Analysis on the US Eduations Dataset

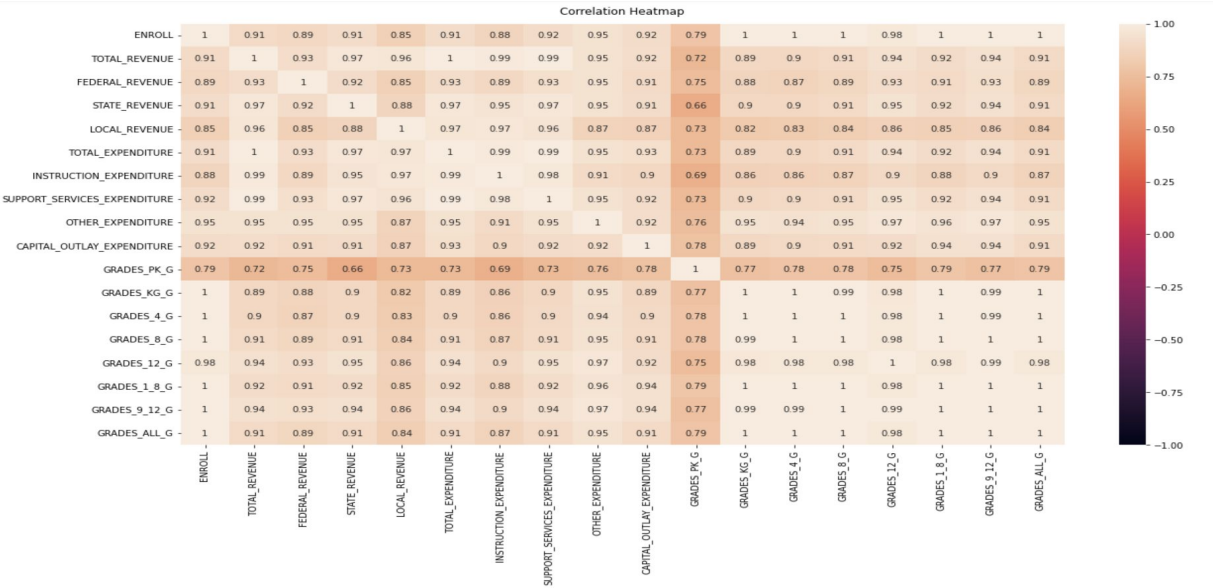
Gokul Sunilkumar	Xiaolin Sima	Lei Liu	Rachel Peng	Ziyao Zhou
gs3214	xs2483	ll3442	xp2197	zz2915

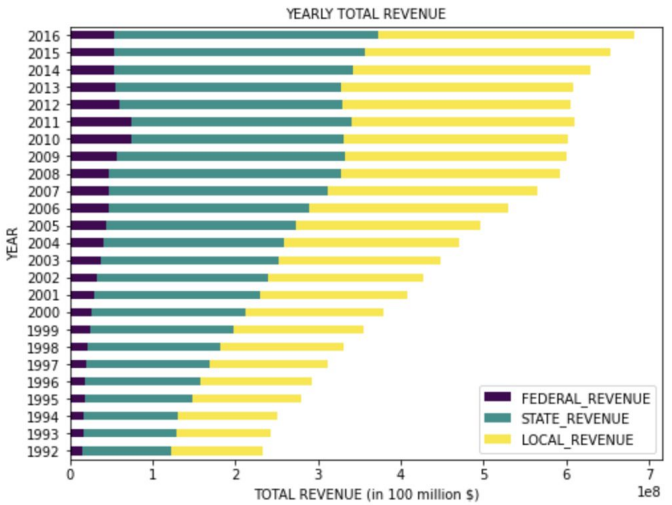
Dataset: US Educations : Unification Project
Source: https://www.kaggle.com/datasets/noriuk/us-education-datasets-unification-project
Type of Problem to be solved: Regression

Features: `'STATE', 'YEAR', 'ENROLL', 'TOTAL_REVENUE', 'FEDERAL REVENUE', 'STATE REVENUE', 'LOCAL REVENUE', 'TOTAL EXPENDITURE', 'INSTRUCTION EXPENDITURE', 'SUPPORT SERVICES EXPENDITURE', 'OTHER EXPENDITURE', 'CAPITAL OUTLAY EXPENDITURE', 'GRADES PK G', 'GRADES KG G', 'GRADES 4 G', 'GRADES 8 G', 'GRADES 12 G', 'GRADES 1 8 G', 'GRADES 9 12 G', 'GRADES ALL G', 'AVG MATH 4 SCORE', 'AVG_MATH_8_SCORE', 'AVG_READING_4_SCORE', 'AVG_READING_8_SCORE'`

