Cecilia Cho
UCLA Extension
COM SCI-X 450.1 Introduction to Data Science
Final Project
GitHub repository URL: https://github.com/CeciliaCali/datascience450

Student Test Scores

March 24, 2021

OVERVIEW

This project aims to find out what factors determine a student's success in tests. By learning which factors play the largest roles in determining student scores, we can better customize learning interventions to raise student's scores.

HYPOTHESES

The hypotheses I start off with are the following:

1. Gender does not affect test scores.
2. Race / ethnicity affect test scores.
3. The higher the level of parents' education, the higher the student test scores.
4. Students receiving free / reduced cost lunches will have lower test scores.
5. Completed test preparation will improve test scores.

DATA SET

Overview

The data set is of U.S. high school students' test scores. The data is available on Kaggle.[1]

[1] https://www.kaggle.com/spscientist/students-performance-in-exams.

Observations

The data set contains 1000 observations of 8 variables. I use as target variables the following:

- Math score - integers.
    - Minimum of 0.
    - Maximum of 100.
- Reading score - integers.
    - Minimum of 17.
    - Maximum of 100.
- Writing score - integers.
    - Minimum of 10.
    - Maximum of 100.

The remaining 5 variables which I use as predictor variables are:

- Gender
    - Female - 518 observations
    - Male - 482 observations
- Race/Ethnicity
    - Group A - 89 observations
    - Group B - 190 observations
    - Group C - 319 observations
    - Group D - 262 observations
    - Group E - 140 observations
- Parental Level of Education
    - some high school - 179 observations
    - high school - 196 observations
    - some college - 226 observations
    - associate's degree - 222 observations
    - bachelor's degree - 118 observations
    - master's degree - 59 observations
- Lunch[2]
    - free/reduced - 355 observations
    - Standard - 645 observations
- Test Preparation Course

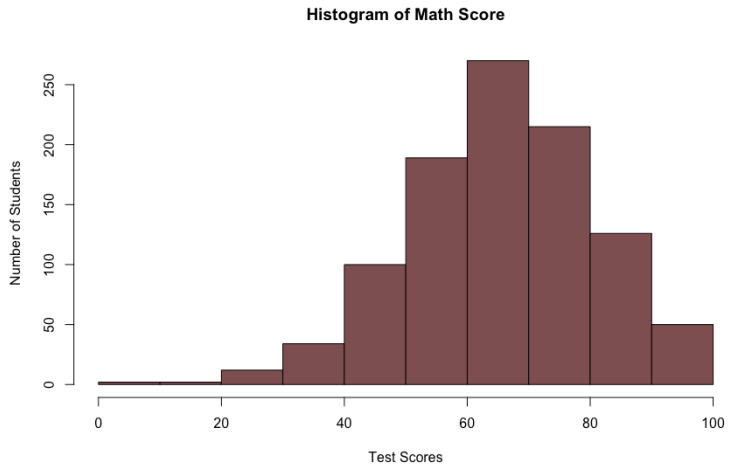
[2] The National School Lunch Program is a federally assisted meal program that provides reduced cost or free lunches to children meeting low-family income requirements. For the purposes of this study, lunch status can be a proxy for family income where free/reduced is low household income and standard lunch is not low household income.

