

# 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 discrepancies 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)
```

## 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
<b>X</b>	1721397	1757428	89494
<b>Y</b>	5644547	5802445	479769
<b>advisorySpeed</b>	54.21	55	25
<b>areaUnitID</b>	546282	536651	53813
bicycle	0.02873	0	0
bridge	0.0137	0	0
bus	0.01593	0	0
<b>carStationWagon</b>	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
<b>houseOrBuilding</b>	0.02353	0	0
kerb	0.03547	0	0
<b>meshblockId</b>	1351845	1177900	1525399
moped	0.007022	0	0
motorcycle	0.03711	0	0
<b>NumberOfLanes</b>	2.332	2	0
<b>objectThrownOrDropped</b>	0.002169	0	0
otherObject	0.02378	0	0
<b>otherVehicleType</b>	0.005422	0	0
overBank	0.0422	0	0
<b>parkedVehicle</b>	0.2575	0	0
pedestrian	1.044	1	0
<b>phoneBoxEtc</b>	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
<b>strayAnimal</b>	0.004231	0	0
suv	0.107	0	0
taxi	0.01064	0	0
<b>temporarySpeedLimit</b>	45.33	40	20
trafficIsland	0.0288	0	0
<b>trafficSign</b>	0.04928	0	0
train	0.001449	0	0
tree	0.1029	0	0
truck	0.0799	0	0
<b>unknownVehicleType</b>	0.003839	0	0
vanOrUtility	0.1783	0	0
vehicle	0.02407	0	0
<b>waterRiver</b>	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
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. :0.000000	Min. :0.000	Min. :0
1st Qu.:0.000000	1st Qu.:0.000000	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.:0.000000	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
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
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

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, sli-pOrFlood, 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:

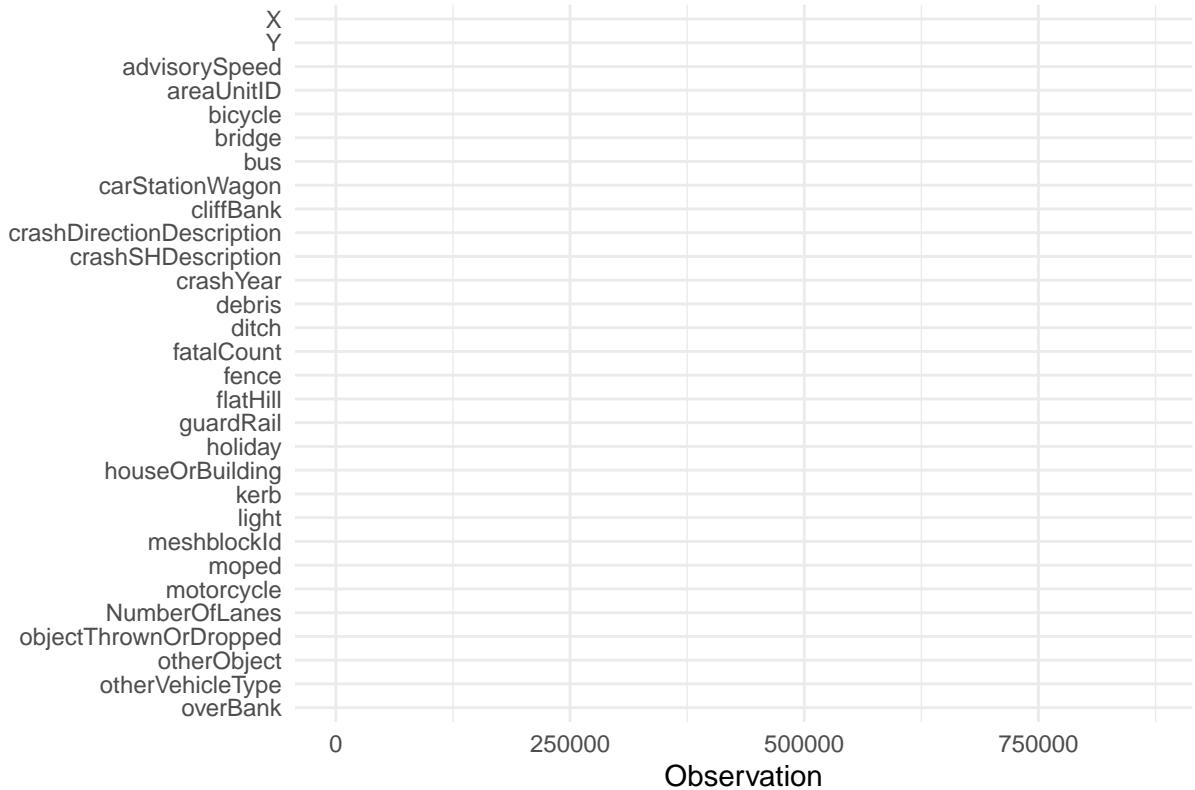
```
Crash$crashSHDescription <- as.factor(Crash$crashSHDescription)
Crash$flatHill <- as.factor(Crash$flatHill)
Crash$holiday <- as.factor(Crash$holiday)
Crash$light <- as.factor(Crash$light)
Crash$region <- as.factor(Crash$region)
Crash$roadCharacter <- as.factor(Crash$roadCharacter)
Crash$roadLane <- as.factor(Crash$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)
```

## 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.

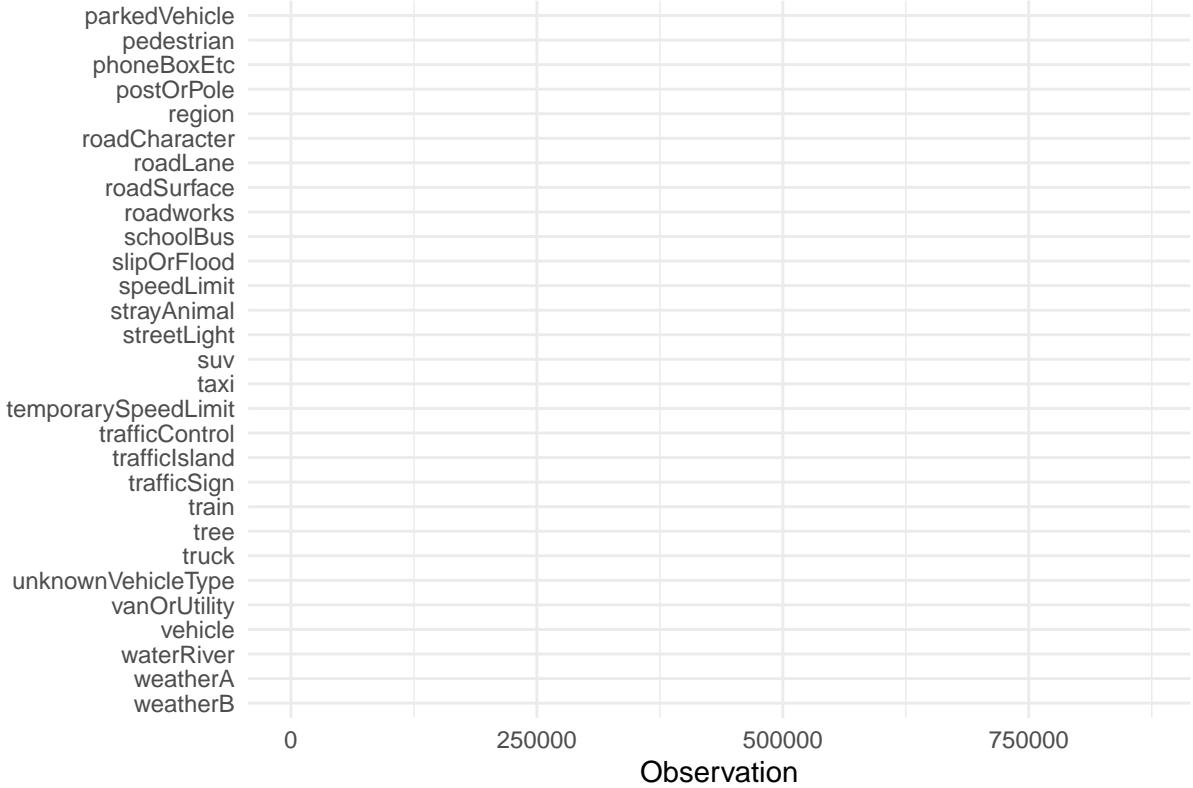
```
# Producing a missing data plot for the data frame.
missing_plot(Crash[1:30], title = "Missing data by observation and variable")
```

## Missing data by observation and variable



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

## Missing data by observation and variable



```
# Saving plots:
# plotA <- missing_plot(Crash[,1:30], title = "Missing data by observation and variable")
# plotB <- missing_plot(Crash[,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 advisorySpeed 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(Crash)){
  if (sum(is.na(Crash[[i]])) > 0){
    percentages <- append(percentages, prop_miss(Crash[[i]]))
    missing_percents_colnames <- append(missing_percents_colnames, colnames(Crash[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	98.35%
## 2	weatherB	96.79%
## 3	pedestrian	96.71%
## 4	advisorySpeed	96.10%
## 5	roadCharacter	95.99%
## 6	holiday	94.49%
## 7	trafficControl	66.31%
## 8	waterRiver	59.02%
## 9	vehicle	59.02%
## 10	tree	59.02%
## 11	train	59.02%
## 12	trafficSign	59.02%
## 13	trafficIsland	59.02%
## 14	strayAnimal	59.02%
## 15	slipOrFlood	59.02%
## 16	roadworks	59.02%
## 17	postOrPole	59.02%
## 18	phoneBoxEtc	59.02%
## 19	parkedVehicle	59.02%
## 20	overBank	59.02%
## 21	otherObject	59.02%
## 22	objectThrownOrDropped	59.02%
## 23	kerb	59.02%
## 24	houseOrBuilding	59.02%
## 25	guardRail	59.02%
## 26	fence	59.02%
## 27	ditch	59.02%
## 28	debris	59.02%
## 29	cliffBank	59.02%
## 30	bridge	59.02%
## 31	crashDirectionDescription	37.61%
## 32	streetLight	34.45%
## 33	weatherA	1.85%

```

## 34           light      0.96%
## 35           flatHill   0.74%
## 36           region     0.38%
## 37           NumberOfLanes 0.24%
## 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
Crash <- Crash %>% select(-missing.table$variable[1:32])

```

There are now have 27 variables in the data set.

