Will the Liberal Party Win the Next Election in 2025?

STA304 - Assignment 3

GROUP 10: Xinjing Guo, Sitong Lin, Chuanfeng Li, Yuqiang Zhang

November 5, 2021

Introduction

In Canada, the electoral system is referred to as a single-member plurality system. This means that during the riding, the political party with the most votes gets a seat in the House of Commons and is the member of Parliament for that constituency. The governor general requests members of Parliament to form a government, which is usually led by the party with the most seats in Parliament; the leader of such party is usually named Prime Minister (Canadian electoral system, 2021).

There are 22 registered political parties in Canada and the major 6 parties are: the Liberal Party, the Conservative Party, the New Democratic Party, the Bloc Québécois, People's Party and the Green Party (Registered Political Parties, n.d.).

Different political parties have their unique promises and approaches to Canadian public concern. With the most previous federal election held on September 20th, 2021, the Liberal Party, led by Justin Trudeau, won the election and is currently the political party that governs the country (2021 Canadian federal election, 2021). Despite the election results, in the past 6 years where Liberals were in power, political parties always have fierce competition during the election, and our community had concerns about whether the Liberal Party benefits the country more than other political parties. With this background information, our question of interest is **who proportion of the Canadian will vote for the Liberal Party in the 2025 Canadian federal election** (45th Canadian federal election, 2021). The importance of this analysis is that it could help the Liberal Party and their supporters to understand their probability of winning in the next federal election, and realize the potential factors that might influence the election results by public voting.

In order to perform this prediction, we are going to use 2 data sets. The first data set is called **2019 Canadian Election Study**, this is a phone survey taken in the Campaign Period for 2019's federal election. The second data set is called **2017 GSS Data**, this is the General Social Survey data being collected nationally in 2017. The reason we are using 2 data sets is because we are using the **2019 Canadian Election Study**, the smaller data set, to form a model and use that model to predict the outcome based on the larger data set, **2017 GSS Data**.

Before performing the analysis, we guess that **the Liberal Party would still have a high chance of getting elected in the 2025 federal election**. It is because, in the past, Justin Trudeau was engaged in solving global problems such as climate change. And he is currently evoking the world to have a standard for pricing carbon (Connolly, 2021). Canadian people, living in a continent with abundant forest resources and having close contact with nature, are likely to support Justin Trudeau's concern, and this might be the driving factor that leads the Canadian community to support the Liberal Party.

Terminology

Single-member Plurality System

- Single-member plurality (SMP) systems are common in countries that have inherited elements of the British parliamentary system; Canadians are most familiar with this type of electoral system. Simple rather than absolute majorities are sufficient to determine the winner of an electoral contest in electoral districts represented by one member in an elected assembly. Each elector places a single "X" (or other similar mark) beside the name of his or her preferred candidate. Although several candidates may run for the seat, the winner only needs to receive the most votes cast. As a result, this type of electoral system is known as a "single-member plurality" (Canadian electoral system, 2021).

House of Commons

- The House of Commons serves as a conduit between Canadians and their government. Farmers, teachers, lawyers, business people, and others who we elect to represent us bring their ideas and experience to their work (Guide to the Canadian House of Commons, n.d.).

Parliament

- The Canadian Parliament is divided into three chambers: the Queen, the Senate, and the House of Commons. They collaborate to create our country's laws. The Queen, the Prime Minister and Cabinet, and the departments of government comprise the executive branch. They carry out the laws (Guide to the Canadian House of Commons, n.d.).

Data

Data collection

Survey data: Firstly, **2019 Canadian Election Study** was used as the survey data in this project. This database mainly contains the information about Canadian Election in 2019, such as voters' age, family income, education, province and party affiliation (Gibbs & Stringer, 2021). This data was collected by Canadian Election Study, and this study is investigated annually (Gibbs & Stringer, 2021). The survey was conducted after the election. A total of 4,021 interviewees were interviewed. First, by telephone, they would be asked whether they were willing to complete the survey. Then, if they were willing to participate in this survey, and they can choose to reply by phone or email (CES, 2020). More specifically, the phone number was selected by Random Digit Dialling (Harell, Loewen, Rubenson, & Stephenson, 2020). Random Digit Dialling is a common procedure to choose the phone number randomly by computer (Random Digit Dialling, n.d.). Also, the number of calls per day is fixed according to the certain proportion by province and phone type, which ensures that the results are average across provinces and call types. (CES, 2020). Moreover, if the system dials six times and the person does not pick up, this sample would be removed, and the system would choose another sample (CES, 2020). If someone answers the phone the first six times and is willing to accept the interview, this person can choose to accept the interview by phone or email (CES, 2020). Therefore, the participants' information was recorded by Canadian Election Study.

Census data: Secondly, **2017 GSS Data** was used as the census data in this project. This database mainly contains information about the Canadians in 2017, such as age, family income, education and province. This data was collected by General Social Survey. The survey mainly studies social problems, such as family problems (The General Social Survey: An overview, 2017). This survey is conducted every five years (The General Social Survey: An overview, 2017). The survey would involve telephone interviews with people over the age of 15 (The General Social Survey: An overview, 2017). The samples were chosen from ten provinces, and Random Digit Dialing was the method to select the phone number (The General Social Survey: An overview, 2017). The interview lasted more than half an hour on average (The General Social Survey: An overview, 2017). The calls were made during business hours, and if the sample declined to participate in the survey, the staff would call another two times to clarify the significance of the survey and motivate them to participate. (Cycle 31, 2020) The purpose of this is to expand the number of samples. Therefore, the participants' information was recorded by General Social Survey.

Data collection

Survey data: Firstly, **2019 Canadian Election Study** was used as the survey data in this project. This database mainly contains the information about Canadian Election in 2019, such as voters' age, income, education, province and party affiliation (Gibbs & Stringer, 2021). This data was collected by Canadian Election Study, and this study is investigated annually (Gibbs & Stringer, 2021). The survey was conducted after the election. A total of 4,021 interviewees were interviewed. First, by telephone, they would be asked whether they were willing to complete the survey. Then, if they were willing to participate in this survey, and they can choose to reply by phone or email (Harell, Loewen, Rubenson, & Stephenson, 2020). More specifically, the phone number was selected by Random Digit Dialling (Harell, Loewen, Rubenson, & Stephenson, 2020). Random Digit Dialling is a common procedure to choose the phone number randomly by computer (Random Digit Dialling, n.d.). Also, the number of calls per day is fixed according to the certain proportion by province and phone type, which ensures that the results are average across provinces and call types (Harell, Loewen, Rubenson, & Stephenson, 2020). Moreover, if the system dials six times and the person does not pick up, this sample would be removed, and the system would choose another sample (Harell, Loewen, Rubenson, & Stephenson, 2020). If someone answers the phone the first six times and is willing to accept the interview, this person can choose to accept the interview by phone or email (Harell, Loewen, Rubenson, & Stephenson, 2020). Therefore, the participants' information was recorded by Canadian Election Study.

Census data: Firstly, only quantified Canadian citizens over the age of 18 have the right to vote (Elections in Canada, 2021); therefore, only Canadian citizens over the age of 18 are selected. Furthermore, the education background can also be divided into three levels; high school or below, college or university, master or beyond, according to the standard of the survey data. Also, since some ages have a decimal point, we rounded them up to a whole number. Then, the missing values from age, family income, educational background, and province were deleted. Finally, in order to match the variables from the survey data, 18750 Canadians were selected to participate in the analysis, along with their age, family income, education level, and province.

Important variable:

Age, education level, family income and province contain in the survey and census data. Vote only contains in the survey data.

Vote: The indicator of whether a vote for liberal or not. (categorical)
Age: Age of each participant. (numerical)
Education level: Education background of each participant. (categorical)
Income: Family income level of each participant. (categorical)
Province: The province in which each participant currently live (categorical)

Data summary

Response Variable: Vote (categorical)

This is a summary information about the vote:

Table 1: Vote Proportion

Option	Vote Liberal	Non-vote Liberal
count	731	2284
percentage	0.2424544	0.7575456

Table 1 shows the specific values of the voting situation in survey data. A total of 3015 participants took part in the survey, of which 731 voted for the Liberal, approximately 24.25% of the total participants. And the remaining 2284 did not vote for the Liberal, which took about 75.75% of the total participants in the survey. Further analysis in more detail will be carried out in connection with the relevant graphs below.

