Stats382: Project2

Erica Castillo

2024-04-14

Load Dataset

```
college_data <- read_csv("/Users/ecasti37/Downloads/college_sample2.csv")
```

Task 1

Converting specific variables in the dataset into factors.

```
college_data$HighestDegree <- as.factor(college_data$HighestDegree)
college_data$FundingModel <- as.factor(college_data$FundingModel)
college_data$Region <- as.factor(college_data$Region)
college_data$Geography <- as.factor(college_data$Geography)
college_data$SAT_Cat <- factor(college_data$SAT_Cat, ordered = TRUE, levels = c("Lower",
"Middle", "Higher"))
```

Task 2

In our dataset, college_data, I will be testing whether the variable HighestDegree varies from the variable FundingModel. In other words, does the variable FundingModel affect the variable HighestDegree.

To see if the variables are independent of one another, I will perform a chi-square test, to test for independence at a 3% significance level.

```
## Chi-squared test results:
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(college_data$HighestDegree, college_data$FundingModel)
## X-squared = 3.0359, df = 1, p-value = 0.08144
```

- **The hypothesis being tested:**
  - Null Hypothesis (H0): HighestDegree and FundingModel are independent.
  - Alternative Hypothesis (H1): HighestDegree and FundingModel are not independent.

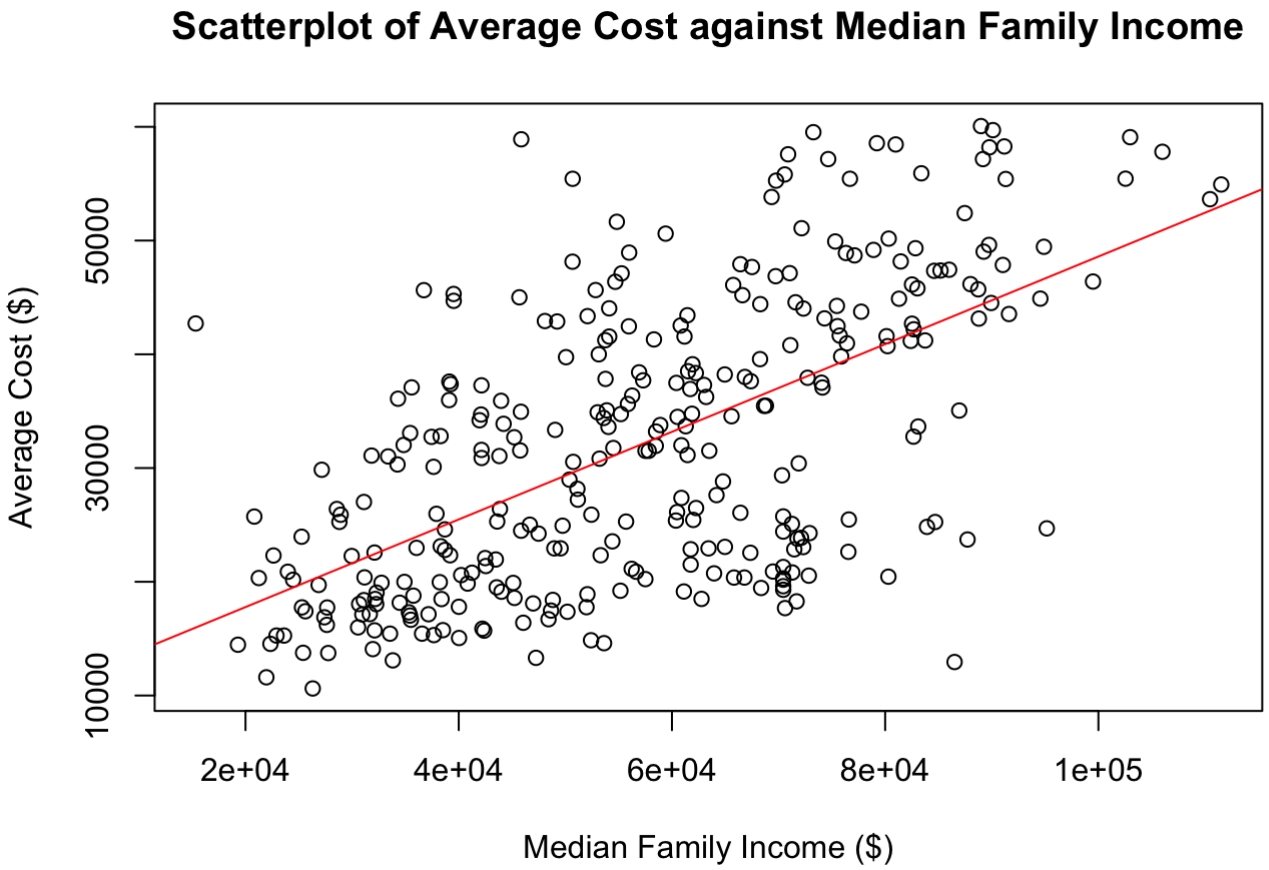
From our results of the chi-sqaured test, our p-value is 0.08144103. Since the p-value 0.08144103 is greater than the significance level of 0.03, we do not reject the null hypothesis. Thus, there is insufficient evidence to conclude that there is a significant association between HighestDegree and FundingModel.

Task 3

In this task, we will perform a simple linear regression analysis to predict AverageCost using MedianFamilyIncome as a predictor variable.

- **State the Model**
  - Y = AverageCost
  - X = MedianFamilyIncome

Scatterplot of the Data w/ Regression Line



- **Interpretation**
  - The scatterplot shows the relationship between our predictor variable (MedianFamilyIncome) and our response variable (AverageCost).
  - Based on the scatterplot, there seems to be no strong indication of linearity. Although, there may be linearity starting from the bottom-left corner going up to the upper-right corner.
  - Need more analysis!

