Victimisation Rates throughout the pandemic

Phase 2 – Individual Project Report

Raphael Dan Gueco – Student ID: 300449479

Yuelin Yao - Student ID: 300504459

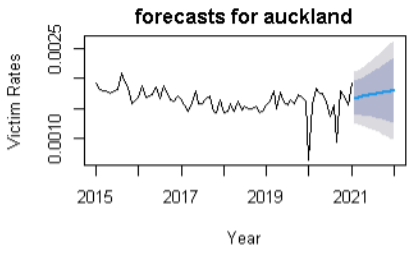
Chantal Blaikie-Salkeld - Student ID: 300443745

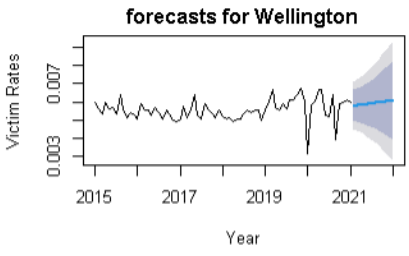
Christopher Visser-Fee - Student ID: 300389996

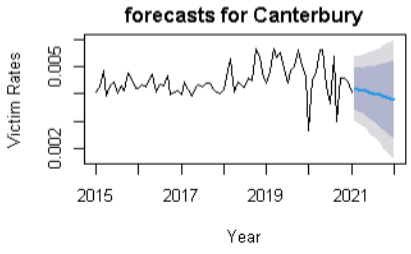
Date of Submission: 11/10/2021

Executive Summary

The research goals of this project aim to predict the victimisation rate in 2020 and 2021 using the three most populated regions in New Zealand from 2015-2020. Predicting the year 2020 is specifically used to test the validity of our forecasting model, in this case Holt-Winters was used, while predicting year 2021 is to see how well the model predicts towards the future. Methods for predicting this rate across three regions include calculating the prediction intervals for each specified region. Prediction intervals specify the uncertainty in forecasting within a specified probability. It is not possible to exclude this uncertainty if we were to find the accuracy of the forecast. Other methods also include testing the validity of the Holt Winters statistical model to assess the overestimating/underestimating of our standard errors. Also note that patterns in future will be the same as the previous periods, so if the circumstances change then the predictions will not be valid. Furthermore, an example finding below suggests that forecasts for Auckland are at an upwards trend, signifying that victimisation increases in this 2021 forecast. The 80% prediction intervals are shown as the closer spread darker area, and the 95% prediction intervals are shown as the wider lighter color.







Background

For background purposes, victimization is an extremely valuable topic of information that can help influence and assist in the government and everyday citizen struggles. Victimization is the unjust perception of being treated poorly such as sexual assault, robbery, etc.

In 2020, the Covid-19 virus emerged. New Zealand reported its first case on 28th February 2020 (RNZ,2020). On the 21st March 2020, the government introduced a 4-tiered Alert Level system to help contain Covid-19. This alert system was created by the government and was designed for public health and social measures to be taken in the fight against COVID-19 (Alert,2020). New Zealand moved to Alert Level 3 on 23rd March for 48 hours (about 2 days) which was then followed by Alert Level 4. The country remained at Alert Level 4 until May 11th (Covid19,2020). The epidemic has caused a significant impact on economies, social productivity, people's daily lives and more. We are curious about how reported victimization has changed during the lockdown period in New Zealand, and to investigate this change a thorough exploratory data analysis has been carried out.

The aim of the analysis is to focus on predicting the victimization rate of 2020 and 2021, using the previous years of data. This analysis was primarily focused by using the three most populated regions. These populated regions are vital and have significant role in the prediction process as they are populated cities/places that are most likely to be heavily affected by the outbreak in New Zealand. High concentrations of close contact with another person.

Data Description

The dataset that data group 8 has been working on was obtained from the police database supplied by the NZ police. In recent years it is evident that Covid-19 has had a detrimental effect on not only people, but has also influenced individuals' actions. This has led to drastic fluctuations in the number of victims per incident and the rates of victimization across regions. Therefore, it is interesting to analyze trends of these specific topics to better understand how Covid changed them over time.

The data used for the project was attained from the unique demographics page of the New Zealand police website (Police, 2021). Population data was retrieved from stats NZ, respective to each region (Stats, 2021). The data source that the police obtained data from was through the Recorded Crime Victims Statistics (RCVS) and the Recorded Crime Offenders Statistics (RCOS) collections (Data,2021). RCVS is about victims of crime, while RCOS is victims of crime but also includes crimes that do not have a clear victim. These specific data sources are regularly updated since 2014 and new RCVS and RCOS data are released at the end of every

month, with crime statistics previously released twice-yearly. (Data,2021). User manuals are also available that go more in-depth about the statistics behind the sources.

Victimization is an extremely important numerical variable, this is because the variable is an indication of not only the type of victims, but also the number of people affected by that one particular incident. It can be further grouped up into totals and divided by the population to be turned into rates for that specific region/area. Another main variable includes the Police District which is important as it us narrow down the regions we want to focus on like Wellington and Auckland.

The exploratory data analysis section below uses data from the unique Victims (demographics) (Police, 2021). The section shows how the number of victimizations has changed between 2015 and 2020.

To compare the result for the number of victimizations throughout the years, files pertaining to victimization demographics across 5 years were used. The data consists of dates and a high number of categorical variables e.g., type, district with few numerical variables included. Since there were a high percentage of categorical columns, this section focused on the more in depth-interpretation of said months during lockdown. The dataset was further narrowed down to the three most populated regions in New Zealand were primarily focused upon, as these were the regions that were most likely to spread faster if the outbreak occurred.

