Analysis of the effect of voter turnout on the 2019 Canadian Federal Election

Jihong Huang 1004742610

18/12/2020

GitHub repo

Code and data supporting this analysis is available at:

https://github.com/KolbeHuang/STA304-Final-Project

Abstract

In this study, I used a survey dataset and a census dataset to build and apply my models, where the survey data was collected in the 2019 Canadian Election Study (CES) while the census data was collected in the General Social Survey (GSS). After cleaning the data, I built 6 logistic regression models on the survey data with justification and then employed the post-stratification technique on the census data to make prediction on the population-level proportion of voters for each party in the 2019 Canadian Election Study if "everyone" had voted. I found that the winner of 2019 Canadian Federal Election would be the Conservatives if "everyone" had voted, which meant that the turnout has influences on the election result. This conclusion might make future policy changes to make more reasonable election policies.

Keywords

Key words: Logistic Model, Post-stratification, Turnout, Proportion, Vote, Election, Everyone

Introduction

Nowadays, statistical analysis plays a very essential role in political and social study. As political and social researches investigate the relationships among people, an appropriate statistical analysis could help with collecting public information, forecasting elections, regulating economy and determining necessary policy changes. With statistics, it is easier for decision-makers to see the potential factors influencing the national policies and determine whether some changes are necessary. Usually, a model with specific assumptions based on current data is an important approach to finding these potential factors, by comparing the prediction and the facts.

One popular and effective modelling method is multilevel regression with post-stratification. Multilevel regression with post-stratification was first introduced in 1997 and got expanded later. Recently, it has been applied for the predictions on the national level by statistic scientists. In this report, multilevel regression with post-stratification is applied to predict the result of 2019 Canadian Federal Election based on the

assumption that everyone had voted and to investigate the potential influence of voter turnout on the election result.

Voter turnout refers to the percentage of qualified voters who cast a ballot in an election. In general, a high voter turnout is considered reflecting the will of people. Therefore, voter turnout is usually a significant factor for an election and a high voter turnout is often desired. However, in recent years, the voter turnouts in Canada's general elections were not high.[8] So, this is the motivation for the study that investigates the possible election result in 2019 if everyone had voted and the influence of voter turnout on 2019 Canadian Federal Election result.

For this study, a survey dataset and a census dataset are required to perform the multilevel regression with post-stratification and make predictions. In the Methodology section, I describe the way that the models are built based on survey data with multilevel regression and how they are used to make predictions with poststratification. The results of the multilevel regression with poststratification are provided in the Result section, and inferences of the results along with conclusions are presented in Conclusion section.

Methodology

Data

This analysis requires a survey dataset to build models and a census dataset to make predictions.

We obtained the census data from the online library of University of Toronto. The census dataset contains 20602 observations with 81 variables. The source of this census dataset is the General Social Survey in Statistics Canada.[1] To collect valid data, researchers applied the Computer Assisted Telephone Interviewing (CATI) technique to make phone calls to interview a random qualified member of a household. During the data processing, researchers extracted data from the linkage with agreement to guarantee precise answers to those questions that interviewees could not give a precise answer to. These collecting and processing techniques ensure the validity and reliability of this dataset. Based on the GSS guide, it is clear that the target population is all non-institutionalized Canadians that are 15 years old and older who are living in private households in the 10 provinces of Canada. The frame population is all Canadians with a telephone or a cell phone. The sampling frames are the lists of telephone numbers in use including landline and cellular that are available to Statistics Canada and the list of dwellings within the ten provinces from Address Register. Thus, the sampled population is all Canadians who are willing to answer a questionnaire from a phone call or telephone call. The sample in this study is the 20602 observations in the GSS dataset.

As mentioned above, researchers found the respondents by making random phone calls with the CATI technique. In this study, there were also non-respondents after reasonable attempts. To handle the non-responses for better overall representativeness of this study, researchers reweighted the responses.

The key features of this GSS study are the usage of the combination of CATI techniques and full-content reviews. For this study, one strength is that this study used detailed and standardized questions to collect precise and useful data. Another strength is the extraction of information from the linkage with agreement, which guaranteed the reliability of the data for the questions that interviewees could not give a clear answer to. However, this study also has weaknesses. Too many questions in the survey would make interviewees less willing to finish the questionnaire, which could potentially reduce the size of the data. Meanwhile, those questions concerning deep privacy could cause many missing values in the data, which possibly prevent further analyses. Finally, this study took 9 months and employed many people to realize the data collection and processing, which is a great cost.

In this dataset, I paid most attention to the variables that concern the age, sex, living province, education levels, the birthplace, importance of personal religion and the household income of each respondent (refer to the appendix). To make the dataset more solid, I removed all the observations with attributes that are "Don't know". Also, for the purpose of generality, I regrouped the education levels, importance of religion and ages so that each categorical variable has only two different categories in general. In this way, the

census data is more general and clearer for further analysis. Besides, some variables were excluded because they are very similar to the chosen ones or not correspondent to the survey data. For example, the variable "income_respondent" is very similar to the household income and it was not considered because there is no correspondence in the survey data.

With the visualization of the selected variables, half of the variables like education and sex have their categories evenly distributed, which means that there is no distinct difference among counts of different levels for each variable. For example, in the plot for sex in Figure 4 and 5, the proportion of male is very similar to the proportion of female. However, in the plot for religion_importance in Figure 4 and 5, we can see that the number of people who think religion is important is approximately twice as that of people who think religion is not important. Also, there are many more people from Ontario and British Columbia than people from Prince Edward Island.

Refer to Appendix Figure 4 and 5 for the whole plots for selected variables in census data.

We obtained the survey data from the cesR package.[2] The survey dataset contains 4021 observations with 278 variables. The source of this survey dataset is the 2019 Canadian Election Study.[3] There were two phases of data collection. During the election campaign, researchers interviewed 4021 Canadian citizens with telephone. After the election, 72% of previous respondents finished the interviews in the way according to their preferences. In both phases, all telephone data collection was completed with Computer Assisted Telephone Interviewing (CATI). In the second phase, all web data collection was completed on the Advanis proprietary platform. During the process of data, researchers weighted the data results to provide unbiased estimates for phone ownership type and province. Based on the phone survey technical report, the target population of this survey is all Canadian citizens who are 18 years of age or older and reside in one of the ten Canadian provinces excluding the territories. The frame population is all Canadians over 17 years old in one of the ten Canadian provinces excluding the territories who have landline or wireless telephone in the households. Ideally, the sampling frames for this study would be a complete list of all residential telephone numbers in Canada. However, the practical sampling frame is a modified form of random digit dialling (RDD) employed by Advanis. Thus, the sampled population is all Canadians over 17 years old in one of the ten Canadian provinces excluding the territories who are willing to answer a survey from a landline or wireless phone call in the household. The sample in this study is the 4021 observations in the cesR dataset.