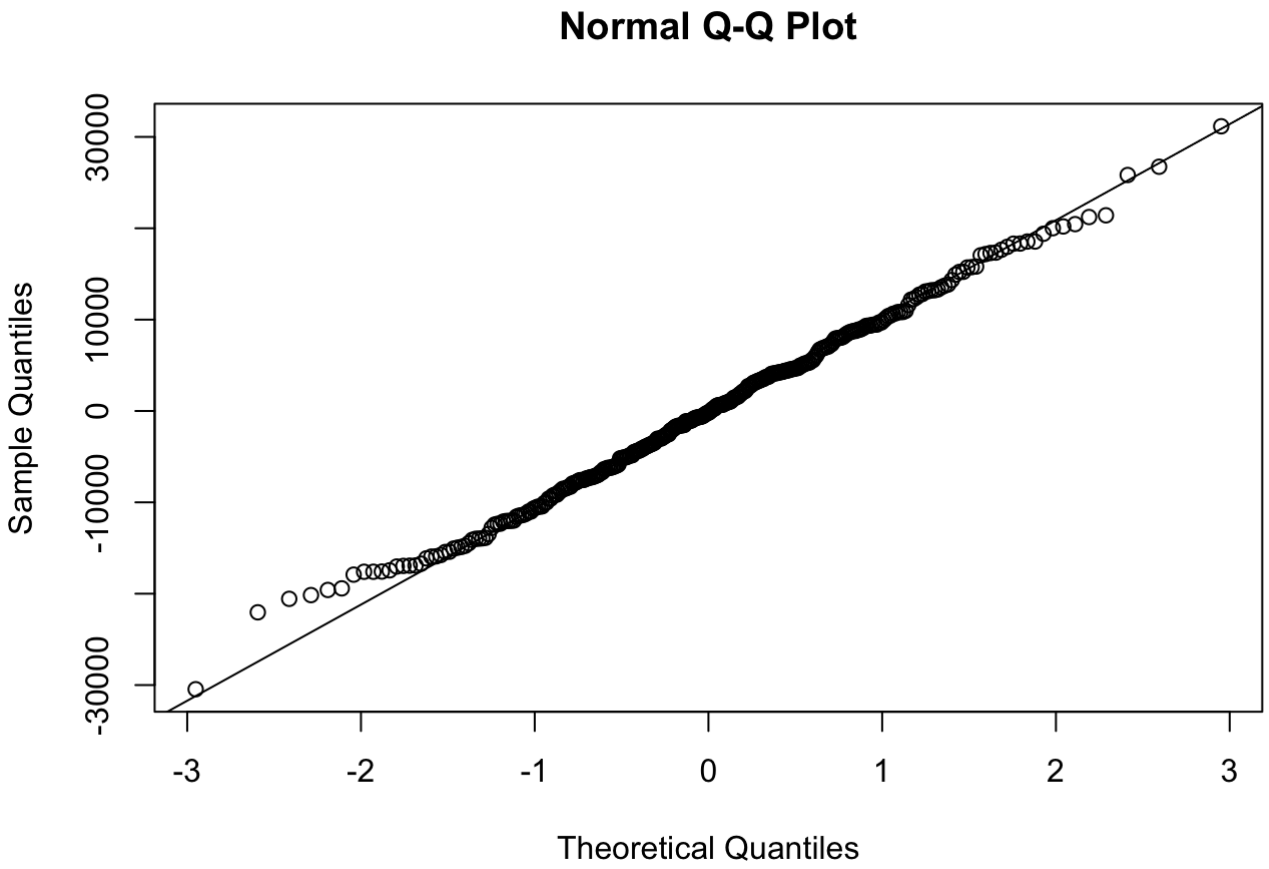
Check Assumptions for Linear Regression

Run Linear Regression Model

```
##
## Call:
## lm(formula = AverageCost ~ MedianFamilyIncome, data = college_data)
##
## Coefficients:
##        (Intercept)  MedianFamilyIncome
##          1.007e+04           3.853e-01
```

- **Equation of the regression line:**
  - Average Cost = 1.007e+04 + 3.853e-01 * Median Family Income
  - This was the red line on the scatterplot above.
  - Similar to y=mx+b
    - m = 3.853e-01 (our slope)
    - b = 1.007e+04 (our intercept)
    - x = MedianFamilyIncome
- **Interpretation:**
  - As MedianFamilyIncome increases by 1, AverageCost increases by approximately 0.39$.
  - In other words, families with higher income have a higher average cost.
  - This shows some relationship between the variables.

