Lab group: Ammara Essa, Curtis Lin, & David Wheeler

Table of Contents

Lab 2: Comparing Means

w203 Statistics for Data Science

The Data

The American National Election Studies (ANES) conducts surveys of voters in the United States. While its flagship survey occurs every four years at the time of each presidential election, ANES also conducts pilot studies midway between these elections. You are provided with data from the 2018 ANES Pilot Study.

An important disclaimer is that the ANES Pilot Study does not represent a random sample of the U.S. population. 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 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. For the purposes of this assignment, however, you are not asked to use the survey weights. Instead, you should view your analysis as pertaining to the population of YouGov participants. (For groups with a strong interest in survey analysis, we recommend that you read about R's survey package (http://r-survey.r-forge.r-project.org/survey/). We will assign a very small number of bonus points (up to 3) to any group that correctly applies the survey weights and includes a clear explanation of how these work).

For a glimpse into some of the intricacies that go into the design of this study, take a look at the introduction to the ANES User's Guide and Codebook (https://electionstudies.org/wp-content/uploads/2019/02/anes_pilot_2018_userguidecodebook.pdf).

In [1]:
```
A = read.csv("anes_pilot_2018.csv")
```

Following is an example of a question asked on the ANES survey:

> How difficult was it for you to vote in this last election?

The variable `votehard` records answers to this question, with the following encoding:

- -1 inapplicable, legitimate skip
- 1 Not difficult at all
- 2 A little difficult
- 3 Moderately difficult
- 4 Very difficult
- 5 Extremely difficult

To see the precise form of each question, take a look at the Questionnaire Specifications (https://electionstudies.org/wp-content/uploads/2018/12/anes_pilot_2018_questionnaire.pdf).

Assignment

You will use the ANES dataset to address five research questions. For each question, you will need to operationalize the concepts (selecting appropriate variables and possibly transforming them), conduct exploratory analysis, deal with non-response and other special codes, perform sanity checks, select an appropriate hypothesis test, conduct the test, and interpret your results. When selecting a hypothesis test, you may choose from the tests covered in the async, including both paired and unpaired t-tests, as well as their nonparametric analogues. You may select a one-tailed or two-tailed test.

Please organize your response according to the prompts in this notebook.

Note that this is a group lab. There is a maximum of three students per team. You are free to form a group of your choice amongst the students in your live section. Although you may work on your own, we do not recommend this (we have found that individuals tend to do worse than teams on past labs).

Please limit your submission to 15 pages in total. This means that you will have to carefully prioritize which visualizations to include.

Submission Guidelines

- Submit *one* report per group.
- Submit *both* your pdf report as well as your source file.
- **Only analyses and comments included in your PDF report will be considered for grading.**
- Include names of group members on the front page of the submitted report.
- Naming structure of submitted files:
  - PDF report: [student_surname*1]*\[student_surname*2][**]_lab_2.pdf
  - Jupyter Notebook: [student_surname*1]*\[student_surname*2][**]_lab_2.ipynb

Research Questions

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

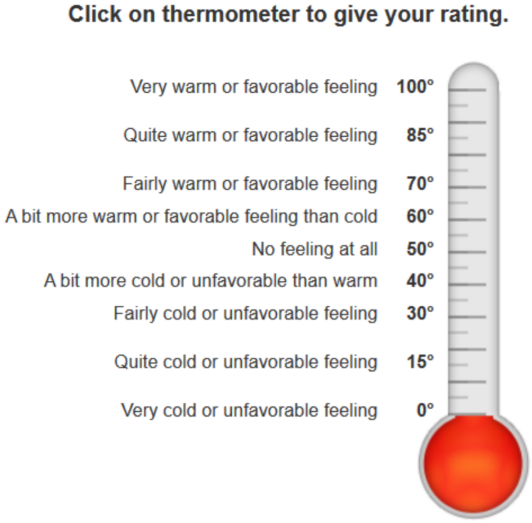
Introduce your topic briefly. (5 points)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

- The subset of data from the 2018 ANES study includes responses from 2500 participants on a variety of questions. Two such question are regarding the rating (sentiment) towards police and journalists.
- Of the 767 available variables, the variables that will help shed light on the matter of whether US voters have more respect for the police or for journalists are

$[\mathbf{ftpolice}]$: How would you rate the police?
$[\mathbf{ftjournal}]$: How would you rate journalists?

Respondents are asked to give their feedback using a feeling thermometer with the following description



Based on the study questionnaire specifications
- Ratings between 50 degrees and 100 degrees mean that the respondents feel favorable and warm towards a particular subject
- Ratings between 0 degrees and 50 degrees mean that the respondents don't feel favorable toward towards a particular subject
- Rating of exactly 50 means that the respondent does not feel particularly warm or cold towards a particular subject

We note that while strictly speaking, there is no separate question for level of respect for either entity, we will use favorability ranking to be indicative of respect throughout this analysis e.g. high favorabilibilty ranking implies high respect etc. Moreover, the variables are ordinal i.e. the difference between option A and option B may not be the same as the difference between option B and option C. As a result, in general, we cannot simply average the responses i.e. $\frac{A+C}{2} \neq B$

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

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

In [3]:
```
# Add relevant libraries
library(gmodels)
library(effsize)
library(powerAnalysis)
library(descr)
```

- As noted above, the [ftpolice] and [ftjournal] are ordinal variables. However, inspection of unique values of both variables shows every possible valid value between 0 and 100 (-7 means missing value) . So while the difference between some sentiments categories is 10 and between other categories is 15, for this analysis we will consider them continous cardinal vairables upon which parametric tests can be performed.

In [4]:
```
paste('Unique values of Police rating')
sort(unique(A$ftpolice))
paste('Unique values of Journalist rating')
sort(unique(A$ftjournal))
```

'Unique values of Police rating'

```
0   1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17  18  19  20  21  22
23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42
43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62
63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82
83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100
```

'Unique values of Journalist rating'

```
-7  0   1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17  18  19  20  21
22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41
42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61
62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81
82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100
```

- Replace any missing values (-7) by NA so that R functions can take appropriate action

In [5]:
```
# Replace any missing values (-7) with null values recognized by R i.e. NA
A$ftpolice[A$ftpolice == -7] <- NA
A$ftjournal[A$ftjournal == -7] <- NA
#sort(unique(A$ftpolice))
#sort(unique(A$ftjournal))
paste("Verify that -7 has been replaced with NA ")
paste("Police data NA values = ", sum(is.na(A$ftpolice)))
paste("Journalist data NA values = ", sum(is.na(A$ftjournal)))
```

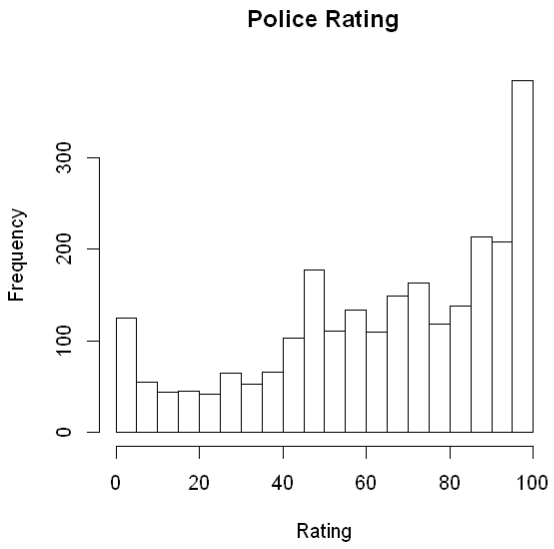
'Verify that -7 has been replaced with NA '

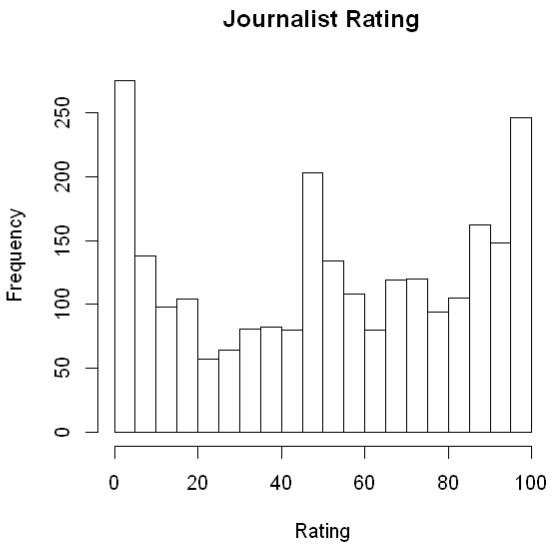
'Police data NA values = 0'

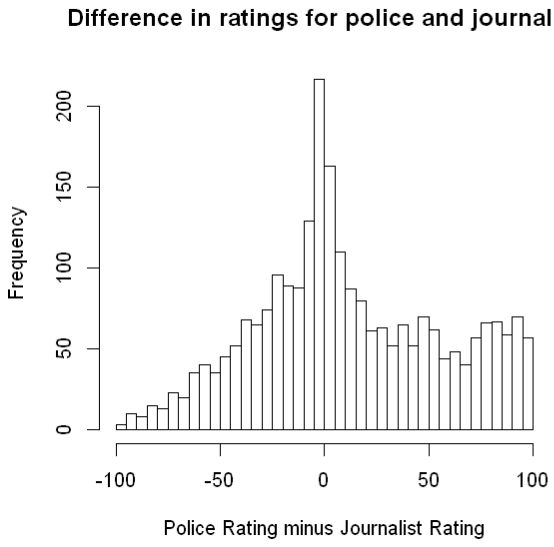
'Journalist data NA values = 2'

- Now that some basic data wrangling has been performed, we move on to inspecting the data to assess which statistical analysis is best suited
  - We see that the distribution for neither [ftpolice] nor [ftjournal] variables are exactly normal, [ftpolice] even seems slightly skewed to the right. However, the sample size is large enough to invoke the Central Limit Theorem
  - Moreover, the difference between [ftpolice] and [ftjournal] appears to be closer to a normal distribution (despite a few spikes), as can be seen by the histogram below. Again, the sample size is enough to invoke CLT to proceed with the appropriate statistical analysis

In [6]:
```
# Illustrate distribution
par(mfrow=c(1,2))
options(repr.plot.width=10, repr.plot.height=5)
hist(A$ftpolice, breaks=30, main="Police Rating", xlab="Rating")
hist(A$ftjournal, breaks=30, main="Journalist Rating", xlab="Rating")
hist(A$ftpolice - A$ftjournal, breaks=30, main="Difference in ratings for poli
ce and journalist", xlab="Police Rating minus Journalist Rating")
```







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

Explain why your test is the most appropriate choice, focusing on its statistical assumptions.

