Predicting the Competition between the Liberal and the Conservative in the Canadian Federal Election 2025

STA304 - Fall 2023 -Assignment 2

Group 120: Ruilin Peng, Johnson Guo, Tony Chen, Yuchen Li

11/7/2023

1. Introduction

1.1 The research question

What will be the predicted outcome for the party that receives the highest amount of votes in the upcoming election 2025 between the Conservative Party and the Liberal Party?

1.2 Importance of the research

Free and fair elections are considered important features contributing to a healthy democracy. Fairness of the election could be ensured as citizens can vote to reflect their wills without political interference. These conditions can contribute to a durable democracy in Canada and also improve the internal political efficacy of Canada (Elections Canada, 2023). The choice of who will form the government and which party's policies will guide the government and legislation. These appointments and policies can influence various aspects of society, including the economy, public health initiatives, and environmental sustainability. The analysis of election outcomes helps encourage citizens to track the political process and promote their participation (Statistics Canada, 2022). Political parties and candidates can learn about competitors' voting tendencies and concerns of voters so adjusting their electoral strategies.

Importantly, our analysis provides a demonstration of the relationship between the selected variables and the binary outcomes of voting decisions (vote or not vote) to see how the change of variables influences the change of probability change in vote. In this case, the result of the analysis can be an essential reference for future study of Canadian elections, and representatives of parties may set up relevant rules and benefits to increase their political reputation and construct campaign strategies according to the favorable factors considered by respondents. Furthermore, the government may refine relevant policy in terms of the races, employment, income, and other socioeconomic data presented in this study.

1.3 Terminology

Turnout in political elections means the probability or proportion of popular votes for a party.

The **45th Canadian federal election** will be held on or before October 20, 2025. This date is determined by the Canada Elections Act (Parliament of Canada, 2006). There are currently 337 members in office in the 44th Parliament. Both the Liberal Party and Conservative Party have 275 members (158 and 117 respectively), and they are the two major Canadian parties(Members of Parliament, 2023).

The **Liberal Party** is the oldest active federal political party in Canada and has dominated federal politics for the majority of Canadian history (Liberal Party, 2019). The party espouses liberal principles and is generally central to center-left. Its main competitor, the Conservative Party, is right-leaned. Liberal Party is currently led by Justin Trudeau who has been the prime minister of Canada since 2015. Signature policies and legislation of the Liberal Party include universal health care, pension plans, bilingualism, gun control, related charters and acts that legalize same-sex marriage and cannabis usage, and expanding access to abortion (Liberal Party of Canada. n.d.).

The **Conservative Party** is one of the major political parties in Canada, known for its center-right political orientation. Its platform typically includes a commitment to fiscal responsibility, free-market principles, and a focus on individual liberties. The party has traditionally attracted support from a broad coalition of conservative voters, including those with social or economic conservative values. The party is currently led by Pierre Poilievre since 2022. Its signature policies involve reducing government debt, lowering taxes, and eliminating the long-gun registry (Dippel, 2016).

The **Liberal Party** leans on progressive policies and social transformation, and the Liberal encourages free-market competition in economics, diversity of minorities, and inclusion of immigrants. As opposed to the Liberal, the Conservative prefers traditional and preservative rules, imposes more government regulation on economics, and disagrees with social activism (Canada Guide, 2023).

Canada includes five distinct **regions**: The Atlantic Provinces (New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island), Central Canada (Ontario and Quebec), The Prairie Provinces (Manitoba, Saskatchewan and Alberta), The West Coast (British Columbia), and The Northern Territories (Nunavut, Northwest Territories, and Yukon Territories) (Government of Canada, 2022). To more easily present the data, adjustments have been applied to the categories of five regions, and the Prairie Provinces have been renamed into western regions.

Visible minority refers to individuals who are non-Caucasian in race or non-white in color, except indigenous population groups. The visible minority group consists of South Asian, Chinese, Black, Filipino, Latin American, Arab, Southeast Asian, West Asian, Korean, and Japanese (Statistics Canada, 2021). To present data concisely, we adjusted the components of the visible minority group by also including the indigenous population.

1.4 Hypothesis

As the 45th Canadian Election is approaching, the competition between "Liberal" and "Conservative" is kicked off again. The Liberal and the Conservative parties dominate the political system of Canada and the show plays between these parties. Moreover, if we look at the recent 20 years of popular vote percentages, the Liberal Party almost beats the Conservative Party for each election with an outstanding difference (Simon Fraser University, 2021). Therefore, we hypothesize that the Liberal party is still going to win the election and it is exciting to see if there is any chance that the Conservative will probably present a surprise.

The goal of this study is to predict which political party in Canada is going to win the most popular votes in the 2025 Canadian Election based on the information from the previous 2021 Canadian Election. This study matches the survey data collected from the 2021 CES (Canadian Election Study) online survey with the census data collected from the Canada GSS (General Social Survey) and predicts the voting decision of respondents in the next 2025 Election in a logistic regression model based on demographic variables of respondents which include age, sex (male/female), language (English/French), income class (Low, Medium and High), the religious or atheist, and the races (the white or the other minorities). (Stephenson et al.,

2021; Canada Statistics, 2020). To enhance the representatives of the population, we apply the stratification method that divides the survey respondents into strata by different variables and conclude the estimated voting probability for both the Liberal and the Conservative to infer the possible outcomes in the 2025 Election.

2. Data

2.1 Data description

Survey Data

The survey consists of the campaign period part and the post-election period part. The collection process started on August 7th, 2021, and ended on October 4th, 2021. To demonstrate, the campaign period survey data was collected on an online sample of 20,968 respondents of the Canadian general population at the Leger Opinion panel. The sampling process included three waves of panels in a modified rolling cross section. In this case, the survey data of three waves were merged into a whole sample at the end, but the respondents of the survey were randomly selected in three different periods. (Stephenson et al., 2021).

Census Data

The census data in this study is the 2020 General Social Survey, which provides detailed information on social trends for addressing interested social issues and improving the living conditions of Canadians, and it was released by September 2021. In addition, data was collected by both computer-based phone calls and electronic questionnaires through mail links from Statistics Canada offices in different regions from August 2020 to February 2021. According to the descriptive statistics, the overall response rate was 40.3%, and a total of 20,602 observations were collected. Moreover, the stratified sampling method was applied where each response was assigned to a province-level (a stratum). Then, a simple random sampling (SRS) was conducted in each stratum (Statistics Canada, 2020).

