PyCitySchools Analysis and Findings

Alexis Perumal, 12/6/19
UCSD Data Science Bootcamp, HW#4 Pandas, Ex. 2

Summary

The PyCitySchools assignment involves analyzing a combined dataset of high schools within a school district, students, demographics, spending and test scores.

The assignment doesn't specify the type of requester is hoping to achieve from the analysis, but presumably it involves:

1. Derive better insights on drivers of student and school success.
2. Understand how to optimize school and student success.
3. Evaluate equity and fairness across schools.

Findings about schools and student performance are below. They provide an initial look at customer demographics and transaction patterns and suggest further analysis that can enable increased revenue and profit maximization.

Beyond the scope of the assignment is evaluation of statistically significant correlation, causality or hypothesis testing. Therefore, the analysis is descriptive and simplistic, essentially a pre-analysis that may be suggestive of areas for a deeper look.

Approach

Analysis involved creating a jupyter notebook file and using Python and Pandas to analyze the CSV dataset, generating tabular outputs.

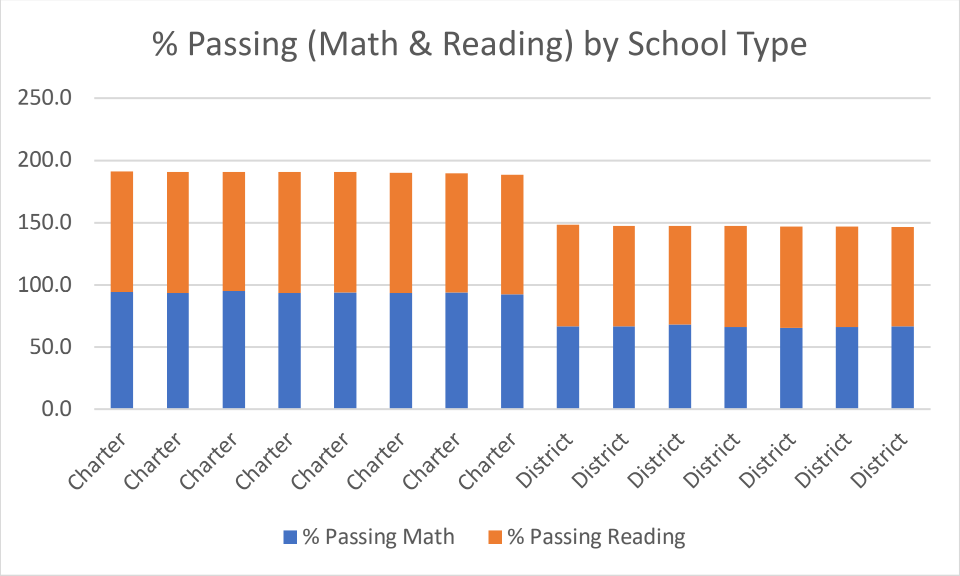
Since the simple descriptive results were boring, I exported the combined dataset to a .csv file, pulled it into Excel, and ran a regression analysis of Math Score, Reading Score, Combined Score, and corresponding passing rates (as the dependent variables) compared to school type, # of students, total school budget, per student budget.

Findings

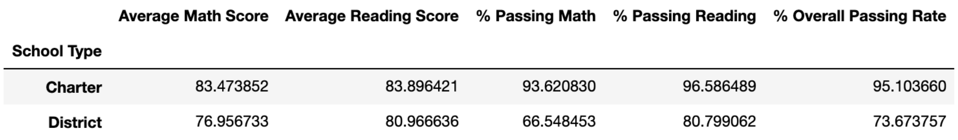
1. **Charter Schools correlate with better results much better than non-charter/district schools.** Linear regression analysis on all 6 performance measures showed statistically significant (95%+ confidence) t-stats. Combined passing rates for charter schools ranged from 94.4% to 95.6% compared to district schools from 73.3% to 74.3%. This produced a

very strong statistically significant results (t-stat of 18.2%, a 10^-9 P-value, R2=99.9%) strongly supporting rejection of the null hypothesis that school type isn't correlated.

