



US Education Funding



Joel Robles, Eric Dally, Kenneth
Priester, Alejandra Alvarez



Description of the project:

Analyzing the relationship between the revenue and expenditures in education by state, and their average NAEP Math/ Reading exam scores.

We would like to compare the funding between higher and lower performing states to find how it might have affected academic performance.

Reason for selecting the topic:

- We are all passionate about education and wanted to see how we could use data to help improve academic performance in the future
- Understanding that proficiency on tests predicts future academic success. In order to improve U.S. student performance, we set out to analyze funding sources and their possible impact on test scores.

How are schools funded?

According to [Education Week](#), public school funding comes from a variety of sources at the local, state and federal level. Approximately 48% of a school's budget comes from state resources, including income taxes, sales taxes, and fees. Another 44% is contributed locally, primarily through the property taxes of homeowners in the area. The last 8% of the public education budget comes from federal sources, with an emphasis on grants for specific programs and services for students that need them.

Description of the source data:

Utilizing Kaggle, we will work with the following [data set](#).

To ensure that our exploration of data is specific and measurable, we will be looking exclusively at the Nation Assessment of Education Progress (NAEP) Math and Reading scores for 4th and 8th grades.

As NAEP is congressionally mandated project that is conducted nationwide, it will likely provide the best data comparison across the states.

Original dataset:

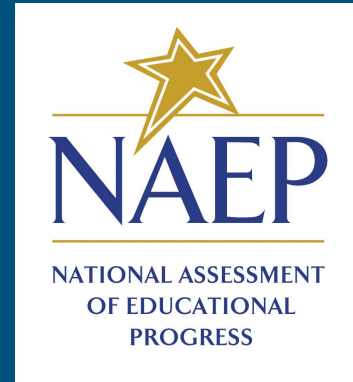
	Id	State	Year	Enrolled	TotalRevenue	FederalRevenue	StateRevenue	LocalRevenue	InstructionExpenditure
0	2003_ALABAMA	ALABAMA	2003	727900.0	5196054.0	567704.0	2966981.0	1661369.0	2817111.0
1	2003_ALASKA	ALASKA	2003	133303.0	1425948.0	259423.0	813371.0	353154.0	763525.0
2	2003_ARIZONA	ARIZONA	2003	875111.0	6529894.0	740579.0	2912629.0	2876686.0	2810907.0
3	2003_ARKANSAS	ARKANSAS	2003	450158.0	3241275.0	379947.0	2394336.0	466992.0	1768713.0
4	2003_CALIFORNIA	CALIFORNIA	2003	6226552.0	59815855.0	5795655.0	33617766.0	20402434.0	29561563.0
...
352	2015_VIRGINIA	VIRGINIA	2015	1279867.0	15857524.0	1012205.0	6240349.0	8604970.0	8755896.0
353	2015_WASHINGTON	WASHINGTON	2015	1072359.0	13709442.0	1036422.0	8293812.0	4379208.0	6508964.0
354	2015_WEST_VIRGINIA	WEST_VIRGINIA	2015	279565.0	3478401.0	362959.0	1979466.0	1135976.0	1819903.0
355	2015_WISCONSIN	WISCONSIN	2015	861813.0	11637376.0	814385.0	5869265.0	4953726.0	5723474.0
356	2015_WYOMING	WYOMING	2015	93867.0	1962874.0	120290.0	1116917.0	725667.0	895910.0

AvgMath4Score	AvgMath8Score	AvgReading4Score	AvgReading8Score
223.0	262.0	207.0	253.0
233.0	279.0	212.0	256.0
229.0	271.0	209.0	255.0
229.0	266.0	214.0	258.0
227.0	267.0	206.0	251.0
...
247.0	288.0	229.0	267.0
245.0	287.0	226.0	267.0
235.0	271.0	216.0	260.0
243.0	289.0	223.0	270.0
247.0	287.0	228.0	269.0

What is NAEP?

The National Assessment of Educational Progress (NAEP) is the only assessment that measures what U.S. students know and can do in various subjects across the nation, states and in some urban districts.

NAEP is given to a representative sample of students across the country rather than the entire student population.



