Chicago Public Schools (CPS)

Svetlana Voda

11/2/2024

Data Introduction and Description

In this analysis, I aim to examine the primary factors influencing college enrollment rates among students from Chicago Public Schools (CPS). Specifically, I intend to identify which characteristics of CPS schools are most strongly linked to college enrollment outcomes. Using a variety of school-related metrics, I aim to predict college enrollment numbers y=College_Enrollment based on a detailed set of predictors, including:
x1 = Average_Student_Attence
x2 = Rate_of_Misconducts
x3 = Average_Teacher_Attence
x4 = X9_Grade_Explore_2009
x5 = X11_Grade_Average_ACT_2011
x6 = College_Eligibility
x7 = Graduation_Rate
x8 = Freshman_on_Track_Rate
x9 = Probation
My goal is to analyze how these predictors correlate with college enrollment rates and to identify the most significant factors.

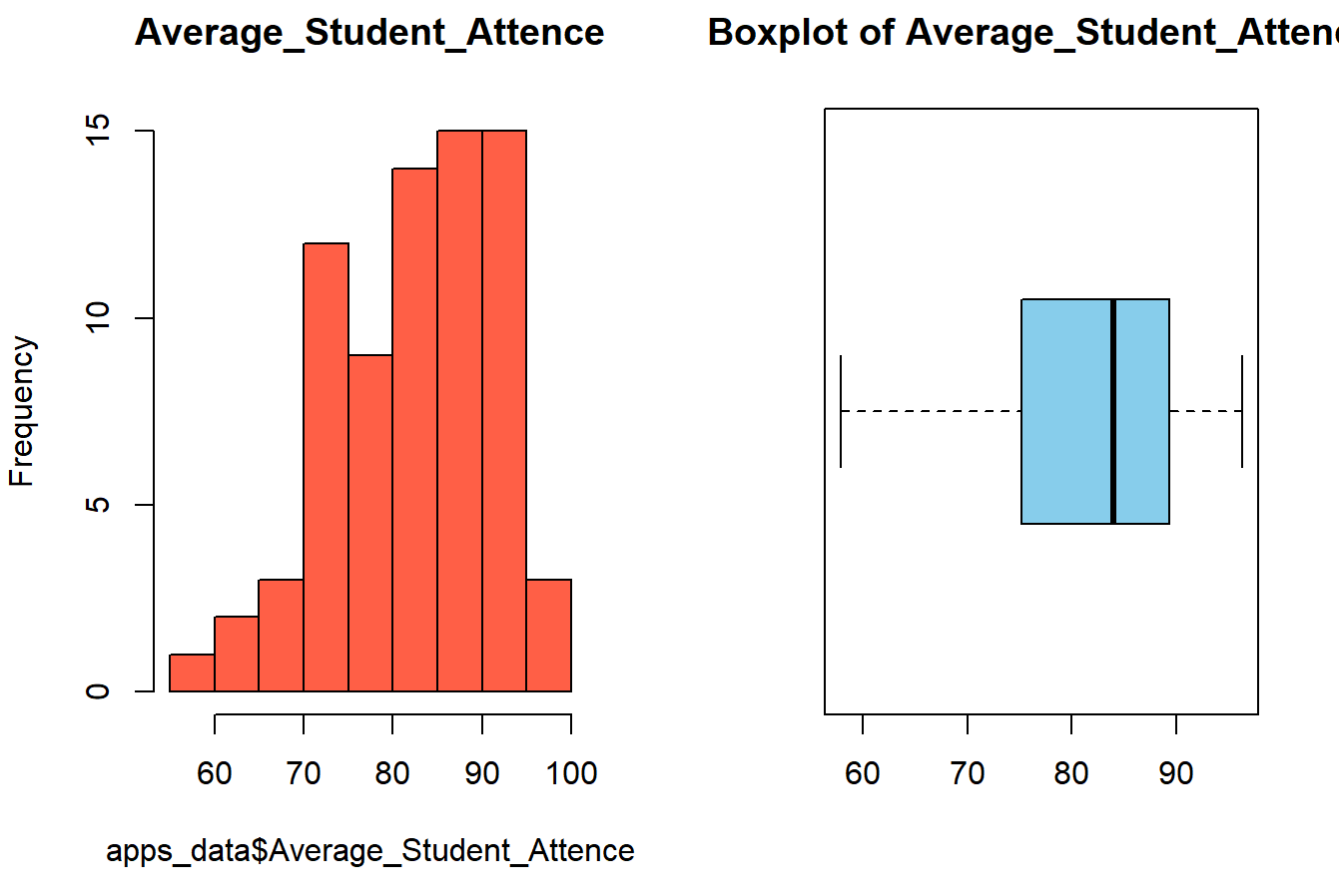
My initial step was to identify and remove any rows with missing values to ensure data completeness for the analysis. I discovered that certain columns, such as Freshman_on_Track_Rate had: 468 missing values, Graduation_Rate: 476 missing values, College_Eligibility had: 469, X11_Grade_Average_ACT_2011 had: 472 , X9_Grade_Explore_2009 had: 466 , and Average_Student_Attence had only 1 missing value, which I addressed by removing those rows.

To better understand the factors influencing college enrollment in Chicago Public Schools, I begin with a detailed analysis of each predictor variable under consideration. For each variable, I will:

- Provide descriptive Statistics – Including the mean, standard deviation, variance, and a five-number summary (minimum, Q1, median, Q3, and maximum) to understand the central tendency and spread.
- Visualize Distributions – Using histograms and boxplots to gain insights into the shape, skewness, and potential outliers in the data.
- Interpret Results – Highlighting observations about each variable’s distribution and variability to provide context for its potential influence on college enrollment.

1. To begin, I will analyze the predictor Average_Student_Attendance. This variable represents the average attendance rate of students at each CPS school and is likely to have a significant impact on college

enrollment rates.

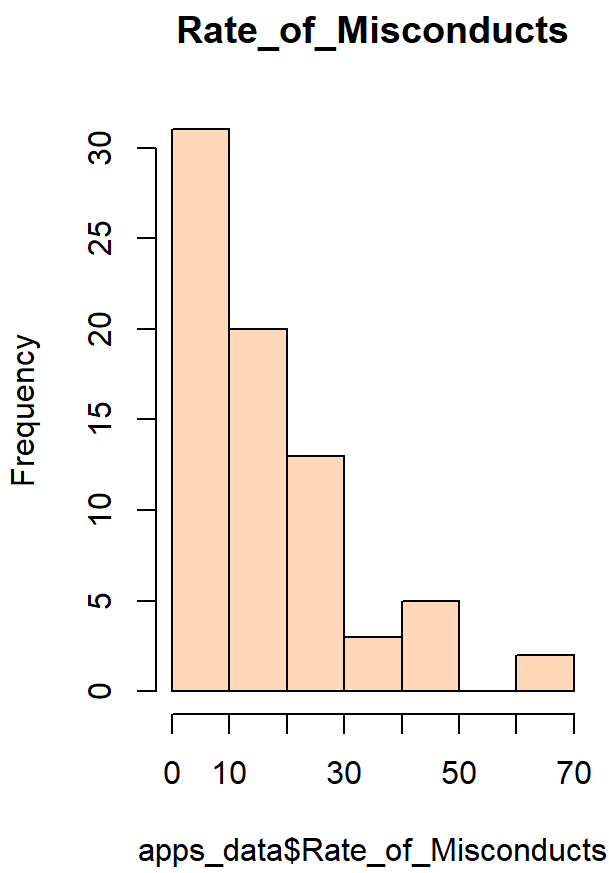


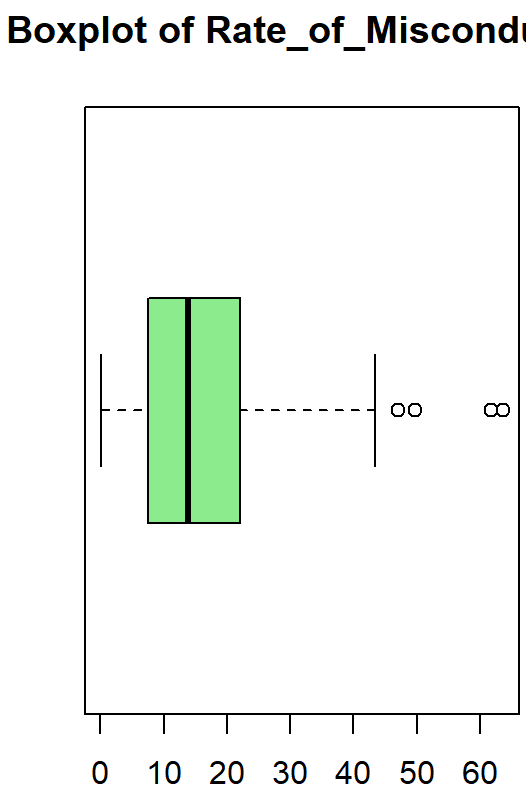
min	Q1	median	Q3	max
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
57.9	75.2	83.95	89.3	96.3

1 row

Analyzing the histogram and boxplot of the quantitative variable Average_Student_Attendance, I observe that the data is left-skewed with two peaks. From the descriptive statistics, I find that the median is 83.95, the mean is 82.5797297, the standard deviation is 8.9232493, and the variance is 79.624378. Additionally, the interquartile range (IQR) is 14.1

2. The subsequent predictor examined was Rate_of_Misconducts.



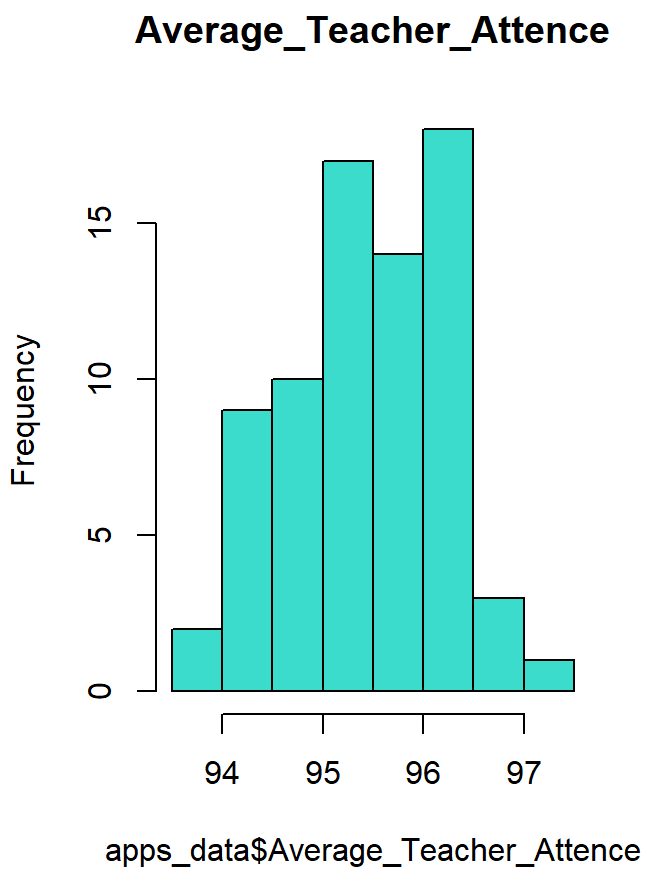


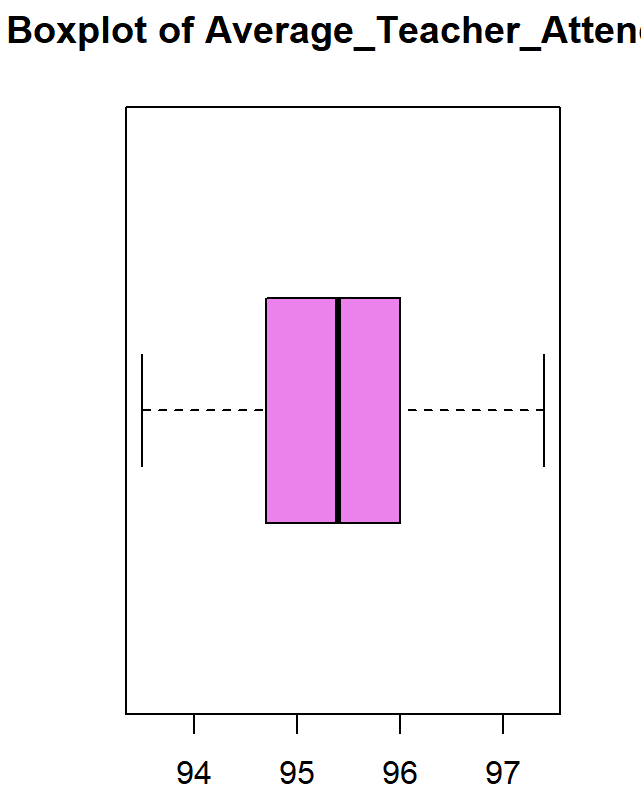
min	Q1	median	Q3	max
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0.1	7.5	13.9	22	63.6

1 row

Analyzing the histogram and boxplot of the quantitative variable Rate_of_Misconducts, we observe that the data is right-skewed with one peak and contains four outliers. From the descriptive statistics, I find that the median is 13.9, the mean is 16.877027, the standard deviation is 13.7742518, and the variance is 189.730013. Additionally, the interquartile range (IQR) is 14.5 .