See:

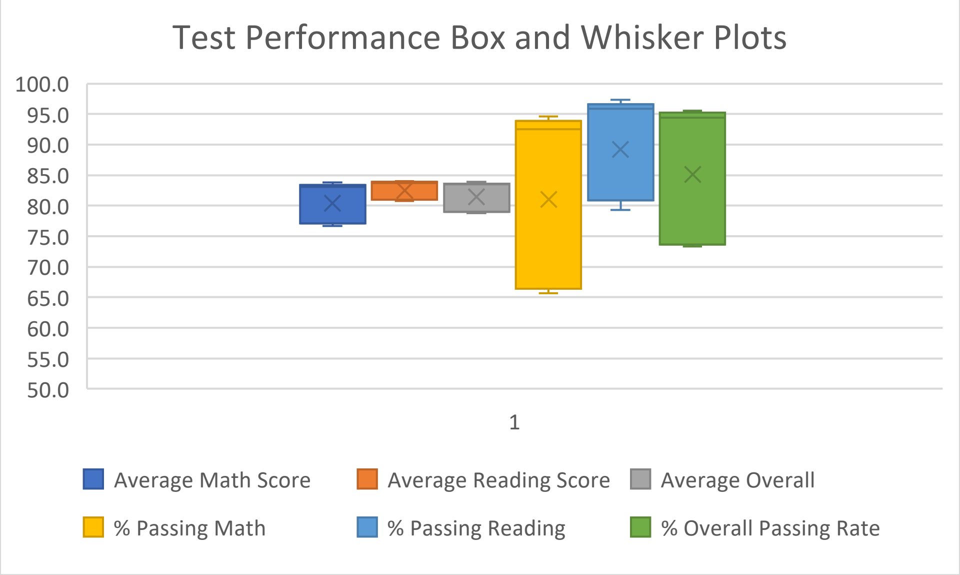


Charter schools correlate with statistically significance on all 6 indicators, but the increase is greater in math than reading, and in test pass rates over test scores.



	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
School Type					
Charter	83.473852	83.896421	93.620830	96.586489	95.103660
District	76.956733	80.966636	66.548453	80.799062	73.673757

2. Evaluating passing rates produces a much greater difference with % passing rates than raw scores.



3. **Other explanatory variables had much smaller regression coefficients and were not statistically significant for most performance indicators.**

   For math scores, total students and total school budget did have statistically significant correlations, but very small coefficients (4 pt. reduction per increase of 1k students, 5 pt. increase per $1M increase in the school budget). Per student budget did not have statistical significance suggesting that the school size and school budget factors may be conflated with each other (larger schools have larger budgets).

   For reading scores, only school type (charter, district) was statistically significant, and as indicated above, the impact was large.

   The three passing rate % metrics (math, reading, overall) had statistically significant correlations only with school type.

   Lastly, regression analysis of student performance by grade was not done, but a quick look at the results suggests minimal variation across grades. If the four grades are taking the same math and reading tests, this would suggest that student performance in these two areas doesn't change significantly across four years of high school.

Recommendations

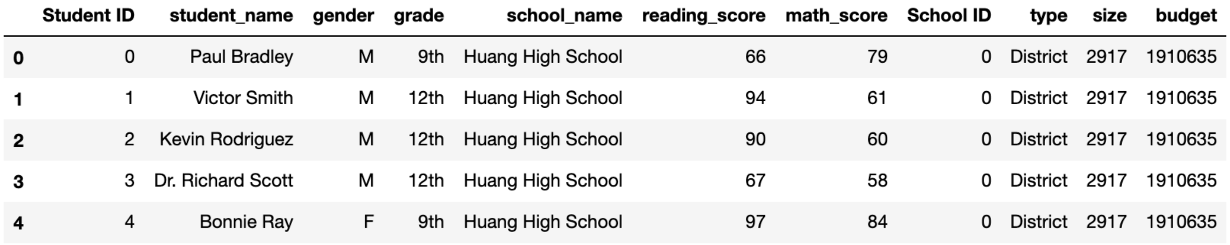
1. Probe further on the impact of charter schools. Can this be understood by other parameters not in this dataset? (Example: student family income.) Additionally, causality wasn't evaluated, but careful analysis should be considered to determine if there is a causal impact. This would require going beyond this dataset. If causality can be demonstrated, this would suggest significant public policy recommendations.

