The Impact of Voters' Demographic Factors in Canadian Federal Elections

STA304 - Assignment 2

Group 109: Isha Sharma, Jenny Oh, Manjima Banerjee

November 24, 2022

Introduction

Forecasting elections has been prominent in political science for the past 40 years. Once supposed to be a part of research and education, forecasting has recently grabbed the attention of rival online and media forecasters. It is integral for the public to be aware of the right predictions for elections as it may subconsciously impact their baseline expectations and ideologies before placing a vote [1]. Therefore, it is key to ensure usage of appropriate statistical models, correct variables, and relevant analysis for predictions. From the political side, forecasting can allow political parties to assess areas of improvement for increasing their chances of success.

Along the same vein, voting behavior, a form of electoral behavior, has been a central concern for political scientists to assess how and why decisions were made by political leaders [2]. Several studies show how Canada's elections are majorly affected by an individual voter's demographic factors. Thus, in order to make electoral predictions, we must take into account factors such as gender, race, religion, age, region of residence, wealth etc. These factors act as determinants for the result of elections for political parties [3].

The objective of this study is to find the overall popular vote for the Canadian federal election by focusing on the likelihood of individuals to vote for Liberal Party. Forecasting the votes using statistical models can help us not only get an idea about the future and what we can expect in terms of the political atmosphere of a country, but also give us insights on the factors that may possibly influence a party's vote. Through this study, we can determine the outcome as well as the uncertainty that comes with our prediction. In a deeper sense, we are able to understand the behavior of voters and their thoughts around governance changes due to certain changes in their personal lives. Lastly, when political conversations face-to-face or on social media become more of an argument rather than a discussion, we may be able to add value to these discussions about the entire voting process, as well as the electoral results, by conducting our analyses.

We are given the census data (General Social Survey - GSS)[4] and survey data (Canadian Election Study 2019 - CES2019) [5]. The survey data has demographic factors of the individuals such as age, marital status, sex, and income, as well as questions related to their voting choices like 'Which party is your first choice to vote for the elections?'.

The research question for our analysis is: How do the demographic factors of voters affect the likelihood of voting for the Liberal Party?. As previously mentioned, studies show that demographic factors affect voter's values and their political attitudes. Our research question focuses on various factors of individual voters such as income, religion, language known, province, and level of education that will help us predict the tendency of an individual voter with specific demographic factors to vote for the Liberal Party. This analysis can help us interpret what factors may affect the overall voting pattern, the perspectives and behaviour of voters in Canada and lastly in determining what aspects to focus on when determining policies for the parties in order to increase their chances of winning. We hypothesize that demographic factors of individual affect their tendency to vote for the Liberal Party.

Data

There are two datasets used in this paper, the Canadian Election Study 2019 (CES2019)[5] and the General Social Survey (GSS)[4], which represent survey data and census data, respectively. The CES2019 data was collected from an online survey that surveyed Canadian citizens and permanent residents, aged 18 or older, before and after the elections in 2019 to determine Canadians' attitudes toward current events, elections, and democracy [6]. In the pre-election period, 37,822 responses were gathered via the survey and the post-election period had 10,337 responses [6]. The creators of the survey aimed to have a sample of 50% men, 50% women, 28% aged 18-34, 33% aged 35-54, 39% aged 55 and higher, 80% French within Quebec, 20% English within Quebec, 10% French within the Atlantic region, and 10% French nationally [6].

The GSS data was collected from telephone surveys, with telephone numbers gathered from the Address Registrar (AR), across Canada that focus on gathering data on social trends to keep track of changing living conditions and the well-being of Canadians [7]. Questions in the survey revolved around studying families and households, such as familial origins, organizational structure within a household, marital status, etc. The target population of the survey was everyone aged 15 years or older in Canada, except residents of the Yukon, Northwest Territories, and Nunavut and full-time residents of institutions [7]. The provinces were divided into strata based on geographic areas and a simple random sample without replacement was performed in each stratum [7]. The target sample size was 20,000 with 20,602 actually responding [7].

In order to determine what demographic factors we wanted to study for our research question, we first looked at both the datasets and made a list of the variables that were common to both datasets or could be cleaned in a way that allowed for use in both datasets. In addition, when selecting these variables we reasoned through ones that we believed might have an actual impact on voting outcome. For example, we believed age should be studied further since there were columns in both datasets that gave us this data and we thought that older people were more likely to vote for conservative parties over liberal parties. Based on this process, the variables we ended up with were age, sex, family income, province, education level, whether someone practiced a religion or not, marital status, languages spoken (English and French), and employment. For the response variable, we chose the variable "vote choice", which was indicative of what party someone said they would vote for. Although there were variables asking similar questions for both the census and survey data, the responses and names of the variables were not entirely identical. Below is a description of each variable and how we manipulated the datasets to arrive at each variable:

1. **Age**: *A numerical variable representing the age of the respondent.*
   The census data had decimals for the age variable so we rounded the ages into integers. For the survey data, there was a variable of 'year of birth' instead of age. Therefore, we subtracted the year of birth from 2019 (when the survey was conducted) in order to calculate ages. There were no NA values in either dataset.

2. **Sex**: *A categorical variable indicating whether the respondent is either male or female.*
   There were no NA values and only binary outcomes (male and female) for the census data so no further cleaning was needed. For the survey data, there was no sex variable, only a gender variable. We assumed that if someone's gender was listed as a woman, then their sex is female and similarly, men were listed as males. This is because we needed the same variable in both datasets. We converted gender(man, woman) to sex(male, female) and removed 'other' genders since there were no corresponding observations in the census data.

3. **Family Income**: *A categorical variable indicating that the respondent's family income falls within the given income range.*
   For the census data, there were 6 categories ("Less than $25,000", "$25,000 to $49,999", "$50,000 to $74,999", "$75,000 to $99,999", "$100,000 to $ 124,999", and "$125,000 and more") for income and no NAs so no further cleaning was needed. For the survey data, we used the numeric 'income' variable. We removed NA values and categorized the observations into the same 6 categories as the census data.