Reducing the data down to complete cases only:

```

Crash <- Crash[complete.cases(Crash), ]
# making the dependent variable the first column
Crash <- Crash %>% 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

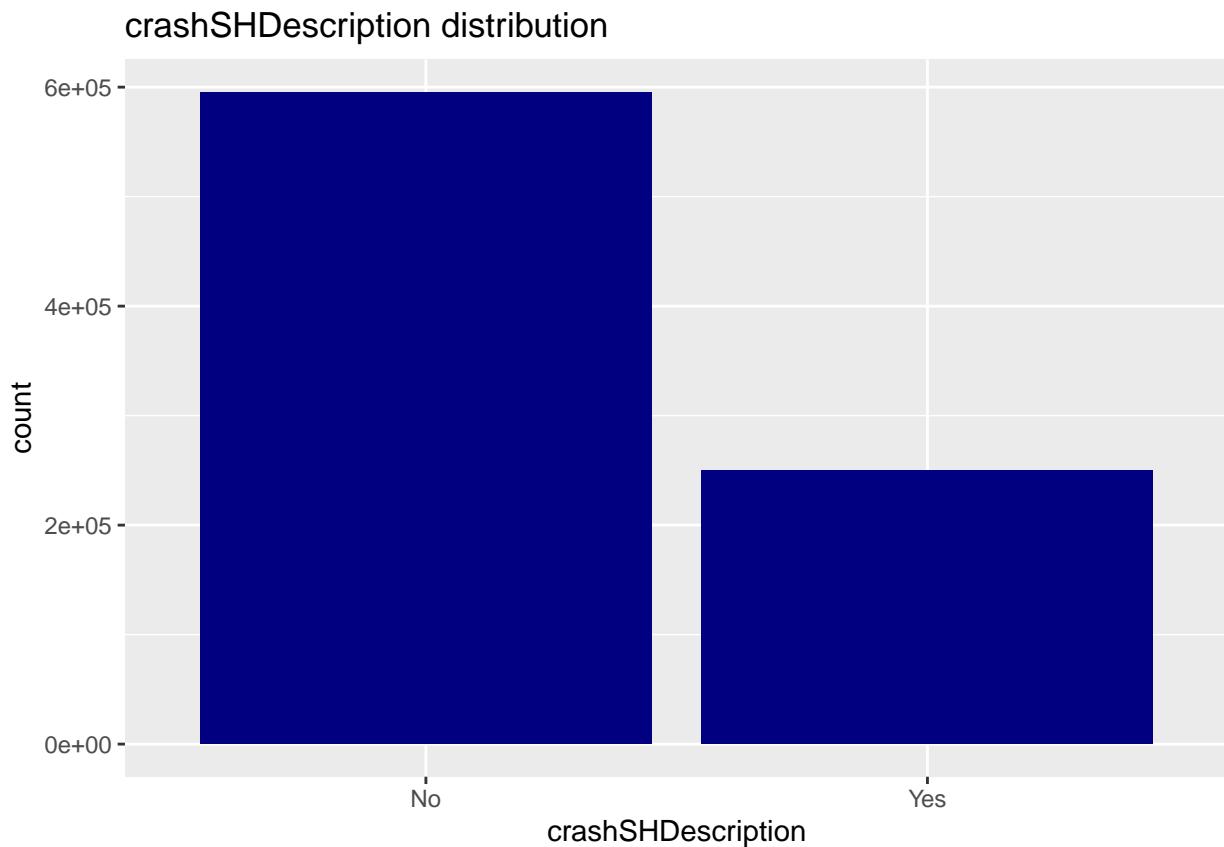
All categorical variables:

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

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

Bar plot of crashSHDescription:

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

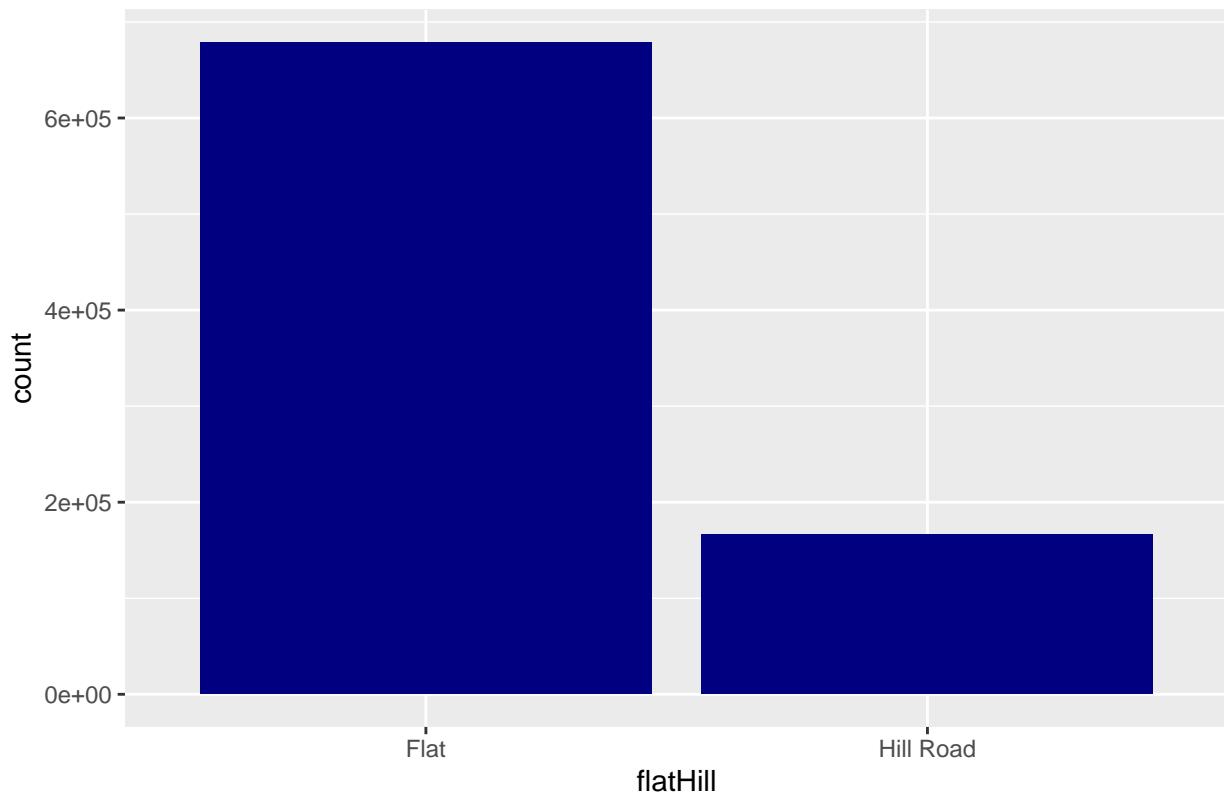


There are less twilight observations (around 500,000).

Bar plot of flatHill:

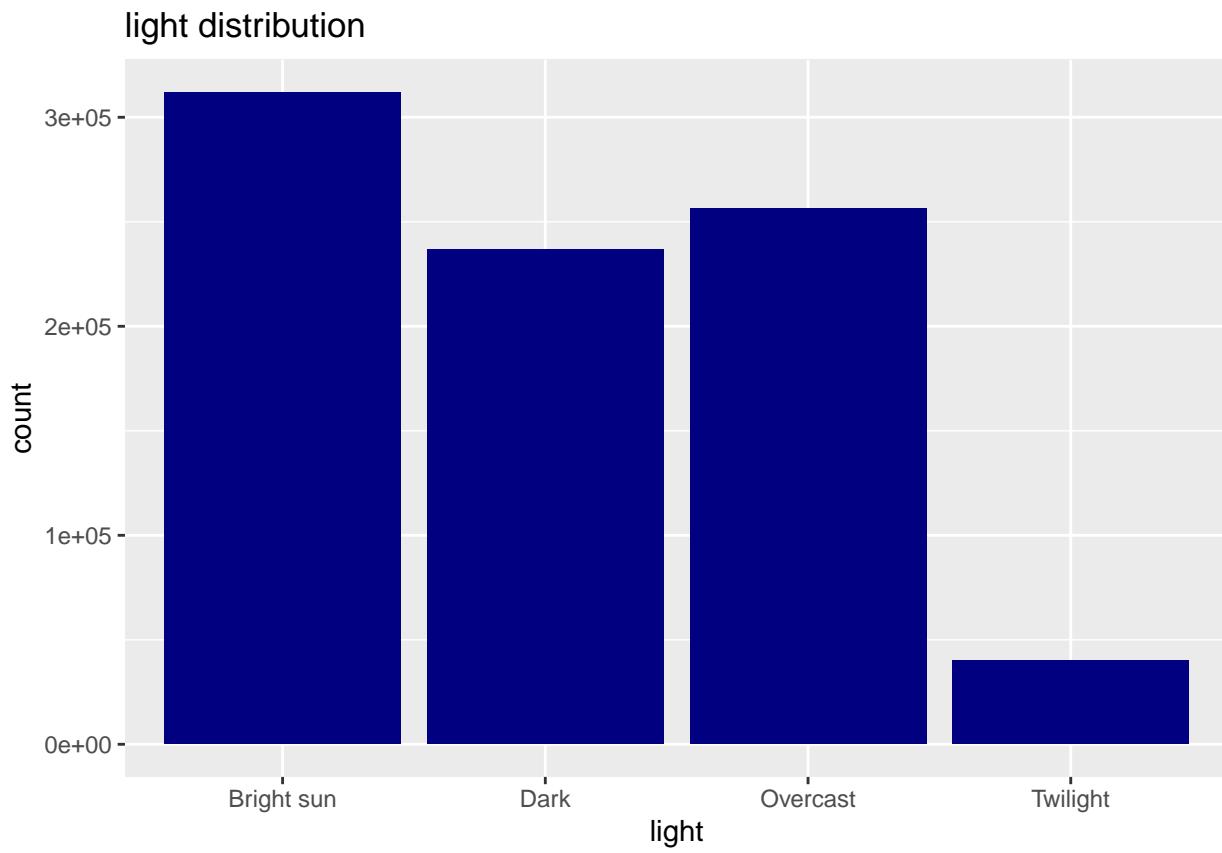
```
ggplot(Crash, aes(x = flatHill)) +
  geom_bar(fill = "navy") +
  ggtitle("flatHill distribution")
