st20258024-CIS6008(1).pdf

by Pathiranalage Dona Himashi Darshika Karunanayaka

Submission date: 14-Aug-2025 08:48PM (UTC+0100)

Submission ID: 263767599

File name: 119746_Pathiranalage_Dona_Himashi_Darshika_Karunanayaka_st20258024-CIS6008_1__2705795_1695585586.pdf (12.46M)

Word count: 10218

Character count: 71884

2

Acknowledgement

In preparation for my Analytics and Business Intelligence assignment, I had to take the help and guidance from some respected persons, who deserve my deepest respect. First, I would like to offer my gratitude to Mr. Roy Ian course instructor, at ICBT Campus, for giving me good guidelines for a better assignment through numerous consultations. I would like to thank all those who helped me in many ways, directly and indirectly, in preparing this assignment.

Thank you.

Table of Contents

Figures

Task A

1.1. Shapiro Test, Anderson-Darling Test and Lillieforce Test

Introduction to Shapiro-Wilk Test

The Shapiro-Wilk test checks whether a dataset follows a normal distribution. It is particularly useful when having small to moderate sample sizes (less than 2000 data points). The test is based on the null hypothesis (H_0) that the sample data originate from a normally distributed population, while the alternative hypothesis (H_1) posits a deviation from normality. (Geeks, 2025)

- W-statistics - Measures how closely the data follows a normal distribution. If the W-statistic is close to 1, it means the data is very close to a normal distribution.
- p-value - Tells us whether the result is statistically significant. If the p-value is greater than 0.05, it means the data does not significantly differ from a normal distribution. If it is less than 0.05, it suggests the data deviates from normality.

Introduction to Anderson-Darling Test

The Anderson-Darling test is used to check if a dataset follows a particular distribution, usually the normal distribution. It improves upon the Kolmogorov-Smirnov test by focusing more on the tails of the distribution (the extremes). The test assesses the null hypothesis (H_0) that the data adhere to the specified distribution, in contrast to the alternative hypothesis (H_1) that they do not. (Pannell, 2022)

- A^2 statistic - This test statistic evaluates the squared differences between the observed and expected cumulative distributions. Elevated values signify a poorer fit.
- p-value - A p-value greater than 0.05 indicates that the data conform to the specified distribution, while a p-value less than 0.05 denotes a significant deviation.

Introduction to Lillieforce Test

The Lilliefors test is a variation of the Kolmogorov-Smirnov test used when the population mean, and variance are unknown. It helps determine if a dataset follows a normal distribution when the parameters are estimated from the data itself. The test assesses the null hypothesis (H_0) that the data originate from a normally distributed population, in contrast to the alternative hypothesis (H_1) that they do not. (Zaiontz, 2025)

- D-statistic - This statistic signifies the maximum discrepancy between the empirical distribution function and the anticipated normal cumulative distribution.
- p-value - A p-value exceeding 0.05 indicates normality, whereas a p-value below 0.05 signifies a notable deviation from normality.

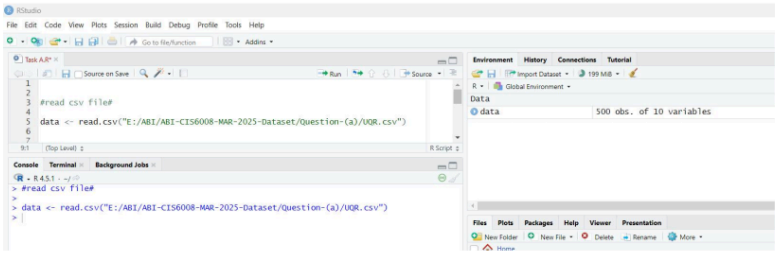


Figure 1- datset

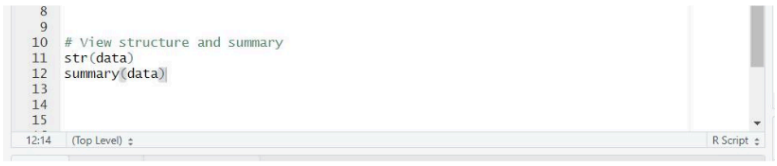


Figure 2- get summry

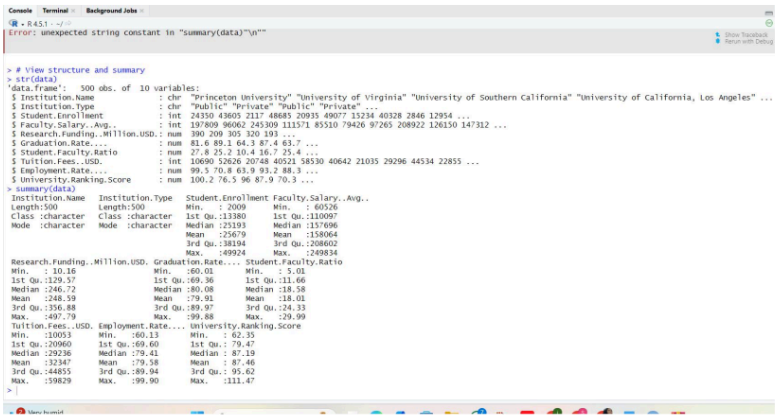


Figure 3- summery result

1.1.1 Student Enrollment

Shapiro Test

```
24
25  # Shapiro-Wilk Test of Student Enrollment
26  shapiro.test(data$Student.Enrollment)
27
```

```
>
> # Shapiro-Wilk Test of Student Enrollment
> shapiro.test(data$Student.Enrollment)

        Shapiro-Wilk normality test

data:  data$Student.Enrollment
W = 0.9529, p-value = 1.557e-11
```

Figure 4 - Shapiro student enrollment

The Shapiro-Wilk test on the student enrollment data yielded W = 0.9529 and p = 1.557e-11, indicating very strong evidence against normality ($p < 0.05$). A W statistic significantly less than 1 suggests a significant deviation from the normal distribution. Thus, the null hypothesis is rejected.

Anderson-Darling Test

```
63
64  # Anderson-Darling Test of Student Enrollment
65  ad.test(data$Student.Enrollment)
66
```

```
>
> # Anderson-Darling Test of Student Enrollment
> ad.test(data$Student.Enrollment)

        Anderson-Darling normality test

data:  data$Student.Enrollment
A = 5.9331, p-value = 1.302e-14
```

Figure 5- AD student enrollment

The Anderson-Darling test for student enrollment yielded A = 5.9331 and p = 1.302e-14, providing strong evidence against normality ($p < 0.05$). The A-statistic measures the difference between the observed and expected distribution, with a high value indicating significant deviation from normality. The large A-value and very small p-value confirm that the student enrollment data does not follow a normal distribution, leading to the rejection of the null hypothesis.

