Graphs

Joyce Gill, Khanh Do

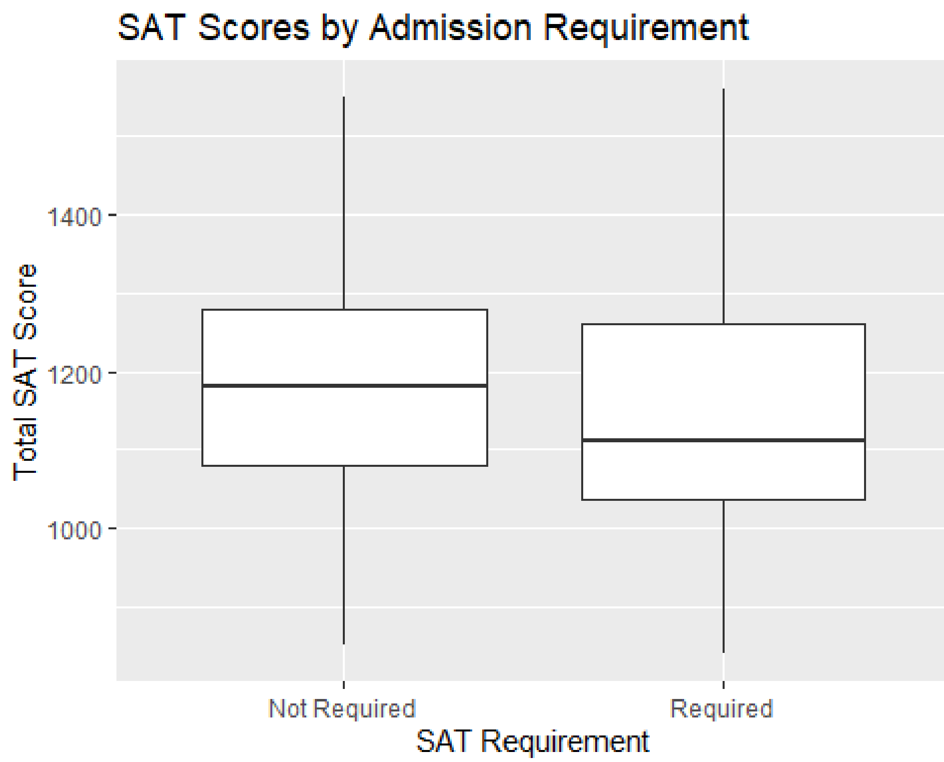
2026-02-10

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
dec,
## : EOF within quoted string
```

SAT Scores by Admission Requirement

Do schools that require SAT scores have higher average SAT scores than schools that do not require them?

```
# Create SAT requirement variable
adm_test_f2024$sat_required <- ifelse(
  adm_test_f2024$ADM2024.Admission.test.scores == "Required to be considered
for admission",
  "Required", "Not Required"
)

# Calculate total SAT score (Reading/Writing + Math)
adm_test_f2024$sat_total <-
adm_test_f2024$ADM2024.SAT.Evidence.Based.Reading.and.Writing.50th.percentile
.score +

adm_test_f2024$ADM2024.SAT.Math.50th.percentile.score

# Remove missing values
adm_test_f2024_clean <- adm_test_f2024[!is.na(adm_test_f2024$sat_total), ]

# Create boxplot
ggplot(adm_test_f2024_clean, aes(x = sat_required, y = sat_total)) +
  geom_boxplot() +
  labs(x = "SAT Requirement", y = "Total SAT Score",
       title = "SAT Scores by Admission Requirement")
```

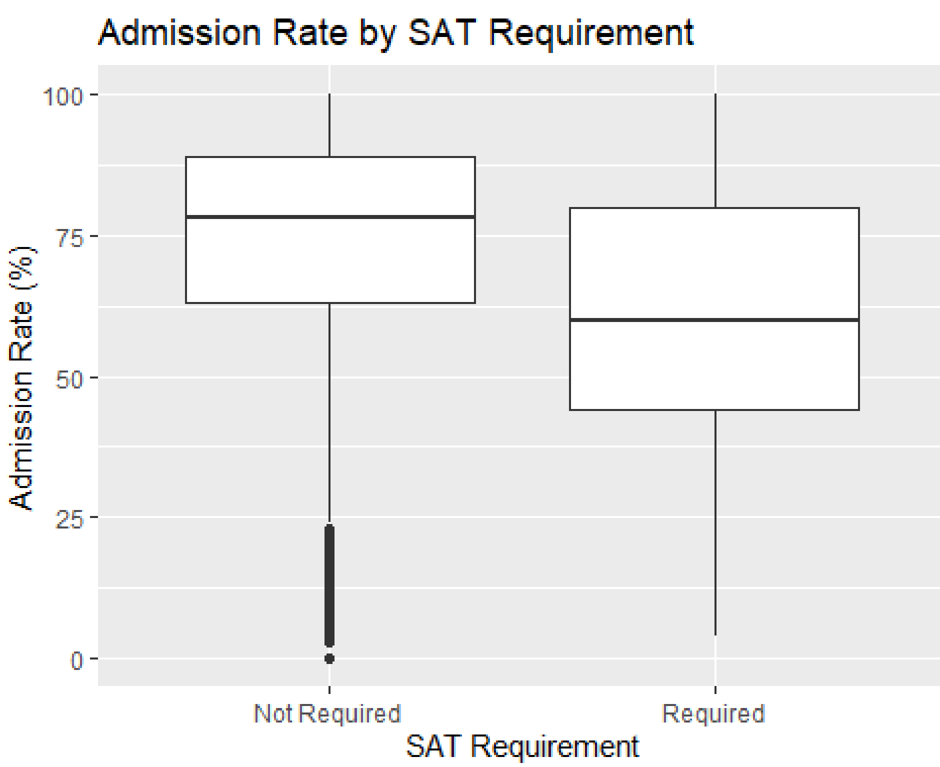



