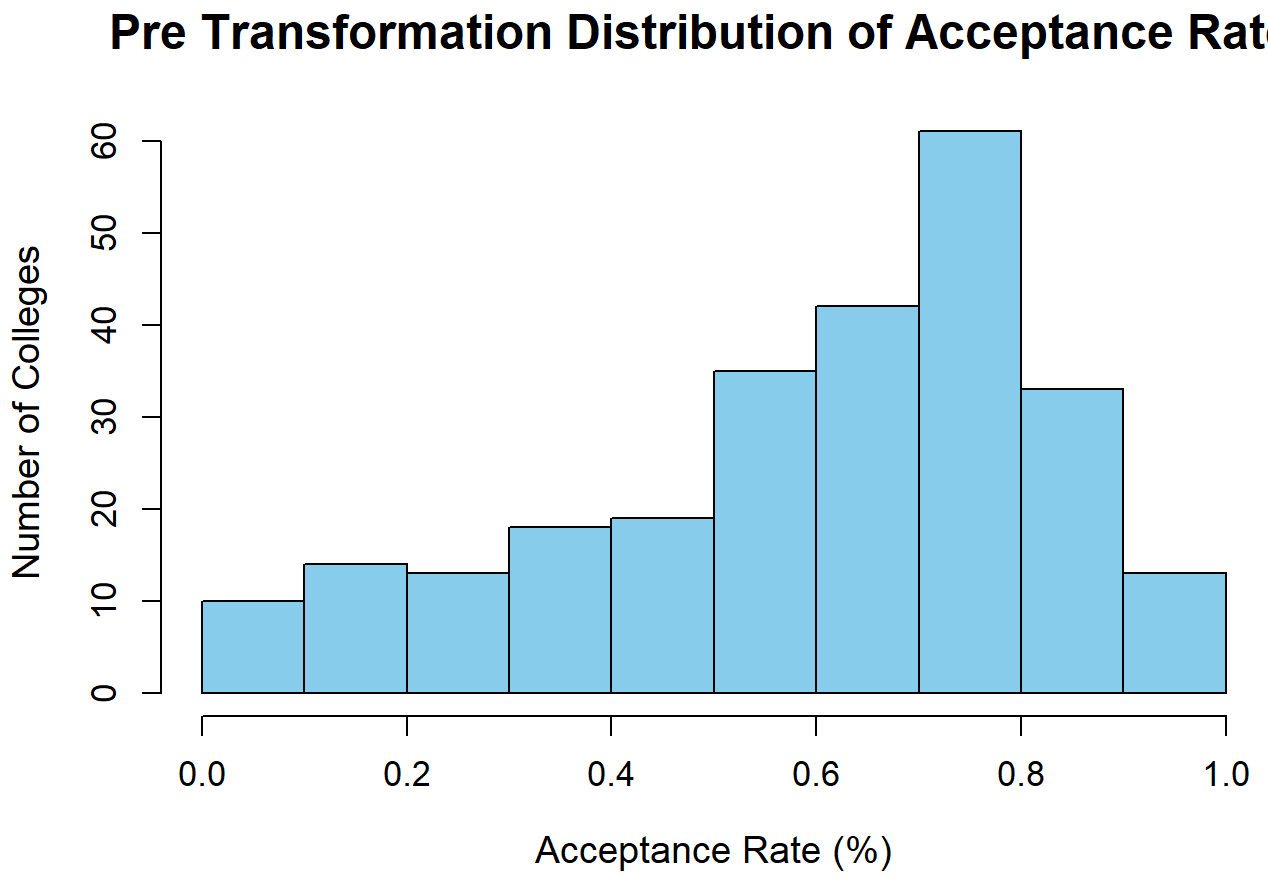
CollegeSAT

Rory Quinlan

```
college = read.table("college.csv", header = TRUE, sep = ",")
```

Create plots and fit model on raw data that uses SAT mean to predict acceptance rate

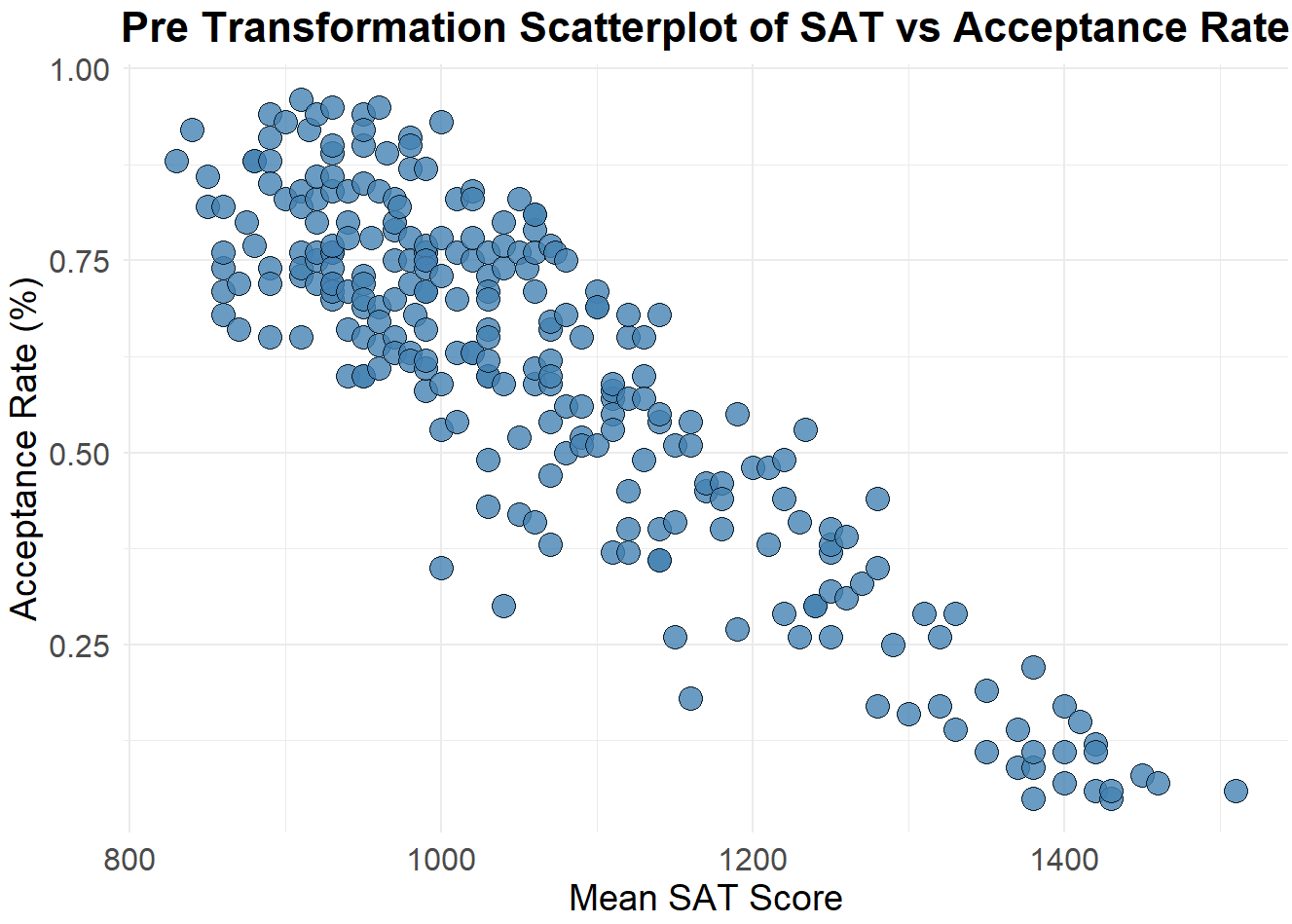
```
# Create a histogram with customized colors and labels
hist(college$acc_rate,
     col = "skyblue",  # Change bar color to a light blue
