Final Project

Nabeel Haider

5/15/2025

Main Questions:

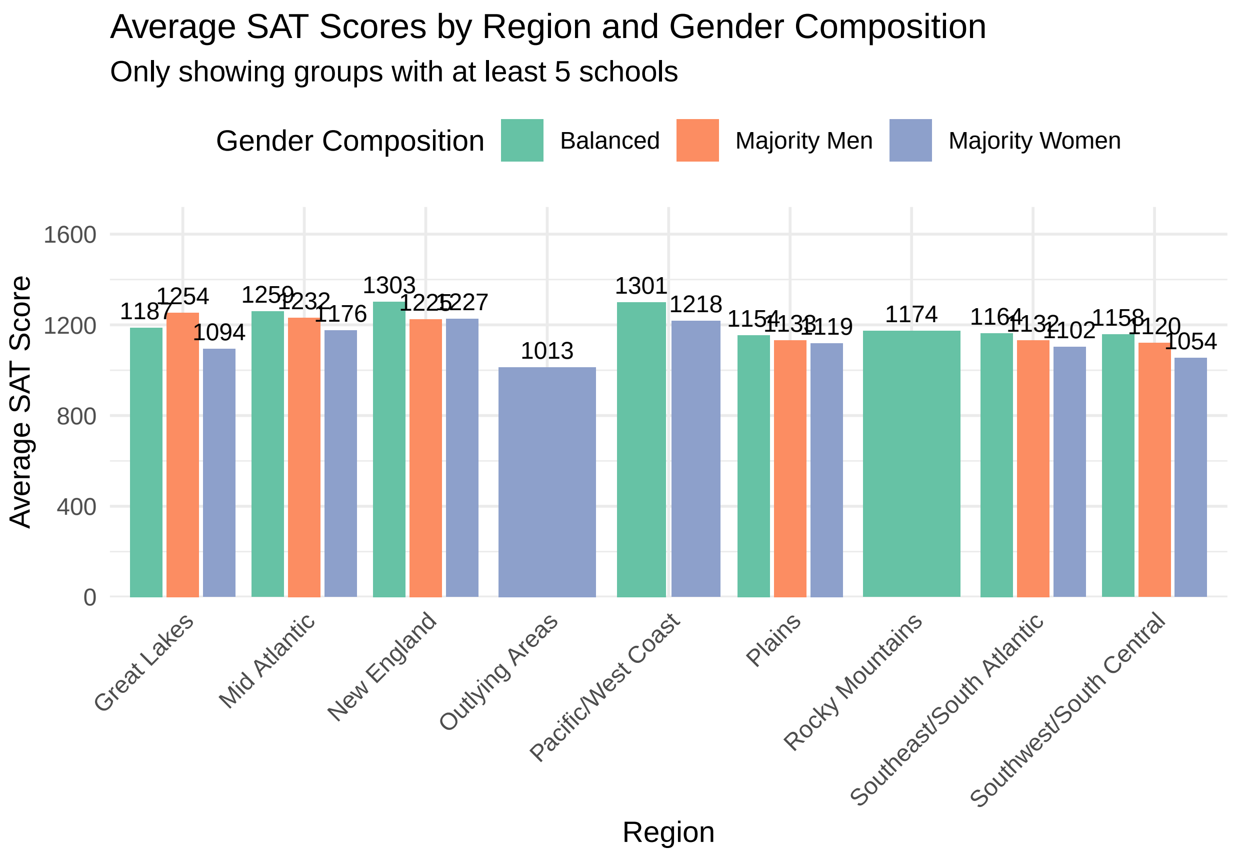
1. How do institutional performance measures vary by region and student demographics in Colleges in the United States?
2. How do the educational outcomes change across the different regions in the United States? How are these differences related to the different factors?

Questions:

1. How do test scores vary across schools by region and gender?
2. How does the cost of attendance relate to graduation rates?
3. Do outcomes differ by school type (public, private, proprietary)
4. Is there a relationship between cost of attendance and median student earnings after graduation?

