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:

```
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)
```

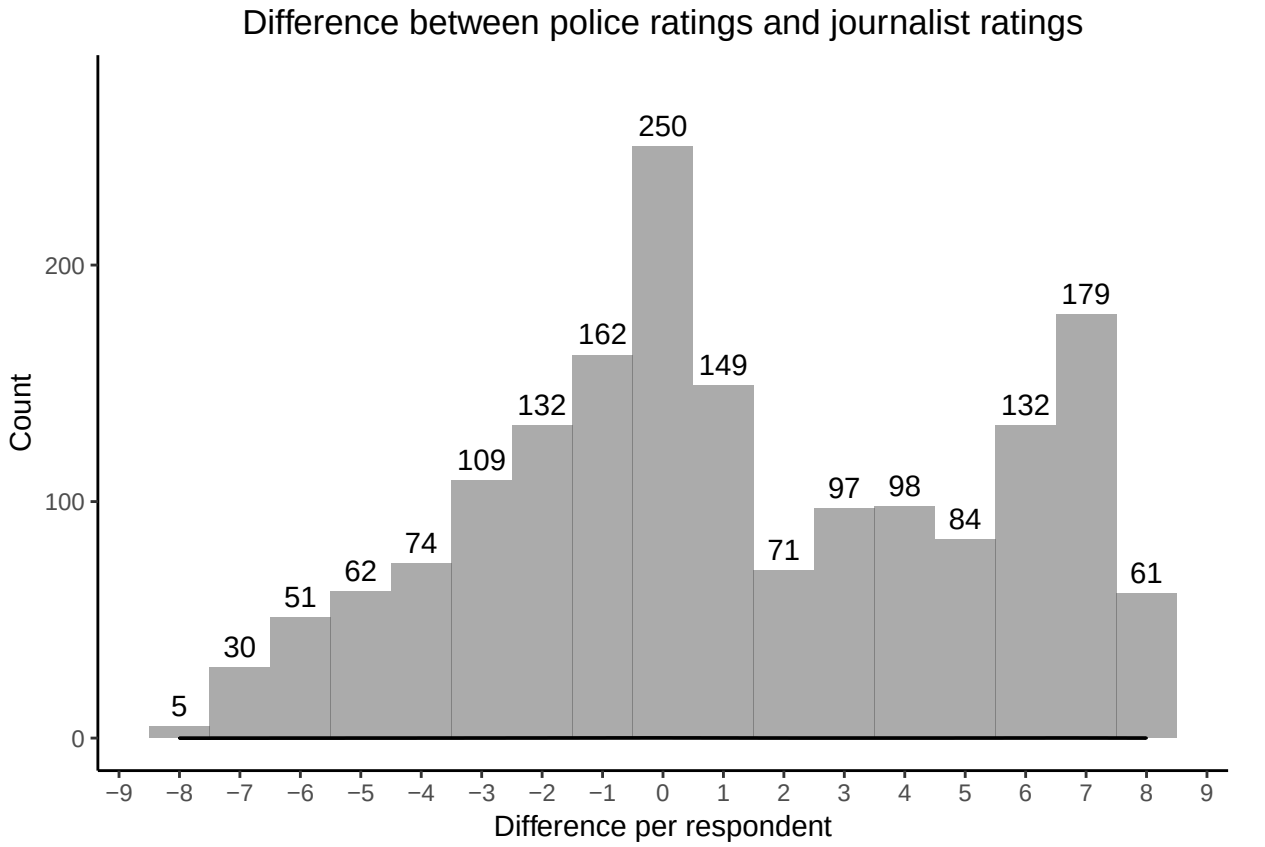
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -8.000  -2.000   0.000   1.152   5.000   8.000
```

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.

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

We also note that we will use a two-tailed test since we do not have prior evidence that the rating of one group should be greater than the other.

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 general approach discussed in Async where we compute a correlation-like value by dividing the z-score of the p-value by sqrt(n). We can do this because our sample size is very large (n=1746 to be exact).

$$r = \frac{z}{\sqrt{n}}$$

Our proportion approach produces an effect size

```
p.value <- 1.169115e-29 #from wilcox.test
z.score <- -qnorm(p.value/2)
effect.correlation <- z.score / sqrt(length(JR_PL_ordinal_data$Police_minus_Journalist))
print(effect.correlation)
```

