Lab 1: Comparing Means

w203: Statistics for Data Science

Alice Hua - Haley Farber - John Schulz

The Data

The American National Election Studies (ANES) conducts surveys of voters in the United States.
Details about the design of this study: ANES User's Guide and Codebook.
Precise form of each question: Questionnaire Specifications.

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. We are using this weight for the assignment.

Assumptions about the data:

1. The data is independent and identically distributed.
   The respondents's answers are independent of each other. In this case, a respondent's answer on one question does not affect how the another respondent answered theirs. In regard to data being identically distributed, it is not possible to have perfect independece because once a respondent is drawn, they cannot be redrawn and this changes the distribution for the next draw. However, the change is very small, we can safely assume that the data is identically distributed.

2. The sample size is 2500 which is large enough to rely on the CLT for normality and account for small amount of skewness.

```r
A = read.csv("anes_pilot_2018.csv")
```

Research Questions

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

Introduce your topic briefly. (5 points)

We operationalize respect as having higher rating. The variables ftpolice and ftjournal recorded the rating for police and journalists on a scale from 0 to 100. We characterize higher rating to more respect. We will consider rating of each individual respondent, if their police rating is higher than their rating for journalist, then the respondent has more respect for the police, if the ratings are or otherwise, then the respondent has more respect for journalist and if the ratings are equal, then we did not include them as it does not help us answer if the US voters have more respect or less respect for either group.

```r
Q1 <- select(A, 'ftpolice','ftjournal') %>%
      mutate(respect = case_when(
        ftpolice > ftjournal ~ 'More respect for Police',
        ftpolice < ftjournal ~ 'More respect for Journalists'))
head(Q1)
```

```
##   ftpolice ftjournal                      respect
## 1      100        50      More respect for Police
## 2       48        85 More respect for Journalists
## 3       45        50 More respect for Journalists
## 4       81        53      More respect for Police
## 5       73        77 More respect for Journalists
## 6       38        23      More respect for Police
```

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

We are expecting 0 to 100 but there might be -1 for skip, -4 for error due to technical problem or -7 for no answer. We will need to remove no answers since we are taking the difference between the ratings. For example, including -7 as a rating would be incorrect.

```
Q1 <- filter(Q1, Q1$ftpolice >= 0  & Q1$ftjournal >= 0)
nrow(Q1)
```

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

```
table(Q1$respect)
```

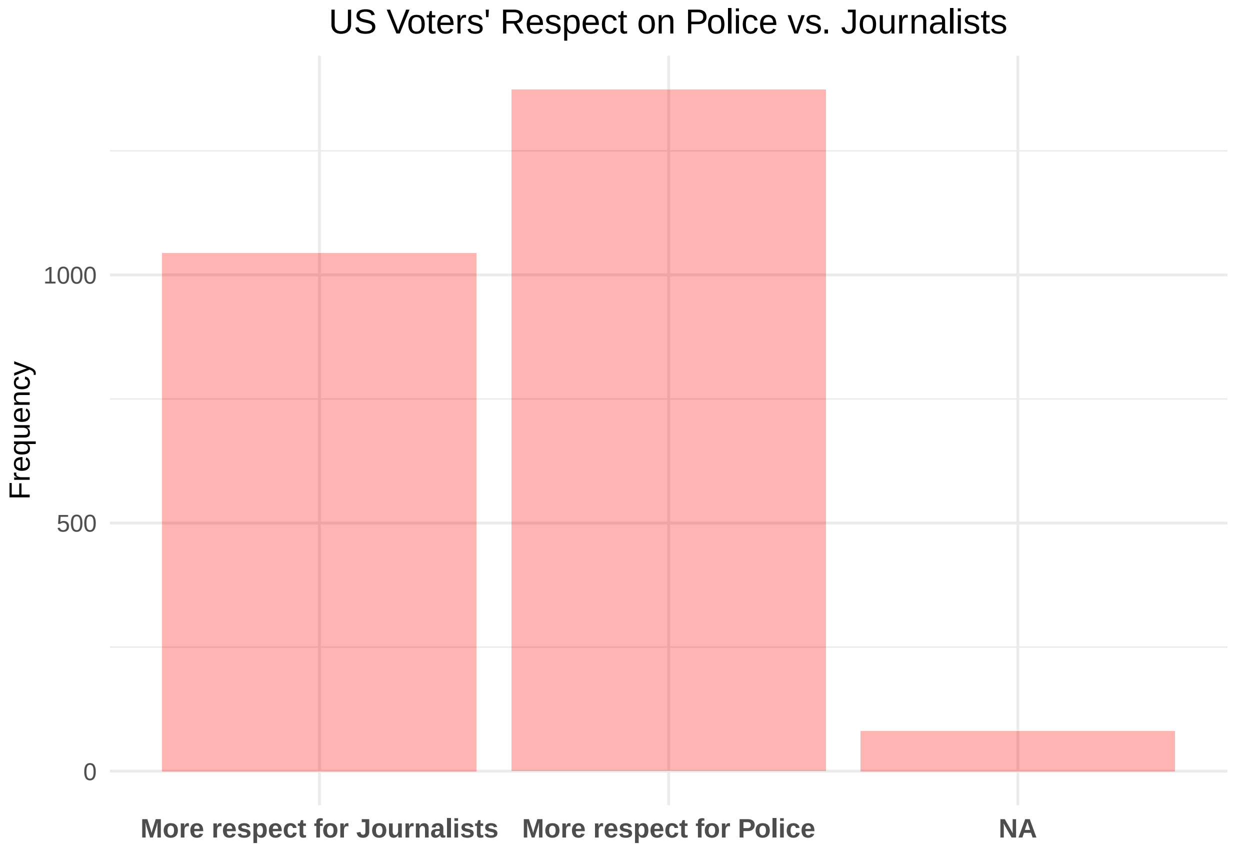
```
##
## More respect for Journalists      More respect for Police
##                         1044                         1373
```

```
summary(Q1)
```

```
##     ftpolice        ftjournal        respect
##  Min.   :  0.00   Min.   :  0.00   Length:2498
##  1st Qu.: 47.00   1st Qu.: 21.00   Class :character
##  Median : 70.00   Median : 52.00   Mode  :character
##  Mean   : 64.67   Mean   : 52.31
##  3rd Qu.: 90.00   3rd Qu.: 82.00
##  Max.   :100.00   Max.   :100.00
```

We can see more voters have higher respect for the police than journalists. If we look at the summary statistics, it seems that the respondents rated higher for the police than the journalists based on the medians but the mean ratings show a closer gap. However, this variable cannot be considered as metric because much like a Likert scale, the intervals are not consistent. For example, it is not possible to say that one's difference between agree and strongly agree is the same as one's difference between neutral and agree.

```
Q1 %>% ggplot(aes(x=respect)) +
       geom_bar(fill='red', alpha=.3) +
       labs(title = "US Voters' Respect on Police vs. Journalists", y = 'Frequency') +
       theme_minimal() +
       theme(axis.title.x = element_blank(),
             axis.text.x = element_text(size=10,face="bold"),
             plot.title = element_text(hjust = 0.5))
```



The bar chart shows a count of US voters who have more respect for either the journalists or the police, there is a small subset of the population where the respect is equal for both group. As we have mentioned earlier, we will only look at those with more respect for either group to answer the question.

