COVID-19:

The effects of a global pandemic on criminal activity

Keiryn Hart 300428418

Date of Submission: 15 / 10 / 2020

Group members:

- Abbey Bremner - 300436003
- Shaun Godinet - 300426053
- Keiryn Hart - 300428418
- Wonsang Yu (Josh) - 300463616

Executive Summary:

The goal of this research report is to build a model that can effectively predict the rate of assault cases in a given region within New Zealand given the current status of COVID-19 and the effects that it has on cases of assault in New Zealand this model will be able to accurately predict the effects that another nation-wide lockdown will have on the rates of assault in New Zealand. This will be beneficial as it will help to identify regions in New Zealand that are more at risk of having higher rates so that more can be done in those regions to help prevent and or maintain these rates to help their communities.

The methods used for these predictions is linear regression, three models have been trained on available data from the victimizations time and place dataset from the police NZ website.

All three models proved to be relatively good predictors of assault in New Zealand but there was one model that through comparison and testing appeared to be the most accurate and reliable model for predictions.

These models however are not completely accurate and are subject to inaccuracy, this is primarily due to the availability of data concerned with COVID-19 and its effects on criminal activity. With more data these models will prove to be better and will be able to accurately predict these rates and be able to identify possible hotspots within New Zealand if we were to go into another nation-wide lockdown.

From this analysis it was clear that the Auckland region experiences the highest rates of Assault both during and not during the COVID-19 lockdown periods, every region in New Zealand however did experience increased rates of assault as a result of these lockdowns, these linear regression models were able to pick up on this happening but there is still room for improvement as the amount of data available for training was not sufficient enough to help the linear regression models learn the trends better.

Background:

As Covid-19 has slowly grown around the world causing mass disruption and change to everyone's daily lives, there has been an increasing amount of studies emerging looking into the effects that the virus has had on crime rates across the globe. This caught myself and my groups interest as we started to see trends from different countries releasing studies focusing on the effect the virus has had on domestic violence and assault.

Because of this we decided that it would be interesting to have a look into these effects on New Zealand's criminal activity and more specifically cases of assault.

The aim of this New analysis in a similar fashion to the exploratory analysis that my group did earlier is to look into the effects that COVID-19 has had on assault in New Zealand, this time however I am looking to identify what the rates of Assault would be if New Zealand were to go into another nationwide lockdown and what regions within New Zealand would be expected to experience the highest rates.

The specific questions I will be looking to answer in the report are:

- What would the rates of Assault be if New Zealand were to go back into a nationwide lockdown?

- Which regions of New Zealand are likely to be effected by another lockdown the most?

These questions are looking at helping identify where in New Zealand and in what volume would this criminal activity occur, the answers to this are important as they could be very useful to the New Zealand police and the general public in helping identify where potential hotspots for assault could be in New Zealand if another lockdown were to occur so that we can look into prevention methods and generate more support where support is needed.

COVID-19 is a virus that has effected everyone in many different ways and it is important that we look to identify the negative implications of the virus so we can help people to overcome the challenges they are facing and help to avoid any more unnecessary stress and problems.

Data Description:

The datasets that are being used in this analysis are the victimizations time and place dataset from the New Zealand police website and the COVID-19 calls for service dataset which is also from the New Zealand police website.

The victimizations time and place dataset contains data that is mostly categorical with the exception of 4 variables that are numeric, there is obviously a data variable to which is indicates that the data is time series.

The COVID-19 calls for service dataset is similar to the victimizations dataset in the sense that a majority of the variables are categorical and only the (No) number of instances at each level is numerical. Shown below is a table of the data from both datasets which has been taken from the initial exploratory analysis report.

Victimisations data:

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

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 ....

The variables that I ended up using for my detailed analysis were:

- Location:
    - This was a variation of the COVID calls for service "District" and victimizations "Area Unit" – these were the different regions in New Zealand.
- Number of records – continuous integer (response variable)
- Alert Level – factor – different COVID-19 Alert Levels
- Month and year – integer – factor
- Unemployment rate – numeric – gathered from stats NZ website

In regard to the completeness of the datasets used in this analysis all of the datasets were complete and had no missing values.

