Wealth/Racial Distribution in Relation to Access to Internet and STEM Jobs

Quynh Doan, Andrew Michaels, and Nolan Kim

Contents

Research Questions

1. How does digital inequality in access to the internet affect school districts within states below the national median household income level?

In the US states below the average median household income line, a similar trend between high school student internet accessibility and STEM degree attainment exists between 2001 to 2011. Although not a strong correlation, there appears to be a positive relationship between a greater percentage of internet accessibility in high school and STEM degree achievement in states where the average median income level is below the national average.

2. Does a correlation exist between living in lower income communities and having quality access to the internet?

Our data indicates that there is a positive correlation between the income level and internet access across the US States. The access to the internet increases when the income level increases.

3. Does a correlation exist between racial groups and low income communities?

Even though there might not seem to be a concrete correlation between racial groups and low income communities, it is important to point out that white and asian racial groups had higher income levels than the “total all races” bar graphs. Black and hispanic groups rated lower on the spectrum.

4. Where do STEM degrees get awarded most frequently and does this depend on the internet access?

It is clearly evident that there is a positive correlation between having quality access to the internet and the possibility of having STEM occupations across US States. The Science and Engineering degrees increase when access to the internet increases.

5. Does a correlation exist between racial groups and internet access?

By looking at our bar plot that compares internet access to each racial group that we were able to find data for, we observed an interesting pattern. Hispanic and black racial groups were below the total average while white and asian groups were above the total average. Although an observation, pointing out this distinction might shed light on the other correlations that we have found from answering the other questions.

Key Takeaways

From our graphs, we can say conclusively that the access to the internet is strongly correlated to income level distribution and also affects highschool students' chance to pursue a STEM degree in university. As income level increases, digital access increases, which leads to higher STEM degrees awarded. However, there was little that we can say about the relation between racial distributions and internet access, other than that White and Asian seem to access more than Black and Hispanic. Interestingly, we can also see this same correlation between racial distributions and income levels. But more research needs to be done to have a clear conclusion about this.

Internet access and STEM degrees distribution in the US is very far from uniform. Rather than being evenly spread out, access to the internet is seen in areas with the highest income populations, such as California, Washington, Wyoming, Colorado, and much of the east coast.

Motivation & Background

As racial injustice and socioeconomic disparity have heightened and grown more transparent in the last decade, Americans have had to face a divided country. With the recent events, such as the BLM protests and the skyrocket advancement of America's 1% elite, understanding and elucidating this divided country are imperative to enact some sort of change. Delving deep into the big picture issues that racial and lower class people face in America, our group has decided to look at digital access and the access to technology and how various groups in America are affected by this issue.

Having access to quality internet holds different meanings for people across the world. For instance, one might need a secure internet connection to attend their classes, while someone else might simply need a connection to access Netflix. Peoples' needs vary. By exploring this topic of quality internet usage in various neighborhoods, we might be able to discover a relationship between certain neighborhoods and the degrees of internet usage.

From our current understanding, many students from lower income communities have less access to technology use and digital classes in schools than others. This results in a downward spiral of a lack of awareness in technology careers and job attainment in the field. Often, by not having this awareness, people are stuck following a single path. There are so many paths that people can choose from, professionally, socially, and personally. Why be limited by where you live and/or the color of your skin. We want to challenge our conceptions to see if what we believe is correct.

We hope that, by exploring these correlations, we will educate ourselves more during the process and for others to see that there exist inequalities in the world that need change.

Datasets

1. https://techdatasociety.asu.edu/broadband-data-portal/dataaccess/statedata

We're using the State2014ACSfactfinder.xls at the bottom of this website. This dataset holds estimates of the percentage of Internet use in the 50 U.S. states along with their 2014 populations. The home internet access is further broken down into race, education, employment status.

2. https://www.census.gov/data/tables/2020/demo/income-poverty/p60-270.html

The dataset presents income levels by households in America. The data is distributed into income levels based on race and ethnicity and presents the population within these categories by year.

3. https://www.nsf.gov/nsb/sei/edTool/data/college-19.html

This dataset presents Science and engineering degrees as a percentage of higher education degrees conferred by state in 2001, 2006 and 2011. We downloaded the data by the Chart data: xls option at the bottom of the site. We would only use the 2011 column, as we reasoned this is the closest to compare to the internet access dataset in 2016.

4. https://github.com/kjhealy/us-county/blob/master/data/geojson/gz_2010_us_040_00_5m.json

This json file presents the Map of US States and territories, with geospatial data. We just need to download it straight from this github account using the Download button.

Methodology

Numbered list correlates to the numbered research question listed earlier.