```
## [1] 0.2706734
```

The effect correlation r is between small (0.1) and medium (0.3) effect size. While the test is highly statistically significant, the effect size is not large, and could be considered weak. As a result, there is not a large practical significance in the results.

Question 2

Introduce your topic briefly. In order to determine age, the variable birthyr (year of birth) will be subtracted from 2018, which is the year that the survey was conducted. This is the best guess for a person age in years since we do not have information on their birth month or other details. As a result of this however, we may be -1 year off from the person's real age in years. Also, because we do not have month information, a person's age in years is the best we can do.

To determine voter status and party affliation, 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 because they happen to strongly favor a particular candidate of another party, or strongly disfavor their own party's candidate. This voting behavior does not change their true party affliation, and the best guess we have as to whether someone is Democrat or Republican is what they consider themselves to be. The following lines will collect age and party information as described:

```
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 first example the ages variable. From the summary below:

```
summary(AG_PY_data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   42.50   57.00   54.18   65.50   91.00
```

we can see that the ages represent a large range, from the minimum voting age 18 up to age 91. There are no missing values in the age. The median is 57, while mean is a little over 54. We can also get a sense of the ages of Democratic versus Republican voters using summary as well:

```
AG_PY_data %>% filter(party == 'D') %>% summary
```

```
##       age           party
##  Min.   :18.00   Length:669
##  1st Qu.:39.00   Class :character
##  Median :55.00   Mode  :character
##  Mean   :52.57
##  3rd Qu.:64.00
##  Max.   :91.00
```

```
AG_PY_data %>% filter(party == 'R') %>% summary
```

```
##       age           party
##  Min.   :18.00   Length:466
##  1st Qu.:47.00   Class :character
##  Median :59.00   Mode  :character
##  Mean   :56.49
##  3rd Qu.:67.00
##  Max.   :90.00
```

We see that the min/max ages for both groups are very similar, but the median/mean seems to indicate that Republicans seem to be on average a bit over. We also want to get a sense of whether there is an equal number of Democrats and Republicans in our sample

```
AG_PY_data %>% filter(party == 'D') %>% dim
```

```
## [1] 669   2
```

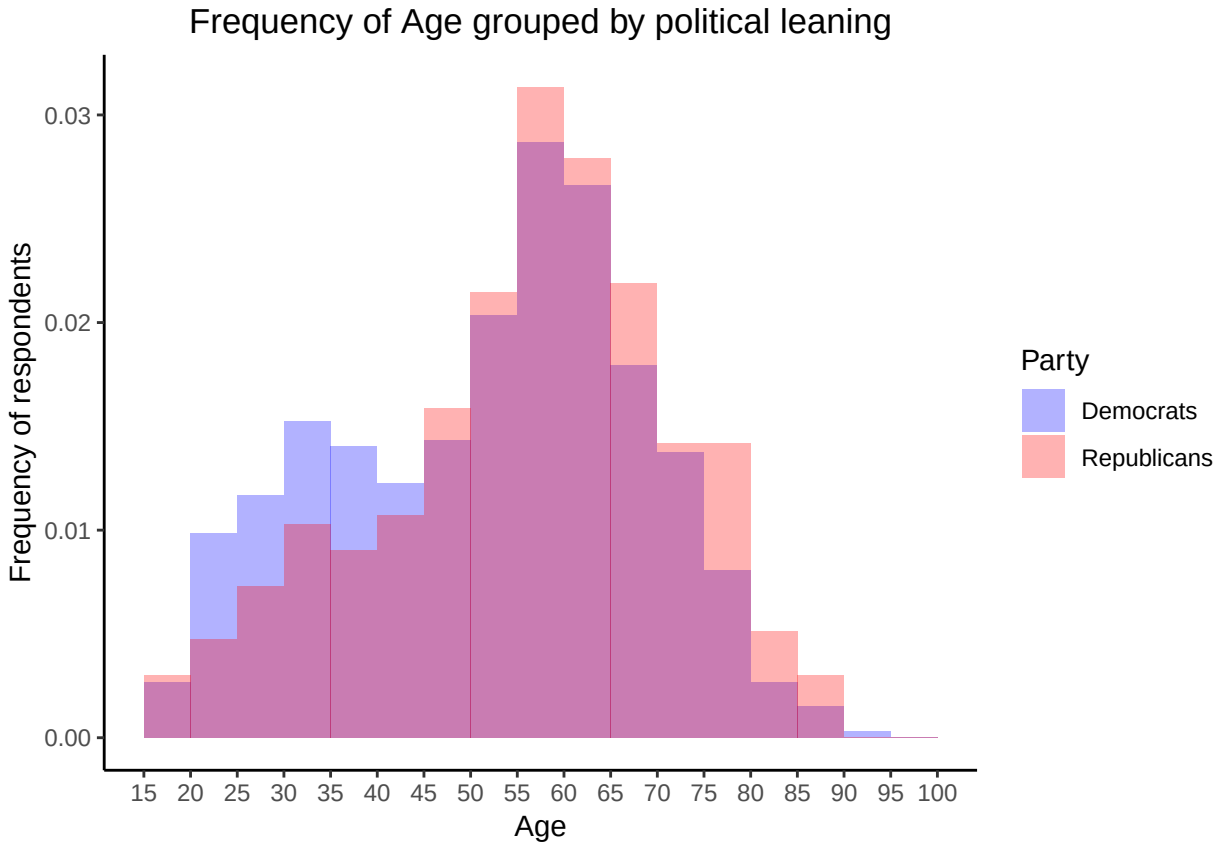
```
AG_PY_data %>% filter(party == 'R') %>% dim
```

```
## [1] 466   2
```

