

# Motivations for contributing to OS: plots

## Overview

This script makes some plots from Q6, which is about participants' reasons for contributing to open source.

## 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")))
```

## Define functions

### get\_df\_for\_job\_category

- Arguments:
  - job: A string.
  - raw\_data: The Qualtrics data, unmodified. Don't mess with this. I only made this an argument just to clarify that there's another data structure going into this function besides job.
- Details:

- A function that takes a particular job category of interest and produces a data frame with counts for all motivation categories. By ‘count’, I mean the number of people who selected that motivation.
- Outputs:
  - A data frame with two columns: **Motivation** and **Count**. Count is the number of ‘yes’ responses for that motivation for the given job category.

```
get_df_for_job_category <- function(job, raw_data = data) {
  df <- raw_data %>%
    filter(job_category == job) %>%
    select(
      starts_with("motivations")
    )
  df <- shorten_long_responses(df, codenames)
  # Remove any columns that are all NA or empty strings
  df <- df[, colSums(is.na(df) | df == "") < nrow(df)]
  df <- rename_cols_based_on_entries(df)
  # Remove any rows where they didn't answer the question about motivations
  df <- make_df_binary(df)
  df <- data.frame(
    Motivation = names(df),
    Count = unname(apply(df, 2, function(x) round(sum(x, na.rm = TRUE))))
  )
  return(df)
}
```

```
line_plot <- function(df, x_var, y_var, title) {
  p <- ggplot(df, aes(x = .data[[x_var]], y = .data[[y_var]])) +
    geom_point(size = 4) + # Adjust dot size
    scale_y_continuous(labels=scales::percent) +
    labs(
      y = "Proportion of Participants Motivated by\nDesire to Learn New Skills",
      title = title
    ) +
    theme(
      axis.title.x = element_blank(),
      axis.title.y = element_blank(),
      axis.text.x = element_text(angle = 60, vjust = 0.6, size = 24),
      axis.text.y = element_text(size = 24),
      axis.ticks.x = element_blank(),
      axis.ticks.y = element_blank(),
    )
}
```

```

    legend.title = element_blank(),
    plot.title = element_text(hjust = 0.5, size = 24),
    plot.margin = unit(c(0.3, 0.3, 0.3, 0.3), "cm"),
    panel.grid = element_line(linetype = "solid", color = "gray90"),
    panel.background = element_blank()
  )
  return(p)
}

```

## Load data

```
data <- load_qualtrics_data("deidentified_no_qual.tsv")
```

## Define codes

```

codenames <- c(
  "Developing open-source" = "Job",
  "To improve the tools" = "Improve Tools",
  "To customize existing" = "Customize",
  "To build a network" = "Network",
  "To give back to" = "Give back",
  "To improve my skills" = "Skills",
  "Because it's fun" = "Fun",
  "Other " = "Other"
)

```

## Basic bar plot of contributor motivations

First we wrangle the data using several functions from my utilities script `scripts/utis.R`.

```

motivations <- data %>% select(
  starts_with("motivations")
)
motivations <- shorten_long_responses(motivations, codenames)

```

```
# Remove any columns that are all NA or empty strings
# (Which means nobody selected that response)
motivations <- exclude_empty_columns(motivations)
# Remove any rows that are all NA or empty strings
# (The participant did not answer the question
# (because they're not UC or not a contributor))
motivations <- exclude_empty_rows(motivations)

motivations <- rename_cols_based_on_entries(motivations)
motivations <- make_df_binary(motivations)
head(motivations)
```

	Job	Improve Tools	Customize Network	Give back Skills	Fun	Other
1	1	1	1	1	1	0
2	0	1	1	1	0	0
3	0	1	1	0	0	1
4	1	1	1	0	1	0
5	0	1	1	0	1	1
7	0	1	1	0	0	0

Now we sum up counts for each motivation.

