Assuming full voter Turnout, Justin Trudeau and the Liberal Party Would win 54% of the popular vote in the 2019 Candian Federal Election

December 20, 2020

Abstract

On October 21st, 2019, the 43rd Federal Election resulted in Prime Minister Justin Trudeau and Liberal Party winning a minority government over Andrew Scheer and the Conservative Party. However, voter turnout numbers decrease to a measly 62%. Using both data from the Canadian Election Study and General Social Survey we conducted a multilevel regression with post-stratification using a Multinomial logistic regression model in order to determine the nation-wide popular vote distribution if voter turnout was assumed to be 100%. We discovered that Trudeau and the liberal party would have won 54% of the popular vote in 2019, most likely claiming a majority government. Our analysis displays the significance of voter turnout in determining elections results, how key demographic variables such as age and education level influence voter preferences and how political groups can target these voters in future elections.

Keywords

Canadian Federal Election, voter turnout, Multilevel Regression with Post Stratification, Liberal Party

1 Introduction

In 2019, Justin Trudeau and the Liberal party won the Canadian Federal Election and sustained a minority government through capturing a total of 157 seats. The race was fairly close as the Conservative Party gained a total of 26 new seats and Justin Trudeau and the Liberal Party lost 20 seats making the election much tighter than the previous year. In addition, the Liberal Party only obtained 33.12% of the popular vote, losing to the conservative party who obtained 34.34% of votes. However, one of the main stories of the election was that voter turnout was down, at 62%, from 68% in the 2015 Federal election.

The political system in Canada is based off of that in the United Kingdom. The house of commons has 338 individual seats which are held by members who are elected by citizens in the general election. The party that obtains the most seats wins the election and if they hold at least 50% of all electoral seats, they hold a majority position in the government. However, the datasets we obtained are not specific to each electoral division, and thus our results will instead be based off the popular vote. Although the popular vote usually falls in line with the party with the most seats, instances such as seen in 2019 prove this is not a guarantee as the Conservative Party won the popular vote and still ended up falling almost 30 seats short of the Liberal Party. Thus, if we see a relative increase in the popular vote for the Liberal Party, we can assume that they still won the election and depending on the size may have captured a majority government. On the other hand, if we see a significant increase in the popular vote for the Conservative Party, and most importantly in Ontario, we may be able to come to the conclusion that the party could have obtained at the least a minority government.

In this report we will use Multilevel Regression with Post Stratification in order to analyze significant character traits of voters which determine their voting preferences and then make a final predication on what would have happened in the 2019 Canadian Federal Election if all eligible voters voted. The variables we will be

analyzing are; age, gender, province, education, if respondent has had children and marital status. In order to conduct our analyze, we collected two large datasets, the CES (Canadian Electoral Study) dataset and the GSS (General Social Survey) Dataset. First, we will apply the CES dataset to our regression and then using the GSS dataset we will conduct a post-stratification. Our results predict that if all eligible Canadians voted in the 2019 Federal Election, the Liberal Party would have captured a majority government through obtaining a 54% of the popular vote and an estimated 174 electoral seats. However, since there is high uncertainty in our model and with voter preferences, we cannot be certain of our results. Despite of this, the results overall will be a good indicator on how important voter turnout is to election results.

The next section will display and thoroughly analyze the datasets used in our report as well as the adjustments made to the datasets in order to fit our regression model appropriately. In section 3 we will present and analyze the model used in our report and inspect its predictions and their accuracy. Furthermore, we demonstrate the strategy in which we used to generate a predication of the change in the voting as well as go over the weaknesses of our model. Next in section 4, we will display the results of the model and the predication from the previous sections accompanied by methods in which these values can be interpreted.

To conclude, in Section 5, we display the significant finding of the report and analyze the importance they have in displaying the significance of voter turnout on elections. In addition, there is an appendix under the discussion in which code for the entire report can be found

2 Data Discussion:

We will use data from the 2019 Canadian Election Study in order to create a model that predicts voting preferences of an individual in the 2019 Federal Election and then we will use data provided by the General Social Survey through post-stratification to make predictions on election results.

2.1 2019 CES data

Since 1965, the Canadian Election Study has provided an unparalleled view into the political behavior and attitude of Canadians. For the 2019 CES study, data was collected between September 13th to October 21st of 2019 via online survey.The survey aimed to collect demographic variables, political views and voting intentions prior to the 2019 Canadian Federal Election. The online sample for the 2019 Canadian Election Study was comprised of a two-wave panel with a modified rolling-cross section during the campaign period and a post-election recontact wave. For our study we will just be using the initial dataset obtained during the campaign period. During the campaign period, a total of 74,548 individual were contacted to take the online survey with 37,822 respondents completing it to a high enough standard to pass the removal criteria. In addition, the survey itself is fairly large, as it is composed of 620 variables which ranged from political preferences to social-demographic questions.

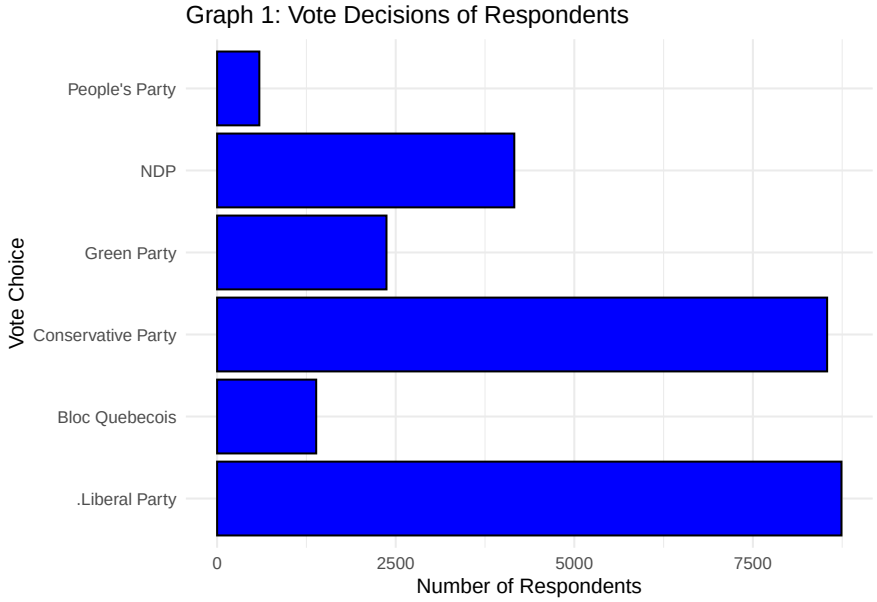
The survey data was weighted using a non-probability sampling method run by Qualtrics in order to mimic the Canadian population, based on factors such as age, sex, language (French or English) and region. For gender they aimed for a 50/50 split between male and female, and for age they aimed to have 28% of our respondents aged 18- 34, 33% aged 35-54 and 39% aged 55 and higher. Next the regions were: Atlantic (Newfoundland and Labrador, New Brunswick, Nova Scotia, Prince Edward Island), Quebec, Ontario Prairies (Manitoba, Saskatchewan, Alberta), and British Columbia. Within each of those regions, the provincial quotas were split evenly. For language, they aimed to have 80% French and 20% English within Quebec, 10% French within the Atlantic region, and 10% French nationally.

In addition, The CES handled the non-response problem very well. All survey responses were thoroughly analyzed and were removed if any problems arose. Some reasons responses were removed were ineligible being under the age of 18, did not consent to survey, if respondents did not complete initial quota demographics or if they were deemed a “speeder” where respondents completed the in less than 500 seconds they were also removed. Incomplete responses, those who “straight-lined” grid questions (“straight liners”), and respondents

whose postal code did not match their province were additionally all removed from the data file. Furthermore, Duplicate responses were also flagged using information such as IP address, year of birth, education and religion. If a duplicate response was identified, the first response was kept and the second response was removed from the data.

