You got Money, You get Citizenship: A Study of Income Distribution among Immigrants Admission Categories in Canada*

Aviral Bhardwaj

April 19, 2024

This paper investigates the relationship between income levels and immigrants' admission categories on citizenship rates in Canada from 2001 to 2021. Utilizing data from Statistics Canada, the study explores the influence of income distribution among top six immigrants admission categories. Results reveal a notable affinity towards the `Economic Immigrants` category, with the highest median income. Positive correlations between immigrant admissions and income signify Canada's success in attracting economically impactful individuals. This research sheds light on Canada's immigration policy implications amidst socio-economic challenges.

Table of contents

*Code and data supporting this analysis is available at:https://github.com/Aviral-03/canada_immigrant_income_distribution/tree/main

1 Introduction

Recently, Immigration, Refugees, and Citizenship Canada (IRCC), along with Canada's liberal government, announced to "stop ramping up immigration amidst challenges such as high inflation and a housing crisis" (Lone 2023). Still, Canada continues to stand out as a top destination for immigrants, attributed to its exceptional quality of life and robust economy. By prioritizing immigrants with high qualifications and skills, Canada is effectively addressing labor market demands, as evidenced by the fact that 6 in 10 immigrants are selected for their positive economic impact" (Immigration and Canada 2020). Moreover, this approach serves to enhance trade relations, bolster the education system through the contributions of international students (e.g., International students contribute more than $21 billion to the economy annually (Immigration and Canada 2020)), and attract foreign investment, thereby further fortifying the economy.

The question of whether reducing the influx of immigrants can effectively address certain challenges arises, especially considering Canada's substantial reliance on immigrants to fuel its economy. One plausible explanation could be that Canada tends to grant citizenship to immigrants with higher income levels, with affinity to certain admission groups. To explore this hypothesis, this paper investigates the influence of income levels among various immigrants admission categories (our estimand) on the citizenship rate in Canada from 2001 to 2021.

Specifically, the focus is on the citizenship rate, defined as the proportion of immigrants who acquire Canadian citizenship after meeting the residency requirement, and the income profiles within these cohorts. Income serves as a key indicator and our encompassing metrics such as

average employment income, median wages, and salaries. To provide a comprehensive analysis, the study examines the income of four main immigrants admission category "Economic immigrant", "Immigrant sponsored by family", "Refugee", and "Other Immigrant".

The paper utilized various data-set from StatCan, in particular 2021 Census Population data to analyze the income distribution among admission categories, and Table 43-10-0010-01 (Statistics Canada 2019) from the Canadian Census to further examine the change in income profiles. This study reveals a significant Canada's Immigration affinity towards "Economic Immigrant, principal applicant" category, which has the highest median income among all admission categories. The study also shows a positive correlation between the total number of immigrants admitted and their income, suggesting that Canada has been successful in attracting people with income to the country.

The paper is structured as follows: Section 2 provides an overview of the data cleaning process, variables of interest and measurement used in the analysis, and tables of observation. Section 3 presents the Poisson and regression models used to estimate the relationship between the total number of immigrants and variables of interest. Section 4 presents the results of the analysis, focusing on the trends in citizenship rates and income levels. Section Section 5 discusses examination of how Canada continues to experience a substantial influx of economic contributions. Finally, Section Section 6 concludes the paper with weakness of the paper and potential extension for future research.

2 Data

The data utilized in the paper is entirely collected from StatCan (Warin and Le Duc 2024), which is accessible to the public. Specifically, Statistics Canada's 2021 Census of Population dataset titled "Total - Admission category and applicant type for the population in private households - 25% sample" is employed to analyze the income distribution among admission categories.

Table 43-10-0010-01 (Statistics Canada 2019) from the Canadian Census is further used to examine the changes in income profiles of immigrants from 2001 to 2016. This table encompasses 1536 observations and 4 variables, including income profiles categorized by admission category, total count, total count with income, and median income.

The data underwent cleaning and analysis using the R programming language (R Core Team 2023). Cleaning was performed with the `tidyverse` package (Wickham et al. 2019), involving the removal of unnecessary columns for the analysis. Subsequently, analysis was conducted utilizing the `dplyr` package (Wickham et al. 2023). Other packages used include `readr` (Wickham, Hester, and Bryan 2023) for reading and importing data, `janitor` (Firke 2023) for cleaning data, `rstan` (Goodrich et al. 2024) for fitting the model, and `modelsummary` (Arel-Bundock 2022) for summarizing the model. The data were then visualized using the `ggplot2` package (Wickham 2016).

2.1 Variables of Interest

2.1.1 Immigrant Admission

The data from the four censuses (i.e., 2001, 2006, 2011, and 2016) were used to calculate citizenship rates among immigrants who met the residency requirements for citizenship during the five years preceding each census. For this analysis, we focused on adult immigrants aged 15 years or above who arrived in Canada between five and nine years before each census. For example, immigrants granted citizenship in 2001 arrived in Canada between 1991 and 1996, making them part of the 2001 census.

To maintain consistency in the study populations across censuses, we defined the lower sample limit as immigrants who had resided in Canada for at least five years.

