The prediction on Canadian Federal Election 2025

STA304 - Assignment 2

Group 20: Peiyu Liu, Ruiqi Sheng, Zhaoqi Li, Huan Xu

November 24, 2022

Introduction

In order to guarantee that the government maximizes the nation's prosperity in all domains of politics, economics, and education, Canadian federal elections are held every four years(Canada Election,2022). In a parliamentary democracy where the government seeks to enact laws and policies that will primarily reflect the preferences of its people, voting is both a right and a duty for Canadian citizens. Political parties differ in their beliefs, and to reflect these convictions, they propose new laws or amend existing ones. Canada's electoral system is distinguished by single-member constituencies as well as a first-past-the-post voting procedure (FPTP). According to this method, the election is won by the candidate who receives the most votes in a certain electoral district(Doody,2021). Participation in the electoral process is thus a crucial way for the government to consider the real voice of the people and select a party that is responsible for the well-being of society and its people.

Two dominant political parties were competing in the last election, the liberal party and the conservative party. The liberal party won in the last election and has a renowned reputation in universal health care, Canada Pension Plan, Canada Student Loans, and immigration policies. The Conservative Party is politically center-right. They now hold the second-largest number of seats in the House of Commons as the official opposition. Apart from the Liberals, the Conservatives are the only party in power in Canada at the federal level. The Conservatives support low taxes and minimal government and oppose initiatives such as the legalization of abortion and prostitution. In this research, we are interested in more in-depth details regarding the election data that contributed to this result and analyzing the background information of the voters to predict the vote for the Liberal Party in the next Canadian election.

We retrieved two sources for this investigation. Specifically, the "General Social Survey (GSS)" and the "CES," which stand for census and survey data, respectively. The sampling population was chosen from the results of the CES survey. 4,021 observations make up the survey data. Every observation includes a summary of the voters' demographics, including their gender, age, educational background, and responses to questions about their thoughts and preferences about various political parties. In the meantime, the GSS census data will serve as the main study's target population.

Regarding how individuals feel about elections across Canada, the census data covers 20,602 observations and 81 variables. Due to sampling bias in the CES survey's small sample size, we were able to get the census data. To determine the final election results using the post-stratification technique, we simulated the election outcomes using the CES survey and then combined it with the census data.

In this research, we believe that age, sex, education level, and religions to be important factors influencing the voting result.

We still predict that on Canadian Federal Election 2025, the Liberal Party will win the election and the following data collection will generally explain the reason of our prediction.

Data

Two datasets are used in this report, the first one the Canadian Census data in 2017 which entirely collected Canadian Citizens living situations and their basic information within the family, for example: personal information, family situation, province, language abilities and etc. The other one is the CES dataset which collected the data about the voters information in 2019 Canadian Federal Election. It has some data that are same to the GSS Canadian Census data in 2017, but includes more information about the election, such as their federal satisfaction and which party they voted in 2019. The following steps explaine how the data will be cleaned.

In order to select the most relevant parameters from the datasets for further analysis of our models, we decided to clean GSS and CES separately. The most important values, in this case, would be votes for the liberal and conservative party. In addition to that, we also find citizens' attitudes towards the domestic economy, liberal and conservative parties significant. Such factors could help us evaluate the general public impression of their performance in terms of economy, policies, decision making, and leadership in various communities. Hence, we created a new dataframe by taking the selected variables out of the original CES and saved it as a cleaned survey database. Since all the variables are restored numerically, we had to find the definition of each number according to the study documentation(Stephenson et al., 2020) of CES and then explain the values in words. For instance, since q31 represents people's attitude towards domestic economy, in which their votes are on a scale from 1 to 3, we had to define that 1 indicates that they think the economy was becoming better in the past year; 2 indicates the economy was worsened; and 3 indicates that the economy stayed the same. For the convenience of coding, we simply named 1,2,3 as "Better", "Worse", and "About the same" respectively.

After filtering the voters, we started to get interested in their background information and wondered about what could be the reasons that influenced their propensity to vote. As a result, their age, sex, education background, working status, and religion drew our attention. Therefore, we adjusted the cleaned survey database by adding these variables to it and, at the same, changed the numbers to descriptive words.