3.Following the analysis of Rate_of_Misconducts, the subsequent variable considered is Average_Teacher_Attence.



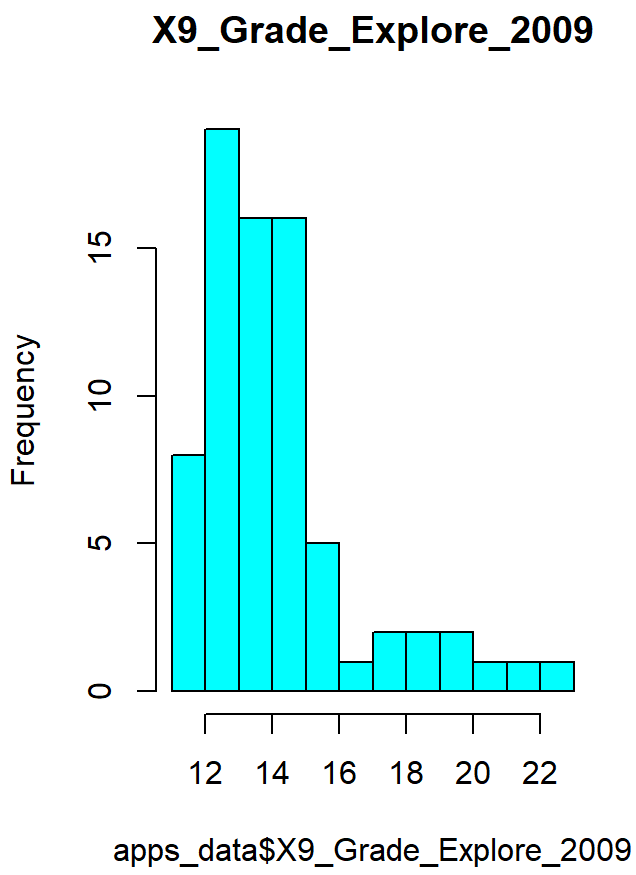


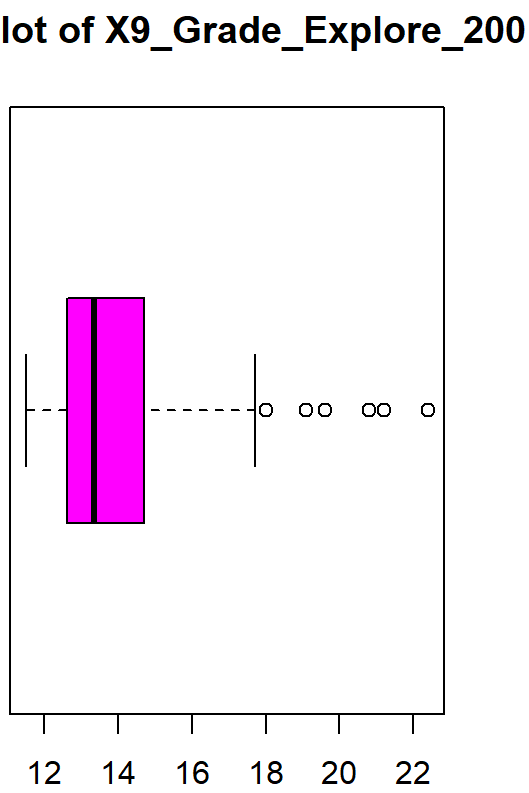
min	Q1	median	Q3	max
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
93.5	94.7	95.4	96	97.4

1 row

Analyzing the histogram and boxplot of the quantitative variable Average_Teacher_Attence, I observe that the data is left-skewed with two peaks. From the descriptive statistics, I find that the median is 95.4, the mean is 95.4121622, the standard deviation is 0.7865222, and the variance is 0.6186172. Additionally, the interquartile range (IQR) is 2.2

4. Following the analysis of Average_Teacher_Attence, the subsequent variable considered is X9_Grade_Explore_2009



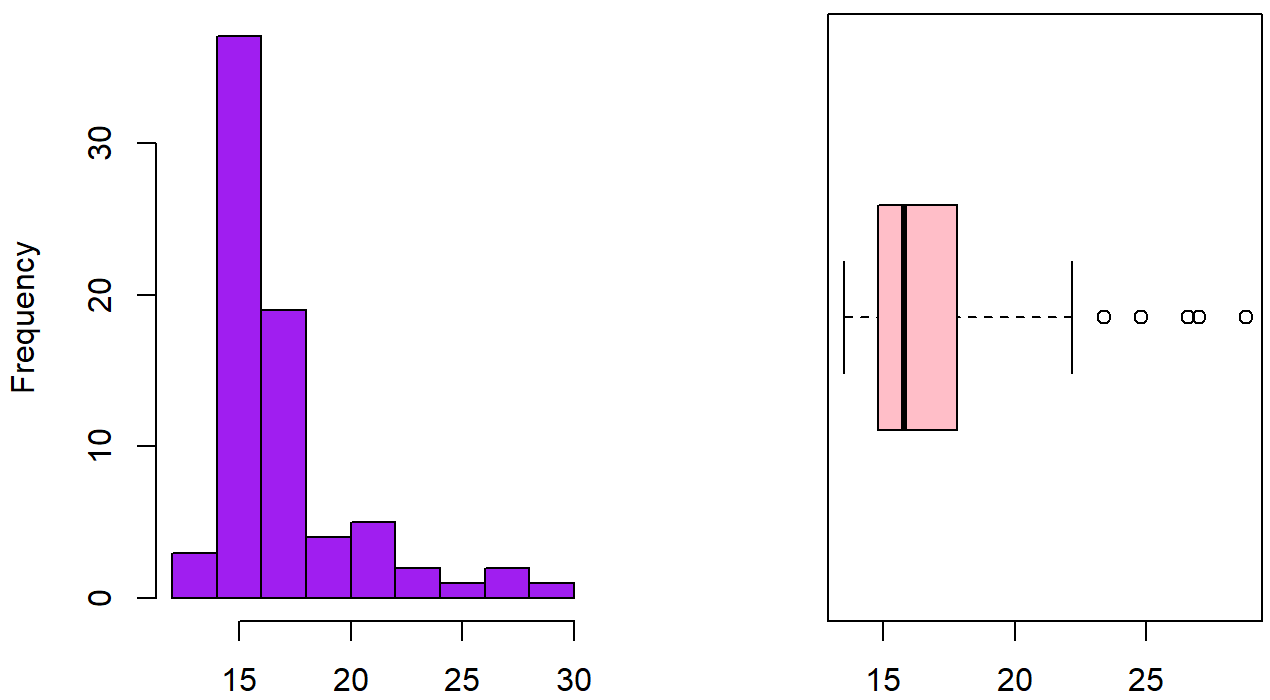


min	Q1	median	Q3	max
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
11.5	12.6	13.35	14.7	22.4

1 row

Analyzing the histogram and boxplot of the quantitative variable X9_Grade_Explore_2009, I observe that the data is right-skewed with one peak and contains six outliers. From the descriptive statistics, I find that the median is 13.35, the mean is 14.0783784, the standard deviation is 2.3515818, and the variance is 5.5299371. Additionally, the interquartile range (IQR) is 2.1

5. The subsequent predictor examined was X11_Grade_Average_ACT_2011

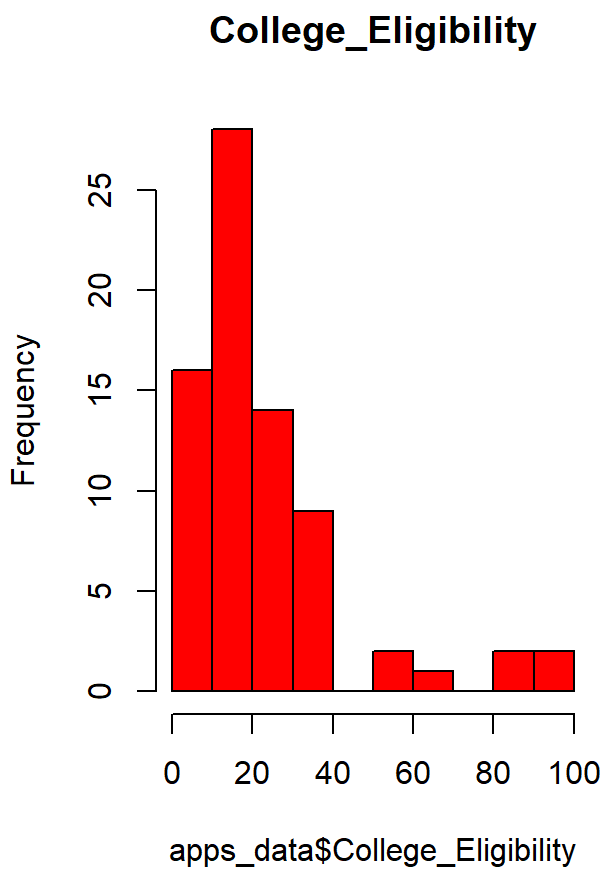


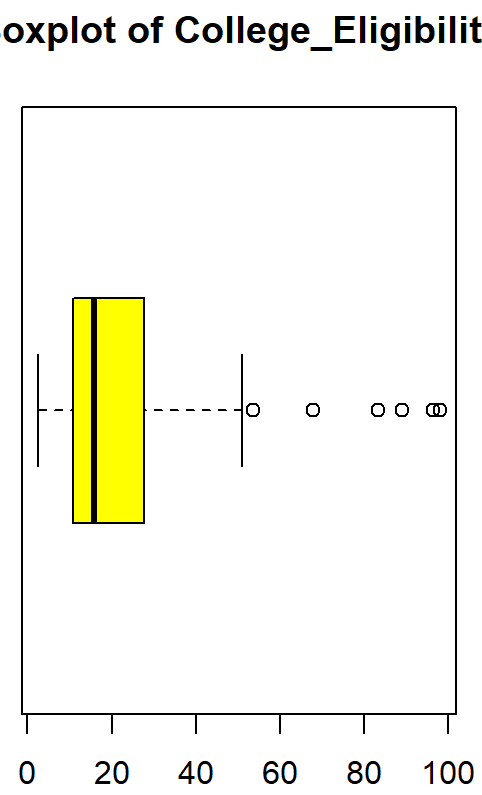
min	Q1	median	Q3	max
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
13.5	14.8	15.8	17.8	28.8

1 row

Analyzing the histogram and boxplot of the quantitative variable X11_Grade_Average_ACT_2011, I observe that the data is right-skewed with one peak and contains five outliers. From the descriptive statistics, I find that the median is 15.8, the mean is 16.8783784, the standard deviation is 3.213681, and the variance is 10.3277453. Additionally, the interquartile range (IQR) is 2.1

6. The subsequent predictor examined was College_Eligibility



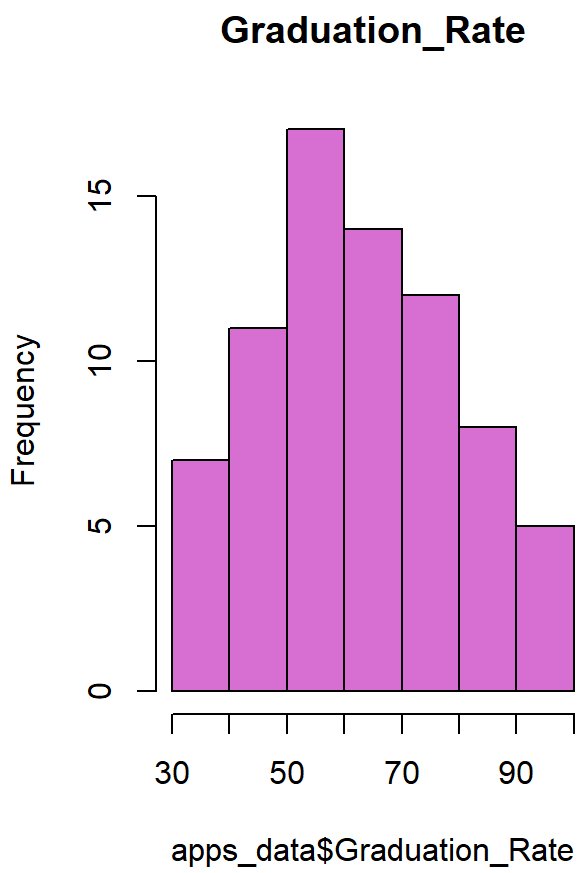


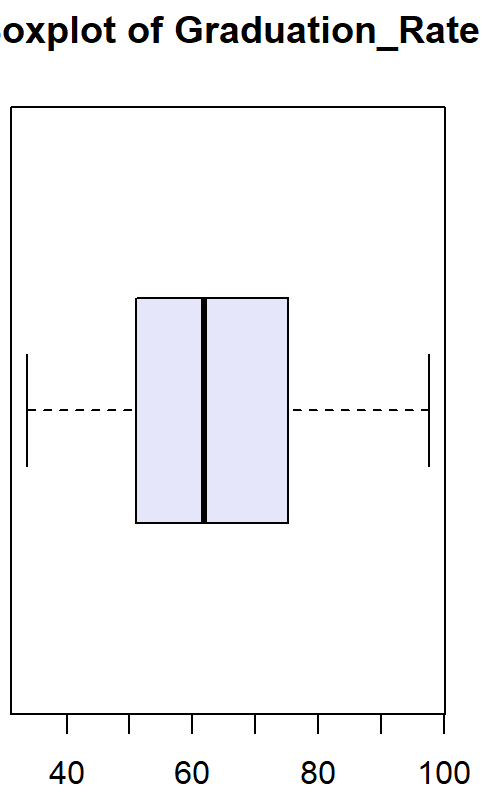
min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>
2.5	10.8	15.9	27.7	98

