Arun Agarwal

Principles of Data Science, CIS 3715

April 25th, 2021

Final Report

Deep Analysis and Prediction of the Effects of Various Factors on the Performance of Students on Examinations

**Introduction:**

The future of every society, great and struggling, is in the hands of its young people. These future leaders and their store of knowledge, their abilities to reason, their worldviews and skills will determine the improvements and problems the rest of the century will bring. Education has been the path to better opportunity for generations of American strivers. By 1970, America had the world’s leading educational system, and until 1990, the gap between minority and white students was significantly closing. However, education gains in this country have plateaued since then, and educational inequality has become one of America’s most vexing problems.

According to Economist Rolan Fryer, in 2016, only 44 percent of American students are proficient in reading and math. Worse, the proficiency of African American students, many of them in underperforming schools, is even lower. And in a study by the Education Commission of the States, black and Hispanic students in kindergarten through 12th grade performed on par with the white students who sit in the lowest quartile of achievement. This gap in education starts as early as kindergarten, where black children are eight months behind their white peers in learning. As each year passes, the gap widens.

The issue extends beyond race, as other key factors causing this inequality—identified by numerous reputable agencies like the United States Department of Education—include poverty rates (which are three times higher for blacks than for whites), parental level of education, gender, diminished teacher and school quality, unsettled neighborhoods, ineffective parenting, personal trauma, and peer group influence, which only strengthens as children grow older.

Around the world, 59 million children of primary school age are being denied an education, and nearly 65 million for secondary school. Yet, even for students with the privilege of having access to a public education system with billions of tax dollars invested into it, many students still come out with poor performance scores across all subjects. This poor performance leads to a competitive atmosphere where students take drastic and, sometimes, counterintuitive measures (cheating) to fight for every percentage point in hopes on increasing their chances for a successful future. However, an understanding of the various causes of poor test scores can benefit both the learning of the student and teaching style of the educator.

The purpose of an examination is to understand the ability and learning of a student; thus, through careful analysis of patterns between parental level of education, gender, race, reading scores, writing scores, math scores, and other factors from the Royce Kimmons Kaggle dataset named “StudentPerformance.csv, one can understand which aspects have the largest impact on test outcomes as well as what are the best ways to improve student scores for the future. Furthermore, the improvement of test scores would benefit the economy at a national level by creating equality, leading to major GDP growth, and building a modern society. Correlated with improved test scores is a stronger education system, which for the learner, creates more employment opportunities, secures a higher income, develops problem-solving skills, provides a prosperous and happy life, and educates them to give back to the community.

Prediction of student academic performance in mathematics, reading, and writing based on various demographic and socioeconomic statistics can be performed through creation of various data science models such as linear regression, logistic regression, and k-NN. While the dataset includes this information for 1000 students, not all useful pieces of data, such as hours spent studying, are provided, so I could only perform analysis on the effects these features have on performance. A model with good RMSE, MSE and MSAE scores signifies the model predicts student’s performance well, making the data useful in identifying methods to improve student performance for the future. For each model, I used different hyperparameters and preprocessing techniques to optimize performance. Various types of graphs and plots will also be utilized for exploratory data analysis. The early detection of students who are vulnerable to suffering academic failure (through use of these models) can subsequently be used to design new teaching/mentoring strategies for an overall strengthening education system and society.

Motivation:

Educational inequality has always been a large passion of mine due to the poor socioeconomic and demographic background I was brought up in that led me to learn in an impoverished education system. Furthermore, while I have always been able to push harder to reach higher levels of success in life, I realize that not all students can obtain such milestones due to the various external factors that prevent them from obtaining educational equality. After doing thorough research into developing a reimagined approach to education to better support educational equality, I have been motivated to utilize the skills I have gained in the data science sector to build models that would help one to understand which factors are the true causes of poor performance in school. After finding a Kaggle dataset, provided by Royce Kimmons, with student data on gender, race/ethnicity, parental level of education, and test scores, I have decided that I could perform analysis myself with the dataset to determine which factors have major impacts on student performance. It is my hope that others with more power over begetting change in the education system will examine my models and results to work to rectify the system.

Related Project:

[Regression Models for Predicting Student Academic Performance in an Engineering Dynamics Course](8.%09https:/peer.asee.org/regression-models-for-predicting-student-academic-performance-in-an-engineering-dynamics-course.pdf) Created by Shaobo Huang and Ning Fang, staff members at Utah State University

By collecting, compiling, and processing data for 239 undergraduate students over three semesters, these researchers were able to create multivariate linear regression models to predict student academic performance in Engineering Dynamics—a high enrollment, high-impact, and core engineering core that almost every mechanical or civil engineering student must take. The purpose of collecting and modeling this data was to help instructors develop a good understanding of how well or how poorly the students in their classes will perform so that instructors can take proactive measures to improve student learning. The inputs/independent variables of the model include the student’s cumulative GPA, gender, race, and other factors; the output/dependent variable of the models is a student’s final exam score in the Dynamics course stated above. Multiple criteria were utilized to evaluate and validate the predictive models, including R-square values, and average prediction accuracy. A good prediction was defined to have prediction error of plus or minor 10%. The results showed that the developed models had average prediction accuracy of 86.8% to 90.7%. This project is related to mine because it does a directly similar task of using models to determine student performance on exams.

My Contribution:

Since I am completing this project by myself, all code, graph creation, model creation, and data analysis was performed by me. One should examine my created Jupyter Notebook to clearly see my contribution.

**Approach:**

Objective: to understand which aspects have the largest impact on test outcomes as well as what are the best ways to improve student scores for the future.

The goal of this project is to use various supervised learning techniques including linear regression and logistic regression to ultimately predict student’s performance on examinations based on multiple socioeconomic and demographic statistics such as gender, race/ethnicity, parental level of education, and provided test scores. The dataset, provided by Royce Kimmons, has such data for 1000 students and includes both features—gender, race/ethnicity, parental level of education, test preparation course, lunch—and dependent variables/targets—scores in reading, writing, and mathematics. Unfortunately, not all essential and useful pieces of data, such as hours spent studying, are provided by this dataset, so I can only perform analysis of the effects these features have on performance. The model was fit with a Linear Regression model (because we are interested in the factors that influence performance). However, prediction of student performance will also be performed through creation of various types of graphs and other models such as Linear Least Squares Regression, Logistic Regression, Support Vector Machine, and Random Forest Regression. Through analysis of these models, we will also be able to determine the effectiveness of test preparation courses and the effect food/lunch has on performance. For each model, I will use different hyperparameters and preprocessing techniques to optimize performance.

The questions I attempted to answer through the exploratory data analysis and various supervised learning techniques are:

1. What effect does having or not having lunch place on student performance?
2. Are there substantial differences seen between males and females in regard to examination performance?
3. Does the parent's level of education impact the success of the student?
4. Does one's race serve as a benefit or hindrance (or neither) to their performance on examinations?
5. Will taking a test preparation score have a positive impact on student's examination scores?
6. Can we predict student performance by creating models with the provided dataset?
7. Can we *predict* one's gender by creating models with the provided dataset?
8. Are their high correlations between any of the features/target variables?

Proposed Approaches:

To accomplish the stated goals and answer the subsequent questions, I completed the following proposed approaches. To start, I proposed to examine similar projects, import all necessary libraries, read in the data/csv file, complete feature description, look for missing values, examine the data types of the features and targets, and remove outliers. For the exploratory data analysis, I proposed to create numerous graphs/plots to get a preliminary understanding of the data including histograms, pair plots, heatmaps, distance plots, pie charts, bar graphs, count plots, and density graphs. From the analysis of these graphs, I proposed to be already able to answer whether there are high correlations between any of the features/target variables and whether a clear distinction exists in performance on examinations based on gender. I also proposed to make an extended dataframe that contains other columns of data, including grades based on letter and pass/fail status. From here, I would start the creation of models, beginning with linear and ridge regression. For these models, I proposed to label encode the categorical features, split the dataset, and train and evaluate the models to be able to determine if one can predict student performance on examinations based on various categorical features. I then proposed to create a logistic regression model to determine which categorical features do the best job at classifying students as passing or failing their examinations (answering questions 1, 3, 4, and 5 above). Furthermore, I proposed to create another logistic regression model to determine if one’s gender could be classified/predicted based on their exam scores, answering question 7 above. For both of these models, I would label encode, split the dataset, train and select the hyperparameter with cross validation, and retrain and evaluate with the best hyperparameter. To explore new models and to go outside of my comfort zone, I proposed to build a Random Forest Regressor and Support Vector Machine (SVM) using Dummy Encoding. These models would help me to validate whether student performance could be predicted based off categorical features. Finally, I proposed to do a principal component analysis to set myself up to do a kNN classification, which included partitioning the dataset with Spectral Clustering, computing NMI, and performing PCA/dimensionality reduction. For all plots and models built, the results would be analyzed to complete my goal and answer the proposed questions.

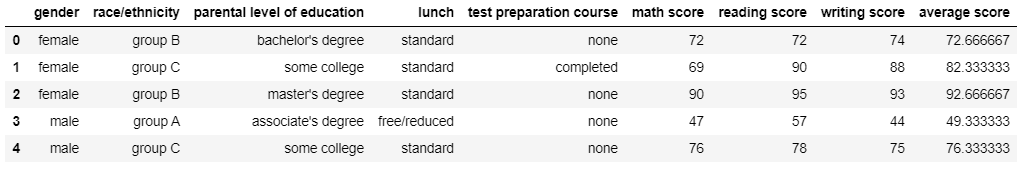
**Results:**

For the first week, I wrote the project proposal and progress report I, imported the dataset, imported required libraries, completed data preprocessing (including cleaning the data, making graphs to understand the data, label encoding, and splitting the dataset), and did exploratory data analysis (for a total of 12 hours). For the second week, I started and finished the linear and ridge regression models as well as all subsequent steps (label encoding, splitting the dataset, training, and evaluating). Model metrics were written, and scores for MSE, MAE, RMSE, and R-squared were coded and outputted in the notebook for both the training set and test set. Model performance was then analyzed by the creation of numerous graphs and plots of the outputted data. Next, I built a logistic regression model using k-fold cross validation to see if categorical features could classify a student’s pass result. I tuned the hyperparameter for the regularization term (I kept the code in for all tested hyperparameters to demonstrate the varying performance with each one). I then evaluated the accuracy, precision, recall, and F1 scores of the classifications for each preprocessing technique for an understanding of performance. I also created another logistic regression model to see if gender could be predicted based on examination scores. The last thing I did in week 2 was create a linear least squares regression model to understand the strength of the relationship between the reading and writing examination scores of students. Residuals—the deviation of the fitted values from the actual values—needed to be created for this to determine if this model was good to predict the reading scores from the same individual’s writing scores. A coefficient of determination (R^2 value) and a plot of the two features graphed was also created and outputted. With completion of these models came all corresponding pattern and model performance analysis. After spending 8 hours on these models and Progress Report II in Week two, I started and finish the support vector machine (SVM), random forest regressor, and principle component analysis in Week three. For SVM and random forest, I utilized dummy encoding and performed a similar procedure to the linear regression models to predict student performance on examinations based on categorical features. For the principal component analysis, I partitioned the dataset with spectral clustering, computed NMI, and performed dimensionality reduction. Thus, after spending 8 hours on this and the lightning talk in Week 3, the final week consisted of adding any finishing details, updating the Jupyter Notebook, writing the conclusion, and completing the final report.

