Data Analysis and Predictive Modeling – NZ Injury Statistics

Jordan, Rawinder

IDS201

6/25/21

Injuries bring an array of both social and economic implications of injuries. For this very reason, the Ministry of Health (MOH) regards injuries as a prominent event. Current statistics by Stats NZ have shown an increase in non-fatal injuries in the period 2012-2019 (Stats NZ, 2020). As a government agency, the MOH needs to closely manage this data, to effectively manage its resources. They need to be able to analyze current trends in the data and be able to draw insight and inferences. Alongside this, they need to be able to predict using various variables to better understand what needs to be done.

The data set we have chosen to analyze shows the number of serious both fatal and non-fatal injuries in New Zealand annually and has indicators to show trends in the rate of the outcomes from these injuries. An injury is considered serious non-fatal when the person injured has a 6.9% chance to be fatal, an injury is considered serious if the person has been badly injured such as fracture, severe cut, or if the person has to remain in the hospital for an extended amount of time if the person ends up dying it is considered fatal.

Since the goal was to predict the outcome of injuries based on a range of indicators, the decision was made to use a predictive analysis model, which predicts outcomes using the KNN regression algorithm. This was because it can look at the features of each injury case and figure out which ones are closest in similarity and make predictions on new data by comparing the new data to the model (Lateef, 2020). This means it is good for providing accurate predictions. This could be considered quite important when it comes to injury statistics as these predictions could be used to see where improvements are needed with injury responses.

The analytical approach to solve the issue of descriptive analysis, inference, and predictive modeling is based on a framework produced by Harvard Data Science Department (Keller et al., 2020). The first step in this framework was to effectively understand the data science issue. This helped with understanding what exactly needed to be achieved. The following step was the data discovery stage. The data used in this case study was retrieved from the Stats NZ website. The next step was Data Wrangling. In this step, the data set was transformed by various preprocessing procedures, to ensure its suitability for modeling. The next stage was Statistical modeling and analyses. This stage used both inferential and descriptive statistical methods. In this stage, the data was visualized into readable graphs, from which you can draw an inference. From this, the data was fed into the KNN model to produce predictions. The final stage was to summarize the presentation and the results. This approach proved to be successful in effective application on both descriptive and inferential statistical methods.

To analyze the data a check had to be done to see if it required cleaning and if so doing that, preprocess it so it is in a trainable format, and then visualize the current data using graphs. The original data set had already been cleaned and there weren't any columns specifically unnecessary so the next step was to preprocess the data, this was done by turning the data that could be categorized into numbers for each column/indicator, for example through the severity column the values got changed from serious non-fatal to 1, serious to 2, and fatal to 3.

```
Severity
Serious Non-fatal = 1 | Serious = 2 | Fatal = 3

Age
0-14 Years = 1 | 0-74 Years = 2 | 75+ Years = 3 | All ages = 4

Population
Maori = 1 | Children = 2 | Whole Population = 3

Validation
Provisional = 1 | Validated = 1

Cause
Assault = 1 | Falls = 2 | Work = 3 | Intentional Self-harm = 4 | Intentional = 5 | Motor Vehicle Traffic Crashes = 6 | Car Occupant= 7 | Pedestrian = 8 | Drowning = 9 | All = 10

Indicator
Number = 1 | Age-standardised rate = 2 | Rate per billion km = 3 | Rate per thousand registered vehicles = 4

Units
Injuries = 1 | Per 100,000 people = 2 | Per 100,000 FTE's = 3 | Per billion km = 4 | Per thousand vehicles registered = 5

Type
Moving average = 1 | Single year = 2
```

This was done using for and if loops in RStudio to check the values in each column then change them accordingly.

```
#Go through severity column and change values to number Serious non-fatal = 1, Serious = 2, Fatal = 3
if(NewInjuryIndicators[i, 13]=="Serious non-fatal"){
  NewInjuryIndicators[i, 13] <- 1
}
else if(NewInjuryIndicators[i, 13]=="Serious"){
  NewInjuryIndicators[i, 13] <- 2
}
else {
  NewInjuryIndicators[i, 13] <- 3
}
```

