Exploring Geographic Factors that Correlate with House Prices in Canterbury

Group TBA

AUTHORS:

Daniel Peach	56155289
Jonathan Coachman	81822865
Allan Ou	31779222
Xuanxi Yin	51169786
Logan Lawson	22709657

NUMBER OF WORDS: 3886

Users can access and download every dataset (except REINZ data)

from the GitHub repository below:

https://github.com/ajcoach/DATA422-Group-Project

Table of Contents

Introduction

This project focuses on creating a relational database that could be used to explore important factors related to the Canterbury property market. The approach taken has been to integrate openly available datasets with the proprietary Real Estate Institution of New Zealand (REINZ) property dataset. Motivation for this project can be found in rapidly rising NZ house prices. House prices in NZ have risen 30% over the last year (REINZ). Closer to home, the median Canterbury house price has increased by $558,000 in the last 30 years. This is growth from $253,850 ten years earlier (McKnight, E 2021). Therefore, our group project research is to integrate a variety of data that could be used to analyse the impact of social, environmental, and economic factors on house prices; hoping to help more people select suitable houses according to their own conditions.

Data sources

We used five datasets relating to Canterbury. One dataset included a school and area dataset from the education counts government website (Ministry of Education 2021) and was downloaded as a csv file. It included school names, suburbs, decile, number of students are more. Another dataset included *Victimisation Time and Place* dataset (NZ Police 2021). This was downloaded as a Tableau Packaged Workbook, then converted into a csv file. This consists of all crimes in NZ involving victims between September 2016 to August 2021. The *Subnational population estimates* dataset was also used, it consisted of the population of every suburb in New Zealand based on SA2 (Statistical area 2) and includes age and sex suburb level population data (Stats NZ 2021). Only SA2 areas in Canterbury were chosen. This was downloaded as an excel (xlsx) file then converted into a csv file. The Canterbury Region Census 2018 data was also included, more specifically income per person, age and gender count, and rent per suburb (Stats NZ 2021). This was downloaded as an excel (xlsx) file, and information was extracted from the relevant sheets in this file. CliFlo weather data was accessed from the National Institute of Water Atmospheric Research (NIWA 2021) via web scraping. These data sources are all openly available from reliable sources.

We are also planning to join all our datasets to Real Estate Institute of New Zealand's (REINZ) dataset which includes information about houses sold in Canterbury. REINZ data was procured through email correspondence and a Zoom meeting with representatives of REINZ and the University of Canterbury. All of the wrangling work for this project up to this point has been done using a fake dataset provided by REINZ. Unfortunately, logistics have prevented us from receiving the full dataset at this stage. Nonetheless, the structure of the fake dataset has guided our data processing, and once the full dataset is received, it can be integrated.

Spatial datasets

The sf library in R was used for manipulation of GIS data. The spatial data used to integrate the various datasets in this project was sourced from Land Information New Zealand (LINZ). Four important spatial datasets were used. These are shown in Table 1. These datasets were all obtained through the LINZ data portal (FENZ 2021, LINZ 2021, Stats NZ 2020, Stats NZ 2021).

Table 1: Spatial data sources

Dataset	**Source**	**Related geographic information**
NZ Land Districts	LINZ	Region boundaries
Fire and Emergency NZ Localities	FENZ	Suburbs (urban) and localities (rural)
Statistical Area 2	Stats NZ	Statistical area boundaries
Meshblock 2021 (generalised)	Stats NZ	Meshblock boundaries

sf uses planar spatial data. This means it uses data based on a coordinate system on a flat surface (plane). To create these datasets, geographical features of the spherical earth are projected onto a plane. The projection used in this project was the NZTM2000 projection (crs2139). Obviously, if the different datasets used differing spatial coordinate systems, they could not be used together accurately.

Wrangling techniques

General techniques

We used a wide range of data wrangling techniques in Julia and R. This included data collection techniques extracting data from csv and xlsx (excel) files, and web scraping CliFlo. Data cleaning involved removing unnecessary columns, removal and replacement of NA values, converting column data types to relevant types. Data structuring included structuring column names as DataFrame appropriate strings, transforming data into new columns such as normalizing data by population, and joining separate dataframes using suburb as the join condition. The following includes more detail about wrangling techniques for specific datasets.

Spatial data integration

The main dataset for this project is the REINZ property sales database. This database includes suburb and region columns. The wrangled datasets were integrated using these columns. The suburb column in the REINZ database is encoded in the form of FENZ localities. All other datasets used in this project, except the weather data, used a spatial geographic encoding of statistical area 2 (SA2). Unfortunately, SA2 areas and FENZ localities don't match together neatly. This is because these datasets are used for different things. The FENZ boundaries are simply required for area identification. The SA2 boundaries were created based on population, to give accurate statistical representations of areas. In urban areas, SA2 areas are smaller than that of FENZ areas but in rural areas, SA2 areas are larger than that of FENZ areas. In other

words, the two geographic encodings are not hierarchical. This creates the confusing situation where in some instances, there are multiple SA2 areas which could be within one FENZ suburb and in other instances, there are multiple FENZ suburbs that could be within one SA2 area. Additionally, the boundary polygons in the shapefiles for Sa2 and FENZ localities do not align. This meant that intersecting the polygons associated many SA2 areas and FENZ localities erroneously. To circumvent these issues, it was decided to simply associate areas by proximity. SA2 area's were assigned to their nearest FENZ locality, and vice versa. This match is not robust from a low-level GIS standpoint, but for this project it is reasonable and should provide adequate accuracy for high level analysis using our data model.

Weather data

Adding suburb and region to CliFlo dataset

Because one of the primary keys for this database is *suburb*, collected CliFlo data needs to have an associated suburb. Initially, a "coordinate to address to suburb", reversed geocoding technique was trialled for joining weather data with the other datasets. This approach failed because of inconsistencies in address formatting which prohibited automated suburb extraction from addresses. After trying and failing with the address based reverse geocoding method, a new method was required. The previous method was always a bit "hacky" and became clear that this was a job for GIS tools.

