Indicators of student test performance

Andrew Walther

10/19/2020

Introduction

During the formative years of every young student's life, they are often inundated by the task of taking exams to assess knowledge in a variety of subject matter areas. Although few students find taking exams to be an enjoyable tasks, the results are often thought to be indicative of one's aptitude for learning and future performance in academics. On the outside of every exam experience, a multitude of outside factors can influence student performance aside from the raw education received in the classroom. For one, does gender play a significant role in exam performance among different subject areas like mathematics, reading, and writing? Additionally, can we use simple models to project how a student will score on standardized exams based on general demographic information like their ethnicity, parents' education, socioeconomic status, and the amount of preparation they have taken on prior to an exam? These are important questions to ask when we consider how student's perform on exams in order to ensure that their performance is an accurate reflection of their academic aptitude and no inherent bias exists due to outside factors.

At the end of the day, the education world likes to use exam scores as a proxy for academic success as a whole so we should advocate for fair and accurate testing. It is also important to accept that although we may hope a test score is an accurate indicator of future success or failure, past performances can never guarantee future results!

In the following analysis, we'll investigate the relationship between the given demographic factors with student testing performance while also considering possibilities for more robust data collection that could improve future analysis.

Data set

The data used in this analysis was acquired from Kaggle and is the StudentsPerformance data set. This data set includes demographic information for 1000 students along with test scores on math, reading, and writing exams. Some of the demographic information collected includes: gender, race/ethnicity, parental level of education, free/reduced lunch, and completion of test preparation course. The data set can be found here: https://www.kaggle.com/spscientist/students-performance-in-exams.

Specific Terminology

We compute a metric in this analysis called `avg.score` that is simply the average score for each student across their math, reading, and writing exam scores.

Additionally, a data dictionary including all variable names with relavant details is included below:

Variable	Description
gender	male, female
race/ethnicity	denoted as A,B,C,D,E
parental level of education	highest level of education completed by a parent

Variable	Description
lunch	standard, free/reduced
test preparation course	none, completed
math score	score on math exam (0-100)
reading score	score on reading exam (0-100)
writing score	score on writing exam (0-100)

Exploratory Analysis

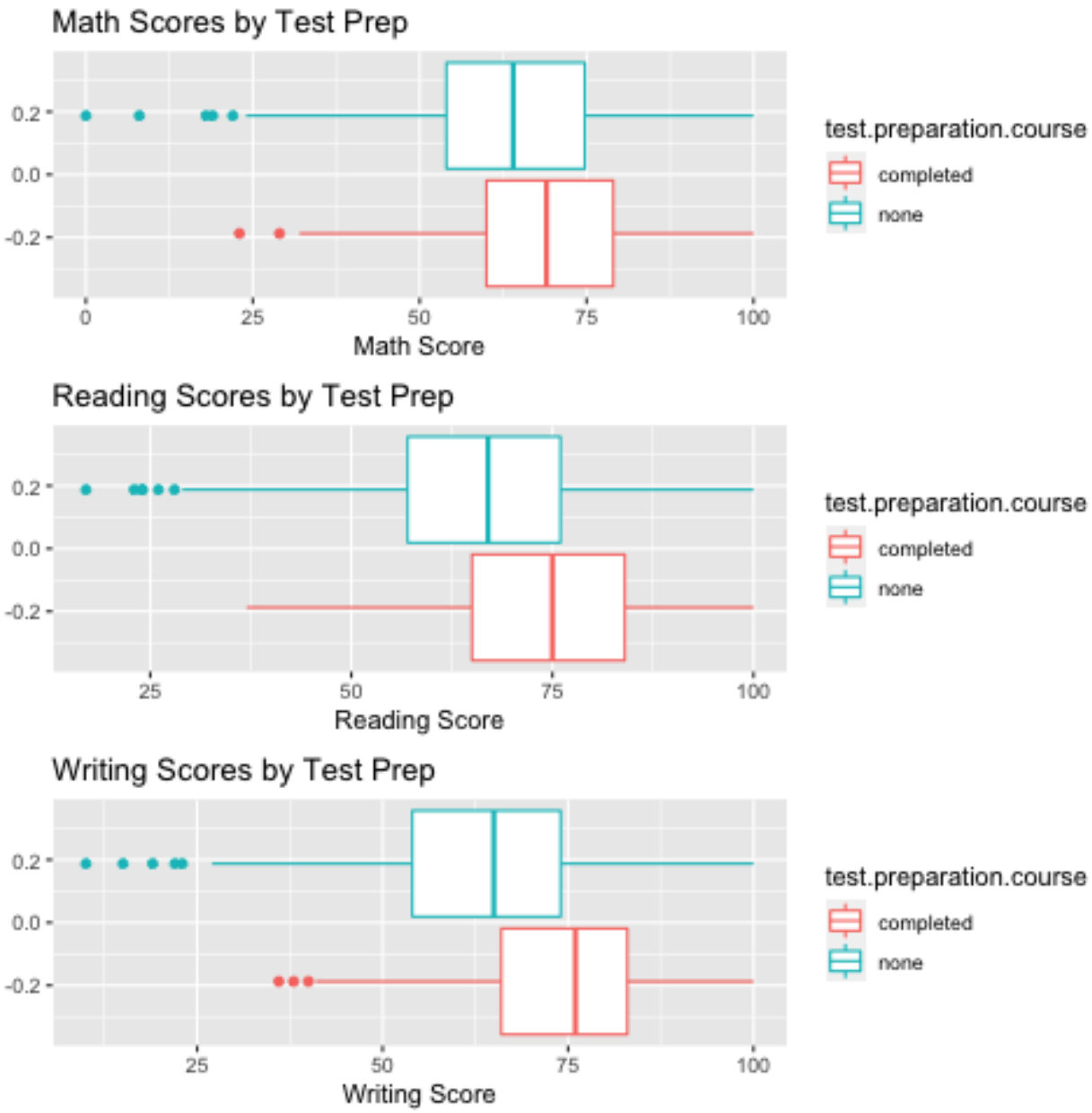
Here, we can see a list of summary statistics from the three exams that students completed in order to get a decent idea of some possible trends in the data.