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

We choose the sign test (which uses a binomial test) with the following assumptions:

1. The rating variables are ordinals in term of established order from 0 for *Very cold or unfavorable feeling* to 100 for *Very warm or favorable feeling*. We transformed the variables to dichotomous and nominal (1 - more respect for police, 0 - more respect for journalists).
2. There there is pairing for the data since each unit of observation is one respondent where there are two measurements (ratings for police and journalist).

3. The data is iid as mentioned in the **Assumptions about the data**.

- The null hypothesis: US voters have the same respect for the police and journalists. The probability of two possible outcomes (more respect for police, more respect for journalists) is 50%.

- The alternative hypothesis: The probability of two possible outcomes (voters respect police more than journalists and vice versa) is not more than 50%.

Conduct your test. (5 points)

```r
more_police <- sum(Q1$ftpolice > Q1$ftjournal)
less_police <- sum(Q1$ftpolice < Q1$ftjournal)
total <- sum(Q1$ftpolice > Q1$ftjournal | Q1$ftpolice <= Q1$ftjournal)
binom.test(more_police, total, p = 0.5)
```

```
##
##  Exact binomial test
##
## data:  more_police and total
## number of successes = 1373, number of trials = 2498, p-value =
## 7.581e-07
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5298803 0.5692826
## sample estimates:
## probability of success
##              0.5496397
```

```
cor <- (more_police - less_police)/total
cat("The correlation is:", cor)
```

```
## The correlation is: 0.1317054
```

Our test is statistically significant because the p-value < 0.05. We can reject the null hypothesis that US voters have the same respect for the police and journalists.
This means that the data support the alternative hypothesis which is that the proportion of US voters who respect police more than journalists is not equal to 50%.
To express practical significance, we use the common language effect size 1373/2498 = 0.55 which is the proportion of voters who have higher respect for police than journalists. Together with the correlation calculation of .13, the effect size is small and thus we think that the difference in population who respect police more than journalists is not practically significant.

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

Introduce your topic briefly. (5 points)

To answer this question, we used the variables birthyr and pid1d. Birthyr is the year of birth of the respondent and pid1d contains a number between 1-4 that represents what political party the respondent identifies as. The numbers represent the following: 1. Democrat 2. Republican 3. Independent 4. Something Else.

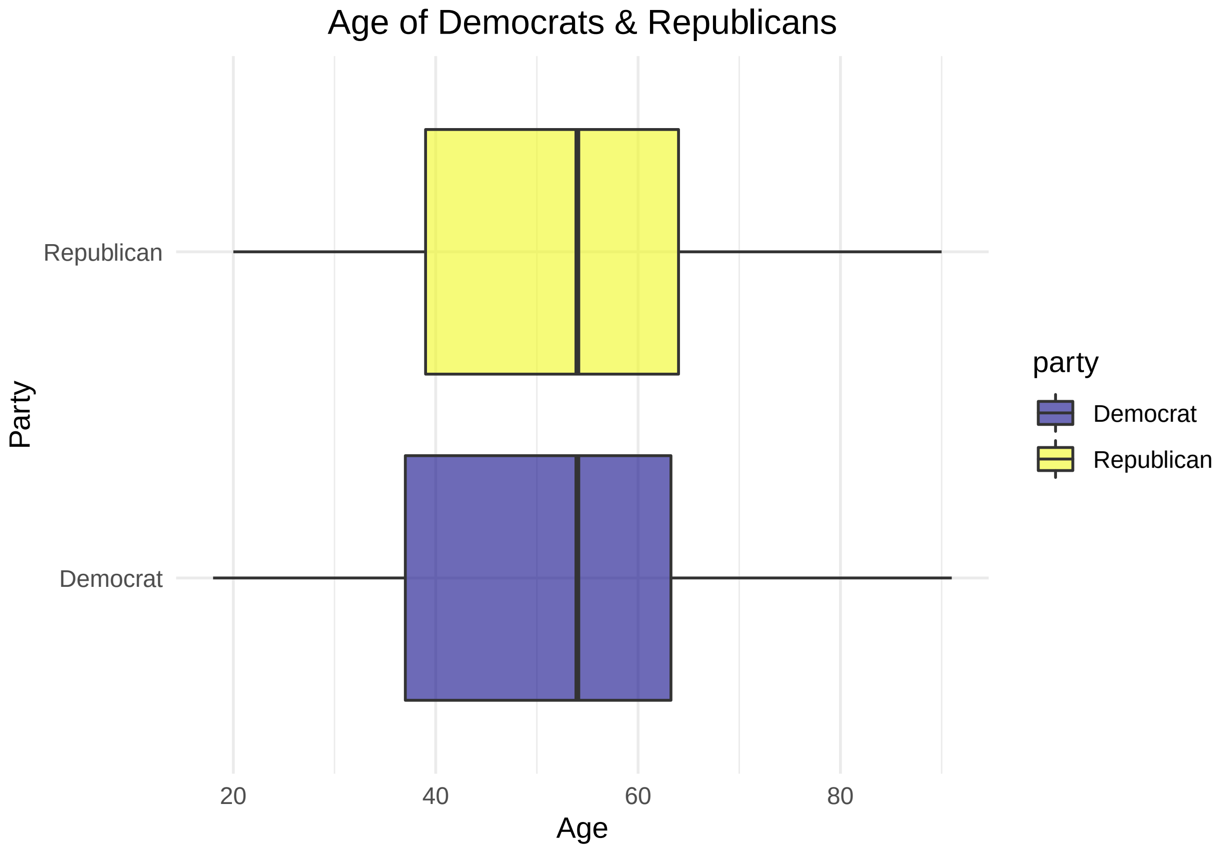
We subtracted the birth year of the respondent to determine the respondents' ages and used these ages as an indicator of how old voters are. The pid1d directly explains the political identity of the respondent so we will utilize the respondents who answered that they are Republican or Democrat to answer the question above.

```
Q2 <- select(A, birthyr, pid1d) %>%
      mutate(age = 2018 - birthyr) %>%
      mutate(party = case_when(
        pid1d == 1 ~ "Democrat",
        pid1d == 2 ~ "Republican"))
head(Q2)
```

```
##   birthyr pid1d age      party
## 1    1986     2  32 Republican
## 2    1972    -1  46       <NA>
## 3    1999    -1  19       <NA>
## 4    1975     3  43       <NA>
## 5    1989    -1  29       <NA>
## 6    1992    -1  26       <NA>
```

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

```
Q2 %>% filter(party == 'Democrat' | party == 'Republican') %>%
      ggplot(aes(y=party, x=age, fill=party)) +
      geom_boxplot() +
      scale_fill_viridis(discrete = TRUE, alpha=0.6, option="C") +
      theme_minimal() +
      labs(title = "Age of Democrats & Republicans", x='Age', y='Party') +
      theme(plot.title = element_text(hjust = 0.5))
```



```
summary(Q2$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   35.00   52.00   49.48   62.00   91.00
```

It appears the youngest respondent who identifies as Democrat or Republican is 18 years old and the oldest is 91 years old.