2. Explore the establishment of passing rate thresholds and if they are true and correct indicators of future student success. As described above, the 70% pass rate threshold resulted in much greater differentiation of performance across the schools than the test score percentages.

3. Deep analysis of score variation by grade should be done, particularly if the tests taken are the same across the grades. Even if the tests are different, careful analysis to understand uplift by school characteristics may lead to policy recommendations for optimizing outcomes.

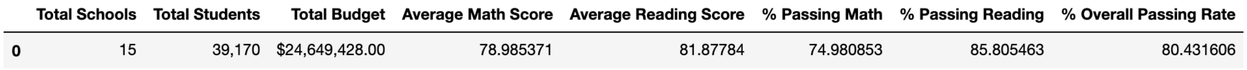
Appendix A: Required Tabular Reports (Pandas/Jupyter Notebook)

Combined Dataset (left join, head)



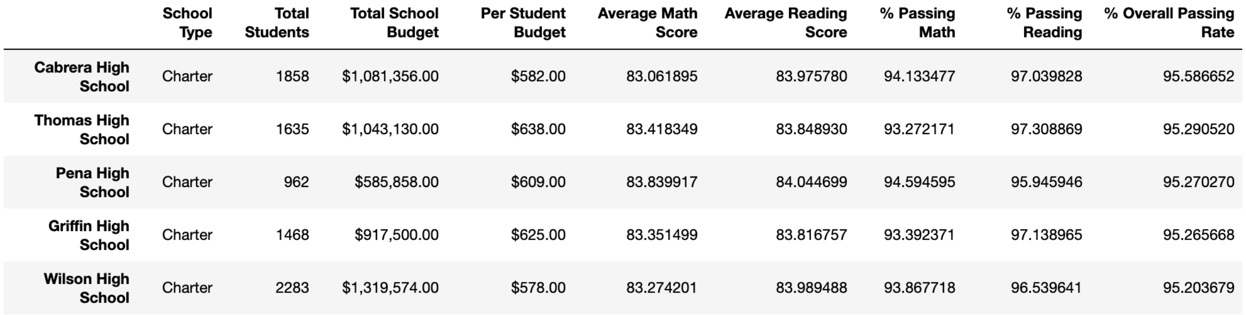
	Student ID	student_name	gender	grade	school_name	reading_score	math_score	School ID	type	size	budget
0	0	Paul Bradley	M	9th	Huang High School	66	79	0	District	2917	1910635
1	1	Victor Smith	M	12th	Huang High School	94	61	0	District	2917	1910635
2	2	Kevin Rodriguez	M	12th	Huang High School	90	60	0	District	2917	1910635
3	3	Dr. Richard Scott	M	12th	Huang High School	67	58	0	District	2917	1910635
4	4	Bonnie Ray	F	9th	Huang High School	97	84	0	District	2917	1910635

