IDS 702 HW 3

Eric Ortega Rodriguez

Load data

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(dplyr)

college <- read.csv("https://raw.githubusercontent.com/anlane611/datasets/refs/heads/main/col
```

Exercise 1

a. **What is the sample size? What does each observation represent?**

The data set contains 797 rows and 39 columns, with each row representing a college or university. There are 39 variables.

```
#commented out to save space
#glimpse(college)
```

b. **How many, if any, missing observations are in each of the following variables: grad_100_value, med_sat_value, endow_value, student_count, basic, control. Go ahead and create a subset that excludes observations with missing data in those variables (note: not the whole dataset).**

Missing values are the following:

- `grad_100_value`: 7 missing values
- `med_sat_value`: 168 missing values
- `endow_value`: 71 missing values
- `student_count`: 0 missing values
- `basic`: 0 missing values
- `control`: 0 missing values

```
missing_col_counts <- college %>%
  summarise(
   grad_100_value_missing = sum(is.na(grad_100_value)),
    med_sat_value_missing = sum(is.na(med_sat_value)),
    endow_value_missing = sum(is.na(endow_value)),
    student_count_missing = sum(is.na(student_count)),
    basic_missing = sum(is.na(basic)),
    control_missing = sum(is.na(control))
  )
missing_col_counts
```

```
  grad_100_value_missing med_sat_value_missing endow_value_missing
1                      7                   168                  71
  student_count_missing basic_missing control_missing
1                     0             0               0
```

```
college_clean <- college %>%
  filter(!is.na(grad_100_value),
         !is.na(med_sat_value),
         !is.na(endow_value),
         !is.na(student_count),
         !is.na(basic),
         !is.na(control))
```

c. There are four different categorical variables for `basic` are:

1. `Baccalaureate Colleges--Arts & Sciences`
2. `Baccalaureate Colleges--Diverse Fields`

3. `Research Universities--high research activity`
4. `Research Universities--very high research activity`

The filled out table can be seen below.

```
basic_summary <- college_clean %>%
  group_by(basic) %>%
  summarise(Count = n()) %>%
  mutate(Proportion = Count / sum(Count))
basic_summary
```

```
# A tibble: 4 x 3
  basic                                              Count Proportion
  <chr>                                              <int>      <dbl>
1 Baccalaureate Colleges--Arts & Sciences              195      0.317
2 Baccalaureate Colleges--Diverse Fields               231      0.376
3 Research Universities--high research activity         84      0.137
4 Research Universities--very high research activity   105      0.171
```

Level	Count (n)	Proportion or %
Baccalaureate Colleges–Arts & Sciences	195	0.3171
Baccalaureate Colleges–Diverse Fields	231	0.3756
Research Universities–high research activity	84	0.1369
Research Universities–very high research activity	105	0.1707

d. There are two different categorical variables for `control`:

1. `Public`
2. `Private not-for-profit`

```
control_summary <- college_clean %>%
  group_by(control) %>%
  summarise(Count = n()) %>%
  mutate(Proportion = Count / sum(Count))


control_summary
```

```
# A tibble: 2 x 3
  control                Count Proportion
  <chr>                  <int>      <dbl>
1 Private not-for-profit   409      0.665
2 Public                   206      0.335
```

Level	Count (n)	Proportion or %
Private not-for-profit	409	0.6650
Public	206	0.3349

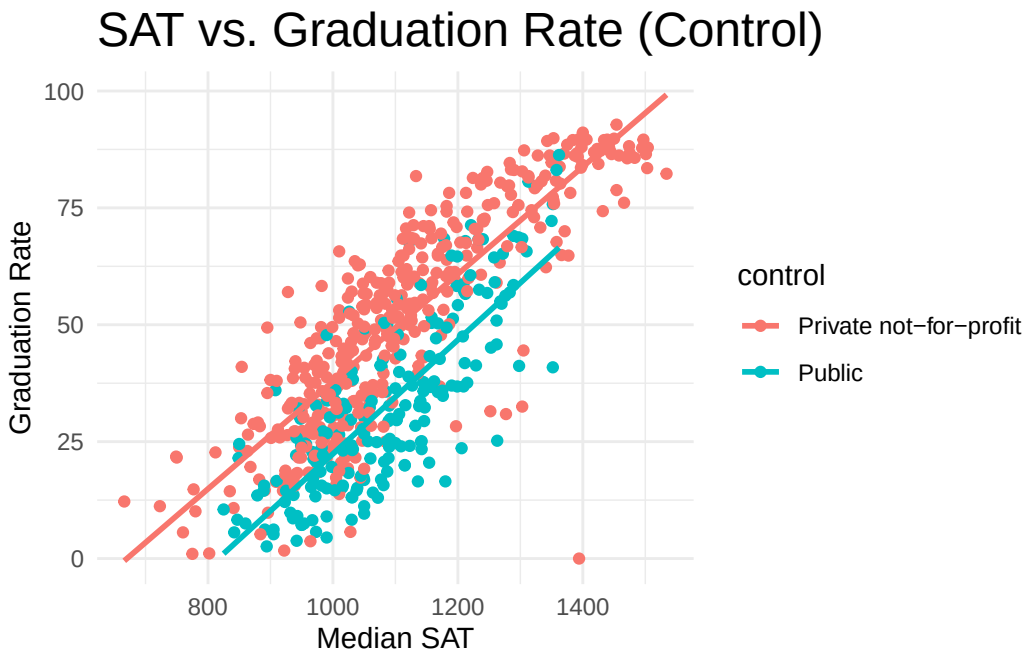
e.Graphs are shown below individually to improve legibility.

1. **Median SAT score (med_sat_value), colored by private vs. public institution (control)**: Private institutions appear to have higher median SAT scores compared to public institutions.
2. **Median SAT score (med_sat_value) colored by institution type (basic)**: Research Universities tend to have higher median SAT scores compared to Baccalaureate Colleges.
3. **Endowment value (endow_value) colored by private vs. public institution (control)**: Private institutions tend to have higher endowment values than public institutions.
4. **Endowment value (endow_value) colored by institution type (basic)**: Research Universities tend have larger endowment values compared to Baccalaureate Colleges.
5. **Enrollment total (student_count) colored by private vs. public institution (control)**: Public institutions tend to have higher enrollment totals compared to private institutions.
6. **Enrollment total (student_count) colored by institution type (basic)**: Research Universities tend to have significantly higher enrollments compared to Baccalaureate Colleges.