```

### flatHill distribution



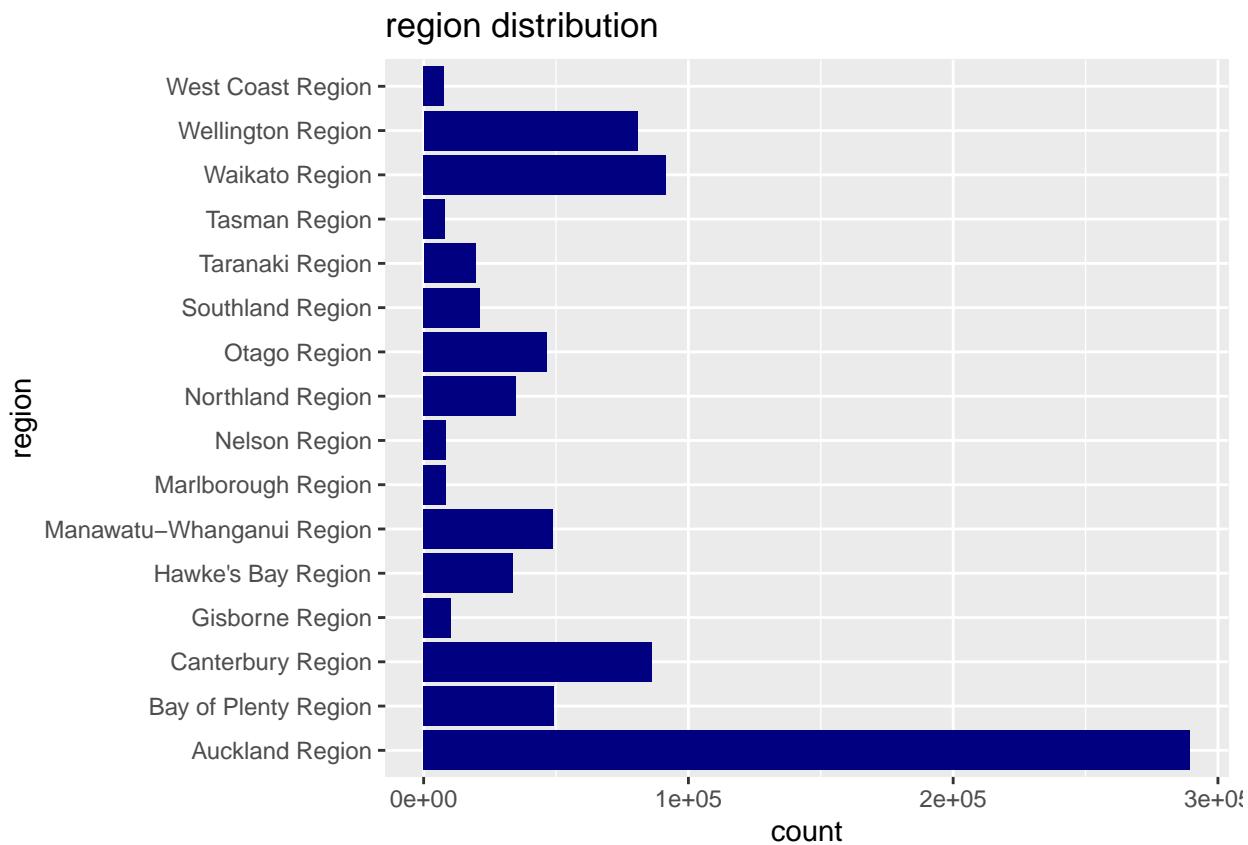
Bar plot of light:

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



Bar plot of region:

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

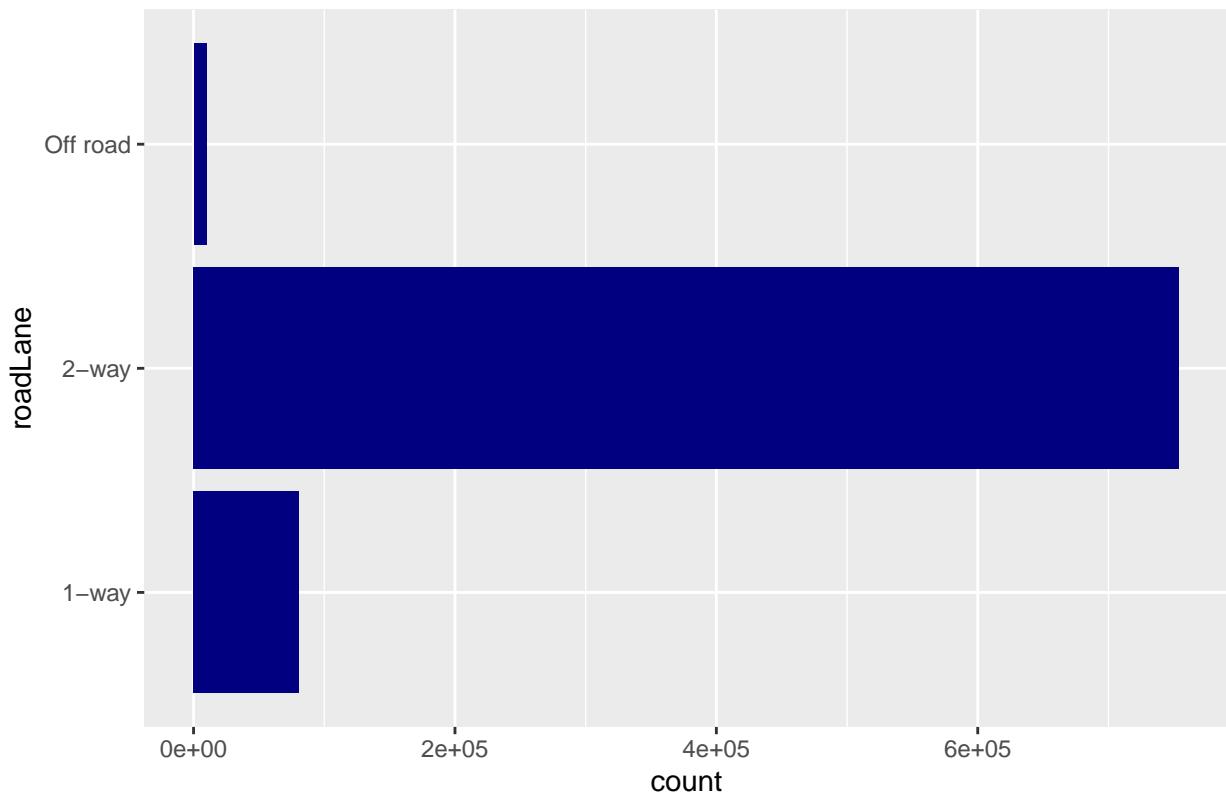


The majority of car crashes occur in Auckland.

Bar plot of roadLane:

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

## roadLane distribution

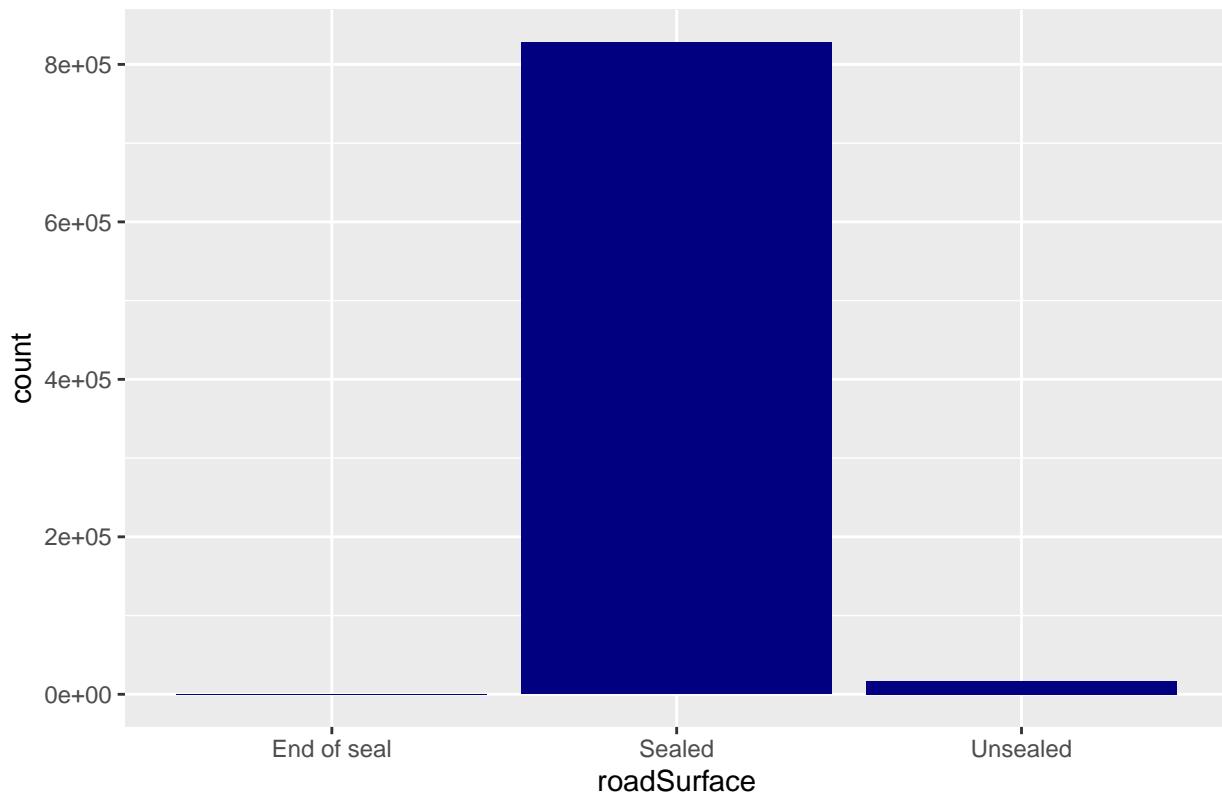


Few Off road observations

Bar plot of roadSurface:

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

### roadSurface distribution



There are very few End of seal observations

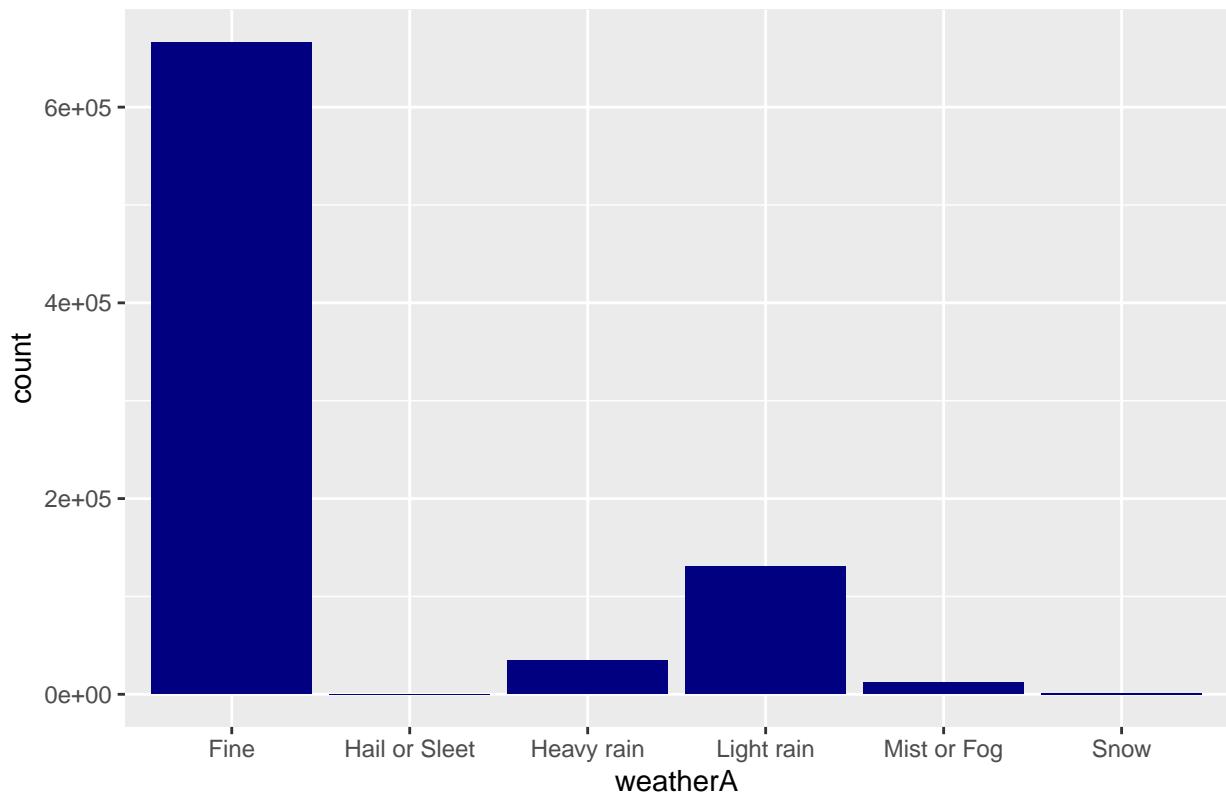
```
sum(Crash$roadSurface == "End of seal")
```