This was attacked by first getting information about all weather stations in NZ, then attaching suburb and region data to this dataset. This dataset with weather station locations was then used as an index to supply suburb and region names to newly scraped weather data.

Normally, station information would be done using the clifro command cf_find_station() but for an unknown reason, this function is unoperational at the time of writing. Instead, a list of operating weather stations was obtained from manually the CliFlo website. This process is shown in Appendix B.

CliFlo provides location data about its weather stations in the form of latitude and longitude coordinates. To assign a suburb and region to each station in NZ, the sf library was required. Here the relevant function is intersect. intersect takes two planar spatial datasets, of the same projection and reports on their relationships (i.e. if one is within another).

First, a list of spatial points was created from the coordinates in the weather station information. The projection for common GPS style coordinates is 4326. EPSG 4326 is not a planar type system. The weather station location points needed to be converted to NZTM, 2193, for comparison with LINZ datasets. After conversion, the list of points was intersected with one of the LINZ sourced spatial datasets. Where a point lay inside the boundaries of a particular polygon in the FENZ dataset, it was ascribed the geographic attribute (suburb/region/SA2/meshblock) associated with said polygon.

Once region and suburb data had been added to the dataset of NZ weather stations, stations in Canterbury were filtered into a set of target stations for weather data extraction.

Querying CliFlo

To extract weather data from CliFlo, cfUser,cfStation, cfDatatype objects and a date were passed into the clifro function cfQuery. clifro queried CliFLo and returned weather information easily converted into a dataframe.

From this target set, the weather station "agent" number was extracted. Agent numbers were used to create a list of cfStation objects. A simple list of numbers specifies the type of data for clifro to reques. This is provided in the form of a cfDatatype object. To query CliFlo, a cfUser object must be passed in cfQuery. This object is created using a CliFlo username and password. CliFlo accounts are free but have row extraction limits. clifro was halted when it tried to perform more than 10 queries in quick succession. clifro was also halted when it tried to query more than 20 stations simultaneously. To avoid these issues we broke the queries into parts. In this instance, subsequent queries were simply hard coded, but this is likely iterable.

From there, the resulting weather dataset was tidied up by renaming and dropping various columns. In this step, maximum and minimum temperatures in the weather data set were "mutated" into a median temperature column. The median temperature column was calculated using the formula: medianTemperature = \dfrac{max Temperature + min Temperature}{2}.

Victimisation Time and Place Data

After the *victimisation time and place* data set was converted from a Tableau Packaged Workbook to a csv file format, the data set was wrangled in R.

In R the following techniques were used for the *Victimisation Time and Place* dataset (but not limited to these):

- The *filter* function was used to select only Canterbury districts and areas.
- The *rename* function was used to rename column headings to more appropriate headings.
- *Subset* was used to remove crimes which did not include a suburb (these suburbs were originally labelled as *999999).*
- All the suburb names in the dataset initially had full stops at the end. *Substr* function was used to remove the full stop at the end of every suburb name.
- *Groupby* was used to make the suburb column the key column. Then the *sum* function was used to get the sum of crime records and crime victims per suburb, and the *n_disctinct* function was used to get the number of crime types per suburb (based on ANZSOC Division).
- A left join was used to join the *Victimisation Time and Place estimates* dataset onto the *Subnational population* dataset. A left join was chosen because we wanted to include all the Canterbury population data from the *Subnational population* dataset.

- *Transform* function was used to add two new columns to normalise the data by population. This was done by dividing the number of crimes (and victims) divided by the year 2020 suburb population times 1000.

Subnational Population Estimates Data

In Julia the following data wrangling techniques were used for the *Subnational population estimates* csv dataset:

- *DataFrame* function was used to convert a csv file into a DataFrame.
- Indexing was used to select only the suburb, 2013 and 2018-2021 suburb population columns.
- *Rename* function was used to rename a column from "June" to "Suburb".
- The *select* function was used to remove the 2013 suburb population column.
- The final dataframe was saved as a csv using *CSV.write* so it could later be joined onto the *victimisation time and place* dataset.

(Note the original dataset was originally an excel file, when the excel file was saved as a csv during the conversation process the first 4 rows and sixth row with the dataset information was removed. The fifth row first column heading changed to "June", and the second column which was empty was removed)

Victimisation Time and Place and Subnational Population Estimates Joined

The *Subnational population estimates* dataset and *victimisation time and place* dataset were joined together using R.

- The *merge* function was used to left join the *victimisation time and place* dataset onto the *Subnational population estimates* dataset.
- *Transform* function was used to normalise the number of crime records and number of crime victims, normalisation was done per 1000 people.

Area Data

The 2013 Canterbury Region dwelling dataset was used to for the area dataset. First the readxl package was used to import the Excel file in R. The Excel sheet of interest was selected and converted to a data frame. From that, columns of interest were selected, remaining got dropped and all columns got renamed. Then a new column was added called "suburb" which is based on the "area" column. Then all duplicates in the "area_name" columns were removed. Finally, the data frame was exported as a CSV.

Canterbury Census Data

The Canterbury census data was structured as individual sheets within a .xlsx format file. In order to read this data appropriately, the *readxl* and *tidyxl* packages were utilised, with the

function *readxl()* allowing the user to specify which sheet is read. After being read, the column names were formatted strangely, with periods replacing any spaces, and the column number was added to the end of each column name. This formatting was removed through use of the *sub()* function.

