STATS 765 - Project Milestone 1: Where are the crimes?

Aidan Jared - ajar667, Simon Loh - sloh166, Sujay Anjankar - sanj402

Objectives

This project aims to predict the types and rates of crimes based on the features of neighbourhoods from the census data.

Data

We propose to use the following data sources:

Crimes	https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz
Census	https://2023census-statsnz.hub.arcgis.com
Meshblock	https://datafinder.stats.govt.nz/layer/111224-meshblock-higher-geographies-2023-generalised/

Using the Victimisation Time and Place data from the police, we can get a list of crimes and the area in which they occurred. This area can be linked to the Census data allowing the two primary data sources to be joined. The specific Census data we will be looking at is from the latest 2023 Census. We will be looking at Statistical Area 2 which collates people into groups of between 1000 and 4000. The following Census datasets contain the most important features about neighbourhoods: individuals, households, dwellings, Extended Families, by topic, and means of travel to work / education, and 2023 Census Population change by ethnic group. The Police data can then be combined with the census data to link the information on different neighbourhoods to the crime data, and build a linear model or Neural Network on it.

Exploratory Idea

- Are certain neighbourhood features (income, house conditions, ages, etc.) more likely to be a predictor of crime?
- Is there a spatial aspect to high-crime areas?
    - Do high crime areas increase crime in their neighbouring regions?
    - Are crimes more likely to occur away from the main predictors?
- Are different features more or less important in predicting specific types of crime?
- Are certain areas subject to police bias (racial profiling, lower income, high police station density)?

Approach

The first step will be to combine our two data sources into a single dataset by tidying and joining the data. Once the data is in an easy-to-parse format, it can be brought into R, where some Exploratory Data Analysis will be performed, and then a model can be trained. We can then use AIC, Mallows Cp, and other model evaluation metrics to get the feature importance for the model.

Challenges

- The first major challenge will be combining the two datasets from different providers, so we will need to find a way to join the data into a comprehensive database.
- Then there will likely be issues with missing, incomplete, confidential, and non-standard data, which we will work with as they appear through different techniques as appropriate.
- There may also be issues with disproportionate police data in certain regions due to bias or overreporting, which may skew the results and overemphasize certain census areas. This challenge will be difficult to fix, but undersampling the data in over-reported neighbourhoods may help reduce the impact of these reporting issues.
- Alongside bias issues, there will be imbalances in the types of crimes, with large amounts of thefts reported, but other major crimes being less common. This challenge will be difficult to overcome, but there may be some benefit in splitting up the model to tackle individual crime types instead of looking at crime as a whole.
- Temporal trends (seasons, yearly trends) may cause some issues. It might be a good idea to look at a single year or a small spread of years while separating the seasonal data into different models.