- Based on the EDA and assumptions above and the fact that we do not know the population variance, a t-test is the appropriate choice for the analysis.
- The null hypothesis H_0 and alternative hypothesis H_A are given as follows

$\mathbf{H_0}$: There is no difference in respect for police and journalist : $\mu_p - \mu_j = 0$
$\mathbf{H_A}$: There is a difference in respect for police and journalist : $\mu_p - \mu_j \neq 0$

We thus conduct a paired t-test to test the hypothesis at a significan level $\alpha = \mathbf{0.05}$. A paired test is appropriate here because the responses for police and journalists ratings are given by the same respondent.

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

In [7]:
```
# Perform a two-sided, paired t-test
t.test(A$ftpolice, A$ftjournal, paired=TRUE, na.action=na.omit )
# na.rm indicates whether NAs should be removed before computation; if paired=
=TRUE then all incomplete pairs are removed.
cohen.d(A$ftpolice, A$ftjournal, paired=T, na.rm=TRUE)
```

```
        Paired t-test

data:  A$ftpolice and A$ftjournal
t = 13.711, df = 2497, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 10.58776 14.12160
sample estimates:
mean of the differences
               12.35468

Cohen's d

d estimate: 0.2743325 (small)
95 percent confidence interval:
    lower     upper
0.2186004 0.3300646
```

Statistical significance

The results can be broken into two parts.

- Firstly, the calculated p-value is extremely small $(p < 2.2e-16)$ and of course $p < 0.05$ and thus we reject the null hypothesis H_0 that there is no difference in US voters respect toward the police and journalists. In fact, H_0 could have been rejected at an even stricter significance level of $\alpha = 0.01$
- Secondly, looking at the calculated value of the resulting 'mean of the differences', we see that $\mu_p - \mu_j > 0$, which is indicative that US voters sampled by the YouGov survey tend to respect police more than journalist. #### Practical Significance
- While the result is statistically signifincant, the effect size calculated by Cohen's d = 0.2743325 indicates that the result may be of less if not neglible practical significance.

∴ In response to the question: "Do US voters have more respect for the police or for journalists?" we conclude that that statistically speaking, US voters sampled from YouGov may respect police more than journalists though the difference may be practically less significant.

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

Introduce your topic briefly. (5 points)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

- To analyze the question above, we need to know two things
  - The party to which the voters associate themselves with
  - Voter age

- We will extract the voter age by using the $[\mathbf{birthyr}]$ ("In what year were you born?") variable in the given data. Since the survey was conducted in 2018, the age will be calculated accordingly.

- The party allegience for the voters is a little trickier to obtain from the data. There are the varaibles that are of interest
  - $[\mathbf{pid1d}]$: Generally speaking, do you usually think of yourself as a Democrat, a Republican
  - $[\mathbf{pid1r}]$: Generally speaking, do you usually think of yourself as a Republican, a Democrat
  - $[\mathbf{pidlean}]$: Do you think of yourself as closer to the Republican Party or to the Democratic

[PARTY ID]

[GENERATE RANDOMIZATION VARIABLE RAND_PID = 1 OR 2]
[IF RAND_PID =1]
[pid1d] Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent, or what?
__ Democrat [1]
__ Republican [2]
__ independent [3]
__ something else [4]

[IF RAND_PID =2]
[NOTE RESPONSE CODE VALUES MATCH pid1d BUT ORDER (2,1,3,4) DIFFERS]
[pid1r] Generally speaking, do you usually think of yourself as a Republican, a Democrat, an independent, or what?
__ Republican [**2**]
__ Democrat [**1**]
__ independent [3]
__ something else [4]

[IF pid1d=3 OR 4 OR NO ANSWER OR pid1r = 3 OR 4 OR NO ANSWER]
[pidlean] Do you think of yourself as closer to the Republican Party or to the Democratic Party?
__ Closer to the Republican Party [1]
__ Closer to the Democratic Party [2]
__ Neither [3]

- Here we note the distinction in how the [pid1d] and [pid1r] are framed. [pid1d] leads the question with 'Democrat' while [pid1r] leads the question with 'Republican'. There is a possibility that somone on the political fence may respond differently when the question is led by one party or another. Further inspection of the questionare specifications shows that in what appears to be an effort to compensate for this possible influence, a random number (1 or 2) is assigned to each respondent and the question is then posed accordingly. If the [pid1d] format question is given to the participant, the [pid1r] entry for the same participant is marked as -1 (legitimate skip) and vice versa. Thus, we can extract the party preference by looking at the valid entries of [pid1d] and [pid1r], where 1 = Democrat and 2 = Republican in both cases.

- As for [pidlean], that question is only given to the participant if they respond Independent [3] or Something else [4]. While this may indicate if a voter is more inclined towards a certain party, it still does not definitively identify the respondant as either.

- Thus, for this analysis, we will only focus on responses that clearly identify the voter as Democrat or Repulican and will only extract party identity from variables [pid1d] and [pid1r]

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

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

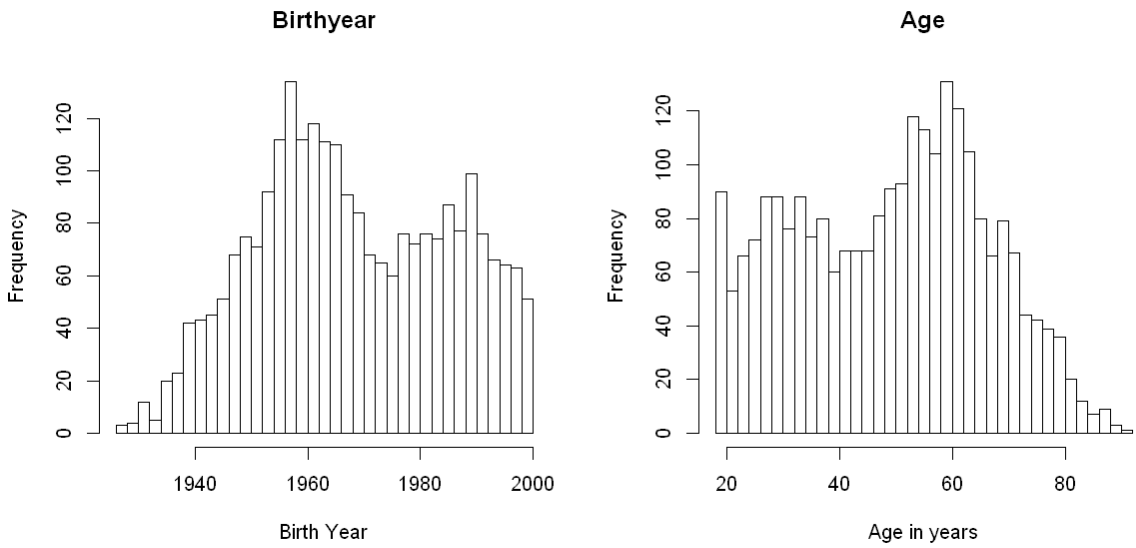
Voter Age

- First, we look at the birthyr variable to ensure it is not too skewed or non normal. While the distribution is not exactly normal and there is a little skew, the large sample size allows us to invoke the CLT.
- It should be noted that the histogram is for all respondents and not just voters with Democratic or Republican allignment. We will sanity check the age distribution for just Republican and Democrat voters prior to the statistical test.
- Since they aren't any drastic anomalies in with the birthyr variable, we will add a column "voterAge" by performing a simple calculation of Age = (2018 - birthyr). No surprise that the histogram of age is a 'flipped' version of the birthyear distribution, thus the same conclusions as birthyr apply. Moreover, the youngest calculated voter age is 18, which indicates the calculation (2018 - birthyr) is valid.
- voterAge is a cardinal variable and is thus valid for a parametric statistical test

In [8]:
```
# Calculate age w.r.t. 2018 survey date
A$voterAge <- (2018 - A$birthyr)
paste("Summary of Voter Age")
summary(A$voterAge)
par(mfrow=c(1,2))
options(repr.plot.width=10, repr.plot.height=5)
options(plot.width=10, repr.plot.height=5)
hist(A$birthyr, breaks=50, main="Birthyear", xlab="Birth Year")
hist(A$voterAge, breaks=50, main="Age", xlab="Age in years")
```

'Summary of Voter Age'

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



Party Identity

- Now we inspect variables pid1d and pidir
- As mentioned above, for every non-missing (-7) response, only pid1d or pid1r will contain valid information. We verify this by looking at the first 6 entries of pid1d and pidir and see that first 6 party identifiers are: [2] [2] [3] [3] [2] [1]

From the tables for pid1d and pid we see the # of Democrats [1] = 432 + 425 = 857 , # of Republicans [2] = 326 + 283 = 609, so on

In [9]:
```
paste("Table for pid1d")
table(A$pid1d)
paste("Table for pid1r")
table(A$pid1r)
paste("First 5 entries for pid1d")
head(A$pid1d)
paste("First 5 entries for pid1r")
head(A$pid1r)
```