1. We will investigate internet accessibility in high schools within US states below the national average median household income to understand how students and school programs are affected by degree of internet usage. We used three datasets to answer this question. We began by cleaning the datasets about internet accessibility in high schools, STEM degrees in each state, and average

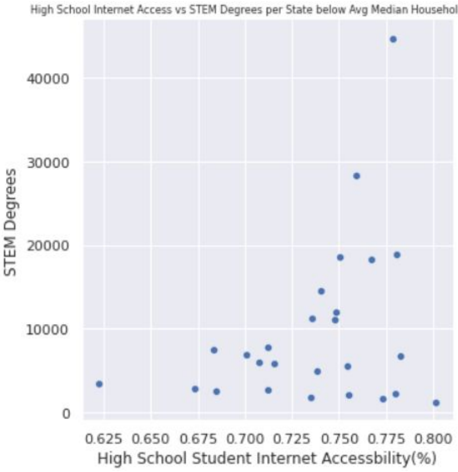
median income levels in each state. In doing so, we removed the excel formatting in these files, corrected the indices, and renamed the columns for easier interpretation. Then, we joined these three datasets together based on the state columns. This merged table was used to create a scatter plot based on internet accessibility in high school by STEM degree attainment in each state. The average median income column should be filtered to those states below the national average. Now, we can see the type of correlation high school internet accessibility may have on achieving a STEM degree.

2. To find a correlation between living in lower income communities and having quality access to the internet, we first plotted a map of the US states besides Hawaii and Alaska, colored by the median household income of each state. Then we plotted them on a scatter plot against the quality of internet access in those states. By including a regression line drawn through them, we would be able to observe if there exists a correlation between the two variables.
3. We will be analyzing how racial group may correlate with income level. In doing so, we created many different bar charts looking at different income categories by the percentage of a racial group that falls into a specific category. We separated our data into five different datasets labeled by the income distribution of each race in our data(asian, black, hispanic, white, all races). Then, we created one function with an input parameter of racial group to generate bar charts separating the income distributions into different income-level categories. These illustrations visualize patterns across racial groups and income levels.
4. By adopting the same concept from the second question, this question was answered by plotting a map of the US states colored by the above average internet access on an empty plot of the US. As this was somewhat difficult to see, we also came up with a different plot. The second plot was a map of the US states colored by above average STEM degrees over all higher degrees achieved on an empty plot of the US. These plots together show the most technically developed states of the United States: like California, Washington D.C, etc with some outliers. We then plotted a scatter plot with a regression line, and found out a positive correlation between these two variables.
5. Taking the dataset about the internet access levels in America by racial category, we calculated the average of the percentage of those with internet access for each state in America by racial category. Then, with this information, we created two different bar graph plots. One was to show the raw average percentage for each racial category with the total average for all races plotted side by side. The other was to show the offset with respect to the total average for all races. Displaying both of these plots next to each other highlights the difference in distribution of the internet access for each race and the patterns that can be found from the information.

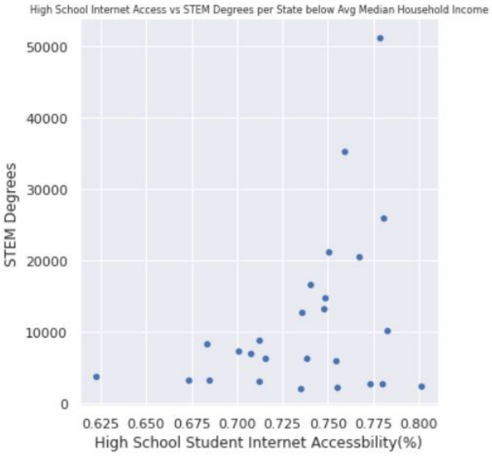
Results

How does digital inequality in access to the internet affect school districts within states below the national median household income level?

