

NAME: Awojobi Oluwaseyitan

COURSE: INFO 5709

STUDENT ID: 11087659

Project 4

## INTRODUCTION

For this project, I got a dataset off Kaggle of the US education unification project. This dataset had 1493 rows and several dimensions. There were several unnecessary variables and missing values in the dataset hence it had to go through a data cleaning process.

We remove all columns where there are large amounts of missing values, or which are not relevant to our analysis. Our goal is to find out the relationship between variables such as the State, the Revenue allocated, total expenditure, scores and the number of enrollments. Different visualization techniques were utilized such as using the filter tool, using different spatial layouts, re-ordering charts and color encoding.

## Data Attributes:

In this dataset, we have various attributes such as the

**State-Variou states within the USA**

**Year- Year involved.**

**Total revenue- Total money assigned to each state per year.**

**Capital outlay expenditure- Expenses incurred on capital projects.**

**Total Expenditure- Total amount spent on expenses**

**Instruction Expenditure- Expenditures for activities related to the interaction between teachers and students. This includes salaries and benefits.**

**Grades PK G- This Is the pre-kindergarten classes for children**

**Grades KG- Kindergarten class**

**Average reading score**

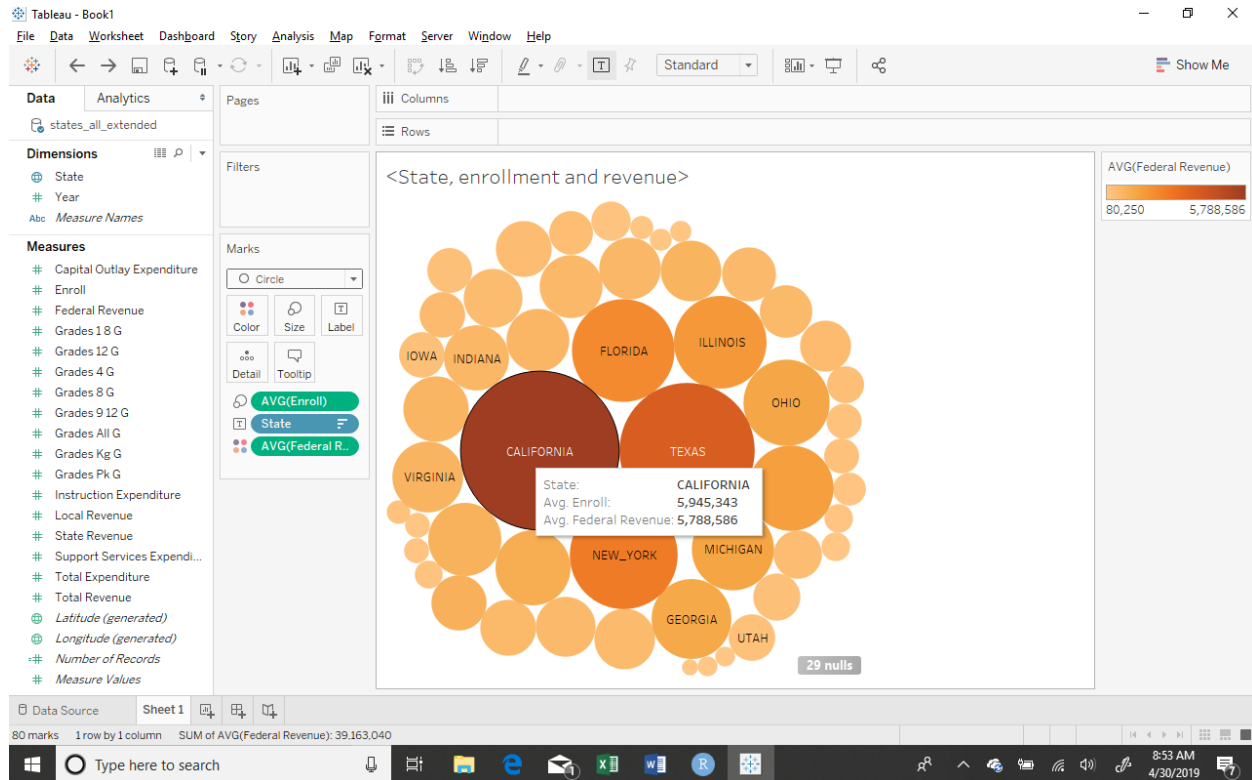