Statistic	Math Exam (male, female)	Reading Exam (male, female)	Writing Exam (male, female)
Minimum	0 (27, 0)	17 (23, 17)	10 (15, 10)
Q1	57 (59, 54)	59 (56, 63.25)	57.75 (53, 64)
Median	66 (69, 65)	70 (66, 73)	69 (64, 74)
Q3	77 (79, 74)	79 (75, 73)	79 (73.75, 82)
Maximum	100 (100, 100)	100 (100, 100)	100 (100, 100)
Mean	66.09 (68.73, 63.63)	69.17 (65.47, 72.61)	68.05 (63.31, 72.47)
Standard Deviation	15.16 (14.36, 15.49)	14.60 (13.93, 14.38)	15.20 (14.11, 14.84)
Count	1000 (482, 518)	1000 (482, 518)	1000 (482, 518)

Since 1000 students exam results were recorded, the mean and median score for each test ended up being approximately the same and at least one student scored a perfect 100 on each of the three exams. The reading exam had the highest average score, followed by the writing exam, and the math exam had the lowest average score. It is slightly surprising how similar the mean, median, and quartile scores are between the three exams and it raises the question of whether or not the scores were normalized to be on the same scale between each of the three exams even if one of the exams was more challenging in reality.

In the graphic below, it is apparent that male students typically score higher than females on the math test, but females score higher than males on both the reading and writing tests.

In the series of box plots below, it is very clear that students that complete a test prep course score higher than those who do not on the tests for each of the three subjects (math, reading, and writing).



Regression model to distinguish examine scores by gender

Considering the summary statistics for the exam score stratified by gender, it is apparent that there is a significant performance gap between males & females on all three exams with regard to the average score for each gender. On the math exam, males score about 5 points higher than females on average. On the reading exam, females outscore males on average by about 7 points, and females also outsocre males by over 9 points on the writing exam. We can set up a linear regression model to see how well we can predict a male's math exam score with their gender, reading score, and writing scores as predictors. The resulting regression model equation is

$$\text{Math Score } = -6.19 + 0.58 \times \text{ Writing Score } + 0.38 \times \text{ Reading Score } + 13.14 \times \text{ Gender(male)} + \epsilon$$