Preparation of the data was necessary for the time series to function. There was the creation of specific columns to match the time series more appropriately. A new month column that displays the full name of the month, the victim rate, the year and lastly the month number. This was so the time series function could interpret and read the data effectively.

To make sure our rates were relevant and updated accordingly. A new variable called population was stored for each individual year for each region. This was so it was accurate and respective to that point in the timeline. There were also date variables that had to be renamed and revolved around the period at which individuals were victimized, since the names of the months were only the first 3 letters of each month. These were fixed to have full names.

Missing data was completely removed from the dataset beforehand, these were two specific columns called variance and variance test. Typically, they consisted of a huge percentage of missing data. It was not an important column as they were just the discrepancies between individual victimisations, they were being grouped together for totals anyway so it would not show any relevance thus needing to be removed.

Ethics, Privacy and Security

Ethics

Ethical concerns talk about how people allow the usage of their data if the data usage being used is unjust or unfair against another person. In a sense of like a system wrongly identifying a specific individual based on their ethnicity. In this case the most concerning issue is the AI in the police database determining what the most appropriate methods to use are. We do not want a repeat of predicted Policing bias. Predictive policing is the situation of crime following regular patterns, this can lead to misdirection such as focusing efforts of police towards areas that are most affected by crime. The bias part was that the system was based on location rather than the individual. This specific type of bias is prevalent throughout police technology, where there is occasionally a misjudgment in a certain case.

The police database does contain demographic data such as ethnicity and sex, and so we do need to be careful to avoid presenting an exploration of the data that suggests some intrinsic differences between ethnicities when it comes to crime rates. However, we have chosen to mostly focus on exploring victimization data, rather than offender data. We are not looking at who commits crimes, but rather who is affected by them. Being the victim of a crime does not have the same moral implications as being an offender, in the eyes of most of New Zealand society. However, there are still potentially negative connotations to presenting a specific demographic as a consistent representation of the victims. We would still need to be careful in how we presented our data to avoid presenting potentially harmful messages. It will suffice to tell a story that has accurate and correct facts, there just needs to be a key reminder for others to not extract irrelevant type of information that would lead to irrelevant and inconclusive conclusions.

Our goal is to use this opportunity to study how lockdowns and quarantines, and to an extent, pandemics themselves, affect us on a national scale. To focus on crime specifically, by researching the relation between Covid-19, the lockdowns and crime, we can potentially help police forces and governing bodies plan ahead and prepare for future pandemics and quarantines. If our research can help police predict and better respond to crime during lockdowns, we should be able to better society for New Zealanders as a whole

Privacy

Privacy concerns explain how each victimization's data has been collected and whether their personal information has been hidden safely. Original victimization data is collected by each victim who dialed 111 or 105. Personal details like their name, phone number, etc. are easily leakable and have to be kept private and confidential. Privacy concerns include the right to not answer intrusive questions, to needlessly provide personal information or the right to have your information kept securely. Because there are categories of "not stated" in the data, we can interpret that not all questions were required to have an answer. This respects individuals' rights to not have to answer intrusive questions and to provide personal information, if they do not feel comfortable in doing so or if it is not relevant.

In terms of privacy considerations, as they apply to our own project, the data we have been working with has been compiled with privacy concerns already in mind, as it has been made publicly available. Entries in the database include as little personal information as possible, excluding names, phone numbers, addresses, specific date of contact or any information that could link an entry directly back to an individual.

Furthermore, data is not taken on individual cases of victimisation, instead being collated into numbers of victimisations over a period of a month, grouped by each of the categories present in the dataset. This also helps mitigate the reverse engineering of an individual identity, although it is notable that for a lot of entries, only a single victimisation is present. These entries do, in practice, refer to only a single account of victimisation, which increases the risk of an individual being identifiable

The exploration of the data we have done has grouped the data even more than the police data does, however in our EDA at least we have generally looked at the number of victimisations countrywide by month, which usually number in the hundreds at minimum. Even more so than the police data itself, the data we are presenting is extremely unlikely to be used to identify people in the database.

