The Study of the possible outcome of 2019 Canadian Federal Election if everyone voted

Yiyang

2020/12/18

Contents

1 Information

Topic: The Study of the possible outcome of 2019 Canadian Federal Election if everyone voted

Author: Yiyang Huang

Data: CES Dataset: `ces2019_web` General Social Survey, Cycle 27: Social Identity dataset

Date: December 20, 2020

Github Link: https://github.com/YiyangHuang028/STA304-Final.git

2 Keywords

Election, 2019 Canadian Federal Election, Education Level, Marital Status, interest in politics, likely to vote, AIC, Stepwise selection, Multiple Logistic Regression, Post-stratification

3 Abstract

In the 2019 Canadian Federal Election, the Liberals only got 33.1% of the popular votes, which is slightly lower than the Conservatives. However, the Liberals still got 157 seats in the House of Commons and successfully formed a minority government. People argued that the result is controversial. This study analyzed the possible outcome of the 2019 Canadian Federal Election if everyone has voted. The study is based on the Canadian Election Study 2019 Online Survey which was conducted before and after the election. Multiple Logistic Regression with Post-stratification is used to give an estimation. The result suggests that the Liberals are not likely to win if everyone has voted during the election.

4 Introduction

In the 2019 Canadian Federal Election, the Liberal Party won a strong minority government with a total of 157 seats in the House of Common, at the same time, Justin Trudeau won a second term as Canada's Prime Minister (Britneff, 2019). Although the Liberals secured 36 more seats than the Conservatives, the Liberals only got 33.1% of the popular vote while the conservatives gained 34.4% (Bowen, 2019). Warren, a journalist of CBC News, stated that "Canadians aren't getting the government they're voting for" (Bowen, 2019). Also, some people argue that the "first-past-the-post" electoral system does not provide a fair representation.

If we take a look at the voter turnout, we can tell it dropped slightly in 2019 compared to the election in 2015. Elections Canada stated that only 65.95% of eligible voters in Canada cast a ballot in comparison to 68.3% in 2015 (Britneff, 2019). According to Statistics Canada, 35% of the non-voters aged below 75 stated that the reason for not voting is low interest in politics (Statistics Canada, 2020). People aged 75 and older usually report illness (49%) and difficulties in the electoral process (9%) (Statistics Canada, 2020). Considering there are 31.7% of the eligible Canadian voters who did not vote in the 2019 Canadian Federal Election due to various reasons, there is reason to believe that the outcome of the election may be different if everyone has voted.

This study gives a prediction of the possible outcome of the 2019 Canadian Federal Election if everyone has voted. The analysis is based on the `ces2019_web` survey dataset and General Social Survey, Cycle 27: Social Identity dataset. The estimation of the outcome is carried out by Multiple Logistic Regression with Post Stratification. The Methodology section will give a brief introduction of the census and survey dataset along with how they are cleaned and selected. Also, it contains how the model is built and how post-stratification is done. In the Result section, tables and plots are used to visualize the relationship between voter's vote choice and other factors (age group, education level, marital status, likely to vote, and interest in politics). The conclusion, relative findings, weaknesses, and further analysis of this study will be included in the Discussion section. The result can be used as a reference for those parties who want to earn support from the Canadians and get representation in the House of Common.

5 Data

`ces2019_web` is used as the survey dataset to estimate the outcome. It contains pre-election and post-election survey data which was conducted online before and after the 2019 Canadian Federal Election. The composition of the sample was considered when collecting the information. According to Table1, 58.8% of the respondent are female and 41.2% are male. The respondent who aged between 18 and 34 take up 24.1% of the sample, those who aged 35-54 take up 34.7%, and those aged 55 and higher take up 41.2%.

The original dataset contains 37822 observations and 620 variables. `cps19_age`, `cps19_gender`, `cps19_marital`, `cps19_education`, `cps19_v_likely`, `cps19_interest_gen_1`, `cps19_votechoice`, and `cps19_vote_unlikely` are chosen to do the analysis in this study. Each variable is renamed and observations are cleaned. The column `vote_choice` is added accordingly as the response variable in our study. If the respondents claimed they are likely to vote, then `vote_choice` equals 1 if they claimed that they will vote for the Liberals(according to the variable `cps19_votechoice`), 0 otherwise. If the respondents claimed they are not likely to vote, then `vote_choice` equals 1 if they favor the Liberals than the other parties (according to the variable `cps19_vote_unlikely`), 0 otherwise. Thus, `vote_choice` takes the thoughts of those non-voters into account, which is why we use this variable as the response variable instead of `cps19_votechoice`. This will give us a better estimation of the outcome under the assumption - everyone has voted in the 2019 Election.

