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. Namely, 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" and "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:

```
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 were this study are stored under ftjournal and ftpolice, 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. As a result, we plan to bin respondents into 1 of 9 categories as defined by the scale.

One deficiency here is that "rating" a group does not necessary mean respect for the group. One can think that a group is effective at their job, but still not have respect for them, for example, if they believe they sometimes perform their jobs unethically. However, we do not have better data to access respect.

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

We wanted to answer this question from the perspective of each individual voter. 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. Even if they did not vote in a particular election, we know that they were least eligible to. There were only two individuals who did not respond to ftjournal, and all responded to ftpolice. In addition, 316 were not registered to vote.

```
A %>% filter(ftjournal >=0 & ftpolice >=0) %>% dim
```

```
## [1] 2062  767
```

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

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

```r
JR_PL_data <- A %>%
  filter(ftjournal >=0 & ftpolice >=0) %>%
  filter(reg<=2) %>% select(caseid, ftpolice, ftjournal)
```

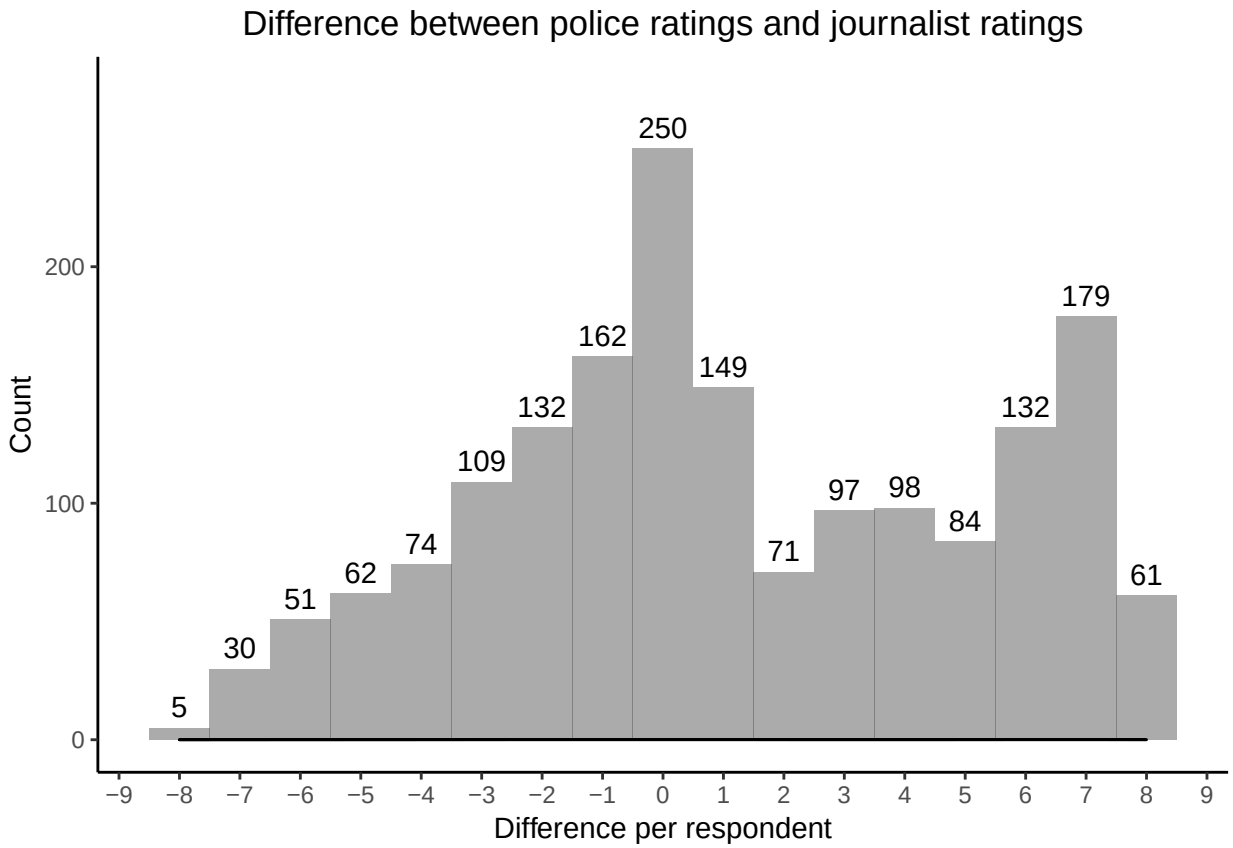
We will bin responses into 1 of 9 categories as presented in the study giving each bin a rank. This is because we cannot confidently say that a score of 99 is definitively better than 98, since they both fall within "Quite Warm" and "Very Warm." The lower the rank, the lower the respondents rated each group.

