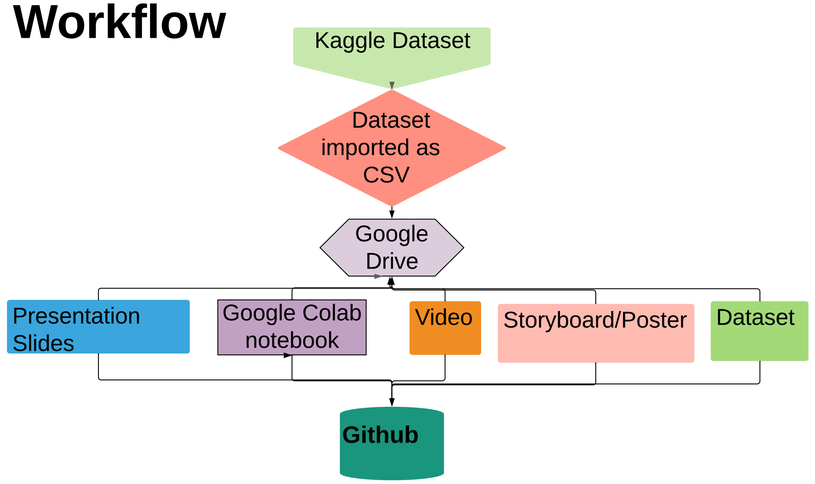
High Score Test Scores

DH 100 Theory and Methods | Channing Lee | 5/30

Workflow



Workflow description: The Dataset was first found on Kaggle and was imported as a CSV into my Google Drive. With the Dataset in my Google Drive, I was able to create the Presentation Slides, Google Colab Notebook, Video, and Storyboard/Poster. All of these, including my dataset, were pushed into Github for public access.

Descriptions:

This Dataset has 8 columns. It includes data on gender, race ethinicity, parent level of eduction, lunch, test preparations, and test score results for math, reading, and writing. I created additional columns such as total score and a percent correct.

I will be using Machine learning methods to figure out which factors affect the test score results the most. I will try methods such as linear regression, random forest, lasso regression, ridge regression, and principal components regression.

Dataset: Student Performance in Exams

Link: https://www.kaggle.com/spscientist/students-performance-in-exams

Questions I will be addressing:

Main question: How can we increase the scores of the high school students?

Sub questions:

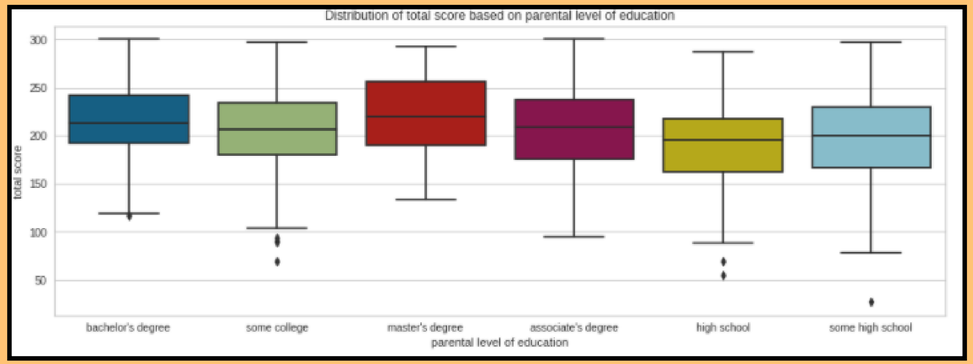
1)How does factors such as gender, race/ethnicity parental level of education, and etc affect student test performance?
2)What factors affect math score the most and the least?
3)What factors affect writing score the most and the least?
4)What factors affect reading score the most and the least?
5)Do the factors that affect each individual score (math, writing, and reading) the most have the same impact on the total score?

