Sarah Deussing & Emily Kirk

ACMS 40210: Scientific Programming

Dr. Adam Volk

April 30, 2024

A Statistical Exploration of Predictors Affecting Students' Math Scores

Introduction

Because we are both mathematics majors, we focused our dataset search on mathematics-related topics. The dataset we chose displays math, reading, and writing standardized scores and several environmental, parental, and social factors. In selecting this dataset, we hoped to determine the factor(s) that most contributed to a higher mathematics score.

The columns of our dataset were: gender, ethnic group, parent education, lunch type, test prep, parent marital status, practices a sport, is a first child, number of siblings, transport means to school, and weekly study hours - as well as reading, writing, and math scores. These factors are interesting because many are not outwardly connected to a student's math performance, but we hope to determine any that are important. Because previous class work analyzed the relationship between grades studying and extracurriculars, the variables of test prep, whether or not a child practices a sport, and weekly study hours were not examined. The data set also did not define the different ethnic groups, so that variable was also not examined.

This data is important for those in academia and many other people because it can help identify students who may need additional help when preparing for mathematics exams. Instead of identifying students after they perform badly on an exam, our conclusions can help identify these students much earlier. The dataset can be accessed through the following link: Math Scores Data Set from Kaggle.

The dataset was created by Kaggle user Royce Kimmons for purely educational purposes. Kimmons specified that the data used in our analysis is fictional data for students at an imaginary public school. He explained that he created his data based on other datasets featured on Kaggle, but that he expanded the number of columns (variables) and rows (entries). He also purposefully added missing values to allow users to practice data cleaning. We assume that his methodology in creating the data is reasonable; however, he did not elaborate on further details as to how exactly this data was created. As a result, we cannot take our final analysis to be generalizable to the real world as discussed later in our conclusion section.

Data Cleaning

Before data analysis could be completed, data cleaning was done to remove null values and format text for readability. The column 'Unnamed 0' was removed from the dataset, as it was a duplicate of the index. This process was followed by the formatting of variables. For standardization, the options for 'WklyStudyHours' were changed from "<5", "5-10", and ">10" to "0-5", "5-10", and "10+". Next the word "group" was removed from the 'EthnicGroup' column, and "degree" was removed from the 'ParentEduc' variable, as these words were unnecessary and hindered readability. In addition, the 'TransportMeans' variable was changed from "school_bus" and "private" to "bus" and "car" respectively, allowing for a clearer distinction between the two transportation options.

Next, null values were removed from the data set. Since all only children are the first children in their family, NaN values of the 'IsFirstChild' column were filled with 'yes' if the corresponding value for the 'NrSiblings' column was 0. This resulted in filling 151 values of the 'IsFirstChild' variable. Next, we calculated conditional probabilities to determine if any strong

relationships existed within the data, specifically between lunch type, parent education, and parent marital status. Calculations and probabilities can be found in the Python code. Because these conditional probabilities were lower than expected, there did not appear to be a relationship between any of these variables that would allow us to fill null values. With such a large dataset and no missing values in math, reading, or writing scores of students, all rows with null values were dropped from the dataset.

