Predicting Probability of Participants' Vote of Conservative and Liberty in 2025 Canadian election*

Jingyi Shen

April 18, 2024

In our study, we use binary logistic regression to analyze voter preferences for Canada's Conservative and Liberal parties based on demographic and socioeconomic factors, adjusting the models with the AIC for optimal predictor selection. The Canadian electoral system elects parliamentarians through a multi-step process where, unless a candidate achieves an outright majority, votes are redistributed from the lowest-ranking candidates until one secures 50%. The Liberal Party plans to increase its refugee quota to 15% of all immigrants and admit 500,000 migrants annually by 2025. The Conservative Party emphasizes economic growth and attracting skilled immigrants. Our hypothesis predicts a likely Conservative victory in the 2025 elections, with the Liberals closely behind, based on trends and electoral mechanics. This analysis uses post-stratification to enhance the representativeness and accuracy of our predictions.

Table of contents

*Code and data supporting this analysis are available at: https://github.com/CSCmaster/Final-Project

1 Introduction

1.1 Background

The Canadian electoral process, epitomizing democratic principles, empowers citizens to elect parliamentarians who shape Parliament's composition and, consequently, the government formation. In the recent 2021 federal election, Justin Trudeau's Liberal Party secured a majority in the 44th Parliament(Nakhaie 2006). Voters' primary concerns include the cost of living, healthcare, climate change, post-pandemic recovery, and immigration, with dissatisfaction towards Trudeau's government prevalent. Trudeau's popularity lags behind UK's Poilievre(Nakhaie 2006). The government's ambitious immigration plan aims to admit 500,000 migrants annually by 2025, albeit facing a backlog. The Conservative Party, with a focus on economic prosperity, emphasizes attracting highly skilled immigrants while addressing concerns like living costs and housing issues(Elections 2023).

1.2 Hypothesis

To enhance control and prevention of illegal immigration, the Liberal Party has suggested modifications to the Immigration and Refugee Protection Act. They propose establishing an annual refugee quota of 250, specifically aimed at supporting human rights defenders, journalists, and humanitarian workers seeking safety and resettlement in Canada. Looking ahead, the Liberal Party intends to incrementally increase this refugee quota, aiming for refugees to comprise 15% of all immigrants. Conversely, the share of economic immigrants is planned to decrease to 58%, while the family reunification immigrant ratio will remain constant. These proportions are expected to be maintained over the next two years. Moreover, the Canadian national government recently unveiled an ambitious plan to admit 500,000 migrants annually by 2025, looking for nearly 1.5 million newcomers in Canada within the next three years (Canada's Immigration Levels Plan 2024-2026). Historically, the Liberal government's immigration strategy tends to be more permissive, offering broader entry routes compared to that of the Conservative Party(Nakhaie 2006). Based on our analysis of historical trends, governmental policies, and electoral procedures, our hypothesis aligns with the prevailing expectation that the Conservatives will emerge as the likely winners in the 2025 election. **We anticipate that the Liberals will likely secure the highest number of votes, with the Conservatives expected to closely follow in second place.** Our research indicates that the vote-counting process in Canada significantly influences the outcome, as voters do not directly elect a specific prime minister candidate. Initially, all votes are tallied based on first preferences. If a candidate obtains 50% of the vote, they win. If not, the candidate with the fewest votes is eliminated(Elections 2023). Then, votes from supporters of the eliminated candidate are redistributed to their second-choice candidate. This iterative redistribution continues until one candidate achieves 50% of the vote. Hence, our objective is to forecast the election probabilities for the Conservative and Liberal parties and compare which is higher(Elections 2023).

1.3 Terminology

In this paper, I used binary logistic regression models to forecast the likelihood of individuals affiliating with the Conservative or Liberal parties.Binary logistic regression is a statistical technique tailored for predicting binary outcomes, where the variable under consideration has two possible results. In our study, these outcomes represent the probability of individuals aligning with each of the two parties.Each party's likelihood will be predicted through separate logistic regression models, with the response variables indicating whether individuals vote for the corresponding party or not. The predictors used in these models include age, province, education, religion, and income before tax. The variable "sex" will be initially included but later removed after AIC model testing, as further elaborated in the model selection section. This regression approach is particularly advantageous when there's a nonlinear relationship between independent and dependent variables, especially in scenarios with categorical dependent variables like ours. Additionally, we will incorporate post-stratification in our analysis to enhance accuracy. Post-stratification helps refine precision estimates by ensuring that the sample is more representative of the entire population. We utilize this technique to predict the probability of voter shares for the Liberal or Conservative parties in the upcoming election.

2 Data

2.1 Raw Data

In both datasets, we've focused on standardizing variables such as age, sex, province, education, religion, and pre-tax income. Conducting a census typically involves higher expenses and more time since it requires reaching out to every individual in the population. Surveys, however, offer a more cost-effective and quicker alternative, especially for large populations, by collecting data from a select group that represents the larger whole.

Census Data: Census data refer to information systematically gathered from the entire population of a defined area at a specific point in time. The census data we have comes from the General Social Survey (GSS), conducted on August 12, 2022. We've made necessary adjustments to ensure it matches our survey data. Data collection aims to provide a detailed snapshot of demographic, economic, and social characteristics.The GSS is a nationwide survey program designed to collect data on societal trends, behaviors, and attributes across the entire population, offering insights into the changing dynamics of communities(GSS 2022).

Survey Data: Survey data, are collected from a sample of the population rather than every individual. This method is often used when it's impractical or unnecessary to include everyone. Survey data aim to infer the characteristics of the larger population based on the responses of the sample. The survey data, obtained from the Canada Election Study (CES) in 2023, includes responses from over 37,000 participants. The CES focuses specifically on voter behavior, attitudes, and the electoral process within the Canadian context. Unlike the GSS, which aims for a comprehensive overview, the CES targets specific topics of interest to derive insights from a segment of the population(CES 2023).

2.2 Data analysis Tools

In this research,the arrow(Richardson et al. 2024)package significantly impacts the way large datasets are handled and processed. It provides high-performance reading and writing of data in the Arrow file format, including support for Parquet files, which are highly efficient for storing and querying large datasets. Tidyr(Wickham, Vaughan, and Girlich 2024) helps in tidying data, meaning it makes it easier to structure datasets so that they are straightforward to work with. It provides functions to transform data into a tidy format, where each variable forms a column, each observation forms a row, and each type of observational unit forms a table. The mass package(Venables and Ripley 2002), It provides a wide range of statistical techniques including linear and nonlinear modeling, statistical tests, time series analysis, classification, and clustering. The package is known for its functions to fit generalized linear models, among many other tools. With summarytools(Comtois 2022), users can easily generate frequency tables, descriptive statistics summaries, cross-tabulations, and more. It makes exploratory data analysis more efficient and is particularly useful for preliminary data analysis, ensuring that researchers and analysts can understand their data before moving on to more complex analyses. Data analysis was conducted using the R programming language (R Core Team 2022), renowned for its open-source nature and robust statistical analysis capabilities. Visualization complexities were addressed with the ggplot2 package (Wickham 2016), which supports the creation of intricate graphics. For data manipulation, dplyr(Wickham et al. 2022) was employed, providing a streamlined grammar that simplifies dataset filtering, summarization, and reorganization. Fast and efficient data importing was achieved through the use of the readr package(Wickham, Hester, and Bryan 2022). The process of generating this report was seamlessly managed by knitr(Xie 2014), facilitating the embedding of R code directly within the text.

2.3 Variable Description

2.3.1 Survey Data Variable

Age: The cps21_age field from the survey data denotes participants' ages at the time they completed the General Social Survey. This variable requires no alterations but should be renamed to "age" to ensure consistency with the census data's corresponding variable.

sex: In the survey dataset, cps21_genderid identifies 9474 respondents as male, 11370 as female, 90 as non-binary, and 34 as other. Given the relatively small number of non-binary and other responses, distribute these respondents between male and female categories based on existing proportions, with females at 54.22% and males at 45.18%.

province: The survey data's cps21_province includes three additional provinces not present in the census data, creating discrepancies. To align the datasets, remove Northwest Territories, Nunavut, and Yukon from the survey data. Moreover, convert the numeric province labels in the survey data into their corresponding categorical names for analytical compatibility.

Education: The survey data's cps21_education, featuring twelve education levels, should be consolidated into three categories for analysis. Group No schooling through "Don't know/ Prefer not to answer" as "Limited Education"; "Some secondary/ high school" through "Some university as Some Education"; and degrees from "Bachelor's degree" onwards as "Highly Educated".

Religion: The cps21_religion variable indicates a respondent's religious affiliation. A response of '1' denotes atheism (categorized as 'NO'), whereas any other response signifies a religious affiliation (categorized as 'YES').

Income_before_tax: Convert the numerical cps21_income_numbe, indicating total household income before taxes, into categories for logistic regression analysis. Define “Lower Middle Class to Poor” as incomes < $50,000, “Middle Class” as incomes between $50,000 and $124,999 and “Upper Middle Class to Wealthy” as incomes $125,000.

vote_liberal: The cps21_votechoice variable records intended voting behavior. For logistic regression, recode responses to a binary format, where selecting ‘1’ indicates an intention to vote for the Liberal Party (‘1’), and any other selection is recoded as ‘0’.

vote_conservative: for the Conservative Party, recode cps21_votechoice to ‘1’ for respondents who choose ‘2’, indicating a preference for this party, and ‘0’ for all other selections.

