Data Analysis of Number of Deaths in New Zealand

2022-11-27

Introduction

This report is an analysis of the number of deaths in New Zealand using data from the COVID-19 Data Portal. This data set was chosen after looking through different indicators on the COVID-19 Data Portal. Looking under the “Health” indicator, the “Total deaths (all causes)” data set looked interesting as the graph showed an unusual peak in the weekly total number of deaths in winter 2022, compared to previous years.

Data handling

The data is available to download directly from the COVID-19 Data Portal application in Excel format, and from the Stats NZ open data API as a JSON file. After examining data from both sources I chose to use the Excel data. The Excel file consists of two sheets, “Metadata” and “Data”. Data from the “Data” sheet was converted to CSV format and loaded into a DataFrame in R.

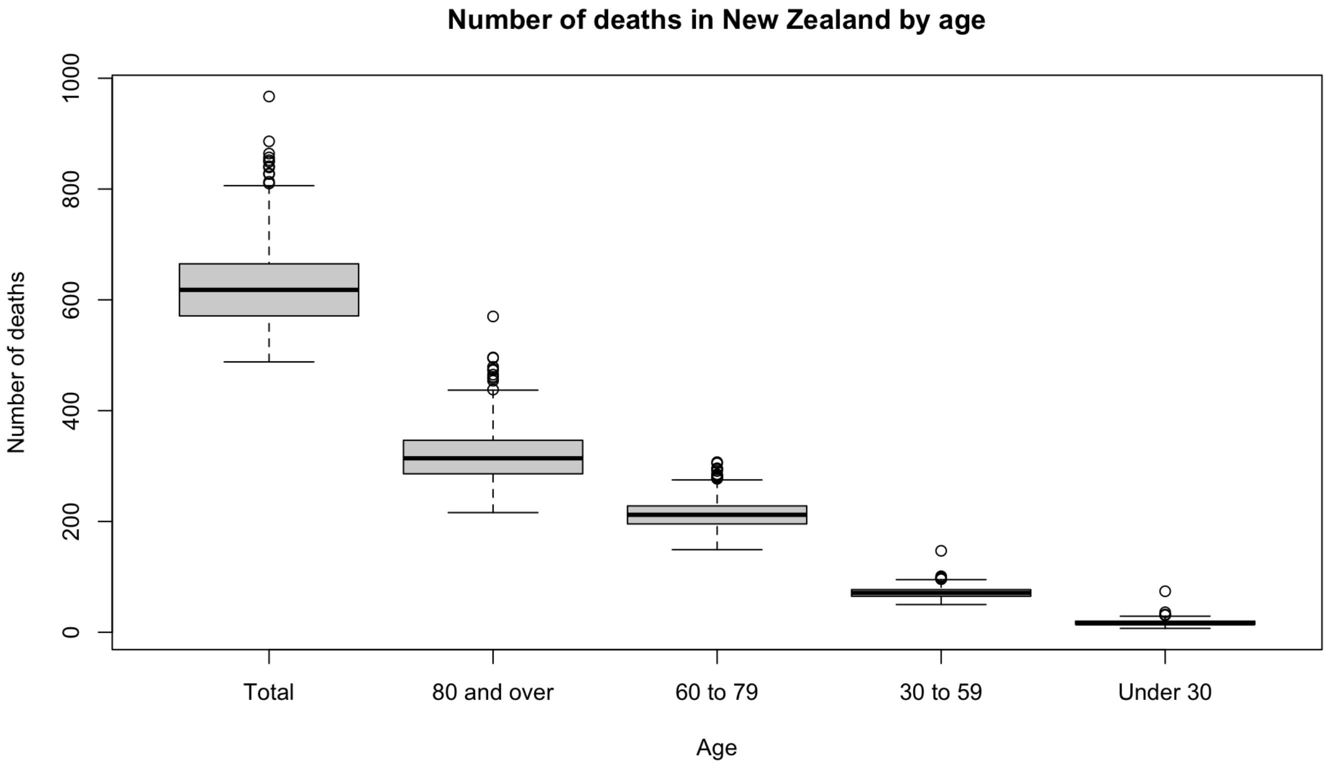
```
## 'data.frame':    3095 obs. of  10 variables:
##  $ ResourceID: chr  "CPWEE1" "CPWEE1" "CPWEE1" "CPWEE1" ...
##  $ Geo       : logi  NA NA NA NA NA NA ...
##  $ Period    : chr  "2011-01-02" "2011-01-02" "2011-01-02" "2011-01-02" ...
##  $ Label1    : chr  "30 to 59" "60 to 79" "80 and over" "Total" ...
##  $ Label2    : logi  NA NA NA NA NA NA ...
##  $ Label3    : logi  NA NA NA NA NA NA ...
##  $ Value     : int  58 202 253 538 25 70 199 278 564 17 ...
##  $ Unit      : chr  "Number" "Number" "Number" "Number" ...
##  $ Measure   : chr  "Registered deaths" "Registered deaths" "Registered deaths" "R
egistered deaths" ...
##  $ Multiplier: int  0 0 0 0 0 0 0 0 0 0 ...
```

The DataFrame consists of 3095 observations with 10 variables. I changed the type of the two variables “Period” and “Label1” from type “chr” to “Date” and “factor” respectively, and renamed “Label1” to “Age”.

