Lab 3: Hypothesis Tests about the Mean.

w203: Statistics for Data Science

```
library(car)
library(effsize)
```

```
## Warning: package 'effsize' was built under R version 3.3.3
```

```
S = read.csv("ANES_2012_sel.csv")
```

Assignment

1. Did voters become more liberal or more conservative during the 2012 election?

```
## First we evaluate the libpo_self and libpre_self columns
(libcpo_tb= table(S$libcpo_self, as.numeric(S$libcpo_self)))
```

```
##
##                                                                  1    2
##   -2. Haven't thought much {do not probe}                      410    0
##   -6. Not asked, unit nonresponse (no post-election interview)   0  252
##   -7. Deleted due to partial (post-election) interview           0    0
##   -8. Don't know                                                 0    0
##   -9. Refused                                                    0    0
##   1. Extremely liberal                                           0    0
##   2. Liberal                                                     0    0
##   3. Slightly liberal                                            0    0
##   4. Moderate; middle of the road                                0    0
##   5. Slightly conservative                                       0    0
##   6. Conservative                                                0    0
##   7. Extremely conservative                                      0    0
##
##                                                                  3    4
##   -2. Haven't thought much {do not probe}                        0    0
##   -6. Not asked, unit nonresponse (no post-election interview)   0    0
##   -7. Deleted due to partial (post-election) interview         152    0
##   -8. Don't know                                                 0   23
##   -9. Refused                                                    0    0
##   1. Extremely liberal                                           0    0
##   2. Liberal                                                     0    0
##   3. Slightly liberal                                            0    0
##   4. Moderate; middle of the road                                0    0
##   5. Slightly conservative                                       0    0
##   6. Conservative                                                0    0
##   7. Extremely conservative                                      0    0
##
##                                                                  5    6
##   -2. Haven't thought much {do not probe}                        0    0
##   -6. Not asked, unit nonresponse (no post-election interview)   0    0
##   -7. Deleted due to partial (post-election) interview           0    0
##   -8. Don't know                                                 0    0
##   -9. Refused                                                   36    0
```

```
##   1. Extremely liberal                                              0  166
##   2. Liberal                                                        0    0
##   3. Slightly liberal                                               0    0
##   4. Moderate; middle of the road                                   0    0
##   5. Slightly conservative                                          0    0
##   6. Conservative                                                   0    0
##   7. Extremely conservative                                         0    0
##
##                                                                     7    8
##   -2. Haven't thought much {do not probe}                           0    0
##   -6. Not asked, unit nonresponse (no post-election interview)      0    0
##   -7. Deleted due to partial (post-election) interview              0    0
##   -8. Don't know                                                    0    0
##   -9. Refused                                                       0    0
##   1. Extremely liberal                                              0    0
##   2. Liberal                                                      646    0
##   3. Slightly liberal                                               0  639
##   4. Moderate; middle of the road                                   0    0
##   5. Slightly conservative                                          0    0
##   6. Conservative                                                   0    0
##   7. Extremely conservative                                         0    0
##
##                                                                     9   10
##   -2. Haven't thought much {do not probe}                           0    0
##   -6. Not asked, unit nonresponse (no post-election interview)      0    0
##   -7. Deleted due to partial (post-election) interview              0    0
##   -8. Don't know                                                    0    0
##   -9. Refused                                                       0    0
##   1. Extremely liberal                                              0    0
##   2. Liberal                                                        0    0
##   3. Slightly liberal                                               0    0
##   4. Moderate; middle of the road                                1756    0
##   5. Slightly conservative                                          0  671
##   6. Conservative                                                   0    0
##   7. Extremely conservative                                         0    0
##
##                                                                    11   12
##   -2. Haven't thought much {do not probe}                           0    0
##   -6. Not asked, unit nonresponse (no post-election interview)      0    0
##   -7. Deleted due to partial (post-election) interview              0    0
##   -8. Don't know                                                    0    0
##   -9. Refused                                                       0    0
##   1. Extremely liberal                                              0    0
##   2. Liberal                                                        0    0
##   3. Slightly liberal                                               0    0
##   4. Moderate; middle of the road                                   0    0
##   5. Slightly conservative                                          0    0
##   6. Conservative                                                 975    0
##   7. Extremely conservative                                         0  188
```

```
(libcpre_tb= table(S$libcpre_self, as.numeric(S$libcpre_self)))
```

```
##
##                                      1    2    3    4    5    6    7
##   -2. Haven't thought much about this  556    0    0    0    0    0    0
```

```
##   -8. Don't know                          0   26    0    0    0    0    0
##   -9. Refused                             0    0   32    0    0    0    0
##   1. Extremely liberal                    0    0    0  195    0    0    0
##   2. Liberal                              0    0    0    0  638    0    0
##   3. Slightly liberal                     0    0    0    0    0  641    0
##   4. Moderate; middle of the road         0    0    0    0    0    0 1828
##   5. Slightly conservative                0    0    0    0    0    0    0
##   6. Conservative                         0    0    0    0    0    0    0
##   7. Extremely conservative               0    0    0    0    0    0    0
##
##                                           8    9   10
##   -2. Haven't thought much about this     0    0    0
##   -8. Don't know                          0    0    0
##   -9. Refused                             0    0    0
##   1. Extremely liberal                    0    0    0
##   2. Liberal                              0    0    0
##   3. Slightly liberal                     0    0    0
##   4. Moderate; middle of the road         0    0    0
##   5. Slightly conservative              789    0    0
##   6. Conservative                         0 1001    0
##   7. Extremely conservative               0    0  208
```

Here we see that the numbering for the same levels in libcpre_self and libcpo_self are different. Also there are two additional level in libcpo_self i.e. -6. Not asked, unit nonresponse (no post-election interview) and "Deleted due to partial (post-election) election survey. In order to get a apples to apples comparision we will remove these data points from our analysis

```
post_sur_del= c("-6. Not asked, unit nonresponse (no post-election election survey)","-7. Deleted due t
S_narrow = subset(S,!(S$libcpo_self %in% post_sur_del ))
```

