Final Project

MinJae Jo

2025-12-02

Introduction

In this project, I look at the question, "Does the competitiveness of a college relate to its degree attainment rate?" I chose C150_4 as a response variable because it shows the percentage of students earning degrees within 150% of a typical period. This is one of the clearest indicators of graduation performance on the dataset.

The SAT_AVG was used as an explanatory variable. SAT scores are often used as indicators of a college's selection level or academic competition rate, although they do not directly measure graduation rates. Colleges with higher SAT averages generally accepted outstanding students and had a more competitive academic environment, so they thought that this method was reasonable for comparing the differences between colleges.

Since both variables are continuous, I want to use a linear model to see if there is a connection between the SAT mean and graduation rate. Many people think that universities with higher competition rates have better outcomes, but this topic comes up interesting because it's hard to know exactly this without looking at real data. This project helped us see what the CollegeScorecard dataset suggests about these patterns.

Preprocessing

```
##  grad_rate_150       sat_avg
##  Min.   :0.0000   Min.   : 564
##  1st Qu.:0.3229   1st Qu.:1044
##  Median :0.4944   Median :1116
##  Mean   :0.4881   Mean   :1131
##  3rd Qu.:0.6453   3rd Qu.:1195
##  Max.   :1.0000   Max.   :1558
##  NA's   :4703     NA's   :5743
```

Visualization

Summary statistics for graduation rate (C150_4)

n_grad	mean_grad	median_grad	sd_grad	iqr_grad	min_grad	max_grad
7058	0.4881164	0.4944	0.2233851	0.3224	0	1

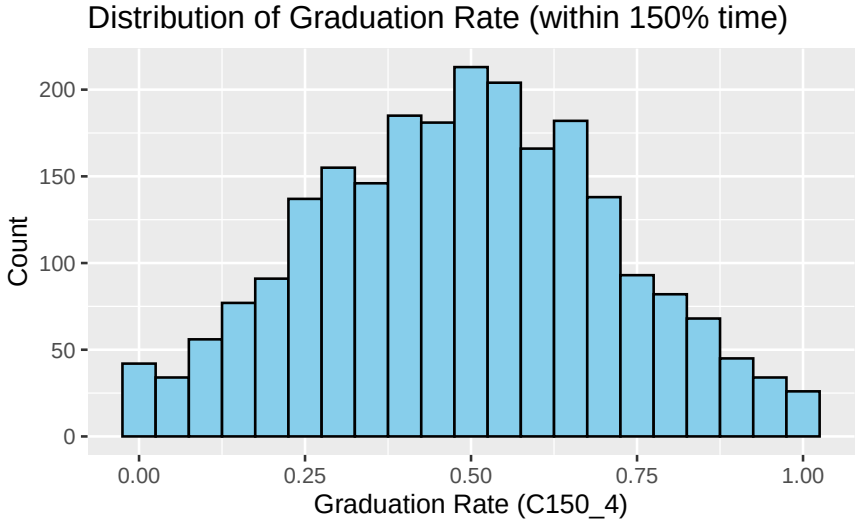
-The graduation rate values range from 0 to 1. The median is around 0.49, so about half of the colleges graduate less than half of their students on time. This suggests that many schools are kind of in the middle rather than extremely high or low.

Summary statistics for SAT_AVG

n_sat	mean_sat	median_sat	sd_sat	iqr_sat	min_sat	max_sat
7058	1131.28	1116	129.6887	150.5	564	1558