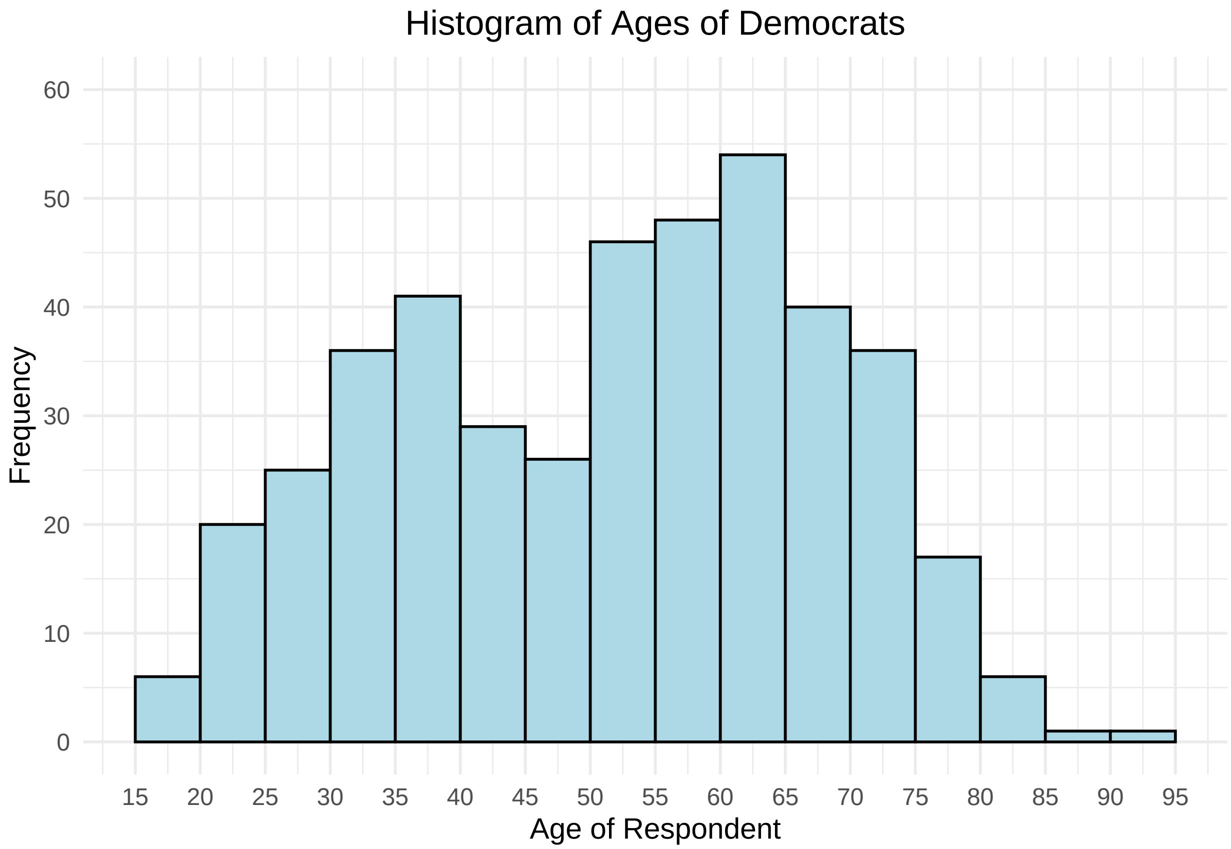
```
table(Q2$party)
```

```
##
##   Democrat Republican
##        432        326
```

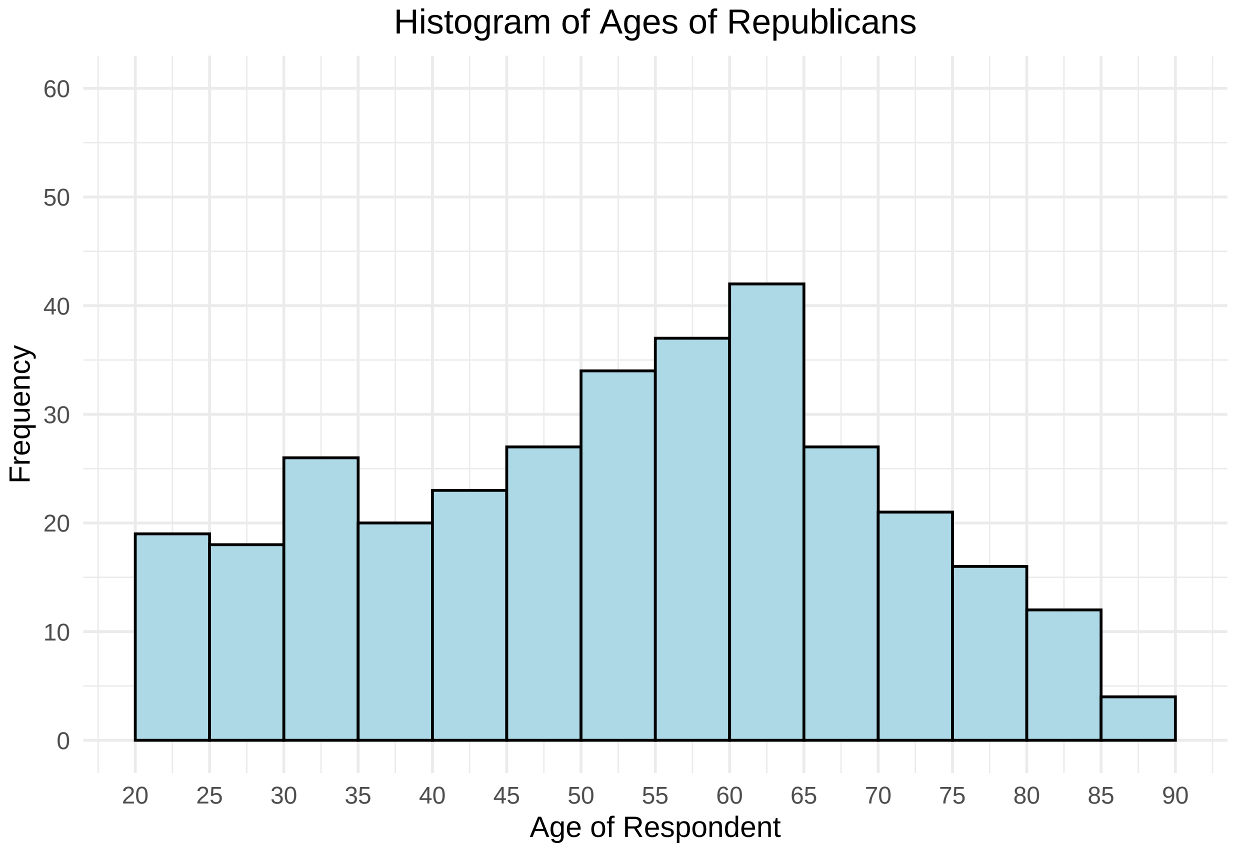
There are 432 respondents who identify as Democrats which is a much larger number than 30, which is the rule of thumb for the CLT. Thus, it appears that the CLT would apply for this variable.
There are 326 respondents who identify as Republicans which is also a much larger number than 30 so the CLT would also apply in this situation. To be safe, let's look at the distribution of ages to see if there's a