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.

Ethics:

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 I am focusing on a broader location rather than fine grained spatial analysis. This helps 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 contacted the disease and protecting the identity of those individuals and communities from ridicule and judgement it is also essential that these people are aware of their information being used for statistical analysis. Although this is not a direct concern of ours since the Covid data being 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 that any disclosure of patterns particularly involving location was accurate and unbiased. The data and information around location needs to be accurate enough that conclusions about geographical locations were true and not misleading or defamatory to those areas. 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.

Another concern 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.

Privacy:

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 this information is protected.

The main privacy concern for this data was looking at such specific locations that information could be matched with other data and possibly identified, this is avoided with the use of regional zones. It 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). This is reassuring for privacy concerns.

Security:

All the analysis and data management done in this project was initially through the original GitHub used in first stage or now on my local computer than only I have access to. This provided security not only for the data but also any exploratory data analysis. It’s important to have good security measures at all stages as not all information may be published, it’s crucial that work stay private until it has gone through a review process to ensure accuracy before it is finalised for publication.

All of my analysis and data management was done through software that is safe and trustworthy or has been handled on my personal laptop, this ensures that the programs in use such as R do not run the risk of leaking information.

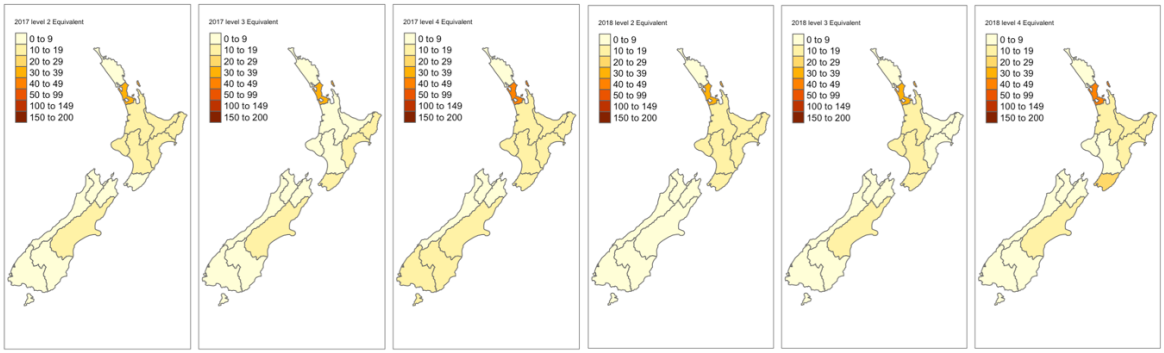
Exploratory Data Analysis:

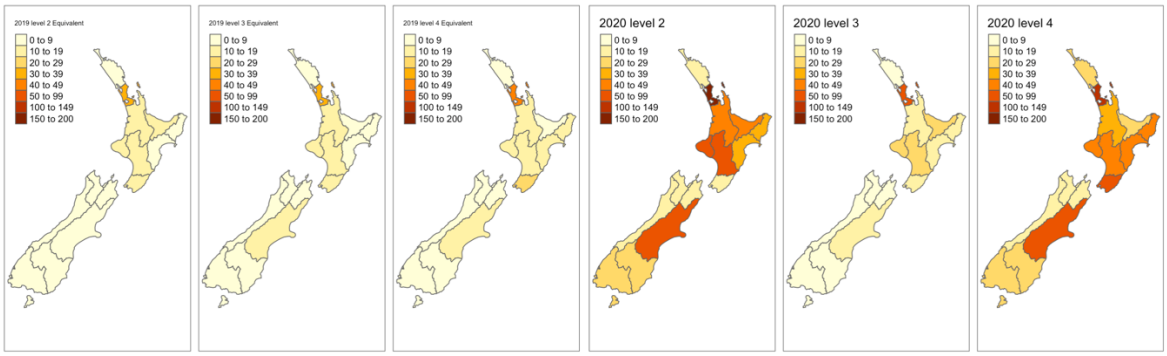
Initially for my exploratory analysis I decided to look into my locational analysis again with the use of the maps that I used last time, the labels have been adjusted to suit the plots a bit better as since there were no actual alert levels in 2017, 2018 or 2019.

