If All Canadian Voters Participated in 2019 the Liberals Would Increase Their Win

Jack Edward Smith

22/12/2020

All code for this analysis was completed in R. The supporting code and data are available at the github repo https://github.com/JackSmith100/STA304-Final-Project-Repo.git

Abstract

This report analyzes the 2019 Canadian election using the 2019 Canadian Election Survey (Stephenson et. al, 2019) and the 2016 Canadian Census (Statistics Canada, 2016). The goal of this report is to predict the winner of the Canadian election if all eligible voters cast a vote in 2019. Binary logistic regression models were used to determine the likelihood of certain demographic groups voting for each major political party. Then a poststratification technique was used to expand the results of the model to the entire population of Canada with the Canadian Census. It was found that the Liberal government would make gains on the Conservative party and potentially increase their seat total, but not enough evidence to conclude the Liberals would win a majority government.

Keywords

Keywords: Poststratification, Census, CES, Logistic Regression, Election, Canada, Conservative, Liberal, Demographic, Model

Introduction

The 2019 Canadian election resulted in a minority government for the Liberal party and the largest opposition party in history (CBC, 2019). With a very close election result, its important to know how the entire country feels about it. What might have been the result if the entire population voted? The voter participation rate was only 66% from eligible voters (CBC, 2019), the goal of this paper is to determine the outcome of the vote if the participation rate was 100%. The data sets that are used in order to conduct the analysis are the Canadian Election Survey 2019 (Stephenson et. al, 2019) and the 2016 Canadian Census data (Statistics Canada, 2016). The Canadian Election Survey 2019 or CES asks people living in Canada their thoughts on the election, important issues and which party they intend to vote, while the 2016 census is the most recent census conducted by the Canadian government.

This paper utilizes a poststratification technique to determine the results if everyone voted. Poststratification was famously used by Wang et. al when they used Xbox gamer surveys to predict the results of the 2010 US presidential election. Poststratification is used to create a more statistically sound way to "extrapolate" results onto a large population of people. The poststratification used in this paper involves portioning into cells based on demographic characteristics such as age, gender and province. The Canadian Election Survey will be used to create weights that represent the likelihood of voting for a party based on their demographic group. Then weighting these groups and their voting probabilities onto the census population total, to predict the winner of the 2019 Canadian election if the entire population voted.

This study will show the importance of voter turnout and how people that did not vote could sway the election results from one party to another. This paper will analyze the CES and Census data sets, the logistic

regression model, the poststratification and find results for the number of people voting each political party. After predicting the winner using the poststratification technique the results of the 2019 election will be compared to the results obtained from this study.

Methodology

Data

The CES is a survey that asks a wide range of topics on the 2019 election including the most important issues, who they intend to vote for and demographic details such as gender, age, location. The Canadian Census provides an overview of the Canadian people's demographic, economic and social characteristics. The CES population is all eligible voters in Canada, the sampling frame is Canadians with access to the internet and the sample is people who voluntarily took the survey. The Census is a survey conducted by the entire population so the population, frame and sample should be all Canadians. The CES had a total of 37822 respondents before the data was cleaned and after the data was cleaned had a total of 26213 respondents, while the Census had a total of 930421 respondents and was cleaned to have 742181 respondents. Cleaning in the Census involved removing anyone under the age of 18, which is the age to vote in Canada, and removing any data that had NA values in the demographics of interest (Age, gender, province). The CES cleaning involved removing anyone who did not list a party they intended to vote and changing the location of people living in the territories to Northern Canada as the census did not have separate territories listed and removing people who were not Canadian Citizens.

