Predicted outcome in the Canadian Federal Election 2025

STA304 - Fall 2023 -Assignment 2

Group 120: Ruilin Peng, Johnson Guo, Tony Chen, Yuchen Li

11/7/2023

1. Introduction

1.1 The research question

What will be the predicted outcome for the party that receives the highest-amount votes in the upcoming election 2025 between Conservative Party and Liberal Party?

1.2 Importance of the research

Free and fair elections are considered as important features contributing to a healthy democracy. Fairness of the election could be ensured as citizens can vote to reflect their own wills without political interference. These conditions can contribute to a durable democracy in Canada and also improve internal political efficacy of Canada (*Elections Canada*, 2023). The choice of who will form the government and which party's policies will guide the government and legislation. These appointments and policies can influence various aspects of society, including the economy, public health initiatives, and environmental sustainability. The analysis of election outcome helps encourage citizens to track the political process and promote their participation (*Statistics Canada*, 2022). Political parties and candidates can learn about competitors' voting tendency and priorities and concerns of voters so adjusting their electoral strategies.

Importantly, our analysis provides a demonstration of the relationship between the selected variables and the binary outcomes of voting decisions (vote or not vote) to see how the change of variables influences the change of probability change in vote. In this case, the result of the analysis can be an essential reference for future study of Canadian elections, and representatives of parties may set up relevant rules and benefits to increase their political reputation and construct campaign strategies according to the favorable factors considered by respondents. Furthermore, the government may refine relevant policy in terms of the races, employment, income, and other socioeconomic data presented in this study.

1.3 Terminology

The 45th Canadian federal election will be held on or before October 20, 2025. This date is determined by the Canada Elections Act (Parliament of Canada, 2006). There are currently 337 members in office in the 44th Parliament. Both Liberal Party and Conservative Party have 275 members (158 and 117 respectively), and they are the two major Canadian parties(Members of Parliament, 2023).

The Liberal Party is the oldest active federal political party in Canada and has dominated federal politics for the majority of Canadian history (Liberal Party, 2019). The party espouses liberal principles and is generally central to center-left. Its main competitor, the Conservative Party, is right-leaned. Liberal Party is currently led by Justin Trudeau who has been the prime minister of Canada since 2015. Signature policies and legislation of the Liberal Party include universal health care, pension plan, bilingualism, gun control, related charters and acts that legalize same-sex marriage and cannabis usage, and expanding access to abortion (Liberal Party of Canada. n.d.).

The Conservative Party is one of the major political parties in Canada, known for its center-right political orientation. Its platform typically includes a commitment to fiscal responsibility, free-market principles, and a focus on individual liberties. The party has traditionally attracted support from a broad coalition of conservative voters, including those with social or economic conservative values. The party is currently led by Pierre Poilievre since 2022. Its signature policies involve reducing government debt, lowering taxes, eliminating the long-gun registry (Dippel, 2016).

The Liberal party leans on progressive policies and social transformation, and the Liberal encourages free-market competition in economics, diversity of minorities, and inclusion of immigrants. As opposed to the Liberal, the Conservative prefers traditional and preservative rules, imposes more government regulation on economics, and disagrees with social activism (Canada Guide, 2023).

Canada includes five distinct regions: The Atlantic Provinces (New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island), Central Canada (Ontario and Quebec), The Prairie Provinces (Manitoba, Saskatchewan and Alberta), The West Coast (British Columbia), and The Northern Territories (Nunavut, Northwest Territories, and Yukon Territories) (Government of Canada, 2022). In order to more easily present the data, adjustments have been applied to the categories of five regions. The Prairie Provinces have been renamed into western regions.

1.4 Hypothesis

As the 45th Canadian Election is approaching, the competition between "Liberal" and "Conservative" is kicked off again. The Liberal and the Conservative parties dominate the political system of Canada and the show plays between these parties. Moreover, if we look at the recent 20 years of popular vote percentages, the Liberal Party almost beats the Conservative Party for each election with an outstanding difference (Simon Fraser University, 2021). Therefore, we hypothesize that the Liberal party is still going to win the election and it is exciting to see if there is any chance that the Conservative will probably present a surprise.

