A Statistical Prediction of the Outcome of Future Canadian Federal Elections

STA304 - Assignment 3

GROUP 137: Ali Kaazempur-Mofrad, Zeynep Karadeniz, Dhanraj Patel

November 5, 2021

Introduction

With a population of over 30 million people, Canada is one of the largest countries in the world with a wide array of diverse and constantly evolving political beliefs among the population. As a multiple system parliamentary democracy Canadians have many different choices of political parties to choose from during elections with the three largest parties being the Liberal Party, the Conservative Party and the New Democratic Party (NDP). As the attitudes that Canadians have about these different political parties is constantly changing, the goal of this study is to use publicly sampled data and apply stratification with census data to predict the winner of the next Canadian federal election. It is important to note that the sampled data contains samples from a wide array of diverse Canadians from a survey taken in 2019.

It is important to analyze sampled data and to use that information to predict the outcome of the next federal election in order to better understand the the population's political beliefs and the political party they support. The data allows us to also analyze the political opinions of Canadians in different regions to understand how politically diverse or how concentrated political opinions are between different provinces. In addition to regional and national relevance, this study also has global relevance as the data allows us to compare our political beliefs with other countries around the world. The main goal of this study is to better understand the political beliefs of Canadians and how they feel about different Canadian political parties.

Terminology

Before getting to the data and methods section, it is first important to understand some of the terminology within the survey.

In terms of Canadian politics, it is important to know that Canada uses a multi-party parliamentary democracy. At the federal level, while Canada has many parties, the three main parties are the Liberal Party, the Conservative Party and the New Democratic Party (NDP). These three parties will be the focus of this study.

Also, for this study it is important to understand that both the survey sampling data and census data that will be used. The sampling data in this study was collected by the Canadian Election Study (CES) group. They specialize in collecting both the the demographics and political opinions of Canadians. In this study, we will be using the survey conducted in 2019 for the sample data in this study. For the census data, we will be using the General Social Survey (GSS) which is an in depth survey conducted by Statistics Canada. By using the sample data we will be able investigate the political opinions of Canadians and by extrapolating the data using post stratification using the census data, we will be able to predict the winner of the next federal election.

Central Question

By using the data from the CES survey and conducting post stratification with the census data, this study aims to answer the question of who will be the winner of the next Canadian federal election. Since the CES

data took place in 2019, we hypothesize that the results of the study will mirror the outcome of the federal election that took place in 2019. In this election the Liberal Party won the most amount of seats, securing the position of Prime Minister, but the Conservative Party won the popular vote. Lastly, we hypothesis that the NDP party will place third place in the results.

Data

Introducing the Data

In order to predict the outcome of the next federal election we will be utilizing data collected by the Canadian Election Study (CES). The Canadian Election Study is a group that aims to collect information on what Canadians think of different Canadian political parties and the demographics of those who respond to their survey. They primarily collect data using telephone surveys in addition to online surveys [1]. Using their 2019 data, we will be able to understand how a diverse set of Canadians view the three largest political parties in Canada (the Liberal Party, the Conservative Party and the the New Democratic Party) and if they would vote for them. An issue arises in the fact that while the survey data is extensive, it is not large enough to extrapolate to all of Canada. In order to amend this issue, we will be using post stratification to extrapolate our data by utilizing the General Social Survey (GSS) conducted by Statistics Canada. The GSS is a large scale survey conducted through phone surveys and internet surveys that collects information on the demographics of many Canadians [2]. By using the CES data in conjunction with the GSS, we will be able to predict the outcome of the next federal election.

Cleaning of the Data

After retrieving the data, it is important to clean the data so that the results generated by the models and summaries are an accurate reflection of the data. In order to clean the data, all columns from the dataset that were not used in the data summaries and models were removed from the CES dataset. After removing the columns, the only columns that remained were information regarding the survey participants, including their age group, gender, family income, highest level of attained education, and whether they would vote for the Liberal, Conservative, NDP parties. In addition, within the CES and GSS surveys, survey participants had the option to skip/refuse to answer any given question. To prevent missing data from impacting the results, any row that contained missing data was removed. Moreover, in the CES survey we modified the age group intervals of the survey participants to match the age intervals collected in the GSS survey, which is necessary for us to implement a post-stratification model in our methodology.

