Data Analytics Project | Ibekwe Ebere

FACTORS THAT DETERMINES THE STUDENTS' ACADEMIC PERFORMANCE IN AN EXAMINATION

Data: Student performance in exam (Excel)

Introduction

A Pupil's academic performance is an essential part in higher learning institutions. This is because one of the criteria for a high-quality university is based on its excellent record of academic achievements. This project is an analysis of the performance of students in a particular examination. I intend to conduct a descriptive analysis of the data set. Descriptive Analysis will enable us to present simple summaries about the sample, the variables, and the observations with the sample. Such summaries could be quantitative, i.e., summary statistics, or visual, which are, simple-to-understand pie charts, histograms, or bar graphs. After which, I will conduct an analysis of association where I will investigate how factors like gender, race, parental levels of education, levels of test preparation influence students' performance. This analysis will be useful to students, educators, parents, and academic institutions to help improve students' achievement and success more effectively in an efficient way.

- ***Dataset***

Our data describe students' performance in an examination. The data is secondary sourced from the internet at www.kaggle.com. The data attributes include student grades which are: math scores, writing scores, reading scores. Also, our data set has other attributes which include demographic, and social characteristics i.e., gender,

race/ethnicity, parental level of education. The dataset also consists of information on other factors that might have an influence on students' performance like if they are provided with lunch at school, and their participation in the preparation course.

- **Features**

	A	B	C	D	E	F	G	H
1	Gender	Race/ethnicity	Parental level of education	Lunch	Test preparation cour	Math score	Reading score	Writing score
2	male	group A	high school	standard	completed	67	67	63
3	female	group D	some high school	free/reduced	none	40	59	55
4	male	group E	some college	free/reduced	none	59	60	50
5	male	group B	high school	standard	none	77	78	68
6	male	group E	associate's degree	standard	completed	78	73	68
7	female	group D	high school	standard	none	63	77	76
8	female	group A	bachelor's degree	standard	none	62	59	63
9	male	group E	some college	standard	completed	93	88	84
10	male	group D	high school	standard	none	63	56	65
11	male	group C	some college	free/reduced	none	47	42	45
12	male	group E	some college	standard	completed	99	83	85
13	female	group D	high school	standard	completed	80	87	90
14	male	group D	associate's degree	standard	completed	77	87	85

Screenshot of the raw

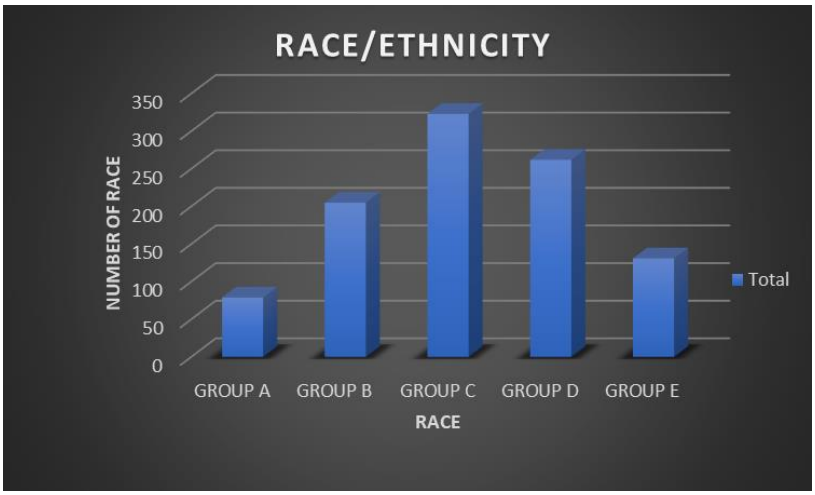
This sample is from a population of students who took a particular examination. It is not a random sample as it is representative of our population of interest, which are students of a particular grade who took a particular examination. The data set consists of 1000 observations with 8 variables. There are 5 categorical variables and 3 numerical variables. Using counta I observed that there was no empty cells in all the columns.

FEATURES OF VARIABLES

❖ RACE/ETHNICITY

This variable informs us about the race or ethnicity of the students who participated in the examination. It is a nominal variable. It comprises 5 categories. The group C has the highest number student (323) while group A has the lowest(79). The chart shows the distribution of the data based on race/ ethnicity.

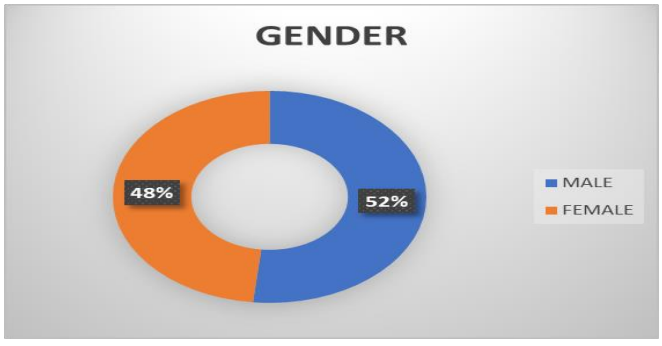
Row Labels	Count of race/ethnicity
GROUP A	79
GROUP B	205
GROUP C	323
GROUP D	262
GROUP E	131



❖ GENDER

The gender variable is a qualitative variable that has both female and male. About 52% of the students are 48% female, while 52% are male. The doughnut chart shows the composition.

