Diplomas and Disadvantage: Mapping U.S. Colleges on County Disadvantage Metrics

Annamarie Warnke

December 16, 2024

1 *Diplomas and Disadvantage* Motivation

Colleges are unequally distributed across the United States; there is a high concentration of colleges along the East half of the country and the West Coast, but sparse options in the Midwest. There is even variation within regions and states. Though more students than ever before are traveling across the country to attend university thanks to air travel, over two-thirds of college students stay within 50 miles of home [1]. Therefore, having a college in close proximity can be a deciding factor for whether a student pursues higher education, and lack of nearby colleges can compound on other community disadvantages. Researchers from the University of Michigan and Princeton University developed an Index of Deep Disadvantage to analyze holistic disadvantage by combining life expectancy, low infant birth weight rate, rates of poverty and deep poverty, and social mobility [2]. To inspire questions about community deep disadvantage and how these disadvantages affect local colleges, the *Diplomas and Disadvantage* project maps the distribution of colleges and universities across the United States in relation to metrics that signal deep disadvantage.

The idea for *Diplomas and Disadvantage* came from my internship this summer at MDRC, a nonprofit doing education and social policy research. Their Postsecondary Education policy area had a Lunch and Learn event earlier this year about community disadvantages, and they were interested in mapping their portfolio of partners on the Index of Deep Disadvantage to determine which disadvantaged communities are underserved by their research. *Diplomas and Disadvantage* is distinct from my summer work because it uses no confidential MDRC data and no code originally written over the summer. In particular, the use of Python and the inclusion of a Dash app are new additions for *Diplomas and Disadvantage*.

2 Data Description

The U.S. colleges data comes from the Integrated Postsecondary Education Data System (IPEDS), specifically the Custom Data Files page. It contains data for every postsecondary institution in the United States. I pulled data on institution names, state, longitude, latitude, highest degree level offered, public status, Historically Black College or University (HBCU) status, tribal college status, and number of students receiving each type of degree. When I generated the data file in October 2024, the most recent data available was from 2023, and the degree conferment data encompassed the 2023-2024 school year.

The data obtained from IPEDS is stored in a comma-separated values (CSV) format. It contains 5,994 postsecondary institutions. Nearly half of these schools offer at least a four-year degree, and about one-third are public. 101 colleges are classified as HBCUs, and 35 colleges are classified as tribal colleges. There is no universal definition of community colleges, but generally they are public colleges that mostly offer certificates and associate's degrees. To determine the cutoff ratio of certificates and associate's degrees, I calculated the percentage of students who received these degrees at each public school in 2022-2023 and ranked the schools according to this ratio. By looking through this ranking, I determined that the colleges switched from primarily community colleges to primarily satellite campuses around the 90% mark. Therefore, I defined community colleges as public institutions where at least 90% of graduating students receive a certificate or associate's degree. Using this definition, there are 1,227 community colleges in the dataset.

The deep disadvantage data comes from the University of Michigan's Understanding Communities of Deep Disadvantage page (see [2] for link), and it is downloadable as an Excel workbook. There are 45 variables, but I decided to include 15 of these variables. The default variable that I chose to focus on is the ranking of the Index of Deep Disadvantage. I also included the Index of Deep Disadvantage, which combines health measures, poverty rates, and social mobility data to holistically measure community well-being and lack thereof, but the process of compiling the index includes principal component analysis, which preserves variation but reduces the interpretability of the index. Therefore, the rank is more informative and interpretable. Five of the included variables are components of the index: life expectancy, low infant birth weight rate, poverty rate, deep poverty rate, and social mobility. I included variables for white, Black, and Native resident percentages because I knew these variables would intersect well with the college subsets I had chosen. I included variables for percentages of residents with less than a high school diploma and with a college degree to better understand another factor other than college location that leads students to pursue postsecondary education: whether the people in their family and social circles attended college. Economic opportunities are also a major factor in this decision, so I included unemployment rate and Gini coefficient to capture a view of these opportunities. Finally, I included a variable for the number of climate disasters that occurred from 1989 to 2017 because I am curious about how climate change will continue to affect American communities and their youth's educational journeys.

All of these variables are available for 3,141 counties (the U.S. has between 3,007 and 3,244 counties, depending on how you define a county [3]) and the 500 largest cities in the U.S. I decided to only include counties because they provide more uniformity for mapping; cities are hard to include on maps because in they usually overlap counties, and city boundaries and definitions can vary depending on the state.