To visualize the data from here bar charts seemed like the best way to display the data, so bar charts were made comparing 2 indicators.

```
#Get a summary of overall dataframe
summary(ModifiedInjuryIndicators)

#Create tables showing totals for Population, Ages, Cause, Severity
table(ModifiedInjuryIndicators$Population)
table(ModifiedInjuryIndicators$Age)
table(ModifiedInjuryIndicators$Cause)
table(ModifiedInjuryIndicators$Severity)

#Create tables comparing between Cause and Severity, Age and Severity, and Population and Severity
table(ModifiedInjuryIndicators$Cause, ModifiedInjuryIndicators$Severity)
table(ModifiedInjuryIndicators$Age, ModifiedInjuryIndicators$Severity)
table(ModifiedInjuryIndicators$Population, ModifiedInjuryIndicators$Severity)


#Splits NewInjStats based on severity column
SplitInjIndicators <- split(ModifiedInjuryIndicators, ModifiedInjuryIndicators$Severity)
lapply(names(SplitInjIndicators), function(x){
  write_csv(SplitInjIndicators[[x]], path = paste(x, ".csv", sep = ""))
})

#Then taking the subsets made from splitting between severity before,
# creates tables comparing population and cause then age and cause for each subset
table(SplitInjIndicators[[1]]$Population, SplitInjIndicators[[1]]$Cause)#Serious non-fatal tables
table(SplitInjIndicators[[1]]$Age, SplitInjIndicators[[1]]$Cause)#Serious non-fatal tables

table(SplitInjIndicators[[2]]$Population, SplitInjIndicators[[2]]$Cause)#Serious Tables
table(SplitInjIndicators[[2]]$Age, SplitInjIndicators[[2]]$Cause)#Serious tables

table(SplitInjIndicators[[3]]$Population, SplitInjIndicators[[3]]$Cause)#Fatal tables
```

This was also all done in RStudio, where the graphs were just chosen based on indicators we decided were most important/we could get the most value out of.

The objective for this data analysis was to visualize the data from the Stats NZ website for the MOH. The next step is to draw insight from the graphs into statistical findings. In the following section, there will be snapshots of graphs produced from the analysis.

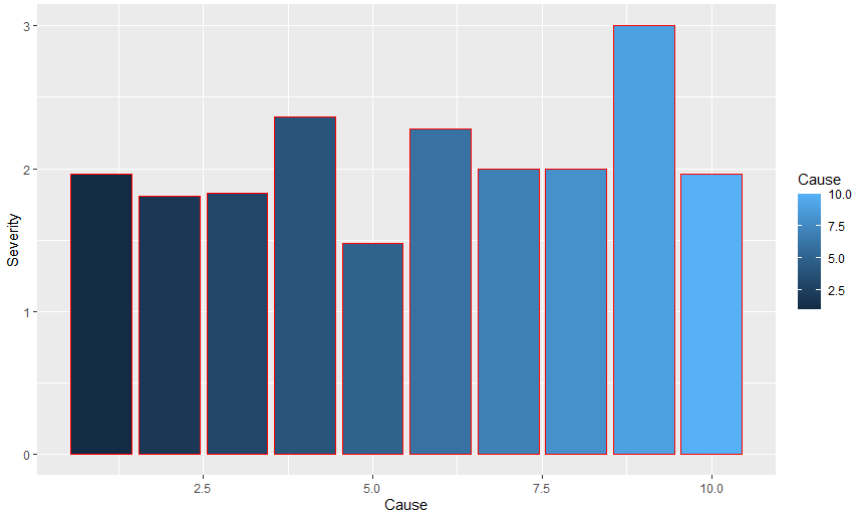

Figure 1 here shows the relationship between the different types of injury causes and the resulting severity level. Some key findings are: The most severe cause of an injury was drowning, alongside this was self-harm and motor vehicle traffic crashes. On the other end of the spectrum, falls and intentional accidents were less serves. From this the MOH and focus more on isolated injury management.

