DATA SCIENCE

The Significance of Educational Performance on School Funding

Srikant Vasudevan*

Correspondence:
vasudevan21@ncssm.edu
North Carolina School of Science and Mathematics, 1219 Broad St., 27705 Durham, NC
Full list of author information is available at the end of the article
*NCSSM Online Program

Abstract

Federal and state resource allocation, primarily in the topic of educational funding, has long been a highly enigmatic topic when it comes to the causal factors and subsequent effects of the variability of revenue for school systems. Legislation in the past has attempted to streamline the process of allocating funds, by prioritizing areas with low income statuses and other limiting factors. Even so, the variability of revenues for state school systems exists and still causes many to ponder the true causes. We collected data from various different indexes and sources that allow us to acquire the academic performance indicators, enrollment data and other variable we needed. The indexes were parsed, cleaned and organized into a tidy dataset, then used to carry the process of thorough statistical analysis. Several linear models and unsupervised machine learning techniques such as the gini index were used to determine the statistical significance of educational performance on educational funding. Accompanying these models were well developed visuals that encompass and properly present the data for a streamlined analysis.

Keywords: education; funding; academic performance

1 Introduction

Public education, the root of the United States' youth and arguably the key to the future success of the country, is an intricate institution that is dependent on several vital factors to maintain manageable environments and ensure the success of its students. Most public education institutions obtain several different sources of revenue, most notably from the federal, state and local governments they are within. These sources of funding fluctuate year to year and can have numerous effects on not only the annual performance of students, but also on their future successes. There have been many studies that look deeper into the classroom and thoroughly evaluate the direct cause-effect relationship between revenue and educational performance. These studies have generally revolved around the consensus that as the revenue of a school, or school system tends to increase, the educational performance of the students in these schools tends to increase as well (seen in test scores such as the end-of-grade exams and math/reading benchmarks) [2, 3]. It is often thought that large decreases in funding for schools are due to poor academic performance from the students in those schools, therefore putting these schools in a downward spiral as low funding leads to poorer academic performance. A landmark act, known as the "No child left behind act", was issued in 2001 and attempted to lift up disadvantaged schools and school systems through increased funding and other means [5]. Throughout its institution, this act was quite successful in establishing a more

balanced success rate in the United States, but major inequality in success rates and quality of schools existed (the law was in effect until 2015) [6]. This raises a few simple but mind-boggling questions. How drastically does school funding actually vary? How can we close these gaps? To what extent will closing these financial divides actually help public education in the United States?

Throughout history, there have been several advocates for the reform of educational funding; in the state of Washington, a series of landmark cases from 1977 to 2007 fought this issue and procured a potential solution. The 2007 decision, McCleary v. Washington, stemmed from a lawsuit by the McCleary family, but was also backed by many students and families in the state of Washington. The lawsuit claimed that Washington schools were underfunded and that there were particular forms of financial neglect in severely underprivileged areas. In the final decision, the state Supreme Court ruled, among other things, that there needed to be a drastically larger operating budget for public schools [8]. Eventually, in 2015 the state legislature settled for 1.3 billion dollars [7, 8]. Though partly a niche case, this suit can be representative of the nationwide unrest for financial divides in school systems, and in a large-scale sense, it can also show the drastic differences in funding for schools within a state.

Though there is increasing attention towards federal and state resource allocation for education, there are a lot of unknown factors involved in determining how much money should be given to state public school institutions. As seen in a 2014 study on the fairness of public schooling, enrollment of impoverished students in public educational institutions can reach as high as 53%-57% of the ¡10% socioeconomic strata (in states such as New Hampshire and New Jersey) [9]. These poverty enrollment rates signify the overall quality of schools in the impoverished areas, but any vision of surefire reform of these areas is hazy at best. In fact, there has been little to no proactive legislation in the past 5 years that directly addresses the economic stabilization of educational institutions [10]. A large part of attempting to institute equality in public schools is understanding the factors leading to fluctuations in revenue as well as understanding the effects they have.