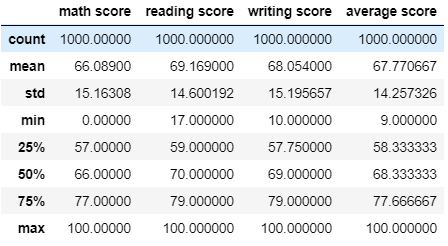
Going into detail of previously completed work, the first thing I did (after finishing the project proposal) was find and download the dataset. The dataset is provided by Royce Kimmons and can be found on the Kaggle website under the Student Performance on Examinations project. I did a preliminary exploration of the features and size of the dataset, and then moved on to setting up the coding environment. For consistency, I utilized the same program that is used in the class’s labs, Jupyter Notebook, accessed through the Anaconda Navigator. Since I created my own environment to work with, I needed to import multiple libraries to be able to perform all analysis that I plan to on this project at some point. To know which libraries I needed to import, I reviewed our class examples (Slides and Labs 3-8) and previous work with this dataset, available from the Kaggle dataset. I also revisited the similar projects listed in the Project Proposal document. I reviewed the code and setup of these notebooks not only to understand the necessary libraries to import but also to get an understanding of how I should go about working on the project and what I can include. Once I learned all the libraries necessary to complete my project, I started the import process in the Anaconda Navigator. I installed plotly, lightgbm, py-xgboost, and seaborn into the environment. Other libraries that I would use for my project that are already imported include numpy, pandas, matplotlib, sklearn, and pylab.

With the environment set up, the Jupyter Notebook software could be opened, and creation of the notebook began. A folder was created for the project, called ‘Final Project 3715’, with the csv dataset uploaded into it. Also added into the folder was a new notebook, which when opened was named ‘Agarwal—001—Final Project Student Performance on Examinations’. Next, due to my lack of experience with the intricacies of Jupyter Notebook, numerous things had to be looked up. For example, I had to learn how to make different headers, how to write plain text, how to write italics, how to make bullet points, how to organize the code well, how to reorder cells, how to add numbered points, how to create links in the document (for a table of contents), and how to do numerous shortcuts (delete cells, add cells above, add cells below, move cells, etc.). I learned this information through simple searches and through the help page in the software itself. Next, I added a title to the beginning of the document, my name, and the class name. Next, I created a cell called ‘Table of Contents’. When examining other projects with the dataset, I found that using a Table of Contents in a Jupyter Notebook is a great way for one to organize their dataset. Thus, as the project went on, I continued to add links to various areas in my projects for ease of access and organization in the notebook. Next, I added sections called ‘Introduction’ and ‘Goal’ which, as their title states, introduces the project (explain what will be done, why it will be done, how it will be done, etc.) and describes the end-goals. Another section, named “Questions to Try to Answer,” lists all of the essential questions I will try to answer through the analysis I will perform on this dataset, stated earlier.

Next, the coding part of the notebook began with importing the libraries stated earlier. The imports are broken down in the notebook based on what they will be used for. After this, I made a section for reading in the csv file, similar to how we do it with every lab. I also called the head() function to print the first few rows to get a superficial understanding of the dataset and its features. Furthermore, as shown below, I added a target column to the data called ‘average score’, which is simply the average of the math, reading, and writing scores. This will be useful for creating models in which having one overarching target column will be helpful.



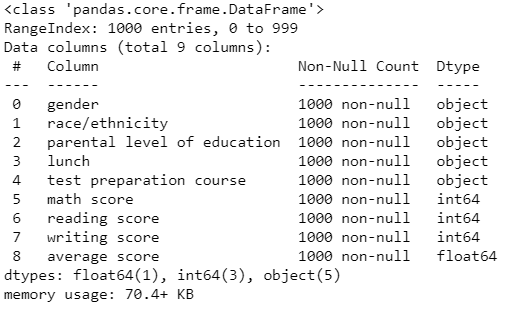
After this, I listed out the targets of the data (the scores) and the features (gender, race/ethnicity, parental level of education, lunch, and test preparation course). I next printed out the size of the dataset; seeing that the dataset has only 1000 points, I chose not to remove any outliers in hopes of preserving/holding on to as much of the data as possible. I still created a remove outlier function, however, in case I wanted to remove any such outliers. I also used the describe method to understand basic information about the numerical features such as the count, mean, std, min, and max and analyzed the meaning of these results:



One can understand a lot about the dataset just from this table output:

1. As expected, only the numerical features are displayed in the table (math, reading, writing, and average scores) because these are the only features that can have mathematical computations performed on them.
2. Examining the counts, all four score variables have a count of 1000, suggesting there are no missing values for the targets.
3. Looking at the mean scores, one can see that there is an average lower performance on math than reading and writing, but only by two to three percent.
4. The standard deviation is very similar for all columns, suggesting there may be some correlation to be found between these targets.
5. The minimum values are 0 for math, 17 for reading, and 10 for writing, suggesting at least one student in the dataset received these scores on the respective exams.
6. Finally, looking at the maximum scores, one can see that there is/are student(s) that achieved the maximum possible percentage, 100, in math, reading, and/or writing.

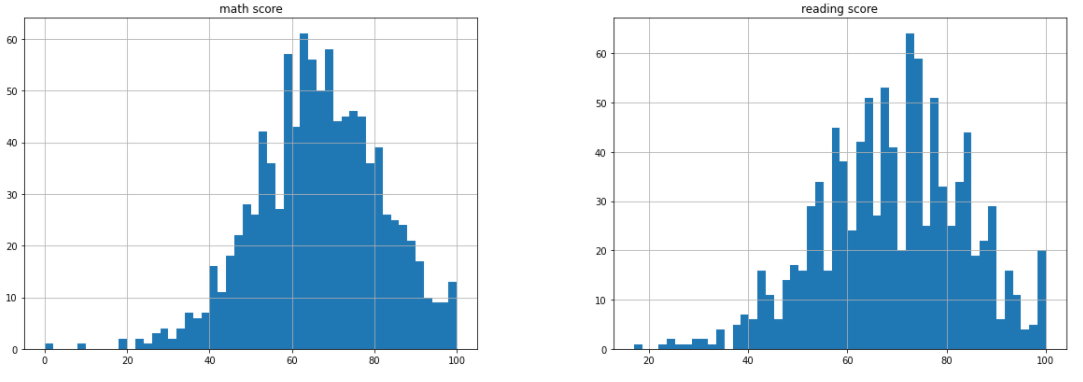
Moving on, I checked for missing values using isnull().sum(), which demonstrated that there are no missing values and no filling in would need to be completed. I also used the info() function to demonstrate which features were categorical and which numerical:

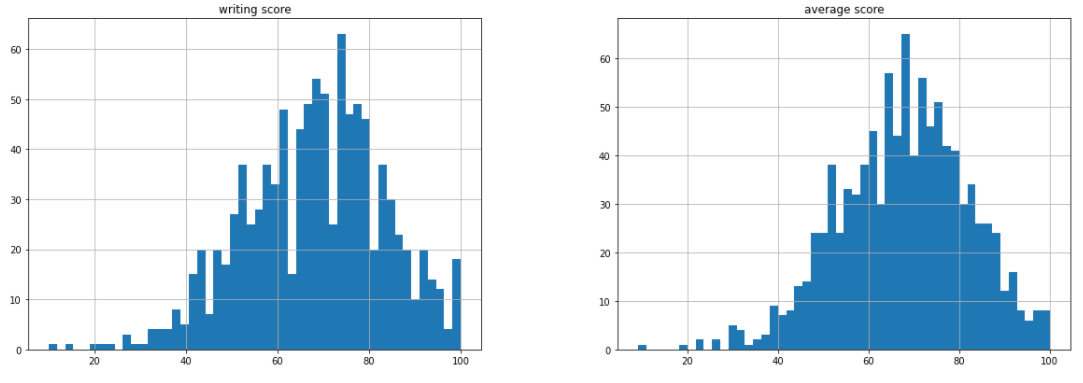


From this, we can see that the current categorical features are gender, race/ethnicity, parental level of education, lunch, and test preparation course, or all the features. All of the targets (math score, reading score, writing score, and average score) are numerical features. Thus, we will use Label Encoding to convert these categorical features to numerical ones later.

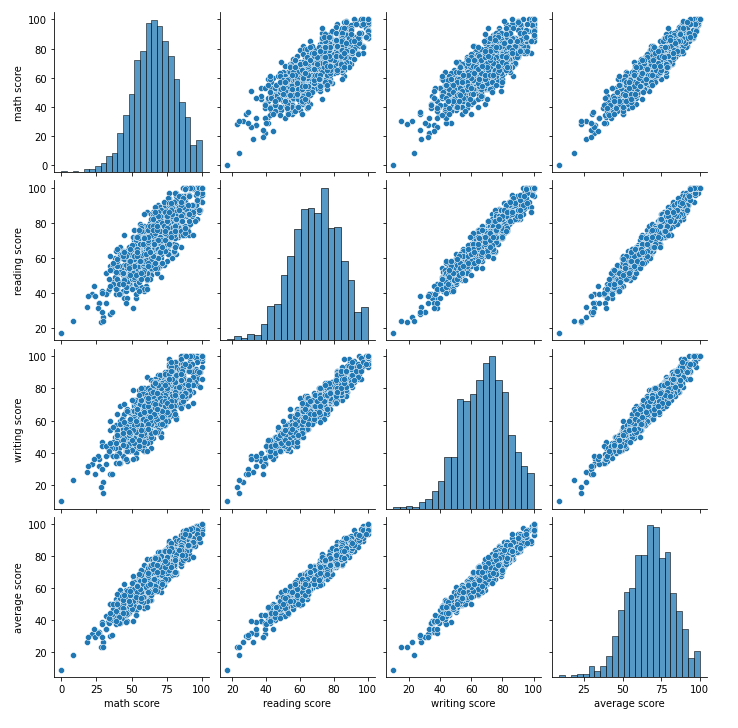
Exploratory Data Analysis:

With preliminary data examination completed, I started the exploratory data analysis/data visualization section of the project:



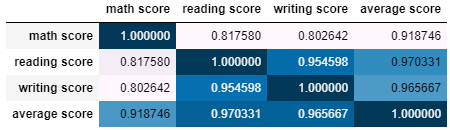


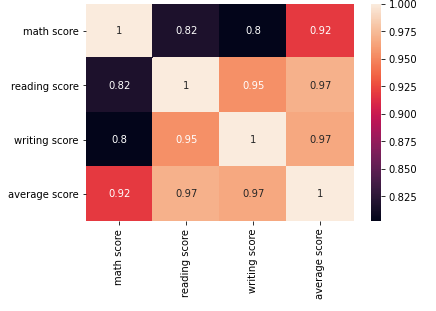
First, I visualized all numerical features with histogram plots to see their distributions and analyzed the results. I decided 50 bins was a decent amount to display the data well. Generally, one can see from these histogram plots that the scores follow qualitatively similar shapes. This will be more evident through other exploratory data analysis.