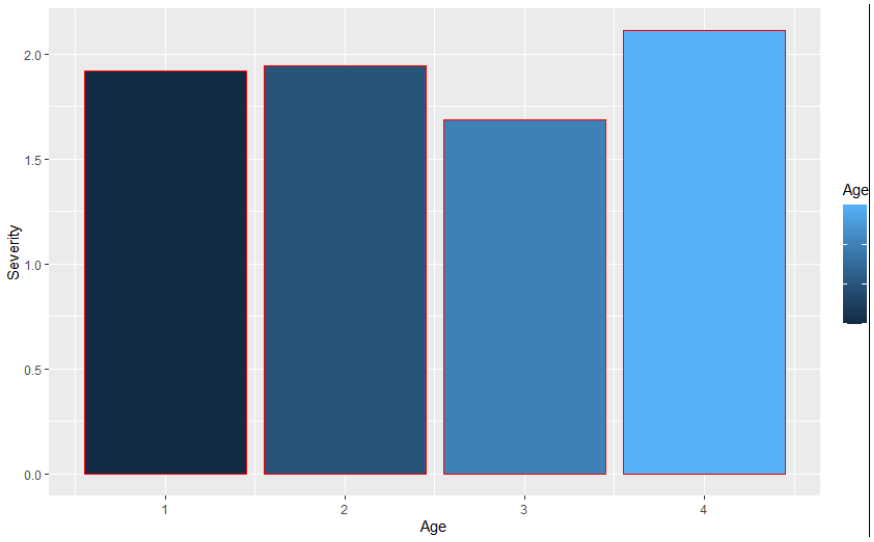

Figure 2 here shows the relationship between the various age levels in this data and the severity of their injury causes. The graph shows that children aged 0-14 years were more likely to sustain severe injuries than 75+ years. From this, the MOH can focus on young age management of injuries.

```
> summary(ModifiedInjuryIndicators)
 Series_reference      Period              Type          Data_value          Lower_CI            Upper_CI
 Length:2606        Length:2606        Min.   :1.000   Min.   :    0.713   Min.   :    0.192   Min.   :    0.801
 Class :character   Class :character   1st Qu.:2.000   1st Qu.:   15.335   1st Qu.:   13.254   1st Qu.:   17.250
 Mode  :character   Mode  :character   Median :2.000   Median :   63.000   Median :   50.924   Median :   72.927
                                       Mean   :1.837   Mean   :  481.944   Mean   :  456.203   Mean   :  507.686
                                       3rd Qu.:2.000   3rd Qu.:  301.250   3rd Qu.:  267.232   3rd Qu.:  335.268
                                       Max.   :2.000   Max.   :13135.000   Max.   :12910.372   Max.   :13359.628
     Units          Indicator         Cause          Validation      Population         Age           Severity
 Min.   :1.000   Min.   :1.000   Min.   : 1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000
 Median :2.000   Median :2.000   Median : 4.000   Median :2.000   Median :2.000   Median :4.000   Median :2.000
 Mean   :1.609   Mean   :1.555   Mean   : 4.577   Mean   :1.946   Mean   :2.165   Mean   :2.996   Mean   :2.008
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.: 6.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:3.000
 Max.   :5.000   Max.   :4.000   Max.   :10.000   Max.   :2.000   Max.   :3.000   Max.   :4.000   Max.   :3.000
```

Figure 3 here shows the statistical summary in table format for all the variables, in the dataset. This table highlights the most important values of min, max, median and mean.

