Maini_Kumar_Robertson_Lab2_2

July 9, 2020

[2]:
```
# Data
A = read.csv("anes_pilot_2018.csv")
```

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

0.2 Topic

Survey used `feeling thermometer` wigit to understand how the respondents **feel** about some persons or groups (which include both journalists and police). Respondents need to answer on an ordinal scale with a number between 0 and 100. This setting, although, gives a good idea about how favorable or unfavorable the respondant feels about blacks and police, it is only a rough proxy for **respect**.

0.2.1 Operational Definition

We are going to use the respondent's feeling about the police and journalists (from columns `ftpolice` and `ftjournal`) as a proxy for the respect they have for these groups. Numbers close to 100 mean high respect and numbers close to 0 mean low respect.

0.3 Exproratory Data Analysis

[3]:
```
# Basic variable summary
sub = A[c('ftjournal', 'ftpolice')]
summary(sub)
paste("Number of rows in the original data: ", nrow(sub))

# Visualizing data ros where no responses were provided
print("Non-responding entries")
(sub[sub['ftjournal'] < 0,])
```

```
   ftjournal         ftpolice
 Min.   : -7.00   Min.   :  0.00
 1st Qu.: 21.00   1st Qu.: 47.00
 Median : 52.00   Median : 70.00
 Mean   : 52.26   Mean   : 64.68
 3rd Qu.: 82.00   3rd Qu.: 90.00
 Max.   :100.00   Max.   :100.00
```

'Number of rows in the original data: 2500'

```
[1] "Non-responding entries"
```

	ftjournal	ftpolice
51	-7	84
597	-7	91

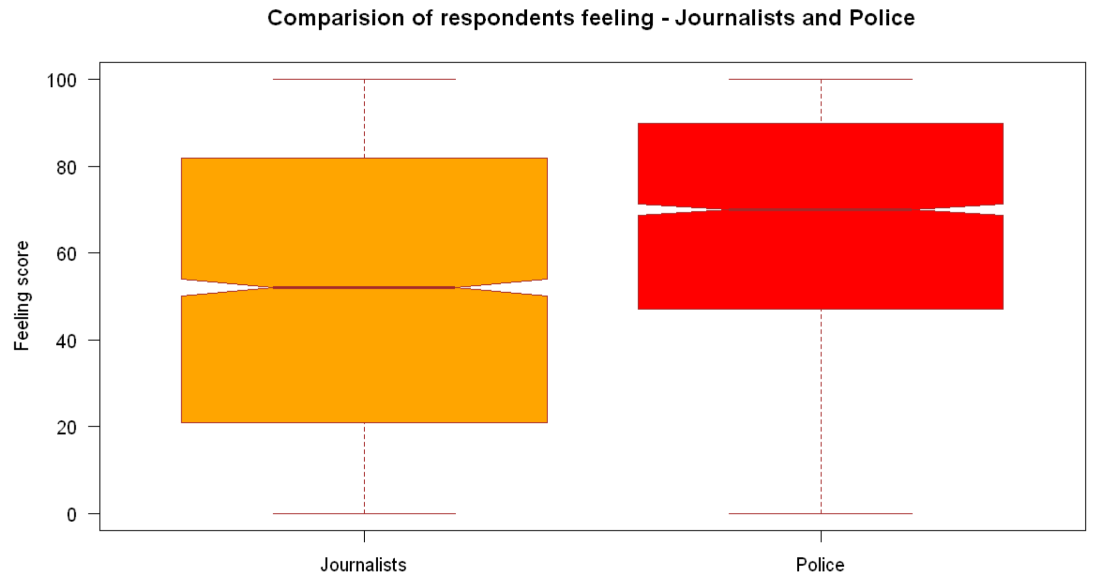
[4]:

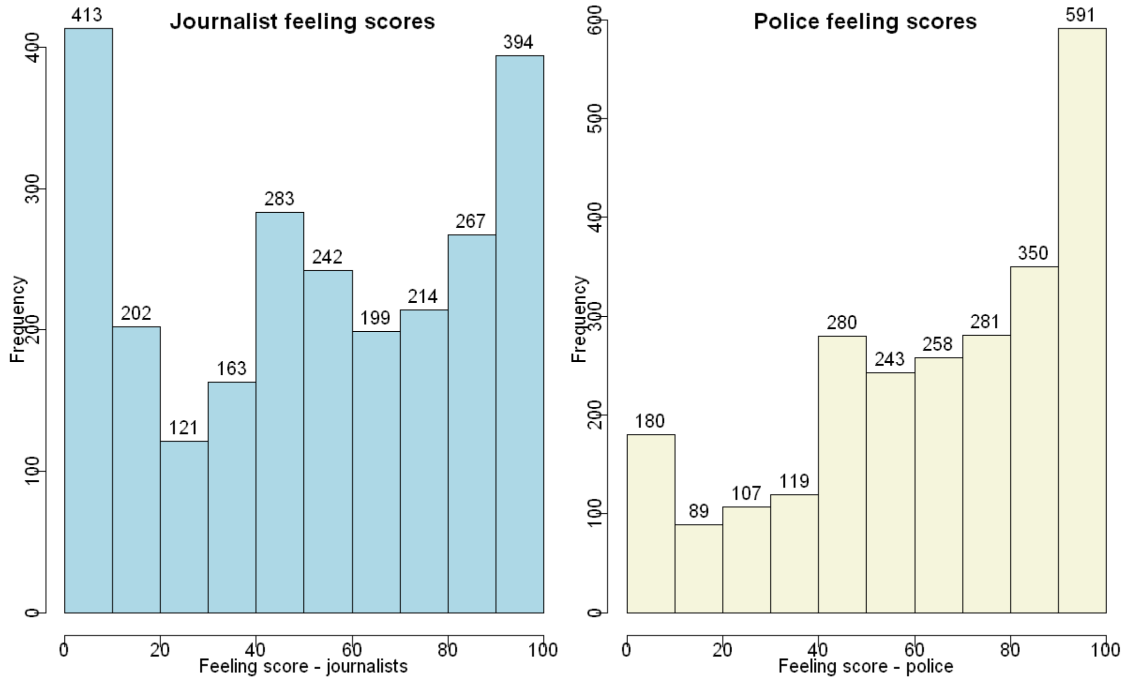
```
# Treating for no-response by removing those respondents
sub = sub[sub['ftjournal']>=0,]
summary(sub)
paste("Number of rows in the transformed data: ", nrow(sub))

# Single variable analysis (Box-plots)
options(repr.plot.height=6, repr.plot.width = 10)
boxplot(sub$ftjournal, sub$ftpolice, main = "Comparision of respondents feeling␣
 ↪- Journalists and Police", at = c(1,2),
        names = c("Journalists", "Police"), las = 1, col = c("orange","red"),␣
 ↪border = "brown", horizontal = F,
        notch = TRUE, ylab='Feeling score')
par(mfrow=c(1,2), mar=c(2,2,0,0), mgp=c(.8,.1,0))
x <- hist(sub$ftjournal, main='', col='lightblue', xlab = "Feeling score -␣
 ↪journalists")
text(x$mids,x$counts,labels=x$counts, adj=c(0.5, -0.5))
title('Journalist feeling scores', line=-1, adj=0.5)
Y <- hist(sub$ftpolice, main='', col='beige', xlab = "Feeling score - police")
text(Y$mids,Y$counts,labels=Y$counts, adj=c(0.5, -0.5))
title('Police feeling scores', line=-1, adj=0.5)
```

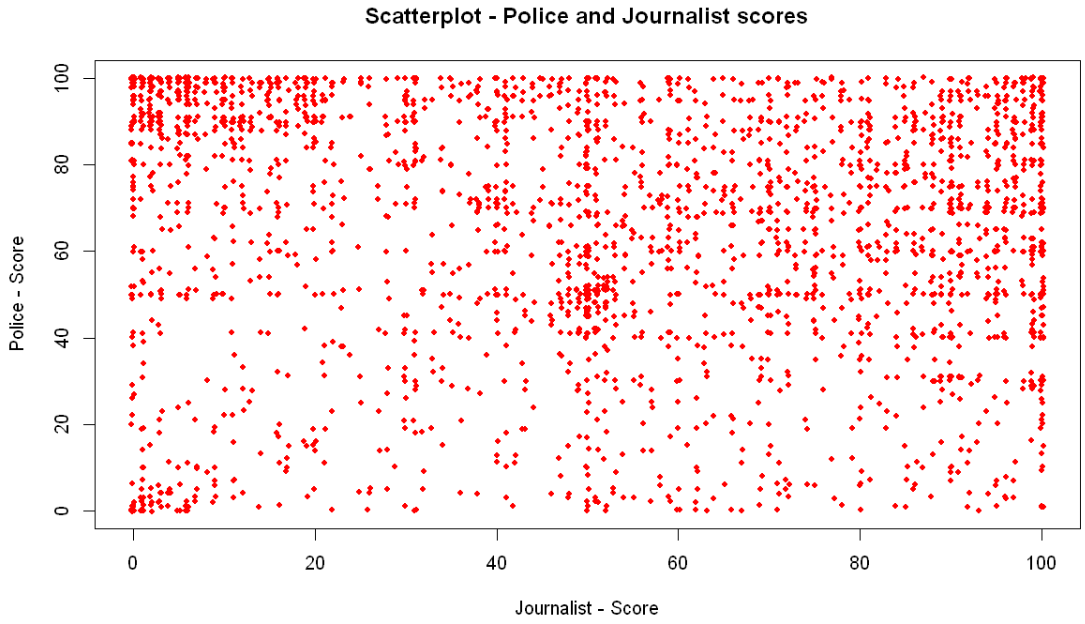
```
   ftjournal          ftpolice
 Min.   :  0.00   Min.   :  0.00
 1st Qu.: 21.00   1st Qu.: 47.00
 Median : 52.00   Median : 70.00
 Mean   : 52.31   Mean   : 64.67
 3rd Qu.: 82.00   3rd Qu.: 90.00
 Max.   :100.00   Max.   :100.00
```

