SOCI 1230

Data Science Across The Disciplines

Winter 2022

Week Two Assignment

In this assignment, you will continue exploring how variables are associated. The assignment will build on the tools we have learned in morning and afternoon sessions to visualize and manipulate data.

You are encouraged to use any of our course materials. You are free to collaborate with other students in our section but each student should submit their own report.

This assignment is due via Canvas by 10:00 AM on Monday, January 24, 2022.

Please submit your .Rmd notebook and a knitted PDF of your notebook that includes your output.

The tfs variables to use are available **here**.

The css variables to use are available **here**.

The matched TFS-CSS variables to use in the bonus question are available **here**.

Part One

The `REASON10` question in the TFS captures responses to the question of whether the respondent attended college "to learn more about things that interest me." Find the correlations between each response to this question and `k_rank` for each type of college. Summarize your findings in a table (formatted with kable) and interpret your findings in a few sentences.

```
tfs_cor_df <- tfs_means |>
  select(campus_id,
         n_responses,
         type,
         k_rank,
         starts_with("REASON10")) |>
  mutate(n_responses_nona = n_responses * (1-REASON10.propna)) |>
  pivot_longer(names_to = "response_level",
               values_to = "prop",
               (starts_with("REASON10") & !contains("propna"))) |>
  mutate(response_level = str_remove(response_level, "_mean")) |>
```

Table 1: Still Need A Title

	Importance		
Institution Type	Not important	Somewhat important	Very important
Private non-profit	-0.346	-0.292	0.309
Public	-0.666	-0.591	0.627

```
  mutate(response_level = str_remove(response_level, "REASON10.")) |>
  mutate(response_level = factor(response_level,
                                 levels = c("notimportant",
                                            "somewhatimportant",
                                            "veryimportant"),
                                 labels = c("Not important",
                                            "Somewhat important",
                                            "Very important")))
```

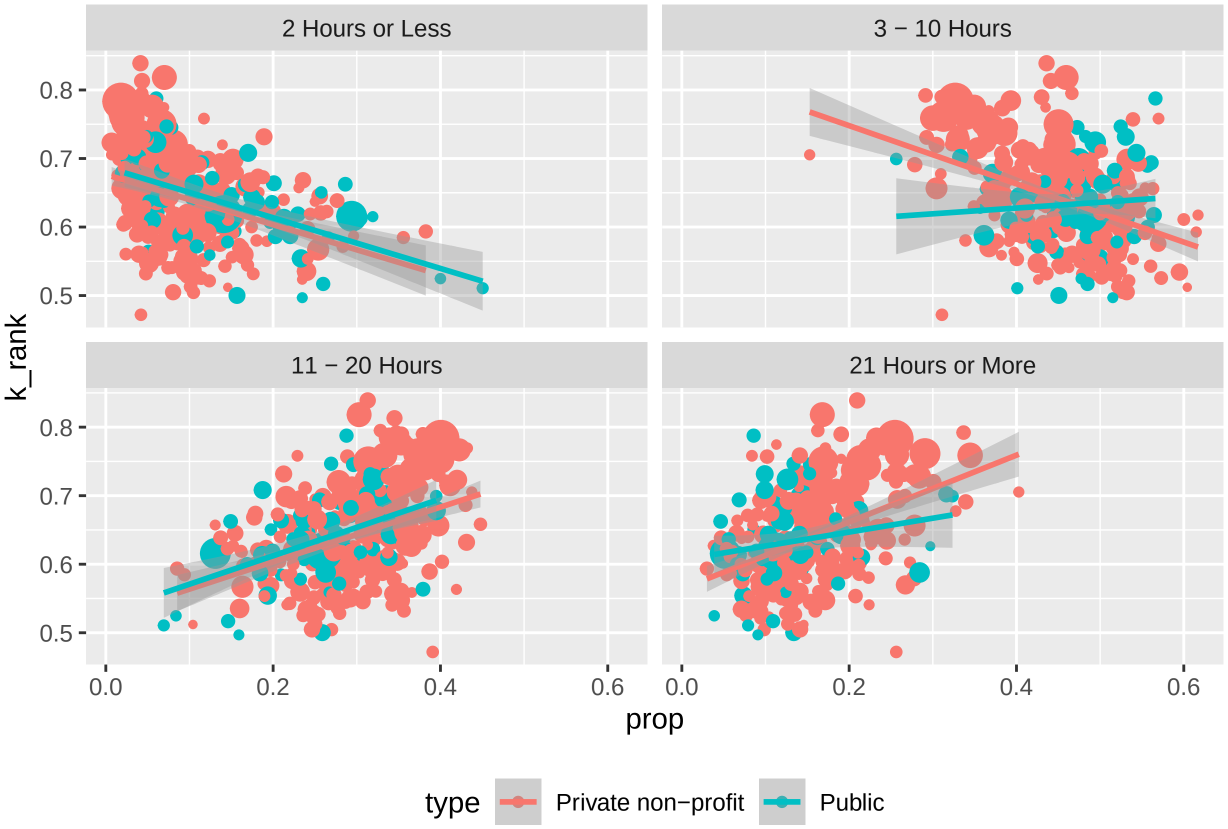
```
tfs_cor_summary <- tfs_cor_df |>
  group_by(type, response_level) |>
  summarise(wtd_cor_krank = wtd.cor(prop,
                            k_rank,
                            w = n_responses_nona)[1]) |>
  mutate(across(where(is.numeric),round,3))

kable_tfs_cor_summary <- tfs_cor_summary |>
  pivot_wider(names_from = "response_level",
              values_from = "wtd_cor_krank") |>
  rename("Institution Type" = type) |>
  kbl(booktabs = TRUE,
       align = "lccccc",
      caption = "Still Need A Title") |>
    add_header_above(c(" " = 1, "Importance" = 3))
```

```
kable_tfs_cor_summary
```

