A Statistical Analysis of the 2019 Canadian Federal Election with a 100% Voter Turnout

Diana Azriel

21/12/2020

Abstract

This paper examines the importance of voter participation by predicting the 2019 Canadian Federal Election Results assuming a 100% voter participation. Using a logistic regression model, we use age, gender, province, and primary home language to predict the likelihood of an individual voting for the Liberal Party in 2019. Then, using the 2017 General Social Survey on Families, we post-stratify to predict the overall popular vote results for the Liberal Party to be 34.4%. Code and data supporting this analysis can be found in the following repository: https://github.com/dianaazriel/STA304-PS4

Keywords

MRP, 2019 Canadian Federal Election, Liberal Party, Conservative Party, Election Turnout, Justin Trudeau, Andrew Scheer

Introduction

Voting is a fundamental right of all Canadian citizens, and the extent to which eligible voters participate in the democratic process is an indicator of societal and political engagement in Canada. In 2019, the Liberal Party did not win the popular vote, gaining 33% of votes while Conservatives gained 34%. However, the voter turnout in the 2019 Canadian Federal Election was only 77%. Interestingly, voter turnout among younger people did not change from 2015 election numbers to those of 2019. This paper will examine the popular vote results if everyone had voted in the election. Furthermore, it will use age, gender, province, and primary home language to determine whether or not an individual will vote for the Liberal Party. This will be done using a logistic regression model. Then, the data will be post-stratified using census data to analyze how a 100% voter turnout would have altered the results of the election.

Data

PES Election Survey

The Consortium on Electoral Democracy (CDEM) released the results of "Canada Election Study 2019", which was an online survey conducted between Sept 13th, to Oct 21st for its campaign period survey, and Oct 24th to Nov 11th for its post-election survey. The survey was offered in both English and French and included 720 variables.

The population in this survey was all Canadian citizens or permanent residents aged 18 or older. The frame included all those who were contacted to answer the survey. Some respondents had to be removed from the dataset to due reasons including, but not limited to, not consenting to the survey, not being a Canadian citizen, not meeting the age requirement of 18 years and older etc. After cleaning, the sample included 37,822 people who were interviewed in the pre-election survey and 10,337 who were interviewed in the post-election survey.

The online sample was produced through Qualtrics and used stratification by region, gender, and age to produce a representative sample. Approximately 10,000 of those surveyed initially were re-contacted after the election to provide post-election information.

This survey was quite strong as it aimed to have an appropriate representation and balance in gender and age groups within each region. It also had proper representation of French and English speakers in different regions, for example, an 80-20 French-English ratio in Quebec. In addition, the survey used respondent's panel ID to match responses between the Campaign Period and Post Election Surveys.

However, the survey also had some duplicate or incomplete responses, which affected the data quality. Furthermore, respondents were able to refuse to answer any given questions, which meant that "Don't know/ Prefer not to answer" was added as an option to required responses. Thus, some responses had to be recorded as missing.

2017 General Social Survey on Families

The 2017 General Social Survey on Families was used for post-stratification analysis. This survey is conducted every 5 years with the objective of gathering data on social trends of Canadians. The 2017 survey in particular was collected between Feb 1st and Nov 30th of 2017. The target population was all non-institutionalized persons 15 years of age or older who live in the 10 provinces of Canada. For the purpose of this analysis, individuals younger than 18 years of age were removed from the dataset. Sampling was done with a cross-sectional design that combined landline and cellular telephone numbers from Statistics Canada's sources. Only one eligible person per household was interviewed. The stratification method, which produced 27 strata in total, was used to carry out sampling.

Some strengths of this data include its rigorous data collection and large sample. Approximately 43,000 people were sampled. Additionally, the survey is quite extensive and reports on a variety of information such as age, sex, family income, family size, and more. The questions in the survey are clear and leave little room for ambiguity.

In terms of weaknesses, since all respondents were interviewed by telephone, households without landline telephones were therefore excluded from the study. Additionally, the exclusion of Canada's 3 territories - Nunavut, Yukon, and Northwest Territories, poses some issues in this analysis. Unfortunately, respondents in the PES survey from the 3 territories had to be removed. Finally, the GSS only collected information on respondents' sex, and not gender. Therefore, to match results with the PES results, gender was imputed on the assumption that males identify as a "Man", and females as a "Woman".

