Exploring Income Inequality Dynamics in Canada: A Geospatial and Temporal Analysis*

Insights from 1976 to 2021

Sirui Tan

April 6, 2024

This paper investigates income inequality in Canada from 1976 to 2021 using data from various surveys. Through exploratory data analysis and regression modeling, we uncover increasing income inequality trends over time and variations across different provinces. Our findings reveal nuanced relationships between temporal trends and regional disparities, and income distribution dynamics. Through rigorous analysis, we uncover two key findings: a consistent rise in income inequality over time, significant regional disparities across provinces. Understanding these dynamics is important for policymakers to develop targeted strategies aimed at mitigating income inequality and promoting social and economic equity in Canada.

Table of contents

*Code and data are available at: https://github.com/siru1366/income.git.

1 Introduction

Income inequality is a notable socioeconomic issue that has gained attention worldwide, including in Canada. Over recent decades, Canada has seen a widening gap in per capita after-tax income, raising concerns about economic disparities and social justice. Understanding the drivers behind this trend is important for policymakers and researchers, providing insights into the factors influencing income distribution and informing strategies to mitigate inequality.

This paper aims to investigate the intricate dynamics of income inequality within Canada, focusing on the period from 1976 to 2021. Using a wide-ranging dataset from surveys like the Survey of Consumer Finances (SCF), the Survey of Labour and Income Dynamics (SLID), and the Canadian Income Survey (CIS), we thoroughly investigate income trends across provinces and over time. By examining key variables such as year, geographical location, income decile, and income range, we seek to uncover patterns and nuances in income distribution dynamics.

Despite numerous studies on income inequality in Canada, there remains a clear gap in understanding the specific drivers and mechanisms contributing to the widening income gap observed over the past four decades. To address this gap, we undertake a multifaceted analysis, encompassing exploratory data analysis and multiple linear regression modeling. Through this approach, we aim to elucidate the role of various factors, including temporal trends, regional disparities, in shaping income distribution patterns.

Temporal trends reveal a consistent increase in income inequality from 1976 to 2021. Geographical disparities highlight varying levels of income inequality across provinces, with Atlantic provinces and Quebec showing higher inequality compared to others.Overall, these findings emphasize the multifaceted nature of income inequality dynamics in Canada, urging policymakers to consider both temporal trends and regional disparities when addressing this pressing socioeconomic issue.

The remainder of this paper is structured as follows. Section 2 introduces the data used for analysis and findings, including visualizations of the variables of interest, Section 3 proposes a straightforward linear regression model to examine and forecasts the effects of time (year) and geographical location (province) on income inequality in Canada. In Section 4, we display the interpretations of the models alongside other findings from analyzing the data. Section 5 provides a discussion on the implications of the findings as well as the weaknesses of this paper and its next steps for further study on this subject.

1.1 Estimand

The estimand of this study is to quantify the relationship between time (year) and income inequality in Canada. Specifically, we aim to estimate the effect of each unit change in the year on income inequality, holding other variables constant. This involves analyzing how income inequality trends have evolved over the period from 1976 to 2021, capturing any systematic changes over time.

Additionally, we seek to understand the impact of geographical location (province) on income inequality and how this interacts with temporal trends. By examining the coefficients associated with both year and province variables in our regression models, we aim to discern the average differences in income inequality between different provinces, while also exploring how these differences may have changed over time.

Overall, our estimand focuses on uncovering the nuanced relationship between time, geographical location, and income inequality in Canada.

2 Data

2.1 Data Source

According to a report by Statistics Canada (2023), titled *Upper income limit, income share and average income by economic family type and income decile*,the data provided valuable insights into income distribution patterns in Canada.

2.2 Data Measurment

The estimates for *Upper income limit, income share and average income by economic family type and income decile* are based on data from various surveys spanning different periods. From 1976 to 1992, data was collected from the Survey of Consumer Finances (SCF). For the period from 1993 to 1997, a combination of SCF and the Survey of Labour and Income Dynamics (SLID) was used. Subsequently, from 1998 to 2011, data solely from SLID was utilized. Starting from 2012, the Canadian Income Survey (CIS) became the primary data source. More details

on these surveys and their revisions can be found in publications by Statistics Canada, including "Revisions to 2006 to 2011 Income Data" (2015) by Statistics Canada and other related papers by Cotton (2000) and Lathe (2005).

The Canadian Income Survey (CIS) aims to provide insights into the income and its sources of Canadians, alongside their individual and household characteristics. It combines data from the Labour Force Survey (LFS) and tax records.

Estimates from the Survey of Consumer Finances cover individuals aged 15 years and over, whereas estimates from SLID and CIS include individuals aged 16 years and over.

The CIS introduced improvements in methodology and data processing, notably from the 2021 reference year onwards. It transitioned to using the Administrative Personal Income Masterfile, incorporating data from both T1 tax returns and associated tax slips. Prior to this, only T1 tax returns were used. These enhancements, including updates to weighting methodology, aim to enhance data quality while minimally impacting key estimates and trends.