'Number of rows in the transformed data: 2498'

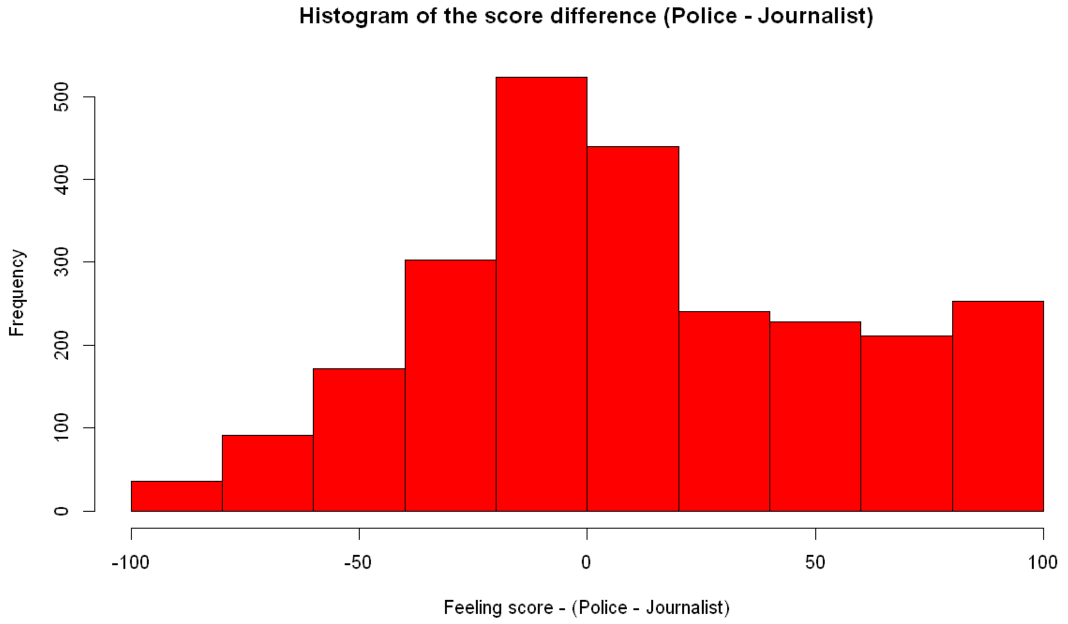




```
[5]: plot(jitter(sub$ftjournal), jitter(sub$ftpolice), main="Scatterplot - Police␣
     ↪and Journalist scores",
         xlab="Journalist - Score", ylab="Police - Score", pch=20, col='red')
```



```
[6]: hist(sub$ftpolice - sub$ftjournal, main='Histogram of the score difference␣
      ↪(Police - Journalist)', col='red',
              xlab = "Feeling score - (Police - Journalist)")
```



0.3.1 Basic EDA summary

1. All the respondents (2500) provided their feelings about `police` but 2 of them didn't provide an answer about `journalists`. **We suggest removing them from the consideration set as we are going to do paired hypothesis test where it will be important to have legitimate responses for both Police and Journalists**.
2. All the provided answers are in the range of 0 to 100 which makes sense (No need for null removal)
3. Looking at single variable summary, we see that:
   1. The range of scores is higher for Journalists than Police. The median score is higher for police than journalists (Refer Box-plot)
   2. Journalist score are more concentrated on both extremes where are police score is concentrated on the high end. The distribution in the middle is pretty much the same for both classes
4. The Scatterplot shows that individual scores for feeling towards police and journalists are not correlated. At every journalist score bucket, there seem to be a bias towards higher police score which might suggest that poeple feel better towards journalists than police in the sample, but we need to perform a hypothesis test to test this hypothesis.
5. The histogram of score difference per individual shows a bias towards positive.

0.4 Considerations for Hypothesis Test

There are a few considerations that we will take into account:

1. The feeling score is an **ordinal variable** (not a metric variable). A score of 20 is bettern than a score of 10 for example but it might not be correct to interpret that the difference in feeling towards any group is same when between a response of 10 and 20, 45 and 55 and 90 and 100. In this regard, the difference between these variable have limited meaning.

2. Every respondent in the sample (excluding 2 which didn't respond to poilce question) have a pair of feeling scores for police and journalists. As per our operational definition, this score is our proxy for respect and we can do a **paired hypothesis test**.

3. We assume that **each row of data (i.e. respondent) is independent**. This may be a strong assumption given the fact that all of the selected respondents share a common trait (completing most number of survey's as per YouGov panel). Although, it is possible that some of teh data rows are dependent, it is not probable.

4. **Sampling is not exactly representative of US voting population**. This has been discussed and the authors have tried to compensate for it using weights. We are however assuming that the weight for each participant is 1 (i.e. the sample is representative of the US voting population) for our hypothesis testing exercise below.

The question asks US voters have more respect for which group (Police or Journalists). As we don't have an strong opinion to support any one of the groups, we will start with the **two-sided hypothesis**. If we find that there is actually a stat-sig difference in respect across the two groups, we can test one sided hypothesis to test which one. Below we state the two-sided hypothesis.

0.5 Two sided Hypothesis

H_0: There is no difference in respect between police and journalists among US voters.

H_A: There is a difference in respect between police and journalists among US voters.

0.6 Appropriate test

Given that we want to do **paired test** and the **variables are Ordinal in nature**, We wil use **sign test** for hypothesis testing. This test has lower power which means there is a higher chance of not being able to detect the effect. The key assumptions are: 1. Differences in Police and Journalist ratings across respondents is independent. As per above discussion, we are assuming this condition to be satisfied 2. The ratings for police and journalists are ordinal (which is obviously true) 3. All the differences are sampled from the same population (which is what we are assuming to be true)

0.7 Significance Level

We want to perform the test at **99% significance level** which means the expected type 1 error rate to be 1%.

0.8 Conducting the test

[7]:

```
library('BSDA')
```

```
Warning message:
"package 'BSDA' was built under R version 3.6.3"Loading required package:
lattice

Attaching package: 'BSDA'

The following object is masked from 'package:datasets':

    Orange
```

[8]:

```
SIGN.test(sub$ftpolice , sub$ftjournal, md = 0, conf.level = 0.99, alternative␣
 ↪= 'two.sided')
```

```
Dependent-samples Sign-Test

data:  sub$ftpolice and sub$ftjournal
S = 1373, p-value = 2.361e-11
alternative hypothesis: true median difference is not equal to 0
99 percent confidence interval:
 2 7
sample estimates:
median of x-y
            4

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.9890      2      7
```

```
Interpolated CI       0.9900      2      7
Upper Achieved CI     0.9902      2      7
```

0.9 Interpretation of the test

We have the following observations:

1. The total sample size is 2417. Out of that, there are 1373 respondents who gave higher score to Police than Jourlalists.
2. The p-value for this test (at 99% significance level) is $2.361 * 10^{-11}$. which is way smaller than our target significance level (of 0.01). Hence, we have statistically significant evidence to reject the Null hypothesis that the US voters repect police and journalists the same.
3. The 99 percent confidence interval for the median difference in Score for proportion of people favoring police over journalists is [0.5418, 0.5941]. We can see that this number doesn't contain 0.5 which was our null hypothesis.
4. The sample median for the difference between police and journalist median scores is 4

As mentioned before, at this point, we can perform a one-sided hypothesis test to see if the US voter prefer police over journalists.

0.9.1 One sided Hypothesis

$H1_0$: There is no difference in respect between police and journalists among US voters.

$H1_A$: US voters respect Police more than Journalists.

```
[9]: SIGN.test(sub$ftpolice , sub$ftjournal, md = 0, conf.level = 0.99, alternative␣
      ↪= 'greater')
```

```
Dependent-samples Sign-Test

data:  sub$ftpolice and sub$ftjournal
S = 1373, p-value = 1.18e-11
alternative hypothesis: true median difference is greater than 0
99 percent confidence interval:
   2 Inf
sample estimates:
median of x-y
            4

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.9893      2    Inf
Interpolated CI       0.9900      2    Inf
Upper Achieved CI     0.9904      2    Inf
```

From the above one sample result, we can conclude that at 99% confidence level, we can reject the null hypothesis $H1_0$in favor of the alternative hypothesis $H1_A$.

