Project: Part I

Adele Chase

School of Information Science, University of South Carolina

ISCI 709: Fundamentals of Data and Digital Communications

Dr. Ryan Rucker

May 20, 2023

Project: Part I

CFPB College Credit Card Marketing Agreements

The dataset "cfpb_college-credit-card-agreements-database-2009-2020.csv" is a database of current information about college credit card marketing agreements made between credit card issuers and higher education institutions and their affiliates from 2009 to 2021. The data were compiled by the Consumer Financial Protection Bureau (CFPB) (CFPB, n.d.-a). The purpose of the dataset is to provide transparency and accountability about the effects of the marketing agreements between credit card companies and educational organizations (CFPB, n.d.-b).In addition to information on the credit card issuer and the organization, the dataset contains detailed information on various aspects of the marketing agreements, including the status of each agreement, the amount of any payments made by the credit card issuer to the organization, and the number of new and existing accounts that were opened by consumers as a result of the marketing agreement. The dataset enables the comparison of marketing agreement terms by type of organization (educational or affiliate), and a comparison of the payment terms per user by year, credit card issuer, organization, or geographic location. It also enables the CFPB and the public to monitor changes to the effects of the marketing agreements over time.

U.S. Education Datasets: Unification Project

The dataset "states_all.csv" provides an overview of financial and student achievement information in the US. It contains information on education-related revenues, expenditures, and selected enrollment numbers and test scores, broken down by state, from 1986 to 2019. It is a compilation of data gathered from the U.S. Census Bureau, the National Center for Education Statistics (NCES), and the National Assessment of Educational Progress (NAEP) (Garrard, 2020).

The dataset can be used to compare educational spending and achievement trends by state on a per-student basis. It can be used to find correlations and perhaps even predict student achievement. However, the dataset lacks information for many states and years, a factor that must be considered when analyzing its contents. The dataset will need to be cleaned in preparation for analysis.

Shakespeare plays

The dataset "Shakespeare_data.csv" contains a line-by-line version of the full text of all of Shakespeare's plays. It also contains metadata about which character speaks the line; which play the line is from; and which act, scene, and line of the play the line is from (LiamLarsen, 2017).

The data could be used for text analysis purposes, including gaining insight into Shakespeare's vocabulary and style. It could also be used for natural language processing, sentiment analysis, topic modeling, and word frequency analysis.

References

Consumer Financial Protection Bureau. (n.d.-a). *College credit card marketing agreements and data.* https://www.consumerfinance.gov/data-research/student-banking/marketing-agreements-and-data/

Consumer Financial Protection Bureau. (n.d.-b). *Student banking and college credit card marketing agreements*. https://www.consumerfinance.gov/data-research/student-banking/

Garrard, R. (2020, April 13). *U.S. Education Datasets: Unification Project*. Kaggle. https://www.kaggle.com/datasets/noriuk/us-education-datasets-unification-project

LiamLarsen. (2017, April 27). *Shakespeare plays*. Kaggle. https://www.kaggle.com/datasets/kingburrito666/shakespeare-plays

Project: Part II

Adele Chase

School of Information Science, University of South Carolina

ISCI 709: Fundamentals of Data and Digital Communications

Dr. Ryan Rucker

June 14, 2023

Project: Part II

For my final project, I analyzed "states_all", a dataset with information on student enrollment numbers, state revenue, state education spending, and student test scores for math and reading in fourth and eighth grade on the National Assessment of Educational Progress (NAEP) exam. The data is broken down by state and runs from 1986 to 2019. This dataset was compiled from three sources: the U.S. Census Bureau, the National Center for Education Statistics (NCES), and the National Assessment of Educational Progress (NAEP). It was created to provide a convenient and comprehensive way to analyze multiple facets of U.S. education in one place (Garrard, 2020).

I chose a dataset on this topic because I am interested in educational policy and how it affects student outcomes. As a future academic librarian, I expect that the data analysis skills I have developed and will continue to develop will assist me in assessing the impact of library collections, facilities, and programming on student achievement.

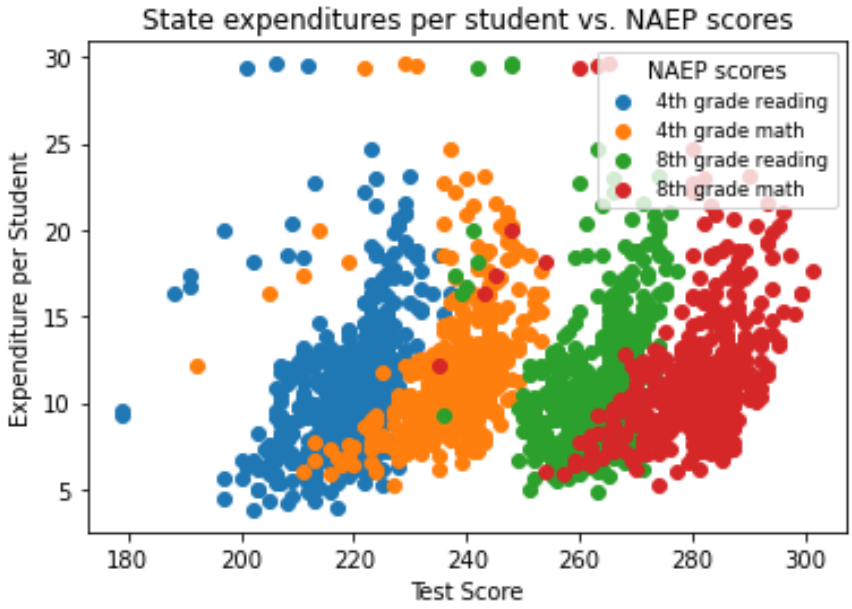
When deciding how to approach analyzing this dataset, I decided to omit several columns in order to get a better handle on the dataset. The information in the columns I omitted was included in the columns I analyzed, however. For example, there were four columns dealing with expenditures: TOTAL_EXPENDITURE, SUPPORT_SERVICES_EXPENDITURE, OTHER_EXPENDITURE, CAPITAL_OUTLAY_EXPENDITURE. I decided to work mainly with the TOTAL_EXPENDITURE column because the data in this column represents the sum of the other three columns. I did the same for the columns related to revenue. I used only the enrollment numbers from the U.S. Census and omitted the enrollment numbers from NCES.