As the variables mentioned above will be used in our models as beta, we have also extracted them from the GSS database and formed a new datafram called cleaned census. In order to make sure that all the relevant data in cleaned survey and cleaned census are mutually corresponding, we need to make some adjustments. For example, the education background in the cleaned census data has a variable "High school diploma or a high school equivalency certificate", whereas it is described as numerical numbers 4 and 5 in survey data. Therefore, we had to change both of them to "High school" for the purpose of consistency. And so did we do the samething for the data in sex, age, and religion.

The following explanations are for the variables used in the models as beta.

Variables

- Age

The first variable we think is important is the age group. People were separated into different numerical age groups in the CES survey, and they had different standards for their ideal leaders. A large percentage of young people who have just reached election age do not know how to choose a party and do not have a deep understanding of the different parties, so most of them choose to vote based on their parents' choice of party. However, unlike teenagers, middle-aged people have already gone through many elections. They know that different parties have different impacts on people's lives, education, and the economy, so middle-aged people will choose the party of their choice, taking into account the impact on their lives. Finally, older people choose parties that are generally more traditional. In their hearts, there are some parties and some sentimental parties to choose their direction is not very easy to change, and they are less likely to choose not to understand the emergence of new parties.

- Sex

As for gender, people of different genders will also have largely different ideas on the general election. The CES survey presented three categories when the question mentions gender. Gender is not binary; some people do not want to reveal or define their gender. Furthermore, for this group of people, our choice is not to subsume it into the discussion for now. The Liberal and Conservative parties have different ideas about women's right to abortion. The Liberal party believes that women should have the right to control their bodies to have an abortion. In contrast, the Conservative party believes that women are generally not allowed to have abortions(Wherry, 2022). This move by the Liberal party had a chance to win votes among women to a great extent. So the gender gap also had a considerable impact on the election results.

- Education level

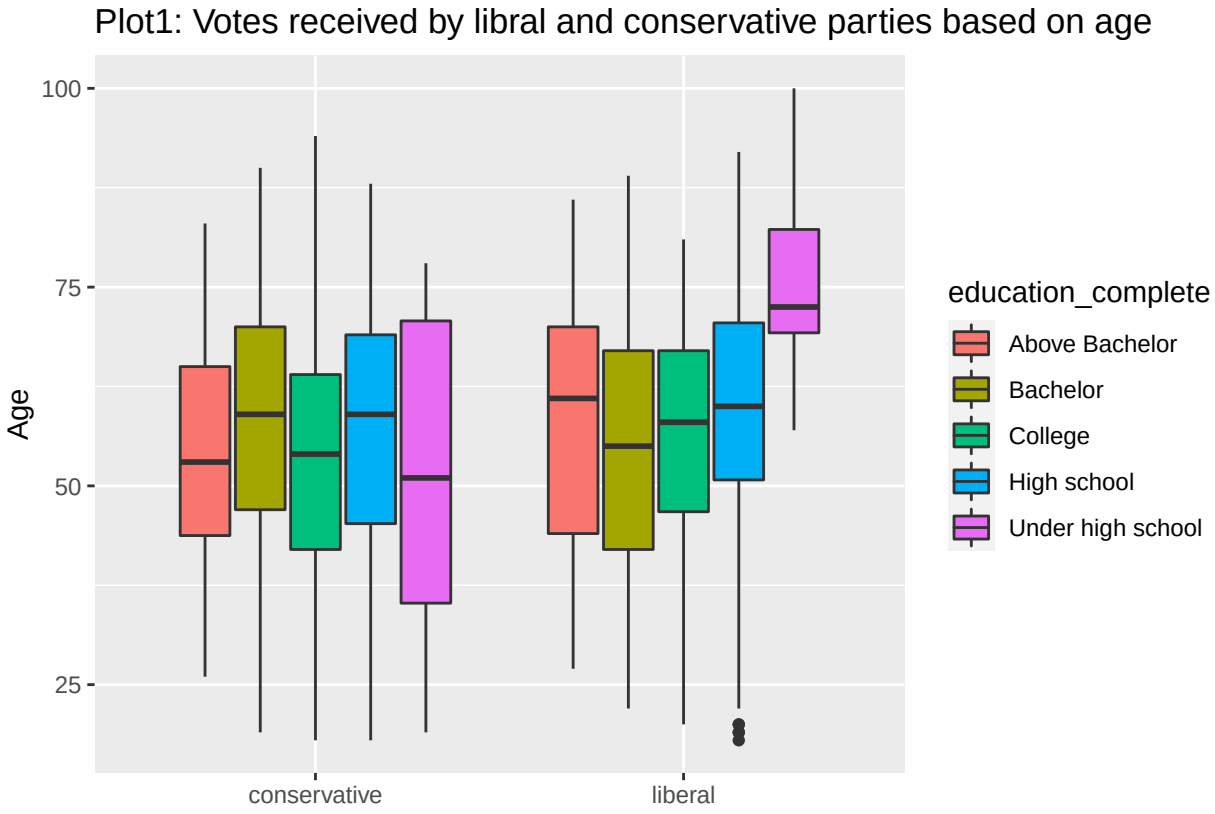
People's education level also largely influences their judgment of political party choices. The CES gives out 11 levels of education to categorize people's s cognitive abilities. Less educated people will focus on their immediate interests, such as basic human needs. However, the more educated will have their own deeper thinking about the facts and the sustainability of humanity and the world. When global warming threatens the environment and the world's survival, the Conservative Party proposes reducing the carbon tax(Cochrane,2021), reducing the economic burden on people. However, on the other hand, it will also liberalize our limits on carbon emissions and allow global warming to continue to threaten our environment. Individuals with varying levels of education place different values on the two possible outcomes.

