Will the Liberal Party Win the Popular Vote for the Next Election?

STA304 - Fall 2023 -Assignment 2

GROUP 24 : Shuwen Chen, Rena Wang, Heylim Myung, Joanna Joseph

2023-11-07

Introduction

Every four years in Canada there is a federal election where those over age 18 are eligible to vote their party of choice among the current 24 political parties (Walker, 2021). In 2021, those 24 political parties contested in Canada's 44th federal election, with the Conservative, Liberal, and New Democratic parties holding the top 3 percentages in the popular vote (Jeudy, 2021). However, events such as the Covid-19 pandemic had a significant impact on the opinion polls for the different leading parties based on the way they tackled the crises (Walker, 2021). Depending on the events preceding the next election and how the different parties address them, the top parties for popular vote are susceptible to change.

We are interested to see whether the 45th federal election set to occur in 2025 will yield similar results as the past elections, and whether the Liberal party will stay in the top percentages of popular vote. Because the Liberal party was only 1.1% behind the Conservatives in popular vote (Jeudy, 2021), the research question is to predict whether the Liberal party will win the popular vote for the next election. Thus, the null hypothesis of this report is that the Liberal party will hold the popular vote for Canada's upcoming federal election in 2025, while the alternate hypothesis is that the Liberal party will not hold the popular vote for the next election.

Data

The data used in this report is from both census and survey data from Canada's 2021 federal elections.

The survey data is obtained from the 2021 Canadian Election Study (CES) by Stephenson et al (2021), with the survey being conducted online using a rolling-cross section sampling method in 2 waves. The first part of the survey was a Campaign Period Survey (CPS), where 20,968 people from the Canadian population were sampled online and stratified by region while keeping age and gender balanced. The total sample number 20,968 in the CPS was also split into 3 parts: CPS, CPS Modules, and CPS Oversample. The target for the CPS wave was 300 respondents per day, and 1500 respondents per day for CPS Modules but these goals were not always met. The final part of the survey was the Post-Election Survey (PES), with 15,069 of the 20,968 people responding to the follow-up survey after the election, meaning the return rate was 72%.

The census data was obtained from the General Social Survey (GSS): Canadians at Work and Home by Statistics Canada, which is conducted every 5 years. The main goals of the GSS is to gain insights on emerging/current social policy issues and to track the living conditions of the population over time by exploring the relationships between the population's views on work, home, and leisure. The survey was conducted with a cross-sectional design, and the target population is people who are non-institutionalized and above the age of 15. The 10 provinces are divided in strata that have a determined minimum sample size to ensure there is an acceptable level of variability between each strata. A two-stage sampling design was used, where the first stage has clusters selected by telephone number and the final stage has individuals selected from the households with the correlated telephone number.

Data Cleaning

To predict the popular vote for the next election, we used the variables Age, Gender, Education and Region.

Age - Age described the age of the voters at the time of voting.

Gender - Gender described the gender that the voter identified with.

Education - Education was the level of education that the voter had at the time they were voting.

Region - Region described which region in Canada the voter was living in at the time of voting.

To clean the data for both the CES survey data and GSS census data, we removed all missing observations and split age, gender and education into 4 levels. Ages were categorized from 15 to 25, 25 to 50, 50 to 75, and 75 to 100. We included participants aged 15-17 because they will be turning 18 before the next election. So, we think their opinion is also important for our prediction. Gender was split into the 4 levels of Male, Female, Non-binary, and Another gender. Education was split into the 4 levels of "Less than high school diploma or its equivalent", "High school diploma or a high school equivalency certificate", "University or College certificate diploma or degree", and "Don't know/ Prefer not to answer". We started by splitting Age 4 levels so that there is enough to make reasonable inferences but not too much where it is tedious and unmanageable to analyze. Then to match the 4 levels for Age, we also chose to split Education, Gender, and Region into 4 levels as well. There are five regions in Canada: Central Canada, The Prairie Provinces, The West Coast, The North and The Atlantic Region. Since we want to keep it to 4 levels, we are combining the West Coast and the North region as one level. Since British Columbia is the only province in the West Coast and geographically speaking it is close to the North region (consisting of 3 provinces) that also has a relatively small amount of voters compared to the other provinces, we decided to group West Coast and the North together. The census data from GSS has the same 4 levels for age. The code for combining the 4 levels into one bar for the bar plot in Figures 1 through 4 is referenced from "Plotting categorical data with ggplot2" (Siddiqui, 2021), in Appendix A. One difference is that in the census data, gender was split into Male and Female, and from those 2 categories we mutated 0.19% of both Male and Female into Another gender, and 0.14% of Male and Female was mutated into Non-binary based on the article "Sex at birth and gender…" (2021). In order to mutate the Female and Male in the census data into Non-binary and Another Gender, we referenced and modified our code based on Siddiqui's tutorial (Siddiqui, 2021).

Numerical Summaries

The proportions for gender is reflected in the numerical summaries: in table 3 in appendix F, there are 7927 female voters, 6563 male voters, 63 who identified as non-binary and 24 who were Another gender. In table 4, there are 1006 voters aged 15 to 25, 5625 voters aged 25 to 50, 6905 aged 50 to 75 and 1041 aged 75 to 100. In table 5, there are 12336 voters with a University or College certificate diploma or degree, 1871 with a High school diploma or a high school equivalency certificate, 346 with Less than high school diploma or its equivalent, and 24 voters who did not answer. In table 6, there are 9564 voters living in Central Canada, 2591 in the prairie provinces, 1610 in The West Coast and The North and 812 in the Atlantic region.

Summary of Plots

All the figures shown below reflect the survey data only. Figures for the census data are in the Appendices B to E and will be referred to accordingly.