Predictive Modeling:

```
> model
  [1] 3 1 2 3 3 1 1 3 1 3 3 3 2 3 3 2 3 1 1 3 3 1 3 1 3 3 1 3 3 3 1 1 2 3 3 3 3 2 2 3 2 3 3 3 2 1 2 3 3 3 1 3 3 3 2 1 3 3 3 3 3 1 1 3 3 3
 [67] 3 1 3 1 3 1 3 2 1 2 3 2 3 2 3 1 3 3 3 3 1 3 1 1 3 1 1 3 3 2 2 1 1 2 3 1 1 3 3 1 1 1 3 2 1 3 3 3 3 3 1 2 3 2 3 1 2 2 1 1 3 3 2 3 3 3
[133] 1 1 3 1 3 1 3 3 3 3 1 1 2 3 2 3 3 3 3 3 3 3 1 3 1 1 1 3 3 3 3 3 1 3 1 1 1 3 3 3 3 1 3 2 3 1 3 3 2 3 3 3 3 3 2 3 3 3 1 3 1 2 3 3 1 1 3
[199] 3 2 2 1 1 2 1 1 3 3 2 3 3 1 3 3 1 3 3 3 3 3 1 1 3 1 3 1 3 1 2 1 2 3 1 3 3 3 3 3 3 3 3 2 3 1 3 3 1 1 3 3 3 3 1 1 3 3 3 3 3 3 3 3
[265] 3 3 3 3 2 3 2 3 2 1 2 2 1 2 1 3 2 2 2 3 1 3 1 1 2 3 3 3 3 3 1 3 3 1 3 3 1 2 3 1 3 1 3 3 2 3 3 3 2 3 2 1 1 3 3 3 3 2 2 1 3 1 3 3 1 3
[331] 3 2 3 3 3 3 2 3 3 3 2 3 1 3 3 2 1 1 2 3 1 2 3 1 3 3 3 3 3 1 3 2 3 3 3 3 1 3 1 1 2 2 3 1 3 2 3 1 3 1 3 2 3 3 2 3 2 2 3 1 1 1 3 3 1 3
[397] 3 2 2 3 3 3 1 3 3 1 3 3 3 3 3 1 3 3 3 3 3 3 3 2 1 3 1 2 3 3 3 1 3 1 3 3 3 1 1 3 3 2 3 2 3 3 1 3 1 3 1 3 2 3 3 1 2 1 2 3 3 3 1 1 3 2
[463] 3 1 1 3 3 1 3 3 1 3 1 3 2 3 3 2 1 2 3 1 3 1 3 3 3 3 3 1 3 3 3 1 1 2 3 3 2 1 3 2 1 3 3 3 2 2 3 3 3 1 3 3 2 1 3 3 3 3 3 1 3 1 3 1 3 1
[529] 2 3 3 3 3 1 2 2 2 3 3 2 3 3 2 3 3 1 3 3 3 1 3 2 3 3 3 3 3 3 2 3 1 2 2 3 1 3 3 1 3 3 3 1 3 3 1 3 3 1 3 3 3 3 3 1 3 2 2 3 3 2 1 3 1 2 2 2
[595] 3 1 3 3 3 3 3 3 1 2 2 1 3 1 1 1 1 2 3 2 3 3 1 2 3 2 3 3 1 3 1 3 1 2 3 3 2 3 2 2 3 2 3 3 3 2 1 1 1 3 3 3 3 3 1 3 3 2 3 1 2 2 3 1 1 3 3
[661] 3 1 1 3 3 3 3 3 1 3 3 1 1 3 3 2 3 1 1 1 3 1 1 3 2 2 3 1 1 3 1 1 3 3 3 3 3 1 2 3 1 3 1 3 3 3 2 3 3 2 2 3 1 1 1 3 3 3 3 3 3 2 3 3 3 3 2 3
[727] 3 3 3 3 3 3 3 3 2 3 2 3 3 3 3 3 2 2 3 3 3 3 3 3 3 2 3 1 3 3 3 3 3 3 3 1 3 1 3 3 3 3 3 1 3 3 3 3 2 1 3 3 3 3 3 3 1 1 3 2 3
Levels: 1 2 3
> table(model, y_test)
     y_test
model   1   2   3
    1  96  77  29
    2  41  91   0
    3 115 109 225
> |
```

Figure 4 shows the results produced by the KNN model. For all the test data and was split in the training stage the model has produced the corresponding labels from the variables. Using the model MOH is able to give new data and be able to predict what level of severity an injury is most likely to be.

References:

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. *2.1*, *2*(1). https://doi.org/10.1162/99608f92.2d83f7f5

Lateef, Z. (2020, May 14). *KNN Algorithm: A Practical Implementation Of KNN Algorithm In R*. Edureka. https://www.edureka.co/blog/knn-algorithm-in-r/

Ministry of Health NZ. (n.d.). *Ministry of Health*. Health Govt. https://www.health.govt.nz/

Stats NZ. (2020, December 11). *Serious injury outcome indicators: 2000–19 | Stats NZ*. https://www.stats.govt.nz/information-releases/serious-injury-outcome-indicators-200019