State Average Test Score Analysis

Anthony Andino

2025-02-14

```
options(repos = "https://cran.rstudio.com/")
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Load Libraries

```
# Install packages
# install.packages("ggrepel")
```

```
# Provides access to case1201 dataset (State Average SAT Scores)
library(Sleuth3)
```

```
## Warning: package 'Sleuth3' was built under R version 4.3.2
```

```
# Visualization package
library(ggplot2)

# Data manipulation package to summarize, filter, arrange, and transform datasets.
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# ggplot2 extension that improves text labeling in visualizations and prevents overlapping
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 4.3.3
```

```
print('Libraries are Downloaded and Ready')
```

```
## [1] "Libraries are Downloaded and Ready"
```

Explore Dataset

```
head(case1201)
```

```
##         State  SAT Takers Income Years Public Expend Rank
## 1        Iowa 1088      3    326 16.79   87.8  25.60 89.7
## 2 SouthDakota 1075      2    264 16.07   86.2  19.95 90.6
## 3 NorthDakota 1068      3    317 16.57   88.3  20.62 89.8
## 4      Kansas 1045      5    338 16.30   83.9  27.14 86.3
## 5    Nebraska 1045      5    293 17.25   83.6  21.05 88.5
## 6     Montana 1033      8    263 15.91   93.7  29.48 86.4
```

```
summary(case1201)
```

```
##         State         SAT             Takers          Income     
##  Alabama   : 1   Min.   : 790.0   Min.   : 2.00   Min.   :208.0  
##  Alaska    : 1   1st Qu.: 889.2   1st Qu.: 6.25   1st Qu.:261.5  
##  Arizona   : 1   Median : 966.0   Median :16.00   Median :295.0  
##  Arkansas  : 1   Mean   : 947.9   Mean   :26.22   Mean   :294.0  
##  California: 1   3rd Qu.: 998.5   3rd Qu.:47.75   3rd Qu.:325.0  
##  Colorado  : 1   Max.   :1088.0   Max.   :69.00   Max.   :401.0  
##  (Other)   :44                                                   
##      Years           Public          Expend           Rank      
##  Min.   :14.39   Min.   :44.80   Min.   :13.84   Min.   :69.80  
##  1st Qu.:15.91   1st Qu.:76.92   1st Qu.:19.59   1st Qu.:74.03  
##  Median :16.36   Median :80.80   Median :21.61   Median :80.85  
##  Mean   :16.21   Mean   :81.20   Mean   :22.97   Mean   :79.99  
##  3rd Qu.:16.76   3rd Qu.:88.25   3rd Qu.:26.39   3rd Qu.:85.83  
##  Max.   :17.41   Max.   :97.00   Max.   :50.10   Max.   :90.60  
## 
```

Summary Insights

SAT Scores

- The median SAT score (966) is slightly higher than the mean (947.9), suggesting a slight skew towards lower scores.

Test Takers

- There is a wide range of test takers per state (Min: 2, Max: 69).
- Some states have a much higher number of test-takers suggested by the mean (26.22) and median (16).

Median Income

- Income range from $208K to $401K can be assumed to be upper-income households
- A close mean (294) and median (295) suggest a relatively symmetric distribution.

Years of Education

- Average years of education per state is 16.21 years (approximately college sophomore)
- The 14.39-17.41 range indicated some states have a much higher average educational attainment than others.

Public School Percentages

- The 81.20% mean suggest indicates that in most states, the majority of students attend public schools.
- The percentages of students in public schools ranges from 44.8% - 97%.

Education Expenditure per State

- The range is $13.84K and $50.10K per student
- Median (21.61) and Mean (22.97) suggest most states allocate similar funding. However, few states invest significantly more.

Median Percentile Rank

- The rank varies from 69.8 to 90.6, meaning there is a clear disparity in performance.
- The mean rank (79.99) and median (80.85) suggest that most states cluster around the middle, but a few outperform significantly.

Questions?

1. Does higher income correlate with higher SAT scores?
2. Do states that spend more per student see better SAT performance?
3. Are states with a higher percentage of public school students performing worse or better on the SAT?
4. Which states outperform their expected SAT scores based on income and education spending?

Question 1: Does higher household income correlate with higher SAT Scores?

The assumption is that higher income suggest more educational resources to prepare for the SAT. We can see if this assumption is true based on a correlation test.

```r
income_SAT_correlation <- cor(case1201$Income, case1201$SAT, use = "complete.obs")
print(paste("Correlation between Median Household Income and SAT Scores is:", income_SAT_correlation))
```

```
## [1] "Correlation between Median Household Income and SAT Scores is: 0.584466574225131"
```

The correlation of 0.5844 indicates a moderate positive relationship between Income and SAT Scores.

- As income increases, SAT scores tend to increase as well.
- The correlation is not extremely strong, but it does indicate that higher-income states generally have higher SAT scores.

But, is this correlation statistical significant?

```r
cor.test(case1201$Income, case1201$SAT)
```

```
##
##  Pearson's product-moment correlation
##
## data:  case1201$Income and case1201$SAT
## t = 4.9904, df = 48, p-value = 8.329e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3655958 0.7420878
## sample estimates:
##       cor
## 0.5844666
```

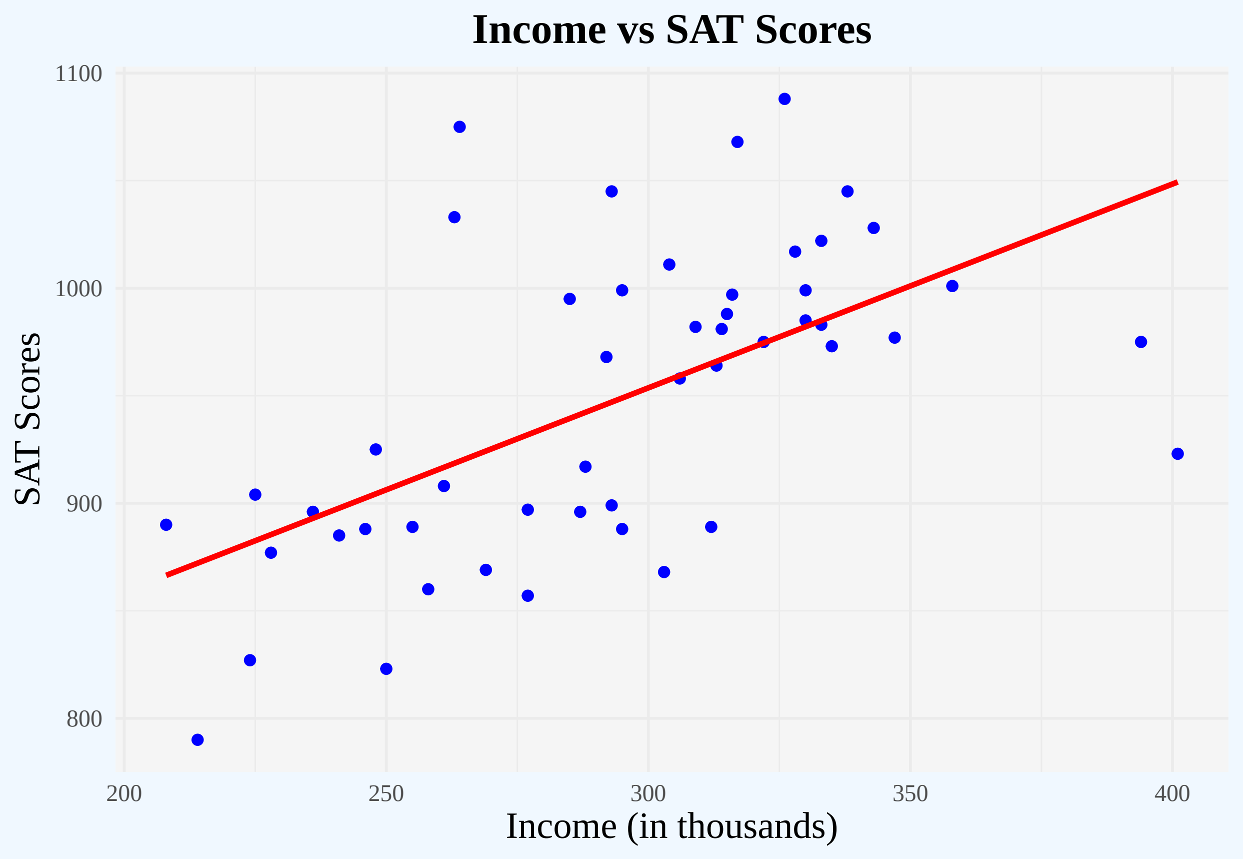
The p-value (8.32e-06) is extremely small, meaning the correlation between income and SAT Scores is statistically significant at any reasonable significance level.

Because P< 0.05, we reject the null hypothesis (which assumes no correlation) and confirm the relationship between income and SAT scores is unlikely due to random chance.

Let’s visualize the results

```
ggplot(case1201,aes(x = Income, y = SAT)) +
  geom_point(color = 'blue') +
  geom_smooth(method = 'lm', color = 'red', se = FALSE) +
  labs(title = 'Income vs SAT Scores', x = 'Income (in thousands)', y = 'SAT Scores') +
  # Theme customization
  theme_minimal(base_family = "Times") +  # Times New Roman for all text
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Center and bold the title
    axis.title = element_text(size = 14),  # Increase axis title font size
    panel.background = element_rect(fill = "#f5f5f5", color = NA),  # Light gray background
    plot.background = element_rect(fill = "#f0f8ff", color = NA)  # Light blue outer background
  )
```

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



Let’s quantify the effect of income on SAT Scores based on the regression model!

```
income_sat_model <- lm(SAT ~ Income, data = case1201 )
summary(income_sat_model)
```

```
##
## Call:
## lm(formula = SAT ~ Income, data = case1201)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.376  -42.705   -1.628   27.030  155.476
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 669.2994    56.4364   11.86 7.15e-16 ***
## Income        0.9478     0.1899    4.99 8.33e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.09 on 48 degrees of freedom
## Multiple R-squared:  0.3416, Adjusted R-squared:  0.3279
## F-statistic:  24.9 on 1 and 48 DF,  p-value: 8.329e-06
```

KEY TAKEAWAYS

1. The **Adjusted R-squared** of 0.3279 confirms that approximately 32% of the variance in SAT scores is explained by Median Household Income.
2. The **Income coefficient** of 0.9478 indicates that for each additional unit increase in Median Household Income, SAT scores increase by approximately 0.95 points.
3. The Residual Standard Error of 58.09 means that acutal SAT scores typically deviate by about 58 points from the predicted values. Not a small error, but suggests other factors influence SAT Scores.

Question 2: Do states that spend more per student see better SAT performance?

Let's start by establishing a statistical trend with all states and then compare the top 5 vs. bottom 5 states to identify patterns and outliers.

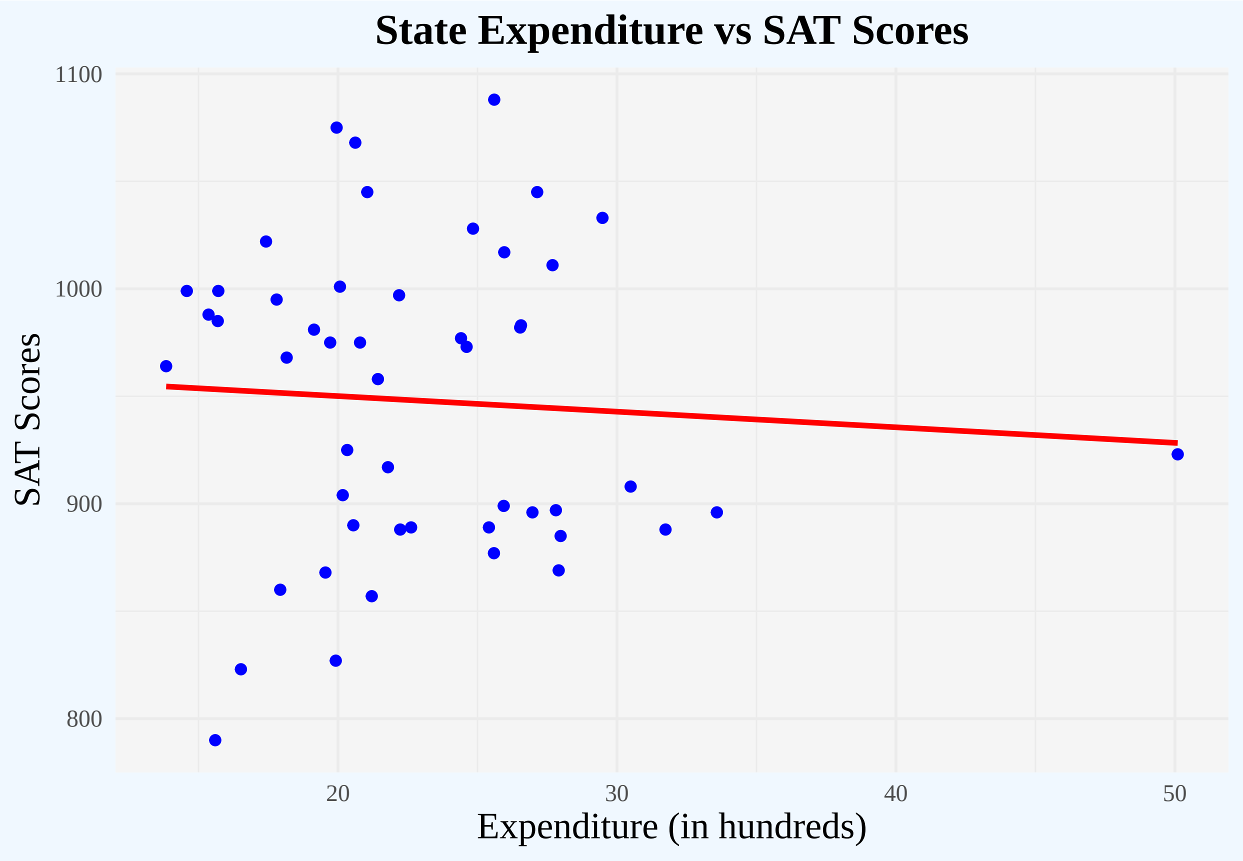
All States Expenditure Analysis

```r
cor.test(case1201$Expend, case1201$SAT)
```

```
##
##  Pearson's product-moment correlation
##
## data:  case1201$Expend and case1201$SAT
## t = -0.43649, df = 48, p-value = 0.6644
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3353560  0.2193084
## sample estimates:
##         cor
## -0.06287764
```

```r
ggplot(case1201,aes(x = Expend, y = SAT)) +
  geom_point(color = 'blue') +
  geom_smooth(method = 'lm', color = 'red', se = FALSE) +
  labs(title = 'State Expenditure vs SAT Scores', x = 'Expenditure (in hundreds)', y = 'SAT Scores') +
  # Theme customization
  theme_minimal(base_family = "Times") +  # Times New Roman for all text
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Center and bold the title
    axis.title = element_text(size = 14),  # Increase axis title font size
    panel.background = element_rect(fill = "#f5f5f5", color = NA),  # Light gray background
    plot.background = element_rect(fill = "#f0f8ff", color = NA)  # Light blue outer background
  )
```

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



KEY TAKEAWAYS

1. Since p-value (0.6644) > 0.05, we fail to reject the null hypothesis, which determines there is no statistically significant relationship between education expenditure and SAT Scores. Spending more per student does not necessarily lead to better SAT performance.
2. Correlation strength is very weak (-0.0629), meaning expenditure and SAT scores are almost uncorrelated. The negative direction suggest that higher expenditure slightly correlates with lower SAT scores, but this effect is minimal and likely due to random chance.

However, this is a general trend from all states. Is there any difference between the top 5 and bottom 5 states?

```r
# Extract top and bottom 5 states by expenditure
top_5_states <- case1201[order(-case1201$Expend), ][1:5,] # Descending order
bottom_5_states <- case1201[order(case1201$Expend), ][1:5,] # Ascenging order

# Convert factors to characters to avoid unnecessary level printing
top_5_states$State <- as.character(top_5_states$State)
bottom_5_states$State <- as.character(bottom_5_states$State)

# Print results in a readable format
cat("Top 5 States that spend the most:\n")
```

```
## Top 5 States that spend the most:
```

```r
print(top_5_states$State)
```

```
## [1] "Alaska"        "NewYork"       "Massachusetts" "Oregon"
## [5] "Montana"
```

```r
cat("\nTop 5 States that spend the least:\n")
```

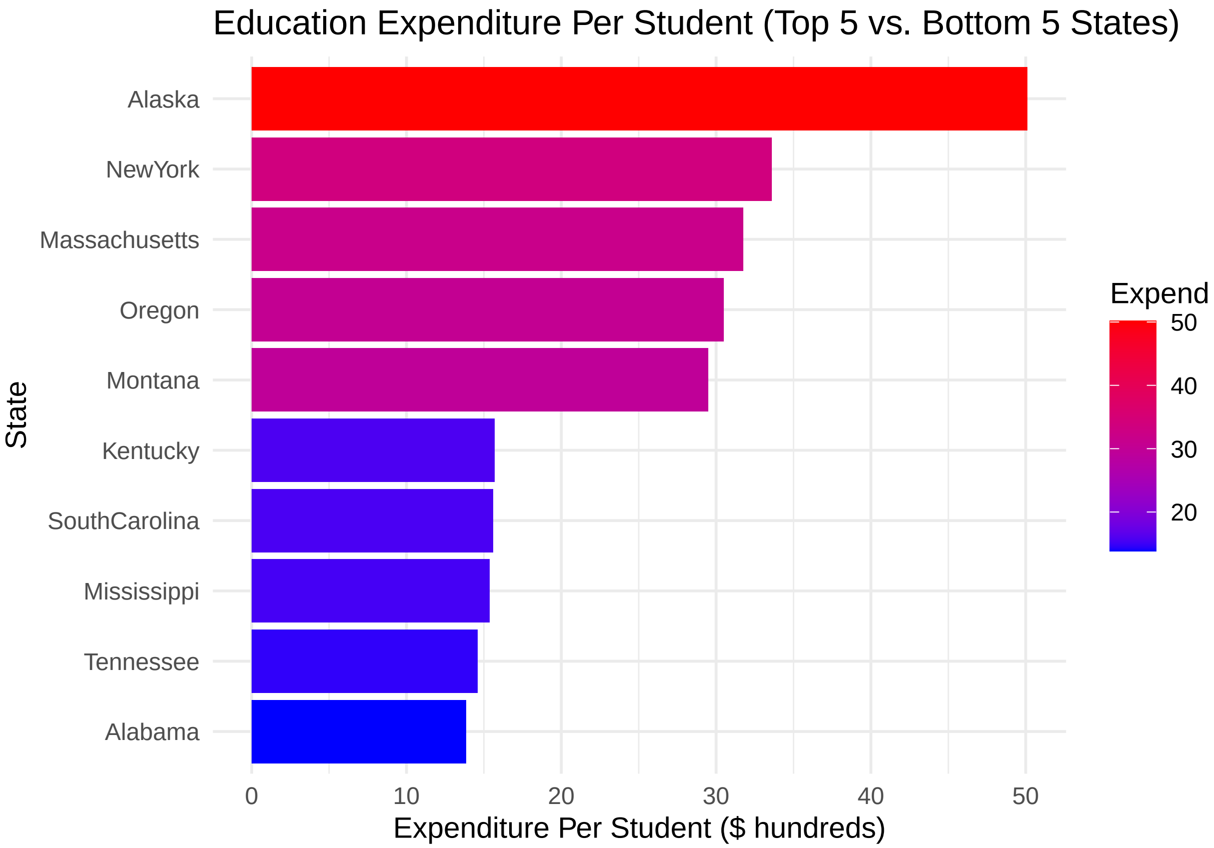
```
##
## Top 5 States that spend the least:
```

```r
print(bottom_5_states$State)
```

```
## [1] "Alabama"        "Tennessee"      "Mississippi"    "SouthCarolina"
## [5] "Kentucky"
```

```r
# Combine top and bottom 5 states
top_bottom_states <- rbind(top_5_states, bottom_5_states)

# Plot
ggplot(top_bottom_states, aes(x = reorder(State, Expend), y = Expend, fill = Expend)) +
  geom_bar(stat = "identity") +
  coord_flip() +  # Flip for readability
  scale_fill_gradient(low = "blue", high = "red") +
  labs(title = "Education Expenditure Per Student (Top 5 vs. Bottom 5 States)",
       x = "State",
       y = "Expenditure Per Student ($ hundreds)") +
  theme_minimal()
```



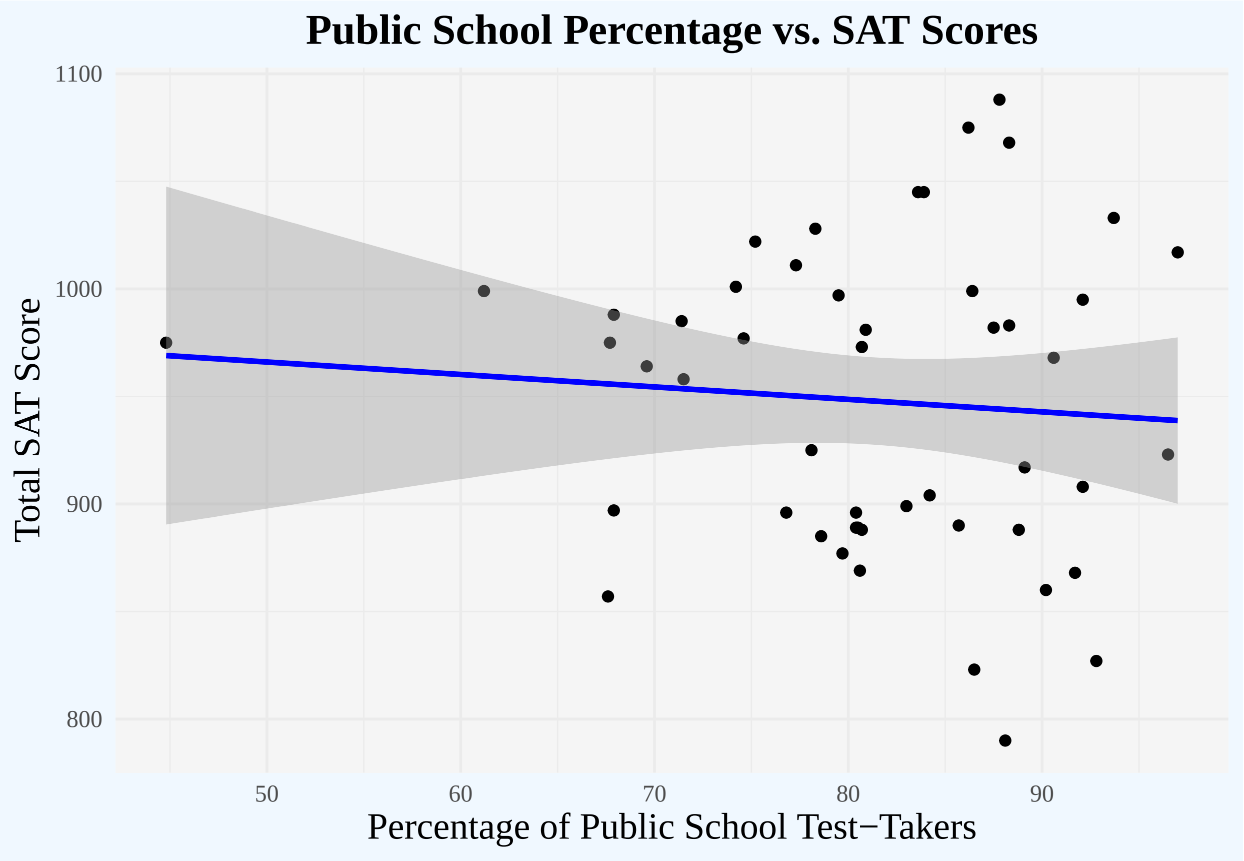
KEY TAKEAWAYS From the graph, we can see there is a significant difference between top 5 and bottom 5 state expenditure.

1. Further research demonstrates that Alaska's high state education expenditure is due to the logistical cost of operating schools in remote, rural areas across the state. Expenditure includes transportation cost, teacher turnover, and healthcare.
2. New York and Massachusetts have significantly higher expenditures compared to the lowest-spending states. This suggests that states in the Northeast prioritize education funding more than Southern states like Alabama and Tennessee.
3. The bottom-spending states (Kentucky, South Carolina, Mississippi, Tennessee, and Alabama) have relatively similar expenditure levels. This could indicate a regional trend where lower education funding is a common characteristic among Southern states.

Question 3 : Are states with a higher percentage of public school students performing worse or better on the SAT?

```r
ggplot(case1201, aes(x = Public, y = SAT)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = 'blue') +
  labs(
    title = "Public School Percentage vs. SAT Scores",
    x = "Percentage of Public School Test-Takers",
    y = "Total SAT Score"
  ) +
  # Theme customization
  theme_minimal(base_family = "Times") +  # Times New Roman for all text
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Center and bold the title
    axis.title = element_text(size = 14),  # Increase axis title font size
    panel.background = element_rect(fill = "#f5f5f5", color = NA),  # Light gray background
    plot.background = element_rect(fill = "#f0f8ff", color = NA)  # Light blue outer background
  )
```

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



```
cor.test(case1201$Public, case1201$SAT, use = "complete.obs")
```

```
##
##  Pearson's product-moment correlation
##
## data:  case1201$Public and case1201$SAT
## t = -0.55854, df = 48, p-value = 0.5791
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3508569  0.2025206
## sample estimates:
##         cor
## -0.08035688
```

KEY INSIGHTS

1. The correlation coefficient (`cor = -0.0804`) suggests a very weak negative relationship between the percentage of public school test-takers and SAT scores. However, since the value is close to 0, the relationship is negligible.
2. The p-value (`0.5791`) is much greater than 0.05, meaning we **fail to reject** the null hypothesis. This means there is **no sufficient evidence** to conclude a meaningful correlation between the percentage of public school students and SAT scores.

Question 4: Which states outperform their expected SAT scores based on income and education spending?

To answer this question, we use a regression model that estimates SAT scores based on:

- Income: higher family income = more educational resources
- Education expenditure: more funding per student might improve their SAT scores.

```
# Build a regression model to predict SAT scores based on Income and education spending
outperform_model <- lm(SAT ~ Income + Expend, data = case1201)
summary(outperform_model)
```

```
##
## Call:
## lm(formula = SAT ~ Income + Expend, data = case1201)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.364 -35.664  -0.359  25.990 151.440
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.0411    60.9868  11.446 3.44e-15 ***
## Income        0.9782     0.1907   5.130 5.41e-06 ***
## Expend       -1.6398     1.3558  -1.209    0.233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.81 on 47 degrees of freedom
## Multiple R-squared:  0.3615, Adjusted R-squared:  0.3343
## F-statistic:  13.3 on 2 and 47 DF,  p-value: 2.641e-05
```

Model Interpretation: **Intercept (698.04)**

- This means that if Income and Expenditure were both 0, the model predicts an SAT score of 698.

Income (0.9782)

- For every additional $1000 in median income, the SAT score increases by 0.98 points.
- The p-value (5.41e-06) is very small, meaning income is a statistically significant predictor of SAT scores.

Expenditure (-1.6398)

- For every additional $1000 spent per student, the SAT score decreases by 1.64 points.
- The p-value (0.233) is greater than 0.05, meaning this effect is not statistically significant.

Adjusted R² = 0.3343 (33.43%)

- Suggests that Income and Expenditure together explain only about 33.43% of SAT score variability and the model is limited as other factors matter more.

The Regression model equation is then:

Expected SAT = 698.04 + (0.9782 * Income) + (1.6398 * Expend)

```
# Apply model to each state and predicts SAT score based on income and spending
Expected_SAT <- predict(outperform_model, case1201)

# Find the residuals by comparing the actual SAT scores vs Expected SAT scores
   # positive values --> states outperform
  # negative values --> states underperform
case1201$Performance_Difference <- case1201$SAT - Expected_SAT
```

```r
# Sort states by performance difference
top_states <- case1201[order(-case1201$Performance_Difference), ] # Descending

# print Bottom States
head(top_states, 10)
```

```
##          State  SAT Takers Income Years Public Expend Rank
## 2  SouthDakota 1075      2    264 16.07   86.2  19.95 90.6
## 6      Montana 1033      8    263 15.91   93.7  29.48 86.4
## 1         Iowa 1088      3    326 16.79   87.8  25.60 89.7
## 5     Nebraska 1045      5    293 17.25   83.6  21.05 88.5
## 3  NorthDakota 1068      3    317 16.57   88.3  20.62 89.8
## 10   Wisconsin 1011     10    304 16.85   77.3  27.69 84.2
## 4       Kansas 1045      5    338 16.30   83.9  27.14 86.3
## 15       Idaho  995      7    285 16.18   92.1  17.80 85.9
## 9      Wyoming 1017      5    328 16.01   97.0  25.96 87.5
## 12    Arkansas  999      4    295 15.49   86.4  15.71 89.2
##    Performance_Difference
## 2               151.44013
## 6               126.04519
## 1               113.05937
## 5                94.87746
## 3                93.69673
## 10               61.00580
## 4                60.84678
## 15               47.37346
## 9                40.69338
## 12               38.16484
```

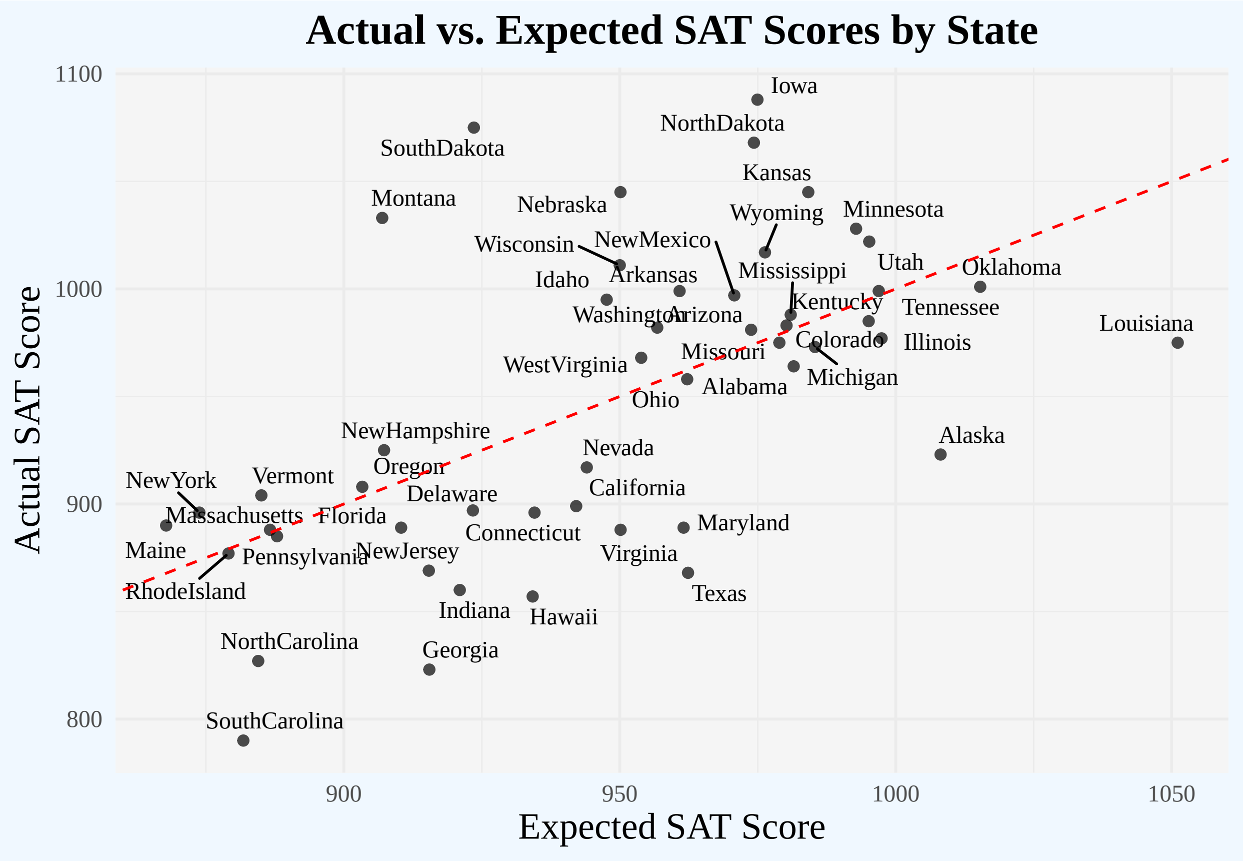
To better understand which states outperform expectations, we created a scatter plot with expected SAT scores on the X-axis and actual SAT scores on the Y-axis. A red reference line represents the points where actual scores match expected scores.

- Points above the red line → States that scored higher than expected
- Points below the red line → States that scored lower than expected

This visualization helps identify states that exceed or underperform relative to their predicted SAT scores, providing insights into potential influencing factors.

```r
ggplot(case1201, aes(x = Expected_SAT, y = SAT, label = State)) +
  geom_point(alpha = 0.7) + # adjust transparency
  geom_text_repel(size = 3, max.overlaps = 15, family = "Times") + # prevent overlap
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Actual vs. Expected SAT Scores by State",
       x = "Expected SAT Score",
       y = "Actual SAT Score") +

# Theme customization
  theme_minimal(base_family = "Times") +  # Times New Roman for all text
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Center and bold the title
    axis.title = element_text(size = 14),  # Increase axis title font size
    panel.background = element_rect(fill = "#f5f5f5", color = NA),  # Light gray background
    plot.background = element_rect(fill = "#f0f8ff", color = NA)  # Light blue outer background
  )
```



KEY TAKEAWAYS

- This is just a basic model and it can be improved by adding other factors.
- States that Outperform - > Iowa , South Dakota, North Dakota, Nebraska, and Kansas
- States that Underperform - > South Carolina, Georgia, North Carolina, Hawaii, and Indiana
- Interestingly, states with highest expenditures such as Massachusetts and New York did not significantly outperform.

Improvements and Next Steps

The current prediction model explains only 33.43% of the variance in SAT scores, indicating that additional factors could be incorporated to enhance its predictive power. Exploring socioeconomic variables, school funding, or student-teacher ratios might provide deeper insights.

This project primarily served as a practice exercise for applying R programming and data analytics techniques. So far, four key questions have been addressed using this dataset. However, further analysis could uncover additional relationships and testing methods to expand our understanding of SAT score patterns across states. Future work may involve experimenting with different modeling techniques, feature engineering, or incorporating external datasets to improve predictions.