Security

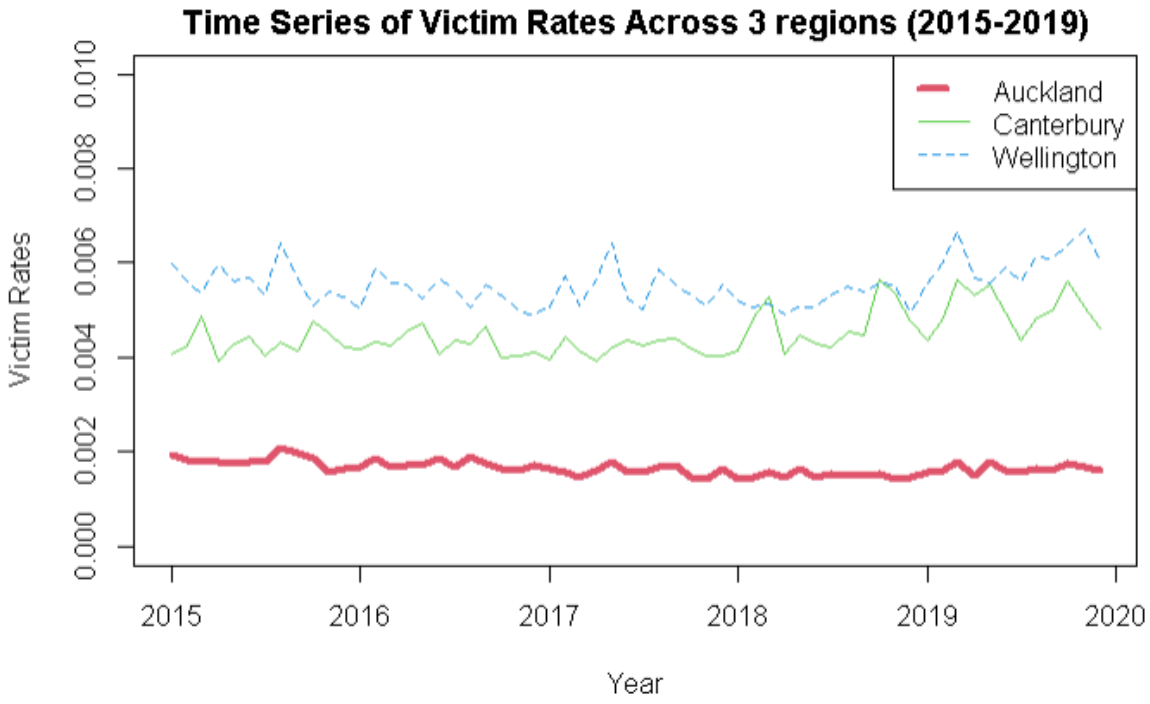
Security in regards to data is the practice of keeping digital data protected and only accessible to authorized agents. It is important to keep this data impenetrable so that it does not get into the wrong hands. If the victims from the police dataset were to be identified by this data being leaked or exploited would mean that their right of security has been compromised.

Security also refers to steps to keep our project data secure. Our project team stored our important data extractions of information on google sheets, where the only way to view the data was through email links through our group chat so that only our group and the host Google would know of this sheet's existence. Note that this is like privacy, in a sense that data is easily leakable if there are cracks in the system. Theoretically a weekly check on the system by human interaction could avoid bias. This makes sure that data integrity is valid. Also keeping our data secure on a cloud or police database. A place that is heavily secured with a system.

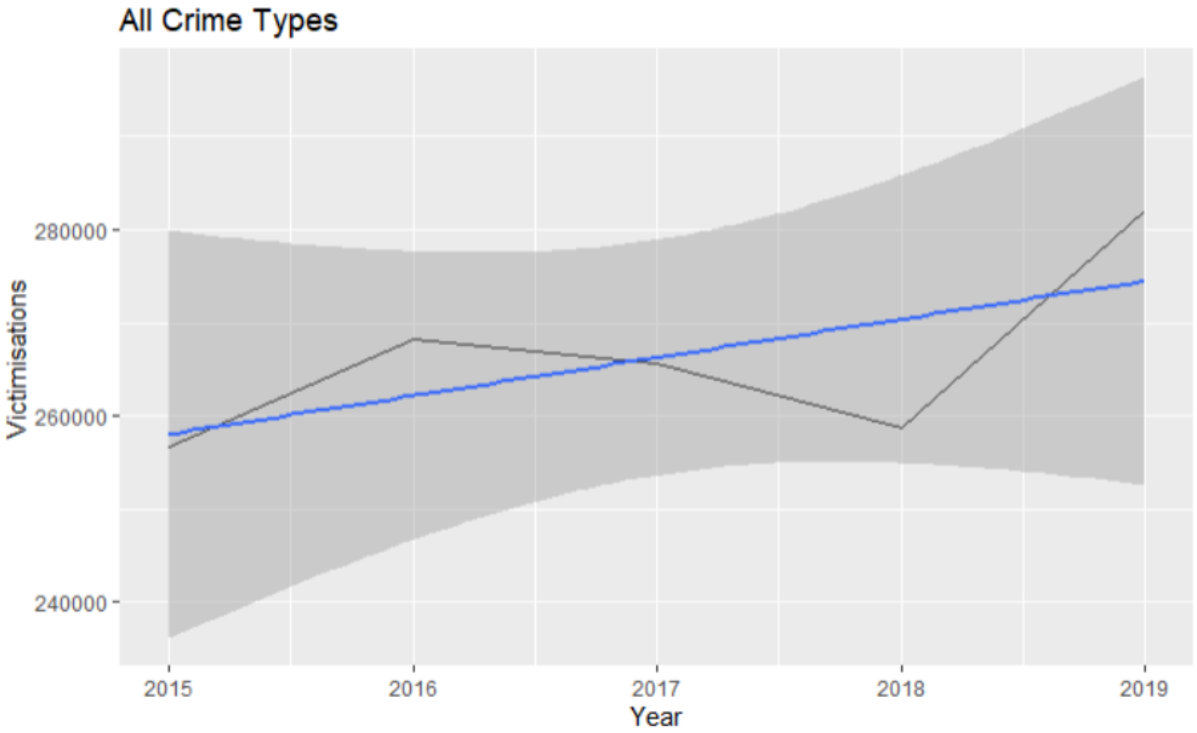
Exploratory Data Analysis

Time series is an important aspect in analysis as it is an excellent way to show underlying trends over certain periods that you would not normally see on a table. To analyze and show significant patterns and irregularities. Using the data from the previous year's 2015-2019, the units displayed below were the victim rates per population.