The data collection process is as follows: Before the 2021 Census, naturalization among immigrants was determined by the question "Of what country is this person a citizen?" It offered two options: "Canada, by birth" and "Canada, by naturalization," along with a write-in box for specifying other countries. In the 2021 Census, two questions are asked: "Is this person a Canadian citizen?" and "Is this person a citizen of a country other than Canada?" The questionnaire clarifies that naturalization refers to the process by which an immigrant obtains Canadian citizenship under the Citizenship Act. This question was asked in the IRCC Immigrant Application, filled out by every new or recent immigrant applying for permanent residency. The data were collected by the IRCC and Statistics Canada, and the citizenship rate was calculated based on the number.

2.1.2 Admission Category

'Admission category' refers to the name of the immigration program or group of programs under which an immigrant has been granted for the first time the right to live in Canada permanently by immigration authorities. For our analysis, we will be focusing on the following admission categories:

- `Total, immigrant admission category`: Represents the total number of immigrants admitted to Canada by combining all admission categories.
- `Economic Immigrant, principal applicant`: This category includes immigrants who have been selected for their ability to contribute to Canada's economy through their ability to meet labour market needs. This also include skill-worker and caregiver programs.
- `Economic Immigrant, spouse and dependent`: This category includes the spouse and dependent children of the principal economic applicant.
- `Immigrants sponsored by family`: This category includes immigrants who were sponsored by a Canadian citizen or permanent resident and were granted permanent resident status based on their relationship either as the spouse, partner, parent, grandparent, child or other relative of this sponsor.

- `Refugee`: This category includes immigrants who were granted permanent resident status based on a well-founded fear of returning to their home country.
- `Other immigrant`: This category includes immigrants who were granted permanent resident status under a program that does not fall in the economic immigrants, the immigrants sponsored by family or the refugee categories.

We choose these categories as they represent the primary groups of immigrants admitted to Canada, as defined by the IRCC, in Census 2021 data. Below Table 1 presents the sample of total number of Immigrants accepted in Canada from 2001 to 2016, categorized by admission category. There were in total 23 admission categories, but for the purpose of this analysis, we have selected the above six categories as they are the most relevant to our research question.

Table 1: Total # of Immigrants by Admission Cateogry (in thousands), Canada, 2001-2016

Year	Total Immigrant Admission	Principal Economic Immigrant	Secondary Economic Immigrant	Immigrants Sponsored by Family	Refugee	Other
2001	7800	8900	4200	8800	11900	18500
2002	8100	9400	4800	8400	11900	14000
2003	7500	8500	4300	7300	13600	15000
2004	8300	9700	4800	7400	13300	17200
2005	8500	10300	5000	7100	14400	17900
2006	9500	13000	5800	8200	12600	17200
2007	9900	14500	6000	8500	12100	18500
2008	10700	17300	6200	8600	11600	17400
2009	11000	18700	6700	8100	11500	18900
2010	11100	17400	6800	8200	13000	21100

2.1.3 Income

Total income is "sum of certain incomes (in cash and, in some circumstances, in kind) of the statistical unit during a specified reference period" (Government of Canada, Statistics Canada 2023). This survey encompasses various statistical units, including persons, private households, census families, economic families, enterprises, companies, establishments, locations, and farm operators and families.

For individuals, total income encompasses receipts from specific sources, prior to income taxes and deductions, over a designated reference period. For this analysis, the reference period is the calendar year 2021, and the data was collected from the Canadian Revenue Agency (CRA), where all tax returns are compiled and stored, however, anonymized for privacy reasons.

For the purpose of analysis, from `Table 43-10-0010-01` (Statistics Canada 2019) we are interested in the following variables:

- `Total with income`: Represents the total number of immigrants who were granted citizenship reported income in thier CRA tax return.
- `Median with income (in $CAD)`: Represents the median income of all

After further cleaning process, below Table 2 presents the total immigrants admitted to Canada from 2001 to 2016, along with the number of people with income, Median Income (in $CAD), and Mean Income (in $CAD). For our analysis we are only interested in the first three variables, as they are the most relevant to our research question. We will not be accounting Mean Income (in $CAD) in our analysis, because it can be very sensitive to outliers and may not be a good representation of the data (Table 4 presents how for certain admission categories, mean income can be biased).

Table 2: Total Immigrants Admission by Income, Canada, 2001-2016

Year	Total Admission	Total Admission with Income	Median Income (in $CAD)	Mean Income (in $CAD)
2001	153850	91120	14600	7800
2002	139590	82030	14400	8100
2003	133755	75315	14500	7500
2004	142090	85085	16800	8300
2005	154625	96250	16200	8500
2006	154640	98615	18100	9500
2007	145895	94800	19900	9900
2008	151290	97585	21100	10700
2009	155340	94445	22100	11000
2010	169745	104555	21500	11100

To analyze the income distribution among admission categories, we utilized the 2021 Census Population data. The data was transposed to provide a clear view of the income distribution among admission categories. The income groups range from less than $20,000 to $150,000 and over, with 7 income groups in total. Below Table 3 presents the total number of immigrants accepted in each category, categorized by income group.

