DATA422 Assignment 3:
Analyzing the Impact of School Count and Crime Rate on Housing Prices in Aotearoa

Group members: Emma MARIE (68268218), Sarah OONG (96301834), Huaipu (Levo) WANG (28629644), Jingcheng (Jacci) ZHOU (82204407）

I. Introduction

The housing market in New Zealand has consistently been a topic of significant interest, as it reflects not only the housing needs of families but also investment opportunities and changes in socioeconomic conditions. Housing prices vary significantly across different regions of New Zealand, and these variations are influenced by a multitude of factors. This project aims to conduct an analysis of two primary influencing factors: schools and crime rates, to investigate how they impact housing prices in New Zealand across various regions.

Indeed, the quality and quantity of schools have consistently been crucial considerations for homebuyers and families. High-quality schools attract families to purchase properties in their proximity to ensure their children receive a quality education. Consequently, the number and reputation of schools can directly influence housing prices. In certain regions, property prices may be higher as they are located near renowned schools or have a greater number of schools, as individuals are willing to invest more for their children's education.

Crime rates are also a critical factor in housing prices. Areas with low crime rates are typically perceived as safer and more desirable, attracting more homebuyers. People prefer to purchase properties in areas with lower crime rates to ensure the safety of their lives and assets. High-crime areas may have a negative impact on housing prices as potential buyers may be concerned about safety issues.

Through regional analysis in this study, the project aims to explore the interplay between schools and crime rates and how they shape the housing price landscape in New Zealand. This will provide valuable insights for homebuyers, investors, and policymakers, aiding in a better understanding of the dynamics and trends in housing prices across different regions.

II. Data Sources

The data was obtained from various different sources.

1. The NZ school directory data was obtained directly from Education Counts. All territorial authorities were selected and the data was exported as a CSV file.

2. The crime data was obtained from PoliceNZ. The data covered the time period from January 2023 to June 2023 and records the number of victimizations for different crime types for each territorial authority during this time frame. The data was downloaded as a CSV file.

3. The housing price data for each territorial authority was scraped from Opes Partners (see Techniques Employed section for scraping details).

4. Population estimates for each territorial authority was also collected from StatsNZ (download data section) in order to normalize the school counts and crime rate. The data was downloaded as an XLSX file.

5. In order to create a map of the territorial authorities, NZ territorial authority polygon data was obtained from StatsNZ.

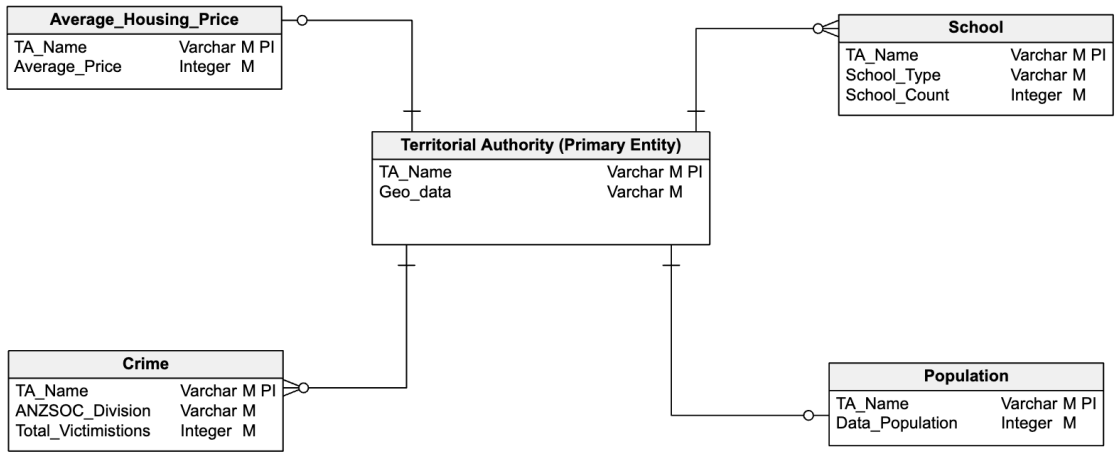
III. Relational Data Model

Vertabelo was used to create the relational data model (see Figure 1). First, five distinct entities were defined, specifically “Average_Housing_Price”, “School”, “Crime”, “Population” and “Territorial Authority”. "Territorial Authority" is the primary key used to connect and establish relationships between these tables. In each table, only the attributes that are relevant to the project were selected.