0.10 Effect size

A measure of difference can be the median difference in the sample between police and journalist score whcih is 4. Additionally, If we want to use a dimentionless effect size estimate, we can calculate PS_{dep} which is relevant in the case of dependent sample non-parametric testing and calculated as

$$PS_{dep} = \frac{n_+}{N}$$

Here n_+ is the total number of cases which have police_score > journalist_score and N is the total number of cases in the consideration set (excluding the ties). The interpretation of this score is that it is the probability that in a randomly sampled pair of scores, score from condition B will be greater than score from condition A. When we put in the numbers, we have

$$PS_{dep} = \frac{1373}{2417} = 0.568$$

PS_{dep} indicates that there is 56.8% chance of police score being better than jornalist score for an individual (based on our sample).

0.11 Practical significance and Discussion

Although we got a highly statistically significant result, the effect size seems to be medium (4 point difference) and it is rather hard to interpret this difference as the feeling score is not a metric. From the effect size estimator PS_{dep} we have 56.8% chance of police score being better than jornalist score for an individual which seems to be significant.

Our research question was specifically about respect rather than likeness hence there is **an asymmetry between out research question and the actual question we are trying to answer**.

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

0.13 Introduction

To address this question, we need to know two things: 1) what constitutes a ”Republican (Democratic) voter?, and 2) How old are the individuals in these two groups.

The first question is the more difficult one to answer. The issue lies in the phrasing of the quesiton, as there are several ways to interpret what a “Republican voter” is. Should we use party membership, although that does not neccessarily mean the person votes? Should we count independent voters who lean Repbulican (indicating that they have likely voted for Republicans in the past and will do so again) as Republican voters?

For the purposes of this analysis, we have operationalized the question in the following way: we consider a person who “generally thinks of themselves as a Republican or Democrat” to be a Republican or Democratic voter. This is not a perfect definition for the reasons stated above, but it gives us the best view of the respective parties voting bases.

Accordingly, we use the pid1d and pid1r variables, which is the answer to the question “Generally speaking, do you consider yourself a Republican, a Democrat…”. These two variables need to be merged into a single table since together they are the full dataset that answers the question posed. We also used the birthyr variable which tells us the year the voter was born.

0.14 Exploratory data analysis

[10]:

```
library(dplyr)
#if(!require(lsr)){install.packages("lsr")}
library(lsr)

party <- A[,c('pid1d', 'pid1r','birthyr')]

# Respondents that declined to answer the party affliation question
print("Rows of data with missing score entry")
party[(party$pid1d == -7)|(party$pid1r == -7),]
party <- party[!((party$pid1d == -7)|(party$pid1r == -7)),]

# Summary of the filtered resulting dataset
print("Overall Summary of the filtered data")
summary(party)

# Tagging the part affliations
party$party <- ifelse(party$pid1d == 2 | party$pid1r == 2, 'REPUBLICAN',␣
 ↪ifelse(party$pid1d == 1 | party$pid1r == 1,
                                                                        ␣

 ↪'DEMOCRAT', 'O'))

# Removing all the data rows representing votes who are neither democrat nor␣
 ↪republicans
party <- party[party$party != 'O',]

# Final dataset size
paste("Number of records (total) :", nrow(party))

#Create a separate dataset for each party
repid <- filter(party, party == 'REPUBLICAN')
demid <- filter(party, party == 'DEMOCRAT')

#Calculating Age

repid$age <- 2018 - repid$birthyr
demid$age <- 2018 - demid$birthyr

#Summarize each party-specific data set
summary(repid$age)
summary(demid$age)
```

```
#Show a histogram of each part-specific dataset
hist(demid$age, main='Democrat Voters - Age Histogram', col = 'blue', xlab =␣
 ↪'Voter Age')
hist(repid$age, main='Republican Voters - Age Histogram', col = 'red', xlab =␣
 ↪'Voter Age')
```

```
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


[1] "Rows of data with missing score entry"
```

	pid1d	pid1r	birthyr
497	-1	-7	1993
596	-7	-1	1953

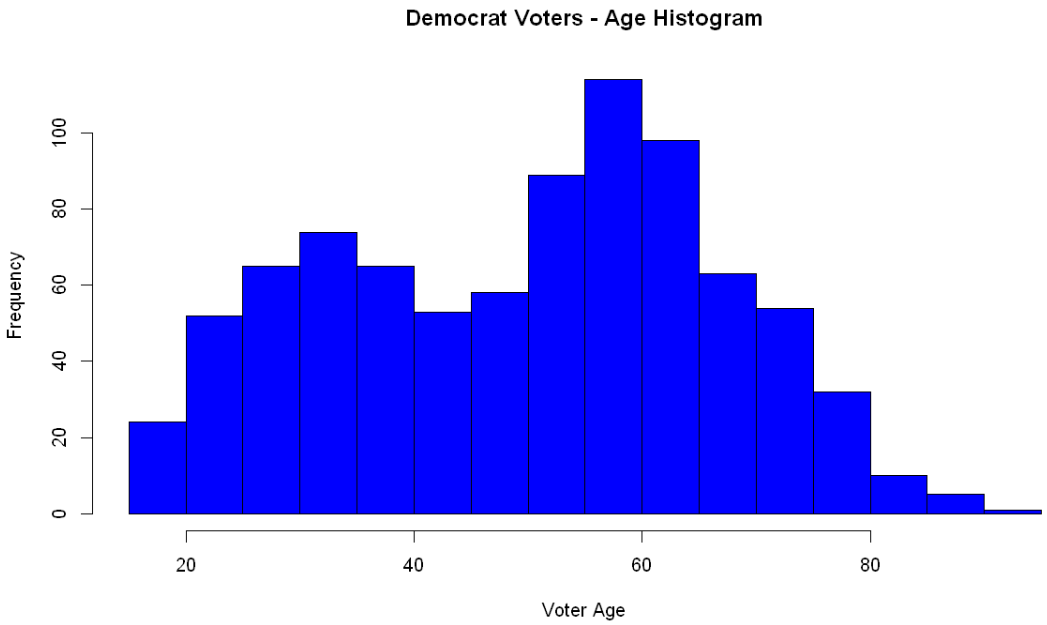
```
[1] "Overall Summary of the filtered data"
```

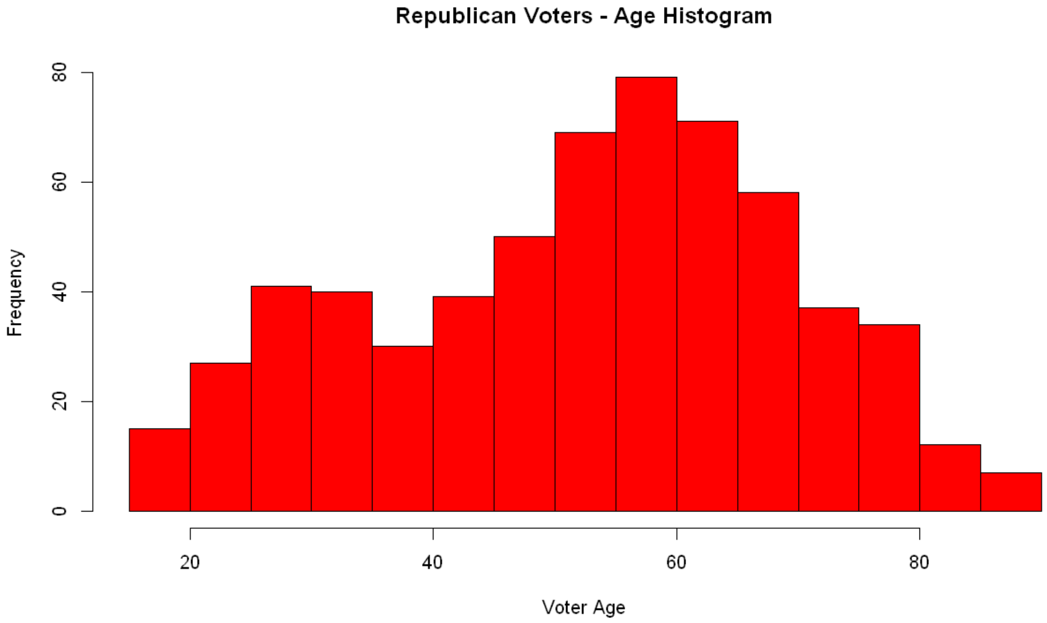
```
     pid1d              pid1r              birthyr
 Min.   :-1.0000   Min.   :-1.0000   Min.   :1927
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:1956
 Median :-1.0000   Median :-1.0000   Median :1966
 Mean   : 0.4155   Mean   : 0.4644   Mean   :1969
 3rd Qu.: 2.0000   3rd Qu.: 2.0000   3rd Qu.:1983
 Max.   : 4.0000   Max.   : 4.0000   Max.   :2000
```

