**Introduction**

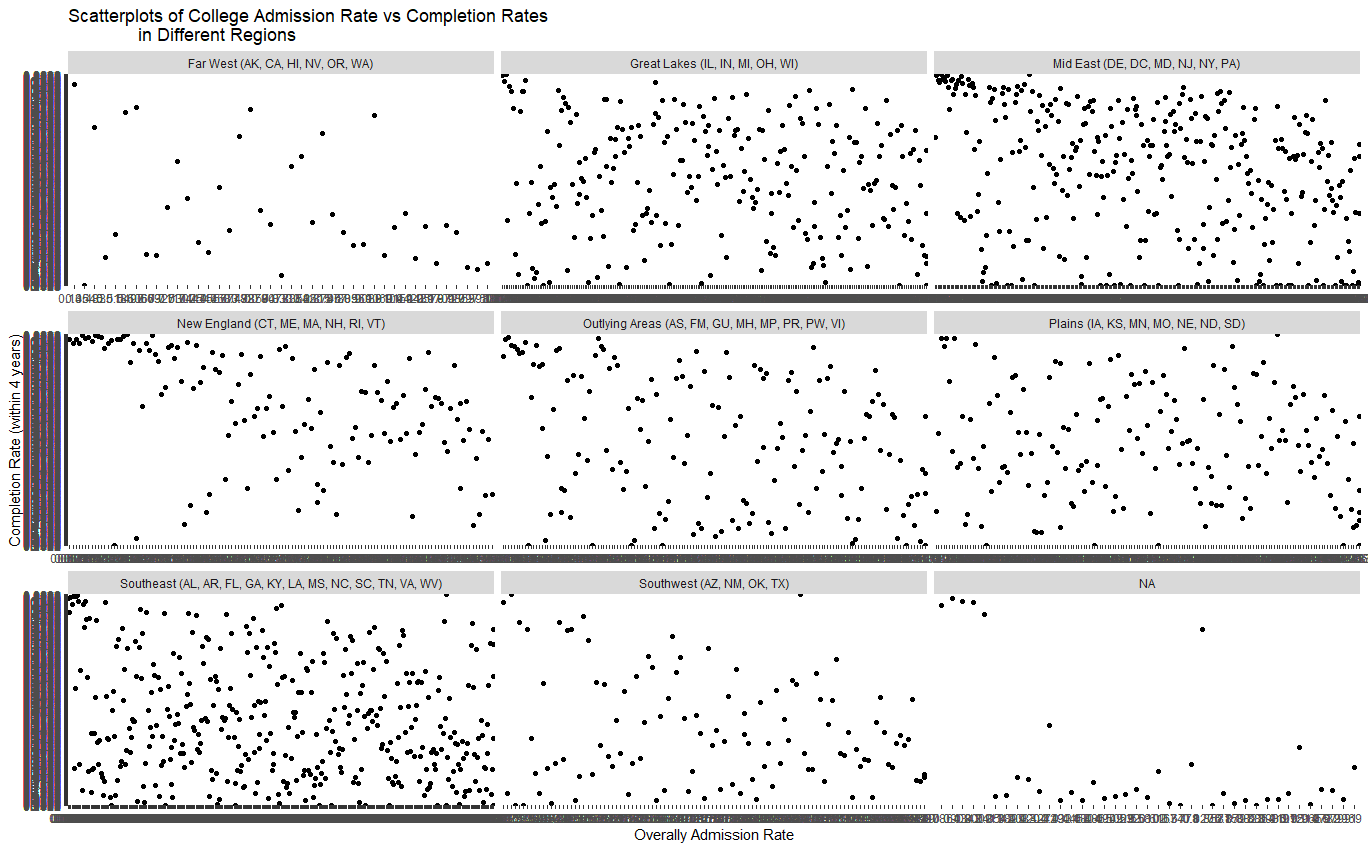
My goal is about determining if getting into colleges with lower acceptance rates actually affects the time it takes for a person to graduate within 4 years. The data I will be using to conduct this research is found from the “Most Recent Institution-Level Data” on the website <https://collegescorecard.ed.gov/data/> . This college scorecard dataset is created by the US government is about collected data on institution costs, graduation rate, employment rate, average amount of money borrowed and loaned, and the loan default rate. Other variables include post-graduate income, percentage of graduates earning more than students with just a high-school degree, race/ethnicity, socio-economic diversity, student/faculty ration, and over 2000 more variables. I thought this would be an original idea because a lot of students and/or parents stress over getting into a good college and most if not all colleges that are considered “good” typically have a low acceptance rate. There seems to be a misconception that schools with a lower acceptance rate ultimately provide a better education to those colleges of higher acceptance rates. This may lead people into believing students who attend lower accepted schools work harder to graduate on time compared to students who attend higher accepted schools. But there are a lot of factors that may impact a student’s ability to graduate within 4 years and those factors can include, their health, part-time job, family obligations, mental and physical state, disabilities, and etc. so it wouldn’t be right to automatically assume that schools with lower accepted schools guarantee or increase the chances of graduating on time. From this, I hypothesize that there will be no correlation between a school’s acceptance rate and a student’s time of graduation. There was some similar research regarding school admissions and post-college aspects, but none of them tried to analyze a school’s acceptance rate on graduation rate. One study analyzed a person’s wealth on their quality of education while another study analyzed students who go to work during college and their income after college.