1 row

Analyzing the histogram and boxplot of the quantitative variable College_Eligibility, I observe that the data is right-skewed with one peak and contains six outliers. From the descriptive statistics, I find that the median is 15.9, the mean is 22.5986486, the standard deviation is 20.5691982, and the variance is 423.091916. Additionally, the interquartile range (IQR) is 16.9

7. The next predictor analyzed was Graduation_Rate



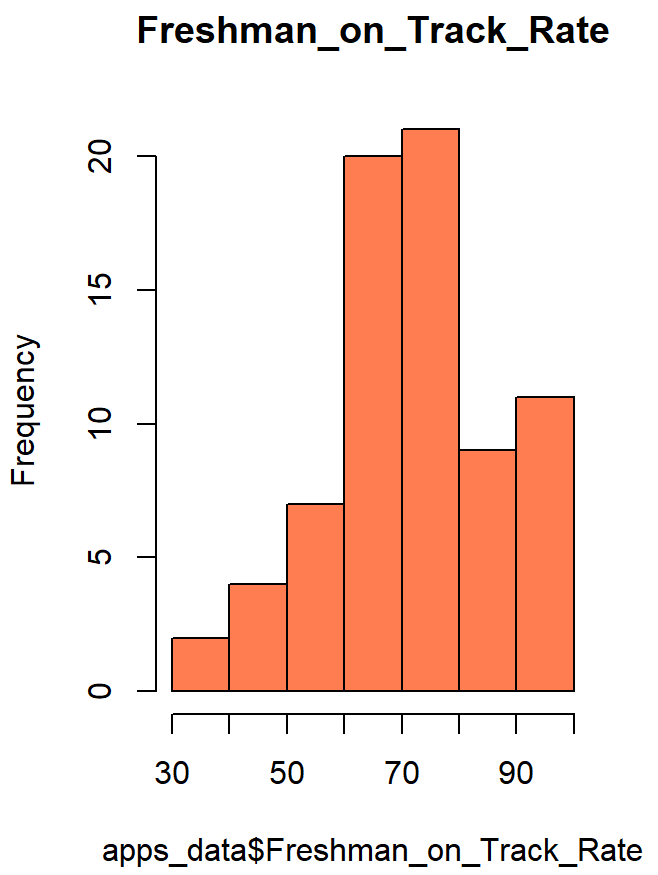


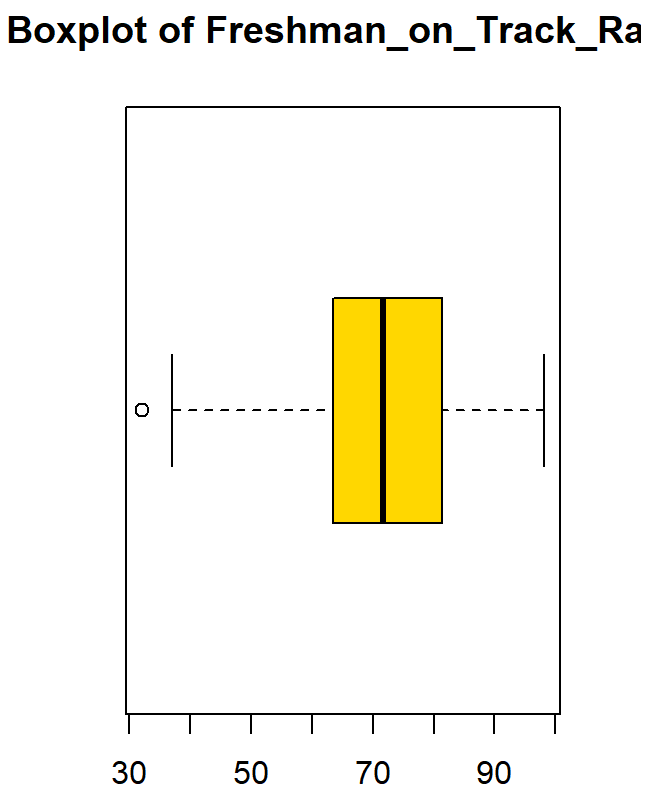
min	Q1	median	Q3	max
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
33.7	51.1	61.9	75.2	97.6

1 row

Analyzing the histogram and boxplot of the quantitative variable Graduation_Rate, I observe that the data is bell shaped with no outliers. From the descriptive statistics, I find that the median is 61.9, the mean is 62.9608108, the standard deviation is 16.7761884, and the variance is 281.440498. Additionally, the interquartile range (IQR) is 24.1.

8. The next predictor analyzed was Freshman_on_Track_Rate



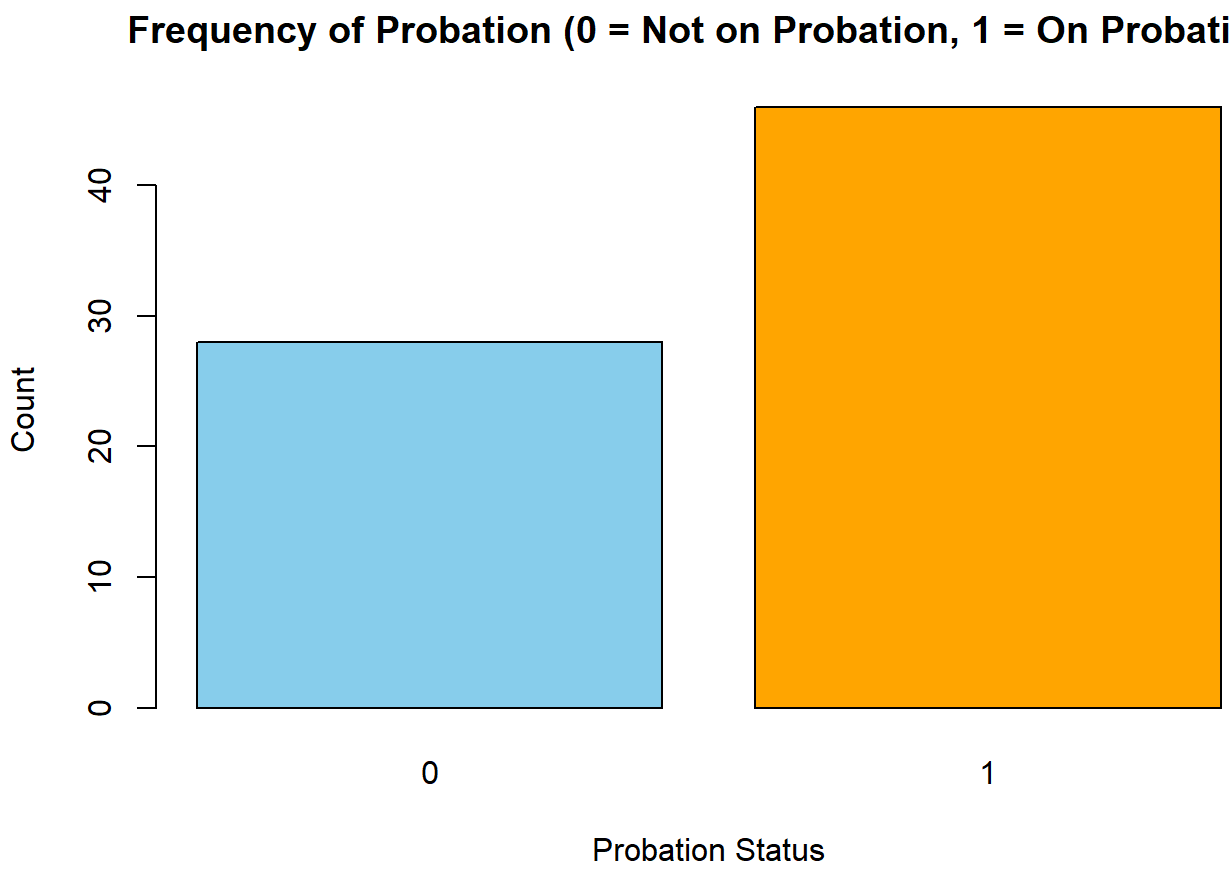


min	Q1	median	Q3	max
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
32.1	63.5	71.65	81.3	98.1

1 row

Analyzing the histogram and boxplot of the quantitative variable Freshman_on_Track_Rate, I observe that the data is LEFT-skewed with one peak and contains one outliers. From the descriptive statistics, I find that the median is 71.65, the mean is 71.9594595, the standard deviation is 14.5520127, and the variance is 211.7610737. Additionally, the interquartile range (IQR) is 17.8

9. The next predictor is Probation



	min	Q1	median	Q3	max
	<int>	<dbl>	<dbl>	<dbl>	<int>
0	28	28	37	46	46

1 row

Table of Probation

0	1
28	46

Analyzing the bar plot of probation status, the most common category is ‘On Probation,’ with 46 students, compared to 28 students who are ‘Not on Probation.’ The frequency range spans from 28 to 46. This represents 62.2% of students on probation compared to 37.8% not on probation, indicating a significant portion of students are on probation.

Multiple Linear Regression Analysis

To predict factors influencing response variable College_Enrollment, I fit a multiple linear regression model with the following predictors:

* x1 = Average_Student_Attendance
* x2 = Rate_of_Misconducts
* x3 = Average_Teacher_Attendance

* x4 = 9th_Grade_Explore_2009
* x5 = X11_Grade_Average_ACT_2011
* x6 = College_Eligibility
* x7 = Graduation_Rate
* x8 = Freshman_on_Track_Rate
* x9 = Probation

The model was fit using an alpha level of 0.05 for significance testing.

Multicollinearity Check

In regression analysis, multicollinearity occurs when two or more predictor variables are highly correlated, which can distort the results and make it difficult to determine the individual effect of each predictor on the response variable. I use Variance Inflation Factor (VIF) to detect multicollinearity. A VIF value above 5 or 10 often indicates a problematic level of multicollinearity, suggesting that one or more variables may need to be removed.

```
## Warning: package 'car' was built under R version 4.1.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.3
```

```
##    apps_data$Average_Student_Attence        apps_data$Rate_of_Misconducts
##                             4.412981                              1.875466
##    apps_data$Average_Teacher_Attence      apps_data$X9_Grade_Explore_2009
##                             1.583707                             29.289103
## apps_data$X11_Grade_Average_ACT_2011        apps_data$College_Eligibility
##                            32.780012                             11.587422
##            apps_data$Graduation_Rate     apps_data$Freshman_on_Track_Rate
##                             3.502900                              2.962817
##                 apps_data$Probation
##                             2.832663
```

Based on our threshold, I reviewed predictors with VIF values greater than 10 and considered removing those with the highest values to improve model stability. For instance, the predictor X11_Grade_Average_ACT_2011 has a VIF of 32.780, X9_Grade_Explore_2009 has a VIF of 29.28, and College_Eligibility has a VIF of 11.587. I first removed X11_Grade_Average_ACT_2011 and re-ran the VIF test to observe any reductions.

```
## apps_data$Average_Student_Attence     apps_data$Rate_of_Misconducts
##                          4.395634                           1.873523
## apps_data$Average_Teacher_Attence   apps_data$X9_Grade_Explore_2009
##                          1.557304                          11.901912
##     apps_data$College_Eligibility         apps_data$Graduation_Rate
##                          9.180288                           3.423573
##  apps_data$Freshman_on_Track_Rate               apps_data$Probation
##                          2.890060                           2.703349
```

