

Final Project

Jaden Pathammavong

2024-12-10

Introduction

The purpose of this project is to determine whether the sort of financial help received by first-time, full-time students has an impact on their completion rate. Though its effect on graduation rates is less evident, financial aid—such as grants and loans—is essential in ensuring that higher education is accessible. The type of financial aid received (categorical: “Grants only,” “Loans only,” “Both,” or “None”) is the explanatory variable, while the response variable is the completion rate (continuous). The mean completion rates for each of these groups will be compared using modeling testing to see if any differences are statistically significant. This analysis can help inform judgments about enhancing educational outcomes and is crucial for comprehending how financial aid affects student performance.

Preprocessing

```
college_reduced <- college %>%  
  select(C150_4_NOLOANNOPELL, C150_4_LOANNOPELL, C150_4_PELL, REGION)
```

The college dataset is filtered to hold only the columns C150_4_NOLOANNOPELL, C150_4_LOANNOPELL, and C150_4_PELL, along with the REGION column, to focus on completion rates and regional information for analysis.

```
college_reduced <- college_reduced %>%  
  rename(  
    No_Aid_Completion_Rate = C150_4_NOLOANNOPELL,  
    Loan_Only_Completion_Rate = C150_4_LOANNOPELL,  
    Pell_Only_Completion_Rate = C150_4_PELL  
  )
```

The columns are renamed to more descriptive names:

No_Aid_Completion_Rate, Loan_Only_Completion_Rate, and Pell_Only_Completion_Rate

```
college_reduced <- college_reduced %>%
  mutate(
    Region_Label = recode(
      REGION,
      `1` = "Northeast",
      `2` = "Southeast",
      `3` = "Midwest",
      `4` = "West",
      `5` = "Southwest",
      `6` = "Mountain",
      `7` = "Other"
    )
  )
```

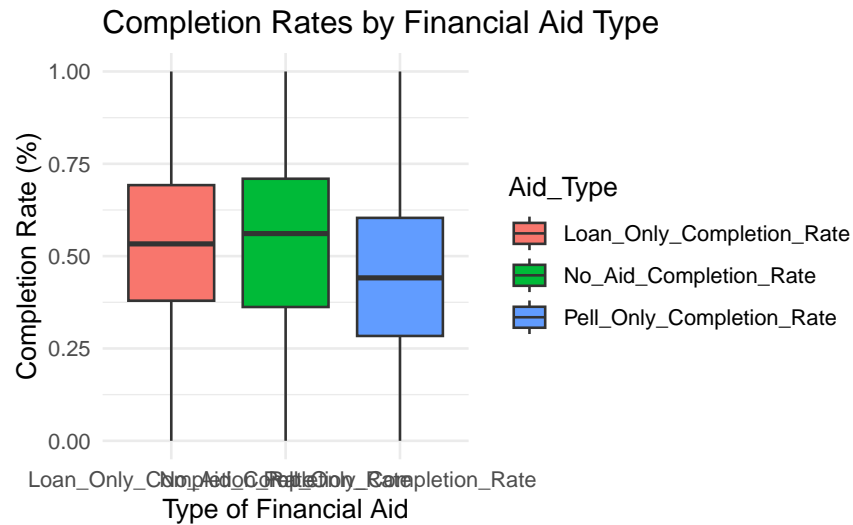
```
## Warning: There was 1 warning in 'mutate()'.
## i In argument: 'Region_Label = recode(...)'.
## Caused by warning:
## ! Unreplaced values treated as NA as '.x' is not compatible.
## Please specify replacements exhaustively or supply '.default'.
```

A new column, `Region_Label`, is created using the `mutate()` function, which recodes the `REGION` variable into more descriptive labels such as “Northeast” and “Southeast” to facilitate easier interpretation of regional data.

