How the voter turnout in the federal election has an impact on the outcome of the election

22/12/2020

Option B

Author: Jiaheng Li(lijiahe5, 1003825088)

Abstract

This project focuses on whether the final result is the same, if "everyone" voted in the 2019 Canadian Federal Election, based on the 2019 Canadian Election Study and Canadian general social surveys. Knowledge points of statistics such as the multiple logistic models, post-stratification and barplot, are used in the analyses. The purpose of this project is to make people know the importance of voting in the general election

Keywords

Binary-logistic model, Post-stratification, Vote, Bar plot, Logistic Regression, Election

Introduction

The Canadian federal election held every four years is an important opportunity for Canadians to express their views on the country's development and contribute to its growth in the next four years. However, although some people are dissatisfied with the current national policies and economic trends, they did not vote for them in the federal election. Maybe they feel that one of their votes does not affect the general election outcome, nor can it affect the country's future development direction, or they feel that the federal election is not essential.

In this project, "Everyone" represents all people who have the right to vote in the 2019 Canadian federal elections. In the process, we will select a couple of variables that may influence the Canadian federal election from the dataset 2019 Canadian Election Study and Canada Census datasets and analyze the dataset using a multi-logistic model with post-stratification. Then use the results to prove to people that everyone's votes in the general election may change the outcome of the final federal election and the country's future direction. People need to exercise their rights in federal elections and use voting to express their views.

In the following part of the report, we will introduce our selected data, variables, and the statistical models that help us predict the result of the election that all people who meet the voting conditions of the general election will vote. The following part also contains more information that introduces how to use CES2019 to make a multi-logistic model and make post-stratification by using census datasets; what the final result is. Eventually, we will conclude based on the statistical model and analysis result, explaining the importance of turnout.

Methodology

Data

The multi-logistic model will be used to re-predict the result of the 2019 Candian Federal election that suppose that all the people who have the right to vote in 2019 vote in this election. Therefore, two datasets are ready for this project.

The 2019 Canadian Election Study online Survey data (sample data)

Source of the data

We select the dataset which is 2019 Canadian Election Study online Survey data as the sample data and uploaded from the website of Canadian Election Study.

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

Taget population, Sampling Frame and Sample population

For this data set, the target population is all people living in Canada. The sampling frame is people living in Canada and usually browses webs. The sample population is people who are in the sampling frame, and noticed the online survey and completed the survey.

Sampling Methodology

The goal of this survey is "to gather enough data to allow for constituency-level analysis as well as proper subgroup analysis of populations that are typically underrepresented in the CES"[1]. Besides, the core questions that emerged in the survey remain unchanged from the previous survey, and "because of the large sample size and the ability of an online platform to easily implement complex designs, investigators can imbed many community projects into the larger online survey"[1].

Key features, Strengths, Weakness

For the accuracy of the established statistical model, the 2019 Canadian Election Study online Survey dataset would be used in this project because "the CES has been a rich source of data on Canadians' political behaviour and attitudes, measuring preferences on key political issues such as free trade with the US, social spending and Quebec's place in Canada"[1]. Moreover, "The 2019 CES will allow for a more fine-grained analysis of barriers to electoral participation and preferences at the micro-level than previous studies have been able to provide."[1] However, more variables can be added to this survey data, such as income, marital status, employment status, and so on. This helps people use the survey results to perform some statistical analysis to get more accurate results.

Variables selected

On the other hand, there are 37822 rows and 620 variables in this data set. The variables gender, age, education, provinces and vote choices were selected from the dataset. Then we created six new binary variables that were respectively vote_literal, vote_ndp, vote_green, vote_conservative, vote_people, and vote_Bloc, based on the data in the column of votechoices. Furthermore, rows in which respondents are under 18 until 2019 cleaned from the dataset. The code in clean_data.R completed the filter and cleaned off the dataset. After filter and clean, the dataset contains 20860 rows and 11 variables, and we used it for building binary-logistic models.

Raw table for sample data

```
##                       Stratified by votechoice
##                        Another     Bloc_Quebecois Conservative_Party Green_Party
##   n                    161         1140           6601               1932
##   gender (%)
##      Female             77 (47.8)   544 ( 47.7)   3283 (49.7)        1170 (60.6)
##      Male               81 (50.3)   591 ( 51.8)   3287 (49.8)         740 (38.3)
##      Others              3 ( 1.9)     5 (  0.4)     31 ( 0.5)          22 ( 1.1)
##   age_group (%)
##      adult              51 (31.7)   272 ( 23.9)   2160 (32.7)         687 (35.6)
##      old_people         71 (44.1)   609 ( 53.4)   2767 (41.9)         615 (31.8)
##      very_old_people    26 (16.1)   181 ( 15.9)    948 (14.4)         191 ( 9.9)
##      young_people       13 ( 8.1)    78 (  6.8)    726 (11.0)         439 (22.7)
##   education (%)
##      high               69 (42.9)   435 ( 38.2)   2885 (43.7)         928 (48.0)
##      low                 6 ( 3.7)    83 (  7.3)    363 ( 5.5)          92 ( 4.8)
##      medium             86 (53.4)   622 ( 54.6)   3353 (50.8)         912 (47.2)
##   province (%)
##      AB                 15 ( 9.3)     0 (  0.0)   1501 (22.7)         122 ( 6.3)
##      BC                 25 (15.5)     0 (  0.0)    744 (11.3)         359 (18.6)
##      MB                  4 ( 2.5)     0 (  0.0)    378 ( 5.7)          88 ( 4.6)
##      NB                  1 ( 0.6)     0 (  0.0)    122 ( 1.8)          92 ( 4.8)
##      NL                  3 ( 1.9)     0 (  0.0)     70 ( 1.1)          11 ( 0.6)
##      NS                  8 ( 5.0)     0 (  0.0)    126 ( 1.9)          78 ( 4.0)
##      NT                  0 ( 0.0)     0 (  0.0)      3 ( 0.0)           2 ( 0.1)
##      ON                 59 (36.6)     0 (  0.0)   2464 (37.3)         735 (38.0)
##      PE                  1 ( 0.6)     0 (  0.0)     15 ( 0.2)          22 ( 1.1)
##      QC                 36 (22.4)  1140 (100.0)    806 (12.2)         377 (19.5)
##      SK                  9 ( 5.6)     0 (  0.0)    367 ( 5.6)          42 ( 2.2)
##      YT                  0 ( 0.0)     0 (  0.0)      5 ( 0.1)           4 ( 0.2)
##                       Stratified by votechoice
##                        Liberal_Party NDP          People_Party p      test
##   n                    7205          3362         459
##   gender (%)                                                   <0.001
##      Female            4057 (56.3)   2235 (66.5)  205 (44.7)
##      Male              3110 (43.2)   1061 (31.6)  251 (54.7)
##      Others              38 ( 0.5)     66 ( 2.0)    3 ( 0.7)
##   age_group (%)                                                <0.001
##      adult             2428 (33.7)   1328 (39.5)  217 (47.3)
##      old_people        2750 (38.2)    857 (25.5)  124 (27.0)
##      very_old_people    951 (13.2)    226 ( 6.7)   27 ( 5.9)
##      young_people      1076 (14.9)    951 (28.3)   91 (19.8)
##   education (%)                                                <0.001
##      high              4011 (55.7)   1545 (46.0)  168 (36.6)
##      low                249 ( 3.5)    198 ( 5.9)   42 ( 9.2)
##      medium            2945 (40.9)   1619 (48.2)  249 (54.2)
##   province (%)                                                 <0.001
##      AB                 451 ( 6.3)    315 ( 9.4)   61 (13.3)
##      BC                 748 (10.4)    509 (15.1)   52 (11.3)
##      MB                 279 ( 3.9)    174 ( 5.2)   14 ( 3.1)
##      NB                 169 ( 2.3)     42 ( 1.2)   15 ( 3.3)
##      NL                 149 ( 2.1)     73 ( 2.2)    5 ( 1.1)
##      NS                 252 ( 3.5)     81 ( 2.4)   11 ( 2.4)
##      NT                   7 ( 0.1)      2 ( 0.1)    1 ( 0.2)
##      ON                3345 (46.4)   1461 (43.5)  178 (38.8)
##      PE                  26 ( 0.4)      6 ( 0.2)    0 ( 0.0)
##      QC                1658 (23.0)    512 (15.2)  101 (22.0)
##      SK                 115 ( 1.6)    180 ( 5.4)   21 ( 4.6)
##      YT                   6 ( 0.1)      7 ( 0.2)    0 ( 0.0)
```