Figure 1: Barplot of Age level response to vote Liberal or not in Survey dat

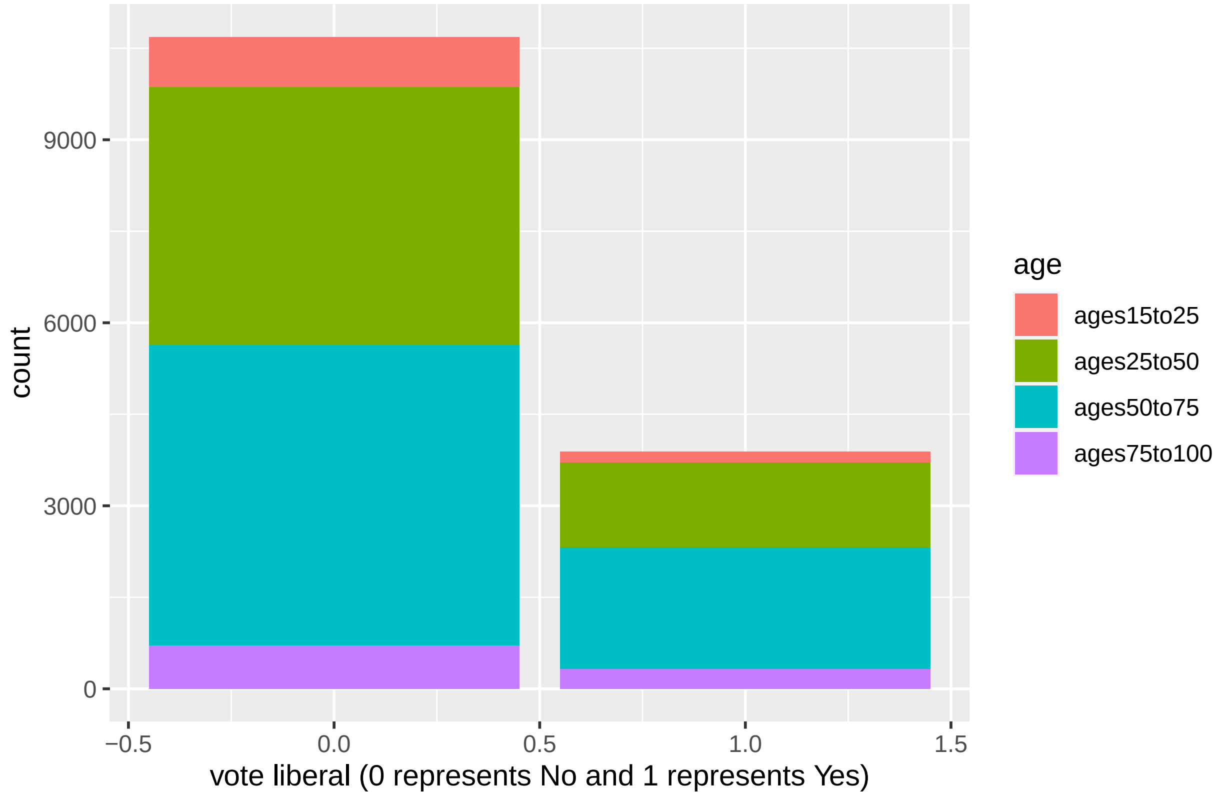
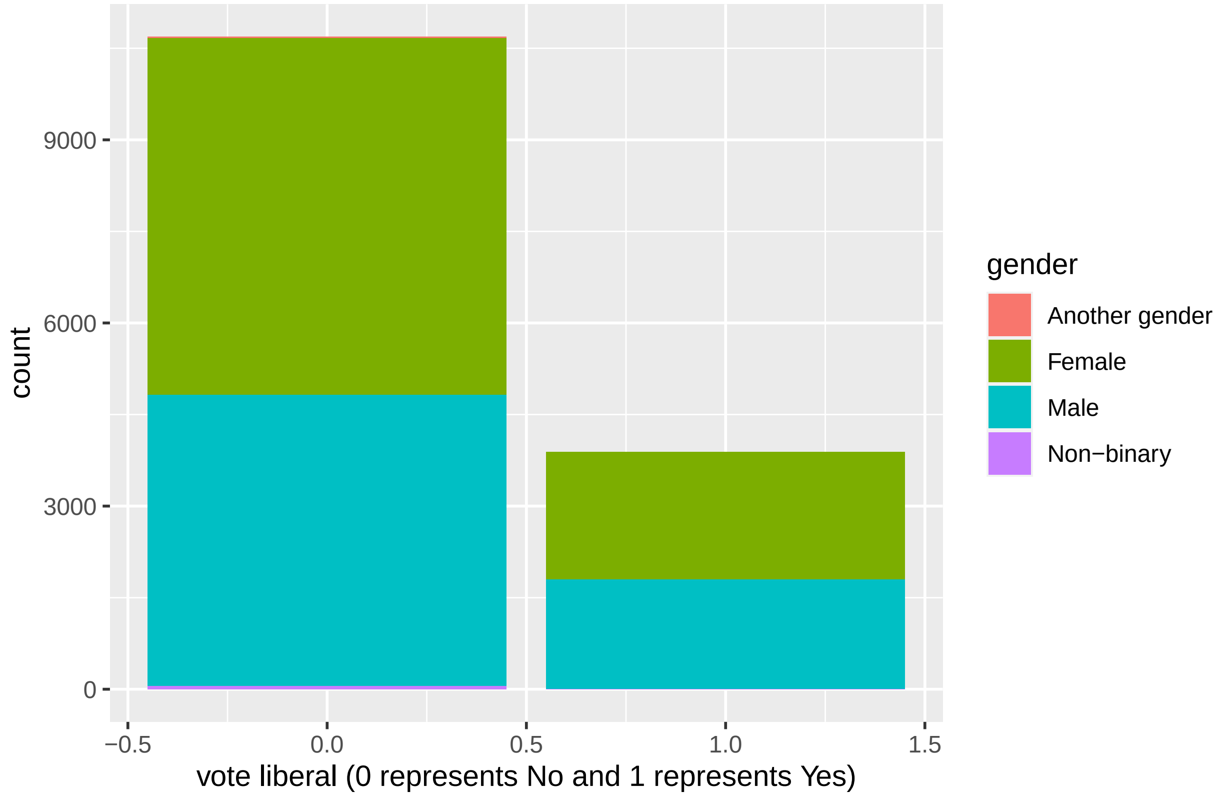