2.2 Data cleaning process

The sample survey data in this study was collected during the 2021 federal election campaign period and intended to gather the opinions of Canadian citizens and permanent residents aged 18 or older about election votes. Data cleaning was conducted on both survey and census datasets to enhance their suitability for analysis. Variables from both sources were aligned, allowing for the extrapolation of survey findings to the broader census population.

Initially, the survey dataset was filtered to include only Canadian citizens, excluding all Permanent Residents. Additionally, we filtered out respondents under 18 years old, as only Canadian citizens above 18 have the right to vote.

In the census dataset, age was already a present variable and was rounded for consistency.

Regarding gender, the survey dataset included an "Other" option, but this was minimally selected. Given its negligible proportion and lack of impact on key variables' distribution, this category was excluded from the dataset. The survey's gender data and the census's sex data were thus streamlined to include only "Male" and "Female" options, with the term 'sex' in the census data being renamed to 'gender' for uniformity.

Both the survey and census datasets included information on respondents' current province of residence. Considering Canada's composition of 10 provinces and 3 territories, we streamlined the data based on geographic and demographic considerations to determine whether they are in the West. As mentioned earlier, we divided the country into five distinct regions. In the codebook, the western region initially included Manitoba, Saskatchewan, Alberta, and British Columbia, but we rearranged it to contain only the Prairie Provinces. It's important to note that there were no responses from the Northern Territories (Nunavut, Northwest Territories, and Yukon Territories).

In both survey and census datasets, respondents provided information on their primary language. We matched the census's 'language_home' data with the survey's 'UserLanguage', as it likely represents the respondent's first language or the language they are most proficient in.

Regarding income, we aligned the census's 'income_family' data with the survey's household income data, as both represent household income. Households earning less than $49,999 were classified as low income, those earning $50,000 to $124,999 as median income, and those earning above $125,000 as high income. (Springfield Financial, 2023).

Considering Canada's diverse immigrant population, respondents were simply classified as having or not having a religious affiliation, without specifying particular beliefs. Responses of 'Don't Know' were ignored.

Regarding visible minorities, we differentiated between 'Not a visible minority' and 'visible minority', disregarding 'Don't Know' responses.

Finally, after the initial data-cleaning process, respondents with missing values were excluded for clarity and accuracy. This step was necessary as incomplete responses could be due to various factors, such as issues in reaching respondents or the survey's design causing discomfort. Missing values could introduce anomalies, potentially reducing prediction accuracy. Although adjustments could have been made with more resources, omitting these respondents was considered the most appropriate approach for this study.

To glimpse the cleaned version of census data and survey data, please refer to Table 13 and Table 14 in the Appendix.

2.3 Data summary

The variables described in this section applied to both the CES survey data and GSS census data because we matched the variables in both datasets in the Data Cleaning section.

Response Variables
vote of conservative: The binary variable that whether the voter will vote for the Conservative Party in the 2025th election.

vote of liberal: The binary variable that whether the voter will vote for the Liberal Party in the 2025th election.

Independent Variables
age: This variable represents the age of the observations. It is a numerical variable.

sex: This variable represents the sex of the observations

language: A categorical variable indicating the primary language of individuals, with categories "FR" for French and "EN" for English.

west: A binary variable indicating whether an individual is from Western provinces (Manitoba, Saskatchewan, Alberta) or not.

income: A categorical variable categorizing individuals into "Low", "Medium", and "High" income groups based on their income category.

religion: A binary variable indicating whether individuals have a religious affiliation or not.

minority: A binary variable indicating whether individuals are part of a visible minority or not.

Table 1: Statistics about the proportions of Yes votes for the Liberal and the Conservative in Survey data

LiberalProportion	ConservativeProportion
0.1719	0.181

Table 1 shows that from survey result, the proportions of respondents indicating their favor for Liberal party versus for Conservative are similar with less than 0.01 difference.

Table 2: Statistics about the ages of voters in 2021 election survey data

min	Q1	median	Q3	max	IQR	mean	sd
18	30	46	63	92	33	47	18.81

Table 3: Statistics about the ages of voters in 2021 election census data

min	Q1	median	Q3	max	IQR	mean	sd
18	39	56	68	80	29	53.59	17.21

Table2 & 3 show that the distribution of age in years from survey and census are slightly different. The mean age from the census is 53.59, yet that from survey is 47. Median age from census is 56 yet it is 46 in the survey. This indicates that in census, 50% of the people are elder than 56 while people from survey are generally younger. In census, a Q3 (the age at 75% of the sample when the ages are ordered from small to large) of 68 also indicates that 25% of the population is elder than 68 years old and in survey,there is 25% of the population is elder than 63 years old . However, standard deviation is similar, with 17.21 in census and 18.81 in survey, which indicates the variations of both data are similar.

Table 4: Statistics about the frequency and proportion of male and female voters in 2021 election survey and cenus data

sex	SurveySexNumber	SurveySexProportion
Female	149	0.67
Male	72	0.33
sex	CensusSexNumber	CensusSexProportion
Female	9250	0.55
Male	7680	0.45

Table 4 shows that there exists a slightly higher proportion of Female individuals in survey compared to census

Table 5: Statistics about the frequency and proportion of French and English speaking voters in 2021 election survey and cenus data

language	SurveyLanguageNumber	SurveyLanguageProportion
EN	158	0.71
FR	63	0.29
language	CensusLanguageNumber	CensusLanguageProportion
EN	13541	0.8
FR	3389	0.2

Table 5 shows that the proportion of English speaker versus French speaker is similar across survey and census data with 0.71 versus 0.8 for English speaker and 0.29 versus 0.2 for French speaker.

Table 6: Statistics about the frequency and proportion of west and non-west voters in 2021 election survey and cenus data

west	SurveyWestNumber	SurveyWestProportion
No	186	0.84
Yes	35	0.16
west	CensusWestNumber	CensusWestProportion
No	13644	0.81
Yes	3286	0.19

Table 6 also shows that the proportion of resident living in west part of the country versus east resident is similar across survey or census data with 0.84 versus 0.81 for east resident and 0.16 versus 1.19 for west resident.