```
## [1] 122
```

Bar plot of weatherA:

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

## weatherA distribution



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

```
## [1] 172
```

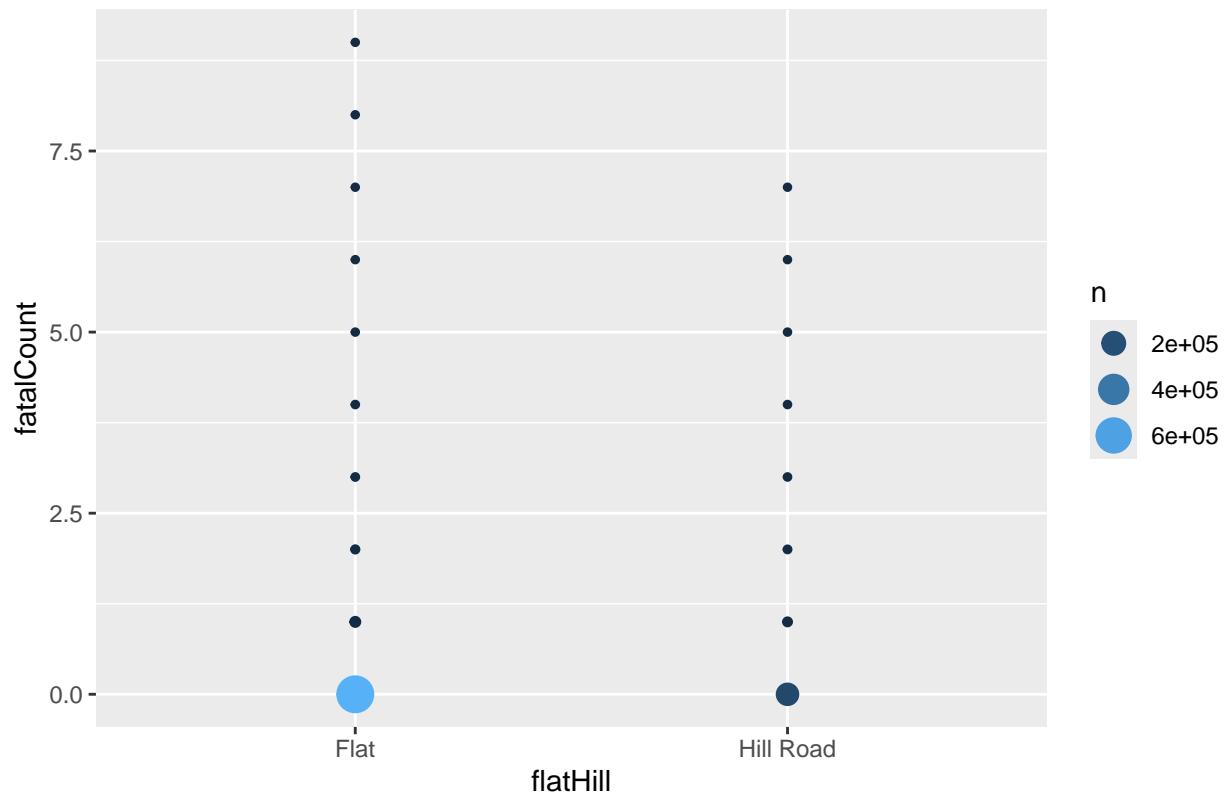
There are only 172 Hail or Sleet observations. Many of the categorical variables are imbalanced.

## Numerical data plots

Count plot of response variable fatalCount:

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

## Count plot of fatalCount by flatHill



There are few observations with at least 1 death.

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

```
## [1] 836997
```

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

```
## [1] 7968
```

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

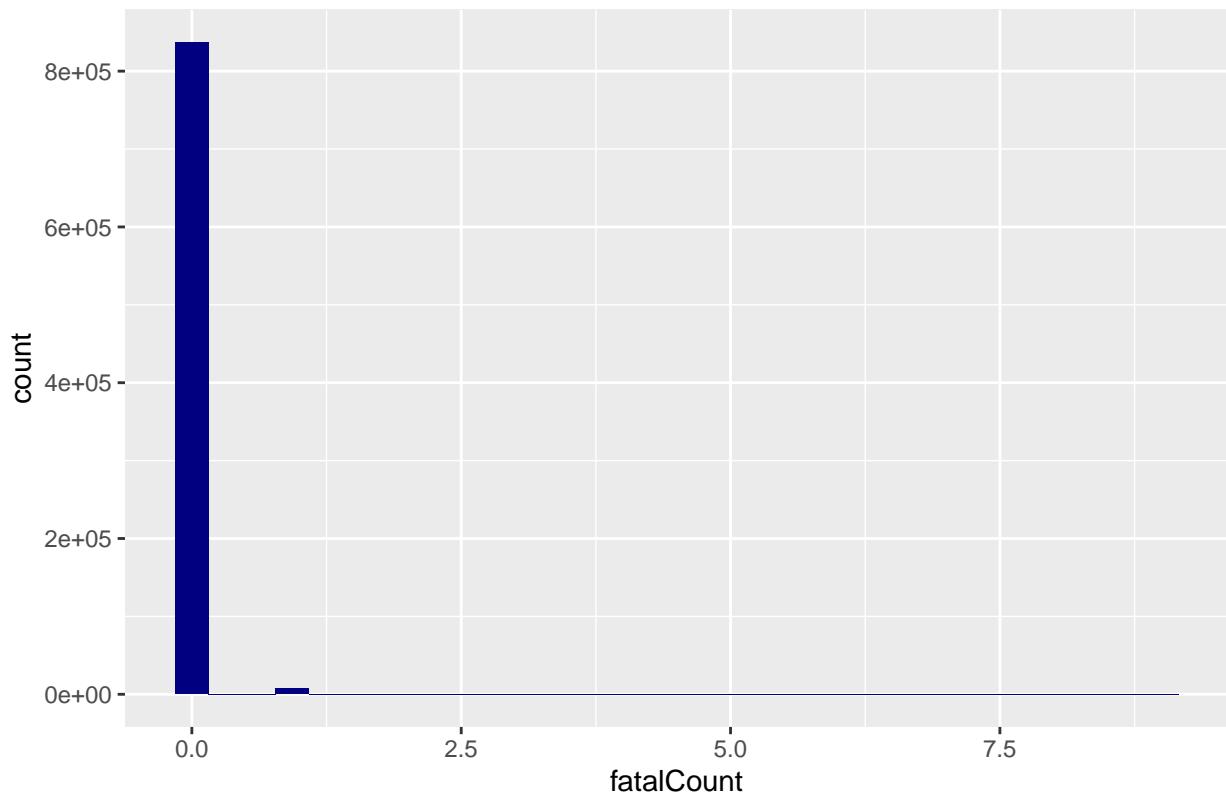
```
## [1] 0.95
```

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

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

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

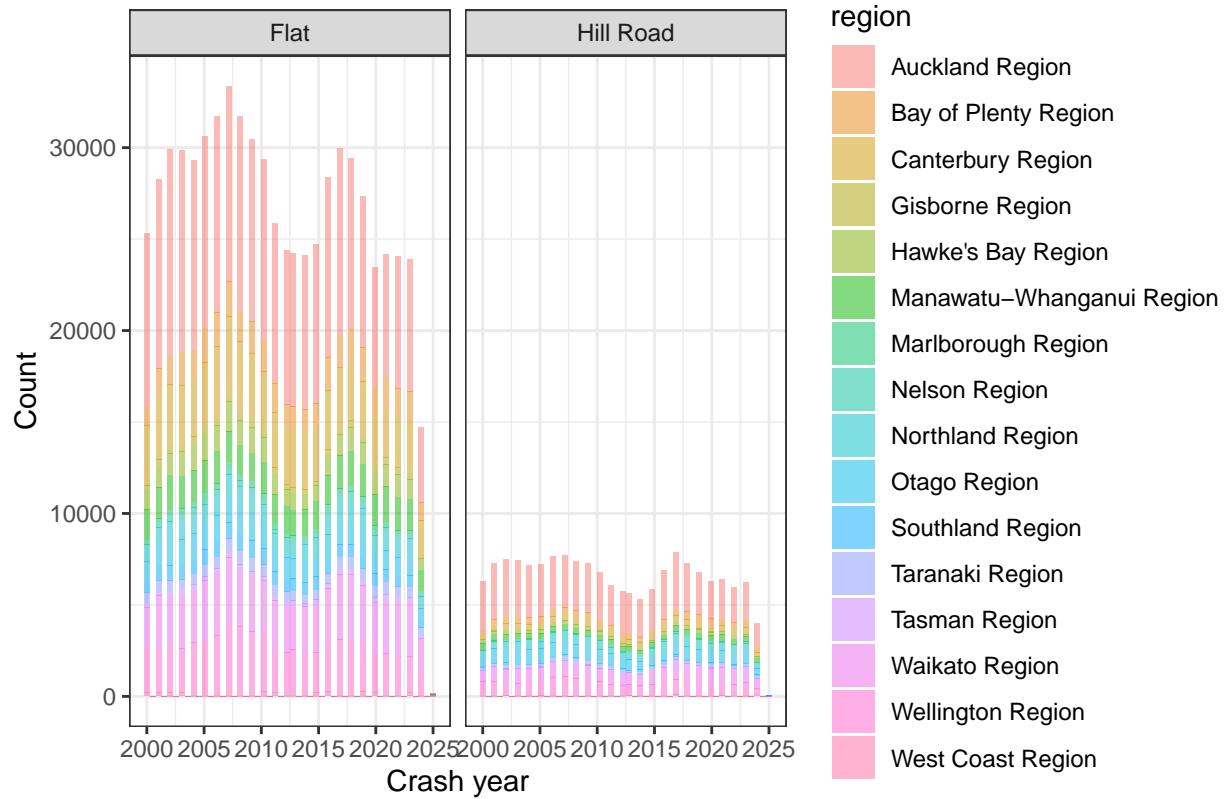
### fatalCount distribution



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