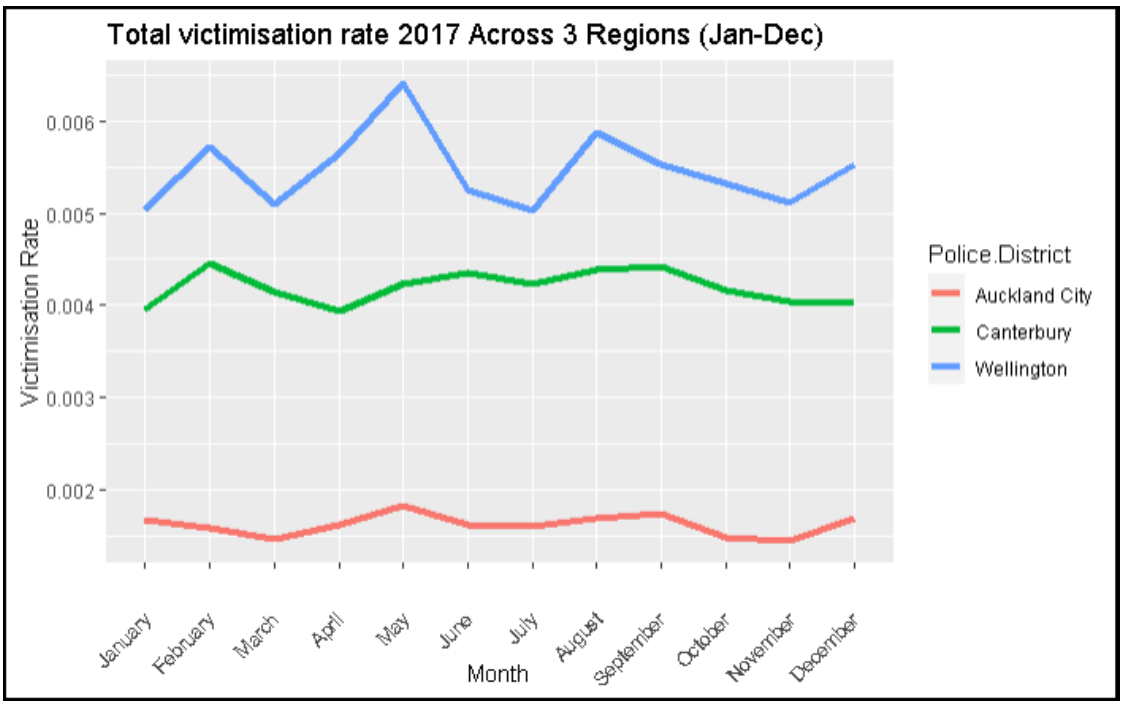
In this graph below, there is a clear indication of seasonality. Seasonality is the regular repeat of a cycle. In this case we can see that there is a trend of increasing victimisation rates every 2-3 years.



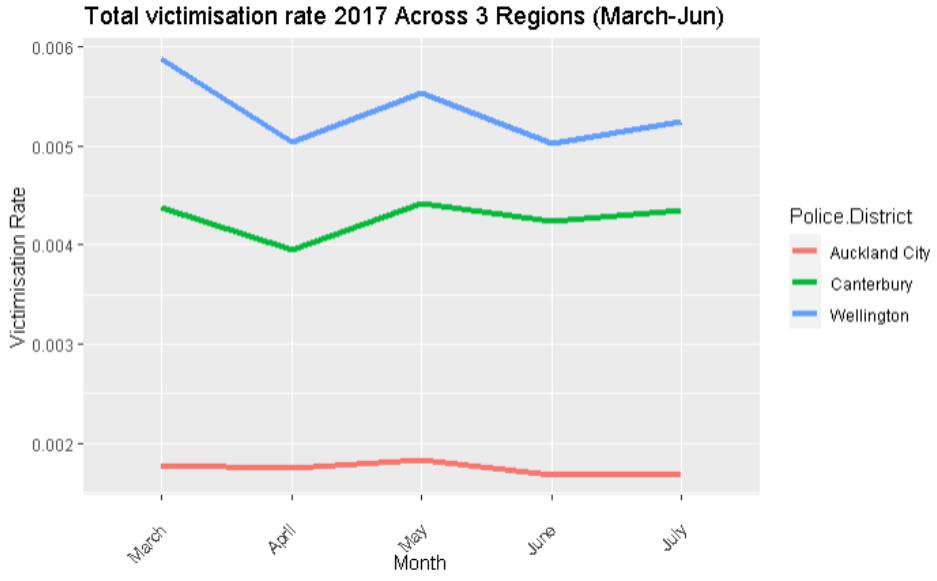
The main variables involved were the victimisation, date of victimisation, year of victimisation and region columns. This plot involved acquiring the total victimisations throughout the years 2015 to 2019. As shown below, there is an evident trend that across the years the number of victimisations grew overall as time increased. The global maximum was during 2019 and the local maximum number of victimisations occurred around 2016. While the global minimum number of victimisations was 2015, and the local minimum amount was in 2018. A line was placed to identify the general trend and direction of the graph overtime, its overall going upwards positively.