The CIS is a sample survey with a cross-sectional design, administered to a sub-sample of LFS respondents. The LFS employs a rotating panel sample design, with selected dwellings remaining in the sample for six consecutive months in the provinces and for two years in the territories. Rotation groups from the LFS are utilized for the CIS sample, with approximately 55,000 households included in the CIS sample for 2021.

Data cleaning and analysis were conducted using the open-source statistical programming language `R` (R Core Team 2023), utilizing functionalities from the `tidyverse` (Wickham et al. 2019), `ggplot2` (Wickham 2016), `dplyr` (Wickham et al. 2023), `readr` (Wickham, Hester, and Bryan 2024), `tibble` (Müller and Wickham 2023),`stringr` (Wickham 2023), `haven` (Wickham, Miller, and Smith 2023), `janitor` (Firke 2023), `knitr` (Xie 2023).

2.3 Variables of Interest

Table 1: Extracting the first ten rows from the Income data

Year	Geographical location	Income decile	Income	Highest-to-Lowest Average Income Ratio	income range
1976	Canada	Total deciles	63300	16.8	148100
1976	Canada	Lowest decile	9400	NA	NA
1976	Canada	Second decile	21300	NA	NA
1976	Canada	Third decile	31500	NA	NA
1976	Canada	Fourth decile	41500	NA	NA
1976	Canada	Fifth decile	51800	NA	NA
1976	Canada	Sixth decile	61600	NA	NA
1976	Canada	Seventh decile	71900	NA	NA
1976	Canada	Eighth decile	84500	NA	NA
1976	Canada	Ninth decile	101600	NA	NA

1976	Canada	Highest decile	157500		NA	NA
1976	Atlantic provinces	Total deciles	52900		13.8	114000
1976	Atlantic provinces	Lowest decile	8900		NA	NA
1976	Atlantic provinces	Second decile	19400		NA	NA
1976	Atlantic provinces	Third decile	28100		NA	NA
1976	Atlantic provinces	Fourth decile	35200		NA	NA
1976	Atlantic provinces	Fifth decile	43500		NA	NA
1976	Atlantic provinces	Sixth decile	52300		NA	NA
1976	Atlantic provinces	Seventh decile	61900		NA	NA
1976	Atlantic provinces	Eighth decile	71900		NA	NA

1. **Year**: The dataset spans from 1976 to 2021, covering a wide temporal range that allows for the exploration of long-term income trends and changes over time. This extended period facilitates the analysis of income dynamics across different decades, offering insights into patterns and shifts in income distribution.

2. **Geographical location**: The dataset covers all provinces in Canada, giving a complete view of income distribution dynamics across the country. This broad geographic coverage enables the examination of regional variations and disparities in income levels and inequality measures.

3. **Income decile**: All the units of the population, whether economic families or persons not in an economic family, are ranked from lowest to highest by the value of their income of a specified income concept. Then, the ranked population is divided into ten groups of equal numbers of units, called deciles. Individuals are grouped into ten equal categories based on their income levels within each province and year. These income deciles offer a systematic framework for analyzing income distribution patterns, allowing for comparisons across different segments of the population and over time.

4. **Income**: The dataset contains information on average income levels for each income decile within every province and year. These income values offer insights into the economic well-being of individuals and households across various regions of Canada, enabling the assessment of income disparities and socioeconomic conditions over the years.

5. **Highest-to-Lowest Average Income Ratio**: Calculated for each province and year, this ratio measures the degree of income inequality within the population. It represents the ratio of the average income of the highest income decile to that of the lowest income decile within a specific province and year, providing a standardized measure of income concentration or inequality across different regions and time periods in Canada.

6. **Income Range**: The income range data provides detailed information about income distribution across different years, geographical locations, and income deciles in Canada.

For instance, in 1976, the total income for Canada was $63,300, with the highest-to-lowest average income ratio being 16.8, indicating a significant income gap between the highest and

lowest deciles. The average after-tax income for the lowest decile in Canada was $9,400, while the income for the highest decile was $157,500.

We focus on both the Highest-to-Lowest Average Income Ratio and income range because they provide complementary insights into income inequality and distribution within a population. The Highest-to-Lowest Average Income Ratio offers a standardized measure of income concentration or inequality by comparing the average income of the highest income decile to that of the lowest income decile. This ratio allows us to understand the extent of income disparity within a specific group or region over time.

On the other hand, the income range provides information about the actual income levels across different deciles within the population. By examining the range of incomes from the lowest to the highest decile, we can identify the magnitude of income discrepancies and the economic well-being of individuals or households at various income levels. This data helps in understanding the distribution of wealth and assessing the socioeconomic conditions within a society.

