Predicting 2025 Canadian Federal Election Situation for the Liberal Party

Lingjun Meng, Fengbai Han, Di Zhang

November 5, 2021

Introduction

Under democratic election, the party that governs a country is not only in power, bringing different benefits and influences to a country, but also reflecting the demands of society and representing the actual needs of the people of that country. As a result, it is important to study the election of Canada and how the votes are spread in different groups of people for a better life for people. The party that we are interested in for this study is the Liberal party because it is one of the major parties in the country since the establishment of the Dominion of Canada in 1867 and has been the governing party at the federal level for most of the period [7]. The Liberals' policies include supportive of social welfare spending, universal health care, Canada Student Loans, gun control, peacekeeping and legalizing same-sex marriage [8]. For studying the votes for the Liberal party, we would have an idea of the division of votes among different groups of people and how likely that the party would win in the next election.

For this report, a study would be conducted by using the "2019 CES phone survey data". The data set provides information about recorded phone survey answers of people in Canada for the election in 2019. It contains 4021 observations and 278 variables, including general information of the participants as well as their voting results in 2015 and voting tendencies in 2019. However, due to the difference between survey data and the actual Canadian population attribute, another "General Social Survey data 2017" is generated for helping a better understanding of the target population. The GSS data contains 20602 observations and records key information that could match the CES data.

According to research, there are several factors that would affect voters' behavior, such as geography, sex, race, social class and religion [9]. Hence, with both data sets, the goal of the study is to predict the probability of Liberal Party winning the next Canadian Election tentatively in 2025. Thus, the research question of the study would be if the variables in the data set have a relationship with the voting to the Liberal party and we can use the model we get to make predictions. **Specifically, if there is any association between votes for the Liberal party with age, sex, education level, province, language and family income.**

For this study, we will make a logistic regression model that represents the relationship between votes for the Liberal party with age, sex, education level, province, language and family income. The hypotheses is that there is an association between votes for the Liberal party with age, sex, education level, province, language and family income. If the hypotheses holds, we can predict if the Liberal Party would win the Canadian Election in 2025.

Data

Data Collection Process

We used 2017 General Social Survey: Families Cycle 31 (hereinafter called census data)[18] and 2019 Canadian Election Study - Phone Survey (hereinafter called survey data)[10] as our data sets.

We would like to first introduce the census data. The 2017 GSS is a sample survey with a cross-sectional design that was performed from February 2nd to November 30th 2017. The target demographic consists of all non-institutionalized people aged 15 and up who live in Canada's ten provinces. [18]

The target population is all Canadians aged 15 and up, with the exception of residents of the Yukon, Northwest Territories, and Nunavut, as well as full-time residents of institutions. [18]

For the collection process, the poll collects data over the phone using a new frame introduced in 2013, which integrates telephone numbers (both landline and cellular) with Statistics Canada's Address Register. Both sampling and non-sampling errors can affect data. Computer-assisted telephone interviews were used to collect data for the 2017 GSS (CATI). Respondents were interviewed in their preferred official language. Interviews by proxy were not permitted. All interviews were conducted utilizing centralized telephone facilities in five of Statistics Canada's regional offices, with calls made Monday through Friday between 9:00 a.m. and 9:30 p.m.

On Saturdays, 10:00 a.m. to 5:00 p.m., and Sundays, 1:00 p.m. to 9:00 p.m., interviews were held. Halifax, Sherbrooke, Sturgeon Falls, Winnipeg, and Edmonton were the five regional offices. [18] Interviewers were taught in telephone interviewing techniques using CATI, as well as survey concepts and processes, by Statistics Canada professionals.

The vast majority of interviewers had previously conducted interviews for GSS cycles. [18] Interviewers were advised to make every reasonable effort to secure a completed interview from a randomly selected household member.

Those who initially declined to participate were called up to two more times to explain the value of the survey and persuade them to do so. When the interviewer's call came at an inconvenient moment, an appointment was made to call again at a more convenient time. Numerous callbacks were made in circumstances where no one was at home. [18]

Here is the introduction for the survey data collection process. The data was collected in two stages. Telephone interviews with 4,021 Canadian people were conducted throughout the election campaign. After the election, all respondents to the Campaign-Period Survey (CPS) were phoned or emailed, depending on their preference, and asked to take the Post-Election Survey (PES), which 2,889 (72%) did, with 2,067 (72%) doing so over the phone and 822 (28%) doing so online. [10] The CPS used a modified random digit dialling (RDD) process to obtain telephone numbers, and the respondent was identified using the birthday selection method for landline samples in households with more than one adult Canadian citizen. Interviewing began on September 10, 2019 (after the election was called) and ended on October 20, 2019, on the eve of the election.

With the exception of September 11, 2019 (to allow the team to review the survey results from the soft launch), September 15, 2019, and Thanksgiving Day (October 14, 2019), interviews were conducted every day. [10] Respondents were asked at the end of the CPS whether they would prefer to complete the PES by phone or online, and those who said they would prefer to finish the survey online were given an email address. The PES registration process began the day after the election, on October 22, 2019, and the first email invitations were sent out on October 24, 2019, the third day following the election.

Within 10 days of the vote, all CPS respondents were contacted. Although not all respondents were accessible when first contacted, more than half of the PES interviews were completed eight days following the election.

(Completing this phase of the PES interviews took nine days in 2015 and 14 days in 2011.) All of the phone calls and emails were completed within 31 days of the election, with the final interviews taking place on November 21, 2019. [10]

For the target population part, the sample for the 2019 Canadian Election Study was designed to represent the adult population of Canada, which is defined as Canadian citizens aged 18 and up who live in one of the

ten provinces (thus excluding the territories).

The small fraction of Canadian households without landline or wireless telephones were omitted from the sample population because the initial survey (the CPS) was done via telephone. [10]

Data Summary

"General Social Survey data 2017":

Firstly, we load the data from gss_cleaning.csv, which has already been processed by gss_cleaning.R. We selected columns including: citizenship status, age, sex, province, education, income of family and language of home to form a data frame for us to clean.

We first looked at the citizenship status, there are four categories of elements in the column: "By birth", "By naturalization", "Don't know" and NA. We replaced the "Don't know" elements with NA. Then, we replaced all of the blank elements (which are equal to "") with NA as well. After that, we kept all the rows that are not"NA" in the citizenship status. We also omit the NAs in the data frame.

After that, we start to map the education levels into a binary variable. When education level is "High school diploma or a high school equivalency certificate" or "Less than high school diploma or its equivalent", it will be recorded as 'No school to high school' Level. We take other education levels as 'College to professional degree' because NAs has already been filtered.

For age, we mapped it into age groups. When age is below or equal to 20, we marked it as '20 or less'. When age is greater than 20 but below or equal to 35, we marked it as '21 to 35'. When age is greater than 35 but below or equal to 50, we marked it as '35 to 50'. When age is greater than 50 but below or equal to 65, we marked it as '50 to 65'. When age is greater than 65, we marked it as 'above 65'.

We also cleaned language of home variable by replacing language other than English or French with "Other" because we focus more on people who speaks English or French.

Finally, we selected province, education variable after process, age groups after process, household income, language variable after process and sex. We renamed province variable with 'Province', education variable with 'Education', age groups variable with 'Age', household income variable with 'Income', sex variable with 'Sex'.

<Here is a resource for grabbing the CES2019 data: https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-canadian-election-study>

"2019 CES phone survey data":

Firstly, we loaded the 2019 phone survey data from ces2019-phone_clean.csv. We selected following columns to match the variables in the census data: 'sample_id', 'age', 'interviewer_gender_CES', 'q61', 'q4', 'q69', 'q67', 'q1', 'q10', 'q11'. Although the column names are vague, we studied the phone survey questionnaire and replaced the column names[10].

We replaced 'sample_id' with id, 'age' with age, 'interviewer_gender_CES' with gender, 'q61' with education, 'q4' with province, 'q69' with income, 'q67' with language, 'q1' with citizenship, 'q10' with intention, 'q11' with vote.

By observing the recorded values in the survey [10], negative values represents "Don't know", "Refused to answer" or other useless values. So we filtered the data with the conditions of age, province, education, income, language and vote are greater than or equal to 0.

Because we are focusing on the winning possibility of Liberal Party, we mutated vote variable into a binary variable. If the value of vote equals 1, which means the person votes for Liberal Party, in the new variable Vote_Liberal will be recorded as "Yes". If the value is not 1, in Vote_Liberal the row will be recorded as "No".