Table 3: Admission Category and Applicant Type

Income	Total Immigrant Admission	Principal Economic Immigrant	Secondary Economic Immigrant	Immigrants Sponsored by Family	Refugee	Other
$< $20,000 (including loss)	4674240	139745	412325	305240	174770	14950

Income	Total Immigrant Admission	Principal Economic Immigrant	Secondary Economic Immigrant	Immigrants Sponsored by Family	Refugee	Other
$20,000 to $39,999	5669385	283450	447885	471040	27737	523130
$40,000 to $59,999	4655070	307055	315825	353100	17248	015585
$60,000 to $79,999	3001050	209950	184740	180135	83375	7645
$80,000 to $99,999	1903800	141795	110740	92800	41830	3545
$100,000 to $149,999	1877265	162755	109875	82705	34285	3040
$150,000 and over	909600	80325	47900	33915	12020	1285

2.2 Relationship between Income Distribution by Admission Category

In order to get a better understanding of the relationship between income distribution and admission category, we present the income distribution among admission categories in 2021. The data is presented in Figure 1, with the x-axis representing the income group and the y-axis representing the total number of Immigrants accepted in each income group. The line plot is color-coded by admission category, allowing for easy comparison between the different categories.

Figure 1 indicates a skewed distribution, with the majority of immigrants falling within the income group of $20,000 to $39,999. Specifically, the category "Economic Immigrant, spouse and dependent" exhibits the highest number within this range. Notably, the "Economic Immigrant, principal applicant" category dominates the income group of $150,000 and over. This suggests a prioritization of high-income immigrants, particularly among economic immigrant categories, in Canada's immigration policies.

2.3 Relationship between Admission Category and Income Change (2001 - 2016)

Figure 2 illustrates the total number of immigrants admitted to Canada from 2001 to 2016, compared with the total number of immigrants with income. The graph shows a steady increase in both the total number of immigrants admitted and the total number of immigrants with income over the years. This suggests that until 2016 Canada has been successful in attracting people with income to the country, with steady increase in the citizenship rate. It is interesting to note that Total Admission with Income accounts for as much as 60% of the Total Admission, suggesting that a significant proportion of Immigrants admitted to Canada have income.

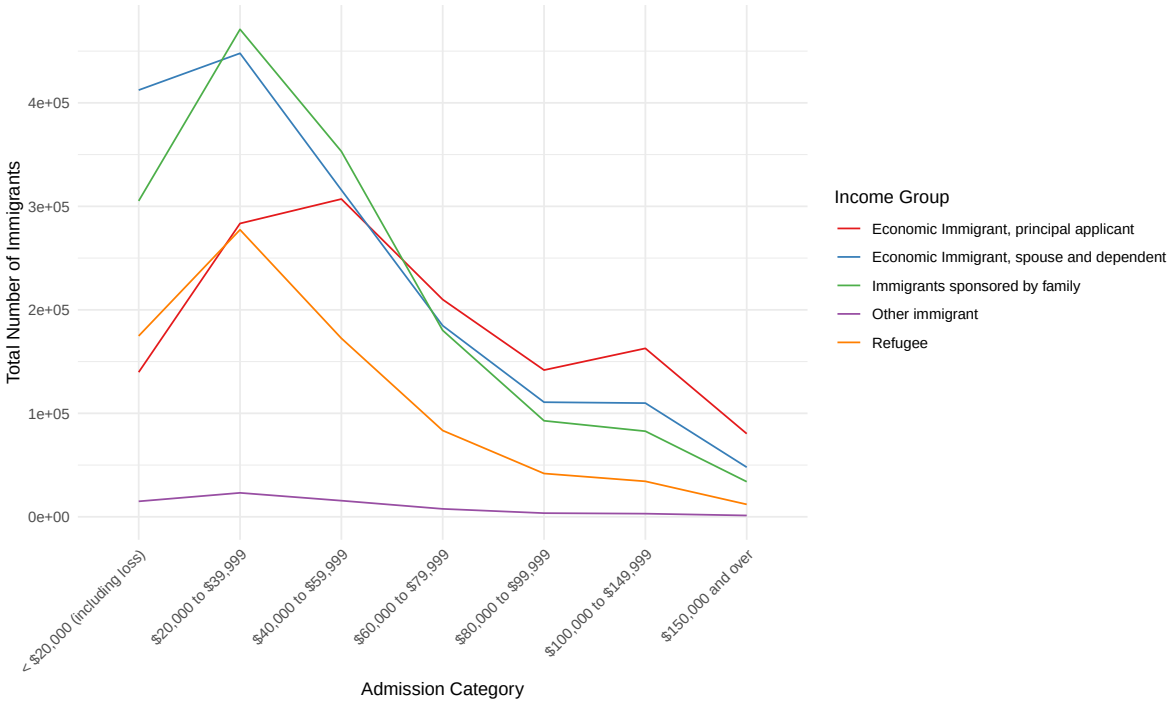

Figure 1: Immigrants Income Distribution by Admission Category (2021)

To better understand the variables of interest, we present summary statistics in Table 4. This table depicts the percentage of total count and count with income for each admission category, relative to the total immigrants admitted to Canada. Additionally, it showcases the mean median income for each admission category, along with the percentage change in median income from 2001 to 2016.

