ADD A DESCRIPTIVE TITLE

STA304 - Fall 2023 -Assignment 2

GROUP NUMBER: ADD YOUR NAMES HERE

Insert Date Here

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.>

```
survey_data %>% group_by(cps21_votechoice) %>% summarise(n = n())
```

```
## # A tibble: 8 x 2
##   cps21_votechoice                          n
##   <dbl+lbl>                             <int>
## 1  1 [Liberal Party]                     3890
## 2  2 [Conservative Party]                3637
## 3  3 [ndp]                               2828
## 4  4 [Bloc Québécois]                    1295
## 5  5 [Green Party]                        352
## 6  6 [Another party (please specify)]     484
## 7  7 [Don't know/ Prefer not to answer]  2091
## 8 NA                                     6391
```

```
survey_data %>% group_by(cps21_province) %>% summarise(n = n())
```

```
## # A tibble: 13 x 2
##    cps21_province                    n
##    <dbl+lbl>                     <int>
##  1  1 [Alberta]                   2527
##  2  2 [British Columbia]          2329
##  3  3 [Manitoba]                   795
##  4  4 [New Brunswick]              410
##  5  5 [Newfoundland and Labrador]  199
##  6  6 [Northwest Territories]       15
##  7  7 [Nova Scotia]                530
##  8  8 [Nunavut]                      4
##  9  9 [Ontario]                   7309
## 10 10 [Prince Edward Island]        59
## 11 11 [Quebec]                    6317
## 12 12 [Saskatchewan]               446
## 13 13 [Yukon]                       28
```

o Alberta (1) o British Columbia (2) o Manitoba (3) o New Brunswick (4) o Newfoundland and Labrador (5) o Northwest Territories (6) reject o Nova Scotia (7) o Nunavut (8) reject o Ontario (9) o Prince Edward Island (10) o Quebec (11) o Saskatchewan (12) o Yukon (13) reject

```
census_data %>% group_by(province) %>% summarise(n = n())
```

```
## # A tibble: 10 x 2
##    province                      n
##    <chr>                     <int>
##  1 Alberta                    1728
##  2 British Columbia           2522
##  3 Manitoba                   1192
##  4 New Brunswick              1337
##  5 Newfoundland and Labrador  1094
##  6 Nova Scotia                1425
##  7 Ontario                    5621
##  8 Prince Edward Island        708
##  9 Quebec                     3822
## 10 Saskatchewan               1153
```

```
survey_data %>% group_by(cps21_age) %>% summarise(n = n())
```

```
## # A tibble: 79 x 2
##    cps21_age     n
##        <dbl> <int>
##  1        18    92
##  2        19   187
##  3        20   209
##  4        21   210
##  5        22   217
##  6        23   224
##  7        24   264
##  8        25   271
##  9        26   296
## 10        27   300
## # i 69 more rows
```

```
census_data %>% group_by(age) %>% summarise(n = n())
```

```
## # A tibble: 651 x 2
##      age     n
##    <dbl> <int>
##  1  15       7
##  2  15.1     9
##  3  15.2    13
##  4  15.3    18
##  5  15.4     7
##  6  15.5    15
##  7  15.6     7
##  8  15.7     9
##  9  15.8    21
## 10  15.9     9
## # i 641 more rows
```

```
sum(is.na(survey_data$cps21_children))
```

```
## [1] 0
```

```
survey_data %>% group_by(cps21_children) %>% summarise(n = n()) #remove 7 Don't know/ Prefer not to ans
```