The goal of this study is to predict which political party of Canada is going to win the most popular votes in the 2025 Canadian Election based on the information of the previous 2021 Canadian Election. This study matches the survey data collected from the 2021 CES (Canadian Election Study) online survey with the census data collected from the Canada GSS (General Social Survey) and predicts the voting decision of respondents in the next 2025 Election in a logistic regression model based on demographic variables of respondents which include age, sex (male/female), language (English/French), income class (Low, Medium and High), the religious or atheist, and the races (the white or the other minorities). (Stephenson et al., 2021; Canada Statistics, 2020). To enhance the representatives of the population, we apply the stratification method that divides the survey respondents into strata by different variables and conclude the estimated voting probability for both the Liberal and the Conservative to infer the possible outcomes in the 2025 Election.

2. Data

2.1 Data description

<Type here a paragraph introducing the data, its context and as much info about the data collection process that you know.>

Survey Data The sample survey data in this study was collected during the 2021 federal election campaign period and intended to gather the opinions of Canadian citizens and permanent residents aged 18 or older about election votes.

The survey consists of the campaign period part and the post-election period part. The collection process started on August 7th, 2021, and ended on October 4th, 2021. To demonstrate, the campaign period survey data was collected on an online sample of 20,968 respondents of the Canadian general population at the Leger Opinion panel. The sampling process included three waves of panel in a modified rolling-cross section. In this case, the survey data of three waves were merged into a whole sample at the end, but the respondents of the survey were randomly selected in three different periods of time. (Stephenson et al., 2021).

Census Data The census data in this study is the 2020 General Social Survey, which provides detailed information of social trends for addressing interested social issues and improving living-conditions of Canadians, and it is released by September 2021. In addition, data was collected by both computer-based phone calls and electronic questionnaires through mail links from Statistics Canada offices in different regions from August of 2020 to February of 2021. According to the descriptive statistics, the overall response rate was 40.3%, and a total of 20,602 observations were collected. Moreover, the stratified sampling method was applied where each response was assigned to a province level (a stratum). Then, a simple random sampling (SRS) was conducted in each stratum (Statistics Canada, 2020).

2.2 Data cleaning process

<Type here a summary of the cleaning process (**only add in stuff beyond my original gss_cleaning.R code**). You only need to describe additional cleaning that you and your group did.>] You will need to describe the cleaning you do to the survey data as well.

2.3 Data summary

<Remember, you may want to use multiple datasets here, if you do end up using multiple data sets, or merging the data, be sure to describe this in the cleaning process and be sure to discuss important aspects of all the data that you used.>

<Include a description of the important variables.>

The variables described in this section applied to both the CES survey data and GSS census data because we matched the variables in both datasets in the Data Cleaning section.

Response Variables vote of conservative: The binary variable that whether the voter will vote for the Conservative Party in the 2025th election.

vote of liberal: The binary variable that whether the voter will vote for the Liberal Party in the 2025th election.

Independent Variables age: This variable represents the age of the observations. It is a numerical variable.

sex: This variable represents the sex of the observations

language: A categorical variable indicating the primary language of individuals, with categories “FR” for French and “EN” for English.

west: A binary variable indicating whether an individual is from Western provinces (Manitoba, Saskatchewan, Alberta) or not.

income: A categorical variable categorizing individuals into “Low”, “Medium”, and “High” income groups based on their income category.

religion: A binary variable indicating whether individuals have a religious affiliation or not.

minority: A binary variable indicating whether individuals are part of a visible minority or not.

Table 1: Statistics about the proportions of Yes votes for the Liberal and the Conservative in Survey data

prop.Liberal	prop.Conservative
0.2037	0.1975

Table 1 shows that

Table 2: Statistics about the ages of voters in 2021 election survey data

min	Q1	median	Q3	max	IQR	mean	sd
18	35	53	65	92	30	50.58	18

Table 3: Statistics about the ages of voters in 2021 election census data

min	Q1	median	Q3	max	IQR	mean	sd
18	39	56	68	80	29	53.59	17.21

Table2 & 3 show that

Table 4: Statistics about the frequency and proportion of male and female voters in 2021 election survey and cenus data

sex	num.surveySex	prop.surveySex
Female	554	0.68
Male	261	0.32

sex	num.censusSex	prop.censusSex
Female	9250	0.55
Male	7680	0.45

Table 4 shows

Table 5: Statistics about the frequency and proportion of French and English speaking voters in 2021 election survey and cenus data

