Evolution of Income Inequality in Canada*

It's becoming a more serious issue as over time, especially in the richest provinces.

Alicia Yang

27 April 2022

Abstract

People have claimed that the income disparity in Canada has increased alongside with the rapid growth of its economy. To explore whether and how income disparity became a serious issue in Canada, this paper aims to investigate the relationship between time, province, and income inequality in Canada. By utilizing the data from Statistics Canada, we found that income inequality continuously increase over time and is a greater issue in more developed provinces. Particularly, it is closely related to the prosperity of a region. In addition, implications and possible reasons behind this correlation is discussed and possible solutions to the problem of income disparity is proposed.

Contents

*Code and data are available at: https://github.com/Alicia-y/Telling-stories-with-data-final-paper

1 Introduction

In recent decades the world as a whole is developing fast economically, and one evidence of which is the rise in income level. As more countries become more developed in terms of their economy and eventually overcome their issues of poverty, their governments have gradually shift their goal to dealing with the problem of economic inequality. It concerns how the wealth and income of the population is distributed. Although we are generally living higher quality lives and getting higher incomes, another problem is being put forward by people who are claiming that the gap between the rich and poor is getting larger due to technological advancement, globalization, countries' governmental structure and policies, which betters off the people who are already wealthy by making it easy for them to seek for more opportunities, thus, snowballing their assets faster. Income disparity can be a huge issue. It makes the income per capita of a country less reflective of its people's true living quality, because income per capita does not tell the distribution of the income. The richest people may pull up the statistics about the average/median income level, but it doesn't mean the country is performing well economically as a whole. In addition to that, income inequality also have other potential negative effects. It could reduce the stability of the economy and increase the risk of financial crisis. It could also result in corruption, mis-allocation of resources as the rich is empowered economically, socially and even politically (Scheffler 2020). Canada is one of the country in the world that provides the most benefits and welfare, yet people still claimed that the economic inequality is rising along with the income over the past 2 decades (Economics 2022). Therefore, I'm interested in whether income inequality is a serious issue in Canada. If so, in which regions and what period of time is it a serious issue? And is it possible to predict how the disparity of income change in future?

The data used in this report is obtained from Open Government Data of Statistics Canada. The dataset is called "Upper income limit, income share and average income by economic family type and income decile", and it provides different kinds information about the income level of the households of Canada between 1976 and 2020 by the Canadian Income Survey. In this dataset, I'm interested in the year the data is from, the geographical location, the income decile so that we could observe the difference between the lowest and highest decile, and the actual value of the income. In section 2, a cleaned dataset is obtained and explained to perform further analysis. Visualizations in the forms of tables and figures are presented to help to explain the possible correlation between year, provinces and the income disparity in Canada. By data visualization, I've primarily assumed that the income disparity in Canada has generally increased as time passed and presents different trends in different provinces before constructing the model. In section 3, a multiple linear regression model is constructed to justify the relationship between year, province and income inequality, and to make future predictions. The interpretation of the final model along with all the findings regarding the evolution of income inequality is presented in section 4. A discussion is carried out in section 5 on the implications of the findings regarding time and geographical location, and possible solutions to this issue in federal and provincial level, as well as the weaknesses and future steps of this paper.

2 Data

2.1 Dataset Description and Methodology

The dataset, which is a summary of the income-related statistics in Canada between 1976 and 2020, is obtained from the Open Government Portal of Statistics Canada. The source of the data is from the Canadian Income Survey conducted annually. The survey covers all the possible population in Canada except for the people and households living in remote areas or indigenous settlements. It doesn't really affect the outcome of the survey, because these people only take up less than 2 percent of the population (Canada 2020). The selected respondents were drawn from the Labour Force Survey samples, which is based on a stratified probability sampling (Canada 2020). To reduce some non-sampling errors, telephone interviews ware conducted prior to the main survey to increase the response rate of the selected participants Once they give consent to completing the survey, they will be able to complete it in an online form. Respondents are protected from the confidentiality rules, so that their privacy won't be disclosed after the survey is made public. One possible bias came with the survey methodology was the over-coverage of the units that are not the target population and the under-coverage of certain sub-population that should be included in the survey,

such as certain remote areas.

The original dataset contains 173043 observations and 18 variables with all kinds of information about income level statistics. This report wants to focus on exploring the trend of income inequality and investigating the possible factors that may affect income inequality. Thus, I will be interested in the variables "Year", "Geographical Location", "Income decile", "Income" which tells how the income level of different regions of Canada has evolved throughout these years. R (R Core Team 2020), and R packages "tidyverse" (Wickham et al. 2019), "janitor" (Firke 2021), "knitr" (Xie 2021), "dplyr" (Wickham et al. 2021), and "kableExtra" (Zhu 2021) are utilized to create an extract of the cleaned dataset (Table 1).