Important Variables

The following are the important variables from the 2019 CES survey

- age - The exact age of the CES survey participant
- pretax_income - The exact reported income (pre-tax) of the CES survey participant
- age_group - The age grouping interval of the CES survey participant
- gender - The gender of the CES survey participant
- Income_family - The family income of the CES survey participant
- education - The highest completed level of education of the CES survey participant
- vote_liberal - Whether the CES survey participant indicated they would vote for the Liberal Party in the next election
- vote_conservative - If the CES survey participant indicated they would vote for the Conservative Party in the next election
- vote_NDP - Whether the CES survey participant indicated they would vote for the NDP Party in the next election
- vote - The party that the CES survey participant reported that they will vote for in the next election

The following are the important variables from the GSS survey

- age_group - The age grouping interval of the GSS survey participants
- gender - The gender of the GSS survey participants
- Income_family - The family income of the GSS survey participants
- education - The highest completed level of education of the GSS survey participants
- province - The province in which the GSS survey participants resides

Numerical Summaries

Before moving onto methods, it is important to understand the data using numerical summaries. It is particularly important to investigate the spread and location of the data to better understand the values.

Comparing income and gender of the 2019 CES survey participants

Table 1: The trimmed mean is trimmed by 10 percent

gender	Count	Min	Q1	Median	Q3	Max	Mean	Trimmed mean	Var	SD
Female	1266	0	40000	75000	122000	2000000	94497.63	94497.63	9570854145	97830.74
Male	1748	0	50000	90000	150000	2120000	113141.30	113141.30	16408681224	128096.37

When analyzing survey participants and their leanings towards a political party, it is important to consider the survey participant income levels as income often plays a heavy factor in an individuals political beliefs. In the above table (see Table 1), we see the numerical summaries of the 2019 CES survey participants income levels for each gender. It is evident that the 2019 CES survey contains more males than females with the data containing responses from 1748 males and 1266 females. In addition, pertaining to their income (pre-tax), we see that from the median, mean, and trimmed mean values that males in the survey have a higher income, on average, compared to females. In addition, the average income for participants of both genders is higher than national average income of $52,600 as of Jan 2019 [3]. Moreover, we see that on the low end, the survey contained individuals, both male and female, who self reported no income. On the high end, the data contains individuals, both male and female, who earn an annual income of about 2 million CAD. This causes a high level of variance and standard deviation in the data.

Comparing gender and age of the 2019 CES survey participants

Table 2: The trimmed mean is trimmed by 10 percent

gender	Count	Min	Q1	Median	Q3	Max	IQR	Mean	Trimmed mean	Var	SD	Range
Female	1266	18	38	50	63	95	25	50.55	50.28	252	15.88	77
Male	1748	18	37	50	63	95	26	50.14	50.08	261	16.17	77

From the above table (see Table 2), we are able to see the numerical summaries of the 2019 CES survey participants age for each gender. From the table, we see that for both male and female survey participants, age ranges from 18 to 95. In addition, we see that both genders have nearly identical values. Notably, the mean and trimmed mean ages for males is slightly lower compared to the females, causing the males to have a slightly higher variance and standard deviation. Despite this, it seems that males and females have a nearly identical distribution of ages in the survey.

Graphical Summaries

In addition to numerical summaries, it is helpful to visualize the distribution of the variables in the data set.