large deviation from normality.

```
Q2 %>% filter(party == 'Democrat') %>%
       ggplot() +
       aes(x=age)+
       geom_histogram(binwidth = 5, boundary = 0, closed = "left",
                      color = 'black', fill = 'lightblue') +
       labs(title ='Histogram of Ages of Democrats',
            x ='Age of Respondent', y="Frequency") +
       scale_x_continuous(breaks =  seq(15,100,5))+
       scale_y_continuous(limits = c(0,60),breaks =  seq(0,60,10))+
       theme_minimal() +
       theme(plot.title = element_text(hjust = 0.5))
```



```
Q2 %>% filter(party == 'Republican') %>%
       ggplot() +
       aes(x=age)+
       geom_histogram(binwidth = 5, boundary = 0, closed = "left",
                      color = 'black', fill = 'lightblue') +
       labs(title ='Histogram of Ages of Republicans',
            x ='Age of Respondent', y="Frequency") +
       scale_x_continuous(breaks =  seq(15,100,5))+
       scale_y_continuous(limits = c(0,60),breaks =  seq(0,60,10))+
       theme_minimal() +
       theme(plot.title = element_text(hjust = 0.5))
```



There is very little skew in these distributions. With 432 and 326 observations for each group respectively, the skew is not enough to violate the validity of a t-test, they look relatively normal.

```
democrats_age <- Q2[Q2$pid1d == 1,]$age
republicans_age <- Q2[Q2$pid1d == 2,]$age

summary(democrats_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   37.00   54.00   51.02   63.25   91.00
```

```
summary(republicans_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.0    39.0    54.0    52.5    64.0    90.0
```

The summary statistics for both groups match each other very closely leading us to believe there will not be a significant difference between the ages of Democrats and Republicans.

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

We will use an unpaired t-test. There are two distinct groups, Republicans and Democrats, and we are measuring the difference between age for both groups. We chose the t-test because the data is metric, iid, and not too unnormal. Below is a further explanation.

The required t-test assumptions are:
1. metric scale - age of respondents is coded numeric, and there is a consistent meaning to a difference of 1 year.
2. The data is iid as mentioned in the **Assumptions about the data**.

3. not too unnormal considering the sample size - there is very little skew and the sample size is large enough to rely on the CLT.

- Null hypothesis: There is no difference in age between Republican and Democrat voters. The true difference in mean age betweenn the two groups is zero.

- Alternative hypothesis: There is a difference in age between Republican and Democrat voters. The true difference in mean age between the two groups is not zero.

Conduct your test. (5 points)

```
t.test(Q2$age ~ Q2$party)
```

```
##
##  Welch Two Sample t-test
##
## data:  Q2$age by Q2$party
## t = -1.2146, df = 691.37, p-value = 0.2249
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.8702977  0.9119644
## sample estimates:
##   mean in group Democrat mean in group Republican
##                 51.02083                 52.50000
```

The p-value is 0.2249 which means we fail to reject the null hypothesis that there is no difference in age between Repubicans and Democrats.

```
mean_dems <- mean(democrats_age)
mean_reps <- mean(republicans_age)
mean_reps - mean_dems
```

```
## [1] 1.479167
```

For a measure of practical significance, the simple difference in means, 52.5 - 51.02083 = 1.479167 is probably the best measure. Republicans on average are 1.48 years older than Democrats. A year and a half is not much compared to the average human life span so it agrees with the results of the t-test that there is no difference in the mean age between Republicans and Democrats.

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)

To answer this question, we used the variables russia16 and pid1d. The russia16 variable contains the respondents’ answers to the question, “Do you think the Russian government probably interfered in the 2016 presidential election to try to help Donald Trump win, or do you think this probably did not happen?”.
We operationalized the variable russia16 as:

Believe that the federal investigations of Russian election interference are not baseless:
1 = Russia probably interfered = not baseless

Believe that the federal investigations of Russian election interference are baseless:
2 = This probably did not happen = baseless

We used the varibale pid1d, which was mentioned in question 2, to determine which respondents are independents. Pid1d = 3 means the respondent is an independent.

```
Q3 <- select(A, russia16, pid1d) %>%
      filter(pid1d == 3) %>%
      mutate(base = case_when(
        russia16 == 1 ~ 'Not Baseless',
        russia16 == 2 ~ 'Baseless'
      ))
head(Q3)
```

```
##   russia16 pid1d         base
## 1        2     3     Baseless
## 2        2     3     Baseless
## 3        1     3 Not Baseless
## 4        1     3 Not Baseless
## 5        1     3 Not Baseless
## 6        1     3 Not Baseless
```

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

Summarize russia16 for independents:

```
summary(Q3$russia16)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   1.000   1.000   1.463   2.000   2.000 
```

It looks like the mean 1.463 which is less than 1.5 where 1 represents those who believe the investigations are not baseless and 2 otherwise. Thus, it appears that more independent voters believe that the investigations are not baseless.