After removing 11th_Grade_Average_ACT_2011 from the model, I conducted a Variance Inflation Factor (VIF) analysis and found that the VIF for College_Eligibility had dropped below 10. However, the predictor 9th_Grade_Explore_2009 still exhibited a high VIF of 11.901, indicating persistent multicollinearity, though at a reduced level. To further address multicollinearity, I decided to remove 9th_Grade_Explore_2009 from the model as well.

```
## apps_data$Average_Student_Attence     apps_data$Rate_of_Misconducts
##                          4.395029                          1.771837
## apps_data$Average_Teacher_Attence     apps_data$College_Eligibility
##                          1.542979                          2.557268
##         apps_data$Graduation_Rate  apps_data$Freshman_on_Track_Rate
##                          3.172587                          2.877598
##               apps_data$Probation
##                          2.680117
```

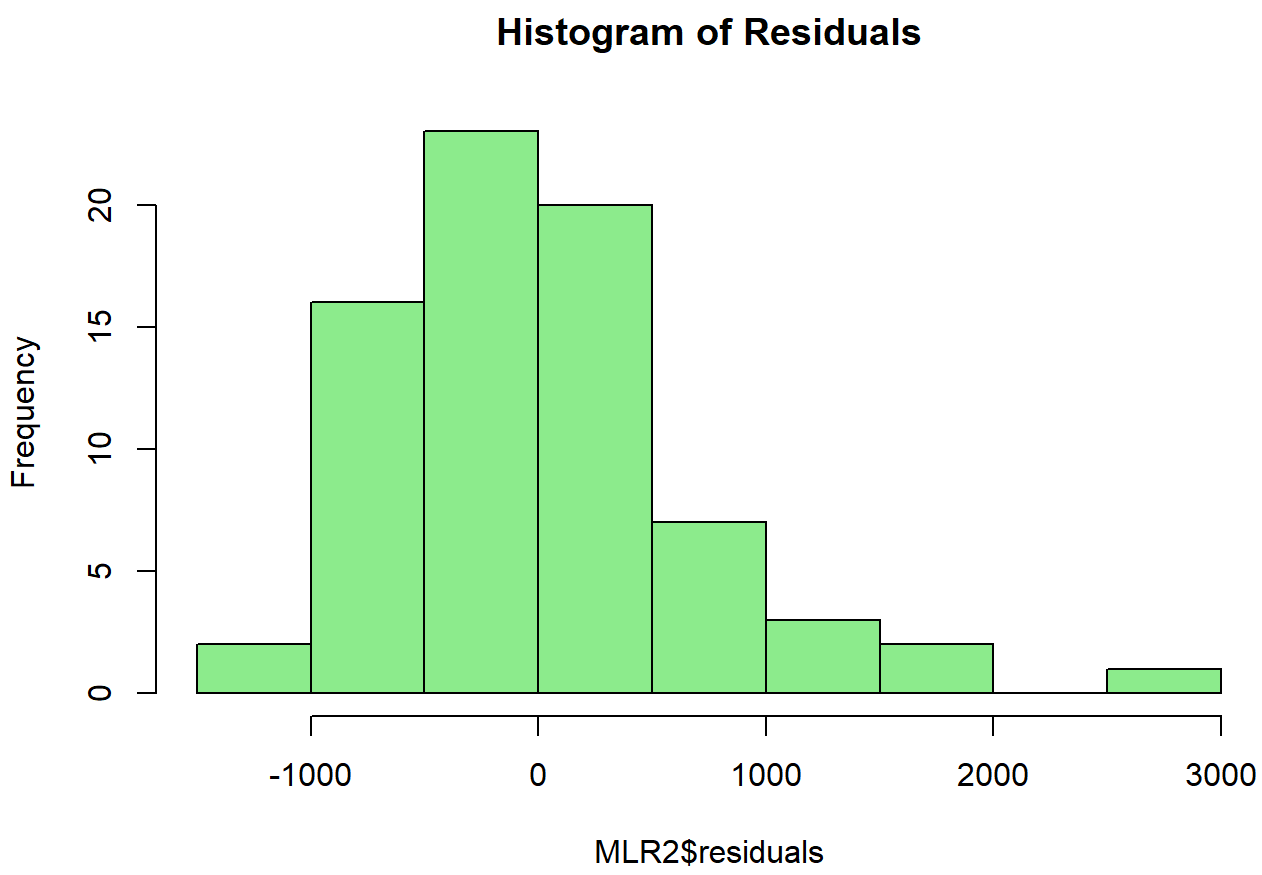
After removing 9th_Grade_Explore_2009 and re-running the VIF analysis, I confirmed that all remaining predictors exhibited lower VIF values, indicating that multicollinearity was no longer a concern. This adjustment significantly improves the model, allowing me to proceed with greater confidence in the stability and interpretability of the regression analysis.

Initial Regression Model

For this analysis, I employed multiple linear regression to examine how various predictors are associated with the dependent variable, College_Enrollment. The predictors included in the model are:
x1 = Average_Student_Attendance
x2 = Rate_of_Misconducts
x3 = Average_Teacher_Attendance
x4 = College_Eligibility
x5 = Graduation_Rate
x6 = Freshman_on_Track_Rate
*x7 = Probation

Multiple linear regression allows us to assess the impact of each predictor on the outcome variable while controlling for the influence of others. This approach helps in understanding the individual and combined effects of these predictors on college_enrollment.

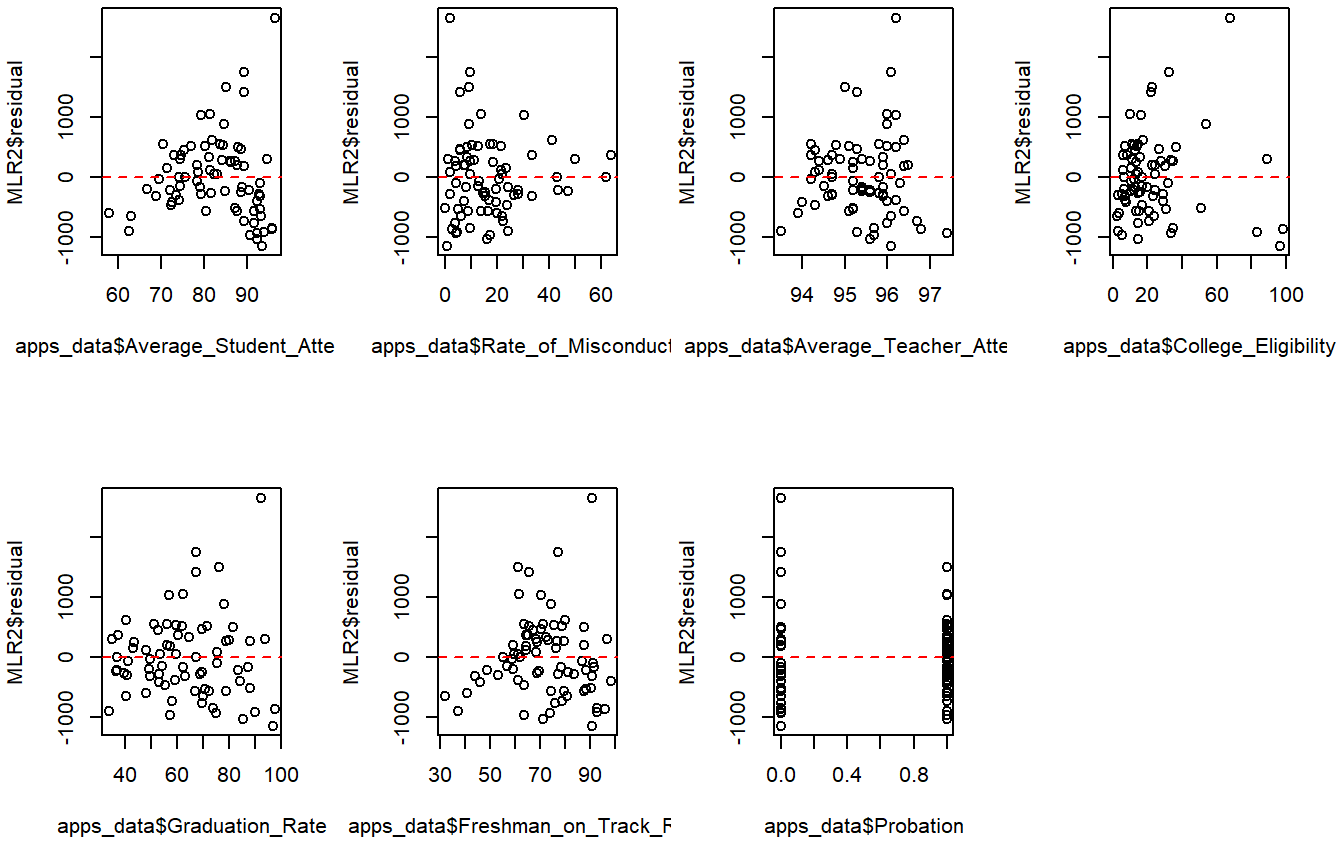
To assess the fit of the model, I first examined the residuals. The histogram below displays the distribution of residuals from the initial full model, providing insight into whether they follow a normal distribution. This check is essential to verify the model's assumptions and to identify any potential issues that may affect the reliability of the regression analysis.



From the histogram, it appears that the residuals are somewhat right-skewed. Most residuals are clustered around 0 and 500. There is also a small portion of negative residuals (below 0), indicating that the model sometimes overpredicts.

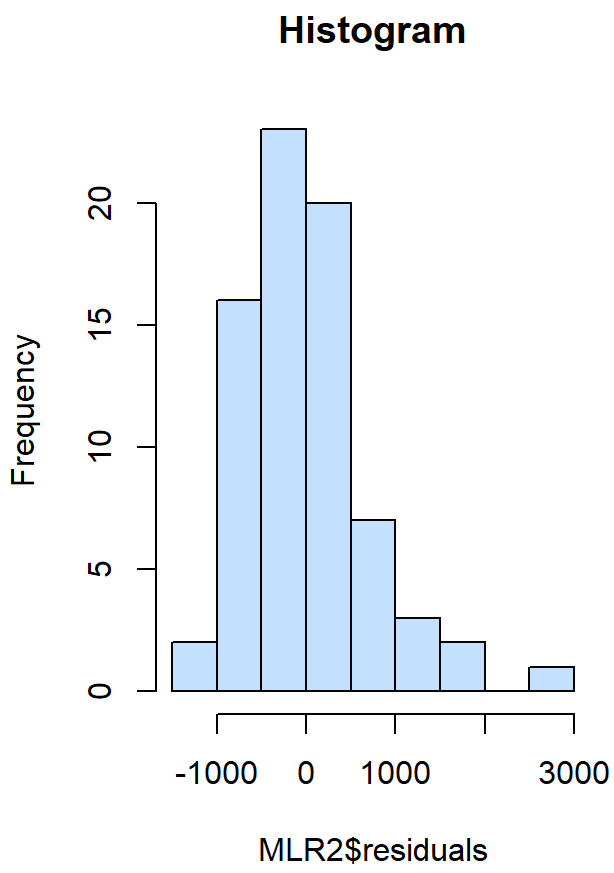
Model Assumptions

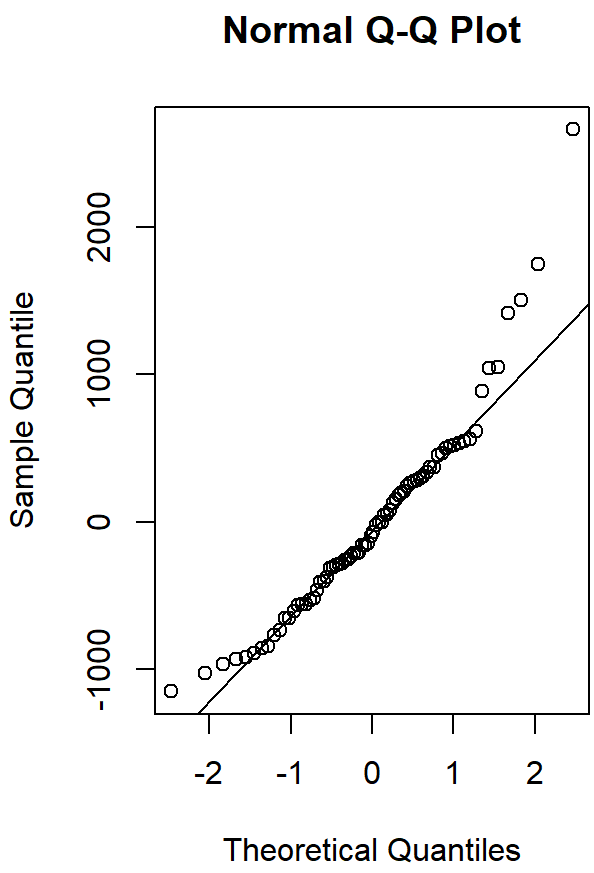
- First I check Linearity



Linearity assumption is not met because I have a list one graph that is not good. Average_Student_Attence has a u-shape patter, Rate_of_Misconducts has a fanned pattern , College_Eligibility is not randomly scattered around the horizontal line at zero.