Comparing the different parties that the 2019 CES survey participants reported to vote for in the next election

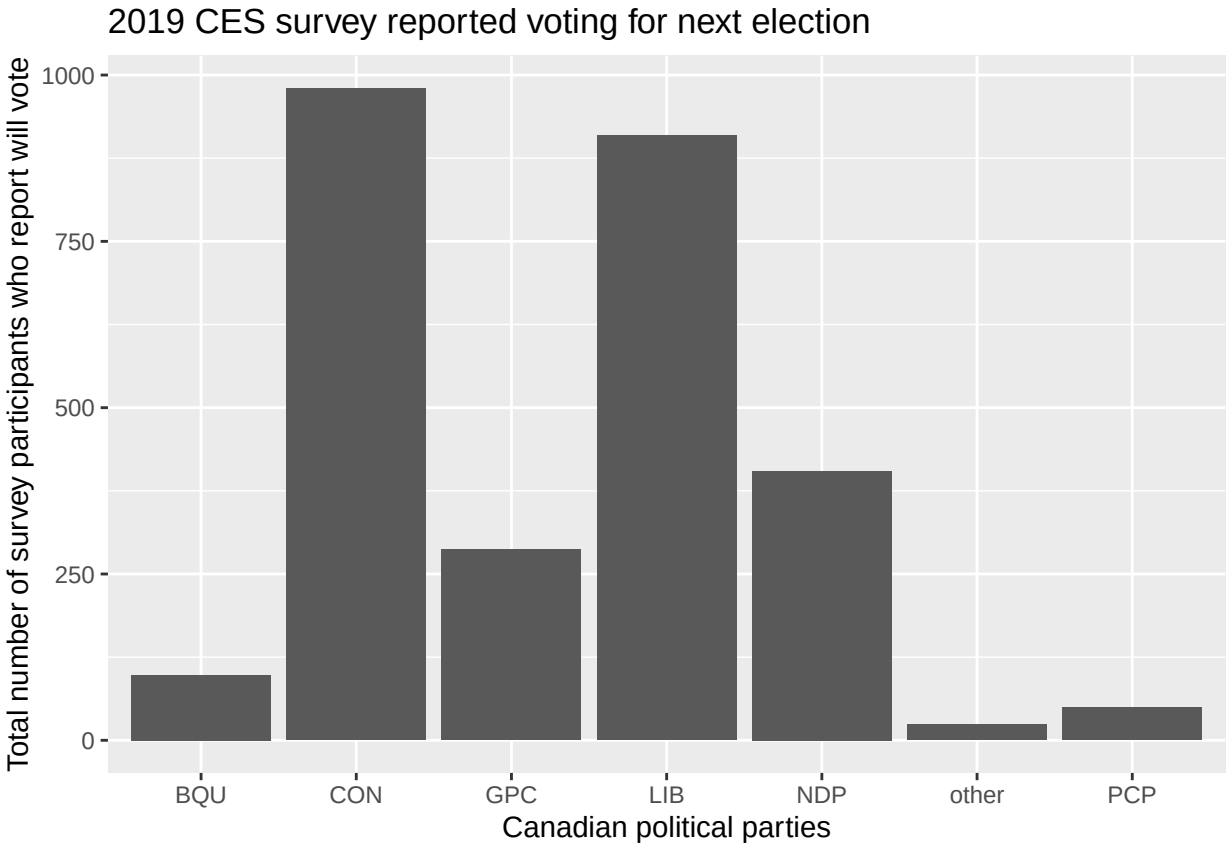

***Figure 1**: Total number of 2019 CES survey participants who plan to vote for each political party in the next Canadian federal election.*

In the above barplot (see Figure 1), we see the total number of 2019 CES survey participants and which party they intend to vote for in the next Canadian election. From the graph, we are able to see that most people in the survey reported that in the next election they would vote for the Conservative Party, followed by the Liberal Party and the New Democratic Party. The number of people who state they would vote for the other parties is noticeably lower. It is for this reason that in this study we will be focusing on the Liberal Party, Conservative Party and the New Democratic Party and the predicted voting turnout.