The locational analysis allows us to see the relationships between the different time periods and the rates of assault in the different regions around New Zealand. These static maps shown below are easy to understand and give a good representation of the changes in reported events over the past 4 years during this small time period.

To ensure that the data was able to be used in making these maps I had to do a bit of data manipulation to the datasets I used in this analysis, the manipulation included ensuring that the labels used for the locations in the datasets were one, the same as each other so they could be combined together and two, the same labels that the mapping software used so that they could also be combined together so that the maps would register the values in the dataset.

Looking at the maps shown below, it is almost instantly clear that there were obvious differences in the rates of assault in previous years when compared to the COVID-19 period. In level 4 of 2020 we can see that in many parts of the country the map is showing a much darker shade than the relative time periods in previous years, this is a good initial indicator that there has been a rise in the rates of Assault during COVID-19. This observation is similar to that of a study in Dallas where they also experienced a spike in the reported cases of domestic assault initially after going into lockdown (Piquero, Riddell, Bishopp, & Narvey, 2020). That same study also reported a decrease in the reported cases after a few weeks of being in lockdown, this is interesting to note because we can also say from the looks of these maps that after moving to level 3 which was roughly a month later, the rates of assault dropped substantially.





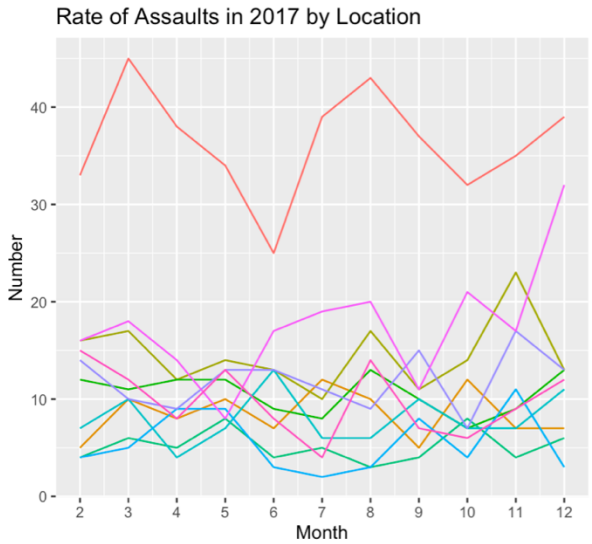
It is also interesting to note that after level 3 there appeared to be a pretty drastic increase when New Zealand moved down to lockdown level 2 which looked to be experiencing a similar amount cases as what was experienced in level 4. When comparing the cases of assault from level 2 to previous years however, we can still see that it is much larger than that of similar time periods in previous years.

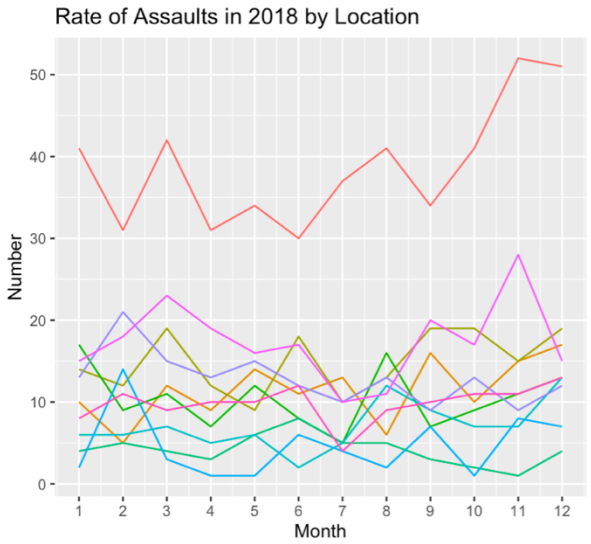
These observations are initially helpful when considering the sub-question of this report which is considering which regions are likely to be most affected by another lockdown, as we can already start to identify regions that have seen more cases than others.

