Introduction

In this project, I will provide election suggestions for the Liberal Party of Canada through the analysis of statistical data.

I'll focus on the factors that affect election results, including age, gender, education, and so on, to see how they affect people's choices. With reliable models, we may be able to predict people's choices and change campaign strategies accordingly.

At the same time, I will also compare the Liberal Party with its main competitor, the conservative party, to see if a certain group of people have a significant preference for the choice of political parties.

The methods I will use are linear regression and hypothesis test.

Question 1

Based on some information that we know about a person, can we predict how do they feel about the Liberal Party?

This question can show which factors determine the rating of the Liberal Party. The population I'm trying to make inference is the whole sample in the survey. I'll use the method of linear regression.

Hypothesis

The rating of the Liberal Party are determined by the age, gender, education level and how people are satisfied with the way democracy works.

Data instruction

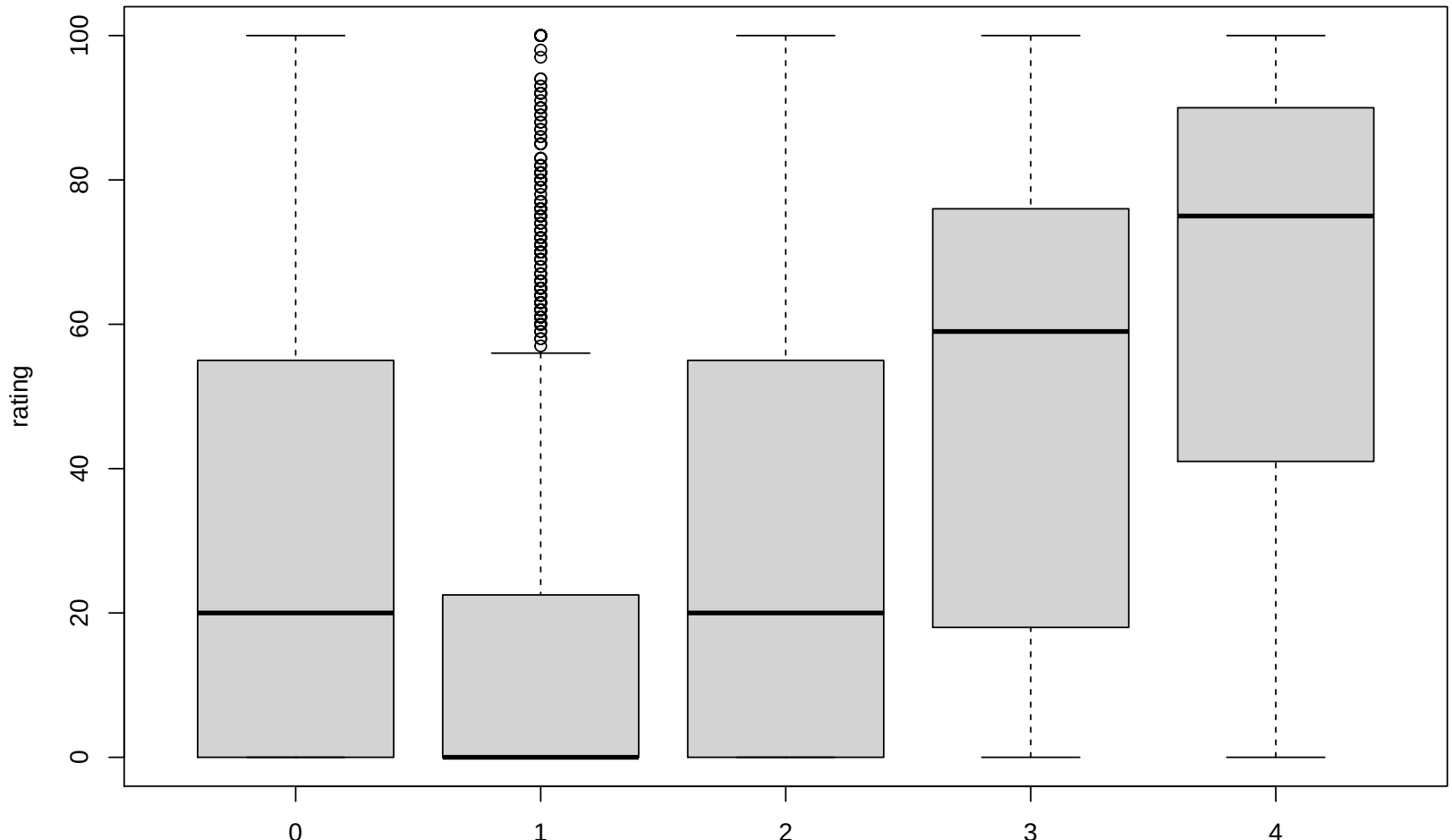
The following columns in *ces19* are concerned:
age , *gender* , *education* , *demsat* , *party_rating_23*

