Nabeel Haider

May 18, 2025

Final Report

In this project, I explore how the educational outcomes change across the different regions in the United States, and how these differences are related to the different factors, or columns, in the data set. More specifically, I answer how the test scores vary across schools by region and gender, I also answer how student outcomes are different across the different regions and institution types. In addition, I explore the patterns or disparities that exist in the educational system when considering race, gender, and location. For the final project, I decided to use the data from the .csv file found on the U.S. Department of Education website. I downloaded the Most Recent Institution-Level Data from the website and loaded the file into a data frame in R; the data was last updated on April 23, 2025. By cleaning, organizing, extrapolating, visualizing, and plotting the given data, I hope to uncover some truths with lots of merit. Given that I am a college student myself and have been exposed to the controversial topic of education and its effect on the individuals who pursue higher level education, I find great value in working and deriving insights under a topic that speaks to issues close to my experience. After reading the data in and getting rid of all of the immediate unnecessary data, the data frame has 6429 rows and 3229 columns.

To answer the main questions that I identified, I identified that there were 59 unique states in the data set, which weren't all in the continental United States; DC and Puerto Rico were also considered to be states in this data set as opposed to the traditional 50 U.S. States. I printed all the unique state and region combinations and uncovered the region that each of the numeric valued region corresponded to. Region 0 wasn't consistently grouped with a state, so I got rid of the 0 region values as I came to the conclusion that it was for unknown regions. After reading the state abbreviations and region numbers side by side, I saw that region 1 is considered to be New

England, region 2 is the Mid-Atlantic, region 3 is the Eastern Midwest, region 4 is the Western Midwest, region 5 is the South Atlantic, region 6 is the Southwest, region 7 is Mountain West, region 8 is the West Coast, and finally, 9 groups the regions outside of the U.S. This first initial step is important for all the geospatial analysis in addition to making the faceted plots region by region used to help us derive insights into how equitable education is across the United States and finding patterns across the board.

After cleaning up the data, figuring out the numeric regions in the United States, and understanding the data, I continued to my first figure, which is a faceted figure where I group the data, which is the SAT Average by region and then by gender. Firstly, I filtered out all the rows where there is not a value for the SAT Average, percentage of women undergraduates, and percentage of men undergraduates as those rows wouldn't help with the faceted plot. In addition, I filter out the rows that don't have a value for region or isn't a known region or in other words a numeric region from 1 to 9. In the faceted plot, majority women mean greater than 60% of the students are women; majority men mean greater than 60% of the students are men. In all other cases, where neither men nor women are greater than 60%, is considered to be balanced. In addition, I make sure to map the numerical region values to the regions which I mentioned before such as 1 being New England and 8 being the Pacific or the West Coast. I calculate the average SAT for each group in each region, and only if the number of schools is greater than or equal to five. After generating the faceted figure, I can better see how the SAT scores are different across the different regions and by the gender composition of the student populations. After analyzing the data via. the faceted plot, I can see that the balanced gender composition group has a better average than either majority men and majority women though it isn't always the case in every region. It's

also important to note that the average SAT scores of the outlying areas is drastically lower than the other nine regions.

For my next figure, I created a geospatial map of the SAT Averages by State in the Continental United States. Using this geospatial map, I can identify the regional trends in the SAT performance from a geographic point of view. To start off, I filtered out the rows where the SAT Average was missing and then I grouped the data all corresponding to its respective state finding its average. I used the map_data function from ggplot2 to match each state to it's name. I ended up merging the average SAT data with the map data which has the coordinates needed to outline and essentially draw out each of the U.S. States. After plotting the data using the map data, it's easy to see the clusters and the trends across the regions. After seeing the geospatial plot, I see that Wyoming is the only state greyed out meaning that it's the only state within the continental United States that didn't have any data in the data frame. The color gradient, as shown in the legend, ranges from purple, which shows a lower average SAT score, to yellow, which is a higher average SAT score. We can see that the states in New England, such as Massachusetts and Vermont, are a lot more yellow in addition to California, which shows that these states have a higher SAT average. As opposed to these states, we see that states in the South Atlantic and the plains regions, such as Mississippi, Kansas, West Virginia, etc. are a lot more purple/blue signifying a lower SAT Average.

