Data Formatting

Yuto Kitano

2024-08-13

Formatting Data

Formatting Semester Data

```
setwd("/Users/kitanoyuuto/Downloads/warmup training package/01_data/raw/semester_dummy")
semester_1 <- read.csv("semester_data_1.csv", header = TRUE, skip = 1)

semester_2 <- read.csv("semester_data_2.csv")

colnames(semester_2) <- colnames(semester_1)

semester_data <- rbind(semester_1, semester_2) %>%
  select(-Y)

semester_data <- semester_data %>%
  group_by(unitid) %>%
  mutate(semester_start_year = ifelse(any(semester == 1), min(year[semester == 1]), NA))

semester_data <- semester_data %>%
  mutate(semester_flag = ifelse(year >= semester_start_year, 1, 0))
```

Formatting Gradrate Data

```
library(readxl)
library(purrr)
library(stringr)

file_paths <- list.files(path = "/Users/kitanoyuuto/Downloads/warmup training package/01_data/raw/outcom
valid_file_paths <- file_paths[!str_detect(file_paths, "~\\$")]

data_list <- map(valid_file_paths, read_excel)


graduate_data <- bind_rows(data_list) %>%
  filter(year <= 2010) %>%
  mutate(
    tot4yrgrads = as.numeric(tot4yrgrads),
    totcohortsize = as.numeric(totcohortsize),
    women_gradrate_4yr = as.numeric(women_gradrate_4yr),
    m_4yrgrads = as.numeric(m_4yrgrads),
    m_cohortsize = as.numeric(m_cohortsize)
```

```
  ) %>%
  mutate(
    w_grad = round(women_gradrate_4yr * 0.01,3),
    t_grad = round(tot4yrgrads / totcohortsize,3),
    m_grad = round(m_4yrgrads / m_cohortsize,3)
  )
```

Formatting Covariates Data

```
setwd("/Users/kitanoyuuto/Downloads/warmup training package/01_data/raw/covariates")

covariates <- read_xlsx("covariates.xlsx") %>%
  rename("unitid" = "university_id")

covariates <- covariates %>%
  mutate(unitid = str_replace_all(unitid, "aaaa", ""))

cov_wide <- covariates %>%
  pivot_wider(names_from = category, values_from = value,
              values_fill = NA,
              id_cols = c(unitid, year)) %>%
  filter(year >= 1991 & year <= 2010)

id <- unique(semester_data$unitid)

cov_wide <- cov_wide %>%
  filter(cov_wide$unitid  %in% id) %>%
  mutate(year = as.numeric(year)) %>%
   mutate(unitid = as.numeric(unitid))
```

Left_Join

```
data <- left_join(semester_data, graduate_data , by = c("unitid", "year"))

data <- left_join(data, cov_wide, by = c("unitid", "year"))
```