**Average Math score**

## Hypothesis:

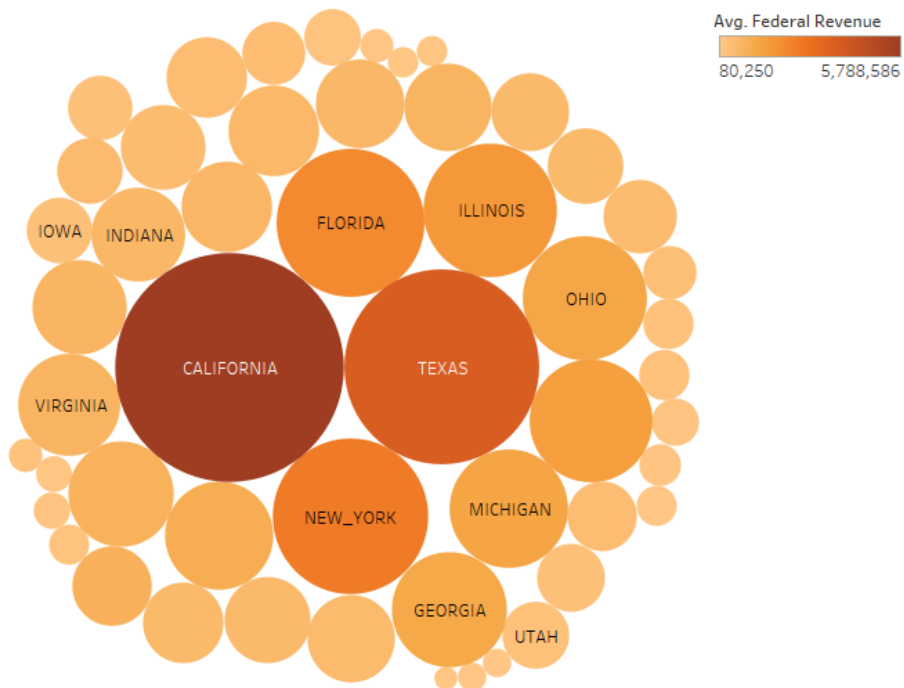
1. Ho: Revenue allocated is directly proportional to expenses incurred.  
H1: Revenue allocated is not directly proportional to expenses incurred
2. Ho: Increase in instruction expenditure does not guarantee an increase in enrolled students  
H1: Increase in instruction expenditure guarantees an increase in enrolled students
3. Ho: Highly populated states receive higher educational funding than lower populated states  
H1: Highly populated states do not receive higher educational funding than lower populated states

These hypotheses can be answered using the given dataset as it contains components such as the State involved, Federal revenue allocation, Enrolled students and applicable scores needed for our analysis.

First, we do an interactive visualization to check the city which has the highest allocation of Federal revenue and highest enrollment rate.



<State, enrollment and revenue>










State. Color shows average of Federal Revenue. Size shows average of Enroll. The marks are labeled by State.

We found that California has the highest allocation followed by to Texas, New York, Florida, Ohio amongst others.