The HPW15 question in the CSS captures responses to the question of how often the respondent socialized with friends in person. Combine the responses into these categories: 2 hours or less, 3 hours - 10 hours, 11 hours - 20 hours, 21 hours or more. Find the correlations between each of the re-categorized responses to this question and k_rank for each type of college. Make one figure that includes the five scatterplots.

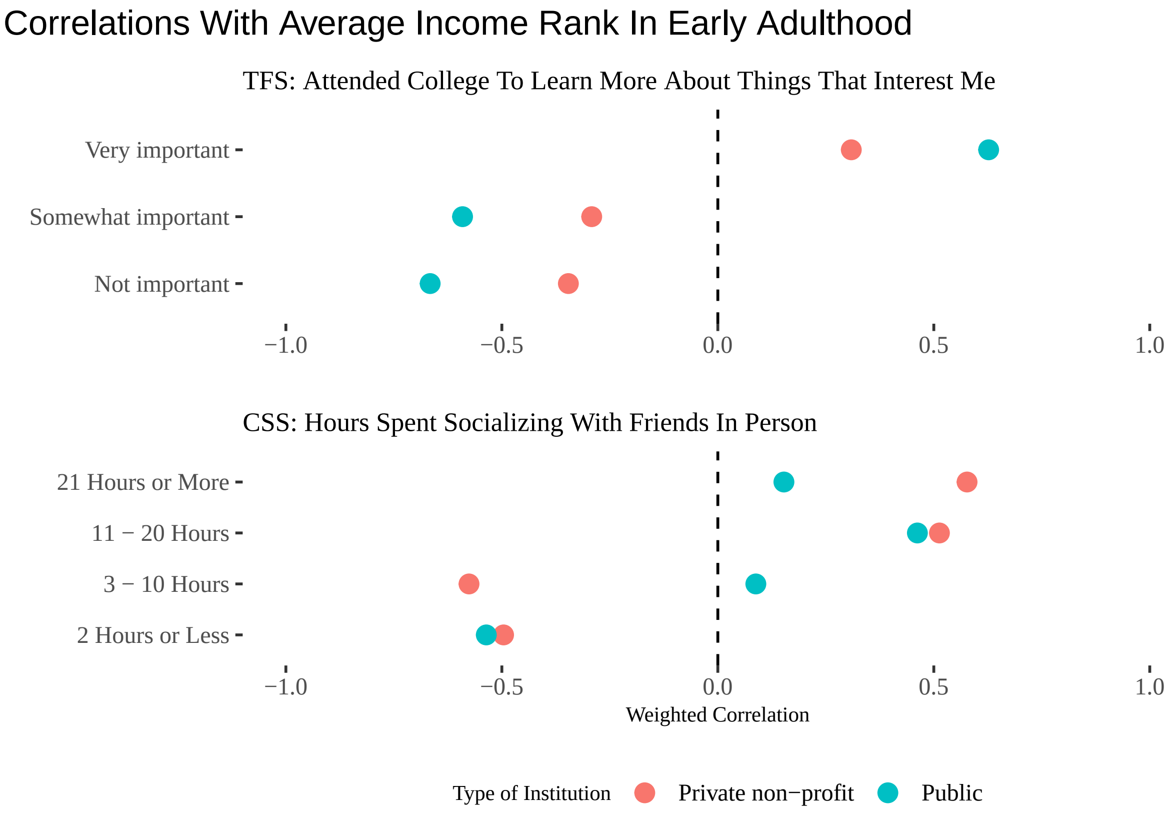
```
css_cor_df <- css_means |>
  select(campus_id,
         n_responses,
         type,
         k_rank,
         starts_with("HPW15")) |>
  mutate(n_responses_nona = n_responses * (1-HPW15.propna)) |>
  mutate(HPW_0_2 = HPW15.hours0_mean + HPW15.hours0to1_mean + HPW15.hours1to2_mean,
         HPW_3_10 = HPW15.hours3to5_mean + HPW15.hours6to10_mean,
         HPW_11_20 = HPW15.hours11to15_mean + HPW15.hours16to20_mean,
         HPW_21_plus = HPW15.hours21plus_mean) |>
  select(-starts_with("HPW15")) |>
  pivot_longer(names_to = "response_level",
               values_to = "prop",
               (starts_with("HPW_"))) |>
  mutate(response_level = factor(response_level,
                                 levels = c("HPW_0_2",
                                            "HPW_3_10",
                                            "HPW_11_20",
                                            "HPW_21_plus"),
                                 labels = c("2 Hours or Less",
                                            "3 - 10 Hours",
                                            "11 - 20 Hours",
                                            "21 Hours or More")))
css_cor_summary <- css_cor_df |>
  group_by(type, response_level) |>
  summarise(wtd_cor_krank = wtd.cor(prop,
                              k_rank,
                              w = n_responses_nona)[1]) |>
  mutate(across(where(is.numeric),round,3))
```

```
css_cor_df |>
  group_by(response_level) |> # we haven't used this with ggplot before
  ggplot(aes(x = prop, y = k_rank,
             size = n_responses_nona,
             color = type)) +
  geom_point() + geom_smooth(method = "lm") +
  theme(legend.position = "bottom") + guides(size = "none") +
  facet_wrap(~response_level) # need this if we have a group_by
```

