Predicting the 2029 Canadian Federal Election Popular Vote Using Post-Stratification

STA304 - Assignment 2

GROUP 51: CAROL XU, ISLA CHEONG, MIA SUNG, VANESSA GABRIELA KWOK

2025-11-13

1 Introduction

Strong demographic and regional variations significantly impact Canadian federal elections, which complicates the estimation of voting results relying solely on survey data. This report seeks to forecast the national popular vote—the share of votes received by each party nationwide—for the approaching 2029 federal election. To achieve this, we merge individual-level survey responses from the Canadian Election Study (CES) with population figures from the 2021 Canadian Census to produce nationally representative estimates (Stephenson et al., 2022; Statistics Canada, 2023).

The ability to anticipate voting results is quite beneficial since it can reveal the larger political landscape, aiding researchers and policymakers in grasping voter perspectives through numerical data about party inclinations (Barnfield et al., 2025). Even though Canada selects governments via a first-past-the-post system (Albert, 2024), the popular vote continues to be an important measure in political discourse and scholarly studies as it indicates national backing for significant parties including the Liberal Party, Conservative Party, and New Democratic Party. In this report, popular vote denotes the percentage of all legitimate votes cast across the country for a specific political party, while post-stratification indicates the method of modifying model forecasts to align with the demographic makeup of the population.

The research question guiding this analysis is: What will the popular vote distribution look like for the Liberal, Conservative, and NDP parties in the 2029 Canadian federal election? To answer this, we apply post-stratification to the results of a regression model that estimates how voting intention varies with demographic factors. Age and province are included as predictors because trends from recent elections suggest notable generational and regional cleavages in party support.

Before conducting the statistical analysis, we proposed the following hypotheses: (1) older voters are more likely to support the Conservative Party; (2) younger voters are more likely to

support the NDP; and (3) meaningful variation in voting preferences exists across provinces. The remaining sections of this report outline the data sources used, describe the modeling and post-stratification procedures, present the projected popular vote estimates, and discuss the interpretation, limitations, and future directions of this work.

2 Data

2.1 Data Description

The analysis uses two publicly available datasets: the 2021 Canadian Election Study (CES) and the 2021 Canadian Census. The CES is a nationally distributed survey that collects information on Canadians' political attitudes, voting intention, and demographic characteristics during federal election cycles. It utilizes sampling and weighting techniques driven by probability to ensure extensive participation from eligible voters across different areas (Stephenson et al., 2022). The 2021 Census offers comprehensive demographic figures for the Canadian populace and is conducted by Statistics Canada every five years to assess age, gender, education, and geographical distribution (Statistics Canada & Statistics Canada, 2023). Together, the CES and Census allow us to combine behavioural data (voting intention) with population counts to correct for survey non-representativeness using post-stratification.

2.2 Data Cleaning

Both datasets required preprocessing before analysis. First, we extracted from the CES only the variables needed for modeling: age, province, gender, education, and vote intention. Age was converted into 15 five-year categories (18–19 to 85+), provinces were standardized to Statistics Canada codes, and the three northern territories were merged into a single "North" category to match Census regional reporting. Gender responses were recoded into a binary classification ("Woman" or "Man") to match the categories available in the census data used for comparison. Non-binary respondents were excluded because we could not validly classify them within the census framework, and retaining them would require misgendering (Kennedy, Khanna, Simpson, Gelman, 2020). Education levels were collapsed into three categories: high school or less, some college, and college degree. We then generated three binary variables for respondents' intended vote choice: Liberal, Conservative, and NDP. The Census data was cleaned using the same groupings so that each demographic combination — age $\times$ province $\times$ gender $\times$ education — aligned across the two datasets. Observations missing or outside the allowable ranges for the four variables were removed. This process ensures that predictions generated from the survey model can be applied directly to Census population counts.

