

Explanability Study

Michael Amodeo, Krista Mar, Mona Iwamoto

August 16, 2017

Data Import and Prepration

Import data

Data was exported from Qualtrics using the Legacy Format CSV export. This allowed for additional fields based on whether questions were seen, even if the questions did not require answers.

```
#Import all data
all_content = readLines("Explainability_Study_legacy_export.csv")

#Delete second and third rows of not useful information
skip_second = all_content[-c(2,3)]

#Create table and data.table
d <- read.csv(textConnection(skip_second), header = TRUE, stringsAsFactors = FALSE)
d <- data.table(d)

remove(all_content, skip_second)

# Create new data table without fields we are not using
dt <- d[, -c('ResponseSet', 'IPAddress', 'StartDate', 'EndDate', 'RecipientLastName',
            'RecipientFirstName', 'RecipientEmail', 'ExternalDataReference', 'Status',
            'Q_TotalDuration', 'Enter.Embedded.Data.Field.Name.Here...', 'LocationLatitude',
            'LocationLongitude', 'LocationAccuracy', 'Q3.5', 'Q4.5', 'Q6.5', 'Q7.5', 'Q8.1',
            'Q9.1', 'Q10.1', 'Q10.3')]

# Rename variables
old_names <- colnames(dt)

## Key to var names: tc = Twitter control group
#                   tt = Twitter treatment group
#                   rc = recidivism control group
#                   rt = recidivism treatment group

new_names <- c("ResponseID", "Finished", "First.Context", "random", "intro", "tweet",
              "tControl", "tcFair", "tcAcc", "tcSat", "tcUseful", "tcClear",
              "tcMeaningful", "tcReqInfo1", "tcReqInfo2", "tcReqInfo3", "tcReqInfo4",
              "tcReqInfo4_txt",
              "tTreat", "ttFair", "ttAcc", "ttSat", "ttUseful", "ttClear",
              "ttMeaningful", "ttReqInfo1", "ttReqInfo2", "ttReqInfo3", "ttReqInfo4",
              "ttReqInfo4_txt",
              "recidivism",
              "rControl", "rcFair", "rcAcc", "rcSat", "rcUseful", "rcClear",
              "rcMeaningful", "rcReqInfo1", "rcReqInfo2", "rcReqInfo3", "rcReqInfo4",
              "rcReqInfo4_txt",
              "rTreat", "rtFair", "rtAcc", "rtSat", "rtUseful", "rtClear",
              "rtMeaningful", "rtReqInfo1", "rtReqInfo2", "rtReqInfo3", "rtReqInfo4",
```

```

      'rtReqInfo4_txt',
      'ageGroup', 'white', 'black', 'native', 'asian', 'pac_isle', 'hispanic',
      'other', 'gender', 'socMed', 'educ', 'feedback')

setnames(dt, old_names, new_names)
colnames(dt)

## [1] "ResponseID"      "Finished"         "First.Context"    "random"
## [5] "intro"           "tweet"            "tControl"         "tcFair"
## [9] "tcAcc"           "tcSat"            "tcUseful"         "tcClear"
## [13] "tcMeaningful"    "tcReqInfo1"       "tcReqInfo2"       "tcReqInfo3"
## [17] "tcReqInfo4"      "tcReqInfo4_txt"   "tTreat"           "ttFair"
## [21] "ttAcc"           "ttSat"            "ttUseful"         "ttClear"
## [25] "ttMeaningful"    "ttReqInfo1"       "ttReqInfo2"       "ttReqInfo3"
## [29] "ttReqInfo4"      "ttReqInfo4_txt"   "recidivism"       "rControl"
## [33] "rcFair"          "rcAcc"            "rcSat"            "rcUseful"
## [37] "rcClear"         "rcMeaningful"     "rcReqInfo1"       "rcReqInfo2"
## [41] "rcReqInfo3"      "rcReqInfo4"       "rcReqInfo4_txt"   "rTreat"
## [45] "rtFair"          "rtAcc"            "rtSat"            "rtUseful"
## [49] "rtClear"         "rtMeaningful"     "rtReqInfo1"       "rtReqInfo2"
## [53] "rtReqInfo3"      "rtReqInfo4"       "rtReqInfo4_txt"   "ageGroup"
## [57] "white"           "black"            "native"           "asian"
## [61] "pac_isle"        "hispanic"         "other"            "gender"
## [65] "socMed"          "educ"            "feedback"

remove(old_names)

```

Data Cleanup

The questions were based on a 5 point Likert scale. For each metric, the answers varied from “Extremely” to “Not at All.” Most Qualtrics questions were set so the extreme positive value was the first choice (1). In order to show an increase in acceptance or trust as positive, we will rescale these values and flip them around the median value (3).

```

# Function to flip the scale to show more positive as larger number
flip <- function(originalScale) {
  x <- originalScale - 3      # 3 is median
  return(3 - x)
}

# For an unknown reason, question 7.2 Fairness for Recidivism Treatment
# the values were offset by 24. This was cross-checked with the text-based responses.

dt$rtFair <- dt$rtFair - 24    # Qualtrics weirdness
dt$educ <- dt$educ - 10       # Qualtrics weirdness

# For the Twitter fairness questions, the Qualtrics survey
# responses were reversed in two instances. All others
# are reversed using the flip function below

# Organize scales so larger values correlate to more fair, more accurate, etc.

flip_cols <- c("tcAcc", "tcSat", "tcUseful", "tcClear", "tcMeaningful",

```

```
"ttFair", "ttAcc", "ttSat", "ttUseful", "ttClear", "ttMeaningful",
"rcAcc", "rcSat", "rcUseful", "rcClear", "rcMeaningful",
"rtFair", "rtAcc", "rtSat", "rtUseful", "rtClear", "rtMeaningful")
```

```
dt[, (flip_cols) := lapply(.SD, flip), .SDcols = flip_cols]
```

Consolidate each metric across treatments

Because the Qualtrics format requires each question to be different, we have to consolidate the responses for each metric into a single value per context. Because respondents either saw control or treatment for each context, we simply make a new metric that is the sum of the old metrics.

Create consolidated data table

```
dc <- data.table(ResponseID = dt[, ResponseID])

dc[, complete := !is.na(dt[, random])] # Did they complete the survey?
dc[, tAssign := dt[, tTreat] == 1 | dt[, tControl] == 1] # Was a Twitter treatment assigned?
dc[, tControl := !is.na(dt[, tControl])]
dc[, tTreat := !is.na(dt[, tTreat])]
dc[, rControl := !is.na(dt[, rControl])]
dc[, rTreat := !is.na(dt[, rTreat])]
dc[, rAssign := dt[, rTreat] == 1 | dt[, rControl] == 1] # Was a recidivism treatment assigned?
dc[, tweet := !is.na(dt[, tweet])] # Did they reach the Twitter context?
dc[, recidivism := !is.na(dt[, recidivism])] # Did they reach the recidivism context?

dc[, tFair := rowSums(dt[, c('tcFair', 'ttFair')], na.rm=T)]
dc[, tAcc := rowSums(dt[, c('tcAcc', 'ttAcc')], na.rm=T)]
dc[, tSat := rowSums(dt[, c('tcSat', 'ttSat')], na.rm=T)]
dc[, tUseful := rowSums(dt[, c('tcUseful', 'ttUseful')], na.rm=T)]
dc[, tClear := rowSums(dt[, c('tcClear', 'ttClear')], na.rm=T)]
dc[, tMeaningful := rowSums(dt[, c('tcMeaningful', 'ttMeaningful')], na.rm=T)]

dc[, rFair := rowSums(dt[, c('rcFair', 'rtFair')], na.rm=T)]
dc[, rAcc := rowSums(dt[, c('rcAcc', 'rtAcc')], na.rm=T)]
dc[, rSat := rowSums(dt[, c('rcSat', 'rtSat')], na.rm=T)]
dc[, rUseful := rowSums(dt[, c('rcUseful', 'rtUseful')], na.rm=T)]
dc[, rClear := rowSums(dt[, c('rcClear', 'rtClear')], na.rm=T)]
dc[, rMeaningful := rowSums(dt[, c('rcMeaningful', 'rtMeaningful')], na.rm=T)]

dc[, tReqInfo1 := rowSums(dt[, c('tcReqInfo1', 'ttReqInfo1')], na.rm=T)]
dc[, tReqInfo2 := rowSums(dt[, c('tcReqInfo2', 'ttReqInfo2')], na.rm=T)]
dc[, tReqInfo3 := rowSums(dt[, c('tcReqInfo3', 'ttReqInfo3')], na.rm=T)]

dc[, rReqInfo1 := rowSums(dt[, c('rcReqInfo1', 'rtReqInfo1')], na.rm=T)]
dc[, rReqInfo2 := rowSums(dt[, c('rcReqInfo2', 'rtReqInfo2')], na.rm=T)]
dc[, rReqInfo3 := rowSums(dt[, c('rcReqInfo3', 'rtReqInfo3')], na.rm=T)]

dc[, white := !is.na(dt[, white])]
dc[, black := !is.na(dt[, black])]
dc[, native := !is.na(dt[, native])]
dc[, asian := !is.na(dt[, asian])]
dc[, pac_isle := !is.na(dt[, pac_isle])]
```

```

dc[, hispanic := !is.na(dt[, hispanic])]
dc[, other := !is.na(dt[, other])]

dc[, female := (dt[, gender]==2)]
dc[, gender_nc := (dt[, gender]==3)]

dt1 <- dt[, c('ageGroup', 'socMed', 'educ', 'First.Context')]

dc <- cbind(dc, dt1)

# Converting tTreat and rTreat to binaries instead of logicals
(to.replace <- names(which(sapply(dc, is.logical))))
for (var in to.replace) dc[, var:= as.numeric(get(var)), with=FALSE]

head(dc)

```

Randomization Check

Were the two contexts assigned equally?

The first randomization assigned which context the respondent would see first. Check to see if that randomization evenly distributed the order of contexts seen.

```
dc[, .N, by = First.Context]
```

```

##      First.Context    N
## 1:      Recidivism 320
## 2:         Twitter 321

```

320 received the 'Recidivism' context first. 321 received the 'Twitter' context first. This was a pretty even split. Next we check if each context received similar assignment to treatment.

```
dc[, .N, by = .(tTreat, tAssign)]
```

```

##      tTreat tAssign    N
## 1:         0         1 315
## 2:         1         1 312
## 3:         0        NA   14

```

In this instance, we see that of those assigned to either treatment or control in the Twitter context (tAssign), it was a pretty even split between treatment and control. However, those 14 that were not assigned to either treatment or control are indicative of attrition that we will need to review in greater detail.

```
dc[, .N, by = .(rTreat, rAssign)]
```

```

##      rTreat rAssign    N
## 1:         0         1 312
## 2:         1         1 316
## 3:         0        NA   13

```

Similarly, we see a pretty even split between recidivism context assignment, with another 13 instances of attrition. These could overlap with the other examples of attrition.

Was treatment assigned equally across contexts?

```
dc[complete == 1, .N, by = .(First.Context, tTreat, rTreat)]
```

```
##      First.Context tTreat rTreat  N
## 1:      Recidivism      0      0 78
## 2:         Twitter      1      0 78
## 3:      Recidivism      1      1 79
## 4:         Twitter      0      0 78
## 5:         Twitter      1      1 78
## 6:         Twitter      0      1 79
## 7:      Recidivism      0      1 77
## 8:      Recidivism      1      0 75
```

Of the eight possible combinations of context order, Twitter treatment, and recidivism treatment, there are a fairly equal number of respondents who completed the survey in each category. This shows that our randomization worked at every level.

Were all questions answered?

```
## Show number of responses for each question
apply(dt, 2, function(x) length(which(!is.na(x))))
```

```
##      ResponseID      Finished First.Context      random      intro
##           641           641           641           622           641
##      tweet      tControl      tcFair      tcAcc      tcSat
##           631           315           315           315           315
##      tcUseful      tcClear      tcMeaningful      tcReqInfo1      tcReqInfo2
##           315           315           315           84           173
##      tcReqInfo3      tcReqInfo4 tcReqInfo4_txt      tTreat      ttFair
##           145           9           641           312           312
##      ttAcc      ttSat      ttUseful      ttClear      ttMeaningful
##           312           312           311           311           311
##      ttReqInfo1      ttReqInfo2      ttReqInfo3      ttReqInfo4 ttReqInfo4_txt
##           93           144           129           15           641
##      recidivism      rControl      rcFair      rcAcc      rcSat
##           636           312           312           312           312
##      rcUseful      rcClear      rcMeaningful      rcReqInfo1      rcReqInfo2
##           310           310           310           90           176
##      rcReqInfo3      rcReqInfo4 rcReqInfo4_txt      rTreat      rtFair
##           175           12           641           316           316
##      rtAcc      rtSat      rtUseful      rtClear      rtMeaningful
##           316           316           315           315           315
##      rtReqInfo1      rtReqInfo2      rtReqInfo3      rtReqInfo4 rtReqInfo4_txt
##           89           165           154           15           641
##      ageGroup      white      black      native      asian
##           622           422           36           14           134
##      pac_isle      hispanic      other      gender      socMed
##           1           30           5           620           622
##      educ      feedback
##           622           637
```

From this, it appears that there were a couple instances of attrition in the middle of answering questions about a treatment. Note the drop from 312 to 311 between ttSat and ttUseful, or the drop from 316 to 315 between rtSat and rtUseful.