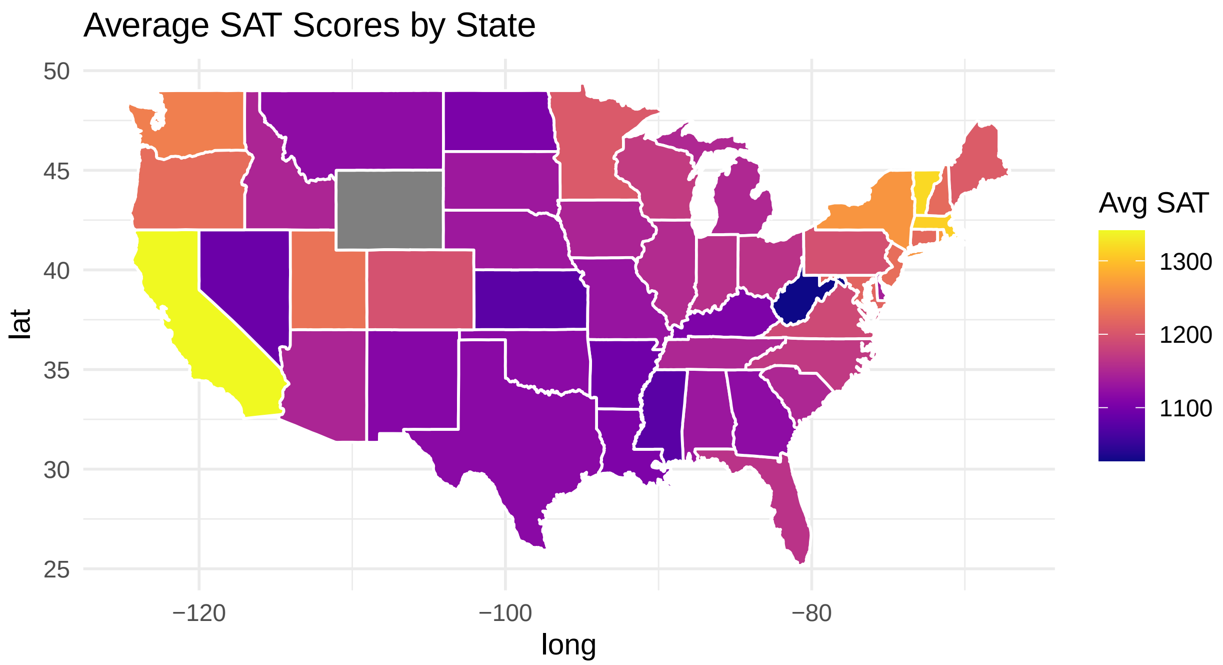
1 Faceted bar plot (Average test scores by region faceted by gender)

```
avg_scores <- data %>%
  filter(!is.na(SAT_AVG), !is.na(UGDS_WOMEN), !is.na(UGDS_MEN)) %>%
  filter(!is.na(REGION), REGION %in% 1:9) %>%
  mutate(
    Gender_Group = case_when(
      UGDS_WOMEN > 0.6 ~ "Majority Women",
      UGDS_WOMEN < 0.4 ~ "Majority Men",
      TRUE ~ "Balanced"
    ),
    REGION = case_when(
      REGION == 1 ~ "New England",
      REGION == 2 ~ "Mid Atlantic",
      REGION == 3 ~ "Great Lakes",
      REGION == 4 ~ "Plains",
      REGION == 5 ~ "Southeast/South Atlantic",
      REGION == 6 ~ "Southwest/South Central",
      REGION == 7 ~ "Rocky Mountains",
      REGION == 8 ~ "Pacific/West Coast",
      REGION == 9 ~ "Outlying Areas"
    )
  ) %>%
  group_by(REGION, Gender_Group) %>%
  summarize(
    Avg_SAT = mean(as.numeric(SAT_AVG), na.rm = TRUE),
    n_schools = n(),
    .groups = "drop"
  ) %>%
  filter(n_schools >= 5)

ggplot(avg_scores, aes(x = REGION, y = Avg_SAT, fill = Gender_Group)) +
  geom_col(position = position_dodge(width = 0.9), width = 0.8) +
  geom_text(aes(label = round(Avg_SAT, 0)),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 3) +
  scale_fill_brewer(palette = "Set2") +
  scale_y_continuous(
    limits = c(0, max(avg_scores$Avg_SAT) * 1.2),

    expand = expansion(mult = c(0, 0.1))

  ) +
  labs(title = "Average SAT Scores by Region and Gender Composition",
       subtitle = "Only showing groups with at least 5 schools",
       x = "Region",
       y = "Average SAT Score",
       fill = "Gender Composition") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "top")
```