The Canadian general social surveys (census data)

Source of the data

The data of the Canadian general social surveys are provided by Statistics Canada under the terms of the Data Liberation Initiative (DLI) and "General social survey on Family (cycle 31), 2017" was selected as the census data.

Welcome to my.access – please choose how you will connect. (n.d.). Retrieved December 20, 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm

Taget population, Sampling Frame and Sample population

The target population is including all people who are 15 years old or older living in Canada. The Sampling Frame contains people in the target population and whose contact information has been stored in Statistics Canada and the Address Register. The sampling population is the people who are included in the sampling frame and complete the survey.

Sampling Methodology

Investigators interviewed the target group and encouraged them to complete the survey by calling at a specific time during the day. For those who refuse to be interviewed for the first time, investigators would call them again and express their wish to participate in the survey.

Key features, Strengths, Weakness

The area covered by this survey is vast, and the number of people accepted is large. These make the results of the survey more convincing and universal. On the other hand, each respondent's information included in the survey results is very rich and diverse, including gender, age, whether they are married, whether they have children, etc.

However, because the survey's primary channel is to make phone calls, some Canadian residents who do not have mobile phones are not given the opportunity to participate in the survey.

Variables selected

In this dataset, there are 20602 rows and 461 variables. We select four variables, age, sex, education, and provinces, that match the sample data to make the prediction using the fitted model. We have modified the expression of values corresponding to some variables to obtain exact matches of the two datasets' variables. Since every person who is a Canadian citizen and who on polling day is 18 years of age or older is qualified as an elector[2], we cleaned the rows which the respondents are under 18 until 2019 or are not Canadian citizens. We used gss_dict.txt, gss_dict.txt, and gss_cleaning.R(provided in the Problem set 2) to clean the dataset and filtered by clean_data.R. Finally, we make the post-stratification for this dataset, and the result will be shown in the following sections.

Raw table for census data

```
##                      Stratified by province
##                       AB          BC          MB          NB
##   n                   24          24          24          24
##   gender = Male (%)   12 ( 50.0)  12 ( 50.0)  12 ( 50.0)  12 ( 50.0)
##   age_group (%)
##      adult             6 ( 25.0)   6 ( 25.0)   6 ( 25.0)   6 ( 25.0)
##      old_people        6 ( 25.0)   6 ( 25.0)   6 ( 25.0)   6 ( 25.0)
##      very_old_people   6 ( 25.0)   6 ( 25.0)   6 ( 25.0)   6 ( 25.0)
##      young_people      6 ( 25.0)   6 ( 25.0)   6 ( 25.0)   6 ( 25.0)
##   education (%)
##      high              8 ( 33.3)   8 ( 33.3)   8 ( 33.3)   8 ( 33.3)
##      low               8 ( 33.3)   8 ( 33.3)   8 ( 33.3)   8 ( 33.3)
##      medium            8 ( 33.3)   8 ( 33.3)   8 ( 33.3)   8 ( 33.3)
##   province (%)
##      AB               24 (100.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      BC                0 (  0.0)  24 (100.0)   0 (  0.0)   0 (  0.0)
##      MB                0 (  0.0)   0 (  0.0)  24 (100.0)   0 (  0.0)
##      NB                0 (  0.0)   0 (  0.0)   0 (  0.0)  24 (100.0)
##      NL                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      NS                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      ON                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      PE                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      QC                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      SK                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##                      Stratified by province
##                       NL          NS          ON          PE
##   n                   24          24          24          24
##   gender = Male (%)   12 ( 50.0)  12 ( 50.0)  12 ( 50.0)  12 ( 50.0)
##   age_group (%)
##      adult             6 ( 25.0)   6 ( 25.0)   6 ( 25.0)   6 ( 25.0)
##      old_people        6 ( 25.0)   6 ( 25.0)   6 ( 25.0)   6 ( 25.0)
##      very_old_people   6 ( 25.0)   6 ( 25.0)   6 ( 25.0)   6 ( 25.0)
##      young_people      6 ( 25.0)   6 ( 25.0)   6 ( 25.0)   6 ( 25.0)
##   education (%)
##      high              8 ( 33.3)   8 ( 33.3)   8 ( 33.3)   8 ( 33.3)
##      low               8 ( 33.3)   8 ( 33.3)   8 ( 33.3)   8 ( 33.3)
##      medium            8 ( 33.3)   8 ( 33.3)   8 ( 33.3)   8 ( 33.3)
##   province (%)
##      AB                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      BC                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      MB                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      NB                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      NL               24 (100.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      NS                0 (  0.0)  24 (100.0)   0 (  0.0)   0 (  0.0)
##      ON                0 (  0.0)   0 (  0.0)  24 (100.0)   0 (  0.0)
##      PE                0 (  0.0)   0 (  0.0)   0 (  0.0)  24 (100.0)
##      QC                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##      SK                0 (  0.0)   0 (  0.0)   0 (  0.0)   0 (  0.0)
##                      Stratified by province
##                       QC          SK          p      test
##   n                   24          24
##   gender = Male (%)   12 ( 50.0)  12 ( 50.0)   1.000
##    age_group (%)                                      1.000
##       adult            6 ( 25.0)   6 ( 25.0)
##       old_people       6 ( 25.0)   6 ( 25.0)
##       very_old_people  6 ( 25.0)   6 ( 25.0)
##       young_people     6 ( 25.0)   6 ( 25.0)
##    education (%)                                      1.000
##       high             8 ( 33.3)   8 ( 33.3)
##       low              8 ( 33.3)   8 ( 33.3)
##       medium           8 ( 33.3)   8 ( 33.3)
##    province (%)                                      <0.001
##       AB               0 (  0.0)   0 (  0.0)
##       BC               0 (  0.0)   0 (  0.0)
##       MB               0 (  0.0)   0 (  0.0)
##       NB               0 (  0.0)   0 (  0.0)
##       NL               0 (  0.0)   0 (  0.0)
##       NS               0 (  0.0)   0 (  0.0)
##       ON               0 (  0.0)   0 (  0.0)
##       PE               0 (  0.0)   0 (  0.0)
##       QC              24 (100.0)   0 (  0.0)
##       SK               0 (  0.0)  24 (100.0)
```

