Final Project

Shruti Valappil

2025-05-05

Introduction

I will explore the relationship between student academic preparedness and institutional success by examining how the average SAT score of admitted students relates to the graduation rate of a college. Specifically, "How does the average SAT score of admitted students relate to the graduation rate of a college?" This question is important because standardized test scores are often used as a key metric in college admissions, and understanding their potential link to graduation outcomes can provide insights for both policymakers and prospective students.

To answer this question, I will use two variables from the College Scorecard dataset: SAT_AVG, which represents the average SAT score of students admitted to each institution, and C150_4, which measures the percentage of students who graduate within 150% of the expected time for 4-year programs. C150_4 will serve as the response variable, and SAT_AVG will be the explanatory variable. I will apply a linear regression model to analyze the strength and direction of the relationship between these variables.

This analysis could offer valuable information on whether higher SAT scores, as a proxy for academic readiness, are associated with better college completion outcomes. If a significant relationship is found, it may support the continued use of standardized tests in admissions—or spark discussion on alternative metrics if the relationship is weak.

Preprocessing

To answer my question, "How does the average SAT score of admitted students relate to the graduation rate of a college?", I will create a reduced dataset containing only the two variables relevant to this analysis: SAT_AVG (average SAT score) and C150_4 (graduation rate within 150% of the expected time for 4-year programs). These variable names are not immediately interpretable, so I will rename them to average_SAT and grad_rate_150pct to make the data more human-readable and easier to work with.

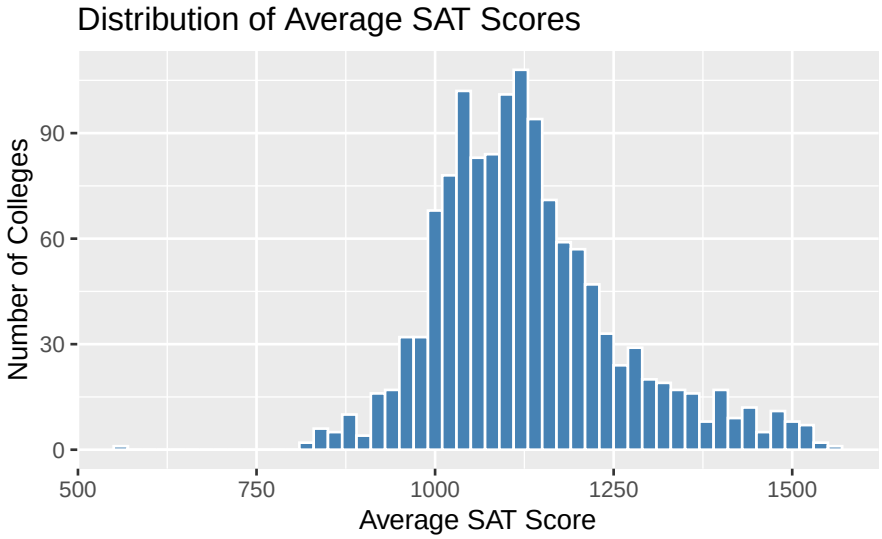
```
college_reduced <- college %>%
  select(SAT_AVG, C150_4) %>%
  rename(
    average_SAT = SAT_AVG,
    grad_rate_150pct = C150_4
    )
```

Visualization

Graph 1: Histogram of average SAT scores

This histogram shows the distribution of SAT scores across colleges. I want to examine the spread of SAT scores and identify any clustering or skewed patterns that may suggest selectivity differences among institutions.

```
college_reduced %>%
  ggplot() +
  geom_histogram(aes(x = average_SAT), binwidth = 20,
                 fill = "steelblue", color = "white") +
  labs(title = "Distribution of Average SAT Scores",
       x = "Average SAT Score",
       y = "Number of Colleges"
       )
```



