Final Project Prospectus

Title: The relationship between the number of international students and COVID cases in states
Notice: Dr. Bryan Runck
Author: Liang-Ting Chen
Date: May 03, 2021

Project Repository:

https://github.com/chen6761/GIS5572

Abstract

Due to the COVID-19, lots of schools decide to practice distance learning instead of in-person classes to help to reduce the risk of spreading the virus. Because of the decision, many international students go back to their country and continue their education via remoting classes. According to the article from Migration Policy Institute, the total number of international students of the fall semester in 2020 decreased 16 percent from 2019. (Batalova, 2021) This project shows the relationship between the number changing of international students and the COVID cases in states.

Problem Statement

In the project, I planned to see if the COVID cases affect the number of international students in each state. Does the higher ratio of COVID cases cause the more obvious decrease of international students?

Table 1. Required data

#	Requirement	Defined As	Spatial Data	Attribute Data	Dataset	Preparation
1	Number of international students	Raw input data of numbers		Numbers/ Country	DHS	
2	States boundaries	Raw input of state boundaries	Boundary geometry		Census Bureau	
3	United States COVID-19 Cases and Deaths by State	Raw input of COVID-19 cases by state		Total cases	CDC	

Input Data

I used the Student and Exchange Visitor Information System (SEVIS) data from the U.S. Department of Homeland Security (DHS) to calculate the number of international students. I used the student amounts in March 2019 and September 2020. The reason that I chose these two times was that I wanted to compare the difference from non-Covid cases to the end of the 2019/20 academic year. About the state boundary and COVID-19 data, I used the one in the census bureau and CDC.

Table 2. Input data

#	Title	Purpose in Analysis	Link to Source
1	March 2019 SEVIS Data Mapping Tool Data	Raw input dataset for the number of international students in March 2019	DHS
2	September 2020 SEVIS Data Mapping Tool Data	Raw input dataset for the number of international students in September 2020	DHS
3	States boundaries	Raw input of state boundaries	Census Bureau
4	United States COVID-19 Cases and Deaths by State	Raw input of COVID-19 cases by state	CDC

Methods

I made two choropleth maps first, one for the changing ratio of international students, the other for the COVID-19 accumulative cases on August 31st, 2020. Both contents were presented based on state boundaries. Then, I had the rank for each state in both data and found the relationship to see whether the higher number of COVID-19 cases caused the fewer international students to come or not.

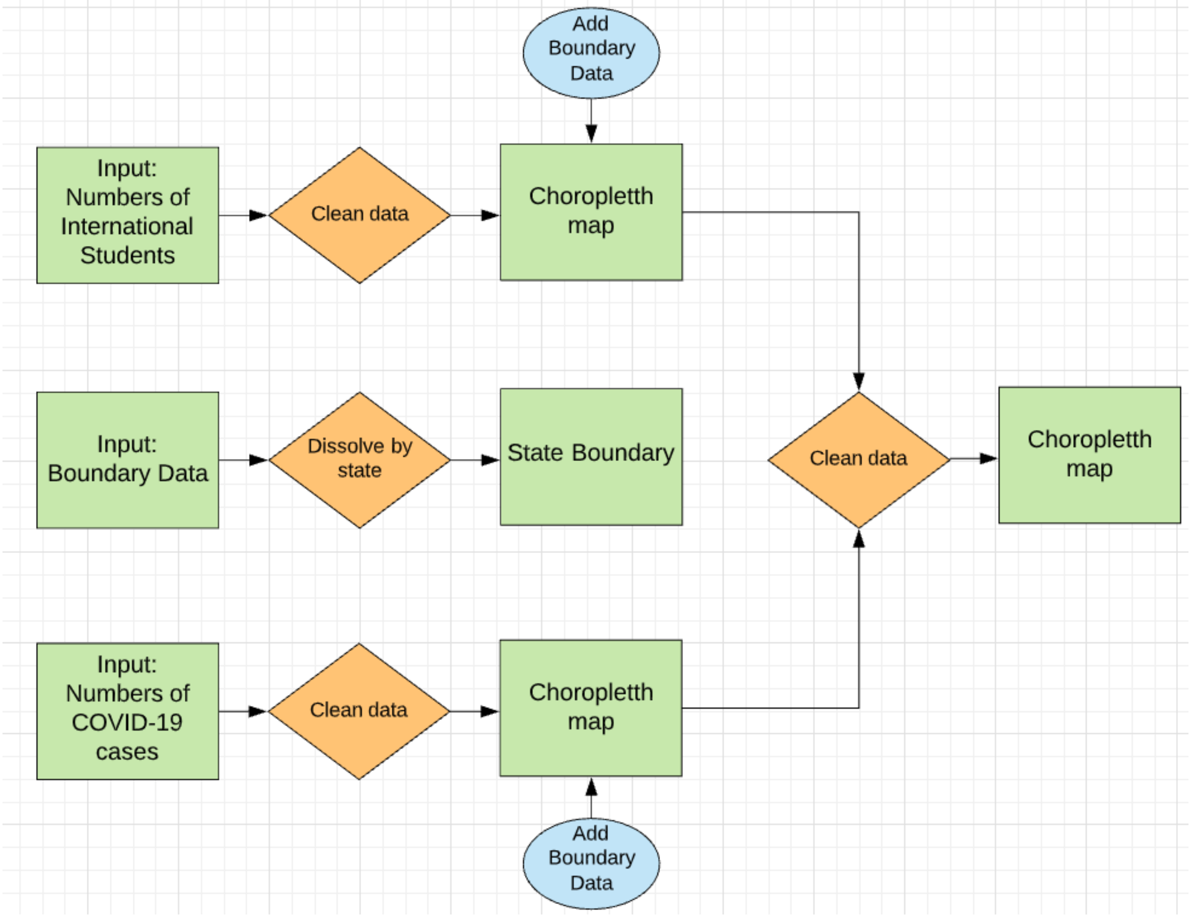


Figure 1. Data flow diagram.

