Estimating the 2029 Canadian Popular Vote: A Poststratification Approach

STA304 - Assignment 2

Mariana Garcia Mejia

2025-11-13

1 Introduction

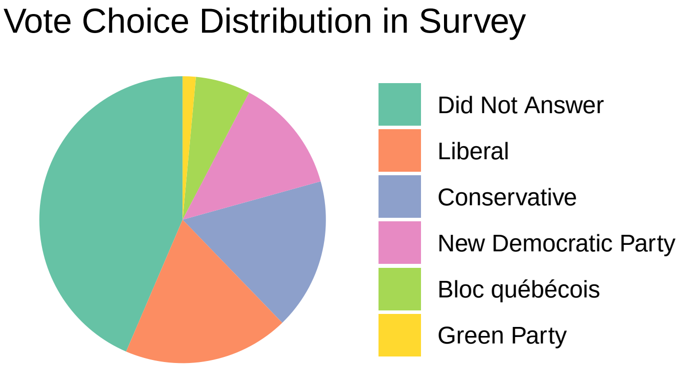
Understanding electoral trends is important not only for political parties but also for voters, businesses, and researchers who rely on predictions to inform decision-making. In Canada, federal elections are scheduled approximately every four years, providing regular opportunities to assess shifts in public opinion and policy making. This report estimates the popular vote for the 2029 Canadian federal election through poststratification and logistic regression, and compares the projected outcomes to the 2025 results, focusing on the Liberal, Conservative, and New Democratic Parties. It is important to note that any forecast is limited by the data available at the time (Williams & Reade , 2016). As a result, predictions of this report will be based on and limited by the 2023 Canadian Election Study survey data from the Consortium on Electoral Democracy (CES, 2023) and the census data from 2021 (Census of Canada, 2021).

Canada's federal elections use a first-past-the-post system, where the party that wins the most seats in the House of Commons forms the government, even if it does not receive a majority of the popular vote. Canadians vote for a representative of the party of choice in their local electoral district, or riding, rather than directly for prime minister. A single member of parliament is selected from each riding, and the leader of the party that wins the most parliament seats becomes prime minister. Because Canada is a multiparty system, it is possible for a party to win the most votes nationwide but not the most seats, as happened in the 2021 election when the Conservative Party received a higher share of the popular vote (33%) than the Liberal Party (32%), but the Liberals won more seats (160 vs. 119), allowing Justin Trudeau to remain prime minister (Heard, 2021). The distribution of votes across ridings, rather than the overall popular vote, is what ultimately determines the prime minister, and when no party wins a majority of seats, a minority government is formed requiring cooperation between parties to pass legislation (BBC, 2025).

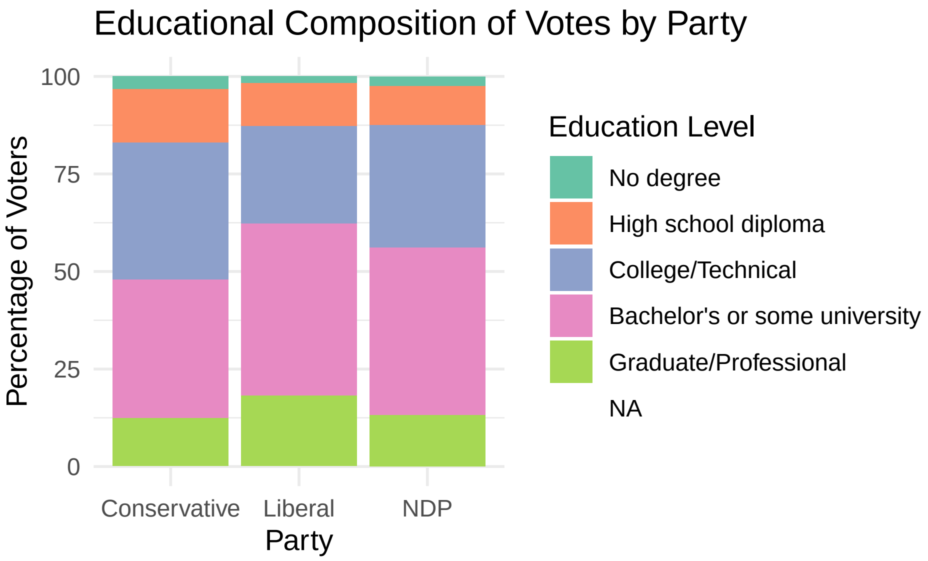
Although the popular vote does not entirely determine who becomes prime minister, it provides a useful overview of which party is leading nationally and offers insights for political parties, businesses, and individuals. As a result, this report aims to predict the popular vote for the three main parties—the Liberal Party, the Conservative Party, and the New Democratic Party (NDP)—in the next Canadian federal election, scheduled for 2029. There are two other large parties, the Bloc Québécois and the Green Party, however we will only focus on the three leading ones. The prediction process will be conducted in two steps. First, a logistic regression model will be fitted to survey data to estimate the probability of voting for each party based on age group, gender, province, and education level. Second, a poststratification analysis will scale these proportions to the population level using census data. Finally, the predicted results will be compared to the 2025 election outcomes to evaluate potential differences and assess the model's reliability given the information available.

