Predicting the Results of the 45th Canadian Federal Election

STA304 - Assignment 3

Group 38: Ethelia Choi - 1005204976 Cameron Dietzel - 1003506191
Minh Nguyen - 1005038594 Andy Vu - 1005244932

November 5, 2021

Introduction

National elections are an essential component of Western democracy, and provides citizens the liberty of influencing the policy and governance of their nation. For a majority of democratic states in the world, national elections are held on a cyclical basis, often between 4-5 years. Out of a large pool of political parties, citizens have the opportunity to vote for the political party that they support. The winning party gains majority control of the government, with the leader of that party being delegated as the head of government.

The 45th Canadian Federal Election will take place on October 20, 2025, with the winner of the election being determined by the political party with the most seats in the House of Commons. There are 338 seats in the House, corresponding to 338 of Canada's electoral districts (better known as ridings). Representatives of major political parties vie for the majority vote within each riding in order to become a Member of Parliament (MP), thus taking up a seat within the House of Commons.

The purpose of this report is to **predict the popular vote leading up to the 45th Canadian Federal Election**. Popular votes are a measure of how popular a candidate is among Canadians, and is determined by tallying the total number of votes that a political party receives across Canada. While often used as a predictor for the most likely party to win the majority of seats in the House, it should be noted that the popular vote is not a foolproof method for determining the majority seat-holder (and thus, the winner of the election) in the House [8].

To predict the popular vote, we construct 7 full multivariate logistic regression models with post-stratification. The output of each model represents the likelihood of 1 of the 7 parties winning the election. This requires us to have access to two datasets: a survey dataset and a census dataset. We proceed to clean the survey dataset and impute/categorize the relevant variables to match their counterparts in the census dataset; specifics will be discussed in the Results section of the report. After constructing our models with the survey data, we input our census data into our models in order to get a probability for a particular party winning the popular vote; this process is known as post-stratification.

We hypothesize that the Conservative Party of Canada (CPC) will have the highest likelihood of gaining the majority vote, solely based on the popular vote results of the 44th and 43rd Canadian Federal Elections [9].

Data

Data Collection

There are two datasets that will be used in this analysis. The first of which is the survey data, which will be used to construct multiple logistic regression models. The second dataset is the census data, which will be used in conjunction with the regression models for post-stratification. Both of these datasets have been collected from different sources.

The survey data used is from the *2019 Canadian Election Study - Phone Survey*. This survey was conducted by phone during and after the Canadian federal election in 2019 by Stephenson et al. [1]. The process used

Computer-assisted telephone interviewing, to collect data from Canadian citizens and permanent residents who were 18 years of age or older [1]. Most importantly, the survey contained data on who the participant will likely vote for in the upcoming election. The survey, although focused on the election, public opinion, and politics, also included personalized questions about the participant, such as their age, gender, income, province of residence, ethnic/cultural identity, etc. These variables are what give us the ability to use the post-stratification process. As a result, it is crucial that these variables are also able to be mapped to the ones found in the census dataset.

The census data used is from the *General Social Survey - Family (GSS) (Cycle 31) 2017*. This census was conducted by the Government of Canada as part of the annual General Social Survey with Cycle 31 in 2017 focusing on Canadian Families [4]. Responding to the census was voluntary, and similar to the survey data, most of the responses were collected through a computer-assisted telephone interview [4]. Unlike the survey data, since the census was conducted by the Government of Canada, linking has also been used to get data without having to ask the respondent questions [4]. That is, the census linked the respondents to their personal tax records to obtain their income data, as well as other personal records to obtain their address, date of birth/ age, sex, etc. [4]. With the census data, we now have all the parts to predict the popular vote using multivariate logistic regression models with post-stratification.

Data Cleaning

The first step of any analysis is to clean the data and extract only the variables of interest from the dataset. In this process, we will also be removing any invalid observations to ensure that the data used for the regression model is complete. Since we will be using a regression model with post-stratification, there are two datasets. In the cleaning process, we also need to ensure that the two datasets result in the same explanatory variables with corresponding factors.

The first dataset that will be cleaned is the survey data. Since the regression model that we plan to use is logistic regression, we need to create a binary variable for the response. The idea is to create a model for each political party, hence a new separate variable will need to be created for each possible party. Then, by using Question 11 in the survey data, which asks "If you decide to vote, which party do you think you will vote for?", we created a binary variable for whether or not the respondent would vote for the Liberal Party, Conservative Party, New Democratic Party, Bloc Québécois, Green Party, People's Party, or an Other Party[1].

The next step is to clean the observational variables that will be used in the model. The first variable of interest is age. For this variable, we used Question 2 in the survey which asks for the respondent's birth year and the year in which the survey was conducted, 2019, to calculate the respondent's age at the time. The second variable of interest is gender. There are three possible choices for gender from the survey: Male, Female, or Transgender. We notice that there is only sex, which is binary, in the census data, hence we will discuss a method to deal with this in the Results section, under Sex and Gender. For now, gender is taken from the "Interviewer gender for CES" question, in the survey data.

The third variable we are interested in is income, specifically family/household income. Since the household income in the census data is given as categorical ranges, we created a new categorical variable for income. Using the numeric responses from Question 69 of the survey, which asks for the respondent's total household income before taxes, we assigned each value to the following categories: "Less than $25,000", "$25,000 to $49,999", "$50,000 to $74,999", "$75,000 to $99,999", "$100,000 to $ 124,999" and "$125,000 and more". The third observational variable is the country of birth. In the census data, we are only given whether or not the respondent is born in Canada. Hence for the survey data, we created a new variable for birthplace and used Question 64 to assign the respondents who responded to being born in Canada or Quebec to "Born in Canada" and those who responded being born in other countries to "Born outside Canada".

The last variable is education. We notice that the census data asks for the highest level of completed education, whereas the survey has options for some of a level of education. Thus, creating a new variable for education, we used the responses from Question 64 of the survey to aggregate the responses to the following categories: "Less than high school diploma or its equivalent", "High school diploma or a high school equivalency certificate", "Completed or Some Technical, College, CEGEP or other non-university certificate

or di...", "Completed or Some University certificate or diploma below the bachelor's level", "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)", and "University certificate, diploma or degree above the bach...". After cleaning and creating the new variables, we filter out all of the unnecessary and unused variables. The last crucial step is to omit any observations with missing values to ensure that the data is complete.

The second dataset that will be cleaned is the census data. Since most of the cleaning in the survey data has been made to complement the census data, there is not much to do here. For example, the variables for age and province did not need any cleaning as the survey data had already been made to match it. Similarly for the variable representing sex, there is no need to change anything in the census data as alterations will need to be made for the gender variable in the survey data.