Results

Choropleth Map

Here was the first choropleth map. (Figure 2) In this map, I presented the covid-19 accumulative cases until September 2020. I used the nature break method to separate the data into five classes. As we can see, the highest class includes California, Texas, Florida, and New York states. The data here is showed by the number of cases.

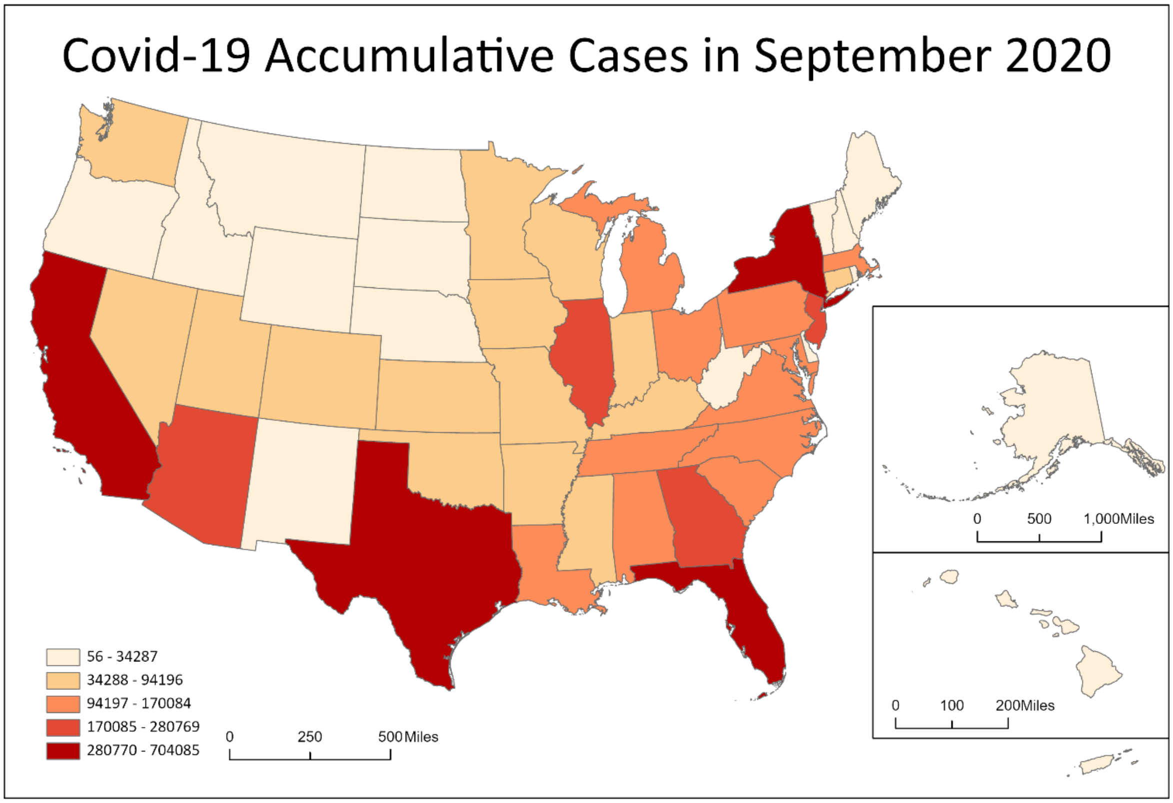


Figure 2. Covid-19 Accumulative Cases in September 2020

The second choropleth map presented the variation of students between March 2019 and September 2020. (Figure 3) I also used the nature break method to separate the student amount into five classes. The highest class has California and New York states. The data here is also shown by the number of students' variation.

