DATA301 - Group Project Report

31/08/2020

Police data group project

Team members

Abbey Bremner - 300436003
Shaun Godinet
Keiryn Hart - 300428418
Josh Yu

1 Background and Data

As the pandemic has unfolded around the world many studies have emerged looking at how the changes in social systems have effected crime and violence. “In the United Kingdom, calls, emails and website visits to”Respect“, the national domestic violence charity, have increased 97%, 185% and 581% respectively”(WHO, 2020). “The UN has described the worldwide increase in domestic abuse as a”shadow pandemic" alongside Covid-19. It’s thought cases have increased by 20% during the lockdown, as many people are trapped at home with their abuser"(Mohan, 2020). Articles such as the above sparked interest among our group and made us want to investigate if the same patterns are present in New Zealand.

Our initial question evolved around looking into domestic violence rates specifically as many countries saw a rise in this crime in particular. However the system used for reporting types of crime (ANSOC - Australian and New Zealand Standard Offence Classification) has domestic violence as a subcategory of Assault, this meant there was not enough data specifically on domestic violence for us to solely focus on it. For this reason we adapted our question to be focused on how the pandemic has affected Assault in New Zealand.

We have been collecting data from several sites mainly from the Ministry of Police, Stats NZ. data.govt.nz and NZ.Stat. This was not limited to researching for related data about crimes and covid-19 in NZ such journals, articles newspaper and news broadcast. The links to these sites is provided below. We have used two datasets for our analysis:

1. Victimisation time and place (Victimisation Time and Place, 2018)
2. Covid_calls for services (Covid-19 Response, 2020)

The Victimisation time and place dataset was downloaded from the information recorded in the Police dynamic operational database. This dataset represents detailed information about where and at what times of the day and week crime victimization occur in NZ. It excludes victimization where meshblock level data and Area Unit is unknown. It includes time and place where the offence occurs, number and type of offence. This data frame contains 699705 number of observations and 17 variables and it is dated from 2017 until March 2020.

The Covid_calls for services dataset was also downloaded from the police website and it contains information about crimes during the different level of Covid-19 Lockdown. These include districts, level of lockdown, type of crime and the number of times a specific crime has been committed. It also shows information about the demand in call for service and prevention during lockdown. The information has been recorded starting from the lockdown period in March 2020 until July.

1.2.1 Inside the data Example of data types in the original Victimisations data set and the basic transformations that were neccasery

Attribute	Original data type	Transformation	About
ANSOC.Division	Character	Factor	Consistent with Stats NZ and police.nz records
ANSOC.Group	Character	Factor	Official groupings based on police.nz
ANSOC.Subdivision	Character	Factor	Subgroupings of the ANSOC.Group
Location.Type	Character	Factor	
Locn.Type.Division	Character	Factor	
Meshblock	int	-	Meshblock number consistent with Stats NZ
Weapon	Character	Factor	Weapon used in crime where applicable
Number.of.Records	Int	-	Number of records

Example of Data types in the original COVID calls for service data.

Attribute	Original data type	Transformation	About
District	Character	Factor	District name as defined by stats NZ
Alert.Level	Int	-	Covid 19 Alert level for New Zealand
Calls.for.Service.and.Prevention	Character	Factor	Demand or ....
crime.and.NonCrime.Demand	Character	Factor	Crime or ....

There is some excess information that was groomed out. From the Victimisations we kept the Original index in case we need to reference back, the date converted to a Date object, the ANSOC groups that related only to Assault, the Area unit and Meshblock number and the Territorial Authority the instance occured in. From the Covid data we have kept the District, Alert level, calls for service and prevention, crime and non-crime demand, event category, event type filtered to be only those related to assault and the No of instances of each type.

1.2.2 Merging the data Merging the data was a longer process that anticipated as the structure of each data source was different. The Victimisations data was in a time series format where there was one row for each instance whereas the Covid dataset accounts for multiple occurances in each row. This meant to merge and compare pre-covid crime and covid we needed to alter the format quite siginificantly.

These alterations included matching the time periods, since Covid is still unfolding it is a short window of time to look at. This has been matched in the previous years and approximate “levels” to the Victimisations data to allow for a more acurate comparison.

