MCAS Data Analysis Examples

MCAS Question Statement

The Massachusetts Comprehensive Assessment System (MCAS) is a standardized exam administered to students in the state of Massachusetts since 1993. Under law, students educated with public funds are required to participate in statewide testing. Students take exams according to their grade level in subjects such as Mathematics and English Language Arts; passing grades on the Grade 10 MCAS are required for high school graduation. Exam results are used to check student progress as well as measure school and district performance.

However, research has shown that variation in standardized test scores is often explained by factors such as race and socioeconomic class; students of color and students from economically disadvantaged backgrounds tend to score lower on standardized tests as a result of having less access to educational resources than their peers. For example, a *New York Times* article from October 2020 reported the results of a study demonstrating that in the United States, students performed worse on standardized tests for every additional day that was 80 degrees Fahrenheit or higher—the association was observed for Black and Hispanic students, in addition to students with lower family income, but not for white students.

In this problem, you will use data from the 2018 Grade 10 Mathematics MCAS to investigate evidence of achievement gaps in standardized test score at the school level. The `mcas` data contains information on 355 schools in Massachusetts. Scores on the MCAS are classified into levels: warning/failing, needs improvement, proficient, and advanced. Students who score in the "proficient" category are said to "demonstrate a solid understanding of challenging subject matter", while those in the "advanced" category are said to "demonstrate a comprehensive and in-depth understanding of rigorous subject matter"; the variable `PA_perc` represents the percentage of students at a school who score at the proficient or advanced level.

The descriptions of the variables are as follows.

- `school_name`: school name
- `district_name`: school district name
- `PA_perc`: percentage of students scoring proficient or advanced
- `number_of_students`: total number of students, including special education beyond grade 12
- `average_class_size`: average class size, regardless of subject
- `average_math_class_size`: average math class size
- `english_learner`: percentage of students for whom English is not their first language and who cannot perform ordinary classroom work in English.
- `students_disabilities`: percentage of students in the school who have an individual education plan (IEP) identifying special learning needs
- `econ_dis`: percentage of students from an economically disadvantaged background
- `student_teacher_ratio`: average student-teacher ratio
- `attendance_rate`: the number of full-time equivalent student-days attended by full-time students in grades 1-10 as a percentage of the total number of possible student-days attended

over the period

- `exp_per_pupil`: amount spent by the district per pupil, in dollars
- `majority`: coded `White` if >= 50% of the students are white, otherwise coded `Minority`
- `african_american`: percentage of students with origins in any of the black racial groups of Africa
- `asian`: percentage of students having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent
- `hispanic`: percentage of students of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race
- `white`: percentage of students having origins in any of the original peoples of Europe, the Middle East, or North Africa
- `native_american`: percentage of students having origins in any of the original peoples of North and South America (including Central America), and who maintain tribal affiliation or community attachment
- `native_hawaiian_pacific_islander`: percentage of students having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands
- `multi_race_non_hispanic`: percentage of students selecting more than one racial category and non-Hispanic
- `largest_minority`: largest minority group

Some additional background on certain variables:

- Whether a student is economically disadvantaged is a proxy measure for student family income; a student is considered economically disadvantaged if they are participating in one or more of the following programs: the Supplemental Nutrition Assistance Program (SNAP), the Transitional Assistance for Families with Dependent Children (TAFDC), the Department of Children and Families' (DCF) foster care program, MassHealth (Medicaid).

- Expenditures per pupil is a proxy measure for the amount of funding available to a school district. This variable is measured by district (i.e., constant for schools in the same district).

Use these data to answer the following questions.

a) Construct a useful visualization for comparing the percentage of economically disadvantaged students between schools where less than 50% of students are white and schools where 50% or more of students are white. Describe what you see.

b) Fit a linear regression model predicting `PA_perc` from `majority`, `average_math_class_size`, `attendance_rate`, and `student_teacher_ratio`. Don't include any interaction terms.

   i. Check whether the model assumptions are reasonably satisfied. Summarize your findings.

   ii. Interpret the estimated coefficient for `majority`, as well as the associated p-value and confidence interval. Be sure to frame interpretations in the context of the data.

c) Add `econ_dis` to the model from part b) and interpret the estimated coefficient for `majority`. Explain how the interpretation of the coefficient for `majority` in this model differs from the previous interpretation in part b), ii., using terms accessible to someone who has not taken a statistics course.

