Analysis on the Educational System in the US

FINAL PROJECT CONTENT

Gold Okorie, Silva Kelly, Masvikeni Tamara| ADTA5130 | Dec 8th 2021

# Business Understanding

The US department of Education has a list of college score cards that provides data that helps prospective students and their advocates find the right fit for postsecondary education. The college score card dataset allows users to search and compare colleges based on a highlighted set of elements such as graduation rates, in-state and out of state tuition rate etc. for both public and private universities. The goals that will be achieved using the dataset are as follows:

* To provide correlation analysis to observe whether a relationship exists between graduation rate and several variables in the data such as expertise of the faculty and all the costs associated with attending a private or public university.
* Provide information to the Governor of the Midwestern region on whether the graduation rate and the in-state tuition in the Midwestern region is representative of the whole United states
* Provide analysis on the proportion of public universities in the Midwestern region
* Provide analysis that investigates whether there is a difference in the average graduation rate among the different regions in the United States.

This study will execute the above analyses tasks by exploring our data and using statistical tools such as Confidence Interval and ANOVA.

# Data Understanding

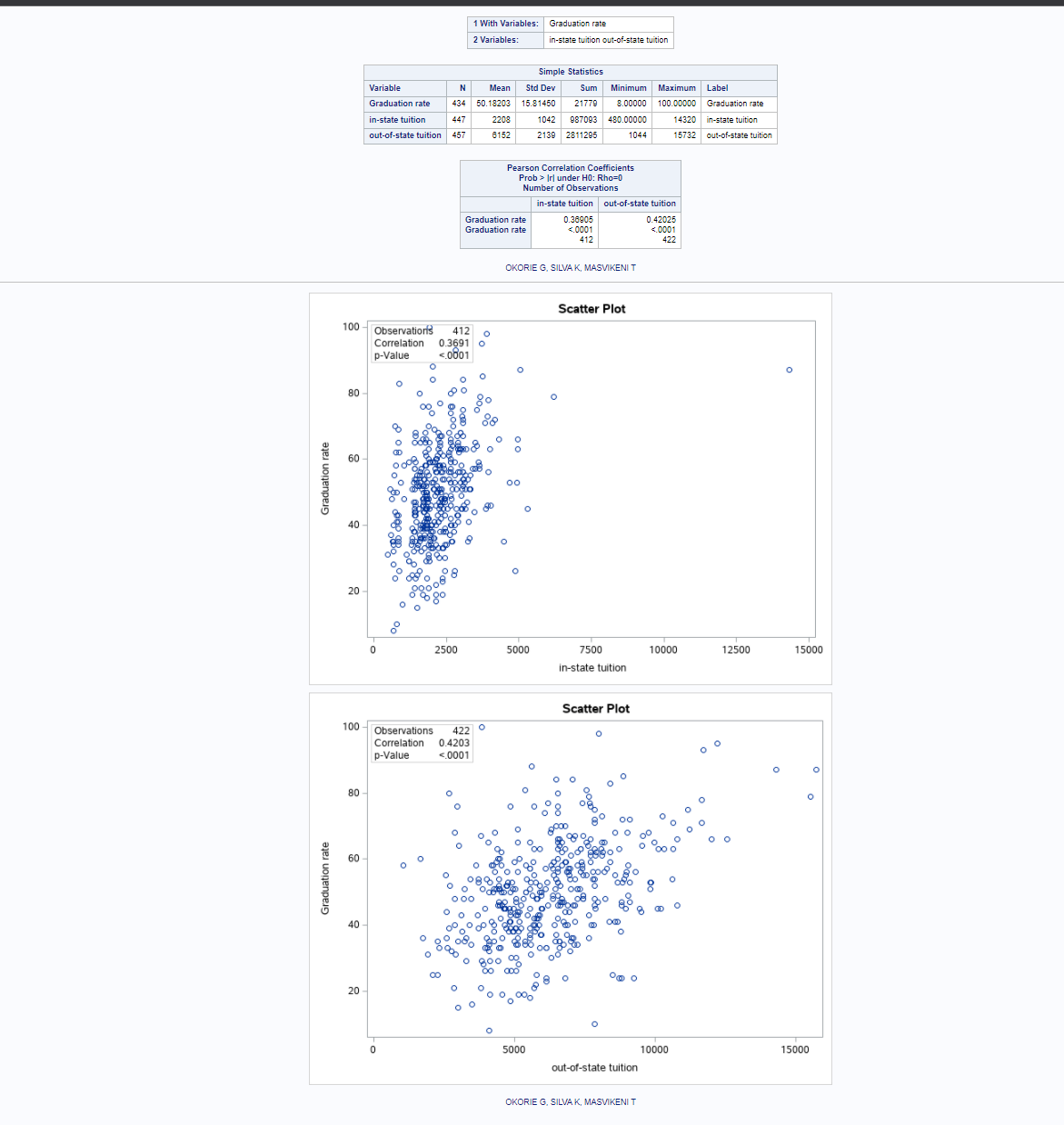
The dataset used was retrieved from the US department of Education and contains the following variables:

* College name
* Regions (new column added to the dataset)
* State
* Public (1)/Private (2)
* Number of Applications received
* Number of Applicants accepted
* Number of students enrolled
* Percentage of new student from top 10%
* Percentage of new student from top 25%
* Number of Full-time undergraduates
* Number of Part time undergraduates
* In-state tuition
* Out-of-state tuition
* Room
* Board, additional fees
* Estimated book cost
* Estimated personal allowance
* Percentage of faculty without PHD
* Student to faculty ratio
* Graduation rate.

The main variables we will be using for our analysis based on the goals we are stated earlier in the Business understanding section are; graduation rates, public or private, in-state tuition and out of state tuition and region.

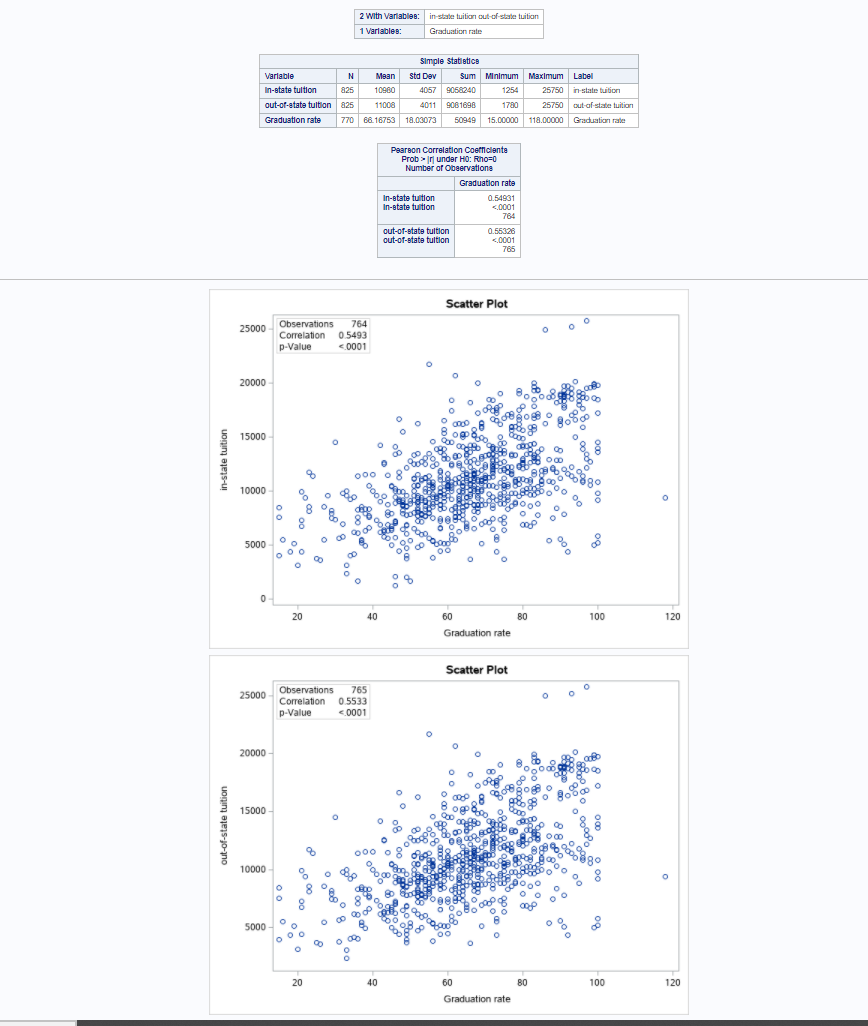
**CORRELATION ANALYSIS**

Our research question was formulated to test whether there is a statistically significant linear relationship between two continuous variables, graduation rate, and tuition fees for public and private universities. We used a bivariate Pearson correlation to test whether there is a statistically significant linear relationship between tuition and graduation rates and to determine the strength and direction of the association

****

The screenshot above shows the correlation for between graduation rates and in-state and out-of state tuition rates for public universities. The first correlation between graduation rate and the in-state tuition is 0.36.