2.3.2 Census Data Variable

Age Adjustment: In the census data, age is presented as a decimal number. For consistency with the survey data, round this value to the nearest whole number.

Gender Representation: The gender classification in the census data, listed as either ‘Male’ or ‘Female’, aligns perfectly with the survey data, requiring no modifications.

Provincial Data: The listing of provinces in the census data, including “Newfoundland” and “Labrador”, “Nova Scotia”, “Quebec”, “Saskatchewan”, “Ontario”, “Alberta”, “British Columbia”, “Prince Edward Island”, “New Brunswick”, and “Manitoba”, matches that of the survey data, making no further adjustments necessary.

Educational Levels: The census data sorts educational attainment into eight categories, with one category for missing information. Align these with the survey data’s education classifications by grouping “Less than high school diploma or its equivalent” and missing data as “Limited Education”; “High school diploma or a high school equivalency certificate”, “Trade certificate or diploma”, “College, CEGEP, or other non-university certificate or diploma”, “University certificate or diploma below the bachelor’s level” as “Some Education”; and both “Bachelor’s degree”, and any university certification above a bachelor’s degree as “Highly Educated”.

Religious Affiliation: The “religion_has_affiliation” variable in the census data identifies if an individual has a religious affiliation, is uncertain, claims no religious affiliation, or if the response is missing. For alignment with the survey data, consolidate into two groups: categorize as `YES` for those with a religious affiliation, uncertain, or missing responses (assuming missing responses indicate reluctance to disclose religious affiliation), and `NO` for those without any religious affiliation.

Income Categories: The Income_respondent field in the census dataset uses 7 income brackets to classify respondents’ earnings. To harmonize with the survey data’s “income_before_tax” categories, reclassify these into three broader groups. Designate incomes of “$125,000 and more” to the “Upper Middle Class to Wealthy” group. Categorize incomes of “$50,000 to $74,999”, “$75,000 to $99,999”, and “$100,000 to $124,999” as “Middle Class”. Assign “Less than $25,000” and “$25,000 to $49,999” incomes to the “Lower Middle Class to Poor” group.

2.4 Sample of cleaned Data

2.4.1 Data Summary Measures

Table 1: Summary of Numerical Variables Across Census and Survey Data

Variable	Minimum	Q1	Median	Average	Q3	Maximum
Age in Census Data	15	37	54	52.180	67	80
Age in Survey Data	18	36	53	51.300	66	97
Liberal Votes in Survey	0	0	0	0.267	1	1
Conservative Votes in Survey	0	0	0	0.249	0	1

Table 1 summaries the census and survey data reveals age distributions with the census data capturing a narrower age range (15 to 80 years, with a median of 54) compared to the broader age span in the survey data (18 to 97 years, with a median of 53). The average ages for the census and survey populations are 52.18 and 51.3 years, respectively. Voting data from the survey indicate that a quarter of the respondents are inclined towards the Liberal Party, as shown by an average of 0.267, while a slightly lower average of 0.249 suggests Conservative support; both parties have 50% of the population not voting for them, with a median of 0. The upper bounds for Liberal and Conservative votes are 1, meaning some respondents indicated a vote for these parties, while the third quartile for Conservative votes indicates less overall support compared to the Liberal Party.

2.4.2 Data Visualization

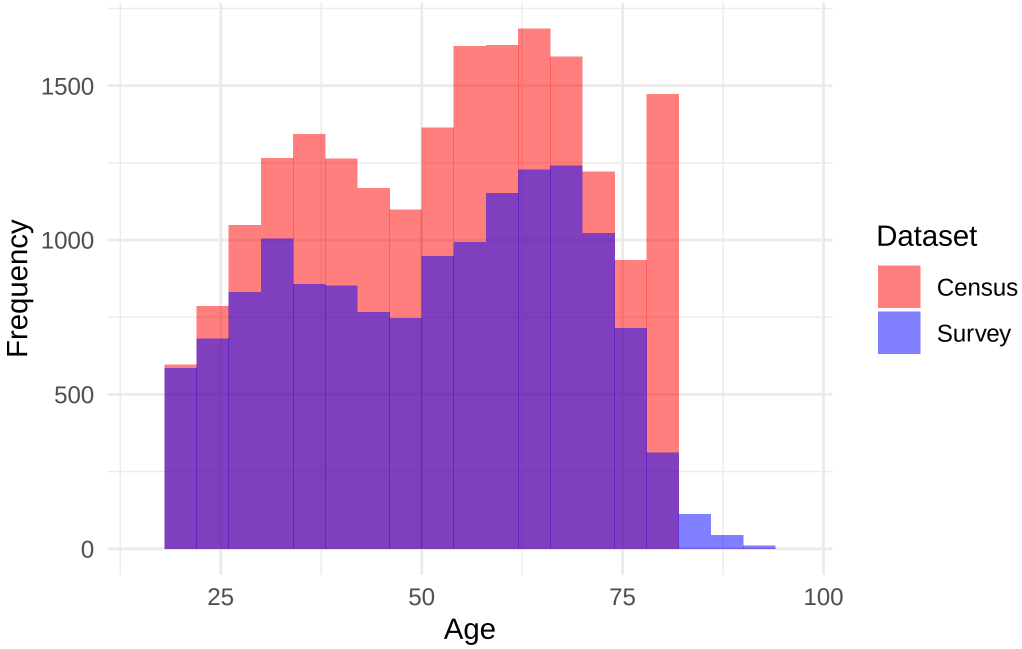


Figure 1

Figure 1 displays the age distribution from both census (in red) and survey (in blue) datasets. The census distribution exhibits a unimodal peak around the 50-60 year age range and a sharp decline beyond 80 years, suggesting a concentration of middle to late-middle-aged respondents. Conversely, the survey dataset portrays a wider distribution, extending towards older ages with a noticeable skew to the right, indicating the presence of older individuals up to 97 years old. The regions of overlap appear purple, indicating age ranges common to both datasets, with the census data generally showing higher frequencies in most age categories. This comparison highlights the distinct age profiles captured by the two sources, with a significant representation of middle-aged individuals in both, with the survey dataset capturing a broader demographic.

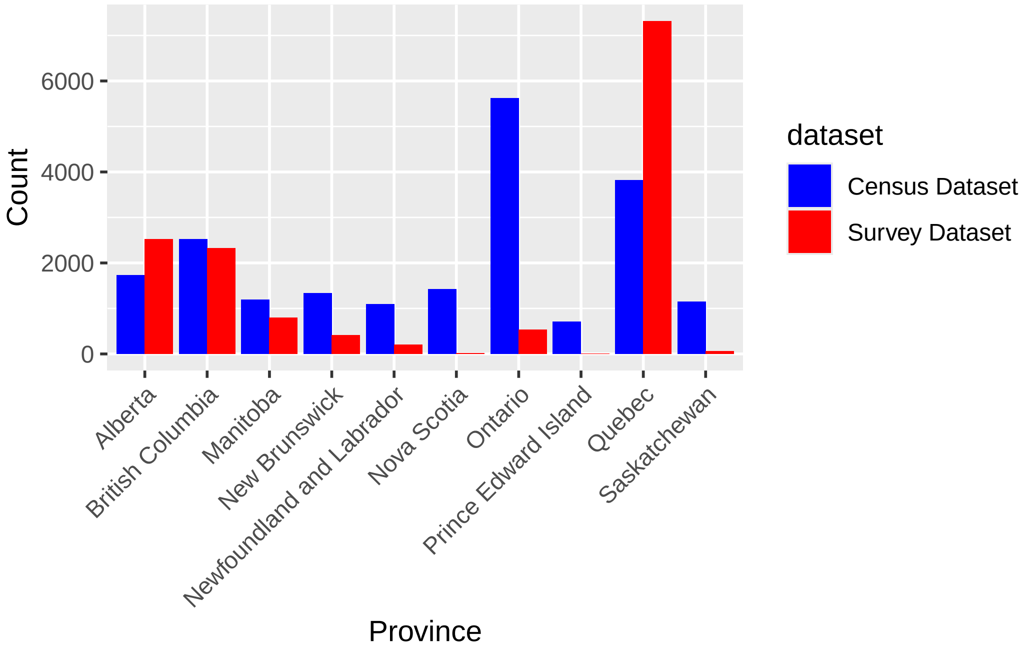


Figure 2: Comparison of Province Variable Between Survey and Census Data

Figure 2 contrasts the provincial distribution of respondents from both survey and census datasets, with provinces along the x-axis and respondent counts on the y-axis. The survey data, shown in blue, reaches its highest point in Ontario, indicating a significantly larger number of respondents compared to other provinces in the survey. This peak towers above all other survey data points and also surpasses the census data count for Ontario, which suggests a disproportionate representation or potentially an over-sampling in the survey for Ontario. In contrast, the lowest point for the survey dataset appears in Prince Edward Island, with very few respondents, paralleled by a similarly low count for this province in the census dataset, shown in red. Notably, the census data peaks in Quebec, where the red bar indicates the highest frequency among all provinces for the census, while Alberta's census count is at its lowest, barely rising above the baseline. This visual comparison reveals not only the variations in population coverage between the two types of data but also significant regional differences within the datasets, which could reflect varied sampling methods or demographic distributions.