Positive percentages in the "Median Income (% Change)" column denote an increase in median income, while negative percentages indicate a decrease. This metric aids in understanding income trends over time across different admission categories.

Notable changes in median income are evident across various admission categories. For instance, the "Economic Immigrant, principal applicant" category experienced a substantial 250% increase in median income, signifying significant growth over the years. Following this trend, it still represented 16% of the total immigrant intake, with as high as 20.31% for immigrants with income.

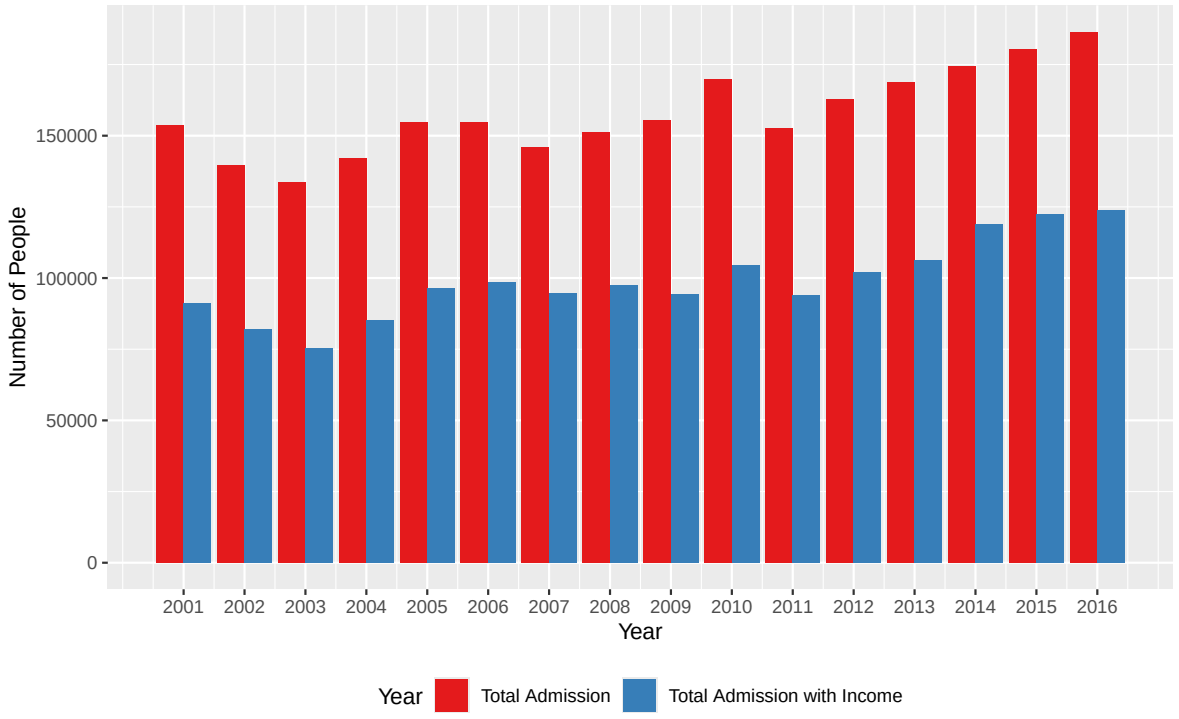

Figure 2: Total Immigrant Admission vs Total Immigrant Admission with Income (2001 - 2016)

Table 4: Summary Statistics: Income Interest by Admission Category

Admission Category	Total Count (%)	Total with Income (%)	Mean of Median Income	Median Income (% Change)
Total, immigrant admission category	100.00	100.00	11431.25	120.51282
Economic Immigrant, principal applicant	16.19	20.31	17725.00	250.56180
Immigrants sponsored by family	16.00	14.14	8493.75	23.86364
Economic Immigrant, spouse and dependent	11.32	10.33	6893.75	169.04762
Refugee	5.71	4.28	12412.50	-26.05042
Other immigrant	0.78	0.94	19150.00	27.02703

Finally, in order to gain better understanding of the underlying data, Table 5 provides an overview of the mean and standard deviation of each variable of each variable (i.e Total Im-

Table 5: Mean and Standard Deviation of each variable, for all Admission Category

Variable	Mean	SD
Total Immigrants	842095.83	878846.399
Total Immigrants with Income	529085.83	562826.095
Median Income (in $CAD)	12684.38	4897.941

migrants, Total Income Immigrants, Median Income (in $CAD)) by admission category. The data are presented in thousands. We can thus use this table to understand the distribution of the data and the variability of the data for each admission category, and create statistical models.

3 Model

Based on our data analysis, we found a correlation between the total number of immigrants, their admission category, and their income. To delve deeper into this relationship, we can employ Poisson and regression models. However, it's crucial to note that the Poisson regression model assumes equal variance and mean, but, we will only be focusing on the admission category and Total Immigrants for that category.