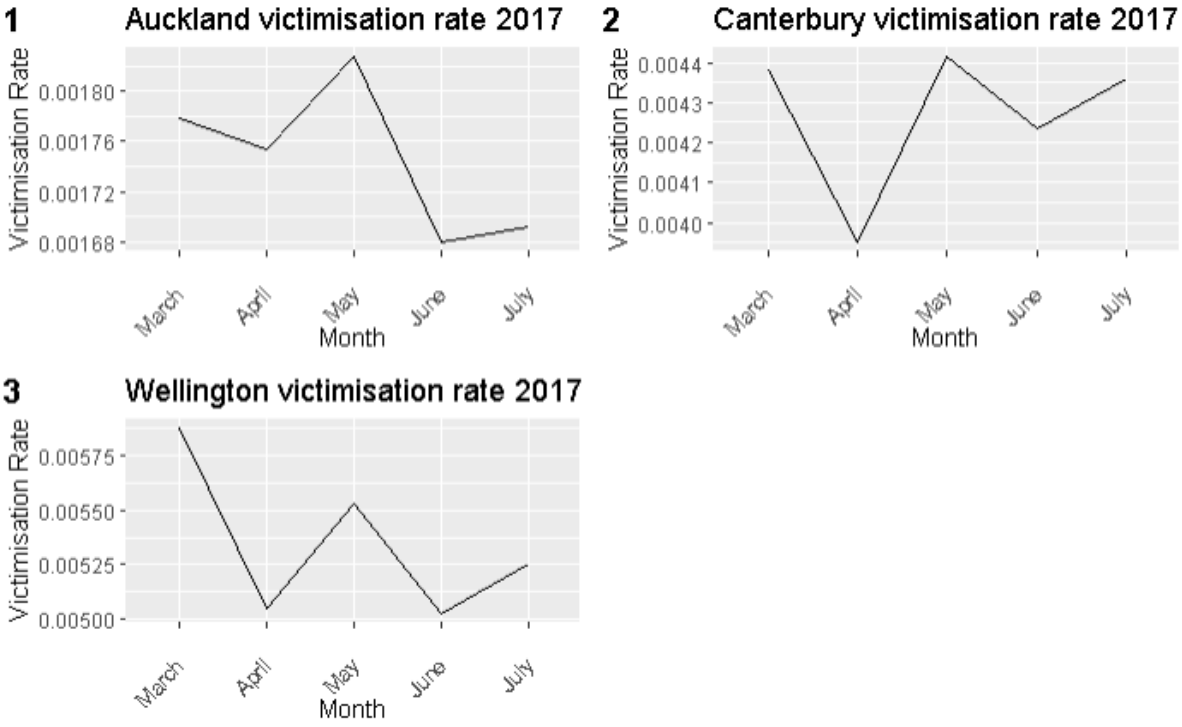
In reality it is not ideal to look at only the total number of victimisations per year as its meaningless. So, most of the plots were done in terms of rates such as below respective of each regions population and year. Rates are important as they represent the per head of population. 2017 was chosen and displayed because most of the years were basically similar pattern wise, and showed that victimisation was clearly at its highest in Wellington and peaked in months April to May. One result that was certain was that Auckland had the lowest victimisation rate, despite having the highest population count in New Zealand. Across all years there was a general trend of Wellington having the highest rates of victimisation, whilst Auckland having the lowest. There were only a few interesting/abnormal occurrences. During months October in 2018, both the victimisation rates of Canterbury and Wellington were similar. This may tell us that Canterbury and Wellington are similar in terms of how often an individual gets victimized even though the populations are completely different. 2019 dataset showed promising results, where Canterbury and Wellington had similar rates during May.



Next part of the analysis revolved around subsetting to only a few specific months, typically because this is the lockdown period. The year 2017 was shown to see if any victimisations in the three most populated regions grew within these specified national lockdown months. 2017 was specifically chosen out of interest as it is roughly the midpoint during in terms of number of total victimisations throughout years 2015-2019. So, this was a subset used for EDA purposes only. The Lockdown period typically started around march and lasted until June to July. As shown below it is still the same compared with the above plots.



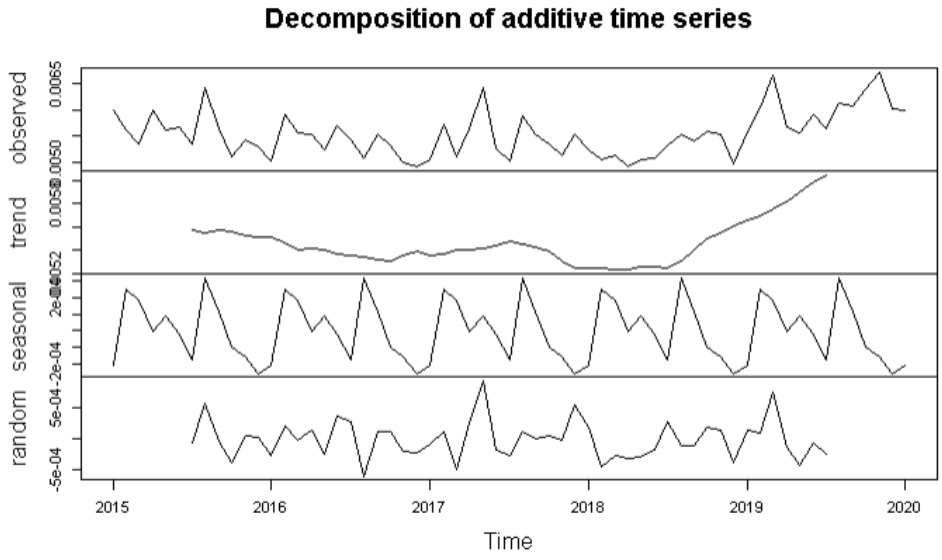
Lastly it is important to see the three most populated regions separately, to see if there is any special pattern found, therefore it is a great way to get an overview of the rates in those specific areas. There is an interesting find during these specific months, Auckland and Wellington had a lower trend for victimisation rate, while Canterbury was generally high most of the time. This may play an important part in the output later on, seeing that Canterbury is the outlier region.