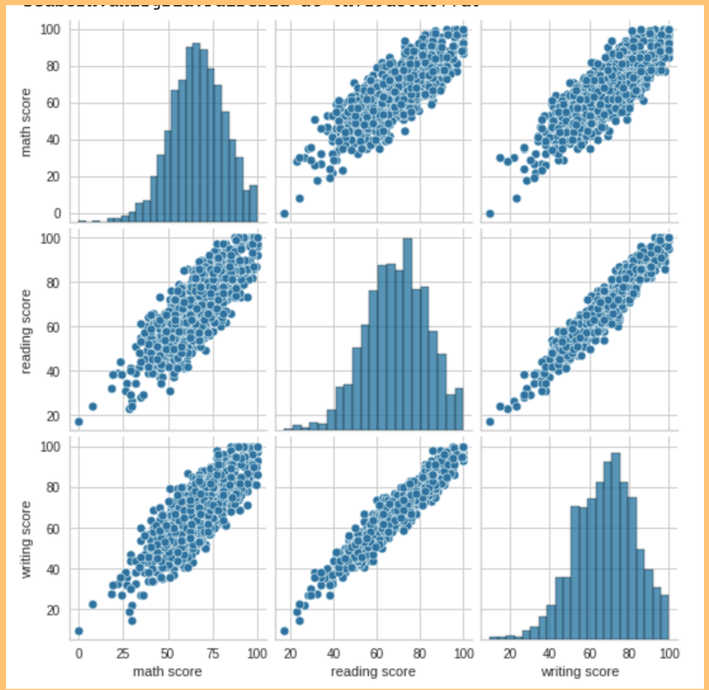
Introduction / Why I chose this dataset

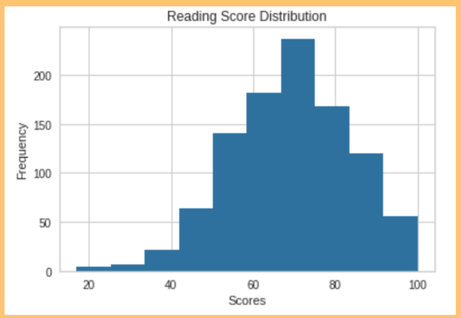
High School is a pivotal time in a student's life and will often determine the path a student wil take towards their future career through attending college. However, college admissions are not easy and often require high gpa and test scores to be admitted. When I was a high school student and even as a college transfer, I researched how to get into prestigious universities and watched countless videos of high schoolers getting admitted. Now, as a UC Berkeley graduate, I mentor High Schools students to get into prestigious colleges and prepare them for the next step in their academic journey. I chose this dataset because it appealed to the High Schooler/ Transfer in me and I wanted to examine how these factors affect standardize testing performance and apply them to the students I mentor.

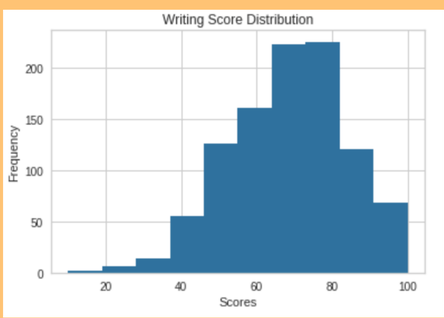
1) High School Test Scores
2) DigHum 100 Theory & Methods in the Digital Humanities
3) Dr. Adam Anderson
4) Channing Lee

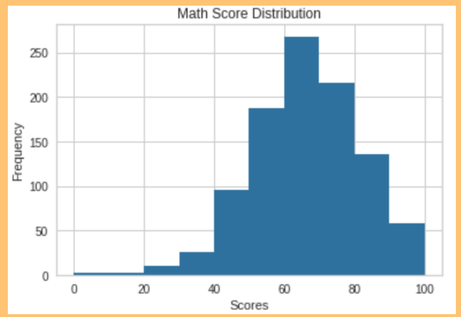
Early Exploratory Graphs to understand the stucture of the Data