Data Visualization

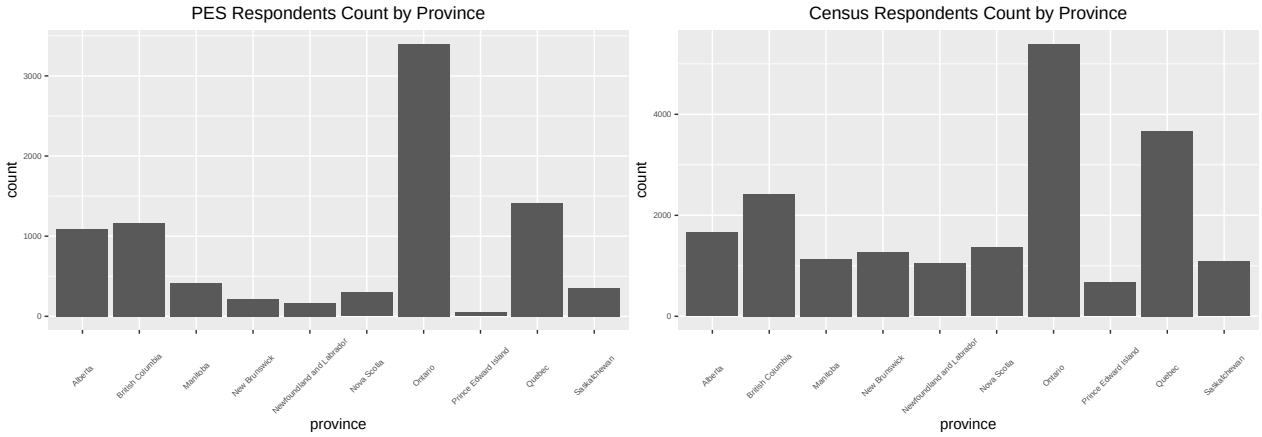

Figure 1: PES and Census Respondents Count by Province

Figure 1 depicts the response breakdown by province for the PES survey and the census. We can see that the proportions are relatively similar between the two surveys, however, some provinces such as Prince Edward Island are under-represented in the Post Election Survey. Therefore, post-stratification will be particularly useful for this analysis.

Model

The following is an outline of the variables used in the model:

Age	Gender	Language_Home	Province
18	Man	French	Alberta
.	Woman	Not French	British Columbia
.			Manitoba
.			New Brunswick
99			Newfoundland and Labrador
			Nova Scotia
			Ontario
			Prince Edward Island
			Quebec
			Saskatchewan

Age is a numerical variable while the rest are categorical. For the purpose of our model, gender, language_home, and province are all treated as dummy variables. Meaning, if the respondent is from Ontario, for example, there is an indicator function that marks $x_{Ontario}$ as 1, while all other province variables get assigned a 0.

Then, a logistic regression model is built of the following form:

$$\begin{aligned} log(\frac{\hat{p}_{Liberal}}{1-\hat{p}_{Liberal}}) = \beta_0 &+ \beta_1 x_{age} + \beta_2 x_{Woman} + \beta_3 x_{Not\ French} + \beta_4 x_{British\ Columbia} \\ &+ \beta_5 x_{Manitoba} + \beta_6 x_{New\ Brunswick} + \beta_7 x_{Newfoundland\&Labrador} \\ &+ \beta_8 x_{Nova\ Scotia} + \beta_9 x_{Ontario} + \beta_{10} x_{PEI} + \beta_{11} x_{Quebec} + \beta_{12} x_{Saskatchewan} \end{aligned}$$

To illustrate how the model works, here is an example:

Given an individual's data:

age: 22
Gender: Woman
Language_home: Not French
Province: Ontario

This would translate to:
$x_{age} = 22$
$x_{Woman} = 1$
$x_{Not\ French} = 1$ $x_{Ontario} = 1$
$x_{all\ other\ provinces} = 0$

We would plug the information into our model to get:

$$\begin{aligned} log(\frac{\hat{p}_{Liberal}}{1-\hat{p}_{Liberal}}) = \beta_0 &+ \beta_1(22) + \beta_2(1) + \beta_3(1) + \beta_4(0) \\ &+ \beta_5(0) + \beta_6(0) + \beta_7(0) \\ &+ \beta_8(0) + \beta_9(1) + \beta_{10}(0) + \beta_{11}(0) + \beta_{12}(0) \end{aligned}$$

Ultimately, we are interested in $\hat{p}_{Liberal}$. So we would solve:

$$\hat{p}_{Liberal} = \frac{e^{\beta_0+\beta_1*22+\beta_2+\beta_3+\beta_9}}{1+e^{\beta_0+\beta_1*22+\beta_2+\beta_3+\beta_9}}$$

Using r language, a general linear model with logistic regression was run to estimate the beta coefficients as listed below. For age, $\beta 1$ indicates that with every unit increase in age, the log likelihood of voting liberal increases by 0.0032. For all other variables, an indication of province, gender, or home_language would increase log likelihood by the appropriate beta. For example, 1.217 increase for Ontario.

