A2

STA304 - Fall 2023 -Assignment 2

Group 120:Ruilin Peng, Yiwei(Johnson) Guo

11/7/2023

1. Introduction

<Here you should have a few paragraphs of text introducing the problem, getting the reader interested/ready for the rest of the report.>

<Introduce terminology.>

<Highlight hypotheses.>

<Optional: You can also include a description of each section of this report as a last paragraph.>

1.1 The research question

What will be the predicted outcome for the party that receive the highest-amount votes in the upcoming election 2025?

1.2 Importance of the research

Free and fair elections are considered as important features contributing a healthy democracy. Fairness of the election could be ensured as citizen can vote to reflect their own wills without political interference. These conditions can contribute to a durable democracy in Canada and also improve 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 outcome helps encourage citizens to track the political process and promote their participation. Political parties and candidates can learn about competitors' voting tendency and priorities and concerns of voters so adjusting their electoral strategies.

1.3 Terminology

balabala..

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. 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). 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 of Canada is going to win the most popular votes in the 2025 Canadian Election based on the information of 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). 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.

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.

2. Data

2.1 Data description

<Type here a paragraph introducing the data, its context and as much info about the data collection process that you know.>

2.2 Data cleaning process

<Type here a summary of the cleaning process (**only add in stuff beyond my original gss_cleaning.R code**). You only need to describe additional cleaning that you and your group did.>] You will need to describe the cleaning you do to the survey data as well.

2.3 Data summary

<Remember, you may want to use multiple datasets here, if you do end up using multiple data sets, or merging the data, be sure to describe this in the cleaning process and be sure to discuss important aspects of all the data that you used.>

<Include a description of the important variables.>

Response Variable vote_

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

gender: This variable represents the gender of the observations

language:

min	Q1	median	Q3	max	IQR	mean	sd
18	35	53	65	92	30	50.57546	17.99608

Table 2: Statistics about the frequency and proportion of male and female voters in 2021 election survey data

