# Predicting the overall popular vote of next Canadian Federal election with 2019 Canadian Election Survey

STA304 - Assignment 2

GROUP 14: Mai Dang Khoi Nguyen, Joyce Nguyen, Yewon Kim, Tshego Kelesitse

December 1, 2022

## Introduction

In recent years, the use of statistical tools and analysis to forecast election outcomes has increasingly become popular. The Canadian parliamentary democracy has electoral districts or ridings which make up the House of Commons. The party with the most members elected into the seats forms the majority government.

Commonly, a random sample of people are asked who they intend to vote for and this is referred to as representative polls. However, this is costly and time-consuming. Additionally, Wang et al., (2015) argues that response rates have declined over the years, resulting in non-representative samples. The importance of the analysis is using over-represented or under-represented samples to predict popular votes for the federal election. Consequently, to quell the costly and time-consuming drawback of representative polling.

Our study is conducted using results from the 2019 Canadian Election online survey, which focused on gathering the attitudes and opinions of Canadians during and after the 2019 Federal election. The goal of our study is to develop predictive models and post-stratify our data to project the percentages of votes the liberal parties will receive in the upcoming 2025 Federal election. Our main research question is "Will the Liberal Party of Canada get a lower or higher percentage of popular vote in 2025 comparing to 2019?". We hypothesized that the Liberal Party of Canada will get a lower percentage of popular vote comparing to 2019, as the results from 2019 and 2021 Federal Election has been on a downward trend comparing to when Justin Trudeau won the election in 2015.

We begin by introducing the Canadian Election survey data and showcasing some exploratory data analysis. Followed by the process of building predictive models to predict the proportion of voters for the liberal and conservative parties respectively. We then conduct a post-stratification of the data. Post-stratification refers to the process of adjusting the estimates, essentially a weighted average of estimates from all possible combinations of attributes (for example age and gender, though there were more). Each combination is sometimes called a "cell" (Multilevel Regression with Post-stratification, 2022). Final results, limitations, and future perspectives conclude the project.

## Data

**Data Collection Process** The data for our analysis was gathered from the public 2019 Canadian Election Study (CES) on Canadians' political behavior and attitudes, aiming to measure their preferences on key political issues. This data for the election in 2019 was collected from two platforms, an online survey and the other one was over the phone, using a rolling cross-section throughout the federal election campaign and a follow-up poll after the election. By noticing the size of answers obtained from the online survey is much larger than the phone-based one (37,822 compared to 4021) and due to the fact that the online survey would give people more time to gauge question details, our analysis decided to go with the survey from the online platform to get a more objective view of voters' intention.

To be more specific about the survey data collection process, we accessed a package for CES surveys which already included all survey information, questions, and answers of the opinions about elections from 1965 to 2019. By knowing exactly which year and what type of survey we were looking for (year 2019 and online platform), we could choose the corresponding dataset code and draw out the desired data frame.

In this analysis, since we want to use the non-representative 2019 CES survey to forecast future election voting results, we also need a census data that represents the population. Keeping this in mind, we grabbed the General Social Survey (GSS) census data from the online databases maintained by Computing in the Humanities and Social Sciences (CHASS) at the University of Toronto. This dataset gives us rich demographic information with 20602 observations from 81 background check questions.

**Data Cleaning Process** At first glance, we see that the 2019 CES survey is extremely exhaustive with information ranging from each individual's demographics to their political knowledge, ideology, and satisfaction. The GSS census data also gives us an abundant amount of details about Canadians' backgrounds. While this is good news because we need their demographic information to predict the voting results, the data is quite all over the place with lots of irrelevant data to the interest of our analysis. So we select the same five variables from both datasets, namely voters' age, gender, province of locating, home language, and education level. After that, we notice both datasets encounter the same issue of containing too many missing values so we emulate all those elements that may challenge our analyzing process. Furthermore, depending on each variable selected, we cleaned and categorized them into more comprehensive subsets in both datasets simultaneously as follow:

- Age was originally in decimals so we round it to the closest integer. Also, since only Canadians by the age of 18 or older are qualified to vote, we excluded all observations with ages under 18. Then we categorized the rest into six age intervals, namely 18-24, 25-34, 35-44, 45-54, 55-64, and 65+, for the benefit of our analysis. Gender: Although the 2019 CES survey data is inclusive of all types of genders, the GSS census data only contains two options of "Male" and "Female" so only these two answers were kept.
- Province: Looking at the Appendix of the survey data, we notice that out of 10 provinces and 3 territories, only Alberta, British Columbia, Ontario, and Quebec have sufficient percentages of responses, and the proportion of the rest falls between 0.1 - 4.5%. So for this variable, we keep data for the five mentioned provinces the same but group the rest together into a category called "Some Others".
- Language: Taking into account the fact that the two official languages of Canada are English and French, we group related options into these two language subsets (e.g. "French and non-official language" in survey data fall into "French" and so on).
- Education level: To compactly capture the education level of each individual, we take the Bachelor's degree as the standard threshold and classify the given data into three categories, namely Below Bachelor's degree level, Achieved Bachelor's degree, and Above Bachelor's degree level.

Moreover, for the purpose of our data analysis, there are some differences in the variables of the two datasets that we need to take into account: * As we are using the CES survey data to build a model to forecast the number of votes for the Liberal party, we also include Canadians' responses to the question: "Which party do you think you will vote for?". Then we created a new variable with binary indicators which shows whether or not the person intended to vote for the Liberal party. * On the other hand, due to the nature of the poststratification method that we will implement in this report (details about the method will be highlighted in the Methods part), we will need to consider the proportion of all possible combinations of age (6 categories), gender (2 categories), province (6 categories), language (2 categories), and education level (3 categories) in in the population. Hence, for the GSS census data, a new variable was created which contains the size of each possible combination.

After a long procedure of cleaning and processing data in R programming language, the 2019 CES survey data has 26405 variables with 7 variables, and the GSS census data has 271 observations and 6 variables.

Here is a resource for grabbing the CES2019 data: <https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-canadian-election-study>

**Data Summary**/ In total, here are all the variables that we will be looking at in this report:

- `cps2019_age`: The age of qualified Canadian voters in the 2019 election, divided into six intervals, namely 18-24, 25-34, 35-44, 45-54, 55-64, and 65+.

- `cps2019_gender`: The gender of qualified Canadian voters in the 2019 election.
- `cps2019_province`: The resident province of qualified Canadian voters in the 2019 election.
- `cps2019_Q_Language`: The home language of qualified Canadian voters in the 2019 election.
- `cps2019_education`: The education level of qualified Canadian voters in the 2019 election.
- `cps2019_votechoice`: The party that qualified Canadian voters intended to vote for in the 2019 election.

We are done with parsing the data, now we can look at some important graphical visualizations and numerical summaries.