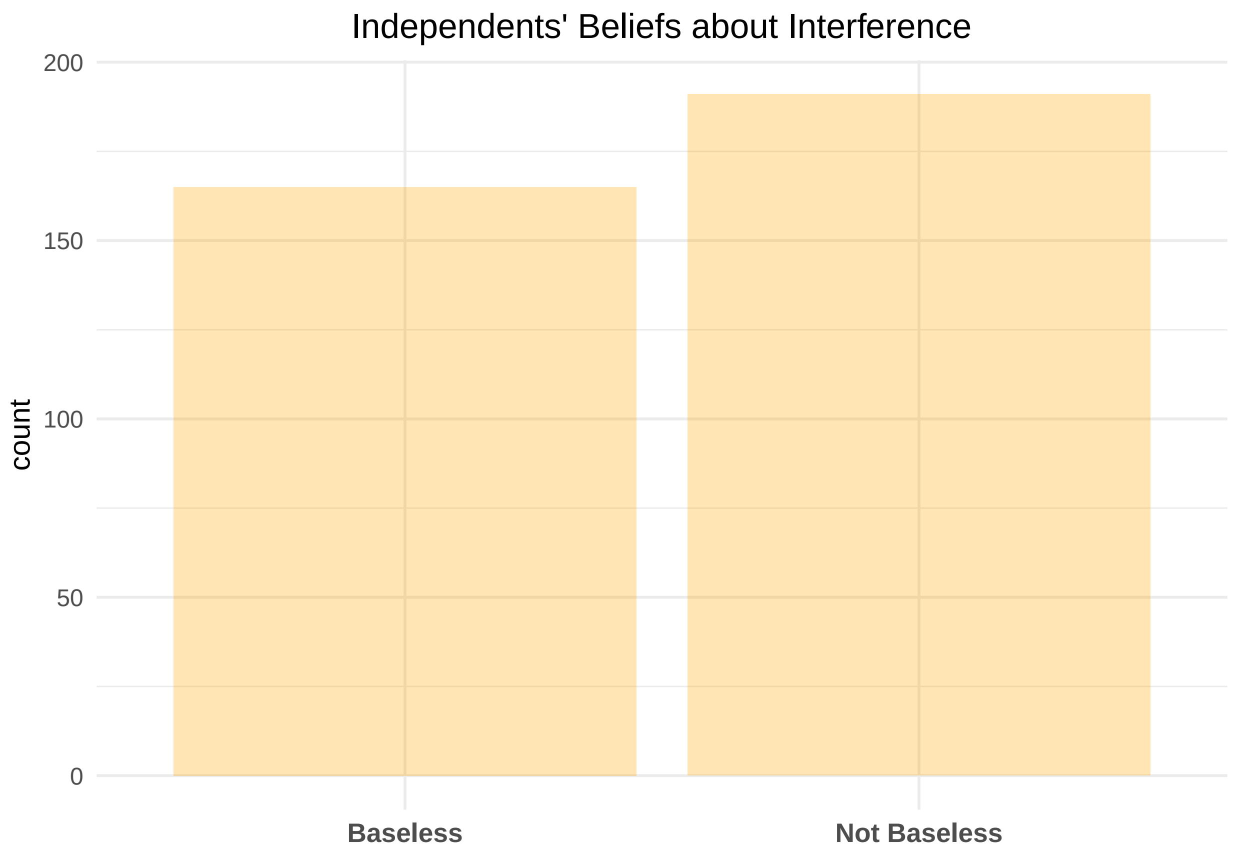
```
#Define numbers of baseless, not baseless, and total independent voters
not_baseless = sum(Q3$russia16 == 1)
baseless = sum(Q3$russia16 == 2)
total = nrow(Q3)

#print these numbers
cat("Number of Not Basless Respondents:", not_baseless, "\n",
    "Number of Baseless Respondents:",baseless, "\n",
    "Total Number of Independent Respondents:",total)
```

```
## Number of Not Basless Respondents: 191 
##  Number of Baseless Respondents: 165 
##  Total Number of Independent Respondents: 356
```

The total number of independent voters is 356. We have 191 independent voters who believe the investigations are not baseless and 165 who believe that the investigations are baseless.

```
Q3 %>% ggplot(aes(x=base)) +
       geom_bar(fill = 'orange', alpha=.3) +
       labs(title = "Independents' Beliefs about Interference") +
       theme_minimal() +
       theme(axis.title.x = element_blank(),
             axis.text.x = element_text(size=10,face="bold"),
             plot.title = element_text(hjust = 0.5))
```



```
#Percent Not Basesless and Percent Baseless
p_not_baseless = not_baseless/total
p_baseless = baseless/total

#print values
cat("Percent Not Basless:",p_not_baseless, "\n",
    "Percent Baseless:", p_baseless, "\n")
```

```
## Percent Not Basless: 0.5365169
##  Percent Baseless: 0.4634831
```

The percentage of the independent voters who believed the claims were not basless is higher than the percent of independent voters who believed the claims were not baseless.

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

We chose to run an unpaired t-test because the data is unpaired and metric. Below is a further explanation.

The required t-test assumptions are:
1. metric var - the data we use is metric in this instance. We assign those who believe the investigations are not baseless a "1" and those who believe the investigations are baseless a "2". We treat these categorical variables as metric because we are interested in the percent of those who fall into either category. In this case, russia16 is our outcome variable, and it's binary. A binary varible can be treated as metric. Our rule is that a metric variable is one in which intervals of the same size are equivalent. In this case of a binary variable, there is only one interval - the one from 1 to 2.
2. The data is iid as mentioned in the **Assumptions about the data**.
3. Not too un-normal of a distribution. It is bimodal, but that is not a major problem for the CLT. In addition, there is a large sample size of 191 independents who believe the investigations are not baseless and

165 respondents who believe the investigations are baseless.

- Null hypothesis: Half of independents believe that the investigations are baseless and half of the independents believe that the investigations are not baseless.
- Alternative hypothesis: Either a majority of independents believe that the investigations are not baseless or a majority of independents believe that the investigations are baseless.

Conduct your test. (5 points)

```
t.test(Q3$russia16, mu = 1.5)
```

```
##
##  One Sample t-test
##
## data:  Q3$russia16
## t = -1.3797, df = 355, p-value = 0.1685
## alternative hypothesis: true mean is not equal to 1.5
## 95 percent confidence interval:
##  1.411433 1.515534
## sample estimates:
## mean of x
##  1.463483
```

We fail to reject the null hypothesis that half of independents believe that the investigations are baseless and half of the independents believe that the investigations are not baseless. The p-value is 0.1685 which is much higher than our 0.05 cutoff for statistical signficance. Thus, we can not conclude that most independents believe the investigations are not baseless.

```
p_not_baseless - p_baseless
```

```
## [1] 0.07303371
```

The practical significance can be represented by a difference in the percent of the independents who believe the allegations are not baseless,about 54% and those who believe the investigations are baseless, about 47%. There is about 7% more independent respondents believe the allegations are not baseless than 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)

We operationalized the variables turnout16, turnout18, geafraid, and geangry as voter turnout for 2016, voter turnout for 2018, fear and anger.
For each respondent, we looked at how angry and afraid they were in 2016 and 2018. The variables geangry and geafraid have ordinal ratings where 1 signifies less anger/afraid and 5 signifies the most anger/afraid.

