# U.S. Regional Collegiate Analysis

McKenzie Fuller, Mariah Moore, Maddie Thomas

MSCI:6060 Data Programming in R

Fall 2019

# The Situation

For our project we chose to analyze data from universities and colleges around the United States. We obtained a data set from The Integrated Postsecondary Education Data System (IPEDS) containing data regarding tuition amounts, graduation rates, diversity percentages, and additional statistics for U.S. colleges from 2010-2014. The main goal of our analysis was to examine how schools from each U.S. region compare to one another on tuition rates, graduation rates, and financial aid awarded.

After cleaning and uploading our data into R, we evaluated the different variables in comparison to the location of the university (state and region). In doing so, we hoped to gain a better grasp of how the geographic location of a university can impact other aspects of that university. We also hoped to find relationships and common characteristics among colleges and universities in the different regions.

It is important to note that this data set is from 2015, therefore, it may not accurately reflect current trends of the universities and regions. Since we reference the geographic regions throughout this report, it is important to note which geographic regions refer to which states:

* Far West – Alaska, California, Hawaii, Nevada, Oregon, Washington
* Great Lakes – Illinois, Indiana, Michigan, Ohio, Wisconsin
* Mid East – Delaware, District of Colombia, Maryland, New Jersey, New York, Pennsylvania
* New England – Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont
* Plains – Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota
* Rocky Mountains – Colorado, Idaho, Montana, Utah, Wyoming
* Southeast – Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, West Virginia
* Southwest – Arizona, New Mexico, Oklahoma, Texas

# Answering the Question

Our group cleaned the dataset in Excel before uploading to R and after making a few minor modifications to the data frame, we set out to analyze our data visually. Our research focused on comparisons of university attributes across the different regions of the United States as well as across all individual states. To explore our main question, we researched the different attributes listed below:

Regional Comparisons:

* Tuition
* Graduation Rates
* Percent of Admittance
* Financial Aid Awarded

For this project, we focused on exploratory analysis to discover trends between colleges from different regions. First, we set out to examine how tuition rates varied between the different regions using a box plot to show the regions’ average tuition rates. We found that New England had the highest of the 8 regions and Rocky Mountains had the lowest (see Figure 1). The higher tuition rates seen in the New England region makes sense because most of the Ivy League schools are located there and they typically have higher tuition rates.

Figure 1


Figure 1

Next, we examined the percent of students receiving financial aid by region using another box plot. We found the regions were comparable, offering financial aid to between 80-100% of students (See Figure 2). The regions that stood out were the Plains and Southeast, which offered financial aid to most of their students, while the Far West and Rocky Mountains seemed to have higher variations in financial aid provided. The lower amount of financial aid in the Rocky Mountain region makes sense due to their low average tuition rates. We initially expected to see the Far West and New England regions offering more financial aid due to their higher tuition rates. Please refer to Appendix – Figure 13 for information on how tuition affects the percent of financial aid for each region and Appendix – Figure 14 for a breakdown of financial aid by state.

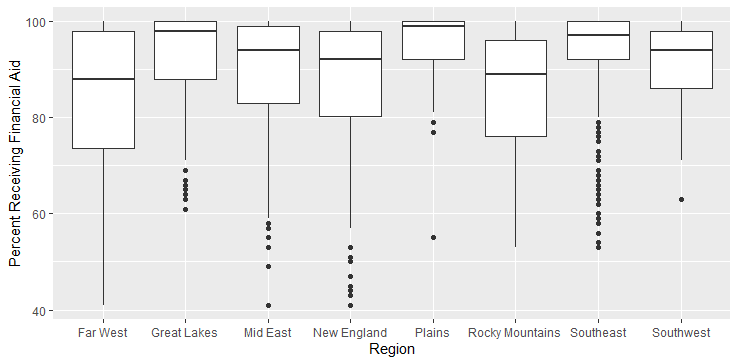


Figure 2

The last section of our main analysis focused on how tuition rates affected the percent of students admitted within each region and that region’s graduation rate. Overall, the regions did not differ greatly when comparing them across the specified variables. We noticed a common trend, shown in Figure 3, where schools with a low percent of students admitted and high graduation rate had the highest tuitions. On the contrast, colleges with high acceptance rates and low graduation rate had lower tuitions. This fits our intuition that colleges with higher academic success cost more, have a higher graduation rate, and a lower acceptance rate. See Appendix – Figure 11 for a general overview of how tuition affects enrollment rates and Appendix – Figure 12 for a breakdown of enrollment numbers by state.