```
# Load necessary libraries
library(tidyverse)

plot1 <- ggplot(college_clean, aes(x = med_sat_value,
                                   y = grad_100_value,
                                   color = control)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
```

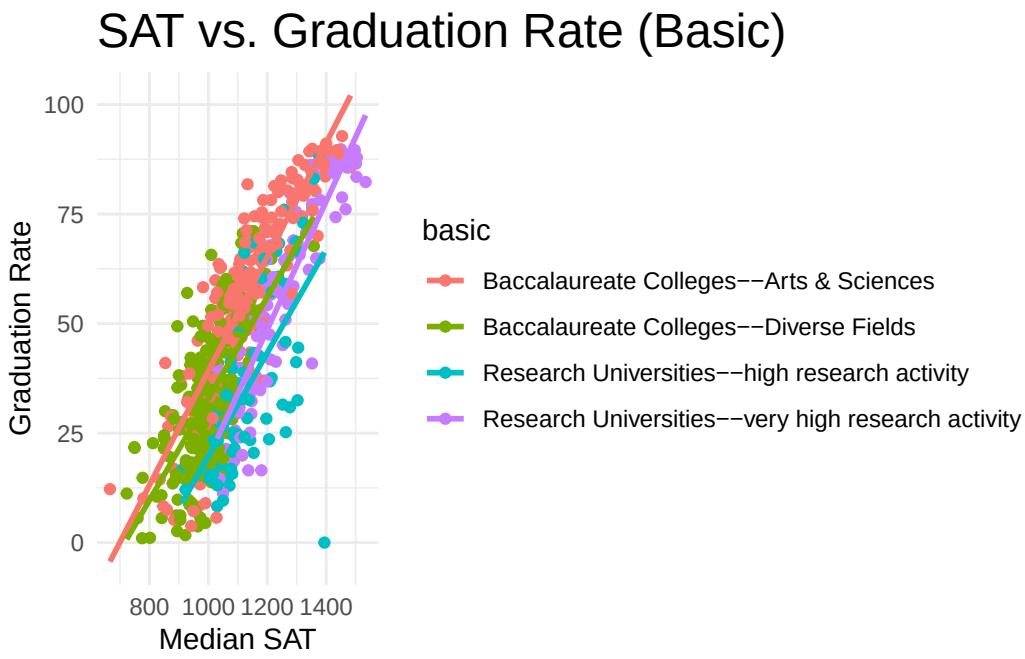
```
  labs(title = "SAT vs. Graduation Rate (Control)",
       x = "Median SAT", y = "Graduation Rate") +
  theme_minimal() +
  theme(plot.title = element_text(size = 18))

plot1
```

```
`geom_smooth()` using formula = 'y ~ x'
```



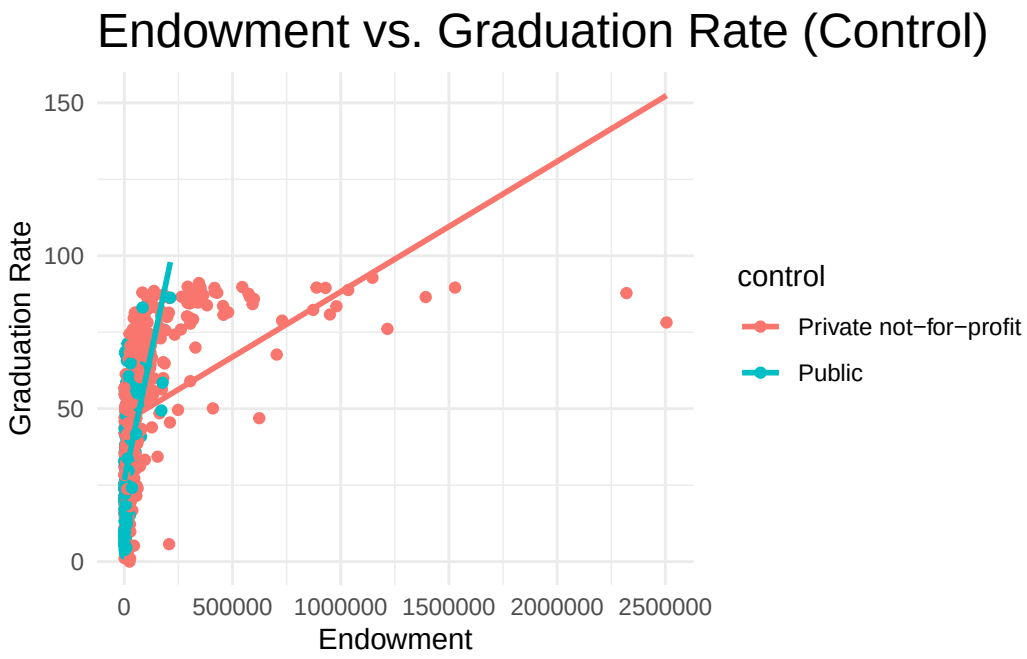