Following this, a variety of data wrangling techniques were employed to extract the census data referencing suburb **income**, **employment status**, and **gender**. These are as follows:

- The *filter()* and *na.omit()* functions were used to remove any rows with NA values.
- The *names()* function was used to rename columns to dataframe-appropriate values. For the income dataset column names, 'X' was added to the beginning of any columns that originally started with a number
- *distinct()* was used to obtain the names of all unique suburbs in Canterbury
- The *sapply()* function was used when changing column data types to appropriate values
- Dataframe indexing was used to extract relevant columns

A similar process was followed to extract the **income** data from the .xlsx file using *Julia*, as an alternative process in case the user wants to avoid the use of R. This required the use of the "XLSX" package, specifically the *readxlx()* function, which allows the specify the desired workbook sheet when indexing referencing data. Additionally, the *rename!* function was used to convert column names to dataframe-appropriate values; *unique()* was used to extract the distinct suburbs from the dataset; and an *innerjoin* was used to join the extracted income data with the suburbs data, creating an enriched dataset containing income by suburb, alongside Territory and Region.

School Data

A directory of New Zealand schools was available on the Education Counts website. The website allows you to filter the data. Filtering by region and selecting only Canterbury was chosen to get a reduced dataset of just schools in Canterbury. The dataset was in a .CSV format and imported using the DataFrames and CSV package within JupyterHub and Julia kernel. The data was already in a tidy format and relevant columns as determined by the group were chosen with the remaining being dropped. Some columns got renamed using the *!rename* function, such as "Total School roll" to "Total Student". Some values in the decile had a value of "99" for missing value. Using the *!replace* function of Julia, they were replaced to "NA". Last but not least, the final dataframe was than exported as school.csv using the *CSV.write* function.

Data structure

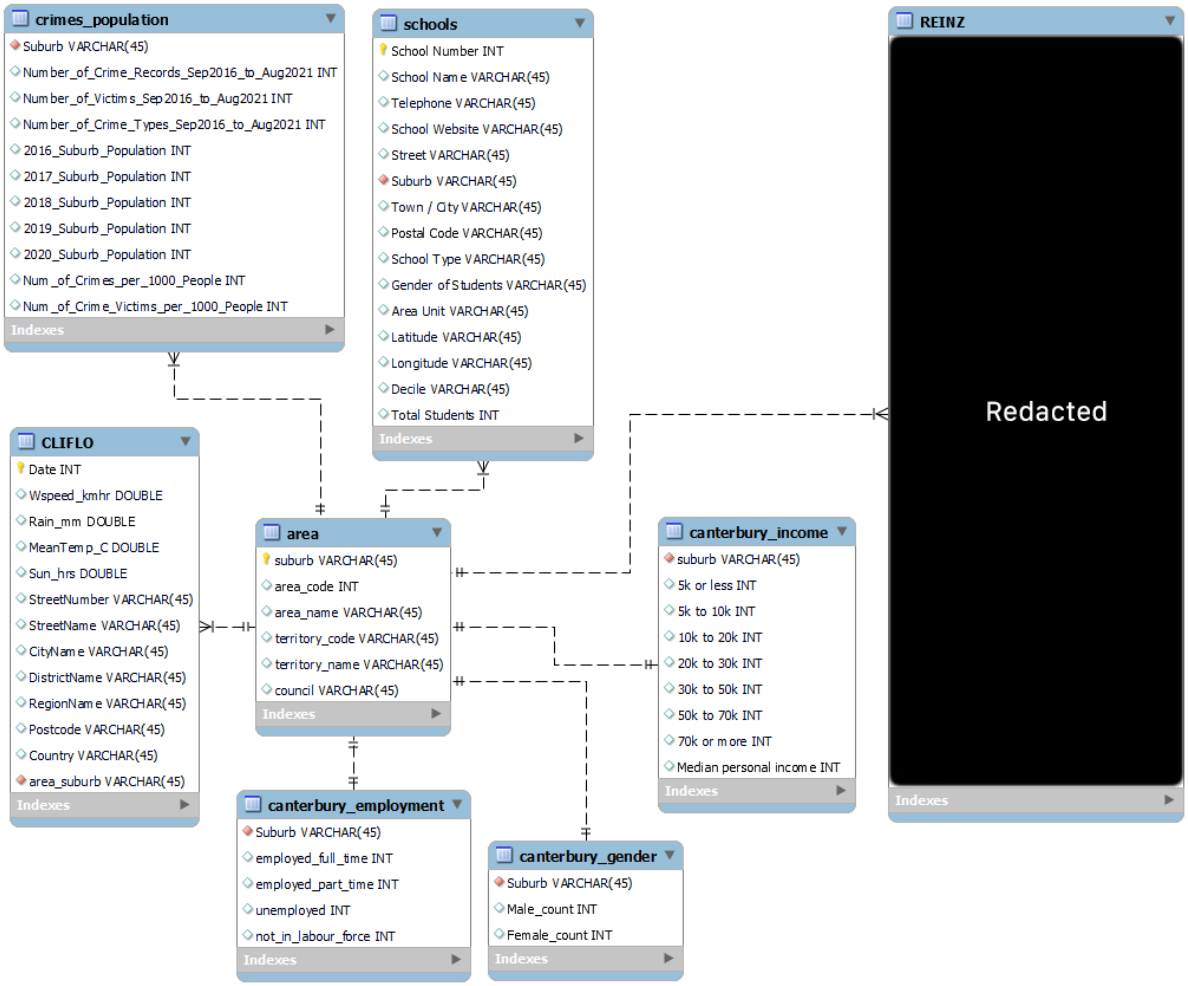

Figure 1: Data model

The data structure of the finalised database can be observed in Figure 1. It consists of a collection of the datasets collected throughout the wrangling process. All datasets (aside from the REINZ data) utilise the "SA2" suburb naming format, which will be changed to the emergency services suburb naming format used by the REINZ data once access has been provided. Subsequently, suburb name is used as the join condition for each of the datasets. The **schools**, **crimes_population_data**, **CLIFLO**, and **REINZ** datasets all have a many-to-one relationship to the **area** dataset, while the **canterbury_income** dataset has a one-to-one relationship. Appendix A contains clarifying information on what each column in each of the remaining datasets refers to, and Appendix B documents the structure of the CLIFLO data.

Please note that the data structure of the REINZ dataset has been redacted due to privacy agreements relating to the use and publication of this data, to ensure all identifying information and potentially valuable information remains private.

