Is Theft Common in New Zealand ?

Hariz Aqil Abd Ghani (47725827)

Assignment 1, DATA401, 2022

Introduction

Is theft common in New Zealand ? According to the Global Peace Index, New Zealand is ranked as the **second safest country** worldwide in 2021 and has been top five since 2008. However, theft always been the most commonly crime recorded and keep increasing every year.

This project will interpret the data from theft cases in New Zealand. The purpose of this study is to view the cases of theft by age-group, ethnicity and demographic. These data will help to plot and summarize the information recorded.

Method and Analysis

1. Load and clean the dataset.

Firstly, I load csv files that are needed as our dataset. In this assignment, I used two csv files as the Police NZ website limits the amount of variable that we can download. In **'data_1'** file, we used 'Year', 'Victimisations', 'Sex', 'Ethnicity' and 'Age Group' variables meanwhile we used 'Police District' in **'data_saya'**.

Before downloading, both dataset were observed to make sure that we downloaded same crime case within same time frame, in this case. In this way, we will avoid of downloading and uploading different dataset.

In the r chunks, we included **head()** function and **str()** function to view the variables that we want to use in cleaning and plotting step.

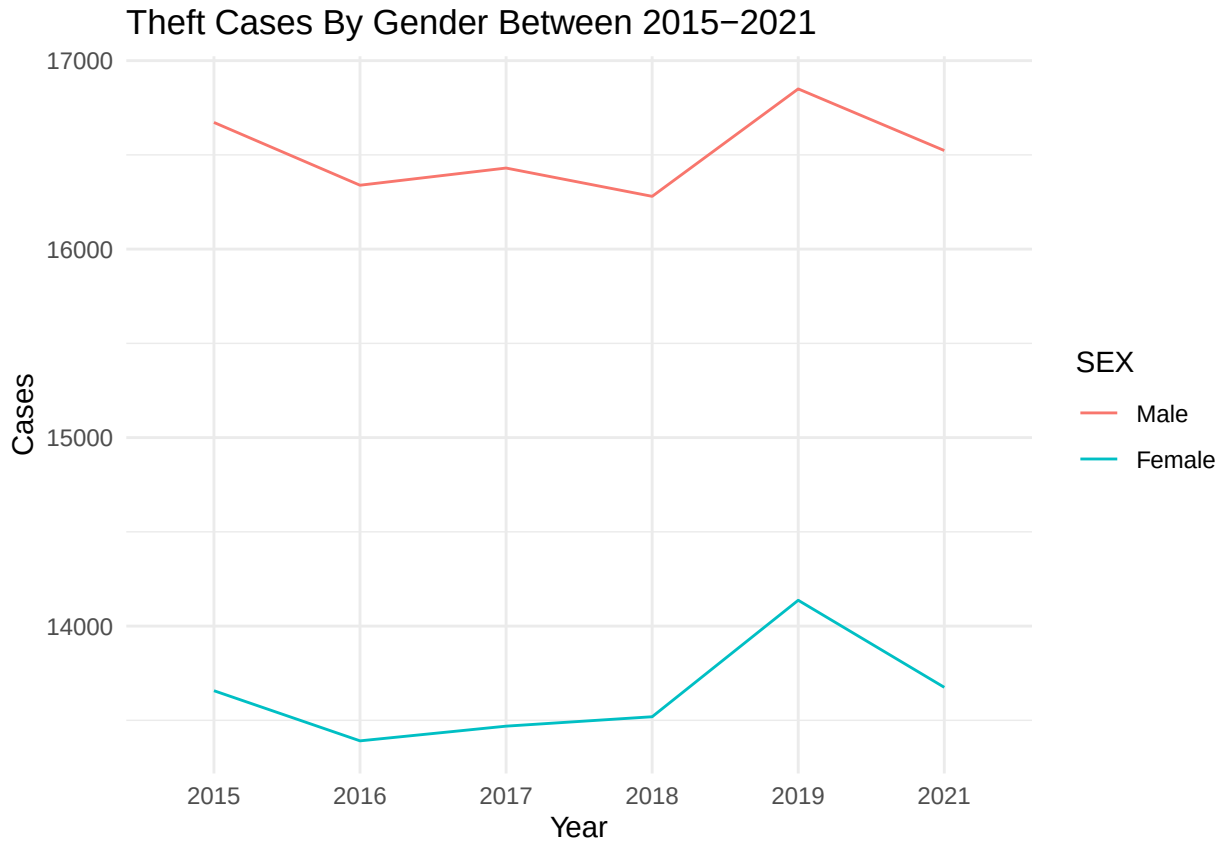
2. Remove any 'Non-Appplicable'(NA) numerical or categorical data in data frame.