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

Year	Geographical location	Income decile	Income	Income range
1976	Canada	Total deciles	72400	184400
1976	Canada	Lowest decile	9200	NA
1976	Canada	Second decile	21200	NA
1976	Canada	Third decile	32400	NA
1976	Canada	Fourth decile	44600	NA
1976	Canada	Fifth decile	57000	NA
1976	Canada	Sixth decile	69100	NA
1976	Canada	Seventh decile	81800	NA
1976	Canada	Eighth decile	97200	NA
1976	Canada	Ninth decile	118300	NA

Table 1 shows the first ten rows of the cleaned dataset. It contains 6395 variables and 5 variables in total. The target population of the dataset is a combination of economic families and unattached individuals in Canada. As we known, the information from the dataset is extracted from the Canada Income Survey. Variable "Year" indicates the year of the survey. Variable "Geographical location" indicates where the group of respondents are from. Variable "Income" gives the average value of after-tax income of the respondents in corresponding year and location. Variable "Income decile" gives more information about the variable "Income", by telling us whether each piece of income information represents the average after-tax income of all respondents, the richest 10 percent respondents, the second richest 10 percent respondents, ..., and the poorest 10 percent respondents. This piece of data is extremely helpful because it also gives extra information about the difference regarding the income level between the poorest and richest groups. The income disparity is measured by the last variable "Income range", which is not in the original dataset. It indicates the difference in income between the poorest and richest groups, which is obtained from manipulating data from the "Income" column. It's only available at the instances of "Total deciles".

2.2 Data Visualization

In order to get further familiarized with the dataset and estimate the possible associations between the nation's average income and other factors, exploratory analysis is carried out by conducting data visualizations to observe whether the patterns between certain factors matches the generally expectations of the trend of income and income inequality in Canada. First of all, we will have a look through the overall trend of income level in Canada.

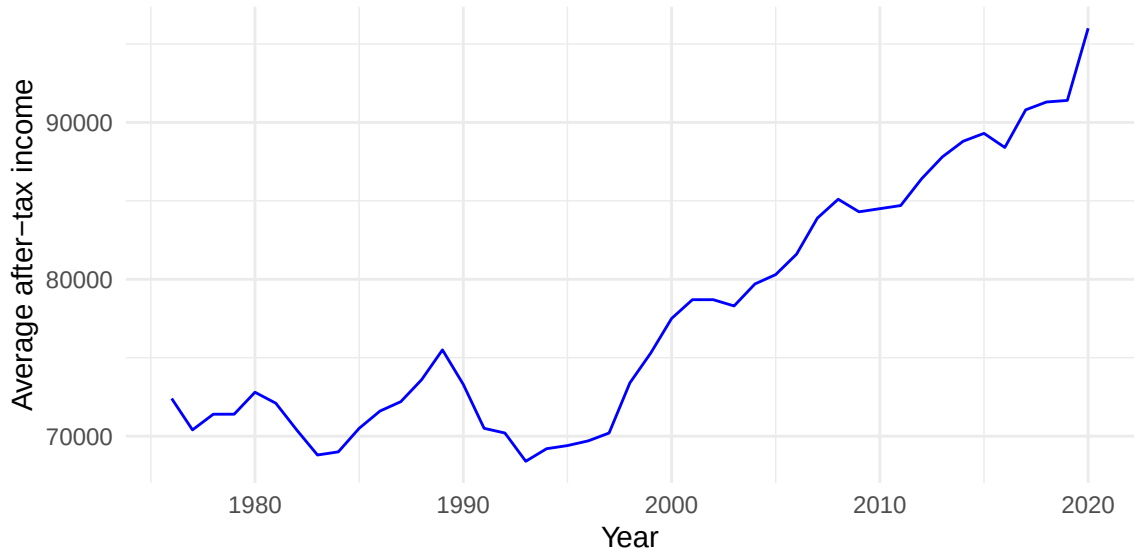

Figure 1: Average after-tax income in Canada per year between 1976 and 2020

As Figure 1 has demonstrated, Canadians have experienced a rapidly growing income level since 1976 to now. Early 1980s and 1990s are basically the only two periods of times that we see an obvious pattern of the decrease in income, and they all lasted for less than five years. This phenomenon is expected as during the time of early 1980s and 1990s, Canada was experiencing recession due to the change in monetary policy and a side effect of the cold war. As a result, the unemployment rate rose and due to an increase in the competitiveness of workforce, people's incomes have dropped. However, during other times, the income level have always demonstrated an increasing trend, indicating that without negative external forces, Canada generally has expanded its economy and enhanced its people's life quality well. The average after-tax income for economic families and unattached individuals in Canada has rose from around $72400 in 1976 to $96000 in 2020.