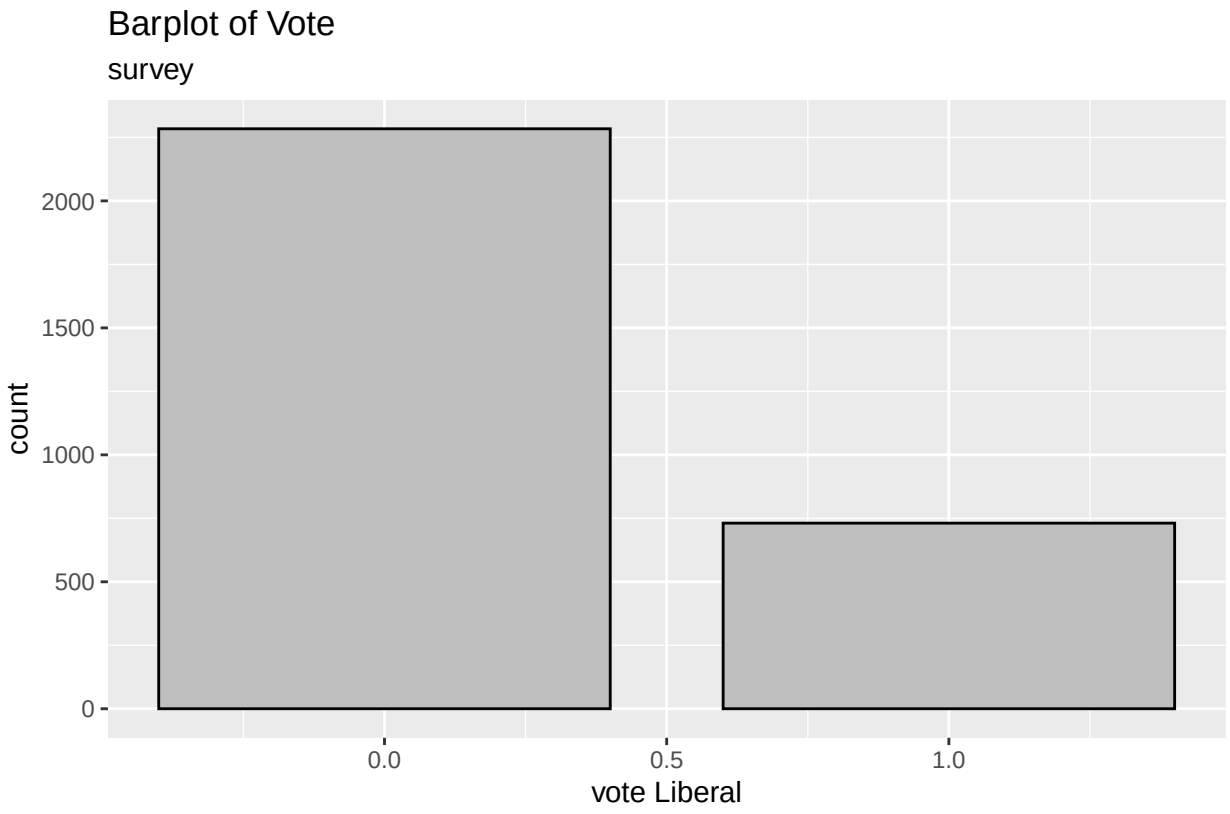

Figure1

Figure 1 shows the counts of voting Liberals and non-voting Liberals. 0 represents the participants who do not vote Liberal, 1 represents participants who voted Liberal. In the plot, the number of those who did not vote Liberal is three times more than the number who voted. Although it seems that fewer people voted Libertarian than did not vote, because the number of people who voted for all the other parties was divided into five other parties. Thus, the distribution of the plot is reasonable.

By the combination of Table 1 and Figure 1, votes of the voting Liberals account for a quarter of all votes.

Through this analysis, we believe that the majority of people still support the Liberals, as we intend to predict the probability of a Liberal victory in 2025. The goal of the project is to predict the probability of voting liberal, thus, vote as a response variable will be included in the model.

Predictor Variable: Age (numerical)

There is a summary information about the numerical variable Age of mean, median, IQR, min and max in two datasets:

Table 2: Numerical summary about Age

Variables	Mean	Median	IQR	Min	Max
age in census	53.65	56.00	28	18	80
age in survey	50.3	50.0	25	18	95

Table 2 shows the numerical information of age in census data and survey data. The mean and median values of age in the survey are around 50, the range of it is from 18 to 95. In addition, the mean and median values of age in the census are around 54, the range of it is from 18 to 80. The IQR values of the 2 datasets are also very close to each other both around about 27.
As we can see the slight the difference between the sample age distribution and the census age distribution. This would suggest that, in the later analysis, using post-stratification could improve the prediction of proportion for voting the Liberal Party.

Further analysis in more detail will be carried out in connection with the relevant graphs below.

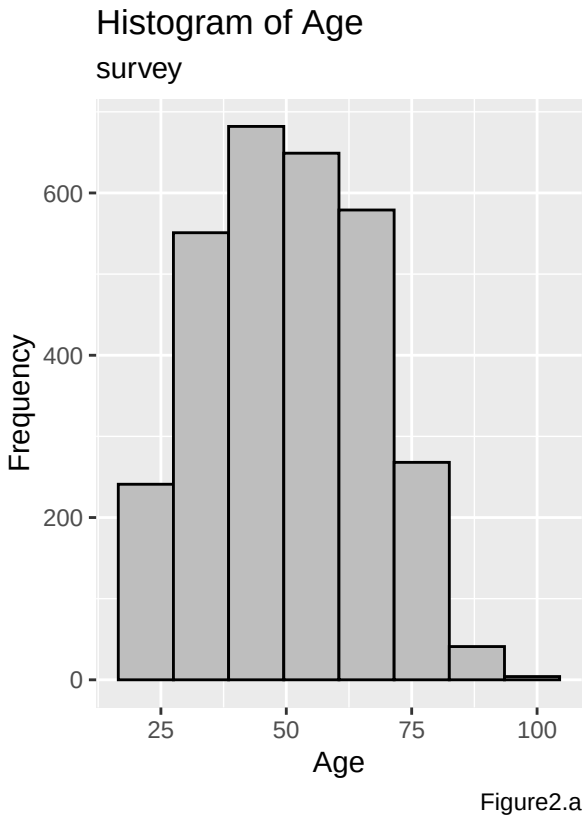

Figure2.a

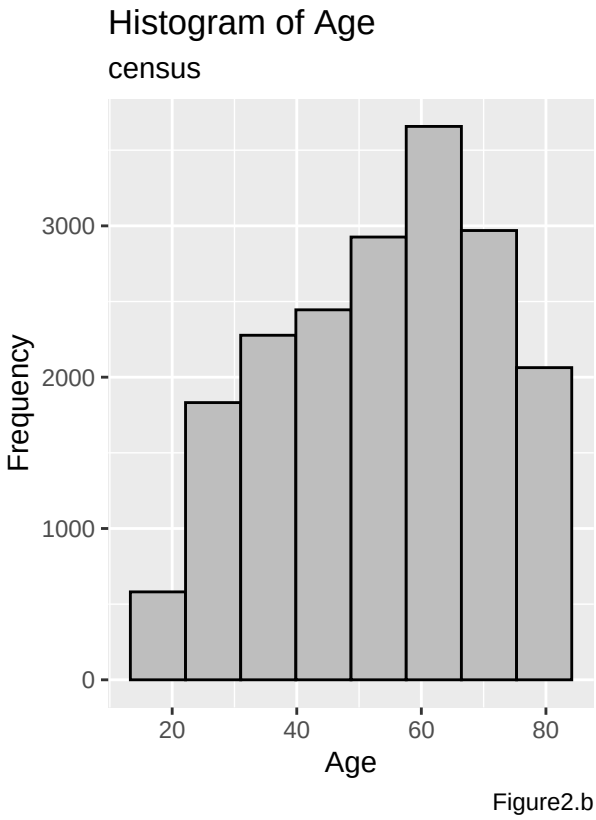

Figure2.b

This histogram (Figure 2. a,b) shows the distribution of age in survey and census. From the plot, both distributions of age in survey and census are unimodal. However, the distribution of age in the survey is slightly right-skewed, with most of the values concentrated around 50. And the distribution of census is slightly left-skewed, with most of the values concentrated around 58. And the distribution of age in the census is a little flatter. Because the amount of data contained in the census is a bit larger.

According to the Demographic population of Canada in 2020 (Jeudy, 2020), there is a large proportion of people between the ages of 50 and 64, which is consistent with the age distribution in Figure 2. So, the distribution about the plot and table are reasonable. By combining Table 2 and Figure 2. a,b, the percentage of age is different, so we include age in the model to predict whether different ages affect the outcome of the voting.

Predictor Variable: Education Level (categorical)

This is a summary information about the education level:

Table 3: Proportion of Education level in survey data and census data

Education Level	Survey	Census	Percentage of Survey	Percentage of Census
College or University	1955	9933	0.6484245	0.52976
HighSchool or Below	563	7164	0.1867330	0.38208
Master or Beyond	497	1653	0.1648425	0.08816

Table 3 shows the specific values of the education level in survey data and census data. There are 3015 participants in survey data, and 18750 in census data. Among the three levels, the proportion of College or University is the largest in both two data, which is about 64.84% and 52.98%. Then followed by HighSchool or Below and Master or Beyond. This implies that more than half of participants with College or University education levels in both two data.

Further analysis in more detail will be carried out in connection with the relevant graphs below.