Var1	Freq
Female	554
Male	261

Sex	Number of Voters	Proportion
Female	554	0.68
Male	261	0.32

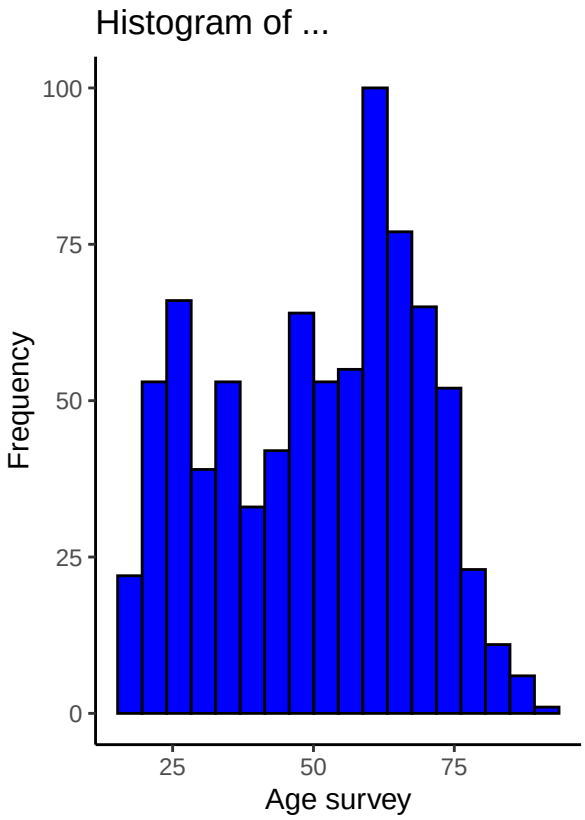
Var1	Freq
EN	622
FR	193

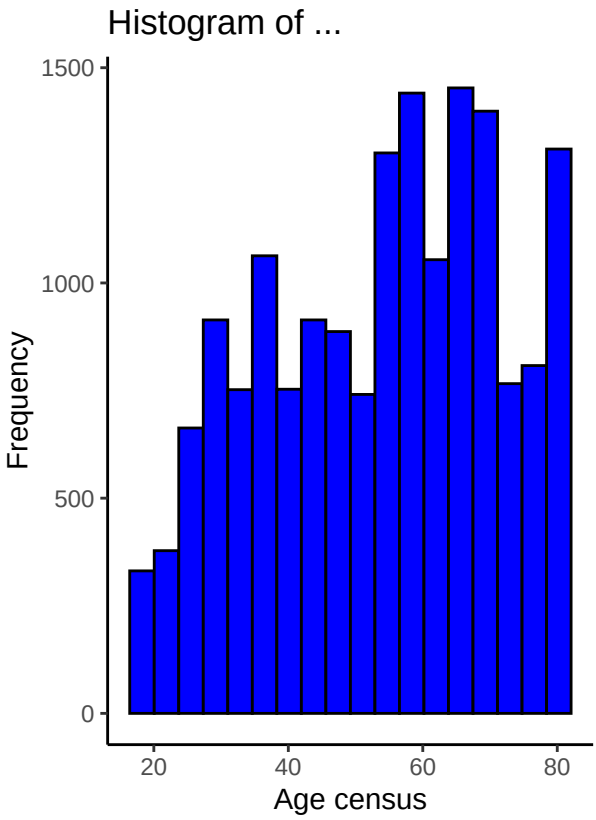
Var1	Freq
High	603

Var1	Freq
Low	120
Medium	92

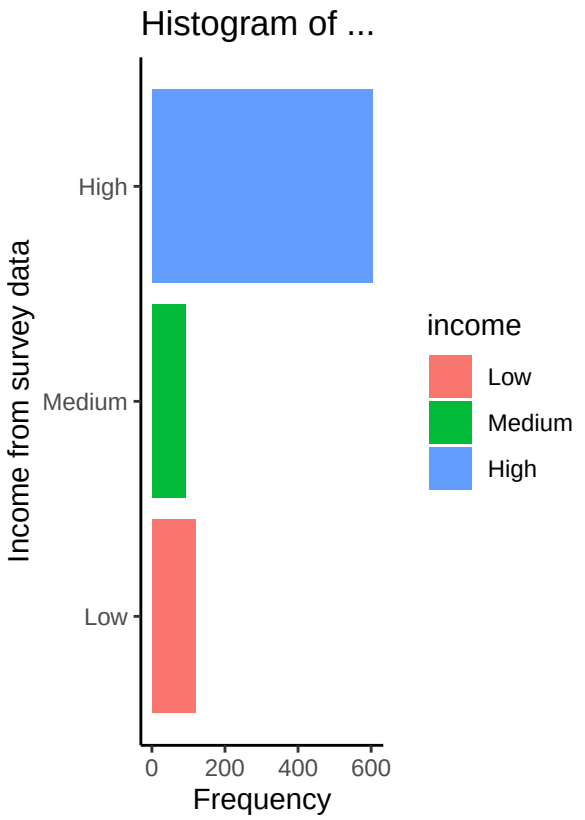
<Include a description of the numerical summaries. Remember you can use `r` to use inline R code.>

