Section 1: Spyros Garyfallos, Paul Petit, Ross MacLean

In [2]:
```
A = read.csv(file='anes_pilot_2018.csv')
```

Research Questions

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

Introduce your topic briefly. (5 points)

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 variables ftploice and ftjournal can be used to infer whether the YouGov population have more respect for the police or for journalists. These variables have been operationalized using a feeling thermometer that records the amount of favorable/unfavorable feelings respondants have towards each of these groups, on a scale ranging from 0-100. A score of 0 reflects the most unfavorable feelings, 50 reflects no feeling at all (neutral/indifferent) and 100 reflects the most favorable feelings toward the group in question. Ftpolice and ftjournal are therefore discrete random variables on a 100-point scale.

One limiting factor is that these variables record the amount of favorable/unfavorable feeling towards certain groups, rather than the amount of respect per se. For the purpose of this analyis however, the definitions of favorable/unfavorable and respect/disrespect will be considered equivalent.

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 [8]:
```
# Install packages
install.packages('effsize')
install.packages('ggpubr')
library(effsize)
library('ggpubr')
library('ggplot2')

# Resize the plots
ratio = 0.75
width = 4
options(repr.plot.width=4, repr.plot.height=3)
```

In [21]:
```
# Check ftpolice values
paste(c('Min police values: ', head(sort(unique(A$ftpolice), decreasing=F),5
)))
paste(c('Max police values: ', tail(sort(unique(A$ftpolice), decreasing=F),5
)))

# Check ftjournal values
paste(c('Min journalist values: ', head(sort(unique(A$ftjournal), decreasing=F
),5)))
paste(c('Min journalist values: ', tail(sort(unique(A$ftjournal), decreasing=F
),5)))

# Replace non-response with NA
for(i in c(-1, -4, -7))
{A$ftjournal <- replace(A$ftjournal, A$ftjournal == i, NA)}

# Remove cases with NA in 'ftpolice' or 'ftjournal
A2 <- subset(A, select=c('ftpolice', 'ftjournal'))
A2 <- A2[complete.cases(A2), ]

# Final sample size
paste(c('Final sample size:', length(A2$ftpolice)))
```

'Min police values: ' '0' '1' '2' '3' '4'

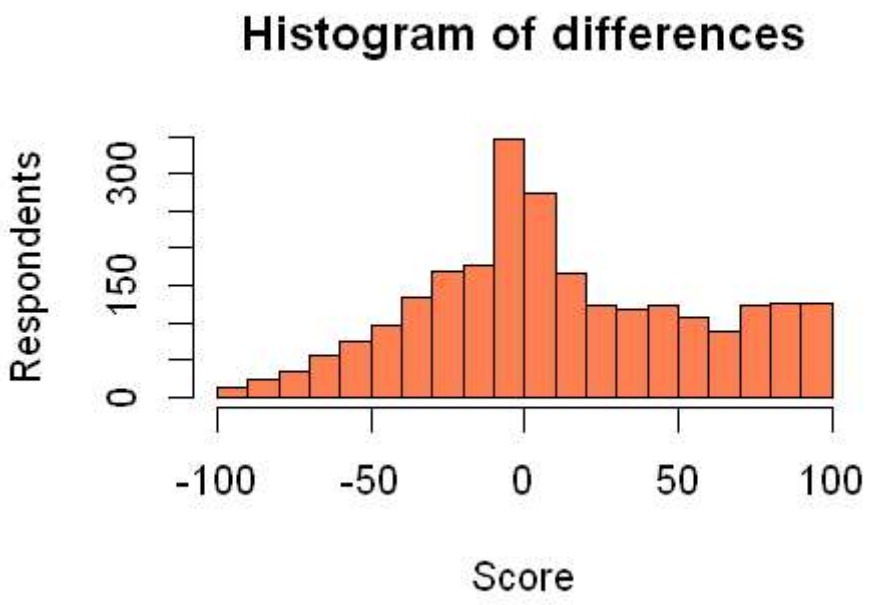
'Max police values: ' '96' '97' '98' '99' '100'

'Min journalist values: ' '0' '1' '2' '3' '4'

'Min journalist values: ' '96' '97' '98' '99' '100'

'Final sample size:' '2498'

In [23]:
```
# Histogram of differences
hist(A2$ftpolice - A2$ftjournal, breaks=20, col = 'coral',
     main = 'Histogram of differences',
     xlab = 'Score', ylab = 'Respondents')
```



The valid values in ftpolice and ftjournal are intergers within the range 0-100. Therefore any negative values indicate non-response (as confirmed by the ANES documentation). A quick check of minimum values identified where these non-response values should be replaced with NA. A subset of the data set was then taken to ensure only complete cases were retained in subsequent analysis (n=2498). The above histogram shows that the distribution of the differences of ftpolice and ftjournals is non-normal with a heavy tail. The distribution is relatively symmetrical, peaking around zero. These factors indicate that there is an underlying relationship between respondents favorability scores towards police and journalists.

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.

The above histogram shows that the distribution of the differences of ftpolice and ftjournals is non-normal with a heavy right hand tail. The sample size is large (n=2498) and so the Central Limit Theorem can be invoked, suggesting a t-test will be appropriate as the sampling distribution of the differences will be approximately normal. The two samples consist of responses from the same respondents and so to minimize the amount of intersubject variability in our test, a paired two-sample t-test has been deemed most appropropriate.