To prepare the data for the visualisation I changed the order of the factor levels for the “Age” variable.

```
## [1] "Total"       "80 and over" "60 to 79"    "30 to 59"    "Under 30"
```

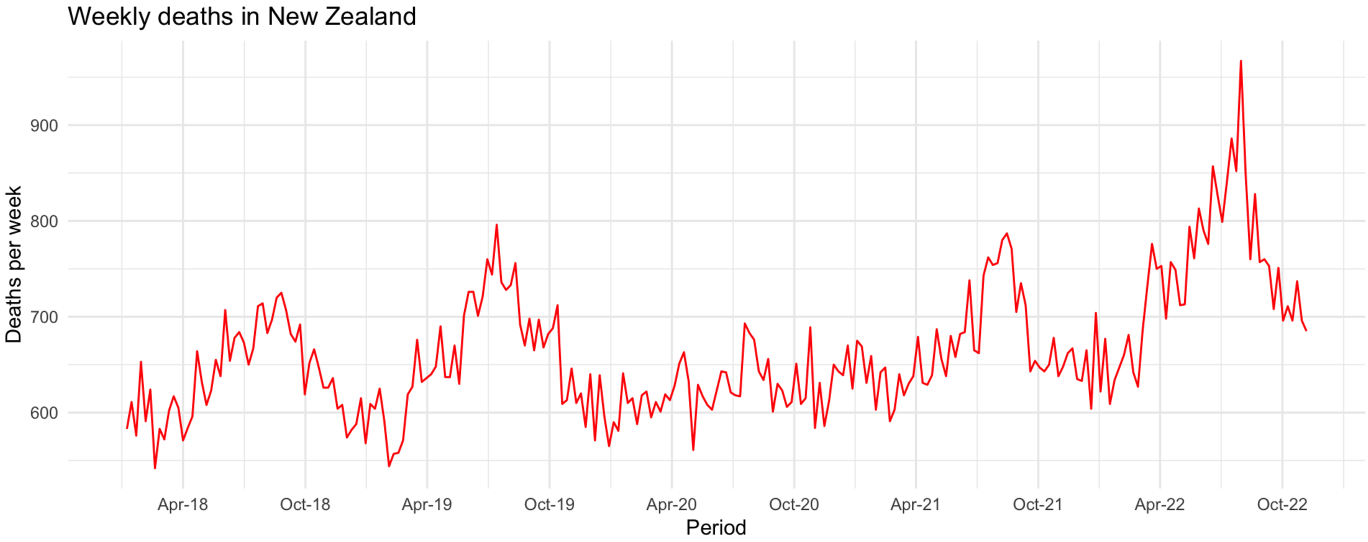
Next, I checked for any outliers in the data by plotting boxplots of number of deaths for each age group. The boxplots showed that there were outliers in each age group but they did not need to be removed as they appear to be part of the legitimate observations. I also performed missing value analysis to check for the presence of any NULL or missing values. There we no missing values found in the data set.