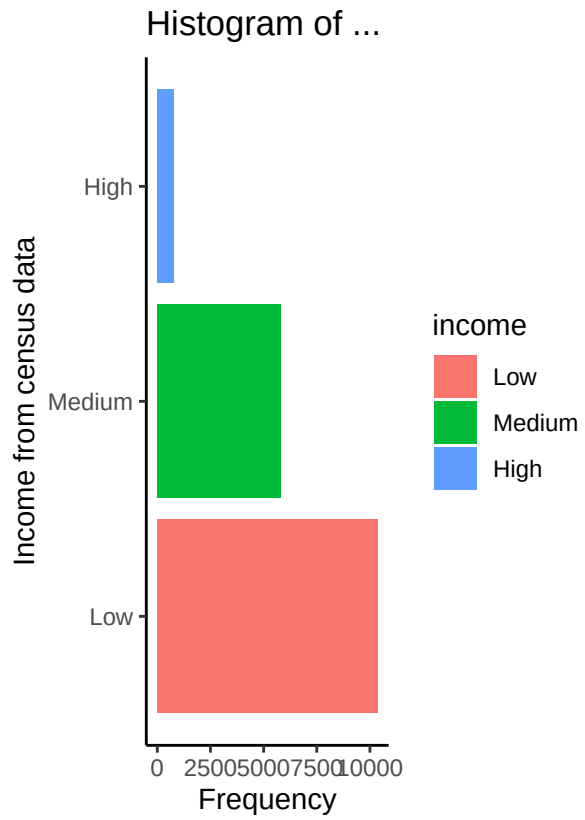
```
survey_data1$income <- factor(survey_data1$income, levels = c("Low", "Medium", "High"))
census_data1$income <- factor(census_data1$income, levels = c("Low", "Medium", "High"))
# Use this to create some plots. Should probably describe both the sample and population.
survey_age <- survey_data1 %>% ggplot(aes(x = age)) + geom_histogram(fill = "blue", color = "black", bi
  labs(x = "Age survey", y = "Frequency", title = "Histogram of ...")

census_age <- census_data1 %>% ggplot(aes(x = age)) + geom_histogram(fill = "blue", color = "black", bi
  labs(x = "Age census", y = "Frequency", title = "Histogram of ...")

survey_income <- survey_data1 %>% ggplot(aes(x = income, fill = income)) +
  geom_bar() + theme_classic() + coord_flip()+
  labs(x = "Income from survey data", y = "Frequency", title = "Histogram of ...")

census_income <- census_data1 %>% ggplot(aes(x = income, fill = income)) +
  geom_bar() + theme_classic() + coord_flip()+
  labs(x = "Income from census data", y = "Frequency", title = "Histogram of ...")

survey_gender_count <- table(survey_data1$gender)

library(gridExtra)
library(patchwork)
survey_age | census_age
```



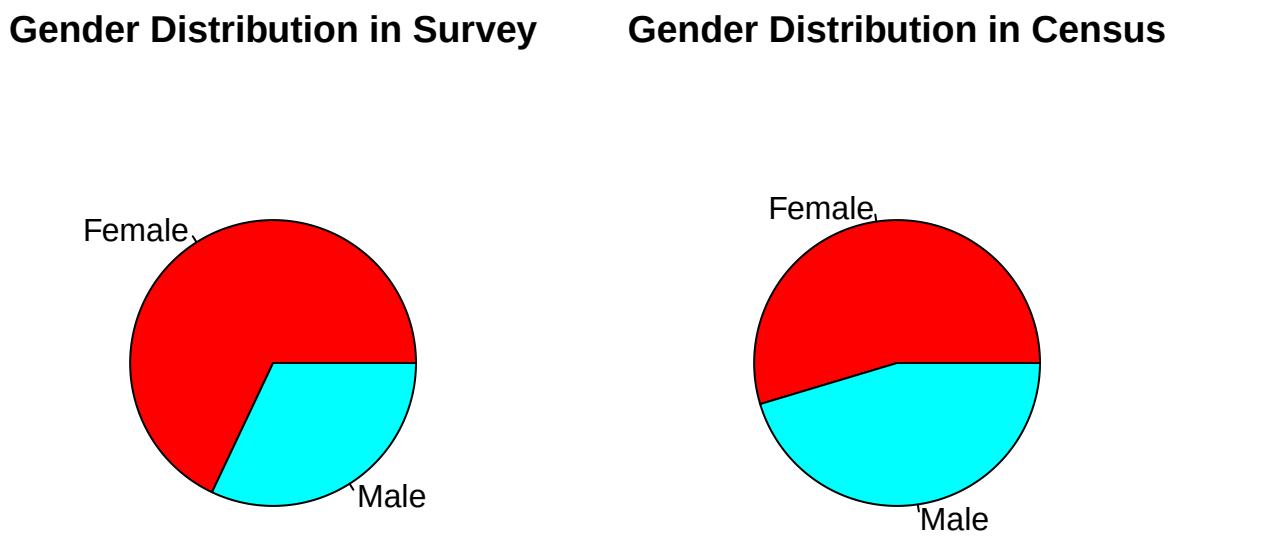



```
survey_income | census_income
```





```
par(mfrow = c(1, 2))
pie(survey_gender_count, main = "Gender Distribution in Survey", col = rainbow(length(survey_gender_cou

census_gender_count <- table(census_data1$gender)
pie(census_gender_count, main = "Gender Distribution in Census", col = rainbow(length(census_gender_cou
```




```
par(mfrow = c(1, 1))
```