language	num.surveyLanguage	prop.surveyLanguage
EN	622	0.76
FR	193	0.24

language	num.censusLanguage	prop.censusLanguage
EN	13541	0.8
FR	3389	0.2

Table 5 shows

Var1	Freq
0	674
1	141

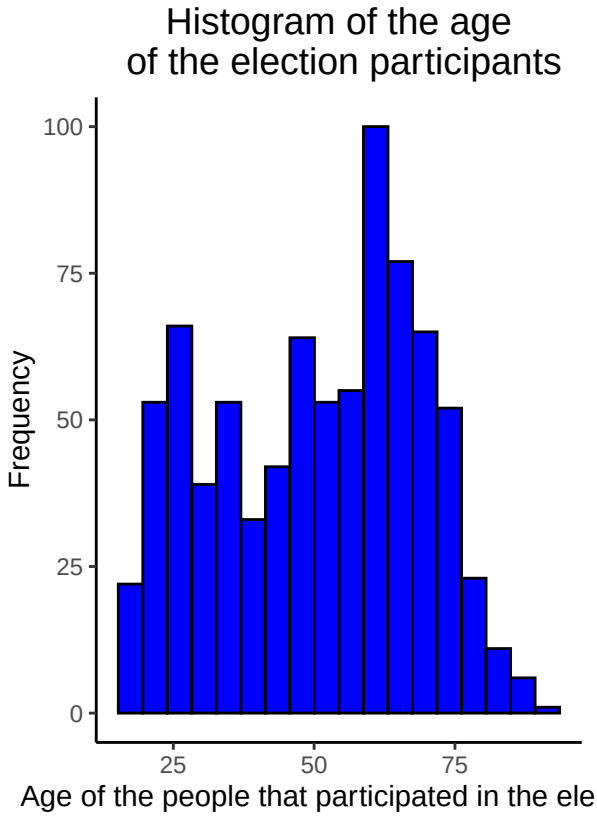
Var1	Freq
High	603
Low	120
Medium	92

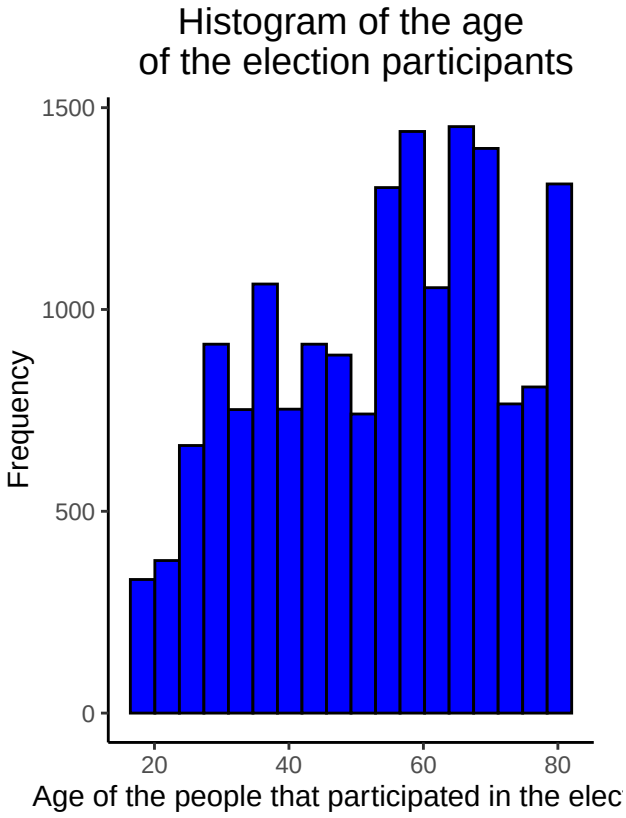