```r
#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`

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 as well (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.
2. Within-pair differences have meaning. The differences between 99 and 0 response (corresponding to bin 1 versus 9) is larger than between 50 and 0 (bins 1 and 4). We do not want 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. I think the flowchart is wrong.

H0 is difference is 0 Ha is that difference is non-zero

```
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
```

Very large statistical difference. Effect size...

question 2

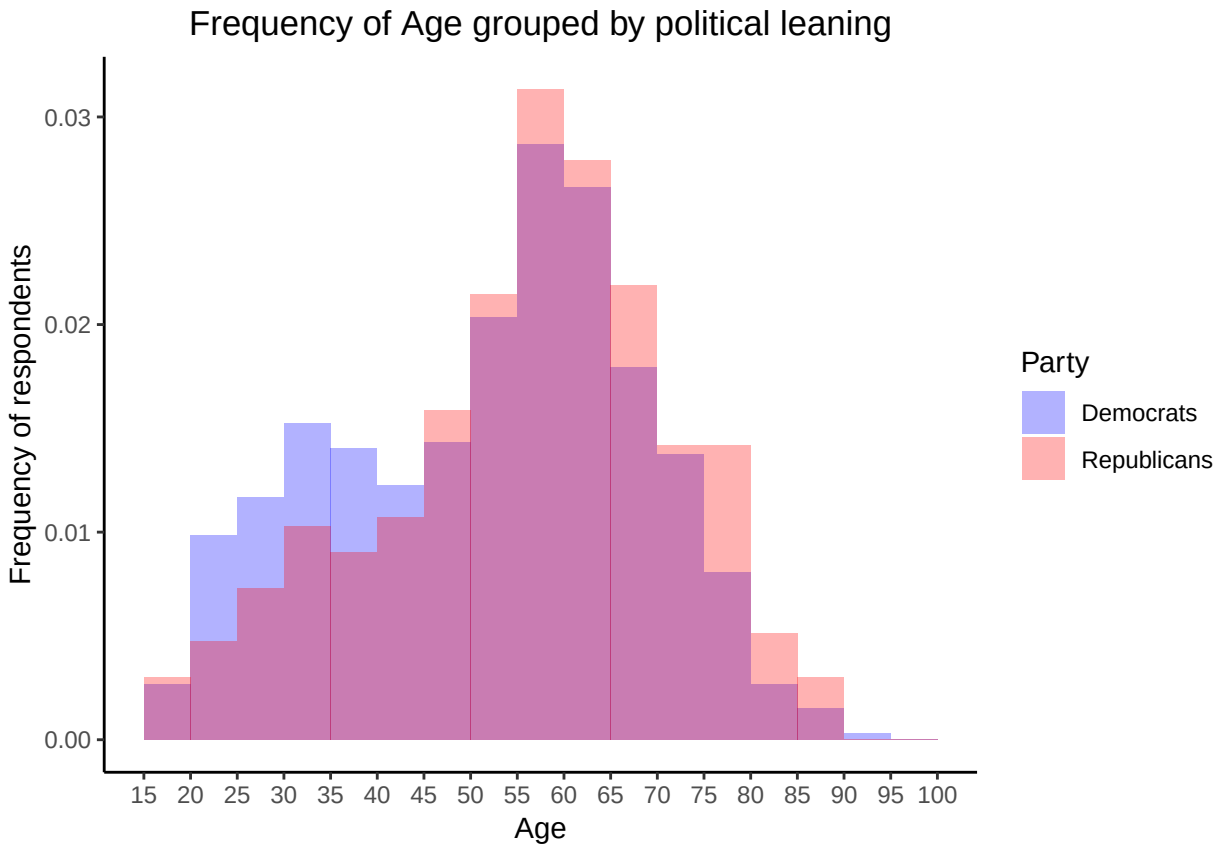
Introduce your topic briefly. We will use self-classified categories from pid7x. The reason we chose this over how participants actually voted in 2016 and 2018 is that sometimes in certain elections, self-identified Democrats can vote Republican and vice-versa. Blah blah blah

```
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