- Second I check Normality





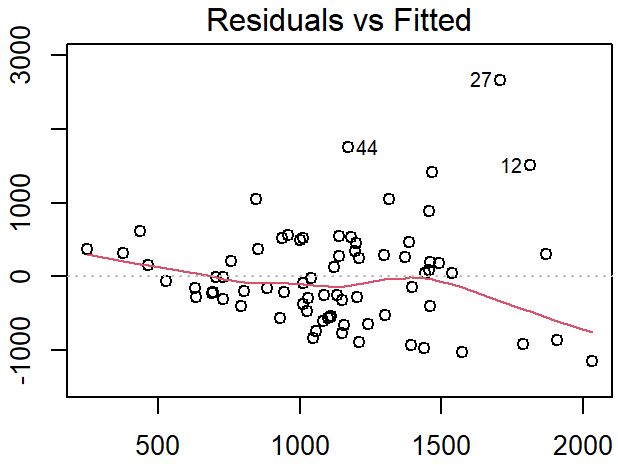
Analyzing histogram I can conclude that it is skewed right so it does not appear that normality is plausible. From QQ Plot I can see that it does not follow reference line very well especially in the tails, indicate departures from normality. Assumption is not met.

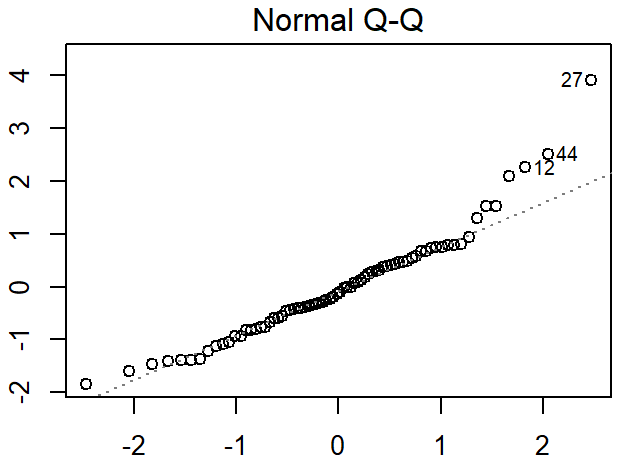
Hypotheses:
H0: Errors are normally distributed
H1: Errors are not normally distributed

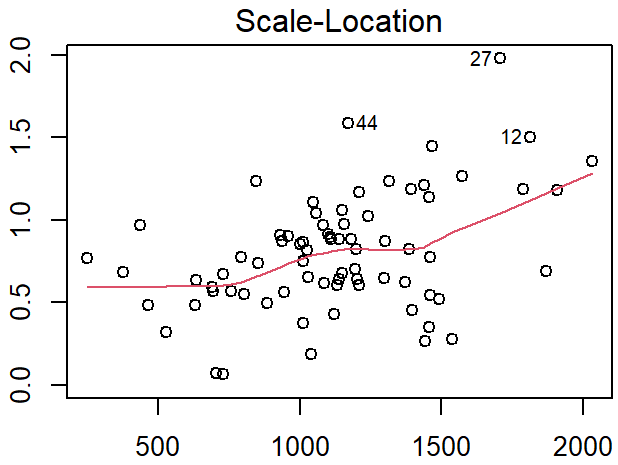
```
##
##  Shapiro-Wilk normality test
##
## data:  MLR2$residuals
## W = 0.93446, p-value = 0.0008413
```

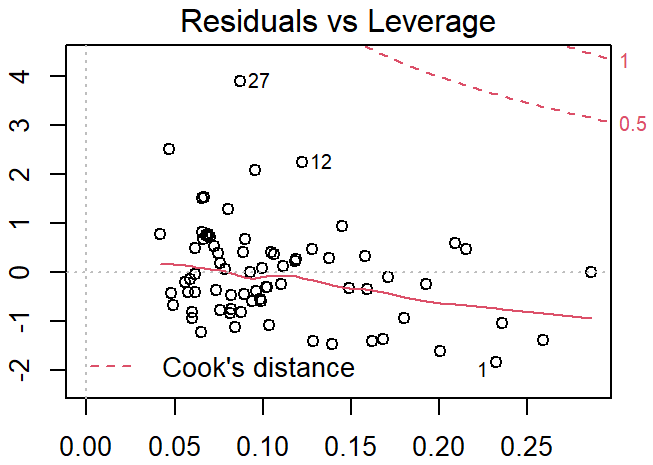
From Shapiro-Wilk Test I can see that the p_value is very small so we Reject H0 (null hypotheses) and can conclude that there are enough evidence that the error are not normally distributed. So, overall this assumption is not met.

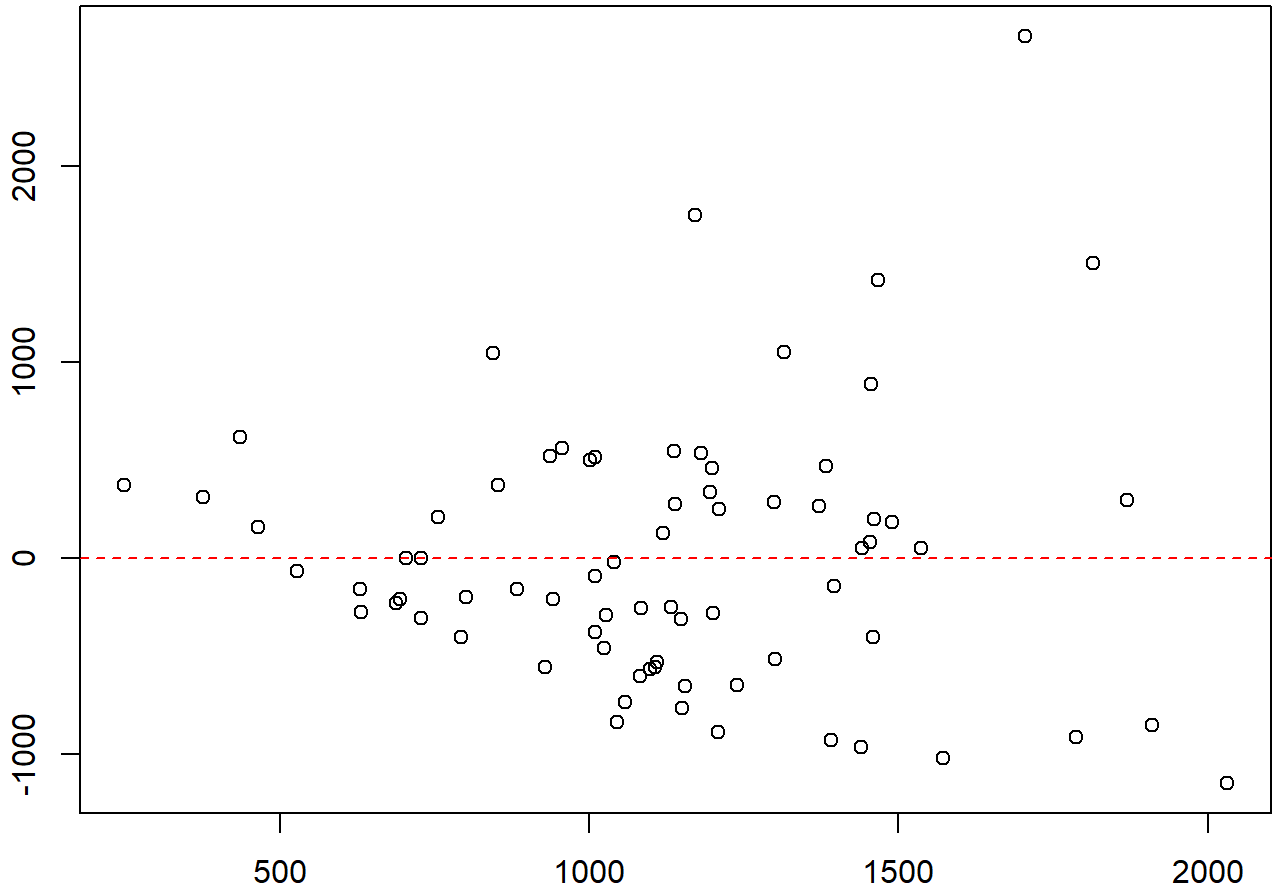
- Third check if assumption of Equal Variance is met


Residuals vs Fitted
27
44
12
3000
1000
0
-1000
500
1000
1500
2000


Normal Q-Q
27
44
12
4
3
2
1
0
-1
-2
-2
-1
0
1
2


Scale-Location
27
44
12
2.0
1.5
1.0
0.5
0.0
500
1000
1500
2000


Residuals vs Leverage
27
12
1
0.5
Cook's distance
1
4
3
2
1
0
-1
-2
0.00
0.05
0.10
0.15
0.20
0.25


2000
1000
0
-1000
500
1000
1500
2000

Analyzing Residual versus Fitted Values plot the residuals do not appear to have a constant spread around zero. There are some potential outliers (points labeled 44, 27, 12) which might be influencing the residual distribution. Also, we can see a pattern. This indicates that the assumption of equal variance is violated.

- Check for Independence
  In this case is not required since it is not time-series data.

Because the assumption of Linearity, Equal Variance, and Normality are not met we will perform Box-Cox transformation

Model Transformation

In our analysis, I identified that the assumptions of linear regression, including linearity, equal variance, and normality of residuals, were not met. To address these violations and improve the model's reliability, it is essential to transform the response variable,Y (in our case, College_Enrollment) using Box-Cox transformation. The Box-Cox transformation is a statistical technique used to improve the suitability of data for analysis, especially in linear regression models. It helps address issues like non-constant variance (when the spread of the data varies) and non-normality (when the data does not follow a bell-shaped distribution). By selecting an appropriate value for λ (such as −2, −1, −0.5, 0, 0.5, 1, or 2), we can transform the data to better meet the model assumptions.

After analyzing the data, I found that the optimal transformation parameter, λ, is 0. Therefore, I will use the logarithmic transformation for College_Enrollment.

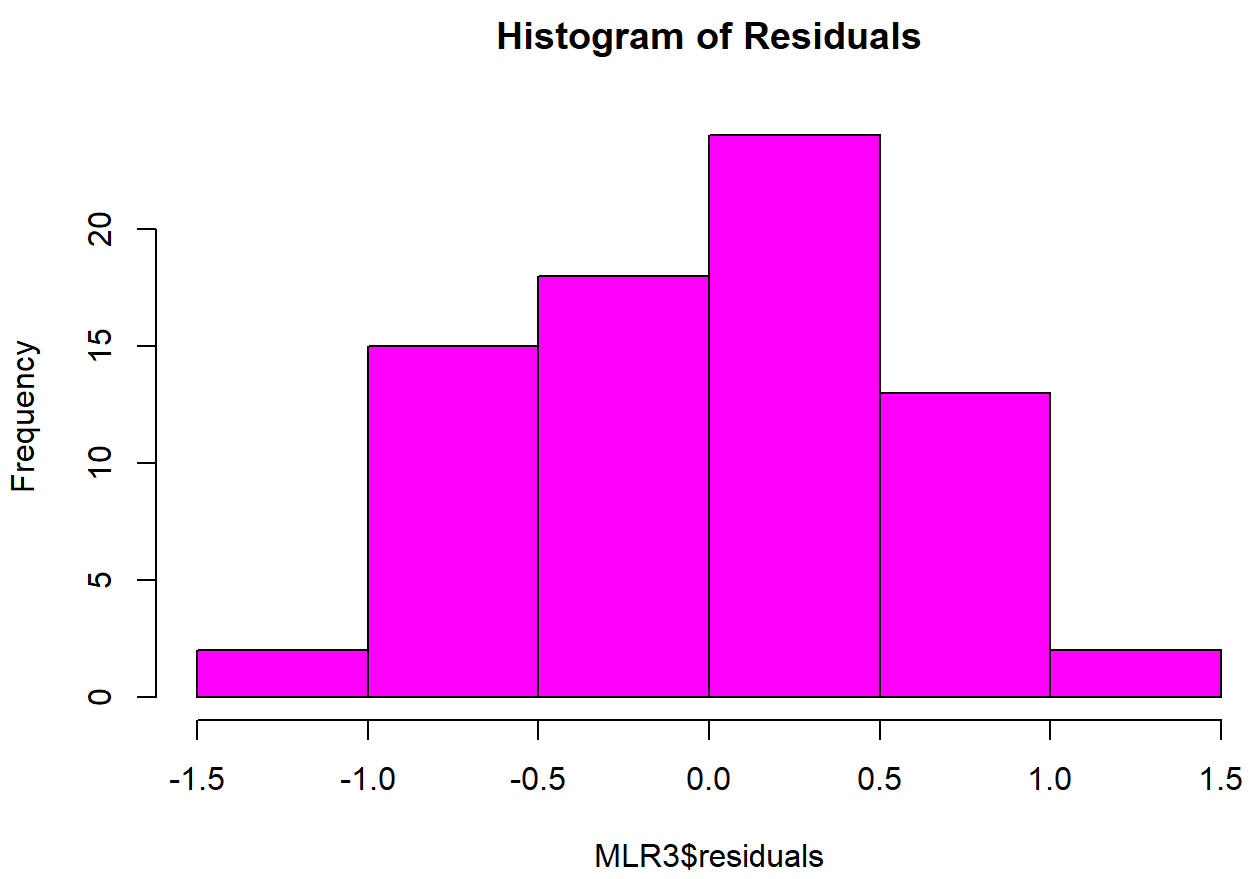
```
## [1] 0
```

After applying the Box-Cox transformation, I created a new column in our data frame called logy, which contains the transformed values of the response variable College_Enrollment. This transformation helps to stabilize variance and improve the normality of the data, aligning it more closely with the assumptions required for linear regression analysis.