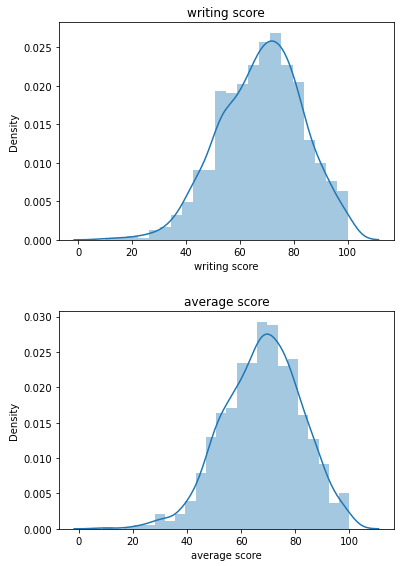
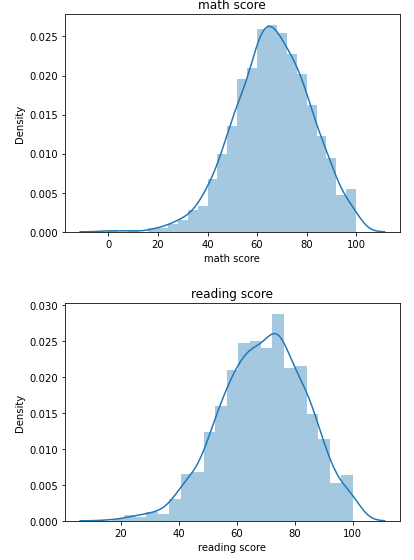
This pair plot provides another way to visualize the numerical features from the dataset.

Next, I used the corr() function of Pandas and the heatmap plot to show the correlation between different numerical features. The high correlation between the reading and writing scores was analyzed:



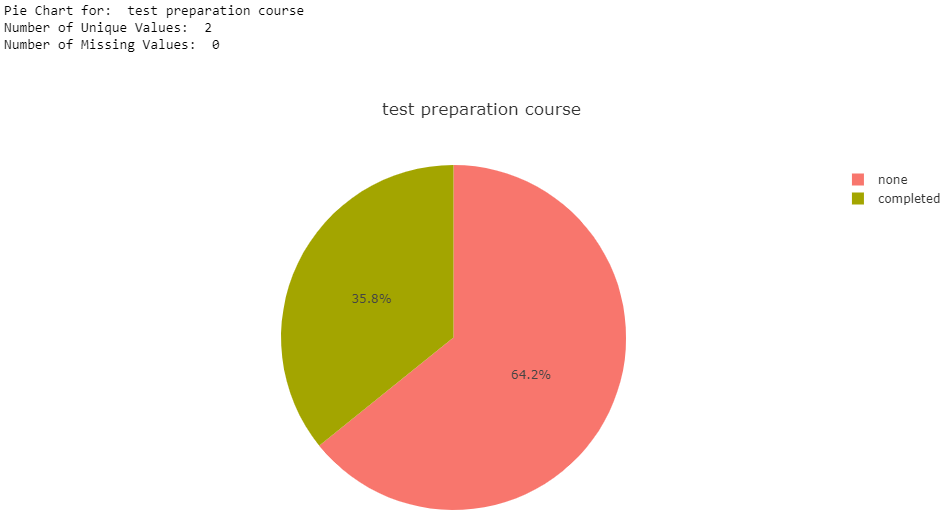
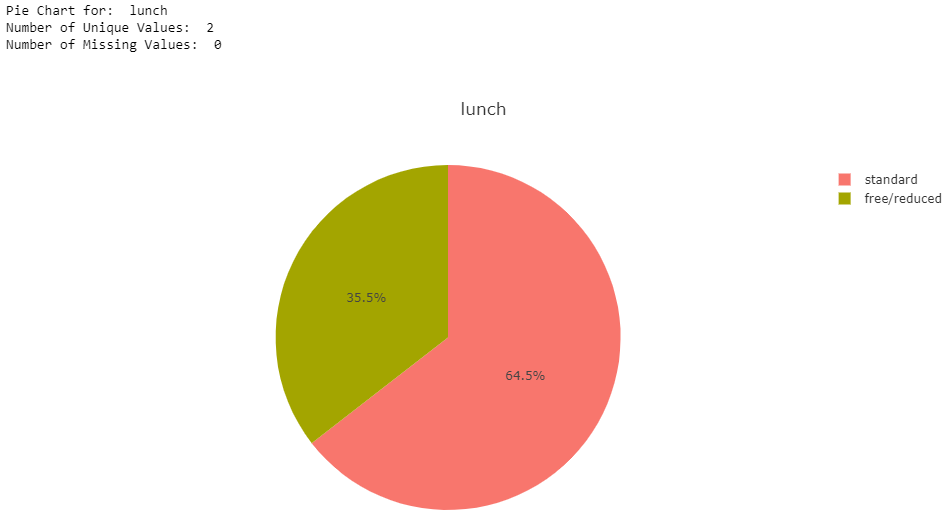
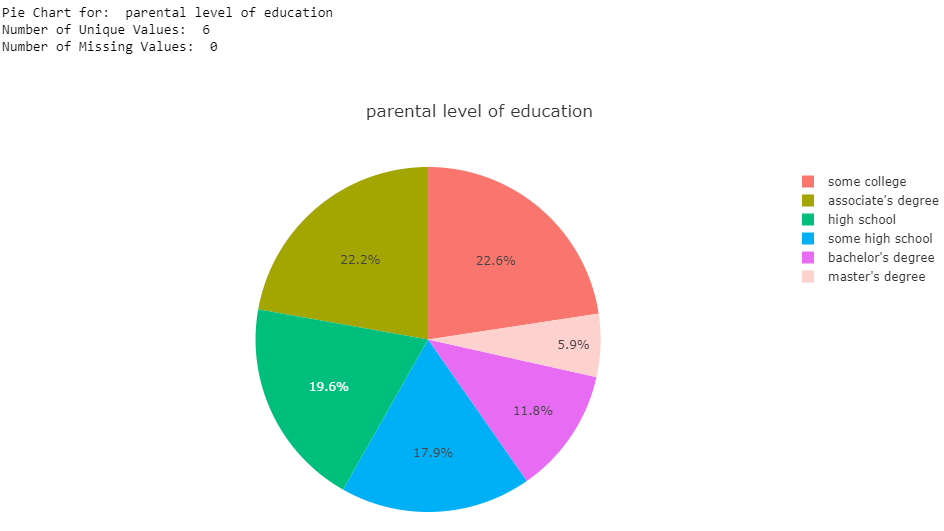
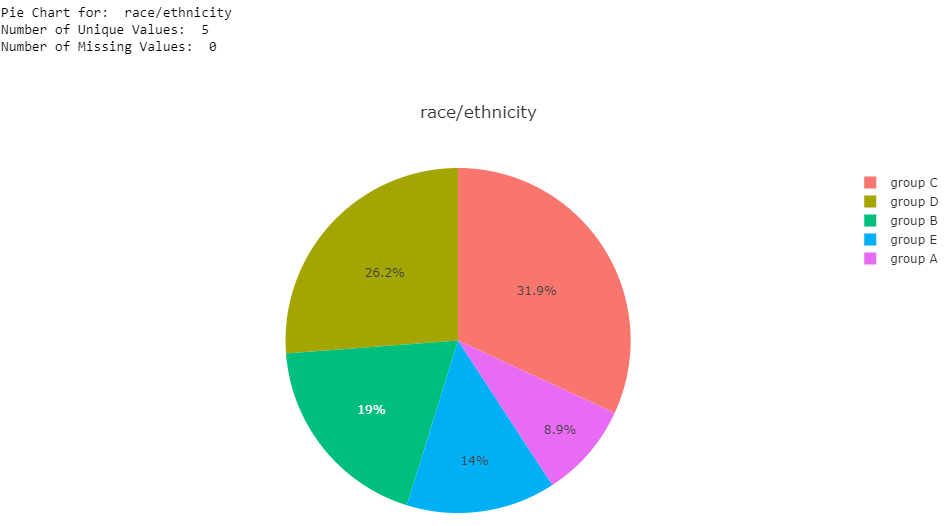
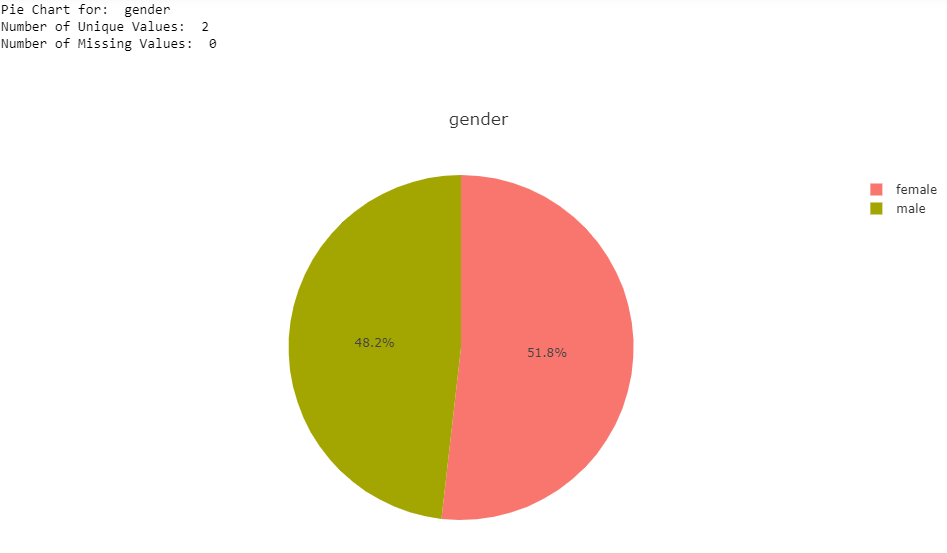


One can see that the highest correlation exists between the reading and writing scores of 0.954598, which is a statistically significant score to consider that they are correlated. We will show this in greater detail later. The correlation score between math and reading as well as math and writing are much lower at around .80 to .81, suggesting a lack of correlation between these. This heatmap then helps to answer one of the eight core questions listed: there does exist a statistically significant correlation between some of the features and/or targets: the reading and writing examination scores. Through the later creation of a linear least squares regression model, I will be able to validate this claim.



These graphs demonstrate another way to visualize the numerical features with the use of a density plots. One can see that they are all qualitatively similar, with the reading and writing curves being the most similar.