Now we examine if the people who did not answer the question in the pre-election interview (i.e. "-2. Haven't thought much about this","-8. Don't know","-9. Refused") changed there answers in the post-election interview otherwise we can remove them from the analysis.

```
non_answers = c( "-2. Haven't thought much about this","-8. Don't know","-9. Refused")

answers = c( "1. Extremely liberal","2. Liberal",  "3. Slightly liberal",  "4. Moderate; middle of the

S_changed =subset(S,libcpre_self %in% non_answers)
nrow(S_changed)
```

```
## [1] 614
```

```
nrow((subset(S_changed,libcpo_self %in% answers)))
```

```
## [1] 283
```

Here we see there are 614 responses which are non-answers in the pre-election interview, out of which 283 responses changed during the post-election interview. We will keep all these data point incudling the above 283 data points, but exculde data points which are non-answers in pre-election interview as well as post-election interview

```
S_narrow =subset(S_narrow,!((libcpre_self %in% non_answers ) & (libcpo_self %in% non_answers )))
```

Let us ennumerate the libpre_slef and libpo_self such that level from Extremely liberal to Extremely conservative are same for both the variables

```
S_narrow$libcpo_self_sc = (as.numeric(S_narrow$libcpo_self)-2)
S_narrow$libcpre_self_sc = as.numeric(S_narrow$libcpre_self)
S$libcpo_self_sc = (as.numeric(S$libcpo_self)-2)
S$libcpre_self_sc = as.numeric(S$libcpre_self)

libcpo_tb= table(S_narrow$libcpo_self, as.numeric(S_narrow$libcpo_self)-2)
libcpre_tb= table(S_narrow$libcpre_self, as.numeric(S_narrow$libcpre_self))

#Let us consider the non-answers (i.e "-2. Haven't thought much about this","-8. Don't know",
#"-9. Refused") as mean. This will help us to know if the non-answer responses shifted right
#(more liberal/ i.e <7) or left( more conservative / i.e > 7)
S_narrow$libcpre_self_sc = ifelse(S_narrow$libcpre_self_sc < 4 , 7, S_narrow$libcpre_self_sc)
S_narrow$libcpo_self_sc = ifelse(S_narrow$libcpo_self_sc < 4 , 7, S_narrow$libcpo_self_sc)

#Since the libcpre_self and the libcpo_self are ordinal varibles we cannot run the
# parameteric test, we will have to use the non-parameteric test.
# since we need to measure the different between the pre-election interview and post-election
#interview we will use the signed Rank
# Our null hypothesis will be that there is not change between the survery reponses before
#and after the election.

wilcox.test(S_narrow$libcpre_self_sc,S_narrow$libcpo_self_sc,paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  S_narrow$libcpre_self_sc and S_narrow$libcpo_self_sc
## V = 1095800, p-value = 0.1064
## alternative hypothesis: true location shift is not equal to 0
```

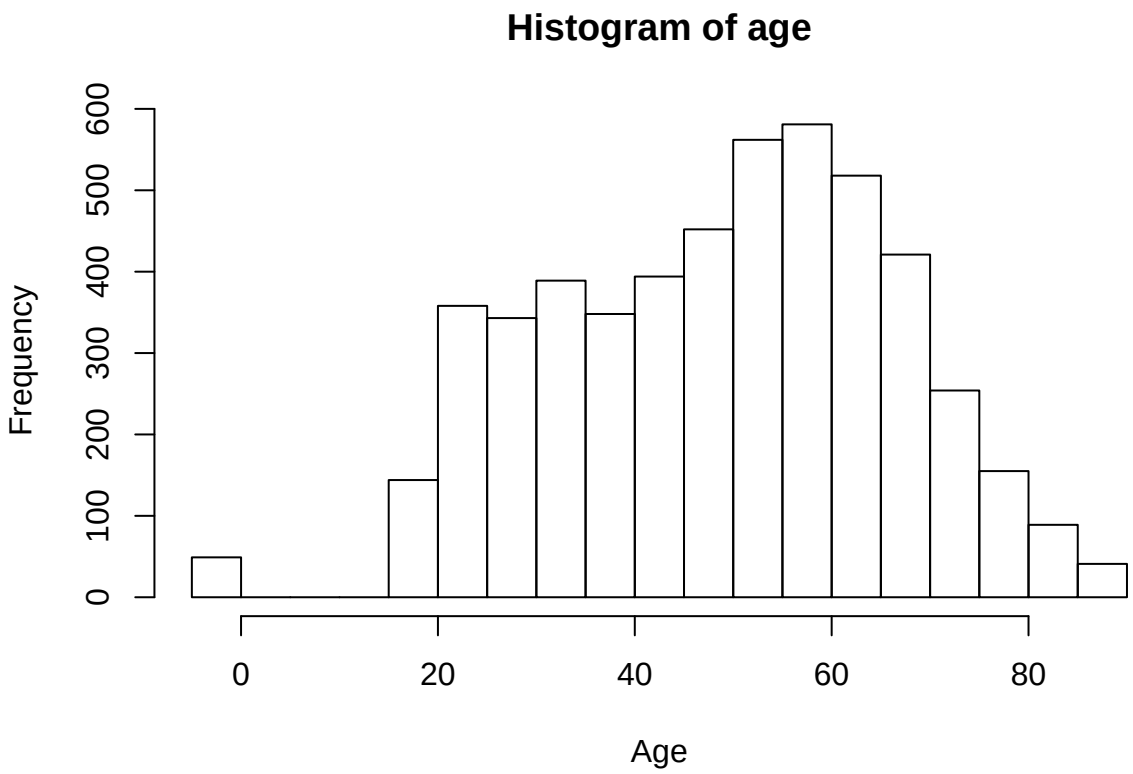
```
#We see that the p-value is 0.1068, so we fail to reject the null hypothesis. This means
#that the survery reponses did not change significantly before and after the election
```