Table 7: Statistics about the frequency and proportion of high, medium, low income voters in 2021 election survey and cenus data

income	SurveyIncomeNumber	SurveyIncomeProportion
Low	120	0.54
Medium	92	0.42
High	9	0.04
income	CensusIncomeNumber	CensusIncomeProportion
Low	5831	0.34
Medium	7256	0.43
High	3843	0.23

From table 7, we can observe that distributions of income from survey and census slightly differ. Specifically, the data of census consist of higher proportion of high-income individuals yet less proportion of low-income individual.

Table 8: Statistics about the frequency and proportion of religious and non-religious voters in 2021 election survey and cenus data

religion	SurveyReligionNumber	SurveyReligionProportion
No	63	0.29
Yes	158	0.71
religion	CensusReligionNumber	CensusReligionProportion
No	3361	0.2
Yes	13569	0.8

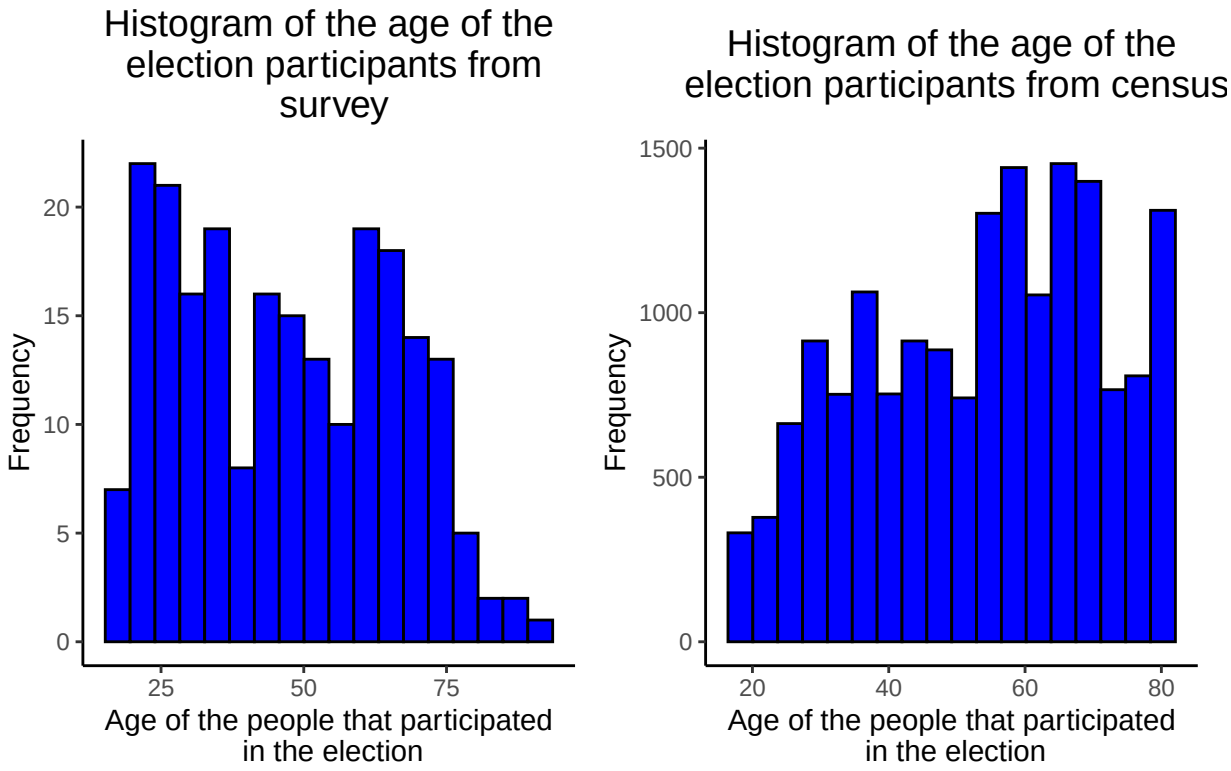
Table 8 also shows that the proportion of religious individuals versus non-religious individuals is similar across survey or census data with 0.71 versus 0.8 for religious and 0.2 versus 0.29 for non-religious.

Table 9: Statistics about the frequency and proportion of minority and non-minority voters in 2021 election survey and cenus data

minority	SurveyMinorityNumber	SurveyMinorityProportion
No	155	0.7
Yes	66	0.3
minority	CensusMinorityNumber	CensusMinorityProportion
No	15793	0.93
Yes	1137	0.07

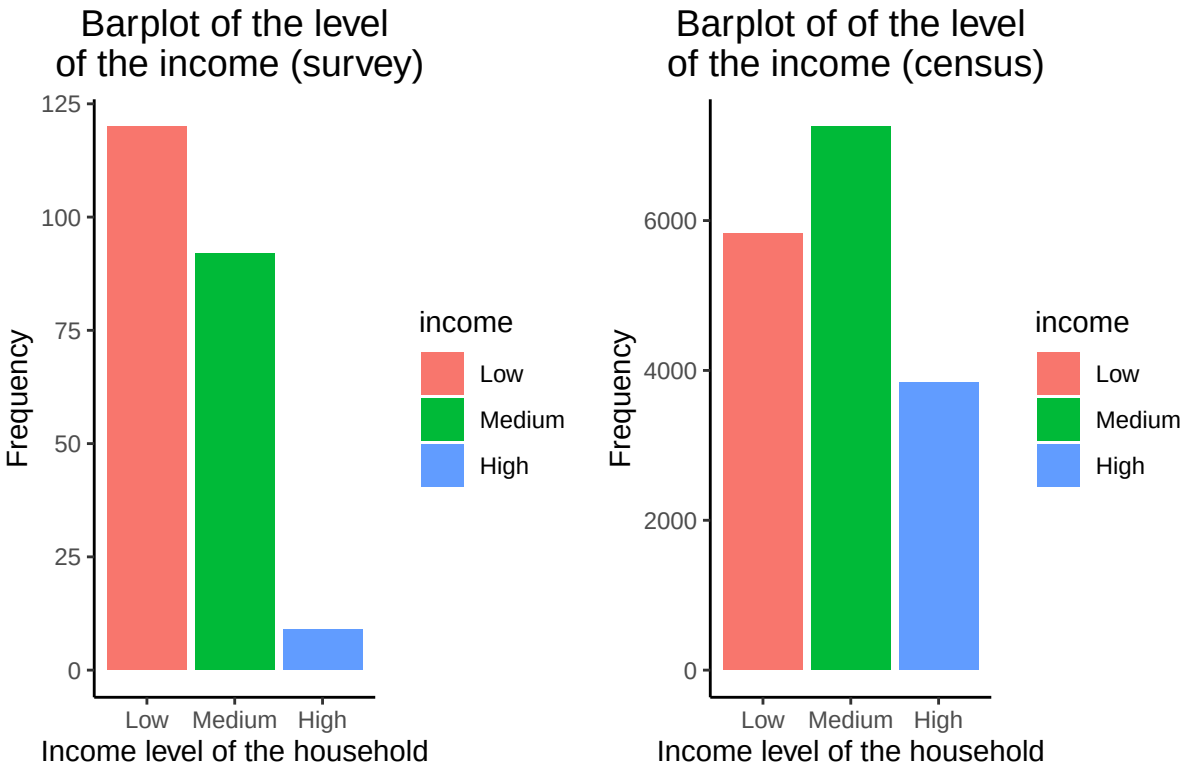
From table 9, we can observe difference between proportions of minority and non-minority between survey and census. The census has a higher proportion of non-minority while lower proportion of minority respondents.