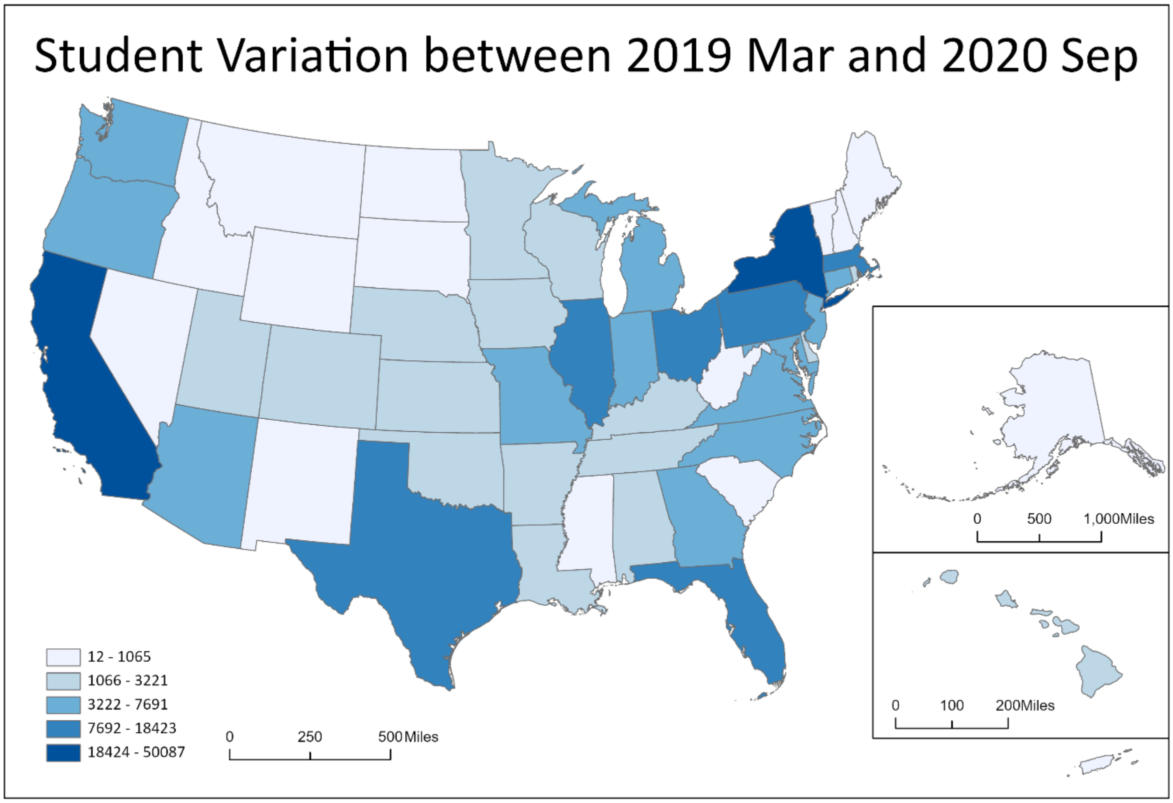


Figure 3. Student Variation between 2019 Mar and 2020 Sep

Then, I ranked the states by the previous two data and found which state in the same ranking. The result shows this choropleth map. (Figure 4) There were only five states in the same rank. However, I thought it was not quite a precise result. Bryan gave me some advice about considering the state population and do the linear regression to find the relationship.

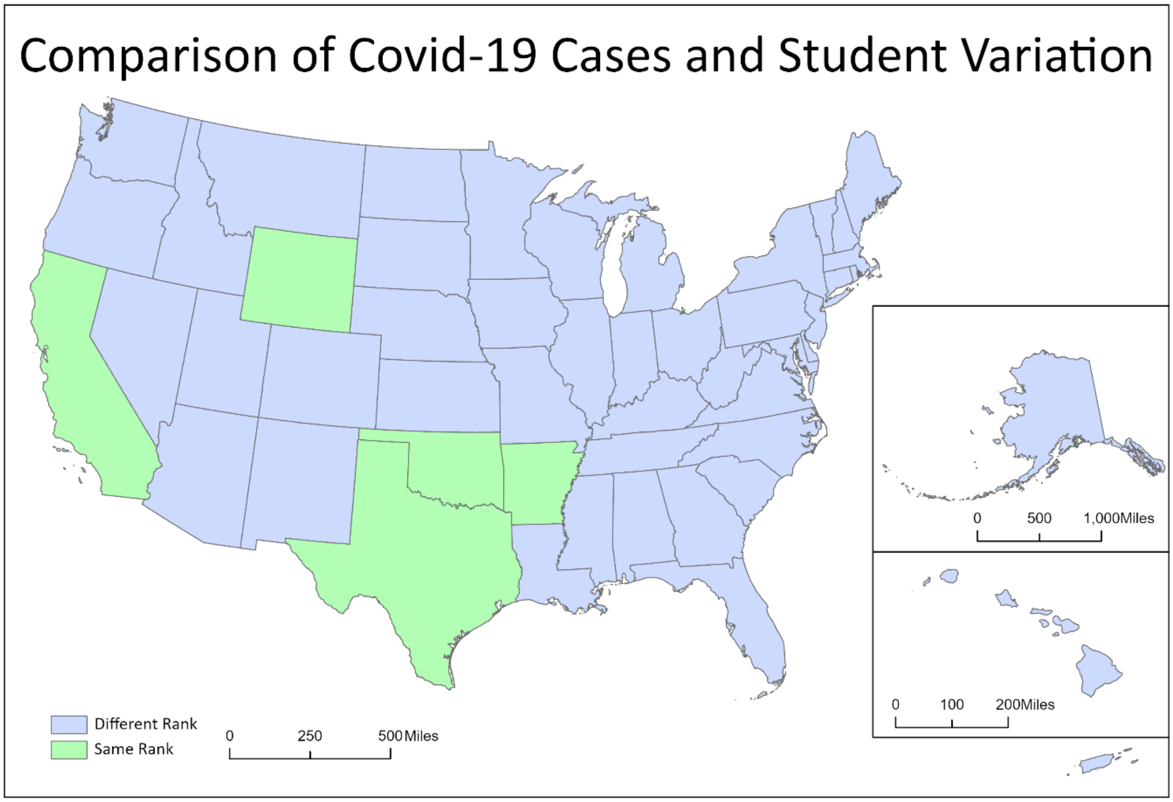


Figure 4. Comparison of Covid-19 Cases and Student Variation

Here was the map of the updated covid-19 accumulative cases, the difference was that I presented it by dividing the cases by state population. (Figure 5) So, the distribution of five classes changed. The highest class includes Arizona, Louisiana, Mississippi, Alabama, Georgia, South Carolina, and the Florida States.