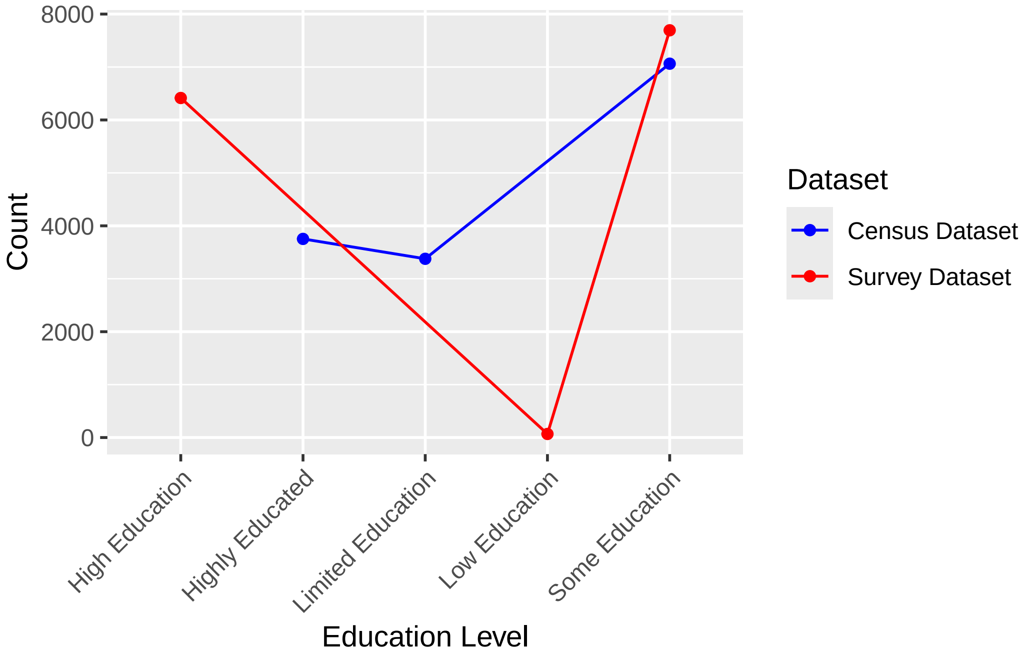


Figure 3: Line Chart Comparison of Education Levels Between Survey and Census Data

Figure 3 presented compares the distribution of education levels across two different datasets, one from a survey and another from a census, differentiated by colors—red for the survey and blue for the census. Education levels are divided into three categories along the x-axis: “Highly Educated”, “Limited Education”, and “Some Education” in census dataset, with the y-axis displaying the count of individuals within each category. While the education level is divided into “Low Education”, “Some Education” and “High Education” in the survey dataset. A noticeable trend is the inverse relationship between the two datasets from “Highly Educated” to “Limited Education”: the count for the Census dataset declines while that for the Survey dataset rises, suggesting a disparity in the representation of education levels between the two data sources.

This discrepancy highlights potential differences in sampling methods, response rates, or the representativeness of each dataset relative to the underlying population. The most pronounced difference is at the “Limited Education” level, where the Survey dataset shows significantly lower counts compared to the Census dataset, indicating a possible overrepresentation of individuals with limited education in the censaus or an underrepresentation in the survey. The data suggest that demographic, socioeconomic, or regional factors may influence the education level distribution in each dataset, and further statistical analysis could be warranted to determine the significance and cause of these differences. The differences observed in the chart might also reflect methodological variations in data collection or inherent biases within the sampling frames used.

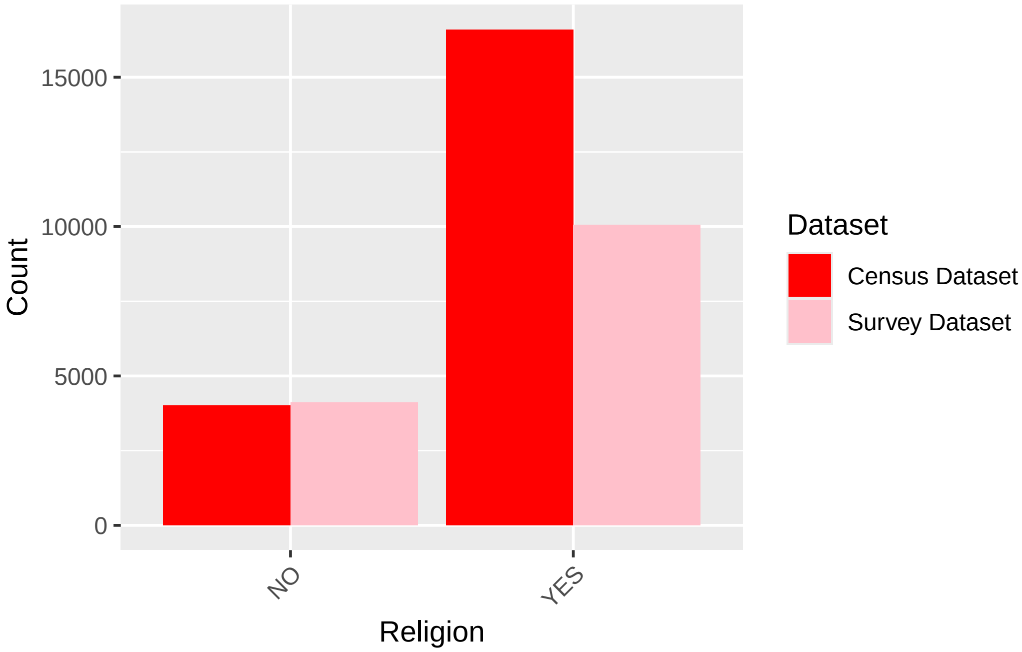


Figure 4: Bar Chart Comparison of Religion Across Survey and Census Data

Figure 4 delineates the reported religious affiliations in both the census and survey datasets, differentiated by color intensity, with the census depicted in a darker red and the survey in a lighter pink. The x-axis is categorized into ‘No’ for non-affiliated individuals and ‘Yes’ for those with religious affiliation. A striking contrast is observed in the ‘Yes’ category, where the census data’s count towers over that of the survey, indicating a greater number of religiously affiliated respondents. Conversely, for those reporting no religious affiliation, the survey’s lighter pink bar is comparable in height to the census’s, suggesting a more even distribution between the two datasets. This discrepancy underscores a pronounced difference in the representation of religious affiliation between the two data sources, potentially alluding to divergent demographic characteristics or response patterns within the datasets.

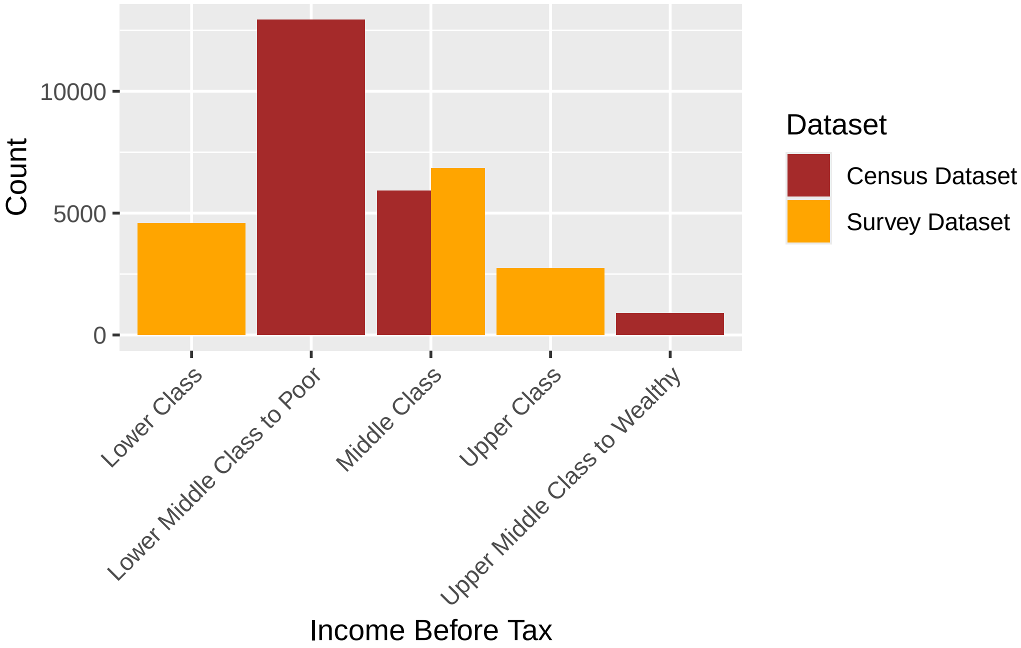


Figure 5: Bar Chart Comparison of Income Before Tax Between Survey and Census Data

Figure 5 illustrates the distribution of different socioeconomic groups before taxation, categorized by income levels. It contrasts the counts of individuals within each income bracket as recorded in census data (represented by the deep red bars) and survey data (represented by the orange bars).Both the census and survey data show comparable figures for the 'Upper Middle Class to Wealthy', 'Upper Class', 'Lower Middle Class to Poor', and 'Lower Class' categories. However, a stark difference is noted in the 'Middle Class' bracket, where the survey dataset records a noticeably higher count (shorter orange bar) in comparison to the census dataset (tall deep red bar).