```
plot2 <- ggplot(college_clean, aes(x = med_sat_value,
                                   y = grad_100_value, color = basic)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title = "SAT vs. Graduation Rate (Basic)", x = "Median SAT",
       y = "Graduation Rate") +
  theme_minimal() +
  theme(plot.title = element_text(size = 18))

plot2
```

```
`geom_smooth()` using formula = 'y ~ x'
```



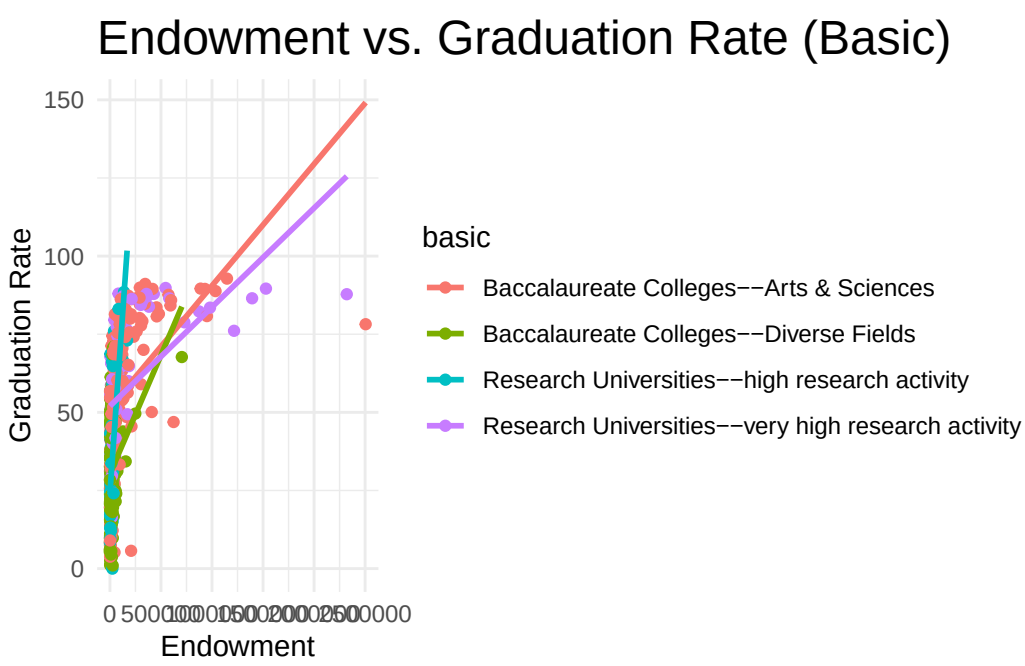

```
plot3 <- ggplot(college_clean, aes(x = endow_value,
                                   y = grad_100_value,
                                   color = control)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title = "Endowment vs. Graduation Rate (Control)",
       x = "Endowment",
       y = "Graduation Rate") +
  theme_minimal() +
  theme(plot.title = element_text(size = 18))

plot3
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
plot4 <- ggplot(college_clean, aes(x = endow_value,
                                   y = grad_100_value,
                                   color = basic)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title = "Endowment vs. Graduation Rate (Basic)",
       x = "Endowment",
       y = "Graduation Rate") +
  theme_minimal() +
  theme(plot.title = element_text(size = 18))

plot4
```

```
`geom_smooth()` using formula = 'y ~ x'
```

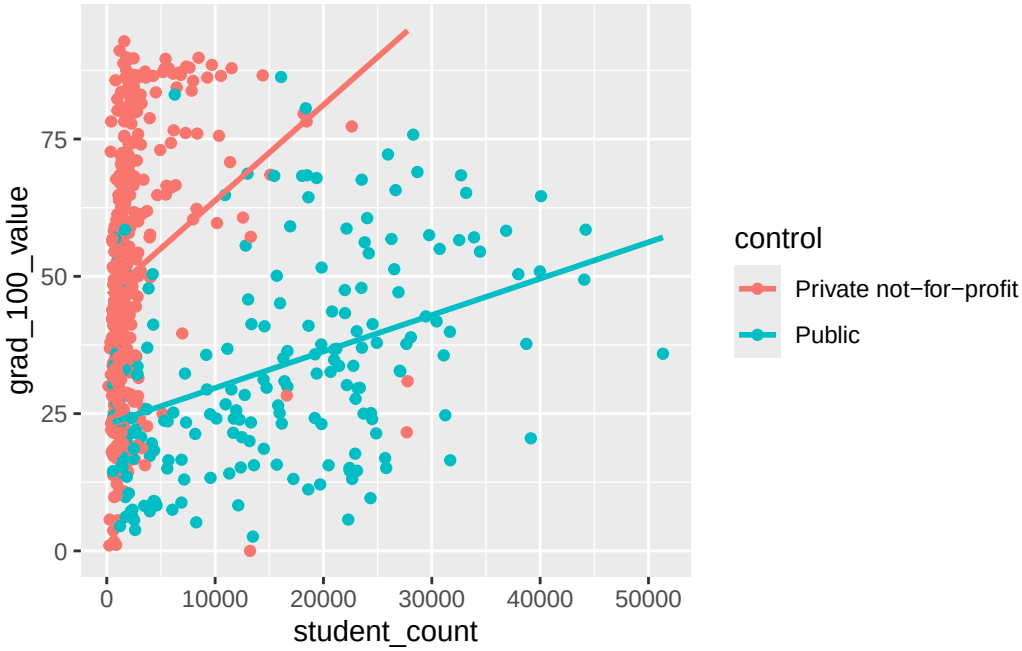



```
plot5 <- ggplot(college_clean, aes(x = student_count,
                                   y = grad_100_value,
                                   color = control)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE)
  labs(title = "Enrollment vs. Graduation Rate (Control)",
       x = "Enrollment",
       y = "Graduation Rate") +
  theme_minimal() +
  theme(plot.title = element_text(size = 18))
```