Visualization

```
college_reduced %>%
  pivot_longer(
    cols = c(No_Aid_Completion_Rate,
              Loan_Only_Completion_Rate,
              Pell_Only_Completion_Rate),
    names_to = "Aid_Type",
    values_to = "Completion_Rate"
  ) %>%
  ggplot(aes(x = Aid_Type, y = Completion_Rate, fill = Aid_Type)) +
  geom_boxplot() +
  labs(
    title = "Completion Rates by Financial Aid Type",
    x = "Type of Financial Aid",
    y = "Completion Rate (%)"
  ) +
  theme_minimal()
```

```
## Warning: Removed 14540 rows containing non-finite outside the scale range
## ('stat_boxplot()').
```

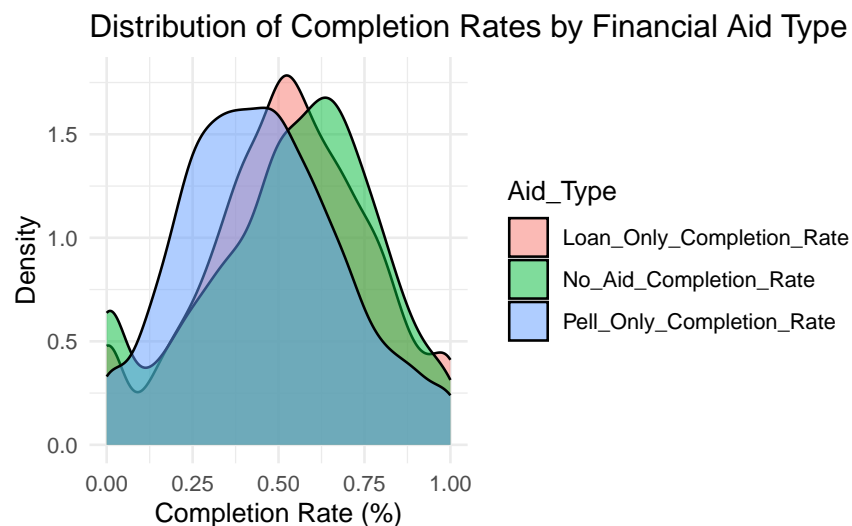


This boxplot compares the completion rates for students with no aid, students receiving only loans, and students receiving only Pell grants. The purpose of this graph is to visualize the distribution of completion rates across different financial aid types to understand how these aid categories influence student outcomes.

In the box plot, the median completion rate for the Pell grant category is lower compared to the other financial aid types (“No Aid” and “Loan Only”). This suggests that colleges with a higher proportion of Pell grant recipients may have lower overall completion rates. The spread of the data (IQR and whiskers) also indicates a larger variance in completion rates for Pell grant recipients, showing a wider range of outcomes among these colleges.

```
college_reduced %>%
  pivot_longer(
    cols = c(No_Aid_Completion_Rate,
              Loan_Only_Completion_Rate,
              Pell_Only_Completion_Rate),
    names_to = "Aid_Type",
    values_to = "Completion_Rate"
  ) %>%
  ggplot(aes(x = Completion_Rate, fill = Aid_Type)) +
  geom_density(alpha = 0.5) +
  labs(
    title = "Distribution of Completion Rates by Financial Aid Type",
    x = "Completion Rate (%)",
    y = "Density"
  ) +
  theme_minimal()
```

```
## Warning: Removed 14540 rows containing non-finite outside the scale range
## ('stat_density()').
```



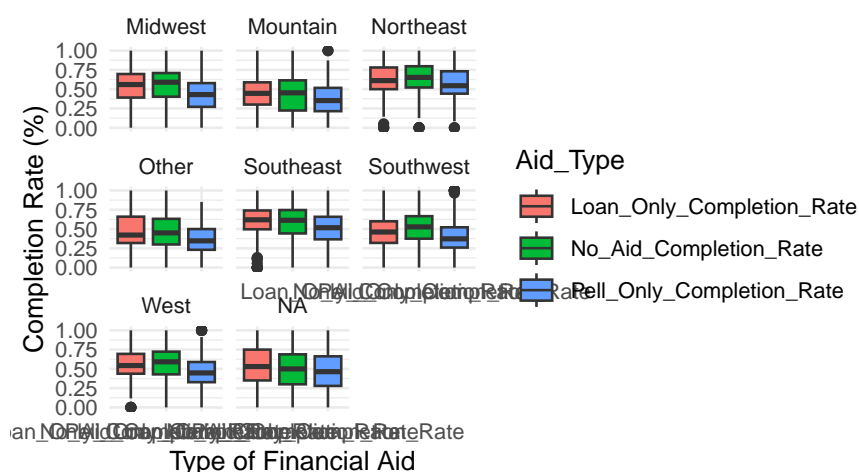
This density plot illustrates the distribution of completion rates for each financial aid type. It is used to understand the spread and concentration of completion rates across the different financial aid groups.