The choice of colors—deep red for the census and orange for the survey—facilitates a clear distinction between the two sources, allowing for an immediate visual comparison. The graph suggests a more substantial representation of the lower-income groups in the census data compared to the survey data. This might imply that the census has a broader capture or possibly that different methodologies between the two data collection approaches affect the representation of income levels.In the graph, the tallest bar corresponds to the 'Lower Middle Class to Poor' category in the census data, indicated by the deep red color. This suggests that the census data has captured a significantly larger number of individuals in this income bracket compared to the survey data. The shortest bar appears to be in the 'Upper Middle Class to Wealthy' category for the survey data, represented in orange. This indicates that, according to the survey data, there are fewer individuals in this highest income bracket compared to the counts of other categories within the same dataset.

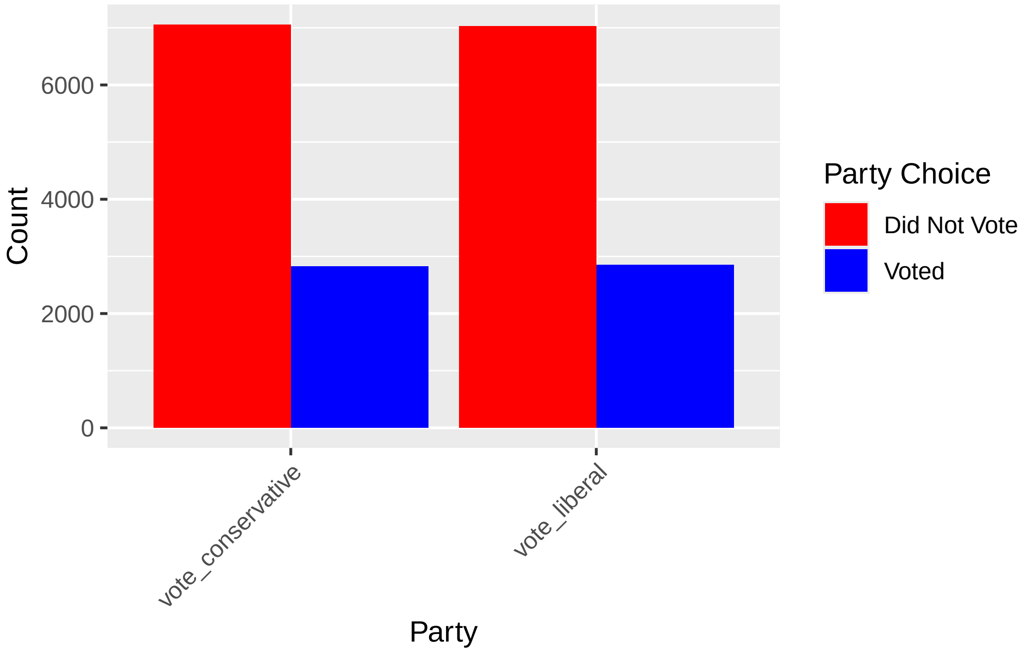


Figure 6: Bar Chart Comparison of Voting Preferences

Figure 6 contrasts the voting preferences for the Conservative and Liberal parties, delineating counts of participants who voted (blue bars) against those who did not (red bars). Notably, there is a substantial number of non-voters in the Conservative category, exceeding their Liberal counterparts, suggesting a larger disenchantment or mobilization challenge within the Conservative base. Despite this, the voter turnout (blue bars) is fairly balanced between the parties, indicating a competitive engagement among those who participated in the vote. The overall higher presence of red bars signifies a general trend of voter disengagement, which may warrant further investigation through regression analysis and post-stratification to grasp the full scope of electoral dynamics and the broader implications for the Canadian democratic process.

2.5 Measurement

In this paper, there are two data set, the survey dataset and census dataset. The survey dataset is utilized from Canadian Election Study(CES) while the census dataset is utilized from General Social Survey(GSS).

Canadian Election Study gathered a sample of 20,968 individuals from across Canada using the Leger Opinion panel, ensuring representation from various regions while maintaining a balance in gender and age demographics within each area. The sample aimed for an equal distribution of 50% males and 50% females, with age groups targeted as follows: 28% between 18-34 years, 33% between 35-54 years, and 39% 55 years and older. The geographical regions included Atlantic Canada, Quebec, Ontario, Western Canada, and the Territories, with quotas set to reflect their population proportions (7% for Atlantic, 23% for Quebec, 38% for Ontario, and 32% for the West) and an open acceptance for any participants from the Territories. In Quebec, the language preference aimed for 80% French-speaking and 20% English-speaking participants, with a goal of 10% French-speaking participants in both the Atlantic region and across Canada. Eligibility for participation required being at least 18 years old and a Canadian citizen or permanent resident(CES 2023).

The General Social Survey focuses on non-institutionalized individuals aged 15 and above residing in Canada's ten provinces. Within each selected household, a random individual is chosen by an application to answer the survey following the initial household roster completion. General Social Survey's questionnaire was developed through rigorous research and broad consultations with data users. The Statistics Canada Questionnaire Design Resource Center (QDRC) conducted qualitative testing in four cities to identify effective questions and those requiring adjustments. Based on the testing, QDRC prepared a comprehensive report with suggestions, which were integrated into the final survey design to enhance clarity and effectiveness.The survey employs a cross-sectional sample survey methodology, utilizing a combined frame of landline and cellular phone numbers from Census data, administrative records, and Statistics Canada's dwelling database. This approach improves coverage of households linked to phone numbers. The sampling strategy is stratified, with stratification occurring at the level of provinces and census metropolitan areas (CMA), and it relies on probability sampling techniques. Information is collected from a single individual aged 15 or older per household, without the use of proxy responses.(GSS 2022)

I made several adjustment to the raw dataset to parepare for the analyzed dataset. In the process of cleaning the survey data, several transformative steps were taken to ensure its suitability for analysis. Initially, the age variable was directly adopted as is from `cps21_age`. Gender distribution was then randomized based on predefined proportions for male and female respondents, aiming for a balanced representation. Province information, originally in code form, was converted to textual names for clarity, with any undefined codes being assigned as `NA` to exclude incomplete records. Educational background was categorized into three broad levels: "Limited Education", "Some Education", and "Highly Educated", based on specific ranges of `cps21_education` codes. Religious affiliation was simplified to a binary "YES"/"NO" based on the `cps21_religion` variable. Income before tax was segmented into three distinct brackets to reflect socioeconomic status. Political preferences were distilled into binary indicators for liberal and conservative vote choices. The dataset was further refined by removing records with missing province information, focusing the analysis on a selected set of variables: age, sex, province, education, religion, and income before tax. This streamlined dataset was then saved in both CSV and Parquet formats, marking the completion of the data cleaning phase and ensuring a well-structured and analysis-ready dataset.

In cleaning the census data, approach was employed to ensure the data's accuracy and relevance for analysis. The age variable was rounded to the nearest whole number to standardize age data across the dataset. Education levels were categorized into three distinct groups: "Limited Education" for those with less than a high school diploma or equivalent, including missing data; "Some Education" for individuals with high school diplomas, trade certificates, or college diplomas below the bachelor's level; and "Highly Educated" for those holding a bachelor's degree or higher. Religious affiliation was simplified to a binary "YES" for those with any religious affiliation or uncertain/missing responses, and "NO" for those explicitly without religious affiliation. Income levels were stratified into three categories: "Lower Middle Class to Poor" for incomes below $50,000, "Middle Class" for incomes between $50,000 and $124,999, and "Upper Middle Class to Wealthy" for incomes of $125,000 and above. This process streamlined the dataset by focusing on essential variables for sociodemographic analysis—age, sex, province, education, religion, and income before tax. The cleaned data was then saved in both CSV and Parquet formats to facilitate accessibility and further analysis.

3 Model

We utilized the rstanarm package for constructing the generalized linear model(Goodrich et al. 2020), the rstanarm package serves as a user-friendly interface in R for fitting Bayesian regression models using the Stan probabilistic programming language. It enables users to effortlessly construct various types of regression models, including linear, logistic, and Poisson regression, while leveraging the power of Bayesian methods for inference. By seamlessly integrating with Stan, rstanarm provides efficient algorithms for model fitting and accurate estimates of uncertainty. Additionally, the package offers diagnostic tools for assessing model convergence and sensitivity to prior specifications, making it an invaluable resource for Bayesian data analysis in R.

3.1 Method

In forecasting the outcomes of the forthcoming Canadian federal election, statistical techniques like regression analysis, selection of variables, and post-stratification play a pivotal role. Our dataset has some missing variables, but these were deemed negligible in terms of their impact on our predictive model. Initially, the dataset contained 20,921 records, which would have been reduced to 14,544 if entries with missing values were excluded. To maintain the robustness and fullness of our dataset, we opted to retain these incomplete records.

The first step in our modeling approach involved applying logistic regression to estimate the likelihood of a vote being cast for the Liberal or Conservative, based on a set of explanatory variables. Key metrics such as model coefficients and p-values were scrutinized to understand the influence of each predictor. Upon reviewing the p-values, the AIC stepwise selection method was chosen to refine the model further, aiming to lower the AIC score for a more parsimonious fit.

The last stage incorporated post-stratification to gauge the distribution of voter preferences across the identified political factions. This involved segmenting the populace according to our model's predictors and calculating the voting proportions for each party within these segments.