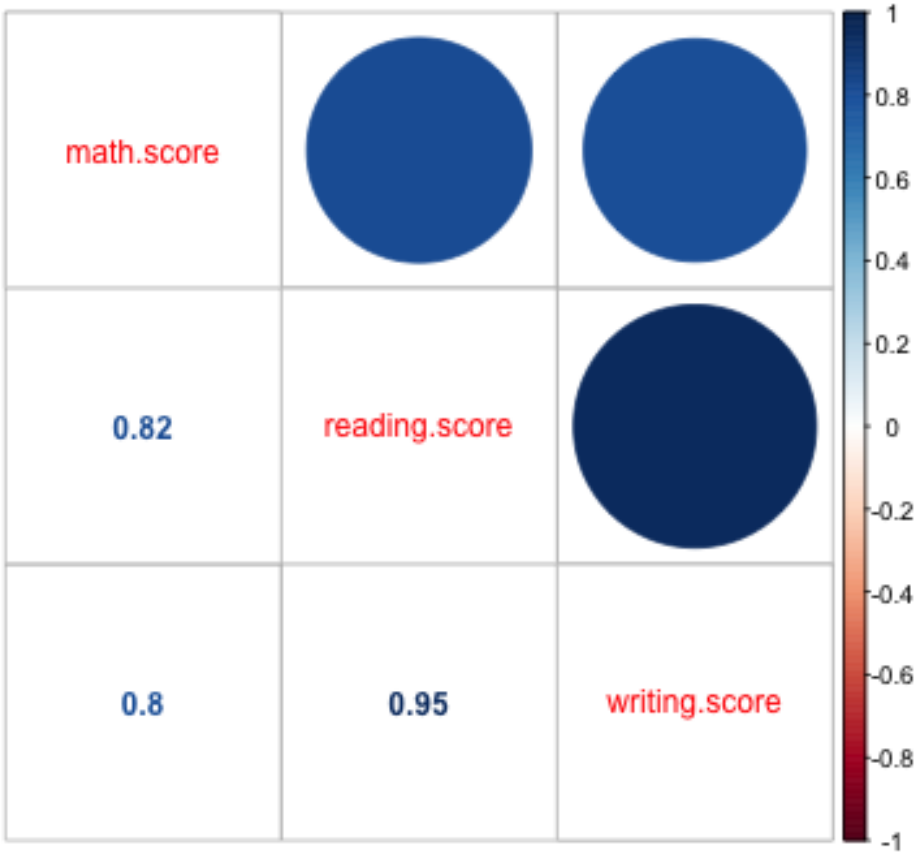
.

In the multiple linear regression model, we find that all explanatory variables (gender ($\beta = 13.138$), reading score ($\beta = 0.382$), and writing score ($\beta = 0.581$)) are significant when predicting the math exam score of a male student. Interestingly enough, we find that between males and females, the model coefficient of being a male increases the predicted math exam score by over 13 points (13.138)! We also note that the multiple regression model is fairly valid since it has an adjusted R-squred value of 0.8401, indicating that the model accounts for about 84% of the variability in the data when predicting math scores. Overall, this model gives a strong indication that on average, males have higher math scores than females, and higher scores in one subject are associated with higher scores in the other two subjects (math, reading, or writing).

Correlation between different tests

Previously, we analyzed a linear regression model that used test score results for reading & writing to predict math scores between males & females. While this model was a strong performer, a major concern arises in how the model used scores that had been recorded to model predictions on other scores that already occurred. Additionally, it may not be very useful to predict scores while using other performance metrics as the predictors. We can see in the correlation plot below (and table), that utilizing scores from one test to predict for another is problematic since the tests are all highly correlated with each other (≥ 0.80). For this reason, it would be more informative to only make predictions on test scores with the prior demographic information that is given for each student.

