Data Assignment 1

Read-in data and prepare for analysis

```
library(ggplot2)
library(readr)
library(ggdag)
```

```

Attaching package: 'ggdag'

The following object is masked from 'package:stats':

    filter
```

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ stringr   1.5.1
✔ forcats   1.0.0     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.4     
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks ggdag::filter(), stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to b
```

```
library(gt)
library(modelsummary)

# read-in data
#3dat = read_csv(here::here("workshops/aau_survey/clean_endline_did.csv" )) %>%
dat = read_csv("https://raw.githubusercontent.com/jrspringman/psci3200-globaldev/main/wo
    # clean home region variable
  mutate(q8_baseline = ifelse(q8_baseline == "Southern Nations, Nationalities, and Peopl
         q8_baseline = str_remove(q8_baseline, " Region"))
```

```
Rows: 825 Columns: 280
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (106): response_id, user_language, q2, q11_1, q11_2, q11_3, q11_4, q12, ...
dbl (158): treatment_status, q13_1_1, q13_2_1, q13_3_1, q13_4_1, q13_5_1, q1...
lgl  (16): q27_1, q27_3, border_ethnic, border_resource, list_treat, q82_spl...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# create color palette for plotting
palette = MetBrewer::met.brewer(name = "Cross")

d <- dat %>%  # select neccessary columns as outlined
    select(response_id, treatment_status, user_language, q3_baseline, q26_civ, q26_civ_b
  rename(
   gender = q3_baseline,
    # Career Plans
  Bcivsociety_plan = q26_civ_baseline,# Baseline measurements recieve a B, so I can see
  Ecivsociety_plan = q26_civ, # Endline measurements recieve an E
  Bpol_plan = q26_politics_baseline,
  Epol_plan = q26_politics,
  Bpub_plan = q26_public_baseline,
  Epub_plan = q26_public,
  Bpuboffice_plan = q27_1_baseline,
  Epuboffice_plan = q27_1,
  Bnongov_plan = q27_3_baseline,
  Enongov_plan = q27_3,

  # Political Efficacy
  Byour_change = q17_3_baseline,
  Eyour_change = q17_3,
  Byouth_engage = q17_1_baseline,
  Eyouth_engage = q17_1,
  Byouth_change = q17_2_baseline,
  Eyouth_change = q17_2
)
```

Part 2: Create Index Measures

Requirement 2 (10%)

An additive index is created by summing the number of positive or consistent responses across several related survey items. It's a straightforward way to combine variables, but it only works well when all items are measured on the same scale and in the same direction. If items vary in scale or meaning, the additive approach becomes misleading. It's best used when the goal is to reflect the total number of favorable responses or behaviors.

An averaged z-score index, by contrast, standardizes each variable before combining them. This means transforming each item into a z-score by subtracting its mean and dividing by its standard deviation, so that all variables are on the same scale — with a mean of zero and a standard deviation of one. Then, these z-scores are averaged to create a composite measure. This method is particularly useful when combining variables that are measured differently or have different ranges.

The benefit of an additive index is its simplicity and intuitive interpretation. The advantage of an averaged z-score index is that it allows for fair combination of variables even when they're originally on different scales. However, z-score indices assume a normal distribution and require interval-level data.

Requirement 3 (20%)

1. Create an additive index for the baseline and endline measures of the "Future plans for a career in public sector or civil society" variables. This should correspond to seperate counts of the number of future plans that each individual has at baseline and endline.

```
d <- d %>%
  drop_na() %>% # drop any NAs that might screw up the index measurements
  mutate(across(c(Ecivsociety_plan, Bcivsociety_plan, Epub_plan, Bpub_plan),
              ~ ifelse(. == TRUE, 1, 0)), # mutate important rows
         Bindex_civsoc_pub = Bcivsociety_plan + Bpub_plan, # pre index
         Eindex_civsoc_pub = Ecivsociety_plan + Epub_plan) # post index
```

2. Create an averaged z-scores for the baseline and endline values of the "Future plans for a career in public sector or civil society" and "Feelings of political efficacy" variables.

```
all_vars <- list(
  base_career = c("Bcivsociety_plan", "Bpub_plan"),
  end_career  = c("Ecivsociety_plan", "Epub_plan"),

  base_eff = c("Byour_change", "Byouth_engage", "Byouth_change"),
  end_eff  = c("Eyour_change", "Eyouth_engage", "Eyouth_change")
)

z_score = function(x, y){ # taken from fp_essential.qmd
  # calculate column mean and sd
  c_mean = mean( as.numeric( unlist(x[, y])) , na.rm = T)
  c_sd = sd( as.numeric( unlist(x[, y])) , na.rm = T)
  # subtract column mean and divide by column SD
  ( as.numeric(x[, y, drop = TRUE]) - c_mean) / c_sd
}


for (group_name in names(all_vars)) {
  group_vars <- all_vars[[group_name]]

