Educational Performance in the Connected Age

Derek He
Comm 318 Final Project

December 23, 2020

1 Abstract

The current pandemic has necessitated internet connectivity. No where is this more clear than in education. This article seeks to zoom in and look at data involving internet connectivity and educational performance as well as other factors and how these may affect each other and report and contextualize those findings. A fairly significant positive relationship is found between educational performance and internet connectivity but it is difficult to separate from various other confounding factors.

2 Introduction

The recent COVID-19 pandemic has exposed an inherent need for connectivity. Whether it's for work or sociability, internet access has become a pillar of society. Nowhere is this more apparent than in education. Education, contextualized in the year 2020, is vital to breaking down structures of inequality, both class and economic. It can thus be hypothesized that inability to access broadband internet can be detrimental towards educational performance in this day and age. In this article I sought to use take a look at some data as well as perform some statistical analysis to see if this hypothesis can be supported based on historical data.

3 Data

3.1 Data Sources

All data was sourced from the United States Department of Agriculture Economic Research Service County-level Data sets [1]. The exception is broadband connectivity data which was sourced from Arizona State University's center on Technology, Data, and Society's County Data for Broadband data [2]. All of these contain statistics by Federal Information Processing Standard or FIPS codes.

3.2 Data Cleaning and Merging

Each of these datasets was cleaned and then merged to a dataframe as a whole. This section provides context on what data was taken.

3.2.1 Education

The county level datasets for education have data on educational attainment for the years 2014-2018. While this is broad, it separates educational attainment into 4 categories: completed less than highschool, graduated highschool, completed some college, graduated college. For each FIPS code, each of these as a percentage were taken. See section 5 for how we turn this into classification and scoring.

3.2.2 Internet Connectivity

The broadband connectivity datasets have the percentage of households connected for a fips code from the years 2000-2018. Since the years 2014-2018 are being worked with, the mean of those years by fips code were taken and returned as the internet connectivity statistic.

3.2.3 Poverty

The confounding factors of poverty were looked at. The poverty dataset gives us for each FIPS code the 2018 amount of people and percentage of people living in poverty as well as upper and lower bounds of a 90 percent confidence interval. For this I chose

3.2.4 Unemployment

The confounding factors of unemployment rate and median household income were also looked at. From this dataset we took the unemployment rate as well as the median household income by FIPS code.

3.2.5 Urbanization Classification

The confounding factors of urbanization class were then also looked at. This is a number 1 to 6 where 1 defines the densest of metropolitan areas and 6 defines a rural area. See section 5 for distributions as well as turning this into a one-hot encoding. This is given by FIPS code as well.

4 Samples of Interest

This dataset has 10 features with 3127 rows currently so some sampling to zoom in by FIPS codes to see interesting things may help.

4.1 Random

First random sampling was taken to see if anything interesting could be found. With a fairly large dataset these seem to return fairly average and expected values for everything.

4.2 Broadband

As we'll see later in Section 6 and this following graphic, there is a large range of connectivity in the united states. Thus, samples from the high and low end of connectivity to see were pulled to see what information could be garnered

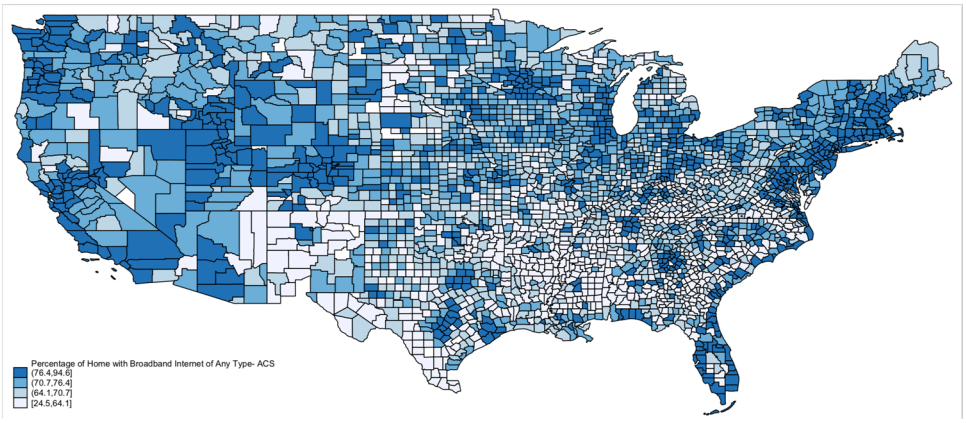

Figure 1: 2017 percentage of homes with broadband internet. From Arizona State University

4.2.1 Minimum

On the low end of broadband connectivity we see places such as Harding County, New Mexico; Saguache County, Colorado; and Kennedy County, Texas. These tended to be more rural areas with higher poverty rates and

lower median-house income. These tended to have higher percentages of people doing less than high school graduation and only high school graduation.