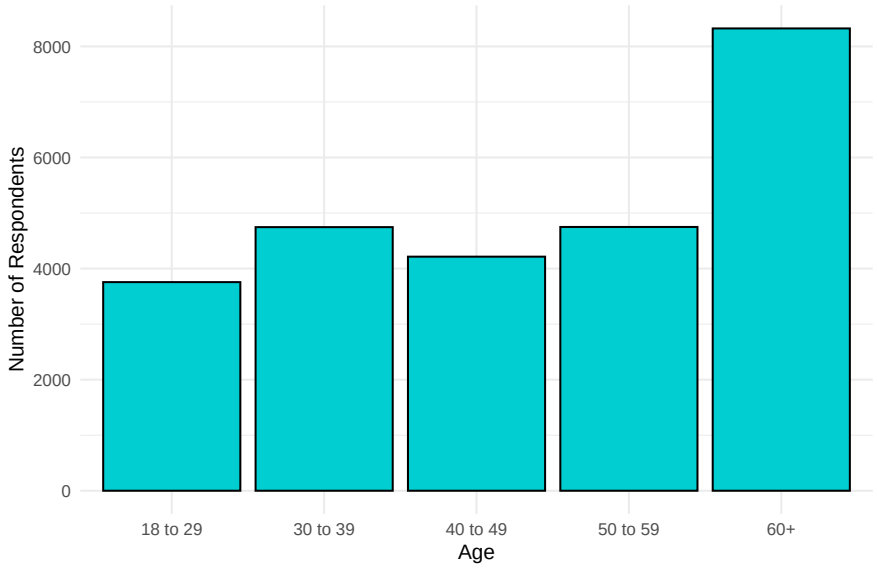
In this study we will be examining the impact of; age, gender, province and education on voting preferences in the 2019 Canadian Federal election, in order to use them to determine election results if voter turnout was 100%. To do so we created a new dataset (see Appendix) that comprises of our variables of interest; age, gender, province, education, children and martial_status. In addition, since we will be using post-stratification for this study, we made few alterations to the data in order for both datasets to have matching column names and response options. First for all categories, the option "Don't know/ prefer to not answer" was removed to focus on the significant response options. For age, we grouped ages 18 to 29, 30 to 39, 40 to 49, 50 to 59 and ages 60+. However, since we were given year of birth instead of age, we had to assume that since the survey was taken between September and October of 2019 that the majority of the respondents already celebrated their birthdays in 2019 previous to taking the survey, so age will be divided between what age respondents are turning in and not their exact age. Next for gender, we removed all responses other than male or female in order to match with the GSS dataset. In addition, for education, we grouped all responses by their highest level of education completed. Options "Master's Degree" and "Professional degree or doctorate" were grouped under "University degree or certificate above bachelor's degree", whereas the responses "Some University" and "Some college" were grouped with "Completed secondary/ high school" and any education levels lower than "Completed secondary/ high school" were grouped under "less than high school diploma". Next, all responses taken in "Northwest Territories", "Yukon" and "Nunavut" were removed as options in order to match with the GSS dataset. Then variables children and marital_status were renamed in order to match our GSS dataset.Finaly we removed all NA values from our cleaned dataset so that we are left with only complete cases and thus, our regression model can run smoothly.

2.2 Display of Survey Data



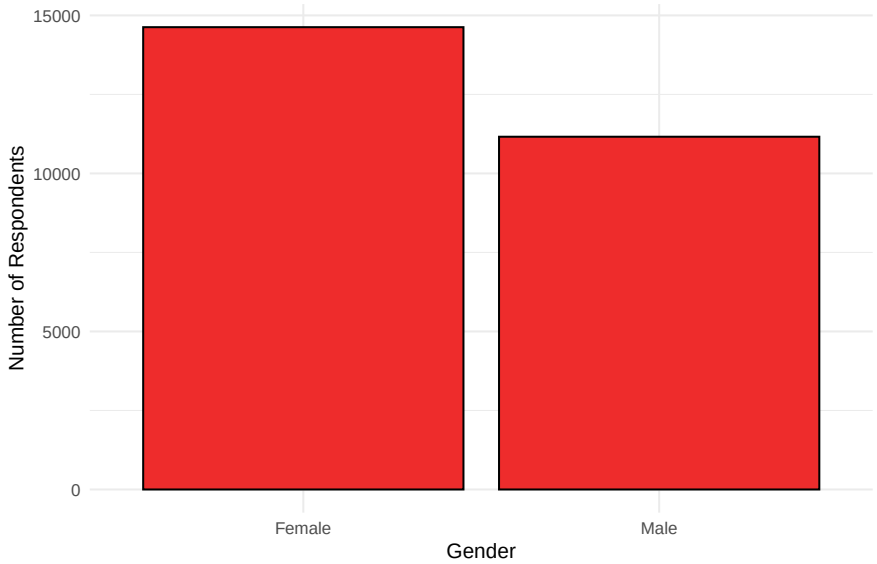
As seen in our first plot, Graph 1, we display the results of who respondents were most likely to vote for in upcoming 2019 Canadian Federal Election. As displayed above, the only two parties that have a chance to win are the Conservatives and the Liberal, where Liberal voters slightly outweigh conservative voters at 35% and 33% respectively.The small difference between the two parties suggests the outcome of the election will be very close and thus requires a very thorough analysis.

Graph 2: Age Proportions of Respondents

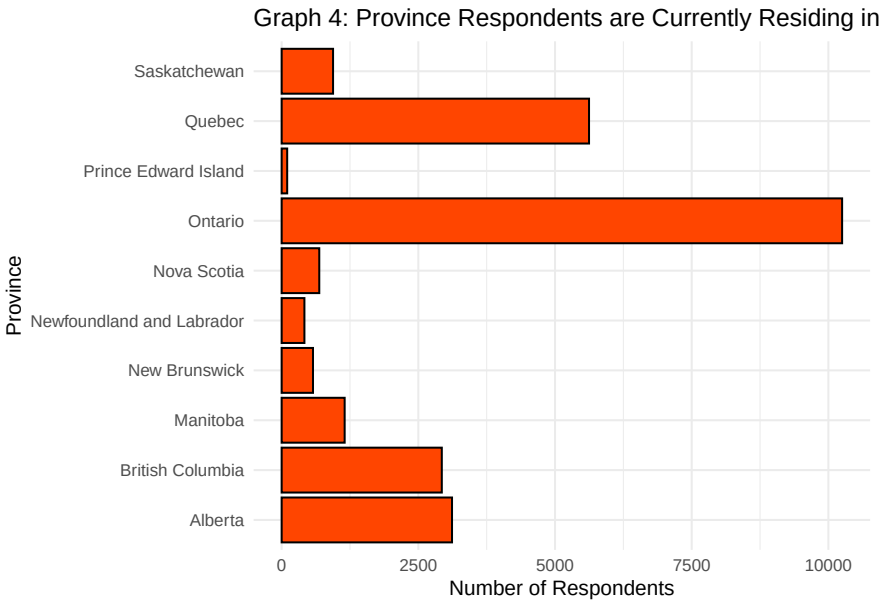


Next in Graph 2, we display how age was distributed throughout the respondents of the survey. We can see that the largest number of respondents fall under the age group "60+" with the other age groups having similar size at about half the number of respondents. Never the less, this inherently makes sense as the age range for this group is much larger than the other groups. Although this was to be expected, it could potentially result in there being a high standard error in the prediction for the "60+" age group as the group may have a high variance between its voters.

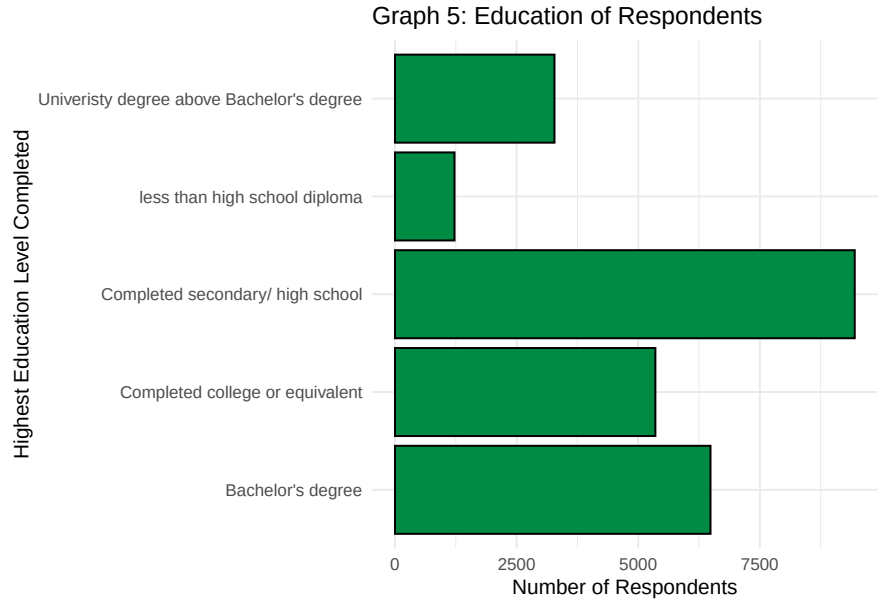
Graph 3: Gender Proportions of Respondents