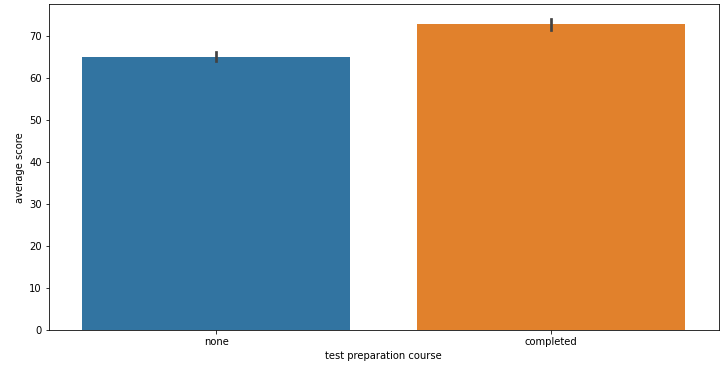
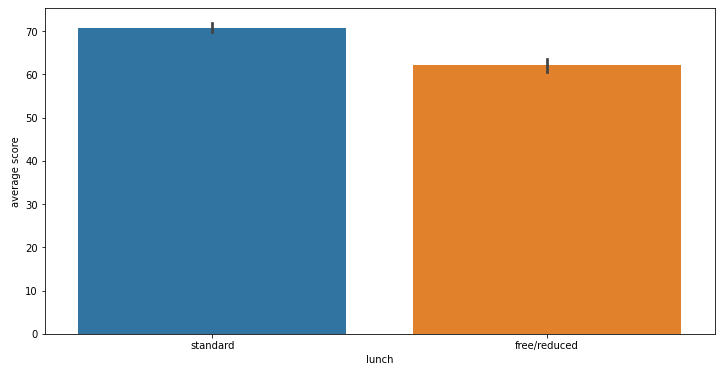
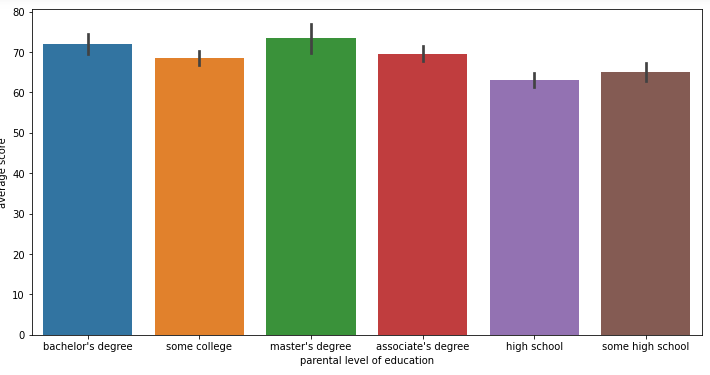
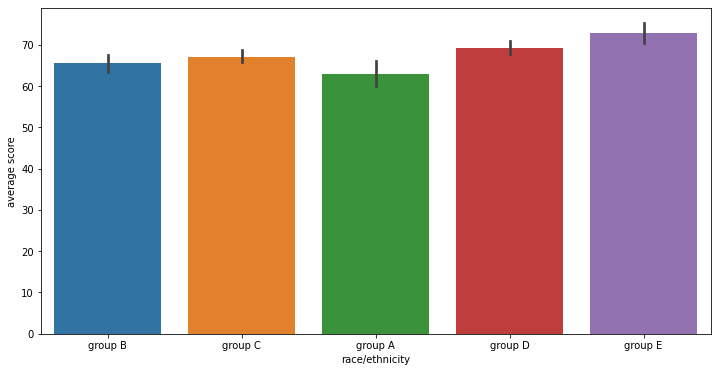
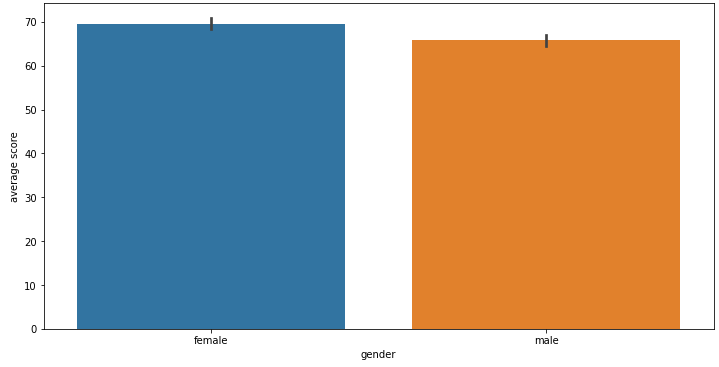
After data visualization of the numerical features, I examined the categorical features using a variety of graphs. By coding a method to create pie charts, I was able to make a pie chart for each categorical feature at the same time:



A lot of information can be gathered from these pie charts. One should take note that one can click on the percentages to see the exact value counts as well.

1. For gender, we can see there are two unique values, male and female, in which the dataset has a slightly larger percentage of females than males. Ideally, these should have been 50/50; however, in the real world, the proportion of males to females in similar (49/51). The data is split evenly enough to most likely not cause any issue.
2. For race/ethnicity, we see there are five unique values, labeled groups A-E, suggesting the actual races/ethnicities is not stated for the dataset, but rather labeled with alphabetical letters. The ordering of the groups from largest to smallest is group C, group D, group B, group E, and finally group A. This suggests that groups E and A are more of minorities than the other groups. Similar to real life, there is not an even distribution of races/ethnicities in the school system, especially in the United States.
3. For parental level of education, we see there are six unique values--some college, associate degree, high school, some high school, bachelor's degree, and master's degree. There is an approximately equal proportion of parents who have some college or an associate degree. Adding along the groups of parents with bachelor's degrees and master's degrees, one can qualitatively see that nearly two-thirds of the data has parents who have completed some college or more. The last third of parents either had some high school or only went up to high school. Like real life, there is not an even distribution of parental level of education. In our society, further education is optional as there are an innumerable number of fields one can pursue, some requiring higher education and others not. However, we would still like to see if there is a correlation between the parent's level of education and their student’s examination performance.
4. For lunch, we see there are two unique values, standard and free-reduced, where nearly two-thirds of students in the dataset had the standard lunch, and one third a free/reduced lunch. We will try to determine if this has any effect on student performance.
5. Finally, for test preparation course, we see there are obviously two unique values, either one completed a test preparation course or did not. Approximately two-thirds of students in the dataset did not complete a prep course, but about one-third did. Coincidentally, these proportions/percentages are extremely like those seen for the unique values for lunch. As said, this is most likely by chance.

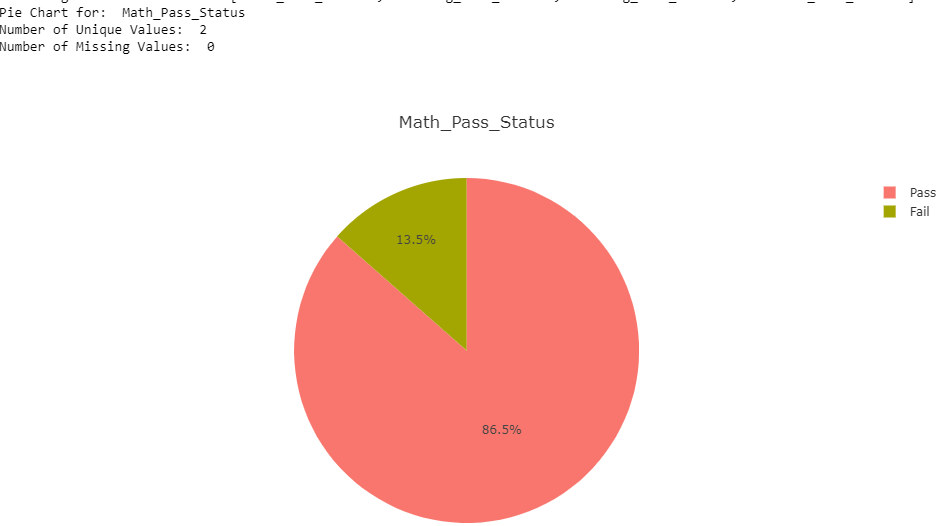
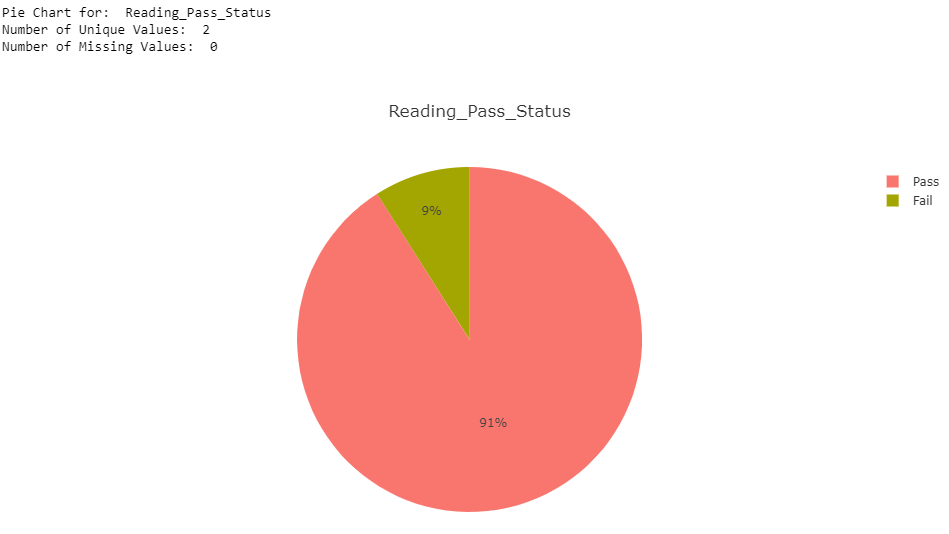
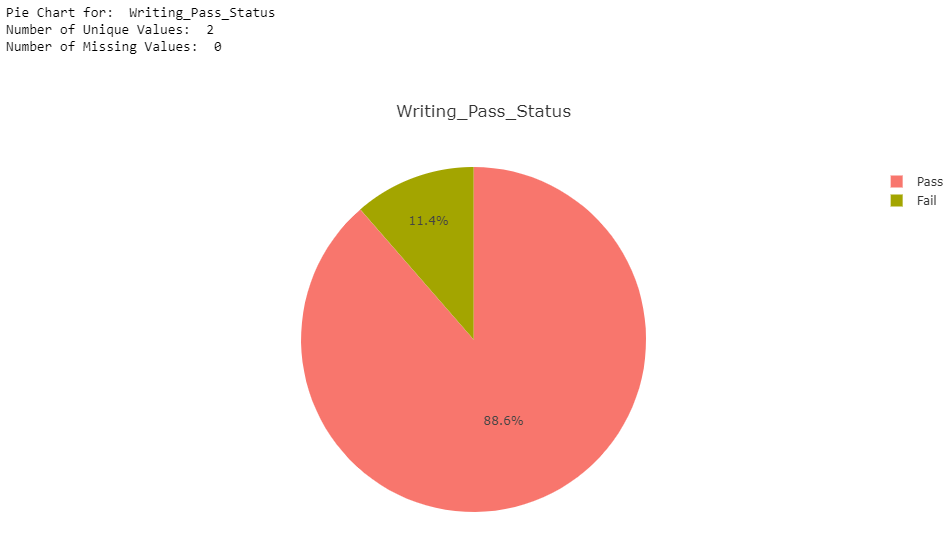
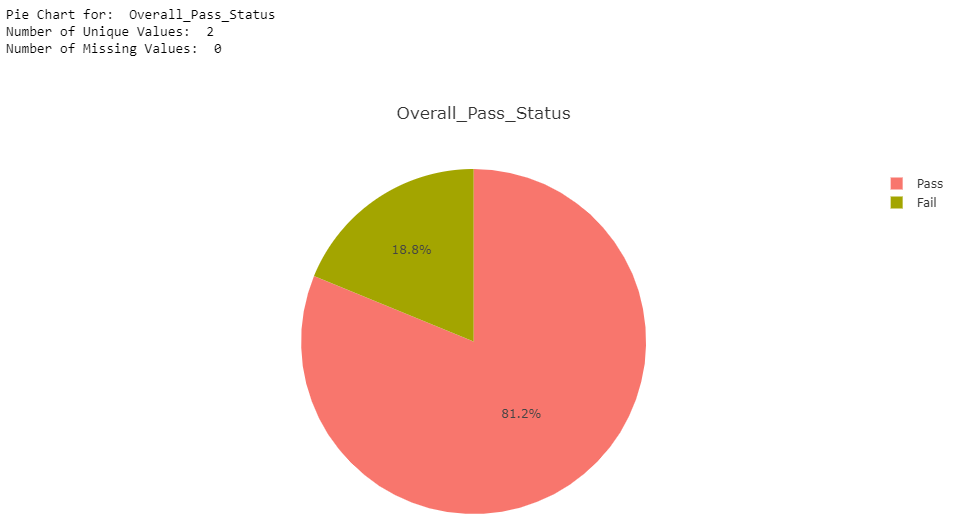
Further categorical feature analysis was performed by creating bar graphs to compare each feature’s unique values to the average score for those values:



A lot of general patterns can be determined from these five bar graphs:

1. For gender, we can see females do slightly better than males on exams.
2. For race/ethnicity, we can see the average scores are very similar no matter the race/ethnicity. However, group A, which we identified earlier as being a minority, is the lowest. Also, we identified group E as a minority, and it demonstrates the highest average score. This suggests that minorities deviate from the standard patterns.
3. For parental level of education, we can see the average scores are again very similar no matter the parental level of education. Still, the average scores increase as the parental level of education increases, suggesting a correlation.
4. For lunch, those students who have the standard lunch have a considerably larger average score than those with a free/reduced lunch.
5. Finally, for test preparation course, those students who completed a test preparation course sensibly had considerably higher average scores than those that did not.

Next, I explicitly coded a pass/fail mark, where 50% or lower is fail and standard letter grades are coded for the rest of the percentages. Then, I was able to see, for each student, which student is passing, and which is failing for each examination score. These new features were added to a copy of the dataframe, named df\_extended because I will only need these columns a few times. Using the previously created pie chart method, I was then able to create pie charts for the newly added categorical features:

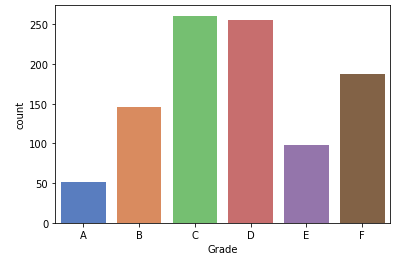
    From these pie charts, we can see that most students passed their exams. The most students passed their reading exams (91%), second most passed their writing exams, (88.6%), and the least passed was the math exams (86.5%). There was an overall pass percentage of 81.2%.

Two more columns were added to the extended dataframe, including ‘Total Marks’ and ‘Percentage’ to later find the percentage of marks.

Constructing a Grading Scale:

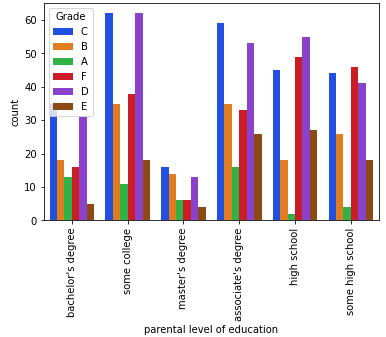
* 90 to 100 = A Grade
* 80 to 90 = B Grade
* 70 to 80 = C Grade
* 60 to 70 = D Grade
* 50 to 60 = E Grade
* Below 50 = F Grade (Fail)

With the created grading scale, I plotted the number of students who obtained each grade to get a general sense of how well student do on examinations:



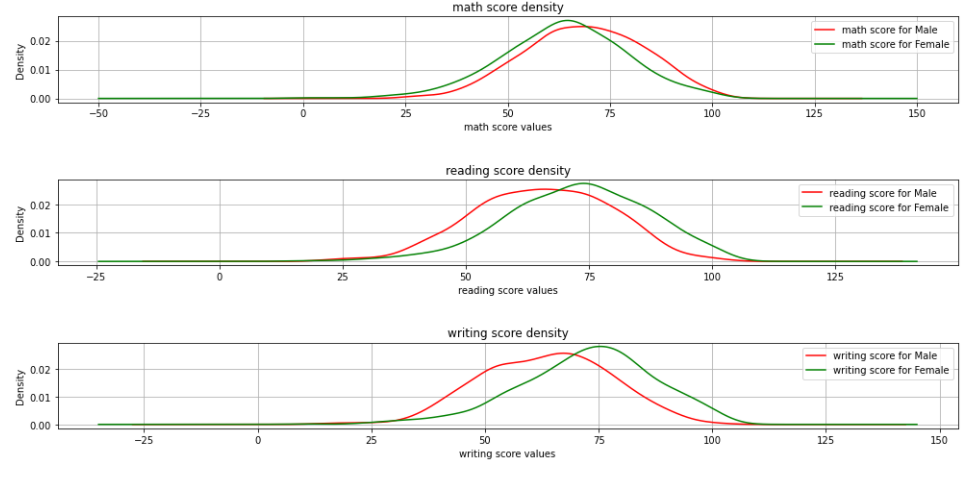
From this bar graph, we can see that the least obtained grades are A and E, and the most obtained grades are C and D (average), which makes sense with what one sees in real life. There is an unusually high number of students who failed (received an F), however.

I also made a count plot for parental level of education broken down by student grade:



Shown here is a more detailed breakdown of score by parental level of education than the bar graph created for such earlier. This breakdown provides counts based on letter grade, offering more information to analyze. One can see that, regardless of parental education level, the largest number of student's grades are average (C or D). One can also see that more students obtain an A or B as the level of education of the parent increases. This then answers one of my eight core questions: parental level of education does not seem to correlate with the success of the student on examinations. This will be validated through one of the created logistic regression models.

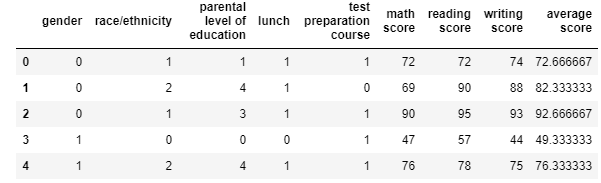
Finally, away from the grading scale, I performed analysis based on gender by developing plot density graphs for the three scores based on gender:



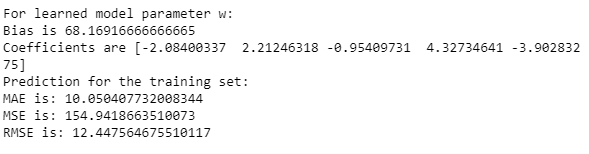
From these plot density graphs, we can generally see that males perform better on average than females on math exams, and females perform better on average than males on both reading and writing exams. These results agree with previous results we saw based on gender. This then answers one of the eight stated questions attempting to be answered in this project: a distinction does exist between males and females in terms of performance on academic examinations.

Linear Regression:

After the Exploratory Data Analysis, I began the creation of my first model, the linear regression model. I am creating a linear regression model to determine if one can use categorical features--gender, race/ethnicity, parental level of education, having or not having lunch, and taking or not taking a test preparation course--to predict student performance on an examination. As done in labs, I converted the categorical features to numerical features using Label Encoding. This had to be done for a total of five columns, and the info method was called on the dataset to confirm that the whole dataset is now numerical values:

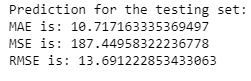


Next, I split the dataset with 80% of the data being in the training set, and the final 20% in the test set. To train the linear regression model to do prediction, I used min𝑤 1/𝑛‖𝑦−𝑋𝐰‖22 (apologies for poor formatting, it is hard to write this in Word) and outputted the learned model parameter w to see how the learned model fit the training set. I of course printed out the MAE, MSE, and RMSE scores for the training set as well:

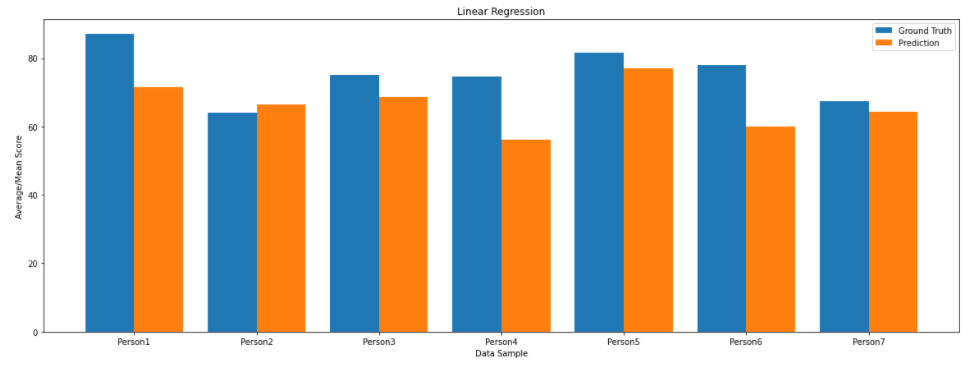


While not much information can be collected from these scores alone, they will be compared to results from other models such as Ridge Regression, SVM, and Random Forest Regression to determine their validity.

I then evaluated the learned model to see how well the model generalized on the testing test.



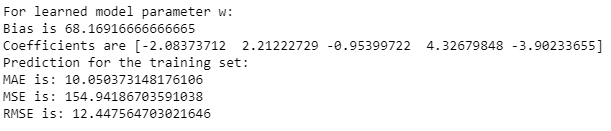
To better understand the results, I coded and outputted a bar graph of the ground truth and prediction:

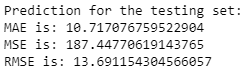


By comparing the ground truth and prediction for a sample of the data, one can see that the learned linear regression model does well to generalize on the testing set. In other words, it is satisfactory to use the categorical features in the dataset to predict student performance on examinations, which answers another one of the main questions of this project.

Ridge Regression:

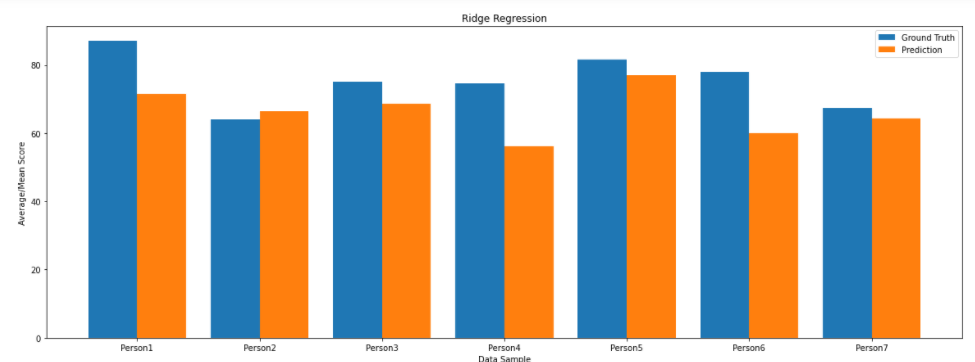
Similar to linear regression, I then began and completed the ridge regression procedure. For ridge regression, I used min𝑤 1/𝑛‖𝑦−𝑋𝑤‖22+𝜆‖𝑤‖22 (again, sorry for poor formatting) instead of the formula stated for linear regression. I also used different lambda values to see their affect on the performance of the model on the testing set. The prediction scores for the ridge regression model on the training set (and later the test set) were outputted:





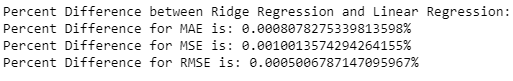
One could see from the scores that the prediction values for the training set for MAE, MSE, and RMSE were almost identical for the values for Linear Regression, as we expect (since the same training and test sets are used).