Attrition effects

Out of 641 surveys, 622 were completed. Was either context more impacted than the other?

```
dc[, sum(complete)/.N, by = First.Context]
```

```
##      First.Context      V1
## 1:      Recidivism 0.9656250
## 2:       Twitter 0.9750779
```

Similar ratios completed the survey regardless of which context they started with. This does not seem indicative of a problem with the experiment, but we will need to be careful about how we calculate effects.

```
dc[, .N, by = .(First.Context, tTreat, rTreat, tAssign, rAssign)]
```

```
##      First.Context tTreat rTreat tAssign rAssign  N
## 1:      Recidivism      0      0      1      1  78
## 2:       Twitter      1      0      1      1  79
## 3:      Recidivism      1      1      1      1  79
## 4:       Twitter      0      0      1      1  78
## 5:       Twitter      1      1      1      1  78
## 6:       Twitter      0      1      1      1  80
## 7:      Recidivism      0      1      1      1  77
## 8:      Recidivism      1      0      1      1  75
## 9:      Recidivism      0      0     NA     NA   7
## 10:     Twitter      0      0     NA     NA   3
## 11:     Twitter      0      0      1     NA   2
## 12:     Twitter      1      0      1     NA   1
## 13:      Recidivism      0      1     NA      1   2
## 14:      Recidivism      0      0     NA      1   2
```

To look again at all possible combinations, we see that the largest number of dropouts we had were the 10 who were assigned a context but did not make it far enough to be assigned to treatment or control for either context. These respondents must have followed the link to Qualtrics but then dropped out without doing anything within Qualtrics. The other instances of attrition are pretty small and even (1 or 2 for each group who did not make it to a second context). Overall, there might be a very small effect because of attrition, but it does not seem to be due to the experiment design or content and does not affect the contexts differently.

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. Again, each of these responses were based on a 5 point Likert scale.

Visual data exploration, grouped bars

Twitter Response Histograms

```
par(mfrow=c(2,3))
hist(dt$tcFair, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Fairness", xlab="Fairness", ylim=c(0,175))
hist(dt$ttFair, col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)
#legend("topright", c("Control", "Treatment"), fill=c("blue", "red"))
```

```

hist(dt$tcAcc, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Accuracy", xlab="Accuracy", ylim=c(0,175))
hist(dt$ttAcc,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)

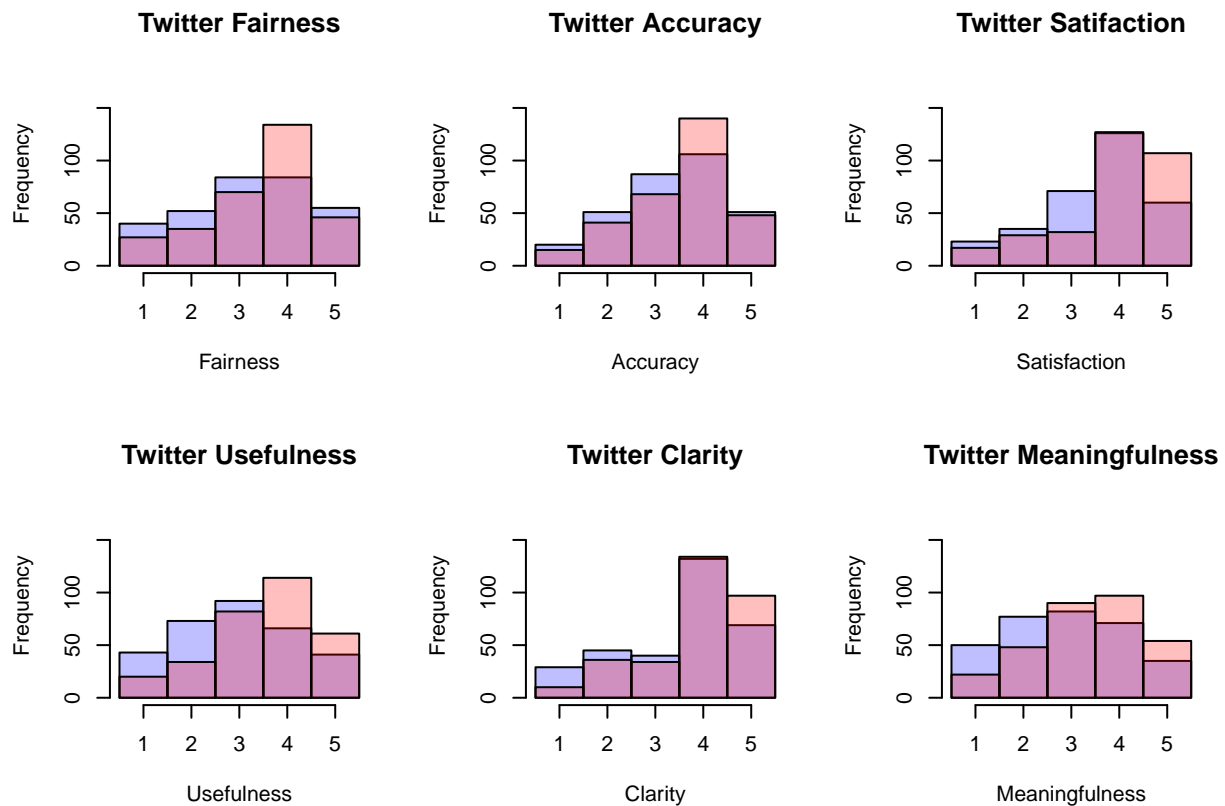
hist(dt$tcSat, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Satisfaction", xlab="Satisfaction", ylim=c(0,175))
hist(dt$ttSat,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$tcUseful, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Usefulness", xlab="Usefulness", ylim=c(0,175))
hist(dt$ttUseful,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$tcClear, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Clarity", xlab="Clarity", ylim=c(0,175))
hist(dt$ttClear,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$tcMeaningful, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Meaningfulness", xlab="Meaningfulness", ylim=c(0,175))
hist(dt$ttMeaningful,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)

```



Recidivism Responses Histogram

```

par(mfrow=c(2,3))
hist(dt$rcFair, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),

```

```

    main= "Recidivism Fairness", xlab="Fairness", ylim=c(0,175))
hist(dt$rtFair,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)
#legend("topright", c("Control", "Treatment"), fill=c("blue", "red"))

hist(dt$rcAcc, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Accuracy", xlab="Accuracy", ylim=c(0,175))
hist(dt$rtAcc,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)

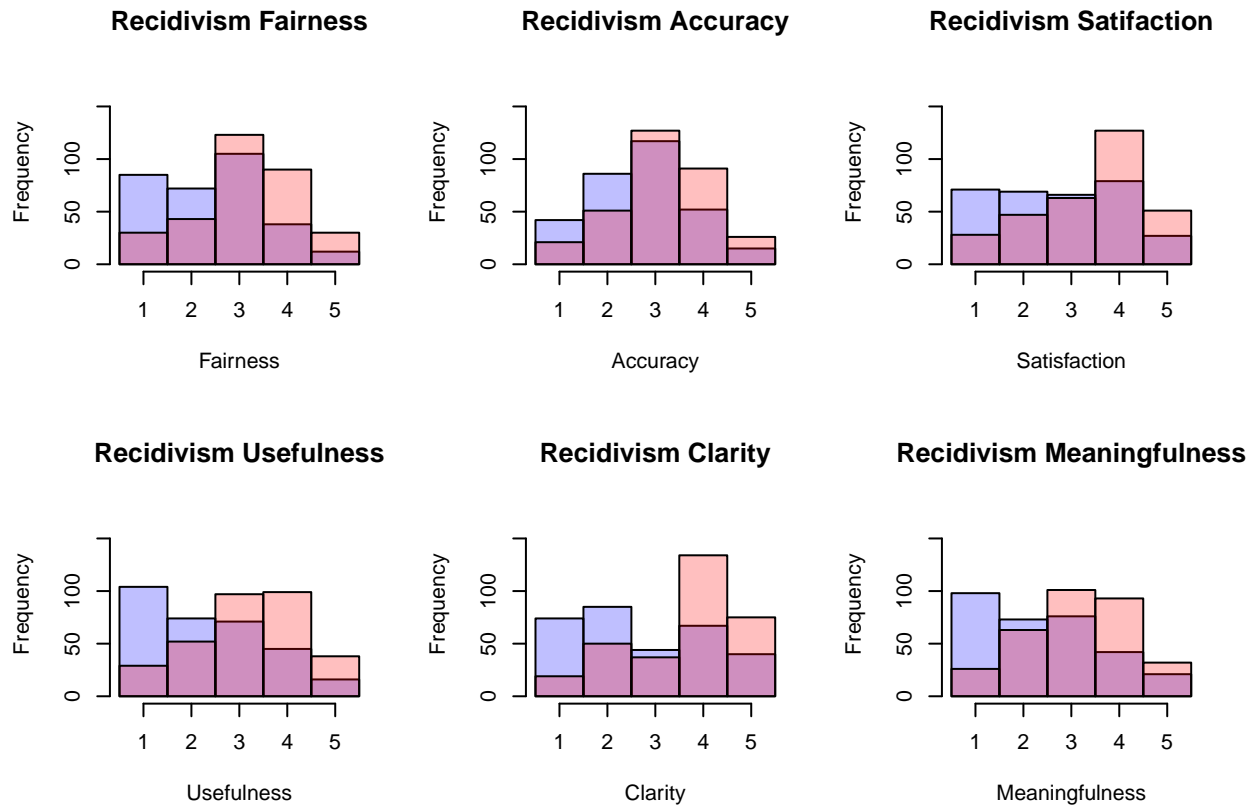
hist(dt$rcSat, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Satisfaction", xlab="Satisfaction", ylim=c(0,175))
hist(dt$rtSat,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$rcUseful, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Usefulness", xlab="Usefulness", ylim=c(0,175))
hist(dt$rtUseful,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$rcClear, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Clarity", xlab="Clarity", ylim=c(0,175))
hist(dt$rtClear,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$rcMeaningful, col=rgb(0,0,1,1/4), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Meaningfulness", xlab="Meaningfulness", ylim=c(0,175))
hist(dt$rtMeaningful,col=rgb(1,0,0,1/4), breaks = seq(0.5, 5.5, 1), add=T)

```



In both contexts, we can see a difference between control and treatment that increases each of the metrics under treatment.

Twitter Histograms Greyscale

```
par(mfrow=c(2,3))
hist(dt$tcFair, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Fairness", xlab="Fairness", ylim=c(0,175))
hist(dt$ttFair,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)
#legend("topright", c("Control", "Treatment"), fill=c("blue", "red"))

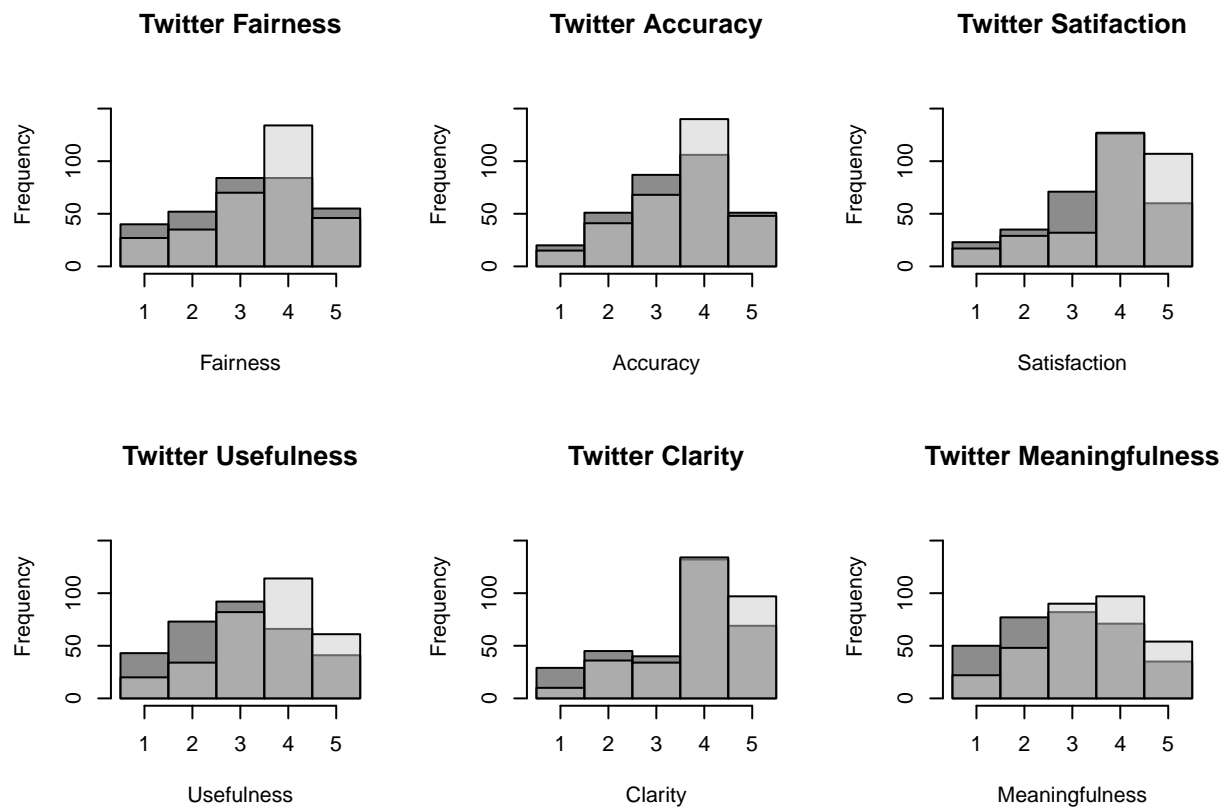
hist(dt$tcAcc, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Accuracy", xlab="Accuracy", ylim=c(0,175))
hist(dt$ttAcc,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$tcSat, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Satisfaction", xlab="Satisfaction", ylim=c(0,175))
hist(dt$ttSat,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$tcUseful, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Usefulness", xlab="Usefulness", ylim=c(0,175))
hist(dt$ttUseful,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$tcClear, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Clarity", xlab="Clarity", ylim=c(0,175))
hist(dt$ttClear,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$tcMeaningful, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Twitter Meaningfulness", xlab="Meaningfulness", ylim=c(0,175))
hist(dt$ttMeaningful,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)
```