```
ggplot(data = Crash,  
       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()
```

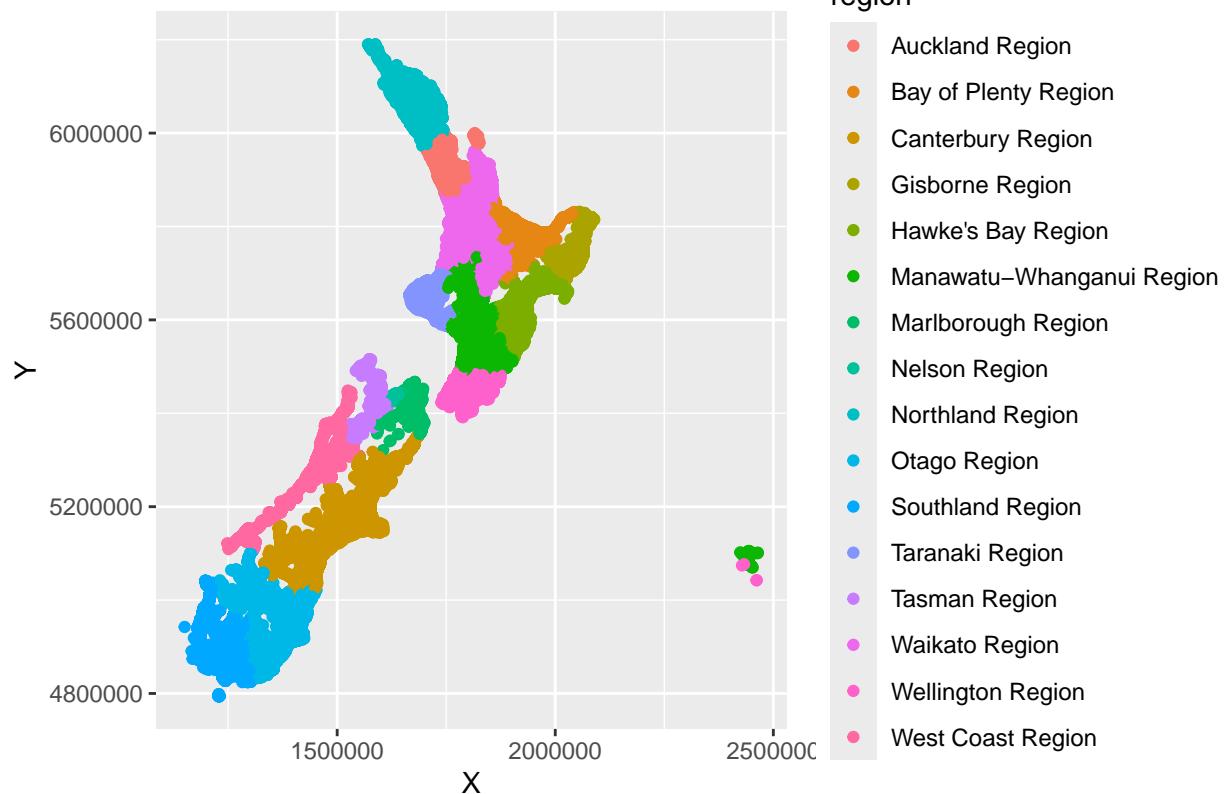
## crashYear by flatHill and Region



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(Crash, aes(x = X, y = Y, color = region)) +
  geom_point() + ggttitle("X by Y and region")
```

## X by Y and region

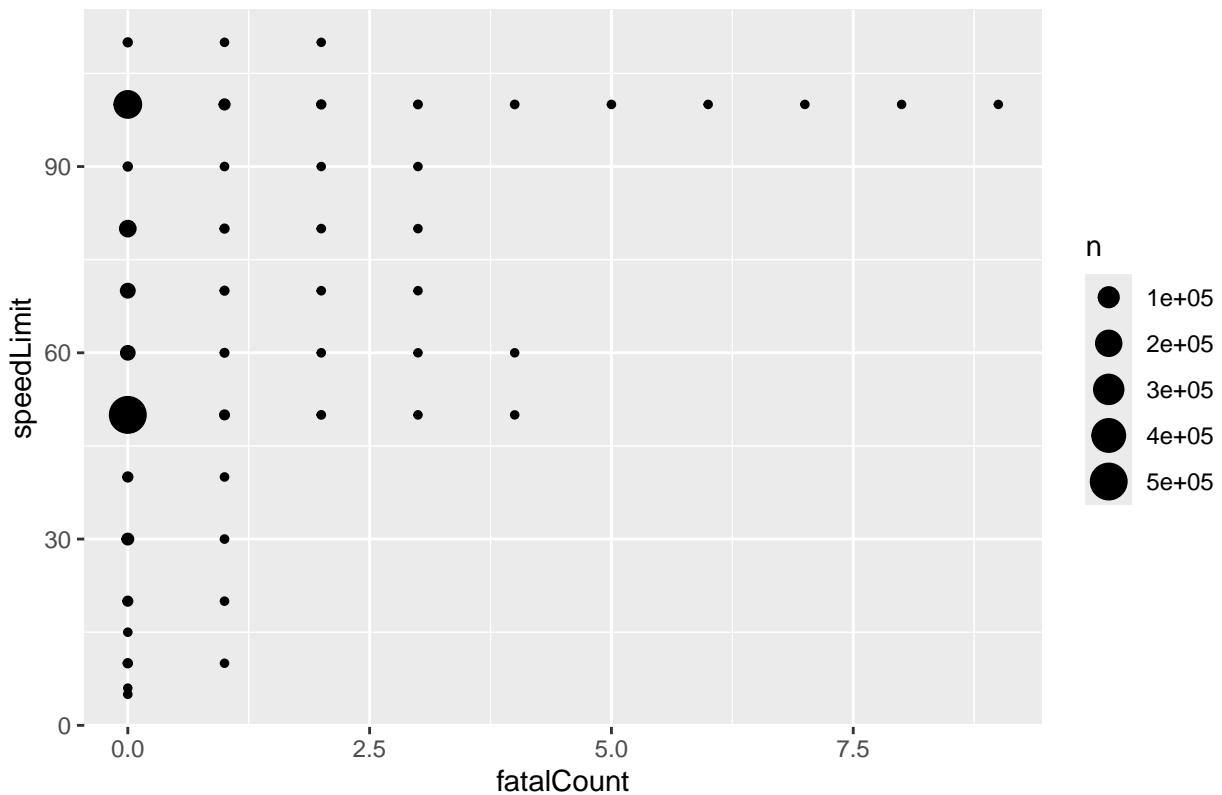


It appears that X and Y contain redundant information that can be modeled by region.

Count plot of response variable fatalCount:

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

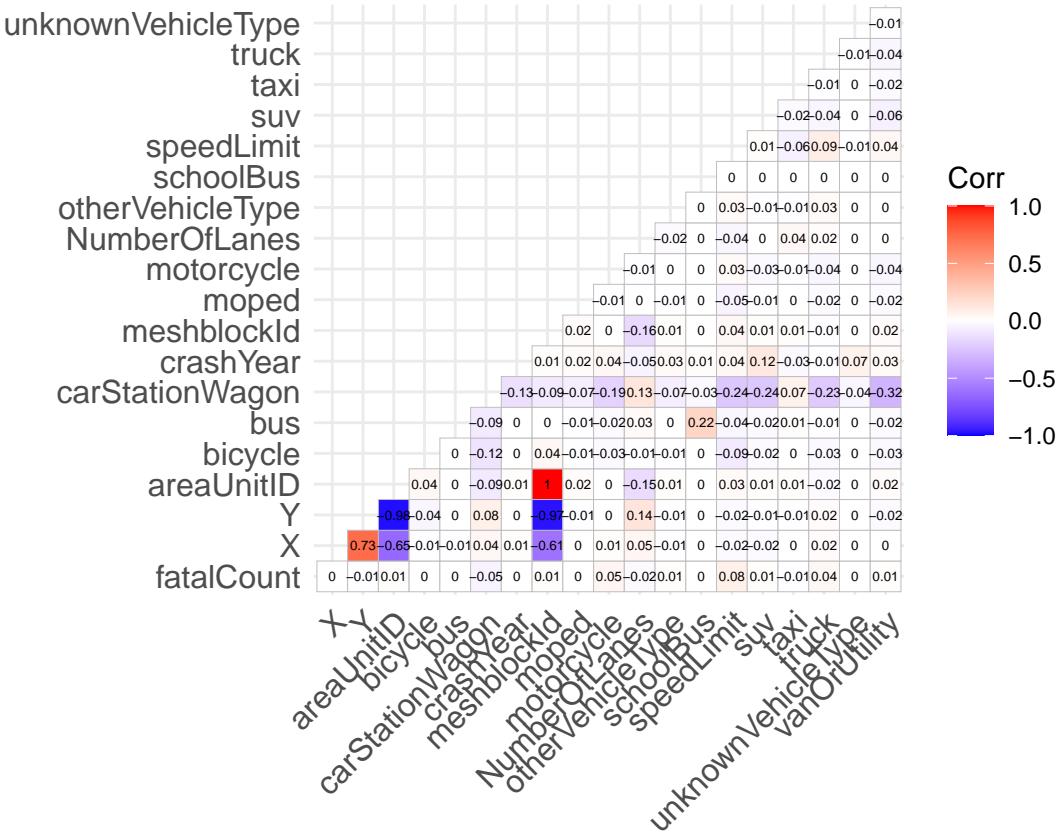
Count plot of fatalCount by speedLimit



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

### Correlation plot

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



```
findCorrelation(cor(select_if(Crash, 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

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

I now have 23 variables in the Crash data set.

## Skewness

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

Table 28: Table continues below

fatalCount	bicycle	bus	carStationWagon	crashYear	moped
15.07	6.129	8.042	0.4287	0.0966	11.93

Table 29: Table continues below

motorcycle	NumberOfLanes	otherVehicleType	schoolBus	speedLimit	suv
5.651	1.802	14.3	36.35	0.6734	3.158
<hr/>					
taxi	truck	unknownVehicleType	vanOrUtility		
10.5	3.555	19.54	2.227		

The response variable fatalCount has a skewness of 15.07 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(Crash, is.numeric)))
```