Visualisations

In order to ensure our final data model was formatted correctly and was working as intended, we created some visualisations to explore the relationship between several of the geographical and suburban factors within our datasets. These are just a few of the possible ways this data can be explored. Future investigation of this data would include application of the REINZ data, such as employing linear regression to determine which geographical features are most influential in impacting house prices.

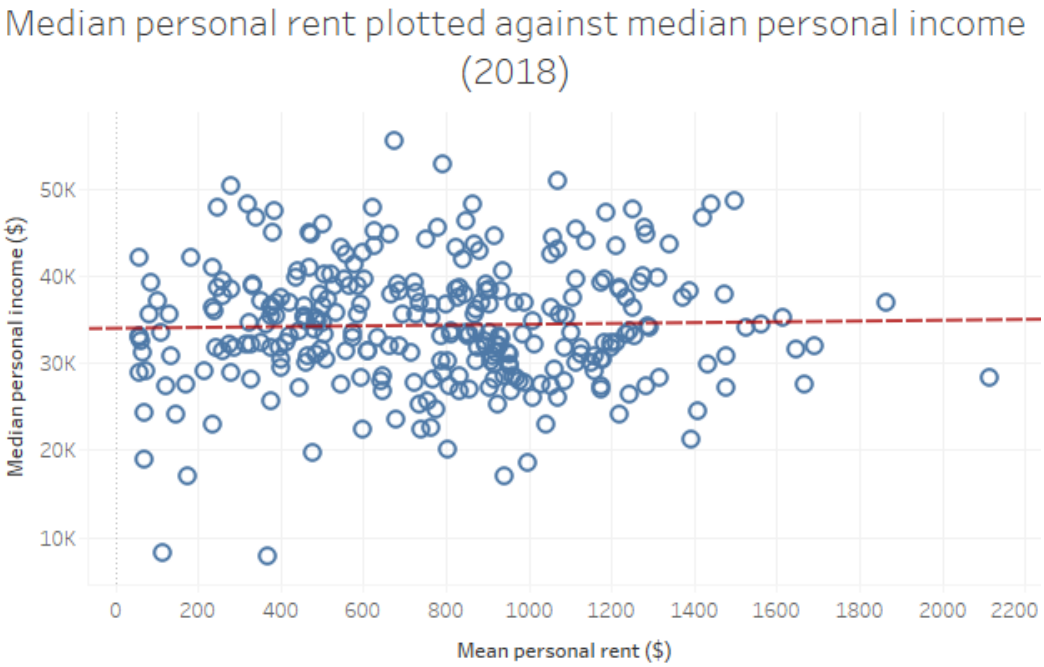

Figure 2: Median personal rent against median personal income, by suburb

The median personal rent against median personal income by suburb is shown in Figure 2. In exploring the correlation between suburb-level median personal rent and income, we expected that rent would increase relatively linearly as income increases. Interestingly, our data appears to show almost no linear correlation, with median personal income values remaining relatively consistent in suburbs regardless of what the median personal rent is.

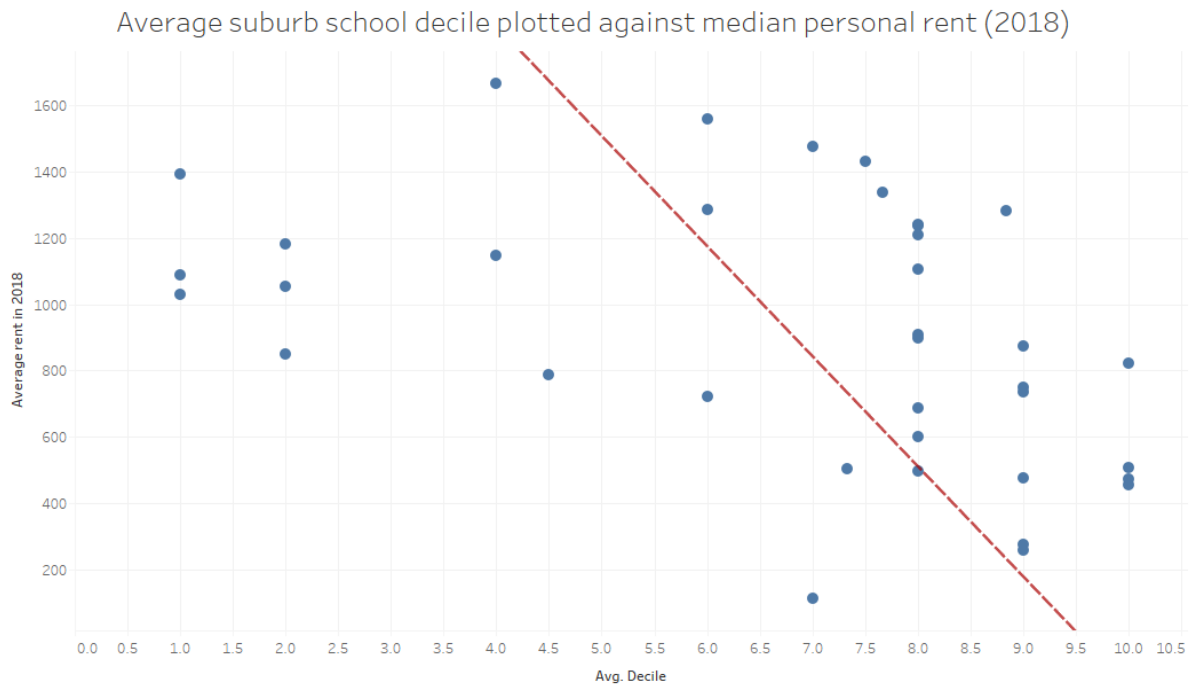

Figure 3. Average school decile against median personal rent, by suburb

Plotting median personal rent against average school decile rating also returned some interesting results. This is shown in Figure 3. As with the previous data exploration, we would expect that median rent would increase as average school decile rating increased. This was not the case with our data, as observation of the trend line shows us that the opposite actually occurs. This linear relationship is weak, however, and we can see that the distribution of median rent is very varied, particularly amongst suburbs with a high average school decile rating. Taking into account the previous visualisation, we can determine that decile rating has virtually no impact on median suburb rent.