```
motivations_to_plot <- data.frame(
  Motivation = names(motivations),
  Count = unname(apply(motivations, 2, function(x) round(sum(x, na.rm = TRUE))))
)
head(motivations_to_plot)
```

	Motivation	Count
1	Job	109
2	Improve Tools	198
3	Customize	161
4	Network	78
5	Give back	156
6	Skills	142

Reorder factor levels based on count.

```
motivations_to_plot <- motivations_to_plot %>%
  mutate(Motivation = fct_reorder(Motivation, Count, .desc = FALSE))
```

And make a plot, using a function in `utils.R`.

```
myplot <- basic_bar_chart(motivations_to_plot,
  x_var = "Motivation",
  y_var = "Count",
  title = "Reasons for Contributing to Open Source",
  horizontal = TRUE,
  show_bar_labels = TRUE,
  show_ticks_y = FALSE,
  color_index = 3,
  show_axis_title_y = FALSE,
  show_grid = TRUE
)
```

Save the plot if you wish.

```
#save_plot("motivations_overall.tiff", 10, 6, p=myplot)
```

## Stacked bar plots of motivations by role

Now let's make some stacked bar plots of motivations by role (job category). We'll make two: one with the absolute number of responses, and one where all roles are normalized to 1, so we can see the relative proportions of each motivation.

```
faculty <- get_df_for_job_category("Faculty")
nrstaff <- get_df_for_job_category("Non-research Staff")
postdocs <- get_df_for_job_category("Post-Doc")
other_researchers <- get_df_for_job_category(
  "Other research staff (e.g., research scientist, research software engineer)"
)
grads <- get_df_for_job_category("Grad Student")
undergrads <- get_df_for_job_category("Undergraduate")
# Example
faculty
```

	Motivation	Count
1	Job	23
2	Improve Tools	54
3	Customize	45
4	Network	16

5	Give back	38
6	Skills	23
7	Fun	32
8	Other	13

For visual clarity, let's combine post-docs and other research staff into one category.

```
postdocs_other <- bind_rows(postdocs, other_researchers) %>%
  group_by(Motivation) %>%
  summarise(Count = sum(Count, na.rm = TRUE), .groups = "drop")
```

Add a Role column and combine these little data frames into one long-format data frame.

```
faculty$Role <- "Faculty"
nrstaff$Role <- "Non-research Staff"
grads$Role <- "Grad Students"
postdocs_other$Role <- "Postdocs and\nStaff Researchers"
undergrads$Role <- "Undergraduates"
composite_df <- rbind(faculty, nrstaff, grads, postdocs_other, undergrads)
head(composite_df)
```

	Motivation	Count	Role
1	Job	23	Faculty
2	Improve Tools	54	Faculty
3	Customize	45	Faculty
4	Network	16	Faculty
5	Give back	38	Faculty
6	Skills	23	Faculty

Create one plot with the absolute number of responses on the y-axis, and another plot where all jobs are scaled to 1.

```
stacked_plot_raw <- stacked_bar_chart(composite_df,
  x_var = "Role",
  y_var = "Count",
  fill = "Motivation",
  title = "Reasons for Contributing to Open Source",
)