'Table for pid1d'

```
  -7   -1    1    2    3    4 
   1 1331  432  326  356   54 
```

'Table for pid1r'

```
  -7   -1    1    2    3    4 
   1 1317  425  283  411   63 
```

'First 5 entries for pid1d'

2 -1 -1 3 -1 -1

'First 5 entries for pid1r'

-1 2 3 -1 2 1

- Now that we know how the data are structured, we can add yet another column "myParty" that will contain the result of a function that checks the values in pid1r and pidid and returns the following:
  $\mathbf{0}$ = Other
  $\mathbf{1}$ = Democrat
  $\mathbf{2}$ = Republican

But before that, we replace any missing values -7 by NA

In [10]:
```
# Replace any missing values (-7) with null values recognized by R i.e. NA
A$pid1d[A$pid1d == -7] <- NA
A$pid1r[A$pid1r == -7] <- NA

# For any given party identification question, if the question is framed as pi
d1d, then pid1r will be -1 and vice versa
# From the data, 1 = Democrat and 2 = Republican

voterparty <- function(varpid1d, varpid1r){
    if (is.na(varpid1d) | is.na(varpid1r)){
        party <- NA }
    else if (varpid1d == 1 | varpid1r == 1){
        party <- 1 }
    else if  (varpid1d == 2 | varpid1r == 2){
        party <- 2 }
    else {
        party <- 0}
    return(party)}

A$myParty <- mapply(voterparty, A$pid1d, A$pid1r)
paste("Table for calculated myParty variable")
table(A$myParty)
paste("Democrat = 1 ; Republican = 2 ; Other = 0")
```

'Table for calculated myParty variable'

```
   0    1    2
1032  857  609
```

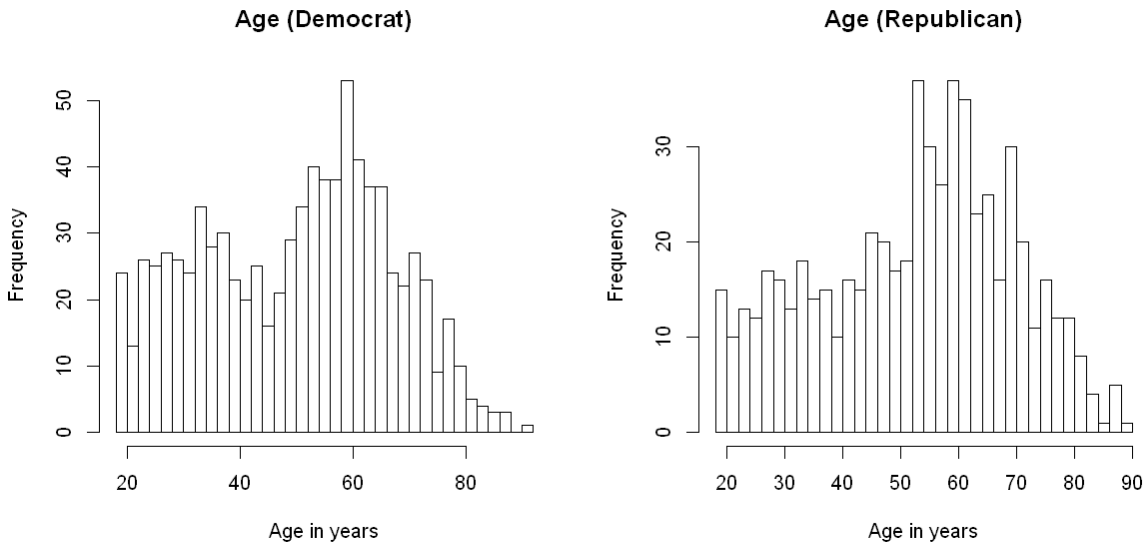
'Democrat = 1 ; Republican = 2 ; Other = 0'

- Finally, we can now extract two subsets of data to perform analysis on, one for Democrats and the other for Republican voters. We perform a final sanity check that number of entries are consistent with the original data and that the age histograms have not dramatically. Again, we note that while the age distribution for neither Democrats nor Republican is exactly normal, the large sample size allows us to employ the Central Limit Theorem for statistical analysis.

In [11]:
```
dem <- subset(A, A$myParty == 1)
rep <- subset(A, A$myParty == 2)
paste('Size of Democrat dataset', nrow(dem),'rows', 'x', ncol(dem), 'columns')
paste('Size of Democrat dataset', nrow(rep),'rows', 'x', ncol(rep), 'columns')
par(mfrow=c(1,2))
options(repr.plot.width=10, repr.plot.height=5)
hist(dem$voterAge, breaks=50, main="Age (Democrat)", xlab="Age in years")
hist(rep$voterAge, breaks=50, main="Age (Republican)", xlab="Age in years")
```

'Size of Democrat dataset 857 rows x 769 columns'

'Size of Democrat dataset 609 rows x 769 columns'



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

Explain why your test is the most appropriate choice, focusing on its statistical assumptions.

- Based on the EDA, we can invoke the CLT and since we do not know the population variance, we will conduct a parametric t-test to analyze the null hypothesis H_0 and alternative hypothesis H_A which are given below.

$\mathbf{H_0}$: There is no difference in the ages of Democratic and Republican voters : $\mu_d - \mu_r = 0$
$\mathbf{H_A}$: There is a difference in the ages of Democratic and Republican voters : $\mu_d - \mu_r \neq 0$

This will be a two-tailed unpaired t-test at a significan level $\alpha = \mathbf{0.05}$. An unpaired test is appropriate here because Democratic and Republican voters are independent.

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

In [12]:
```
t.test(dem$voterAge , rep$voterAge)
cohen.d(dem$voterAge , rep$voterAge)
```

```
        Welch Two Sample t-test

data:  dem$voterAge and rep$voterAge
t = -2.939, df = 1309.7, p-value = 0.00335
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.3723921 -0.8718651
sample estimates:
mean of x mean of y
 50.23337  52.85550

Cohen's d

d estimate: -0.155755 (negligible)
95 percent confidence interval:
      lower       upper
-0.25987016 -0.05163986
```

Statistical significance

The results can be broken into two parts.

- Based on the calculated p-value $(p = 0.00335)$ and $p < 0.05$ we reject the null hypothesis H_0 that there is no difference in the ages of Democratic and Republican voters
- Secondly, since $H_A = \mu_d - \mu_r \neq 0$, the -ve values of the 95% confidence interval and the values of the mean of voter age for Democrats and Republicans indicate that Republican voters are slightly older than Democratic voters.

Parctical significance

- While the result is statistically signifincant, the effect size calculated by Cohen's d = -0.155755 is extremly low and indicates that difference in ages is of negligible practical significance.

∴ In response to the question: "Are Republican voters older or younger than Democratic voters?" we conclude that that statistically speaking, of the respondents of YouGov survey, Republican voters are slightly older than their Democratic counterparts but practically speaking, the difference has neglible significance in the real world.

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)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

- The construct we want to measure is: the belief of independent voters in the baselessness of investigations around the 2016 election interference by Russian officials.
- To estimate this construct, we argue that 3/767 variables ([muellerinv],[russia16], and [coord16]) provide different but appropriate abstractions. Of these three [muelleriv] is the closest to the construct of interest. This variable categorizes the approval of the Mueller probe on a scale from 1-7.

[muellerinv]	Do you approve, disapprove, or neither approve nor disapprove of Robert Mueller's investigation of Russian interference in the 2016 election?
	__ Approve extremely strongly [1]
	__ Approve moderately strongly [2]
	__ Approve slightly [3]
	__ Neither approve nor disapprove [4]
	__ Disapprove slightly [5]
	__ Disapprove moderately strongly [6]
	__ Disapprove extremely strongly [7]

Gaps between the construct and variables:

- Just because voters, for example, dissaprove with Mueller probe, does not mean that they think it is baseless, $sensu\ stricto$. It is possible that voters think that the probe is well-founded but that Mueller's leadership, execution, etc. leave something to be desired. The baselessness of the probe and the approval thereof by the voters, therefor, can be divorced into two separate constructs. Other examples are abound. For example, it is possible that some voters think that the probe is baseless but nevertheless approve of it in order to let the justice system run its course.
- Given the possible discrepancies between the construct described in the question and the [muellerinv] variable, all future inferences with be conditioned on these limitations.

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

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

- Subset only indepedent voters who (i) identified as Independent [3] in response to pid1d OR (ii) identified as Independent [3] in response to pid1r AND (iii) for those identified as Independent for either pid1d or pid1r, identified as neither Democrat or Republican in response to pidlean.