'Number of records (total) : 1466'

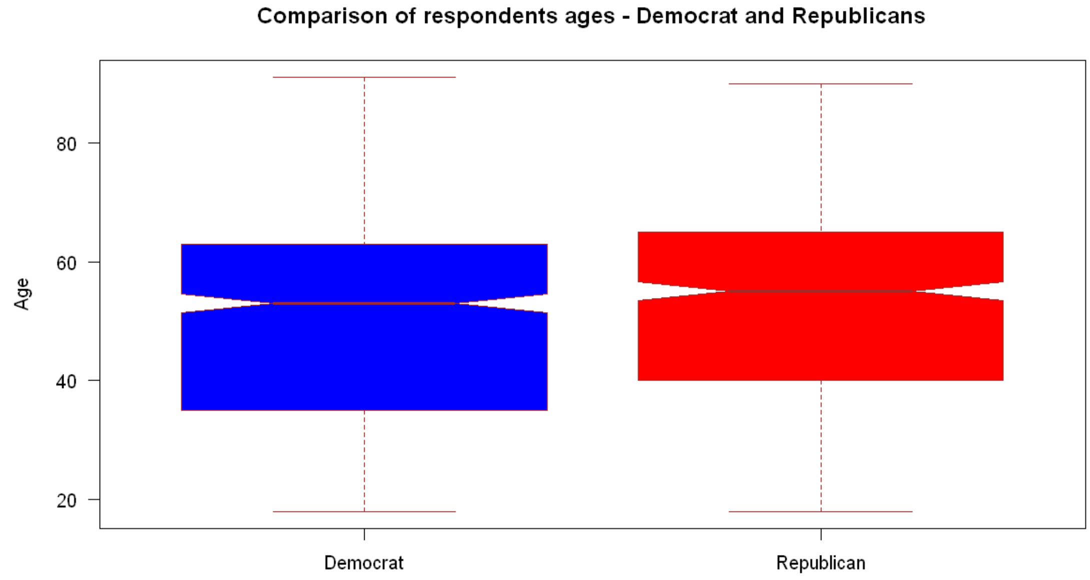
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   40.00   55.00   52.86   65.00   90.00


   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   35.00   53.00   50.23   63.00   91.00
```





[12]:
```
boxplot(demid$age, repid$age, main = "Comparison of respondents ages - Democrat␣
↪and Republicans", at = c(1,2),
        names = c("Democrat", "Republican"), las = 1, col = c("blue","red"),␣
↪border = "brown", horizontal = F,
        notch = TRUE, ylab='Age')
```



Basic EDA summary Steps:

1) Merge pid1d and pid1r, then create a data frame that only has the responses to the pid1d/pid1r and birth year
2) Do a simple summary of the new data frame. The answer to pid1d/pid1r is not meaningful, but the answer to birth year is.
3) Break the relevant groups up using a filter command. Here we filter on response 1 (people who identify as Democrats) and 2 (people who identify as Republicans).
    a) 1466 of 2500 (~59%) of people responded in one of these two ways
    b) Since the question is restricts the population to necessarily belonging to one of these two groups, dropping ~41% of responses is appropriate.
4) Do a summary of the two groups
5) Show a histogram of the Age of these two groups (with 2018 as the reference year). A brief look at these graphs shows that they have similar meansand the shape of the distributions is also similar. It seems roughly normal and is not skewed.

0.15 Hypothesis test

A simple two-tailed t-test is the most appropriate test in this case. Since we don't have a hypothesis for why one group "should" be older than another, we should stick to testing whether the means are different or not.

Assumptions of the t-test

1: Republican and Democratic voter ages are independent. We hold this to be obviously true.

2: The means of the two populations are approximately normally distributed. We see that this is generally true through the hisotgrams of each variable. On top of that, sample sizes are all > 400 and the data is not skewed in any extreme manner, so we can rely on the Central Limit Theorem to ensure that this assumption holds true.

3: Age is a metric variable

4: Variance of the two samples seem similar from the histogram

5: For this analysis, we are assuming that the sample is random and representative of general US voting population. This is a strong assumption (specially given the concerns of the survey authors) and weights are available to somewhat correct for this. But for this analysis, we will not be using the weights.

Two sided Hypothesis Test H_0: There is no difference in birth year between Republican and Democratic voters.

H_A: There is a difference in birth year between Republican and Democratic voters.

Significance Level We will perform this test to a 99% confidence level.

Follow up One-sided Hypothesis Test If we reject the null hypothesis based on the two-tailed t test, we will follow up with a one-tailed t-test in the direction that we suspect there is a difference. Based on an initial look at the historgrams, we suspect that Republicans are older than Democrats, so we may find it useful to run a subsequent one-sided test to see if the mean age of Republicans is older than that of Democrats.

H_0: There is no difference in age between Republican and Democratic voters.

H_A: The mean age of Democratic voters is less than the mean age of Republican voters.

[13]:

```r
#Conduct a two-tailed test to a 99% confidence level
t.test(demid$age, repid$age, conf.level = .99)

#Cohen's D to assess practical significance of the result
paste("Cohen's D : ",cohensD(demid$age, repid$age))

#Follow up with a one-tailed test to determine if Republicans are older than␣
 ↪Dems, to a 99% conf level
t.test(demid$age, repid$age, alternative = "less", conf.level = .99)
```

```
Welch Two Sample t-test
```

```
data:  demid$age and repid$age
t = -2.939, df = 1309.7, p-value = 0.00335
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -4.9235927 -0.3206645
sample estimates:
mean of x mean of y
 50.23337  52.85550
```

'Cohen\'s D : 0.15575500851897'

```
Welch Two Sample t-test

data:  demid$age and repid$age
t = -2.939, df = 1309.7, p-value = 0.001675
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
      -Inf -0.544058
sample estimates:
mean of x mean of y
 50.23337  52.85550
```

Explanation of Results We reject the null hypothesis, with a p-value of 0.00335 at 99% confidence level. This means that there is a statistically significant difference in the means of the birth year's of the Republican and Democratic sample voters.

We also found a low practical significance (Cohen's D = 0.1558). This means that while the difference is real, it is not a hugely significant feature of the two groups. This is not surprising based even on a quick look at the histrograms in the EDA section.

We then ran a one-sided t-test, where we reject the null hypothesis (at 99% confodence level) in favor of the alternative hypothesis that Democratic voter's mean age is less than that of Republican voters. Here, our p-value was .00168.

In conclusion, based on this data, we can say that Republican voters are older than Democratic voters on average. However, the effect size is small.

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

0.16.1 Introduce your topic briefly. (5 points)

Topic Introduction We shall attempt to answer the question , do a majority of independent voters believe that the federal investigations of Russian election interference are baseless.

We shall use variable pid1d and pid1r to identify the independent voters i.e. we shall only get voters who indicated that they are independent = 3 in either of these fields.

We shall use the field muellerinv to get identify voters sentiment on federal investigations of Russian election interference esp for this question we shall focus on proportion of voters who selected 7 : Disapprove extremely strongly as response to this survey question.

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

[14]:

```
# get all voters
voters <- A[,c('pid1d', 'pid1r','muellerinv')]
#review unique values of pid1d field
print("Unique Values of pid1d field :")
unique(voters$pid1d)
#review unique values of pid1r field
print("Unique Values of pid1r field :")
unique(voters$pid1r)
#review unique values of muellerinv field
print("Unique Values of muellerinv field :")
unique(voters$muellerinv)

# only get voters who indicated that they are independent = 3
indpVoters <- voters[(voters$pid1d == 3)|(voters$pid1r == 3),]
print("Summary of Independent Voters :")
summary(indpVoters)
paste("Number of Independent Voters : ", nrow(indpVoters))
```

```
[1] "Unique Values of pid1d field :"
```

```
1. 2 2. -1 3. 3 4. 1 5. 4 6. -7
```

```
[1] "Unique Values of pid1r field :"
```

```
1. -1 2. 2 3. 3 4. 1 5. 4 6. -7
```

```
[1] "Unique Values of muellerinv field :"
```

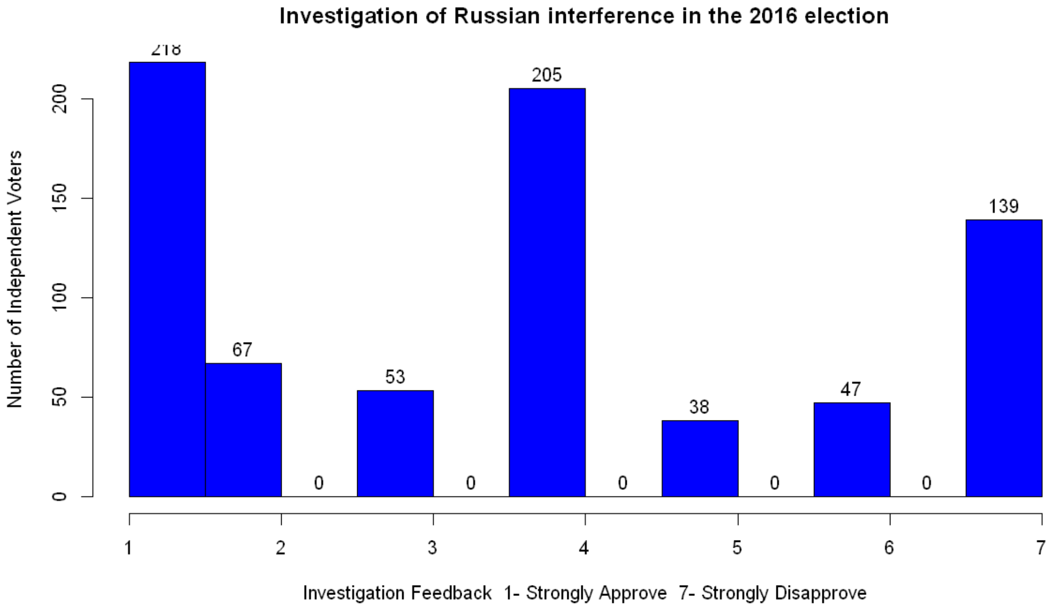
```
1. 4 2. 1 3. 5 4. 2 5. 7 6. 3 7. 6 8. -7
```