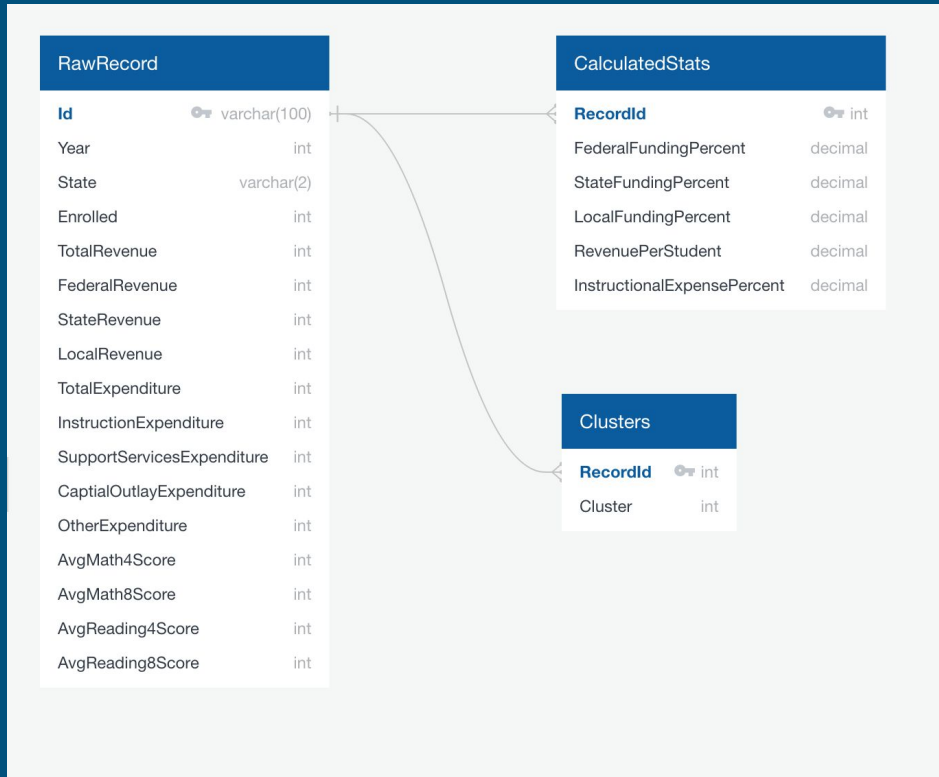
*From nces.ed.gov

Questions to answer:

1. Are there correlations between funding and/or expenditures and the average state NAEP Math/Reading test scores for 4th and 8th grade?
2. How does the funding and expenditures of higher performing states (as defined by higher average NAEP Math/Reading scores) compare to lower performing states?

By exploring the questions, we hope to provide advice/guidance to state decision makers as they advocate for funding or spend dollars on school budgets.

After investigating the CSV file (and the respective columns contained within), the team decided to create the following ERD:

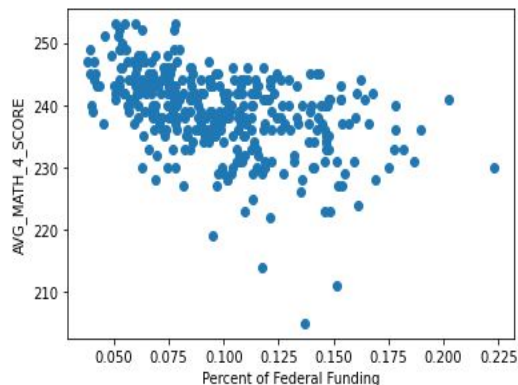


Exploring the Data:

What relationship exists between funding and 4th grade math Scores?

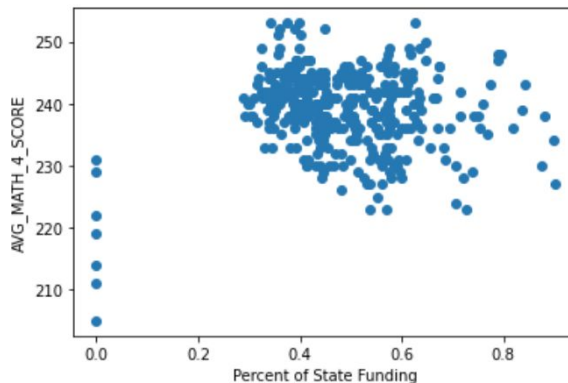
Federal Funding

```
# How does the percentage of federal funding correlate to math grades?  
plt.scatter(df2['FEDERAL_REVENUE']/df2['TOTAL_REVENUE'], df2['AVG_MATH_4_SCORE'])  
plt.xlabel('Percent of Federal Funding')  
plt.ylabel('AVG_MATH_4_SCORE')  
plt.show()
```