One difference between the two data sets that had to be adjusted is that the CES contained other genders than male and female, they gave the option to pick Other (e.g. Trans, non-binary, two-spirit, gender-queer). While the Census only contains male and female as options. There had to be a decision on what to do with the people who picked "other", whether to remove them or impute them to be male or female. The decision to remove them felt like the best option as they made up a small amount of the data set (291 of 37822) and it avoids assigning the respondent to a gender that they do not want to be assigned. Removing the people listing "other" could be called discrimination by some and could be a weakness of the study, but with the small number of people identifying as "other" it was roughly equal to assigning half to each gender, so to avoid misgendering they were removed.

The variables chosen to complete the model were age, gender and province. These variables were picked as they are 3 characteristics of a voter that are likely to influence how they vote. Using these 3 variables also allowed for a large amount of data to be available for most groups of people, which helps the model as there is enough data to get sufficient estimates for each group. A key strength for each of the data sets is that they obtain a lot of data that can be used to get statistically significant results. Graphs of the data are included in the results section with the parties each demographic intends to vote for.

Model

In order to analyze the data, a frequentist approach was used with the CES 2019 and Census data sets, the software used to complete the analysis was R. Conclusions are based on the frequency and proportion of the votes for each party. Using the CES 2019, binary multiple regressions were developed to explore the Canadian election results. The binary response was designed to take the value of 1 when someone votes for the selected party and the value 0 when the intend to vote for a different party. The independent variables of the regression are age, gender and province. The age variable is a categorical variable, as well as the province and gender variables. The gender variable takes values male and female while province has a category for each province and one for Northern Canada. A binary logistic regression model is a good fit for the data analysis as it meets all the expected criteria and assumptions. The assumptions needed to complete a logistic regression include having a binary response variable, independent data, uncorrelated predictor variables and a linear relationship. The binary response is created, by assumption the CES is an independent survey, there should be no correlation between where someone lives, their age and gender, and lastly the linear relationship is shown in Equation. 1. A Binary regression works better than linear regression as the response variable

fits perfectly with 0 and 1 (vote for the party or not), which would not be a good for linear. Binary logistic regression models were developed for the Conservative, Liberal, Green, NDP and People's Party.

Equation 1

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 Age_{25-34} + \beta_2 Age_{35-44} + \beta_3 Age_{45-54} + \beta_4 Age_{55-64} + \beta_5 Age_{65-74} + \beta_6 Age_{75-84} + \beta_7 Male + \beta_8 Prov_{BC} + \beta_9 Prov_{MB} + \beta_{10} Prov_{NB} + +\beta_{11} Prov_{NL} + \beta_{12} Prov_{NC} + \beta_{13} Prov_{NS} + \beta_{14} Prov_{ON} + \beta_{15} Prov_{PEI} + \beta_{16} Prov_{QC} + \beta_{17} Age_{SK} + \epsilon$$

Equation 1 represents the logistic model and p represents the proportion of voters who will vote for the political party in the model. The coefficients of the regression (betas and intercept) have the same interpretation in each of the models. The intercept or β_0 represents the log odds of an 18-24-year-old female from Alberta voting for the chosen political party of the model, the slope coefficients or the betas 1 through 17 then represent the change in log odds of voting for the party with respect to the intercept. β_1 to β_6 represent the slope coefficient and change in log odds for each age group. β_7 is the gender slope coefficient and represents the change in log odds for being male relative to female. β_8 through β_{17} then represent the slope coefficient for the province variables.