Table 2: Fitting generalized (binomial/logit) linear model: vote_liberal ~ age + gender + province + language_home

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.23	0.1523	-14.64	1.594e-48
age	0.003208	0.001517	2.114	0.03447
genderWoman	0.174	0.04731	3.677	0.0002363
provinceBritish Columbia	0.7687	0.1023	7.514	5.742e-14
provinceManitoba	0.6852	0.1341	5.11	3.225e-07
provinceNew Brunswick	1.08	0.1646	6.564	5.237e-11
provinceNewfoundland and Labrador	1.173	0.1764	6.651	2.909e-11
provinceNova Scotia	1.294	0.1417	9.133	6.671e-20
provinceOntario	1.217	0.0876	13.89	6.838e-44
provincePrince Edward Island	1.057	0.3092	3.417	0.0006325
provinceQuebec	1.283	0.1169	10.98	4.697e-28
provinceSaskatchewan	-0.189	0.1692	-1.117	0.2641
language_homeNot French	0.3973	0.09185	4.326	1.521e-05

In the leftmost column of table 2, we can see that all beta coefficients, except for that of Saskatchewan's, have a p-value of less than 0.05. This indicates that those coefficients are significant in determining the log likelihood of an individual voting conservative.

Post-Stratification

To estimate the popular vote results for the Liberal Party given a 100% voter turnout (in the 10 provinces), we use our logistic model to predict voting probabilities in the population. Then, we multiply the probability for each group's voting by the size of the group. For example, there are 7 people who are 18 years old, men, speak French at home, and live in Quebec.Our model predicts $\hat{y}$ for this group to be 0.291. Then, this is multiplied by 7, and is similarly added to all the the other products of bin size and voting estimate. This represents the numerator in the equation below:

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

Then, we divide by the total population in the census.

Results

Using logistic regression with post stratification, we estimate that with a 100% voter turnout, the Liberal Party would have received 34.4% of the popular vote in the 2019 Canadian Federal Election. In reality, Liberals only won 33.1% of the popular vote, while Conservatives won 34.4%. Our model was rerun to predict the likelihood of voting conservative, and after similar post-stratification, we estimate that only 29.5% of Canadians would have voted for the Conservative Party.

For model results of voting likelihood for the Conservative Party, please see appendix.

Experts note that this election was historically significant, marking only the second time in Canadian history that a governing party takes power with such a low share of the votes. In the past, there was only one time that a party formed government having earned less than 35% of the popular vote - this incident was the 1867 election in which John A. Macdonald won. Thus, despite the Liberal Party gaining more votes if everyone had voted, they would still earn one of the lowest popular vote proportion given their win.

