W203 Statistics for Data Science: Comparing Means

Jade Hou, Skyler Roh, Cecily Sun, Richard Wang

Contents

The Data **1**

Assignment **2**
Submission Guidelines 2

Research Questions **3**
Question 1: Do US voters have more respect for the police or for journalists? 3
Question 2: Are Republican voters older or younger than Democratic voters? 7
Question 3: Do a majority of independent voters believe that the federal investigations of Russian election interference are baseless? 15
Question 4: Was anger or fear more effective at driving increases in voter turnout from 2016 to 2018? 19
Question 5: Select a fifth question that you believe is important for understanding the behavior of voters 26

The Data

The American National Election Studies (ANES) conducts surveys of voters in the United States. While its flagship survey occurs every four years at the time of each presidential election, ANES also conducts pilot studies midway between these elections. You are provided with data from the 2018 ANES Pilot Study.

For a glimpse into some of the intricacies that go into the design of this study, take a look at the introduction to the ANES User's Guide and Codebook.

It is important to consider the way that the ANES sample was created. Survery participants are taken from the YouGov panel, which is an online system in which users earn rewards for completing questionnaires. This feature limits the extent to which results generalize to the U.S. population.

To partially account for differences between the YouGov panel and the U.S. Population, ANES assigns a survey weight to each observation. This weight estimates the degree to which a citizen with certain observed characteristics is over- or under-represented in the sample. For the purposes of this assignment, however, you are not asked to use the survey weights. (For groups with a strong interest in survey analysis, we recommend that you read about R's survey package. We will assign a very small number of bonus points (up to 3) to any group that correctly applies the survey weights and includes a clear explanation of how these work).

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
anes <- read.csv("data/raw/anes_pilot_2018.csv")
```

Following is an example of a question asked on the ANES survey:

> How difficult was it for you to vote in this last election?

The variable `votehard` records answers to this question, with the following encoding:

- `-1` inapplicable, legitimate skip
- `1` Not difficult at all
- `2` A little difficult
- `3` Moderately difficult
- `4` Very difficult
- `5` Extremely difficult

To see the precise form of each question, take a look at the Questionnaire Specifications.

Assignment

You will use the ANES dataset to address five research questions. For each question, you will need to operationalize the concepts (selecting appropriate variables and possibly transforming them), conduct exploratory analysis, deal with non-response and other special codes, perform sanity checks, select an appropriate hypothesis test, conduct the test, and interpret your results. When selecting a hypothesis test, you may choose from the tests covered in the async videos and readings. These include both paired and unpaired t-tests, Wilcoxon rank-sum test, Wilcoxon signed-rank test, and sign test. You may select a one-tailed or two-tailed test.

Please organize your response according to the prompts in this notebook.

Note that this is a group lab. There is a **maximum of three students per team.** Although you may work on your own, we do not recommend this (we have found that individuals tend to do worse than teams on past labs).

Please limit your submission to 4500 words, not counting code or figures. We will use some python code like the following to perform the wordcount on your notebook or Rmd file.

```
import nbformat

with open('W203_Lab_2.ipynb') as f:
    nb = nbformat.read(f, as_version=4)
    total_words = 0

    for cell in nb['cells']:
        if cell['cell_type'] == 'markdown':
            total_words += len(cell['source'].split())
    print("Total words:", total_words)
```

Hint: When answering questions about the potential gaps between conceptual and operational definitions it is often helpful to ask yourself "If I could do anything I wanted, including borrowing Rick's portal gun, or doing an FMRI scan on all survey takers, to operationalize this particular concept how would I do it" and then explain how that is different than what is actually measured.

Submission Guidelines

- Submit *one* report per group.
- Submit *both* your pdf report as well as your source file.
- **Only analyses and comments included in your PDF report will be considered for grading.**
- Include names of group members on the front page of the submitted report.
- Naming structure of submitted files:
  - PDF report: [student_surname_1]_[student_surname_2][_*]_lab_2.pdf

- Jupyter Notebook: [student_surname_1]_[student_surname_2][_*]_lab_2.ipynb (or)
- Rmd file: [student_surname_1]_[student_surname_2][_*]_lab_2.Rmd

Research Questions

Question 1: Do US voters have more respect for the police or for journalists?

Introduce your topic briefly. (5 points)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

> The 'respect' variable is operationalized as the question *'How would you rate journalists/the police?'* in the survey. To respond this question, the respondent was given a thermometer to give the rating on how warm/cold and favorable/unfavorable they are. The biggest gap between the variable we want to study and the survey question is that, warm/cold and favorable/unfavorable does not necessarily mean respect. For example, people may think of police as cold and unfavorable for the nature of their job, but it doesn't mean that they don't respect the police. Looking at the results from this question may not fully reflect the concept that we are trying to study.

Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

```
# Summary statistics of responses for journalists
summary(anes$ftjournal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -7.00   21.00   52.00   52.26   82.00  100.00
```

```
# How many of the responses were negative?
length(anes[anes$ftjournal<0])
```

```
## [1] 2
```

```
# Is there any missing values(NA)?
sum(is.na(anes$ftjournal))
```

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

```
# Summary statistics of responses for the police
summary(anes$ftpolice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   47.00   70.00   64.68   90.00  100.00
```

```
# Is there any missing values(NA)?
sum(is.na(anes$ftpolice))
```

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

```
# How many of the responses were negative?
length(anes[anes$ftpolice<0])
```

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

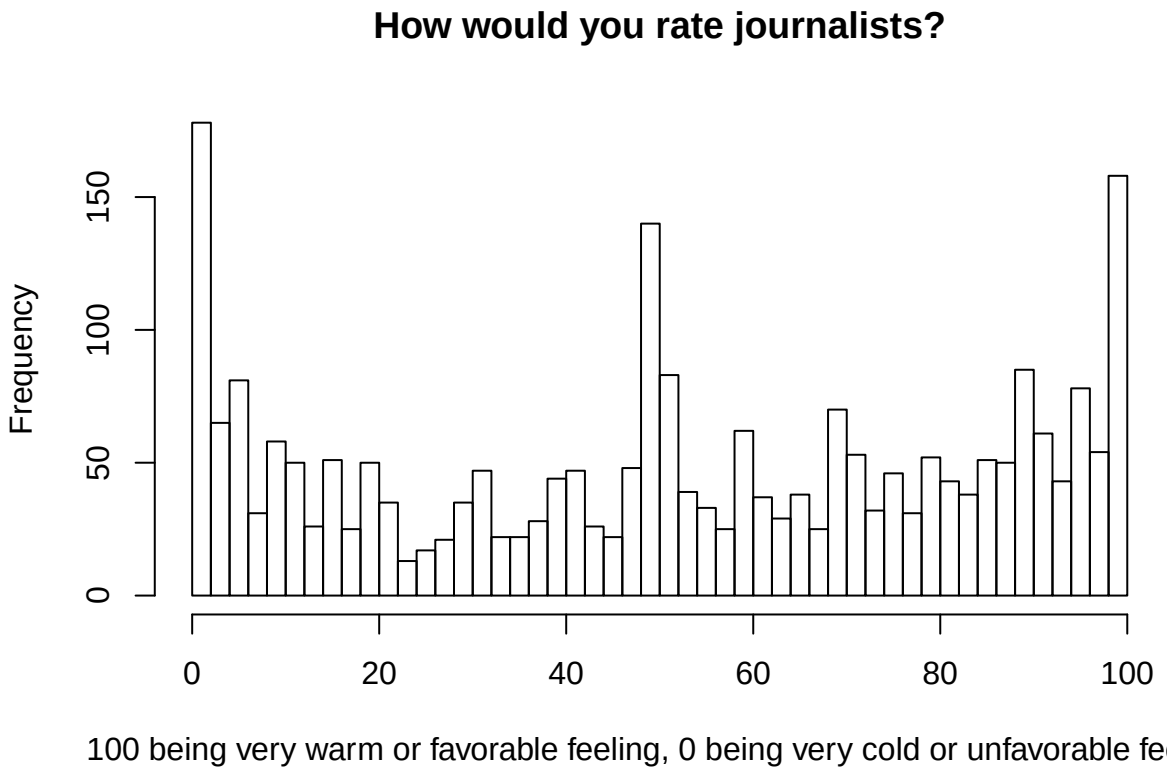
> Looking at the summary statistics, I found that there are 2 *"no answer(-7)"* in *'How would you rate journalists?'*. Since the size of "no answer" is very small, and removing these 2 observations

won't have too much impact on the overall sample mean(2 out of 2500), I will remove these 2 observations from the sample.

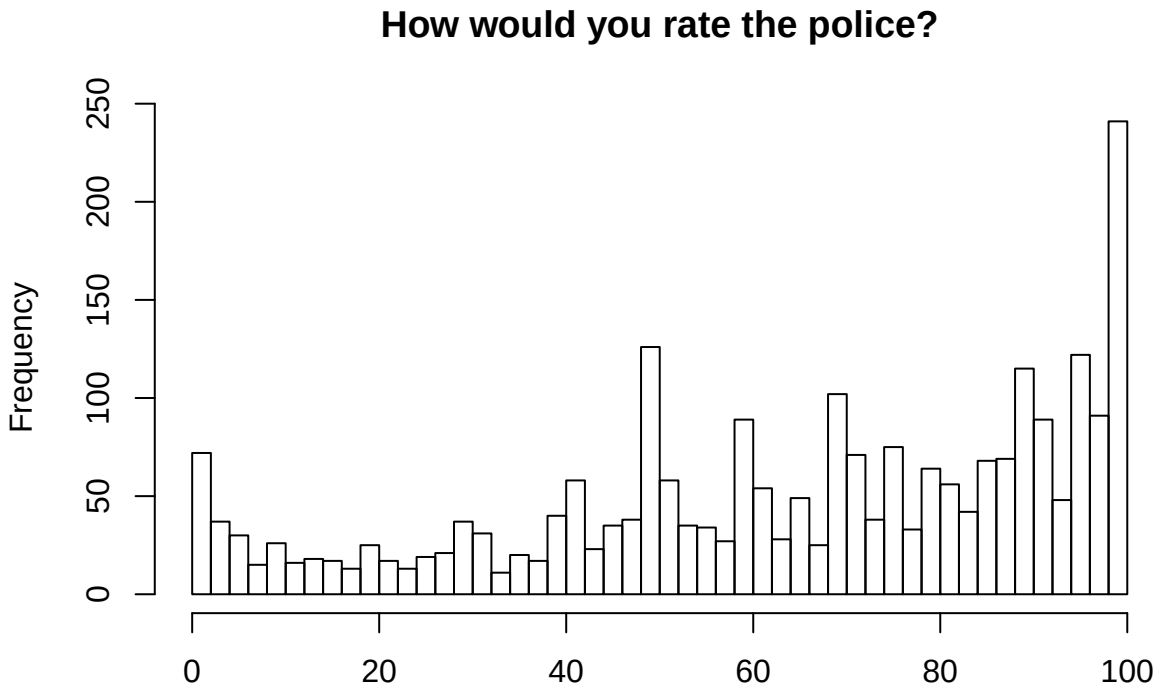
```
anes_Q1 <- anes[anes$ftjournal>=0,]
length(anes_Q1$ftjournal)
```

