Lab 2

W203 Statistics for Data Science

Team 1
Section 6

General data analysis:

Data quality is very important in statistical analysis. Since incorrect data can lead to incorrect results, we wanted to only perform analysis on individuals whose responses were trustworthy. Namely, we pre-filtered our participants based on their "RESPONSE QUALITY" questions. We only considered those who were "Never" or "Some of the time" insincere in their responses, and those who were answered questions honestly "Most of the time" or "Always." For those who were most of the time honest and sincere, a majority of their answers were reliable, so they will provide correct information more often than incorrect information. To do this, we generated a new dataset stored under A. The reduced the sample from 2500 samples to 2064 samples:

```r
setwd("~/Desktop/Stone/Berkeley_MIDS/Statistics/github/lab_2/")
full_data <- read.csv('anes_pilot_2018.csv')

library(ggplot2)
suppressMessages(library(dplyr))

A <- full_data %>% filter(honest >= 4 & nonserious <= 2)
```

Question 1

Introduce your topic briefly.

The relevant questions asked in this study are stored under ftjournal and ftpolice variables, which represent responses from the participants on how highly they would rate journalists and police respectively. Responses range from 1 to 100, but the survey was presented as a thermometer with 9 labels ranging from very cold (numerically 0) and very warm (numerically 100). We feel that because of this presentation, the rankings are similar to a Likert scale, meaning we do not have a metric variable. For example, we do not believe that those who responded 75 was different than those who responded 80, both numbers are between the "Fairly Warm" and "Quite Warm categories." However, those who responded 86 would favor a particular group more than someone who responded 84, because the one who ranked 86 made a conscious choice to rate the group as "Quite Warm," whereas the one who rated 84 made the choice to rate the group as "Fairly Warm." As a result, even though the numeric gap is larger in the first case, we feel it is less indicative of a difference in perception than the small gap in the second case, where the gap crosses the boundary between two thermometer levels. As a result, we plan to bin respondents into 1 of 9 categories as defined by the survey scale, and give each bin a rank starting from 1 and giving to 9. The lower the rank, the lower the respondents rated each group.

One deficiency here is that participants are asked to "rate" a group, which does not necessarily translate to respect for the group. For example, one can rate the police highly because one thinks the group is effective at responding to emergencies; however, one may not have a high level of respect for them if one believes that they sometimes perform their jobs unethically (such as discriminating against certain racial group). However, we do not have better data to access respect in this survey, so we must make do with these variables.

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

We wanted to answer this question from the perspective of each individual voter, as each voter was asked how they would rate the police versus journalists (a natural pairing). Namely, for each voter, do they respect the police or journalists more? To address this, we require that all respondents answer both questions, and filtered out those that did not answer either or both question (response -7). Also, to narrow the population down to US voters, we only looked at individuals who were registered to vote. These are many reasons why someone might choose to not vote in an election, including lack of interest in the voter decision, lack of strong support for either opposition, or simply forgetting. However, if one is eligible to vote, they should be classified as US voters, since these individuals have the ability and opportunity to cast their decision in upcoming elections. There were only two individuals who did not respond to ftjournal, and all responded to ftpolice. In addition, 316 were not registered to vote. This reduces our dataset to 1746 participants, still a sizable number.

```
#A %>% filter(ftjournal >=0 & ftpolice >=0) %>% dim
A %>% filter(ftjournal >=0 & ftpolice >=0) %>% filter(reg<=2) %>% dim
```

```
## [1] 1746  767
```

```
JR_PL_data <- A %>%
  filter(ftjournal >=0 & ftpolice >=0) %>%
  filter(reg<=2) %>% select(caseid, ftpolice, ftjournal)
#Ensures we get the same bins as in the study
ftbins <- c(-0.01,15,30,40,50,60,70,85,99.99,100)

JR_PL_ordinal_data <- data.frame(
  cut(JR_PL_data$ftpolice, breaks=ftbins) %>% as.numeric(),
  cut(JR_PL_data$ftjournal, breaks=ftbins) %>% as.numeric()
)
colnames(JR_PL_ordinal_data) <- c("Police Rating", 'Journalist Rating')

JR_PL_ordinal_data$Police_minus_Journalist <-
  JR_PL_ordinal_data$`Police Rating` -
  JR_PL_ordinal_data$`Journalist Rating`
```

“‘

In order to perform EDA on the distribution of the difference between ranks, we have first look at summary statistics for the difference of Police rating minus Journalist rating for each voter.