The null hypothesis (H_0) is that there is no difference between the favorability scores towards police and journalists. The alternate hypothesis (H_a) is that there is a difference between the favorability scores towards police and journalists.

$$\text{Null hypothesis:} \quad H_0 : \mu_1 - \mu_2 = \Delta_0$$
$$\text{Alternate hypothesis:} \quad H_a : \mu_1 - \mu_2 \neq \Delta_0$$

While the orignal question ask whether "voters have more respect for the police than journalists", a non-directional alternate hypothesis was selected because it is quite possible that the mean favorability score for journalists is greater than that for the police, therefore we must test for statistical significance in both directions. In the event of being able to reject the null hypothesis, the mean difference in respondents' scores will indicate the direction of this difference. A significance level of 0.05 has been selected for this test.

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 [11]:
```
# Two sample t-test (paired)
t.test(A2$ftpolice, A2$ftjournal, paired=TRUE)

# Practical effect size, Cohen's D
cohen.d(A2$ftpolice, A2$ftjournal, paired=TRUE)
```

```
        Paired t-test

data:  A2$ftpolice and A2$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:
      inf       sup
0.2186004 0.3300646
```

The test was found to be statistically significant, t = 13.711 and p < 2.2e-16, at the p < 0.05 significance level. This means we can reject the null hypothesis that there is no difference between the favorability scores for police and journalists. The mean (paired) difference of favorability scores was found to be 12.35 'degrees' warmer (using the feeling themometer terminology) for police than for journalists. This result should however be viewed in light of the small practical effect size obtained using Cohen's D (d = 0.274).

The large sample size (n = 2498) may have contributed to the highly significant result. As the sample size increases, so too does the precision of sample estimates and the power of the test (through smaller margins of error), which can lead to an increase in the number of significant results. It is important to therefore consider both the p-value and the practical effect size when reporting results.

In conclusion, we can reject the null hypothesis that there is no difference in the amount of respect US voters have for police than for journalists. Police received greater favorability/respect scores than journalists but the difference was not found to be practically significant. Therefore the test supports the notion that there is not meaningful difference in the amount of respect US voters have for the police and for journalists.

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.

A key consideration when assessing how best to answer the question "are Republican voters older or younger than Democratic voters?" is determining which variables allows us to best define respondents as either Republican or Democratic voters. Following an assessment of potentially suitable variables, the variables of 'pid1d', 'pid1r' and 'birthyr' were identified as allowing us to best answer this question.

The ANES survey includes several questions which ask voters how they voted in different elections (e.g. U.S House, U.S Senate and the 2016 Presidential Election) however complications arise when categorizing voters as either Republican or Democratic in the event that they voted for different parties during these various elections. For this reason, questions asking more generally about which political party voters identify with, have been considered the best indicator of political party allegiance ('pid1d' and 'pid1r'). Pid1d and pid1r are both categorical variables.

Questions 'pid1d' and 'pid1r' both pose the same question to subjects however the ordering of responses is different - 'pid1d' presents the responses in the order of 'Democratic', 'Republican', 'independent', 'something else'. Question 'pid1r' flips the ordering of 'Democratic' and 'Republican', with 'Republican' presented first. Based on the ANES documentation, questions 'pid1d' and 'pid1r' were presented randomly to subjects, presumably to minimize the potential for primacy bias (the tendency for respondents to select the first option presented to them).

The variable 'birthyr' contains the birth year of respondents. It is a discrete random variable and can be used in combination with the survey year (2018) to calculate the age of subjects.

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 [14]:
```
# Check pid1d and pid1r values
paste(c('Check pid1d values: ', sort(unique(A$pid1d), decreasing=F)))
paste(c('Check pid1r values: ', sort(unique(A$pid1r), decreasing=F)))
A$pid <- ifelse(A$pid1d >= 1, A$pid1d, A$pid1r)

# Replace pid non-response with 0
for(i in c(-1, -4, -7))
{A$pid <- replace(A$pid, A$pid == i, NA)}

paste('NAs removed from pid')
paste('Total valid responses: ', length(A$pid[!is.na(A$pid)]))
paste('Final sample: ', length(A$pid[A$pid %in% c(1, 2)]))

# Check birthyear values
paste(c('Min birth years: ', head(sort(unique(A$birthyr), decreasing=F), 5)))
paste(c('Max birth years: ', tail(sort(unique(A$birthyr), decreasing=F), 5)))

# Create Age variable
A$age <- as.numeric(2018 - A$birthyr)

# Check range of ages
paste(c('Min ages: ', head(sort(unique(A$age), decreasing=F), 5)))
paste(c('Max ages: ', tail(sort(unique(A$age), decreasing=F), 5)))
```

'Check pid1d values: ' '-7' '-1' '1' '2' '3' '4'

'Check pid1r values: ' '-7' '-1' '1' '2' '3' '4'

'NAs removed from pid'

'Total valid responses: 2350'

'Final sample: 1466'