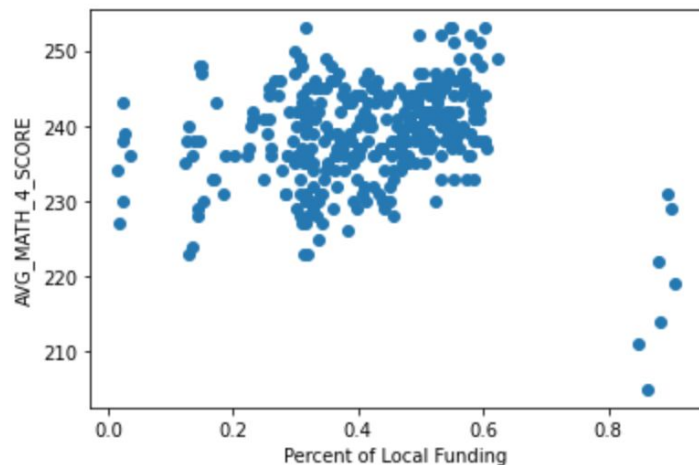
State Funding

```
# How does the percentage of state funding correlate to math grades?  
plt.scatter(df2['STATE_REVENUE']/df2['TOTAL_REVENUE'], df2['AVG_MATH_4_SCORE'])  
plt.xlabel('Percent of State Funding')  
plt.ylabel('AVG_MATH_4_SCORE')  
plt.show()
```



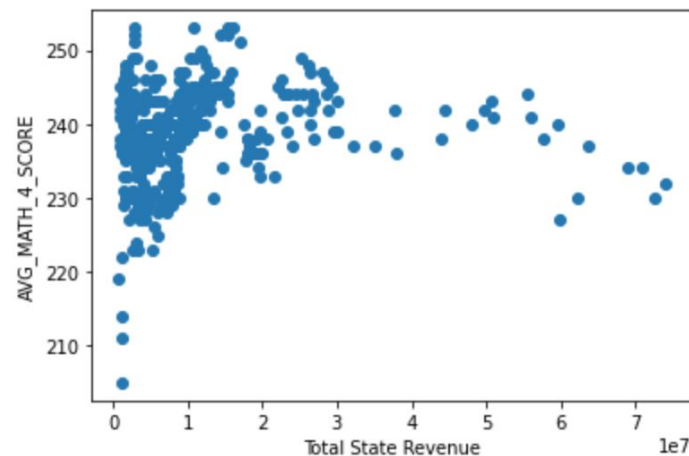
Local Funding

```
# How does the percentage of local funding correlate  
plt.scatter(df2['LOCAL_REVENUE']/df2['TOTAL_REVENUE'],  
plt.xlabel('Percent of Local Funding')  
plt.ylabel('AVG_MATH_4_SCORE')  
plt.show()
```



Total Revenue


```
#Looking for relationship between total state revenue and a  
plt.scatter(df2['TOTAL_REVENUE'], df2['AVG_MATH_4_SCORE'])  
plt.xlabel('Total State Revenue')  
plt.ylabel('AVG_MATH_4_SCORE')  
plt.show()
```



Model Factors

We transformed the raw data by calculating the percentage of total funding received from federal, state and local. We did this so that we could more evenly compare state of different sizes.

We also calculated the revenue per student and the percent of expenditures devoted to instruction.

CalculatedStats	
RecordId	 int
FederalFundingPercent	decimal
StateFundingPercent	decimal
LocalFundingPercent	decimal
RevenuePerStudent	decimal
InstructionalExpensePercent	decimal

Normality

Shapiro-Wilk Test

```
from scipy.stats import shapiro
#note: used the following link for guidance on how to interpret the results
#https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-stat
```

```
#AVG 4th Grade Math Score
stat, pvalue = shapiro(df3['AVG_MATH_4_SCORE'])
print(f'Statistics = {stat}, pvalue = {pvalue}')

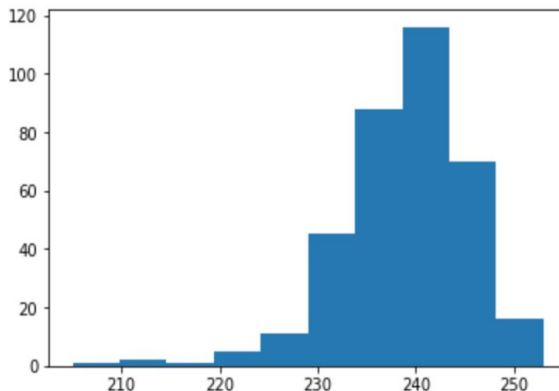
#interpret Results
alpha = 0.05
if pvalue > alpha:
    print('Sample looks normally distributed (fail to reject null hypothesis)')
else:
    print('Sample does not look normally distributed (reject null hypothesis)')
```

```
Statistics = 0.9533835053443909, pvalue = 3.596219988466487e-09
Sample does not look normally distributed (reject null hypothesis)
```