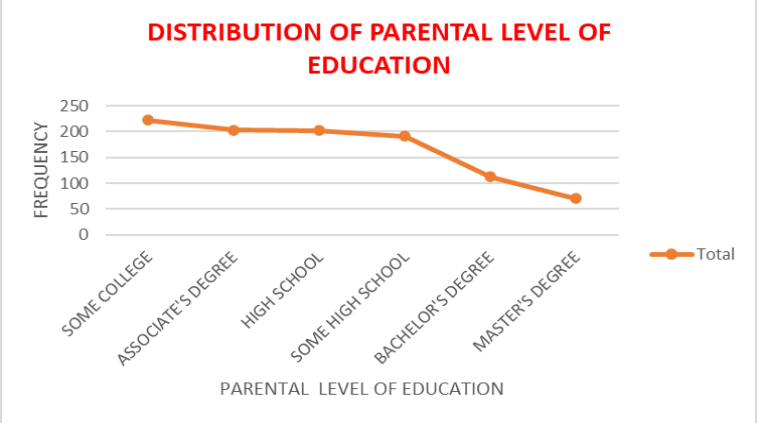
Row Labels	Count of gender
MALE	517
FEMALE	483



❖ Parental level of Education

The parental level of education of parents of the students who participated in this examination has five possible outcomes from my dataset. The line chart shows the distribution of the data base on the parents' level of education. Parents with Some 'college degree' has the highest number of 222 while master's degree has the lowest number.

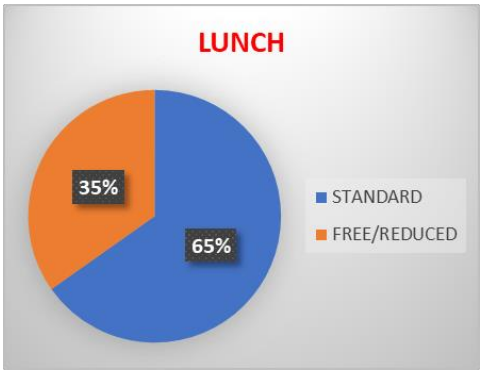
Row Labels	Count of parental level of education
SOME COLLEGE	222
ASSOCIATE'S DEGREE	203
HIGH SCHOOL	202
SOME HIGH SCHOOL	191
BACHELOR'S DEGREE	112
MASTER'S DEGREE	70
Grand Total	**1000**



❖ Lunch

The 2 levels of the lunch variable(dichotomous variable) are standard and free/reduced. 65% of the students that participated in the exams had standard lunch, while 35% had free/reduced lunch.

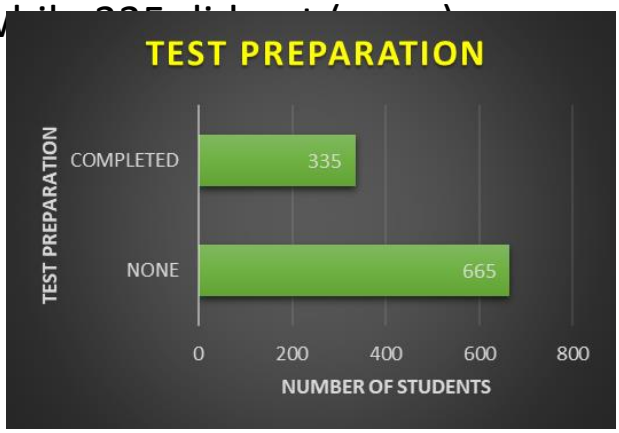
Lunch	Count of lunch
STANDARD	652
FREE/REDUCED	348
Grand Total	**1000**



❖ Test Preparation

The 2 levels of the test preparation are none and completed. 665 of the students that participated in the exams completed their test, w

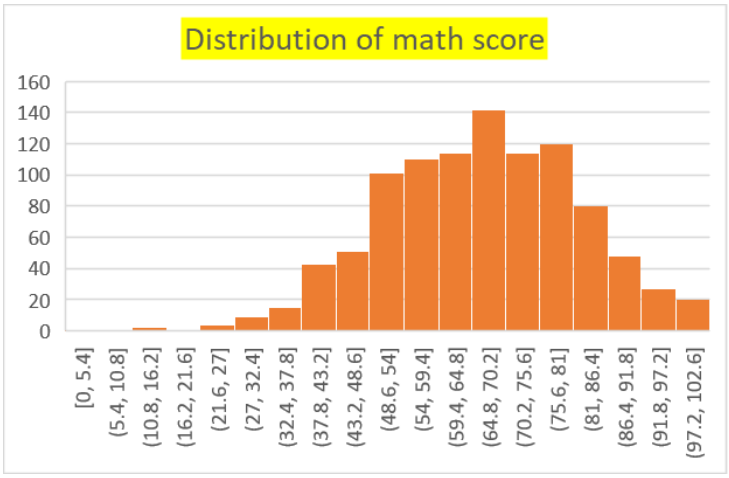
Test Preparatio	Count of test preparation course
NONE	665
COMPLETED	335
Grand Total	**1000**



❖ Math Score

This variable ranges from 0 to 100. From my five number summary in the table below, I observed that the mean math score is 66.396 ,median is 66.5, and mode is 63,maximum is 100 and minimum is 13 .The distribution of the Math Score Variable is left-skewed.