**Data Exploration & Methods**

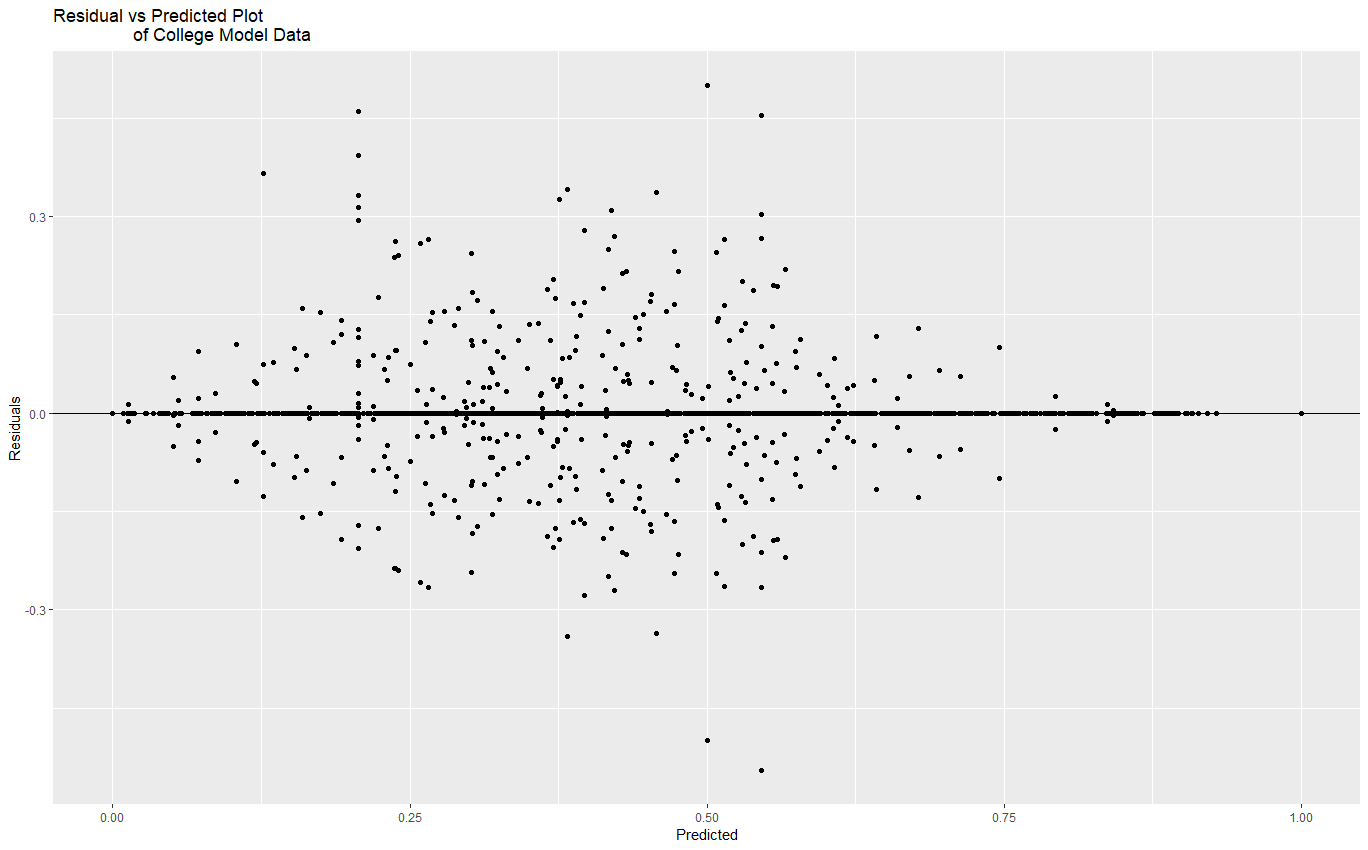
From the dataset chosen, there a total of 2989 variables and 6681 observations. The variables of interest include “INSTNM” (institution name), “CITY”, “STABBR” (state), “REGION”, “ADM\_RATE” (admission rate), and “C100\_4” (completion rate within 4 years). All other variables from the original data will not be used since I will only be focusing on analyzing completion rate and admission rate of different institutions. Any variables including SAT scores, ACT scores, number of degrees, gender, race, finances, affiliations, etc. will not be used since they are not necessary for my goal. All coding and visualizations will be done through R. One method I will be using to achieve my goal is creating a scatter plot of different regions to see if there is any correlation with admission rate and completion rate in different regions. If there are any strong correlations among the regions, we can conclude that the certain regions have a stronger correlation compared to the rest. Another method I will use will be creating a linear regression prediction model. This method will allow me to better visualize the correlation of admission rate and completion rate as a whole. If there is no strong positive or negative correlation between the two variables, then we can conclude that there is no correlation. The multiple scatterplots can allow us to easily see the correlation between admission rate and completion rate as well as any clustering effects. However, scatterplots may not be able to represent the conclusion well if they are not read correctly and they may not be completely accurate since numerical values of variables are often rounded off or omitted depending on the situation. For the linear regression model, it can be good for continuous linear and quantitative data and it is easier to implement overall. The biggest pro would be by making estimations simple and easy to interpret. However, linear regression models are highly susceptible to noise, mistakes, and outliers.