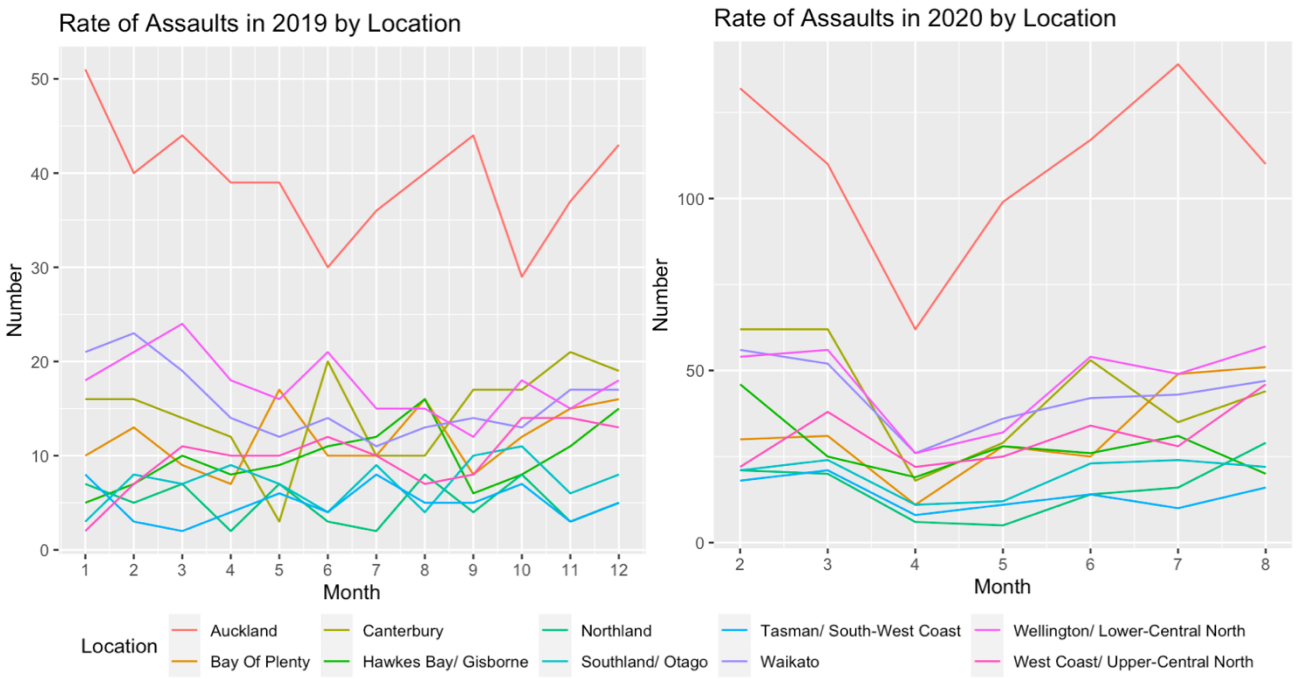
After analysing these maps I also thought it would be good to analyse some yearly time series plots of the rates in cases of assault over the course of a year for different locations. This was just a basic case of using ggplot2 and some line graphs to plot the lines for the individual locations for year of the 4 years.

When looking at the plots it is clear that the all 4 years appear to follow relatively similar trends where Auckland experiences a much larger rate of cases per month than the likes of Hawkes bay and Northland, the most likely cause for this is the fact that Auckland's population is much larger than that of anywhere else in the country and as a result they are more likely to experience these higher rates.

We can see though, similarly to that of the maps above that there is a clear increase in the number of reported cases in 2020 than there is in the previous years, these time series plots also help to show that although the rates have increase in 2020 the trends relative to each other are still the same where Auckland always seems to experience more cases but that is pushed to an extreme in 2020 in comparison.