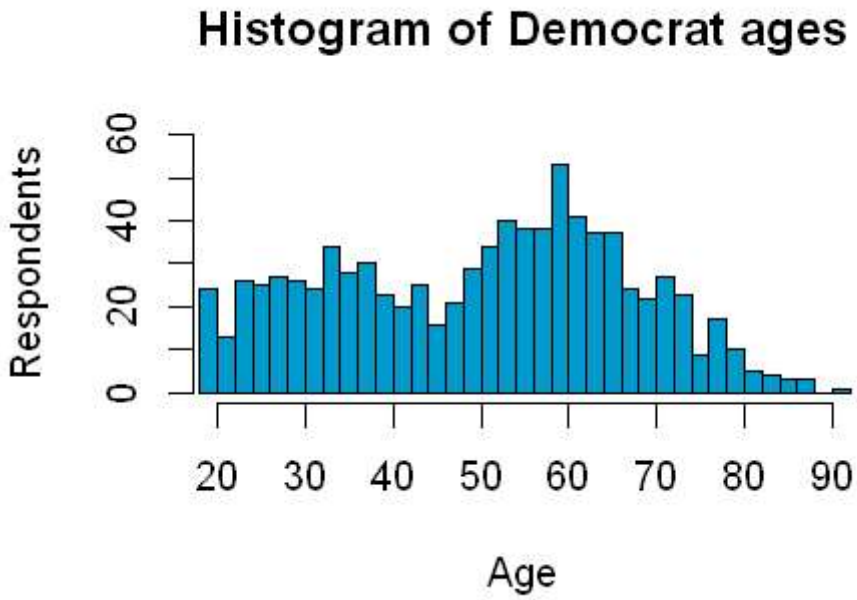
'Min birth years: ' '1927' '1928' '1929' '1930' '1931'

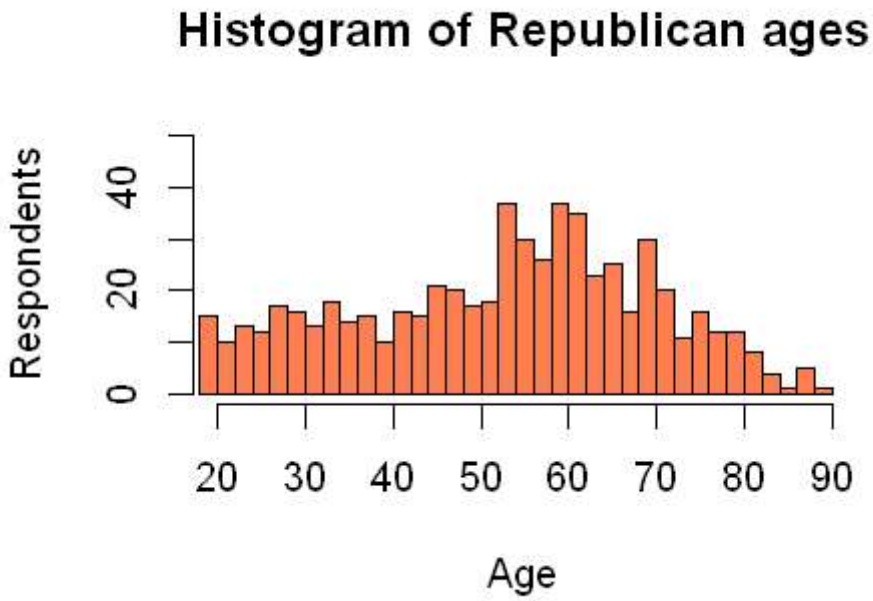
'Max birth years: ' '1996' '1997' '1998' '1999' '2000'

'Min ages: ' '18' '19' '20' '21' '22'

'Max ages: ' '87' '88' '89' '90' '91'

In [24]:
```
# Histogram of Democrat ages
hist(A$age[A$pid == 1], breaks=40, col = 'deepskyblue3',
     main = 'Histogram of Democrat ages',
     xlim = c(20,90), xlab = 'Age',
     ylim = c(0,60), ylab = 'Respondents')

# Histogram of Republican ages
hist(A$age[A$pid == 2], breaks=40, col = 'coral',
     main = 'Histogram of Republican ages',
     xlim = c(20,90), xlab = 'Age',
     ylim = c(0,50), ylab = 'Respondents')
```





As respondents were randomly assigned either question 'pid1d', 'pid1r' it is necessary to combine the responses into a single variable which can then be used to determine the political party allgience of the entire sample. Once this was completed, 150 responses were classified as non-response. A further 884 respondents identified with parties other than the Democratic or Republican party and so were excluded from subsequent analysis. The removal of this data was deemed warranted because it does not pertain to the question we are seeking to answer. This left 1466 respondents in our final sample that identified with either the Democratic or Republican party.

The full range of birth years (and corresponding ages) were assessed to ensure all values look plausible given the nature of the sample data. Ages range from 18-91 years which makes sense given that the minimum voting age in the U.S. is 18 years.

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.

The underlying distribution of Republican voter ages and Democractic voters ages are somewhat non-normal but the sample size is large (n=1466) and so the Central Limit Theorem can be invoked. Given the aforementioned characteristics the sampling distribution of Republican voter ages and Democractic voters ages will be approximately normally distributed, indicating that a t-test will be appropriate. Respondents could only identify with either the Democractic or Republican party. The two samples are therefore distinct and an unpaired t-test will be selected.