```
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

```
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.

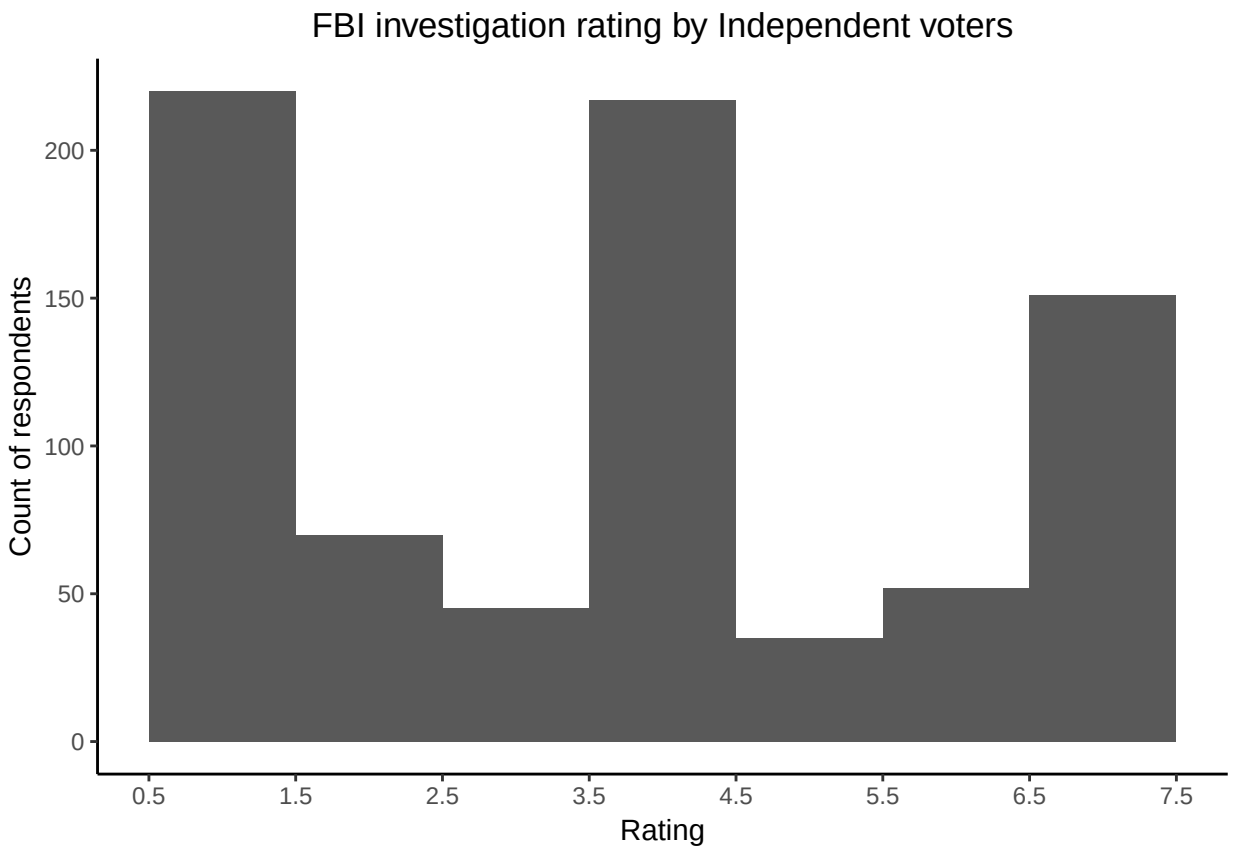
We will again take only those who are registered to vote, and this time identifies as Independent. The response

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.

```
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.

```
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.

```r
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
```