Informative Title Name

STA304 - Assignment 3

GROUP NUMBER: ADD YOUR NAMES HERE

November 5, 2021

Introduction

The Canadian Election Study is an annual survey of voting and other preferences and demographics which are thought to pertain to political behavior of Canadian voters.This survey is used to predict the overall popular vote of the next Canadian federal election (tentatively 2025) using a regression model with post-stratification.

Some of the important terminologies is the Majority government, which is the party that won the most number of votes and were elected to become the ruling government. Next, is the Minority party, it's the party with the least votes and they are elected to become the opposition to the majority party. Also, some of the parties mentioned below are Bloc Quebecois, Conservatives, Greens, Liberal, NDP, and People's Party.

The research question we would be investigating is that does age, religion and location enough to predict the number of votes for each party?

In this paper we will use post-stratification and multilevel regression to investigate if certain features of the population population such as: age, religion and location can be used to predict how many votes each party gets in a federal election

Data

Data Collection Process

The Data is a collection of two datasets. The first dataset contains data from the General Social Survey on Family (cycle 31) on 2017. Canada's General Social Survey (GSS) program conducts annual survey covering one topic in depth (citation). As such, this dataset contains mostly contains information pertaining to families. However, we will later investigate some common variables between this dataset and the Canada Election Study (CES) dataset. This will give us a list of factors in the general population that could be potentially associated to political affiliations

the Canada Election Study (CES) data which was collected in 2019. This data was collected from a questionnaire delivered to the people living in Canada through the Computer Assisted Telephone Interviewing(CATI). Phone calls were made potential interviewees during both the day and evening for every weekday. These questions asked some personal information such as their age and also political opinions such as "what is your opinion on Justin Trudeau?" (citation)

Data cleaning

The regression model to be constructed, will be applied to both datasets, therefore only variables that appeared in both the datasets could be used. One way was examine if there were common terms present in the columns for both the dataset. Unfortunately most the columns the CES dataset were just question numbers. However, each column also had a label that stated the question itself. For example column "q2" had the label "In what year were you born?". Hence we collected common words that appeared in the columns of the GSS data and the labels of the CES data to find possible topics that were common in both datasets.

Then for each topic we search the columns for both the dataset to see if any two column were describing identical or similar variables. These are the variables we found:

- Age: The age of the person was recorded in both data with the same column names.
- Sex: Gender was recorded in CES data and sex was recorded in the GSS data. Here we assumed the people who reported their gender to be male or female also would have the same sex as their gender. Hence we removed all genders that were not male or female and renamed the q3 column in the CES data to sex.
- Province: Both the datasets recorded the province a person lived in.
- religious importance: Both datasets recorded how the level of importance for religion. The categorical values in both the datasets were the same with the exception of an additional value "Refused" being present in the CES data. We removed rows containing this value from the CES data.
- Aboriginal: Both datasets recorded if a person was aboriginal or not. This was an option for question 66a in CES data which asked which ethnicity the person belonged to. q66a_15 recorded if the person belonged to an Aboriginal group. the value "(1) Selected" indicated the person identified as original and "(0) Not Selected". We filtered out other values as the indicated that the question was skipped or if the person was not sure. Then we mutated "(1) Selected" to "Yes" and "(0) Not Selected" to "No" to match the values of the variable in both dataset. Both datasets had values "Don't know" to indicate that a person was not sure. However, there very few rows with this value so we removed these rows.
- Education Level: Both datasets recorded the highest level of education completed by a person. However, these categories had slightly different names. For example, "University certificate, diploma or degree above the bach.." indicated that a person has a qualification above a bachelor's degree. In the CES data there are Master's degree as well as other degree above bachelor's degree. Hence we put the education level in both datasets into three groups: "Above Bachelor", "Below Bachelor" and "Bachelor".
- Household size: Both the datasets recorded household size.

After finding the common variables, we removed rows with missing values in these columns for both the datasets. Then we renamed the columns of the CES data to match the names of the columns of the GSS data. This was done so that models constructed on the CES data could also be used on the GSS data for post-stratification. Then we took the common variables and constructed two datasets. One dataset containing the rows from the CES data and the other dataset containing the rows of the GSS data. Futhermore, in the first dataset, voting data was added to the CES variables. This is the outcome of interest

Table 1: Variable description table

Variable Name	Description
province	Sex of the person
education	Province that a person is currently residing in
religion_importance	Level of education divided into three categories: 'Above Bachelor', 'Below Bachelor' and 'Bachelor'
aboriginal	How important is their religion to a person: 'Not important at all', 'Somewhat important', 'Very important', 'Not very important' 'Don't know'
hh_size	Is the person aboriginal
age	How many people does a person share their living space with including themselves
sex	Age of the person in years
vote	Party that the person will/might/has vote(d) for

Numerical Summaries

Table 2: Frequency for each category in the CES data

		Count
Sex	Female	895
	Male	1057
Province	Alberta	152
	British Columbia	296
	Manitoba	142
	New Brunswick	102
	Newfoundland and Labrador	114
	Nova Scotia	106
	Ontario	436
	Prince Edward Island	102
	Quebec	367
	Saskatchewan	135
Education	Above Bachelor	304
	Bachelor	534
	Below Bachelor	1114
Religion Importance	Don't know	4
	Not at all important	184
	Not very important	355
	Somewhat important	742
	Very important	667
Aboriginal	No	1889
	Yes	63
vote	Bloc Quebecois	74
	Conservatives	764
	Greens	163
	Liberal	673
	NDP	245
	People's Party	33

From the above table we can gain the following insights on the survey data:

- There seems to slightly more males and females in this dataset.
- Most of the people in the survey have an education level below a bachelor's degree.
- Most of the people in give a lot of importance or at least some importance to their religion.
- The large majority of the people in the survey are not Aboriginal. This perhaps suggests that this variable might not be of use for the model.
- It seems that the most popular party was the Conservative party followed closely by the Liberal party in 2019
- There does not seem to be an significant difference in the number of answered the survey for each province. This indicates that this survey represents the data of each province fairly well.