District Summary



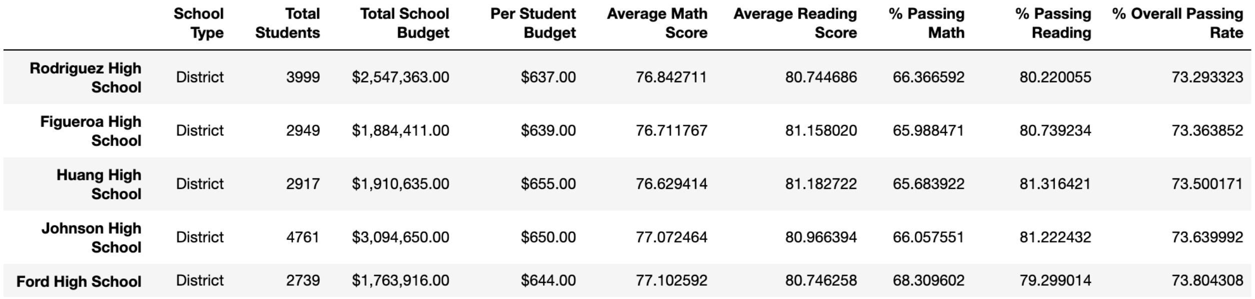
	Total Schools	Total Students	Total Budget	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
0	15	39,170	$24,649,428.00	78.985371	81.87784	74.980853	85.805463	80.431606

School Summary: Top Performing Schools (By Passing Rate)



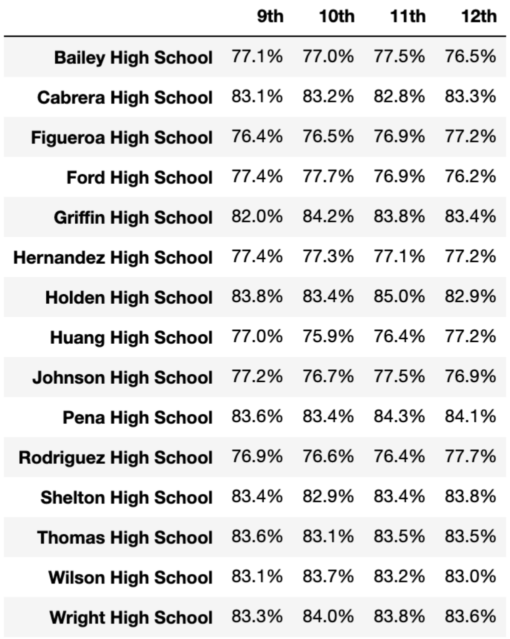
	School Type	Total Students	Total School Budget	Per Student Budget	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
Cabrera High School	Charter	1858	$1,081,356.00	$582.00	83.061895	83.975780	94.133477	97.039828	95.586652
Thomas High School	Charter	1635	$1,043,130.00	$638.00	83.418349	83.848930	93.272171	97.308869	95.290520
Pena High School	Charter	962	$585,858.00	$609.00	83.839917	84.044699	94.594595	95.945946	95.270270
Griffin High School	Charter	1468	$917,500.00	$625.00	83.351499	83.816757	93.392371	97.138965	95.265668
Wilson High School	Charter	2283	$1,319,574.00	$578.00	83.274201	83.989488	93.867718	96.539641	95.203679

School Summary: Bottom Performing Schools (By Passing Rate)



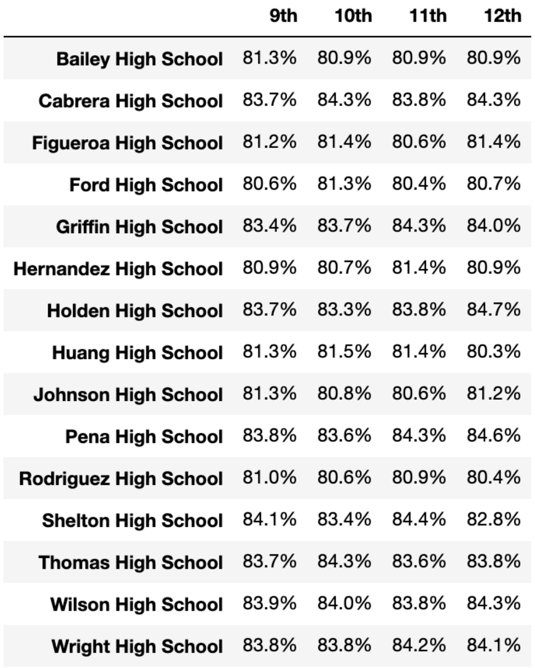
	School Type	Total Students	Total School Budget	Per Student Budget	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
Rodriguez High School	District	3999	$2,547,363.00	$637.00	76.842711	80.744686	66.366592	80.220055	73.293323
Figueroa High School	District	2949	$1,884,411.00	$639.00	76.711767	81.158020	65.988471	80.739234	73.363852
Huang High School	District	2917	$1,910,635.00	$655.00	76.629414	81.182722	65.683922	81.316421	73.500171
Johnson High School	District	4761	$3,094,650.00	$650.00	77.072464	80.966394	66.057551	81.222432	73.639992
Ford High School	District	2739	$1,763,916.00	$644.00	77.102592	80.746258	68.309602	79.299014	73.804308