2 Data

The poststratification analysis will employ two datasets: the 2023 Canadian Election Study from the Consortium on Electoral Democracy (CODEM) and the 2021 Census of Canada. The survey includes responses from 20,968 Canadians, with data quality ensured by removing duplicates, failed attention checks, and invalid postal codes (CES,2023). The key variables of interest are vote choice, province, gender, age, and education, which were selected for their consistency with available census variables and their relevance in explaining voting behavior. Education has been shown to be an important predictor, as those with college education are more likely to vote than those without (Ahearn, Brand & Zhou, 2022), as well as age and gender (Aribowo, Nurbasari & Hadianto, 2021). The same demographic variables were extracted from the census data, except for vote choice which is only present in the survey data.



We start by looking at the survey data before cleaning to assess the overall votes. Before cleaning, the survey data showed that about 41% of respondents did not report their vote choice, either skipping the question or selecting "Prefer not to answer." This low response rate suggests potential nonresponse bias, meaning the survey results may not fully represent the broader population's preferences. After cleaning (removing respondents with missing education, non-binary gender, or unknown/underage entries) the response rate improved, with 70% reporting a vote choice. As shown in Figure 1, the Liberal Party received the highest share of votes, followed closely by the Conservative Party, with the New Democratic Party (NDP) in third place, mirroring the 2025 election results (CBC, 2025). Figure 2 illustrates that individuals with higher education levels (Bachelor's degree or above) were more likely to support the Liberal Party, consistent with previous research (Kiss, Polacko & Graefe, 2023). Similarly, over half of NDP voters hold higher education credentials, while more than 50% of Conservative voters have less than a Bachelor's degree.



Next, we compare the cleaned data sets. We observe the survey is slightly more representative of Female than Male respondents. Additionally, the share of Non-Binary individuals in the survey data was very small, making their exclusion reasonable for consistency with census data. Table 2 suggests that voting preferences vary by gender, suggesting that both gender and education (as observed in Figure 2) could be strong predictors in the logistic regression models. Interestingly, slightly more women reported their vote choice than men. Among those who did, the Liberal party had the most votes, while the Conservative party had the highest share of votes among men. Additionally, a notably larger share of women voted for the NDP, and about 12–15% of both genders supported other parties, votes that will not be captured in the poststratified estimates.