```
# Statistical tests
t.test(sat_total ~ sat_required, data = adm_test_f2024_clean)

##
##  Welch Two Sample t-test
##
## data:  sat_total by sat_required
## t = 2.083, df = 63.157, p-value = 0.04131
## alternative hypothesis: true difference in means between group Not
Required and group Required is not equal to 0
## 95 percent confidence interval:
##   1.90872 91.94720
## sample estimates:
## mean in group Not Required     mean in group Required
##                   1194.560                   1147.632

cohens_d(sat_total ~ sat_required, data = adm_test_f2024_clean)

## Cohen's d |       95% CI
## ------------------------
## 0.31      | [0.04, 0.58]
##
## - Estimated using pooled SD.
```

SAT Requirement vs Admission Rate

Do schools that require SAT scores have different admission rates compared to schools that do not require them?

```r
# Create SAT requirement variable and convert admission rate to numeric
sat_admission_rate <- adm_test_f2024 %>%
  mutate(
    sat_required = ifelse(
      ADM2024.Admission.test.scores == "Required to be considered for
admission",
      "Required", "Not Required"
    ),
    admit_rate = as.numeric(DRVADM2024.Percent.admitted...total)
  ) %>%
  filter(between(admit_rate, 0, 100))

# Create boxplot
ggplot(sat_admission_rate, aes(x = sat_required, y = admit_rate)) +
  geom_boxplot() +
  labs(
    x = "SAT Requirement",
    y = "Admission Rate (%)",
    title = "Admission Rate by SAT Requirement"
  )
```



```r
# Statistical tests
t.test(admit_rate ~ sat_required, data = sat_admission_rate)
```

```
##
##  Welch Two Sample t-test
##
## data:  admit_rate by sat_required
## t = 3.763, df = 64.747, p-value = 0.0003639
## alternative hypothesis: true difference in means between group Not
Required and group Required is not equal to 0
## 95 percent confidence interval:
##   6.048416 19.731583
## sample estimates:
## mean in group Not Required     mean in group Required
##                    72.3818                    59.4918
```

```r
cohens_d(admit_rate ~ sat_required, data = sat_admission_rate)
```

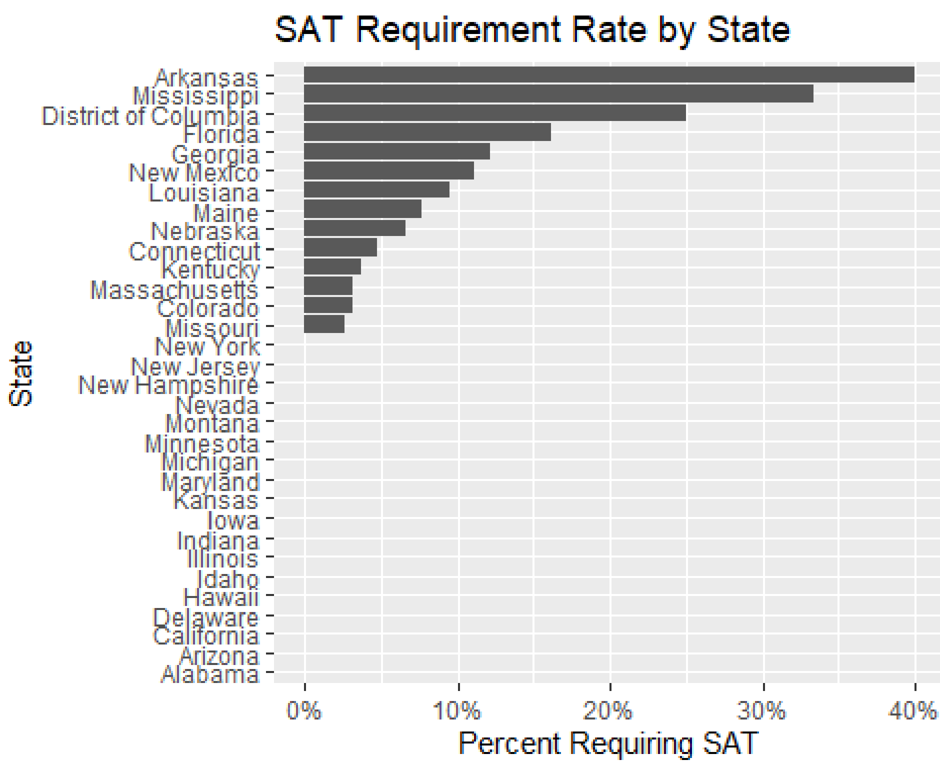
```
## Cohen's d |       95% CI
## ------------------------
## 0.56      | [0.30, 0.82]
##
## - Estimated using pooled SD.
```

SAT Requirements by State

Are SAT requirements geographically clustered across states?