  #create columns for each z score using z_score function
    for (var in group_vars) {
    d[[paste0(var, "_z")]] <- z_score(d, var)
  }

   #create an averaged index for the group
  z_vars <- paste0(group_vars, "_z")  # List of z-score column names
  d[[paste0("z_index_", group_name)]] <- rowMeans(d[, z_vars], na.rm = TRUE)
}
```

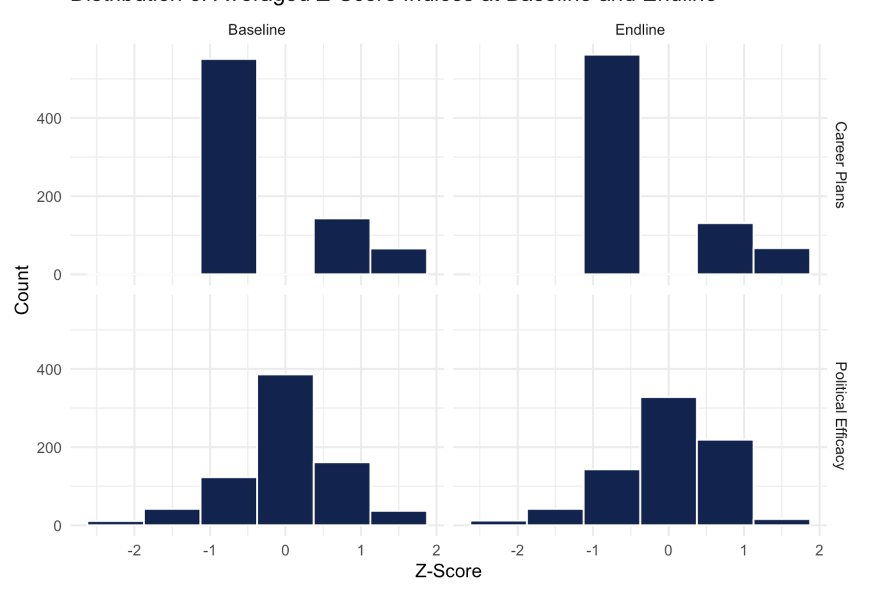
Requirement 4 (20%)

To make sure that these scores look as you'd expect, create a ggplot visualizing the distribution of the z-scores at baseline and endline. You should have 4 figures: one corresponding to each z-score at baseline and endline. In words, describe whether the figures tell us anything about changes over time.

```
z_plot_data <- d %>% # reshape data for easier facet
  select(z_index_base_career, z_index_end_career,
         z_index_base_eff, z_index_end_eff) %>% # select cols
  pivot_longer(
    everything(),
    names_to = "IndexType",
    values_to = "ZScore"
  ) %>%
  mutate(
    Time = case_when(
      grepl("_base_", IndexType) ~ "Baseline",
      grepl("_end_", IndexType) ~ "Endline"
    ),
    Dimension = case_when(
      grepl("career", IndexType) ~ "Career Plans",
      grepl("eff", IndexType) ~ "Political Efficacy"
    )
  )
```

```
ggplot(z_plot_data, aes(x = ZScore)) +
  geom_histogram(bins = 12, fill =  palette[9], color = "white", binwidth = 0.75) +
  facet_grid(Dimension ~ Time) +
  theme_minimal() +
  labs(
    title = "Distribution of Averaged Z-Score Indices at Baseline and Endline",
    x = "Z-Score",
    y = "Count"
  )
```

Distribution of Averaged Z-Score Indices at Baseline and Endline



The figure displays the distributions of two z-score indices — career plans and political efficacy — measured at baseline and endline. The purpose is to assess whether there are meaningful changes in youth attitudes over time.

Political Efficacy: The bottom row shows that the distributions of political efficacy z-scores are approximately normal at both time points. There is a slight rightward shift from baseline to endline, suggesting a modest increase in perceived political efficacy. This could imply that youth participants felt slightly more empowered or confident in their ability to participate in politics by the end of the time period.

Career Plans: In contrast, the top row shows distributions for career plans that are highly clustered and discrete, with most values stacked into a few narrow bars. This is a direct result of using binary TRUE/FALSE variables (e.g., whether a respondent plans to work in civil society or not). Due to this limited variation, it's hard to interpret the results or whether or not there was any change.

