Final Project Report

Crime Rates and Income Levels in British Columbia: Insights from Linear Regression, Classification, and Clustering Techniques

Dewang Marya, Jasjeet Singh, Parvir Gill

https://sc-gitlab.ufv.ca/202305comp381on1parvirgill/projectgroup13.git

Comp 381

Prof. Carl Janzen

19th June, 2023

Executive Summary

The questions that we aimed to answer with this project are:

1) Does higher income (or literacy rate) lead to lower crime?
2) After finding an answer to the above question, based on that we wish to categorize various cities/jurisdictions into low, medium, or high risk.

We aimed to use our chosen datasets (Crime rates in various jurisdictions and Median income) to train the models to answer the above questions. For regression, we chose to use Linear Regression, we aimed to train a dataset and then apply it to a test set to see if there was some correlation. In the case of Classification, we sought to use Logistic Regression to classify whether a municipality had a high crime rate or a low crime rate. Using Logistic Regression, the goal is to train the model on training datasets and test it on one, and we would set a threshold that would give us a binary responding variable as to whether or not a municipality has a high crime rate. Finally, for Clustering, the goal is to define and characterize occurrences of low/medium/high crime rates. We chose to use K-means to cluster the occurrences, and then plot them to see which municipalities are characterized as one of three types of crime rates.

We chose not to deviate too much from our original proposal, though we abandoned the idea of including literacy rate as a variable.

Findings, Approach, and Deviations

We have discovered that regions have a higher chance of higher crime rates with lower median family income. We have also clustered the regions into low, medium, and high danger(crime). We also discovered that no matter the danger level in a particular region the crime rates have risen over the 5 year period for all the regions.

We also chose to deviate away from the original proposal of including literacy rate as a variable. We deviated from it due to the lack of usable data within the scope of our project, we were unable to implement literacy rate as one of the variables.

Crime Data from the government of British Columbia and income data from Statistics Canada have been used to perform regression, classification, and clustering with data from 2015 to 2018 as the training set and data from 2019 as the test set. These models can be used to predict crime rates for regions across British Columbia, classify the regions into low and high crime rates, and cluster the regions together based on how dangerous they are.

Data Sets

The income dataset(Statistics Canada, 2022) provides us with the median family income based on different cities in British Columbia. This data is also recorded annually, however, some new regions were added to it only after 2015, hence the limit was set to 2015 for all regions for preciseness in the answer for all regions. The online table was filtered to only show the data for different cities in BC and the median family income for those cities was then used as a median income for the policing jurisdiction, since there were more cities in the list than the jurisdictions, new medians were formed for such jurisdictions by adding the incomes (using median addition formula).

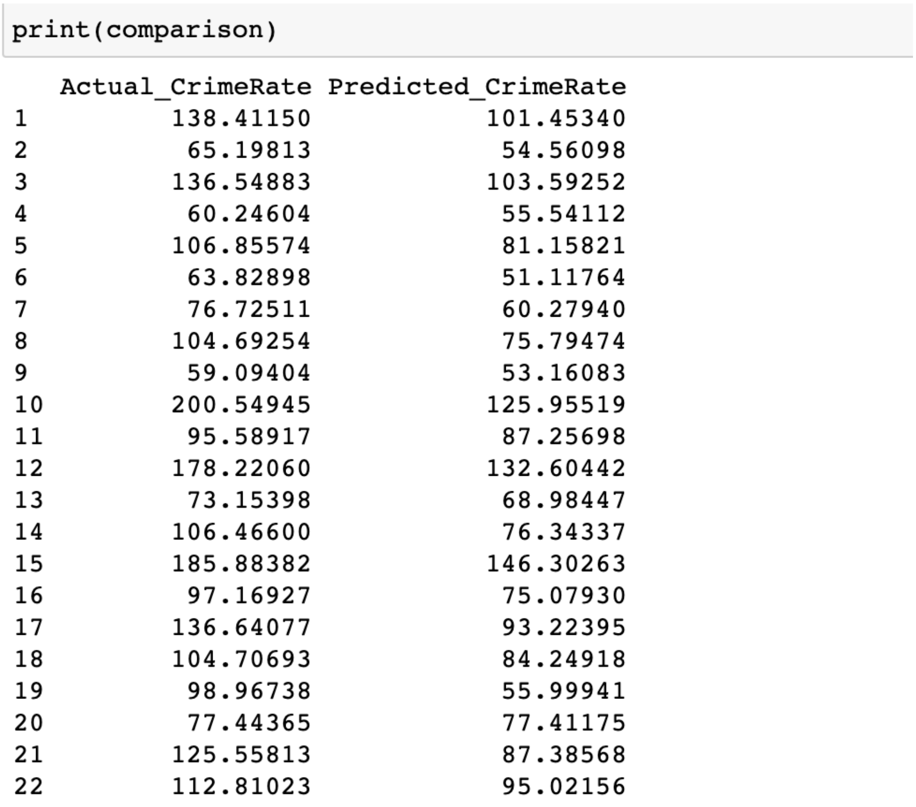
The crime rate dataset(Gov. BC, 2021) contains crime statistics and trends for regional districts in British Columbia from the years 2012 through 2021. Due to inconsistency in the median income data set, the crime rate data for years before 2015 was dropped. The data had various types of crimes and their frequencies, rates, and other relevant measurements over this time period. The data we specifically used out of these was the total criminal code offenses reported in the jurisdiction per 1000 people in that jurisdiction excluding the traffic violations.