To find the respondents, researchers obtained landline records from ASDE which is a sample provider and obtained wireless sample generated internally by Advanis. For the purpose of randomness, a modified random digit dialling (RDD) procedure was applied to select telephone numbers and the birthday selection method was applied to determine the respondent in the landline sample in households with more than one adult Canadian citizen. In this way, researchers found the respondents with randomness. In the first phase, an interviewee would be considered as "non-response" if no eligible respondent was reached after 6 attempts. Then, researchers would retire that non-response sample record and select a new sample record to call. Therefore, non-responses were directly excluded in the first phase. In the second phase, those non-responses (1132 in 2019 study) would have missing values for the post-election questions.

The key features of this study are that there are two phases of data collection divided by the election, where the second phase allowed the online or telephone completion. The features required more techniques and methods to guarantee the validity of collected data. One advantage of this study is the employment of a dual sample frame, which is widely used for general RDD sampling. As researchers were reaching interviewees by landline and wireless phone calls, there would be overlaps between the two frames. This dual sample frame would determine the overlap information and make correction for the biased selection probability of the overlap group during the process of weighting, making our data unbiased. Another advantage is that researchers considered the fact that people from different provinces had different likelihoods to get interviewed and then weighted the data results to correct for these unequal probabilities. However, this study also has its weaknesses concerning the missing values. In both phases, some questions are depending on other questions and some questions are province-specific. For example, only Quebec residents were asked questions relating to the Bloc Québécois. This dependency and specificity would lead to many missing values in the dataset. Also, not all respondents in the first phase would complete the survey in the second phase. The absence could cause many missing values in all questions in the second survey. These missing values in the dataset would make future analyses hard to realize because of the inadequate information.

In this dataset, I focused on the variables that concern the party to support, age, sex, living province, education levels, the birthplace, importance of personal religion and the household income of each respondent (refer to the appendix). To make the data more suitable for the purpose of analysis, I removed all the observations with attributes that respondents refused to answer or did not know. To simplify the predictor – party to support, I only focused on the 6 specific parties in the list to gain concrete predictions. To obtain the corresponding age groups like in the census data, I calculated the ages based on the variable "year_of_birth" and regroup the ages. Also, I regrouped the education levels, importance of religion and ages so that each categorical variable has only two different categories in general. In this way, the survey dataset is more general and clearer.

With the visualization of the selected variables, we can see that most variables have their own major category, which means that the counts of categories of each variable are not evenly distributed. For example, in the plot for education in Figure 6 and 7, around 80% interviewees in the survey dataset have attended university or college. Also, in the plot for vote_party in Figure 6 and 7, most people preferred liberal or conservatives, where only a small part of people support other parties.

Refer to Appendix Figure 6 and 7 for the whole plots for selected variables in survey data.

Model

In this study, the goal is to investigate the possible results of 2019 Canadian Federal Election if "everyone" had voted and the potential importance of turnout. Here, I define "everyone" to be every Canadian that is 18 years old or older and is eligible for voting. Also, I assume that in this study, every individual must vote for one of the six parties: Liberal, Conservatives, NDP, Green Party, Bloc Québécois and People's Party, which simplifies the modeling. The goal would be achieved by building six logistic regression models based on the survey data and applying a post-stratification technique for each model on the census data. In the following subsections, I would describe and explain the process of building and selecting the final models along with the application of the post-stratification technique.

I run my code in the Rstudio.

Model Specifies

After the data cleaning and regrouping where I removed the observations that did not support the 6 main parties, the response is a categorical variable with 6 categories corresponding to 6 different parties. Notice that I also directly removed the observations that responded "will spoil ballet" or "will not vote" because these two kinds of answer failed to satisfy the assumption that "everyone had voted". For example, if an individual in the survey data chose to support the Liberal Party in the questionnaire, the corresponding response of that individual will be "Liberal". Meanwhile, to avoid making the models over-complex and hard to interpret, I removed all the missing values and regrouped some categorical variables into fewer but more representative groups. Take the variable "education" as an example. Previously in the census data, there were 7 categories of education, specifically describing the highest education level of each interviewee. To make the model simpler and easier for future interpretation, I regrouped them into two categories: "Not attended university/college" and "Attended university/college". The new grouping will make the models more meaningful and concise. Besides, I chose to use age groups as a categorical variable instead of ages as a numerical variable, which allows us to divide cells based on the age groups in the later post-stratification.

Since our response is a categorical variable with 6 categories, the ideal model would be a multinomial regression model. However, to simplify the modeling and make the models more interpretable, I decided to employ 6 frequentist logistic regression models, where each model corresponds to one unique value of the response. Therefore, I added 6 binary variables to indicate whether an observation would vote for a party or not. Take the Liberal Party as an example. Since in the cleaned survey data an individual could either support the Liberal Party or not, the variable vote_liberal, which describes whether this individual supports the Liberal Party, is obviously a binary variable. Thus, a frequentist logistic regression was used to model the proportion of voters who will vote for the Liberal Party. The same idea was also applied to the rest

parties. In this way, I was able to predict the proportion of voters for each party with 6 different logistic regression models.

Based on the requirement of the later post-stratification, I had to guarantee that the variables included in the logistic models were also included in the cell characteristics of the census data. By comparing the variables in the survey data and censes data along with the relevant researches, I finally selected 7 most relevant variables for the initial full models, which were age, sex, education, province, whether born in Canada, religion importance and household income. Then, I employed the stepwise selection techniques according to AIC and BIC criterion on each logistic model. It turned out that AIC criterion could preserve more variables in every model to ensure a valid modeling process. Therefore, I kept the 6 models with around 2 to 4 variables from the AIC selection, and saved the initial complex models as the alternative models. These refined logistic models were chosen because the fewer number of variables makes it easier to describe and understand them, and also reduces the computation cost for predictions. Also, some variables had many categories in the initial full models, which might cause overfitting and make the model less general.