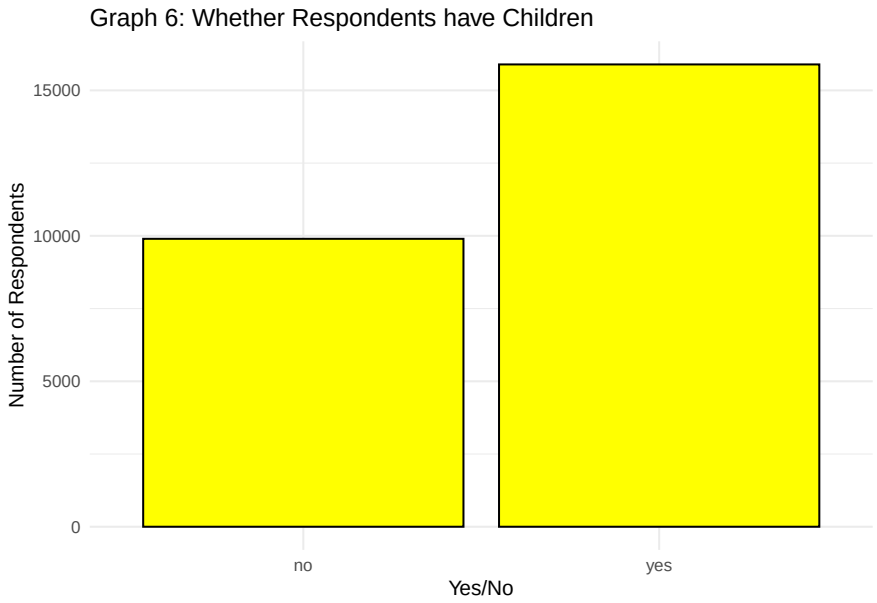
From Graph 3, we can see a significant difference between the proportions of Male and Female respondents with approximately 56% of respondents being female and 44% of respondents being male. Although the CES survey was weighted to mimic the nation wide gender ratio of 50% male and 50% female, as seen above females were over represented in the survey. However, we will be using a post-stratification dataset which more accurately represents the distribution of male and female voters across Canada and thus this will most likely have no effect on our prediction results.



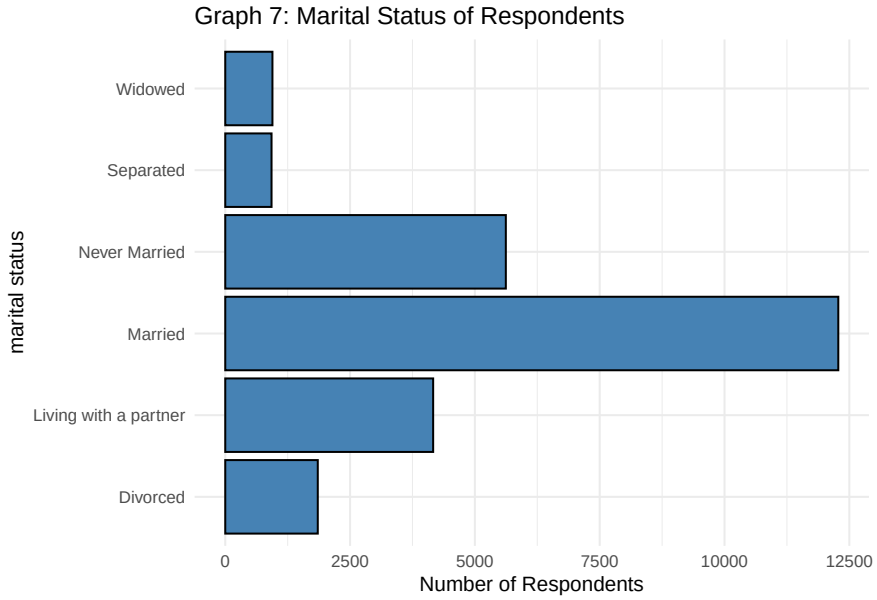
As seen in Graph 4, we display the provincial distribution of respondents, with exception of Yukon, Nunavut and the NorthWest Territories. With about 40% of the populous, Ontario accounts for the largest amount of respondents with Quebec in second at around 22%. Next Alberta and British Columbia each represent about 12% of the populous and the remaining 6 provinces account for only around 15% of the populous collectively. These results are fairly similar to the real-world distribution of the Canadian population, further proving that the CES sample is properly representative of the Canadian population.



Next in Graph 5, we display the distribution of the highest education level obtained by respondents. With 37% of the sample, respondents with high school as their highest level of completed education accounted for the largest portions out of the other responses.However, this populous does include respondents that completed some university or college and students over the age of 18 currently enrolled in post secondary school. The distribution of these responses were similar to reports publish by Statistics Canada (see References) and thus further provided the validity of the CES dataset.



In graph 6, we display the distribution of respondents that have had children and respondents that have not. Respondents that have had children outweigh respondent that have not in by a little over 50% which is similar to real-world statistics provided by Statistics Canada.



Lastly, in Graph 7, we display the distribution of marital status among respondents. Not surprisingly we see that around 45% of respondents are married which is similar to results pulled by statistics Canada stating that 46% of Canadians over the age of 15 were legally married. People that were never married had the second largest portion of our sample, followed by living with a partner, then divorced, then widowed and finally separated. This provides further evidence that the CES dataset is representative of the Canadian population over the age of 18.

2.3 GSS Data

The General Social Survey or GSS data we used in this paper was collected back in 2017 by the Canadian General Social Survey Program. The 2017 GSS is a sample survey with a cross-sectional design which sampled over 20,000 people above the age of 15 that currently live in the 10 provinces in Canada. This survey was conducted between February 1, 2017 to November 30, 2017 via telephone interviews. This survey included a wide variety of over 400 important demographic and political opinionated statistics. For examples some of these statistics included were marital status, highest education level obtained, living arrangements and total number of children. As seen below we have displayed the 2017 GSS data.

The sampling for this survey was based on a stratified design using probability sampling. The stratification was conducted on the province/Census Metropolitan Area (CMA) level for a total of 27 different strata. The sampling design was conducted in two stages. First, sampling units are initially grouped by telephone number. Then during the second phase of sampling, one representative over the age of 15 from each household which meets the criteria for selection is randomly selected to complete the questionnaire. This created an initial survey sample of around 43,000 individuals.

The non-responsive from this survey was particularly high at around 52.4%. To reduce the effects of non-responsive, the GSS implemented a three-stage plan in which they extensively analyzed responses. First, complete non-response submissions were removed from the data and this was completed within each stratum. Next, In the second phase, using auxiliary information from sources available to Statistics Canada, further adjustments were made to remove non-responsive. These households had some auxiliary information which was used to model propensity to respond. Finally, in the third phase, adjustments for partial non-response were made through using auxiliary information of households. In addition, through using well-tested questionnaire, proven methodology, strict quality control and specialized interviews, the overall level of bias was minimized significantly.

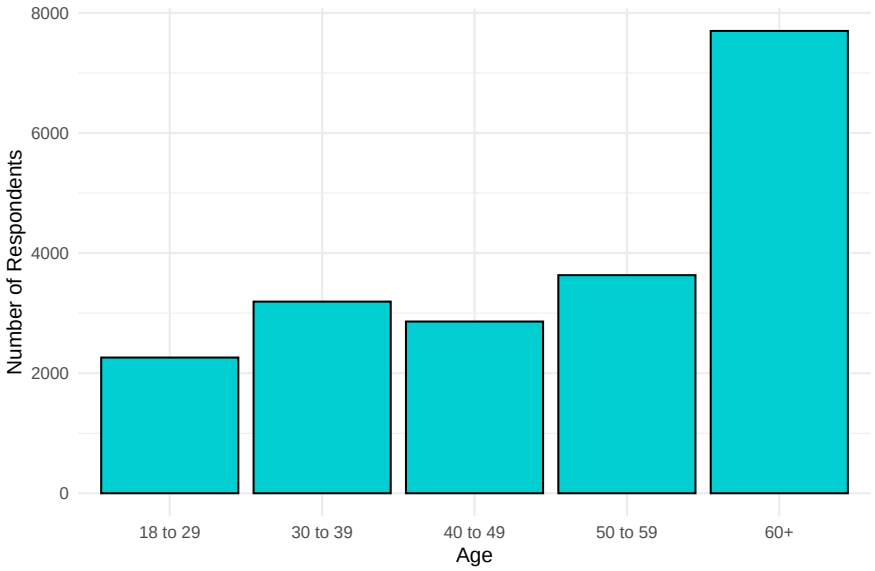
However, there are still some limitations this dataset faced in which the GSS did not properly account for. To begin territories Yukon, Nunavut and the Northwest territories were excluded from the dataset. This can be significant in our analysis as each territory accounts for 1 electoral seat and could have greatly impacted election results in determining what level of government the winner receives.In addition since the survey was conducted via telephone, households without telephones were underrepresented in the data.