```
## [1] 2498
```

```
# check the distribution of responses for journalists
hist(anes_Q1$ftjournal, breaks = 50,
     main = "How would you rate journalists?",
     xlab = "100 being very warm or favorable feeling, 0 being very cold or unfavorable feeling")
```

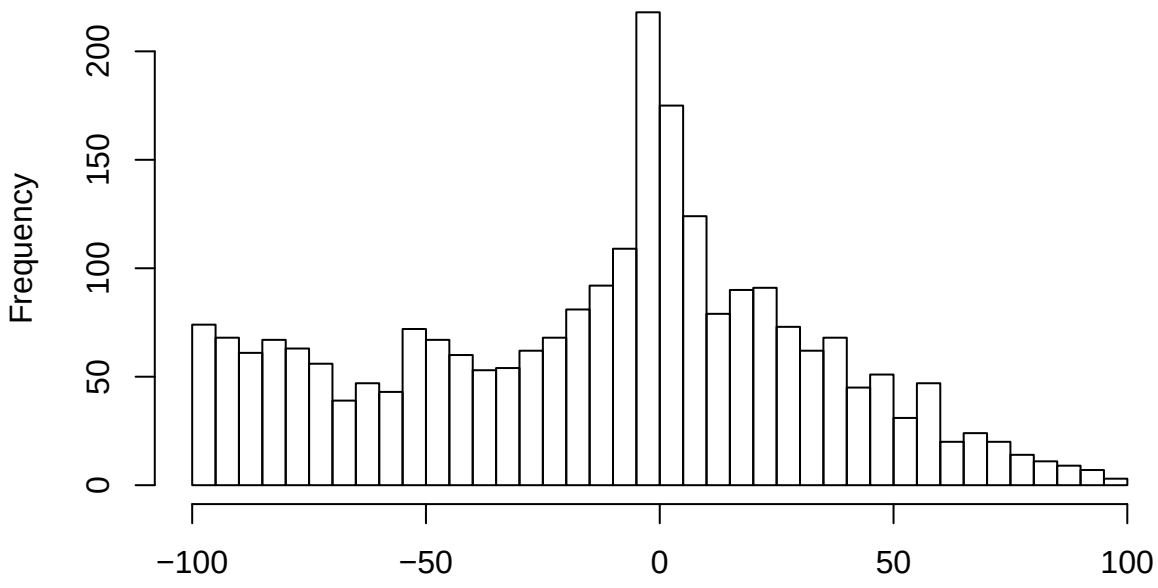



```
# check the distribution of responses for the police
hist(anes_Q1$ftpolice,  breaks = 50,
     main = "How would you rate the police?",
     xlab = "100 being very warm or favorable feeling, 0 being very cold or unfavorable feeling")
```



100 being very warm or favorable feeling, 0 being very cold or unfavorable feeling

```r
# check the distribution of the difference in responses between journalists and the police
hist(anes_Q1$ftjournal - anes_Q1$ftpolice,  breaks = 50,
     main = "Difference between responses towards journalist and the police",
     xlab = "Positive being higher rating for journalist, negative being higher rating for the police")
```



Positive being higher rating for journalist, negative being higher rating for the police

Based on your EDA, select an appropriate hypothesis test. (5 points)

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

> I will be using **dependent sample(paired) t-test** for this study for the following reasons - 1. There is a natural pairing between the two data points because they are collected from the same individual. The assumption. 2. The measurements follow the same scale/interval(between 0 to 100). 3. From the above histograms, we cann see that even thought neither does responses for journalists or the responses for the police have a normal curve, the difference between the two does. So instead of comparing the distributions between two responses, we will test for the distribution of the difference between the two responses, which will be much more likely to be statistically significant. The assumptions will then be - * Null hypothesis H_0: The difference between the two responses is 0 * Alternative hypothesis H_a: The difference between the two responses is not 0

Conduct your test. (5 points)

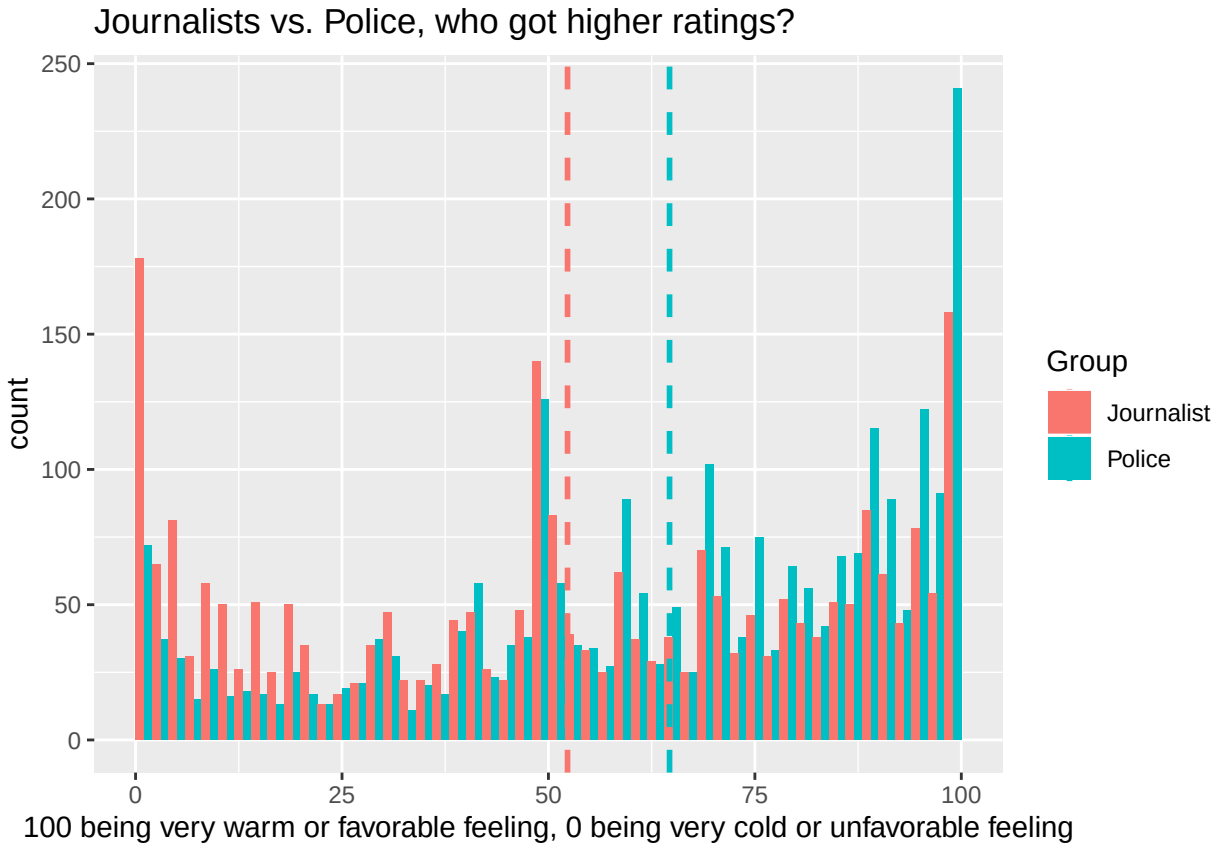
Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

```
# statistical significance test
t.test(anes_Q1$ftjournal,anes_Q1$ftpolice, paired = T)
```

```
##
##  Paired t-test
##
## data:  anes_Q1$ftjournal and anes_Q1$ftpolice
## t = -13.711, df = 2497, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -14.12160 -10.58776
## sample estimates:
## mean of the differences
##               -12.35468
```

> The t-test is highly statistically significant, with the p-value being very small(almost 0). This means that statistically, we can reject the null hypothesis that the difference between the two responses is 0.

```
# Practical significance
Q1_plot_df <- rbind(data.frame(Group="Journalist", rating = anes_Q1$ftjournal),
            data.frame(Group="Police", rating = anes_Q1$ftpolice))
Q1_Group_by_mean <- rbind(data.frame(Group="Journalist", avg_rating = mean(anes_Q1$ftjournal)),
            data.frame(Group="Police", avg_rating = mean(anes_Q1$ftpolice)))
ggplot(Q1_plot_df, aes(x = rating, fill = Group)) +
  geom_histogram(binwidth=.5, breaks = seq(0, 100, by = 2), position="dodge") +
  geom_vline(data=Q1_Group_by_mean, aes(xintercept = avg_rating,  colour = Group),
               linetype="dashed", size=1)+
  labs(title="Journalists vs. Police, who got higher ratings?") +
  labs(x="100 being very warm or favorable feeling, 0 being very cold or unfavorable feeling")
```



The histogram above represents the distribution of ratings by journalists and police, adding the two dotted lines as the average rating for each of the group. In addition to the Police received higher average ratings(12) than the journalists, the distribution of the ratings for the police is skewed to the left, meaning that the majority of people who surveyed gives higher than average ratings to the police. While the distribution of the ratings for journalists looks like a "E"-shaped - with 0, 50, 100 sticking out from the rest. It's clear to see that the difference has its practical meaning here - fewer people holds unfavorable feelings towards the police than the journalists. **To conclude, we think that US voters have more respect for the police than for journalists.**

Question 2: Are Republican voters older or younger than Democratic voters?

Introduce your topic briefly. (5 points)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

#####Question description: (1) **Sub-question1**: Are Republican voters older or younger than Democratic voters for the U.S. Senate? *H0*: True difference between the mean of Republican and Democratic voters in senate election is equal to 0 *H1*: True difference between the mean of Republican and Democratic voters in senate election is not equal to 0 (2) **Sub-question2**: Are Republican voters older or younger than Democratic voters for governor? *H0*: True difference between the mean of Republican and Democratic voters in governor election is equal to 0 *H1*: True difference between the mean of Republican and Democratic voters in governor election is not equal to 0

#####Anes variable: (1) **senate18p**: For the U.S. Senate, did you vote Democrat, Republican, or another party(, or di -1 inapplicable, legitimate skip 1 Democrat 2 Republican 3 another party 4 two different parties (2) **gov18p**: For governor of [INPUTSTATE], did you vote for a Democrat, Republican, or another -7 No Answer -1 inapplicable, legitimate skip 1 Democrat 2 Republican 3

another party (3) **birthyr**: profile data Birth Year Some gaps in the operationalization of the question include the likely difference in age distribution of voters depending on which type of election in question: gubenatorial, congressional, presidential. Additionally, not every state has the same election cycle and may not have had these elections in 2018.

Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

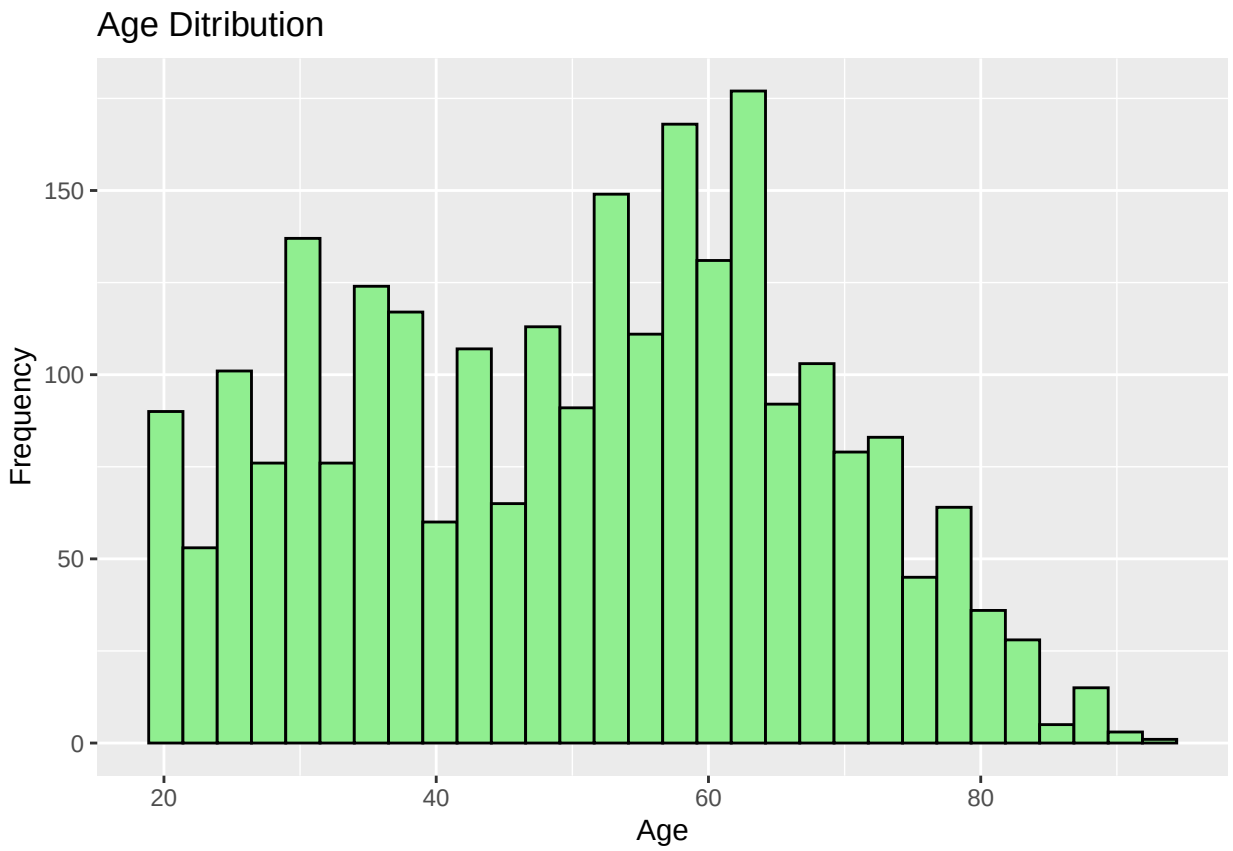
We examine the age distributions for each of the combinations w.r.t. party affiliation and election type for approximate normality.

Variable: Age