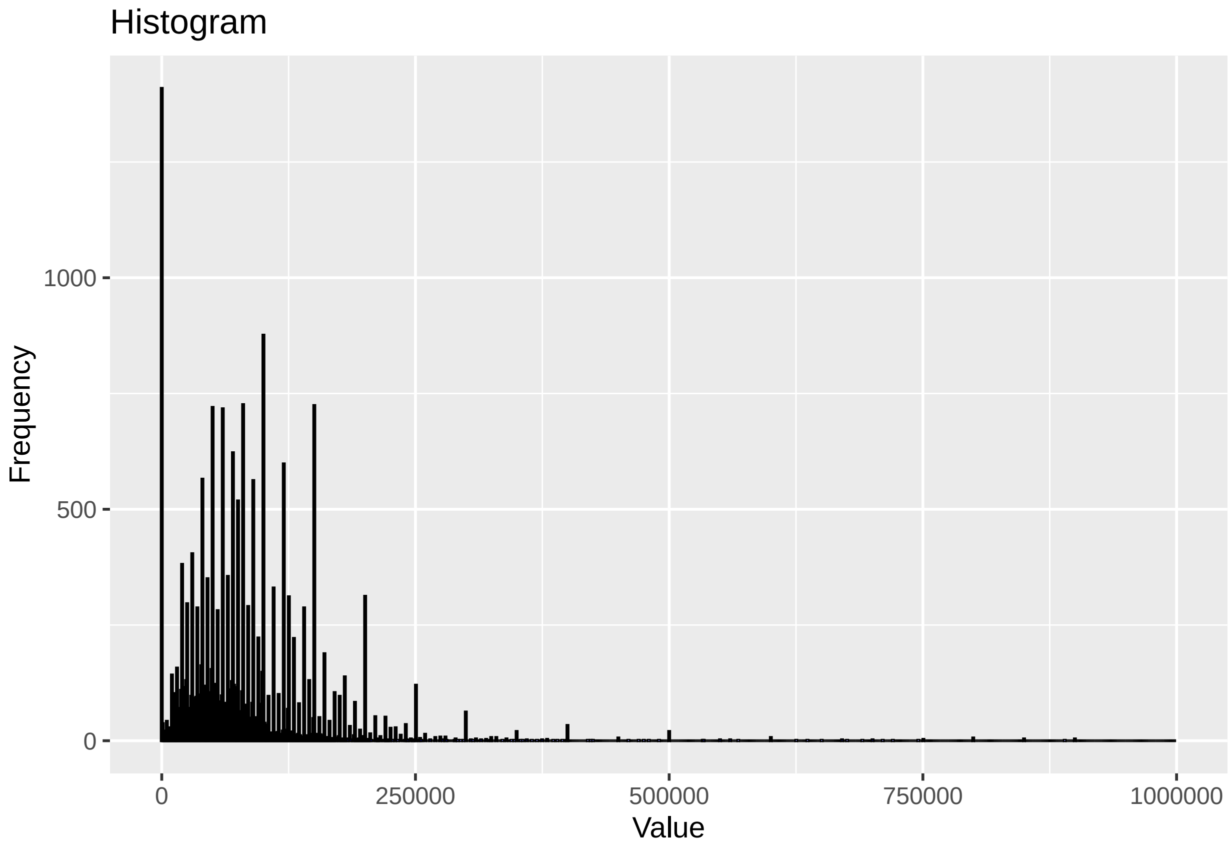
```
## # A tibble: 7 x 2
##   cps21_children                             n
##   <dbl+lbl>                              <int>
## 1 1 [0]                                   8520
## 2 2 [1]                                   3403
## 3 3 [2]                                   5800
## 4 4 [3]                                   2252
## 5 5 [4]                                    634
## 6 6 [5 or more]                            284
## 7 7 [Don't know/ Prefer not to answer]      75
```

```
# in survey_data: 1 means 0 child, 2 means 1 child, ... 6 means five or more kids. 7 means unknow
census_data %>% group_by(total_children) %>% summarise(n = n())
```

```
## # A tibble: 9 x 2
##   total_children     n
##            <dbl> <int>
## 1              0  6224
## 2              1  2907
## 3              2  6199
## 4              3  3156
## 5              4  1259
## 6              5   468
## 7              6   184
## 8              7   186
## 9             NA    19
```