2. Were Republican voters (examine variable pid_x) older or younger (variable dem_age_r_x), on the average, than Democratic voters in 2012?

```
## Let us examine the pid_x and dem_age_r_x variables
summary(S$pid_x)
```

```
##                   -2. Missing                1. Strong Democrat
##                            24                              1485
##  2. Not very strong Democract          3. Independent-Democrat
##                           871                               747
##                4. Independent       5. Independent-Republican
##                           792                               610
## 6. Not very strong Republican            7. Strong Republican
##                           623                               762
```

```
## we see there are 24 missing values. let us remove them since they so not help our analysis
## Also Since we have to see if age difference between democratic and republican votes we
#can set aside the independent voters from the analysis

S_new  = subset(S,!(pid_x %in% c("4. Independent")))
S_new= S_new[S_new$pid_x!="-2. Missing",]

hist(S_new$dem_age_r_x, xlab = "Age" ,main = "Histogram of age")
```




```
summary(S_new$dem_age_r_x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -2.00   35.00   51.00   49.32   62.00   90.00
```

```
## here the distribution is almost normal. So we can apply the central limit therom here.
## We will be performing a parameteric t-test.
## We also see that there is data points with age -2. This is not possible and represents
## a corrupt data point.
summary(S_new[S_new$dem_age_r_x == -2,1:8])
```

```
##        X                                profile_educ
##  Min.   :  11   -1. Inapplicable              :49
##  1st Qu.: 641   1. Less than high school      : 0
##  Median :1114   2. High school                : 0
##  Mean   :1040   3. Some college               : 0
##  3rd Qu.:1453   4. Bachelor's degree or higher: 0
##  Max.   :2045
##
##           profile_gender
##  -1. Inapplicable:49
##  1. Male         : 0
##  2. Female       : 0
##
##
##
##
```

```
##                                                    profile_homeown
##  -1. Inapplicable                                              :49
##  1. Owned or being bought by you or someone in your household: 0
##  2. Rented for cash                                            : 0
##  3. Occupied without payment of cash rent                      : 0
##
##
##
##                profile_hhincome               profile_marital  dem_age_r_x
##  -1. Inapplicable      :49      -1. Inapplicable      :49     Min.   :-2
##  1. Less than $5,000   : 0      1. Married            : 0     1st Qu.:-2
##  10. $35,000 to $39,999: 0      2. Widowed            : 0     Median :-2
##  11. $40,000 to $49,999: 0      3. Divorced           : 0     Mean   :-2
##  12. $50,000 to $59,999: 0      4. Separated          : 0     3rd Qu.:-2
##  13. $60,000 to $74,999: 0      5. Never married      : 0     Max.   :-2
##  (Other)               : 0      6. Living with partner: 0
##               profile_region9
##  -1. Inapplicable     :49
##  1. New england       : 0
##  2. Mid-atlantic      : 0
##  3. East-north central: 0
##  4. West-north central: 0
##  5. South atlantic    : 0
##  (Other)              : 0
```

```
## from running the summary on the firt 8 columns it looks like around 49 survey
## participants did not fill the personal information but filled in the
##questionaire associated with the survey.

## For this for the analysis of the  age we will be replacing -2 with the  mean age of 49.
S_new$dem_age_r_x = ifelse(S_new$dem_age_r_x ==-2, 49,S_new$dem_age_r_x)

## we will assign a new variable party which identifies whether the participaint is
## democratic or republican
vote_class_demo = c("1. Strong Democrat","2. Not very strong Democract","3. Independent-Democrat")

S_new$party = factor(ifelse(S_new$pid_x %in% vote_class_demo, "Democrat","Republican"))
table(S_new$party)
```

```
##
##   Democrat Republican
##       3103       1995
```

```
# We will run a levelne test to check if the meet the assumption of Homogeneity of Variance
leveneTest(S_new$dem_age_r_x, S_new$party,center = mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##         Df F value Pr(>F)
## group    1  0.0142 0.9051
##       5096
```

```
# We get the value of .9051 with is not Statistically significate suggesting that both the
# distributions are almost similar variance

# Since the dem_age_r_x is normal and the variance are similar we will use the two tailed
# parameteric t.test The null hypothesis here is that that there is no difference between
```

```r
# the age of republicans and democrats
t.test(dem_age_r_x ~ party ,data=S_new ,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  dem_age_r_x by party
## t = -5.1817, df = 5096, p-value = 2.284e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.411523 -1.538673
## sample estimates:
##   mean in group Democrat mean in group Republican
##                 48.83919                 51.31429
```

```r
# Here we see that p values is 2.42e-07. So the mean age of republicans and democrats is
#not the same and is statitically significant We see that the mean age of Repulicans(=51.31)
#is higher than the democrats(=48.83)

# Now let is check for pratical significance for the same
# We will use Cohen's d for the determining the pratical signifiance
cohen.d(dem_age_r_x ~ party ,data=S_new)
```

```
##
## Cohen's d
##
## d estimate: -0.148699 (negligible)
## 95 percent confidence interval:
##        inf        sup
## -0.2050316 -0.0923664
```

```r
# Here we see that the means of two groups are just 0.14 SD apart. So even though age is
#statitically significant it is not pratically significant

cor(S_new$dem_age_r_x,as.numeric(S_new$party))
```

```
## [1] 0.07239602
```