Comparing residing province and highest achieved education level of GSS survey participants (separated by gender)

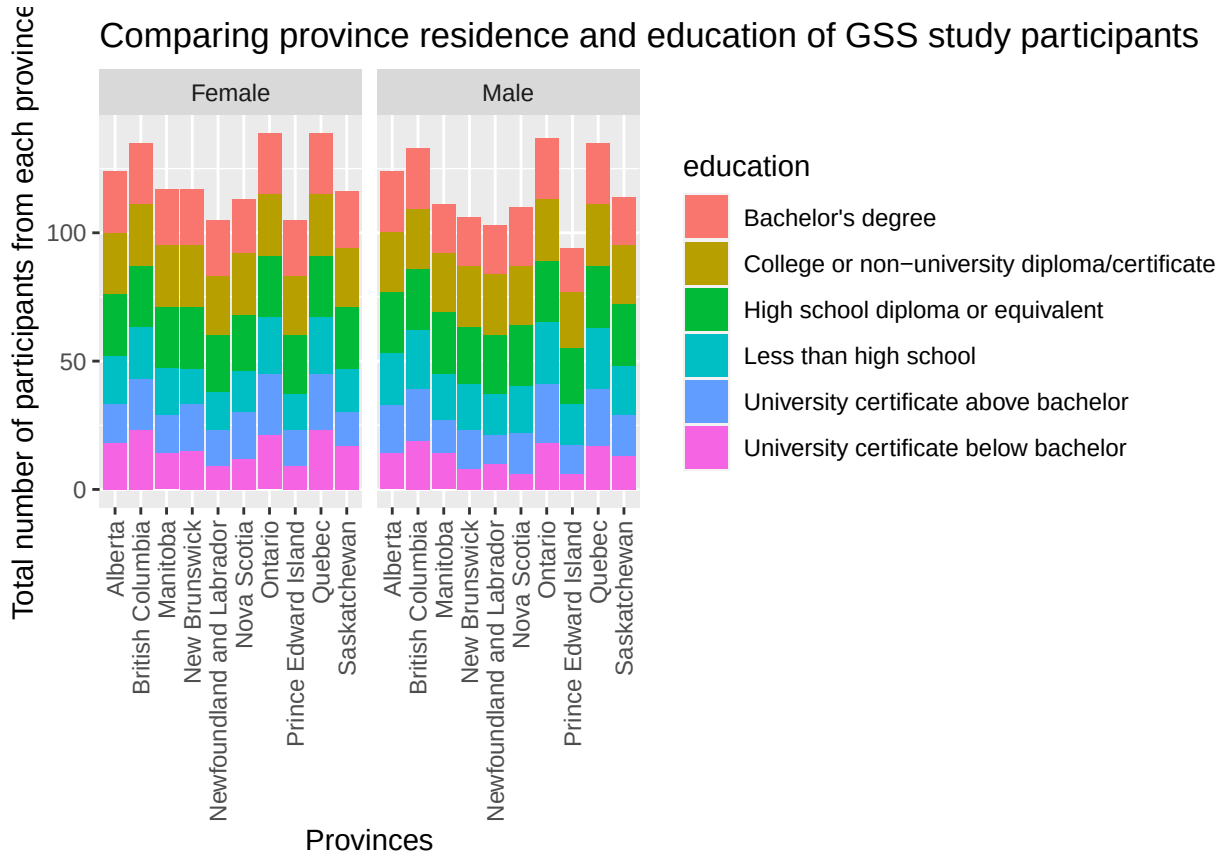

***Figure 2**: Breakdown of gender and education level of survey participants in each province.*

From the plot above (see Figure 2), we see the total number of participants in the GSS study from each province and their highest level of completed education for each gender. From the graph we can see that for both males and females, most people in the GSS study are from Ontario and Quebec, followed by British Columbia and Alberta. This makes reasonable sense as these four are the highest populated provinces in Canada. In contrast, for both males and females, we see that Newfoundland and Labrador has the lowest population. In terms of the highest level of education achieved, we see that both males and females seem to have a roughly even distribution across all levels of education represented in each province. Some notable exceptions are that it appears that Nova Scotia and Prince Edward Island have a low amount of individuals with a university certificate below bachelor's (particularly for men). Lastly, we can see that Ontario and Quebec have a relatively higher amount of people with a University certificate higher than a bachelor's in comparison to the other provinces.