Graph 1: Age Distributions: Survey vs Census

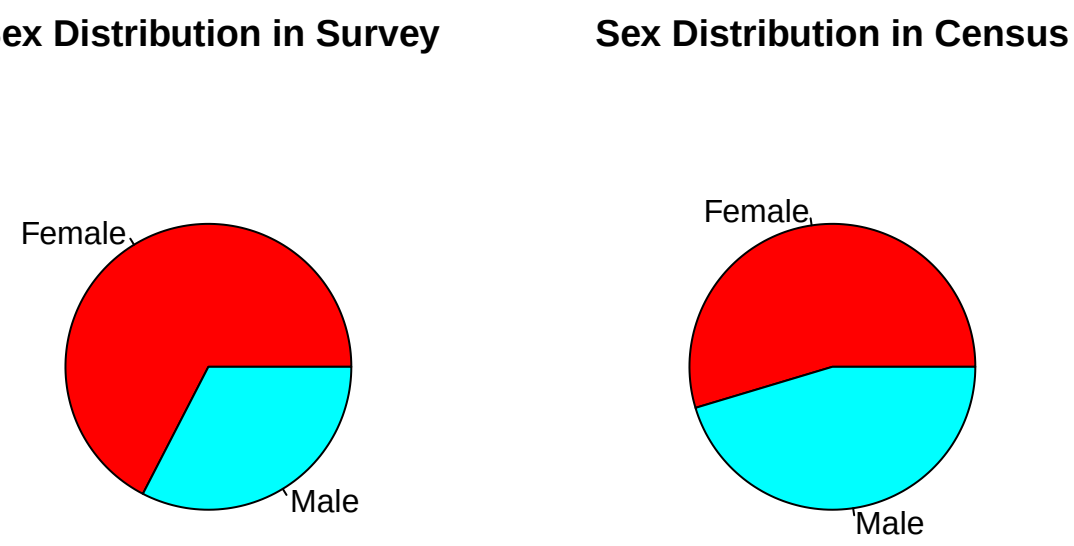


From Graph 1, we can observe that the distribution of age in survey is relatively uniform with no outliers while the distribution of the census is approximately symmetrically distributed with major mode around 60s and 80s and minor mode around late 30s which represent elder and middle aged. Also, the distribution of the census is slightly right-skewed which corresponds to our previous observation of slightly higher median than mean.

Graph 2: Income Distributions: Survey vs Census



From Graph 2, we can observe from the barplots that distributions of survey and census slightly differ. Specifically, the data of census consist of more individuals of medium income while in the survey, we can observe a mode of low income group.



Graph 3: Sex Distributions: Survey vs Census

From the pie charts of Graph 3, there exists a higher proportion of Female individuals in survey compared to census, and the the proportion of Females is generally greater than males in both datasets.

3. Methods

3.1 Model selection and rationale

Our study will apply the logistic regression model to estimate the probability of voting for both the Liberal Party and the Conservative Party. Logistic regression is used for investigating the relationship between a binary outcome and independent variables. The outcome is categorical and represents two classes, coded as 0 and 1. The outcomes of interest, in our case, are whether to vote for the Liberal Party or not and whether to vote for the Conservative Party or not. In this case, 0 represents not voting for the Party, while 1 represents voting for that.

The model follows

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k i$$

The model outcome is the log of odds where p is the probability of the event of interest occurring, and those coefficients reflect the change in log odds. Based on factors that influence individuals' voting reported by Statistics Canada, we chose variables listed as the following: age, sex, language, living in particular western regions, overall household income, religious status, and minority identity (2012). In 2011, the voting rate among the 18-24 age group was about 50%. The elderly, by contrast, are more likely to vote with a 70% rate from the 45-54 age group. Turnout increased to 82% among the 65-74 age group (Statistics Canada, 2012). Rutgers University reported that there are generally about 7.4 million more registered female voters in the US than registered male voters (2023). Meanwhile, the female group is usually more left-leaning compared to the male group, so females tend to vote less for the Conservative Party.

Regions can influence voting decisions, the Prairie Provinces such as Saskatchewan and Alberta, or western

regions in this case, tend to vote for the Conservative Party (Keller, 2021). People with religious beliefs and Canadian French speakers are more likely to support the Conservative Party as well (Grenier, 2017; Keller, 2021).

Economic well-being also is positively related to voting participation. People with higher household incomes are more likely to vote for the right-leaning party (McMaster University, 2023). Research also showed that immigrants from Central/South America or East Asia tend to vote less in the election (Statistics Canada, 2012).

Then, we will use the post-stratification which divides the survey data in terms of the variables common in the census data and the survey, and yields the estimated overall probability of voting for the Liberal Party and Conservative Party respectively. In the end, we compare the estimated voting probability between the two parties to decide which one is going to win the election (i.e. the party has the higher estimated probability). In our model, to ensure the qualification of voting, we only keep the respondents who are at least 18 years old. In order to remove the influence of missing data, we delete the observations (record of respondents) which contain missing values of variables.