```
NULL
```

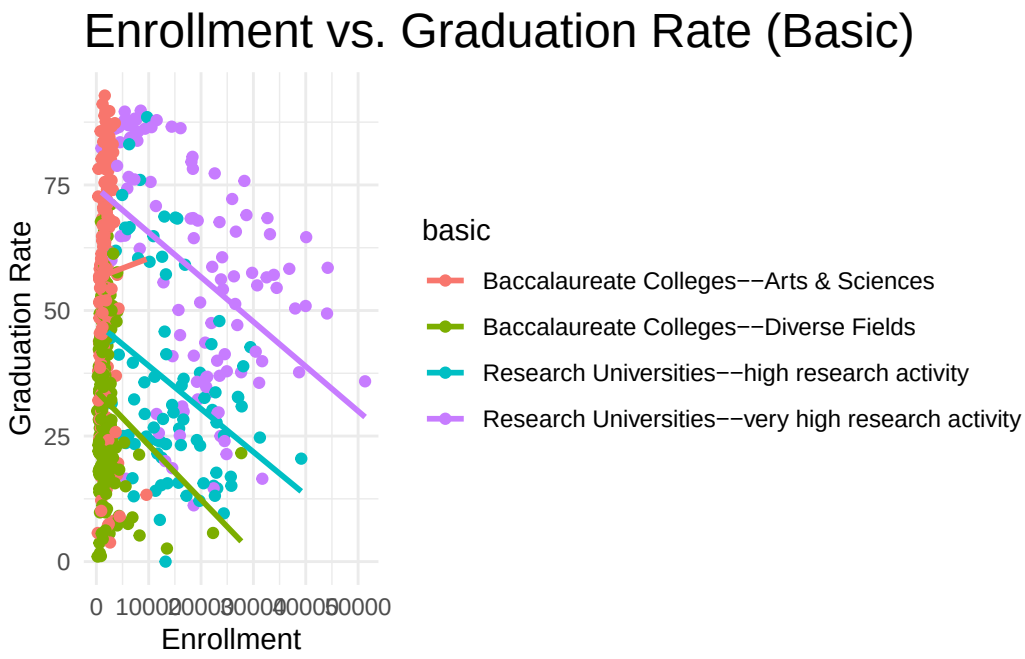
```
  plot5
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
plot6 <- ggplot(college_clean, aes(x = student_count,
                                   y = grad_100_value,
                                   color = basic)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title = "Enrollment vs. Graduation Rate (Basic)",
       x = "Enrollment",
       y = "Graduation Rate") +
  theme_minimal() +
  theme(plot.title = element_text(size = 18))

plot6
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Exercise 2

b.**Considering both the number of observations in each level and the model interpretation, decide whether or not you should combine the levels of institution types into 2 levels: baccalaureate colleges and research universities. Justify your choice.**

I think that in this case, it would be fine to combine the observations into two levels: baccalaureate colleges and research universities. This is because we do not have a large data set/number of observations. In addition to this, it would allow for a more simple model.

b.**Write the theoretical model based on your decision in part a.**

$$\text{Grad}_i = \beta_0 + \beta_1 \cdot \text{Endowment}_i + \beta_2 \cdot \text{Type}_i + \epsilon_i$$

Meaning:Grad(i) represents the graduation rate for institution, Endowment(i) represents the endowment value of the institution (i), Type(i) is a binary value which indicates which institution (0 for Baccalaureate Colleges and 1 for Research Universities), 0 represents the intercept of the model, 1 represents the effect of endowment on graduation rates, next is 2 which is the effect of being a research organization on graduation rates compared to baccalaureate colleges, and last, (i) is the error term.

c. **Fit the model in R and display the summary table. Write the full fitted model and the fitted model for each institution type.**

```
library(ggplot2)
library(dplyr)

college_clean$institution_type <- ifelse(college_clean$basic %in% c("Baccalaureate Colleges--
                                                                    "Baccalaureate Colleges--Diverse Fields
                                         0,
                                         1)
model_one <- lm(grad_100_value ~ endow_value + institution_type, data = college_clean)
summary(model_one)
```

```
Call:
lm(formula = grad_100_value ~ endow_value + institution_type,
    data = college_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-88.279 -15.707  -2.138  16.069  41.601

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.960e+01  1.025e+00  38.633   <2e-16 ***
endow_value      5.064e-05  3.806e-06  13.306   <2e-16 ***
institution_type 2.560e+00  1.775e+00   1.443     0.15
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.28 on 612 degrees of freedom
Multiple R-squared:  0.229, Adjusted R-squared:  0.2265
F-statistic:  90.9 on 2 and 612 DF,  p-value: < 2.2e-16
```

Fitted Model:

$$\text{Grad}_i = 3.960 \times 10^1 + 5.064 \times 10^{-5} \cdot \text{Endowment}_i + 2.560 \cdot \text{Type}_i$$

Fitted Model for Baccalaureate Institutions:

$$\text{Grad}_{\text{Bacc}} = 3.960 \times 10^1 + 5.064 \times 10^{-5} \cdot \text{Endowment}_i$$

Fitted Model for Research Universities:

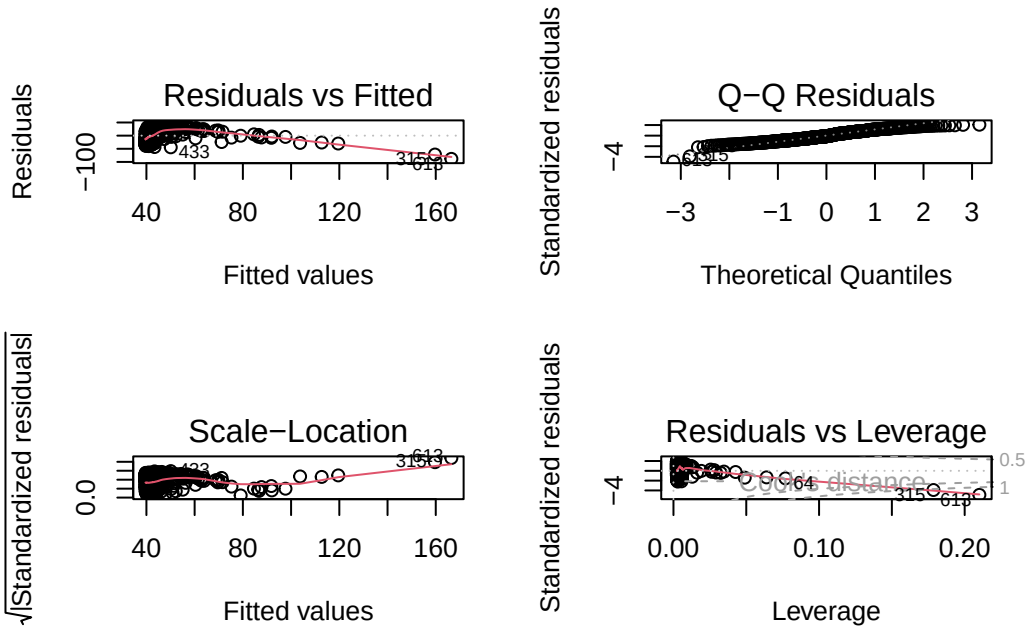
$$\text{Grad}_{\text{Research}} = 3.960 \times 10^{1} + 2.560 + 5.064 \times 10^{-5} \cdot \text{Endowment}_{i}$$

d.**Generate the diagnostic plots for your model. Comment on what you observe in the residuals vs fitted plot and the QQ-plot. Based on the plot you generated in #1, are these diagnostic plots surprising? Why or why not?**

The Residuals vs Fitted Plot shows a curved pattern, particularly at higher fitted values, which suggests potential non-linearity in the model. Regarding the Q-Q Plot, the deviation from the diagonal line shown in the graph at the ends indicates that the residuals are not perfectly normally distributed.

This is not surprising, as these patterns support what we observed in #1. Specifically, it's expected because we are fitting a simplified linear model, whereas, in reality, graduation rates are likely influenced by many complex factors not captured by the model. Graduation rates are affected by numerous variables beyond those included in this model, which contributes to the non-linearity and deviations in the residuals.

```
par(mfrow = c(2, 2))
plot(model_one)
```



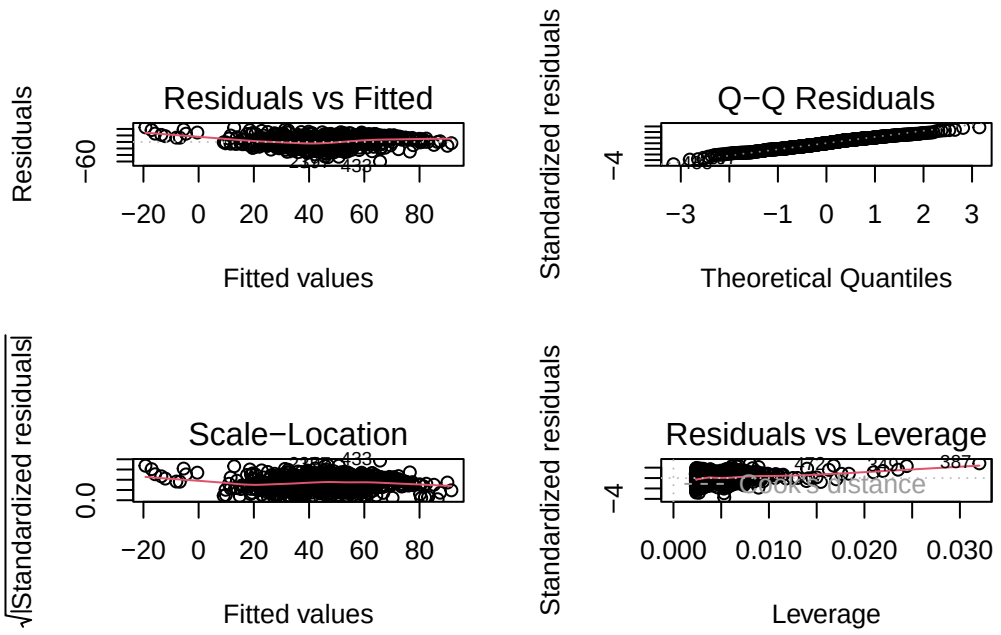
e.**Fit a model regressing graduation rate on log(endowment value) and institution type (again based on your decision in part a) and show the summary table. Generate the diagnostic plots for this model. Comment on what you observe in the residuals vs fitted plot and QQ-plot here compared to what you saw in part c.**

The Residuals vs Fitted Plot shows some improvement compared to the initial model, but there is still a slight curved pattern, particularly at higher fitted values, which suggests that non-linearity may still exist. This indicates that the log transformation of endowment value helped but did not fully resolve the issue. Next is the Q-Q Plot, while the residuals show a better alignment with the diagonal line than in the previous model, there is still deviation at the ends, indicating that the residuals are not perfectly normally distributed, though they are closer to normality than before. This was not surprising, as these patterns continue to support what we observed in the earlier model. Specifically, it's expected because even though we applied a log transformation to endowment, we are still fitting a relatively simplified linear model. Graduation rates are likely influenced by many complex factors beyond just endowment value and institution type.

```
par(mfrow=c(2,2))
model_log<- lm(grad_100_value ~ log(endow_value) + institution_type, data = college_clean)
summary(model_log)
```

```
Call:
lm(formula = grad_100_value ~ log(endow_value) + institution_type,
    data = college_clean)

Residuals:
   Min     1Q Median     3Q    Max
-60.01 -11.66  -0.04  12.30  44.86

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -54.0509     3.9329 -13.743   <2e-16 ***
log(endow_value)   9.7831     0.3866  25.305   <2e-16 ***
institution_type   2.1912     1.4083   1.556     0.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.1 on 612 degrees of freedom
Multiple R-squared:  0.5142,    Adjusted R-squared:  0.5126
F-statistic: 323.9 on 2 and 612 DF,  p-value: < 2.2e-16
```

```
plot(model_log)
```



f.**What are the adjusted R^2 values for the two models?**

The adjusted R-Squared for the first model is 0.2265014 and for the second model it is 0.5126438.

```
summary(model_one)$adj.r.squared
```

```
[1] 0.2265014
```

```
summary(model_log)$adj.r.squared
```

```
[1] 0.5126438
```

g.**Based on the results in parts e and f, which of these two models do you think is better? Write the final fitted model for each institution type based on your choice.**