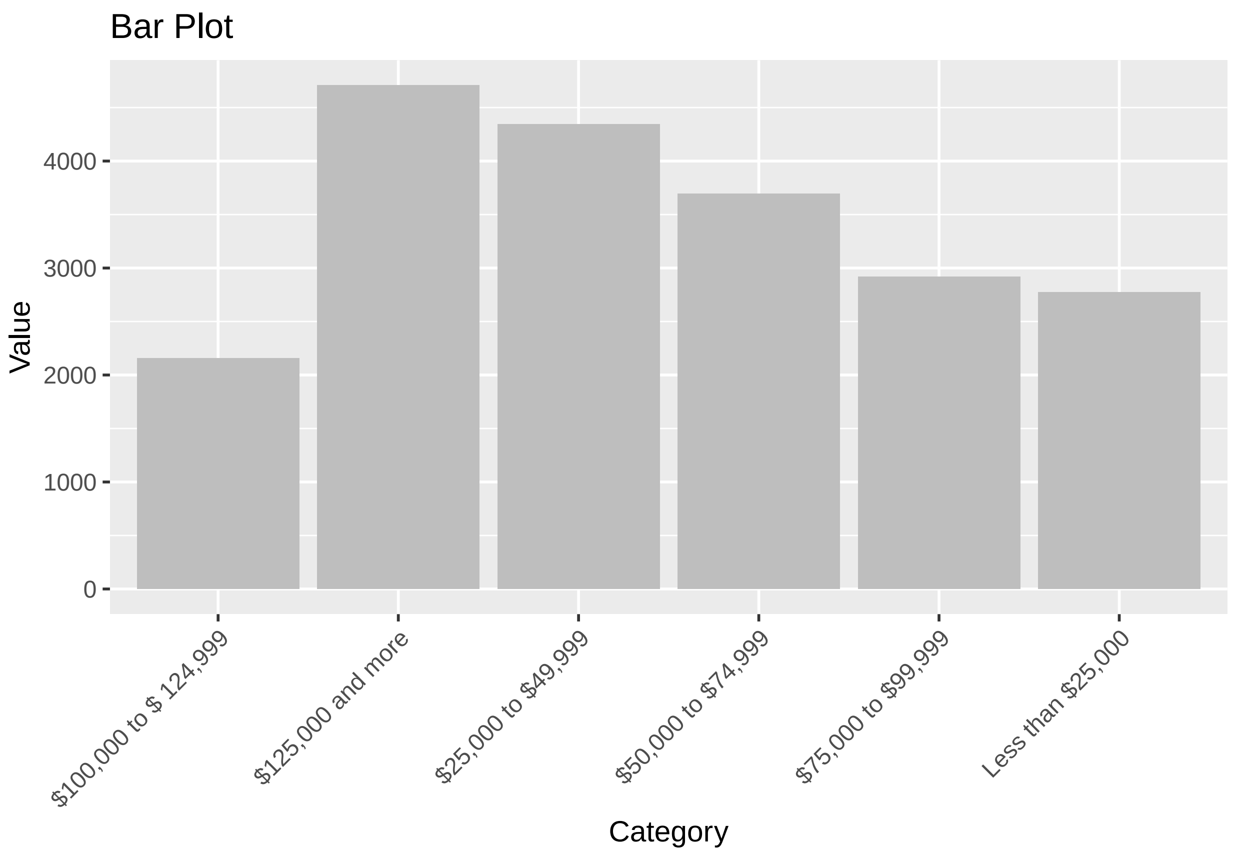
```
df_sv_income <- tibble(x=survey_data$cps21_income_number) # famaily income

ggplot(data = df_sv_income, aes(x = x)) +
  geom_histogram(bins = 1000,fill = "blue", color = "black") +
  xlim(c(-1000,1000000)) +
  labs(title = "Histogram",
       x = "Value",
       y = "Frequency")
```



```r
df_ces_income <- census_data %>% group_by(income_family) %>% summarise(n = n()) # famaily income

ggplot(df_ces_income, aes(x = income_family, y = n)) +
  geom_bar(stat = "identity", fill = "grey") +
  labs(title = "Bar Plot",
       x = "Category",
       y = "Value") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
df_ces_income
```

```
## # A tibble: 6 x 2
##   income_family             n
##   <chr>                 <int>
## 1 $100,000 to $ 124,999  2158
## 2 $125,000 and more      4707
## 3 $25,000 to $49,999     4345
## 4 $50,000 to $74,999     3696
## 5 $75,000 to $99,999     2921
## 6 Less than $25,000      2775
```

Data

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

<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.