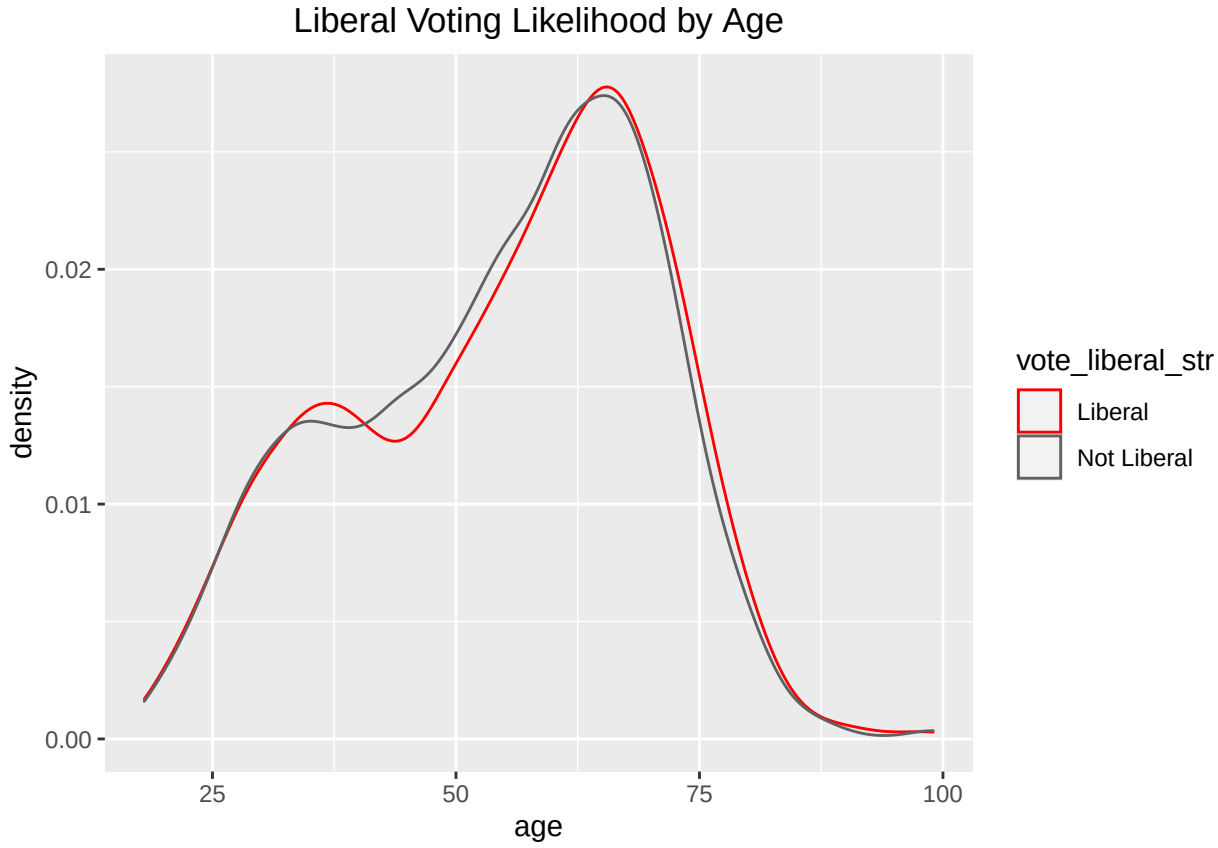

Figure 2: Liberal Voting Likelihood by Age

Figure 2 depicts Liberal party voting density by age. Interestingly, the liberal (red) curve and not liberal (grey) curve follow a somewhat similar pattern. It is worth noting that up until approximately age 40, one is slightly more likely to vote liberal. Then, from age 40 to approximately 60, the grey curve lies above the red curve, indicating that individuals in that age group are more likely not to vote liberal.