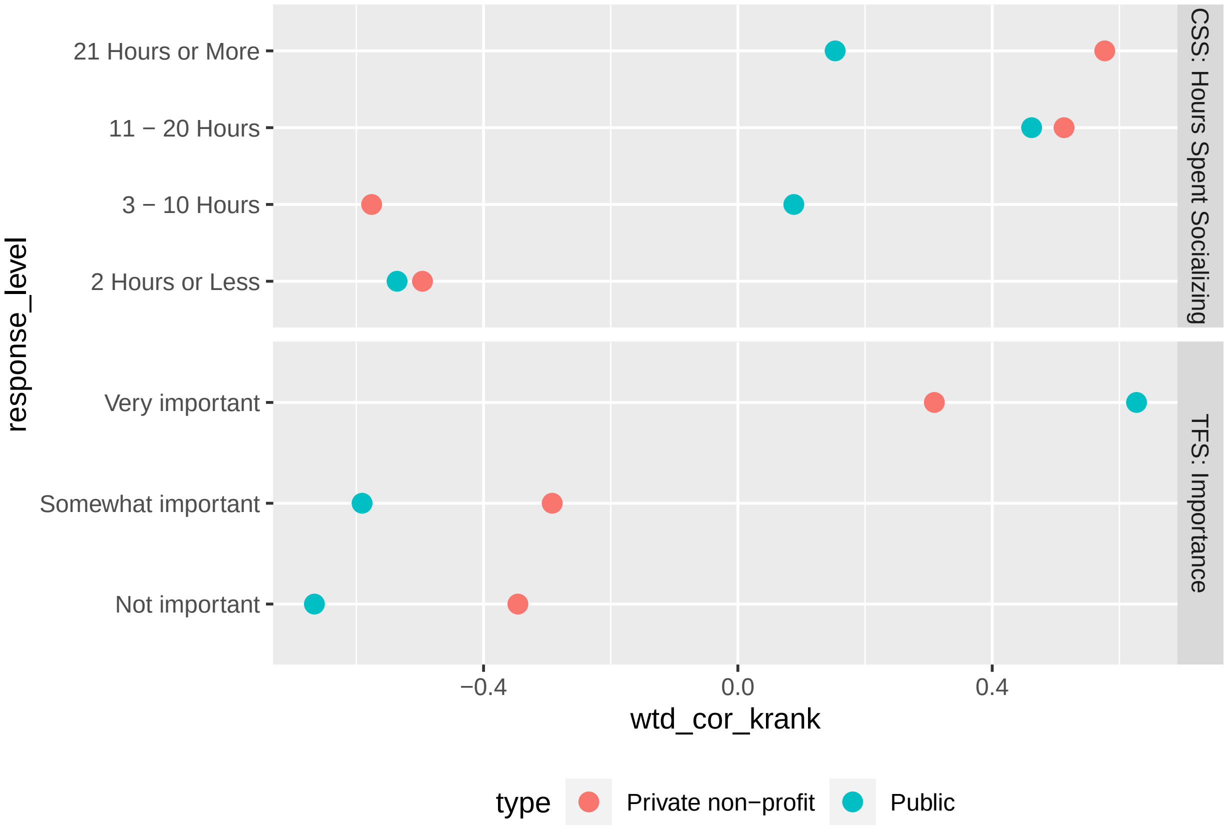


Part Two

Take all the correlations with `k_rank` that you found in Part One. Make one figure that combines plots of the coefficients (by type of college) in a plot similar to the Opportunity Insights summary plot. My example is on the top of the next page:



```r
tfs_css_correlations <- bind_rows(tfs_cor_summary,
                                  css_cor_summary) |>
  mutate(question = ifelse(str_detect(response_level, "Hours"),
                           "CSS: Hours Spent Socializing",
                           "TFS: Importance")) |>
  ggplot(aes(x = response_level,
             y = wtd_cor_krank,
             color = type)) +
  geom_point(size = 3) + coord_flip() +
  facet_grid(question~., scales = "free_y") +
  theme(legend.position = "bottom")

tfs_css_correlations
```