The null hypothesis (H_0) is that there is no difference between the ages of Democractic voters and Republican voters. The alternate hypothesis (H_a) is that there is a difference between the ages of Democractic voters and Republican voters.

$$\begin{aligned} \text{Null hypothesis:} \quad & H_0 : \mu_1 - \mu_2 = \Delta_0 \\ \text{Alternate hypothesis:} \quad & H_a : \mu_1 - \mu_2 \neq \Delta_0 \end{aligned}$$

The question that we are seeking to answer is "are Republican voters older or younger than Democratic voters?". A non-directional alternate hypothesis has once again been selected. It is quite plausible for the Republicans voters to either be older or younger than Democractic voters, and so it is necessary to test to for significance in both directions. In the event of being able to reject the null hypothesis, the mean ages of Republican voters and Democractic voters can be compared, allowing us to directly answer the question posed.

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 [16]:
```
# Democrats
d <- A$age[A$pid %in% c(1)]

# Republicans
r <- A$age[A$pid %in% c(2)]

# Two sample t-test (unpaired)
t.test(d, r, paired=FALSE)

# Practical effect size, Cohen's D
cohen.d(d, r, paired=FALSE, na.rm=TRUE)
```

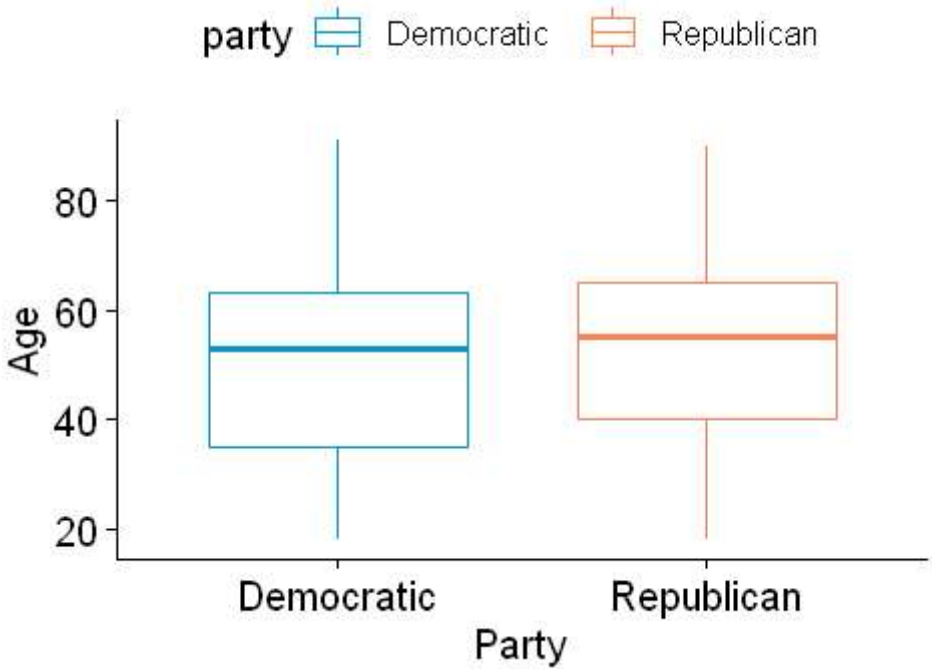
```
        Welch Two Sample t-test

data:  d and r
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:
        inf         sup
-0.25987016 -0.05163986
```

The test was found to be statistically significant, t = -2.939 and p = 0.003, at the p < 0.05 significance level. The means we can reject the null hypothesis that there is no difference between the ages of Democractic voters and Republican voters. The effect size for this analysis (d = 0.156) was found to be negligible (small), based on Cohen's convention for effect sizes. The mean of age of Democratic voters ($\bar{x}$ = 50.23) is less than the mean age of Republican voters ($\bar{y}$ = 52.86).

Once again the large sample size (n = 1466) will have contributed to this highly significant result by essentially reducing the margin of error within the sample, and therefore increasing the likelihood of detecting statistically significant differences in the mean ages of voters in each sample. The significant result should therefore be viewed in light of the practical effect size to understand if the difference is truly meaningful.

In conclusion, we can reject the null hypothesis that there is no difference in the ages of Republican voters and Democratic voters. While Republican voters were found to be older than Democractic voters, the difference was not practically significant and so observed difference in ages is not meaningful. Intuitively, this is confirmed by the following box plot showing voter ages by party.

In [17]:
```
# Select Democrats and Republicans
A3 <- A[A$pid %in% c(1, 2), ]
A3$party <- ifelse(A3$pid == 1, 'Democratic', 'Republican')

# Plot boxplot
ggboxplot(A3, x = 'party', y = 'age',
          color = 'party', palette = c('deepskyblue3', 'coral'),
          order = c('Democratic', 'Republican'),
          xlab = 'Party', ylab = 'Age',
          shape = 'party')
```



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

Introduce your topic briefly. (5 points)

To answer this question, we'll refer to responses to the russia_16 question that asks, "Do you think the Russian government probably interfered in the 2016 presidential election?"

While this variable in our dataset is our best operationalized option, it's not perfect. It's conceivable that a respondent to our survey or a voter in America would say that federal investigations into Russian interference were baseless or were not baseless regardless of whether or not they believed Russia interfered with the election. In other words, the semantics of the question make the focus of the question on federal investigation's baselessness and not Russian interference. A voter might think that Russia did indeed interfere with the election, but that the federal investigations were baseless because, for example, the federal investigations shouldn't have been conducted by the government. Conversely, a voter might think that Russia didn't interfere with the election but that the investigations weren't baseless because it was the federal government's job to confirm that Russia didn't interfere.