Model 1

term	estimate	std.error	statistic	p.value
(Intercept)	0.8223431	0.0841262	9.775113	0.0000000
as.factor(Age)25 to 34	-0.2776648	0.0815562	-3.404581	0.0006627
as.factor(Age)35 to 44	-0.2497149	0.0801266	-3.116506	0.0018301
as.factor(Age)45 to 54	-0.2696440	0.0813787	-3.313448	0.0009215
as.factor(Age)55 to 64	-0.4136814	0.0815877	-5.070386	0.0000004
as.factor(Age)65 to 74	-0.6950996	0.0822619	-8.449835	0.0000000
as.factor(Age)75 to 84	-1.1122448	0.0927558	-11.991103	0.0000000
as.factor(Gender)Male	0.3815332	0.0283861	13.440823	0.0000000
as.factor(Province)British Columbia	-1.4065801	0.0549963	-25.575916	0.0000000
as.factor(Province)Manitoba	-0.9027487	0.0709817	-12.718042	0.0000000
as.factor(Province)New Brunswick	-1.5146769	0.1006480	-15.049254	0.0000000
as.factor(Province)Newfoundland and Labrador	-1.7292460	0.1200031	-14.410006	0.0000000
as.factor(Province)Northern Canada	-1.5475697	0.2899835	-5.336752	0.0000001
as.factor(Province)Nova Scotia	-1.9045912	0.1003734	-18.975052	0.0000000
as.factor(Province)Ontario	-1.3793172	0.0431694	-31.951248	0.0000000
as.factor(Province)Prince Edward Island	-1.9386954	0.2500081	-7.754530	0.0000000
as.factor(Province)Quebec	-2.1405266	0.0513685	-41.670009	0.0000000
as.factor(Province)Saskatchewan	-0.3697916	0.0756066	-4.890995	0.0000010

Model 2

term	estimate	std.error	statistic	p.value
(Intercept)	-1.3522739	0.0882113	-15.3299378	0.0000000
as.factor(Age)25 to 34	-0.0079593	0.0811041	-0.0981368	0.9218237
as.factor(Age)35 to 44	-0.0891067	0.0798814	-1.1154875	0.2646416
as.factor(Age)45 to 54	-0.2086453	0.0812633	-2.5675210	0.0102429
as.factor(Age)55 to 64	-0.1741637	0.0810993	-2.1475362	0.0317506
as.factor(Age)65 to 74	-0.2566330	0.0813545	-3.1545043	0.0016077
as.factor(Age)75 to 84	-0.3457126	0.0880769	-3.9251218	0.0000867
as.factor(Gender)Male	-0.0966111	0.0273062	-3.5380620	0.0004031
as.factor(Province)British Columbia	0.7326513	0.0614569	11.9213818	0.0000000
as.factor(Province)Manitoba	0.6819198	0.0795940	8.5674746	0.0000000
as.factor(Province)New Brunswick	1.0745993	0.0976969	10.9993219	0.0000000
as.factor(Province)Newfoundland and Labrador	1.4886889	0.1073965	13.8616191	0.0000000
as.factor(Province)Northern Canada	1.0445641	0.2673996	3.9063783	0.0000937
as.factor(Province)Nova Scotia	1.3599611	0.0891772	15.2500994	0.0000000
as.factor(Province)Ontario	1.1028494	0.0509744	21.6353472	0.0000000
as.factor(Province)Prince Edward Island	1.1945222	0.2068534	5.7747272	0.0000000
as.factor(Province)Quebec	0.9549720	0.0544565	17.5364319	0.0000000
as.factor(Province)Saskatchewan	-0.2967544	0.1050914	-2.8237736	0.0047462

Here are two examples of the logistic regression models that were run for each party. The Conservative Party is represented by model 1 and the liberal party by model 2. As you can see in the models, the majority of the p-values are very statistically significant, this is a strength of the model and the data set. Values with a negative coefficient mean that the group is less likely to vote for the party relative to the intercept (18-24-year-old female from Alberta), where as positive values mean the group is more likely to vote for the party relative to the intercept. The Conservative and Liberal models were shown as they are the two parties with an opportunity to win the election.