The relationships between these entities were illustrated as follows:

1. **Average_Housing_Price and Territorial Authority (One-to-One):** This relationship means that for each record in the "Territorial Authority" table, there is exactly one corresponding record in the " Average_Housing_Price " table. It suggests that house price data is specific to each individual territorial authority, creating a direct one-to-one relationship.
2. **Crime and Territorial Authority (Many-to-One):** In this relationship, there are multiple records in the "crime" table associated with a single record in the "Territorial Authority" table. This suggests that multiple victimisations are reported within a single territorial authority, creating a many-to-one relationship.
3. **School and Territorial Authority (Many-to-One):** Similarly, this relationship implies that there are multiple school records associated with a single territorial authority. It suggests that multiple schools are located within the boundaries of a single territorial authority.

4. **Population and Territorial Authority (One-to-One):** This relationship means that population data is specific to each individual territorial authority, and there is one population record for each territorial authority, indicating a one-to-one relationship.



(Figure 1: Relational Data Model)

Therefore, the relational data model reflects the link between the five data sources (section II) together. One important point to mention is that the final data visualization involved creating map-based reports in the Shiny application. These reports can depict the crime rate, the number of schools, and house prices for different territorial authorities. To ensure that the crime rate and the number of schools could be comparable between territorial authorities, normalization using population data was applied (section V. E).

IV. Techniques Employed

Scraping - Julia

HTTP Package - The http package was used to send GET requests to the specific URL to get the CSV file.

Data Wrangling - Julia

CSV Package - The csv package was used to read and write the CSV file.
DataFrames Package - Similar to the dplyr library in R, the dataframes package was used for data manipulation.

Data Wrangling - R

Readxl library- The readxl package was used to load/read excel files.

Tidyverse library- Only one package required as tidyverse is a collection of packages designed for data science. (Tidyverse) 2 core tidyverse packages were used: dplyr and readr. Dplyr was used mainly to perform data manipulations (e.g., slice(), filter(), slice(), filter(), select(), group_by(), summarize(), left_join(), merge()) and readr was used for saving the clean data frame to a CSV format (write_csv()). Moreover, the magrittr pipe operator was used throughout to increase code readability.

Sf library- The sf package was used for dealing with spatial data. More specifically, functions such as st_read(), st_write() and st_transform() were used to read/write shape files, and to convert spatial coordinates into the required format, respectively.

Base library - The base library was also used to perform specific tasks such as dealing with non-consistent string variables (e.g., gsub()) or to obtain more information about the data itself (e.g., unique(), length(), sum()).

Data Visualization (R)

R Shiny was used to create an interactive dashboard to display the data. Using ggplot2, Plotly and Leaflet, interactive plots were generated to display the impact of the number of schools and the crime rate on housing price as well as to illustrate the differences in housing price, number of schools and crime rate between territorial authorities (through a choropleth map).

GitHub

In modern data-driven environments, effective collaboration is paramount in ensuring the success of any group project. GitHub, a widely adopted version control platform, was thus used to facilitate seamless teamwork in the data wrangling process. This approach enhances transparency, version control, and accountability throughout the data wrangling process, ultimately leading to more robust and reliable analyses (Juviler, 2022)

A new repository on GitHub was dedicated to this data wrangling project, and a clear directory structure for organizing the data, code, and outputs was established to ensure a systematic approach. Each team member contributed to the repository by making regular commits. A ReadMe file detailing the structure of the repository and its contents was also created.

V. Data Wrangling

In this extensive data processing and cleaning section, five datasets from the mentioned sources were acquired. Both Julia and R were used for this data wrangling process.

Careful attention was paid to data exploration to familiarize with the data. Data consistency was addressed, ensuring that each dataset was up to standards. Transformations were applied, refining the data to the right data type. Integration efforts were made to synthesize information from various sources, creating a cohesive dataset. Rigorous validation and quality checks were implemented to fortify the reliability of the data.

A. Housing Price (Julia)

Both data acquisition and cleaning for the housing price dataset (‘housedata’) were executed entirely using Julia. The data was scraped by sending an HTTP GET request to the specified URL and converted the data from binary format into a DataFrame.