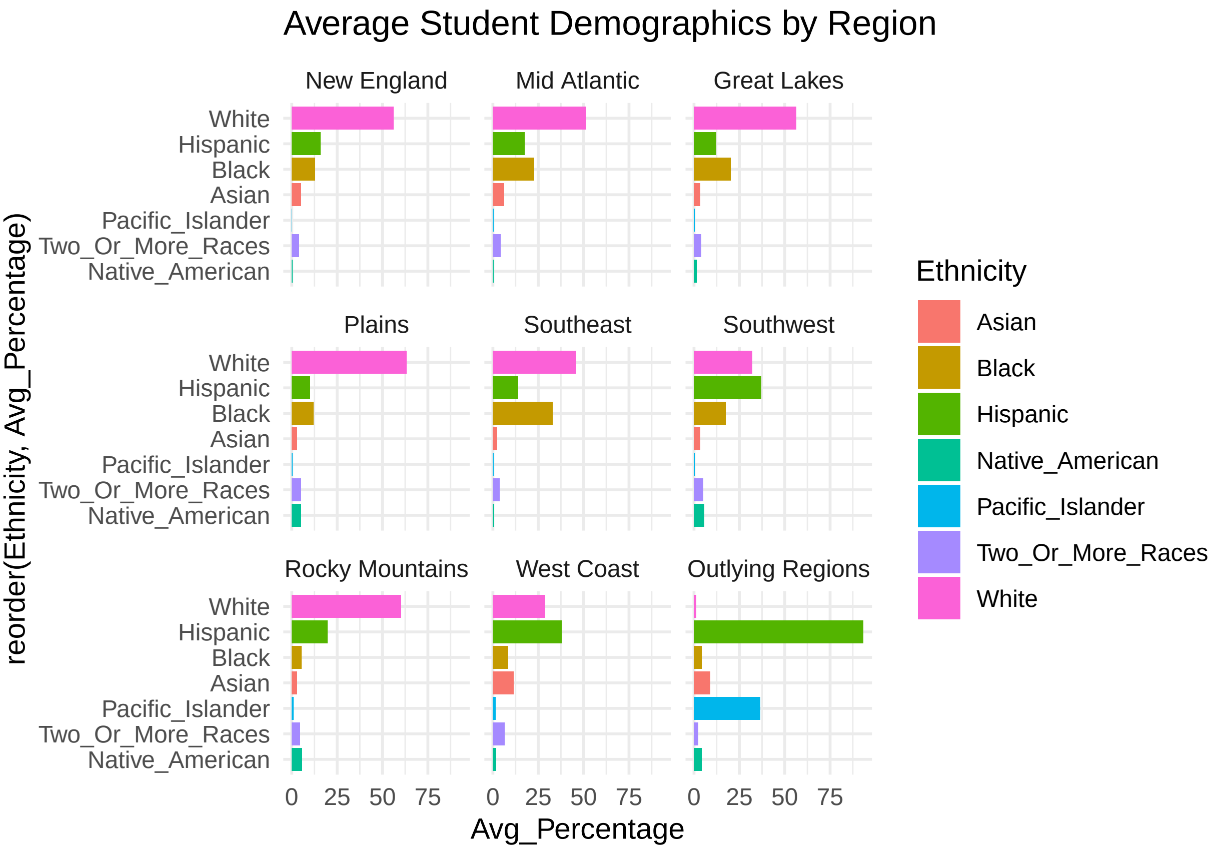
2 Second figure (Geospatial Map) of SAT Averages by State :

```
state_scores <- data %>%
  filter(!is.na(SAT_AVG)) %>%
  group_by(STABBR) %>%
  summarize(avg_sat = mean(as.numeric(SAT_AVG), na.rm = TRUE))

state_scores$region <- tolower(state.name[match(state_scores$STABBR, state.abb)])

us_map <- map_data("state")


map_df <- left_join(us_map, state_scores, by = "region")

ggplot(map_df, aes(x = long, y = lat, group = group, fill = avg_sat)) +
  geom_polygon(color = "white") +
  coord_fixed(1.3) +
  scale_fill_viridis(option = "plasma", name = "Avg SAT") +
  labs(title = "Average SAT Scores by State") +
  theme_minimal()
```