The second model, which incorporated the log-transformed endowment value, appears to perform better. Applying the log transformation helped to "unclutter" the observations in the Residuals vs. Fitted plot, reducing the curvature and adjusting the residuals. Additionally, the

transformation improved the linearity in the Q-Q plot, with fewer deviations from the diagonal line, suggesting a better fit to the assumption of normality. Overall, the second model better captured the relationship between endowment and graduation rates.

Baccalaureate Colleges:

$$\widehat{\text{Grad}}_{\text{Baccalaureate}} = -54.0509 + 9.7831 \cdot \log(\text{Endowment}_i) + \epsilon$$

Research Universities:

$$\widehat{\text{Grad}}_{\text{Research}} = (-54.0509 + 2.1912) + 9.7831 \cdot \log(\text{Endowment}_i) + \epsilon$$

h.**Write an interpretation for the coefficient estimates, p-values, and 95% confidence intervals in the context of the problem.** The log of endowment value is a significant predictor of graduation rates based on the model above. For each unit increase in the log of endowment value, the expected graduation rate increases by 9.78. The p-value is less than 2e-16, which indicates that this result is highly statistically significant. As for the 95% confidence interval, it is between 9.05 and 10.58. This highlights an accurate estimate. Overall, it shows that larger endowments generally have higher graduation rates, likely due to better financial resources that can improve the overall quality of education. On the other hand, the coefficient for institution type (Research Universities vs. Baccalaureate Colleges) is 2.19 percentage points, indicating that Research Universities are predicted to have graduation rates that are higher than Baccalaureate Colleges. However, this result was shown to not be statistically significant (p-value of 0.12). The 95% confidence interval for this estimate is between -0.57 and 4.97. This includes zero, making it not accurate.

Exercise 3

a.**Decide which interaction term(s) for private vs public institution to include in your model, and justify your decision (you must include at least one interaction term). You should base your decision on the exploratory analysis you conducted in #1.**

In my exploratory analysis from #1, it can be seen that private institutions typically have larger endowments, which can significantly influence graduation rates by providing better resources to students. The relationship between endowment and graduation rates appears to differ between public and private institutions, as indicated by the intersecting slopes in the graphs from exercise one. This suggests that the effect of endowment on graduation rates is not uniform across institution types.Including an interaction term between private/public institutions and endowment in the model is needed to fully capture how endowment impacts graduation rates differently. Private institutions may benefit more from larger endowments

than public ones, as they tend to rely more heavily on endowment funding to improve student outcomes. This interaction term will allow for a more accurate representation of how endowment affects graduation rates based the different institution type.

b.**Write the full theoretical model for a model regressing graduation rate on median SAT score, endowment value, total enrollment, private vs public institution, and the interaction term(s) you selected in part a. Be sure to define the predictor terms (i.e., median SAT score). Then, write the separate theoretical models for private and public institutions.**

Fitted Model:

$$\text{grad\_rate}_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \beta_5 \cdot (x_2 \times x_4) + \epsilon_i$$

Public Institution:

$$\text{grad\_rate}_{\text{public}} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \epsilon_i$$

Private Institution:

$$\text{grad\_rate}_{\text{private}} = \beta_0 + \beta_1 \cdot x_1 + (\beta_2 + \beta_5) \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 + \epsilon_i$$

Definitions:

- (beta_0): The intercept
- (beta_1): Median SAT score
- (beta_2): Endowment value
- (beta_3): Total enrollment
- (beta_4): Private institution vs public institution
- (beta_5): Endowment value and institution type (private institution vs public institution).
- (epsilon): Error term

c.**Fit your model, show the summary table, and generate the diagnostic plots. Comment on what you observe. Specifically, address the following:**

- **Based on the diagnostic plots, does the linearity assumption appear to be violated? If so, how could the model be improved?** Regarding linearity, Residuals vs Fitted plot highlight a random scatter, which is a good indication that the linearity assumption is largely met. However, there are some areas where non linearity can be seen. This could be improved my adding additional terms to capture non-linear data.