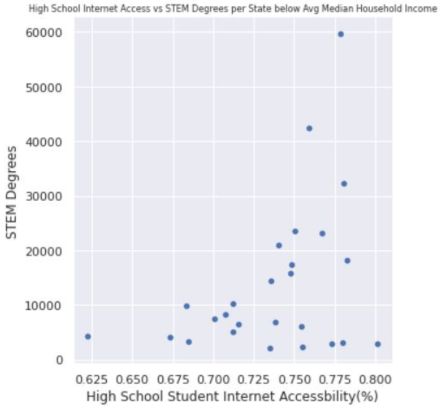
2001



2006



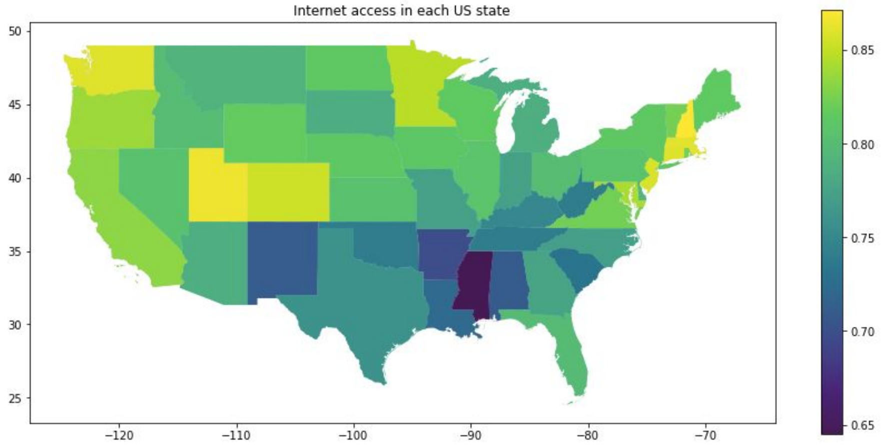
2011

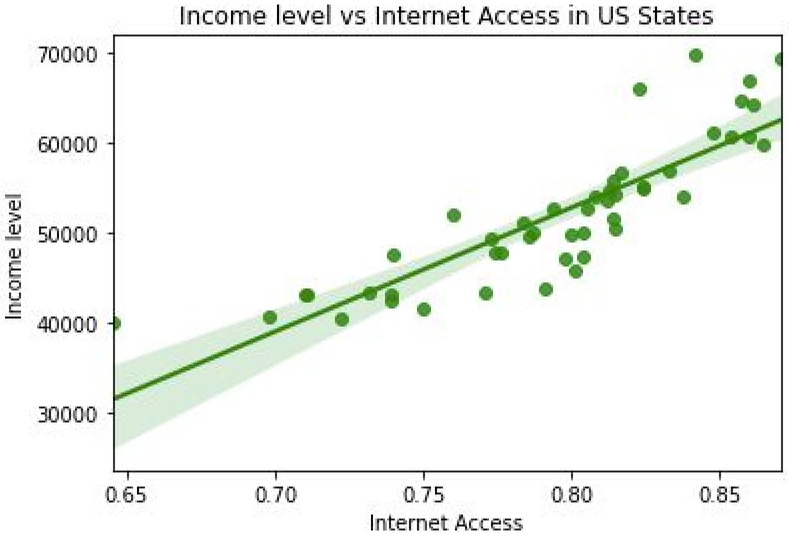


The scatterplots above portray a connection between high school internet accessibility and STEM degree attainment in states below the national average income level between 2001 and 2011. There appears to be a positive relationship between a greater percentage of internet accessibility in high school and STEM degree achievement in states where the average median income level is below the national average, especially as the percentage of high school internet accessibility increases. The correlation also appears to be consistent between the time span. These results are expected as STEM classes, especially those involving computer usage, may require a greater degree of internet usage. Therefore, schools with greater accessibility will be able to provide their students with a better variety of STEM classes if they have the correct technological resources.

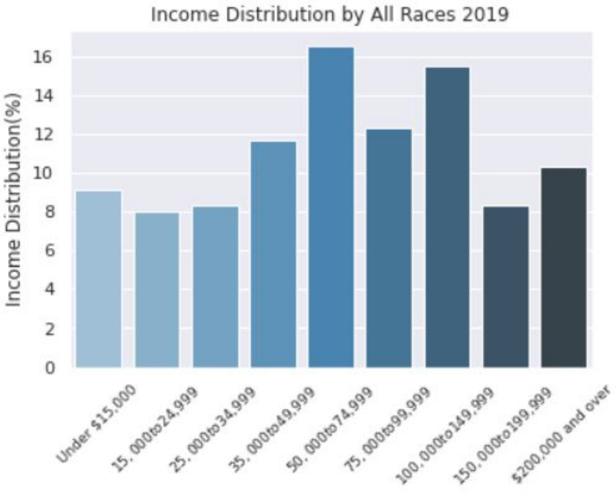
How closely is living in lower income communities related to having quality access to the internet?

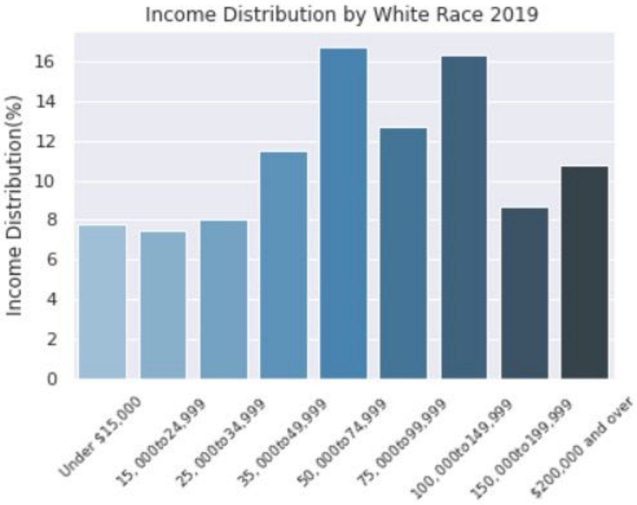
We plotted the access to the internet to a Geospatial map of US States to have an overall view of areas that have high versus low internet access. We can see that internet access is high in states such as California, Colorado, Connecticut, Massachusetts, etc, which are all high income level states. We then plotted a regression plot of median income level per state over the internet access. The correlation of this plot is over 0.8, indicating a moderate to strong correlation.