Recidivism Histogram Greyscale

```
par(mfrow=c(2,3))
hist(dt$rcFair, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Fairness", xlab="Fairness", ylim=c(0,175))
hist(dt$rtFair,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)
#legend("topright", c("Control", "Treatment"), fill=c("blue", "red"))

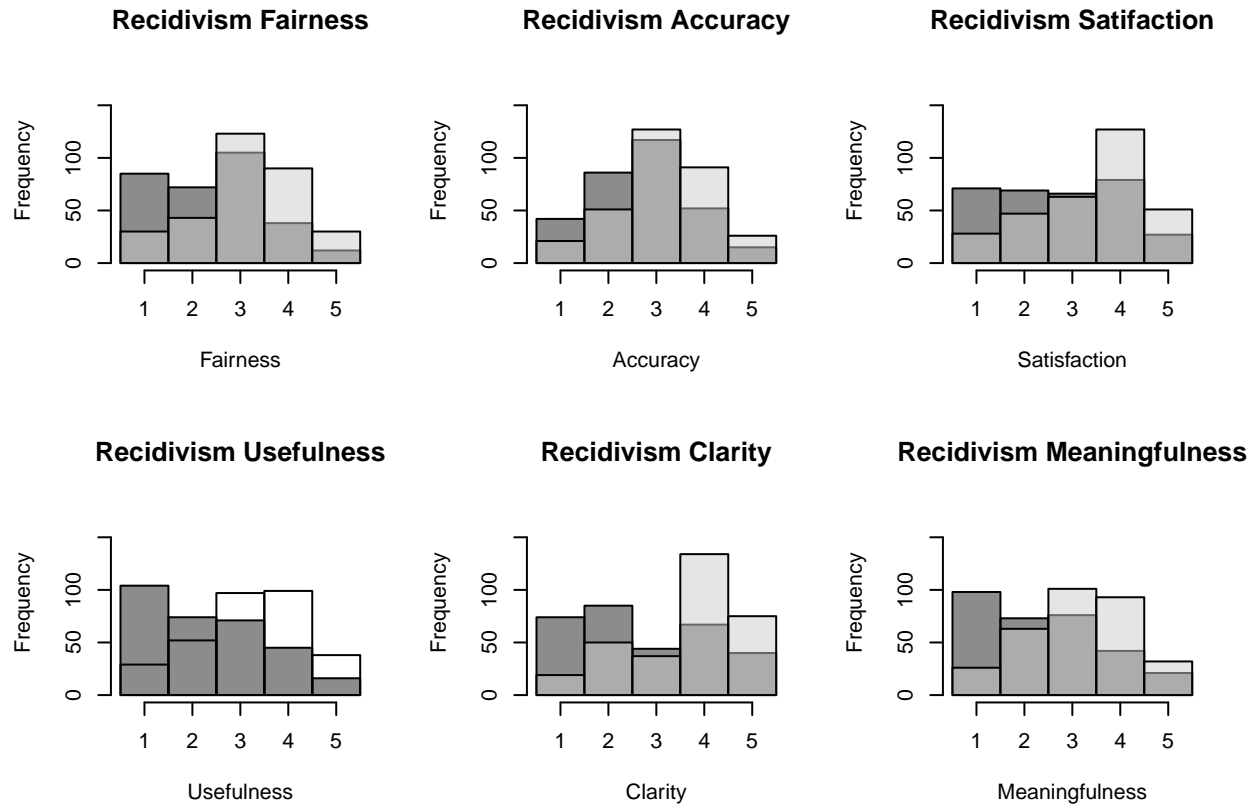
hist(dt$rcAcc, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Accuracy", xlab="Accuracy", ylim=c(0,175))
hist(dt$rtAcc,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$rcSat, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Satisfaction", xlab="Satisfaction", ylim=c(0,175))
#legend(4,9, Treat(df),lwd=4, col=c())
hist(dt$rtSat,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$rcUseful, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Usefulness", xlab="Usefulness", ylim=c(0,175))
hist(dt$rtUseful,colx=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)

hist(dt$rcClear, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Clarity", xlab="Clarity", ylim=c(0,175))
hist(dt$rtClear,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)
```

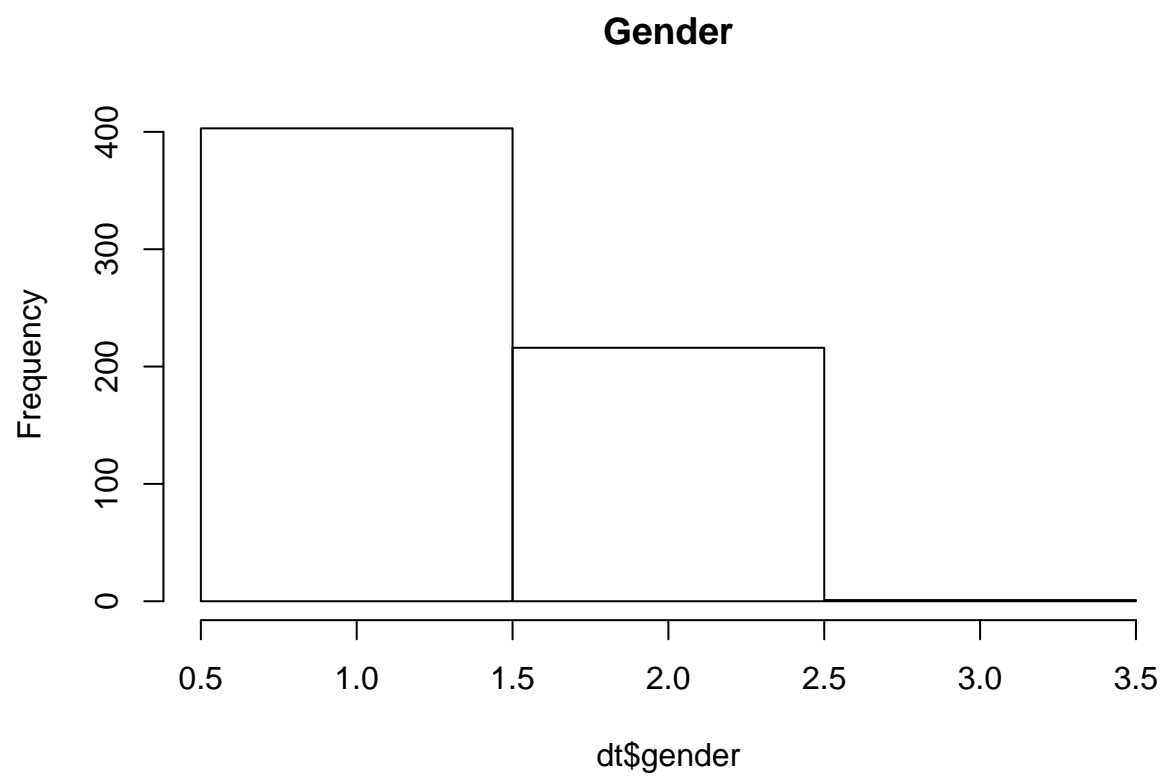
```
hist(dt$rcMeaningful, col=rgb(0.1,0.1,0.1,0.5), breaks = seq(0.5, 5.5, 1),
     main= "Recidivism Meaningfulness", xlab="Meaningfulness", ylim=c(0,175))
hist(dt$rtMeaningful,col=rgb(0.8,0.8,0.8,0.5), breaks = seq(0.5, 5.5, 1), add=T)
```



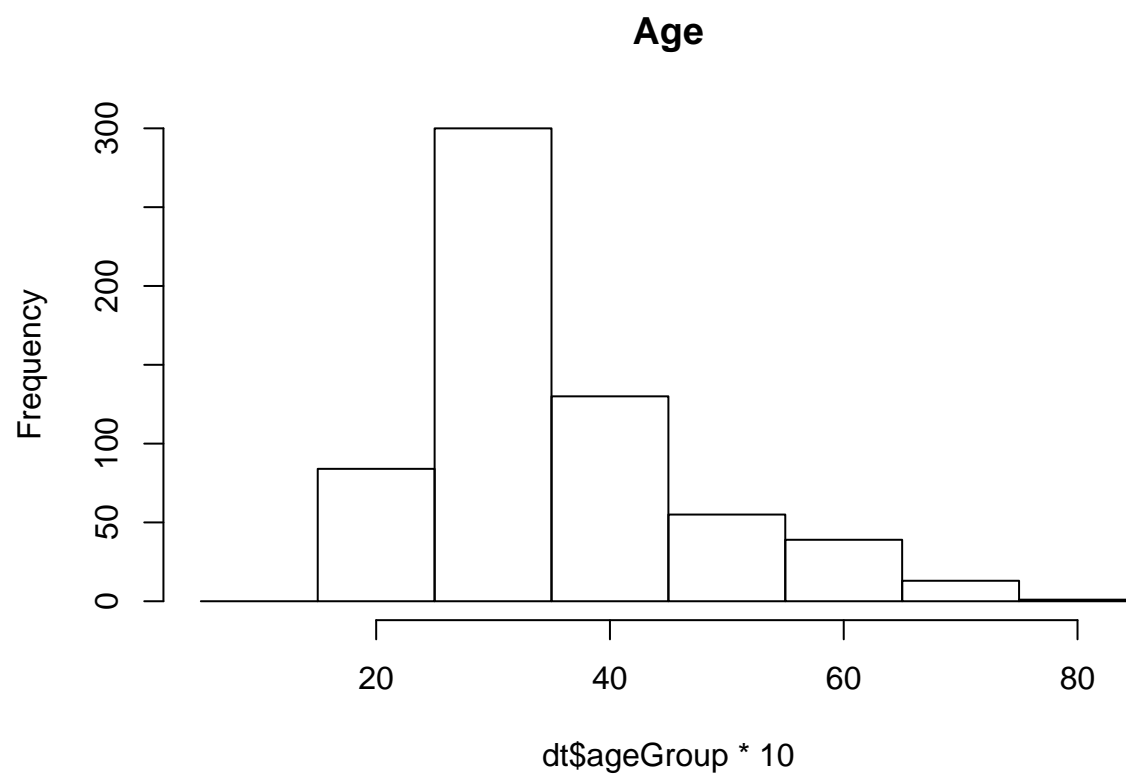
Demographic Data Review

```
#ageGroup', 'race', 'gender', 'socMed', 'educ', 'feedback', 'duration'
# par(mfrow=c(2,2))

hist(dt$gender, main = "Gender", breaks = seq(0.5, 3.5, 1))
```