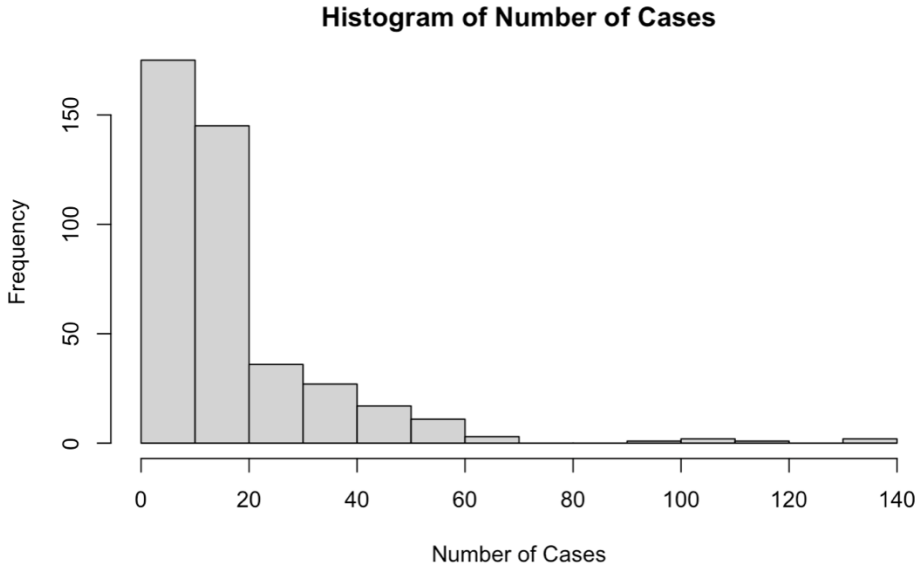






After Looking at these different plots and maps, I thought it would be a good idea to get a better understanding of the distribution of the data as this could possibly help inform decisions going forward when looking into regression and prediction.

Shown is a histogram of the frequency of the different number of cases reported in a given month over the 4 years over all locations. It is quite obvious Immediately that the distribution is skewed to the right and that the distribution is a poisson distribution. Almost all of the reported cases for a month are contained within the 0-20 cases range and from there it gradually decreases up until around 70 and then only experiences a number of cases around 100 or above which as we can see from above are from Auckland during the COVID-19 Lockdown.



Detailed Analysis:

For my detailed analysis I decided that a suitable technique to use for this data would be regression, given I am trying to find a good way to predict the number of cases that would arise from another COVID-19 Lockdown. Regression seemed suitable as it is a good way of making predictions about a singular outcome variable and given the fact that the outcome variable in question is continuous regression should work well.

There were no missing values in the data which meant that there was no need for imputation of data, this also means that all of the data is true in theory and this should prove beneficial when training models and making predictions.

Before building my models and making predictions I split my dataset into training and test sets with 70% being training and 30% being testing, the testing set does not have a 'Numbers' variable as it has been removed and is kept separately to compare results.

To begin with I created a simple linear regression model, this model included all of the variables in the dataset as the predictors aside from the 'Numbers' variable which was the target.

The first model is shown below:

$$Number \sim year + month + Location + AlertLevel + UnemploymentRate$$

This model initially showed promise with an R-squared value of 0.80 which indicates that the model is fitting quite well, given that there is quite a number of variables included in the model it is good to check adjusted R-squared too as it adjusts for this, the value for that was 0.78 meaning that around 80% of the variance can be explained by the model. The F-statistic is relatively large in this case sitting at 38 on 28 degrees of freedom but, it has a p-value of < 2.2e-16 indicating that we can reject the null hypothesis and say that there is a relationship between these variables and the Number of cases.

```
Residual standard error: 9.138 on 265 degrees of freedom
Multiple R-squared:  0.8006,    Adjusted R-squared:  0.7795
F-statistic:    38 on 28 and 265 DF,  p-value: < 2.2e-16
```