```
[1] "Summary of Independent Voters :"
```

```
     pid1d              pid1r           muellerinv
 Min.   :-1.0000   Min.   :-1.000   Min.   :1.000
 1st Qu.:-1.0000   1st Qu.:-1.000   1st Qu.:1.000
 Median :-1.0000   Median : 3.000   Median :4.000
 Mean   : 0.8566   Mean   : 1.143   Mean   :3.619
 3rd Qu.: 3.0000   3rd Qu.: 3.000   3rd Qu.:5.000
 Max.   : 3.0000   Max.   : 3.000   Max.   :7.000
```

’Number of Independent Voters : 767’

```
[15]: histVal <- hist(indpVoters$muellerinv, main='', col='blue',
               xlab = "Investigation Feedback  1- Strongly Approve  7- Strongly␣
       ↪Disapprove",
               ylab ="Number of Independent Voters" )
      text(histVal$mids,histVal$counts,labels=histVal$counts, adj=c(0.5, -0.5))
      title('Investigation of Russian interference in the 2016 election', line=1,␣
       ↪adj=0.5)
```



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

Hypothesis Testing We shall use two sided proportion test to access sentiment of independent voters at significance level =0.05.

Null Hypothesis H_0: Independent voters niether approve nor disapprove of federal investigations of Russian election interference i.e. proportion of voters who strongly disapprove = 0.5

Alternate Hypothesis H_A: prop of voters who strongly disapprove != 0.5

The key conditions for this test are:

1. Random sample - we can assume this is met as survey respondedents are random voters who elect to take YouGov survey.
2. The sample is Independent and identically distributed - which is true as these are voters who elect to take YouGov survey and are not related to each other esp. in some extreme cases when members of same family might be takign survey even in this condition their responses would be deemed independent.

3. Sample size restrictions - number of independent voters who stronly disapprove and do not disapprove should be greater than 10
    a) number of independent voters who strongly disapprove > 10 (exact value : 139)
    b) number of independent voters who do not strongly disapprove > 10 (exact value : 767 - 139 = 628)

0.16.4 Conduct your test. (5 points)

[17]:
```
# count of number of voters who strongly disapprove of federal investigations␣
 ↪of Russian election interference
dispVoters <- indpVoters$muellerinv == "7"
summary(dispVoters)

votersWhoDisapprove = sum(dispVoters)
totalVoters = nrow(indpVoters)

paste("Voters who disapprove : ", votersWhoDisapprove)
paste("Total Number of Voters : ", totalVoters)

prop.test(x = votersWhoDisapprove, n= totalVoters,
          conf.level = 0.95, alternative = 'two.sided')
```

```
   Mode   FALSE    TRUE
logical     628     139
```

’Voters who disapprove : 139’

’Total Number of Voters : 767’

```
1-sample proportions test with continuity correction

data:  votersWhoDisapprove out of totalVoters, null probability 0.5
X-squared = 310.49, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1549640 0.2107399
sample estimates:
        p
0.1812256
```

Result Interpretation Since p-value (2.2e-16) < significance level (0.05) hence we reject the null hypothesis.

The results are staitically significant.

Also since 0.5 does not lie in the confidence interval hence we shall conduct a new
less than one sided proportion test with significance level of 0.05

Null Hypothesis $H1_0$: Independent voters niether approve nor disapprove of federal investigations of Russian election interference i.e. prop of voters who disapprove =0.5

Alternate Hypothesis $H1_A$: prop of voters who disapprove of federal investigations of Russian election interference < 0.5

[18]:

```
prop.test(x = votersWhoDisapprove, n= totalVoters,
          conf.level = 0.95,alternative = 'less', correct=F)
```

```
1-sample proportions test without continuity correction

data:  votersWhoDisapprove out of totalVoters, null probability 0.5
X-squared = 311.76, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.2052115
sample estimates:
        p
0.1812256
```

Final Results Since p-value (2.2e-16) $<$ significance level (0.05) hence we reject the null hypothesis in favor of alternate hypothesis.

Majority of Independent voters do not think that federal investigations of Russian election interference is baseless.

using cohen h to measure effect size h = 0.20: "small effect size" h = 0.50: "medium effect size" h = 0.80: "large effect size"

Details Available At: https://en.wikipedia.org/wiki/Cohen%27s_h

[19]:

```
# cohen h function
propVotersWhoDisapprove <- votersWhoDisapprove/totalVoters
paste("Proportion of Voters Who Disapprove : " , propVotersWhoDisapprove)
propVotersWhoApprove <- (totalVoters - votersWhoDisapprove)/totalVoters
paste("Proportion of Voters Who Approve : " , propVotersWhoApprove)
cohenH <- 2*(asin(sqrt(propVotersWhoApprove)) -␣
 ↪asin(sqrt(propVotersWhoDisapprove)))
paste("Cohen H : " , cohenH)
```

'Proportion of Voters Who Disapprove : 0.18122555410691'

'Proportion of Voters Who Approve : 0.81877444589309'

'Cohen H : 1.38262499410043'

Since value of cohenH $>$ 0.8 hence results have large effect size and is practically significant.

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

0.18 Topic Introduction

We shall attempt to answer the question , was anger or fear more effective at driving increases in voter turnout from 2016 to 2018. We shall use variable turnout16 to identify voters who definelty voted in 2016. We shall use variable turnout18 to identify voters who definelty voted in 2018.

To determine increase in number of voter turnout from 2016 to 2018, we shall choose value of turnout16 != 1 i.e. did not vote or probably didn't vote in 2016 and turnout18 values of 1, 2 or 3 i.e. definelty voted in 2018. We shall use the field geangry to access voter sentiment on anger. We shall use the field geafraid to access voter sentiment on fear.

0.19 Exploratory data analysis

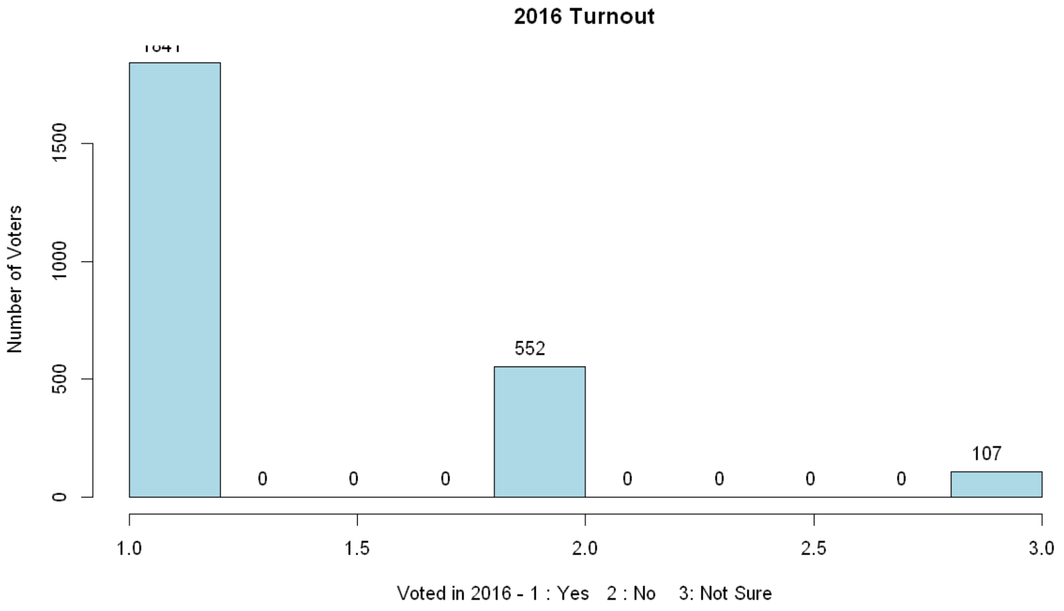
[20]:

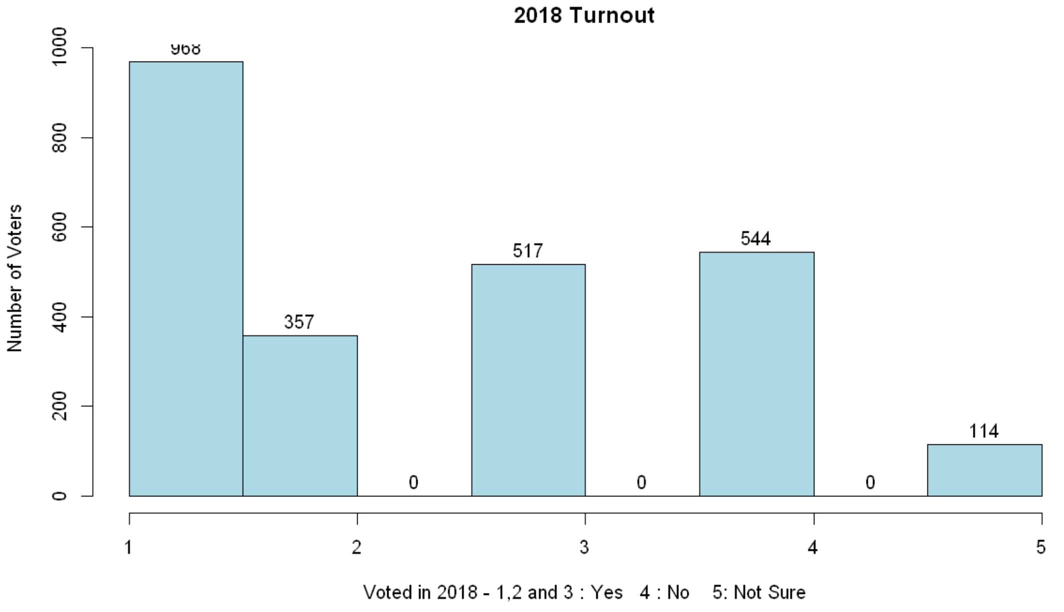
```
angerFearAnalysis <- A[,c('turnout16', 'turnout18','geangry' , 'geafraid')]