To carry out the post-stratification, we need a census dataset that contains similar variables as the variables selected in the CES dataset. The General Social Survey, Cycle 27: Social Identity dataset is used to construct the census dataset. The data was collected from June 2013 to March 2014, and the target population is the Canadians and permanent residents aged above 15. The census data is constructed and cleaned in a separate Rscript file called `gss_cleaning_2013`.

The data originally contains 27534 observations and 720 observations. `agegr10`, `sex`, `marstat`, `dh1ged`, `vbr_25`, `rep_05` are selected to match with the variables selected in CES dataset. Then, variables are renamed and observations are cleaned using the same way as above. Then, the variables are grouped, and the column `n` is added to contain the count result of each group for post-stratification.

Table 1: Baseline Characteristics of the CES dataset

	N=31378
Age	
15 to 24 years	2345 (7.5%)
25 to 34 years	5289 (16.9%)
35 to 44 years	5582 (17.8%)
45 to 54 years	5295 (16.9%)
55 to 64 years	6363 (20.3%)
65 to 74 years	5084 (16.2%)
75 years and over	1420 (4.5%)
sex	
Female	18241 (58.1%)
Male	13137 (41.9%)

6 Model

To estimate the possible outcome of the election if everyone voted, the Multiple Logistic Regression (MLR) Model with Post-stratification is used in this study because our response variable `vote_choice` is a binary variable. In comparison to the multiple linear regression model, which is better at giving numeric estimation, the MLR is better at predicting discrete binary outcomes.

First, a logistic regression model called `full_model` is built with `age`, `sex`, `marital_status`, `grouped_education`, `likely_to_vote`, and `interest_in_politics`. I include `sex` in the model because Justin Trudeau is a feminist who "believes men and women should be equal" (Carpenter, 2018), and this might affect how male and female vote during the election. According to Statistics Canada, people aged 55-74 always have a higher voter turnout in comparison to other age groups (Statistics Canada, 2020). Thus, `age` is included in the model. `likely_to_vote` directly people's willingness to vote during the election, which is closely related to the voter turnout. Also, considering "low interest in politics" is the main reason why people don't vote in an election, `interest_in_politics` definitely affects the voter turnout, and so these two variables are also included. A study conducted by myself in 2020 based on the CES dataset suggests that "people who attained higher education level have a higher satisfaction level with Justin Trudeau" (Huang, 2020), so `grouped_education` is also included in the model.

Then, a backward regression is performed to select the best model for the estimation. In this process, `sex` is removed from the previous model and the new model is called `reduced_model`. The AIC for the `full_model` is 36303, and the AIC for the `reduced_model` is 36249, which suggests the `reduced_model` has a better fit. The AIC mentioned here measures how the goodness of fit and simplicity are balanced in a model (Chen & Yang, n.d.)

The MLR model here is run by R, and here is the model:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 X_{1_{age25-34}} + \beta_2 X_{2_{age35-44}} + \beta_3 X_{3_{age45-54}} + \beta_4 X_{4_{age55-64}}$$

$$+\beta_5 X_{5_{age65-74}} + \beta_6 X_{6_{age75+}} + \beta_7 X_{7_{NotMarried}} - \beta_8 X_{8_{highschool}} + \beta_9 X_{9_{postsec}} + \beta_{10} X_{10_{uni}}$$

$$+\beta_{11} X_{11_{likely}} + \beta_{12} X_{12_{unlikely}} + \beta_{13} X_{13_{notveryinter.}} + \beta_{14} X_{14_{interested}} + \beta_{15} X_{15_{veryinter.}}$$