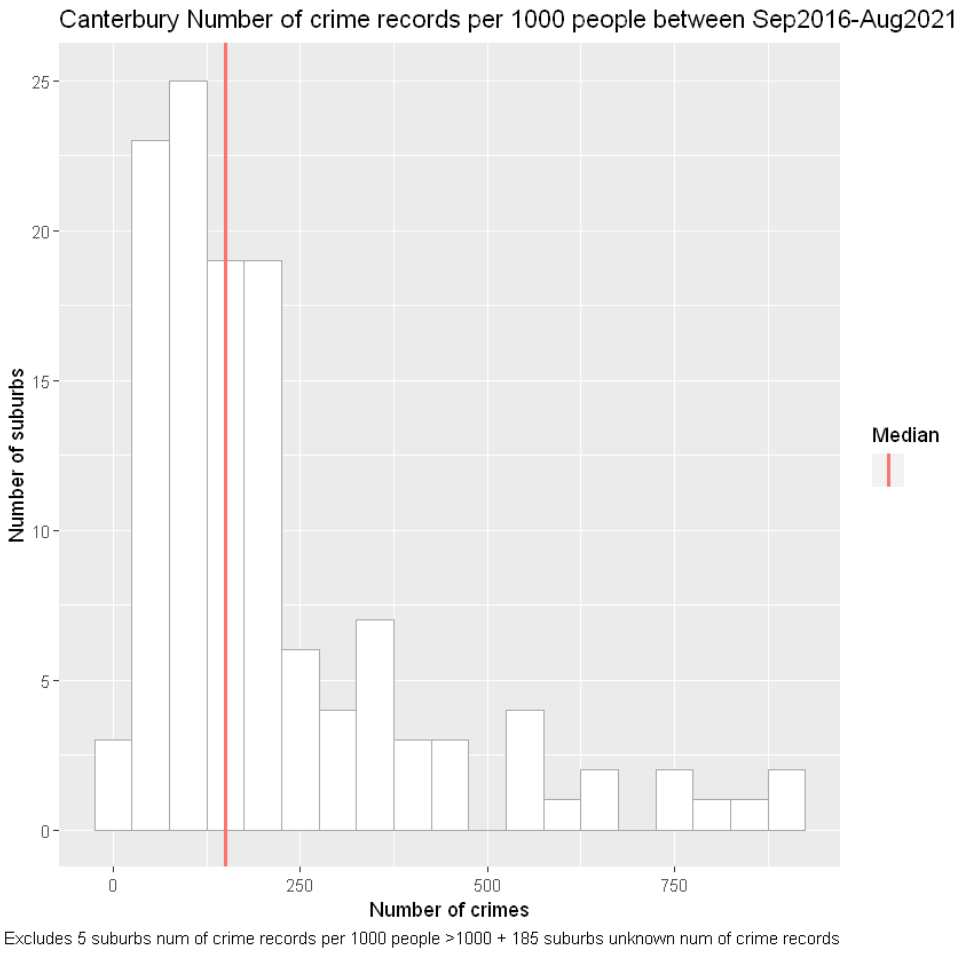

Figure 4. Number of crimes recorded per 1000 population by suburb

Finally, we explored the Canterbury crimes dataset, to determine whether there were any patterns in the data. Unlike the previous two visualisations, the results in Figure 4 are closer to what would be expected. We can see that the distribution of crimes in Canterbury suburbs has a strong right skew, entailing that most suburbs have a relatively low level of crime, with a median number of 156 crimes committed in each suburb in the 5-year period between Sep 2016 and Aug 2021 per 1000 population. However, there are still a number of suburbs reporting high levels of crime, with 12 suburbs reporting over 500 crime records per 1000 people in a 5-year period.

Achievements

After successfully wrangling and scraping the chosen datasets, we have created a data model consisting of a variety of suburban and geographical features that correlate with house prices (Figure 1). By using suburb as a central focus for our data model, we have managed to unify all datasets under a single join condition in which suburb acts as the primary or secondary key

for each, making querying tables particularly easy. This collection of datasets can act to enrich the confidential REINZ database, providing greater depth to any investigation or analysis that may be done into factors that may influence housing prices in Canterbury, in addition to any other investigations into suburban and geographical features.

Challenges & limitations

Challanges

Residential block naming structure:

We had problems in residential block naming structure, which meant that the dataset use had multiple methods of structuring residential blocks. For example, the 2013 census data provided variable location names in the dataset, such as suburb and statistical area.

Solution:

We had to identify rows not fitting the suburb naming and convert these to conventional format and aggregate values.

Multiple datasets format:

We used three different suburb area types, Fire and Emergency NZ Localities, Statistical Area 2 (SA2), and Police Territorial Authority. These different datasets join on the suburb column, but the datasets use different suburb area types so many of the rows had N/A and Null values.

Solution:

Identify the different names of suburbs and unified name to drop off the suburbs which are unverifiable.

We realised that the SA2 and the Police Territorial Authority frequently had compass directions (south, east, west, north) at the end of suburb names; so we added compass directions at the end of suburb names and aggregated columns with the same suburb name.

Limitations

Data ethics in variable selection:

Although the census datasets collect a wide variety of variables, there are ethical reason to not use some of the variables in public investigation. Therefore, we had to consider which variables could be informative in a non-sensitive way. The data analysis of certain factors can cause the unintended harm to some individual, groups or regardless of intentions.

Conclusion

We developed a dataset that contains a variety of different suburban factors that relate with house price.

Income and average rent

From the median personal rent and income plots, median personal income remains relatively stable, and does not change the median suburbs rent much depending on income.

Average rent nearby school

Based the survey data and plotting showed that rent increases have not followed in school scores. The linear relationship is weak, there is non-relationship between the average rent and nearby schools.