Government of British Columbia. (2021). Columbia Regional District Crime Trends, 2012–2021. https://www2.gov.bc.ca/gov/content/justice/criminal-justice/policing-in-bc/publications-statistics-legislation/crime-police-resource-statistics#regional

Statistics Canada. (2022). Selected income characteristics of census families by family type (Table 11-10-0009-01). Annual Income Estimates for Census Families and Individuals (T1 Family File). https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1110000901

Models

Regression



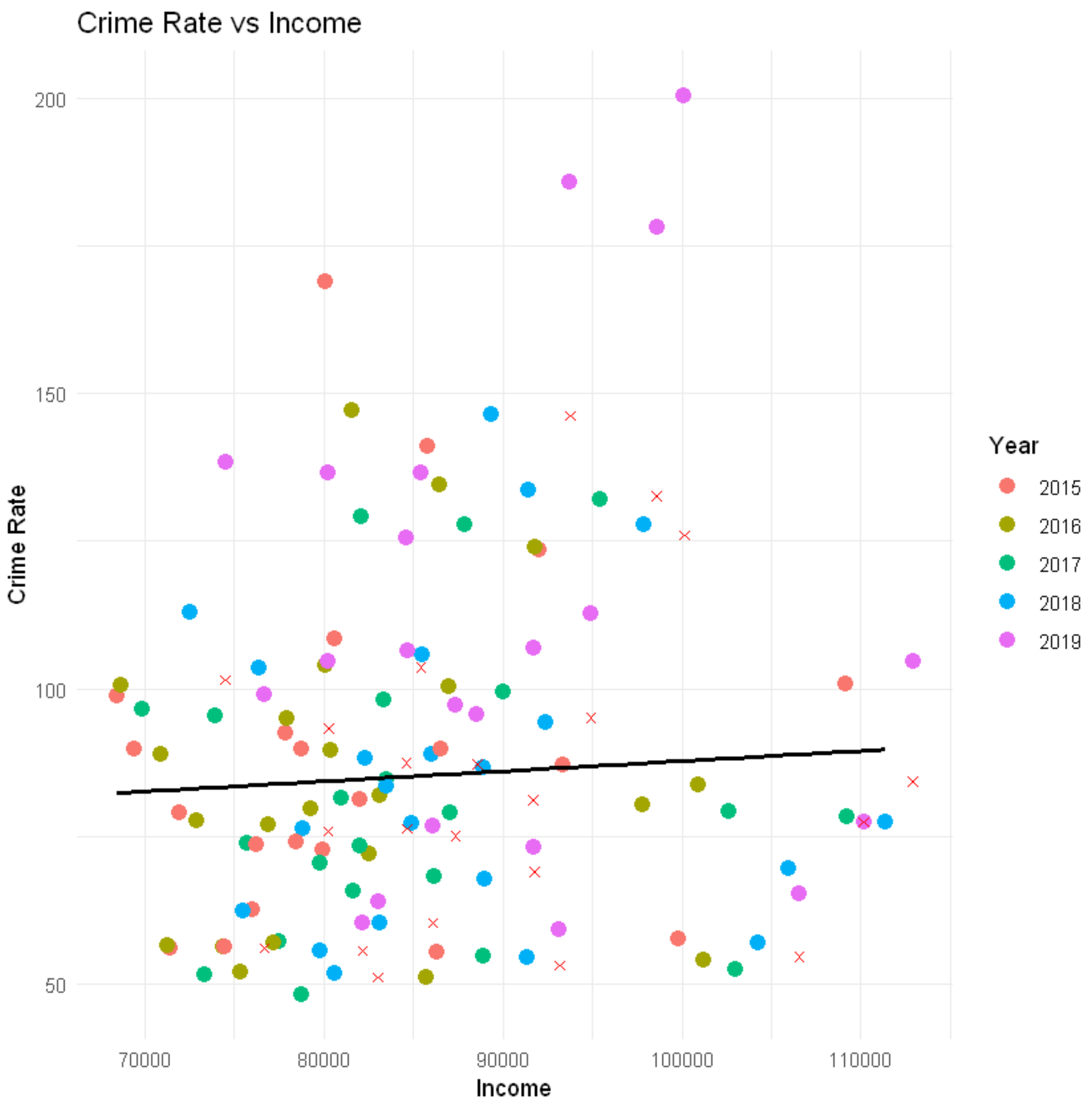
```
print(comparison)

   Actual_CrimeRate Predicted_CrimeRate
1         138.41150           101.45340
2          65.19813            54.56098
3         136.54883           103.59252
4          60.24604            55.54112
5         106.85574            81.15821
6          63.82898            51.11764
7          76.72511            60.27940
8         104.69254            75.79474
9          59.09404            53.16083
10        200.54945           125.95519
11         95.58917            87.25698
12        178.22060           132.60442
13         73.15398            68.98447
14        106.46600            76.34337
15        185.88382           146.30263
16         97.16927            75.07930
17        136.64077            93.22395
18        104.70693            84.24918
19         98.96738            55.99941
20         77.44365            77.41175
21        125.55813            87.38568
22        112.81023            95.02156
```