Model

The final purpose of this project is to predict the results of the 2019 Canadian Federal Elections that assume all people who are qualified as electors vote in this election. Therefore, to achieve this, we built five different binary logistic models by using data from the 2019 Canadian Election Study online Survey data. We choose to build binary logistic models because the value corresponding to the variables vote_literal, vote_ndp, vote_green, vote_conservative, vote_people, and vote_Bloc are all 0 or 1(vote is 1, not vote is 0). On the other hand, Logistic regression is "a classification algorithm used to find the probability of event success and event failure. It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions."[3] After building models, we make post-stratification that applies fitted models to the census data from The Canadian general social surveys to predict the outcome of each party's general election and compute the probability of each party.

Model Introduction

We used four variable as explanatory variable to build five different binary logistic model. x_{gender}: the gender of the respondent in the sample data. There are two categories:

1. male;
2. female.

x_{age}:Age ranges of respondents in the sample data. It has four categories:

1. Very old people (>70);
2. old people (50 - 70);
3. adult (30 - 50);
4. young people(18 - 30).

$x_{education}$:the level of education identify of respondents in the sample data. It has three categories:

1. high(Bachelor degree or above);

2. medium(A high school degree or above, but no bachelor's degree);
3. low(Didn't get a high school degree).

$x_{provinces}$:provinces where respondents live in the sample data. It has 13 categories:

1. AB (Alberta);
2. BC (British Columbia);
3. MB (Manitoba);
4. NB (New Brunswick);
5. NL (Newfoundland and Labrador);
6. NS (Nova Scotia);
7. ON (Ontario);
8. PE (Prince Edward Island);
9. QC (Quebec);
10. SK (Saskatchewan);
11. NT (Northwest Territories);
12. NU (Nunavut);
13. YT (Yukon.)

For the independent variable age, we change it from numerical to categorical. In other words, we select to use age groups to build the binary logistic model and make the bar plots. The reason is that the categorical variable age can be better interpreted and analyzed in the table and some kinds of plots. Besides, for the variable education, it is categorical in the initial data set, but the types of values in this variable are too miscellaneous. Therefore, we categorize all its values into three levels(high, medium and low). The other two variables remain unchanged

- Liberal party

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{genderMale} + \beta_2 X_{genderOthers} + \beta_3 X_{provinceBritishColumbia} + \beta_4 X_{provinceManitoba} +$$
$$\beta_5 X_{provinceNewBrunswick} + \beta_6 X_{provinceNewfoundlandAndLabrador} + \beta_7 X_{provinceNorthwestTerritories} +$$
$$\beta_8 X_{provinceNovaScotia} + \beta_9 X_{provinceOntario} + \beta_{10} X_{provincePrinceEdwardIsland} + \beta_{11} X_{provinceQuebec} + \beta_{12}$$
$$X_{provinceSaskatchewan} + \beta_{13} X_{provinceYukon} + \beta_{14} X_{educationlow} + \beta_{15} X_{educationmedium} + \beta_{16} X_{yobOldPeople} +$$
$$\beta_{17} X_{yobveryOldPeople} + \beta_{18} X_{yobyoungPeople}$$

β_0 represents the intercept of this binary logistic model for Liberal Party

β_1 to β_2 represents the relationship between the age of the respondents and which party they have chosen to vote for

β_3 to β_{13} represents the relationship between the provinces where the respondents live and whether they voted for the Liberal Party

β_{14} to β_{15} represents the relationship between the level of education of respondents and whether they choose to support Liberal Party in the elelction.

β_{16} to β_{18} represents the relationship between different age groups of respondents and whether they voted for the Liberal Party

$log(\frac{p}{1-p})$ represent the probability of the Liberal Party's total number of votes in the 2019 Canadian Federal Elections. $vote_liberal$=1 means that this respondents voted for Liberal Party, while $vote_liberal$=0 means that this respondents did not vote for Liberal Party.

The models for the other four parties are the same as this binary logistic model for the Liberal Party but build by using different dependent variables. Once these five models were set up, we made post-stratification with them one by one and calculate the probability of the total number of votes of different parties in the 2019 Canadian Federal Elections.

Post Stratification

The technique of Post Stratification is beneficial, “it allows the estimating of preference within a specific locality based on a survey taken across a wider area that includes relatively few people from the locality in question, or where the sample may be highly unrepresentative”[2]. On the other hand, it makes statistical analysis more effective, accurate, and less cost. This is the reason why we implement this technique in this project.

After getting six binary logistic models, we first use R function predict() to compute the log-odds estimate(p) for each group in the census dataset(at the beginning, we use group_by to divide all rows in the census dataset into different groups). Then we implement the equation $log(\frac{p}{1-p})$ to compute the average probability of voting for a particular each group in the dataset. Finally, we sum all probability of voting for this party of each voter and divide by the total number of voters to get the probability of voting for this particular party in the 2019 Canadian Federal Election.

Albernative models

About alternative models, since the question has been stipulated to use MRP with post-stratification, the only model types we can choose are binary logistic regression model and normal multilevel logistic regression. Finally, we decided on the binary logistic regression model because we think this model allows us to change better to analyze and get more desired results.

Results

Results of building models and making post stratification

As mentioned in section Model, we firstly built binary logistic models for five different parties and applied these fitted models to make the post-stratification based on the census dataset. For this part, we use function glm() to help to build these five binary logistic model.

$$model_{literal} = glm(vote\_Literal \sim gender + province + education + age\_group, data = survey, binomial)$$

$$model_{conserative} = glm(vote\_Conserative \sim gender+province+education+age\_group, data = survey, binomial)$$

$$model_{green} = glm(vote\_Green \sim gender + province + education + age\_group, data = survey, binomial)$$

$$model_{bloc} = glm(vote\_Bloc \sim gender + province + education + age\_group, data = survey, binomial)$$

$$model_{people} = glm(vote\_People \sim gender + province + education + age\_group, data = survey, binomial)$$

Table 1: Literal Party

term	estimate	std.error	statistic	p.value
(Intercept)	-1.2700320	0.0584600	-21.7248179	0.0000000
genderMale	-0.1004581	0.0306050	-3.2824115	0.0010292
genderOthers	-0.5774408	0.1890706	-3.0541016	0.0022574
provinceBC	0.6596848	0.0685983	9.6166393	0.0000000
provinceMB	0.6347471	0.0889372	7.1370239	0.0000000
provinceNB	1.0033775	0.1116637	8.9857055	0.0000000
provinceNL	1.4121740	0.1258167	11.2240595	0.0000000
provinceNS	1.2765072	0.1005455	12.6958226	0.0000000