In order for the GSS data to be entered in our regression function we matched up both the question and response options from this GSS dataset to those in The CES dataset. Thus, all groups for these graphs will appear the exact same except with slightly different distributions.For example, like the survey data, age was grouped into 5 categories; “18 to 29”, “30 to 39”, “40 to 49”, “50 to 59” and “60 +”. Further information on regrouping can be seen in the appendix. The variables we are going to be focusing on in this study are age, gender, province, education, children and marital_status and our objective is to determine the relationship between these six variables and respondents voting preferences regarding the 2019 Canadian Federal Election.

Although there are limitations to this dataset, the GSS accurately represents the Canadian population and thus will be reliable post-stratification dataset for out report.

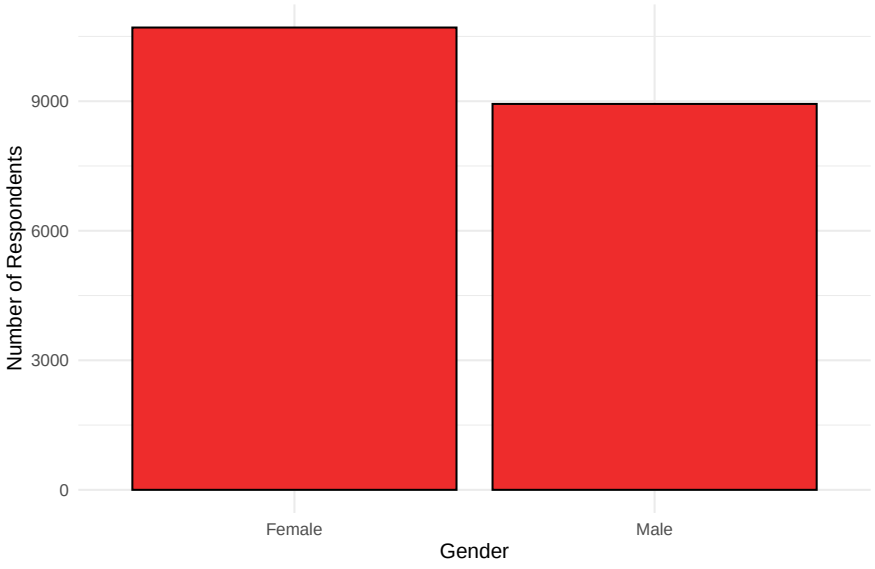
2.4 Display of GSS data

Graph 8: Age Proportions of Respondents

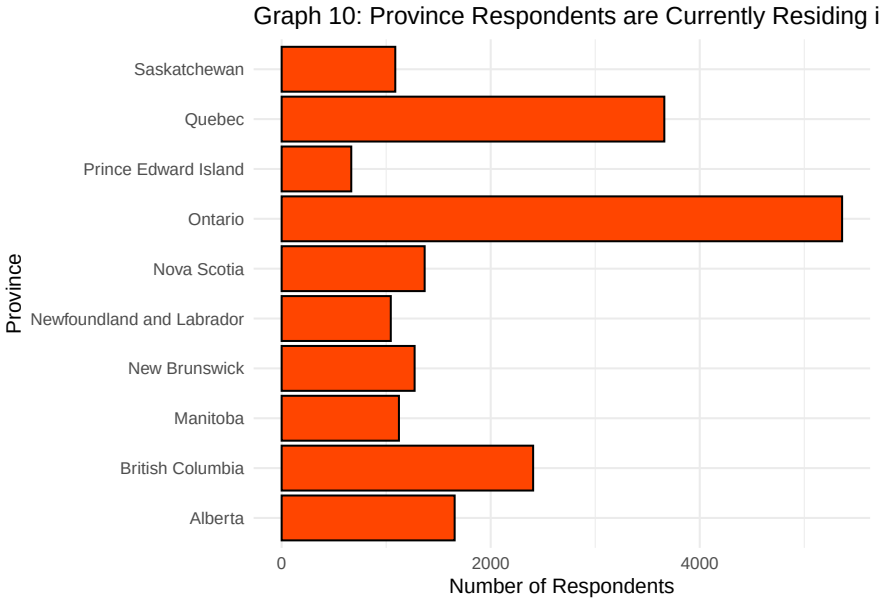


Graph 8 plots the distribution in age groups we see across our data. Much like the CES survey data we see a large portion of respondents fall into the 60 + age group totaling over 35%. We can see that the remaining age groups also hold the similar distribution as our initial dataset ranging between 11% to 17%. As mentioned previously this is to be expected as the age range for 60+ is much lager than the other age groupings and we expect this to have the same possible effect on the standard error in our predictions that we must pay attention to.

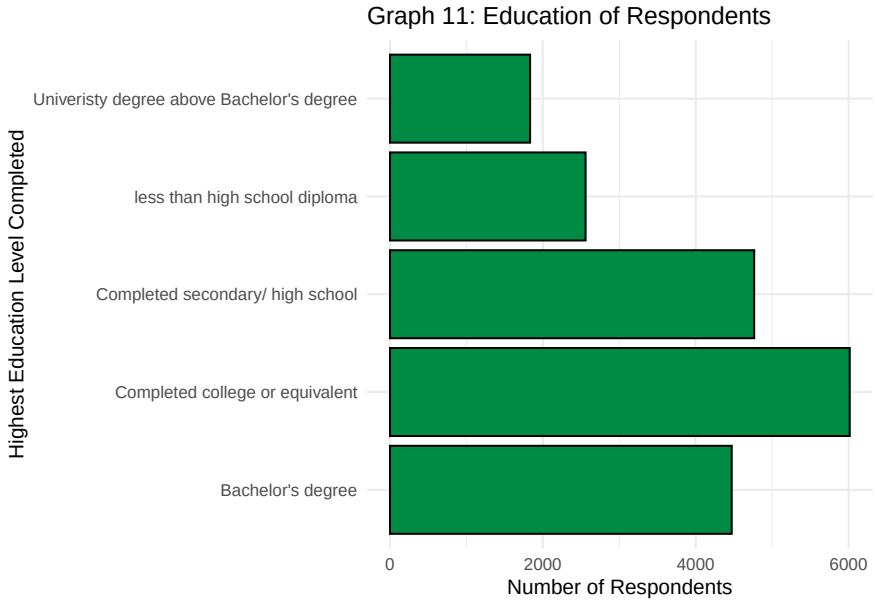
Graph 9: Gender Proportions of Respondents



Next,in Graph 9, we display the distribution of gender in the GSS dataset. Again, we notice a significant difference between Female and Male respondents, as females account for 53% of the data. However this distribution was closer to the national average of 50% female and 50% Male in comparison to our initial dataset.The low proportional difference between male and females was most likely due to a simple variation in the distribution of individuals contacted. As a result, we do not believe that this should cause any problem with our model's performance.

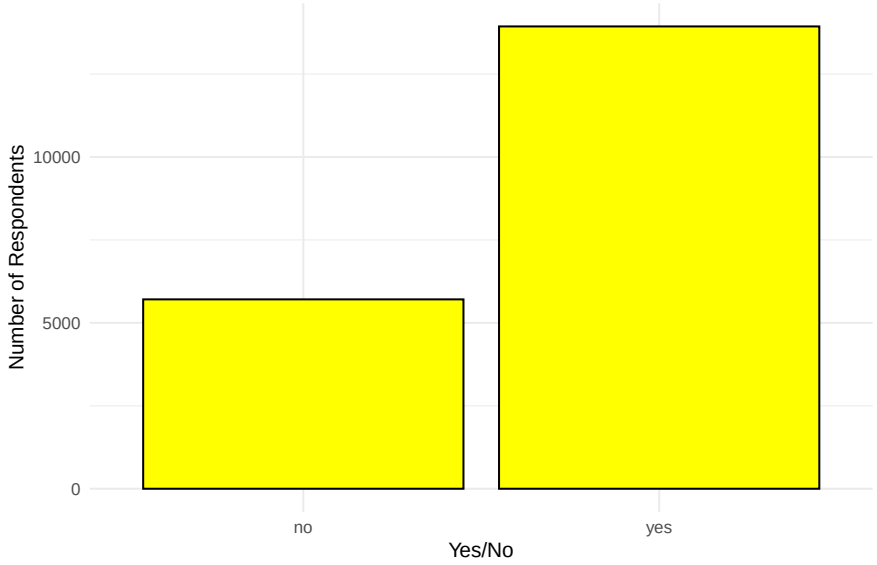


In Graph 10, we have displayed the regional distribution of our respondents. As we can see, the large proportion of respondents live in Ontario as seen in the CES dataset, as well with the other provinces making up similarly proportions with exception to British Columbia having a larger proportion of respondent than Alberta. However, this is more accurate when looking at the nation wide distribution of population. This fact combined with the similarities between the GSS data set and the distribution of the Canadian population provided by Statistics Canada strengthens the case for the GSS being using as the post-stratification dataset.