The average rent showed a different mode of housing price. According to the National Bureau of Economic Research there is a certain correlation between school expenditure and family value. For every dollar spent on public schools in community, the value of the house increases by $20. These findings suggest that additional school spending may benefit everyone in the community, whether they have children in the local public school system (Gorman, L. 2003).

Crime population

According to the police crimes_population_data dataframe data output. Southbridge has the lowest of number of crime records and number of victims per 1000 people. The five suburbs with the lowest number of crimes per 1000 people are Southbridge, Ashburton East, Rolleston South West, Rolleston North West, and Prebbleton.

Some studies showed that violent crime leads to the decline of house prices, but the relationship between property crime and house values is more blurred. Reducing the standard deviation of local criminal damage density by a one-tenth will increase the average property price 1% (Steve, 2004).

Households

From 2016 to 2018 the top 10 suburbs' household Tibble, the Linwood West has the highest total households in occupied private dwellings since 2016. In 2008 and 2006, it owned more than 2000 private houses.

Locality weather

From the CliFlo dataset, we are purposed to explore the potential future impact in housing price. In an equilibrium model of housing choice, price shows different elasticity to climate risk (Markus et al., 2020). Our results find that house prices reflect the heterogeneity of long-term climate change risk.

References

Gorman, L (2003). School spending raises property values. National Bureau of Economic Research. Retrieved 29 October 2021, from nber.org/digest/jan03/school-spending-raises-property-values

LINZ (2021). NZ Land Districts. Retrieved 29 October 2021, from https://data.linz.govt.nz/layer/50785-nz-land-districts/

FENZ (2021). Retrieved 29 October 2021, from https://data.linz.govt.nz/layer/104830-fire-and-emergency-nz-localities/

Markus, B., Lorenzo, G., Constantine, Y. (2020). Does climate change affect real estate price? Only if you believe in it. *The Review of Financial Studies, 33*(3) 1256-1295. Retrieved 30 October 2021 , from https://doi.org/10.1093/rfs/hhz

McKnight, E. (2021). Canterbury Property Market. OPES PARTNERS. Retrieved 30 October 2021, from https://www.opespartners.co.nz/property-markets/canterbury

Ministry of Education (2021). New Zealand Schools. Retrieved 21 October 2021, from https://www.educationcounts.govt.nz/directories/list-of-nz-schools

NIWA (2021). CliFlo – The National Climate Database. Retrieved 27 October 2021, from https://cliflo.niwa.co.nz/

NZ Police (2021). Victimisation Time and Place. Retrieved 31 October 2021, from https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz/victimisation-time-and-place?nondesktop

REINZ (2021). Residential Statistics Report for September 2021. Retrieved 29 October 2021, from https://www.reinz.co.nz/residential-property-data-gallery

Stats NZ (2020). Statistical Area 2 2020 (generalised). Retrieved 29 October 2021, from https://datafinder.stats.govt.nz/layer/104271-statistical-area-2-2020-generalised/

Stats NZ (2021). Statistical area 1 dataset for 2018 Census – updated March 2020. Retrieved 31 October 2021, from https://www.stats.govt.nz/information-releases/statistical-area-1-dataset-for-2018-census-updated-march-2020

Stats NZ (2021). Subnational population estimates. Retrieved 31 October 2021, from http://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7993&_ga=2.104052355.2017186628.1634336143-719107025.1632194594&_gac=1.187780826.1634341688.CjwKCAjwzaSLBhBJEiwAJSRokp1ezN-rHwcbENfgmCOaqB5ZV1PjF8PHuwo4KUjET-ASdu_WBRRo5xoCnIgQAvD_BwE

Stats NZ (2021). Meshblock 2021 (generalised). Retrieved 29 October 2021, from https://datafinder.stats.govt.nz/layer/105176-meshblock-2021-generalised/

Steve, G (2004). The Costs of Urban Property Crime. *The Economic Journal*, *114*(499), F441-F463. Retrieved 30 October 2021, from https://doi.org/10.1111/j.1468-0297.2004.00254.x073

Appendix A – Data structure

Canterbury census (updated 2018)

canterbury_income – contains counts of individuals within specific yearly income brackets in 2018:

suburb (string / chr) – respective suburb from which data was collected

X5k_or_less (integer) – count of individuals within the income bracket $0-5000

X5k_to_10k (integer) – count of individuals within the income bracket $5000-10000

X10k_to_20k (integer) – count of individuals within the income bracket $10000-20000

X20k_to_30k (integer) – count of individuals within the income bracket $20000-30000

X30k_to_50k (integer) – count of individuals within the income bracket $30000-50000

X50k_to_70k (integer) – count of individuals within the income bracket $50000-70000

X70k_or_more (integer) – count of individuals within the income bracket $70000+

median_personal_income (integer) – median personal income for referenced suburb

canterbury_region – contains regional location information for the Canterbury region:

Suburb (string / chr) – contains the names of all suburbs in Canterbury

Territory (string / chr) – contains the names of all territories in Canterbury

canterbury_gender – contains counts of individuals' genders in the Canterbury region. Please note that this data was collected in 2018, and due to data collection methods at the time does not reflect the counts of transgender, non-binary and other genders in the region.