4.2.2 Maximum

On the high end of broadband connectivity we see places such as Hamilton County, Indiana; Loudoun County, Virginia; and Douglas County, Colorado. These had lower rates of poverty, and tended to be more metropolitan while not fully the highest class of population density. These also had lower unemployment rates and higher median house income with the highest percentage of educational attainment being college graduation.

4.3 Education

For education we sampled the counties that were on the higher end of each percentage. Regression analysis will be needed to actually see if there is a significant relationship but from these samples we tended to see decreases of unemployment, increases of broadband connectivity, increases of income, increases of urbanization, and decreases of poverty.

5 Variance and Grouping

5.1 Min, Max, Range, Mean, Variance

For many statistics the minimum, maximum, range, mean, and variance of the data was pulled but there was not that much to learn from these statistics. Instead, looking at distributions is more key to our understanding.

5.2 Distributions

5.2.1 Education

To help with problem formulation it's key to rethink how educational attainment is defined as one label instead of 4.

5.2.2 Max Education Percent

One way to think of turning education into one label is to classify in terms of which educational attainment was reached the most: 1, 2, 3, or 4 based upon

whether the most common was some high school, high school graduation, some college, or college graduation. The result was a bar distribution as such.

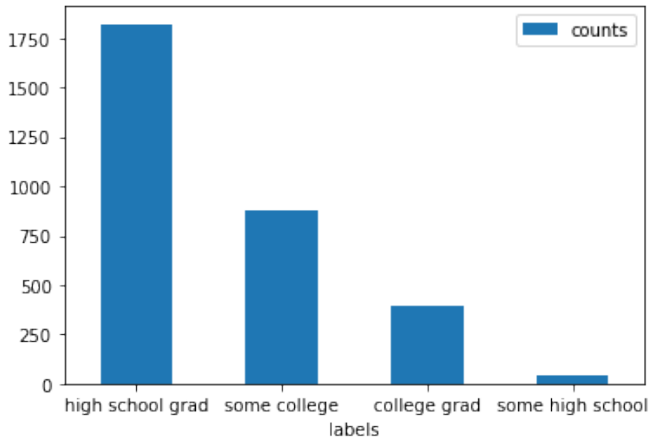

Figure 2: Distribution of most common educational attainment

This is not particularly indicative of a good label as it's highly unbalanced but may prove useful for classification visualization.

5.2.3 Education Score

The other way to create one label for education is to think in terms of an education score, with higher levels of education contributing more. Essentially if the percentage of some high school, high school graduation, some college, and college graduation are considered W, X, Y, and Z, we can create a sore formula as such.

$$score = 1 * W + 2 * X + 3 * Y + 4 * Z$$

As a rather simple method of creating a score label, the result is the following distribution.

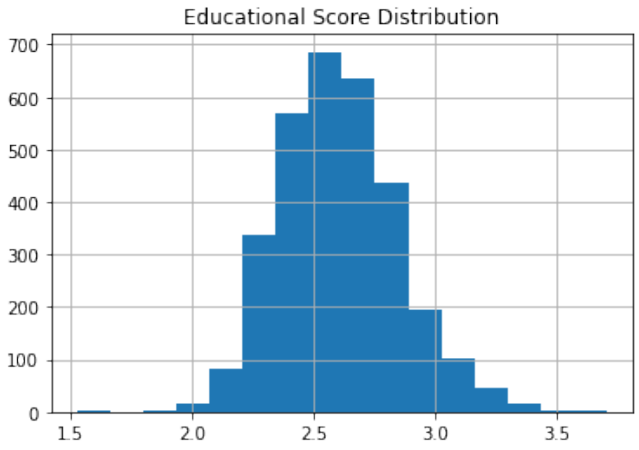

Figure 3: Distribution of educational attainment score

This distribution is fairly normal and seems usable for the sake of correlation and regression.

5.2.4 Broadband

We also get to see the distribution of broadband connectivity.