where log represents the natural logarithm and p represents the probability of the Liberals winning the Election. $\frac{1}{1-p}$ represents the odd ratio, and $log(\frac{p}{1-p})$ represents the log odds ratio. X_i on the left-hand side of the equation represents the predictor variables in the model. All of the predictor variables in this model are categorical, which means each X_is takes the value 0 or 1 (dummy variables). For example, $X_{1_{age25-34}}$ equals to 1 if the respondent aged between 25-35 and 0 otherwise. The coefficients for each of the subcategories are β_is (for i from 1 to 15), which represent the average difference in the log of odds ratio between $X_i = 0$ and $X_i = 1$ holding other variables constant. For example, $\beta_7 = 0.086$ represents the average difference in the log of odds ratio between married and unmarried respondents holding other variables constant. β_0, on the other hand, is the constant term that represents the intercept at time zero. $logit(p)$ equals to β_0 when every $X_i = 0$. ϵ is the error term of this model.

At last, we performed a post-stratification analysis to make sure the prediction is as accurate as possible. The census data is partitioned into demographic cells according to the subcategories of each predictor variable. All subcategories are divided during the post-stratification to pursue the accuracy of our estimation. Then, the chances of the Liberals winning the election is calculated within each cell, and the final estimation is calculated by the formula:

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

(Caetano, 2020)

where $\hat{y}^{PS}$ on the left-hand side of the equation is the chances of the Liberals winning the 2019 Election if everyone voted. $\hat{y}_j$ is the estimated chances within j^{th} cell, and N_j represents the population size of j^{th} cells. $\sum N_j$ represents the entire population.