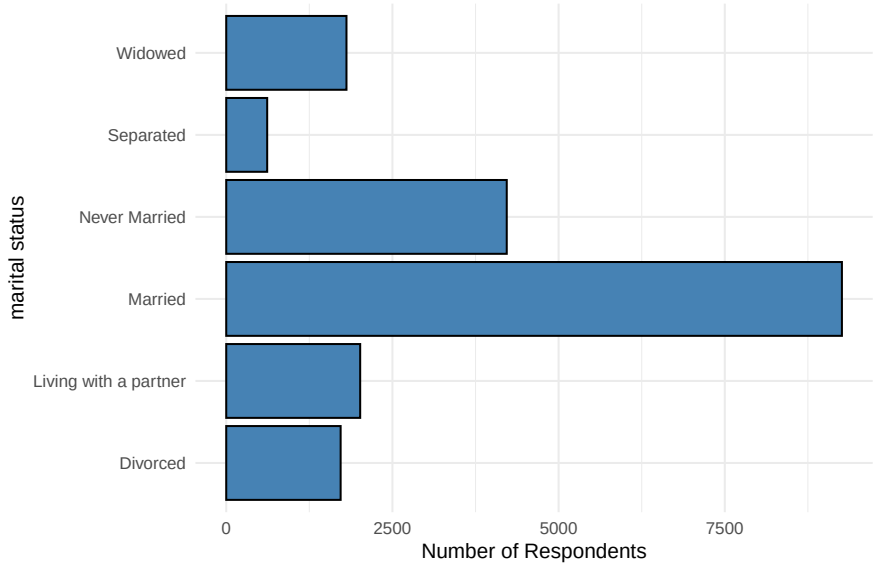
Next, from Graph 11, we see College or equivalent is the largest response at around 30% followed second by high school at around 25% as their highest level of education completed. Next 23% of respondents have a bachelors degree, 11% have not completed high school and finally 9% of respondents have completed a degree above the Bachelor's level. Through analyzing data from the 2016 census, out of around 19 million people aged 25-65, the distribution of education was extremely similar within 2% for each response category. This further validates the accuracy of the GSS dataset and strengths its case for being a concrete dataset to use for post-stratification.

Graph 12: Whether respondents have children



In graph 12, Again,we display the distribution of respondents that have had children and respondents that have not. Respondents that have had children similarly outweigh respondent that have do not have children in by a little over 50% which is similar to the nation wide distribution providied by Statistics Canada.

Graph 13: marital status of respondents



Finally,in Graph 13, we display the distribution of marital status among respondents. Similar we see that around 47% of respondents are married which is very close to the national wide average of 46%. In similar fashion to the CES dataset, People that were never married had the second largest portion of our sample, followed by living with a partner, then widowed, then divorced and finally separated. However we see that Widowed respondents outweigh Divorced respondents slightly which is not similar to the distribution among the Canadian population. We will keep a close eye on the significant of this factor to determine if it results in any misrepresentation of our predicted votes.

Now that we have viewed all our our predictor variables and confirmed that there were no troubling observations in the data, we are ready to move on to developing our model and begin to make forecasts for the election based on this.

First, however, we must explain the method we will be using to analyze the CES data.

2.5 Multilevel Modelling With Post-Stratification

In this section we will be conducting multilevel modeling through developing a regression model using data from our sample population in order to predict the outcome of target population's response. To use this strategy, first we will use the CES data to create a Multinomial regression model that can estimate the relationship between our response variable, vote selection, and our predictor variables, age, gender, province, education , children and marital status. Then we will conduct a post-stratification using the GSS dataset by grouping unique responses into cells. These cells are mutually exclusive and are used as a representation of our entire dataset. Then using the unique responses in our cells and our regression model we will predict which party respondents for each unique cell would have voted for in the 2019 Canadian Federal Election.

Non probability sampling, similar to the strategy used in our report is a great and very practical technique used by many statistics around the world. This type of sampling is very cost efficient when comparing other techniques such as probability sampling. Furthermore, through conducting a post-stratification using data that is representative of the entire population, we will be able to recognize and gain information on misrepresented groups. However, in order to make an accurate prediction, we need to have well defined cells that are proportional to the target population.As seen in real world statistics, it is very possible that our cells could have misrepresented certain groups so we must account for that in our analysis.

3 Model Discussion and Development

3.1 Regression Model

For the purposes of this study, using R, we will be creating and using a Multinomial logistic regression model to both analyze the Canadian Election Survey data and predict election results using the GSS data. We will be using MRP to come to our result as the GSS more accurately depicts key demographic characteristics of the Canadian population in comparison to the CES. Simply using the CES data to come to our results could lead to bias as the survey sample is not representative of the Canadian population.

Multinomial logistic regression is an extension of binary logistic regression which allows for a dependent variables to have more than 2 categorical outcomes. Similar to binary logistic regression, Multinomial logistic regression utilizes maximum likelihood estimation to solve the probability of the different categorical responses of the dependent variable.

In addition, Multinomial logistic regression does not assume things such as normality, linearity or homoscedasticity and since our predictor variables are categorical, they would not pass any of the strict assumptions that linear models require. Linear regression models are use more in theory as they are not effective when tackling issues that require the use of real world statistics. They require many unrealistic requirements unlike our multinomial logistic regression model.For this reason coupled with with the fact that our dependent variable has more than 2 possible responses, we chose to use a Multinomial logistic regression model for our study. Although there are only 2 parties that have a chance at obtaining at least a minority government, the Liberal Party and the Conservative Party, votes for smaller parties such as the Bloc Quebecois and NDP have a large effect on determining the strength of which government is obtained by the either the Liberals or the Conservatives.

To begin, using the CES dataset we must create a model to predict voter choice. Our dependent variable describes the possible voting choices in which Canadians can make in the 2019 Federal election.The variable contain 6 possible categorical options in which respondents can choose from; "Liberal Party", "Conservative Party", "NDP","Bloc Quebecois" and "People's Party". We then have 6 key demographic characteristics of respondents which will be used to determine their voting preferences; "age", "province", "gender", "education", "children", and "martial_status".

3.2 Model Validation

in order to validate the model, we will first look at the accuracy of the model in predicting voter preferences of respondents for our survey dataset.Through running code in R as seen in the appendix, our model only predicted 44% of the respondents from the survey data correctly. This number is definitely concerning and will be taken into consideration when determining the accuracy of the results produce through applying our model to our post-stratification dataset. However, given the number of responses to our dependent variable we are using for this study, 44% accuracy is not a horrible result, although it will most likely lead to many inaccurate predictions.

To further validate our model we will determine and analyze the MSPE and MSRes of our model. We concluded that 20% of the dataset should be used to calculate the MSPE, which total 4972 observations and the remaining 19889 were used to calculate the MSRes. Through running code in R, we found the MSPE of our model to be 0.679 and MSres came out to 0.677, leaving a very smaller difference of less than 0.002. An MSPE value that is close to the MSRes value based on the regression fit to the model-building data set, indicates a high predictive ability of our model. This rivals are previously results and thus proves the validity of our model selection being a multinomial regression.

