - speedLimit           1    38841 71717
##
## Step:  AIC=69469.08
## fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +
##     NumberOfLanes + otherVehicleType + crashSHDescription + carStationWagon
##
##                       Df Deviance   AIC
## - otherVehicleType     1    36590 69464
## <none>                      36582 69469
## - region              15    36789 69475
## + flatHill             1    36580 69481
## - crashSHDescription   1    36622 69495
## - NumberOfLanes        1    36649 69523
## - carStationWagon      1    36694 69568
## - truck                1    36944 69817
```

```
## - roadLane            2    37066 69926
## - motorcycle          1    37136 70009
## - speedLimit          1    38869 71742
##
## Step:  AIC=69464.06
## fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +
##     NumberOfLanes + crashSHDescription + carStationWagon
##
##                       Df Deviance   AIC
## <none>                     36573 69464
## + otherVehicleType    1    36565 69469
## - region             15    36782 69471
## + flatHill            1    36572 69476
## - crashSHDescription  1    36613 69490
## - NumberOfLanes       1    36641 69518
## - carStationWagon     1    36690 69567
## - truck               1    36938 69815
## - roadLane            2    37058 69921
## - motorcycle          1    37125 70002
## - speedLimit          1    38867 71744


##
## Call:  glm.nb(formula = fatalCount ~ speedLimit + motorcycle + roadLane +
##     truck + region + NumberOfLanes + crashSHDescription + carStationWagon,
##     data = Train_set2, init.theta = 0.1200149088, link = log)
##
## Coefficients:
##                      (Intercept)                       speedLimit
##                         -8.26813                          0.03283
##                       motorcycle                     roadLane2-way
##                          1.02148                          1.45975
##                 roadLaneOff road                            truck
##                          1.55933                          0.67296
##      regionBay of Plenty Region          regionCanterbury Region
##                          0.43315                          0.38124
##           regionGisborne Region         regionHawke's Bay Region
##                          0.02220                          0.24440
## regionManawatū-Whanganui Region         regionMarlborough Region
##                          0.30946                          0.27700
##             regionNelson Region           regionNorthland Region
##                         -0.17587                          0.39045
##              regionOtago Region           regionSouthland Region
##                         -0.13589                          0.06196
##           regionTaranaki Region              regionTasman Region
##                          0.23592                         -0.03927
##            regionWaikato Region          regionWellington Region
##                          0.37216                         -0.12532
##         regionWest Coast Region                     NumberOfLanes
##                          0.17971                         -0.17029
##           crashSHDescriptionYes                    carStationWagon
##                          0.19150                         -0.21625
##
## Degrees of Freedom: 676024 Total (i.e. Null);  676001 Residual
## Null Deviance:       44340
```