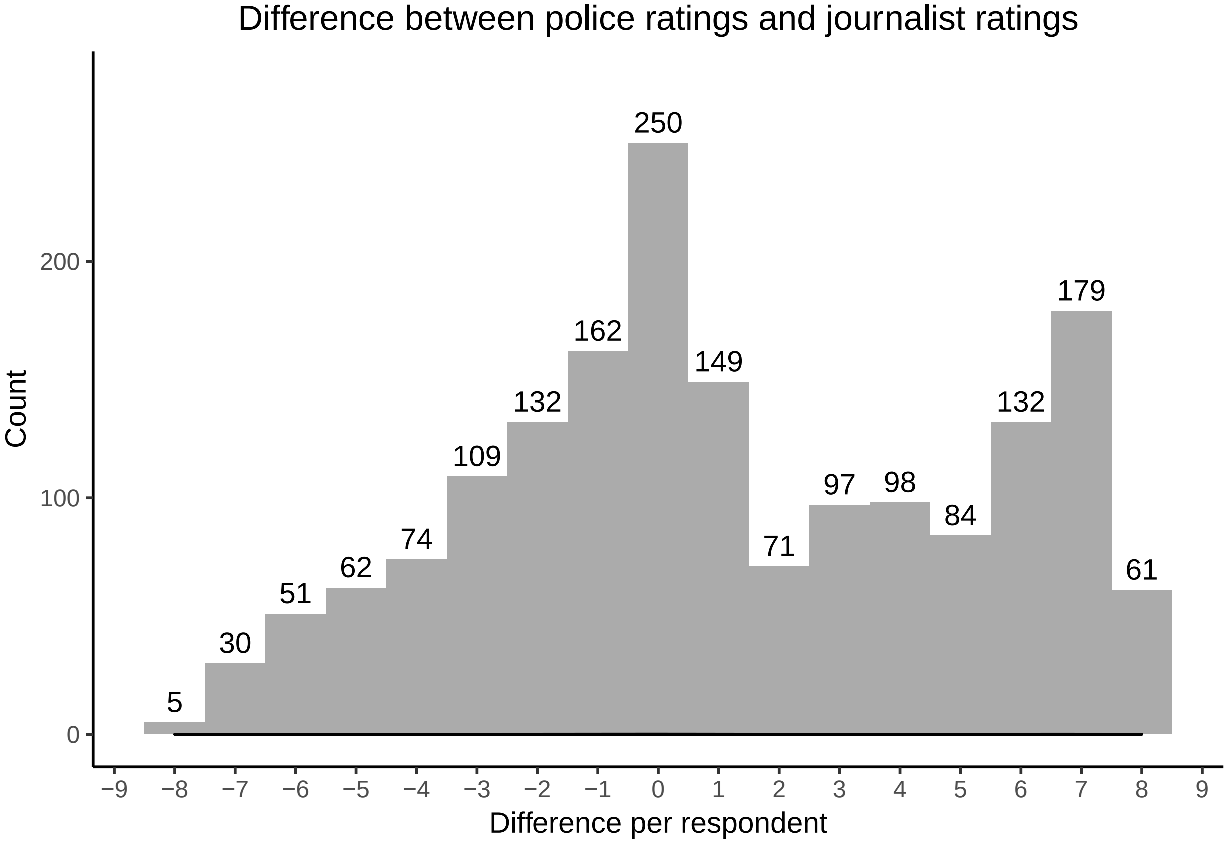
```
summary(JR_PL_ordinal_data$Police_minus_Journalist)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -8.000  -2.000   0.000   1.152   5.000   8.000
```

We see that the difference takes on both negative and positive values with the same minimum and maximum difference in ranks, so there is not a unanimous agreement across our sample on which group is rated higher. In fact, there are individuals who rated Police 1 and Journalists 9, and those who rated Police 9 and journalists 1. However, our 3rd Quartile is larger in magnitude than our 2nd Quartile, with the mean shifted upwards, indicating that there is skew toward the side of the police.

To visualize this skew, we plot a histogram of the difference in ranks.

```
ggplot(data = JR_PL_ordinal_data, aes(x=Police_minus_Journalist)) +
  geom_histogram(alpha=0.5,
                 binwidth = 1) +
  geom_density(alpha=0.5)+
  labs(title='Difference between police ratings and journalist ratings',
       x='Difference per respondent',
       y = "Count") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))+
  ylim(0,275)+
  scale_x_continuous(breaks=seq(-10, 10, by=1))+
  stat_bin(aes(y=..count.., label=ifelse(..count.. > 0, ..count.., "")),
           geom="text",
           vjust=-.5,
           binwidth=1
           )
```



We see that while the distribution mode is 0, but there are sharp peaks in the positive region skewing the data to the right (meaning these individuals rated the police higher than the journalists).