Math Scores by Grade



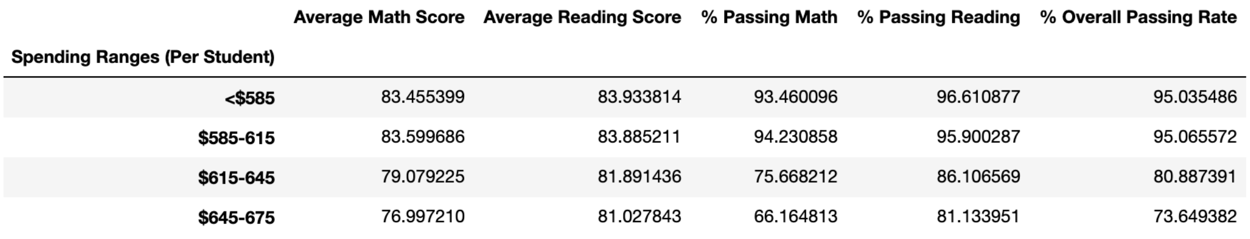
	9th	10th	11th	12th
Bailey High School	77.1%	77.0%	77.5%	76.5%
Cabrera High School	83.1%	83.2%	82.8%	83.3%
Figueroa High School	76.4%	76.5%	76.9%	77.2%
Ford High School	77.4%	77.7%	76.9%	76.2%
Griffin High School	82.0%	84.2%	83.8%	83.4%
Hernandez High School	77.4%	77.3%	77.1%	77.2%
Holden High School	83.8%	83.4%	85.0%	82.9%
Huang High School	77.0%	75.9%	76.4%	77.2%
Johnson High School	77.2%	76.7%	77.5%	76.9%
Pena High School	83.6%	83.4%	84.3%	84.1%
Rodriguez High School	76.9%	76.6%	76.4%	77.7%
Shelton High School	83.4%	82.9%	83.4%	83.8%
Thomas High School	83.6%	83.1%	83.5%	83.5%
Wilson High School	83.1%	83.7%	83.2%	83.0%
Wright High School	83.3%	84.0%	83.8%	83.6%

Reading Scores by Grade



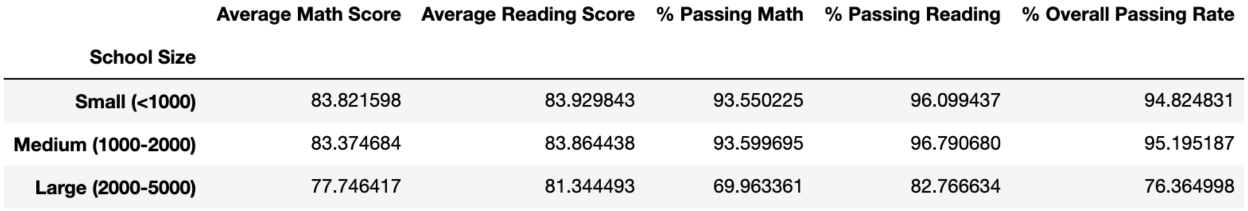
	9th	10th	11th	12th
Bailey High School	81.3%	80.9%	80.9%	80.9%
Cabrera High School	83.7%	84.3%	83.8%	84.3%
Figueroa High School	81.2%	81.4%	80.6%	81.4%
Ford High School	80.6%	81.3%	80.4%	80.7%
Griffin High School	83.4%	83.7%	84.3%	84.0%
Hernandez High School	80.9%	80.7%	81.4%	80.9%
Holden High School	83.7%	83.3%	83.8%	84.7%
Huang High School	81.3%	81.5%	81.4%	80.3%
Johnson High School	81.3%	80.8%	80.6%	81.2%
Pena High School	83.8%	83.6%	84.3%	84.6%
Rodriguez High School	81.0%	80.6%	80.9%	80.4%
Shelton High School	84.1%	83.4%	84.4%	82.8%
Thomas High School	83.7%	84.3%	83.6%	83.8%
Wilson High School	83.9%	84.0%	83.8%	84.3%
Wright High School	83.8%	83.8%	84.2%	84.1%