Table 31: Table continues below

fatalCount	bicycle	bus	carStationWagon	crashYear	moped
363.7	45.68	69.04	4.26	1.804	146.3

Table 32: Table continues below

motorcycle	NumberOfLanes	otherVehicleType	schoolBus	speedLimit	suv
44.27	7.582	219.8	1421	1.726	13.28
<hr/>					
taxi	truck	unknownVehicleType	vanOrUtility		
128.8	15.91	410.4	7.649		

The response variable fatalCount has a kurtosis of 363.7, 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 = Crash[,-1], y = Crash$fatalCount, nonpara = TRUE)
roc_imp <- data.frame(cbind(variable = rownames(roc_imp), score = roc_imp[,1]))
roc_imp$score <- as.double(roc_imp$score)
roc_imp <- roc_imp[order(roc_imp$score,decreasing = TRUE),]
pander(roc_imp)
```

	variable	score
16	speedLimit	0.006846
3	carStationWagon	0.002264
9	motorcycle	0.002249
4	crashSHDescription	0.002016
19	truck	0.001523
12	region	0.0005017
10	NumberOfLanes	0.000402
13	roadLane	0.0003068
6	flatHill	0.0001869
11	otherVehicleType	0.0001188
21	vanOrUtility	6.696e-05
17	suv	5.021e-05
18	taxi	4.436e-05
14	roadSurface	4.328e-05
8	moped	2.239e-05
15	schoolBus	1.841e-05
20	unknownVehicleType	1.511e-05
2	bus	9.198e-06
5	crashYear	6.401e-06
7	light	2.277e-06
22	weatherA	6.409e-07
1	bicycle	5.928e-07

Reducing the data set to the ten most important variables:

```
Crash2 <- Crash %>% 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 = Crash2, family = poisson)
pander(summary(model_glm_P))

```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-8.228	0.09641	-85.34	0
speedLimit	0.03274	0.0006049	54.13	0
motorcycle	0.8625	0.02125	40.59	0
roadLane2-way	1.406	0.06512	21.59	2.436e-103
roadLaneOff road	1.594	0.1403	11.36	6.532e-30
truck	0.6601	0.02501	26.39	1.586e-153
regionBay of Plenty Region	0.439	0.04721	9.3	1.403e-20
regionCanterbury Region	0.3516	0.0431	8.157	3.435e-16
regionGisborne Region	0.1095	0.09288	1.178	0.2387
regionHawke's Bay Region	0.1822	0.05687	3.203	0.001359
regionManawatū-Whanganui Region	0.3092	0.04726	6.542	6.062e-11
regionMarlborough Region	0.2506	0.09092	2.756	0.005849
regionNelson Region	-0.227	0.1538	-1.476	0.1399
regionNorthland Region	0.3633	0.04938	7.358	1.874e-13
regionOtago Region	-0.1165	0.05793	-2.011	0.04431
regionSouthland Region	0.04354	0.06932	0.6282	0.5299
regionTaranaki Region	0.2113	0.06679	3.163	0.00156
regionTasman Region	-0.08718	0.09632	-0.9051	0.3654
regionWaikato Region	0.3614	0.03952	9.145	5.962e-20
regionWellington Region	-0.1379	0.05483	-2.515	0.0119
regionWest Coast Region	0.2326	0.08259	2.817	0.004851
NumberOfLanes	-0.168	0.01778	-9.45	3.401e-21
otherVehicleType	0.3089	0.08474	3.645	0.0002674
crashSHDescriptionYes	0.1974	0.02434	8.112	4.968e-16
carStationWagon	-0.1577	0.01641	-9.608	7.412e-22
flatHillHill Road	-0.02073	0.0247	-0.8395	0.4012

(Dispersion parameter for poisson family taken to be 1 )

Null deviance:	85086 on 844964 degrees of freedom
Residual deviance:	74307 on 844939 degrees of freedom

## Checking model assumptions:

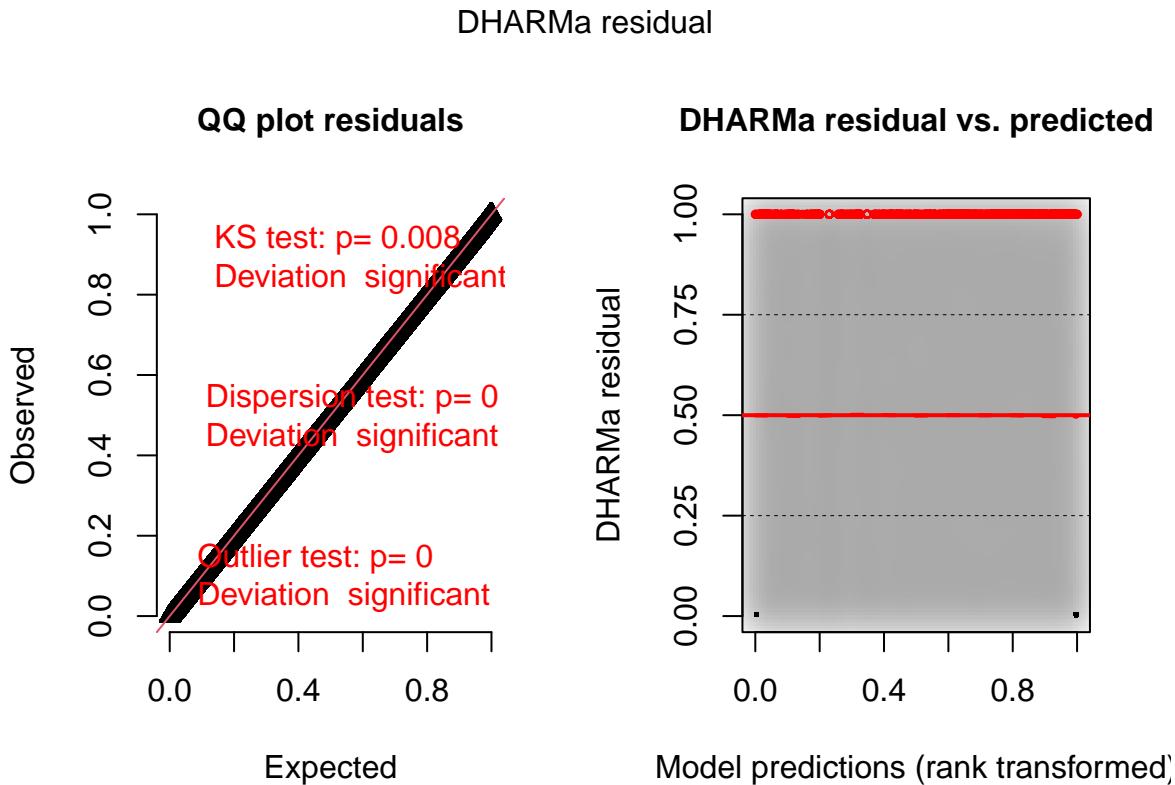
I need to check if the response variable fatalCount follows a poisson distribution:

```

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 = 13.104, p-value < 2.2e-16  
## alternative hypothesis: true dispersion is greater than 1  
## sample estimates:  
## dispersion  
## 1.199071
```

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 quasipoisson 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 = Crash2)
```

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

```
##           df      BIC
## model_glm_P 26 91166.21
## model_nb    27 87974.53
```

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: 836997
##   Predicted zeros: 837009
##             Ratio: 1.00