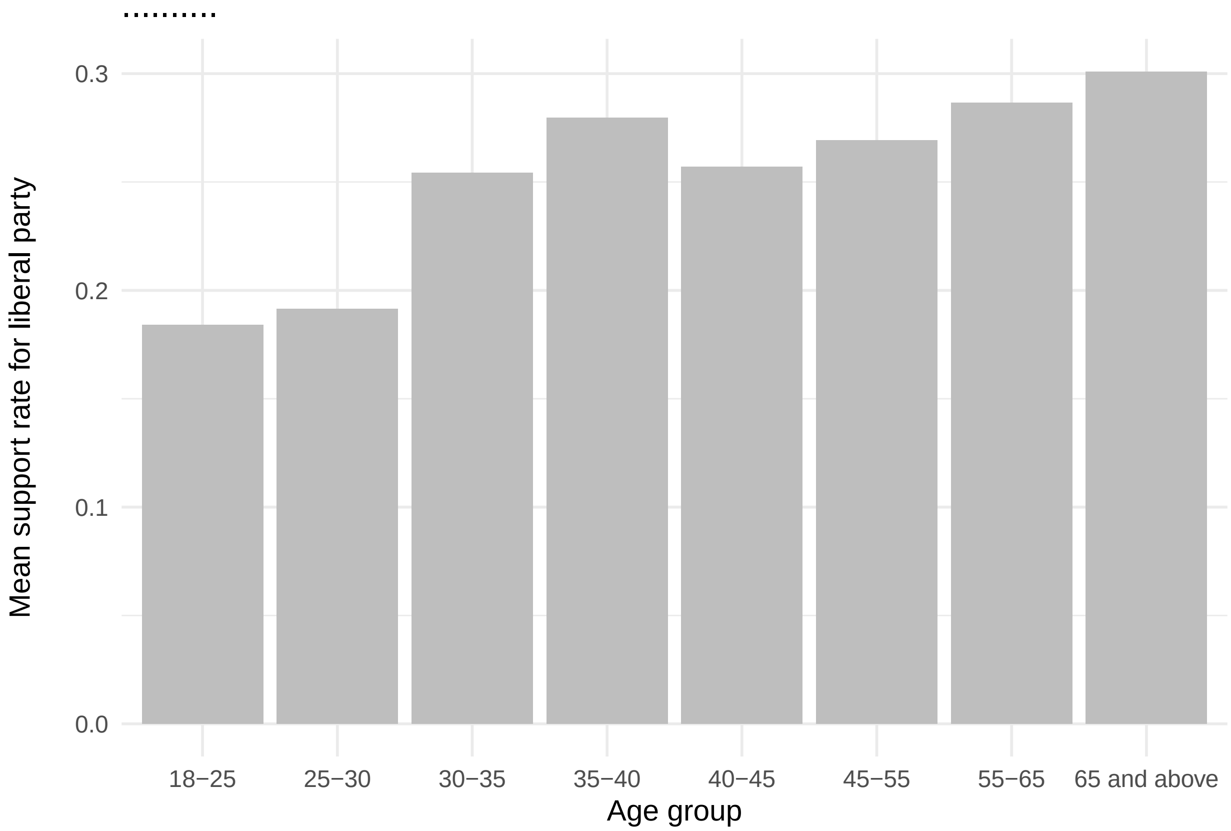
<Remember, you may want to use multiple datasets here, if you do end up using multiple data sets, or merging the data, be sure to describe this in the cleaning process and be sure to discuss important aspects of all the data that you used.>

<Include a description of the important variables.>

```
# Use this to calculate some summary measures.
df_agetest <-
```

```
  survey_data_clean %>%
  group_by(age) %>%
    summarise(p=mean(vote_liberal))

ggplot(df_agetest,
       aes(y= age, x = p))+
  geom_bar(stat = "identity", fill = "grey") +
  coord_flip() +  # Flip the plot horizontally
  labs(title = ".........",
       x = "Mean support rate for liberal party
       ",
       y = "Age group") +  # Flip x and y axis labels
  theme_minimal()
```



```
tablesex <-
  survey_data_clean %>%
  group_by(sex) %>%
  summarise("Counts" = n(),
            "Numver of Voting liberal party" =  sum(vote_liberal),
            "Support rate" = sum(vote_liberal)/Counts)
kable(tablesex, caption = "")
```