# Visual analysis of this data set , before filtering for voters who voted both␣
 ↪in 2016 and 2018

x <- hist(angerFearAnalysis$turnout16, main='', col='lightblue',
          xlab = "Voted in 2016 - 1 : Yes   2 : No    3: Not Sure",
          ylab = "Number of Voters",
          xlim=c(1,3)
         )
text(x$mids,x$counts,labels=x$counts, adj=c(0.8, -0.8))
title('2016 Turnout', line=1, adj=0.5)

y <- hist(angerFearAnalysis$turnout18, main='', col='lightblue',
         xlab = "Voted in 2018 - 1,2 and 3 : Yes   4 : No    5: Not Sure",
         ylab = "Number of Voters",
         xlim=c(1,5)
         )
text(y$mids,y$counts,labels=y$counts, adj=c(0.5, -0.5))
title('2018 Turnout', line=1, adj=0.5)
```





[21]:

```
summary(angerFearAnalysis)
print("Total Number of Voters :")
paste(nrow(angerFearAnalysis))
```

```
  turnout16        turnout18        geangry          geafraid     
 Min.   :1.000   Min.   :1.000   Min.   :-7.00   Min.   :-7.000  
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 2.00   1st Qu.: 2.000  
 Median :1.000   Median :2.000   Median : 3.00   Median : 3.000  
 Mean   :1.306   Mean   :2.392   Mean   : 2.93   Mean   : 2.662  
 3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.: 4.00   3rd Qu.: 4.000  
 Max.   :3.000   Max.   :5.000   Max.   : 5.00   Max.   : 5.000  
```

```
[1] "Total Number of Voters :"
```

'2500'

[22]:

```
# get voters who indiacted they did not vote in 2016 i.e. turnout16 !=1
# and who indicated they definetly voted in 2018 i.e. turnout18 should be 1 ,2␣
 ↪or 3
# this will gove the data set needed for anger and fear sentiment analysis

angerFearAnalysis <- angerFearAnalysis[(
                                          (angerFearAnalysis['turnout16'] !=␣
 ↪1) &
                                          (angerFearAnalysis['turnout18'] < 4)

                                      ),]




summary(angerFearAnalysis)
print("Increase In Number of Voters from 2016 to 2018:")
paste("Total number of respondents:",nrow(angerFearAnalysis))
```

```
  turnout16        turnout18       geangry           geafraid     
 Min.   :2.000   Min.   :1.00   Min.   :-7.000   Min.   :-7.000  
 1st Qu.:2.000   1st Qu.:1.00   1st Qu.: 2.000   1st Qu.: 2.000  
 Median :2.000   Median :1.00   Median : 3.000   Median : 3.000  
 Mean   :2.147   Mean   :1.75   Mean   : 2.793   Mean   : 2.603  
 3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.: 4.000   3rd Qu.: 4.000  
 Max.   :3.000   Max.   :3.00   Max.   : 5.000   Max.   : 5.000  
```

```
[1] "Increase In Number of Voters from 2016 to 2018:"
```

'Total number of respondents: 116'

[29]:

```
# Remove the rows from this filtered list where voters did not answer either␣
 ↪angry or afraid question which is ndicated by -7
angerFearAnalysis <- angerFearAnalysis[(
                                          (angerFearAnalysis['geafraid'] !=␣
 ↪-7) &
                                          (angerFearAnalysis['geangry'] != -7)
```

```
                                       ),]


print("Increase In Number of Voters from 2016 to 2018:")
paste(nrow(angerFearAnalysis))

# View unique value of Afraid to access if there are any issues with data.
print ("Unique Values of Afraid")
unique(angerFearAnalysis$geafraid)

# View unique value of Angry to access if there are any issues with data.
print ("Unique Values of Angry")
unique(angerFearAnalysis$geangry)


anger <- hist(angerFearAnalysis$geangry, main='', col='lightblue',
              xlab = "Anger Rating 5 : Extremely and 1 Not at all",
              ylab = "Number of Voters",
              xlim=c(0,6)
         )
text(anger$mids,anger$counts,labels=anger$counts, adj=c(0.5, -0.5))
title('How Angry does the voter feel?', line=1.5, adj=0.5)

afraid <- hist(angerFearAnalysis$geafraid, main='', col='lightblue',
               xlab = "Afraid Rating 5 : Extremely and 1 Not at all",
               ylab = "Number of Voters",
               xlim=c(0,6)
         )
text(afraid$mids,afraid$counts,labels=afraid$counts, adj=c(0.5, -0.5))
title('How Afraid does the voter feel?', line=1, adj=0.5)
```