stacked_plot_proportional <- stacked_bar_chart(composite_df,
  x_var = "Role",
```

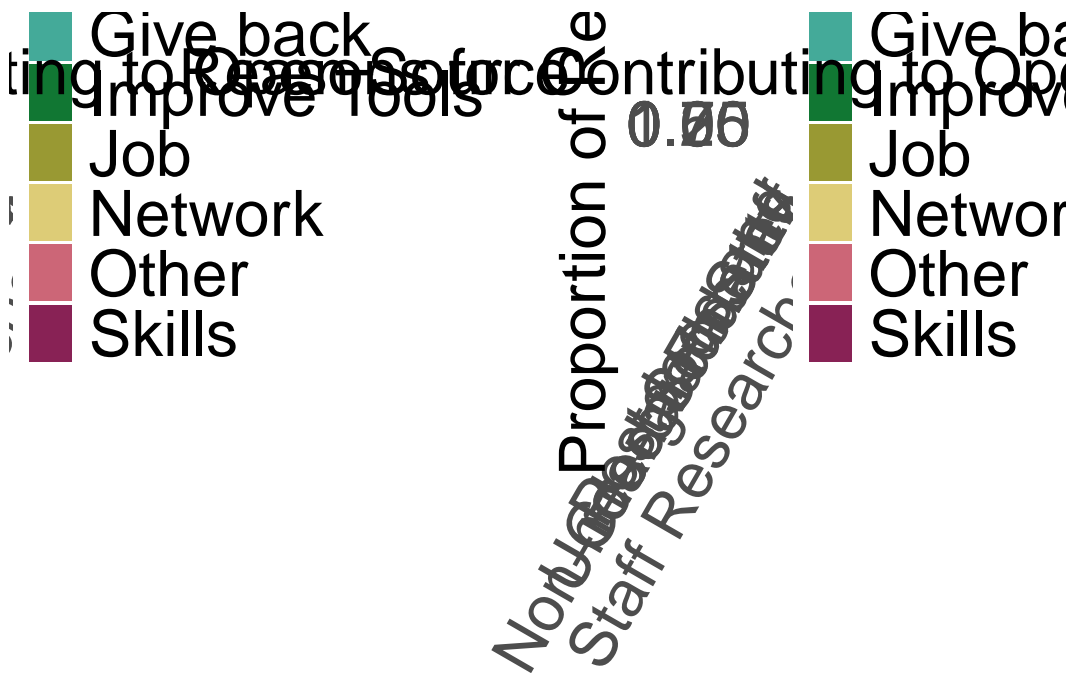
```

y_var = "Count",
fill = "Motivation",
title = "Reasons for Contributing to Open Source",
proportional = TRUE
)

```

Visualize

```
stacked_plot_raw + stacked_plot_proportional
```



Save

```
save_plot("motivations_stacks.tiff", 16, 8)
```

Request from Greg: What about IT vs. academics? (Students, Teachers, and Researchers)

```

it <- data %>%
  filter(staff_categories == "Information Technology (IT)") %>%
  select(
    starts_with("motivations")
  )

```

```

it <- shorten_long_responses(it, codenames)

# Remove any columns that are all NA or empty strings
# (Which means nobody selected that response)
it <- exclude_empty_columns(it)
# Remove any rows that are all NA or empty strings
# (The participant did not answer the question
# (because they're not UC or not a contributor))
it <- exclude_empty_rows(it)

it <- rename_cols_based_on_entries(it)
it <- make_df_binary(it)
head(it)

```

	Job	Improve Tools	Customize Network	Give back Skills	Fun	Other
1	1	0	0	1	1	0
2	1	1	1	1	1	1
3	0	1	1	1	1	1
4	0	1	0	0	0	0
5	0	1	0	0	1	1
6	0	1	1	0	1	1

```
dim(it)
```

```
[1] 33 8
```

```

it <- data.frame(
  Motivation = names(it),
  Count = unname(apply(it, 2, function(x) round(sum(x, na.rm = TRUE))))
)
it$Role <- "IT"
it

```

	Motivation	Count	Role
1	Job	7	IT
2	Improve Tools	26	IT
3	Customize	23	IT
4	Network	10	IT
5	Give back	27	IT
6	Skills	21	IT
7	Fun	20	IT
8	Other	1	IT



```

academics <- composite_df %>%
  filter(
    Role == "Faculty" |
    Role == "Grad Students" |
    Role == "Postdocs and Staff Researchers" |
    Role == "Undergraduates"
  )
academics$Role <- "Academic"