Detailed Analysis Results

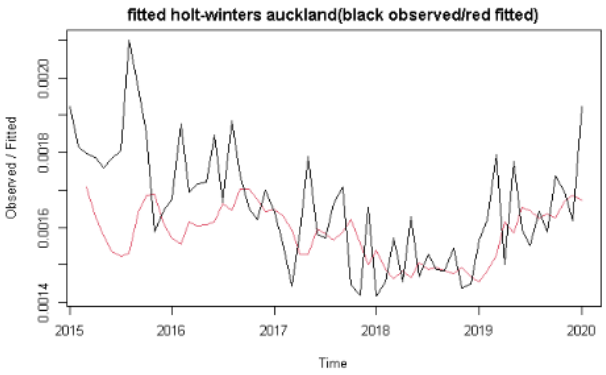
To understand these rates more clearly a time series forecast analysis needs to be done. The time series above is an additive model that has an increasing or decreasing trend with seasonality, therefore Holt winters exponential smoothing is useful to make short-term forecasts.

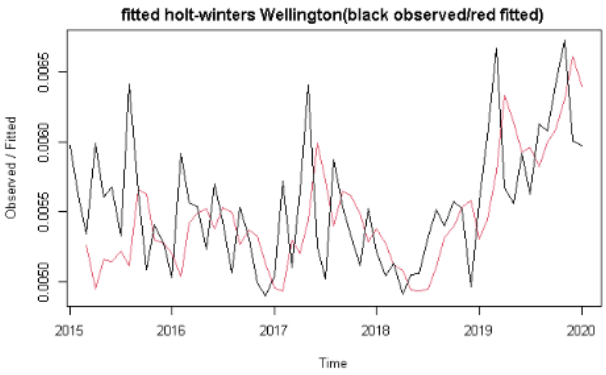
Decomposing time series is an important aspect as it more clearly shows the various and not so clear patterns that can help us solve the problem more easily. The time series is broken down into 4 estimates, the observed, trend, seasonal, and irregular components of a time series. The graph below is a representation of all three regions as they were similar in general. Wellington and canterbury had the same trend display, with Auckland being the opposite trend-wise and decreasing instead of increasing. It is clear to see that troughs and peaks in seasonality occur ever 2-3 years again.

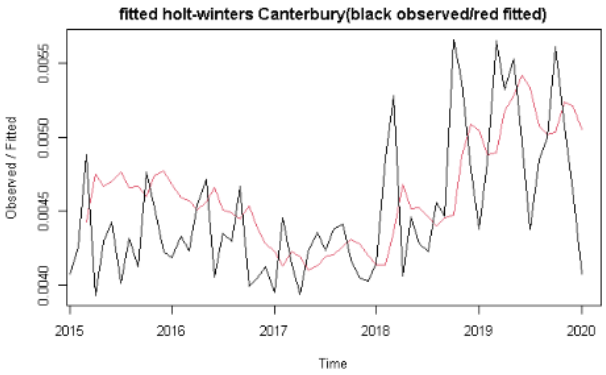


This next part involves fitting data from 2015-2019 for each region. Fitting up to the end of 2019 was done because we want to check the validity of the model.

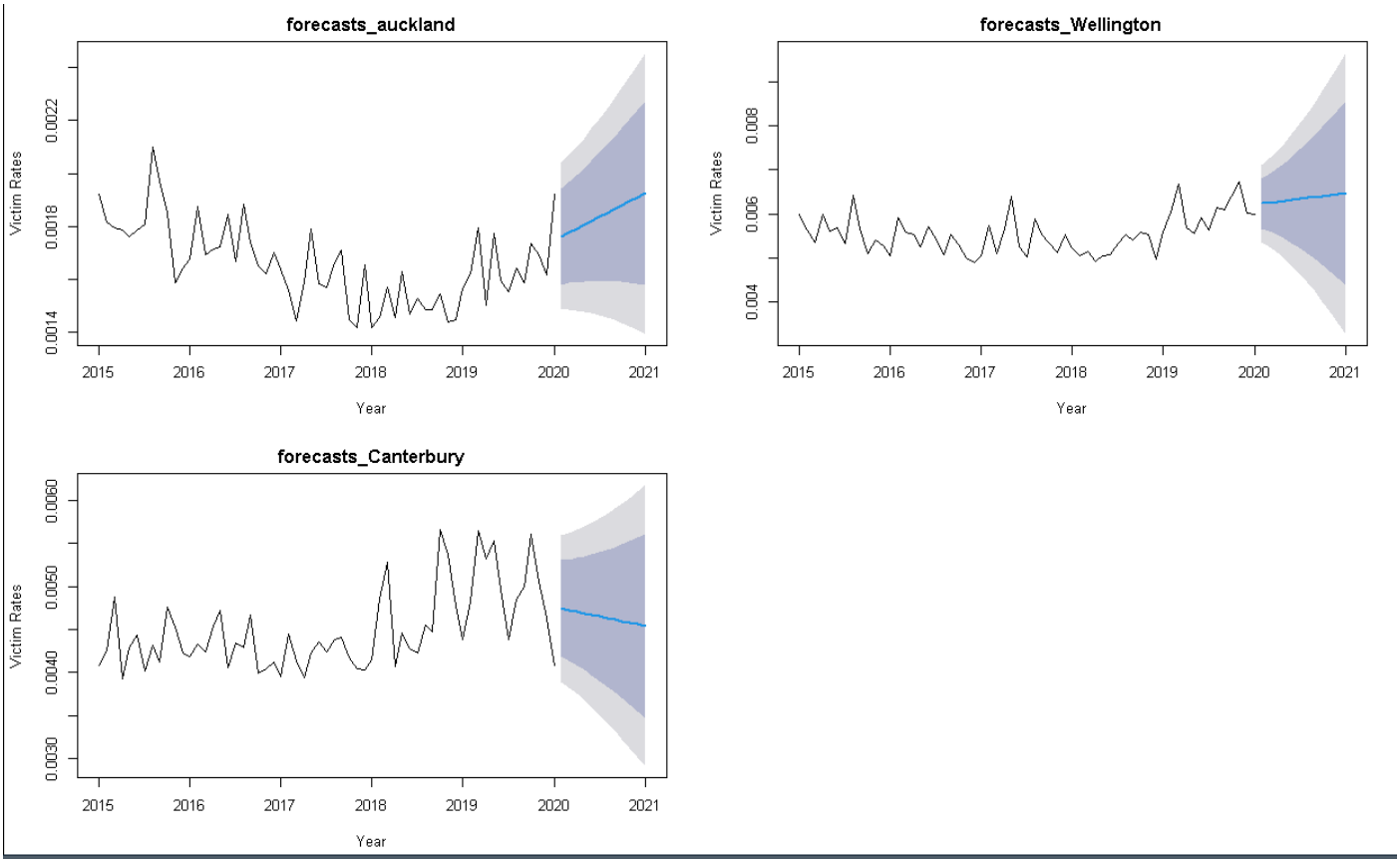
The results below show that the graph is not as accurate in comparing the observed black line data to the red fitted by holt winters. In general, it seems that the fitted model is not amazing.