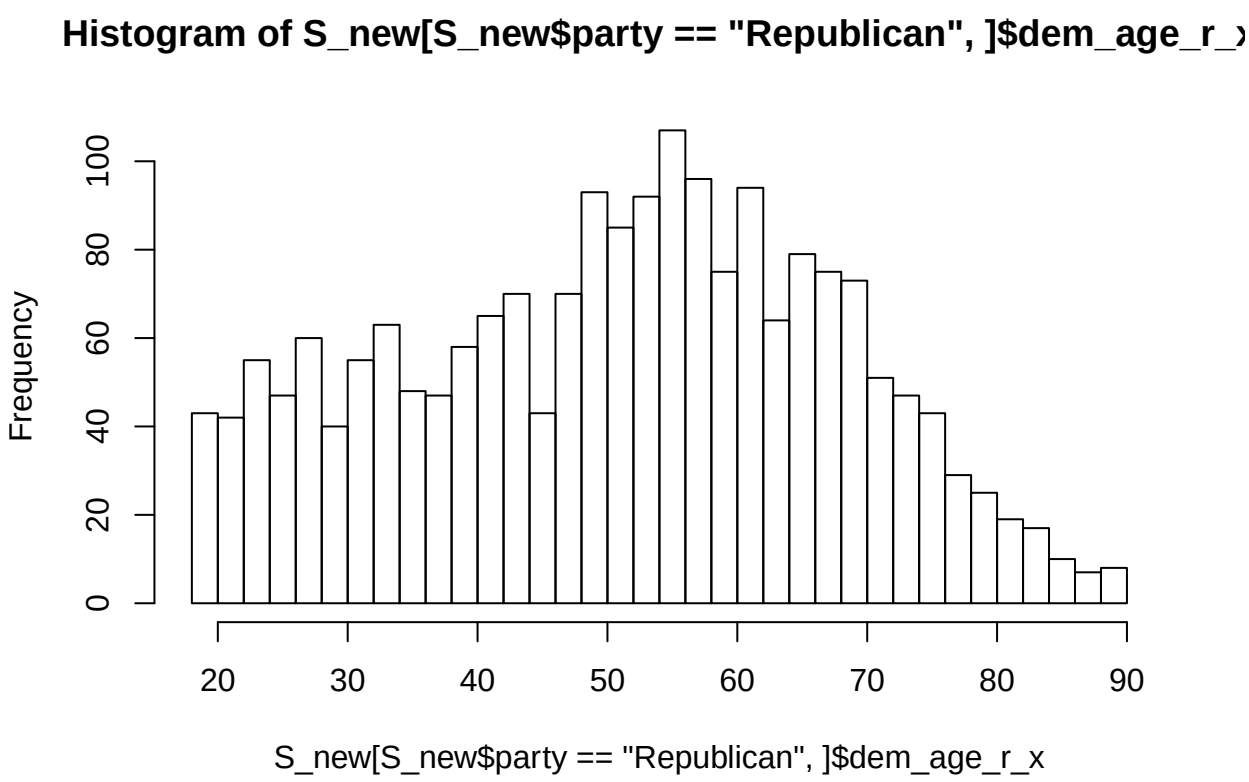
```r
# Here as well we see that the correlation between the paryt(groups) and the
# age(dem_age_r_x) is low at 0.07.
```

3. Were Republican voters older than 51, on the average in 2012? We need to test if the Republican voters were old than 51 Let us evaluvate the republican voters

```r
summary(S_new[S_new$party=='Republican',]$dem_age_r_x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   38.00   53.00   51.31   64.00   90.00
```

```r
hist(S_new[S_new$party=='Republican',]$dem_age_r_x, breaks = 50)
```



From standard mean we can see that the mean is equal to 51.31 We can see that the data is almost normal, so we can do parameteric t.test.

Since the dem_age_r_x is normal we can run a two tailed parameteric t.test with mean =51. We can do this because the we need to see if the difference between avg age of repulican is different from 51

The null hypothesis is that that mean age of repulican voters is 51

```
t.test(S_new[S_new$party=='Republican',]$dem_age_r_x , mu=51, var.equal=TRUE)
```

```
##
##  One Sample t-test
##
## data:  S_new[S_new$party == "Republican", ]$dem_age_r_x
## t = 0.83909, df = 1994, p-value = 0.4015
## alternative hypothesis: true mean is not equal to 51
## 95 percent confidence interval:
##  50.57972 52.04885
## sample estimates:
## mean of x
##  51.31429
```

```
t.test(S_new[S_new$party=='Republican',]$dem_age_r_x ,alternative = "greater" , mu=51, var.equal=TRUE)
```

```
##
##  One Sample t-test
##
## data:  S_new[S_new$party == "Republican", ]$dem_age_r_x
## t = 0.83909, df = 1994, p-value = 0.2008
```

```
## alternative hypothesis: true mean is greater than 51
## 95 percent confidence interval:
##  50.69791      Inf
## sample estimates:
## mean of x
##  51.31429
```

Here we see that we fail to reject the null hypothesis i.e. avg age of Republicans is 51

4. Were Republican voters more likely to shift their political preferences right or left (more conservative or more liberal), compared to Democratic voters during the 2012 election?

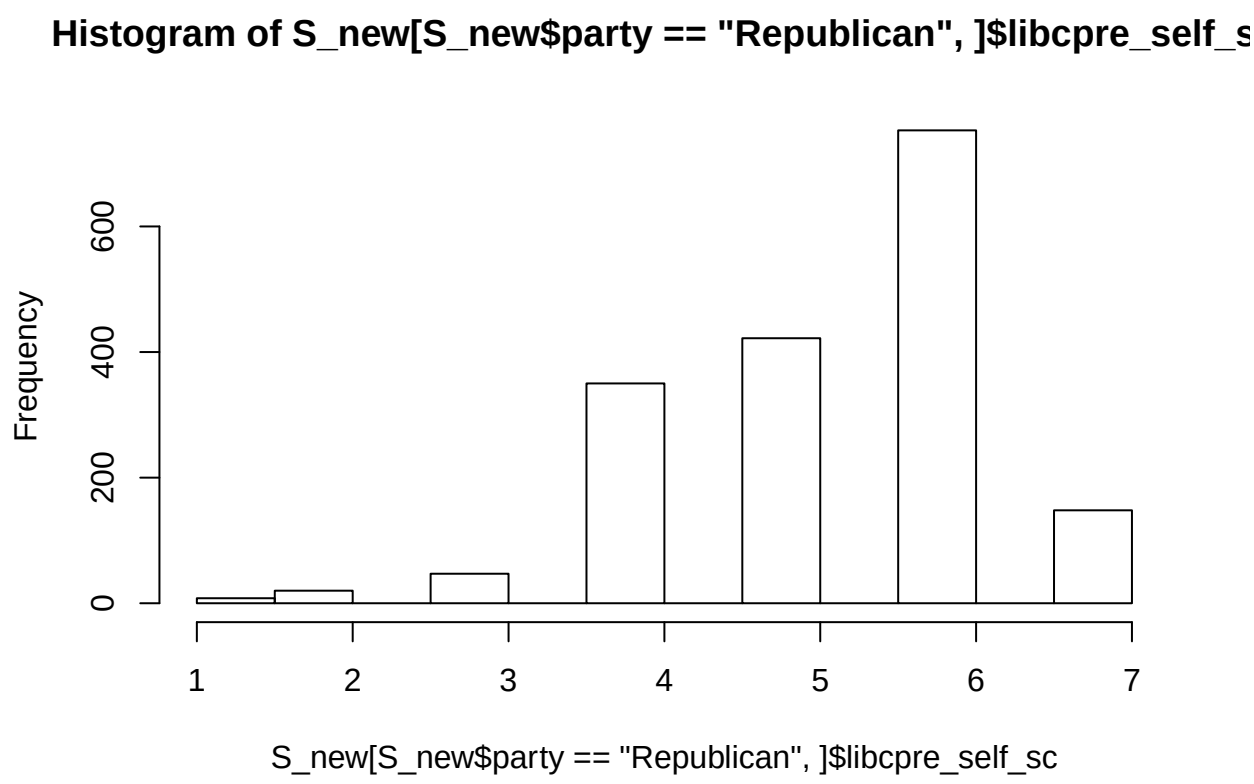
```
S_new  = subset(S,!(pid_x %in% c("4. Independent")))
S_new= S_new[S_new$pid_x!="-2. Missing",]
## Let us remove all the non-answers from the data
S_new = subset(S_new,libcpre_self %in% answers)
S_new = subset(S_new,libcpo_self %in% answers)
## we will assign a new variable party which identifies whether the participaint is
## democratic or republican
S_new$party = factor(ifelse(S_new$pid_x %in% vote_class_demo, "Democrat","Republican"))
## Let us reduce the nubmer of factors and created new varaibles libcpre_self_sc and
## libcpo_self_sc which enumerates them
S_new$libcpo_self =factor(S_new$libcpo_self)
S_new$libcpre_self = factor(S_new$libcpre_self)
S_new$libcpre_self_sc = as.numeric(S_new$libcpre_self)
S_new$libcpo_self_sc = as.numeric(S_new$libcpo_self)

## Let us examine the new variables libcpo_self for repulican vs Democrats
summary(S_new[S_new$party=='Republican',]$libcpre_self_sc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.000   6.000   5.293   6.000   7.000
```