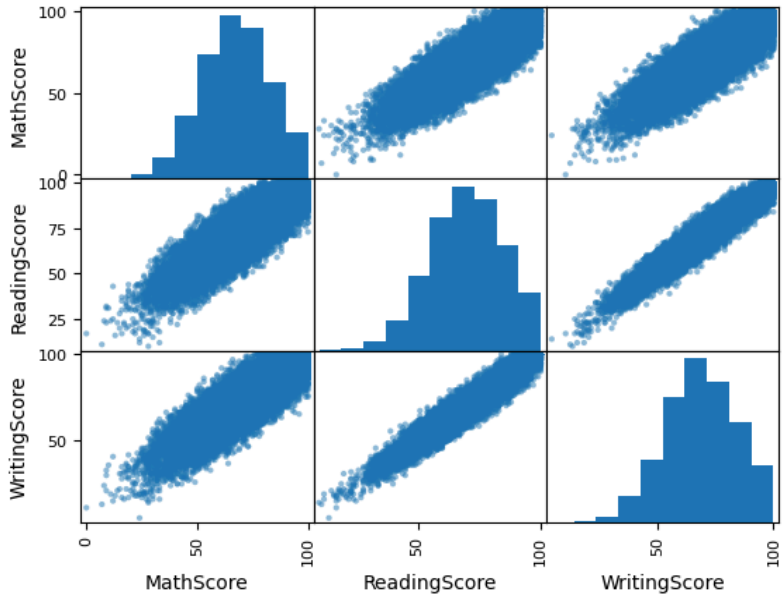
Data Analysis

Relationship between reading writing and math scores

Our analysis began with understanding the correlation between math, reading, and writing scores. We found that all were highly correlated, with reading and writing being the most correlated of the three scores. To provide a more mathematical conclusion, we performed a correlation calculation for these variables which is located in the adjacent table. This was followed by a visual representation of the relationship between the three variables, as seen in the figure below. The histograms of all three variables showed an approximately normal distribution that was slightly skewed toward higher scores. The scatterplot of math scores with reading and writing scores reveal a strong positive linear relationship between the variables, so increases in math scores resulted in increased reading and

Correlation Matrix			
	Math Score	Reading Score	Writing Score
Math Score	1.00	-	-
Reading Score	0.819	1.00	-
Writing Score	0.809	0.953	1.00

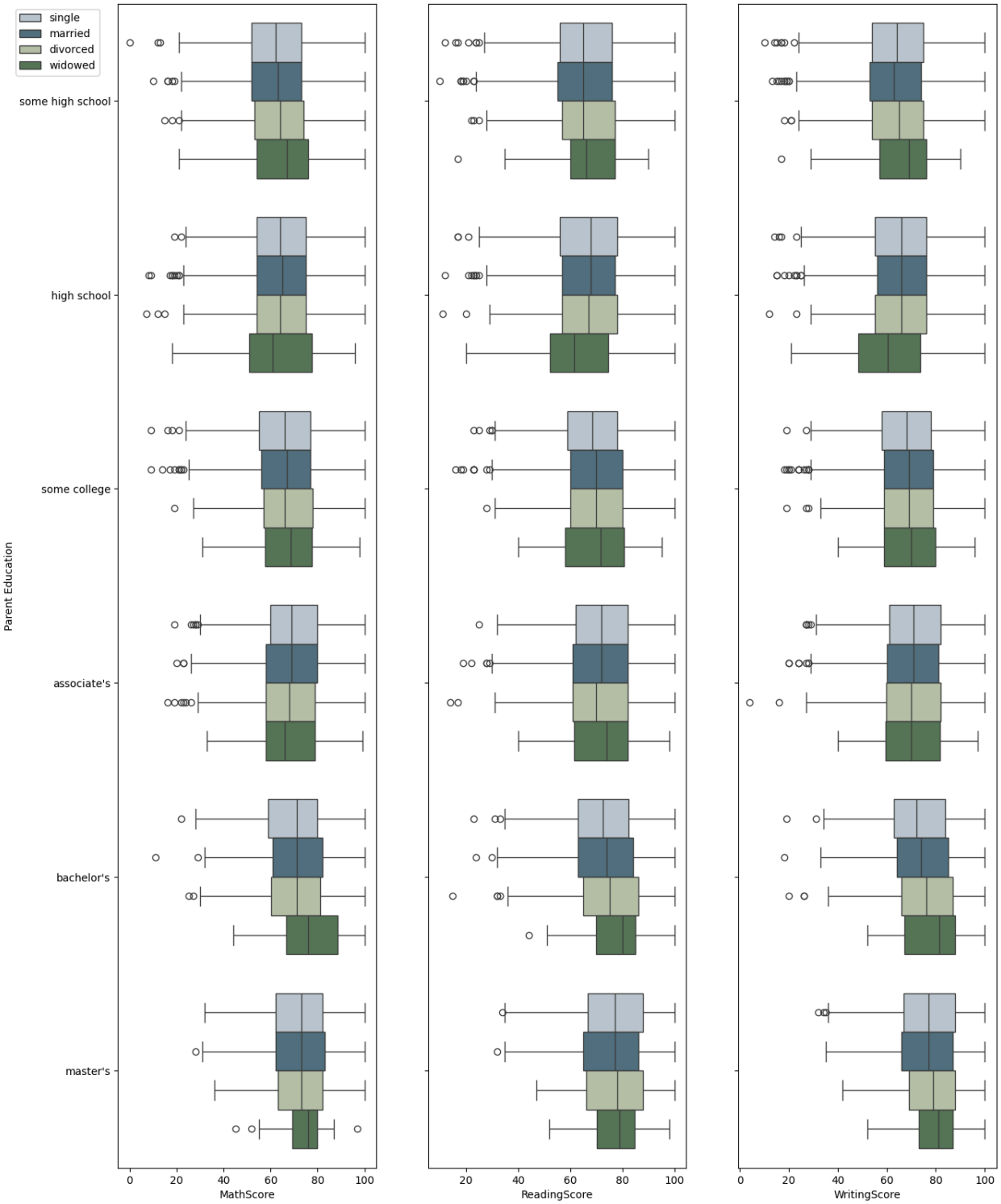
writing scores. The scatterplot of reading and writing scores reveals an even stronger linear positive relationship compared to that of the math scores, which is confirmed in the correlation matrix above.