**Final Results**

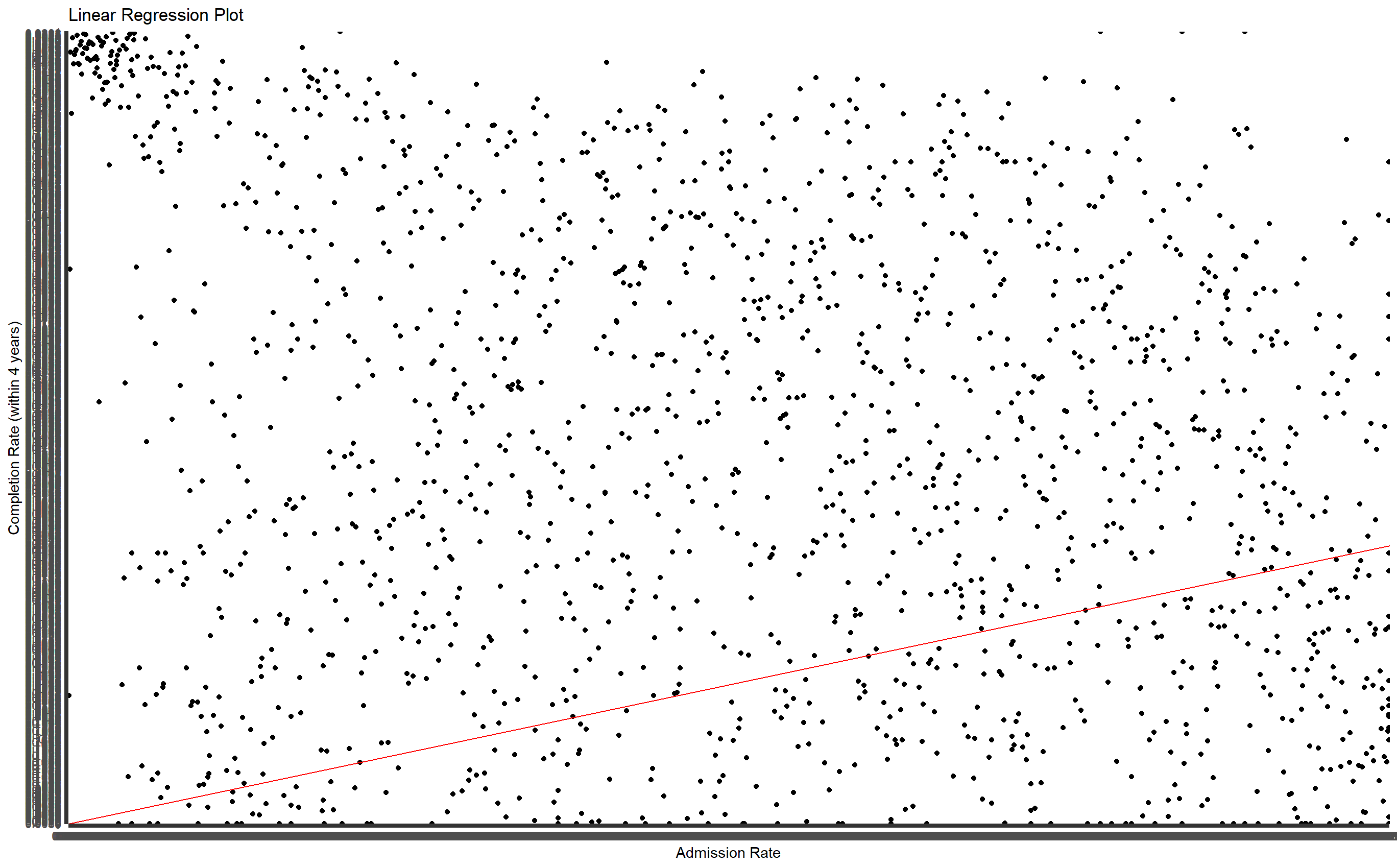
From the original dataset, variables of interest were first selected out from the rest of the unneeded variables. Then, any observations without any numerical values regarding admission rates or completion rates were omitted and filtered out. Once the new dataset was less crowded and noisy, the column names were rewritten so that they were more understandable. After, the states variables were grouped into different regions so that we could visualize the different areas of the states better.



For the first visualization, I decided to create faceted scatter plots for the different regions and from the plots, we can see that the points from each region are scattered evenly and there aren’t any clusters or strong positive or negative trends. This method was done because I wanted to know if the effects of admission rate on completion rate would vary from region to region. From this specific visualization, we can say that there is no correlation between admission rates and completion rates in different regions.



For the second plot, I decided to create a residual vs. predicted plot to see if the dataset would be suitable for a linear regression. Since there are an even number of points above and below the regression line, we can conclude that a linear model would be appropriate.



Lastly, I created a linear regression plot of all the points rather than separating them into regions in the first plot. This was done so we could visualize the dataset as a whole. From this visualization, we can see a small cluster of points located in the top left where admission rates are low and completion rates are high. However, the rest of the plot show an even distribution of points regardless of where those points are located. In conclusion, there seems to be some affect of admission rate on completion rate, but there may not be a correlation between the two variables overall according to the plot as a whole.

**Citations**

Dahill-Brown, S. E., Witte, J. F., & Wolfe, B. (2016). Income and Access to Higher Education: Are High Quality Universities Becoming More or Less Elite? A Longitudinal Case Study of Admissions at UW-Madison. The Russell Sage Foundation journal of the social sciences : RSF, 2(1), 69–89. https://doi.org/10.7758/RSF.2016.2.1.04

Douglas, D., & Attewell, P. (2019). The Relationship Between Work During College and Post College Earnings. Frontiers in sociology, 4, 78. https://doi.org/10.3389/fsoc.2019.00078