Table 3: Frequency for each category in the GSS data

		Count
Sex	Female	8961
	Male	7391
Province	Alberta	1352
	British Columbia	1807
	Manitoba	974
	New Brunswick	1225
	Newfoundland and Labrador	1030
	Nova Scotia	1283
	Ontario	3859
	Prince Edward Island	645
	Quebec	3193
	Saskatchewan	984
Education	Above Bachelor	1249
	Bachelor	2787
	Below Bachelor	12316
Religion Importance	Don't know	167
	Not at all important	3335
	Not very important	2708
	Somewhat important	5003
	Very important	5139
Aboriginal	No	15572
	Yes	780

In the census data the differences observed in the categorical variables are similar to the differences observed in the survey data, with the exception of Sex. There are more females than males in the census data. This indicates that there is an underrepresentation of females in the data survey data.

Table 4: Comparing Age variable in both dataset

	mean age	minmum age	maximum age	standard deviation in years
From GSS dataset	52.52854	15	80	17.70734
From CES Dataset	53.71721	18	100	16.60615

Even though the mean age is similar in both datasets, the magnitude of the min, max and standard deviation values indicates that age has a wider distribution in GSS dataset.

Table 5: Comparing Age variable in both dataset

	25th qauntile household size	median household size	75th quantile household size	min household size	max household size
From GSS dataset	1	2	3	1	6
From CES Dataset	2	2	4	1	15

From this table, we can see that household size has similar distributions, however CES data seems to indicate that people in this dataset tend to have more household members. The max size 15 could be a potential outlier

Graphical Summaries

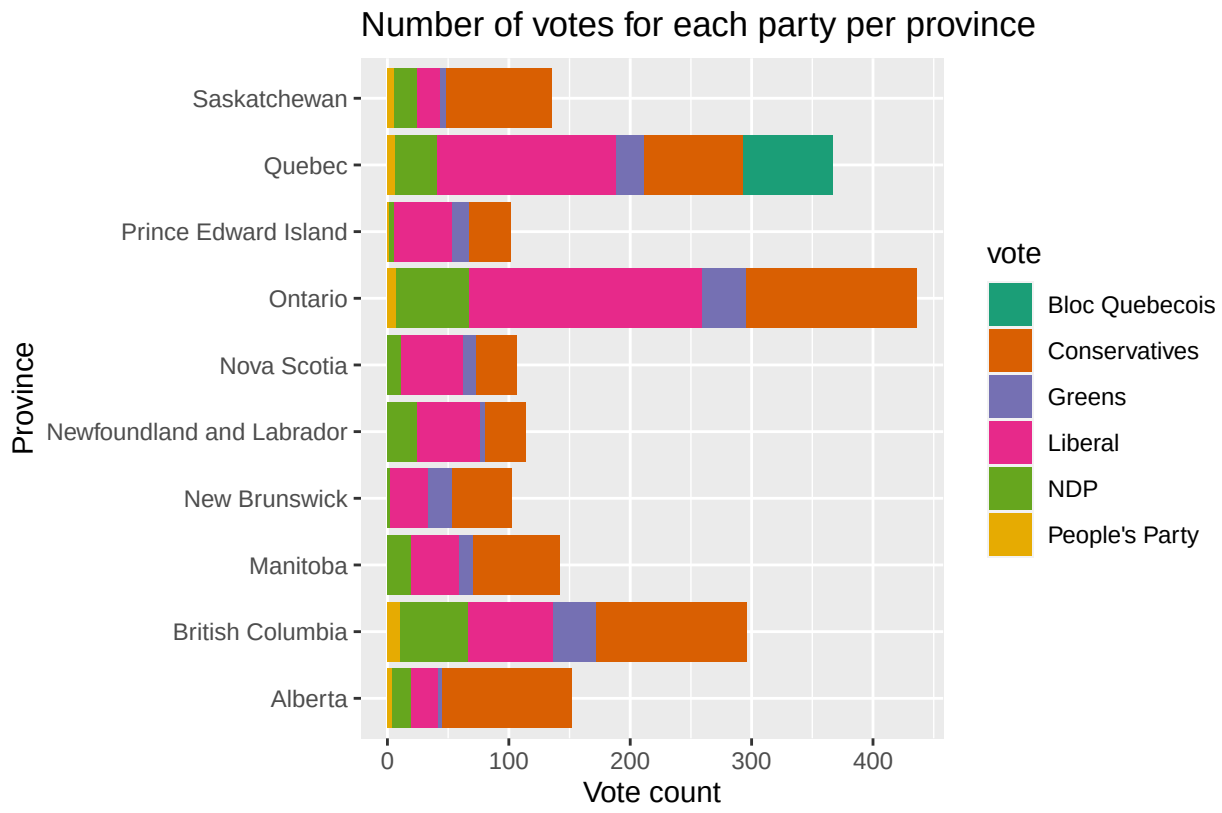

Figure 1

From this table, it looks that the Liberal party is the most popular party in Ontario and all the provinces to its left. Manitoba and all the provinces on its right favor the Conservatives most. It is perhaps somewhat surprising that Bloc Quebecois are not the most popular party in Quebec as it is a party that is most focused on Quebec.