```
age = 2019 - anes$birthyr
age2 = as.data.frame(age)

ggplot(age2, aes(x=age)) +
  geom_histogram(color="black", fill="lightgreen") +
  xlab("Age") +
  ylab("Frequency") +
  ggtitle("Age Ditribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



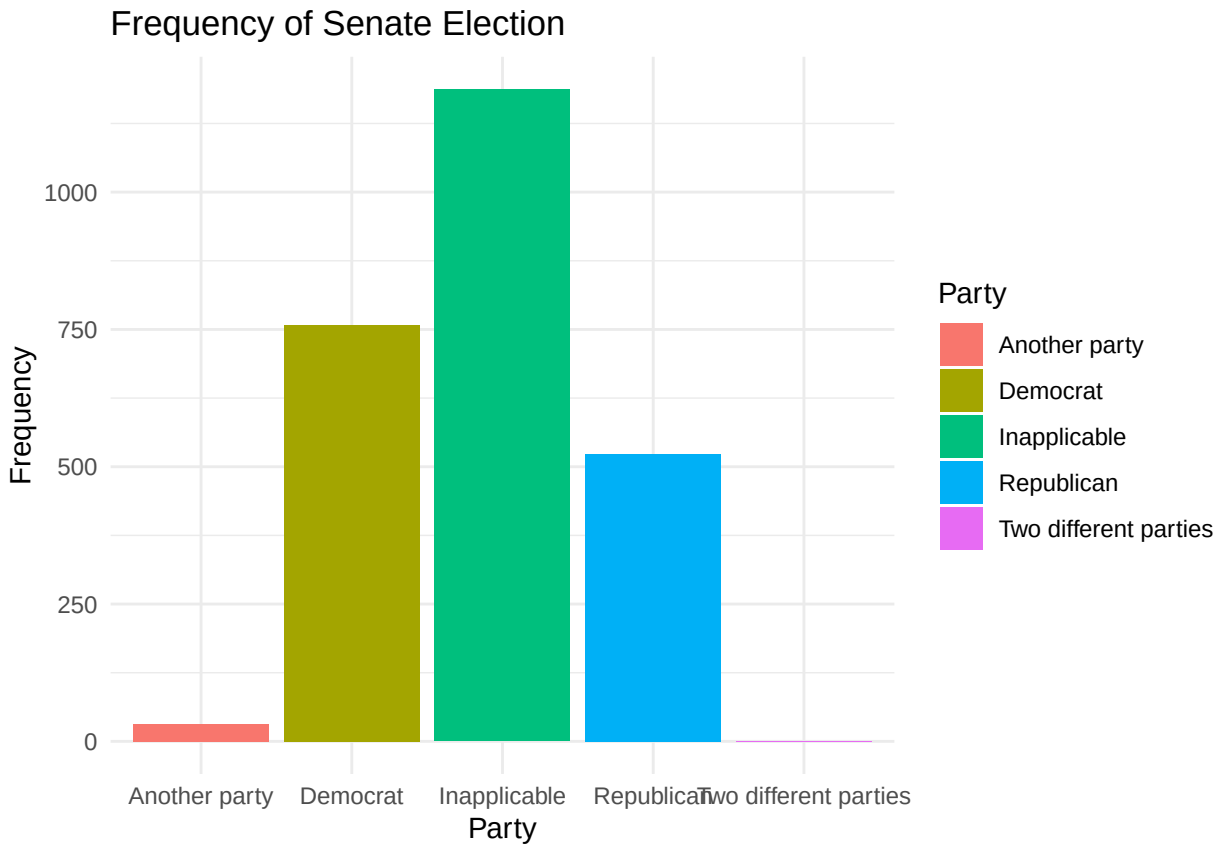
Variable: senate18p

```r
senate = anes$senate18p

senate_count = as.data.frame(table(senate))
Party = c("Inapplicable", "Democrat", "Republican", "Another party", "Two different parties")
senate_count = cbind(Party, senate_count)
senate_count
```

```
##                   Party senate Freq
## 1          Inapplicable     -1 1187
## 2              Democrat      1  758
## 3            Republican      2  523
## 4         Another party      3   31
## 5 Two different parties      4    1
```

```r
# plot the frequency of Senate Election
p<-ggplot(senate_count, aes(x = Party, y = Freq, fill = Party)) +
  geom_bar(stat="identity")+theme_minimal() +
  xlab("Party") +
  ylab("Frequency") +
  ggtitle("Frequency of Senate Election")
p
```