Table 1: Literal Party *(continued)*

term	estimate	std.error	statistic	p.value
provinceNT	1.3810831	0.5224717	2.6433642	0.0082087
provinceON	1.0758443	0.0570726	18.8504377	0.0000000
provincePE	0.9679746	0.2543500	3.8056794	0.0001414
provinceQC	0.9060068	0.0608233	14.8957149	0.0000000
provinceSK	-0.1906156	0.1145975	-1.6633478	0.0962428
provinceYT	0.4955794	0.4845673	1.0227257	0.3064375
educationlow	-0.6834414	0.0768602	-8.8920050	0.0000000
educationmedium	-0.4118299	0.0309309	-13.3145061	0.0000000
age_groupold_people	0.1197445	0.0356242	3.3613201	0.0007757
age_groupvery_old_people	0.1904263	0.0495407	3.8438336	0.0001211
age_groupyoung_people	-0.0429827	0.0460806	-0.9327723	0.3509375

Table 2: Conservative Party

term	estimate	std.error	statistic	p.value
(Intercept)	0.1175120	0.0506603	2.319607	0.0203622
genderMale	0.3796164	0.0321192	11.818988	0.0000000
genderOthers	-0.5231439	0.2081669	-2.513099	0.0119676
provinceBC	-1.3543442	0.0616119	-21.981845	0.0000000
provinceMB	-0.8798496	0.0798093	-11.024400	0.0000000
provinceNB	-1.4733291	0.1157355	-12.730144	0.0000000
provinceNL	-1.7352611	0.1436867	-12.076697	0.0000000
provinceNS	-1.7335077	0.1108596	-15.636965	0.0000000
provinceNT	-1.9051371	0.6508975	-2.926939	0.0034232
provinceON	-1.3137789	0.0488313	-26.904438	0.0000000
provincePE	-1.8356344	0.2979799	-6.160262	0.0000000
provinceQC	-2.1253377	0.0579376	-36.683231	0.0000000
provinceSK	-0.4780513	0.0860505	-5.555473	0.0000000
provinceYT	-1.6997598	0.5148003	-3.301785	0.0009607
educationlow	0.2983828	0.0733090	4.070208	0.0000470
educationmedium	0.2828135	0.0327824	8.627005	0.0000000
age_groupold_people	0.2372111	0.0372219	6.372883	0.0000000
age_groupvery_old_people	0.3120910	0.0512424	6.090488	0.0000000
age_groupyoung_people	-0.5211098	0.0519266	-10.035500	0.0000000

Table 3: Green Party

term	estimate	std.error	statistic	p.value
(Intercept)	-2.8550116	0.1023893	-27.8838945	0.0000000
genderMale	-0.1856083	0.0504374	-3.6799730	0.0002333
genderOthers	0.1905034	0.2337751	0.8149002	0.4151294
provinceBC	1.2474129	0.1095007	11.3918212	0.0000000
provinceMB	0.7055087	0.1458615	4.8368400	0.0000013
provinceNB	1.6588344	0.1501431	11.0483544	0.0000000

Table 3: Green Party *(continued)*

term	estimate	std.error	statistic	p.value
provinceNL	-0.3434369	0.3209119	-1.0701906	0.2845335
provinceNS	1.1759643	0.1539846	7.6368947	0.0000000
provinceNT	1.1126666	0.7671377	1.4504131	0.1469433
provinceON	0.6344080	0.1009772	6.2826886	0.0000000
provincePE	2.2132243	0.2755294	8.0326239	0.0000000
provinceQC	0.5827702	0.1076280	5.4146716	0.0000001
provinceSK	0.1682025	0.1843749	0.9122852	0.3616186
provinceYT	1.4776557	0.5623754	2.6275257	0.0086008
educationlow	-0.0508994	0.1162385	-0.4378874	0.6614679
educationmedium	0.0038145	0.0502356	0.0759315	0.9394736
age_groupold_people	-0.2197857	0.0591874	-3.7133888	0.0002045
age_groupvery_old_people	-0.3029257	0.0867699	-3.4911386	0.0004810
age_groupyoung_people	0.3324834	0.0666119	4.9913498	0.0000006

Table 4: NDP

term	estimate	std.error	statistic	p.value
(Intercept)	-1.7028503	0.0695789	-24.4736713	0.0000000
genderMale	-0.4597435	0.0414058	-11.1033661	0.0000000
genderOthers	0.8089854	0.1651009	4.8999461	0.0000010
provinceBC	0.7179405	0.0800996	8.9630919	0.0000000
provinceMB	0.4956671	0.1057285	4.6881138	0.0000028
provinceNB	-0.2673475	0.1751878	-1.5260624	0.1269943
provinceNL	0.7920969	0.1504973	5.2631951	0.0000001
provinceNS	0.2322653	0.1370093	1.6952525	0.0900275
provinceNT	0.0855901	0.7704489	0.1110912	0.9115441
provinceON	0.4294894	0.0683157	6.2868347	0.0000000
provincePE	-0.3820699	0.4360648	-0.8761770	0.3809338
provinceQC	-0.0517859	0.0777401	-0.6661412	0.5053208
provinceSK	0.8580450	0.1073320	7.9943085	0.0000000
provinceYT	1.2171954	0.4757261	2.5586053	0.0105093
educationlow	0.2266485	0.0869714	2.6060104	0.0091604
educationmedium	0.1090597	0.0405090	2.6922316	0.0070976
age_groupold_people	-0.5955289	0.0482328	-12.3469776	0.0000000
age_groupvery_old_people	-0.8104108	0.0772056	-10.4967912	0.0000000
age_groupyoung_people	0.4605511	0.0502867	9.1585120	0.0000000

Table 5: Bloc Quebecois

term	estimate	std.error	statistic	p.value
(Intercept)	-22.1324538	580.5506398	-0.0381232	0.9695894
genderMale	0.0826319	0.0708259	1.1666902	0.2433355
genderOthers	-0.0575905	0.5045230	-0.1141485	0.9091201
provinceBC	-0.0578965	823.8247371	-0.0000703	0.9999439
provinceMB	-0.0493322	1105.3790027	-0.0000446	0.9999644

