DSCI 310: Predicting Canada's Community Well-Being Index Scores

Shawn Li, Selena Shew, Sri Chaitanya Bonthula, Lesley Mai

Table of contents

Introduction:

Across Canada, we are able to evaluate the socio-economic health of communities by using the Community Well-Being (CWB) Index, which provides a unique perspective through which we may do so[1]. The Community Well-Being Index (CWB Index) is a comprehensive score that serves to reflect the overall well-being of each community[2]. It does this by incorporating factors such as education, income, labour force activity, and housing values. In addition to facilitating comparisons between communities of First Nations and Inuit, these scores also allow for comparisons to be made against a more general Canadian context.

Our dataset originates from Indigenous Services Canada. **How do these socio-economic factors influence CWB scores, and what can they reveal about community health disparities?** This is the question that our research investigates as we investigate the predictive power of these factors. The purpose of this project is to disentangle the intricate connections that exist between these indicators and the well-being of the community by making use of data obtained from the Census of Population conducted by Statistics Canada. The findings of this project could provide valuable insights that could be used to guide policy and intervention strategies.

Table 1: Full data in CWB 2021

```
# A tibble: 6 x 9
  CSD_Code_2021 CSD_Name_2021  Census_Population_2021 Income_2021 Education_2021
          <dbl> <chr>                           <dbl>       <dbl>           <dbl>
1       1001113 Trepassey                         405          76              48
2       1001124 Division No. ~                   1373          81              63
3       1001126 Cape Broyle                       499          78              57
4       1001131 Renews-Cappah~                    280          79              59
5       1001149 Ferryland                         371          83              56
6       1001155 Division No. ~                    469          78              58
# i 4 more variables: Housing_2021 <dbl>, Labour_Force_Activity_2021 <dbl>,
#   CWB_2021 <dbl>, Community_Type_2021 <chr>
```

Methods

The Python programming language (R Core Team 2023) and the following R packages were used to perform the analysis: (Hadley Wickham and al. 2023), (Wickham 2023), (Barret Schloerke 2023a), (Barret Schloerke 2023b), (Wickham et al. 2023), (Kuhn and Vaughan 2023), as well as the following data from official websites of Canadian government and departments: (Government of Canada 2024), (National Collaborating Centre for Indigenous Health 2016), (Government of Canada 2019), (Statistics Canada 2021)

Data

We have stored the dataset into a csv file, and import the file into Jupyter Notebook for analysis. We directly used the function `read_csv` to view the csv file downloaded.

Some of the observations are not including certain information, therefore, we use function `filter` to exclude rows with NA values in key columns beforing actually analyzing and obtain a clean data frame, ensuring the dataset's integrity for analysis.

Here is the first few rows of our data:

Data Preprocessing

The data was first split into training and testing sets so that the model could be validated. Then we made histograms and summary statistics to see how 4 variables are spread out. These insights help choose the features of the model.

Here is the head of our training data:

Table 2: CWB_2021 Training Dataset

```
# A tibble: 6 x 9
  CSD_Code_2021 CSD_Name_2021  Census_Population_2021 Income_2021 Education_2021
          <dbl> <chr>                           <dbl>       <dbl>          <dbl>
1       4602032 De Salaberry                     3918          74             57
2       4605032 Boissevain-Mo~                   2309          77             62
3       3521024 Caledon                         76581          85             69
4       4717062 Beaver River ~                   1277          75             56
5       1310011 Canterbury                        552          80             60
6       2428015 Sainte-Aur lie                    856          81             50
# i 4 more variables: Housing_2021 <dbl>, Labour_Force_Activity_2021 <dbl>,
#   CWB_2021 <dbl>, Community_Type_2021 <chr>
```