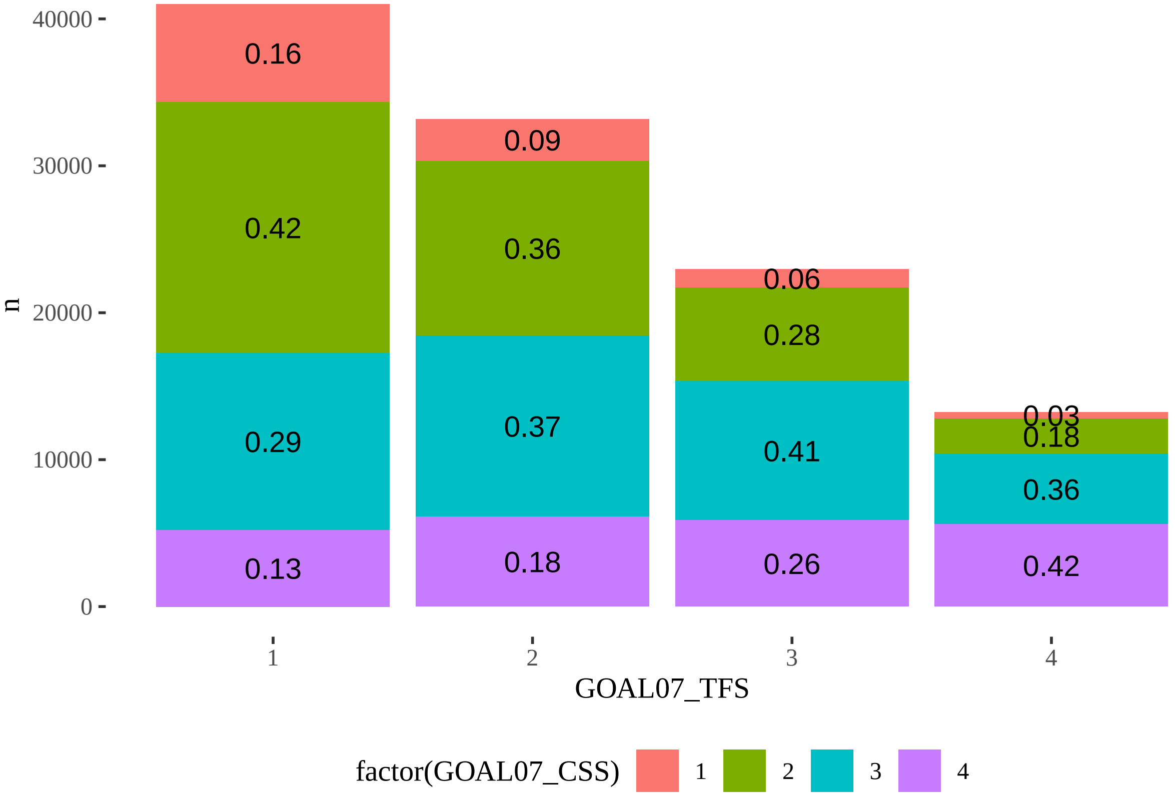
Bonus (if you make good progress on the above parts during Thursday's class and want more practice)

The matched CSS-TFS file links surveys across years by student. Take individuals' responses to the GOAL07 questions in both surveys and make a figure showing how the distribution of responses changed over time. The values are: 1 = "Not important", 2 = "Somewhat important", 3 = "Very important", 4 = "Essential". Write a few sentences explaining why you made your visualization decisions.

```
# Maybe something like this?

matched_df <- read_csv("https://raw.githubusercontent.com/mjclawrence/soci1230_w22/main/

prop.table(table(matched_df$GOAL07_TFS,
                 matched_df$GOAL07_CSS, exclude = NULL),1)
```

```
##
##                1          2          3          4       <NA>
##   1   0.16053327 0.41035311 0.28899832 0.12531828 0.01479702
##   2   0.08463021 0.35251223 0.36519935 0.18144360 0.01621461
##   3   0.05412569 0.27165455 0.40517315 0.25083369 0.01821291
##   4    0.03342927 0.17703491 0.35436499 0.41273707 0.02243377
##   <NA> 0.07894193 0.29468899 0.34841909 0.23186609 0.04608390
```

```
matched_df |>
  filter(!is.na(GOAL07_TFS), !is.na(GOAL07_CSS)) |>
  group_by(GOAL07_TFS) |>
  count(GOAL07_CSS) |>
  na.omit() |>
  mutate(prop = n/sum(n)) |>
  ungroup() |>
  ggplot(aes(x = GOAL07_TFS, y = n, fill = factor(GOAL07_CSS))) +
  geom_col() +
  theme_tufte() +
  theme(legend.position = "bottom") +
  geom_text(aes(label = round(prop,2)),
            position = position_stack(vjust = .5))
```


n
GOAL07_TFS
factor(GOAL07_CSS)