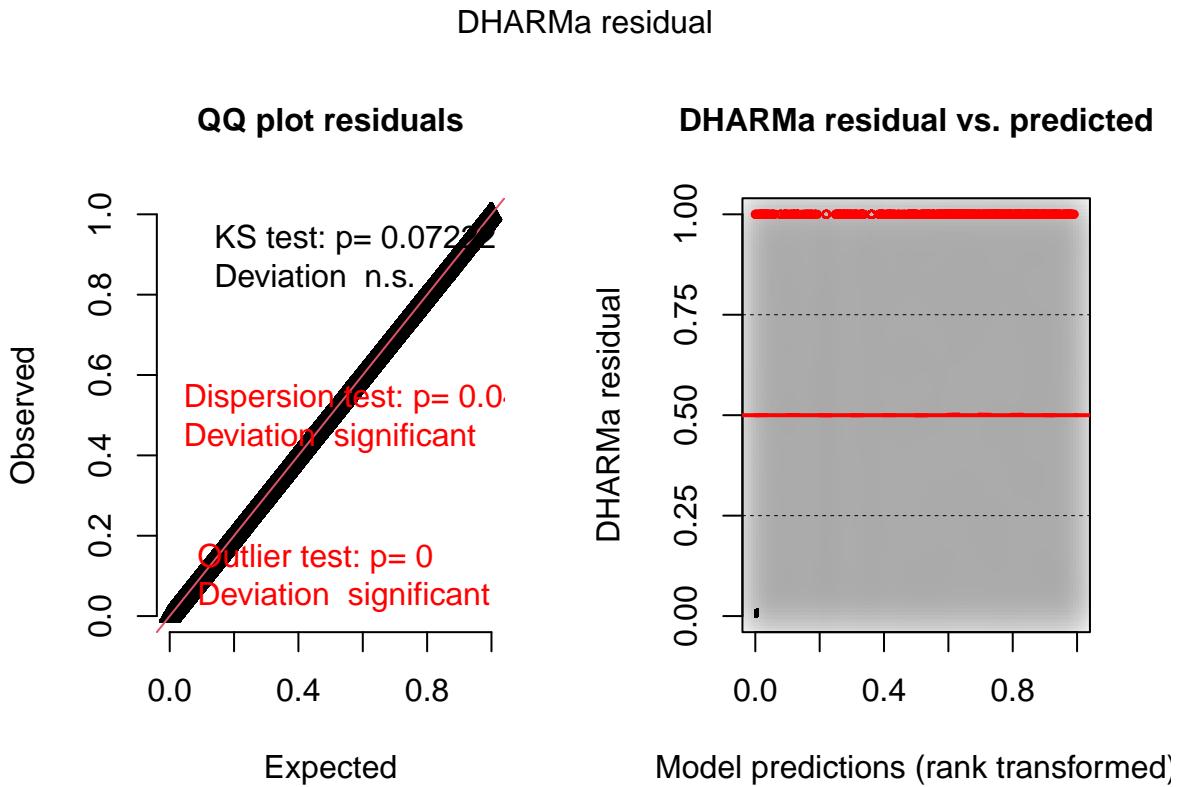
## Model seems ok, ratio of observed and predicted zeros is within the
## tolerance range (p = 0.840).
```

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 is 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.

The dispersion test is statistically significant at a significance level of 0.05, I may need to consider additional methods to deal with overdispersion.

## Diagnostic measures

Is there severe multicollinearity in the data set:

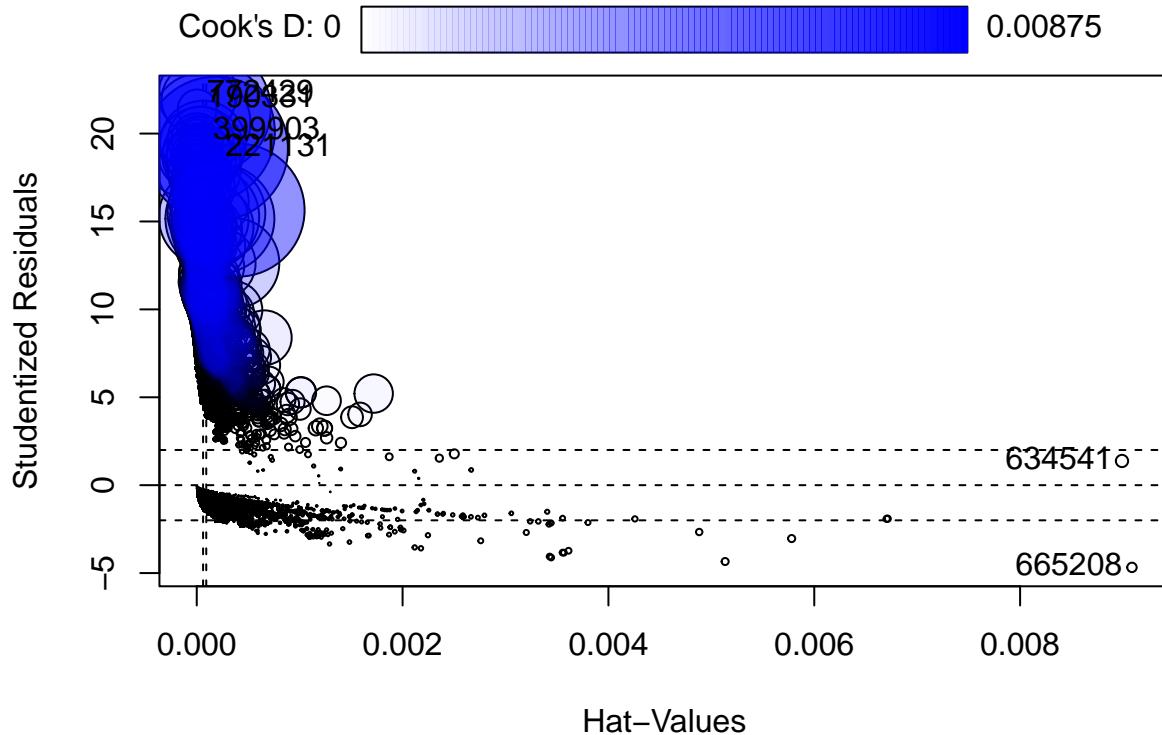
```
pander(vif(model_nb))
```

	GVIF	Df	GVIF^(1/(2*Df))
speedLimit	1.466	1	1.211
motorcycle	1.112	1	1.055
roadLane	1.237	2	1.055
truck	1.141	1	1.068
region	1.401	15	1.011
NumberOfLanes	1.195	1	1.093
otherVehicleType	1.009	1	1.004
crashSHDescription	1.355	1	1.164
carStationWagon	1.326	1	1.152
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  21.934360 1.531365e-05 2.278637e-03
## 221131  19.138611 1.864857e-04 8.753667e-03
## 399903  20.106231 6.764269e-05 8.546707e-03
## 634541   1.376228 8.990071e-03 6.349918e-05
## 665208  -4.676156 9.086875e-03 3.759999e-05
## 772429  22.223502 8.801727e-06 2.604160e-03
```

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

```
Crash2 <- Crash2[-c(190331, 221131, 399903, 634541, 665208, 772429),]
```

Finding outliers using the standardized residuals:

```
model.diag.metrics <- augment(model_nb)
```

```

## Warning: The 'augment()' method for objects of class 'negbin' is not maintained by the broom team,
## This warning is displayed once per session.

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] 208

```

There are 208 observations that have standardized residuals larger than 3 in absolute value.

I am going to remove these from the data set:

```
Crash2 <- Crash2[-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 = Crash2)

```

Stepwise BIC:

```

step(model_nb, direction = "both", k = log(nrow(Crash2)))

## Start: AIC=86673.35
## fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +
##   NumberOfLanes + otherVehicleType + crashSHDescription + carStationWagon +
##   flatHill
##
##                               Df Deviance    AIC
## - flatHill             1   45841 86660
## - otherVehicleType     1   45850 86669
## <none>                  45841 86673
## - crashSHDescription  1   45888 86707
## - region                15  46092 86720
## - NumberOfLanes          1   45923 86742
## - carStationWagon        1   45983 86802
## - truck                  1   46303 87122
## - roadLane                2   46415 87220
## - motorcycle              1   46566 87385
## - speedLimit              1   48611 89430
##
## Step: AIC=86660.44

```

```

## fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +
##      NumberOfLanes + otherVehicleType + crashSHDescription + carStationWagon
##
##                                     Df Deviance   AIC
## - otherVehicleType     1    45842 86656
## <none>                      45833 86660
## + flatHill            1    45832 86673
## - crashSHDescription  1    45880 86694
## - region              15   46088 86710
## - NumberOfLanes       1    45915 86729
## - carStationWagon     1    45975 86789
## - truck               1    46296 87110
## - roadLane            2    46407 87207
## - motorcycle          1    46558 87372
## - speedLimit          1    48657 89471
##
## Step: AIC=86655.84
## fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +
##      NumberOfLanes + crashSHDescription + carStationWagon
##
##                                     Df Deviance   AIC
## <none>                      45819 86656
## + otherVehicleType     1    45810 86660
## + flatHill            1    45818 86669
## - crashSHDescription  1    45866 86689
## - region              15   46075 86707
## - NumberOfLanes       1    45901 86724
## - carStationWagon     1    45965 86789
## - truck               1    46285 87108
## - roadLane            2    46393 87202
## - motorcycle          1    46541 87364
## - speedLimit          1    48651 89475
##
## Call: glm.nb(formula = fatalCount ~ speedLimit + motorcycle + roadLane +
##      truck + region + NumberOfLanes + crashSHDescription + carStationWagon,
##      data = Crash2, init.theta = 0.121324822, link = log)
##
## Coefficients:
##                                     (Intercept)           speedLimit
##                                         -8.20080             0.03260
##                                     motorcycle         roadLane2-way
##                                         1.04516             1.40543
##                                     roadLaneOff road          truck
##                                         1.59591             0.67747
##                                     regionBay of Plenty Region regionCanterbury Region
##                                         0.45064             0.36263
##                                     regionGisborne Region regionHawke's Bay Region
##                                         0.11214             0.18990
##                                     regionManawatū-Whanganui Region regionMarlborough Region
##                                         0.33031             0.26455
##                                     regionNelson Region regionNorthland Region
##                                         -0.20138            0.37398
##                                     regionOtago Region regionSouthland Region

```

```

##          -0.10684          0.04611
## regionTaranaki Region      regionTasman Region
##          0.21910          -0.09001
## regionWaikato Region      regionWellington Region
##          0.36364          -0.14249
## regionWest Coast Region    NumberOfLanes
##          0.21637          -0.16827
## crashSHDescriptionYes     carStationWagon
##          0.18715          -0.21691
##
## Degrees of Freedom: 844750 Total (i.e. Null); 844727 Residual
## Null Deviance: 55460
## Residual Deviance: 45820 AIC: 86380

```

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.

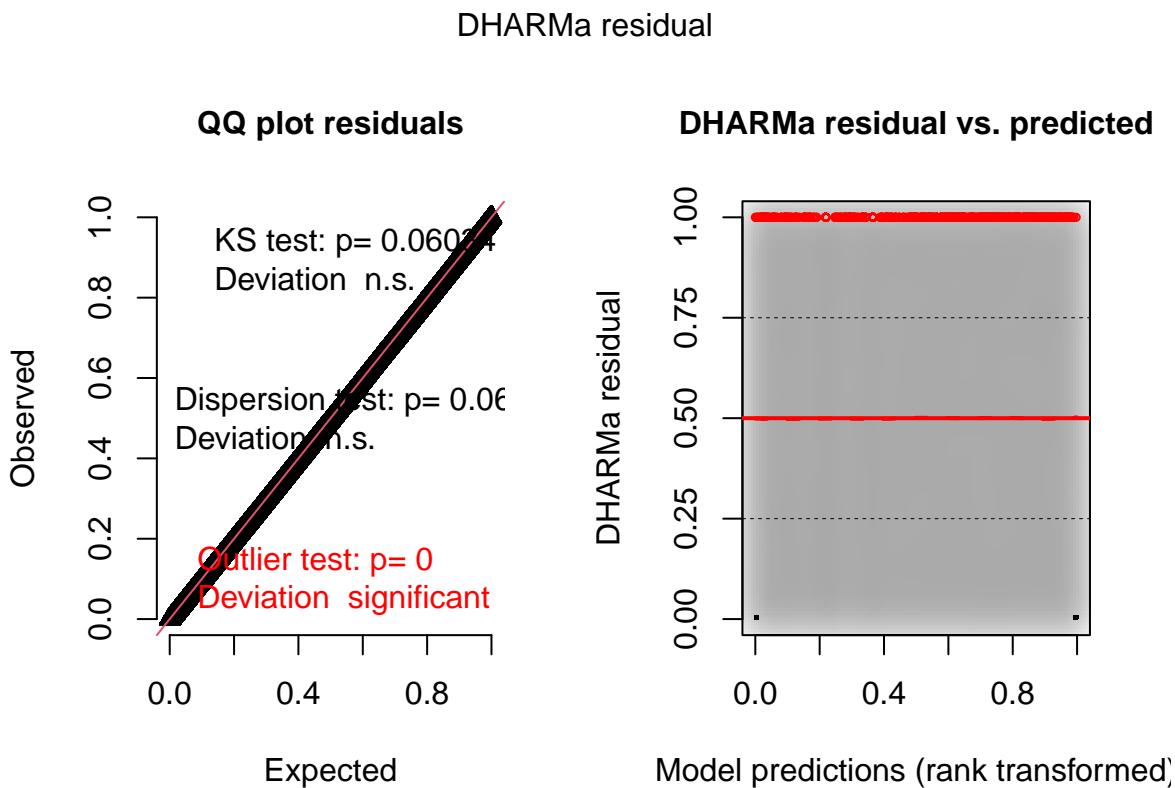
## Residual plots

```

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