Table 3: CWB_2021 Testing Dataset

```
# A tibble: 6 x 9
  CSD_Code_2021 CSD_Name_2021  Census_Population_2021 Income_2021 Education_2021
          <dbl> <chr>                           <dbl>       <dbl>          <dbl>
1       1001197 Mount Carmel-~                    382          81             61
2       1001308 Whiteway                          351          79             59
3       1001325 Heart's Conte~                    330          72             52
4       1001332 Winterton                         436          74             48
5       1001339 Division No. ~                   1673          73             54
6       1001374 Division No. ~                    320          84             72
# i 4 more variables: Housing_2021 <dbl>, Labour_Force_Activity_2021 <dbl>,
#   CWB_2021 <dbl>, Community_Type_2021 <chr>
```

Here is the head of our testing data:

Visualization of Data

Overall Data

Now the first several rows of the data has been shown, and it seems well-organized and tidy. Since the goal is to predict the CWB index by 4 predictors below, we need to have a glance at each of the variables for a deeper insight.

- Income: The average income level in 2021. Higher income levels are typically associated with better access to resources and may positively influence the CWB score.

- Education: The average education level attained. Education is often linked to better socioeconomic outcomes and may be a strong predictor of CWB.
- Housing: The average housing quality or status. Adequate housing is a critical component of well-being, affecting both physical and mental health.
- Labour force: The level of employment or labor force participation. Higher labor force activity might correlate with higher economic productivity and, consequently, higher CWB scores.

First of all, we create a plot for the entire database, this plot involves all 4 predictors and the response variable CWB index. Specifically, we used multiple lines in the plot to indicate the counts of different variables in different levels (or values). Each line of different colour represents one of the variable in the data. By looking at the plot, we can easily find potential relationships between predictors and the response variable, as well as what does the entire data look like.

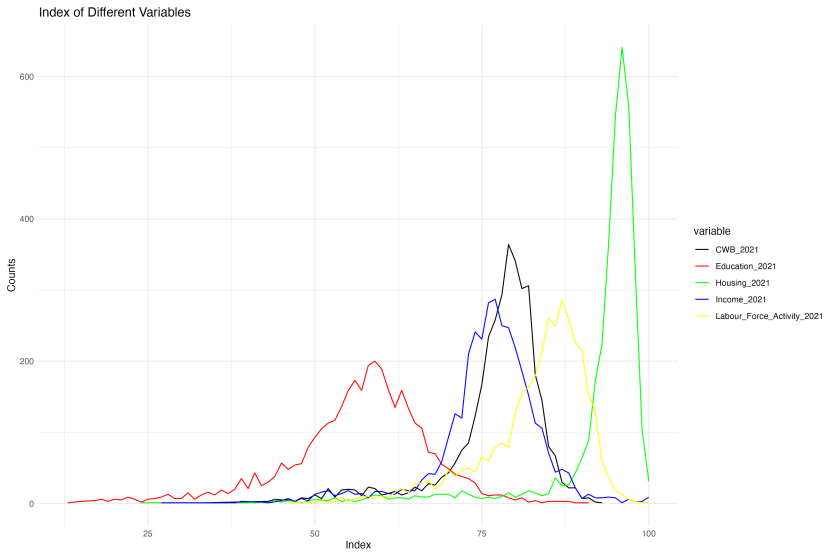

Figure 1: Index of Different Variables for the entire dataset of CWB_2021

By looking at Figure 1, we see that all the predictors and the response variable forms a normal distribution, with mean between 50 - 100. Since the distribution of CWB index is right in the middle of all other variable, it might be possible that CWB index is actually the mean of all other variables.

Viewing the Mean

Below, we looked at the mean value of the four predictors and the response variable.

Table 4: The mean value of the four predicting variables (house, education, labour, income) and the response variable (CWB Index)

```
# A tibble: 1 x 5
  Income_2021 Education_2021 Housing_2021 Labour_Force_Activity_2021 CWB_2021
        <dbl>          <dbl>        <dbl>                      <dbl>    <dbl>
1        75.9           56.5         92.4                       83.0     77.0
```

Difference between Different Communities

Since we have neglected the type of communities in our predictors, it is essential to take them into account. By splitting the data into different parts, each have one type of community, we generated three plots of counts of different variables in different communities just like what we did in the previous plot. Then we can make comparison between communities. Here is the plots for all three communities:

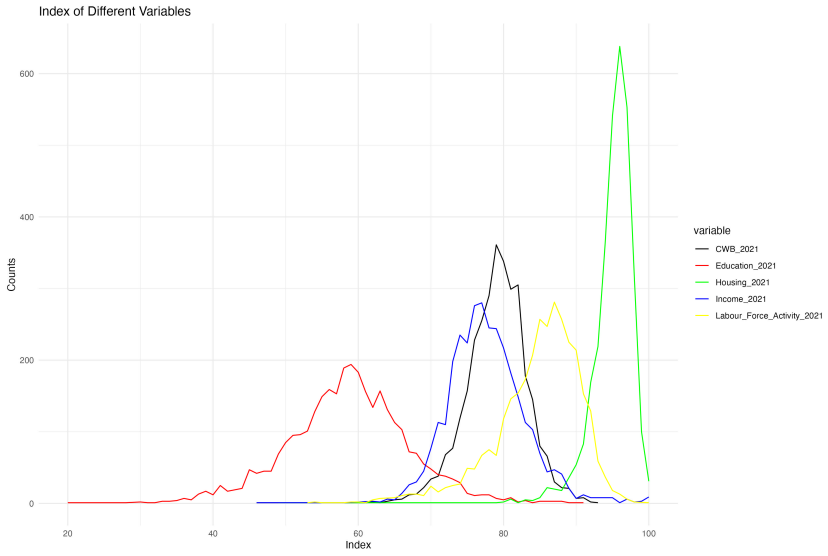

Figure 2: Index of Different Variables for Non-Indigenous Community

It seems like the quality of first nation people' lives from Figure 3 is a bit lower than non-indigenous people's reflected by Figure 2 since the distribution of variables in Figure 3 is shows a lower mean. We did not clearly see the distribution of data in Figure 4 since there are only a few data points, all we see is their housing index is very high.

Since there are only a few data points in first nation and inuit communities, even if there is a small difference between different communities, we can nearly neglect the difference when creating the predicting model.