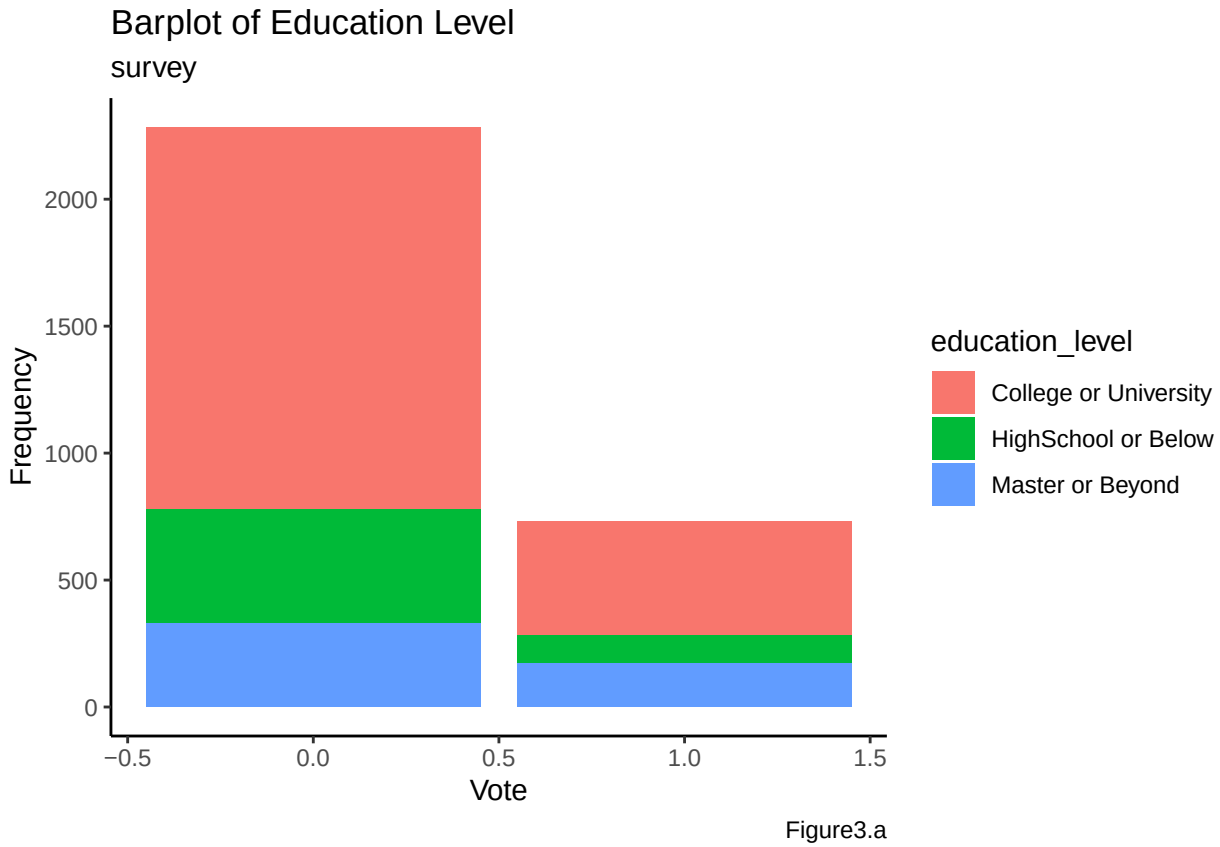

Figure3.a

This bar chart (Figure 3. a) further compares the education levels of the voting choices in the survey. Both for those who voted Libertarian and those who did not, the percentage of those whose education level was college or university was large. For participants who did not vote Libertarian, more than half have an education level of College or University, followed by those with an education level in Highschool or Below, and the smallest is Master or Beyond.

On the other hand, for participants who do not vote Libertarian, the percentage of those with an education

level of HighSchool or Below is higher than the percentage of those with a Master or Beyond. But for participants who vote Libertarian, the proportions of these two levels are almost equal.

By Combination of Table 3 and Figure 3.a,b,c (Figure 3.b,c explain the detailed distribution information about education level in survey and census data in Appendix section), the proportion of education levels are different, which means that education level does affect the voting results. The education level will be considered in setting modeling as a predictor variable is reasonable.

Predictor Variable: Income level (categorical)

This is a summary information about the income level:

Table 4: Proportion of Information Level in survey and census data

Income Level	Survey	Census	Percentage of Survey	Percentage of Census
Less than $ 25,000	345	2444	0.1144279	0.1303467
$ 25,000 to $ 49,999	449	3954	0.1489221	0.2108800
$ 50,000 to $ 74,999	537	3370	0.1781095	0.1797333
$ 75,000 to $ 99,999	403	2676	0.1336650	0.1427200
$ 100,000 to $ 124,999	390	2009	0.1293532	0.1071467
$ 125,000 and more	891	4297	0.2955224	0.2291733

Table 4 shows the specific counts of the income level in survey data and census data. There are 3015 participants in survey data, and 18750 in census data. Among the six levels, the number of income levels with $ 125,000 and more is the largest in both two data. And the least income level is Less than $ 25,000 in both data. And the proportion is similar for other income levels.

Further analysis in more detail will be carried out in connection with the relevant graphs below.

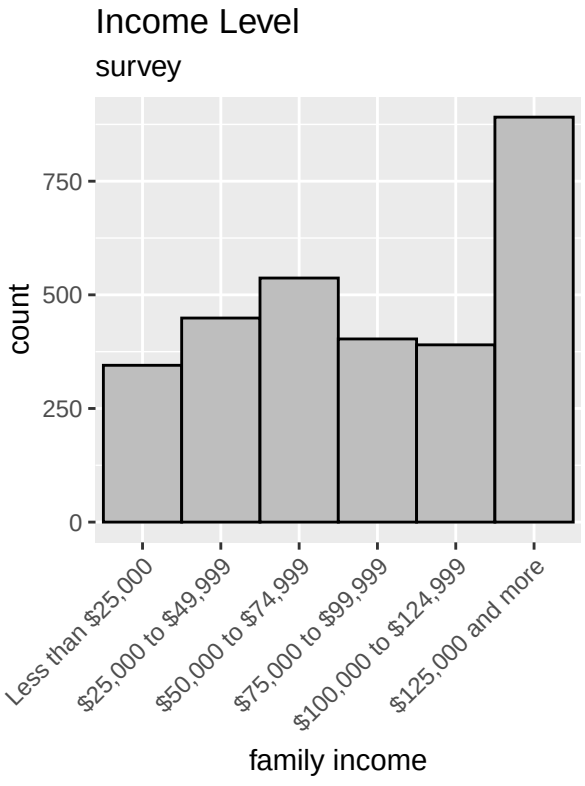

Figure4.a

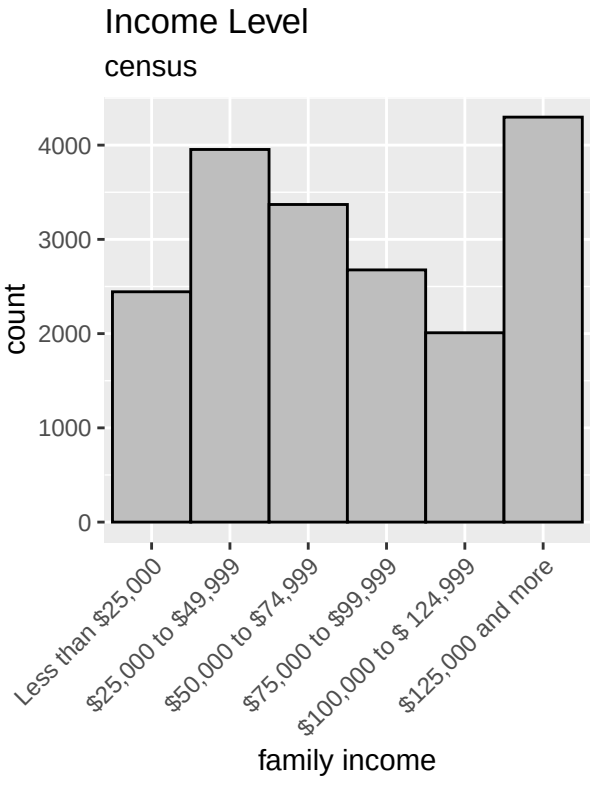

Figure4.b

This bar chart (Figure 4. a,b) shows the distribution of participants' family income levels in survey data and census data. As shown in the ranking of income levels from small to large, the overall distribution of both datasets is basically similar, except for the level of $ 25,000 to $ 49,999. The highest income level option $ 125,000 and more, accounts for the largest proportion, while the other income levels account for a relatively even split. However, in the survey data, the proportion of income levels from $ 25,000 to $ 49,999 is significantly lower than the proportion in the census.

Therefore, it can be seen that the distribution of income levels in the survey is not exactly the same as in the census, so the variable of income levels will be considered in the post-stratification in the next method section. It is possible that the sample collected in the survey is incomplete and does not contain all cases, but the distribution of the other parts is very similar, and the distribution of income levels is in line with common sense. Thus, it is reasonable for the distribution of income level from plot and table.

By the combination of Table 4 and Figure 4.a,b, the income level may have some influence on the voting. Therefore, income level will be a predictor variable in the model and determine in the result section whether the income level has an impact on the voting result based on the p-value of the income level.

Predict Variable: Province (categorical)

This is a summary information about the province:

Table 5: Proportion of Province in survey data and census data

Province	Survey	Census	Percentage of Survey	Percentage of Census
Alberta	198	1548	0.06567164	0.08256000
British Columbia	589	2247	0.19535655	0.11984000
New Brunswick	151	1249	0.05008292	0.06661333
Newfoundland and Labrador	146	1034	0.04842454	0.05514667
Ontario	601	5064	0.19933665	0.27008000
Prince Edward Island	153	650	0.05074627	0.03466667
Saskatchewan	208	1034	0.06898839	0.05514667
Manitoba	197	1064	0.06533997	0.05674667
Nova Scotia	157	1344	0.05207297	0.07168000
Quebec	615	3516	0.20398010	0.18752000

Table 5 shows the specific counts of the province in survey data and census data. There are 3015 participants in survey data, and 18750 in census data. Among the 10 provinces, the number of participants from Ontario is the largest in both two data, which take about 19.93% in survey data and 27.01% in census data. The difference between count and proportion of survey and census are slightly different in British Columbia and Quebec.