The responses are as follows for turnout16 and turnout18:

turnout16
1 = Definitely voted
2 = Definitely did not vote
3 = Not completely sure

turnout18
1 = Definitely vote in person on Nov 6th
2 = Definitely voted in person before Nov 6th
3 = Deinitely voted by mail

4 = Definitely did not vote
5 = Not completely sure

We categorized those who 1 as 'voted' and 2 as 'did not vote' for turnout16. For turnout18, we categorized those with responses from 1 to 3 as 'voted' and 4 as 'did not vote'.

Those who were unsure if they voted are excluded as we cannot categorize them in either group with certainty. With these two voter turnout for 2016 and 2018, we filtered our target group to be composed of voters who didn't vote in 2016 but did vote in 2018. This target population will help us identify whether anger or fear increased voters' turnout in 2018.

We also categorized people who feel more afraid than angry as 'More Afraid', those who feel more angry than afraid as 'More Angry', those with equal feeling of angry and afraid are not included as they does not help us answer the question of whether or not anger or fear was more effective.

```
Q4 <- select(A, 'turnout18','turnout16', 'geafraid', 'geangry') %>%
      mutate(vote_18 = case_when(
        turnout18 <= 3 ~ 'voted',
        turnout18 == 4 ~ 'did not vote')) %>%
      mutate(vote_16 = case_when(
        turnout16 == 1 ~ 'voted',
        turnout16 == 2 ~ 'did not vote'))

Q4_f <- Q4 %>%
  filter(vote_18 == 'voted' & vote_16 == 'did not vote') %>%
  mutate(feeling = case_when(
    geafraid > geangry ~ 'More Afraid',
    geafraid < geangry ~ 'More Angry'
  ))

head(Q4_f)
```

```
##   turnout18 turnout16 geafraid geangry vote_18      vote_16     feeling
## 1         1         2        2       2   voted did not vote        <NA>
## 2         3         2        4       4   voted did not vote        <NA>
## 3         1         2        3       1   voted did not vote More Afraid
## 4         3         2        4       3   voted did not vote More Afraid
## 5         1         2        5       3   voted did not vote More Afraid
## 6         1         2        3       3   voted did not vote        <NA>
```

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

This next section previews information describing voter participation differences from 2016 to 2018 and how afraid or angry survey respondents were.
There are some non-responses to these questions which are indicated by a -7. We removed these values from our target data as they are not useful in the analysis.

```
Q4_f <- Q4_f %>%
  filter(geafraid != -7 & geangry != -7)
```

```
nrow(Q4)
```

```
## [1] 2500
```

```
nrow(Q4_f)
```

```
## [1] 96
```

We can see that there were 2500 respondents total but only 96 who responded to 'did not vote' in 2016 and 'voted' in 2018. This is our target population.

```
# Counts of many many people who  angry
table(Q4_f$geafraid)
```

```
##
##  1  2  3  4  5
## 17 21 31 10 17
```

```
table(Q4_f$geangry)
```

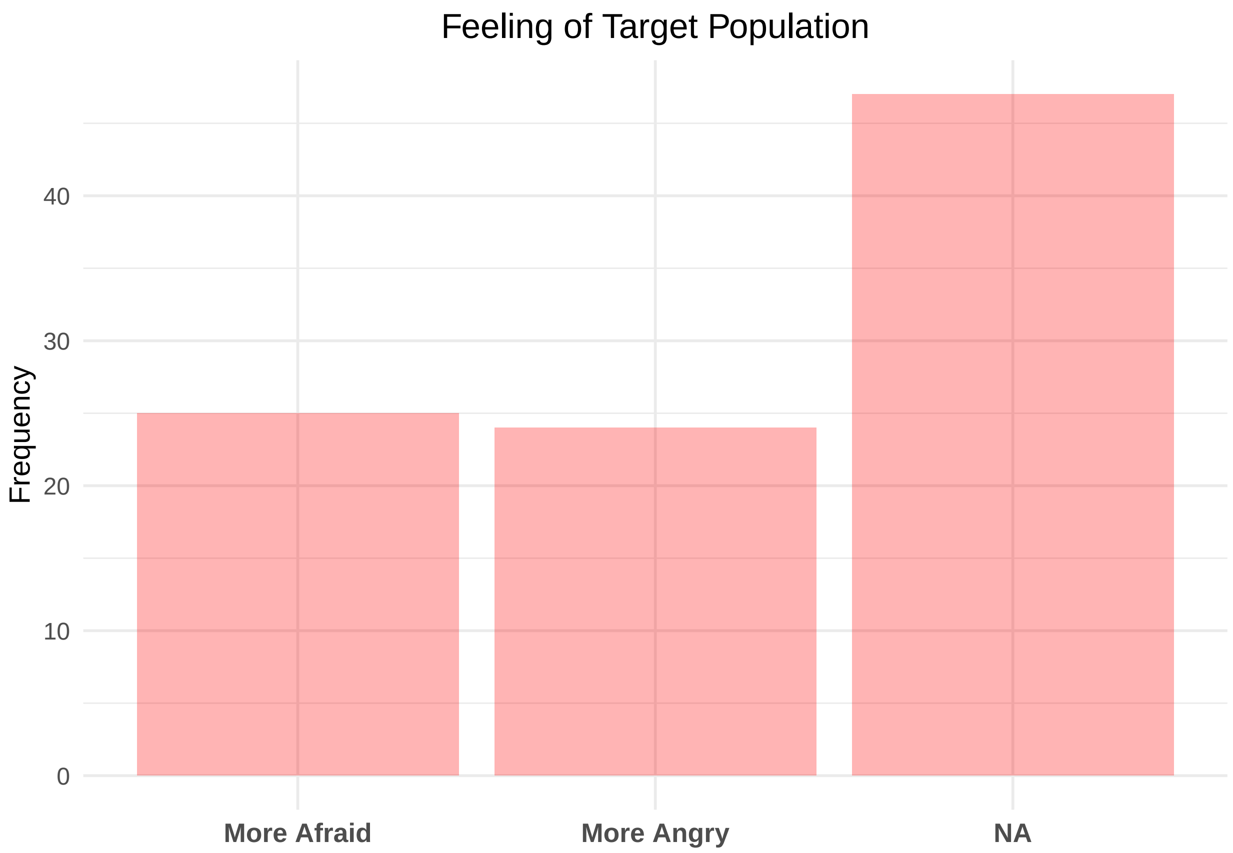
```
##
##  1  2  3  4  5
## 17 22 28 12 17
```

```
table(Q4_f$feeling)
```

```
##
## More Afraid  More Angry
##          25          24
```

We can see that prior to categorizing each respondent as more angry or more afraid, the original counts for each variable are similar with most people answered 3 on a scale of 1 to 5, 5 being the most fear or anger. Our grouped counts for more angry or more afraid voters are also almost identical.

```
Q4_f %>% ggplot(aes(x=feeling)) +
       geom_bar(fill='red', alpha=.3) +
       labs(title ="Feeling of Target Population", y="Frequency") +
       theme_minimal() +
       theme(axis.title.x = element_blank(),
       axis.text.x = element_text(size=10,face="bold"),
       plot.title = element_text(hjust = 0.5))
```



The bar chart shows what we have confirmed above, these 2 groups have almost the same counts. Notice a lot more people have the same feeling for fear and anger.