The second plot indicates a correlation between the out-of-state tuition fees is 0.42. The direction of the relationship is positive, meaning that these variables tend to increase together.



The screenshot above shows the correlation for between graduation rates and in-state and out-of state tuition rates for private universities. The first correlation between graduation rate and the in-state tuition is 0.5436 which is a positive relationship between

The second plot indicates a correlation between the out-of-state tuition fees is 0.533. The direction of the relationship is positive, meaning that these variables tend to increase together.

**DESCRIPTIVE ANALYSIS OF DATASET:**

Analysis of Graduation Rate in the United States:

# Chart, box and whisker chart Description automatically generated

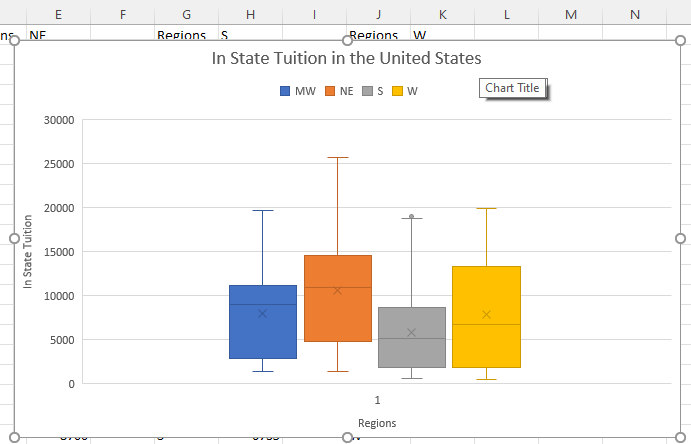
Shown above is an image of box plot representations of the graduation rates of the public and private universities for the four major regions in the United States. The box plots display the variance of the data, the values for their means, maximum values, minimum values, and interquartile ranges. The plots show the Northeast Region with the highest values for the average graduation rate, maximum value, and minimum value.

Graphical user interface, application, table, Excel

Description automatically generated

Shown above is a bar chart displaying the average graduation rate the four regions of the United States for both public and private universities.

Shown above is a bar chart displaying the average graduation rate for the four regions of the United States for both public and private Universities.



Shown above is an image of box plot representations of the in-state tuition of the public and private universities for the four major regions in the United States. The box plots display the variance of the data, the values for their means, maximum values, minimum values, and interquartile ranges. The plots show the Northeast Region with the highest values for the average graduation rate, maximum value, and minimum value.

**Strengths and Weaknesses of the data**

A clear strength in this data is the number of cases and variables. Our data has 1303 cases in total and 20 variables as mentioned before. This gives room for greater options for statistical tools that can be implemented on the data to answer business questions.

A disadvantage of our data is that it has 2002 missing values. These missing values result in reduced confidence that our data is fully representative of the entire population.

# Data Preparation

We performed our data wrangling and preparation using the Excel software tool. The dataset consisted of missing values but we didn’t replace the values with any other value to avoid distorting our analysis. The dataset was free from any duplicate values, typos or errors. However, we added a new column to our dataset called ‘Region’ that indicates region of the Universities located in the United state.

# Modeling

Confidence Interval

Confidence Interval is a useful statistical tool that allows us to see the likely range of values of our population mean. The Governors in the Mid-West want to know whether the universities in their region are representative of all universities in the United States. As their data analysts, we have decided to use the Graduation rate and the in-state tuition for comparison. The Governors also believe that most of the universities in the Mid-West are public universities and have estimated it to be about 55% of their universities. This hypothesis is their rationale that funding does not need to be poured into public universities and they have asked the data scientists to confirm this.