- Religion importance

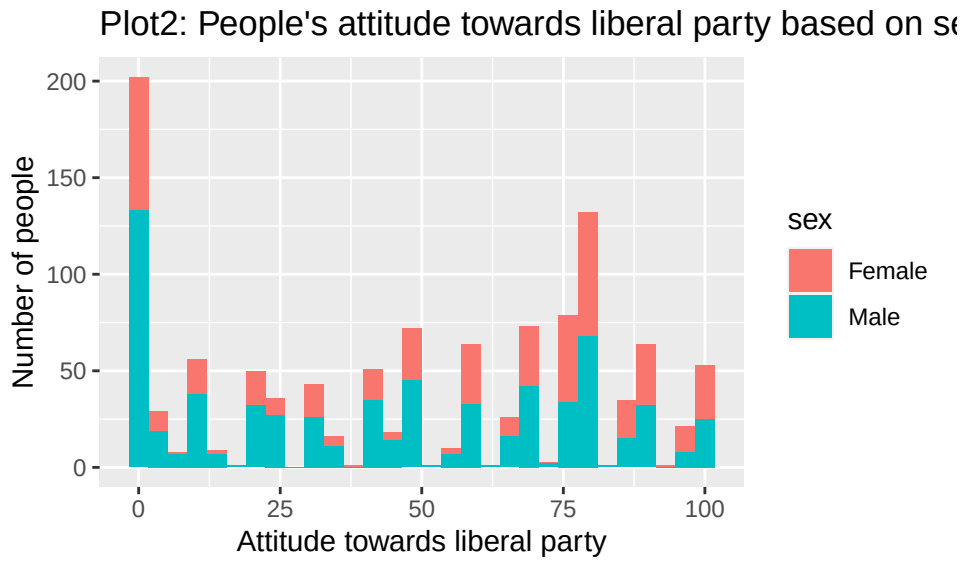
Since colonization, Christianity has been the dominant faith in this country (Evason, 2016). According to the CES survey, it categorized people's attitudes toward the importance of religions into four different levels. We assume people who believe in religion would participate more in the election, as people who have religion are more focused on social unity and stability(Emerson, 2011). At the same time, political views for the common progress of society and for the promotion of public security are more likely to be favored by people who believe in religion. So, we wonder how people who value religion differently would have different political views and vote for different parties.

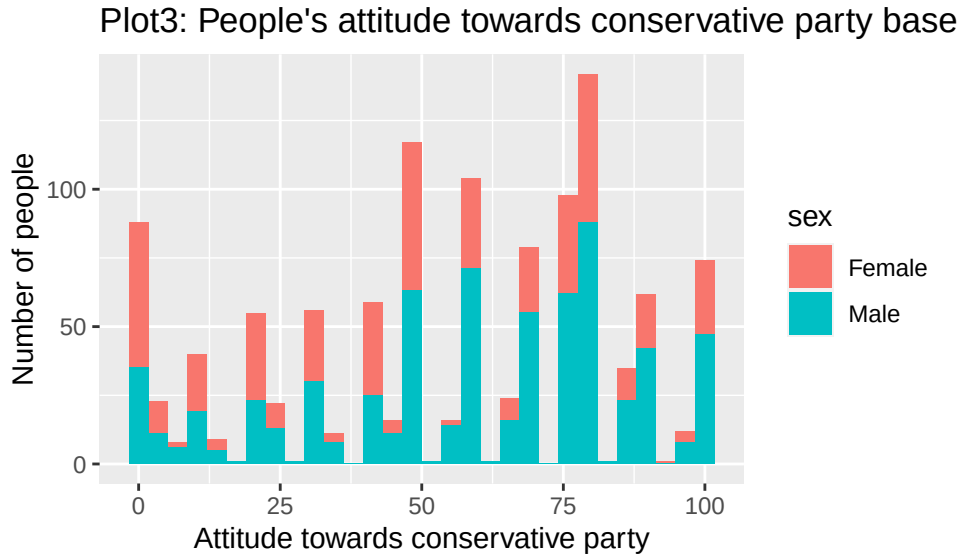
votes_count	liberal	consv	rate
1151	622	529	54%:46%