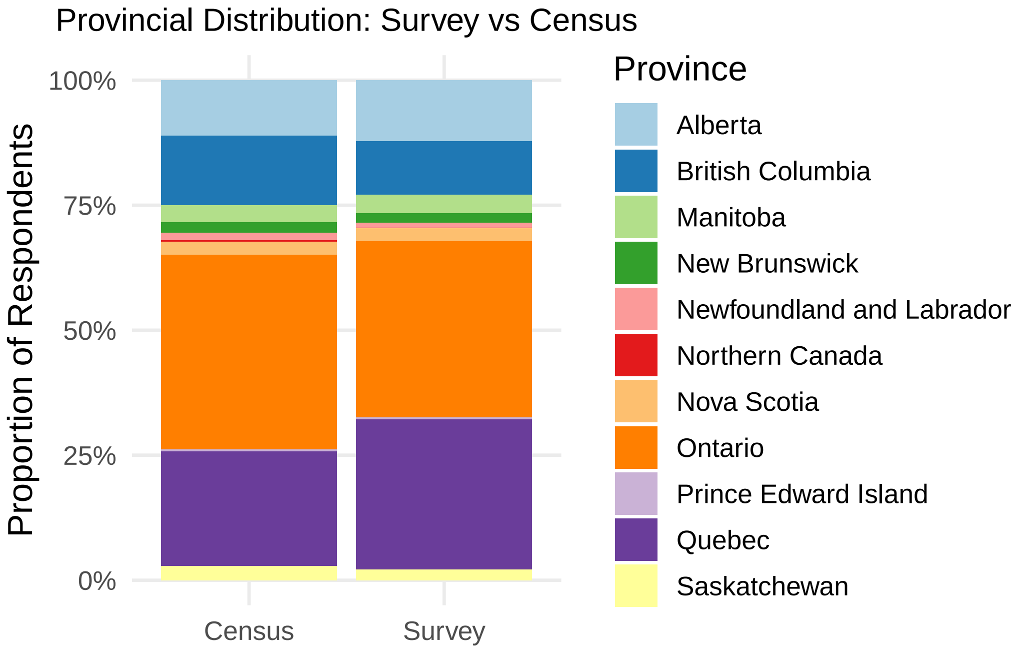
Table 1: The proportions of different Genders in the Census vs. Survey data.

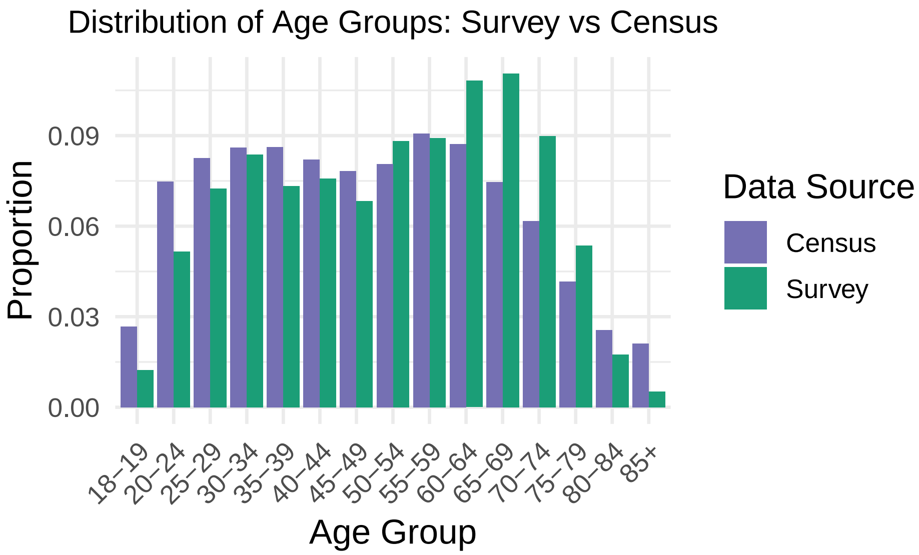
	Male	Female	Non-Binary
Survey data	4062	4974	43
Census data	381266	398787	0

Table 2: The proportions of each Gender that reported their vote in the survey data and percentage of votes for the main parties.

Gender	% That reported vote choice	% Liberal	% Conservative	% NDP	% Other
Male	68.7%	33.5%.	35.3%	15.6%	15.6%
Female	69.7%	33%	26%	29%	12%

Other variables used to predict the proportion of votes for each party were province and age group. Figures 4 and 5 illustrate how well the survey data represents the population captured by the census. The main difference in provincial representation is that the cleaned survey slightly overrepresents Quebec and underrepresents Ontario, the two most populous provinces. British Columbia also appears somewhat underrepresented. In terms of age distribution, both datasets show similar overall patterns, but the survey skews older age groups, with a mode around the 65–69 age group, whereas the census peaks around 55–59. This suggests that older respondents were somewhat more likely to participate in the survey. While these differences are not extreme, they may influence the weighted estimates in the poststratification step and will be considered when interpreting the final predictions.