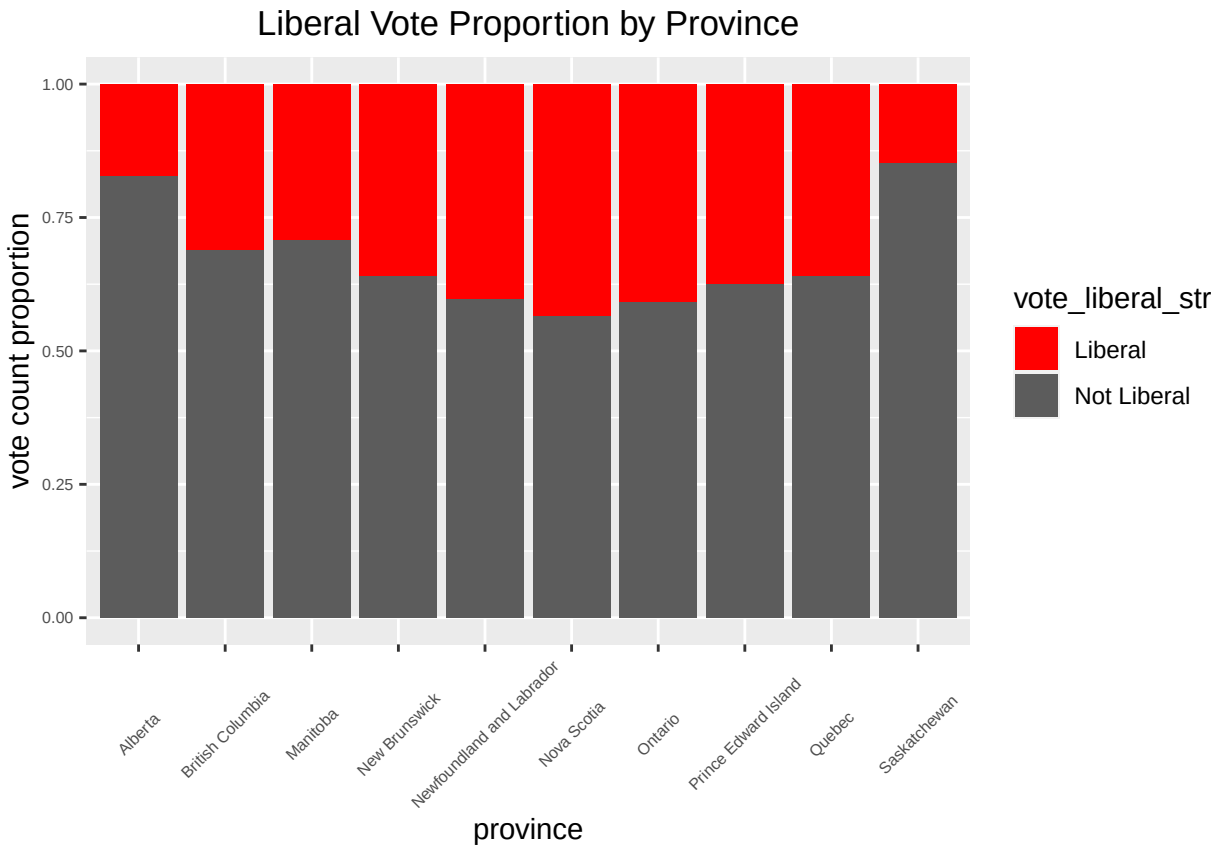

Figure 3: Liberal Vote Proportion by Province

Above is a graph depicting the Liberal Vote by Province. It is apparent that in maritime provinces, namely New Brunswick, Nova Scotia, and PEI, almost 50% of voters indicate they voted liberal. Alberta and Saskatchewan have the lowest proportion of liberal voters.