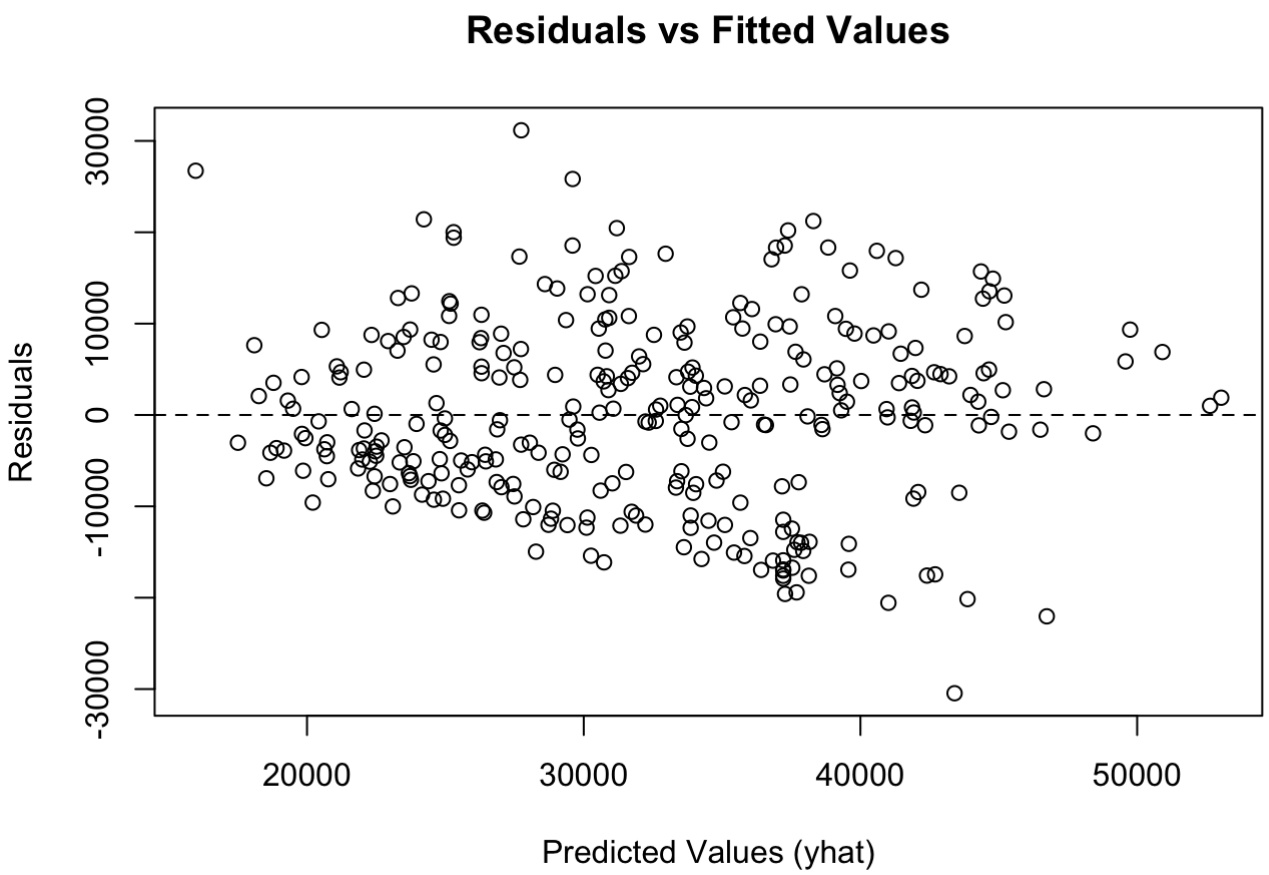
QQ Plot to check for Normality



- **Interpretation:**
  - The Q-Q plot tapers off towards the left end of the scatterplot which can suggests that the residuals are not normally distributed.

- Although, the main portion stay on the line which can still indicate it is normally distributed.

Residuals vs Fitted Values for Equal Variance



- **Interpretation:**
  - In the Residuals vs Fitted Values plot, we observe no distinct pattern, indicating equal variance of errors.

R-squared Value

```
## [1] "R-squared value below:"
```

```
## [1] 0.3686619
```

- **Interpretation:**
  - Our R-squared value is 0.3686619. This means the model explains about 36% of the variability in Average Cost using Median Family Income as a predictor variable.

Pearson Correlation Coefficient

```
## [1] "Correlation coefficient value below:"
```

```
## [1] 0.6071753
```

- **Interpretation:**

- Our correlation coefficient (r value) is 0.6071753.
- This means there's a moderate positive relationship: when Median Family Income goes up, Average Cost will go up too.
- In simpler terms, it means that when families make more money, the cost of college goes up too.

Hypothesis Test for Linear Relationship Between the Variables

To determine if there is a significant linear relationship between Median Family Income and Average Cost, we'll perform a hypothesis test at a 5% significance level.

```
##
## Call:
## lm(formula = AverageCost ~ MedianFamilyIncome, data = college_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30456.0  -7244.4   -182.7   6949.6  31162.5
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.007e+04  1.726e+03   5.836 1.33e-08 ***
## MedianFamilyIncome 3.853e-01  2.845e-02  13.541  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10100 on 314 degrees of freedom
## Multiple R-squared:  0.3687, Adjusted R-squared:  0.3667
## F-statistic: 183.4 on 1 and 314 DF,  p-value: < 2.2e-16
```

- **Hypotheses & Results:**:
  - Null Hypothesis (H0): There is no linear relationship between Median Family Income and Average Cost.
  - Alternative Hypothesis (H1): There is a linear relationship between Median Family Income and Average Cost.
  - p-value: 2.2e-16
  - Decision: Since the p-value is smaller than our significance level, we reject the null hypothesis.
  - Conclusion: There is a significant linear relationship between MedianFamilyIncome and AverageCost.

Task 4

In this task, we will create a multi-linear regression model (with no interactions) to predict the variable AdmissionRate using variables: ACTMedian, MedianDebt, AverageCost, and AverageFacultySalary.

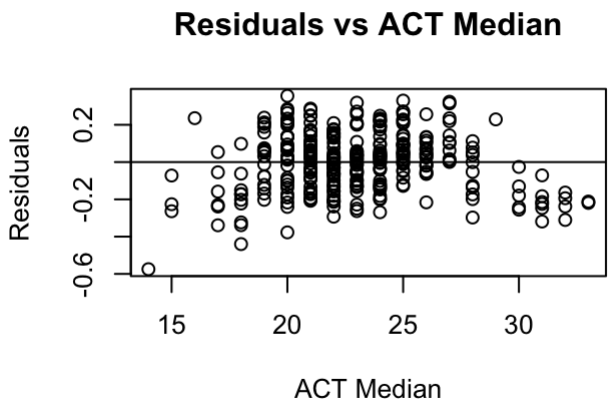
- **State the Model**
  - Y = AdmissionRate
  - X1 = ACTMedian
  - X2 = MedianDebt
  - X3 = AverageCost
  - X4 = AverageFacultySalary

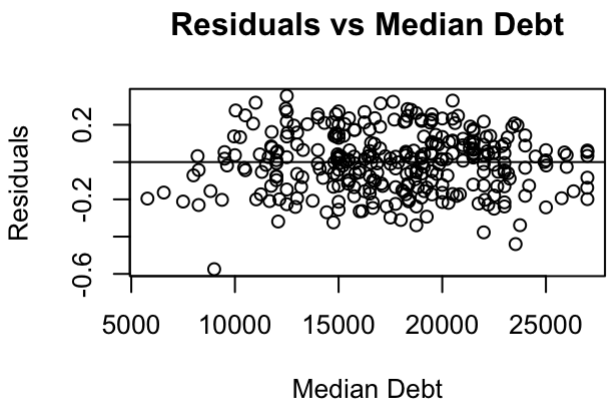
Run the Multi-Linear Regression Model w/o Interactions