The variables used in the post-stratification model and their roles are summarized as follows: • Age — demographic predictor strongly linked to political preference trends • Province — captures well-established regional differences in Canadian voting patterns • Gender —

included as a demographic control variable • Education — captures socioeconomic status differences that correlate with political behaviour • Vote intention (Liberal, Conservative, NDP) — outcome variables of interest

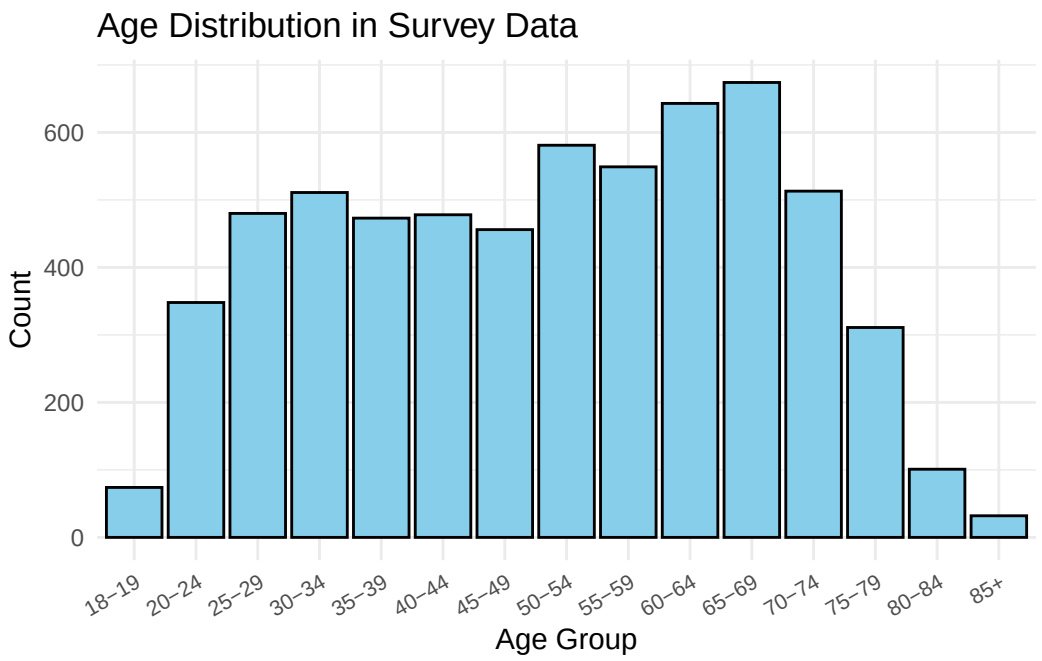
2.3 Numerical Summaries

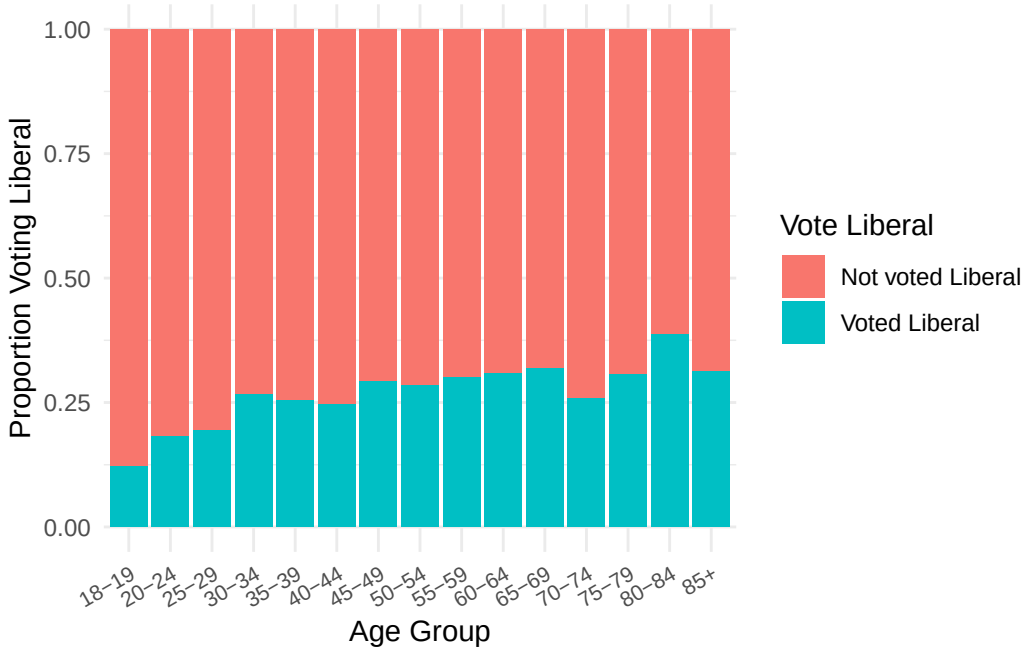
Numerical summaries help illustrate the composition of the CES sample prior to post-stratification. After the cleaning procedure, 6224 respondents remained. The gender distribution is moderately imbalanced, with women making up about 55% of the sample and men about 45%. Provincial representation is uneven: Ontario (36%) and Quebec (30%) constitute the largest shares of respondents, while provinces such as Alberta (12%) and British Columbia (11%) are represented to a lesser extent, and the Atlantic provinces and the North collectively account for only small fractions of the sample.