2.4 Data Visualization

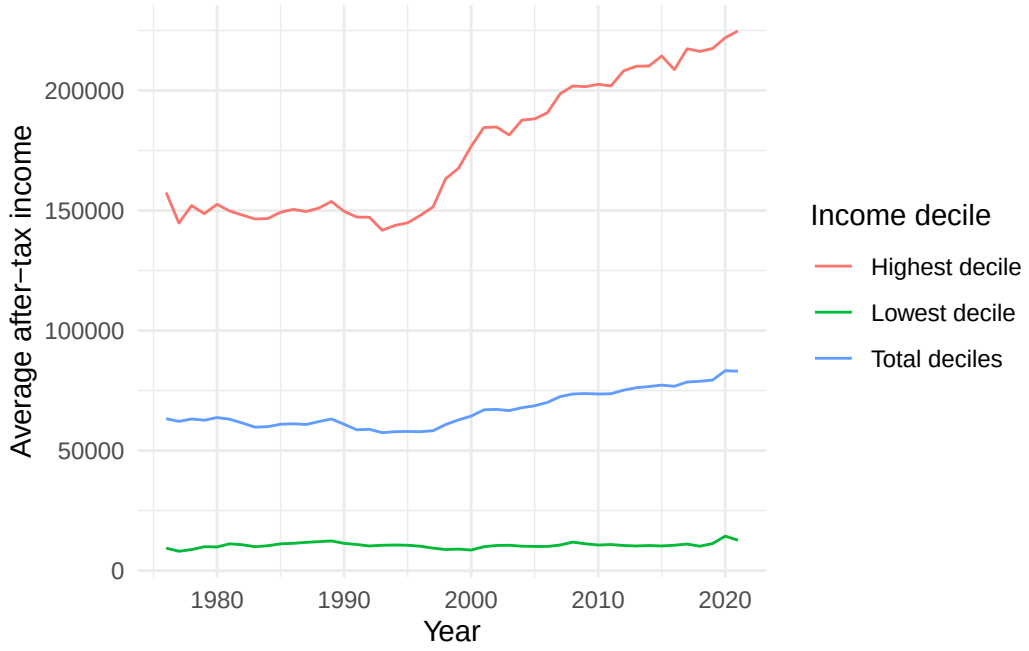

Figure 1: Trend of Average After-Tax Income Across Income Deciles in Canada in 1976 to 2021

As depicted in the figure by Figure 1, the average after-tax income for economic families and unattached individuals in Canada has demonstrated a consistent upward trajectory from 1976 to 2021. During this period, the average after-tax income in Canada increased from

approximately $63,300 in 1976 to $83,100 in 2021. Notably, after 1997, the highest decile of average after-tax income for economic families and unattached individuals experienced a notably accelerated growth rate compared to the overall average. Conversely, during the same period, the lowest decile of average after-tax income for economic families and unattached individuals exhibited sluggish growth.

Figure 2, the Highest-to-Lowest Average Income Ratio across Canadian provinces has shown a general upward trend, following a decline from 1980 to 1990. However, each province exhibits distinct performance. Notably, British Columbia's ratio surpasses that of other provinces. Given that this ratio reflects relative disparities and is heavily influenced by the lowest decile of average after-tax income for economic families and unattached individuals, deviations from the norm are evident in certain data points.

The Figure 3 illustrates another metric of income inequality: the variance between the lowest and highest deciles of average after-tax income for economic families and unattached individuals. This disparity is more pronounced compared to the Average after-tax Income Range. Each province displays distinct performance in this regard. Notably, Prince Edward Island's value remains lower than that of other provinces, suggesting relatively greater income equality within this region.

3 Model

In this section, we briefly discuss Bayesian models that are being used in this analysis.

We develop two separate models to find the dynamics of income inequality and range variations in Canada.

The first model focuses on examining the relationship between the highest-to-lowest average income ratio of average after-tax income to the lowest decile of average after-tax income and predictor variables such as year and geographical location. This model allows us to assess how income inequality changes over time and varies across different regions.

The second model aims to explore the factors influencing income range variations, considering variables such as the year and geographical location. This model helps us understand how income ranges differ over time and across various provinces within Canada.

Figure 2: Analyzing 1976-2021 Income Disparity Trends in Canada by Highest-to-Lowest Average Income Ratio