With **unique()** function, we are able to view all type of observation in the each of the variables. Then, using **'as.factor'** function to change into factor and **'levels'** function to set the level. For examples, we sort the level of age group variable from lowest age range (youngest) to highest age range (oldest). However, in the case of gender where we used **'Sex'** variable between *Male* and *Female*, there are no gender better than another. We can used another way of cleaning and removing unwanted data in dataset by using *Tidyverse* packages. Due to my lack of knowledge of using tidyverse packages, I proceed by using **'as.factor'** and **'levels'** functions.

3. Load *ggplot2* and *tidyverse*, then plot the graph.

With the dataset that we uploaded in R, we plot and categorized into several graphs.

(a) Year vs Victimisations

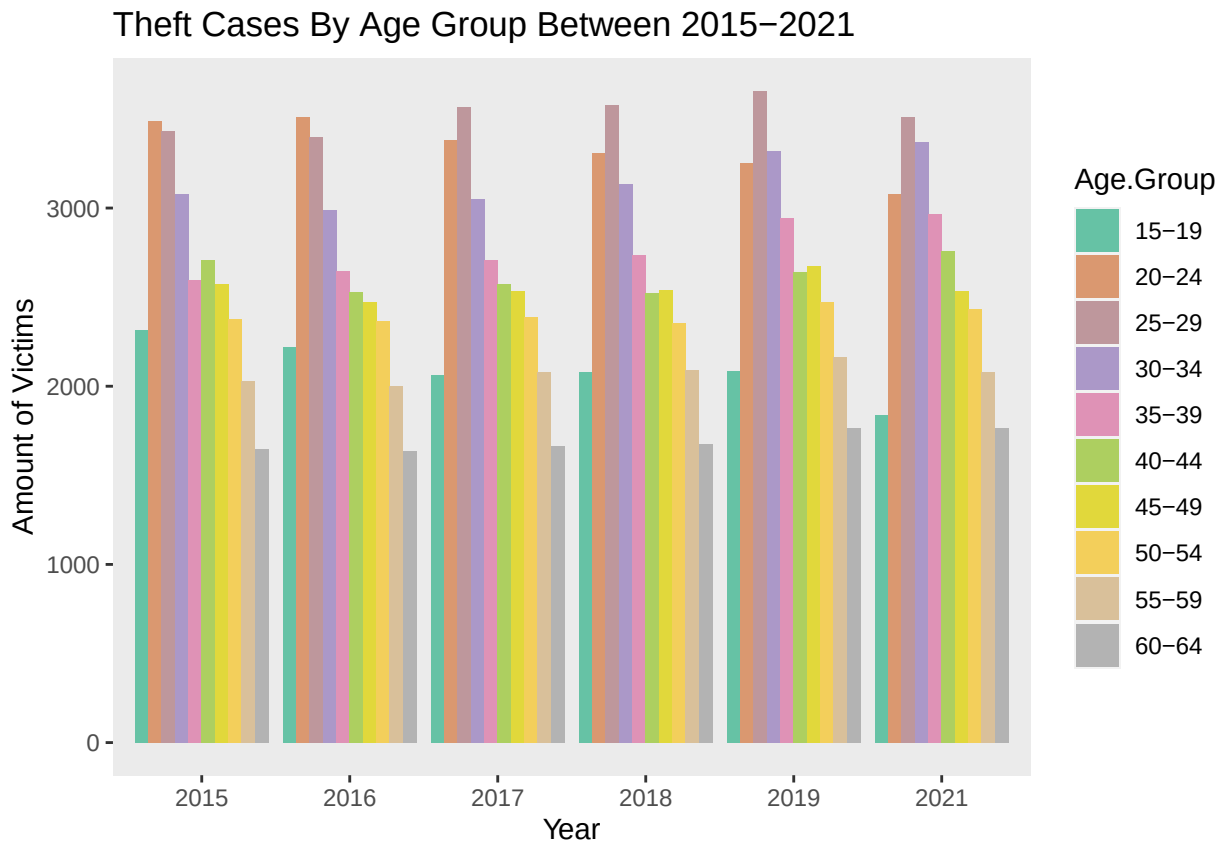


Bases on the above graphic, we decided to use line graph with the group of sex variable to show the differences between two lines in one graph. We filtered two years, which are 2014 and 2022. Why ? It is because these two years of data were downloaded with insufficient amount of months to make it a year. For examples, year 2014 contains only between June until December. By removing two years, we managed to reduce bias in handling the data. *The year of 2020 are not included in our downloaded dataset as that was the year of pandemic of Covid-19, as New Zealand undergoes months of lockdown.*

Based on analysis on line graph above, *theft* cases have huge differences between *male* and *female*. Starting in the year of 2015, differences between male and female is more than 2,900 cases (~16,600 - ~13,700). We can view whole line in the graph from 2015 to 2022 between male and female **does not intercept**, which means majority of theft cases across the year were conducted by male and the amount of cases were never the same as female.

An interesting observation is from 2018 to 2021, the trend and slope of graph shows almost the same although the numbers are different. However, it is slightly different between 2016 and 2018. To illustrate my point, we can see in the graph in year 2016, trend of male cases increase sightly and decrease in the following year. Meanwhile, female cases show that the trend continuously increase from 2016 to 2018, then spiking from 2018 to 2019.

