Data Cleaning

Khanh Do, Joyce Gill

2026-02-03

Download data

Instruction:

1. Go to IPEDS Data Center, and click on **Complete Data Files**
2. Download Data File HD2024, EF2024D, etc.
3. Open the zipped file, extract and put the raw csv into this repo's data/raw/ folder

```
# Directory
hd2024 <- read.csv("data/raw/hd2024.csv")

# Fall Enrollment
ef2024d <- read.csv("data/raw/ef2024d.csv")
ef2024b <- read.csv("data/raw/ef2024b.csv")

# Cost
cost1_2024 <- read.csv("data/raw/cost1_2024.csv")
cost2_2024 <- read.csv("data/raw/cost2_2024.csv")
```

Process tables with multiple id

```
ef2024b_inst <- ef2024b %>%
  group_by(UNITID) %>%
  summarize(
    total_enrollment = sum(EFAGE09, na.rm = TRUE),
    total_men = sum(EFAGE07, na.rm = TRUE),
    total_women = sum(EFAGE08, na.rm = TRUE)
  )
```

Get removed UNITIDs

```
remove_unitids <- bind_rows(
  # 2-year schools
  hd2024 %>%
    filter(SECTOR %in% c(4, 5, 6, 7, 8, 9, 99)) %>%
```

```
    select(UNITID),

  # < 500 population
  ef2024b_inst %>%
    filter(total_enrollment < 500 | is.na(total_enrollment)) %>%
    select(UNITID)
) %>%
  distinct()
```

Helper function to remove rows

```
filter_and_write_ipeds_2024 <- function(df, remove_tbl, out_path) {
  df_clean <- df %>%
    dplyr::anti_join(remove_tbl, by = "UNITID")

  write.csv(df_clean, out_path, row.names = FALSE)

  invisible(df_clean)
}
```

Remove rows and write tables

```
# All IPEDS 2024 tables are filtered to exclude:
# (1) two-year institutions
# (2) institutions with <= 500 total enrollment
# using UNITID as the join key

tables_2024 <- list(
  hd2024       = hd2024,
  ef2024d      = ef2024d,
  ef2024b      = ef2024b,
  cost1_2024   = cost1_2024,
  cost2_2024   = cost2_2024
)

cleaned_tables <- lapply(names(tables_2024), function(name) {
  filter_and_write_ipeds_2024(
    tables_2024[[name]],
    remove_unitids,
    paste0("data/cleaned/", name, "_clean.csv")
  )
})

names(cleaned_tables) <- paste0(names(tables_2024), "_clean")
```