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

We chose a sign test with the following assumptions met:
1. The rating for both angry and afraid are ordinal and in decreasing order which meets the first assumption.
2. The data is iid as mentioned in the **Assumptions about the data**.
3. The data is paired since each respondent that didn't vote in 2016 but did in 2018 each have their own angry and afraid ratings.

- Null hypothesis is that anger or fear are equally as effective at driving increases in voter turnout from 2016 to 2018.
- Alternative hypothesis is that either anger or fear was more effective at driving increases in voter turnout from 2016 to 2018.

Conduct your test. (5 points)

```
more_afraid = sum(Q4_f$geafraid > Q4_f$geangry, na.rm=T)
trials = sum(Q4_f$geafraid < Q4_f$geangry | Q4_f$geafraid > Q4_f$geangry, na.rm=T)
binom.test(more_afraid , trials)
```

```
##
##  Exact binomial test
##
## data:  more_afraid and trials
## number of successes = 25, number of trials = 49, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.3633783 0.6557526
## sample estimates:
## probability of success
##              0.5102041
```

```
r = (more_afraid - (trials - more_afraid))/trials
r
```

```
## [1] 0.02040816
```

We fail to reject the null hypothesis that anger or fear were equally as effective at driving increases in voter turnout from 2016 to 2018. The p-value of 1 is greater than our predetermined cut off value of 0.05. There are 25/49 = .51 of people who didn't vote in 2016 but did in 2018 responded that they were more afraid than angry and 24/49 = .49 responded more angry than afraid. In other words, there is 2% of the people who didn't vote in 2016 but voted in 2018 who felt more fearful than angry. This is not a large difference in effect size, we cannot say that either anger or fear was more effective at driving increases in voter turnout. In addition, the practical significance also be represented by the correlation of people who more afraid and the people who were more angry. A correlation of 0.02 suggests a very small effect.

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)

We chose the question "Is there a difference in education level of US voters who support building the wall in Mexico and those who don't?"
This question is relevant to understanding the behavior of voters since this issue is highly debated and has gained traction among voters which could caused an effect on the election in 2016 and 2018. The issue of building the wall between US and Mexico got considerable coverage in the media and who would like to know who are supporting or against this order, in particular, we are interested in voters' education levels.

To answer this question, we used the educ variable and the wall variable. The education variable represents how much education a respondent has and the wall variable represents how much a respondent disapproves or approves of the wall.

Educ Variable: "What is the highest level of education you have completed?" 1 = No HS
2 = High school graduate
3 = Some college
4 = 2-year
5 = 4-year
6 = Post-grad

Wall variable: Do you favor, oppose, or neither favor nor oppose building a wall on the U.S. border with Mexico?
1 = Favor a great deal
2 = Favor moderately
3 = Favor a little
4 = Neither favor nor oppose
5 = Oppose a little
6 = Oppose moderately
7 = Oppose a great deal

We group those who favor the wall a great deal, moderately, and little into an "approved" the wall category and those who opposte the wall a great deal, moderately, and a little into a "disapproved" the wall category.

We do not include voters who neither favor or oppose the wall. Thus, we create a binary variable for wall approval.

```
Q5 <- select(A, educ, wall) %>%
      mutate(level = case_when(
        educ == 1 ~ 'No HS',
        educ == 2 ~ 'High school graduate',
        educ == 3 ~ 'Some college',
        educ == 4 ~ '2-year',
        educ == 5 ~ '4-year',
        educ == 6 ~ 'Post-grad')) %>%
      filter(wall != 4 ) %>%
      mutate(approval = case_when(
        wall <= 3 ~ 'Approved',
        wall >= 5 ~ 'Disapproved'))

head(Q5)
```

```
##   educ wall                level    approval
## 1    6    5            Post-grad Disapproved
## 2    2    7 High school graduate Disapproved
## 3    3    2         Some college    Approved
## 4    1    6                No HS Disapproved
## 5    6    7            Post-grad Disapproved
## 6    3    1         Some college    Approved
```

Perform EDA and select your hypothesis test (5 points)