Relationship between reading writing and math scores by parent education and marital status

We then began to analyze some factors that we hypothesized may influence exam scores. We first focused on parent information - marital status and the highest level of education received. To do so, we graphed side-by-side boxplots of math, reading, and writing scores by parent education level, with different boxplots for marital status. The figure shown below revealed that students whose parents had more education had generally higher scores, however, further data analysis was needed to see the importance of these factors. A groupby table, shown on page 6, was used to determine the average math, reading, and writing score for each parent education group. The lowest average scores were students whose parents had the lowest education level of some high school, while the highest average scores were students whose parents had a master's degree, the highest educational level. This same conclusion was also revealed in the side-by-side boxplot below.

Side-by-side Boxplots of Math, Reading, and Writing Scores by Education Level and Marriage Status



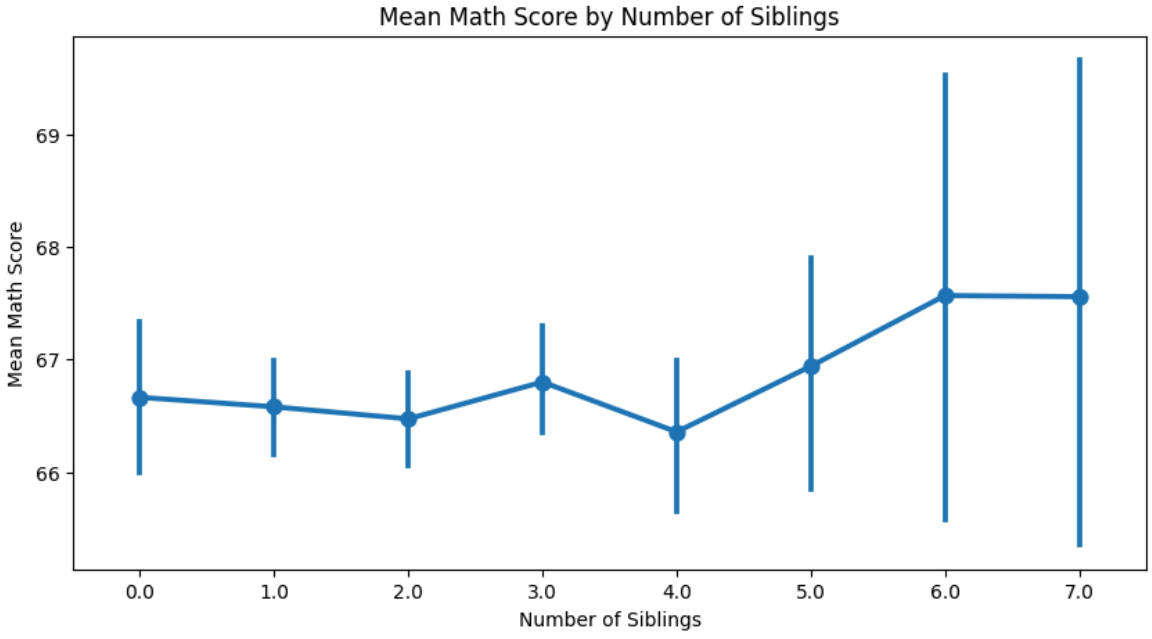
Parent Education	Math Score	Reading Score	Writing Score
Master's	72.322	75.964	76.465
Bachelor's	70.728	73.552	73.860
Associate's	68.552	71.315	70.553
Some College	66.535	69.254	68.609
High School	64.278	67.269	65.484
Some High School	62.519	66.420	63.563

The side-by-side boxplot reveals differing distributions of math, reading, and writing scores by parent's marital status. By comparing within parent education, the median scores across differing marital status remain similar, however the shape of the distribution changes, particularly within the lowest quartile of scores. In order to further examine the relationship between parent's marital status and student's scores, the groupby() function was used to determine the average math, reading, and writing score for each parent marital status. Shown in the table below, a comparison of mean scores across parent marital status reveals that average scores do not vary by a large amount. However, this difference may or may not be significant. Because of the differing distributions shown in the side-by-side boxplot, parent marital status was still included as a possible predictor, as the relationship between math scores and parent marital status, particularly when other variables are included such as parent education, remains unclear.