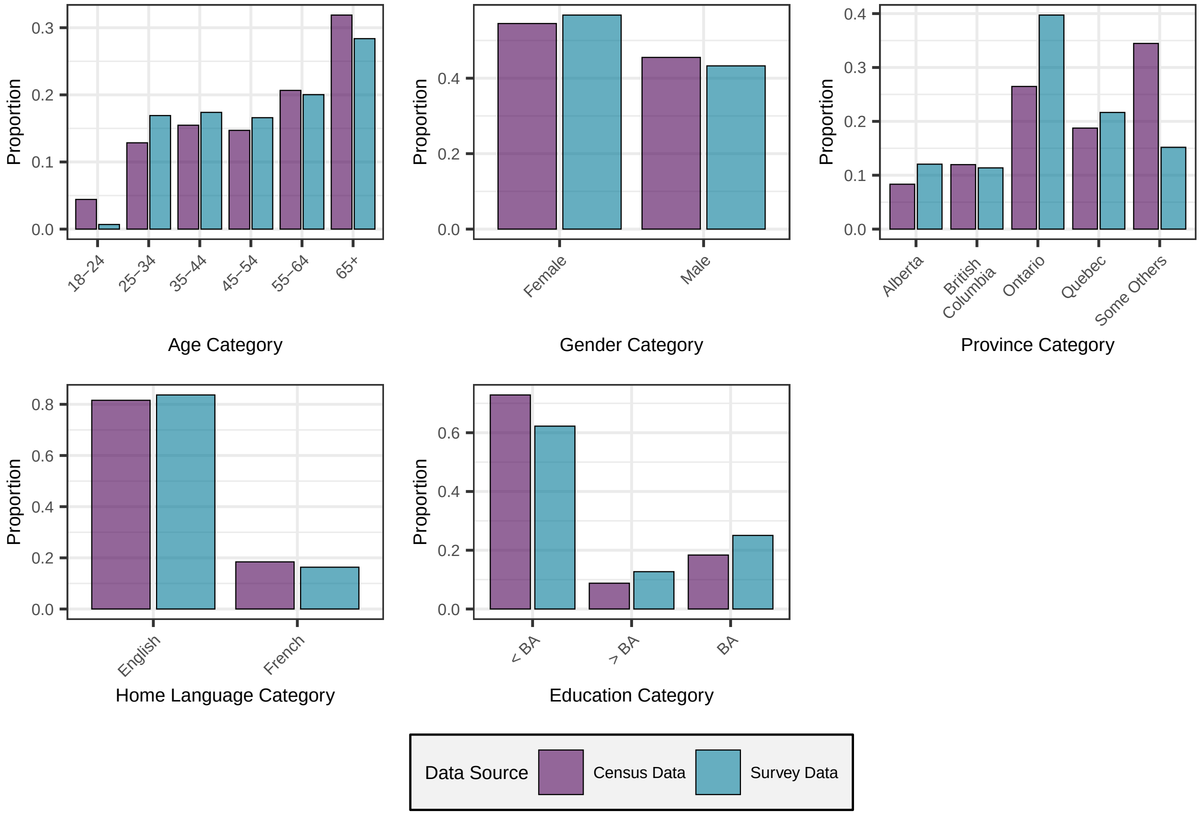


Fig. 1: A comparison of the demographic information in the 2019 CES survey data set and the GESS census dataset.

Fig. 1 compares the demographic composition of the CES survey participants and the GESS census dataset. The most striking difference is for the province category. As one might expect, people residing in Ontario comprise 40% of the CES survey dataset, compared to only approximately 25% in the GSS census data; and while the CES survey only has 15% of respondents from other provinces, people from those places took up 35% of the GSS census data. It has been known that the location will be one of the strong indicators for election voting, and we will have to implement rigorous methods to explore this association in Canada.