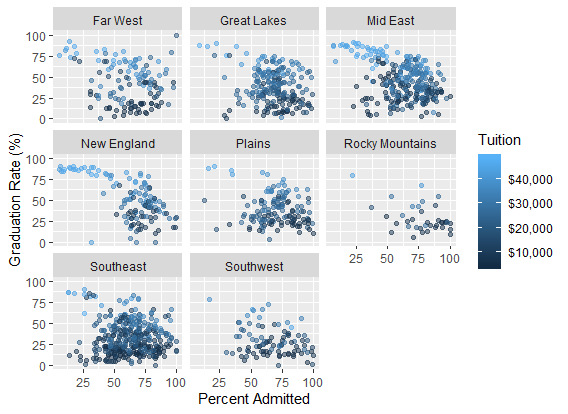


Figure 3

After completing our initial analysis looking at how each region compared regarding tuition, graduation rate, percent admitted, and financial aid provided, we ultimately found that the New England region was one of the most competitive regions for academic success. They demonstrated higher tuition rates, high graduation rates, and a low percent of admitted students. However, we observed similar trends in the Mid East and Far West regions. The rest of our analysis looks at other characteristics of U.S. colleges.

# Further Analysis

To continue our exploratory analysis looking at different aspects of U.S. colleges, we compared the following with visualizations created in R:

* Regional: Number of Institutions, Public vs. Private Schools, Religious Affiliation, and Level of Urbanization
* State Comparisons: Diversity Rates, Graduation Rates, and Financial Aid Rates

We believe that these comparisons and results are useful to students seeking more information about prospective universities. This information is also useful for universities who want to make changes based on their own statistics or regional statistics. Finally, this information could be useful for state budgets and financial aid distribution concerning various institutions and regions.

## Number of Institutions

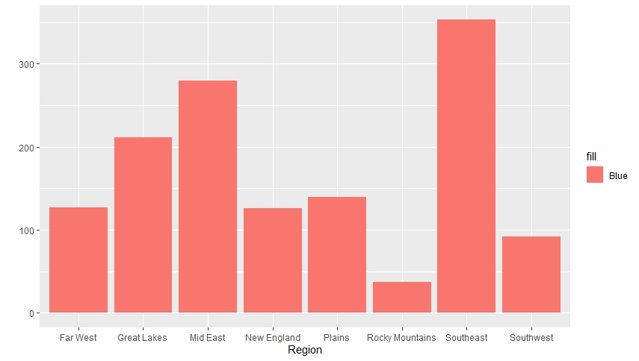


Figure 4

The chart above (Figure 4) represents the number of institutions in each region. The Southeast has the highest number of institutions across the 12 states in the region. The Rocky Mountain region has the lowest number of institutions across its 5 states. The number of institutions is relatively consistent with the regions’ size and population.

## Public vs. Private

The charts below display the number of private and public schools (Figure 5), as well as religious affiliation (Figure 6), in each region. Overall, the relationship between number of private institutions and number of religious institutions seems to be strong.

Notable results include:

* The Southeast has highest count of private and religious schools
* In the Great Lakes, Mid East, New England, and Plains there is more than double the number of private institutions compared to public
* Only the Rocky Mountains and Southwest have more public than private institutions

Potential explanations for these results include:

* The states in the Southeast are historically known for being more religious than other areas of the United States, thus influencing the universities in those states
* New England and the Mid East surround Washington D.C., where there is a higher need for lawyers and political education, and private institutions are often associated with these types of education
* Rocky Mountains, Southwest, and Far West have lower levels of private institutions because the west of the United States is much newer than the east

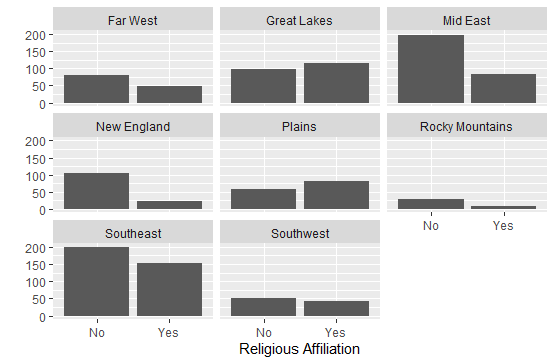


Figure 5

Figure 6

## Diversity

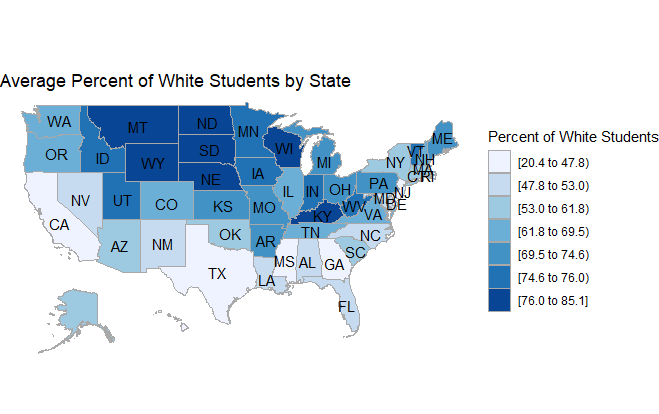


Figure 7

The map in Figure 7 depicts the average diversity of schools in each state by displaying the average percentage of white students at that state’s universities. The darker blue the state is, the higher the average percentage of white students.

