Colorado Public School Academic Performance

A Study of Financing versus Student Performance

Jarryd Allison
CSPB 4502: Data Mining
University of Colorado
Boulder, CO 80309
allisonj@colorado.edu.com

Problem Statement

Public education in the United States has come under increased scrutiny of late, due to consistently poor trends in academic performance. This has hit the Denver metropolitan area as well, with numerous articles bemoaning the state of public education[1,2].

The most recent data from US News for Denver area schools shows that while the operating budget grows considerably over the past decade[3], test scores continue to decline or at least fail to improve. Proficiency scores for reading hover in the low 30s to upper 40 percentages, but math remains stuck below 30% proficiency across all age groups[4].

The combination of low test scores but accompanied with a growing budget seems to be counterintuitive, and demands closer inspection to better understand the various relationships that exist within the available data. For example, Denver does allocate funding differently based on perceived and measured needs of students[5]. Could a focused look at financing across Denver public schools reveal additional measures to better allocate funding, or provide links between funding and increased performance? The aim of this research topic is to answer this critical question.

Additionally, there may be other underlying correlations and interesting findings that reveal themselves through the conduct of this research. Additional findings will be included in the final report.

Specifically, I wish to address the following questions:

1. Does a correlation exist between school funding and performance. If so, what is this correlation, and what are its most important variables.
2. How has funding and performance changed given the dataset from 2014 - 2023. It is noted that funding data may not be readily accessible prior to 2019, which may alter the study's range to the range of funding information available.
3. What variables appear to influence performance the most?
4. What variables appear to influence funding the most?

CCS Concepts

Social and professional topics → Professional Topics

Keywords

Education, finance, data mining, analysis

ACM Reference format:

Jarryd Allison. 2024. Colorado Public School Academic Performance: A Study of Financing versus Student Performance. In *. ACM, New York, NY, USA, 3 pages.* https://doi.org/10.1145/XXXX

Literature survey

There exists a large number of articles and research that describe the issues that Denver Public Schools have with regards to test scores and low performance[1,2,6], and financial data exists to describe the growth in the budget[3], but very little reporting and analysis exists to justify increases, at least from what appears to be publicly available at the time of the writing of this paper.

Aside from the sources listed, no analysis exists that intends to identify patterns and provide a holistic view of the budget versus student performance, and possibly help inform public policy to level set expectations, understand expenditures, and identify programs that may increase performance and educational outcomes in Denver public schools. There do exist compilation reports describing funding and performance, though these fail to describe significant patterns and a deeper analysis regarding the data[9].

Proposed Work

The proposed work involves a careful process of cleaning, preprocessing, and integration of multiple datasets to better understand the problem set. This differs from the literature survey conducted simply because the depth of analysis is lacking. The literature survey indeed studies the test performance dataset, and also arrays anecdotal evidence atop it to determine whether performance is good or bad.

But the literature review is severely lacking in objective data mining techniques to truly develop a deeper understanding, through pattern exploration and analysis. Additionally, layering in financial data also appears to be completely ignored by the majority of the literature reviewed for this project.

The following is the proposed work to complete this project, including data cleaning, data preprocessing, and data integration.

1. Data Cleaning

Much of the data currently exists in excel documents, specific to individual school years, with omissions and various discrepancies for multiple attributes across the data sets. The data cleaning step will infer missing data, using either the mean, median, to apply the central tendency, or from determining the most likely values from surrounding data points. Because the dataset is at times inconsistent, this step will also involve identifying discrepancies, and developing measures to process the data to derive meaningful insights relevant to the research topic. Notably, null values should be propagated as much as possible, instead of removed. More often than not, the median values should suffice, though, time permitting, more relevant methods, such as accounting for similar size, location, funding, etc. will be used.

2. Data Preprocessing

Data preprocessing will first involve data reduction, given the size of the dataset and the scope of the project. Multiple attributes simply will not apply, and will be removed. Duplicates are not an anticipated issue, as the data is structured by schools in the Denver area, meaning that duplicates are highly unlikely, but will be screened to ensure duplicates do not exist. Sampling and clustering will be attempted, both to better refine the processes developed in this research, as well as to determine if significant patterns emerge. Finally, data normalization will be used to identify relevant trends amongst various factors, such as test scores across subjects, aggregate pass/fail percentages, and so on. For example, the dataset includes such information as English and Spanish performance. Time permitting, these will also be included to determine performance in those respects, but mathematics will be the primary focus for this report. Additionally, state and district aggregations are also available. However, these will not be used as they tend to mask individual schools that are vastly different, and may provide additional insights important to the outcome of this research.