Suburb (string / chr) – contains the names of all suburbs in Canterbury

Male_count (integer) – count of individuals who are male

Female_count (integer) – count of individuals who are female

canterbury_employment – contains counts of residents' employment status at the time of the 2018 census collection:

Suburb (string / chr) – contains the names of all suburbs in Canterbury

employed_full_time (integer) – count of individuals employed full-time

employed_part_time (integer) – count of individuals employed part-time

unemployed (integer) – count of individuals not in employment but able to work

not_in_labour_force (integer) – count of individuals not in the labour force

area – contains all areas in the Canterbury region

suburb (string / chr) – Suburb names in the Canterbury region

area_code (integer) – Unique area code

area_name (string / chr) – Unique area name

territory_code (integer) – Code of the territory

territory_name (string / chr) – Code of the territory

council (string / chr) – Council responsible for the suburb

School Directory – Education Counts

schools – contains all schools in the Canterbury region and its contact details and location

School name (string / chr) – Name of the school

Telephone (integer) – Telephone number of the school

School Website (string / chr) – Public website address of the school

Street (string / chr) – Street name where the school is located

Suburb (string / chr) – Suburb name where the school is located

Town / City (string / chr) – Street name where the school is located

Postal Code (integer) – Postal code where the school is located

School Type (string / chr) – School type, whether its primary, secondary, composite or special

Gender of Students (string / chr) – Co-educational or single-sex (either male or female only)

Area Unit (string / chr) – Area name where the school is located

Latitude (float) – street name where the school is located

Longitude (float) – street name where the school is located

Decile (integer) – School decile

Total Students (integer) – Number of total students

Victimisation Time and Place and Subnational population estimates:

crimes_population_data – contains counts of crime records, number of crime victims, and number of crime types between September 2016 - August 2021, 2018-2021 suburb population data, and normalised crime data:

Number_of_Crime_Records_Sep2016_to_Aug2021 (dbl) – contains the number of crime records between September 2016- August 2021 in Canterbury

Number_of_Victims_ Sep2016_to_Aug2021 (dbl) - contains number of victims of crimes between September 2016- August 2021 in Canterbury

Number_of_Crime_Types_ Sep2016_to_Aug2021 (int) – contains number of distinct types of crimes between September 2016- August 2021 in Canterbury (based off ANZSOC Division)

2018_Suburb_Population (dbl) - Number of residents in each suburb in Canterbury in 2018

2019_Suburb_Population (dbl) - Number of residents in each suburb in Canterbury in 2019

2020_Suburb_Population (dbl) - Number of residents in each suburb in Canterbury in 2020

2021_Suburb_Population (dbl) - Number of residents in each suburb in Canterbury in 2021

Num_of_Crimes_per_1000_People (dbl) - Number of crimes recorded in Canterbury between September 2016- August 2021 for each suburb divided by its corresponding suburb population in 2020.

Num_of_Crime_Victims_per_1000_People (dbl) - Number of crime victims recorded in Canterbury between September 2016- August 2021 for each suburb divided by its corresponding suburb population in 2020.

Subnational population estimates

suburbs - Subnational population estimates (DHB, DHB constituency). Contains suburb population counts between the years 2018-2021:

Suburb (string / chr) - suburb name

2018 (int) - suburb population in the year 2018

2019 (int) - suburb population in the year 2019

2020 (int) - suburb population in the year 2020

2021 (int) - suburb population in the year 2021

Appendix B – CliFlo Station List Acquisition

To acquire all stations in NZ, the center of NZ was chosen as a coordinate, then a radius specified of 1000 km. CliFlo returned comma seperated text containing all of the weather stations in NZ. Figure B 1 and Figure B 2 show the CliFlo web portal interactions.

It turned out that some of the station names had commas in them. In those rows the last two columns had been squished together. To fix this, the following code was run.

```
fix station list
readLines("collectedData/clifloStationList-Active.csv") %>%
  str_replace(pattern = ", ", replacement = "") %>%
  writeLines("collectedData/clifloStationList-Active.csv")
read.csv("collectedData/clifloStationList-Active.csv") %>% as_tibble %>%
  separate(col = Long.dec_deg., into = c("Long.dec_deg.","drop"), sep = " ") %>%
  rename(lat = Lat.dec_deg., lon = Long.dec_deg.) %>%
  select(-drop) %>% select(-`Dist_Km`) %>%
  write.csv(file = "collectedData/clifloStationList-Active.csv")
```