4. **Province**: *A categorical variable indicating the province that the respondent resides in.*
   There were no NAs for both the census and survey data. However, the census data included only provinces but not territories, while the survey data included both provinces and territories. Therefore,

we removed territory observations from the survey data.

5. **Education**: *A categorical variable indicating the highest level of education the respondent obtained at the time.*
   For the census data, there were 7 categories and NA values, while there were 12 categories without any NAs for the survey data. We classified similar categories for both the survey and census data into 6 new categories. First, we removed NAs from the census data. Then we converted "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" to "Bachelor's degree"; any of "College, CEGEP or other non-university certificate or di...", "Trade certificate or diploma", or "University certificate or diploma below the bachelor's level" to "College or certificate"; "High school diploma or a high school equivalency certificate" to "High school diploma"; "Less than high school diploma or its equivalent" to "Less than high school"; and "University certificate, diploma or degree above the bach..." to "Above bachelor" for the census data. For the survey data, we converted "No schooling", "Some elementary school", "Completed elementary school", "Some secondary/ high school" to "Less than high school"; "Completed secondary/ high school" to "High school diploma"; any of "Some technical, community college, CEGEP, College Classique", "Completed technical, community college, CEGEP, College Classique", or "Some university" to "College or certificate"; "Bachelor's degree" to "Bachelor's degree"; and "Master's degree" and "Professional degree or doctorate" to "Above bachelor". We removed observations that answered "Don't know/ Prefer not to answer" from the survey data.

6. **Religion**: *A binary, categorical variable indicating whether the respondent is religious or not.*
   For the census data we used the variable asking if they have religious affiliation, which had 3 categories and no NAs, as our religion variable indicator. For the survey data, we used the variable asking what their religion is, which had no NAs. For the census data, we removed the NA values and re-categorized respondents into whether they are religious or not based on their answers. If someone answered "Don't know", "No", or "No religious affiliation", we categorized them as "Not religious" and if someone answered "Has religious affiliation", we categorized them as "Religious". For the survey data, if someone answered "None/ Don't have one/Atheist" or "Don't know/ Prefer not to answer", we categorized them as "Not religious" and the other observations that listed specific religions were categorized as "Religious".

7. **Marital Status**: *A binary, categorical variable indicating whether the respondent has a partner or is single.*
   For both the census and survey data, there were the same 6 categories indicating a respondent's marital status, with the exception of the survey data having an extra "Don't know" category. First, we removed NAs from the census data and "Don't know" observations from the survey data. Then, for both datasets, we classified each category into a binary outcome: "Married" and "Living with a partner" as "Has a partner" and the other variables ("Widowed", "Separated", etc.) as "Single".

8. **Language Knowledge**: *A categorical variable indicating whether the respondent speaks English, French, or both.*
   For the census data, we used the variable asking about knowledge of Canada's official languages (French and English), which had 5 categories including a "Don't know" category and NA values. We removed the NAs and "Don't know" observations from the census data. For the survey data, we used two variables asking about their knowledge of English and/or French. We used the same four categories as the census data. Therefore, if someone knows both English and French, we categorized them into "Both English and French". If someone knows one of either English or French, we categorized them into "English only" or "French only", respectively. If someone knows neither English nor French, we categorized them into "Neither English nor French".

9. **Occupation**: *A binary, categorical variable indicating whether the respondent is employed or unemployed.*
   For the census data, we categorized people who stated their occupations as 'employed' and people who did not state their occupations as 'non-employed'. For the survey data, if someone answered "Working for pay full-time" or "Working for pay part-time", we categorized them into the "employed" group and the other groups were put into the "non-employed" group.

10. **Vote Liberal**: *A binary, categorical variable indicating whether the respondent indicated they would vote for the Liberal Party or not.*
    For the survey data, we used the variable asking about their potential voice choice for the next election. Since our research question is interested in whether people would vote for the Liberal party or not, we categorized people who said "Liberal Party" into the "Vote for Liberal" group and the other as the "Not Vote for Liberal" group.

Next, we decided to make some plots and tables of the various variables we were interested in. Some of them are shown below.

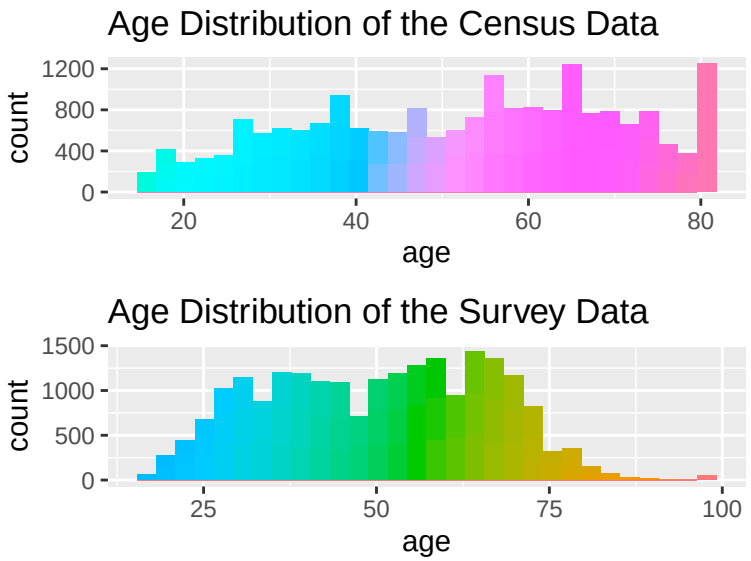

Figure 1: Histograms of age distributions in census data and survey data.