Whenever there is a value that represents the true parameter, the confidence interval provides a range of values in which it is contained. This can be calculated with statistical tools. Confidence interval is useful because it provides evidence of statistical significance and provides a direction of the element being studied. Du Prel et. al., (2009) A disadvantage of confidence interval is the common misconception that are associated with its interpretation. The biggest misconception when confidence intervals is interpreted is that the percentage given is representing the percentage of data that falls in the bounds given. For example, a 99% percent interval does not mean that 99% of the data is contained in the given bounds. What it actually means is that one can be 99% certain that the range will contain the population mean. Hayes (2018)

The following are the analytical tasks and questions that were formulated to answer our business questions:

1. Summary statistics including the mean, standard deviation, standard error, number of   
   observations and 95% confidence interval for Graduation rate, In-state tuition, and faculty without PhD rate for all the Universities in the entire United states.
2. Hypothesis comparing the average Graduation rate in the Mid-west universities to that of the rest of the United States.
3. Based on average Graduation Rate, could the Governors be 95% confident that the average Graduation rate for universities in the Mid-west is representative of the graduation rate for all the universities in the United States?
4. Hypothesis comparing the average In-state tuition in the Mid-west universities to that of the rest of the United States.
5. Based on average In State tuition, could the Governors be 95% confident that the average In-state tuition for universities in the Mid-west is representative of the In-state tuition for all the universities in the United States?
6. The Governors believe that 55% of Universities in the Mid-West are public universities. What is the hypothesis for this claim?
7. Constructing a 95% confidence level for the proportion of public universities in the Mid-West.
8. What is proportion of public universities in the Mid-west?
9. Using the Wald interval to determine the 95% confidence interval for the proportion of public universities in the Midwest.
10. Our conclusion about the Governors in the Midwest’s claim that 55% of the universities in the Mid-West are public universities
11. Based on the results of the analysis, Our recommendations for the states in the Mid-West about their graduation rates, their In-State tuition and the proportion of public universities.

**The Results**

1. **Summary statistics including the mean, median, standard deviation, maximum value, minimum value, number of observations, number of missing values and 95% confidence interval for Graduation rate and In-state tuition for all the Universities in the entire United states.**

Graphical user interface, text, application, Word

Description automatically generated

1. **Hypothesis comparing the average Graduation rate in the Mid-west universities to the rest of the United States.**

H0: μmw = μus the average Graduation rate of universities in the Mid-West is the same as the average Graduation rate for universities in the entire United States.

H1: μmw ≠ μus the average Graduation rate of universities in the Mid-West is different from the average Graduation rate for universities in the entire United States.

1. **Based on average Graduation Rate in the Mid-West, could the Governors be 95% confident that the average Graduation rate for universities in the Mid-west is representative of the graduation rate for all the universities in the United States?**

Graphical user interface, application, Word

Description automatically generated

The above image shows the 95% confidence interval for the Graduation Rate for all the Universities in the U. S., the number of observations and the number of missing values.

Graphical user interface, text, application, Word

Description automatically generated

The above image shows the mean, standard deviation, and number of observations for the Graduation Rate for universities in the Mid-West.The above results show that the universities in the Mid-West show a mean Graduation rate of 60.9542683. The mean is within the confidence limits of the Graduation rate of all the universities in the United States. We fail to reject the null hypothesis. This means that the Governors can be 95% confident that the Graduation Rate in the Mid-West is representative of the Graduation rate in the United States.

1. **Hypothesis comparing the average In-state tuition in the Mid-west universities to that of the rest of the United States.**

H0: μmw = μus (the average in state tuition for universities in the Mid-West is equal to the in-state tuition in the whole United States)

H1: μmw ≠ μus (the average in state tuition for universities in the Mid-West is different from the in-state tuition in the whole United States)

1. **Based on average In State tuition, could the Governors be 95% confident that the average in state tuition for universities in the Mid-west is representative of the in-state tuition for all the universities in the United States?**

Graphical user interface, text, application, Word

Description automatically generated

The above image shows the total number of observations (N) for universities in the United States, the number of missing values (N Miss) and the 95% confidence limit.

Graphical user interface, application, Word

Description automatically generated

The average In-State tuition for universities in the Mid-West is shown in the above image, as well as the standard deviation (Std Dev), total number of observations (N), number of missing values (N Miss).