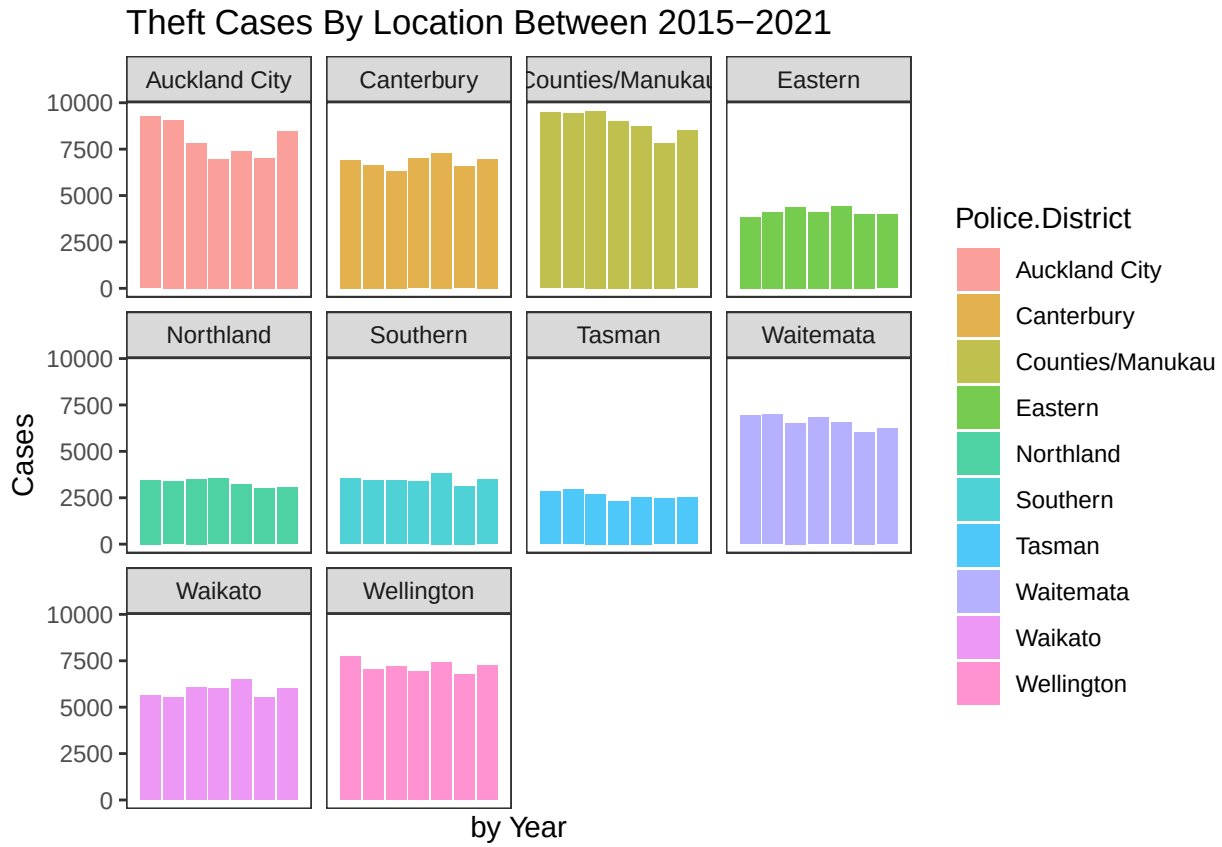
(b) Age Group vs Year



We take different **method** in this chart. Bar chart were used to shows more differences between age group. Although **'position = fill'** can be use to present the proportions, it may confused the reader as it is very hard to read because we have huge group numbers, which are 10 age groups. Also, *'RColorBrewer'* packages were loaded because default color palette cannot interpret data that have more than 8 or 9 groups in one graph. In R, we filtered unwanted age group which below than 15 years old and above 64 years old.

In the bar chart above, 15-19 age group declining consistently compared to other age groups. In the year 2015 and 2016, this graph shows 20-24 age group committed the highest number theft crime. Meanwhile from 2017 to 2021, 25-29 age group consistently showing highest amount of committed crime although there is decreasing trend in 2021. If you look at 60-64 age group, we can see that it is does not shows any rapid changes either increase or decrease in trend, make it seems like it is uniform distribution. Let's turn to 30-34 age group, which showing *S-curve* trend. The graph started high then decreasing the following year and started to grow from 2017 to 2020, later climb slightly in 2021.