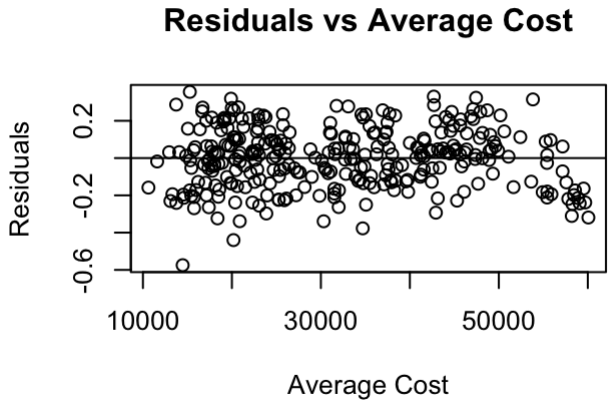
```
##
## Call:
## lm(formula = AdmissionRate ~ ACTMedian + MedianDebt + AverageCost +
##     AverageFacultySalary, data = college_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57469 -0.12311  0.01571  0.10517  0.35523
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           8.333e-01  7.113e-02  11.715  < 2e-16 ***
## ACTMedian            -5.278e-03  3.792e-03  -1.392  0.16488
## MedianDebt            8.366e-06  2.386e-06   3.507  0.00052 ***
## AverageCost          -2.749e-06  9.539e-07  -2.882  0.00423 **
## AverageFacultySalary -1.805e-05  5.471e-06  -3.299  0.00108 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1601 on 311 degrees of freedom
## Multiple R-squared:  0.1475, Adjusted R-squared:  0.1366
## F-statistic: 13.46 on 4 and 311 DF,  p-value: 3.979e-10
```

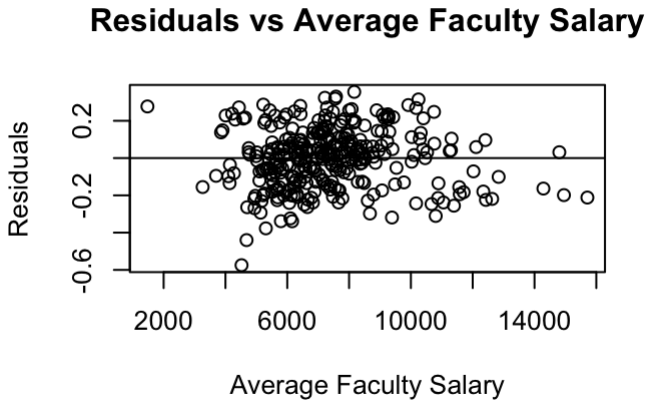
Assumptions for Multi-Linear Regression

Checking Linearity









- **Interpretation:**
  - Residuals vs ACTMedian
    - Indicator Variable
  - Residuals vs MedianDebt
    - Points are scattered, no pattern.
    - Appropriate linear model.
  - Residuals vs AverageCost
    - Points are scattered, no pattern.
    - Appropriate linear model.
  - Residuals vs AverageFacultySalary
    - Points are scattered, no pattern.
    - Approriate linear model.

Checking Independence

Lets perform a hypothesis test for errors of independence using the Durbin-Watson test with a 1% significance level.

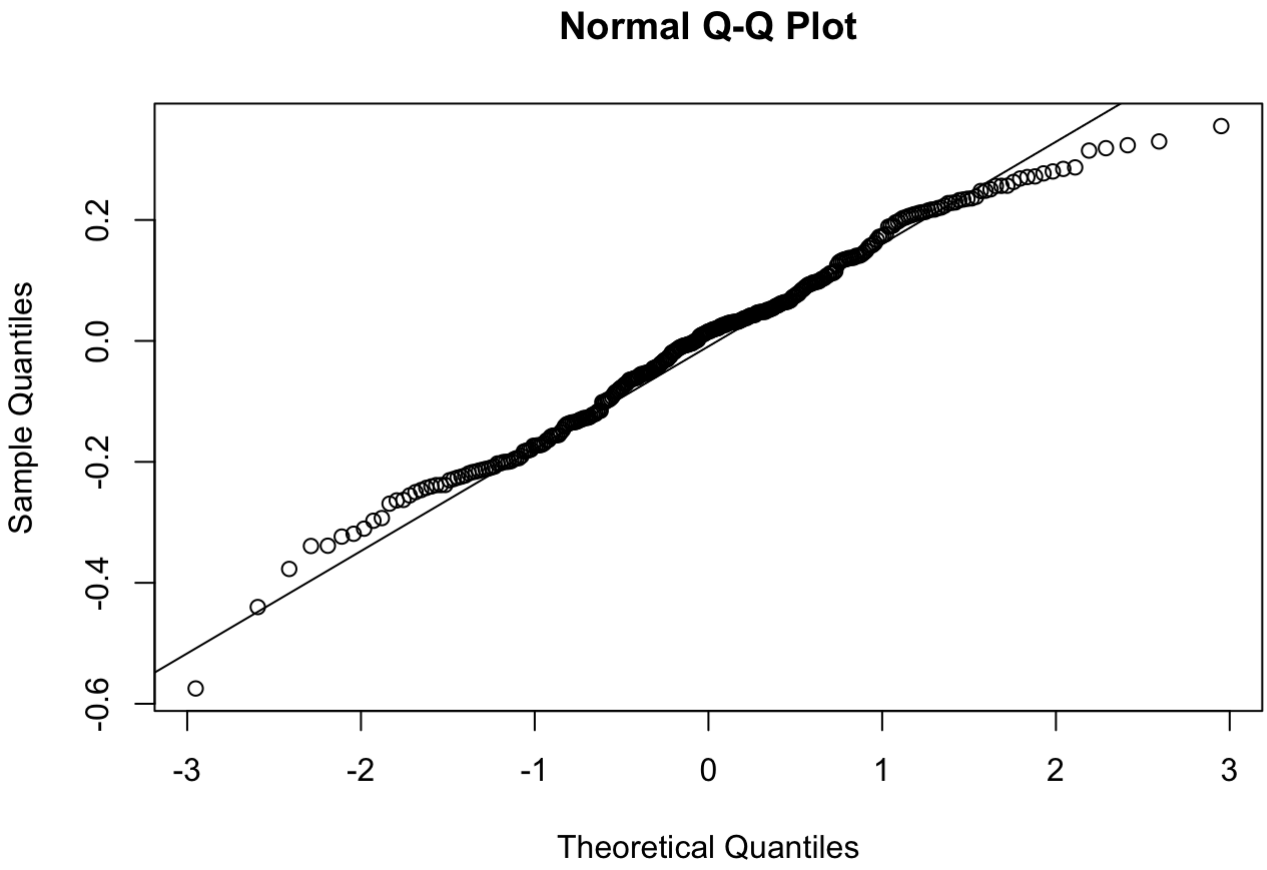
```
##
##  Durbin-Watson test
##
## data:  mlr_model
## DW = 2.0962, p-value = 0.3867
## alternative hypothesis: true autocorrelation is not 0
```

- **Hypotheses & Results:**:
  - Null Hypothesis (H0): Errors are independent
  - Alternative Hypothesis (H1): Errors are not independent
  - p-value: 0.3867
  - Decision: Since the p-value is larger than our significance level, we do not reject H0.
  - Conclusion: There is not enough evidence to show that the errors are not independent.