7 Result

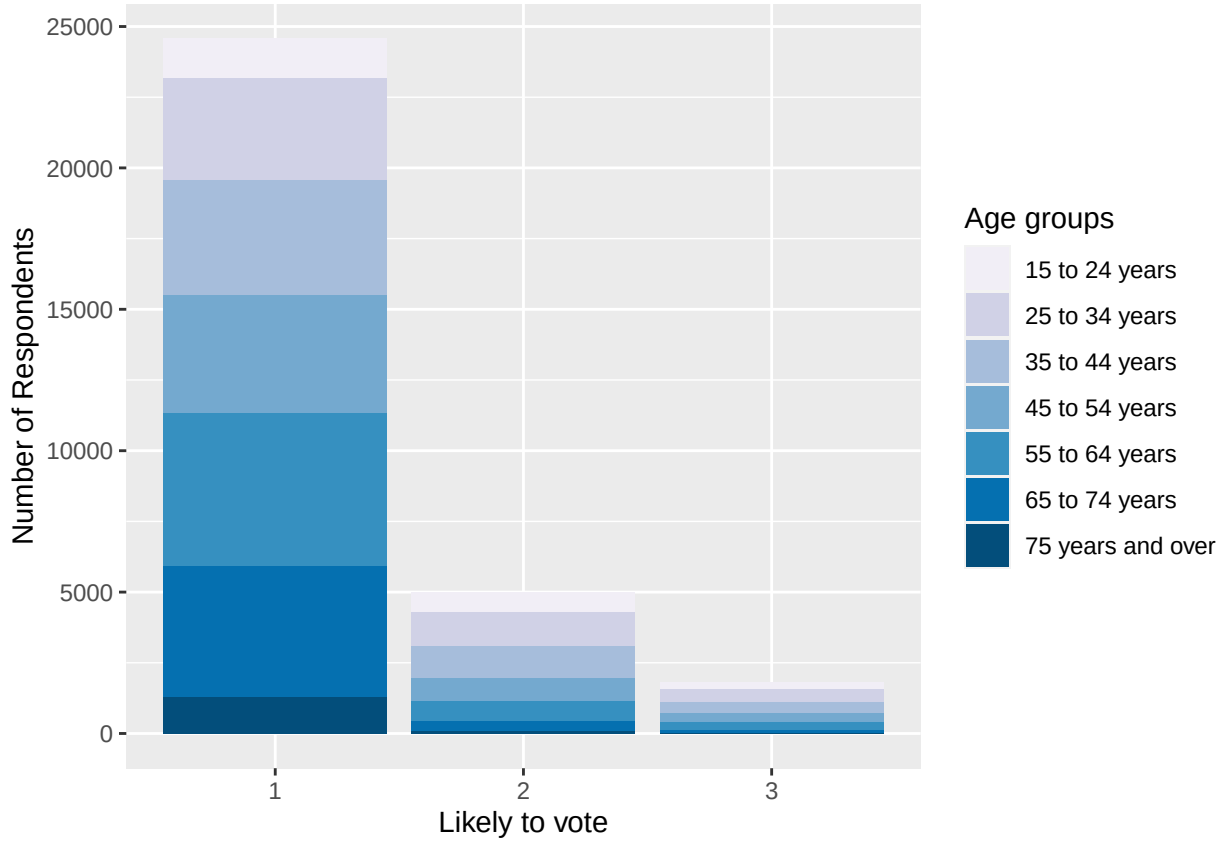

Figure 1: Relationship Between Respondents' Interest to Vote and Age Group

Figure1 shows respondents' interest to vote within each age group. `Likely_to_vote` is a variable that collects respondents' willingness to vote during the election. 1 on the x-axis means "very likely", 2 means "possibly", and 3 means "not very likely". Considering there are relatively even amount of respondents within each age group (25-74), the plot suggests that people who aged between 55 to 74 are more likely to vote in the election.