3. Data Integration

The dataset identified for school performance is clean, well structured, and relatively complete. However, integrating this data with financial data will be a comprehensive step in the work required to align finance information alongside the academic performance dataset. To do this, a web scraper will likely need to be developed in order to collect the finite information required for this research.

Datasets

The datasets identified for the scope of this research are as follows:

1. Colorado Measures of Academic Success (CMAS) - Mathematics, English Language Arts, Science and Social Studies Data and Results (2014 - 2023)

The CMAS performance dataset[7] is compiled and provided by the Colorado Department of Education annually. It is provided in multiple formats, to include .csv format, and is available for download. Limitations are myriad in this dataset. While it does present the most comprehensive dataset available for this information, it only gives mean scale score and the standard deviation, while omitting high/low scores and more accurate representations of the scores across each school as a whole. While valuable metrics can still be gleaned from this, they would be better served with more scoring information. This report will attempt to quantify these restrictions by using the scores and standard deviations to determine if any interesting trends emerge.

2. Colorado Department of Education Financial Transparency Office

The Colorado Department of Education also offers Financial Transparency, both in the form of an online tool[8] as well as comprehensive quarterly financial statements[3]. These datasets are not as well structured as the performance data, and will need to be heavily processed in order to correctly align them with the performance dataset described above. This processing will be done with the creation of a web scraper. Initially, 2023 data will be utilized, but, given additional time, historic data through 2019 is present, giving a much richer data set that will help reveal insights to the correlations between performance and funding. The largest issue with this dataset is the omission of reporting prior to 2019. This coil be gleaned by parsing PDFs also available, but that would be well outside the scope of this course.

Evaluation Methods

While the body of literature is relatively small, the Common Sense Institute report[9] compiles a list of generalized information regarding performance and finances across Colorado schools. Additionally, historic[10] and more current research[11] seems to be mixed, with previous studies finding no correlation between funding and performance, but more focused studies finding a strong correlation. To date, there does not seem to be any Denver-focused study on this topic. However, the existing research can help measure and provide additional techniques to better understand the relationships that may exist.

Finally, the data set itself provides data that can be used to confirm initial results and ensure that the underlying functions are correctly being utilized across each row. For example, 2023 reporting includes changes from 2022 and 2019 (2021 and 2020 data was mysteriously absent), which can be used to confirm findings year over year in some small regards.

Tools

To conduct this research, the following tools will be utilized:

1. Jupyter notebooks will be heavily utilized to run various data mining techniques and produce images to support the findings.
2. Python will be used extensively for data mining, as it is adept at formatting CSVs to run advanced data mining techniques against the underlying data.
3. Numpy/pandas will be used in python, as these libraries contain multiple powerful features to simplify the data processing steps. Specifically, pandas will be used to create dataframes, which are much easier to utilize when conducting analysis on datasets.
4. BeautifulSoup, a python library used to help parse HTML data, will be used to scrape public information regarding Colorado school financial data.
5. Matplotlib is a popular python plotting library that will be used extensively to create data visualizations in support of the research goals of this project.
6. Looker studio may also be utilized as well to produce high quality data visualizations, time permitting.

Milestones

July 23: Continue financial data cleaning, preprocessing, and integration.

July 28: Complete all data cleaning, preprocess, and integration.

August 3: First draft of final report complete.

August 9: Final report complete.

1. Milestones Completed

The current progress has been slightly delayed due to the data inaccuracies and various different formats encountered during the data processing step. Colorado shifted its reporting formats several times, from 2014 to 2015, from 2015 to 2018, and from 2019 and on. This has resulted in significant work formatting the csv files into a comprehensive format that matches year over year and allows for standardization of results in the long run.

The current work is progressing well in spite of these setbacks. A web scraping tool that collects public information on Colorado school financial data[8] has been completed for the year 2023. More work will be accomplished to secure the financial data up to 2019, after which a PDF parser would be required, which is outside of the scope of this project given the time constraints. However, even the past four years of data will be instrumental in understanding funding vs. performance metrics across the Colorado school data set. The only significant problem with the web scraper is it takes quite a while to run, given that for each school it is making an API call to retrieve the financial data html, and then parsing the HTML. A test file for 2023 only has been created, and time permitting, a full file for all available financial data (2019-2023) from the website will be gathered.

Additionally, work is nearing completion in formatting the data for more extensive research and comparisons. Multiple columns exist in the data set that will not be used in the conduct of this research. Removing those is relatively simple, but each grade must be reformatted into a separate column to attempt to identify any significant trends outside of the "All Grades" category, which is relatively simple as this item can be pulled directly from each school. Some schools support grades 3-5, others 6-8, and still others 3-8, which requires some manipulation of the data to ensure it matches across each school year.

1. Milestones To Do

Complete Data Preprocessing

As stated above, data preprocessing has continued to prove challenging, given the data discrepancies and formatting issues. However, this is going much faster as familiarization with the data increases.