Correlation	Math	Reading	Writing
Math	1.000	0.818	0.803
Reading	0.818	1.000	0.855
Writing	0.803	0.955	1.000

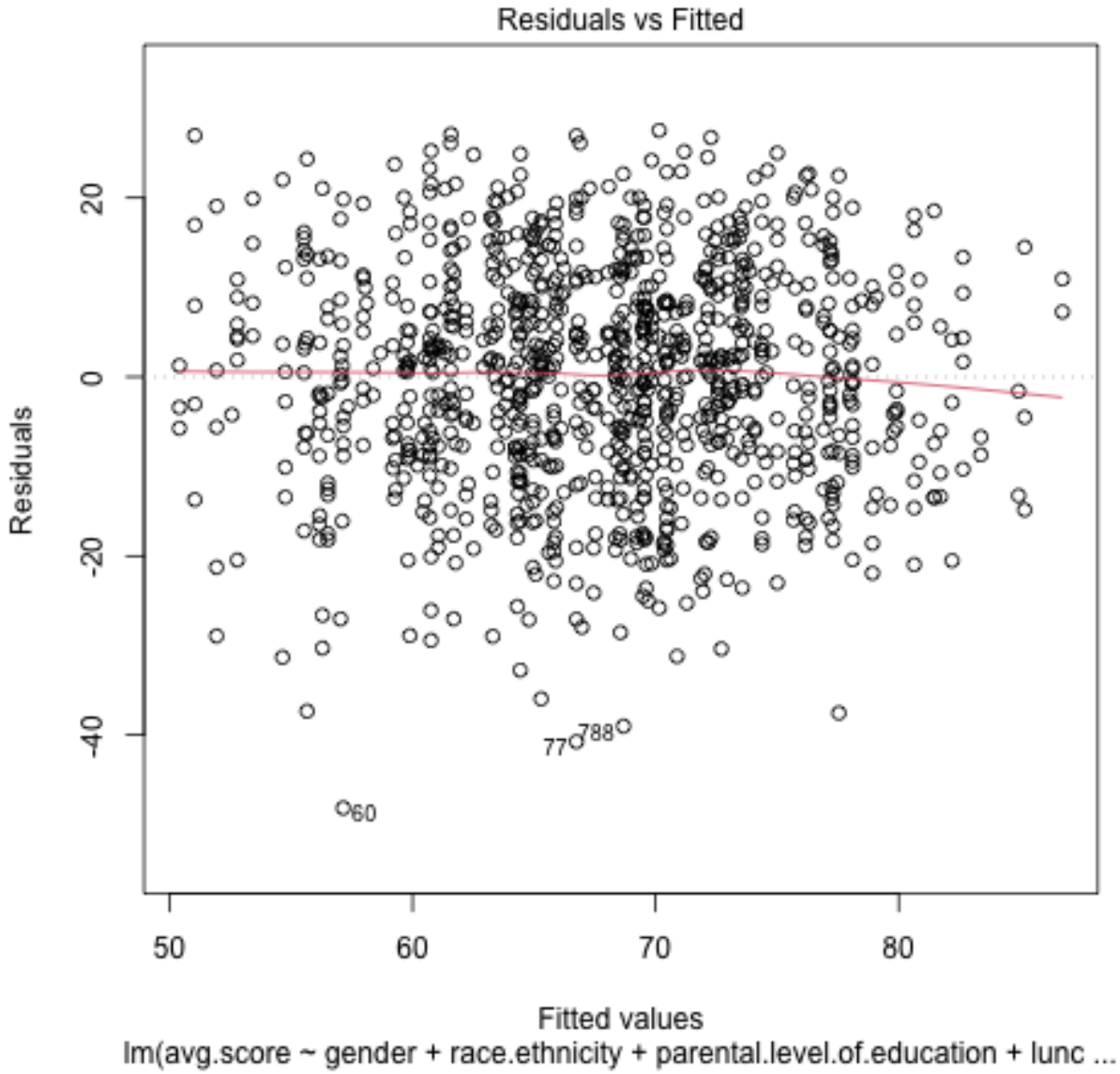


Regression model with demographic predictors

We're really interested in modeling how students score on the exams when the only predictors that are used are the demographic information given along with the three test scores (math, reading, & writing). These demographic factors are the student's gender, their race/ethnicity, the highest level of education that their parents completed, whether or not they have free/reduced lunch at school, and whether or not their completed a test preparation course before taking the exams. By only considering the demographic variables, we're ignoring any potential confounding information from performance on other exams. Additionally, we can just try to model the average score of each student between the three exams combined. The equation for this regression model is:

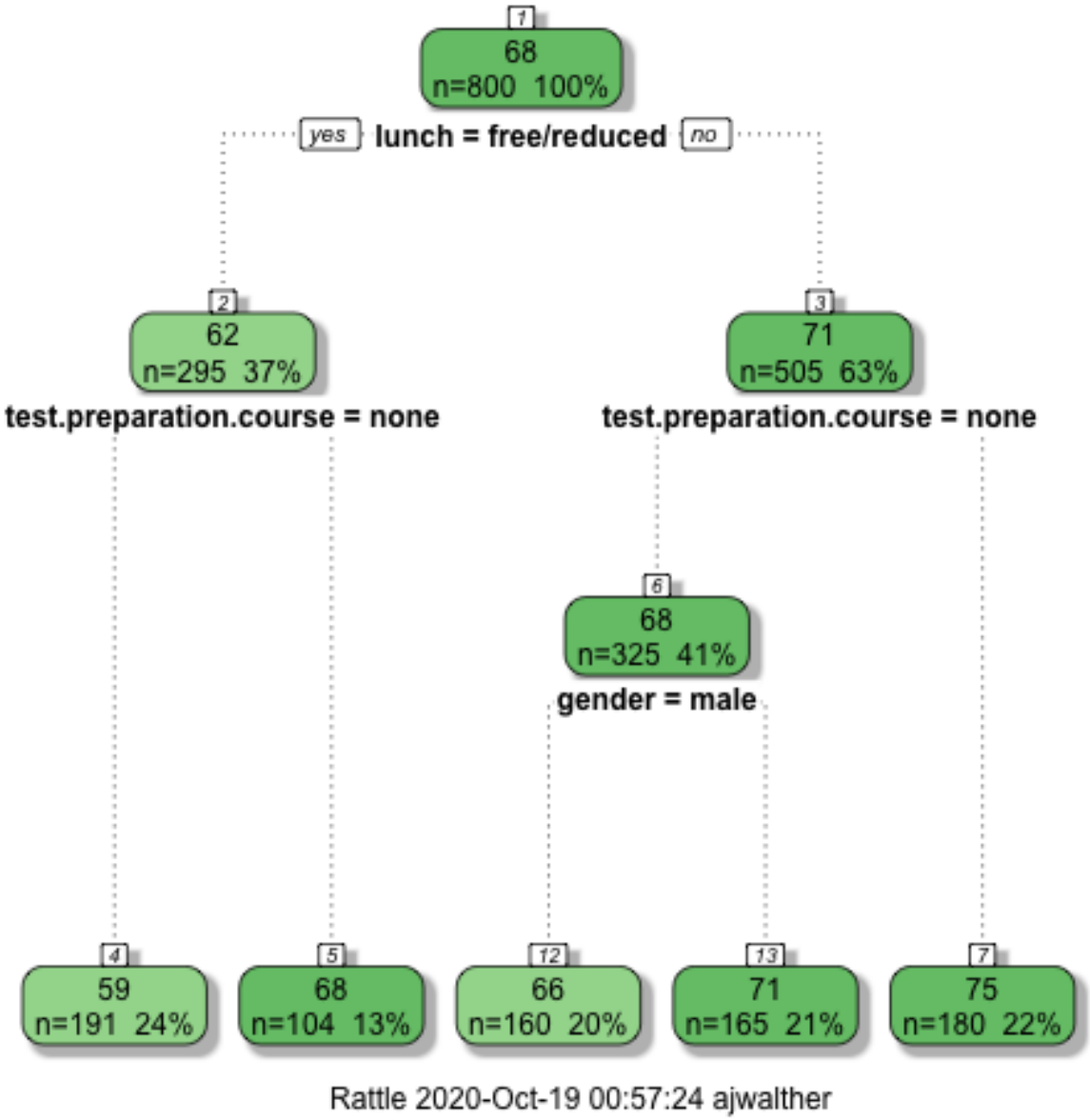
$$\text{Average Score} = 66.94 - 3.72 \cdot \text{Gender(male)} + \beta_2 \cdot \text{Race} + \beta_3 \cdot \text{Education} + 8.78 \cdot \text{lunch(standard)} - 7.64 \cdot \text{no test prep} + \epsilon$$