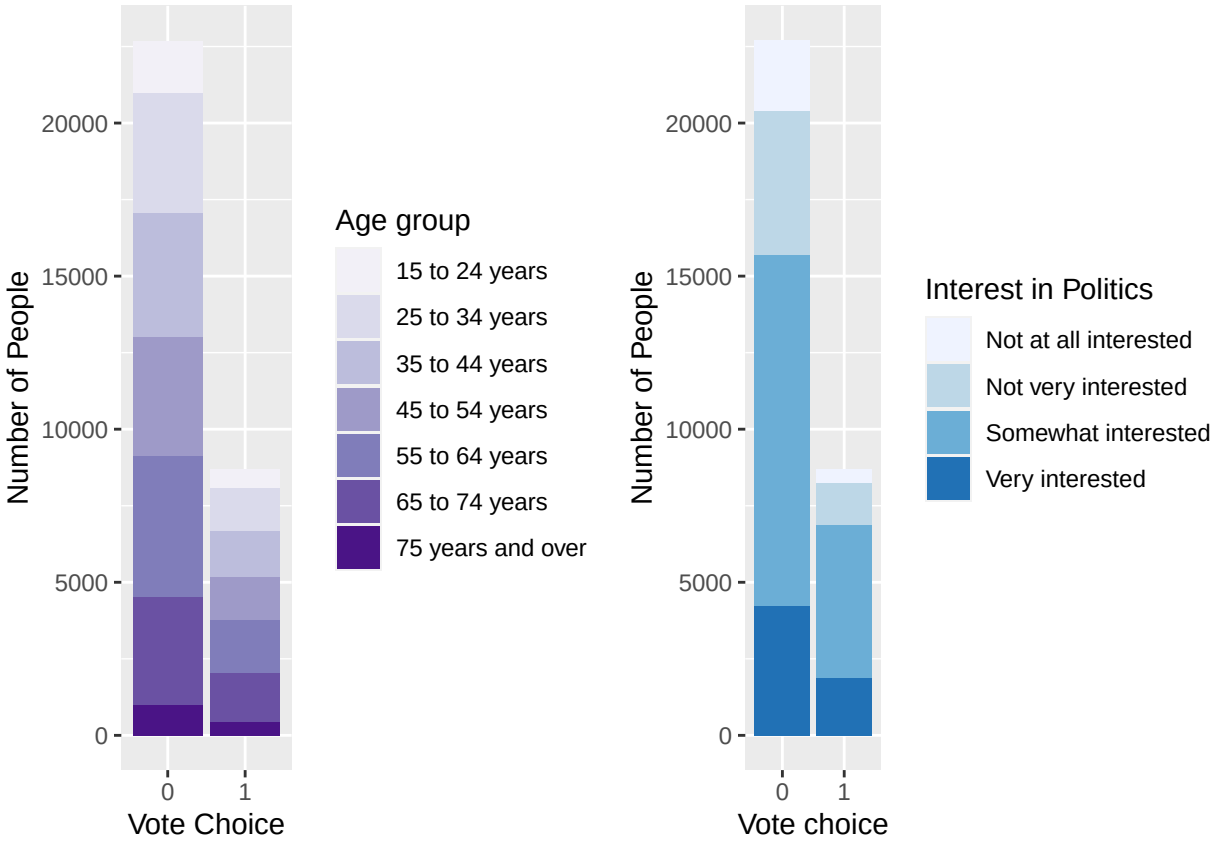

Figure 2: Relationship Between Vote Choice and the Number of voters

Figure2 contains two plots that visualize the number of people with different vote choice, each filled with a factor (`age` & `interest_in_politics`). Both plots show that there are only less than a half of the respondents that is going to vote for the Liberals. From Figure2-1, we can tell people aged 55-64 are less likely to vote for the Liberals than people from other age groups. From Figure 2-2, we can tell more than half of the respondents show at least some interest in politics, but there is still a great portion of the respondents who claimed "not interested".

Table 2: The comparison of the AIC values of two models

Models	Values
Full Model	36303.23
Final Model	36248.92

Table2 contains two AIC values. The first model we built has an AIC value of 36303. After performing the backward selection, variable `sex` is removed, and the AIC of the reduced(final) model is 36249, which suggests a better fit.

Table 3: The Summary of Final Logistic Regression Model for Election Result Estimation

Factors	Estimates	SD Error	T-values	P-value
(Intercept)	-1.614	0.080	-20.262	0.000
age25 to 34 years	-0.135	0.058	-2.324	0.020
age35 to 44 years	-0.099	0.058	-1.700	0.089
age45 to 54 years	-0.106	0.059	-1.801	0.072
age55 to 64 years	-0.055	0.057	-0.950	0.342
age65 to 74 years	0.062	0.059	1.053	0.293
age75 years and over	0.052	0.077	0.670	0.503
marital_statusOthers	0.086	0.027	3.202	0.001
grouped_educationLess than High School	-0.128	0.068	-1.882	0.060
grouped_educationPost-secondary diploma	0.137	0.035	3.897	0.000
grouped_educationUniversity degree	0.515	0.033	15.414	0.000
as.factor(likely_to_vote)2	0.017	0.037	0.450	0.652
as.factor(likely_to_vote)3	-0.676	0.075	-9.061	0.000
interest_in_politicsNot very interested	0.249	0.061	4.089	0.000
interest_in_politicsSomewhat interested	0.557	0.057	9.861	0.000
interest_in_politicsVery interested	0.524	0.062	8.497	0.000

Table3 contains the estimates, standard error, T-values, and P-values of the model. The p-value shows that some subcategories of the preditor is not statistically significant. `likely_to_vote` and `interest_in_politics` has a stronger relationship with the vote choice. By substituting the intercept and the coefficient, the final model is shown as below:

$$log(\frac{\hat{p}}{1-\hat{p}}) = -1.614 - 0.135X_{1_{age25-34}} - 0.099X_{2_{age35-44}} - 0.106X_{3_{age45-54}} - 0.055X_{4_{age55-64}}$$

$$+0.062X_{5_{age65-74}} + 0.052X_{6_{age75+}} + 0.086X_{7_{NotMarried}} - 0.128X_{8_{highschool}} + 0.137X_{9_{postsec}} + 0.515X_{10_{uni}}$$

$$+0.017X_{11_{likely}} - 0.676X_{12_{unlikely}} + 0.249X_{13_{notveryinter.}} + 0.557X_{14_{interested}} + 0.524X_{15_{veryinter.}}$$

By using the model above, we performed the post-stratification to estimate the likelihood of the Liberals winning the election if everyone have voted.

Table 4: The Likelihood of The Liberals Winning the Election If Everyone Have Voted

predict
0.2549516

As shown in Table4, the estimated value is 0.255, which suggests that the Liberals have 25.5% chance of winning the election.

8 Discussion

The goal of this study is to estimate the possible outcome of the 2019 Canadian Federal Election if everyone has voted. The study is based on the Canadian Election Study 2019 Online Survey which was conducted before and after the election. Also, the General Social Survey (Cycle 27: Social Identity) is used as the census data for post-stratification. To get an accurate estimation, we need to get as many observations as possible, especially those people who are not likely to vote during the election. `cps19_votechoice` is a variable that collects the vote choice of those respondents who claimed to vote, which is not enough for the study. Thus, I included `cps19_vote_unlikely`, which collects the vote choice of those respondents who are not likely to vote during the election. By combining these two variables, the data is more comprehensive and ready to be analyzed.

