Louis K Finney
COSC 5610
11/5/20

Data Exploration on Alternative and Public-School Districts in the U.S.

The dataset I have chosen is ProPublica's Alternative Schools in U.S. School Districts. This dataset compares several different variables between public and alternative school districts across America. There are seven broad categories of variables: alternate ratio school enrollment, per pupil state and local funding, percent of schools with counselor, average student teacher ratio, percent of teachers in their first/second year, percent teacher absent ratio, and graduation rate. These seven categories have three different measurements: the alternative district, the public district, and the percent difference between them.

The first part of the dataset I wanted to investigate was the missing data. When examining the data, it became clear that nearly every school district had one or more missing variables from the 28 possible. The dataset contains over 1900 school districts in total, but because of missing variables, my dataset could be considerably smaller than the total. The number of valid datapoints for each variable is plotted in figure 1.

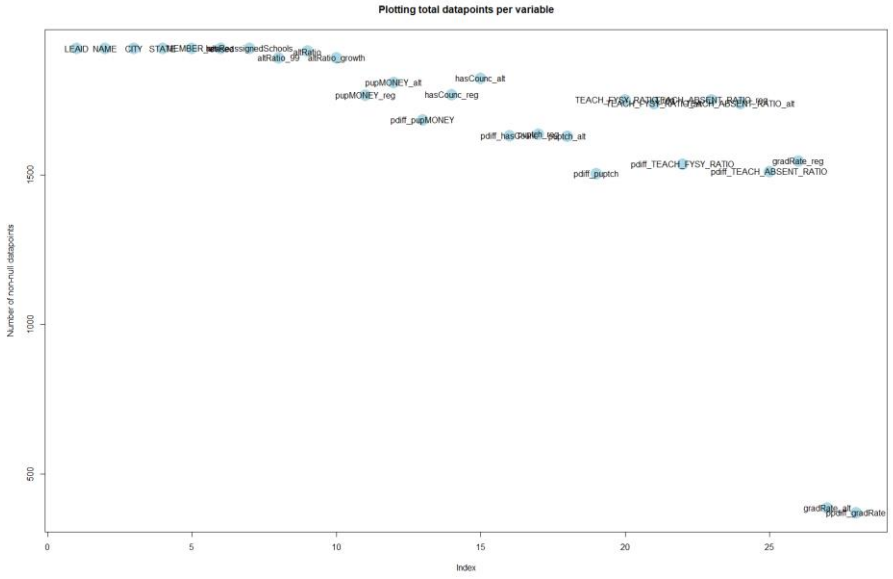

Figure 1. Plot of number of non-null values for each variable.

The biggest thing to standout in this plot is the variables at the bottom right corner, representing the variables graduation rate in alternative district and percent difference graduation rate between alternative and public district. It makes sense that since the reported graduation rate of graduation being low, there cannot be a comparison to the public-schools in the same district. Originally, I was interested in how most of the other variables affected graduation rate. That could still be done, but it is unlikely I would have a representative enough size for that to work.

The next question I asked was how evenly distributed are the number of school districts across the U.S? Using the tool plot_usmap, demonstrated in [Di Lorenzo], I was able to visually represent how many school districts had information collected about them by state and displayed in figure 2.

Louis K Finney
COSC 5610
11/5/20

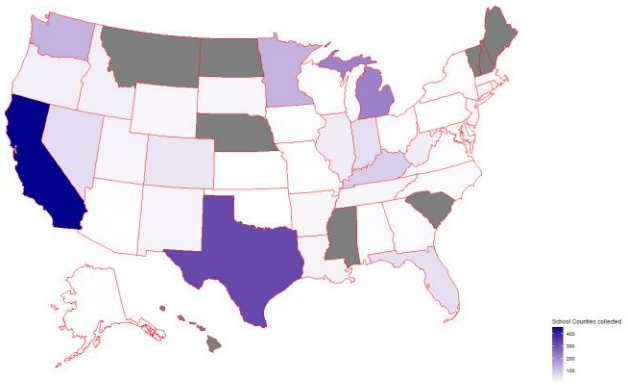

Figure 2. Map of number of school district data points in each state.