Lilliefors Test

```
102
103  # Lilliefors Test of Student Enrollment
104  lillie.test(data$Student.Enrollment)
105
```

```
>
> # Lilliefors Test of Student Enrollment
> lillie.test(data$Student.Enrollment)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Student.Enrollment
D = 0.075994, p-value = 2.819e-07
```

Figure 6- liliforce student enrollment

The Lilliefors test for student enrollment yielded D = 0.075994 and p = 2.819e-07, indicating significant deviation from normality. The D-statistic measures the maximum distance between the empirical and normal distribution CDF. With a p-value much less than 0.05, the null hypothesis is rejected, confirming that the data does not follow a normal distribution.

Histogram with curve

```r
x <- data$Student.Enrollment

# Define range for x-axis
x_min <- min(x, na.rm = TRUE)
x_max <- max(x, na.rm = TRUE)
x_range <- c(x_min - 1000, x_max + 1000)

# Generate normal curve
x_seq <- seq(x_range[1], x_range[2], length.out = 1000)
normal_curve <- dnorm(x_seq, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE))

# Compute y-axis limit
hist_vals <- hist(x, breaks = 30, plot = FALSE, probability = TRUE)
y_max <- max(max(hist_vals$density), max(normal_curve)) * 1.1

# Draw histogram
hist(x,
     breaks = 30,
     probability = TRUE,
     main = "Histogram of Student Enrollment",
     xlab = "Student Enrollment",
     col = "lightgreen",       # Your preferred color
     border = "black",         # Your preferred border
     xlim = x_range,
     ylim = c(0, y_max))

# Add red normal curve
lines(x_seq, normal_curve, col = "red", lwd = 2)
```

```
> x <- data$Student.Enrollment
>
> # Define range for x-axis
> x_min <- min(x, na.rm = TRUE)
> x_max <- max(x, na.rm = TRUE)
> x_range <- c(x_min - 1000, x_max + 1000)
>
> # Generate normal curve
> x_seq <- seq(x_range[1], x_range[2], length.out = 1000)
> normal_curve <- dnorm(x_seq, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE))
>
> # Compute y-axis limit
> hist_vals <- hist(x, breaks = 30, plot = FALSE, probability = TRUE)
Warning message:
In hist.default(x, breaks = 30, plot = FALSE, probability = TRUE) :
  argument 'probability' is not made use of
> y_max <- max(max(hist_vals$density), max(normal_curve)) * 1.1
>
> # Draw histogram
> hist(x,
+      breaks = 30,
+      probability = TRUE,
+      main = "Histogram of Student Enrollment",
+      xlab = "Student Enrollment",
+      col = "lightgreen",       # Your preferred color
+      border = "black",         # Your preferred border
+      xlim = x_range,
+      ylim = c(0, y_max))
>
> # Add red normal curve
> lines(x_seq, normal_curve, col = "red", lwd = 2)
```

Figure 7 histogram student enrollment

The histogram with an overlaid normal curve visually confirms the deviation from normality. The distribution is right-skewed, with more data concentrated around lower admission values and a long tail extending toward higher values. This aligns with the results of the Shapiro-Wilk, Anderson-Darling, and Lilliefors tests, all rejecting the assumption of normality.

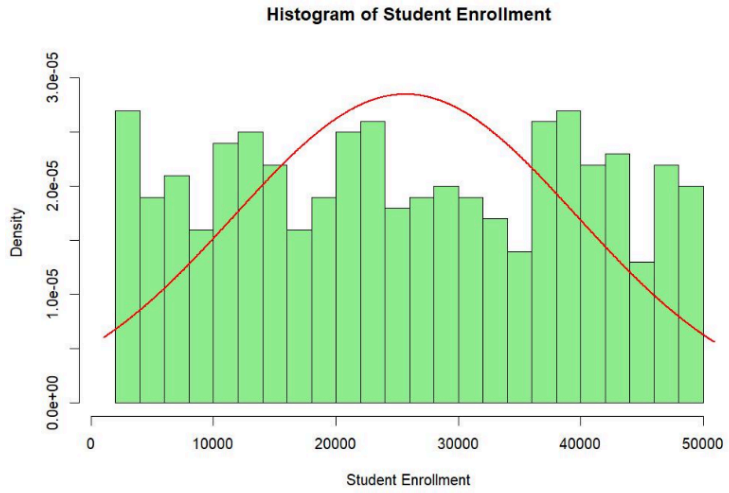


Figure 8 histogram result student enrollment

Summary Of Student Enrollment

The student admissions variable deviates significantly from normality, as indicated by the very low p-values in all three tests. The histogram shows a right-skewed distribution, indicating that most universities have moderate enrollments and a few outliers have very high enrollments. The null hypothesis of normality is strongly rejected.

- ✓ Conclusion - Student enrollment is not normally distributed.

1.1.2. University Ranking Score

Shapiro Test

```
28
29  # Shapiro-Wilk Test of University Ranking Score
30  shapiro.test(data$University.Ranking.Score)
31
```

```
> # Shapiro-Wilk Test of University Ranking Score
> shapiro.test(data$University.Ranking.Score)

        Shapiro-Wilk normality test

data:  data$University.Ranking.Score
W = 0.98619, p-value = 0.0001109
```

Figure 9- Shapiro university ranking score

The Shapiro-Wilk test for University Ranking Score resulted in W = 0.98619 and p = 0.0001109, indicating significant evidence against normality. While the W-statistics is close to 1, the low p-value leads to the rejection of the null hypothesis, suggesting the data does not follow a normal distribution.

Anderson-Darling Test

```
67
68  # Anderson-Darling Test of University Ranking Score
69  ad.test(data$University.Ranking.Score)
70
```

```
> # Anderson-Darling Test of University Ranking Score
> ad.test(data$University.Ranking.Score)

        Anderson-Darling normality test

data:  data$University.Ranking.Score
A = 1.9032, p-value = 7.335e-05

>
```

Figure 10- AD university ranking score

Further supporting this conclusion, the Anderson-Darling test for this variable returned an A value of 1.9032 with a p-value of 7.335e-05, indicating strong disagreement between the observed and theoretical cumulative distribution functions particularly at the tails of the distribution. This large A value implies pronounced deviations, warranting rejection of normality.