The first variable that needs a simple change is the variable for family income. The variable name for family income in the census data is changed to income for simplification. The second variable that needs work is the variable for birthplace. Since this variable included invalid responses, we made a new variable and made sure to only keep those who responded either "Born in Canada" or "Born outside Canada". The last variable that needs cleaning is the one for education. For this variable, we did the same thing as we did for the survey data, aggregating the responses to the categories: "Less than high school diploma or its equivalent", "High school diploma or a high school equivalency certificate", "Completed or Some Technical, College, CEGEP or other non-university certificate or di...", "Completed or Some University certificate or diploma below the bachelor's level", "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)", and "University certificate, diploma or degree above the bach...". Lastly, we filter out all of the unnecessary and unused variables and omit any observations with missing values to ensure the data is complete.

Description of Variables

Survey Data

Table 1: Description of the Important Variables in the Survey Data

Variable	Description
Age	The age of the respondent
Gender	The gender identity of the respondent (Male/Female/Transgender)
Province	The province in which the respondent is living in
Income	The respondent's household/family income
Birth Place	Whether or not the respondent was born in Canada or outside of Canada
Education	The highest level of education that the respondent has completed
Vote for Liberal	1 if they would vote for the Liberal Party, 0 if not
Vote for Conservative	1 if they would vote for the Conservative Party, 0 if not
Vote for NDP	1 if they would vote for the New Democratic Party, 0 if not
Vote for Bloc Québécois	1 if they would vote for the Bloc Québécois, 0 if not
Vote for Green Party	1 if they would vote for the Green Party, 0 if not
Vote for People's Party	1 if they would vote for the People's Party, 0 if not
Vote for some other Party	1 if they would vote for some other party, 0 if not

Census Data

Table 2: Description of the Important Variables in the Census Data

Variable	Description
Age	The age of the respondent
Sex	The biological sex of the respondent (Male/Female)
Province	The province in which the respondent is living in
Income	The respondent's household/family income
Birth Place	Whether or not the respondent was born in Canada or outside of Canada

Variable	Description
Education	The highest level of education that the respondent has completed

Numerical Summaries and Plots

Table 3: Age summaries for the survey and census data

Variable	Min	First.Quartile	Median	Mean	Third.Quartile	Max
Age - Survey	18	38.0	51.0	50.84	64.0	100
Age - Census	15	37.4	54.3	52.21	66.7	80

Table 3 outlines the summary measures for the numeric variable, age, with respect to both the survey and census data. It should be noted that the median age of the survey population is roughly 3 years younger than the census population, while there is also a much greater spread of ages in the survey data. Here we also notice that the minimum age that appears in the census data is 15, whereas the minimum age in the survey data is 18. The reason for this is because the survey was aimed towards eligible voters, who in Canada must be at least 18 years of age. In hindsight, it may not make sense to post-stratify to an age group that is not eligible to vote. However, we are predicting the popular vote of the 45th Canadian Federal Election, hence the 15 year olds in the census would be eligible to vote then. This highlights some discrepancies between the ages of those within the two data sets, which will be corrected when they are grouped in the post-stratification procedure.

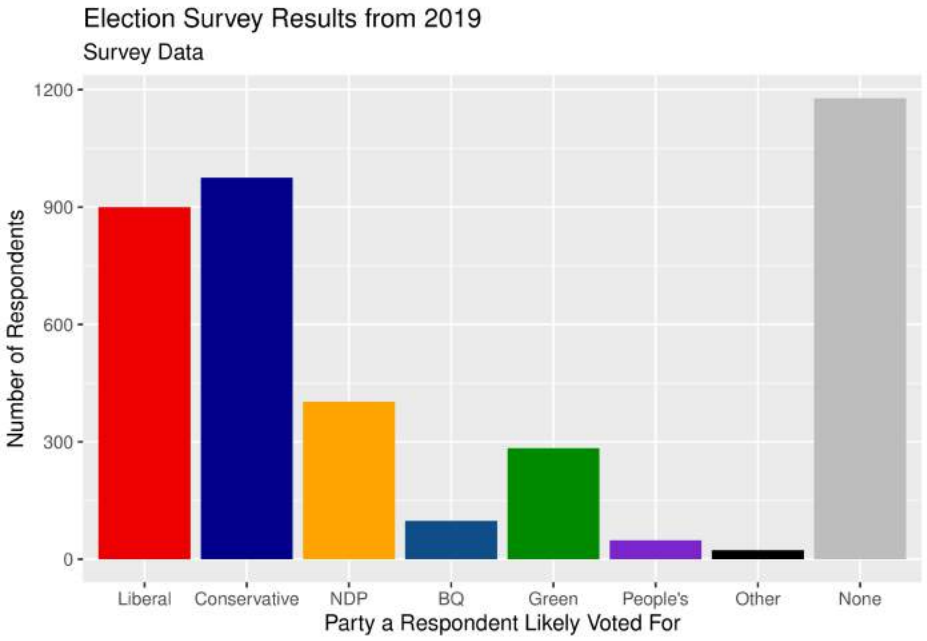

Figure 1: Election Survey Results from 2019

Figure 1 displays the distribution of the party each respondent was most likely to vote for in the 2019 election, based on the survey results. The key takeaway from this figure is the dominance of the liberal and conservative parties among the rest, which indicates a high potential for one of these two parties to win the popular vote in the upcoming election. Another key point is that there is a large chunk of the sample who responded to voting for none of the parties. It is crucial to not remove these observations because it provides a true representation of the population. That is, we do not expect all Canadians to vote, even if they are eligible.