The final logistic regression models we obtained is:

$$\begin{aligned} log(\frac{p_1}{1-p_1}) = \beta_{01} &+ \beta_{11}x_{age:50-70} + \beta_{21}x_{age:70+} + \beta_{31}x_{age:30-} + \beta_{41}x_{sex:Male} + \beta_{51}x_{province:British\ Columbia} \\ &+\beta_{61}x_{province:Manitoba} + \beta_{71}x_{province:New\ Brunswick} + \beta_{81}x_{province:Newfoundland\ and\ Labrador} \\ &+\beta_{91}x_{province:Nova\ Scotia} + \beta_{101}x_{province:Ontario} + \beta_{111}x_{province:Prince\ Edward\ Island} \\ &+\beta_{121}x_{province:Quebec} + \beta_{131}x_{province:Saskatchewan} + \beta_{141}x_{place\_birth\_canada:Born\ outside\ Canada} \end{aligned} \tag{1}$$

$$\begin{aligned} log(\frac{p_2}{1-p_2}) = \beta_{02} &+ \beta_{12}x_{religion\_importance:importantreligion} + \beta_{22}x_{sex:Male} + \beta_{32}x_{province:British\ Columbia} \\ &+\beta_{42}x_{province:Manitoba} + \beta_{51}x_{province:New\ Brunswick} + \beta_{62}x_{province:Newfoundland\ and\ Labrador} \\ &+\beta_{72}x_{province:Nova\ Scotia} + \beta_{82}x_{province:Ontario} + \beta_{92}x_{province:Prince\ Edward\ Island} + \beta_{102}x_{province:Quebec} \\ &+\beta_{112}x_{province:Saskatchewan} + \beta_{122}x_{place\_birth\_canada:Born\ outside\ Canada} \end{aligned} \tag{2}$$

$$\begin{aligned} log(\frac{p_3}{1-p_3}) = \beta_{03} &+ \beta_{13}x_{age:50-70} + \beta_{23}x_{age:70+} + \beta_{33}x_{age:30-} + \beta_{43}x_{sex:Male} + \beta_{53}x_{province:British\ Columbia} \\ &+\beta_{63}x_{province:Manitoba} + \beta_{73}x_{province:New\ Brunswick} + \beta_{83}x_{province:Newfoundland\ and\ Labrador} \\ &+\beta_{93}x_{province:Nova\ Scotia} + \beta_{103}x_{province:Ontario} + \beta_{113}x_{province:Prince\ Edward\ Island} \\ &+\beta_{123}x_{province:Quebec} + \beta_{133}x_{province:Saskatchewan} + \beta_{143}x_{religion\_importance:important\ religion} \end{aligned} \tag{3}$$

$$\begin{aligned} log(\frac{p_4}{1-p_4}) = \beta_{04} &+ \beta_{14}x_{age:50-70} + \beta_{24}x_{age:70+} + \beta_{34}x_{age:30-} + \beta_{44}x_{province:British\ Columbia} \\ &+\beta_{54}x_{province:Manitoba} + \beta_{64}x_{province:New\ Brunswick} + \beta_{74}x_{province:Newfoundland\ and\ Labrador} \\ &+\beta_{84}x_{province:Nova\ Scotia} + \beta_{94}x_{province:Ontario} + \beta_{104}x_{province:Prince\ Edward\ Island} \\ &+\beta_{114}x_{province:Quebec} + \beta_{124}x_{province:Saskatchewan} + \beta_{134}x_{place\_birth\_canada:Born\ outside\ Canada} \end{aligned} \tag{4}$$

$$\begin{aligned} log(\frac{p_5}{1-p_5}) = \beta_{05} &+ \beta_{15}x_{province:British\ Columbia} + \beta_{25}x_{province:Manitoba} + \beta_{35}x_{province:New\ Brunswick} \\ &+\beta_{45}x_{province:Newfoundland\ and\ Labrador} + \beta_{55}x_{province:Nova\ Scotia} + \beta_{65}x_{province:Ontario} \\ &+\beta_{75}x_{province:Prince\ Edward\ Island} + \beta_{85}x_{province:Quebec} + \beta_{95}x_{province:Saskatchewan} \\ &+\beta_{105}x_{place\_birth\_canada:Born\ outside\ Canada} \end{aligned} \tag{5}$$

$$log(\frac{p_6}{1-p_6}) = \beta_{06} + \beta_{16}x_{province:British\ Columbia} + \beta_{26}x_{province:Manitoba} + \beta_{36}x_{province:New\ Brunswick} + \beta_{46}x_{province:Newfoundland\ and\ Labrador} + \beta_{56}x_{province:Nova\ Scotia} + \beta_{66}x_{province:Ontario} + \beta_{76}x_{province:Prince\ Edward\ Island} + \beta_{86}x_{province:Quebec} + \beta_{96}x_{province:Saskatchewan} + \beta_{106}x_{place\_birth\_canada:Born\ outside\ Canada} \quad (6)$$

Since the logistic models above are very similar to each other, I would explain one of them to introduce the general idea. Notice that each equation number corresponds to a party to support where 1 for Liberal, 2 for Conservatives, 3 for NDP, 4 for Bloc Québécois, 5 for Green Party and 6 for People's Party.

Let's take the Liberal Party as an example. In the equation (1), on the left hand side, p_1 represents the proportion of voters who would vote for the Liberal Party and $log\frac{p_1}{1-p_1}$ represents the logodds of the proportion of voters who would vote for the Liberal Party. On the right hand side, β_{01} is the intercept of the model which represents the logodds of probability of voting for the Liberal Party if the voter is a femle between 30 and 50 years old who was born in Canada and is living in Alberta. Besides, β_{141} represents the expected change of logodds of probability of voting for the Liberal Party for a voter born outside Canada ($x_{place\_birth\_canada:Born\ outside\ Canada} = 1$) compared to a voter born in Canada ($x_{place\_birth\_canada:Born\ outside\ Canada} = 0$), with all the other variables (age, sex, province) fixed. The other coefficients have similar meanings as the first two coefficients β_{01}, β_{141} described above. Notice that at the end of the subscript of each coefficient in equation (1), there is a 1, which is used to make distinction from the similar coefficients in the other models.

At last, I justified the logistic models with their Areas Under the Curve (AUC) value and Receiver Operating Characteristic (ROC) curves in the Figure 1 to check their binary classification abilities. In the Figure 1, I plotted sensitivity rate against false positive rate for each threshold. Suppose a model is very good at detecting true positives, then the false positive rate would be 0 and the true positive rate would be 1, which would produce an upper-triangular curve with AUC = 1. However, if the classification ability of a model is weak, then its curve would be close to the 45 degree line with AUC = 0.5. From the Figure 1 below, we can see that the trends of ROC curves indicate that our models all have a relatively good classification ability.

Post-stratification

In order to estimate the proportions of voters who would vote for each party in order to estimate the overall winner if "everyone" had voted, a post-stratification analysis was employed. First, I divided the census dataset into 2716 demographic cells based on the selected features – age, sex, province, education, religion importance, birthplace and the household income. Then, a specific party was fixed for the following procedures. For each demographic cell, the proportion of voters for the specific party was estimated by its corresponding logistic model above. Then, each proportion estimate in the cell was weighted by the respective population size of that cell. At last, the weighted estimates were summed and divided by the entire population size to estimate the proportion of people in the population that would vote for that specific party. After that, I repeated the procedures for the other 5 parties to get their estimated population-level proportions of voters.