Data analysis and visualisation

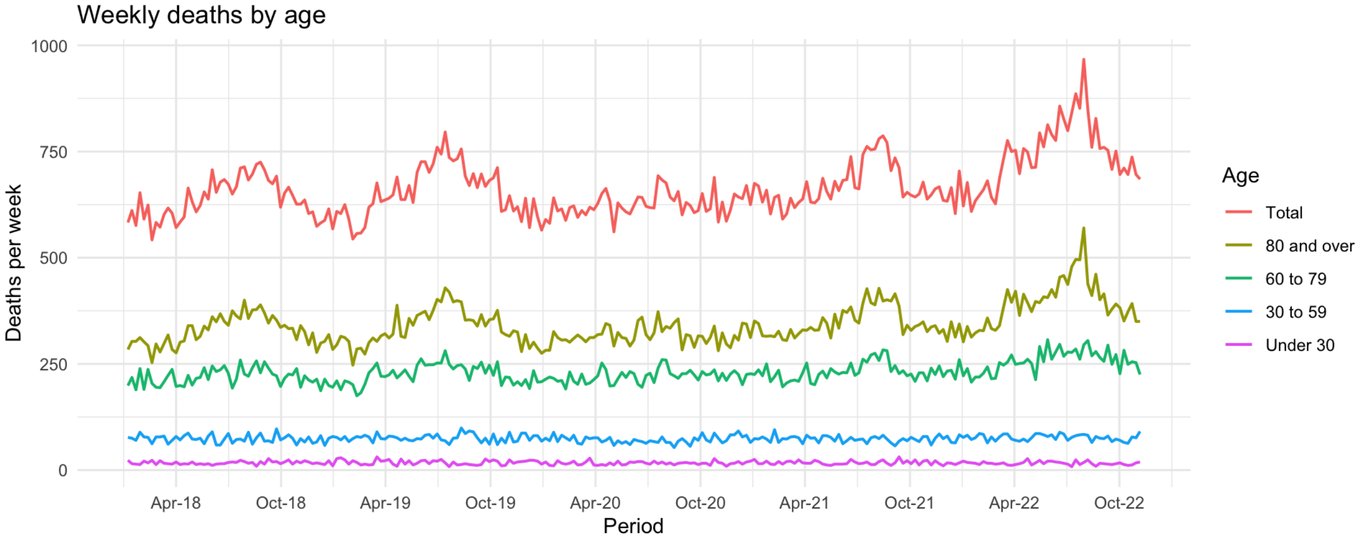
To begin the data analysis I selected the variables “Period”, “Age” and “Value” and filtered the data from 2018 to have a closer look at the recent data.

I plotted a simple graph showing the total weekly number of deaths in New Zealand (week ending on Sunday). To make the graph more readable, I chose to show the x-axis label for the “Period” six monthly, displaying only the name of the month and the year.

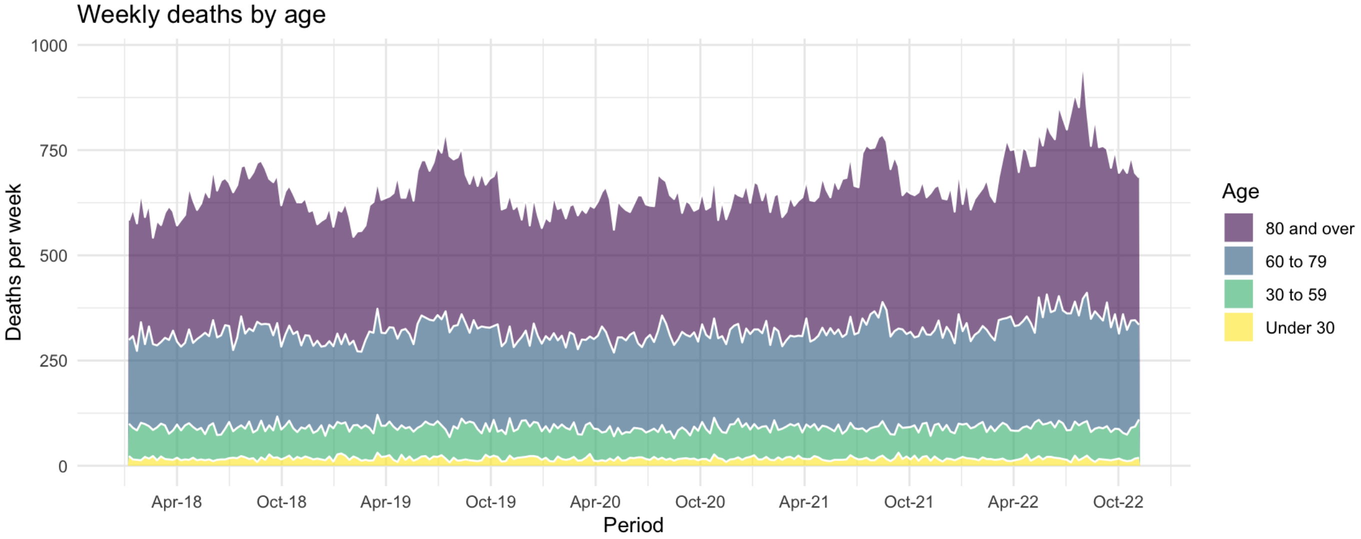


Findings

As it can be observed in above graph, the total number of weekly deaths shows regular peaks during each winter, with a strong peak in the 3rd quarter of 2022 and a flattened peak in the winter of 2020. To investigate further, I plotted a graph of the weekly number of deaths for each age group shown by different colours.



Interestingly, we can observe that there is not a large variation in the number of deaths in the age groups “under 30” and “30 to 59” and even “60 to 79”. The graph shows the increase in the total number of weekly deaths in the third quarter of 2022 results from the increased number of weekly deaths in the age group 80 and over. We can see this more clearly in an area chart.



Improvements and suggestions

To further investigate the results of the increased number of deaths for 80 and over age group, I would like to include an analysis of the number of weekly COVID-19 related deaths by age.