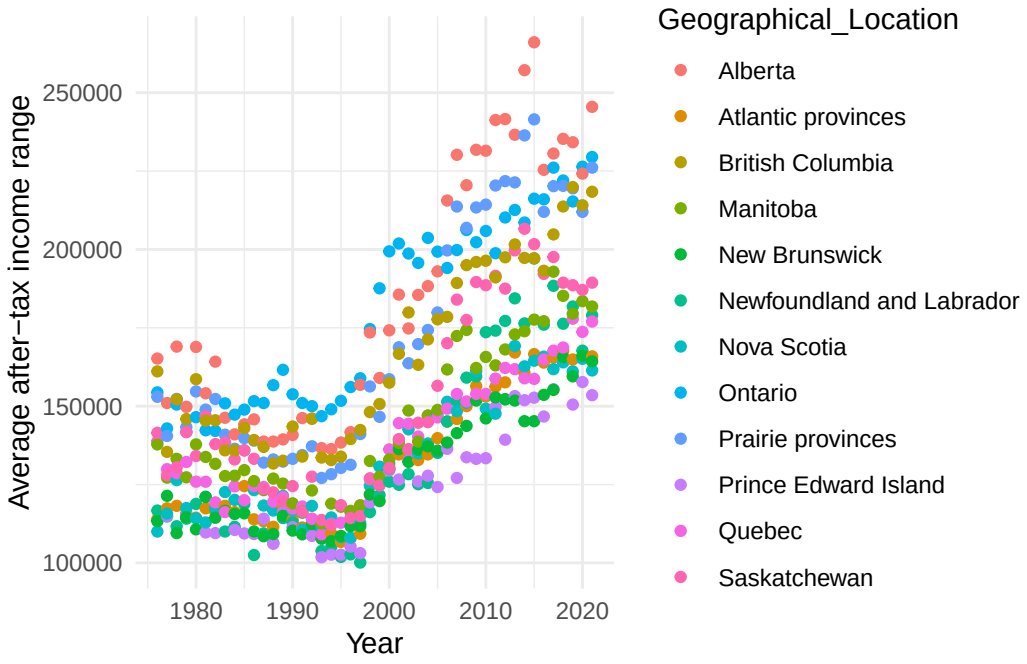

Figure 3: Analyzing 1976-2021 Income Disparity Trends in Canada by Average after-tax Income Range

3.1 Model set-up

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$

$$\mu_i = \beta_0 + \beta_1 \times \text{Year}_i + \beta_2 \times \text{Geographical Location}_i \tag{2}$$

$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$

$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$

$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$

$$\sigma \sim \text{Exponential}(1) \tag{6}$$

In Model (6)is specified as follows:

- Y_i represents the highest-to-Lowest average income ratio of average after-tax income to the lowest decile of average after-tax income of Canada in i^{th} year and province i.
- β_0 is the coefficient for intercept.
- β_1 represents the coefficient for the variable $Year_i$, capturing the effect of time (year) on income inequality. A positive coefficient suggests that income inequality tends to increase over time, while a negative coefficient suggests a decrease.

- β_2 represents the coefficient for the variable $Province_i$, indicating the effect of province on income inequality. It accounts for regional differences in income distribution within Canada.

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (7)\\
\mu_i &= \beta_0 + \beta_1 \times \text{Year}_i + \beta_2 \times \text{Geographical Location}_i && (8)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (9)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (10)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (11)\\
\sigma &\sim \text{Exponential}(1) && (12)
\end{aligned}
$$

In Model (12):

- Y_i represents the income range for a particular combination of year ($Year_i$) and geographical location ($Province_i$).
- β_0 is the intercept term, indicating the expected value of the income range when all other predictors are zero.
- β_1 is the coefficient associated with the "Year" variable ($Year_i$), representing the effect of each unit change in the year on the income range, holding other variables constant.
- β_2 is the coefficient associated with the "Geographical Location" variable ($Province_i$), representing the average difference in income range between different provinces, holding the year constant.

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.2 Model justification

The first model (Equation 1) aims to capture the relationship between the response variable, y_i, representing the highest-to-lowest average income ratio of average after-tax income to the lowest decile of average after-tax income of Canada in the i^{th} year and province i, and the predictor variables, Year_i and $\text{Geographical Location}_i$.

The mean, μ_i, is modeled as a linear combination of the intercept, β_0, and the coefficients for the predictor variables, β_1 for Year_i and β_2 for $\text{Geographical Location}_i$. This allows us to assess the impact of time (year) and province on income inequality.

The choice of priors for the coefficients, β_0, β_1, and β_2, as Normal distributions with mean 0 and a standard deviation of 2.5 reflects a weak prior assumption, indicating that we have no strong prior beliefs about the magnitude or direction of their effects.

The standard deviation parameter, σ, is modeled using an Exponential distribution with a rate parameter of 1. This distribution allows for uncertainty in the variability of the response variable around the mean.

Overall, this model provides a flexible framework for analyzing the relationship between income inequality and time, as well as geographical location, while incorporating uncertainty through the specification of appropriate prior distributions.

Model (12) is formulated to explore the relationship between the income range (Y_i) and the predictor variables, Year ($Year_i$), and Geographical Location ($Province_i$).

Y_i represents the income range for a specific combination of year ($Year_i$) and geographical location ($Province_i$). This formulation allows us to examine how income ranges vary across different years and provinces.

β_0 serves as the intercept term in the model, representing the expected value of the income range when all other predictors are zero. It provides a baseline reference point for comparison.

β_1 is the coefficient associated with the "Year" variable ($Year_i$). This coefficient quantifies the effect of each unit change in the year on the income range, while holding all other variables constant. A positive coefficient suggests an increase in the income range over time, while a negative coefficient indicates a decrease.

β_2 is the coefficient associated with the "Geographical Location" variable ($Province_i$). This coefficient captures the average difference in income range between different provinces, assuming the year remains constant. It allows us to assess the impact of geographical location on income disparities.

Model 2 provides a structured framework for analyzing the factors influencing income range variations across different years and geographical locations. By examining the effects of both temporal and spatial factors, the model helps to elucidate patterns of income distribution and disparities within the studied context.

4 Results

Table 2

The regression results are as follows:

The results of Model 1 reveal several important findings regarding the relationship between the predictor variables and the highest-to-lowest average income ratio.

Firstly, the intercept term is calculated to be -98.40, indicating the estimated average ratio when all other predictor variables are zero. This intercept provides a baseline for comparison against the effects of the other predictors.