d) Investigate whether the association between the percentage of students scoring at the proficient/advanced levels and the percentage of students who are economically disadvantaged differs by whether less than 50% of the students are white, after adjusting for average math class size, attendance rate, and student-teacher ratio. Summarize the results:

- State the null and alternative hypotheses.
- State the test statistic and p-value.
- Interpret the relevant estimates (for answering the scientific question of interest) in context of the data.
- *Note*: For this context, consider interpreting the coefficient estimates in terms of per 10-unit change rather than 1-unit change.

e) Consider schools with the following features: an average math class size of 20 students, attendance rate of 93%, and student to teacher ratio of 11, where 60% of students are African American and 45% of students are economically disadvantaged. Based on the model from part c), what is the predicted average percentage of students scoring at the proficient/advanced level on the Mathematics Grade 10 MCAS for such schools? Report and interpret an appropriate interval estimate as part of your answer that could be used to estimate the average PA percentage for schools with such features.

f) Suppose that the results from these analyses will be discussed at a future meeting of the Racial Imbalance Advisory Council, which advises the Massachusetts Commissioner of Education and the Board of Education on matters related to providing access to effective educational programs for all students in the state regardless of race or socioeconomic class.

Prepare a statement, no more than ten sentences long, summarizing the main findings of the analyses with respect to understanding whether these data show indication of poverty and/or race-based achievement gaps in Grade 10 Mathematics MCAS scores. Be sure to use language that is accessible to a general audience and make specific references to previous numerical results.

Example Code

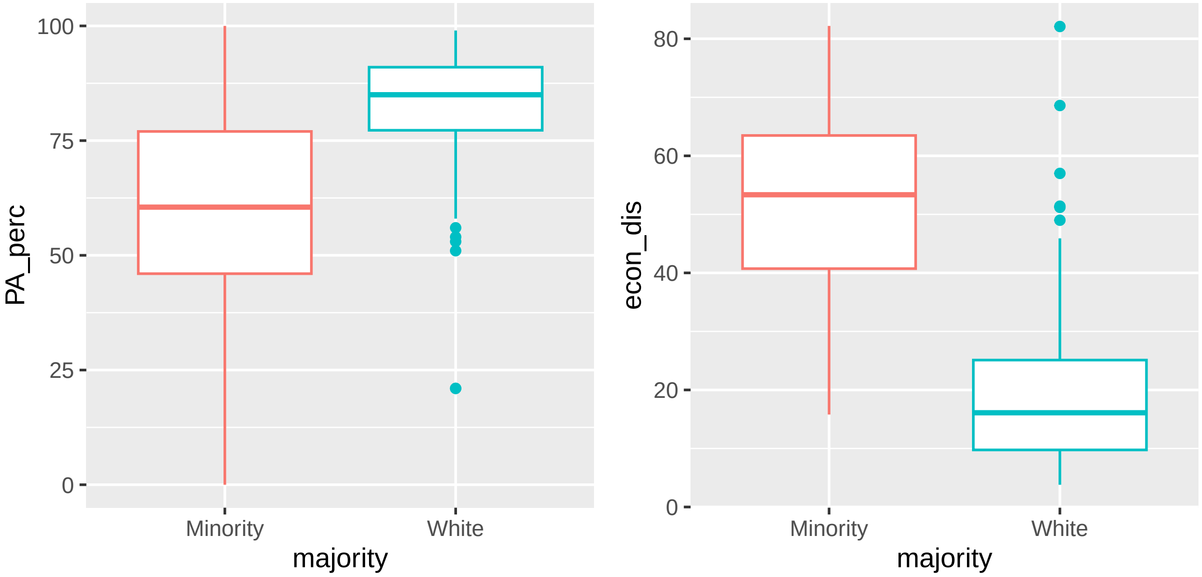
```r
#load any necessary packages
library(tidyverse)
library(patchwork)
library(moderndive)

#load the data
load("mcas.Rdata")
```

```r
p1 <- mcas %>%
  ggplot(aes(y = PA_perc, x = majority, col = majority)) +
  geom_boxplot() +
  guides(col = "none")

p2 <- mcas %>%
  ggplot(aes(y = econ_dis, x = majority, col = majority)) +
  geom_boxplot() +
  guides(col = "none")

p1 + p2
```



```r
mcas %>%
  group_by(majority) %>%
  summarize(mean_pa_perc = mean(PA_perc),
            mean_econ_dis = mean(econ_dis))
```

```
## # A tibble: 2 x 3
##   majority mean_pa_perc mean_econ_dis
##   <chr>           <dbl>         <dbl>
## 1 Minority         61.3          52.2
## 2 White            82.8          19.1
```

```r
#fit initial model
mod <- lm(PA_perc ~ majority + average_math_class_size + attendance_rate +
             student_teacher_ratio, data = mcas)

#model summary
get_regression_table(mod)
```