Checking Normality

Lets observe a QQ Plot to check for normality.

QQ Plot



- **Interpretation:**
  - The Q-Q plot tapers off on both ends of the scatterplot which can suggests that the residuals are not normally distributed.

For a better look on normality, lets perform a… #### Shapiro-Wilk Normality Test

Lets perform a Shapiro-Wilk test for normality at a significance of 5%.

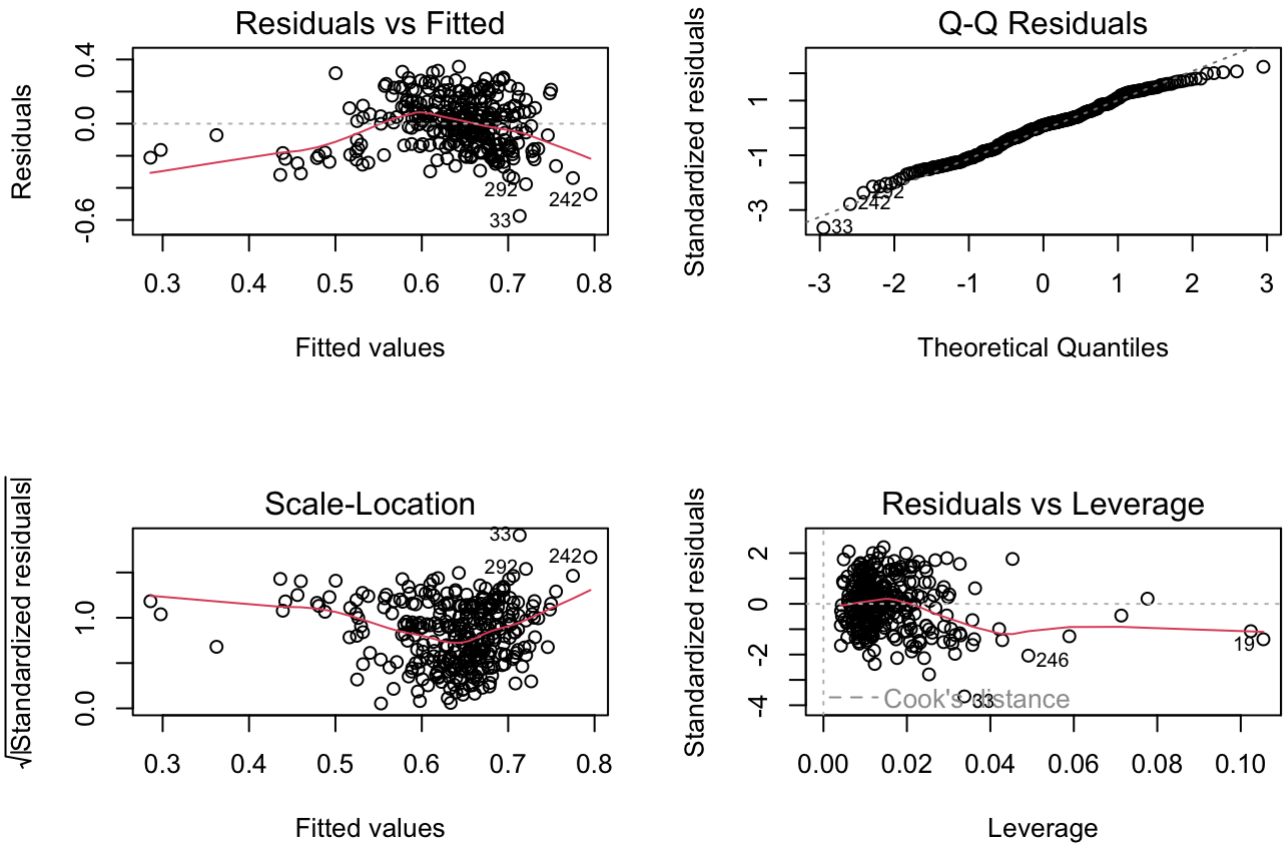
```
##
##  Shapiro-Wilk normality test
##
## data:  mlr_model$residuals
## W = 0.98944, p-value = 0.02189
```

- **Hypotheses & Results:**:

- Null Hypothesis (H0): Errors are normal
- Alternative Hypothesis (H1): Errors are not normal
- p-value: 0.02189
- Decision: Since the p-value is smaller than our significance level, we reject H0.
- Conclusion: There is enough evidence to show that the errors are not normal. Our assumption is not met.

