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")
```

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_ipeds_2024 <- function(df, remove_tbl) {
  df %>%
    anti_join(remove_tbl, by = "UNITID")
}
```

Remove rows and write tables

```
hd2024_clean <- filter_ipeds_2024(hd2024, remove_unitids)
ef2024d_clean <- filter_ipeds_2024(ef2024d, remove_unitids)
ef2024b_clean <- filter_ipeds_2024(ef2024b, remove_unitids)
cost1_2024_clean <- filter_ipeds_2024(cost1_2024, remove_unitids)

write.csv(hd2024_clean, "data/cleaned/hd2024_clean.csv", row.names = FALSE)
write.csv(ef2024d_clean, "data/cleaned/ef2024d_clean.csv", row.names = FALSE)
write.csv(ef2024b_clean, "data/cleaned/ef2024b_clean.csv", row.names = FALSE)
write.csv(cost1_2024_clean, "data/cleaned/cost1_2024_clean.csv", row.names = FALSE)
```