Table 5: Bloc Quebecois *(continued)*

term	estimate	std.error	statistic	p.value
provinceNB	-0.0080786	1490.9540871	-0.0000054	0.9999957
provinceNL	0.0104526	1736.3629040	0.0000060	0.9999952
provinceNS	-0.0452813	1353.9784828	-0.0000334	0.9999733
provinceNT	-0.0617987	7451.4519663	-0.0000083	0.9999934
provinceON	0.0332530	661.6070399	0.0000503	0.9999599
provincePE	-0.0583195	3481.7411856	-0.0000168	0.9999866
provinceQC	20.4102203	580.5506358	0.0351567	0.9719548
provinceSK	-0.0288847	1212.0024521	-0.0000238	0.9999810
provinceYT	0.0594749	6190.1056620	0.0000096	0.9999923
educationlow	0.8165719	0.1518198	5.3785603	0.0000001
educationmedium	0.2291899	0.0735551	3.1158954	0.0018339
age_groupold_people	0.7444358	0.0844674	8.8132922	0.0000000
age_groupvery_old_people	0.7382495	0.1135279	6.5028014	0.0000000
age_groupyoung_people	-0.4106833	0.1399065	-2.9354130	0.0033310

Table 6: People's party

term	estimate	std.error	statistic	p.value
(Intercept)	-3.9437513	0.1594403	-24.7349674	0.0000000
genderMale	0.5776647	0.0966390	5.9775542	0.0000000
genderOthers	-0.0917450	0.5888295	-0.1558091	0.8761835
provinceBC	-0.0671508	0.1919891	-0.3497635	0.7265162
provinceMB	-0.4547619	0.2997428	-1.5171738	0.1292228
provinceNB	0.3630863	0.2946184	1.2323953	0.2178015
provinceNL	-0.4332215	0.4704923	-0.9207833	0.3571636
provinceNS	-0.1174928	0.3322610	-0.3536161	0.7236266
provinceNT	1.0943479	1.0472639	1.0449591	0.2960419
provinceON	-0.0700565	0.1513750	-0.4628009	0.6435071
provincePE	-12.8237292	282.2927441	-0.0454271	0.9637669
provinceQC	-0.0920586	0.1650884	-0.5576322	0.5770956
provinceSK	0.1571665	0.2577613	0.6097366	0.5420363
provinceYT	-12.8540005	505.4170511	-0.0254325	0.9797100
educationlow	0.9418921	0.1779093	5.2942278	0.0000001
educationmedium	0.5173083	0.1027592	5.0341815	0.0000005
age_groupold_people	-0.7726228	0.1151834	-6.7077628	0.0000000
age_groupvery_old_people	-1.2334225	0.2068358	-5.9632928	0.0000000
age_groupyoung_people	-0.1711273	0.1290269	-1.3262913	0.1847432

Since post-stratification is the process of adjusting the estimates, essentially a weighted average of estimates from all possible combinations of attributes[4], we made the post-stratification by using the census dataset and these five completed fitted binary logistic regression models to predict the probability that each Canadian party would get the votes in the popular vote. In this part, We use the steps of making poststratification mentioned in the Model section to get the final vote rate of each party in the general election.

The below table shows the predicted probability of each party in the 2019 Canadian Federal Election which "everyone" voted, after making the post-stratification

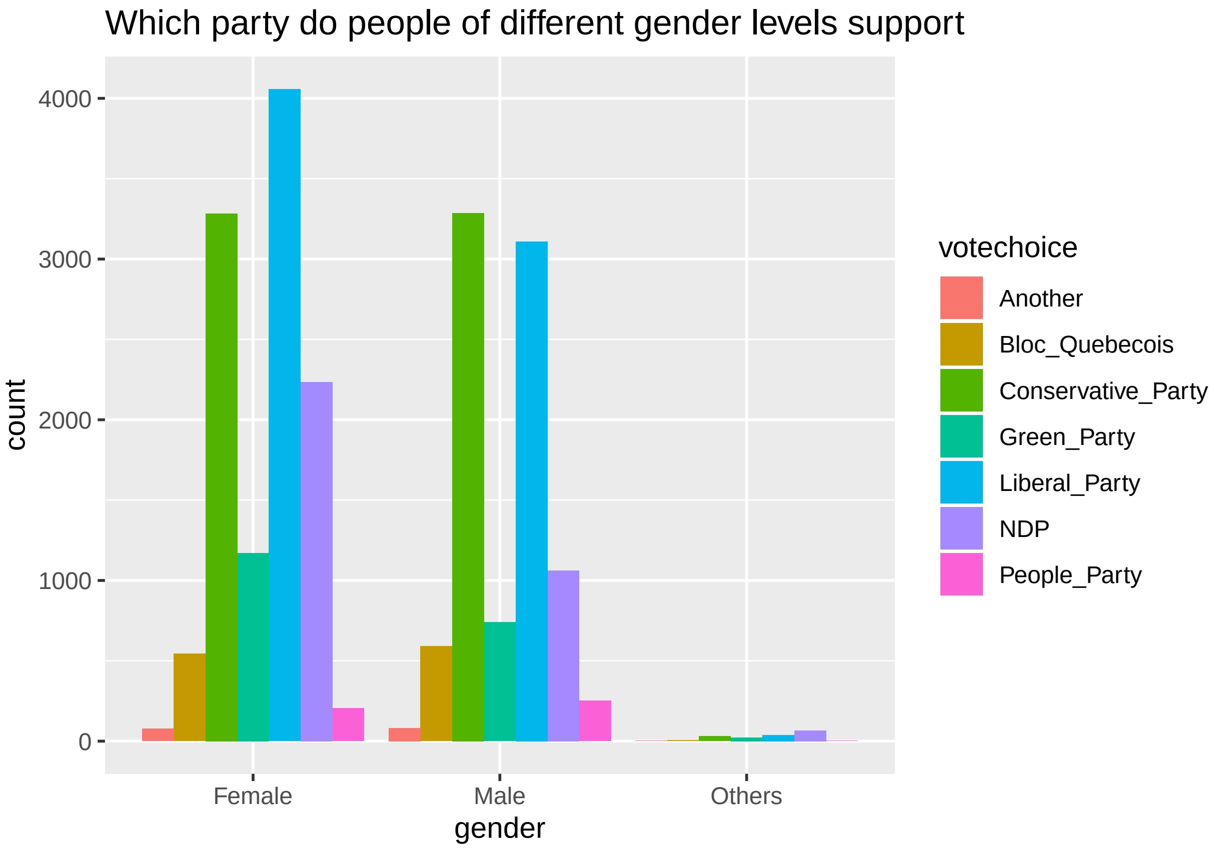
Table1

Liberal.Party	Conservative.Party	NDP	Green.Party	People.s.Party	Bloc.Quebecois	Others
33.23 %	32.83 %	14.97 %	10.16 %	2.31 %	5.14 %	1.36%

From the table, the Liberals won the general election, with the Conservative party trailing them by less than one percent. Besides, The NDP, The Green Party, the Bloc Quebecois, and the People's Party are in third, fourth, fifth, and sixth places.