3.3 Final Model

From the output provided by table 1, as seen in the output section under "table_1.pdf", we are able to determine the final formula for our regression model. Since there are 6 possible categorical responses to our dependent variable voter choice, using basic concepts of multinomial logistic regression provided by Charles Zaiontz, as seen in References, and through selecting the Liberal Party as the conditional response, we are able to provide the mathematical notation of probability of respondents selecting each of the 6 possible parties as their vote for the 2019 Federal Election:

$$Pi(vote\_2019 = Liberal\ Party) = \frac{1}{1 + \sum_j^r e^{\beta_j \cdot x_i}}$$

$$Pi(vote\_2019 = Bloc\ Quebecios) = \ P(vote\_2019 = Liberal\ Party) \cdot e^{\beta_2 \cdot x_i}$$
$$Pi(vote\_2019 = Conservative\ Party) = \ P(vote\_2019 = Liberal\ Party) \cdot e^{\beta_3 \cdot x_i}$$
$$Pi(vote\_2019 = Green\ Party) = \ P(vote\_2019 = Liberal\ Party) \cdot e^{\beta_4 \cdot x_i}$$
$$Pi(vote\_2019 = NDP) = \ P(vote\_2019 = Liberal\ Party) \cdot e^{\beta_5 \cdot x_i}$$
$$Pi(vote\_2019 = People's\ Party) = \ P(vote\_2019 = Liberal\ Party) \cdot e^{\beta_6 \cdot X_i}$$

Bj is a vector of the coefficient for the explanatory variables for the jth outcome for j = {1,2,3,4,5,6}. For our model B1j is the coefficient for age, B2j is the coefficient for gender, B3j is the coefficient for province, B4j is the coefficient for province, B5j is the coefficient for children and B6j is the coefficient for marital_status. xi represents the values of the ith observation of each explanatory variable. Thus, x1i would be the value of age, x2i would be the value of gender, x3i would be the value of province, x4i would be the value of education, x5i would be the value of children, x6i would be the value of marital_status for the ith observation.Finally r represents the fact that we have r+1 possible outcomes for our dependent variable.

4 Results

4.1 Regression Model Results

As seen in Table 1, using the summ() function from jtools package, we have displayed a table of summary statistics which includes the predictor's; estimate, standard error, z-value, and p-value. Through analyzing

these values, we can the significance of our predictor variables in determining voter preferences of Canadians in the 2019 Federal Election. Since we are using a multinomial model, one response option from the dependent variable will be used to condition the results for the other 5 possible responses for our dependent variables. We chose for the Liberal Party to be the conditional response for our model. Thus, the estimate value tells us the change in the log odds of the respondent's vote going toward the Liberal party in comparison to another Federal political party, given their response to the question. Next, the standard error shows us the expected error we observe in a particular estimate value. The z-value and p-value will be used to determine if we can reject the null hypothesis that our estimate value is zero. This allows us to determine whether or not our predictors are significant. For this analysis will we consider predictor variables significant if their z-value is more than 1.96 and the p-value is smaller than 0.5.

Since our predictors variables are set up categorically, each variable must be conditioned on the response that appear first in alphabetical order.This means that is estimate values indicate the difference we expect to see between a certain response and the conditioned response. For our analysis We will mainly focus on the Liberal and Conservative Party as they are the only 2 political parties that had a chance at winning the election.

Age group Estimates: The responses for this variable were conditioned on the age group "18 to 29". As we can see, the values of all estimates when comparing the Liberal Party and the Bloc Quebecois are positive, which means that the younger voters in the 18 to 29 age group are more likely to vote for the Liberal and as respondents get older, they become more likely to vote for Bloc Quebecois. Next for the NDP, Green Party and the People's Party all values were negative and in descending order, so as respondents get older, they become less likely to vote for these parties and more likely to vote for the Liberal party. When analyzing the Conservative party, we see for age groups 18 to 29 and 60+ respondents are more likely to vote for the Liberal party, and for age groups 30 to 39, 40 to 49 and 50 to 59, respondents are more likely to vote for the conservative party. P and z-values: For Bloc Quebecois all age groups except 30 to 39 met the requirements of the z and p value confirming their significance as predictors. Next for the Conservative party we noticed that only age group 50 to 59 passed the requirement for the z and p values, and thus we don't believe age is a strong predictor when determining if a respondent chooses between Liberal and Conservative party. For the Green Party and the NDP, all categories met the requirements of the z and p values confirms that age is a significant factor in determining if a voter selects one of these parties. Finally, for the People's party only age groups 50 to 59 and 60+ passed the z and p value requirements, thus showing that age only has a significance in determining if people are voting for the people party if respondents are over the age of 50.

Gender Estimates: For this variable the conditional group is "female" respondents.Sorting parties from most to least likely to vote for if voters are a female we have; NDP, Green Party, Liberal Party, Bloc Quebecois, Conservative Party, People's Party. P and z-values: Through analyzing the p and z values, we see that all responses satisfy the requirements and thus for all parties, gender is a significant factor in determining the vote preferences of the respondent.

Province Estimates: The province predictor variable uses "Alberta" as the conditional response. For Bloc Quebecois as expected, we can see the highest probability of a voter choosing this party if they live in Quebec with an estimate of 7.908. However as seen in the next paragraph, Quebec is the only province of significance in determining if a person will vote for the Bloc Quebecois, so we will not analyze the estimates for the other province. Next when analyzing the Conservative party, ranking provinces from most to least likely to vote conservative we have Saskatchewan, Alberta, Manitoba, British Columbia, Ontario, New Brunswick, Quebec, Nova Scotia, Newfoundland and Labrador, Prince Edward Island. We see that people who live in West Canada are far more likely to vote for the Conservative party than compared to East Canada. For Parties NDP,Green Party and People's Party, their voting preferences discussed later when analyzing the distribution of predicted votes across provinces. P and z values: Most importantly, the p and z values for the province in determining if a voter would choose the Conservative party all passed the requirements, making province a significant predictor on whether respondents will vote for the Conservative party or Liberal party. For Bloc Quebecois, Quebec was the only province that passed requirements and proved to be a significant predictor. For NDP, Green Party and People's Party only around half of the possible provinces proved to be significant and thus we can conclude when determining if voter would select one of these parties, province was a moderately strong predictor.

Education Estimates:The conditional group used for education is "bachelor's degree". If we ranked parties on most to least likely to vote for respondents who have not completed high school the list produce would be as follows, People's Party, Bloc Quebecois, Conservative Party, NDP, Green Party, Liberal Party. We can see this trend and something very similar for all education levels, as the higher the respondent's education level, the more likely they are to vote for parties on the end of the list provided. However, for education levels above the bachelor level this trend is not exactly the same although these estimate values are very small and do not seem to be as significant.
P and z-values: Through studying the p and z values of education levels above bachelor's degree, we can see for the majority of parties this option appears to be insignificant as it does not pass the requirements. However, all other response options for each party satisfy the p and z value requirement, so in general we can label this predictor variable as significant in terms of determining voter's preferences.

Children Estimates: for our children variable, "No", people who don't have children will be the conditional group. If we ranked parties from most to least likely to vote for based on if they have children the list would be Bloc Quebecois, People's Party, Conservative Party, NDP, Liberal Party, and then Green Party. P and z values: As we can see from the p-value and z-value for People's Party, Conservative Party, and Bloc Quebecois, having children is a significant factor in determining if a person will vote for one of these options. On the other hand, as seen from the p and z values, having children is not a significant factor in determining if a voter will select either the NDP or the Green Party.

Marital Status Estimates:For marital status, the conditional response will be "Divorced". Out of all response for marital status,we see that being married has the greatest significance on voting preferences. If we ranked parties from most to least likely to vote for based on if respondents are married the list would be Conservative Party, People's Party, Liberal Party, Bloc Quebecois, NDP and then Green Party. P and z values: For the strong majority of responses besides "Married", we see that marital status was not a strong predictor as it failed the z and p value requirements.

Although many options of predictor values did not pass our z and p value criteria, most importantly when distinguishing whether on not a respondent will vote Conservative or Liberal the strong majority of predictor variables proved to be significant. This is important as the strong majority of votes will either be cast toward the Liberal or the Conservative party and we need to be able to accurately predict the ratio between these parties.