The average in-state tuition for Universities in the Mid-West shown above **8012.43**, falls within the confidence limits of the in-state tuition for the universities of the entire US with lower limit **7603.09** and upper limit **8191.46**. This shows that the in-state tuition in the Mid-West is representative of the in-state tuition in the United States. We fail to reject the null hypothesis. Therefore, the Governors can be 95% confident that the in-state tuition in the Mid-West is representative of the in-state tuition in the entire U.S.

1. **The Governors believe that 55% of Universities in the Mid-West are public universities. What is the null hypothesis and alternate hypothesis for this claim?**

H0: p= 0.55 (55% of universities in the Mid-West are public universities)

H1: p≠ 0.55 (the proportion of public universities in the Mid-West is not 55%)

1. **Constructing a 95% confidence level for the proportion of public universities in the Mid-West.**
2. **What is the proportion of public universities in the Mid-west?**
3. **Using the Wald interval to determine the 95% confidence interval for the proportion of public universities.**
4. **Our conclusion about the Governors in the Mid West’s claim that 55% of the universities in the Mid-West are public universities.**

Graphical user interface, application

Description automatically generated

The results above show that the proportion of Public Universities is **32.10%** and that of Private Universities is **67.90%.** The results also show the proportions represented in the form of the bar chart shown above.

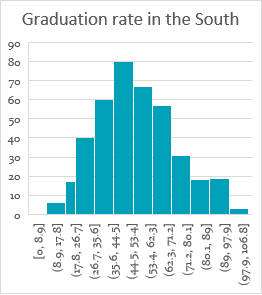
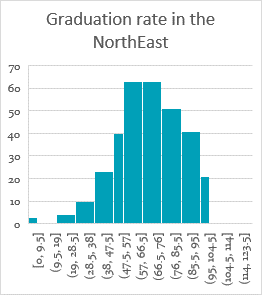
Graphical user interface, application, Word

Description automatically generated

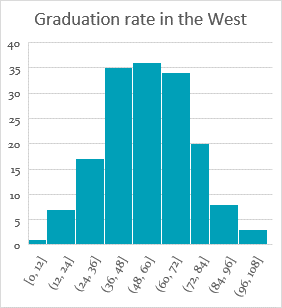
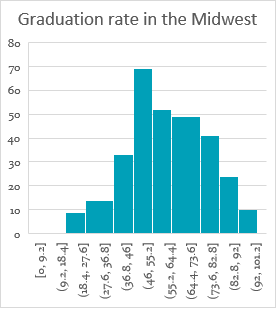
The above image shows the Proportion of Public Universities in the Mid-West as **0.3210.** According to Wald we are 95% confident that the proportion of public universities in the Mid-west falls between **0.2723** and **0. 3698**.We can conclude that there is insufficient evidence to support the Governor’s claim that 55% of universities in the Mid-west are public universities. We therefore reject the null hypothesis in favor of the alternate hypothesis.

**ANOVA model analysis**

We are using the ANOVA modeling tool to answer the question of whether there is a significant difference between the average graduation rates in the different regions of the United States. This is crucial information for the United States department of Education to have because it can lead to further investigation to assess why a difference exists if such a difference were to be found. ANOVA is an appropriate statistical tool that allows the analysis of multiple means. This is a case in which ANOVA is relevant for us to answer these questions in our analysis. Before running the ANOVA analysis, we ensured that the ANOVA analysis was appropriate by checking for independence, normal distributions in each region or parameter and by checking if the data had comparable variances by assessing box plots as shown in the images below:

Histogram for the Southern region Histogram for the Northeastern Region

Histogram diagram for the West Histogram diagram for the Midwest

The histograms display normal distributions for the graduation rate in the four regions.

Chart, box and whisker chart

Description automatically generated

The box plots above show that the variances for the graduation rate in all four regions are comparable.

The graduation rates for each region are independent.

