

DATA301 - Group Project Report

Abbey Bremner

18/08/2020

Police data group project

Team members

Abbey Bremner - 300436003

Shaun

Kieryn

Josh

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. ** Insert relevant stats ** We are interested to look at whether 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 souley focus on it. For this reason we adapted our question to be focused on how the pandemic has effected 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. Victimization time and place (<https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz/victimisation-time-and-place>)
2. Covid_calls for services (<https://www.police.govt.nz/about-us/statistics-and-publications/data-and-statistics/covid-19-response>)

The Victimization time and place dataset was downloaded from the information recorded in the Police dynamic operational database. (www.police.govt.nz) 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. The dataset 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 lockdown during covid-19. 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 Data types in the original Victimisations data set.

Attribute	Original data type	Transformation	About
ANSOC.Division	Character	Factor	Consistent with Stats NZ and police.nz records
Year.Month	Character	Date	yyyy/mm/dd format
Month.Year	Character	Date	yyyy/mm/dd format
ANSOC.Group	Character	Factor	Official groupings based on police.nz
ANSOC.Subdivision	Character	Factor	Subgroupings of the ANSOC.Group
Area.Unit	Character	Factor	Stats NZ area unit names
Location.Type	Character	Factor	
Locn.Type.Division	Character	Factor	
Meshblock	int	-	Meshblock number consistent with Stats NZ
Occurrence.Day.Of.Week	Character	Factor	7 levels corresponding to day of the week
Occurrence.Hour.Of.Day	Character	-	24 hour time representing approximate hour
Table.1	Character	Factor	
Territorial.Authority	Character	Factor	TA consistent with Stats NZ outputs
Weapon	Character	Factor	Weapon used in crime where applicable
Number.of.Records	Int	-	Number of records
Victimisations	Int	Int	Number of victimisations for this crim in this area

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
Event.category	Character	Factor	Administrative or
Event.type	Character	Factor	Crime type
No	Int	-	Number of instances of each Event.type

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 occurred 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 occurrences in each row. This meant to merge and compare pre-covid crime and covid we needed to alter the format quite significantly.

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 accurate comparison.

There was also a difference in the location measures. ***** Kiernan do you want to fill this in *****

After prepping the data unnecesary 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 doesn’t 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. (<https://www.tandfonline.com/doi/full/10.1080/2578983X.2019.1662595>) 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. <http://wellington.scoop.co.nz/?p=130187> 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.” <https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz/victimisation-time-and-place>

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

Example of code in report

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. Therefore, I thought if we explore “ANZ-SOC.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.

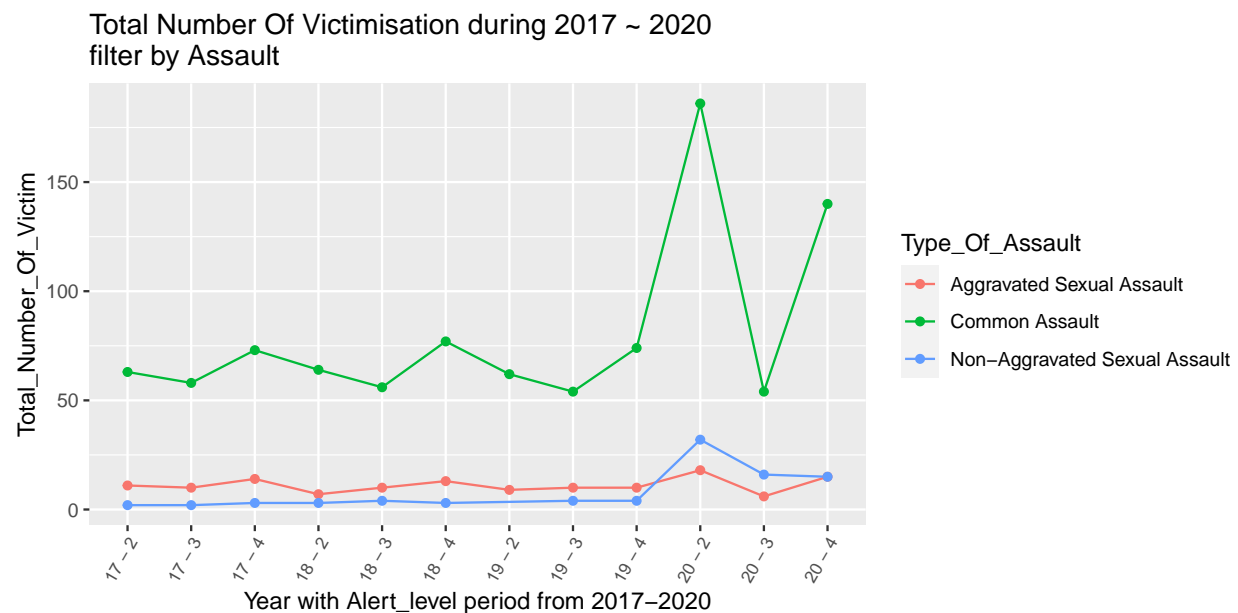
Secondly, I explored the “Covid calls for service cleaned” dataset same as “Victimisation dataset” but it shows four categories and only three of them are commom category with victimisation. I did not handle it as

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 period. 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 article “<https://link.springer.com/article/10.1007/s12103-020-09551-3>” which proved that during covid-19 period some type of the crime actually increased in USA, which can support this line graph result.



4 Individual 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 finding suitable data and discussing possible ideas and issues around potential data sets. We also wrote the Ethics and Privacy section as a group alongside a discussion around the topics.

Abbey Repository manager

Shaun Data manager

Kieryn Biblio manager

Wonsang Yu (Josh) onsang 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. Also when our scientific question develop, I was trying to find relative article or scientific journal to support our project or find connection.