Our total sample size is 1135. It appears that there are 203 more Democrats than Republicans in our sample. In plotting a histogram, since there is class imbalance, it would make more sense to plot a density histogram rather than a count histogram in order to get a better sense of the distribution of age within the parties. If we were to plot only the count, we would expect Democrats to dominate many of the bins simply due to the fact that there are many more of them.

To see the distribution of age for Democrats versus Republicans, we can plot a density histogram of the ages differentiated by parties.

```
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))
```



We see that there appears to be a much larger population from 20 to 40 for Democrats than Republicans, while from ages 45 and up, there are at least as many Republicans as there are Democrats for all bins, with the exception of the very last bin, where the oldest candidate (91 years) was Democratic.

Based on your EDA, select an appropriate hypothesis test

Since we have a metric variable, we do not know the population variable, and there is no natural pairing between the participants across parties, we would like to perform an independent two sample t test.

First, we confirm that all of the assumptions are met for an independent sample t test.

1. We have ratio data (age), since we are on a metric scale, and there is an absolute zero (age = 0). We can add/subtract/take ratios of age and get sensible statistics.
2. iid observations. Like in question 1, one subject's age will not influence another's age, so there is independence. They are from the same population of individuals who responded to the survey.
3. normal population: We know that since our sample size is very large (n=1135, >>30), by the Central Limit Theorem, the sampling distribution of our mean will be very close to normal. This makes the t-test valid even if the population distribution is not exactly normal. To show that the sampling distribution of the mean is normal, we can apply a bootstrap technique. The basic approach is to first treat the sample as the population, and sample from that "population" with replacement some number of N samples, each with sample size the original size. Calculating the statistic (in this case the mean) on each of these N samples gives a rough estimation for the sampling distribution of the statistic. Bootstrapping also provides a consistent estimate, meaning as the original sample size grows, the estimated sampling distribution approaches the true sampling distribution.