Based on your EDA, select an appropriate hypothesis test. The most appropriate test here is the Wilcoxon signed-rank test. The assumptions of this test are:

1. Data is paired, but each pair is iid. The data is certainly paired since it is the same respondent on two different questions. The samples certainly represent the pool of voters who responded to the survey, even if it does not represent all eligible US voters. We do not expected one participants responses from this survey to affect response from others, so we also safely claim independence in the responses.

2. The differences are measured on at least an ordinal scale. Note that since this is the case, even though sample size is large, we cannot use the paired t-test. We have binned our variables into an ordinal scale that we believe best reflects the meaning in the respondent's answers. As a result, we can subtract the ranks and look at, for each respondent, what the difference between ranking of each group is between police rating and journalist rating.

3. The distribution is symmetric around its mean and median. This is true under our null hypothesis, which is that the difference between the ranks of the two groups is 0. Possible values that the difference takes from integers going from -8 to +8.

More explicitly, let D represent the r.v. of the difference between ratings of the police and ratings of journalists. Finally, we have to decide on our allowed probability of type I error, or equivalently, the level of significance.

alpha = 0.05 is a commonly used value, which is what we will set here.

$$H_0 : D = 0$$
$$H_a : D \neq 0$$

To conduct the test, we use the following:

```
wilcox.test(JR_PL_ordinal_data$`Police Rating`, JR_PL_ordinal_data$`Journalist Rating`, paired = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  JR_PL_ordinal_data$`Police Rating` and JR_PL_ordinal_data$`Journalist Rating`
## V = 748390, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

We see that there's a very large statistical difference between the rating of the two groups, with a p-value << 0.001, and much less than alpha. This means that since p-value = 2.2e-16, if H0 was true, we had a 2.2e-16 chance of seeing data as least as extreme as what was observed. This is an extremely low value. We can then reject our null hypothesis in favor of the alternative that the difference in rating between the two groups is in fact not 0. Since the paired difference shows that the rating for police is typically higher than for journalists, the answer to our original question is that US voter have more respect for the police than they do for journalists.

In order to compute the effect size, we will use the proportion approach described in the Async where we look at:

$$r = \frac{\sum \text{positive ranks} - \sum \text{negative ranks}}{\sum \text{all ranks}}$$

Our proportion approach produces an effect size

```
(sum(JR_PL_ordinal_data$Police_minus_Journalist > 0) -
sum(JR_PL_ordinal_data$Police_minus_Journalist < 0)) /
length(JR_PL_ordinal_data$Police_minus_Journalist)
```