Target Variable: `TOTAL_SCORE`

Initial Data Exploration





- The original data set consists of over 20 different features and we intended to reduce the dimensionality without decreasing the explanatory power of our model. Hence, the natural first step would be to explore the correlations between these variables. According to the graph above, many of the features are actually highly correlated.

- Therefore, we should drop some of the features with respect to the threshold of Corr = 0.97 to enhance the efficiency of our training process, and dropping the highly correlated features would not decrease the model accuracy significantly as they do not bring in much additional information.

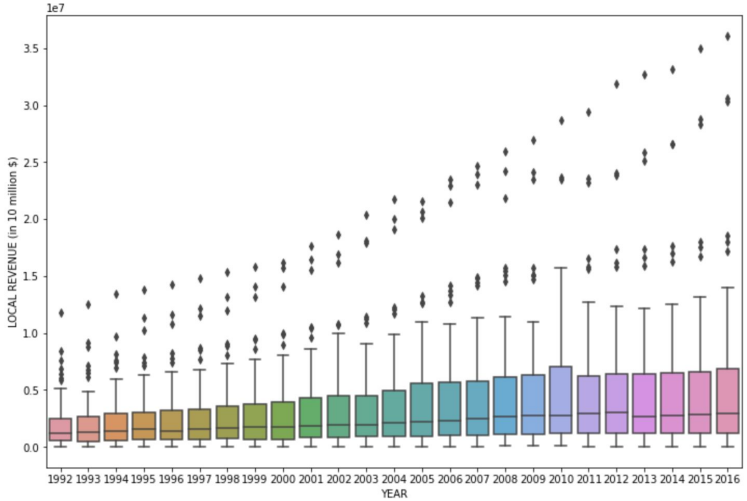
- Looking at the distribution of the total revenue across the years from 1992 to 2016, we see that there hasn't been a significant increase in the federal revenue as compared to that of state and local revenue, although the total revenue has risen steeply. This means that the local and state governments are the predominant ones interested in improving infrastructure and facilities for the enhancement of students' educational quality as against the federal government.

Cleaning and Sampling

Features	18
Datapoints	1715
Correlated features (with threshold 0.97)	11

Column-wise missing values

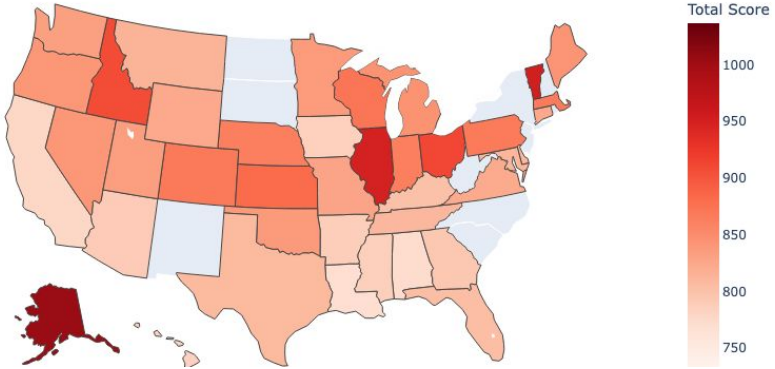
```
PRIMARY_KEY                    0
STATE                          0
YEAR                           0
ENROLL                       491
TOTAL_REVENUE                440
FEDERAL_REVENUE              440
LOCAL_REVENUE                440
OTHER_EXPENDITURE            491
CAPITAL_OUTLAY_EXPENDITURE   440
GRADES_PK_G                  173
AVG_MATH_4_SCORE            1150
AVG_MATH_8_SCORE            1113
AVG_READING_4_SCORE         1065
AVG_READING_8_SCORE         1153
dtype: int64
```