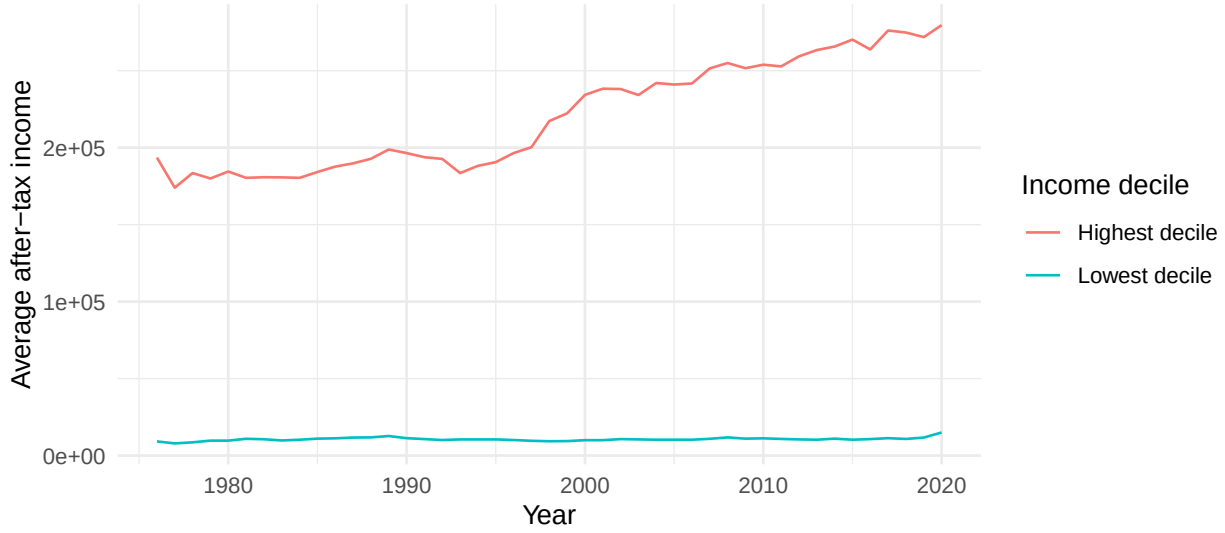

Figure 2: Trend of lowest and highest decile of sverage after-tax income in Canada between 1976 and 2020

Figure 2 is an illustration of how the income level of the richest and poorest groups of economic families and unattached individuals differ between the year of 1976 and 2020. The blue line indicates the trend of the lowest decile of the average after-tax income in Canada, which barely increased. In fact, the lowest decile of average after-tax income was $9200 in 1976, but only $15000 after 44 years in 2020. On the other hand, the red line indicates the trend of the highest decile of the average after-tax income in Canada, which has a very clear increasing pattern, by looking at the data, it increased for $86000 between 1976 and 2020, which is more than 10 times the income level increase of the lowest decile. By comparing the two lines, we can see that the income inequality in Canada as a whole is getting larger as the two lines drift further away from each other as time passed.

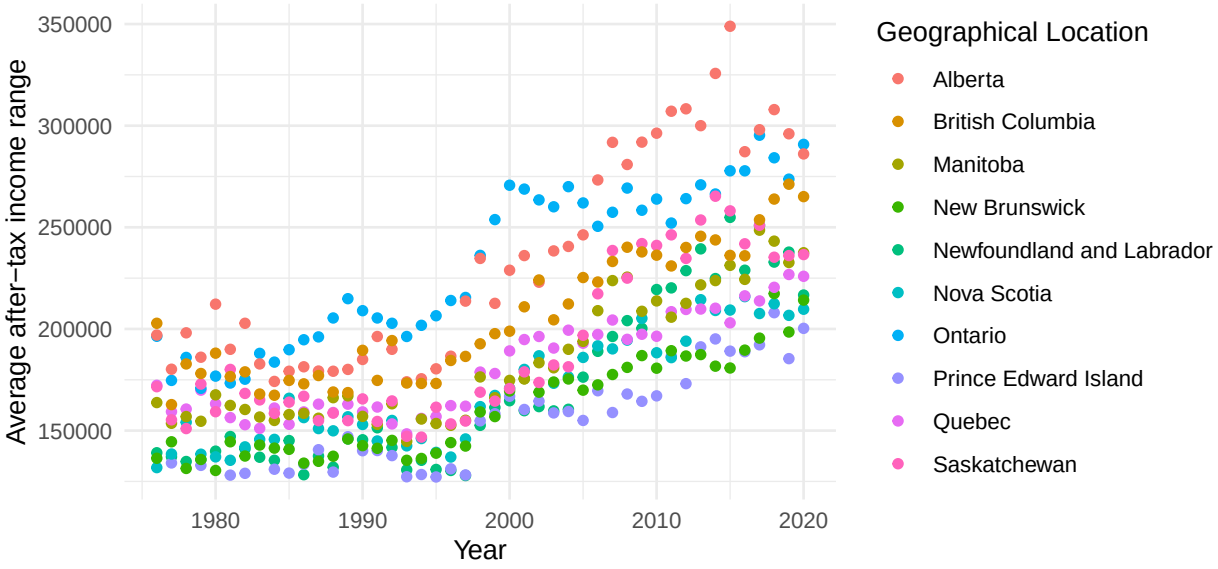

Figure 3: Difference between lowest and highest average after-tax income in the 10 provinces of Canada between 1976 and 2020

Figure 3 has demonstrated the difference between the lowest and highest decile of the average after-tax income of economic families and unattached individuals in the ten provinces of Canada between 1976 and 2020. Overall, there is an increasing pattern in every province, indicating that the gap between the incomes earned by richest and poorest people is getting larger in every region of Canada. By just examining this figure, some particular provinces seem to experience a greater increase in income inequality than the others as time passed. For instance, as shown by the red dots, Alberta clearly has the greatest increase among all the provinces of Canada. On the other hand, Quebec seems to see the smallest increase in income inequality as it has a flatter slope in the figure. The cause the difference in income inequality between these provinces is likely to be difference in their provincial policy and economy.