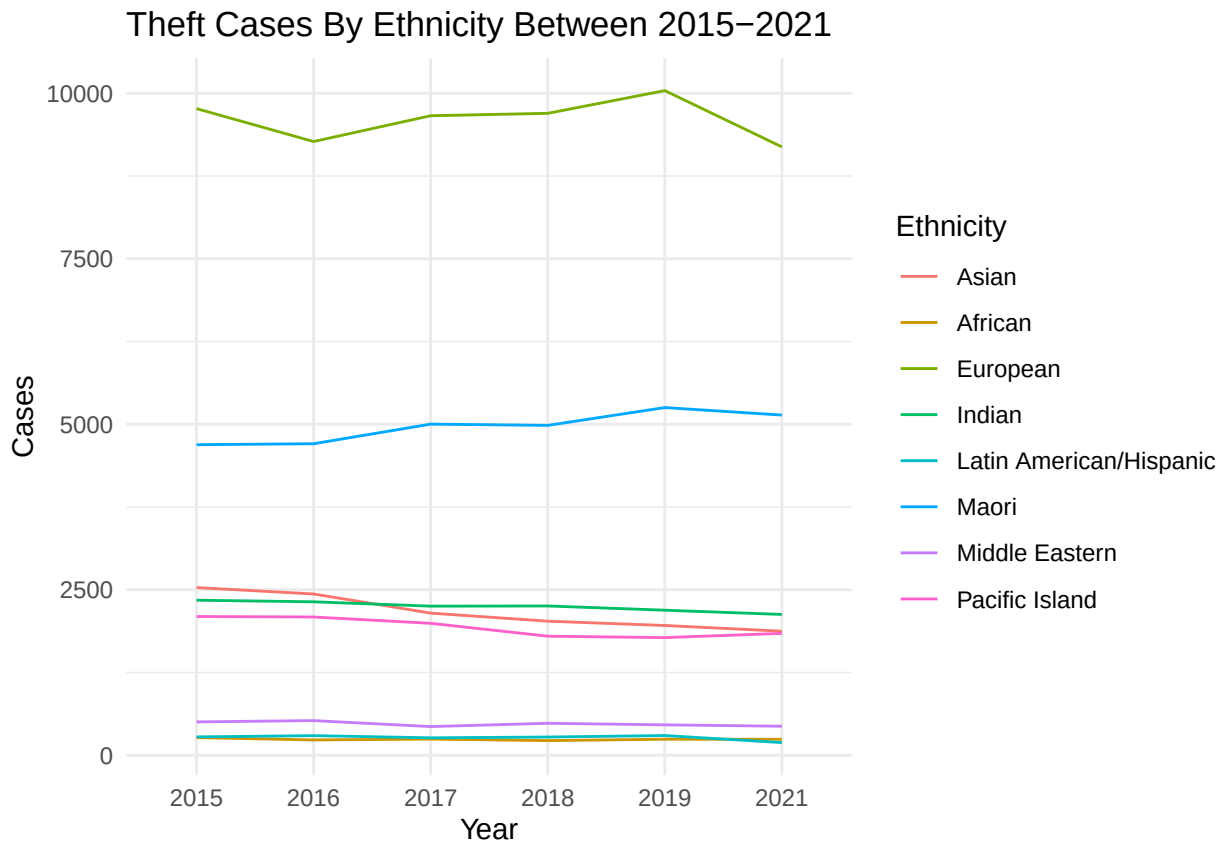
(c) Location vs Year



Based on the data that we have, we know that this graph contains large number of group in one graph as previous graph. This time, we used different approach by using **'facet_wrap()'** function to split the bar graph into several tables. This way we can interpret the data individually with more neat and less messy. We used Police District variable to separate the location to district because it is easier rather than using city variable, which rural areas will be excluded.

As you can see from above, we can interpret that high density of population in the demographic has higher amount of crime committed compared to other locations such as Auckland City, Counties/Manukau and Wellington compared to Northland, Tasman and Southern. High population area shows almost the same trend which decreasing after 2015 then increase after 2019 or 2020. In the other hand, location such as Eastern, Canterbury, Northland, Southern, Tasman, Waitemata and Waikato shows inconsistency in the graph trend. Also, Canterbury shows the highest number of theft cases recorded among South Island as Canterbury has highest population compared to other districts.

(d) Ethnicity vs Year



In our last graph, we included *ethnicity* variable to compared. Previously, we used bar graph with *'position = stack'*. It is less messy, however it is harder to read and does not make any sense to interpret. Due to our large number of groups, we use line graph to show more differences between the ethnicity. With the same function and code in *r-chunks*, we plotted the graph.