**N: B: The darker the color, the higher the average Federal revenue allocated and the larger the size, the higher the number of students enrolled. This spatial layout was chosen because it conveys the message of our hypothesis better. The color also is bold and differentiates effectively the various categories/states**

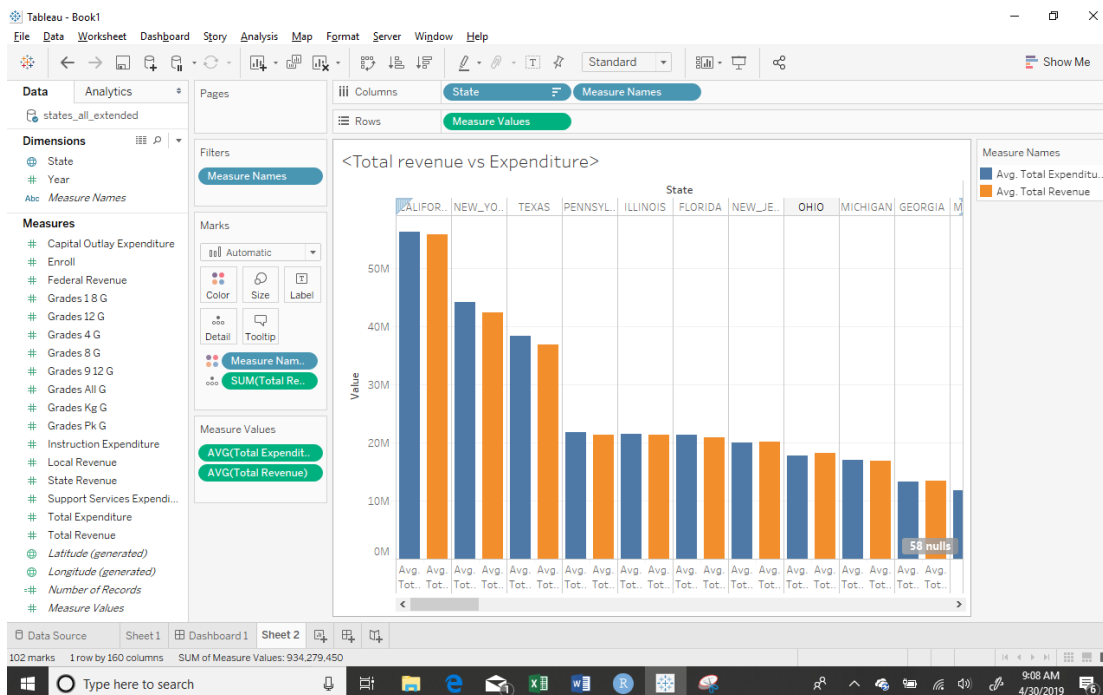
We check this data by finding the states with the highest population in the United States. As shown below, we see that each state is allocated revenue based on the population.

RANK	STATE	POPULATION	AREA (SQ MILES)
1	 California	39,536,650 (2017)	163,695.73
2	 Texas	28,304,600 (2017)	268,820.15
3	 Florida	20,984,400 (2017)	65,754.74
4	 New York	19,849,400 (2017)	54,556.24
5	 Pennsylvania	12,805,540 (2017)	46,055.42
6	 Illinois	12,802,020 (2017)	57,914.55
7	 Ohio	11,658,610 (2017)	44,824.92

## Conclusion:

***This answers our third hypothesis and confirms that highly populated cities receive more funding hence we accept the null hypothesis.***

Next, we seek to find out how funds are being spent in comparison with the amount received.

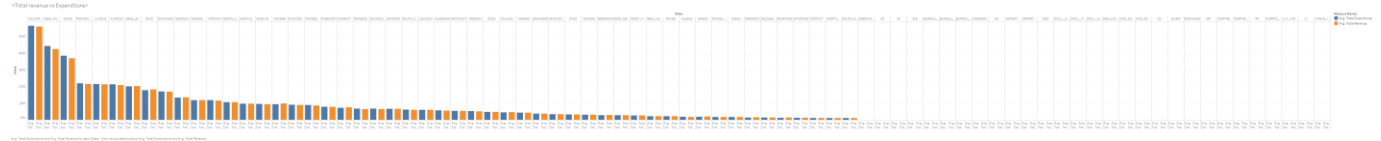


**N: B- The blue tab represents the average total expenditure while the orange tab represents the average total revenue.**

**Here, a bar graph is used because it differentiates and visualized the different dataset (revenue and expenditure) accurately. It also makes it easy to compare the relationship between both the revenue and the expenditure.**

Upon exploring the image above, we see that most states spend almost as much funds as they were allocated per year. We see California, New york, Texas and Pennsylvania toping the chart here.