In province cleaning, in the original data, province stands for the province participants are currently living, which is different from the province of birth in the census data. Here, we assume that province participants are currently living equals to the province of birth.(See limitations) For the cleaning process, the value equals 1 will be replaced by "Newfoundland and Labrador", the value equals 2 will be replaced by "Prince Edward Island", the value equals 3 will be replaced by "Nova Scotia", the value equals 4 will be replaced by "New Brunswick", the value equals 5 will be replaced by "Quebec", the value equals 6 will be replaced by "Ontario", the value equals 7 will be replaced by "Manitoba", the value equals 8 will be replaced by "Saskatchewan", the value equals 9 will be replaced by "Alberta", the value equals 10 will be replaced by "British Columbia", the value equals 11 will be replaced by "Northwest Territories", the value equals 12 will be replaced by "Yukon", the value equals 13 will be replaced by "Nunavut" and other numbers will be replaced by NA.

For education, value greater than or equal to 1 and less than or equal to 5 will be recorded as 'No school to high school' in the newly created variable education_binary. Education value greater than 5 and less than or equal to 11 will be recorded as 'No school to high school' in education_binary. In education cleaning, the value equals 1 will be replaced by "No schooling", the value equals 2 will be replaced by "Some elementary school", the value equals 3 will be replaced by "Completed elementary school", the value equals 4 will be replaced by "Some secondary / high school", the value equals 5 will be replaced by "Completed secondary / high school", the value equals 6 will be replaced by "Some technical, community college, CEGEP, College Classique", the value equals 7 will be replaced by "Completed technical, community college, CEGEP, College Classique", the value equals 8 will be replaced by "Some university", the value equals 9 will be replaced by "Bachelor's degree", the value equals 10 will be replaced by "Master's degree", the value equals 11 will be replaced by "Professional degree or doctorate", and other numbers will be replaced by NA. For age, we divided it into age groups.

For Age variable, when age is below or equal to 20, we marked it as '20 or less'. When age is greater than 20 but below or equal to 35, we marked it as '21 to 35'. When age is greater than 35 but below or equal to 50, we marked it as '35 to 50'. When age is greater than 50 but below or equal to 65, we marked it as '50 to 65'. When age is greater than 65, we marked it as 'above 65'.

For household income, we also divided it into income groups. When income is below 25000, we marked it as 'Less than $25,000'. When income is greater than or equal to 25000 but below 50000, we marked it as '$25,000 to $49,999'. When income is greater than or equal to 50000 but below 75000, we marked it as '$50,000 to $74,999'. When income is greater than or equal to 75000 but below 100000, we marked it as '$75,000 to $99,999'. When income is greater than or equal to 100000 but below 125000, we marked it as '$100,000 to $ 124,999'. When income is greater than or equal to 125000, we marked it as '$125,000 and more'.

We also created a new language group variable by replacing value 1 with English and value 4 with French and values other than English or French with "Other" because we focus more on people who speaks English or French.

For citizenship variable, we cleaned it by replacing value 1 with "Yes", value 2 with "No", and replaced others with "NA". For intention variable, we cleaned it by replacing value 1 with "Certain", value 2 with "Likely", value 3 with "Unlikely", value 4 with "Certain not to vote", value 5 with "Already voted in advanced poll" and replaced others with "NA". We replaced blank values which are equal to "" with "NA" and then we omit all NAs in the cleaned data frame.

There are three genders in gender column including male, female and transgender. We would like to replace it with biological sex. What we did is we calculated the proportion of male and female without transgender, and assign people who are transgender randomly using for loop and the proportion of male and female. The random seed we used here is 1005712579 (see Limitation part for more details of mapping gender to Sex).

Finally, we filtered the eligible voter using the data collected. We considered that people who are citizen, age is greater than or equal to 18 are able to vote. We also filtered people with negative voting intentions: people answering "Unlikely" and "Certain not to vote" are also excluded. We created survey data final data frame to conclude the parts we need.

We included Vote Liberal, education binary, age group, province, simulated sex, income group and language group. We renamed education binary with Education, age group with Age, province with Province, simulated

sex with Sex, income group with Income and language group with Language.

Discription of important variables in the data:

There are 6 categorical variables selected in the census data and 7 categorical variables selected in the survey data, and both data frames shared 6 of the variables.

We will first start with the shared variables with the same categories and names in two data frames:

1. Province: Indicates the province of birth. There are 13 choices in the survey, only 10 provinces are included in the data. (See data introduction) 10 categories of provinces are: Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec and Saskatchewan.
2. Education: Indicates if the participant receives the education higher than high school. There are 2 categories: No school to high school and College to professional degree.
3. Age: Indicates the age group participant is in. There are 5 categories of age groups: 20 or less, 21 to 35, 35 to 50, 50 to 65 and above 65.
4. Income: Indicates the income group participant is in. There are 6 categories of income groups: 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.
5. Language: Indicates the language participant mainly speaks. There are 3 categories in Language: English, French and Other.
6. Sex: Indicates the biological sex of the participant. For transgender people, see the cleaning process for more information about assigning biological sex. There are 2 categories in Sex: Female and Male.

There is also one categorical variable in the survey data not shared by the census data.

7. Vote_Liberal: Indicates if the participant is most likely voting for Liberal Party. It is a categorical variable can be only found in the survey data. There are two categories in Vote_Liberal: Yes and No.

All the variables involved are categorical variables, there is no need to do numerical summaries.

Graphical Summaries of Variables of interest

Barplot of variable of interest: vote for Liberal in CES data

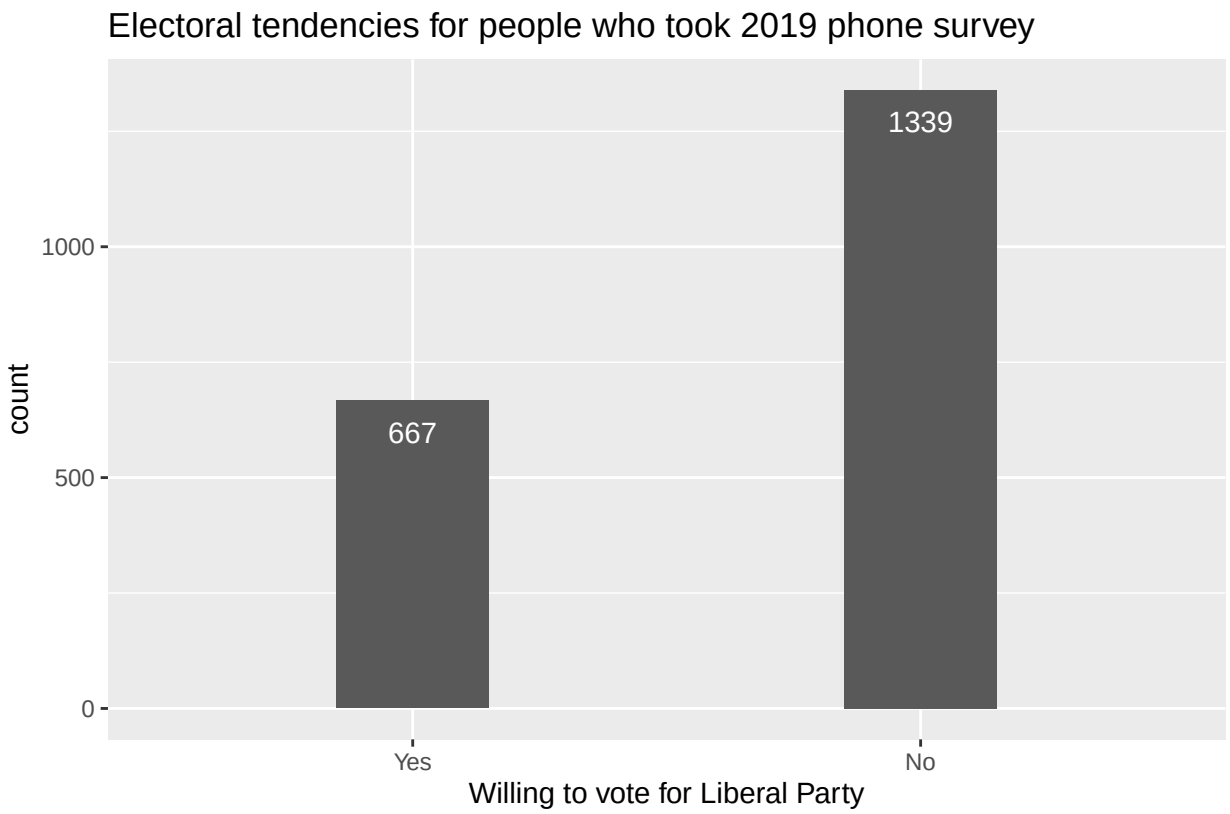

Figure 1. Electoral tendencies for people who took 2019 phone survey

In this paper, we are predicting the probability of Liberal Party winning the election, so we will first take a look at the tendency of people voting Liberal Party. In 2019 phone survey, 667 people were most likely voting for Liberal Party and 1339 people were most likely voting for other parties, which means 33.250% of people said they were possibly voting for Liberal Party in 2019 phone survey.