```
## [1] 0.1408935
```

This means that approximately 14% more of total US voters from this sample rated police higher than those that rated police lower (about 50% ranked police higher, and 36% ranked journalists higher). This is on the same scale as the correlation r, and between a small and medium effect size. While the test is highly statistically significant, we see that about 14% more of the total sample ranked police higher than journalists, which is a decent proportion of individuals, but also not gigantic.

question 2

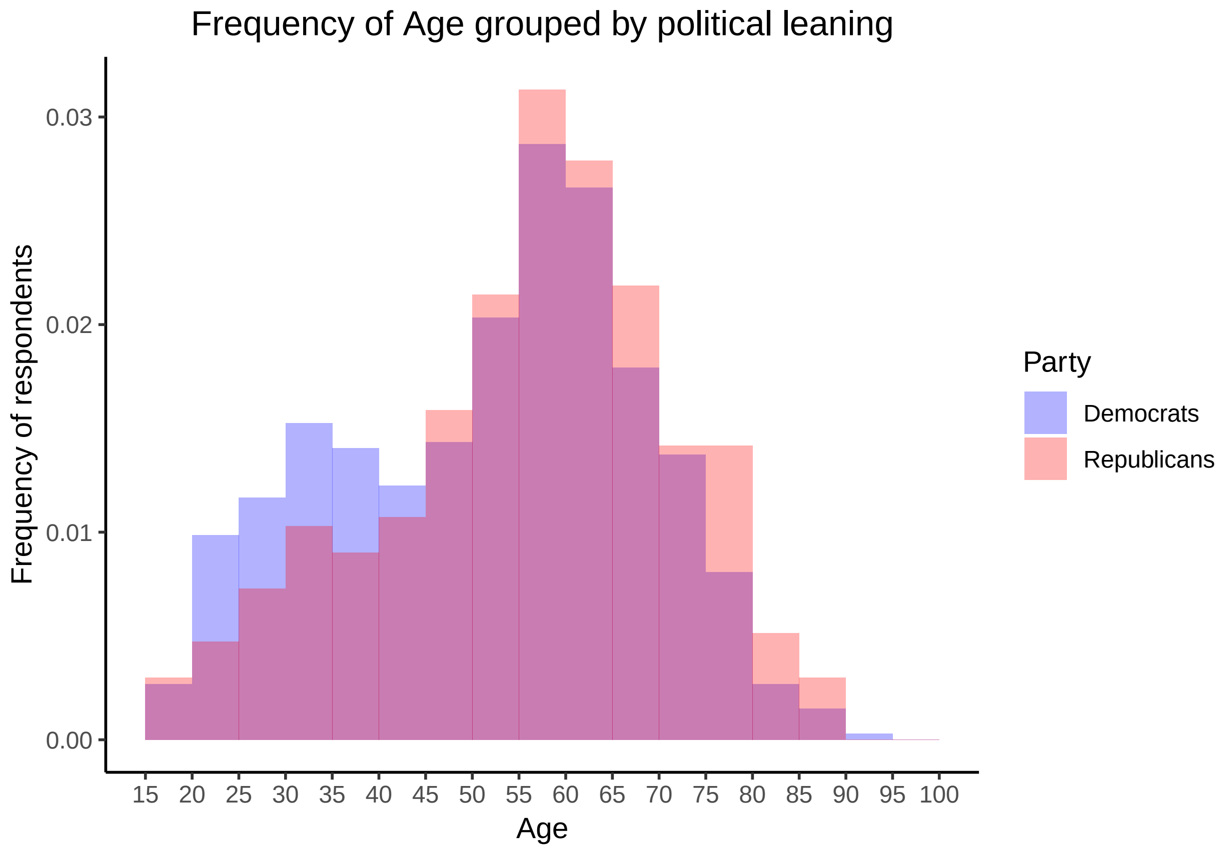
Introduce your topic briefly. We will use self-classified categories from pid7x. The reason we chose this over how participants actually voted in 2016 and 2018 is that sometimes in certain elections, self-identified Democrats can vote Republican and vice-versa. Blah blah blah

```
AG_PY_data <- A %>% mutate(party =
                                  ifelse(pid7x == 1 | pid7x == 2, 'D',
                                         ifelse(pid7x == 6 | pid7x == 7, 'R', 'I'))) %>%
  mutate(age = 2018 - birthyr) %>%
  filter(party != 'I', reg == 1 | reg == 2) %>%
  select(age, party)
```

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

We can plot a histogram of the ages differentiated by parties. We see that there appears to be a large population from 20 to 40 for democrats

```r
ggplot(data = AG_PY_data) +
  geom_histogram(aes(x=age, y=(..density..), fill=party),
                 alpha=0.3,
                 breaks=seq(15,100,5),
                 position='identity') +
  labs(title='Frequency of Age grouped by political leaning',
       x='Age',
       y = "Frequency of respondents",
       fill = "Party") +
  scale_fill_manual(labels = c("Democrats", "Republicans"),
                    values=c("blue1", "red1"))+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks=seq(15,100, by=5))
```



Based on your EDA, select an appropriate hypothesis test

Independent two sample t test. Metric variable, large sample size.

H0 is that the two samples have the same means H1 is that the means are not equal