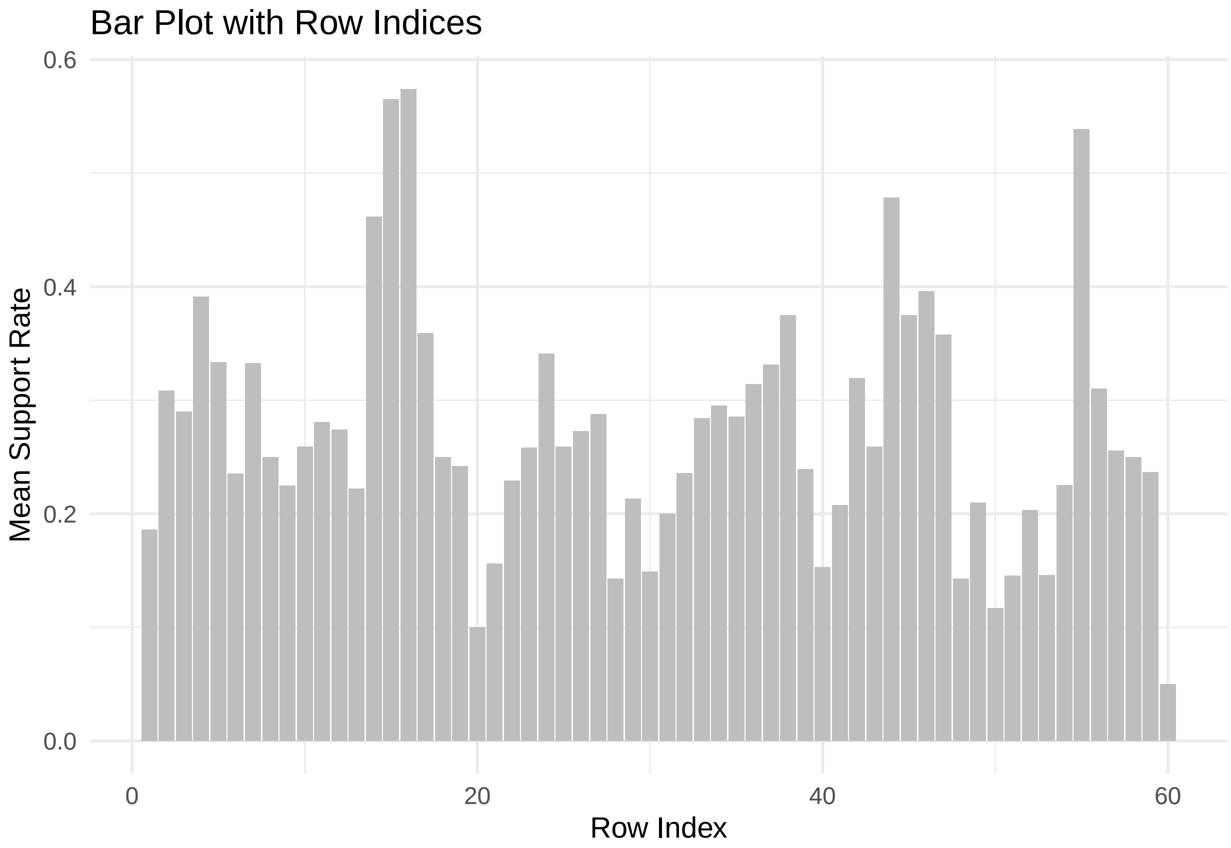
sex	Counts	Numver of Voting liberal party	Support rate
Female	7912	2079	0.2627654
Male	6545	1793	0.2739496

```
df_incometest <-
  survey_data_clean %>%
  group_by(income_family, province) %>%
    summarise(p=mean(vote_liberal), .groups = "drop") %>% mutate(Index = row_number())

tb1 <- df_incometest %>% summarise(
                                  min = min(p),
                                  Q1 = quantile(p,0.25),
                                  median = median(p),
                                  Q3 = quantile(p,0.75),
                                  max = max(p),
                                  sd = sd(p))
kable(tb1)
```

min	Q1	median	Q3	max	sd
0.05	0.2124714	0.2586778	0.3223981	0.5740741	0.1068243

```
ggplot(df_incometest, aes(x=Index, y=p))+
  geom_bar(stat = "identity", fill = "grey") +
  labs(title = "Bar Plot with Row Indices",
       x = "Row Index",
       y = "Mean Support Rate") +
  theme_minimal()
```



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

```
# Use this to create some plots. Should probably describe both the sample and population.
```

<Include a clear description of the plot(s). I would recommend one paragraph for each plot.>

Methods

To predict which parties will win the selection, we fit a logistic regression with random-effect on intercept using the glmer function from the "lme4" package and logit as link function.