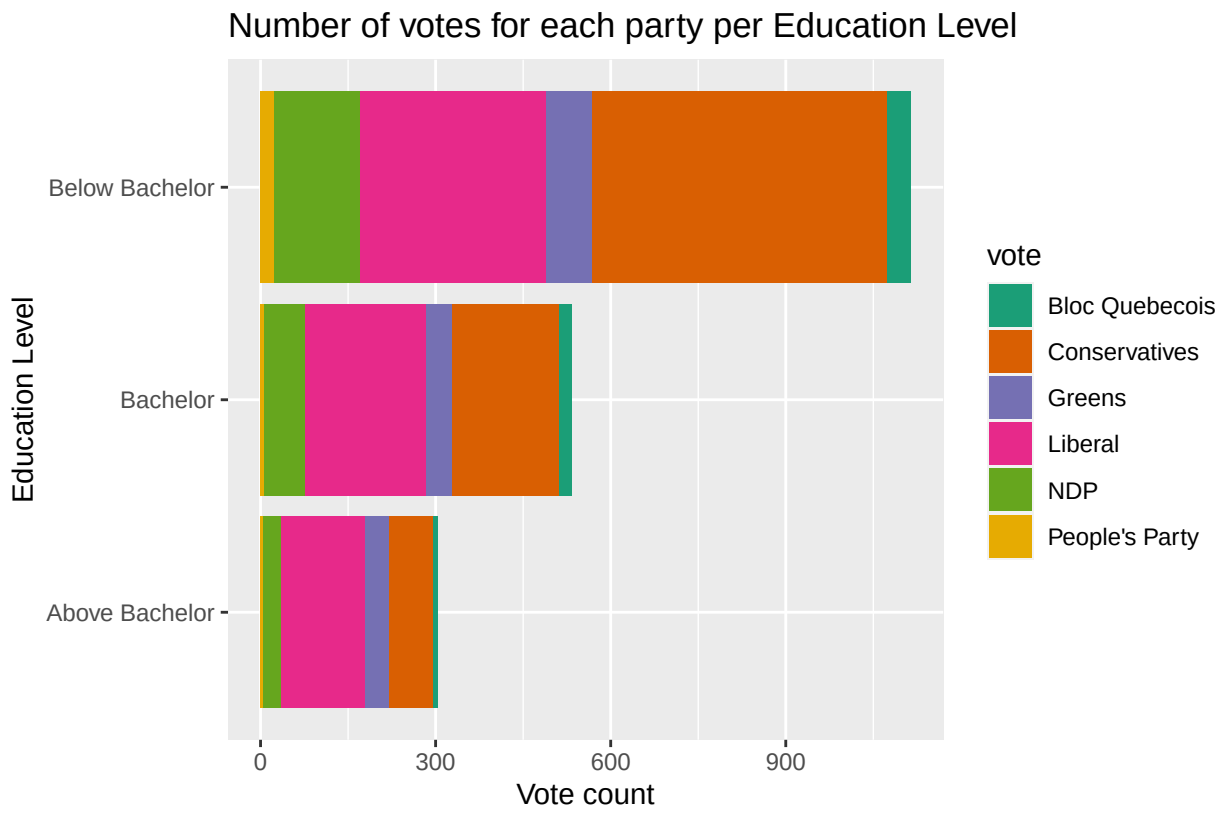

Figure 2

From this plot we can see, that the ratio of votes for each party does not change much across education levels. This seems to suggest that political affiliation is independent of education.