```r
# Merge SAT requirement data with institutional characteristics
sat_state_requirements <- adm_test_f2024 %>%
  mutate(
    unitid = as.character(unitid),
    sat_required = ADM2024.Admission.test.scores == "Required to be
considered for admission"
  ) %>%
  select(unitid, sat_required) %>%
  inner_join(
    inst_chars_dir_2024_25 %>%
      mutate(unitid = as.character(unitid)) %>%
      select(unitid, state = HD2024.State.abbreviation),
    by = "unitid"
  ) %>%
  filter(!is.na(state))

# Calculate SAT requirement rate by state (only states with at least 5
schools)
state_rates <- sat_state_requirements %>%
  group_by(state) %>%
  summarise(rate = mean(sat_required), n = n(), .groups = "drop") %>%
  filter(n >= 5)
```

```r
# Create bar chart
ggplot(state_rates, aes(x = reorder(state, rate), y = rate)) +
  geom_col() +
  coord_flip() +
  scale_y_continuous(labels = scales::percent) +
  labs(
    x = "State",
    y = "Percent Requiring SAT",
    title = "SAT Requirement Rate by State"
  )
```

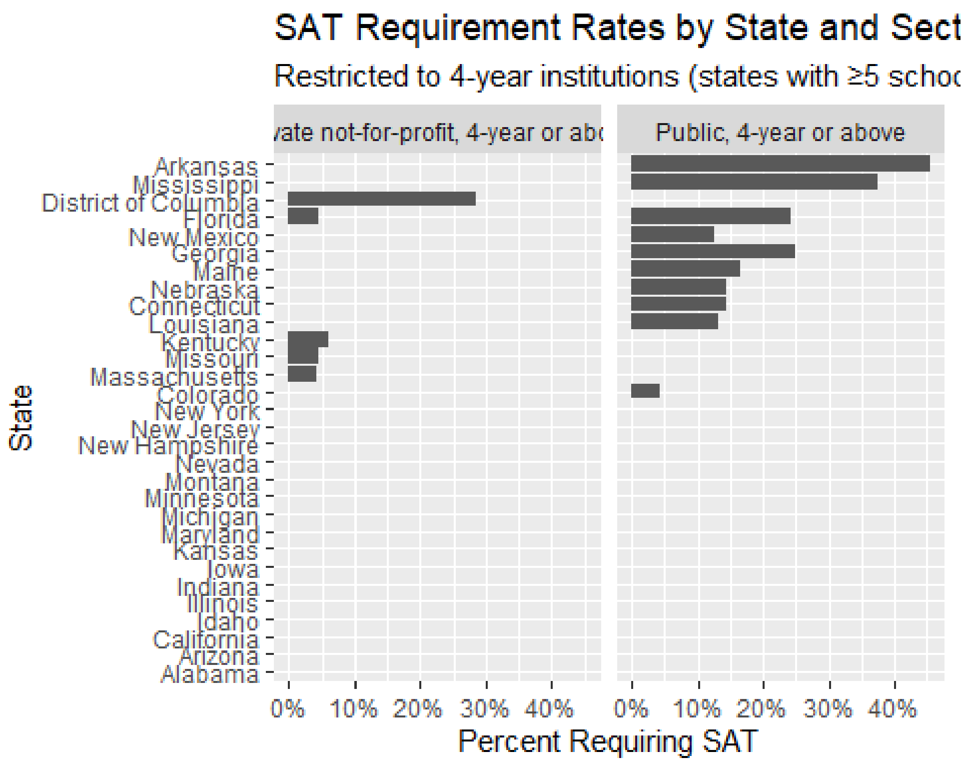


SAT Requirements by State and Sector

Are SAT requirements geographically clustered even after accounting for school type (public vs. private)?

```r
# Merge SAT requirement data with state and sector information
sat_state_sector <- adm_test_f2024 %>%
  mutate(
    unitid = as.character(unitid),
    sat_required = ADM2024.Admission.test.scores == "Required to be
considered for admission"
  ) %>%
  select(unitid, sat_required) %>%
  inner_join(
```

```r
    inst_chars_dir_2024_25 %>%
      mutate(unitid = as.character(unitid)) %>%
      select(unitid, state = HD2024.State.abbreviation, sector =
HD2024.Sector.of.institution),
    by = "unitid"
  ) %>%
  filter(
    !is.na(state),
    sector %in% c("Public, 4-year or above", "Private not-for-profit, 4-year
or above")
  )

# Calculate SAT requirement rate by state and sector
state_sector_rates <- sat_state_sector %>%
  group_by(state, sector) %>%
  summarise(rate = mean(sat_required), n = n(), .groups = "drop") %>%
  filter(n >= 5)

# Create faceted bar chart
ggplot(state_sector_rates, aes(x = reorder(state, rate), y = rate)) +
  geom_col() +
  coord_flip() +
  facet_wrap(~ sector) +
  scale_y_continuous(labels = scales::percent) +
  labs(
    x = "State",
    y = "Percent Requiring SAT",
    title = "SAT Requirement Rates by State and Sector",
    subtitle = "Restricted to 4-year institutions (states with ≥5 schools per
sector)"
  )
```

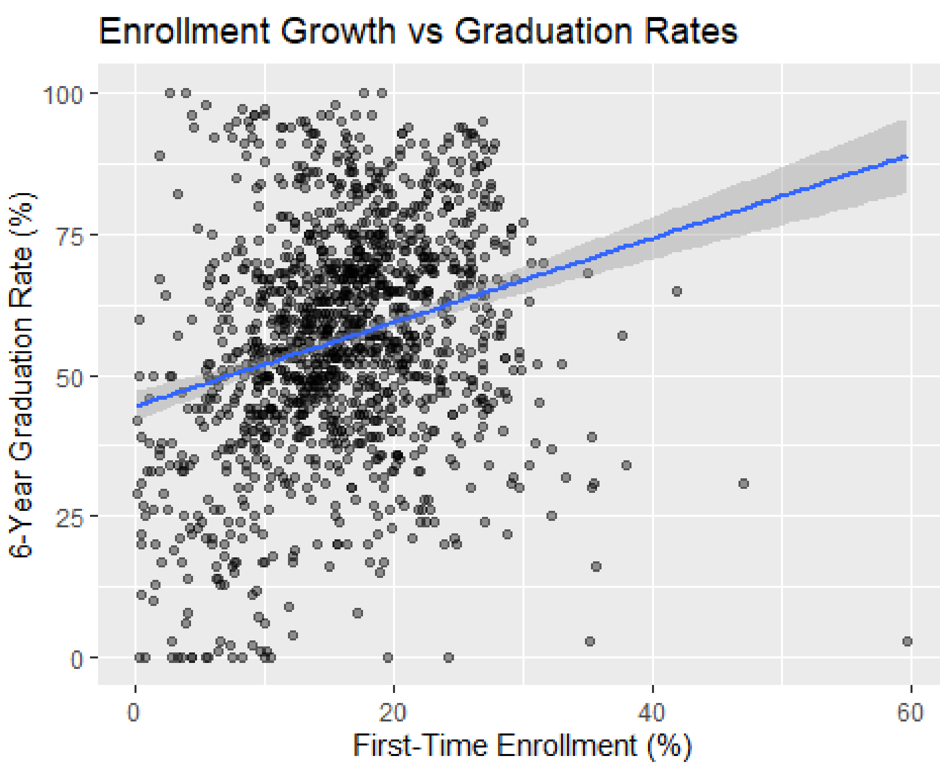


Enrollment Growth vs Graduation Rates

Do schools that grow enrollment faster sacrifice graduation rates, or do some manage both?