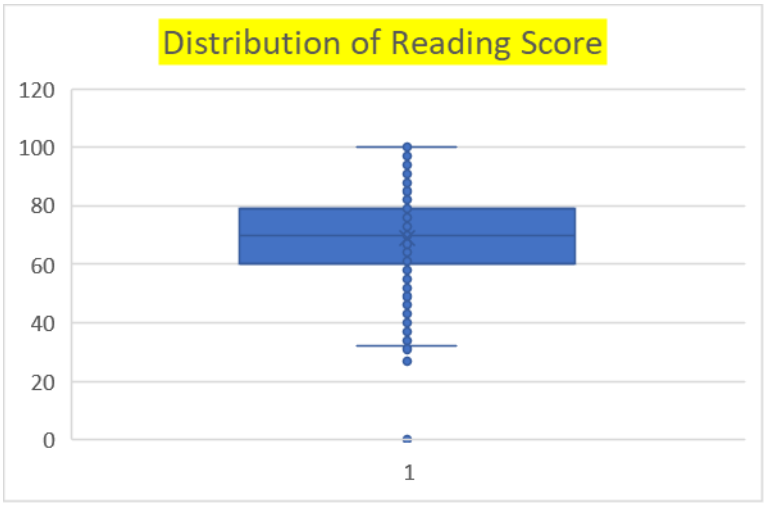
Descriptive summary of math score	
Mean	66.396
Standard Error	0.487082
Median	66.5
Mode	63
Standard Deviation	15.40287
Sample Variance	237.2484
Kurtosis	-0.22485
Skewness	-0.15115
Range	87
Minimum	13
Maximum	100
Sum	66396
Count	1000



❖ Reading Score

The distribution of the Reading Score is left-skewed. From our variable summary table, I discovered that the mean reading score is 69. The spread of the distribution is 14.7. The median score is 70 and the mode is 71.

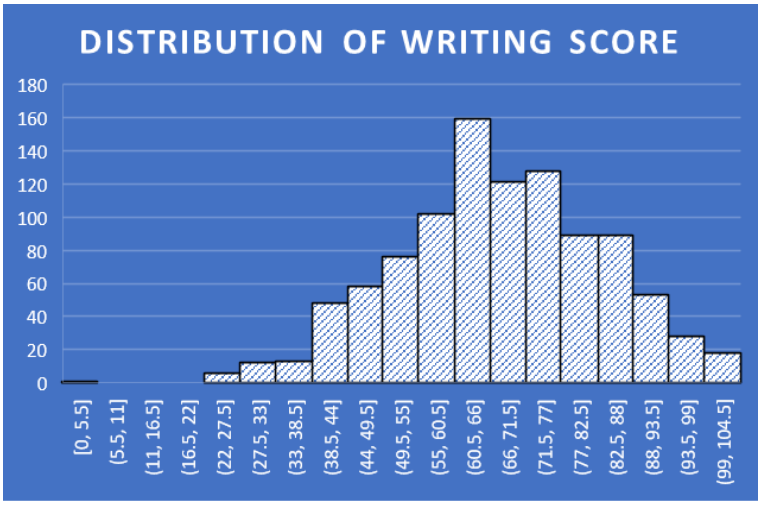
Descriptive summary of the Reading score	
Mean	69.002
Standard Error	0.466033
Median	70
Mode	71
Standard Deviation	14.73727
Sample Variance	217.1872
Kurtosis	-0.29252
Skewness	-0.19166
Range	73
Minimum	27
Maximum	100
Sum	69002
Count	1000



❖ Writing Score

This variable ranges from 0 to 100. The distribution is left-skewed. From our variable summary table, I discovered that the mean writing score is 67.7. The spread of the distribution is 15.6, median score is 68 and mode is 71 .

Descriptive summary of writing score	
Mean	67.738
Standard Error	0.493346
Median	68
Mode	71
Standard Deviation	15.60099
Sample Variance	243.3907
Kurtosis	-0.34857
Skewness	-0.15362
Range	77
Minimum	23
Maximum	100
Sum	67738
Count	1000



RELATIONSHIIP BETWEEN QUALITATIVE AND QUANTITATIVE VARIABLES

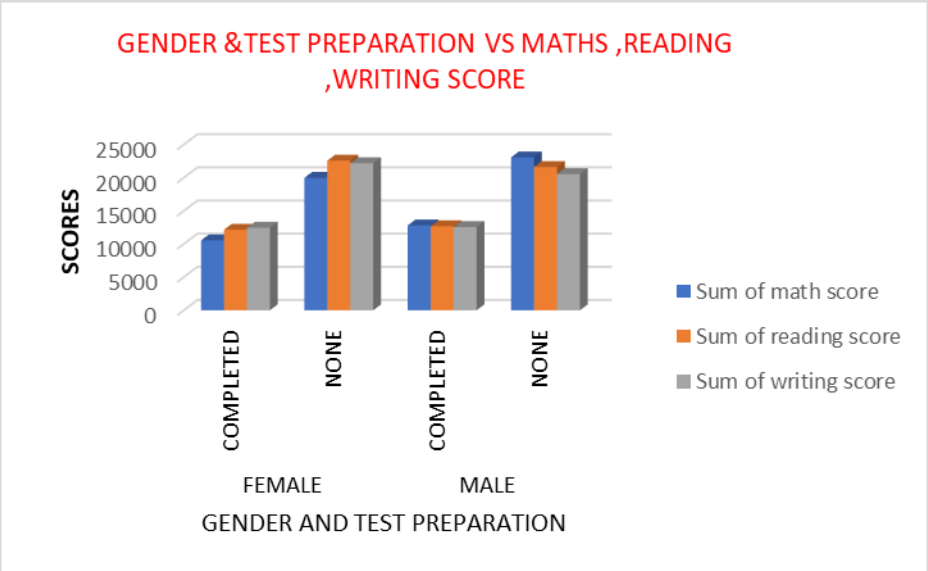
Pivot Charts were used to determine :