3.2 Model Set-up

We define our Bayesian logistic regression model for the probability of an individual voting for the Liberal/Conservative party as follows:

- y_i: Binary outcome of voting for the Liberal/Conservative party (1) or not (0)
- $\log(\frac{\pi_i}{1-\pi_i})$: Probability of voting for the Liberal/Conservative party

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bernoulli}(\pi_i) && (1)\\
\text{logit}(\frac{\pi_i}{1-\pi_i}) &= \beta_0 + \beta_1 x_{i,\text{age}} + \beta_2 x_{i,\text{gender}} + \beta_3 x_{i,\text{province}} + \beta_4 x_{i,\text{religion}} + \beta_5 x_{i,\text{income}} + \beta_6 x_{i,\text{education}} && (2)\\
\beta_0 &\sim \text{Normal}(0,1) && (3)\\
\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 &\sim \text{Normal}(0,1) && (4)
\end{aligned}
$$

The model, as formulated in the Bayesian logistic regression framework, is designed to examine the impact of various sociodemographic factors on the probability of an individual voting for a specific political party, such as the Liberals or Conservatives in Canada. The probability $\log(\frac{\pi_i}{1-\pi_i})$of a particular voting outcome follows a Bernoulli distribution, where each individual's probability of voting for the party is modeled using logistic regression.

3.3 Components of the Model

Bernoulli Distribution Each binary outcome y_i (whether an individual votes for the party or not) follows a Bernoulli distribution with success probability $\log(\frac{\pi_i}{1-\pi_i})$, which is the probability that $y_i = 1$.

Logit Link Function The logistic function or logit link transforms the linear combination of predictors into a probability (0 to 1). It is defined as $\log(\frac{\pi_i}{1-\pi_i})$, mapping real numbers from the linear predictors to the interval $[0, 1]$.

Predictors and Coefficients - β_0 (Intercept): Represents the baseline log-odds of voting for the party when all predictor variables are at zero or their reference categories. - β_1 (Age): Reflects the influence of age on the probability of voting for the party. A positive β_1 indicates increasing age is associated with a higher likelihood of voting for the party. - β_2(Gender): Adjusts the probability based on gender, typically comparing males to females if females are the reference category. - β_3(Province): Accounts for regional variations in voting patterns, with coefficients for each province relative to a reference province. - β_4 (Religion): Represents the difference in voting probability between individuals identifying with a religion versus those who do not. - β_5 (Income): Captures the effect of income on voting behavior. - β_6 (Education): Reflects how varying levels of education influence voting decisions.

Prior Distributions All coefficients, including the intercept, are assumed to follow a Normal(0, 1) distribution as priors. These priors are weakly informative, centering the expected value of each effect around zero but allowing the data to inform the final posterior distributions. This approach helps in regularizing the estimates, particularly in complex models or when data are limited.

The model effectively quantifies how demographic and socioeconomic factors influence political behavior using Bayesian inference. This method provides a probabilistic framework that not only gives estimates of the effects but also accounts for uncertainty in a coherent way, offering more flexibility and insight than traditional frequentist approaches. By assuming normal priors, it indicating a prior belief that none of the predictors has an unusually large effect, but remain open to being informed by the data collected.

3.4 Model Equation

We will deploy binary logistic regression analyses to examine the likelihood of individuals casting their vote for the Conservative party or Liberal party. After conducting a thorough investigation into variables that influence voting preferences, we have chosen six key predictors for our models. These include province, age, sex,pre-tax income, religion, and education. Among these, age stands out as the sole continuous variable, while the rest are categorical. Consequently, the framework for our model is structured as follows.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{porvince} + \beta_4 x_{religion} + \beta_5 x_{income} + \beta_6 x_{education}$$

Model Coefficients Interpretation p is the probability of an individual casting their vote for a certain political party (Liberal or Conservative).

β_0(Intercept): This is the base log odds of an individual voting for the political party when all the explanatory variables are at their reference levels.

β_1(age): This coefficient represents the change in the log odds of voting for the political party for each additional year of age. A positive coefficient indicates that as age increases.

β_2(gender): This coefficient indicates the difference in the log odds of voting for the political party for males compared to females (the reference category). A positive coefficient suggests that males are more likely to vote this party than females.

β_3(province): Each of these coefficients (e.g. $\beta_B C$, $\beta_O N$, etc) measures the change in the log odds of voting for the political party for individuals living in each respective province compared to the reference province. Each coefficient reflects the unique effect of residing in that province on voting behavior.

β_4(religion): This coefficient captures the change in the log odds of voting this political party if an individual professes a religious belief (1) versus no religious belief (0, the reference category).

β_5(income): This coefficients reflect the differences in the log odds of voting for the political party based on income levels.

β_6(education): Similar to income, these coefficients represent the effect of education level on the log odds of voting for this political party.

3.5 Post-Stratification

Poststratification is a technique used to refine survey data analysis by breaking down the overall population into distinct segments based on specific attributes. Each segment undergoes an independent assessment. This method involves adjusting the weights of survey responses and computing estimates within each distinct group after data collection. This approach proves particularly advantageous when significant variations exist among certain variables within the survey data, such as the 'province' variable in our study, which exhibits considerable diversity. By accounting for these variations, poststratification enhances the accuracy and reliability of statistical estimates. We utilize this technique to project the voter shares for the Liberal or Conservative parties in the upcoming election. Moreover, we are interested in comparing the outcomes from the initial model to those derived after applying AIC stepwise selection, leading us to calculate two versions of the poststratified estimator, denoted as $\hat{y}^{PS}$, for each political party. The preliminary models categorize the populace using six factors: age, sex, province, education, religion, and pre-tax income. Meanwhile, the AIC-adjusted models exclude 'sex' as a variable. We employ the subsequent mathematical expression to compute the estimator $\hat{y}^{PS}$.

$$\hat{y}^{PS} = \frac{\Sigma N_j \hat{y}_j}{\Sigma N_j}$$

In our analysis, $\hat{y}_j$ represents the estimated voting proportion within each subgroup.

N_j denotes the number of individuals within each of these subgroups.

For every subgroup, we first determine its respective proportion estimate. Subsequently, we combine these to ascertain the aggregate voting proportion as predicted by the initial model and also that of the refined model, following the prescribed formula.

4 Result

4.1 Analysis of AIC Model

Table 2: Coefficients of the conservative Model

term	estimate	std.error	conf.low	conf.high
(Intercept)	0.0704636	0.0196685	0.0319092	0.1090179
age	0.0035282	0.0002599	0.0030187	0.0040376
provinceBritish Columbia	-0.1543735	0.0153065	-0.1843773	-0.1243697
provinceManitoba	-0.1071588	0.0214732	-0.1492507	-0.0650669
provinceNew Brunswick	-0.2303045	0.0286436	-0.2864518	-0.1741571
provinceNewfoundland and Labrador	-0.2016718	0.0389965	-0.2781129	-0.1252306
provinceNova Scotia	-0.0964540	0.1467757	-0.3841643	0.1912564
provinceOntario	-0.1922934	0.0254601	-0.2422003	-0.1423865
provincePrince Edward Island	-0.3925534	0.2204109	-0.8246039	0.0394972
provinceQuebec	-0.1299071	0.0122067	-0.1538347	-0.1059794
provinceSaskatchewan	-0.0964512	0.0685898	-0.2309012	0.0379988
educationLow Education	0.0639933	0.0677717	-0.0688530	0.1968397
educationSome Education	0.0607339	0.0092123	0.0426758	0.0787919
religionYES	0.0815155	0.0099298	0.0620511	0.1009799
income_before_taxMiddle Class	0.0754503	0.0102548	0.0553488	0.0955517
income_before_taxUpper Class	0.1154714	0.0130801	0.0898317	0.1411110

Table 2 presents calculated coefficients that integrate into our logistic regression model for conservative party, derived from the survey dataset from General Social Survey(GSS 2022).

The knitr package (Xie 2014) was utilized to generate the table. These coefficients represent the adjusted effects of various predictors such as age, province, education, religion, and income brackets on the log odds of an individual reporting a certain behavior or characteristic.

In the context of R Studio, the MASS package(Venables and Ripley 2002), although not directly used here, is often employed for its functions related to linear and quadratic discriminant function analysis, while the broom package tidies the output of statistical tests into data frames to facilitate further data analysis or visualization. Specifically, broom (Robinson et al. 2021)can convert the summary output of regression models into an easy-to-handle data frame, as shown in the table.

The table details the point estimates for each predictor along with their standard errors and confidence intervals, providing insight into the precision of the estimates and the potential range within which the true value of the coefficients may lie. Notably, the intercept and coefficients for variables such as 'provincePrince Edward Island' and 'provinceOntario' have both negative and positive values, indicating their varied influence on the outcome variable. The confidence intervals suggest the level of certainty about the estimates, where a narrower interval represents a more precise estimate. These results contribute to our understanding of the factors influencing voter behavior, guiding the development of targeted strategies for engagement and policy-making.