Utilizing the models above, the poststratification technique was then performed with the Canadian Census. In poststratification the population is divided into cells based on the demographic characteristics (Age, Province, Gender) that were used to develop the binary logistic regression model above. The cells each represent a specific group of people, for example there will be a cell for 24-34-year-old males from Ontario and a cell for 35-44-year-old females from BC, each set of characteristics is separated into a cell for a total of 154 cells. This is why changing the age variable into categorical is beneficial as it vastly reduces the number of cells, this is effective in making sure there is a sufficient amount of data in each cell to complete the poststratification and make estimates. The logistic regression estimates are made to create weights for each cell on how likely they are to vote for each political party. With the weights of each cell from the logistic regression the Canadian Census is used to standardize the groups to a more realistic representation of the true Canadian population. This is the important part of the analysis as it creates a more statistically sound way to extrapolate the results onto the entire population, which will increase the accuracy of the prediction. The weights/estimates for each cell, combined with the frequency at which the cell occurs in the census is what gets the final vote prediction and establishes a more statistically accurate percentage of the population voting for each party.

From the logistic regression models it can be concluded the estimated vote for each political party is:

Conservative = 33.70%
Liberal = 34.17% NDP = 14.35% Green = 8.85%
Peoples = 2.20%

Results

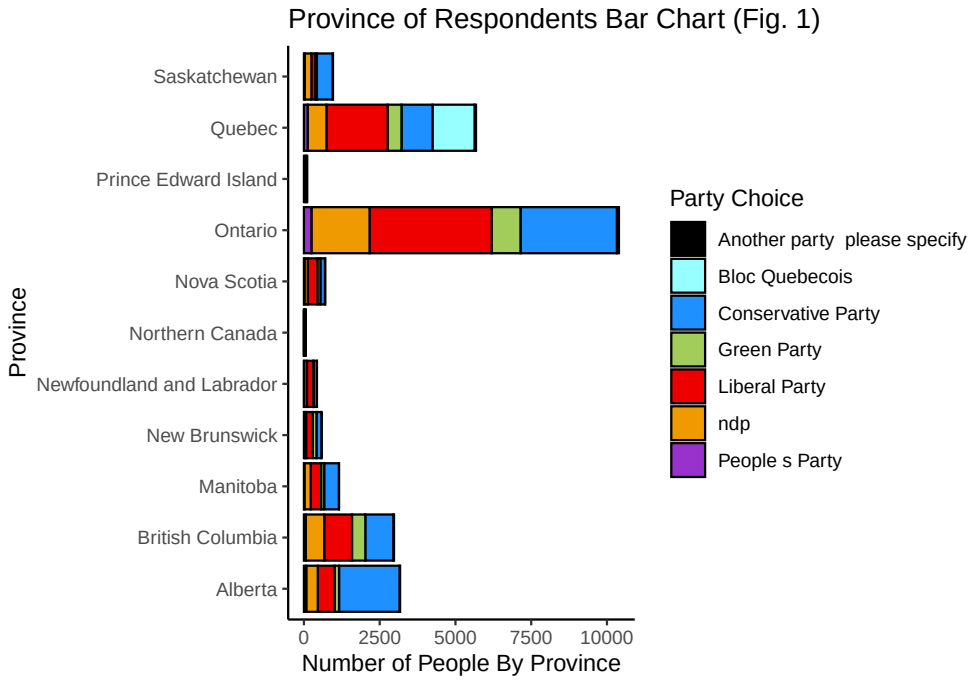


Figure 1 shows the breakdown of the number of votes per province and the number of votes of each province that went to each political party.