Lilliefors Test

```
106
107 # Lilliefors Test of University Ranking Score
108 lillie.test(data$University.Ranking.Score)
109
110
```

```
>
> # Lilliefors Test of University Ranking Score
> lillie.test(data$University.Ranking.Score)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$University.Ranking.Score
D = 0.047407, p-value = 0.009257
```

Figure 11- lilli force university ranking score

The Lilliefors test for normality gave a D-statistic of 0.07599 and a p-value =0.009257 , which is well below 0.05. This test, which adapts the Kolmogorov-Smirnov method for estimated parameters, evaluates the distance between the empirical and theoretical cumulative distribution functions. The significant result confirms that the distribution of University Ranking Score is not normal.

Histogram with curve

```
143
144 # Histogram of University Ranking Score
145
146 x <- as.numeric(data$University.Ranking.Score)
147
148 hist(x, breaks = 30, probability = TRUE,
149      main = "Histogram of University Ranking Score",
150      xlab = "University Ranking Score", col = "lightgreen", border = "black")
151
152 curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
153       add = TRUE, col = "red", lwd = 2)
154
```

```
> # Histogram of University Ranking Score
>
> x <- as.numeric(data$University.Ranking.Score)
>
> hist(x, breaks = 30, probability = TRUE,
+      main = "Histogram of University Ranking Score",
+      xlab = "University Ranking Score", col = "lightgreen", border = "black")
>
> curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
+       add = TRUE, col = "red", lwd = 2)
```

Figure 12- histogram university ranking score

The histogram with a fitted normal curve visually confirms this, showing slight asymmetry and mild peaks, with the curve failing to align perfectly, especially in the tail areas. This further supports the test results indicating deviations from normality.

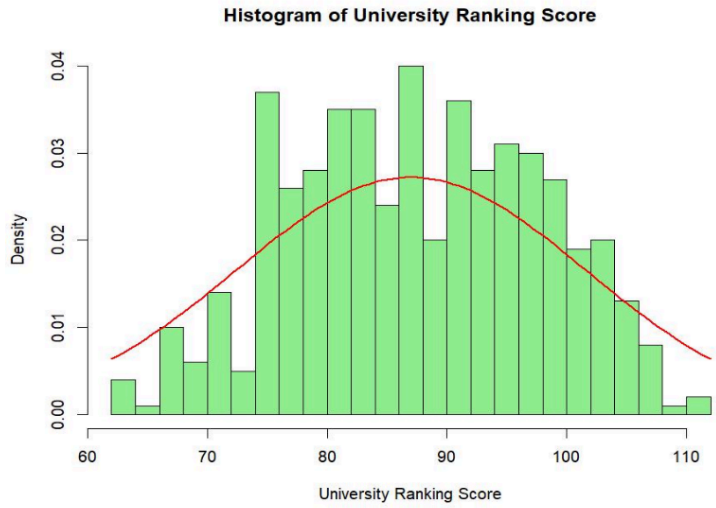


Figure 13 histogram result university ranking score

All three statistical tests Shapiro-Wilk, Anderson-Darling, and Lilliefors produced p-values far below 0.05, confirming that the variable deviates from a normal distribution. The histogram shows mild skewness and irregularities in shape, which do not conform to the expected normal bell curve. These results indicate that the distribution of university ranking scores among the observed universities does not satisfy the assumption of normality.

- ✓ Conclusion – University Ranking Score is not normally distributed.

1.1.3. Faculty Salary

Shapiro Test

```
32
33 # Shapiro-Wilk Test of Faculty.Salary..Avg
34 shapiro.test(data$Faculty.Salary..Avg.)
35
```

```
> # Shapiro-Wilk Test of Faculty.Salary..Avg
> shapiro.test(data$Faculty.Salary..Avg.)

        Shapiro-Wilk normality test

data:  data$Faculty.Salary..Avg.
W = 0.95169, p-value = 1.025e-11
```

Figure 14- Shapiro faculty salary

The Shapiro-Wilk test conducted on the faculty salary data produced a W statistic of 0.95169 and a p-value of 1.025e-11 which indicates very strong evidence against the assumption of normality ($p < 0.05$). The W statistic, being significantly lower than 1, suggests a distinct deviation from the normal distribution. Consequently, the null hypothesis of normality is rejected.

Anderson-Darling Test

```
71
72  # Anderson-Darling Test of Faculty Salary (Avg)
73  ad.test(data$Faculty.Salary..Avg.)
74
```

```
>
> # Anderson-Darling Test of Faculty Salary (Avg)
> ad.test(data$Faculty.Salary..Avg.)

        Anderson-Darling normality test

data:  data$Faculty.Salary..Avg.
A = 6.4205, p-value = 8.971e-16
```

Figure 15- AD faculty salary

The Anderson-Darling test applied to the faculty salary data yielded an A statistic of 6.4205 and a p-value of 8.971e-16 further confirming very strong evidence against normality. The A statistic highlights discrepancies in the tails of the distribution; therefore, a high A-value coupled with a very small p-value substantiates that the data significantly diverges from a theoretical normal distribution. The null hypothesis is unequivocally rejected.

Lilliefors Test

```
110
111  # Lilliefors Test of Faculty Salary (Avg)
112  lillie.test(data$Faculty.Salary..Avg.)
113
```

```
>
> # Lilliefors Test of Faculty Salary (Avg)
> lillie.test(data$Faculty.Salary..Avg.)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Faculty.Salary..Avg.
D = 0.074401, p-value = 5.846e-07

>
```

Figure 16- liliforce faculty salary

In a similar manner, the Lilliefors test resulted in a D statistic of 0.074401 with a p-value of 5.846e-07, indicating that the maximum deviation between the empirical and normal distribution cumulative distribution functions (CDFs) is statistically significant. Given that p <0.05, we reject the null hypothesis and conclude that the faculty salary data does not conform to a normal distribution according to the Lilliefors criterion.

Histogram with curve

```
158 # Histogram of Faculty Salary
159
160 x <- as.numeric(data$Faculty.Salary..Avg.)
161
162 hist(x, breaks = 30, probability = TRUE,
163      main = "Histogram of Faculty Salary (Avg)",
164      xlab = "Faculty Salary (Avg)", col = "lightgreen", border = "black")
165
166 curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
167       add = TRUE, col = "red", lwd = 2)
```

```
> # Histogram of Faculty Salary
>
> x <- as.numeric(data$Faculty.Salary..Avg.)
>
> hist(x, breaks = 30, probability = TRUE,
+      main = "Histogram of Faculty Salary (Avg)",
+      xlab = "Faculty Salary (Avg)", col = "lightgreen", border = "black")
>
> curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
+       add = TRUE, col = "red", lwd = 2)
```

Figure 17- histogram faculty salary

Histogram, overlaid with a normal distribution curve, provides visual corroboration for statistical tests. The distribution exhibits mild right skewness, with many faculty salaries clustered around mid-range values; however, some institutions offer significantly higher salaries, resulting in a long tail to the right. This asymmetry is consistent with the results of all three tests.