Barplot of variables of interest: income level and vote to Liberal in CES data

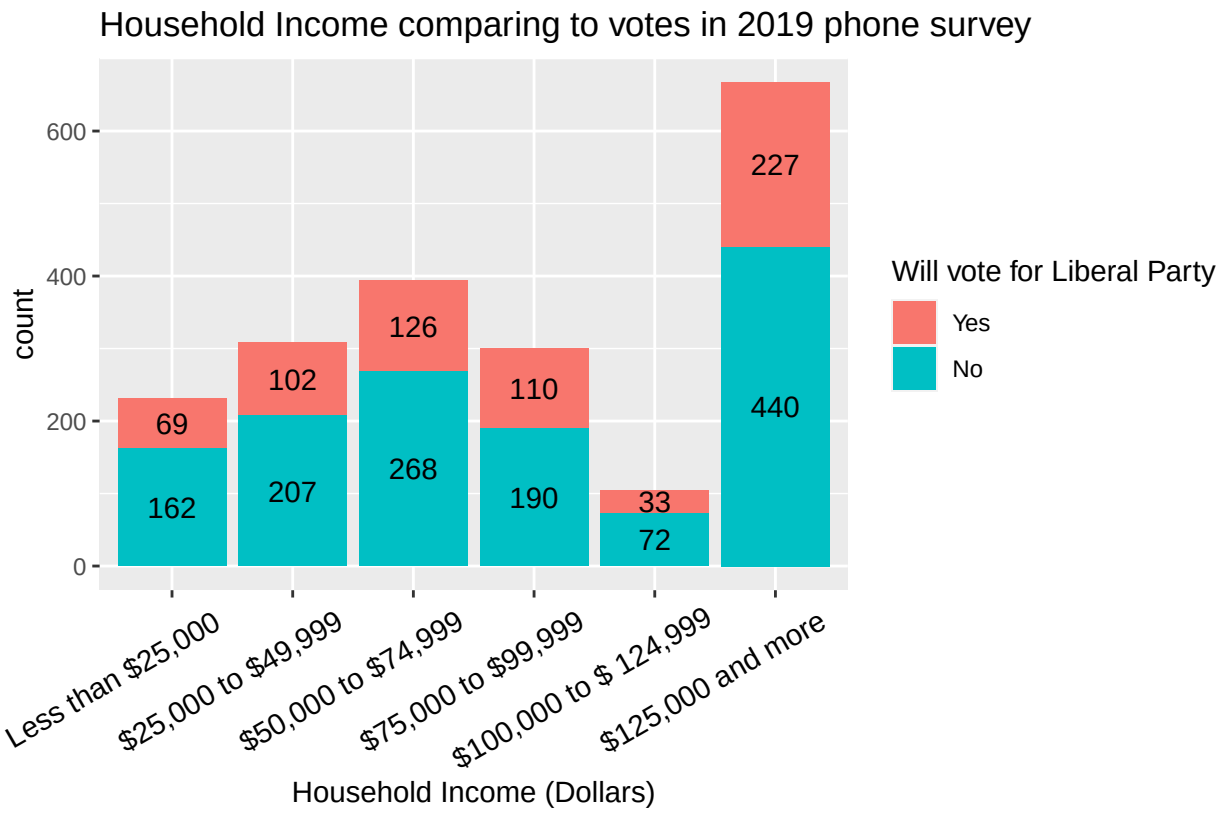

Figure 2. Household Income comparing to votes in 2019 phone survey

In order to find out the relationship between the votes for Liberal Party and the household income, we created figure 2 as a bar plot. From 2006 observations in the phone survey, we can find there are significant portion differences. Arranged by the household income from low to high, the percentage of people voting Liberal Party is: 29.870%, 33.010%, 31.980%, 36.667%, 31.429%, 34.033%. It shows the possibility that people from different classes will be benefited differently under Liberal Party's policies. It is important to consider household income as one of the factors in the regression model below.

Graphical Summaries of Important Variables.

Barplot of variables of interest: age and vote to Liberal in CES data

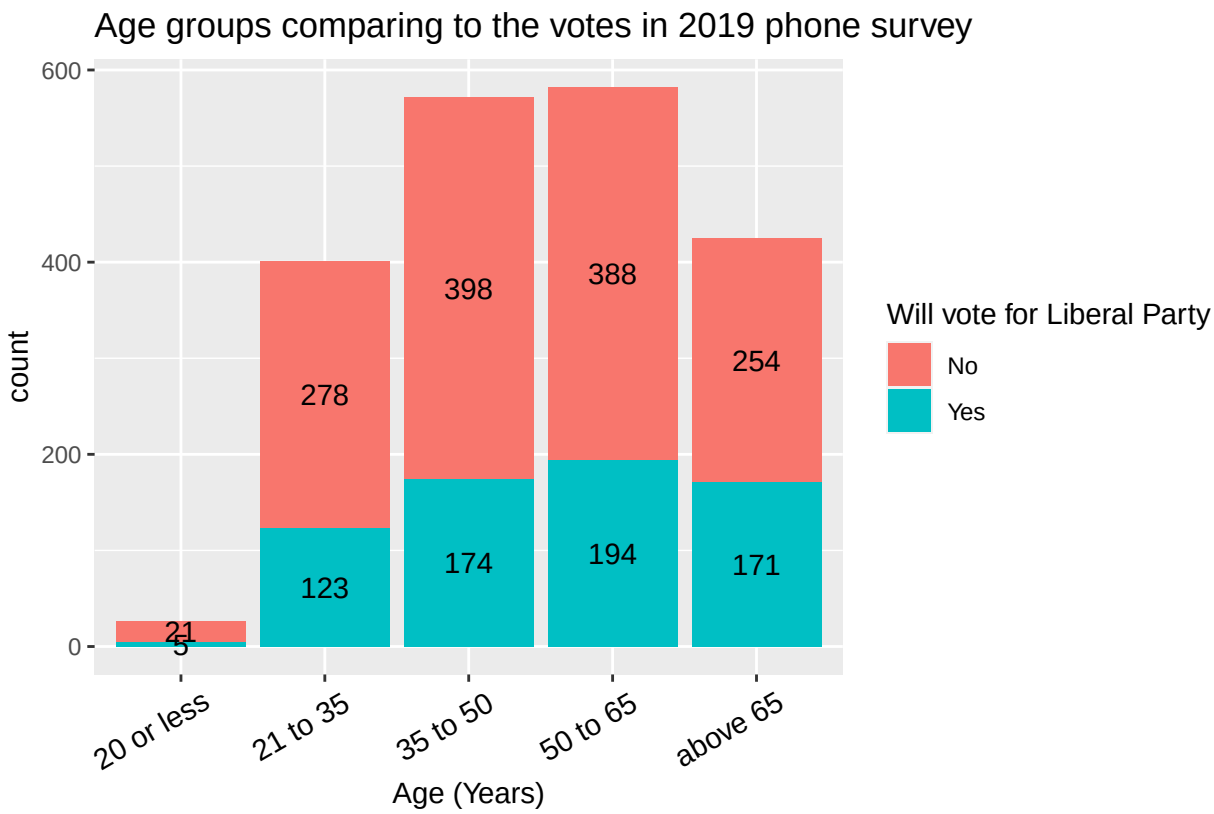

Figure 3. Age groups comparing to the votes in 2019 phone survey

In order to find out the relationship between the votes for Liberal Party and the age groups, we created figure 3 as a bar plot. From 2006 observations in the phone survey, we can find there are also significant portion differences. Arranged by the age groups from young to old, the percentage of people voting Liberal Party is: 19.231%, 30.673%, 30.420%, 33.333%, 40.235%. It shows the possibility that people of different age groups have different reactions to the policies of Liberal Party. It is important to consider age groups as one of the factors in the regression model below.