3. Methods

<Include some text introducing the methodology, maybe restating the problem/goal of this analysis.>

3.1 Model selection and rationale

<Include some text introducing the methodology, maybe restating the problem/goal of this analysis.> Our study applies the logistic regression model to estimate the probability of voting of each voter for the Liberal Party and Conservative Party. Then, we 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 in both the census and the survey data, and the variability of the variables is similar in survey data compared to the census data as shown in the plots of Data Summary section (...). 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.

$$log(\frac{p_c}{1-p_c}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{French} + \beta_4 x_{west} + \beta 5 x_{incomeMedium} + \beta 6 x_{incomeHigh} + \beta 7 x_{religion} + \beta 8 x_{minority}$$

$$log(\frac{p_l}{1-p_l}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{French} + \beta_4 x_{west} + \beta 5 x_{incomeMedium} + \beta 6 x_{incomeHigh} + \beta 6 x_{religion} + \beta 7 x_{minority}$$

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).

```
# Creating the Model
model1 <- glm(vote_liberal ~ age + gender + language + west + income + religion + minority, data=survey


model2 <- glm(vote_conservative ~ age + gender + language + west + income + religion + minority, data=su

summary(model1)
```

```
##
## Call:
## glm(formula = vote_liberal ~ age + gender + language + west +
##     income + religion + minority, family = "binomial", data = survey_data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9905  -0.7389  -0.6055  -0.4375   2.2393
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.267205   0.419979  -5.398 6.72e-08 ***
## age           0.009968   0.005425   1.837  0.06616 .
## genderMale   -0.227448   0.199097  -1.142  0.25329
## languageFR   -0.484627   0.225499  -2.149  0.03162 *
## west         -0.866868   0.281246  -3.082  0.00205 **
## incomeMedium  0.810730   0.380566   2.130  0.03314 *
## incomeHigh    0.643312   0.304973   2.109  0.03491 *
## religion      0.152323   0.216766   0.703  0.48224
## minority     -0.035841   0.203118  -0.176  0.85994
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 823.91  on 814  degrees of freedom
## Residual deviance: 798.18  on 806  degrees of freedom
## AIC: 816.18
##
## Number of Fisher Scoring iterations: 4
```

```
summary(model2)
```

```
##
## Call:
## glm(formula = vote_conservative ~ age + gender + language + west +
##     income + religion + minority, family = "binomial", data = survey_data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3418  -0.6977  -0.5380  -0.3608   2.5210
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.357670   0.454871  -7.382 1.56e-13 ***
## age           0.017478   0.005735   3.048 0.002307 **
## genderMale    0.326888   0.195041   1.676 0.093739 .
## languageFR   -0.637058   0.257533  -2.474 0.013372 *
## west          0.798567   0.220004   3.630 0.000284 ***
## incomeMedium  0.646471   0.379296   1.704 0.088307 .
## incomeHigh    0.356121   0.295946   1.203 0.228849
## religion      0.734006   0.250475   2.930 0.003385 **
## minority     -0.179578   0.214871  -0.836 0.403298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 810.08  on 814  degrees of freedom
## Residual deviance: 752.12  on 806  degrees of freedom
## AIC: 770.12
##
## Number of Fisher Scoring iterations: 4
```

```
#predict(model1,type = "response")

# Model Results (to Report in Results section)
# summary(model)
# OR
# broom::tidy(model)

### Don't show the results/output here...
```

3.3 Post-Stratification

<In order to estimate the proportion of voters…..>

<To put math/LaTeX inline just use one set of dollar signs. Example: $\hat{y}^{PS}$ >

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.