4.2 election results with all voters

Table 2: Popular Vote Predictor for each province

	Liberal Party	Conservative Party	Bloc Quebecois	NDP	Green Party
Alberta	0	100.0	0	0.0	0.0
British Columbia	36	51.6	0	12.3	0.0
Manitoba	12	82.0	0	6.1	0.0
New Brunswick	69	29.3	0	0.0	2.0
Newfoundland and Labrador	94	1.8	0	4.0	0.0
Nova Scotia	96	3.2	0	0.3	0.0
Ontario	63	32.9	0	3.7	0.0
Prince Edward Island	88	0.9	0	0.0	11.4
Quebec	68	2.5	30	0.0	0.0
Saskatchewan	0	92.5	0	7.5	0.0
Total Popular Vote	54	36.2	4	5.5	0.5

Table 3: Seats Won by each Party Prediction Broken Down By Province

	Liberal Party	Conservative Party	Bloc Quebecois	NDP	Green Party	Total Seats
Alberta	0	34	0	0	0	34
British Columbia	15	22	0	5	0	42
Manitoba	2	11	0	1	0	14
New Brunswick	7	3	0	0	0	10
Newfoundland and Labrador	7	0	0	0	0	7
Nova Scotia	11	0	0	0	0	11
Ontario	75	41	0	5	0	121
Prince Edward Island	4	0	0	0	0	4
Quebec	53	2	23	0	0	78
Saskatchewan	0	13	0	1	0	14
Total Seats	174	126	23	12	0	335

Table 4: Popular Vote prediction distribution broken down by predictor variables

	Liberal Party	Conservative Party	Bloc Quebecois	NDP	Green Party
Age					
18 to 29	52	20.1	0	24.3	3.6
30 to 39	63.2	33	0.1	3.2	0.5
40 to 49	57.2	40.3	1	1.3	0.2
50 to 59	48.2	45.7	6	0.1	0
60+	52.9	36.4	10.7	0	0
Gender					
Male	45.4	46.5	6.2	1.4	0.5
Female	61.6	27.7	4.8	5.3	0.5
Education					
less than high school diploma	31.8	42.3	21.9	3.5	0.4
Completed secondary/ high school	45	42.3	5.7	6.1	0.9
Completed college or equivalent	46.1	45.3	3.9	4	0.7
Bachelor's degree	74.8	23.6	0.3	1.4	0
Univeristy degree above Bachelor's degree	85.9	13.5	0.2	0.4	0
Children					
Has Children	50	42.4	6.9	1	0.2
Does not have Children	65.8	21.2	2.1	9.7	1.2
Marital Status					
Never Married	64.8	17.3	4.1	12.8	1
Living with a partner	59.7	21.8	10.9	5.2	2.3
Serparated	71.6	18.3	6.8	2.4	0.8
Married	45.1	51.4	3.3	0.2	0
Divorced	65.8	27.1	6	0.8	0.3
Widowed	53.2	33.9	12.8	0.11	0

Using our regression model, we were able to predict voting selection of each of the 3071 individuals post-stratification cells and we were able to come to the following results. As seen in table 2, our model predicts that the Liberal Party will win the majority of the popular vote at 54.2%, followed by Conservatives with 36.3%. The NDP placed third with 5.5% of votes, followed closely by Bloc Quebecois with 3.5%, Green party with 0.5% and the Peoples Party did not obtain any votes.

As mentioned previously, the Conservative Party and the Liberal Party were the only political groups that had a shot at winning the election and our prediction had Justin Trudeau and the Liberal Party winning the popular vote by a very large margin.

However, as previously stated, the election is not won by popular vote and the geographic distribution of votes are very significant. Although we cannot separate votes into specific electoral districts to determine the results of each seat, we can distribute votes by province and calculate the number of seats for each province as a percentage of the popular vote. As seen in table 3, the Conservative Party dominated the west coast capturing the majority of seats for British Columbia, Alberta, Saskatchewan and Manitoba by a large margin. Also similar to the election, the Liberal Party dominated the East coast capturing the majority of votes for the remaining 6 provinces. As expected, the Bloc Quebecois only received votes in Quebec, capturing a total of about 30% of votes or 23 seats. We can see for the NDP, the majority of their votes came from both British Columbia and Ontario, where we predicted they would obtain 5 seats each. The Green party only obtained votes in only Prince Edward Island and Newfoundland and Labrador, however these 2 provinces only combined for a total of 11 seats, and the votes towards the green party in these provinces were not enough to obtain any seats. Finally the People's Party won zero seats and did not obtain any votes as our model did not predict that any of our cells would produce this party as their voting choice.

Distribution of votes between our key demographic variables as seen in table 4 will now be analyzed. Through looking at the distribution of votes between ages, we can see Parties such as the NDP and Green Party preforming well against voters ages 18 to 29, but having a great struggle in obtaining votes from individuals over the age of 40. For the Bloc Quebecois this was the opposite as they received almost zero votes from individuals under the age of 40 and they recieved almost 10 percent of all votes from respondents over the age of 60. This resulted in the Bloc Quebecois obtaining more overall votes than the NDP and the Green party combined, as the distribution of ages was largely skewed so that voters ages 60+ outnumbered voters under the age of 40. When comparing our main two parties, the Liberal Party and the Conservative party, we see that the Liberal Party out preformed the Conservatives in every age group, with the largest spread among the younger populous were the Conservatives struggled to win votes.

Through looking at the distribution of votes among gender, most notably we see that the Liberal party performed much better among females, earning 61.6% of total female votes where in comparison the Conservatives struggled and only obtained 27%. However when we look the distribution of votes among males, the conservatives actually outperformed the Liberals gaining 46.5% of male votes in comparison to their 45.4%. In addition, in our dataset females outweighed males 53% to 47%, which could prove why the Liberals had such a large portion of the popular vote.

When we analyze the distribution of education among voters we see a very significant trend where voters at a college or lower level of education were equally drawn to vote for Conservatives or Liberals, with exception to less than a high school diploma where conservatives obtained around 10% votes than the Liberals. However when we looking at the voting preferences of Canadians with a Bachelors degree or better, we see a significant change in voting preferences as Liberal votes outweigh conservative votes more than 3 to 1.

Through analyzing voting preferences of individuals with and without children we can see that, although the distribution of voters with children was only won by liberals by an 8% margin, 61.2% of votes from people without children went to the Liberal party where as only 21.5% went to the Conservatives.

Finally for marital status, we see that the Conservatives obtained the majority of votes from individuals who were married at 51.4% where and Liberals only obtained 45.1% of married peoples votes. However for every other category, Liberals dominated the conservatives by more than 20% for each response.

Through analyzing total number of cells, Liberal party obtained 1771, Conservative Party had 1027, NDP had 141,Bloc Quebecois had 88 and Green Party had 42. When looking at the larger parties, we noticed that the cells with the highest populations seemed to be evenly distributed among the Liberal Party and Conservative Party with a slight edge to the Liberals, however cells with lower populations appeared to dominated by the Liberal Party in comparison to Conservatives. This suggests that the smaller more unique groups may be the driving factor for the Liberal Party's success and may in fact provide the Liberals with enough votes to hold a majority government. On the other hand, smaller parties such as the NDP,Bloc Quebecois, Green

Party and the People's Party all severely unperformed in comparison to their real totals in 2019. Through analyzing cells with high populations, these parties all performed poorly as the majority were occupied by Liberals and the Conservatives. This may be the reason why these parties performed so poorly as many of their votes from voters were part of large and demographically similar groups were absorbed by the larger parties. Reasoning for why thing happened will be further analyzed in the discussion section below.

5 Discussion

Preamble

Forecasting election results is extremely difficult as there are so many factors that influence an individuals voting preference. In our analysis we utilized key demographic variables to determining voting preferences, however there are many subtle characteristics that affect someone's political preference which cannot be accounted for when performing simple analyses on limited survey data. In addition, since Canada determines the winners of its elections by total number of electoral seats held, it is almost impossible to determine a the accurate results of an election merely based off of popular vote using purely statistical methods. Regardless of this, we have come to a prediction using multi-level regression on how the 2019 Canadian Federal Election would change if voter turnout was assumed to be 100%.

5.1 Regression Model Discussion

Applying our regression model to our original dataset provided us with key insights of how demographic characteristics have a large impact on voter preferences in the 2019 Federal Election. We found that out of all the demographic variables used in our model, a voters age and the province they are living held the greatest significance in determining voter preference.

We noticed that out of all age groups, the Liberal Party, NDP and the Green Party performed best among younger voters and the Conservative and Bloc Quebecois performed better in the older demographic.

Firstly, Jagmeet Singh and the NDP have proposed education plans that would make college and university tuition free for. To do so they have said they will initially start by working with provinces and territories to reduce prices of tuition. They also aim to eliminate federal interest rates on outstanding student loans and increase grants for Canadian students. This has resulted in many young voters casting their vote for the NDP as our model predicts almost 25% of the voters aged 18 to 29 will vote for this party. We also see this trend among Green Party voters as the majority of their votes came from voters aged 18 to 29. The Green Party's would as well like to abolish college and University tuition which is a very attractive promise for young voters. Although liberals performed well against all age ranges they did practically well among the millennial voters. This however is not surprising as millennials were the Liberals target audience. One major reason people vote for the Liberal or Conservative party which has been researched extensively, is so that the other party will not win. A plausible theory for explaining the success of the liberal party among young voters is that although the benefits the Liberals proposed weren't as targeted towards the youth as the NDP or the Green Party, many young voters chose to vote Liberal because they made for more of an effort to accommodate the youth than the Conservative Party. For example, the Liberal Party pledged to raise minimum wage to $15, where as the Conservative Party had no plans to raise the minimum wage. Many voters feel that casting a vote towards a smaller party is a waste of a vote as the only Parties that have a rel chance at winning the election are the Liberal and the Conservative Party. They could have mainly voted for the Liberal Party, just so that the Conservative Party would not win. The Conservatives comparatively performed better in older demographics as their political policies were gears towards middle class families. They accomplish this by make policy recommendations such as increasing the age tax credit. This method has proved effective for them in the past as voter turnout for older people is much higher, however in this 100% voter turnout scenario it is not as effective in capturing a win.