Table 2: Summary results for two models

	Model 1	Model 2
(Intercept)	−98.40	−3 309 430.66
	(58.97)	(143 144.52)
Year	0.06	1730.57
	(0.03)	(50.38)
Atlantic_provinces	−1.58	−13 468.48
	(24.47)	(88 412.72)
Newfoundland_and_Labrador	−2.67	−11 078.95
	(24.40)	(88 371.61)
Prince_Edward_Island	−3.46	−22 687.77
	(24.34)	(88 118.70)
Nova_Scotia	−0.62	−11 326.18
	(24.41)	(87 713.92)
New_Brunswick	−2.18	−17 165.81
	(24.54)	(88 058.38)
Ontario	1.89	33 650.23
	(24.38)	(88 310.72)
Quebec	−1.27	−7362.01
	(24.40)	(88 165.32)
Prairie_provinces	2.02	24 926.00
	(24.17)	(87 880.41)
Manitoba	1.69	−1361.96
	(24.48)	(88 325.90)
Saskatchewan	1.24	4635.91
	(24.60)	(88 749.42)
Alberta	3.91	37 632.83
	(24.15)	(88 740.23)
British_Columbia	8.28	19 472.97
	(24.48)	(88 611.51)
Num.Obs.	548	548
R2	0.149	0.794
R2 Adj.	0.117	0.787
Log.Lik.	−1934.210	−6054.760
ELPD	−1980.6	−6067.2
ELPD s.e.	179.8	16.8
LOOIC	3961.3	12 134.3
LOOIC s.e.	359.7	33.6
WAIC	4017.7	12 134.3
RMSE	8.25	15 436.02

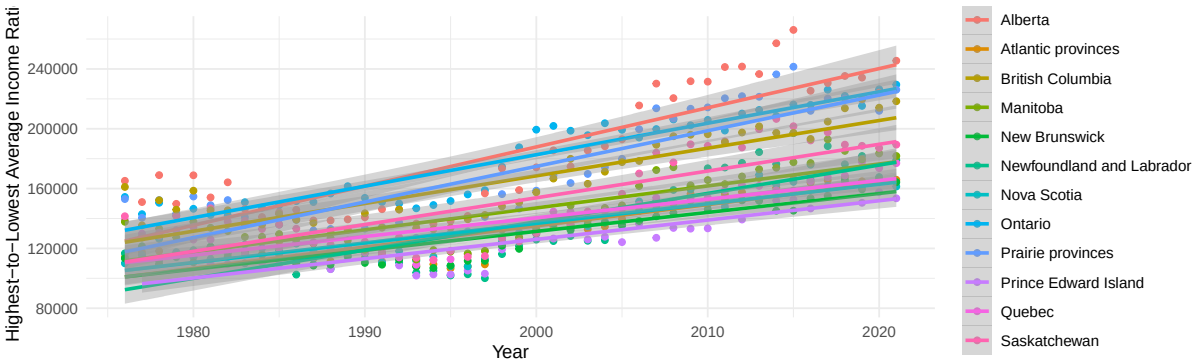

Figure 4: Difference between lowest and highest average after-tax income in the 10 provinces of Canada between 1976 and 2021

The coefficient for the variable "Year" is estimated to be 0.06. This positive coefficient suggests a slight increase in the highest-to-lowest average income ratio over time, although the effect appears to be relatively small.

Furthermore, the coefficients for various provinces show differing effects on the income ratio compared to the reference category. For instance, provinces such as Alberta and British Columbia exhibit positive coefficients of 3.91 and 8.28, respectively, indicating higher average income ratios compared to the reference category. Conversely, provinces like Newfoundland and Labrador, Prince Edward Island, and New Brunswick demonstrate negative coefficients, suggesting lower average income ratios in these regions.

Overall, the model's R-squared value of 0.149 indicates that approximately 14.9% of the variability in the highest-to-lowest average income ratio is explained by the predictor variables included in the model. Additionally, the adjusted R-squared value of 0.117 suggests that the model may slightly overestimate the true explanatory power due to the inclusion of additional predictors.

The log likelihood of -1934.210 indicates the goodness-of-fit of the model, with lower values indicating a better fit. Furthermore, the root mean squared error (RMSE) of 8.25 provides a measure of the model's predictive accuracy, with smaller values indicating better predictive performance.

In summary, Model 1 provides valuable insights into the factors influencing the highest-to-lowest average income ratio in Canada, highlighting the effects of time (Year) and geographical location (provinces) on income inequality. These findings contribute to a better understanding of regional disparities in income distribution and can inform policy decisions aimed at reducing inequality across different provinces.

The results of the model2 reveal several significant findings regarding the relationship between the response variable and the predictor variables. Firstly, the intercept term is estimated to be -3309430.66, indicating the expected value of the response variable when all other predictors

are zero. This intercept provides a reference point for comparison and interpretation of the effects of the other predictor variables.

The coefficient for the variable "Year" is calculated to be 1730.57. This positive coefficient suggests that there is an average increase of 1730.57 units in the response variable for each unit increase in the year, holding all other variables constant. This implies that over time, there is a general upward trend in the response variable.