The formula we used to calculate the post-stratified proportion is:

$$\hat{y}^{PS} = \frac{\sum_j N_j \hat{y}_j}{\sum_j N_j} \quad (7)$$

where $\hat{y}_j$ is the estimated proportion of voters for a specific party in the j^{th} cell, N_j is the population size of the j^{th} cell based off demographics (from census data). For example, if I would like to get the population-level proportion of voters for the Liberal Party, $\hat{y}_j$ is the estimated proportion of voters for the Liberal Party in the j^{th} cell.

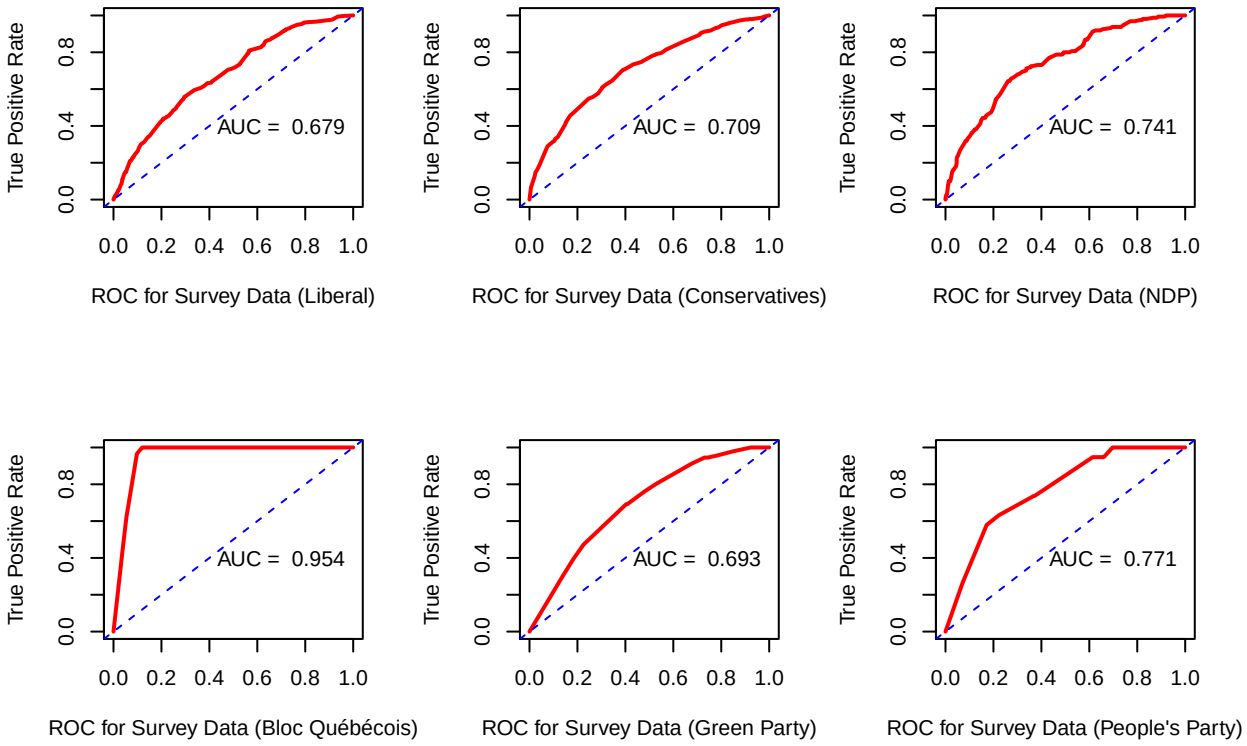

Figure 1: ROC Curve. AUC value was calculated for model performance

Here, I chose the 7 variables that were mentioned above to divide the cells in the census data because they all occurred in the survey data and they are likely to influence the voters' choices. First of all, different provinces usually have different economic conditions and political preferences, which might affect the voters' choice of their favourite parties.[6] Similarly, the household income would influence people's attitudes to the policies proposed by different parties, deciding their votings. In addition, education levels could affect an individual's understanding and judgment to each policy's policies and ideas. Besides, age, sex and birthplace would affect an individual's feeling to this country and the hometown, which can be great factors concerning the voters' own benefits. For example, an old individual would probably prefer to a party that prioritize the benefits of old people. Finally, the importance of the religion to an individual also affect the voter's attitudes towards different policies by different parties.

The post-stratification technique could help to reduce the bias of representativeness in the population by accounting for non-major groups with non-probability based sampling. With post-stratification, the groups that were underrepresented groups would correctly contribute to the final result and the variance in the prediction would reduce.

All variables described can be referred to Table 4 in Appendix.

Result

From the previous sections, I have shown the 6 logistic regression models derived by equations (1) - (6) and I would use them to predict the proportions of voters for each party and see what would be the result if "everyone" had voted. Since the models are similar to each other, I would introduce the model for the Liberal Party as an example.

Table 1: Coefficients Table

coefficient	predictor variable	value
β_{01}	intercept	-1.8854
β_{11}	$x_{age:50-70}$	0.2864
β_{21}	$x_{age:70+}$	0.6612
β_{31}	$x_{age:30-}$	-0.2020
β_{41}	$x_{sex:Male}$	-0.2837
β_{51}	$x_{province:BritishColumbia}$	0.4109
β_{61}	$x_{province:Manitoba}$	0.8520
β_{71}	$x_{province:NewBrunswick}$	1.1742
β_{81}	$x_{province:NewfoundlandAndLabrador}$	1.6335
β_{91}	$x_{province:NovaScotia}$	1.5052
β_{101}	$x_{province:Ontario}$	1.3851
β_{111}	$x_{province:PrinceEdwardIsland}$	1.5406
β_{121}	$x_{province:Quebec}$	1.2112
β_{131}	$x_{province:Saskatchewan}$	-0.2165
β_{141}	$x_{placeBirthCanada:BornOutsideCanada}$	0.8942

All the important predictors are listed in the Table 1 above, along with the corresponding coefficients. Then, the model derived by the equation could be written as

$$
\begin{aligned}
log(\frac{p_1}{1-p_1}) = -1.8854 + 0.2864x_{age:50-70} + 0.6612x_{age:70+} - 0.2020x_{age:30-} - 0.2837x_{sex:Male} \\
+0.4109x_{province:BritishColumbia} + 0.8520x_{province:Manitoba} + 1.1742x_{province:NewBrunswick} \\
+1.6335x_{province:NewfoundlandAndLabrador} + 1.5052x_{province:NovaScotia} + 1.3851x_{province:Ontario} \\
+1.5406x_{province:PrinceEdwardIsland} + 1.2112x_{province:Quebec} \\
+ - 0.2165x_{province:Saskatchewan} + 0.8942x_{placeBirthCanada:BornOutsideCanada}
\end{aligned} \tag{8}
$$