- **Gender and Test preparation versus math, reading and writing score.**

From the pivot table and pivot chart, the male students have the score is highest score (23070) under none for test preparation. Under test preparation (none) , girls have the highest reading score of 22571. The girls also have the highest writing score of 22178.

This implies that the boys academic performance in math is higher than the girls while the girls performed excellently in their reading and writing score.

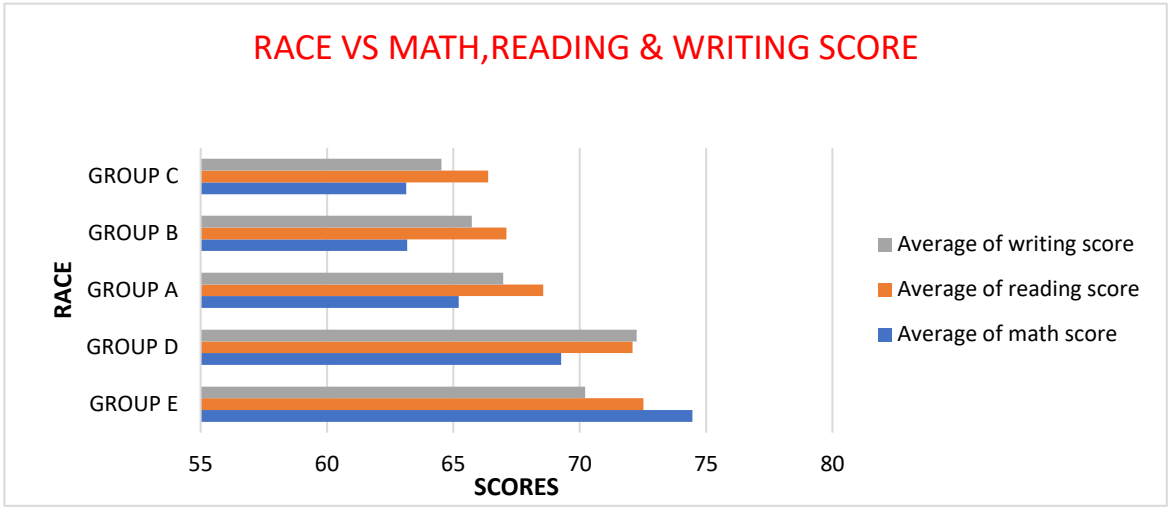
Gender/Test preparation	Sum of math score	Sum of reading score	Sum of writing score
FEMALE	**30524**	**34722**	**34635**
COMPLETED	10543	12151	12457
NONE	19981	22571	22178
MALE	**35872**	**34280**	**33103**
COMPLETED	12802	12669	12558
NONE	23070	21611	20545
Grand Total	**66396**	**69002**	**67738**



❖ Race/Ethnicity Vs Math Score, Reading Score, and Writing Score

Ethnicity group E on average perform best in math, reading, and writing when compared to other ethnicity groups. Group C has the lowest in math, reading and writing score.

Race	Average of math score	Average of reading score	Average of writing score
GROUP E	74.46564885	72.52671756	70.21374046
GROUP D	69.26717557	72.08778626	72.25954198
GROUP A	65.21518987	68.55696203	66.97468354
GROUP B	63.17073171	67.10243902	65.73658537
GROUP C	63.13003096	66.38390093	64.52321981
Grand Total	**66.396**	**69.002**	**67.738**