3 Model

By the exploratory analysis of the data, we found that the year and province does have some correlation with the extent of income inequality in Canada. The relationship seems to be linear as the figures generally demonstrate increasing trends. Therefore, to further proceed the analysis and predict the future situation regarding income disparity, a multiple linear regression model will be constructed.

Prior to constructing the model, the dataset is split into the training set and the testing set with a proportion of 8:2. The training set is used to build the multiple linear regression models, and the testing set is used to test the accuracy and unbiasness of the model. R package "tidymodels"(Kuhn and Wickham 2020) is used to split the dataset.

Two models are initially constructed based on the different possible factors that might have an effect on income inequality in Canada. The first model has continues variable year, categorical variable province as its predictor variables, and the difference between lowest and highest decile of income as its response variable. I would also like to check whether the variables year and province are implicitly related. Therefore, the second model has continues variable year, categorical variable province, and the interaction between them as its predictor variables, and the difference between lowest and highest decile of income as its response variable. To compare these two models to find out which is more accurate and is better at prediction, AIC and BIC tests are carried out, and R^2 is examined. Furthermore, the testing dataset is used to test how the two models perform when extra data is involved. RMSE is measured by the testing data to compare the prediction power of the two models. In Appendix A, these tests are performed and the test statistics of the two models are compared to select the model with better performance.

As a result, model 2 has a better performance than model 1, and is selected as the final model to model the behavior of Income inequality in Canada as time passed. The assumption check for the model is done in Appendix B.

The final model is displayed below:

$$Y_{ij} = \beta_0 + \beta_1 Year_i + \beta_2 Province_j + \beta_3 Year_i Province_j \tag{1}$$

In Model (1):

- Y_{ij} is the difference between the highest and lowest decile of average after-tax income of economic families and unattched individuals of Canada in i^{th} year and province j.
- β_0 is the coefficient for intercept.
- β_1 is the coefficient for the continuous year variable.
- β_2 is the coefficient corresponding to province j.
- β_3 is the coefficient for the interaction term between i^{th} year and province j.
- The baseline of this model is year 0 and Alberta province.

P-value is an essential value used to determine whether a predictor variable has significant impact on the response variable. It measures the probability that the observed data would occur given that the null hypothesis is true. In the context of a multiple linear regression model in this paper, null hypothesis is the hypothesis that there is no relationship between the predictor variable and the response variable (Bevans 2020b). The common threshold of p-value is 0.05, so we will use this threshold in this paper. By examining the p-values of all the predictor variables of the model, we found that they are all less than 0.05, indicating that the hypothesis that there is no relationship between the variables of interest is rejected. As a result, none of the predictor variables of this model will be reduced because they all have some signifant impact on the response variable.

4 Results

Model (1) is eventually selected that best demonstrates the correlation between the year, the province of interest, and the difference between highest and lowest decile of average after-tax income of the economic families and unattached individuals in Canada. Figure 4 illustrated a fitted graph of the multiple linear regression model, where each line represents the trend of income inequality as time passed in respect to each corresponding province. As the model indicates, as the year increases, the income gap increases. In the meanwhile, the difference in province also does have an effect on income inequality.

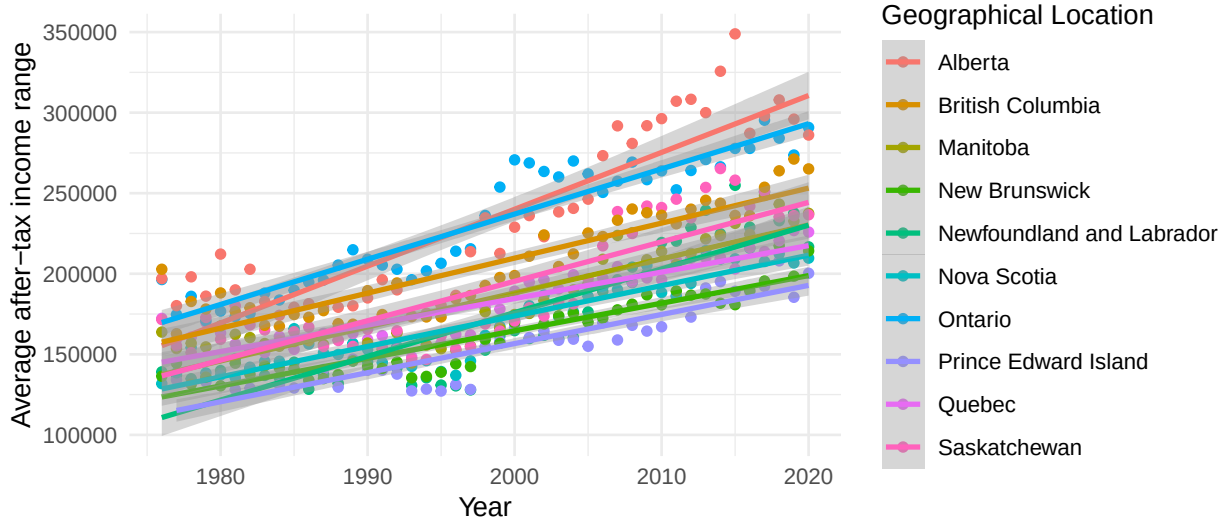

Figure 4: Difference between lowest and highest average after-tax income in the 10 provinces of Canada between 1976 and 2020