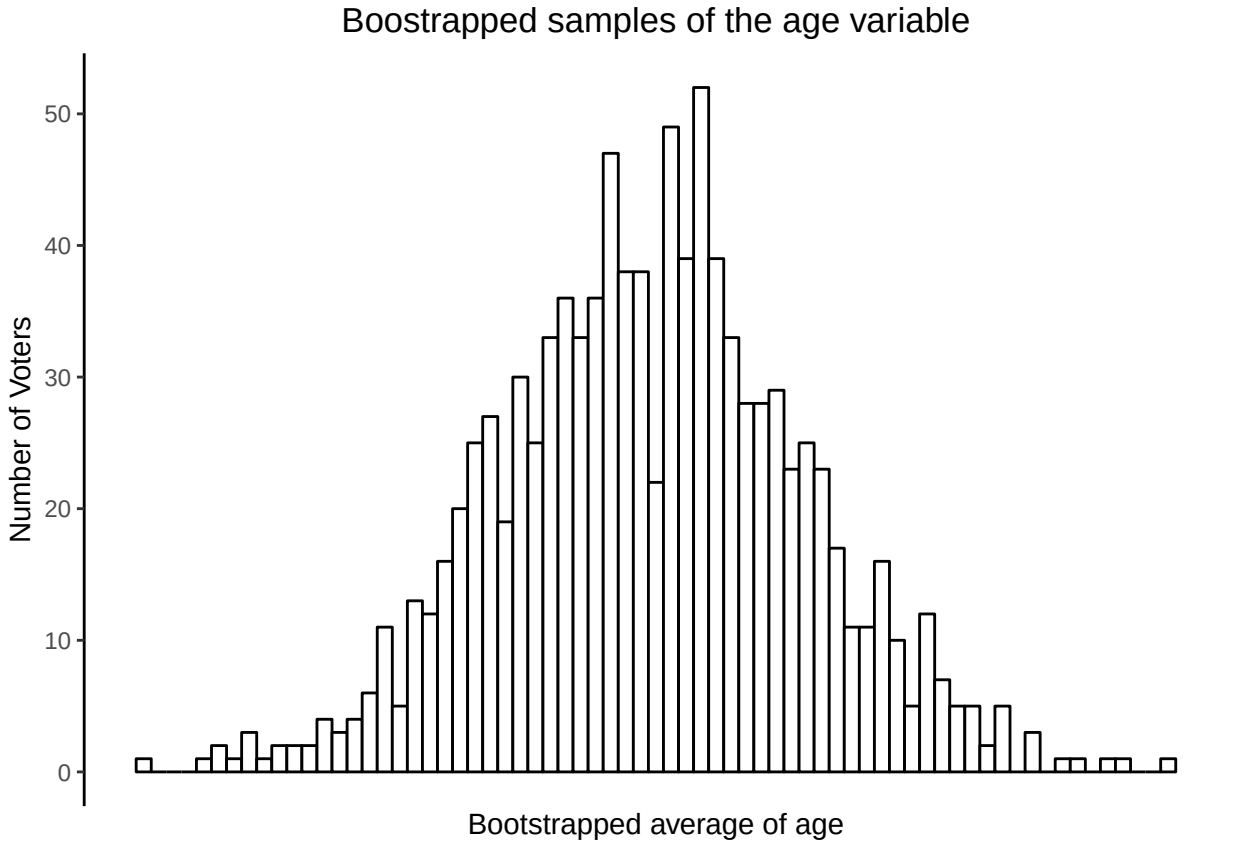
```
library(bootstrap)

#using skipped variable:
```

```r
#sample with replacement length(AG_PY_data$age) sample size
#1000 times
#report the mean of each sample

bootstrapped_samples <- bootstrap(AG_PY_data$age, 1000, function(x){mean(x)})

p <- ggplot() +
  aes(bootstrapped_samples$thetastar)+
  geom_histogram(fill="white",
                 color="black",
                 binwidth = 0.05)+
  labs(title='Boostrapped samples of the age variable',
       x='Bootstrapped average of age',
       y = 'Number of Voters') +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks=seq(0.0, 2.5, by=0.1))
p
```