Scores by School Spending



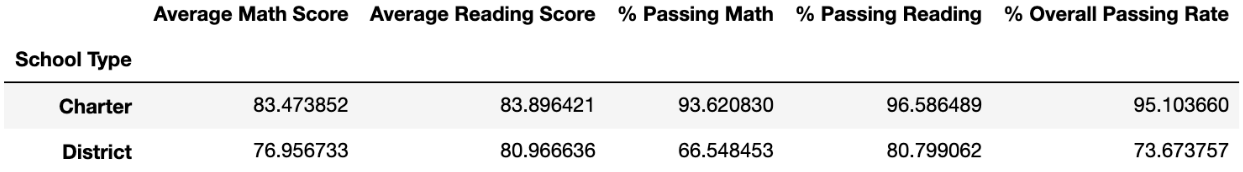
	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
Spending Ranges (Per Student)					
<$585	83.455399	83.933814	93.460096	96.610877	95.035486
$585-615	83.599686	83.885211	94.230858	95.900287	95.065572
$615-645	79.079225	81.891436	75.668212	86.106569	80.887391
$645-675	76.997210	81.027843	66.164813	81.133951	73.649382

Scores by School Size



	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
School Size					
Small (<1000)	83.821598	83.929843	93.550225	96.099437	94.824831
Medium (1000-2000)	83.374684	83.864438	93.599695	96.790680	95.195187
Large (2000-5000)	77.746417	81.344493	69.963361	82.766634	76.364998

Scores by School Type



	Average Math Score	Average Reading Score	% Passing Math	% Passing Reading	% Overall Passing Rate
School Type					
Charter	83.473852	83.896421	93.620830	96.586489	95.103660
District	76.956733	80.966636	66.548453	80.799062	73.673757

Appendix B: Regression Analysis Results

Average Math Score

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99820905
R Square	0.9964213
Adjusted R Square	0.99498982
Standard Error	0.23885938
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	158.855499	39.7138748	696.077626	3.5114E-12
Residual	10	0.57053802	0.0570538		
Total	14	159.426037			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	90.3550682	3.81775455	23.6670711	4.1156E-10	81.848581	98.8615554	81.848581	98.8615554
Type	-6.6858137	0.30785395	-21.717486	9.5772E-10	-7.3717551	-5.9998724	-7.3717551	-5.9998724
Total Students	-0.003624	0.00160886	-2.2525532	0.04797128	-0.0072088	-3.928E-05	-0.0072088	-3.928E-05
Total School Budget	5.8129E-06	2.5775E-06	2.25526108	0.04775123	6.9907E-08	1.1556E-05	6.9907E-08	1.1556E-05
Per Student Budget	-0.0111136	0.00621838	-1.7872227	0.10419665	-0.024969	0.00274178	-0.024969	0.00274178