In addition to the unascertained facts surrounding resource allocation to both local school systems and larger school systems, the variety of socioeconomic statuses in a concentrate or cluster leads to the inability to pinpoint areas that require the most (and the least) financial aid [11]. This is a minor issue when considering the grand scheme of economic distribution. It more or less is a cascading stream of revenue from the federal level, state level and then finally the local level and donations. When distributing money to the states, differences in economic status within the state does matter to an extent, but viewing the state more wholly will allow for easier classification and distribution of money for the proper use toward educational institutions in each state. Much of the action taken towards maintaining economic stability in schools is by state legislators and officers and an enumeration of data can only take values into account, such as fluctuations in money and enrollment, rather than annual political influence [12].

In this paper, we will use data from the “Unification Project” dataset by Kaggle and several different indexes with important values such as number of schools per state and cost of living data. Our aim was to develop models and visuals that

would not only clarify the gray areas surrounding educational funding, but would also reach a consensus on causes and effects of federal, state and local funding on student performance, expenditures and future stability of state school systems. The primary dataset, the "Unification Project" includes annual revenues and expenditures, enrollment numbers, average test scores and several other summative values from 1992-2017 that can help us develop accurate and conclusive models. We will mainly view the potential for a cause-effect relationship through the use of advanced mathematical models and graphics. In the case of this paper, there will be a large use of scatterplots and linear models to streamline the analysis and interpretation of the data.

2 Dataset(s)

In this study, our aim was to use previously aggregated data on financial revenues and expenditures, enrollment data, test score data, and other indexes for each of the 50 states in the United States to determine the most influential factors that cause fluctuations in revenue and the subsequent effects of these fluctuations. Certain specific values such as school district locations, revenues of school districts, and average scores for those districts were not included in the large dataset or any of the indexes we pulled our data from. The data, as aforementioned, was largely preformatted for ease of use. It primarily consists of summative values of each year from 1992 to 2017 retrieved from several different contributors and was congregated into a dataset (later available from Kaggle). Enrollment data was acquired from the National Center for Education Statistics (NCES), the financial data, both revenues and expenditures, was acquired from the United States Census Bureau and the academic achievement data was acquired from an academic statistics and reports institution called The Nation's Report Card. The academic performance indicator used in this data is the National Assessment of Educational Progress (NAEP) for both math and reading in grades 4 and 8.

Preparation of the dataset involved two key processes of data science: exploratory data analysis (EDA) and data munging. The EDA, data munging, and thorough analysis of the dataset were done using the R programming language using a free online integrated development environment (IDE), RStudio. On initial review of the cumulative dataset, there were over 1200 data points containing information for several different variables. Before cleaning the data or managing the information, a separate, aesthetic dataset from the SocViz (social visualization) library, which contains the latitude and longitude coordinates to create a scale map of the continental United States, was annexed to the primary dataset. SocViz is a larger part of the book "Data Visualization" by Kieran Healy, which outlines the premise of using R and goes in depth to streamline the processes of data science [13]. To merge the datasets, we first replaced all spaces with underscores for the names of each state in the SocViz dataset. Secondly, we changed the variable name from "STATE" to "region" in the primary dataset to match the names of the variable containing state data in each dataset. Finally, both datasets were merged into one dataset and that dataset was ready for exploratory data analysis.

In order to thoroughly understand the data, we looked at several descriptive statistics of the data: summary, dim, glimpse and head. The summary statistics contained

all 29 variables with the minimum value, first-quarter value, median, third-quarter value, and maximum value of each quantitative variable and frequencies for different "factors" or strings for each categorical variable. The "dim" statistics show the dimensions of the final dataset, 1331 data points with 29 variables. The "glimpse" statistics contain several values similar to the summary statistics but also contain the type of each variable (factor, integer, numeric, etc.) in the dataset. Finally, the "head" command displayed the first 5 (this number can be changed) values of the dataset; the command showed the first 5 states: Alabama, Alaska, Arizona, Arkansas, and California in the year 1992 and all the financial, enrollment and academic achievement data associated with them.

To begin the cleaning of the dataset, we attempted to eliminate all NA (either missing or unreported) data. In doing this, we ran into a fatal error: every observation in the dataset contained at least one missing variable, therefore eliminating all NAs actually eliminated the entire dataset. To work around this error the NA variables would be sorted out or neglected when honing in on specific variables. After dealing with missing data, variable names were altered to aesthetically fit the plots. Categorical variables such as "STATE" would be set to full caps and any "_" characters would be changed to space characters, for example, "north_carolina" would be changed to "NORTH CAROLINA". This change is primarily so that on models and graphics, categorical variables are easier to understand.