3 Methods

app.py is the only file that the user will interact with if they wish to use the Dash app. Upon running the file, *app.py* first creates the Dash app instance and sets the title and favicon (browser icon) for the page. Then, it calls the *collect_and_clean* function with the deep disadvantage data and IPEDS data. The *collect_and_clean* function is located in *src/create_map.py* and loads in these two files as pandas dataframes along with a county geometries file from the Internet. After some minor cleaning of both dataframes, the function deals primarily with the IPEDS data. The only subset of colleges that is not directly present in the IPEDS data is community colleges, so the function adds the different degree level columns together and calculated the ratio to compare to the cutoff described in the Data Description section. Finally, this pandas dataframe is transformed into a geopandas dataframe, and each school point is matched with its respective county, allowing for the IPEDS data to be merged with the deep disadvantage data before all three dataframes (county geometries, deep disadvantage, and IPEDS) are returned to *app.py*.

app.py then uses the *layout* function to add all of the elements of the page, including a header, a dropdown menu for the disadvantage metrics, a description of the chosen metric, radio buttons for the college subsets, a description of the chosen subset (if applicable), a map, check buttons for the tooltips, and a link to the original Index of Deep Disadvantage page. Some limited CSS styling is included in the HTML elements; CSS also controls the text and spacing through the files in the *assets* folder.

Once the layout is added to the app, the first callback creates the map. The callback passes the selected metric, subset, and combination of tooltips to the *create_map* function, along with the three dataframes from *collect_and_clean*. The *create_map* function can be found in *src/create_map.py*. First, it filters the schools dataframe using the *subset_schools* function, found in *src/util.py*. Then, it sets up the schools and disadvantage dataframes with the chosen metric using the *county_settings* function, also found in *src/util.py*. This function creates a "metric_of_interest" column in both dataframes that is a copy of the column containing the appropriate variable, and also returns the appropriate colorscale for that particular metric. Most of the metrics use the *deep_r* colorscale, which shows that a lower value indicates more disadvantage because lower values lean towards a darker blue, while higher values lean towards a light green. Some metrics are reversed, so they use the *deep* colorscale with higher values being darker blue. Finally, the three racial demographic

metrics use the *ice_r* colorscale, which has a sequential palette and therefore tried to portray a different connotation the *deep* and *deep_r* colorscales.

The *create_map* function continues with the resulting dataframes and colorscale by setting the tooltip text and options along with the default map formatting. As part of this map formatting, *create_map* calls the *create_title* function in *src/util.py* with the chosen metric and subset. This function combines these two elements into a descriptive title for the plot, and the title is returned to *create_map*. Two map types are layered to create the visualization: first the choropleth map with the disadvantage data, then the scatterplot map with the school data. Finally, the function returns the figure back to *app.py*.

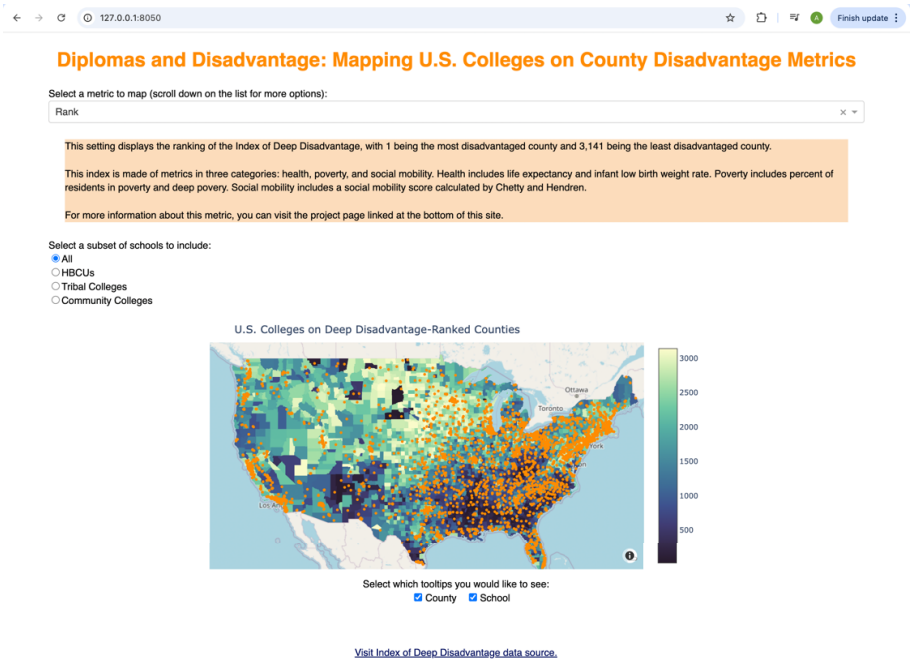
There are two remaining callbacks in *app.py*, and both call functions in *src/util.py*. The first calls the *metric_inner_html* function, which returns a pre-written description for the user about their chosen metric. The second calls the *subset_inner_html* function, which returns a pre-written description for the user about their chosen subset. Both return strings, which are passed through the *DangerouslySetInnerHTML* function (part of *dash_dangerously_set_inner_html*) so that line breaks are followed.