Var1	Freq
High	603
Low	120
Medium	92

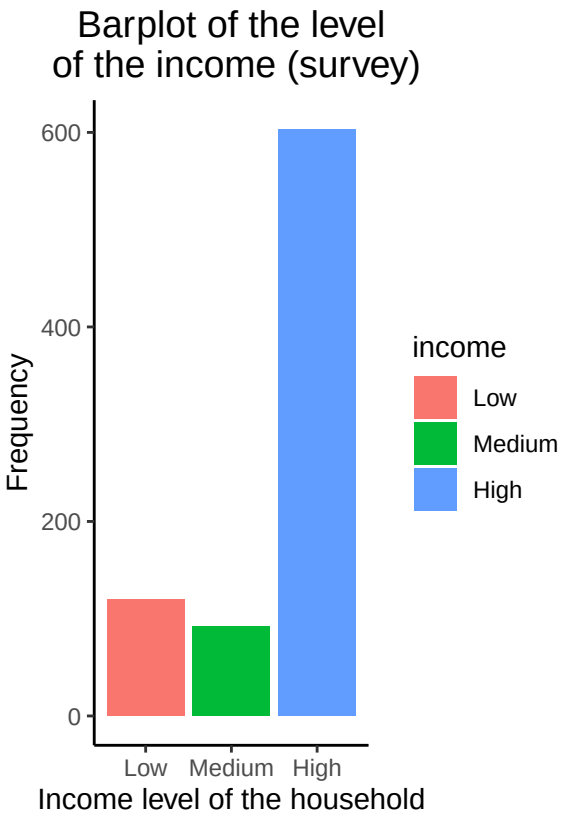
Var1	Freq
High	603
Low	120
Medium	92

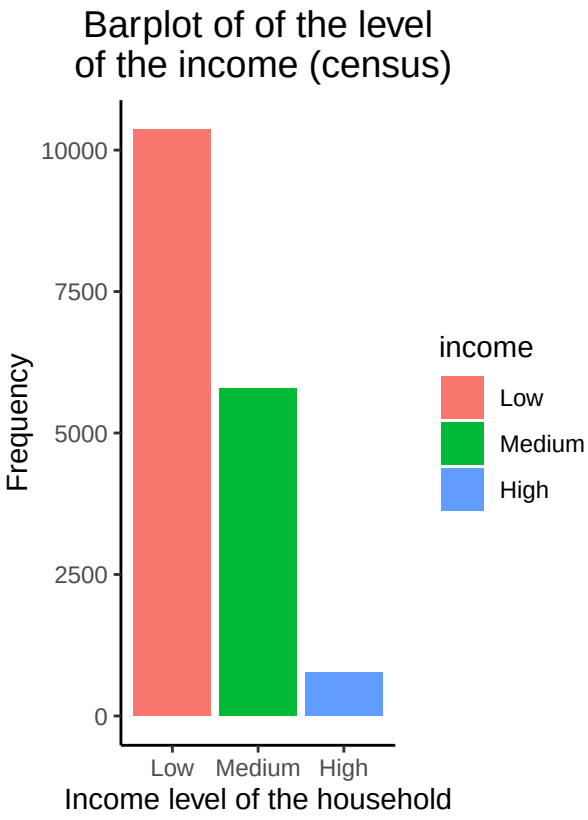
Text

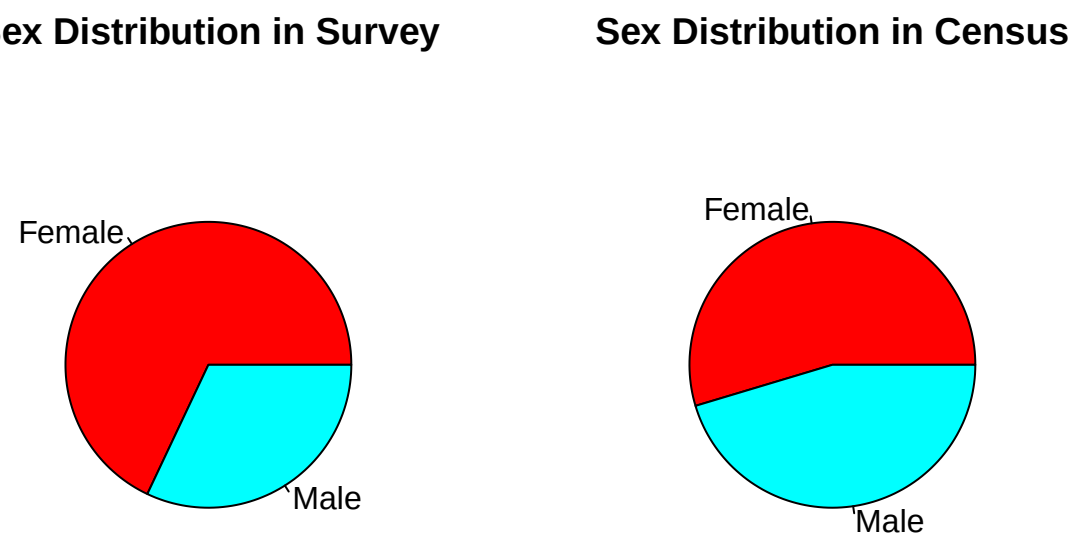
<Include a description of the numerical summaries. Remember you can use `r` to use inline R code.>