Checking Equal Variance

To check for equal variance, let plots our multi-linear regression model



- **Interpretation:**
  - Based on our four given models:
    - There is no linearity aside from the QQ Residuals plot
    - Equal variance is met in plots:
      - Residuals vs Fitted
      - Scale-Location
      - Residuals vs Leverage

Hypothesis Test for Overall

To see if independent variables explain variation in dependent variables, I will perform a hypothesis test with a 5% significance level.

```
## Analysis of Variance Table
##
## Response: AdmissionRate
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## ACTMedian              1 0.7688 0.76883 30.0001 8.925e-08 ***
## MedianDebt             1 0.1313 0.13132  5.1241  0.024284 *
## AverageCost            1 0.2002 0.20025  7.8137  0.005508 **
## AverageFacultySalary   1 0.2789 0.27894 10.8844  0.001082 **
## Residuals            311 7.9702 0.02563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **ACTMedian:**
  - Null Hypothesis (H0): The coefficient for ACTMedian is zero (no effect on the dependent variable).
  - Alternative Hypothesis (H1): The coefficient for ACTMedian is not zero (some effect on the dependent variable).
  - F-statistic: 30.0001
  - p-value: 8.925e-08
  - Decision: Since the p-value is much smaller than our significance level, we reject the null hypothesis.
  - Conclusion: There is enough evidence to conclude that ACTMedian has a significant effect on the dependent variable.
- **MedianDebt:**
  - Null Hypothesis (H0): The coefficient for MedianDebt is zero (no effect on the dependent variable).
  - Alternative Hypothesis (H1): The coefficient for MedianDebt is not zero (some effect on the dependent variable).
  - F-statistic: 5.1241
  - p-value: 0.024284
  - Decision: Since the p-value is smaller than our significance level (usually 0.05), we reject the null hypothesis.
  - Conclusion: There is enough evidence to conclude that MedianDebt has a significant effect on the dependent variable.
- **AverageCost:**
  - Null Hypothesis (H0): The coefficient for AverageCost is zero (no effect on the dependent variable).
  - Alternative Hypothesis (H1): The coefficient for AverageCost is not zero (some effect on the dependent variable).
  - F-statistic: 7.8137
  - p-value: 0.005508
  - Decision: Since the p-value is smaller than our significance level (usually 0.05), we reject the null hypothesis.
  - Conclusion: There is enough evidence to conclude that AverageCost has a significant effect on the dependent variable.
- **AverageFacultySalary:**
  - Null Hypothesis (H0): The coefficient for AverageFacultySalary is zero (no effect on the dependent variable).
  - Alternative Hypothesis (H1): The coefficient for AverageFacultySalary is not zero (some effect on the dependent variable).
  - F-statistic: 10.8844
  - p-value: 0.001082
  - Decision: Since the p-value is much smaller than our significance level, we reject the null hypothesis.

- Conclusion: There is enough evidence to conclude that AverageFacultySalary has a significant effect on the dependent variable.

R-squared and Adjusted R-squared

```
## [1] "R-squared result:"
```

```
## [1] 0.1475301
```

```
## [1] "Adjusted r-squared result:"
```

```
## [1] 0.1365659
```

- **Interpretation:**
  - Based on our results, the regression line in our mlr_model explains about 15% of the pattern we see in the data (based on our r-squared value).
  - For our adjusted r-square value, it suggests that our model explains roughly 14% of the pattern we see in the data.

Hypothesis Partial t-tests for Individual Variable Significance

```
##
## Call:
## lm(formula = AdmissionRate ~ ACTMedian + MedianDebt + AverageCost +
##     AverageFacultySalary, data = college_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57469 -0.12311  0.01571  0.10517  0.35523
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            8.333e-01  7.113e-02  11.715  < 2e-16 ***
## ACTMedian             -5.278e-03  3.792e-03  -1.392  0.16488
## MedianDebt             8.366e-06  2.386e-06   3.507  0.00052 ***
## AverageCost           -2.749e-06  9.539e-07  -2.882  0.00423 **
## AverageFacultySalary -1.805e-05  5.471e-06  -3.299  0.00108 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1601 on 311 degrees of freedom
## Multiple R-squared:  0.1475, Adjusted R-squared:  0.1366
## F-statistic: 13.46 on 4 and 311 DF,  p-value: 3.979e-10
```

Partial T-test results:

1. **ACTMedian:**
   - Hypotheses:
     - H0: The coefficient of ACTMedian is equal to 0 (the variable is not important).

        - H1: The coefficient of ACTMedian is not equal to 0 (the variable is important).
    - p-value: 0.165
    - Decision: Fail to reject the null hypothesis.
    - Conclusion: ACTMedian is not statistically significant (p > 0.05), indicating that it may not have a significant impact on AverageCost.
2. **MedianDebt:**
    - Hypotheses:
        - H0: The coefficient of MedianDebt is equal to 0 (the variable is not important).
        - H1: The coefficient of MedianDebt is not equal to 0 (the variable is important).
    - p-value: 5.202×10^{-4}
    - Decision: Reject the null hypothesis.
    - Conclusion: MedianDebt is statistically significant (p < 0.05), indicating that it has a significant impact on AverageCost.
3. **AverageCost:**
    - Hypotheses:
        - H0: The coefficient of AverageCost is equal to 0 (the variable is not important).
        - H1: The coefficient of AverageCost is not equal to 0 (the variable is important).
    - p-value: 4.227×10^{-3}
    - Decision: Reject the null hypothesis.
    - Conclusion: AverageCost is statistically significant (p < 0.05), which is unusual and may suggest some form of multicollinearity or other issue.
4. **AverageFacultySalary:**
    - Hypotheses:
        - H0: The coefficient of AverageFacultySalary is equal to 0 (the variable is not important).
        - H1: The coefficient of AverageFacultySalary is not equal to 0 (the variable is important).
    - p-value: 1.082×10^{-3}
    - Decision: Reject the null hypothesis.
    - Conclusion: AverageFacultySalary is statistically significant (p < 0.05), indicating that it has a significant impact on AverageCost.

Overall:

The variables AverageCost and AverageFacultySalary are important. Throughout the testing, both variables showed significance to the variable AdmissionRate.

Task 5

In this task, I will conduct a One-Way ANOVA test to see if the mean value of AdmissionRate varies by Region at a 3% significance level

Conducting One-Way ANOVA Test

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Region        7  0.358 0.05114   1.752 0.0967 .
## Residuals   308  8.992 0.02919
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assumptions Check

Check Normality Assumption Using Shapiro-Wilk Test w/ 5% Significance Level

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_result)
## W = 0.98308, p-value = 0.0008873
```

Shapiro-Wilk Normality Test:

- **Hypotheses:**
  - Null Hypothesis (H0): The data follows a normal distribution.
  - Alternative Hypothesis (H1): The data does not follow a normal distribution.
- **Test Result:**
  - Significance Level: 0.03
  - W statistic: 0.93737
  - p-value: 2.721e-10
- **Decision:**
  - Since the p-value (2.721e-10) is smaller than the significance level (0.03), we reject the null hypothesis.
- **Conclusion:**
  - There is enough evidence to show that the data does not follow a normal distribution.