If the user wants just the plotly HTML file of the plot instead of the Dash app, they can run *src/create_map.py* directly. When they do, the two functions in the file are run to create a basic plot with the index of deep disadvantage rankings and all colleges. They can also create their own python file and import the functions in *src/create_map.py* to be able to load their own data files and choose a different metric and subset for their plotly file.

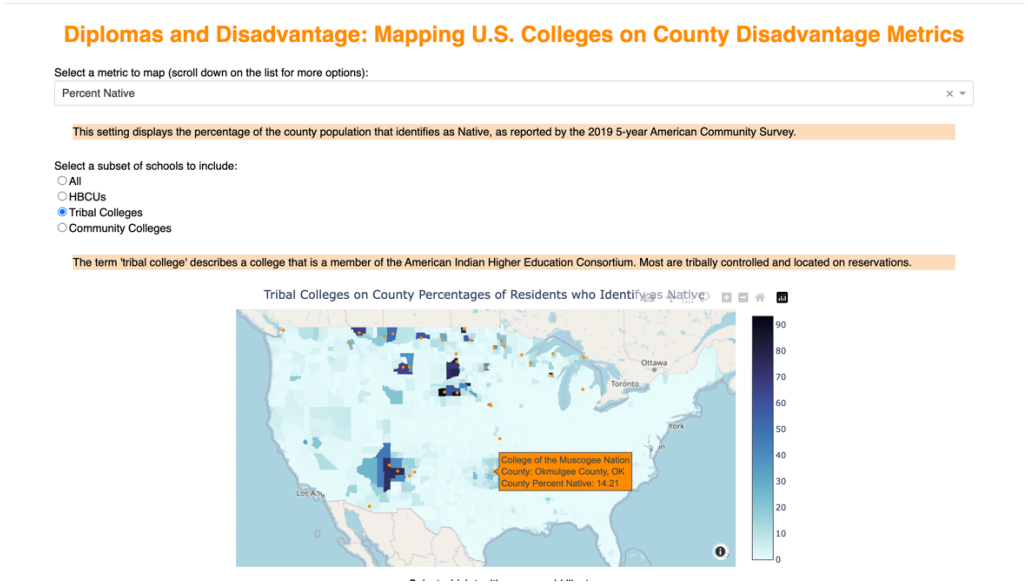
4 Output

When *app.py* is run, it starts the Dash app. This app includes a metric dropdown menu, a set of radio buttons to control the subset of colleges, and two check buttons to toggle the available tooltips on the map. The map is automatically updated every time that any of these settings are changed. Metric and subset descriptions also appear to explain what the user has selected. The bottom of the site has a link to the deep disadvantage data if the user wants to explore the relationships they find on this page more deeply.