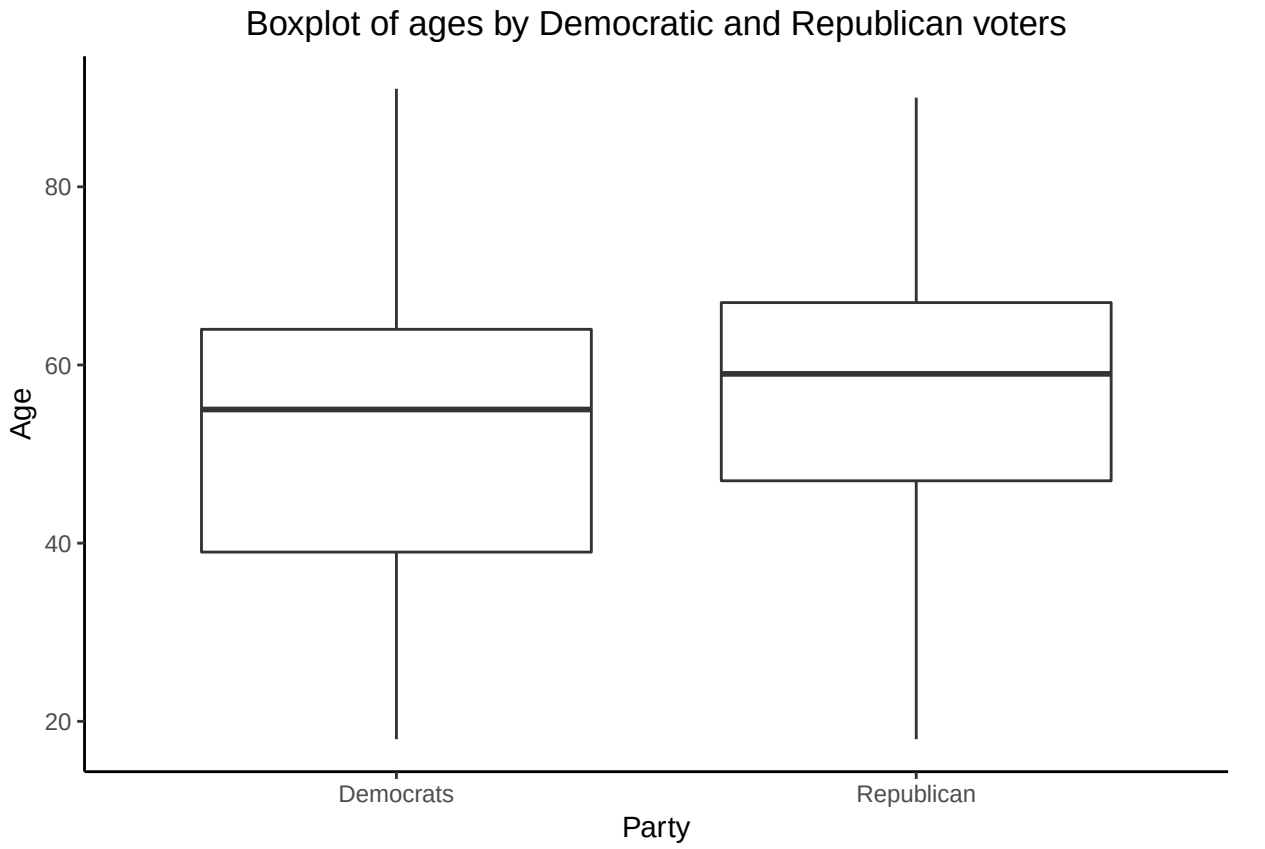

```r
shapiro.test(bootstrapped_samples$thetastar)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bootstrapped_samples$thetastar
## W = 0.99915, p-value = 0.9402
```

We see that the shapiro.test has a large p-value, meaning we have support for the null hypothesis that the sampling distribution of the mean age is normal.

4. No outliers in the data. A common way to detect outliers is using a boxplot. The interquartile range represents the value of the 3rd quartile minus the 1st quartile. Outliers can be viewed as data point lying outside of 1.5 times the interquartile range above the upper quartile or below the lower quartile. ggplot's geom_boxplot will automatically draw boxplots with outliers, which we will display as red dots. We can see from the plot generated that there are no outliers.

```
ggplot(data = AG_PY_data) +
  geom_boxplot(aes(x=party, y=age),
  position = "dodge2",
  outlier.color = 'red')+
  labs(title='Boxplot of ages by Democratic and Republican voters',
       x='Party',
       y='Age') +
  scale_x_discrete(
         labels=c("Democrats","Republican")) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```



5. Homogeneity of Variance. We need to determine whether the standard deviation of our two independent samples are equivalent or different. To do this, we'll apply Levene's test at a significance level of 0.05 (a typically accepted value). The setup as follows (letting sigma_D/sigma_R representing population standard deviations of Democrats and Republicans):

$$H_0 : \sigma_D = \sigma_R$$
$$H_a : \sigma_D \neq \sigma_R$$

We apply Levene's test for Equality Of Variances from the inferr package, and return the p-value for the test:

```
library(inferr)

age.levene <- infer_levene_test(AG_PY_data,age, group_var = party)
print(age.levene$p_lev)
```

```
## [1] 0.0586
```

Since the p-value is >0.05, we fail to reject the null hypothesis that the population of ages come from the same population. As a result, when performing our two sample t-test, we will use treat the variances between the Republican voter age population and Democratic voter age population the same.

Finally, we can setup our t test as follows using a significance level of 0.05 (using mu to replace sigma from above, and same letters denoting party):

$$H_0 : \mu_D = \mu_R$$
$$H_a : \mu_D \neq \mu_R$$

We also note that we will use a two-tailed test since we do not have prior evidence that the age of one group should be greater than the other.

To conduct the t test:

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