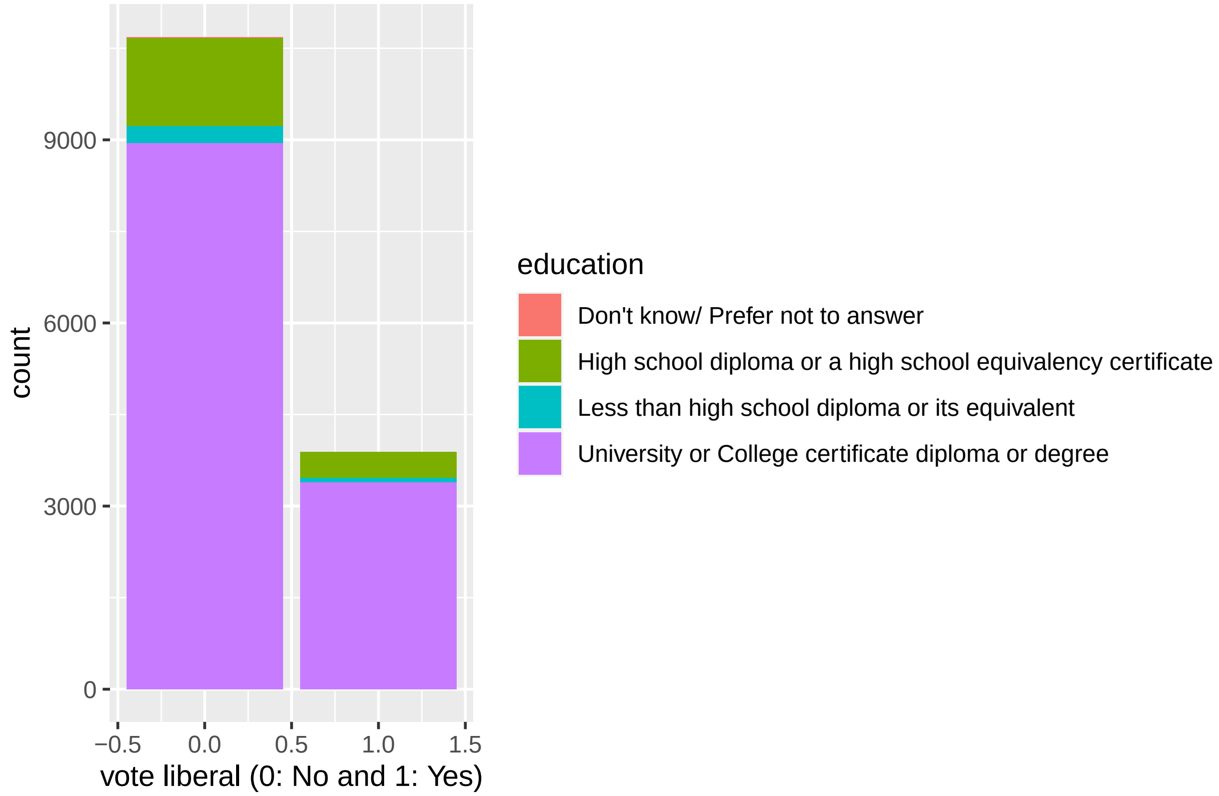
Figure 1 is centered at ages 25 to 50 and ages 50 to 75 since those categories have the most Liberal voters, which means the shape of this data for this is unimodal (see census data in Figure 5 in Appendix B). In Figure 1, the age proportions of the bars representing 'No' and 'Yes' are similar, except more people aged 15 to 25 voted no compared to those aged 75 to 100 who voted no.

Figure 2: Barplot of Gender level response to vote Liberal or not in Survey



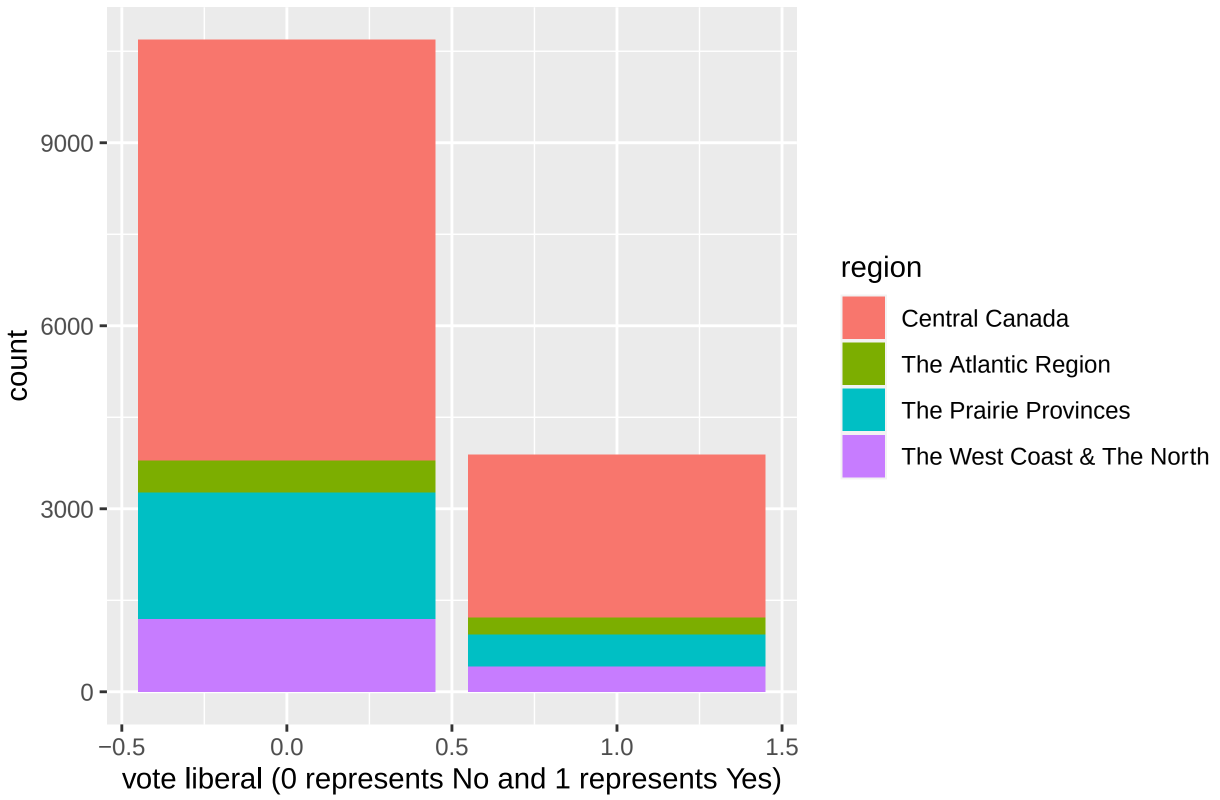
In Figure 2, The proportions between the different levels for gender among those who voted Yes and No are similar. As shown in Figure 6 in Appendix C, Females outnumber Males, and this is reflected in Figure 2 where for both Yes and No votes, Females also outnumber Males. This figure shows the number of liberal voters in the survey data based on Gender is expected, similar to the census data in Figure 6 in Appendix C: Male and Female had the highest Liberal voters since Non-Binary and Another Gender combined together is only 0.33% of both Male and Female voters.