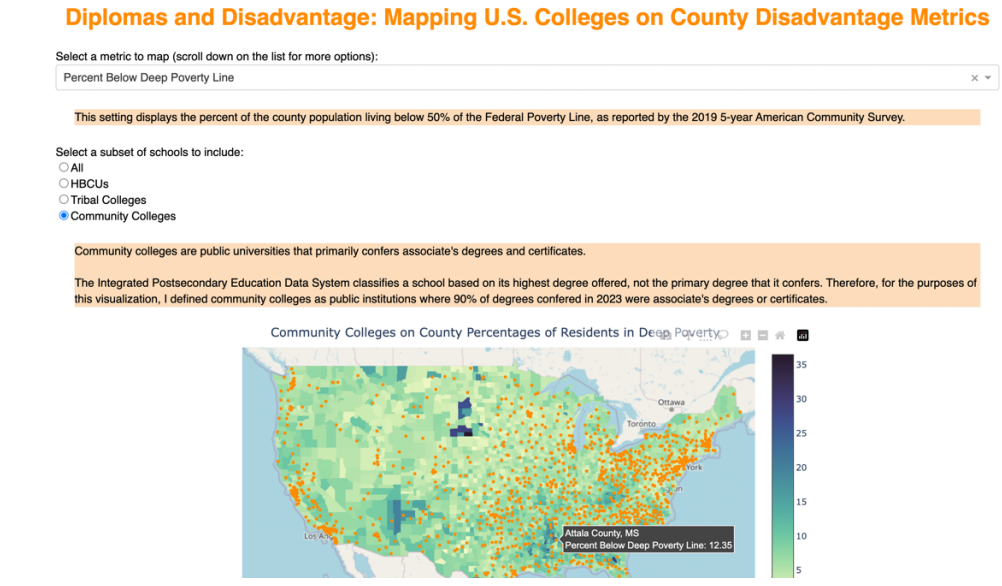
What the page looks like when the app is started:



Here is what the page might look like once the user has chosen a different metric and subset and is hovering over the map to see the tooltip:

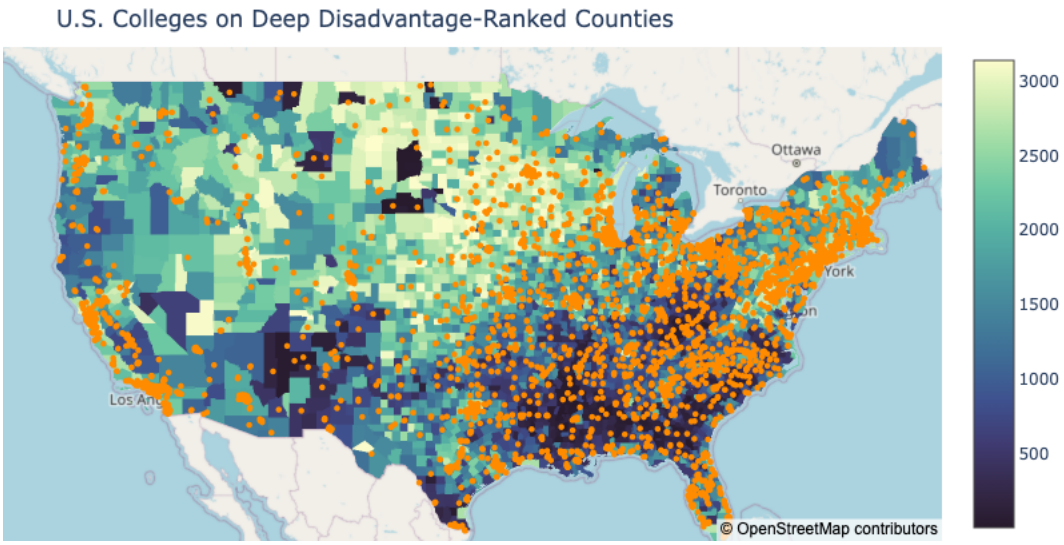


Another example of different settings, this time with school tooltips off:



If the user runs *create_map.py*, they will generate the plotly figure alone instead of the Dash app.

Plotly figure generated by *create_map.py*:



5 Future Work

If I revisit this project in the future, I hope to implement a state-level feature that would filter the data by state and zoom into that specific state on the map. I would also consider a place on the page to upload college data or toggle specific colleges if the user wanted a non-default subset of schools, such as the MDRC portfolio of postsecondary partners I was working with this summer. Finally, I would also consider adding a new tool for statistical analysis. The current tool is useful for generating questions, but not necessarily answering those questions, so I would want to see the use cases for the page in its current form to know what questions should be answered by an analysis tool.

References

[1] Nick Hillman. How Far Do Students Travel for College? *The Institute for College Access & Success*. October 2023. https://ticas.org/wp-content/uploads/2023/11/HIllman-Geography-of- Opportunity-Brief-2_2023.pdf

[2] University of Michigan Poverty Solutions. *Understanding Communities of Deep Disadvantage.* https://poverty.umich.edu/projects/understanding-communities-of-deep-disadvantage/

[3] Wikipedia. *List of United States Counties and County Equivalents*. https://en.wikipedia.org/wiki/List_of_United_States_counties_and_county_equivalents