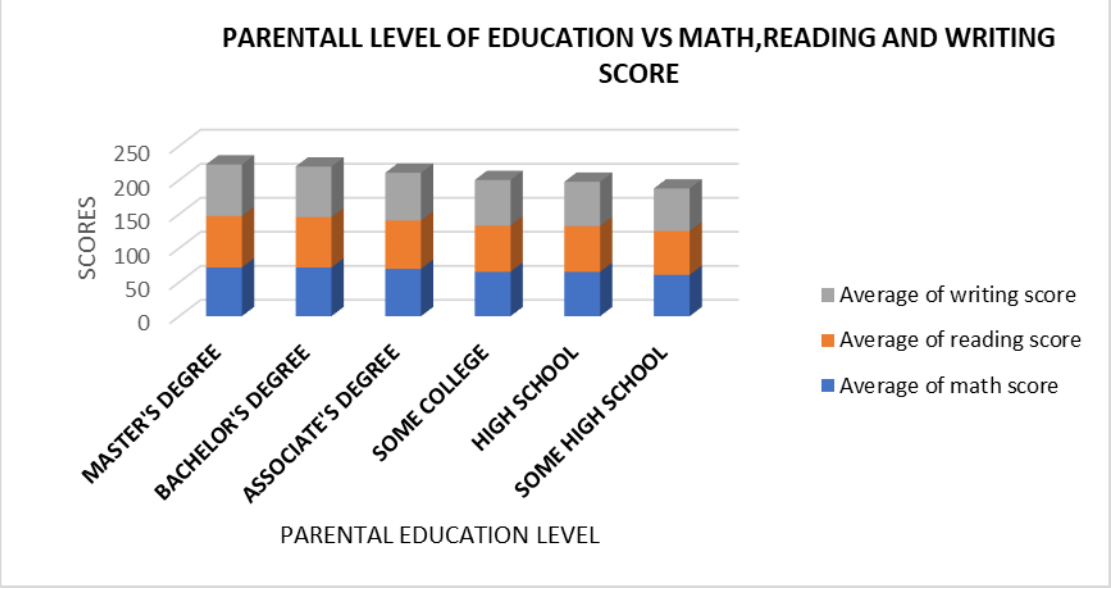


❖ Parental level of education VS math. reading and writing score

The math, reading, and writing score increases as the parents' level of education increases. The student whose parents have master's degree obtain the highest average score in the 3 subjects. While the parents with some high school level of education have the lowest scores.

This implies that the students whose parents have higher level of education pay more attention to their children's academic performance .

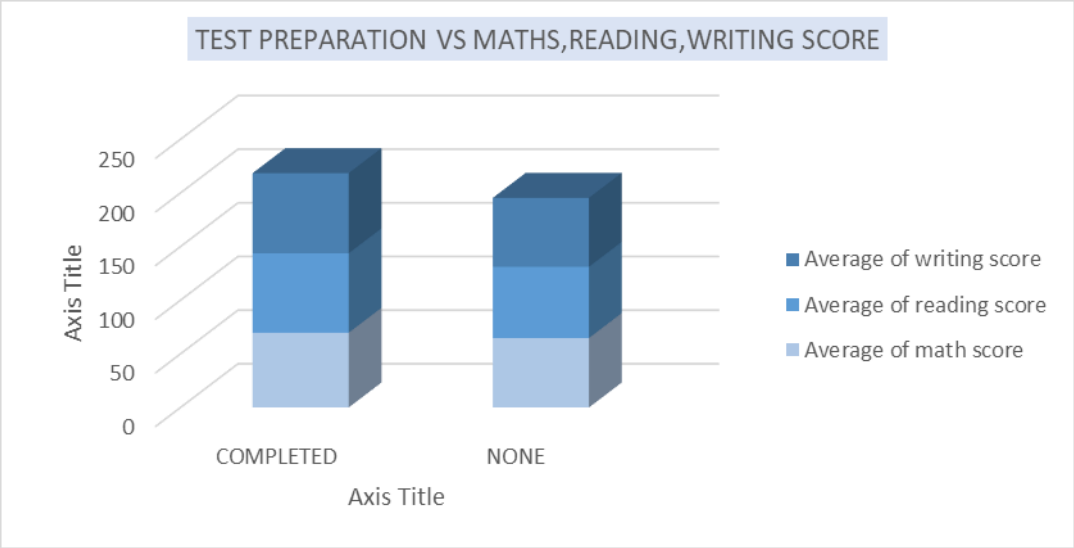
Parents education level	Average of math score	Average of reading score	Average of writing score
MASTER'S DEGREE	71.58571429	75.42857143	75.88571429
BACHELOR'S DEGREE	71.49107143	74.00892857	74.41071429
ASSOCIATE'S DEGREE	69.53694581	70.98522167	70.11330049
SOME COLLEGE	65.2972973	68.04504505	66.73423423
HIGH SCHOOL	65.20792079	67.4009901	64.84653465
SOME HIGH SCHOOL	60.70157068	64.40837696	62.53926702
Grand Total	**66.396**	**69.002**	**67.738**



❖ TEST PREPARATION VS MATH, READING,WRITING SCORE

The test preparation affected the outcome of the average math and reading score. Student who completed the test preparation course score higher on average than the student who did not complete the test preparation.

Test preparation	Average of math score	Average of reading score	Average of writing score
COMPLETED	69.68656716	74.08955224	74.67164179
NONE	64.73834586	66.43909774	64.24511278
Grand Total	**66.396**	**69.002**	**67.738**

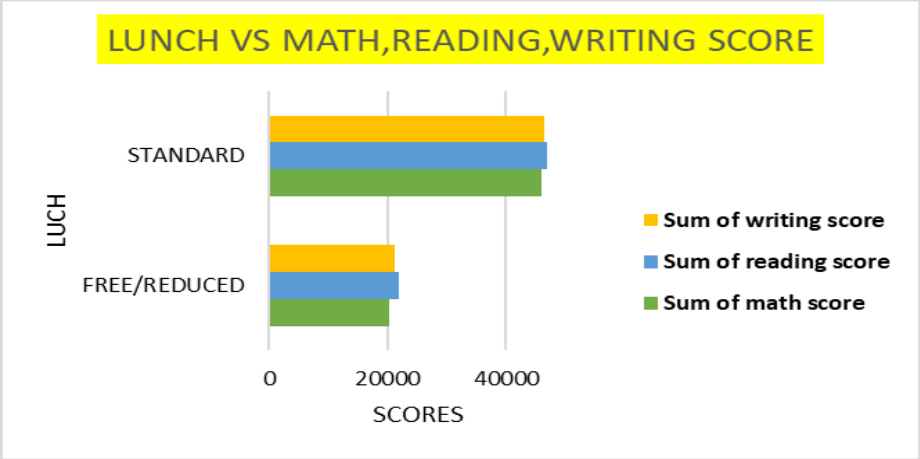


❖ LUNCH VS MATH, READING,WRITING SCORE

The student who had standard lunch scored higher than the other students who ate free/reduced lunch in all the subjects. The significant difference can be seen on Math score.