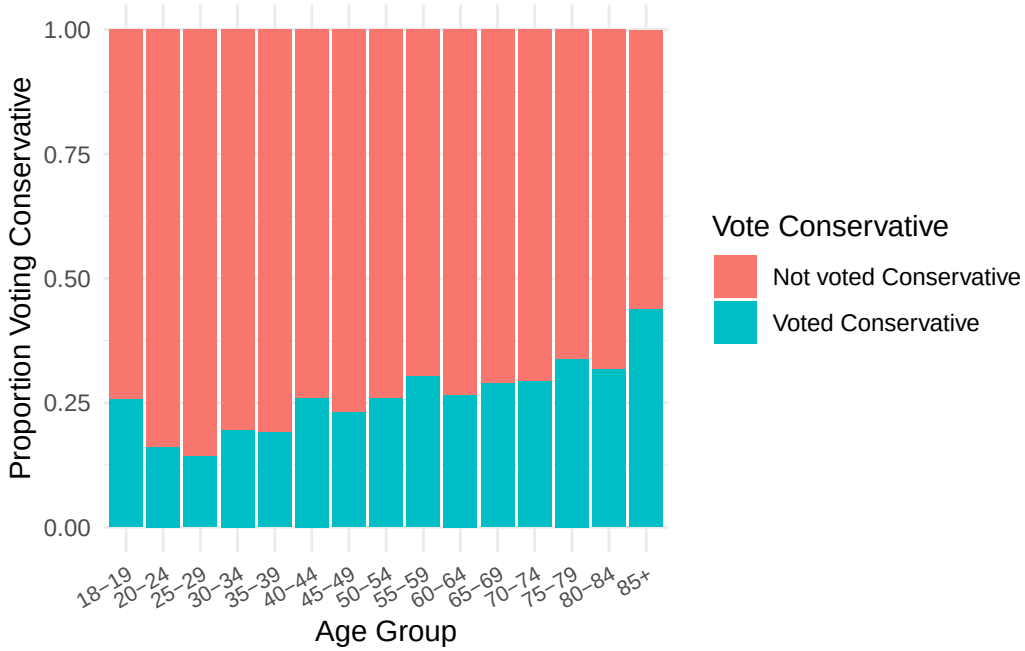
The age distribution is also skewed, with relatively few respondents in the youngest (18–19) and oldest (85+) categories. Participation increases sharply among individuals in their 20s and early 30s, remains substantial through midlife, and reaches its highest levels among those aged 60–69 before declining again among the oldest respondents.