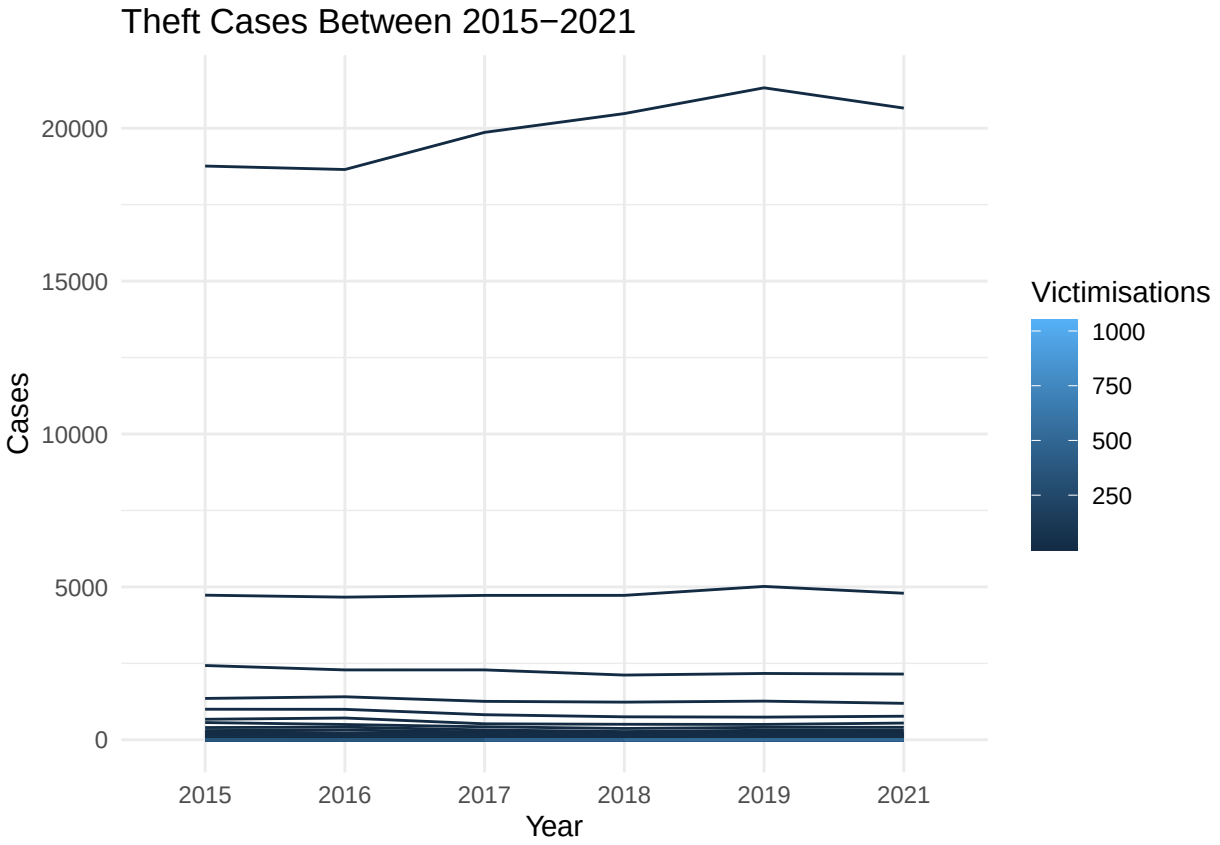
This is interesting if we look at the graph, major differences that we can see is *European* and *Māori* compared to other ethnicity. *European* indicates to Kiwi which also indicates to Australian, British, American and Europe European. That explained why *European* has big different of line graph compared to *Māori*.

But why these two shows higher number than other ?

We can look at different perspective, both *Māori* and *European (Kiwi)* are both local giving us indicator that they have higher proportion of population compared to other ethnicity. For *Māori*, the trend is rise up along the year but later decline slightly after 2019, meanwhile for *European*, the trend falls after 2015 but increases to 2019 before falling down rapidly on 2021.

If we take a look of other ethnicity such as *Asian*, the trend decreasing from 2015 to 2021. In the other hand, the other 5 ethnicity shows inconsistency of trend which almost like no changes, as it is like a straight line with zero degree of slope.

Conclusion



We can conclude the relationship between number of theft cases committed are closely related to age group, location and ethnicity. Each of this variable will give different results.

On above graph, we can make a conclusion that theft cases rises every year since and theft cases is common in New Zealand. By looking at current data, we can predict that year 2022 might show less amount of cases compared to 2021.