The plots about provinces are explained in Appendix section, by combination of Table 5. The analysis in more detail will be carried out in connection with the relevant graphs below.

The plots (Figure 5. a,b) about the province explain the distribution about the province of survey data and census data, and Figure 5.d,e shows the comparison of participants who vote Liberals and who did not, which in the Appendix section. These plots show the correlation between the province and votes, and there is a huge gap in votes for the different provinces, it is meaningful for us to research the voting in different provinces.

In the 2019 federal general election, Fournier indicates that 79% of aged 65 to 74 years old people vote liberal but 57% of voters age between 18 to 34. Then, it shows there has a correlation between voting and age. Furthermore, based on Kurtzleben (2016), points out that there has a huge gap of partisan for highly educated and non-highly educated which more highly educated students support the Liberal. Then, it means the education and the liberal voter have a linear relationship. Also, Hopper (2021) shows that "wealthier Canadians are turning out to be one of the Liberals' most reliable constituencies." It illustrates that income and the support Liberal has a linear relationship. Moreover, according to CBC-News (2019), Ontario and Quebec have better voting than others, which the provinces have a correlation with the voting. Therefore, we analyzed these variables and included them as predictor variables in our model.

Methods

Our goal of the research is to predict the proportion of Canadian who will vote for the Liberal Party in the coming 2025 federal election. In order to do so, we have to first build a model and then estimate the parameters in the model based on the **2019 Canadian Election Study** data set.

Because the possible outcomes of voting the Liberal Party would only be *yes* or *no*, with this kind of binary variable type, it is suitable to use a logistic regression model.

Model Specifics

The logistic regression model we are using is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_{2,edu_i} x_{edu_i} + \beta_{3,inc_j} x_{inc_j} + \beta_{4,prv_k} x_{prv_k}$$

For: $i = 1, 2,\ j = 1, 2, 3, 4, 5,\ k = 1, 2, 3, 4, 5, 6, 7, 8, 9$

Where:

- p is the probability of the a qualified Canadian voting for the Liberal Party.

- x_{age} is the numerical variable for people's age.
- x_{edu_1} is the indicator variable for people having highest education level of college or university. It equals to 1 if the person's highest education level is college or university, and it equals to 0 if the person has a different education level.
- x_{edu_2} is the indicator variable for people having highest education level of master or beyond. It equals to 1 if the person's highest education level is master or beyond, and it equals to 0 if the person has a different education level.
- x_{inc_1} is the indicator variable for people having family income between $ 100,000 ~ $ 124,999. It equals to 1 if the person's family income is between $ 100,000 ~ $ 124,999, and it equals to 0 if the person has a different family income level.
- x_{inc_2} is the indicator variable for people having family income $ 125,000 and more. It equals to 1 if the person's family income is $ 125,000 and more, and it equals to 0 if the person has a different family income level.
- x_{inc_3} is the indicator variable for people having family income between $ 25,000 ~ $ 49,999. It equals to 1 if the person's family income is between $ 25,000 ~ $ 49,999, and it equals to 0 if the person has a different family income level.
- x_{inc_4} is the indicator variable for people having family income between $ 50,000 ~ $ 74,999. It equals to 1 if the person's family income is between $ 50,000 ~ $ 74,999, and it equals to 0 if the person has a different family income level.
- x_{inc_5} is the indicator variable for people having family income between $ 75,000 ~ $ 99,999. It equals to 1 if the person's family income is between $ 75,000 ~ $ 99,999, and it equals to 0 if the person has a different family income level.
- x_{prv_1} is the indicator variable for people in British Columbia. It equals to 1 if the person is in British Columbia, and it equals to 0 if the person residents in other provinces.
- x_{prv_2} is the indicator variable for people in Manitoba. It equals to 1 if the person is in Manitoba, and it equals to 0 if the person residents in other provinces.
- x_{prv_3} is the indicator variable for people in New Brunswick. It equals to 1 if the person is in New Brunswick, and it equals to 0 if the person residents in other provinces.

- x_{prv_4} is the indicator variable for people in Newfoundland and Labrador. It equals to 1 if the person is in Newfoundland and Labrador, and it equals to 0 if the person residents in other provinces.
- x_{prv_5} is the indicator variable for people in Nova Scotia. It equals to 1 if the person is in Nova Scotia, and it equals to 0 if the person residents in other provinces.
- x_{prv_6} is the indicator variable for people in Ontario. It equals to 1 if the person is in Ontario, and it equals to 0 if the person residents in other provinces.
- x_{prv_7} is the indicator variable for people in Prince Edward Island. It equals to 1 if the person is in Prince Edward Island, and it equals to 0 if the person residents in other provinces.
- x_{prv_8} is the indicator variable for people in Quebec. It equals to 1 if the person is in Quebec, and it equals to 0 if the person residents in other provinces.
- x_{prv_9} is the indicator variable for people in Saskatchewan. It equals to 1 if the person is in Saskatchewan, and it equals to 0 if the person residents in other provinces.

- β_0 represents the intercept of the model, and it is the expected log odds of voting for the Liberal Party when the individual is at age 0 with the education level of high school or below and with family income of less than $ 25,000. This is meaningless to us because the qualified voting age is at least 18 (Elections in Canada, 2021).
- β_1 represents the slope of age in our model. For every one unit increase in age, we expect a β_1 increase log odds of voting for the Liberal Party, for a fixed **education level** and **family income level** and **province located**.
- β_{2,edu_1} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals with the education level of university or college equivalent **to** individuals with the education level of high school or below, for a fixed **age** and **family income level** and **province located**.
- β_{2,edu_2} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals with the education level of master or above **to** individuals with the education level of high school or below, for a fixed **age** and **family income level** and **province located**.
- β_{3,inc_1} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals with the family income level of $ 100,000 ~ $ 124,999 **to** individuals with the family income level of less than $ 25,000, for a fixed **age** and **education level** and **province located**.
- β_{3,inc_2} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals with the family income level of $ 125,000 and more **to** individuals with the family income level of less than $ 25,000, for a fixed **age** and **education level** and **province located**.
- β_{3,inc_3} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals with the family income level of $ 25,000 ~ $ 49,999 **to** individuals with the family income level of less than $ 25,000, for a fixed **age** and **education level** and **province located**.
- β_{3,inc_4} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals with the family income level of $ 50,000 ~ $ 74,999 **to** individuals with the family income level of less than $ 25,000, for a fixed **age** and **education level** and **province located**.
- β_{3,inc_5} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals with the family income level of $ 75,000 ~ $ 99,999 **to** individuals with the family income level of less than $ 25,000, for a fixed **age** and **education level** and **province located**.
- β_{4,prv_1} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals located in British Columbia **to** individuals located in Alberta, for a fixed **age** and **education level** and **family income level**.

- β_{4,prv_2} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals located in Manitoba **to** individuals located in Alberta, for a fixed **age** and **education level** and **family income level**.
- β_{4,prv_3} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals located in New Brunswick **to** individuals located in Alberta, for a fixed **age** and **education level** and **family income level**.
- β_{4,prv_4} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals located in Newfoundland and Labrador **to** individuals located in Alberta, for a fixed **age** and **education level** and **family income level**.
- β_{4,prv_5} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals located in Nova Scotia **to** individuals located in Alberta, for a fixed **age** and **education level** and **family income level**.
- β_{4,prv_6} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals located in Ontario **to** individuals located in Alberta, for a fixed **age** and **education level** and **family income level**.
- β_{4,prv_7} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals located in Prince Edward Island **to** individuals located in Alberta, for a fixed **age** and **education level** and **family income level**.
- β_{4,prv_8} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals located in Quebec **to** individuals located in Alberta, for a fixed **age** and **education level** and **family income level**.
- β_{4,prv_9} represents the expected difference in log odds of voting for the Liberal Party when comparing individuals located in Saskatchewan **to** individuals located in Alberta, for a fixed **age** and **education level** and **family income level**.

We use this model because we have a binary outcome for voting the Liberal Party, 'yes' or 'no', and our survey data satisfies the logistic model assumption in the following ways:

- First, we have sample including 3015 individual records, which is quite large (Statistics Solutions, 2021).
- Second, the respondents in the survey are being randomly selected and we believe that they are independent to each other (Statistics Solutions, 2021).

After we have calculated those estimates of β and their p-values from the Result section, we will check if there is any individual or a specific group-associated non-significant predictors (i.e. x) in the models by comparing their p-values with the significance level of 0.05. If the situation happens, we will have another reduced-variable logistic model and will use AIC and BIC criterion to select the optimal model (Prabhat, 2010).

Model selection criteria such as AIC and BIC are commonly employed. The acronyms AIC and BIC stand for Akaike's Information Criteria and Bayesian Information Criteria, respectively. Despite the fact that both phrases refer to model selection, they are not interchangeable. When comparing the Bayesian and Akaike's Information Criteria, the penalty for additional parameters is higher in the BIC than in the AIC (Prabhat, 2010).

Post-Stratification

After having our logistics model ready and have those estimates of β, we will use the model to predict the proportion of Canadian voting the Liberal Party in the coming 2025 federal election based on the matching version of **2017 GSS Data**. The method we will use is called post-stratification.