```

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



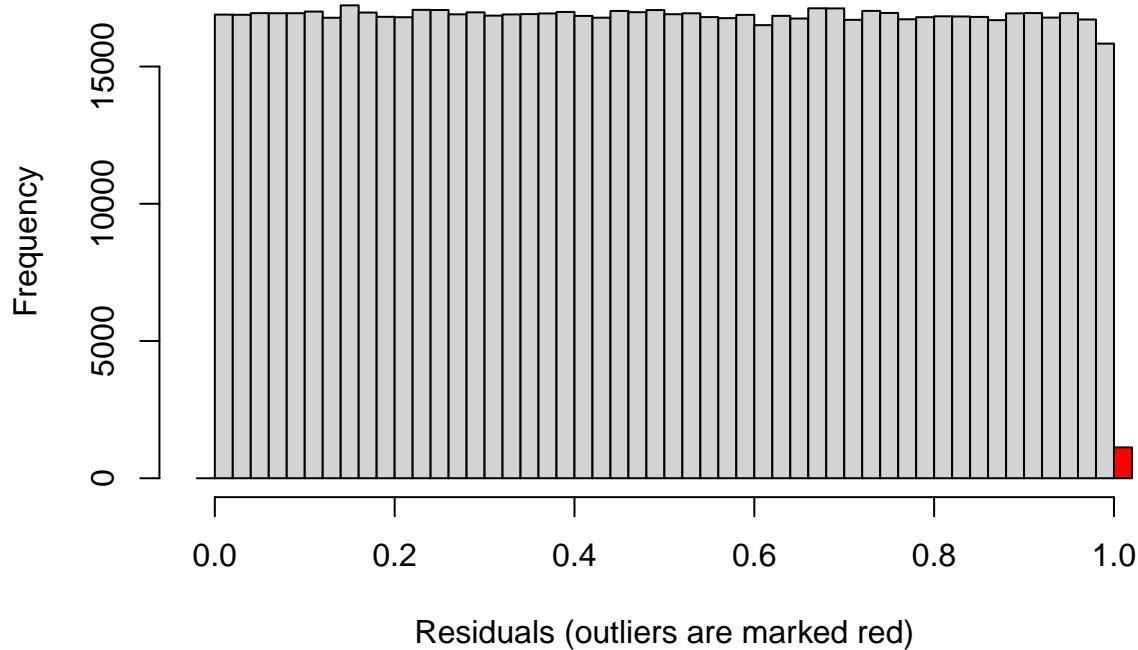
The KS test and dispersion test are not statistically significant at a significance level of 0.05, so I conclude that the data does fit a negative binomial model.

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)
```

## Outlier test significant



```
outliers
```

```
##  
##  DHARMA bootstrapped outlier test  
##  
##  data: simulationOutput2  
##  outliers at both margin(s) = 1124, observations = 844751, p-value <  
##  2.2e-16  
##  alternative hypothesis: two.sided  
##  percent confidence interval:  
##  0.001915712 0.002085259  
##  sample estimates:  
##  outlier frequency (expected: 0.00201148030603101 )  
##                                         0.00133057
```

There are 1153 more outliers than expected in the data set.

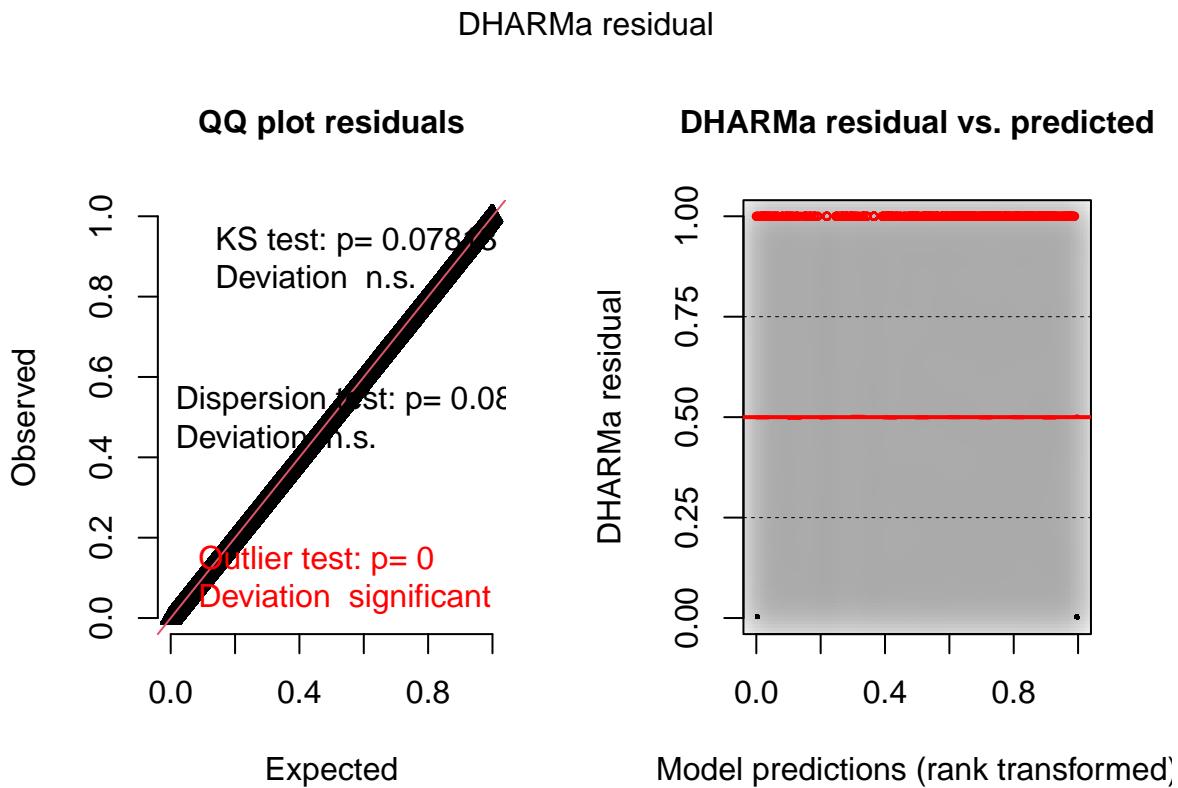
Checking the residual plots to see if removing these outliers fixes the outlier problem:

```
Potential_outliers <- sort(simulationOutput2$scaledResiduals, decreasing = TRUE)[1:1153]  
Crash3 <- Crash2[-c(Potential_outliers), ]
```

```
model_nb2 <- glm.nb(fatalCount ~ speedLimit + motorcycle + roadLane + truck + region +  
NumberofLanes + otherVehicleType + crashSHDescription +  
carStationWagon + flatHill, data = Crash3)
```

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

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



```
##
##  DHARMA bootstrapped outlier test
##
##  data:  simulationOutput3
##  outliers at both margin(s) = 1166, observations = 844750, p-value <
##  2.2e-16
##  alternative hypothesis: two.sided
##  percent confidence interval:
##  0.001928085 0.002183753
##  sample estimates:
##  outlier frequency (expected: 0.00202770050310743 )
##                                0.00138029
```

Removing the outliers identified by the first outlier test did not fix the outlier issue, given this result I have decided to keep the outliers in the data set and to continue with my original negative binomial model.

## Interpreting model coefficients

```
pander(summary(model_nb))
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-8.203	0.1025	-80.03	0
speedLimit	0.03264	0.0006492	50.28	0
motorcycle	1.048	0.03598	29.14	1.242e-186
roadLane2-way	1.408	0.06854	20.55	7.968e-94
roadLaneOff road	1.593	0.1537	10.36	3.725e-25
truck	0.6753	0.03008	22.45	1.215e-111
regionBay of Plenty Region	0.4476	0.05202	8.604	7.681e-18
regionCanterbury Region	0.3569	0.04691	7.609	2.77e-14
regionGisborne Region	0.1109	0.1027	1.08	0.2803
regionHawke's Bay Region	0.1867	0.0626	2.983	0.002856
regionManawatū-Whanganui Region	0.3259	0.05196	6.272	3.561e-10
regionMarlborough Region	0.257	0.1023	2.512	0.01199
regionNelson Region	-0.2016	0.1624	-1.242	0.2143
regionNorthland Region	0.3716	0.05469	6.795	1.084e-11
regionOtago Region	-0.1064	0.06268	-1.697	0.08974
regionSouthland Region	0.03776	0.07605	0.4966	0.6195
regionTaranaki Region	0.2153	0.07406	2.908	0.00364
regionTasman Region	-0.09333	0.1058	-0.8821	0.3777
regionWaikato Region	0.3603	0.04326	8.327	8.287e-17
regionWellington Region	-0.1435	0.05858	-2.45	0.01427
regionWest Coast Region	0.2149	0.09393	2.288	0.02215
NumberOfLanes	-0.1688	0.01911	-8.83	1.043e-18
otherVehicleType	0.3191	0.1022	3.122	0.001797
crashSHDescriptionYes	0.1878	0.02725	6.893	5.461e-12
carStationWagon	-0.2145	0.01847	-11.61	3.548e-31
flatHillHill Road	-0.02383	0.02773	-0.8594	0.3901

(Dispersion parameter for Negative Binomial(0.1217) family taken to be 1 )

Null deviance:	55490 on 844750 degrees of freedom
Residual deviance:	45841 on 844725 degrees of freedom

Interpretation of coefficients:

For a one unit change in the speedLimit, the log of expected counts of fatalCount changes by 0.03264, 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 1.048, 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 1.408 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 1.593 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 0.6753, 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 0.4476 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 0.3569 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 0.1109 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 Hawke's Bay region is 0.1867 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 0.3259 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 0.257 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 Nelson region is -0.2016 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 0.3716 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 -0.1064 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 Southland region is 0.03776 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 0.2153 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 -0.09333 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 Waikato region is 0.3603 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 -0.1435 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 0.2149 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 -0.1688, 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 0.3191, 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 0.1878, 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 -0.2145, 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 -0.02383, 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 regions Gisborne, Marlborough, Nelson, Otago, 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

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.

For future projects I will split my data into a training set and a testing set so that I can use model evaluation metrics such as AUC-ROC, accuracy, precision and recall.

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.