```
# Calculate first-time enrollment as percentage of total enrollment
enroll_prep <- fe_f2024 %>%
  mutate(
    unitid = as.character(unitid),
    total_enroll = as.numeric(DRVEF2024.Total..enrollment),
    first_time =
as.numeric(DRVEF2024.First.time.degree.certificate.seeking.undergraduate.enro
llment)
  ) %>%
  filter(!is.na(total_enroll), !is.na(first_time), total_enroll > 0,
first_time > 0) %>%
  mutate(growth_pct = 100 * first_time / total_enroll) %>%
  select(unitid, growth_pct)

# Get 6-year graduation rates
grad_prep <- grad_feq_var_cohort_2018_21 %>%
  mutate(unitid = as.character(unitid)) %>%
  mutate(grad_rate =
as.numeric(DRVGR2024.Graduation.rate...Bachelor.degree.within.6.years..total)
) %>%
  filter(!is.na(grad_rate), between(grad_rate, 0, 100)) %>%
```

```r
  select(unitid, grad_rate)

# Merge enrollment and graduation data
enrollment_growth_graduation <- enroll_prep %>%
  inner_join(grad_prep, by = "unitid")

# Create scatter plot
ggplot(enrollment_growth_graduation, aes(x = growth_pct, y = grad_rate)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  labs(
    x = "First-Time Enrollment (%)",
    y = "6-Year Graduation Rate (%)",
    title = "Enrollment Growth vs Graduation Rates"
  )

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



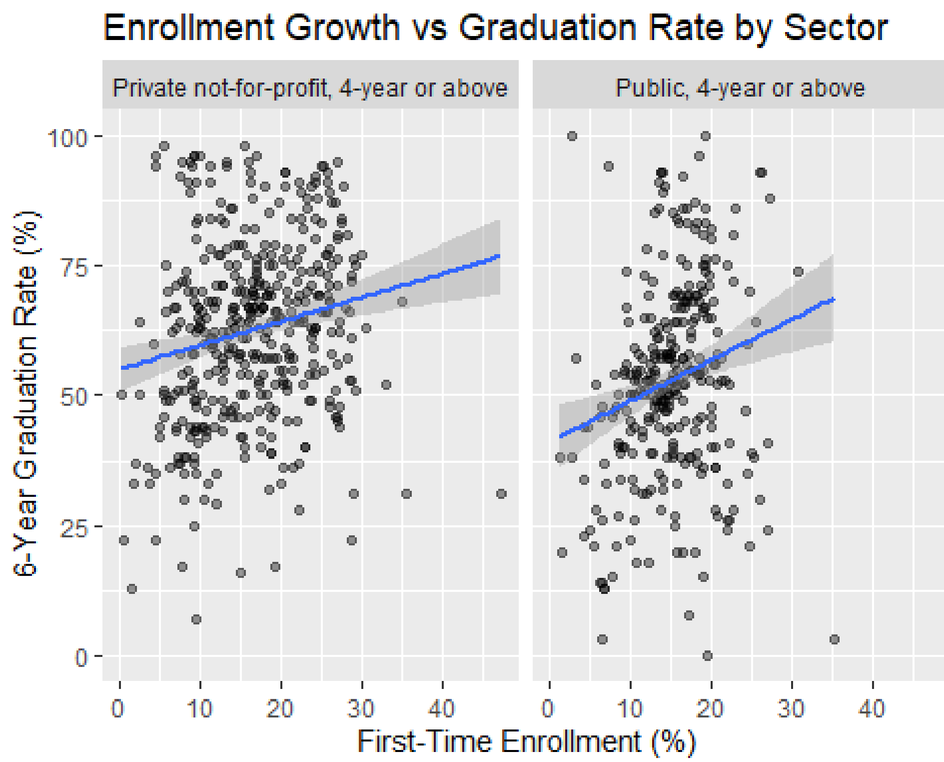

```r
# Correlation analysis
cor.test(enrollment_growth_graduation$growth_pct,
enrollment_growth_graduation$grad_rate)

##
##  Pearson's product-moment correlation
##
## data:  enrollment_growth_graduation$growth_pct and
enrollment_growth_graduation$grad_rate
```

```
## t = 9.7651, df = 1383, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2040288 0.3025982
## sample estimates:
##       cor
## 0.2539728
```

Enrollment Growth vs Graduation Rates by Sector

Is the relationship between enrollment growth and graduation rates consistent across public and private institutions?