2.1. Data Cleaning.

The cleaning process aimed to ensure consistent formatting between the survey and census datasets to enable poststratification. All cleaning was performed in `R (version 4.0.2)` using functions from the `tidyverse` package and guided by each dataset's documentation. In the survey data, numeric province codes were replaced with province names using `mutate()` and `casewhen()`, and northern territories were combined into `Northern Canada` consistent with census data. Gender was recorded into a binary variable in a new `male` column and observations with non-binary or undeclared gender were dropped to match the census data. The age variable was grouped into `agegrp` to match census age group categories, and a binary indicator column was created for each party's vote choice (`vote_liberal`, `vote_conservative`, `vote_ndp`, `vote_bc` and `vote_gp`). The education variable required harmonization across both datasets, so five categories were created (`No schooling`, `High school diploma`, `College/Technical`, `Bachelor's or some university`, and `Master's/Professional`) using grouping rules detailed in the Appendix. Lastly, `filter()` was used to remove observations with age under 18 or undeclared, or missing education, and `select()` was used to extract the relevant columns. Only 0.65% of survey observations were dropped during the cleaning process, and about 20.5% of the census observations.

3 Methods

3.1 Model Specifics

The predicted proportion of votes for each party will be estimated using a two-step process. First, we fit three logistic regression models to the survey data to predict the probability of voting for the Liberal, Conservative, or NDP party based on Age group, Gender, Province, and Education. Logistic regression is the most suitable model when the dependent variable is binary, in our case, whether or not a respondent voted for a given party. It also allows to include the previously mentioned variables by converting them into indicator variables, with one category for each variable omitted as the baseline. For instance, the baseline group in our models includes females from Alberta with no degree in the age group 17-18.

The general model estimated for each party is:

$$\begin{aligned}\log\left(\frac{p_{\text{party}}}{1-p_{\text{party}}}\right) = {} & \beta_0 + \beta_1 I_{\text{age:20-24}} + \beta_2 I_{\text{age:25-29}} + \beta_3 I_{\text{age:30-34}} + \beta_4 I_{\text{age:35-39}} + \beta_5 I_{\text{age:40-44}} \\ & + \beta_6 I_{\text{age:45-49}} + \beta_7 I_{\text{age:50-54}} + \beta_8 I_{\text{age:55-59}} + \beta_9 I_{\text{age:60-64}} + \beta_{10} I_{\text{age:65-69}} \\ & + \beta_{11} I_{\text{age:70-74}} + \beta_{12} I_{\text{age:75-79}} + \beta_{13} I_{\text{age:80-84}} + \beta_{14} I_{\text{age:85+}} + \beta_{15} I_{\text{male}} \\ & + \beta_{16} I_{\text{BC}} + \beta_{17} I_{\text{MB}} + \beta_{18} I_{\text{NB}} + \beta_{19} I_{\text{NL}} + \beta_{20} I_{\text{NC}} \\ & + \beta_{21} I_{\text{NS}} + \beta_{22} I_{\text{ON}} + \beta_{23} I_{\text{PEI}} + \beta_{24} I_{\text{QC}} + \beta_{25} I_{\text{SK}} \\ & + \beta_{26} I_{\text{HighSchool}} + \beta_{27} I_{\text{College}} + \beta_{28} I_{\text{University}} + \beta_{29} I_{\text{Graduate}}\end{aligned}$$

Here, p_{party} represents the probability of voting for a given party. The coefficients $\beta_1, \beta_2, ...$ represent the estimated change in log-odds of voting for a given party for that category relative to the baseline group. For example, β_{15} captures the change in log-odds of voting for the given party for males compared to females. Lastly, the value for each variable is denoted by an I, as it can only take values 0 or 1 due to being categorical. It is important to note that because the logistic model expresses results in log-odds, we must solve for p_{party} to obtain the actual predicted probability of voting for that party.

3.2 Post-Stratification

Poststratification is a statistical adjustment technique used to improve the accuracy of survey estimates, especially when the sample is not perfectly representative of the population. This step is crucial because the survey, despite having a large sample size, slightly overrepresented certain demographics (e.g., younger respondents and some provinces).