Figure 1 shows the distribution of the age variable for both the survey data and the census data. From the plot of census data, we see that there is a fairly even distribution among the different ages, with slightly higher numbers for ages 55 and higher. In contrast, the survey data seems to have more counts for ages between around 30 to 75 years when compared to the census data. In addition, there seems to be limited counts for ages greater than 75 for the survey data, whereas the census data has some of the highest counts for ages slightly above 80.

These highlighted differences indicate that the survey data may not be truly representative of the population. This is possibly further supported by the methods of data collection used for the survey and census data. As mentioned previously, the survey data was collected from an online survey and the census data was collected from telephone surveys. Although Internet usage has increased over the years amongst Canadian seniors, it is still well below the average for the overall Canadian population [8]. In addition, in 2017, more Canadian households had subscribed to mobile services over Internet services [9]. These two factors combined, decreased Internet usage amongst seniors and the overall greater prevalence of mobile services in households, could be a reason for more seniors partaking in the census data over the survey data. Thus, this would support our claim that the survey data may not be representative of the Canadian population.

Table 1: Voting for the Liberal Party based on sex.

	Female	Male
Vote for Liberal	0.706	0.697
Not Vote for Liberal	0.294	0.303

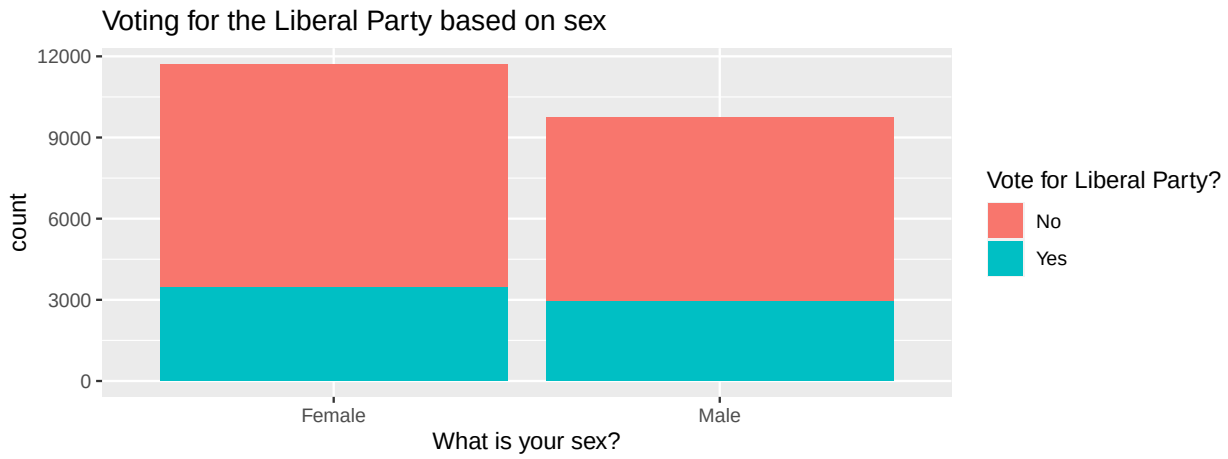

Figure 2: Bar plot of voting for the Liberal Party based on sex.

Figure 2 shows us the counts of each sex that responded to the CES survey and includes a breakdown of which party they voted for, either the Liberal Party or other parties. From the plot, we see that more people of the female sex responded to the survey and it seems that slightly more females than males indicated that they would vote for the Liberal Party. To further study the effects of sex as a demographic factor on the likelihood of voting for the Liberal Party. Table 1 was created to show the proportions of each group, females or males, and their likelihood of voting for the Liberal Party or not. The table shows us that around 70.5% of females and 69.7% of males surveyed are likely to vote for the Liberal Party, whereas 29.4% of females and 30.3% percent of males would not. These respective proportions are fairly similar between males and females, with less than 1% of a difference. Thus, it does not seem that sex as a demographic factor has much of an effect on a person's likelihood for the Liberal Party or some other party and so, it will not be studied further. Any other demographic variables, such as age, that were excluded from our final study were excluded for similar reasons, that is not having enough of a difference, less than 1%, between the various groups in the demographic factor.

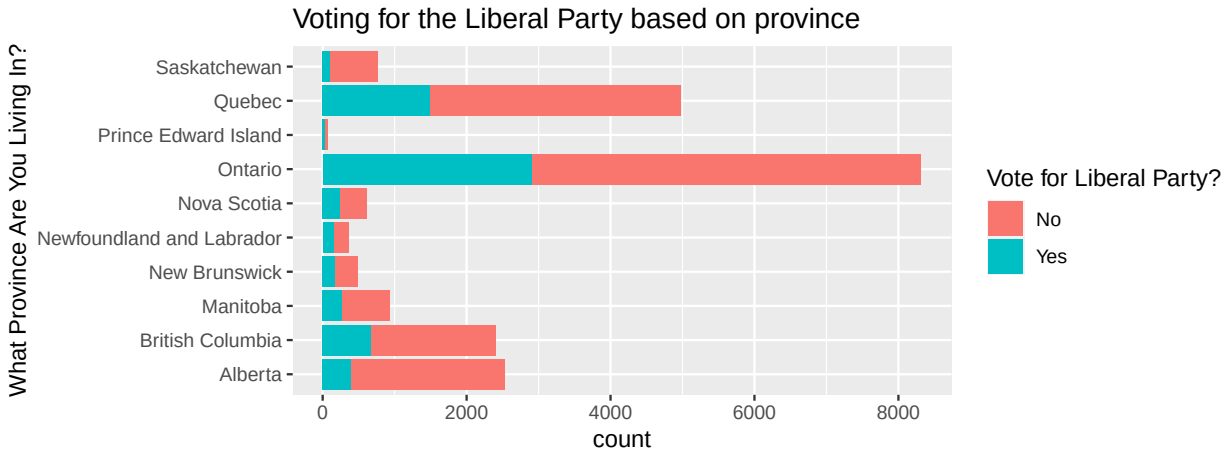

Figure 3: Bar plot of voting for the Liberal Party based on province.