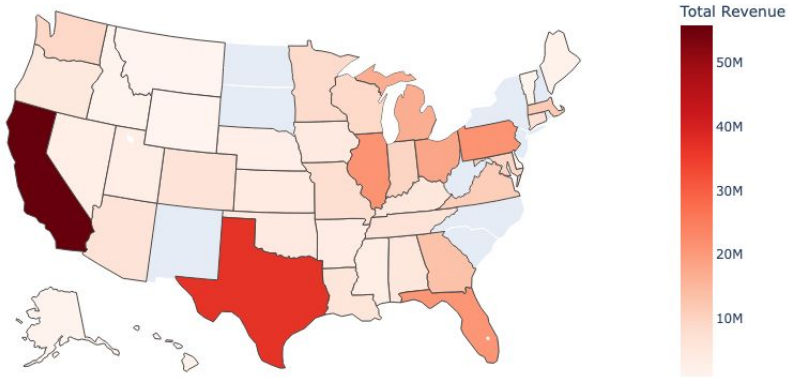
- For the purpose of model training, we have created a target variable called TOTAL_SCORE, that is the sum of the average math and reading scores in grades 4 and 8.
- Please note that we have not used the GRADES_ALL column from the extended dataset as it is the total score obtained by all students in the state. This is a biased feature to predict as it does not take into account the number of student enrollments in the state into account. Hence, it'll always be biased towards states like California, which although having less-performing students could will be predicted to be the best-performing by the model due to its huge population.
- If we drop the rows with missing values, we will end up with a dataset that is very small in size (approximately 300 rows), which is a very small number and would underfit due to curse of dimensionality.
- In order to better use the distribution of the existing data at hand and not lose information for training, we resort to imputation of the missing values based on the mean value of the non-missing datapoints within the same state.
- We have scaled the data using a Robust Scaler as the distribution of some of the features contain significant amount of outliers. For instance, a boxplot of the distribution of local revenue across the years help us in visualising this. Hence, using a Standard or Min-Max Scaler could potentially introduce skewness in the data towards the direction of the outliers.

States with high total revenue are states with large number of students, but not necessarily states with high total score