Table 2: Model Coefficients and 95 percent Confidence Interval

	Coefficients	Confidence Interval Lower Bound	Confidence Interval Upper Bound
Intercept	-7313468.1	-8125023.3	-6501912.8
Year	3777.0	3371.1	4182.9
British Columbia	3176361.3	2064871.5	4287851.2
Manitoba	2596133.4	1347663.3	3844603.5
New Brunswick	3880833.2	2741920.7	5019745.6
Newfoundland and Labrador	2035618.5	933920.7	3137316.3
Nova Scotia	3517201.0	2386007.1	4648394.9
Ontario	2107020.6	953353.5	3260687.7
Prince Edward Island	3738771.0	2507489.4	4970052.7
Quebec	3876947.7	2694150.1	5059745.3
Saskatchewan	2727094.3	1560977.5	3893211.1
Year: British Columbia	-1603.4	-2159.5	-1047.4
Year: Manitoba	-1325.2	-1949.3	-701.0
Year: New Brunswick	-1978.2	-2548.0	-1408.4
Year: Newfoundland and Labrador	-1049.5	-1600.5	-498.4
Year: Nova Scotia	-1792.4	-2358.3	-1226.6
Year: Ontario	-1055.5	-1632.8	-478.1
Year: Prince Edward Island	-1911.8	-2527.4	-1296.2
Year: Quebec	-1967.1	-2558.8	-1375.4
Year: Saskatchewan	-1386.4	-1970.0	-802.8

Table 2 shows the coefficients of the predictor variables of Model (1). Due to the fact that none of their corresponding p-values have a value more than 0.05, they are all kept in the final model. The baseline of the model is year 0 and Alberta province, and it also accounts for the interaction term between year and Alberta, which means that an intercept-only model presents the difference between lowest and highest decile of average after-tax income at year 0 in province Alberta. This is of course unrealistic because we only account for the data after 1976. Therefore, the "Year" variable has a strict minimum value of 1976 for this model. If the goal is to estimate the income inequality in the provinces other than Alberta, coefficient β_2 indicating which province it is and β_3 indicating the interaction between year and the specified province come into play, and their corresponding coefficient values can be found in Table 2 under "Coefficients".

As Table 2 shows, in all the provinces of Canada with no exceptions, income inequality is continuously increasing and becoming more serious issues. Among all provinces, Alberta has the most increase in income inequality. As the "Year" variable increases by 1, the difference between lowest and highest decile of the average after-tax income in Alberta can increase by up to $3777, while all the other provinces demonstrate some level of increase in income inequality, but none of them reach the level of increase in Alberta. The provinces following Alberta with the second and third fastest increase in the difference between lowest and highest decile of the average after-tax income is Newfoundland and Labrador and Ontario. As the "Year" variable increases by 1, their estimated income inequality increase respectively by $2727.5 and $2721.5. Quebec, on the other side, is the province with the lowest increase in the difference between lowest and highest decile of the average after-tax income in Canada. As the "Year" variable increases by 1, the estimated difference between lowest and highest average after-tax income decile in Quebec increases by $1809.9. As a result of this, we can see that even though early in 1976, Ontario province had the highest estimated income inequality, with a difference between highest and lowest income decile of approximately $171236.5, but at the end of 2020, Alberta is currently the province with the highest estimated difference between lowest and highest average after-tax income decile, with a estimated value of $316071.9. If in real life, the situation keeps following the trend of our estimated model, by the end of 2050, the predicted income inequality in Alberta will reach a point where the difference between lowest and highest average after-tax income decile will be approximately $429381.9, which is incredibly high.

The confidence interval is a type of statistic that gives an upper bound and a lower bound of the value that we expect that if we rerun the test with different samples of the population, the value will fall into this range for at least a certain percent of times (Bevans 2020a). In this report, I decided to use a two-tailed 95% confidence interval for all the coefficients of Model (1). In the context of this model, I'm confident that if I re-sample the population and construct the same model for 100 times, for 95 times the value of coefficients will fall in the confidence interval. The 95% confidence interval of all coefficients in presented in Table 2.

5 Discussion

5.1 Evolution of Income Disparity by Year

In section 4, we have talked about the outcome of the model, which is that time is the main factor expanding the issue of income disparity in Canada. Between 1976 and 2020, the difference in income level between the poorest group and the richest group has increase from $184400 to $264600 for approximately $80200, which is even more than what an average Canadian earns in 2020. This implies that even though people are aware of this for quite some time, income disparity is still becoming a larger issue, and is likely to grow more in future. Some people claimed that the uneven distribution of income is inevitable if we want the country's economy to keep expanding; it's just a side effect of economic growth. It seems true in the case of Canada, since in Figure 3, we can see that the only times when the difference between highest and lowest income decile decreased, is around 1980s and 1990s when Canada was experiencing recession. However, constant gdp growth does not mean that the average people of Canada are doing better, because commodity prices grow as well along with the income and may be a bigger pressure for the poorest people whose income had not increased much (Economics 2022). The richest countries in the world typically share similar issues with respect to income inequality. On the contrary, countries that have good geographical locations and a potential to be more developed but choose to sacrifice that advantage for the well-being of the general public, such as lots of European countries, usually have pretty low income inequality and poverty level by their tax, welfare

system and government policies.