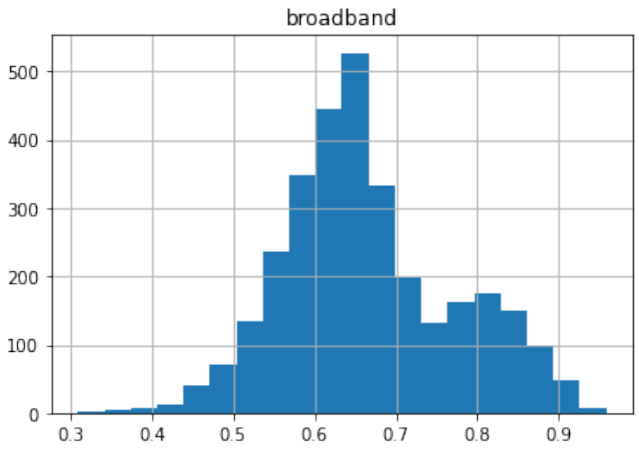

Figure 4: Distribution of broadband connectivity percentage

This also looks to be a fairly normal distribution albeit with some local maxima at higher percentages.

5.2.5 Urbanization

We also take a look at the distribution of urbanization classification.

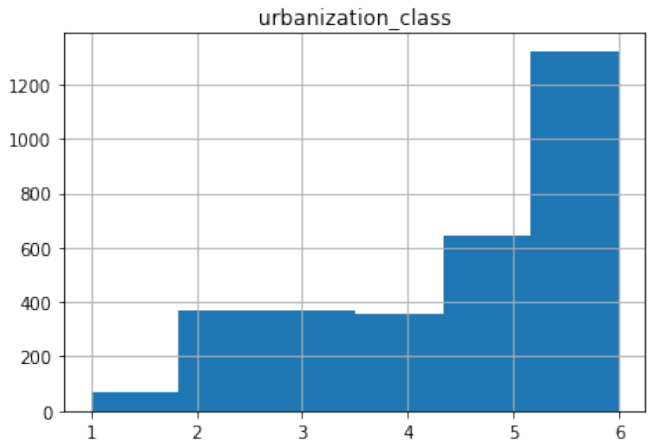

Figure 5: Distribution of urbanization classification

As is the distribution of America, there is heavy skewing towards more rurally classified FIPS codes. This may indicate that the usage of one hot vectors as a feature for this classification better suits it.

5.2.6 Poverty

We also look at the distribution of poverty since we take two poverty classifications to figure out which we should use.

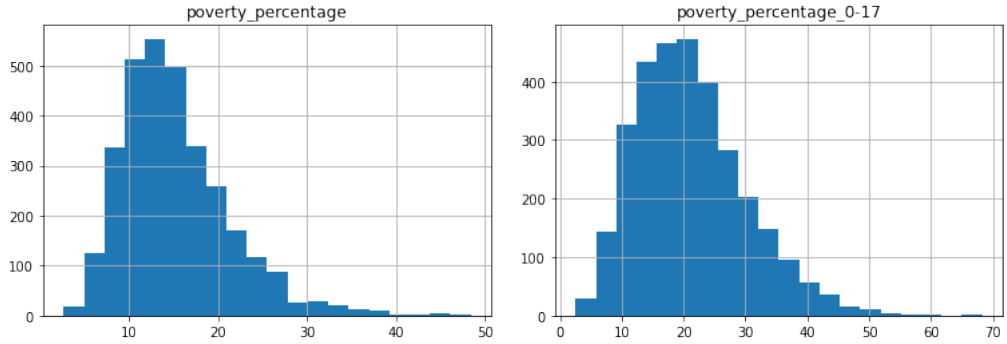

Figure 6: Distribution of poverty rate for both all population and ages 0-17

From this we see these are both fairly left skewed normal distributions but with the wider range of poverty rate from 0-17 as well as the fact that we're dealing with education poverty rate for 0-17 year olds seemed like a better choice

5.3 Correlation

We then look at a correlation matrix to look for significant correlations with education score as well as with broadband connectivity. From the matrix we got a .648 correlation value of educational score with broadband which is significant enough to look further into and indicates a positive relationship between these variables. However, correlations also exist with median house income and urbanization class.

5.3.1 Broadband and Median House Income

We can then look at a scatterplot of broadband connectivity against median house income. The figure 7 is below with median house income as the y axis and connection as the x axis.