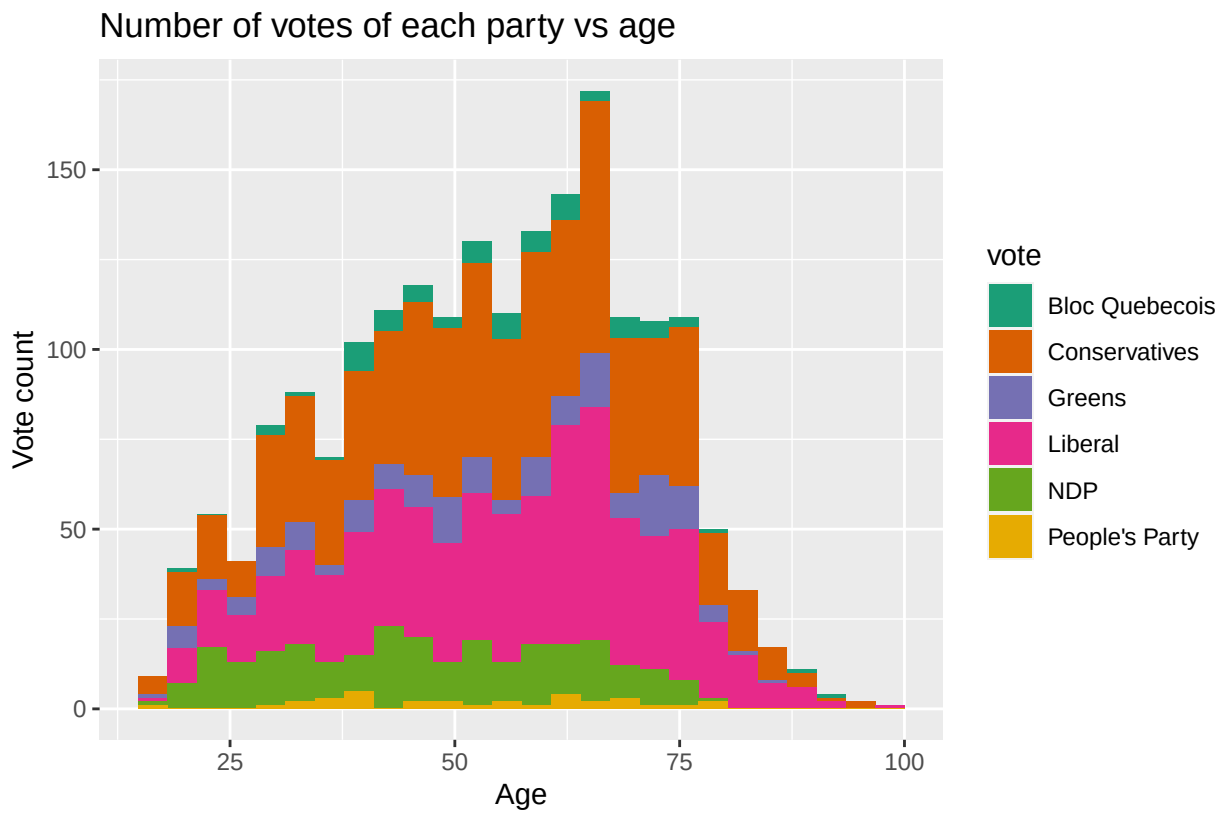

Figure 3

From this table we can see the number of votes with respect to age, follows a roughly normal distribution. With most of the votes coming from people aged between 50 and 75. There is not clear ratio differences in vote ratio across the ages ages upto 75. However, the votes for parties other than Liberal or Conservatives decrease significantly for age groups older than 75.

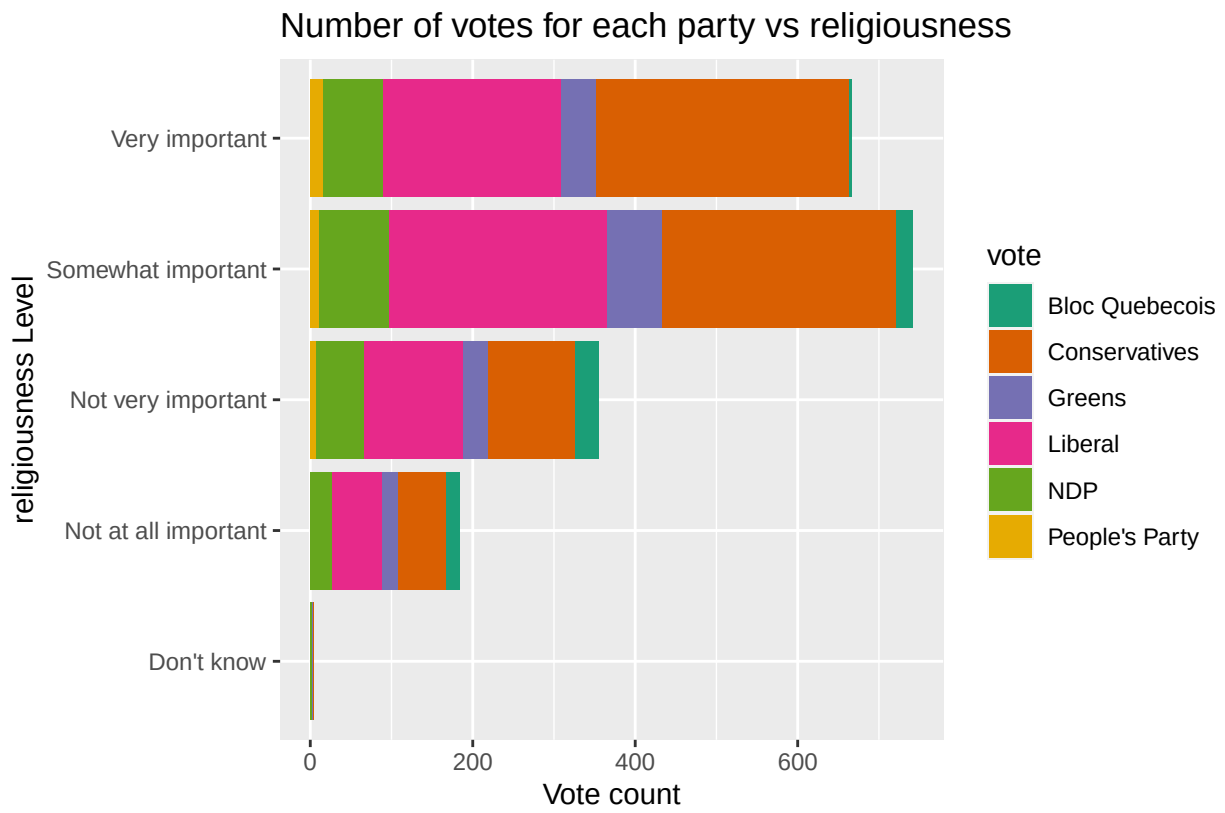

Figure 4

From this table, we again observe that the vote ratio between parties don't change across the groups. This suggests that political affiliation is independent of religion.