In [13]:
```
Ind_Voters = subset(A, pid1d == 3 | pid1r == 3 & pidlean == 3)
```

- Remove non-response cases [-7]:

In [14]:
```
Ind_Voters = Ind_Voters[!(Ind_Voters$muellerinv == "-7"),]
```

- Convert integers to factors

In [15]:
```
Ind_Voters[,'muellerinv']<-factor(Ind_Voters[,'muellerinv'])
```

- Tabulate data

In [16]:
```
table(Ind_Voters$muellerinv)
```

```
  1   2   3   4   5   6   7 
127  44  35 155  23  28  91 
```

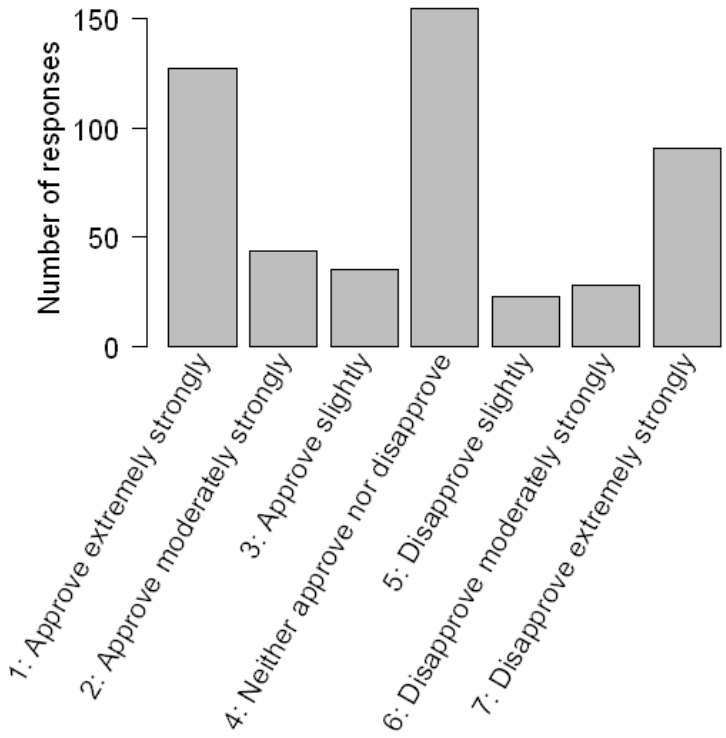
- Visualize the number of responses for each category

  - x-axis labels

In [17]:
```
x_labs =c("1: Approve extremely strongly",
                  "2: Approve moderately strongly",
                  "3: Approve slightly",
                  "4: Neither approve nor disapprove",
                  "5: Disapprove slightly",
                  "6: Disapprove moderately strongly",
                  "7: Disapprove extremely strongly")
```

  - Plot

In [18]:
```
options(repr.plot.height = 6, repr.plot.width = 7, repr.plot.pointsize = 10)
par(mar = c(7, 4, 2, 2) + 8) #add room for the rotated labels
bp=barplot(table(Ind_Voters$muellerinv),
        las=1,
        names.arg="",
        ylab = "Number of responses")
text(bp[,1], -3.7,
     srt=60, adj=1,
     xpd = TRUE,
     labels = x_labs, cex=1)
options(repr.plot.height = 9, repr.plot.width = 7, repr.plot.pointsize = 10)
```



- Sanity checks

In [19]:
```
sum(table(Ind_Voters$muellerinv)) == sum(summary(Ind_Voters$muellerinv)) # do
 sums equate?
```

TRUE

In [20]:
```
sum(is.na(Ind_Voters$muellerinv)) # any NA's?
```

0

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

Explain why your test is the most appropriate choice, focusing on its statistical assumptions.

- Given that these data are ordinal- the distance between 1 and 2 is not necessarily the same as the distance between 2 and 3, 3 and 4, etc... a t-test is therefore not valid.
- Similarly, since we do not have two continuous or categorical variables we can not (i) apply correlations or OLS regressions or (ii) a χ^2 test for independence to determine if political party is independent of the reported beliefs about federal investigations into Russian interference. Although interesting, the latter option would really not address our null hypothesis, H_0: the majority of independent voters do not believe that the federal investigations of Russian election interference are baseless.

$\star$ To test H_A: the majority of independent voters believe that the federal investigations of Russian election interference are baseless, we really need to use a rank based test, since the hypothesis posits about the $majority$ of independent voters. This create new changes, however. We have 7 groups of respondants within the Indepedent voters, but have only discussed tests that differentiate means or ranks of 2 groups. In order to reconile this conflict between the tests we have learned and the number of levels (7) within the independent variable we need to collapse the 7 categories into 2 groups and proceed with the Wilcoxon rank-sum test. This approach, however, creates yet another issue- in order to collapse these data into two groups, those who support and those who do not support the investigation, we need to ignore the voters who with an answer of [4], since they are indifferent to the investigation and do not contribute to either group.

$\star$ For the Wilcoxon rank-sum test, the H_0: there is no difference in ranks between independent voters who support or do not support the federal investigation into Russian interference of the 2016 US election. Further, we can justify a uni-directional test, because our $a\ priori\ H_A$ is about the **majority** of voters.

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

- First, polarize thethe groups of 7 respondants into two groups, those that support the investigations, "Yes_Inv", and those that do not, "No_Inv".

In [21]:
```
Yes_Inv = subset(Ind_Voters, muellerinv == "1" |
                 muellerinv == "2" |
                 muellerinv == "3")
```

In [22]:
```
No_Inv = subset(Ind_Voters, muellerinv == "5" |
                muellerinv == "6" |
                muellerinv == "7")
```

- Rename factor levels from 5,6,7 to 3,2,1 to match those from the voters who supported the investigation.

In [23]:
```
No_Inv$muellerinv[No_Inv$muellerinv == "5"] <- 3
No_Inv$muellerinv[No_Inv$muellerinv == "6"] <- 2
No_Inv$muellerinv[No_Inv$muellerinv == "7"] <- 1
```

- Conduct the Wilcoxon rank-sum test & recover statistical significance

In [24]:
```
wilcox.test(as.numeric(No_Inv$muellerinv),
            as.numeric(Yes_Inv$muellerinv),
            alternative = "greater")
```

```
        Wilcoxon rank sum test with continuity correction

data:  as.numeric(No_Inv$muellerinv) and as.numeric(Yes_Inv$muellerinv)
W = 14286, p-value = 0.666
alternative hypothesis: true location shift is greater than 0
```

- Practical significance

In [25]:
```
cohen.d(as.numeric(No_Inv$muellerinv),as.numeric(Yes_Inv$muellerinv))
```

```
Cohen's d

d estimate: -0.04219915 (negligible)
95 percent confidence interval:
     lower      upper
-0.2567491  0.1723508
```

⋆ Interpretation: fail to reject H_0 there is no difference (p=0.66 > α = 0.05) in ranks between Independent voters that supported or refuted the Mueller investigation. This provides very very weak evidence in support of the original H_A the ranks of each group are different and, more specifically, the rank of the group who did not support the investigation > the rank of the group who did not. This claim is corroborated by negligible practical signficance (Cohen's d = -0.042).

∴ In response to the initial question: "Do a majority of independent voters believe that the federal investigations of Russian election interference are baseless?" we conclude that there is no evidence the majority of Independent voters sampled from YouGov believe the Mueller probe is 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)

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

- The constructs we want to measure are anger, fear, voter turnout in 2016, and voter turnout 2018.
- To estimate these constructs, we argue that several variables provide adequate approximations of these contructs.

- For anger and fear, [geangry] and [geafraid] are appropriate approximations, in response to:

> "Generally speaking, how do you feel about the way things are going in the country these days?"

		Not at all [1]	A little [2]	Somewhat [3]	Very [4]	Extremely [5]
[geangry]	How angry do you feel?	○	○	○	○	○
[geafraid]	How afraid do you feel?	○	○	○	○	○

- For voter turnout in 2016 and 2018, [turnout16] and [turnout18] are appropriate approximations to the constructs of interest.

[turnout16] In 2016, the major candidates for president were Donald Trump for the Republicans and Hillary Clinton for the Democrats. In that election, did you definitely vote, definitely not vote, or are you not completely sure whether you voted?
_ Definitely voted [1]
_ Definitely did not vote [2]
_ Not completely sure [3]

[turnout18] In the election held on November 6, did you definitely vote in person on election day, vote in person before Nov 6, vote by mail, did you definitely not vote, or are you not completely sure whether you voted in that election?

__ Definitely voted in person on Nov 6 [1]
__ Definitely voted in person, before Nov 6 [2]
__ Definitely voted by mail [3]
__ Definitely did not vote [4]
__ Not completely sure [5]

Gaps between the construct and variables:

- These emotions can co-occur, as they did in this dataframe where some individuals responded to both [geangry] and [geafraid] with "5". Indeed, these emotions tend to be closely related- see the EDA, below. This false dichotomy prevents us from detecting the real effects of both emotions on voter turnout, say with multiple regression.
- Both [geangry] and [geafraid] assume respondents are submitting honest answers, this may not always be a safe assumption and may introduce some distance/gaps between the contructs of anger and fear and these actualized variables.
- Similiarly, there is no direct association between responses to [geangry] and [geafraid] and [turnout16] and [turnout18]- just because voters are emotive (angry and or fearful) does not mean they will vote differently than being happy or sad. Voters may be happy, sad, angry, or fearful for a variety of reasons which may or may not have affected voter turnout in 2016 or 2018.

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

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

Wrangle data

Angry voters

> Convert response [-7], no Answers, to NAs since they do not contribute to testing to our H_0 & cell counts for Pearson's χ^2 test for independence

In [26]:
```
A$geangry[A$geangry == "-7"] <- NA
```

Fearful voters

> Convert response [-7], no Answers, to NAs since they do not contribute to testing to our H_0 & cell counts for Pearson's χ^2 test for independence

In [27]:
```
A$geafraid[A$geafraid == "-7"] <- NA
```

Voter turnout for 2016

> Convert response [3], Not completely sure, to NAs since they do not contribute to testing to our H_0 & cell counts for Pearson's χ^2 test for independence

In [28]:
```
A$turnout16[A$turnout16 == "3"] <- NA
```

Voter turnout for 2018

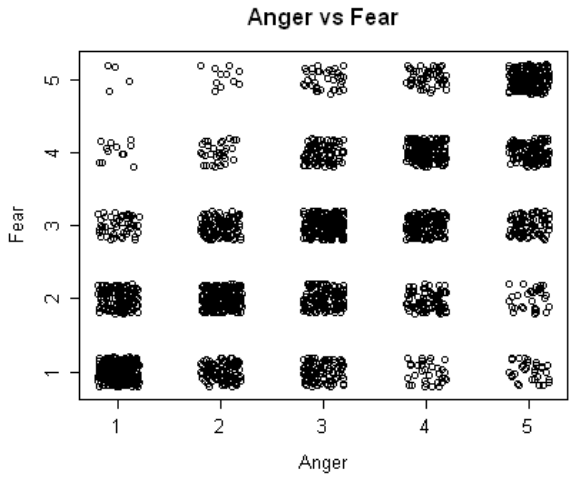
> Convert response [5], Not completely sure, to NAs since they do not contribute to testing to our H_0 & cell counts for Pearson's χ^2 test for independence