There are several takeaways from this graph. First is that no data is collected from eight states: HI, ME, MS, MT, ND, NE, NH, SC, and VT. DC was included in the dataset, so there are a total of 43 state labels. The next takeaway is that there is a huge disparity between CA, WA, TX, MN, and MI compared to the rest of the states. Most states have less than a dozen school counties with data. Based on this, it would be difficult to use state as a predictor in any model, due to the imbalance and small sample size. I could attempt to use SMOTE or ROSE but need to see if this dataset would be applicable.

My next step was to try to find correlations between variables. Most of the variables were inputs to the school – funding, teacher ratio, counselor. The graduation rate was the only "outcome" variable given in the data. Average test scores or percent attending college would be examples of other outcomes many of these variables could be tested for. I began to test what variables affected the outcome. There was not a strong relationship. Figure 3 shows the plot of per pupil funding in alternative schools and graduation rate. This led me to pursue other options besides graduation rate.

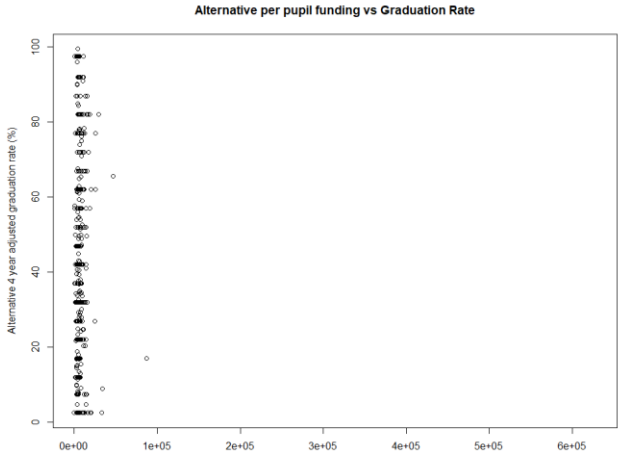

Figure 3. Alternative per pupil funding vs graduation rate

Louis K Finney
COSC 5610
11/5/20

The next approach I took was to examine the variables more closely to see what the differences were between public and alternative. I focused on comparing the public to the alternative school district datapoints. This means that I was not using many of the percent differences between districts. I begin to attempt to visualize these differences using box and whisker plots. This proved to be ineffective, as shown in figure 4.

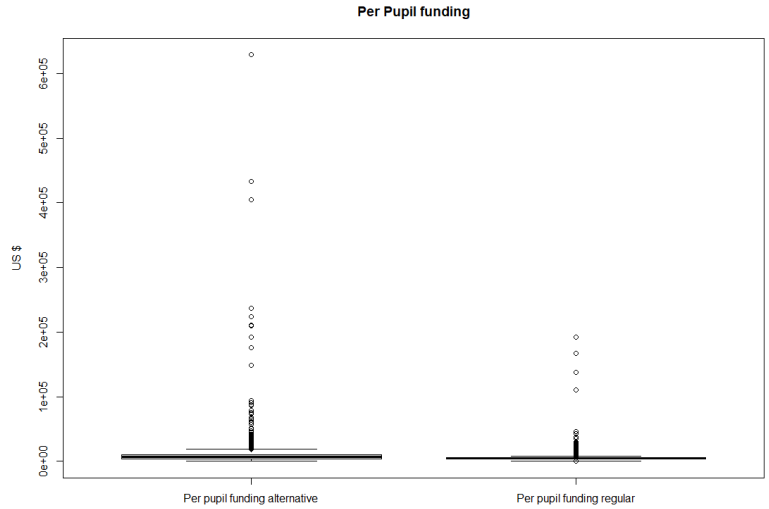

Figure 4. Boxplot of per pupil funding based on type of school district.