From this table, we selected that valid votes counts towards both Liberal and Conservative parties are 1151 and the Liberal has 622 votes which occupied 54% in total. Otherwise, the Conservative own the rest 46% of votes.

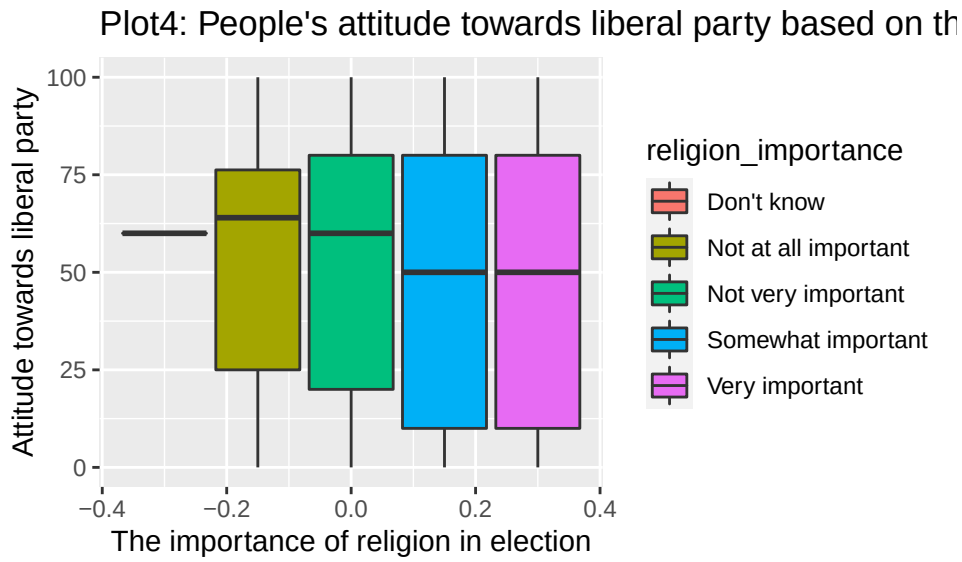


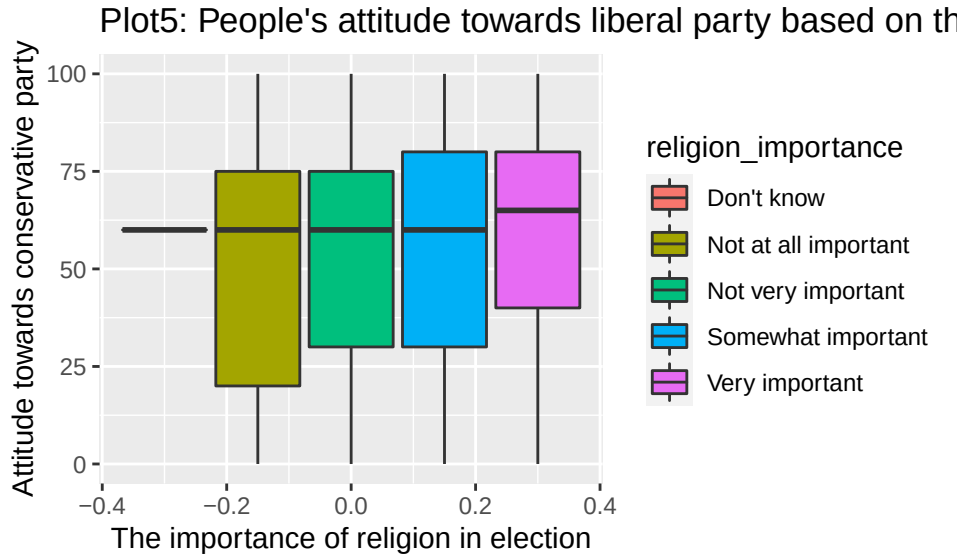
This is a boxplot that depicts the people's age distribution and their education background based on the votes received by liberal and conservative party. As we can see from the plot, people who received an above bachelor degree and voted for liberal party have a higher average age than those who voted for conservative party. Besides, people who have under high school education background and voted for conservative party are much younger(age median around 51) than those who voted for liberal party(age median around 72).