Barplot of variable of interest: income level and vote to Liberal in GSS data

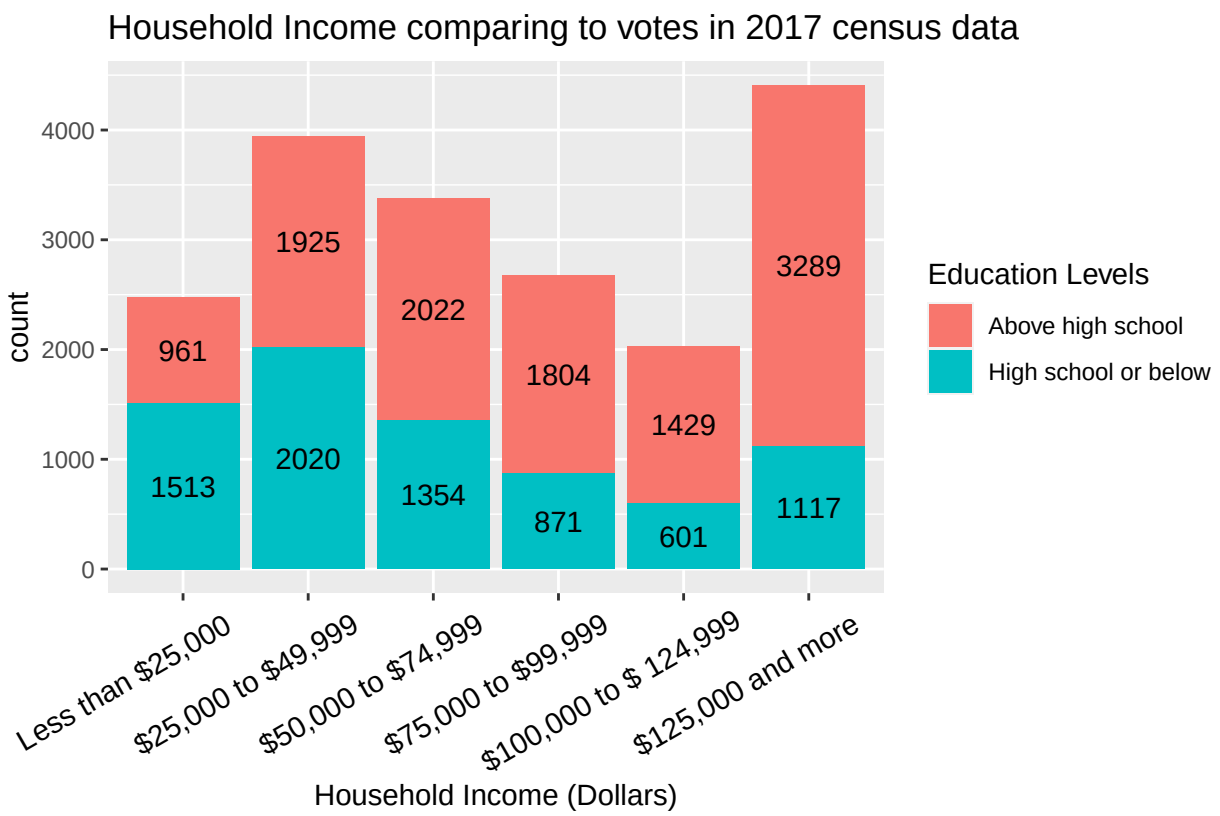

Figure 4.Household Income comparing to votes in 2017 census data

Comparing to figure 3, assuming that all of the Canadians participated in the census, it is easy to find out that people with less than 125000 household income are less likely to participate in the phone survey. It points out a limitation: the samples are not randomly selected. People with higher income are more likely to participate in the phone survey. We can also find out the differences among education levels of different classes. It somehow shows the relationship between income and education, which are both taken as factors in the model below (see limitations part for more details about violation of non-zero correlation assumption).

Methods

Here our goal is to predict the overall popular vote for the next Canadian Federal Election (tentatively in 2025). We decided to use a regression model and then make prediction using post-stratification.

The specific regression model we used is a logistic regression model as our response variable is a probability of a party will be voted. If the probability of one get voted is P, then the logistic regression model will give the distribution for y, where y = log(p/(1-p)). Please see next part for more detailed explanation.

The general idea of a regression model is to find the possible relationship between a response and one or many predictors that is (are) in our interests using existing data set where we have both the inputs (the predictors) and the output (the response). Here we used the information in 2019 Canadian Election Study - Phone survey data to train the regression model.

After a research was done, we found that the result for election can be associated with Age, Sex, Education, Family status, Province that born and Languages that individual speak. There are actually other factors that can affect vote, for example it is indicated by statistic Canada [11] that the immigration status and employment status can be associated with individuals vote. However, we did not include other predictors

other than those listed above as we do not have the data from the census data we will be using for prediction. That is, age, sex, education level, province, language and family income are the variables that can be derived from both data set.

Fitting a logistic regression model with the output and predictors indicated above using survey data will produce a series of coefficient. The model then can be used to make prediction using unseen data for the popular vote for the next election.

However we do have to pay attention to the difference between the data set we use to train the model and the data set we use to make prediction. In the real world, there is no training data that is completely representative of the population except the data consist every example in the population. In our example we want to predict the situation of votes among Canadians, meaning our population is every individual that participates in the Canadian federal election. Here our sample only consists of 4021 individuals, which is far from the entire population. And the data set we will be using to make prediction consists of 20602 individuals. Thus we have to introduce post-stratification method into our model to adjust the estimated coefficient so that we can use the model trained using CES data set to make predictions using GSS data set.

generally speaking, post-stratification is the method of taking a weighted average. However, how to divide the sample into different parts is a tricky question. (Please see Post-Stratification part for more detailed explanation) For example if we are going to group the sample by age, then the probability of each age group voting some party will be calculated and the the weighted average is calculated among different age groups as the probability of post-stratification.

Model Specifics

We are trying to predict the probability of Liberal Party winning the vote by predictors including: Age, Sex, Education, Province, Language, Family income. Each of these variables are shown to be associated with the outcome: whether one will vote Liberal Party.

Thus a multiple logistic regression model is introduced as follows:

$$y = \beta_0 + \beta_1 x_{age} + \beta_2 x_{Sex} + \beta_3 x_{Education} + \beta_4 x_{Province} + \beta_5 x_{Language} + \beta_6 x_{Income} + \epsilon$$

In logistic regression model, y is represented as follows:

$$y = \log(P/(1-P))$$

Where P is the probability of one will vote Liberal Party,

β_0 represents the intercept, which is the value of y where all the numerical predictors are 0 and all the categorical predictors in the first one in the category.

β_1 represents the coefficient associated with predictor: Age,meaning that with are other predictors staying constant, the difference of y between each age groups and the first age group. Here each age group will have a individual β_1 except for the first age group(the difference of y between the first age group and it self is 0), and each individual β_1 is the difference between this particular age group and the first age group.

β_2 represents the coefficient associated with predictor: Sex,meaning that with all other predictors staying constant, the y difference between Male and Female is expected to be β_2

β_3 represents the coefficient associated with predictor: Education, meaning that with all other predictors staying constant, the difference of y between Male and Female are expected to be β_3

β_4 represents the coefficient associated with predictor: Province, meaning that with all other predictors staying constant, the difference of y between each Province and the First Province in the category, which is Newfoundland and Labrador. Similar as the coefficient for age, each Province will have a individual $\beta 4$, which is the expected difference of y between each Province and Newfoundland and Labrador.

β_5 represents the coefficient associated with predictor: Language, meaning that with all other predictors staying constant, the expected difference of y between English speakers and non-English speakers is β_5

β_6 represents the coefficient associated with predictor: Income, meaning that with all other predictors staying constant, the difference of y between each Income group and the Less than 24,999 group. Each Income group will have a individual β_6, showing the expected difference of y between this group and Less than 24,999 Income group.

Post-Stratification

The reason why we are using post-stratification is that, for the survey data, we only have a sample size of 4201 and after cleaning and selection we only have 2006 observations left. However, the sample we will be using to make prediction have a size of 18906 after cleaning. When the sample is not representative enough for the population, we will have to introduce post-stratification otherwise we will be getting a biased prediction.