Not all information was available for all columns of the dataset I chose. I know that this skewed my results somewhat, but I wasn't sure what the best way to handle this data would be. I considered cleaning my dataset by eliminating rows with incomplete columns, but many of these rows had all the information I needed for my analysis. To make the pie chart, I had to eliminate rows with incomplete information.

I encountered many challenges when it came to preparing this analysis. Primarily, I was not sure what types of analysis I could perform. My knowledge of math and statistics is remedial at best, so I could only really utilize basic functions, such as finding the means of certain values. I also wish I had had more time to analyze the data in more detail. When I was stuck on Python syntax, I was able to use resources I found via Googling, in particular, StackOverflow, when I was stuck and could not find the answer in *A Hands-On Introduction To Data Science* by Chirag Shah or the lectures. In terms of minor issues, I was not able to figure out how to eliminate trailing and leading spaces from data in Python; going forward, I would probably preprocess this data to get rid of those instead of attempting to use Python. I also didn't know how to resolve a recurring error that I got, which was that "A value is trying to be set on a copy of a slice from a DataFrame." This error did not prevent my code from running, however.

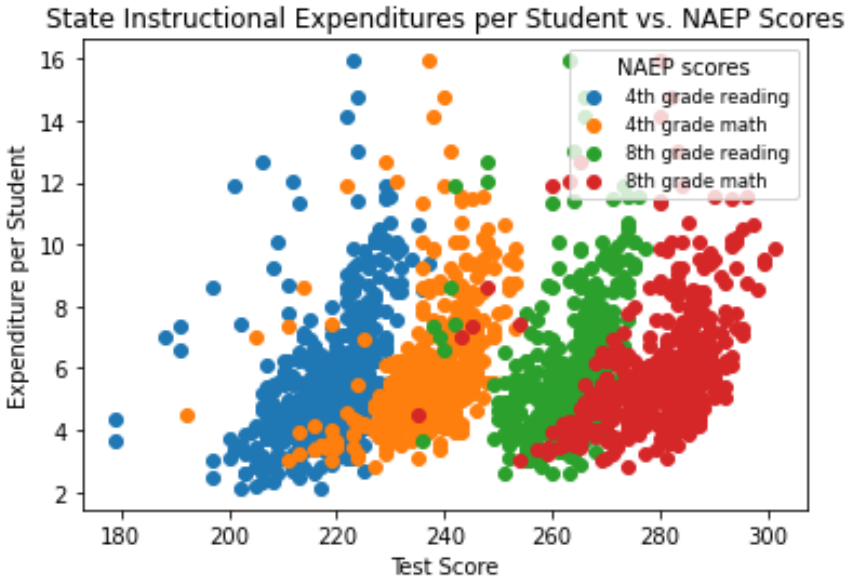
Going into the project, I expected to see that education expenditures have a measurable effect on the outcomes of student test scores. My preliminary analysis suggests that there is a weak positive association between education expenditures and NAEP test scores, but that after a certain point, additional spending has little impact on student achievement. See Figure 1, below, for a scatterplot detailing the relationship between test scores and state expenditures.

Figure 1.



Because the correlation was so weak, I decided to look at only the portion of expenses devoted to instruction costs as well. I found barely any difference between the scatterplot I created to map the relationship between test scores and instructional expenditure, shown below in Figure 2.

Figure 2.



I also expected to see that South Carolina was below average in both academic achievement and education expenditures when compared to national averages, and unfortunately, I confirmed that this is the case. Interestingly, South Carolina spends 102.98% of its revenue on education, a higher percentage than the national average, 101.14%. When I initially found this result, I thought it boded well; this is complicated by my finding that educational spending beyond a certain amount has little to no impact on student test scores, and further complicated by the fact that South Carolina's per capita student spending is below that threshold. South Carolina could therefore stand to raise taxes to increase revenue and improve student outcomes while spending a lower percentage of its revenue. Though South Carolina's deficit spending is necessary, the state still lags behind in average spending per student; the national average is $10.03, while South Carolina's average is only $8.96.

The creator of this dataset has also made another dataset in which the demographics of students are broken down by race and gender. I chose not to submit this dataset for Part I of this project because I was daunted by the amount of data and thought I wouldn't be able to get a handle on it. Now that I have worked with the smaller dataset, I wish I had submitted the larger dataset because I realize now how easy it is to analyze only certain columns and set the rest aside. I am passionate about diversity, equity, and inclusion, and how to promote policies that lead to better outcomes for everyone in the educational system. By analyzing the complete data, I could have gained insights into the relationship between race and gender as they pertain to education spending and test scores, and how different states perform in terms of equity and achievement gaps.

References

Garrard, R. (2020, April 13). *U.S. Education Datasets: Unification Project*. Kaggle. https://www.kaggle.com/datasets/noriuk/us-education-datasets-unification-project