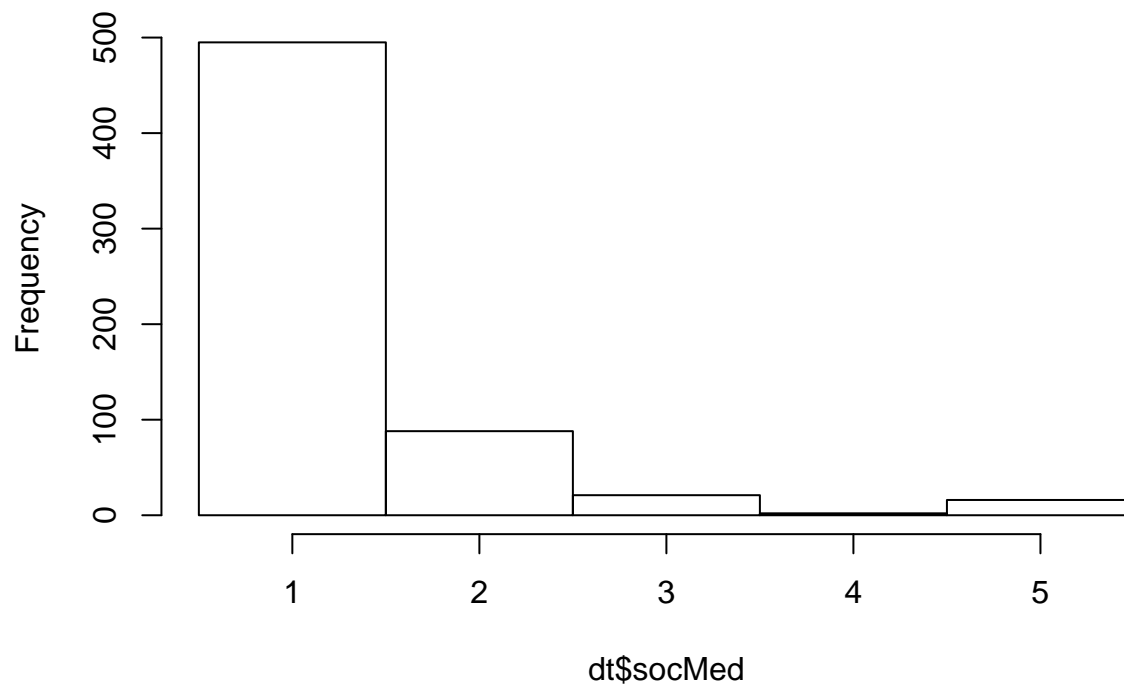


```
hist(dt$AgeGroup*10, main = "Age", breaks = seq(5, 85, 10))
```



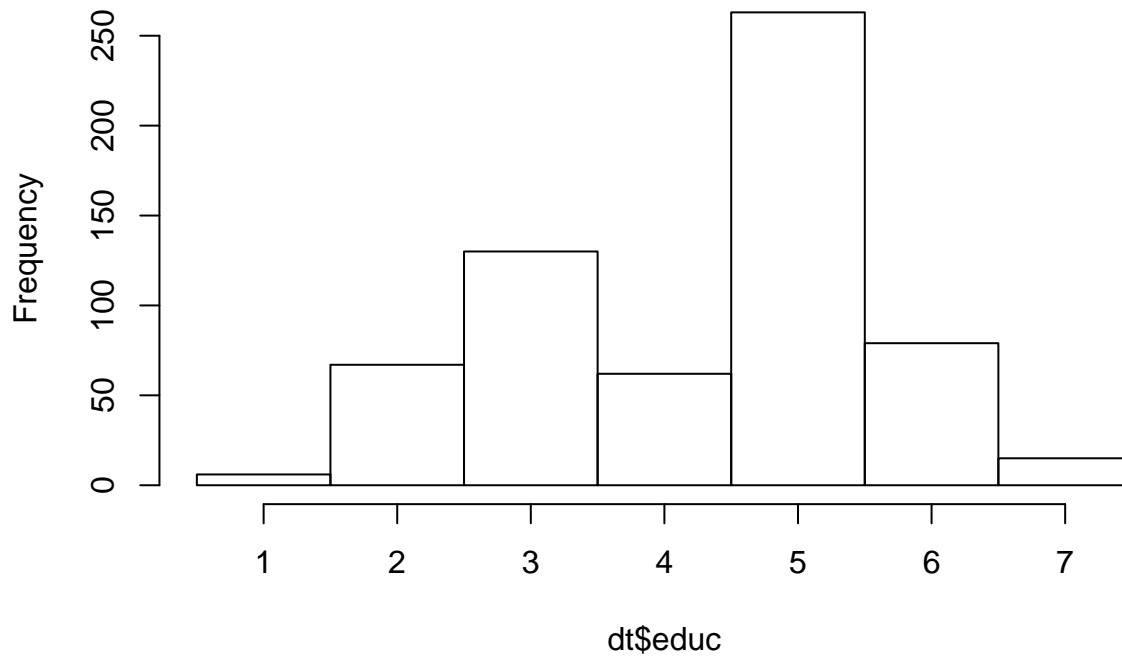
```
hist(dt$socMed, main= "Social Media Usage", breaks = seq(0.5, 5.5, 1))
```

Social Media Usage

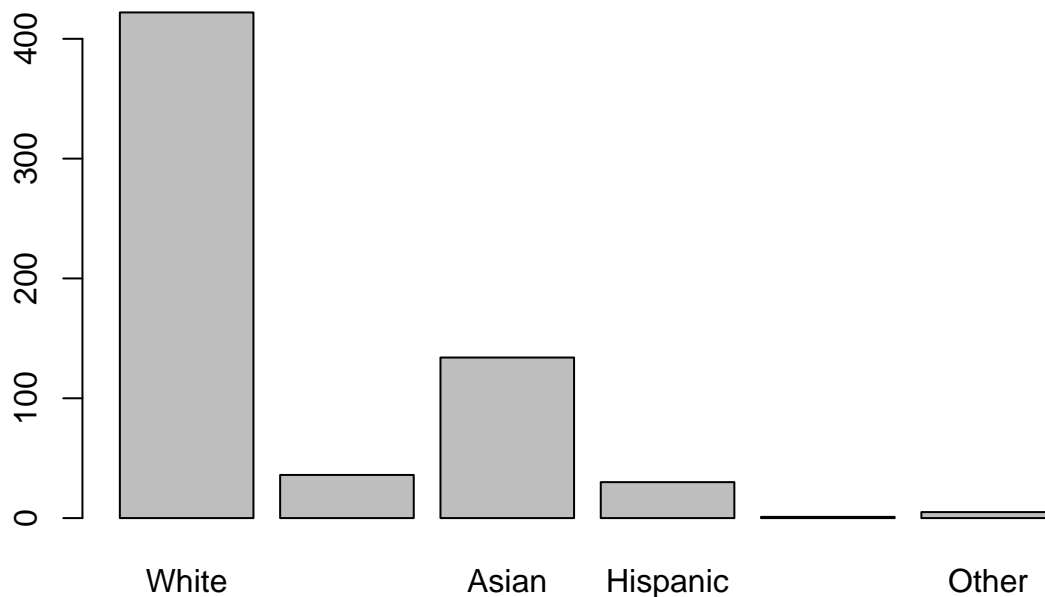


```
hist(dt$educ, main = "Educational Level", breaks = seq(0.5, 7.5, 1))
```

Educational Level



```
ethnic <- c("White", "African American", "Asian", "Hispanic", "Pacific Islander", "Other")
ethnicities <- data.table(sum(!is.na(dt$white)), sum(!is.na(dt$black)), sum(!is.na(dt$asian)),
                          sum(!is.na(dt$hispanic)), sum(!is.na(dt$pac_isle)), sum(!is.na(dt$other)))
ethnicities2 = transpose(ethnicities)
barplot(ethnicities2$V1, names = ethnic)
```



```
dc[, .N, by = .(white, black, asian, hispanic, pac_isle, other)]
```

```
##      white black asian hispanic pac_isle other   N
## 1:      1      0      0          0          0      0 409
## 2:      0      1      0          0          0      0 30
## 3:      0      0      1          0          0      0 127
## 4:      0      0      0          1          0      0 26
## 5:      1      1      0          0          0      0  4
## 6:      0      0      0          0          0      1  5
## 7:      0      0      0          0          0      0 29
## 8:      1      0      1          0          0      0  4
## 9:      1      0      0          1          0      0  4
## 10:     1      0      1          0          1      0  1
## 11:     0      1      1          0          0      0  2
```

Based on a quick look at our survey demographic responses, we see that approximately 2/3 of the respondents are male. The respondents also skew young, as almost half are between 25 and 34. Nearly 500 of our 641 respondents use social media daily, which may be biasing our results. More than half have completed a 4 year degree or higher. The respondents are also over 2/3 white. Only 5 identify as none of the given options, 1 selected Pacific Islander, but also identifies as white and Asian. 29 respondents (5%) respondents did not answer the question.

Regression Models

A basic view of the data showed there was a change in the distributions. Regression models will allow us to gauge the significance of these changes. We have create linear models for each question for each context

(Twitter and recidivism). The models subset the data to look only at those respondents that were assigned to either treatment or control for that context. In this way, someone who attrited in the first context will not count against the second context.

Twitter Moderation

```
mtFair <- lm(tFair ~ tTreat, data = dc[tAssign == 1])
mtFair$se <- sqrt(diag(vcovHC(mtFair)))
mtFair$p <- coeftest(mtFair, vcovHC(mtFair))[ , 4]

mtAcc <- lm(tAcc ~ tTreat, data = dc[tAssign == 1])
mtAcc$se <- sqrt(diag(vcovHC(mtAcc)))
mtAcc$p <- coeftest(mtAcc, vcovHC(mtAcc))[ , 4]

mtSat <- lm(tSat ~ tTreat, data = dc[tAssign == 1])
mtSat$se <- sqrt(diag(vcovHC(mtSat)))
mtSat$p <- coeftest(mtSat, vcovHC(mtSat))[ , 4]

mtUseful <- lm(tUseful ~ tTreat, data = dc[tAssign == 1])
mtUseful$se <- sqrt(diag(vcovHC(mtUseful)))
mtUseful$p <- coeftest(mtUseful, vcovHC(mtUseful))[ , 4]

mtClear <- lm(tClear ~ tTreat, data = dc[tAssign == 1])
mtClear$se <- sqrt(diag(vcovHC(mtClear)))
mtClear$p <- coeftest(mtClear, vcovHC(mtClear))[ , 4]

mtMeaningful <- lm(tMeaningful ~ tTreat, data = dc[tAssign == 1])
mtMeaningful$se <- sqrt(diag(vcovHC(mtMeaningful)))
mtMeaningful$p <- coeftest(mtMeaningful, vcovHC(mtMeaningful))[ , 4]

stargazer(mtFair, mtAcc, mtSat, mtUseful, mtClear, mtMeaningful,
  type = 'text',
  se = list(mtFair$se, mtAcc$se, mtSat$se, mtUseful$se, mtClear$se, mtMeaningful$se),
  p = list(mtFair$p, mtAcc$p, mtSat$p, mtUseful$p, mtClear$p, mtMeaningful$p),
  covariate.labels = c("Explanation"),
  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)
Explanation	0.242** (0.096)	0.157* (0.087)	0.367*** (0.091)	0.545*** (0.095)	0.332*** (0.094)	0.467*** (0.096)
Constant	3.197*** (0.071)	3.371*** (0.063)	3.524*** (0.064)	2.965*** (0.069)	3.530*** (0.070)	2.886*** (0.070)

```
## Observations          627      627      627      627      627      627
## R2                    0.010    0.005    0.025    0.050    0.020    0.036
## Adjusted R2           0.008    0.004    0.024    0.049    0.018    0.035
## Residual Std. Error (df = 625) 1.203    1.091    1.139    1.183    1.170    1.202
## F Statistic (df = 1; 625)   6.357**   3.266*   16.292***  33.215*** 12.623*** 23.643***
```

```
## =====
## Note:                                                         *p<0.1; **p<0.05; ***p<0.01
```

```
dc[(tAssign == 1 & tSat == 0), .N]
```

```
## [1] 0
```

```
dc[(tAssign == 1 & tMeaningful == 0), .N]
```

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

This shows that 1 person dropped out between seeing the treatment and responding in the Twitter context. This is probably not affecting our last few metrics, but they are all statistically significant by a large margin anyway. This represents our intent to treat effect for the questions they did not answer. However, they get through all of the first three questions, so those responses are not affected by attrition.