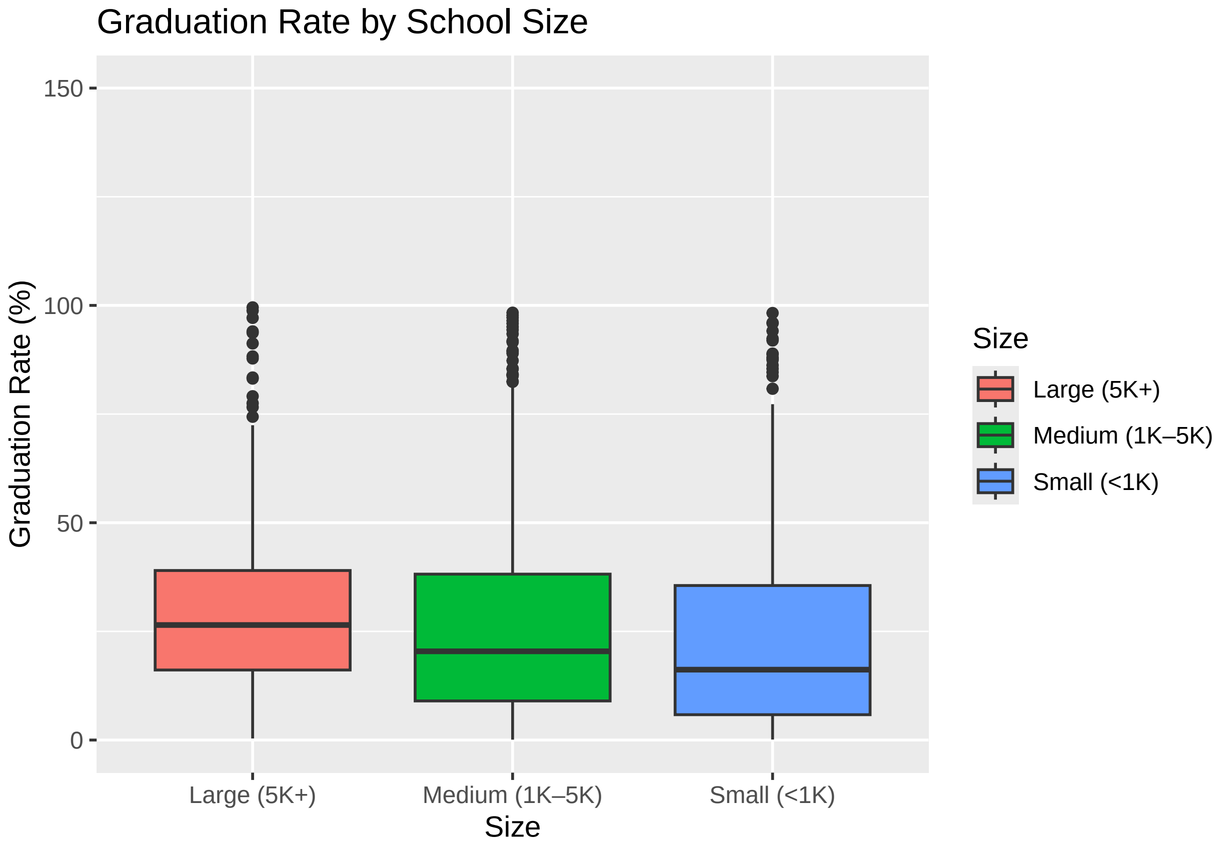
3 Faceted Figure of Average Student Demographics by Region

```
final_plot <- data %>%
  select(REGION,
         White = UGDS_WHITE,
         Black = UGDS_BLACK,
         Hispanic = UGDS_HISP,
         Asian = UGDS_ASIAN,
         Native_American = UGDS_AIAN,
         Pacific_Islander = UGDS_NHPI,
         Two_Or_More_Races = UGDS_2MOR,
         Nonresident = UGDS_NRA,
         Unknown = UGDS_UNKN) %>%
  pivot_longer(
    cols = -REGION,
    names_to = "Ethnicity",
    values_to = "Percentage"
  ) %>%
  mutate(
    REGION = factor(REGION,
                    levels = 1:9,
                    labels = c("New England", "Mid Atlantic", "Great Lakes", "Plains",
                              "Southeast", "Southwest", "Rocky Mountains", "West Coast", "Outlying Region
  ) %>%
  filter(
    !is.na(REGION),
    !is.na(Percentage),
    Percentage > 0,
    Ethnicity %in% c("White", "Black", "Hispanic", "Asian",
                    "Native_American", "Pacific_Islander", "Two_Or_More_Races")
  ) %>%
  group_by(REGION, Ethnicity) %>%
  summarize(
    Avg_Percentage = mean(Percentage, na.rm = TRUE) * 100,
    .groups = "drop"
  )

ggplot(final_plot, aes(x = reorder(Ethnicity, Avg_Percentage),
                      y = Avg_Percentage,
                      fill = Ethnicity)) +
  geom_col() +
  coord_flip() +
  facet_wrap(~REGION) +
  labs(title = "Average Student Demographics by Region") +
  theme_minimal()
```