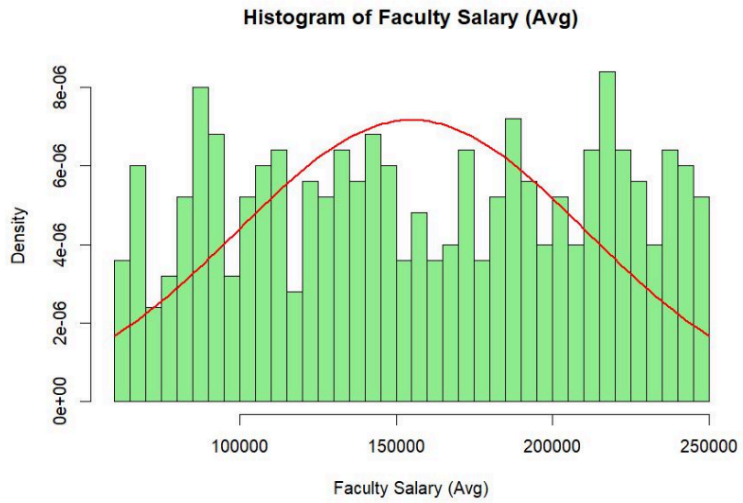


Figure 18- histogram result faculty salary

Summary of Faculty Salary

The faculty salary variable shows a significant deviation from normality, as evidenced by the extremely small p-values obtained from all three tests. The histogram also reinforces this observation by displaying a slight skew and a lack of symmetry. Therefore, all evidence supports the rejection of normality assumption.

- ✓ Conclusion – Faculty Salary (Avg.) is not normally distributed.

1.1.4. Research Funding Million SD

Shapiro Test

```
36
37  # Shapiro-Wilk Test of Research.Funding..Million.USD
38  shapiro.test(data$Research.Funding..Million.USD)
39
```

```
>
> # Shapiro-Wilk Test of Research.Funding..Million.USD
> shapiro.test(data$Research.Funding..Million.USD)

        Shapiro-Wilk normality test

data:  data$Research.Funding..Million.USD
W = 0.96099, p-value = 3.038e-10
```

Figure 19- Shapiro research funding

The Shapiro-Wilk test conducted for research funding resulted in W = 0.96099 and p = 3.038e-10, which provides strong evidence against the assumption of normality ($p < 0.05$). While the W value is relatively close to 1 in comparison to other variables, the significantly low p-value necessitates the rejection of the null hypothesis.

Anderson-Darling Test

```
75
76  # Anderson-Darling Test of Research Funding (Million USD)
77  ad.test(data$Research.Funding..Million.USD)
78
```

```
>
> # Anderson-Darling Test of Research Funding (Million USD)
> ad.test(data$Research.Funding..Million.USD)

        Anderson-Darling normality test

data:  data$Research.Funding..Million.USD
A = 4.5677, p-value = 2.448e-11
```

Figure 20- AD research funding

The Anderson-Darling test yielded an A value of 4.5677 and a p-value of 2.448e-11, suggesting that the empirical distribution significantly deviates from a normal distribution, especially in the tails. The high A-value and extremely low p-value reinforce the conclusion of non-normality in the data. Thus, the null hypothesis is firmly rejected.

Lilliefors Test

```
114
115  # Lilliefors Test of Research Funding (Million USD)
116  lillie.test(data$Research.Funding..Million.USD)
117
```

```
>
> # Lilliefors Test of Research Funding (Million USD)
> lillie.test(data$Research.Funding..Million.USD)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Research.Funding..Million.USD
D = 0.069798, p-value = 4.357e-06
```

Figure 21- Lil force research funding

The Lilliefors test resulted in a D statistic of 0.069798 and a p-value of 4.357e-06, indicating a statistically significant departure from the theoretical normal distribution. Given that the p-value is substantially below 0.05, the null hypothesis is rejected, signifying non-normal behavior in the research funding data.

Histogram with curve

```
184
185 # Histogram of Research Funding (Million USD)
186
187 x <- as.numeric(data$Research.Funding..Million.USD.)
188
189 hist(x, breaks = 30, probability = TRUE,
190      main = "Histogram of Research Funding (Million USD)",
191      xlab = "Research Funding (Million USD)", col = "lightgreen", border = "black")
192
193 curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
194       add = TRUE, col = "red", lwd = 2)
195
```

```
> # Histogram of Research Funding (Million USD)
>
> x <- as.numeric(data$Research.Funding..Million.USD.)
>
> hist(x, breaks = 30, probability = TRUE,
+      main = "Histogram of Research Funding (Million USD)",
+      xlab = "Research Funding (Million USD)", col = "lightgreen", border = "black")
>
> curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
+       add = TRUE, col = "red", lwd = 2)
```

Figure 22 histogram research funding

The histogram, when displayed with an overlaid bell curve, reveals a right-skewed distribution. A significant number of institutions are concentrated around low-to-moderate funding levels, while a few receive exceptionally high funding, extending the tail. This visual asymmetry aligns with the patterns suggested by the statistical analyses.

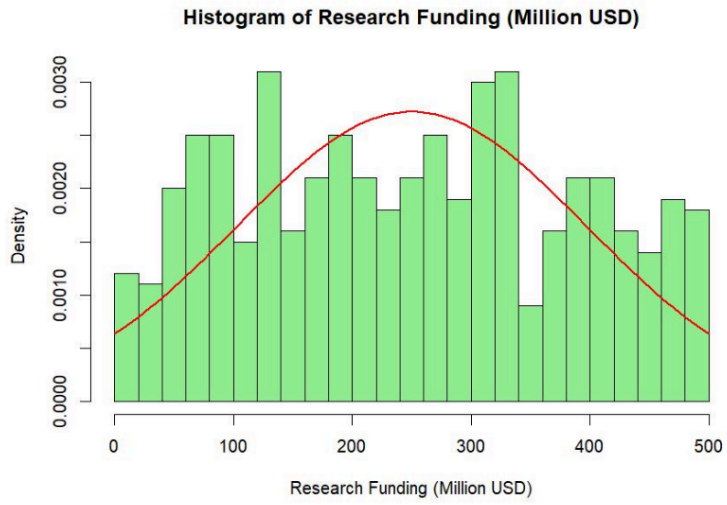


Figure 23- histogram result research funding

Summary of Research Funding (Million USD)

The research funding variable provides clear evidence of non-normality, as indicated by the consistent rejection of the null hypothesis across the Shapiro-Wilk, Anderson-Darling, and Lilliefors tests. The histogram further corroborates this with a skewed distribution pattern, reflecting an unequal distribution among institutions.