This shows that a complete diet influenced the students' academic performance.

lunch /test prep	Sum of math score	Sum of reading score	Sum of writing score
FREE/REDUCED	20360	21990	21202
STANDARD	46036	47012	46536
Grand Total	**66396**	**69002**	**67738**



❖ Correlation of math, reading and writing score

Correlation	*math score*	*reading score*	*writing score*
math score	1		
reading score	0.819397545	1	
writing score	0.805944439	0.954274434	1

In this project, I discovered that the numeric values of the three subjects are strongly positively related to each other. The larger the absolute value of correlation points signify a strong relationship between the three variables, while a low correlation means that the variables are weakly related. It also shows that when there is a great improvement in reading and writing score the math score increases.

Also, this tells us that a student who does well in math will most probably also do well in the other two subjects, while a student who fails at math will most probably also fail at

writing and reading. It shows I can use the math score as my dependent variable Y in representing students' performance.

❖ Regression Analysis

Reading and writing score is the independent variable which is the predictor or explanatory while Math score is my target or response.

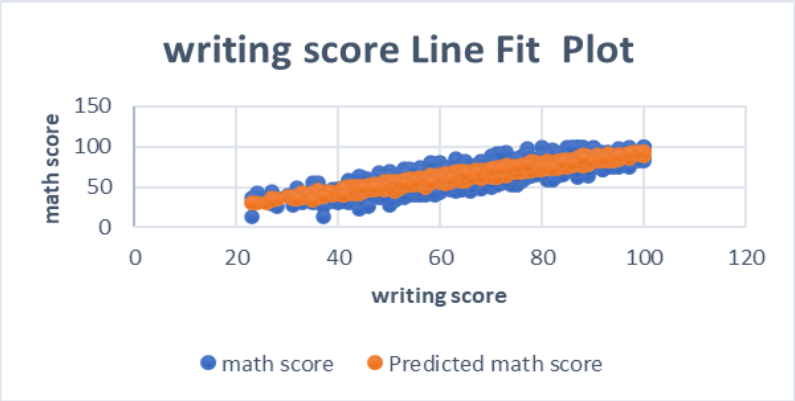
Standard error of 4 shows the certainty of my regression because of the smaller value the smaller the value the more certain.

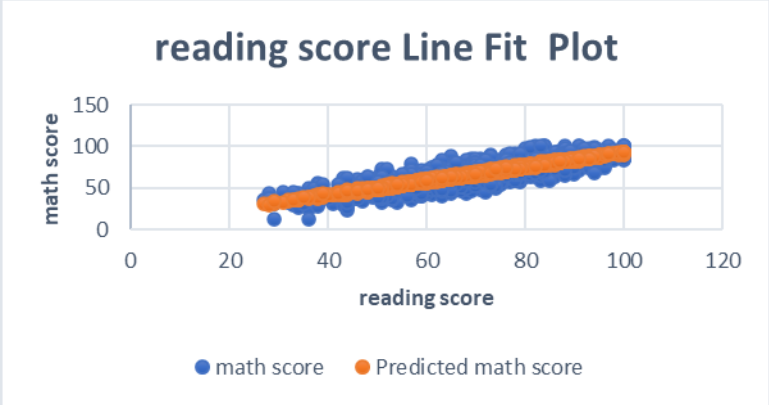
R square as the sum of squares deviation from thee mean. 91% fits my regression analysis model. It signifies that 91% of the dependent variable are explained by the independent variable (x). It is considered a good fit for my model. It shows that the students with good writing skills and reading skill performed better in math.

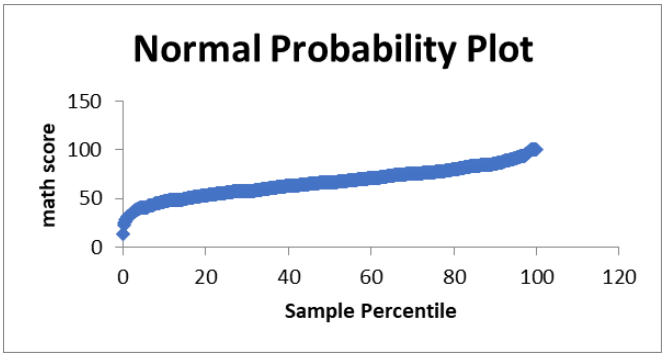
The significance f is less than 5% which shows that our dependent and independent variable were rightly choosed.