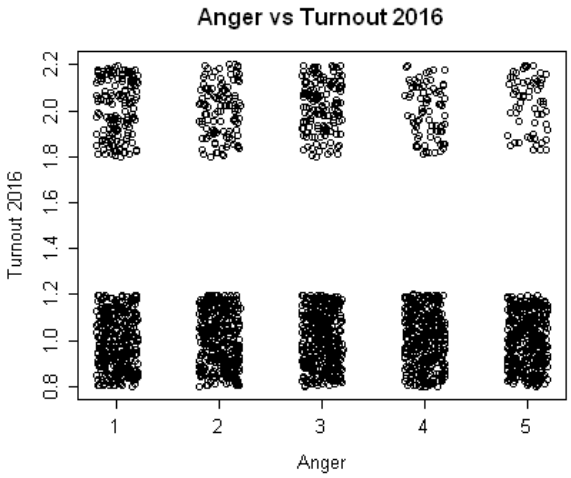
In [29]:
```
A$turnout18[A$turnout18 == "5"] <- NA
```

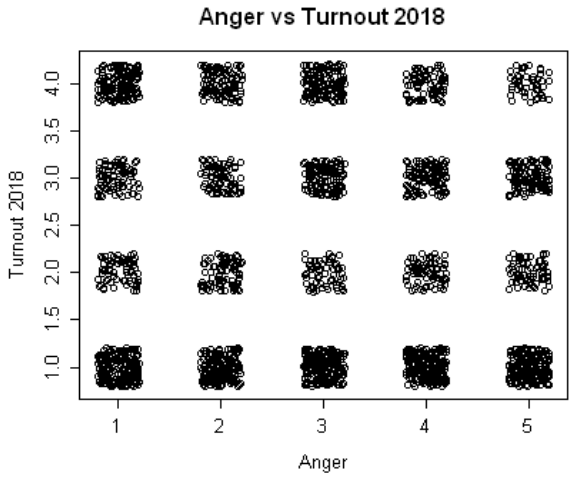
Exploratory data analyses

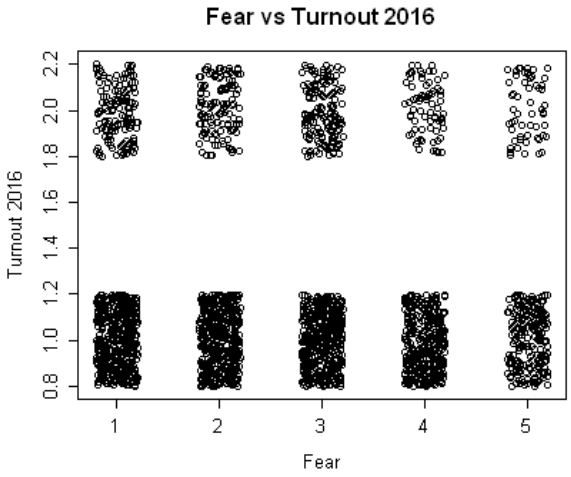
> - Pairwise relationships between all variables

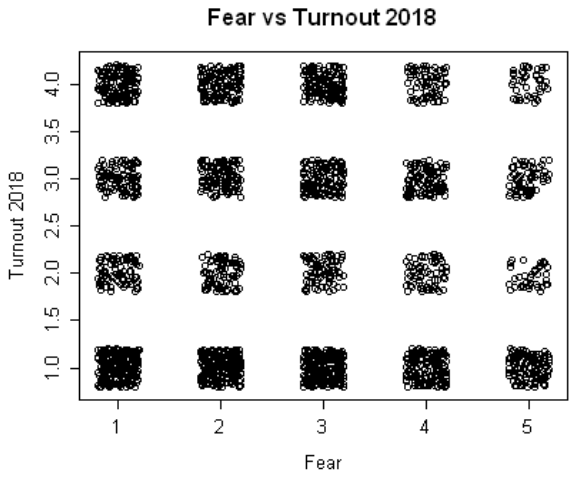
In [30]:
```
par(mfrow=c(3,2)) # specify plotting space
plot(jitter(A$geangry),jitter(A$geafraid),
    xlab="Anger",ylab="Fear", main="Anger vs Fear")
plot(jitter(A$geangry),jitter(A$turnout16),
    xlab="Anger",ylab="Turnout 2016",main="Anger vs Turnout 2016")
plot(jitter(A$geangry),jitter(A$turnout18),
    xlab="Anger",ylab="Turnout 2018",main="Anger vs Turnout 2018")
plot(jitter(A$geafraid),jitter(A$turnout16),
    xlab="Fear",ylab="Turnout 2016",main="Fear vs Turnout 2016")
plot(jitter(A$geafraid),jitter(A$turnout18),
    xlab="Fear",ylab="Turnout 2018",main="Fear vs Turnout 2018")
plot(jitter(A$turnout16),jitter(A$turnout18),
    xlab="Turnout 2016",ylab="Turnout 2018",main="Turnout 2016 vs Turnout 201
8")
options(repr.plot.height = 9, repr.plot.width = 7, repr.plot.pointsize = 10)
```

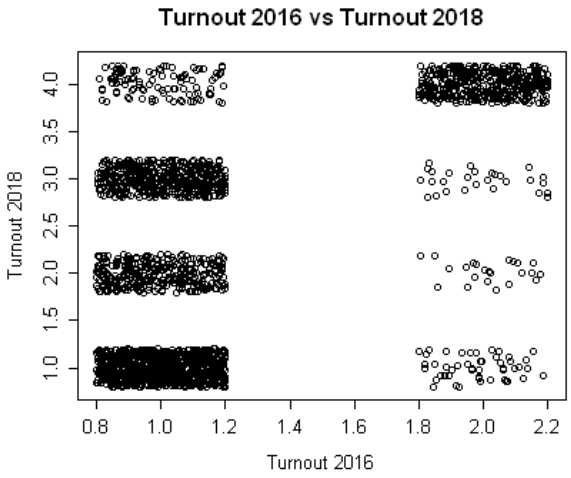

Anger vs Fear
Fear
Anger
1
2
3
4
5


Anger vs Turnout 2016
Turnout 2016
Anger
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2


Anger vs Turnout 2018
Turnout 2018
Anger
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Fear vs Turnout 2016
Turnout 2016
Fear
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2


Fear vs Turnout 2018
Turnout 2018
Fear
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Turnout 2016 vs Turnout 2018
Turnout 2018
Turnout 2016
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2

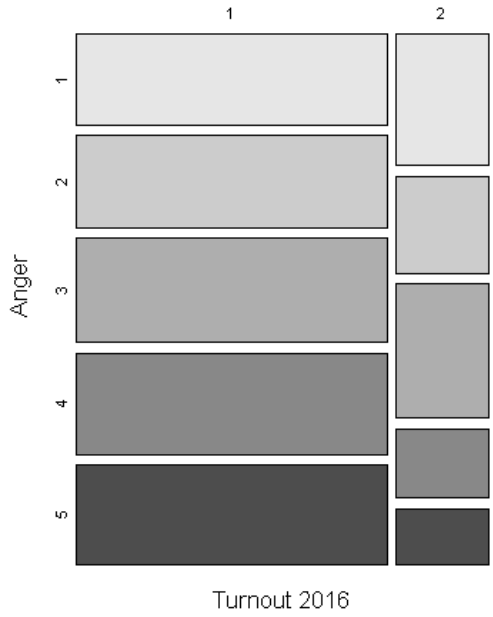
In [31]:
```
par(mfrow=c(2,2))

crosstab(A$geangry,A$turnout16,
         xlab = "Turnout 2016", ylab = "Anger",
         expected = TRUE, prop.c = TRUE,
         prop.r = TRUE)

crosstab(A$geangry,A$turnout18,
         xlab = "Turnout 2018", ylab = "Anger",
         expected = TRUE, prop.c = TRUE,
         prop.r = TRUE)

crosstab(A$geafraid,A$turnout16,
         xlab = "Turnout 2016", ylab = "Fear",
         expected = TRUE, prop.c = TRUE,
         prop.r = TRUE)

crosstab(A$geafraid,A$turnout18,
         xlab = "Turnout 2018", ylab = "Fear",
         expected = TRUE, prop.c = TRUE,
         prop.r = TRUE)
```

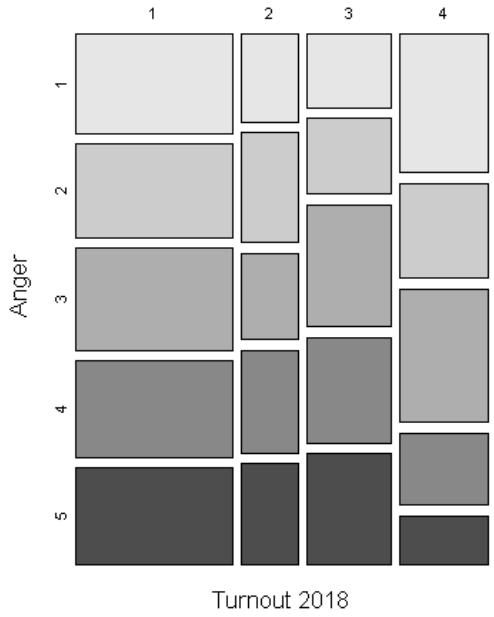
```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
|             Row Percent |
|          Column Percent |
|-------------------------|

==================================
             A$turnout16
A$geangry        1       2   Total
----------------------------------
1              346     149     495
             380.9   114.1
              69.9%   30.1%   20.7%
              18.8%   27.0%
----------------------------------
2              347     110     457
             351.7   105.3
              75.9%   24.1%   19.1%
              18.9%   20.0%
----------------------------------
3              393     151     544
             418.6   125.4
              72.2%   27.8%   22.8%
              21.4%   27.4%
----------------------------------
4              380      78     458
             352.5   105.5
              83.0%   17.0%   19.2%
              20.7%   14.2%
----------------------------------
5              374      63     437
             336.3   100.7
              85.6%   14.4%   18.3%
              20.3%   11.4%
----------------------------------
Total         1840     551    2391
                77%     23%
==================================
```