it_academics <- rbind(it, academics)

```

Plot

```

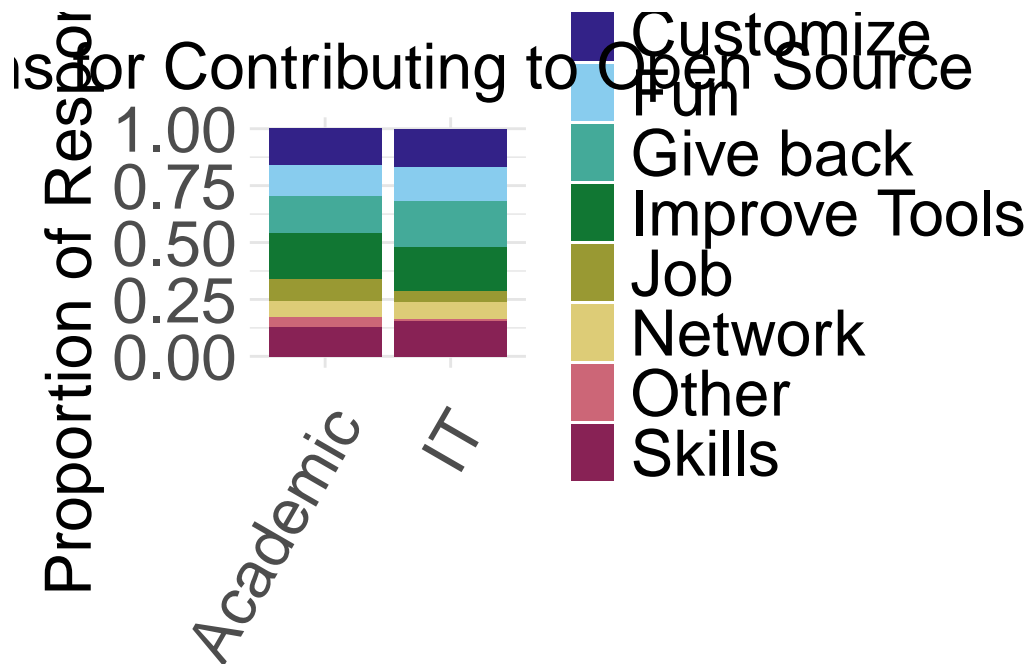
stacked_plot_raw_it <- stacked_bar_chart(
  it_academics,
  x_var = "Role",
  y_var = "Count",
  fill = "Motivation",
  title = "Reasons for Contributing to Open Source",
)

stacked_plot_proportional_it <- stacked_bar_chart(
  it_academics,
  x_var = "Role",
  y_var = "Count",
  fill = "Motivation",
  title = "Reasons for Contributing to Open Source",
  proportional = TRUE
)

```

Visualize

```
stacked_plot_proportional_it
```



Save

```
#save_plot("motivations_stacks_it_academics.tiff", 8, 8)
```

## Line plots for particular motivations

All 7 undergraduates selected “Skills” and “Give back” as motivations. This made me curious about whether these motivations decrease as we get older and advance in our careers. Let’s make some line plots to investigate.

```
motivations_raw <- data %>% select(
  starts_with("motivations")
)
motivations_raw <- shorten_long_responses(motivations_raw, codenames)
motivations_raw <- rename_cols_based_on_entries(motivations_raw)
motivations_raw$Role <- data$job_category
motivations_raw <- shorten_long_responses(motivations_raw, c("Other research staff" = "Other

# Remove any rows where they didn't answer the question about motivations
```

```

motivations_raw <- motivations_raw %>%
  filter(if_any(Job:Other, ~ .x != ""))

motivation_cols <- as.vector(codenames)
motivations_raw <- make_df_binary(motivations_raw, cols = motivation_cols)

skills_by_role <- motivations_raw %>%
  group_by(Role) %>%
  summarise(
    n_yes = sum(Skills == 1), # number of 1s
    n_tot = n(), # total rows
    Proportion = n_yes / n_tot
  )