As the process of data munging continued, distribution plots were developed and analyzed to understand the data. It is important to note that during this process, each variable was looked at individually to prevent premature analysis (leading to false conclusions). This was primarily so that we could determine what the distribution is for each variable and if it can provide accurate results when modeled and plotted. All numeric variables in the dataset follow an approximately normal or bimodal (depending on the variable) distribution when displayed in a 6-bin frequency histogram. The significance of these distributions is that they provide more accurate findings than skewed distributions and irregular distributions. Minor adjustments, such as replacement of NA data in revenue or expenditure variables to fit the distribution, were made to further prevent misrepresentation of the data. Once the data munging process concluded, the dataset was reviewed several times to prevent errors during analysis and to make sure all the required data was present.

3 Methods

The beginning of analysis was an overall look at the dataset and a quick analysis of the likely correlations between different variables. After the aforementioned data setup, we decided to first look at expenditures and revenues. More specifically, the percentage of annual revenue spent by each state each year. Realistically, we were looking for numbers less than 100% (1 if using numerical probability), as this would mean each state was keeping within their revenue stream and not going into debt that year. The year and state were taken into account when analyzing the percentages and based on if annual revenues were exceeded or not, we looked at subsequent revenues to see if they were altered consistently.

The primary library used was ggplot2 from the tidyverse package. This library is often referred to as the "grammar of graphics" due to its large variety of graphics

and parameters to construct aesthetically pleasing and informative graphics. For 5 randomly selected school years (i.e 2009-2010) the longitude and latitude data from the SocViz dataset were used to create a scale map of the US with each state shaded differently based on different factors such as revenues, expenditures, academic success, enrollment, and number of public schools. These graphics served as general, non-numeric comparisons for visual analysis. Though much could be extrapolated from the data, it was inadvisable to assume correlation or even a population-representative relationship without sound numerical models that backed up the graphic. Along with the United States graphics, ggplot2 was used to construct scatterplots that visually present potential linear relationships between two separate variables. In the case of the scatterplots, all years from 1992 to 2017 were used; these were distinguished by the size of the points in each scatterplot. Another library that was used extensively throughout the graphical representation of the data was gridExtra. This library, simply put, is a versatile tool that helps arrange and organize graphics in R. GridExtra was used to arrange scatter plots and polygon graphics (the US maps) side by side for easy visual comparison.

Other technical aspects to the R programming language were utilized in the analysis as well, these include loops (for, while, etc.) and if else statements. These methods were used primarily to expedite the sorting of the dataset by computationally running through the observations and assigning attributes to them. The tidyverse package, a modern aggregation of several useful libraries including ggplot2 and dplyr was used extensively throughout the data analysis process. Dplyr, a library in the tidyverse package, was a crucial element in the manipulation and simplification of the data. Many of the translations, mutations and creations of data frames with specific values were done with this library.