- ✓ Conclusion – Research Funding (Million USD) is not normally distributed.

1.1.5. Graduation Rate

Shapiro Test

```
40
41 # Shapiro-Wilk Test of Graduation.Rate..
42 shapiro.test(data$Graduation.Rate..)
43
```

```
> # Shapiro-Wilk Test of Graduation.Rate..
> shapiro.test(data$Graduation.Rate..)

        Shapiro-Wilk normality test

data:  data$Graduation.Rate..
W = 0.94947, p-value = 4.874e-12
```

Figure 24- Shapiro graduation rate

The results of the Shapiro-Wilk normality test for the Graduation Rate variable returned a W statistics of 0.94947 and a p-value of 4.874e-12. These figures provide statistically significant evidence to reject the null hypothesis of normality at the 5% significance level. Although the W value is moderately close to 1, the extremely low p-value implies that the observed distribution deviates substantially from a normal distribution.

Anderson-Darling Test

```
79
80  # Anderson-Darling Test of Graduation Rate (%)
81  ad.test(data$Graduation.Rate..)
82
83
```

```
>
> # Anderson-Darling Test of Graduation Rate (%)
> ad.test(data$Graduation.Rate..)

        Anderson-Darling normality test

data:  data$Graduation.Rate..
A = 6.8091, p-value < 2.2e-16
```

Figure 25- Ad graduation rate

The Anderson-Darling test produced an A value of 6.8091 and a p-value less than 2.2e-16. This clearly reinforces the conclusion that the Graduation Rate variable exhibits non-normal characteristics, especially in the distribution tales. The magnitude of the A statistics reflects substantial deviation between empirical and theoretical cumulative distributions.

Lilliefors Test

```
118
119  # Lilliefors Test of Graduation Rate (%)
120  lillie.test(data$Graduation.Rate..)
121
```

```
> 
> # Lilliefors Test of Graduation Rate (%)
> lillie.test(data$Graduation.Rate..)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Graduation.Rate..
D = 0.07992, p-value = 4.327e-08
```

Figure 26- Lilli force graduation rate

The Lilliefors test for Graduation Rate resulted in a D statistic of 0.07992 and a p-value of 4.327e-08, which also provides strong statistical evidence against normality. The observed D-value indicates a notable distance between the actual and expected normal cumulative distribution functions.

Histogram with curve

```
199  # Histogram of Graduation Rate
200
201  x <- as.numeric(data$Graduation.Rate..)
202
203  hist(x, breaks = 30, probability = TRUE,
204       main = "Histogram of Graduation Rate",
205       xlab = "Graduation Rate (%)", col = "lightgreen", border = "black")
206
207  curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
208        add = TRUE, col = "red", lwd = 2)
```

```
> # Histogram of Graduation Rate
> 
> x <- as.numeric(data$Graduation.Rate..)
> 
> hist(x, breaks = 30, probability = TRUE,
+      main = "Histogram of Graduation Rate",
+      xlab = "Graduation Rate (%)", col = "lightgreen", border = "black")
> 
> curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
+       add = TRUE, col = "red", lwd = 2)
```

Figure 27- histogram graduation rate

The histogram of Graduation Rate, overlaid with a fitted normal curve, illustrates an uneven and flattened shape, lacking the symmetry of a bell-shaped distribution. The mismatch between the histogram bars and the red normal curve, particularly at the distribution tails, visually supports the statistical test results that indicate non-normality.

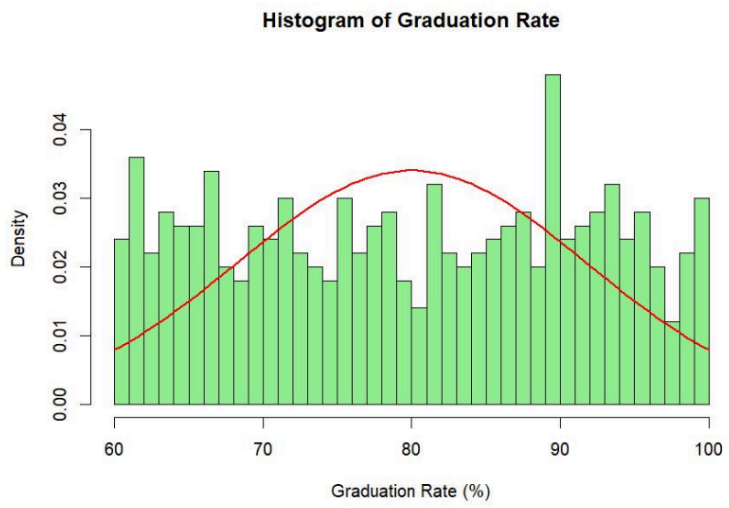


Figure 28- histograme result graduation rate

All three statistical tests consistently reject the null hypothesis, supported by visual evidence of skewness in the histogram. The Graduation Rate data demonstrates clear non-normal characteristics, particularly due to deviations in tail behavior.

- ✓ Conclusion – Graduation Rate (%) is not normally distributed.

1.1.6.Student Faculty Ratio

Shapiro Test

```
44
45  # Shapiro-Wilk Test of Student.Faculty.Ratio
46  shapiro.test(data$Student.Faculty.Ratio)
47
48
```

```
>
> # Shapiro-Wilk Test of Student.Faculty.Ratio
> shapiro.test(data$Student.Faculty.Ratio)

        Shapiro-Wilk normality test

data:  data$Student.Faculty.Ratio
W = 0.95244, p-value = 1.325e-11
```

Figure 29- shapiro student faclty ratio

The Shapiro-Wilk test for the student-Faculty Ratio returned a W statistic of 0.95244 and a p-value of 1.325e-11, indicating strong evidence against the assumption of normality. Although the W value is relatively close to 1, the extremely low p-value confirms a statistically significant deviation from the normal distribution, thereby warranting the rejection of the null hypothesis.

Anderson-Darling Test

```
83
84  # Anderson-Darling Test of Student-Faculty Ratio
85  ad.test(data$Student.Faculty.Ratio)
86
```

```
>
> # Anderson-Darling Test of Student-Faculty Ratio
> ad.test(data$Student.Faculty.Ratio)

        Anderson-Darling normality test

data:  data$Student.Faculty.Ratio
A = 5.8075, p-value = 2.596e-14
```

Figure 30- AD student faculty ratio

The Anderson-Darling test reported an A statistic of 5.8075 with a p-value of 2.596e-14, suggesting considerable divergence from the expected normal cumulative distribution, particularly in the tails. The high A value, along with the very small p-value, strongly supports the conclusion that the distribution does not follow normality. Hence, the null hypothesis is decisively rejected.

Lilliefors Test

```
122
123  # Lilliefors Test of Student-Faculty Ratio
124  lillie.test(data$Student.Faculty.Ratio)
125
```

```
>
> # Lilliefors Test of Student-Faculty Ratio
> lillie.test(data$Student.Faculty.Ratio)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Student.Faculty.Ratio
D = 0.069104, p-value = 5.823e-06
```

Figure 31- Lil force student faculty ratio

In the Lilliefors test, the D statistic was 0.069104 with a p-value of 5.823e-06, offering further statistical support for non-normality. Since the p-value is far below 0.05, the null hypothesis of normality is rejected.

Histogram with curve

```
213  # Histogram of Student-Faculty Ratio
214
215  x <- as.numeric(data$Student.Faculty.Ratio)
216
217  hist(x, breaks = 30, probability = TRUE,
218       main = "Histogram of Student-Faculty Ratio",
219       xlab = "Student-Faculty Ratio", col = "lightgreen", border = "black")
220
221  curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
222        add = TRUE, col = "red", lwd = 2)
```

```
> # Histogram of Student-Faculty Ratio
>
> x <- as.numeric(data$Student.Faculty.Ratio)
>
> hist(x, breaks = 30, probability = TRUE,
+      main = "Histogram of Student-Faculty Ratio",
+      xlab = "Student-Faculty Ratio", col = "lightgreen", border = "black")
>
> curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
+       add = TRUE, col = "red", lwd = 2)
```

Figure 32- histogram student faculty ratio

The histogram for the student-Faculty Ratio, overlaid with a normal distribution curve, reveals a flattened, irregular shape with multiple local peaks. The distribution lacks symmetry and does not conform to the expected bell curve, visually supporting the results from the three statistical tests.

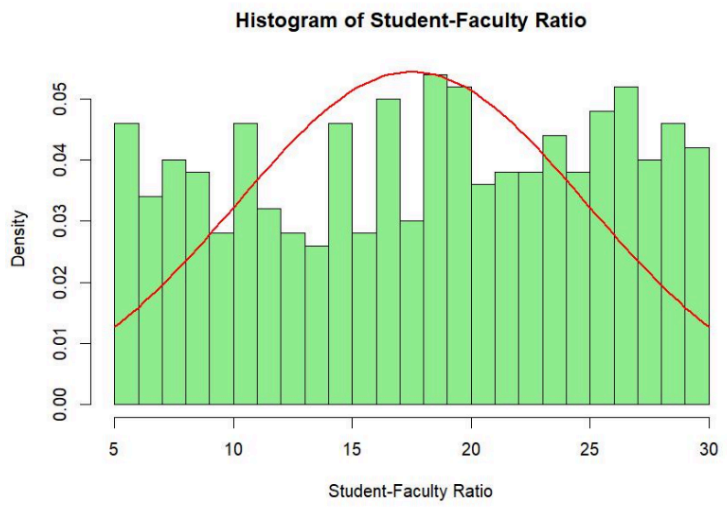


Figure 33- histogram result student faclty ratio

Summary of Student-Faculty Ratio

The student-Faculty Ratio variable shows a clear deviation from a normal distribution. All three normality tests yielded p-values significantly below 0.05, and the histogram visually confirms this non-normal pattern with uneven dispersion across the observed range.

- ✓ Conclusion – Student-Faculty Ratio is not normally distributed.

1.1.7. Tuition Fees USD

Shapiro Test

```
48
49  # Shapiro-Wilk Test of Tuition.Fees..USD
50  shapiro.test(data$Tuition.Fees..USD)
51
```

```
>
> # Shapiro-Wilk Test of Tuition.Fees..USD
> shapiro.test(data$Tuition.Fees..USD)

        Shapiro-Wilk normality test

data:  data$Tuition.Fees..USD
W = 0.94085, p-value = 3.262e-13
```

Figure 34- Shapiro tuition fees

The Shapiro-Wilk test on Tuition Fees data yielded a W value of 0.94085 and a p-value of 3.262e-13, indicating strong statistical evidence against the assumption of normality. The W statistic's deviation from 1, together with the low p-value, necessitates rejecting the null hypothesis.

Anderson-Darling Test

```
87
88  # Anderson-Darling Test of Tuition Fees (USD)
89  ad.test(data$Tuition.Fees..USD)
90
```

```
>
>
> # Anderson-Darling Test of Tuition Fees (USD)
> ad.test(data$Tuition.Fees..USD)

        Anderson-Darling normality test

data:  data$Tuition.Fees..USD
A = 8.7858, p-value < 2.2e-16
```

Figure 35- AD tuition fees

The Anderson-Darling test resulted in an A statistic of 8.7858 and a p-value < 2.2e16, suggesting a significant departure from the normal distribution. The high A value reflects notable tail discrepancies, especially at the higher end of tuition values.

Lilliefors Test

```
126
127  # Lilliefors Test of Tuition Fees (USD)
128  lillie.test(data$Tuition.Fees..USD)
129
```

```
>
>
> # Lilliefors Test of Tuition Fees (USD)
> lillie.test(data$Tuition.Fees..USD)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Tuition.Fees..USD
D = 0.09542, p-value = 9.272e-12
```

Figure 36-Lilliforce tuition fees

The Lilliefors test yielded a D statistic of 0.09542 with a p-value of 9.272e-12, reinforcing the conclusion that the Tuition Fees data does not conform to a normal distribution.

Histogram with curve

```
228  # Histogram of Tuition Fees (USD)
229
230  x <- as.numeric(data$Tuition.Fees..USD.)
231
232  hist(x, breaks = 30, probability = TRUE,
233       main = "Histogram of Tuition Fees (USD)",
234       xlab = "Tuition Fees (USD)", col = "lightgreen", border = "black")
235
236  curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
237        add = TRUE, col = "red", lwd = 2)
```

```
> # Histogram of Tuition Fees (USD)
>
> x <- as.numeric(data$Tuition.Fees..USD.)
>
> hist(x, breaks = 30, probability = TRUE,
+      main = "Histogram of Tuition Fees (USD)",
+      xlab = "Tuition Fees (USD)", col = "lightgreen", border = "black")
>
> curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
+       add = TRUE, col = "red", lwd = 2)
```

Figure 37 - histogram tuition fees

The histogram exhibits a positively skewed distribution, with most observations clustered around lower-to-moderate tuition levels, and a long right tail extending toward higher values. The overlaid bell curve does not fit the data, particularly in the upper range, visually confirming non-normality.

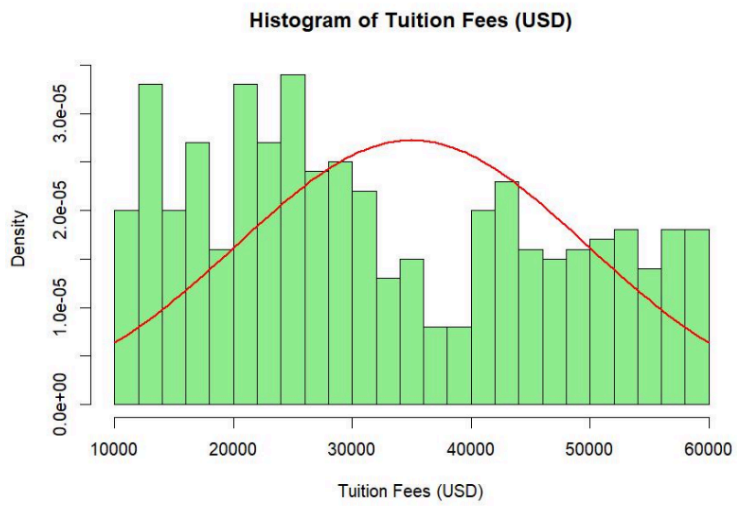


Figure 38- histogram result tuition fees

Summary of Tuition Fees (USD)

All three tests Shapiro-Wilk, Anderson-Darling, and Lilliefors strongly reject the assumption of normality. The histogram further confirms a right-skewed, asymmetric pattern, indicating the presence of high-fee outliers among institutions.

- ✓ Conclusion – Tuition Fees (USD) is not normally distributed.

1.1.8. Employment Rate

Shapiro Test

```
52
53 # Shapiro-Wilk Test of Employment.Rate..
54 shapiro.test(data$Employment.Rate..)
55
```

```
>
> # Shapiro-Wilk Test of Employment.Rate..
> shapiro.test(data$Employment.Rate..)

        Shapiro-Wilk normality test

data:  data$Employment.Rate..
W = 0.95163, p-value = 1.006e-11
```

Figure 39- Shapiro employment rate

The Shapiro-Wilk test on the Employment Rate produced a W statistic of 0.95163 and a p-value of 1.006e-11, demonstrating strong statistical grounds for rejecting the null hypothesis of normality. The W-value, being significantly lower than 1, indicates deviation from the expected normal pattern.

Anderson-Darling Test

```
91
92  # Anderson-Darling Test of Employment Rate (%)
93  ad.test(data$Employment.Rate..)
94
```

```
>
> # Anderson-Darling Test of Employment Rate (%)
> ad.test(data$Employment.Rate..)

        Anderson-Darling normality test

data:  data$Employment.Rate..
A = 6.5412, p-value = 4.632e-16
```

Figure 40- AD employment rate

The Anderson-Darling test recorded an A value of 6.5412 and a p-value of 4.632e-16, suggesting considerable mismatch between the observed and theoretical cumulative distribution functions. The result strongly rejects the null hypothesis, highlighting distribution irregularities, especially in the tails.

Lilliefors Test

```
130
131  # Lilliefors Test of Employment Rate (%)
132  lillie.test(data$Employment.Rate..)
133
134
```

```
> 
> # Lilliefors Test of Employment Rate (%)
> lillie.test(data$Employment.Rate..)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Employment.Rate..
D = 0.082084, p-value = 1.471e-08
```

Figure 41- Lillie force employment rate

The Lilliefors test produced a D statistic of 0.082084 with a p-value of 1.471e-08, which also confirms a statistically significant deviation from normal distribution. The null hypothesis of normality is therefore rejected under this criterion as well.

Histogram with curve

```
243  # Histogram of Employment Rate
244
245  x <- as.numeric(data$Employment.Rate..)
246
247  hist(x, breaks = 30, probability = TRUE,
248       main = "Histogram of Employment Rate",
249       xlab = "Employment Rate (%)", col = "lightgreen", border = "black")
250
251  curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
252        add = TRUE, col = "red", lwd = 2)
```

```
> # Histogram of Employment Rate
>
> x <- as.numeric(data$Employment.Rate..)
>
> hist(x, breaks = 30, probability = TRUE,
+      main = "Histogram of Employment Rate",
+      xlab = "Employment Rate (%)", col = "lightgreen", border = "black")
>
> curve(dnorm(x, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE)),
+       add = TRUE, col = "red", lwd = 2)
```

Figure 42- histogram employment rate

The histogram of Employment Rate (%) displays an irregular, wide distribution without a central peak. Frequencies are dispersed across the entire range, and the red bell curve fails to align with the actual shape particularly at the lower and upper ends further validating the statistical findings.

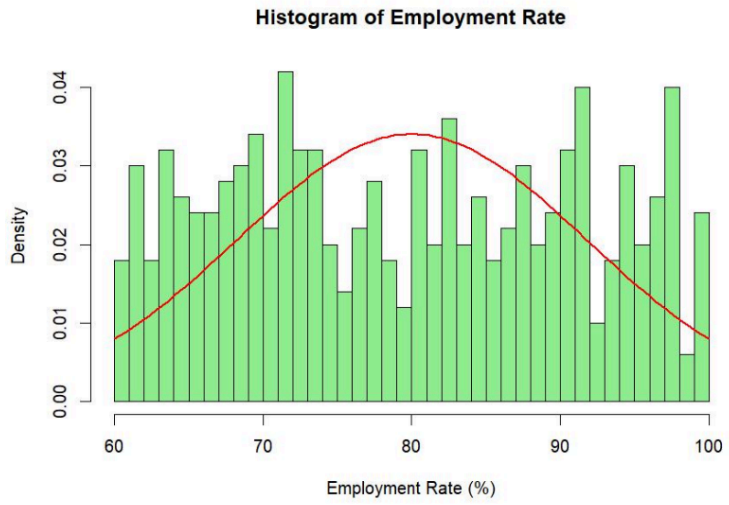


Figure 43- histogram result employment rate

Summary of Employment Rate (%)

The Employment Rate variable does not meet the assumptions of normality. This conclusion is supported by all three tests and further evidenced by the visual skewness and irregular distribution pattern shown in the histogram.