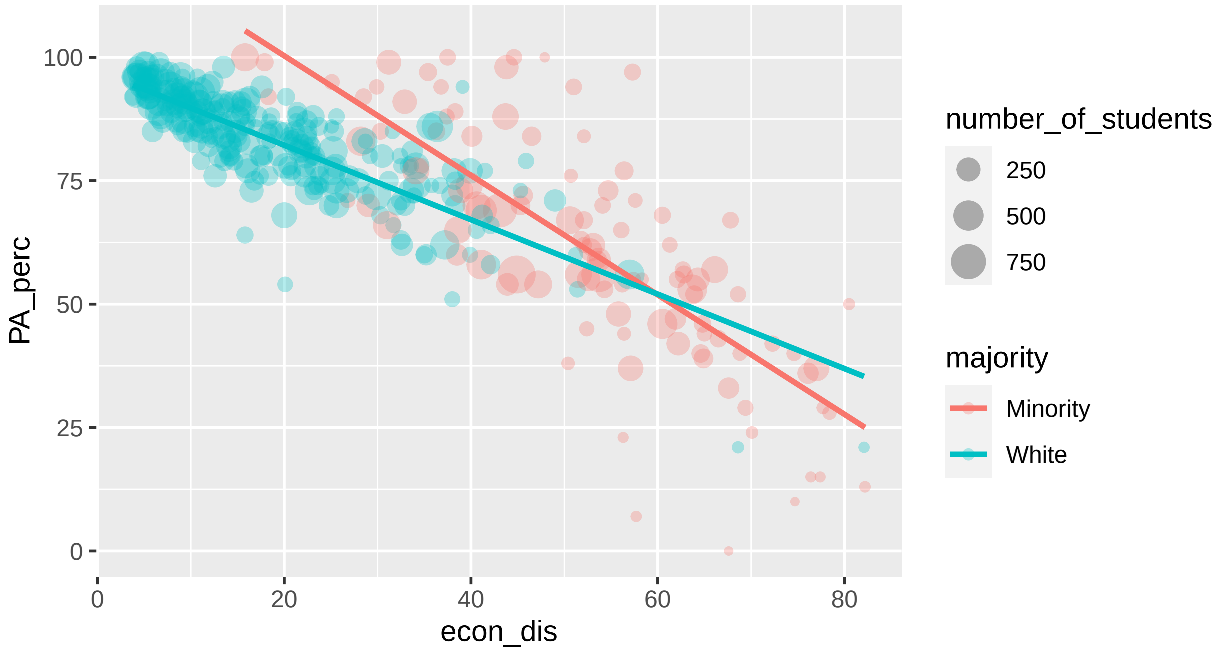
```
## # A tibble: 5 x 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept               -100.        13.2      -7.60     0    -126.      -74.2
## 2 majority: White           15.0       1.60      9.36     0      11.8     18.1
## 3 average_math_class_size    1.35      0.218     6.18     0       0.921    1.78
## 4 attendance_rate            1.52      0.153     9.98     0       1.22     1.82
## 5 student_teacher_ratio      0.133     0.239     0.554    0.58   -0.338    0.604
```

```r
#fit model with econ_dis
mod2 <- lm(PA_perc ~ majority + average_math_class_size +
             attendance_rate + student_teacher_ratio + econ_dis,
           data = mcas)

#model summary
get_regression_table(mod2)
```

```
## # A tibble: 6 x 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 31.6      12.8       2.46    0.014    6.34    56.9
## 2 majority: White           -5.89      1.76     -3.34    0.001   -9.36    -2.42
## 3 average_math_class_size    0.709     0.17      4.18    0        0.375    1.04
## 4 attendance_rate            0.643     0.128     5.04    0        0.392    0.894
## 5 student_teacher_ratio     -0.091     0.182    -0.504   0.615   -0.449    0.266
## 6 econ_dis                  -0.748     0.046   -16.2     0       -0.838   -0.657
```

```r
#visualize (without adjusting for additional confounders)
ggplot(mcas, aes(y = PA_perc, x = econ_dis, col = majority)) +
  geom_point(alpha = 0.30, aes(size = number_of_students)) +
  geom_smooth(method = "lm", se = FALSE)
```



```r
#fit model with interaction term
mod_interact = lm(PA_perc ~ average_math_class_size + attendance_rate +
             student_teacher_ratio + majority*econ_dis, data = mcas)

#model summary
get_regression_table(mod_interact)
```

```
## # A tibble: 7 x 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 54.6      13.6       4.00    0         27.8     81.4
## 2 average_math_class_size    0.634     0.167     3.81    0          0.306    0.962
## 3 attendance_rate            0.54      0.127     4.26    0          0.29     0.789
## 4 student_teacher_ratio     -0.114     0.177    -0.646   0.519   -0.463    0.234
## 5 majority: White          -19.6       3.63     -5.40    0        -26.7    -12.4
## 6 econ_dis                  -0.98      0.07    -13.9     0         -1.12    -0.841
## 7 majority: White:econ_d~    0.337     0.079     4.29    0          0.183    0.492
```