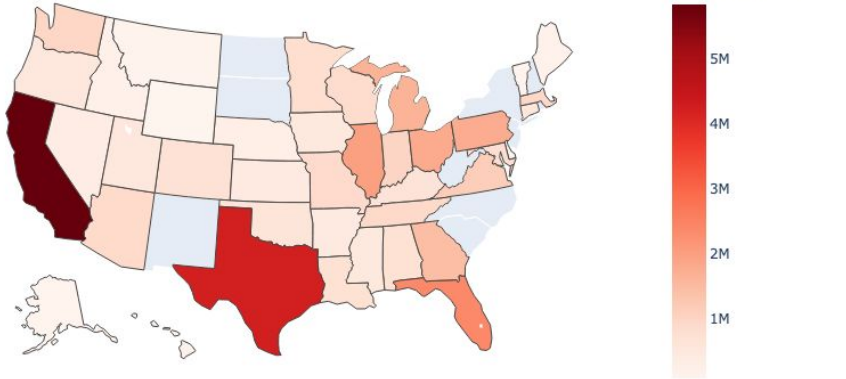
Total Score by State



Total Revenue by State

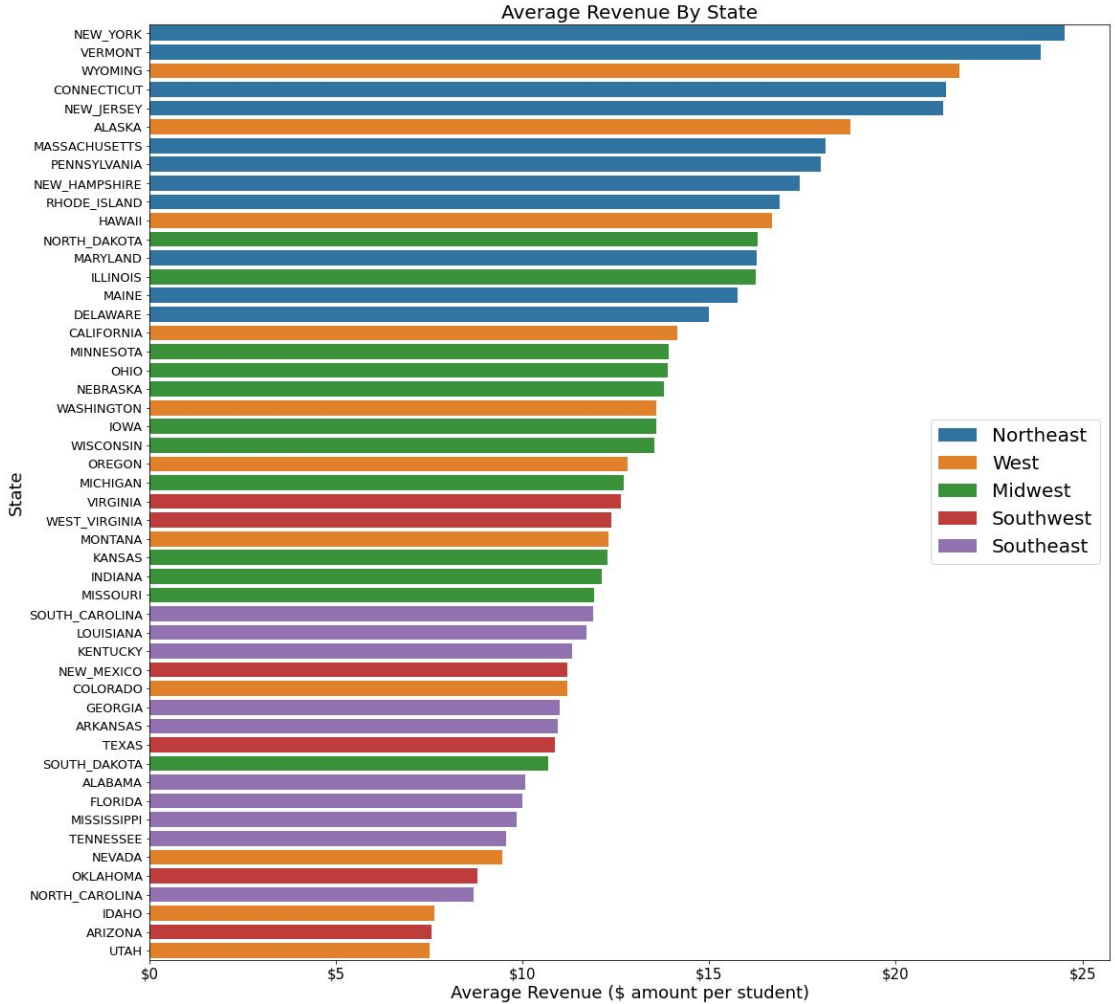


Total Students by State



States with highest total score across the years are Alaska, Illinois, and Vermont.

States with high total revenue across the years such as California, Texas and Florida are not the states with the highest total score. This may due to the fact that these states also rank among the top of the states that have a large number of total students.



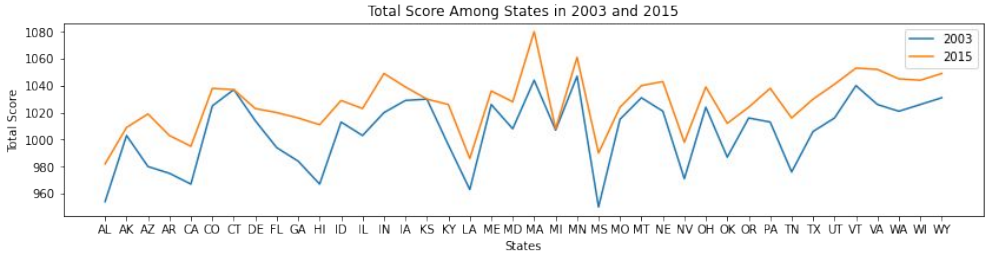
Average revenue per student were the highest in Northeast and lowest in South in 2016

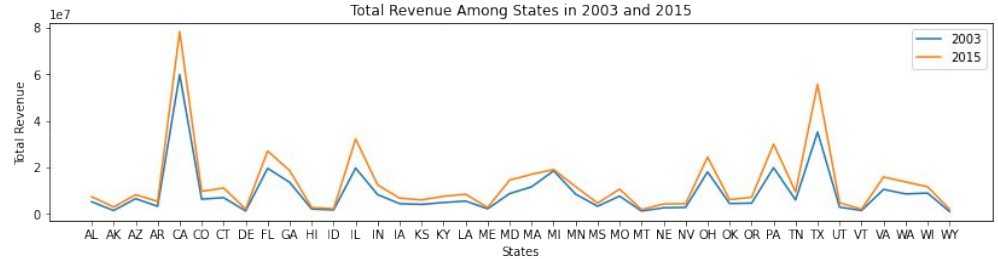
As a first step, we explored the geographic patterns for average revenue per student by state in 2016. The average revenue is calculated by using total amount of revenue for the state divided by the count of all students in that state, both of which are columns in the data.