where β_2 for race/ethnicity is 1.53 for group B, 2.39 for group C, 5.13 for group D, and 6.93 for group E. Additionall, β_3 for education is 2.54 for a bachelor's degree, -5.17 for high school, 4.09 for a master's degree, -0.93 for some college, and -4.54 for some high school. In the linear model output where "average score" is the model response, gender, ethnicity (D), ethnicity (E), education (high school), education (master's degree), education (some high school), lunch, and test prep course are all significant predictors of a student's average score. The intercept for this model is an average score of 66.9 and the model coefficient for no test prep course is -7.6. This coefficient indicates that, based on this regression model, a student who does not complete a test preparation course is predicted to score over 7 points fewer than a student who does complete a test prep course. This is an interesting takeaway from a simple regression model and it may provide evidence to suggest that students should invest their time and resources into taking a course to prepare before taking importnat exams. Unfortunately, we might not be able to trust this model too much since the adjusted R-squared value of 0.233 means that only about 23% of the variation in the response (average score) data is accounted for by the model. Interestingly enough, if we check out the residual plot for this regression model, we find that the model is actually in agreement with the assumptions required of linear regression since the residuals are normally distributed and show no signs of heteroscedasticity. Finally, when we make predictions on data set aside for testing, we get a Mean Squared Error on the predictions of 258.59.



Decision tree to predict average exam score from demographic factors

Another option to predict a student's average test score based on all of the demographic factors provided is with a decision tree. In the decision tree below, we can see that the students are first separated based on if they qualify for free/reduced lunch, then by if they took a test preparation course, and finally by gender. We can see that students with the lowest predicted scores (59) qualify for free/reduced lunch and did not take a test prep course. Both of these factors can serve as a proxy for socioeconomic status where we see that students who are less well off are not able to afford lunch and also do not take test prep courses. This is associated with lower scores on their exams. On the other hand, we see that students who do not qualify for free/reduced lunch and who did take a test prep course are predicted to have the highest average scores at 75 points. The predictions for this decision tree model returned a mean squared error value of 229.75. Compared to the multiple regression model with demographic predictors, the decision tree had a slightly smaller MSE from the model predictions.



Conclusions

In this analysis of the student test performance dataset, we set out to investigate if there is an association between gender and test performance and whether or not we could make viable predictions of exam scores for students based on some provided background demographic information. We first found that when we consider the math exam, for a given pair of reading & writing scores, male students are predicted to score about 13 points higher on the math test than female students. On the other hand, for a given math score, when we compare males and females on the reading and writing tests, female students score significantly higher than male students. These conclusions were made based off of the results of a multiple linear regression model.

Our primary interest was if we can predict scores based off demographic data that is provided for each student that took the exams. We built another multiple linear regression model and a decision tree model to train and make predictions on where we "threw the kitchen sink" of the demographic variables in the dataset at

the models. In the regression model, all of the demographic predictors were significant toward predicting the response of average score, but only lunch, test prep, and gender were significant for the decision tree prediction model. An interesting takeaway from both models is that students who have free/reduced lunch and who do not complete a test prep course are predicted to score significantly lower than students who have standard lunch and who do take a test prep course. Additionally, students whose parents do not have a bachelor's degree or higher are also predicted to score lower on average than students whose parents have a college and/or graduate degree. These factors all could be considered proxies for socioeconomic status, so we're really seeing that students with lower socioeconomic status, are on average projected to have lower test scores. This obviously raises an alarm to make sure that underprivileged students aren't left behind in their academic pursuits.