Table 2: Voting for the Liberal Party based on province.

	AB	BC	MB	NB	NL	NS	ON	PE	QC	SK
Not Vote for Liberal	0.844	0.724	0.717	0.654	0.587	0.609	0.651	0.613	0.701	0.87
Vote for Liberal	0.156	0.276	0.283	0.346	0.413	0.391	0.349	0.387	0.299	0.13

Figure 3 shows us the numbers of people from each province that responded to the CES survey and whether they voted for the Liberal Party or not. From the plot, we see that Ontario had the highest number of respondents and that most of the provinces indicated that they were more likely to vote for other parties. The provinces of Newfoundland and Labrador and Nova Scotia seem to have more of an even split between voting for the Liberal Party or some other party. Table 2 gives us more of an insight into how a voter's province may affect their chances of voting for the Liberal Party. Across the provinces, we see that people are more likely to not vote for the Liberal Party, as indicated by the higher proportions of respondents for this. Moreover, we see that some provinces, such as Alberta and Saskatchewan, indicated by their provincial codes AB and SK, have a very high proportion of respondents indicating that they would not vote for the Liberal Party, 84.4% and 87% respectively.

On the other hand, as we predicted based on the plot, Newfoundland and Labrador had closer to an even split at 58.7% stating that they would not vote for the Liberal Party and 41.3% saying they would. Between the provinces, there is a wide range in likelihood of voting for the Liberal Party or not, ranging from 0.4% to 26%. With the exception of the voting proportion difference between Nova Scotia and Prince Edward Island, all of the proportion differences amongst the provinces are over 1%. This indicates that province, as a demographic factor, may have an effect on the likelihood of someone voting for the Liberal Party. Thus, it will be studied further along with any other demographic factors that have similar varied differences of at least over 1% between categories within a demographic factor.

We conducted EDA for all 9 predictors with the same process that we used for the factors of sex and province. From these results, we decided to include the variables of family income, province, level of education, whether someone is religious or not, and languages spoken in our analysis and not include age, marital status, and occupation. Please see the Appendix for additional plots and numerical tables. All analysis for this report was programmed using `R version 4.0.2`.

Methods

Based on our research question, we want to see how demographic factors affect a person's likelihood to vote for the Liberal Party in the next election. Since we are interested in seeing whether someone will vote for the Liberal Party or not, this is a binary outcome. As such, a logistic regression model will be most appropriate for our purposes since these are used in situations with binary outcomes. The logistic regression model is a statistical model that models the probability of an event happening by having the log-odds for the event be a combination of one or more independent variables. These models can be used to predict the value of a dependent variable, i.e. voting for the Liberal Party, based on one or more independent variables, i.e. the demographic factors we are interested in. Therefore, we will fit a logistic model to predict the likelihood of an individual voting for the Liberal Party.

Model Specifics

For our logistic regression model, the predictor variables are family income, province, level of education, whether someone is religious or not, and the languages they speak (English, French, or both), which was decided based on our exploratory data analysis (EDA). The response variable is the probability of an individual voting for the Liberal Party. The logistic regression model is shown here:

$$log(\frac{P}{1-P}) = \beta_0 + \beta_1 x_{income} + \beta_2 x_{province} + \beta_3 x_{education} + \beta_4 x_{religious} + \beta_5 x_{language}$$

Here, P represents the probability of an individual voting for the Liberal Party, the coefficients of β_1, β_2, β_3, β_4, and β_5 represent the change in log odds for one unit increase in x_{income}, $x_{province}$, $x_{education}$, $x_{religious}$, and $x_{language}$ respectively, and β_0 represents the intercept term.

Logistic regression models generally rely on a few assumptions. These are listed below and their applicability to our data is also explained.

1. **Independence of variables**
   This assumption has to ensure that our variables are not dependent on one another which signifies that a change in one variable should not affect another. If this is the case, it may affect our prediction for the likelihood of Liberal Party voting. In our case, our demographic factors that we have chosen as variables are all independent of one another and do not rely on each other.

2. **Influential Values/Outliers**
   Individual values that have the tendency to alter the outcomes of our regression model. These can be checked by using Cook's distance values. The outliers that we detect can be either removed or transformed to fit into log scale. Since the variables we are using for our logistic regression model are all categorical and we have removed any NA values, this assumption does not really apply.

3. **Multicollinearity**
   Multicollinearity occurs when two or more predictor variables are highly correlated or associated with one another. This suggests that variable columns are dependent on each other which might make the coefficients sensitive, making our model difficult to generalize. Generally, our predictor variables are not associated with each other. For example, languages spoken are not directly correlated with whether someone is religious or not or their family income. Any variables that could be associated with each other, such as level of education and its impacts on income, are not directly related and can be influenced by other variables, which may or may not be included in our model. Thus, this assumption is met.

4. **Outcomes are binary**
   Outcomes have to be binary for a logistic regression model to be applicable, that is there are only two possible values for our response variable. We see that our model consists of binary outcomes as people can either vote for the Liberal Party or not vote for it. Thus, we know that this assumption is met.

Post-Stratification

From the EDA, we noticed that the sample from the survey data may not be representative of the population. Thus, we need some way to account for this in the construction of our model. Poststratification is a technique that is used to correct model estimates when there are differences between a sample population and a target population, such as in our scenario. This process involves proportionally adjusting our estimates by matching a weighted sample cell count to the population cell count total.