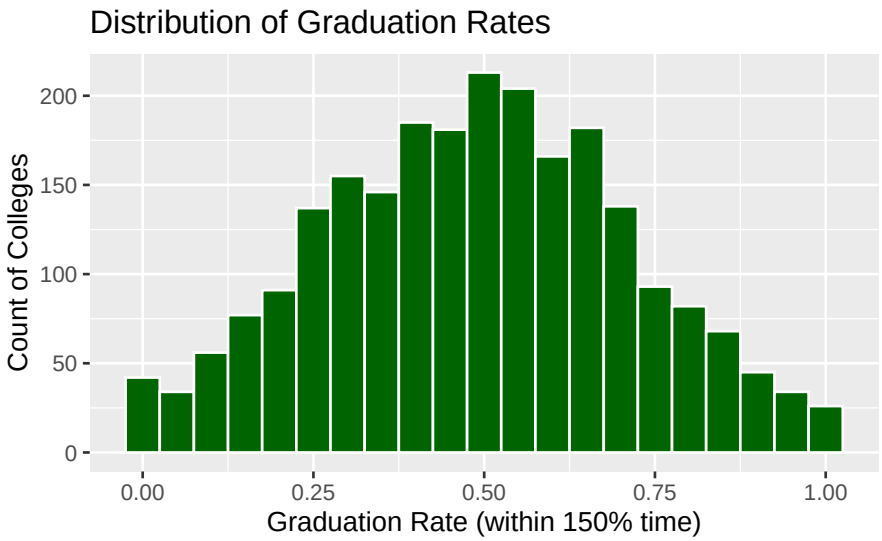
The distribution is roughly bell-shaped but shows a slight right skew, with a concentration of colleges reporting average SAT scores between 1000 and 1200. The most common SAT averages fall near 1050–1100, indicating this is where most institutions cluster. There are relatively fewer institutions with averages below 900 or above 1300. The tail extends more on the right, suggesting a few highly selective institutions report significantly higher average SAT scores, though these are rarer.

Overall, the histogram illustrates that most colleges attract students with mid-range SAT scores, while both very low and very high SAT averages are less frequent. This insight helps frame the SAT score variable before analyzing its relationship with graduation rates.

Graph 2: Histogram of Graduation Rates

This graph helps understand how graduation rates vary across colleges, which is important to assess before modeling.

```r
college_reduced %>%
  ggplot() +
  geom_histogram(aes(x = grad_rate_150pct), binwidth = 0.05,
                 fill = "darkgreen", color = "white") +
  labs(title = "Distribution of Graduation Rates",
       x = "Graduation Rate (within 150% time)",
       y = "Count of Colleges"
       )
```



The distribution appears to be approximately normal but slightly left-skewed, with the majority of institutions clustering around the 0.50 mark—indicating that many colleges have around a 50% graduation rate. There is a noticeable decline in the number of colleges as we move toward the extremes, especially below 0.2 and above 0.8. A small number of colleges have graduation rates near zero or one, but these are rare.

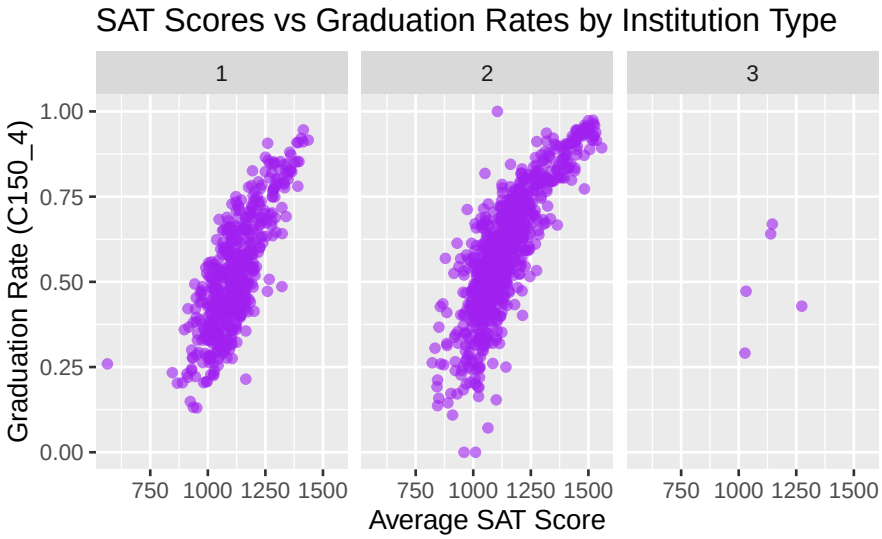
The center of the distribution, the modal graduation rate range, lies between 0.4 and 0.6, suggesting this is a common outcome across U.S. colleges. This histogram effectively shows the overall shape and spread of graduation rates, laying the foundation for understanding how this variable behaves before exploring its relationship with predictors like SAT scores or institutional type.