skills_by_role_clean <- skills_by_role %>%
  # drop the staff categories
  filter(!Role %in% c("Non-research Staff", "Other research staff")) %>%
  # drop the unnecessary columns
  select(Role, Proportion) %>%
  # order the factor levels
  mutate(Role = factor(Role,
    levels = c(
      "Undergraduate",
      "Grad Student",
      "Post-Doc",
      "Faculty"
    ),
    ordered = TRUE
  )) %>%
  arrange(Role)

```

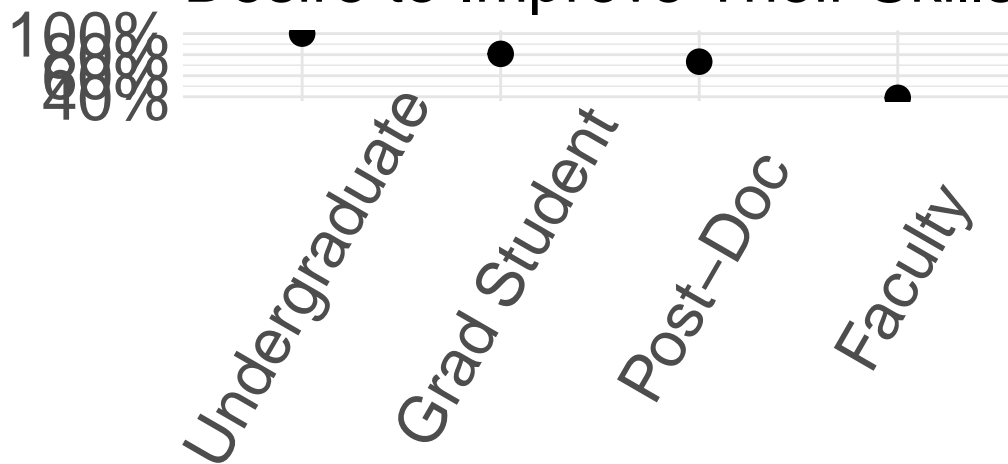
Plot and visualize

```

skills_plot <- line_plot(skills_by_role_clean,
  x_var = "Role",
  y_var = "Proportion",
  title = "Percent of Participants Motivated by\nDesire to Improve Their Skills"
)
skills_plot

```

## Percent of Participants Motivated Desire to Improve Their Skills



Save

```
save_plot("motivations_skill_by_role.tiff", 10, 8, p=skills_plot)
```

What about giving back?

```
give_by_role <- motivations_raw %>%
  group_by(Role) %>%
  summarise(
    n_yes = sum(`Give back` == 1), # number of 1s
    n_tot = n(), # total rows
    Proportion = n_yes / n_tot
  )

give_by_role_clean <- give_by_role %>%
  # drop the staff categories
  filter(!Role %in% c("Non-research Staff", "Other research staff")) %>%
  # drop the unnecessary columns
  select(Role, Proportion) %>%
  # order the factor levels
  mutate(Role = factor(Role,
    levels = c(
      "Undergraduate",
```

```

    "Grad Student",
    "Post-Doc",
    "Faculty"
),
ordered = TRUE
)) %>%
arrange(Role)