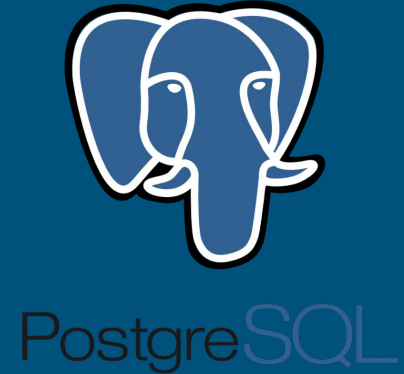
Is our data normalized?

```
: #creating Histogram for the AVG 4th Grade Math Score
pyplot.hist(df3['AVG_MATH_4_SCORE'])

: (array([ 1.,  2.,  1.,  5., 11., 45., 88., 116., 70., 16.]),
   array([205. , 209.8, 214.6, 219.4, 224.2, 229. , 233.8, 238.6, 243.4,
         248.2, 253. ]),
   <BarContainer object of 10 artists>)
```



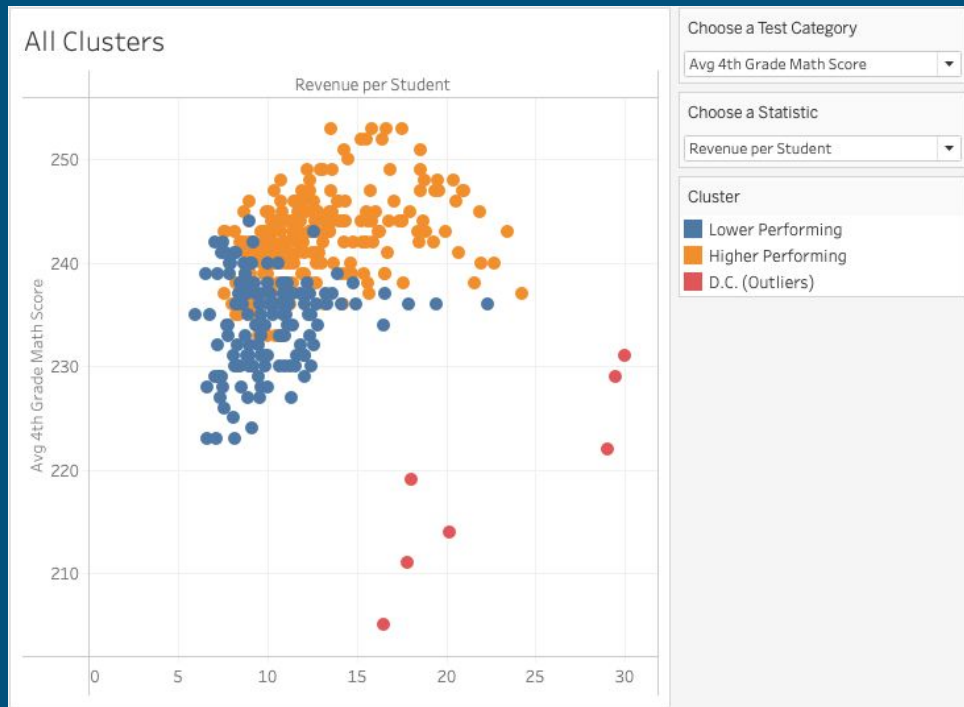
Technologies



Analysis

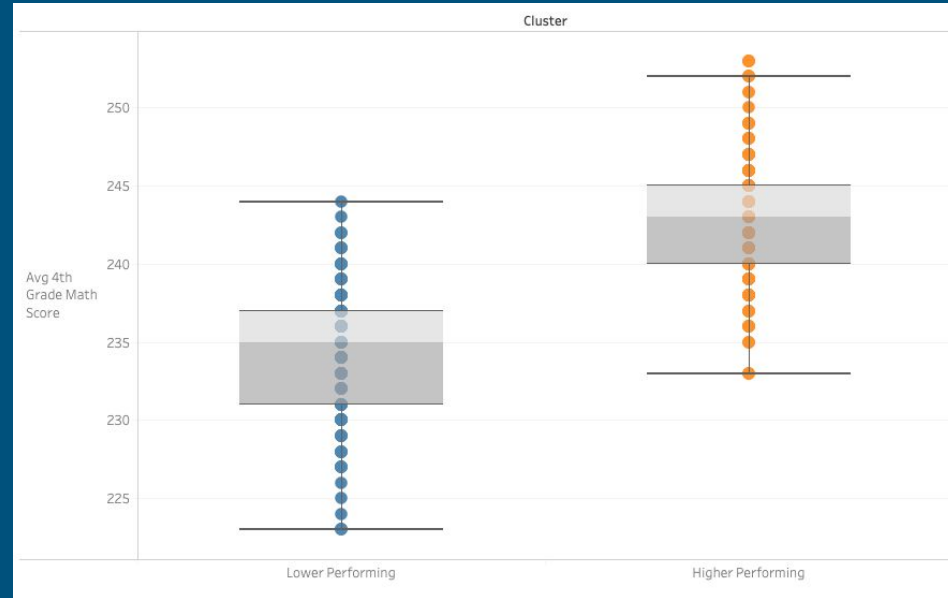
We utilized a KMeans unsupervised classification model to identify clusters of data points with similar funding and test scores in an attempt to identify higher and lower performing samples for our comparisons.

After some experimentation we decided using a $K = 3$ was ideal. 1 of the clusters was all data points from D.C. which we decided to treat as outliers and the remaining clusters functioned as our higher and lower performing samples.