The process of post-stratification is that:

First, we will separate the population into unique groups, that each individual will belong to their specific group that has unique **age**, **education level**, **family income level** and **province** characteristics. For example, one of the group would be **people with the age of 21 from British Columbia that have university degrees(highest education level they have completed) and have family income between $25,000 ~ $49,999**.

Second, for each group, we are going to use the model to predict the proportion of voting the Liberal Party for those people. We will denote those predicted values by $\hat{p}_{i,j}$, and it means the estimated probability of choosing the Liberal Party for the people from the jth group. We will denote the jth group's estimate as $\hat{p}_j$. Note that for a specific group, group members have the exactly same information regarding to **age**, **education level**, **income level** and **province**. This means that they will have the same estimate from the model. For a kth group, $\hat{p}_{1,k} = \hat{p}_{2,k} = \cdots = \hat{p}_{n,k}$.

Note that in here we are using $\hat{p}_{i,j}$ as the estimated probability of an individual to vote the Liberal Party. But from the logistic regression model, we could only derive $\hat{y}_{i,j}$ directly, which is equal to $log(\frac{p}{1-p})$. Therefore, before doing the post-stratification process, we need to convert $\hat{y}_{i,j}$ into $\hat{p}_{i,j}$.

At this moment, we will have the information of **each group's estimated proportion of voting the Liberal Party** and **the size of these group**. We will use those two information to calculate the weighted average at the population level, which would be the estimated proportion of the population would vote for the Liberal Party, and we will denote this weighted average by $\hat{p}^{PS}$.

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

Here:

- N_j means the group size for the jth group
- $\hat{p}_j$ means the average probability of voting the Liberal Party for the people from jth group

The post-stratification technique is useful here because the large data set, **2017 GSS Data**, does not reflect the information about people's voting preference, and we have to use the smaller data set,**2019 Canadian Election Study**, which contains that information to train the model and will later to be used on **2017 GSS Data** for predicting the proportion of people that will vote for the Liberal Party at a national level. Also, post-stratification accounts for the weights of the individual groups, and this would result more precise probability estimation rather than averaging the entire estimation without the account of weights.

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

Results

Table 6: Summary of Measures for the Model_1

term	estimate	std.error	statistic	p.value
(Intercept)	-2.9265546	0.3140567	-9.3185537	0.0000000
age	0.0092933	0.0028020	3.3167007	0.0009109
education_levelCollege or University	0.1624274	0.1222927	1.3281853	0.1841169
education_levelMaster or Beyond	0.6306254	0.1491304	4.2286843	0.0000235
income_family$100,000 to $124,999	0.1008622	0.1850685	0.5449995	0.5857539
income_family$125,000 and more	0.2541115	0.1595997	1.5921802	0.1113442
income_family$25,000 to $49,999	0.1842731	0.1774821	1.0382629	0.2991477
income_family$50,000 to $74,999	0.2253546	0.1700820	1.3249762	0.1851790
income_family$75,000 to $99,999	0.3474495	0.1784921	1.9465822	0.0515848
provinceBritish Columbia	0.6893179	0.2491231	2.7669776	0.0056579
provinceManitoba	0.8116465	0.2846134	2.8517504	0.0043479
provinceNew Brunswick	0.8677603	0.2971807	2.9199749	0.0035006
provinceNewfoundland and Labrador	1.1057776	0.2930766	3.7729991	0.0001613
provinceNova Scotia	1.0839622	0.2898761	3.7393983	0.0001845
provinceOntario	1.3251016	0.2433671	5.4448677	0.0000001
provincePrince Edward Island	1.1315627	0.2901025	3.9005613	0.0000960
provinceQuebec	0.9086574	0.2466025	3.6847047	0.0002290
provinceSaskatchewan	-0.0648017	0.3206921	-0.2020684	0.8398632

From Table 6, we can see that p-values of β_0, β_{age}, β_{2,edu_2}, β_{3,prv_1}, β_{3,prv_2}, β_{3,prv_3}, β_{3,prv_4}, β_{3,prv_5}, β_{3,prv_6}, β_{3,prv_7}, β_{3,prv_8} are less than 0.05. In other word, for the hypothesis test of $H_0 : \beta_i = 0$ vs $H_a : \beta_i \neq 0$ for $i = 0, age, edu_2, prv_1, prv_2, prv_3, prv_4, prv_5, prv_6, prv_7, prv_8$, there is strong evidence to reject the alternative hypothesis of $H_a : \beta_i \neq 0$ at the significance level of 0.05.

On the other hand, we noticed that the p-values of the remainder parameters are larger than 0.05. This means for the hypothesis test of those β, we do not have the evidence to reject the alternative hypothesis of $H_a : \beta_i \neq 0$ at the significance level of 0.05 for $i = edu_1, inc_1, inc_2, inc_3, inc_4, inc_5, prv_9$. Especially for all the β associated with family income levels.

Hence, we will form a reduced model that does not include the family income as one of our predictor. The reduced model is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_{2,edu_i} x_{edu_i} + \beta_{4,prv_k} x_{prv_k}$$

Where all the β interpretations would be same but without the consideration of family income.

For example, β_0 represents the intercept of the model, and is the expected log of odds of voting for the Liberal Party when the individual is at age 0 with the education level of high school or below. This is meaningless to us because the qualified voting age is at least 18 (Elections in Canada, 2021).

Table 7: Summary of Measures for the Model_2

term	estimate	std.error	statistic	p.value
(Intercept)	-2.7161738	0.2844300	-9.5495349	0.0000000
age	0.0089191	0.0027594	3.2322579	0.0012282
education_levelCollege or University	0.1785351	0.1210101	1.4753741	0.1401120
education_levelMaster or Beyond	0.6720132	0.1441064	4.6633131	0.0000031
provinceBritish Columbia	0.6862876	0.2490231	2.7559192	0.0058527
provinceManitoba	0.8099982	0.2841973	2.8501263	0.0043702
provinceNew Brunswick	0.8680408	0.2965993	2.9266448	0.0034264
provinceNewfoundland and Labrador	1.0991018	0.2924684	3.7580187	0.0001713
provinceNova Scotia	1.0712419	0.2891753	3.7044722	0.0002118
provinceOntario	1.3257813	0.2432394	5.4505200	0.0000001
provincePrince Edward Island	1.1232324	0.2891799	3.8841993	0.0001027
provinceQuebec	0.8949072	0.2460158	3.6376007	0.0002752
provinceSaskatchewan	-0.0586531	0.3203811	-0.1830731	0.8547407

Table 7 is the summary table for the reduced model. We can see that most of the p-values are less than 0.05 which means that we have passed the p-value test for those parameters at the significance level of 0.05. However, this does not imply that our reduced model will have a better fit for the data set. In order to test out if the reduced model fits better than the original model, we will use the AIC and BIC criterion (Prabhat, 2010).

Table 8: AIC and BIC for the two models

Model	AIC	BIC
Original	3251.372	3359.576
Reduced	3246.343	3324.491

From Table 8, for the reduced model, we can see both the AIC and BIC measures are smaller than those of the original model. Therefore, by the AIC and BIC criterion, the reduced model fits better and we will use this model in the following analysis (Prabhat, 2010).

Back to Table 7 for the reduced model:

- $\hat{\beta}_0$ = -2.7161738 this means that the estimated log of odds of voting for the Liberal Party when the individual is at age 0 with the education level of high school or below is -2.7161738. In other word, the probability of voting for the Liberal Party when the individual is at age 0 with the education level of high school or below is 0.0620257. This is meaningless to us because the qualified voting age is at least 18 (Elections in Canada, 2021).
- $\hat{\beta}_1$ = 0.0089191 means that for every one unit increase in age, by estimation, there is a 0.0089191 increase log odds of voting for the Liberal Party, for a fixed **education level** and **province located**.
- $\hat{\beta}_{2,edu_1}$ = 0.1785351 is the estimated difference in log odds of voting for the Liberal Party when comparing individuals with the education level of university or college equivalent **to** individuals with the education level of high school or below, for a fixed **age** and **province located**.
- $\hat{\beta}_{2,edu_2}$ = 0.6720132 is the estimated difference in log odds of voting for the Liberal Party when comparing individuals with the education level of master or above **to** individuals with the education level of high school or below, for a fixed **age** and **province located**.

- $\hat{\beta}_{4,prv_1} = 0.6862876$ is the estimated difference in log odds of voting for the Liberal Party when comparing individuals located in British Columbia **to** individuals located in Alberta, for a fixed **age** and **education level**.
- $\hat{\beta}_{4,prv_2} = 0.8099982$ is the estimated difference in log odds of voting for the Liberal Party when comparing individuals located in Manitoba **to** individuals located in Alberta, for a fixed **age** and **education level**.
- $\hat{\beta}_{4,prv_3} = 0.8680408$ is the estimated difference in log odds of voting for the Liberal Party when comparing individuals located in New Brunswick **to** individuals located in Alberta, for a fixed **age** and **education level**.
- $\hat{\beta}_{4,prv_4} = 1.0991018$ is the estimated difference in log odds of voting for the Liberal Party when comparing individuals located in Newfoundland and Labrador **to** individuals located in Alberta, for a fixed **age** and **education level**.
- $\hat{\beta}_{4,prv_5} = 1.0712419$ is the estimated difference in log odds of voting for the Liberal Party when comparing individuals located in Nova Scotia **to** individuals located in Alberta, for a fixed **age** and **education level**.
- $\hat{\beta}_{4,prv_6} = 1.3257813$ is the estimated difference in log odds of voting for the Liberal Party when comparing individuals located in Ontario **to** individuals located in Alberta, for a fixed **age** and **education level**.
- $\hat{\beta}_{4,prv_7} = 1.1232324$ is the estimated difference in log odds of voting for the Liberal Party when comparing individuals located in Prince Edward Island **to** individuals located in Alberta, for a fixed **age** and **education level**.
- $\hat{\beta}_{4,prv_8} = 0.8949072$ is the estimated difference in log odds of voting for the Liberal Party when comparing individuals located in Quebec **to** individuals located in Alberta, for a fixed **age** and **education level**.
- $\hat{\beta}_{4,prv_9} =$ -0.0586531 is the estimated difference in log odds of voting for the Liberal Party when comparing individuals located in Saskatchewan **to** individuals located in Alberta, for a fixed **age** and **education level**.