The process of doing post-stratification is quite simple to follow. First of all the sample should be divided into different groups (cells) based on some variable. Here we choose to create a individual cell for each of the categories in predictor Age and Sex. In our model, there are 5 categories for predictor Age, and the predictor Sex is binary. Therefore, through Post-stratification, the 18906 observations in the data set that we are going to use to make predictions will be divided in to totally 10 possible groups.The predictors in each individual group (cell) will be unique from other groups (cells). Here in our example, since we decided to use Age and Sex predictors to do the post-stratification, one possible cell could be all the individuals that satisfy: Male and Aged aged 21 to 25 (specific categories in variable Age and Sex).

The reason why these two variables were chosen is easy to understand. (Please see Appendix for the exact distributions of Age and Sex of two different data set) The distribution of Age and Sex in the 2019 CES data is significantly different from that of 2017 GSS data. Like mentioned above, Post-Stratification is required when the sample is not representative enough to be used to make prediction for the whole population. Here the Age and Sex variables in the data we used to train the regression model is not representative enough (vastly different from that of GSS data). However, other variables like Language, Income, Education and Province do not seem to differ for different data set or at least can not tell a significant difference from the distribution. (Please see next steps for more information). Thus, only Age and Sex variables are chosen to do the Post-Stratification.

For each individual group (cell), the estimated value will be calculated to be $\hat{y}_j$. Lastly the result that Post-stratification will provide us with is as follows:

$$\hat{y}^{PS} = \sum(N_j \hat{y}_j) / \sum(N_j)$$

Where $\hat{y}^{PS}$ is the result we want from Post-Stratification, $\hat{y}_j$ is the estimated value of y in j^{th} cell, N_j is the size of the population in j^{th} cell.

All analysis for this report was programmed using `R version 4.0.5`.

Results

The variables that will be used for the regression model is whether an individual will vote for the Liberal party, which is the independent and a binary variable. The dependent variables would be age, sex, education level, province, language and family income, where all the variables are adjusted to categorical variables. Age and Sex variable are used to split the data into different cells, and thus the estimated coefficient will not be shown together with the common predictors like Income, Language, Province and Educational level.

Note that since all the predictors used here are categorical, the result that one can get could be different if different base categories are used. To be more specific, the coefficients generated are expected difference between the category and the base category. Here the Base category for Income is: $100,000 to $124,999; base category for Language is: English; base category for Education is: College to Professional and the base category for Province is: Alberta.

Table of regression model outputs:

Variable	Estimate	Std. Error	Statistics	p-value
(Intercept)	-1.906677	0.332836	-5.729	1.01e-08
Education No school to high school	-0.245561	0.134149	-1.831	0.067175
Income $125,000 and more	0.168476	0.235687	0.715	0.474714
Income $25,000 to $49,999	0.030493	0.251486	0.121	0.903492
Income $50,000 to $74,999	-0.007897	0.244880	-0.032	0.974275
Income $75,000 to $99,999	0.229364	0.250925	0.914	0.360678
Income Less than $25,000	-0.181890	0.264998	-0.686	0.492473
Language French	0.023304	0.203764	0.114	0.908947
Language Other	0.714179	0.156987	4.549	5.38e-06
Province British Columbia	0.781879	0.268909	2.908	0.003642
Province Manitoba	0.828384	0.309464	2.677	0.007432
Province New Brunswick	1.272819	0.334613	3.804	0.000142
Province Newfoundland and Labrador	1.566520	0.325499	4.813	1.49e-06
Province Nova Scotia	1.326506	0.322681	4.111	3.94e-05
Province Ontario	1.497664	0.264181	5.669	1.44e-08
Province Prince Edward Island	1.504751	0.320499	4.695	2.67e-06
Province Quebec	1.339177	0.309025	4.334	1.47e-05
Province Saskatchewan	0.080692	0.345273	0.234	0.815215

Note that the Estimate value in the table is the in the logistic form, for example if the Probability of getting the vote is p then the estimated value is shown as $log(p/(1-p))$ and is known as "log odds". Transforming that into probability is easy, if the log odds is y, then the probability p is equal to $exp(y)/(1+exp(y))$.

The regression model tells us that when an individual is from the base category (come from Newfoundland and Labrador, Speak English, Have a higher level of education and have a family income between $100,000 and $124,999),then the estimated log odds for the probability that individuals will vote for the Liberal party is -1.906677. (or equivalently, probability p = 0.1293546)

The log odds of people with a education level of no school to high school is expected to be 0.244341 smaller than that of a higher education level (college to professional). That means people with higher educational level are expected to have a higher probability to vote Liberal Party. However, the p-value is 0.067175, which is larger than the commonly use significance level 5%, we can conclude that there is no statistically significant evidence to show that education level will influence an individual's vote considering whether they will vote for Liberal Party or not.

Similarly, although people from different Income group showed different attitudes regarding their probability of voting Liberal Party, the p-values for different income levels are not significant (larger than the commonly use significance level 5%), indicating that is no evidence to show that people from different income level will have a different probability of voting Liberal Party. Or equivalently, we fail to show that there is an association between Income and the probability of voting Liberal Party.

For the variable Language, we found that people that speak English behave very similar to those who speak French (small difference with large p-value). However, there is significant difference considering the probability of whether they will vote for Liberal Party. People who don not speak English or French are expected to be more supportive for Liberal Party.The p-values calculated is 5.69e-06 (extremely significant result). This is a reasonable result as people that don't speak English or French might be immigration, who will be benefited under Liberal Party's governance [13].

For voting condition to Liberal in different provinces in Canada, we found that all provinces except Saskatchewan behave differently compared to Alberta (base category). Among the Provinces, Newfoundland and Labrador is the most supportive ones for Liberal Party. This is also consistent with the beneficial policy that Liberal Party promises to impose [14]. All p-values expect for Saskatchewan are significant (smaller than 5%), which shows that there is an evidence that an association exists between province that people live and whether they will vote for Liberal party.

Scatterplot of age and sex with vote to Liberal:

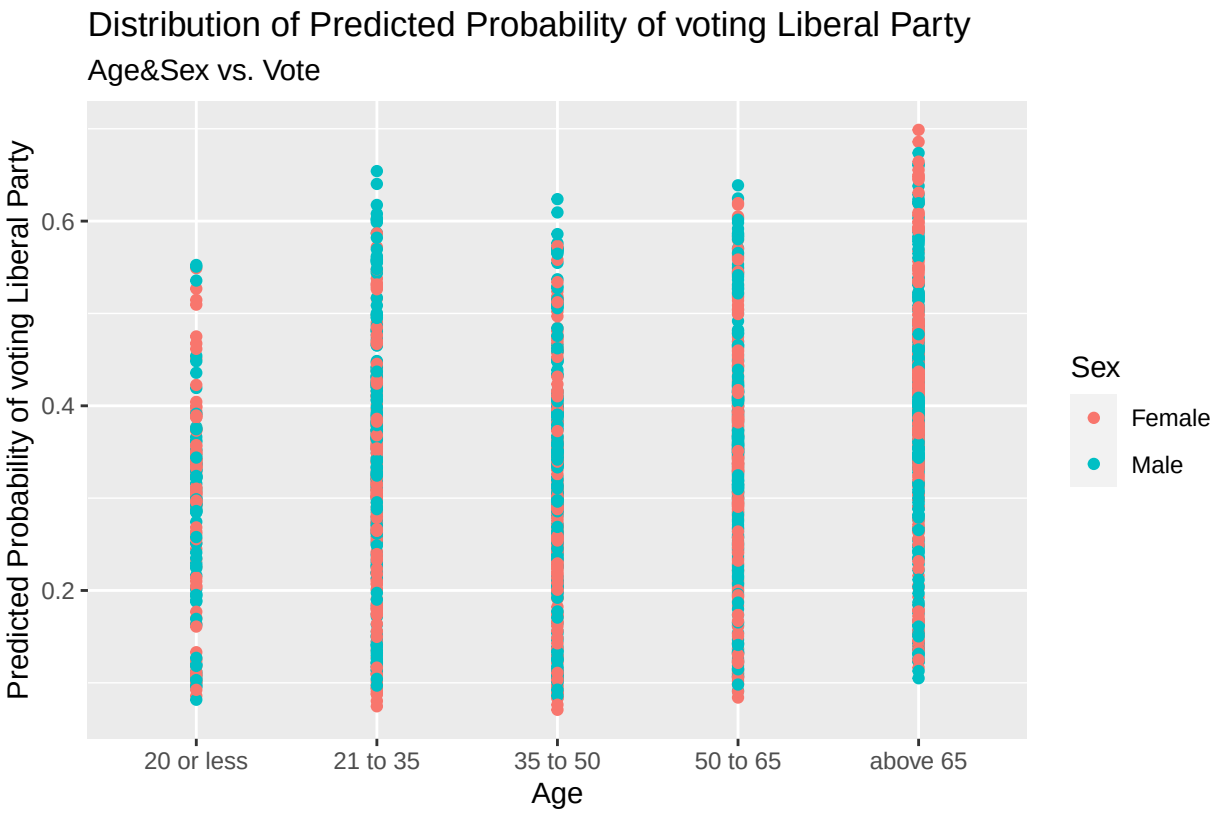

Figure 5.Distribution of Predicted probability of voting Liberal Party

The plot above shows the distribution of the predicted probability of whether individuals will vote Liberal Party among different age groups and Sexes. From the plot we can see that the overall predicted probability of voting Liberal Party is going up as age goes up. And it is obvious for us to find that Females that aged above 65 are more likely to vote Liberal Party.

Table 2: Predicted percentagt of population that will vote Liberal Party

liberal_predict
0.3336636

The result we get after doing post-stratification is shown above. That is the estimated probability of Liberal Party getting voted from the entire population (assuming there will only be 18906 samples in the population). It shows the percentage of population that will vote for the Liberal party will be 33.37%. According to the past two elections' popularity result, Liberal has won 39.47% in 2015 and 33.12% vote in 2019 [16]. Therefore, the prediction is reasonable as there is not much deviation from the voting results of previous years.

Conclusions

Summary of the Hypotheses, Methods and Results

The hypotheses of the research question is that there is an association between votes for the Liberal party with age, sex, education level, province, language and family income. According to the "2019 CES phone survey data" and "General Social Survey data 2017" data set, logistic regression is being used to find the relationship between votes for the Liberal party with age, sex, education level, province, language and family income. In addition, the post-stratification method is also being used to calculate the weighted average of age and sex for different groups since the sample is not representative enough to be used to make prediction for the whole population.

The result of the regression shows that there is an association between average votes for Liberal and the age, sex of people as well as the province they live in. However, there is not enough evidence that shows votes of the Liberal party relates to people's education level, language they speak and an individual's family income.

In addition, the overall goal for our study is to predict the probability of the Liberal Party would win the Canadian Federal Election tentatively in 2025. The result we got from the study is 33.37%, meaning that 33.37% of the population (samples in GSS data) will vote for Liberal Party. Thus, there will be a great chance for them to win the Election.

With the result of the study, the Liberal party could have a better understanding of the population groups who vote for them. For younger generation, they are less likely to vote for Liberal compared to older generations. Moreover, for female age 21 to 50, they are less likely to vote for Liberal than male. Through the distribution of votes, the party could know specific groups of the people that might not satisfied with the decisions and policies they made so that some corresponding adjustments could be made.

Limitations

1. Both the "General Social Survey data 2017" and "2019 CES phone survey data" are a little out of date as the 2021 Canadian election has finished. It might not best predict the 2025 election result.
2. During the data cleaning process, the observations with NA are removed from analysis, the reasons might be people prefer not to answer specific questions, which omits some of the population. It could not give people a more intuitive feeling about the whole voting condition of the Canadian election.
3. When defining if the vote is valid during the data cleaning process, we are setting people who are unlikely to vote, but if they were to vote, which party they would vote for as NA. The reason is that it is difficult to separate the group with the design of the survey. If people answer unlikely to vote, but eventually vote for a party, that would make a different result for our prediction.
4. In data cleaning process, we had to map Gender variable in GSS data set into Sex to be able to make predictions. Because there is only Sex variable in CES data set which we used to train the model. However, Gender variable in GSS data set has a category of "Transgender".The assumption of transgender group has the same proportion of male and female is made, which might be inconsistent with reality. There are other ways how you can map Gender to Sex, for more detailed information of other mapping method and their limitations, please refer to 17.
5. In order to avoid making the coefficient we get from the model to be biased, we assumed the predictors to be completely independent from each other. But we have to include both variables as educational level are shown to be associated with vote while individuals with a higher Income are more likely to participate in the survey. The model provided us with the result that both predictors are not associated with whether on will vote for Liberal Party. However, since the correlation between Income and Educational level is not zero, we might end up getting a biased estimate of coefficients for Educational Level or for different Income categories.
6. According to the research, there are other factors that would influence the voters' behavior, such as religion and race. However, not all variables that are potentially related to voting results are included in the model, which makes the result less accurate.

7. In the data part, we assumed that the choices of the province where people are living in from the survey data(CES data)to be the same as the choice of province of birth in the census data(GSS data). In reality there are definitely people who are living in a Province that is different than that which the people were born. Violation of this assumption might make the result less accurate.

Next Steps

The "General Social Survey data 2017" and "2019 CES phone survey data" share more common variables that have potential relationship with votes for the Liberal party and do not take account of this study, we could take a deeper investigation to see all variables that have association with the votes for Canadian election.

In the study we used Alberta as the base category to study the vote among different Provinces and got the result that nearly every Province has a different attitudes towards Liberal Party. Later on, a more detail analysis could be done do study the difference of voting among different Provinces by setting a different bast province.

In addition, during the post-stratification process, we use the predictors age and sex. For further analysis, we could choose other predictors and adjust the weight in a population so that the weighted totals within mutually exclusive cells equal the known population totals. The result of the prediction would be more accurate. (For example include the Province predictor in the Post-Stratification process).

Discussion

An election gives people the opportunity to make their own choice. The result of the election and the subsequent impact represent the choice of the majority of the people and the public's recognition of the result of the election. Through studying the association between different demographic groups with the party they vote for, parties could have a better understanding of the division of their votes. For our study, the Liberal Party can consolidate the groups that would vote for them and strengthen welfare to persuade those who will not vote for them, and thus win the election.

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. Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://github.com/sfirke/janitor

5. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

6. Mine Çetinkaya-Rundel, David Diez, Andrew Bray, Albert Y. Kim, Ben Baumer, Chester Ismay, Nick Paterno and Christopher Barr (2021). openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs. http://openintrostat.github.io/openintro/, https://github.com/OpenIntroStat/openintro/.

7. Rayside, D. (2021). Liberal Party of Canada - Policy and structure. Encyclopedia Britannica. Retrieved 3 November 2021, fromhttps://www.britannica.com/topic/Liberal-Party-of-Canada/Policy-and-structure.

8. Liberal Party of Canada - Wikipedia. En.wikipedia.org. (2021). Retrieved 3 November 2021, from https://en.wikipedia.org/wiki/Liberal_Party_of_Canada.

9. Factors Affecting Voter Behavior. Everettsd.org. (2010). Retrieved 3 November 2021, from https://www.everettsd.org/cms/lib07/WA01920133/Centricity/Domain/506/6%20Factors%20Affecting%20Voter%20Behavior%2010.pdf.

10. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Phone Survey", https://doi.org/10.7910/DVN/8RHLG1, Harvard Dataverse, V1, UNF:6:eyR28qaoYlHj9qwPWZmmVQ== [fileUNF] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8RHLG1

11. Statistic Canada. 2021. Factors associated with voting. [online] Available at: https://www150.statcan.gc.ca/n1/pub/75-001-x/2012001/article/11629-eng.htm [Accessed 3 November 2021].

12. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

13. El-Assal, K., 2021. Election 2021: What Canada's parties say about immigration. [online] CIC News. Available at: https://www.cicnews.com/2021/08/election-2021-what-canadas-parties-say-about-immigration-0818986.html#gs.fhrn7o [Accessed 5 November 2021].

14. CBC. 2021. Liberals claim slim majority in Newfoundland and Labrador, as voters tap Furey to lead | CBC News. [online] Available at: https://www.cbc.ca/news/canada/newfoundland-labrador/nl-election-results-2021-1.5966912 [Accessed 5 November 2021].

15. Yihui Xie.(2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.34

16. 2019 Canadian federal election - Wikipedia. En.wikipedia.org. (2021). Retrieved 3 November 2021, from https://en.wikipedia.org/wiki/2019_Canadian_federal_election#2015.