```
mrFair <- lm(rFair ~ rTreat, data = dc[rAssign == 1])
mrFair$se <- sqrt(diag(vcovHC(mrFair)))
mrFair$p <- coeftest(mrFair, vcovHC(mrFair))[ , 4]
```

```
mrAcc <- lm(rAcc ~ rTreat, data = dc[rAssign == 1])
mrAcc$se <- sqrt(diag(vcovHC(mrAcc)))
mrAcc$p <- coeftest(mrAcc, vcovHC(mrAcc))[ , 4]
```

```
mrSat <- lm(rSat ~ rTreat, data = dc[rAssign == 1])
mrSat$se <- sqrt(diag(vcovHC(mrSat)))
mrSat$p <- coeftest(mrSat, vcovHC(mrSat))[ , 4]
```

```
mrUseful <- lm(rUseful ~ rTreat, data = dc[rAssign == 1])
mrUseful$se <- sqrt(diag(vcovHC(mrUseful)))
mrUseful$p <- coeftest(mrUseful, vcovHC(mrUseful))[ , 4]
```

```
mrClear <- lm(rClear ~ rTreat, data = dc[rAssign == 1])
mrClear$se <- sqrt(diag(vcovHC(mrClear)))
mrClear$p <- coeftest(mrClear, vcovHC(mrClear))[ , 4]
```

```
mrMeaningful <- lm(rMeaningful ~ rTreat, data = dc[rAssign == 1])
mrMeaningful$se <- sqrt(diag(vcovHC(mrMeaningful)))
mrMeaningful$p <- coeftest(mrMeaningful, vcovHC(mrMeaningful))[ , 4]
```

```
stargazer(mrFair, mrAcc, mrSat, mrUseful, mrClear, mrMeaningful,
  type = 'text',
  se = list(mrFair$se, mrAcc$se, mrSat$se, mrUseful$se, mrClear$se, mrMeaningful$se),
  p = list(mrFair$p, mrAcc$p, mrSat$p, mrUseful$p, mrClear$p, mrMeaningful$p),
  covariate.labels = c("Explanation"),
  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)
## -----
## Explanation 0.726*** 0.440*** 0.649*** 0.872*** 0.906*** 0.736***
## (0.088) (0.082) (0.099) (0.095) (0.104) (0.095)
##
## Constant 2.423*** 2.718*** 2.750*** 2.324*** 2.705*** 2.388***
## (0.064) (0.059) (0.073) (0.070) (0.079) (0.071)
##
## -----
## Observations 628 628 628 628 628 628
## R2 0.098 0.044 0.064 0.118 0.109 0.088
## Adjusted R2 0.097 0.042 0.063 0.117 0.108 0.086
## Residual Std. Error (df = 626) 1.102 1.029 1.239 1.192 1.295 1.189
## F Statistic (df = 1; 626) 68.079*** 28.723*** 43.073*** 84.045*** 76.767*** 60.139***
## =====
## Note: *p<0.1; **p<0.05; ***p<0.01
```

```
dc[(rAssign == 1 & rSat == 0), .N]
```

```
## [1] 0
```

```
dc[(rAssign == 1 & rMeaningful == 0), .N]
```

```
## [1] 3
```

This shows that 3 people dropped out between seeing the treatment and responding in the recidivism context. This could be throwing off the last few metrics, but those are all statistically significant by a large margin. This represents our intent to treat effect.

Comparison of Contexts

Our second hypothesis asked if there was a difference between how respondents evaluated the explanation in two different contexts of varying importance or personal significance.

```
dc2 <- melt(dc, id.vars = c('ResponseID', 'tAssign', 'tControl', 'rAssign',
                           'rControl', "tweet", "recidivism", "tFair", "tAcc",
                           "tSat", "tUseful", "tClear", "tMeaningful", "rFair",
                           "rAcc", "rSat", "rUseful", "rClear", "rMeaningful"),
           measure.vars = c('tTreat', 'rTreat'))

dc2[, Fair := (variable == 'tTreat')*tFair + (variable == 'rTreat')*rFair]
dc2[, Acc := (variable == 'tTreat')*tAcc + (variable == 'rTreat')*rAcc]
dc2[, Sat := (variable == 'tTreat')*tSat + (variable == 'rTreat')*rSat]
dc2[, Useful := (variable == 'tTreat')*tUseful + (variable == 'rTreat')*rUseful]
dc2[, Clear := (variable == 'tTreat')*tClear + (variable == 'rTreat')*rClear]
dc2[, Meaningful := (variable == 'tTreat')*tMeaningful + (variable == 'rTreat')*rMeaningful]
names(dc2)[names(dc2) == "variable"] = "Context"
names(dc2)[names(dc2) == "value"] = "treat"

dc2[,c("tFair", "tAcc", "tSat", "tUseful", "tClear", "tMeaningful", "rFair", "rAcc", "rSat",
       "rUseful", "rClear", "rMeaningful"):=NULL]

mFair <- lm(Fair ~ factor(Context) + treat + treat*factor(Context),
           data = dc2[rAssign == 1 & tAssign == 1])
```

```

mFair$se <- sqrt(diag(vcovHC(mFair)))
mFair$p <- coeftest(mFair, vcovHC(mFair))[, 4]

mAcc <- lm(Acc ~ factor(Context) + treat + treat*factor(Context),
           data = dc2[rAssign == 1 & tAssign == 1])
mAcc$se <- sqrt(diag(vcovHC(mAcc)))
mAcc$p <- coeftest(mAcc, vcovHC(mAcc))[, 4]

mSat <- lm(Sat ~ factor(Context) + treat + treat*factor(Context),
           data = dc2[rAssign == 1 & tAssign == 1])
mSat$se <- sqrt(diag(vcovHC(mSat)))
mSat$p <- coeftest(mSat, vcovHC(mSat))[, 4]

mClear <- lm(Clear ~ factor(Context) + treat + treat*factor(Context),
            data = dc2[rAssign == 1 & tAssign == 1])
mClear$se <- sqrt(diag(vcovHC(mClear)))
mClear$p <- coeftest(mClear, vcovHC(mClear))[, 4]

mUseful <- lm(Useful ~ factor(Context) + treat + treat*factor(Context),
            data = dc2[rAssign == 1 & tAssign == 1])
mUseful$se <- sqrt(diag(vcovHC(mUseful)))
mUseful$p <- coeftest(mUseful, vcovHC(mUseful))[, 4]

mMeaningful <- lm(Meaningful ~ factor(Context) + treat + treat*factor(Context),
                 data = dc2[rAssign == 1 & tAssign == 1])
mMeaningful$se <- sqrt(diag(vcovHC(mMeaningful)))
mMeaningful$p <- coeftest(mMeaningful, vcovHC(mMeaningful))[, 4]

stargazer(mFair, mAcc, mSat, mClear, mUseful, mMeaningful, type = 'text',
          se = list(mFair$se, mAcc$se, mSat$se, mUseful$se, mClear$se, mMeaningful$se),
          p = list(mFair$p, mAcc$p, mSat$p, mUseful$p, mClear$p, mMeaningful$p),
          covariate.labels = c("Recidivism Context", "Treatment", "Recidivism Treatment"),
          dep.var.labels = c("Fairness", "Accuracy", "Satisfaction", "Usefulness",
                             "Clarity", "Meaningfulness"),
          dep.var.caption = c("Context Comparison"))

```

```
##
## =====
```

	Context Comparison					
	Fairness	Accuracy	Satisfaction	Usefulness	Clarity	Meaningfulness
	(1)	(2)	(3)	(4)	(5)	(6)
## Recidivism Context	-0.795***	-0.671***	-0.798***	-0.830***	-0.649***	-0.504***
##	(0.096)	(0.087)	(0.097)	(0.099)	(0.105)	(0.100)
## Treatment	0.233**	0.147*	0.354***	0.335***	0.546***	0.469***
##	(0.096)	(0.087)	(0.091)	(0.094)	(0.093)	(0.096)
## Recidivism Treatment	0.510***	0.306**	0.315**	0.577**	0.332***	0.275**
##	(0.131)	(0.120)	(0.135)	(0.134)	(0.139)	(0.135)
## Constant	3.204***	3.380***	3.534***	3.540***	2.974***	2.895***
##	(0.072)	(0.063)	(0.064)	(0.069)	(0.070)	(0.070)

```
##
## -----
## Observations          1,248      1,248      1,248      1,248      1,248      1,248
## R2                    0.101      0.078      0.110      0.113      0.121      0.084
## Adjusted R2           0.098      0.076      0.108      0.111      0.119      0.082
## Residual Std. Error (df = 1244) 1.151      1.059      1.186      1.223      1.179      1.187
## F Statistic (df = 3; 1244)  46.403*** 35.233*** 51.221*** 52.900*** 57.212*** 38.257***
## =====
## Note:                                                         *p<0.1; **p<0.05; ***p<0.01
```

When comparing contexts, we see statistical significance in the baseline constant for all metrics in the fourth row. This represents the constant in the Twitter control group. In the first row of the regression, we see the change to recidivism has a significant negative effect in all metrics. This means that respondents were less accepting of the algorithm's decision without an explanation than they were in the Twitter context. This may be partly attributable to the design of the survey. Criminal recidivism is a complicated problem with more inputs than a 140 character Tweet. Because we included the full Tweet, people were able to judge the appropriateness of the decision by themselves. In the recidivism context, respondents were only given a brief description of the case, with just the offense the defendant was being charged with. Including some information about criminal history or other factors may have made this a more appropriate comparison.

In the second row, we see what is effectively the same significance of the Twitter treatment that we saw in the Twitter only models. The numbers are slightly different here because this analysis looks only at individuals who were assigned to treatment or control in both contexts. So a few instances are missing where an individual did not make it to the second half of their survey. In the third row, we see the effect of the recidivism treatment compared to the effect of the Twitter treatment. Again, we have high statistical significance in all metrics as we did in the recidivism only models. However, the significance of the differences is less than of the treatment itself, dropping in 4 of the 6 cases below the 0.01 level, although still significant at a level of 0.05.

Difference in Order

We also discussed looking at the difference in responses depending on the order of contexts. Significant effects here would show whether answering one context first created a bias in the response to the second context.

```
otFair <-      lm(tFair ~ First.Context + tTreat + rTreat + tTreat*rTreat,
                  data = dc[tAssign == 1])
otFair$se <- sqrt(diag(vcovHC(otFair)))
otFair$p <- coeftest(otFair, vcovHC(otFair))[ , 4]

otAcc <-      lm(tAcc ~ First.Context+ tTreat + rTreat + tTreat*rTreat,
                  data = dc[tAssign == 1])
otAcc$se <- sqrt(diag(vcovHC(otAcc)))
otAcc$p <- coeftest(otAcc, vcovHC(otAcc))[ , 4]

otSat <-      lm(tSat ~ First.Context+ tTreat + rTreat + tTreat*rTreat,
                  data = dc[tAssign == 1])
otSat$se <- sqrt(diag(vcovHC(otSat)))
otSat$p <- coeftest(otSat, vcovHC(otSat))[ , 4]

otUseful <-    lm(tUseful ~ First.Context+ tTreat + rTreat + tTreat*rTreat,
                  data = dc[tAssign == 1])
otUseful$se <- sqrt(diag(vcovHC(otUseful)))
otUseful$p <- coeftest(otUseful, vcovHC(otUseful))[ , 4]
```

```

otClear <- lm(tClear ~ First.Context+ tTreat + rTreat + tTreat*rTreat,
              data = dc[tAssign == 1])
otClear$se <- sqrt(diag(vcovHC(otClear)))
otClear$p <- coeftest(otClear, vcovHC(otClear))[ , 4]

otMeaningful <- lm(tMeaningful ~ First.Context+ tTreat + rTreat + tTreat*rTreat,
                  data = dc[tAssign == 1])
otMeaningful$se <- sqrt(diag(vcovHC(otMeaningful)))
otMeaningful$p <- coeftest(otMeaningful, vcovHC(otMeaningful))[ , 4]