The two plots combined illustrate the relationship between people's attitude towards the two different parties and their opinion of how much religion would influence the decision in election. Apparently, people who believe that religion is important to a certain degree(somewhat and very important) would be inclined to have a better attitude towards conservative party. And people who are in favour of liberal party hold different opinion, as the blue and purple box in plot 3 are longer than in plot 2. As a result, people who think that religion is important have a lower median attitude towards liberal party than conservative party.





These two graphs together explained people's attitudes towards liberal and conservative parties based on their sex. Plot 4 is right skewed, which indicates that the left bar is the highest in the graph. It demonstrates that most of man and women are not satisfied with liberal party at all. However, there are two modes in plot 5, in which people's attitude towards conservative party are centered around 50 and 80. The distribution of the attitude of male towards conservative party shows a reverse trend in plot 5, which indicates that it is left skewed. In other words, more males tend to support conservative party than liberal party in elections, where as female's attitudes are more generally distributed.

Methods

In this research, we will predict the voting situation of the liberal party in Canada which is the response variable. There are only two possible outcomes: voting or not voting, which is a binary variable. Therefore we should use a logistic regression model.

Model Specifics

For logistic regression model, we assume the observations are independent and the sample size is large. According to GSS and CES, we can know the observations are independent, because each individual can only fill in one response. Additionally, the sample size is large enough with 4,021 observations of CES and 20,602 observations of GSS. We also assume it is linearity which means there is a linear relationship between predictors and the logit of the response variable. Additionally, there is no multicollinearity and no outliers.

By comparing the AIC, we can determine which model is more suitable for predicting the proportion of people who will vote for the liberal party. We choose model with lower AIC.

We will use frequentist logistic regression model to model the proportion of people who will vote for the liberal party. We will use age, sex, education, and religion as predictors.

Logistic regression frequentist:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{education} + \beta_4 x_{religion} + \epsilon$$

Where $\log \frac{p}{1-p}$ represents the proportion of people who will vote for the liberal party. β_0 represents when age, sex, education situation and religion situation is 0 (do not have influence of liberal party vote), the response of the proportion of people who will vote for the liberal party. β_1 represents the coefficient of age. β_2 represents the coefficient of sex. β_3 represents the coefficient of education level. β_4 represents the coefficient of religion importance situation.

Post-Stratification

In order to estimate the proportion of voters, the post-stratification will be used to accurate the estimate the of the proportion. We think that four categories of variables: age, sex, education levels and religions would impact the accuracy of the data. After the generation of logistic regression models above and the comparing of four different models, we can used the result to weight the average estimator of all possible combination of attributes. The general formula $\hat{y}^{PS}$ will be specified with the explanation of variables to estimate the proportion of the original dataset which is the 2019 census data. In the expended formula, where has four variables, i is age, j is sex, k is education levels and l is religion importance. The reason why we have to use the post stratification is that the variables in survey data didn't include many indication like census data had. For example: the education variable in survey data did have "Trade Certificate" option, which means the survey data can't represent the entire population. And the result for the next Federal Election won't be accurate if the census data don't count into the post-stratification. Therefore, all i, j, k and l, those variables would help us to stratify the original data as well and the result would be more reliable

$$\hat{y}^{PS} = \frac{\sum_i \sum_j \sum_k \sum_l N_{ijkl} \hat{y}_{ijkl}}{N} = \sum_i \sum_j \sum_k \sum_l W_{ijkl} \hat{y}_{ijkl}$$

Where i represents the age, j represents the sex, k represents the education level, l represents the importance of religion.