```
## Residual Deviance: 36570     AIC: 69190
```

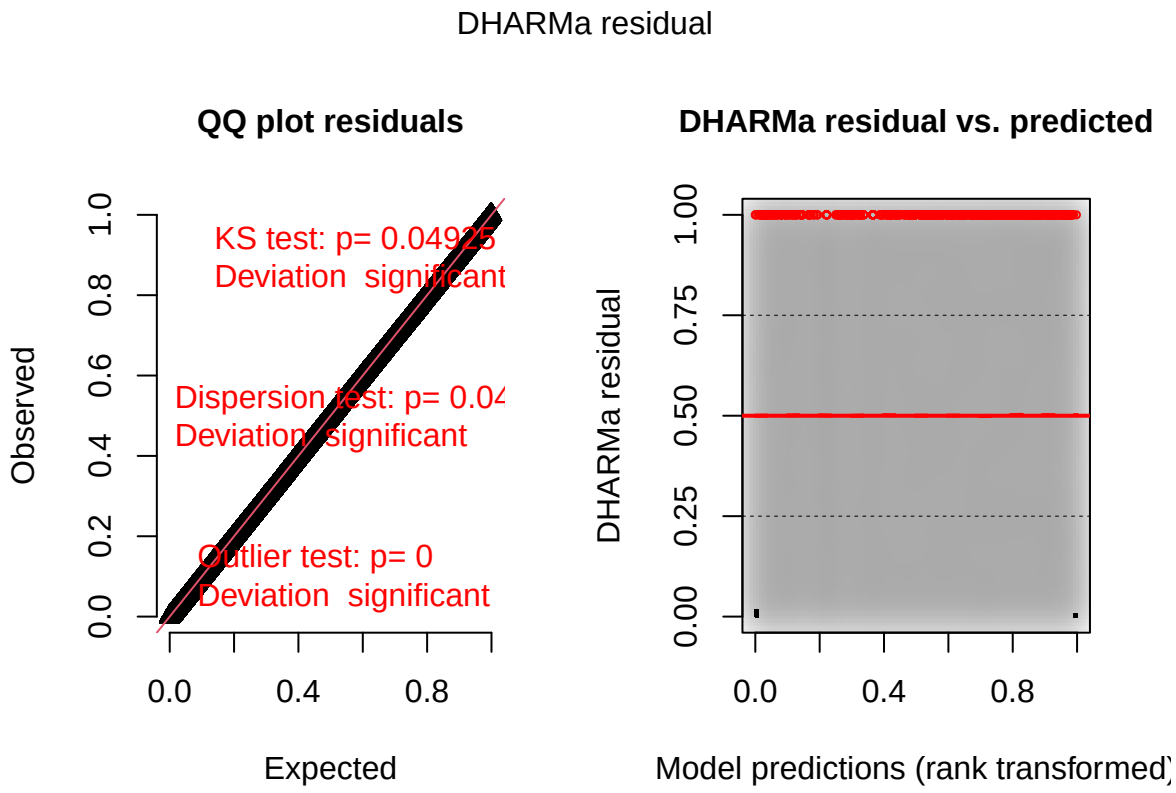
Stepwise BIC initially removed flatHill and otherVehicleType, but then added these variables back into the model. This suggests that the inclusion of these variables improve (decrease) the BIC of the model.

BIC stepwise regression eventually selected all variables already in the model.

## Residual plots

```
simulationOutput2 <- simulateResiduals(fittedModel = model_nb)
plot(simulationOutput2)
```

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```

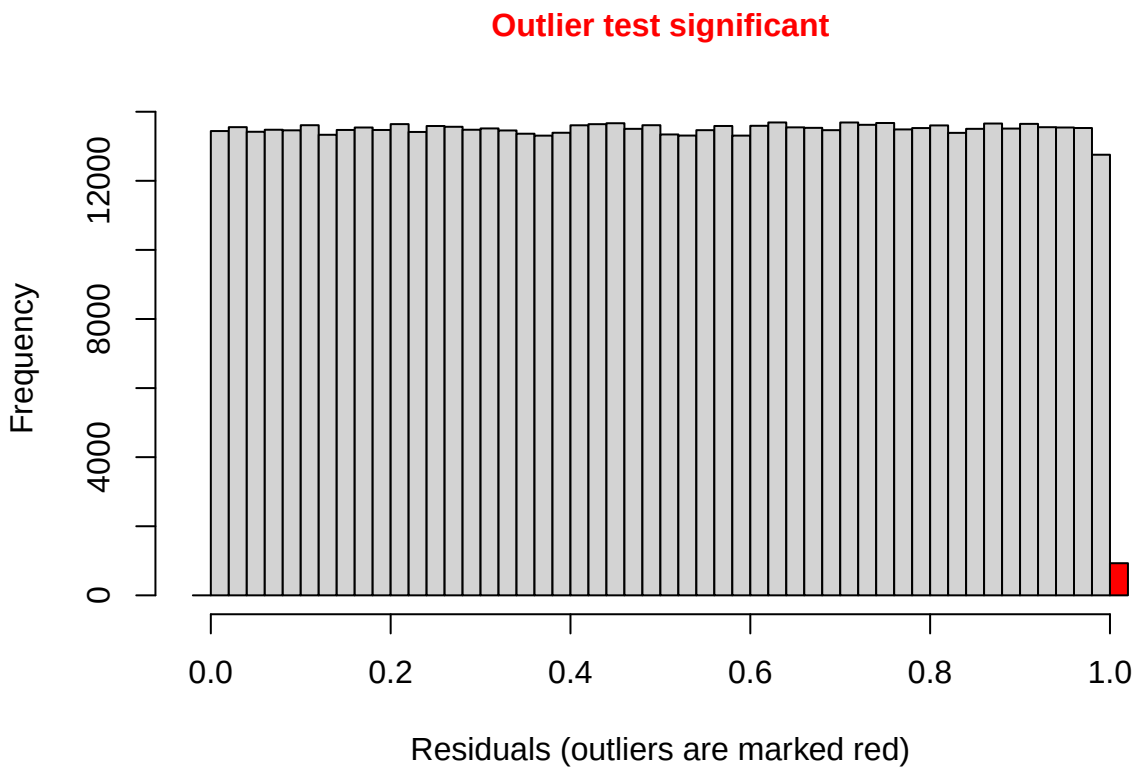



The KS test and dispersion test are statistically significant at a significance level of 0.05, so I conclude that the data does not fit a negative binomial model. I may need to consider other models.

The outlier test has a p-value of 0 which is concerning, this indicates that there could potentially be outliers in the data which could negatively affect the model by leading to overfitting. It's possible that using type = binomial may have inflated the amount of estimated outliers, so I will use a bootstrap method next to determine the amount of outliers.

## Handling Outliers