- completed - 358 observations
- None - 642 observations

DISTRIBUTION OF TEST SCORES

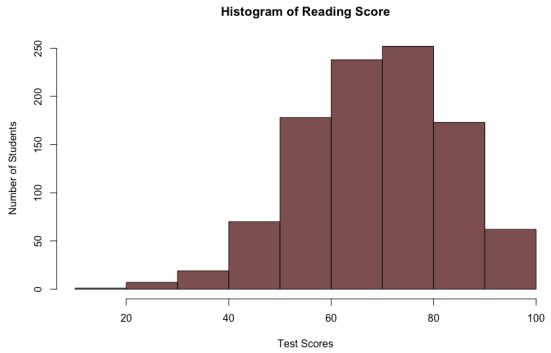
Math

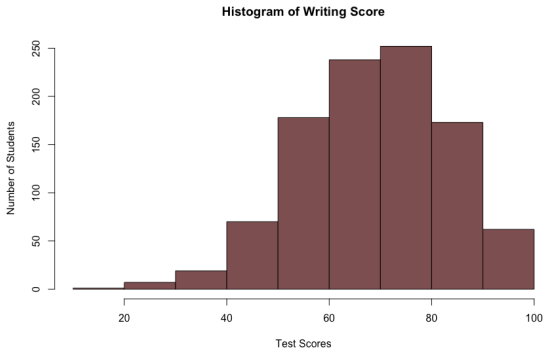
The median math score is 66, and the histogram bar for the range of 60-69 is the tallest.



Reading and Writing

The median reading score is 70, and the histogram bar for the range of 70-79 is the tallest. The median writing score is 69, and the histogram bar for the range of 60-69 is the tallest.





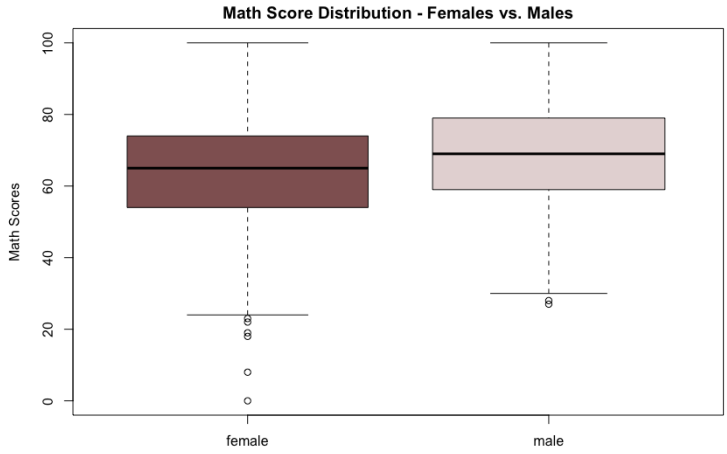
Reading and writing scores are highly correlated at 95%, so I address reading and writing scores together.

	math.score	reading.score	writing.score
math.score	1.0000000	0.8175797	0.8026420
reading.score	0.8175797	1.0000000	0.9545981
writing.score	0.8026420	0.9545981	1.0000000

MATH SCORES

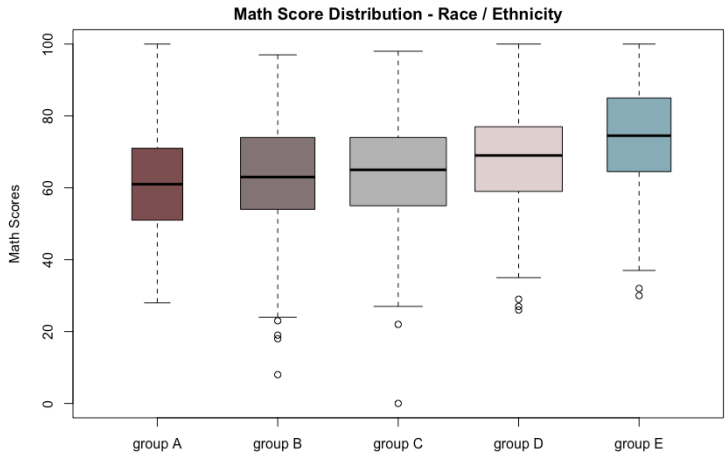
Comparison by Gender

It appears that males generally score higher than females in math. While both female and male cohorts have students that attain the maximum score of 100, there are more lower math scores among the female students.