```
[1] "Increase In Number of Voters from 2016 to 2018:"
```

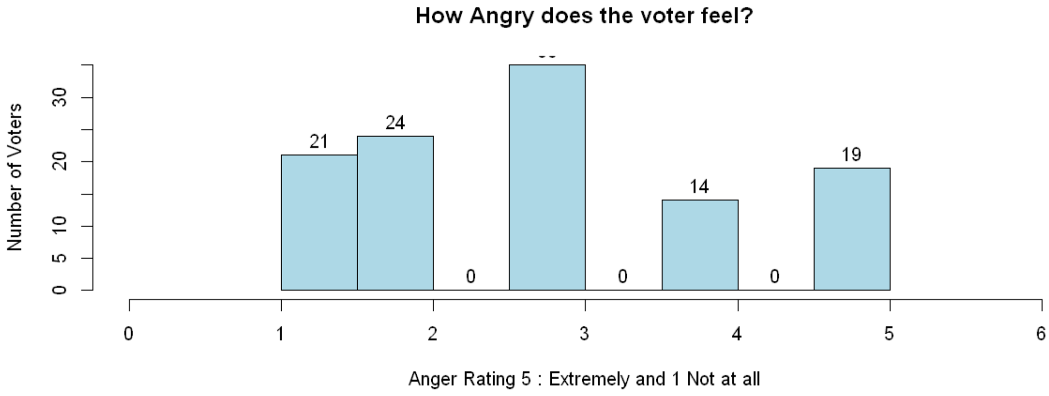
’113’

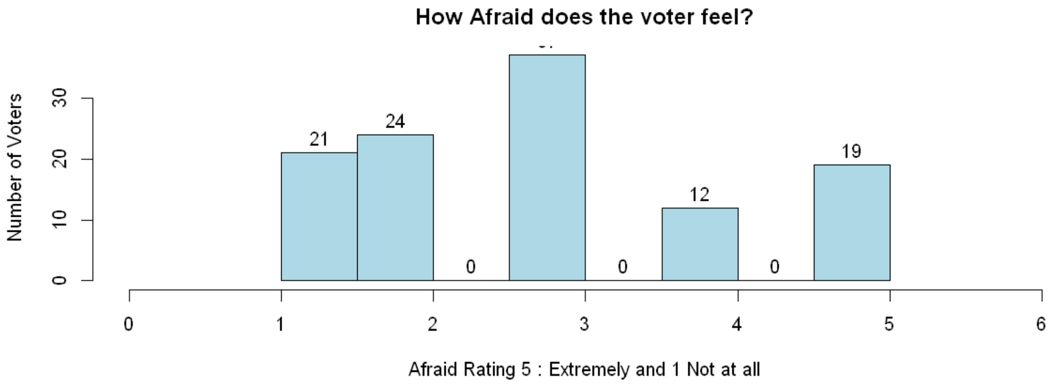
```
[1] "Unique Values of Afraid"
```

1. 2 2. 4 3. 3 4. 5 5. 1

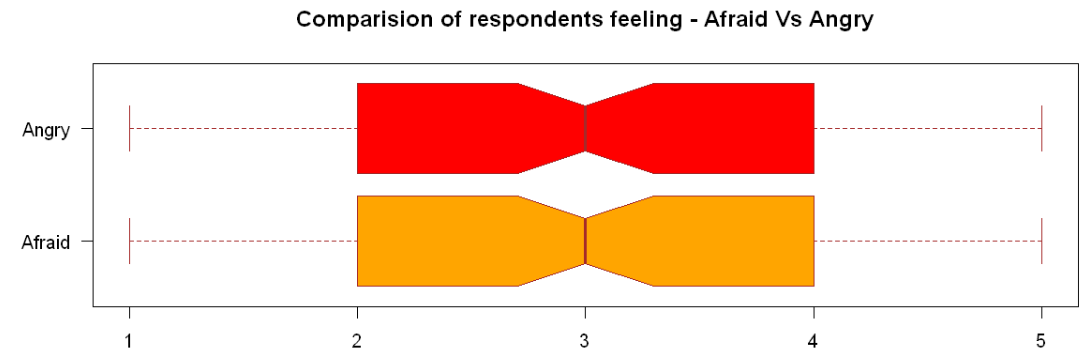
```
[1] "Unique Values of Angry"
```

1. 2 2. 4 3. 1 4. 3 5. 5



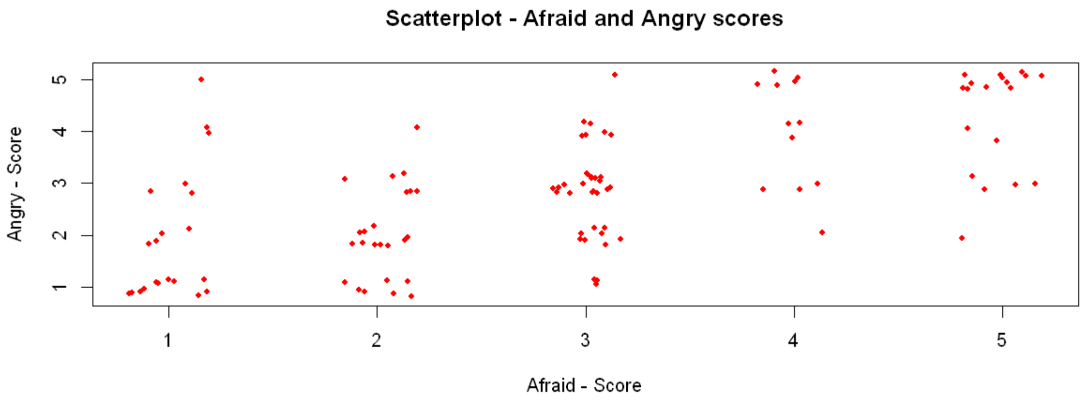


[30]:
```
# Box-plots
boxplot(angerFearAnalysis$geafraid, angerFearAnalysis$geangry,
        main = "Comparision of respondents feeling - Afraid Vs Angry", at =␣
 ↪c(1,2),
        names = c("Afraid", "Angry"), las = 1, col = c("orange","red"), border␣
 ↪= "brown", horizontal = T,
        notch = TRUE)
```



There does not seem to be much difference in median value based on above chart, though anger seem to be slightly higher.

```
[31]: # Correlation analysis
plot(jitter(angerFearAnalysis$geafraid), jitter(angerFearAnalysis$geangry),␣
 ↪main="Scatterplot - Afraid and Angry scores",
     xlab="Afraid - Score", ylab="Angry - Score", pch=20, col='red')
```



There does not seem to be correlation in the data

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

Hypothesis Testing We shall Sign Test for the **paired samples** of anger and afraid survey questions as we have ordinal variables in form of anger sentiment and fear sentiment coming from the same respondent.

Assumptions of the Sign Test are :

1. Differences in anger and fear ratings **across respondents** is independent. As per above discussion, we are assuming this condition to be satisfied
2. The ratings for fear and anger are ordinal (which is obviously true)
3. All the differences are sampled from the same population (which is what we are assuming to be true)

Null Hypothesis: H_0: There is no difference between medians of anger sentiment and fear sentiment for voters who did not vote in 2016 but did in 2018. i.e. median of diff = 0

Alternate Hypothesis: H_A: There is difference between medians of anger sentiment and fear sentiment for voters who did not vote in 2016 but did in 2018. i.e. median of diff $\neq$ 0

Significance Level = 0.05

0.19.2 Conduct your test. (5 points)

```
[32]: SIGN.test(angerFearAnalysis$geafraid, angerFearAnalysis$geangry, md=0, conf.
       ↪level = 0.99, alternative = 'two.sided')
```

```
Dependent-samples Sign-Test

data:  angerFearAnalysis$geafraid and angerFearAnalysis$geangry
S = 29, p-value = 1
alternative hypothesis: true median difference is not equal to 0
99 percent confidence interval:
 0 0
sample estimates:
median of x-y
            0

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.9859      0      0
Interpolated CI       0.9900      0      0
Upper Achieved CI     0.9919      0      0
```

0.20 Results

Since p-value (1) > significance level (0.05), we **fail to reject null hypothesis**. The results are not statistically significant. Neither anger nor fear was the reason for increase in voter turn out in 2018. Since we fail to reject null hypothesis hence practical significance effect does not have to be computed here.

0.20.1 Discussion

Embedded assumptions in this question are: 1. The fear and anger scores respondents gave represent their true feelings 2. Those feelings had causal effect on their voting behavior (very strong assumption)

0.21 Question 5

0.22 Introduction

Climate change is a politically charged issue, one of the most politically dividing in the context of US politics. On the party lines, Democrats and Republicans have very different opinions about this issue. As this is a global issue of significance, it is important to note the voter sentiment around it. In politics, nothing is permanent and if enough voter sentiment changes around the subject (one way or the other) we can expect party leadership to change the tune on this matter.

In this regard, we expect the liberal-conservative bias to show up in the public opinion. We want to go one step further and see if we can come up with something more interesting.

0.23 Political question

The key question we want to answer is that **between the Democrat and Republican voters, is there a difference between the proportion of voters with extreme opinions about climate change**? The intuition behind this question is that if a large proportion of voters are on a extreme on some policy position, there is less motivation for the political parties to differ from that extreme. Hence the difference in the extreme opinion holding voting population should give some idea about the rigidity of the political party of the matter in the immediate future.

0.24 Design and Operational Definitions

0.24.1 Opinion Tagging

`Extreme` is a loaded word and we define it differently based on the cohort. There are five choices for the question `How important is the issue of climate change to you personally?`. We define the following encoding for Democrat and republican voters:

Answer Option	Democrat encoding	Republican Encoding
Not at all important (1)	0	1
A little important (2)	0	0
Moderately important (3)	0	0
Very important (4)	0	0
Extremely important (5)	1	0

Via this encoding, we are trying to understand that within each voting cohort (D and R), what is the proportion of extreme response.

0.24.2 Party affliation tagging

We are using the combination of `pid1r` and `pid1d` columns for this. We will only consider the rows where each of these column values are either 1 or 2 (with 1 being democrat and 2 being republican). We will disregard all other values.

CAUTION: In this way, we are disregarding every respondent who has declined to answer this question, independent and supports other party. The last are ok because we explicitely are interested in Democrat and Republican split but the first is concerning as this means we might be missing some undecided candidate. (Analysis show that we wil be disregarding only 2 people thsi was so that is ok)

[33]:

```
party <- A[,c('pid1d', 'pid1r','warmyou')]

# Respondents that declined to answer the party affliation question
print("Rows of data with missing score entry")
party[(party$pid1d == -7)|(party$pid1r == -7),]
party <- party[!((party$pid1d == -7)|(party$pid1r == -7)),]

# Summary of the filtered resulting dataset
print("Overall Summary of the filtered data")
summary(party)

# Tagging the part affliations
party$party <- ifelse(party$pid1d == 2 | party$pid1r == 2, 'REPUBLICAN',␣
 ↪ifelse(party$pid1d == 1 | party$pid1r == 1,
                                                                          ␣
 ↪'DEMOCRAT', 'O'))

# Removing all the data rows representing votes who are neither democrat nor␣
 ↪republicans
party <- party[party$party != 'O',]

# Final dataset size
paste("Number of records (total) :", nrow(party))
```

```
[1] "Rows of data with missing score entry"
```

	pid1d	pid1r	warmyou
497	-1	-7	5
596	-7	-1	1