There was also a difference in the location measures, the Covid dataset for a start was alot more generalised than the victimisations dataset and had already been classified into 12 regions which made it relatively easy to create an initial set of location categories that could be used for both datasets. When it came to the victimisation’s dataset the location variable was far more specific containing 65 different districts and cities, because of this it took a while to sort through all the locations in order to create a matching location column for each dataset where the location analysis could be based on the same areas for all the data.

As touched on above, the victimisations dataset was in a time series format and the Covid dataset accounted for multiple events in one row with a count. Obviously when doing a merge we could not have half the data in time series format and half not, to counter this we decided to create a ‘sum’ of all the individual instances

based on the Location, Event, year and Alert level for each of the datasets so that they both had a column with a relatable scale to compare.

After prepping the data unnecasery and duplicate columns the merge was completed using rbind.

---

2 Ethics, Privacy and Security

There are many ethical and privacy issues to consider when dealing with any police data, for this reason published data on the police.nz website is already cleaned and anonymised. This was useful for our process in that we could be confident with the level of ethical preparation that went into our raw data. We didn't have to do any anonymisation prior to merging our data, however there were still some possible issues we discussed.

2.1 Ethical considerations An ethical responsibility when publishing results is to ensure that they are accurate and fairly representative. Due to this being an unfolding pandemic where data and research is still emerging we decided to focus on a broader location rather than fine grained spatial analysis. This helps us ensure that we are accounting for possible variability in the data that may arise from limited and evolving data.

A huge ethical concern when looking at Covid data currently is identifying those who have contracted the disease and protecting the identity of those individuals and communities from ridicule and judgement. Although this is not a direct concern of ours since the Covid data we have used doesnt specifically identify cases it only focuses on the time period based on area, it is important to be aware of the current social climate around publications involving covid.

Assault is a sensitive topic to many people especially victims, with this being said it was a critical concern of ours that any disclosure of patterns particularly involving location was accurate and unbiased. We also wanted the data and information around location to be accurate enough that our conclusions about geographical locations were true and not misleading or defamatory to those areas. Misleading facts about crime rate can lead to decreased house price as well as a fear for the safety of residents and visitors in that area (Ceccato & Wilhelmsson, 2019). It can also cause unnecessary panic for not only individuals but also businesses, primarily those in hospitality such as businesses that operate late at night.

There was recently an article published about the increased rates of assault in Courtney place, Wellington. This increase was linked to alcohol use and large gatherings of public. (Zealand, 2020) Articles like this, although true may discourage new businesses or customers from this area and may create a negative perception of not only the area but also the businesses operating. We wanted to avoid running the risk of something like this being a result of our research.

A concern of ours is that inaccurate or misleading information can cause unnecessary stress and distrust between communities and their local law enforcement, this is a serious concern as public opinion often influences authorities to investigate or make changes. If these changes are based off of inaccurate facts, not only does this waste time and money but it also runs the risk of creating a biased system of justice.

2.2 Privacy concerns The biggest privacy concern when working with victimisations data is ensuring that victims cannot be identified and their personal information is protected.

Since the data is already anonymised when it is published onto the police NZ website there isn't a lot of concern for us around privacy.

The main privacy concern for our data was looking at such specific locations that information could be matched with other data and possibly identified, we have addressed this concern by looking at spaces no

smaller than regions. This keeps the data anonymous and protects people in those areas from unnecessary fear as discussed in the above section.

The police website states that “To protect privacy of individuals, sensitive details that cannot be released at a detailed “time and place” level have been removed. Such details include victim demographics, homicides and other than burglary victimizations that occurred in dwellings." (Victimisation Time and Place, 2018)

Because the data is collected from a reputable source we are confident in the quality of the privacy measures that have been taken.

2.3 Secure project management We used a private github repository to manage our project and store data, this ensured that only those with an access invitation may access the files. This provided security not only for the data but also any exploratory data analysis that we produced. Its important to have good security measures at all stages as not all information may be published, its crucial that work stay private until it has gone through a review process to ensure accuracy before it is finalised for publication.

By using github we had clear documentation of who edited what files, ensuring that accountability for changes was taken and any errors were easily traceable. The repositories were also stored locally and updated from the shared space regularly, meaning there was always multiple back ups of the project information protecting us against accidental loss or internet problems.

We used licensed and well known programs such as github, Microsoft teams and trello for communication and management, by not using unknown third party applications it helps keep everything secure.