stargazer(otFair, otAcc, otSat, otUseful, otClear, otMeaningful,
           type = 'text',
           se = list(otFair$se, otAcc$se, otSat$se, otUseful$se, otClear$se, otMeaningful$se),
           p = list(otFair$p, otAcc$p, otSat$p, otUseful$p, otClear$p, otMeaningful$p),
           covariate.labels = c("Twitter First", "Twitter Treatment",
                                "Recidivism Treatment", "Both Treatments" ),
           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)
## -----
## Twitter First                    0.102    0.153*    0.160*    0.012    0.018    -0.112
##                                (0.096)  (0.087)  (0.091)  (0.095)  (0.094)  (0.096)
##
## Twitter Treatment                0.268**   0.224*    0.517***   0.656***  0.448***  0.629***
##                                (0.136)  (0.127)  (0.131)  (0.138)  (0.131)  (0.140)
##
## Recidivism Treatment            -0.050    0.110     0.123     0.108    0.124     0.254*
##                                (0.144)  (0.128)  (0.129)  (0.139)  (0.140)  (0.140)
##
## Both Treatments                 -0.050   -0.133    -0.298    -0.223   -0.231   -0.324*
##                                (0.193)  (0.175)  (0.182)  (0.190)  (0.188)  (0.192)
##
## Constant                       3.170***  3.239***  3.381***  2.905***  3.459***  2.816***
##                                (0.113)  (0.100)  (0.103)  (0.112)  (0.109)  (0.112)
##
## -----
## Observations                    627      627      627      627      627      627
## R2                             0.013     0.011     0.035     0.053     0.022     0.044
## Adjusted R2                    0.007     0.005     0.028     0.047     0.016     0.038
## Residual Std. Error (df = 622) 1.204     1.090     1.136     1.184     1.171     1.200
## F Statistic (df = 4; 622)      2.040*    1.801     5.585***  8.635***  3.541***  7.214***
## =====
## 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. There is significance to receiving the Twitter treatment, but there is no significance to receiving the Twitter context first. There is also not significance to whether or not the recidivism treatment was received.

```

orFair <- lm(rFair ~ First.Context + rTreat + tTreat + rTreat*tTreat,
            data = dc[rAssign == 1])
orFair$se <- sqrt(diag(vcovHC(orFair)))
orFair$p <- coeftest(orFair, vcovHC(orFair))[, 4]

orAcc <- lm(rAcc ~ First.Context + rTreat + tTreat + rTreat*tTreat,
            data = dc[rAssign == 1])
orAcc$se <- sqrt(diag(vcovHC(orAcc)))
orAcc$p <- coeftest(orAcc, vcovHC(orAcc))[, 4]

orSat <- lm(rSat ~ First.Context + rTreat + tTreat + rTreat*tTreat,
            data = dc[rAssign == 1])
orSat$se <- sqrt(diag(vcovHC(orSat)))
orSat$p <- coeftest(orSat, vcovHC(orSat))[, 4]

orUseful <- lm(rUseful ~ First.Context + rTreat + tTreat + rTreat*tTreat,
              data = dc[rAssign == 1])
orUseful$se <- sqrt(diag(vcovHC(orUseful)))
orUseful$p <- coeftest(orUseful, vcovHC(orUseful))[, 4]

orClear <- lm(rClear ~ First.Context + rTreat + tTreat + rTreat*tTreat,
              data = dc[rAssign == 1])
orClear$se <- sqrt(diag(vcovHC(orClear)))
orClear$p <- coeftest(orClear, vcovHC(orClear))[, 4]

orMeaningful <- lm(rMeaningful ~ First.Context + rTreat + tTreat + rTreat*tTreat,
                  data = dc[rAssign == 1])
orMeaningful$se <- sqrt(diag(vcovHC(orMeaningful)))
orMeaningful$p <- coeftest(orMeaningful, vcovHC(orMeaningful))[, 4]

stargazer(orFair, orAcc, orSat, orUseful, orClear, orMeaningful,
          type = 'text',
          se = list(orFair$se, orAcc$se, orSat$se, orUseful$se, orClear$se, orMeaningful$se),
          p = list(orFair$p, orAcc$p, orSat$p, orUseful$p, orClear$p, orMeaningful$p),
          covariate.labels = c("Twitter First", "Recidivism Treatment", "Twitter Treatment",
                                "Both Treatments" ),
          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)
Twitter First	-0.063	-0.043	-0.140	-0.191**	-0.263**	-0.110
	(0.088)	(0.083)	(0.099)	(0.095)	(0.103)	(0.095)
Recidivism Treatment	0.608***	0.430***	0.658***	0.862***	0.885***	0.716***
	(0.128)	(0.113)	(0.142)	(0.133)	(0.147)	(0.134)
Twitter Treatment	0.012	0.019	0.201	0.159	0.100	-0.020

	(0.128)	(0.119)	(0.147)	(0.140)	(0.157)	(0.143)
## Both Treatments	0.236	0.020	-0.020	0.019	0.039	0.040
##	(0.176)	(0.165)	(0.198)	(0.191)	(0.207)	(0.191)
## Constant	2.449***	2.730***	2.721***	2.341***	2.788***	2.453***
##	(0.100)	(0.089)	(0.113)	(0.104)	(0.122)	(0.106)

## Observations	628	628	628	628	628	628
## R2	0.105	0.045	0.073	0.128	0.120	0.090
## Adjusted R2	0.099	0.038	0.067	0.123	0.115	0.084
## Residual Std. Error (df = 623)	1.101	1.031	1.236	1.188	1.290	1.190
## F Statistic (df = 4; 623)	18.199***	7.255***	12.252***	22.951***	21.312***	15.347***
=====						
## Note:	*p<0.1; **p<0.05; ***p<0.01					

In most cases, we see that the only statistical significance is the base rating and the effect of the recidivism treatment. In two cases (Clarity and Usefulness), there is a significant decrease in the metric if Twitter was viewed first. This is perhaps concerning, but the fact that it is negative shows that the actual effect of the recidivism treatment was more positive than we had previously shown.

Influence of Other Factors - Demographics, etc

Gender and Race

```
dtFair <- lm(tFair ~ tTreat + white + female + tTreat*white + tTreat*female, data = dc)
dtFair$se <- sqrt(diag(vcovHC(dtFair, type = "HC1")))
dtFair$p <- coeftest(dtFair, vcovHC(dtFair, type = "HC1"))[, 4]

dtAcc <- lm(tAcc ~ tTreat + white + female + tTreat*white + tTreat*female, data = dc)
dtAcc$se <- sqrt(diag(vcovHC(dtAcc, type = "HC1")))
dtAcc$p <- coeftest(dtAcc, vcovHC(dtAcc, type = "HC1"))[, 4]

dtSat <- lm(tSat ~ tTreat + white + female + tTreat*white + tTreat*female, data = dc)
dtSat$se <- sqrt(diag(vcovHC(dtSat, type = "HC1")))
dtSat$p <- coeftest(dtSat, vcovHC(dtSat, type = "HC1"))[, 4]

dtUseful <- lm(tUseful ~ tTreat + white + female + tTreat*white + tTreat*female, data = dc)
dtUseful$se <- sqrt(diag(vcovHC(dtUseful, type = "HC1")))
dtUseful$p <- coeftest(dtUseful, vcovHC(dtUseful, type = "HC1"))[, 4]

dtClear <- lm(tClear ~ tTreat + white + female + tTreat*white + tTreat*female, data = dc)
dtClear$se <- sqrt(diag(vcovHC(dtClear, type = "HC1")))
dtClear$p <- coeftest(dtClear, vcovHC(dtClear, type = "HC1"))[, 4]

dtMeaningful <- lm(tMeaningful ~ tTreat + white + female + tTreat*white + tTreat*female, data = dc)
dtMeaningful$se <- sqrt(diag(vcovHC(dtMeaningful, type = "HC1")))
dtMeaningful$p <- coeftest(dtMeaningful, vcovHC(dtMeaningful, type = "HC1"))[, 4]

stargazer(dtFair, dtAcc, dtSat, dtUseful, dtClear, dtMeaningful,
          type = 'text',
          se = list(dtFair$se, dtAcc$se, dtSat$se, dtUseful$se, dtClear$se, dtMeaningful$se),
```



```

p = list(dtFair$p, dtAcc$p, dtSat$p, dtUseful$p, dtClear$p, dtMeaningful$p),
covariate.labels = c("Explanation", "White", "Female"),
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)
## -----
## Explanation      0.567***   0.272*   0.563***   0.639***   0.504***   0.670***
##                  (0.170) (0.159) (0.163) (0.174) (0.167) (0.176)
##
## White            0.506***   0.228*   0.102      -0.024   0.102      0.010
##                  (0.150) (0.135) (0.132) (0.152) (0.148) (0.154)
##
## Female           0.231      0.218*   0.334***   0.117      0.230      0.081
##                  (0.143) (0.130) (0.128) (0.149) (0.144) (0.149)
##
## tTreat:white     -0.531***  -0.232   -0.195     -0.139   -0.196     -0.245
##                  (0.201) (0.186) (0.190) (0.202) (0.199) (0.205)
##
## tTreat:female     0.058      0.104   -0.220     0.025    -0.077     -0.063
##                  (0.195) (0.180) (0.186) (0.198) (0.193) (0.202)
##
## Constant         2.783***   3.151*** 3.350***   2.947*** 3.390***   2.856***
##                  (0.128) (0.117) (0.116) (0.134) (0.129) (0.137)
## -----
## Observations      620      620      620      620      620      620
## R2                 0.040     0.025     0.036     0.057     0.029     0.044
## Adjusted R2        0.032     0.017     0.028     0.049     0.022     0.036
## Residual Std. Error (df = 614) 1.190     1.083     1.135     1.176     1.159     1.196
## F Statistic (df = 5; 614) 5.122*** 3.111*** 4.608*** 7.424*** 3.727*** 5.620***
## =====
## Note:                                                     *p<0.1; **p<0.05; ***p<0.01

```

Women were statistically significantly more likely than men to agree with the Twitter decision, although they had no significantly different response to the explanation. The respondents who replied as non-confirming gender also showed highly significant effects in 5 metrics, but this is suspect because it is a very small sample. Only 1 respondent identified as non-confirming.

When looking at ethnicity, we see some significance, but some problems emerge. Previous regressions showed significant negative effects of treatment on Pacific Islanders. However, there were very few Pacific Islanders who responded. In fact, this group represents only 1 person. Also, with so many metrics and variables, it is highly likely to see significance somewhere at a 0.05 level.

Age

Age was defined by groupings of mostly 10 year increments. This requires a regression on the factors of the variable as they are not continuous.

```

dtFair <- lm(tFair ~ tTreat + factor(ageGroup) + tTreat*factor(ageGroup), data = dc)
dtFair$se <- sqrt(diag(vcovHC(dtFair, type = "HC1")))
dtFair$p <- coeftest(dtFair, vcovHC(dtFair, type = "HC1"))[, 4]

dtAcc <- lm(tAcc ~ tTreat + factor(ageGroup) + tTreat*factor(ageGroup), data = dc)
dtAcc$se <- sqrt(diag(vcovHC(dtAcc, type = "HC1")))
dtAcc$p <- coeftest(dtAcc, vcovHC(dtAcc, type = "HC1"))[, 4]

dtSat <- lm(tSat ~ tTreat + factor(ageGroup) + tTreat*factor(ageGroup), data = dc)
dtSat$se <- sqrt(diag(vcovHC(dtSat, type = "HC1")))
dtSat$p <- coeftest(dtSat, vcovHC(dtSat, type = "HC1"))[, 4]

dtUseful <- lm(tUseful ~ tTreat + factor(ageGroup) + tTreat*factor(ageGroup), data = dc)
dtUseful$se <- sqrt(diag(vcovHC(dtUseful, type = "HC1")))
dtUseful$p <- coeftest(dtUseful, vcovHC(dtUseful, type = "HC1"))[, 4]

dtClear <- lm(tClear ~ tTreat + factor(ageGroup) + tTreat*factor(ageGroup), data = dc)
dtClear$se <- sqrt(diag(vcovHC(dtClear, type = "HC1")))
dtClear$p <- coeftest(dtClear, vcovHC(dtClear, type = "HC1"))[, 4]

dtMeaningful <- lm(tMeaningful ~ tTreat + factor(ageGroup) + tTreat*factor(ageGroup), data = dc)
dtMeaningful$se <- sqrt(diag(vcovHC(dtMeaningful, type = "HC1")))
dtMeaningful$p <- coeftest(dtMeaningful, vcovHC(dtMeaningful, type = "HC1"))[, 4]

stargazer(dtFair, dtAcc, dtSat, dtUseful, dtClear, dtMeaningful,
  type = 'text',
  se = list(dtFair$se, dtAcc$se, dtSat$se, dtUseful$se, dtClear$se, dtMeaningful$se),
  p = list(dtFair$p, dtAcc$p, dtSat$p, dtUseful$p, dtClear$p, dtMeaningful$p),
  covariate.labels = c("Explanation"),
  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)
## -----
## Explanation                      -0.111    0.083    0.028    0.028    -0.181    -0.153
##                                (0.261)  (0.220)  (0.221)  (0.222)  (0.199)  (0.238)
##
## factor(ageGroup)3                -0.190    0.050   -0.250   -0.444** -0.444**   -0.325
##                                (0.213)  (0.196)  (0.168)  (0.192)  (0.178)  (0.209)
##
## factor(ageGroup)4                -0.161    0.013   -0.176   -0.509** -0.209    -0.354
##                                (0.241)  (0.215)  (0.197)  (0.211)  (0.193)  (0.229)
##
## factor(ageGroup)5                -0.189    0.011    0.071   -0.285   -0.303    -0.194
##                                (0.339)  (0.310)  (0.268)  (0.307)  (0.295)  (0.314)
##
## factor(ageGroup)6                -0.061    0.367   -0.322   -0.639** -0.639*   -0.544*
##                                (0.311)  (0.257)  (0.314)  (0.276)  (0.326)  (0.317)

```

```
##
## factor(ageGroup)7          0.239    0.267    -0.322    -0.789    -0.889*    -0.394
##                          (0.641)   (0.637)   (0.629)   (0.634)   (0.515)   (0.689)
##
## factor(ageGroup)8         -2.361*** -2.333*** -2.722*** -2.389*** -2.889*** -2.194***
##                          (0.187)   (0.173)   (0.141)   (0.160)   (0.142)   (0.181)
##
## tTreat:factor(ageGroup)3    0.362    0.020    0.396    0.553**   0.586**   0.673**
##                          (0.296)   (0.255)   (0.256)   (0.264)   (0.244)   (0.277)
##
## tTreat:factor(ageGroup)4    0.329    0.115    0.298    0.601**   0.410     0.713**
##                          (0.336)   (0.292)   (0.303)   (0.297)   (0.284)   (0.314)
##
## tTreat:factor(ageGroup)5    0.323    0.033    0.141    0.330     0.556     0.768*
##                          (0.427)   (0.385)   (0.384)   (0.405)   (0.376)   (0.410)
##
## tTreat:factor(ageGroup)6    0.916**   0.322     0.730    1.327***  1.194***  1.240***
##                          (0.425)   (0.360)   (0.456)   (0.380)   (0.406)   (0.444)
##
## tTreat:factor(ageGroup)7    0.011    -0.058    0.322    0.497     1.181*    0.353
##                          (0.756)   (0.720)   (0.780)   (0.794)   (0.642)   (0.811)
##
## tTreat:factor(ageGroup)8
##
##
## Constant                  3.361***  3.333***  3.722***  3.389***  3.889***  3.194***
##                          (0.187)   (0.173)   (0.141)   (0.160)   (0.142)   (0.181)
##
## -----
## Observations              622      622      622      622      622      622
## R2                        0.028     0.025     0.040     0.077     0.046     0.059
## Adjusted R2              0.009     0.006     0.021     0.059     0.027     0.041
## Residual Std. Error (df = 609) 1.202     1.088     1.139     1.169     1.154     1.192
## F Statistic (df = 12; 609)  1.455     1.322     2.091**   4.222***  2.461***  3.199***
## =====
## Note:                                     *p<0.1; **p<0.05; ***p<0.01
```

There seem to be some significance to the effects, although many of these age groups are not well represented as shown previously. It might be more appropriate to bin these groups in a different way to test whether or not the largest group (ages 25-34) is somehow different.

Education

Education was not collected as years of education, but as levels of education. It is appropriate to look at the factors for treatment effects.