```
##
##  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 = -4.0837, df = 1133, p-value = 4.744e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.809705 -2.038791
## sample estimates:
## mean of x mean of y
##  52.56502  56.48927
```

Since the p-value is <<0.001 (and <0.05), we reject our null hypothesis in favor of the alterative hypothesis that the difference in age between Democratic voters and Republican voters are not 0. In fact, since the mean for Democratic voters are younger, the test also tests us that Democratic voters are on average statistically significantly younger.

To evaluate effect size, we will use the the Cohen's D statistic from the library effsize

```
library(effsize)
cohen.d(AG_PY_data[AG_PY_data$party == 'R',]$age,
        AG_PY_data[AG_PY_data$party == 'D',]$age)
```

```
##
## Cohen's d
##
## d estimate: 0.2464019 (small)
## 95 percent confidence interval:
##     lower     upper
## 0.1275807 0.3652230
```

Since our Cohen's D value is 0.246, it is between what is typically considered small (0.2) and medium (0.5). As a result, we have a reasonable small effect size. Even though the difference in age was significant due to the large sample size, the practical significance is small, meaning there is a significant overlap in age between Democratic and Republican voters.

Question 3

Introduce your topic briefly.

As in question 2, 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. These responses represent strongly disapprove to mildly disapprove of the investigation, while the rest of the variables represents no opinion, or some degree of approval. Since this question asks whether the respondent approves or disapproves, this may not fully capture whether the respondent thought the investigation was baseless. For example, these individuals who disapproved could have found that the investigation was necessary, but that the execution was poor and therefore disapprove this. Some degree of disapproval is necessary for considering the investigation baseless, but not necessarily sufficient. However, this is the closest we could get. Below, we filter the respondents based on their party, and select their

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

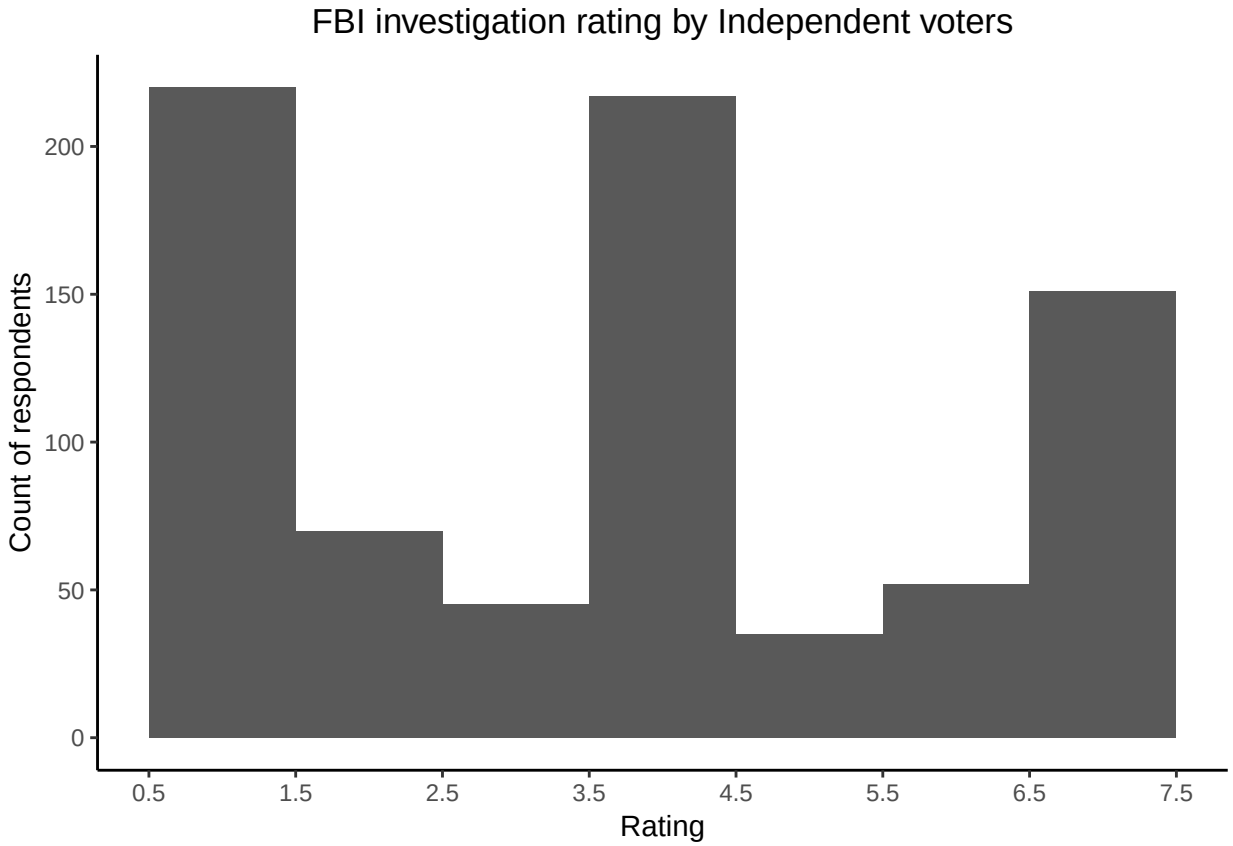
First, we generate a summary of the muellerinv variable for Independent voters.