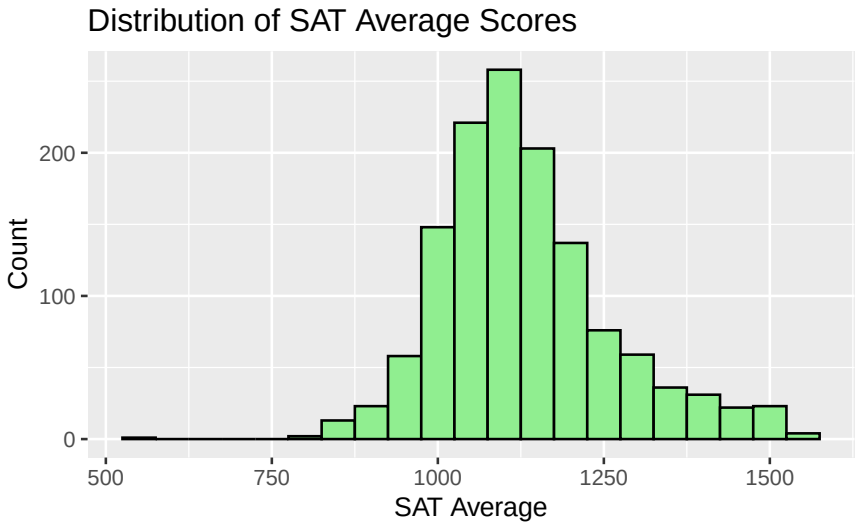
-The SAT averages go from the mid-500s to around 1550. The median is about 1116, which means most schools are not super selective, but not weak either. The scores are pretty spread out, so colleges in this dataset vary a lot in competitiveness.

Graduation Rate of Histogram



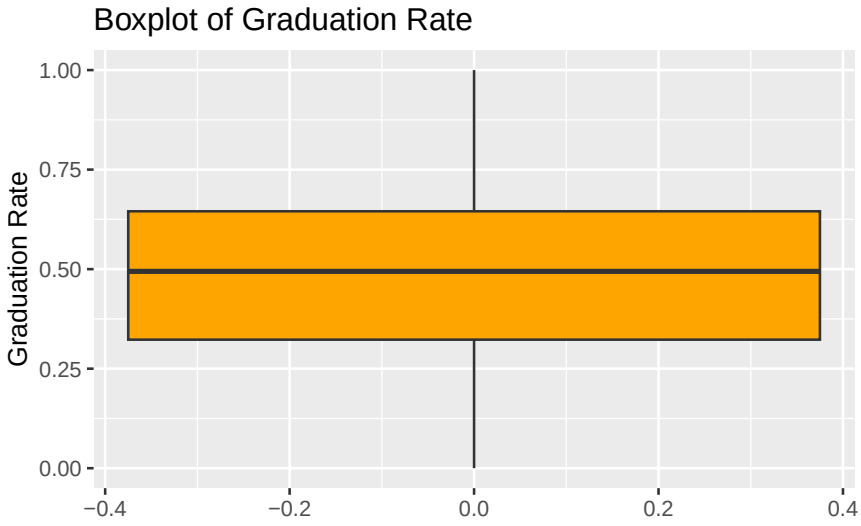
-The graduation rate histogram shows that most colleges have graduation rates clustered around the middle, roughly between 0.3 and 0.7. The highest bars appear near the center, which means many schools graduate around half of their students within 150% of the normal time. Very low and very high rates exist, but they are much less common.

SAT Average of Histogram



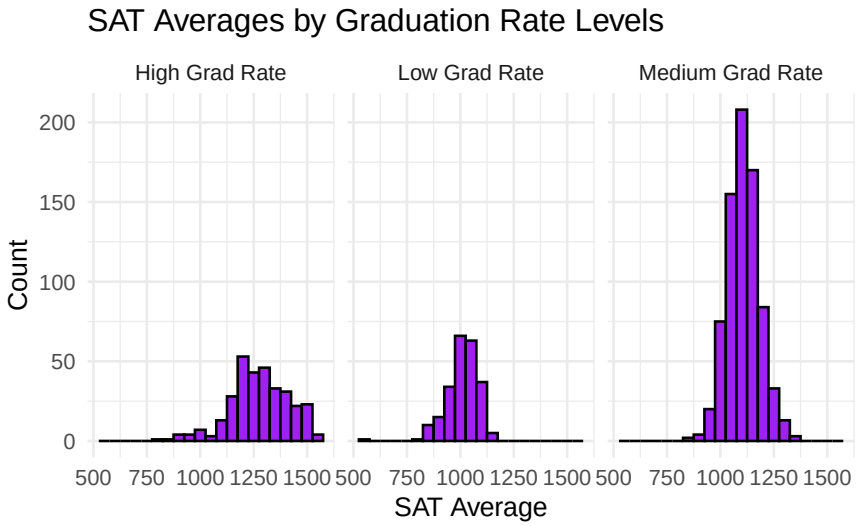
-The SAT average scores form a distribution that looks close to a normal shape. Most schools have SAT averages between about 1000 and 1200, and the center is around 1100. There are a few schools with much lower or much higher scores, but overall the scores are tightly grouped, suggesting similar academic selectivity for many colleges.