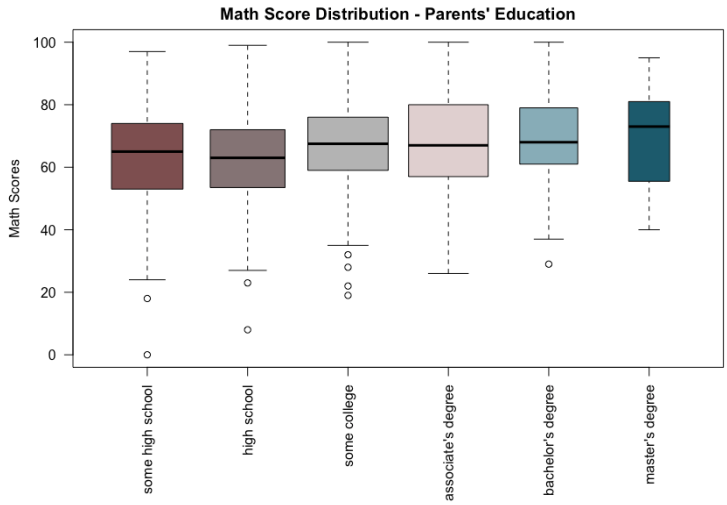
Comparison by Race / Ethnicity

It appears Group A slightly scored lower in math than the other racial / ethnic groups. Group E appears to score higher in math than the other groups.



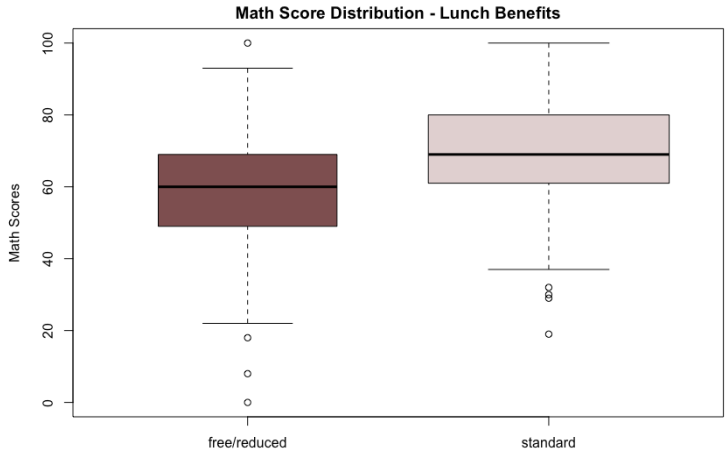
Comparison by Parental Level of Education

It appears that students whose parents' highest level of education was from having some high school education to holding a bachelor's degree had similar math scores. The median math score of students whose parents have master's degree is the highest of the set.



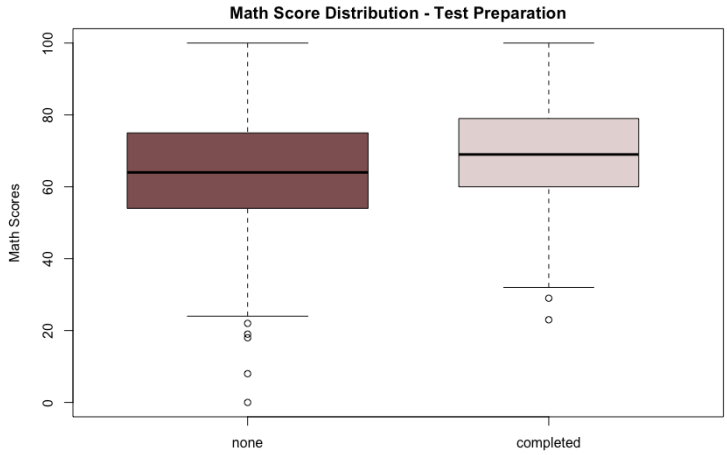
Comparison by Lunch Benefits

It appears that students whose family income is above the free / reduced cost lunch requirement score higher than students whose family income meet the free / reduced cost lunch requirement. The median (50%) score for students who pay the standard cost of lunch is about at the third quartile (75%) of students who qualify for free / reduced cost lunch.



Comparison by Test Preparation Course

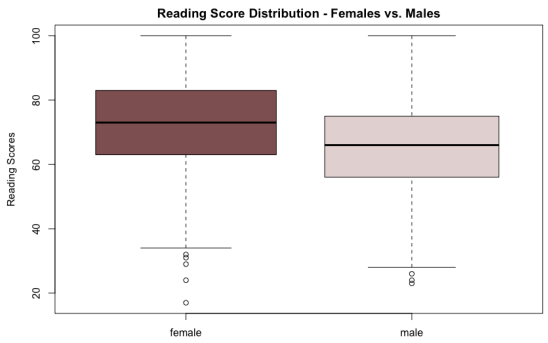
It appears that students who completed test preparation courses scored slightly higher than students who did not do test preparation courses.