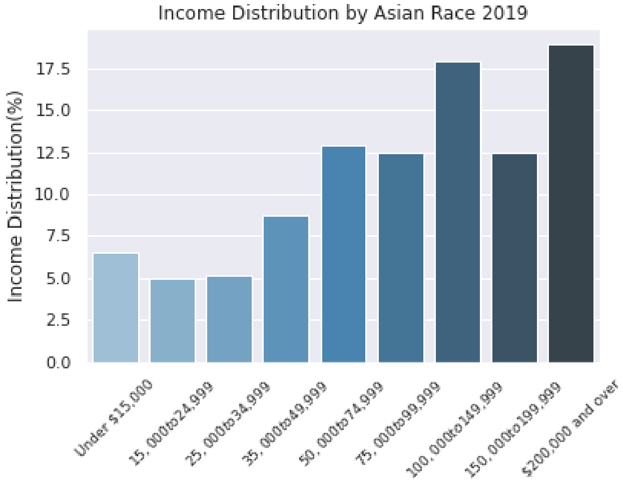


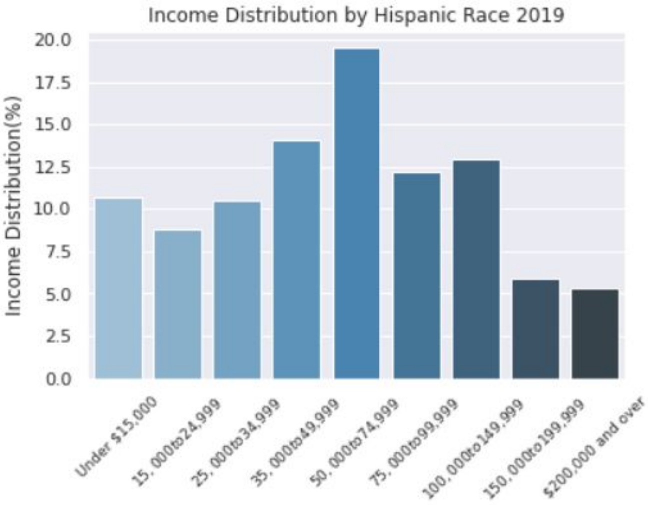


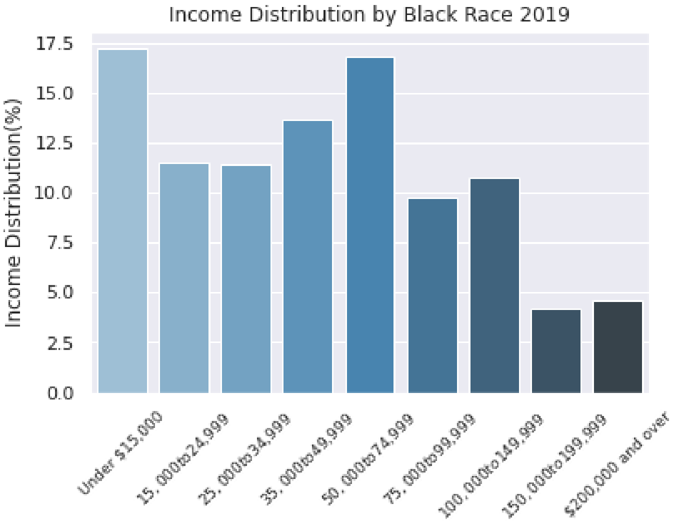
Does a correlation exist between racial groups and low income communities?