There is almost no insight that can be gained from this plot. Other forms of visualization proved ineffective as well because of the distribution similarity and long tails of data. Instead, summaries of the data proved to be much more effective at discerning slight differences. Table 1 provides the summaries of all data points used in

	Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Number of NA's
Per Pupil State and local Funding	Public	0	3754	4414	5717	5646	191881	157
	Alternative	0	4172	6620	10355	10273	629001	116
Percent of schools within district with counselor	Public	0	87.50	100	87.76	100	100	156
	Alternative	0	0	50	47.57	100	100	101
Average Student	Public	1.507	14.965	17.654	19.910	21.366	2691.462	288
	Alternative	0.667	8.295	13.396	22.019	19.534	4400	295

Teacher Ratio								
Percent of teachers in their 1st /2nd year teaching	Public	0	5.664	9.919	11.399	15.177	99.625	174
	Alternative	0	0	0	10.57	15.79	100	186
Percent of Teachers absent more than 10 school days	Public	0	12.9	23.95	25.70	34.44	98.87	1174
	Alternative	0	0	15.38	22.22	36.74	100	186
Four year adjusted cohort Graduation Rate (%)	Public	2.50	86.0	92.0	89.23	95.0	99.5	378
	Alternative	2.50	22.01	42.0	45.39	67.0	99.5	1537

Table 1. Comparing Public and Alternative school statistics

Table 1 provides much more substance and information than any type of visualization I attempted performs. We can see that public schools generally have higher percentages of teachers absent, graduation rate*, and schools with counselor. Alternative schools have higher per pupil funding. After seeing this, I had a realization that trying to predict one variable with another would not be the best option to pursue. The preferred method would be to try and predict whether a district is public or alternative based on the values in table 1. The data would need to be reshaped by separating the alternative and public columns in each row, then adding a label of “public” or “Alternative” to each of them. This fits what we have discussed in class of creating a binary linear/logistic regression.

Louis K Finney
COSC 5610
11/5/20

References

Anna J. Egalite and Patrick J. Wolf. 2016. A Review of the Empirical Research on Private School Choice. *Peabody Journal of Education* 91, 4 (2016), 441–454. DOI:http://dx.doi.org/10.1080/0161956x.2016.1207436

Christopher Lubienski and Sarah Theule Lubienski. 2006. Charter, Private, Public Schools and Academic Achievement: New Evidence from NAEP Mathematics Data. *National Center for the Study of Privatization in Education* (January 2006). DOI:http://dx.doi.org/https://nepc.colorado.edu/sites/default/files/EPRU-0601-137-OWI%5B1%5D.pdf

Haifeng (Charlie) Zhang and David J. Cowen. 2009. Mapping Academic Achievement and Public School Choice Under the No Child Left Behind Legislation. *Southeastern Geographer* 49, 1 (2009), 24–40. DOI:http://dx.doi.org/10.1353/sgo.0.0036

Hannah Fresques, Heather Vogell, and Olga Pierce. 2017. Alternative Schools in U.S. School Districts. (March 2017).

Heather Vogell Hannah Fresques. 2017. Methodology: How We Analyzed Alternative Schools Data. (2017). Retrieved October 7, 2020 from https://www.propublica.org/article/alternative-schools-methodology

Manyee Wong, Thomas D. Cook, and Peter M. Steiner. 2011. No Child Left Behind: An Interim Evaluation of Its Effects on Learning Using Two Interrupted Time Series Each With Its Own Non-Equivalent Comparison Series. *Institute for Policy Research Northwestern University* (September 2011). DOI:http://dx.doi.org/https://www.ipr.northwestern.edu/documents/working-papers/2009/IPR-WP-09-11.pdf

Matthew M. Chingos, Daniel Kuehn, Tomas Monarrez, Patrick J. Wolf, John F. Witte, and Brian Kisida. 2019. The Effects of Means-Tested Private School Choice Programs on College Enrollment and Graduation. Research Report. *Urban Institute* (July 2019). DOI:http://dx.doi.org/https://eric.ed.gov/?id=ED601791

Paolo Di Lorenzo. 2020.(October 2020). Retrieved November 6, 2020 from https://cran.r-project.org/web/packages/usmap/vignettes/mapping.html

Zach. 2020. How to Plot Multiple Boxplots in One Chart in R. (April 2020). Retrieved November 6, 2020 from https://www.statology.org/multiple-boxplots-r/