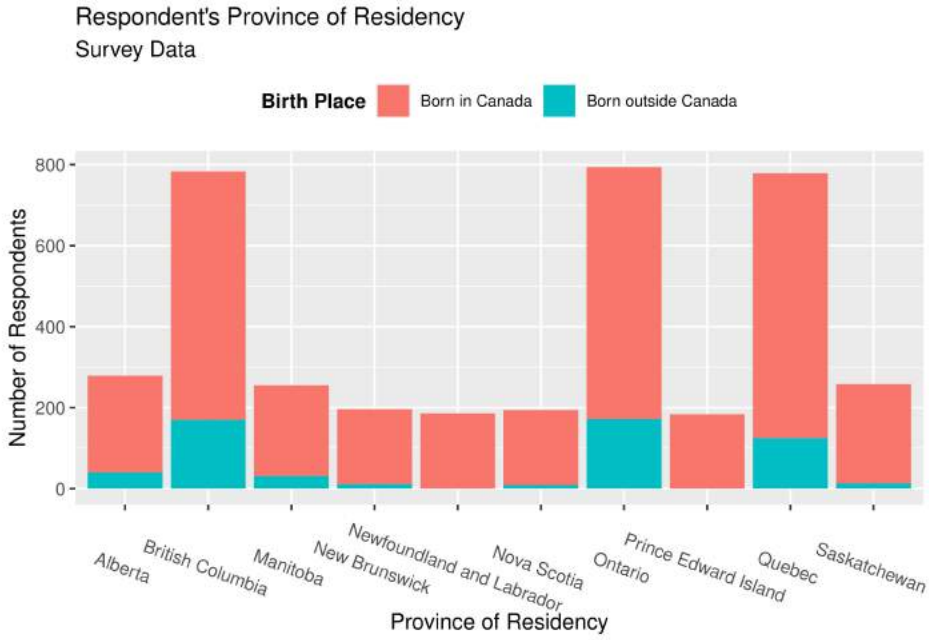

Figure 2: Respondent's Province of Residency

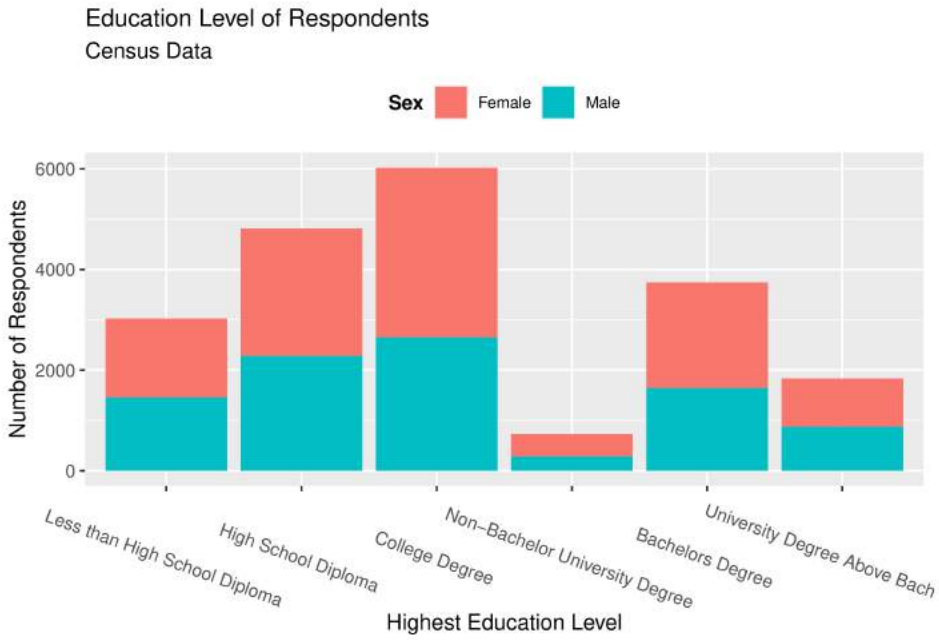

Figure 3: Education Level of Respondents

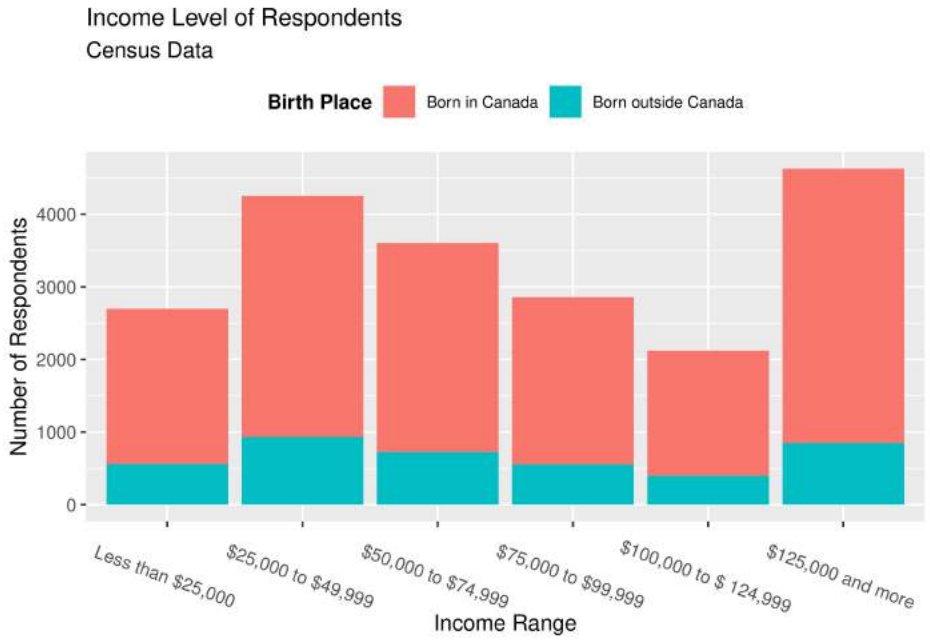

Figure 4: Income Level of Respondents

Figure 2 displays the distribution of each respondent's province of residency, separated by their birth location. The key takeaway from figure 2 would be that not only are the majority of respondents from the provinces of British Columbia, Ontario, and Quebec, but these provinces also appear to have the highest percentage of those who do not claim to be born in Canada.

Figure 3 analyzes the education level of those in the population, separated by sex. It should be noted that roughly 70% of all individuals within this sample have a post-secondary education, a figure which is right around the national average [5].

Figure 4 displays the range in household income of those within the population. The figure shows that the majority of individuals are not earning the nation's average income, as the highest frequency ranges consist of those in the $25,000 to $49,999, and $125,000 and up ranges, respectively. The figure also indicates consistency among the income earned for those born in and outside of Canada.

Methods

In each cycle of the Canadian Federal Election, there are six main political parties that compete for seats in the House of Commons. The leader of the party with the most seats becomes the acting Prime Minister. As the purpose of this study is to determine which of these parties will win the popular vote in the upcoming election, it is a question of whether or not a party would receive a vote. As such, we will apply a logistic regression model to accomplish this goal. This form of the model is appropriate as logistic regression models are suitable for using multiple categorical and numerical predictors to determine a binomial outcome. Since there are six main parties, as well as a seventh consisting of politicians who are running independently, we will invoke seven different models, all independent of each other, to determine the probability of the Canadian population voting for each respective party. The binomial distribution follows as each model determines the percentage of votes the reference party will receive, with the opposite outcome being a vote for any of the remaining six parties or none at all. Each model will then be used to post-stratify the census data in order to make voting inferences on the greater population.