Methods

The question that we are trying to answer is whether factors such as sex, province, age, education, religion importance, aboriginal and household size help us predict someone's voting preference.

Model Specifics

We will be using multinomial logistic regression to build our model as our dependent variable, or our response variable is a nominal variable. Multinomial logistic regression is an extension of binary logistic regression, but instead of two categories of the dependent variable it allows for multiple. Therefore it is suitable for our model as we are predicting which party will tend to have the overall popular vote of the next Canadian federal election (tentatively 2025). The formula can be written as follows: Suppose our outcome of interest is Y which has the levels $y_1, \ldots, y_j$ and we want to find predict Y using the predictors $X_1, \ldots, X_k$. Then we can express the log odds of each level of Y as:

$$log\left(\frac{P(y_i)}{1-P(y_i)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon$$

where $log\left(\frac{P(y_i)}{1-P(y_i)}\right)$ is the log-odd of observing the outcome y_i. In this case y_i outcome that the ith party gets the vote. β_0 represents the default value of the log-odds. The coefficient B_l for $l \in \{1, 2, \ldots k\}$represent the change in log odds for one unit change in X_l assuming that all other predictor values being constant. ϵ is a random variable that has a standard normal distribution. This term represents the error term in our prediction.

To get a prediction from this model we can do the following process: Let $\eta = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$. Now we have that $log\left(\frac{P(y_i)}{1-P(y_i)}\right) = \eta$. Therefore we get that $P(y_i) = \frac{e^\eta}{1+e^\eta}$. we now have the probability of outcome y_i from predictors. Now we can set a threshold γ where if $P(y_i) \geq \gamma$ we predict 1 and 0 otherwise. Usually the γ is set to 0.5 which means 50% probability

Assumptions

- We assume choice of vote given to one party is not related to the choice of vote given to another party. This is know as assumption of independence among dependent variable choices
- We also assume that we cannot perfectly separate all the cells by outcome groups. Each group would the party in this case
- We also assume that there is no natural ordering present between the party, meaning that we treat each party as equal.

Model construstion and testing

To get the estimates of each variable, we will be using the iterative maximum log likelihood estimation procedure (citation needed). For variable selection, one of the methods we will use is backward selection using Akaike Information Criteria (AIC). The Akaike information criterion (AIC) is a mathematical method used for evaluating the best fit model. We use it by comparing multiple variables and their combination and select which combination is the best fit for the data. The AIC considers two things, the number of independent variables we used and the maximum likelihood estimate of the model. Therefore according to the AIC the best fit model would be one that considers the most variation and the least amount of variables. We have made use of the backward selection to select the right model. Backward selection works with all variables and keeps reducing one along the way to choose the best fit model. We have also took in consideration the accuracy of the model in our code, therefore choosing the model with the highest accuracy as well as lowest AIC.

Furthermore, we will also split the dataset into testing and training set and test the accuracy of each model on the testing set to further evaluate which model is more suitable

Post-Stratification

We will post stratification method in order to estimate the proportion of voters in each province.The reason post stratification is preferred is because stratified sample design increases the accuracy of estimates compared to a simple random sample especially for non-probability based sampling. We partition the data into thousands of demographic cells, in the case of our study these cells are grouped by provinces and then we estimate response variable for each cell using a multilevel regression model. The sum of this cell levels estimates population-level estimate, $\hat{y}^{PS}$, by weighting each cell by its relative proportion in the population. The formula for this is:

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

Where $\hat{y}_j$ is the estimate in each cell. And $N_j $ is the population size of the J^{th} cell based off demographics.

Results

The original multinomial logistic regression model (with all the variables) had an AIC of 3880.9002249 but has a lower training accuracy than another model which had an AIC of 4233.1022562. However on the test set, the full model had an test accuracy of 49.87 and the other model had an test accuracy of 43.99. Hence we picked the original model as the final model.

