# HarvardX PH125.9x Capstone Project - Education Data

Elan Weingarten

2025-06-08

```
data_dir = file.path(".", "data")

#data is split into four datasets since ElSi can only ouput 75 columns at a time
get_elsi_data = function (filename) {
  zip_filename = paste0(filename, ".zip")
  zip_path = file.path(data_dir, zip_filename)

  ed_data_filename = paste0(filename, ".csv")
  ed_data_path = file.path(data_dir, ed_data_filename)

  if (!file.exists(ed_data_path)) unzip(zip_path, exdir = data_dir)
  read_csv(ed_data_path, skip = 5)
}



filenames = c("1 ElSi School Info",
              "2 ElSi Characteristics",
              "3 ElSi Race",
              "4 ElSi Teacher and Race")

raw_data = lapply(filenames, get_elsi_data)

kable(
  sapply(raw_data, dim) %>%
    `rownames<-`(c("Rows", "Columns")) %>%
    `colnames<-`(filenames),
  caption = "Raw education data loaded with the following dimensions"
)
```

Table 1: Raw education data loaded with the following dimensions

| | 1 ElSi School Info | 2 ElSi Characteristics | 3 ElSi Race | 4 ElSi Teacher and Race |
|---|---|---|---|---|
| Rows | 114306 | 114306 | 114306 | 114306 |
| Columns | 75 | 71 | 73 | 33 |

```
make_clean_data = function (raw_data) {
  raw_data %>%
    rename( #easier names to work with
      school_name = `School Name`,
      state = `State Name [Public School] Latest available year`,
```

```
      full_id = `School ID (12-digit) - NCES Assigned [Public School] Latest available year`
    ) %>%
     # Restructure to turn school year from column name into a separate column
    pivot_longer(
      cols = -c(school_name, state, full_id) & !contains("Latest available year"),
      values_to = "value",
      names_to = "metric"
      ) %>%
    mutate(  #splits at the last space before the school year
      year = as.numeric(paste0("20", str_sub(metric, nchar(metric) - 1))),
      metric = str_sub(metric, 1, nchar(metric) - 8)
      ) %>%
    pivot_wider(
      names_from = metric,
      values_from = value
    )
}

clean_data = lapply(raw_data, make_clean_data)

#combine all four datasets into one big one
ed_data = clean_data[[1]] %>%
  left_join(select(clean_data[[2]], -c(school_name, state)), by = c("full_id", "year")) %>%
  left_join(select(clean_data[[3]], -c(school_name, state)), by = c("full_id", "year")) %>%
  left_join(select(clean_data[[4]], -c(school_name, state)), by = c("full_id", "year")) %>%
  rename_with(~ .x %>% #clean up column names
    str_remove_all("\\s\\[(District|Public School)\\]|\\.| Latest available year| [---] NCES Assigned") %>%
    str_trim() %>%                                    # remove leading/trailing spaces
    str_replace_all("[ /\\--]", "_") %>%                 # replace space, /, -, or - with _
    str_to_lower() %>%                                    # make all lower case
    str_replace_all("agency", "district") %>%            # replace agency with district
    str_replace_all("\\(sy_2017_18_onward\\)", "2018_onward") %>% #replace note about sy so that all pa
    str_remove_all("_\\([^\\)]*\\)")                       # remove all parentheticals
  ) %>%
  mutate( #merge school level (since changed after 2018)
    school_level = str_replace(str_replace(str_replace(
      str_replace(school_level, "1-Primary", "Elementary"),
      "2-Middle", "Middle"),
      "3-High", "High"),
      "4-Other", "Other")
  ) %>%
  mutate(
    school_level = if_else(is.na(school_level), school_level_2018_onward, school_level)
    ) %>%
  select(-school_level_2018_onward) %>%
  # Replace missing data with NA
  mutate(across(everything(), ~ ifelse(. %in% c("†", "-", "‡"), NA, .))) %>%
  filter( # remove all rows where there is no meaningful data
  !if_all(
    -c(school_name, state, full_id, district_id, year),
    ~ is.na(.)
    )
  ) %>%
```

```r
  # guess type of data
  type.convert(as.is = TRUE) #%>%
  # convert flags into boolean columns
  # mutate(
  #
  # )

kable(
  as.data.frame(t(dim(ed_data))) %>%
    `colnames<-`(c("Rows", "Columns")),
  caption = "Clean data is created with the following dimensions"
)
```