The procedure has three steps. First, the data was partitioned into 1650 demographic cells with all the unique combinations of demographic variables and denoted using j in the formula below. An example of a cell is Females with ages 18-19, from Quebec with a Master's or Doctorate degree. Then, the logistic regression model was used to estimate the probability of voting for each party based on the demographic characteristics of that specific cell. Lastly, the formula below and the census data were used to aggregate the cell-level estimates up to a population-level estimate, by weighting each cell estimate by its proportion in the population.

$$\hat{p}^{PS}_{party} = \frac{\sum_j N_j \hat{p}_{party}}{\sum_j N_j}$$

We use this formula to calculate the final predicted proportions of votes for each party in the population, $\hat{p}^{PS}_{party}$. Here, $\hat{p}_{party}$ is the estimated proportion of votes for the given party for the jth demographic cell. N_j represents the counts of individuals in the census data with the characteristics of the jth cell. The formula is weighting the probability of each cell by how large that cell is in the population, allowing for population estimates that more accurately reflect the true demographic composition of Canada rather than the survey sample.

4 Results

4.1. Logistic Regression Results

	Liberal	Conservative	NDP
(Intercept)	-2.961***	-0.266	-0.083
	(0.425)	(0.336)	(0.340)
as.factor(agegrp)8	0.437	-0.410	-0.015
	(0.386)	(0.312)	(0.271)
as.factor(agegrp)9	0.502	-0.563+	-0.239
	(0.378)	(0.306)	(0.267)
as.factor(agegrp)10	0.931*	-0.249	-0.601*
	(0.374)	(0.298)	(0.268)
as.factor(agegrp)11	0.864*	-0.332	-0.516+
	(0.375)	(0.300)	(0.269)
as.factor(agegrp)12	0.848*	0.130	-0.942***
	(0.376)	(0.296)	(0.275)
as.factor(agegrp)13	1.116**	-0.026	-1.224***
	(0.375)	(0.298)	(0.281)
as.factor(agegrp)14	1.059**	0.096	-1.327***
	(0.372)	(0.292)	(0.275)
as.factor(agegrp)15	1.178**	0.262	-1.588***
	(0.372)	(0.291)	(0.282)
as.factor(agegrp)16	1.230***	0.018	-1.391***
	(0.370)	(0.289)	(0.273)
as.factor(agegrp)17	1.301***	0.133	-1.687***
	(0.369)	(0.288)	(0.278)
as.factor(agegrp)18	0.924*	0.185	-1.518***
	(0.374)	(0.293)	(0.284)
as.factor(agegrp)19	1.135**	0.351	-1.621***
	(0.381)	(0.301)	(0.307)
as.factor(agegrp)20	1.551***	0.262	-1.628***
	(0.416)	(0.353)	(0.408)
as.factor(agegrp)21	1.259*	0.628	-2.966**
	(0.529)	(0.462)	(1.047)
male	0.037	0.476***	-0.434***
	(0.060)	(0.063)	(0.073)
provBritish Columbia	0.208	-0.836***	0.439***
	(0.130)	(0.121)	(0.131)
provManitoba	0.274	-0.765***	0.291
	(0.179)	(0.174)	(0.188)
provNew Brunswick	0.768***	-1.047***	-0.235
	(0.218)	(0.245)	(0.282)
provNewfoundland and Labrador	0.734*	-1.142**	0.817**
	(0.307)	(0.364)	(0.310)
provNorthern Canada	1.227*	-1.113+	0.046
	(0.555)	(0.669)	(0.676)
provNova Scotia	0.828***	-0.816***	0.134
	(0.191)	(0.205)	(0.215)
provOntario	0.607***	-0.604***	-0.048
	(0.102)	(0.090)	(0.110)
provPrince Edward Island	0.192	-0.887+	-1.585
	(0.575)	(0.535)	(1.045)