As 'Territorial_Authority' acts as the primary key in the relational data model, the process commenced with the renaming of the 'Location' column to 'Territorial_Authority'. Following this modification, the data was cleansed by removing the initial row, which contained an aggregate of New Zealand data, not relevant to this specific analysis. Subsequently, the 'Price' column underwent refinement by removing commas from numeric representations and converting them into integers. This transformation was accomplished using the parse() function after comma removal. To streamline the dataset further, only the 'Territorial_Authority' and 'Price' columns were selected, creating a more focused subset.

The resulting processed dataset was then saved in a new CSV file, 'housedata.csv,' in preparation for subsequent steps such as joining tables.

B. Crime Data (R)

Data processing for the "crime_data" dataset was carried out using the R. As mentioned previously, the original data was downloaded from the NZ Police website in CSV format. However, the file was in UTF16 format, which led to anomalies during the import into Jupyter Notebook. In particular, the "Territorial Authority" column exhibited irregular formatting and lacked clear delimiters between columns, resulting in improper data display. To address this issue, the CSV file was converted to an Excel format to ensure data integrity and consistency.

Subsequently, only columns of interest were selected, including "Territorial Authority," "Anzsoc Division," "Year Month," and "Victimisations". In an effort to improve clarity and maintain a consistent naming convention, the columns were renamed by replacing spaces with underscores.

Furthermore, trailing periods in the "Territorial Authority" column were present. These were eliminated to enhance data accuracy. A check to verify if the dataset contained all "Territorial Authority" values was also conducted.

A summary dataset named "victims_by_TA" was then created by grouping the data based on "Territorial Authority" and "Anzsoc Division," calculating the total victimizations for each specific crime type within each territorial authority.

It should be also noted that data missing issues when joining tables (see section E) were encountered. After a careful examination of each dataset, a spelling mistake was identified as the cause of this issue: "Whanganui District" was misspelled as "Wanganui District" in this "crime_data" dataset. This was therefore corrected to "Whanganui District."

C. School Data (R)

Using R, the school dataset ('School_Data.csv') was read using the read.csv() function. The first 15 rows of the data were hidden as they were header introductions in the original CSV data. In addition, only specific columns of interest were selected, including 'Territorial.Authority', 'School.Type', and 'Total.School.Roll'.

Examination of the data table revealed that the Auckland territorial authority was divided in multiple suburbs (in the format Auckland-Suburb - e.g., Auckland - Hibiscus and Bays). The 'Territorial.Authority' column thus underwent some data cleaning using the gsub() function to remove the Auckland suburb.

Unique values from 'Territorial.Authority' and 'School.Type' were then extracted, displaying them along with the total count of unique values. This step is to know the unique content of each column, especially to check if the values in the 'Territorial.Authority' column are consistent with the other datasets as this is the primary key.

Finally, a summary dataset named 'group_by_data_draft' was created by grouping the data based on 'Territorial.Authority' and 'School.Type', calculating school counts and total school roll numbers. The first five rows were removed as they contained missing values for 'Territorial.Authority'. The columns were also renamed and the dataframe was written to a CSV file named 'Final_School_Data.csv'.

D. Population Data (R)

The population data (Excel file) was loaded into R using read_excel() and by specifying the tab containing the data ("Table 4"). As the Excel file contained some formatting, the resulting dataframe contained some unnecessary rows at the beginning and end of the dataset. These were thus removed. Moreover, the columns of the dataframe were assigned more intuitive and descriptive names.

The dataset also contained information for each Auckland local board. Since the primary focus of this project lays on territorial authorities, data pertaining to Auckland local boards was excluded.

Next, an irregularity within the year values was addressed: some entries containing an extraneous "P" were removed for uniformity. Subsequently, the data was filtered using the year column, extracting only the records pertinent to 2022 and only relevant columns (i.e., "Territorial_Authority" and "Total_Population") were selected.

To account for potential character encoding issues, variations in the "Territorial_Authority" column were addressed, whereby characters were standardized to ensure uniformity and accuracy in the dataset.

The last row containing a summary for the entirety of New Zealand was also deleted.