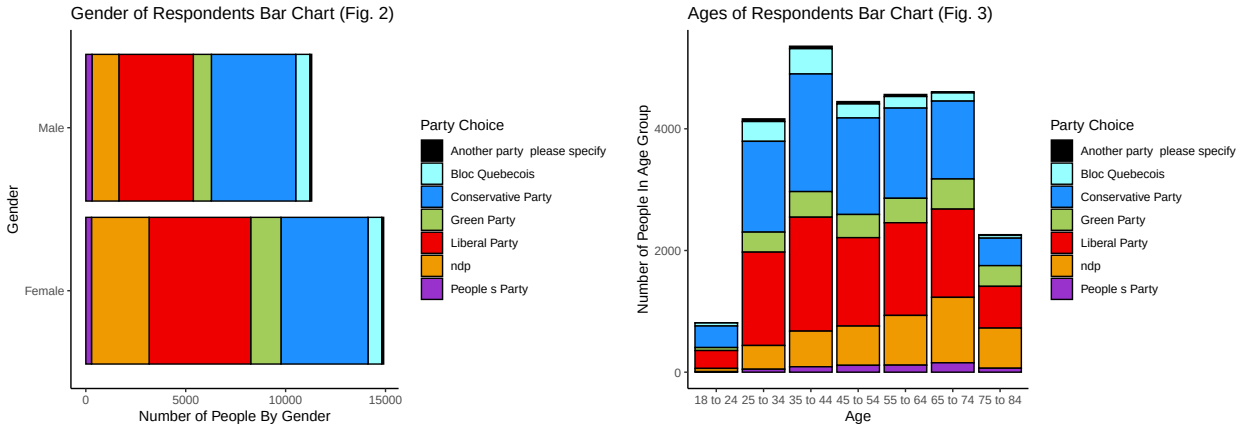
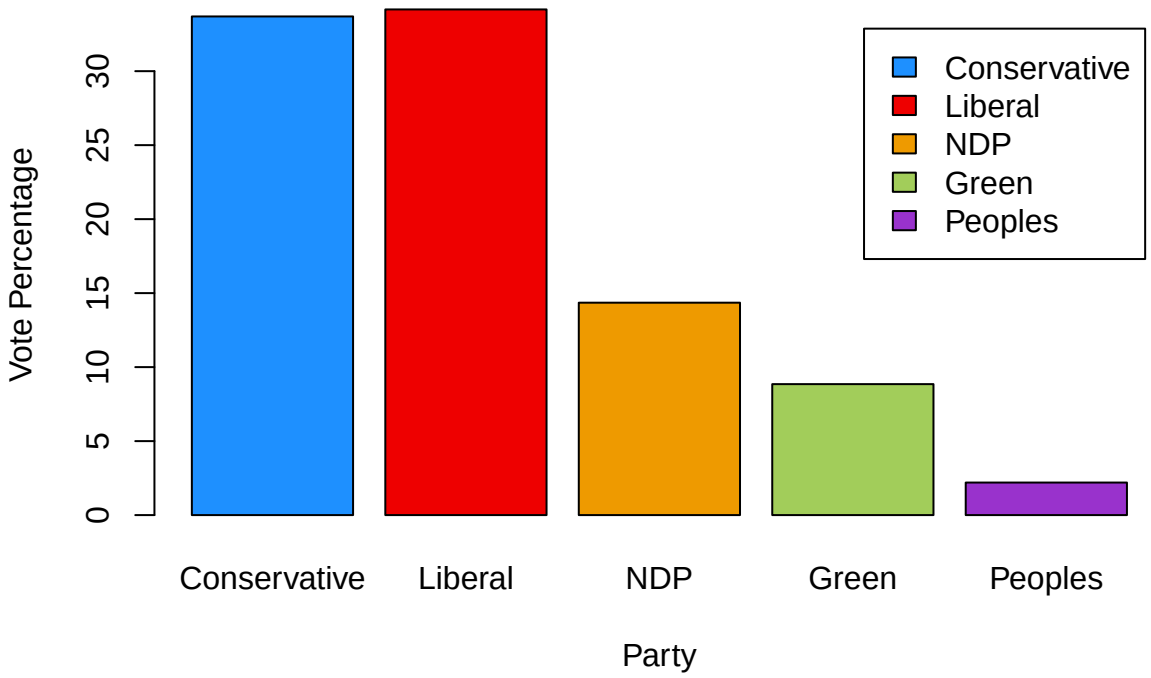


Figure 2 represents the number of votes by gender and the political parties they intend to vote. This is consistent with actual results as women voted in larger numbers than men in the election (Elections Canada).