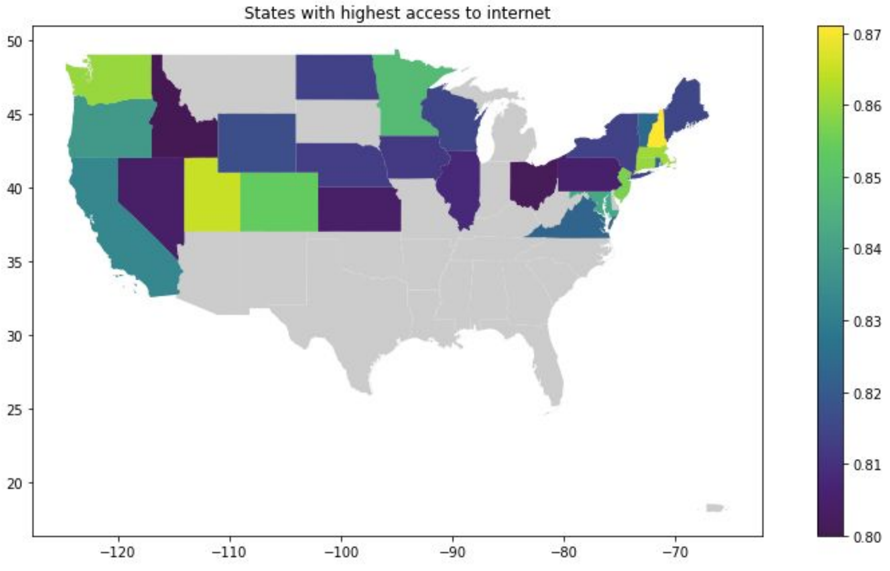




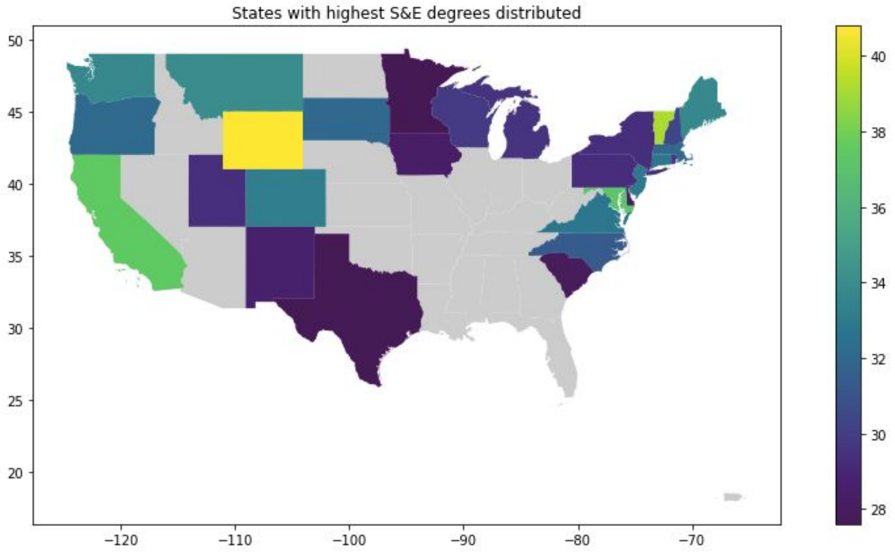
Each of the five graphs above portray a specific story of how each racial group rates against one another for each income category. By the difference in the bar graphs, we can see that white and asian racial groups have, on average, more people in the upper half of the income distribution compared to black and hispanic populations. Based on political and historical fact, these comparisons make sense. Especially with the recent events of BLM, it has been elucidated how the American justice and political system does not support the black population of America. Additionally, for the last eight years, ICE has been mistreating and deporting the Hispanic population. These blatant acts of bias towards these two races might serve as the reason for why the hispanic and black populations rank lower in income levels than white and asian people.

Where do STEM degrees get awarded most frequently and does this depend on the internet access?

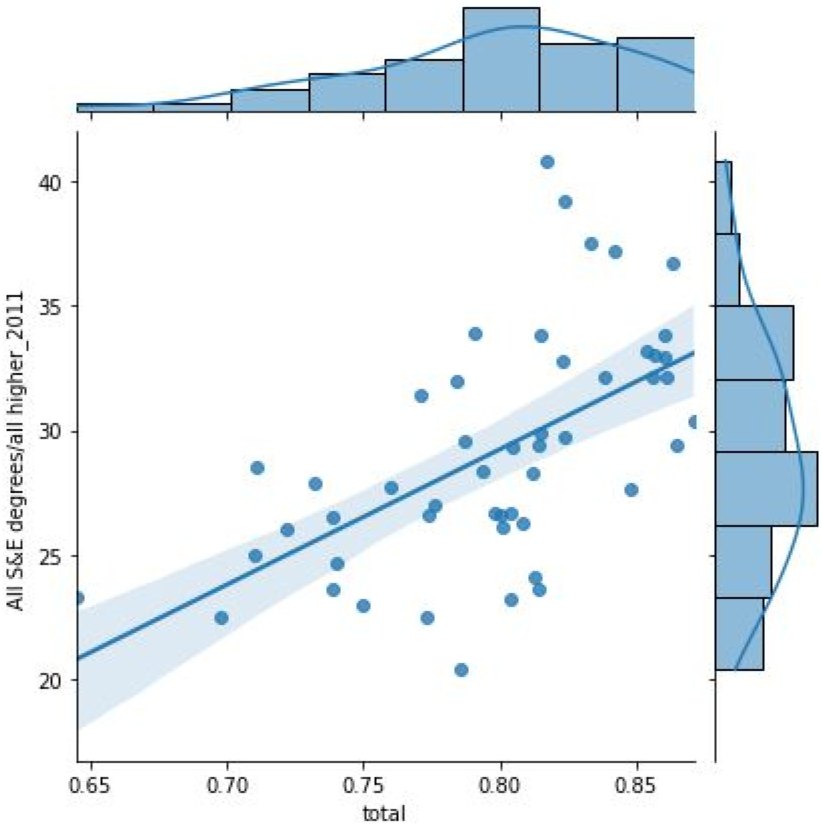
The areas highlighted below are the states with the highest access to the internet in the U.S., highlighted by percentage. The highlighted regions seem to match most of the U.S. states with the highest amount of STEM degrees awarded: California, District of Columbia, Maryland, Vermont, Wyoming, and other higher developed states are all represented below. I further plotted a regression plot between STEM degrees and access to the internet in each State and saw that they are closely related: as internet access increases, the number of STEM degrees awarded increases.