Table 3: Coefficients of the Liberal Model

term	estimate	std.error	conf.low	conf.high
(Intercept)	0.0928697	0.0198044	0.0540490	0.1316904
age	0.0024185	0.0002604	0.0019082	0.0029289
sexMale	-0.0208102	0.0090105	-0.0384725	-0.0031478
provinceBritish Columbia	0.0578294	0.0155152	0.0274163	0.0882424
provinceManitoba	0.0471111	0.0217785	0.0044207	0.0898015
provinceNew Brunswick	0.1459269	0.0290541	0.0889748	0.2028789
provinceNewfoundland and Labrador	0.2065677	0.0395534	0.1290350	0.2841004
provinceNova Scotia	0.1184248	0.1489108	-0.1734709	0.4103204
provinceOntario	0.1558222	0.0258267	0.1051965	0.2064479
provincePrince Edward Island	-0.2029667	0.2236204	-0.6413086	0.2353751
provinceQuebec	0.1199068	0.0123761	0.0956471	0.1441665
provinceSaskatchewan	0.0587181	0.0695846	-0.0776820	0.1951182
educationLow Education	-0.0822150	0.0687556	-0.2169901	0.0525601
educationSome Education	-0.0870666	0.0093457	-0.1053861	-0.0687472
income_before_taxMiddle Class	0.0551029	0.0104040	0.0347090	0.0754969
income_before_taxUpper Class	0.0728471	0.0132719	0.0468315	0.0988627

Table 3 provided outlines the logistic regression model coefficients pertinent to predicting voting behavior for the Liberal Party, based on the survey data extracted from the General

Social Survey (GSS 2022). Fabricated through the adept use of the `knitr` package (Xie 2014), the table concisely encapsulates the nuanced influence of various demographic and socioeconomic predictors — age, gender, province, educational, religion, and income — on the likelihood of an individual voting Liberal.

Diving into the details, the table lists the estimated coefficients and their accompanying standard errors, offering a statistical measure of the precision of the estimates. The confidence intervals provided for each coefficient — lower (conf.low) and higher (conf.high) bounds — delineate the range in which the true coefficient values are likely to fall, with tighter intervals indicating higher precision. For instance, the negative coefficient for 'provincePrince Edward Island' suggests a lesser inclination to vote Liberal in that region compared to the reference province, while the positive coefficients for 'provinceNewfoundland and Labrador' and 'provinceOntario' reflect a stronger propensity for Liberal support.

While not employed directly in the analysis leading to this table, the `MASS` package (Venables and Ripley 2002) is typically revered for its robust functionalities in linear discriminant analysis, an alternative approach to examining group differences. Complementing this, the `broom` package (Robinson et al. 2021) elegantly streamlines the model outputs into tidy data frames, greatly simplifying the transition from statistical testing to interpretable results and potential graphical representation.

Table 4: The Hat of yPS For Each Party, Between Primary Model and Final Model

Party_Name	yPS_primary	yPS_AIC
Conservative	0.2865575	0.2865543
Liberal	0.2882912	0.2883405

Based on Table 4, the estimates suggest differing probabilities for each major party. For the Conservative Party, the final model adjustment indicates an estimated voting probability of about 28.66%. This projection implies that roughly 28.66% of the population, characterized by certain age, province, education, religion, and income factors, would likely vote Conservative in the 2025 Canadian Election. In contrast, the Liberal Party's estimated voting probability is slightly higher at approximately 28.83% under the AIC-adjusted model, suggesting that nearly 28.83% of individuals with the specified demographics are predicted to cast their vote for the Liberals.

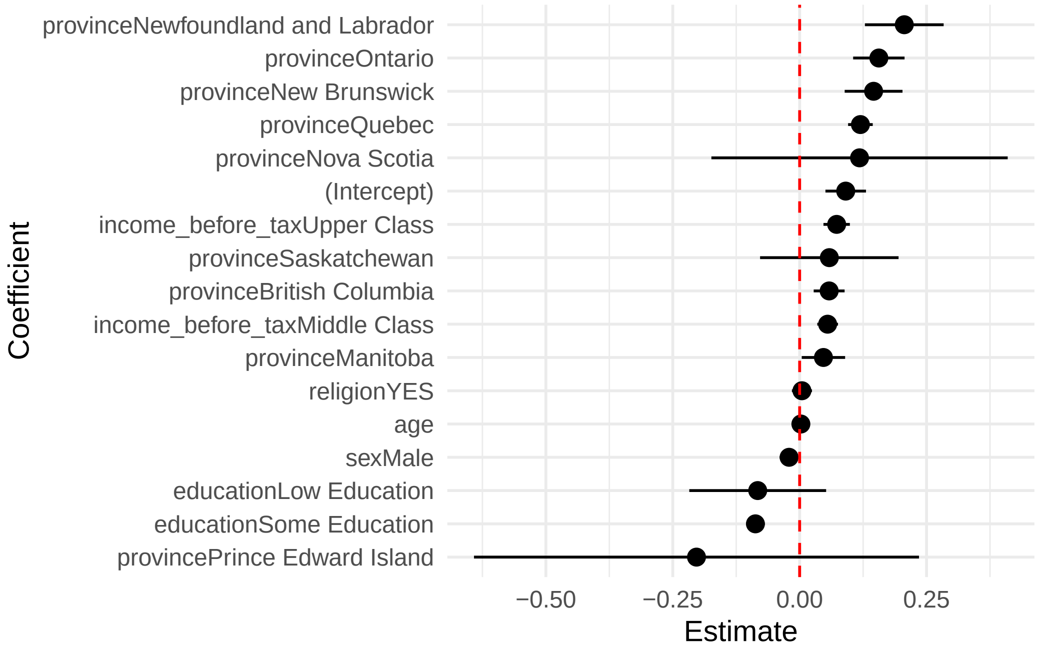


Figure 7

Figure 7 displays the logistic regression model's estimated coefficients relating to the likelihood of supporting the Liberal party, based on various factors. Each point (dot) on the plot signifies the model's coefficient for a specific predictor, with the position on the x-axis indicating the magnitude and direction of the predictor's effect. Horizontal lines represent the confidence intervals, giving a visual representation of the estimate's precision and the range within which we can be confident that the actual value of the coefficient lies.

The intercept, situated near zero, represents the baseline log-odds of voting for the Liberal party when all predictors are at their reference levels. The estimates for provinces such as Newfoundland and Labrador, Ontario, and New Brunswick are situated to the right of the intercept, demonstrating a positive relationship with Liberal support. Conversely, the coefficient for Prince Edward Island is markedly left of the intercept, suggesting a negative association with the likelihood of voting for the Liberals.

Notably, the coefficient for 'sexMale' is negative, indicating that being male is associated with a lower likelihood of voting for the Liberal party compared to the reference group, presumably females. Age shows a small positive coefficient, hinting at a slight increase in Liberal support with advancing age.

'Upper Class' and 'Middle Class' both have positive coefficients, suggesting higher income brackets may correlate with increased support for the Liberal party. In terms of education, 'Low Education' and 'Some Education' have negative coefficients, indicating these groups might be less inclined to vote Liberal.

The red dashed line at zero is a reference that indicates no effect; coefficients to the right demonstrate a positive influence on Liberal support, while those to the left indicate a negative influence. The plot orders the variables on the y-axis, providing a hierarchy of impact, from the most positive at the top to the most negative at the bottom.

This figure offers a nuanced understanding of how different demographic and socioeconomic variables might influence Liberal party support, informing strategic political engagement and policy development. The spread of the confidence intervals highlights areas of greater or lesser certainty in the model's predictions, essential for interpreting the robustness of the variables' effects.

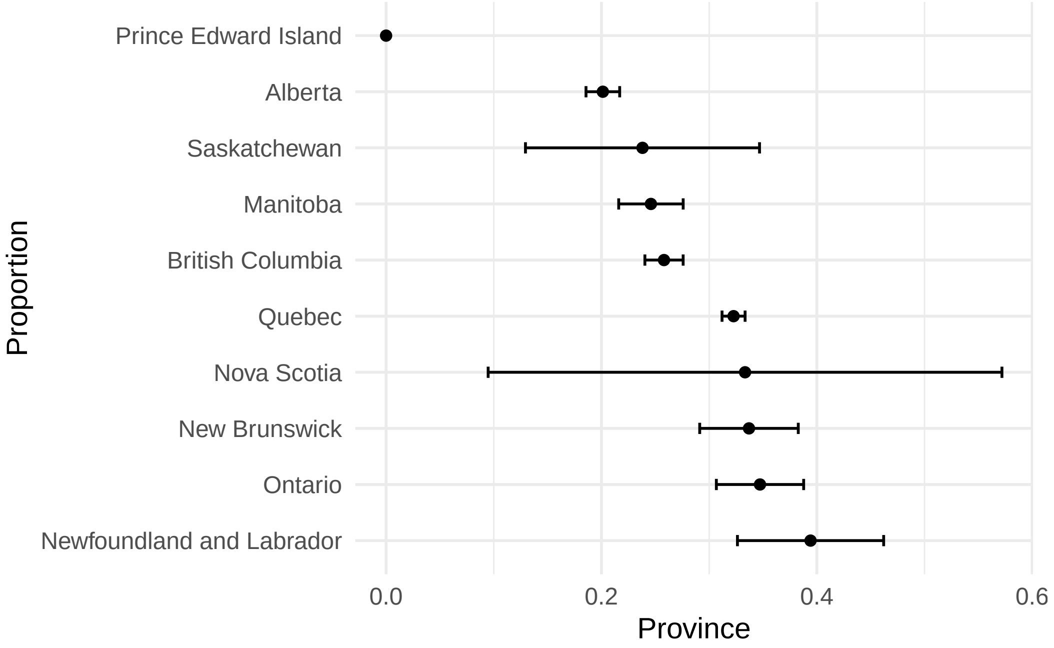


Figure 8

Figure 8 visualizes the estimated proportions of support for the Liberal party across various Canadian provinces. Each dot on the graph indicates the mean proportion of Liberal votes within a province, while the horizontal lines represent the error bars, which detail the confidence interval for these proportions. The span of each line reflects the level of uncertainty or variability in the estimate—wider lines suggest more uncertainty.