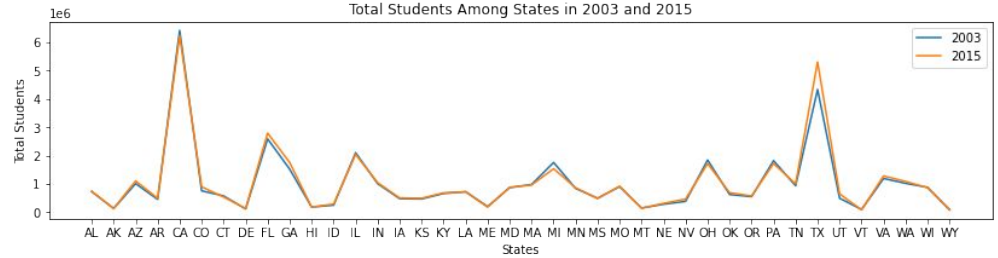
The plot shows that Northeast states had the highest average revenue, followed by West states, while South states had the least average revenue. We performed the same analysis for average expenditure and observed a similar pattern.

The results were in line with our hypothesis that dispersion exists in both revenue and expenditure across states. In later analysis, we looked at if they have any indication for students' academic performance.

Total revenue and number of students remain relatively the same between 2003 and 2015 among the states, but changes occur for total score





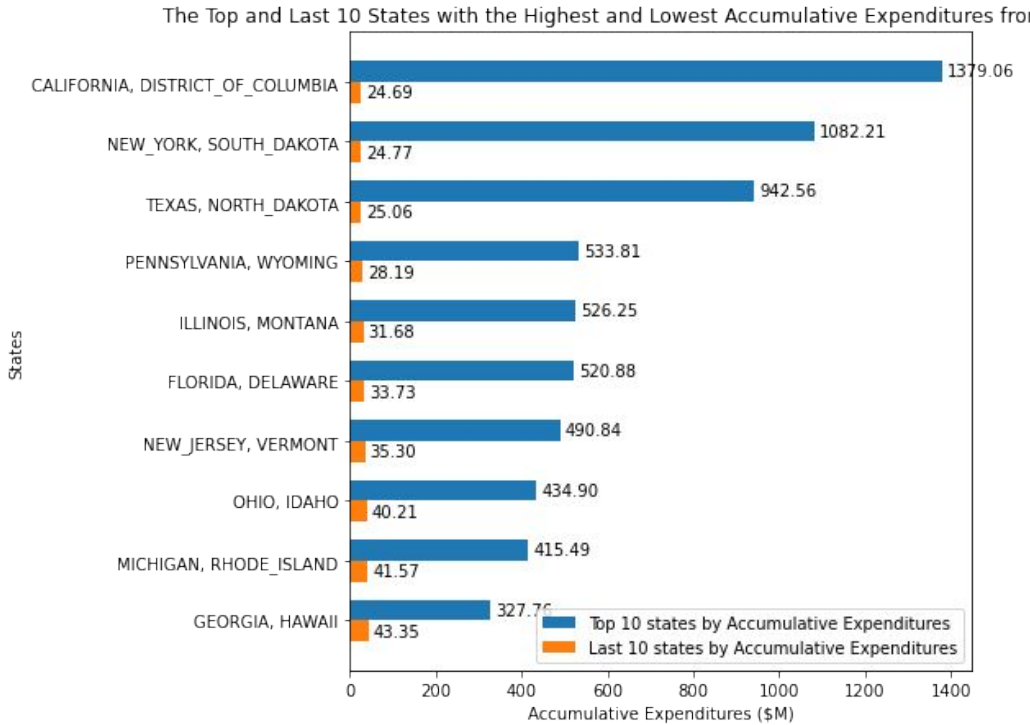


For the count of total score, we can see that overall, the scores for different states increased from 2003 to 2015. However, there are vibrations. For example, some of the states had a large increase. The top being Hawaii, with a total score increase of 44 points. It was followed by Mississippi and Tennessee, each of which increased by 40 points. Some of the states did not change much, especially Connecticut and Kansas, where total score remained the same. Relative ranking between states changed as an result.