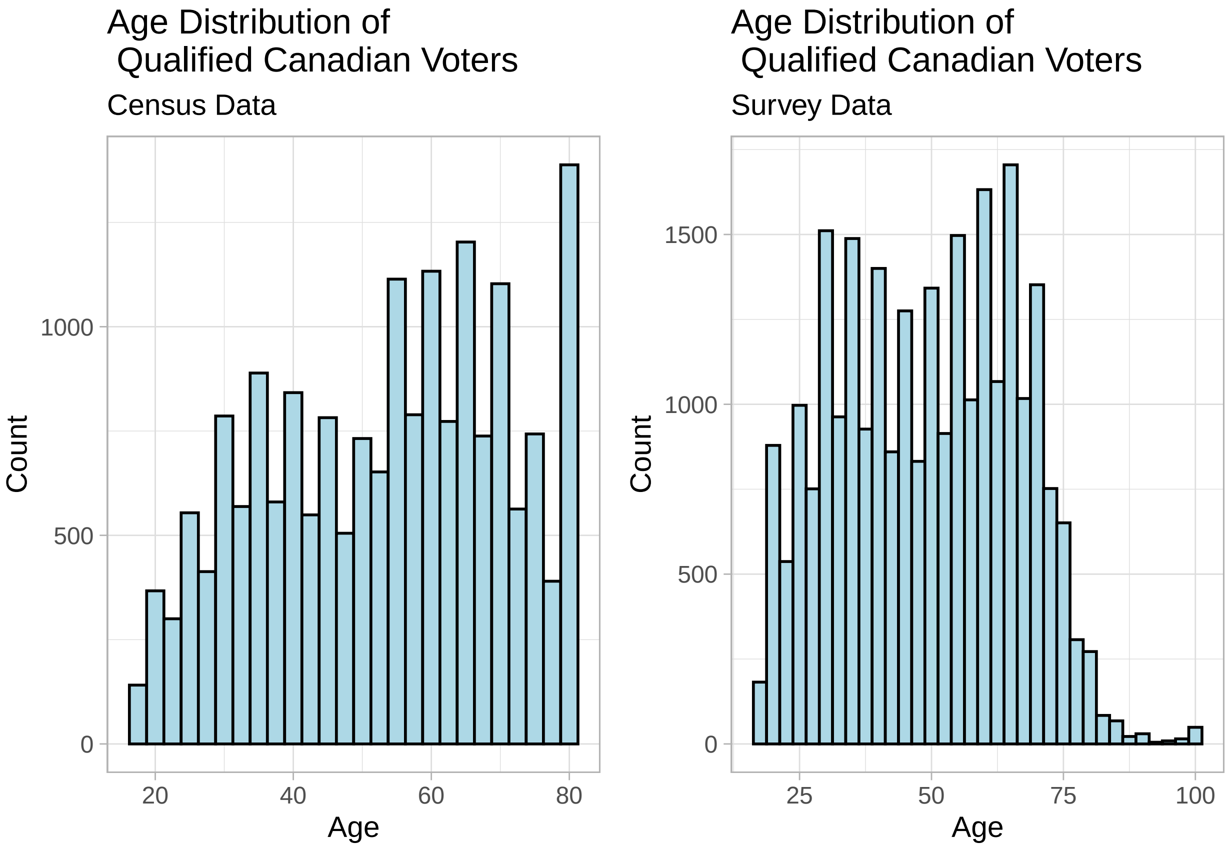


Table 1: Table 1: Age Summaries for Census and Survey Datasets

| Variable | Mean | S.D. |
|---|---|---|
| Census Data | 53.22 | 17.212 |
| Survey Data | 49.193 | 16.68 |

Other than age, other variables are characteristic so we want to take a more thorough look at the differences in voter's age between census and survey datasets. As the histograms in Fig. 2 show, voter's age in census data is quite left-skewed while in survey data it is slightly right-skewed. This might be due to the fact that the age range in survey data is slightly longer. Diving deeper into numerical values in Table 1, we notice that there is not much of a huge difference between the means and spread of age of both datasets. This confirms that the CES survey data set is quite representative of the population.

## Methods

To predict the results of the 2025 Election and whether the Liberal Party has majority vote, we first created a multiple logistic regression model from the Canadian Election Study survey, and used post-stratification onto the General Social Survey to extrapolate our data to the population.

### Model Specifics

Logistic regression shows the odd ratio of event happening with the predictor variables, and we are interested in whether the respondent will vote for Liberal or not and the probability of Liberal being elected. Thus the outcome should represent the probability, and logistic regression is the correct model that gives the outcome we want. As there are many factors considered in election, we use more than one variable thus it will be multiple logistic regression. Hence, we will use multiple logistic regression model as we are interested in the probability of Liberal getting voted on.

To build our model we dichotomize the votechoice variable, 1 indicates the Liberal party vote and 0 indicates a non-Liberal vote. We then group the provinces according to the four largest by area in Canada (Provinces and Territories of Canada, 2022), namely Alberta, British Columbia, Ontario, and Quebec. Respondents outside these provinces are placed into the Other category. The provinces are assigned 1, 2, 3, 4, and 5 respectively. Similarly, the highest level of education completed is grouped into three categories, below a Bachelor's Degree, Bachelor's Degree, and above a Bachelor's Degree. Language is changed to a binary variable, 1 for English and 2 for non-English speakers. We use sex instead of gender, hence we assign 1 to male respondent and 2 to female respondent. We fit a multiple logistic regression model with the following predictors, age, sex, province, education, and language. Our outcome variable is voter choice, liberal or non-liberal vote.

The following is the multiple logistic regression model we are using:

$$y = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{province} + \beta_4 x_{education} + \beta_5 x_{language}$$

$y$ represents whether the respondent votes for the Liberal party or not, in which 1 represents he votes for the Liberal and 0 is for non-Liberal. $\beta_0$ represents the intercept of the model, when all of the predictors have value of 0. $\beta_1$ represents the coefficient of the age predictor, which shows the expected change in log probability when there is a change in the predictor. $\beta_2$, $\beta_3$, $\beta_4$, and $\beta_5$ are the coefficient of sex, province, education, and language, respectively. They all represent the association with the predictor variables and the outcome that is the log probability.