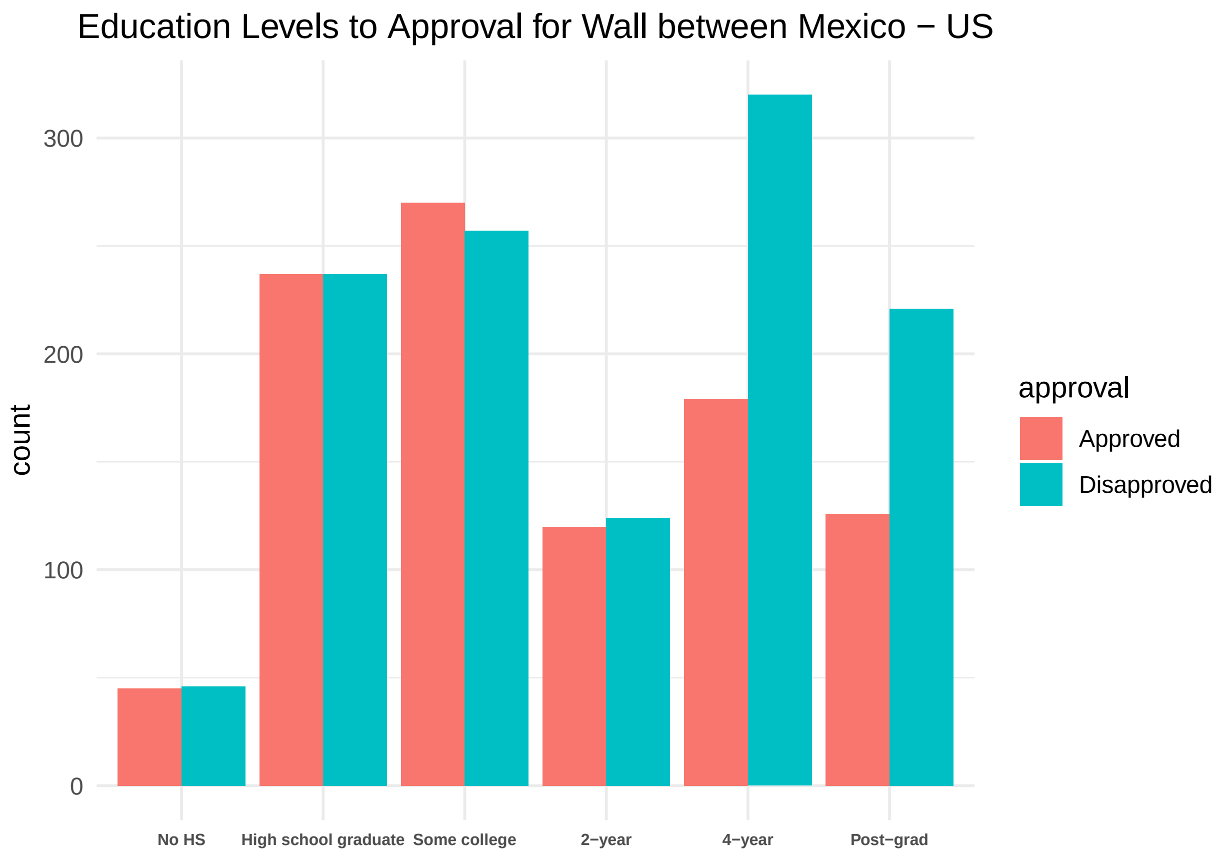
```
table(Q5$level)
```

```
##
##               2-year               4-year High school graduate
##                  244                  499                  474
##                No HS            Post-grad         Some college
##                   91                  347                  527
```

```
table(Q5$approval)
```

```
##
##    Approved Disapproved
##         977        1205
```

```
Q5 %>% mutate(reorder = fct_reorder(level, educ)) %>%
       ggplot(aes(x=reorder, fill = approval)) +
       geom_bar(position = 'dodge') +
       labs(title = 'Education Levels to Approval for Wall between Mexico - US') +
       theme_minimal()  +
       theme(axis.title.x = element_blank(),
             axis.text.x = element_text(size=6,face="bold"),
             plot.title = element_text(hjust = 0.5))
```



It appears from the histogram above that the more education a respondent has, the less likely they are to approve of the wall. A large difference in the number of those who approve and disapprove can be seen once a respondent has 4-years of college education. For lower levels of education, the number of those who disapprove and approve of the wall are almost equal to each other. Thus, it appears voters are much more likely to disapprove of the wall if they have a bachelor's degree or higher.

Based on your EDA, select an appropriate hypothesis test.

We chose the Wilcoxon rank-sum test because there is no pairing in the data and the data meets the following assumptions:

1. The variables of interest (education and approval of building a wall) are ordinal. The order of education level is the order of lowest to highest from 1 - 6, our wall responses are in order of favor to oppose from 1 - 7.
2. The data is iid as explained in the **Assumptions about the data**.

- Null hypothesis: There is no difference in education level between US voters who disapprove or approve building a wall on the U.S border with Mexico.

- Alternative hypothesis: There is a difference in education level between US voters who disapprove or approve building a wall on the U.S border with Mexico.

Conduct your test. (2 points)

```
wilcox.test(Q5$educ ~ as.numeric(factor(Q5$approval)))
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  Q5$educ by as.numeric(factor(Q5$approval))
## W = 508770, p-value = 2.467e-08
## alternative hypothesis: true location shift is not equal to 0
```

Our p-value is 2.467 e-08 which means our test is very highly significant so we reject the null hypothesis that there is no difference in education level between US voters who disapprove or approve building a wall on the U.S. border with Mexico.

```
cor.test(Q5$educ, as.numeric(factor(Q5$approval)), method='spearman')
```

```
## Warning in cor.test.default(Q5$educ, as.numeric(factor(Q5$approval)), method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  Q5$educ and as.numeric(factor(Q5$approval))
## S = 1524743778, p-value = 2.223e-08
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1193893
```

To measure practical signifiance, we calculated the correlation between education level and the approval/disapproval of building the wall. The correlation coefficient is 0.12 which suggests a small effect. We also computed a common language effect size below.

```
agg_tbl <- select(Q5, level, approval) %>%
       group_by(level, approval) %>%
       count() %>%
       pivot_wider(names_from = level, values_from = n) %>%
       ungroup() %>%
       adorn_totals()
colnames <- c("approval", "No HS", "High school graduate", "Some college", "2-year", "4-year", "Post-gra
agg_tbl <- agg_tbl[colnames]
agg_tbl
```

```
##     approval No HS High school graduate Some college 2-year 4-year Post-grad
##     Approved    45                  237          270    120    179       126
##  Disapproved    46                  237          257    124    320       221
##        Total    91                  474          527    244    499       347
```

These are the counts of voters in six different education levels and whether they approved or disapproved of the wall. It would be more helpful to look at the porportion of these counts. Thus, we computed them below.

```
disapprove_portion <- round(as.numeric(agg_tbl[2,2:7])/as.numeric(agg_tbl[3,2:7]), 2)
approve_portion <- round(as.numeric(agg_tbl[1,2:7])/as.numeric(agg_tbl[3,2:7]), 2)
Education_level <- c("No HS", "High school graduate", "Some college", "2-year", "4-year", "Post-grad")
portion_tbl <- data.frame(
          Disapprove = c(disapprove_portion),
          Approve =c(approve_portion)) %>%
          cbind(Education_level)
portion_tbl
```

```
##   Disapprove Approve      Education_level
## 1       0.51    0.49                No HS
## 2       0.50    0.50 High school graduate
## 3        0.49    0.51         Some college
## 4        0.51    0.49               2-year
## 5        0.64    0.36               4-year
## 6        0.64    0.36            Post-grad
```

In term of common language effect size, the table above shows that between a voter who approves of the wall and a voter who disapproves of the wall, the person who disapproves of the wall will have more education than the person who approves of the wall.

Conclusion (3 points)

We conclude that there is a difference between education level between US voters who disapprove or approve building a wall on the U.S border with Mexico. However, the corelation between education level and the approval/disapproval is small. It also appears that the proportion of No HS up to 2 year college voters have similar approval/dispproval rates from 49% - 51% but voters with education level of 4-year and above have doubled disapproval rates compared to their approval rates. This tells us that more voters with higher education level (4-year and above) disapproved of building the wall between Mexico and US. The person who disapproves of the wall is likely to have more education than the person who approves of the wall.