Parent Marital Status	Math Score	Reading Score	Writing Score
Divorced	66.712	69.803	68.947
Married	66.700	69.460	68.477
Single	66.256	69.351	68.506
Widowed	68.108	70.701	69.679

Relationship between math scores by number of siblings

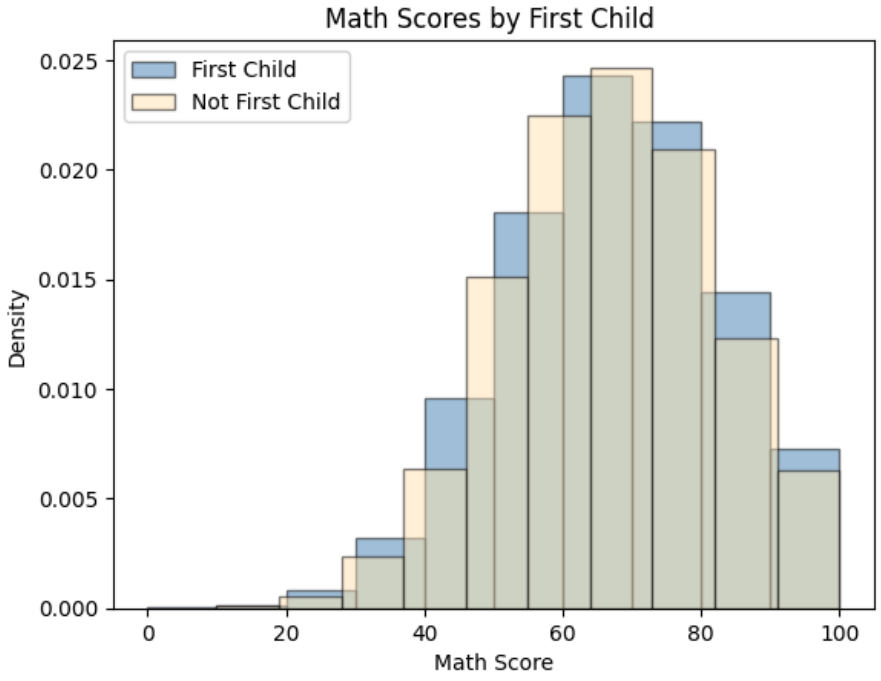
Using a point plot, the relationship between mean math scores and the number of siblings was analyzed. Because the goal of this project was specifically to analyze predictors of math scores, the associations between number of siblings and both reading and writing scores were not examined. The figure below depicts the mean math scores and confidence intervals for students with zero to seven siblings. With similar means and overlapping confidence intervals, there is not sufficient evidence to suggest a relationship between math scores and the number of siblings. This is confirmed in an examination of means for students with different numbers of siblings as seen in the table. Further analysis, therefore, did not include the number of siblings as a predictor.



Number of Siblings	0	1	2	3	4	5	6	7
Mean Math Scores	66.664	66.579	66.473	66.800	66.357	66.940	67.568	67.558

Relationship between math scores by first child

Although the number of siblings a student had did not appear to be significantly related to math scores, we next analyzed the effect of whether or not a student was the first child in their family on math scores. The histogram below displays the distribution of math scores for students who are the first child in their family and students who are not the first child in their family. The distribution of students who are the first child in their family is approximately normal with a slight skew toward higher scores and most students getting a score between 60 and 80. Similarly, the distribution of students who are not the first child in their family is also approximately normally distributed with a slight skew toward higher scores and most students receiving a score



between 55 and 85. The mean math score of students who were the first child was 66.709, while students who were not the first child had a mean score of 66.464. With similar distributions, ranges, and means of math scores, it was determined that 'IsFirstChild' did not have a significant relationship with math scores and was, therefore, not included as a predictor in models.

