Ingest DOE Student Record Demographics

## LOAD LIBRARIES AND FUNCTIONS

library(data.table)  
library(dataplumbr)  
library(here)  
library(inspectdf)  
library(maditr)

## DOE Student Record Demographics

### Load data file

doe\_student\_records <- fread(here("data/original/q5/DOE/Student Records.csv"), colClasses = "character")

### Standardize column names

colnames(doe\_student\_records) <- name.standard\_col\_names(colnames(doe\_student\_records))

### Check if more than one record per unique\_id and calendar\_year

multiples <- nrow(doe\_student\_records[, .N,.(unique\_id, school\_year)][N > 1])  
multiples

## [1] 2810

### Apply Deduplication algorithm to get Demographics By Year

doe\_student\_dmgs <- doe\_student\_records[, .(birth\_month, birth\_year, race\_type, ethnic\_flag, prek\_funding\_code),.(unique\_id, school\_year)]  
  
set.dedup\_choice <- function(df) {  
 dt <- data.table::setDT(df)  
 for (j in colnames(dt)) {  
 data.table::set(dt, j = j, value = dt[get(j) != "", .N, j][order(-N)][, ..j][1])  
 }  
 dt[1]  
}  
  
set.dedup\_choice\_by\_key <- function(df, key = "uid") {  
 if (exists("out\_dt") == TRUE) rm(out\_dt, envir = globalenv())  
  
 dt <- data.table::setDT(df)  
 unique\_keys <- unique(dt[, get(key)])  
 key\_cnt <- length(unique\_keys)  
 pb <- progress::progress\_bar$new(format = "[:bar] :current/:total :percent eta: :eta", total = key\_cnt)  
  
 for (k in unique\_keys) {  
 pb$tick()  
 g <- dt[get(key)==k]  
 r <- set.dedup\_choice(g)  
 if (exists("out\_dt") == FALSE) out\_dt <- r else out\_dt <- rbindlist(list(out\_dt, r))  
 }  
  
 out\_dt  
}  
  
doe\_student\_dmgs\_dedup <- set.dedup\_choice\_by\_key(doe\_student\_dmgs, "unique\_id")  
  
# verify only one code per id per year  
nrow(doe\_student\_dmgs\_dedup[, .N, .(unique\_id)][N > 1])

## [1] 0

### write to csv

fwrite(doe\_student\_dmgs\_dedup, here("data/working/DOE/doe\_student\_records\_by\_year\_dmgs\_prek.csv"))