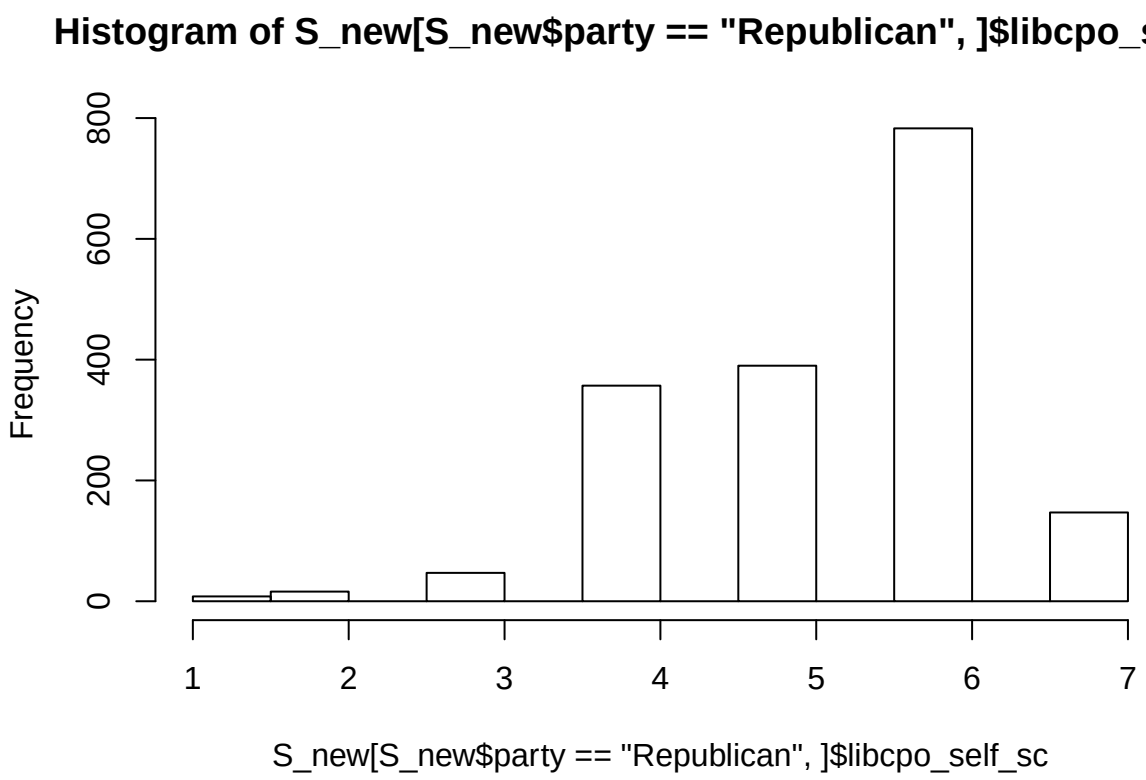
```
summary(S_new[S_new$party=='Republican',]$libcpo_self_sc )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.000   6.000   5.312   6.000   7.000
```

```
hist(S_new[S_new$party=='Republican',]$libcpre_self_sc ,breaks = 13)
```




```
hist(S_new[S_new$party=='Republican',]$libcpo_self_sc,breaks = 16 )
```




```
table(S_new$libcpo_self)
```

```
##
##            1. Extremely liberal                       2. Liberal
##                             142                              595
##             3. Slightly liberal 4. Moderate; middle of the road
##                             556                             1247
##        5. Slightly conservative                 6. Conservative
##                             562                              887
##       7. Extremely conservative
##                             169
```