```
dtFair <- lm(tFair ~ tTreat + factor(educ) + tTreat*factor(educ), data = dc)
dtFair$se <- sqrt(diag(vcovHC(dtFair, type = "HC1")))
dtFair$p <- coeftest(dtFair, vcovHC(dtFair, type = "HC1"))[, 4]

dtAcc <- lm(tAcc ~ tTreat + factor(educ) + tTreat*factor(educ), data = dc)
dtAcc$se <- sqrt(diag(vcovHC(dtAcc, type = "HC1")))
dtAcc$p <- coeftest(dtAcc, vcovHC(dtAcc, type = "HC1"))[, 4]
```

```

dtSat <- lm(tSat ~ tTreat + factor(educ) + tTreat*factor(educ), data = dc)
dtSat$se <- sqrt(diag(vcovHC(dtSat, type = "HC1")))
dtSat$p <- coeftest(dtSat, vcovHC(dtSat, type = "HC1"))[, 4]

dtUseful <- lm(tUseful ~ tTreat + factor(educ) + tTreat*factor(educ), data = dc)
dtUseful$se <- sqrt(diag(vcovHC(dtUseful, type = "HC1")))
dtUseful$p <- coeftest(dtUseful, vcovHC(dtUseful, type = "HC1"))[, 4]

dtClear <- lm(tClear ~ tTreat + factor(educ) + tTreat*factor(educ), data = dc)
dtClear$se <- sqrt(diag(vcovHC(dtClear, type = "HC1")))
dtClear$p <- coeftest(dtClear, vcovHC(dtClear, type = "HC1"))[, 4]

dtMeaningful <- lm(tMeaningful ~ tTreat + factor(educ) + tTreat*factor(educ), data = dc)
dtMeaningful$se <- sqrt(diag(vcovHC(dtMeaningful, type = "HC1")))
dtMeaningful$p <- coeftest(dtMeaningful, vcovHC(dtMeaningful, type = "HC1"))[, 4]

stargazer(dtFair, dtAcc, dtSat, dtUseful, dtClear, dtMeaningful,
  type = 'text',
  se = list(dtFair$se, dtAcc$se, dtSat$se, dtUseful$se, dtClear$se, dtMeaningful$se),
  p = list(dtFair$p, dtAcc$p, dtSat$p, dtUseful$p, dtClear$p, dtMeaningful$p),
  covariate.labels = c("Explanation"),
  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)
Explanation	-1.500**	-0.750	-2.250**	-0.500	-0.500	-1.250*
	(0.715)	(0.748)	(0.903)	(0.912)	(0.619)	(0.748)
factor(educ)2	-0.556	-0.278	-1.333***	-0.194	0.250	-1.028
	(0.426)	(0.411)	(0.402)	(0.752)	(0.417)	(0.752)
factor(educ)3	-0.327	-0.073	-0.820**	-0.013	0.047	-1.187
	(0.388)	(0.381)	(0.380)	(0.727)	(0.382)	(0.728)
factor(educ)4	0.210	0.403	-0.823**	0.290	-0.081	-0.839
	(0.416)	(0.393)	(0.411)	(0.757)	(0.449)	(0.753)
factor(educ)5	-0.230	-0.206	-0.960**	-0.040	-0.000	-1.167
	(0.374)	(0.372)	(0.374)	(0.724)	(0.375)	(0.724)
factor(educ)6	-0.643	-0.300	-1.129***	-0.143	0.014	-1.171
	(0.408)	(0.397)	(0.391)	(0.743)	(0.404)	(0.744)
factor(educ)7	-0.214	0.357	-0.786**	-0.000	0.214	-0.857
	(0.532)	(0.433)	(0.397)	(0.744)	(0.491)	(0.756)
tTreat:factor(educ)2	2.072***	1.205	3.212***	1.307	0.815	1.600**

```
##          (0.778) (0.797) (0.937) (0.958) (0.672) (0.809)
##
## tTreat:factor(educ)3      1.727**  0.796  2.225**  0.895  0.699  1.582**
##          (0.751) (0.774) (0.927) (0.935) (0.651) (0.778)
##
## tTreat:factor(educ)4      1.210  0.395  2.540***  0.500  0.919  1.185
##          (0.765) (0.787) (0.943) (0.964) (0.710) (0.805)
##
## tTreat:factor(educ)5      1.654**  0.924  2.652***  1.094  0.832  1.877**
##          (0.730) (0.760) (0.914) (0.924) (0.637) (0.762)
##
## tTreat:factor(educ)6      2.325***  1.345*  2.924***  1.461  1.168*  2.035**
##          (0.760) (0.782) (0.931) (0.944) (0.661) (0.795)
##
## tTreat:factor(educ)7      1.214  0.018  1.786*  1.000  0.536  1.607*
##          (0.960) (0.943) (1.014) (1.062) (0.825) (0.935)
##
## Constant      3.500***  3.500***  4.500***  3.000***  3.500***  4.000***
##          (0.358) (0.358) (0.358) (0.715) (0.358) (0.715)
##
## -----
## Observations      622      622      622      622      622      622
## R2      0.040  0.033  0.061  0.071  0.035  0.057
## Adjusted R2      0.019  0.013  0.041  0.051  0.015  0.037
## Residual Std. Error (df = 608)  1.196  1.085  1.127  1.174  1.162  1.194
## F Statistic (df = 13; 608)  1.944**  1.617*  3.019***  3.552***  1.710*  2.820***
## =====
## Note:                                     *p<0.1; **p<0.05; ***p<0.01
```

Satisfaction seems to be significant across all levels of education.

Social Media Usage

```
dtFair <- lm(tFair ~ tTreat + factor(socMed) + tTreat*factor(socMed), data = dc)
dtFair$se <- sqrt(diag(vcovHC(dtFair, type = "HC1")))
dtFair$p <- coeftest(dtFair, vcovHC(dtFair, type = "HC1"))[, 4]

dtAcc <- lm(tAcc ~ tTreat + factor(socMed) + tTreat*factor(socMed), data = dc)
dtAcc$se <- sqrt(diag(vcovHC(dtAcc, type = "HC1")))
dtAcc$p <- coeftest(dtAcc, vcovHC(dtAcc, type = "HC1"))[, 4]

dtSat <- lm(tSat ~ tTreat + factor(socMed) + tTreat*factor(socMed), data = dc)
dtSat$se <- sqrt(diag(vcovHC(dtSat, type = "HC1")))
dtSat$p <- coeftest(dtSat, vcovHC(dtSat, type = "HC1"))[, 4]

dtUseful <- lm(tUseful ~ tTreat + factor(socMed) + tTreat*factor(socMed), data = dc)
dtUseful$se <- sqrt(diag(vcovHC(dtUseful, type = "HC1")))
dtUseful$p <- coeftest(dtUseful, vcovHC(dtUseful, type = "HC1"))[, 4]

dtClear <- lm(tClear ~ tTreat + factor(socMed) + tTreat*factor(socMed), data = dc)
dtClear$se <- sqrt(diag(vcovHC(dtClear, type = "HC1")))
dtClear$p <- coeftest(dtClear, vcovHC(dtClear, type = "HC1"))[, 4]
```

```

dtMeaningful <- lm(tMeaningful ~ tTreat + factor(socMed) + tTreat*factor(socMed), data = dc)
dtMeaningful$se <- sqrt(diag(vcovHC(dtMeaningful, type = "HC1")))
dtMeaningful$p <- coeftest(dtMeaningful, vcovHC(dtMeaningful, type = "HC1"))[, 4]

stargazer(dtFair, dtAcc, dtSat, dtUseful, dtClear, dtMeaningful,
  type = 'text',
  se = list(dtFair$se, dtAcc$se, dtSat$se, dtUseful$se, dtClear$se, dtMeaningful$se),
  p = list(dtFair$p, dtAcc$p, dtSat$p, dtUseful$p, dtClear$p, dtMeaningful$p),
  covariate.labels = c("Explanation"),
  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)
## -----
## Explanation                      0.305*** 0.223** 0.419*** 0.596*** 0.375*** 0.500***
##                                (0.109) (0.097) (0.101) (0.105) (0.105) (0.107)
##
## factor(socMed)2                   0.193   0.216   0.008   -0.206  -0.139  -0.257
##                                (0.178) (0.162) (0.175) (0.198) (0.216) (0.198)
##
## factor(socMed)3                   0.253   -0.082  -0.207  -0.075  -0.318  -0.091
##                                (0.410) (0.325) (0.388) (0.202) (0.280) (0.341)
##
## factor(socMed)4                  -1.164   0.668   0.960*** 1.008   0.432   0.576
##                                (0.717) (0.716) (0.363) (0.717) (0.717) (1.071)
##
## factor(socMed)5                   0.836   1.001*** -0.207   0.175  -0.068   0.076
##                                (0.587) (0.315) (0.617) (0.558) (0.466) (0.676)
##
## tTreat:factor(socMed)2            -0.315  -0.336  -0.249  -0.012   0.044   0.072
##                                (0.264) (0.248) (0.257) (0.284) (0.271) (0.274)
##
## tTreat:factor(socMed)3            -0.611  -0.029  -0.530  -0.290  -0.514  -0.0005
##                                (0.509) (0.461) (0.525) (0.375) (0.441) (0.471)
##
## tTreat:factor(socMed)4
##
##
## tTreat:factor(socMed)5            -1.005  -1.256** -0.253  -0.762  -0.775  -0.900
##                                (0.731) (0.590) (0.784) (0.677) (0.577) (0.780)
##
## Constant                         3.164*** 3.332*** 3.540*** 2.992*** 3.568*** 2.924***
##                                (0.081) (0.073) (0.072) (0.080) (0.079) (0.079)
## -----
## Observations                      622      622      622      622      622      622
## R2                                0.020    0.017    0.036    0.064    0.038    0.050
## Adjusted R2                       0.008    0.004    0.024    0.051    0.026    0.038

```

```
## Residual Std. Error (df = 613) 1.203 1.089 1.137 1.174 1.155 1.194
## F Statistic (df = 8; 613) 1.590 1.325 2.896*** 5.211*** 3.038*** 4.037***
## =====
## Note: *p<0.1; **p<0.05; ***p<0.01
```

Again, there is some scattered significance throughout these models. The explanation has high significance for the control group, which is daily users. However, this is an incredibly large portion of the respondent population, so that is not surprising given our overall results. And it is difficult to trust the small sample sizes of the other groups. They also have larger errors, showing great variation, so we cannot draw conclusions from these groups.

Race and Gender in the Recidivism Context

Given the results shown above for the Twitter context, we are not going to examine all of the effects in the recidivism context. However, we will look at the role that race and gender play. Minorities have been shown to have been the victims of bias in the recidivism algorithms, so this is important to review.