A linear regression model was used to predict the crime rates for the year 2019, a training set was used from the years 2015-1018 and the year 2019 was predicted, to make it interesting, we can see that the above predictions were compared to the actual rates. This shows some regions like number 20 were very closely predicted. We can see none of the predicted values was an overshoot, the gap in the prediction and actual numbers is big in some cases, but that could be due to various other factors, increased gang activity, and population increase due to immigration, and should be looked at further.

The predicted values can be seen represented by crosses in the image below. Comparing them to the pink values which are the actuarial crime rates, we can see that none of the values was an overshoot of the actual crime rate observed.

This model can be used to predict crime rates in various other regions of BC. In this model, income was considered as an independent variable and crime rate was a dependent variable.

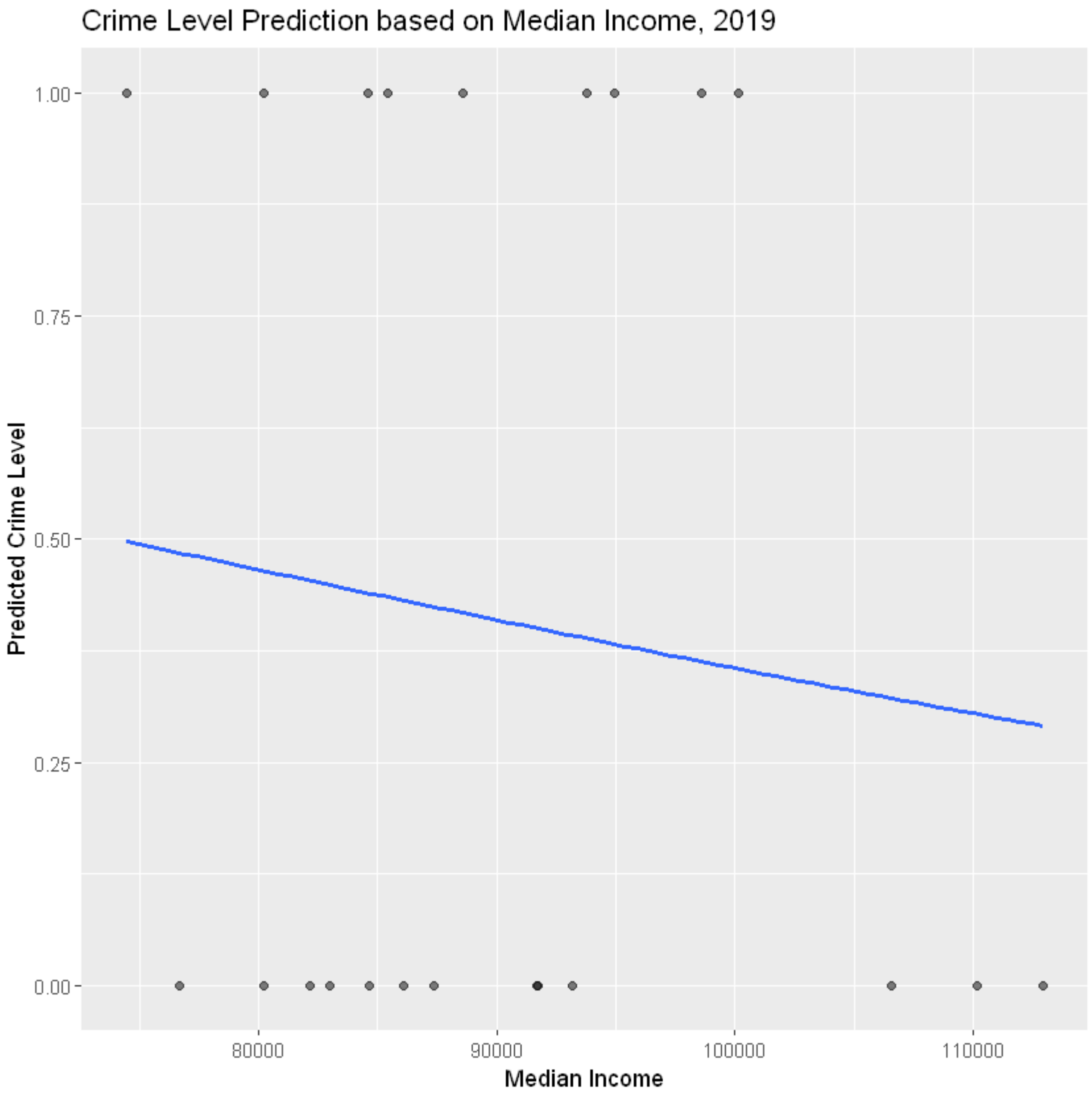
In order to achieve the outcome we have in regression we had to learn to wrangle the data, especially pivoting to change the data into long format to apply the linear model to the data which took some time.