Comparing income and age of GSS survey participants (separated by gender)

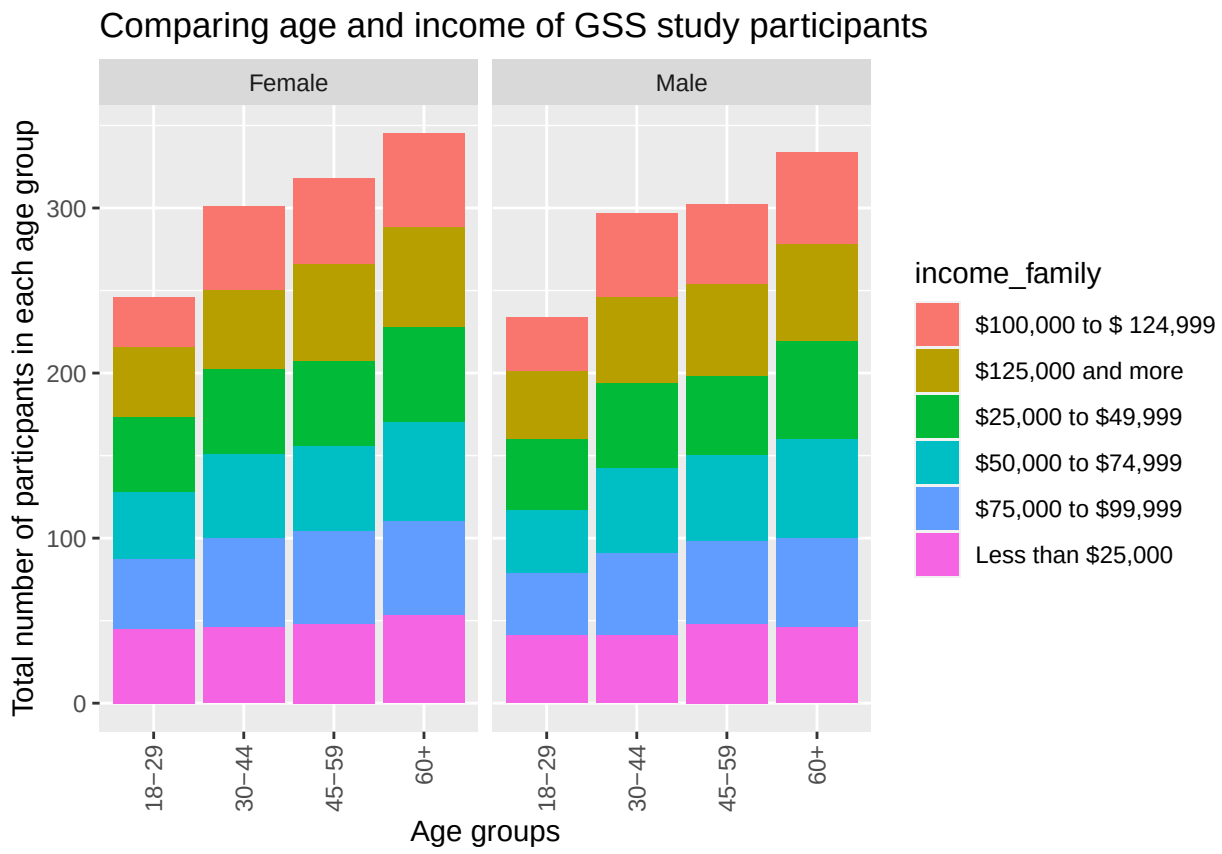

***Figure 3**: Breakdown of survey age and family income for male and female survey participants.*

From the plot above (see Figure 3), we see the total number of participants in the GSS study in each age group compared to their pre-tax income levels separated by gender. From the graph, we are able to see a pattern where for both male and female individuals, there are more GSS survey participants as the age interval increases. As a result, those between 18-29 have the lowest level of representation in the GSS survey and those above the age of 60 have the highest representation. In terms of income, we see that both males and females have a similar distribution of income level. As for the distribution itself, there seems to be a pattern where as the income level increases, the disparity between the younger and older age groups becomes more pronounced. As such, we see that there seems to be an equal amount of people making less than 25 thousand across all age groups, but there are a noticeably larger amount of people above the age of 60 making above 100 thousand CAD compared to individuals between ages 18 and 29.