In the density plot, the density curve for Pell grant recipients is shifted to the left and right skewed, further indicating lower completion rates. This pattern implies that, on average, colleges with more Pell grant recipients experience lower completion rates compared to those with “No Aid” or “Loan Only” students.

```
college_reduced %>%
  pivot_longer(
    cols = c(No_Aid_Completion_Rate,
              Loan_Only_Completion_Rate,
              Pell_Only_Completion_Rate),
    names_to = "Aid_Type",
    values_to = "Completion_Rate"
  ) %>%
  ggplot(aes(x = Aid_Type, y = Completion_Rate, fill = Aid_Type)) +
  geom_boxplot() +
  facet_wrap(~Region_Label) +
  labs(
    title = "Completion Rates by Financial Aid Type and Region",
    x = "Type of Financial Aid",
    y = "Completion Rate (%)"
  ) +
  theme_minimal()
```

```
## Warning: Removed 14540 rows containing non-finite outside the scale range
## ('stat_boxplot()').
```

Completion Rates by Financial Aid Type and Region



This facet-wrapped boxplot provides a more granular view of completion rates by financial aid type within different regions. It is intended to explore whether the influence of financial aid on student outcomes varies by geographic region.

Higher completion rates for Pell grant recipients and more variability for students without aid are consistent across different regions, as shown in the facet-wrapped boxplot. However, each region experiences these changes to varying degrees. For example, the advantages of Pell grants in raising completion rates appear to be more noticeable in some areas than others. This implies that financial aid and regional factors may interact to affect student success. Greater variation in completion rates is shown by the broader IQRs in some locations for students who solely received loans; this could be the result of other regional factors that the model did not account for.

Summary Statistics

```
college_reduced %>%
  group_by(Region_Label) %>%
  summarize(Count = n()) %>%
  arrange(desc(Count))
```

Region_Label	Count
Southwest	1778
NA	1143
Southeast	1118
Midwest	1041
Mountain	740
West	583
Northeast	389
Other	266

The first summary provides the count of institutions in each region labeled by Region_Label, arranged by the highest count. This breakdown helps to understand the distribution of data across different regions.

```
college_reduced %>%
  pivot_longer(
    cols = c(No_Aid_Completion_Rate,
              Loan_Only_Completion_Rate, Pell_Only_Completion_Rate),
    names_to = "Aid_Type",
    values_to = "Completion_Rate"
  ) %>%
  group_by(Region_Label, Aid_Type) %>%
  summarize(
    Count = n(),
    Mean = mean(Completion_Rate, na.rm = TRUE),
    Median = median(Completion_Rate, na.rm = TRUE),
    Range = range(Completion_Rate, na.rm = TRUE) %>% diff(),
    SD = sd(Completion_Rate, na.rm = TRUE),
    IQR = IQR(Completion_Rate, na.rm = TRUE)
  ) %>%
  arrange(Region_Label, Aid_Type)
```

'summarise()' has grouped output by 'Region_Label'. You can override using the
'.groups' argument.