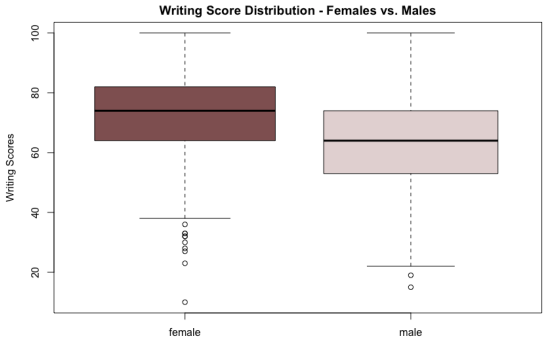


READING AND WRITING SCORES

Comparison by Gender

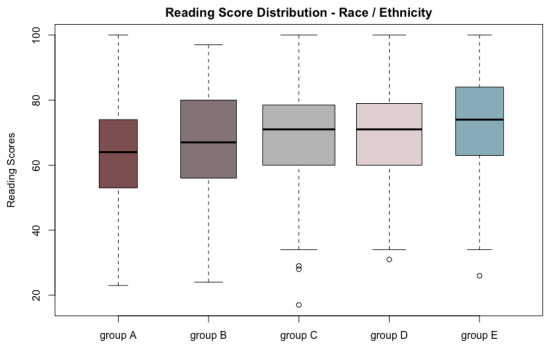
It appears that females generally score higher than males in reading and writing.

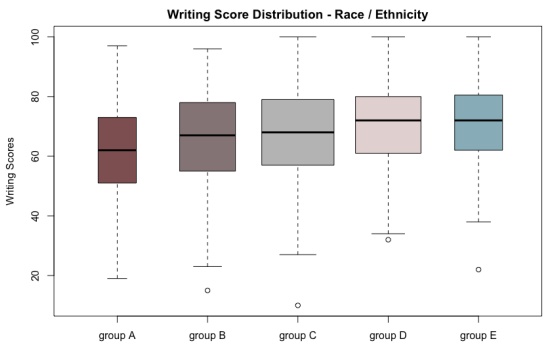




Comparison by Race / Ethnicity

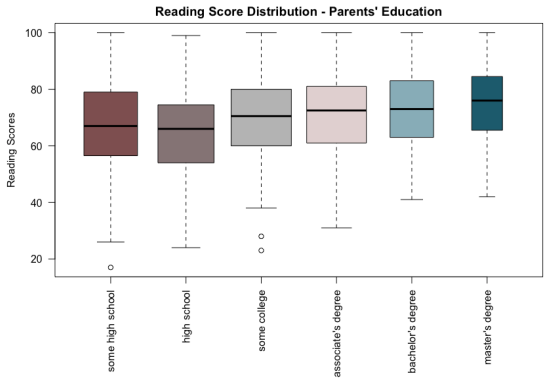
It appears that Group A scored the lowest in reading and writing out of the five groups. Group E appears to have scored the highest in both reading and writing.