3.1 Model Setup

We created a Poisson regression model in R (R Core Team 2023), utilizing the `rstanarm` package (Goodrich et al. 2024). The model aims to estimate the relationship between the various admission categories (Categorical data) and the total number immigrants in each category (Count data). We have allowed auto-scaling for the priors, which are set to normal distributions with a mean of 0 and a standard deviation of 2.5. The model also includes a prior for the intercept, which is set to a normal distribution with a mean of 0 and a standard deviation of 2.5. The model is fitted using the `stan_glm` function, with the family set to `neg_binomial_2` and the link function set to `log`.

3.2 Model Description

3.2.1 Poisson Regression Model

Define y_i as the log count of total immigrants for Admission category i. Then admission category$_i$ is the applicant type of the immigrant:

$$y_i|\lambda_i, \alpha \sim \text{Poisson}(\lambda_i, \alpha) \tag{1}$$
$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{Admission Category}_i \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 6.21) \tag{4}$$
$$\alpha \sim \text{Exponential}(1) \tag{5}$$

In Model 1:

- y_i is the log count of total immigrants for Admission category i
- β_0 is the coefficient for the intercept
- β_1 is the coefficient for the Admission Category variable

3.2.2 Model Parameters

The independent variable Admission Category is a categorical variable with four levels:

- Economic Immigrant, principal applicant
- Economic Immigrant, spouse and dependent
- Immigrants sponsored by family
- Refugee
- Other Immigrants

The dependent variable "Total Count" is the total number of immigrants granted citizenship in each Admission Category.

3.3 Model Justification

In our analysis, our primary goal was to understand the relationship between admission categories and the total number of immigrants admitted into Canada. Given the nature of our data, which involves count data (the number of immigrants) and categorical independent variables (admission categories), we initially considered employing a Negative Binomial regression model. The Negative Binomial regression model is well-suited for count data that exhibit overdispersion, a common characteristic where the variance of the data exceeds the mean.

Upon conducting our analysis, we found that both the Poisson and Negative Binomial regression models yielded similar coefficient estimates. This observation suggested that both models could effectively capture the relationship between our independent and dependent variables. However, as the Poisson regression model is a simpler and more interpretable model, and since it provided comparable results to the Negative Binomial model, we ultimately chose to utilize the Poisson regression model for our analysis.

By selecting the Poisson regression model, we aimed to provide a more straightforward interpretation of the relationship between admission categories and the total number of immigrants while still adequately addressing the overdispersion observed in the data.

4 Results

Table **?@tbl-modelresults** presents the summary of the Poisson regression model, and multiple regression model.

In the above table, "Economic Immigrant, spouse and dependent" serves as the reference category and is omitted from the table. In this baseline model, we observe positive coefficients for the admission categories "Immigrants sponsored by family" and "Economic Immigrant, principal applicant". Specifically, $\beta_1 = 1.49$ suggests that being an "Economic Immigrant, Principal Applicant" is positively associated with the log count of total immigrants compared to the reference category. This positive association indicates that individuals falling under this admission category are more likely to be granted citizenship compared to those in the reference category and other categories with negative coefficients. This relationship aligns with the findings in Table 4, where we observe that the percentage of total count for "Economic Immigrant, principal applicant" is 16%, which is the highest among all admission categories. Additionally, the admission category "Immigrants sponsored by family" has a positive coefficient of 0.35, indicating a positive association as well, which accounted for 16% of the total count as well as observed in Table 4.

We also see that Other Immigrant has a negative coefficient of -2.67, indicating that this category has a negative association with the log count of immigrants, consistent with 0.78% of the total count. This suggests that individuals in this category are less likely to be granted citizenship compared to the reference category.

Figure 3 presents intriguing insights into the Median Income (in $CAD) by Admission Category from 2001 to 2016. The graph reveals a positive correlation between the year and median income, with incomes steadily rising over the period. This trend aligns with expectations, as economic growth typically corresponds with higher median incomes among immigrants.

Additionally, the trend lines for each admission category provide further clarity, indicating the direction of their respective income trends. Notably, the line for "Economic Immigrant, principal applicant" stands out, consistently surpassing other categories and even exceeding the median income for the "Total, immigrant admission category," reaching $30,000 CAD. This observation suggests that principal applicants are significantly above the average income level, reflecting Canada's strategy of attracting high-income immigrants. Close behind is "Immigrants sponsered by family", whose Median income has also increased over the years, reaching $25,000 CAD in 2016. Similar results were also observed Figure 1, where "Immigrants sponsored by family" was ranked second in terms of income distribution among admission categories.

Note that "Refugee" and "Other Immigrants" almost never exceed the median income of the $15,000 CAD, indicating that these categories have lower median incomes compared to other admission categories, and do not positively contribute to the economy.

These findings can be cross-referenced with the Poisson regression model, which indicates similar results.