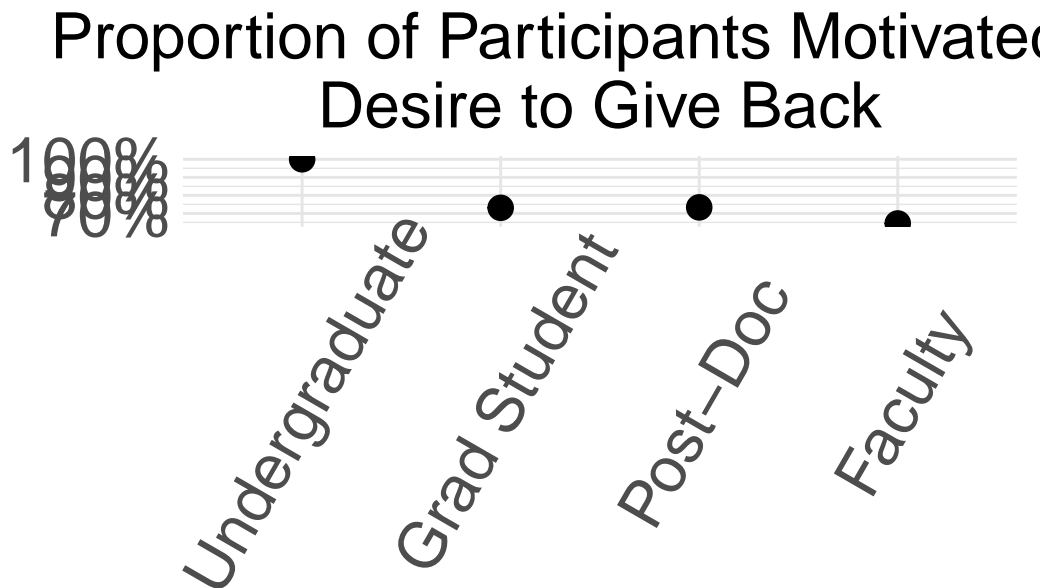
```

Plot and visualize

```

line_plot(give_by_role_clean,
  x_var = "Role",
  y_var = "Proportion",
  title = "Proportion of Participants Motivated by\nDesire to Give Back"
)

```



Save

```

#save_plot("motivations_giveback_by_role.tiff", 8, 6)

```

## Session Info

```
sessionInfo()
```

```
R version 4.4.2 (2024-10-31)
Platform: aarch64-apple-darwin20
Running under: macOS Sequoia 15.4.1
```

```
Matrix products: default
```

```
BLAS: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
```

```
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;
```

```
locale:
```

```
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
```

```
time zone: America/Los_Angeles
```

```
tzcode source: internal
```

```
attached base packages:
```

```
[1] tools      stats      graphics  grDevices datasets  utils      methods
```

```
[8] base
```

```
other attached packages:
```

```
[1] treemap_2.4-4      tidyr_1.3.1        stringr_1.5.1
[4] scales_1.4.0       readr_2.1.5        pwr_1.3-0
[7] patchwork_1.3.0    mvabund_4.2.1      languageserver_0.3.16
[10] here_1.0.1         gtools_3.9.5       fpc_2.2-13
[13] forcats_1.0.0      factoextra_1.0.7   ggplot2_3.5.2
[16] dplyr_1.1.4        corrplot_0.95      cluster_2.1.8.1
```

```
loaded via a namespace (and not attached):
```

```
[1] gtable_0.3.6      xfun_0.52          ggrepel_0.9.6
[4] processx_3.8.6    lattice_0.22-6     callr_3.7.6
[7] tzdb_0.5.0        vctrs_0.6.5        ps_1.9.1
[10] generics_0.1.4    stats4_4.4.2       parallel_4.4.2
[13] flexmix_2.3-20    tibble_3.2.1       DEoptimR_1.1-3-1
[16] pkgconfig_2.0.3   data.table_1.17.6  RColorBrewer_1.1-3
[19] lifecycle_1.0.4   compiler_4.4.2     farver_2.1.2
[22] statmod_1.5.0     httpuv_1.6.16      htmltools_0.5.8.1
[25] class_7.3-22      yaml_2.3.10        later_1.4.2
[28] pillar_1.10.2     prabclus_2.3-4     MASS_7.3-61
```

[31]	diptest_0.77-1	mclust_6.1.1	mime_0.13
[34]	robustbase_0.99-4-1	tidyselect_1.2.1	digest_0.6.37
[37]	stringi_1.8.7	purrr_1.0.4	kernlab_0.9-33
[40]	labeling_0.4.3	rprojroot_2.0.4	fastmap_1.2.0
[43]	grid_4.4.2	colorspace_2.1-1	cli_3.6.5
[46]	magrittr_2.0.3	withr_3.0.2	promises_1.3.3
[49]	tweedie_2.3.5	rmarkdown_2.29	igraph_2.1.4
[52]	nnet_7.3-19	modeltools_0.2-24	hms_1.1.3
[55]	shiny_1.11.0	evaluate_1.0.3	knitr_1.50
[58]	rlang_1.1.6	Rcpp_1.0.14	xtable_1.8-4
[61]	gridBase_0.4-7	glue_1.8.0	xml2_1.3.8
[64]	renv_1.1.4	jsonlite_2.0.0	R6_2.6.1