4 Results

4.1 Revenues and Expenditures

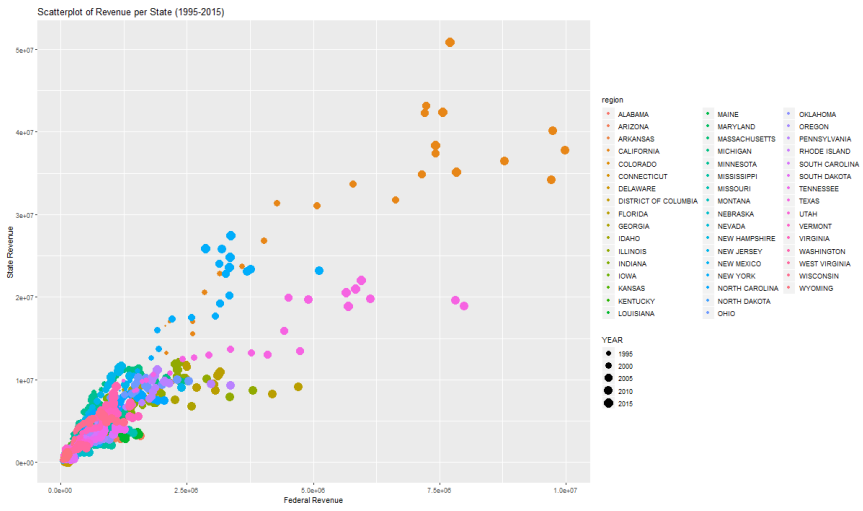

Figure 1 Scatterplot of State and Federal Revenues (1992-2017)

```
Call:
lm(formula = dataset$STATE_REVENUE ~ dataset$FEDERAL_REVENUE)

Residuals:
      Min        1Q    Median        3Q       Max
-14661856  -1229706   -423064   1130114  18451385

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.446e+06  5.943e+03   243.3   <2e-16 ***
dataset$FEDERAL_REVENUE 4.022e+00  3.046e-03  1320.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2905000 on 389921 degrees of freedom
  (15537 observations deleted due to missingness)
Multiple R-squared:  0.8173,    Adjusted R-squared:  0.8173
F-statistic: 1.744e+06 on 1 and 389921 DF,  p-value: < 2.2e-16
```

Figure 2 Linear Model Output for State and Federal Revenues

This first set of results consists of primarily linear models and scatterplot graphics. This set of results primarily covers the holistic data in the dataset and doesn't delve into the deeper aspects of the dataset. A key component of the analysis is the relationship between the different types of revenue: federal, state, and local. Local revenue data is far less important in this case due to the nature of the dataset (the data contains enrollment, academic benchmark and expenditure data only for the state levels, thus requiring the state level to be the narrowest level that is analyzed). The first model created was a linear model between all federal and state revenue data. The y-value would be state revenue and the x-value would be federal revenue. Each point on the scatterplot would be a value containing both the state revenue and federal revenue for a unique state-year combination (i.e one point on the scatterplot represents Alabama in 1992 and no other point on the plot would contain the same year and state combination). Figure 1 is the fully compiled scatterplot with state and federal revenues. Upon viewing the scatterplot, it is apparent that there is a strong, positive, linear correlation between federal revenue and state revenue. This is both expected and is corroborated by the findings of the National Center for Education Statistics [14].

Though the scatterplot provided us with strong visual evidence of the correlation between state and federal revenue, we required numerical evidence of a strong linear relationship. Figure 2 is the output of a linear model of state revenues vs federal revenues. The key numbers we looked for were the R-squared and adjusted R-squared values (in this case they are equal to each other). In statistics, the r value represents the correlation of data where 1 is a perfect, positive, linear relationship and -1 is a perfect, negative, linear relationship. The R-squared value, on the other hand, represents the amount of variability in the y-value (state revenue) that can be accounted for by the x-value (federal revenue). Looking at the R-squared value in the linear model output, we determined that 81.73% of the variability in state

revenue can be accounted for by the distribution of federal revenue. Our consensus of the strong positive linear relationship of federal and state revenues is evidenced by this value. Due to this relationship, it is safe to assume that as the federal revenue provided to a state tends to increase, the state revenue for that state also tends to increase. This assumption will be crucial to the rest of the analysis.

Switching angles, another key relationship to look at is that of the total revenue to the total expenditure. The key value extrapolated is the percentage of times (each state from 1992 to 2017) that total expenditure for the year exceeds the total revenue for the year. In order to do this, we computationally ran through each type of revenue (local, state and federal) for each state and collected the sum under the variable total revenue; we did the same for expenditures per state per year. After collecting the values, a separate dataset was created with the total revenue and total expenditure as well as percentage of revenue spent, which was calculated by using the following formula:

$$(TOTAL_{EXPENDITURE})/(TOTAL_{REVENUE})$$

Upon initial review of this data, the median value of the percentage of revenue spent per state per year was greater than 100%, this means that over 50% of the time, a state has overspent its revenue from 1992-2017. After digging deeper, by finding the amount of over-spending and under-spending cases, we found that approximately 58.2% of annual revenues per state per year from 1992-2017 were exceeded by the states' expenditures. This signifies that about 58.2% of the time, the revenue for a state school system is exceeded by its expenditures.