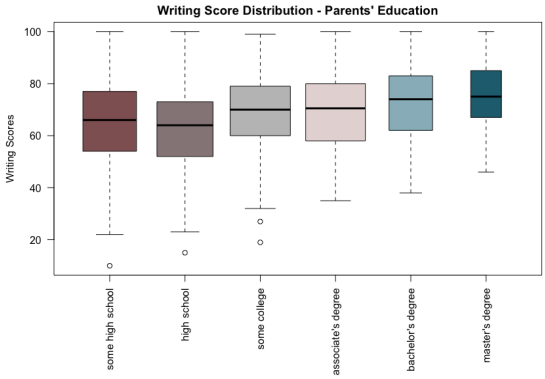




Comparison by Parental Level of Education

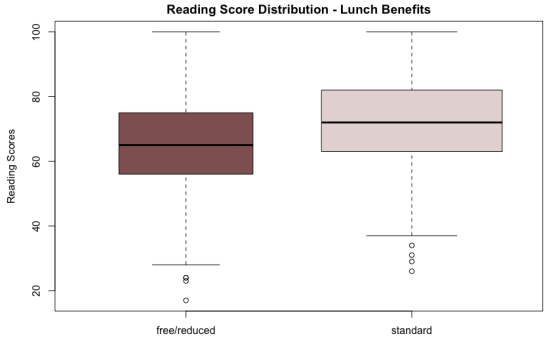
It appears that as parents' highest level of education increases, reading and writing scores also slope upwards.

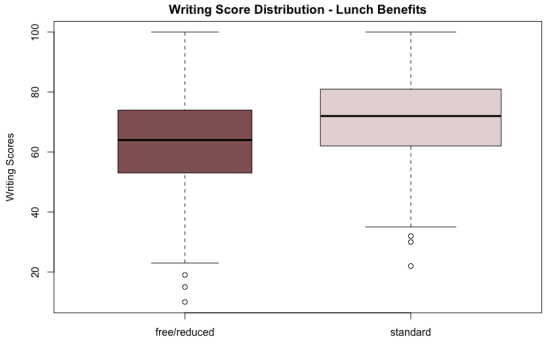




Comparison by Lunch Benefits

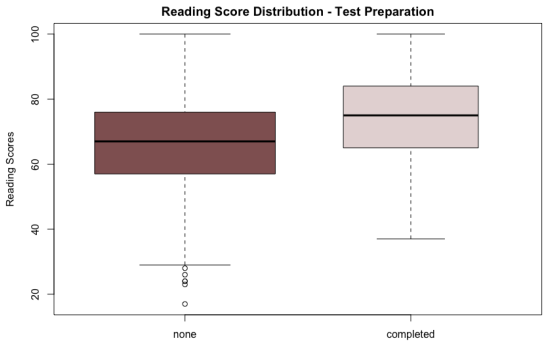
Again, it appears that students whose family income is above the free / reduced cost lunch requirement score higher than students whose family income meet the free / reduced cost lunch requirement. This is true for reading scores and writing scores.

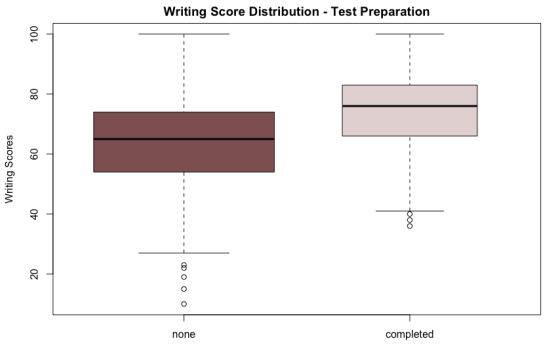




Comparison by Test Preparation Course

It appears that test preparation courses are helpful to scoring higher and reading and writing.





REGRESSIONS - SUPERVISED MACHINE LEARNING

Linear Regression Model and Supervised Machine Learning

The regression models I ran are:

$$Test\ Score = \alpha + \beta_1 \times (Gender) + \beta_2 \times (Race.ethnicity) + \beta_3 \times (Parental\ level\ of\ education)$$

$$+ \beta_4 \times (Lunch) + \beta_5 \times (Test\ prep\ course)$$

I ran linear and logistic multivariate regressions using only the math scores as the target variable to see if the data appeared to be a linear relationship or a logistic one. Based on the residuals vs. fitted values, and QQ Plot, it appeared that the data was linear and that logistic regression did not appear to offer a better fit. Therefore, I used linear regressions to find the predictors for all test scores (math, reading, writing).

I set the training / test set split into 75-25 because there were 1000 observations which would still leave 250 observations for testing.