3.2 Model Specifics

Our study creates two logistic regression models to predict the proportion of voters who will vote for the Liberal Party and the Conservative Party. Some assumptions are made before constructing our logistic regression models: The voting outcome of respondents is binary, that is, the respondents either vote for the party or not, and the selected categorical variables are independent in both the logistic models built for the Liberal and the Conservative. We select the following models because the variables are matched with similar proportions of each outcome- in both the census and the survey data, as shown in the tables of the Data Summary section (Tables 2, 3, 4, 5, 8). In addition, a previous study by Statistics Canada proposes that economic status, belief, and personal inborn identities such as sex, language, and race are the dominant factors that potentially influence the intention of voting, so we select the following variables as predictors to investigate the probability of each voter to vote (Uppal, S.& LaRochelle-Côté S., 2012). In the logistic regression models for both parties, we use the same variables as predictors.

$$log(\frac{p_c}{1-p_c}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{French} + \beta_4 x_{west} + \beta_5 x_{incomeMedium} + \beta_6 x_{incomeHigh} + \beta_7 x_{religion} + \beta_8 x_{minority}$$

$$log(\frac{p_l}{1-p_l}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{French} + \beta_4 x_{west} + \beta_5 x_{incomeMedium} + \beta_6 x_{incomeHigh} + \beta 7 x_{religion} + \beta_8 x_{minority}$$

p_c represents the probability of voting for the Conservative Party and p_l represents the probability of voting for the Liberal Party. The $log(\frac{p_l}{1-p_l})$ and $log(\frac{p_c}{1-p_c})$ are the log odds in both models.

β_0 = is the intercept of the model. It represents the log odds of voting for the candidate or party when all the predictor variables are at their baseline level (which means that the respondent is aged at 0 years, sex category female, speaking the language of English, not living in Quebec, the default income class at low level of income, does not believe in religions, and is a minority other than the White.).

β_1 represents the relationship between the age of the respondent and the log odds of voting for the parties. For every one-unit increase in age (typically one year), there is an expected β_1 change in the log odds of voting for the parties, assuming all other variables are held constant.

β_2 quantifies the average difference in the log odds of voting for the parties between two sex groups (e.g., male and female), controlling for other factors in the model. It shows how sex influences the likelihood of voting for the candidate or party.

For β_3, this coefficient measures the average difference in log odds of voting for the parties between respondents who speak French and English when other predictors stay the same. It reflects how changes in language proficiency between English and French can affect the voting behavior of a respondent, with other factors being equal.

β_4 represents the average difference of being in the western inland provinces (Alberta, Manitoba, Saskatchewan) or not on the log odds of voting for the candidate or party while holding other predictors constant. This could reflect regional differences in voting preferences, with all other variables held constant.

Forβ_5, This coefficient measures the average difference of log odds of voting between the respondent who possesses a low and medium income level. Similarly, β_6 measures the average difference of log odds of voting between the respondent who possesses a low and high-income level when other predictors stay the same. They indicate how changes in income level impact the log odds of voting for the candidate or party. Both coefficients capture the relationship between economic status and voting behavior, controlling for other variables in the model.

β_7 assesses the average difference in the log odds of voting for the parties between the respondents who are atheists or possess religious beliefs with the other predictors fixed. This factor might capture how the difference between a religious believer and a non-religious person on voting preferences.

Finally, β_8 measures the average difference in the log odds of voting for the parties if a respondent is identified as a minority member compared to a non-minority one. This could reflect how membership in certain demographics or ethnic minorities influences voting patterns compared to the majority component of White people in the population of Canadian citizens (Statistics Canada, 2022).

3.3 Post-Stratification

Post-stratification is a method to ensure that the results can accurately represent different groups within a population. It involves adjusting the weights for each estimated parameter within specific post-strata based on their corresponding weights in the census population size (Open AI, 2023).

In order to estimate the proportion of voters who will vote for the Liberal Party and the Conservative Party we will perform two post-stratification analyses by applying the following formula:

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

According to the equation above, $\hat{y}$ is the estimate in each cell, and N_j is the population size of the j^{th} cell based on demographic variables of age, biological sex, language speaking, whether or not living in the particular western provinces, classes of overall household income, religion status and minority identity because we assume that different levels or values of these variables can represent different groups of eligible voters in the Canadian population and the variables are all matched both in the survey data and the census data.

Moreover, the estimate ($\hat{y}$) is $\hat{p}$ that refers to the proportion of voting for the Party. Our goal is to predict the probabilities of popular votes ($\hat{y}^{PS}$) for the Liberal Party and the Conservative Party.

Post-stratification will conduct logistic regression in each cell and use the logistic model to estimate the $\hat{p}$ within each cell. We will create cells based on different age, sex, region, religion, income, language, and minority. Using the logistic regression models described previously, we will estimate the proportion of voters in each cell to be $\hat{p}$. Since there are 2 categories in sex, 2 in western regions, 2 in religion, 3 in income, 2 in language, and 2 in minority, we will have 96 cells for each age group. We will subsequently weight each proportion estimate within each cell by the respective population size of that cell and will then sum those values and divide that by the entire population size, which gives us the estimated probabilities $\hat{y}^{PS}$ of vote for the parties.

4. Results

4.1 Result of Logistic Regression Models

The following are the logistic regression model results (coefficients rounded to 3 decimals). For Liberal

$$log(\frac{\hat{p_l}}{1-\hat{p_l}}) = -2.580 + 0.017x_{age} + 0.030x_{sex} - 0.938x_{French} - 1.115x_{west} + 0.822x_{incomeMedium}$$
$$+0.504x_{incomeHigh} + 0.362x_{religion} - 0.590x_{minority}$$

According to the liberal model For Conservative

$$log(\frac{\hat{p_c}}{1-\hat{p_c}}) = -3.024 + 0.007x_{age} - 0.368x_{sex} - 1.102x_{French} + 1.105x_{west} + 0.512x_{incomeMedium}$$
$$+1.509x_{incomeHigh} + 1.221x_{religion} + 0.030x_{minority}$$

where $\hat{p_c}$ and $\hat{p_l}$ are the estimated probability of voting to the Conservative and Liberal Party.

The following section compares and interprets the coefficients of the same predictors in different models.