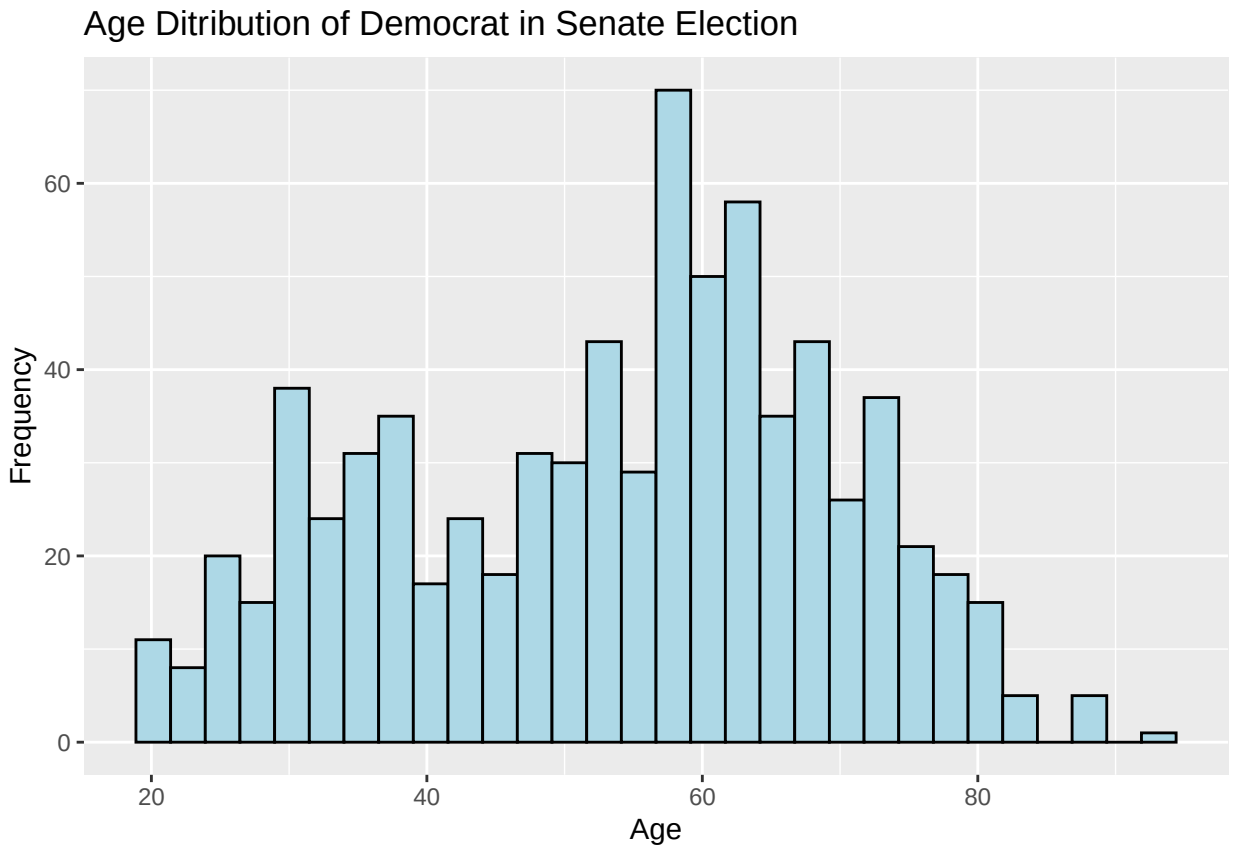


Age Ditribution of Democrat in Senate Election

```r
democrat = which(anes$senate18p == 1)
age = 2019 - anes$birthyr
d_age_s = age[democrat]
d_age_s = as.data.frame(d_age_s)
```

```
ggplot(d_age_s, aes(x=d_age_s)) +
  geom_histogram(color="black", fill="lightblue") +
  xlab("Age") +
  ylab("Frequency") +
  ggtitle("Age Ditribution of Democrat in Senate Election")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

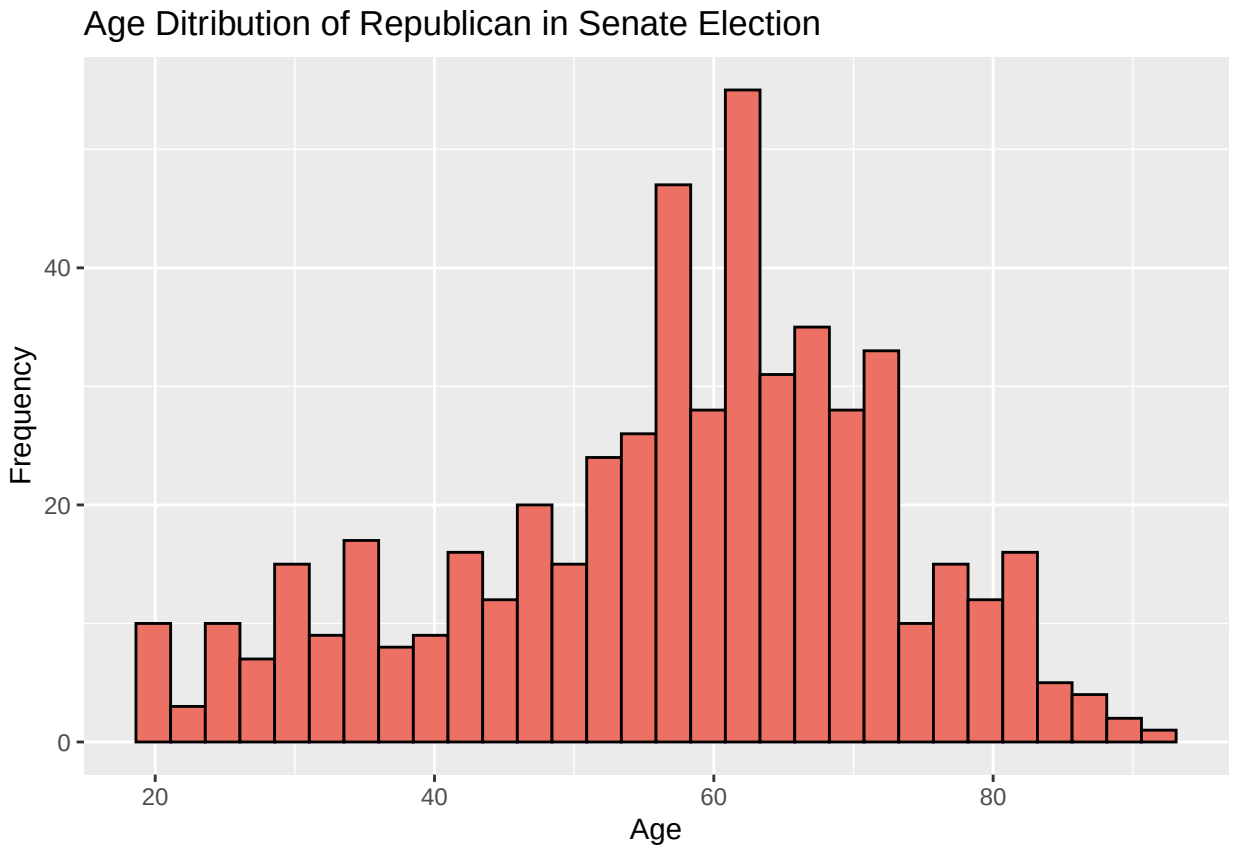


Age Ditribution of Republican in Senate Election

```
republican = which(anes$senate18p == 2)
age = 2019 - anes$birthyr
r_age_s = age[republican]
r_age_s = as.data.frame(r_age_s)

ggplot(r_age_s, aes(x=r_age_s)) +
  geom_histogram(color="black", fill="#EC7063") +
  xlab("Age") +
  ylab("Frequency") +
  ggtitle("Age Ditribution of Republican in Senate Election")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



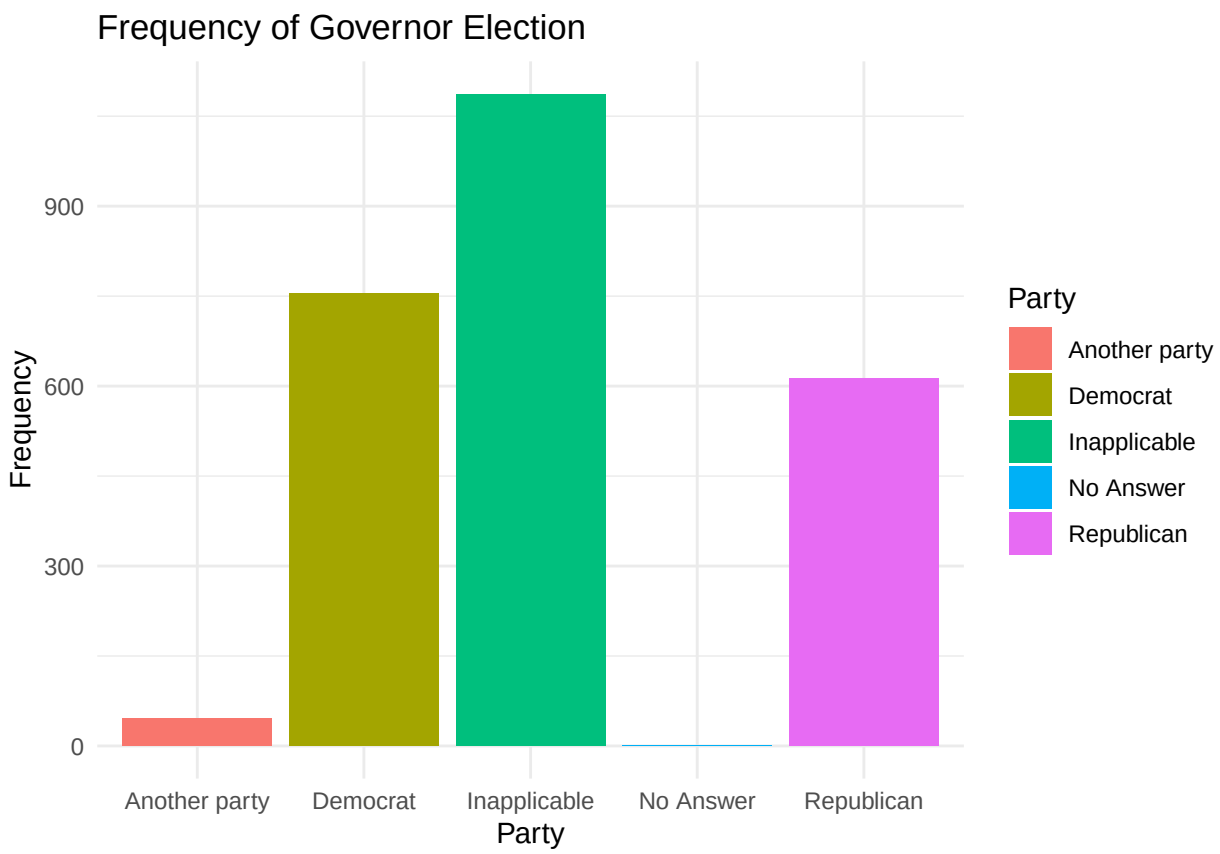
Variable: gov18p

```
gov = anes$gov18p

gov_count = as.data.frame(table(gov))
Party = c("No Answer", "Inapplicable", "Democrat", "Republican", "Another party")
gov_count = cbind(Party, gov_count)
gov_count
```

```
##           Party gov Freq
## 1     No Answer  -7    1
## 2  Inapplicable  -1 1087
## 3      Democrat   1  754
## 4    Republican   2  612
## 5 Another party   3   46
```

```
# plot the frequency of Senate Election
p<-ggplot(gov_count, aes(x = Party, y = Freq, fill = Party)) +
  geom_bar(stat="identity")+theme_minimal() +
  xlab("Party") +
  ylab("Frequency") +
  ggtitle("Frequency of Governor Election")
p
```