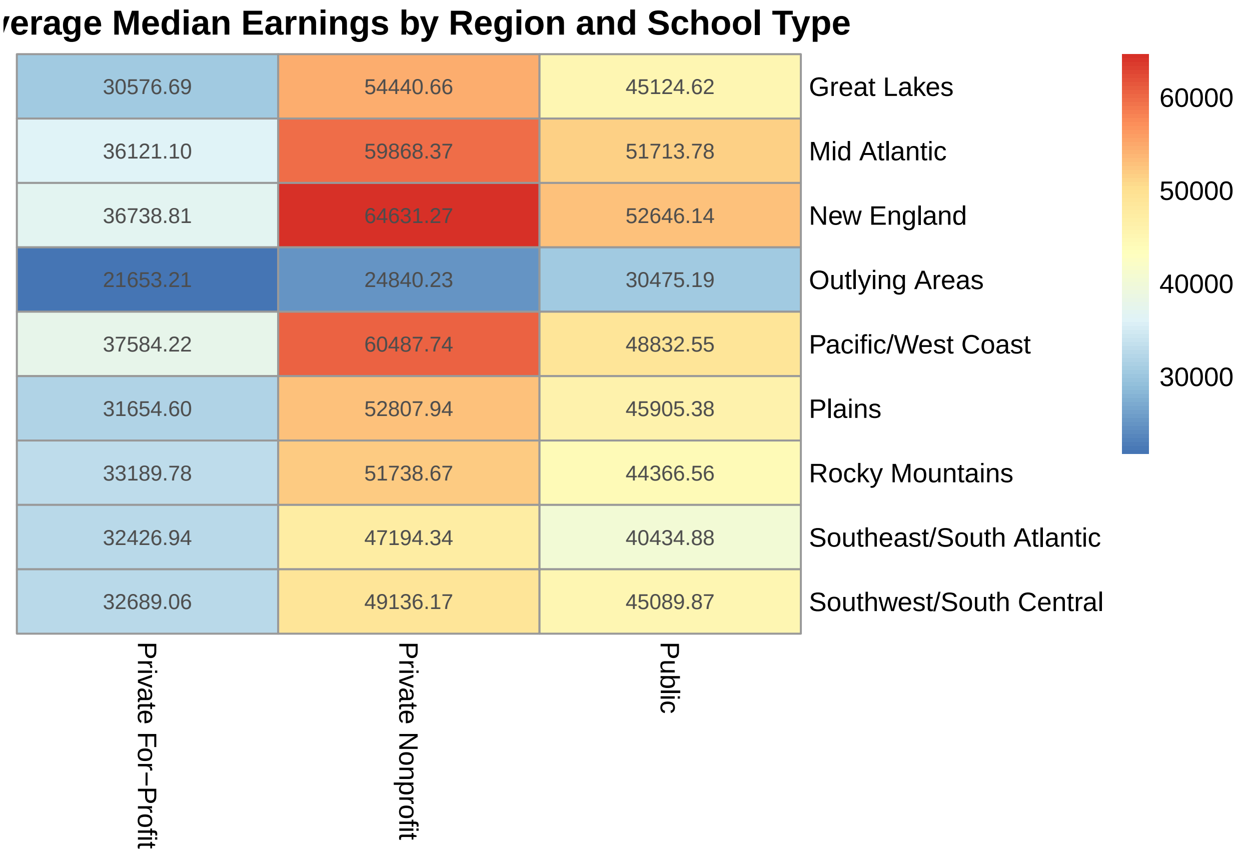
4 Box Plot of Graduation Rate by School Size

```
ggplot(
  data %>%
    filter(!is.na(GRADS), !is.na(UGDS), UGDS > 0) %>%
    mutate(
      Grad_Rate = GRADS / UGDS,
      Size = case_when(
        UGDS < 1000 ~ "Small (<1K)",
        UGDS < 5000 ~ "Medium (1K-5K)",
        TRUE ~ "Large (5K+)"
      )
    ) %>%
    filter(Grad_Rate <= 1),
  aes(x = Size, y = Grad_Rate * 100, fill = Size)
) +
  geom_boxplot() +
  labs(title = "Graduation Rate by School Size", y = "Graduation Rate (%)") +
  ylim(0, 150)
```