We use Akaike information criterion(AIC) in order to check if the model we chose is good. Since the model has to have at least two variables and we have five selected variables, we run AIC with the models with three predictors to see which one is the most suitable. The smallest AIC represent the model fits well. Based on the result, there were two model with smallest AIC, one with four predictors except sex variable and one with five predictors that includes all of the predictors we selected. We choose the one with five predictors since the AIC value are very close (36988.47 and 36989.9), and we are interested in whether there is a big difference in between two sex groups.

## Post-Stratification

Next, we use the method of post-stratification to apply our model and predict the voters' choice based on their characteristics. Post-stratification is a process of dividing our demographic into cells, where each cell presents a different combination of socioeconomic attributes, such as their age, sex, language, education or current provinces (Wang, 2015). For each cell, we then estimate each cell's response (ex: probability of voting Liberal) with the model created from our survey data. Finally, we multiply the estimated responses of each cells with their proportion of the targetted total population.

The main benefit of post-stratification is its ability to correct for non-probability sampling, allowing survey data to speak on behalf of the bigger population. As we can see from the previous section, the population of Ontario is underrepresented in the CES 2019 response when compared to the actual population. Post-stratification allows us to adjust the responses that are either oversampled or undersampled, therefore creates a better extrapolation of the population actual response. The mathematical formula of post-stratification can be defined as:

$$\hat{y}^{PS} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

In the above formula, $\hat{y}_j$ is the estimate response for a specific cell, and $N_j$ is the population size of the jth cell. The generated cells were a combination of the following attributes:

- `cps19_age`: The individual's age at time of census. We grouped them into 6 categories, with the "18-24" for the average age group still in post-secondary, the age groups of "25-34", "35-44", "45-54", "55-64" are those in the working ages, and "65+" for the ones that have retired. Each age group has a different socioeconomic situation and upbringing, therefore might be attracted to certain parties separately based on different factors.

- `cps19_gender`: The individual's biological sex. There are 2 categories, "Men" and "Women".
- `cps19_province`: The individual's located province at the time of census. There are 5 categories: "Ontario", "Alberta", "British Columbia", "Quebec", and "Others". We grouped the data for other provinces as there isn't sufficient population in the sample data for the model. Choosing provinces as an attribute is important, as there will be favorites for certain parties in a particular provinces during election, which was influenced from past elections.
- `cps19_education`: The individual's highest level of education at the time of survey. There are 3 categories: "<Bachelor Degree", "Bachelor Degree" and ">Bachelor Degree". We want to see whether there are any differences regarding the different level of education, and whether policies from certain parties resonate more with each groups.
- `cps19_Q_Language`: The individual's language spoken at home. We grouped them into 2 categories, "English" and "French". Partitioning by language will help us highlight any bias with family upbringing, and whether there is a lower interest in Liberal/Conservative due to having another French-speaking party (Bloc Quebecquois).

We believed that the above attributes can represent well the socioeconomic status of specific individuals, with those of similar characteristics to have similar responses. We didn't include any politically related variables, as such data in the General Social Survey wasn't readily available. We also filtered out any data points which has an `N/A` in any of the columns and excluded individuals under 18, as they are not eligible to vote. The division of the population into cells with the above characteristics would give us a total of 360 cells. In reality, we were only able to extract an estimate for 271 cells, as not all cells were covered from our General Social Survey data set.

The post-stratification method assumed that the sample for each cell is drawn at random from a larger population, hence will create an unbiased estimate (Wang, 2015). Both the General Social Survey and the 2019 Canadian Election Study used a Random Digit Dialing (RDD) method to collect cross-sectional data from random sample of Canadians, therefore allow us to meet the required assumption(Canadian Election Study, 2019). It is also optimal to make sure that the cells weren't broken down into very small levels, as our model will run the risks of over-fitting with our survey data, and hence perform poorly in our population estimation (Wang, 2015). As discussed above, we grouped certain categories which had a low sample counts but still maintain similar characteristics, therefore preventing the data to be over-fitted.

Our variable of interest for post-stratification is the proportion of population which would vote for a certain political party. We created the model for Liberal Party of Canada as they are the current Government.

With the discussed methods, variables and assumptions discussed above, post-stratification is an appropriate methodology to predict the result of the next Canadian Federal Election. This is backed by the similar researches which utilized the same method, including those which seek for public opinions and voter turnouts by demographic groups.