As previously stated Conservatives dominated the west coast winning British Columbia, Alberta, Saskatchewan and Manitoba by a large margin and the Liberals won all other provinces. However when examining the distribution of electoral seats across these regions, we see that being the most populated regions in Canada, Ontario and Quebec hold the largest amount of seats with 124 and 76 respectively. The fact that Liberally performed particularly well in these provinces and the Conservatives performed very poorly was a large factor in producing the Liberals large victory. If the Conservatives want a chance at winning at the very least a minority government they must perform better in these 2 provinces.

When we analyze the distribution of education among voters we see a very significant trend where voters at a college or lower level of education were equally drawn to vote for Conservatives or Liberals, with exception to less than a high school diploma where conservatives obtained around 10% votes than the liberals. However when we looking at the voting preferences of Canadians with a bachelors degree or better, we see a significant change in voting preferences as liberal votes outweigh conservative votes more than 3 to 1. These results are most likely due to the fact that the education standard in provinces that the conservatives perform well against, such as Alberta,Saskatchewan and Manitoba, are much lower than provinces where the Liberals perform well in such as Ontario.

Finally, through analyzing voting preferences of voters with and without children we can see that, although the distribution of voter with children was only won by liberals by an 8% margin, 61.2% of votes from people without children went to the liberal party where only 21.5% went to the conservatives. For marital status, we see that for voter who are married, Conservatives took the majority of their votes at 51.4% where as Liberals only obtained 45.1% of their votes. However for every other category, Liberals dominated the conservatives by more than 20% for each response. These results make sense as voters get older they are more likely to get married or have a child and the Conservatives perform well against the aging population in compariosn to younger voters.

5.2 Importance of voter turnout

In the end, our prediction suggests that Justin Trudeau and the Liberal Party would have the popular vote by a very large margin capturing 54% of the popular vote if voter participation was as assumed to be 100%. Based on the distribution of his votes across the 10 provinces, we estimated that he captured a total of 174 seats, which would be enough to sustain a majority to government. Even through factoring in possible error that our model could have sustained, we can see that at the very least Justin Trudeau and the Liberal Party would have sustain a minor government as they did in 2019. One of the main driving factor in this result is that voter turnout among the youth is very low in comparison to old age demographic. This results in older demographic having a larger impact on voting outcomes, which is beneficial to the Conservative party as previously description. However with full voter-turnout assumed, we see that the Liberal Party drastically increase their lead against the conservative party as voters aged 18 to 29 were more than twice as likely to vote for the Liberal party than the Conservatives.

These results are very significant, as unlike the 2019 Canadian federal election results, we predicted that the Liberal Party would obtain a majority government which is very different from the minority government they sustained previously. The power that a majority government holds in comparison to a minority government its drastically higher. When a minority government is won in Canadian Politics, the winning party has one of two choice. First opting for another election were they hope to win a majority of seats or form a coalition with another party in order to sustained the minimum of 50% of electoral seats. The difference with a majority government is they don't need to form a block with another party and can thus purse policy options which they value and not compromise for any other party. A majority government can also make quick decisions on pressing matters where as when there is a minority government issues are usually encompass long drawn out negotiations. Thus we can see how important voter turnout is in Canada as it has a significant effect on election results which could have resulted in a very different form of government in 2019.

However, there are many factors which could have lead to inaccuracy in our forecasts. The distribution of our predictor variables and the heavily outweighed groups we see can cause problems with our models as errors in the estimates of highly populated groups will create a much larger shift in the number of votes each

Party gets. Furthermore, the underrepresented groups can be problematic as our regression model has less data to work with here, leaving us with less overall confidence in our predictions for these groups.

5.3 How to increase voter turnout

As described previously, voter participation has a large effect on election outcomes, and is important in order to accurate represented the distribution of political views of the population of a country. However, voter turnout rates have remained relatively low in Canada as of recent fluctuating between 58 and 68 percent. One possible remedy to this predicament is to make voting mandatory for all eligible Canadians. One country that has had great success in implementing this strategy is Australia. In Australia first time offenders are penalized a small fee of 20 dollar and second time offender are issued a 50 dollar fine. Although these penalties do not seem large, since Australia made voting mandatory in 1924, the voter participation rate has been over 90% with exception to 2014 where voter turnout was a still very impressive 88.5%. If Canada can acknowledge the great success Australia has had from implementing this policy, they could too have very high voter turnout rates, which will help display more accurately the political views and opinions of Canadians.

5.4 Weaknesses and Future Work

The most significant limitation we encountered in our study was the lack of similar explanatory variables in the CES and GSS datasets. Through only using information on a respondents age, gender, province they are currently living in, highest education level obtained, whether or not they children and their marital status, we only did not have much data in which we could analyze a respondents preferences. In addition, the GSS lacked data on questions regarding political views or opinions on certain policies. The responses from these question could link respondents to a party that shares those same views more accurately then our model describes. Votes are not determined merely by observational characteristics as a they have a lot to do with an individuals opinions and views.

While this study provides key insights on what observational characteristics effect an individuals voting preference and how important voter turnout is, there is still many improvements that can be made and future. As previously mentioned, to obtained more accurate results, additional variables should be added to our model. Factors such as an individuals opinions on certain policies, income level, political views, current place of employment, and more geographical specific variables would make our model more reliable and in turn, increase the accuracy of predicting a persons voting preferences. Adding additional variables would also increase the number of post-stratified cells that would be produced. This is important as smaller parties such as the Green Party, NDP, Bloc Quebecois and the People's Party were underrepresented in cells that had a significant weight, resulting in an under representation of these parties in our study. In addition, as previously displayed our model was only capable of predicting 44% of respondents voter preferences. Through adding more variables as mentioned above, we would be able to increase this number and thus the accuracy of our results.

Furthermore, The General Social Survey data is too small to be an effective post-stratification dataset. Although it was more representative of the Canadian population then the CES dataset, it was smaller which very counteractive to have a training dataset that is larger than your post-stratification dataset. In addition, since Canadian election results are determined by the number of electoral seats won and not the popular vote, including information on which specific regions a person lives and which seat their vote is going towards, would allow us to predict which party is likely to win each seat. This would result in a much more realistic idea of the winner and what form of government they obtained.

Appendix

Code for this study can be found at:

References

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

CBC Radio. 2019. "In Australia, voting is mandatory, easy and often fun. Is there a lesson for Canada?". CBC. https://www.cbc.ca/radio/day6/mandatory-voting-canada-s-weediversary-fighting-alongside-the-kurds-atwood-archives-dolly-parton-more-1.5324795/in-australia-voting-is-mandatory-easy-and-often-fun-is-there-a-lesson-for-canada-1.5324822

datasciencebeginners. 2020. "Multinomial logistic regression With R". R-Bloggers. https://www.r-bloggers.com/2020/05/multinomial-logistic-regression-with-r/

Government of Canada, S. C. 2017. The General Social Survey: An Overview. Government of Canada, Statistics Canada. https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm

Larmarange, Joseph. 2020. Labelled: Manipulating Labelled Data. http://larmarange.github.io/labelled/.

Long, Jacob A. 2020. Jtools: Analysis and Presentation of Social Scientific Data. https://cran.r-project.org/package=jtools.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Statistics Canada. 2019. "Census Profile, 2016 Census Canada". from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/Page.cfm?Lang=E

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

Venables, W. N.; Ripley, B. D. 2002. Modern Applied Statistics with S. Fourth Edition. Springer,New York. ISBN 0-387-95457-0

Wickham, Hadley; Averick, Mara; Bryan, Jennifer; Chang, Winston; D'Agostino McGowan, Lucy; François, Romain; Grolemund,Garrett; et al. 2019. "Welcome to the tidyverse." Journal of Open Source Software 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley; Miller, Evan. 2020. Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files.

Zaiontz, Charles. 2014. "Basic concepts of multinomial logistic regression". Real Statistics Using Excel. https://www.real-statistics.com/multinomial-ordinal-logistic-regression/basic-concepts-of-multinomial-logistic-regression-basic-concept/

Zhu, Hoa. 2020. kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.1 https://CRAN.R-project.org/package=kableExtra