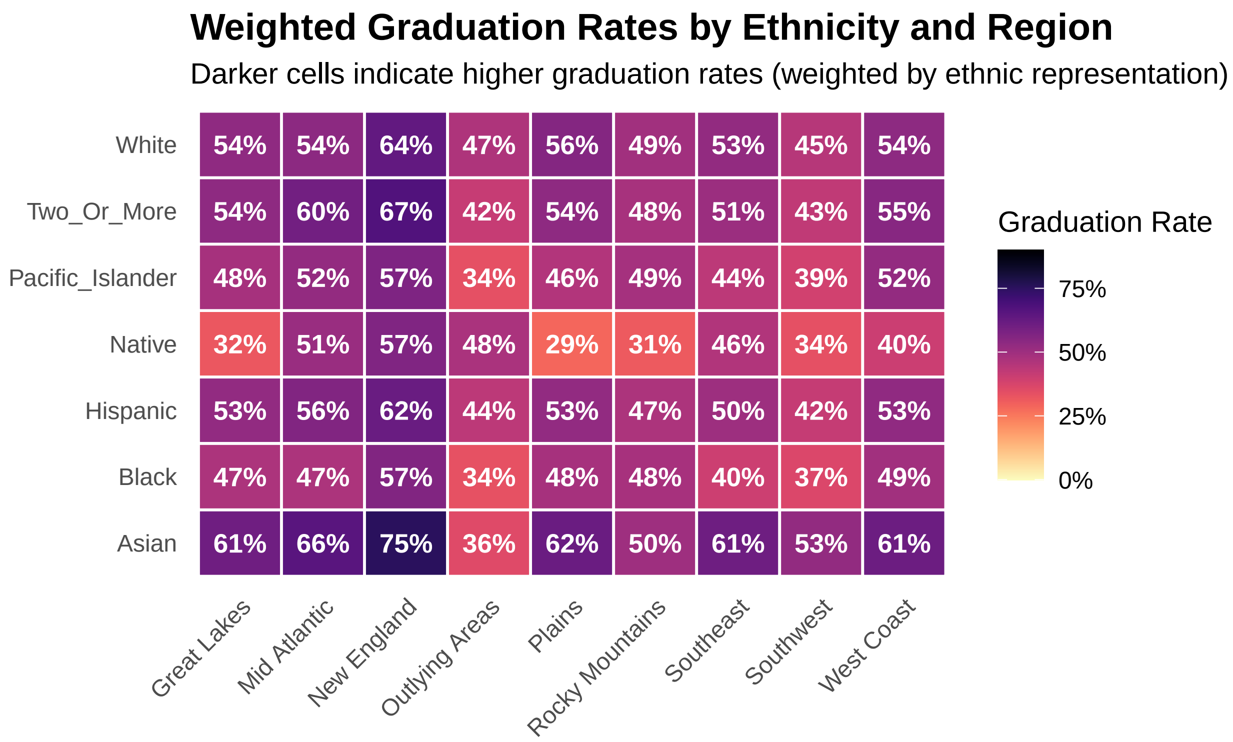
5 Heatmap 1 - Average Median Earnings by Region and School Type

```r
earnings_summary <- data %>%
  filter(!is.na(MD_EARN_WNE_P10), !is.na(REGION), !is.na(CONTROL)) %>%
  mutate(
    REGION = case_when(
      REGION == 1 ~ "New England",
      REGION == 2 ~ "Mid Atlantic",
      REGION == 3 ~ "Great Lakes",
      REGION == 4 ~ "Plains",
      REGION == 5 ~ "Southeast/South Atlantic",
      REGION == 6 ~ "Southwest/South Central",
      REGION == 7 ~ "Rocky Mountains",
      REGION == 8 ~ "Pacific/West Coast",
      REGION == 9 ~ "Outlying Areas"
    ),
    School_Type = case_when(
      CONTROL == 1 ~ "Public",
      CONTROL == 2 ~ "Private Nonprofit",
      CONTROL == 3 ~ "Private For-Profit"
    )
  ) %>%
  group_by(REGION, School_Type) %>%
  summarize(Avg_Median_Earnings = mean(MD_EARN_WNE_P10, na.rm = TRUE), .groups = "drop") %>%
  filter(!is.na(REGION))

earnings_matrix <- earnings_summary %>%
  pivot_wider(names_from = School_Type, values_from = Avg_Median_Earnings) %>%
  filter(!is.na(REGION)) %>%
  column_to_rownames("REGION")


pheatmap(earnings_matrix,
         main = "Average Median Earnings by Region and School Type",
         cluster_rows = FALSE, cluster_cols = FALSE,
         display_numbers = TRUE)
```