Model Specifics

The model will implement one numerical predictor and five categorical predictors from the survey data to determine the outcome of the popular vote of the next election. The numerical predictor includes the age of the respondent, while the categorical predictors include the gender, family income level, province of residency, place of birth, as well as the highest education level completed by/of the respondent. Apart from these variables being of general interest for us to analyze, they also represent reasonable predictors as voter intent likely varies based on these demographics. The variables were also chosen as they allowed an accurate map from the survey data to the census data in order to perform post-stratification, as mentioned in the latter part of this section.

Therefore, the logistic regression model for each party will have the following form:

$$
\begin{aligned}
\hat{y}_{\text{party}} = {} & \beta_0 + \beta_{\text{age}} x_{\text{age}} + \beta_{\text{male}} x_{\text{male}} + \beta_{\text{transgender}} x_{\text{transgender}} \\
& + \beta_{\text{BC}} x_{\text{BC}} + \beta_{\text{MB}} x_{\text{MB}} + \beta_{\text{NB}} x_{\text{NB}} + \beta_{\text{NL}} x_{\text{NL}} + \beta_{\text{NS}} x_{\text{NS}} \\
& + \beta_{\text{ON}} x_{\text{ON}} + \beta_{\text{PE}} x_{\text{PE}} + \beta_{\text{QC}} x_{\text{QC}} + \beta_{\text{SK}} x_{\text{SK}} \\
& + \beta_{\text{More than 125,000}} x_{\text{More than 125,000}} + \beta_{\text{25,000 to 49,999}} x_{\text{25,000 to 49,999}} \\
& + \beta_{\text{50,000 to 74,999}} x_{\text{50,000 to 74,999}} + \beta_{\text{75,000 to 99,999}} x_{\text{75,000 to 99,999}} \\
& + \beta_{\text{Less than 25,000}} x_{\text{Less than 25,000}} + \beta_{\text{unknown income}} x_{\text{unknown income}} \\
& + \beta_{\text{born outside Canada}} x_{\text{born outside Canada}} \\
& + \beta_{\text{non-uni certificate}} x_{\text{non-uni certificate}} \\
& + \beta_{\text{uni certificate below bachelor's degree}} x_{\text{uni certificate below bachelor's degree}} \\
& + \beta_{\text{high school diploma}} x_{\text{high school diploma}} \\
& + \beta_{\text{less than high school}} x_{\text{less than high school}} \\
& + \beta_{\text{above bachelor's degree}} x_{\text{above bachelor's degree}} + \epsilon
\end{aligned}
$$

In the above model, β_0 is the vertical intercept, which represents the default/reference group for all of the categorical variables. $\hat{y}_{\text{party}}$ is a log-odds binomial response variable, where 1 denotes a definite vote for the corresponding party, and 0 denotes a definite no vote for the corresponding party. For the numerical variable, x_{age} represents the age of the respondent, and β_{age} represents the effect that a unit change in the age of an individual has on the outcome of their log-odds to vote for the corresponding party, holding all else constant. ϵ represents some random error in the model.

Gender: Gender is a categorical variable with male, female and transgender distinctions. For our model the female gender is chosen as the reference category. Therefore, x_{male} is an indicator variable where $x_{\text{male}} = 1$ if the respondent identifies as male, and $x_{\text{male}} = 0$ if the respondent does not identify as male. β_{male} represents

the change in the log-odds of voting for the party of interest, relative to if the individual identified as female and holding all else constant. Similar implications for $x_{\text{transgender}}$ and $\beta_{\text{transgender}}$.

Province: The respondent's province of residency is a categorical variable with 10 feasible categories. Alberta is chosen to be the reference category. Therefore, x_{BC}, x_{MB}, x_{NB}, x_{NL}, x_{NS}, x_{ON}, x_{PE}, x_{QC}, and x_{SK} represent indicator variables which identify whether the respondent is from British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec, and Saskatchewan respectively. Therefore, $x_{\text{BC}} = 1$, if the province that the respondent resides in is British Columbia, and $x_{\text{BC}} = 0$, if the respondent does not reside in the province of British Columbia. Similarly for x_{MB}, x_{NB}, x_{NL}, x_{NS}, x_{ON}, x_{PE}, x_{QC}, x_{SK}. If the respondent is from Alberta then $x_i = 0$ for all $i \in$ {BC, MB, NB, NL, NS, ON, PE, QC, SK}(Please refer to the Appendix for the abbreviation of the Provinces). β_{BC}, β_{MB}, β_{NB}, β_{NL}, β_{NS}, β_{ON}, β_{PE}, β_{QC}, and β_{SK} represent the respective change in the log-odds of voting for the party of interest, relative to if the respondent was from Alberta, holding all else constant.

Income: The household income of the respondent is a categorical variable with seven distinct income intervals. For this variable, the chosen reference category will be the household income range of \$100,000 to \$124,999. Therefore, $x_{\text{More than 125,000}}$, $x_{\text{25,000 to 49,999}}$, $x_{\text{50,000 to 74,999}}$, $x_{\text{75,000 to 99,999}}$, $x_{\text{Less than 25,000}}$ and $x_{\text{unknown income}}$ represent indicator variables which identify if the respondent's household income is: More than \$125,000, \$25,000 to \$49,999, \$50,000 to \$74,999, \$75,000 to \$99,999, Less than \$25,000 or unknown respectively. Therefore, $x_i = 1$ where $i \in$ {More than 125,000, 25,000 to 49,999, 50,000 to 74,999, 75,000 to 99,999, Less than 25,000, unknown income}, represents the respondent's household income being in the interval i, and $x_i = 0$ if the respondent's household income is not in the interval i. If the respondent's household income is in the range \$100,000 to \$149,999 then $x_i = 0$ for all income ranges i. $\beta_{\text{More than 125,000}}$, $\beta_{\text{25,000 to 49,999}}$, $\beta_{\text{50,000 to 74,999}}$, $\beta_{\text{75,000 to 99,999}}$, $\beta_{\text{Less than 25,000}}$ and $\beta_{\text{unknown income}}$ represent the respective change in the log-odds of voting for the particular party, relative to if the respondent's household income was in the \$100,000 - \$124,999 interval, holding all else constant.

Birth Place: The birth location of the respondent is a categorical variable with two distinctions: born in Canada, or born outside of Canada. For our model, being born in Canada is chosen as the reference category. Therefore, $x_{\text{born outside Canada}}$ is an indicator variable where $x_{\text{born outside Canada}} = 1$ if the respondent is born in any country other than Canada, and $x_{\text{born outside Canada}} = 0$ if the respondent is born in Canada. $\beta_{\text{born outside Canada}}$ represents the change in the log-odds of voting for the particular party of interest, relative to if the individual was born in Canada and holding all else constant.