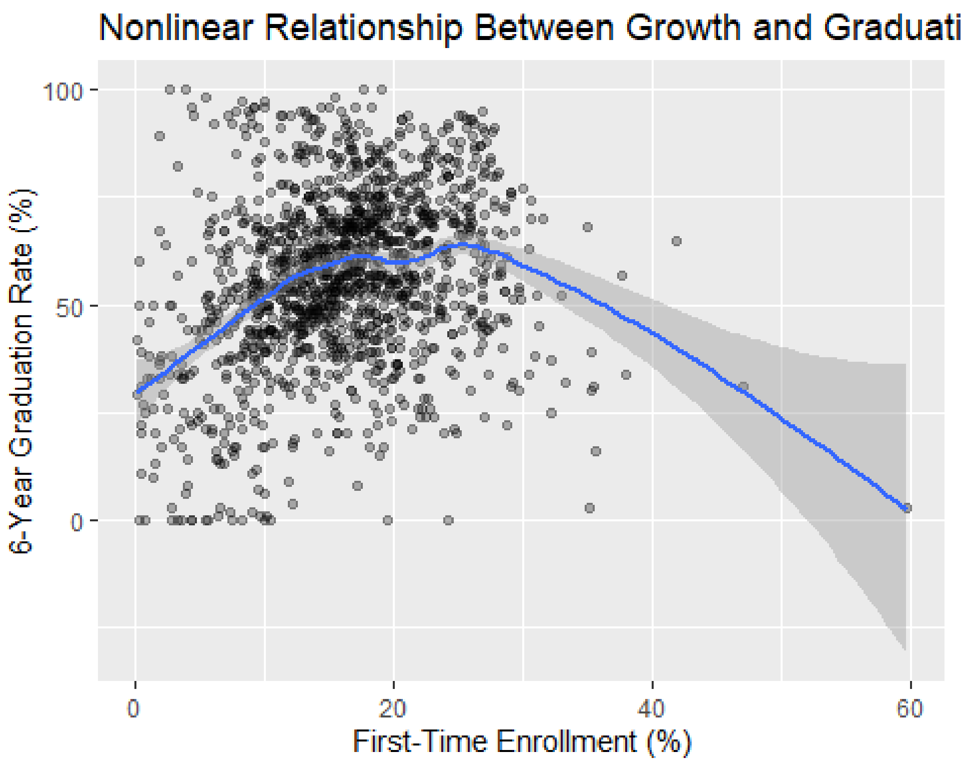
```r
# Add sector information to enrollment-graduation data
growth_grad_sector <- enrollment_growth_graduation %>%
  inner_join(
    inst_chars_dir_2024_25 %>%
      mutate(unitid = as.character(unitid)) %>%
      select(unitid, sector = HD2024.Sector.of.institution),
    by = "unitid"
  ) %>%
  filter(sector %in% c("Public, 4-year or above", "Private not-for-profit, 4-
year or above"))

# Create faceted scatter plot
ggplot(growth_grad_sector, aes(x = growth_pct, y = grad_rate)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  facet_wrap(~ sector) +
  labs(
    x = "First-Time Enrollment (%)",
    y = "6-Year Graduation Rate (%)",
    title = "Enrollment Growth vs Graduation Rate by Sector"
  )

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



Nonlinear Relationship: Enrollment Growth vs Graduation Rates

Is there a nonlinear relationship between enrollment growth and graduation rates, suggesting an optimal growth rate?

```
# Create scatter plot with loess smooth to explore nonlinearity
ggplot(enrollment_growth_graduation, aes(x = growth_pct, y = grad_rate)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "loess", span = 0.5) +
  labs(
    title = "Nonlinear Relationship Between Growth and Graduation Rates",
    x = "First-Time Enrollment (%)",
    y = "6-Year Graduation Rate (%)"
  )

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



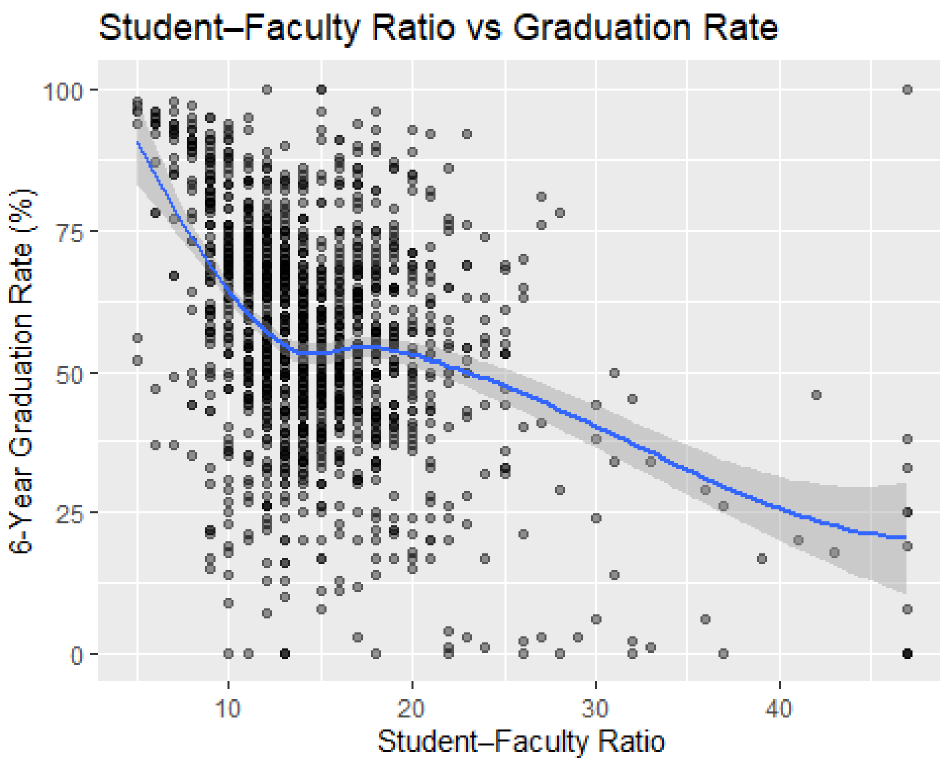

```r
# Fit quadratic model to test for nonlinear relationship
quad_model <- lm(grad_rate ~ growth_pct + I(growth_pct^2), data =
enrollment_growth_graduation)
summary(quad_model)
```

```
##
## Call:
## lm(formula = grad_rate ~ growth_pct + I(growth_pct^2), data =
enrollment_growth_graduation)
##
## Residuals:
##    Min      1Q Median     3Q    Max
## -62.53 -10.96   0.43  11.63  63.97
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     28.574417   1.992024   14.34   <2e-16 ***
## growth_pct       2.947677   0.223069   13.21   <2e-16 ***
## I(growth_pct^2) -0.063833   0.006101  -10.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.73 on 1382 degrees of freedom
## Multiple R-squared:  0.1332, Adjusted R-squared:  0.1319
## F-statistic: 106.1 on 2 and 1382 DF,  p-value: < 2.2e-16
```