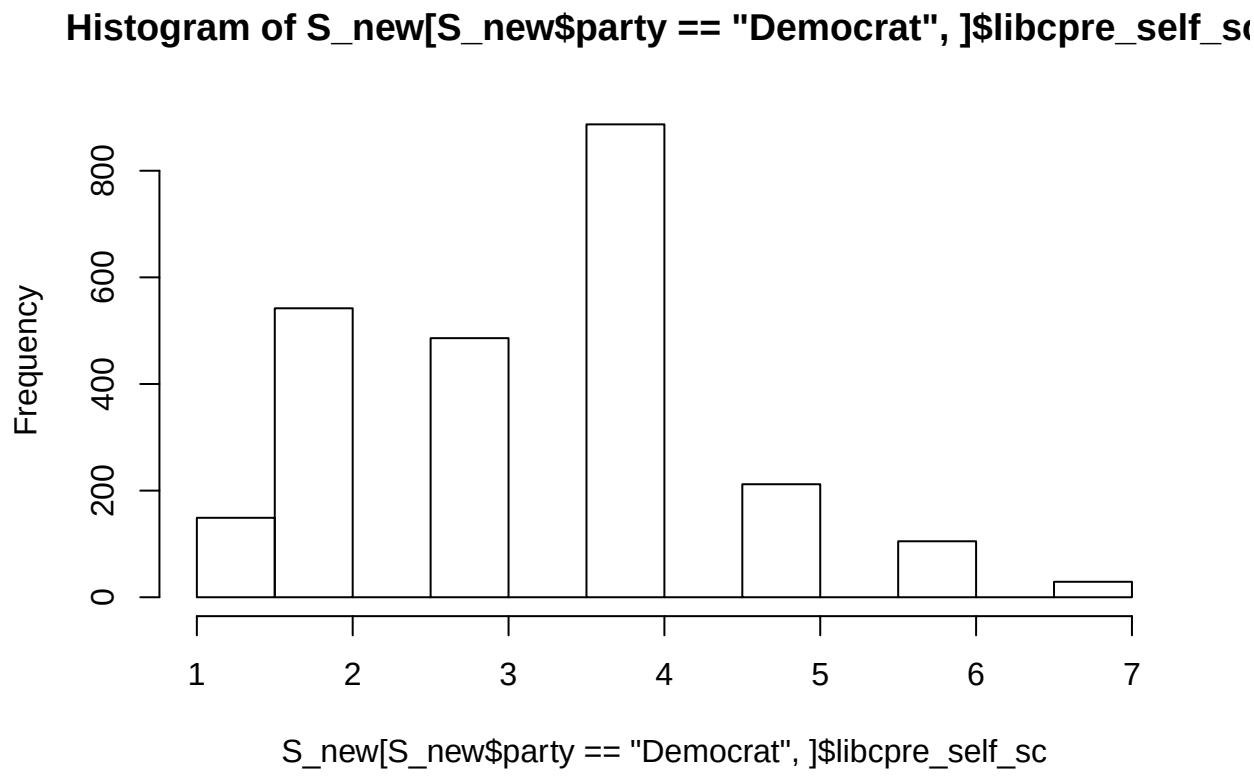
```
## similary for democrats

summary(S_new[S_new$party=='Democrat',]$libcpre_self_sc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   4.000   3.374   4.000   7.000
```

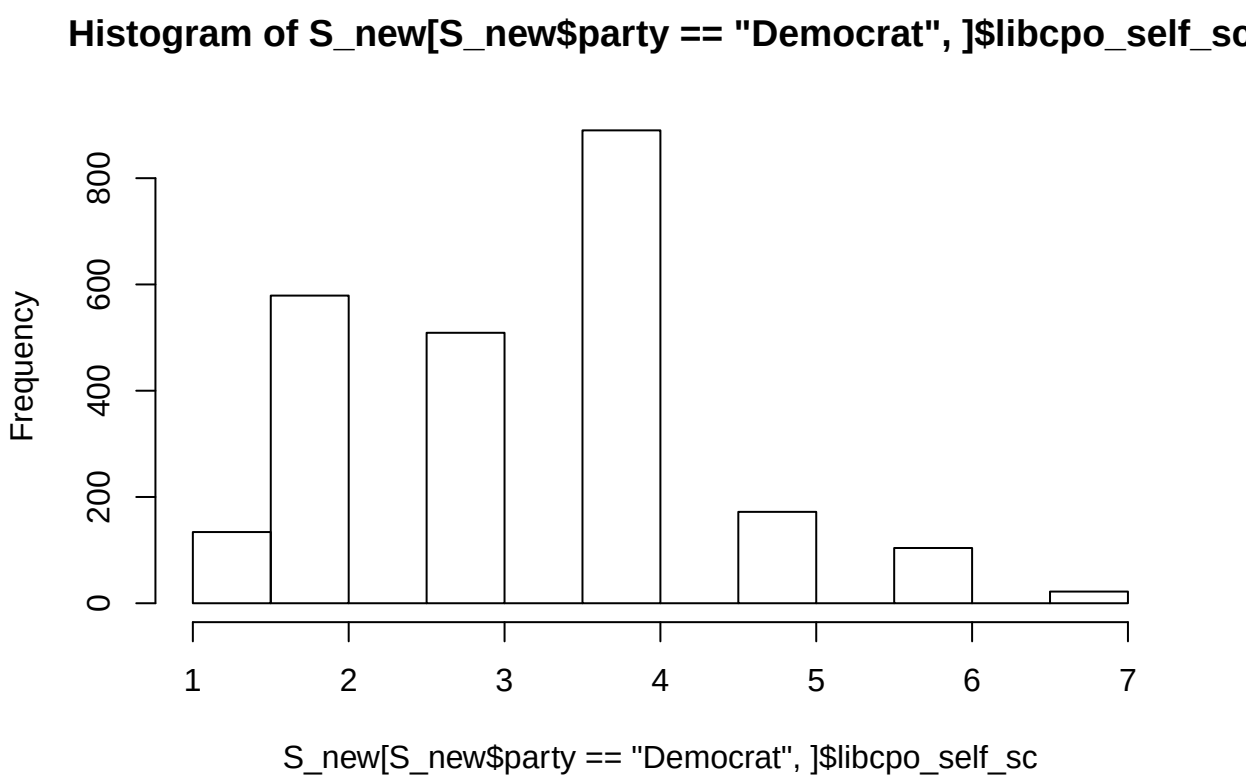
```
summary(S_new[S_new$party=='Democrat',]$libcpo_self_sc )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.327   4.000   7.000
```

```
hist(S_new[S_new$party=='Democrat',]$libcpre_self_sc ,breaks = 13)
```




```
hist(S_new[S_new$party=='Democrat',]$libcpo_self_sc ,breaks = 16 )
```




```
nrow(S_new[S_new$party=='Republican',])
```

```
## [1] 1748
```