Most of the parameter estimates do not have straight forward meanings. But it is crucial to calculate those estimates and use these with the model to have a prediction on the entire population.

Table 9: Postratification Estimate

Estimate	Value
Vote Prop for Liberal Party	0.2426362

From Table 9, through the post-stratification process, we estimate that 24.2636224% proportion of the Canadian would vote for the Liberal Party in the coming 2025 federal election.

Our prediction seems reasonable in a way that even if the Liberal Party wins the election it is likely that it wins other parties without a significant gap in the number of votes, since the other political parties are also very competitive. However, our prediction for the 2025 election is constrained by the available data set. For a potential better estimation, we would suggest using the data being collected in the future time.

To answer our research goal, having 24.2636224% proportion of the Canadian voters would provide solid benefits to the Liberal Party in the election. But with this percentage, we could only say that the chances to get elected is pretty high.

Conclusions

The Canadian election is very critical to Canadian citizens. Different ruling parties have different emphases on social issues. Choosing the appropriate ruling party can better deliver services to the people, such as health care, education, and security. The Liberal party's leader, Justin Trudeau, has made big contributions to the epidemic, such as providing free vaccines to Canadians. And he guaranteed more money to improve the vaccine passport system (Canada election, 2021). It shows that the Liberal Party cares about citizens' lives seriously. Therefore, in this project, we believe that the Liberal party can win the 2025 federal election. In order to predict the probability of the Liberal Party winning the Canadian election in 2025 based on the 2019 Canadian Election Study data and 2017 GSS data, we use the logistic regression model to measure the proportion of supporting the Liberal Party according to the voters' age, income, education level, and province located. Then, we use the p-value, BIC, and AIC to determine whether the logistic model is appropriate or not. As the ratio of the number of people who are being characterized by the levels of the variables is different between census and survey data, we would also need to use the Post-Stratification to determine the proportion of voting Liberal party at a population level. Furthermore, income is not a proper predictor based on p-value, BIC, and AIC. Therefore, we use voters' age, education level, and province location as predictors to predict the outcome of supporting the Liberal Party. Next, according to the Post-Stratification, we estimate that a 24.2636224% proportion of Canadians would vote for the Liberal Party in the coming 2025 federal election. In other words, more than one-quarter of Canadian citizens support Liberal Party, which is quite an optimistic number. Also, it is reasonable because according to other polls, approximately half are satisfied with the Liberal Party, and about twenty-seven percent are willing to continue supporting Justin Trudeau (Canadian Federal Politics, 2021).

Limitation

There are several limitations to this project. First, since we use the 2017 census data and 2019 survey data, there are some mismatches on income, and province. These variables are dynamic, which means people may have different incomes in two years and some people may move to other provinces. All these unpredictable factors can lead to imprecision in the calculation and estimation steps. Another limitation is that in order to match two data, we only choose 10 provinces in Canada to predict the voting on Liberal, but three territories are not included, which are Northwest Territories, Nunavut, and Yukon. This is because these 3 territories were not included in the residency question from the 2017 census data. Although the population in these territories is quite small, the combined population ratio does not exceed 0.5% among the Canadian population (Province and territories, 2021), it is still essential to know their voting so that the outcome can be more accurate. But for statistical and analytical reasons, we have to do so in order to perform post-stratification techniques. Moreover, for the survey and census, only select interviewees from people who have phone calls and this might result in potential biases to our estimation result. For example, some people who do not have phones would not be interviewed and are less likely to be drawn from the population poll. Lastly, we have used a logistic regression model and have satisfied several assumptions. But besides intuition, we did not check the multicollinearity assumption between independent variables in a statistical way. It is mainly because we do not acquire the proper tool with the given time frame (Statistics Solutions, 2021).

Future study

In the future study, with the research goal of predicting the chances of winning the next Canadian federal election in 2025 for the Liberal Party, one would find more updated datasets and use that as a source to perform the analysis and prediction. Also, besides variables such as education, income, age, and province, one can try other variables as predictors, comparing different models by certain model selection techniques. Some of the interesting variables could not be observed in the available data set and as the data set updates throughout the time, one could also consider those for modeling. Moreover, one can further check the model assumption of independent variables' multicollinearity by their valid methods. If a significant violation happens, one could try other assumption-satisfied models and calculate the outcomes. Thus, in future studies, one could further improve this research by observing and using updated data, trying different models and including different predicting variables.

Bibliography

Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/

Barr, C., Bray, A. , Baumer, B., Çetinkaya-Rundel, M., Diez, D., Ismay, C., Kim, A. Y. & Paterno, K.(2021). *openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs. R package version 2.2.0*. https://CRAN.R-project.org/package=openintro

Canadian electoral system. Wikipedia. (2021, September 23). Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/Canadian_electoral_system

Canada election: Complete list of promises made during the 2021 campaign. Global News. Retrieved November 04, 2021, from https://globalnews.ca/news/8106833/canada-election-promises-made-2021/

Canadian Federal Politics - August 3, 2021. (2021, August 3). Leger. Retrieved November 04, 2021, from https://leger360.com/surveys/legers-north-american-tracker-august-3-2021/

Chan, K. (2021, October 1). *More people moved to BC than anywhere else in Canada over the past year: statistics*. DailyHive. https://dailyhive.com/vancouver/canada-interprovincial-migration-statistics-october-2021

Connolly, A. (2021, November 3). *COP26: Trudeau says world needs a 'standard' for pricing carbon. What might that look like?* Global News. https://globalnews.ca/news/8340498/cop26-justin-trudeau-global-carbon-price/

Coughlan, S. (2019, September 26). *The symbolic target of 50% at university reached*. BBC News. Retrieved November 04, 2021, from https://www.bbc.com/news/education-49841620

Cycle 31 : Families Public Use Microdata File Documentation and User's Guide. (2020, April). Computing in the Humanities and Social Sciences. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

Elections in Canada. Wikipedia. (2021, October 2). Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/Elections_in_Canada.

Fournier, P. J. (2021, April 25). *338Canada: The Liberals are winning over older-normally Conservative-voters*. Macleans.ca. https://www.macleans.ca/politics/ottawa/338canada-the-liberals-are-winning-over-older-normally-conservative-voters/

Gibbs, A., Stringer, A.(2021). 16.2 Canadian Election Study. *Probability, Statistics, and Data Analysis* . University of Toronto Press.

Grenier, E. (2019, October 22). *Ontario and Quebec keep Liberals in power and Conservatives out*. CBC News. https://www.cbc.ca/news/politics/grenier-election-results-1.5330105

Grolemund, G. (2014, July 16). *Introduction to R Markdown*. R Studio. https://rmarkdown.rstudio.com/articles_intro.html

Guide to the Canadian House of Commons. (n.d.). Parliament of Canada. https://learn.parl.ca/sites/Learn/default/en_CA/Guide-to-the-Canadian-House-of-Commons

Harell, A.,Loewen, P. J. , Rubenson, D., Stephenson, L. B. (2020). *CES 2019 Phone Technical Report Final*. https://doi.org/10.7910/DVN/8RHLG1.

Hpooer, T. (2021, September 10). *ELECTION INSIGHTS: Why rich Canadians are all-in for the Liberals*. nationalpost. https://nationalpost.com/news/canada/election-insights-why-rich-canadians-are-all-in-for-the-liberals

Jeudy, L. (2021,July 6). *Resident population of Canada in 2020, by age*. Statista. https://www.statista.com/statistics/444868/canada-resident-population-by-age-group/

Kurtzleben, D. (2016, April 30). *Why Are Highly Educated Americans Getting More Liberal?* npr. https://www.npr.org/2016/04/30/475794063/why-are-highly-educated-americans-getting-more-liberal

Pedersen, L.T. (2020). *patchwork: The Composer of Plots*. https://patchwork.data-imaginist.com, https://github.com/thomasp85/patchwork.

Prabhat, S. (2010, October 3). *AIC and BIC*. Difference Between. http://www.differencebetween.net/miscellaneous/difference-between-aic-and-bic/

Provinces and territories of Canada. Wikipedia. (2021, October 26). Retrieved November 03, 2021, from https://en.wikipedia.org/wiki/Provinces_and_territories_of_Canada

Random Digit Dialling (RDD). djsresearch. (n.d.). Retrieved November 4, 2021, from https://www.djsresearch.co.uk/glossary/item/Random-Digit-Dialling-RDD.

Registered Political Parties and Parties Eligible for Registration – Elections Canada. (n.d.). Elections Canada. https://www.elections.ca/content.aspx?section=pol&dir=par&document=index&lang=e

Statistics Solutions. (2021, August 11). *Assumptions of Logistic Regression*. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

Tables. (n.d.). *R Markdown*. https://rmarkdown.rstudio.com/lesson-7.html

The General Social Survey: An overview. Government of Canada, Statistics Canada. (2017, February 27). Retrieved November 5, 2021, from https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm#a2.

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

2021 Canadian federal election. Wikipedia. (2021, November 1). Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/2021_Canadian_federal_election.

45th Canadian federal election. Wikipedia. (2021, November 4). Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/45th_Canadian_federal_election

Appendix

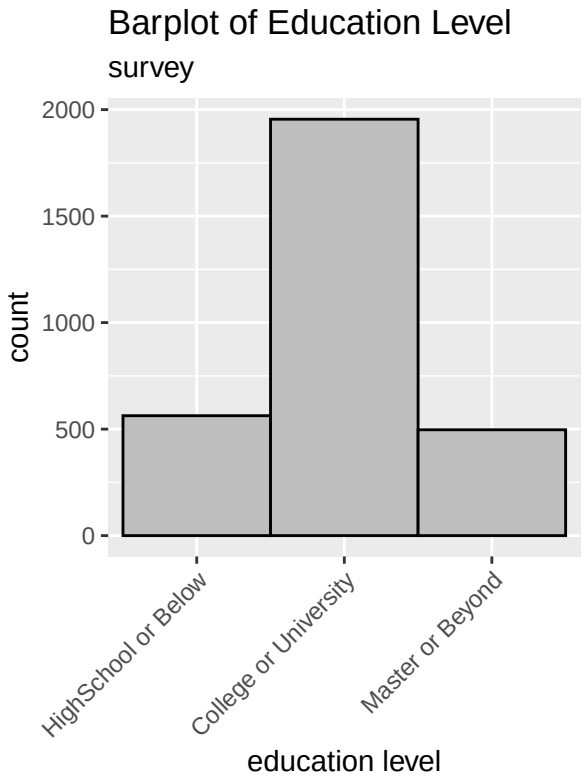

Figure3.b

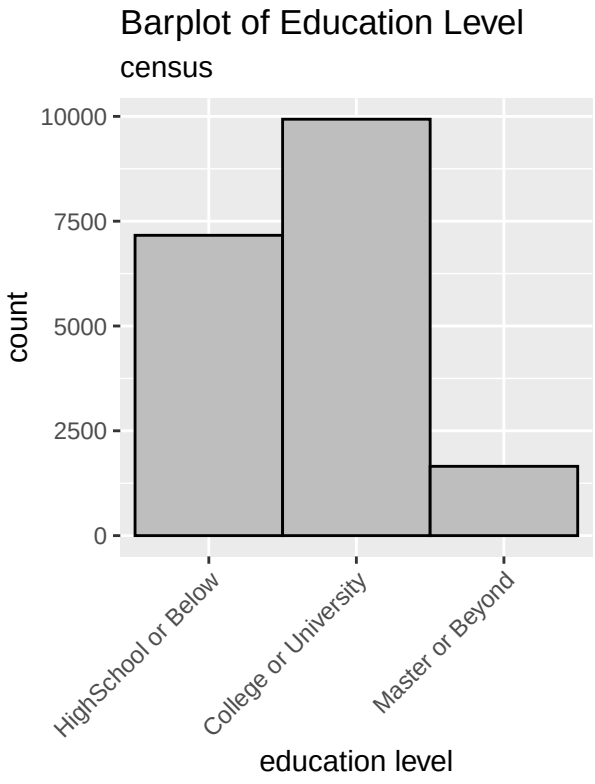

Figure3.c

These two bar plots (Figure 3.b,c) show the distribution of people's education levels in the survey data and census data. Both datasets showed that the participants had the largest proportion of education level in College or University, followed by High school or Below and Master or Beyond. However, the proportion of HighSchool or Below in census data is more than in survey. And the proportion of participants with Master or Beyond in the survey is slightly higher than the census.

Thus, from the plots (Figure 3.b,c), the distributions of the two data are still different, and the percentage of participants with HighSchool or Below included in the survey is smaller than the percentage of the census. This means that the sample collection in the survey may not contain the full situations and is incomplete. Thus, education level will be considered in the post-stratification in the next method section.

According to BBC News (Coughlan, 2019), we can learn that the percentage of people who do have a college or university degree is now almost 57%, which is in line with Figure 3. This illustrates the validity of our graphs and data, and we can use that for analysis and prediction in the method section.

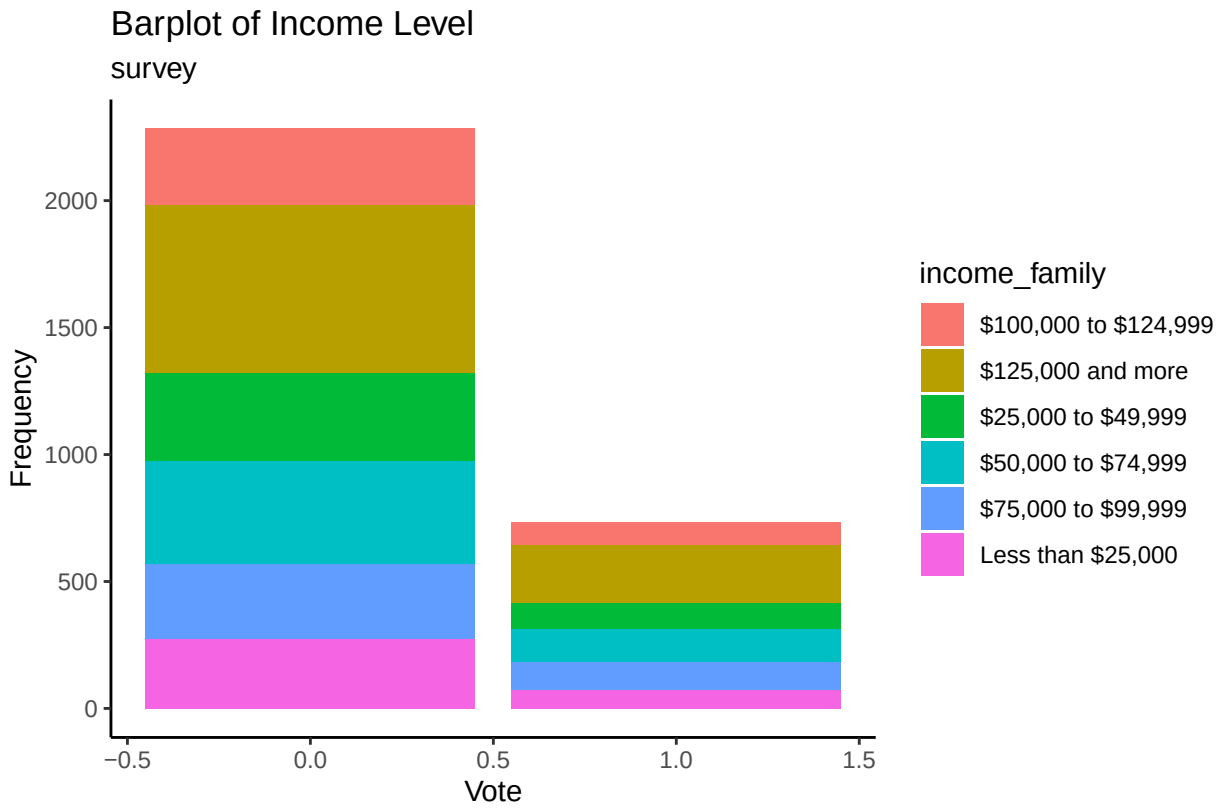

Figure4.c

This bar chart (Figure 4. c) further compares the income levels of the voting choices in the survey. This plot shows that these 6 different income levels are relatively similar in proportion to voting liberal and non-voting liberal, with no huge differences in proportion. But since the previous image (Figure 4.a,b) shows that the distribution of income levels in the survey and census are not exactly the same, the survey is not fully representative of the census. Therefore, the income level may have some influence on the voting. Therefore, income level can be a predictor variable in the model and determine in the result section whether the income level has an impact on the voting result based on the p-value of the income level.