Relationship between reading writing and math scores by lunch type

Reading, writing, and math scores were then analyzed by a student's lunch type. Lunch type was classified as either 'standard' or 'free/reduced' lunch. The side-by-side violin plot shows all three scores by type of lunch. This plot reveals a similar distribution of scores by lunch type across the three scores of reading, writing, and math. In addition, reading, writing, and math scores are higher for students with standard lunch compared to students with free or reduced lunch.

To better understand the relationship between scores and lunch type, the mean scores of students for both types of lunch were calculated. These scores are shown in the table below. This table reveals a higher mean score for math, reading, and writing for students with standard lunch compared to free or reduced lunch. As a result of the analysis, 'LunchType' was included as a possible predictor of math scores.

	Mean Score		
Lunch Type	Math Score	Reading Score	Writing Score
Free/Reduced	58.844	64.249	62.732
Standard	70.847	72.375	71.765

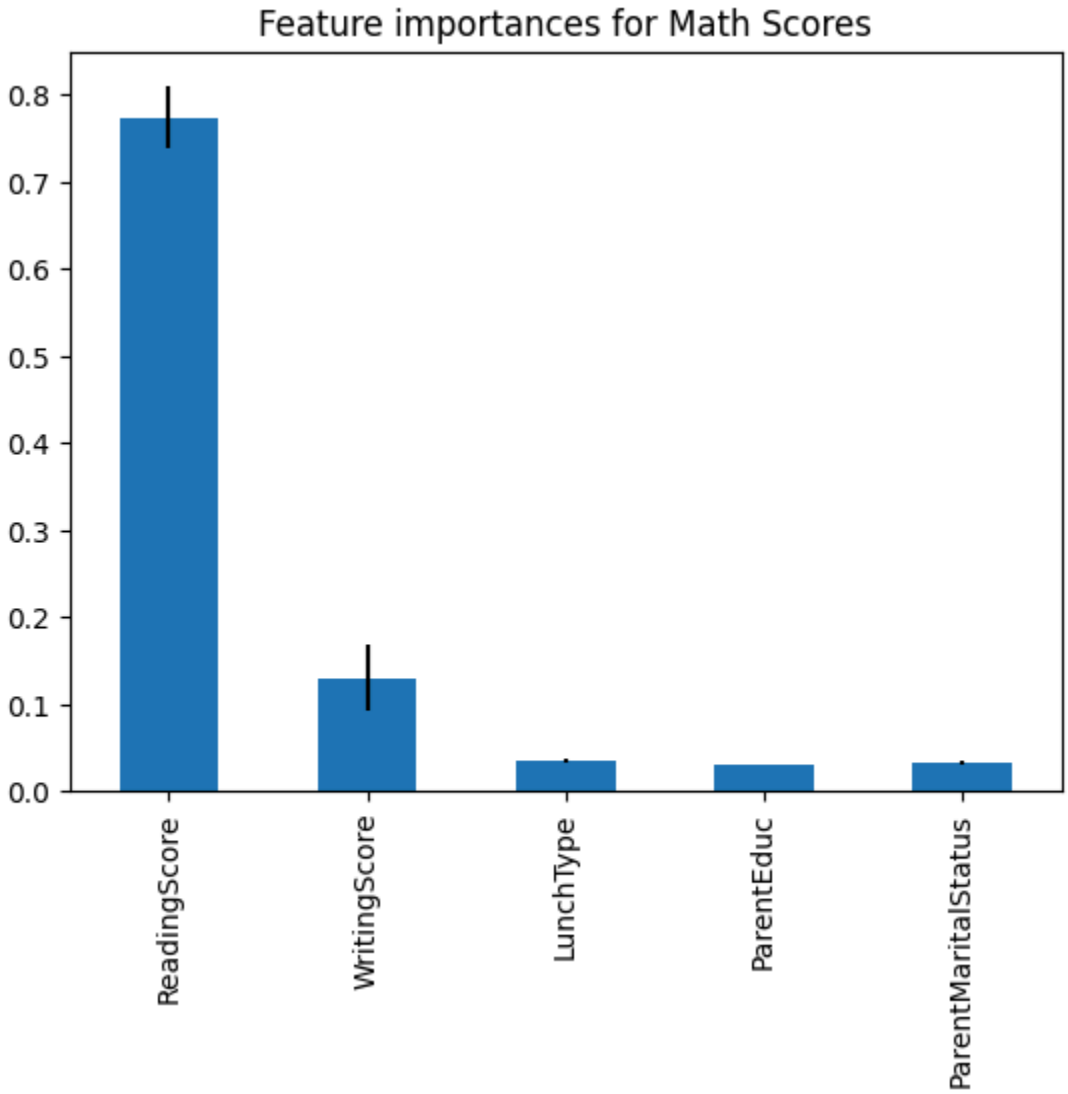
We hypothesize that this relationship may be the result of other factors. Parent income is not included in the dataset, but receiving free/reduced lunch may indicate that a student's parents cannot afford the standard lunch; therefore, we may hypothesize that these families may not be able to afford other educational benefits to their children, such as tutoring. In turn, this may result in lower scores, and thus the association shown above between lunch type and scores.