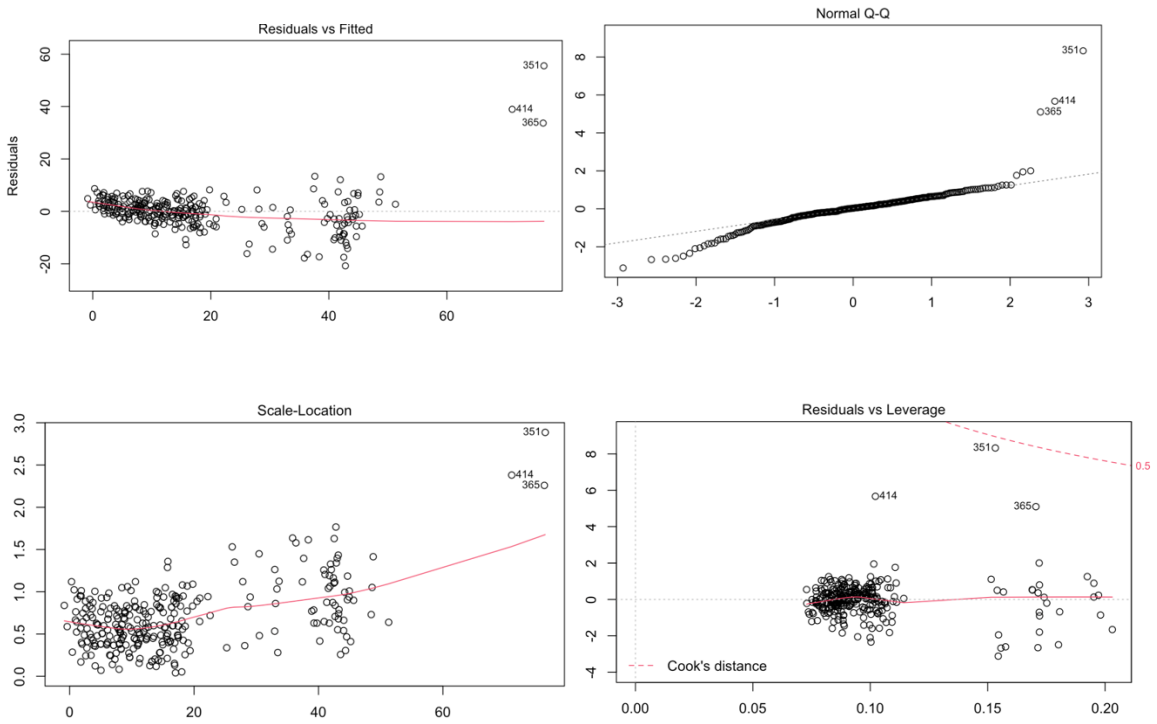
Plots:

Firstly, looking at the residuals vs fitted values plot it seems that for the most part shows not much evidence of non-linearity however there are a few points in the top right corner that may be causing the plot to skew, the red line in the middle is relatively horizontal though which is a good indication.

The Q-Q plot does happen to show a little bit of deviance from the straight line which could indicate evidence of non-normality, this is a good indication for this. A transformation of the response variable could help this issue.

The scale location plot immediately shows evidence of non-constant variance in the data, the plot is clearly skewed and the line through the middle is not horizontal at all, again, a transformation of the response variable could help this issue.

The cooks distance plot or residuals vs leverage plot shows that there doesn't appear to be any clear evidence of influential observations



After training this initial model I decided to try out some other models with minor adjustments to see If I could eliminate some of non-normality and non-constant variance.

Results with test data will come at the end on the analysis.

The second model trained is shown below:

$$Number \sim year + month + Location + AlertLevel$$

In this model I just removed the variable Unemployment rate that I decided to add to the data earlier, in the end this variable was deemed to be unimportant and useless as when looking at the output of this second model we can see that it almost produced the same R-squared value at 0.8 and multiple squared at 0.78.

```
Residual standard error: 9.121 on 266 degrees of freedom
Multiple R-squared:  0.8006,    Adjusted R-squared:  0.7803
F-statistic: 39.55 on 27 and 266 DF,  p-value: < 2.2e-16
```

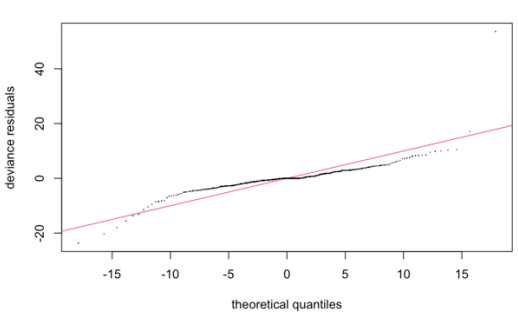
The diagnostic plots in this case were basically Identical to the previous model, this however does mean that this second model is a better model as it is less complex than the previous and does appear to show a similar performance.