Graph 3: Scatter plot of SAT Scores vs Graduation Rates (Faceted by Control)

This scatter plot directly visualizes the relationship between SAT scores and graduation rates. I also facet by CONTROL (public, private nonprofit, private for-profit) to explore how institution type might influence this relationship.

```r
college %>%
  ggplot() +
  geom_point(aes(x = SAT_AVG, y = C150_4), alpha = 0.6,
             color = "purple") +
  facet_wrap(~ CONTROL) +
  labs(title = "SAT Scores vs Graduation Rates by Institution Type",
```

```
       x = "Average SAT Score",
       y = "Graduation Rate (C150_4)"
       )
```



This faceted scatter plot shows a strong positive relationship between average SAT scores and graduation rates across institution types. In both Public (1) and Private Nonprofit (2) institutions, higher SAT scores are clearly associated with higher graduation rates, forming tight upward trends. The relationship is particularly strong and linear among Private Nonprofit institutions.

In contrast, Private For-profit (3) institutions display few data points, and the pattern is much more scattered and weak, suggesting that SAT scores in this group may not predict graduation outcomes as clearly—likely due to different admission or academic support practices.

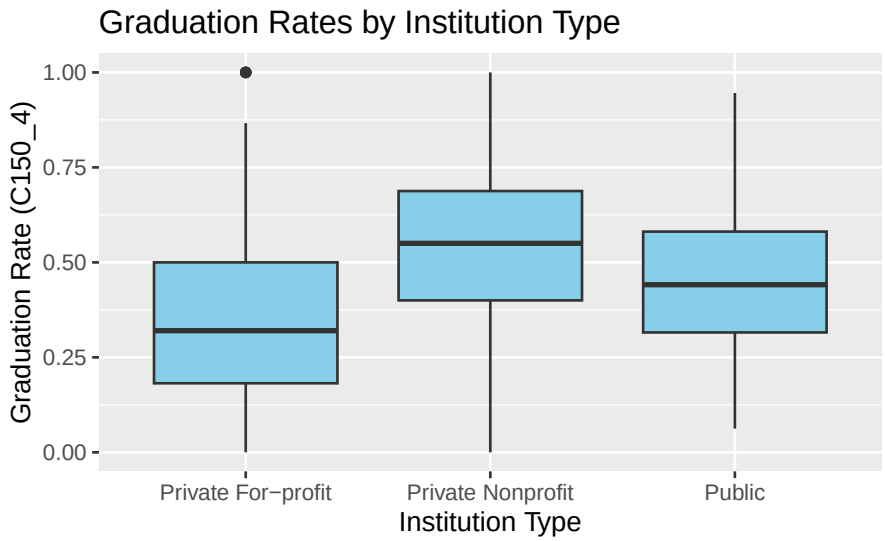
By using faceting, this graph effectively highlights how the relationship between student preparedness (SAT scores) and success (graduation rates) varies by institutional control.

Graph 4: Boxplot of Graduation Rates by Institution Types

This boxplot shows how graduation rates vary across different types of institutions. It helps identify whether institution control (public, private nonprofit, or for-profit) is associated with different performance outcomes.

```
college_box <- college %>%
  select(C150_4, CONTROL) %>%
  mutate(
    control_type = recode(CONTROL,
                          `1` = "Public",
                          `2` = "Private Nonprofit",
                          `3` = "Private For-profit")
  )
```

```r
college_box %>%
  ggplot() +
  geom_boxplot(aes(x = control_type, y = C150_4), fill = "skyblue") +
  labs(title = "Graduation Rates by Institution Type",
       x = "Institution Type",
       y = "Graduation Rate (C150_4)"
       )
```



The boxplot visualizes the variance in graduation rates (C150_4) across three categories of institution types: Private For-profit, Private Nonprofit, and Public. Each box represents the interquartile range (middle 50% of values), with the line inside indicating the median graduation rate.