Faculty-to-Student Ratio vs Graduation Rates

Does having more faculty per student always translate to better graduation outcomes?

```
# Get student-faculty ratio data
faculty_data <- fe_f2024 %>%
  mutate(
    unitid = as.character(unitid),
    student_faculty_ratio = as.numeric(EF2024D.Student.to.faculty.ratio)
  ) %>%
  filter(!is.na(student_faculty_ratio), between(student_faculty_ratio, 5,
50)) %>%
  select(unitid, student_faculty_ratio)

# Get graduation rates
grad_data <- grad_feq_var_cohort_2018_21 %>%
  mutate(unitid = as.character(unitid)) %>%
  mutate(grad_rate =
as.numeric(DRVGR2024.Graduation.rate...Bachelor.degree.within.6.years..total)
) %>%
  filter(!is.na(grad_rate), between(grad_rate, 0, 100)) %>%
  select(unitid, grad_rate)

# Merge data
faculty_ratio_graduation <- faculty_data %>%
  inner_join(grad_data, by = "unitid")

# Create scatter plot with loess smooth
ggplot(faculty_ratio_graduation, aes(x = student_faculty_ratio, y =
grad_rate)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess") +
  labs(
    x = "Student-Faculty Ratio",
    y = "6-Year Graduation Rate (%)",
    title = "Student-Faculty Ratio vs Graduation Rate"
  )

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




```
# Correlation analysis
cor.test(faculty_ratio_graduation$student_faculty_ratio,
faculty_ratio_graduation$grad_rate)