The first step is data collection and data cleaning. We initialize a new frame to store the data in Question 1, and put the data into the frame. The following code is an example of dealing with the gender data.

```
gender_vec <- numeric(0)
n <- 0
i <- 1
for (gender_data in ces19$gender){
  if (gender_data == "A man"){
    gender_vec[i] <- 1
    i <- i+1
  } else if (gender_data == "A woman"){
    gender_vec[i] <- 0
    i <- i+1
  } else {
    gender_vec[i] <- 1.5
    i <- i+1
    n <- n+1
  }
```

Take a look at the data. The following figure shows that the rating may increase with more satisfaction level of democracy. Here I chose the **boxplot** because the sample size is too large.

```
boxplot(rating~demsat,data = q1_frame)
```



Statistically, we use the linear model:

$$\widehat{party\_rating\_23} = \widehat{b_0} + \widehat{b_1} \times age + \widehat{b_2} \times gender$$
$$+\widehat{b_3} \times education + \widehat{b_4} \times demsat$$

R codes (run in background)

```
fit <- lm(rating~age+gender+education+demsat,
data=q1_frame)
summary(fit)
```

Model result

$$\widehat{party\_rating\_23} = 4.39 - 0.17 \times age - 2.2 \times gender$$
$$+1.8 \times education + 13.7 \times demsat$$

From the summary of the result, all the coefficients are significant. The model shows that the following kind of person will be more likely to support the Liberal Party:

- Young people;
- Women;
- People with higher educational level; and
- People who are satisfied with the way democracy works

Question 2

Does the age factor significantly influence how satisfied are people with the performance of the federal government under Justin Trudeau?

The question concerns whether people of different ages have different satisfaction with the Liberals.

Hypothesis

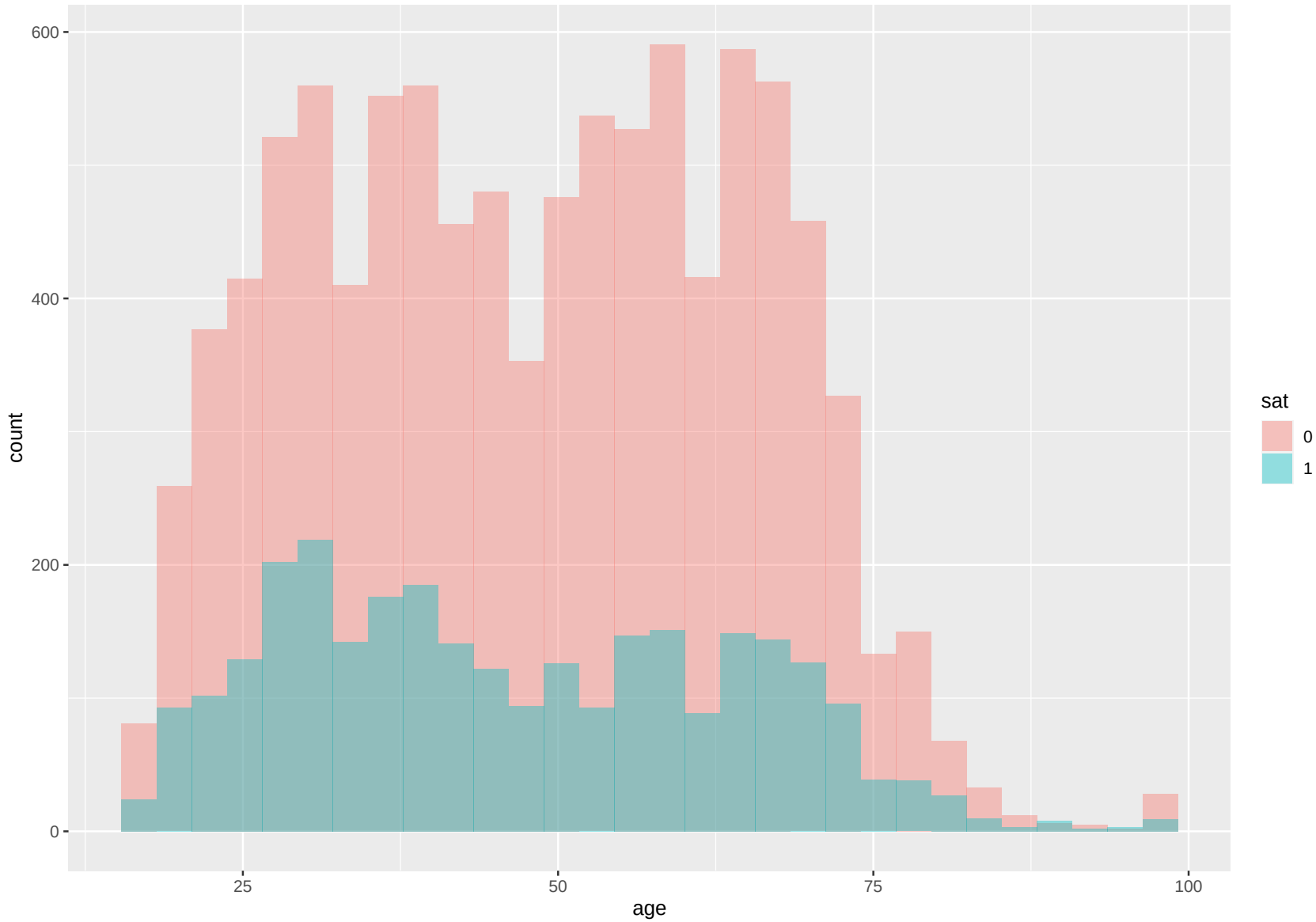
People who are very satisfied with the government have different age with those who are not very satisfied.