I also created a bar graph for ridge regression to pictorially see how well the learned model did to generalize on the test set (and it did do well).



By comparing the ground truth and prediction for a sample of the data, one can see that the learned ridge regression model does well to generalize on the testing set. In other words, it is satisfactory to use the categorical features in the dataset to predict student performance on examinations.

To compare the performance of the ridge regression and linear regression models, I used a percent difference calculations (Percent Difference = (|Value 1 – Value 2|)/([Value 1 + Value 2]/2) \* 100%):

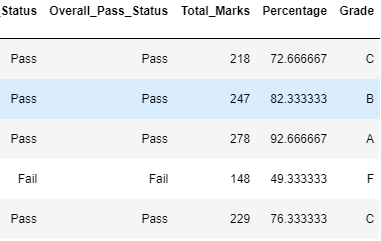
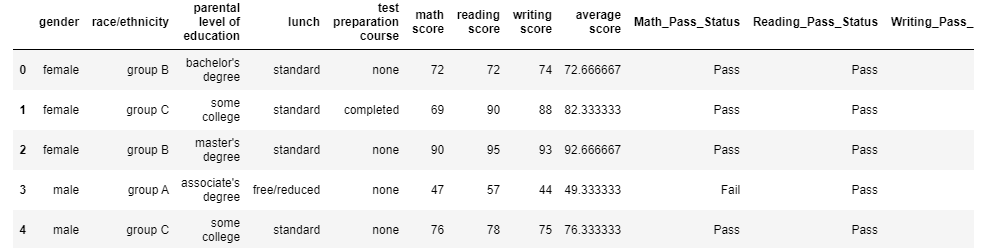


The small percent difference values between the ridge and linear regression models suggest they perform very similarly.

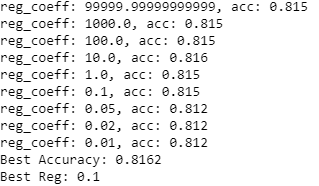
One could see that, when testing different lambda values, the performance of the ridge regression model on the testing set worsened as the lambda value increased. This is because if one's lambda value is too high, the model will be simple, but they run the risk of underfitting the data (it did not learn enough about the training data to make useful predictions). Lambda values becoming incrementally smaller than 0.1 were also tested. Once the lambda values became extremely small, the model became more complex and ran the risk of overfitting the data.

Logistic Regression #1: Student Exam Pass Status Based on Categorical Features:

Next, I built a logistic regression model using k-fold cross validation to determine which categorical features do the best job at classifying students as passing or failing their examinations. I used the extended dataframe I made so that I have access to Pass/Fail status:



After completing label encoding and dataset splitting for this extended dataframe, I used the formula  to do classification. I tuned the hyperparameter for the regularization term (I kept the code in for all tested hyperparameters to demonstrate the varying performance with each one). With different hyperparameters, I got different model parameters w, resulting in different prediction performance. Thus, I used 10-fold cross-validation to select the best hyperparameter.



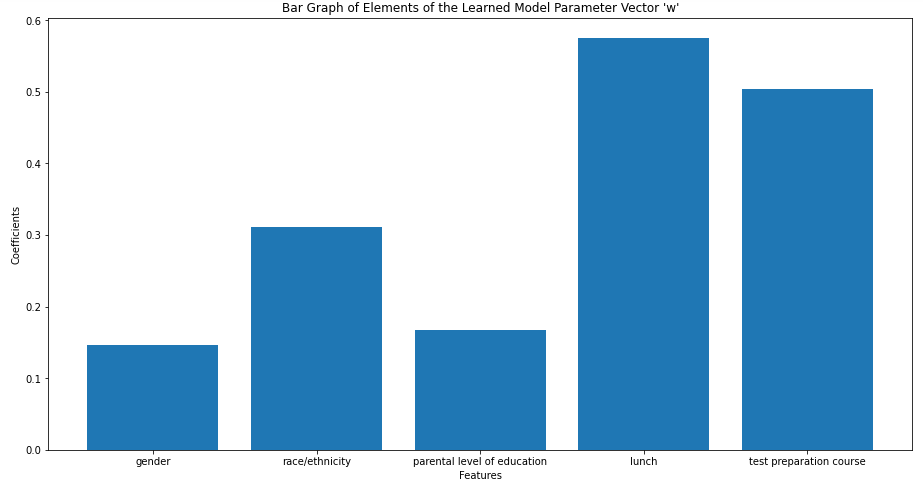
Thus, the best regularization coefficient was found to be 0.1 and was used for subsequent model building.

From there, I retrained the model with the best hyperparameter on the training validation set. I then evaluated the accuracy, precision, recall, and F1 scores of the classifications for each preprocessing technique for an understanding of performance:



These scores do not help us learn much about the performance of the model, but the generally high accuracy, recall, precision, and f1 scores suggest the model did well.

I then used a bar plot to visualize the elements of the learned model parameter vector w:

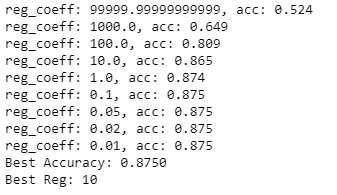


From this plot, we can see that some features have much greater coefficients in the learned model parameter than others. Specifically, lunch and test preparation course have the higher coefficient values. This means that these features were able to fit the logistic regression model with the hyperparameter best of all of the features. One could say having or not having lunch and/or a test preparation course are then good predictors of whether a student passes or fails an exam. It also tells us that lunch and test preparation course are more statistically significant than the other features. The closer the coefficient values are to 0, the less correlation exists (that is why we took the absolute value).

It should be noted that, although we took the absolute value of the coefficients, a positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease.

Logistic Regression #2: Gender Prediction Based on Examination Scores:

Next, I built another logistic regression model; however, this time I determined if one's gender can be classified/predicted based on their examination scores. I again used the extended dataframe I made so that I have access to Pass/Fail status. Following the same procedure as above, a best regularization coefficient was found:



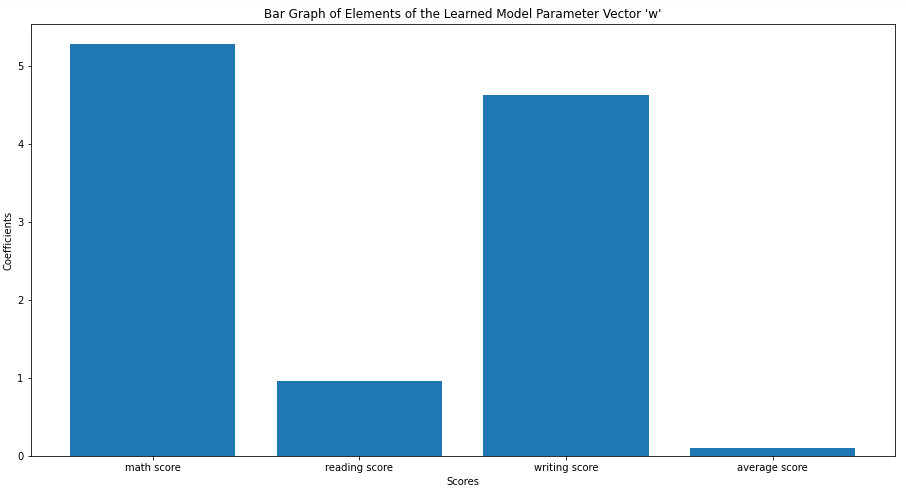
Thus, the best regularization coefficient was found to be 10 and was used for subsequent model building.

From there, I retrained the model with the best hyperparameter on the training validation set. I then evaluated the accuracy, precision, recall, and F1 scores of the classifications for each preprocessing technique for an understanding of performance:



These scores do not help us learn much about the performance of the model, but the generally high accuracy, recall, precision, and f1 scores suggest the model did well.

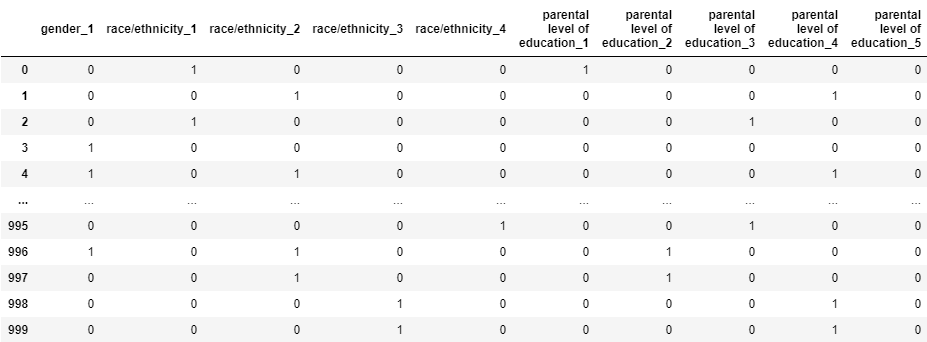
I again used a bar plot to visualize the elements of the learned model parameter vector w:



From this bar graph, we can see that some scores have much greater coefficients in the learned model parameter than others. Specifically, math and writing scores have the higher coefficient values. This means that these scores were able to fit the logistic regression model with the hyperparameter best of all of the features. One could say that, of the scores provided, the math and writing scores are the best predictors of a student's gender. The closer the coefficient values are to 0, the less correlation exists (that is why we took the absolute value). Thus, the reading and average scores are not good predictors of one's gender. One can then say that there is not a clear distinction between which gender performs better on reading examinations (no correlation). It makes sense for the average score to not be a good predictor of one's gender because this score value is not representative of a single subject/category nor the person's actual score (it is an average).

