Contributor roles

Overview

This script explores Q4: "Which of these open source contributor roles has ever applied to you?".

Import packages and utilities

```
project_root <- here::here() # requires that you be somewhere in the
# project directory (not above it)
# packages
suppressMessages(source(file.path(project_root, "scripts/packages.R")))
# functions and objects used across scripts
suppressMessages(source(file.path(project_root, "scripts/utils.R")))
```

Load data

```
roles_raw <- load_qualtrics_data("clean_data/contributor_roles_Q4.tsv")
other_quant <- load_qualtrics_data("clean_data/other_quant.tsv")
head(roles_raw)
```

```
  Maintainer Contributor Bug/Issue Reporter Community Manager Educator Other
1          1           1                  1                 1        1     0
2          0           1                  0                 0        0     0
3          1           1                  1                 0        1     0
4          1           1                  1                 0        1     0
5          1           1                  1                 0        1     0
```

```
6          0           0                   0                  0        0     0
  Supervisor IT/Systems administrator UI/UX Designer Technical support
1          1                         0              0                 1
2          0                         0              0                 0
3          0                         0              1                 0
4          1                         0              0                 0
5          1                         0              0                 0
6          0                         0              0                 0
```

```
nrow(roles_raw)
```

```
[1] 332
```

Wrangle data

Drop rows with all 0s (NAs were already converted to 0s during data cleanup).

```
roles <- filter_all(roles_raw, any_vars(. != 0))
nrow(roles)
```

```
[1] 233
```

Explore data

```
counts <- colSums(roles)
counts <- data.frame(Role = names(counts), Count = as.integer(counts), row.names = NULL)
counts <- counts %>%
    arrange(desc(Count))

counts
```

```
                    Role Count
1        Bug/Issue Reporter   189
2               Contributor   187
3                  Educator   138
4                Maintainer   134
5         Technical support   102
```

```
6                Supervisor    96
7         Community Manager    49
8            UI/UX Designer    46
9  IT/Systems administrator    44
10                    Other    13
```

That's mildly interesting. There's about 2.7-3 maintainers for every community manager, sys admin, or UI designer. I'm curious how many people are EXCLUSIVELY one thing. Mainly interested in how many people are EXCLUSIVELY a maintainer, contributor, bug reporter, or supervisor.

```
indices <- which(rowSums(roles) == 1 )
exclusive <- roles[indices,]
nrow(exclusive)
```

```
[1] 16
```

Let's repeat what we did for the whole data frame, but now with just people who filled one role.

```
counts_exc <- colSums(exclusive)
counts_exc <- data.frame(Role = names(counts_exc), Count = as.integer(counts_exc), row.names
counts_exc <- counts_exc %>%
    arrange(desc(Count))

counts_exc
```

```
                      Role Count
1        Bug/Issue Reporter     5
2               Contributor     4
3                Supervisor     3
4                  Educator     2
5                Maintainer     1
6         Community Manager     1
7                     Other     0
8  IT/Systems administrator     0
9            UI/UX Designer     0
10        Technical support     0
```

So, out of 233 contributors, only 16 identified with exactly one contributor role. 5 of those were Bug/Issue Reporter, the most common exclusive-role.

```r
# Calculate total roles for each participant
roles2 <- cbind(roles, total_roles = rowSums(roles))

# Get mean number of roles per column (role)
get_mean_num_roles <- function(df, colnum) {
    # drop rows where the entry in this col is 0
    filtered <- df[df[[colnum]] != 0, ]
    return(
        mean(filtered$total_roles)
    )
}
num_roles <- data.frame(
  role = names(roles),
  avg_num_roles_per_participant = sapply(
    seq(ncol(roles)),
    function(x) get_mean_num_roles(roles2, x)
  )
)

num_roles
```

```
                    role avg_num_roles_per_participant
1             Maintainer                       5.194030
2            Contributor                       4.737968
3      Bug/Issue Reporter                       4.682540
4       Community Manager                       6.163265
5                Educator                       5.195652
6                   Other                       6.384615
7             Supervisor                       5.583333
8  IT/Systems administrator                       6.386364
9          UI/UX Designer                       6.369565
10      Technical support                       5.715686
```

Huh. So I guess no one role is particularly enriched for people who hold many roles.