Figure 3: Barplot of response to vote Liberal or not in Survey data based o



In Figure 3, the survey data shows that most voters possess a college certificate, diploma, or degree, and this is reflected in Figure 3 where those who went to college outnumber those who did not and have the highest proportion of voters for both 'Yes' and 'No'. The proportions are similar for levels of education among those who voted 'Yes' and 'No', with there being more than twice the amount of voters for 'No' regardless of their educational background.

Figure 4: Barplot of response to vote Liberal or not in Survey data based o



In Figure 4, most of the voters are from Central Canada as seen in Figure 8 (Appendix E), so that region has the highest number of voters that chose 'Yes' or 'No' for the Liberal party. However, among the voters who voted 'No', the proportion of those in Prairie provinces is relatively higher compared to other provinces than the proportion of voters who voted 'Yes' and are also from the Prairie provinces.

Methods

The objective of the analysis is to predict the overall popular vote of the next Canadian federal election (tentatively 2025) using a regression model with post-stratification.In our research, we aim to investigate the impact of variables such as age, gender, education, and region on an individual's probability of supporting the Liberal Party in the upcoming election.

Model Specifics

We have used multiple regression with post stratification for this purpose. Multiple regression is a statistical method used to examine the relationship between a dependent variable and two or more independent variables. It extends simple linear regression, which involves only one independent variable, to a scenario where multiple factors may influence the outcome.

The reason we chose a multiple regression model is because we assumed it to be the most appropriate choice given that we were dealing with multiple variables and wished to examine the effect of each variable on the outcome (the likelihood that an individual will vote for the Liberal Party) i.e we wish to determine how individuals in each age group (15-25, 25-50, 50-75, 75-100), gender(male, female, non binary, another gender), education level(Less than high school diploma or its equivalent, High school diploma or a high school equivalency certificate, University or College certificate diploma or degree, Don't know/ Prefer not to answer) and region(The Prairie Provinces, The Atlantic Region, Central Canada, The West Coast & The North) will vote in the next elections (particularly with respect to the Liberal Party).

For our multiple regression model, the independent (predictor) variables are age, gender, education and region. The dependent (response) variable is the probability that an individual will vote for the Liberal Party in the upcoming elections. The multiple linear regression model is shown here:

$$y = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{education} + +\beta_4 x_{region} + \epsilon$$

Here y represents the likelihood that an individual will vote for the Liberal Party. β_0 is the intercept, β_1 represents a change in the expected value of the dependent variable (likelihood to vote for Liberal Party) for a one-unit change in the corresponding independent variable (age). β_2 represents a change in the expected value of the dependent variable (likelihood to vote for Liberal Party) for a one-unit change in the corresponding independent variable(gender). β_3 represents a change in the expected value of the dependent variable (likelihood to vote for Liberal Party) for a one-unit change in the corresponding independent variable(education). β_4 represents a change in the expected value of the dependent variable (likelihood to vote for Liberal Party) for a one-unit change in the corresponding independent variable(region). ϵ is the error term which captures the difference between the theoretical value of the model and the actual observed results.

Multiple linear regression models rely on certain assumptions which we have discussed below. Their pertinence to our data is also discussed.

1. Linear relationship: There exists a linear relationship between each predictor variable and the response variable (Zach, 2011). All of our predictor variables : age, gender, education level and region have a linear relationship with the dependent variable, i.e likelihood to vote for Liberal Party.
2. No Multicollinearity: None of the predictor variables are highly correlated with each other (Zach, 2011). There should not be perfect linear relationships among the independent variables. Perfect multicollinearity occurs when one independent variable is a perfect linear function of others, making it challenging to estimate the individual contributions of each variable. Our predictor variables:- age, gender, education level and region are not highly correlated with each other.
3. Independence: The observations are independent from each other (Zach, 2011). A change in one of the variables will not affect any other variable. All of the variables we have chosen(age, gender, education and region) are non dependent on each other i.e are not affected by changes in the remaining variables.
4. Homoscedasticity: The residuals have constant variance at every point in the linear model (Zach, 2011). The variability of the residuals remains consistent across all levels of the independent variables. In other

words, the dispersion of the residuals should stay relatively uniform as the values of the independent variables change.

5. Multivariate Normality: The residuals of the model are normally distributed (Zach, 2011). This assumption holds significance in ensuring precise statistical inferences, especially in the context of hypothesis testing and the construction of confidence intervals.