	Liberal	Conservative	NDP
provQuebec	0.207+	-1.516***	-0.700***
	(0.106)	(0.102)	(0.121)
provSaskatchewan	-0.154	0.328+	-0.051
	(0.249)	(0.193)	(0.242)
educationHigh school diploma	0.412+	-0.165	-0.200
	(0.225)	(0.197)	(0.239)
educationCollege/Technical	0.272	-0.132	0.030
	(0.215)	(0.186)	(0.223)
educationBachelor's or some university	0.769***	-0.416*	-0.147
	(0.213)	(0.185)	(0.221)
educationGraduate/Professional	0.908***	-0.593**	-0.106
	(0.220)	(0.197)	(0.233)
Num.Obs.	6224	6224	6224
AIC	7114.7	6564.8	5545.7
BIC	7316.8	6766.8	5747.8
Log.Lik.	-3527.372	-3252.380	-2742.875
RMSE	0.44	0.42	0.37

Note: ^^ + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

The results of the logistic regression models estimating the probability of voting for each major party based on Age, Gender, Province and Education can be found above. Several coefficients were statistically significant and aligned with observed patterns in Section 2. In the Liberal model, older age groups showed significantly higher log-odds of Liberal support compared to the youngest group, with coefficients increasing steadily from age 30–34 through 80–84. Provincial effects were also notable, with higher support observed in Ontario and Nova Scotia. Lastly, education was positively associated with Liberal voting: respondents with graduate or professional degrees had, on average, 0.91 higher log-odds of supporting the Liberal Party compared to those with no schooling, holding all other else constant.

In contrast, Gender appeared significant in the Conservative model, where male respondents had 0.48 higher log-odds of supporting the Conservative Party compared to female respondents, holding all other variables constant. Additionally, the party observed a significantly lower support in Quebec compared to Alberta holding all else fixed. Lastly, for the NDP, male respondents were less likely to support the party, and support declined sharply with age, as seen in large negative coefficients for older age groups. For example, the Age group 45–49 had 1.22 lower log-odds of voting for the NDP, than the 17-18 age group holding all else fixed. Regionally, NDP support was strongest in British Columbia, but significantly lower in Quebec. Overall, these results suggest that age, education, and province were strong predictors of party preference, while gender effects were party-dependent.

4.2. Poststratification Results

Table 4: The proportions of popular vote for the Liberal, Conservative and New Democratic Party (NDP).

	Liberal Party	Conservative Party	NDP Party
Predicted Percentage of Popular Vote	24.3%	27.3%	20.1%

The table above displays the predicted proportions obtained for each party after the poststratification. While some votes remain unaccounted for due to the exclusion of smaller parties such as the Bloc Québécois and the Green Party, the missing share (around 30%) is higher than expected compared to the 12–15% expected missing share from the raw survey by gender. Interestingly, Table 4 shows the Conservative Party with the highest estimated proportion of votes, even though the Liberal Party led in the survey. This shift is reasonable because post-stratification adjusts for how demographic characteristics are distributed in the population. So it is possible that while the survey sample contained more Liberal voters, the census-weighted population included a greater share of groups that tend to vote Conservative. Lastly, the NDP's estimated support increased relative to the survey, though it remained below that of both major parties. Overall, these results appear consistent with expectations and highlight how accounting for demographic composition of the population can substantially alter survey-based voting proportions.

5 Discussion

Table 5: 2025 Election Results (Elections Canada)

	Liberal Party	Conservative Party	NDP Party
Percentage of Popular Vote	43.7%	41.3%	6.3%

5.1. Overview

When comparing the estimated proportions from the poststratification analysis with the most recent 2025 election results, we observe notable differences. First, both the Liberal and Conservative parties received substantially higher proportions of votes in the actual election (43.7% and 41.3%) than in our estimates (24.4% and 27.2%). Several technical factors could explain these gaps. Since the poststratification process depends on the representativeness of the survey data, the small imbalances observed in Section 2 may have biased the weighted estimates. Moreover, the assumptions of independence and linearity within the regression model may not hold perfectly in the electoral data leading to inaccurate estimates.

A large discrepancy appears in the NDP's predicted support, which our model estimated at approximately 20% compared to only 6% in the 2025 results. This difference may be due to over-weighting demographics that are more likely to express support for the NDP, but less likely to turn out on election day, like for example younger individuals. It also highlights how applying a uniform model across parties may have ignored key predictors that influence each party's vote share differently. We must also acknowledge that there are differences between stated preferences in surveys and actual voting, which is often influenced by last-minute shifts, so it might be reasonable for survey data and actual outcomes to be different.