Methods

In this analysis, we aim to predict the overall popular vote for the next Canadian Election. In order to predict the outcome, we decided to create a model which estimates the probability of voting for different parties in each Province of Canada. Our response variable is voting for a selected party, which is a binary variable. In order to estimate a binary variable, we utilize a multi-variable logistic regression model. There are 20,602 observations in our data, but it is not a good representation of the national population. Therefore, we will partition data into demographic cells (in our case by province) and estimate the response variable for each province. Then, we will apply a post-stratification model and aggregate the cell-level estimates up to a population-level estimate, and predict the selected party's votes in the next federal elections. Finally,

we will compare the percentage of the total votes for each party in order to predict which party will win the election.

Model Specifics

We will be using a multi-variable logistic regression model to estimate the proportion of voters who will vote for the selected party. A logistic regression model is used for binary classification models, where the response is either 1 or 0. In the case of our model, the dummy variable is set to 1 if an individual plans to vote for the seleceted party, and set to 0 otherwise. The logistic regression model used in the study is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{education} + \beta_3 x_{family\ income} + \beta_4 x_{gender}$$

Where:

- β_0 = y-intercept (The value of dependent variable when independent variables are 0)
- β_1 = For every increase in age group we expect the log odds to increase by β_1
- β_2 = For every increase in education level we expect the log odds to increase by β_2
- β_3 = For every increase in family income category we expect the log odds to increase by β_3
- β_4 = Differential intercept coefficient for gender (measures the effect of the gender on logit of voting to selected party)

Assumptions for Logistic Regression

- We assume that our explanatory variables are independent of each other
- There are no extreme outliers
- There is a linear relationship between our explanatory variables and the logit of the response variable

Issue: One problem with standard logistic regression is that we assume that our observations about different variables are independent, but we know it is not the case in reality. Some group behaviors have an effect on each other. We will use multi-level regression and a post-stratification model.

Multi-Level Modelling

The multi-level regression model specifies a logit transformation of the mean for a binary response variable, voting for a selected party. In the model, there are two different levels of independent variables. Level 1 is for individual characteristics, which may vary between all the individuals. We will use these characteristics for predicting our result with the logistic function, as mentioned above. Level 2 is group characteristics, which also have an effect on the individual characteristics. In our case, we believe that people living in the same province have a similar perspective in life and tend to vote for a specific party according to their individual characteristics.

In the multi-level regression we model our data in two different levels:

- Level 1 (individual level):
  - Age group, education, family income, gender
- Level 2 (group level):
  - Province

Assumptions for Multi-Level Modelling

- Level 1 variables are independent
- The random effects are normally distributed

Post-Stratification

We constructed a hierarchical logistic regression model in order to calculate the mean of a binary response variable conditioned on post-stratification cells, which are age group, education, family income, and gender as the first layer and province for layer two. The hierarchical model allows us to fit more cells than it is possible to fit using classical methods, thus allowing us to include more population-level information while also allowing us to include all the information used in standard survey sampling inferences [4]. We are combining the small-area estimations with the population information. The post-stratification model to find this collective proportion is:

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

Where: $\hat{y}^{PS}$ = estimated proportion of voters for the population N_j = sample size in each province $\hat{y}_i$ = estimated proportion of voters in each province

In simple terms, in order to estimate the proportion of voters we use the cells obtained from the multi-level regression model. In the previous section, we decided to divide the population per age group, education level, family income, gender. With the post-stratification model we are aiming to find an aggregate result which will give us a more reliable proportion about the population's votes for the selected party.

We will implement the same methodology and model construction for the three largest parties, the Liberal Party, the Conservative Party, and the New Democratic Party (NDP). Comparing the results for each model will enable us to make a prediction of the outcome of the next Canadian federal election

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

Results

The post stratification computations on the logistic regression model provide predictions for the next federal election in Canada. For each Canadian province, the predicted percentages of the total vote for the Liberal Party, Conservative Party, and New Democratic Party (NDP) are highlighted below (see Table 3).