For the amount of total revenue, we can see that almost all the states had a higher total revenue in 2015 as compared to 2003, but the overall trend remained the same.

For the number of total students, we can see that the number barely changed among the states in 2003 compared to 2015, as the two lines almost align with each other. Some minor differences occurred such as Michigan had slightly less students and Texas had slightly more students in 2015.

Accumulated expenditures shows educational disparities among states

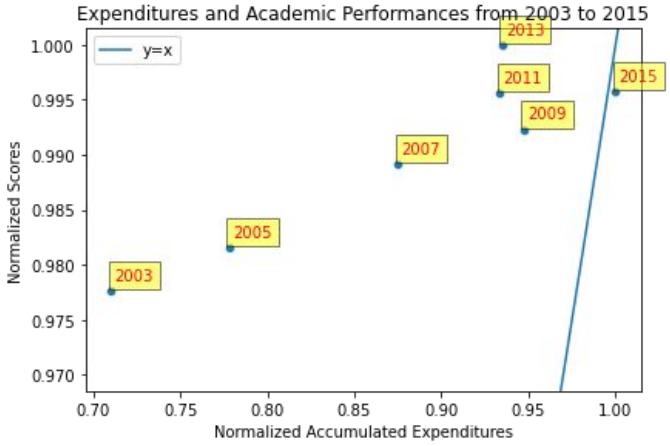


The figure compares the accumulated total expenditures, where the top 10 states with the highest expenditures are compared with the lowest 10 states.

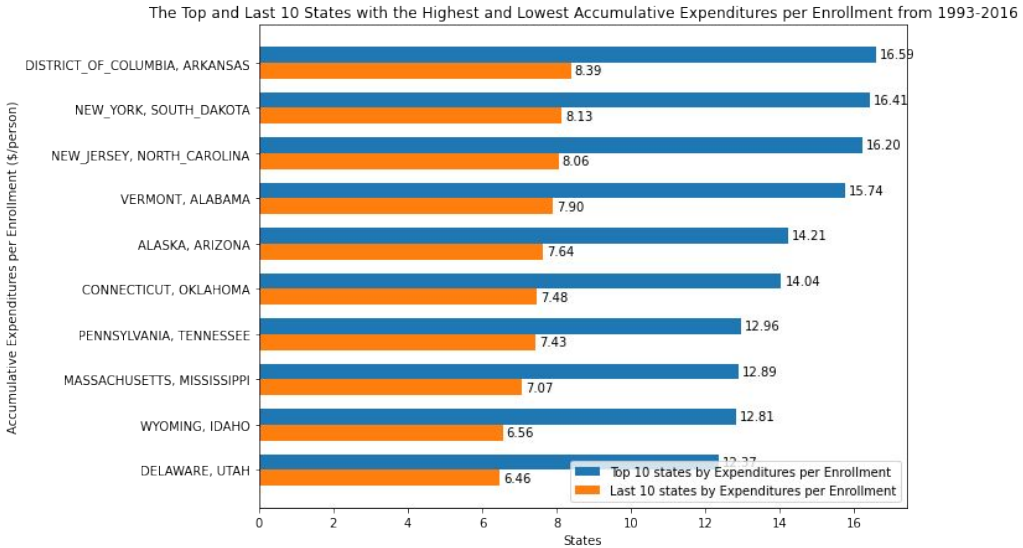
From 1993 to 2016, California, the state with the highest accumulated expenditures, has spent about 50 times more than the District of Columbia, the state with the lowest accumulated expenditures, on educational instruction, support services, and capital overlay. The disparity among the top 10 and last 10 states decreases rather drastically from about 50 to 7 times.

Among the lowest 10 states, the expenditures are different within less than 2 folds. However, more than 4 folds disparities exist among the highest 10 states. This suggests that more disparities exist among the top spending states, but in the meanwhile, the low spending states have much more consistent spendings.