Figure 3 shows the breakdown of the number of votes by each age group and the number of votes from that age group that go to each political party. This shows that the age groups of 18-24 and 75-84 could be underrepresented in the CES survey. This could result in a problem if results from the CES survey were interpreted as the final results, this exactly part of the reason why the poststratification technique was used. As shown by Wang et. al when they used Xbox surveys to predict the US presidential election, a poststratification technique is good a correcting a sample that is misrepresented of the true population. The Xbox surveys were likely skewed toward young men, but they corrected for it by using the US census in poststratification.

Figure 4: Poststatification Election Results



The estimated proportion of voters in favour of voting for the Liberal Party is expected to be 34.17%. This calculation is based on the poststratification analysis of the proportion of Liberal party voters modeled by binary logistic regression, which accounted for peoples age, gender and province. Very similar models were also calculated for the Conservative, Green, NDP and People's Party. The estimated percent of the vote going to the Conservative Party is 33.70%, the estimated percentage for the NDP is 14.35%, while the vote going to the Green and Peoples Party is at 8.85% and 2.20% respectively.

Discussion

Summary

The model used in this paper was designed to predict the 2019 Canadian election if the everyone eligible to vote had voted. The Canadian Election Study 2019 was used to build logistic regression models to predict the likelihood of people voting for each party based on age, gender and province. Then the census was used in a poststratification to expand the results. With the census poststratification it was found that the Liberal party would win the popular vote instead of the Conservative party which came in a close 2nd place. The difference between the Conservatives and the Liberals was only 0.47%, when the entire population votes.

Conclusions

Comparing the results from the 2019 election and the poststratification model, it can be determined if there would be change in the results of the election if everyone voted. The 2019 election resulted in a minority Liberal Government, but the conservatives won the popular vote (CBC, 2019). In the 2019 election the Conservative Party finished with 34.4% of the popular vote, the Liberals with 33.1% of the vote, NDP with 15.9%, Green with 6.5% and Peoples with 1.6% (CBC, 2019). The results of the actual election and the results of the poststratification are very similar. One interesting change is the Liberal Party wins the

popular vote instead of the Conservatives. The Liberal Party, Green Party and Peoples Party made gains in percentage, while the Conservative Party and the NDP have decreases in percentage of the vote. A model for Bloc Québécois was not run but given the total vote percentage already accounted for is 93.33%, they are not going to win the election but will likely receive most of the percentage not already accounted for as there were not many votes for other parties.

Although the popular vote is normally indicative of the winner of the election, this is not how the Canada election system works, as they use a first past the post system to determine the winner of the election, winning the popular vote does not determine who is Prime Minister. Comparing the results of the poststratification with the 2015 election where the Liberals won a majority government, the Liberals finished with 39.5% of the vote and the conservatives with 31.9% (CBC, 2015). Since the poststratification model predicts 33.70% for the Conservatives and 34.17% for the Liberals it is unlikely the Liberals would win a majority and the result would likely again end up in a minority government, As the vote percentages are still very similar to the 2019 results. Although a majority may not be guaranteed it is still predicted that the liberals would win the popular vote by 0.47% over the Conservatives and would most likely have an increase in the number of seats they hold. A 0.47% difference is a small amount and part of the reason that the Conservatives won the popular vote in the actual election, 34% of the population did not vote and that benefited the conservatives. This shows the importance of voter turnout as the result that the majority of the population would vote for may not come true if a large portion does not show up.

Weaknesses

One potential weakness of the model is that the CES survey did not contain any information of voters over the age of 84. The poststratification technique is great at taking samples that are misrepresentative of the population and weighting them to be representative, but the model does not consider weights that have zero representation at all. "In 2016, people aged 85 and older represented 2.2% of the Canadian population" (Government of Canada, 2019), this is enough to sway the vote. In such a close result with the vote predicted for the Conservatives to be less than 1 percent different than the Liberal party, it would be best if the data were available for those over 84. One potential reason to why people over 84 did not participate is that the study was conducted online, compared to a similar phone survey done by CES that had people over 84 participate. The older population may not be well captured by the sampling frame of the CES.