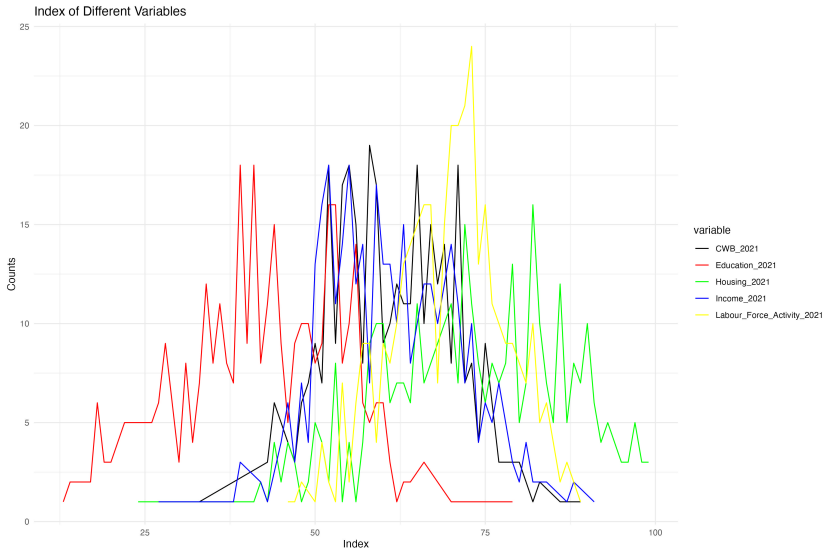

Figure 3: Index of Different Variables for First Nations Community

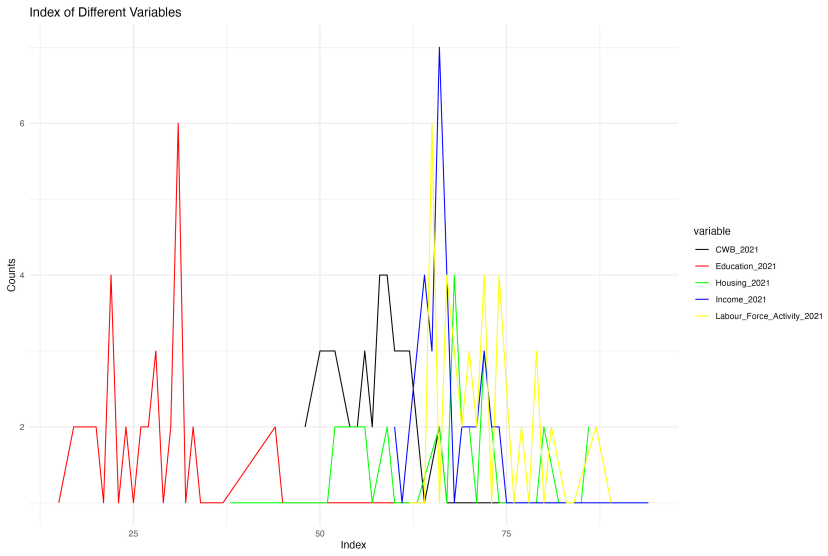

Figure 4: Index of Different Variables for Inuit Community

General Relationship

We use Barret Schloerke (2023a) from the Barret Schloerke (2023b) package to see how key variables are distributed and how they are related to each other. This gives us a full picture of possible linear correlations and points out any outliers or strange patterns.

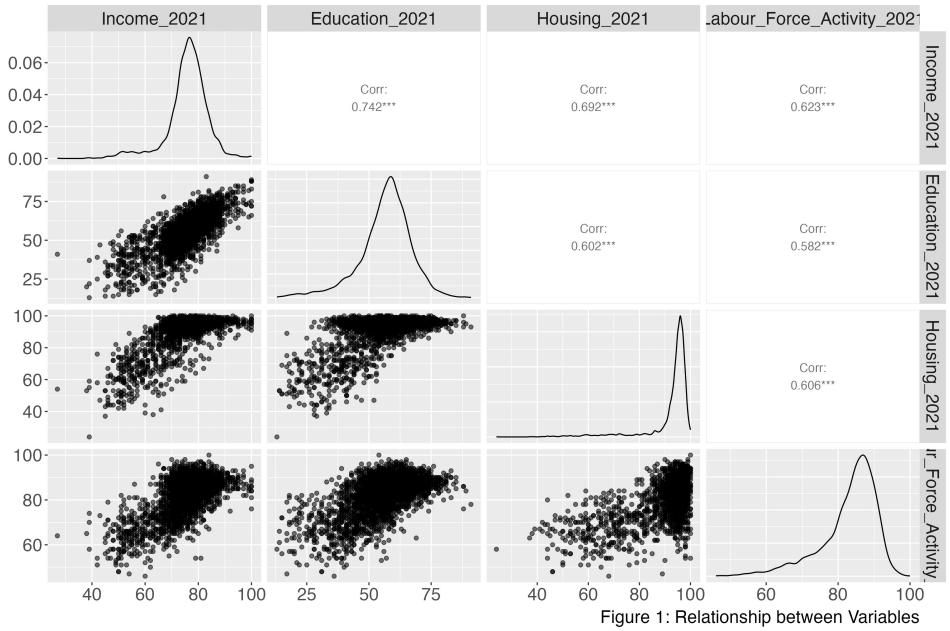

Figure 5: Correlations between different Variables

Observation: Figure 5 show that all 4 predictors have strong linear relationships with the CWB score, and the correlation coefficients show that these relationships are very strong.

Analysis

In order to avoid the weakness of KNN Regression, a linear regression model from DSCI100 is used. The workflow is to build up a formula, or a linear model trained from the training dataset, and use the trained model to test its prediction on testing data.

Thus, we proceeded with building our linear Regression model.

- We first specified our linear regression model using the `lm()`, capitalizing on its simplicity and robustness for our type of data. Then we define the recipe with 4 chosen variables.
- Next, fitted the model to the training data using `fit()`, which allowed us to estimate the relationship between our predictors and the CWB score.