In this model for the Liberal Party, $\beta_{01} = -1.8854$ represents that the logodds of the proportion of voters for the Liberal Party would be -1.8854 if voters are all females between 30 and 50 years old who were born in Canada and are living in Alberta. $\beta_{30-} = -0.2020$ means that the expected change of the logodds of the proportion of voters for the Liberal Party would be -0.2020 for voters younger than 30 years old compared to voters between 30 and 50 years old, with all other variables fixed. $\beta_{101} = 1.3851$ represents that the expected change of the logodds of the proportion of voters for the Liberal Party would be 1.3851 for voters living in Ontario compared to voters living in Alberta, with all other variables fixed. $\beta_{141} = 0.8942$ means that the expected change of the logodds of the proportion of voters for the Liberal Party would be 0.8942 for voter born outside Canada compared to voters born in Canada, with all other variables fixed. The other coefficients have similar meanings as the coefficients described above.

The other logistic models for the other parties are very similar to the one described above. With these logistic models, I finally reached the proportions of voters for each party in the population level if "everyone" had voted. It is summarized in the Table 2 below.

Table 2: Estimated Proportion of Parties

Party Name	Population-level Proportion
Liberal	0.3675301
Conservatives	0.3793230
NDP	0.1370669
Bloc Québécois	0.0463317
Green Party	0.0652423
People' Party	0.0121437

The proportions in the Table 2 and Figure 2 are the estimated population-level proportions of voters for different parties if "everyone" had voted, which came from the combination of multiple ligistic models and the post-stratification. For example, it is estimated that in total 37.9323% Canadian would vote for the Conservatives if "everyone" had voted, which is based off our post-stratification analysis of the proportion of voters in favour of the Conservatives modelled by a logistic regression model, which accounted for sex, province, religion importance and birthplace.

Discussion

Summary

During the whole study, I used a survey data and a census data to build up the general model. The survey data was collected by calling the interviewees once before the 2019 Canadian Federal Election and once after the election and asking questions concerning many aspects in the 2019 Canadian Election Study (CES). The census data was collected by calling the interviewees about the private information by the General Social Survey (GSS) in Statistics Canada.

I first removed the missing values and standardized the categorical variables in the survey and census datasets to ensure the validity of the data. Then, I built 6 logistic regression models for 6 parties based

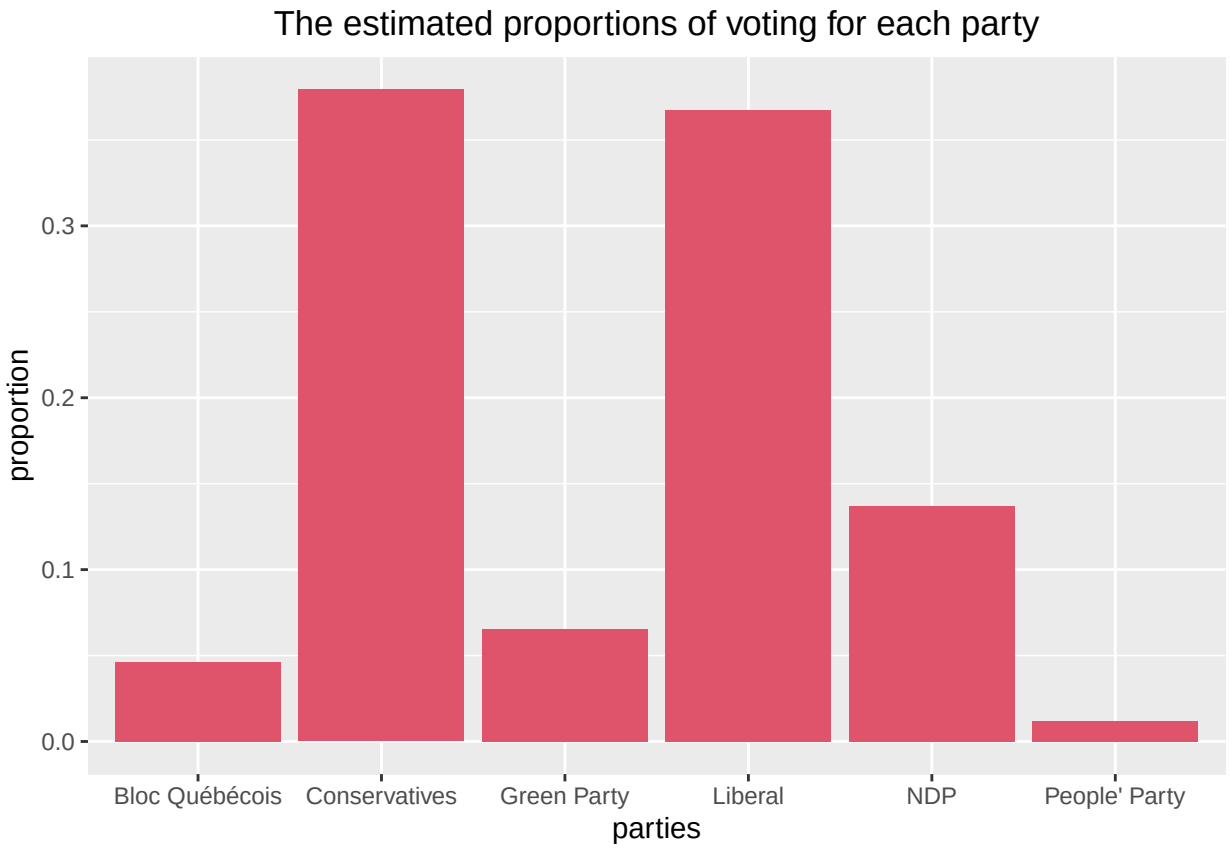

Figure 2: Estimated Proportion of Parties

on the survey data to replace a multinomial regression model. To justify the models, I employed AUC values and ROC curves to guarantee the appropriateness of my models. Later, I used the census data for the post-stratification for each logistic regression model to predict the overall population-level proportion of voters for each party in the 2019 Canadian Election Election if "everyone" had voted. As the result, the Conservatives would win the election with 37.93% of all votes in that case.