Barplot of the level
of the income (survey)
600
400
200
0
Frequency
Low
Medium
High
Income level of the household
income
Low
Medium
High


Barplot of of the level
of the income (census)
10000
7500
5000
2500
0
Frequency
Low
Medium
High
Income level of the household
income
Low
Medium
High



3. Methods

3.1 Model selection and rationale

Our study applies the logistic regression model to estimate the probability of voting of each voter for the Liberal Party and Conservative Party. Then, we use the post-stratification which divides the survey data in terms of the variables common in the census data and the survey, and yields the estimated overall probability of voting for the Liberal Party and Conservative Party respectively. In the end, we compare the estimated voting probability between the two parties to decide which one is going to win the election (i.e. the party has the higher estimated probability). In our model, to ensure the qualification of voting, we only keep the respondents who are at least 18 years old. In order to remove the influence of missing data, we delete the observations (record of respondents) which contain missing values of variables.

3.2 Model Specifics

Our study creates two logistic regression models to predict the proportion of voters who will vote for the Liberal Party and the Conservative Party. Some assumptions are made before constructing our logistic regression models: The voting outcome of respondents is binary, that is, the respondents either vote for the party or not, and the selected categorical variables are independent in both the logistic models built for the Liberal and the Conservative. We select the following models because the variables are matched in both the census and the survey data, and the variability of the variables is similar in survey data compared to the census data as shown in the plots of Data Summary section (...). In addition, a previous study by Statistics Canada proposes that economic status, belief, and personal inborn identities such as sex, language, and race are the dominant factors that potentially influence the intention of voting, so we select the following variables as predictors to investigate the probability of each voter to vote (Uppal, S.& LaRochelle-Côté S., 2012). In the logistic regression models for both parties, we use the same variables as predictors.

$$log(\frac{p_c}{1-p_c}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{French} + \beta_4 x_{west} + \beta_5 x_{incomeMedium} + \beta_6 x_{incomeHigh} + \beta_7 x_{religion} + \beta_8 x_{minority}$$

$$log(\frac{p_l}{1-p_l}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{French} + \beta_4 x_{west} + \beta_5 x_{incomeMedium} + \beta_6 x_{incomeHigh} + \beta 6 x_{religion} + \beta_7 x_{minority}$$

p_c represents the probability of voting for the Conservative Party and p_l represents the probability of voting for the Liberal Party. The $log(\frac{p_l}{1-p_l})$ and $log(\frac{p_c}{1-p_c})$ are the log odds in both models.

β_0 = is the intercept of the model. It represents the log odds of voting for the candidate or party when all the predictor variables are at their baseline level (which means that the respondent is aged at 0 years, sex category female, speaking the language of English, not living in Quebec, the default income class at low level of income, does not believe in religions, and is a minority other than the White.).

β_1 represents the relationship between the age of the respondent and the log odds of voting for the parties. For every one-unit increase in age (typically one year), there is an expected β_1 change in the log odds of voting for the parties, assuming all other variables are held constant.

β_2 quantifies the average difference in the log odds of voting for the parties between two sex groups (e.g., male and female), controlling for other factors in the model. It shows how sex influences the likelihood of voting for the candidate or party.

For β_3, this coefficient measures the average difference in log odds of voting for the parties between respondents who speak French and English when other predictors stay the same. It reflects how changes in language proficiency between English and French can affect the voting behavior of a respondent, with other factors being equal.

β_4 represents the average difference of being in the western inland provinces (Alberta, Manitoba, Saskatchewan) or not on the log odds of voting for the candidate or party while holding other predictors constant. This could reflect regional differences in voting preferences, with all other variables held constant.

Forβ_5, This coefficient measures the average difference of log odds of voting between the respondent who possesses a low and medium income level. Similarly, β_6 measures the average difference of log odds of voting between the respondent who possesses a low and high income level when other predictors stay the same. They indicate how changes in income level impact the log odds of voting for the candidate or party. Both coefficients capture the relationship between economic status and voting behavior, controlling for other variables in the model.

β_7 assesses the average difference in the log odds of voting for the parties between the respondents who are atheists or possess religious beliefs with the other predictors fixed. This factor might capture how the difference between a religious believer and a non-religious person on voting preferences.