As we did above, we'll use responses to the pid1d and pid1r question to gauge political identification. The question is, "Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent, or what?"

It's worth noting that this question doesn't ask how a respondent is registered; it just asks how they think of themself, which is likely fluid and may not be as meaningful for generalization to the broader voting population. But it sufficiently suits our investigation here.

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

In [3]:
```
# summary of pid1d and pid1r
paste('pid1d: ')
table(A$pid1d)
paste('pid1r: ')
table(A$pid1r)
```

```
'pid1d: '

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

'pid1r: '

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

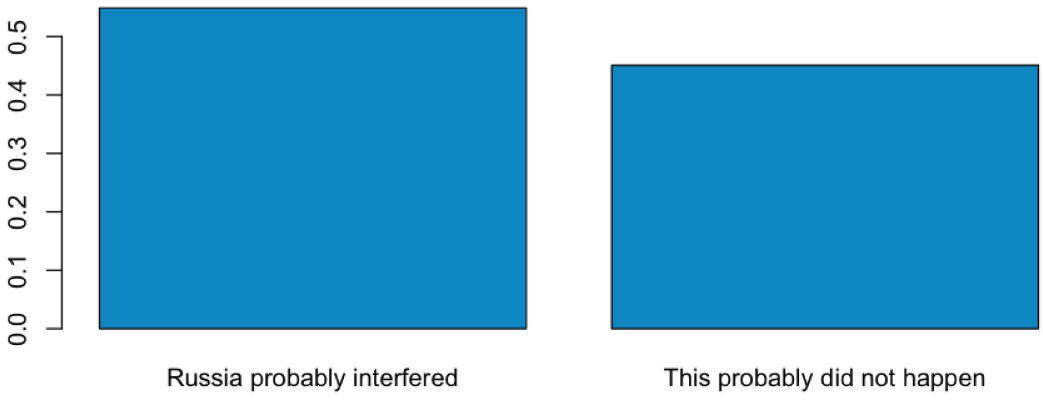
In [4]:
```
# constrain dataset to just voters who identify as independent
# B is data.frame with just voters who marked that they were independent (opti
on 3) according to pid1d or pid1r
B <- A[A$pid1d == 3 | A$pid1r == 3, ]

# check that count of independent voters in dataframe B is equal to sum of ind
ependent voters from pid1d and pid1r in A
paste('Independent voter filter check: ',
      ifelse(length(B$caseid) == sum(ifelse(A$pid1d == 3, 1, 0)) + sum(ifelse(
A$pid1r == 3, 1, 0)), 'Yes', 'No'))

## create variable russia_16_bp
## russia_16_bp is a factor (categorical) variable
russia_16_bp <- factor(B$russia16, labels = c("Russia probably interfered", "T
his probably did not happen"))

# summary of russia16
paste('Response count: ', length(B$russia16))
paste('Mean of responses: ', mean(B$russia16))
table(B$russia16)
options(repr.plot.width=8, repr.plot.height=4)
barplot(prop.table(table(russia_16_bp)), col = 'deepskyblue3')
```

'Independent voter filter check: Yes'

'Response count: 767'

'Mean of responses: 1.45110821382008'

```
  1   2
421 346
```



When the dataset is filtered to include just respondents who identified as independent, there are no cases of non-response, which means we don't have to filter out non-responses above.

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

To answer our question, the most appropriate test is a binomial test. We can think of our survey responses as independent trials in which each trial can result in one of two possible binary outcomes: Russia probably interfered, which we can think of as success, and Russia probably did not interfere, which can think of as failure. Our null hypothesis posits that the probability of success is constant from trial to trial. The number of trials is the number of responses (767) we have in our sample.

To state it more formally, our null hypothesis will be that a majority of independent voters think that Russia probably did NOT interfere in the 2016 election. Our test will show us how extreme our data is assuming that hypothesis is true. Our alternative hypothesis will be that a majority of independent voters think that Russia probably did interfere in the 2016 election. Our threshold for "a majority of voters" will be over 50%.

Since we're concerned about whether or not more than 50% of voters think Russia interfered, we'll use a one-sided test. We're not concerned about testing whether or not exactly 50% of voters believe Russia interfered, just that more than 50% believe Russia interfered. So, we'll use a one-sided test as opposed to a two-sided test.

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 [5]:
```
# one-sided binomial test
binom.test(length(B$russia16[B$russia16 == 1]), length(B$russia16), p = 0.5,
           alternative = "greater",
           conf.level = 0.95)
```