I'd be kind of interested in the 3-way venn diagram of contributors, but reporters and maintainers. (If we do plot this, an UpSet plot would probably be better than a Venn diagram)

```r
# maintainers, contributors, and bug reporters in that order

three_way_counts <- with(roles, {
  M <- Maintainer == 1
```

```
  C <- Contributor == 1
  B <- `Bug/Issue Reporter` == 1

  c(
    "M, !C, !B"   = sum(M & !C & !B),
    "M, C, !B"  = sum(M & C & !B),
    "M, !C, B"  = sum(M & !C & B),
    "M, C, B" = sum(M & C & B),
    "!M, !C, B"   = sum(!M & !C & B),
    "!M, C, B"  = sum(!M & C & B),
    "!M, C, !B"   = sum(!M & C & !B),
    "!M, !C, !B" = sum(!M & !C & !B)
  )
})

three_way_counts
```

```
 M, !C, !B   M, C, !B   M, !C, B    M, C, B  !M, !C, B   !M, C, B  !M, C, !B
         5         14          4        111         27         47         15
!M, !C, !B
        10
```

```
# This should be 233, the number of experienced contributors
sum(three_way_counts)
```

```
[1] 233
```

```
# Number of people who contributed as at least one of these three roles
at_least_one <- sum(three_way_counts)-three_way_counts["!M, !C, !B"]
at_least_one
```

```
!M, !C, !B
       223
```

```
# Percent of total who identified as at least one of the three
round(three_way_counts/at_least_one*100, digits=1)
```

```
 M, !C, !B   M, C, !B   M, !C, B    M, C, B  !M, !C, B   !M, C, B  !M, C, !B
       2.2         6.3         1.8        49.8        12.1        21.1         6.7
!M, !C, !B
       4.5
```

Interesting. So 96% of the survey respondents identified as at least one of Maintainer, Contributor, or Bug Reporter (223/233). 50.0% of those folks identified as all three (111/223). 21.1% identified as a contributor and bug reporter, but not a maintainer, and 12.1% identified as a bug reporter only. Together, these three groups make up 50+21.1+12.2 83.3% of the total.

Out of curiosity, how many maintainers also selected tech support?

```
two_way_counts <- with(roles, {
  M <- Maintainer == 1
  TS <- `Technical support` == 1

  c(
    "M, !T"   = sum(M & !TS),
    "M, T"  = sum(M & TS)
  )
})

two_way_counts
```

```
M, !T  M, T
   66    68
```

Eh, not that interesting. I thought more maintainers would also be tech supporters.

What's the ratio of maintainers to other contributors?

```
nrow(subset(roles, Maintainer == 1))
```

```
[1] 134
```

```
nrow(subset(roles, Maintainer == 0))
```

```
[1] 99
```

That's somewhat interesting, and I guess it's consistent with the "three-way venn-diagram" above. 134 people identified as maintainers, and only 99 people identified as contributors of some sort, but not maintainers. So 134/233 = 57.5% of experienced contributors in this survey are maintainers. I'm getting the sense, basically, that there are a lot of maintainers in this survey.