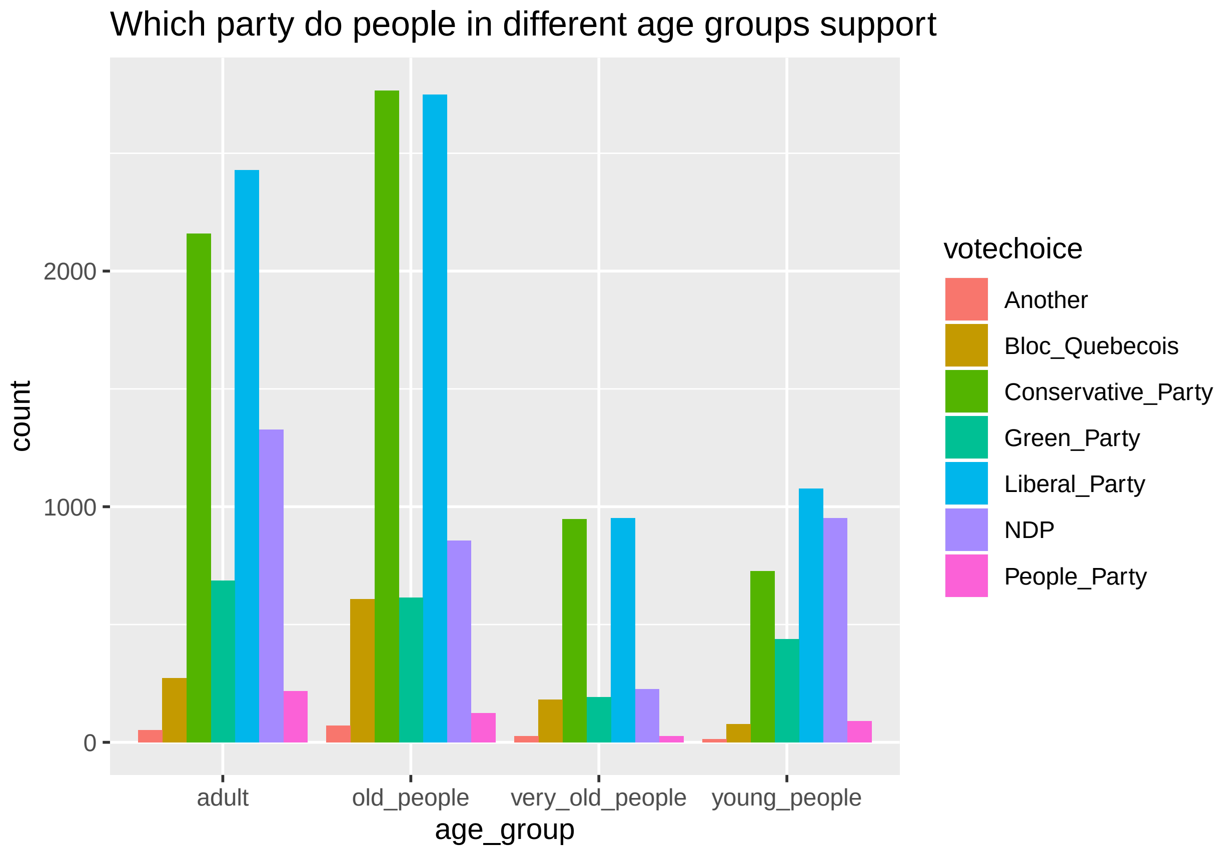
More plots:

***Figure1*:**



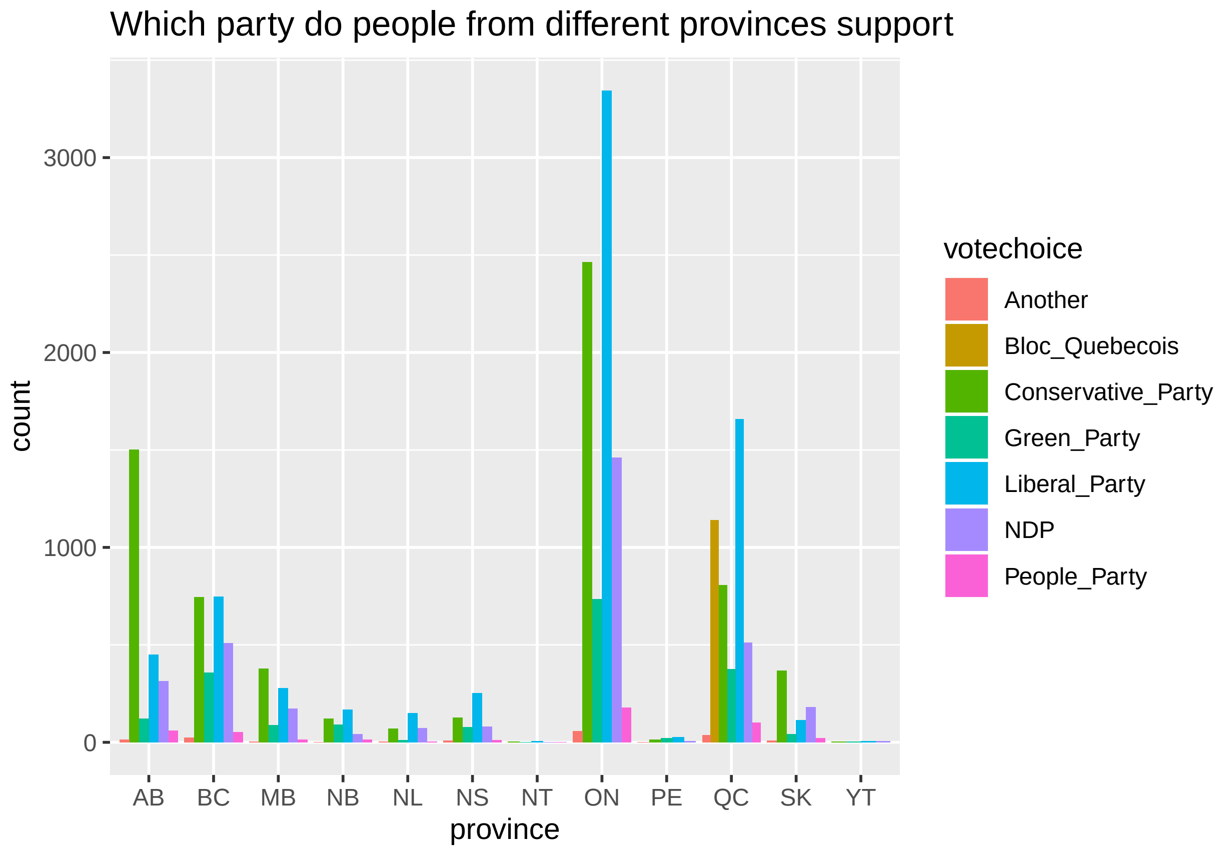
From the above plot, for female voters, most of them are more supportive of the Liberal Party followed by the Conservative Party. For male voters, support for the Liberal Party and the Conservative Party is evenly split, but overall, there is more support for the Conservative Party. The other parties are far behind the Liberal Party and the Conservative Party At last, for others, more people vote for the NDP.