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

## Results

Table 2: Predicted Liberal Party's vote turnout for each province

| Province | Predicted Vote Share |
|---:|---:|
| 1 | 0.258 |
| 2 | 0.300 |
| 3 | 0.335 |
| 4 | 0.311 |
| 5 | 0.400 |

Out of the four major provinces, Ontario had the highest vote share for the Liberal party which was 33.499%. Quebec and British Columbia showed approximately similar results 31.098% and 30.01% of vote turnout for Liberal respectively, whereas Alberta had the lowest among the selected provinces that was only 25.765%. Although majority of population live in those four provinces, it was the combined provinces that had the highest predicted result of 40.013%. The result seemed reasonable as Ontario is the most populated province in Canada(Statistics Canada, 2022), thus there is a higher probability compared to the other provinces that Liberal supporters live in Ontario. Moreover, the Liberal party also won majority in Ontario and Quebec in the last election, whereas the Conservative party winning both BC and Alberta, which explain the predicted values here.

Table 3: Predicted Liberal Party's vote turnout for each age group

| Age | Predicted Vote Share |
|---|---:|
| 18-24 | 0.297 |
| 25-34 | 0.309 |
| 35-44 | 0.339 |
| 45-54 | 0.333 |
| 55-64 | 0.353 |
| 65+ | 0.361 |

For the age category, the age group that had the highest vote share of 36.139% was people who are older than 64. Meanwhile, it was the age group from 18 to 24 that had the lowest vote share which was 29.706%. The age group in between showed relatively uniform vote share, from 30.88% to 35.283%. While this was unexpected as we assumed that the younger population will vote for the Liberal party based on the article we found that showed there was a tendency that liberals are more likely to convert to conservatives compared to vice versa, across the lifespan(Peterson et al., 2020). However, it was reasonable as based on the past election data, the older population who were over 64 participated in elections more and voted for Liberals than the younger population who were younger than 65 (Elections Canada). It matched with the past data which showed that the older population had a vote share of over 60% whereas only 54% of the younger population who were from 18 to 24 voted for the Liberal party.

Table 4: Predicted Liberal Party's vote turnout for each education level

| Education | Predicted Vote Share |
|---:|---:|
| 1 | 0.320 |
| 2 | 0.378 |
| 3 | 0.454 |

People who had more than Bachelor's degree showed the highest vote share on the Liberal party that was 45.399% whereas it was only 31.983% for people who did not get Bachelor's degree. This was a reasonable result based on the research conducted by the Pew Research Center showed that the more education one received, the more liberal one became (2016).

Table 5: Predicted Liberal Party's vote turnout for each sex

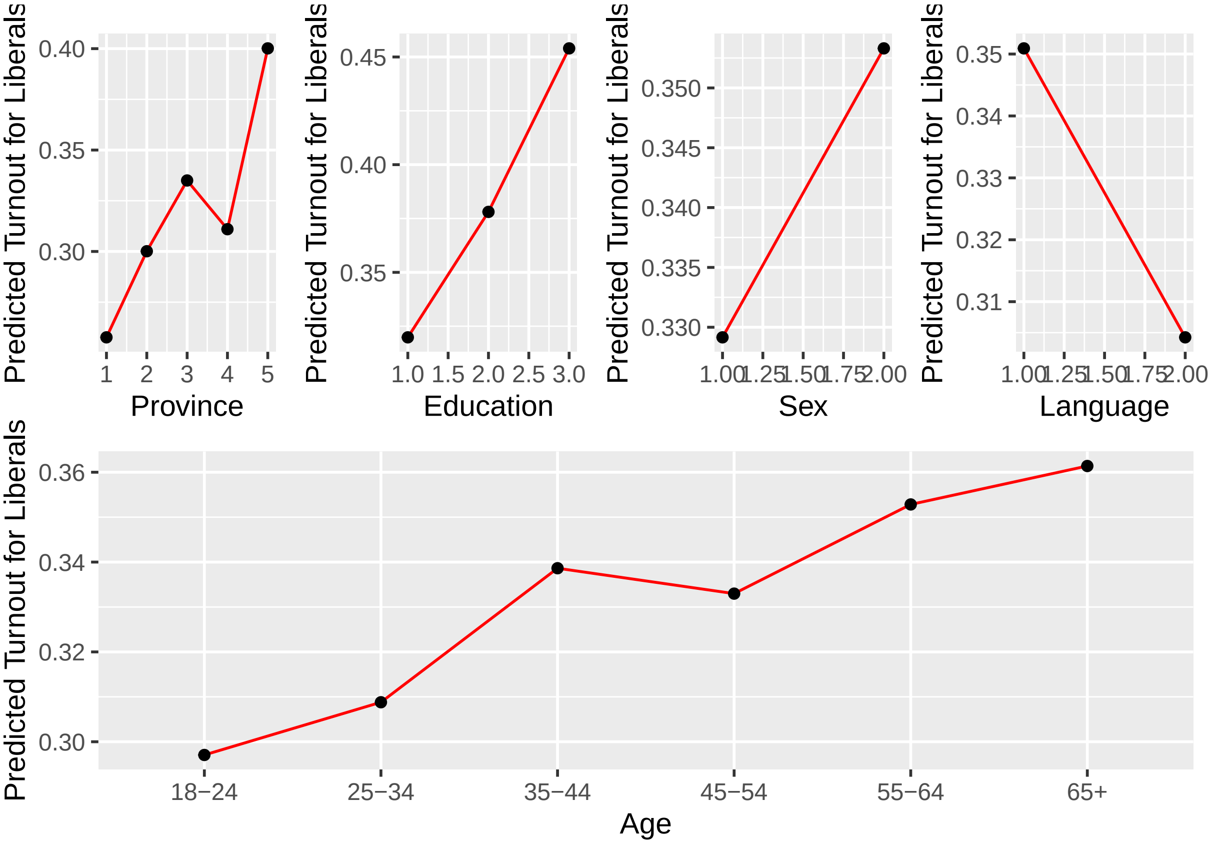
| Sex | Predicted Vote Share |
|---:|---:|
| 1 | 0.329 |
| 2 | 0.353 |