In my third figure, I highlight the average student demographics including racial and ethnic groups across the different regions using a faceted plot. The data used the following racial/ethnic categories to group by which are: White, Black, Hispanic, Asian, Native American, Pacific Islander, and multiracial students. Each facet is a different region; the different bars show the different percentages of each group which allows to compare the different groups a lot more easily.

Nabeel Haider

May 18, 2025

From the plot, we can see from the different plots that the demographics themselves shift from region to region. One major find is that the outlying regions have significantly more Hispanic and pacific islanders compared to the rest of the faceted regions. In addition, the South Eastern United States has more Black individuals relative to the rest of the regions. This visualization helps highlight the ethnic and racial composition of the schools in the different regions across the United States.

For my next figure, I created a boxplot comparing the graduation rate by the school size. The three different school types are: large schools, which have greater than or equal to 5,000 students, medium schools, which range from 1,000 to 5,000 students, and small schools, which are less than 1,000 students. Using the boxplot, we can see that there is an evident correlation between the school size and the graduation rate. We see that larger schools have the highest graduation rate followed by medium schools followed by small schools and the variance also increases from large schools to small schools. In other words, the possible outcomes of going to a smaller school is a lot higher than the possible outcomes of going to a larger school. Large schools are a lot more consistent in terms of the outcome. One factor that could be influencing this is that larger schools typically have more funding which translates to more and better resources used to help students stay engaged with the courses. Another possibility is that the students at the smaller institutions may be more inclined to go to the bigger institutions for better resources and opportunities. With more students, it's possible that students at these larger universities are motivated by their peers which influences their drive to graduate. Being around more students creates an environment conducive to collaboration which could also increase the graduation rate indirectly.

Nabeel Haider

May 18, 2025

My fourth figure is a heatmap comparing the average median earnings by region and school type. The median earnings are recorded 10 years after graduation. The three school types are: Private For-Profit, Private Nonprofit, and public and the regions are labeled how they are in the previous figures. After creating the heat plot I note certain trends; I see that the median earnings after going to school in New England and the Mid Atlantic is a lot higher on average for all three school types compared to all of the other regions. In addition, I also see that private nonprofit institutions have the highest median earnings in every region except for the outlying areas but has a significantly higher average median earning across all the regions. Conversely, private for-profit institutions have the lowest median earnings across all the regions. Another interesting find is that the median earnings in outlying areas is a lot lower in all the school types. In addition, the highest median earnings in the outlying areas are public schools, which generally isn't the case, and the median earning is $30,475.19. Using this data, it seems that one shouldn't go to school in the outlying areas as the median earnings is a lot lower or the cost of living in the outlying areas is a lot lower. The best place to go to school based on median earnings is either New England, the Mid Atlantic, or the Pacific/West Coast.

For my next figure, I created a heatmap showing the weighted average graduation rates across the different U.S. regions and ethnic groups. The graduation rates are weighted by the fraction of the ethnicity attending the university in the region, which ensures both the performance and representation. In the figure, the darker squares show the higher graduation rates, while the lighter, or more yellow squares show the lower graduation rates. We see that the graduation rates in New England and the Mid-Atlantic are the highest compared to all of the other regions whereas the graduation rates in areas such as the outlying areas, or the southwest are lower. In addition, we

can also see that White people and Asians tend to have higher graduation rates compared to the other ethnicities and races. On the other hand, we see that the graduation rates are a lot lower for Natives, Hispanics, and Blacks across all the regions. It's important to note this disparity as it highlights something deeper. These gaps could be a result of the historical marginalization, the support of the institutions, or socioeconomic status. By creating this figure, in addition to the disparities, we also see that there needs to be a more equitable approach to education to reduce the marginalization that we see among the different races.