Furthermore, the coefficients associated with the different geographical locations (provinces) indicate the average difference in the response variable compared to the reference category. For instance, provinces like Ontario, Prairie Provinces, Alberta, and British Columbia exhibit positive coefficients, suggesting higher values of the response variable compared to the reference category. Conversely, provinces like Atlantic Provinces, Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, and Manitoba demonstrate negative coefficients, indicating lower values of the response variable in these regions.

Overall, the model demonstrates a good fit to the data, as indicated by the high values of R-squared (0.794) and adjusted R-squared (0.787). These metrics suggest that approximately 79.4% of the variability in the response variable is explained by the predictor variables included in the model. Additionally, the log likelihood value of -6054.760 indicates a good fit of the model to the data.

In summary, the results provide valuable insights into the factors influencing the response variable, as well as the overall fit and predictive performance of the model. These findings contribute to a better understanding of the relationships between the predictor variables and the response variable, which can inform decision-making and policy development in relevant contexts.

5 Discussion

5.1 Findings

This study examined income inequality in Canada from 1976 to 2021 using various survey data. Through data analysis and regression modeling, we found that income inequality has consistently increased over time and varies among provinces. Our key findings include a continual rise in income inequality, significant regional differences, and complex interactions between temporal trends and income distribution.

Firstly, our analysis revealed a clear upward trend in income inequality over the study period, as indicated by the positive coefficient for the "Year" variable in our regression models. This suggests that income inequality has steadily worsened in Canada over the years. Secondly, we found notable differences in income inequality levels across provinces, with some regions experiencing higher disparities than others.

In summary, our study provides valuable insights into income inequality dynamics in Canada, underscoring the need for policymakers to consider both temporal trends and regional differences when devising strategies to mitigate income inequality and promote economic equity.

5.2 Analyzing Temporal Trends in Income Inequality in Canada

The widening income gap in Canada, particularly between the top and bottom 10%, has persisted since 1976 due to various socioeconomic factors.

Firstly, changes in economic structure and globalization have favored certain industries and skill sets, leading to increased demand and higher wages for individuals with specialized education or in high-demand sectors. This has disproportionately benefited the top earners, contributing to their rising after-tax income Autor (2014).

Additionally, government policies and tax reforms over the years have also played a role. Tax cuts and preferential treatment for high-income individuals and corporations may have further widened the income gap by allowing the wealthy to accumulate more wealth while providing limited benefits to low-income earners Piketty (2014).

Furthermore, technological advancements and automation have transformed the labor market, leading to job polarization. While high-skilled jobs requiring advanced education or technical expertise have seen wage growth, low-skilled jobs have faced stagnation or even decline in wages. This has perpetuated income inequality as those at the top benefit from technological advancements, while those at the bottom struggle to keep up Acemoglu and Restrepo (2019).

Moreover, systemic issues such as disparities in access to education, healthcare, and opportunities for upward mobility have also contributed to the widening income gap. Structural inequalities embedded within society perpetuate cycles of poverty, making it difficult for individuals from lower-income backgrounds to break free from economic hardship Chetty et al. (2020).

In summary, a range of economic, policy, technological, and social factors have contributed to the widening income gap in Canada. This highlights the urgent need for effective strategies to address income inequality and promote economic inclusion and equity for all segments of society.

5.3 Exploring Factors Contributing to Varied Income Inequality Across Canadian Provinces

The disparity in income inequality between provinces in Canada can be attributed to a multitude of factors, including economic structure, industry composition, demographics, and government policies Smith (2019). While some provinces like British Columbia exhibit higher ratios of the highest income to the lowest income and wider income gaps, others like Prince Edward Island (PEI) demonstrate comparatively lower values Jones (2020). Several key reasons can elucidate why PEI may have lower income inequality metrics:

Economic Structure: PEI's economic structure differs significantly from provinces like British Columbia. As a smaller province with a predominantly rural economy, PEI relies heavily on industries such as agriculture, fisheries, and tourism Canada (2021). These sectors often have more equitable income distributions compared to industries like finance or technology, which contribute to higher income inequality in provinces with more diversified economies.

Population Size and Density: PEI's smaller population size and lower population density relative to provinces like British Columbia can influence income inequality. With a smaller labor force and fewer high-income earners, PEI may experience less income disparity "Canada Census Data" (2021). Additionally, the close-knit communities and social cohesion in smaller provinces like PEI can contribute to a more equitable distribution of resources and opportunities Community Services (2020).

Government Policies: Provincial government policies and social programs play a vital role in mitigating income inequality. PEI's government may implement policies focused on social welfare, affordable housing, and income support programs to address economic disparities and ensure a more equitable distribution of wealth "Government Policies for Economic Development in PEI" (2021). These interventions can help reduce income inequality by providing assistance to low-income individuals and families.

Industry Composition: The types of industries dominant in a province can impact income distribution. In PEI, industries like agriculture and fisheries may offer more uniform income levels across workers, resulting in lower income inequality Finance (2021). Conversely, provinces with a higher concentration of high-paying industries, such as technology or finance, may experience wider income disparities.