We will use our logistic regression model with poststratification to assign different weights to our demographic factor variables such that the survey data becomes representative of the target population. We will poststratify each group and adjust the prediction, that is the probability that an individual in a certain demographic subgroup votes for the Liberal Party. based on the population size and the estimate of each cell. This is illustrated by the mathematical formula:

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

Here, $\hat{y_j}$ is the cell-level estimate, that is the probability that a person with the specified demographic factors in this subgroup of the population will vote for the Liberal Party, using the logistic model we built. N_j is the size of the j^{th} subgroup of the census data. We calculated $\hat{y}^{PS}$, which is the probability that an individual in the specified subgroup votes for the Liberal Party, adjusted for the target population and not the sample.

To implement these two outlined methods, we gathered the census data counts, as the population cell count, after grouping by the five demographic factors we outlined earlier. The logistic regression model was then

used to calculate the estimates, which was stated earlier to be the probability that someone from the subgroup specified by a row in our table would vote for the Liberaly Party. These estimates were then corrected for with poststratification by accounting for the census data counts and our five demographic factors and the data was put into a poststratification table that includes values for each of the five demographic factors and the corresponding probability of someone with those demographic factors voting for the Liberal Party. A section of this table can be seen in the next section, as well as in the Appendix.

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

Results

Table 3: Poststratification Table (in decscending order of proability of voting for the Liberal Party)

Household Income	Province	Education	Religious	Language	Probability of Voting for Liberal
$100,000 to $ 124,999	Newfoundland and Labrador	Above bachelor	Yes	Both English and French	0.5314416
$125,000 and more	Newfoundland and Labrador	Above bachelor	Yes	Both English and French	0.5228185
$75,000 to $99,999	Newfoundland and Labrador	Above bachelor	Yes	Both English and French	0.5192912
$100,000 to $ 124,999	Newfoundland and Labrador	Bachelor's degree	Yes	Both English and French	0.5163959
$100,000 to $ 124,999	Prince Edward Island	Above bachelor	Yes	Both English and French	0.5133301
$100,000 to $ 124,999	Newfoundland and Labrador	Above bachelor	Yes	English only	0.5095230
$100,000 to $ 124,999	Nova Scotia	Above bachelor	Yes	Both English and French	0.5082857
$125,000 and more	Newfoundland and Labrador	Bachelor's degree	Yes	Both English and French	0.5077524
$25,000 to $49,999	Newfoundland and Labrador	Above bachelor	Yes	Both English and French	0.5070588
$50,000 to $74,999	Newfoundland and Labrador	Above bachelor	Yes	Both English and French	0.5052659

After building our logistic regression model and performing poststratification on the data, we have estimates for the different $\hat{\beta}$ in the regression model (displayed in the Appendix), as well as a poststratification table. The poststratification table give us thousands of different subgroups of the population based on a combination of the five variables we focused on in our study: income, province, educational level, whether the person is religious or not, and whether they speak English, French, or both. The combination of these demographic factors yielded an estimate for each subgroup that gives us the probability of someone from this subgroup voting for the Liberal Party. The specific categories under each demographic factor have been listed previously, such as income ranging from $100,000 to 124,999, being from Ontario, etc.

From a portion of the data in Table 3, the poststratification table, ordered in descending order of probability of voting for the Liberal Party, we see that the subgroup of the population with the highest probability of containing someone that votes for the Liberal Party consists of the following factors: income ranging from $100,000 to $124,999, their province being Newfoundland and Labrador, education is above a Bachelor's degree, they are religious, and they speak both English and French. The likelihood that someone from this subgroup votes for the Liberal Party is 53.1%. Similar conclusions can be drawn from each subgroup by looking at the appropriate demographic factors for that group and the estimate, which is the probability of voting for the Liberal Party. In general, from the first 15 rows, we see that those with the highest likelihood

of voting for the Liberal Party generally have an income of $100,000 or more, come from Newfoundland and Labrador, have a degree higher than a Bachelor's, are religious, and speak both English and French.

These results seem reasonable as generally Newfoundland and Labrador tends to lean towards the Liberal Party, as evidenced by the 2015 federal election [10]. Thus, it would make sense that someone from this province has a higher probability than other provinces to vote for the Liberal Party. In addition, Atlantic Canada leans towards the Liberal Party as a whole, so it would make sense to have people from Nova Scotia and Prince Edward Island also have a higher probability of voting for the Liberal Party. In the past, higher family incomes have been linked to a greater tendency to vote more liberally [11] and so, higher income groups in Newfoundland and Labrador or any Atlantic province would have a greater chance of voting for a liberal group such as the Liberal Party. While there are no distinctly similar links between the educational background, being religious, or speaking a certain language and voting more in one direction, Newfoundland and Labrador has a predominantly religious population, with a majority being Christian [12], so the incidence of religion in our subgroups makes sense. Thus, overall, our results do seem to make sense with certain subgroups with a combination of our studied demographic factors having a greater chance of voting for the Liberal Party as a result of these factors.

Conclusion

The research question that we specified focused on whether demographic factors of an individual voter affect their tendency of voting for the Liberal Party. We hypothesized that demographic factors do indeed impact a voter's likelihood of voting for the Liberal Party. To test this, we used a logistic regression model that included five predictor variables of family income, province, educational level, whether the voter is are religious, and languages they speak (English, French, or both). These variables were used to find the probability of individuals in specific demographics subgroups voting for the Liberal Party in the next Canadian Federal Election in 2025.

In addition to this, we performed poststratification on our survey data to ensure our survey data becomes representative of the population by reweighting the population from the census data. This allowed us to come up with predictions about how likely voters with different combinations of demographic factors were to vote for the Liberal Party. From our estimates, we see that the highest probability of an individual voting for the Liberal party is from a subgroup with an income ranging from $100,000 to $124,999, from the province Newfoundland and Labrador, education is above a Bachelor's degree, they are religious, and speak both English and French. Using our table, we see that the probability of an individual from this group will vote for the Liberal party is 53.1% A few common demographic factors for the first few groups with the highest probability are - they are all religious from Newfoundland and Labrador, know both English and Fresh, have an income above $100,000 and their education level is Bachelor's degree or above.