As for Canada, it's just a matter of seeking a balance between economic development and equality. The federal income-tax system we currently have is progressive, which already favors the average people more than the rich (Funds 2018). However, one of the solution to reduce income inequality more in federal level is to further increase the income-tax for the richest people. This could be the most effective approach, but even though it's enhancing the whole society's welfare, it's also harming the interest of some individuals, which could lead to further ethical issues and a whole lot judgments and discussions around this approach. Therefore, this solution would be too difficult and ideal to carry out.

5.2 The role of Province

As we examined the issue of income inequality in different provinces earlier, we found that Alberta has the highest rate of increase in income inequality, followed by Newfoundland and Labrador and Ontario, while Quebec has the lowest increase in income inequality. These statistics in a way implies the economic development and opportunities in these provinces. For instance, Alberta and Ontario are the few provinces with the most GPD per capita and Market income per capita in recent years. Alberta has a whole profitable oil and energy industry due to its rich oil resources, and Ontario, being the province with the largest population in Canada, developed its economy based on lots of sectors including manufacturing, exports, service, high-tech industries and more. In the provinces with higher income inequality, the residents are likely to experience more competition to seek for better opportunities, and are likely to be under higher level of stress, which could lead to more social and health problems. It's unrealistic to suspend the provinces' economic developments just to promote income equality. One approach to reduce it is to manipulate provincial income tax to favor the groups with lower incomes. Another possible solution is to encourage different provinces to carry out different policies to provide benefits and welfare to the lower-income groups to even out. These approaches could be challenging to carry out and unrealistic at the moment, but if actions are not taken, income disparity will likely to grow into more challenging problems in future.

5.3 Weaknesses and next steps

Strictly speaking, income level of individuals is counted as private information, and it's usually not allowed to publicly disclose it unless granted. Therefore, it would be hard to find individual's income information. The data used to justify income inequality in this paper is the difference between the lowest and highest decile of average after-tax income. This doesn't reflect the real gap between the richest and the poorest, because this dataset gathers the average of the top ten percent and the bottom ten percent. In reality, the issue of income inequality will only be worse that what's predicted by this paper. In the Canadian Income Survey, it indicates that less the people who live in remote areas which accounts for less than 2% of the population are not included in the target population of the survey. In the meanwhile, these people are usually the ones with lower income levels. As a result, a systematic error may occur with this dataset due to the survey's methodology and data collection process, which probably increase the lowest decile of average income and further weaken the issue of income inequality.

This paper analyzed how income diaparity in Canada is affected by time and different regions. However, income inequality is only an aspect of economic inequality. Other aspects of economic inequality includes uneven distribution of the wealth and assets of the Canadians, which also causes similar social challenges A possible future step of this paper is to focus on the wealth inequality and investigate whether it is a problem in Canada as well as its relationships with income inequality. In addition, in a few years from now, we could come back to this paper and see if the trend of income inequality is really what it predicted to be, or if the issue is lightened by potential government policies.

Appendix

A Model Testing

R package "modelsummary"(Arel-Bundock 2022) helps to display the coefficients of the two models as well as the result of a series of tests to compare which model performs better.

R^2 measures how well the model explains its response variable's variation. If R^2 is low, it indicates that the model doesn't fit the data well. By Table 3, in Model 1, 84.5% of the variability is explained, and in Model 2, 87.4% of the variability is explained. Both models demonstrate a pretty high R^2 value.

AIC and BIC measure the prediction ability of the multiple linear regression model. AIC(Akaike's Information Criteria) focuses on how well the model fits unknown data, while BIC(Bayesian Information Criteria) focuses on the true model and favours simpler models (Kellen 2010). Lower AIC and BIC both indicate that the model has better prediction power. By Table 3, Model 2 has slightly lower AIC and BIC than Model 1, implying that Model 2 has more prediction power.

