Data Cleaning

Khanh Do, Joyce Gill

2026-01-31

Download data

```
hd2024 <- read.csv("https://raw.githubusercontent.com/joycegill/Advanced-Statistical-Modeling/main/data,
ef2024d <- read.csv("https://raw.githubusercontent.com/joycegill/Advanced-Statistical-Modeling/main/data
dim(hd2024)
```

```
## [1] 6072   72
```

```
dim(ef2024d)
```

```
## [1] 5578   33
```

```
# report 2-yr schools
two_yr_schools <- hd2024 %>%
  filter(SECTOR %in% c(4, 5, 6, 7, 8, 9, 99)) %>%
  select(UNITID)

# report < 500 population schools
less_than_500_pop_schools <- ef2024d %>%
  filter(UGENTERN <= 500 | is.na(UGENTERN)) %>%
  select(UNITID)

dim(two_yr_schools)
```

```
## [1] 3205   1
```

```
head(two_yr_schools)
```

```
##    UNITID
## 1 100760
## 2 101028
## 3 101143
## 4 101161
## 5 101240
## 6 101277
```

```
dim(less_than_500_pop_schools)
```

```
## [1] 3712    1
```

```
head(less_than_500_pop_schools)
```

```
##     UNITID
## 1 100690
## 2 100812
## 3 101116
## 4 101277
## 5 101301
## 6 101365
```