Regression of math score	Multiple R	R Square	Adjusted R Square	Standard Error	Observations			
	0.955207292	0.912420971	0.91224511	4.623651579	999			
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	221832.2436	110916.1218	5188.292786	0			
Residual	996	21292.64131	21.37815392					
Total	998	243124.8849						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-2.504400202	0.710766994	-3.523517867	0.000445199	-3.89917284	-1.109627565	-3.89917284	-1.109627565
Reading Score	0.074109481	0.016568762	4.47284365	8.6067E-06	0.041595794	0.106623169	0.041595794	0.106623169
Writing Score	0.946707594	0.017317226	54.66854861	4.0165E-302	0.91272516	0.980690028	0.91272516	0.980690028

RESIDUALS FOR PREDICTING STUDENTS SCORES







❖ Backward Multiple Regression Model:

The dependent variable is Math Score, and the explanatory variables are Reading Score, Writing Score, whether test preparation course was taken or not, gender dummy (male dummy), Parents' education level -- Some High School, High School, Some College, associate degree, bachelor's degree and master's degree.

I chose a backward regression model where all explanatory variables were first included into the regression table. I then used p values to determine in which order the explanatory

variables should be removed from the model. The elimination decision was made using decreasing order of the p values which can be observed in the tables below:

Regression of mathscore	Multiple R	R Square	Adjusted R Square	Standard Error	Observations
	0.92561068	0.85675513	0.85530675	5.85903122	1000
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	10	203060.5479	20306.05479	591.5261237	0
Residual	989	33950.63612	34.32824684		
Total	999	237011.184			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-4.862752586	0.998186892	-4.871585295	1.28855E-06	-6.8215601	-2.90394504
reading score	0.289836834	0.044481497	6.515896599	1.14899E-10	0.20254788	0.377125791
writing score	0.64808506	0.045012541	14.39787768	8.23424E-43	0.559754	0.736416119
male dummy	12.887051	0.39739566	32.42876636	9.8729E-158	12.1072155	13.66688654
completed	-4.203599003	0.432252328	-9.724873011	2.07242E-21	-5.0518361	-3.35536193
standard	3.31482641	0.416798882	7.953059741	4.95077E-15	2.49691465	4.132738168
high school	0.573606783	0.595296244	0.963565265	0.335499464	-0.594582	1.741795613
some college	-0.341604263	0.585444363	-0.583495691	0.559692694	-1.4904601	0.807251575
bachelor's degree	-0.438541275	0.721465073	-0.607848241	0.54342768	-1.8543195	0.977236915
master's degree	-0.786067184	0.846189848	-0.928948965	0.353142201	-2.446601	0.874466606
associate's degree	0.038856924	0.60767676	0.06394341	0.949028208	-1.153627	1.231340851

Parents Education associate degree dummies have very high p-values of 0.9490 here. This shows that the data is not statistically relevant.

Regression of mathscore	Multiple R	R Square	Adjusted R Square	Standard Error	Observations	
	0.92561036	0.856754538	0.855452307	5.856083471	1000	
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	9	203060.4075	22562.2675	657.9126353	0	
Residual	990	33950.77648	34.29371362			
Total	999	237011.184				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-4.854610184	0.989533316	-4.905959309	1.08644E-06	-6.79643385	-2.91278652
reading score	0.289653234	0.044366397	6.528662616	1.05845E-10	0.202590255	0.376716214
writing score	0.648445694	0.044635346	14.52762774	1.72147E-43	0.560854938	0.736036451
male dummy	12.89122037	0.391812634	32.90149236	5.229E-161	12.12234172	13.66009903
completed	-4.205249145	0.431264215	-9.750980945	1.63598E-21	-5.05154613	-3.35895216
standard	3.31293766	0.415541748	7.972574782	4.26439E-15	2.497493868	4.128381453
high school	0.553961402	0.509632563	1.086981959	0.277309357	-0.44612274	1.554045539
some college	-0.361861973	0.49207185	-0.735384421	0.462279582	-1.32748561	0.603761669
bachelor's degree	-0.460017233	0.638201034	-0.720803021	0.471200883	-1.71239939	0.792364927
master's degree	-0.807706583	0.775182269	-1.041956989	0.297685997	-2.32889566	0.713482498

Parents' Education bachelor's degree has the second highest p value; hence it will be removed next from the linear model. I noticed that the predictive power of the model (R-adjusted) moved from 0.8553 to 0.8554.

Regression of mathscore	Multiple R	R Square	Adjusted R Square	Standard Error	Observations	
	0.92556975	0.856679362	0.855522384	5.854663766	1000	
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	8	203042.59	25380.32375	740.4457428	0	
Residual	991	33968.59402	34.27708781			
Total	999	237011.184				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-4.86661682	0.989153245	-4.91998267	1.01299E-06	-6.8076923	-2.9255414
reading score	0.29248876	0.044180938	6.620247784	5.86617E-11	0.20578982	0.3791877
writing score	0.64413965	0.044223041	14.56570223	1.07875E-43	0.55735809	0.7309212
male dummy	12.8808962	0.391455819	32.90510834	4.4321E-161	12.1127186	13.649074
completed	-4.18176029	0.429927032	-9.72667448	2.0308E-21	-5.0254322	-3.3380884
standard	3.32018204	0.415319477	7.994284443	3.61106E-15	2.50517543	4.1351886
high school	0.6480978	0.492495193	1.315947464	0.188495995	-0.3183554	1.614551
some college	-0.26281284	0.472380774	-0.556358038	0.578091719	-1.1897943	0.6641686
master's degree	-0.68981821	0.757548758	-0.91059249	0.362731556	-2.1764021	0.7967657