##
##  Pearson's product-moment correlation
##
## data:  faculty_ratio_graduation$student_faculty_ratio and
faculty_ratio_graduation$grad_rate
## t = -15.277, df = 1376, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4250743 -0.3347397
## sample estimates:
##        cor
## -0.3808153
```

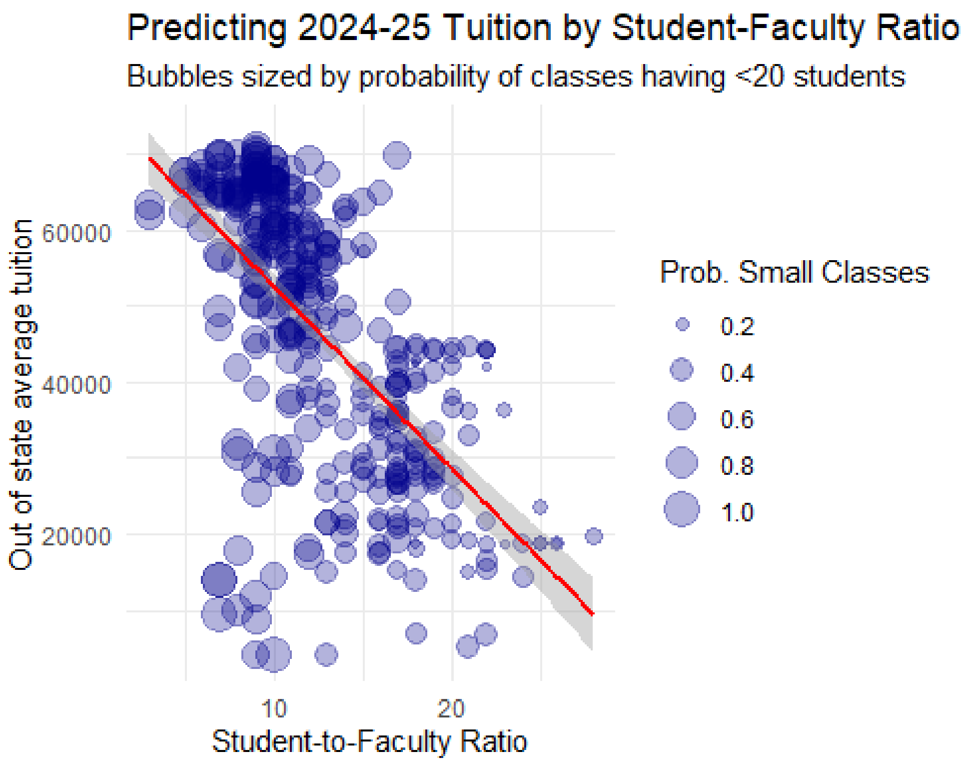
Potential Graphs:

Tuition vs Post-Graduation Salary: Are Higher Tuition Levels Associated with Higher Earnings?

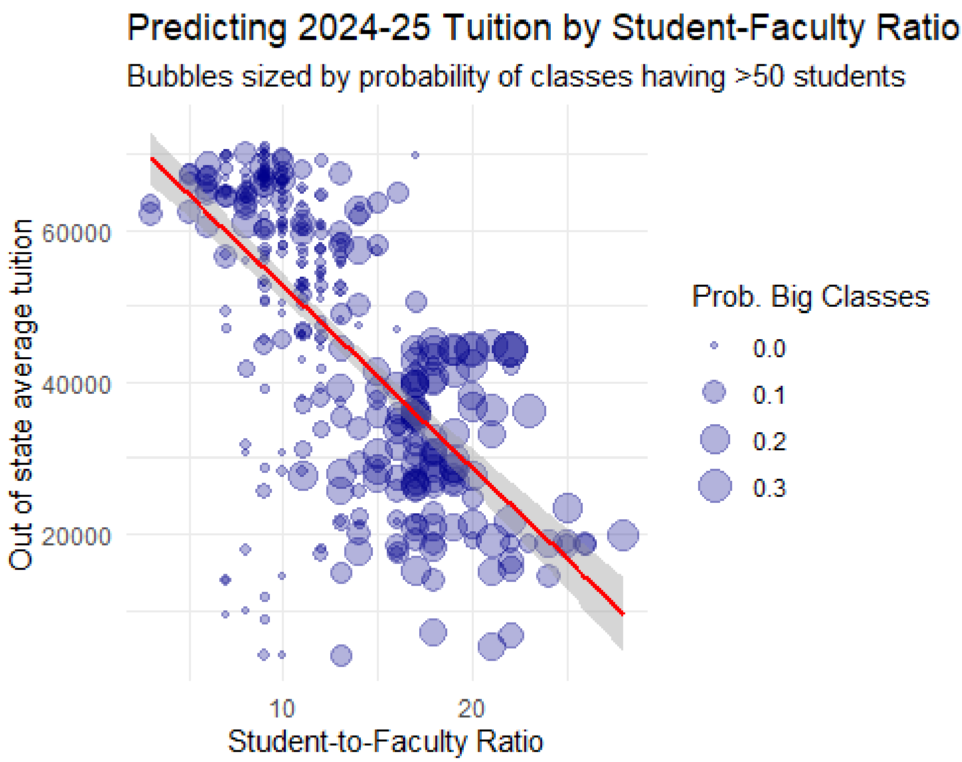
Tuition vs Instructional Spending: Are Higher-Tuition Schools Actually Investing More?

Class-size vs Tuition Cost

```
# MERGE THE DATASETS
# We join your already loaded 'class_size' and 'costs_f2024' on 'UNITID'
# UNITID is the unique identifier for all IPEDS institutions [3, 4].

analysis_df <- class_size %>%
  inner_join(costs_f2024, by = c("UNITID" = "unitid")) %>%
  select(UNITID, STUFACR, CLASIZUND20, CLASIZOVE50,
`COST1_2024.Out.of.state.average.tuition.for.full.time.undergraduates`) %>%

filter(!is.na(`COST1_2024.Out.of.state.average.tuition.for.full.time.undergra
duates`) & !is.na(STUFACR))

# Predicting Out of state average tuition for full time undergraduates
# based on Student-to-Faculty Ratio and Class Size
ggplot(analysis_df, aes(x = STUFACR, y =
`COST1_2024.Out.of.state.average.tuition.for.full.time.undergraduates`)) +
  geom_point(aes(size = CLASIZUND20), alpha = 0.3, color = "darkblue") +
  geom_smooth(method = "lm", formula = y ~ x, color = "red", se = TRUE) +
  labs(title = "Predicting 2024-25 Tuition by Student-Faculty Ratio",
       subtitle = "Bubbles sized by probability of classes having <20
students",
       x = "Student-to-Faculty Ratio",
       y = "Out of state average tuition",
       size = "Prob. Small Classes") +
  theme_minimal()

## Warning: Removed 5 rows containing missing values or values outside the
scale range
## (`geom_point()`).
```




```
ggplot(analysis_df, aes(x = STUFACR, y =
`COST1_2024.Out.of.state.average.tuition.for.full.time.undergraduates`)) +
  geom_point(aes(size = CLASIZOVE50), alpha = 0.3, color = "darkblue") +
  geom_smooth(method = "lm", formula = y ~ x, color = "red", se = TRUE) +
  labs(title = "Predicting 2024-25 Tuition by Student-Faculty Ratio",
       subtitle = "Bubbles sized by probability of classes having >50
students",
       x = "Student-to-Faculty Ratio",
       y = "Out of state average tuition",
       size = "Prob. Big Classes") +
  theme_minimal()

## Warning: Removed 5 rows containing missing values or values outside the
scale range
## (`geom_point()`).
```




```
# Correlation Test
cor.test(analysis_df$STUFACR,
analysis_df$COST1_2024.Out.of.state.average.tuition.for.full.time.undergradua
tes)

##
##  Pearson's product-moment correlation
##
## data:  analysis_df$STUFACR and
analysis_df$COST1_2024.Out.of.state.average.tuition.for.full.time.undergradua
tes
## t = -15.167, df = 355, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6862031 -0.5597043
## sample estimates:
##        cor
## -0.6270704

cor.test(analysis_df$CLASIZUND20,
analysis_df$COST1_2024.Out.of.state.average.tuition.for.full.time.undergradua
tes)