Math Scores Model

In the case of math scores, all predictor values were statistically significant. The model was:

$$Math\ Score\ =\ 55\ +\ 4.8\ \times (Male)\ +4.4\ \times (Race.ethnicity = Group\ D)\ +8.8\ \times (Race.ethnicity = Group\ E)$$

$$+3.06\ \times (Parent\ edu\ level\ =\ associate's\ degree)\ +5.4\ \times (Parent\ edu\ level\ =\ bachelor's\ degree)$$

$$+\ 5.9\ \times (Parent\ edu\ level\ =\ master's)\ +10.2\ \times (Lunch = standard)\ -\ 5.7\ (No\ test\ prep\ course)$$

The adjusted R-squared is only 0.2212 which means that 22.12% of the variation in math score is explained by the model.

The predictor variable with the largest coefficient is lunch. Being from a household with average to above-average income adds 10.2 points to a student's math scores. The next most impactful predictor variable is being of the race / ethnic group E which adds 8.8 points to a student's math score.

```
Call:
lm(formula = math.score ~ gender + race.ethnicity + parental.level.of.education +
    lunch + test.preparation.course, data = traindf)

Residuals:
    Min      1Q  Median      3Q     Max
-44.896  -9.225   0.486   9.609  31.106

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                55.6596     2.1411  25.996  < 2e-16 ***
gendermale                                  4.8107     0.9654   4.983 7.80e-07 ***
race.ethnicitygroup B                       1.1610     2.0064   0.579  0.56299
race.ethnicitygroup C                       2.3909     1.8490   1.293  0.19639
race.ethnicitygroup D                       4.4073     1.8911   2.331  0.02005 *
race.ethnicitygroup E                       8.8119     2.0780   4.241 2.51e-05 ***
parental.level.of.educationhigh school     -1.0843     1.5923  -0.681  0.49613
parental.level.of.educationsome college     2.5689     1.5167   1.694  0.09073 .
parental.level.of.educationassociate's degree   3.0666     1.5424   1.988  0.04716 *
parental.level.of.educationbachelor's degree    5.4318     1.8370   2.957  0.00321 **
parental.level.of.educationmaster's degree      5.9427     2.2428   2.650  0.00823 **
lunchstandard                              10.2103     1.0120  10.089  < 2e-16 ***
test.preparation.coursenone                -5.7036     1.0110  -5.641 2.41e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.15 on 737 degrees of freedom
Multiple R-squared:  0.2337,    Adjusted R-squared:  0.2212
F-statistic: 18.73 on 12 and 737 DF,  p-value: < 2.2e-16
```

To evaluate the model performance, I compare the R-squared of the model using the training data to the R-squared using the test data. In my case (using set.seed to replicate results) the R-squared values were similar at:

Training R2 : 23.36729

Test R2: 29.45597

Therefore the model appears to perform well.

Reading Scores Model

In the case of reading scores, all predictor values were statistically significant. The model was:

$$Reading\ Score = 69 - 7.2 \times (Male) + 4.2 \times (Race.ethnicity = Group\ E)$$

$$+ 3.04 \times (Parent\ edu\ level = associate's\ degree) + 5.2 \times (Parent\ edu\ level = bachelor's\ degree)$$

$$+ 7.5 \times (Parent\ edu\ level = master's) + 6.7 \times (Lunch = standard) - 7.7\ (No\ test\ prep\ course)$$

The adjusted R-squared is only 0.2185 which means that 21.85% of the variation in reading score is explained by the model.