Parents' Education some college has the third highest p value of 0.5780; hence it will be removed next from the linear model. Also, the predictive power of the model (R-adjusted) moved from 0.8554 to 0.8555

Regression of mathscore	Multiple R	R Square	Adjusted R Squar	Standard Error	Observations	
	0.92554557	0.856634597	0.855622946	5.852625889	1000	
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	7	203031.98	29004.56858	846.7688668	0	
Residual	992	33979.20396	34.25322979			
Total	999	237011.184				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-4.95835983	0.974971947	-5.085643591	4.37931E-07	-6.871604082	-3.045115581
reading score	0.29203456	0.044158019	6.613398172	6.12927E-11	0.20538071	0.378688417
writing score	0.64475571	0.044193787	14.58928408	8.04494E-44	0.558031671	0.731479757
male dummy	12.8771842	0.391262721	32.91186076	3.5759E-161	12.10938655	13.64498181
completed	-4.1885319	0.429605123	-9.749725231	1.64769E-21	-5.031571063	-3.345492743
standard	3.32865441	0.414895735	8.022869672	2.90009E-15	2.514480335	4.142828476
high school	0.72869522	0.470543421	1.548624819	0.121790886	-0.194679547	1.652069987
master's degree	-0.61281162	0.744537542	-0.823076855	0.410662104	-2.073861013	0.848237776

In the final model, male dummy, Parents' Education Master's Degree, Parents' Education high school , taken test preparation course, reading score and writing score are good linear predictors of Math Score. This means that parents' education overall is not a statistically significant variable to predict Math Score unless the parent has higher education such as a bachelor's degree or high school.

Also, p values of all variables included in this model is less than 0.05 (alpha). Please note that the reference variable here is female dummy .This would mean that, keeping Parents' Education masters degree, Parents' education high school degree, taken test preparation course, reading course and writing score values constant, a male student would score about 7.6 higher in math than a female student. The

current model shows 0.8556 predictive power and hence, is a very good model for predictive analysis.

❖ Excluding lunch from my analysis

Regression of mathscore	Multiple R	R Square	Adjusted R	Standard Error	Observations	
	0.920648659	0.847593953	0.846208	6.040430911	1000	
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	9	200889.2465	22321.03	611.7561411	0	
Residual	990	36121.93753	36.48681			
Total	999	237011.184				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-4.952989244	1.029024971	-4.813284	1.7161E-06	-6.972309877	-2.93366861
reading score	0.25388489	0.045621246	5.565058	3.3733E-08	0.164359441	0.34341034
writing score	0.721115484	0.045430254	15.87302	1.01412E-50	0.631964831	0.81026614
male dummy	13.15282626	0.408248132	32.21773	2.4646E-156	12.35169519	13.9539573
completed	-4.582222281	0.442924072	-10.34539	6.87458E-24	-5.451400136	-3.71304443
Associate's degree	-0.303641679	0.624915659	-0.485892	0.627151083	-1.529953109	0.92266975
high school	0.547776053	0.613717896	0.892553	0.372313229	-0.656561301	1.75211341
master's degree	-1.351219051	0.869306893	-1.554364	0.120417297	-3.057114821	0.35467672
some college	-0.684009028	0.601935851	-1.136349	0.256085644	-1.86522573	0.49720767
bachelor's degree	-0.750352217	0.742703048	-1.010299	0.312598871	-2.207805272	0.70710084

From this regression summary output, I see that the Adjust R-square is 0.846208. This is low, showing that removing the variable lunch decreases our predictive analysis and influences students' performance in school. I also noticed that p-values for parents' education is very low. This shows that the data is not statistically relevant. Other explanatory variable has very high p-values: gender, lunch, test prep, reading score and writing score. This confirms their statistical relevance to students' performance.

Conclusion

Our analysis of data on students' performance is mostly useful to help the educators and learners to improve their learning and teaching process and overall learning outcomes. This project has reviewed data showing math scores, reading scores, test preparation, lunch, parents education levels.

I found out that as the level of parental education increases, the average score for math, reading, and writing, also increases. This tells us that parents' educational levels directly influence the performance of kids. I conclude that education provides the knowledge and skills necessary for students to advance themselves and their nation economically. Socioeconomic factors, such as family income level, parents' level of education, race and gender, all influence students' performance, and the quality of their education as well as how students can take advantage of this education to improve their lives. As per our analysis with backward regression, I started with relevant explanatory variable, which affects the performance of math score, I chose parental education, gender, reading score, writing score, whether test preparation score was considered or not. I started with the above factors, I considered p-values to determine whether the variable is statistically significant or not by using p-values. The variable with the highest p-value will be eliminated first and the lowest p-value is a good predictor for the analysis.

Data Analyst | Ibekwe Ebere

The p-value less than alpha as 0.05 are considered as statistically significant, linear predictors of math score. In the final model, gender, whether test preparation course is taken or not, reading score, writing score, parent's higher education level (master's or high school) was found to be statistically significant predictors of math performance score. Predicting score with regression model I want to bring math score higher than 65 for the boys, so then our recommendations would be to encourage students to take preparations course, do more reading hours, writing hours and encourage female students to practice math questions as I found female students scored 7.6 points less as compared to male students to increase math score.

Thank you.