Finally, β_8 measures the average difference in the log odds of voting for the parties if a respondent is identified as a minority member compared to a non-minority one. This could reflect how membership in certain demographics or ethnic minorities influences voting patterns compared to the majority component of White people in the population of Canadian citizens (Statistics Canada, 2022).

3.3 Post-Stratification

Post-stratification is a method to ensure that the results can accurately represent different groups within a population. It involves adjusting the weights for each estimated parameter within specific post strata based on their corresponding weights in the census population size.

In order to estimate the proportion of voters who will vote for … we will perform a post-stratification analysis by applying the following formula:

Table 6: Table x.Summary of Logistic Regression Model for Liberal

	Estimates	SE	Test statistics	p-value
(Intercept)	-2.2672	0.4200	-5.3984	0.0000
age	0.0100	0.0054	1.8373	0.0662
sexMale	-0.2274	0.1991	-1.1424	0.2533
languageFR	-0.4846	0.2255	-2.1491	0.0316
west	-0.8669	0.2812	-3.0822	0.0021
incomeMedium	0.8107	0.3806	2.1303	0.0331
incomeHigh	0.6433	0.3050	2.1094	0.0349
religion	0.1523	0.2168	0.7027	0.4822
minority	-0.0358	0.2031	-0.1765	0.8599

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

$\hat{y}$ is the estimate in each cell, and N_j is the population size of the j^{th} cell based on demographics.
The estimated ($\hat{y}$) is $\hat{p}$ that refer to the proportion of voting for… Post-stratification will conduct logistic regression in each cell and use the logistic model to estimate the $\hat{p}$ within each cell. We firstly will create cells based on different ages, sex, and working status. Using the model described in the previous sub-section, we will estimate the proportion of voters in each cell. _Since there are 2 categories in sex, ? in region, ? in education, and ? in religion, we will have ? cells for each age.
We will subsequently weight each proportion estimate within each cell by the respective population size of that cell and will then sum those values and divide that by the entire population size.

4. Results

4.1 Result of Logistic Regression Models

The following are the logistic regression model results (coefficients rounded to 3 decimals). For Liberal

$$log(\frac{\hat{p_l}}{1-\hat{p_l}}) = -2.267 + 0.010x_{age} - 0.227x_{sex} - 0.485x_{French} - 0.867x_{west} + 0.811x_{incomeMedium}$$

$$+0.643x_{incomeHigh} + 0.152x_{religion} - 0.036x_{minority}$$

According to the liberal model For Conservative

$$log(\frac{\hat{p_c}}{1-\hat{p_c}}) = -3.358 + 0.017x_{age} + 0.327x_{sex} - 0.637x_{French} + 0.799x_{west} + 0.646x_{incomeMedium}$$

$$+0.356x_{incomeHigh} + 0.734x_{religion} - 0.180x_{minority}$$

where $\hat{p_c}$ and $\hat{p_l}$ are the estimated probability of voting to the Conservative and Liberal Party.

Tablex shows the summary statistics of the logistic regression for the Conservative, which includes the estimated coefficients (slopes) of the logistic regression models, the corresponding standard errors, test statistics and p-values for the coefficients t-tests. Using a significance level of 0.05, we can see that p-values for the coefficient estimate of numerical variable age and dummy variables: sex of Male or not, Language of French or not, living in Western inland provinces (Alberta, Manitoba, Saskatchewan) or not, income class at Medium or not, income class at High or not, possess a religion or not, identified as a minority or not. Among these predictors, we can see that the coefficients of age, language speaking and religion identification can form a significant relationship with the voting decision of voters (whether or not they will vote the Conservative and the Liberal) because their p-values are smaller than the 0.05 significant level cutoff. Similarly, **Tablex** shows the summary statistics of the logistic regression for the Liberal, and we can see that the coefficients

Table 7: Table x.Summary Logistic Regression Model for Conservative

	Estimates	SE	Test statistics	p-value
(Intercept)	-3.3577	0.4549	-7.3816	0.0000
age	0.0175	0.0057	3.0476	0.0023
sexMale	0.3269	0.1950	1.6760	0.0937
languageFR	-0.6371	0.2575	-2.4737	0.0134
west	0.7986	0.2200	3.6298	0.0003
incomeMedium	0.6465	0.3793	1.7044	0.0883
incomeHigh	0.3561	0.2959	1.2033	0.2288
religion	0.7340	0.2505	2.9305	0.0034
minority	-0.1796	0.2149	-0.8357	0.4033

Table 8: Table x.Estimated probability of voting for the Parties

Liberal	Conservative
0.1711	0.2149

of language speaking, living in Western inland provinces (Alberta, Manitoba, Saskatchewan) or not, income class at Medium or not, income class at High or not, and possess religion or not are significant with the voting decision.