Table 3: Comparing Model 1 and Model 2's Statistics

	Model 1	Model 2
(Intercept)	−4 530 561.568	−7 313 468.082
	(144 712.606)	(412 579.342)
Year	2385.103	3777.033
	(72.368)	(206.356)
‘Geographical Location‘British Columbia	−29 153.586	3 176 361.343
	(3952.739)	(565 060.434)
‘Geographical Location‘Manitoba	−53 167.674	2 596 133.391
	(4135.049)	(634 698.589)
‘Geographical Location‘New Brunswick	−73 689.343	3 880 833.151
	(4007.291)	(579 001.545)
‘Geographical Location‘Newfoundland and Labrador	−62 722.733	2 035 618.468
	(3952.200)	(560 082.330)
‘Geographical Location‘Nova Scotia	−66 254.944	3 517 200.998
	(4006.849)	(575 077.618)
‘Geographical Location‘Ontario	−3990.664	2 107 020.605
	(4102.720)	(586 502.528)
‘Geographical Location‘Prince Edward Island	−84 204.393	3 738 771.008
	(4067.718)	(625 960.320)
‘Geographical Location‘Quebec	−55 405.063	3 876 947.689
	(4007.010)	(601 311.925)
‘Geographical Location‘Saskatchewan	−44 722.266	2 727 094.301
	(4009.997)	(592 831.728)
Year × ‘Geographical Location‘British Columbia		−1603.406
		(282.684)
Year × ‘Geographical Location‘Manitoba		−1325.160
		(317.298)
Year × ‘Geographical Location‘New Brunswick		−1978.226
		(289.664)
Year × ‘Geographical Location‘Newfoundland and Labrador		−1049.480
		(280.150)
Year × ‘Geographical Location‘Nova Scotia		−1792.449
		(287.669)
Year × ‘Geographical Location‘Ontario		−1055.467
		(293.522)
Year × ‘Geographical Location‘Prince Edward Island		−1911.819
		(312.972)
Year × ‘Geographical Location‘Quebec		−1967.060
		(300.812)
Year × ‘Geographical Location‘Saskatchewan		−1386.376
		(296.685)
Num.Obs.	357	357
R2	0.845	0.874
R2 Adj.	0.841	0.867
AIC	7990.5	7935.2
BIC	8037.0	8016.7
Log.Lik.	−3983.236	−3946.617
F	188.922	122.937
RMSE	17 227.00	15 753.80

Table 4: Comparing RMSE between two models

Model 1	Model 2	Dataset
17227.0	15753.8	train
17293.8	15052.8	test

The RMSE measures how far the predicted values of the multiple linear regression model are from their actual values on average. The lower the RMSE, the better the model preforms regarding prediction. Based on Table 4, the RMSE for both two models are very similar for training and testing dataset, indicating that the dataset is unbiased and the two models are performing as expected. Moreover, Model 2 has lower RMSE than Model 2 for both training and testing dataset, indicating that Model 2 predicts data more accurately.

B Model Assumption Check

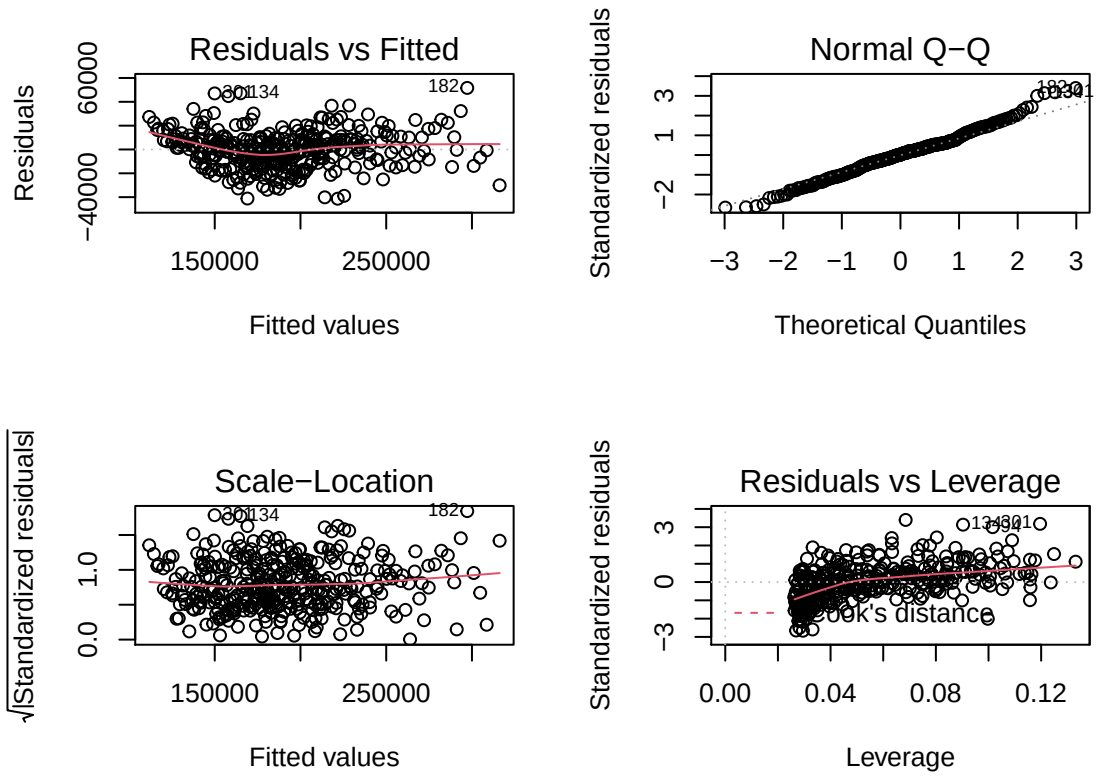

Figure 5: Checking the assumptions of linear model

Check for assumption: Assumptions to the multiple linear regression model are checked to ensure that the model is valid for this dataset. Figure 5 shows the plots used to check for the assumption.