This specific model is designed to analyze the relationship between the dependent binary variable (vote the party or not) and several independent variables, including age, sex, and a random effect related to the combination of province and income_family.

The response variable: vote_liberal/vote_Conservative/vote_ndp/vote_Bloc/vote_Green is binary, indicating whether an individual voted for the that corresponding party (1 for yes, 0 for no).

Thus, we will have 5 models with same structure but different responses. Each of them can estimate the proportion of voting its corresponding party.

We will apply post-stratification for the estimates from 5 models to adjusting for the issues of non-response or non-probability sampling. After that, we can get 5 $\hat{y}^{PS}$: the overall weighted estimated proportion of voting for each party. The party with highest proportion will possibly win the selection.

Model Specifics

We fitted a logistic mixed model to predict the proportion of voting with age and sex (vote ~ age + sex). The model included a random effect related to the combination of province and income_family as random effects on intercept (1 | province:income_family).

$$
\begin{aligned}
log(\frac{p}{1-p}) = \beta_{0j} + \beta_1 I_{\text{ij}}^{\text{25-30age}} + \beta_2 I_{\text{ij}}^{\text{30-35age}} + \beta_3 I_{\text{ij}}^{\text{35-40age}} + \beta_4 I_{\text{ij}}^{\text{40-45age}} + \beta_5 I_{\text{ij}}^{\text{45-55age}} + \beta_6 I_{\text{ij}}^{\text{55-65age}} + \beta_7 I_{\text{ij}}^{\text{over65}} + \beta_8 SEX_{\text{ij}} \\
\beta_{0j} \sim \text{Normal}(\mu_{\beta_0}, \sigma^2_{\beta_0})
\end{aligned}
$$

p represents the proportion of voting for one party. β_{0j} represents a Random Intercept that affected by the combination of province and income_family. I_{ij} is the indicator that if the observation is in the particular age group (Yes = 1, No = 0). SEX_{ij} is a categorical variable: male or female. β_1 to β_8 are the coefficients of these variables.

Since we want to fit a logistic regression with random-effect on intercept, we will have following assumptions:

Linearity of the Log-Odds: The logistic link function used in the model assumes that the relationship between the predictors and the log-odds of the response variable is linear.

Independence of Observations: Each observation in your dataset should be independent of the others.

Appropriate Random Effects Structure: The specification of random effects, in this case, (1|province:income_family), should accurately reflect the structure of the data. You need to ensure that it is justified and that it accounts for any correlations or nesting in the data

After we build the initial model, we will consider how to choose a better model or drop the variable that is not significantly impact the response. We will use AIC and BIC to compare different models and choose the one that best balances goodness of fit. If AIC and BIC are lower, after a change for the model, it typically indicates an improvement in the model's fit to the data.

Post-Stratification

Post-stratification is a technique used in survey analysis to improve the accuracy of estimates by adjusting for differences in sampling probabilities and non-response. It involves dividing the population into strata based on certain characteristics (variables) and then weighting the observations in each stratum to account for different estimates and sum them up to get a overall average.

In our logistic regression with random-effect on intercept for predicting voting proportion, we can apply post-stratification by using survey weights.

In order to estimate the proportion of voting one party, we use group_by function to group our cleaned census data by the variables we use as predictors in our regression models. For example, if we use age, province and income_family as our predictors, then we will group our data by these predictors. Each rows after grouping is cell/strata and has its own combination of age group, province and family income level. Use summarise function to record the strata size for each strata.

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

Then, fit our regression model for each strata to get the estimates and weight these estimates by its sample size based on the formula above. $\widehat{y}_j$ is the estimated proportion of voting one specific party for jth strata. N_j is the sample size for jth strata. $\sum N_j$ is the total size in cleaned census data. Last but not least, $\hat{y}^{PS}$ is the overall weighed proportion of voting one specific party by Post-Stratification.