4.2 Result of post-stratifications

<Here you present your results. You may want to put them into a well formatted table. Be sure that there is some text describing the results.>

<Note: Alternatively you can use the `knitr::kable` function to create a well formatted table from your code. See here: https://rmarkdown.rstudio.com/lesson-7.html.>

<Remember you can use `r` to use inline R code.>

<Include an explanation/interpretation of the visualizations. Make sure to comment on the appropriateness of the assumptions/results.>

5. Conclusions

<Here you should give a summary of the Hypotheses, Methods and Results>

<Highlight Key Results.>

<Talk about big picture.>

<Comment on any Weaknesses.>

<End with a concluding paragraph to wrap up the report.>

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: April 4, 1991)
2. RStudio Team. (2020). *RStudio: Integrated Development for R.* RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: April 4, 1991)

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

Liberal Party of Canada. (n.d.). *2021 Platform.* [https://liberal.ca/our-platform/]

Canada Guide. (2023). *Canadian Political Parties.* https://thecanadaguide.com/government/political-parties/ (Last Accessed: November 11, 2023)

Dippel S. (2016). *As Stephen Harper leaves politics, record shows mixed results for Calgary.* CBC. [https://www.cbc.ca/news/canada/calgary/harper-resigns-mp-calgary-analysis-1.3734081]

Elections Canada. (2023). *The electoral system of Canada.* [https://www.elections.ca/content.aspx?section=res&dir=ces&document=part2&lang=e]

Government of Canada, (2022). *Provinces and territories - Intergovernmental Affairs.* [https://www.canada.ca/en/intergovernmental-affairs/services/provinces-territories.html]

Liberal Party. (2019). *The Canadian Encyclopedia.* [https://www.thecanadianencyclopedia.ca/en/article/liberal-party]

Members of Parliament. (2023). *Current Members.* [https://www.ourcommons.ca/Members/en]

Parliament of Canada. (2006). *Amendment to Canada Elections Act.* [https://www.parl.ca/]

Simon Fraser University. (2021). *Canadian Election Results by Party 1867 to 2021.* https://www.sfu.ca/~aheard/elections/1867-present.html (Last Accessed: November 11, 2023)

Statistics Canada. (2022). Chapter 6: Political participation, civic engagement and caregiving among youth in Canada. *Statistics Canada* [https://www150.statcan.gc.ca/n1/pub/42-28-0001/2021001/article/00006-eng.htm] (Last Accessed: November 20, 2023)

Statistics Canada. (October, 2022). The Canadian census-A rich portrait of the country's religious and ethnocultural diversity. *Statistics Canada* https://www150.statcan.gc.ca/n1/daily-quotidien/221026/dq221026b-eng.htm (Last Accessed: November 12, 2023)

Statistics Canada. (2020). General Social Survey - Social Identity (SI): Detailed information for 2020 (Cycle 35). *Statistics Canada.* https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5024 (Last Accessed: November 20, 2023)

Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. (2021). *The 2021 Canadian Election Study.* Consortium on Electoral Democracy. [dataset]

Uppal, S. & LaRochelle-Côté S. (February, 2012). *Factors associated with voting.* Statistics Canada. https://www150.statcan.gc.ca/n1/en/pub/75-001-x/2012001/article/11629-eng.pdf?st=nuV9FTRH (Last Accessed: November 12, 2023)

Tossutti, L. & Elections Canada. (2007). *La participation électorale des membres des communautés ethnoculturelles.* Élections Canada.

Appendix

Generative AI Statement

Here is where you can explain your usage of Generative AI tool(s). Be sure to reference it. For instance, including something like:

I used the following generative artificial intelligence (AI) tool: Bing AI Version 2.0 for Chrome [4]. I used the tool only in the Results section of this assignment and I gave it the following prompt of `What should I eat for breakfast?` and it gave me a list of 10 breakfast items which I then asked it to: `Please only list breakfast items that do not include eggs`. I then chose my 3 favourite items from the produced list and included those in the Results section.

Supplementary Materials

<Here you can include any additional plots, tables, derivations, etc.>