     border = "black",  # Bar border color
     main = "Pre Transformation Distribution of Acceptance Rates",  # Title
     xlab = "Acceptance Rate (%)",  # X-axis label
     ylab = "Number of Colleges",  # Y-axis label
     breaks = 10,  # Adjust number of bins as needed
     cex.main = 1.5,  # Title size
     cex.lab = 1.2,  # Axis label size
     cex.axis = 1.1  # Axis text size
)
```



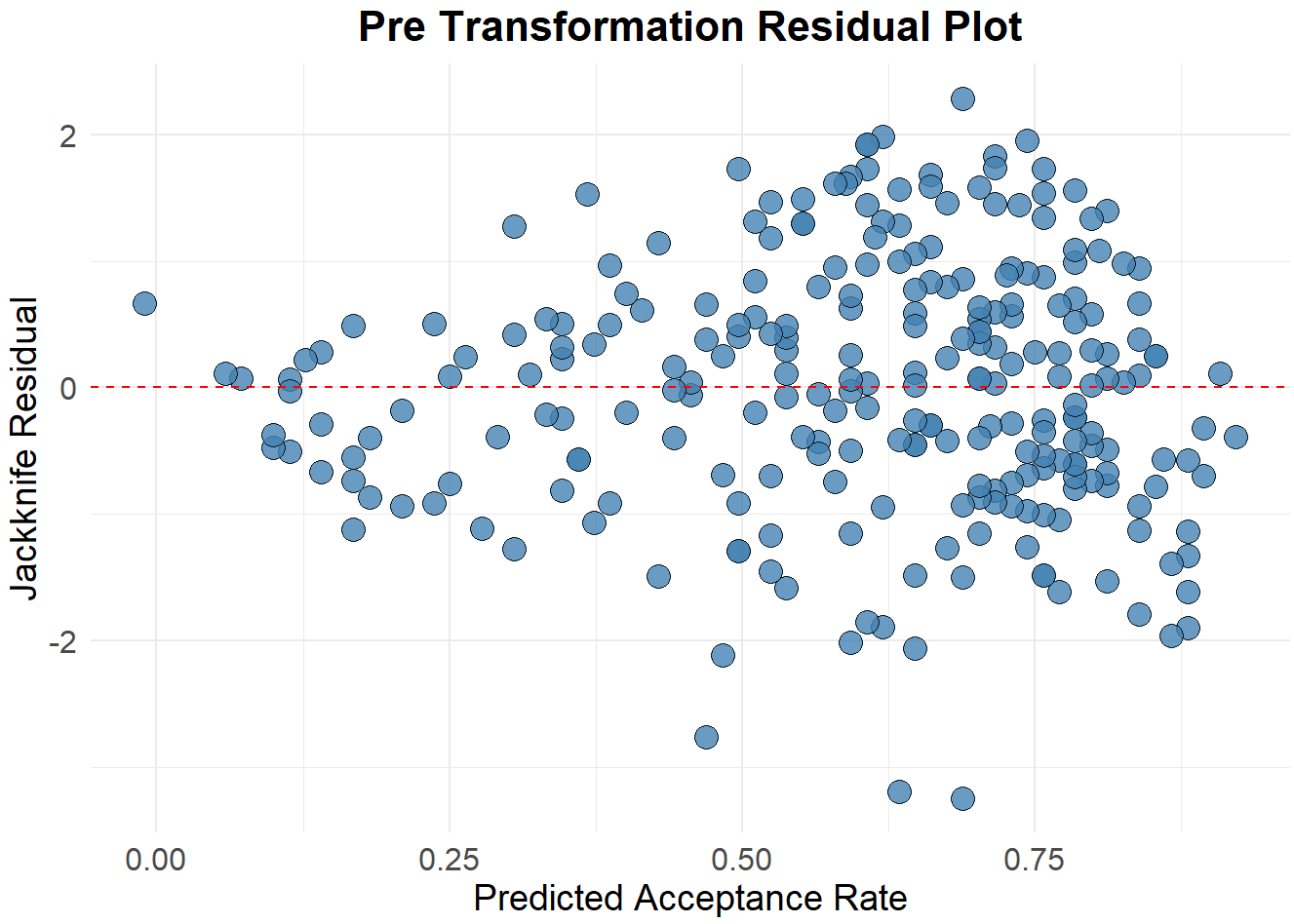

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
# Scatterplot of raw data
# Create a scatterplot with black outlines
# Create a scatterplot with black outlines
ggplot(college, aes(x = sat, y = acc_rate)) +
  geom_point(shape = 21, color = "black", fill = "steelblue", size = 4, stroke = 0.5, alpha = 0.