Regression Model After Transformation

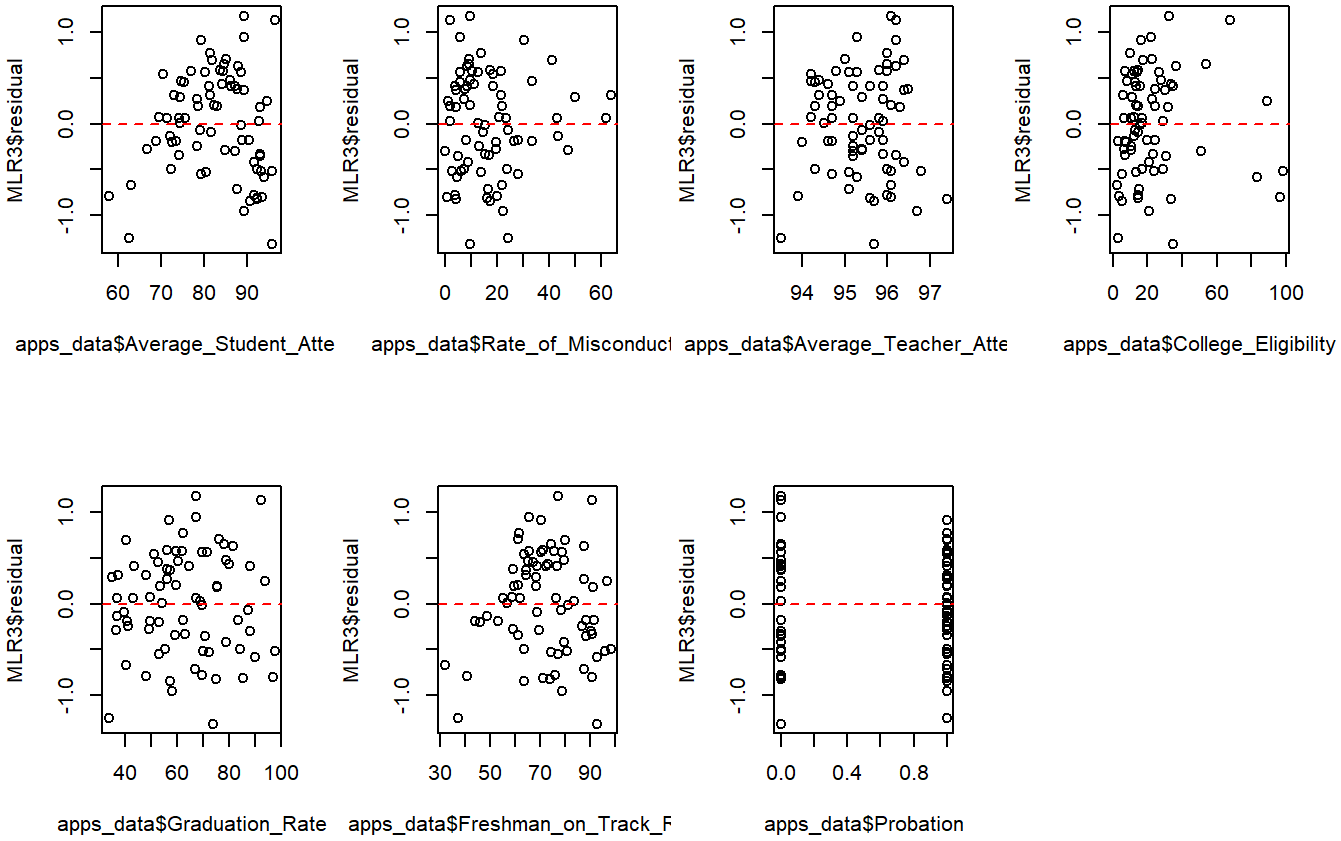
With this new column logy established, I re-ran our regression model, using y = logy as the response variable and next predictors: * x1 = Average_Student_Attence * x2 = Rate_of_Misconducts * x3 = Average_Teacher_Attence * x4 = College_Eligibility * x5 = Graduation_Rate * x6 = Freshman_on_Track_Rate * x7 = Probation

To evaluate the adequacy of the regression model, it is important to assess the behavior of the residuals. The histogram below illustrates the distribution of residuals from the transformed full model, offering a visual indication of whether they approximate a normal distribution. Examining the residual distribution is a critical step in validating the model's assumptions, as deviations from normality can suggest issues that may impact the robustness and reliability of the analysis.



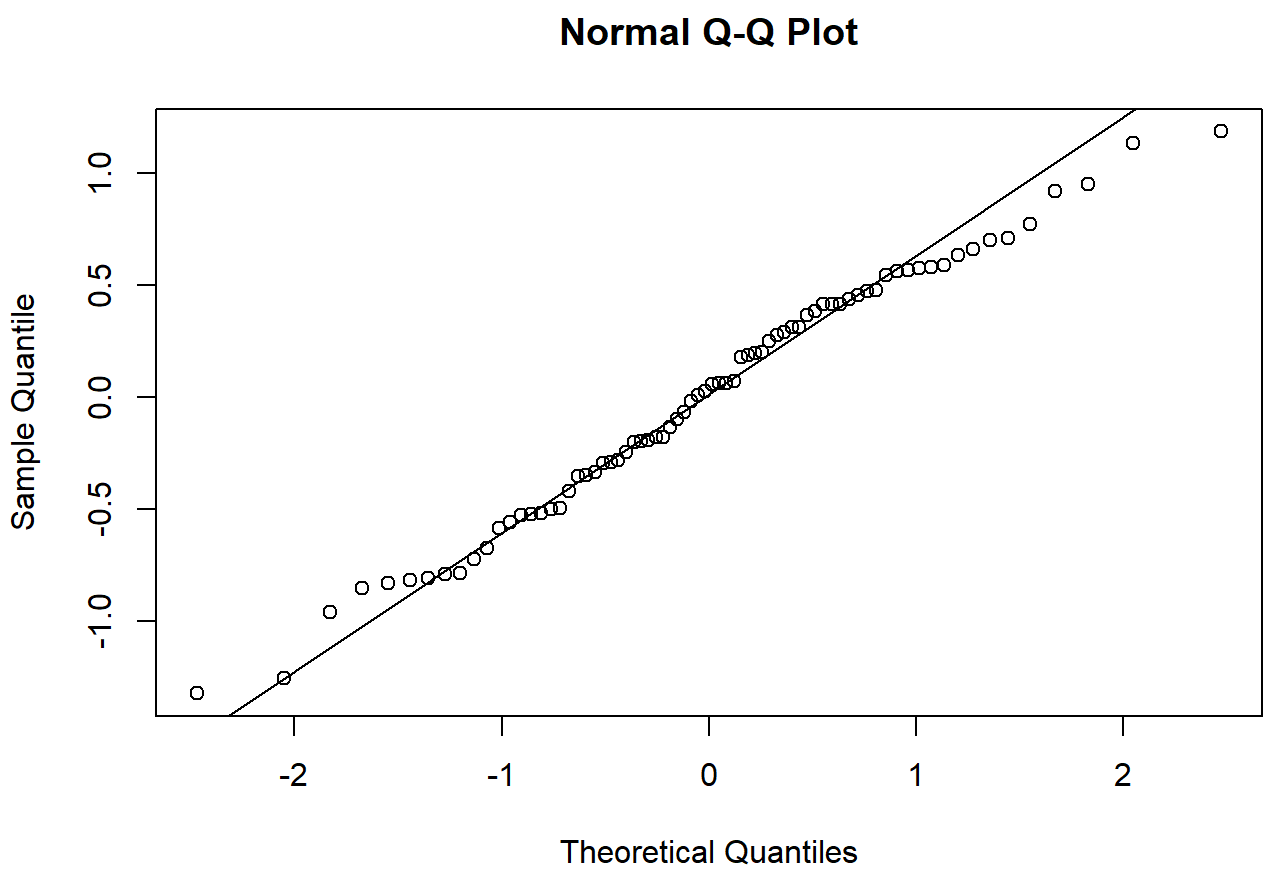
From the histogram, it appears that shows some deviations from normality, as it appears more uniform than bell-shaped. This lack of a clear bell shape suggests that the residuals may not be perfectly normally distributed, which could indicate potential issues with the model fit.

- First I check Linearity



Linearity assumption is not met because I have a list one graph that is not good. Some of the residual plots show noticeable patterns.Some plots appear to have groups of residuals with a potential curve, which might indicate some deviation from linearity in certain predictors. Also, there are some mild trends or clusters in a few variables on Average_Teacher_Attence and Probation, which could suggest slight deviations from linearity.

- Second I check Normality



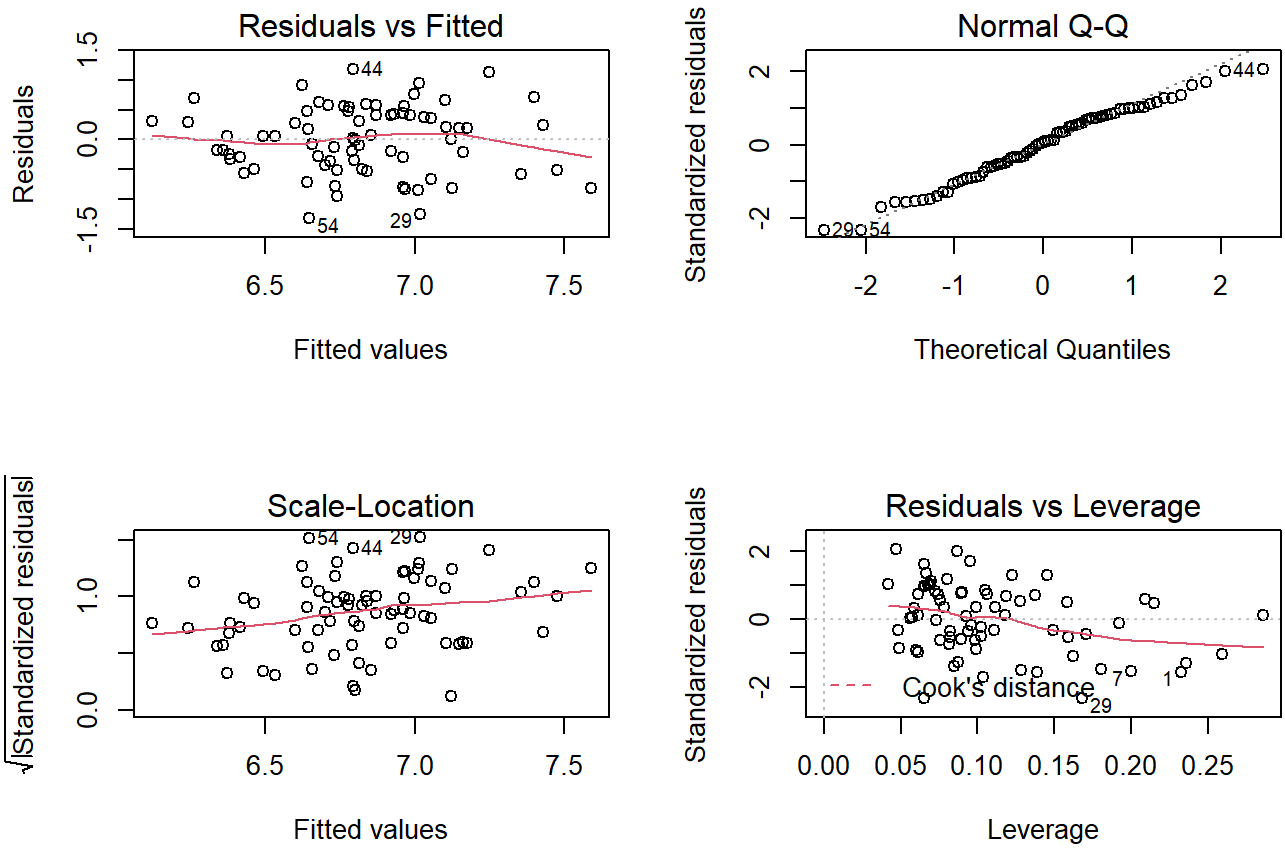
From the QQ plot, I can see that the points lie approximately along the reference line, with some deviation at the tails. Since most points closely follow the line, I can conclude that the assumption of normality is reasonably met.