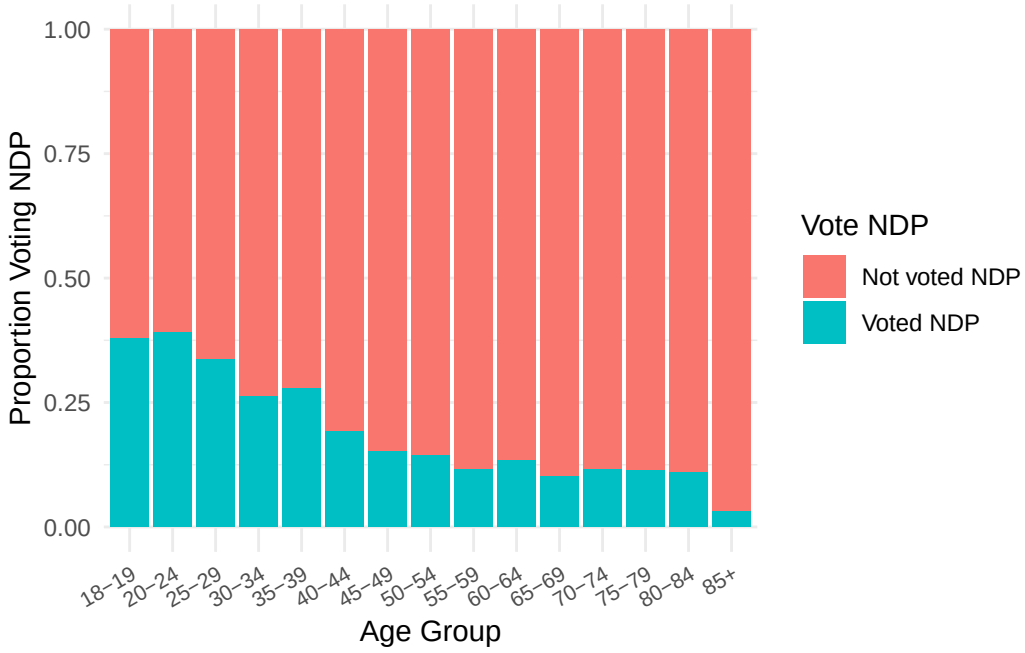
The distribution of education levels is highly concentrated in the two higher categories. Nearly 56% of respondents reported completing some college, and about 44% held a college degree, while only a very small fraction had a high school education or less.

The raw (unadjusted) proportions reporting vote intention were approximately 27.2% for the Liberal Party, 24.8% for the Conservative party, and 18.6% for the NDP. These figures suggest a competitive partisan landscape within the sample, with no single party attracting a majority of stated vote intentions.


Age Distribution in Survey Data
Count
600
400
200
0
18–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
70–74
75–79
80–84
85+
Age Group


Proportion Voting Liberal
1.00
0.75
0.50
0.25
0.00
18–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
70–74
75–79
80–84
85+
Age Group
Vote Liberal
Not voted Liberal
Voted Liberal


Proportion Voting Conservative
1.00
0.75
0.50
0.25
0.00
18–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
70–74
75–79
80–84
85+
Age Group
Vote Conservative
Not voted Conservative
Voted Conservative


Proportion Voting NDP
1.00
0.75
0.50
0.25
0.00
18–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
70–74
75–79
80–84
85+
Age Group
Vote NDP
Not voted NDP
Voted NDP

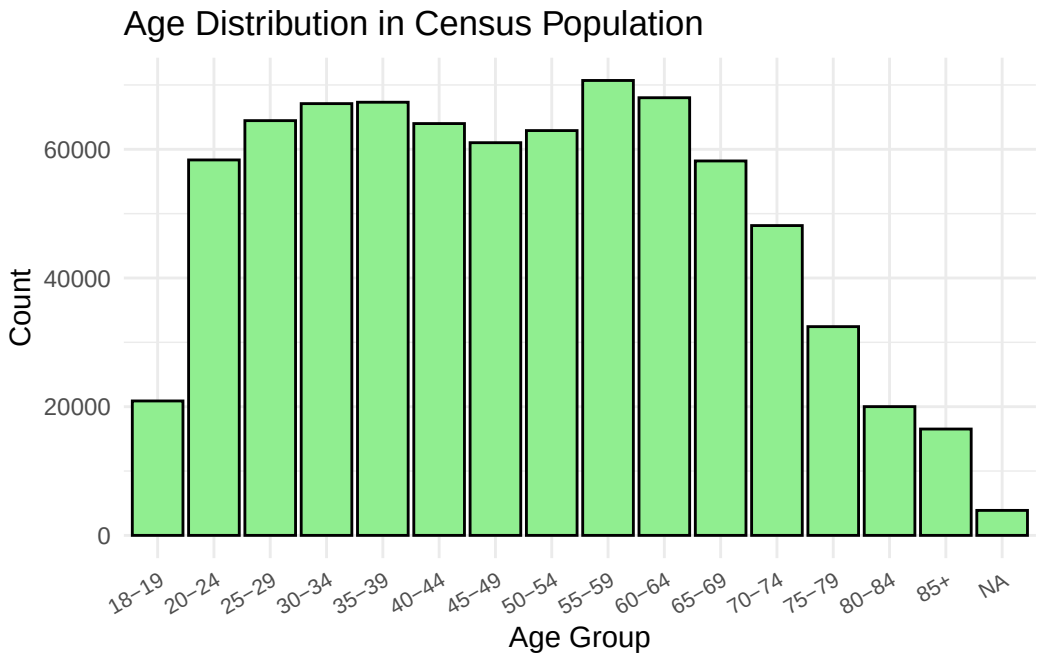


Figure 1 displays the distribution of CES respondents across age categories. The figure shows that the majority of the CES respondents were in their late-30s to late-60s compared to young adults and seniors, highlighting the vitality of adjusting estimates based on population data.