Education: The highest education level of the respondent is a categorical variable with six distinctions. For this variable, the reference category is that the respondent's highest education level is a bachelor's degree. Therefore, $x_{\text{non-uni certificate}}$, $x_{\text{uni certificate below bachelor's degree}}$, $x_{\text{high school diploma}}$, $x_{\text{less than high school}}$, $x_{\text{above bachelor's degree}}$ represent indicator variables which identify if the respondent's highest education level is a non-university certificate, a university certificate below a bachelor's degree, a high school diploma, less than a high school diploma, or a degree above the bachelor's level respectively. Therefore, $x_i = 1$ where $i \in$ {non-uni certificate, uni certificate below bachelor's degree, high school diploma, less than high school, above bachelor's degree} represents the highest level of education completed by the respondent, i, with the remaining variables identically equal to zero. If the respondent's highest education level is a bachelor's degree, then $x_i = 0$ for all education levels i. Finally, $\beta_{\text{non-uni certificate}}$, $\beta_{\text{uni certificate below bachelor's degree}}$, $\beta_{\text{high school diploma}}$, $\beta_{\text{less than high school}}$, $\beta_{\text{above bachelor's degree}}$ represent the respective change in the log-odds of voting for the particular party of interest, relative to if the respondent's highest education level was a bachelor's degree, holding all else constant.

Variable Selection

In order to process the gender of the individual, and to accurately complete the post-stratification method outlined below, the gender variable needed to be imputed. The survey data includes a wider scale of gender options, compared to the binary sex options for that of the census data. Therefore, two methods will be performed: the first is to impute all non-binary genders to male, and the second is to impute all non- binary

genders to female. Model selection is performed on the two methods within each of the seven models identified above, and the imputation which results in the highest adjusted R-squared for the majority of the parties will be selected for the final model.

AIC: With gender imputed to its appropriate male/female distribution, we now can determine the most appropriate model for estimating the probability of a respondent voting for each party. Each multiple logistic regression model will have six identical predictors: age, sex, province of residence, individual income, birthplace, and education level.

Our goal in performing variable selection is to be rid of insignificant variables that are not useful to the overall prediction. To do this, we will implement backward AIC. Backward AIC is the combination of backward selection with the Akaike information criterion. It is the statistical method that examines a model using all predictors under consideration, and determines which, if any, of the statistically insignificant predictors should be removed [10]. In each step of the algorithm, it considers the effect of removing each predictor independently from the full model by comparing the AIC scores. Since AIC is a measure of how much the model is not fitting, it is ideal to have a model with the lowest AIC score. That is, the lower the score, the fewer data that the model is not fitting. Continue the algorithm until the remaining set of predictors all have an AIC score that cannot be minimized further by removing predictors [10]. If all variables initially meet this threshold, then none are removed.

Likelihood Ratio Test: The Likelihood Ratio Test allows us to test whether a simple model is sufficient enough for model accuracy or a more complex model is preferred. A simple model is usually a nested model of the complex one. In other words, all of the predictors in the simple model are in the complex one, but the complex model contains some other predictors that are not in the simple model. By using the likelihood ratio test, essentially, we are testing if those extra variables are necessary.

Our hypothesis is that the simple model is good enough for the purpose of accuracy and prediction. We will use a threshold of 0.05. If the p-value of a likelihood ratio test is below 0.05, we have sufficient, strong evidence that the simple model is not good enough and we should use the complex model, and vice versa.

The Likelihood ratio test is used alongside AIC as a second criterion for variable selection. If we have doubts about the results of the AIC, we will implement the test to double-check and make our decision wisely.

Post-Stratification

An important part of the analysis and model is the post-stratification process conducted between the survey data and the census data. To perform this, we post-stratify the survey data to develop a representative sample of the entire population, which for our purposes is the census data. Post-stratification is the statistical technique that involves weighing mutually exclusive groups based on their proportion of the overall population [6]. This is a key method to determine the impact which underrepresented groups have on the outcome of a study, and can correct for differences between the two populations. This approach is often appropriate when a simple random sample is not properly balanced by the representation, and thus we assume that the survey data displays a better representation of the overall population [7]. In this instance, the need for post-stratification arises due to the nature of how the telephone survey was conducted. With this combination, there is the potential for sampling bias, as it is likely that those who are willing to answer the phone and provide information towards which party which they intend to vote for share a common demographic. To overcome this, we use the survey data to adjust the weights in the census data under the variables age, gender, income, province of residency, birthplace, and education level.

The bins are split by the variables identified above with the following categories/distinctions: sex (male, female), age (18 to 80), income(CAD) (Less than 25,000, 25,000 to 49,999, 50,000 to 74,999, 75,000 to 99,999, 100,000 to 124,999, 125,000 and more), province of residency (Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec, Saskatchewan), birthplace (Born in Canada, Born outside Canada) and highest education level completed ("Less than high school diploma or its equivalent", "High school diploma or a high school equivalency certificate", "Completed or Some Technical, College, CEGEP or other non-university certificate or di...", "Completed or Some University certificate or diploma below the bachelor's level", "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)",

"University certificate, diploma or degree above the bach. . . "). Once the model selection process is conducted and the set of reliable predictors are chosen, the final number of bins used for post-stratification will be all combinations of the factor levels for each predictor. Therefore, the final post-stratification model will have the following form:

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

In the above model, $\hat{y}_i$ represents the estimate of each cell under consideration, and N_j represents the population size of the j^{th} cell based on the census data. Therefore, $\hat{y}^{PS}$ consists of a distinct sum of all cells, factoring in their respective proportion of the population.

In summary, the survey data will initially be partitioned into cells of all different demographic types for our variables of interest. The estimated intent of the voter will be formed using the weights of each cell and a multiple logistic regression model. Variable selection is performed to determine whether those who identify as non-binary should be imputed to male or female, and then performed again in order to determine the set of reliable predictors for the final model. Finally, we will aggregate the estimates of each cell to the census population to determine the estimated outcome of the next election. Since these estimates will be in the form of log-odds, they will then be converted respectively so that it is in the form of a probability.

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

Results

Sex vs. Gender

As aforementioned, in the census data, sex is a binary variable: male and female, whereas in the sample data, gender can take on values: male, female, and non-binary. Specifically, the only non-binary response in our data is transgender. To deal with this dilemma of sex and gender, our options are: 1) impute all non-binary respondents as male, 2) impute all non-binary respondents as female, 3) remove all non-binary respondents. Since we hypothesize that there may be a difference in people's political preferences depending on their gender, removing all non-binary responses is unjustified and may significantly decrease the accuracy of our prediction. Hence, we are left with Option 1 and 2.