Hypotheses:
H0: Errors are normally distributed
H1: Errors are not normally distributed

```
##
##  Shapiro-Wilk normality test
##
## data:  MLR3$residuals
## W = 0.98505, p-value = 0.5327
```

From the Shapiro-Wilk Test, I can see that the p-value is large, so I do not reject H0. Therefore, I can conclude that there is not enough evidence to suggest that the errors are not normally distributed. Thus, the assumption is met.

- Third check if assumption of Equal Variance is met



Analyzing Residual versus Fitted Values plot I can conclude the assumption of equal variance seems to be reasonably met, as the spread of residuals appears fairly consistent across the fitted values.

- Check for Independence
  In this case is not required since it is not time-series data.

Checking the Significance of the Overall Model(Hypothesis Tests)

To assess whether the overall model is significant, I employed two approaches to find out : the summary() function and the analysis of variance (ANOVA). The summary() function provides detailed information about the coefficients, standard errors, and statistical significance of each predictor, and p-value.

```
##
## Call:
## lm(formula = apps_data$logy ~ apps_data$Average_Student_Attence +
##     apps_data$Rate_of_Misconducts + apps_data$Average_Teacher_Attence +
##     apps_data$College_Eligibility + apps_data$Graduation_Rate +
##     apps_data$Freshman_on_Track_Rate + apps_data$Probation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32369 -0.40525  0.04081  0.43117  1.18506
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     11.214254  10.085473   1.112  0.27021
## apps_data$Average_Student_Attence  0.016331   0.016327   1.000  0.32085
## apps_data$Rate_of_Misconducts     -0.011972   0.006716  -1.783  0.07925 .
## apps_data$Average_Teacher_Attence -0.049627   0.109754  -0.452  0.65264
## apps_data$College_Eligibility      0.012136   0.005403   2.246  0.02804 *
## apps_data$Graduation_Rate          0.004370   0.007378   0.592  0.55572
## apps_data$Freshman_on_Track_Rate  -0.021661   0.008101  -2.674  0.00944 **
## apps_data$Probation                0.331823   0.232997   1.424  0.15911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5938 on 66 degrees of freedom
## Multiple R-squared:  0.2192, Adjusted R-squared:  0.1364
## F-statistic: 2.647 on 7 and 66 DF,  p-value: 0.01784
```

The ANOVA test helps evaluate the significance of the entire model by comparing the model's fit against a null model.

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(> <db
apps_data$Average_Student_Attence	1	0.217351732	0.217351732	0.616500664	0.435160
apps_data$Rate_of_Misconducts	1	2.053614573	2.053614573	5.824912174	0.018584
apps_data$Average_Teacher_Attence	1	0.109166590	0.109166590	0.309642231	0.579781
apps_data$College_Eligibility	1	1.080279291	1.080279291	3.064125116	0.084685
apps_data$Graduation_Rate	1	0.001734496	0.001734496	0.004919759	0.944293
apps_data$Freshman_on_Track_Rate	1	2.356342645	2.356342645	6.683575944	0.011946
apps_data$Probation	1	0.715060555	0.715060555	2.028211616	0.159113
Residuals	66	23.268773461	0.352557174	*NA*	*N*

8 rows

By interpreting the results from both functions, I can determine whether the model as a whole is statistically significant and whether the predictors are collectively meaningful.

Hypotheses:
H0: beta_1= beta_2 = beta_3= beta_4= beta_5= beta_6 = beta_7 = 0
H1: at least one beta_j is different 0

As result we see that F-statistic is 2.647 and p-value is 0.01784. We Reject H0 because p-value(0.01783) < alpha(0.05) and can conclude that there is There is enough evidence that the x-variables explain some of the variability in Y. Coefficient of Determination is 0.2192 and Adjuster R-Square is 0.1364. 21.92% of the variability in Y is explained by the regression in X.

Because overall model is significant I have to test all predictors separately.

- x1 = Average_Student_Attence

Hypotheses:
H0: beta_1 = 0
H1: beta_1 different 0

From the Partial T-Test values, I see that for Average_Student_Attendance, the t-value is 1.000 and the p-value is 0.32085. Therefore, we Do Not Reject H0 because the p-value (0.32085) is greater than alpha (0.05). To conclude, there is not enough evidence to suggest that X1 (Average_Student_Attendance) is important in explaining some of the variability in Y (logY).

- x2 = Rate_of_Misconducts

Hypotheses:
H0: beta_2 = 0
H1: beta_2 different 0

From the Partial T-Test values, I see that for Rate_of_Misconducts, the t-value is -1.783 and the p-value is 0.07925. Therefore, we Do Dot Deject H0 because the p-value (0.07925) is greater than alpha (0.05). There is not enough evidence to conclude that X2 (Rate_of_Misconducts) is important in explaining some of the variability in Y.

- x3 = Average_Teacher_Attence

Hypotheses:
H0: beta_3 = 0
H1: beta_3 different 0

From the Partial T-Test values, I see that for Average_Teacher_Attendance, the t-value is -0.452 and the p-value is 0.65264. Therefore, we Do Not Reject H0 because the p-value (0.65264) is greater than alpha (0.05). There is not enough evidence to conclude that X3 (Average_Teacher_Attendance) is important in explaining some of the variability in Y.

- x4 = College_Eligibility

Hypotheses:
H0: beta_4 = 0
H1: beta_4 different 0

From the Partial T-Test values, I see that the statistic for College_Eligibility is 2.246 and the p-value is 0.02804. We Reject H0 because the p-value (0.02804) is less than alpha (0.05). In conclusion, there is enough evidence to suggest that X4 (College_Eligibility) is important in explaining some of the variability in Y.

- x5 = Graduation_Rate

Hypotheses:
H0: beta_5 = 0
H1: beta_5 different 0

From Partial T-Test Values we see that for Graduation_Rate t-value is 0.592 and p-value is 0.55572 so, Do Not Reject H0 because p-value(0.55572) > alpha(0.05). There is not enough evidence that X5(Graduation_Rate) is important in explaining some of the variability in Y.

- x6 = Freshman_on_Track_Rate

Hypotheses:
H0: beta_6 = 0
H1: beta_6 different 0

From the Partial T-Test values, I see that the statistic for Freshman_on_Track_Rate is -2.674 and the p-value is 0.00944. We Reject H0 because the p-value (0.00944) is less than alpha (0.05). In conclusion, there is enough evidence to suggest that X6 (Freshman_on_Track_Rate) is important in explaining some of the variability in Y.

- x7 = Probation

Hypotheses:
H0: beta_7 = 0
H1: beta_7 different 0

From the Partial T-Test values, I see that for Probation, the t-value is 1.424 and the p-value is 0.15911. Therefore, we Do Not Reject H0 because the p-value (0.15911) is greater than alpha (0.05). There is not enough evidence to conclude that X7 (Probation) is important in explaining some of the variability in Y.

The analysis shows that the overall model is statistically significant, meaning that some predictors collectively explain part of the variability in college enrollment. However, only College_Eligibility and Freshman_on_Track_Rate are individually significant, indicating they are the most meaningful predictors.

Variable Selection

Since our model indicates statistical significance overall but includes several predictors that are not individually significant, we proceed with backward selection to refine the model. I use backward elimination based on the Akaike Information Criterion (AIC). Unlike traditional backward selection methods that rely on p-values, backward elimination with AIC optimizes the model by minimizing AIC values—where a lower AIC indicates a better fit while balancing model complexity and predictive power. This approach systematically removes non-significant predictors, allowing us to isolate the most impactful variables for predicting College_Enrollment.

I begin by fitting the full model with y = logy as the response variable all potential predictors: * x1 = Average_Student_Attence * x2 = Rate_of_Misconducts * x3 = Average_Teacher_Attence * x4 = College_Eligibility * x5 = Graduation_Rate * x6 = Freshman_on_Track_Rate * x7 = Probation

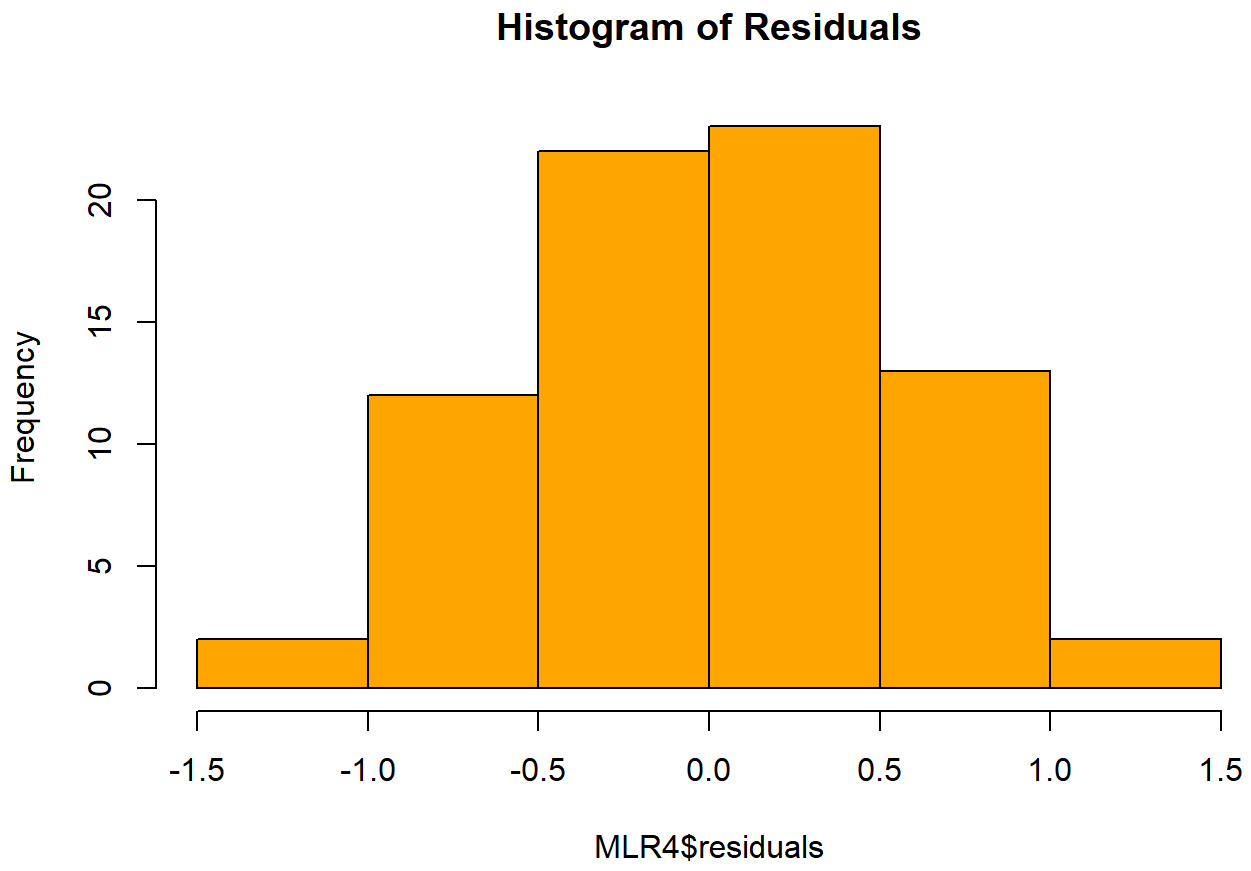
The backward elimination process then iteratively removes the predictor that, when eliminated, results in the largest reduction in AIC. This process continues until no further reduction in AIC is possible, resulting in a model that retains only the most informative predictors.

```
## Start:  AIC=-69.61
## logy ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
##
##        Df Sum of Sq    RSS     AIC
## - x3    1   0.07208 23.341 -71.386
## - x5    1   0.12365 23.392 -71.222
## - x1    1   0.35272 23.622 -70.501
## <none>              23.269 -69.615
## - x7    1   0.71506 23.984 -69.375
## - x2    1   1.12033 24.389 -68.135
## - x4    1   1.77877 25.047 -66.163
## - x6    1   2.52066 25.789 -64.003
##
## Step:  AIC=-71.39
## logy ~ x1 + x2 + x4 + x5 + x6 + x7
##
##        Df Sum of Sq    RSS     AIC
## - x5    1   0.16167 23.503 -72.875
## - x1    1   0.28176 23.623 -72.498
## <none>              23.341 -71.386
## - x7    1   0.74641 24.087 -71.056
## - x2    1   1.12240 24.463 -69.910
## - x4    1   1.75655 25.097 -68.016
## - x6    1   2.50395 25.845 -65.845
##
## Step:  AIC=-72.87
## logy ~ x1 + x2 + x4 + x6 + x7
##
##        Df Sum of Sq    RSS     AIC
## - x1    1   0.35593 23.858 -73.763
## <none>              23.503 -72.875
## - x7    1   0.69976 24.202 -72.704
## - x2    1   1.56658 25.069 -70.100
## - x6    1   2.36154 25.864 -67.790
## - x4    1   2.62434 26.127 -67.041
##
## Step:  AIC=-73.76
## logy ~ x2 + x4 + x6 + x7
##
##        Df Sum of Sq    RSS     AIC
## - x7    1   0.42897 24.287 -74.444
## <none>              23.858 -73.763
## - x2    1   1.86050 25.719 -70.206
## - x6    1   2.17476 26.033 -69.307
## - x4    1   2.66494 26.523 -67.927
##
## Step:  AIC=-74.44
## logy ~ x2 + x4 + x6
##
##        Df Sum of Sq    RSS     AIC
## <none>              24.287 -74.444
## - x2    1    1.4745 25.762 -72.083
```

```
## - x4    1    2.2361 26.523 -69.927
## - x6    1    2.8446 27.132 -68.248
```