Figure 2 illustrates the proportion of each age category indicating a desire to vote Liberal. Support is lowest among the youngest respondents, increases through early and middle adulthood, and remains relatively stable across the older age groups.

Figures 3 and 4 show how preference for the Conservative and NDP parties differ amongst age categories; Conservatives are most supported by older age groups, whereas the NDP party is most supported by young respondents. These illustrative visualizations encourage the addition of age and province in the regression model and emphasize the demographic gradients that post-stratification seeks to address.

Figure 5 presents the age distribution in the Census population. Compared with the survey sample, the population shows a more even spread across most work-aged categories and substantially larger counts in both the youngest and oldest groups. These differences between the sample and population distributions underscore why post-stratification is necessary: estimates based solely on the survey would overrepresent certain age ranges and misrepresent the national electorate.

All analysis, such as data cleaning, numerical summaries, and visualization, was conducted utilizing the R programming language (version 4.5.1) along with the tidyverse and ggplot2 packages for data manipulation and plotting.

3 Methods

To address our research question—estimating the overall popular vote in the next Canadian federal election—we implemented a two-stage analytic framework. Initially, we created a model of voting intention utilizing survey data to connect demographic traits to political inclinations. Second, we applied post-stratification to combine model predictions with census population counts, producing estimates that better reflect the national electorate rather than the (non-representative) survey sample.

3.1 Model Specifics

Voting intention is a binary outcome (support for a given party vs. not), making logistic regression an appropriate modeling choice because it models probabilities bounded between 0 and 1 and is widely used for categorical dependent variables. Logistic regression also allows us to quantify how demographic factors relate to voting behaviour while enabling prediction for population subgroups.

For the Liberal Party model, we specified:

$$logit(P(Y_i = 1)) = \beta_0 + \beta_1(age_i) + \beta_2(prov_i) + \beta_3(gender_i) + \beta_4(edu_i),$$

where Y_i indicates whether respondent i intends to vote Liberal (1 = yes, 0 = no), β_0 is the intercept, and β_1, β_2, β_3, and β_4 represent the effects of age, province, gender, and education respectively. Parallel logistic regression models were also fit for the Conservative Party and NDP.

This model assumes: (1) the relationship between predictors and the log-odds of voting for a party is linear, (2) observations are independent, and (3) all relevant predictors are measured without systematic error. The parameters of interest are the estimated coefficients and the predicted probabilities from each model, which later enter the post-stratification step.

3.2 Post-Stratification

Post-stratification is a statistical technique that adjusts model-based predictions from a survey to the demographic composition of the full population. In simple terms, it addresses the question: "If each demographic group in Canada votes similarly to our survey and according to the census proportions, what would the election outcome appear like?"

Cells were defined by all combinations of age × province × gender × education, since all four predictors were included in the logistic models (they are expected to be associated with political preference) and were consistently measured across both the survey and census datasets.

For each demographic cell i:

1. Predict the probability of support for each party using the corresponding logistic model.
2. Multiply that probability by the census count of individuals in the cell (N_i).
3. Aggregate predictions across all cells to obtain a national estimate.

The post-stratified estimate for the Liberal party is:

$$\hat{y}^{PS} = \frac{\sum_i P(Y = 1 | age_i, prov_i, gender_i, edu_i) \cdot N_i}{\sum_i N_i}$$

Post-stratification is appropriate here because the survey is not perfectly representative of the Canadian electorate; without adjustment, predictions would be biased toward the survey's demographic composition. By integrating census counts, we generate estimates that better approximate real-world voting intentions.

4 Results

Table 1: Post-Stratified Estimates of Popular Vote for Major Canadian Political Parties

Party	Predicted Popular Vote (%)
Liberal	30.2
Conservative	21.2
NDP	19.9

Table 1 presents the post-stratified national estimates of the popular vote for the Liberal Party, Conservative Party, and NDP. These values result from combining logistic regression predictions with census population totals, producing adjusted estimates that reflect the demographic composition of the Canadian electorate rather than the survey sample alone.