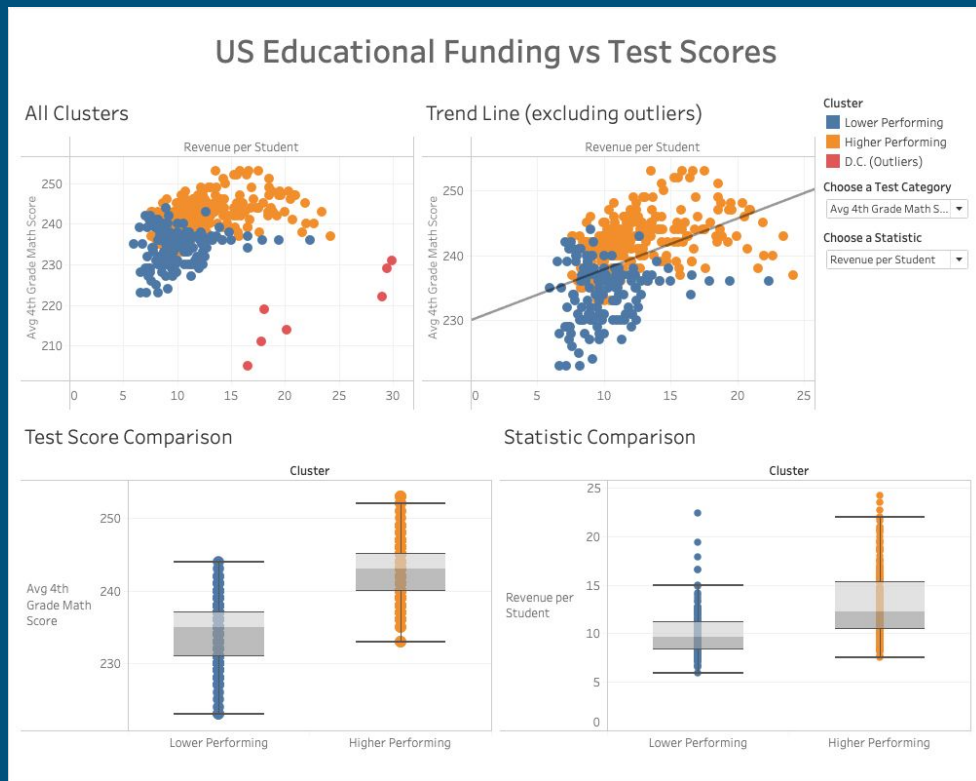


Test Score Comparisons

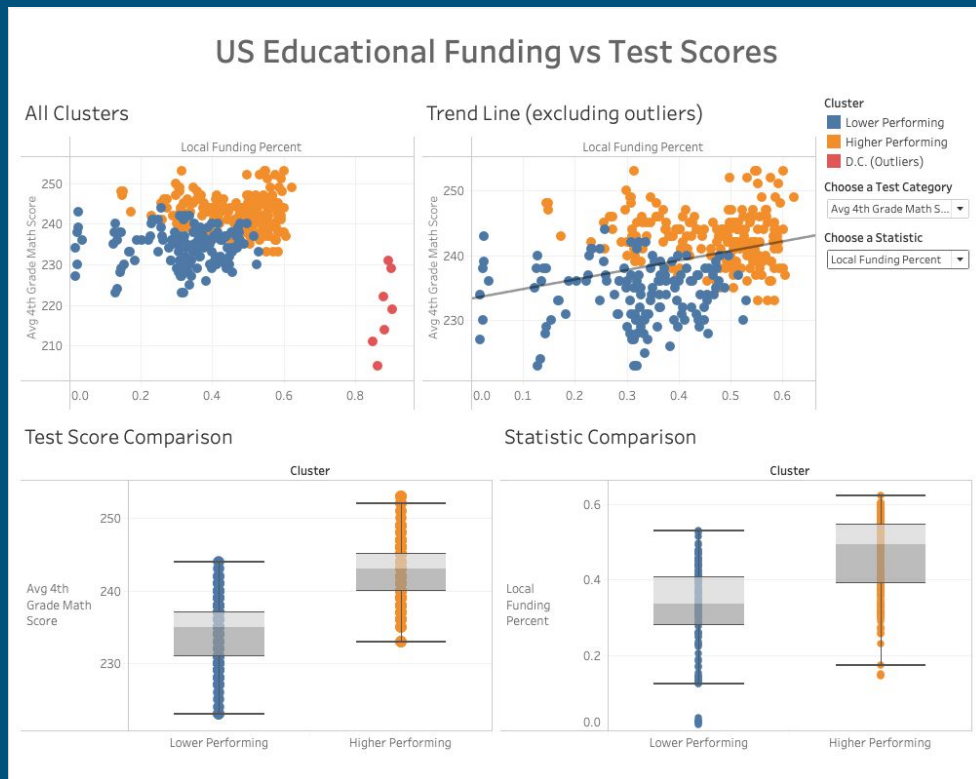
- 4th Grade Math
 - Mean test scores of 234 vs 243
 - p-value of $7.05e-54$
- 4th Grade Reading
 - Mean test scores of 214 vs 224
 - p-value of $7.80e-66$