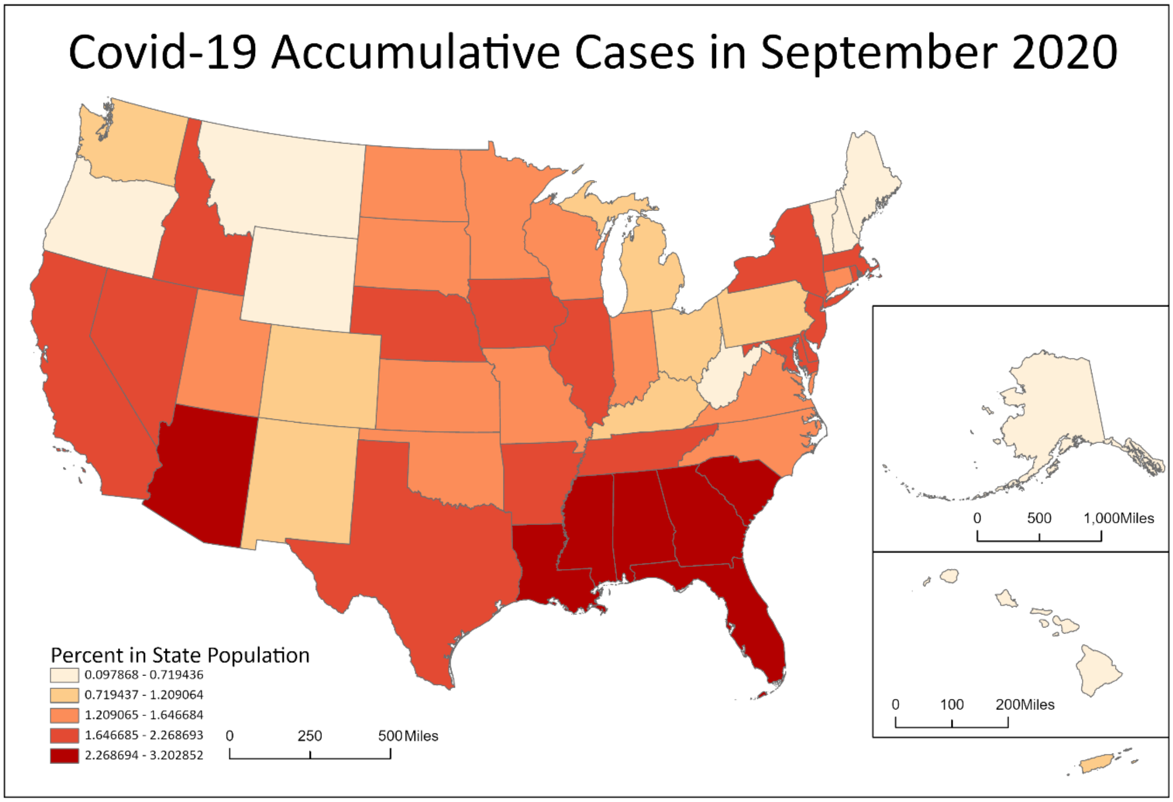


Figure 5. Covid-19 Accumulative Cases in September 2020 in percent

This student variation choropleth map was also updated considering the state population. (Figure 6) New York, Massachusetts State, and DC were in the highest class that the percent of international students dropped down the most.