17. Kennedy, L., Khanna, K., Simpson, D., & Gelman, A. (2020, September 30). Using sex and gender in survey adjustment. arXiv.org. Retrieved November 5, 2021, from https://arxiv.org/abs/2009.14401

18. Statistics Canada, 2020, "General Social Survey Cycle 31: Family, 2017", https://hdl.handle.net/11272.1/AB2/G3DUFG, Abacus Data Network, V2, UNF:6:H+FqbrJ37CGYLaeXt4LBkw== [fileUNF]

Appendix

Choice of Variables fot Post-Stratification Continued

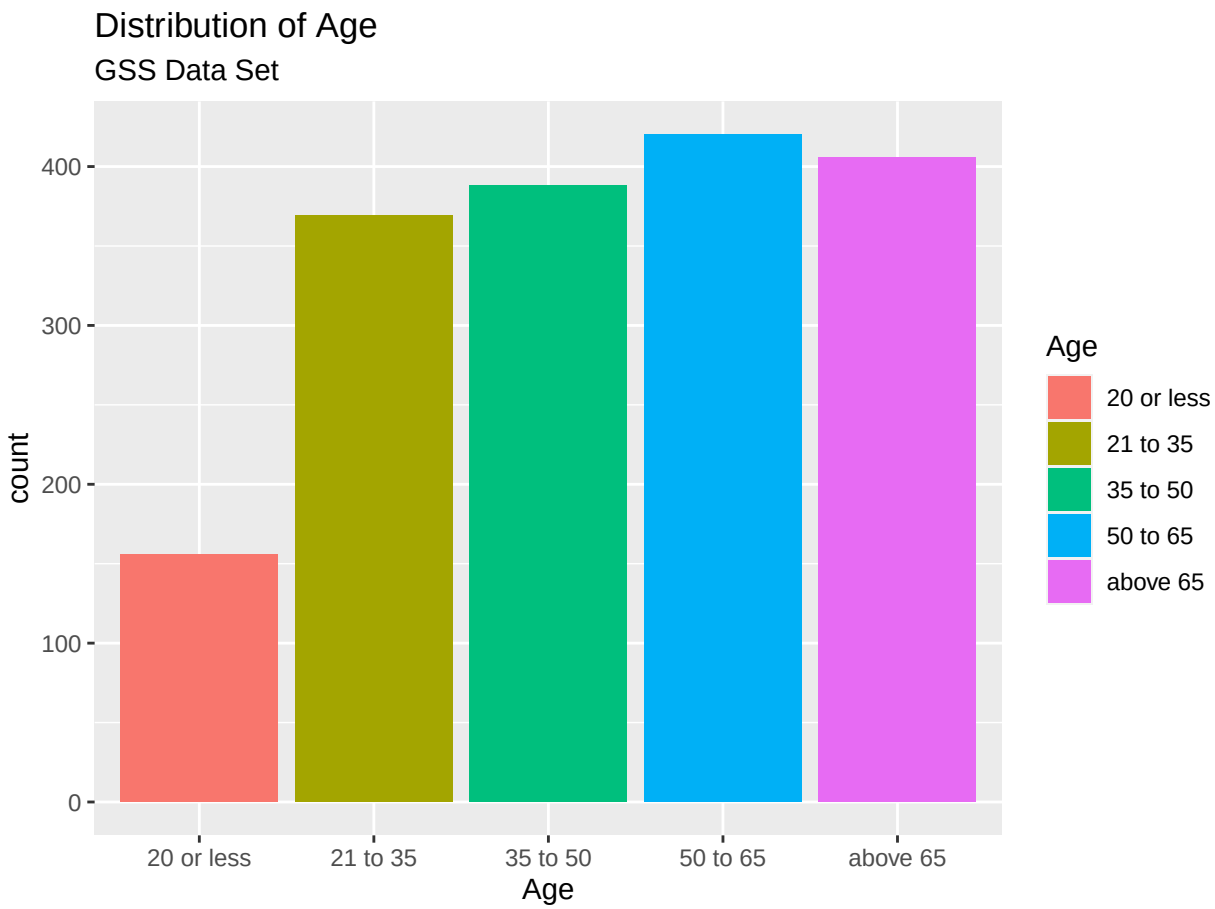

Figure 5.1 Distribution of Age among GSS Data Set

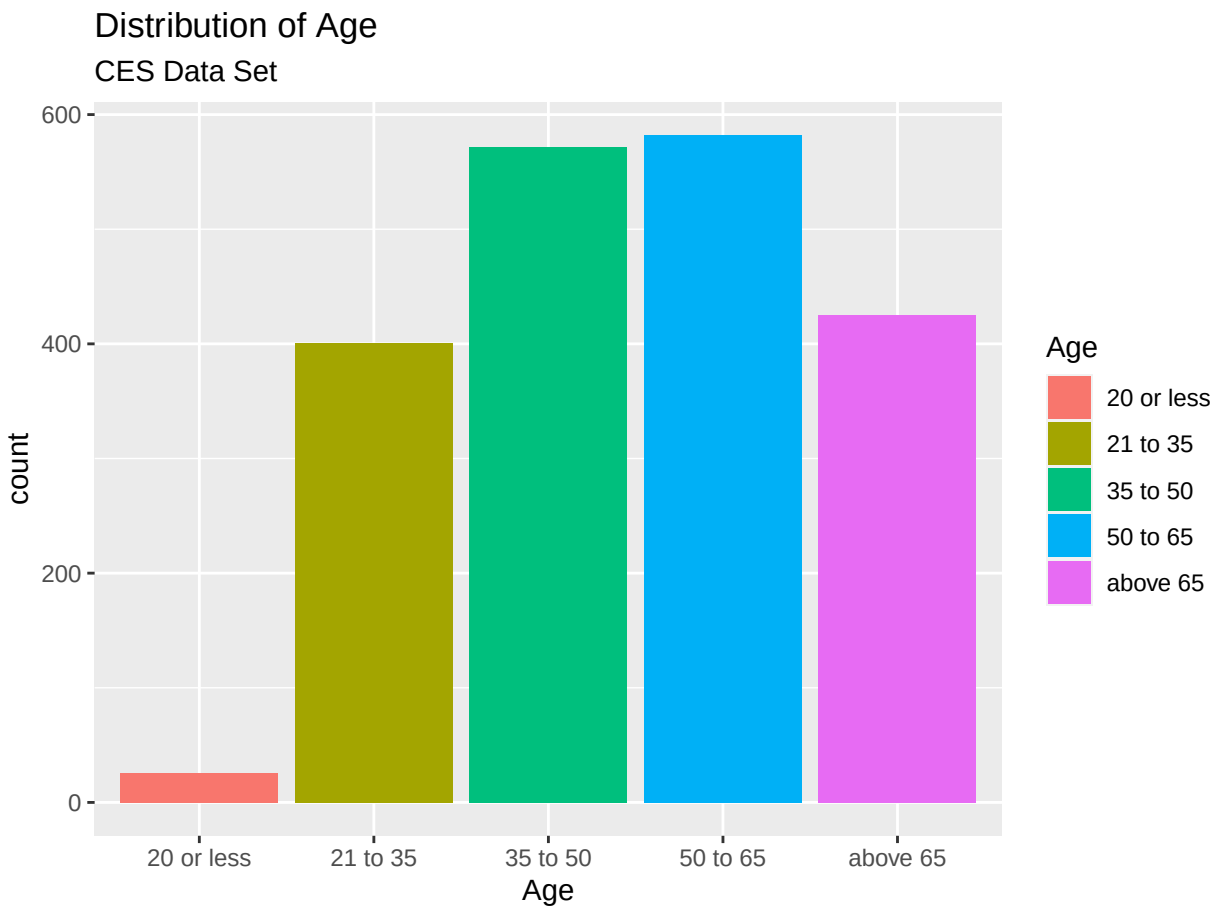

Figure 5.2 Distribution of Age among CES Data Set

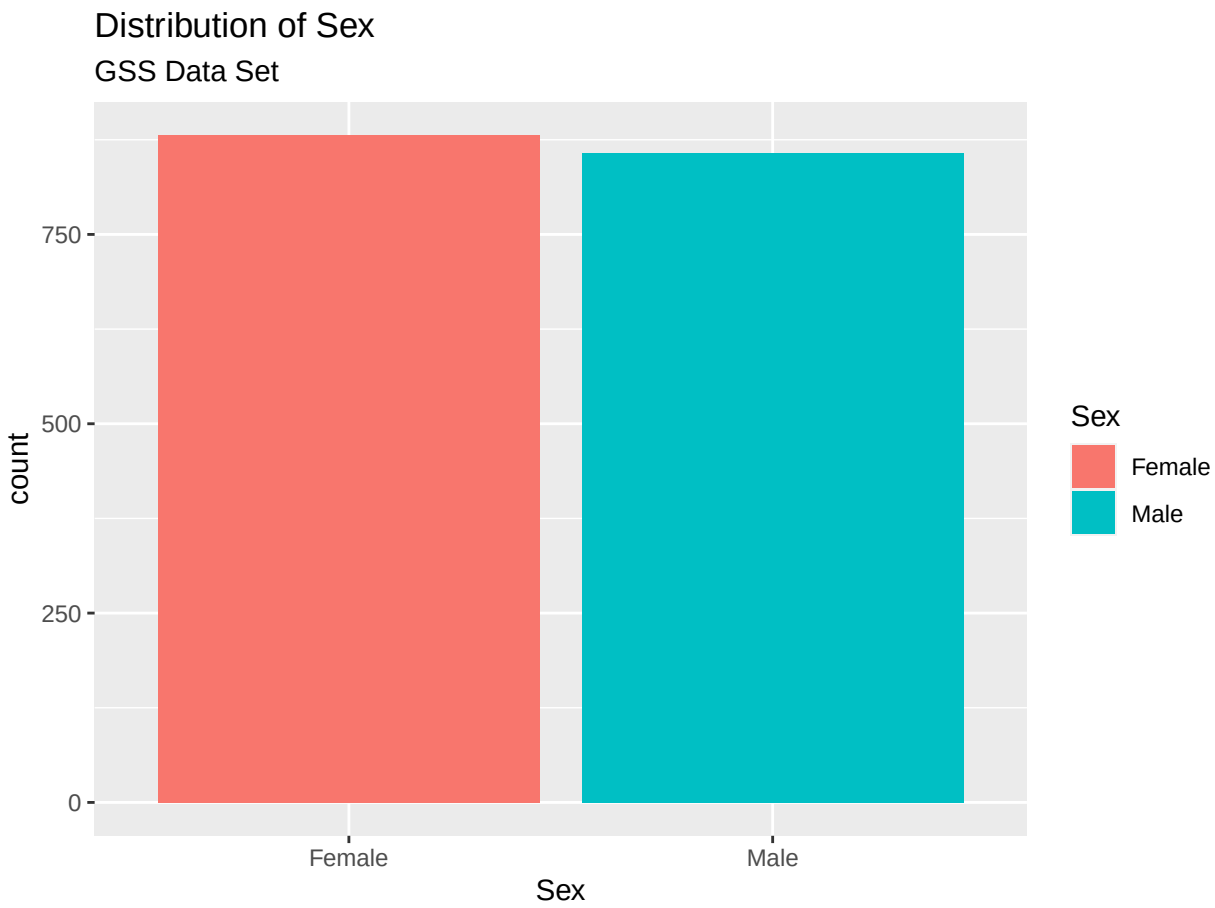

Figure 5.3 Distribution of Sex among GSS Data Set

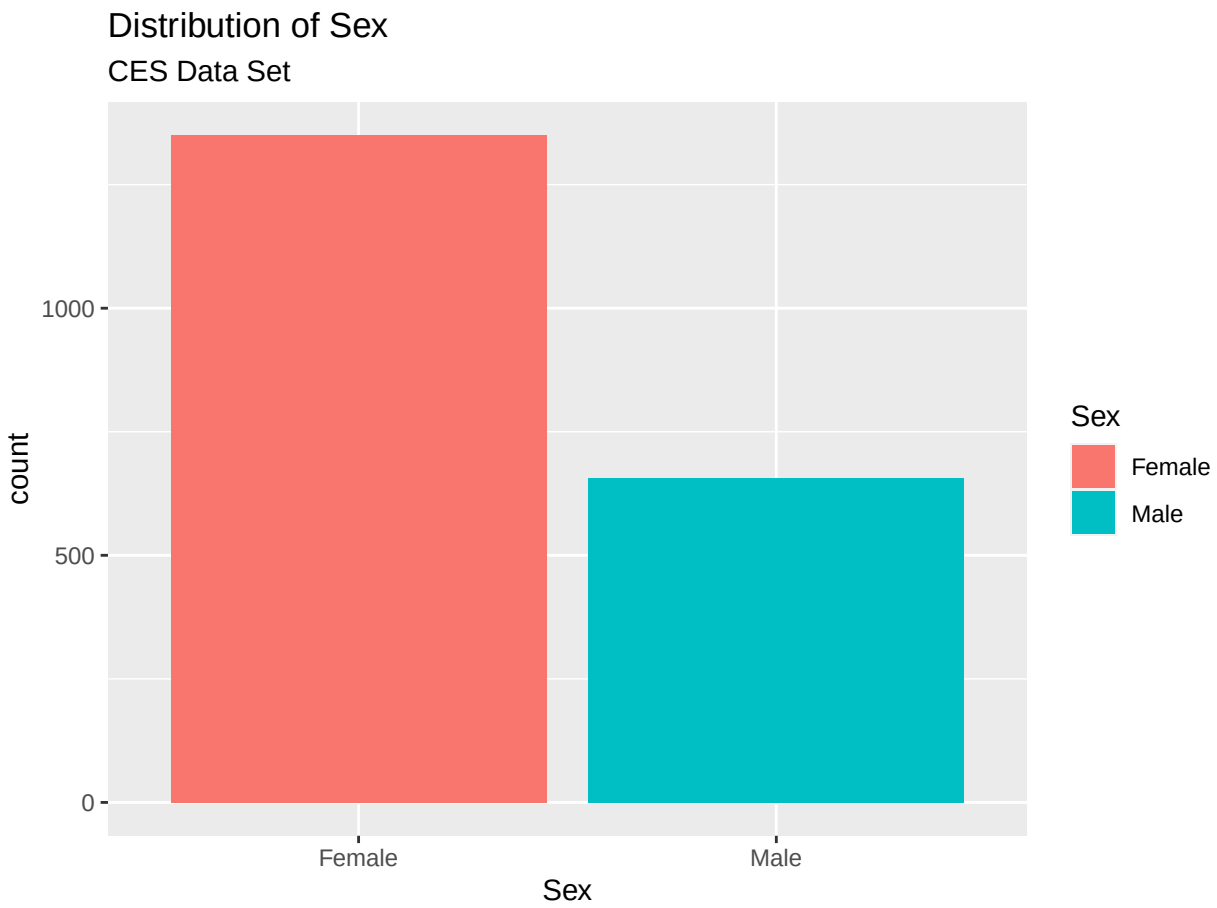

Figure 5.4 Distribution of Sex among CES Data Set

Here we can find out the difference of distribution of Sex and Age in our sample(CES data) and target population(GSS data) we will use to make predictions. Therefore these two variables are chosen to split the data to do the Post-Stratification.

Glimsp of data used to make predictions

Please see below for a glimpse of the cleaned data used for making predictions.

Table 3: Glimpse of the cleaned GSS data

Province	Education	Age	Income	Language	Sex	cell
Quebec	No school to high school	50 to 65	$25,000 to $49,999	French	Female	50 to 65 Female
Manitoba	College to professional degree	50 to 65	$75,000 to $99,999	English	Male	50 to 65 Male
Ontario	College to professional degree	50 to 65	$75,000 to $99,999	French	Female	50 to 65 Female
Alberta	No school to high school	above 65	$100,000 to $ 124,999	English	Female	above 65 Female
Quebec	College to professional degree	21 to 35	$50,000 to $74,999	French	Male	21 to 35 Male
Quebec	No school to high school	50 to 65	$50,000 to $74,999	French	Female	50 to 65 Female

Province	Education	Age	Income	Language	Sex	cell
Nova Scotia	No school to high school	50 to 65	Less than $25,000	English	Female	50 to 65 Female
Quebec	No school to high school	above 65	Less than $25,000	French	Female	above 65 Female
British Columbia	No school to high school	50 to 65	Less than $25,000	English	Female	50 to 65 Female
Saskatchewan	No school to high school	21 to 35	Less than $25,000	English	Male	21 to 35 Male