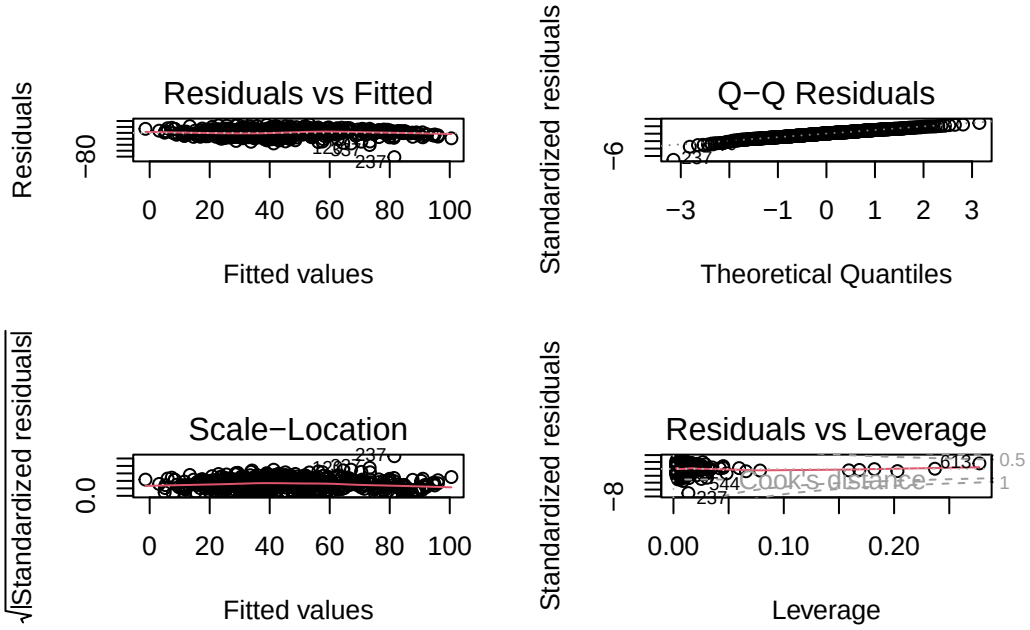
- **Based on the diagnostic plots, does the normality assumption appear to be violated? If so, how could the model be improved?** The Q-Q plot shows deviation from the diagonal line in both tails, indicating some deviation from the assumption of normally distributed residuals. This can be improved by log-transforming the graduation rate.
- **Based on the diagnostic plots, does the homoscedasticity assumption appear to be violated? If so, how could the model be improved?** The Scale-Location plot (standardized residuals vs fitted values) shows a fairly flat red line, which suggests that the variance of residuals is relatively constant across different fitted values. However, fanning can be seen which shows minor heteroscedasticity. A way to adjust this would be to use the weighted least squares regression which would help with the variance in the residuals.
- **Do you notice any residual values that stand out? If so, investigate the relevant observation(s). Are there data entries that appear to be unusual? Consider what might be the issue here and whether or not the observation(s) should be excluded.** Observations 613 and 237 stand out as unusual data points that may be exerting undue influence on our regression results. These outliers could be skewing the model's accuracy, making it necessary to investigate further. We should first verify whether these points are the result of data entry errors or if they are extreme but valid cases. If the data is accurate, further steps will be needed to adjust the model, such as applying more robust regression techniques or transformations to mitigate the impact of these influential points.

```
par(mfrow=c(2,2))

college_clean$institution_profit_type <- ifelse(college_clean$control %in% c("Public"),
                                0,
                                1)
updated_model <- lm(grad_100_value ~ med_sat_value + student_count + institution_profit_type
summary(updated_model)
```

```
Call:
lm(formula = grad_100_value ~ med_sat_value + student_count +
    institution_profit_type * endow_value, data = college_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-81.544  -7.267   0.632   7.840  33.468

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                          -9.364e+01  4.007e+00 -23.366  < 2e-16 ***
med_sat_value                         1.177e-01  3.963e-03  29.705  < 2e-16 ***
student_count                        -2.244e-04  7.426e-05  -3.021 0.002622 **
institution_profit_type               1.404e+01  1.413e+00   9.940  < 2e-16 ***
endow_value                           1.070e-04  2.953e-05   3.625 0.000313 ***
institution_profit_type:endow_value -1.073e-04  2.936e-05  -3.655 0.000279 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.69 on 609 degrees of freedom
Multiple R-squared:  0.7452,    Adjusted R-squared:  0.7431
F-statistic: 356.2 on 5 and 609 DF,  p-value: < 2.2e-16
```

```
plot(updated_model)
```

