Explanations of Algorithms

Michael Amodeo, Krista Mar, Mona Iwamoto

August 22, 2017

Import data

```
#Load data from csv
d <- read.csv("Explainability_Pilot_Study.csv")

# Remove row 1 that contains question text and any incomplete responses
d1 <- d[-c(1), -seq(2,9)]
d1 <- d1[d1$random != "", ]
```

Note: there are more than 100 columns, which is a limit for display in R studio. However, they are there and I will use them by calling directly.

Randomization Check

Did all of the treatments receive similar numbers of respondents?

```
# Naming Conventions
# Q3 - Twitter Control
# Q4 - Twitter Treatment 1
# Q5 - Twitter Treatment 2
# Q6 - Twitter Treatment 3
# Q8 - Recidivism Control
# Q9 - Recidivism Treatment 1
# Q10 - Recidivism Treatment 2
# Q11 - Recidivism Treatment 3

# Calculate number who saw the first question of each of the paths
# Q#.1 - Description of context. Using as an indicator of treatment assignment
nrow(d1[d1$Q3.1 == 1, ])
```

```
## [1] 21
```

```
nrow(d1[d1$Q4.1 == 1, ])
```

```
## [1] 20
```

```
nrow(d1[d1$Q5.1 == 1, ])
```

```
## [1] 20
```

```
nrow(d1[d1$Q6.1 == 1, ])
```

```
## [1] 20
```

```
nrow(d1[d1$Q8.1 == 1, ])
```

```
## [1] 21
```

```
nrow(d1[d1$Q9.1 == 1, ])
```

```
## [1] 19
```

```
nrow(d1[d1$Q10.1 == 1, ])
```

```
## [1] 21
```

```
nrow(d1[d1$Q11.1 == 1, ])
```

```
## [1] 20
```

Yes, pretty close. All received between 19 and 21 respondents.

Were the two contexts assigned equally?

```
nrow(d1[d1$First.Context == "Twitter", ])
```

```
## [1] 44
```

```
nrow(d1[d1$First.Context == "Recidivism", ])
```

```
## [1] 37
```

Several more were assigned Twitter first. This could be because of non-compliance, where 6 respondents did not complete the study. We can check that.

```
nrow(d[d$First.Context == "Twitter", ])
```

```
## [1] 44
```

```
nrow(d[d$First.Context == "Recidivism", ])
```

```
## [1] 43
```

In fact, all 6 non-compliers were assigned to Recidivism first. Perhaps there is a reason for this we should address before rerunning the experiment. The first step in the survey after random assignment was the introduction of the task How many saw the initial description of the task?

```
nrow(d[d$First.Context == "Recidivism" & d$Q7.1 == 1, ])
```

```
## [1] 41
```

41. So that means 2 left without seeing what they were assigned to. Now, how many made it to treatment?

```
nrow(d[d$First.Context == "Recidivism" & d$Q8.1 == 1, ])
```

```
## [1] 9
```

```
nrow(d[d$First.Context == "Recidivism" & d$Q9.1 == 1, ])
```

```
## [1] 9
```

```
nrow(d[d$First.Context == "Recidivism" & d$Q10.1 == 1, ])
```

```
## [1] 10
```

```
nrow(d[d$First.Context == "Recidivism" & d$Q11.1 == 1, ])
```

```
## [1] 10
```

This shows 38 respondents received Recidivism first, saw the introduction, and saw the treatment. Because 37 completed the survey, only 1 did not comply after seeing the treatment. But 3 left the survey after seeing the description of the task but not the treatment. Perhaps we need to reword the introduction. Interestingly, this covers all 6 non-compliers, which means nobody went through the Twitter example and left upon seeing the recidivism example. Perhaps we do Twitter first in all cases if there is no difference in responses based on order.

Define Metrics

The metrics we evaluated were split into two groups. The first three asked respondents to rate the decision that was made with respect to fairness, accuracy, and their satisfaction with the decision. The second three asked specifically about the explanation itself. Respondents were asked if the explanation was useful, clear, and meaningful.

Consolidate each metric across treatments

```
# For the Twitter fairness question, nobody selected value 1, so it created some weirdness
# with the factors. The -4 accounts for a shift between value and factor
# variables starting with a T indicate Twitter. Variables starting with an R indicate recidivism.

# Q#.2 - Fairness
# Q#.3 - Accuracy
# Q#.4 - Satisfaction
# Q#.6 - Usefulness
# Q#.7 - Clarity
# Q#.8 - Meaningfulness


# Fairness questions
d1$T_fair <- 0
for (i in 1:nrow(d1)){
  if (as.numeric(d1[i, ]$Q3.2) == 1){
    d1[i, ]$T_fair = (as.numeric(d1[i, ]$Q3.2) +  as.numeric(d1[i, ]$Q4.2) +
                           as.numeric(d1[i, ]$Q5.2) + as.numeric(d1[i, ]$Q6.2) - 4)
  } else {
    d1[i, ]$T_fair = (as.numeric(d1[i, ]$Q3.2) +  as.numeric(d1[i, ]$Q4.2) +
                           as.numeric(d1[i, ]$Q5.2) +as.numeric(d1[i, ]$Q6.2) - 3)
  }
}
d1$R_fair <- (as.numeric(d1$Q8.2) + as.numeric(d1$Q9.2) + as.numeric(d1$Q10.2)
              + as.numeric(d1$Q11.2) - 4)

# Accuracy questions
d1$T_accurate <- (as.numeric(d1$Q3.3) + as.numeric(d1$Q4.3) + as.numeric(d1$Q5.3)
                  + as.numeric(d1$Q6.3) - 4)
d1$R_accurate <- (as.numeric(d1$Q8.3) + as.numeric(d1$Q9.3) + as.numeric(d1$Q10.3)
                  + as.numeric(d1$Q11.3) - 4)

# Satisfaction
d1$T_satisfied <- (as.numeric(d1$Q3.4) + as.numeric(d1$Q4.4) + as.numeric(d1$Q5.4)
                   + as.numeric(d1$Q6.4) - 4)
d1$R_satisfied <- (as.numeric(d1$Q8.4) + as.numeric(d1$Q9.4) + as.numeric(d1$Q10.4)
                   + as.numeric(d1$Q11.4) - 4)

#Usefulness of explanation
d1$T_useful <- (as.numeric(d1$Q3.6) + as.numeric(d1$Q4.6) + as.numeric(d1$Q5.6)
                + as.numeric(d1$Q6.6) - 4)
d1$R_useful <- (as.numeric(d1$Q8.6) + as.numeric(d1$Q9.6) + as.numeric(d1$Q10.6)
                + as.numeric(d1$Q11.6) - 4)
```

```
# Clarity of explanation
d1$T_clear <- (as.numeric(d1$Q3.7) + as.numeric(d1$Q4.7) + as.numeric(d1$Q5.7)
               + as.numeric(d1$Q6.7) - 4)
d1$R_clear <- (as.numeric(d1$Q8.7) + as.numeric(d1$Q9.7) + as.numeric(d1$Q10.7)
               + as.numeric(d1$Q11.7) - 4)

# Meaningfulness of explanation
d1$T_meaningful <- (as.numeric(d1$Q3.8) + as.numeric(d1$Q4.8) + as.numeric(d1$Q5.8)
                    + as.numeric(d1$Q6.8) - 4)
d1$R_meaningful <- (as.numeric(d1$Q8.8) + as.numeric(d1$Q9.8) + as.numeric(d1$Q10.8)
                    + as.numeric(d1$Q11.8) - 4)

# Treatment group
T_treat = ((as.numeric(d1$Q3.1) - 1) + (as.numeric(d1$Q4.1) - 1)*2
           + (as.numeric(d1$Q5.1) - 1)*3 + (as.numeric(d1$Q6.1) - 1)*4)
R_treat = ((as.numeric(d1$Q8.1) - 1) + (as.numeric(d1$Q9.1) - 1)*2
           + (as.numeric(d1$Q10.1) - 1)*3 + (as.numeric(d1$Q11.1) - 1)*4)
```

Regression Models

Twitter Moderation

Create linear models for each question for both Twitter and recidivism. Regress on assignment groups, which is 1/0 for Q3.1, Q4.1, Q5.1, and Q6.1, with Q3.1 corresponding to control.

```
# Perform each regression on the variable for the metric regressed on the three different
# treatment assignments. This makes the control the constant value in the regression table.
# Again, the Q#.1 indicates assignment to specific treatment category.
# Q4 - Twitter Treatment 1
# Q5 - Twitter Treatment 2
# Q6 - Twitter Treatment 3

m_T_fair <- lm( T_fair ~ Q4.1 + Q5.1 + Q6.1, data = d1)

m_T_accurate <- lm( T_accurate ~ Q4.1 + Q5.1 + Q6.1, data = d1)

m_T_satisfied <- lm( T_satisfied ~ Q4.1 + Q5.1 + Q6.1, data = d1)

m_T_useful <- lm( T_useful ~ Q4.1 + Q5.1 + Q6.1, data = d1)

m_T_clear <- lm( T_clear ~ Q4.1 + Q5.1 + Q6.1, data = d1)

m_T_meaningful <- lm( T_meaningful ~ Q4.1 + Q5.1 + Q6.1, data = d1)

library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
stargazer(m_T_fair, m_T_accurate, m_T_satisfied, m_T_useful, m_T_clear, m_T_meaningful,
          type = 'text',
          covariate.labels = c("Detailed Description", "Similarity to Humans",
                               "Graphic Representation"),
          dep.var.labels = c("Fairness", "Accuracy", "Satisfaction", "Usefulness",
                             "Clarity", "Meaningfulness"),
          dep.var.caption = "Twitter Moderation")
```

```
##
## ================================================================================================
##                                                        Twitter Moderation
##                               ------------------------------------------------------------------
##                               Fairness  Accuracy Satisfaction Usefulness Clarity  Meaningfulness
##                                  (1)      (2)        (3)         (4)       (5)         (6)
## ------------------------------------------------------------------------------------------------
## Detailed Description          -0.767**   -0.531     -0.402      -0.329    -0.345      -0.117
##                                (0.377)  (0.348)    (0.284)     (0.369)   (0.290)     (0.391)
##
## Similarity to Humans          -1.217***  -0.081     0.048       -0.029    -0.095      0.133
##                                (0.377)  (0.348)    (0.284)     (0.369)   (0.290)     (0.391)
##
## Graphic Representation        -1.117***  -0.481     -0.202      -0.179    -0.445      -0.567
##                                (0.377)  (0.348)    (0.284)     (0.369)   (0.290)     (0.391)
##
## Constant                      4.667***  2.381***   1.952***    2.429***  2.095***    2.667***
##                                (0.264)  (0.243)    (0.198)     (0.258)   (0.202)     (0.273)
##
## ------------------------------------------------------------------------------------------------
## Observations                     81        81         81          81        81          81
## R2                              0.144    0.045      0.039       0.013     0.039       0.044
## Adjusted R2                     0.110    0.008      0.001       -0.026    0.001       0.007
## Residual Std. Error (df = 77)   1.208    1.114      0.909       1.182     0.927       1.252
## F Statistic (df = 3; 77)      4.304***   1.209      1.036       0.333     1.031       1.182
## ================================================================================================
## Note:                                                                *p<0.1; **p<0.05; ***p<0.01
```

In the case of Twitter moderation, no explanations show significance in all questions. Interestingly, all explanations suffer a statistically significant decrease in fairness compared to the control. This is surprising and could be attributable to the small sample size. However, it is also possible that because the tweet we are using is so clearly offensive, any explanation causes questions about fairness. In other words, the tweet itself is the best explanation for why it was deemed offensive.

On the other metrics, both the detailed description and the graphic description show improvements to all other metrics (negative is an improvement because option 1 was the most positive attribute, ie "Extremely Meaningful"). The Similarity to Humans explanation showed some improvements and some decreases, and their magnitudes were less than either of the other explanations. This seems to clearly be the least impactful explanation in our limited study.

For reference: fairness - positive is more fair accuracy - negative is more accurate satisfied - negative is more satisfied useful - negative is more satisfied clear - negative is more clear meaningful - negative is more meaningful

Risk of Recidivism

```
# Perform each regression on the variable for the metric regressed on the three different
# treatment assignments. This makes the control the constant value in the regression table.
```

```r
# Again, the Q#.1 indicates assignment to specific treatment category.
# Q9 - Recidivism Treatment 1
# Q10 - Recidivism Treatment 2
# Q11 - Recidivism Treatment 3

m_R_fair <- lm( R_fair ~ Q9.1 + Q10.1 + Q11.1, data = d1)

m_R_accurate <- lm( R_accurate ~ Q9.1 + Q10.1 + Q11.1, data = d1)

m_R_satisfied <- lm( R_satisfied ~ Q9.1 + Q10.1 + Q11.1, data = d1)

m_R_useful <- lm( R_useful ~ Q9.1 + Q10.1 + Q11.1, data = d1)

m_R_clear <- lm( R_clear ~ Q9.1 + Q10.1 + Q11.1, data = d1)

m_R_meaningful <- lm( R_meaningful ~ Q9.1 + Q10.1 + Q11.1, data = d1)

library(stargazer)
stargazer(m_R_fair, m_R_accurate, m_R_satisfied, m_R_useful, m_R_clear, m_R_meaningful,
          type = 'text',
          covariate.labels = c("Detailed Description", "Similarity to Humans",
                               "Graphic Representation"),
          dep.var.labels = c("Fairness", "Accuracy", "Satisfaction", "Usefulness",
                             "Clarity", "Meaningfulness"),
          dep.var.caption = "Recidivism Risk Assessment")
```

```
## 
## ===============================================================================================
##                                                  Recidivism Risk Assessment                    
##                               -----------------------------------------------------------------
##                               Fairness Accuracy Satisfaction Usefulness Clarity  Meaningfulness
##                                 (1)      (2)        (3)         (4)       (5)         (6)      
## -----------------------------------------------------------------------------------------------
## Detailed Description           0.075    -0.459     -0.637      -0.185    -0.331      -0.175    
##                               (0.376)  (0.351)    (0.395)     (0.393)   (0.378)     (0.419)    
##                                                                                                
## Similarity to Humans           -0.048   0.048      0.048       -0.048    -0.381      -0.190    
##                               (0.367)  (0.342)    (0.385)     (0.383)   (0.368)     (0.408)    
##                                                                                                
## Graphic Representation         0.136    -0.293     -0.352     -0.838**  -0.907**    -0.833**   
##                               (0.371)  (0.346)    (0.390)     (0.388)   (0.373)     (0.413)    
##                                                                                                
## Constant                      2.714*** 3.143***   2.952***    3.238***  2.857***    3.333***   
##                               (0.259)  (0.242)    (0.272)     (0.271)   (0.260)     (0.289)    
##                                                                                                
## -----------------------------------------------------------------------------------------------
## Observations                     81       81         81          81        81          81      
## R2                             0.004    0.036      0.049       0.072     0.073       0.057     
## Adjusted R2                    -0.035   -0.002     0.012       0.035     0.037       0.020     
## Residual Std. Error (df = 77)  1.188    1.108      1.248       1.242     1.192       1.323     
## F Statistic (df = 3; 77)       0.095    0.947      1.321       1.978     2.016       1.547     
## ===============================================================================================
## Note:                                                               *p<0.1; **p<0.05; ***p<0.01
```

Viewing the recidivism responses, we do see some statistical significance start to show up for the Graphic Representation of the explanation. This explanation is statistically significant at a 0.05 level for explanation usefulness, clarity, and meaningfulness. It improves scores for decision metrics of fairness, accuracy, and satisfaction, although not significantly. As in the Twitter example, the Similarity to Humans explanation has the smallest magnitude effect in nearly all cases, with some values showing as improvements over control and others as declines compared to control. The detailed description improves all metrics, although with less magnitude than the graphic representation in most cases, with exceptions being decision accuracy and satisfaction. However, it is much lower for metrics based around the explanation itself with no statistical significance.

Difference in Order

We also discussed looking at the difference in responses depending on the order of contexts.

```r
o_T_fair <- lm( T_fair ~ factor(First.Context), data = d1)

o_T_accurate <- lm( T_accurate ~ factor(First.Context), data = d1)

o_T_satisfied <- lm( T_satisfied ~ factor(First.Context), data = d1)

o_T_useful <- lm( T_useful ~ factor(First.Context), data = d1)

o_T_clear <- lm( T_clear ~ factor(First.Context), data = d1)

o_T_meaningful <- lm( T_meaningful ~ factor(First.Context), data = d1)

library(stargazer)
stargazer(o_T_fair, o_T_accurate, o_T_satisfied, o_T_useful, o_T_clear, o_T_meaningful,
          type = 'text',
          dep.var.labels = c("Fairness", "Accuracy", "Satisfaction", "Usefulness",
                             "Clarity", "Meaningfulness"),
          dep.var.caption = "Twitter Moderation")
```

```
##
## ===============================================================================================
##                                                      Twitter Moderation
##                               -----------------------------------------------------------------
##                               Fairness Accuracy Satisfaction Usefulness Clarity  Meaningfulness
##                                 (1)      (2)        (3)         (4)       (5)         (6)
## -----------------------------------------------------------------------------------------------
## factor(First.Context)Twitter   0.117    0.404      0.057       0.048     -0.028      0.181
##                               (0.287)  (0.247)    (0.204)     (0.262)   (0.208)     (0.281)
##
## Constant                      3.838*** 1.892***   1.784***    2.270***  1.892***    2.432***
##                               (0.212)  (0.182)    (0.150)     (0.193)   (0.153)     (0.207)
##
## -----------------------------------------------------------------------------------------------
## Observations                     81       81         81          81        81          81
## R2                             0.002    0.033      0.001       0.0004    0.0002      0.005
## Adjusted R2                   -0.011    0.020      -0.012      -0.012    -0.012      -0.007
## Residual Std. Error (df = 79)  1.287    1.107      0.915       1.174     0.933       1.260
## F Statistic (df = 1; 79)       0.165    2.673      0.078       0.033     0.018       0.415
## ===============================================================================================
## Note:                                                               *p<0.1; **p<0.05; ***p<0.01
```

In the case of Twitter moderation, there does not seem to be a difference based on order of context.

```
o_R_fair <- lm( R_fair ~ factor(First.Context), data = d1)

o_R_accurate <- lm( R_accurate ~ factor(First.Context), data = d1)

o_R_satisfied <- lm( R_satisfied ~ factor(First.Context), data = d1)

o_R_useful <- lm( R_useful ~ factor(First.Context), data = d1)

o_R_clear <- lm( R_clear ~ factor(First.Context), data = d1)

o_R_meaningful <- lm( R_meaningful ~ factor(First.Context), data = d1)

library(stargazer)
stargazer(o_R_fair, o_R_accurate, o_R_satisfied, o_R_useful, o_R_clear, o_R_meaningful,
          type = 'text',
          dep.var.labels = c("Fairness", "Accuracy", "Satisfaction", "Usefulness",
                             "Clarity", "Meaningfulness"),
          dep.var.caption = "Recidivism Risk Assessment")
```

```
##
## ===============================================================================================
##                                                  Recidivism Risk Assessment
##                               -----------------------------------------------------------------
##                               Fairness Accuracy Satisfaction Usefulness Clarity  Meaningfulness
##                                 (1)      (2)        (3)         (4)       (5)         (6)
## -----------------------------------------------------------------------------------------------
## factor(First.Context)Twitter   0.093    -0.045     0.097       -0.045    0.045       0.118
##                               (0.262)  (0.248)    (0.282)     (0.284)   (0.273)     (0.300)
##
## Constant                      2.703*** 3.000***   2.676***    3.000***  2.432***    2.973***
##                               (0.193)  (0.183)    (0.207)     (0.209)   (0.201)     (0.221)
##
## -----------------------------------------------------------------------------------------------
## Observations                     81       81         81          81        81          81
## R2                             0.002    0.0004     0.002       0.0003    0.0003      0.002
## Adjusted R2                    -0.011   -0.012     -0.011      -0.012    -0.012      -0.011
## Residual Std. Error (df = 79)  1.174    1.113      1.262       1.272     1.222       1.344
## F Statistic (df = 1; 79)       0.125    0.034      0.119       0.026     0.027       0.155
## ===============================================================================================
## Note:                                                               *p<0.1; **p<0.05; ***p<0.01
```

There also seems to be no effect of order on the recidivism metrics either.

Conclusions

Based on this limited pilot study, it seems as though detailed explanations of the reasons for a decision are helpful for people, and that a graphic representation of likelihood and the strength of different decision factors further enhance the explanation. Therefore, we would like to proceed with our experiment using the graphic representations of explanations.

There does seem to be differences across contexts, with the Twitter example not showing significant effects. But this could have to do with the example shown. Perhaps we should choose something more ambiguous.

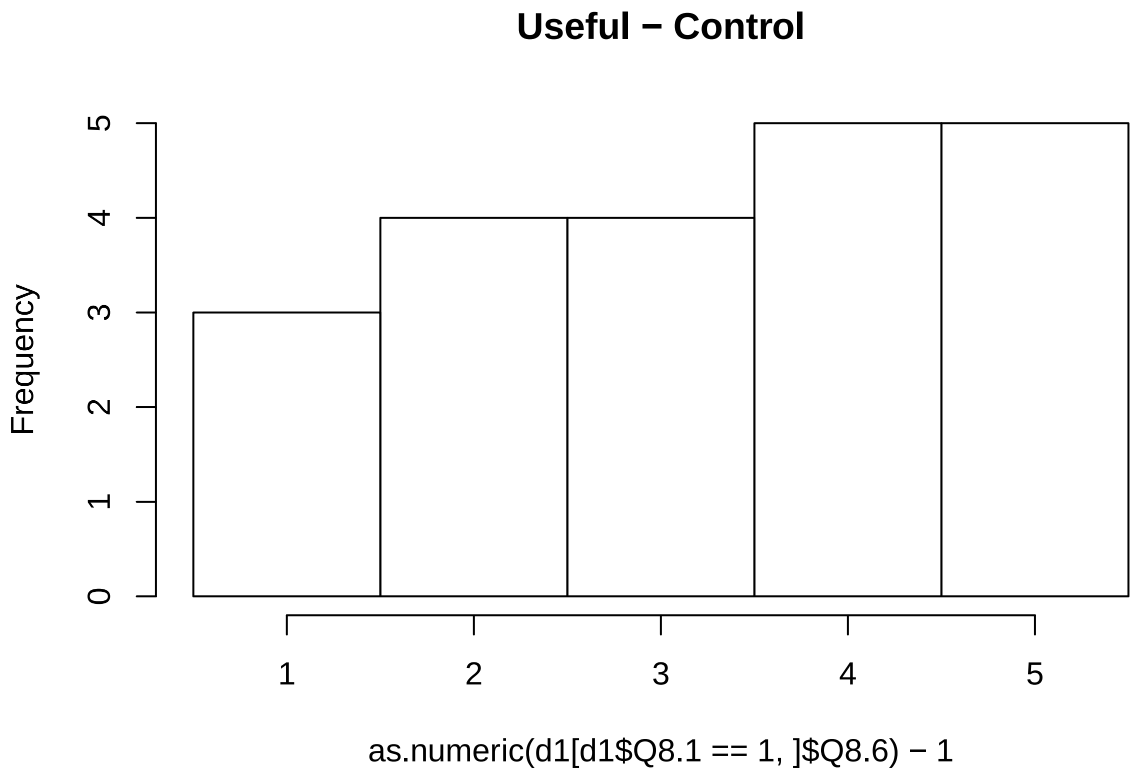
Lastly, the order does not seem to make a difference, so we will continue showing each respondent both contexts in a randomized order. Although perhaps to address the non-compliance, we should go Twitter first always.

Data Checks

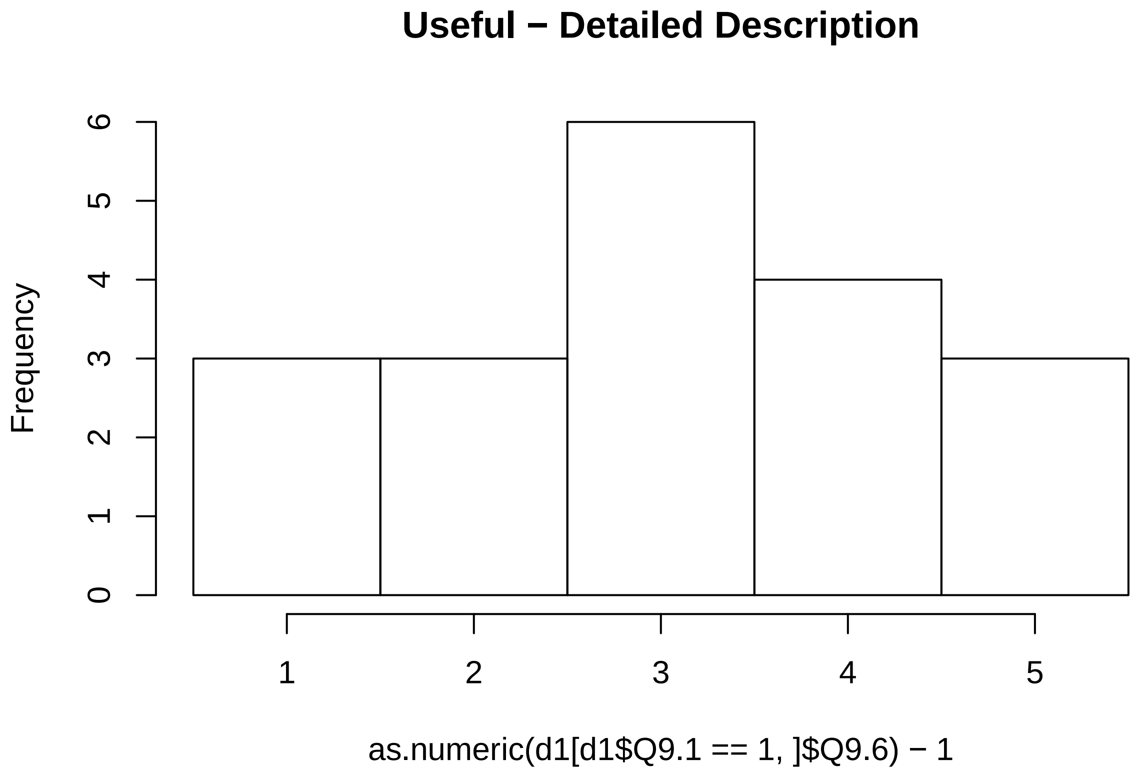
Just to check a couple of these outputs.

Usefulness of Recidivism Explanations

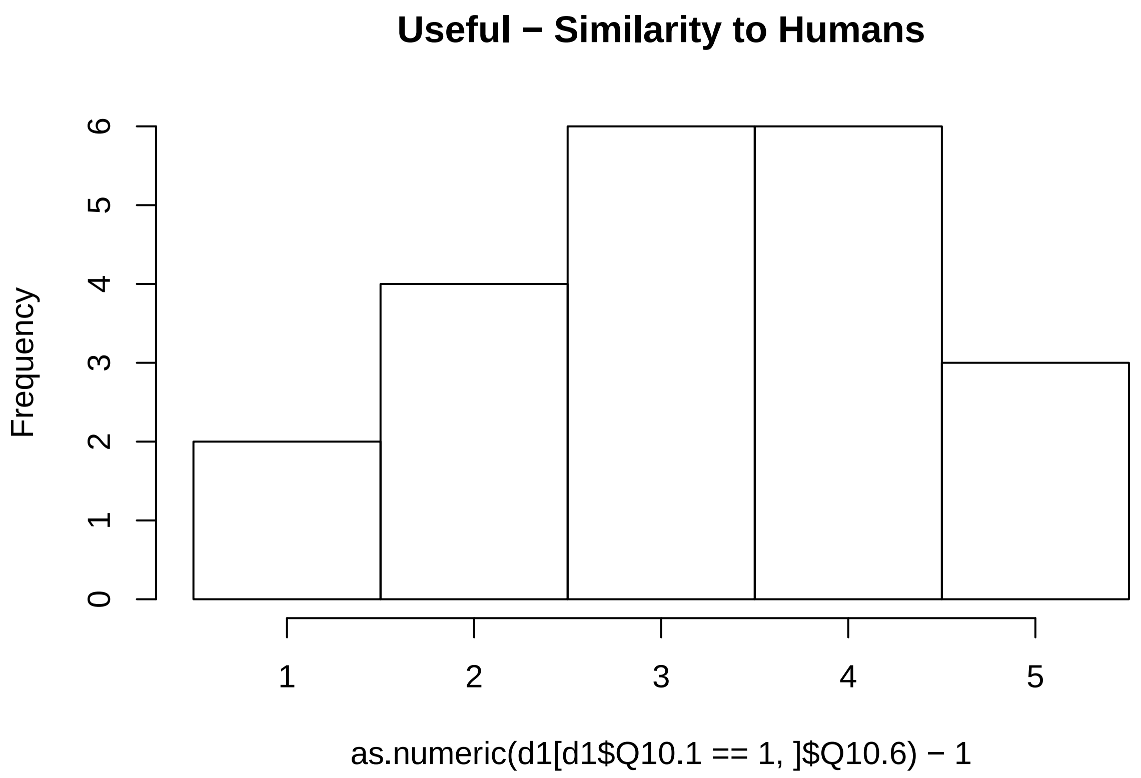
```
hist(as.numeric(d1[d1$Q8.1 == 1, ]$Q8.6)-1, main = "Useful - Control",
     breaks = seq(0.5,5.5, 1))
```



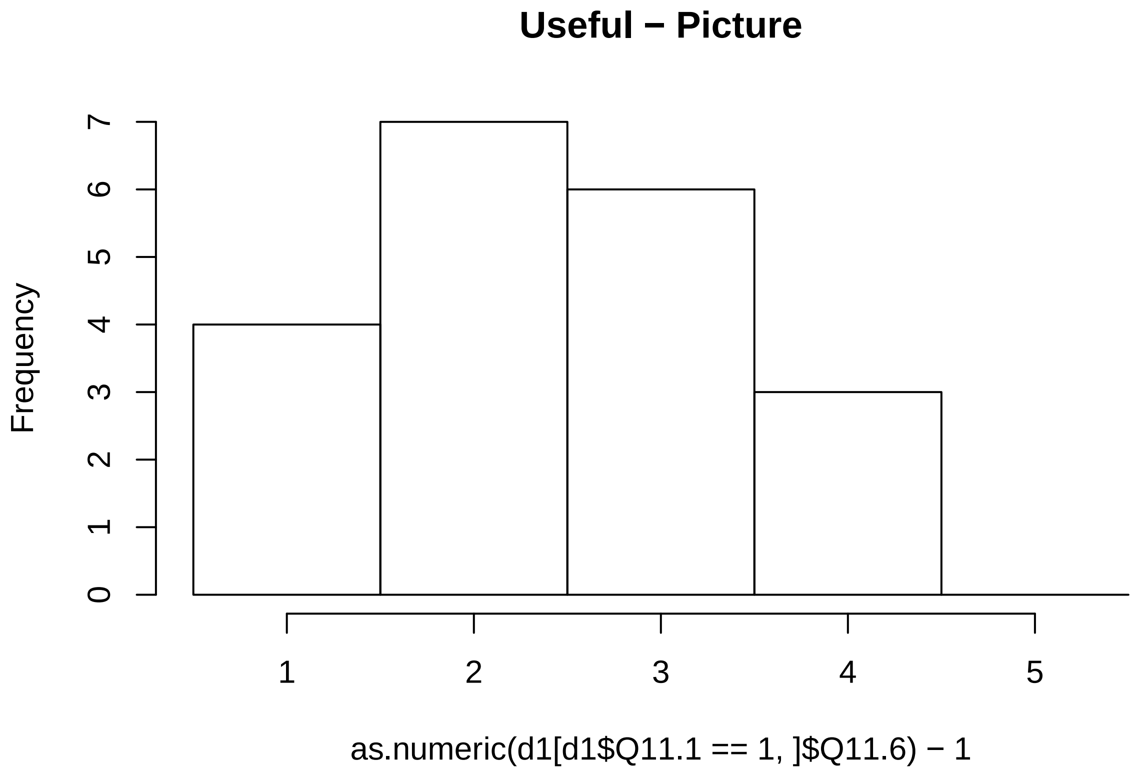
```
hist(as.numeric(d1[d1$Q9.1 == 1, ]$Q9.6)-1, main = "Useful - Detailed Description",
     breaks = seq(0.5,5.5, 1))
```



```
hist(as.numeric(d1[d1$Q10.1 == 1, ]$Q10.6)-1, main = "Useful - Similarity to Humans",
     breaks = seq(0.5,5.5, 1))
```

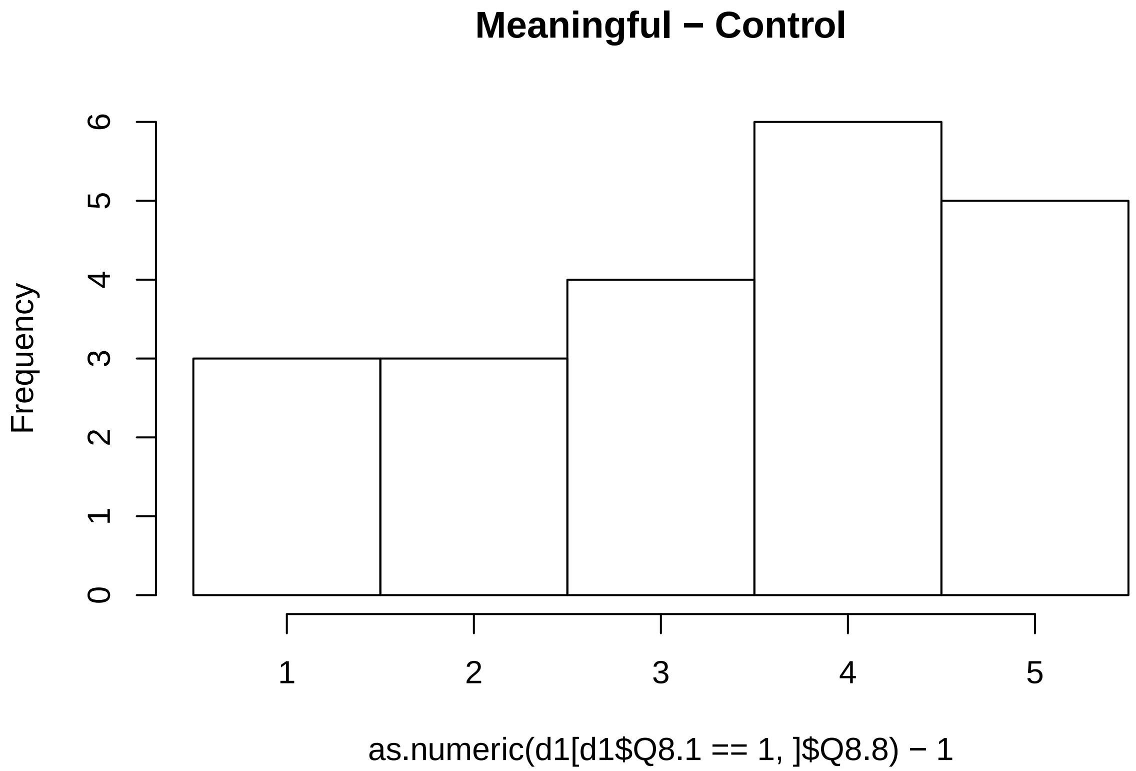


```r
hist(as.numeric(d1[d1$Q11.1 == 1, ]$Q11.6)-1, main = "Useful - Picture",
     breaks = seq(0.5,5.5, 1))
```

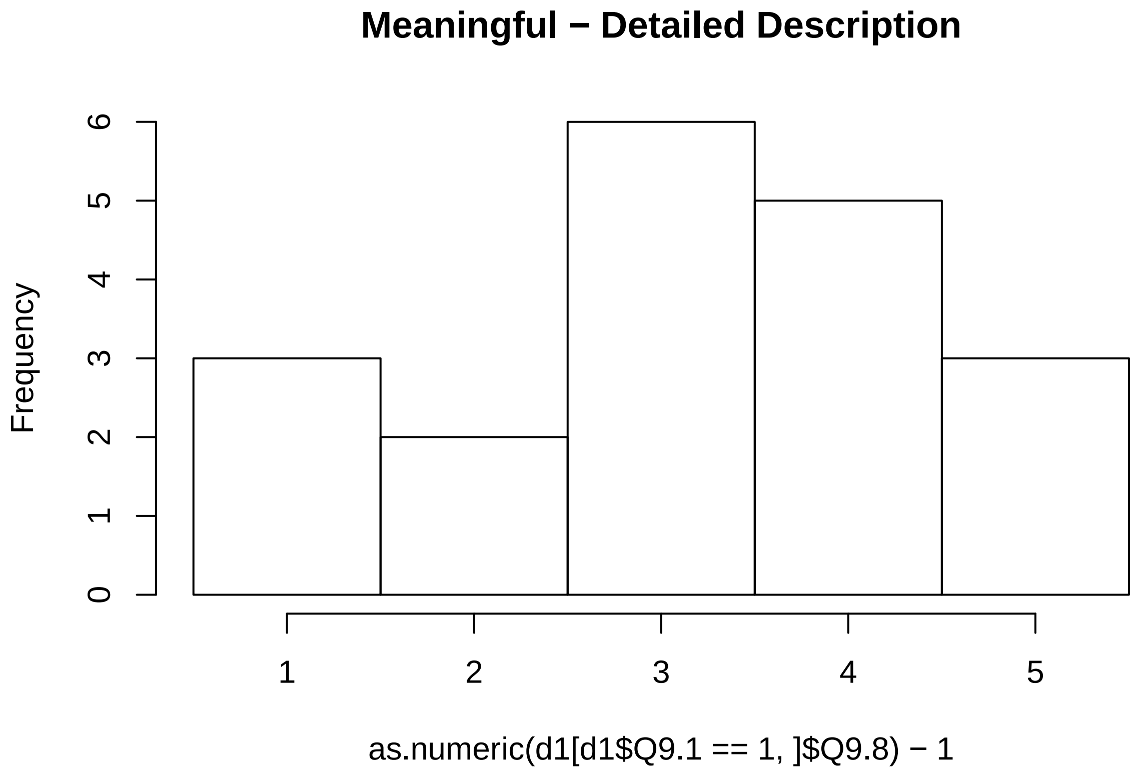


Meaningfulness of Recidivism Explanations

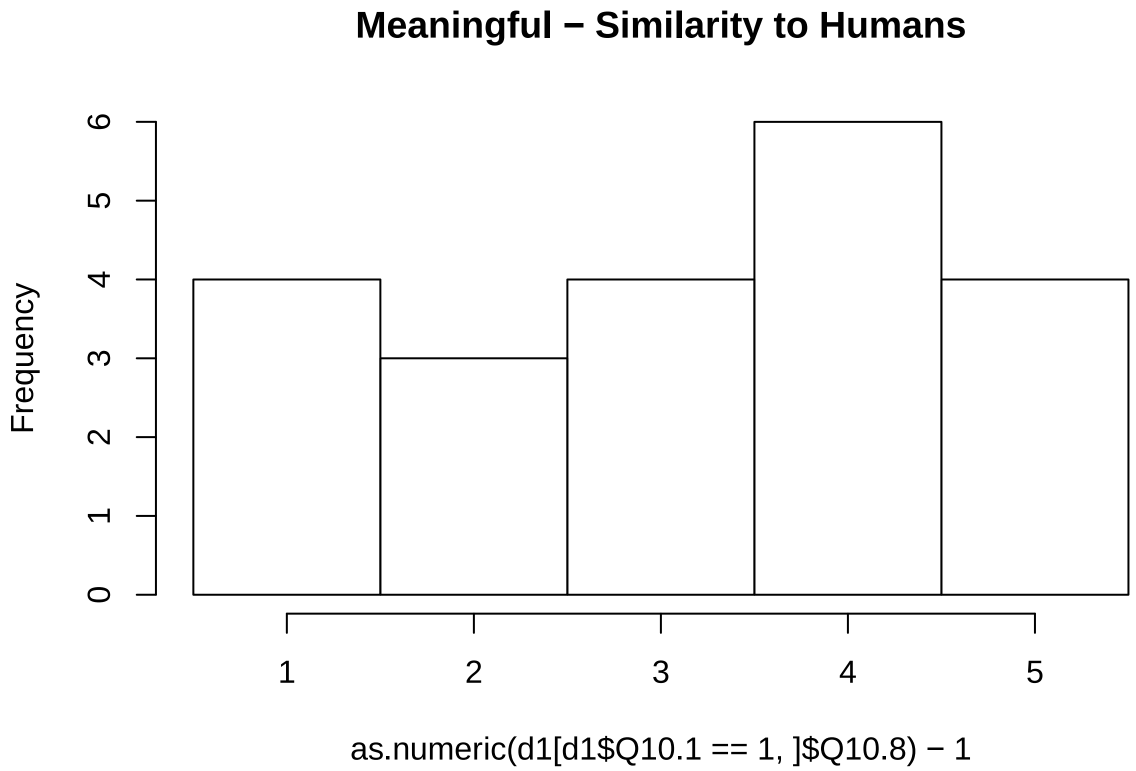
```
hist(as.numeric(d1[d1$Q8.1 == 1, ]$Q8.8)-1, main = "Meaningful - Control", breaks = seq(0.5,5.5, 1))
```



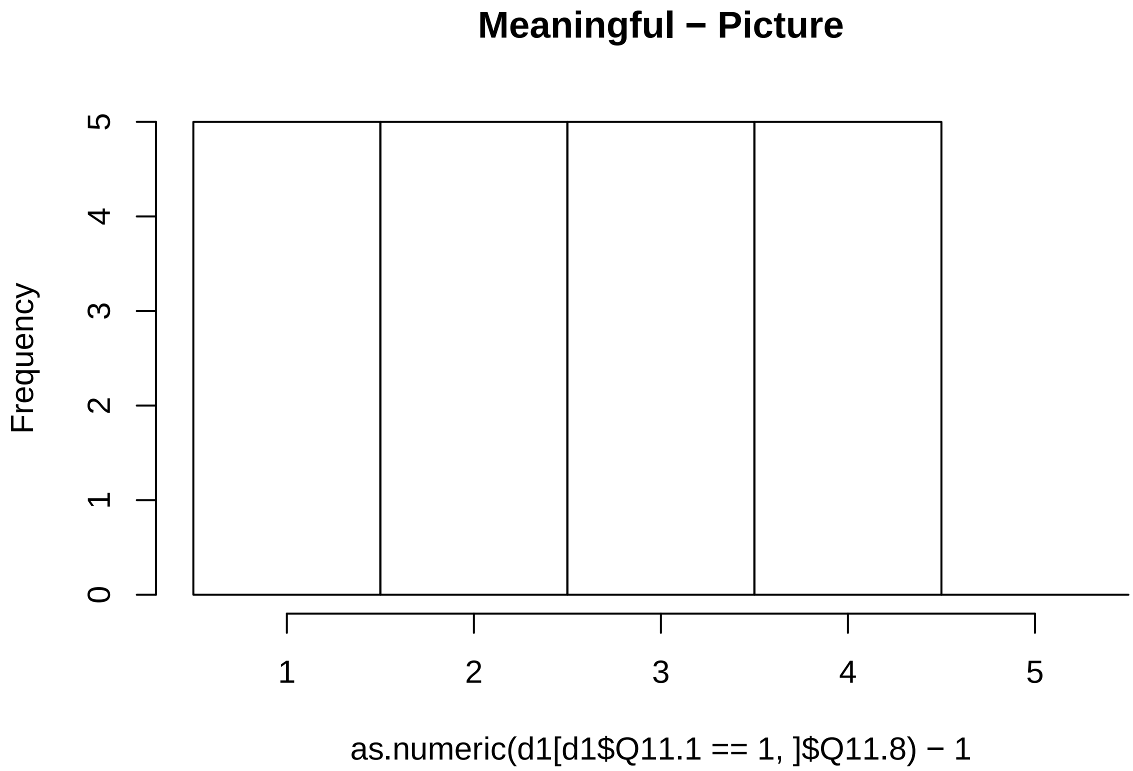
```
hist(as.numeric(d1[d1$Q9.1 == 1, ]$Q9.8)-1, main = "Meaningful - Detailed Description", breaks = seq(0.5
```



```
hist(as.numeric(d1[d1$Q10.1 == 1, ]$Q10.8)-1, main = "Meaningful - Similarity to Humans", breaks = seq(
```



```
hist(as.numeric(d1[d1$Q11.1 == 1, ]$Q11.8)-1, main = "Meaningful - Picture", breaks = seq(0.5,5.5, 1))
```


Meaningful − Picture
Frequency
0
1
2
3
4
5
1
2
3
4
5
as.numeric(d1[d1$Q11.1 == 1,]$Q11.8) − 1