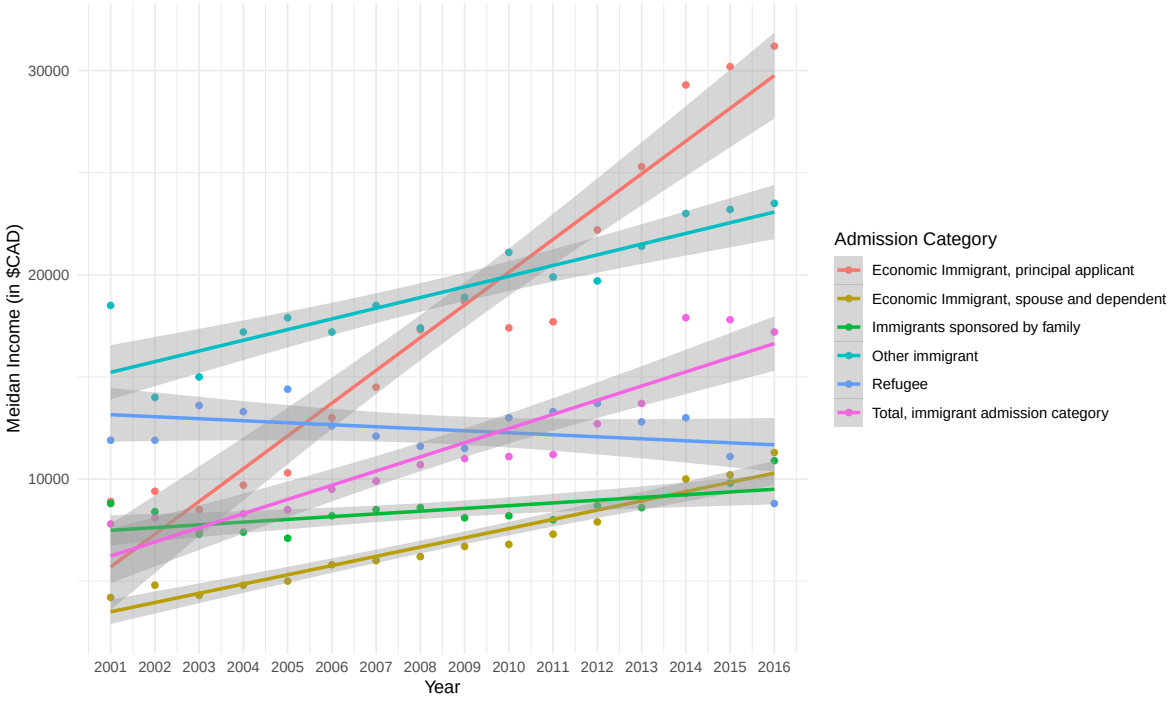

Figure 3: Immigrants Income by Admission Category 2001 - 2016

By utilizing this model both of these models we can estimate that some categories have a higher likelihood of being granted citizenship compared to others. Using the regression model, we have identified `Median Income` as a significant predictor that influences the chances of being granted citizenship.

5 Discussion

In this paper, we analysed the trends between the citizenship rate among various Immigrant admission category and their income levels in Canada from 2001 to 2016. In addition, to that we also examined the income distribution among admission categories in 2021, allowing us to understand the correlation between these variables of interest. Here are some of the key finding that we observed:

5.1 Key Finding between Income Distribution and Admission Category

The analysis reveals a positive correlation between income distribution and admission category, impacting the total number of citizenships granted to immigrants. In Section 2, it's evident that the "Economic Immigrant" category, including principal applicants and spouses, accounted for the highest number of accepted immigrants across all income distributions. This aligns with Canada's strategic focus on attracting skilled workers and economic immigrants to bolster the economy and address labor market demands, which can be explained by Canada's Provinical Nominiee Program (PNP) (Immigration and (IRCC), n.d.) allows provinces to nominate individuals for permanent residency based on their ability to contribute to the economy. This program is designed to address the specific labor market needs of each province, and it is not surprising that the economic immigrant category is the most popular among all the categories.

An intriguing observation is the prominence of "Immigrants sponsored by family" as the second-highest category for income distributions ranging from $20,000 to $59,999, albeit dropping to the third-highest for income distributions between $60,000 to $99,999. This underscores Canada's preference for wealthier family-sponsored immigrants with higher income levels, despite the emphasis on economic contribution in immigrant selection criteria. This trend is consistent with the fact that family sponsorship remains a significant pathway to recieveing citizship and bringing foreign investment directly into Canadian bank.

5.2 Key Finding between Citizenship Rate disparity between Admission Category based on Income

In terms of citizenship distribution, the highest rate is observed among "Economic Immigrant, principal applicant" category, showing a 75% increase from 2001 to 2016. This finding is supported by the Poisson regression model, indicating a higher likelihood of citizenship for individuals in this category compared to others. Additionally, Table 4 reveals that "Economic Immigrants" constitute 27.51% of the total count, the highest among all admission categories. Moreover, the table illustrates that a significant proportion of immigrants granted citizenship also have income, mirroring the total count percentages.

Furthermore, immigrants admitted in 2018 had a median wage of $31,900 in 2019, marking a 4.2% increase from those admitted in 2017. Notably, 2018 admissions had the highest median entry wage reported one year post-admission since 1981. Figure 3 depicts the substantial increase in median income for "Economic Immigrant" category over the years, followed by "Immigrant sponsored by family," corroborating the Poisson regression model's positive coefficients for these admission categories.

It should be noted that Economic principal applicants are more likely to file taxes due to their selection based on their economic contribution potential, thus increasing their likelihood of having higher incomes. Nevertheless, the rising median income across all admission categories

over the years underscores the positive economic contribution, highlighting Canada's successful immigration policies.