```
summary(FBI_IN_data$muellerinv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    4.00    3.68    6.00    7.00
```

We can see that the minimum value is 1 and maximum is 7, meaning a response for everyone in this sample was recorded (since -7 means no response). The median is at 4, which would also be the mean if the distribution were normal. However, the mean is slightly smaller at 3.68. We can generate a histogram of this variable.

```
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))
```



We can see that the distribution peaks 3 times, at strongly disapprove (7), strongly approve (1), and "Neither approve nor disapprove," (4). There is a smaller population that strongly disapprove than strong approve. Since we classify a neural response (4) as not baseless, we see that there are two big peaks for not baseless, and one large peak for thinking the investigation was baseless.

For this question, since we care about the proportion of individuals who believe the investigation was baseless, we will use a binomial test for population proportion. This test compares the number of successes (in this case we call that the number of people who believe the investigation is baseless) to the hypothesized number of successes for a sample of particular size. The assumptions for this test includes:

1. Samples are dichotomous and nominal. This is certainly the case here: the subjects are classified as either believing the investigation was baseless or not, which does not have a numerical value associated with them.
2. iid samples. This is the case since each subject's decision in the survey does not influence other subjects. In terms of identical distribution, this can be assumed since all subjects are drawn from the population of survey takers.

Our hypothesis will be setup as follows, with the variable "p" representing the proportion of the population. Our null hypothesis is that the same proportion of individuals believe the investigation was baseless as those who did not. We perform a 2-tailed test because we do not have a strong prior that the population is strongly leaning in either direction. We also set our confidence level at 0.05, a standard level.

$$H_0 : p = 0.5$$
$$H_a : p \neq 0.5$$

To conduct the test, we can use the binom.test function in R. This takes a parameter for number of successes (number of people who believe the investigation was baseless), number of trials (total number of people in the sample), and 't' for two-sided test:

```r
binom.test(sum(FBI_IN_data$muellerinv >=5),
           length(FBI_IN_data$muellerinv),
           p=0.5,
           alternative = 't')
```

```
##
##  Exact binomial test
##
## data:  sum(FBI_IN_data$muellerinv >= 5) and length(FBI_IN_data$muellerinv)
## number of successes = 238, number of trials = 790, p-value <
## 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2694311 0.3345918
## sample estimates:
## probability of success
##              0.3012658
```

We see that since the p-value is much less than 0.05, we can reject the null hypothesis and say that the proportion of individuals think the investigation was baseless versus those who think otherwise. To answer the original question, since the median was 4 (representing neutral decision), we can say that the majority of Independent voters believed that the investigation was in fact not baseless.

In order to calculate the practical significance, we will use the metric of Cohen's g, which is valid for the one-sample binomial test where the null hypothesis states that the proportion is 0.5. This is exactly our case here. Cohen's g is simply the difference of sample proportion of individuals who believed the investigation was baseless with 0.5, and then taking the absolute value. The closer the sample proportion is to 0.5, the smaller the effect size.

```r
cohen.g <- abs(sum(FBI_IN_data$muellerinv >=5) / length(FBI_IN_data$muellerinv) - 0.5)
print(cohen.g)
```

```
## [1] 0.1987342
```

We get an effect size close 0.2, which is within a medium effect size. Typically, the ranges are <0.15 is small, 0.15-0.25 is medium, and >=0.25 is large. We can conclude that our study is highly statistically significant with medium level of practical significance.