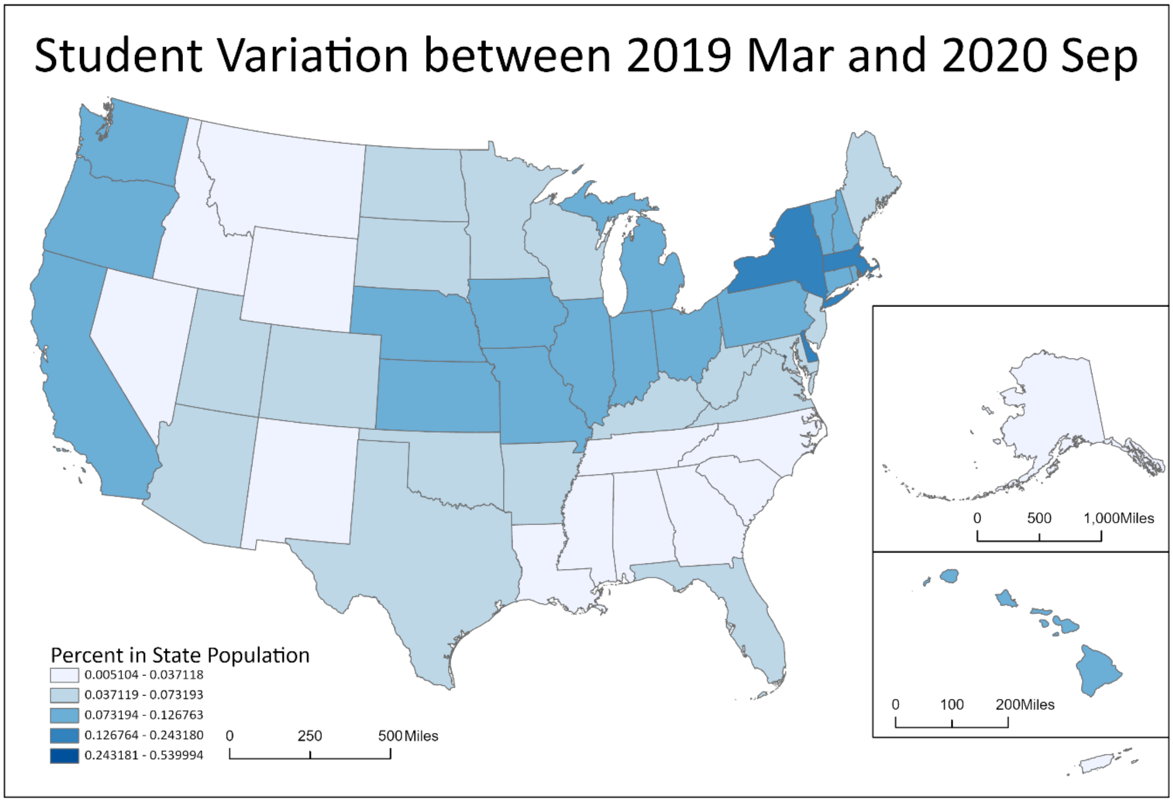


Figure 6. Student Variation between 2019 Mar and 2020 September in percent

<u>Generalized Linear Regression (GLR)</u>

To more understand this analysis, I googled it and realized the R-squared value could present that the variables existing a significant relationship or not. (Minitab Blog Editor, 2013) I used case percent in the state as the dependent variable and student number variation percent in the state as the explanatory variable to get the first analysis. (Figure 7) The relationship between them showed a low R squared value which meant the model does not fit in goodness.

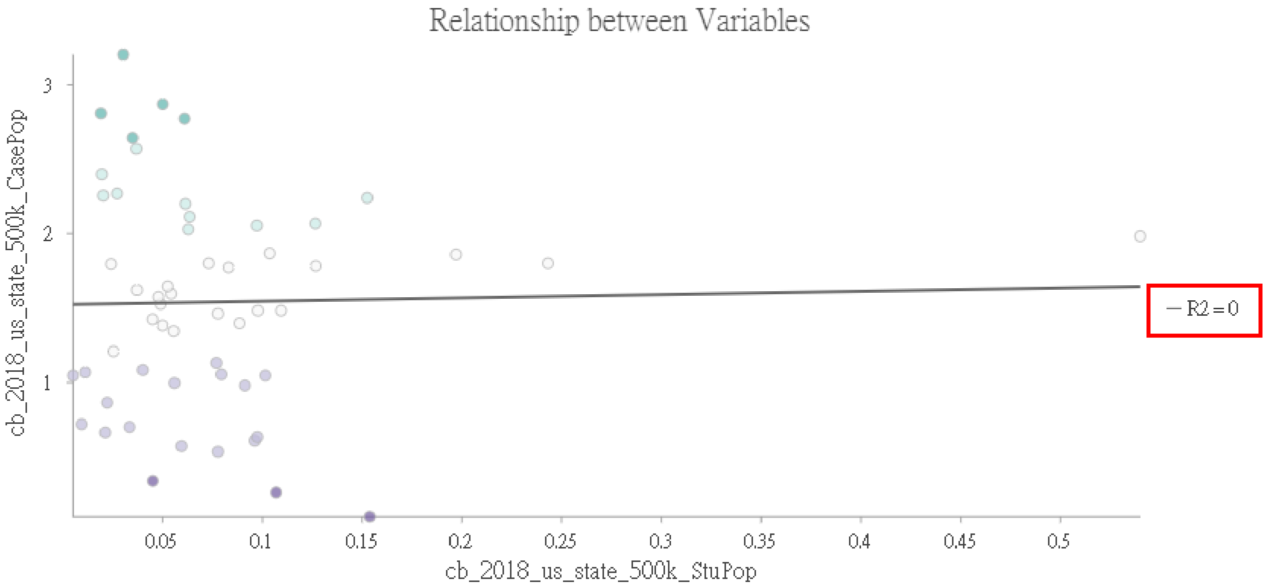


Figure 7

I converted two variables to check the result and found it was still in the low R squared value. (Figure 8) I thought it meant that there did not exist any significant relationship between them.