These factors are reasonably interpreted by our model as the past elections give a similar pattern of results where people with higher incomes tend to vote for the Liberals [11]. We can find that Newfoundland and Labrador elected six Liberal and one Conservative party leaders and similar results are seen in Nova Scotia where eight Liberals and only three Conservative party leaders won [13]. This analysis can be done for all respective subgroups from our dataset and can be confirmed with other existing data from previous election cycles. For example, another section of the data in the Appendix shows randomized selections from our posstratification table and we can see that votes from provinces such as Ontario, British Columbia, and Quebec are much less likely to vote for the Liberal Party, which is generally true as these provinces have voted for the Conservative Party and the NDP in more recent elections [10]. Thus, our overall results do support our hypothesis and show us that demographic factors do play a part in predicting whether someone will vote for the Liberal Party.

A drawback of our model that may affect its reliability can be that we have not considered the effect of any correlated variables or confounding variables. An example of correlated variables from our dataset can be an individual's known languages and province, such as Quebec having a large French speaking population. There may also be other confounding variables, such as social issues, societal pressures, current events, etc., that may impact a demographic subgroup to vote for the Liberal Party aside from the actual demographic factors of the group. These types of variables could lead to our coefficients in the logistic regression model

being quite sensitive and overall, making it difficult to generalize our model.

A future analysis could involve checking for the highest probability estimates for subgroups and finding out what demographic factors may be generalized and show a common pattern. This can help us find a base trend for forecasting future elections. We can check for specific provinces and see which groups need more income-based and employment policies by looking at the income subgroups and assessing their probability of voting for the Liberal Party. Another future direction to explore would be to look into demographic factors for other parties, such as the Conservative Party or NDP, and see what the commonalities and differences are in demographic factors for subgroups that are more likely to vote for these other parties. In addition, we could further assess our conclusions in this paper by performing other statistical tests, such as a hypothesis test or conducting confidence intervals, to truly understand the effect of demographic factors on voting for the Liberal Party.

Bibliography

1. Writer, J. S. H. S. (2020, November 23). Students study past elections to predict future outcomes. Harvard Gazette. https://news.harvard.edu/gazette/story/2020/11/students-study-past-elections-to-predict-future-outcomes/

2. Voting Behaviour in Canada | The Canadian Encyclopedia. (n.d.). Www.thecanadianencyclopedia.ca. https://www.thecanadianencyclopedia.ca/en/article/electoral-behaviour

3. Dowding, K. (2020). Why Forecast? The Value of Forecasting to Political Science. PS: Political Science & Politics, 1–3. https://doi.org/10.1017/s104909652000133x

4. Canada, S. (2020). General Social Survey Cycle 31: Family, 2017. Abacus.library.ubc.ca. https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/G3DUFG

5. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study (CES) - Online Survey", https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1.

6. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study (CES) - Online Survey Codebook", https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1.

7. Canada, S. (2020). General Social Survey Cycle 31: Family, 2017. Abacus.library.ubc.ca. "User's Guide", https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/G3DUFG

8. Government of Canada, Statistics Canada. (2019). The Daily — Study: Evolving Internet Use Among Canadian Seniors. Statcan.gc.ca. https://www150.statcan.gc.ca/n1/daily-quotidien/190710/dq190710d-eng.htm

9. Government of Canada, Canadian Radio-television and Telecommunications Commission (CRTC. (2019). Communications Monitoring Report 2019 - Communications Services in Canadian Households: Subscriptions and Expenditures 2013-2017 | CRTC. Crtc.gc.ca. https://crtc.gc.ca/eng/publications/reports/policymonitoring/2019/cmr1.htm

10. Ibbitson, J. (2015, August 2). Canada's electoral geography: Where parties are likely to gain seats. The Globe and Mail. https://www.theglobeandmail.com/news/politics/canadas-electoral-geography-where-parties-are-likely-to-gain-seats/article25816225/

11. NW, 1615 L. S., Suite 800Washington, & Inquiries, D. 20036USA202-419-4300 | M.-8.-8. | F.-4.-4. | M. (2017, October 24). 10. Financial well-being, personal characteristics and lifestyles of the political typology. Pew Research Center - U.S. Politics & Policy. https://www.pewresearch.org/politics/2017/10/24/10-financial-well-being-personal-characteristics-and-lifestyles-of-the-political-typology/

12. Labrador, S., & John's Cma. (n.d.). Population by Religion and Sex. https://www.stats.gov.nl.ca/Statistics/Topics/census2011/PDF/REL_Religion_2011_NHS.pdf

13. Martin, L. (2021). Canada's 2021 federal election | Live results. The Globe and Mail. https://www.theglobeandmail.com/politics/federal-election/2021-results/

14. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

15. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

16. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

Appendix

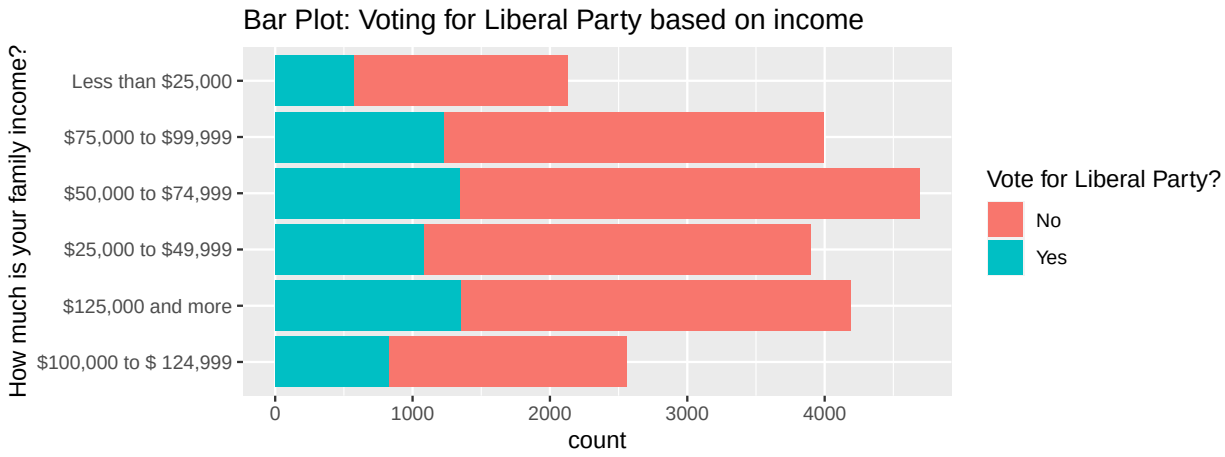

Figure 4: Bar plot of voting for the Liberal Party based on family income.

Table 4: Voting for Liberal Party Based on the Income Factor

	$100,000 to $ 124,999	$125,000 and more	$25,000 to $49,999	$50,000 to $74,999	$75,000 to $99,999	Less than $25,000
Not Vote for Liberal	0.678	0.677	0.724	0.713	0.694	0.73
Vote for Liberal	0.322	0.323	0.276	0.287	0.306	0.27

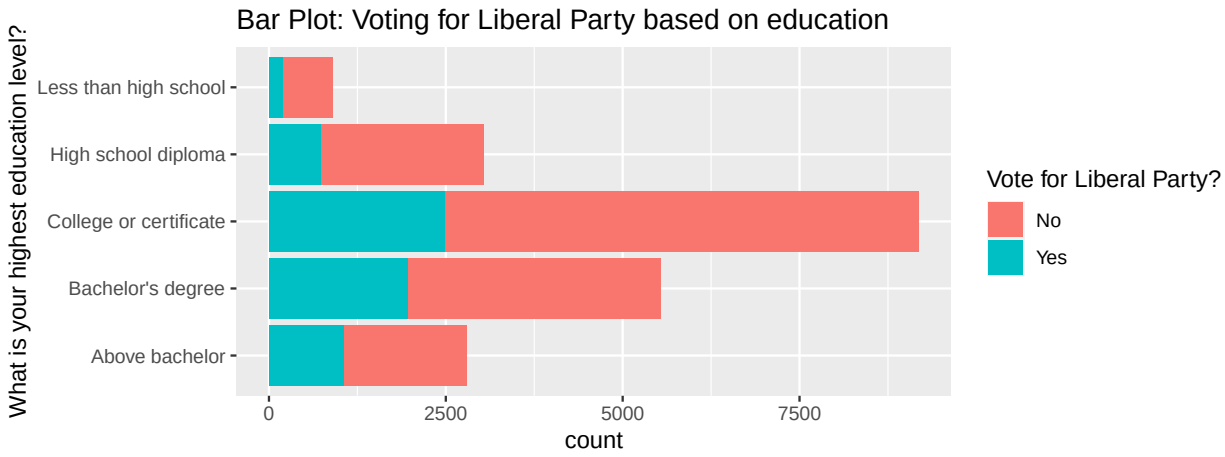

Figure 5: Bar plot of voting for the Liberal Party based on family education level.

Table 5: Voting for Liberal Party Based on the Education Factor

	Above bachelor	Bachelor's degree	College or certificate	High school diploma	Less than high school
Not Vote for Liberal	0.624	0.647	0.73	0.761	0.788
Vote for Liberal	0.376	0.353	0.27	0.239	0.212

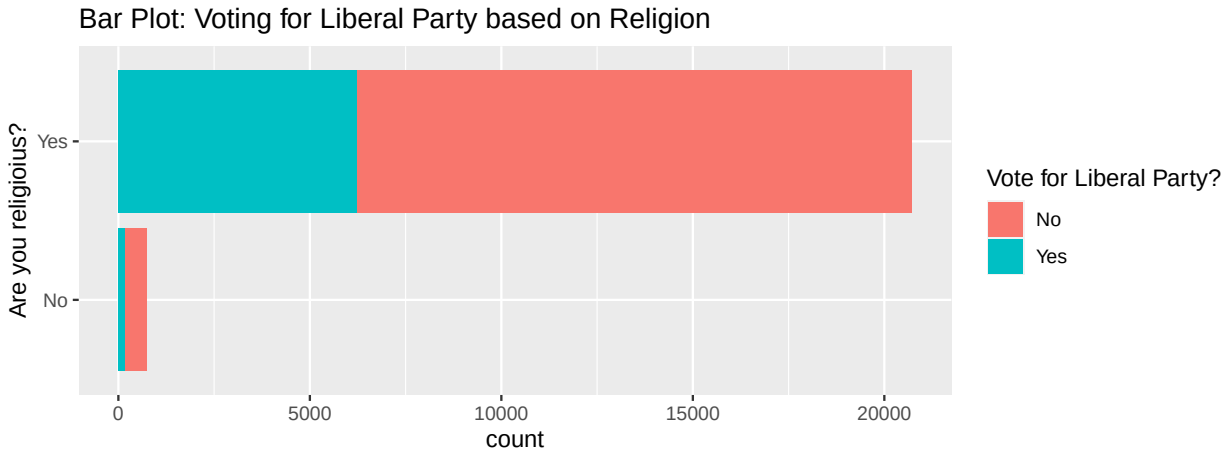

Figure 6: Bar plot of voting for the Liberal Party based on being religious or not.