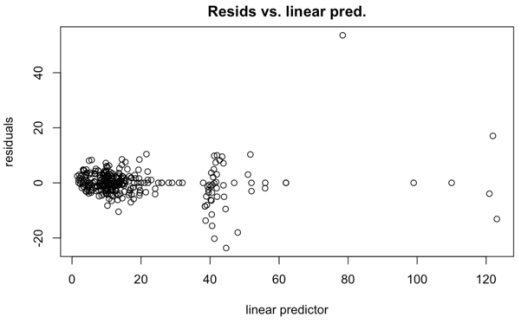
For the third model I decided to try a similar model again but this time include an interaction term between Location and Alert Level as I believed that there may be an interesting relationship between those two variables. That model is shown below.

$$Number \sim year + month + Location : AlertLevel + AlertLevel$$

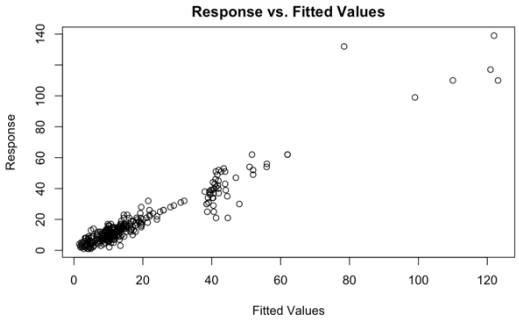
This model, with the inclusion of the interaction term showed some good promise with a reported R-squared value of 0.92 and an adjusted R-squared value of 0.9, this essentially means that around 92% of the deviance in the data can be explained by this model and this now indicates that this seems to be a better model for the data than the previous two.

Shown below are the relative diagnostic plots associated with this model, as well as the model statistics.





statistics	values
RSE	37.329
AdjRsq	0.901
Rsq	0.921



Looking at the Q-Q plot on the top left it is quite apparent that the residuals do not follow a very straight line which means that there is still non-normality in the data, the residuals vs linear plot shows that for the most part the model appears have a little bit of non-linearity but this is not completely clear. The response vs fitted plot shows a very straight line with one or two points veering to each side, this is a good indication that the model is predicting quite well which is promising.

Model Comparisons:

To compare these 3 models I decided to do likelihood ratio tests on each of the models comparing them to each other. The first comparison was between the first and second model, the p-value for this comparison was 0.84 which means that we cannot confirm that there is a significant difference between the two models, because of this the residual deviance becomes the decider and in this case it is slightly lower for the first model and therefore the first is favoured.

The next comparison was between the first and the third model, the p-value in this case is extremely low (< 2.2e - 16) which indicates a significant difference between the two models, the residual deviance value for the third model with the interaction term is significantly lower than that of the first model therefore the model with the interaction term is favoured.

Testing:

	1	2	3	5	6	7	9	11	16	17
m1	49	6	2	16	4	11	10	6	1	9
m2	49	6	2	16	4	11	10	6	1	10
m3	42	8	5	15	6	12	12	8	5	11
t1	33	7	4	16	4	5	12	10	6	11

The above table is the first 10 values for the predictions that the three models made on the test set, m1 is the first model, m2 is the second, m3 is the third and t1 is the actual values for those instances.

All three models appear to be predicting quite well based on these 10 instances, it is interesting to note that all three models over predicted by quite a lot on the first instance which was Auckland during February of 2017. It is not quite clear why this is but all three models appear to be doing so although the third model with the interaction term seems to be close in almost all of the cases of these first 10.