---

3 Exploratory data analysis

3.1 Time period analysis For us, we are interesting whether some type of crimes rate are higher during Alter_level in 2020 compared with 2017,2018,2019. Therfore, I thought if we explore “ANZSOC.Group”(victimisation) “Event”(Covid) columns which is category of Assault with “Year” together,it would be interesting to have a look rather than exploring other columns. I tried to explore three different way, one way is explore the victimisation dataset(each year) and the other way is explore covid dataset(2020) and lastly, explore together by using merged dataset which are row bind (victimisation with covid).

Before explore the dataset I used the several library such as library(dplyr) ,library(tidyr),library(knitr), library(tidyverse),library(ggplot2)and library(ggpubr) to extract particular data and to visualise plot.

Firstly I explore the “Victimisations” dataset by taking several step , I seperate the day-month-year column to extract three different year. and before visualise plot, I generated the number of Assault(Five different categories)in each year by using count() function. As expected, result shows “Common Assault” categories has the dominant all three different years. However, I could found Assault “Serious Assaults resulting in Injury” is hugely increasing by every year compared with other Assault categories.

I visualised three year with number of Assault reported as using bar chart in same grid, so we can see more clearly what I explained above. It shows “Serious Assaults resulting in Injury” Assault how tdiffer from 2017 to 2019 more clearly.

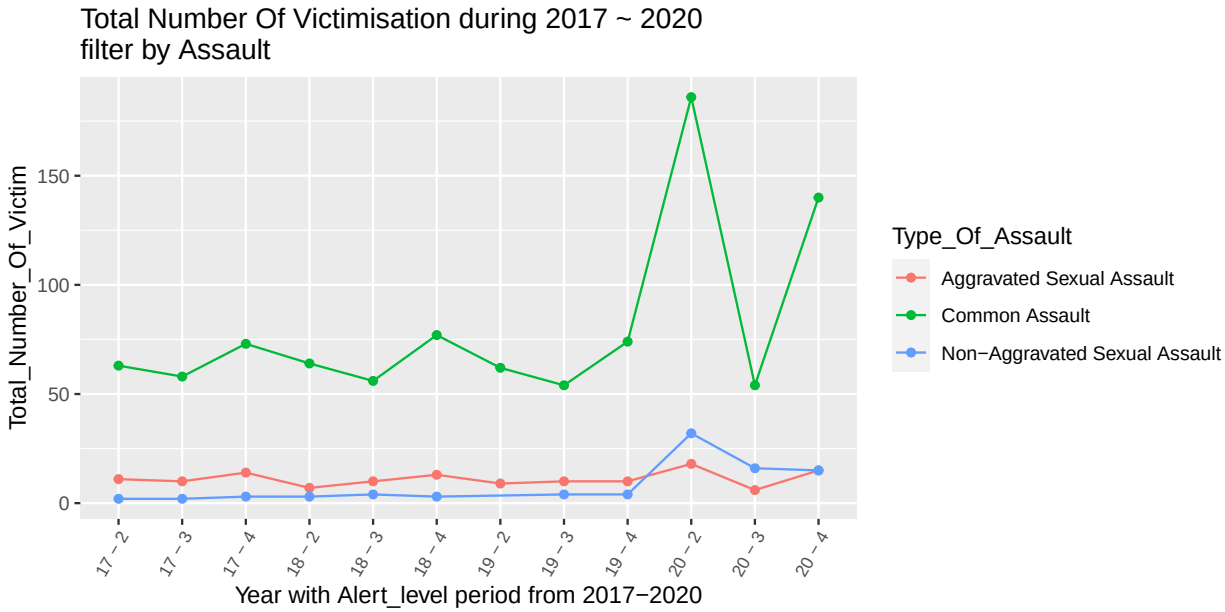
Secondly, I explored the “Covid calls for service cleaned” dataset same as “Victimisation dataset” but it shows four categories and only three of them are commom category with victimisation. I did not handle it as we want to explore just “Covid” dataset. The plot shows “Non-Aggravated Sexual Assault” and “Aggravated Sexual Assault” categories are around 30~40 times happen.