Since we have 5 regression models, we will get the proportion of voting the 5 party by Post-Stratification. The party with highest proportion will possibly win the selection.

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

Results

After we fit the initial model:

$$log(\frac{p}{1-p}) = \beta_{0j} + \beta_1 I_{\text{ij}}^{\text{25-30age}} + \beta_2 I_{\text{ij}}^{\text{30-35age}} + \beta_3 I_{\text{ij}}^{\text{35-40age}} + \beta_4 I_{\text{ij}}^{\text{40-45age}} + \beta_5 I_{\text{ij}}^{\text{45-55age}} + \beta_6 I_{\text{ij}}^{\text{55-65age}} + \beta_7 I_{\text{ij}}^{\text{over65}} + \beta_8 SEX_{\text{ij}}$$

$$\beta_{0j} \sim \text{Normal}(\mu_{\beta_0}, \sigma^2_{\beta_0})$$

The outcome of the model for liberal party is showed on this table:

Table 3: Summary of model 1 for liberal party

Parameter	Coefficient	95% CI	z	p	Effects
(Intercept)	-1.51	[-1.71, -1.31]	-14.66	< .001	fixed
age [25-30]	0.05	[-0.17, 0.28]	0.48	0.633	fixed
age [30-35]	0.37	[0.16, 0.58]	3.40	< .001	fixed
age [35-40]	0.51	[0.29, 0.72]	4.66	< .001	fixed
age [40-45]	0.41	[0.20, 0.62]	3.77	< .001	fixed
age [45-55]	0.47	[0.28, 0.65]	4.85	< .001	fixed
age [55-65]	0.57	[0.38, 0.75]	6.04	< .001	fixed
age [65 and above]	0.65	[0.47, 0.83]	7.12	< .001	fixed
sex [Male]	-0.03	[-0.10, 0.05]	-0.70	0.484	fixed
province:income_family	0.39				random
AIC	16535.3				
BIC	16611.1				
logLik	-8257.6				

We choose age and sex as the fixed effects and the combination of province and family income level as the random intercept. It is noticeable that the p-value for predictor sex is quite large, which means we have strong evidence that ses is not a significant predictor for the response. Thus, we drop the variable sex. And refit the models.

There is our reduced model look like:

$$log(\frac{p}{1-p}) = \beta_{0j} + \beta_1 I_{\text{ij}}^{\text{25-30age}} + \beta_2 I_{\text{ij}}^{\text{30-35age}} + \beta_3 I_{\text{ij}}^{\text{35-40age}} + \beta_4 I_{\text{ij}}^{\text{40-45age}} + \beta_5 I_{\text{ij}}^{\text{45-55age}} + \beta_6 I_{\text{ij}}^{\text{55-65age}} + \beta_7 I_{\text{ij}}^{\text{over65}}$$

$$\beta_{0j} \sim \text{Normal}(\mu_{\beta_0}, \sigma^2_{\beta_0})$$

The outcome of the reduced model for liberal party is showed on this table:

Table 4: Summary of reduced model for liberal party

Parameter	Coefficient	95% CI	z	p	Effects
(Intercept)	-1.52	[-1.72, -1.32]	-14.79	< .001	fixed
age [25-30]	0.05	[-0.17, 0.28]	0.47	0.641	fixed
age [30-35]	0.37	[0.15, 0.58]	3.38	< .001	fixed
age [35-40]	0.50	[0.29, 0.72]	4.63	< .001	fixed
age [40-45]	0.40	[0.19, 0.62]	3.73	< .001	fixed
age [45-55]	0.46	[0.27, 0.65]	4.81	< .001	fixed
age [55-65]	0.56	[0.38, 0.74]	6.00	< .001	fixed
age [65 and above]	0.65	[0.47, 0.82]	7.09	< .001	fixed
province:income_family	0.39				random
AIC	16533.8				
BIC	16602.0				
logLik	-8257.9				

After dropping the sex, the model has lower AIC and BIC. Indicating that the modified model is a better fit to the data and may be a more suitable choice for analysis or prediction. So we choose the reduced model to be our final model.

<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.>

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)

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.>