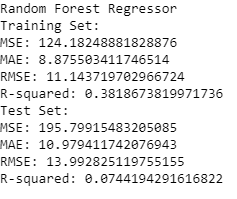
Random Forest Regressor:

With four models done, I moved on to explore new models and to go outside of my comfort zone; thus, I built a Random Forest Regressor and Support Vector Machine (SVM) using Dummy Encoding. These models would help me to validate whether student performance could be predicted based off categorical features. The resulting matrix after dummy encoding is shown below:



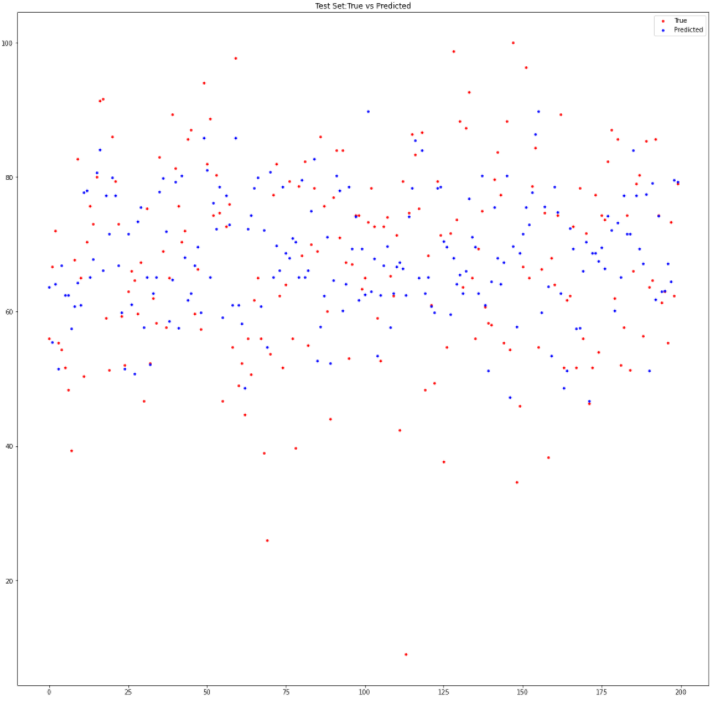
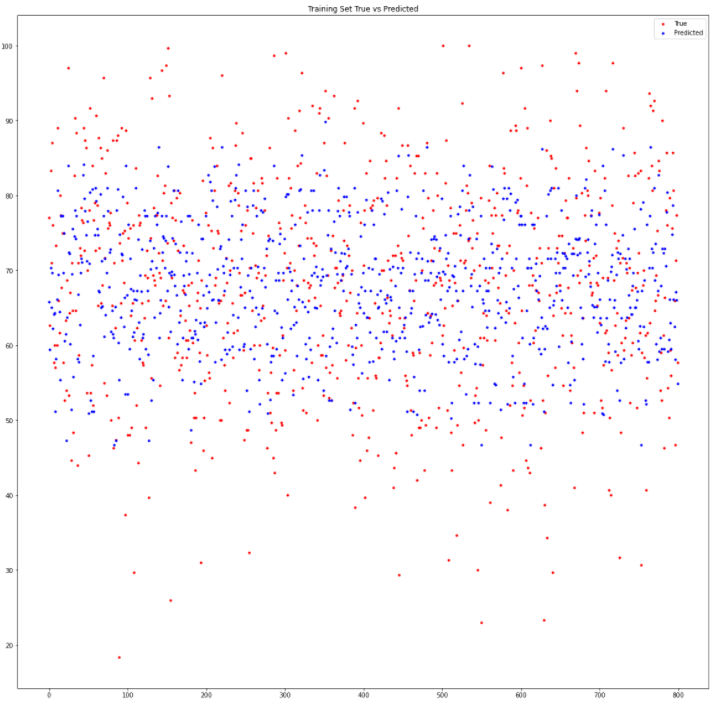
Through dummy encoding, the dataset now consists of only values of 1 or 0, important for the following two models. It should be noted that there only needs to be one column for gender, lunch, and test preparation course because there were already only two possibilities for these columns.

Following a similar procedure to linear and ridge regression (splitting the dataset and building the model), MAE, MSE, RMSE, and R-squared values were outputted for the training and test sets:



One could see from the scores that the prediction values for the training set for MAE, MSE, and RMSE were almost identical for the values for Linear Regression and Ridge Regression, as we expect (since the same training and test sets are used).

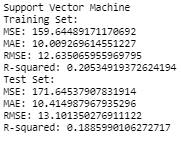
A scatterplot was also created to pictorially demonstrate the performance of the model on both the training set and the test set:



It makes sense that the scatterplot for the test set is a lot more sparse because it only uses 200 data values while the training set contains 800 values. In general, it appears as though the model does not do the best job to account for scores that deviate from the norm, which makes sense.

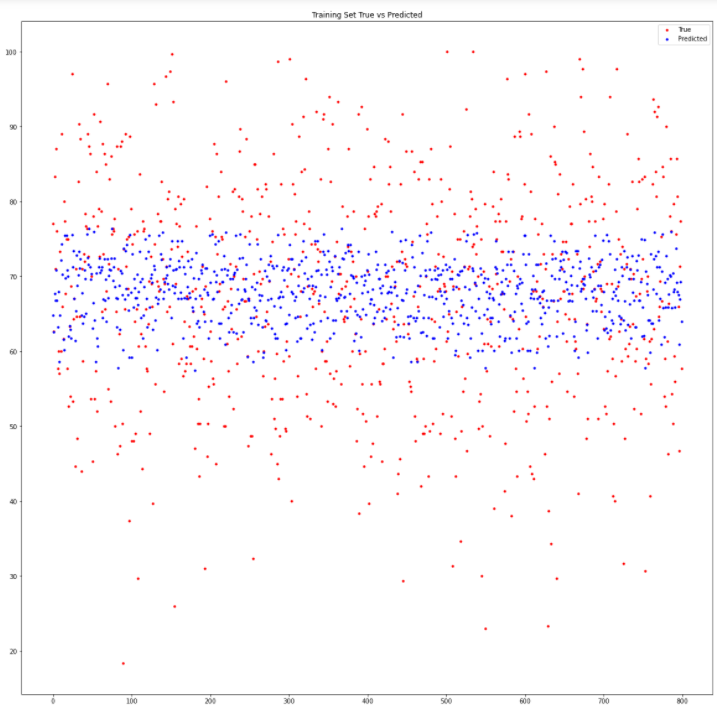
Support Vector Machine (SVM):

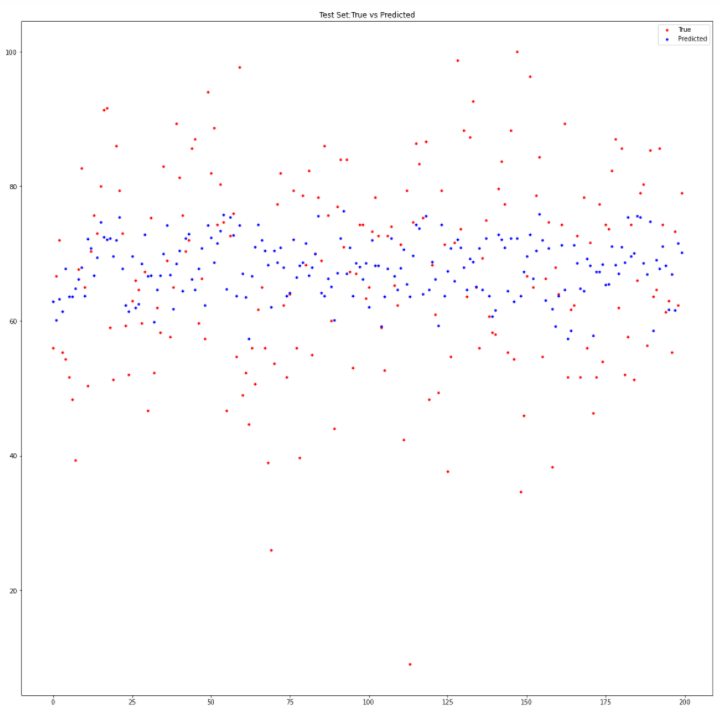
As stated earlier, I followed an almost identical procedure to build a support vector machine (SVM). The results were as follows:



One could again see from the scores that the prediction values for the training set for MAE, MSE, and RMSE were almost identical for the values for Linear Regression, Ridge Regression, and Random Forest Regressor, as we expect (since the same training and test sets are used).

A scatterplot was again created to pictorially demonstrate the performance of the model on both the training set and the test set:





It makes sense that the scatterplot for the test set is a lot more sparse because it only uses 200 data values while the training set contains 800 values. In general, it appears as though the model does not do the best job to account for scores that deviate from the norm, just like the Random Forest Regressor. It is safe to assume the Linear Regression and Ridge Regression models, if graphed as scatterplots, would have produced similar results.

Comparing the MAE, MSE, and RMSE Scores for Linear Regression, Ridge Regression, Random Forest, and Support Vector Machine, one can see that the values are all around the same range, suggesting they perform similarly. It then seems that it may have not been the best decision to create four models that would be trying to answer the same question (can the categorical features be used to predict student performance on examinations).

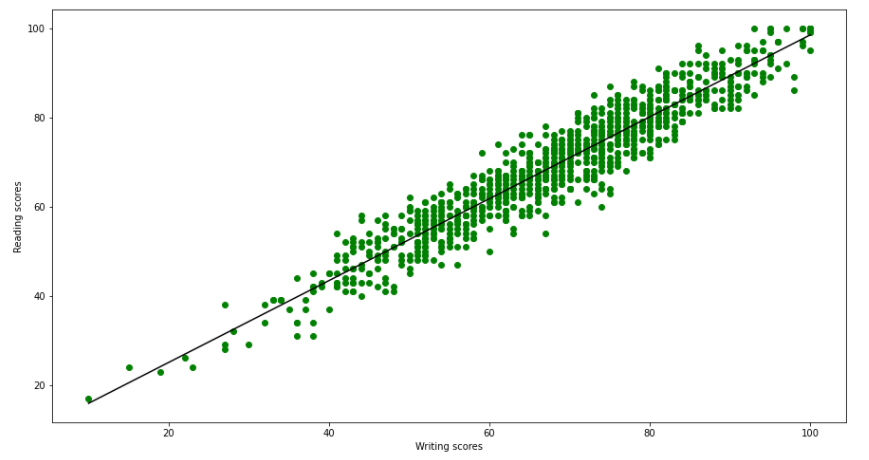
Linear Least Squares Regression:

Moving on, I created an extra linear regression model, called a linear least square regression model, to understand the strength of the relationship between the reading and writing scores. It should be noted that this does not measure the slope. I first got the intercept and slope of the line using linear least squares to predict the reading scores. The intercept was 6.688023759635357 and the slope was 0.9181087994881217. Residuals—the deviation of the fitted values from the actual values—needed to be created for this to determine if this model was good to predict the reading scores from the same individual’s writing scores. The actual equation of the line that used in regression was ys = intercept + slope xs + residuals. To determine if it is good to predict the reading scores from the writing scores or without it, I outputted the RMSE scores:



The Root Mean Square Error (RMSE) is more than 4 if I use the writing scores to predict the values of the reading scores and is more than 14 if I do not use the writing scores to predict the values of the reading scores. Therefore, as the RMSE is less in the former situation, it proves effective to use the writing scores to predict the reading scores. The high r^2 value validates that there is a statistically significant correlation between the reading and writing scores.

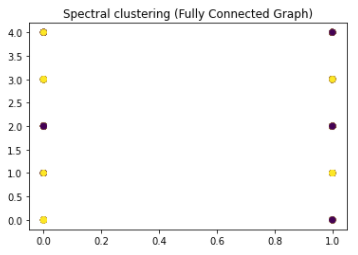
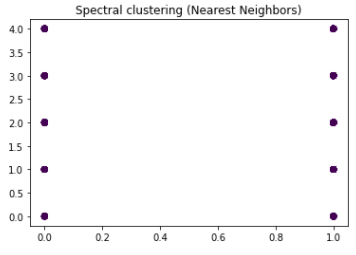
I then created and outputted a plot of the linear least squares line to pictorially validate that a statistically significant correlation exists between reading and writing scores:



From this plot, one can easily see that the model fits the data well. Thus, one can use writing scores to predict reading scores, as predicted in the exploratory data analysis section.