Conclusion

Table 3: Estimated Proportion of Parties

Party Name	Population-level Proportion
Liberal	0.3675301
Conservatives	0.3793230
NDP	0.1370669
Bloc Québécois	0.0463317
Green Party	0.0652423
People' Party	0.0121437

In conclusion, my study suggested that the Conservatives would win the 2019 Canadian Federal Election with 37.93% of all votes if "everyone" had voted. This conclusion is quite different from the fact that Justin Trudeau from the Liberal Party won the 2019 Canadian Federal Election. This difference between facts and statistical analysis may suggest that the turnout did have important influences on the election results since we assumed that "everyone" had voted in our study. Comparatively, the actual turnout in 2019 Canadian Federal Election was only 66% [8], which was far from the "everyone". Therefore, it is reasonable to assume that the turnout did play an important role in determining the 2019 Canadian Federal Election result. However, in my study, the estimated population-level proportion of voters for the Liberal was also 36.75301%, which was only 1.179% lower than that for the Conservatives. Considering that I cleaned the data with my own assumptions, it is possible that some missing values could change the estimated result. Besides, I can also conclude that age, sex, province, education, birthplace, household income and religion importance have some compacts on the voting preference and the turnout.

Another obvious pattern in the result of my study is that the Liberal and Conservatives are still dominant in the Canadian Federal Election while neither of them could have obvious advantages over the other. In this case, the change of turnout seems not able to help parties other than the Liberal and Conservatives to win the election.

Weakness

As mentioned above, the survey dataset came from the 2019 Canadian Election Study (CES) by phone interviewing and online survey in two phases and the census dataset came from the General Social Survey (GSS) in Statistics Canada where the data was also collected by phone calls. Even I made my efforts to clean the two datasets, there are still some biases and drawbacks there. First, the two datasets were actually collected in different years. According to the guide books, the survey data from CES was collected on October 2019 while the census data from GSS was collected from February to November in 2017. This mismatch of time means that the models trained on the survey dataset are actually inappropriate for predictions on the census data, since the census dataset does not contain the same attributes for its observations. Depending on the economic and political changes from 2017 to 2019 in Canada, my conclusion of this study might be much biased. Second, my cleaning procedures might also cause some potential biases in the data. During the cleaning, I removed all the observations that said "Don't know" or "Spoil ballet" in the question about the supporting party. However, there might be other ways to refer to their choices or I could weight the known observations to correct the biases. Third, the census dataset is actually not large enough. With only 20602 observations inside, the census dataset might miss some people with specific characteristics. For example, because of the sampling frame, people without a landline and wireless phone could not get interviewed. The initial bias of the census data could also affect my study result. Finally, when I selected the predictor variables, I thought some variables such as "employment status" were also significant. However, I did not select those variables in the end because they did not occur in the both datasets, which is resulted from the questions posted in each survey study.

Another drawback about the data is my choice of predictor variables. The selection of predictor variables only depended on the AIC selection, which was supported by statistics. However, in the census data in Figure 3, some variables are actually evenly distributed among their categories, which means that the selection of them lacks the support from the social science.

There were also shortcomings in the analysis procedure. In this study, I employed multiple logistic regression models to replace the multinomial regression model for convenience. However, my logistic models were built in a parallel way instead of sequentially. This would cause the potential loss of dependent relationships among data, which might bias the final result and conclusion.

Next Step

In order to improve the study, I would like to find a better dataset as the census data which does not have time mismatch with the survey dataset so I could eliminate the potential bias and increase the validity of my study. Meanwhile, I am considering to apply a multilevel regression model or a multinomial regression model [9] in the analysis to improve the modeling to capture more patterns of the data and avoid biases. Besides, more researches could be done to investigate the factors that influence the voting choices and the turnout. Thus, I would have a more socially reasonable model for better predictions.