- 8th Grade Math
 - Mean test scores of 275 vs 286
 - p-value of $1.26e-57$
- 8th Grade Reading
 - Mean test scores of 258 vs 267
 - p-value of $4.86e-68$



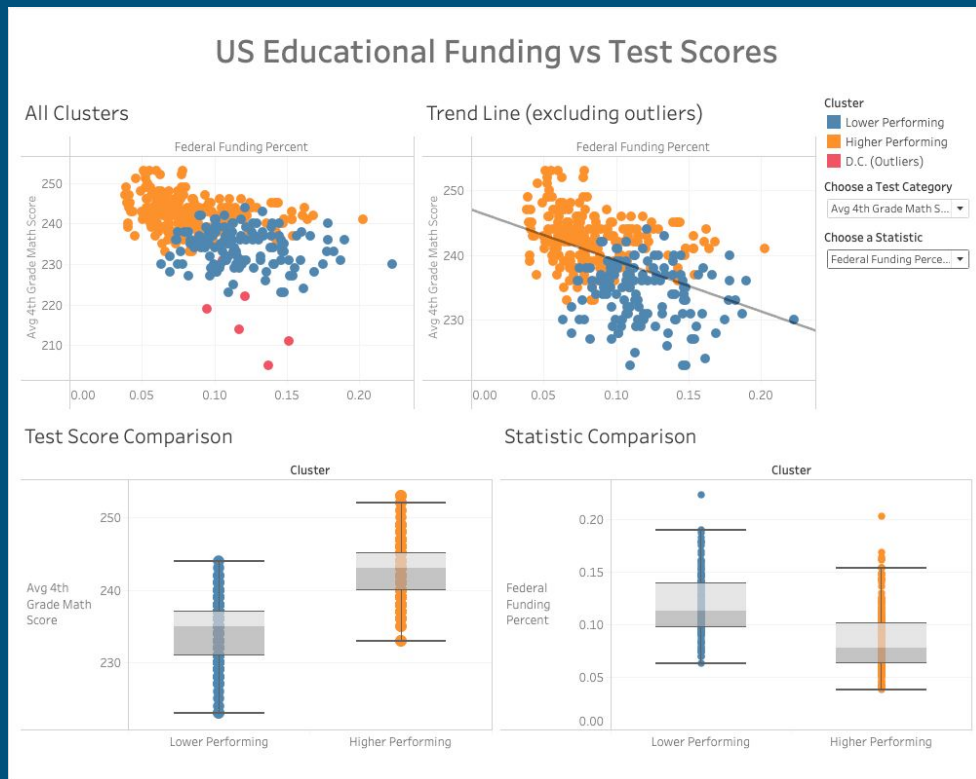
Positive Correlation - Revenue per Student