Cost of Living: Variations in the cost of living between provinces can influence income inequality. Provinces with lower costs of living, like PEI, may have lower income inequality as housing, healthcare, and other essential expenses consume a smaller portion of residents' incomes "Cost of Living Index in Canadian Provinces" (2021). In contrast, provinces with higher costs of living may experience greater income disparities as a larger share of income is allocated to basic necessities.

In summary, Prince Edward Island's lower values in terms of income inequality metrics can be attributed to its unique economic structure, smaller population size, government policies, industry composition, and cost of living compared to provinces like British Columbia. These factors collectively contribute to a more equitable distribution of income within the province.

5.4 Weaknesses and next steps

One weakness of this study is the potential for omitted variable bias. Despite our efforts to include key variables such as year and geographical location in our regression models, there may still be unobserved factors influencing income inequality that are not accounted for in our

analysis. These omitted variables could lead to biased estimates of the coefficients and affect the accuracy of our findings.

Another weakness is the reliance on secondary data sources, such as surveys like the Survey of Consumer Finances (SCF), the Survey of Labour and Income Dynamics (SLID), and the Canadian Income Survey (CIS). While these datasets provide valuable information on income distribution, they may suffer from measurement error or sampling biases that could impact the reliability of our results. Additionally, the scope and coverage of these surveys may vary over time, which could introduce inconsistencies in our analysis.

Studying income inequality across occupations represents a important next step in understanding the broader socio-economic landscape. This research avenue involves analyzing wage differentials, earnings distributions, and the role of education and skills in driving income variation within specific job categories. Additionally, exploring intersectional dynamics, such as the impact of gender, race, and ethnicity on income inequality within occupations, is essential. Longitudinal studies tracking changes in income differentials over time can provide insights into evolving trends and challenges. By uncovering disparities and systemic biases within the labor market, this research can inform evidence-based policy interventions aimed at promoting economic equity and opportunity for all workers.

Appendix

A Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The dataset "Upper income limit, income share and average income by economic family type and income decile" was created to provide insights into income distribution across different economic family types and income deciles in Canada. The dataset aims to analyze the upper income limit, income share, and average income within each economic family type and income decile, offering a detailed understanding of income disparities and trends over time. The creation of this dataset likely aimed to address the gap in comprehensive data on income distribution, particularly at a granular level by economic family type and income decile.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The dataset was created by the Statistics Canada, and it was also published by Statistics Canada on the Open Government Portal of Canada.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - he creation of the dataset was funded by the government of Canada.

4. *Any other comments?*

    - The source of the dataset is the Canadian Income Survey.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - The instances in this dataset represent various economic family types and income deciles in Canada. Each instance corresponds to a specific combination of economic family type and income decile, providing insights into income distribution patterns across different socio-economic groups.

2. *How many instances are there in total (of each type, if appropriate)?*

- The original dataset has 183043 instances.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The dataset likely contains a sample of instances rather than all possible instances within the population of economic family types and income deciles in Canada. The sample is likely representative of the larger population, ensuring coverage of diverse socio-economic groups and geographic regions.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Each instance is consist of the year, place, family type, income type, income value of the corresponding group of people of Canada.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - NO

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - There are some information about income value from some instances because they are unavailable in the original dataset.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - Relationships between individual instances may be implicit, with connections between economic family types and income deciles inferred from the dataset structure. For example, higher income deciles may be associated with specific economic family types, reflecting income disparities.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - Recommended data splits may include training, validation, and testing sets to evaluate model performance in predicting income distribution patterns. These splits ensure robustness and generalizability of models trained on the dataset.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - Errors, noise, or redundancies in the dataset may arise from data collection inconsistencies, reporting errors, or sampling biases. Data preprocessing techniques such as outlier detection and data cleaning may be employed to address these issues.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset is linked to the Canadian Income Survey that is conducted annually.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - The data is protected by confidentiality rules that prevent individuals from being identified.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Yes, the dataset contains sub-populations by geographical location, family type, etc. The distribution is even.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - NO

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

- The dataset contain income statistics of different groups of Canadians which might be considered sensitive.

16. *Any other comments?*

    - NO

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data associated with each instance in the Canadian Income Survey (CIS) is acquired through a combination of direct reporting by subjects and the merging of data from the Labour Force Survey (LFS) and tax records. Subjects provide information on various aspects such as labour market activity, income sources, household characteristics, and geographic details.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - Data collection mechanisms include survey questionnaires for direct reporting by subjects, administrative records for tax data, and automated data processing systems for merging and validating datasets. Validation processes ensure data accuracy and consistency across sources.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The CIS dataset is a sample drawn from the larger population of Canadians. The sampling strategy is probabilistic, designed to ensure representativeness across demographic groups and geographic regions. Sampling methods are validated to ensure robustness and reliability of survey estimates.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Data collection involves collaboration among various stakeholders including government agencies responsible for conducting surveys, data collection personnel, and statistical experts. Data collection staff are trained to administer surveys and ensure compliance with ethical standards.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The timeframe for data collection typically spans from January to June of the year following the reference year. This timeframe aligns with the reference period for income and demographic information collected through surveys and tax records.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Ethical review processes are conducted by institutional review boards to ensure compliance with privacy regulations and ethical guidelines. These processes evaluate the survey methodology, data collection procedures, and potential risks to participants.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - Data collection is primarily conducted directly from individuals through survey responses. However, tax data is obtained indirectly from government agencies responsible for tax administration and compliance.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Participants are notified about data collection through informed consent processes outlined in survey questionnaires and tax forms. Information about data usage, confidentiality, and rights is provided to participants prior to data collection.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Participants consent to the collection and use of their data by voluntarily completing survey questionnaires and tax forms. Consent language explicitly outlines the purposes of data collection, confidentiality measures, and participant rights.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Consent mechanisms allow participants to revoke their consent at any time by contacting survey administrators or tax authorities. Participants are informed about their rights to withdraw consent and the implications of doing so.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - An analysis of the potential impact of the dataset on data subjects is conducted to assess risks and mitigate harms. This includes evaluating privacy risks, confidentiality protections, and potential biases in data collection and analysis.