And the columns in *ces19* concerned are *age* and *fed_gov_sat*

The data collection is similar to Question 1.

To show the data in the **histogram**, I generated a frame for question 2 *q2_frame* with columns of the age and the satisfaction level, 1 for *Very satisfied* and 0 for *Not very satisfied*.

```r
ggplot(q2_frame, aes(x = age, fill = sat)) +
  geom_histogram(position = "identity", alpha = 0.4)
```



It seems that younger people may be more satisfied with the government.

Statistically, we use the hypothesis test.

$H_0 : \ \mu_1 = \mu_2$

$H_1 : \ \mu_1 \neq \mu_2$

Where μ_1 and μ_2 are the average age of people who are or aren't satisfied with the government.

```r
t_test = t.test(q2_sat_vec,q2_nsat_vec)
print(t_test)
```

```
##
##  Welch Two Sample t-test
##
## data:  q2_sat_vec and q2_nsat_vec
## t = -4.4067, df = 4595, p-value = 1.073e-05
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -2.3085051 -0.8869134
## sample estimates:
## mean of x mean of y
##  46.52630  48.12401
```

Since $\mu_1 = 46.53$, $\mu_2 = 48.12$ and p-value$<$0.05, it's statistically significant that younger people are more satisfied with the government.

Question 3

Does the feeling about the immigrants significantly influence the first choice to vote for?

The population I'm trying to make inference is the people whose first choice to vote for is Liberal Party or Conservative Party, the main competitors of the election.