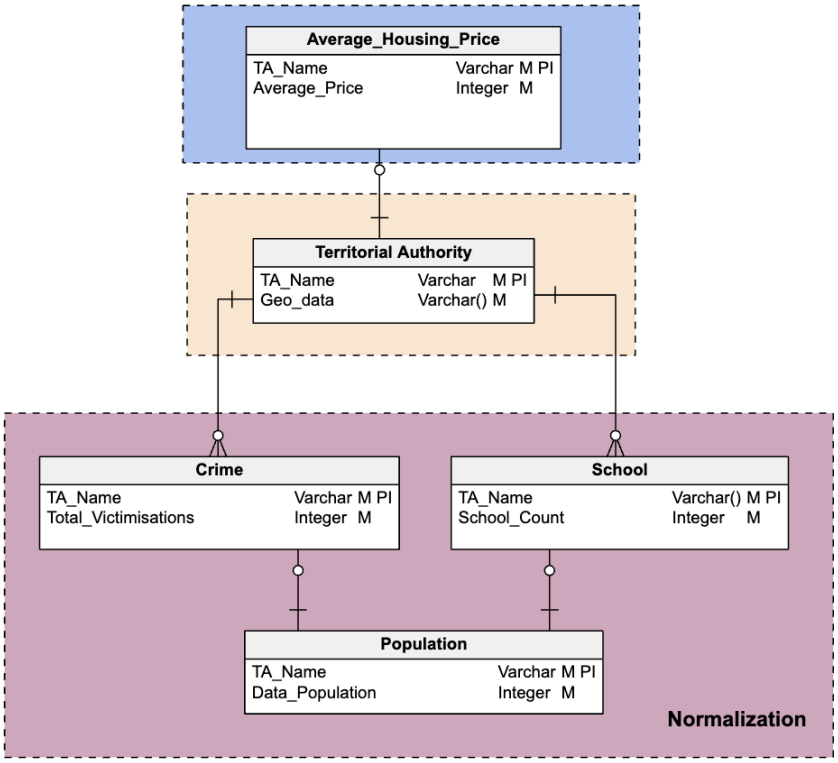
A comprehensive check was conducted to validate the unique value of the "Territorial_Authority" column. This step provided an added layer of assurance regarding the accuracy and consistency of the data.

Finally, the dataframe was written and saved in a CSV format.

E. Normalize the table and join together

Different territorial authorities can exhibit significant variations in population size, with some being notably larger than others. When comparing crime rates and school counts across these territorial authorities, directly using raw counts can lead to unfair comparisons, as larger territorial authorities tend to naturally have higher crime rates and more schools due to their larger populations. Therefore, the victimisations and school counts data were standardized by adjusting them to a common scale (per 10,000 inhabitants) separately, allowing for a more equitable comparison. This adjustment ensures that comparisons are not skewed by population size and enables a fair assessment of relative differences in crime rates and school counts.

Finally, the normalized school data, normalized crime data and house price data were joined together based on the same territorial authority (see Figure 2). Left join was used to ensure that all the territorial authorities were kept even if there was no direct match with school data, crime data or house price data.



(Figure 2: Table Join Process)

F. Mapping Data (R)

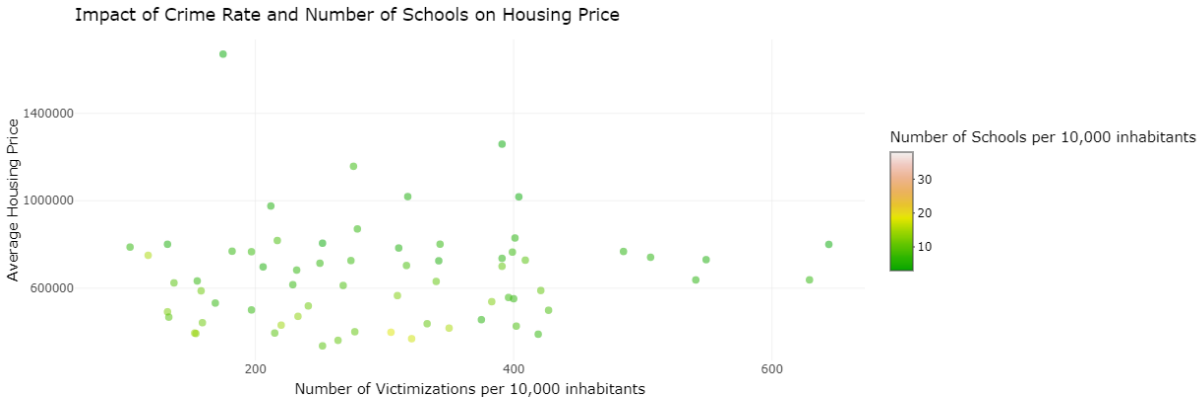
In order to create a map visualization of the data, the territorial authority shapefile was joined with the normalized data table (section V). The geospatial data was first read in R using st_read() from the sf library, and the normalized data was read using read_csv().