Lastly, I explored the dataset which is merged above two dataset. I extract the common Assault which both original dataset has, and I added new column “year-level” to compared time pariod. The reason why I have taken this step is covid dataset does have “date” column like month or year, but we can still assume the

month by considering alter level such as level4 = March, level3 = April and level 2 = May. Later part when I visualise time-series plot I used year_level column as my time period.

I visualised time-period plot by using line graph and I generated three line which are representing different type of Assault categories. From the line graph result, we can find interesting features that all type of Assault are increased during Covid-19 period,especially if we look at 20-2 axis(which is alter_level2 in 2020),all three type of crime pick highest number of victimisation compared with 2017,2018,2019. During 2017, 2018 and 2019 all the assault shows not really different from each other but when they compared with 2020 we can clearly see the difference.

This result could suggest that Covid-19 period changed the crime rate or has some correlation with crime. I found ariticle (Boman IV & Gallupe, 2020) which proved that during covid-19 period some type of the crime actually increased in USA, which can support this line graph result.



3.2 Investigating predictability through regression An important part of research is application to the real world, often this arises in the form of prediction. Investigating crime around covid having the ability to make good predictions around problem areas could prove to be very useful as the pandemic unfolds. Predicting crime rates is nothing new, the following research project found that population number, bars, churches and schools were the most important 4 predictors in a research project comparing linear regression, logistic regression and gradient boosting. (Ingilevicha & Ivanov, 2018)

I focused on 3 models of varying complexity to investigate the relationships within data and its predictors with a goal to predict the numer of victimisations. The three models were a linear regression:

$$NoOfVictimisations \sim Event + Year + AlertLevel + Location$$

A linear regression with log transformation applied to numeric predictors:

$$log(NoOfVictimisations) \sim Event + log(Year) + log(AlertLevel) + Location$$

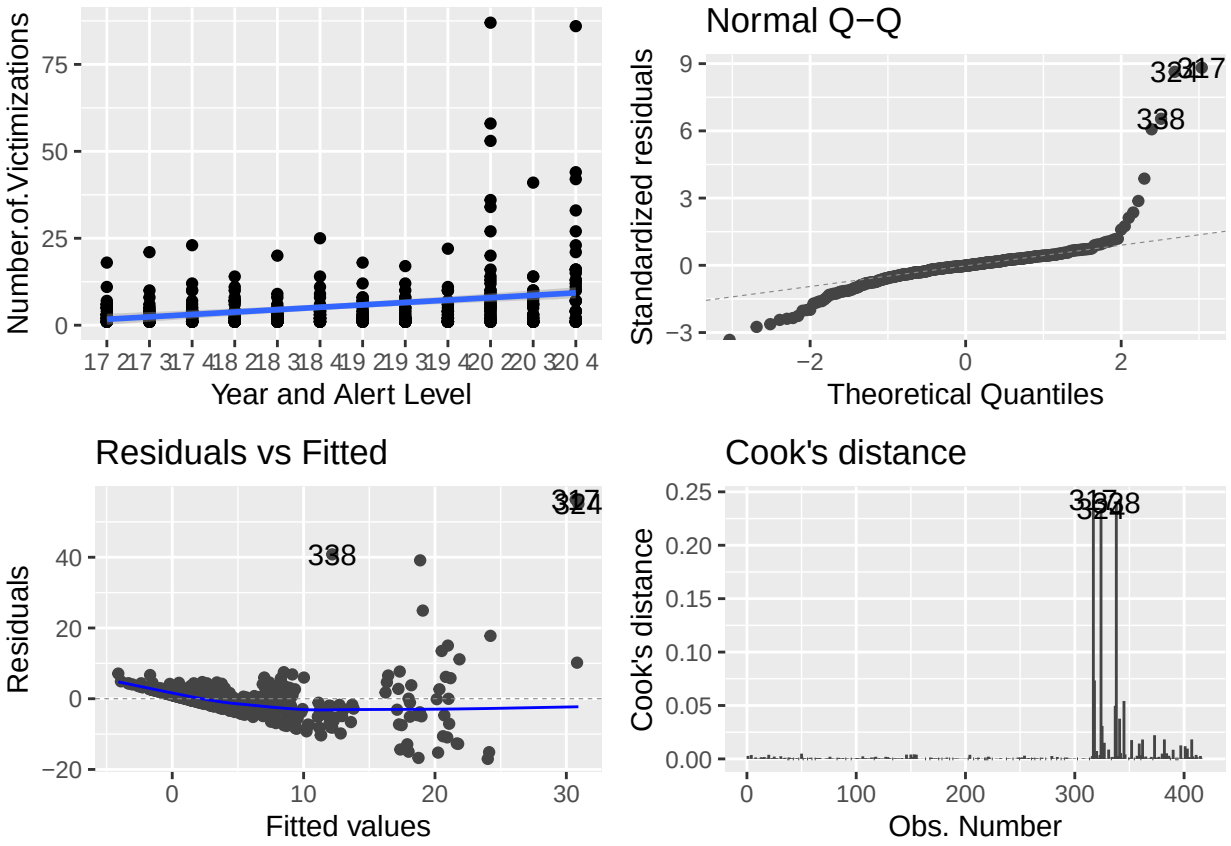
And a generalised additive model (GAM):