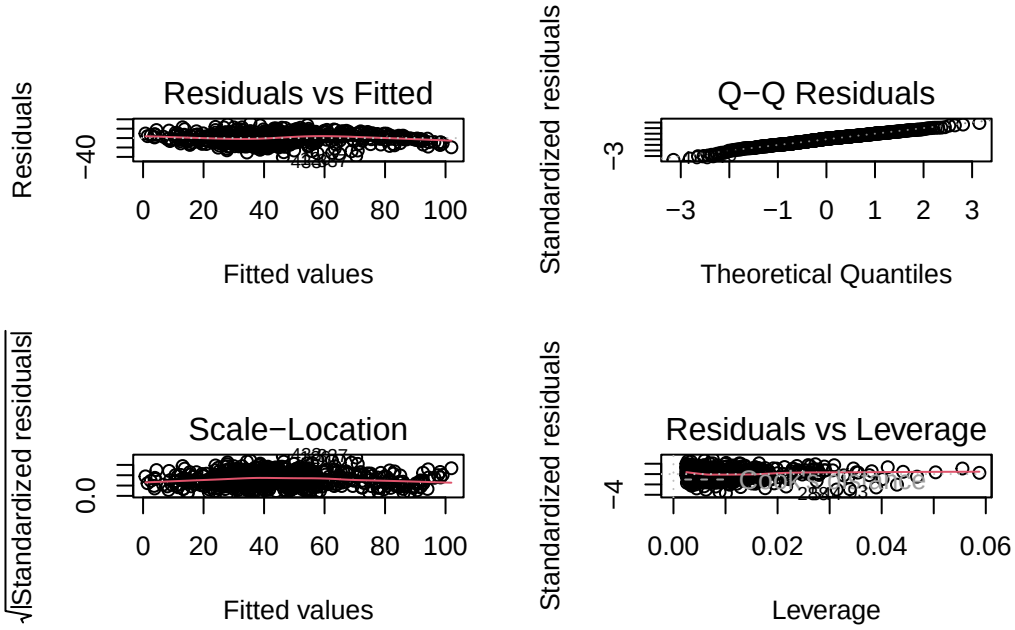


d.**Fit your model again, this time incorporating any changes you made based on your answers to part c. Show the summary table. Write the fitted model for private and public institutions. Generate the diagnostic plots for this model and comment on the difference(s) you observe compared to part c. Has the model improved?**

```r
college_clean <- college_clean[!(rownames(college_clean)
                                 %in% c("237", "613")), ]

model_d <- lm(grad_100_value ~ med_sat_value * control +
                log(endow_value) + student_count, data = college_clean)

summary(model_d)
```

```
Call:
lm(formula = grad_100_value ~ med_sat_value * control + log(endow_value) +
    student_count, data = college_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-40.510  -7.584   1.086   7.416  32.106

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -8.711e+01  3.919e+00 -22.230  < 2e-16 ***
med_sat_value               1.043e-01  4.293e-03  24.306  < 2e-16 ***
controlPublic              -2.357e+01  8.469e+00  -2.784  0.00554 **
log(endow_value)            2.133e+00  4.099e-01   5.204 2.67e-07 ***
student_count              -2.059e-04  7.611e-05  -2.706  0.00700 **
med_sat_value:controlPublic  1.296e-02  8.021e-03   1.616  0.10656
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.05 on 607 degrees of freedom
Multiple R-squared:  0.7706,    Adjusted R-squared:  0.7687
F-statistic: 407.8 on 5 and 607 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(model_d)
```



Public Institution:

$$\hat{\text{Grad}}_{\text{public}} = -87.11 + 1.043 \cdot \text{SAT}_i + 2.133 \cdot \log(\text{Endowment}_i) - 0.206 \cdot \text{Enrollment}_i + \epsilon$$

Private Institution:

$$\hat{\text{Grad}}_{\text{private}} = (-87.11 - 2.357) + (1.043 + 1.296) \cdot \text{SAT}_i + 2.133 \cdot \log(\text{Endowment}_i) - 0.206 \cdot \text{Enrollment}_i + \epsilon$$

Based on diagnostic plots, we can see that the model has improved. Based on the graph above,I can see a more random scatter in the Residuals vs Fitted plot. Closer fit to the diagonal line in the Q-Q plot.A consistent spread of residuals in the Scale-Location plot. Finally, within the The Residuals vs Leverage plot we can see that the outliers from the previous model have been reduced.

e.**Conduct a nested F test to assess whether or not your interaction term(s), as a whole, significantly contribute to the model. What do you conclude based on this test?**

```
with_inter <- lm(grad_100_value ~ med_sat_value * control +
                   log(endow_value) + student_count, data = college_clean)

without_int <- lm(grad_100_value ~ med_sat_value + control +
```

```
                    log(endow_value) + student_count, data = college_clean)

anova(without_int, with_inter)
```

```
Analysis of Variance Table

Model 1: grad_100_value ~ med_sat_value + control + log(endow_value) +
    student_count
Model 2: grad_100_value ~ med_sat_value * control + log(endow_value) +
    student_count
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    608 74473
2    607 74154  1    319.12 2.6122 0.1066
```

Given the nest F test above, we fail to reject the null hypothesis. This is because our p-value (0.1066) is larger than 0.05. In conclusion, there is not enough evidence to conclude that the interaction term between SAT score and institution type significantly contributes to the model.

f.**Interpret the results of your final model. Which terms are statistically significant? Write interpretations for the coefficient estimates that are statistically significant and include 95% confidence intervals. Write an interpretation of the adjusted R^2 value for your model.**

The model shows statistically significant predictors of graduation rates. The median SAT score has a positive effect, with each one-point increase in SAT score resulting in a 0.1058 percentage point increase in the graduation rate. Additionally, public institutions tend to have lower graduation rates compared to private institutions, with public schools having graduation rates that are, on average, 10.07 percentage points lower. The log of endowment value is also a significant predictor, with an increase in the log of endowment associated with a 2.172 percentage point increase in the graduation rate. Total enrollment has a small but statistically significant negative impact, with larger enrollments associated with slightly lower graduation rates. The model explains a substantial amount of the variance in graduation rates, as indicated by the adjusted R-squared value of 0.7681, meaning approximately 77% of the variation in graduation rates is explained by the predictors included in the model. In conclusion, the significant predictors highlight the importance of institutional characteristics such as SAT scores, endowment size, and public versus private control in explaining differences in graduation rates.

```
model_f <- lm(grad_100_value ~ med_sat_value + control +
                log(endow_value) + student_count, data = college_clean)

summary(model_f)
```

```
Call:
lm(formula = grad_100_value ~ med_sat_value + control + log(endow_value) +
    student_count, data = college_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-40.337  -7.675   1.030   7.609  32.043

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -8.927e+01  3.690e+00 -24.190  < 2e-16 ***
med_sat_value     1.058e-01  4.203e-03  25.170  < 2e-16 ***
controlPublic    -1.007e+01  1.399e+00  -7.200 1.79e-12 ***
log(endow_value)  2.172e+00  4.098e-01   5.300 1.62e-07 ***
student_count    -1.515e-04  6.835e-05  -2.217    0.027 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.07 on 608 degrees of freedom
Multiple R-squared:  0.7696,    Adjusted R-squared:  0.7681
F-statistic: 507.8 on 4 and 608 DF,  p-value: < 2.2e-16
```

g.**Imagine that the president of a large public institution approaches you to inquire about which factors are related to graduation rate. Write 1-2 sentences to explain to them, in non-technical terms, the results of your analysis. Include next steps that could be taken to improve the analysis (i.e., should more data be collected? If so, what kinds of information could be helpful?). No code is needed for this question.**

Graduation rates are strongly influenced by SAT scores, endowment size, and total enrollment, with public institutions generally having lower graduation rates compared to private ones. To improve this analysis, more data would be needed. Additional factors, such as the number of commuter students, first-generation status, faculty-to-student ratios, availability of teaching assistants (TAs), and financial aid awarded, could provide deeper insights into the factors impacting graduation rates.