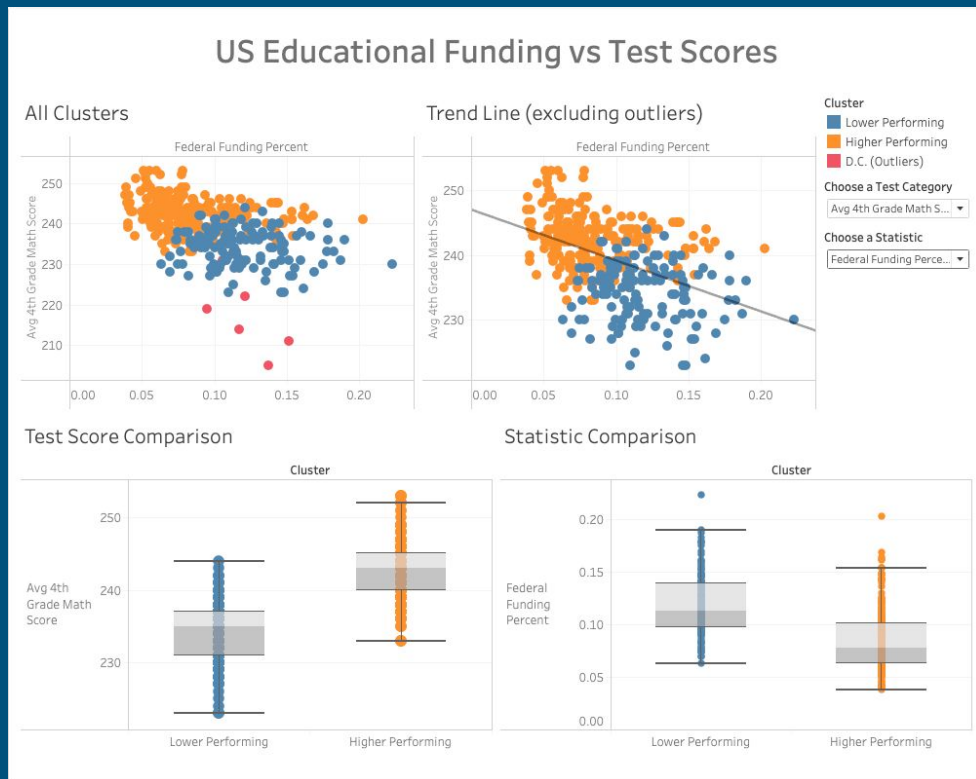
Positive Correlation - Local Funding Percentage



Negative Correlation - Federal Funding Percentage



Negative Correlation - State Funding Percentage



Recommendation for future analysis

- With the understanding that state and federal funding appear to have a negative correlation, more analysis would be needed to understand why this relationship exists.
- Additionally, while we understand that local funding and revenue per student have a positive correlation, focusing in on a specific state (specifically those in a higher performing cluster) and drilling down into how funds are exactly spent when coming from different revenue streams would likely provide further insight.
- Finally, taking a look at state specific testing grades (e.g., TEKS for the state of Texas) could provide another avenue for analysis.

What we could have done differently

1. Curriculum standards
2. Added a Multi Linear Regression Model to improve fit
3. Inflation
4. We could have explored how this analysis may have changed over time utilizing the year factor of our data.

Acknowledgements

- [Kaggle.com](https://www.kaggle.com)
- publicschoolreview.com
- nces.ed.gov

*Data is sourced from the U.S. Census Bureau and the National Center for Education Statistics (NCES)

Q & A