For my seventh figure, I created a violin/box plot to show the relation between the graduation rate by the three different school types, which are private for-profit schools, private nonprofit schools, and public schools. The violins show the density and variance of the graduation rates across the three groups and the boxplot is on top to better visualize the exact statistics of the grouped data. Using the figure, we see that private nonprofit universities have the highest graduation rates followed by private and followed by public. This shows that nonprofit schools tend to perform better compared to the other groups, especially the private for-profit university group. In addition, private for-profit institutions have the most inconsistent graduation rate as the variance is a lot higher whereas the graduation rate of public schools is slightly lower than the graduation rate of private institutions but with less variance which points at a flaw in private for profit universities. This could indicate that the students at private for profit institutions aren't receiving enough resources or support compared to the other institution types. In addition, this information could highlight the issue of for profit institutions focusing on profit a lot more compared to the education for its students.

Nabeel Haider

May 18, 2025

For my last figure, I created a scatter plot to find the relationship between the percent of Pell Grant recipients enrolled at university and the median earnings of the graduates. The percentage of Pell Grant recipients is meant to highlight the percent of low income students and serves as a proxy. In addition, the median earnings of the graduates is the median earning 10 years after graduation. The different points in the figure are individual institutions which are either red, green, or blue. In the legend, we see that the blue points are public institutions, the red points are private for profit institutions, and the green points are private nonprofit institutions. The figure shows a linear regression for each of the three groups colored the same as their respective groups and uncovers some trends. The regression line for all three groups have a negative slope, but some groups are less negative than others. The more negative a slope is, the more it is that low income is associated with a larger drop in median earnings meaning income is more of a determinant of median earnings. At a 0% Pell Grant rate, the median earnings were higher for private nonprofit universities followed by public universities and then private for profit universities. At a Pell Grant Rate of 1%, the median earnings were higher for public school students followed by private for profit universities followed by private nonprofit universities. This data highlights that public universities show a lot more consistency across the different Pell Grants giving more stability as opposed to the other institution types. The median earnings of private for profit universities after graduation are generally worse than public universities irrespective of income and the lower income individuals have a higher median earning after graduation highlighting that it's likely that public universities do the most to support low income individuals. Public universities may be better suited to handle low income students whereas private nonprofit institutions don't do enough to support low income students. This finding shows the flaws that exist in private nonprofit universities as they don't address equity nor access in the realm of higher education.

Nabeel Haider

May 18, 2025

All in all, through my analysis of graduation rates, earnings, and the institution types, there are many different patterns that reveal systematic inequalities and trends across the board in higher education institutes across the United States. By analyzing the data via. faceted plots, geospatial maps, box plots, violin plots, heat maps, and k linear regressions, we were able to uncover findings and draw some conclusions. By interpreting a lot of these plots, I could compare the outcomes across graduation rates, low income students via. Pell Grant percentage, and the median post-graduation earnings after ten years.

The violin plot and box plot showed me that public universities had the most consistent graduation rates and has a more predictable outcome compared to the other two institution types. The violin plot showed me that the private for profit data is a lot more skewed meaning that the students aren't getting the adequate resources that they need to help them through their degrees. My final figure, the three independent linear regressions, shows how socioeconomic status can influence the median earnings after graduation for the three different school types. The relationship between the different school type and the Pell Grant percentage is negative, but we see that private nonprofit institutions have a more negative relation compared to private nonprofit or private for profit schools. This shows that there are disparities in the support structure or the resources at the university level. Public universities, on the other hand, are a lot more stable and the slope is a lot less negative and shows that the median earnings is the highest for higher fractions of Pell Grant recipients.

The data reveals something really concerning: low-income individuals experience worse outcomes across the institution types and some schools are better suited to handle low income individuals. For profit universities should be regulated a lot more to ensure that all students,

regardless of their socioeconomic background or status, are given a fair and equitable shot at not only their educational success but their post-collegiate success.