Table 5: Linear Predicting Model

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	0.0597	0.0787	0.758	0.448
Income_2021	0.251	0.00136	184.	0
Education_2021	0.249	0.000922	271.	0
Housing_2021	0.250	0.00100	249.	0
Labour_Force_Activity_2021	0.249	0.00110	227.	0

Table 6: Performance of the Linear Predicting Model (Evaluation based on RMSE, RSQ, MAE)

.metric	.estimator	.estimate
<chr>	<chr>	<dbl>
rmse	standard	0.321
rsq	standard	0.998
mae	standard	0.260

- We employed the `predict` on the test set to assess the model's predictive accuracy, followed by calculating performance metrics such as RMSE and R-squared, which indicate how well our model generalizes to new data.
- Then, interpreted the model's coefficients to understand the impact of each predictor on the CWB score. We know that a positive coefficient suggests a direct relationship, whereas a negative coefficient indicates an inverse relationship. Finally, reviewed diagnostic plots and validation metrics to ensure the assumptions of linear regression are met.

Here is the linear model based on the training data:

Here is the performance of the model:

With the established linear regression model, our prediction is significantly improved, and we can now clearly see what our model looks like.

The expression of our model is:

$$Community\ Well\text{-}Being\ Index = 0.06391 + 0.25011 * Income + 0.24988 * Education + 0.24929 * Housing + 0.25001 * Labour\ Force\ Activity$$

Here is an interesting observation from Table 5:

We found that indeed, CWB index is just the average of all other variables.

Conclusion

As a result of our investigation, we discovered that socio-economic factors have a significant influence on the Community Well-Being Index scores received by Canadian communities. It has been determined that the most important factors that determine the well-being of a community are income, education, housing, and participation in the labour force. The fact that we were able to improve our predictive accuracy by switching from KNN to linear regression modelling is evidence of the complexity of the relationships between these variables. This research lays the groundwork for the development of specific interventions that are aimed at enhancing the well-being of the community.

We hypothesized that socio-economic indicators have a significant influence on the well-being of communities, and the findings supported this hypothesis. It is clear from this validation that income, education, housing, and employment all play a significant role in determining the health of a community. The findings of our study have a wide range of implications, and they suggest that strategically targeted policy interventions in these areas could significantly improve the CWB scores, particularly in communities that are economically disadvantaged[3]. With such insights, policymakers and community planners have the potential to be guided in the development of more effective strategies for enhancing the well-being of communities across a wide range of landscapes in Canada[4].

Future exploration: 1. How might other, less traditionally quantified socio-economic factors, such as community engagement, access to technology, and environmental sustainability, contribute to the overall well-being of Canadian communities? 2. What role do technological advancements and digital access play in shaping the CWB scores, particularly in remote or underserved communities?

Reference

Barret Schloerke, Di Cook, Jason Crowley. 2023a. *GGally: Extension to 'Ggplot2'*. https://CRAN.R-project.org/package=GGally.

———. 2023b. *GGally: Extension to 'Ggplot2'*. https://CRAN.R-project.org/package=GGally.

Government of Canada. 2019. "Five-Year Evaluation Plan 2019-2020 to 2023-2024." https://www.sac-isc.gc.ca/eng/1570114159516/1570114187006.

———. 2024. "About the Community Well-Being Index." https://www.sac-isc.gc.ca/eng/1421245446858/1557321415997.

Hadley Wickham, Jim Hester, and Romain Francois et al. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Kuhn, Max, and Davis Vaughan. 2023. *Rsample: General Resampling Infrastructure*. https://CRAN.R-project.org/package=rsample.

National Collaborating Centre for Indigenous Health. 2016. "NCCIH - National Collaborating Centre for Indigenous Health > Home > NCCIH PUBLICATIONS." https://www.nccih.ca/495/Health_inequalities_and_the_social_determinants_of_Aboriginal_peoples___health_.nccih?id=46.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Statistics Canada. 2021. "Moving Forward on Well-Being (Quality of Life) Measures in Canada." https://www150.statcan.gc.ca/n1/pub/11f0019m/11f0019m2021006-eng.html.

Wickham, Hadley. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.