Another issue results when attempting to complete a poststratification on the Bloc Quebecois party. As the Bloc party only receives votes from the province of Quebec, using the province as a variable to complete the logistic regression results in an error as the model will over exaggerate the percentage of the vote for the Bloc. Although this weakness will not affect the overall results as the Bloc Quebecois will does not run enough candidates to win the election, it is still something that should be corrected to get the best overall model. This could be a weakness as well with the CES data set as it did not contain much personal information about people in the study making it difficult to pick other variables to perform a poststratification.

Next Steps

The next step could be fitting a model without the provincial variable to get an accurate prediction on the number of people that will vote for the Bloc Québécois. The provincial variable works well for predicting the winner of the popular vote with parties like the Conservative and Liberals as they run candidates in every province but not well for the parties only in one province. Another possible next step would be to add data from the CES 2019 phone survey of people over 85 to the CES web survey which was used in the paper (Stephenson et. Al, 2019). The phone survey contained information and voting preferences for people over the age of 85 and this could be used to account for the vote of people over 85 and fix the weakness in the web survey.

With a current minority government, a potential election could happen at anytime and likely in 2021. Now more than ever its important to understand who would have won in 2019 if everyone voted as it will measure as a potential indicator of who will win the next election with more people becoming politically engaged. With that being said, another study can be done using the CES to determine the important issues why people voted for each party. These issues may have changed due to Covid-19 but can still be measured up against

current party platforms to see who is talking about the important issues and be used to estimate who will be doing well in the polls in 2021.

References

Caetano, S., (2020) , 01-data_cleaning-survey1.R. 01-data_cleaning-post-strat1.R. ProblemSet3 - template-logistic.Rmd., https://q.utoronto.ca/courses/184060/modules.

CBC News: Canadian election drew nearly 66% of registered voters | (2019, October 22). Retrieved December 15, 2020, from https://www.cbc.ca/news/canada/voter-turnout-2019-1.5330207

CBC News: Election 2015 roundup. Retrieved December 15, 2020, from https://www.cbc.ca/news2/interactives/results-2015/ CBC News: Federal election 2019 live results. (n.d.). Retrieved December 15, 2020, from
https://newsinteractives.cbc.ca/elections/federal/2019/results/

Estimation of Voter Turnout by Age Group and Gender at the 2015 General Election. Retrieved December 14, 2020, from https://www.elections.ca/content.aspx?section=res&dir=rec/part/estim/42ge&document=p1&lang=e

Government of Canada, S. (2019, April 03). A portrait of the population aged 85 and older in 2016 in Canada. Retrieved December 22, 2020, from https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016004/98-200-x2016004- eng.cfm

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.Rproject.org/package=dplyr

Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

Hadley Wickham, Jim Hester and Winston Chang (2020). devtools: Tools to Make Developing R Packages Easier. R package version 2.3.2. https://CRAN.R project.org/package=devtools

Joseph Larmarange (2020). labelled: Manipulating Labelled Data. R package version 2.7.0. https://CRAN.R-project.org/package=labelled

Paul A. Hodgetts, Rohan Alexander. cesR: Access the Canadian Election Study Datasets a Little Easier. https://hodgettsp.github.io/cesR/index.htm

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria., https://www.R-project.org/.

Statistics Canada. 2016 Census of Canada (database). Using CHASS (distributor). https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sdacensus/hsda?harcsda+cc16i (accessed December 1 2020).

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

Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Phone Survey", https://doi.org/10.7910/DVN/8RHLG1, Harvard Dataverse, V1

Wang, Wei; Rothschild, David; Goel, Sharad; Gelman, Andrew (2015). "Forecasting elections with non-representative polls". International Journal of Forecasting. 31 (3): 980–991.

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