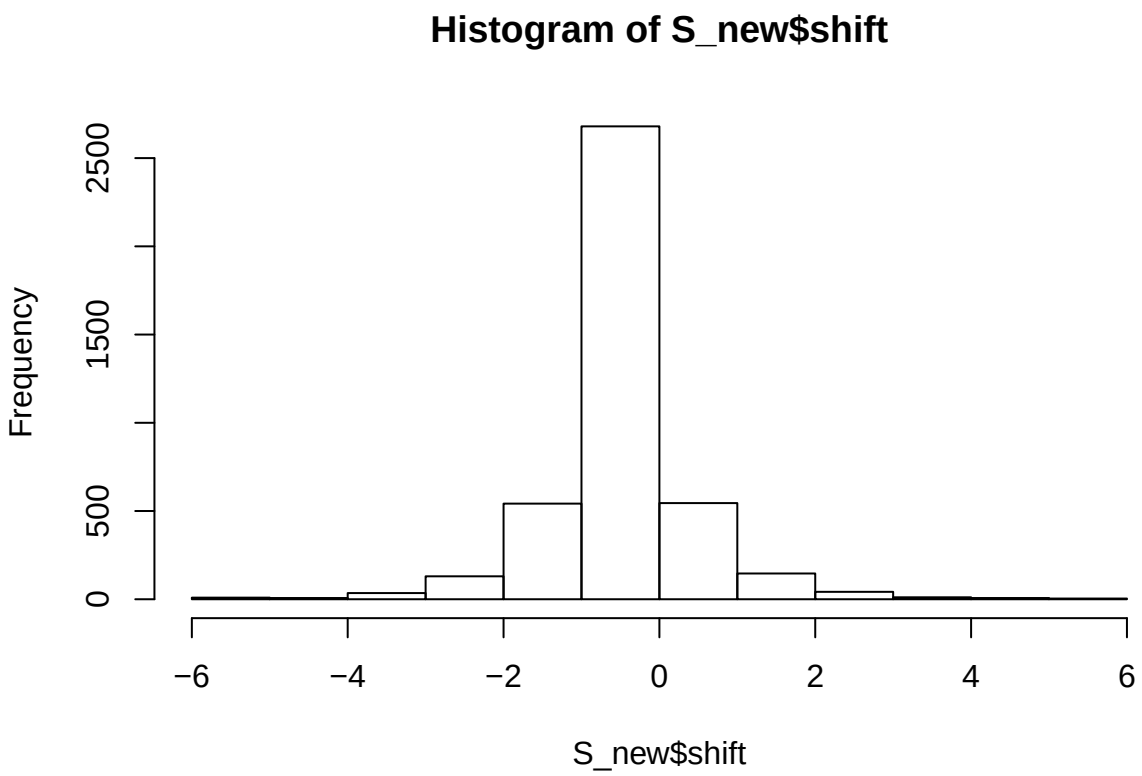
```
nrow(S_new[S_new$party=='Democrat',])
```

```
## [1] 2410
```

```
## To measure the shift we create a new variable called shift which
## is the difference between $libcpre_self and $libcpo_self
S_new$shift = S_new$libcpre_self_sc-S_new$libcpo_self_sc

summary(S_new$shift)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -6.00000  0.00000  0.00000  0.01972  0.00000  6.00000
```

```
hist(S_new$shift, breaks = 12)
```




```
## we see a high concentration in the at 0 and it does not look like a normal plot

## In additoin to that Since the libpre_self and libpo_self are ordinal varible the shift ordinal.
## Hence we need to do non-parameteric test.
##Since we need to check if republician or democrats are more likely to shift we
##need to evaulate the difference between libcpre_self and libcpo_self for republican
# vs democrats.Here we will be using a
##wilcon signed sum rank test for the shift variable between  republican or democrats.
##The null hypothesis is that repulicans are as likely as Democrats to shift thier
## political prefernce
## THe Alternate is that either republican or democratic are more likely to shift
## their prefernce
wilcox.test(S_new$shift~(S_new$party),paired = F)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  S_new$shift by S_new$party
## W = 2189300, p-value = 0.01096
## alternative hypothesis: true location shift is not equal to 0
```

```
## Here we see that the p-value is 0.01096 so we can reject the null hypothesis and
## that democrat or republicans are more likely to shift thier preference. This is
## because two-tailed test do not tell you directional differences

#We can see that taking the difference of means of libpre_self and libpo_self we see
# that democrats are more likely to shift there political opinion that Republicans
```

```
(mean(S_new[S_new$party=='Republican',]$libcpre_self_sc )-mean(S_new[S_new$party=='Republican',]$libcpo
```

```
## [1] -0.01887872
```

```
(mean(S_new[S_new$party=='Democrat',]$libcpre_self_sc )- mean(S_new[S_new$party=='Democrat',]$libcpo_se
```