It was almost similar for both sex groups, but the female group showed a slightly higher vote share which was 35.33% while it was 32.916% for the male group. This was also supported by the actual data on past elections, which showed almost 47% of females voted for the Liberal party while only 35% of males voted for the Liberal(EKOS Research Associates, 2020).

Table 6: Predicted Liberal Party's vote turnout for each language

| Language | Predicted Vote Share |
|---:|---:|
| 1 | 0.351 |
| 2 | 0.304 |

Similar to the result for bot sex groups, there was no big difference in both language groups. We assumed that French speakers would not vote for the Liberal party as Bloc Québécois is the party made in Quebec, where many French speakers reside. Thus, it was reasonable that the group of English speakers had 35.092% of vote share that was approximately 5% higher than the vote share for non-English speakers, 30.424%.



The graphs show the predicted voter turnout for the Liberal party. Each dot represents each group's predicted turnout for the Liberal party. The graphs for sex, age, and language showed less than a 10 percentage points difference between each group, but the graphs for province and education showed more than a 10 percentage points difference between the groups. It is noticeable that no single group within the predictors has more than a 50% of turnout rate, and the highest turnout rate is 45.399% which is the group of people with more than a Bachelor's degree.

## Conclusions

Our paper looks to predict the result of the next Canadian Federal election in 2025, with our main focus on predicting the percentage of popular vote in which the current government, the Liberal Party of Canada, will get from the population. Our methodology included a two-parts process. First, we fitted a multiple logistic regression model from the online Canadian Election Survey in 2019 to predict whether an individual with specific characteristics would vote for the Liberal Party. Next, we used the post-stratification technique to apply our model into the General Social Survey, therefore predicting each cells' in the population and whether they will vote Liberal.

With our research question "Will the Liberal Party of Canada get a lower or higher percentage of popular vote in 2025 comparing to 2019?", we hypothesized that the Liberal Party of Canada will get a lower percentage of popular vote comparing to 2019, as the results from 2019 and 2021 Federal Election has been on a downward trend comparing to when Justin Trudeau won the election in 2015.

Our result predicted that in 2025, of the 4 most populated provinces, Ontario will have the highest percentage of votes for Liberal at 33.499%. It was surprising to see that voters in the older age groups had a higher percentage vote for Liberal Party comparing to younger age groups, which is reversed from other studies. Voters which had a higher education degrees (Masters or above) also has a much higher percentage of those who voted Liberals, which is at 45.399%. Female voters are slightly more likely to vote Liberals comparing to male voters. Similarly, English-speaking voters are more likely to vote Liberals at 35.092% compared to French speaking voters, which is reasonable as the Bloc Quebecois party is popular among French speakers.