Province	Liberal Vote	Conservative Vote	NDP Vote
Alberta	22.35%	26.60%	11.20%
British Columbia	23.45%	25.96%	10.69%
Manitoba	22.26%	27.09%	10.41%
New Brunswick	22.28%	26.21%	10.35%
Newfoundland and Labrador	21.72%	27.20%	10.10%
Nova Scotia	22.76%	25.88%	10.56%
Ontario	23.43%	25.72%	10.76%
Prince Edward Island	22.35%	25.89%	10.80%
Quebec	22.13%	25.64%	10.85%
Saskatchewan	21.90%	27.14%	10.72%

Table 3: Predicted Votes *for the Liberal, Conservative and New Democratic parties for each province (listed in alphabetical order). According to the model used in this study, the Conservative party is projected to obtain the highest percentage (approximately 25-27%) of the total vote in each Canadian province. Meanwhile,*

the Liberal party (approximately 21-24%) and New Democratic Party (approximately 10-11%) are projected to obtain a smaller portion of the total votes in each Canadian province.

As a better visualization of the results, the barplot below (Figure 4) illustrates the projected votes for the Liberal, Conservative, and NDP parties respectively. The results are shown as a percentage of the total vote in each Canadian province.

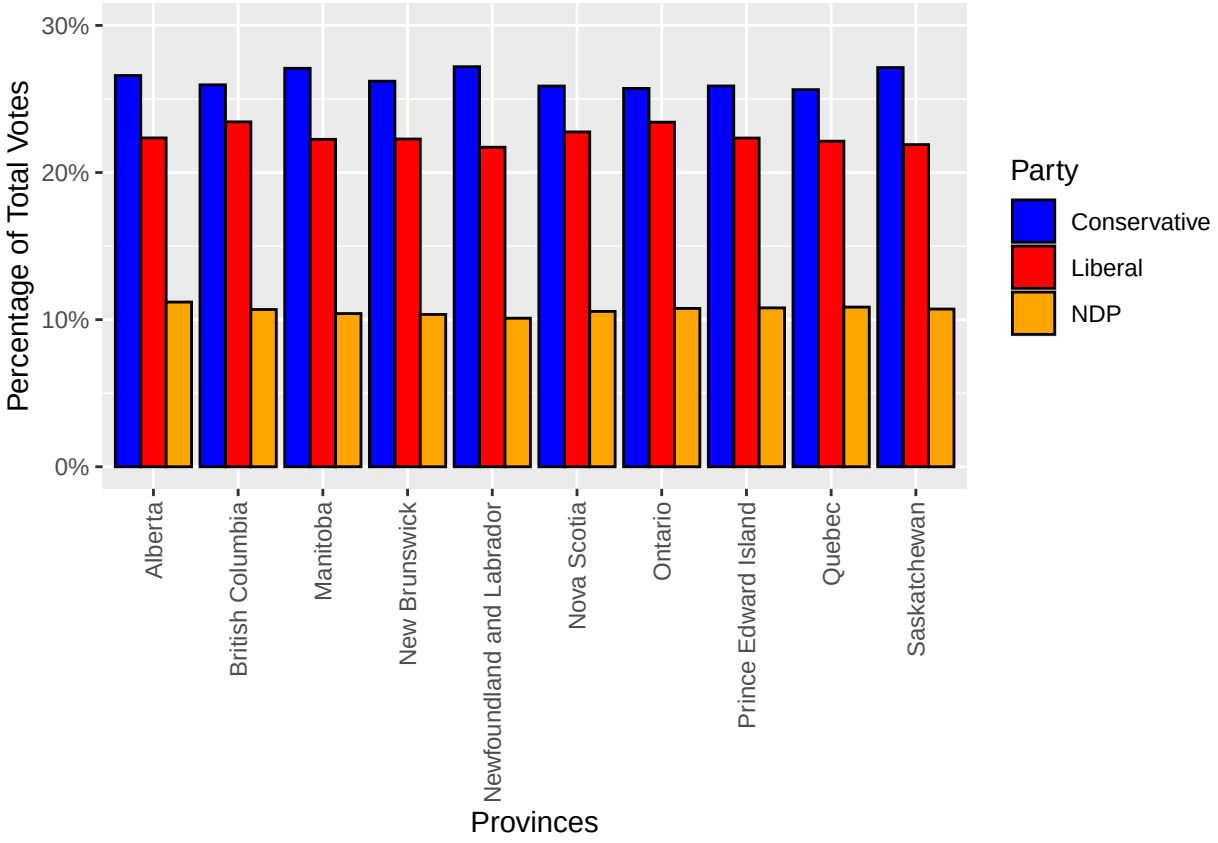

Figure 4*: The barplot illustrates that the Conservative Party is projected to win the popular vote in each Canadian province. Meanwhile, the Liberal Party is projected to obtain a slightly lower percentage of the popular vote and the New Democratic Party is projected to obtain the lowest percentage of the votes between the three parties.*

The results appear to indicate that the Conservative Party is projected to obtain the highest percentage of the total vote in each province, suggesting that the Conservative Party is predicted to win the next Canadian federal election. These results seem to be consistent with recent federal elections where the Conservative and Liberal parties have received the largest portion of the popular vote.

Conclusions

To test our hypothesis of the Conservative Party winning the popular vote in the next federal election, we utilized a multi-level poststratification model to predict the odds for voting for the top three parties, namely the Liberal, Conservative, and NDP parties.

With Canadian federal elections following the parliamentary system, the results of our data analysis provide us with projections on how each of the provinces will vote in the next election. From the results of the post-stratification model, we can infer that the Conservative Party will win the popular vote, but since the Canadian federal election is not determined by the popular vote, we would also need to take into account the