Table 6: Model output for each province

Party	term	estimate	std.error	statistic	p.value
Conservatives	(Intercept)	20.6139109	0.9665995	2.132622e+01	0.0000000
Conservatives	provinceBritish Columbia	0.4674381	0.2885618	1.619889e+00	0.1052561
Conservatives	provinceManitoba	2.6789875	0.3252041	8.237865e+00	0.0000000
Conservatives	provinceNew Brunswick	4.3635676	0.4085819	1.067979e+01	0.0000000
Conservatives	provinceNewfoundland and Labrador	2.2445164	0.3546746	6.328382e+00	0.0000000
Conservatives	provinceNova Scotia	3.3346023	0.3424493	9.737507e+00	0.0000000
Conservatives	provinceOntario	-0.8000263	0.2892250	-2.766103e+00	0.0056731
Conservatives	provincePrince Edward Island	3.6158454	0.3839634	9.417162e+00	0.0000000
Conservatives	provinceQuebec	-17.9532777	0.5163787	-3.476766e+01	0.0000000
Conservatives	provinceSaskatchewan	3.5759682	0.3326847	1.074882e+01	0.0000000
Conservatives	educationBachelor	-0.3008241	0.5578342	-5.392715e-01	0.5896995
Conservatives	educationBelow Bachelor	-0.0161024	0.5240459	-3.072710e-02	0.9754872
Conservatives	religion_importanceNot at all important	-1.9662360	0.8557698	-2.297623e+00	0.0215833
Conservatives	religion_importanceNot very important	-2.6053512	0.8472346	-3.075124e+00	0.0021042
Conservatives	religion_importanceSomewhat important	-1.6177037	0.8502688	-1.902579e+00	0.0570955
Conservatives	religion_importanceVery important	-0.6392330	0.9122755	-7.007017e-01	0.4834892
Conservatives	aboriginalYes	-0.7849667	0.9570745	-8.201730e-01	0.4121175
Conservatives	hh_size	0.2046693	0.1509814	1.355592e+00	0.1752289
Conservatives	age	-0.0239982	0.0117074	-2.049833e+00	0.0403807
Conservatives	sexMale	0.2512105	0.3257900	7.710810e-01	0.4406589
Greens	(Intercept)	19.0024599	1.2617218	1.506074e+01	0.0000000
Greens	provinceBritish Columbia	3.8353295	0.8307563	4.616672e+00	0.0000039
Greens	provinceManitoba	5.4612271	0.8026973	6.803595e+00	0.0000000
Greens	provinceNew Brunswick	8.0926672	0.8222940	9.841575e+00	0.0000000
Greens	provinceNewfoundland and Labrador	4.5542930	0.8765722	5.195571e+00	0.0000002
Greens	provinceNova Scotia	6.6117466	0.8048948	8.214423e+00	0.0000000
Greens	provinceOntario	2.3843226	0.8318575	2.866263e+00	0.0041535
Greens	provincePrince Edward Island	7.1187045	0.8832591	8.059588e+00	0.0000000
Greens	provinceQuebec	-14.9690848	0.9415334	-1.589862e+01	0.0000000
Greens	provinceSaskatchewan	4.9726079	0.9461748	5.255485e+00	0.0000001
Greens	educationBachelor	-0.9554639	0.5892960	-1.621365e+00	0.1049394
Greens	educationBelow Bachelor	-1.1979217	0.5532344	-2.165306e+00	0.0303643
Greens	religion_importanceNot at all important	-2.9457046	0.8858440	-3.325308e+00	0.0008832

Party	term	estimate	std.error	statistic	p.value
Greens	religion_importanceNot very important	-3.3145729	0.8603945	-3.852387e+00	0.0001170
Greens	religion_importanceSomewhat important	-2.4953807	0.8570283	-2.911666e+00	0.0035951
Greens	religion_importanceVery important	-2.1854731	0.9243030	-2.364455e+00	0.0180566
Greens	aboriginalYes	-0.5194670	1.0588072	-4.906153e-01	0.6236986
Greens	hh_size	0.1284141	0.1634337	7.857261e-01	0.4320279
Greens	age	-0.0343957	0.0128231	-2.682313e+00	0.0073115
Greens	sexMale	-0.1580697	0.3627589	-4.357430e-01	0.6630232
Liberal	(Intercept)	-3.0247458	0.5175957	-5.843839e+00	0.0000000
Liberal	provinceBritish Columbia	1.7701572	0.3468576	5.103412e+00	0.0000003
Liberal	provinceManitoba	3.9499426	0.3737032	1.056973e+01	0.0000000
Liberal	provinceNew Brunswick	5.8635121	0.4463751	1.313584e+01	0.0000000
Liberal	provinceNewfoundland and Labrador	4.5422903	0.3836305	1.184028e+01	0.0000000
Liberal	provinceNova Scotia	5.3583685	0.3746632	1.430183e+01	0.0000000
Liberal	provinceOntario	1.3318810	0.3395643	3.922323e+00	0.0000877
Liberal	provincePrince Edward Island	5.7896462	0.4193419	1.380650e+01	0.0000000
Liberal	provinceQuebec	-15.5643498	0.5401497	-2.881488e+01	0.0000000
Liberal	provinceSaskatchewan	4.0272458	0.4115902	9.784600e+00	0.0000000
Liberal	educationBachelor	-0.7828025	0.5390669	-1.452143e+00	0.1464617
Liberal	educationBelow Bachelor	-1.1243778	0.5091714	-2.208250e+00	0.0272268
Liberal	religion_importanceNot at all important	20.3823536	0.3089071	6.598214e+01	0.0000000
Liberal	religion_importanceNot very important	20.2080182	0.2897223	6.974961e+01	0.0000000
Liberal	religion_importanceSomewhat important	21.0684632	0.3076141	6.848992e+01	0.0000000
Liberal	religion_importanceVery important	21.7142162	0.4531525	4.791813e+01	0.0000000
Liberal	aboriginalYes	-0.6930453	0.9479903	-7.310679e-01	0.4647377
Liberal	hh_size	0.2130256	0.1494154	1.425728e+00	0.1539470
Liberal	age	-0.0169714	0.0115368	-1.471061e+00	0.1412745
Liberal	sexMale	-0.4069675	0.3193906	-1.274200e+00	0.2025926
NDP	(Intercept)	22.8250927	0.9294080	2.455874e+01	0.0000000
NDP	provinceBritish Columbia	1.9357194	0.3682696	5.256257e+00	0.0000001
NDP	provinceManitoba	3.4466938	0.4121908	8.361890e+00	0.0000000
NDP	provinceNew Brunswick	2.9592972	0.8317496	3.557918e+00	0.0003738
NDP	provinceNewfoundland and Labrador	3.8867646	0.4185329	9.286641e+00	0.0000000
NDP	provinceNova Scotia	3.9732970	0.4478481	8.871974e+00	0.0000000
NDP	provinceOntario	0.5631734	0.3679999	1.530363e+00	0.1259269
NDP	provincePrince Edward Island	3.6899947	0.5760213	6.406004e+00	0.0000000
NDP	provinceQuebec	-16.8723786	0.5644219	-2.989321e+01	0.0000000
NDP	provinceSaskatchewan	3.8914868	0.4450255	8.744412e+00	0.0000000
NDP	educationBachelor	-0.4891944	0.5863961	-8.342388e-01	0.4041464
NDP	educationBelow Bachelor	-0.3527141	0.5492779	-6.421414e-01	0.5207814