$$NoOfVictimisations \sim Event + s(Year) + s(AlertLevel) + Location$$

The below output is the fitted model and diagnostic plots for the linear regression.

```
## Warning: `arrange_()` is deprecated as of dplyr 0.7.0.
## Please use `arrange()` instead.
## See vignette('programming') for more help
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

## Warning: Removed 1 rows containing missing values (geom_segment).
```



From this output you can see that the assumption of normality is obviously violated as there is a significant difference between 2020 and the previous years, this is reflected in the normal QQ plot. The variance is also non-constant as depicted by the funneling in the residuals vs fitted. There arent any significant outliers with a cooks distance greater than 1. From this Imoved onto a log transformed model which helped alot with the assumptions of normality and constant variance.

Before looking at the metrics of model performance I explored some feature selection using forward and backward feature selection on the log transformed model and the gam, due to the violations of the assumptions of normality and constant variance of the plain linear model I have excluded it.

There is severe multicolinearity for the gam model between all predictors, values around 181 were found, in the context of a generalised linear model this means there will be concurvity present which describes nonlinear dependencies among the predictor variables. For this reason i am drawing the conclusion that the linear model with the log transformation is the best from the selection i have explored.

The below output is the summary output from the reduced linear model with the log transformation.

```
## Warning: package 'gtsummary' was built under R version 4.0.2
```

Characteristic	Beta	95% CI	p-value
log(year)	1.8	0.73, 2.9	0.001
Event			
Aggravated Sexual Assault			
Assault not further defined	2.2	1.9, 2.4	<0.001
Common Assault	1.4	1.2, 1.6	<0.001
Non-Aggravated Sexual Assault	-0.13	-0.35, 0.09	0.2
Serious Assault Not Resulting in Injury	0.67	0.48, 0.85	<0.001
Serious Assault Resulting in Injury	0.48	0.29, 0.68	<0.001
Location			
Auckland			
Bay Of Plenty	-1.0	-1.3, -0.80	<0.001
Cantebury	-0.67	-1.0, -0.29	<0.001
Canterbury	-0.94	-1.2, -0.67	<0.001
Hawkes Bay/ Gisborne	-1.1	-1.4, -0.88	<0.001
Northland	-1.6	-1.8, -1.3	<0.001
Other	-1.5	-2.7, -0.34	0.011
Southland/ Otago	-1.4	-1.7, -1.2	<0.001
Tasman/ South-West Coast	-1.7	-2.0, -1.4	<0.001
Waikato	-0.99	-1.2, -0.76	<0.001
Wellington/ Lower-Central North	-0.79	-1.0, -0.56	<0.001
West Coast/ Upper-Central North	-0.99	-1.2, -0.75	<0.001

For the linear model with a log transformation the RSE of 0.5827 tells us the average amoun that the response will deviate from the line of regression. The r squared value tells us that approximately 58.27% of variability found in the no of victimisations (response) is explained by the predictors. This tells us that the model is doing okay but not great. There is likely an important predictor that is missing that would help in better explaining the model. In the referenced article (Ingilevicha & Ivanov, 2018) they found that population was an important predictor, it is possible that adding in population to this model may improve the R squared value and improve the model. The predictor F stats and p-values mostly all suggest that they are significant predictors, a low p-value indicates that when testing the null hypothesis that the given beta = 0 will be rejected.