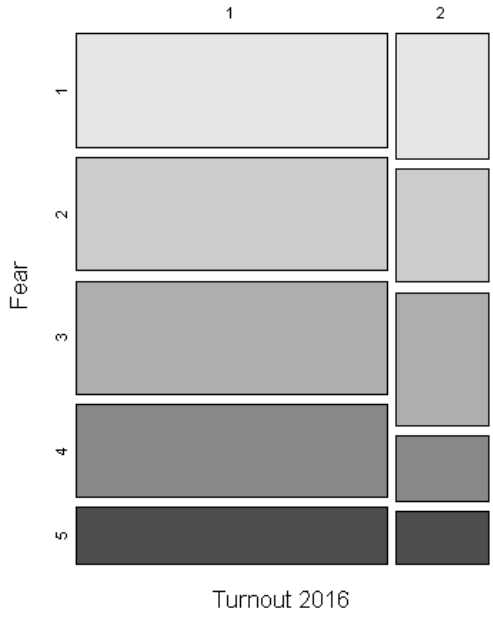
```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
|             Row Percent |
|          Column Percent |
|-------------------------|

==================================================
             A$turnout18
A$geangry        1       2       3       4   Total
--------------------------------------------------
1              198      65      79     155     497
             201.6    74.2   107.8   113.4
              39.8%   13.1%   15.9%   31.2%   20.8%
              20.5%   18.3%   15.3%   28.5%
--------------------------------------------------
2              186      80      80     106     452
             183.3    67.5    98.0   103.1
              41.2%   17.7%   17.7%   23.5%   19.0%
              19.2%   22.5%   15.5%   19.5%
--------------------------------------------------
3              201      63     129     149     542
             219.8    80.9   117.5   123.7
              37.1%   11.6%   23.8%   27.5%   22.7%
              20.8%   17.7%   25.0%   27.4%
--------------------------------------------------
4              192      75     112      80     459
             186.2    68.5    99.5   104.7
              41.8%   16.3%   24.4%   17.4%   19.3%
              19.9%   21.1%   21.7%   14.7%
--------------------------------------------------
5              190      73     117      54     434
             176.0    64.8    94.1    99.0
              43.8%   16.8%   27.0%   12.4%   18.2%
              19.6%   20.5%   22.6%    9.9%
--------------------------------------------------
Total          967     356     517     544    2384
              40.6%   14.9%   21.7%   22.8%
==================================================
```

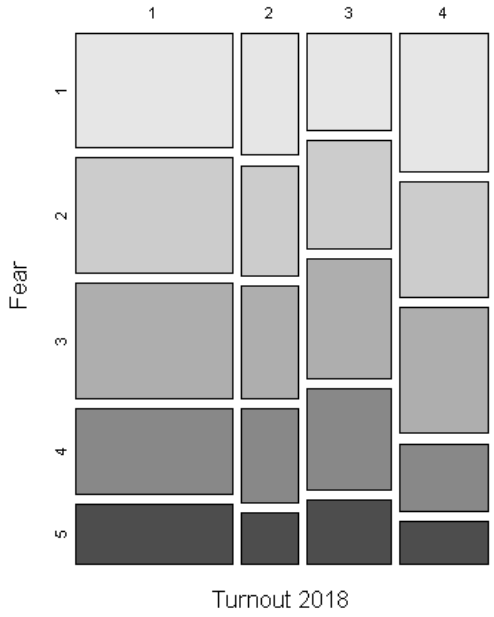
```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
|             Row Percent |
|          Column Percent |
|-------------------------|

==================================
              A$turnout16
A$geafraid        1       2   Total
----------------------------------
1              431     141     572
             440.7   131.3
              75.3%   24.7%   24.0%
              23.4%   25.7%
----------------------------------
2              425     127     552
             425.3   126.7
              77.0%   23.0%   23.1%
              23.1%   23.2%
----------------------------------
3              426     149     575
             443.0   132.0
              74.1%   25.9%   24.1%
              23.2%   27.2%
----------------------------------
4              346      73     419
             322.8    96.2
              82.6%   17.4%   17.5%
              18.8%   13.3%
----------------------------------
5              212      58     270
             208.0    62.0
              78.5%   21.5%   11.3%
              11.5%   10.6%
----------------------------------
Total         1840     548    2388
              77.1%   22.9%
==================================
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
|             Row Percent |
|          Column Percent |
|-------------------------|

==================================================
               A$turnout18
A$geafraid         1       2       3       4   Total
--------------------------------------------------
1                226      89     103     154     572
              232.3    85.3   124.0   130.4
               39.5%   15.6%   18.0%   26.9%   24.0%
               23.4%   25.1%   20.0%   28.4%
--------------------------------------------------
2                229      80     114     128     551
              223.8    82.2   119.4   125.7
               41.6%   14.5%   20.7%   23.2%   23.1%
               23.7%   22.5%   22.1%   23.6%
--------------------------------------------------
3                226      81     126     140     573
              232.7    85.4   124.2   130.7
               39.4%   14.1%   22.0%   24.4%   24.1%
               23.4%   22.8%   24.4%   25.8%
--------------------------------------------------
4                169      68     106      74     417
              169.4    62.2    90.4    95.1
               40.5%   16.3%   25.4%   17.7%   17.5%
               17.5%   19.2%   20.5%   13.6%
--------------------------------------------------
5                117      37      67      47     268
              108.8    40.0    58.1    61.1
               43.7%   13.8%   25.0%   17.5%   11.3%
               12.1%   10.4%   13.0%    8.7%
--------------------------------------------------
Total            967     355     516     543    2381
               40.6%   14.9%   21.7%   22.8%
==================================================
```









- Sanity checks

In [32]:
```
length(A$geangry) == length(A$geafraid)
```

TRUE

In [33]:
```
length(A$turnout16) == length(A$turnout18)
```

TRUE

In [34]:
```
sum((is.na(A$geangry))) # any NA's?
sum((is.na(A$geafraid))) # any NA's?
sum((is.na(A$turnout16))) # any NA's?
sum((is.na(A$turnout18))) # any NA's?
```

3

6

107

114

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

Explain why your test is the most appropriate choice, focusing on its statistical assumptions.

H_0: anger and fear were equally effective in driving voter turnout in 2016 and 2018 elections.

- If the data were continuous instead of ordinal we could use a t-test to test the two H_0: anger and fear were equally effective in driving voter turnout for each year, 2016 and 2018.
- Similarly, if these data were continuous, we could use regression to relate voter turnout for each year to anger and fear either with one multiple linear regression for each year or with two OLS models for each year. Because the data are oridinal t-tests and or regression is not appropriate.
- We could use a Wilcoxson rank sum test to test the H_0 the ranks of angry and fearful voters are not different for each election year but (i) some voters are both fearful and afraid (see the plots above),(ii) this does not address the hypothesis of interest and (iii) if we collapsed the responses to [geangry] and [geafraid] into binary variables that represented angry/fearful (2-5) and not angry/fearfull (1) and ran the test we would introduce a drastic difference in sample sizes between angry and not angry voters since the later is only represented by 1 response while the former is represented by 4.
- Finally, although Pearson's χ^2 test for independence of two categorical variables tests a broader H_0 the two variables are independent, this approach allows us to answer our desired question, and more with the null hypotheses tested below.

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

- H_0 : **Voter anger is independent of voter turnout in 2016**

In [35]:
```
CrossTable(A$geangry,A$turnout16,
           prop.r=F, prop.c=F, expected=T,
           chisq=T,
           prop.chisq=F, prop.t=T)
```

```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
|         N / Table Total |
|-------------------------|

==================================
             A$turnout16
A$geangry        1       2   Total
----------------------------------
1              346     149     495
             380.9   114.1
             0.145   0.062
----------------------------------
2              347     110     457
             351.7   105.3
             0.145   0.046
----------------------------------
3              393     151     544
             418.6   125.4
             0.164   0.063
----------------------------------
4              380      78     458
             352.5   105.5
             0.159   0.033
----------------------------------
5              374      63     437
             336.3   100.7
             0.156   0.026
----------------------------------
Total         1840     551    2391
==================================

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 = 48.66756      d.f. = 4      p = 6.85e-10
```

- Practical significance:

In [36]:
```
ES.chisq.assoc(chisq=48.66756,
               n=2391,
               df=4,
               p=6.849734e-10,
               mindf=1)
```

```
     effect size of chi-squared test of association

           phi = 0.1426691
         chisq = 48.66756
             p = 6.849733e-10
             n = 2391
            df = 4
         mindf = 1

NOTE: small effect size:  phi = 0.1
      medium effect size: phi = 0.3
      large effect size:  phi = 0.5
```

- H_0 : **Voter anger is independent of voter turnout in 2018**

In [37]:
```
CrossTable(A$geangry,A$turnout18,
           prop.r=F, prop.c=F, expected=T,
           chisq=T,
           prop.chisq=F, prop.t=T)
```

```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
|         N / Table Total |
|-------------------------|

=================================================
             A$turnout18
A$geangry        1       2       3       4   Total
-------------------------------------------------
1              198      65      79     155     497
             201.6    74.2   107.8   113.4
             0.083   0.027   0.033   0.065
-------------------------------------------------
2              186      80      80     106     452
             183.3    67.5    98.0   103.1
             0.078   0.034   0.034   0.044
-------------------------------------------------
3              201      63     129     149     542
             219.8    80.9   117.5   123.7
             0.084   0.026   0.054   0.062
-------------------------------------------------
4              192      75     112      80     459
             186.2    68.5    99.5   104.7
             0.081   0.031   0.047   0.034
-------------------------------------------------
5              190      73     117      54     434
             176.0    64.8    94.1    99.0
             0.080   0.031   0.049   0.023
-------------------------------------------------
Total          967     356     517     544    2384
=================================================

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 = 78.16281      d.f. = 12      p = 9.24e-12
```