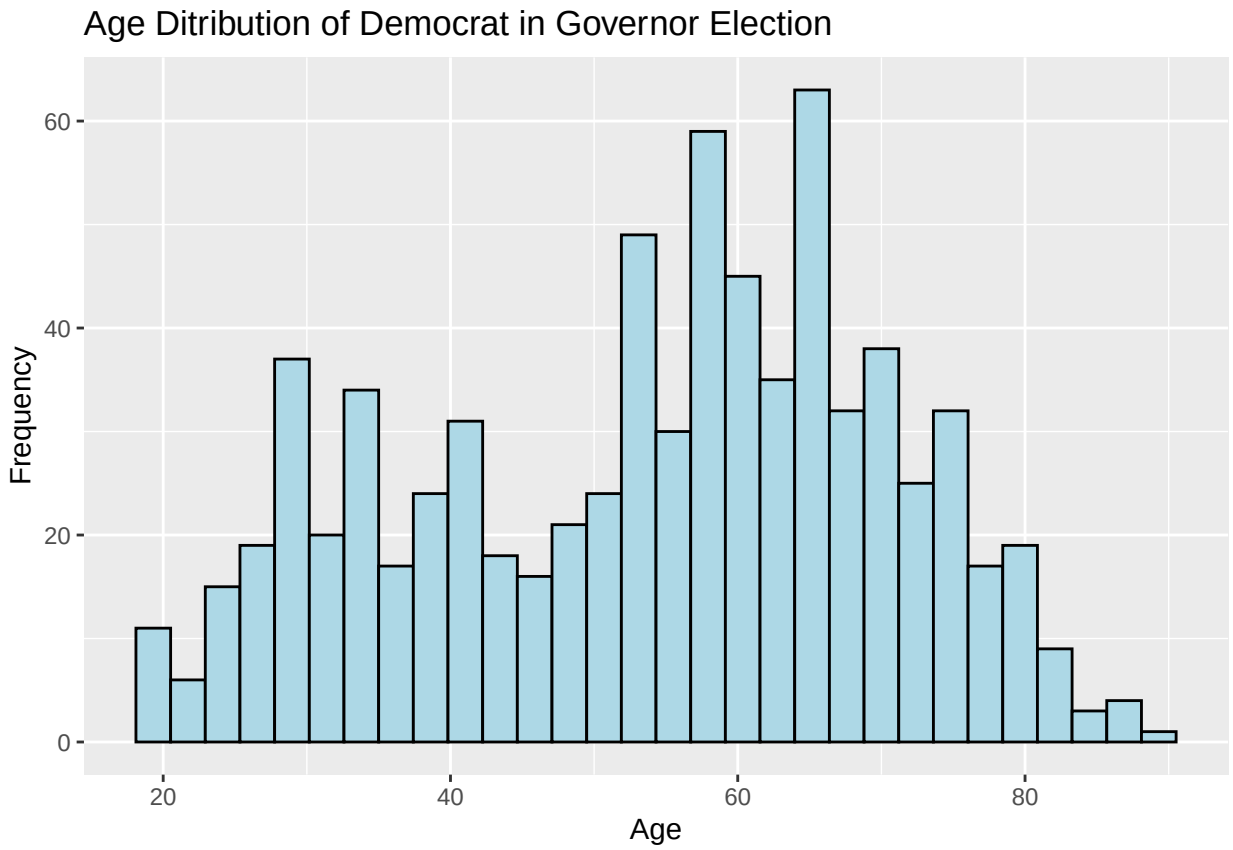


Age Ditribution of Democrat in Governor Election

```
democrat = which(anes$gov18p == 1)
age = 2019 - anes$birthyr
d_age_g = age[democrat]
d_age_g = as.data.frame(d_age_g)

ggplot(d_age_g, aes(x=d_age_g)) +
  geom_histogram(color="black", fill="lightblue") +
  xlab("Age") +
  ylab("Frequency") +
  ggtitle("Age Ditribution of Democrat in Governor Election")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

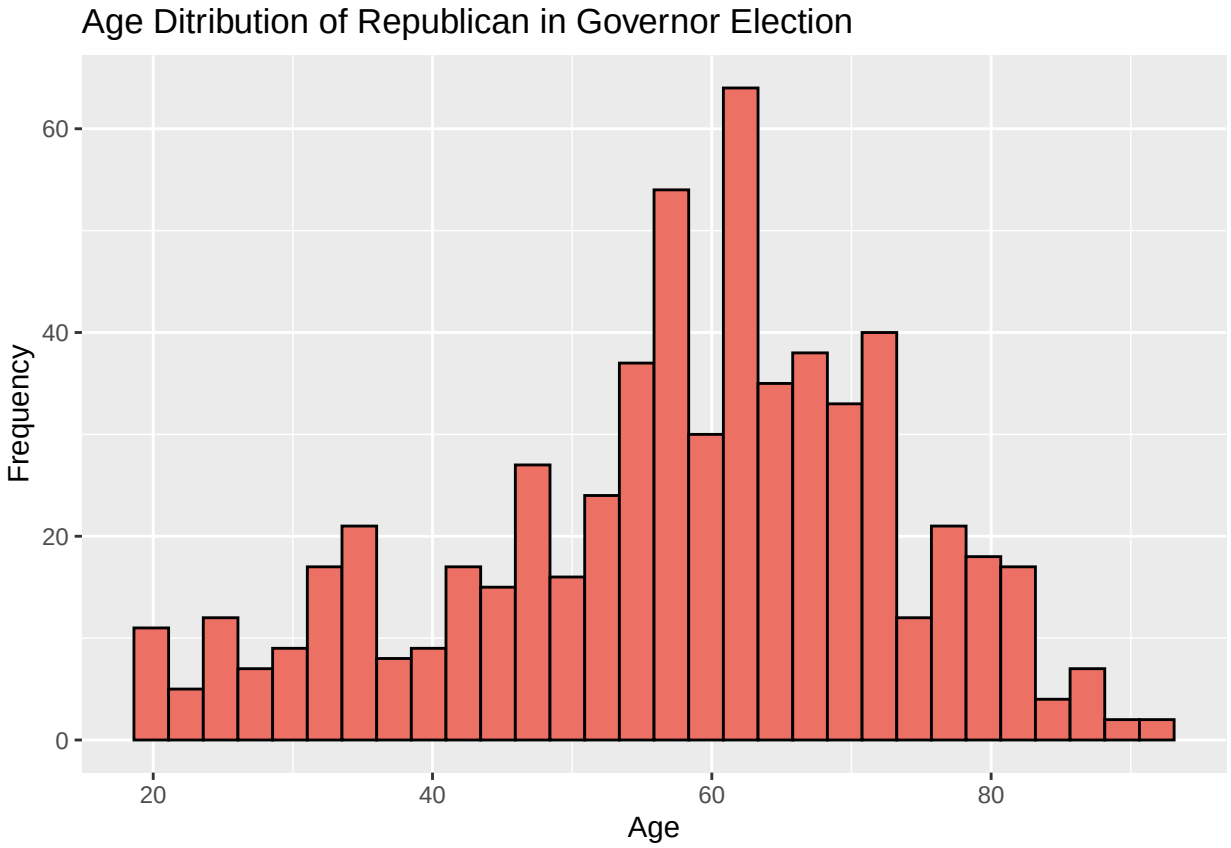


Age Ditribution of Republican in Governor Election

```
republican = which(anes$gov18p == 2)
age = 2019 - anes$birthyr
r_age_g = age[republican]
r_age_g = as.data.frame(r_age_g)

ggplot(r_age_g, aes(x=r_age_g)) +
  geom_histogram(color="black", fill="#EC7063") +
  xlab("Age") +
  ylab("Frequency") +
  ggtitle("Age Ditribution of Republican in Governor Election")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Based on your EDA, select an appropriate hypothesis test. (5 points)

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

> 2 sample T-test for sub-question 1 and 2 (1) The age distribution of Republican and Democratic voters in senate and governor elections both are similiar to normal curve and are discrete interval variables (2) 500 - 700 respondents in each group, CLT applies (3) Samples are i.i.d. and groups are not dependent Based on these reasons and we do not have an assumption on which party is older, we conduct two-sided test for sub-question 1 and 2.

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

> **Sub-question1**: Are Republican voters older or younger than Democratic voters for the U.S. Senate? (1) Statistical significance: We reject the null hypothesis that the average age of Republican and Democratic voters who voted for senate in 2018 is the same by conducting the two-side t-test. The result of the t-test is statistical significant because p-value = 0.0001422 which is smaller than 0.05. (2) Practical significance: In practical significance, the average age of Democratic voters is 53.71768 and Republican voters is 57.13767 which are notably different by nearly 3.5 years.

```
t.test(d_age_s, r_age_s)
```

```
##
##  Welch Two Sample t-test
##
## data:  d_age_s and r_age_s
```

```
## t = -3.8173, df = 1137.8, p-value = 0.0001422
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.177824 -1.662155
## sample estimates:
## mean of x mean of y
##  53.71768  57.13767
```

Sub-question2: Are Republican voters older or younger than Democratic voters for governor? (1) Statistical significance: We reject the null hypothesis that the average age of Republican and Democratic voters who voted for governor in 2018 is the same by conducting the two-side t-test. The result of the t-test is statistical significant because p-value = 3.199e-05 which is smaller than 0.05. (2) Practical significance: In practical significance, the average age of Democratic voters is 53.87268 and Republican voters is 57.50817 which are notably different by over 3.5 years.

```
t.test(d_age_g, r_age_g)
```

```
##
##  Welch Two Sample t-test
##
## data:  d_age_g and r_age_g
## t = -4.1732, df = 1328.7, p-value = 3.199e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.344469 -1.926513
## sample estimates:
## mean of x mean of y
##  53.87268  57.50817
```

Question 3: Do a majority of independent voters believe that the federal investigations of Russian election interference are baseless?

Introduce your topic briefly. (5 points)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

Measurement of the population disapproval of the federal investigations of the Russian election interference is operationalized in the variables `coord16` and more specifically `muellerinv` which ask for whether they believe coordination between Russia and Trump in 2016 and their level of approval towards Special Counsel Robert Mueller's investigation respectively. Using the `coord16` or `muellerinv` variables as an operationalization for identifying whether people believe that the Russian election interference investigation is reasonable does have some conceptual gaps. First, the `coord16` does directly ask respondents if they believe that collusion is probable, but this is not the same as asking about whether they believe an investigation is reasonable or baseless, as one can be unsure or even believe it is improbable that an event happened, but approve of an investigation to gain more evidence and be more confident. Second, the `muellerinv` asks for level approval with respect to the investigation but this may not be based solely on the allegation's reasonableness. Costs, sense of priority, and political strategy may all play a role in one's approval. To encode one's view on whether the investigation is baseless, we may use both these variables as baseless = probable coordination $\cap$ dissapproval of Mueller investigation

Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment

on any features you find.

> Given the high proportion of those for which the `pid1d` variable has an value of inapplicable or skipped, `pid7x` is a more appropriate choice for identifying independent identification. Its Likert scale also expresses degree of identification and can help represent any slight political leanings within the independent group as well. Independents in this variable are represented as `3,4,5` from closer to dem to closer to rep. Within this variable, there are 98 respondents for whom there is no answer, less than 4% of all responses. After filtering for those that indentify as independent, there are 937 survey responses.

```
table(anes$pid7x) # 7 value party identification scale 1-strong dem to 7-strong rep
```

```
##
##  -7   1   2   3   4   5   6   7
##  98 581 276 279 417 241 200 408
```

```
table(anes$pid1d) # Generally speaking where people affiliate themselves
```

```
##
##   -7   -1    1    2    3    4
##    1 1331  432  326  356   54
```

```
independent_surveys = anes[anes$pid7x %in% c(3,4,5),]
sprintf("There are %d surveys from independents", nrow(independent_surveys))
```

```
## [1] "There are 937 surveys from independents"
```

```
#summary(independent_surveys$coord16)
print("table of survey responses on Trump/Russia coordination")
```

```
## [1] "table of survey responses on Trump/Russia coordination"
```

```
table(independent_surveys$coord16)
```

```
##
##   1   2
## 458 479
```

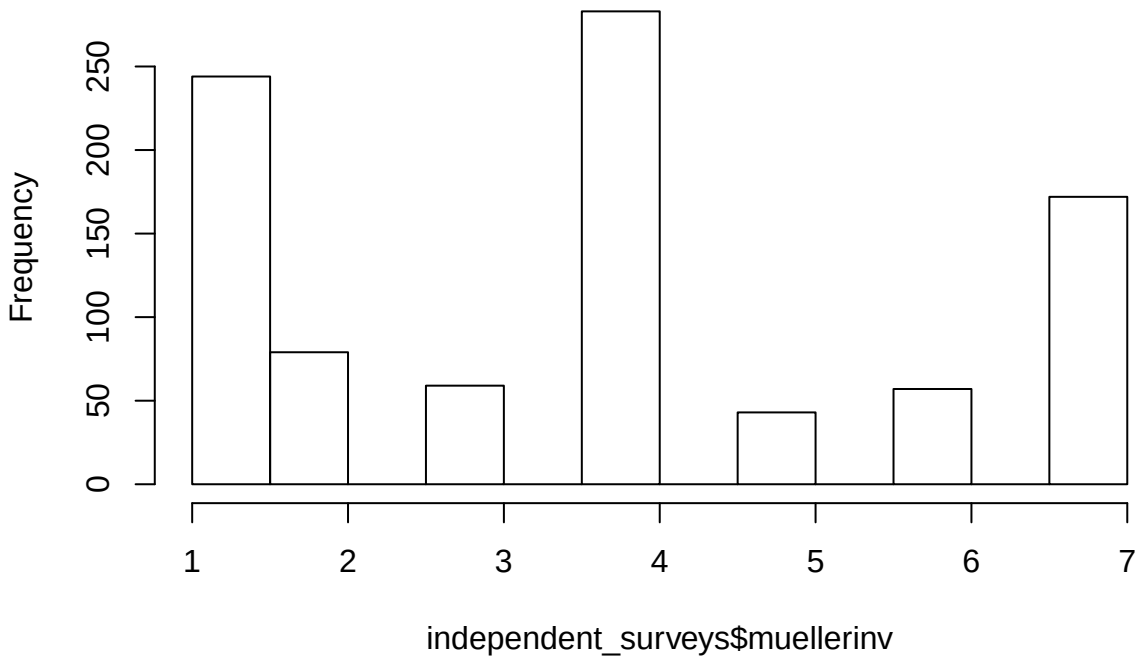
```
#summary(independent_surveys$muellerinv)
print("table of survey responses on approval of Mueller investigation")
```

```
## [1] "table of survey responses on approval of Mueller investigation"
```

```
table(independent_surveys$muellerinv)
```