% Passing Math

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99883753
R Square	0.99767642
Adjusted R S	0.99674699
Standard Err	0.79842551
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	2737.16408	684.291021	1073.4258	4.056E-13
Residual	10	6.374833	0.6374833		
Total	14	2743.53892			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	82.1677593	12.7614527	6.43874653	7.454E-05	53.7334707	110.602048	53.7334707	110.602048
Type	-26.737776	1.02905086	-25.982949	1.6404E-10	-29.030644	-24.444908	-29.030644	-24.444908
Total Studen	0.00652759	0.00537786	1.21378875	0.25271678	-0.005455	0.0185102	-0.005455	0.0185102
Total School	-1.044E-05	8.6156E-06	-1.2122696	0.25327243	-2.964E-05	8.7523E-06	-2.964E-05	8.7523E-06
Per Student I	0.01840853	0.02078591	0.88562514	0.39661615	-0.0279054	0.06472242	-0.0279054	0.06472242

Average Reading Score

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99591857
R Square	0.99185381
Adjusted R Square	0.98859533
Standard Error	0.16225316
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	32.0537842	8.01344605	304.391835	2.1378E-10
Residual	10	0.26326087	0.02632609		
Total	14	32.3170451			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	83.6083986	2.59333648	32.2397033	1.9415E-11	77.8300848	89.3867123	77.8300848	89.3867123
Type	-2.8520595	0.20912001	-13.638386	8.6936E-08	-3.3180079	-2.3861111	-3.3180079	-2.3861111
Total Students	0.0004424	0.00109287	0.4048054	0.69414398	-0.0019927	0.00287746	-0.0019927	0.00287746
Total School Budget	-7.206E-07	1.7508E-06	-0.4116012	0.6893149	-4.622E-06	3.1805E-06	-4.622E-06	3.1805E-06
Per Student Budget	0.00045274	0.00422404	0.1071822	0.9167643	-0.008959	0.00986449	-0.008959	0.00986449

% Passing Reading

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.997632
R Square	0.9952696
Adjusted R Square	0.99337744
Standard Error	0.66573518
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	932.4936	233.1234	525.996501	1.4156E-11
Residual	10	4.43203329	0.44320333		
Total	14	936.925633			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	89.1099334	10.6406269	8.37450032	7.8699E-06	65.4011392	112.818728	65.4011392	112.818728
Type	-17.287151	0.8580329	-20.147423	1.9974E-09	-19.198968	-15.375335	-19.198968	-15.375335
Total Students	0.00101761	0.00448411	0.22693579	0.82504561	-0.0089736	0.01100883	-0.0089736	0.01100883
Total School Budget	-8.718E-07	7.1838E-06	-0.121354	0.90581454	-1.688E-05	1.5135E-05	-1.688E-05	1.5135E-05
Per Student Budget	0.01121123	0.0173315	0.64687022	0.53228214	-0.0274058	0.04982822	-0.0274058	0.04982822

Overall Math + Reading Score

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99896831
R Square	0.99793769
Adjusted R Square	0.99711276
Standard Error	0.13124437
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	83.3508232	20.8377058	1209.73027	2.2345E-13
Residual	10	0.17225084	0.01722508		
Total	14	83.523074			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	86.9817334	2.09771453	41.4650002	1.5949E-12	82.3077342	91.6557326	82.3077342	91.6557326
Type	-4.7689366	0.16915433	-28.192815	7.3253E-11	-5.1458359	-4.3920373	-5.1458359	-4.3920373
Total Students	-0.0015908	0.00088401	-1.7995534	0.10212651	-0.0035605	0.00037887	-0.0035605	0.00037887
Total School Budget	2.5461E-06	1.4162E-06	1.79781682	0.10241575	-6.094E-07	5.7017E-06	-6.094E-07	5.7017E-06
Per Student Budget	-0.0053304	0.00341677	-1.5600831	0.14980093	-0.0129435	0.00228259	-0.0129435	0.00228259