Party	term	estimate	std.error	statistic	p.value
NDP	religion_importanceNot at all important	-3.9446095	0.7713761	-5.113731e+00	0.0000003
NDP	religion_importanceNot very important	-4.1286356	0.7496681	-5.507285e+00	0.0000000
NDP	religion_importanceSomewhat important	-3.8002377	0.7545360	-5.036522e+00	0.0000005
NDP	religion_importanceVery important	-3.1044915	0.8249462	-3.763265e+00	0.0001677
NDP	aboriginalYes	-0.2372872	0.9775717	-2.427312e-01	0.8082136
NDP	hh_size	0.1920298	0.1555313	1.234669e+00	0.2169536
NDP	age	-0.0526179	0.0123110	-4.274064e+00	0.0000192
NDP	sexMale	-0.5904570	0.3454759	-1.709112e+00	0.0874302
People's Party	(Intercept)	10.7637961	0.9347068	1.151569e+01	0.0000000
People's Party	provinceBritish Columbia	1.3564706	0.5946633	2.281073e+00	0.0225441
People's Party	provinceManitoba	-8.3808396	0.0000797	-1.052045e+05	0.0000000
People's Party	provinceNew Brunswick	-12.1615881	0.0000002	-5.580362e+07	0.0000000
People's Party	provinceNewfoundland and Labrador	-9.3555574	0.0000249	-3.754618e+05	0.0000000
People's Party	provinceNova Scotia	-10.4372663	0.0000029	-3.573129e+06	0.0000000
People's Party	provinceOntario	-0.1660275	0.6130012	-2.708437e-01	0.7865112
People's Party	provincePrince Edward Island	3.5373764	0.9652617	3.664681e+00	0.0002476
People's Party	provinceQuebec	-17.0793129	0.7993737	-2.136587e+01	0.0000000
People's Party	provinceSaskatchewan	3.6624084	0.7116821	5.146130e+00	0.0000003
People's Party	educationBachelor	-0.6136605	0.9096604	-6.746040e-01	0.4999274
People's Party	educationBelow Bachelor	-0.2078916	0.8114277	-2.562048e-01	0.7977927
People's Party	religion_importanceNot at all important	-3.3909247	0.0001823	-1.859909e+04	0.0000000
People's Party	religion_importanceNot very important	5.5396618	0.5497286	1.007709e+01	0.0000000
People's Party	religion_importanceSomewhat important	6.1915027	0.5023836	1.232425e+01	0.0000000
People's Party	religion_importanceVery important	7.5907380	0.5803198	1.308027e+01	0.0000000
People's Party	aboriginalYes	0.8124896	1.1257954	7.217027e-01	0.4704773
People's Party	hh_size	0.0549632	0.2119359	2.593387e-01	0.7953739
People's Party	age	-0.0416927	0.0173870	-2.397928e+00	0.0164881
People's Party	sexMale	0.0461285	0.5173728	8.915910e-02	0.9289554

Here the reference level is the party "Bloc Quebecois". This means that all coefficient estimates values are relative to this party. Consider the estimated coefficient of age for example. This can be interpreted the following: If all other variables are held constant then increasing age by one unit decreases the log-odds of the person voting for Conservative by 0.01697 relative to the log odds of them voting for "Bloc Quebecois". The standard error is the estimated deviance of this this estimate and the other statistic is the Z-score for the estimate. This is the z-score under then null hypothesis that the real coefficient value is 0. The p.value is the probability of deriving the observed estimated value under the null hypothesis.

We then performed our post-stratification where we first found the total number of individuals that voted for each party in each province. After, we took the party which had the most number of votes in each province to complete our post-stratification.