5.2. Limitations

Although the findings of this report are valuable, several limitations must be acknowledged. A significant number of observations had to be removed from both the survey and census datasets to ensure consistency, mainly due to missing information on education, age or vote choice, which meant that some high-quality data could not be used. Still, the initial datasets were large enough to maintain adequate sample sizes after cleaning. In addition, the regression model relies on assumptions such as the independence of observations and a linear relationship between predictors and the response variable, meaning that any violations may reduce the accuracy of the estimates obtained. We also assumed that the same predictors are relevant to predicting the probability of voting for the Liberal, Conservative and NDP party, which might not be the case. Finally, while the popular vote offers insight into general electoral trends, it might not be the most important predictor of election outcomes in Canada's first-past-the-post system, where success depends on winning seats in the House of Commons through individual ridings, rather than achieving a high overall vote share (Zadorsky, 2025).

5.3. Future Recommendations

Future research in this area or based on this report should focus on addressing the limitations described previously. Rather than deletion of missing variables, as was used in this report, different imputation techniques should be considered to make the best use of the data available. For example, using the sample data to predict sex in the census, could be explored to avoid dropping non-binary observations (Kennedy et al., 2022). Moreover, future work should assess the assumptions underlying a logistic regression model in the context of electoral data to improve the reliability of these results. In addition, further work could explore popular vote in specific areas or ridings,to gauge the number of seats won by each party, rather than just estimating popular vote across all Canada.

Another recommendation would be to perform hypothesis tests on the final proportion estimates to assess whether they are statistically significantly different from those of the 2025 election, and increase the robustness of the comparison performed earlier in the discussion, as well as using more recent data when predicting 2029 results; the next Census will take place in 2026. Finally, future research should aim to develop party-specific models that incorporate additional variables to better capture the distinct factors shaping each party's support and improve the overall predictive accuracy of election forecasts.

6 Generative AI Statement

I made use of Generative AI, more specifically ChatGPT, to help me summarize my ideas into more concise sentences, find alternative vocabulary for redundant words, and to polish my graphs to look more professional. For example, on some occasions I struggle to write concisely and to avoid repeating myself so much, so I gave ChatGPT the parts where I was struggling to be concise, and asked to refine the sentence structure.

Additionally, I used ChatGPT to find synonyms for overly repeated connectors like "therefore" and it returned a new way to write the sentence. I also used generative AI to improve my code, especially my graphs. For example, I asked ChatGPT how to change the color of the pie chart and the stacked bar graph for all my plots to look more professional. I also used it when I was stuck cleaning the data. For example, I wasn't sure how to use the mutate() function to divide the data into Education Groups, and ChatGPT recommended which groups to create based on the available categories in the census data and in the survey data.

Since I recognize the limitations of AI to produce content and the risk of hallucinations, I did not use AI to produce complete paragraphs, full chunks of code, brainstorm research question ideas or do any research on the topic. Moreover, I made sure to review the relevant content as I included it in my text, and do my own research and analysis of the data before using AI for any purpose. As a result, I believe my submission is fully my own work and that I only used AI to be more efficient and to make the submission better than what it was initially.

7 Ethics Statement

To ensure the analysis is fully reproducible, all code is clearly documented in the Qmd file using `#` comments within each R code chunk, explaining the purpose of every major step. Additionally, the workflow described in this report follows a structured sequence, from data cleaning to modeling and poststratification, so others can easily trace and replicate the analysis. The Data Cleaning section (Section 2) provides a detailed description of how the raw data was processed (both for census and survey), while additional details such as the education recoding scheme are included in the Appendix for reference. The report explicitly lists all R packages and functions used (e.g., `tidyverse`, `glm()`, `filter()`), along with the corresponding formulas and definitions of each variable. Furthermore, standardized statistical methods were thoroughly described, and since there were no random processes in our analysis, anyone reproducing the analysis should get the same results given they use the same datasets, which are referenced in the Bibliography section of this report.