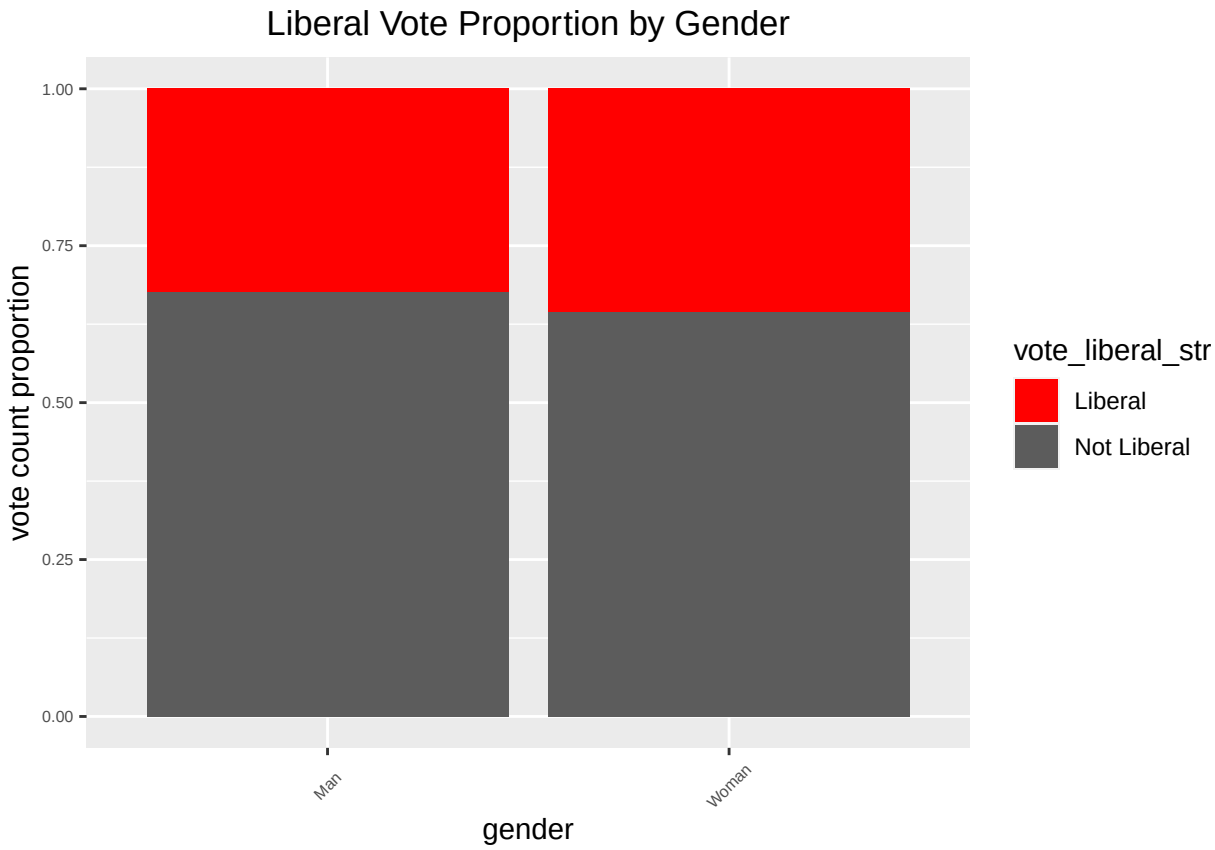

Figure 4: Liberal Vote Proportion by Gender

Figure 4 indicates women are slightly more likely to vote Liberal. This is consistent with our model results, which showed that being a woman increases the log likelihood of voting liberal by 0.174.

Discussion

In the interest of predicting the popular vote outcome of the 2019 Canadian Federal Election with a 100% voter turnout, we ran a logistic regression model with the response variables age, gender, province, and language spoken at home to determine the likelihood of an individual voting for the Liberal Party.

Based off the estimated proportion of voters in favour of voting of the Liberal Party being 34.4%, our model predicted that the 2019 election would have earned the Liberal Party a bigger share of the popular vote if everyone had voted. However, 34.4% of the popular vote would still have been a record low number for popular vote share for a party that takes power.

Individuals living in maritime provinces, as well as Ontario, were most likely to vote Liberal. Men and women had relatively similar proportions of voting for the Liberal Party, but women were slightly more likely to do so. Individuals whose primary home language was not French were also more likely to vote Liberal. Although age was a significant factor in our model, voters had similar voting patterns in all age groups, except 40-60 year olds who were slightly more likely not to vote Liberal.

Weaknesses

The biggest weakness in our 2 data sets was the inconsistency of gender/sex categorization. In the GSS dataset, only sex was coded, with exclusively two options: male and female. This fact, alone, presents a major weakness in the dataset as those two options may not represent all possible representations of an individual. In the Post Election Survey dataset, the gender question had 3 options: Woman, Man, or Other. For the purpose of post-stratification, the two categories had to be the same, so all "Other" responses were imputed as woman. And in the GSS dataset, all female responses were assumed to be equivalent to "woman", while male responses were equivalent to "man". Naturally, this imputation and assumption may create a bias that does not represent the Canadian population accurately.

In addition, the GSS dataset had no information about individuals from the 3 territories of Canada. Thus, our 100% voter participation rate only accounts for those living in the 10 provinces of Canada.

Next Steps

In the future, this analysis should be rerun using a different set of census data - preferably one that has information about respondents' gender.

Appendix

Below are model results for our logistic regression model run to predict conservative voting probabilities.

Table 3: Fitting generalized (binomial/logit) linear model: vote_conservative ~ age + gender + province + language_home

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.4878	0.1603	-3.044	0.002337
age	0.009489	0.001617	5.866	4.454e-09
genderWoman	-0.4637	0.04992	-9.288	1.573e-20
provinceBritish Columbia	-1.262	0.08974	-14.06	6.38e-45
provinceManitoba	-0.9977	0.12	-8.312	9.427e-17
provinceNew Brunswick	-1.164	0.1652	-7.046	1.835e-12
provinceNewfoundland and Labrador	-1.443	0.1863	-7.749	9.25e-15
provinceNova Scotia	-2.009	0.166	-12.11	9.801e-34
provinceOntario	-1.285	0.0734	-17.51	1.195e-68
provincePrince Edward Island	-1.906	0.3778	-5.045	4.531e-07
provinceQuebec	-1.96	0.1238	-15.83	1.89e-56
provinceSaskatchewan	-0.2421	0.1244	-1.946	0.05163
language_homeNot French	0.6657	0.1136	5.861	4.59e-09

References

Alexander, R. and Caetano, S. (2020). "gss_cleaning,.R". Retrieved from: https://q.utoronto.ca/courses/184060

Brean, J. (2019). "All-time low share of popular vote is enough for Liberals to win power". National Post. Retrieved from https://nationalpost.com/news/politics/election-2019/canadian-federal-election-2019-liberals-justin-trudeau-win

Statistics Canada. (2020). General Social Survey Cycle 31: Families. 45250001 Issue no. 2019001. pages 3-11.

Statistics Canada. (2020). "Reasons for not voting in the federal election, October 21, 2019". Retrieved from: https://www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm

Stephenson, L. et al. (2020). "2019 Canadian Election Study - Online Survey", https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1