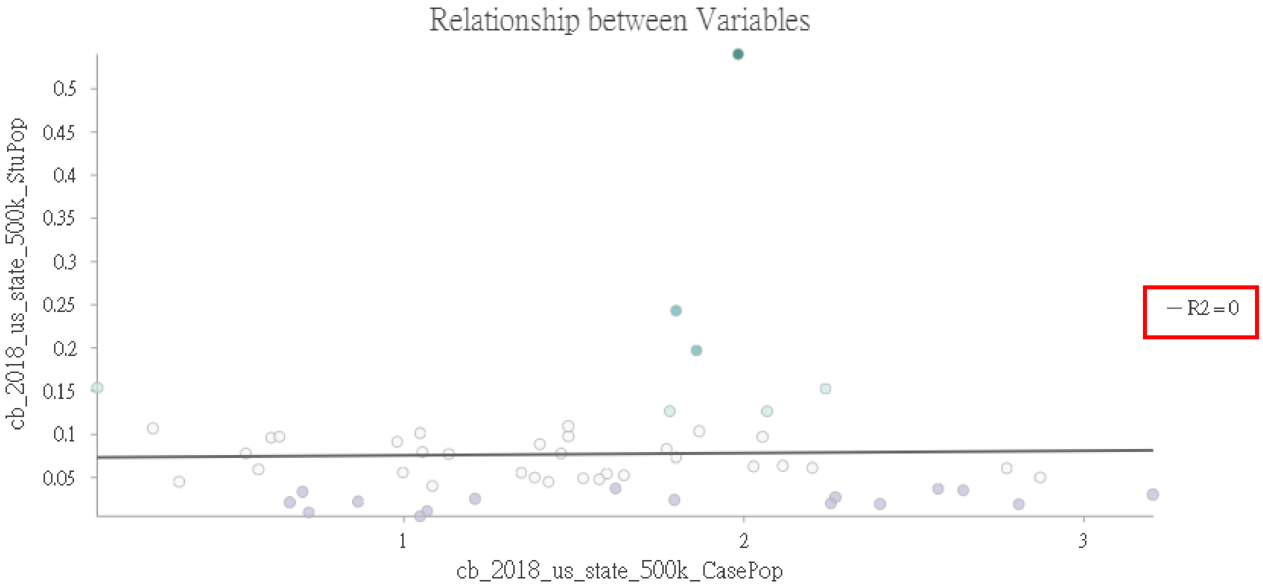
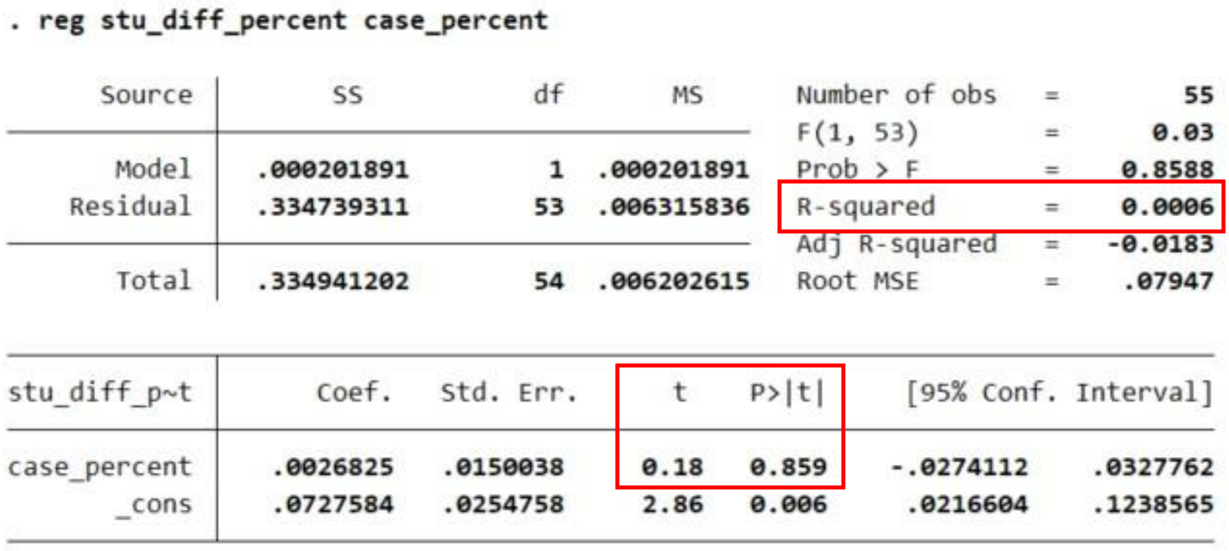


Figure 8

Results Verification

To double-check the result. My friend helped me to do the regression analysis in Stata to see the correlation between case percent in state and student number variation percent in the state. (Figure 9) I found that the t value was low, and the p-value was high, which implied there was no significant effect of the proportion of COVID cases on the proportion of student number variation.



```
. reg stu_diff_percent case_percent

      Source |       SS           df       MS      Number of obs   =        55
-------------+----------------------------------   F(1, 53)        =      0.03
       Model |  .000201891         1  .000201891   Prob > F        =    0.8588
    Residual |  .334739311        53  .006315836   R-squared       =    0.0006
-------------+----------------------------------   Adj R-squared   =   -0.0183
       Total |  .334941202        54  .006202615   Root MSE        =    .07947

------------------------------------------------------------------------------
stu_diff_p~t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
case_percent |   .0026825   .0150038     0.18   0.859    -.0274112    .0327762
       _cons |   .0727584   .0254758     2.86   0.006     .0216604    .1238565
------------------------------------------------------------------------------
```

Figure 9

Then, I used student number in 2019 March as the dependent variable and student number in 2020 September as the explanatory variable. I found the R squared was almost 1. (Figure 10)

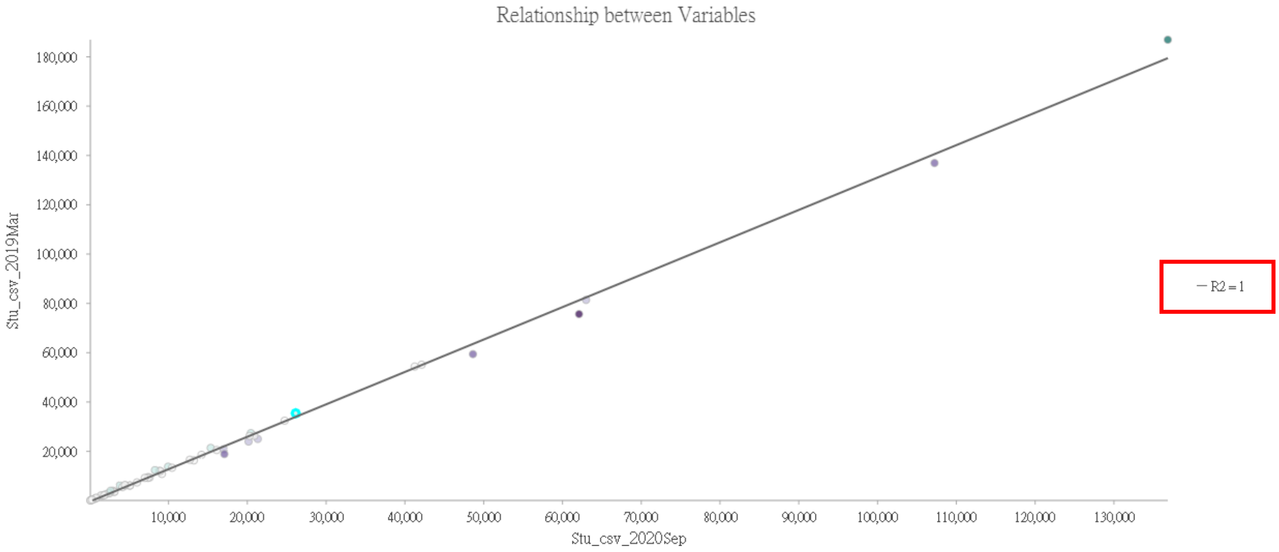
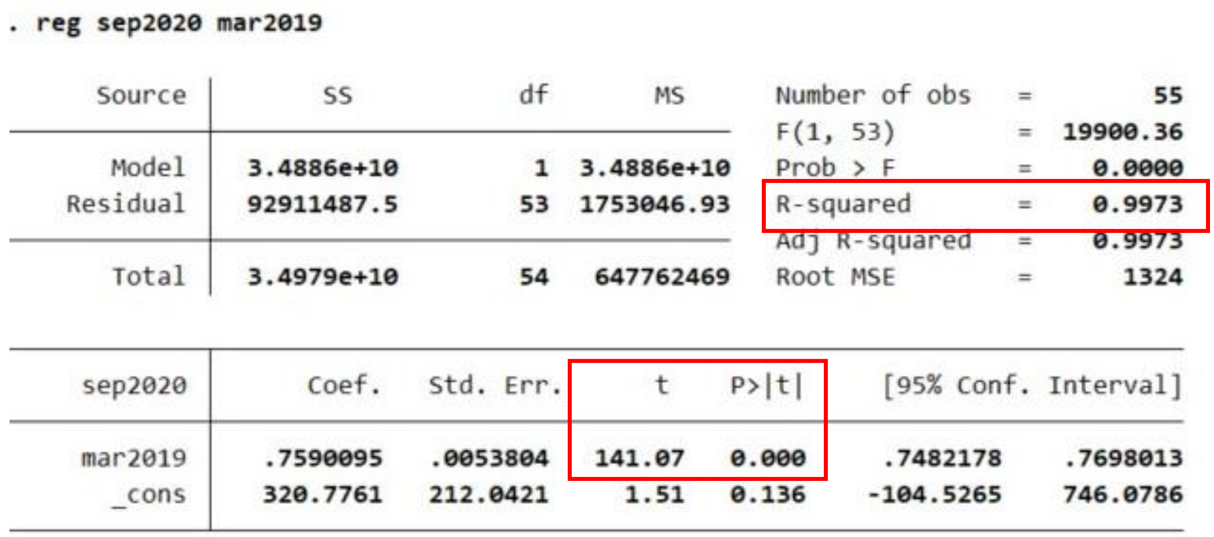


Figure 10

Executing the same variables in the Stata, the t value was high and the p-value was low. (Figure 11) I think it means there is a difference in the number of students from one year to the next, and some factors affect the amount of student enrollment. However, as I mentioned before, there is no relationship and no clustering between covid case and student number variation. This suggests that covid didn't have an impact on enrollment. To find what factor related to the change, I tried to analyze the state population and student number variation.



```
. reg sep2020 mar2019

      Source |       SS           df       MS      Number of obs   =        55
-------------+----------------------------------   F(1, 53)        =  19900.36
       Model |  3.4886e+10         1  3.4886e+10   Prob > F        =    0.0000
    Residual |  92911487.5        53  1753046.93   R-squared       =    0.9973
-------------+----------------------------------   Adj R-squared   =    0.9973
       Total |  3.4979e+10        54   647762469   Root MSE        =      1324

------------------------------------------------------------------------------
     sep2020 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mar2019 |   .7590095   .0053804   141.07   0.000     .7482178    .7698013
       _cons |   320.7761   212.0421     1.51   0.136    -104.5265    746.0786
------------------------------------------------------------------------------
```

Figure 11

I found that if I set the student number variation as the dependent variable and state population as the explanatory variable, the R squared became high to 0.81. (Figure 12) I also got the same result after I converted two variables. (Figure 13)