We first fit 7 models using a data set that imputes all transgender respondents as male. Each model has a response variable of a different party and we compute the corresponding Residual Deviance/Pseudo "R^2" values. The Residual Deviance explains how well the variance of the y variable is explained by all x variables. Thus, a lower Residual Deviance indicates a better-fitted model. Then, in a similar fashion, we fit 7 models using a data set that imputes all transgender respondents as female. Again, we have each model's Residual Deviance values. We compare the Residual Deviance of the two models that use different data sets but have the same response variable, and we select the one with a lower Residual Deviance.

Among all 7 comparisons, 4 of the selected models imputes transgender respondents as female and 3 imputes them as male. Taking the majority, we will impute non-binary gender responses as female for all the models.

Variable Selection

AIC: Upon using backward AIC on each model, we found that age, province are both kept as variables in 5 out of 7 of the models. In 4 of the models, income is kept whereas birth place and education level are kept in 3 of the models. To our surprise, sex only kept in 2 of the 7 models.

By taking the majority again, we remove the variables that are kept in less than 4 of the models, i.e. birth place, education, sex. However, despite the results given by AIC, we still speculate that sex is a somewhat useful predictor and we would like to further investigate this variable. At this stage, we will only remove two variables, birth place and education level, and keep sex in the model.

Likelihood Ratio Test: From the conclusions we made using AIC, we narrowed down our predictors to age, province, income, sex. To further check if sex should be kept, we propose two models: a complex model

uses all four predictors and a simple model uses all four but sex. Again, we fit 14 models where each of the 7 political parties have both a complex and a simple model.

We run 7 likelihood ratio tests where each of them compares the complex and simple model of the same political party and we obtain 7 p-values. We found that all p-values we got from the tests are far above 0.05, which indicates that we have insufficient evidence to reject the null hypothesis. Again, we take the majority; all of the 7 results agree that the simpler model is favored. Therefore, sex is not a significant predictor and we will remove it in our final model.

As a result of the model and variable selection process, our final model comes out as:

$$\begin{aligned}\hat{y}_{\text{party}} = \beta_0 &+ \beta_{\text{age}} x_{\text{age}} \\ &+ \beta_{\text{BC}} x_{\text{BC}} + \beta_{\text{MB}} x_{\text{MB}} + \beta_{\text{NB}} x_{\text{NB}} + \beta_{\text{NL}} x_{\text{NL}} + \beta_{\text{NS}} x_{\text{NS}} \\ &+ \beta_{\text{ON}} x_{\text{ON}} + \beta_{\text{PE}} x_{\text{PE}} + \beta_{\text{QC}} x_{\text{QC}} + \beta_{\text{SK}} x_{\text{SK}} \\ &+ \beta_{\text{More than 125,000}} x_{\text{More than 125,000}} + \beta_{\text{25,000 to 49,999}} x_{\text{25,000 to 49,999}} \\ &+ \beta_{\text{50,000 to 74,999}} x_{\text{50,000 to 74,999}} + \beta_{\text{75,000 to 99,999}} x_{\text{75,000 to 99,999}} \\ &+ \beta_{\text{Less than 25,000}} x_{\text{Less than 25,000}} + \beta_{\text{unknown income}} x_{\text{unknown income}} + \epsilon\end{aligned}$$

Post-Stratification

With our final model, we conduct post-stratification by applying the census data to our final model. This gets us the following probabilities:

Table 4: Post-stratified Predicted Probabilities of Voting for Each Party

Liberal Party of Canada	Conservative Party of Canada	New Democratic Party	Bloc Québé-cois	Green Party of Canada	People's Party of Canada	Other Political Parties
0.2500184	0.2570117	0.1099116	0.02867652	0.07148674	0.009756855	0.004959796

As seen in Table 4, our final model with post-stratification has determined that **the Conservative Party of Canada has the highest probability of being the 45th Canadian Federal Election Popular Vote, with a 25.7% chance of the greater population voting for the Conservative Party of Canada.** Trailing behind the CPC by only 0.7% chance is the Liberal Party, with a 25% chance of the overall population voting for them in the popular vote. As stated in our introduction, these results seem in line with prior results, with the 43rd and 44th federal elections both resulting in a Conservative win in the overall popular vote.

Conclusions

There are several key results in our analysis that we will reiterate. We hypothesized that the Conservative Party would win the popular vote based on recent elections. While cleaning our dataset and categorizing the relevant variables, we encountered difficulty in determining how to map the gender data in our survey dataset (which has male, female, and non-binary) to our sex variable in our census dataset (which only has male and female). Given that the post-stratification process requires variables in both datasets to be identical, we are left to decide what to do with the non-binary datapoints.

We determined that the best approach to the problem would be to impute non-binary to both female and male datapoints, then construct 14 models, with 7 of them being non-binary to female, and the other 7 being non-binary to male. Using the "R^2" approach described in the Results section, we eliminate the 7 models we had with non-binary imputed to male, which leaves us the 7 models with non-binary imputed to female.

In spite of this result, the results of an AIC test determined that three variables were not strong predictors for likelihood: birthplace, education, and gender. We eliminated birthplace and gender but attempted to conduct another test on the resulting model to see if gender is as weak a predictor as the AIC made it out to be. The results of a Likelihood Ratio Test ultimately removed gender from our final model, as the test dictated that models with and without the gender variable had very similar outputs.

Using this result, our final model had only three predictors: income, province, and age. Running our census data through this model eventually gave us strong evidence in support of our hypothesis.

One thing to note is that the closeness in which the Conservative and Liberal parties differ in the likelihood of receiving a vote in the overall popular vote is also reflected in the percentage total of popular votes that both parties accumulated within the last two elections. More specifically, Liberals trailed Conservatives by 1.22% in total popular votes in the 43rd election, and by 1.12% in the 44th election [9]. This may be intuitively explained by the fact that a majority of people tend to vote for the party that they say they will vote for prior to an election. However, a separate study should be conducted to rigorously evaluate such a claim.

Weaknesses and Limitations

Despite the work done within this report, from a practical standpoint, the popular vote serves very little purpose in a full election. While there may be some correlation between the popular vote and the probability of winning the election, we cannot conclusively say there exists one as we would require datasets from all 338 ridings in Canada (and the resulting models could get very complicated). We are ultimately limited by the amount of data we have and the complexity level of our report.