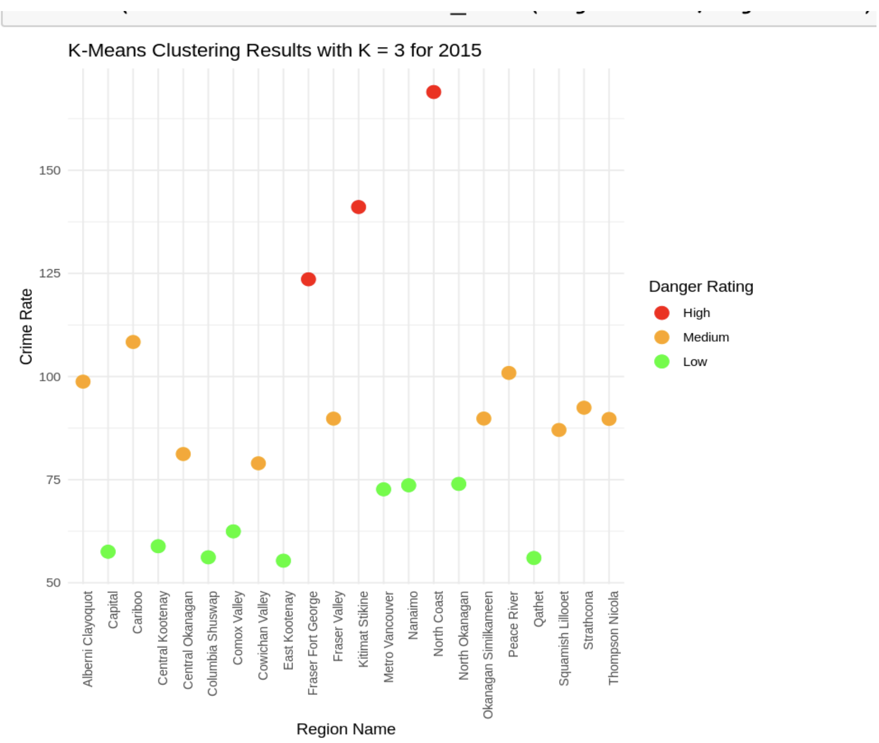
Classification

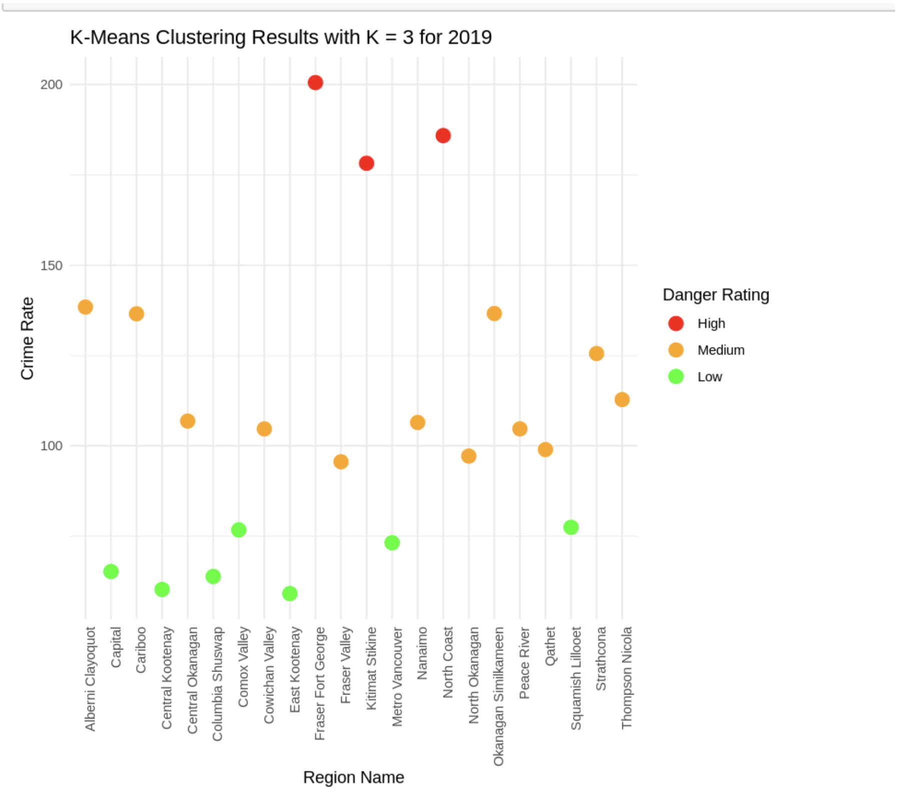
Classification was used within our project to identify a binary response of whether or not a region has a high crime rate or not dependent on the region’s median income. A thresh hold was set to identify what constitutes a high or low crime rate, in this case, the thresh hold set was the median crime rate across all regions. A responding variable was created which would indicate where 1 is a high crime rate, and 0 is a low crime rate. We used Logistic Regression to get our result. The Logistic Regression Model was trained on the dataset from the years 2015 to 2018, to provide a diverse range of data points which would increase the model’s accuracy. Once the model was trained, we picked 2019 as the testing set, as that is the last year of accurate data before the COVID-19 pandemic (which perhaps would result in anomalies). Using the test set results, we compared the actual categorized crime type to the predicted crime type. In the end, we plotted a graph that displays the correlation between median income and the predicted crime type for each region.

While completing Logistic Regression for Classification, we had to research how to accurately portray our result visually. Simplistic bar or line graphs would not have been applicable to our results, thus we had to incorporate a scatter plot (indicate region pointers) with a line graph (portray the slope of the correlation between median income and crime rate).