The Private Nonprofit institutions show a higher center (median) and a relatively narrower spread, suggesting more consistent and generally better graduation outcomes in this group. Public institutions have a slightly lower median, with a wider range, indicating more variability in outcomes. Private For-profit institutions have the lowest median graduation rate and the widest spread, including an outlier with a perfect 100% rate. This broad variation suggests that student outcomes are much less predictable in the for-profit sector.

This pattern highlights how institutional control influences variation in graduation outcomes, with Private Nonprofits generally achieving higher and more consistent performance, while For-profits show the greatest inconsistency and lower typical outcomes.

Summary Statistics

For each continuous variable:

```r
college_reduced %>%
  summarize(
    count_SAT = sum(!is.na(average_SAT)),
    mean_SAT = mean(average_SAT, na.rm = TRUE),
```

```
    median_SAT = median(average_SAT, na.rm = TRUE),
    range_SAT = max(average_SAT, na.rm = TRUE) - min(average_SAT, na.rm = TRUE),
    sd_SAT = sd(average_SAT, na.rm = TRUE),
    iqr_SAT = IQR(average_SAT, na.rm = TRUE),

    count_grad = sum(!is.na(grad_rate_150pct)),
    mean_grad = mean(grad_rate_150pct, na.rm = TRUE),
    median_grad = median(grad_rate_150pct, na.rm = TRUE),
    range_grad = max(grad_rate_150pct, na.rm = TRUE) - min(grad_rate_150pct,
                                                           na.rm = TRUE),
    sd_grad = sd(grad_rate_150pct, na.rm = TRUE),
    iqr_grad = IQR(grad_rate_150pct, na.rm = TRUE)
  ) %>%
  select(count_grad:sd_grad)
```

count_grad	mean_grad	median_grad	range_grad	sd_grad
2355	0.4881164	0.4944	1	0.2233851

1. SAT Scores (average_SAT) Count: 1315 colleges have reported SAT scores. Mean: The average SAT score across colleges is about 1131.3. Median: The middle SAT score is 1116, suggesting a slightly left-skewed distribution. Range: The difference between the highest and lowest SAT scores is 994, showing substantial variability in student preparedness. Standard Deviation: Around 129.7, indicating a moderately wide spread around the mean. Interquartile Range (IQR): 150.5, confirming that middle 50% of SAT scores are relatively tightly packed.

2. Graduation Rates (grad_rate_150pct) Count: 2355 colleges have reported graduation rates. Mean: The average graduation rate is about 48.8%, which is slightly below the median. Median: The median graduation rate is 49.4%, suggesting a fairly symmetric distribution. Range: The full range is 1 (from 0 to 1), indicating maximum possible variation. Standard Deviation: 0.223, showing a relatively high dispersion for a proportion-based variable. IQR: 0.3224, which indicates that 50% of colleges fall within a 32.24 percentage point range.

Data Analysis

In this analysis, I will determine whether the average SAT score of admitted students (SAT_AVG) is linearly related to a college's graduation rate (C150_4), measured as graduation within 150% of the expected time. Since both variables are continuous, I will use linear regression modeling to assess the strength and direction of this relationship.

Handle missing values:

```
college_model_data <- college_reduced %>%
  filter(!is.na(average_SAT), !is.na(grad_rate_150pct))
```

Linear model:

```r
model <- lm(grad_rate_150pct ~ average_SAT, data = college_model_data)
```

```r
tidy(model)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.7307384	0.0258762	-28.23975	0
average_SAT	0.0011419	0.0000227	50.32447	0