Graduation Rate of Boxplot



-The boxplot shows that the middle half of graduation rates is fairly wide, meaning schools vary a lot. The median is close to the middle of the box, so the distribution is balanced. There are also some lower values stretching down toward 0, showing that a number of colleges have low graduation performance.

Graduation Rate of Facting



-This graph divides the graduation rate into three levels (High, Medium, and Low) and compares how the SAT scores distribute differently in each group. The graph shows that groups with higher graduation rates tend to have higher SAT scores. In contrast, groups with lower graduation rates tend to have more widespread SAT scores and lower median scores. This pattern visually shows the relationship that "the more competitive schools have, the higher the graduation rate" than a simple scatterplot.

Summary Statistics

Graduation Rate Summary

n_grad	mean_grad	median_grad	sd_grad	iqr_grad	min_grad	max_grad
7058	0.4881164	0.4944	0.2233851	0.3224	0	1

SAT Summary

n_sat	mean_sat	median_sat	sd_sat	iqr_sat	min_sat	max_sat
7058	1131.28	1116	129.6887	150.5	564	1558

Data Analysis

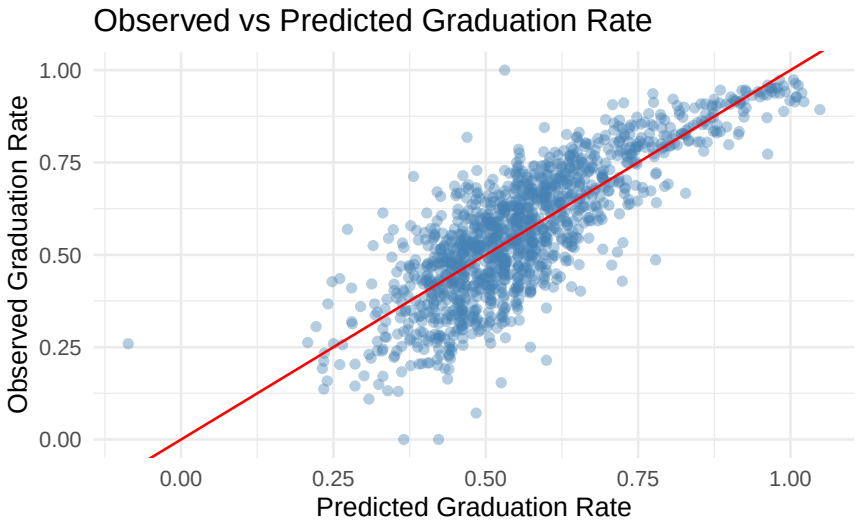
term	estimate	std.error	statistic	p.value
(Intercept)	-0.7307384	0.0258762	-28.23975	0

term	estimate	std.error	statistic	p.value
sat_avg	0.0011419	0.0000227	50.32447	0

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.662702	0.6624409	0.104931	2532.553	0	1	1079.643	-2153.286	-2137.796	14.19269	1289	1291

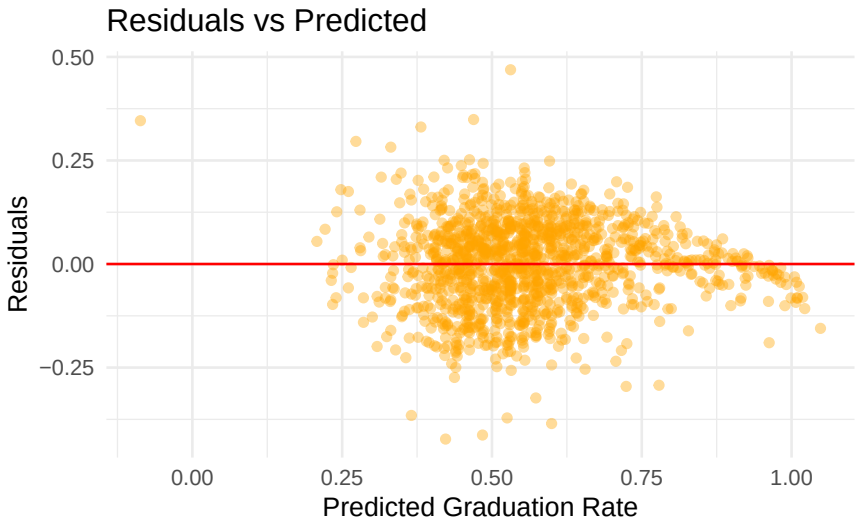
-Overall, linear models show a slight positive correlation between the SAT averages and graduation rates. Higher SAT averages tend to have slightly higher graduation rates, but R squared values are not very high. This suggests that SAT scores alone do not fully explain graduation scores, and other factors are likely to play a large role.