***Figure2*:**



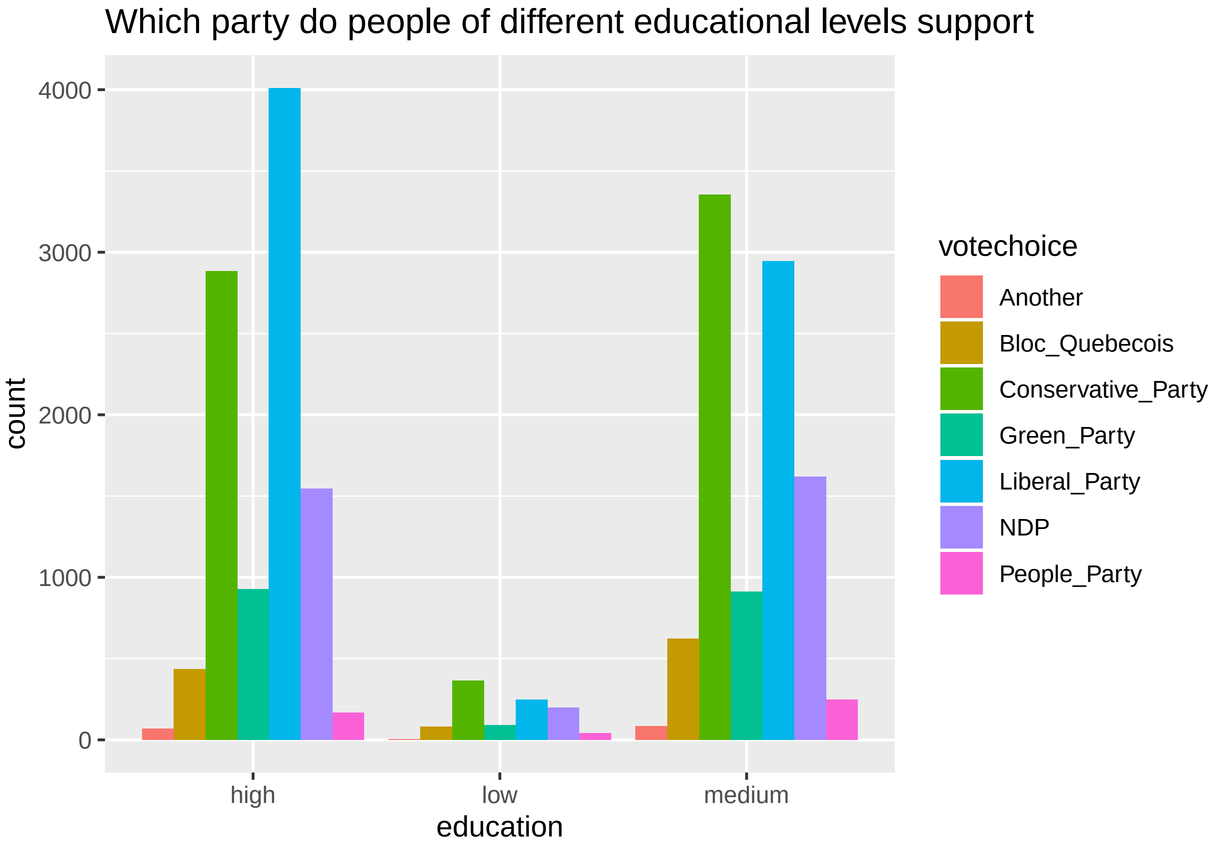
From the above plot, in age groups which are old people and very old people, the two parties with the highest approval ratings are the Liberal party and the Conservative Party, and the approval ratings of these two parties are almost the same. Moreover, for young people and adults, more people voted for the Liberal Party in the election. In particular, the second most popular party among the young is the NDP, while the second most popular party among the other three age groups is always the Conservative Party.

***Figure3*:**



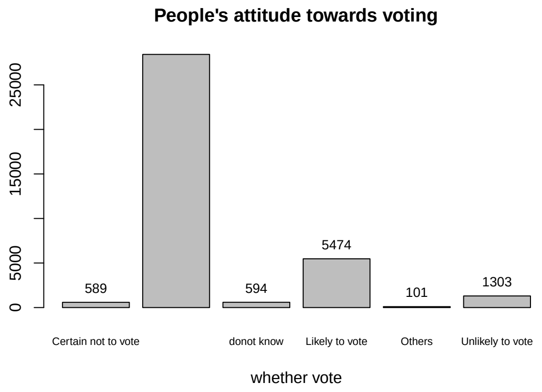
From the above plot, Ontario has the most voters, followed by Quebec. In these 12 provinces, except for Alberta, Manitoba and Saskatoon, where more people will support the Conservative Party, the Liberal Party has always been in the leading position.

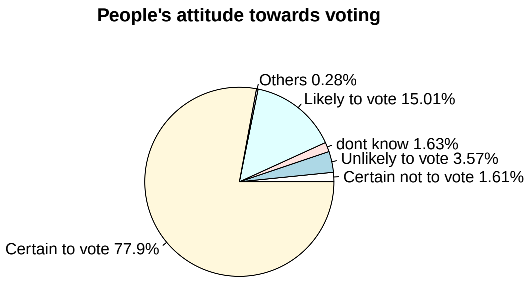
***Figure4*:**



From the above plot, the Liberal Party has a higher support rate among people with higher education, while the Conservative Party has a higher support rate among people with secondary education and lower education.

***Figure5* and *Figure6*:**





The above two plots show the attitude of all respondents in the sample data on whether to vote on election day. Others contain people who are not eligible to vote.

We can see from the graph that 77.9% of people are sure to vote in the general election. However, 1.61% of people will not vote on election day. And more than 20% of people are not sure whether they will vote. At the same time, among the more than 20% of people, more than a quarter tend not to vote.

Discussion

Summary

The main goal of this project is to identify how the 2019 Canadian Federal Election would have been different if "everyone" had voted and to figure out how important every person's vote for the election. We defined the meaning of "everyone" is that all people who are Canadian citizens and who on polling day are 18 years of age or older[2]. Therefore, the main task is to re-predict the result of the 2019 Canadian Federal Election based on the assumption that all people who have the right to vote for the election attend to vote for the popular vote.

At first, we select four independent variables: gender, education, provinces, and age. We value these four variables as independent variables to build the model and make the post-stratification because people with different age groups, different gender and levels of education live in different provinces could have different views of the current government's political policies and their political pursuits. Besides, we create six new binary variables in the sample data set which these variables' names are vote_literal, vote_conservative, vote_green, vote_ndp, vote_Bloc, and vote_people, corresponding to six federal political parties in Canada, based on the value of variable "votechoice". Then we make these six variables as dependent variables for six different models to predict the result of six main federal political parties in the 2019 Canadian Federal Election. Besides, since these six variables are binary, we decided to make six binary logistic regression models and use the post-stratification technique to apply six fitted models to census data.

Conclusions

Gender

As Figure 1, we can discover that The Conservative Party has slightly more male supporters than the Liberal Party, but among women, supporters of the Liberal Party account for more supporters. This may be related to the Liberal Party's strong support, and active implementation of the protection of women's rights just like Justin Trudeau said, "Canadians of all views are welcome within the liberal party of Canada, but when it comes to activiely supporting women's rights, our parties is committed to speaking with one voice"[7].