Again, Pacific Islanders rated the explanation worse than others. Women were more likely to agree with the decision than men. There were also other significant effects throughout the table, but they do not fit a pattern across metrics.

```
drFair <-lm(rFair ~ rTreat + white + female + black + white*rTreat + female*rTreat + black*rTreat, data = data)
drFair$se <- sqrt(diag(vcovHC(drFair, type = "HC1")))
drFair$p <- coeftest(drFair, vcovHC(drFair, type = "HC1"))[, 4]

drAcc <-lm(rAcc ~ rTreat + white + female + black + white*rTreat + female*rTreat + black*rTreat, data = data)
drAcc$se <- sqrt(diag(vcovHC(drAcc, type = "HC1")))
drAcc$p <- coeftest(drAcc, vcovHC(drAcc, type = "HC1"))[, 4]

drSat <-lm(rSat ~ rTreat + white + female + black + white*rTreat + female*rTreat + black*rTreat, data = data)
drSat$se <- sqrt(diag(vcovHC(drSat, type = "HC1")))
drSat$p <- coeftest(drSat, vcovHC(drSat, type = "HC1"))[, 4]

drUseful <-lm(rUseful ~ rTreat + white + female + black + white*rTreat + female*rTreat + black*rTreat, data = data)
drUseful$se <- sqrt(diag(vcovHC(drUseful, type = "HC1")))
drUseful$p <- coeftest(drUseful, vcovHC(drUseful, type = "HC1"))[, 4]

drClear <-lm(rClear ~ rTreat + white + female + black + white*rTreat + female*rTreat + black*rTreat, data = data)
drClear$se <- sqrt(diag(vcovHC(drClear, type = "HC1")))
drClear$p <- coeftest(drClear, vcovHC(drClear, type = "HC1"))[, 4]

drMeaningful <- lm(rMeaningful ~ rTreat + white + female + black + white*rTreat + female*rTreat + black*rTreat, data = data)
drMeaningful$se <- sqrt(diag(vcovHC(drMeaningful, type = "HC1")))
drMeaningful$p <- coeftest(drMeaningful, vcovHC(drMeaningful, type = "HC1"))[, 4]

stargazer(drFair, drAcc, drSat, drUseful, drClear, drMeaningful,
  type = 'text',
  se = list(drFair$se, drAcc$se, drSat$se, drUseful$se, drClear$se, drMeaningful$se),
  p = list(drFair$p, drAcc$p, drSat$p, drUseful$p, drClear$p, drMeaningful$p),
  covariate.labels = c("Explanation", "White", "Female"),
  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)
## -----
## Explanation      0.842***  0.470***  0.610***  0.677***  0.752***  0.661***
##                  (0.186)  (0.178)  (0.192)  (0.200)  (0.208)  (0.195)
##
## White            -0.143    -0.220    -0.332*   -0.473***  -0.246    -0.370**
##                  (0.161)  (0.155)  (0.171)  (0.173)  (0.190)  (0.176)
##
## Female           0.201     0.135     0.257*    0.181     -0.068     0.020
##                  (0.136)  (0.127)  (0.152)  (0.152)  (0.168)  (0.151)
##
## black            -0.616**   -0.473*    -0.201    -0.594**   -0.028     -0.402
##                  (0.264)  (0.260)  (0.292)  (0.265)  (0.342)  (0.293)
##
## rTreat:white      -0.002     0.070     0.287     0.484**    0.287     0.200
##                  (0.213)  (0.204)  (0.223)  (0.226)  (0.237)  (0.223)
##
## rTreat:female     -0.291     -0.207    -0.375*   -0.444**   -0.066     -0.199
##                  (0.189)  (0.178)  (0.208)  (0.203)  (0.222)  (0.203)
##
## rTreat:black      -0.109     0.041     -0.180     0.091     -0.474     -0.065
##                  (0.367)  (0.362)  (0.434)  (0.372)  (0.476)  (0.382)
##
## Constant          2.476***  2.842***  2.887***  2.638***  2.916***  2.675***
##                  (0.138)  (0.136)  (0.149)  (0.151)  (0.170)  (0.154)
## -----
## Observations      620      620      620      620      620      620
## R2                 0.123     0.061     0.082     0.143     0.119     0.105
## Adjusted R2        0.113     0.050     0.071     0.134     0.109     0.095
## Residual Std. Error (df = 612)  1.093     1.027     1.235     1.172     1.284     1.174
## F Statistic (df = 7; 612)      12.246***  5.658***  7.770***  14.625***  11.808***  10.309***
## =====
## Note:                                                     *p<0.1; **p<0.05; ***p<0.01
```

African American respondents show a negative baseline difference with respect to fairness. Interestingly,

What Additional Information did respondents want

In each context, respondents were asked what information they would like to see as part of an explanation. This question was multiple choice with multiple selection allowed as well as a text write-in section. The three options were:

- Examples of other levels of decision output
- Relative importance of the characteristics that led to the decision
- Detailed description of how the algorithm works.

Find total number of those who selected each option.

```
tcsumReqInfo1<-sum(dt$tcReqInfo1, na.rm=TRUE)
tcsumReqInfo2<-sum(dt$tcReqInfo2, na.rm=TRUE)
tcsumReqInfo3<-sum(dt$tcReqInfo3, na.rm=TRUE)
```



```

ttsumReqInfo1<-sum(dt$ttReqInfo1, na.rm=TRUE)
ttsumReqInfo2<-sum(dt$ttReqInfo2, na.rm=TRUE)
ttsumReqInfo3<-sum(dt$ttReqInfo3, na.rm=TRUE)

rcsumReqInfo1<-sum(dt$rcReqInfo1, na.rm=TRUE)
rcsumReqInfo2<-sum(dt$rcReqInfo2, na.rm=TRUE)
rcsumReqInfo3<-sum(dt$rcReqInfo3, na.rm=TRUE)

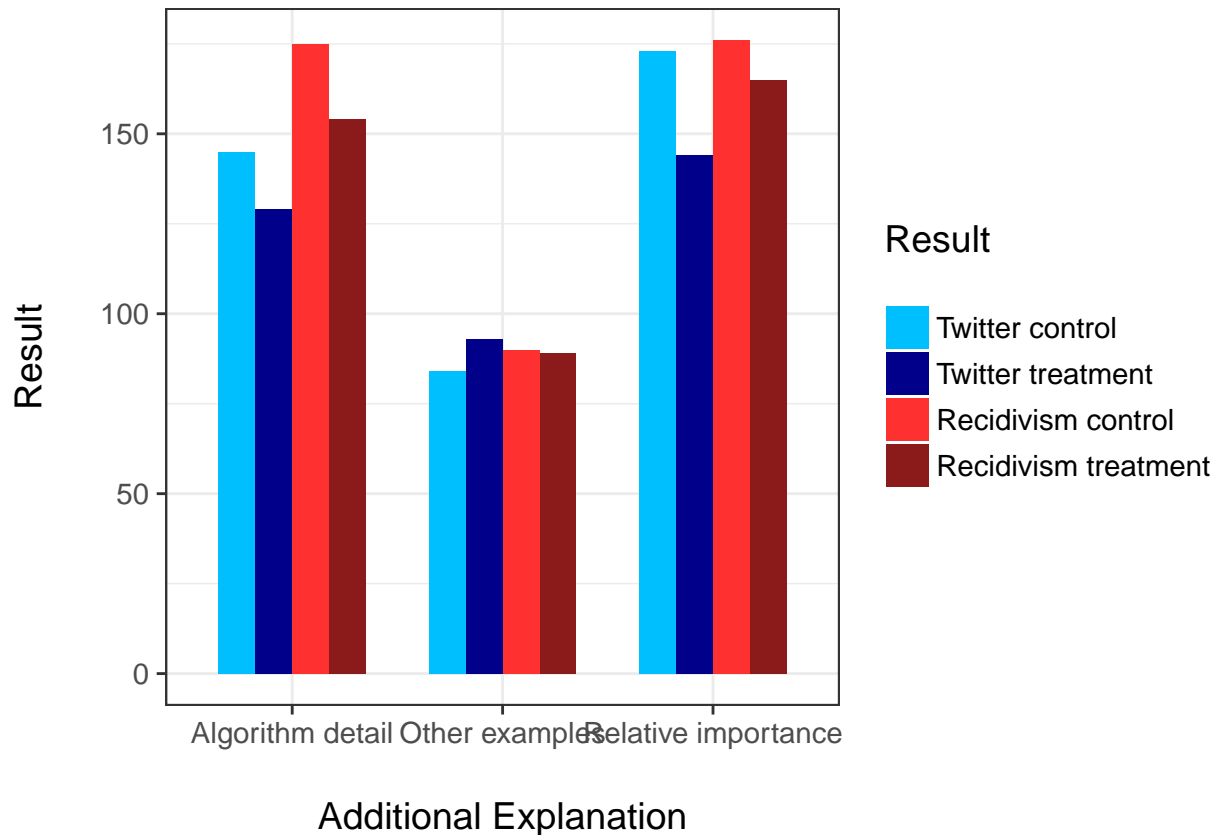
rtsumReqInfo1<-sum(dt$rtReqInfo1, na.rm=TRUE)
rtsumReqInfo2<-sum(dt$rtReqInfo2, na.rm=TRUE)
rtsumReqInfo3<-sum(dt$rtReqInfo3, na.rm=TRUE)

Number <- c("Other examples","Relative importance","Algorithm detail")
tc <- c(tcsumReqInfo1,tcsumReqInfo2,tcsumReqInfo3)
tt <- c(ttsumReqInfo1,ttsumReqInfo2,ttsumReqInfo3)
rc <- c(rcsumReqInfo1,rcsumReqInfo2,rcsumReqInfo3)
rt <- c(rtsumReqInfo1,rtsumReqInfo2,rtsumReqInfo3)
nyx <- data.frame(Number,tc,tt, rc, rt)

# reshape your data into long format
nyxlong <- melt(nyx, id=c("Number"))

# make the plot
ggplot(nyxlong) +
  geom_bar(aes(x = Number, y = value, fill = variable),
    stat="identity", position = "dodge", width = 0.7) +
  scale_fill_manual("Result\n", values = c("deepskyblue","blue4", "firebrick1","firebrick4"),
    labels = c("Twitter control", "Twitter treatment","Recidivism control",
      "Recidivism treatment")) +
  labs(x="\nAdditional Explanation",y="Result\n") +
  theme_bw(base_size = 14)

```



While not everyone answered this question, we see a consistent distribution of choices. The Relative Importance of factors was the most often selected in each combination of context and treatment. The Algorithm Details was a close second in many combinations of context and treatment, nearly the same as Relative Importance in the recidivism control. The control group almost always asked for more explanation than the treatment group, which can likely be attributed to the effectiveness of the explanation. A regression would show whether or not this is a significant effect. The only exception to this is that the Twitter treatment group selected Other Examples more than the control. This is a small difference in the least populated selection, and it is a smaller difference than the other options.

Not to include in report, but text outputs of Other q 4

Twitter control

again, whether there's any oversight into the decision or any appeals (though in this specific case it was clearly valid) Explanation of why they feel it's right to limit free speech. Freedom of speech infringement, you can not like the guy but twitter shouldn't silence his viewpoint, individuals should block him. When you are a platform for social interaction you shouldn't be allowed to restrict who can say what. None, I am against the restriction of the freedom of speech. none, moderation should be done by humans Statistics on pattern of behavior of banned individual, whether wrong people have ever been flagged What conduct rule in particular was deemed violated by the algorithm.

Twitter treatment

A contextual analysis of the actual subject of the comment an explanation of the 70% threshold Definition of the characteristics Explanation of Sentiment Analysis Explanation on what is sentiment analysis and how it was judged I am perfectly satisfied with the explanation exactly as it is. I'm satisfied with the information. It should be stacked. All of the bars should add together. The way it's set up now, it could be at 69% threshold for all criteria and still would not have any action taken against it. more detailed explanation of the graph No other information required. nothing, it is ridiculous Proper spelling and explanation of sentiment decisions What "sentiment analysis" means. Is this thing reacting to everything it views as expressing a negative or argumentative attitude? when does using offensive vocabulary mean you are guilty, then pretty much all of us would be in jail and not got out of college Why Twitter feels the need to implement an algorithm like this at all.

Recidivism control

A human isn't simple enough, there are going to be many factors that won't be considered. Again - ridiculous just as previous answer - doesn't take "person" into account, just numbers An explanation of why and how they use an algorithm in court—and who allowed it. basically, it needs tons more information to make a decision like that none past success percentage Statistics on accuracy the specific information the algorithm uses to make the assessment what happened to innocent until proven guilty? you can not make assessment of someone based on algorithm when they did not do anything wrong whether there's any human oversight into the decision

Recidivism treatment

A possibility of a human psychologist or social worker weighing in on the algorithm results too. An explanation of how level of criminal personality was determined. biases of those who wrote the algorithm Definitions of each How it can assume that everyone is the same based on answers. human evaluation I want to review the source code, the questions, the regression test results (when it was tested on repeat offenders) I would like to know what information the algorithm considered when making the decision. I'm already satisfied with the explanation. More detailed breakdown of what is meant in this context by phrases like "criminal personality" NONE Numbers records showing other uses that turned out to be accurate and the percentage of accuracy overall