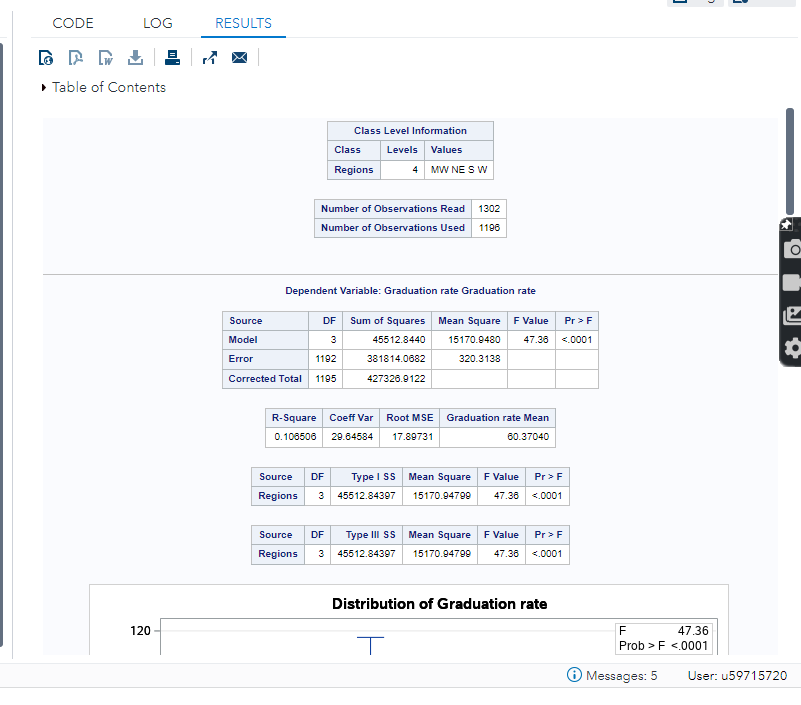
With the conditions for ANOVA satisfied we then ran our analysis and incurred the following results:

ANOVA – graduation rate with regions

The following is the hypothesis:

HO: The mean graduation rates are equal for all the regions (𝜇MW= 𝜇W= 𝜇S= 𝜇NE)   
HA: There is a difference in the graduation rate among the different regions.

We got the following result after using statistical tools in SAS:



Since the value for our F critical value for our model with α = 0.05 is **2.61** and the calculated F value is 47.36 which is greater than the critical value, we reject null hypothesis and conclude that there is a significant difference in the average graduation rate among the regions.

ANOVA – graduation rate and region with public and private as a blocking variable

Graphical user interface, application, table

Description automatically generated

HO: The mean graduation rates are equal for public vs private universities (𝜇PUB= 𝜇𝜇PVT)  
HA: the mean graduation rate is different

The F crit value with α = 0.05 is 3.85 is less than the F calculated value of 208.08 we reject the null hypothesis and conclude that the mean graduation rate between public and private universities is different

&

HO: The mean graduation rates are equal for all the regions (𝜇MW= 𝜇W= 𝜇S= 𝜇NE)   
HA: There is a difference in the graduation rate among the different regions.

Since the F crit value with α = 0.05 is 2.61 which is less than the F calc value of 38.34. We reject the null hypothesis and conclude that the mean graduation rate among the different regions is different.

# Evaluation

Our recommendation for the Governors based on all the results is that they can be 95% confident in that the graduation rate and in-state tuition in the Mid-west is representative of the entire United states. They can therefore use the Mid-west as a representation for the rest of the United States based on those parameters and limits. We also recommend that the Governors put more funding into the building of public universities because their assumption that most of the universities in the Mid-west are public was not proven by the statistics. The proportion was found to be 32.10% which is much lower than their prediction. There are weaknesses in the data however, because there are missing values. The missing values in the graduation rate and in-state tuition could have greatly impacted the outcome of the results.

The results from the ANOVA analysis show that there is a difference in the graduation rates among the four regions of the United States and there is also a difference in the average graduation rate between public and private universities. This information is useful for the United States department of Education because it can encourage further investigation. Our recommendation to them would be to conduct further analysis to find out why certain regions have higher graduation rates than others and to also find out why public and private institutions also have statistically significant differences. Could it be due to other variables such as the expertise of the faculty, the resources in the school, the previous graduation rates in previous years in the different regions and in public vs private universities? The difference in graduation rates for the different regions and in public and private universities could also be impacted by the other variables in the data set such as the number of students accepted from the top 10% or 25% of students. This would need further analysis.

References:

Du Prel, J. B., Hommel, G., Röhrig, B., & Blettner, M. (2009). Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. Deutsches Arzteblatt international, 106(19), 335–339. https://doi.org/10.3238/arztebl.2009.0335

Hayes A. (2018) Tools for Fundamental Analysis: Confidence Interval. https://www.investopedia.com/terms/c/confidenceinterval.asp