* Least Diverse: Montana, Wyoming, South Dakota, North Dakota, Nebraska, Wisconsin, Kentucky
* Most Diverse: California, Mississippi, Georgia, Texas, New Jersey, Maryland

The least diverse states on average contain the highest percentage of white students (above 76%) than other states in the U.S. In the states with the most diversity, the percentage of other races, including African American, Asian, and Native American, are higher on average.

Other insights to the data show that southern states tend to have much higher levels of diversity than northern states, which are typically very blue. Some outliers to this include Illinois, which has much higher diversity than other Midwest region states. Our results seem to have a high correlation with the diversity of the actual population in each state.

## Urbanization



Figure 8

The bar chart in Figure 8 compares the urbanization level (city, rural, suburb, or town) of the institutions among the different regions. Most institutions appear to be located in cities. The next highest percentages of colleges were located in towns or suburbs, and the fewest were located in rural areas.

Some outliers to this result included:

* Plains: The number of institutions in towns outnumbered the schools located in cities.
* Mid East: There is a higher number of schools located in suburbs than in cities.

We also found that the Southeast region by far had the most schools located in cities compared to other regions. The Rocky Mountain region was the only region that did not have any schools in rural areas, but also has the lowest number of institutions. While still a small proportion, there were more rural schools in the New England area than in towns. Overall, the largest or second largest grouping of schools between all regions were in urban areas. This is most likely due to the growing desire to be in an urban area as it presents more opportunities and resources for students.

# Getting the Data

We obtained our data set from The Integrated Postsecondary Education Data System (IPEDS) website at <http://nces.ed.gov/ipeds/datacenter/Default.aspx>. The dataset contained 146 variables for 1,535 colleges in the form of an Excel spreadsheet with the data, variable definitions, notes on the data, and data definitions. In Excel, we filtered out any schools that did not offer 4-year degrees and removed several columns that we did not for our analysis.

Once we saved the Excel file as a .csv, we read the file as a data frame in R and removed all records with any missing data. After cleaning the data, we were left with 1,366 cases and 20 variables including: Name, Religious Affiliation, Percent Accepted, Tuition, State location, Region location, Private or public, Level of urbanization, Enrollment, Race and Ethnicity of students, Graduation rates, and Portion of students receiving financial aid. We also added a column in R which grouped graduation rates into 4 bins between 0 and 100 to aid our analysis.

# Obstacles Faced

One of the first obstacles our group faced was with our original dataset that we were going to use in our proposal. The file was a .data file with difficult formatting that made it impossible to convert into a .csv or read into R. We found the IPEDS dataset, which was much more compatible with R and Excel and decided to use it as our data source for this project.

Another obstacle we ran into was when we were trying to create our Choropleth Maps in R. In our dataset we had a column titled “Region”, which indicated the region within the U.S. where the college was located. When we tried running the state\_choropleth command in R, we experienced some difficulty getting the command to return the data by state because of the alternative column named Region. Once we changed the name of the State column to “region”, we were able to get the results that we wanted.

Additionally, we had some initial difficulty cleaning our dataset and deciding which variables to keep for our analysis. With 250 variables originally, we had to go through each one and determine its importance and relevance to the objective of our analysis. We went through the dataset several times in order to reduce the number of columns that we wanted to work with. We also had to decide which colleges to keep in our dataset since the data covered 2-year, 4-year, and doctoral schools. We ended up deciding to just include 4-year colleges by filtering out the remaining ones.

# Function Documentation

The function we used for this project was designed to help us examine the graduation rates by state. We wanted to be able to see each state’s distribution of graduation rates to more easily compare one another. To use the function, you simply input the state name that you would like to examine and the function returns a bar graph of the state’s binned graduation rates. The code is as follows:

graduation.by.state <- function(state = "Washington") {

tempdf <- subset(df, State == state)

temp <- qplot(graduation\_rate\_binned, data = tempdf, geom = "bar",

main = paste("Graduation Rates in", state, sep = " "),

xlab = "Graduation Rate Percentage")

temp

}

graduation.by.state()

Below are a few examples of our function using Washington and Tennessee as the inputs. As you can see in Figures 9 and 10, the visualizations show how many colleges are within the specified graduation rate bins.