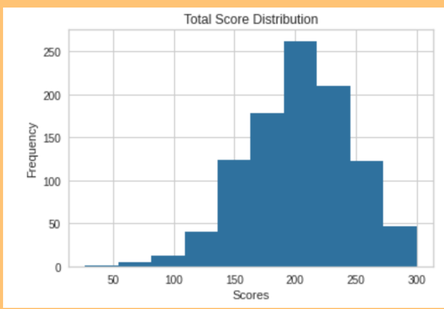












Discussion of results/ Interpretation

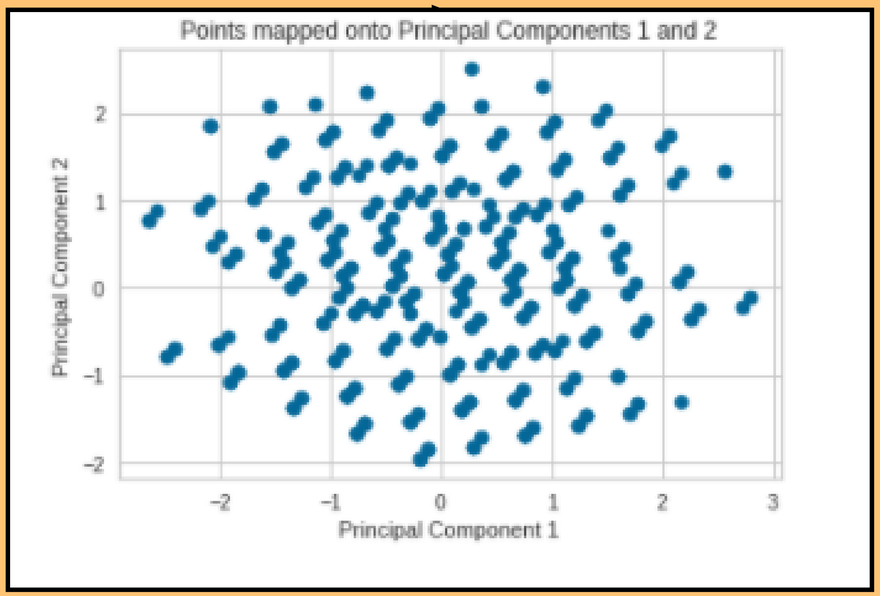
Linear Regression:
Math Score: From the analysis we could see that the Parents level of eduction had the least amount of impact on Math scores while whether the students had lunch or not made the most amount of difference.
Reading Score: For reading score, gender had barely any impact at all while lunch and test preparation mattered the most
Writng Score: Gender had a negative impact on the scores and Once again Preparation had the most impact on Writing Performance.

Ridge Regression:
Math Score: Parental level of education had the least amount of impact while lunch had the most amount of impace
Reading Score: Gender had a negative impact on Reading and test preparation mattered the most
Writing Score: Gender also had a negative impact on Reading and Test preparation mattered the most.

Lasso Regression:
Math Score: Parent level of education had the least impact. Lunch had the most impact
Reading Score: Gender had a negative impact, Test prep most impact
Writing Score: Gender had negative impact, test prep most impact

Random Forest
Instad of using Random forest for classificiation, I used it for regression to find predictors of my model. Running Random Forest produced subpar results. Using Root Mean Squared Error as a metric, running Random Forest resulted in a lower Root Mean Squared Error than all the other methods.

PCA:
I was able to find split my dataset into components to reduce the deimisnsion of the data. While plotting the the points on the first and second components of principal components, I found a pattern as seen on the left with the points forming three separate planes. This made it apparent to me that the dataset was fabricated and not real.