This shows us that budgets are being well defined hence no severe excesses.

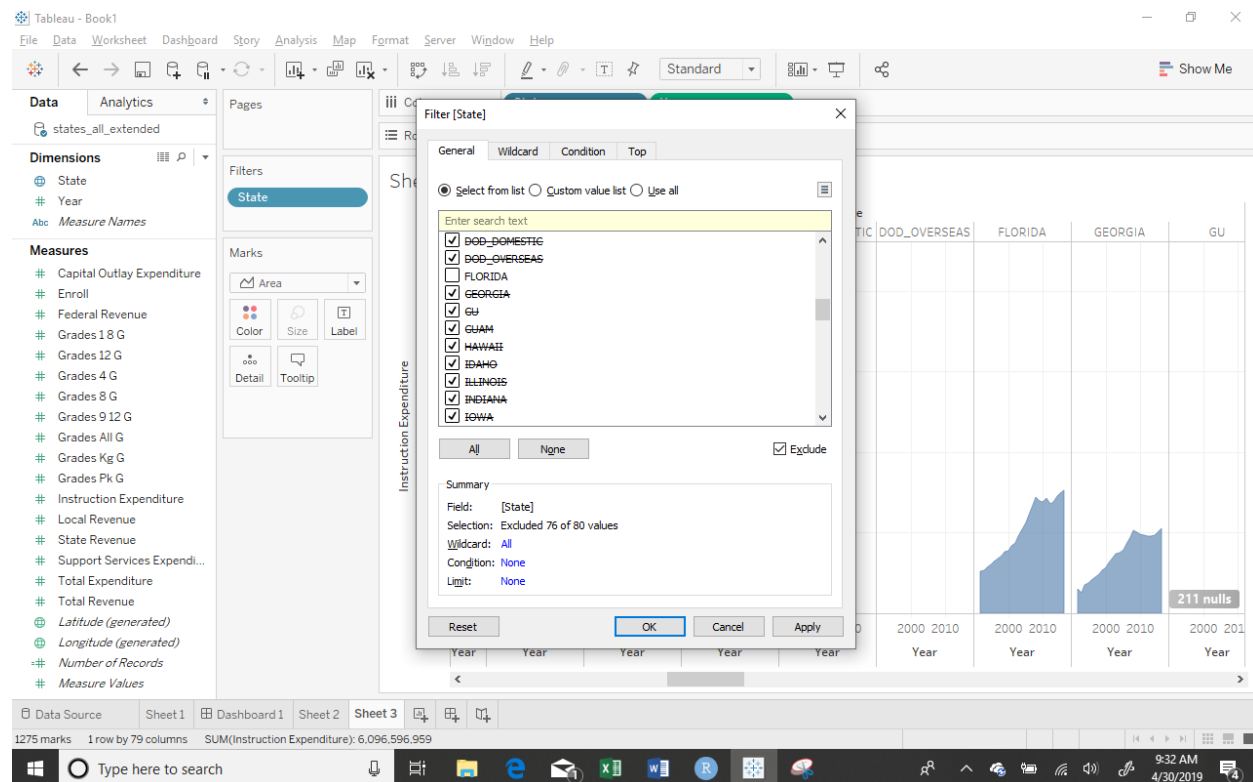


## Conclusions:

**The visualization above answers our first hypothesis and confirms that revenue allocated is directly proportional to expenses incurred hence we accept the null hypothesis.**