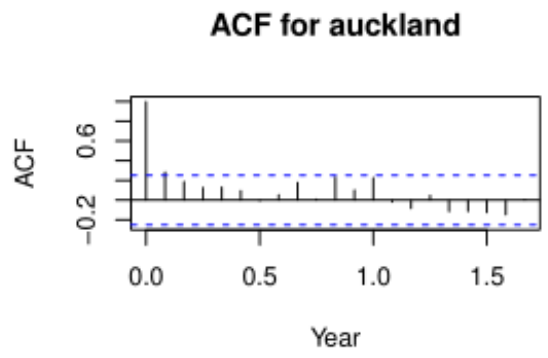


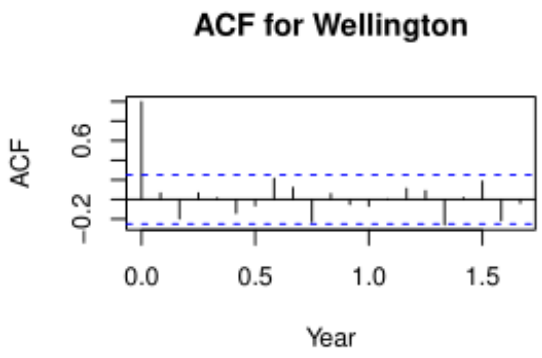
This next section is to see the forecasted general trend, of the holt winters forecast for 2020. Prediction intervals were shown below using Holt Winters. The trend of the forecast shows that there will be a rise in the victimisation rates for Auckland and Wellington, but for Canterbury it decreases. Wellington's interval spread was the largest compared with the rest of the regions, while Auckland and Canterbury were much smaller.

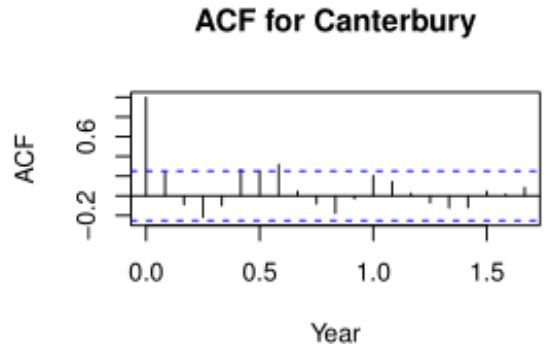


To further check if the validity of the model, an auto correlation function (correlogram) test would need to be carried out. ACF test is checking how the present value of the series is related with its past values. An analysis would need to be done on the in-sample forecast errors that show non-zero autocorrelations at lags 1-20.

Shown below its expected to see 1 in 20 lags being out of bounds for all three. What is abnormal is canterbury having a lag outside the boundary near the halfway point. Although in general all ACF values lie within the bounds.