way each regional area votes with respect to the Members of Parliament. The data and model implemented in this study is limited to a provincial breakdown of votes and therefore, we are not able to determine which region the survey participants belong to. The results of this study provide us with insight into making a prediction for the result of the popular vote, which largely but not always coincides with the election results.

While this poststratification model was constructed using a multi-level logistic regression model, the available variables which were accessible to both the census data and the CES 2019 survey were quite limited. As such, the model implemented in this study was not able to evaluate the prediction using other information that may have been useful, such as views on certain politicians and policies. By integrating more independent variables into the model, we would be able to increase the cell number and ultimately reach a more accurate prediction.

To be able to further strengthen the model implemented in the study, it would be beneficial to obtain data in the CES survey that had more of an overlap with the census data. Additionally, having information on the regional breakdown of survey participants and the corresponding Members of Parliament would enable us to make an accurate prediction on the overall outcome of the federal election.

Bibliography

1. Canadian Election Study https://ces-eec.arts.ubc.ca/english-section/home/. (Last Accessed: Nov 4th, 2021)

2. (2017, Feb 27). *The General Social Survey: An Overview*. Statistics Canada. https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm. (Last Accessed: Nov 4th, 2021)

3. Dodge M. (2020). *The Average Canadian Salary in 2019*. Jobillico. https://www.jobillico.com/blog/en/average-canadian-salary/. (Last Accessed: Nov 4th, 2021)

4. Buttice, MK, and B Highton. "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" Political Analysis, 6 July 2017, https://escholarship.org/uc/item/5wc2g12h. (Last Accessed: Nov 4th, 2021)

5. Zhu, Hao. Create Awesome HTML Table with Knitr::Kable and Kableextra. 19 Feb. 2021, https://haozhu233.github.io/kableExtra/awesome_table_in_html.html#Table_Footnote. (Last Accessed: Nov 4th, 2021)

6. "Legends in Graphs and Charts. Statistics for Ecologists Exercises." Data Analytics, 13 Nov. 2019, https://www.dataanalytics.org.uk/legends-on-graphs-and-charts/. (Last Accessed: Nov 4th, 2021)