Crime Level Prediction based on Median Income, 2019
Predicted Crime Level
1.00
0.75
0.50
0.25
0.00
80000
90000
100000
110000
Median Income

Clustering


K-Means Clustering Results with K = 3 for 2015
Crime Rate
150
125
100
75
50
Region Name
Alberni Clayoquot
Capital
Cariboo
Central Kootenay
Central Okanagan
Columbia Shuswap
Comox Valley
Cowichan Valley
East Kootenay
Fraser Fort George
Fraser Valley
Kitimat Stikine
Metro Vancouver
Nanaimo
North Coast
North Okanagan
Okanagan Similkameen
Peace River
Qathet
Squamish Lillooet
Strathcona
Thompson Nicola
Danger Rating
High
Medium
Low


K-Means Clustering Results with K = 3 for 2019
Crime Rate
200
150
100
Region Name
Alberni Clayoquot
Capital
Cariboo
Central Kootenay
Central Okanagan
Columbia Shuswap
Comox Valley
Cowichan Valley
East Kootenay
Fraser Fort George
Fraser Valley
Kitimat Stikine
Metro Vancouver
Nanaimo
North Coast
North Okanagan
Okanagan Similkameen
Peace River
Qathet
Squamish Lillooet
Strathcona
Thompson Nicola
Danger Rating
High
Medium
Low

Clustering was used in our project to depict the crime trends in different policing jurisdictions in British Columbia, a GGPLOT was used to compare the crime rate of the 22 regions for the years 2015 and 2019 to see where the stats are headed and compare it to our other models. To achieve this, we used the K -means clustering method, and hit and the trial was used to select 'K', initially a value of 2 was selected, but it was later decided a value of 3 would best fit our case to classify them into 3 danger ratings, low, medium and high crime jurisdictions.

From clustering, it was found that there was a general increase in the level of crime. From this cluster, you can see that the average crime rate for low, medium, and high in 2015 was 63, 91.7, and 145, while in 2019 it went up to 68, 114, and 188. Also, the group population of the medium crime rate group went up by 2 in 2019 as compared to 2015, that shift was from the low-income group to the medium-income group, justifying that the average crime rate was on the increase in those years.

Initially to depict clustering, a simple plot was used, it was not much descriptive, then looking at some examples online, it was decided to use GGPLOT, so the GGPLOT library was researched up by the group members to visualize the output properly.

Ethical Concerns

1) Bias

   During the collection of our data, we faced the dilemma of gathering results that would skew the results. Considering being residents of the Lower Mainland and having inner knowings of the dynamics of the municipalities, we avoided choosing cities and rather focused on regions. For us, there is some common knowledge that indicates that there are factors other than median income that affects crime rates, so we decided to instead use regional data so these biases were not present and we could just have the averages of the region itself.

2) Data Accuracy

While finding the data for our project, it was important that we find accurate and reliable datasets that are trusted. Initially, when we started searching from datasets, we couldn't find anything from the BC website so we decided to search other databases, but we came across a lot of user-submitted data that had not been verified or was unreliable. To avoid using these datasets, we returned back to the BC database, narrowed down results, and spent more time locating datasets that were reliable but also applicable to our questions within the context of the project.

Conclusion

We found that the crime rate decreases with median family income in the different regions of British Columbia. We also discovered that over the five-year period, the crime rate has increased across all regions. We also have been able to cluster regions together on the basis of low, medium, and high danger (Crime Rates).

These findings imply that there were more trouble-causing elements in British Columbia in 2019 than there were in 2014 (increasing every year). These could be a product of the income disparity and/or leading people to negative activities in order to survive.

We had limitations regarding the regions we could look into since the jurisdictions for the crime data and the regions for the income data were differently divided.

We suggest that further analysis should be done into how poverty rates and education might affect the crime rate in different regions as these might also be effective predictors.

Appendix

Government of British Columbia. (2021). Columbia Regional District Crime Trends, 2012–2021. https://www2.gov.bc.ca/gov/content/justice/criminal-justice/policing-in-bc/publications-statistics-legislation/crime-police-resource-statistics#regional

Statistics Canada. (2022). Selected income characteristics of census families by family type (Table 11-10-0009-01). Annual Income Estimates for Census Families and Individuals (T1 Family File). https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1110000901

Statology (n.d.). How to Use pivot_longer().
How to Use pivot_longer() in R - Statology

Tutorialspoint. (n.d.). ggplot2 - Quick Guide. Retrieved from https://www.tutorialspoint.com/ggplot2/index.htm