Another weakness within our report is a potential bias that may appear in the sampling survey. Our survey data presumes that respondents maintain their political opinions 4 years into the future. Due to the ever-changing political climate in Canada, we would likely obtain more accurate results should we have survey data for the 44th federal elections (which took place on September 2021), and census data some arbitrary time closer to the date of the election.

In our data cleaning process, we also made the critical assumption that only people born in Canada were in our sample population of respondents. According to Bybee and McCrae, nearly 20% of Canadian residents were not born in the 2000s, where a resident is defined to have the status of Permanent Resident or Full Citizen [12]. As we do not know the proportion of this 20% that is a PR or FC, this would loosely imply that we are missing $<20\%$ or so of our potential data points, which could inhibit the accuracy of our models.

As mentioned in the paper by Kennedy et al. [3], there is no perfect approach when dealing with the sex and gender mismatch in the two datasets. We imputed all non-binary respondents to female. This decision is violating the respondents' rights because they explicitly identified themselves as non-binary and we changed their responses by imputing it to female, which they consciously did not select in the survey. Plus, this decision will lead to some implicit error in our model that cannot be easily eliminated nor identified. Furthermore, looking at the big picture, we are more interested in the variable gender because we speculate that the gender one identifies with is likely to affect their political choice, rather than a biological component.

Next Steps

There are three primary directions that future papers may expand on, which we discussed in this report. A future study would benefit to look into the correlation between the popular vote and the likelihood of a party winning a majority of seats in the House (and thus winning the election). The results of such a study could help bring into question the necessity of having such a measure in the election process.

A second direction that may be pursued is the relationship between the probability of winning the popular vote for a party, and the total percentage of popular votes won per party in previous elections. As stated in the conclusion, prior elections seem to point at a relationship between how close a win is between the popular votes of leading parties, and how probable it is for a party to win the popular vote based on data prior to the election. A study that pursues such a direction would benefit from historical survey and census data dating back further than 2019. Finally, the gender-sex mismatch may be addressed with a separate study on the percentage of people that associate their biological sex with their gender identity. The results of such a

study could help us draw conclusions on the effect of sex on gender, and consequently the effect of gender on political orientation. This may affect the accuracy of our model, as should the study conclude that a very large proportion of people base their gender identity on their biological sex, then non-binary data points would have a lesser impact on our model accuracy.

Bibliography

1. Stephenson, L. B., Harell, A., Rubenson, D., & Loewen, P. J. (2020, May 1). *2019 Canadian Election Study - Phone Survey.* Harvard Dataverse. Retrieved October 28, 2021, from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2F8RHLG1.

2. Stephenson, L., Harell, A., Rubenson, D., & Loewen, P. (2021). *Measuring Preferences and Behaviours in the 2019 Canadian Election Study.* Canadian Journal of Political Science, 54(1), 118-124. doi: 10.1017/S0008423920001006

3. Kennedy, L., Khanna, L., Simpson, D., Gelman, A. (2020, October 1). *Using sex and gender in survey adjustment.* Retrieved November 3, 2021

4. Government of Canada, S. C. (2019, February 6). *General Social Survey - Family (GSS).* Surveys and statistical programs. Retrieved November 1, 2021, from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501.

5. Workopolis. (2014, October 12). *Here's how much Canadians are earning by Province.* Workopolis Blog. Retrieved November 3, 2021, from https://careers.workopolis.com/advice/how-much-canadians-are-earning-by-province/.

6. Poststratification. *Poststratification - an overview | ScienceDirect Topics.* (n.d.). Retrieved November 3, 2021, from https://www.sciencedirect.com/topics/mathematics/poststratification.

7. *6.3 - poststratification and further topics on stratification: Stat 506.* PennState: Statistics Online Courses. (n.d.). Retrieved November 3, 2021, from https://online.stat.psu.edu/stat506/lesson/6/6.3.

8. CBC/Radio Canada. (n.d.). *Federal election 2019 live results.* CBCnews. Retrieved November 3, 2021, from https://newsinteractives.cbc.ca/elections/federal/2019/results/.

9. Siekierska, A. (n.d.). *Conservatives and Erin O'Toole didn't lose completely: Party wins popular vote for second election in a row.* Yahoo! Retrieved November 3, 2021, from https://ca.style.yahoo.com/erin-otoole-conservatives-popular-vote-canada-election-054033279.html.

10. Choueiry, G. (n.d.). *George Choueiry.* Quantifying Health. Retrieved November 3, 2021, from https://quantifyinghealth.com/stepwise-selection/.

11. Government of Canada, S. C. (2018, September 17) *Table 8 Abbreviations and codes for provinces and territories, 2011 Census.* Retrieved November 5, 2021, from https://www150.statcan.gc.ca/n1/pub/92-195-x/2011001/geo/prov/tbl/tbl8-eng.htm

12. Google. (n.d.). *Pisa Science 2006.* Google Books. Retrieved November 5, 2021, from https://books.google.ca/books?id=iZ3zu2130AUC&pg=PA92&redir_esc=y#v=onepage&q&f=false.