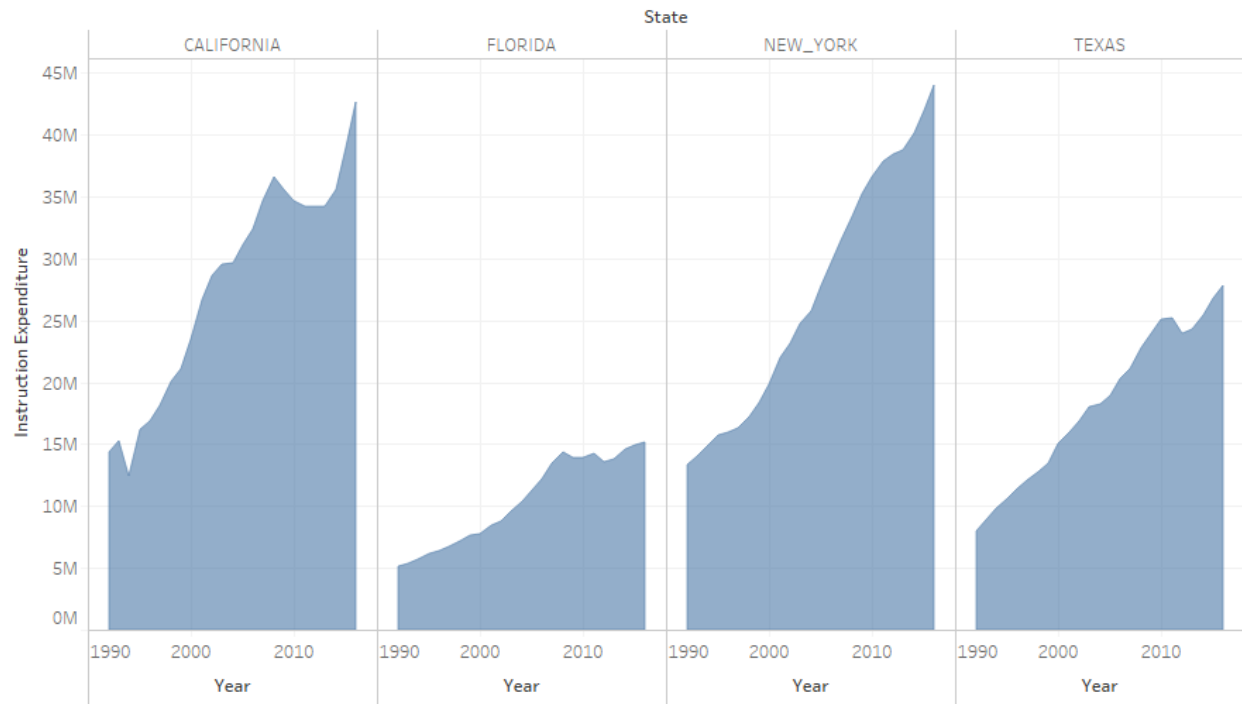
Having streamlined the top 4 states, I would be finding how instructure expenditure over the years has increased and how this has affected the grades of the students. The states to be considered are California, Texas, New York and Florida.

This is done by using the filter tool and excluding all states except the four listed above.



First, we check the how the instruction expenditure has increased over the years in these states. This is shown below;