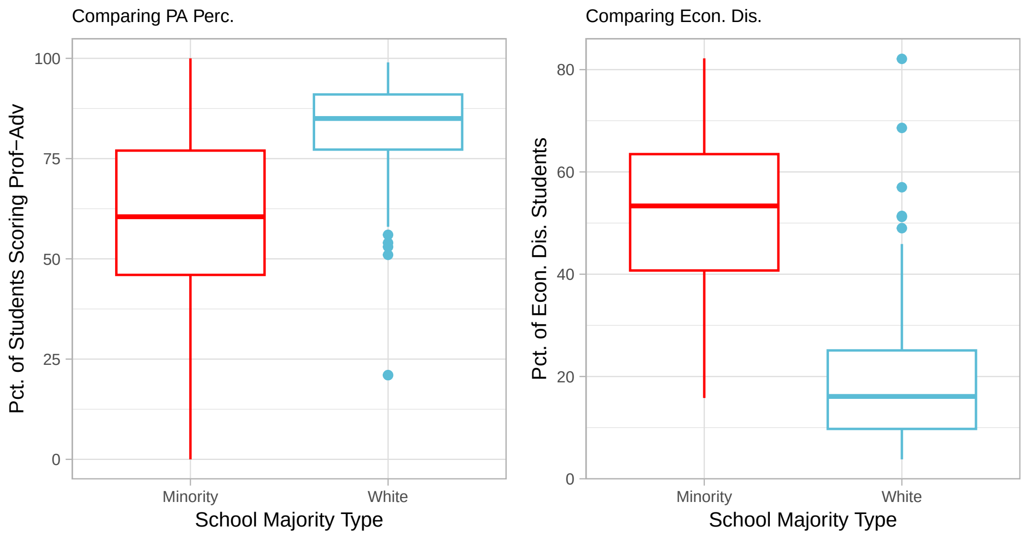
Figures from Slides

```r
library(wesanderson)
wes_colors <- wes_palette("Darjeeling1")[c(1, 5)]
```