8) +  # Thinner outline
  labs(title = "Pre Transformation Scatterplot of SAT vs Acceptance Rate",
       x = "Mean SAT Score",
       y = "Acceptance Rate (%)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
        axis.title = element_text(size = 14),
        axis.text = element_text(size = 12))
```



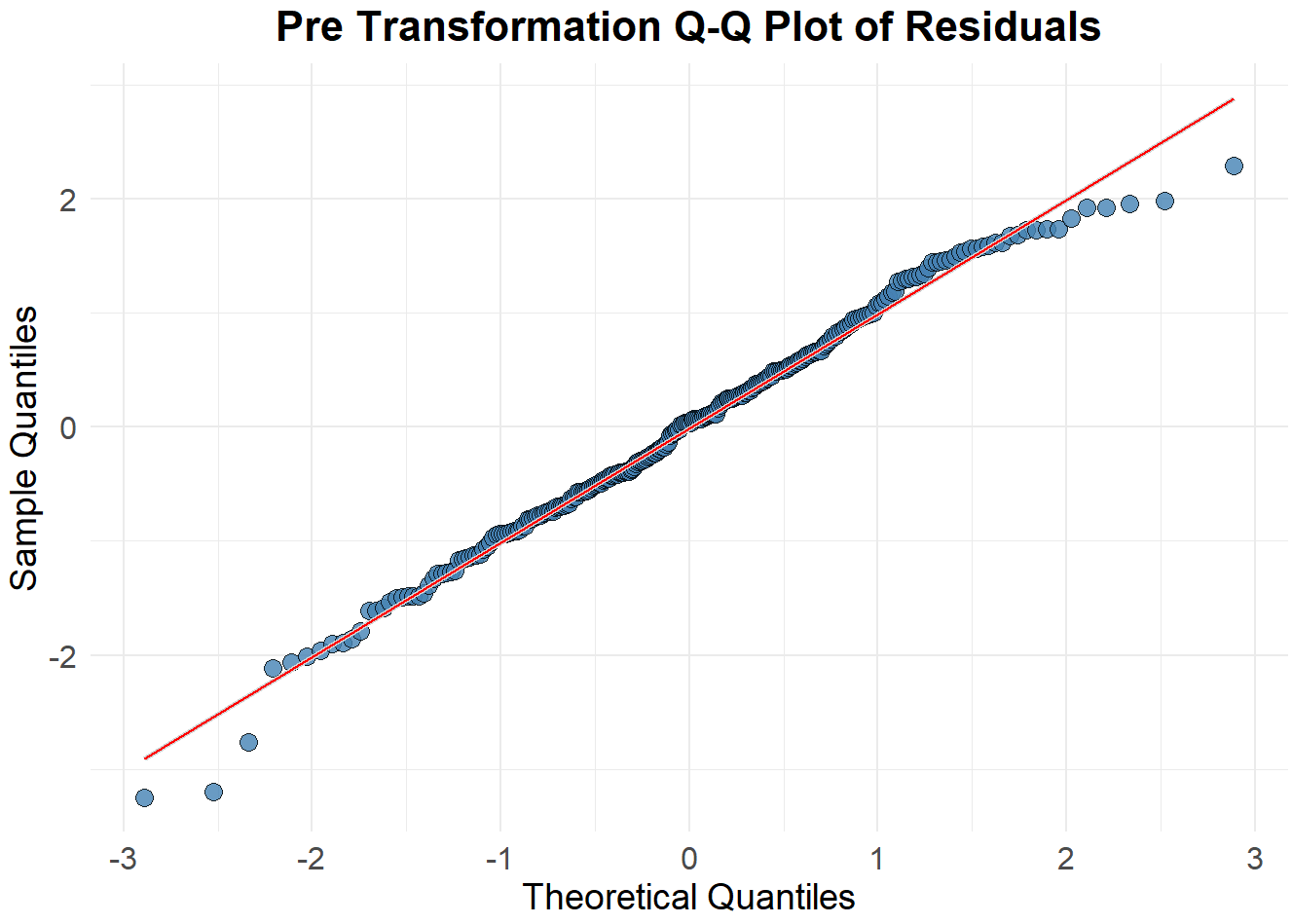
```r
# Fit model on raw data, make predictions, and save residuals
model = lm(acc_rate ~ sat, data = college)
college$pred = predict(model)
college$residuals = rstudent(model)

# Residual plot
ggplot(college, aes(x = pred, y = residuals)) +
  geom_point(shape = 21, color = "black", fill = "steelblue", size = 4, stroke = 0.5,alpha = 0.