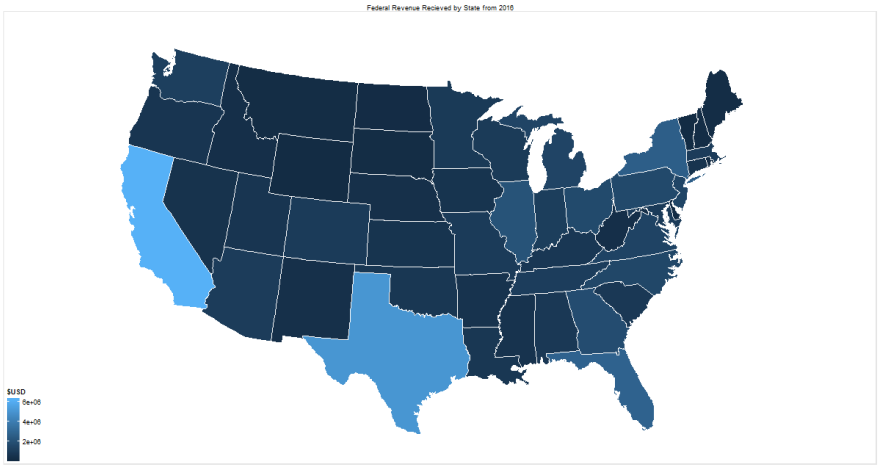

Figure 3 Federal Revenue Per State (2016)

Along with viewing scatterplots and raw numbers, it was helpful to view a map of the continental US with different shades of a color to represent different values. As seen in figure 3, a map of the continental United States, which has a scale of shades of blue representing varying amounts of federal revenue (in USD) received by each

state, the states of California and Texas seem to stand out the most as receiving substantially more federal revenue than the rest of the states.

4.2 Pinpointing Factors

Since the relationship between state and federal revenues has been made clear, we dove further into the data to extrapolate relationships and correlations that tell us more about the factors and effects of educational funding. In evaluating the effect of revenue increases (and decreases), many studies have concluded that as revenues of a school, district and state increase, the relative academic success of the respective educational institutions also tends to increase [3]. This consensus is disproved by examining the relationship between state revenue and NAEP scores (reading and math) from the same year (Figure 4 and Figure 5). It is clear that though there is slight extrapolation of data in the points from the District of Columbia, as DC doesn't receive state revenue. Disregarding those values, it is clear that there is no strong, or even moderately strong, linear relationship between state revenue and NAEP scores. It is important to understand the meaning of these graphs, since the year of the NAEP exams and the state revenue were matched, the state revenue variable is dependent on the NAEP exam score in these graphics. As aforementioned, as state revenue tends to increase, federal revenue tends to increase as well, so the approximate linear relationship between NAEP scores and state revenue (very weak) can be assumed with federal revenue as well. (to be continued)

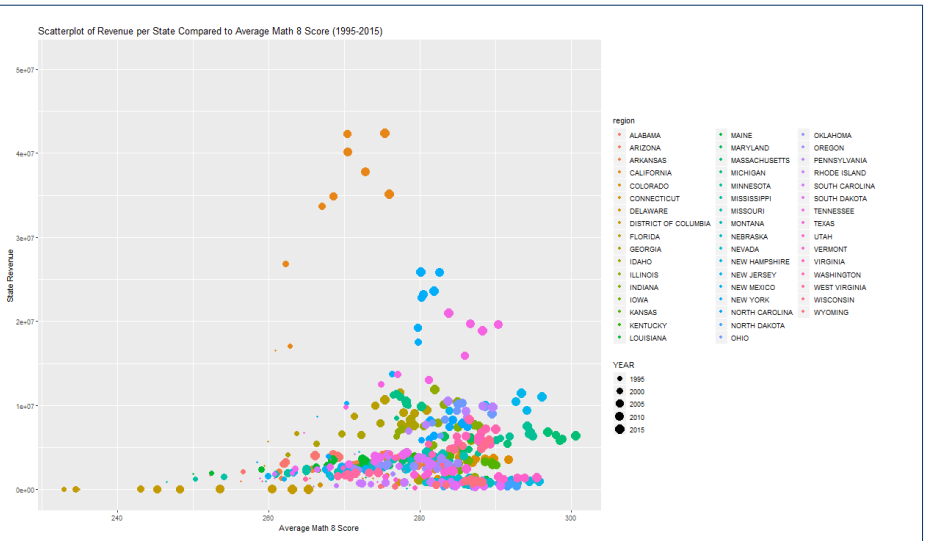

Figure 4 State Revenue vs Average Math 8 Score Per State

There are innately a different number of public schools in one state than the other states. It is important to address this nuance when analyzing relationships which cause altering levels of revenue. There seemed to be either an exponential relationship or a moderately strong linear relationship with few strong outliers. California, as expected, is at the top right of the graph, containing the most public schools and the most federal revenue. Other states fall in line approximately according to the amount of public schools in that state. As the amount of public schools in a