```
d_sub <- d %>%
  select(response_id, gender,
         z_index_base_career, z_index_end_career,
         z_index_base_eff, z_index_end_eff) # subsetting

model1 <- lm(z_index_base_career ~ z_index_base_eff, data = d_sub)
#summary(model1) regular visualization

modelsummary(model1, stars = TRUE,  title = "Regression Model: Career Plans against Feel
```

	(1)
(Intercept)	0.000
	(0.024)
z_index_base_eff	0.096**
	(0.034)
Num.Obs.	760
R2	0.010
R2 Adj.	0.009
AIC	1505.1
BIC	1519.0
Log.Lik.	-749.572
RMSE	0.65
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

Regression Model: Career Plans against Feelings of Political Efficacy

The regression results reveal a statistically significant, positive association between political efficacy and future career plans in civil society or public sector roles. Here, the alpha of 0.024 represents the expected value of z_index_base_career when z_index_base_career. For every one-unit increase in political efficacy or the Beta (standardized), the career plan index increases by approximately 0.096 standard deviations. This suggests that youth who feel more politically empowered are slightly more inclined to express interest in civic-oriented careers. Although the effect is small, it aligns with the broader idea that a sense of political agency can help shape long-term aspirations. However, interpreting this relationship causally requires strong assumptions—namely, that political efficacy is not influenced by unobserved confounders and that no omitted variable drives both outcomes. In this case, such assumptions may not hold. Participants were invited to a one-day workshop explicitly designed to connect them with civil society leaders and provide opportunities for political engagement. Those who chose to attend may have already been more interested in politics and public service than the average person. Effecrtively, individuals who are already drawn to a topic are more likely to want to pursue it as a career, independent of the program's influence. In the absence of proof of these ideal conditions, the findings reflect a meaningful association but not a definitive causal pathway.

```
d <- d %>%
  mutate(
    eff_base_binary = ifelse(z_index_base_eff >= mean(z_index_base_eff), 1, 0), # ifelse
    eff_end_binary = ifelse(z_index_end_eff >= mean(z_index_end_eff), 1, 0),
    gender = factor(gender)  # ensure gender is treated as a factor
  )


model2 <- lm(z_index_base_career ~ eff_base_binary + gender + eff_base_binary:gender, d)

modelsummary(model2, stars = TRUE,  title = "Regression Model: Career Plans against Feel
)
```

	(1)
(Intercept)	-0.080
	(0.064)
eff_base_binary	0.126
	(0.092)
genderMale	0.013
	(0.077)
eff_base_binary × genderMale	-0.001
	(0.107)
Num.Obs.	760
R2	0.009
R2 Adj.	0.005
AIC	1509.7
BIC	1532.9
Log.Lik.	-749.854
RMSE	0.65
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

Regression Model: Career Plans against Feelings of Political Efficacy with Gender Indicator

The regression results examine whether the relationship between political efficacy and future civic career plans differs by gender. Here, alpha represents the expected value of future plans index for the reference group (low political efficacy females). Beta1 (eff_base_binary) at 0.126 represents the difference in career interests between high and low efficacy females. Beta2(genderMale) is the difference between low efficacy males and low efficacy females. Beta3(eff_base_binary x genderMale) is an interaction term, describing how the efficacy effect differs by gender.

The coefficient for the binary efficacy indicator suggests that, among female respondents, those with high political efficacy score approximately 0.126 standard deviations higher on the future career plan index than their low-efficacy peers. However, this difference is not statistically significant. Similarly, male respondents do not significantly differ from female respondents at low levels of efficacy, as indicated by the small and non-significant gender coefficient. Most notably, the interaction term between efficacy and gender is nearly zero and far from significant, indicating that the relationship between efficacy and career interest does not meaningfully vary by gender. Substantively, this implies that while higher political efficacy may be modestly associated with greater civic career aspirations, this effect is relatively uniform across male and female youth, at least in the baseline data.

-Convert the data from 'wide' to 'long' format, so that each respondent (response_id) has two rows of data; one row is baseline and one row is endline.

```
d_long<- d_sub %>%
  pivot_longer(
    cols = c(z_index_base_career, z_index_end_career,
             z_index_base_eff, z_index_end_eff),
    names_to = c("timepoint", "variable"),
    names_pattern = "z_index_(base|end)_(career|eff)",
    values_to = "value"
  ) %>%
  pivot_wider(
    names_from = variable,
    values_from = value
  ) %>%
  mutate(timepoint = recode(timepoint, base = "baseline", end = "endline"))


fe_model <- lm(career ~ eff + factor(response_id), data = d_long)# using base r regressi



modelsummary(fe_model,
             title = "Fixed Effects Regression Model: Future Plans Against Political Effi
             stars = TRUE,
             coef_omit = "response_id")
```

	(1)
(Intercept)	-0.371
	(0.339)
eff	0.020
	(0.031)
Num.Obs.	1520
R2	0.733
R2 Adj.	0.466
AIC	2535.5
BIC	6594.2
Log.Lik.	-505.738
RMSE	0.34
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

Fixed Effects Regression Model: Future Plans Against Political Efficacy (with Response ID Fixed Effects)

After adding fixed effects, Beta1 captures the within person effect of political efficacy on future plans. This controls for all unobserved, time invariant characteristics of each individual (like personality or background), isolating how changes in efficacy over time are associated with a person's respective career interests. Every one unit increase in the z score index is associated with a 0.02 standard deviation increase in their future civic career interests. However, since the relationship is not statistically significant, this relationship could be due to change.