From Figure1, we can tell most of the respondents are likely to vote during the election. As mentioned at the beginning of the report, 65.95% of the eligible voters voted during the 2019 Canadian Federal Election, which matches what is shown in the plot. Also, we can tell from the plot that older voters are more willing to vote, which shows that they care more about who gets to lead the government in comparison to young people. That being said, if a party supports seniors policies and values those old Canadians who shaped the country in the past, they are very likely to gain more support during the election.

According to Figure2, less than half of the respondents favor the Liberals. There is no obvious difference in vote choice within each age group. People aged 55-64 are slightly less likely to vote for the Liberals. Also, from Figure2 -2, we can tell there is still a big portion of the population who are not interested in politics. As mentioned at the beginning, "not interested in politics" is a major reason why some eligible voters don't vote. Those people don't have a belief about political parties that is deep-rooted in their minds, and they might not know much about politics. For most of the parties, those votes are relatively easy to get if they put the effort in it.

The summary table of the model suggests that most of the subcategories of variables are related to the vote choice. Surprisingly, many subcategories under the variable `age` have a p-value >0.05, which suggests that they are not statistically significant. The final estimation calculated using post-stratification is **0.255**, which suggests that only 25.5% of the voters may vote for the Liberal if everyone participated. This result is not very close to the number of popular votes but demonstrates a similar idea - the Liberal might not win this election if everyone has voted.

8.1 Weaknesses

This study is based on the CES dataset which contains survey data. Although the survey is conducted online, it is still very likely to have response bias. Many useful variables contain a large amount of missing values, which are not able to use. The accuracy of the model may improve if we can have more observations for some of the variables. Also, the information in our census dataset is collected during 2013 and 2014, but the survey dataset is built-in 2019. The population structure and many other factors must change during these years, which may also affect the accuracy of the prediction.

8.2 Next Steps

This study can be used as a reference for those parties who want to gain more supporters and get representation in the House of Common. As discussed earlier, people aged 55-75 have a higher voter turnout, which suggests that giving some necessary support to seniors may be beneficial for getting more supports. Also, there are still approximately 25% of the eligible voters who did not participate in the 2019 Canadian Federal Election. The party that can get these votes might be the "winner" of the next election.

9 References

Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Angelo Canty and Brian Ripley (2020). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-25.

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Bowen, W. (2019). Do the math — Canadians aren't getting the government they're voting for. Retrieved from https://www.cbc.ca/news/opinion/opinion-warren-bowen-electoral-change-1.53 33743

Britneff, B. (2019). Canada election: The 2019 results by the numbers. Retrieved from https://globalnews.ca/news/6066524/canada-election-the-2019-results-by-the-numbers/

Caetano,S. (2020). STA304: Multilevel Regression & Poststratification. Pg.4.

Cao, S., Fu, X., Su, Y.& Huang, Y. (2020). The Study of How Family Status Positively Affects Life Satisfaction.

Cao, S., Fu, X., Su, Y.& Huang, Y. (2020). The Study of the Likelihood of Donald Trump Winning the 2020 Election.

Carpenter, J. (2018). Justin Trudeau is a feminist. For him, that's a given. Retrieved from https://www.cnn.com/2018/11/12/success/justin-trudeau-feminism/index.html

Chen,H.W., Hu,X. & Yang,Z.C. (n.d.). Model Selection for Linear Regression Model. Retrieved from https://jbhender.github.io/Stats506/F17/Projects/Group21_Model_Selection.html

Davison, A. C. & Hinkley, D. V. (1997) Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge. ISBN 0-521-57391-2

Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2020). skimr: Compact and Flexible Summaries of Data. R package version 2.1.2. https://CRAN.R-project.org/package=skimr

General Social Survey, cycle 27, 2013 (version 2): Social Identity.Retreived from https://https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/subsda3

Hadley Wickham, Romain Fran?ois, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.0. https://CRAN.R-project.org/package=dplyr

Huang, Y. (2020). The Analysis of the Relationship Between Population's Level of Satisfaction with Justin Trudeau and Their Education Level.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

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

Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.

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

Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2019 Canadian Election Study – Online Collection. [dataset]

Tierney N (2017). "visdat: Visualising Whole Data Frames." *JOSS*, *2*(16), 355. doi: 10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355>.

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

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595