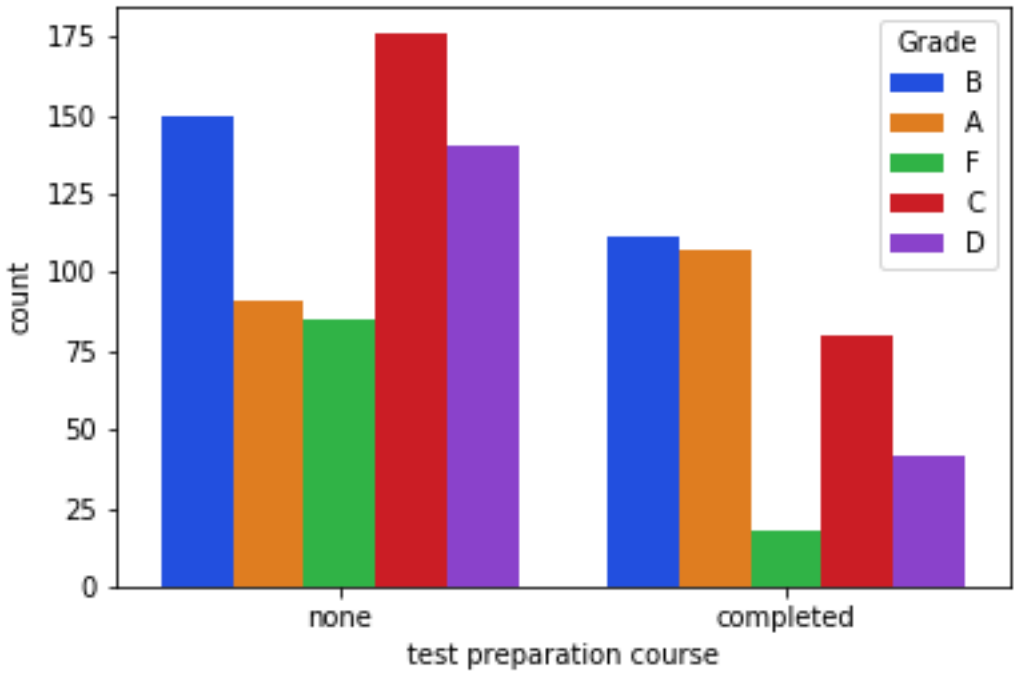
Future Ideas

This analysis ended up providing some interesting insights on the exam score dataset. Its interesting how there is such a wide gap in average performance on the three exams, where males perform better on math and females are much better performers on writing and reading. Additionally, regression & decision tree models are able to predict a student's average score based on demographic data. We found that the 3 exam scores reported for each student were strongly correlated with each other, essentially relegating us to a single numeric response in the average score. Going forward, it would be interesting to take advantage of a much large set of data, like demographic & performance metrics for students who take the ACT & SAT each year. Some additional factors that could make further analysis interesting are: student age, household income, hometown or state of residence, hours spent preparing for a particular exam, and coursework grade point average, among others.

Given more information about a student's background and potential aptitude for an exam, it would be interesting to work on building a prediction model using variables like parental education, household income, GPA in coursework, and hours spent studying to determine if these factors can be used as valid predictors of a student's outcome on an exam. This could be a simple as a multiple regression model or as complex as a deep learning neural network, depending on how clear of a pattern there might be in the data. Additionally, if data from the SAT or ACT exams is considered, it would be very interesting to record data regarding a student's future in higher education like: institution attended, rank of institution attended, selected academic major, undergraduate GPA, and grades across different disciplines that can be related back to scoring in various portions of the SAT/ACT exam to determine if a test score is a good benchmark for past performance (in high school) and a strong indicator of future aptitude for success in university coursework. For example, does a high math score on the ACT translate into a student electing to be a math or statistics major and have success in that particular area of coursework?

Ideas addressed with python

In addition to the future directions mentioned above, I was also curious about what the distribution of average scores would look like separated by if a student completed a test preparation course if we separated scores into typical letter grade bins such as A(80-100), B(70-80), C(60-70), D(50-60), & F(<50). The figure below illustrates the distribution of math scores from the given students under these conditions and we can see that it appears students who completed a test preparation course tended to get higher overall grades. In green, we see that very few of the students who completed a preparation course ended up with an 'F' grade and most had either an A or B. On the otherhand, the predominant grade for students who did not take a test prep course was C, followed by B and D. In the group where students didn't take a test prep course, the number of students who got an A was about the same as the number who got an F. On the other hand, in the "completed" group, about 4 times as many students got an A than students who got an F grade.


Grade
B
A
F
C
D
175
150
125
100
75
50
25
0
count
none
completed
test preparation course