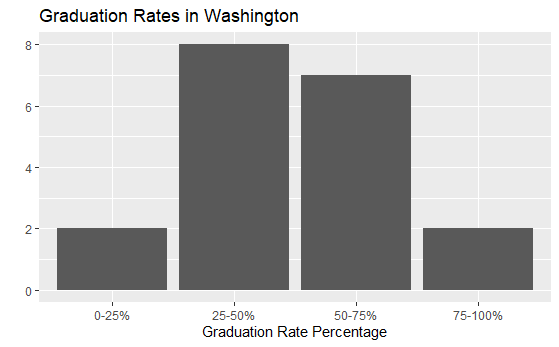


Figure 9

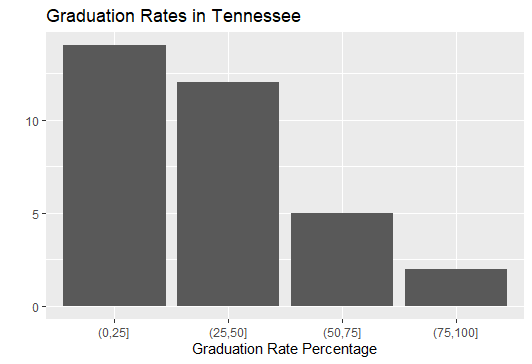


Figure 10

# Appendix

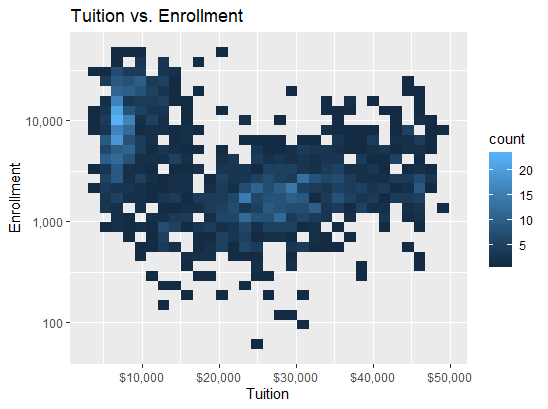


Figure 11

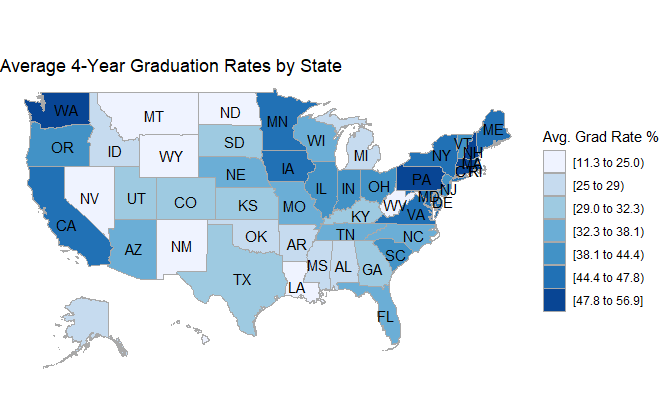


Figure 12

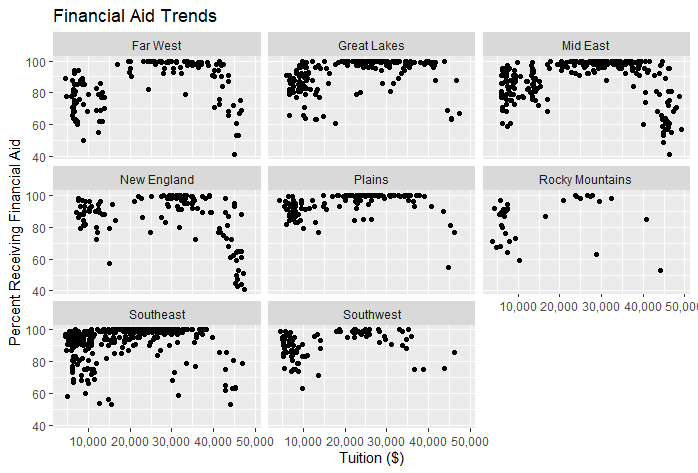


Figure 13

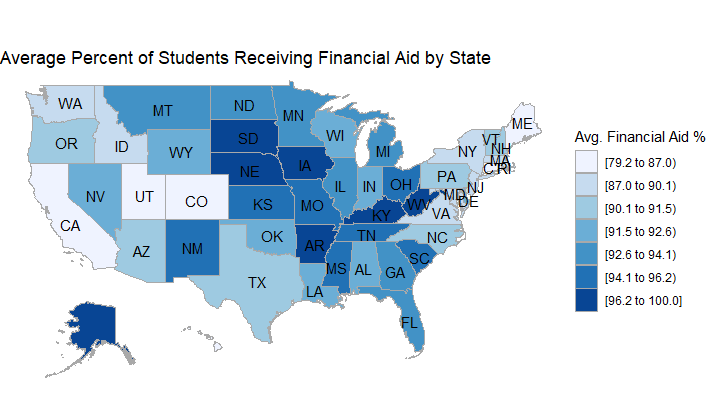


Figure 14