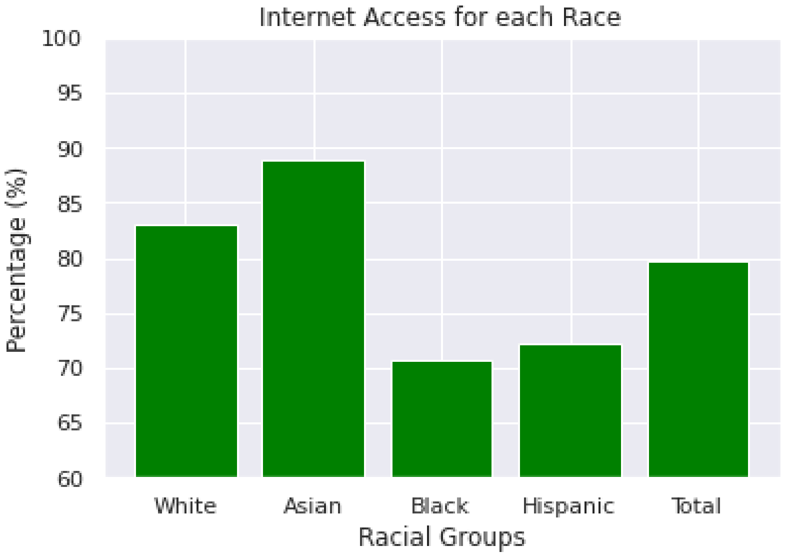
States with highest access to internet



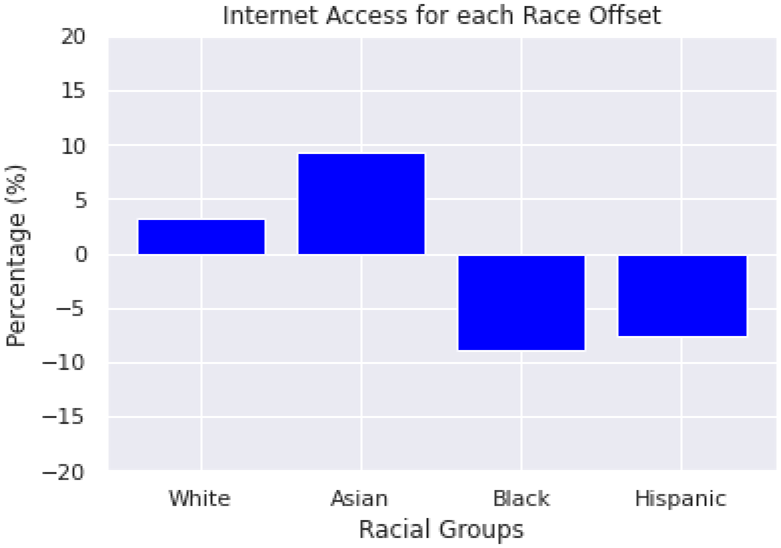
States with highest S&E degrees distributed



Does a correlation exist between racial groups and internet access?



Observing the bar graph, we find it clear that there is a clear difference between hispanic and black racial groups and white and asian racial groups. Although we do not like dividing racial groups in such a manner, it is clear that there is a difference. By taking the “total” column as the reference, which represents the averaged values of all the races, we can see that the averaged percentage values for each race with respect to this reference.



As we can see in this second graph, there is a more distinct difference between the two sets of races: asian and white vs black and hispanic. The conclusions drawn from these two graphs can mean a multitude of things. However, what we can conclude is that there exists a divide in how different races have access to the internet. Knowing this information, it is important to see what the root cause is. Is it income levels? Is it the amount of STEM degrees? The fact that we see the same correlation between racial groups and income levels seems to be an interesting parallel. Which causes which? We seem to have “a chicken or the egg” situation here, and further research should be done to see what came first.

Challenge Goals

Merging Multiple datasets:

- We will use multiple datasets and join them together, completing the multiple datasets challenge goal.
- We can join the internet access accessibility by state with STEM degrees, and Median Household Incomes data tables. These three datasets include information about US States on STEM degrees and internet access across racial and educational lines. This merged table will help us understand internet access and STEM degrees by income level and race/ethnicity in households in America.

Then, we will join these merged tables to the geospatial map of the US states to include more data on internet access and STEM degrees distribution in each state in America.

- This data table will give us insights into possible patterns between income level, race/ethnicity, and internet accessibility. Using this combined data table, we will try to understand an inequality in one dimension of technology (internet accessibility) between racial lines and income levels.

Handling messy data:

- Much of our data is not in a clear csv file format, and needs to be processed and filtered, which will meet the messy data challenge goal.
- Because the numbers in our datasets are mostly in percentage as strings we have to convert them into floats instead to calculate the mean and median.
- We will need to create columns in our datasets to join by in order to create our complete merged dataset.
- We will need to clean our data tables that were made in excel and remove all excel formatting and reconstruct the indices.

Work Plan Evaluation

For our testing and coding, we are going to use a Google Colab notebook that will be shared amongst all the group members.

1. Load and Clean datasets (approximately 5 hours)
    a. Figure out meaning of columns and data by going to each website where the datasets are found
    b. Convert data in State vs Internet Access from strings into floats that would be useful for analysis.
    c. Having various formats after merging these datasets together, as we needed it in DataFrame and GeoDataFrame for different plots.
2. Join datasets and Data manipulation (approximately 7-8 hours)
    a. Combine datasets Internet Accessibility by state, STEM degrees, and Median Household Incomes data table by state, educational levels, and racial/ethnicity.
    b. Combine the merged dataset from part a with the dataset d (json file of the map of the US) by internet accessibility and STEM degrees distributed in each state of the US.
    c. Filter through datasets for below average and above average income level, internet accessibility, and STEM degrees over overall higher degrees achieved by each US state.

3. Plot datasets (approximately 2 hours)
    a. Question 1: Plot data over time for highschool internet access by STEM degrees for each US state
    b. Question 2: Plot the overall internet access by state on a map of the U.S and a regression plot between internet access and median income level.
    c. Question 3: Plot bar charts of income distribution of all races, white, asian, black, and hispanic.
    d. Question 4: Plot the above average internet accessibility and S&E degrees distribution by state on a map of the U.S to compare between them. Then a regression plot and bar graphs between these variables.
    e. Question 5: Plot a bar graph of internet access by race/ethnicity and another bar graph offset average for all races.
4. Report Writing (approximately 5 hours)
    a. Perform analysis as necessary, trying to find a correlation between wealth/racial inequality and internet accessibility.