```
outliers <- testOutliers(simulationOutput2, type = "bootstrap", nBoot = 20)
```




```
outliers
```

```
##
##  DHARMa bootstrapped outlier test
##
## data:  simulationOutput2
## outliers at both margin(s) = 930, observations = 676025, p-value <
## 2.2e-16
## alternative hypothesis: two.sided
##  percent confidence interval:
##  0.001914056 0.002200843
## sample estimates:
## outlier frequency (expected: 0.00201346104064199 )
##                                         0.001375689
```

Outliers according to the DHARMa bootstrapped outlier test are defined as observations that fall outside the range of simulated values, meaning they have scaled residuals of 0 or 1.

There are 930 more outliers than expected in the data set.

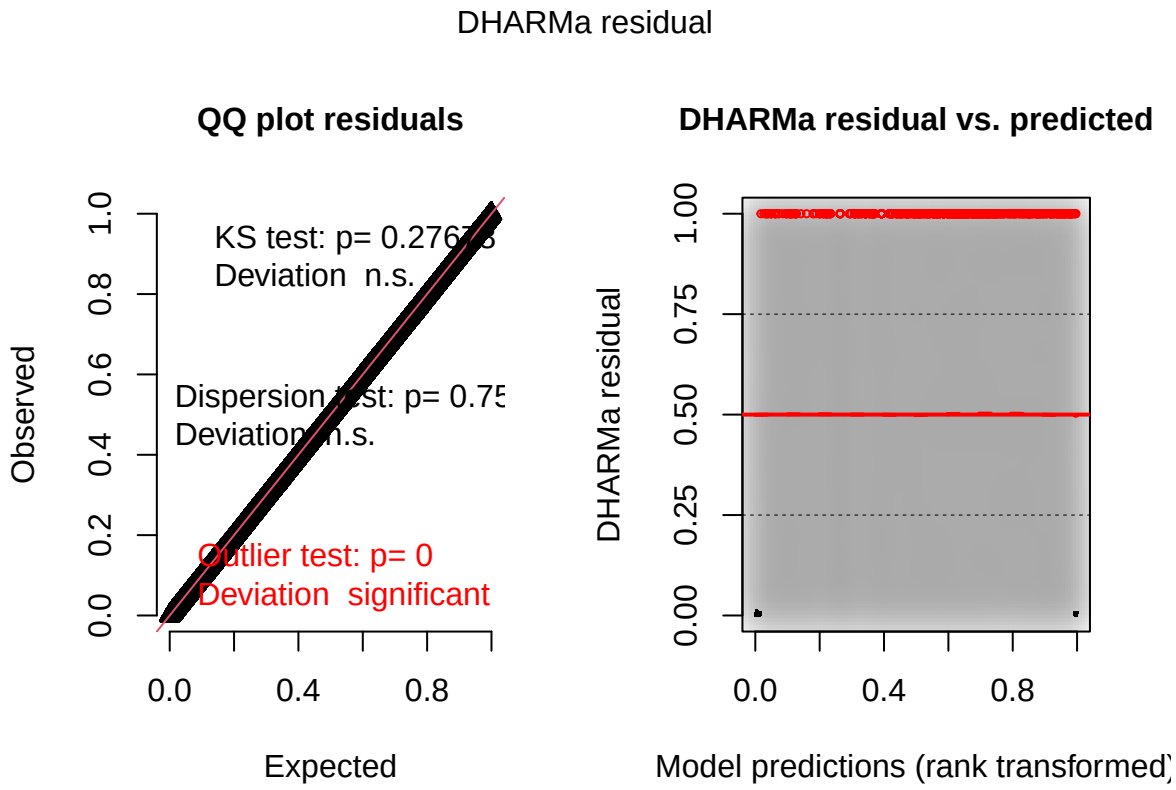
Checking the residual plots to see if removing these outliers fixes the outlier problem:

```
Potential_outliers <- which(simulationOutput2$scaledResiduals == 1 | simulationOutput2$scaledResiduals =
Train_set3 <- Train_set2[-c(Potential_outliers), ]
```

```
model_nb2 <- glm.nb(fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +
                      NumberOfLanes + otherVehicleType + crashSHDescription +
                      carStationWagon + flatHill, data = Train_set3)
```

```
simulationOutput3 <- simulateResiduals(fittedModel = model_nb2)
plot(simulationOutput3)
```

```
## DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-valued dis
```




Removing the outliers identified by the first outlier test did not fix the outlier issue, however it did affect the KS test and Dispersion test.

The KS test now has a non-significant p-value which informs me that the negative binomial distribution does fit the data.

The dispersion test also has a non-significant p-value.

Given this result it seems that model_nb2: the negative binomial model that excludes the outliers identified using the bootstrap method is a better fit to the data than model_nb, I am going to use the MAE, MSE and RMSE to compare the models

## Evaluation of model

Applying the same method for missing data removal on the test set:

```
Test_set2 <- Test_set %>% select(-missing.table$variable[1:32])
Test_set2 <- Test_set2[complete.cases(Test_set2), ]
# making the dependent variable the first column
Test_set2 <- Test_set2 %>% relocate(fatalCount)

# Selecting variables used for the model
Test_set2 <- Test_set2 %>% select(fatalCount, roc_imp$variable[1:10])
```

```
# predicting response data based off model and test data.
prediction <- predict(model_nb, Test_set2, type = "response")
prediction2 <- predict(model_nb2, newdata = Test_set2, type = "response")

MAE <- mae(actual = Test_set2$fatalCount, predicted = prediction)
MAE2 <- mae(actual = Test_set2$fatalCount, predicted = prediction2)
MSE <- mse(actual = Test_set2$fatalCount, predicted = prediction)
MSE2 <- mse(actual = Test_set2$fatalCount, predicted = prediction2)
RMSE <- rmse(actual = Test_set2$fatalCount, predicted = prediction)
RMSE2 <- rmse(actual = Test_set2$fatalCount, predicted = prediction2)

df <- data.frame(. = c("MAE", "MSE", "RMSE"),
                 model_nb = c(MAE, MSE, RMSE),
                model_nb2 = c(MAE2, MSE2, RMSE2))