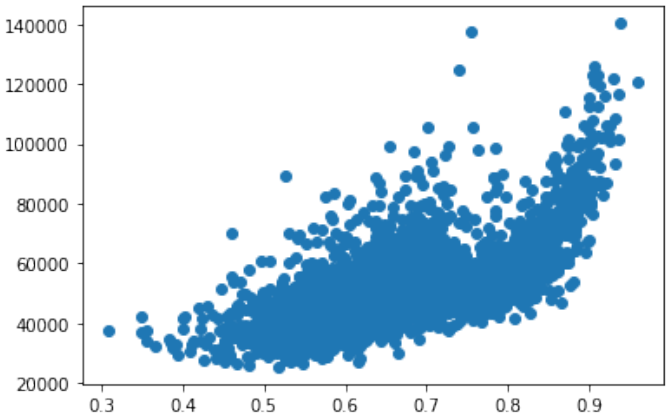

Figure 7: Median House income Versus Broadband connectivity.

This seems to show a somewhat positive relationship which can be contextualized with the education classes created off of most common educational attainment.

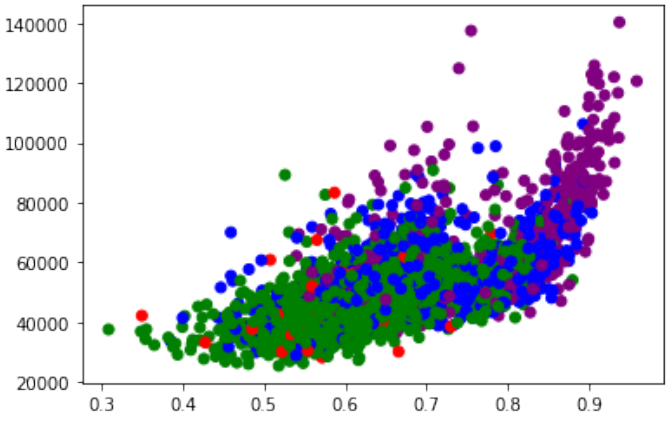

Figure 8: Median House income Versus Broadband connectivity. Education Context

From lower to higher levels of educational attainment the colors go red, then green, then blue, then purple. This seems to show that the increase of educational attainment tends to follow that positive trend of both median household income and connectivity.

5.3.2 Broadband and Education Score

We can also look at a scatterplot of connectivity percentage versus education score.

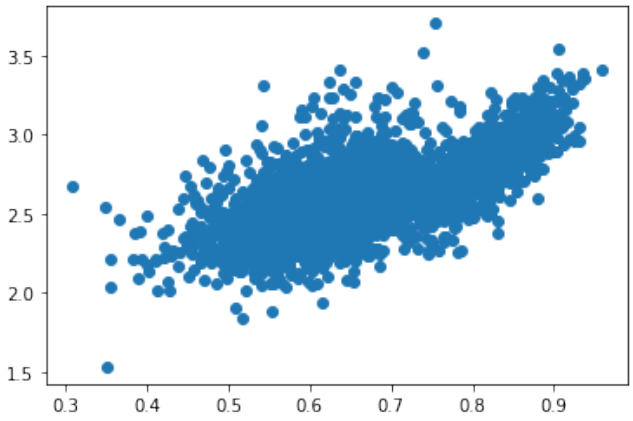

Figure 9: Education score Versus Broadband connectivity

There is a fair bit of variance but there is a positive relationship between the education score and connectivity percentage. We can also contextualize this with the urban classifications.

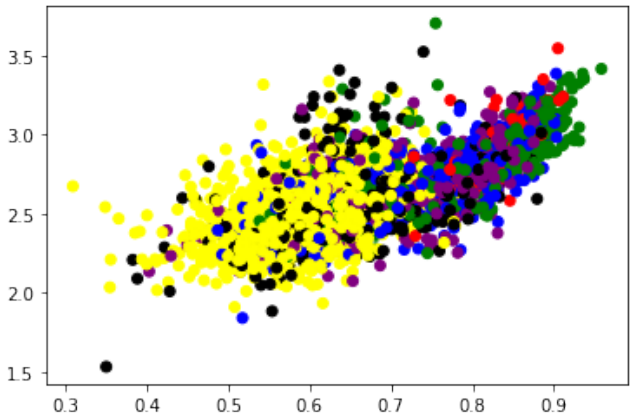

Figure 10: Education score Versus Broadband connectivity. Urbanization context

From rural to urban classifications the colors go yellow, black, purple, blue, green, then red. This groupings do seem to indicate that the increase in urbanization does tend to correlate with the relation between the increase in education score and connectivity percentage.

5.4 Principle Component Analysis

We can also use PCA to perform a dimensionality reduction to two dimensions. First we cleaned the dataset to consist of poverty for only 0-17 year olds, education score instead of percentages, and one-hot encodings of urbanization type.