```r
t.test(AG_PY_data[AG_PY_data$party == 'D',]$age,
       AG_PY_data[AG_PY_data$party == 'R',]$age,
       paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  AG_PY_data[AG_PY_data$party == "D", ]$age and AG_PY_data[AG_PY_data$party == "R", ]$age
## t = -4.1128, df = 1025, p-value = 4.221e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.796550 -2.051946
## sample estimates:
## mean of x mean of y
##  52.56502  56.48927
```

Effect size…

Question 3

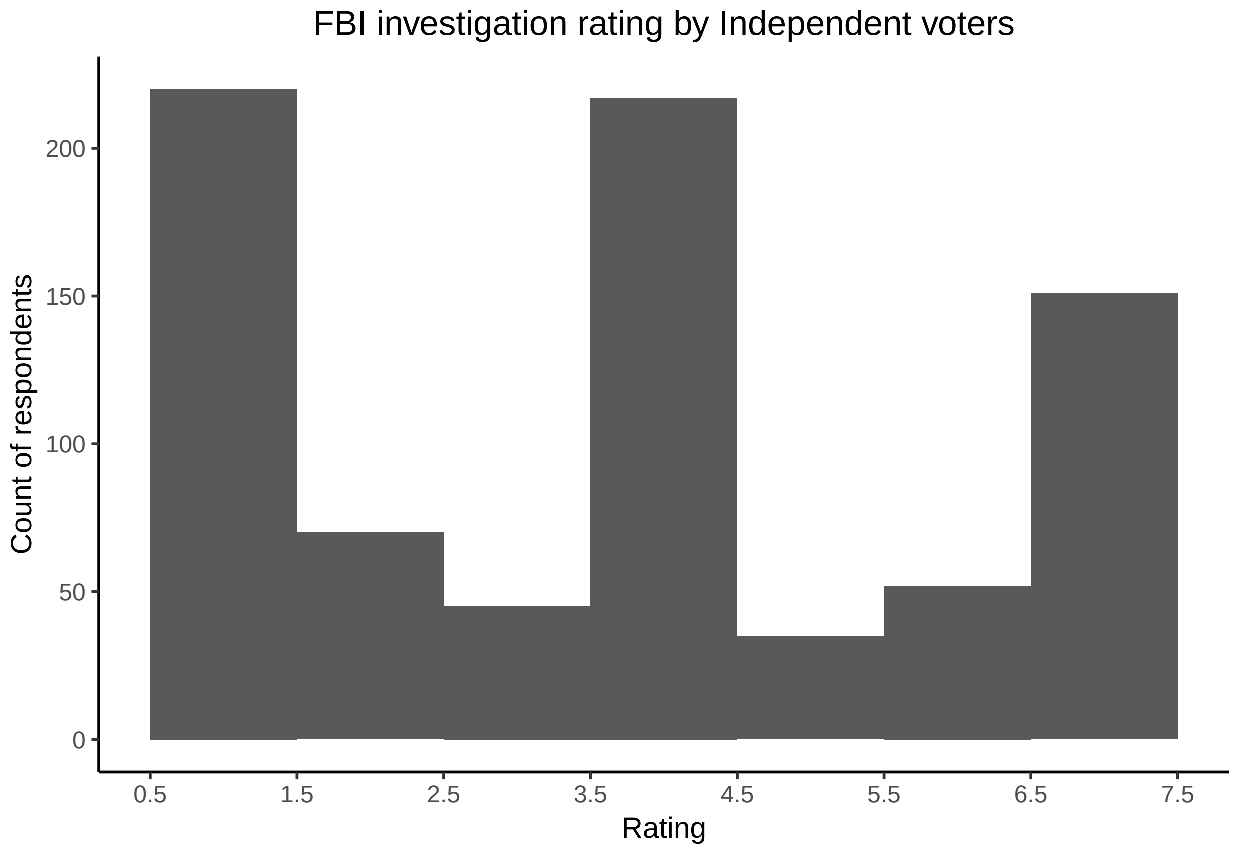
Introduce your topic briefly.

We will again take only those who are registered to vote, and this time identifies as Independent. The response we will analyze is whether these individuals approve of the Mueller investigations stored in the muellerinv variable. To address the question, we will consider a response of 5,6, or 7 as thinking the investigation was baseless (or strongly disapprove the decision to mildly disapprove the investigation). This question asks whether the respondent approves or disapproves which may not fully capture whether the respondent thought the investigation was baseless. They could have found a reason for the investigation but did not like the approach taken for example. However, this is the closest we could get.

```r
FBI_IN_data <- A %>% filter(pid7x <= 5 & pid7x >=3) %>%
  select(muellerinv)
```

We can see from the histogram that the distribution peaks 3 times, at strongly disapprove, strongly approve, and "Neither approve nor disapprove," although there appears to be a smaller population that is strongly disapprove.

```r
ggplot(data = FBI_IN_data) +
  geom_histogram(aes(x=muellerinv, y=(..count..)),
                 breaks=seq(0.5,7.5,1)) +
  labs(title='FBI investigation rating by Independent voters',
       x='Rating',
       y = "Count of respondents")+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks=seq(0.5,7.5, by=1))
```



Best test is the Wilcoxon Rank-Sum test because we have an ordinal variable with no metric structure. It is also extremely skewed to the tails for ratings of 1 and 7.

H0 is that the mean is 4 (think it is baseless) H1 is that mean is not 4.

```
wilcox.test(FBI_IN_data$muellerinv, mu = 4)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  FBI_IN_data$muellerinv
## V = 67364, p-value = 0.0001008
## alternative hypothesis: true location is not equal to 4
```