- Practical significance

In [38]:
```
ES.chisq.assoc(chisq=78.16281,
               n=2384,
               df=12,
               p=9.237008e-12 ,
               mindf=3)
```

```
     effect size of chi-squared test of association

           phi = 0.1045409
         chisq = 78.16281
             p = 9.237014e-12
             n = 2384
            df = 12
         mindf = 3

NOTE: small effect size:  phi = 0.173205080756888
      medium effect size: phi = 0.519615242270663
      large effect size:  phi = 0.866025403784439
```

- H_0 : **Voter fear is independent of voter turnout in 2016**

In [39]:
```
CrossTable(A$geafraid,A$turnout16,
           prop.r=F, prop.c=F, expected=T,
           chisq=T,
           prop.chisq=F, prop.t=T)
```

```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
|         N / Table Total |
|-------------------------|

==================================
             A$turnout16
A$geafraid        1       2   Total
----------------------------------
1               431     141     572
             440.7   131.3
             0.180   0.059
----------------------------------
2               425     127     552
             425.3   126.7
             0.178   0.053
----------------------------------
3               426     149     575
             443.0   132.0
             0.178   0.062
----------------------------------
4               346      73     419
             322.8    96.2
             0.145   0.031
----------------------------------
5               212      58     270
             208.0    62.0
             0.089   0.024
----------------------------------
Total          1840     548    2388
==================================

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 = 11.36088      d.f. = 4      p = 0.0228
```

- Practical significance

In [40]:
```
ES.chisq.assoc(chisq=11.36088,
               n=2388,
               df=4,
               p=0.02279404,
               mindf=1)
```

```
     effect size of chi-squared test of association

           phi = 0.06897454
         chisq = 11.36088
             p = 0.02279404
             n = 2388
            df = 4
         mindf = 1

NOTE: small effect size:  phi = 0.1
      medium effect size: phi = 0.3
      large effect size:  phi = 0.5
```

- H_0 : **Voter fear is independent of voter turnout in 2018**

In [41]:
```
CrossTable(A$geafraid,A$turnout18,
           prop.r=F, prop.c=F, expected=T,
           chisq=T,
           prop.chisq=F, prop.t=T)
```

```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
|         N / Table Total |
|-------------------------|

=================================================
               A$turnout18
A$geafraid         1       2       3       4   Total
-------------------------------------------------
1                226      89     103     154     572
               232.3    85.3   124.0   130.4
               0.095   0.037   0.043   0.065
-------------------------------------------------
2                229      80     114     128     551
               223.8    82.2   119.4   125.7
               0.096   0.034   0.048   0.054
-------------------------------------------------
3                226      81     126     140     573
               232.7    85.4   124.2   130.7
               0.095   0.034   0.053   0.059
-------------------------------------------------
4                169      68     106      74     417
               169.4    62.2    90.4    95.1
               0.071   0.029   0.045   0.031
-------------------------------------------------
5                117      37      67      47     268
               108.8    40.0    58.1    61.1
               0.049   0.016   0.028   0.020
-------------------------------------------------
Total            967     355     516     543    2381
=================================================

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 = 23.1056      d.f. = 12      p = 0.0268
```

- Practical significance

In [42]:
```
ES.chisq.assoc(chisq=11.36088,
               n=2388,
               df=4,
               p=0.02279404,
               mindf=1)
```

```
     effect size of chi-squared test of association

           phi = 0.06897454
         chisq = 11.36088
             p = 0.02279404
             n = 2388
            df = 4
         mindf = 1

NOTE: small effect size:  phi = 0.1
      medium effect size: phi = 0.3
      large effect size:  phi = 0.5
```

⋆ Reject all four H_0 voter fear and anger were independent of voter turnout in 2016 and 2018 since all four p-values were $\leq 0.02 < \alpha = 0.05$. This evidence supports H_A: voter anger and fear were not independent of YouGov sampled voter turnout in 2016 and 2018. Despite these low p-values, this effect sizes ϕ observed were all small, $\approx$ < 0.1. Moreover, the low p-values might be especially low because of the large sample sizes ($n \geq$2384).Finally, in response to the initial question:"Was anger or fear more effective at driving increases in voter turnout from 2016 to 2018?" we can say that, since p=0.02 < p<0.0001, anger was probably more of a motivating emotion and mobilized the population of YouGov voters sampled here for both 2016 and 2018. That said, both anger and fear are not independent of voter turnout in 2016 and 2018.

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)

In words, clearly state your research question and argue why it is important for understanding the recent voting behavior. Explain it as if you were presenting to an audience that includes technical and non technical members.

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

Study question: Does eduction background affect the election choice between Donald Trump and Hillary Clinton?

Importance for understanding the recent voting behavior

Education level is one of the key factors which affects the decision of making choices. Therefore, it is important to investigate whether education background has any effect on voting Donald Trump and Hillary Clinton in 2016 Presidential election.

Variables

To investigate whether education background has the effect on election choices of Donald Trump and Hillary Clinton. Two variables in ANES are important. First variable is [educ] which contains the education of respondents. Second variable is [vote16] which contains the election choice of respondents. The survey quesitons which are in the ANES are listed below

- [educ] Education: (1) No HS (2) High school graduate (3) Some college (4) 2-year (5) 4-year (6) Post-grad
- [vote16] In the 2016 presidential election, who did you vote for? (-7) No Answer (-1) inapplicable, legitimate skip (1) Donald Trump (2) Hillary Clinton (3) someone else

Study approach

The study only foucs the respondents who voted Donald Trump and Hillary Clinton. Therefore, the respondents who chose (1) Donald Trump and (2) Hillary Clinton in [vote16] variable were selected. Next, the [educ] and subset [vote16] were tested for independence between two categorical variables.

Study plan

1. Perform EDA
2. Select the test hypothesis
3. Create the contingency table for displaying the number of observations of each possible outcome
4. Compute the test statistics to understand statistical significance and practical significance

Perform EDA and select your hypothesis test (5 points)

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

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

Based on your EDA, select an appropriate hypothesis test. Explain why your test is the most appropriate choice, focusing on its statistical assumptions.

Exploratory data analysis (EDA)

In [43]:
```
# check the format of the data
paste("Education background [educ]")
head(A$educ)

paste("Voting choices [vote16]")
head(A$vote16)


# check two variables with table
paste("Table of two variable")
table(A$vote16, A$educ)
```

'Education background [educ]'

4 6 2 3 3 1

'Voting choices [vote16]'

1 -1 -1 -1 1 2

'Table of two variable'

```
     1   2   3   4   5   6
-7   0   1   1   1   0   0
-1 108 326 100  34  46  16
1   20 148 212 112 167 111
2   10 129 209  96 258 196
3    7  23  48  32  59  30
```

In [44]:
```
# subset the respondents who either vote Donald Trump or Hillary Clinton
vote_choice <- subset(A, vote16 == 1 | vote16 == 2)

# check subset data
table(vote_choice$vote16, vote_choice$educ)

# check the summary statistics of the education background of respondents
paste("Whole population")
summary(vote_choice$educ)

paste("Vote Hillary Clinton")
summary(vote_choice$educ[vote_choice$vote16 == 2])

paste("Vote Donald Trump")
summary(vote_choice$educ[vote_choice$vote16 == 1])
```

```
   1   2   3   4   5   6
1 20 148 212 112 167 111
2 10 129 209  96 258 196
```

'Whole population'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   3.984   5.000   6.000
```

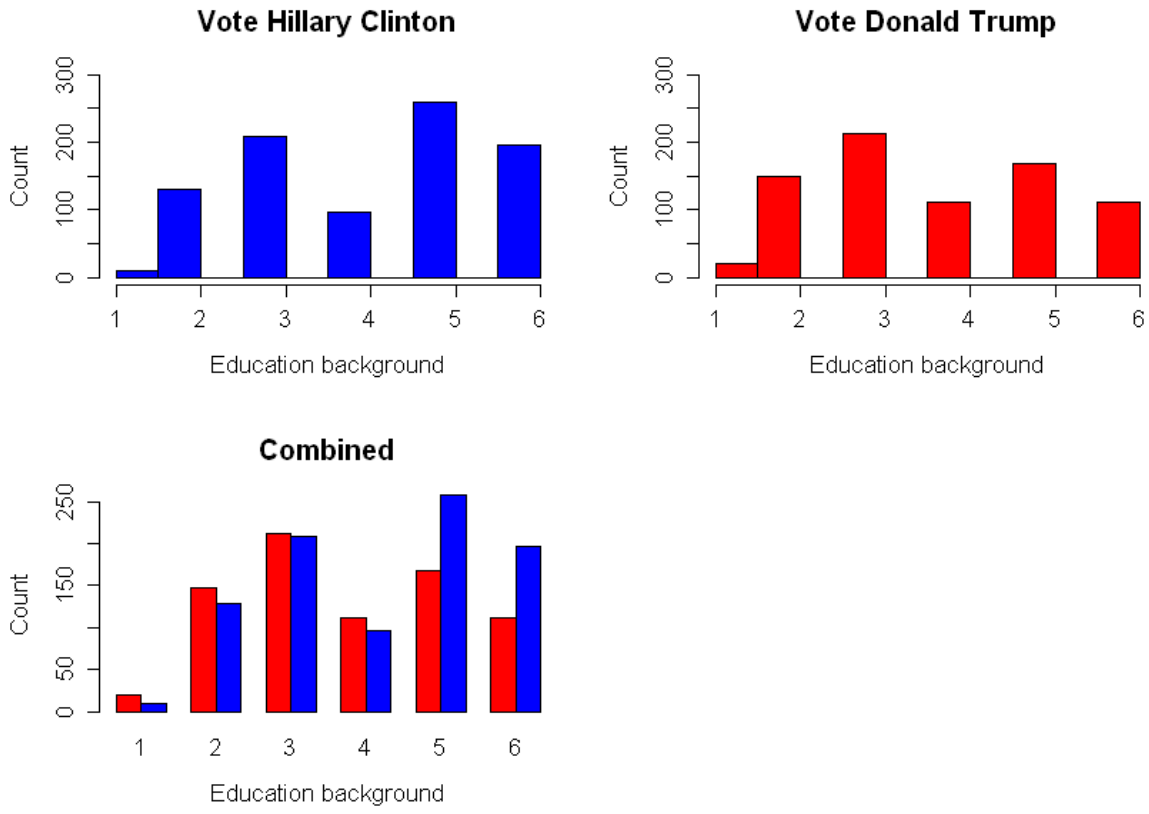
'Vote Hillary Clinton'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.00    5.00    4.17    5.00    6.00
```