Region_Label	Aid_Type	Count	Mean	Median	Range	SD	IQR
Midwest	Loan_Only_Completion_Rate	4041	0.5274953	0.56000	1.0000	0.2434785	0.305700
Midwest	No_Aid_Completion_Rate	4041	0.5447707	0.59030	1.0000	0.2473406	0.307200
Midwest	Pell_Only_Completion_Rate	4041	0.4384048	0.42985	1.0000	0.2238941	0.307875
Mountain	Loan_Only_Completion_Rate	740	0.4474333	0.44555	1.0000	0.2306626	0.288650
Mountain	No_Aid_Completion_Rate	740	0.4249395	0.45260	1.0000	0.2506290	0.391550
Mountain	Pell_Only_Completion_Rate	740	0.3800106	0.35225	1.0000	0.2060053	0.302600
Northeast	Loan_Only_Completion_Rate	389	0.6265848	0.61110	1.0000	0.2184760	0.281700
Northeast	No_Aid_Completion_Rate	389	0.6395427	0.65270	1.0000	0.2200819	0.275675
Northeast	Pell_Only_Completion_Rate	389	0.5747423	0.54495	1.0000	0.2301194	0.290550
Other	Loan_Only_Completion_Rate	266	0.4829076	0.42190	1.0000	0.2567688	0.341850
Other	No_Aid_Completion_Rate	266	0.4541831	0.45000	1.0000	0.2703613	0.332500
Other	Pell_Only_Completion_Rate	266	0.3636185	0.34680	0.8511	0.2030261	0.268700
Southeast	Loan_Only_Completion_Rate	1118	0.6020280	0.61970	1.0000	0.2071035	0.243950
Southeast	No_Aid_Completion_Rate	1118	0.5620106	0.61335	1.0000	0.2566842	0.303575
Southeast	Pell_Only_Completion_Rate	1118	0.5108619	0.51710	1.0000	0.2285273	0.292400
Southwest	Loan_Only_Completion_Rate	778	0.4566171	0.46150	1.0000	0.2302495	0.280050
Southwest	No_Aid_Completion_Rate	778	0.5095960	0.52780	1.0000	0.2435499	0.291700
Southwest	Pell_Only_Completion_Rate	778	0.3995128	0.37265	1.0000	0.2100650	0.263050
West	Loan_Only_Completion_Rate	583	0.5605205	0.54290	1.0000	0.2153546	0.257400

Region_Label	Aid_Type	Count	Mean	Median	Range	SD	IQR
West	No_Aid_Completion_Rate	583	0.5475987	0.59150	1.0000	0.2595946	0.294650
West	Pell_Only_Completion_Rate	583	0.4662396	0.44760	1.0000	0.2135967	0.262150
NA	Loan_Only_Completion_Rate	1143	0.5301765	0.53020	1.0000	0.2740080	0.400000
NA	No_Aid_Completion_Rate	1143	0.4904753	0.50000	1.0000	0.2625052	0.388950
NA	Pell_Only_Completion_Rate	1143	0.4764567	0.46380	1.0000	0.2451394	0.383450

The second summary calculates key statistical measures, including the mean, median, range, standard deviation (SD), and interquartile range (IQR), for each aid type across different regions. These statistics provide insights into the variability and central tendencies of completion rates, helping to identify differences and commonalities across regions and aid types for better decision-making regarding aid strategies.