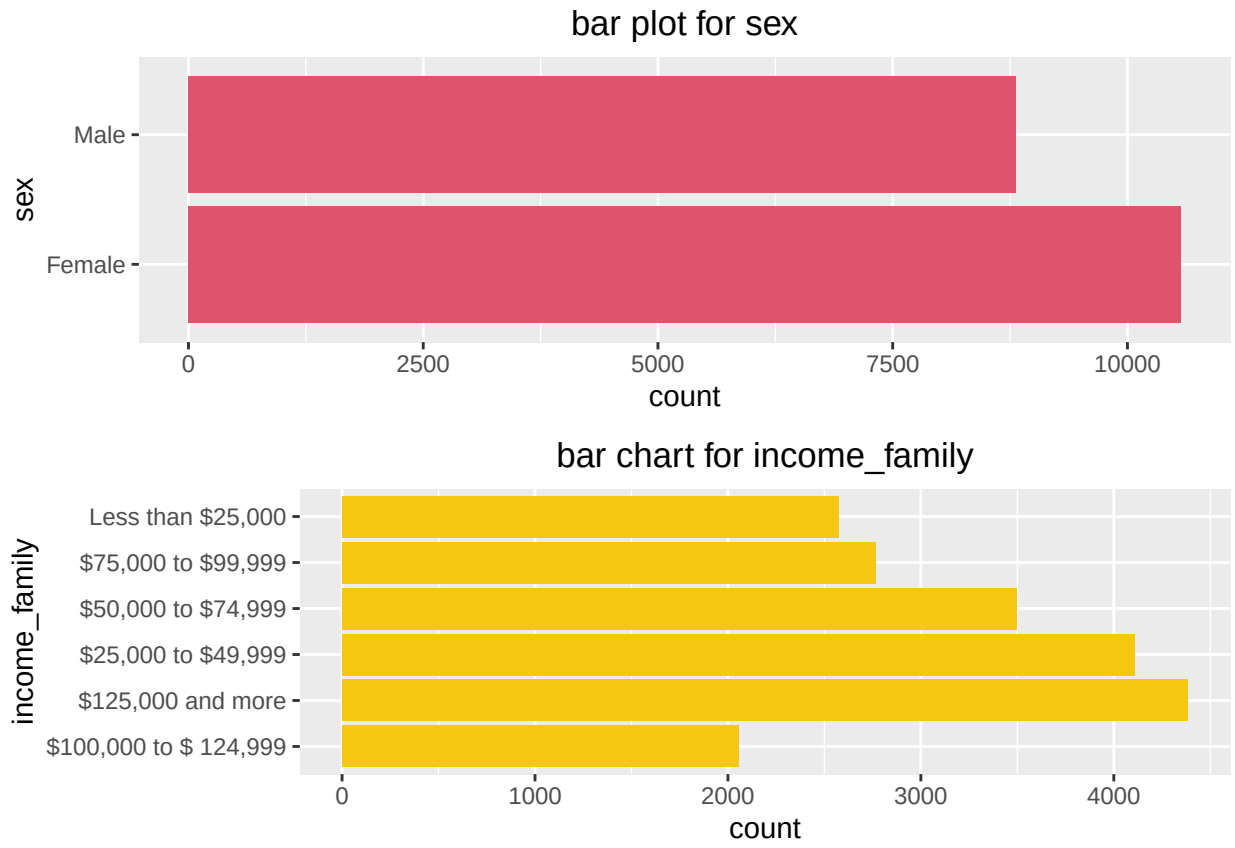

Figure 3: Parts of the census data

Reference

1. CHASS Data Centre, Faculty of Art and Science of University of Toronto, http://dc.chass.utoronto.ca/myaccess.html
2. Paul A. Hodgetts; Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.
3. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Phone Survey", https://doi.org/10.7910/DVN/8RHLG1
4. Malone Mullin, 2019, "Here's what happens when you spoil a ballot", https://www.cbc.ca/news/canada/newfoundland-labrador/spoiled-ballots-1.5136461
5. David Rayside; Jerald Sabin; Paul E.J. Thomas, 2017, "Religion and Canadian Party Politics", https://www.ubcpress.ca/asset/20215/1/9780774835589_Excerpt.pdf
6. Uppal, Sharanjit and LaRochelle-Côté, Sébastien. 2012. "Factors associated with voting". Perspectives on Labour and Income. Spring 2012, vol. 24, no. 7. Statistics Canada Catalogue no. 75-001-XIE. https://www150.statcan.gc.ca/n1/pub/75-001-x/2012001/article/11629-eng.htm
7. Wikipedia contributors. "Voter turnout in Canada." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 2020. https://en.wikipedia.org/w/index.php?title=Voter_turnout_in_Canada&oldid=990862986
8. Wikipedia contributors. Multinomial logistic regression. In Wikipedia, The Free Encyclopedia, 2020. https://en.wikipedia.org/w/index.php?title=Multinomial_logistic_regression&oldid=984397922
9. Rohan Alexander, Sam Caetano, 2020, https://q.utoronto.ca/courses/184060/files/9422740/download?download_frd=1
10. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
11. Course Notes for IS 6489, Statistics and Predictive Analysis, September 03, 2017, https://bookdown.org/jefftemplewebb/IS-6489/logistic-regression.html
12. Wu, Changbao, and Mary E. Thompson. "Basic Concepts in Survey Sampling." Sampling Theory and Practice. Springer, Cham, 2020. 3-15.
13. Alan Agresti. "Categorical Data Analysis (3rd edition)". Wiley, 2011.

Appendix

Table 4: Census Variables

names	types	type number	meanings
age	categorical	4	the age groups
sex	categorical	2	male or female in biological sense
province	categorical	13	living places by provinces of Canada
education	categorical	2	education levels: go to university/college or not
place_birth_canada	categorical	2	born in Canada or outside Canada
religion_importance	categorical	2	whether the religion is important
income_family	categorical	6	the household income groups
vote_party	categorical	6	the party to support and vote for