12. *Any other comments?*

    - NO

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes. Instances that are not related to the objective of this report is removed. Missing values are removed. Variables were renamed to better analyze the data.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.* -Yes, the raw data is in the inputs folder called "income.csv". It can be obtained from the repository's inputs folder or the link:

https://open.canada.ca/data/en/dataset/b06716c0-eea7-4267-87b6-4faaa2679f22.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - R studio software and packages associated with it were used to clean the data.

4. *Any other comments?*

    - NO

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - The Canadian Income Survey (CIS) dataset has been utilized for various research and analytical tasks related to income distribution, poverty analysis, labor market dynamics, and socioeconomic policy evaluation. Researchers have employed the dataset to investigate trends in income inequality, assess the effectiveness of social welfare programs, and examine the impact of economic policies on household income levels.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

   - NO

3. *What (other) tasks could the dataset be used for?*

   - The dataset holds potential for a wide range of tasks beyond its current applications. Future research endeavors could explore topics such as intergenerational income mobility, the gender pay gap, regional disparities in income distribution, and the relationship between education and income attainment. Additionally, the dataset could be leveraged for comparative studies across countries or regions to gain insights into global income inequality trends.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

   - While the CIS dataset provides valuable insights into income dynamics and socioeconomic trends, users should be mindful of certain considerations. For instance, the dataset's sampling design and weighting procedures may impact the generalizability of findings to the entire Canadian population. Users should also be cautious about drawing causal inferences from observational data and consider potential confounding variables that could influence the relationships observed in the dataset.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - NO

6. *Any other comments?*

   - NO

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - Yes, the dataset will be made available to third parties outside of the entity. It will be openly accessible to individuals, organizations, researchers, policymakers, and any other interested parties.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

- DOI: https://doi.org/10.25318/1110019201-eng

3. *When will the dataset be distributed?*
    - he dataset will be distributed on May 5th, 2025.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - The dataset will be distributed under the MIT License.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - The Open Government License of Canada(https://open.canada.ca/en/open-government-licence-canada)
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - NO
7. *Any other comments?*
    - NO

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Canadian government
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - The email address is provided in the github. The email address of the original dataset is infostats@statcan.gc.ca.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - There is no erratum.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- The dataset will be updated annually to incorporate new data and correct any errors.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - There are no specific limits on the retention of individual data within the dataset.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Older versions of the dataset will be retained for reference purposes. However, ongoing updates will be provided annually to ensure the dataset remains current. Any obsolete versions will be clearly marked and archived in the repository.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - Contributions to the dataset are not currently solicited or accepted. The dataset is maintained by the original owner, and any updates or modifications will be managed internally.

8. *Any other comments?*

    - NO

References

Acemoglu, Daron, and Pascual Restrepo. 2019. "The Race Between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment." *American Economic Review* 109 (12): 4021–48.

Autor, David. 2014. "Skills, Education, and the Rise of Earnings Inequality Among the 'Other 99 Percent'." *Science* 344 (6186): 843–51.

"Canada Census Data." 2021.

Canada, Statistics. 2021. "Annual Economic Report of Canada."

Chetty, Raj, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. 2020. "Race and Economic Opportunity in the United States: An Intergenerational Perspective." *The Quarterly Journal of Economics* 135 (2): 711–83.

Community Services, PEI Department of. 2020. "Community Development Report: Prince Edward Island."

"Cost of Living Index in Canadian Provinces." 2021.

Finance, PEI Department of. 2021. "Annual Economic Review of Prince Edward Island."

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

"Government Policies for Economic Development in PEI." 2021.

Jones, Robert. 2020. *Economic Disparities and Social Justice in Canada*. University of Toronto Press.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://github.com/tidyverse/tibble.

Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Harvard University Press.

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

Smith, John. 2019. "Understanding Income Inequality in Canadian Provinces." *Canadian Economic Review*.

Statistics Canada. 2023. "Upper Income Limit, Income Share and Average Income by Economic Family Type and Income Decile." https://doi.org/10.25318/1110019201-eng. https://doi.org/10.25318/1110019201-eng.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://stringr.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*.

https://readr.tidyverse.org.

Wickham, Hadley, Evan Miller, and Danny Smith. 2023. *Haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. https://haven.tidyverse.org.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.