8) +  # Points with outline
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  # Horizontal line at y = 0
  labs(title = "Pre Transformation Residual Plot",
       x = "Predicted Acceptance Rate",
       y = "Jackknife Residual") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
        axis.title = element_text(size = 14),
        axis.text = element_text(size = 12))
```



```
# Normal probability plot
qq_plot <- ggplot(data.frame(residuals = college$residuals), aes(sample = residuals)) +
  stat_qq(shape = 21, color = "black", fill = "steelblue", size = 3, stroke = 0.5, alpha = 0.8)
+  # Points with outline and alpha
  stat_qq_line(color = "lightgray", size = 1) +  # Light gray reference line
  stat_qq_line(color = "red", size = 0.5) +  # Black outline for the line
  labs(title = "Pre Transformation Q-Q Plot of Residuals",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
        axis.title = element_text(size = 14),
        axis.text = element_text(size = 12))
```

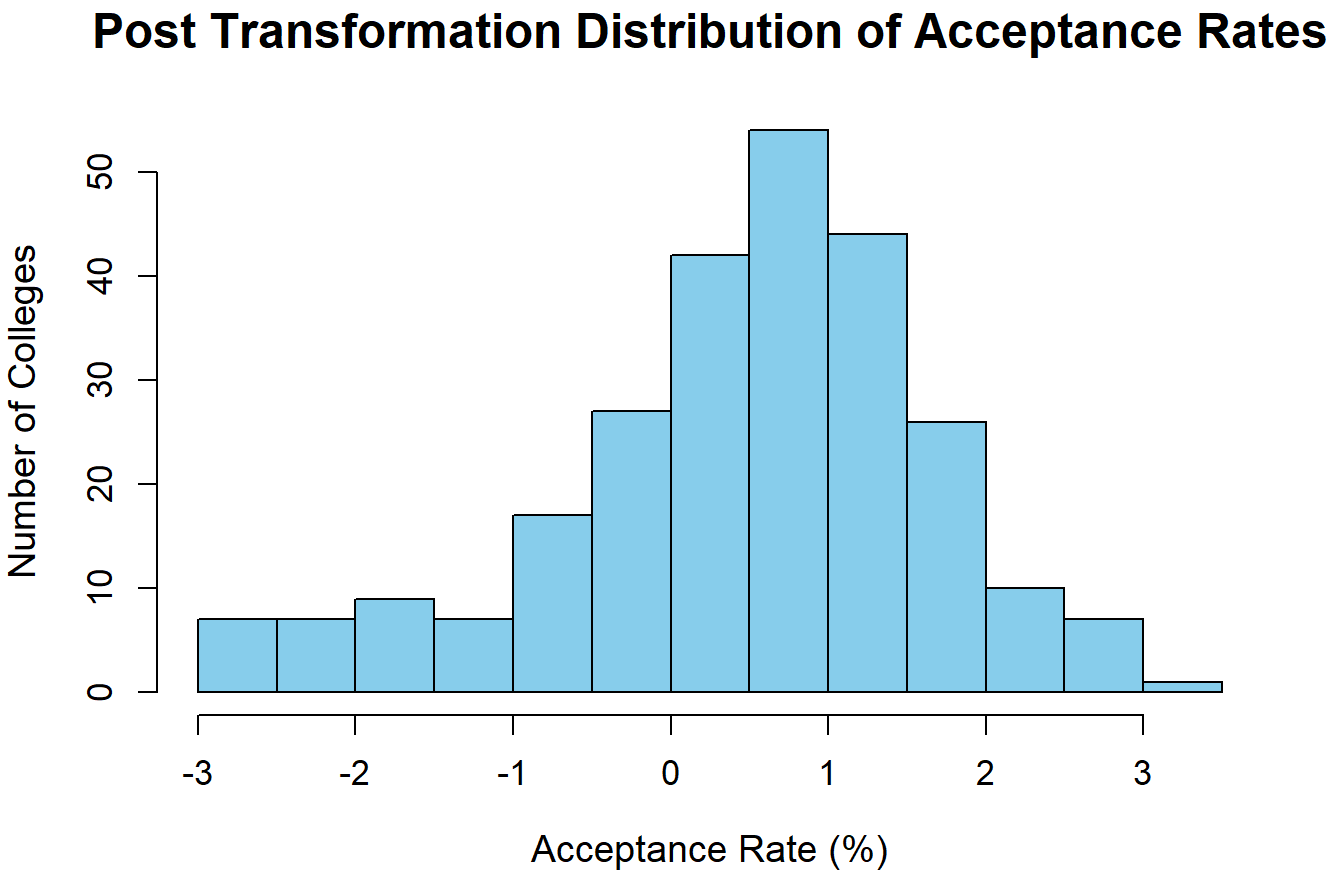
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
qq_plot
```

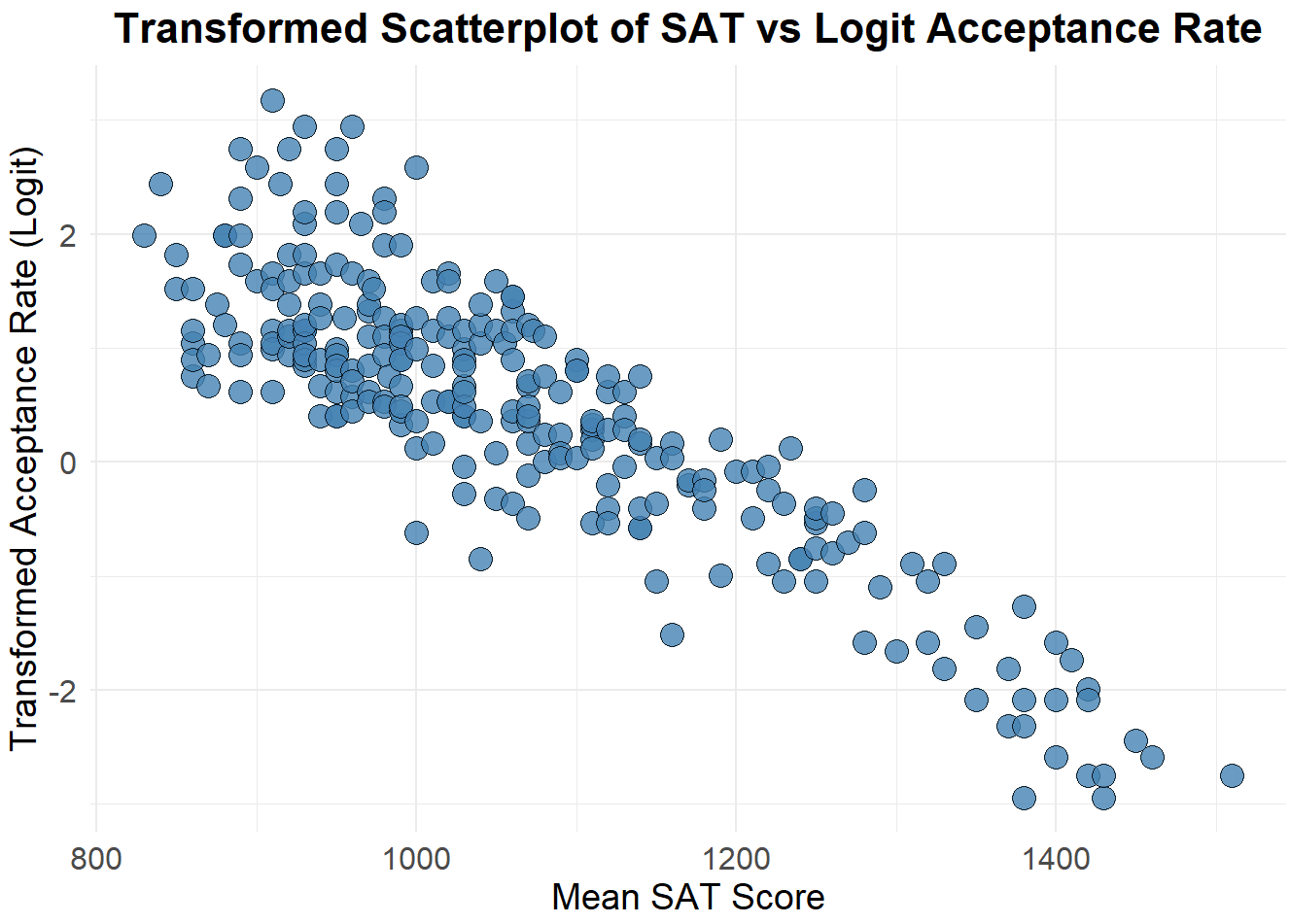


Create plots and fit model on transformed data that uses SAT mean to predict acceptance rate

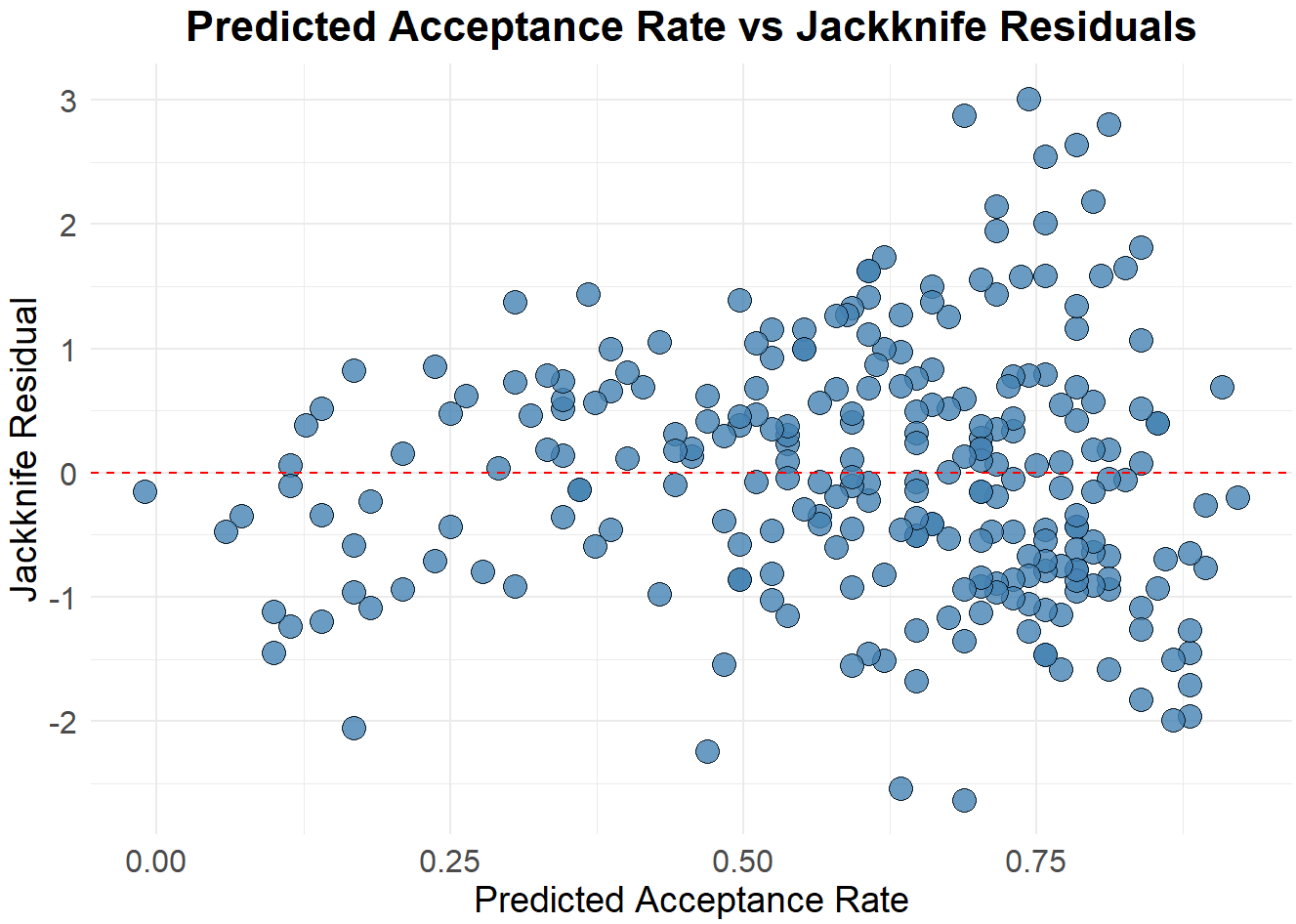
```r
# Perform logit transformation on acceptance rate
college$logit_rate = log(college$acc_rate/(1-college$acc_rate))

# Histogram of raw responses
hist(college$logit_rate,
     col = "skyblue",  # Change bar color to a light blue
     border = "black",  # Bar border color
     main = "Post Transformation Distribution of Acceptance Rates",  # Title
     xlab = "Acceptance Rate (%)",  # X-axis label
     ylab = "Number of Colleges",  # Y-axis label
     breaks = 10,  # Adjust number of bins as needed