The predictor variable with the largest effect completing a test preparation course. Not having taken a test preparation course reduces a student's reading scores by 7.7. The next most impactful predictor variable is having a parent with a master's degree which increases a student's reading score by 7.5. Gender also has a large impact with males scoring 7.2 lower than females. All three variables are statistically significant with the largest P-value being for lunch at 1.58×10^{-11}.

```
Call:
lm(formula = reading.score ~ gender + race.ethnicity + parental.level.of.education +
    lunch + test.preparation.course, data = traindf)

Residuals:
    Min      1Q  Median      3Q     Max
-40.163  -8.369   0.548   9.533  29.427

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    69.2299     2.0710  33.429  < 2e-16 ***
gendermale                                     -7.2915     0.9338  -7.809 1.99e-14 ***
race.ethnicitygroup B                           1.0925     1.9407   0.563 0.573643
race.ethnicitygroup C                           2.6230     1.7885   1.467 0.142910
race.ethnicitygroup D                           3.2025     1.8292   1.751 0.080399 .
race.ethnicitygroup E                           4.2117     2.0100   2.095 0.036479 *
parental.level.of.educationhigh school         -1.6441     1.5402  -1.067 0.286118
parental.level.of.educationsome college         2.1726     1.4670   1.481 0.139041
parental.level.of.educationassociate's degree   3.0487     1.4919   2.043 0.041359 *
parental.level.of.educationbachelor's degree    5.2406     1.7769   2.949 0.003285 **
parental.level.of.educationmaster's degree      7.5861     2.1694   3.497 0.000499 ***
lunchstandard                                   6.7038     0.9789   6.848 1.58e-11 ***
test.preparation.coursenone                    -7.7500     0.9779  -7.925 8.45e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.72 on 737 degrees of freedom
Multiple R-squared:  0.231,     Adjusted R-squared:  0.2185
F-statistic: 18.45 on 12 and 737 DF,  p-value: < 2.2e-16
```

To evaluate the model performance, I compare the R-squared of the model using the training data to the R-squared using the test data. In my case (using set.seed to replicate results) the R-squared values were similar at:

Training R2 : 23.09751

Test R2: 20.13765

Therefore the model appears to perform well.

Writing Scores Model

In the case of reading scores, all predictor values were statistically significant. The model was:

$$Writing\ Score = 68 - 9.1 \times (Male) + 4.9 \times (Race.ethnicity = Group\ D)$$

$$+ 4.01 \times (Parent\ edu\ level = some\ college) + 4.5 \times (Parent\ edu\ level = associate's\ degree)$$

$$+ 7.9 \times (Parent\ edu\ level = bachelor's\ degree) + 10.1 \times (Parent\ edu\ level = master's\ degree)$$

$$+ 7.7 \times (Lunch = standard) - 10.2\ (No\ test\ prep\ course)$$

The adjusted R-squared is 0.3208 which means that 32.08% of the variation in writing score is explained by the model.

For writing scores, the predictor variable with the largest effect completing a test preparation course. Not having taken a test preparation course reduces a student's writing scores by 10.2. The next most impactful predictor variable is having a parent with a master's degree which increases a student's reading score by 10.1. Gender is also has a large impact with males scoring 9.1 lower than females. All three variables are statistically significant with the largest P-value being for having a parent with a master's degree at 1.93×10^{-6}.

For writing scores, a parent's education level is a statistically significant predictor for many levels of education. A student's writing score will be higher if a parent has some college, associate's degree, bachelor's degree, or master's degree. Being from an average to above-average income family (as indicated by receiving a free / reduced cost lunch) also increases writing scores by 7.7 and is statistically significant at 2.02×10^{-15}.

```
Call:
lm(formula = writing.score ~ gender + race.ethnicity + parental.level.of.education +
    lunch + test.preparation.course, data = traindf)

Residuals:
    Min      1Q  Median      3Q     Max
-39.476  -7.588   0.423   8.791  28.048

Coefficients:
                                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 68.2653     2.0186  33.819  < 2e-16 ***
gendermale                                  -9.1385     0.9102 -10.041  < 2e-16 ***
race.ethnicitygroup B                        0.8379     1.8916   0.443  0.65792
race.ethnicitygroup C                        2.6210     1.7432   1.504  0.13313
race.ethnicitygroup D                        4.9000     1.7829   2.748  0.00614 **
race.ethnicitygroup E                        3.7655     1.9591   1.922  0.05499 .
parental.level.of.educationhigh school      -1.3376     1.5012  -0.891  0.37321
parental.level.of.educationsome college      4.0162     1.4299   2.809  0.00511 **
parental.level.of.educationassociate's degree  4.5567     1.4542   3.134  0.00180 **
parental.level.of.educationbachelor's degree   7.9031     1.7319   4.563 5.90e-06 ***
parental.level.of.educationmaster's degree    10.1482     2.1145   4.799 1.93e-06 ***
lunchstandard                                 7.7432     0.9541   8.116 2.02e-15 ***
test.preparation.coursenone                 -10.2013     0.9532 -10.702  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.4 on 737 degrees of freedom
Multiple R-squared:  0.3317,    Adjusted R-squared:  0.3208
F-statistic: 30.48 on 12 and 737 DF,  p-value: < 2.2e-16
```