The dataframes were then joined on the territorial authority columns, using a left join to ensure that all the geospatial data was kept in (even if there was no corresponding school/crime data).

The resulting dataframe was written and saved as a shapefile.

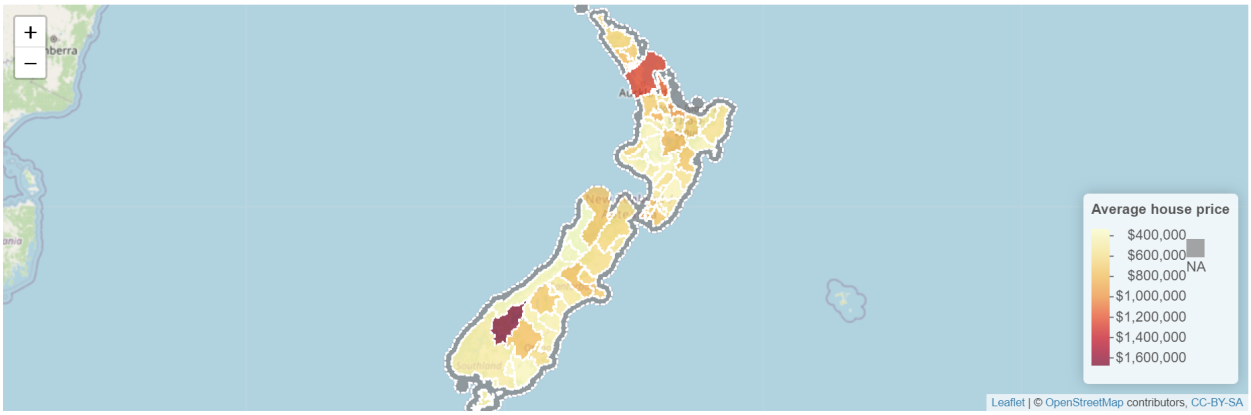
VI. Data Visualization and Insights

Using the mapping data table (section V. F), a scatter plot of the average house price for each territorial authority as a function of both the normalized number of schools and the normalized crime rate was generated (see Figure 3). The plot shows that, based on the collected data, there is no evidence of a correlation between crime rate or number of schools and the average house price.