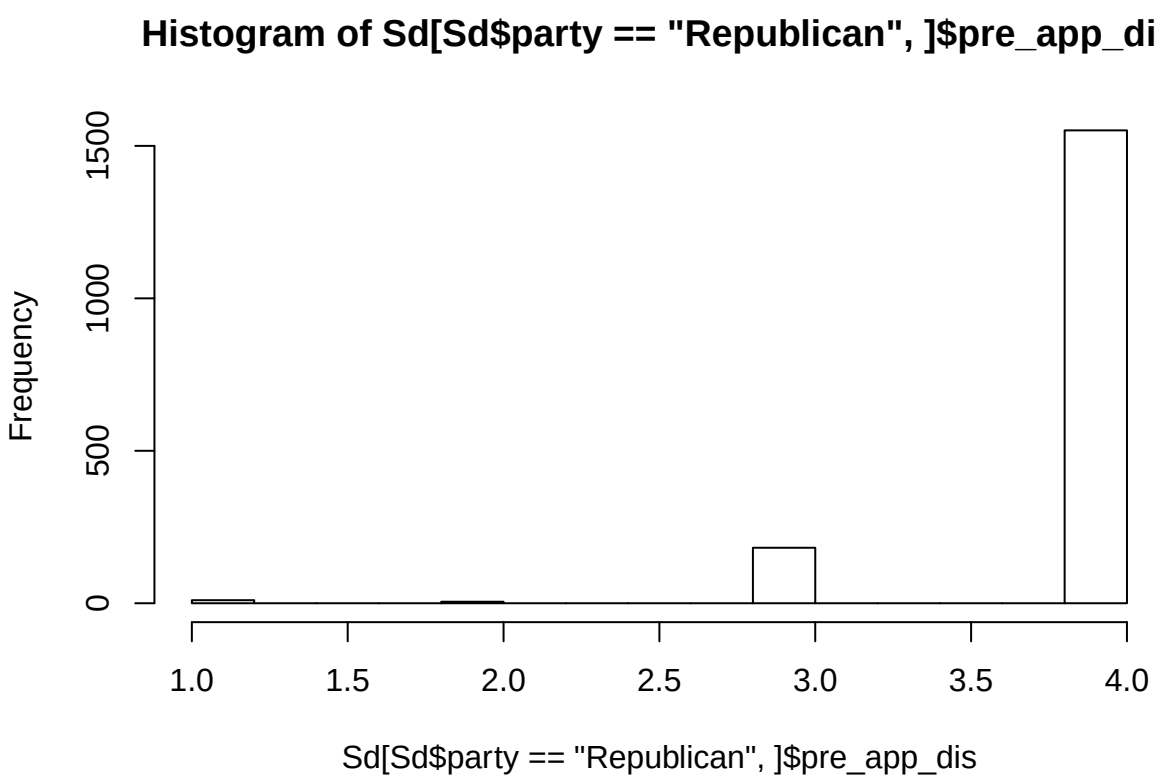
```
## [1] 0.04771784
```

5. Select a fifth question that you are interested in investigating.

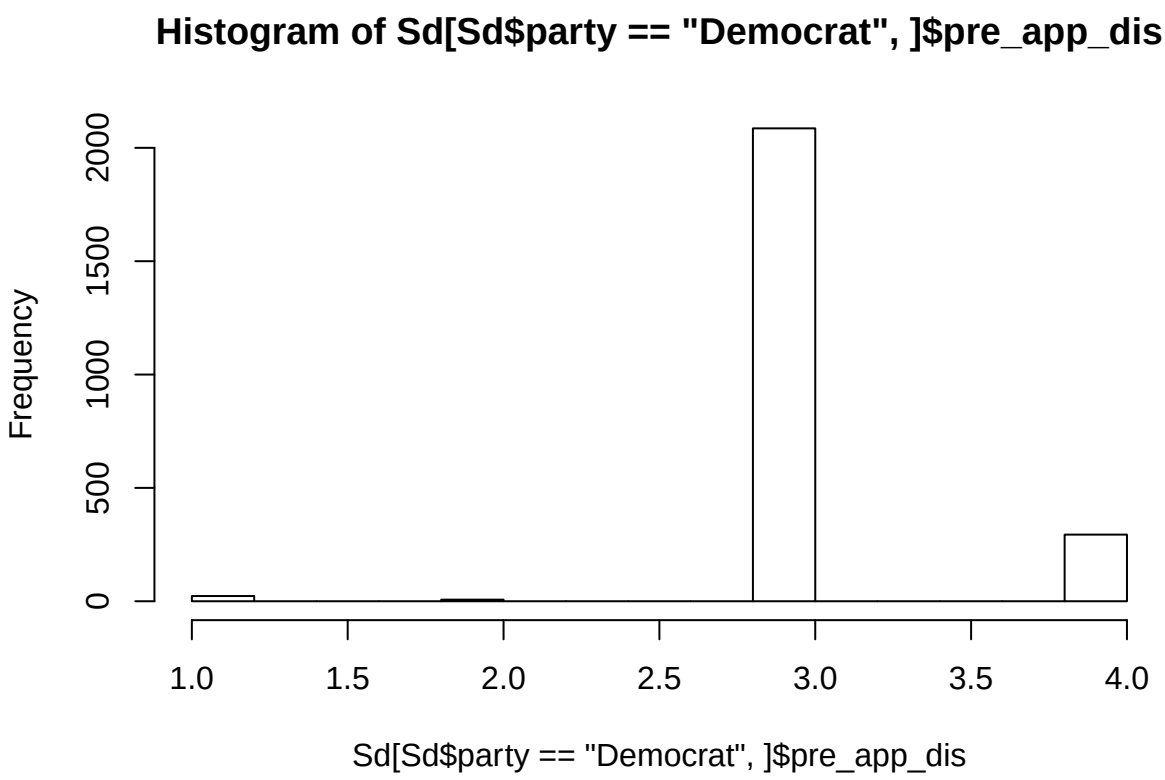
```
## Here we will evalualte if democrats approve the present
##more than the republcians.
##Here we willuse the presapp_job i.e Approve or disapprove
##President handling job as President

S_new$pre_app_dis = as.numeric(S_new$presapp_job)

table(S_new$presapp_job,as.numeric(S_new$presapp_job))
```

```
##
##                      1    2    3    4
##   -8. Don't know    33    0    0    0
##   -9. Refused        0   12    0    0
##   1. Approve         0    0 2268    0
##   2. Disapprove      0    0    0 1845
```

```
## we will only consider data points which are approve or disapprove
Sd = S_new
hist(Sd[Sd$party=='Republican',]$pre_app_dis)
```




```
hist(Sd[Sd$party=='Democrat',]$pre_app_dis)
```




```
## Here we see a clear separation that repulican disapprove
##Obama(4) and Democratic approve Obama(3) where are Demo
## Since we are intrested in evaluating if the democrates
##approve/disapprove more or less the
## same as the republicans.
## our null hypothesis is that  approval for the president is same
##for republicans and democrats.
## Since the presapp_job is an ordinal varaible we will have have to use non-parameteric test.
## Also we are just interest of the means are different, so we will
##run a independent wilcoxcin rank sum test.
wilcox.test(Sd[Sd$party=='Republican',]$pre_app_dis,Sd[Sd$party=='Democrat',]$pre_app_dis,paired = F)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Sd[Sd$party == "Republican", ]$pre_app_dis and Sd[Sd$party == "Democrat", ]$pre_app_dis
## W = 3705400, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
## We can see that the p-value is very small. So we can reject
##the null hypothesis that the
##republicans approve/disapprove the president same as the democrates
## we can take the means to get a estimate to see if democrates approve
##more than the republicans
mean(Sd[Sd$party=='Republican',]$pre_app_dis)
```

```
## [1] 3.872998
```

```r
mean(Sd[Sd$party=='Democrat',]$pre_app_dis)
```

```
## [1] 3.1
```