```
        Exact binomial test

data:  length(B$russia16[B$russia16 == 1]) and length(B$russia16)
number of successes = 421, number of trials = 767, p-value = 0.00375
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.5185836 1.0000000
sample estimates:
probability of success
             0.5488918
```

Our test gives us a p-value of 0.00375, so using a significance level of $p < 0.05$, we'll reject the null hypothesis that the majority of independent voters think that Russia probably did NOT interfere in the 2016 election. This result leads us to believe that it is likely not due to chance that 55% of independent respondents believe that Russia interfered in the election despite a true mean of 50% in the population of YouGov participants.

This question doesn't involve comparing means between groups but rather focuses on whether or not it's defensible to make a claim about a population given a single sample of mutually exclusive, binary outcomes. As such, testing for practical significance isn't applicable here. However, looking at the concept of practical significance as a measure of whether the magnitude of a phenomenon is meaningful, we might think of practical significance in this scenario as a measure of whether or not the 55% positive response here is a meaningful departure from a hypothetical 50% positive response per our question. With that interpretation, we could argue, though arbitrary, that the 5% difference we observed in our positive response rate does constitute a meaningful difference from the 50% we expected under the null hypothesis.

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

Introduce your topic briefly. (5 points)

We will be using the variables **geafraid** and **geangry** to do our analysis. These variables indicate the general level of fear and anger that the population feels **about the way things are going in the country these days**. Moreover, we will use two more variables that show the voter turnout in 2016 and 2018, **turnout16** and **turnout18** respectively.

The turnout variables are categorical variables.

turnout18

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

turnout16

1. Definitely voted
2. Definitely did not vote
3. Not completely sure

> For the purpose of this analysis, we will **exclude all the uncertain and invalid responses** and we will transform the **turnout18** variable to keep only if the voter definitely voted or not.

The variables **geafraid** and **geangry** on the other hand are likert (ordinal) variables with values:

geafraid and geangry

1. Not at all
2. A little
3. Somewhat
4. Very
5. Extremely

One limiting factor with these variables is the difficulty of comparison between the two feelings. Indeed, it is very difficult to compare the magnitude and impact of the two feelings on the population decisions based on these values. First, because of the nature of the responses, a likert value generally is not equally scaled between its values. In addition to that, one could argue that for some people that have responded with the same values to the fear and anger questions, fear might have a bigger impact to their decision making process. In other words, we can study the statistical and practical significance, but in the real world we would also have to study the psychological significance in a decision making process.

Furthermore, to answer this question we are forced to ignore the rest of the voters feelings responses which in total are 10. This means that we might be missing the true driver in the population decision.

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

First, we will prepare our data. We will remove all uncertain or NA responses.

In [285]:
```
A$definate_turnout16 <- A$turnout16
A$definate_turnout18 <- A$turnout18

# remove the uncertain responses
A[A$definate_turnout16 == 3,]$definate_turnout16 = NA

#move voted to value 1
A[A$definate_turnout18 %in% c(1, 2, 3),]$definate_turnout18 = 1

# move the not voted to value 2
A[A$definate_turnout18 == 4,]$definate_turnout18 = 2

# remove the uncertain
A[A$definate_turnout18 == 5,]$definate_turnout18 = NA

# make it zero based. Zero voted, one did not
A$definate_turnout18 <- A$definate_turnout18 - 1
A$definate_turnout16 <- A$definate_turnout16 - 1
```

In [286]:
```
'%ni%' <- Negate('%in%')
A[A$geafraid %ni% c(1, 2, 3, 4, 5), ]$geafraid = NA
A[A$geangry %ni% c(1, 2, 3, 4, 5), ]$geangry = NA

#remove NAs records
CompleteA <- A[complete.cases(A$geangry) & complete.cases(A$geafraid) &
             complete.cases(A$definate_turnout16) & complete.cases(A$definate_
turnout18),]
```

Next, we'll explore the turnout difference

In [263]:
```
paste('how many did definitely vote in 2018 that did not vote in 2016?')
B <- CompleteA[CompleteA$definate_turnout18 == 0 & CompleteA$definate_turnout1
6 == 1, ]
paste('Answer: ',nrow(B))
```

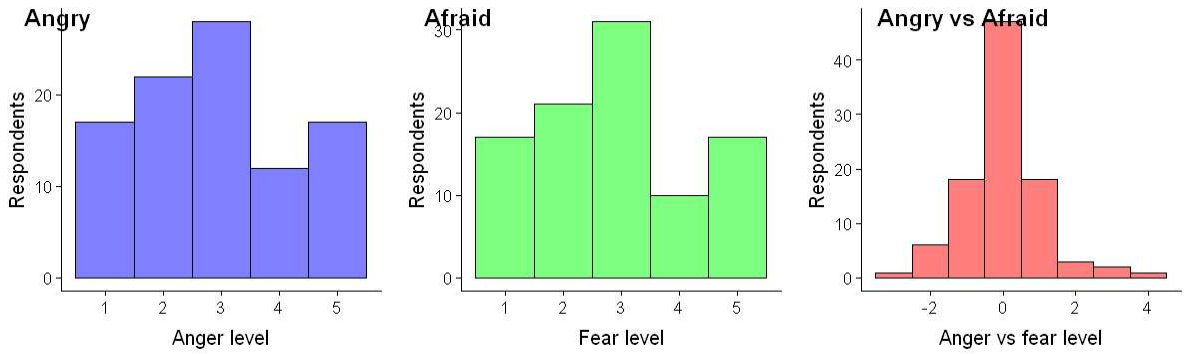
'how many did definitely vote in 2018 that did not vote in 2016?'

'Answer: 96'

Next, we will explore the anger and fear feelings data for only the portion of the sample that did definitely vote in 2018 and did not vote in 2016

In [334]:
```
theme_set(theme_cowplot(font_size=12))
options(repr.plot.width=10, repr.plot.height=3)