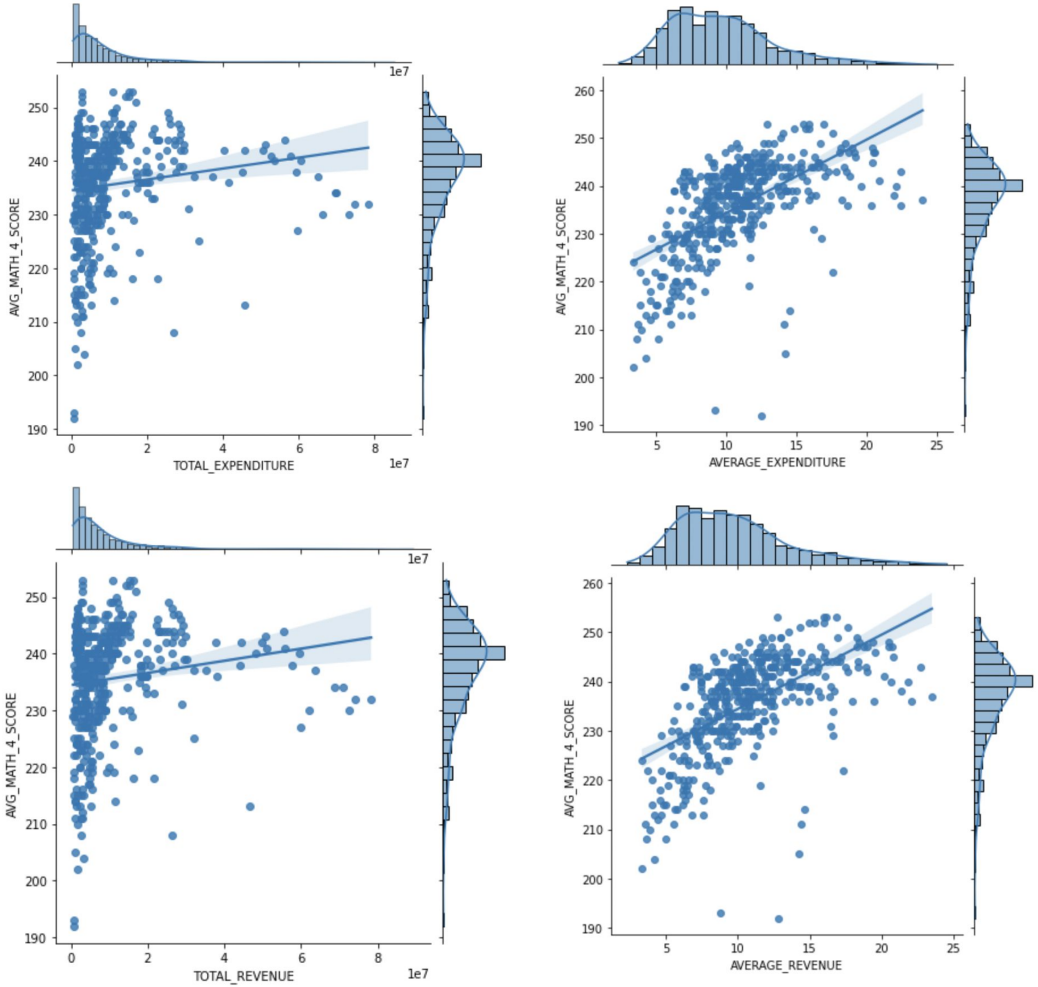
Expenditures Per Enrollment shows less but constant disparity, while academic performance is unaffected over years



This scatter plot shows the lack of significant correlation between national accumulated expenditures and scores per year, indicating that the amount spent in all states on education doesn't affect student's test scores, as they stay constant over years.



The figure above takes into account of enrollment and compares accumulated expenditures per enrollment among states. New York, New Jersey, and Pennsylvania remain to be in the top 10, and South Dakota remains in the last 10. The top 10 states' accumulated expenditures are roughly consistent twice as much as the last 10 states. This consistency suggests that disparity on money spent per student is large and prevalent in the US.



Average revenue & expenditure positively correlate with scores

We're interested in examining the relationship between academic performance (score) and revenue as well as expenditure. Interestingly, the distributions of total revenue and total expenditure are heavily right skewed, and there is no clear relationship vs. score. However, once we adjusted for the count of students in each state, we saw a clear linear relationship between score and average revenue as well as average expenditure. This finding suggests that the average statistics can be useful features for fitting a regression model down the road.