pander(df)
```

| . | model_nb | model_nb2 |
|---|---|---|
| MAE | 0.02047 | 0.01837 |
| MSE | 0.01311 | 0.01319 |
| RMSE | 0.1145 | 0.1148 |

The mean absolute error is better for model_nb2 than model_nb.

The root mean squared error is worse for model_nb than model_nb2.

The improvement in MAE but worse RMSE for model_nb2 informs me that model_nb2 is making more large scale errors and fewer small scale errors than model model_nb.

Considering that the difference in RMSE is 0.0003 between the models, and the previous KS and dispersion tests on the models, I will move forward with model_nb2.

Interpretation of MAE for model_nb2:

MAE: On average the models predictions are around 0.01837 deaths away from the true death counts for a car crash.

## Interpretting model coefficients

```
pander(summary(model_nb2))
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | -9.116 | 0.1414 | -64.46 | 0 |
| **speedLimit** | 0.03749 | 0.0008166 | 45.91 | 0 |
| **motorcycle** | 1.13 | 0.034 | 33.24 | 3.358e-242 |
| **roadLane2-way** | 1.825 | 0.1019 | 17.91 | 1.058e-71 |
| **roadLaneOff road** | 1.829 | 0.1985 | 9.215 | 3.124e-20 |
| **truck** | 0.7649 | 0.0323 | 23.68 | 5.653e-124 |
| **regionBay of Plenty Region** | 0.4855 | 0.06277 | 7.734 | 1.039e-14 |
| **regionCanterbury Region** | 0.3924 | 0.05757 | 6.816 | 9.359e-12 |
| **regionGisborne Region** | -0.03194 | 0.1278 | -0.25 | 0.8026 |
| **regionHawke's Bay Region** | 0.2978 | 0.07254 | 4.105 | 4.041e-05 |
| **regionManawatū-Whanganui Region** | 0.34 | 0.06238 | 5.451 | 5.019e-08 |
| **regionMarlborough Region** | 0.2019 | 0.1222 | 1.652 | 0.09853 |
| **regionNelson Region** | -0.3173 | 0.2153 | -1.474 | 0.1405 |
| **regionNorthland Region** | 0.4413 | 0.06423 | 6.87 | 6.423e-12 |
| **regionOtago Region** | -0.1847 | 0.07874 | -2.346 | 0.01898 |
| **regionSouthland Region** | 0.1297 | 0.08776 | 1.477 | 0.1396 |
| **regionTaranaki Region** | 0.2836 | 0.08585 | 3.303 | 0.0009549 |
| **regionTasman Region** | 0.09087 | 0.115 | 0.7904 | 0.4293 |
| **regionWaikato Region** | 0.3702 | 0.05304 | 6.98 | 2.944e-12 |
| **regionWellington Region** | -0.1627 | 0.07509 | -2.166 | 0.0303 |
| **regionWest Coast Region** | 0.2578 | 0.1059 | 2.434 | 0.01494 |
| **NumberOfLanes** | -0.2502 | 0.02549 | -9.816 | 9.639e-23 |
| **otherVehicleType** | 0.4191 | 0.1048 | 4 | 6.321e-05 |
| **crashSHDescriptionYes** | 0.1618 | 0.03128 | 5.174 | 2.291e-07 |
| **carStationWagon** | -0.227 | 0.02228 | -10.19 | 2.202e-24 |
| **flatHillHill Road** | -0.06764 | 0.03219 | -2.101 | 0.0356 |

(Dispersion parameter for Negative Binomial(0.4147) family taken to be 1 )

| | |
|---|---|
| Null deviance: | 46066 on 675094 degrees of freedom |
| Residual deviance: | 37441 on 675069 degrees of freedom |

Interpretation of coefficients:

For a one unit change in the speedLimit, the log of expected counts of fatalCount changes by 0.03749, given that the other predictor variables in the model are held constant. This change is statistically significant at a significance level of 0.05

For a one unit change in the motorcycle, the log of expected counts of fatalCount changes by 1.13, given that the other predictor variables in the model are held constant. This change is statistically significant at a significance level of 0.05