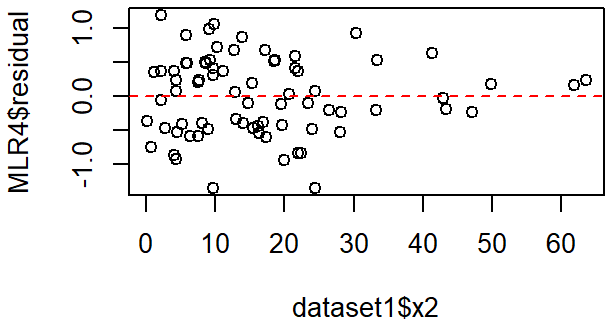
The first predictor removed was Average_Student_Attence, followed by Graduation_Rate, Average_Student_Attence, and Probation. Thus, my final best statistical model is:
log(Y)= beta_0 + beta_2(X2) + beta_4(X4) + beta_6(X6)

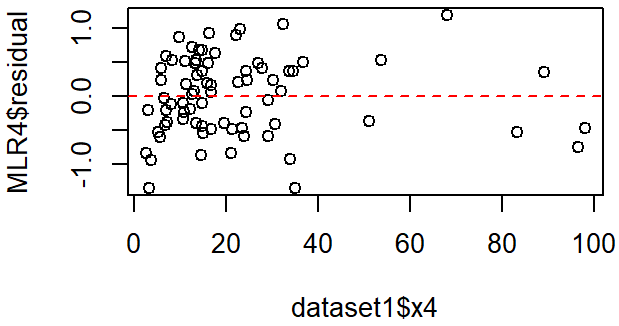
To check if the final best model is good, we need to look at the residuals. The histogram below shows their distribution of residuals from the transformed model, helping us see if they follow a normal distribution. This check is important because if the residuals deviate from normality, it could affect the model's reliability.

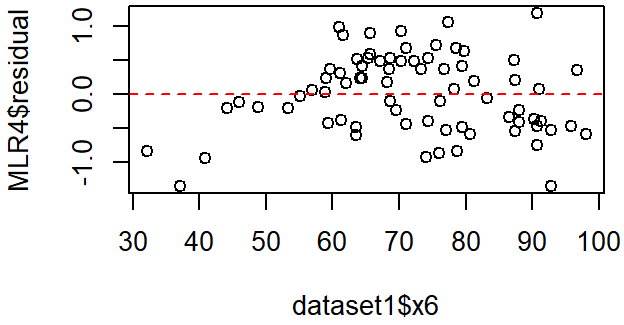


The histogram shows that the residuals are somewhat symmetric and bell-shaped.

- First I check Linearity

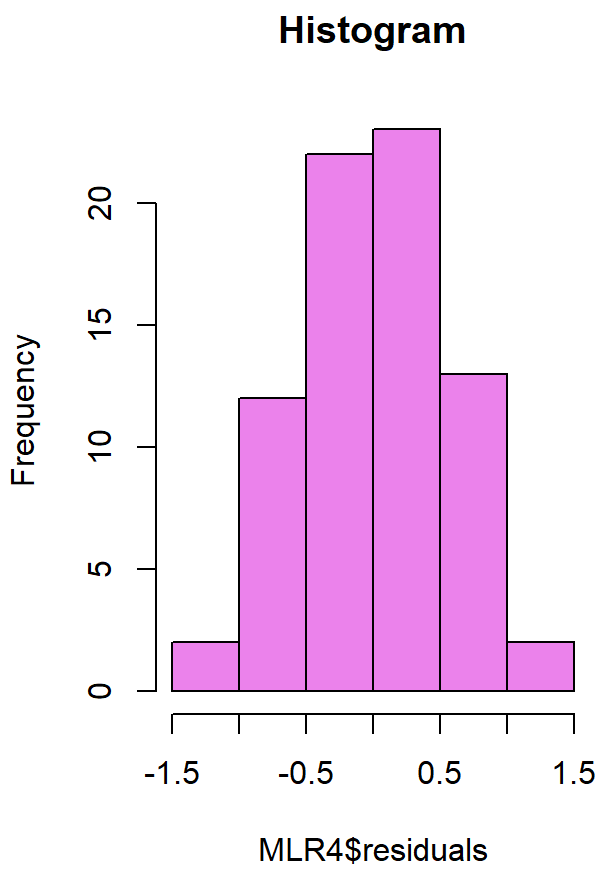


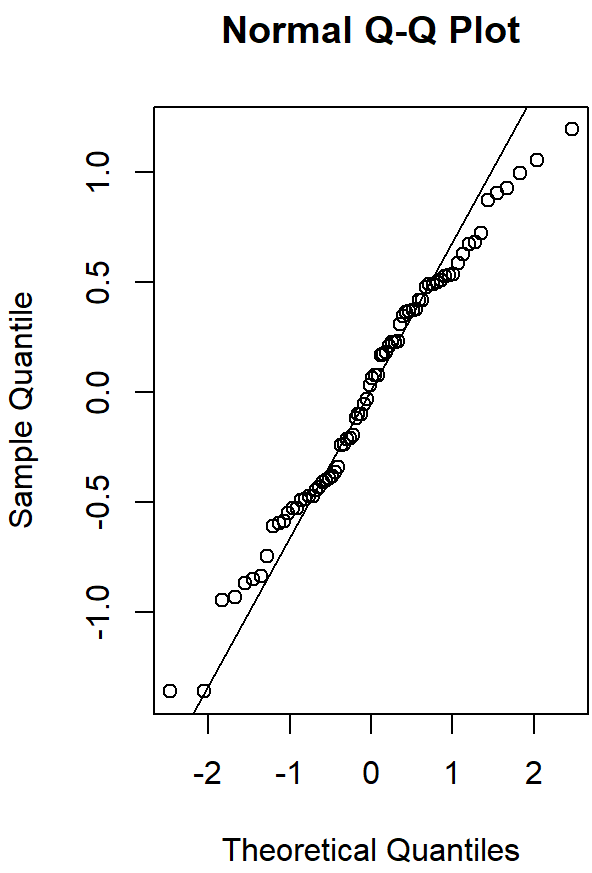




The residuals appear randomly scattered without any clear pattern, suggesting that the linearity assumption is likely met.

- Second I check Normality





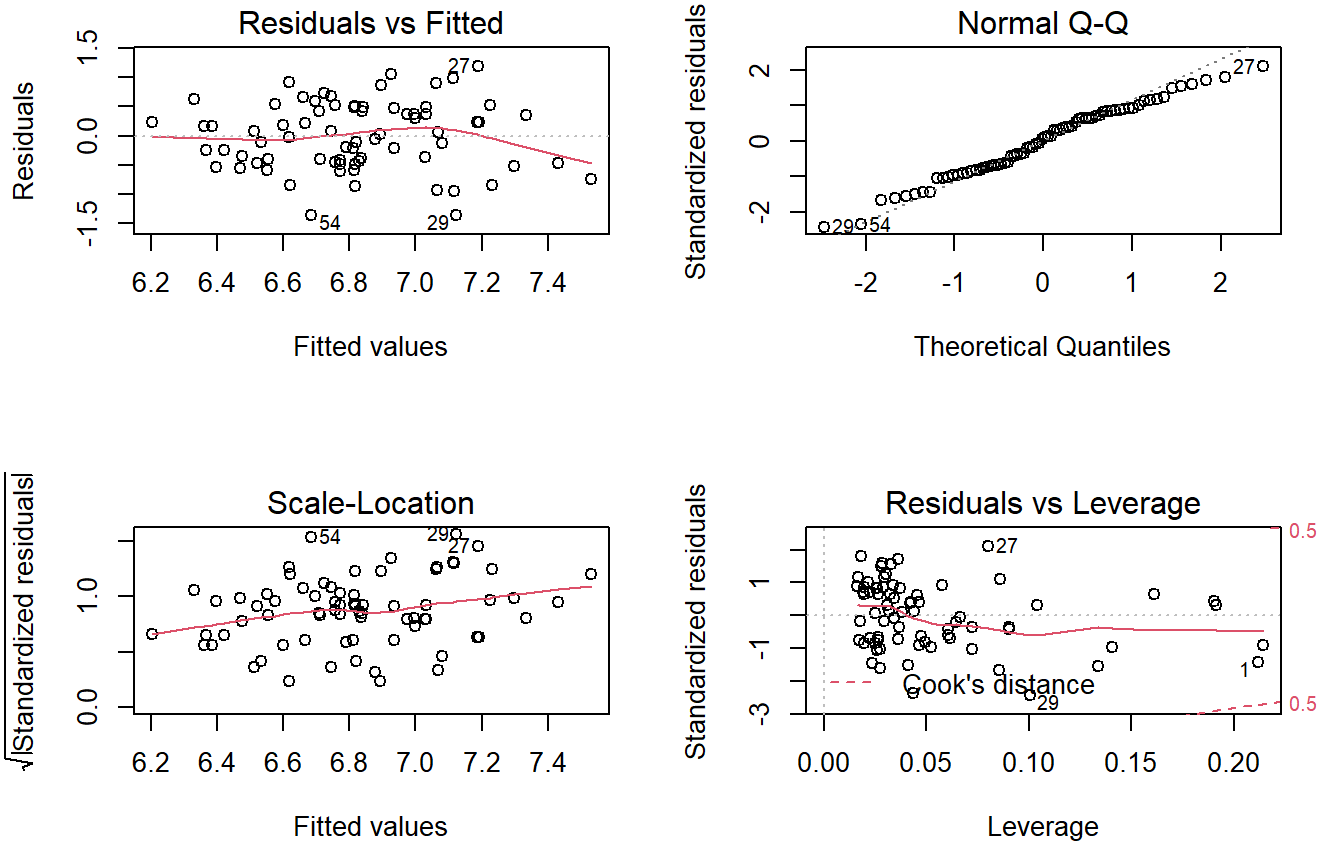
Analyzing histogram I can conclude that it is bell-shaped so it does appear that normality is plausible. From QQ Plot I can see that the points lie approximately along the reference line, except for some deviation at the tails. If most points closely follow the line, we can say that the assumption of normality is reasonably met.

Hypotheses:
H0: Errors are normally distributed
H1: Errors are not normally distributed

```
##
##  Shapiro-Wilk normality test
##
## data:  MLR4$residuals
## W = 0.98242, p-value = 0.3934
```

From Shapiro-Wilk Test I can see that the p_value is large so we Do Not Reject H0. We can conclude that there is not enough evidence that the error are not normally distributed. Assumption is met.

- Third check if assumption of Equal Variance is met



Analyzing the Residuals versus Fitted Values plot, the residuals do not show a clear funnel shape, so the assumption is met.

- Check for Independence
  In this case is not required since it is not time-series data.

Conclusion

The final model for predicting Y is given by the following equation:
log(Y)= 8.033117-0.012105(X2)+0.011393(X4)-0.017559(X6) + error

- log(Y) represents College_Enrollment
- X2 = Rate_of_Misconducts
- x4 = College_Eligibility
- x6 = Freshman_on_Track_Rate

This equation demonstrates that College_Enrollment is influenced by several factors, with the Rate_of_Misconducts, College_Eligibility, and Freshman_on_Track_Rate each playing a distinct role. According to the overall model, 21.92% of the variability in College_Enrollment is explained by these predictors collectively. However, when selecting the best subset of predictors, the model explains 18.50% of the variability in College_Enrollment.

As X2 increase by 1 unit, OR(logy) decrease by 0.0121 units.
As X2 increases by 1 unit, yhat descrease by (exp(-0.0121) - 1)*100% = -1.202%

As X4 increase by 1 unit, OR(logy) increase by 0.0114 units.
As X4 increases by 1 unit, yhat increase by (exp(0.0114) - 1)*100% = 1.147%

As X6 increase by 1 unit, OR(logy) decrease by 0.0176 units.
As X6 increases by 1 unit, yhat descrease by (exp(-0.0176) - 1)*100% = -1.745%