$\hat{\beta_0}$ is the estimated intercept of the models. It represents the log odds of voting for the candidate or party when all the predictor variables are at their baseline level is -2.580 for the Liberal and -3.024 for the Conservative (which means that the respondent is aged at 0 years, sex category female, speaking the language of English, not living in the western inland provinces (Alberta, Manitoba, Saskatchewan), the default income class at the low level of income, does not believe in religions, and is a minority other than the White.).

$\hat{\beta_1}$ represents the relationship between the age of the respondent and the log odds of voting for the parties. For every one-unit increase in age (typically one year), there is an expected 0.017 unit increase in the log odds of voting for the Liberal and 0.007 unit of change for the Conservative, assuming all other variables are held constant. We may infer that although the increase in age may increase the probability of voting, the change is similar for both parties.

$\hat{\beta_2}$ quantifies the average difference in the log odds of voting for the parties between two sex groups (e.g., male and female), controlling for other factors in the model. As we can see that the average difference on log odd between the two sex groups in the Liberal is positive at 0.030 but negative at -0.368 for the Conservative, we may expect that males are more likely to vote for the Liberal party and less likely to vote for the Conservative party.

For $\hat{\beta_3}$, this coefficient measures the average difference in log odds of voting for the parties between respondents who speak French and English when other predictors stay the same. The average difference is -0.939 for the Liberal and -1.102 for the Conservative, so French speakers may vote for the Liberal with more likelihood but it seems neither of the two parties will receive more popular votes from French speakers. This is perhaps due to the fact that a sufficient number of voters who use French as their home language are likely to vote for the local party Bloc Québécois (Canada Guide, 2023).

$\hat{\beta_4}$ represents the average difference of being in the western inland provinces (Alberta, Manitoba, Saskatchewan) or not on the log odds of voting is -1.115 for the Liberal but 1.105 for the Conservative while holding other predictors constant. The distinct signs of estimated coefficients imply that voters in the specific western provinces are not likely to vote for the Liberal and vote for the Conservative instead, which contrasts with the other eastern provinces that may contribute more votes to the Liberal.

For $\hat{\beta_5}$, this coefficient measures the average difference of log odds of voting between the respondent who possesses a low and medium household income level, which are both positive for the Liberal at 0.822 and the Conservative at 0.512. Similarly, $\hat{\beta_6}$ measures the average difference of log odds of voting between the respondent who possesses a low and high household income level when other predictors stay the same, which are both positive for the Liberal at 0.504 and the Conservative at 1.509. We can see that the medium and high-income classes people have a higher probability of voting for the election than the ones with low income. In addition, voters with higher income are more likely to vote for the Conservative than the Liberal with

a substantiated difference compared to the medium income class where the change of average differences is not obvious between the Liberal and the Conservative.

$\hat{\beta}_7$ shows that the voters with religion are more likely to vote for both parties. Furthermore, the Conservative Party will have a higher probability of being voted by the religious voters since the average difference of log odd between atheists and theists is 1.221 for the Conservative which is higher than the Liberal Party at 0.362.

Finally, $\hat{\beta}_8$ measures the average difference in the log odds of voting for the parties if a respondent is identified as a minority member compared to a non-minority one. The different signs show that minority people are more likely to vote for the Conservative Party and less likely to vote for the Liberal Party since the average difference of log odd is positive at 0.030, but it is negative at -0.590.

Table 10: Summary of Logistic Regression Model for Liberal

	Estimates	SE	Test statistics	p-value
(Intercept)	-2.580	0.716	-3.604	0.000
age	0.017	0.011	1.497	0.134
sexMale	0.030	0.433	0.069	0.945
languageFR	-0.939	0.484	-1.938	0.053
west	-1.115	0.660	-1.690	0.091
incomeMedium	0.822	0.394	2.084	0.037
incomeHigh	0.504	0.892	0.565	0.572
religion	0.362	0.490	0.739	0.460
minority	-0.590	0.483	-1.223	0.221

Table 11: Summary Logistic Regression Model for Conservative

	Estimates	SE	Test statistics	p-value
(Intercept)	-3.024	0.758	-3.991	0.000
age	0.007	0.011	0.604	0.546
sexMale	-0.368	0.451	-0.815	0.415
languageFR	-1.102	0.585	-1.885	0.059
west	1.105	0.462	2.393	0.017
incomeMedium	0.512	0.391	1.310	0.190
incomeHigh	1.509	0.824	1.832	0.067
religion	1.221	0.554	2.203	0.028
minority	0.030	0.444	0.068	0.946

Table 11.shows the summary statistics of the logistic regression for the Conservative, which includes the estimated coefficients (slopes) of the logistic regression models, the corresponding standard errors, test statistics, and p-values for the coefficient t-tests. The t-tests help us to determine whether there exists a significant linear relationship between the voting decisions of voters and their demographic variables. Using a significance level of 0.05, we can see that p-values for the coefficient estimate of numerical variable age and dummy variables: sex of Male or not, Language of French or not, living in Western inland provinces (Alberta, Manitoba, Saskatchewan) or not, income class at Medium or not, income class at High or not, possess a religion or not, identified as a minority or not. Among these predictors, the western inland provinces' identification and religious identification can form a significant relationship with the voting decision of voters because their p-values are smaller than the 0.05 significant level cutoff. Similarly, Table 10. shows the summary statistics of the logistic regression for the Liberals, and we can see that merely the coefficient of income class at Medium or not is significant with the voting decision.

4.2 Result of post-stratifications

Table 12: Estimated probability of voting for the Parties

Liberal	Conservative
0.2278	0.255

From Table 12, we use post-stratification to predict the overall weighted probability of popular votes for the Liberal Party and the Conservative Party. The table demonstrates that the Conservative Party is going to win the Liberal Party in the next election because the estimated turnout for the Conservative Party is 0.255 which is slightly larger than the estimated turnout of the Liberal Party at 0.2278. From the statistics perspective of this research, the result of logistic regression models has shown that males vote less for the Conservative Party than females, whereas females are the dominant voting population in the election because Graph 3. in the data summary indicates a higher proportion of females than males in both the survey data of elections and the census data of the general social survey. From the perspective of economics and society, Canada is calming down from the inflation period and is expected to enter the economic slowdown period in 2025. The relatively high unemployment and decrease in productivity may trigger people's attitude toward saving rather than spending, so the economic market will change from an active state to a relatively smooth one. Therefore, the public wish a less progressive government to regulate society, and the Conservative Party will be a more suitable candidate (Aiello, November 2023).