This was only a brief exploration but from this we can see that the year, event and location are important predictors for the number of victimisations. Further predictors should be explored such as population to develope a better model should prediction be pursued in further study.

3.3 Locational Analysis with Maps Another interesting thing we looked to investigate was the relationship between the different time periods, the amount of reported accounts of assault and the various regions around New Zealand.

In order to do this I used packages including (tmap) and (dplyr), (tmap) is a powerful map making package that allowed me to map the differences between location, time and rate of assault with relative ease and some minor changes creating easy to understand static maps that show the changes in reported events over the past 4 years and relative 'Alert levels'.

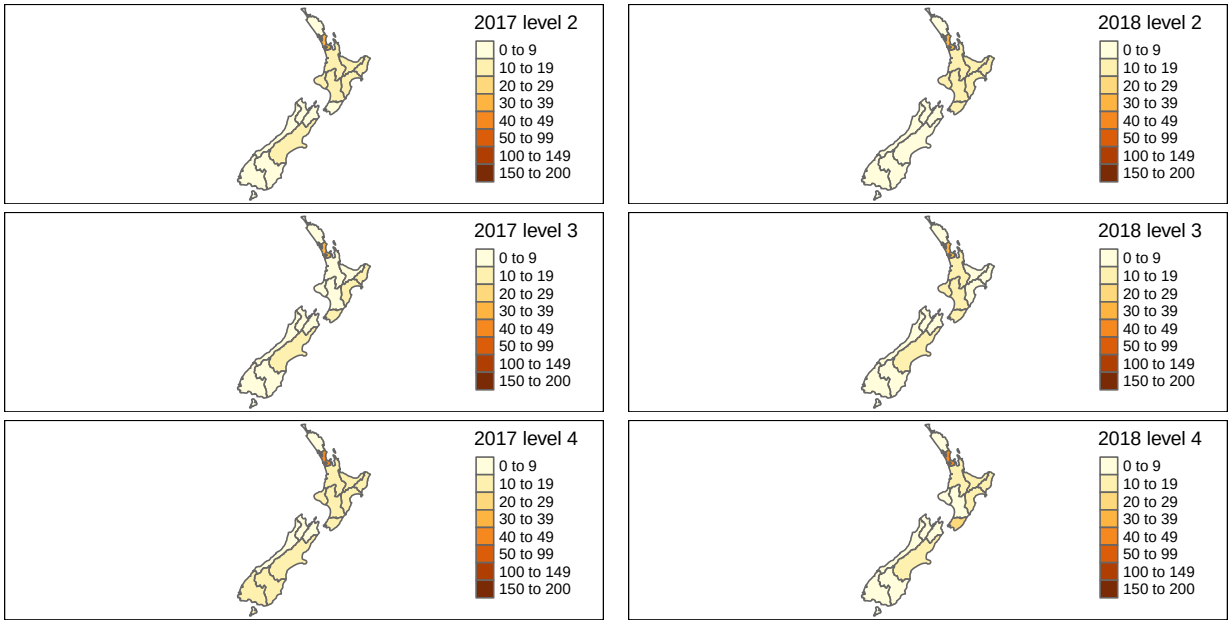
the dataset from tmap was a dataframe that contained geometric and other spatial information that made it possible for me to have an accurate map of New Zealand, in order to have the maps projecting the information from our datasets I had to do a bit of manipulation to both our data and the map information in order for the two to have similar categories and specified data to merge together. once this was completed I was able to create 12 individual maps which showed the changes in reported cases of assault over the 3 'Alert Levels' during the 4 different years under investigation.