Check Equal Variance Assumption Using Levene's Test

```
# Equal Variance Assumption
levene_test <- leveneTest(anova_result)
levene_test
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   7  0.9495 0.4685
##       308
```

Levene's Test for Variance:

- **Hypotheses:**
  - Null Hypothesis (H0): The variances are equal across all groups.
  - Alternative Hypothesis (H1): The variances are not equal across all groups.
- **Test Result:**
  - Significance Level: 0.03
  - F value: 12.211
  - p-value: 7.824e-06
- **Decision:**
  - Since the p-value (7.824e-06) is smaller than the significance level (0.03), we reject the null hypothesis.
- **Conclusion:**
  - There is enough evidence to conclude that the variances are different across all groups.

Hypothesis & Results:

- **Hypotheses:**
  - Null Hypothesis (H0): Mean Admission Rate is the same across all regions.
  - Alternative Hypothesis (H1): Mean Admission Rate varies by region.
- **ANOVA Results:**
  - Significance Level: 0.03
  - F value: 1.752
  - p-value: 0.0967
- **Decision:**
  - Since p-value (0.0967) > 0.03 (significance level), we fail to reject the null hypothesis.
- **Conclusion:**
  - There is insufficient evidence to conclude that mean Admission Rate varies by region at a 3% significance level.

Tukey Test?

- In this case, the ANOVA test for the Admission Rate by Region did not indicate significance (p-value = 0.0967), meaning there is insufficient evidence to conclude that the mean Admission Rate varies by region at a 3% significance level.
- Since the ANOVA test did not find significant differences between the groups, there's no need to test further with the Tukey test. The Tukey test is typically conducted when the ANOVA test shows significant differences between groups, allowing us to compare multiple groups and identify which groups are different from each other.

Task 6: One-Way ANOVA Test for Average Age of Entry by SAT Category

In this task, I will conduct a One-Way ANOVA test to see if the mean value of AverageAgeofEntry varies by SAT_Cat at a 7% significance level.

1. Conducting One-Way ANOVA Test

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## SAT_Cat       2  178.4   89.19    24.7 1.1e-10 ***
## Residuals   313 1130.2    3.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Assumptions Check

Check Normality Assumption Using Shapiro-Wilk Test w/ 5% Significance Level

```
# Normality Assumption
shapiro_test2 <- shapiro.test(residuals(anova_result2))
shapiro_test2
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_result2)
## W = 0.93737, p-value = 2.721e-10
```

Shapiro-Wilk Normality Test:

- **Hypotheses:**
  - Null Hypothesis (H0): The data follows a normal distribution.
  - Alternative Hypothesis (H1): The data does not follow a normal distribution.
- **Test Result:**
  - Significance Level: 0.05
  - W statistic: 0.93737
  - p-value: 2.721e-10
- **Decision:**
  - Since the p-value (2.721e-10) is smaller than 0.05, we reject the null hypothesis.
- **Conclusion:**
  - There is enough evidence to show that the data is not normal.

Check Equal Variance Assumption Using Levene's Test

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  12.211 7.824e-06 ***
##       313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene's Test for Variance:

- **Hypotheses:**
  - Null Hypothesis (H0): The variances are equal across all SAT categories.
  - Alternative Hypothesis (H1): The variances are not equal across all SAT categories.
- **Test Result:**
  - Significance Level: 0.07
  - F value: 12.211`
  - p-value: 7.824e-0
- **Decision:**
  - Since the p-value (7.824e-0) is smaller than 0.07, we reject the null hypothesis.
- **Conclusion:**
  - There is enough evidence to suggest that the variances might not be equal across all SAT categories.

Hypotheses Testing & Conclusion:

- **Hypotheses Tested:**
  - Null Hypothesis (H0): Mean Average Age of Entry is the same across all SAT categories.
  - Alternative Hypothesis (H1): Mean Average Age of Entry varies by SAT category.
- **Results:**
  - Significance Level: 0.07

- p-value: 1.1e-10
- Decision: Since the p-value (1.1e-10) is much smaller than the significance level, we reject the null hypothesis.
- Conclusion: There is strong evidence to suggest that the mean Average Age of Entry varies by SAT category.

4. Tukey Test?

Since a significance effect was observed, I will conduct a Tukey Test.

```
# Perform Tukey's HSD Test
tukey_result <- TukeyHSD(anova_result2)
tukey_result
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = AverageAgeofEntry ~ SAT_Cat, data = college_data)
##
## $SAT_Cat
##                      diff        lwr        upr     p adj
## Middle-Lower  -0.2714339 -0.8695096  0.3266419 0.5341895
## Higher-Lower  -2.0645743 -2.8192610 -1.3098877 0.0000000
## Higher-Middle -1.7931405 -2.4600059 -1.1262751 0.0000000
```

- **Interpretation:**
  - Middle vs. Lower: There's no big difference in the ages students with middle SAT scores and those with lower scores start college.
  - Higher vs. Lower: Students with higher SAT scores usually start college much younger than those with lower scores.
  - Higher vs. Middle: Students with higher SAT scores also start college earlier than those with middle scores.
  - Overall, students with higher SAT scores tend to start college earlier, while middle and lower scorers start at later ages.