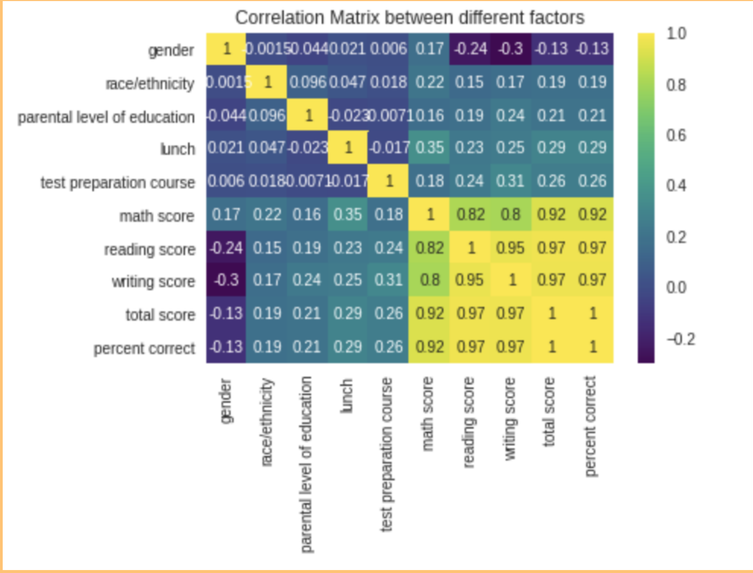
Predictor Coefficients

	Predictor	Linear math coefficients	Ridge math coefficients	Lasso math coefficients
0	gender	8.450756	5.903078	5.935474
1	race/ethnicity	3.483269	2.645235	2.646150
2	parental level of education	2.375814	1.691194	1.694456
3	lunch	13.534315	10.445933	10.514000
4	test preparation course	9.247094	6.472872	6.516533

	Predictor	Linear total coefficients	Ridge total coefficients	Lasso total coefficients
0	gender	6.145644	-8.284617	-8.340255
1	race/ethnicity	11.044390	6.219324	6.222640
2	parental level of education	9.957712	6.026474	6.033271
3	lunch	42.372528	24.805696	24.976868
4	test preparation course	41.498888	25.602848	25.772400

	Predictor	Linear reading coefficients	Ridge reading coefficients	Lasso reading coefficients
0	gender	0.295253	-6.030107	-6.067670
1	race/ethnicity	3.787759	1.668664	1.669694
2	parental level of education	3.601566	1.876152	1.877670
3	lunch	14.539285	6.849392	6.897704
4	test preparation course	15.256613	8.293980	8.348255

	Predictor	Linear writing coefficients	Ridge writing coefficients	Lasso write coefficients
0	gender	-2.600365	-8.157588	-8.208036
1	race/ethnicity	3.773361	1.905426	1.906679
2	parental level of education	3.980331	2.459127	2.461037
3	lunch	14.298928	7.510371	7.564256
4	test preparation course	16.995180	10.835996	10.906709



Root Mean Squared Error Table (Used to compare model performance)

		Math Score		
	Linear Regression	Ridge Regression	Lasso Regression	Random Forest
RMSE	13.630749359091372	13.072340055562838	13.074523665318557	15.111820053140164

		Reading Score		
	Linear Regression	Ridge Regression	Lasso Regression	Random Forest
RMSE	15.427214220485427	12.88787827308032	12.887376769191185	14.861138065798864

		Writing Score		
	Linear Regression	Ridge Regression	Lasso Regression	Random Forest
RMSE	14.717378952128481	12.637762224086966	12.63699088964377	14.471262075423892

		Total Score		
	Linear Regression	Ridge Regression	Lasso Regression	Random Forest
RMSE	42.3533256099155	37.3823489987393	37.383579282526064	42.991653577151006

Conclusions/Further steps

By looking at the coefficients for each response variable we are able to see that lunch and test preparation have the largest impact on test scores. To put my results into action I would recommend the school to offer lunch to students that may not have access to it. This would signficantly inhance a student's cognitive ability and allow them to focus in class resulting in high test scores. Another recommendation would be to put more time into standardize test preparation. Though standarize testing is dying out, it is almost impossible to receive admission to a top ranked school without a good ACT or SAT score. By providing students with a class period every week, students will be prepared for the test and have a bright future ahead of them.

Some other future steps I would like to try in the future is binning the responce variables so I can run Random Forest classification instead of regression. Often, Random Forest classification out performs Linear, Lasso, and Ridge. However, it did not out perform the rest.

Upon learning that my dataset is fake, I learned from the professor that it was probably used as a project proposal for research. Since I am interested in attending a Graduate School in Data Science, I may try to use /build off this dataset and build intituation on factors that really affect test score performance.

Work Cited

1) High School Test Scores
2) DH 100
3) Instructor: Dr Anderson
4) Student: Channing Lee.

Work cited:
Seshapanpu, Jakki. “Students Performance in Exams.” *Kaggle*, 9 Nov. 2018, www.kaggle.com/spscientist/students-performance-in-exams

Link to Github Repository:
https://github.com/earthimmortal/DigHum100