Before post-stratification, the logistic regression models were used to estimate voting intention based on age, province, gender, and education. Across all three models, several predictors showed statistically significant coefficients. In the Liberal model, many age categories exhibited positive and significant associations, while provincial differences also appeared substantial. For the Conservative model, gender had a clear positive association with Conservative support, and several provinces showed strong positive or negative coefficients. In the NDP model, coefficients for older age categories were consistently negative and frequently significant, with multiple provincial indicators also showing strong effects. Education, in contrast, showed no statistically significant associations in any of the models. These results serve as the basis for the predicted probabilities used in the post-stratification step.

Based on the post-stratified results, the predicted popular vote shares are approximately 30.2% for the Liberal Party, 21.2% for the Conservative Party, and 19.9% for the NDP. These estimates indicate that the Liberal Party holds the largest share of predicted support, while the Conservative Party and NDP follow with lower but broadly comparable levels of support. Although the Liberal Party appears to lead, the overall distribution does not reflect a dominant advantage for any one party.

The predicted pattern aligns reasonably well with historical trends in Canadian federal elections, where the Liberal and Conservative parties typically capture the largest portions of the popular vote, followed by the NDP. Given that the survey sample was not fully representative across key demographic groups, the post-stratified estimates likely provide a more credible approximation of national voting tendencies compared with the raw survey proportions.

However, these estimates do not incorporate measures of uncertainty, such as confidence intervals or standard errors. As a result, the true variability around these predictions remains unknown. Future work could account for sampling and model-based uncertainty by implementing resampling techniques (e.g., bootstrapping) or using Bayesian hierarchical post-stratification to quantify the range of plausible outcomes.

Overall, the results underscore the importance of demographic adjustment when analyzing survey-based election forecasts. By incorporating census population counts, post-stratification helps ensure that predicted vote shares better reflect the structure of the Canadian electorate, mitigating biases introduced by uneven sample composition.

5 Discussion

This report aimed to predict the national popular vote in the upcoming Canadian federal election using logistic regression combined with post-stratification. Our hypotheses included: (1) older demographics are more inclined to support the Conservative Party, (2) younger demographics are more inclined to support the NDP, and (3) voting tendencies differ among provinces. To investigate these hypotheses, we first modeled voting intention for each major party using survey data and demographic predictors (age, province, gender and education). We subsequently post-stratified the resulting forecasts utilizing census population totals to derive national estimates that represent the demographic makeup of Canada.

The post-stratified results indicated that the predicted national popular vote shares are approximately 30.2% for the Liberal Party, 21.2% for the Conservative Party, and 19.9% for the NDP. These findings suggest a competitive electoral landscape in which no party is projected to win a dominant share of the vote. In broad terms, the results reflect historical voting trends; the Liberal and Conservative parties remain the two major competitors, with the NDP drawing substantial—but comparatively smaller—support nationally.

The regression results offer additional insight into how demographic characteristics relate to reported voting intentions. Many age categories displayed significant positive associations with support for the Liberal Party, whereas support for the NDP declined sharply across older age groups, with several coefficients strongly negative and highly significant. The Conservative model highlighted distinct regional patterns, including notable positive coefficients for provinces such as Saskatchewan and Alberta, and a substantial effect for gender, with men more likely to support the Conservatives. Education, on the other hand, showed no significant effect in any of the models. Together, these patterns illustrate meaningful demographic gradients in political preference, reinforcing the importance of adjusting for population composition when interpreting survey-based measures.

Even with these advantages, the analysis has multiple shortcomings. Initially, the models rely solely on age, province, gender and education as predictors, which implies that significant variables like income, urbanicity, and race/ethnicity are overlooked due to either data availability issues or inconsistencies between the survey and the census. Besides, one limitation is that the "high school or less" category contains very few respondents, which likely reduces the model's ability to detect education effects. Additionally, survey data naturally carry uncertainty, and sampling bias can remain even following post-stratification. Ultimately, we did not calculate standard errors or confidence intervals for the final popular vote estimates, meaning the uncertainty linked to the predictions cannot be formally assessed.