Post-Stratification

Usually surveys are conducted on a sample of the population(for convenience purposes) and the sample may not always be representative of the entire population. In order to make the results of the survey sample apply to the entire population, researchers take individuals from each strata and multiply them by the total proportion of such people in the entire population. This technique is called post stratification. As can be seen from our data, the sample survey is not representative of the entire population. Thus, we need to undertake post stratification to account for this. Post-stratification estimation is a technique used in sample surveys to improve efficiency of estimators. Survey weights are adjusted to force the estimated numbers of units in each of a set of estimation cells to be equal to known population totals. The resulting weights are then used in forming estimates of means or totals of variables collected in the survey (Richard Valliant, 1992). The main idea is to refine the analysis by accounting for known characteristics of the population to make the estimates more representative.

We have used our multiple linear regression model with post stratification to assign different weights to our independent variables in order to make the survey data representative of the total population. Based on the overall population size and the estimate of each cell, we will post stratify each group and adjust the result (likelihood that an individual in a particular stratum will vote for the Liberal Party). This is demonstrated in the formula:

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

Where $\hat{y}_j$ is the estimate in each cell(probability that an individual from a particular age group, gender, education level and region will vote for the Liberal Party). And N_j is the population size of the j^{th} cell based off demographics.

Thus, we took the population cell count(given by the census data) and grouped it by the four variables age, gender, education and region. We then used the multiple regression model to estimate what the probability of an individual in each age group, gender, education level and region to vote for the Liberal Party was. These estimates were then post stratified by multiplying them by the actual proportion of such individuals in the population.

All analysis for this report was programmed using `R version 4.0.2`.

Results

Table 1 gives us a summary of the results of the multiple regression model we have used. For each of our variable sub groups- ages25to50, ages50to75, ages75to100 ; male, female, non binary; high school diploma or a high school equivalency certificate, less than high school diploma or its equivalent, university or college certificate diploma or degree; the Atlantic Region, the Prairie Provinces, the West Coast & The North we have run the regression model and found the estimate, standard error (difference between actual and observed mean), t statistic and p value.

As can be seen from table 1, for each unit increase in the variable "ageages25to50," the estimated value of the dependent variable increases by 0.0626671. The p-value is very small (0.0000353), suggesting that this effect is statistically significant. Similarly, for each unit increase in the variable "ageages50to75," and "ageages75to100," the estimated value of the dependent variable increases by 0.1064708 and 0.1369239 respectively. The p-values are also very small, suggesting that this effect is statistically significant. This means that there is strong statistical significance that individuals in these age groups have a lower probability of voting for the Liberal Party.

Table 1 also shows us that being female is associated with a 0.0901699 increase in the dependent variable, but the result is not statistically significant as the p-value is 0.3159507. Results for males and non binary individuals are also not statistically significant. This means that the particular gender of the individual does not play a significant role in determining their probability in voting for the Liberal Party.

Table 1 also shows us that having a high school diploma or equivalent is associated with a 0.1288722 increase in the dependent variable, but the result is not statistically significant. Similar interpretation for less than high school diploma or equivalent. For individuals with university or college certificate, diploma, or degree, there is a statistically significant 0.1820816 increase in the dependent variable. This means that individuals with a higher education level have a lower probability of voting for the Liberal Party.

Table 1 also shows us that being in the Atlantic Region is associated with a 0.0658658 increase in the dependent variable, and the result is statistically significant. Being in the Prairie Provinces is associated with a 0.0783731 decrease in the dependent variable, and the result is highly statistically significant. Being in the West Coast & The North is associated with a 0.0246187 decrease in the dependent variable, and the result is statistically significant. This means that the geographical location of an individual does play a role in determining their likelihood to vote for the Liberal Party. Low p values also mean that individuals have a lower probability to vote for the Liberal Party (per our hypothesis).

Thus, overall it seems like we are in a position to reject the null hypothesis (the Liberal Party will hold the popular vote in the upcoming federal elections in Canada). These results do seem reasonable as the significant regional differences align with the diverse political landscapes in Canada. Voting patterns can vary substantially across provinces and territories due to historical, cultural, and economic factors. The Prairie Provinces have historically not been big supporters of the Liberal Party as is evidenced by the 2021 election results. Our regression model too shows very little support for the Liberal Party from the Prairie Provinces in keeping with historical records. Also, more educated people too are now withdrawing their support for the Liberal Party (Dyk, S.V. , 2023). Our regression model supports this theory as there is a lower probability for highly educated people to vote for the Liberal Party.

After conducting post stratification on our multiple regression model, we have determined the probability of the Liberal Party winning the next federal election in Canada. As seen in Table 2, according to our predictions, the Liberal Party has a 20.88 percent chance of winning the next federal elections. This goes against our null hypothesis and we are able to predict that the probability of the Liberal Party winning the next federal election is very small.

Conclusions

To summarize, this report hypothesized that Liberals would win the popular vote based on past election data. However, this hypothesis is not supported based on the prediction for the number of Liberal voters among the population, which the results showed was 20.88%. This is lower than expected considering that Liberals have been among the top parties for popular vote in the past, winning 32.6% of the votes in the last election which is only 1.6% behind the Conservatives who won the popular vote with 33.7% (Jeudy, 2021). Although there are some statistically significant positive associations between all levels for the Age group and whether or not they vote Liberal, the p-values were insignificant for all levels in Gender and Education except for individuals with university or college certificate, diploma, or degree who are significantly positively associated with voting Liberal. The results show a significant negative association for levels in Region such as the Prairies and The West and The North except for Atlantic, which has a significantly positive association. Despite the significantly positive associations within some levels for all variables, most levels show either insignificant association or a significant negative association. This leads us to reject the null hypothesis and accept the alternate hypothesis. Although this percentage doesn't reflect what we predicted, it still gives useful insights as to how the next federal election will turn out by providing information on which levels within age, education, gender and region.