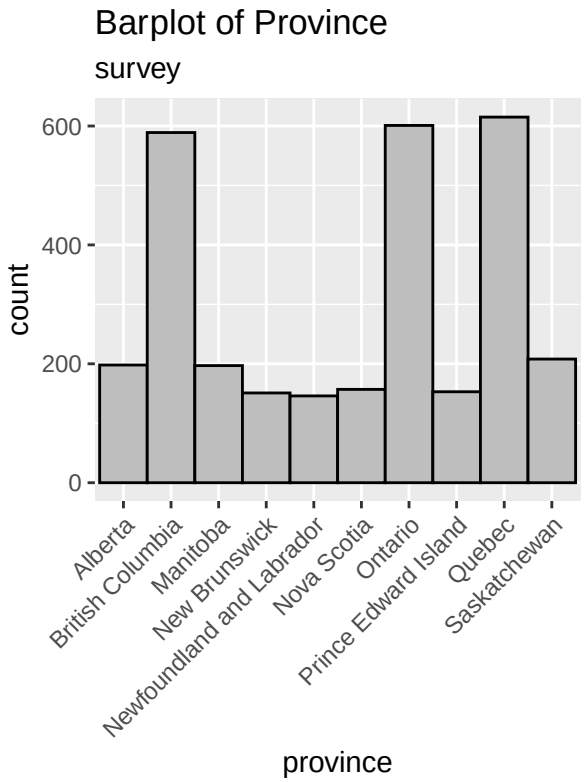

Figure5.a

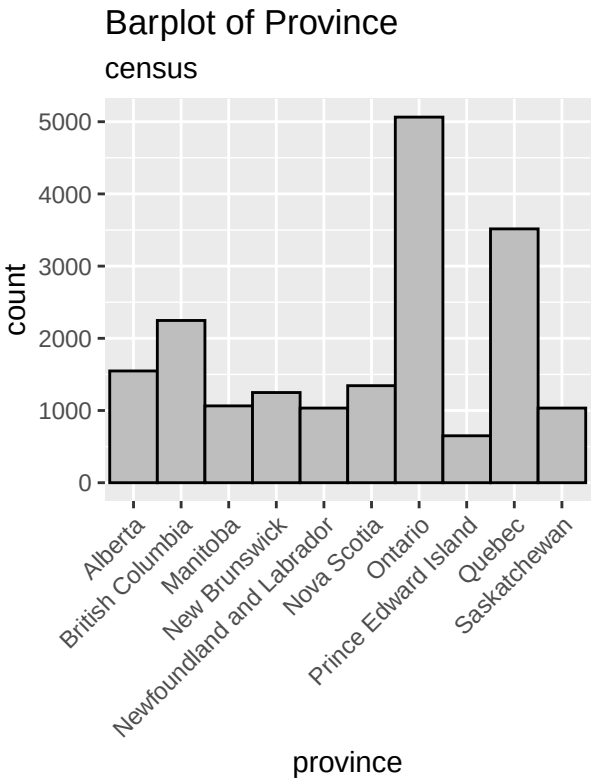

Figure5.b

These bar plots (Figure 5. a,b) show the distribution of observation from different provinces in survey and census data. Since the number of provinces in both survey and census data should be matched, there are 10 provinces selected. For the survey plot (Figure 5.a), there are three provinces that have high populations, which are British Columbia, Ontario, and Quebec. These three provinces have three times as many voters as other provinces supporting the Liberal party. However, the Census data (Figure 5. b), shows Ontario has the highest population and Quebec also has far more people than other provinces. The third-largest province is British Columbia, but the population is close to others compared to Ontario and Quebec.

The difference between the two plots (Figure 5. a,b) is the number of populations in British Columbia. The census dataset is the year 2019 and the survey data is about the year 2017. During two years, it is reasonable that the population has some changes. According to Chan (2021), he points out that in 2019, there are more people moved in B.C province. Then, it illustrates the reason why more people move to British Columbia.

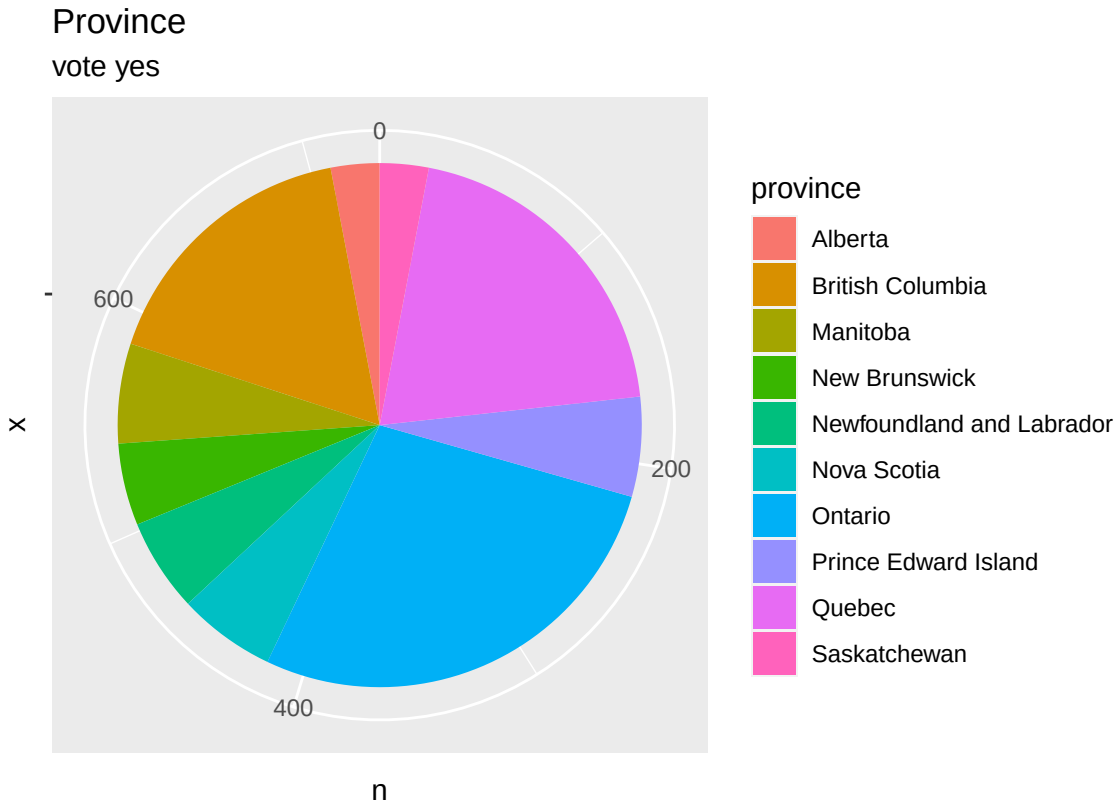

Figure5.c

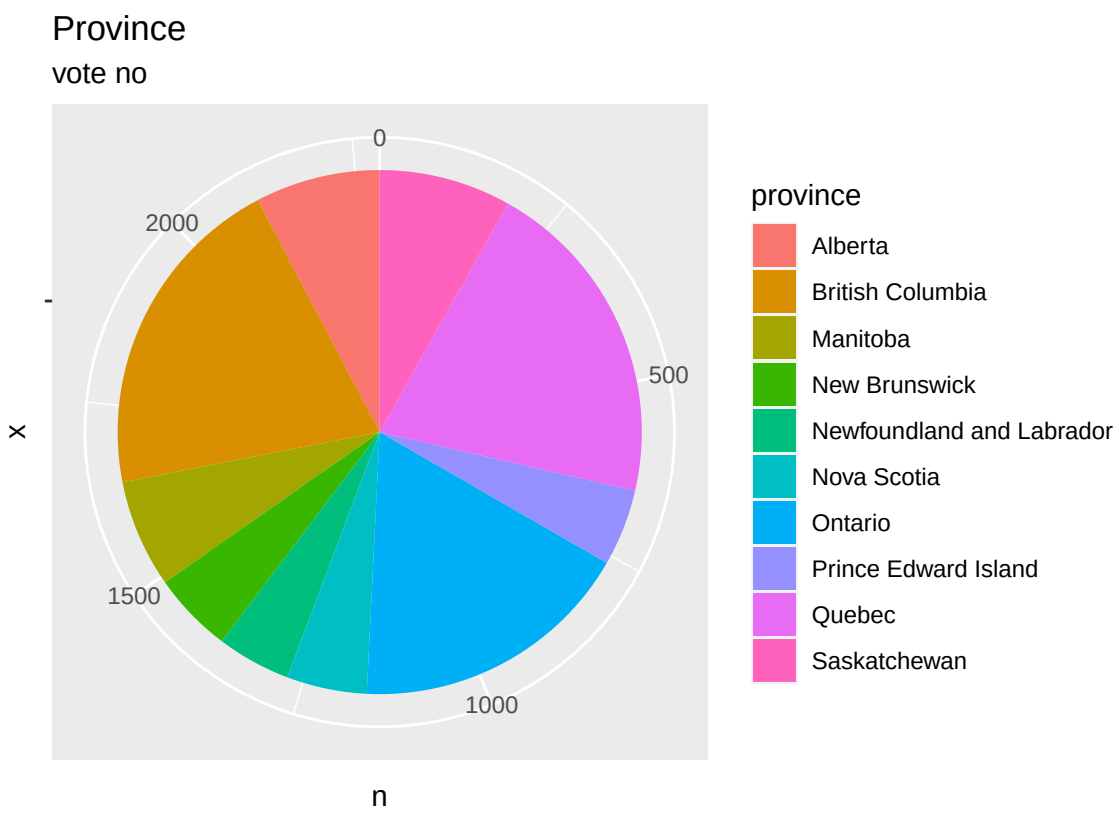

Figure5.d

Figure 5.c shows the voting in different provinces on the Liberal party. It represents the situation in which people vote Liberal. Ontario has the largest proportion which is about a quarter of the Liberal votes and the second largest in Quebec. British Columbia also has a larger number of votes. For the provinces in which people do not vote Liberal (Figure 5.d), Quebec and British Columbia have the highest proportion among 10 provinces. Ontario is below them but the proportion is not small. Since Ontario, Quebec, and British Columbia have the most populous in Canada, it is reasonable that these provinces have a high number of votes and non-votes.

Since the province and votes have a correlation, and there is a huge gap in votes for the different provinces, it is meaningful for us to research the voting in different provinces.