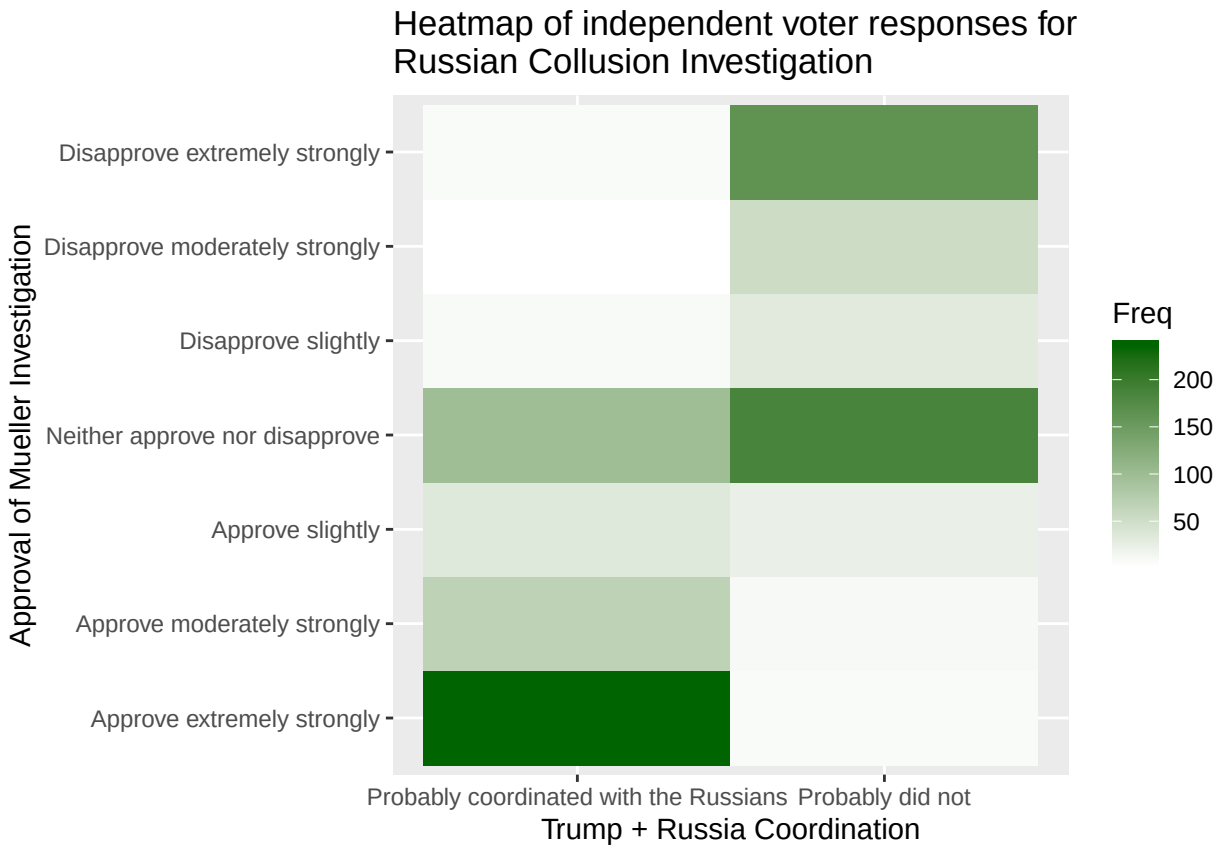
```
##
##   1   2   3   4   5   6   7
## 244  79  59 283  43  57 172
```

```
hist(independent_surveys$muellerinv, main="Histogram of response values to Mueller investigation")
```




```
coord16 = factor(independent_surveys$coord16, levels=c(1,2), labels=c("Probably coordinated with the Rus
muellerinv = factor(independent_surveys$muellerinv, levels=1:7, labels=c("Approve extremely strongly",
                                                                          "Approve moderately strongly",
                                                                          "Approve slightly",
                                                                          "Neither approve nor disapprove
                                                                          "Disapprove slightly",
                                                                          "Disapprove moderately strongly
                                                                          "Disapprove extremely strongly

df = as.data.frame(table(muellerinv, coord16))
ggplot(data=df, aes(coord16, muellerinv)) +
  geom_tile(aes(fill=Freq)) +
  scale_fill_gradient(low="white", high="darkgreen") +
  ggtitle("Heatmap of independent voter responses for\nRussian Collusion Investigation") +
  labs(x="Trump + Russia Coordination") +
  labs(y="Approval of Mueller Investigation")
```



As seen in the histogram of survey responses on the Mueller Investigation, there are three modal groups of responses around 1, 4, and 7. These responses correspond approximately to `approve extremely strongly`, `neither approve not dissapprove`, and `disapprove extremely strongly`. The heatmap represents the joint distribution of respondents feelings about the russia investigation and belief about whether the coordination did happen. The responses do seem to follow an expected distribution that generally those who did not think the collusion happened do not support the investigation, those who do believe collusion happened do support the investigation, and a neutral group. Given the nature of the question that is trying to answer whether a majority of independent voters believe that the investigation is baseless, the voters will be encoded as believing the investigation is baseless if they both dissapprove of the investigation and believe that coordination did not happen.

```
disapprove_and_not_probable = as.numeric(independent_surveys$coord16==2 & independent_surveys$muellerin
table(disapprove_and_not_probable)
```

```
## disapprove_and_not_probable
##   0   1
## 687 250
```

```
sprintf("Among independents, there are %d out of %d voters who rate Robert Mueller Investigation below
```

```
## [1] "Among independents, there are 250 out of 937 voters who rate Robert Mueller Investigation below
```

Given the nature of the question is asking about those whore are independents, it is especially unknown which direction the group will lean with regard to the federal investigations into russian election interference, therefore a two-sided t-test for the proportion of voters that rate Special Counsel Robert Mueller's investigation below neutral is appropriate. The null hypothesis is that the proportion of independents that have below neutral ratings of Robert Mueller is 0.5.

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

```
t.test(disapprove_and_not_probable, mu=0.5)
```

```
##
##  One Sample t-test
##
## data:  disapprove_and_not_probable
## t = -16.13, df = 936, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.2384376 0.2951804
## sample estimates:
## mean of x
##  0.266809
```

$H_O : \mu = 0.5; H_A : \mu \neq 0.5$. For this two sided t-test with null hypothesis that the proportion of independent voters who believe that the Russia investigation by special counsel Robert Mueller is baseless is 0.5, the t statistic of our data with 936 degrees of freedom is -16.13 and p-value of $2.2e-16$. This p-value is less than our α value of 0.05 and thus we thus reject the null hypothesis that the proportion is 0.5. Alternatively, the 95% confidence interval (0.2384, 0.2952) calculated does not include our null hypothesis value which also suggests rejecting the null. Lastly, we observe that with a sample mean of 0.2668 compared to the null of 0.5, this difference seems practically significant and suggests that only approximately 26.7% of independent voters in this sample actually do believe that the Mueller investigation is unreasonable. This result is opposite of the original research question which posed whether a majority of independents believe the investigation is baseless.

Question 4: Was anger or fear more effective at driving increases in voter turnout from 2016 to 2018?

Introduce your topic briefly. (5 points)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

The 'anger/fear' variable is operationalized in 3 questions in the survey that asked the respondent about their emotional feelings, which include anger and fear. The 3 questions are - 1. *Generally speaking, how do you feel about the way things are going in the country these days?* 2. *Think about Donald Trump. How often would you say you've felt each of the following ways because of the kind of person Donald Trump is or because of something he has done?* 3. *Think about immigrants coming from other countries to live in the United States. How often would you say you've felt each of the following ways because of immigrants coming from other countries to live in the United States?* To respond to these questions, the respondent will need to answer on a scale of 1 to 5, 1 being Not at all/Never, and 5 being Extremly/Always. To understand whether anger or fear was more effective at driving increases in voter turnout from 2016 and 2018, I will look at the 'new voters' who didn't vote in 2016, but did vote in 2018, and see if they have the same scale in anger or fear feelings. The gaps between the responses to these questions and the concept that we want to study are as follows- 1. This will take into account of people who were not eligible to vote in 2016 but became eligible to vote in 2018. Their emotional feelings was not the major deciding factor on their turnout. This can be solved by eliminating people who were born before 1998 from the sample. 2. The question didn't ask about their anger/fear in 2016. We only have the data as of 2018 when the survey was conducted, which asked about their current state of feelings. We won't be able to know what were their feelings back in 2016. Chances are

that the 'new voters' have been having the same feeling in 2016 and 2018, but decided to vote in 2018 for some other reasons. 3. Correlation does not necessarily mean causation. There could be an external factor that drove both people's willingness to vote and fear/anger level. It could also be the other way around that their turnout to vote made them angry/afraid.

Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

```
# First look at the total number of respondent who voted in 2016
voter_turnout16 <- anes[anes$turnout16 == 1, ]
length(voter_turnout16$turnout16)
```

```
## [1] 1841
```

```
# Then look at the total number of respondent who voted in 2018
voter_turnout18 <- anes[anes$turnout18 <= 3, ]
length(voter_turnout18$turnout18)
```

```
## [1] 1842
```

From the above result, we don't see a big increase in the number of voters from 2016 to 2018. However, we can look at those who didn't vote in 2016, and see how many of them voted in 2018. I will also remove those who were born after 1998 in the next step because they only became eligible to vote in 2018.

```
increased_voters <- anes[anes$turnout16 == 2 & anes$turnout18 <= 3 & anes$birthyr <= 1998, ]
length(increased_voters$turnout18)
```

```
## [1] 83
```

These 83 'new voters' can be seen as the operationalized concept of increases in voter turnout. In the following analysis, I will compare their score in anger and fear and see if there's a difference.