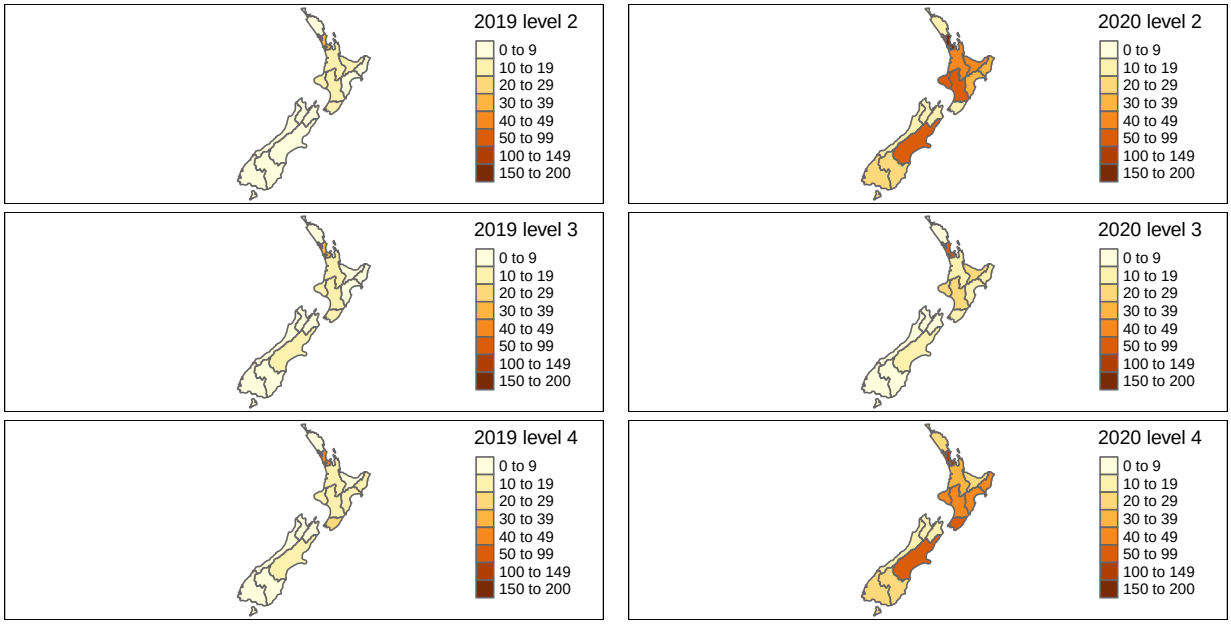
Once the maps were created it was instantly clear that there were some major differences between the rate of assault during Covid-19 in comparison to the same time periods in previous years. Firstly in level 4 of

2020 we can see that alot of the country is showing much darker shades of brown than the relative time periods in all of the previous years, indicating a spike in the level of assault in New Zealand. this observation is supported by a recent study in Dallas which observed an initial spike in the reported cases of domestic assault in the initial weeks after going into lockdown (Piquero, Riddell, Bishopp, & Narvey, 2020). The study also notes a decrease in the rate of reports after a few weeks of being in lockdown, this trend can also be seen on the maps below where the rate of victimization in New Zealand once going to level 3 has decreased dramatically although the lockdown rules from level 4 to 3 don't change much, this could also be due to the fact that level 3 was a slightly shorter period of time in comparison to the other levels.

another interesting thing to note is that the rate of assault appears to increase again once moving from level 3 down to level 2 where it appears that there seems to be a similar rate of assault to that of level 4 in 2020. Once again, when comparing this map of level 2 in 2020 to similar time periods in previous years we can see a massive difference in the rate of assault which again indicates the ongoing impact that the covid 19 lockdowns have had on assault in New Zealand.

These observation helps to support our initial question of whether the lockdown has impacted the rate of assault in New Zealand as there is clearly an elevated rate of assault over a large portion of the country during the lockdown periods in comparison to previous years and this is especially apparent during levels 4 and 2.





4 Individual and group contributions

As a team we all contributed to the research portion of the project, completing mindmaps around discussions as a group. We also worked collaboratively on findind suitable data and discussing possible ideas and issues around potnetial data sets. We also wrote the Ethics and Privacy section as a group alongside a discussion around the topics.

Abbey As the repository manager i was responsible for setting up and maintaining the git repository, i also created the report framework and documented the grooming and merging process with input from Shaun and Kieryn. I completed section X.X exploring relationships between our variables and took responsibility for ensuring the final markdown file knitted properly and everything pushed to the git was included in the final report.

Shaun Data manager

Kieryn As the biblio Manager I was in charge of ensuring that all our sources we used to inform our report and questions were correctly referenced and cited in the report. I also handled the process of merging our datasets together and produced the maps seen in the exploratory analysis.

Wonsang Yu (Josh) I have contributed on this project mostly explore the data such as visualise plot to find interesting feature or some cleaning stuff for dataset, this is documented in section 3.1 .Also when developing our scientific question, I was trying to find relative article or scientific journal to support our project or find connection.