Task 7: Two-Way ANOVA Test with Interactions for Average Cost

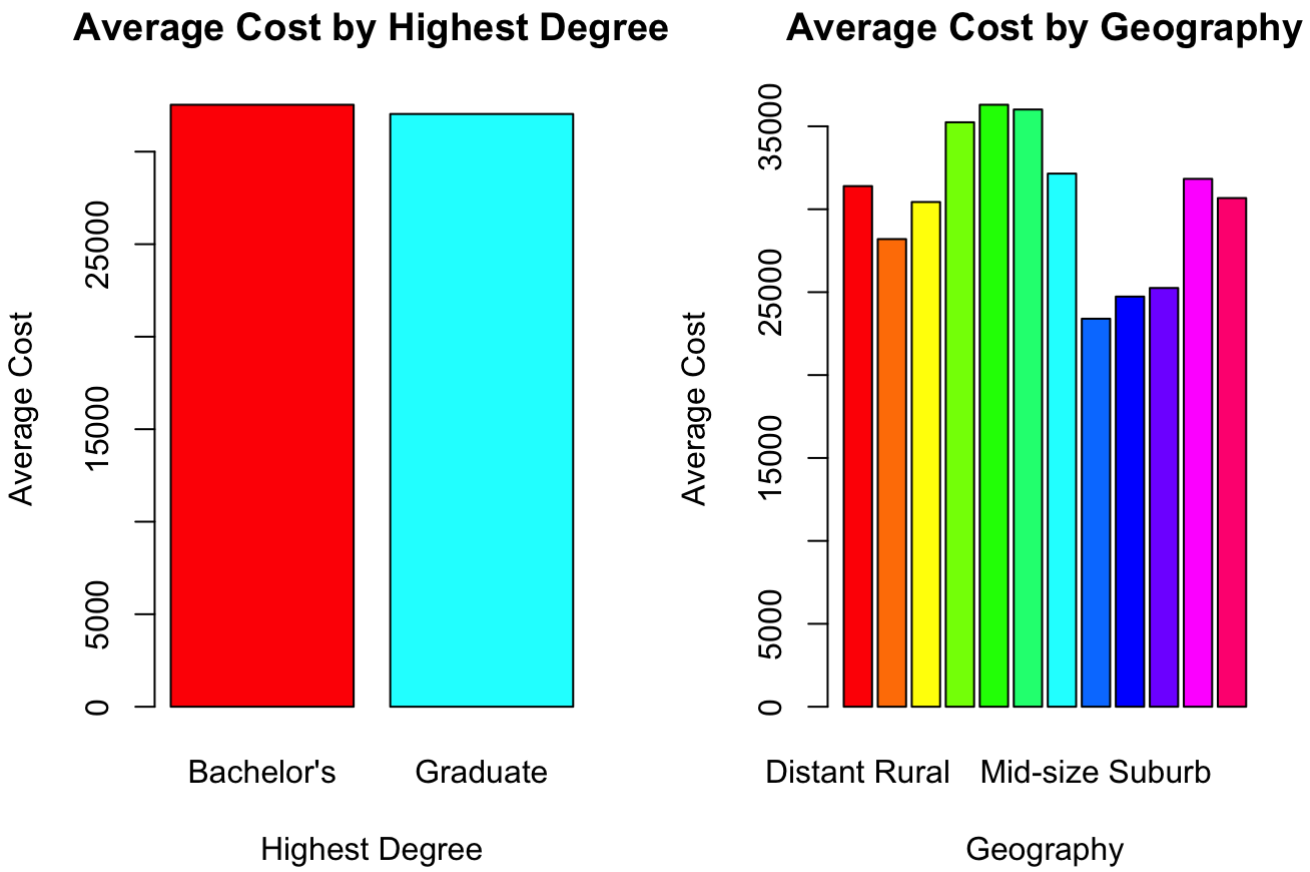
In this task, I will conduct a Two-Way ANOVA test with interactions to test the effects of HighestDegree and Geography on the variable AverageCost.

1. Conduct Two-Way ANOVA Test

```
##                           Df    Sum Sq   Mean Sq F value   Pr(>F)
## HighestDegree              1 1.222e+07  12218044   0.081 0.775514
## Geography                 11 5.363e+09 487568203   3.251 0.000329 ***
## HighestDegree:Geography   10 1.450e+09 144994872   0.967 0.472493
## Residuals                293 4.394e+10 149966629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Assumptions Check

Bar Chart Of Average Cost For Each Predictor




```
## [1] "Average Cost by Highest Degree:"
```

```
## Bachelor's   Graduate
##   32534.79   32036.58
```

```
## [1] "Average Cost by Geography:"
```

```
##   Distant Rural    Distant Town    Fringe Rural     Fringe Town      Large City
##        31394.71        28199.59        30435.00        35242.77        36299.20
##    Large Suburb   Mid-size City Mid-size Suburb    Remote Rural     Remote Town
##        36013.07        32147.97        23398.33        24733.25        25252.21
##      Small City    Small Suburb
##        31831.70        30673.73
```

- **Interpretation:**
  - The bar chart illustrates the average cost for each category of predictors (Highest Degree and Geography).
  - The bar chart helps us understand cost trends.
  - Overall, there does not seem to be a pattern in both charts.
  - More analysis is required!

Check Normality Assumption Using Shapiro-Wilk Test w/ 5% Significance Level

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_result3)
## W = 0.97764, p-value = 7.772e-05
```

Shapiro-Wilk Normality Test:

- **Hypotheses:**
  - Null Hypothesis (H0): The data follows a normal distribution.
  - Alternative Hypothesis (H1): The data does not follow a normal distribution.
- **Test Result:**
  - Significance Level: 0.04
  - W statistic: 0.97764
  - p-value: 7.772e-05
- **Decision:**
  - Since the p-value (7.772e-05) is smaller than 0.04, we reject the null hypothesis.
- **Conclusion:**
  - There is enough evidence to show that the data is not normal.

Check Equal Variance Assumption Using Levene's Test

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  22  1.6777 0.0309 *
##       293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene's Test for Variance:

- **Hypotheses:**
  - Null Hypothesis (H0): The variances are equal across all SAT categories.
  - Alternative Hypothesis (H1): The variances are not equal across all SAT categories.
- **Test Result:**
  - Significance Level: 0.04
  - F value: 1.6777
  - p-value: 0.0309
- **Decision:**
  - Since the p-value (0.0309) is smaller than 0.07, we reject the null hypothesis.
- **Conclusion:**
  - There is enough evidence to suggest that the variances might not be equal across all SAT categories.

3. Hypotheses Testing & Conclusion

- **Hypotheses Tested:**
  - Null Hypothesis (H0): No interaction effect between Highest Degree and Geography on Average Cost.
  - Alternative Hypothesis (H1): Interaction effect exists.
- **Results:**

- Significance Level: 0.04
- p-value for interaction term: 0.472493
- Decision: Since p-value > 0.04 (significance level), we do not reject the null hypothesis.
- Conclusion: There is not enough evidence to conclude that the effects of Highest Degree and Geography on Average Cost are not independent.