Age Group

As Figure 2, we can see that the vast majority of voters over 30 years old voted for the Liberal Party and the Conservative Party. In the three age groups (30-50, 50-70, 70 >), the number of supporters of the Liberal Party and the Conservative Party is basically the same. However, among young people, the top two approval ratings are the Liberal Party and NDP. The main reason why the NDP's approval rate in this age group exceeds that of the Conservative Party is that the NDP places great importance on young people in this election, just as the leader of NDP said that "he has made it a big part of his campaign to reach out to younger voters in ways that are relevant to them, and on platforms they regularly use"[8].

Provinces and Education

Firstly, similar to the last election, in Ontario and Quebec, the approval rate of the Liberal Party is far ahead of other parties. The Conservative Party also has a big lead in the three central provinces of Canada (AB, SK, MB). It is worth mentioning that the supporters of Bloc_Quebecois are mainly concentrated in Quebec, and the number of supporters in this province exceeds that of the Conservative Party.

Then for the variable education, from the Figure4 we can see that the Liberal Party is supported by the majority of people in the group with high education, while the Conservative Party has received a lot of support from the group with secondary education and below.

Final conclusion

Finally, the primary purpose of this analysis is to show and prove how the voter turnout in the federal election has an impact on the outcome of the election. At first, we see Figure 5 and Figure 6, which are shown people's different attitudes about whether to vote in the popular vote. Figure 5 shows the total number of people whose attitudes “Likely to vote” and “Unlikely to vote” ranks second and third in all variables in this bar plot, respectively. At the same time, from Figure 6, we can see that the number of people who meet these two variables, plus those who are determined not to vote in the general election, account for 20% of the total. In other words, the number of people who are not certain not to vote or who are certain not to vote is a quarter of the number of people who are certain to vote. Furthermore, from the results of the 2019 Canadian Federal Election, which predict by making the binary logistic models with post-stratification, we can know that the Liberal Party defeated the Conservative Party in the popular election by a slight advantage of 1%. Nevertheless, in the actual 2019 Canadian Federal Election popular vote, the Conservative Party leads the Liberal Party by 1% and takes the lead in the popular vote. Besides, in the re-forecast data, the Green Party rose from 3.45% in reality to 10.16% , and the People's Party also got 1% more votes than in reality. The probability of voting for The NDP and Bloc Quebecois remains the same as in fact.

In general, when we assume that the turnout is 100%, the predicted election results are different from the actual election results, such that the Liberal Party has changed from a loser in the popular vote to a winner. From the above analysis of the results of this project, we can clearly know how the voter turnout in the federal election has an impact on the outcome of the election. Although the results of the two general elections are the Liberal Party's victory, compared to the actual results, the Liberal Party will be more certain of winning the victory because, in the real election, the Liberal Party has a lower total number of votes than the Conservative Party, and finally won the final victory by winning the number of constituencies. In the actual 2019 Canadian Federal Election, “three-quarters (77%) of Canadians reported voting”[5], and this data is similar to the turnout rate, which is shown in Figure 6. Therefore, if the voter turnout rate can be increased, the result will be different from the actual result, as we predicted in the project. Some parties may get more votes in popular voting, thereby taking more seats in the House of Commons, and may even turn defeat into victory in the general election. Therefore, except “some people who cite for not voting in electoral politics because of philosophical, moral, and practical reasons”[6], the rest of citizenships who are 18 years old or older should have their voice in the elections, maybe your vote can improve the final result of the party you support. At last, the conclusion of this whole analysis is the voter turnout in the federal election will not have a significant impact on the election results, but it may change the final result.

Weakness & Next Steps

The Weakness of this project is the census dataset. We use the GSS dataset as the census dataset, yet the total number of rows of the census dataset was similar to the number of rows of Sample Data which we used to build the binary logistic regression model. This situation may lead to getting inaccurate prediction of the election results after using this census dataset to make the post-stratification. Furthermore, the GSS dataset we used is 2017 data. Although only two years apart, compared with the real-time data in 2019, the accuracy is still not high.

For the next step, we will look for a better dataset with massive data, high credibility, full coverage of information, as close as possible to the 2019 Canadian Federal Election, as the census data. Moreover, we also will check out more information about the 2019 Canadian Federal Election, including the policies guaranteed by each party during the election, election declarations, etc. And try to find as many variables that have a great impact on the results of the 2019 Canadian Federal Election so that our model can be more comprehensive and accurate.

Reference

[1] Welcome to the 2019 Canadian Election Study. (n.d.). Retrieved December 08, 2020, from http://www.ces-eec.ca/

[2] Branch, L. (2020, December 15). Consolidated federal laws of canada, Canada Elections Act. Retrieved December 22, 2020, from https://laws.justice.gc.ca/eng/acts/e-2.01/page-2.html

[3] GeeksForGeeks, A., AmiyaRanjanRout, & GeeksForGeeks, T. (2020, September 02). Advantages and Disadvantages of Logistic Regression. Retrieved December 22, 2020, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

[4] Multilevel regression with poststratification. (2020, December 18). Retrieved December 22, 2020, from https://en.wikipedia.org/wiki/Multilevel_regression_with_poststratification

[5] Government of Canada, S. (2020, February 26). Reasons for not voting in the federal election, October 21, 2019. Retrieved December 22, 2020, from https://www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm

[6] Voter turnout. (2020, December 02). Retrieved December 22, 2020, from https://en.wikipedia.org/wiki/Voter_turnout

[7] Stand with Justin Trudeau: Support a woman's right to choose: Liberal Party of Canada. (n.d.). Retrieved December 22, 2020, from https://liberal.ca/stand-with-justin-trudeau-support-a-womans-right-to-choose/

[8] Here are some of the big campaign promises from the major parties so far | CBC News. (2019, October 07). Retrieved December 22, 2020, from https://www.cbc.ca/news/politics/major-campaign-promises-federal-election-1.5311181

[9] Yihui Xie andJ.J. Allaire and Garrett Grolemund (2018). R Markdown:The Definitive Guide. Chapman and Hall/CRC. URL https://bookdown.org/yihui/rmarkdown.

[10] 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

[11] Statistics Canada. (2017). General Social Survey On Family (Cycle 31), 2017 - Canadian General Social Surveys (GSS).. [online] Available at: https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm

[12] Yihui Xie (2020). tinytex: Helper Functions to Install and Maintain 'TeX Live', and Compile 'LaTeX' Documents. R package version 0.21.

[13] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

[14] Statistics Canada. (2017). General Social Survey On Family (Cycle 31), 2017 - Canadian General Social Surveys (GSS).. [online] Available at: https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm

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

[16] Kazuki Yoshida and Alexander Bartel (2020). tableone: Create 'Table 1' to Describe Baseline Characteristics with or without Propensity Score Weights. R package version 0.12.0. https://CRAN.R-project.org/package=tableone

[17] Gergely Daróczi and Roman Tsegelskyi (2018). pander: An R 'Pandoc' Writer. R package version 0.6.3. https://CRAN.R-project.org/package=pander

[18] Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra