AnalysisReport1_Pranga_Marcus

Marcus Pranga

2024-09-23

Data

```
# Load data
# change the path to where my dataset is located
college_admissions <-
  read.csv("C:\\Users\\djmar\\OneDrive\\Desktop\\DS 3100\\college admissions.csv")

# Report dimensions of the data
dim(college_admissions)
```

```
## [1] 1534  108
```

For this assignment, I worked with the College Admissions dataset from Kaggle, which can be found here: https://www.kaggle.com/datasets/samsonqian/college-admissions. It provides valuable insight into the college admissions process and the dataset contains 1,517 colleges and their records related to college admissions, enrollment, and financial aid for various universities across the United States. Key variables include the total number of applicants, admissions, and enrolled students for each institution, along with SAT/ACT scores, financial aid statistics, and other relevant metrics.

Research Question

In the selected dataset, I aim to determine which variable—the percentage of freshmen submitting SAT scores, the percentage of freshmen submitting ACT scores, or the percentage of freshmen receiving financial aid—has the highest correlation with a university's admission rate.

Variables of interest

1. **Percent of freshmen submitting SAT scores**:
   - **Variable type**: Numeric (percentage)
   - **Description**: The percentage of freshmen at each university who submitted SAT scores.
   - **Range**: 0-100 (percent).
   - **Units**: Percentage of freshmen.
2. **Percent of freshmen submitting ACT scores**:

- **Variable type**: Numeric (percentage)
- **Description**: The percentage of freshmen at each university who submitted SAT scores.
- **Range**: Typically 0-100 (percent).
- **Units**: Percentage of freshmen.

3. **Percent of freshmen receiving any financial aid**:

- **Variable type**: Numeric (percentage)
- **Description**: The percentage of freshmen receiving any form of financial aid at each institution.
- **Range**: 0-100 (percent).
- **Units**: Percentage of freshmen.

4. **Percent admitted - total**:

- **Variable type**: Numeric (percentage)
- **Description**: The percentage of total applicants admitted to the university.
- **Range**: 0-100 (percent).
- **Units**: Percentage of applicants.

Data Wrangling

```
# Load the necessary libraries
library(readr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## 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
## -- 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 become errors
```

```
# Select relevant columns
college_clean <- college_admissions %>%
  select(Percent.of.freshmen.submitting.SAT.scores,
         Percent.of.freshmen.submitting.ACT.scores,
         Percent.of.freshmen.receiving.any.financial.aid,
         Percent.admitted...total)
```

```
# Check for missing values
colSums(is.na(college_clean))
```

```
##       Percent.of.freshmen.submitting.SAT.scores
##                                             277
##       Percent.of.freshmen.submitting.ACT.scores
##                                             275
## Percent.of.freshmen.receiving.any.financial.aid
##                                              42
##                         Percent.admitted...total
##                                             158
```

```r
# Remove rows with missing data
college_clean <- college_clean %>%
  drop_na()
```

Checked for missing values and removed rows with any missing data to ensure a complete dataset for analysis.

```r
# Rename columns for easier reference
colnames(college_clean) <- c("SAT_percent", "ACT_percent",
                             "Financial_Aid_percent", "Admission_Rate")
```

Renamed columns for clarity and easier reference during analysis.

```r
# Recreate Admission Rate categories after filtering
college_clean <- college_clean %>%
  mutate(Admission_Rate_Category = case_when(
    Admission_Rate >= 70 ~ "High",
    Admission_Rate >= 40 & Admission_Rate < 70 ~ "Medium",
    Admission_Rate < 40 ~ "Low"
  ))

# Check the categories
table(college_clean$Admission_Rate_Category)
```

```
##
##   High    Low Medium
##    533    126    596
```

Renamed columns for clarity and easier reference during analysis.

```r
# Log transform SAT and ACT percentages (handling zero values)
college_clean <- college_clean %>%
  mutate(SAT_percent_log = log1p(SAT_percent),
         ACT_percent_log = log1p(ACT_percent))
```

Applied a log transformation to SAT and ACT scores to address potential skewness and stabilize variance.

```r
# Ensure all columns are numeric
college_clean <- college_clean %>%
  mutate(across(everything(), as.numeric))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `across(everything(), as.numeric)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```r
# View summary of cleaned data and transformation
summary(college_clean)
```

```
##   SAT_percent      ACT_percent     Financial_Aid_percent Admission_Rate
##  Min.   :  0.00   Min.   :  0.00   Min.   : 41.00        Min.   :  6.00
```

```
##  1st Qu.: 16.00   1st Qu.: 26.00   1st Qu.: 87.00        1st Qu.: 54.00
##  Median : 65.00   Median : 48.00   Median : 95.00        Median : 66.00
##  Mean   : 53.88   Mean   : 53.01   Mean   : 90.82        Mean   : 64.39
##  3rd Qu.: 86.00   3rd Qu.: 87.00   3rd Qu.: 99.00        3rd Qu.: 77.00
##  Max.   :100.00   Max.   :100.00   Max.   :100.00        Max.   :100.00
##
##  Admission_Rate_Category SAT_percent_log ACT_percent_log
##  Min.   : NA             Min.   :0.000   Min.   :0.000
##  1st Qu.: NA             1st Qu.:2.833   1st Qu.:3.296
##  Median : NA             Median :4.190   Median :3.892
##  Mean   :NaN             Mean   :3.538   Mean   :3.663
##  3rd Qu.: NA             3rd Qu.:4.466   3rd Qu.:4.477
##  Max.   : NA             Max.   :4.615   Max.   :4.615
##  NA's   :1255
```

Ensured all columns are of numeric data type for accurate calculations in subsequent analyses.Moreover, I also generated a summary of the cleaned and transformed dataset to verify the structure and readiness for analysis.

Analysis

I will be conducting a correlation analysis. Specifically, we will calculate the Pearson correlation between the admission rate (dependent variable) and the three independent variables: Percent of freshmen submitting SAT scores, Percent of freshmen submitting ACT scores, and Percent of freshmen receiving financial aid.

```r
# Calculate correlation between Admission Rate and independent variables
correlation_matrix <- college_clean %>%
  select(SAT_percent, ACT_percent, Financial_Aid_percent, Admission_Rate) %>%
  cor()

# Display the correlation matrix
correlation_matrix
```

```
##                       SAT_percent ACT_percent Financial_Aid_percent
## SAT_percent             1.0000000  -0.8908586            -0.2280235
## ACT_percent            -0.8908586   1.0000000             0.2079630
## Financial_Aid_percent  -0.2280235   0.2079630             1.0000000
## Admission_Rate         -0.1379879   0.1024765             0.3603441
##                       Admission_Rate
## SAT_percent              -0.1379879
## ACT_percent               0.1024765
## Financial_Aid_percent     0.3603441
## Admission_Rate            1.0000000
```

```r
# Extract correlations with Admission Rate
admission_rate_correlations <- correlation_matrix["Admission_Rate", ]
admission_rate_correlations
```

```
##           SAT_percent           ACT_percent Financial_Aid_percent
##            -0.1379879             0.1024765             0.3603441
##        Admission_Rate
##             1.0000000
```

```
#The resulting correlation matrix will show how strongly each variable correlates with the others.
```

Extra Data Wrangling

I will be standardizing the variables because they are on different scales (percentages for SAT, ACT submissions, and financial aid vs. admission rate). Standardizing allows for a direct comparison of the effects of each variable on the admission rate by placing them on the same scale (mean = 0, standard deviation = 1). This helps determine which variable has the most relative impact on admission rates, making the analysis more robust and interpretable.

```
# Standardize the variables and store in a new dataframe
college_clean_standardized <- college_clean %>%
  mutate(SAT_percent_scaled = scale(SAT_percent),
         ACT_percent_scaled = scale(ACT_percent),
         Financial_Aid_percent_scaled = scale(Financial_Aid_percent),
         Admission_Rate_scaled = scale(Admission_Rate))

# View the summary of standardized data
summary(college_clean_standardized)
```

```
##   SAT_percent      ACT_percent     Financial_Aid_percent Admission_Rate
##  Min.   :  0.00   Min.   :  0.00   Min.   : 41.00        Min.   :  6.00
##  1st Qu.: 16.00   1st Qu.: 26.00   1st Qu.: 87.00        1st Qu.: 54.00
##  Median : 65.00   Median : 48.00   Median : 95.00        Median : 66.00
##  Mean   : 53.88   Mean   : 53.01   Mean   : 90.82        Mean   : 64.39
##  3rd Qu.: 86.00   3rd Qu.: 87.00   3rd Qu.: 99.00        3rd Qu.: 77.00
##  Max.   :100.00   Max.   :100.00   Max.   :100.00        Max.   :100.00
##
##  Admission_Rate_Category SAT_percent_log ACT_percent_log SAT_percent_scaled.V1
##  Min.   : NA             Min.   :0.000   Min.   :0.000   Min.   :-1.5666428
##  1st Qu.: NA             1st Qu.:2.833   1st Qu.:3.296   1st Qu.:-1.1014364
##  Median : NA             Median :4.190   Median :3.892   Median : 0.3232582
##  Mean   :NaN             Mean   :3.538   Mean   :3.663   Mean   : 0.0000000
##  3rd Qu.: NA             3rd Qu.:4.466   3rd Qu.:4.477   3rd Qu.: 0.9338416
##  Max.   : NA             Max.   :4.615   Max.   :4.615   Max.   : 1.3408972
##  NA's   :1255
##  ACT_percent_scaled.V1 Financial_Aid_percent_scaled.V1 Admission_Rate_scaled.V1
##  Min.   :-1.6744314    Min.   :-4.340426               Min.   :-3.1420864
##  1st Qu.:-0.8531480    1st Qu.:-0.332544               1st Qu.:-0.5591267
##  Median :-0.1582160    Median : 0.364479               Median : 0.0866132
##  Mean   : 0.0000000    Mean   : 0.000000               Mean   : 0.0000000
##  3rd Qu.: 1.0737091    3rd Qu.: 0.712991               3rd Qu.: 0.6785415
##  Max.   : 1.4843507    Max.   : 0.800118               Max.   : 1.9162097
##
```

Visualization

I chose scatterplots for most of my graphs because they are effective for visualizing relationships between two continuous variables, such as the percentage of freshmen submitting SAT or ACT scores and the admission rate. Scatterplots allow me to see patterns, trends, or outliers in the data. By adding regression lines, I

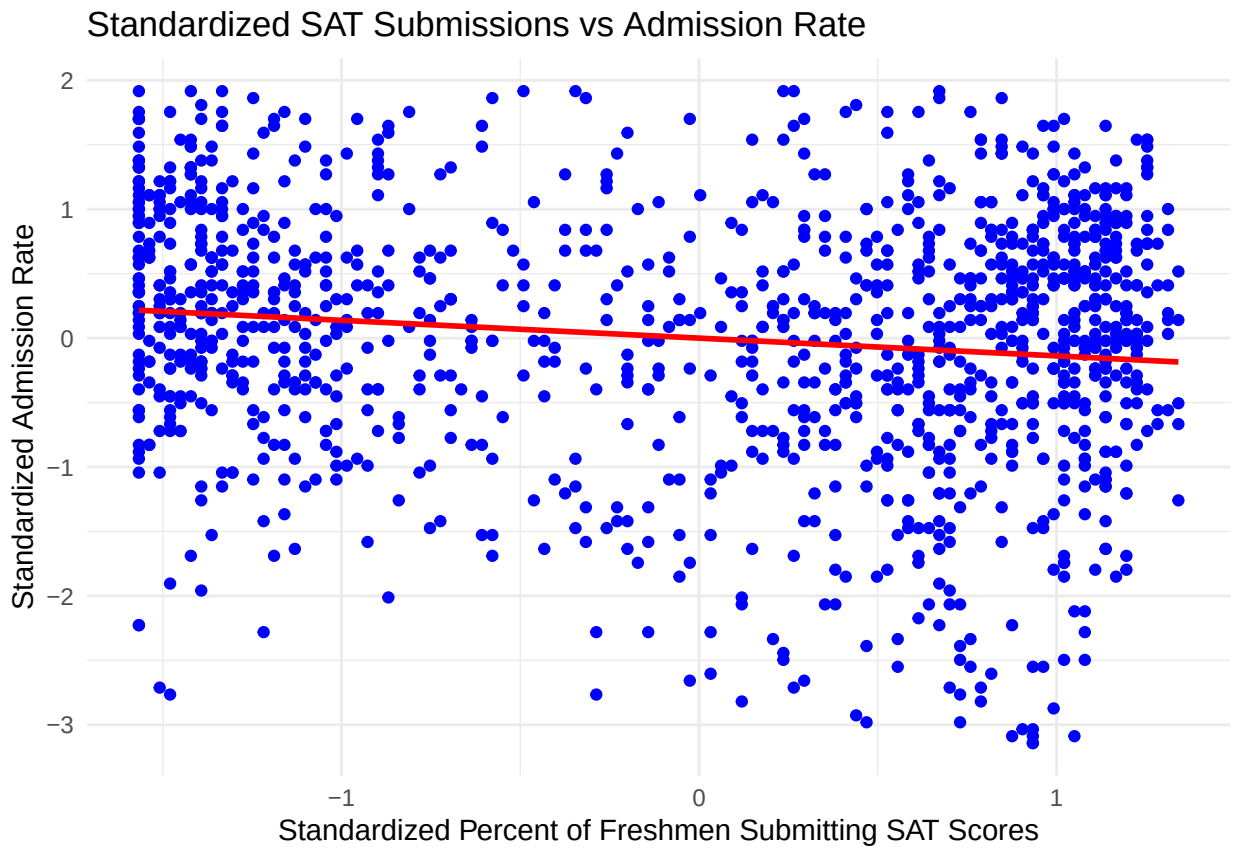
could clearly represent the strength and direction of these relationships, which was key to answering my research question about which factor—test submissions or financial aid—had the strongest correlation with admission rates.

Scatter Plot: SAT Submissions vs Admission Rate (Standardized)

```
# Scatter plot of SAT Submissions vs Admission Rate (standardized)
library(ggplot2)

ggplot(college_clean_standardized, aes(x = SAT_percent_scaled,
                                       y = Admission_Rate_scaled)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Standardized SAT Submissions vs Admission Rate",
       x = "Standardized Percent of Freshmen Submitting SAT Scores",
       y = "Standardized Admission Rate") +
  theme_minimal()
```

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



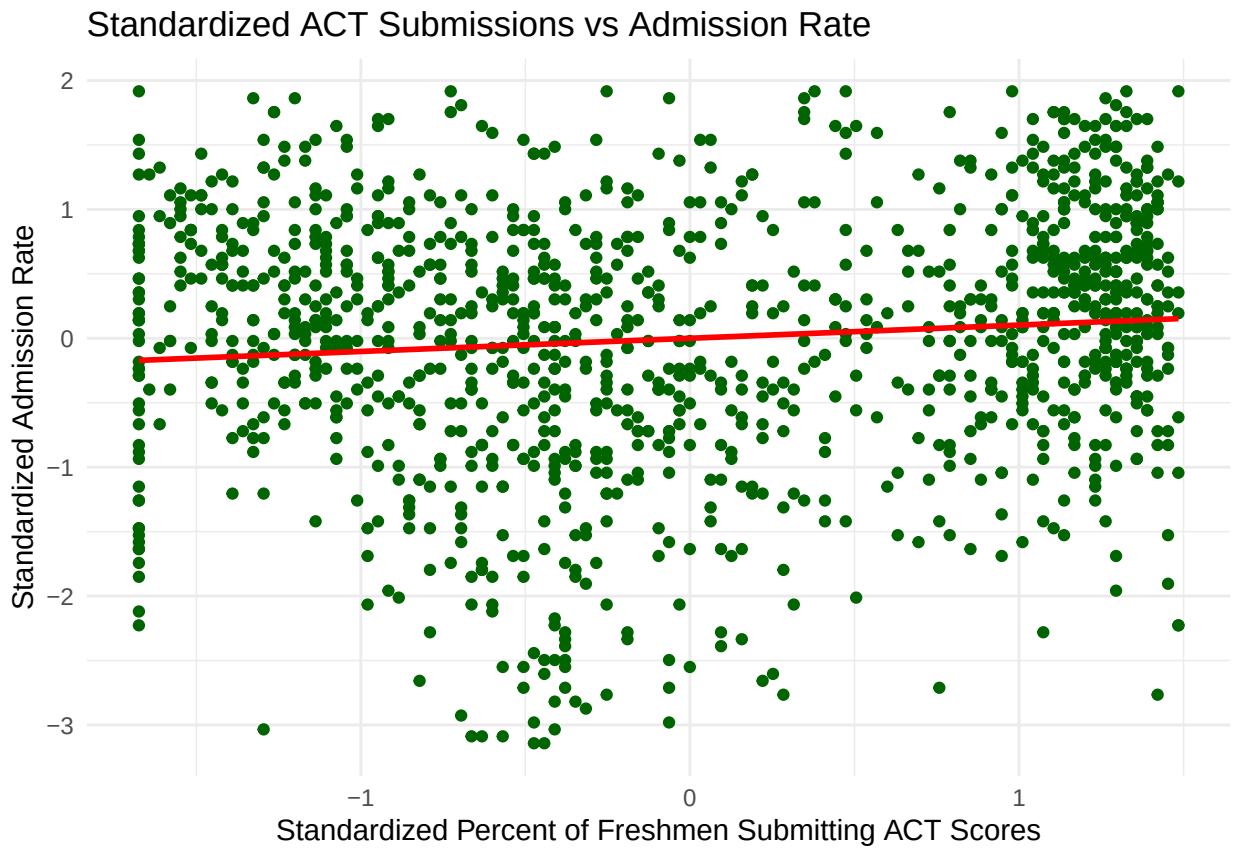
This scatter plot indicates a weak and slightly negative correlation between the percentage of freshmen submitting SAT scores and the admission rate. While standardized tests have historically been a key component of admissions, the nearly flat red trend line suggests that SAT submissions in this dataset are not strongly associated with admission rates. This could reflect recent trends where universities de-emphasize

the SAT in favor of holistic admissions criteria. The wide scatter of points also indicates high variability in how different institutions weigh SAT scores.

Scatter plot of ACT Submissions vs Admission Rate (standardized)

```
# Scatter plot of ACT Submissions vs Admission Rate (standardized)
ggplot(college_clean_standardized, aes(x = ACT_percent_scaled,
                                       y = Admission_Rate_scaled)) +
  geom_point(color = "darkgreen") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Standardized ACT Submissions vs Admission Rate",
       x = "Standardized Percent of Freshmen Submitting ACT Scores",
       y = "Standardized Admission Rate") +
  theme_minimal()
```

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

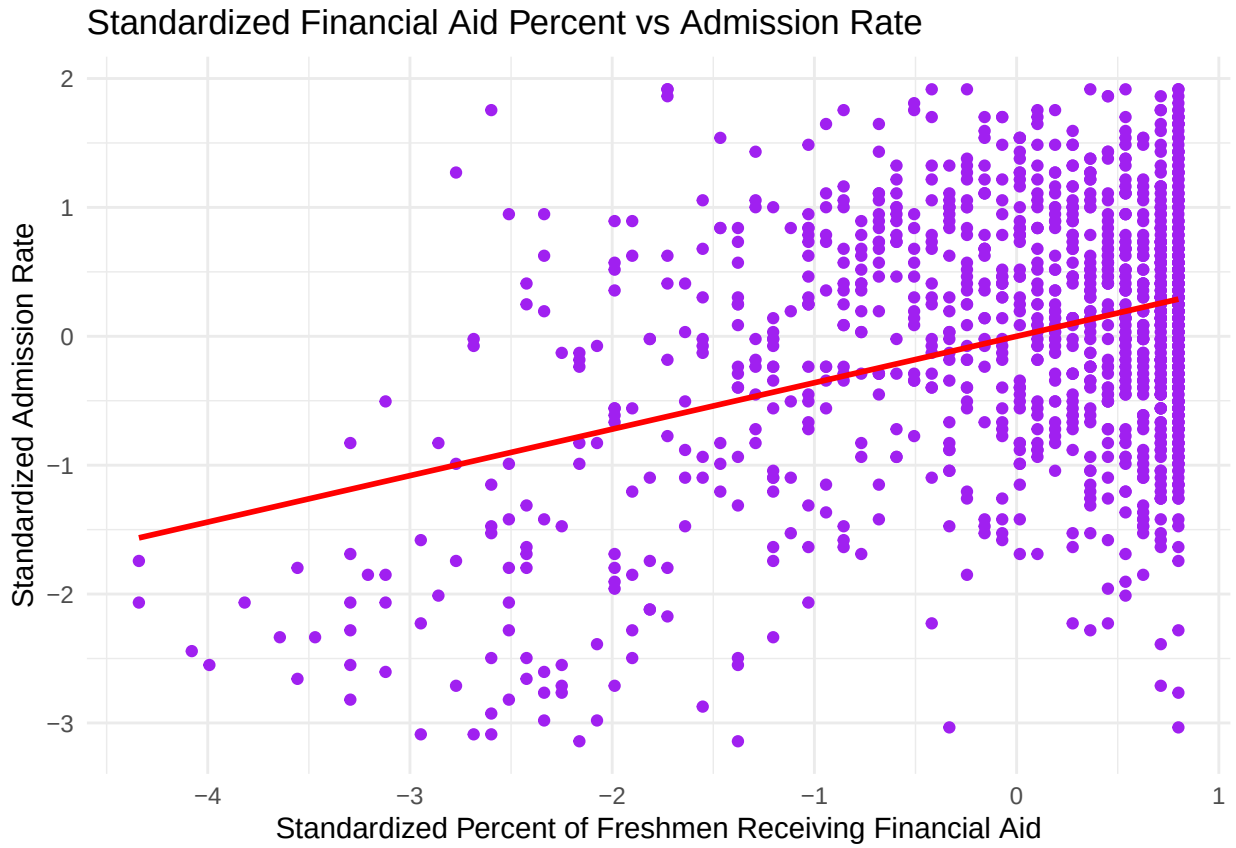


Similar to the SAT plot, this scatter plot also shows a weak positive correlation between ACT submissions and admission rates, indicated by the slight upward trend of the red line. However, the correlation is weak, and the data points remain broadly dispersed.Although there is a small positive correlation between ACT submissions and admission rates, this relationship is not particularly strong. It suggests that the percentage of students submitting ACT scores might slightly increase admission rates, but it is not a key determining factor.Although there is a small positive correlation between ACT submissions and admission rates, this relationship is not particularly strong. It suggests that the percentage of students submitting ACT scores might slightly increase admission rates, but it is not a key determining factor.

Scatter plot of ACT Submissions vs Admission Rate (standardized)

```
# Scatter plot of Financial Aid Percent vs Admission Rate (standardized)
ggplot(college_clean_standardized, aes(x = Financial_Aid_percent_scaled,
                                       y = Admission_Rate_scaled)) +
  geom_point(color = "purple") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Standardized Financial Aid Percent vs Admission Rate",
       x = "Standardized Percent of Freshmen Receiving Financial Aid",
       y = "Standardized Admission Rate") +
  theme_minimal()
```

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



In contrast to the SAT and ACT graphs, this plot shows a clearer upward trend in the relationship between the percentage of freshmen receiving financial aid and the admission rate. The red trend line indicates a moderate positive correlation, suggesting that as financial aid percentages increase, admission rates tend to increase as well.Financial aid appears to have a stronger influence on admission rates compared to SAT or ACT submissions. This suggests that universities that offer more financial aid may attract more students, leading to higher admission rates. The financial accessibility of universities is likely a significant factor influencing their admissions.

Additional Visualization

Heatmap

This visualization provides a comprehensive look at the correlations between the variables. It helps in visually comparing the strength of the relationships all in one look.

```
# Load library for heatmap
library(reshape2)
```

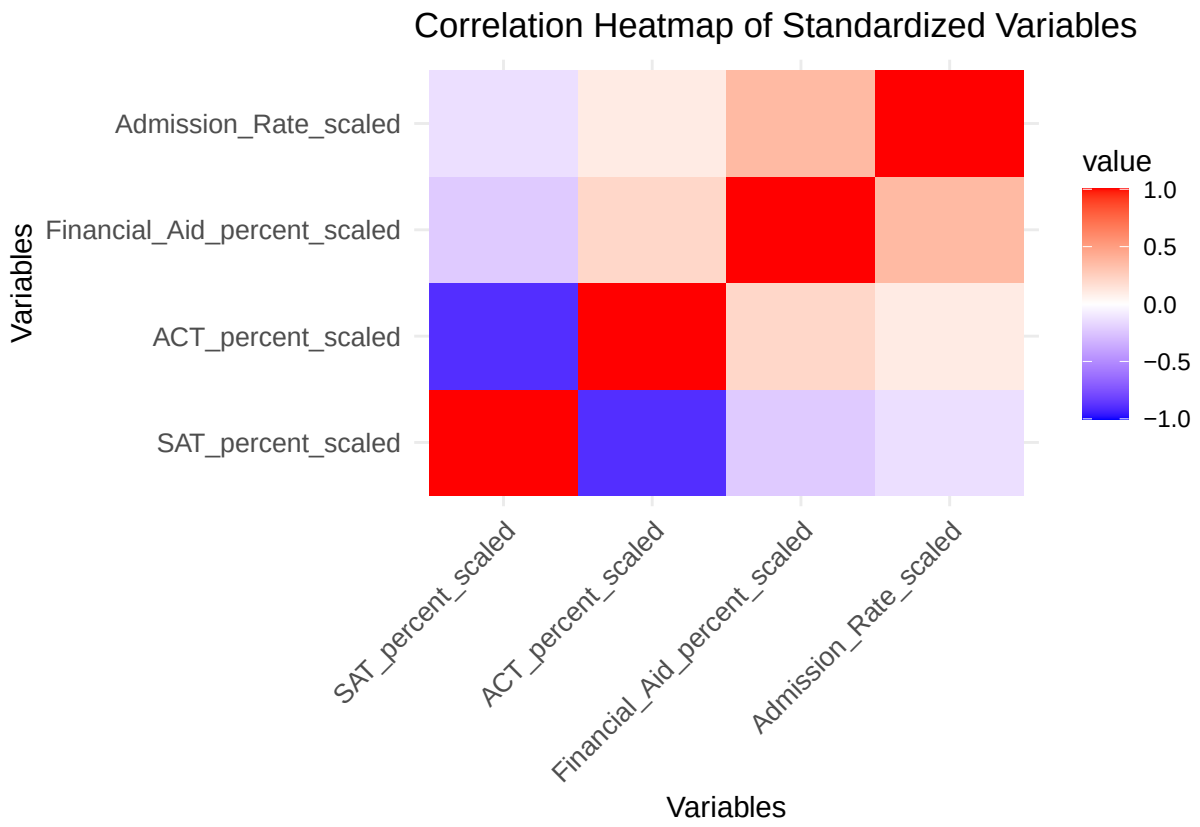
```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
# Create a correlation matrix for standardized variables
cor_matrix <- cor(select(college_clean_standardized,
                         SAT_percent_scaled, ACT_percent_scaled,
                         Financial_Aid_percent_scaled, Admission_Rate_scaled))

# Melt the correlation matrix
cor_melted <- melt(cor_matrix)

# Plot heatmap with rotated axis labels and adjusted font size
ggplot(data = cor_melted, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0,
                       limit = c(-1, 1)) +
  labs(title = "Correlation Heatmap of Standardized Variables",
       x = "Variables", y = "Variables") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 10),# Rotate x-axis
        axis.text.y = element_text(size = 10), # Adjust y-axis label size
        plot.margin = margin(10, 10, 10, 10)) # Increase plot margins
```



The heatmap shows the correlations between standardized variables. There is a strong negative correlation between SAT and ACT submissions, indicating that universities tend to prioritize one test over the other. Financial aid shows a moderate positive correlation with admission rates, suggesting that institutions offering more financial support tend to admit more students. Additionally, financial aid has a slight negative correlation with SAT submissions and a slight positive correlation with ACT submissions, reflecting some variability in the relationship between financial aid and test preferences. Overall, financial aid appears to be a more significant factor influencing admission rates compared to standardized test submissions.

Discussion & Summary

This analysis aimed to explore the relationship between university admission rates and three key factors: the percentage of freshmen submitting SAT scores, the percentage of freshmen submitting ACT scores, and the percentage of freshmen receiving financial aid. The motivation for this research stemmed from ongoing debates regarding the relevance of standardized test scores versus financial aid in the admissions process. With many universities de-emphasizing standardized tests and placing more focus on accessibility through financial aid, the goal was to identify which factor has the strongest correlation with admission rates.

The findings revealed weak correlations between the submission of standardized test scores and admission rates. Specifically, the scatterplots showed that both SAT and ACT submissions have minimal impact on the admission rates, with correlation coefficients of -0.137 for SAT submissions and 0.10 for ACT submissions. These results suggest that universities in the dataset do not heavily weigh standardized test scores when making admissions decisions. In contrast, financial aid demonstrated a stronger positive correlation with admission rates, with a correlation coefficient of 0.36, indicating that universities offering more financial aid tend to admit more students. This is a crucial finding, highlighting the growing emphasis on financial accessibility as a determinant in admissions.

A correlation heatmap further reinforced these trends, showing that financial aid had the most substantial positive correlation with admission rates among the variables analyzed. Interestingly, the heatmap also revealed a negative correlation between SAT and ACT submissions (-0.89), suggesting that universities may prioritize one test over the other. However, both test submissions had relatively little impact on admission rates compared to financial aid.

In conclusion, this analysis indicates that financial aid plays a more significant role in influencing university admission rates than standardized test submissions. This finding supports the notion that universities may prioritize financial accessibility over test scores, which could inform future admissions policies aimed at expanding access to higher education. Nonetheless, there are limitations to this analysis. The dataset does not capture other important factors such as institutional reputation, geographical location, or program-specific competitiveness, which could also impact admission rates. Future research could incorporate these additional variables and further explore how different universities balance financial aid and test scores in their admissions processes. Understanding these dynamics could lead to more equitable admissions practices that prioritize accessibility while maintaining academic standards.