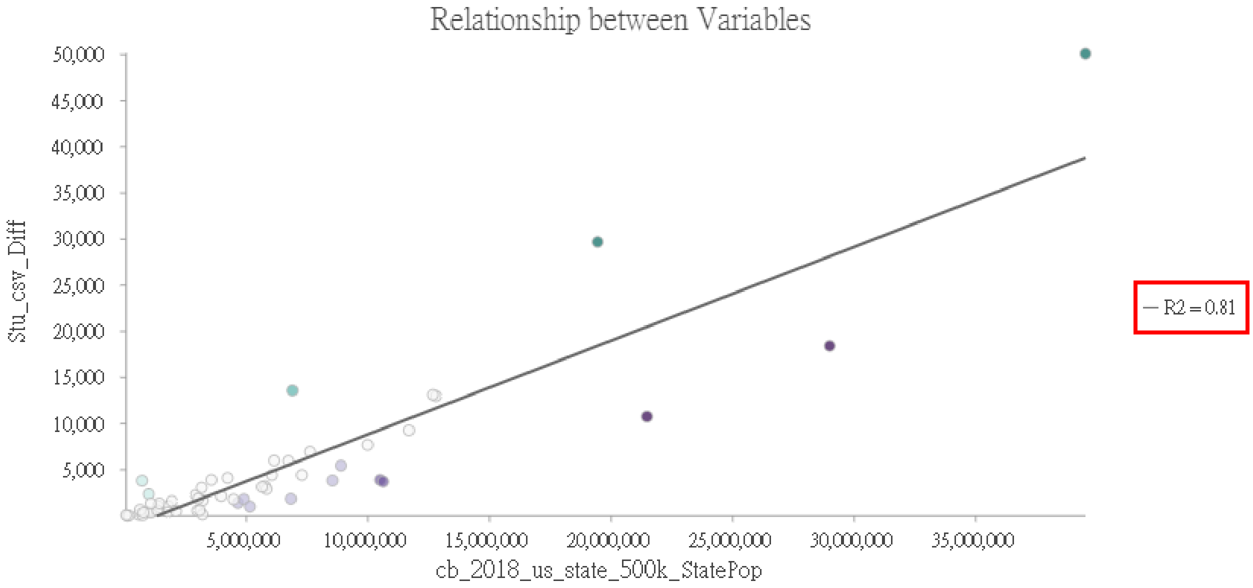


Figure 12

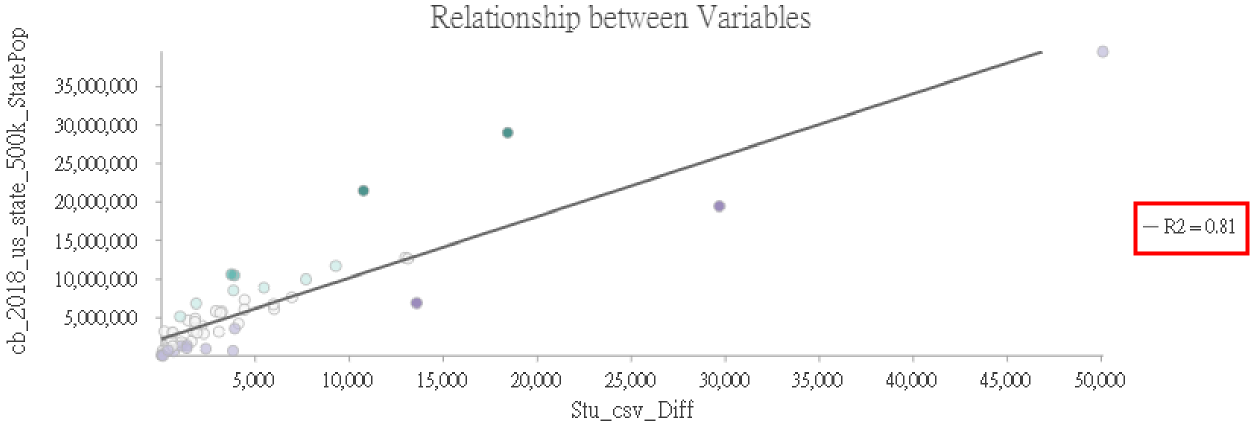


Figure 13

Discussion and Conclusion

I think there is no relationship and no clustering which means Covid didn't have an impact on enrollment. Although there is a difference in the number of students from one year to the next, the result does not show spatial variation. I think there might be two reasons. The first reason is that the covid case does not increase proportionally to the state population through the states. The first case that appeared in each state was at a different time, some states got in pandemics earlier and caused the students to decrease earlier. But this does not mean the state includes more infected covid cases. The second reason is that according to the different population densities and government orders, the infected speed and range were all different. The student ratios are also different in each state, so the decreasing percent will not correspond to the increasing case percent. In the end, I think population might be a factor to impact the student variation because a higher population normally includes more students in an area.

References

Batalova, J. B. E. I. A. J. (2021, February 2). *International Students in the United States*. Migrationpolicy.Org. https://www.migrationpolicy.org/article/international-students-united-states-2020

Minitab Blog Editor. (2013, May 30). Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit? Minitab. https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Self-score

Fill out this rubric for yourself and include it in your lab report. The same rubric will be used to generate a grade in proportion to the points assigned in the syllabus to the assignment.

Category	**Description**	**Points Possible**	**Score**
Structural Elements	All elements of a lab report are included **(2 points each)**: Title, Notice: Dr. Bryan Runck, Author, Project Repository, Date, Abstract, Problem Statement, Input Data w/ tables, Methods w/ Data, Flow Diagrams, Results, Results Verification, Discussion and Conclusion, References in common format, Self-score	28	**28**
Clarity of Content	Each element above is executed at a professional level so that someone can understand the goal, data, methods, results, and their validity and implications in a 5 minute reading at a cursory-level, and in a 30 minute meeting at a deep level **(12 points)**. There is a clear connection from data to results to discussion and conclusion **(12 points)**.	24	**23**
Reproducibility	Results are completely reproducible by someone with basic GIS training. There is no ambiguity in data flow or rationale for data operations. Every step is documented and justified.	28	**27**
Verification	Results are correct in that they have been verified in comparison to some standard. The standard is clearly stated **(10 points)**, the method of comparison is clearly stated **(5 points)**, and the result of verification is clearly stated **(5 points)**.	20	**19**
		100	**97**