B$geangry_minus_geafraid <- B$geangry-B$geafraid
plot.angry<-ggplot(B, aes(x=geangry), xlab = 'test') + geom_histogram(bins=5,
 fill=rgb(0,0,1,1/2), color="black")
plot.afraid<-ggplot(B, aes(x=geafraid)) + geom_histogram(bins=5, fill=rgb(0,1,
0,1/2), color="black")

plot.angry_minus_afraid <- ggplot(B, aes(x=geangry_minus_geafraid)) + geom_his
togram(bins=8, fill=rgb(1,0,0,1/2), color="black")
```

In [335]:
```
plot_grid(plot.angry + labs(x = "Anger level", y = "Respondents"),
          plot.afraid + labs(x = "Fear level", y = "Respondents"),
          plot.angry_minus_afraid + labs(x = "Anger vs fear level", y = "Respo
ndents"),
          nrow = 1,
          labels = c('Angry', 'Afraid', 'Angry vs Afraid'))
```



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.

We can see in that data that while the distributions of anger and fear are non-normal, the distribution of their differences seems to be somewhat normal, with a slight right skew.

Because we are dealing with ordinal values and the distributions of anger and fear are not normal, we will use the **Wilcoxon signed rank test**. We choose to perform a Wilcoxon over a Mann–Whitney, because these values come from the same subject (paired/matched samples).

The null hypothesis (H_0) is that there is no difference between the two feelings of fear and anger that acted as a motivation for the increased turnout in 2018.

Although we are asked to answer which implies a directional test, we will use a two tailed test because we can see that the difference of the two scores is very symmetrically distributed around the zero. This makes us assume that there is no clear direction that can justify a significant result, and because of that, the risk of getting a Type 1 error is increased. For this reason, we choose to test both tails. If we get a two tailed statistically significant result, we will continue our analysis to find the direction. For these reasons, our alternate hypothesis (H_a) is that there is a difference between the two feelings of fear and anger that acted as a motivation for the increased turnout in 2018.

$$\begin{aligned} \text{Null hypothesis:} \quad & H_0 : \mu_1 - \mu_2 = \Delta_0 \\ \text{Alternate hypothesis:} \quad & H_a : \mu_1 - \mu_2 \neq \Delta_0 \end{aligned}$$

We will use the typical significance level 0.05.

Conduct your test. (5 points)

In [204]:
```
#Wilcoxon two sided signed rank (paired)
wilcox.test(B$geafraid, B$geangry, alternative="two.sided", paired=TRUE, exact
=FALSE)
```

```
        Wilcoxon signed rank test with continuity correction

data:  B$geafraid and B$geangry
V = 626, p-value = 0.8917
alternative hypothesis: true location shift is not equal to 0
```

We get a p-value = 0.8917 > 0.05. We failed to reject the null hypothesis in this case and of course we won't examine the practical significance.

> Note: In R if exact is not specified, an exact p-value is computed if the samples contain less than 50 finite values and there are no ties. Otherwise, a normal approximation is used. In our case, we have multiple ties in our data, and the number of samples (96) is close to the minimum number required. Nevertheless, it is save to do this because in our case we don't have to make a marginal decision (p-value >> 0.05)

Conclusion:

Based on the data, we find no statistical significance in the difference between fear and anger in terms of increasing the turnout.

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.

Do people that voted in 2018 trust the media more than the people that didn't vote in 2018?

The fifth question we have chosen to investigate is "Do people that voted in 2018 trust the media more than the people that didn't vote in 2018?". American politics is becoming increasingly polarized, and the election of Donald Trump as President seems only to have acted as a further catalyst for this polarization of attitudes. The President often portrays the media as being dishonest and so we feel it's important to understand the effect of trust on voter turnout. This also has a trickle down effect in terms of how certain media outlets chose to convey the news, i.e. Fox often presents news with a Republican perspective, while NBC adopt a Democratic perspective. The same story is often reported differently between media networks (particularly when remotely related to politics/policy), which when coupled with the President's discouraging remarks, creates real potential for a lack of trust in the media to prevail amongst the U.S. public.

We have specifically chosen to investigate trust amongst voters in the 2018 elections because this ANES survey was also conducted in 2018. Responses to the survey reflect respondents feelings at the time of the completing the survey in 2018, and so are better indicators of their feelings during the 2018 elections.

The variable trustmedia will be used to measure the amount of trust voters have in the media. Specifically, the question asks "In general, how much trust and confidence do you have in the news media when it comes to reporting the news fully, accurately, and fairly?". Responses are on a 5-point Likert Scale, indicating that this is an ordinal variable:

Trust media respones: (1) None, (2) A little, (3) A moderate amount, (4) A lot, (5) A great deal

The variable turnout18 will be used to identify whether respondents turned-out (voted) in the 2018 elections. Turnout18 is a categorical variable but only those respondents who definitely voted will be considered in this analysis (whether voting in-person or by mail). Those who are unsure whther they voted have been excluded on the grounds that if you can't remember whether you voted, then you are likely providing unreliable information, as voting is something one generally does not forget.

A potential gap between the concept of media and the variable trustmedia, is the definition of media. Media today is a broad spectrum of sources, ranging from the traditional sources of news (newspapers, TV, radio) to the likes of social media. The variable trust media is focusing on the traditional news sources. Increasingly social media sites like Facebook have been coming under increased scrutiny because they control the types of adverts users are exposed to, including political messages. Furthermore, in 2016 election there was the issue of 'Fakenews' stories. As such, the distrust generated by social media is not captured by trustmedia which should be noted as a potential gap.

Perform EDA and select your hypothesis test (5 points)

We will split the sample in two groups, the group that definatelly did vote in 2018 and the group that definatelly did not vote in 2018.

In [289]:
```
#remove NAs records
CompleteA <- A[complete.cases(A$trustmedia) & complete.cases(A$definate_turnou
t18),]
paste('how many definitely did vote in 2018?')
paste('Answer: ',nrow(CompleteA[CompleteA$definate_turnout18 == 0, ]))