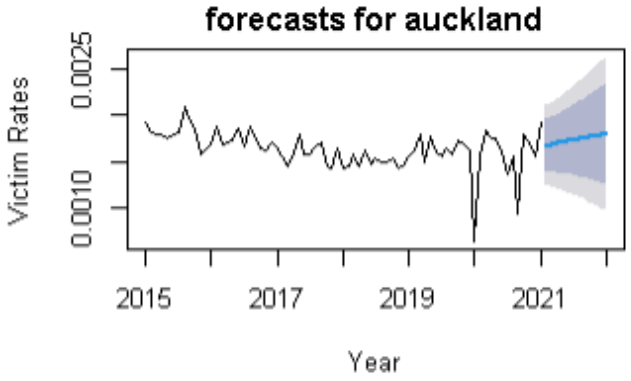


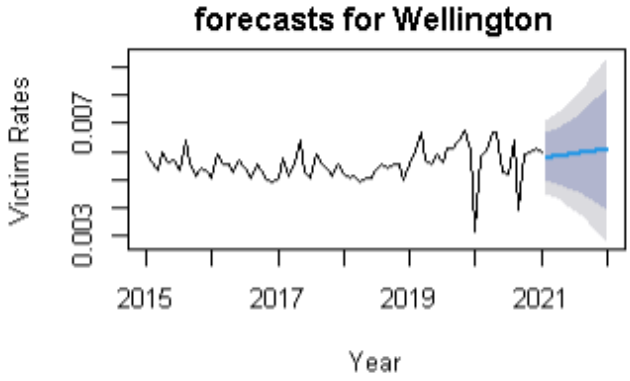
To further get a grasp if the model forecast were a good fit for each region, a l-Jung box test would be needed to be carried out. The Ljung-Box test is statistical test to check if any autocorrelations are different from zero. Results show that the p-value is greater than the null hypothesis for Wellington (0.09869) and Auckland (0.1089) meaning that we do not have evidence to reject null hypothesis and conclude that the model shows some evidence of non-zero autocorrelations at lags 1-20. Since autocorrelations are not different to zero, the model is therefore a good fit.

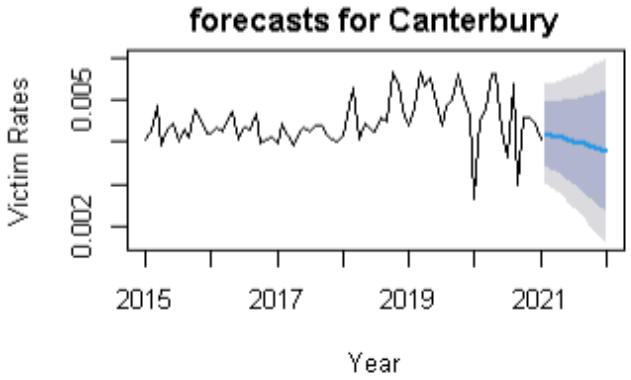
The only time the model was not a good fit was for the Canterbury data, with a p-values of 0.014 which is lesser than the significance level, therefore we have enough evidence to reject the null hypothesis. Meaning that the model was a poor fit as auto correlations were different from zero. The standard error for Canterbury would be possibly overestimated due to the test showing dependence, although it is definitely fine as the ACF plot for canterbury was weak in terms of positivity.

The Holt-winters model is a simple model with a weak ACF plot signifying that there is not much concern as no ACF values are extremely positive or out of bounds in the plot. Therefore, this validates that the Holt winters model is overall ok and no changes need to be made. In terms of uncertainty for all regions, even if the standard of error is larger due to independent uncertainty, all plots for ACF were weak which means that the model is still valid.

This last section below is to check how well the model does towards predicting between 2021-2022 given data from 2015-2020. This is a short section where evidence below suggests that the trends for each region are on the right track and are like the forecasts for 2020, where Wellington and Auckland had an increasing trend and Canterbury was decreasing. Carrying out a l-Jung box test, again the p-values for Wellington and Auckland are both above the significance level, whilst Canterbury is still below the significance level. The ACF plot was also similarly weak again with no extreme positive values showing us that Holt winters model is overall ok to use.







Conclusions and Recommendations

In conclusion, time and the amount of data plays a crucial factor into how we can interpret results. Wellington and Auckland had decent model fits which resulted in a good prediction for the forecast. Although for Canterbury, the ACF had out of bounds and rejected the null hypothesis making it a poor fit. Overall, as stated previously the ACF plots for all were weak, meaning that the were no changes needed to be made for the model. 2020 predictions gave us a reassurance that the model and data were a good fit. Whilst the 2021 forecasts showed us that the holt winters model did a decent job at predicting the future.

Although biases might arise in the data or results. In this case if a region ends up sending more police units towards an area with more crime, it would be bias if they did not account for previous years beforehand. Only dealing with data that is readily available. For example, only 5 years in the past does not represent all crime, thus it would not make sense to send units based on limited years of data and regions.

Recommendations include, making sure there is enough data to be able to forecast more accurately, as of now there is only data from 2014 to around 2020 in the police database. More years would help the forecasting. Especially for the Canterbury region due to its poor fit. Limitations would be that this is only respective of data within those regions. Typically, if we were to expand and get more accurate fits then predicting using all regions would solve this. Future work implementations would include predicting towards 2022 as that is an important year where uncertainty is extremely high in both a statistical and non-statistical sense during the covid era.

Reference:

News, R. N. Z. (2021, February 28). Timeline: The year of COVID-19 in New Zealand RNZ

https://www.rnz.co.nz/news/national/437359/timeline-the-year-of-Covid-19-in-new-zealand

Unique victims (demographics). New Zealand Police. (n.d.).

https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz

History of the COVID-19 alert system. 19. (n.d.).

https://covid19.govt.nz/alert-levels-and-updates/history-of-the-Covid-19-alert-system/

Alert Levels. Covid-19 Government. (n.d.).

https://covid19.govt.nz/alert-levels-and-updates/about-the-alert-system/

Data Source. New Zealand Police. (n.d.).

https://www.police.govt.nz/about-us/publication/data-and-statistics-user-guides

Population. Stats Nz. (n.d.).

https://www.stats.govt.nz/indicators/population-of-nz
gclid=CjwKCAjwtfqKBhBoEiwAZuesiDS13d3bWHCyZQzwahFfMRRKtbjNJIgxF2MACBe9olfTpyEF4NdPHhoC4kwQAvD_BwE