Results

Table 2: The AIC value of frequentist and frequentist multilevel logistic regression model

aic_model	aic_model_multi
1527.7	1541.7

Model fitting

In the project, we fit frequentist logistic regression model and frequentist multilevel logistic regression model. By comparing the AIC (Table 2), the frequentist logistic regression model has a AIC value of 1527.7 which is lower than the AIC value of the frequentist multilevel logistic regression model (1541.7). Therefore, the frequentist logistic regression model is more suitable for predicting the proportion of people who will vote for the liberal party with lower AIC value.

```
## # A tibble: 11 x 5
##    term                                  estimate std.error statistic p.value
##    <chr>                                    <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)                           -12.2     325.        -0.0375 9.70e-1
##  2 age                                     0.00466   0.00385    1.21   2.26e-1
##  3 sexMale                                -0.636     0.126     -5.05   4.32e-7
##  4 education_completeBachelor             -0.561     0.183     -3.06   2.20e-3
##  5 education_completeCollege              -1.18      0.194     -6.09   1.10e-9
##  6 education_completeHigh school          -0.958     0.207     -4.64   3.52e-6
##  7 education_completeUnder high school    -0.851     0.487     -1.75   8.05e-2
##  8 religion_importanceNot at all important 13.0     325.         0.0401 9.68e-1
##  9 religion_importanceNot very important   13.1     325.         0.0403 9.68e-1
## 10 religion_importanceSomewhat important   12.9     325.         0.0398 9.68e-1
## 11 religion_importanceVery important       12.6     325.         0.0387 9.69e-1
```

Model interpretation

According to the table, we know the coefficient estimate of age is 0.004658, which means with the same sex, education level, and religion importance, 1 year of age increase will increase the proportion of people who vote for the liberal party by 0.004658. For the variable sex, the coefficient estimate of male is -0.636216 which is less than 0. Therefore, this variable play a negative role for the probability of people vote liberal party.

liberal_predict_result
0.4473536

Post-Stratification

In the calculation of post stratification above, the y value of each part is calculated by referring to the influence of age, gender, education level and religion. After integrating the sum and finally bringing it into the forecast of the original data, we calculated that the Liberal Party will win the 2025 Canadian federal election by about 44.7%. This result is roughly accurate and can be reproduced. However, there are still some deviations in this result. These deviations come from the errors of data statistics, because there are differences in data statistics problems in the comparison items in the two data. We also try our best to eliminate these errors in data clean, but the obtained ratio of 44.7% is still only an ideal value.

Conclusions

In this research, we aim to analyze how the four variables(age, sex, education level and religion) would influence the voting result of liberal party. People whose highest education level is under high school or above bachelor are more likely give the vote to liberal parties, which gives us two outcomes

All the data we applied in this study are from 2019, and next time, we could try more recent data, as people's attitudes towards different political parties are likely to change depending on the performance of the prime minister in the area currently in power. In addition, we chose to obtain our findings only from cell phones, which limited the actual responses we collected from the public.

To make our result more reliable, it is useful to include more variables in the regression model, in this study, we only include four variables to model the result; there are 81 variables in total from the CES.

The data collection method was also crucial for conducting a detailed study. When we selected the variables to be studied further, we found drawing graphs with categorical variables difficult. Suppose we want to generate the next study. In this case, we can expand the categorical variables to numerical results, making it easier to observe trends in the graphs. In addition, use poststratification to make the data more reliable and accurate. Use the census data model to establish a comparison and bring it into the original data to analyze the lack of original data, and try to complete the original data to make it perfect. Therefore, poststratification can better predict whether the results of the Canadian federal election in 2025 are accurate.

Based on the prediction of our research, political parties can analyze their promotion target and directions,