One limitation is that the predictors (such as gender) are only split into 4 levels. In the Gender variable the level “Another gender” groups several other genders into one category, which means there may be some prediction error since it is not a completely accurate representation of the population. Furthermore, the article we used to obtain the percentages for Another gender and Non-binary is from census data in 2021. Because this report aims to predict the 2025 election results, the data from 2021 may not reflect the true gender proportions among the voters in the next election.

One confounding variable that could also affect the outcome of whether someone votes Liberal or not is the number of events preceding or during the election. Similar to the Covid-19 pandemic that required the parties to tackle the crisis in different ways, another event preceding the election could change how many popular votes they get based on the way they handle the situation.

For future studies of this nature, it would be ideal to collect survey and census data closer to the next election date, preferably a few months before or during the next election. That way, the prediction for certain variables such as Gender with levels for ‘Another gender’ or ‘Non-binary’ would be more accurate, since the number of voters in those categories change every year. Furthermore, by collecting data closer to the election date we would also include impactful events (such as another pandemic) that would affect the public’s opinion on a certain party based on how they deal with the crisis.

Bibliography

1. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: April 4, 1991)

2. Dyk, S. V. (2023, August 30). “getting squeezed on both sides”: Liberals a distant third among younger voters. CTVNews. https://www.ctvnews.ca/politics/getting-squeezed-on-both-sides-liberals-a-distant-third-among-younger-voters-1.6539949#:~:text=Those%20who%20are%20feeling%20the,moving%20toward%20the%20N

3. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: April 4, 1991)

4. Heiss, A. (2022, June 23). Quick and easy ways to deal with long labels in ggplot2. Retrieved from Andrew Heiss: https://www.andrewheiss.com/blog/2022/06/23/long-labels-ggplot/

5. Introduction to post-stratification - world food programme. (2017). Retrieved from World Food Programme: https://docs.wfp.org/api/documents/WFP-0000121326/download/

6. Jeudy, L. (2021, December 6). Federal election results in Canada 2021. Retrieved from Statista: https://www.statista.com/statistics/1066361/federal-election-results-canada/#statisticContainer

7. OpenAI. (2023). *ChatGPT (September 13 version) [Large language model].* https://chat.openai.com/chat (Last Accessed: September 13, 2023)

8. RStudio Team. (2020). *RStudio: Integrated Development for R.* RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

9. Siddiqui, N. (2021, November 10). How to randomly replace values in an R data frame column? Retrieved from tutorialspoint: https://www.tutorialspoint.com/how-to-randomly-replace-values-in-an-r-data-frame-columnhttps://www.tutorialspoint.com/how-to-randomly-replace-values-in-an-r-data-frame-column

10. Statistics Canada. (2016, July 12). General Social Survey: Canadians at Work and Home (GSS). Retrieved from Statistics Canada: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5221

11. Statistics Canada. (2022, April 27). Sex at birth and gender – 2021 Census promotional material. Retrieved from Statistics Canada: https://www.statcan.gc.ca/en/census/census-engagement/community-supporter/sex-birth-gender

12. Stephenson, L., Harell, A., Rubenson, D., Loewen, P. (2022). 2021 Canadian Election Study (CES). Retrieved from Consortium on Electoral Democracy, Harvard Dataverse, V3: https://doi.org/10.7910/DVN/XBZHKC

13. Valliant, R. (1993). Poststratification and conditional variance estimation. Journal of the American Statistical Association, 88(421), 89. https://doi.org/10.2307/2290701

14. Walker, N. (2021, September 29). Canada: 2021 general election. Retrieved from House of Commons Library: https://researchbriefings.files.parliament.uk/documents/CBP-9323/CBP-9323.pdf

15. Zach. (2021, November 16). The five assumptions of multiple linear regression. Retrieved from Statology: https://www.statology.org/multiple-linear-regression-assumptions/