Observed vs Predicted Plot



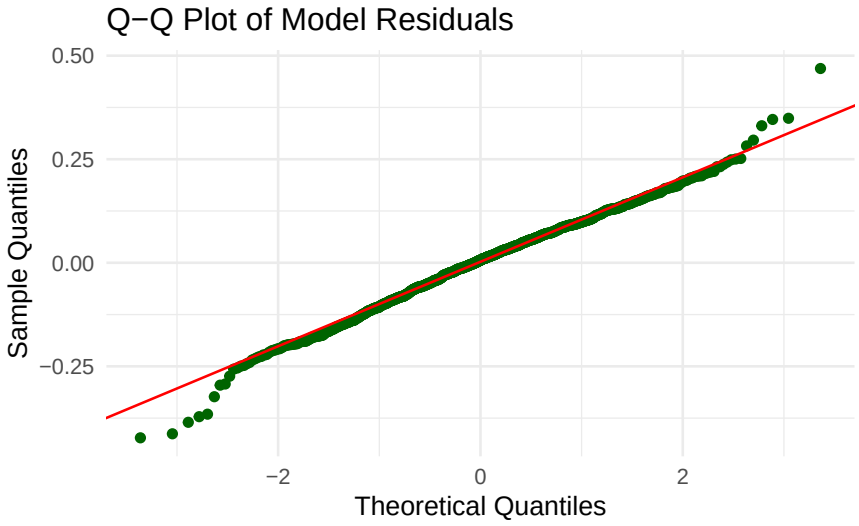
-This graph compares the actual graduation rate with the one predicted by the model. A closer the dot is to the diagonal line (red line), the closer it is to the line in general, but not quite the same. In other words, the SAT averages account for graduation rates to some extent, but show that there are quite a few other effects.

Residuals vs Predicted plot



-The residual graph is used to check whether the error is uniformly spread according to the predicted value. Here, the dots are scattered around the centerline, but they also show some patterns, so they are not completely ideal. This means that the linear model does not account for all the variations.

Q-Q plot



-A Q-Q plot is a graph that verifies that the residuals are normally distributed. Although the points are usually on a straight line, they appear to deviate slightly from the ends. In other words, the model's assumption of normality is not significantly broken, but it is not a perfect fit.

Conclusion

In this project, I compared the average SAT score with the college graduation rate (within 150 percent of the time), and looked at how much SAT scores could explain college graduation performance. Overall, it is true that there is a positive relationship in which the higher the SAT average, the higher the graduation rate. Visualization, summary statistics, and linear model results all show similar patterns.

However, it was also confirmed that the SAT scores did not fully explain the graduation rate, as the model's R-squared value was about 0.66. In other words, schools with higher SAT scores generally have higher graduation rates, but there is a limit to predicting college graduation outcomes based on SAT scores alone. Other factors, such as student aid services, financial conditions, school size, and program quality, are likely to play an important role.

When looking at the three model diagnostic graphs (Observed vs Predicted, Residuals vs Predicted, Q–Q Plot), the linear model is generally not very wrong, but it is difficult to say that it is a perfect model. In particular, the fact that the residuals are not completely evenly spread, and the pattern appears in some sections may be a signal that additional variables are needed.

Overall, I conclude that SAT scores do some good to explain graduation rates, but this alone is not appropriate to evaluate or predict school performance. This analysis suggests that colleges should consider various support systems and environmental factors together in order to improve student performance, rather than simply recruiting high-scoring students.