Machine Learning Model

Model 1 with reading and writing scores as predictors

The goal of our machine learning model was to predict mathematics scores based on the variables we had analyzed above. We found that the number of siblings and whether a student was a first child did not have a significant relationship to math scores, but reading scores, writing scores, parent education, parent marital status, and lunch type did (via our data analysis above). We chose to implement a random forest regression model because compared to a decision tree regressor, random forest regressors use multiple decision trees, making the model robust to differences in individual data and outliers. The data was split using the split_test_train() function in the sklearn.model_selection package in Python. The regressor model was then created and used to predict student math scores using the functions RandomForestRegressor() and fit() from

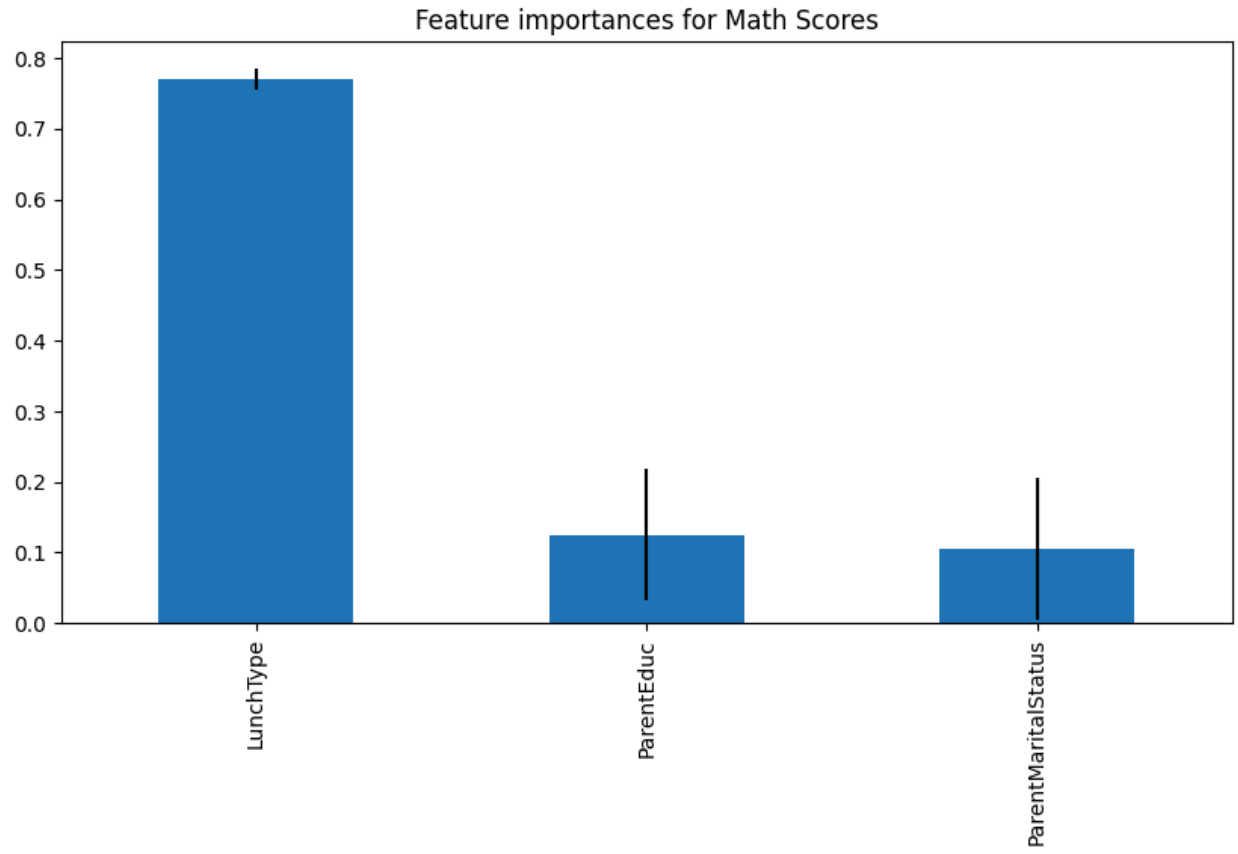
the sklearn.ensemble package. Calculated using the function mean_absolute_error() from the sklearn.metrics package, the mean absolute error from this model was 8.029. After creating our model, we calculated and graphed the feature importance using model.feature_importances_, shown in the figure below, and found that reading scores heavily dominated the model. This feature importance function gives a numeric value to how important a variable was in predicting a dependent variable in a machine learning model. With a value between zero and one, a higher feature importance implies a larger influence in predicting the dependent variable. The figure reveals a higher feature importance for reading and writing scores, with confidence intervals that