Principal Component Analysis (PCA):

Finally, I conducted Principal Component Analysis to set myself up to do a kNN classification, which included partitioning the dataset with Spectral Clustering, computing NMI, and performing PCA/dimensionality reduction.

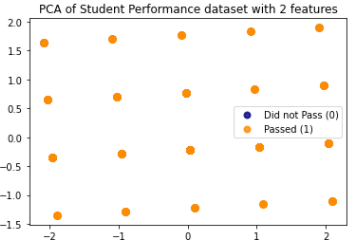


The normalized mutual information score of the Spectral Clustering (Nearest Neighbors) method is 0.0085

The normalized mutual information score of the Spectral Clustering (Fully Connected Graph) method is 0.0002

Based on these plots, we see that both Spectral clustering similarity graphs do very poor to represent the data. The data we are fitting is everything except the Overall Pass Status column, and the y values are from this column, which only contains values of 0 or 1. Thus, it makes sense that the graph data seems to be split into two sections, either 0 or 1. Unfortunately, neither graph appropriately splits the 0's and 1's, as there some belonging to each color/group in both the 1's and the 0's. Again, this result is to be expected. The bad performance is validated by the extremely low normalized mutual information score for nearest neighbors and the fully connected graph. While bad, this result is unfortunately expected.

After performing dimensionality reduction from 5 features to 2, a graph was created to visualize the dataset once more:



Based on the produced graph, one can see that performing PCA does not provide any useful information (the graph is only showing equally spaced points of students who passed the exam). Therefore, it is best not to perform any further type of PCA analysis. I have also then decided that doing k-NN Classification would be futile here.

**Conclusion:**

To conclude, by performing exploratory data analysis, building multiple models (Linear Regression, Ridge Regression, Logistic Regression, Random Forest Regressor, Support Vector Machine, Linear Least Squares), and performing Principal Component Analysis, I was ultimately able to predict student performance on examinations based on multiple socioeconomic and statistics such as gender, race/ethnicity, parental level of education, and provided test scores.

The exploratory data analysis provided a lot of useful information in and of itself. First, the histograms demonstrated that the scores followed qualitatively similar shapes, especially the reading and writing scores, which was validated through the creation of the linear regression model later. The heatmap also validated this claim by demonstrating that the highest correlations exist between the reading and writing scores with a correlation value of 0.954598, which is a statistically significant score to consider that they are correlated.

From the creation of the pie charts, one could see there are two unique values for gender, male and female, in which the dataset has a slightly larger percentage of females than males. Ideally, these should have been 50/50; however, in the real world, the proportion of males to females in similar (49/51). The data is split evenly enough to most likely not cause any issue. For race/ethnicity, we see there are five unique values, labeled groups A-E, suggesting the actual races/ethnicities is not stated for the dataset, but rather labeled with alphabetical letters. The ordering of the groups from largest to smallest is group C, group D, group B, group E, and finally group A. This suggests that groups E and A are more of minorities than the other groups. Similar to real life, there is not an even distribution of races/ethnicities in the school system, especially in the United States. For parental level of education, we see there are six unique values--some college, associate degree, high school, some high school, bachelor's degree, and master's degree. There is an approximately equal proportion of parents who have some college or an associate degree. Adding along the groups of parents with bachelor's degrees and master's degrees, one can qualitatively see that nearly two-thirds of the data has parents who have completed some college or more. The last third of parents either had some high school or only went up to high school. Like real life, there is not an even distribution of parental level of education. In our society, further education is optional as there are an innumerable number of fields one can pursue, some requiring higher education and others not. For lunch, we see there are two unique values, standard and free-reduced, where nearly two-thirds of students in the dataset had the standard lunch, and one third a free/reduced lunch. Finally, for test preparation course, we see there are obviously two unique values, either one completed a test preparation course or did not. Approximately two-thirds of students in the dataset did not complete a prep course, but about one-third did. Coincidentally, these proportions/percentages are extremely like those seen for the unique values for lunch. As said, this is most likely by chance.

From the creation of bar graphs, one could see that for gender, females do slighter better than males on examinations, answering one of our eight key questions. For race/ethnicity, we can see the average scores are very similar no matter the race/ethnicity. However, group A, which we identified earlier as being a minority, is the lowest. Also, we identified group E as a minority, and it demonstrates the highest average score. This suggests that minorities deviate from the standard patterns. For parental level of education, we can see the average scores are again very similar no matter the parental level of education. Still, the average scores increase as the parental level of education increases, suggesting a correlation. For lunch, those students who have the standard lunch have a considerably larger average score than those with a free/reduced lunch. Finally, for test preparation course, those students who completed a test preparation course sensibly had considerably higher average scores than those that did not.

By adding a pass/fail status and creating pie charts based on each subject score, we can see that most students passed their exams. The most students passed their reading exams (91%), second most passed their writing exams, (88.6%), and the least passed was the math exams (86.5%). There was an overall pass percentage of 81.2%. Similarly, after creating a grading scale and correlated bar graph, we can see that the least obtained grades are A and E, and the most obtained grades are C and D (average), which makes sense with what one sees in real life. There is an unusually high number of students who failed (received an F), however. By breaking down the grades by parental level of education, one can see that, regardless of parental education level, the largest number of student's grades are average (C or D). One can also see that more students obtain an A or B as the level of education of the parent increases. This then answers one of my eight core questions: parental level of education does not seem to correlate with the success of the student on examinations. This was validated through one of the created logistic regression models.

Finally, through creation of density plots, we can generally see that males perform better on average than females on math exams, and females perform better on average than males on both reading and writing exams (which adds to the conclusion made generally from the bar graphs). This then answers one of the eight stated questions attempting to be answered in this project: a distinction does exist between males and females in terms of performance on academic examinations.

From here, a linear regression model was created with an 80/20 train/test split. While the MAE, MSE, and RMSE scores for the training and test set were outputted, not much information could be gathered from them alone. After graphing the ground truth versus prediction, one could see that the learned linear regression model did well to generalize on the testing set. In other words, it is satisfactory to use the categorical features in the dataset to predict student performance on examinations, which answers another one of the main questions of this project. The ridge regression model followed a similar procedure, outputting almost identical MAE, MSE, and RMSE scores for the training and test sets, validated by the extremely small percent difference values. The small percent difference values between the ridge and linear regression models suggest they perform very similarly.

One could see that, when testing different lambda values, the performance of the ridge regression model on the testing set worsened as the lambda value increased. This is because if one's lambda value is too high, the model will be simple, but they run the risk of underfitting the data (it did not learn enough about the training data to make useful predictions). Lambda values becoming incrementally smaller than 0.1 were also tested. Once the lambda values became extremely small, the model became more complex and ran the risk of overfitting the data.

Next, two logistic regression models were created to help answer whether student exam pass status could be classified based on categorical features and whether one’s gender could be predicted based on their examination scores for each subject. After obtaining the best hyperparameter (0.1 for the first named model and 10 for the second model), accuracy, recall, precision, and f1 scores were outputted. These scores did not help us learn much about the performance of the model, but the generally high accuracy, recall, precision, and f1 scores suggest the model did well. The learned model parameter vectors ‘w’ had their components graphed to reach general yet significant conclusions. For the classification of exam pass status, one learned that some features had much greater coefficients in the learned model parameter than others. Specifically, lunch and test preparation course had the higher coefficient values. This means that these features were able to fit the logistic regression model with the hyperparameter best of all of the features. One could say having or not having lunch and/or a test preparation course are then good predictors of whether a student passes or fails an exam. It also tells us that lunch and test preparation course are more statistically significant than the other features. The closer the coefficient values are to 0, the less correlation exists (that is why we took the absolute value). For the gender classification, one learned that math and writing scores had the higher coefficient values. This means that these scores were able to fit the logistic regression model with the hyperparameter best of all the features. One could say that, of the scores provided, the math and writing scores are the best predictors of a student's gender. Thus, the reading and average scores are not good predictors of one's gender. One can then say that there is not a clear distinction between which gender performs better on reading examinations (no correlation). It makes sense for the average score to not be a good predictor of one's gender because this score value is not representative of a single subject/category nor the person's actual score (it is an average).

After this, the Random Forest Regressor and Support Vector Machine were able to confirm the results of the Linear and Ridge Regression models as their MAE, MSE, and RMSE scores were almost identical. The scatterplots created were also helpful to demonstrate that all four models would not do well to account for scores that deviate from the norm, which makes sense. It also seems that it may have not been the best decision to create four models that would be trying to answer the same question (if the categorical features can be used to predict student performance on examinations).

Moving on, the linear least squares regression model was able to validate that there exists a statistically significant relationship between the reading and writing scores. The intercept was 6.688023759635357 and the slope was 0.9181087994881217. The Root Mean Square Error (RMSE) was more than 4 if I used the writing scores to predict the values of the reading scores and was more than 14 if I did not use the writing scores to predict the values of the reading scores. Therefore, as the RMSE is less in the former situation, it proves effective to use the writing scores to predict the reading scores. The high r^2 value validates that there is a statistically significant correlation between the reading and writing scores. The plotting of the linear least squares line provides one final validation that the model fits the data well.

Finally, by conducting Principal Component Analysis and reducing the features from five to two, it was found that performing PCA did not provide any useful information (the graph was only showing equally spaced points of students who passed the exam). Therefore, I determined it was best not to perform any further type of PCA analysis. I then also decided that doing k-NN Classification would be futile here.

Overall, all stated questions were able to be answered throughout this project, and the goals of this project were reached successfully. As stated earlier, educational inequality has always been a large passion of mine due to the poor socioeconomic and demographic background I was brought up in that led me to learn in an impoverished education system. Furthermore, while I have always been able to push harder to reach higher levels of success in life, I realize that not all students can obtain such milestones due to the various external factors that prevent them from obtaining educational equality. After finding this dataset from Royce Kimmons, I have been able to understand having lunch and taking a test preparation course have major benefits for the success of students on examinations. It is my hope that others with more power over begetting change in the education system will examine the models I have created to work to create a stronger educational system for the future.

**Acknowledgements:** (clearly acknowledge people that helped you finish the project and the web resources you used)

I did not have a single person help me to finish the project: all work was completed by myself. The web resources that I used as examples and as sources of information for new concepts are listed here:

1. <https://www.kaggle.com/spscientist/students-performance-in-exams/tasks>
2. <http://roycekimmons.com/tools/generated_data/exams>
3. <https://www.kaggle.com/ritikpnayak/regression-model-1-the-best-fit-line-with-r2-0-9>
4. <https://www.kaggle.com/josephchan524/studentperformanceregressor-rmse-12-26-r2-0-26>
5. <https://www.globalcitizen.org/en/content/9-facts-about-education/>
6. <https://www.habitatbroward.org/benefits-of-education/#:~:text=It%20helps%20people%20become%20better,rights%2C%20laws%2C%20and%20regulations>.
7. <https://files.eric.ed.gov/fulltext/EJ1151836.pdf>
8. <https://peer.asee.org/regression-models-for-predicting-student-academic-performance-in-an-engineering-dynamics-course.pdf>
9. <https://www.kdnuggets.com/2020/01/beginners-guide-nearest-neighbors-r.html>
10. <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NeighborhoodComponentsAnalysis.html>
11. <https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a>