Data Analysis

```
completion_model <- lm(No_Aid_Completion_Rate ~
  Loan_Only_Completion_Rate + Pell_Only_Completion_Rate,
  data = college_reduced)

model_summary <- tidy(completion_model)

model_summary
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.2178695	0.0105978	20.557934	0
Loan_Only_Completion_Rate	0.1532803	0.0265427	5.774849	0
Pell_Only_Completion_Rate	0.5419959	0.0288964	18.756547	0

According to the data, Pell grant recipients' completion rates are more strongly correlated (coefficient 0.542) with those of students who did not receive any aid than with those who only received loans (coefficient 0.153). These correlations are unlikely to be the result of chance because both predictors have very low p-values and are highly statistically significant. The intercept indicates that some students achieve without financial assistance, but at lower rates, with a baseline completion rate of 21.79% for those without aid. Overall, the results highlight the strong correlation between financial aid—especially Pell grants—and better outcomes for those who do not receive it, suggesting possible systemic implications.

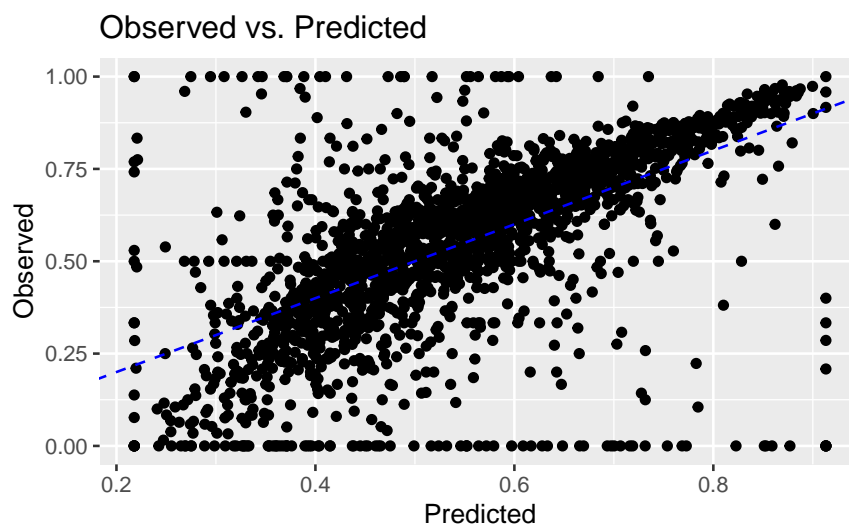
```
model_performance <- glance(completion_model)
model_performance
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	cdf.residual	nobs
0.3784722	0.3778574	0.1867288	615.6368	0	2	526.2987	-1044.597	-1022.144	70.5024	2022	2025

The model explains 37.85% of the variation in no-aid completion rates, as shown by the adjusted R^2 , indicating that Pell grant and loan-only completion rates are meaningful predictors. The model is statistically significant overall, with an F-statistic of 615.64 and an extremely low p-value, confirming the predictors' strong contribution. The residual standard deviation ($\sigma = 0.187$) indicates the average prediction error, while the AIC (-1044.60) and BIC (-1022.14) suggest a good model fit. With 2025 observations, the model captures significant variability but leaves some unexplained, pointing to potential contributions from additional factors.

```
college_clean <- college_reduced %>%
  drop_na(No_Aid_Completion_Rate,
    Loan_Only_Completion_Rate,
    Pell_Only_Completion_Rate)

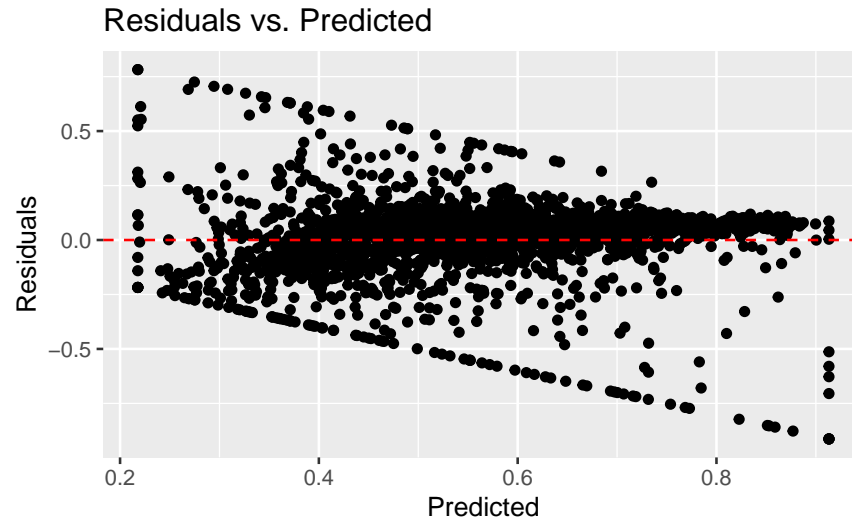
completion_model <- lm(No_Aid_Completion_Rate ~
  Loan_Only_Completion_Rate + Pell_Only_Completion_Rate,
  data = college_clean)
ggplot(data = data.frame(Observed = college_clean$No_Aid_Completion_Rate,
  Predicted = predict(completion_model)),
  aes(x = Predicted, y = Observed)) +
  geom_point() +
  geom_abline(color = "blue", linetype = "dashed") +
  labs(title = "Observed vs. Predicted", x = "Predicted", y = "Observed")
```