With the predicted percentage of voters voting for Liberals, the party is likely to be at risk of losing the election in 2025, given that in 2021, the Liberal Party of Canada lost the majority vote to the Conservative Party of Canada. An evaluation of predictions for other parties would be needed for us to come up with a conclusion, but given the results we found today, our hypothesis that the Liberals Party of Canada would decrease in the percentage of popular vote is accepted.

A big weakness of our study is the size of the census data. Given that the data we use for our census, the General Social Survey, is indeed even smaller than our survey data, it would not allow us to properly predict the population's vote. The above fact risks our proportion to not be representative of the Canadian's actual population, hence will give us a skewed prediction. Our paper also only focused on the Liberal party, and lack the numbers in the other competing parties in the Canadian election. This is a shortcoming, as it will also not give us a full picture of the potential result, therefore not enough for us to predict the actual result of the election as accurately as possible.

For future works, it is advised that we use the 2021 Canada Census, which will give us data that is much more complete comparing to the General Social Survey, hence allowing us to see a better proportion of our population votes. It is also ideal for us to build models for other parties, therefore extracting more precised insights in which party will be predicted to win the Election in 2025. A model using multi-level regression should also be tested in future works, as it is commonly used in similar works globally.

The Federal Election is an important day for Canada, and it is difficult for us to know exactly what will happen before the Election day. However, the post-stratification method is a powerful tool will allow us to extrapolate such insights about the population from non-representative surveys, and hence help policy researchers and makers make more accurate predictions in a cheaper, more efficient method.

## Bibliography

1. Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). *Forecasting elections with non-representative polls.* International Journal of Forecasting, 31(3), 980–991. https://doi.org/10.1016/j.ijforecast.2014.06.001

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

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

4. Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980–991. https://doi.org/10.1016/j.ijforecast.2014.06.001. (Last Accessed: December 1, 2021)
5. Provinces and territories of Canada. (2022, November 26). Wikipedia. https://en.wikipedia.org/wiki/Provinces_and_territories_of_Canada#:~:text=Its%20four%20largest%20provinces%20by. (Last Accessed: December 1, 2021)
6. Tables. (n.d.). Rmarkdown.rstudio.com. https://rmarkdown.rstudio.com/lesson-7.html. (Last Accessed: December 1, 2021)
7. ggplot2 line plot : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA. (n.d.). Www.sthda.com. http://www.sthda.com/english/wiki/ggplot2-line-plot-quick-start-guide-r-software-and-data-visualization. (Last Accessed: December 1, 2021)
8. Statistics Canada. (2019). Population Estimates, Quarterly. Statcan.gc.ca. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901. (Last Accessed: December 1, 2021)
9. Canada, E. (n.d.). Voter Turnout by Sex and Age. Www.elections.ca. https://www.elections.ca/content.aspx?section=res&dir=rec/eval/pes2019/vtsa&document=index&lang=e. (Last Accessed: December 1, 2021)
10. Fournier, P. J. (2021, April 25). 338Canada: The Liberals are winning over older—normally Conservative—voters - Macleans.ca. Macleans.ca. https://www.macleans.ca/politics/ottawa/338canada-the-liberals-are-winning-over-older-normally-conservative-voters/. (Last Accessed: December 1, 2021)
11. Pew Research Center. (2016, April 26). Ideological Gap Widens Between More, Less Educated Adults. Pew Research Center - U.S. Politics & Policy. https://www.pewresearch.org/politics/2016/04/26/a-wider-ideological-gap-between-more-and-less-educated-adults/. (Last Accessed: December 1, 2021)
12. Update on the Political Landscape and the Issues of Race, Policing, and the Three Ms in the Canada-China Affair. (2020, June 26). EKOS Politics. https://www.ekospolitics.com/index.php/2020/06/update-on-the-political-landscape-and-the-issues-of-race-policing-and-the-three-ms-in-the-canada-china-affair/. (Last Accessed: December 1, 2021)
13. Statistics Canada (2017) *General Social Survey: An Overview*. https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm
14. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John. (2019) *2019 Canadian Election Study - Online Survey* https://doi.org/10.7910/DVN/DUS88V