Future extensions of this work could address these limitations by incorporating additional demographic variables, applying multilevel models to account for partial pooling across small population cells, or using bootstrapped and Bayesian post-stratification frameworks to quantify uncertainty. Expanding the geographic granularity of the analysis—from provinces to federal electoral districts—would also allow closer alignment with how federal elections are ultimately decided. Finally, including temporal dynamics, such as month-by-month trends leading up to an election, would shift the analysis from static estimation to genuine forecasting.

In sum, the combined use of logistic regression and post-stratification provides a practical way to adjust survey-derived voting intentions to better reflect the demographic structure of the Canadian population. While the resulting estimates are not free from limitations, the approach succeeds in highlighting systematic demographic patterns in political support and offers a transparent and reproducible foundation for improving electoral prediction techniques.

6 Generative AI Statement

Generative AI tools were used to support the organization of this report, help improve the clarity of explanations, and refine the overall quality of writing. AI assistance was also used in a limited way to enhance grammar and suggest alternative phrasing. All analysis, results, and interpretations were completed independently by the authors. While AI can improve efficiency and communication, it also has limitations. Its responses may contain inaccuracies or lack the critical judgment required for academic work. To ensure reliability, all AI-assisted content was carefully reviewed, cross-checked with course materials, and revised to reflect the authors' own understanding.

7 Ethics Statement

To ensure the reproducibility of our analysis, all data cleaning, modeling, and post-stratification procedures were fully documented within the report and implemented using version-controlled R code. We used consistent statistical methods, clearly defined variable transformations, and provided enough detail for another researcher to replicate the workflow using the same publicly available datasets.

The 2021 Canadian Election Study (CES) and 2021 Census data used in this report are publicly accessible and contain no identifiers that could reveal the identity of individual participants. Because the data are already anonymized and released for public research purposes, the work conducted in this report does not require Research Ethics Board approval. Our analysis uses only aggregated demographic information and does not attempt to re-identify or make inferences about specific individuals. As such, participant privacy and confidentiality remain fully protected throughout the project.

We excluded non-binary respondents from our gender variable so that it would align with the census data, which categorizes gender as only "Man" and "Woman". Since our project relies on population-level census comparisons, we are limited to using the same binary classification. Non-binary respondents represent less than 1% of our sample, and retaining them would force us to misclassify their gender in the analysis. Their exclusion is therefore a methodological decision to maintain consistency with population data and to avoid inappropriate recoding. Excluding them avoids misgendering and ensures that our results are reported in a clear and consistent way with the available population data.

8 Bibliography

1. Stephenson, L. B., Harell, A., Rubenson, D., & Loewen, P. J. (2022). 2021 Canadian Election Study (CES) (Version 3.1, pp. 1615642, 59581151, 361922,

1884147, 225743, 2180672, 2057321, 59190631) [Application/octet-stream,text/tab-separated-values,application/octet-stream,application/pdf,application/pdf,text/tab-separated-values,application/pdf,text/tab-separated-values]. Harvard Dataverse. https://doi.org/10.7910/DVN/XBZHKC

2. Statistics Canada & Statistics Canada. (2023). 2021 Census of Canada, Age, Household type, Gender, Disability, and Marital Status, by Census Tracts [custom tabulation] (Version 1.1, p. 102686142) [Application/x-beyond2020]. Borealis. https://doi.org/10.5683/SP3/UHEPKL

3. Barnfield, M., Phillips, J., Stoeckel, F., Lyons, B., Szewach, P., Thompson, J., Mérola, V., Stöckli, S., & Reifler, J. (2025). The Effects of Forecasts on the Accuracy and Precision of Expectations. Public Opinion Quarterly, 89(1), 185–200. https://doi.org/10.1093/poq/nfaf003

4. Albert, N. (2024, February 11). The First-Past-the-Post Way of Voting is Better-than-the-Rest. C2C Journal. https://c2cjournal.ca/2024/02/the-first-past-the-post-way-of-voting-is-better-than-the-rest/

5. OpenAI (2025, November 20). Proofread election report [Generative AI chat]. ChatGPT. https://chatgpt.com/share/691fba51-3404-800f-9fb4-4d5dc4c78182

6. Kennedy, K., Khanna, K., Simpson, D., Gelman, A. (2020, October 1). Using sex and gender in survey adjustment.