##
##  Pearson's product-moment correlation
##
## data:  analysis_df$CLASIZUND20 and
```

```
analysis_df$COST1_2024.Out.of.state.average.tuition.for.full.time.undergradua
tes
## t = 9.1044, df = 350, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3490192 0.5184045
## sample estimates:
##       cor
## 0.4375858
```

```
cor.test(analysis_df$CLASIZOVE50,
analysis_df$COST1_2024.Out.of.state.average.tuition.for.full.time.undergradua
tes)
```

```
##
##  Pearson's product-moment correlation
##
## data:  analysis_df$CLASIZOVE50 and
analysis_df$COST1_2024.Out.of.state.average.tuition.for.full.time.undergradua
tes
## t = -7.3643, df = 350, p-value = 1.287e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4534500 -0.2721699
## sample estimates:
##        cor
## -0.3662803
```

Average Tuition vs Combined Food and Housing Charge

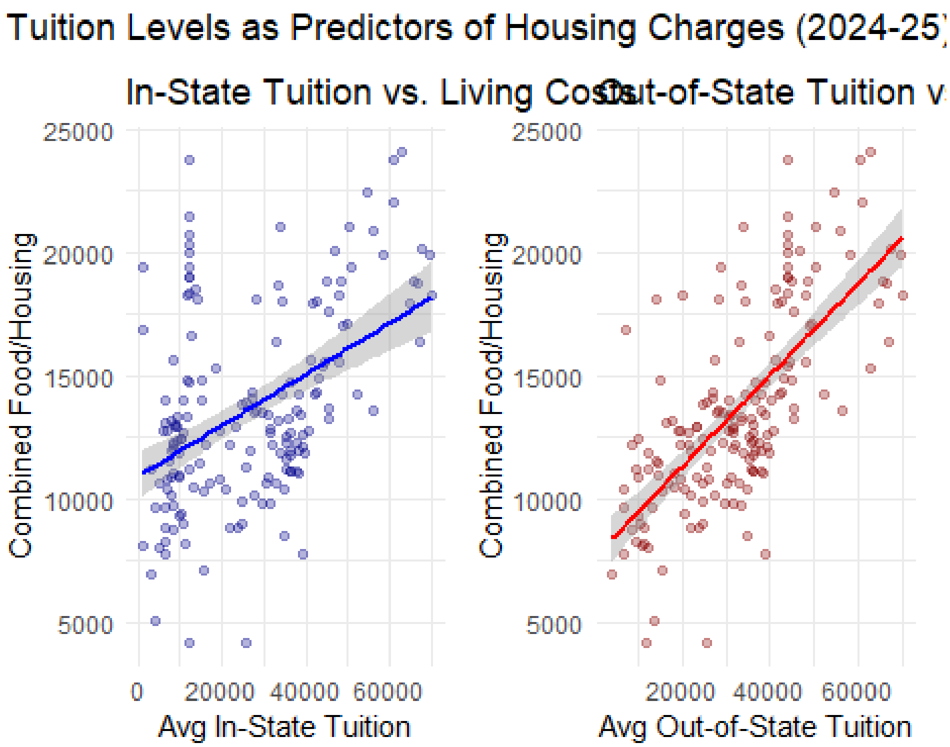
```
df_costs <- costs_f2024 %>%
  select(
    unitid,
    InState =
COST1_2024.In.state.average.tuition.for.full.time.undergraduates,
    OutState =
COST1_2024.Out.of.state.average.tuition.for.full.time.undergraduates,
    RoomBoard = COST1_2024.Combined.food.and.housing.charge
  ) %>%
  filter(!is.na(InState) & !is.na(OutState) & !is.na(RoomBoard))

# MULTIPLE LINEAR REGRESSION
# Model: Does tuition (In vs Out) predict housing costs?
cost_model <- lm(RoomBoard ~ InState + OutState, data = df_costs)
summary(cost_model)
```

```
##
## Call:
## lm(formula = RoomBoard ~ InState + OutState, data = df_costs)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -7850.8 -1997.8  -297.1  1788.7  8158.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7362.26287  526.08973  13.994  < 2e-16 ***
## InState       -0.06060    0.02042  -2.968  0.00344 **
## OutState       0.24246    0.02412  10.053  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2818 on 168 degrees of freedom
## Multiple R-squared:  0.5072, Adjusted R-squared:  0.5013
## F-statistic: 86.44 on 2 and 168 DF,  p-value: < 2.2e-16
```

```r
# 3. VISUALIZATION: Dual-Panel Comparison
# Plot A: In-State Tuition vs Room & Board
p1 <- ggplot(df_costs, aes(x = InState, y = RoomBoard)) +
  geom_point(alpha = 0.3, color = "darkblue") +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "In-State Tuition vs. Living Costs",
       x = "Avg In-State Tuition", y = "Combined Food/Housing") +
  theme_minimal()

# Plot B: Out-of-State Tuition vs Room & Board
p2 <- ggplot(df_costs, aes(x = OutState, y = RoomBoard)) +
  geom_point(alpha = 0.3, color = "darkred") +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Out-of-State Tuition vs. Living Costs",
       x = "Avg Out-of-State Tuition", y = "Combined Food/Housing") +
  theme_minimal()

p1 + p2 + plot_annotation(title = "Tuition Levels as Predictors of Housing
Charges (2024-25)")
```

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


Tuition Levels as Predictors of Housing Charges (2024-25)
In-State Tuition vs. Living Costs
Out-of-State Tuition v
Combined Food/Housing
25000
20000
15000
10000
5000
0
20000
40000
60000
Avg In-State Tuition
Avg Out-of-State Tuition