paste('how many definitely did not vote in 2018?')
paste('Answer: ',nrow(CompleteA[CompleteA$definate_turnout18 == 1, ]))
```

'how many definitely did vote in 2018?'

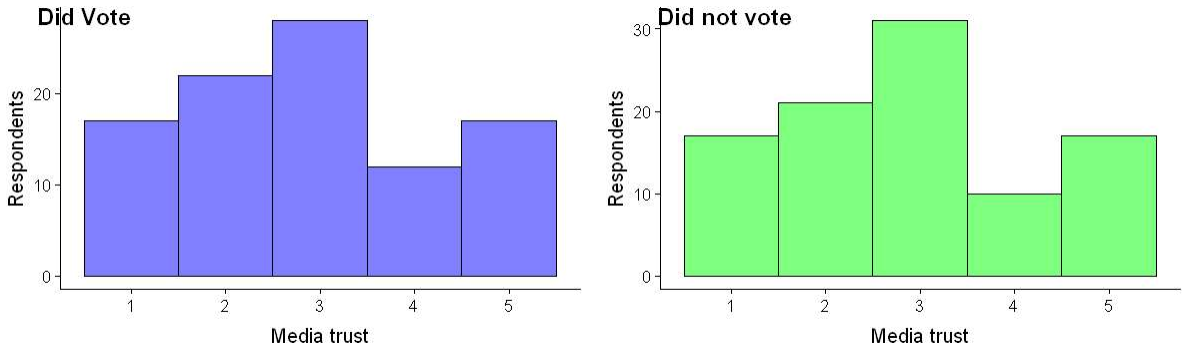
'Answer: 1842'

'how many definitely did not vote in 2018?'

'Answer: 544'

In [336]:
```
turn_out_trustmedia<-CompleteA[CompleteA$definate_turnout18 == 0, ]$trustmedia
no_turn_out_trustmedia <-CompleteA[CompleteA$definate_turnout18 == 1, ]$trustm
edia
df1 <- data.frame(turn_out_trustmedia)
df2 <- data.frame(no_turn_out_trustmedia)

plot.turn_out_trustmedia<-ggplot(df1, aes(x=turn_out_trustmedia), xlab = 'tes
t') + geom_histogram(bins=5, fill=rgb(0,0,1,1/2), color="black")
plot.no_turn_out_trustmedia<-ggplot(df2, aes(x=no_turn_out_trustmedia)) + geom
_histogram(bins=5, fill=rgb(0,1,0,1/2), color="black")
```

In [337]:
```
plot_grid(plot.angry + labs(x = "Media trust", y = "Respondents"),
          plot.afraid + labs(x = "Media trust", y = "Respondents"),
          labels = c('Did Vote', 'Did not vote'))
```



Because we are working with an ordinal variable, and the two variables do do not belong to the same subjects, we will be using the a **Mann-Whitney one-tail unpaired test** and the Cliff's delta for the effect size.

The null hypothesis (H_0) is that there is no difference between the people that turnout in 2018 and the people that did not turnout in 2018 in terms of media trust.

Our alternate hypothesis (H_a) is that the people that did turnout in 2018, trust the media more.

$$\begin{aligned} \text{Null hypothesis:} \quad & H_0 : \mu_1 - \mu_2 = \Delta_0 \\ \text{Alternate hypothesis:} \quad & H_a : \mu_1 - \mu_2 > \Delta_0 \end{aligned}$$

To examine the effect size we will use **Cliff's delta** which is a measure of how often the values in one distribution are larger than the values in a second distribution

> Cliff, Norman (1993). "Dominance statistics: Ordinal analyses to answer ordinal questions". Psychological Bulletin. 114 (3): 494–509. doi:10.1037/0033-2909.114.3.494.

Conduct your test. (2 points)

In [308]:
```
#Mann-Whitney one-tail unpaired test
wilcox.test(turn_out_trustmedia,
            no_turn_out_trustmedia,
            alternative="greater",
            paired=FALSE,
            exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  turn_out_trustmedia and no_turn_out_trustmedia
W = 541150, p-value = 0.001734
alternative hypothesis: true location shift is greater than 0
```

In [309]:
```
# Practical effect size, Cliff's delta
cliff.delta(turn_out_trustmedia,no_turn_out_trustmedia)
```

```
Cliff's Delta

delta estimate: 0.08009397 (negligible)
95 percent confidence interval:
       inf        sup
0.03128259 0.12852423
```

Conclusion (3 points)

We reject the null hypothesis by finding a high statistical significance but the practical significance is negligible. The interpretation of this is that we see a clear indication in our data that people that don't trust the media generally vote less than the people that trust the media, something that seems intuitively right. Nevertheless, the "size" of this correlation is not big enough to be able to classify this finding as practically significant. Perhaps if we examined more parameters related to trust and voting intentions, we would be able to find something more practically significant.