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)
4. Canada, E. (n.d.). Home. – Elections Canada. https://www.elections.ca/content.aspx?section=vot&dir=faq&document=index&lang=e#elec0. (Last Accessed: November 29, 2022)
5. Liberal Party of Canada. (n.d.). https://liberal.ca/. (Last Accessed: November 29, 2022)
6. Wherry, A. (2022, May 5). Analysis | the fate of Roe V. Wade puts both liberals and conservatives on the spot | CBC news. CBC news. https://www.cbc.ca/news/politics/roe-wade-abortion-supreme-court-canada-1.6441654. (Last Accessed: November 29, 2022)
7. Cochrane, D., Shivji, S., & Wherry, A. (2021, April 16). Conservatives announce plan to replace liberal carbon tax with a lower levy of their own | CBC news. CBC news. https://www.cbc.ca/news/politics/carbon-tax-conservatives-1.5988407. (Last Accessed: November 29, 2022)
8. Evason, N. (2016). Canadian culture - religion. Cultural Atlas. https://culturalatlas.sbs.com.au/canadian-culture/canadian-culture-religion. (Last Accessed: November 29, 2022)
9. Doody, B. (2021, August 12). Canada's electoral system. repolitics. https://repolitics.com/features/canadas-electoral-system/#election. (Last Accessed: November 29, 2022)
10. Stringer, A. G. A. A. (2021, January 20). Chapter 16 Short tutorial on pulling data for Assignment 1 | Probability, Statistics, and Data Analysis. https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-canadian-election-study. (Last Accessed: November 25, 2022)
11. General Social Survey: An Overview, 2019. (2019, February 20). https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm. (Last Accessed: November 25, 2022)
12. Stephenson, L. B., Harell, A., Rubenson, D., & Loewen, P. J. (2020). Canadian Election Study, 2019, Phone Survey: Study Documentation. https://github.com/hodgettsp/ces_data/blob/master/documentation/ces2019/Canadian%20Election%20Study%2C%202019%2C%20Phone%20Survey.pdf. (Last Accessed: November 25, 2022)
13. Emerson, M. O., Monahan, S. C., & Mirola, W. A. (2011). Religion matters: What sociology teaches us about religion in our world. Upper Saddle River, NJ: Prentice Hall.

Appendix

The summary data of models:

```
##
## Call:
## glm(formula = vote_liberal ~ age + sex + education_complete +
##     religion_importance, family = "binomial", data = survey_cleaned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7376  -1.0656  -0.7749   1.1719   1.7629
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -12.167388 324.743787  -0.037   0.9701
## age                                       0.004658   0.003851   1.210   0.2264
## sexMale                                  -0.636216   0.125872  -5.054 4.32e-07
## education_completeBachelor               -0.560675   0.183156  -3.061   0.0022
## education_completeCollege                -1.179520   0.193557  -6.094 1.10e-09
## education_completeHigh school            -0.957783   0.206517  -4.638 3.52e-06
## education_completeUnder high school      -0.851380   0.487098  -1.748   0.0805
## religion_importanceNot at all important  13.007218 324.743791   0.040   0.9681
## religion_importanceNot very important    13.101309 324.743764   0.040   0.9678
## religion_importanceSomewhat important    12.938003 324.743746   0.040   0.9682
## religion_importanceVery important        12.564368 324.743747   0.039   0.9691
##
## (Intercept)
## age
## sexMale                                 ***
## education_completeBachelor              **
## education_completeCollege               ***
## education_completeHigh school           ***
## education_completeUnder high school     .
## religion_importanceNot at all important
## religion_importanceNot very important
## religion_importanceSomewhat important
## religion_importanceVery important
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1594.9  on 1155  degrees of freedom
## Residual deviance: 1510.7  on 1145  degrees of freedom
## AIC: 1532.7
##
## Number of Fisher Scoring iterations: 11
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## vote_liberal ~ age + religion_importance + (1 | sex) + (1 | education_complete)
##    Data: survey_cleaned
```

```
##
##      AIC      BIC   logLik deviance df.resid
##   1546.7   1587.1   -765.4   1530.7     1148
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7817 -0.8774 -0.6072  1.0011  1.8831
##
## Random effects:
##  Groups              Name        Variance Std.Dev.
##  education_complete (Intercept) 0.1832   0.4281
##  sex                (Intercept) 0.1183   0.3440
## Number of obs: 1156, groups:  education_complete, 5; sex, 2
##
## Fixed effects:
##                                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -13.155947  29.270303  -0.449    0.653
## age                                       0.004784   0.003838   1.246    0.213
## religion_importanceNot at all important  12.973434  29.269766   0.443    0.658
## religion_importanceNot very important    13.056929  29.269747   0.446    0.656
## religion_importanceSomewhat important    12.895365  29.269545   0.441    0.660
## religion_importanceVery important        12.526628  29.269514   0.428    0.669
##
## Correlation of Fixed Effects:
##             (Intr) age    r_Naai rl_Nvi rlg_Si
## age         -0.005
## rlgn_mpNaai -1.000 -0.002
## rlgn_mprNvi -1.000 -0.002  1.000
## rlgn_mprtSi -1.000 -0.002  1.000  1.000
## rlgn_mprtVi -1.000 -0.003  1.000  1.000  1.000
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```