6 Weighted Graduation Rates by Ethnicity and Region

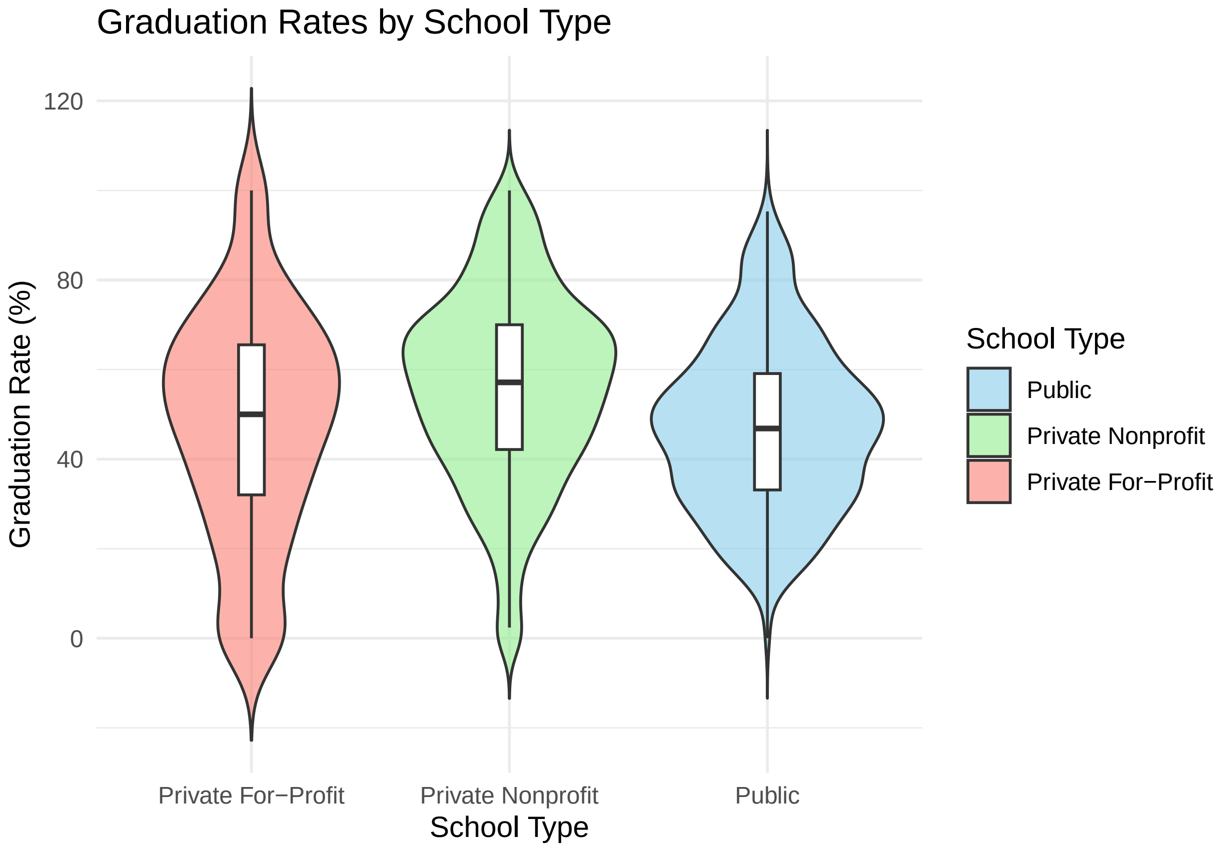
```
grad_rates <- data %>%
  select(
    REGION,
    GRAD_RATE = C150_4,
    White = UGDS_WHITE,
    Black = UGDS_BLACK,
    Hispanic = UGDS_HISP,
    Asian = UGDS_ASIAN,
    Native = UGDS_AIAN,
    Pacific_Islander = UGDS_NHPI,
    Two_Or_More = UGDS_2MOR
  ) %>%
  mutate(
    REGION = case_when(
      REGION == 1 ~ "New England",
      REGION == 2 ~ "Mid Atlantic",
      REGION == 3 ~ "Great Lakes",
      REGION == 4 ~ "Plains",
      REGION == 5 ~ "Southeast",
      REGION == 6 ~ "Southwest",
      REGION == 7 ~ "Rocky Mountains",
      REGION == 8 ~ "West Coast",
      REGION == 9 ~ "Outlying Areas",
      TRUE ~ NA_character_
    )
  ) %>% filter(!is.na(REGION)) %>%
  pivot_longer(
    cols = -c(REGION, GRAD_RATE),
    names_to = "ETHNICITY",
    values_to = "PERCENT"
  ) %>%
  filter(!is.na(GRAD_RATE), !is.na(PERCENT)) %>%

  group_by(REGION, ETHNICITY) %>%
  summarize(
    WEIGHTED_GRAD_RATE = weighted.mean(GRAD_RATE, w = PERCENT, na.rm = TRUE),
    .groups = "drop"
  )

ggplot(grad_rates, aes(x = REGION, y = ETHNICITY, fill = WEIGHTED_GRAD_RATE)) +
  geom_tile(color = "white", linewidth = 0.5) +
  geom_text(
    aes(label = scales::percent(WEIGHTED_GRAD_RATE, accuracy = 1)),
    color = "white",
    size = 3.5,
    fontface = "bold"
  ) +
  scale_fill_viridis(
    name = "Graduation Rate",
    option = "magma",
    direction = -1,
    labels = scales::percent,
    limits = c(0, 0.9)
  ) +
  labs(
    title = "Weighted Graduation Rates by Ethnicity and Region",
    subtitle = "Darker cells indicate higher graduation rates (weighted by ethnic representation)",
    x = NULL,
    y = NULL,
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1),
    panel.grid = element_blank(),
    plot.title = element_text(face = "bold", size = 14),
    legend.position = "right"
  ) +
  coord_fixed(ratio = 0.8)
```



# Violin and Box Plot of Graduation Rate By School Type

```
ggplot(data, aes(
  x = factor(case_when(
    CONTROL == 1 ~ "Public",
    CONTROL == 2 ~ "Private Nonprofit",
    CONTROL == 3 ~ "Private For-Profit"
  )),
  y = C150_4 * 100,
  fill = factor(CONTROL)
)) +
  geom_violin(trim = FALSE, alpha = 0.6) +
  geom_boxplot(width = 0.1, outlier.shape = NA, fill = "white") +
  labs(
    title = "Graduation Rates by School Type",
    x = "School Type",
    y = "Graduation Rate (%)"
  ) +
  scale_fill_manual(
    values = c("1" = "skyblue", "2" = "lightgreen", "3" = "salmon"),
    labels = c("Public", "Private Nonprofit", "Private For-Profit"),
    name = "School Type"
  ) +
  theme_minimal()
```

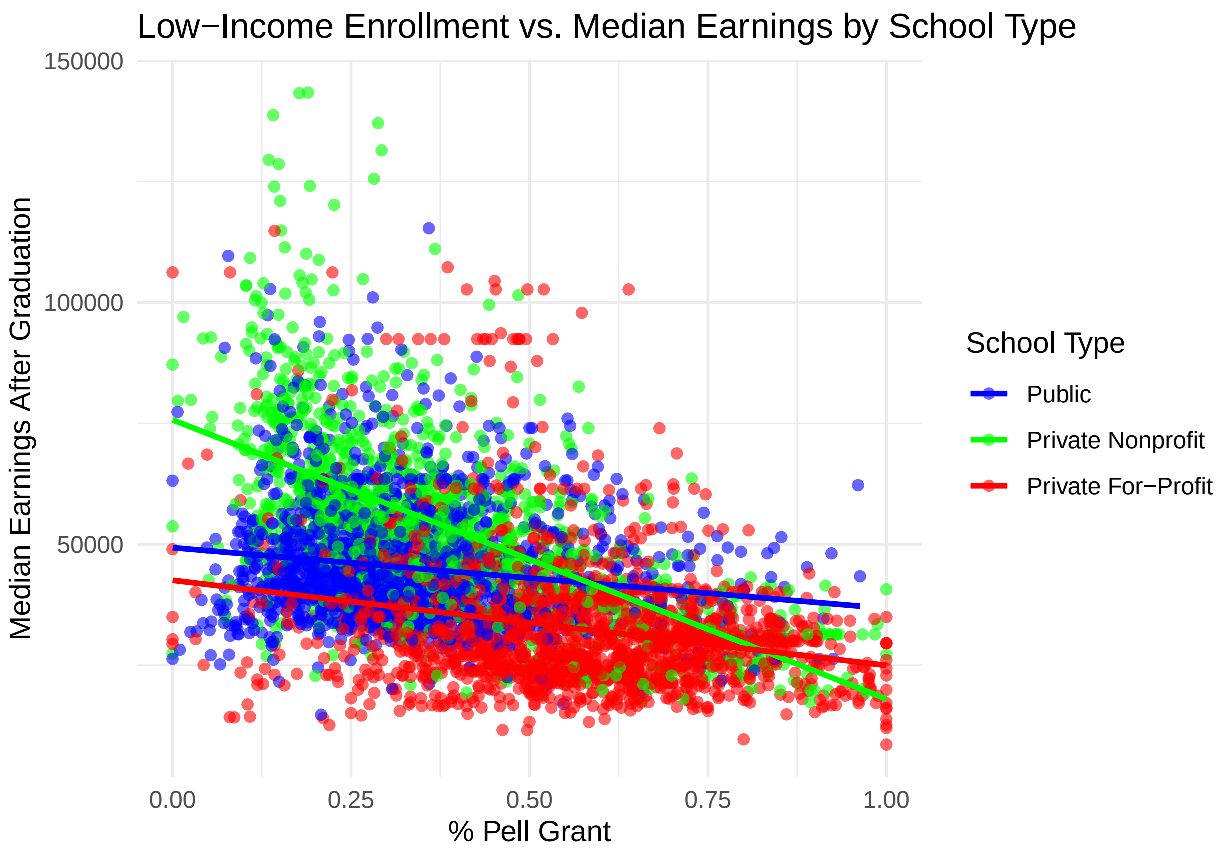
```
## Warning: Removed 4157 rows containing non-finite outside the scale range
## (‘stat_ydensity()‘).
## Warning: Removed 4157 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



#K Regressions - Low Income Enrollment vs. Median Earnings by School Type

```
ggplot(
  data = data %>% filter(!is.na(PCTPELL), !is.na(MD_EARN_WNE_P10), !is.na(CONTROL)),
  aes(x = PCTPELL, y = MD_EARN_WNE_P10, color = as.factor(CONTROL))
) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_manual(
    values = c("1" = "blue", "2" = "green", "3" = "red"),
    labels = c("Public", "Private Nonprofit", "Private For-Profit"),
    name = "School Type"
  ) +
  labs(
    title = "Low-Income Enrollment vs. Median Earnings by School Type",
    x = "% Pell Grant",
    y = "Median Earnings After Graduation"
  ) +
  theme_minimal()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```


Low−Income Enrollment vs. Median Earnings by School Type
Median Earnings After Graduation
150000
100000
50000
0.00
0.25
0.50
0.75
1.00
% Pell Grant
School Type
Public
Private Nonprofit
Private For−Profit