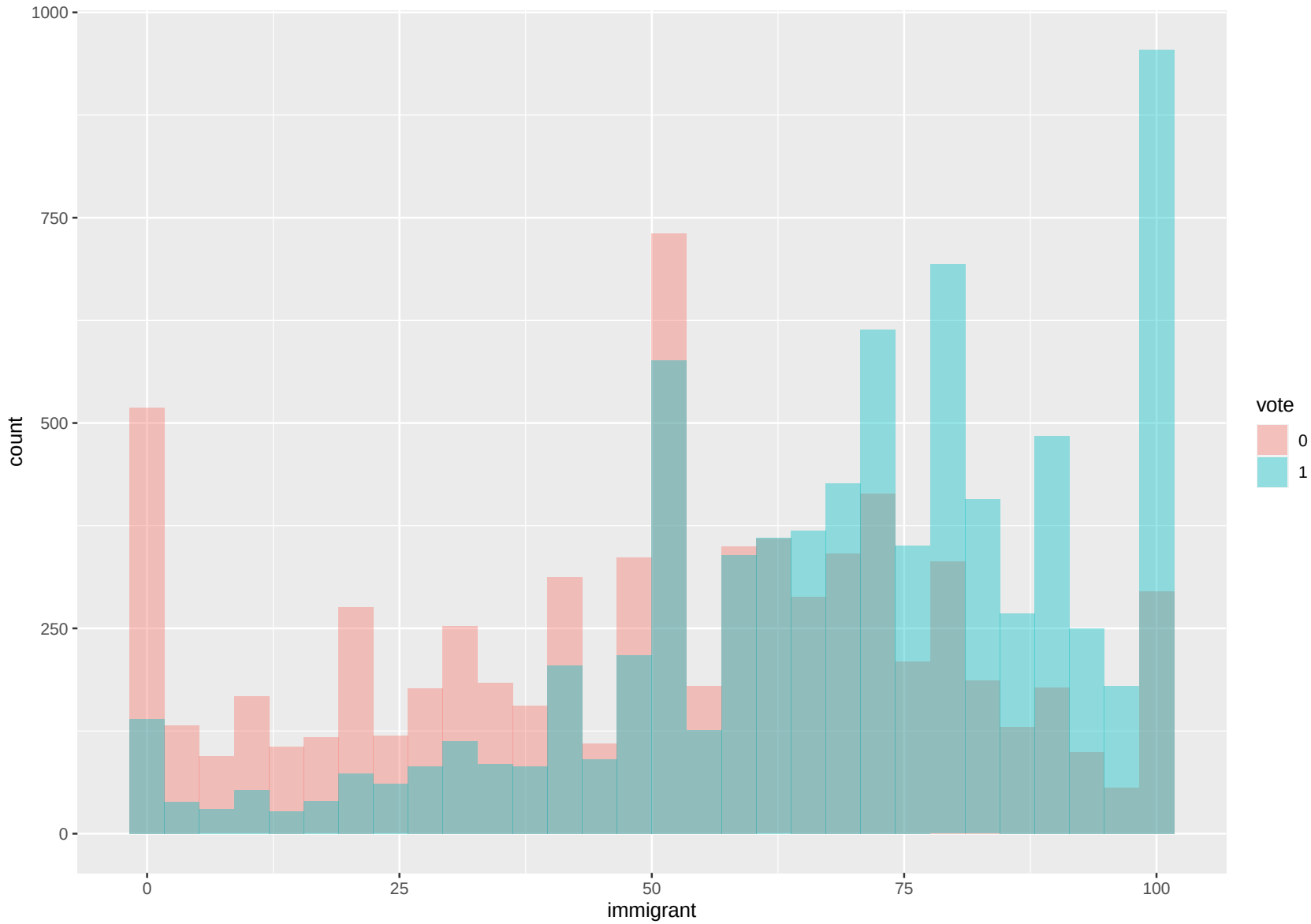
Hypothesis

People who vote for is Liberal Party or Conservative Party have different opinions with immigrants.

And the columns in *ces19* concerned are *votechoice* and *groups_therm_2*

Also, to show the data in the **histogram**, I generated a frame for question 3 *q3_frame* with columns of the feeling about the immigrants and the vote, 1 for *Liberal Party* and 0 for *Conservative Party*.

```r
ggplot(q3_frame, aes(x = immigrant, fill = vote)) +
  geom_histogram(position = "identity", alpha = 0.4)
```



It seems that people who vote for Liberal Party have better feelings about immigrants.

Statistically, we use the hypothesis test.

$H_0 : \ \mu_1 = \mu_2$

$H_1 : \ \mu_1 \neq \mu_2$

Where μ_1 and μ_2 are the average feeling score of people who vote for Liberal Party or Conservative Party.

```r
t_test3 = t.test(q3_lib_vec,q3_con_vec)
print(t_test3)
```

```
##
##  Welch Two Sample t-test
##
## data:  q3_lib_vec and q3_con_vec
## t = 43.613, df = 14220, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  17.54540 19.19674
## sample estimates:
## mean of x mean of y
##  69.17268  50.80161
```

Since $\mu_1 = 69.17$, $\mu_2 = 50.80$ and p-value$<$0.05, it's statistically significant that people who vote for Liberal Party have better feelings about immigrants.