6 Weakness and Next Steps

While the paper aims to offer a comprehensive analysis of the trends in citizenship rates and income levels among various Admission Categories and Applicant Types in Canada from 2001 to 2016, several limitations need acknowledgment. Firstly, the data utilized only extends until 2016, potentially limiting the breadth of our analysis regarding current trends. Moreover, while income level by admission category is a significant factor influencing citizenship rates, other variables such as Generation Status, Education, Language, Place of Birth, and Socio-economic Status may also play pivotal roles. While considering these factors could enhance the analysis, it falls beyond the scope of this paper and suggests a direction for future research.

Furthermore, directly estimating the relationship between the number of applicants and median income may not offer a comprehensive analysis of their correlation. This method cannot conclusively determine whether the increase in applicants results from a rise in median income or a bias towards higher-income applicants. Future research should explore additional factors influencing applicant numbers within each admission category and applicant type.

However, despite the declining citizenship rates, the rise in applicants with higher income levels has substantially contributed to the economy, as evidenced by the increase in median income across admission categories and applicant types over the years. Thus, it remains imperative for the Canadian government to persist in attracting skilled workers and economic immigrants to further fortify the economy and address labor market demands.

7 Appendix

7.1 Datasheet

Extract of the questions from Gebru et al. (2021).

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 part of the Canadian Census 2021. The dataset was created to provide information about the population of Canada, and report other population related metrics. Table 43-10-0010-01 was created as a part of Special Interest Tables, which are released as part of the Census Program Data. This combined Canada Census data as well as Longitudinal immigration database (imdb).
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Statistics 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.*
    - Canadian Government
4. *Any other comments?*
    - TBD

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 demographic information, such as population counts, immigration data, and other related metrics for various regions within Canada. It includes details about age, gender, ethnicity, education, employment, and other demographic factors.
2. *How many instances are there in total (of each type, if appropriate)?*
    - In total 63 census subdivisions defined.

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).*

    - This dataset is a sample of the larger set of Canadian Census data. Subdivisions were selected based on the appropriateness of data, such as provinces, territories, and census metropolitan areas subdivision were removed, and other census not important for the analysis was also removed. The selection was based on the need to provide a representative sample of the population which was 25%.

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 likely consists of demographic features such as population counts, age distributions, immigration statistics, education levels, employment rates, and other relevant demographic indicators. These may be presented in tabular format or as structured data. These are all

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

    - Yes, a unique identifier is associated with each instance, which is the census subdivision code. This code is unique to each census subdivision and is used to identify the region to which the data pertains.

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.*

    - No information is missing from the instances. The dataset is complete and contains all the relevant demographic information for each census subdivision.

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.*

    - Yes, a new README file was created to explain the relationships between the instances. The README file provides information about the structure of the dataset, the meaning of each column, and how the data is organized. This helps users understand the relationships between the instances and how they can be used for analysis.

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.*

    - No, the dataset is self-contained and does not rely on external resources. All the data is included in the dataset itself, and there are no external links or dependencies. This makes it easy to use and analyze the data without any additional requirements.

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.*

    - Yes, the dataset contained information regarding the Income of the population, which is considered confidential. Therefore, only the aggregated data was included in the dataset, and no individual-level information was provided. This ensures that the privacy of individuals is protected and that the data is used in a responsible manner.

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 likely identifies subpopulations based on various demographic factors such as age, gender, ethnicity, education level, employment status, and immigration status. These subpopulations are typically identified through the breakdown of demographic statistics for different demographic categories within each geographic region (i.e Provinces in Canada)

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.*

- Direct identification of individuals from this dataset is unlikely since it deals with aggregated demographic statistics at a regional level.

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.*

    - No

16. *Any other comments?*

    - TBD

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 Census Canada dataset was acquired through self-reported survey responses from individuals residing in Canada. Participants provided information about various demographic factors, including age, gender, ethnicity, education, occupation, and household composition. The data collection process involved direct reporting by subjects, and efforts were made to ensure data accuracy through validation and verification procedures by Statistics Canada.

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 collection for the Census Canada dataset primarily relied on manual human curation through survey forms distributed to Canadian residents. Statistics Canada employed various mechanisms, including online surveys, paper questionnaires, and assisted interviews, to collect demographic information from respondents across the country.

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 Census 2021 Canada dataset is not a sample but represents the entire population of Canada. The census aims to collect information from every individual residing in the country, making it a complete enumeration rather than a sample.

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 data collection process for the Census Canada involved various individuals, including enumerators, field supervisors, and support staff hired by Statistics Canada.

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 is aggregated every 5 years. The data is collected in the year of the census.

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.*

    - Statistics Canada adheres to strict ethical standards and undergoes regular reviews to ensure compliance with legal and ethical guidelines governing data collection, storage, and usage.

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

    - No, the data was collected directly from individuals residing in Canada through self-reported survey responses. Statistics Canada is the primary agency responsible for collecting census data in Canada. Data was downloaded from the Statistics Canada website.