do not overlap with the other features. As a result, we decided to eliminate reading and writing scores as predictors and create another model to better understand how lunch type, parent education, and parent marital status affect math scores.

Model 2 without reading and writing scores as predictors

An additional machine learning model was created and analyzed to determine the relationship between math scores and other predictors including parent education, parent marital status, and lunch type. This model was created because of the strong correlation between math, reading, and writing scores which resulted in very high feature importance of writing and reading scores in predicting math scores. In order to analyze variables without the dominance of reading scores and writing scores, a separate machine learning model was created using a random forest regressor with a dependent variable of math score and independent variables of the predictors mentioned previously, which was created and fit using the same process as described above. The mean absolute error of the model was 11.369. Although this error was higher than that of the



previous model, it is still relatively low, as math scores were measured out of 100. The feature importance, as shown in the figure above reveals lunch type as an important predictor of math scores. Parent education and parent marital status had similar feature importance with overlapping confidence intervals, but they were much less than that of lunch type.

Conclusion

In this project, we examined a dataset that contained math, reading, and writing scores as well as various possible related environmental, parental, and social factors for over 30,000 students. Before the data could be analyzed, various techniques were used to clean the data in order to remove null values, ensure readability, and provide more clarity in the meaning of variables. This was followed by the analysis of the relationship between math, reading, and writing scores with lunch type, parent education, parent marital status, number of siblings, and whether or not the student was the first child in the family. Analysis using various visualizations revealed that math, reading, and writing scores were very strongly positively correlated with each other. In addition, scores differed by parent marital status, parent education, and lunch type. The number of siblings and whether or not the student was the first child in the family did not appear to have a significant relationship with math scores. This analysis was followed by the creation of machine learning models using random forest regression in order to predict math scores using writing scores, reading scores, lunch type, parent marital status, and parent education as independent variables. This revealed reading scores as a dominant predictor in the model, while an additional model without writing and reading scores revealed lunch type as an important predictor of math scores.

Therefore, the major predictors of math scores included reading and writing scores followed by lunch type. This information can be utilized by schools in order to improve student's overall performance and understanding of mathematics. By targeting improvement of not only math scores but also reading and writing scores, overall scores of all three subjects will increase, as a result of the strong positive correlation between reading, writing, and math scores. In addition, our analysis showed that students with free or reduced lunch perform worse on the given math assessment. By targeting these students and giving them more resources for math understanding as well as overall academic success, scores for this group of students can increase. It is important to note that this relationship may not be the result of actual lunch type, but rather the parent's socioeconomic status which would qualify a student for free or reduced lunch. Future research could examine the relationship between math scores and parents' socioeconomic status.

While our analysis gives a better understanding of predictors of math scores, there are limitations in our analysis that should be discussed. This data was specifically created for the use of data exploration, so the findings are not generalizable to real-world applications; however, this analysis gives insight into future directions that could be explored such as examining lunch types or parent socioeconomic status on students' academic achievement. There are also factors that could affect math scores that were not included in this data set and not explored. Because the data is fabricated, certain variables could also not be explored like race and ethnicity as the factors of the variable were not given a description other than a letter from 'A' through 'E.' In addition, conditional probabilities between categorical variables were lower than expected, resulting in the removal of many null values. Although this project gives a good example of analyses that could be conducted in order to determine predictors of math scores and potential

predictors that may influence math scores, analysis must be conducted on data from actual students to maintain a generalizable conclusion that could be used by schools and educators.