I decided to test on another set of data this time the covid dataset that was touched on earlier in the assignment, I adjusted the months to be September October and November for the purpose if trying to predict in a different time of the year shown below are the first 10 predictions for that set.

	1	2	3	4	5	6	7	8	9	10
m1_c	67	58	81	29	19	43	33	24	47	28
m2_c	67	58	81	29	19	43	33	24	47	28
m3_c	99	63	110	99	12	31	29	19	62	28
t1_c	160	63	135	48	20	26	64	18	57	34

M1_c is the first model, m2_c is the second and m3_c is the third with t1_c being the actual values. We can see here that on this occasion all three models appear to be under predicting the larger values on this occasion, it is interesting to note that m1 and m2 are completely identical similarly in the first test but this is obviously because as we found in the likelihood ratio tests there is no significant difference between the two and their R-squared values are almost identical aswell. The third model however still appears to be making better predictions overall although we can see on the 4th instance that it has severely over predicted that value which is interesting.

Conclusion:

Overall the aim of this report was to identify what the rates of assault in New Zealand would be if New Zealand went back into another Lockdown, obviously this is theoretical and subject to many different factors but we can still make predictions about what the possible outcomes may be in these situations.

This analysis has identified 3 different models that with the use of this data can to an extent predict the rate of assault in New Zealand at different times during the year, with those predictions we can also identify which locations in New Zealand may be more or less likely to experience higher rates of assault.

Overall the best model to use out of the three tested in this report would be the third linear regression model with the interaction between Location and Alert Level, this model appeared to make the best predictions and although there were cases where it was underpredicting or over predicting by a large margin, through analysis and comparison it looks to be the best model to make predictions regarding this issue.

My findings in this sense are quite limited as I do not have excessive amounts of data on this issue, COVID-19 has only been a problem in New Zealand since February of 2020 and as a result it is hard to build a model that can accurately identify patterns and predict these rates of assault when there only available data from 1 nationwide lockdown over a 3 month period. Trends in regard to the rates of assault over previous years can be found and can used to help better the model but overall this analysis was extremely limited to the amount of data that was available.

The data that was available was also not extremely specific, the Covid dataset only had data based on alert level which meant that information from any other dataset had to be combined into monthly segments as this was the closest possible time period for comparison. The victimizations dataset was slightly more specific but the date variable was not good and this essentially resulted as the data from this dataset only being useful in monthly format anyway (each date was the first of the month), weekdays (e.g. Monday) were included but this does not suffice for an actual date.

Overall both datasets did not have much data other than what was used that seemed like it would realistically be useful for this analysis, I tried to implement unemployment rates into the data to see if the percentage of unemployment had an effect but it did not. The covid dataset was really only useful for having new data to test the models on because since the time of our initial report, the victimisations dataset had updated itself to include all the information of victimisations throughout the COVID-19 lockdown.

A recommendation I would make in regard to an analysis of this would be to try and find more specific data related to COVID-19, the victimizations dataset is almost good enough for making valid predictions about the rates of assault if there was not a nationwide COVID-19 Lockdown, but given the fact that is not the case more information is needed so that the models can be better trained and make better predictions although the models were not terrible still, there is just room for improvement.

References:

Piquero, A., Riddell, J., Bishopp, S., & Narvey, C. (2020, June 14). Staying Home, Staying Safe? A Short-Term Analysis of COVID-19 on Dallas Domestic Violence. Retrieved from Springer Link: https://link.springer.com/article/10.1007/s12103-020-09531-7

Covid-19 Response. (2020, June 12). Retrieved from New Zealand Police: https://www.police.govt.nz/about-us/statistics-and-publications/data-and-statistics/covid-19-response

Victimisation Time and Place. (2018, January). Retrieved from New Zealand Police: https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz/victimisation-time-and-place

Zealand, R. N. (2020, August 11). Increase in late-night violence on Courtenay Place. Retrieved from Wellington Scoop: http://wellington.scoop.co.nz/?p=130187

Unemployment rate | *Stats NZ*. (2020). Stats NZ. https://www.stats.govt.nz/indicators/unemployment-rate