The Residuals vs Fitted plot checks for the linear relationship assumption. Since the red line is almost horizontal and there isn't any pattern, the model satisfies the linearity assumption. The Normal QQ plot checks for the residual normality assumption. Since almost all the dots are on the dashed line, the residuals follow a normal distribution. The Scale-Location plot checks for the homoscedasticity assumption. The red line is horizontal and the dots are evenly scattered, indicating that the variance of the residuals is constant (Kassambara 2018).

C 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 was created to analyze the relationship between time period, geographical location, and the income of the households in Canada.
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.*
   - The 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 os the dataset represent the people who are selected to take the survey, these people are respectively grouped by year, region, family type, income type.
2. *How many instances are there in total (of each type, if appropriate)?*
   - The original dataset has 173043 instances. After it is cleaned, it contains 9395 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 does contain all possible instances.
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.*
   - The relationships between individual instances are explicit. Some instances represent the same group of people, summarize different income statistics(average, highest decile, lowest decile, etc).
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.*
   - No.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - No.
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 is extracted from the Canada Income Survey responses and is validated.
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?*
    - The data was collected through manual human curation, by conducting online surveys to the sampled respondents.
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 respondents of the surveys were drawn from stratified probability sampling.
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)?*
    - The telephone interviews prior to the official survey were conducted by interviewers working in regional offices. The surveys which were online have data collected automatically.
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 data was collected annually from 1976 to 2020. Each round of survey was lasted for 6 consecutive months.
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.
   - No.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - I obtained the data via the Open Government Portal of Statistics Canada.
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.*
   - Yes. The repondents of the surveys were notified via a telephone interview that they were asked to complete an online servey.
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.*
   - Yes, the response to the survey was voluntary.
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).*
   - It doesn't seem like the individuals were provided with a mechanism to revoke their consent in future, because the survey data is made public already.
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 and its use on data subjects was not conducted.
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 dataset hasn't been used for any other tasks.
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.*
   - Repository for this paper that used the dataset: https://github.com/Alicia-y/Telling-stories-with-data-final-paper
3. *What (other) tasks could the dataset be used for?*
   - Analyze the difference in income level between income types and family types.
4. *Is there anything about the composition of the dataset or the way it was collected and prepro-*

cessed/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?

- No.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - Should not be used for disclosure of personal information.
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.*
   - The dataset is public and available to everyone. The Open Government License granted everyone's rights to access and use the data.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - The dataset will be distributed by Github.
3. *When will the dataset be distributed?*
   - The dataset will be distributed on April 27th, 2022.
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) granted the rights to use and modify the data with several exemptions.
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.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - The owner of the repository where the dataset is in.
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 original dataset will be updated annually. Followed by that, the cleaned dataset in the github will be updated.
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.*
   - No.

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.*
    - Yes, the older versions of the dataset will be maintained annually by adding more instances because the Canada Income Survey is conducted annually.
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.*
    - No.
8. *Any other comments?*
    - No.

References

Arel-Bundock, Vincent. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready.* https://vincentarelbundock.github.io/modelsummary/.

Bevans, Rebecca. 2020a. *Understanding Confidence Intervals | Easy Examples & Formulas.* https://www.scribbr.com/statistics/confidence-interval/#:~:text=A%20confidence%20interval%20is%20the,another%20way%20to%20describe%20probability.

———. 2020b. *Understanding P-Values | Definition and Examples.* https://www.scribbr.com/statistics/p-value/#:~:text=The%20p%2Dvalue%20is%20a,to%20reject%20the%20null%20hypothesis.

Canada, Statistics. 2020. *Canadian Income Survey - 2020 (Cis).* https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5200.

Economics, Canadian. 2022. *Canadian Income Inequality Is Canada Becoming More Unequal?* https://www.conferenceboard.ca/hcp/hot-topics/canInequality.aspx.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://github.com/sfirke/janitor.

Funds, Dynamic. 2018. *The Tax System in Canada.* https://dynamic.ca/eng/snapshots/newcomer/newcomer_taxsystem.html#:~:text=In%20Canada%2C%20the%20tax%20system,steps%2C%20or%20%22brackets.%22.

Kassambara, Alboukadel. 2018. *Regression Model Diagnostics.* http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/.

Kellen, Dave. 2010. *Is There Any Reason to Prefer the Aic or Bic over the Other?* https://stats.stackexchange.com/questions/577/is-there-any-reason-to-prefer-the-aic-or-bic-over-the-other.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.* https://www.tidymodels.org.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Scheffler, Samuel. 2020. *Is Economic Inequality Really a Problem?* https://www.nytimes.com/2020/07/01/opinion/economic-inequality-moral-philosophy.html#:~:text=Enough%20economic%20inequality%20can%20transform,society%20ruled%20by%20the%20rich.&text=Large%20inequalities%20of%20inherited%20wealth,and%20undercuts%20equality%20of%20opportunity.

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, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://dplyr.tidyverse.org,%20https://github.com/tidyverse/dplyr.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.* https://yihui.org/knitr/.

Zhu, Hao. 2021. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* http://haozhu233.github.io/kableExtra/,%0Ahttps://github.com/haozhu233/kableExtra.