```r
glance(model) %>%
  select(r.squared:p.value)
```

r.squared	adj.r.squared	sigma	statistic	p.value
0.6627025	0.6624409	0.1049315	2532.553	0

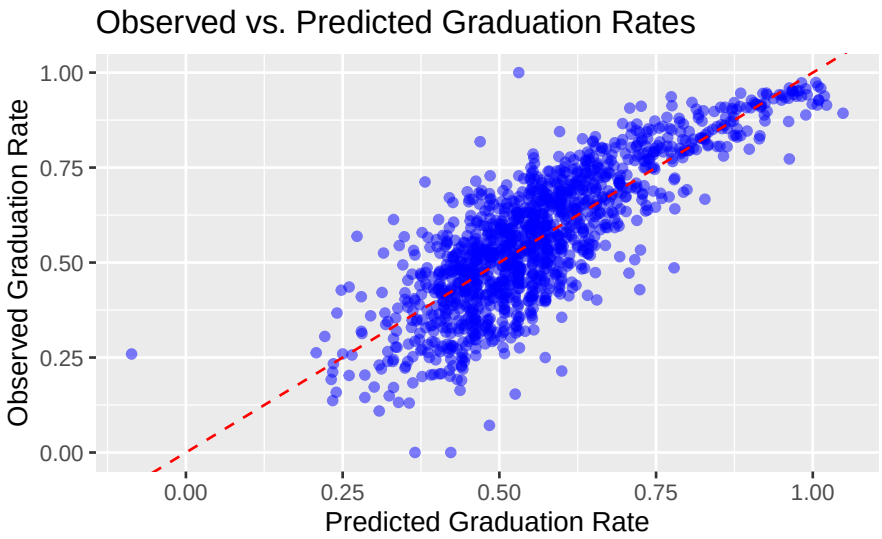
These results of tiny and glance function suggest that average SAT score is a strong and significant predictor of college graduation rates. The positive coefficient means that institutions with higher SAT scores tend to have higher graduation rates. Given the high R-squared value and the extremely low p-value, we can conclude that SAT scores explain a substantial portion of the variability in graduation outcomes among colleges in the dataset.

Model Assumptions using diagnostic plots:

```r
aug_data <- augment(model)
```

1. Observed vs Predicted Plot

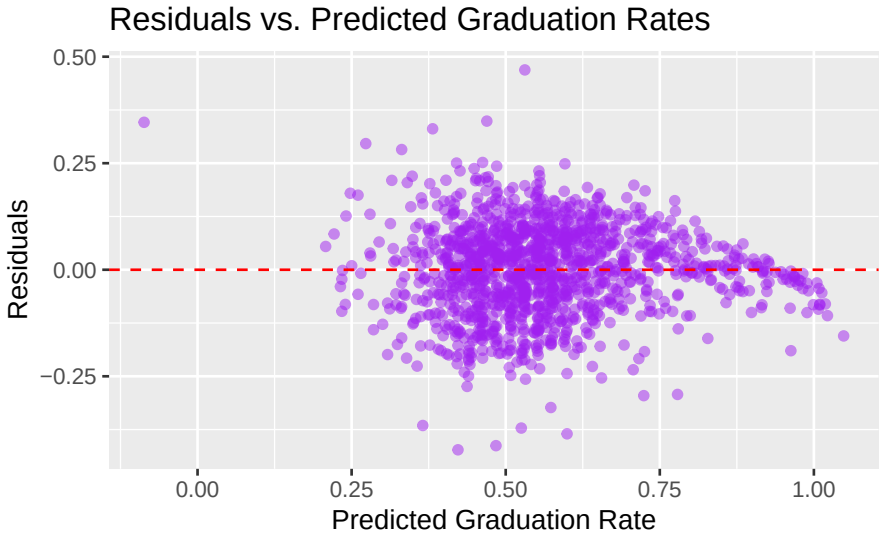
```r
aug_data %>%
  ggplot(aes(x = .fitted, y = grad_rate_150pct)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Observed vs. Predicted Graduation Rates",
    x = "Predicted Graduation Rate",
    y = "Observed Graduation Rate"
  )
```



Linearity: The observed vs. predicted plot shows a strong, approximately linear relationship between the predicted graduation rates from the model and the actual observed values. The points are clustered closely around the 45-degree reference line in red, indicating that the model is effectively capturing the linear trend in the data. This supports the assumption of a linear relationship between average SAT score and graduation rate.

2. Residuals vs Predicted Plot

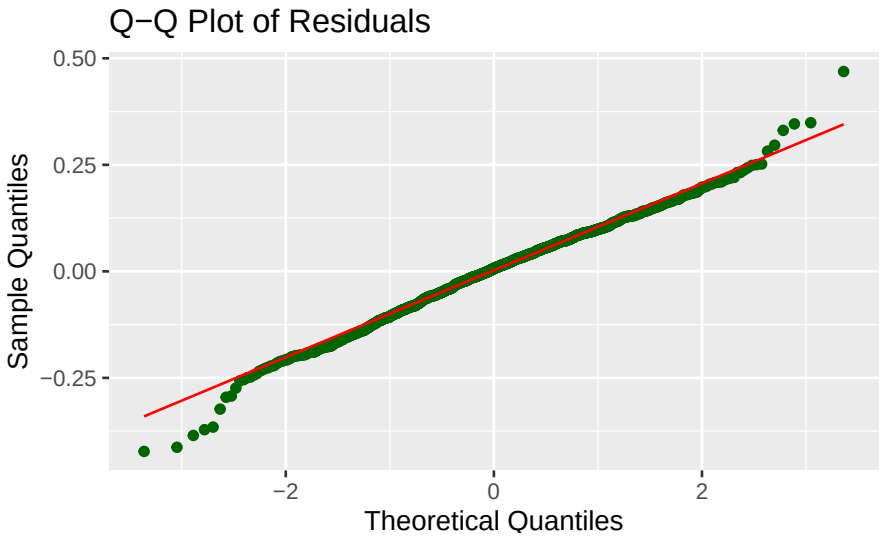
```
aug_data %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5, color = "purple") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Residuals vs. Predicted Graduation Rates",
    x = "Predicted Graduation Rate",
    y = "Residuals"
  )
```



Homoscedasticity: The residuals vs. predicted plot indicates that residuals are generally centered around zero, but the spread of residuals slightly decreases as the predicted graduation rate increases. This "funnel" shape suggests some mild heteroscedasticity—i.e., non-constant variance in residuals—which is a minor violation of the homoscedasticity assumption. However, the overall pattern is not extreme and the residuals remain fairly symmetrically distributed, so the model is still reasonably reliable.

3. Q-Q Plot

```
aug_data %>%
  ggplot(aes(sample = .resid)) +
  stat_qq(color = "darkgreen") +
  stat_qq_line(color = "red") +
  labs(
    title = "Q-Q Plot of Residuals",
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
  )
```



Normality of Residuals: The Q-Q plot shows that the residuals closely follow the red diagonal line, which is the expected distribution under normality. While there are some deviations at the tails, especially on the right, the residuals overall appear to be approximately normally distributed. This supports the normality assumption needed for accurate hypothesis testing and confidence intervals in the linear model.

These diagnostics suggest that the linear model is generally appropriate, with minor deviations in variance that do not significantly undermine the overall conclusions.

Conclusion

The results of this analysis indicate a strong and statistically significant positive relationship between the average SAT score of admitted students and the graduation rate of a college. The linear

regression model showed that as average SAT scores increase, so do graduation rates, with an R-squared value of approximately 0.66, indicating that about 66% of the variation in graduation rates can be explained by differences in average SAT scores. Visualizations such as the scatterplot and faceted plots further supported this relationship, showing a clear upward trend across different institution types. Summary statistics also revealed that colleges with higher average SAT scores tend to have higher median graduation rates.

These findings suggest that student academic preparedness, as measured by SAT scores, may be a key factor influencing graduation outcomes at the institutional level. While this does not imply causality, the strong association underscores the potential importance of admissions criteria in institutional performance metrics. One limitation is that other variables—such as financial aid availability, faculty-student ratios, or institutional support programs—were not included in the model but may confound the results. Additionally, because the data are aggregated at the institutional level, we cannot draw conclusions about individual students. Still, the analysis has implications for policy discussions around admissions standards, equity in standardized testing, and institutional accountability.

Academic Integrity statement

I have not used an AI tool for this assignment.