Draft of Final Report

Per the updated timeline, the draft for the final report is still on schedule, to ensure all processing is complete and results are finalized into a cohesive report.

Final Report

All signs point to being able to complete the finalized report before or on August 9th.

Results So Far

Initial attempts to identify and replace null values have been very successful. For example, if preparing the 2023 data for processing and integration, multiple null values exist, which is surprising given that performance data is mandated to be reported at each school level. Regardless, an attempt was made to replace the null values with the mean value of the test scores for the reporting period. Unfortunately, the data, as indicated in Figure 1, showed that the mean (for 2023 mathematics at All Grades, the mean score was 692.8) lay far outside the expected values, as a noticeable line at the bottom of the scatterplot clearly indicating null values. While justifiable, the median (729) fared much better, visually matching the scatterplot expectations. Time permitting, a more comprehensive replacement of null values incorporating financial, population, and grade data will be utilized instead.

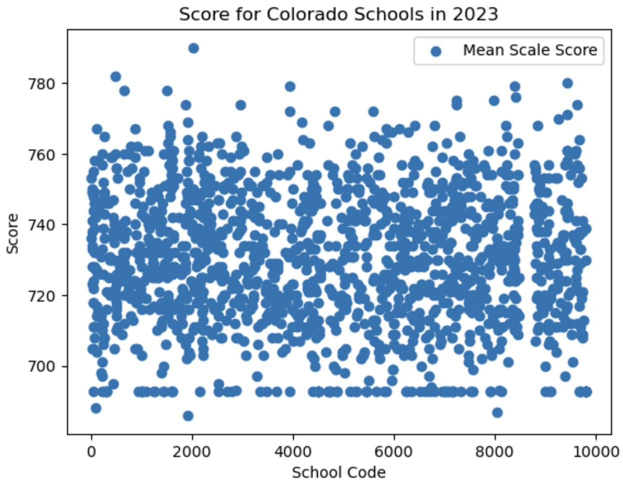

Figure 1: Mean Scale Score for Mathematics tests in Colorado Schools, 2023 with mean score used for null values

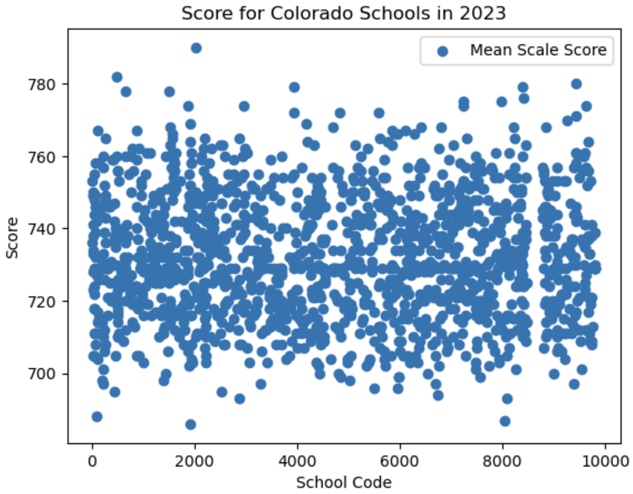

Figure 2: Mean Scale Score for Mathematics tests in Colorado Schools, 2023 with median score used for null values

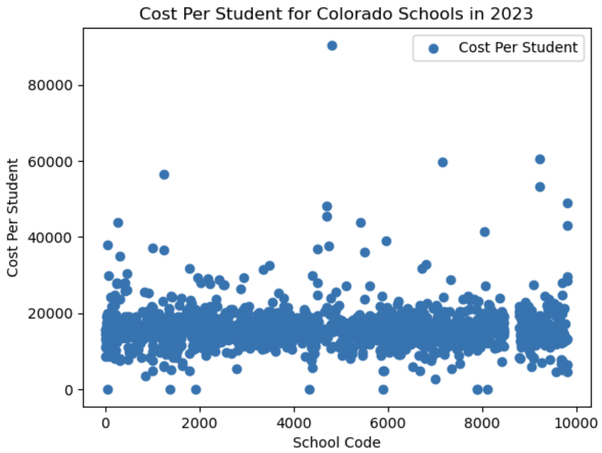

Figure 3: Cost Per Student in Colorado Schools in 2023

Figure 3 shows the initial data for cost per student (mean value of $16,045.63) across all public Colorado schools, which notes interesting outliers that will be explored in subsequent research. Again, the median values replacing null values does enrich the dataset, though not nearly as much as the test scores dataset. This is shown in Figure 4.