Sheet 3



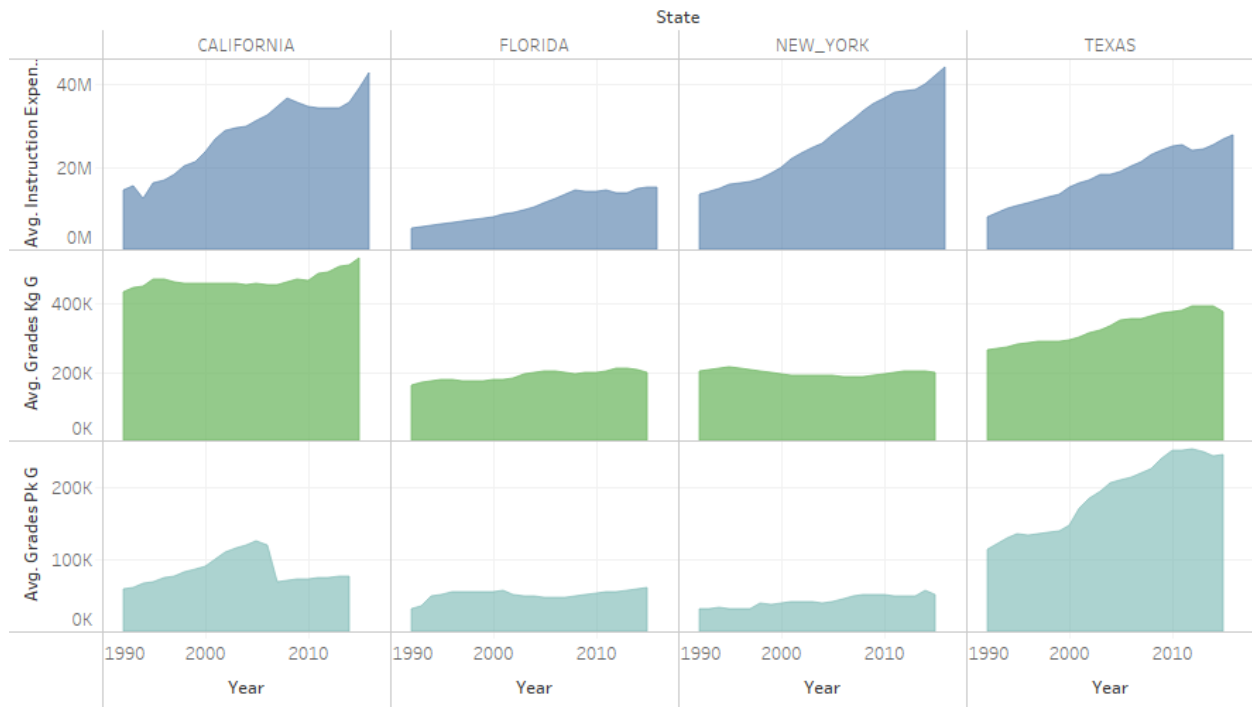
The plot of sum of Instruction Expenditure for Year broken down by State. The view is filtered on State, which keeps CALIFORNIA, FLORIDA, NEW\_YORK and TEXAS.

We see above that California and New York have experienced the highest increase in Instruction expenditure.

The spatial layout chosen is used because it represents effectively the gradual increase or decrease of funds allocated over the years.

Next, we need to find out how this has positively or negatively affected the number of enrolled students in these states.

## <Expenditure Vs Grades>



The plots of average of Instruction Expenditure, average of Grades Kg G and average of Grades Pk G for Year broken down by State. The view is filtered on State, which keeps CALIFORNIA, FLORIDA, NEW\_YORK and TEXAS.

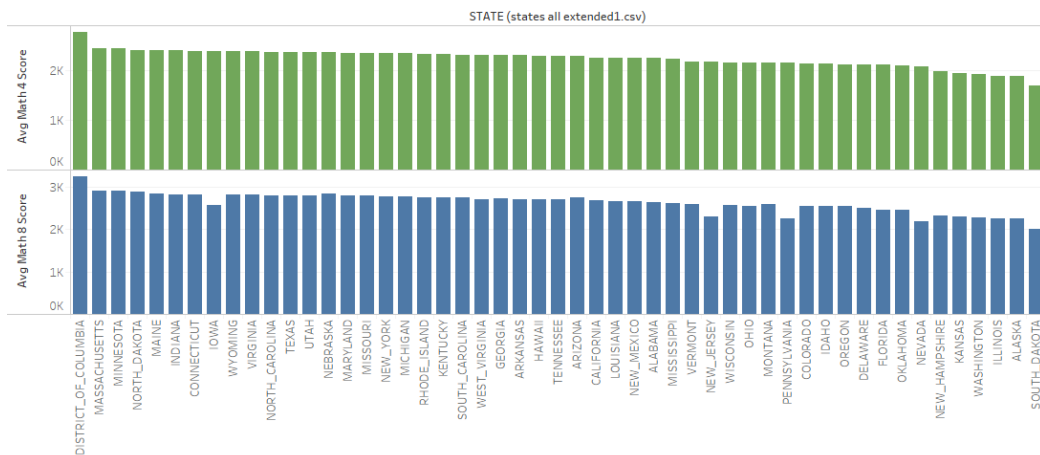
### Conclusions:

**The above visualization answers to our second hypothesis that the Increase in instruction expenditure does not guarantee an increase in enrolled students hence we accept the null hypothesis.**

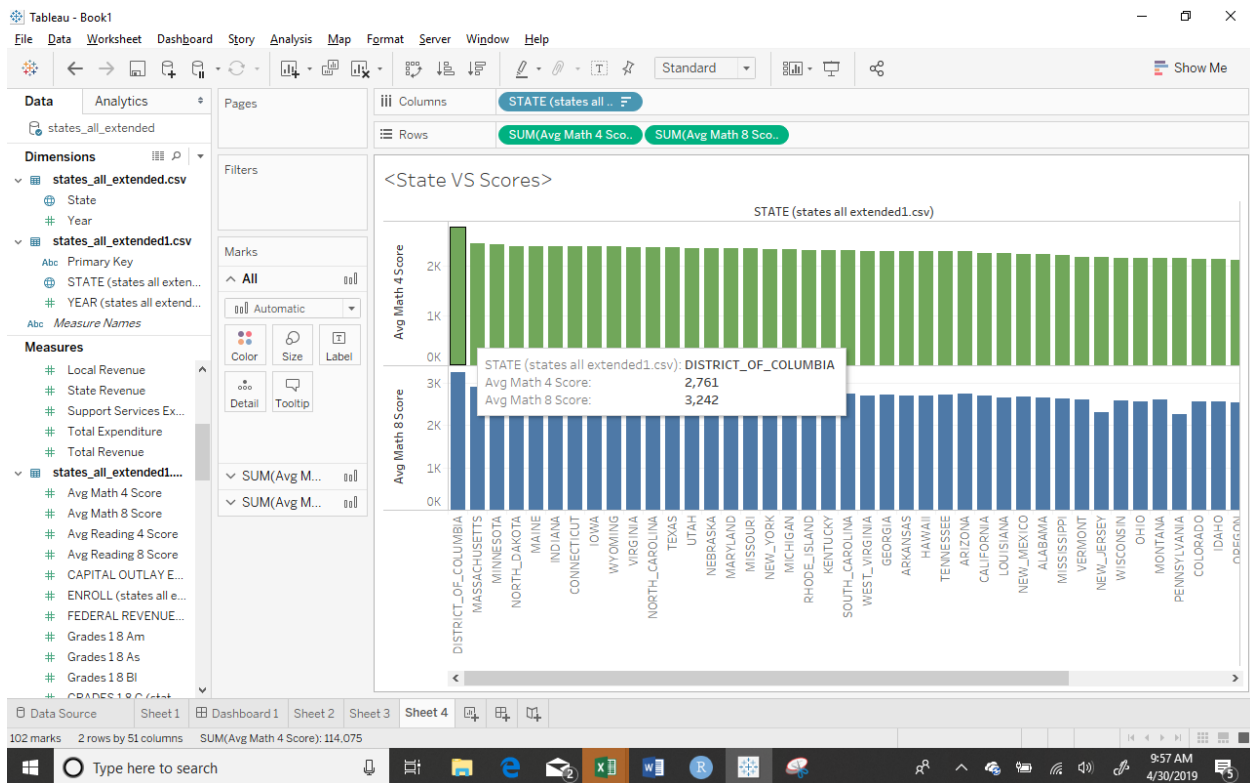
Next, we want to find which states are performing at the optimum amongst all states in our dataset. We unfortunately have very sparse data available as there are several missing values in the Average math and reading score columns. These missing values are specific to certain years.