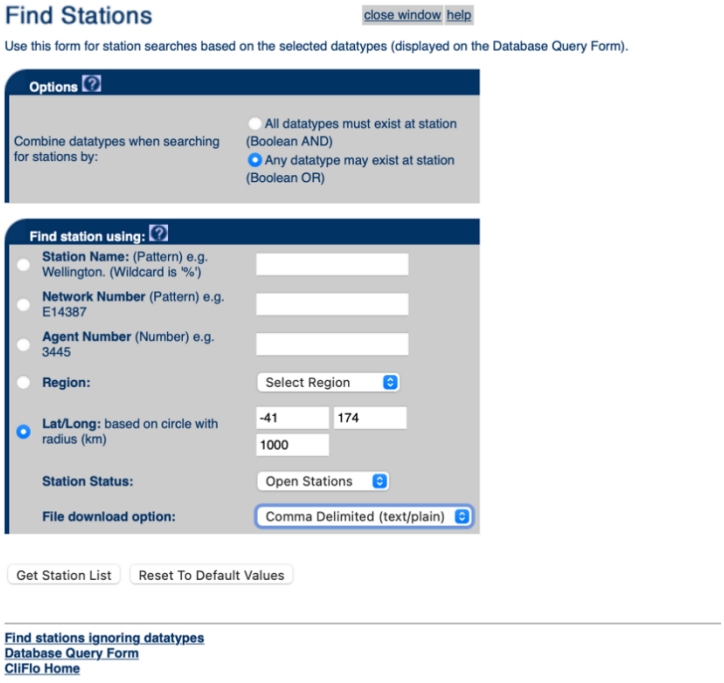

Figure B 1: CliFlo web portal station finder

```
Station Listing
Selected DataTypes were combined by: "ANY selected datatypes MAY exist (OR'd)"
Selected Datatypes are:
Code,Description
312,F301_Data: Climate Obs From Stns With Extended Instrumentation

Agent,Network,Start_Date,End_Date,Percent_Complete,Name,Lat(dec_deg),Long(dec_deg),Dist_Km
4232,G13195,19-Jun-2006,19-Jun-2006,100,Pelorus Sd, Crail Bay,-41.10308,173.96406,   11.8
26169,G04602,01-Oct-2018,27-Oct-2021,90,Stephens Island Aws,-40.66511,173.99969,   37.2
4395,G14142,01-May-1997,28-Oct-2021,100,Brothers Island Aws,-41.1033,174.44174,   38.7
12430,G13495,01-Jun-1996,29-Oct-2021,100,Blenheim Research Ews,-41.49891,173.96285,   55.5
4322,G13581,01-Jan-1972,31-Mar-1987,100,Blenheim Aero,-41.52,173.872,   58.7
4326,G13585,01-Oct-1990,28-Oct-2021,100,Blenheim Aero Aws,-41.52135,173.86432,   59.0
25531,E14073,01-Dec-2006,28-Oct-2021,100,Mana Island Aws,-41.08676,174.77996,   66.0
41559,E1418H,01-May-2017,26-Oct-2021,100,Porirua, Elsdon Park Aws,-41.12683,174.83531,   71.4
25354,E1427P,13-Jul-2004,29-Oct-2021,100,Wellington, Kelburn Aws,-41.28445,174.76794,   71.7
3385,E14272,01-Jan-1954,31-Aug-2005,100,Wellington, Kelburn,-41.286,174.767,   71.7
18468,G14604,01-Sep-2000,31-Aug-2013,100,Awatere Valley, Dashwood Raws,-41.64724,174.07274,   72.1
4241,G13222,01-Dec-1961,30-Oct-2021,100,Nelson Aero,-41.299,173.226,   72.8
4271,G13322,01-Feb-1992,29-Oct-2021,100,Nelson Aws,-41.30081,173.21608,   73.7
41212,E1438H,01-Nov-2015,29-Oct-2021,100,Wellington, Greta Point Cws,-41.30243,174.80574,   75.4
36106,G13543,01-Nov-2008,03-Oct-2021,80,Wairau Valley, Mill Road Cws,-41.572,173.497,   76.2
3445,E14387,01-Jan-1972,28-Oct-2021,100,Wellington Aero,-41.322,174.804,   76.2
41229,E1438I,01-Nov-2013,28-Oct-2021,100,Wellington Aero Backup Aws,-41.3313,174.80566,   76.9
40751,G13314,01-May-2015,30-Oct-2021,100,Richmond Ews,-41.32773,173.18615,   77.3
4420,G14711,01-Jan-1972,31-Aug-2021,100,Grassmere Salt Works,-41.729,174.145,   81.8
3145,E04991,01-Jan-1972,30-Oct-2021,100,Paraparaumu Aero,-40.907,174.984,   83.1
12442,E04995,01-May-2016,30-Oct-2021,100,Paraparaumu Ews,-40.90392,174.98437,   83.2
8567,E04994,01-Apr-1993,29-Oct-2021,100,Paraparaumu Aero Aws,-40.90455,174.98517,   83.2
21937,G13305,01-Apr-2001,29-Oct-2021,100,Appleby 2 Ews,-41.31727,173.09482,   83.6
4424,G14722,01-Sep-1999,28-Oct-2021,100,Cape Campbell Aws,-41.7265,174.2762,   83.9
18234,D14482,01-Oct-2001,29-Oct-2021,70,Baring Head,-41.40828,174.87146,   85.9
12429,G12195,01-Apr-1995,29-Oct-2021,100,Motueka, Riwaka Ews,-41.09798,172.97165,   86.8
40750,E1510M,03-Jan-2015,30-Oct-2021,100,Upper Hutt, Trentham Ews,-41.14027,175.04283,   88.7
3798,F03502,01-May-1982,28-Oct-2021,90,Farewell Spit Aws,-40.54483,173.00097,   97.8
23849,F02885,01-May-2003,30-Oct-2021,100,Takaka Ews,-40.86364,172.80568,  101.2
41352,E0562B,01-May-2016,28-Oct-2021,100,Levin Ews,-40.62699,175.26193,  113.5
3275,E05620,03-Nov-1990,29-Oct-2021,100,Levin Aws,-40.61986,175.25954,  113.6
21938,D15234,01-May-2001,30-Oct-2021,100,Martinborough Ews,-41.25231,175.38985,  119.8
2685,D15521,01-Nov-1999,28-Oct-2021,100,Ngawi Aws,-41.5885,175.2337,  122.3
31850,F12885,01-Jan-2007,31-Jul-2012,100,St Arnaud Raws,-41.80779,172.84213,  132.1
36735,D0596H,01-Mar-2009,29-Oct-2021,100,Masterton Aero Aws,-40.97617,175.63899,  137.3
40984,D0596I,01-Nov-2015,29-Oct-2021,100,Masterton Ews,-40.98155,175.67932,  140.7
37662,D05974,01-Oct-2009,29-Oct-2021,90,Masterton, Te Ore Ore Cws,-40.957,175.707,  143.1
3716,E95903,01-Aug-1978,28-Feb-1987,100,Wanganui Aero,-39.96287,175.02564,  143.6
3719,E95906,01-Aug-1987,28-Oct-2021,100,Wanganui Aws,-39.96122,175.02581,  143.8
37257,E05136,01-May-2019,28-Oct-2021,100,Ohakea Aws,-40.20301,175.37203,  145.0
3206,E05231,01-Jan-1954,29-Nov-1991,100,Ohakea Aero,-40.201,175.373,  145.2
3715,E95902,01-Jan-1972,30-Oct-2021,100,Whanganui, Spriggens Park Ews,-39.93704,175.04509,  146.9
21963,E0536D,01-Apr-2001,30-Oct-2021,100,Palmerston North Ews,-40.38195,175.60915,  151.3
3236,E05361,01-Jan-1972,30-Sep-1988,100,Palmerston N Aero,-40.327,175.616,  154.6
3243,E05368,01-Sep-1991,28-Oct-2021,100,Palmerston North Aws,-40.31858,175.61459,  155.0
25222,E94622,01-Feb-2004,29-Oct-2021,100,Hawera Aws,-39.60991,174.2916,  156.2
25777,F12215,26-Jun-2015,30-Oct-2021,100,Arapito Ews,-41.27058,172.15568,  157.4
```

Figure B 2. CliFlo output station list