MCAS Dataset Construction

About the Data: https://profiles.doe.mass.edu/help/data.aspx?section=assess

Statewide Reports: https://profiles.doe.mass.edu/statereport/

Variables needed from 2018 Grade 10 Mathematics MCAS:

- Percentage of students scoring proficient or advanced
- Class size
- Math class size
- Student-to-teacher ratio
- Race, largest minority group
- Expenditures per pupil
- Percentage of economically disadvantaged students
- Attendance rate

Downloading the data:

- To generate the desired reports, choose Report Type (School), Year, Subject.
- MCAS Achievement Results: percent of students at each achievement level by grade and subject
  - Saved as `mcas_achievement.csv`
- Class Size by Gender and Selected Population Data: number of classes and average class size for each subject by gender, Limited English Proficient and Low Income
  - Saved as `class_features.csv` and `math_class_features.csv`
- Enrollment by Selected Population: number and percent of public school students in student groups First Language Not English (FLNE), English Learners (Els), Students with Disabilities, High Needs, Low-income (2022 to present), Economically Disadvantaged (2015 -2021), and Low-income (prior to 2015)
  - Saved as `selected_populations.csv`
- Enrollment by Race/Gender: percent of public school students by race and gender
  - Saved as `race_gender.csv`
- Per Pupil Expenditures: per pupil expenditures for each district
  - Saved as 'per_pupil_expenditures.csv'
  - School code's first four digits correspond to district code
- Teacher Data: Student-teacher ratio
  - saved as `teacher_data.csv`

- Student Attendance Report: student attendance rate

```
# read in achievement data, clean names, extract info
achievement <- read_csv("statewide_reports/mcas_achievement.csv", skip = 1)
achievement <- clean_names(achievement)
achievement <- achievement %>%
  filter(subject == "MATHEMATICS") %>%
  select(school_name, school_code, p_a_percent,
         no_of_students_included)

# add district code
achievement$district_code <- substr(achievement$school_code,
                        start = 1,
                        stop = 4)

# clean the school name
achievement$school_name <- str_split(achievement$school_name,
                                     "- ", simplify = TRUE)[, 2]

# read in class features, clean names, extract info
class_features <- read_csv("statewide_reports/class_features.csv", skip = 1)
class_features <- clean_names(class_features)
class_features <- class_features %>%
  select(school_code, average_class_size)

# read in math class features, clean names, extract info
math_class_features <- read_csv("statewide_reports/math_class_features.csv", skip = 1)
math_class_features <- clean_names(math_class_features)
math_class_features <- math_class_features %>%
  select(school_code, average_class_size) %>%
  rename(average_math_class_size = average_class_size)

# read in selected populations, clean names, extract info
selected_populations <- read_csv("statewide_reports/selected_populations.csv", skip = 1)
selected_populations <- clean_names(selected_populations)
selected_populations <- selected_populations %>%
  select(school_code, english_language_learner_percent,
         students_with_disabilities_percent, economically_disadvantaged_percent)

# read in teacher data, extract info, clean data
teacher_data <- read_csv("statewide_reports/teacher_data.csv", skip = 1)
teacher_data <- clean_names(teacher_data)
teacher_data <- teacher_data %>%
  select(school_code, student_teacher_ratio)
teacher_data$student_teacher_ratio <- str_extract(teacher_data$student_teacher_ratio, "[^ to]+

# read in enrollment by race/gender, clean names, extract info
race <- read_csv("statewide_reports/race_gender.csv", skip = 1)
```

```
race <- clean_names(race)
race <- race %>%
  select(-school_name, -males, -females)

# add variable indicating largest minority
largest_minority <- rep(NA, nrow(race))
minorities_colnames <- c("african_american", "asian", "hispanic",
                         "native_american", "native_hawaiian_pacific_islander",
                         "multi_race_non_hispanic")
minorities_values <- c("African American or Black", "Asian",
                       "Hispanic or Latino", "Native American",
                       "Native Hawaiian or Other Pacific Islander",
                       "Multi-race, Non-Hispanic")

for(k in 1:nrow(race)){

  ind <- which.max(race[k, minorities_colnames])

  largest_minority[k] <- minorities_values[ind]
}

race$largest_minority <- largest_minority

# read in per pupil expenditures, clean names, extract info
expenditures <- read_csv("statewide_reports/per_pupil_expenditures.csv", skip = 1)
expenditures <- clean_names(expenditures)
expenditures <- expenditures %>%
  select(district_name, district_code, total_expenditures_per_pupil)

# trim district code
expenditures$district_code <- substr(expenditures$district_code,
                       start = 1,
                       stop = 4)

# clean dollar value
expenditures <- expenditures %>%
  mutate(total_expenditures_per_pupil = parse_number(total_expenditures_per_pupil))

# read in attendance, clean names, extract info
attendance <- read_csv("statewide_reports/attendance.csv", skip = 1)
attendance <- clean_names(attendance)
attendance <- attendance %>%
  select(school_code, attendance_rate)
```

```
# joining the data
mcas <- left_join(achievement, class_features,
                  join_by("school_code" == "school_code"))
```

```r
mcas <- left_join(mcas, math_class_features,
                  join_by("school_code" == "school_code"))
mcas <- left_join(mcas, selected_populations,
                  join_by("school_code" == "school_code"))
mcas <- left_join(mcas, teacher_data,
                  join_by("school_code" == "school_code"))
mcas <- left_join(mcas, race,
                  join_by("school_code" == "school_code"))
mcas <- left_join(mcas, attendance,
                  join_by("school_code" == "school_code"))
mcas <- left_join(mcas, expenditures,
                  join_by("district_code" == "district_code"))
```

```r
# cleaning data frame
mcas <- mcas %>%
  select(-school_code, -district_code) %>%
  relocate(school_name, district_name) %>%
  rename(PA_perc = p_a_percent,
         number_of_students = no_of_students_included,
         english_learner = english_language_learner_percent,
         students_disabilities = students_with_disabilities_percent,
         econ_dis = economically_disadvantaged_percent,
         exp_per_pupil = total_expenditures_per_pupil) %>%
  mutate(student_teacher_ratio = as.numeric(student_teacher_ratio)) %>%
  filter(school_name != "State Totals")

mcas <- mcas %>%
  mutate(majority = case_when(
    white < 50 ~ "Minority",
    white >= 50 ~ "White"
  ))

save(mcas, file = "mcas.Rdata")
```