Appendix

Data

Here is a short glimpse of the original datasets:

```
## # A tibble: 6 x 278
##   sample_id survey_end_CES      survey_end_month_CES survey_end_day_CES
##       <dbl> <dttm>                             <dbl>              <dbl>
## 1        18 2019-09-23 21:48:29                    9                 23
## 2        32 2019-09-13 00:02:30                    9                 12
## 3        39 2019-09-11 00:02:33                    9                 10
## 4        59 2019-10-10 21:20:13                   10                 10
## 5        61 2019-09-12 22:28:58                    9                 12
## 6        69 2019-09-17 23:56:26                    9                 17
## # ... with 274 more variables: num_attempts_CES <dbl>,
## #   interviewer_id_CES <dbl>, interviewer_gender_CES <chr>, language_CES <dbl>,
## #   phonetype_CES <dbl>, survey_end_PES <dttm>, survey_end_month_PES <dbl>,
## #   survey_end_day_PES <dbl>, num_attempts_PES <dbl>, interviewer_id_PES <dbl>,
## #   interviewer_gender_PES <chr>, language_PES <dbl>, phonetype_PES <dbl>,
## #   mode_PES <dbl>, phone_type <dbl>, weight_CES <dbl>, weight_PES <dbl>,
## #   c1 <dbl>, c2a <dbl>, c3 <dbl>, q1 <dbl>, q2 <dbl>, q3 <dbl>, q4 <dbl>,
## #   q6 <dbl>, q7 <chr>, q72 <chr>, q73 <chr>, q74 <chr>, q8 <dbl>, q8_70 <chr>,
## #   q9 <dbl>, q10 <dbl>, q11 <dbl>, q11_70 <chr>, q12 <dbl>, q12_70 <chr>,
## #   q13 <dbl>, q14 <dbl>, q15 <dbl>, q16 <dbl>, q17 <dbl>, q18 <dbl>,
## #   q19 <dbl>, q20 <dbl>, q21 <dbl>, q22 <dbl>, q23 <dbl>, q24 <dbl>,
## #   q25 <dbl>, q27_a <dbl>, q27_b <dbl>, q27_c <dbl>, q27_d <dbl>, q27_e <dbl>,
## #   q31 <dbl>, q32 <dbl>, q33 <dbl>, q33_70 <chr>, q34 <dbl>, q34_70 <chr>,
## #   q35 <dbl>, q35_70 <chr>, q36 <dbl>, q36_70 <chr>, q37 <dbl>, q37_70 <chr>,
## #   q38 <dbl>, q38_70 <chr>, q39 <dbl>, q40 <dbl>, q75 <dbl>, q44 <dbl>,
## #   q76 <dbl>, q45 <dbl>, q46 <dbl>, q47 <dbl>, q48 <dbl>, q49 <dbl>,
## #   q52 <dbl>, q52_70 <chr>, q53 <dbl>, q54 <dbl>, q59 <dbl>, q60 <dbl>,
## #   q60_70 <chr>, q77 <dbl>, q43 <dbl>, q61 <dbl>, q62 <dbl>, q62_220 <chr>,
## #   q63 <dbl>, q64 <dbl>, q64_130 <chr>, q65 <dbl>, q66a_1 <dbl>, q66a_2 <dbl>,
## #   q66a_3 <dbl>, q66a_4 <dbl>, q66a_5 <dbl>, ...
```

```
## # A tibble: 6 x 81
##   caseid   age age_first_child age_youngest_chi~ total_children age_start_relat~
##    <dbl> <dbl>           <dbl>              <dbl>          <dbl>            <dbl>
## 1      1  52.7              27                 NA              1               NA
## 2      2  51.1              33                 NA              5               NA
## 3      3  63.6              40                 NA              5               NA
## 4      4  80                56                 NA              1               NA
## 5      5  28                NA                 NA              0             25.3
## 6      6  63                37                 NA              2               NA
## # ... with 75 more variables: age_at_first_marriage <dbl>,
## #   age_at_first_birth <dbl>, distance_between_houses <dbl>,
## #   age_youngest_child_returned_work <dbl>, feelings_life <dbl>, sex <chr>,
## #   place_birth_canada <chr>, place_birth_father <chr>,
## #   place_birth_mother <chr>, place_birth_macro_region <chr>,
## #   place_birth_province <chr>, year_arrived_canada <chr>, province <chr>,
## #   region <chr>, pop_center <chr>, marital_status <chr>, aboriginal <chr>,
## #   vis_minority <chr>, age_immigration <chr>, landed_immigrant <chr>,
## #   citizenship_status <chr>, education <chr>, own_rent <chr>,
## #   living_arrangement <chr>, hh_type <chr>, hh_size <dbl>,
## #   partner_birth_country <chr>, partner_birth_province <chr>,
```

```
## #   partner_vis_minority <chr>, partner_sex <chr>, partner_education <chr>,
## #   average_hours_worked <chr>, worked_last_week <chr>,
## #   partner_main_activity <chr>, self_rated_health <chr>,
## #   self_rated_mental_health <chr>, religion_has_affiliation <chr>,
## #   regilion_importance <chr>, language_home <chr>, language_knowledge <chr>,
## #   income_family <chr>, income_respondent <chr>, occupation <chr>,
## #   childcare_regular <chr>, childcare_type <chr>,
## #   childcare_monthly_cost <chr>, ever_fathered_child <chr>,
## #   ever_given_birth <chr>, number_of_current_union <chr>,
## #   lives_with_partner <chr>, children_in_household <chr>,
## #   number_total_children_intention <dbl>, has_grandchildren <chr>,
## #   grandparents_still_living <chr>, ever_married <chr>,
## #   current_marriage_is_first <chr>, number_marriages <dbl>,
## #   religion_participation <chr>, partner_location_residence <chr>,
## #   full_part_time_work <chr>, time_off_work_birth <chr>,
## #   reason_no_time_off_birth <chr>, returned_same_job <chr>,
## #   satisfied_time_children <chr>, provide_or_receive_fin_supp <chr>,
## #   fin_supp_child_supp <dbl>, fin_supp_child_exp <dbl>, fin_supp_lump <dbl>,
## #   fin_supp_other <dbl>, fin_supp_agreement <chr>,
## #   future_children_intention <chr>, is_male <dbl>, main_activity <lgl>,
## #   age_diff <chr>, number_total_children_known <dbl>
```

Here is a glimpse of the cleaned survey data to use in the model:

```
## Rows: 3,908
## Columns: 13
## $ age               <dbl> 56, 46, 25, 19, 35, 80, 20, 24, 56, 49, 41, 20, 56, ~
## $ gender            <chr> "Female", "Male", "Female", "Female", "Male", "Femal~
## $ province          <chr> "Quebec", "Quebec", "Quebec", "Quebec", "Quebec", "Q~
## $ income            <chr> "$100,000 to $ 124,999", "$75,000 to $99,999", "Less~
## $ birth_place       <chr> "Born in Canada", "Born in Canada", "Born in Canada"~
## $ education         <chr> "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)", "Comp~
## $ vote_liberal      <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1~
## $ vote_conservative <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0~
## $ vote_ndp          <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ vote_bq           <dbl> 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ vote_green        <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ vote_peoples      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
## $ vote_other        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0~
```

Here is a glimpse of the cleaned census data to use in the post-stratification:

```
## Rows: 20,161
## Columns: 6
## $ age         <dbl> 52.7, 51.1, 63.6, 80.0, 28.0, 63.0, 58.8, 80.0, 63.8, 25.2~
## $ sex         <chr> "Female", "Male", "Female", "Female", "Male", "Female", "F~
## $ province    <chr> "Quebec", "Manitoba", "Ontario", "Alberta", "Quebec", "Que~
## $ income      <chr> "$25,000 to $49,999", "$75,000 to $99,999", "$75,000 to $9~
## $ birth_place <chr> "Born in Canada", "Born in Canada", "Born in Canada", "Bor~
## $ education   <chr> "High school diploma or a high school equivalency certific~
```

Methods

The following provincial codes are sourced from [11.]

Abbreviation	Province
AB	Alberta
BC	British Columbia
MB	Manitoba
NB	New Brunswick
NL	Newfoundland and Labrador
NS	Nova Scotia
ON	Ontario
PE	Prince Edward Island
QC	Quebec
SK	Saskachewan