Table 2: Clean data is created with the following dimensions

| Rows | Columns |
|---|---|
| 998926 | 30 |

## Description of all variables used

```
## Column:      school_name
## Type:        character
## Description: Name of the school
## -----------------------------------------
## Column:      state
## Type:        character
## Description: U.S. state where the school is located
## -----------------------------------------
## Column:      full_id
## Type:        character
## Description: Full NCES-assigned 12-digit school ID
## -----------------------------------------
## Column:      district_id
## Type:        character
## Description: NCES-assigned district identifier
## -----------------------------------------
## Column:      year
## Type:        character
## Description: Simplified school year (e.g., 2024 for the 2023-24 school year)
## -----------------------------------------
## Column:      district_name
## Type:        character
## Description: Name of the school district
## -----------------------------------------
## Column:      school_id
## Type:        character
## Description: 7-digit NCES school ID
## -----------------------------------------
## Column:      county_name
## Type:        character
## Description: County where the school is located
## -----------------------------------------
## Column:      county_number
```

```
## Type:        character
## Description: County FIPS or internal number
## -----------------------------------------
## Column:      school_type
## Type:        character
## Description: Type of school operating (e.g., regular, alternative, special ed, etc.)
## -----------------------------------------
## Column:      district_type
## Type:        character
## Description: Type of district (e.g., regular district, regional education service agency (RESA), ind
## -----------------------------------------
## Column:      start_of_year_status
## Type:        character
## Description: Operational status at start of school year
## -----------------------------------------
## Column:      magnet_school
## Type:        character
## Description: Flag if the school is a magnet school
## -----------------------------------------
## Column:      charter_school
## Type:        character
## Description: Flag if the school is a charter school
## -----------------------------------------
## Column:      title_i_school_status
## Type:        character
## Description: Flag the type of Title I program
## -----------------------------------------
## Column:      reconstituted_flag
## Type:        character
## Description: Flag for reconstitution under school improvement
## -----------------------------------------
## Column:      school_level
## Type:        character
## Description: Level served (e.g., elementary, middle, high)
## -----------------------------------------
## Column:      total_students_all_grades
## Type:        numeric
## Description: Total student enrollment across all grades
## -----------------------------------------
## Column:      male_students
## Type:        numeric
## Description: Count of male students
## -----------------------------------------
## Column:      female_students
## Type:        numeric
## Description: Count of female students
## -----------------------------------------
## Column:      american_indian_alaska_native_students
## Type:        numeric
## Description: Count of American Indian/Alaska Native students
## -----------------------------------------
## Column:      asian_or_asian_pacific_islander_students
## Type:        numeric
## Description: Count of Asian/Pacific Islander students
```

```
## ------------------------------------------
## Column:      hispanic_students
## Type:        numeric
## Description: Count of Hispanic students
## ------------------------------------------
## Column:      black_or_african_american_students
## Type:        numeric
## Description: Count of Black/African American students
## ------------------------------------------
## Column:      white_students
## Type:        numeric
## Description: Count of White students
## ------------------------------------------
## Column:      nat_hawaiian_or_other_pacific_isl_students
## Type:        numeric
## Description: Count of Native Hawaiian/Pacific Islander students
## ------------------------------------------
## Column:      two_or_more_races_students
## Type:        numeric
## Description: Count of students of two or more races
## ------------------------------------------
## Column:      total_race_ethnicity
## Type:        numeric
## Description: Total of all race/ethnicity student counts
## ------------------------------------------
## Column:      full_time_equivalent_teachers
## Type:        numeric
## Description: FTE count of teachers
## ------------------------------------------
## Column:      pupil_teacher_ratio
## Type:        numeric
## Description: Ratio of students to teachers
## ------------------------------------------
```