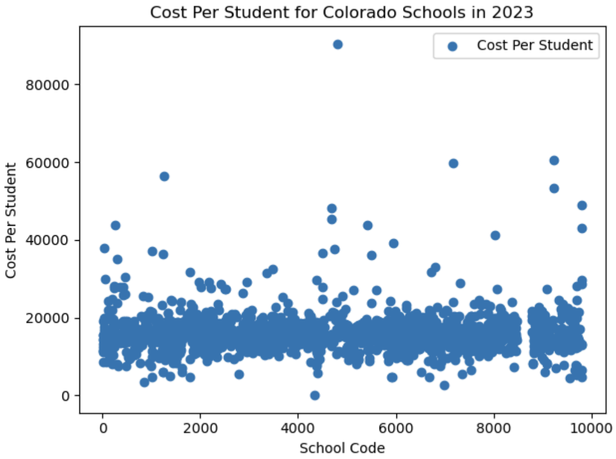

Figure 4: Cost Per Student in Colorado Schools in 2023 with null values replaced with median cost per student

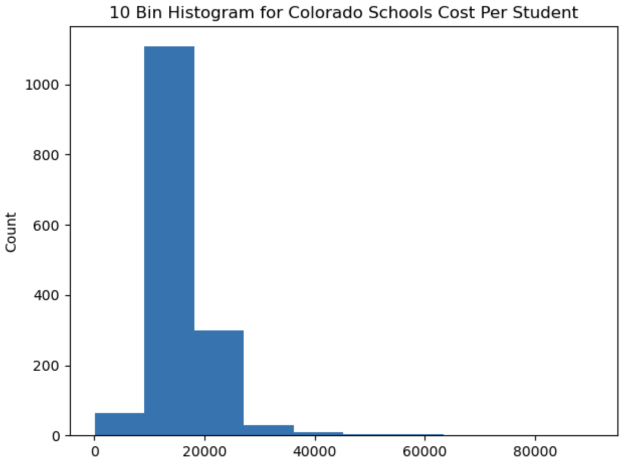

Figure 5: Cost Per Student in Colorado Schools in 2023 with null values replaced with median cost per student

Figure 5 notes that the vast majority of student costs are between the $15,000 - $20,000 mark, with outliers noted on both the low and high end.

Interestingly, the Pearson coefficient, calculated for 2023, indicates a negative correlation (-0.184). Because this is so weak, more data is needed and will follow to surface other interesting trends, but at the very least we can infer (from 2023 only) that there is either no correlation between spending and performance, or that indeed in some cases spending may decrease performance. However, this result is incredibly weak and very preliminary. More information is needed before issuing a final report.

REFERENCES

[1] Melanie Asmar, 2022. More Colorado schools and districts earn low state ratings. https://www.chalkbeat.org/colorado/2022/9/8/23343341/colorado-school-performance-framework-ratings-2022/.

[2] Alan Gottlieb, 2022. Performance plummets in Denver's schools. https://gazette.com/opinion/denver-columns/column-performance-plummets-in-denver-s-schools/article_54f8631a-b4dd-11ec-ac74-c36f0357135b.html

[3] Colorado Financial Services Reports, accessed 2024 for years 2014-2024. https://financialservices.dpsk12.org/o/financialservices/page/financial-transparency

[4] US News. 2024. Overview of Denver Public Schools. https://www.usnews.com/education/k12/colorado/districts/school-district-no-1-in-the-county-of-denver-and-state-of-c-112125

[5] Melanie Asmar, 2022. The $3,500-per-student difference between two Denver schools. https://www.chalkbeat.org/colorado/2022/4/28/23045997/denver-student-based-budgeting-smith-carson-elementary/.

[6] Ari Armstrong, 2023. Denver Public Schools celebrates lowering achievement bar. https://pagetwo.completecolorado.com/2023/09/26/armstrong-denver-public-schools-celebrates-lowering-the-bar/

[7] CMAS - Mathematics, English Language Arts, Science and Social Studies Data and Results, 2024. https://www.cde.state.co.us/assessment/cmas-dataandresults

[8] Financial Transparency for Colorado Schools, 2024. https://www.cde.state.co.us/schoolview/financialtransparency/homepage

[9] Common Sense Institute, 2022. Dollars and Data: A look at PK-12 Funding and Performance in Colorado. https://commonsenseinstituteco.org/dollars-and-data-a-look-at-pk-12-funding-and-performance-in-colorado/

[10] Lawrence O. Picus, 1995. Does Money Matter in Education? A Policymaker’s Guide. https://nces.ed.gov/pubs97/web/97536-2.asp

[11] C. Kirabo Jackson, Rucker C. Johnson, Claudia Persico, 2015. The effects of school spending on Educational and Economic Outcomes: Evidence from School Finance Reforms. https://www.nber.org/system/files/working_papers/w20847/w20847.pdf