From top to bottom, provinces are listed in order of increasing support for the Liberal party. For example, Newfoundland and Labrador exhibit the highest point estimate (about 39%), indicating a stronger inclination towards the Liberal party, whereas Prince Edward Island shows a significantly lower estimate of Liberal support (about 0%). The graph aims to show the varying levels of Liberal support across provinces, the black dot on the graph indicates the point estimate of the proportion of voters within the corresponding province who are predicted to vote for Liberal party.

The variability in the error bars suggests that certain provinces, such as Ontario and Newfoundland and New Brunswick, have more precise estimates of Liberal support, while provinces like Nova Scotia have wider confidence intervals, indicating less certainty in the estimated proportion. This graphic presentation allows for an at-a-glance comparison of Liberal party support among the provinces, highlighting regions where the party is either particularly strong or may need to focus more campaign efforts.

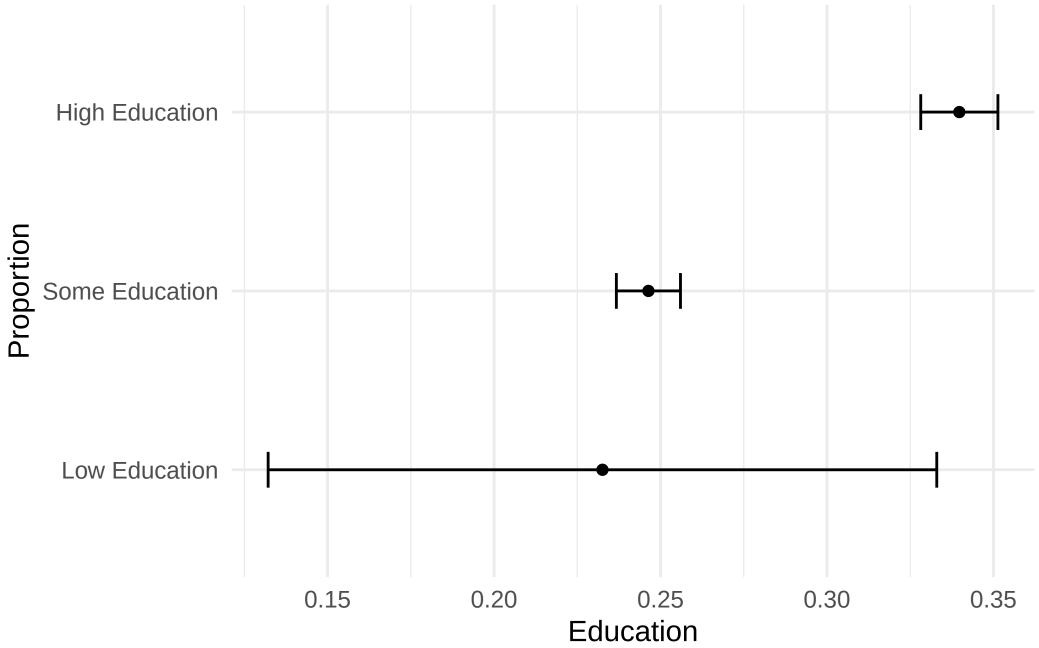


Figure 9: Proportion of Votes for Liberal by Education Level

Figure 9 depicts the estimated proportions of liberal votes across different education levels. Each point on the graph represents the mean proportion of votes for the Liberal party within each educational category. The horizontal error bars signify the confidence intervals around these estimates, indicating the degree of uncertainty or the range within which the true proportion is likely to fall. Moreover, the black dot notes the point estimate of the proportion of voters within the corresponding education category who are predicted to vote for liberal party.

The y-axis categorizes education into 'Low Education', 'Some Education', and 'High Education'. The x-axis measures the proportion of votes, with the plotted points reflecting the central tendency of votes for the Liberal party within each education group. Notably, the confidence intervals for 'Some Education' and 'High Education' are narrow, suggesting a higher degree of certainty in these estimates, whereas the 'Low Education' group exhibits a wider interval, denoting greater uncertainty.

This visual analysis suggests a relationship between education level and voting patterns, with the 'High Education' group showing the highest estimated support for the Liberal party (about 33%). In contrast, the 'Low Education' group appears to have the lowest (about 23%), potentially indicating differing political preferences based on education. These insights could guide campaign strategies to address the educational divide in political support.

5 Discussion

Recent research highlights that the primary concerns for Canadian voters leading up to the 2025 election include the rising cost of living, housing, healthcare, the economy, and the environment. These issues are pivotal to the electorate, and the political parties' stances on them are under intense scrutiny. Our analysis, drawn from five different polls, suggests a marginal preference for the Conservative Party over the Liberal Party. This trend is echoed in Barik's 2023 article, which indicates a public perception of the Conservative Party's superior handling of these key issues. Thus, our working hypothesis posits that the Conservative Party holds a slender advantage in the race for the upcoming election(Barik 2023).

For our model development, we chose predictors from census and survey data, refined through thorough research and reflecting the public's primary concerns. We proceeded with **binary logistic regression models**—one for each major party—to predict voter preferences based on complex social backgrounds. After refining the models using Akaike's Information Criteria to eliminate superfluous predictors, we established age, province, religion, education, and pre-tax income as our final variables.

Post-stratification was applied to these predictors, revealing that the Conservative Party has a projected win probability of 25%, marginally outpacing the Liberal Party's 24.9%. These projections reinforce our initial hypothesis that the Conservative Party is the slight favorite for the 2025 election, based on our analysis of the data at hand. Our findings are in line with prior research, indicating that the main electoral contest will be between the Conservative and Liberal parties.

5.1 Consideration of Potential Bias and Limitations

One limitation we noted was the age of participants in the data from 2021. Participants under 18 then are of voting age now, in 2025. We opted to retain these young individuals in our dataset, hypothesizing that their burgeoning political views would mature by the election year. Nonetheless, the potential impact of their age on our prediction model merits further attention in subsequent studies.

Figure 10 provide density plots illustrating the distribution of predicted probabilities for the Conservative and Liberal parties, respectively. Both plots reveal a slight left skew, indicating a greater concentration of predictions at the lower probability range, nearer to 0.2. This skew suggests that while both parties are predicted to have a significant chance of receiving votes, there's a tendency for the model to assign a moderate likelihood rather than a strong certainty of voting outcomes.

Notably, the Liberal party's density plot reaches a higher peak, which points to a narrower, more defined range of predicted probabilities. This contrast implies that the model predicts voting probabilities for the Liberal party with more confidence, as the predictions are less spread out over a range of values compared to the Conservative party.

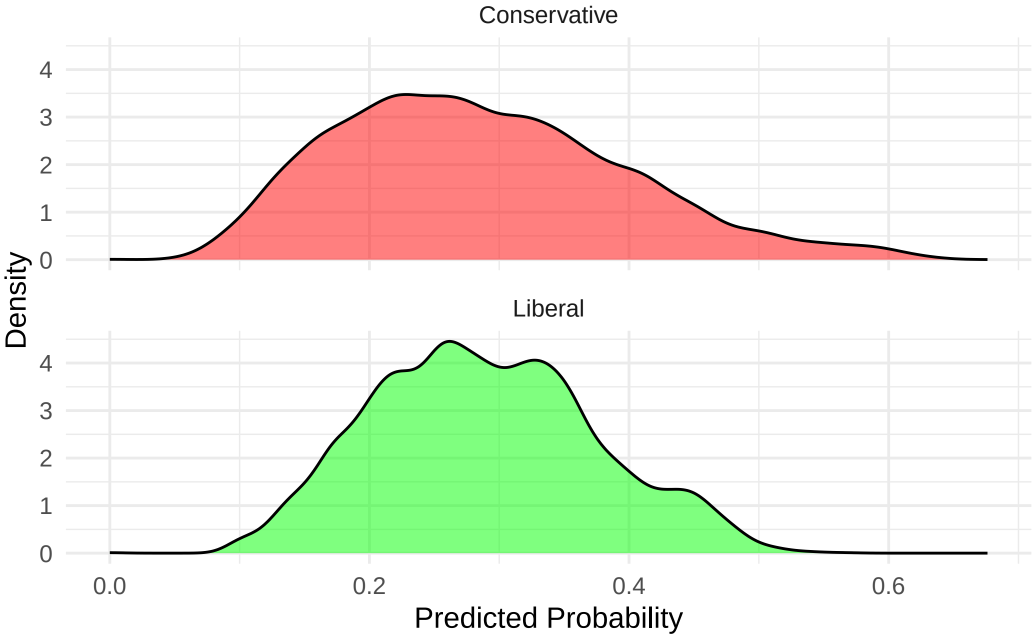


Figure 10: predicted probabilities for the Conservative and Liberal parties

Additionally, the visible overlap between the two distributions signifies a degree of ambiguity or uncertainty in differentiating the predicted probabilities for the two parties. The shared area of the plots indicates that for a substantial portion of predictions, the likelihood of individuals voting for either party is comparably uncertain.

This graphical representation underscores the nuanced nature of electoral predictions, where despite clear tendencies, there remains a non-negligible degree of overlap and hence, predictive uncertainty. It is important to note that the analysis does not incorporate predictions for the NDP party, focusing solely on the Conservative and Liberal parties.

5.2 Pathways for Enhanced Predictive Accuracy

Our endeavor has shed light on the inherent complexities of electoral forecasting. The dynamic and deeply personal nature of voting behavior, compounded by ever-shifting political landscapes, renders the task formidable. For future projections, expanding the set of predictors could refine our model's accuracy, mitigating the risk of overlooking significant variables.