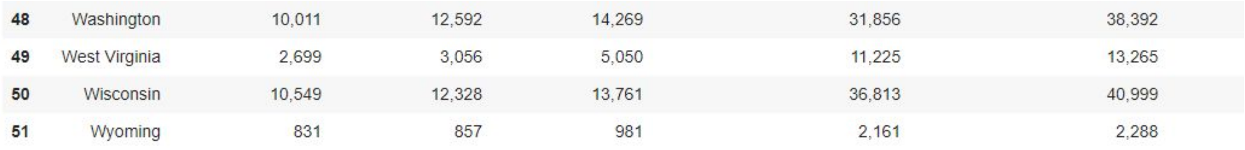
We were reasonably accurate in dividing our work and estimating the time for each part. The part to clean our datasets, join them, and manipulate our data to perform analysis took the most of our time, since we basically have to find all our datasets from scratch again after realizing that our datasets originally didn't have enough data needed for calculated columns. We knew that the quality of our datasets would need extra attention and care since this determines the quality of our plots and analysis in the end.

We also ended up only finding one dataset for median household income for each US State in 1 year, originally we wanted to find datasets over the years to investigate the strength of the correlations we might find in the plots. However, we found data on STEM degrees distributed in each state and different educational internet accessibility, so we could do more in-depth analysis with that.

Testing

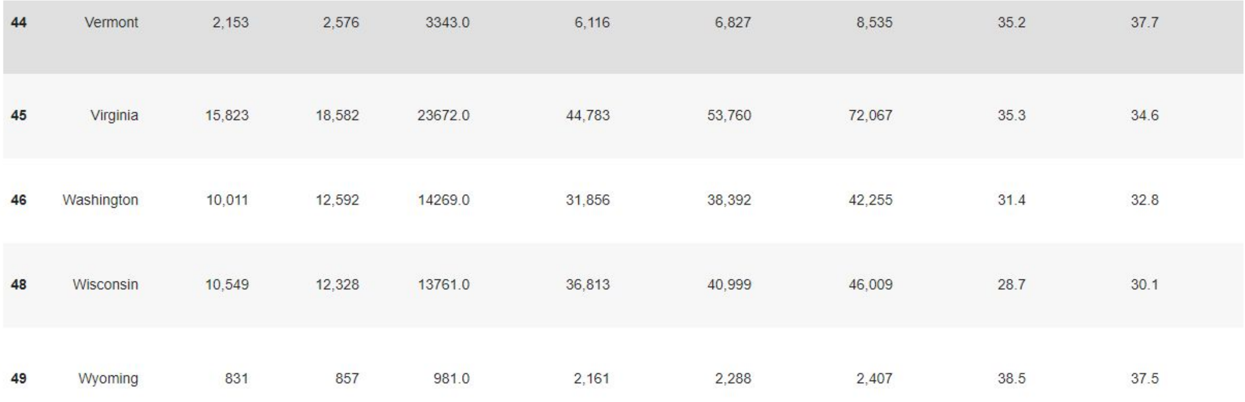
For all of our tests for each python script, we used the “.head()” and “.tail()” functions to compare our researched data file with our desired output, whether that is a cleaned or joined dataset or a plot. For each case, we have a display of the before and after to see if our functions worked properly. We couldn’t test out outputs of plots because we weren’t entirely sure what to expect. Therefore, the best tests that we could implement were those that were concerned with confirming whether or not the data that we cleaned was plotted properly.

Clean Test



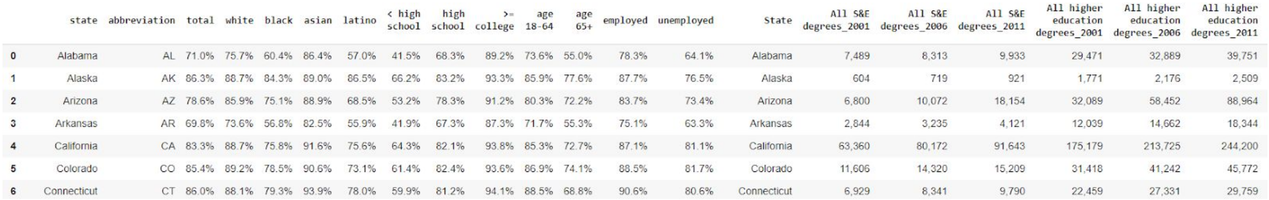
48	Washington	10,011	12,592	14,269	31,856	38,392
49	West Virginia	2,699	3,056	5,050	11,225	13,265
50	Wisconsin	10,549	12,328	13,761	36,813	40,999
51	Wyoming	831	857	981	2,161	2,288

To test our cleaning functions, we used the .tail() function to compare the values to the original excel data files. One can spot the difference between the upper and lower pictures. These are the bottom values of the dataframes, and after the filtering, the upper picture became the lower pictured dataframe. Using these types of checks, we were able to validate whether or not our masks and other filtering methods functioned properly.



