Rodrigo Ivan Rodriguez

Final Project

DSC 530 Fall  
11/15/23

Investigating the Relationship Between School Location and Graduation Rates in Texas

**Statistical/Hypothetical Question**

In this data science project, I aim to explore the potential impact of school location on graduation rates in the state of Texas. Specifically, we will investigate whether urban schools exhibit higher graduation rates compared to rural schools. By conducting a comprehensive analysis of graduation data from the Texas Education Agency state government body.

I will analyze four-year longitudinal graduation data for the graduating class of 2018, 2019, and 2020. I will analyze the data to discover if there are any correlations, patterns, or disparities that may exist between graduation rates and other available factors.

There are a few questions I am hoping to answer with the available data. However, my primary hypothesis is that Urban Schools have higher graduation rates than Rural Schools.

**Outcome of the EDA**

The EDA helped me gage some basic insight into the contents of the data set used for this project. I found the unique counts of campuses (1924), districts (1074), counties (252), and metropolitan areas (20). Additionally, I found the distribution counts of the main classification variable of Region Type (63% Rural vs. 37% Urban).

Lastly, the histograms produced for the continuous variables showed me that some data cleaning had to be done. There were outliers in the data. For example, the variable of Hispanic Graduation Rates has a high 0 count. This occurs because some rural schools have a 0 count of Hispanic students and thus a graduation rate of 0; these instances need to be removed when analyzing the Hispanic subset. Additionally, the variable for total class size should not be 0; zero values for this variable need to be filtered out as well.

**What I Felt Was Missing During the Analysis**

One of the goals was to see if any factors in the data could be used to predict graduation rates. As I started analyzing the relationships between graduation rates and some of the variables, I found statistically significant relationships. However, always with low R-squared values. This means that there are possibly other factors not available in this dataset that more strongly impact graduation rates.

**Variables That Could Have Helped in the Analysis**

My assumption is that the following variables could be better predictors of graduation rate, rather than Location, Class Size, Ethnicity, or Economic-Status:

* Quality of Education
* Natural Aptitude (IQ)
* Mental Disorder Diagnosis
* English as a Second Language (Foreign Students)

**Outcome of my Assumptions**

Throughout the course of this analysis, I was able to answer the following research questions:

1. Do urban schools have higher graduation rates than rural schools?

No, on average rural schools have higher graduation rates than urban schools. This can be seen in the results of the hypothesis test done between school-location and grad rates.

1. Does class size affect high school graduation rates?

While there was a statistically significant relationship between these two variables. The amount of graduation rate variability that could be explained by this relationship was extremely low (R-squared of 0.003), leading me to believe that Class Size does not have much impact on graduation rates.

1. Does cultural ethnicity or economic status affect high school graduation rates?

When plotting the relationship between graduation rates of these subsets and graduation rates of the broader dataset, I found both subsets to be strongly correlated to overall graduation rates (Pearson’s Correlation of 0.98 for each subset). What this implies is that ethnicity and economic status do not play a factor in determining a student's academic performance, rather the broader school performance is a more reliable predictor of student performance within these subsets.

**Challenges and Future Questions**

The variables explored in this data all have a statistically significance impact on graduation rates. However, we have also determined that some of these relationships are weak and that there are variables outside of those available in this data that may have a greater impact on graduation rates. This project is a good starting point but there exists an opportunity to expand on the analysis conducted for this project. There is an opportunity to expand on this work by introducing additional variables into the data, and by experimenting with non-linear models to improve the chances of predicting graduation rates.