Table 6: Voting for Liberal Party Based on the Religion Factor

	Not Religious	Religious
Not Vote for Liberal	0.767	0.699
Vote for Liberal	0.233	0.301

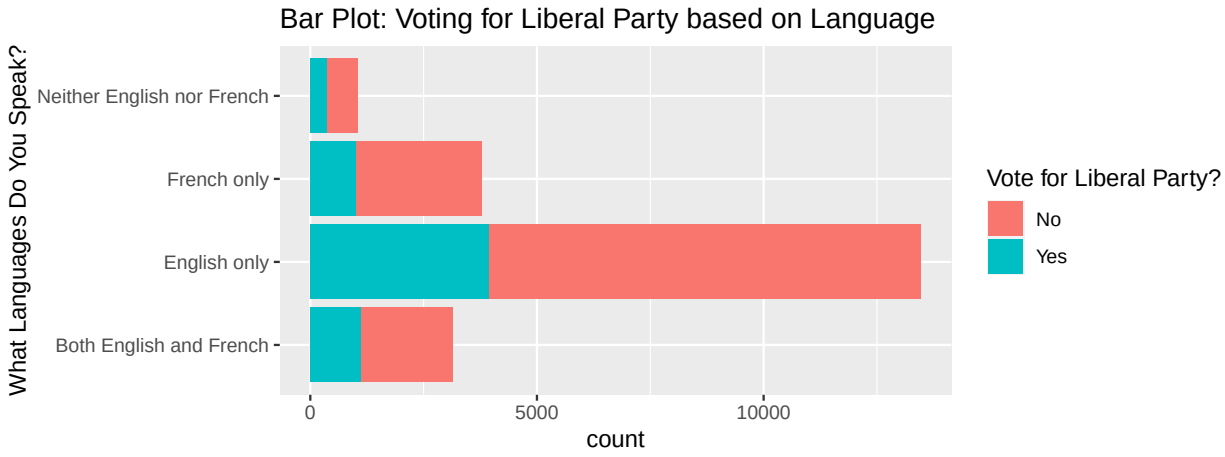

Figure 7: Bar plot of voting for the Liberal Party based on knowledge of English, French, or both languages.

Table 7: Voting for Liberal Party Based on the Language Factor

	Both English and French	English only	French only	Neither English nor French
Not Vote for Liberal	0.648	0.708	0.737	0.66
Vote for Liberal	0.352	0.292	0.263	0.34

This shows the estimated $\hat{\beta}_0$,$\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, and $\hat{\beta}_5$ values.

```
## # A tibble: 23 x 5
##    term                                estimate std.error statistic  p.value
##    <chr>                                  <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                          -1.53      0.123    -12.5   7.03e-36
##  2 income_family$125,000 and more       -0.0346    0.0550    -0.629 5.29e- 1
##  3 income_family$25,000 to $49,999      -0.0977    0.0576    -1.70  8.99e- 2
##  4 income_family$50,000 to $74,999      -0.105     0.0547    -1.92  5.50e- 2
##  5 income_family$75,000 to $99,999      -0.0487    0.0557    -0.875 3.81e- 1
##  6 income_familyLess than $25,000       -0.112     0.0671    -1.67  9.55e- 2
##  7 provinceBritish Columbia              0.711     0.0717     9.91  3.84e-23
##  8 provinceManitoba                      0.762     0.0915     8.33  8.08e-17
##  9 provinceNew Brunswick                 1.08      0.112      9.65  5.15e-22
## 10 provinceNewfoundland and Labrador     1.34      0.121     11.1   1.15e-28
## # ... with 13 more rows
```

Table 8: Post-stratification Table: Randomized Order

income_family	province	education	religious	language_knowledge	liberal_predict
$25,000 to $49,999	British Columbia	High school diploma	No	Both English and French	0.1869166
$50,000 to $74,999	Saskatchewan	College or certificate	Yes	Both English and French	0.1224582
$100,000 to $ 124,999	Manitoba	Less than high school	Yes	Both English and French	0.2405276
$75,000 to $99,999	Quebec	College or certificate	Yes	French only	0.2498104
$100,000 to $ 124,999	Ontario	Above bachelor	Yes	Both English and French	0.4512182
$25,000 to $49,999	Quebec	High school diploma	No	Both English and French	0.2526362
$75,000 to $99,999	Ontario	Less than high school	Yes	English only	0.2638005
Less than $25,000	Alberta	High school diploma	No	Neither English nor French	0.0984733
$125,000 and more	New Brunswick	Bachelor's degree	Yes	Both English and French	0.4414894
$100,000 to $ 124,999	Nova Scotia	Above bachelor	Yes	Both English and French	0.5082857
$75,000 to $99,999	Newfoundland and Labrador	High school diploma	No	English only	0.2941357
$125,000 and more	New Brunswick	College or certificate	Yes	English only	0.3345900
$100,000 to $ 124,999	British Columbia	Bachelor's degree	No	Both English and French	0.2923507
$50,000 to $74,999	Ontario	Less than high school	No	Both English and French	0.2123943

income_family	province	education	religious	language_knowledge	liberal_predict
$25,000 to $49,999	Newfoundland and Labrador	Bachelor's degree	No	Both English and French	0.4138601
$25,000 to $49,999	Ontario	College or certificate	No	Both English and French	0.2622601
Less than $25,000	British Columbia	High school diploma	Yes	English only	0.2216312
$100,000 to $ 124,999	British Columbia	College or certificate	Yes	Both English and French	0.2823987
$125,000 and more	Manitoba	High school diploma	Yes	Both English and French	0.2613047
$100,000 to $ 124,999	British Columbia	High school diploma	Yes	English only	0.2415190
$25,000 to $49,999	British Columbia	Less than high school	No	Both English and French	0.1658461
$100,000 to $ 124,999	Alberta	Above bachelor	No	Both English and French	0.1773841
$100,000 to $ 124,999	Manitoba	Above bachelor	Yes	Both English and French	0.3879835
$25,000 to $49,999	Prince Edward Island	Less than high school	No	English only	0.2419289
$100,000 to $ 124,999	Ontario	Above bachelor	No	English only	0.3544517