'Vote Donald Trump'

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   3.768   5.000   6.000
```

In [45]:
```
# define figure size
options(repr.plot.height = 5, repr.plot.width = 7, repr.plot.pointsize = 10)

# plot the education background between two candidates
par(mfrow=c(2,2))

hist(vote_choice$educ[vote_choice$vote16 == 2], col = 'blue', ylim = c(0, 300
), main = "Vote Hillary Clinton",
    xlab = "Education background", ylab = "Count")
hist(vote_choice$educ[vote_choice$vote16 == 1], col = 'red', ylim = c(0, 300),
main = "Vote Donald Trump",
    xlab = "Education background", ylab = "Count")


results_table <- table(vote_choice$vote16, vote_choice$educ)
barplot(results_table, main="Combined",
  xlab="Education background", ylab = "Count", col=c("red","blue")
       ,beside=TRUE)
```



Summary of EDA

- Base on sanity check, no missing data is observed in the subset data.
- The results showed that the variables of education and voting choices are categorical variables.
- Based on the EDA, no obvious skewness is observed.

***Hypothesis**

The study is to investigate whether education background has the effect on election choices of Donald Trump and Hillary Clinton in 2016 Presidential election. After EDA, we understand the variable of education is a categorical variable. The EDA furhter suggested that it wouldn't be proper to perform the average of education variable and compare the their difference between voting Hillary Clinton and Donald Trump using t-test. Therefore, it would be a better choice to use a chi-squared test for independence between categorical variables. The null hypothesis and alternative hypothesis are listed below.

- H_o: Voting choice between Hillary Clinton and Donald Trump and education level are independent.
- H_A: Voting choice between Hillary Clinton and Donald Trump and education level are not independent.

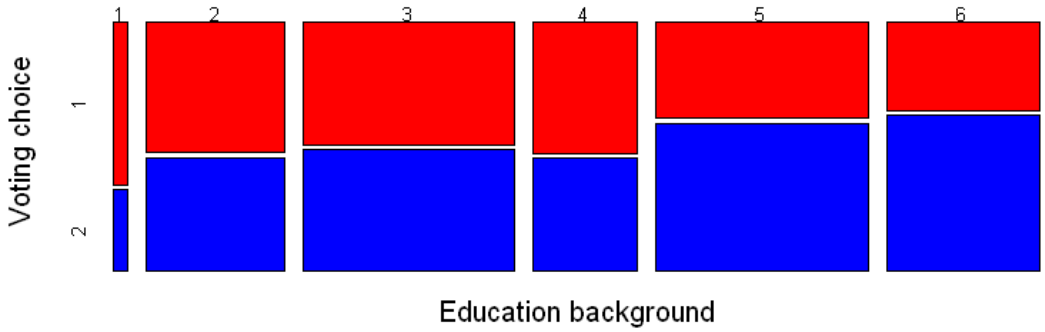
Conduct your test. (2 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result.

Statistical significance - Pearson's chi-squared test

In [46]:
```
# define figure size
options(repr.plot.height = 3, repr.plot.width = 7, repr.plot.pointsize = 10)

# perform chi-square test
crosstab(vote_choice$vote16, vote_choice$educ, chisq=T, xlab = "Education back
ground",
         ylab = "Voting choice", col = c('red', 'blue'), prop.t = TRUE, expect
ed=TRUE)
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
|           Total Percent |
|-------------------------|

=============================================================================
                      vote_choice$educ
vote_choice$vote16       1       2       3       4       5       6   Total
-----------------------------------------------------------------------------
1                       20     148     212     112     167     111     770
                      13.8   127.9   194.3    96.0   196.2   141.7
                       1.2%    8.9%   12.7%    6.7%   10.0%    6.7%
-----------------------------------------------------------------------------
2                       10     129     209      96     258     196     898
                      16.2   149.1   226.7   112.0   228.8   165.3
                       0.6%    7.7%   12.5%    5.8%   15.5%   11.8%
-----------------------------------------------------------------------------
Total                   30     277     421     208     425     307    1668
=============================================================================

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 = 39.31662      d.f. = 5      p = 2.05e-07

        Minimum expected frequency: 13.84892
```



Practical significance - effect size

In [47]:
```
# create table
vote_edu_table = table(vote_choice$vote16, vote_choice$educ)

# calculate effect size
ES.chisq.assoc(ct=vote_edu_table)
```

```
     effect size of chi-squared test of association

           phi = 0.1535289
         chisq = 39.31662
             p = 2.050621e-07
             n = 1668
            df = 5
         mindf = 1

NOTE: small effect size:  phi = 0.1
      medium effect size: phi = 0.3
      large effect size:  phi = 0.5
```

Summary of the test

- Based on the test statistic alone, we obatined $p = 2.05e-07$ and could reject the null hypothesis of independece at a 95% confidence interval. Therefore, the test results showed statistical significance.
- The counts of each group has larger or equal to 30 , indicating that we should have sufficient observations to apply the test and therefore to evaluate the hypothesis in this data.
- Based on the test of effect size, we obatined ϕ = 0.1535289 which is close to a small effect size ($\phi = 0.1$). Therefore, the test results did not show strong practical significance.

Conclusion (3 points)

Clearly state the conclusion of your hypothesis test and how it relates to your research question.

Finally, briefly present your conclusion in words as if you were presenting to an audience that includes technical and non technical members.

- In this study, the statistical analysis showed the voting choice between Hillary Clinton and Donald Trump and education levels are not independent and statistically significant. The results suggested that education has effects on voting choice between Hillary Clinton and Donald Trump based on the statistical standpoint. However, based on the measure of effect size, the difference is too small to have pratical implication.

Appendix

Alternative analyses for Question 3

> - This was not presented because of the reasons described above. Note it does provide additional evidence that anger was more important than fear in mobilizing YouGov sampled voters to turnout in 2016 and 2018.

- Polarize anger responses (1-5) into anger (2-5) and no anger (1)

In [49]:
```
A$anger = factor(ifelse(A$geangry > A$geangry[A$geangry == "1"],
                        "Angry", "Not angry"))
```

```
Warning message in A$geangry > A$geangry[A$geangry == "1"]:
"longer object length is not a multiple of shorter object length"
```

- Statistical significance

In [50]:
```
wilcox.test(A$turnout16~A$anger)
```

```
        Wilcoxon rank sum test with continuity correction

data:  A$turnout16 by A$anger
W = 422990, p-value = 2.35e-05
alternative hypothesis: true location shift is not equal to 0
```

- Practical significance

In [51]:
```
cohen.d(A$turnout16~A$anger)
```

```
Cohen's d

d estimate: -0.2146679 (small)
95 percent confidence interval:
     lower      upper
-0.3140465 -0.1152894
```

Anger vs turnout in 2018

- Statistical significance

In [52]:
```
wilcox.test(A$turnout18~A$anger)
```

```
        Wilcoxon rank sum test with continuity correction

data:  A$turnout18 by A$anger
W = 434270, p-value = 0.01904
alternative hypothesis: true location shift is not equal to 0
```

- Practical significance

In [53]:
```
cohen.d(A$turnout18~A$anger)
```

```
Cohen's d

d estimate: -0.1243723 (negligible)
95 percent confidence interval:
      lower       upper
-0.22351707 -0.02522758
```

Fear vs turnout in 2016

- Polarize fear responses (1-5) into fearful (2-5) and no fear (1)

In [54]:
```
A$fear = factor(ifelse(A$geafraid > A$geafraid[A$geafraid == "1"],
                       "Fearfull", "Not fearfull"))
```

```
Warning message in A$geafraid > A$geafraid[A$geafraid == "1"]:
"longer object length is not a multiple of shorter object length"
```

- Statistical significance

In [55]:
```
wilcox.test(A$turnout16~A$fear)
```

```
        Wilcoxon rank sum test with continuity correction

data:  A$turnout16 by A$fear
W = 496050, p-value = 0.2057
alternative hypothesis: true location shift is not equal to 0
```

- Practical significance

In [56]:
```
cohen.d(A$turnout16~A$fear)
```

```
Cohen's d

d estimate: -0.06100458 (negligible)
95 percent confidence interval:
      lower       upper
-0.15552724  0.03351807
```

Fear vs turnout in 2018

In [57]:
```
wilcox.test(A$turnout18~A$fear)
```

```
        Wilcoxon rank sum test with continuity correction

data:  A$turnout18 by A$fear
W = 487390, p-value = 0.1422
alternative hypothesis: true location shift is not equal to 0
```

- Practical Significance

In [58]:
```
cohen.d(A$turnout18~A$fear)
```

```
Cohen's d

d estimate: -0.07022724 (negligible)
95 percent confidence interval:
      lower       upper
-0.16479946  0.02434498
```

In [ ]:
```

```