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.

**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.

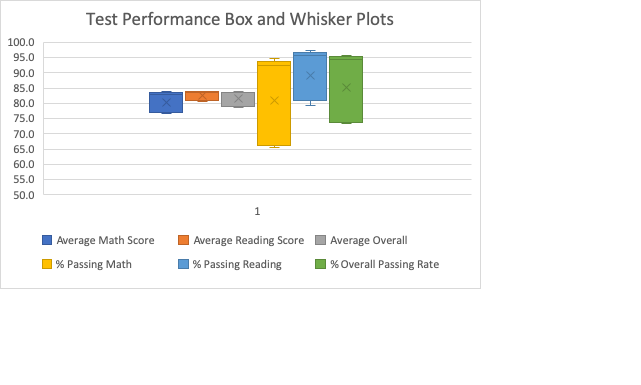
Charter schools correlate with statistically significance on all 6 indicators, but the increase is greater in math than reading, and in test pass rates over test scores.

A screenshot of a cell phone

Description automatically generated

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

Passing thresholds filter a much larger number of students in non-charter schools with the impact of multiplying the overall performance difference by school type.



1. **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)

A screenshot of a cell phone

Description automatically generatedDistrict SummaryA picture containing screenshot

Description automatically generated

School Summary: Top Performing Schools (By Passing Rate)A screenshot of a cell phone

Description automatically generated

School Summary: Bottom Performing Schools (By Passing Rate)A screenshot of a cell phone

Description automatically generated

Math Scores by Grade

A screenshot of a cell phone

Description automatically generated

Reading Scores by Grade

A screenshot of a cell phone

Description automatically generated

Scores by School Spending

A screenshot of a cell phone

Description automatically generated

Scores by School Size

A screenshot of a cell phone

Description automatically generated

Scores by School Type

A screenshot of a cell phone

Description automatically generated

**Appendix B: Regression Analysis Results**

Average Math Score



% Passing Math



Average Reading Score



% Passing Reading



Overall Math + Reading Score



% Passing Overall



% Passing Overall, Regression with school type only