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, individuals residing in Canada were notified about the data collection process through various communication channels, including mail, online announcements, and public awareness campaigns. Statistics Canada provides detailed information about the census process, the purpose of data collection, and the importance of participation to ensure accurate and representative data.

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, individuals were asked to provide consent for the collection and use of their data as part of the census process. Statistics Canada ensures that participants are

informed about the purpose of data collection, the confidentiality of their information, and the importance of their participation in the census. Consent is obtained through various means, including online forms, paper questionnaires, and assisted interviews.

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).*

    - Individuals were assured of confidentiality and privacy protections regarding their data, as outlined by Statistics Canada.

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.*

    - Statistics Canada conducts thorough assessments of the potential impact of census data collection on data subjects, including privacy implications and data protection measures, which can be found in their official documentation.

12. *Any other comments?*

    - TBD

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.*

    - Refer to https://www12.statcan.gc.ca/census-recensement/2021/ref/98-304/2021001/chap8-eng.cfm for information regarding the data processing and cleaning.

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.*

    - NA

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

    - A few Structured Query Languages (SQL) and statistical analysis system (SAS) modules are also part of the census edit and imputation processing flow.

4. *Any other comments?*

- TBD

Uses

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

   - As of now, there is no specific information available regarding the specific tasks for which this dataset has been used. However, given its comprehensive nature and the involvement of Statistics Canada, it's likely that researchers, policymakers, and analysts have utilized this dataset for various purposes such as demographic research, population projections, immigration policy analysis, socioeconomic studies, and urban planning.

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.*

   - Stats Canada: https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/about-apropos/about-apropos.cfm?Lang=E

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

   - The dataset could be utilized for a wide range of tasks and analyses, including:
   - Studying population trends and dynamics over time.
   - Assessing the impact of immigration on demographic composition and economic development.
   - Identifying disparities in education, employment, and income across different demographic groups and regions.

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?*

   - No. The data is structured and aggregated in a way that protects individual privacy and confidentiality. However, users should be aware of the limitations of the dataset and the potential biases that may arise from the sampling strategy or data collection methods. It is important to interpret the data in context and avoid making generalizations or assumptions that could lead to unfair treatment or misinterpretation of the results.

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

- The dataset should not be used for individual-level analysis or identification. It is not suitable for making decisions about specific individuals or groups, as it does not contain personal information or detailed demographic data at the individual level. The dataset is intended for aggregate analysis and population-level studies.

6. *Any other comments?*

   - TBD

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 distributed to third parties outside of Statistics Canada for research, analysis, and policy development purposes. Researchers, policymakers, analysts, and other stakeholders may access the dataset through official channels and data repositories.

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 likely be distributed through official channels such as the Statistics Canada website, data repositories, and other authorized platforms. StatsCan r library is available for R users to access the data.

3. *When will the dataset be distributed?*

   - NA

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.*

   - Yes, the dataset will likely be distributed under a specific license or terms of use that govern its usage, distribution, and sharing. Users may be required to comply with certain conditions, restrictions, or fees associated with the dataset. The specific licensing terms and conditions will be provided by Statistics Canada.

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.*

   - No

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, general census is available to the public.

7. *Any other comments?*

    - TBD

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - Statistics Canada

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - Official contact information for Statistics Canada can be found on their website: https://www.statcan.gc.ca/eng/start

3. *Is there an erratum? If so, please provide a link or other access point.*

    - N0. 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)?*

    - Done by Statistics Canada

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.*

    - Any applicable limits on data retention and associated enforcement mechanisms will be specified in accordance with relevant privacy regulations and guidelines.

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.*

    - Old versions of the dataset may be archived and maintained for historical reference, but they may not be actively supported or updated. Any obsolescence of older versions will be communicated to dataset consumers through official channels and documentation.

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?*

    - TBD

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

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

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of ACM* 64 (12): 86–92.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Government of Canada, Statistics Canada. 2023. "Canadian Income Survey - 2021 (CIS)." Surveys and Statistical Programs.

Immigration, Refugees, and Citizenship Canada. 2020. "2020 Annual Report to Parliament on Immigration." https://www.canada.ca/content/dam/ircc/migration/ircc/english/pdf/pub/annual-report-2020-en.pdf.

Immigration, Refugees, and Citizenship Canada (IRCC). n.d. "Provincial Nominee Programs: Apply and Get Nominated." Government of Canada. https://ircc.canada.ca/english/helpcentre/answer.asp?qnum=736&top=6.

Lone, Wa. 2023. "Canada Caps Immigration Target Amid Housing Crunch, Inflation | Reuters." *Canada Caps Immigration Target Amid Housing Crunch, Inflation*. https://www.reuters.com/world/americas/canada-keeps-immigration-target-unchanged-next-2-years-amid-housing-crunch-2023-11-01/.

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

Statistics Canada. 2019. "Immigrant Income by Admission Year and Immigrant Admission Category, Canada and Provinces." Table 43-10-0010-01. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=4310001001.

Warin, Thierry, and Romain Le Duc. 2024. *statcanR: Client for Statistics Canada's Open Economic Data*. https://github.com/warint/statcanR/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.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://CRAN.R-project.org/package=dplyr.

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