Responses from question "Generally speaking, how do you feel about the way things are going in the country these days?"

```
# Summary statistics of responses for anger
summary(increased_voters[increased_voters$geangry>0, 'geangry'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.927   4.000   5.000
```

```
# How many of them did not answer?
length(increased_voters$geangry[increased_voters$geangry<0])
```

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

```
# Summary statistics of responses for fear
summary(increased_voters[increased_voters$geafraid>0, 'geafraid'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.862   4.000   5.000
```

```
# How many of them did not answer?
length(increased_voters$geafraid[increased_voters$geafraid<0])
```

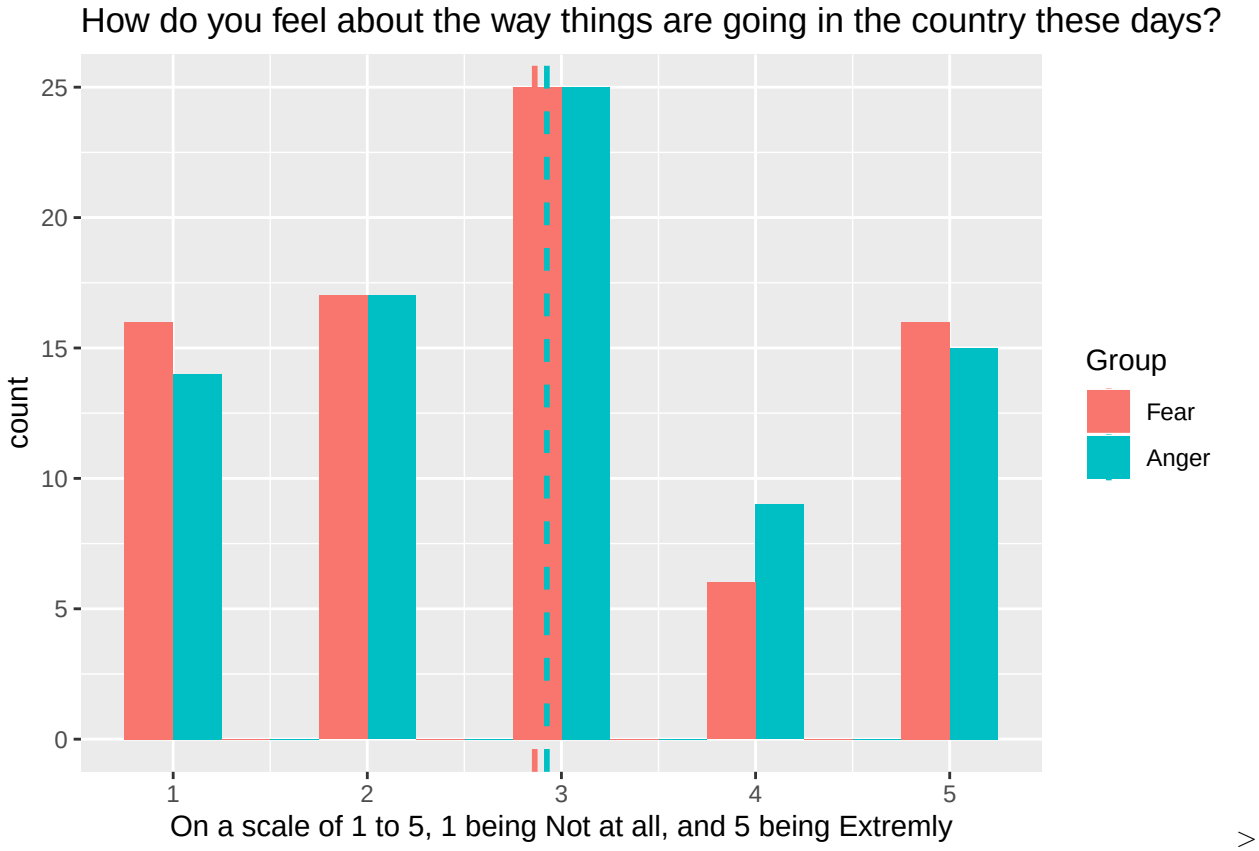
```
## [1] 3
```

```
# How many of them did answered both at the same time?
length(increased_voters$geafraid[increased_voters$geafraid > 0 & increased_voters$geangry > 0])
```

```
## [1] 80
```

Since there are only 3 of the increased voters didn't answer both anger and fear, I will remove them from the following analysis

```
Q4_sample1 <- increased_voters[increased_voters$geafraid > 0 & increased_voters$geangry > 0, ]
# check the distribution of the responses - fear vs. angry
Q4_plot_df <- rbind(data.frame(Group="Fear", rating = Q4_sample1$geafraid),
           data.frame(Group="Anger", rating = Q4_sample1$geangry))
Q4_Group_by_mean <- rbind(data.frame(Group="Fear", avg_rating = mean(Q4_sample1$geafraid)),
           data.frame(Group="Anger", avg_rating = mean(Q4_sample1$geangry)))
ggplot(Q4_plot_df, aes(x = rating, fill = Group)) +
  geom_histogram(binwidth=.5, position="dodge") +
  geom_vline(data=Q4_Group_by_mean, aes(xintercept = avg_rating,  colour = Group),
             linetype="dashed", size=1)+
  labs(title="How do you feel about the way things are going in the country these days?") +
  labs(x="On a scale of 1 to 5, 1 being Not at all, and 5 being Extremly")
```



The histogram above represents the distribution of frequency score for fear/anger feelings on a scale of 1 to 5, adding the two dotted lines as the average frequency score for each of the group. From the graph above, we can see that the distribution does not follow a normal curve, and the difference between the fear and anger feeling was very small.

Responses from question "Think about Donald Trump. How often would you say you've felt each of the following ways because of the kind of person Donald Trump is or because of something he has done?"

```
# Summary statistics of responses for anger
summary(increased_voters[increased_voters$dtangry>0, 'dtangry'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.500   3.000   3.395   5.000   5.000
```

```
# How many of the responses were no answer or skip?
length(increased_voters$dtangry[increased_voters$dtangry<0])
```

```
## [1] 40
```

```
# Summary statistics of responses for fear
summary(increased_voters[increased_voters$dtafraid>0, 'dtafraid'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.933   4.000   5.000
```

```
# How many of the responses were no answer or skip?
length(increased_voters$dtafraid[increased_voters$dtafraid<0])
```

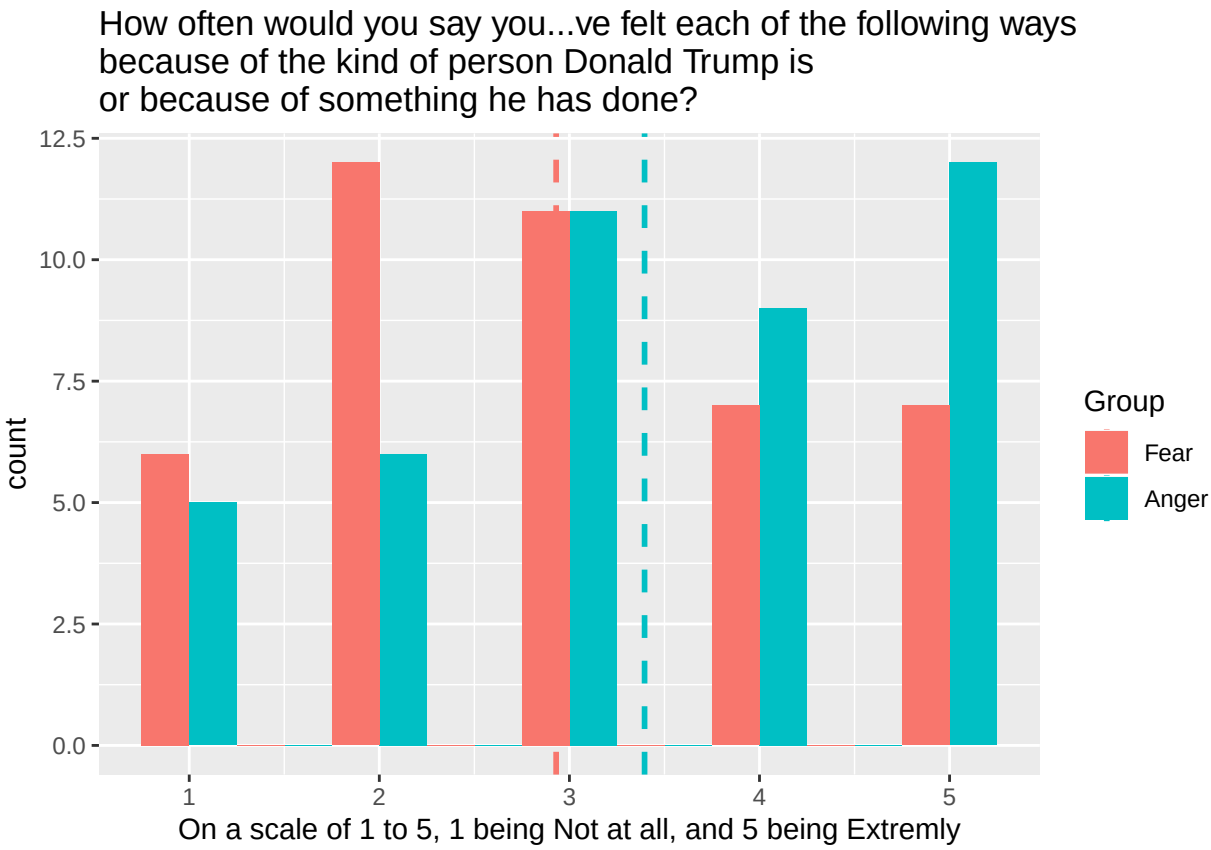
```
## [1] 38
```

```
# How many of them answered both at the same time?
length(increased_voters$dtafraid[increased_voters$dtafraid > 0 & increased_voters$dtangry > 0])
```

```
## [1] 43
```

Even though 40 increased voters did not respond to both anger and fear, the sample size for people who answered both of the questions is greater than 30, which makes it suitble for the following test.

```
Q4_sample2 <- increased_voters[increased_voters$dtafraid > 0 & increased_voters$dtangry > 0, ]
# check the distribution of the responses - fear vs. angry
Q4_plot_df2 <- rbind(data.frame(Group="Fear", rating = Q4_sample2$dtafraid),
            data.frame(Group="Anger", rating = Q4_sample2$dtangry))
Q4_Group_by_mean2 <- rbind(data.frame(Group="Fear", avg_rating = mean(Q4_sample2$dtafraid)),
            data.frame(Group="Anger", avg_rating = mean(Q4_sample2$dtangry)))
ggplot(Q4_plot_df2, aes(x = rating, fill = Group)) +
  geom_histogram(binwidth=.5, position="dodge") +
  geom_vline(data=Q4_Group_by_mean2, aes(xintercept = avg_rating,  colour = Group),
               linetype="dashed", size=1)+
  labs(title="How often would you say you've felt each of the following ways \nbecause of the kind of p
  labs(x="On a scale of 1 to 5, 1 being Not at all, and 5 being Extremly")
```



The histogram above represents the distribution of frequency score for fear/anger feelings on a scale of 1 to 5, adding the two dotted lines as the average frequency score for each of the group. From the graoh above, we can see that the distribution does not follow a normal curve either. However, the distribution for 'fear' feeling skewed towards right, and the distribution for 'anger' feeling skewed towards left. There's a much bigger difference in the average frequency score.