5 Conclusion

5.1 Key Results and Big Picture

Our goal of this research is to predict the probability of winning between the Liberal and the Conservative Party in the 2025 Canadian Federal Election. Specifically, we hypothesize that there is a relationship between the probability of voting for parties and the demographic variables including age, sex, region, religion, income, language, and minority, and the turnover of the Liberal Party is higher than the Conservative Party, but the result illustrates that the turnout of Conservative is estimated to be 25.5% which is slightly greater than the turnout of Liberal that is estimated to be 22.78%. This result broke the stereotype that the Liberal Party is always sitting on the political throne.

In methodology, we use logistic regression models to predict the proportion of voting for Conservatives and Liberals as this response variable is binary and apply post-stratification on the model to estimate this proportion to get a more representative result for the census.

5.2 Weaknesses

The model our team performed cannot involve all variables that affect the election outcome, such as education level which can affect individual voting decisions. McMaster University reported that educated individuals tend to vote for the left-leaning parties (2023).

One weakness was made as we modified the categories of chosen variables. The BC province was excluded which used to involve in western regions of the Survey Data. We also merged indigenous groups into visible minorities. For the independent variable of gender, we adjusted into sex and removed observations with "other" options. Those modifications can potentially increase bias and reduce the model predictions, outputting in-stable predicted election outcomes. Missing data from the Northern Territories could result in non-response bias and decrease the model precision.

Confounding factors may also exist which affect both variables and the predicted outcomes. The social status of current candidates may be a confounding factor since they often have an advantage in reputation, experience, and resources, which could affect their popularity and election results. Economic factors such as unemployment, inflation, and GDP growth can be confounders as well. Changes in economic conditions can affect public criticism toward candidates and related election outcomes.

In addition, some p-values of included variable coefficients are not statistically significant. Therefore, we cannot fully explain that those variables are associated with the outcome of voting probability.

5.3 Future works and Wrap up

For the future step, We will consider removing variables with non-significant p-values from the model. This simplifies the model and may improve its overall performance. We will evaluate and add other variables to identify the fitted model for predicting the probability of voting with the inclusion of confounding adjustment.

We modified the classifications of variables to present data more easily. In the future, we can classify cells more specifically to make more generalized and precious estimates of the outcomes of interest. We can also collect relative data from the Northern Territories to reduce the non-response bias.

Bibliography

Aiello, R. (2023, November 16). Trudeau's Liberals trailing Poilievre's conservatives in power index: Nanos. CTVNews. https://www.ctvnews.ca/politics/trudeau-s-liberals-trailing-poilievre-s-conservatives-in-power-index-nanos-1.6648263

Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: April 4, 1991)

Canada Guide. (2023). *Canadian Political Parties*. [https://thecanadaguide.com/government/political-parties/] (Last Accessed: November 11, 2023)

Canada, E. (n.d.). Facts about voter registration, citizenship and voter ID. – Elections Canada. https://www.elections.ca/content.aspx?section=med&dir=c76%2Fcitizen&document=index&lang=e

Center for American Women and Politics (CAWP). (2023). *Gender Differences in Voter Turnout* New Brunswick, NJ: Center for American Women and Politics, Eagleton Institute of Politics, Rutgers University-New Brunswick. [https://cawp.rutgers.edu/facts/voters/gender-differences-voter-turnout] (Accessed November 22, 2023)

Country's religious and ethnocultural diversity. *Statistics Canada* https://www150.statcan.gc.ca/n1/daily-quotidien/221026/dq221026b-eng.htm (Last Accessed: November 12, 2023)

Coward, H., & Slater, P., & Chagnon, R. (2022). Religion. In The Canadian Encyclopedia. Retrieved from https://www.thecanadianencyclopedia.ca/en/article/religion

Dippel S. (2016). *As Stephen Harper leaves politics, record shows mixed results for Calgary*. CBC. [https://www.cbc.ca/news/canada/calgary/harper-resigns-mp-calgary-analysis-1.3734081]

Elections Canada. (2023). *The electoral system of Canada*. [https://www.elections.ca/content.aspx?section=res&dir=ces&document=part2&lang=e]

Erickson, L., & O'Neill, B. (2002). The Gender Gap and the Changing Woman Voter in Canada. International Political Science Review / Revue Internationale de Science Politique, 23(4), 373–392. http://www.jstor.org/stable/1601539

Government of Canada, (2022). *Provinces and territories - Intergovernmental Affairs*. [https://www.canada.ca/en/intergovernmental-affairs/services/provinces-territories.html]

Grenier E. (2017). *French a major advantage for Conservative leadership contenders*. CBC. [https://www.cbc.ca/news/politics/grenier-conservative-leadership-french-1.3940956] (Accessed November 22, 2023)

Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: April 4, 1991)

Keller, J. (2021). *How Canada voted: A look at how the parties, the regions and the issues shaped our next government*. The Globe and Mail. [https://www.theglobeandmail.com/politics/article-takeaways-about-who-voted-and-the-issues-and-tensions-that-helped/]

Liberal Party of Canada. (n.d.). *2021 Platform*. [https://liberal.ca/our-platform/]

Liberal Party. (2019). *The Canadian Encyclopedia*. [https://www.thecanadianencyclopedia.ca/en/article/liberal-party]

McMaster University. (2023). *Analysis: Educated voters in Canada tend to vote for left-leaning parties while richer voters go right*. McMaster University. [https://brighterworld.mcmaster.ca/articles/analysis-educated-voters-in-canada-tend-to-vote-for-left-leaning-parties-while-richer-voters-go-right/]

Members of Parliament. (2023). *Current Members*. [https://www.ourcommons.ca/Members/en]

OpenAI. (2023). *ChatGPT (September 13 version) [Large language model]*. https://chat.openai.com/chat (Last Accessed: September 13, 2023)

Parliament of Canada. (2006). *Amendment to Canada Elections Act*. [https://www.parl.ca/]

RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Simon Fraser University. (2021). *Canadian Election Results by Party 1867 to 2021*. https://www.sfu.ca/~aheard/elections/1867-present.html (Last Accessed: November 11, 2023)

Statistics Canada. (2012). *Factors associated with voting*. https://www150.statcan.gc.ca/n1/pub/75-001-x/2012001/article/11629-eng.htm#a4 (Last Accessed: November 12, 2023)