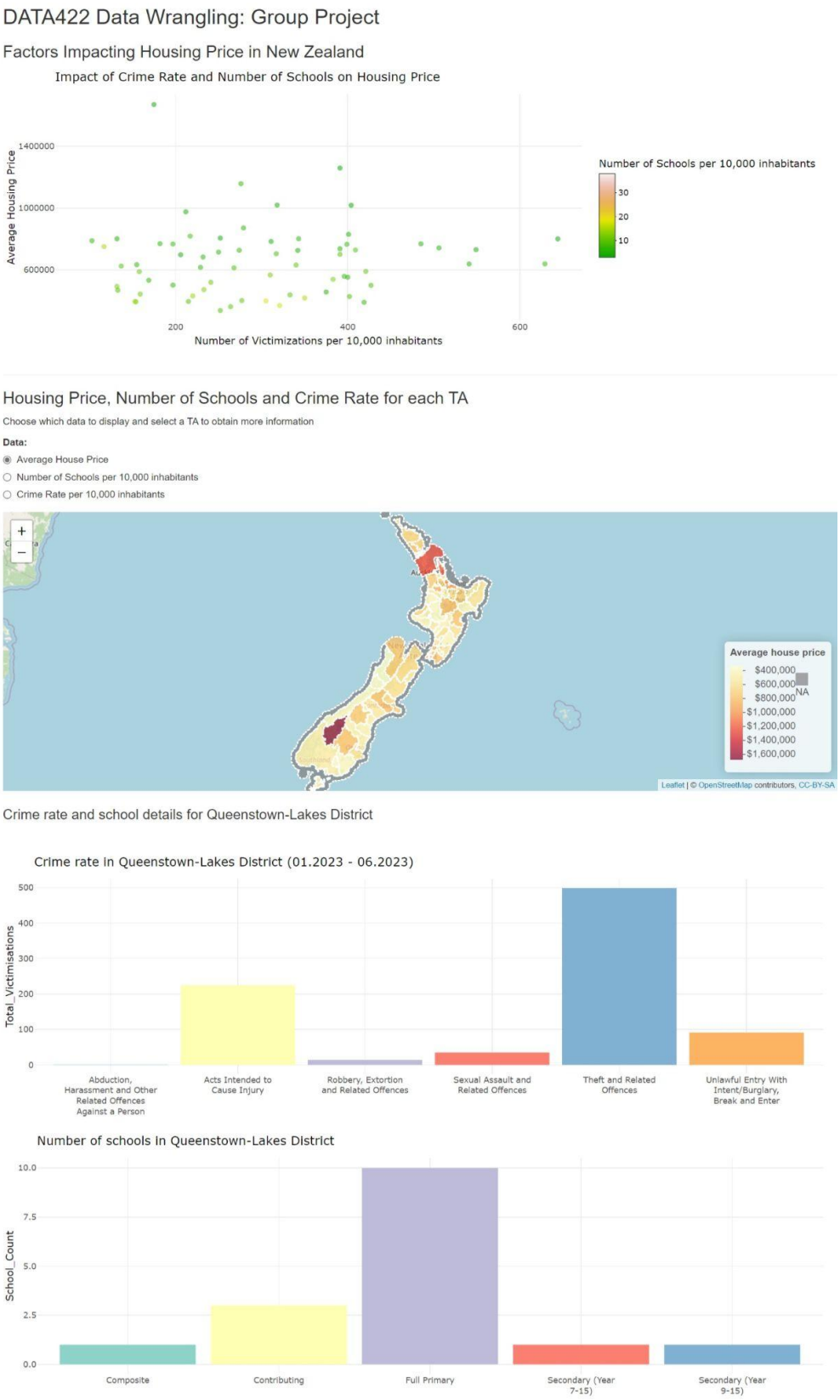
(Figure 3: Scatter plot of the average house price and normalized number of schools and crime rate)

Moreover, choropleth maps displaying the normalized number of schools, the normalized crime rate and the average house price for each territorial authority were created (see Figure 4 for example). The R Shiny dashboard allows the user to choose which data to display (see Figure 5).



(Figure 4: Average house price choropleth map)

In addition, the user can select a specific territorial authority to display further information about the school and crime data (see Figure 5). More specifically, using the cleaned crime and school data, bar charts showing the distribution of the crime types (e.g., theft, sexual assault, etc.) and school types (e.g., primary, secondary, etc.) for the selected region were generated.



(Figure 5: Screenshot of R Shiny application - example for Queenstown-Lakes District)

VII. Challenges

A challenge while trying to scrape housing price data was encountered. Indeed, the required data was within a table on the website, which could not be scraped. To address this, the 'Fetch' panel in the developer tools was employed to monitor the network requests made during the webpage loading process. Through this inspection, it was discovered that the table data was stored in CSV format on the website's backend. Then, the URL associated with the request was located and the HTTP library in Julia was used to perform an HTTP GET request to this specific URL.

Moreover, some challenges when developing the R shiny app were encountered. Specifically, the territorial authority shapefile did not contain the longitude and the latitude of the territorial authorities, but rather variables such as land area. The shapefile thus needed to be transformed using st_transform("+proj=longlat +datum=WGS84") in order for this data to work with Leaflet.

In addition, as some territorial authorities had some missing data, detailed plots of the crime rate or school number could not be generated for them.

VIII. Conclusion

In conclusion, the aim of this project was to investigate the impact of factors such as crime rate and the number of schools on housing prices in Aotearoa. To achieve this, data from 5 different sources was collected and wrangled using various techniques, including scraping, data cleaning, aggregation and joining. An interactive dashboard was then produced using the wrangled and cleaned data to display findings. Overall, based on the data collected, there is no evidence of correlation between crime rate or number of schools and housing price. Further investigation that takes into account temporality and the changes of these factors over time would be useful to confirm this.

IX. References

Juviler, J. (2022). *What Is GitHub?* Retrieved from hubspot: https://blog.hubspot.com/website/what-is-github-used-for

Tidyverse. (n.d.). *R packages for data science*. Retrieved from Tidyverse: https://www.tidyverse.org/packages/