Proceeding with the visualization for the years available, we see that the District of Columbia has the highest scores in Mathematics for Grade 4 and 8. This poses the question: Why do we not have higher performing students in California, New York and Texas given they are allocated more funds?

## <State VS Scores>

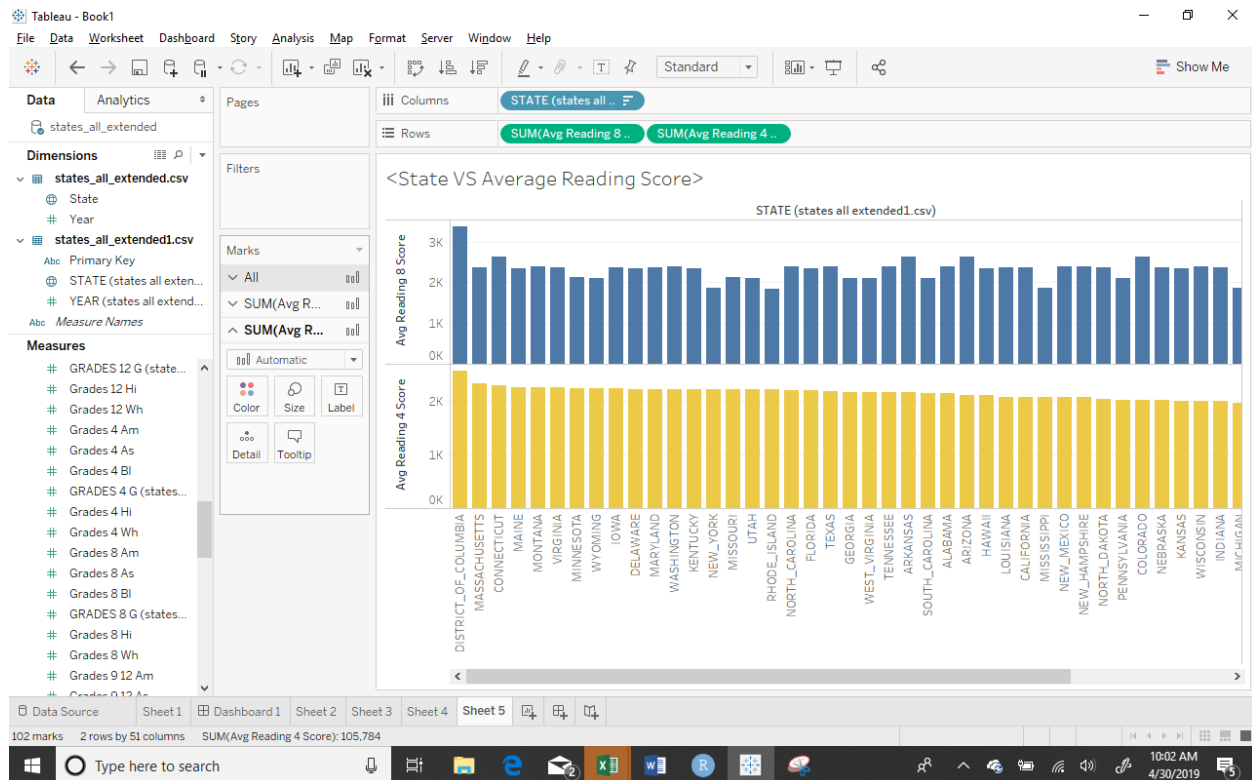


Sum of Avg Math 4 Score and sum of Avg Math 8 Score for each STATE (states all extended1.csv).



We observe a similar trend in the Average reading score with the District of Columbia topping the chart.





## CONCLUSION:

From the above, we see how population of a location contributes to the expenses incurred and income generated. We also see the different interactive techniques used.