```
[1] "Overall Summary of the filtered data"
```

```
     pid1d              pid1r              warmyou
 Min.   :-1.0000   Min.   :-1.0000   Min.   :1.000
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:2.000
 Median :-1.0000   Median :-1.0000   Median :3.000
 Mean   : 0.4155   Mean   : 0.4644   Mean   :3.142
 3rd Qu.: 2.0000   3rd Qu.: 2.0000   3rd Qu.:4.000
 Max.   : 4.0000   Max.   : 4.0000   Max.   :5.000
```

'Number of records (total) : 1466'

0.25 Potential issues

1. We are using self declaration by respondents on the face value. We also have the data available around 2016 voting and it might be more relevant to use that as the source of party affliation. Our core assumption is that the self reported affliations are ok
2. Voter opinion should be an imporant factor in deciding the position a political party take but there can be multiple reasons for a delay between change in public opinion and change in party leadership position. This analysis doesn't account for that
3. This analysis takes into cosideration that high proportion if voters of respective parties hold extreme opinions about climate change (whcih came from the data). If in coming time that changes (i.e. less percent of voters have extreme opinion), the basis of this study will have to change.
4. We are **assuming** that party position on global warming is a factor which affects voting behavior. If it doesn't then the proportion of people with extreme opinions on thsi subject will not matter at all to party leadership.

0.26 Exploratory analysis

As per our discussion above, both democrat and republicans have different connotations for extreme opinions. Following the schema discussed above, we will transfrom the `warmyou` variable into a binary vector. There are total 1466 rows of data in our consideration set (609-Republican| 857-Democrat). Overall, the data quality after appropriate tagging is good and no Nulls are present in the data. We have already extracted out the respondents that didn't provide information about their party affliation. The data ranges for all three variables is as expected.

```
[34]: # tagging the extreme results appropriately in line with the party lines

      party$extreme_opinion <- ifelse((party$party ==␣
       ↪'REPUBLICAN')&(party$warmyou==1),
                                      1, ifelse((party$party ==␣
       ↪'DEMOCRAT')&(party$warmyou==5), 1, 0))
```

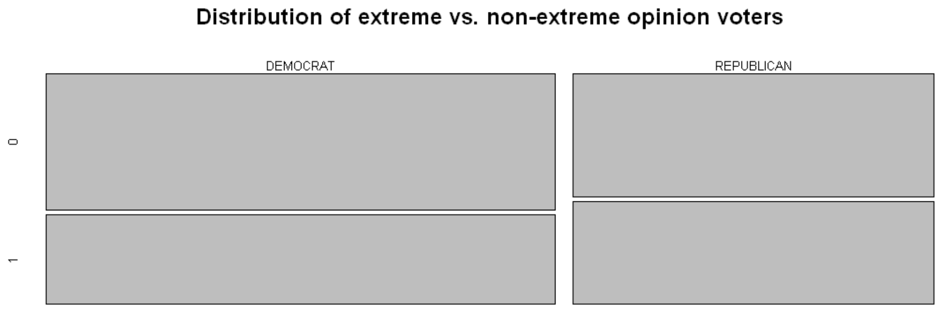
0.27 Visualization

```
[35]: print('The distribution across classes')
      table(party$party, party$extreme_opinion)

      plot(table(party$party, party$extreme_opinion), main = 'Distribution of extreme␣
       ↪vs. non-extreme opinion voters')
```

```
[1] "The distribution across classes"

              0   1
  DEMOCRAT   518 339
  REPUBLICAN 332 277
```



0.28 Hypothesis - Two sided

H_0: Proportion of voters with extreme responses between democrat and republican voters is the same

H_A: Proportion of voters with extreme responses between democrat and republican voters is NOT the same

0.29 Significance level

We want to perfrom the test at 95% significance level (i.e. Expected FPR of 5%)

0.30 Appropriate test selection and conditions

2-proportion z-test seems to be appropriate for this case. The key conditions for this test are:

1. We have a simple random sample - This condition is not met exactly but for all the analysis before this section we are assuming this to be true. Hence we can assume this condition to be met here as well
2. Independent sample - Sample should be independent. As mentioned in previous analysis, this might be a strong assumption but we are anyways considering it to be true
3. Sample size restrictions - Each sample has > 10 respondents and population is atleast 20 times bigger than the sample

0.31 Conduction of the test

```
[36]: prop.test(x = c(sum(party[(party$party ==␣
       ↪'DEMOCRAT')&(party$extreme_opinion==1), 'extreme_opinion']),
                     sum(party[(party$party ==␣
       ↪'REPUBLICAN')&(party$extreme_opinion==1), 'extreme_opinion'])),
               n=c(nrow(party[party$party == 'DEMOCRAT', ]), nrow(party[party$party␣
       ↪== 'REPUBLICAN', ])),
          conf.level = 0.95,alternative = 'two.sided', correct=F)
```

```
2-sample test for equality of proportions without continuity
correction

data:  c(sum(party[(party$party == "DEMOCRAT") & (party$extreme_opinion ==  out of c(nrow(part
X-squared = 5.1348, df = 1, p-value = 0.02345
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.11061828 -0.00793788
sample estimates:
   prop 1    prop 2
0.3955659 0.4548440
```

We tested the two sided hypothesis that the proportion of extreme opinion voters in Demotratic and Republican party are same. Against the alternate hypothesis that they are different. From the 2-sample z-test we see that

1. The p-value is 0.02345
2. 95% confidence interval for the difference in proportions is [-0.1106, -0.0079] (which doesn't contains 0)

At 95% confidence level, we have sufficient evidence coming from the data to **reject the Null Hypothesis**.

0.32 One sided Hypothesis test

We can now test the hypothesis that among Republican voters, prevelance of extreme opinion is higher than among Democrat voters.

0.33 Hypothesis - One sided

H_0: Proportion of voters with extreme responses between democrat and republican voters is the same

H_A: Proportion of voters with extreme responses among democrat voters is less than republican voters

[37]:

```
prop.test(x = c(sum(party[(party$party ==␣
↪'DEMOCRAT')&(party$extreme_opinion==1), 'extreme_opinion']),
                sum(party[(party$party ==␣
↪'REPUBLICAN')&(party$extreme_opinion==1), 'extreme_opinion'])),
          n=c(nrow(party[party$party == 'DEMOCRAT', ]), nrow(party[party$party␣
↪== 'REPUBLICAN', ])),
          conf.level = 0.95,alternative = 'less', correct=F)
```

```
2-sample test for equality of proportions without continuity
correction
```

```
data:  c(sum(party[(party$party == "DEMOCRAT") & (party$extreme_opinion ==  out of c(nrow(party
X-squared = 5.1348, df = 1, p-value = 0.01173
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.01619203
sample estimates:
   prop 1    prop 2
0.3955659 0.4548440
```

0.34 Results

At 95% confidence level, we have sufficient evidence to reject the null hypothesis in favor of one sided alternative hypothesis.

0.35 Effect size

The difference of proportion (which stands at -0.0592) in the sample between democrat voters with extreme opinions and republicans can be a good proxy for the effect size. Converting this to percentages, this says that the affinity to extreme with respect to climate change is ~6% higher among republican voters than democrat voters.

If we want to use a dimentionless measure of effect size, we can use Cohen's H. More details about this can be found from https://en.wikipedia.org/wiki/Cohen%27s_h. The effect size table showcases:

Cohen's H	Interpretation
h = 0.20	Small effect size
h = 0.50	Medium effect size
h = 0.80	Large effect size

From the below analysis, we see that **the effect size if small**.

[38]:

```
cohens_h = function( prop_1, prop_2, n1, n2, ci = 0.95 ){
    x1 = asin(sign(prop_1) * sqrt(abs(prop_1)))
    x2 = asin(sign(prop_2) * sqrt(abs(prop_2)))
    es = x1 - x2
    se = sqrt(0.25 * (1 / n1 + 1 / n2 ))
    ci_diff = qnorm(1 - (1-ci) / 2) * se
    return( c( h = es*2, h_low = (es-ci_diff)*2, h_upp = (es+ci_diff)*2 ) )
}

cohens_h(prop_1=0.4548440, prop_2=0.3955659, n1=332+277, n2=518+339, ci = 0.95)
```

h 0.119982269428239 **h_low** 0.0161060777970168 **h_upp** 0.223858461059461

0.36 Conclusion

We tried to understand the opinions of US voters vis-a-vis party opinions along the topic of climate change. What we see is that a huge proportion of voters on both sides cling to the extremes (democrats to the importance of this topic and Republicans to the unimportance). Also there is a **statistically significant difference** between the extreme opinion voters across the party lines although the effect size is small.

There are significantly more extreme opinion holding voters on republican side than democrat (~ 6%). In the long run, this analysis give rise to the idea that atleast in the short run, there is less chance for Republican party line to change its narrative around Climate change (in absence of public support).

More work need to be done to understand this further in terms of respondant age and other demographic factors.