Table 7: Post-Stratification Results showing Popular Vote by Province

Province	Predicted Party	Count
Quebec	Liberal	2017
Ontario	Liberal	2233
Alberta	Conservatives	1339
British Columbia	Conservatives	1270
Saskatchewan	Conservatives	964
Manitoba	Conservatives	835
Prince Edward Island	Liberal	437
New Brunswick	Conservatives	1017
Newfoundland and Labrador	Liberal	659
Nova Scotia	Liberal	708

The above table shows that among the ten provinces in the dataset, the Liberal Party is the most favored overall among individuals surveyed. The Liberal and Conservative Party are most favored in five provinces each. [https://www.datanovia.com/en/lessons/identify-and-remove-duplicate-data-in-r/]

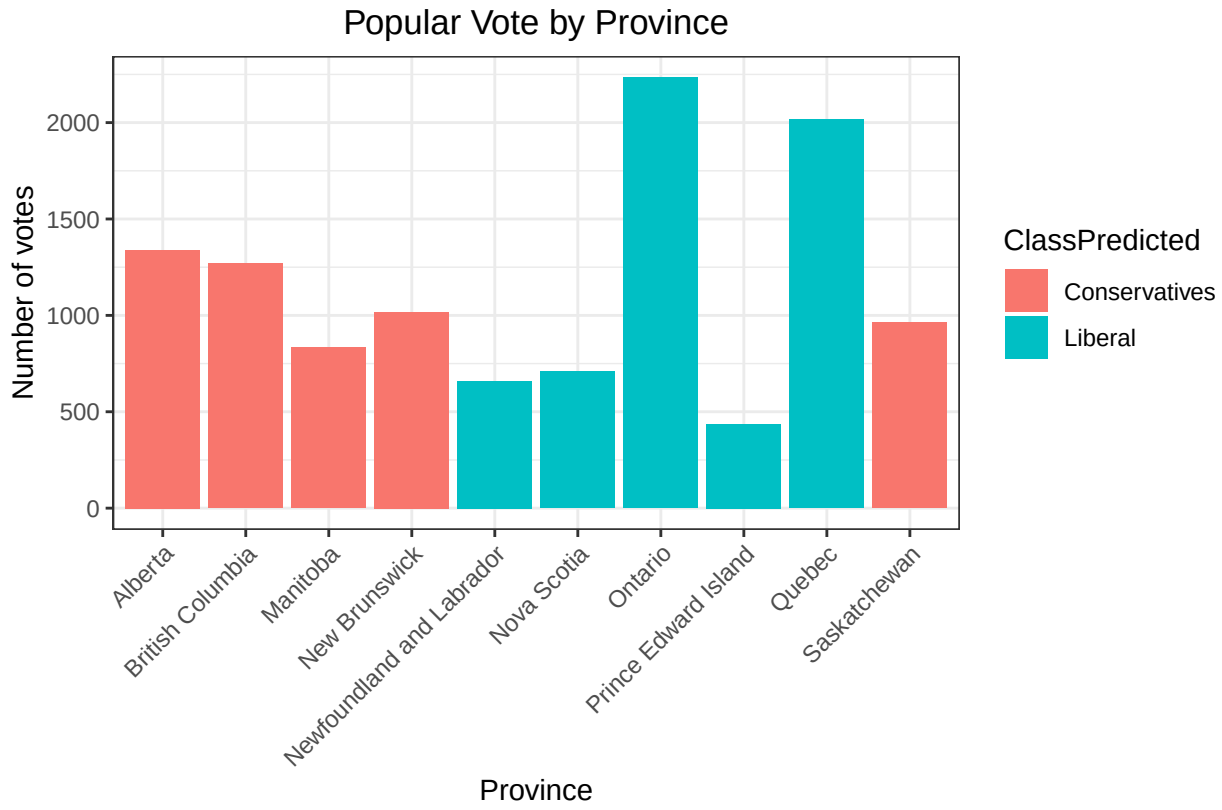

Figure 5

The bar graph above shows that the Liberal Party was predicted to win by most individuals in Ontario, Quebec, Nova Scotia, Newfoundland, and Prince Edward Island. The Conservative Party had the popular

vote in the remaining provinces. Ontario and Quebec are the two largest provinces in Canada and as seen in the graph have a much larger number people voting for the Liberal Party as compared to other Liberal and Conservative provinces. Hence, it makes sense that the Liberal party is predicted to win the election across all provinces in Canada. [https://www.biostars.org/p/464531/]

Conclusions

Our aim was to analyze whether factors such as sex, province, age, education, religion importance, aboriginal, and household size help us predict an individual's voting preference. We expected an overall preference towards the Liberal Party among individuals since this was the majority party in Canada in 2019. We built a multinomial logistic regression model. We divided our data into training and testing tests to measure the accuracy of our model in order to accurately predict the voting trend of the public. We used the backward selection method and considered the to select the best model.

We used post stratification method in order to estimate the proportion of voters in each province. We found that the Liberal Party was favored in five provinces and the Conservative party was favored in the other five. The Liberal Party was the most favored overall among all provinces in Canada.

A few weaknesses in our approach include:

- The AIC approach only provides a relative test of model quality. AIC would provide no indication if the statistical models used are equally a poor fit for the data [https://www.thoughtco.com/introduction-to-akaikes-information-criterion-1145956].
- When working with high dimensional datasets such as this one, there could be an over-fitting of the model on the training set. This may lead to inaccurate results in the model. We can use regularization techniques to avoid over-fitting, however, this will make the model much more complex [https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/].
- In the future, we can aim to conduct a post-stratification to estimate proportion of votes for other variables such as sex, education, or household size and analyze their voting preferences. This would help us broaden our analysis and draw up other conclusions on our dataset.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
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)