The expected log count for 2-way road lane is 1.825 higher than the expected log count for a 1-way road lane, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

The expected log count for Off road lane is 1.829 higher than the expected log count for a 1-way road lane, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

For a one unit change in the truck, the log of expected counts of fatalCount changes by 0.7649, given that the other predictor variables in the model are held constant. This change is statistically significant at a significance level of 0.05

The expected log count for the Bay of Plenty region is 0.4855 higher than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

The expected log count for the Canterbury region is 0.3924 higher than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

The expected log count for the Gisborne region is 0.03194 lower than the expected log count for the Auckland region given that the other predictor variables in the model are held constant. This is not statistically significant at a significance level of 0.05

The expected log count for the Hawke's Bay region is 0.2978 higher than the expected log count for the Auckland region given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

The expected log count for the Manawatū-Whanganui region is 0.34 higher than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

The expected log count for the Marlborough region is 0.2019 higher than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is not statistically significant at a significance level of 0.05

The expected log count for the Nelson region is 0.3173 lower than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is not statistically significant at a significance level of 0.05

The expected log count for the Northland region is 0.4413 higher than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

The expected log count for the Otago region is 0.1847 lower than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

The expected log count for the Southland region is 0.1297 higher than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is not statistically significant at a significance level of 0.05

The expected log count for the Taranaki region is 0.2836 higher than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

The expected log count for the Tasman region is 0.09087 higher than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is not statistically significant at a significance level of 0.05

The expected log count for the Waikato region is 0.3702 higher than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

The expected log count for the Wellington region is 0.1627 lower than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

The expected log count for the West Coast region is 0.2578 higher than the expected log count for the Auckland region, given that the other predictor variables in the model are held constant. This is statistically significant at a significance level of 0.05

For a one unit change in the NumberOfLanes, the log of expected counts of fatalCount changes by -2502, given that the other predictor variables in the model are held constant. This change is statistically significant at a significance level of 0.05

For a one unit change in the otherVehicleType, the log of expected counts of fatalCount changes by 0.4191, given that the other predictor variables in the model are held constant. This change is statistically significant at a significance level of 0.05

For a one unit change in the crashSHDescriptionYes, the log of expected counts of fatalCount changes by 0.1618, given that the other predictor variables in the model are held constant. This change is statistically significant at a significance level of 0.05

For a one unit change in the carStationWagon, the log of expected counts of fatalCount changes by -0.227, given that the other predictor variables in the model are held constant. This change is statistically significant at a significance level of 0.05

For a one unit change in the flatHillHill, the log of expected counts of fatalCount changes by -0.06764, given that the other predictor variables in the model are held constant. This change is not statistically significant at significance level of 0.05

# Conclusion

The coefficients of the the regions Gisborne, Marlborough, Nelson, Southland and Tasman aren't statistically significant which means that there is no statistically significant difference in the count of deaths for a car crash (fatalCount) between those regions and the reference level Auckland. However other regions have a statistically significant difference in the count of deaths for a car crash when compared to Auckland.

The variable flatHill is statistically insignificant however I will not be removing this variable from the data due to the fact that stepwise BIC keeps this variable in the model.

The variables are on extremely different scales, which could be affecting the specific values of regression coefficients, however this does not affect the statistical significance or interpretation of the coefficients.

Each observation in this data set could have a different number of people in the car at the time of the crash which will affect the total number of possible fatalities for that car crash, this is difficult to account for. Unfortunately the total number of people in the car at the time of the crash for each observation is unreported and I cannot determine a way to derive this variable from the variables in the data set. If the number of people in the car for each observation had been recorded then I would have treated this variable as an exposure variable and I would have used this variable as an offset in my model.

I started this project with the intent of understanding the relationships between fatalCount (the count of fatalities associated with a crash) and the other variables in the data set. I have fit a model that provides information about the relationship each variable has with the response (whether it is negative or positive, and whether it is statistically significant). I also found that the data follows a negative binomial distribution.