To evaluate the model performance, I compare the R-squared of the model using the training data to the R-squared using the test data. In my case (using set.seed to replicate results) the R-squared values were similar at:

Training R2 : 33.17003

Test R2: 32.84946

Therefore the model appears to perform well.

SUMMARY

The R-squared values for all three models to predict test scores were low, ranging from 23.09 to 33.17. Still, the models did find statistically significant predictor variables. For future investigation, other variables such as curriculum and grades might be helpful.

Given the available data, there were some informative outputs that could help direct learning interventions to raise students' test scores. This section reviews the initial hypotheses outlined in the beginning of this study and makes recommendations to raise test scores.

Gender

Hypothesis: Gender does not affect test scores.

Outcome: Gender is a significant predictor variable for all three tests. In math, being male predicted an additional 4.8 points to a student's scores. In the verbal subjects of reading and writing, being male predicted a reduction of 7.2 points (reading) to 9.1 (writing) points.

Recommendation: This suggests that assistance in math would be more useful for female students and that assistance for reading and writing should be targeted towards males.

Race / Ethnicity

Hypothesis: Race / ethnicity affect test scores.

Outcome: Being a member of Group E increased student scores in math and reading tests. In math, Group E students scored 8.8 points higher than members of other groups. In reading, Group E students scored 4.2 points higher than members of other groups.

Group D students also had higher scores. In math, they were predicted to score an additional 4.4 points and an additional 4.9 points in writing.

Recommendation: In math, there should be outreach to students in Groups A, B, and C. In reading, there should be outreach to students in Groups A, B, C, and D. In writing, there should be outreach to students in Groups A, B, C, and E. Outreach could include role models from the same groups succeeding in the subjects that are being highlighted.

Parental Level of Education

Hypothesis: The higher the level of parents' education, the higher the student test scores.

Outcome: In math, having a parent with an associate's degree or higher predicted an additional 3 to 5.9 points on the score. In reading, having a parent with an associate's degree or higher predicted an additional 3 to 7.5 points on the score. In writing, having a parent with additional years of education beyond high school predicted an additional 4 to 7.9 points on the score. In general this suggests a significant generational effect of education and how obtaining an associate's degree and beyond supports children's future educational attainment.

Recommendation: Schools with parents with lower percentages of education beyond high school should be prioritized for educational assistance.

Lunch

Hypothesis: Students receiving free / reduced cost lunches will have lower test scores.

Outcome: Receiving free / reduced cost means that a student's family income is low enough to qualify for the food assistance. Students from families with higher incomes were predicted to score higher across all three subjects than students from families with lower incomes. In math, the benefit from being from an average to above-average income household was particularly pronounced with an additional 10.2 points going to students with more financial resources.

In reading and writing, the students who were not receiving food assistance were predicted to score an additional 6.7 to 7.7 points.

In all three subjects, the lunch coefficients were statistically significant at the 0.001 level.

Recommendation: Educational assistance should be directed to lower-income schools as indicated by the percentage of students receiving free / reduced cost lunches. Priority should go to math assistance.

Test Preparation

Hypothesis: Completed test preparation will improve test scores.

Outcome: Not having had preparation was statistically significant at the 0.001 level. None having completed a test preparation course predicted a math test score that was 5.7 points lower than students who did complete a test preparation course. The effect on reading and writing was more pronounced. Not having done a test preparation course predicted a reading score that was 7.7 points lower and a writing score that was 10.2 points lower.

Recommendation: Test preparation courses are helpful across all subjects but particularly for reading and writing. Providing test preparation opportunities can help raise student's scores.