44	Vermont	2,153	2,576	3343.0	6,116	6,827	8,535	35.2	37.7
45	Virginia	15,823	18,582	23672.0	44,783	53,760	72,067	35.3	34.6
46	Washington	10,011	12,592	14269.0	31,856	38,392	42,255	31.4	32.8
48	Wisconsin	10,549	12,328	13761.0	36,813	40,999	46,009	28.7	30.1
49	Wyoming	831	857	981.0	2,161	2,288	2,407	38.5	37.5

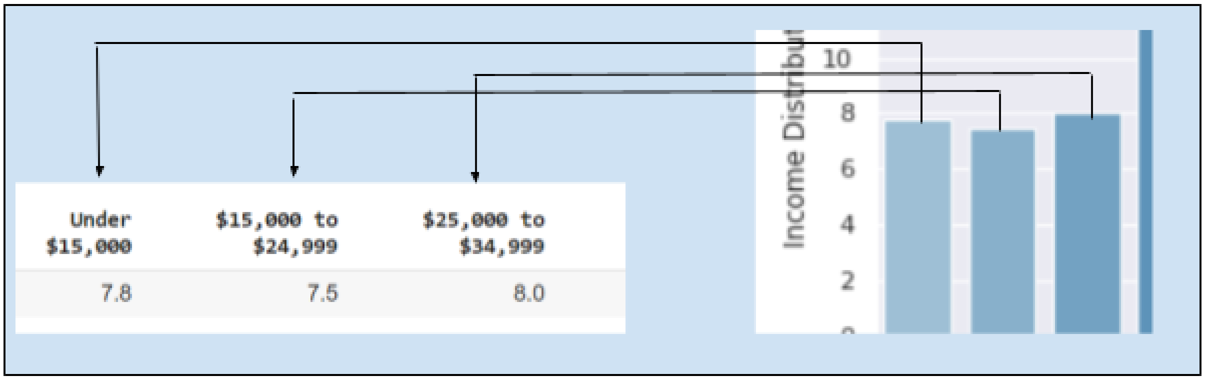
Join Test



	state	abbreviation	total	white	black	asian	latino	< high school	high school	>= college	age 18-64	age 65+	employed	unemployed	State	All S&E degrees_2001	All S&E degrees_2006	All S&E degrees_2011	All higher education degrees_2001	All higher education degrees_2006	All higher education degrees_2011
0	Alabama	AL	71.0%	75.7%	60.4%	86.4%	57.0%	41.5%	68.3%	89.2%	73.6%	55.0%	78.3%	64.1%	Alabama	7,489	8,313	9,933	29,471	32,889	39,751
1	Alaska	AK	86.3%	88.7%	84.3%	89.0%	86.5%	66.2%	83.2%	93.3%	85.9%	77.6%	87.7%	76.5%	Alaska	604	719	921	1,771	2,176	2,509
2	Arizona	AZ	78.6%	85.9%	75.1%	88.9%	68.5%	53.2%	78.3%	91.2%	80.3%	72.2%	83.7%	73.4%	Arizona	6,800	10,072	18,154	32,089	58,452	88,964
3	Arkansas	AR	69.8%	73.6%	56.8%	82.5%	55.9%	41.9%	67.3%	87.3%	71.7%	55.3%	75.1%	63.3%	Arkansas	2,844	3,235	4,121	12,039	14,662	18,344
4	California	CA	83.3%	88.7%	75.8%	91.6%	75.6%	64.3%	82.1%	93.8%	85.3%	72.7%	87.1%	81.1%	California	63,360	80,172	91,643	175,179	213,725	244,200
5	Colorado	CO	85.4%	89.2%	78.5%	90.6%	73.1%	61.4%	82.4%	93.6%	86.9%	74.1%	88.5%	81.7%	Colorado	11,606	14,320	15,209	31,418	41,242	45,772
6	Connecticut	CT	86.0%	88.1%	79.3%	93.9%	78.0%	59.9%	81.2%	94.1%	88.5%	68.8%	90.6%	80.6%	Connecticut	6,929	8,341	9,790	22,459	27,331	29,759

When we joined different datasets, looking at the first seven elements of the dataset, we were able to determine if two datasets were able to successfully join. In the above picture, we have one of our join examples. We implemented an inner join involving two datasets, one about state vs internet access and the other about state vs degree types. We joined on the ‘State’ and ‘state’ columns, and as we can see, the join worked. This type of try and check is how our joins were tested.

Plot Test



Under $15,000	$15,000 to $24,999	$25,000 to $34,999
7.8	7.5	8.0

This is an example of a visual check using .head(1) for the income distribution data tables to compare the values to the bar charts of income distributions by race. For this example, we chose to compare the “7.5” value in the “Under $15,000” column of the “Income Distribution by White Race 2019” bar chart. As we can see above, the values match, along with the other column values for “$25,000 to $24,999$” and “$25,000 to $34,000.”

Collaboration

The work was done by Quynh Doan, Andrew Michaels, and Nolan Kim. We seeked outside help with the complexity of Geopandas, Panda datasets, Matplotlib, Seaborn through consultation of tutorials and message boards such as Stackoverflow, etc, but no outside help was given with the data analysis.