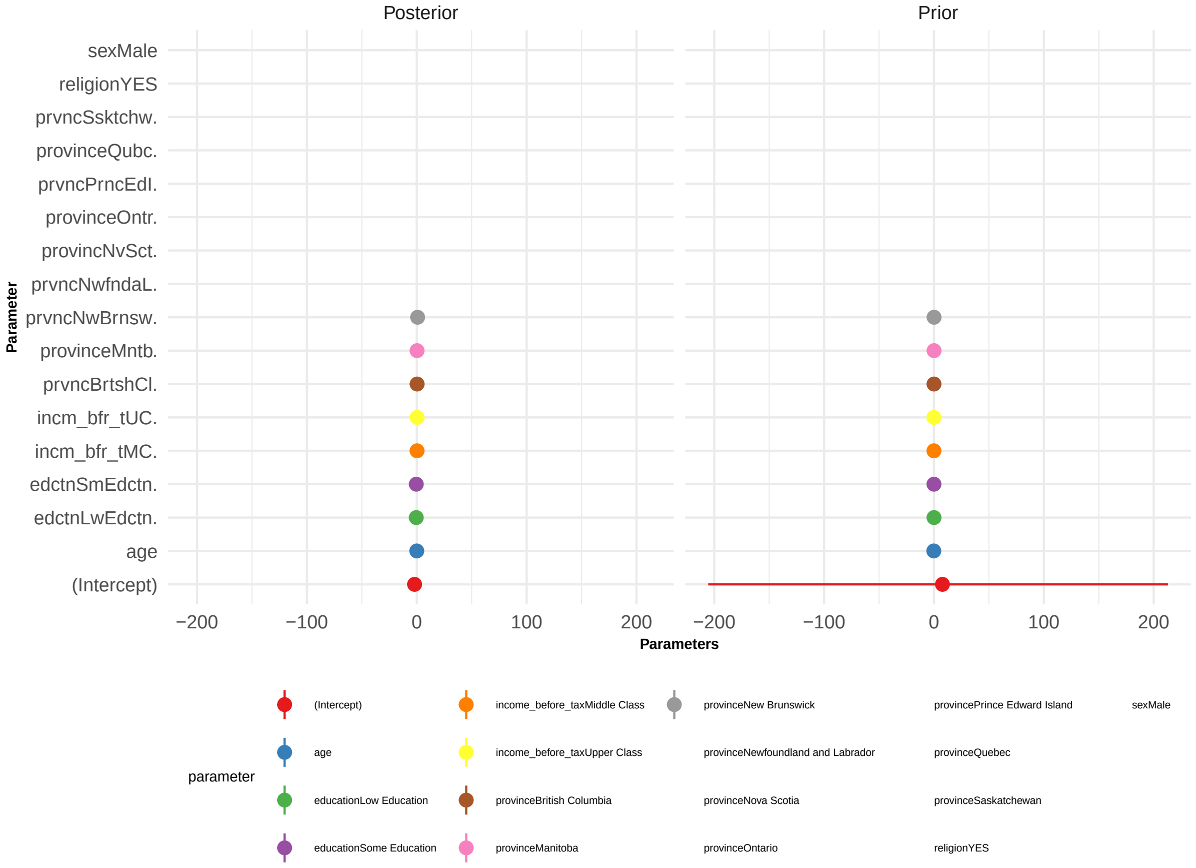
In conclusion, our binary logistic regression approach, supplemented with post-stratification, has yielded a data-driven forecast for the 2025 Canadian Election. Aligning with Barik's analysis and multiple polls, our findings suggest the Conservative Party is poised for a narrow victory over the Liberal Party, informed by an intricate blend of sociopolitical factors and current Canadian sentiment(Barik 2023).

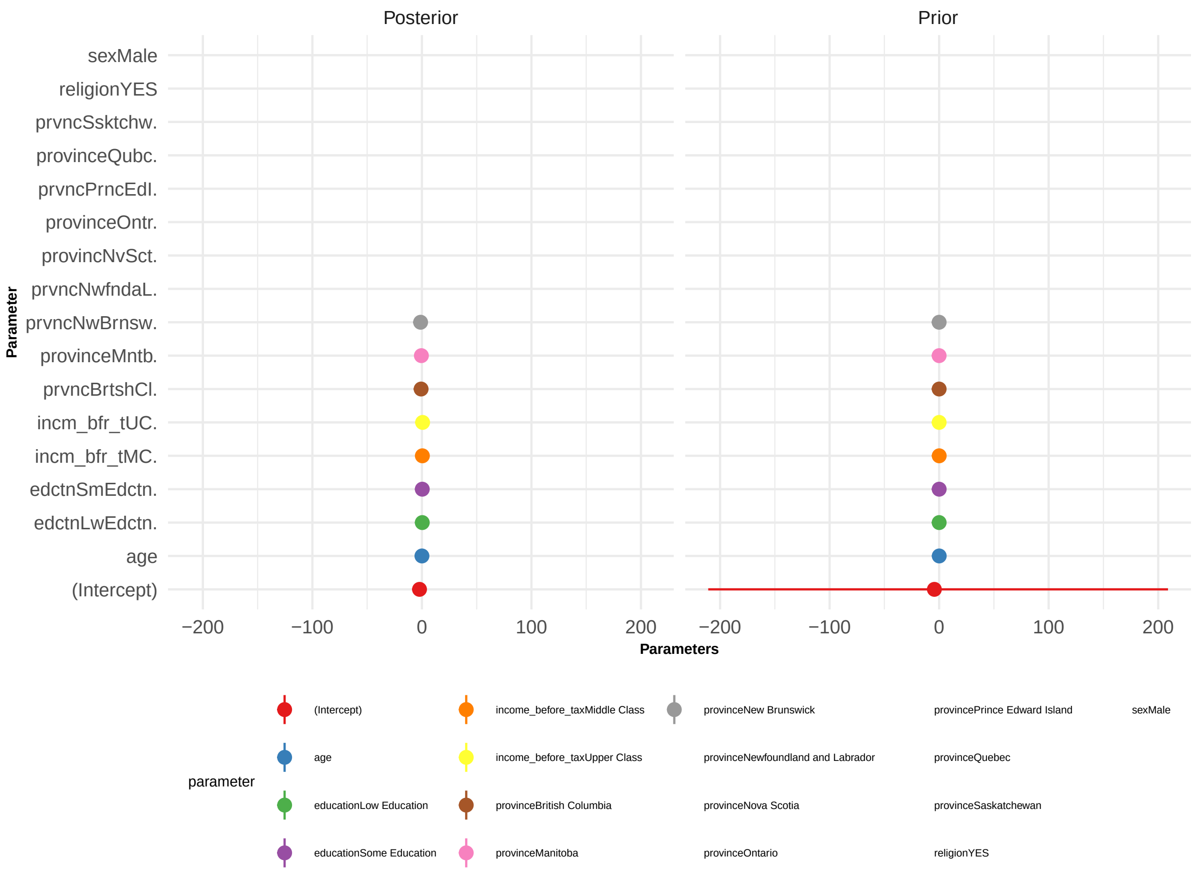
Appendix

6 Additional data details

7 Model details

7.1 Posterior predictive check


Posterior
Prior
sexMale
religionYES
prvncSsktchw.
provinceQubc.
prvncPrncEdI.
provinceOntr.
provincNvSct.
prvncNwfndaL.
prvncNwBrnsw.
provinceMntb.
prvncBrtshCl.
incm_bfr_tUC.
incm_bfr_tMC.
edctnSmEdctn.
edctnLwEdctn.
age
(Intercept)
Parameter
−200
−100
0
100
200
−200
−100
0
100
200
Parameters
parameter
(Intercept)
age
educationLow Education
educationSome Education
income_before_taxMiddle Class
income_before_taxUpper Class
provinceBritish Columbia
provinceManitoba
provinceNew Brunswick
provinceNewfoundland and Labrador
provinceNova Scotia
provinceOntario
provincePrince Edward Island
provinceQuebec
provinceSaskatchewan
religionYES
sexMale



?@fig-liberal-model

?@fig-conservative-posterior

7.2 Diagnostics

References

Barik, A. 2023. "Citation and Authors." *Latest Federal Opinion Polls Canada 2025: Game over for Justin Trudeau?**. The PoliticalPulse. %5Bhttps://politicalpulse.net/politics/federal-opinion-polls-canada/%5D.

CES. 2023. "Citation and Authors." *The Canada Election Study*. The University of Chicago. http://www.ces-eec.ca/.

Comtois, Dominic. 2022. *Summarytools: Tools to Quickly and Neatly Summarize Data*. https://github.com/dcomtois/summarytools.

Elections, Canada. 2023. "Past Elections." – *Elections Canada*. https://www.elections.ca/content.aspx?section=ele&dir=pas&document=index&lang=e.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2020. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

GSS. 2022. "Citation and Authors." *The General Social Survey*. The University of Chicago. https://gss.norc.org/.

Nakhaie, R. 2006. "Electoral Participation in Municipal, Provincial and Federal Elections in Canada." *Cambridge Cor*. Cambridge Core. https://www.cambridge.org/core/publications/journals%5D(https://www.cambridge.org/core/publications/journals.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Robinson, David, Alex Hayes, Simon Couch, Max Kuhn, and tidymodels contributors. 2021. *Broom: Convert Statistical Analysis Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4.

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

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://tidyr.tidyverse.org.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.