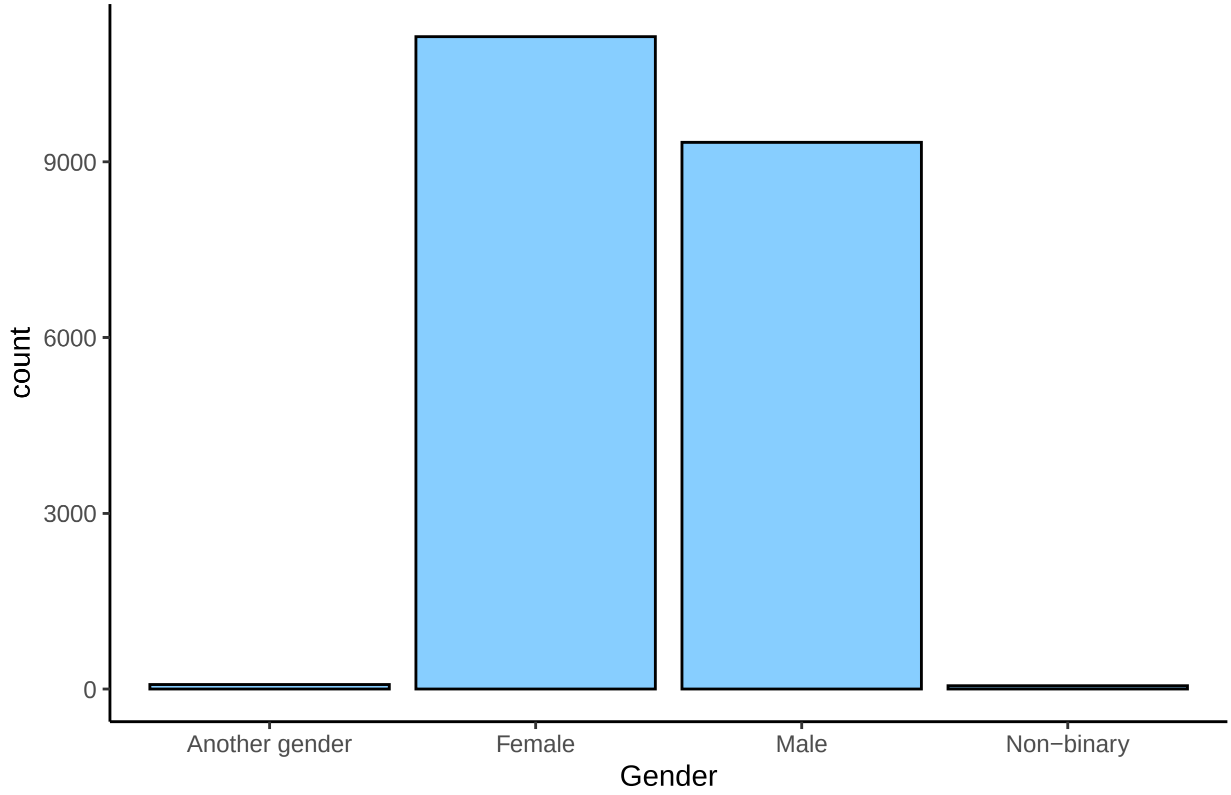
Appendix

Appendix A:Reference for the code that we changed and mutated based on Siddiqui's tutorial(Siddiqui, 2021).

```
census_data$gender[sample(nrow(census_data), 0.0019 * nrow(census_data))] = "Another gender"
census_data$gender[sample(nrow(census_data), 0.0014 * nrow(census_data))] = "Non-binary"
census_data$education[is.na(census_data$education)] = "Don't know/ Prefer not to answer"
```

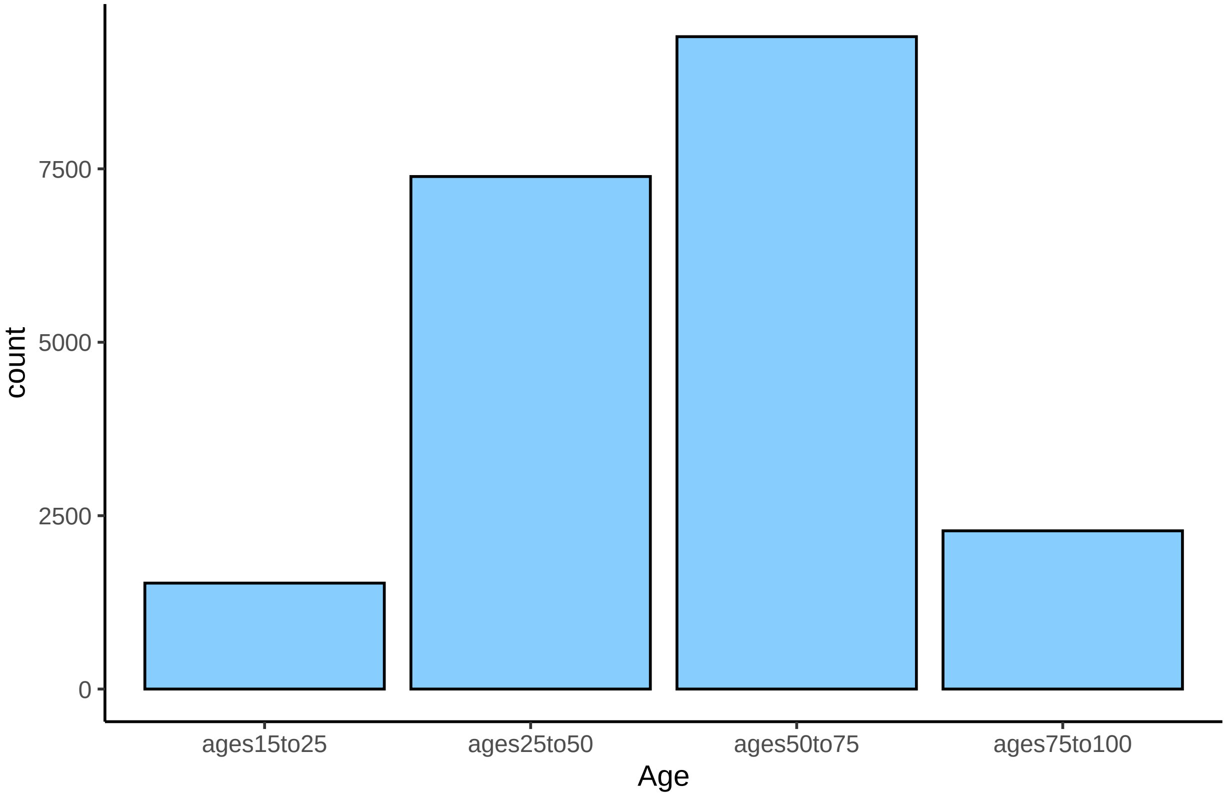
Appendix B:

Figure 5: Barplot of Gender counts in Census data



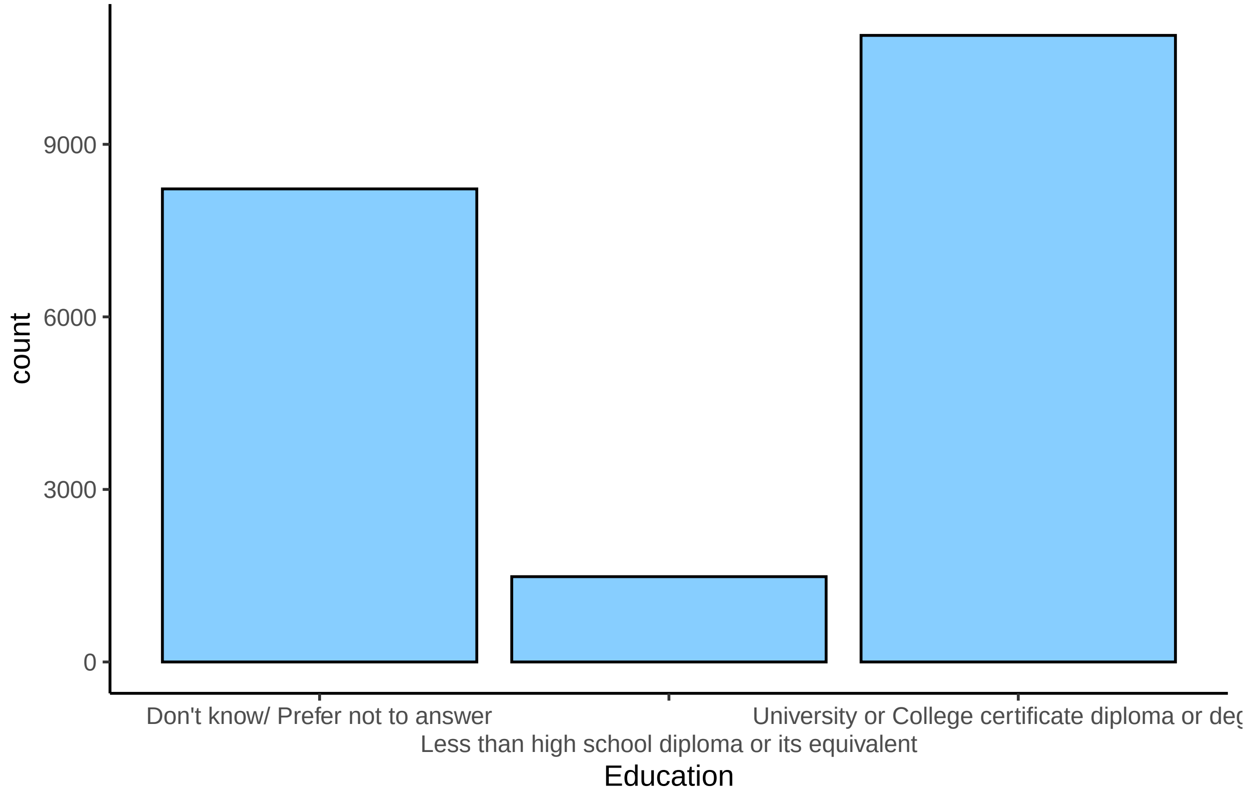
Appendix C:

Figure 6: Barplot of age counts in Census data



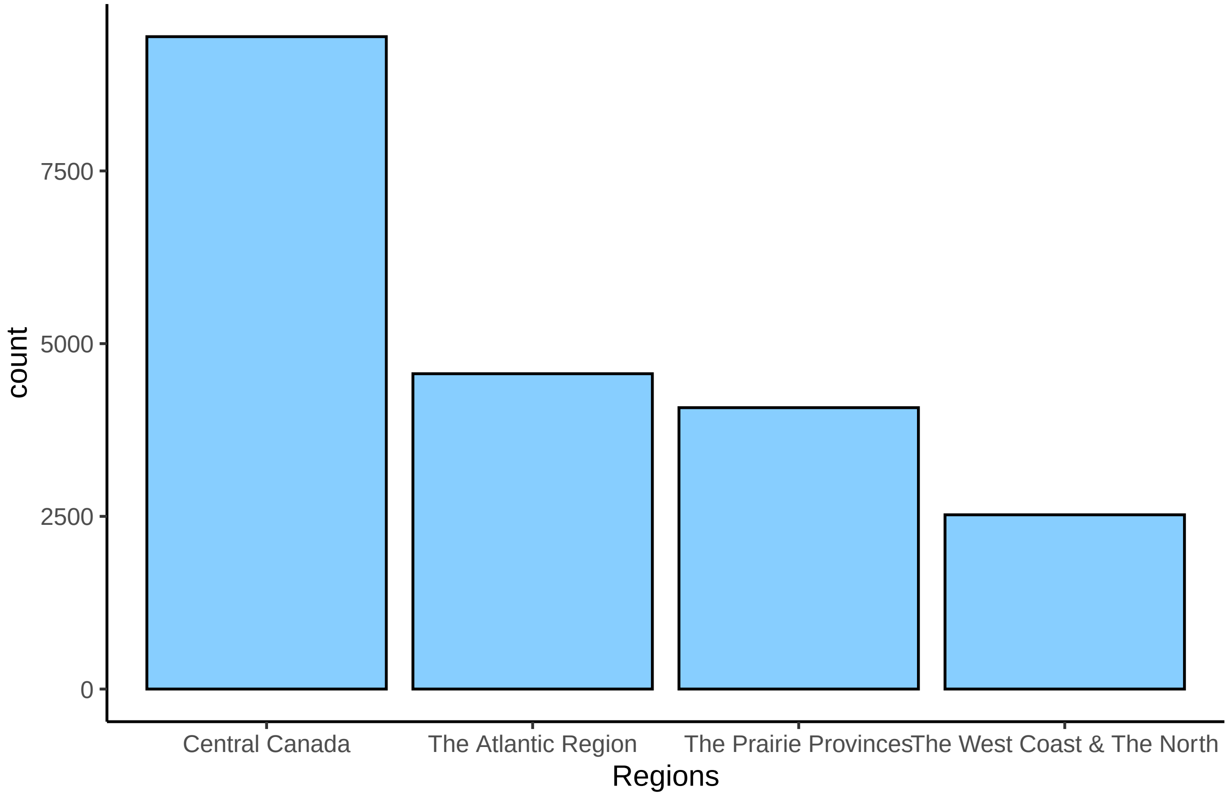
Appendix D:

Figure 7: Barplot of education counts in Census data



Appendix E:

Figure 8: Barplot of region counts in Census data



Appendix F:

Table 1: Gender summary

gender	num_lines
Another gender	24
Female	7927
Male	6563
Non-binary	63

Table 2: Age summary

age	num_lines
ages15to25	1006
ages25to50	5625
ages50to75	6905
ages75to100	1041

Table 3: Education summary

education	num_lines
Don't know/ Prefer not to answer	24
High school diploma or a high school equivalency certificate	1871
Less than high school diploma or its equivalent	346
University or College certificate diploma or degree	12336

Table 4: Region summary

region	num_lines
Central Canada	9564
The Atlantic Region	812
The Prairie Provinces	2591
The West Coast & The North	1610

Generative AI Statement

We have not used any generative AI for the purpose of this assignment.