Bring in job category

Are Math/CS people more likely to be maintainers than biologists?

```
roles_field <- cbind(roles_raw, other_quant$field_of_study)
# Rename column
names(roles_field)[ncol(roles_field)] <- "field_of_study"
# Filter out non-academics (people who didn't answer the field of study question)
roles_field <- roles_field %>%
    filter(field_of_study != "")

nrow(roles_field)
```

```
[1] 188
```

```
# Remove people who did not select any role
roles_field_clean <- roles_field[rowSums(roles_field[, -ncol(roles_field)]) != 0,]

nrow(roles_field_clean)
```

```
[1] 147
```

```
# Sanity check: make sure none of the rows sums to 0
unname(rowSums(roles_field_clean %>% select(-field_of_study)))
```

```
  [1]  7  1  5  5  5  4  5  3  6  4  3  2  3  3  5  2  3  4  2  4  8  3  5  3  3
 [26]  4  1  4  2  3  3  3  8  4  2  3  6  1  8 10  4  5  2  3  5  4  3  3  3  4
 [51]  4  7  2  6  6  5  4  2  7  7  3  4  3  5  2  2  2  5  9  5  6  2  7  6  1
 [76]  2  8  2  5  4  7  4  3  4  3  3  4  2  5  1  2  3  3  4  4  4 10  4  6  4
[101]  4  6  5  4  3  1  1  2  6  7  1  1  7  8  5  4  2  4  6  2  6  5  8  4  8
[126]  7  4  7  7  3  1  2  6  4  4  1  4  5  3  5  6  2  8  4  3  6  2
```

Okay, so we have a total of 188 academics. It looks like 147 of those are experienced contributors–people who selected at least one role.

```
maintainer_props <- roles_field_clean %>%
  group_by(field_of_study) %>%
  summarise(
    n_people          = n(),
```

```
    n_maintainers     = sum(Maintainer == 1),
    prop_maintainers  = mean(Maintainer == 1)
  ) %>%
  arrange(desc(prop_maintainers))

maintainer_props
```

```
# A tibble: 5 x 4
  field_of_study    n_people n_maintainers prop_maintainers
  <chr>                <int>         <int>            <dbl>
1 Math and CS             72            53            0.736
2 Physical sciences       27            18            0.667
3 Life sciences           34            18            0.529
4 Social sciences         10             4            0.4
5 Humanities               4             1            0.25
```

Meh, mildly interesting. 50% or more of our survey respondents from STEM fields are maintainers. There's really not sufficient respondents from the non-STEM fields to draw any conclusions there.

Campus

What proportion of respondents are maintainers, per campus?

```
roles_campus <- cbind(roles_raw, other_quant$campus)
# Rename column
names(roles_campus)[ncol(roles_campus)] <- "campus"
# Filter out non-UC
roles_campus <- roles_campus %>%
    filter(campus != "I'm not affiliated with UC")
# Remove people who did not select any role
roles_campus <- roles_campus[rowSums(roles_campus[, -ncol(roles_campus)]) != 0,]

nrow(roles_campus)
```

```
[1] 233
```

```r
maint_by_campus <- roles_campus %>%
  group_by(campus) %>%
  summarise(
    n_rows         = n(),
    n_maintainer   = sum(Maintainer == 1),
    prop_maintainer = mean(Maintainer == 1)
  ) %>%
  arrange(desc(prop_maintainer))


maint_by_campus <- maint_by_campus[order(maint_by_campus$prop_maintainer, decreasing = TRUE)
maint_by_campus
```

```
# A tibble: 10 x 4
   campus           n_rows n_maintainer prop_maintainer
   <chr>             <int>        <int>           <dbl>
 1 Other UC             19           15           0.789
 2 UC Santa Cruz        32           22           0.688
 3 UC Berkeley          26           16           0.615
 4 UC San Francisco      7            4           0.571
 5 UC Santa Barbara     61           34           0.557
 6 UC Los Angeles       40           21           0.525
 7 UC Merced             8            4           0.5
 8 UC Davis             29           14           0.483
 9 UC San Diego          9            4           0.444
10 UC Irvine             2            0           0
```

Meh, again, only mildly interesting. I suppose it's an interesting add-on to the stat that about half of our respondents are maintainers. The campuses with fewer than ten respondents just feel like noise to me, since we can't really draw conclusions. Let's filter those out.

```r
subset(maint_by_campus, n_rows > 10)
```

```
# A tibble: 6 x 4
  campus           n_rows n_maintainer prop_maintainer
  <chr>             <int>        <int>           <dbl>
1 Other UC             19           15           0.789
2 UC Santa Cruz        32           22           0.688
3 UC Berkeley          26           16           0.615
4 UC Santa Barbara     61           34           0.557
5 UC Los Angeles       40           21           0.525
6 UC Davis             29           14           0.483
```

That's a little clearer.