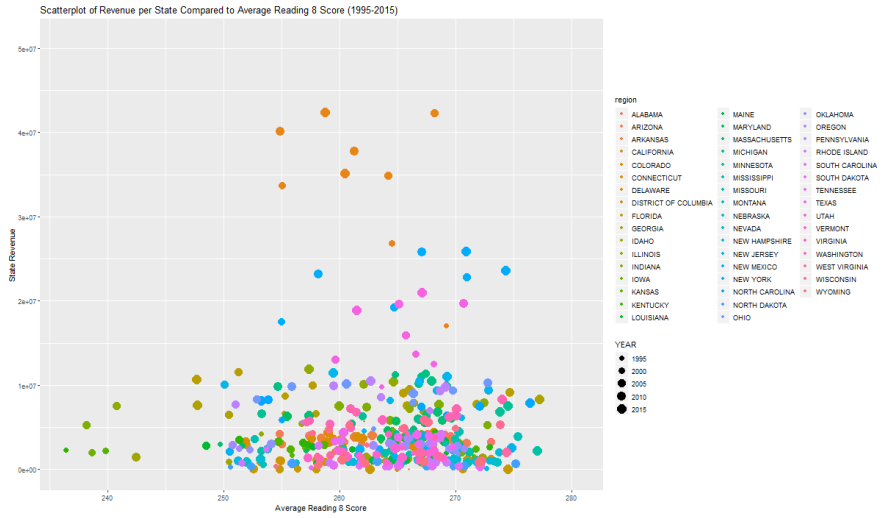

Figure 5 State Revenue vs Average Reading 8 Score Per State

state tends to increase, the amount of revenue given to the state also tends to increase, but this relationship is contained outside the realm of influential factors as it expected and an innate relationship.

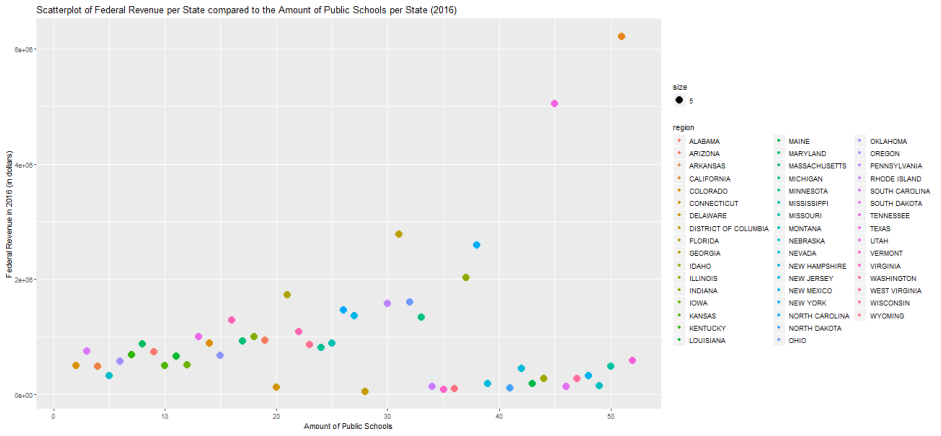

Figure 6 Federal Revenue vs the Number of Public Schools in Each State (2016)

4.3 GINI Index

The gini index is a large part of statistical measurement and inequality determination. The value ranges from a scale of 0-1, where a generally lower score signifies higher statistical significance and a variable that is more determinate to the variability in revenue per state (state and federal). To evaluate the gini index of each variable, a few variables had to be sorted out, so we created a new dataset. All the non-numeric variables (mostly factors) were not included in the dataset as a gini index for a factor cannot be calculated. The new dataset contained 20 total variables, all of which are numeric (integer, float or double). We created a for loop which ran through an arbitrary variable "i" which started at 1 and added 1 to itself until it reached 20. In the for loop, each variable in the new dataset had its gini index calculated, if the index was lower than the preceding index, then the "value" variable would become the new lower gini index and the "number" variable would record the value of "i" when the lowest gini index was recorded. Figure 7 contains the "value" variable, "number" variable, and the names of each variable in the numerics dataset to match the "number" with the name of the variable. The lowest gini index was produced by the 1st variable in the numerics dataset: "TOTAL_EXPENDITURE". The index produced, .9988062, though, is approximately 1, which signifies little to no statistical significance. This means that all numeric variables in the dataset have a gini index greater than or equal to .9988062, and are therefore, not statistically significant when determining subsequent federal or state revenues for a state school system