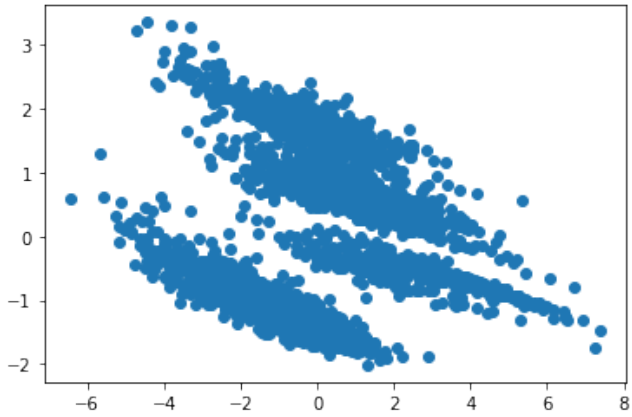

Figure 11: Dimensionality Reduction to 2 dimensions

As you can see from this there's variance in two main directions. However, the explained variance ratios are only 0.34407379 and 0.12928377. Meaning, there is an indication that the two directions of variance don't account for a lot of the variance of the data.

We can actually see below this dimensionality reduction contextualized to various classifications. For instance, educational attainment classes.

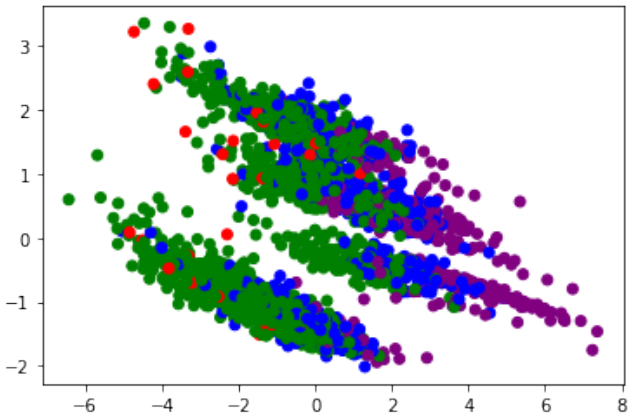

Figure 12: Dimensionality Reduction with education classes

As we see. A lot of the variance in one direction is explained by education which correlates pretty heavily with broadband connectivity. We see though that we can't consider it a simple relation as below we contextualize with urbanization classes.

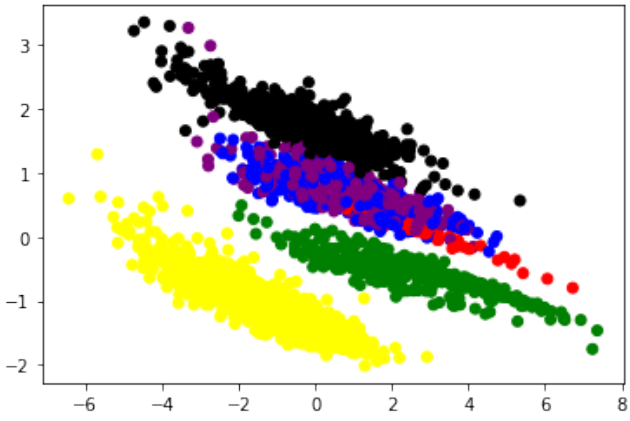

Figure 13: Dimensionality Reduction with Urbanization classes

From this we see a lot of the variance inherent to which urbanization class an observation belongs to.

6 Regression Significance

Finally we performed an OLS regression of broadband connectivity against education score. The regression returned coefficients of 1.495 for broadband connectivity percentage and a constant of 1.609. The p-score given indicated that this is a statistically significant relationship in which the null hypothesis that these are not related can be rejected. However, when OLS regression is also opened to other features such as median house income and urbanization class the relationships with other features are also found to be statistically significant.

7 Conclusions

The question we sought to answer was is there a significant correlation between internet connectivity and educational performance. From our sampling, correlation, and regression we can indeed say that we found a correlation that could be considered statistically significant. However, a lot of other features and statistics could be considered confounding as shown with our distribution analysis and principle component analysis. This may not have answered all the questions but the existence of a relationship here does indicate that there is possible value towards helping internet connectivity across the united states.

References

[1] County-level Data Sets. (n.d.). Retrieved December 23, 2020, from https://www.ers.usda.gov/data-products/county-level-data-sets/

[2] County Data. (2020, April 23). Retrieved December 23, 2020, from https://techdatasociety.asu.edu/broadband-data-portal/dataaccess/countydata