Aside from reproducibility, another important ethical concern is the source of the data and how it was collected, especially when talking about surveys. Since survey and census data both involve demographic information of living human beings, we must assess our responsibilities towards the respondents of the survey and participants of the census when using their data for an analysis. The Research Ethics Board at the University of Toronto is in charge of reviewing the ethical aspects of research studies done at the university and it requires commencement of review protocol for a project if it involves living human beings and the secondary use of personal information (even if it is de-identifiable). However, there are some exceptions, and since the CES2021 survey data is publicly available our report falls under Exception 1: Research that relies exclusively on information publicly available through a mechanism set out by a regulation and that is protected by law (University of Toronto, 2022).

Even though this project would likely not go through a REB because the survey data is publicly available, it is still important to consider our responsibilities when using this data, namely the fact that information can be linked to a specific individual even if it is not their names. Although the survey did not contain names or contact information, this report avoided including information like Zip code to ensure anonymity. Additionally, the survey data had thousands of observations, making it very difficult to identify specific individuals from summary statistics which compose a big part of this report. Considering that vote choice can be very personal information, these decisions were taken to prioritize the right to privacy of the respondents by ensuring they cannot be identified from our report. Lastly, it is important to mention that the respondents of the survey were volunteers and that they had the choice to not disclose their party of choice by selecting the "Prefer not to answer" option.

8 Bibliography

1. Activities exempt from human ethics review. (2022). Utoronto.Ca. Retrieved November 8, 2025, from https://research.utoronto.ca/ethics-human-research/activities-exempt-human-ethics-review

2. Ahearn, C.E., Brand, J.E. & Zhou, X. (14 September, 2022). How, and For Whom, Does Higher Education Increase Voting? Research in Higher Education 64, 574–597 (2023). https://doi.org/10.1007/s11162-022-09717-4

3. Aribowo, A., Nurbasari, A., & Hadianto,B. (December 28, 2021). Participation in Voting Parties Based on Gender and Ages. Journal of Social and Political Sciences, Vol.4 No.4 (2021), Available at SSRN: https://ssrn.com/abstract=3995065

4. BBC News. (2025, March 10). How does Canada's general election work? A simple guide. BBC. https://www.bbc.com/news/articles/cwydlr3reqpo

5. Canada Votes. (May 23, 2025.). CBC News. Retrieved November 8, 2025, from https://newsinteractives.cbc.ca/elections/federal/2025/results/

6. Census of Canada, 2021: individual public use microdata file. [dataset].

7. Heard, A. (n.d.). Canadian election results: 1867-2021. Sfu.Ca. Retrieved November 8, 2025, from https://www.sfu.ca/~aheard/elections/1867-present.html

8. Kennedy, L. et al. (23 March, 2022) He, she, they: Using sex and gender in survey adjustment. Retrieved November 9, 2025 from https://arxiv.org/pdf/2009.14401

9. Kiss, S., Polacko, M., & Graefe, P. (October 25, 2023) Analysis: Educated voters in Canada tend to vote for left-leaning parties while richer voters go right. Labour Studies; McMaster University. https://labourstudies.mcmaster.ca/news-article-example-1/

10. R Core Team. (2020). R: A language and environment for statistical computing (Version 4.0.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

11. Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2021 Canadian Election Study. [dataset]

12. Williams, L. V., and Reade, J. J. (11 November, 2015) Forecasting Elections. Journal of Forecasting, 35: 308–328. doi: https://doi-org.myaccess.library.utoronto.ca/10.1002/for.2377

13. Zadorsky, J. (March 26, 2025). Expert explainer: How Canadian federal elections work. Westernu.Ca. Retrieved November 8, 2025, from https://news.westernu.ca/2025/03/how-canadian-federal-elections-work/

9 Appendix

Education categories from each data set that were merged into new `education` categories for consistency: * Variable name in Census data: `hdgree` * Variable name in Survey data: `cps21_education`

Category	Census codes & labels	Survey codes & labels
No schooling	1 — No certificate, diploma or degree	1–4 — No schooling to some high school
High school diploma	2 — High school diploma or equivalent	5 — Completed secondary/high school
College/Tech.	3–7 — College/Technical programs	6–7 — Some or completed college/CEGEP
Bachelor's or some university	8–10 — University below to above bachelor level	8–9 — Some university or bachelor's
Master's/Prof.	11–13 — Professional, master's, or doctorate	10–11 — Master's or professional degree
Removed categories	88, 99 — Not available/applicable	12 — Don't know/Prefer not to answer