Figure 4: Census Data Visualization 1

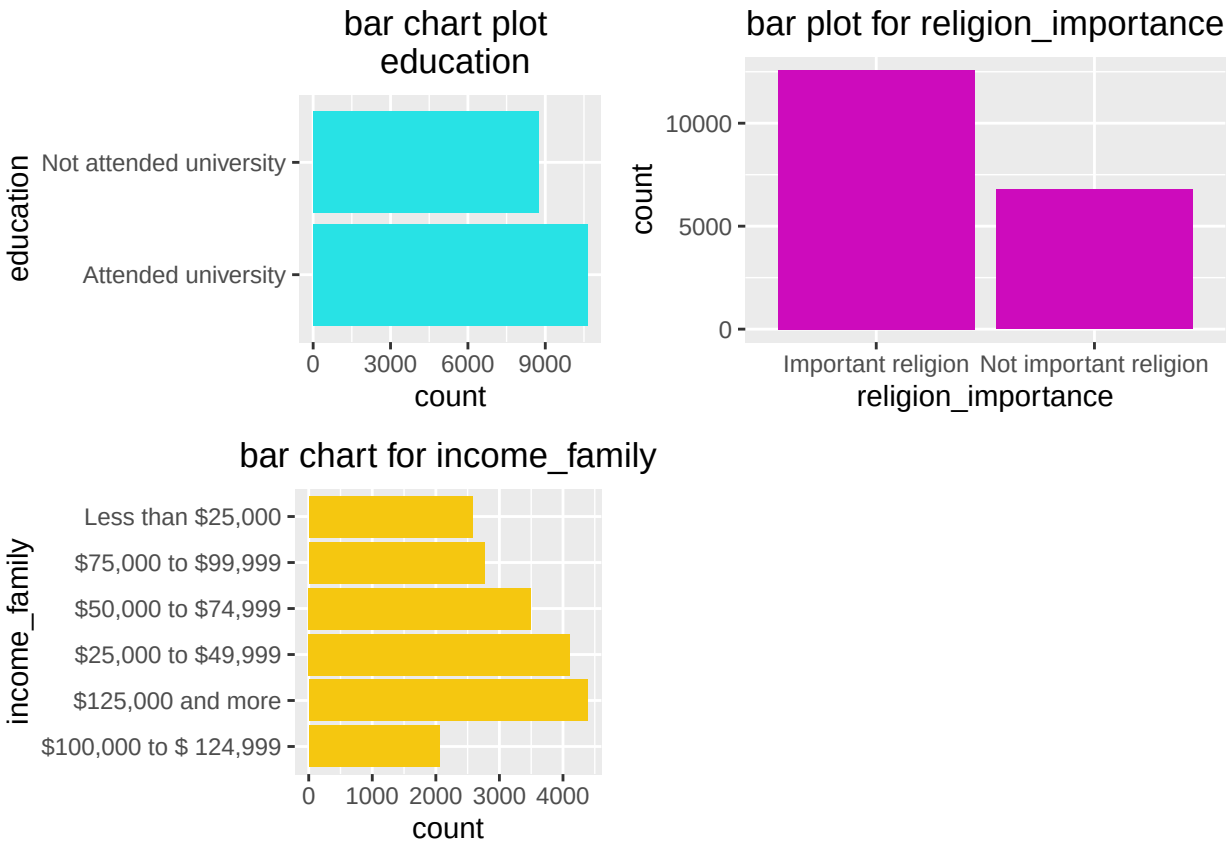

Figure 5: Census Data Visualization 2

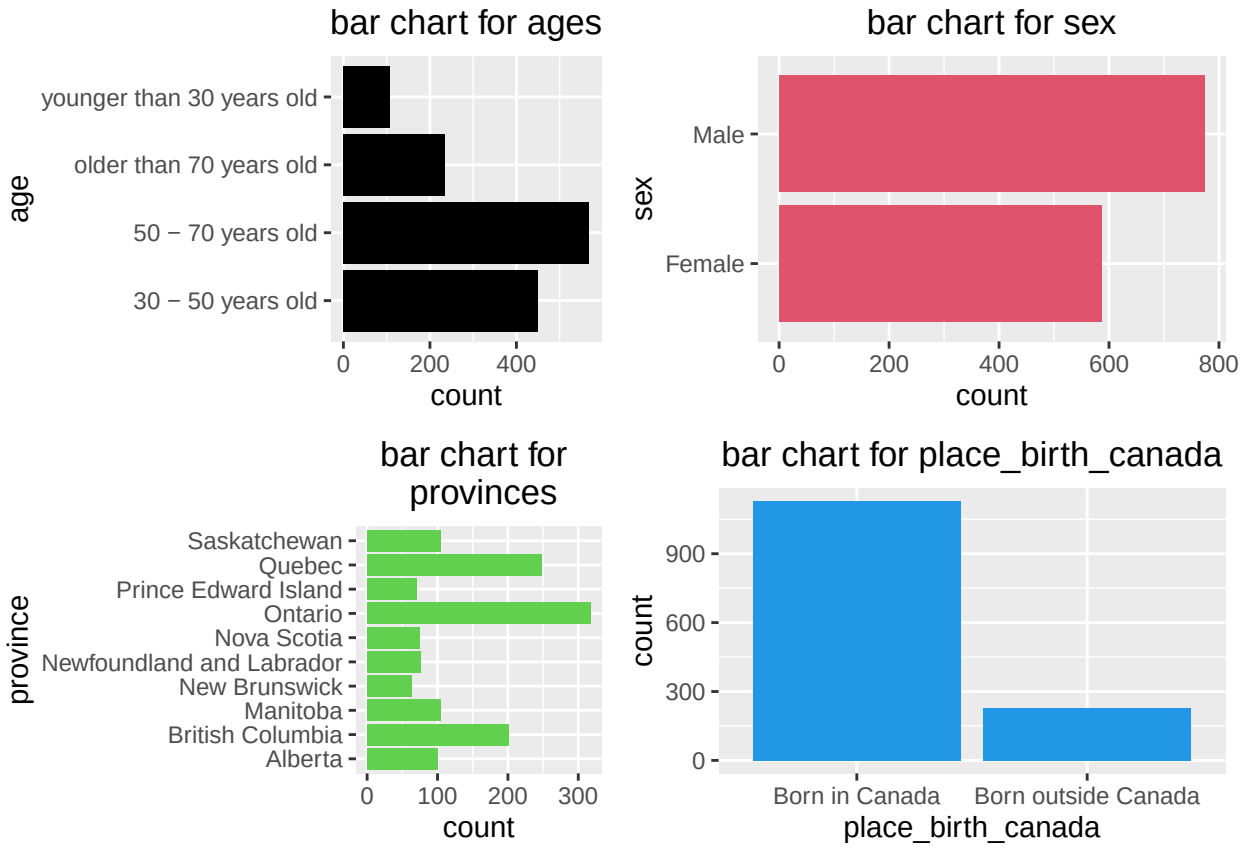

Figure 6: Survey Data Visualization (partial)

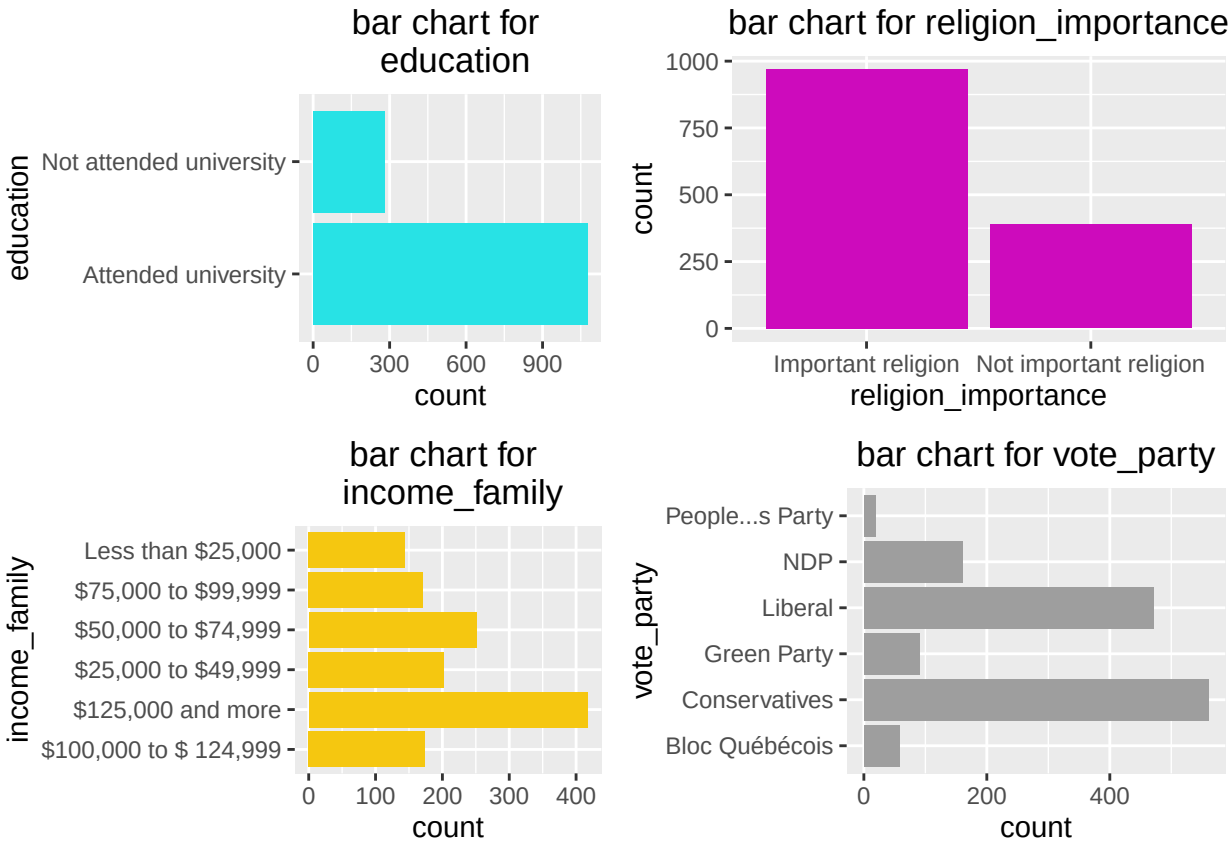

Figure 7: Survey Data Visualization (partial)