5 Conclusion

As aforementioned, research concludes that higher levels of federal revenue, state revenue and even local revenue tend to lead to more successful schools in the realm of academic achievement. The purpose of this study was to determine if there was a significant correlation between the level of academic achievement and federal revenue, where academic achievement acts as the influence, or causal factor. The data was acquired from several different sources, but all these sources proved to be reliable due to their position as government or scholarly papers. Several different linear models and indexes were calculated and congregated to measure the significance of several different variables and accompanying these models, many different visuals were presented to provide strengthening evidence to back our claims.

Looking at the linear models created, there is a strong positive linear relationship between state and federal revenue, the two types of revenue we examined closely, but on the contrary, there isn't such a determinate relationships between either type of revenue and an NAEP exam. For this paper, we chose to compare state revenues because the variance is slightly more predictable than federal revenues, and the NAEP exam average per state(math and reading) did not show significant effects on revenue changes. Furthermore, the linear relationship between the two variables (Average NAEP exam scores and state revenues per state) was a weak and indeterminate relationship with a slow positive curve.

The gini index only strengthened the developing idea that academic performance may have little to no effect on the changes in federal and state funding for states. The index determined that the most statistically determinate variable, though having little to no statistical significance, is not related to academic performance, and

rather, expenditures. The weakness of the academic performance variables have led to a rather anticlimactic consensus.

It seems that there is little to no statistical significance of academic performance indicators when it comes to the variability of educational funding. Though there have been studies that prove the converse relationship, this specific relationship seems to be out of the bounds of education-specific variables (academic performance) and more in the realm of inflation, differences in costs of living, and pertinent fiscal occurrences.

Availability of data and materials
The data for this work was obtained from
https://www.kaggle.com/noriuk/us-education-datasets-unification-project.

Competing interests
The author declares that they have no competing interests.

Authors' contributions
This paper is solely the work of the author. All references are included in the bibliography and are cited appropriately.

Funding
Funding for this program is provided by the North Carolina School of Science and Mathematics, the University of North Carolina General Administration, and the General Assembly for the State of North Carolina.

References

1. Lessinger, Allen: Performance proposals for education funding: a new approach to federal resource allocation. The Phi Delta Kappan **51**(3), 2 (1969)
2. Honu: United States' Higher Education Costs (2019)
3. NCES: Public Elementary Schools and Jurisdiction (2004)
4. Hanushek: School Resources and Student Performance (2000)
5. K12: No Child Left Behind Act (2001)
6. Whitney, Candelaria: Effects of the No Child Left Behind Act (2017)
7. Cornwell: Key Events of the McCleary School-Funding Decision (2016)
8. Barlett: Short Story: McCleary Case (2018)
9. et al., B.: Is School Funding Fair? (2014)
10. NCSL: Education Legislation — Bill Tracking (2008)
11. Moses, Bireda: Reducing Student Poverty in the Classroom (2010)
12. Lynch: Allocating Resources to Improve Student Learning (2011)
13. Healy: Social Visualization (2018)
14. NCES: Public School Revenue Sources (2019)

[1] [?] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14]

```
Call:
value
number
names(numerics)

[1] 0.9988062


[1] 1


[1] "TOTAL_EXPENDITURE"            "INSTRUCTION_EXPENDITURE"      "SUPPORT_SERVICES_EXPENDITURE" "OTHER_EXPENDITURE"
[5] "CAPITAL_OUTLAY_EXPENDITURE"   "GRADES_PK_G"                  "GRADES_KG_G"                  "GRADES_4_G"
[9] "GRADES_8_G"                   "GRADES_12_G"                  "GRADES_1_8_G"                 "GRADES_9_12_G"
[13] "GRADES_ALL_G"                 "AVG_MATH_4_SCORE"              "AVG_MATH_8_SCORE"             "AVG_READING_4_SCORE
[17] "AVG_READING_8_SCORE"          "X2016_F_REVENUE"               "X2015_AVG_M8"                 "S_2016_REV"
```

Figure 7 Gini index for numerics in the dataset