Responses from question "Think about immigrants coming from other countries to live in the United States. How often would you say you've felt each of the following ways because of immigrants coming from other countries to live in the United States?"

```
# Summary statistics of responses for anger
summary(increased_voters[increased_voters$imangry>0, 'imangry'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.105   3.000   5.000
```

```
# How many of the responses were no answer or skip?
length(increased_voters$imangry[increased_voters$imangry<0])
```

```
## [1] 45
```

```
# Summary statistics of responses for fear
summary(increased_voters[increased_voters$imafraid>0, 'imafraid'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.895   3.000   5.000
```

```r
# How many of the responses were no answer or skip?
length(increased_voters$imafraid[increased_voters$imafraid<0])
```

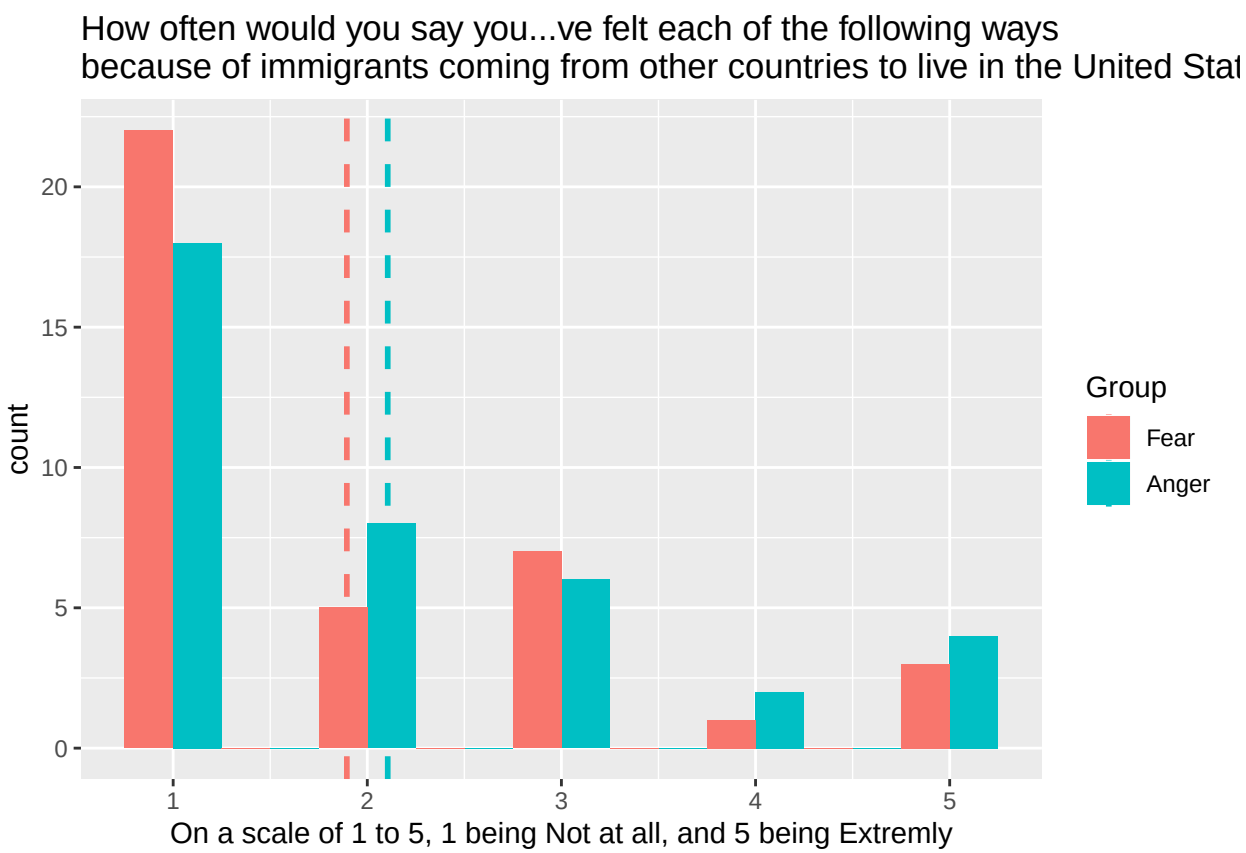
```
## [1] 45
```

```r
# How many of them answered both at the same time?
length(increased_voters$imafraid[increased_voters$imafraid > 0 & increased_voters$imangry > 0])
```

```
## [1] 38
```

Same as above, Even though 45 increased voters did not respond to both anger and fear, the sample size for people who answered both of the questions is greater than 30, which makes it suitble for the following test.

```r
Q4_sample3 <- increased_voters[increased_voters$imafraid > 0 & increased_voters$imangry > 0, ]
# check the distribution of the responses - fear vs. angry
Q4_plot_df3 <- rbind(data.frame(Group="Fear", rating = Q4_sample3$imafraid),
            data.frame(Group="Anger", rating = Q4_sample3$imangry))
Q4_Group_by_mean3 <- rbind(data.frame(Group="Fear", avg_rating = mean(Q4_sample3$imafraid)),
            data.frame(Group="Anger", avg_rating = mean(Q4_sample3$imangry)))
ggplot(Q4_plot_df3, aes(x = rating, fill = Group)) +
  geom_histogram(binwidth=.5, position="dodge") +
  geom_vline(data=Q4_Group_by_mean3, aes(xintercept = avg_rating,  colour = Group),
               linetype="dashed", size=1)+
  labs(title="How often would you say you've felt each of the following ways \nbecause of immigrants co
  labs(x="On a scale of 1 to 5, 1 being Not at all, and 5 being Extremly")
```



The histogram above represents the distribution of frequency score for fear/anger feelings on a scale of 1 to 5, adding the two dotted lines as the average frequency score for each of the group.

In this case, both of the distributions skewed towards right.

Based on your EDA, select an appropriate hypothesis test. (5 points)

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

I will be using *Wilcoxon signed test* for the first sample for the following reasons - 1. There is a natural pairing between the two data points because they are collected from the same individual. 2. From the histograms above, none of the two measurements have a normal curve. Nonparametric tests can be used as distribution-free tests, and data has already been ranked from lowest to highest across the two measurements. The assumptions will then be - * Null hypothesis H_0: The difference between the two responses is 0 * Alternative hypothesis H_a: The difference between the two responses is not 0

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

```
# Wilcoxon signed test for sample 1 - feelings towards country today
wilcox.test(Q4_sample1$geafraid, Q4_sample1$geangry, paired = TRUE, exact = FALSE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Q4_sample1$geafraid and Q4_sample1$geangry
## V = 355, p-value = 0.8179
## alternative hypothesis: true location shift is not equal to 0
```

```
# Wilcoxon signed test for sample 2 - feelings towards Donald Trump
wilcox.test(Q4_sample2$dtafraid, Q4_sample2$dtangry, paired = TRUE, exact = FALSE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Q4_sample2$dtafraid and Q4_sample2$dtangry
## V = 48, p-value = 0.008414
## alternative hypothesis: true location shift is not equal to 0
```

```
# Wilcoxon signed test for sample 3 - feelings towards immigrants
wilcox.test(Q4_sample3$imafraid, Q4_sample3$imangry, paired = TRUE, exact = FALSE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Q4_sample3$imafraid and Q4_sample3$imangry
## V = 25, p-value = 0.2782
## alternative hypothesis: true location shift is not equal to 0
```

Based on the test above, we can reject the null hypothesis that there's no difference between US voter's anger and fear feelings towards Donald trump(*sample 2, p-value is less than .05*). However, we cannot reject the null hypothesis that there's no difference between US voter's anger and fear feelings towards the country(*sample 1, p-value is greater than .05*) and immigrants(*sample 3, p-value is greater than .05*). What does this mean? This means that for people who could've voted in 2016 but didn't and decided to vote in 2018, their anger towards Donald Trump is statistically higher than their fear towards Donald Trump. Referring back to the histgram from our EDA above, we can see that there are more respondents having extreme feelings(*scale 4&5*) of anger than fear, and less respondents have no feelings(*scale 1&2*) of anger than fear. This

strong angry feeling towards Donald Trump could've been the driver for them to vote. Another drawback from this test is that we have ommited those who chose not to answer the corresponding questions in the survey. In the above test, we assumed that those who didn't answer the questions were similar to those who did answer the questions, which is absolutely not a sound assumption. However, we have no way to figure out why these people chose not to answer the questions. It could be that they had "more than extreme" feelings towards the question, or very neutural to the questions that they don't have a 'feeling' at all. For whatever reason it is, we couldn't figure it out from the survey itself. That's why we will need to collect more data and do further analysis to determine whether dropping out the 'non-respondents' could affect the test results. **To conclude, US voters anger towards Donald Trump could have been an effective driver that increased voter turnout from 2016 to 2018. However, further tests and analysis need to be done to prove the causal inference between them.**

Question 5: Select a fifth question that you believe is important for understanding the behavior of voters

Clearly argue for the relevance of this question. (10 points)

In words, clearly state your research question and argue why it is important for understanding the recent voting behavior. Explain it as if you were presenting to an audience that includes technical and non technical members.

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

> Question: Are republican voters more or less likely to persuade others to voter than democratic voters? This queistion is important because if one party's voter is more likely to persuade others to vote, we know that party's voter is probably more passionated about the voting and small number of voters can snowball into large number of voters. Also, if the voters are more likely to convince others to vote. The candidate might be more advantagious at a close campaign.

variable: pid1d. This variable help me identify if the voter is republican or democrat.

Variable: persuade. During past 12 month, have you tried to persuade anyone to vote one way or another? This variable captures what I want to get fairly well. However, the question is asking if the voter has tried. So we are not sure aout the success rate of the result. Maybe certain party's voters are more willing to try but have lower success rate.

Perform EDA and select your hypothesis test (5 points)

Perform an exploratory data analysis (EDA) of the relevant variables.

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

Based on your EDA, select an appropriate hypothesis test. Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

```
### Variable definition
### 1 - I have done this in past 12 month, 2 - I have not done this in the past 12 months
### modify no to 0 so mean is the % of people who tried to persuade

anes[anes$persuade == 2,]$persuade = 0
unique(anes$persuade)
```

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

```r
table(anes$persuade)
```

```
##
##    0    1
## 1568  932
```

```r
summary(anes$persuade)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3728  1.0000  1.0000
```

```r
### variable definition
### 1 - Democrat, 2 - Republican, 3 - Independent, 4 - Something else
unique(anes$pid1d)
```

```
## [1]  2 -1  3  1  4 -7
```

```r
table(anes$pid1d)
```

```
##
##   -7   -1    1    2    3    4
##    1 1331  432  326  356   54
```

```r
count = table(anes$pid1d, anes$persuade)

count/rowSums(count)
```

```
##
##              0         1
##   -7 1.0000000 0.0000000
##   -1 0.6393689 0.3606311
##   1  0.5185185 0.4814815
##   2  0.6319018 0.3680982
##   3  0.7106742 0.2893258
##   4  0.6111111 0.3888889
```

Conduct your test. (2 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result.

```r
t.test(anes[anes$pid1d == 1,]$persuade, anes[anes$pid1d == 2,]$persuade, paired = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  anes[anes$pid1d == 1, ]$persuade and anes[anes$pid1d == 2, ]$persuade
## t = 3.1508, df = 712.19, p-value = 0.001696
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.04273317 0.18403347
## sample estimates:
## mean of x mean of y
## 0.4814815 0.3680982
```

```r
### the p-value of the test is 0.0017 which is significant. So I would reject the null hypothesis that

### This is also practical significant. The ratio of democratic voters who tried to persuade others is
```

Conclusion (3 points)

Clearly state the conclusion of your hypothesis test and how it relates to your research question.

Finally, briefly present your conclusion in words as if you were presenting to an audience that includes technical and non technical members.

> The student t-test was significant. And the result indicates that the democratic voters and republican voters are not equally likely to persuade others to vote. 48.1% of democratic voters tried to convince others to vote and 36.8% of republican voters tried to convince others to vote. The result is really interesting. Even though the test was the two-tail test. It is likely that the democratic voters are more likely to try to convince others to vote.