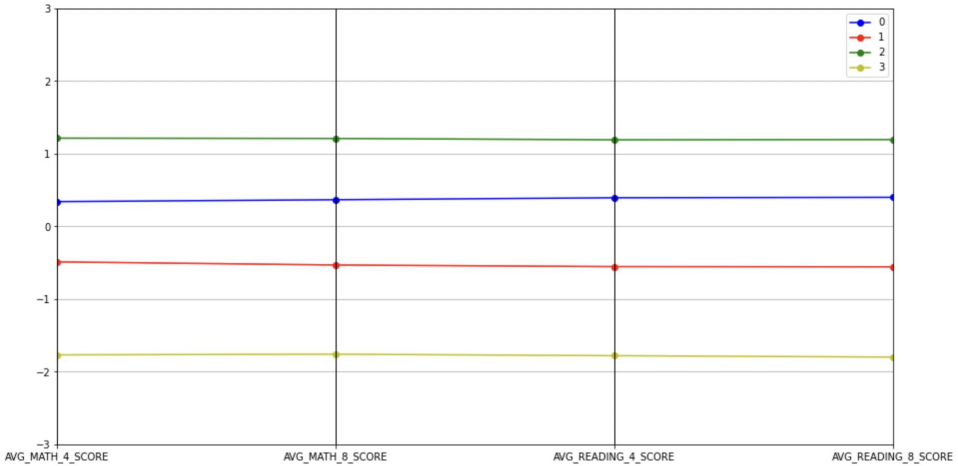
Y-axis: Average score for fourth graders taking the NAEP math exam (Patterns are similar for math score for eighth graders and reading scores for both graders based on our work)

X-axis:

- Top left: Total expenditure
- Top right: Average expenditure
- Bottom left: Total revenue
- Bottom right: Average revenue

Clustering analysis using K-Means

	AVG_MATH_4_SCORE	AVG_MATH_8_SCORE	AVG_READING_4_SCORE	AVG_READING_8_SCORE	prediction
0	0.341530	0.367440	0.394586	0.400491	0
1	-0.487957	-0.530855	-0.553217	-0.556145	1
2	1.214785	1.210982	1.192256	1.195740	2
3	-1.767774	-1.758712	-1.779371	-1.798412	3



Finally, we performed a clustering analysis to further understand the structure of our data.

First, we used K-Means algorithm to perform clustering analysis:

1. We removed all the missing data. Each row in our dataset represents the education daa of a state in a given year.
2. We selected the academic scores as the features to fit K-Means. Therefore we have 4 clusters. The cluster ID now becomes a column, say “prediction”, in our data.

Next, we used a parallel coordinate to visualize the results:

- In a parallel coordinates plot, variables are represented through vertical parallel lines. These lines form the axes for the plot.
- Each data point of the multivariate data is marked in these vertical axes which result in a polyline. One data point corresponds to one polyline, which is a cluster in our case.

From this plot, we can see that there is no line crossing each other, indicating that clusters that have low math scores also have low reading scores. Since points in clusters are just scores by state, we can make a conclusion that states that score poorly on one metric score poorly in all categories.

Machine Learning Techniques Proposed to be Implemented

Linear Regression	Baseline Model
Ridge Regression	Ridge Regression model is retained since it can handle correlated features by shrinking them evenly without dropping them
Elastic Net Regression	L1 regularization is not ideal for our project. However, we still want to see how our model interacts with L1 regularization and thus we include Elastic Net and bring in a mix of L1 and L2 regularization to see whether it could outperform the vanilla Ridge Regression model.
~~**Lasso Regression**~~	Lasso Regression has been crossed out for two reasons. Firstly, we have already narrowed down to seven features after dropping the highly correlated features, and we shall train our model based on those features instead of possibly dropping more in the training process. Secondly, although we have dropped those highly correlated ones, some of the pairs still pertain to a high correlation of around 0.9
Random Forest	Although tree-based models have drawbacks (mentioned below), Random Forest has been retained in order to compare the performance of the linear models against that of a tree based model and pick the best (our intuition at this point, based on the characteristics of the dataset is that the linear model will generalise well and perform better.)
~~**XGBoost**~~	One of the most critical drawbacks of tree-based models is that they cannot extrapolate. They are making inferences based on the observed means. However, if we tend to apply our model to datasets beyond the scope of our training set, especially to apply it to years or certain states with better performances, the tree-based model could underperform compared to other models. Furthermore, from our data exploration, it is clear that if we plot the average expenditure against the total score, we could get a better-fitted line, indicating that our focus should mainly be linear models. Hence, we drop XGBoost.