     cex.main = 1.5,  # Title size
     cex.lab = 1.2,  # Axis label size
     cex.axis = 1.1  # Axis text size
)
```



```
# Scatterplot of transformed data
ggplot(college, aes(x = sat, y = logit_rate)) +
  geom_point(shape = 21, color = "black", fill = "steelblue", size = 4, stroke = 0.5, alpha = 0.
8) +  # Thinner outline
  labs(title = "Transformed Scatterplot of SAT vs Logit Acceptance Rate",
       x = "Mean SAT Score",
       y = "Transformed Acceptance Rate (Logit)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
        axis.title = element_text(size = 14),
        axis.text = element_text(size = 12))
```



```
# Fit model on transformed data, make predictions, and save residuals
tr_model = lm(logit_rate ~ sat, data = college)
college$tr_pred = predict(tr_model)
college$tr_residuals = rstudent(tr_model)

# Residual plot
ggplot(college, aes(x = pred, y = tr_residuals)) +
  geom_point(shape = 21, color = "black", fill = "steelblue", size = 4, stroke = 0.5, alpha = 0.
8) +  # Thinner outline
  labs(title = "Predicted Acceptance Rate vs Jackknife Residuals",
       x = "Predicted Acceptance Rate",
       y = "Jackknife Residual") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +  # Horizontal line at y = 0
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
        axis.title = element_text(size = 14),
        axis.text = element_text(size = 12))
```



```
# Normal probability plot
ggplot(college, aes(sample = tr_residuals)) +
  geom_qq(shape = 19, color = "steelblue") +
  geom_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Jackknife Residuals",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
        axis.title = element_text(size = 14),
        axis.text = element_text(size = 12))
```