% Passing Overall

SUMMARY OUTPUT

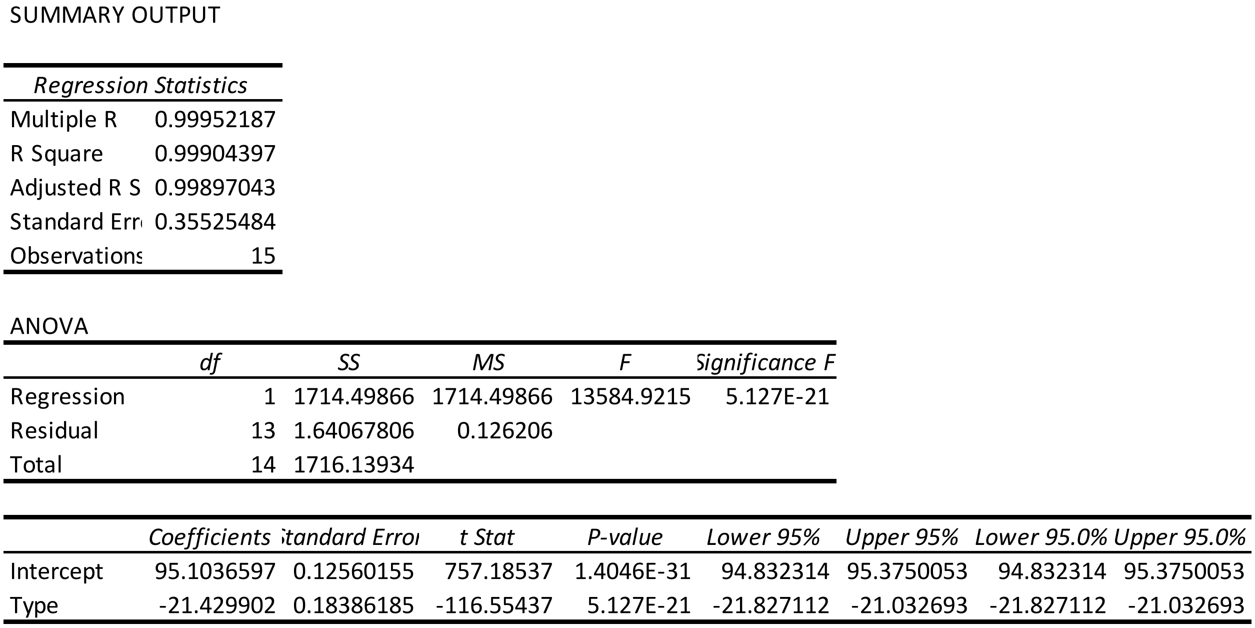
Regression Statistics	
Multiple R	0.9997491
R Square	0.99949826
Adjusted R Square	0.99929756
Standard Error	0.29343827
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	1715.27828	428.81957	4980.13482	1.9071E-16
Residual	10	0.86106016	0.08610602		
Total	14	1716.13934			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	85.6388463	4.69010383	18.2594777	5.2165E-09	75.1886438	96.0890489	75.1886438	96.0890489
Type	-22.012464	0.37819796	-58.20355	5.4418E-14	-22.855141	-21.169786	-22.855141	-21.169786
Total Students	0.0037726	0.00197648	1.90874739	0.08538125	-0.0006313	0.00817646	-0.0006313	0.00817646
Total School Budget	-5.658E-06	3.1664E-06	-1.7869118	0.10424934	-1.271E-05	1.3971E-06	-1.271E-05	1.3971E-06
Per Student Budget	0.01480988	0.00763926	1.93865303	0.08126204	-0.0022115	0.03183122	-0.0022115	0.03183122

% Passing Overall, Regression with school type only



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99952187
R Square	0.99904397
Adjusted R S	0.99897043
Standard Err	0.35525484
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1714.49866	1714.49866	13584.9215	5.127E-21
Residual	13	1.64067806	0.126206		
Total	14	1716.13934			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	95.1036597	0.12560155	757.18537	1.4046E-31	94.832314	95.3750053	94.832314	95.3750053
Type	-21.429902	0.18386185	-116.55437	5.127E-21	-21.827112	-21.032693	-21.827112	-21.032693