The “Observed vs. Predicted” plot shows a linear relationship between observed and predicted completion rates, demonstrating that the model captures the overall trend. The points are spread

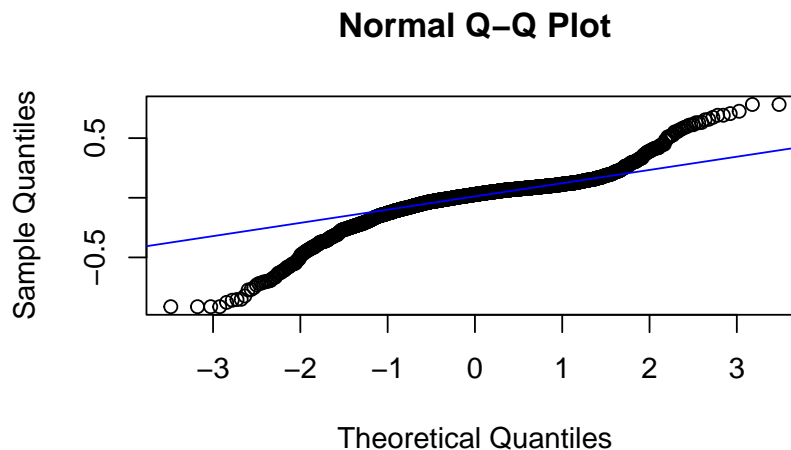
consistently around the line of equality, but there is some widening at lower to median predicted values, indicating potential areas where the model's accuracy could be improved.

```
ggplot(data = data.frame(Residuals = resid(completion_model),
                          Predicted = predict(completion_model)),
       aes(x = Predicted, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs. Predicted", x = "Predicted", y = "Residuals")
```



The residual vs. predicted plot reveals that the residuals are roughly evenly scattered around zero across the range of predicted values. This suggests that the assumption of constant variance holds, though any noticeable funnel shape would indicate that the residuals are not constant.

```
qqnorm(resid(completion_model))
qqline(resid(completion_model), col = "blue")
```



The Q-Q plot displays residuals that fall close to the diagonal reference line, indicating that the residuals are approximately normally distributed. Any significant deviations would suggest a violation of this assumption.

Conclusion

Several important insights into the variables affecting college completion rates for both financially aided and unfinancial students are provided by the analyses in this project. Financial help is clearly a major factor in raising student completion rates, as shown by the summary statistics and visualizations. The significant benefits of financial help are demonstrated by the greater completion rates among Pell grant recipients when compared to those who only received loans. This finding is supported by the linear model, which demonstrates that while Pell awards and loans are both important predictors of student outcomes, their correlation is much stronger. These results are supported by the diagnostic plots, which demonstrate that the linear model assumptions are satisfied. The data already present in the “Grants only” and “Loans only” categories is essentially combined in the “Both” category. There would be no new information or unique data regarding the impact of financial help on completion rates if it were included. Instead, by overlapping with the knowledge of the other variables, it may weaken their impact.

However, it is important to remember that the dataset is designed so that each row represents an entire institution or university, rather than individual students. The capacity to immediately infer information about certain student groupings from the data is limited by this aggregate. Nevertheless, because the data is aggregated, it is possible to draw conclusions at the institutional level, indicating that institutions with larger percentages of Pell grant applicants might have higher overall completion rates. The conclusions that financial aid—particularly Pell grants—is essential for raising student achievement are supported by the analyses that follow one another. This research has significant implications for legislators and academic institutions, highlighting the necessity of strengthening and growing financial aid programs to improve overall student retention and completion rates.