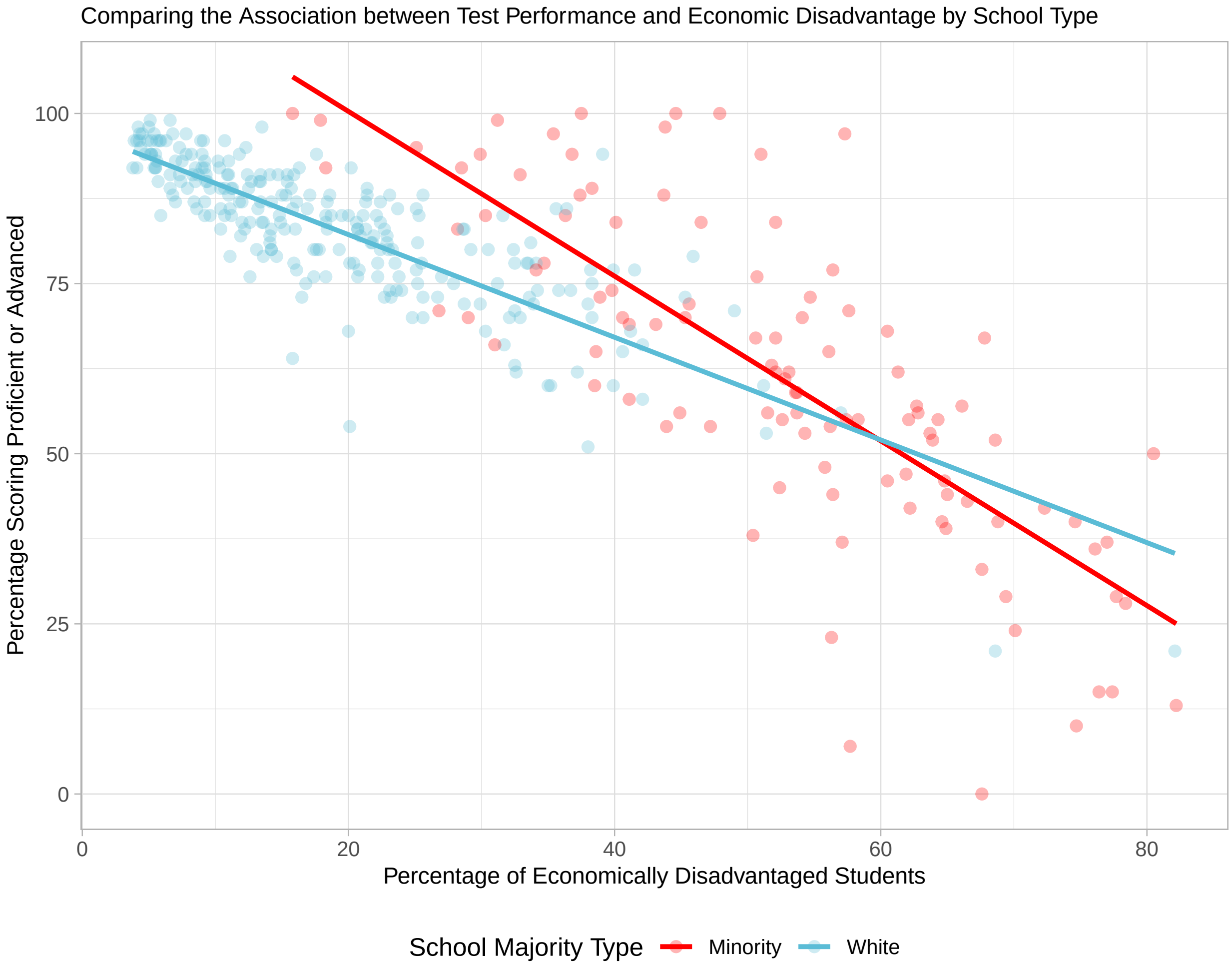
```r
p1 <- mcas %>%
  ggplot(aes(y = PA_perc, x = majority,
             col = majority)) +
  geom_boxplot() +
  guides(col = "none") +
  labs(title = "Comparing PA Perc.",
       y = "Pct. of Students Scoring Prof-Adv",
       x = "School Majority Type") +
  scale_color_manual(values = wes_colors) +
  theme_light() +
  theme(axis.title = element_text(size = 9),
        plot.title = element_text(size = 8),
        axis.text = element_text(size = 7))

p2 <- mcas %>%
  ggplot(aes(y = econ_dis, x = majority,
             col = majority)) +
  geom_boxplot() +
  guides(col = "none") +
  labs(title = "Comparing Econ. Dis.",
       y = "Pct. of Econ. Dis. Students",
       x = "School Majority Type") +
  scale_color_manual(values = wes_colors) +
  theme_light() +
  theme(axis.title = element_text(size = 9),
        plot.title = element_text(size = 8),
        axis.text = element_text(size = 7))

p1 + p2
```



```
mcas %>%
  ggplot(aes(y = PA_perc, x = econ_dis,
             col = majority)) +
  geom_point(alpha = 0.30, size = 2) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_manual(values = wes_colors) +
  theme_light() +
  theme(axis.title = element_text(size = 10),
        plot.title = element_text(size = 10),
        legend.position = "bottom") +
  labs(y = "Percentage Scoring Proficient or Advanced",
       x = "Percentage of Economically Disadvantaged Students",
       title = "Comparing the Association between Test Performance and Economic Disadvantage by
  guides(col = guide_legend(title = "School Majority Type"))
```



Comparing the Association between Test Performance and Economic Disadvantage by School Type