Statistics Canada. (2020). *General Social Survey - Social Identity (SI): Detailed information for 2020 (Cycle 35)*. [https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5024] (Last Accessed: November 20, 2023)

Statistics Canada. (2021). *Visible minority of person*. [https://www23.statcan.gc.ca/imdb/p3Var.pl?Function=DECI&Id=257515]

Statistics Canada. (2022). Chapter 6: Political participation, civic engagement and caregiving among youth in Canada. *Statistics Canada*. [https://www150.statcan.gc.ca/n1/pub/42-28-0001/2021001/article/00006-eng.htm] (Last Accessed: November 20, 2023)

Statistics Canada. (October, 2022). The Canadian census-A rich portrait of the

Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. (2021). *The 2021 Canadian Election Study*. Consortium on Electoral Democracy. [dataset]

The Globe and Mail. (n.d.). Motley Fool Press Releases & Markets News. https://www.theglobeandmail.com/investing/markets/markets-news/Motley%20Fool/19986251/does-your-income-make-you-upper-class-middle-class-or-lower-class/

Tossutti, L. & Elections Canada. (2007). *La participation électorale des membres des communautés ethnoculturelles*. Élections Canada

Uppal, S. & LaRochelle-Côté S. (February, 2012). *Factors associated with voting*. Statistics Canada. https://www150.statcan.gc.ca/n1/en/pub/75-001-x/2012001/article/11629-eng.pdf?st=nuV9FTRH (Last Accessed: November 12, 2023)

Appendix

Generative AI Statement

We used the following generative artificial intelligence (AI) tool: Chat GPT3.5 (Open AI, 2023). We used the tool only in the Method section of this assignment and I gave it the following prompt of `what is the definition of post-stratification` and it gave me the explanations of post-stratification.

`Post-stratification is a statistical technique used in survey sampling and analysis. It involves the division of a population into groups or strata based on certain characteristics after the data has been collected. This technique is used to improve the precision of estimates by ensuring that the sample reflects the known population characteristics more accurately.`

`Here's how it generally works:`

`Initial Data Collection: Surveys or data collection exercises are conducted without considering specific stratifications. This initial dataset might not completely represent the population's various characteristics.`

`Stratification: After collecting the data, it is sorted or divided into subgroups or strata based on certain variables such as age, gender, income, or any other relevant characteristic.`

`Analysis and Weighting: Each stratum is then analyzed separately. Statistical weights are often applied to adjust for differences in the population sizes of the strata. This weighting helps in making estimates more accurate by giving appropriate importance to each subgroup according to its representation in the population.`

`Estimation: Finally, estimates or inferences are made by combining the results from all strata according to the population distribution.`

`Post-stratification is particularly useful when the initial sampling methods do not ensure a balanced representation of certain characteristics within the sample. By dividing the data into strata after collection and adjusting the analysis accordingly, it helps improve the accuracy and representativeness of the findings, especially when dealing with diverse or unevenly distributed populations.`

Then we paraphrased the explanations of post-stratification into section 3.3 of Methods.

Supplementary Materials

The following tables show the glimpse of cleaned survey and census data.

Table 13: Glimpse of Survey Data

age	vote_liberal	vote_conservative	sex	language	west	religion	minority	income
34	1	0	Male	EN	0	0	0	Low
76	0	0	Female	FR	0	1	0	Low
63	0	0	Female	EN	0	0	0	Low
64	1	0	Female	EN	1	1	0	Medium
65	1	0	Male	EN	1	1	0	Medium
42	1	0	Female	EN	0	0	0	Low
71	1	0	Female	EN	0	1	1	Medium
71	1	0	Female	EN	0	1	0	Medium
19	0	0	Male	EN	0	0	0	Low
29	0	0	Female	FR	0	1	1	Medium
63	0	0	Male	FR	0	1	0	Low
57	0	1	Female	EN	1	1	1	Medium
19	1	0	Female	FR	0	1	0	Low
25	0	0	Male	FR	0	0	0	Low
22	0	0	Female	EN	0	1	1	High
26	0	0	Female	EN	1	0	1	Low
45	0	1	Male	FR	0	1	1	Medium
35	0	0	Female	EN	0	0	0	Medium
25	0	0	Female	EN	1	0	0	Low
20	0	1	Female	EN	1	1	0	Low
92	0	1	Female	EN	0	1	0	Low
22	0	1	Female	EN	0	1	0	Medium
21	0	0	Male	EN	0	1	1	Low
81	0	0	Male	FR	0	1	0	Low
47	1	0	Male	FR	0	0	0	Medium
46	0	0	Female	FR	0	0	0	Medium
39	0	0	Male	FR	0	1	0	Medium
75	0	0	Male	EN	0	0	1	Low
46	0	0	Female	EN	0	1	1	Medium
57	0	1	Female	EN	1	1	0	Medium

Table 14: Glimpse of Census Data

age	sex	language	west	income	religion	minority
53	Female	FR	0	Low	1	0
51	Male	EN	1	Medium	1	0
64	Female	FR	0	Medium	1	0
80	Female	EN	1	Medium	1	0
28	Male	FR	0	Medium	1	0
63	Female	FR	0	Medium	1	0
59	Female	EN	0	Low	1	0
80	Female	FR	0	Low	1	0
64	Female	EN	0	Low	1	0
25	Male	EN	1	Low	1	0
40	Female	EN	1	Low	1	0
57	Female	EN	0	Medium	1	0
27	Female	FR	0	Medium	1	0
67	Female	FR	0	Low	1	0
69	Male	EN	0	High	1	0
34	Female	EN	0	Medium	1	0
73	Male	FR	0	Medium	1	0
65	Male	EN	0	High	1	0
49	Male	EN	0	Medium	0	0
64	Female	FR	0	Low	1	0
78	Male	EN	0	Medium	0	1
31	Female	EN	0	Medium	0	1
44	Male	EN	1	Medium	1	0
41	Female	FR	0	High	1	0
80	Male	EN	1	Low	1	0
80	Female	FR	0	Low	1	0
71	Male	EN	0	Medium	1	0
72	Male	EN	0	Low	1	0
75	Female	EN	0	Medium	1	0
19	Male	EN	1	High	1	0

Graph 4: Distributions of whether from western provinces: Survey vs Census