In order to estimate the proportion of voters who will vote for … we will perform a post-stratification analysis by applying the following formula:

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

- N_j.. - $\hat{y}_j$…

$\hat{y}$ is the estimate in each cell, and N_j is the population size of the j^{th} cell based on demographics.
The estimated ($\hat{y}$) is $\hat{p}$ that refer to the proportion of voting for… Post-stratification will conduct logistic regression in each cell and use the logistic model to estimate the $\hat{p}$ within that cell. We firstly will create cells based on different ages, sex, and working status. Using the model described in the previous sub-section, we will estimate the proportion of voters in each cell. _Since there are ? categories in sex, ? in region, ? in education, and ? in religion, we will have ? cells for each age.
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.

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

4. Results

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

$$log(\frac{p_l}{1-p_l}) = -2.267 + 0.010x_{age} - 0.227x_{sex} - 0.485x_{French} - 0.867x_{west} + 0.811x_{incomeMedium}$$

$$+0.643x_{incomeHigh} + 0.152x_{religion} - 0.036x_{minority}$$

$$log(\frac{p_c}{1-p_c}) = -3.358 + 0.017x_{age} + 0.327x_{sex} - 0.637x_{French} + 0.799x_{west} + 0.646x_{incomeMedium}$$

$$+0.356x_{incomeHigh} + 0.734x_{religion} - 0.180x_{minority}$$

```
# Here I will perform the post-stratification calculation for Conservative
census_data_counts <- census_data1 %>% group_by(age, gender, language, west, income, religion, minority
  summarise(N=n())

census_data_counts$estimate2 <-
  model2 %>%
  predict(newdata = census_data_counts, type = "response") # get proportion
# prof's code can only provide log ratio

result_conservative <- census_data_counts %>%
  mutate(conservative_predict_prop = estimate2*N) %>% ungroup() %>%
  summarise(conservative_predict = sum(conservative_predict_prop)/sum(N))

table <- tibble(result_liberal, result_conservative)

knitr::kable(table, col.names = c("Liberal", "Conservative"), caption = "Estimated probability of votin
```

Table 6: Estimated probability of voting for the Parties

Liberal	Conservative
0.1711	0.2149

<Here you present your results. You may want to put them into a well formatted table. Be sure that there is some text describing the results.>

<Note: Alternatively you can use the `knitr::kable` function to create a well formatted table from your code. See here: https://rmarkdown.rstudio.com/lesson-7.html.>

<Remember you can use `r` to use inline R code.>

<Include an explanation/interpretation of the visualizations. Make sure to comment on the appropriateness of the assumptions/results.>

5. Conclusions

<Here you should give a summary of the Hypotheses, Methods and Results>

<Highlight Key Results.>

<Talk about big picture.>

<Comment on any Weaknesses.>

<End with a concluding paragraph to wrap up the report.>

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: April 4, 1991)
2. RStudio Team. (2020). *RStudio: Integrated Development for R.* RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: April 4, 1991)

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

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

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

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

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)

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

Appendix

Generative AI Statement

Here is where you can explain your usage of Generative AI tool(s). Be sure to reference it. For instance, including something like:

I used the following generative artificial intelligence (AI) tool: Bing AI Version 2.0 for Chrome [4]. I used the tool only in the Results section of this assignment and I gave it the following prompt of `What should I eat for breakfast?` and it gave me a list of 10 breakfast items which I then asked it to: `Please only list breakfast items that do not include eggs`. I then chose my 3 favourite items from the produced list and included those in the Results section.

Supplementary Materials

<Here you can include any additional plots, tables, derivations, etc.>