- ✓ Conclusion – Employment Rate (%) is not normally distributed.

Summery table of hypothesis formulations and the results of the normality tests

Variable	Shapiro-Wilk Test (W, p-value)	Anderson-Darling Test (A, p-value)	Lilliefors Test (D, p-value)	Conclusion
Student	W = 0.9529	A = 5.9331	D = 0.075994	Not normally

Enrollment	p = 1.557e-11	p = 1.302e-14	p = 2.819e-07	distributed
University Ranking Score	W = 0.98619 p = 0.0001109	A = 1.9032 p = 7.335e-05	D = 0.047407 p = 0.009257	Not normally distributed
Faculty Salary (Avg.)	W = 0.95169 p = 1.025e-11	A = 6.4205 p = 8.971e-16	D = 0.074401 p = 5.846e-07	Not normally distributed
Research Funding	W = 0.96099 p = 3.038e-10	A = 4.5677 p = 2.448e-11	D = 0.069798 p = 4.357e-06	Not normally distributed
Graduation Rate (%)	W = 0.94947 p = 4.874e-12	A = 6.8091 p < 2.2e-16	D = 0.07992 p = 4.327e-08	Not normally distributed
Student-Faculty Ratio	W = 0.95244 p = 1.325e-11	A = 5.8075 p = 2.596e-14	D = 0.069104 p = 5.823e-06	Not normally distributed
Tuition Fees (USD)	W = 0.94085 p = 3.262e-13	A = 8.7858 p < 2.2e-16	D = 0.09542 p = 9.272e-12	Not normally distributed
Employment Rate (%)	W = 0.95163 p = 1.006e-11	A = 6.5412 p = 4.632e-16	D = 0.082084 p = 1.471e-08	Not normally distributed

1.2 Spearman Correlation Matrix Analysis

This study used the Spearman correlation method to investigate the relationships between university ranking scores and a range of institutional variables in the University Quality Ranking (UQR) dataset. Since the assumptions of data normality were not met, it was appropriate to use Spearman's rank-order correlation, a non-parametric statistical technique, to assess the strength and direction of univariate associations between variables.

```r

# Select Relevant Numeric Columns

numeric_data <- data[, c("University.Ranking.Score",
                         "Faculty.Salary..Avg..",
                         "Research.Funding..Million.USD.",
                         "Graduation.Rate....",
                         "Student.Faculty.Ratio",
                         "Tuition.Fees..USD.",
                         "Employment.Rate....",
                         "Student.Enrollment")]



# Rename Columns

colnames(numeric_data) <- c("University Ranking Score",
                            "Faculty Salary (Avg.)",
                            "Research Funding (Million USD)",
                            "Graduation Rate (%)",
                            "Student-Faculty Ratio",
                            "Tuition Fees (USD)",
                            "Employment Rate (%)",
                            "Student Enrollment")



#Compute Spearman Correlation Matrix

cor_matrix <- cor(numeric_data, method = "spearman", use = "complete.obs")


#Melt Correlation Matrix for ggplot2

cor_melt <- melt(cor_matrix)

```

```r
ggplot(cor_melt, aes(x = Var2, y = Var1, fill = value)) +
  geom_tile(color = "white", linewidth = 0.8) +
  geom_text(aes(label = round(value, 2)), color = "black", size = 4.2, fontface = "bold") +
  scale_fill_gradient2(
    low = "blue",
    mid = "#F7F7F7",
    high = "red",
    midpoint = 0,
    limits = c(-1, 1),
    name = "Correlation"
  ) +
  labs(
    title = "Spearman Correlation Matrix - UQR Dataset",
    x = NULL,
    y = NULL
  ) +
  theme_minimal(base_size = 13) +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 1, size = 10, face = "bold", color = "black"),
    axis.text.y = element_text(size = 10, face = "bold", color = "black"),
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold", color = "black"),
    panel.grid = element_blank(),
    plot.background = element_rect(fill = "white", color = NA),
    panel.border = element_blank()
  )

```

```
> # Select Relevant Numeric Columns
>
> numeric_data <- data[, c("University.Ranking.Score",
+                          "Faculty.Salary..Avg..",
+                          "Research.Funding..Million.USD.",
+                          "Graduation.Rate....",
+                          "Student.Faculty.Ratio",
+                          "Tuition.Fees..USD.",
+                          "Employment.Rate....",
+                          "Student.Enrollment")]
>
>
>
> # Rename Columns
>
> colnames(numeric_data) <- c("University Ranking Score",
+                             "Faculty Salary (Avg.)",
+                             "Research Funding (Million USD)",
+                             "Graduation Rate (%)",
+                             "Student-Faculty Ratio",
+                             "Tuition Fees (USD)",
+                             "Employment Rate (%)",
+                             "Student Enrollment")
>
>
>
> #Compute Spearman Correlation Matrix
>
> cor_matrix <- cor(numeric_data, method = "spearman", use = "complete.obs")
>
>
> #Melt Correlation Matrix for ggplot2
>
> cor_melt <- melt(cor_matrix)
```

```
> ggplot(cor_melt, aes(x = Var2, y = Var1, fill = value)) +
+   geom_tile(color = "white", linewidth = 0.8) +
+   geom_text(aes(label = round(value, 2)), color = "black", size = 4.2, fontface = "bold") +
+   scale_fill_gradient2(
+     low = "blue",
+     mid = "#f7f7f7",
+     high = "red",
+     midpoint = 0,
+     limits = c(-1, 1),
+     name = "Correlation"
+   ) +
+   labs(
+     title = "Spearman Correlation Matrix - UQR Dataset",
+     x = NULL,
+     y = NULL
+   ) +
+   theme_minimal(base_size = 13) +
+   theme(
+     axis.text.x = element_text(angle = 90, hjust = 1, vjust = 1, size = 10, face = "bold", color = "black"),
+     axis.text.y = element_text(size = 10, face = "bold", color = "black"),
+     plot.title = element_text(hjust = 0.5, size = 16, face = "bold", color = "black"),
+     panel.grid = element_blank(),
+     plot.background = element_rect(fill = "white", color = NA),
+     panel.border = element_blank()
+   )
```

Figure 44- Correlation matrix

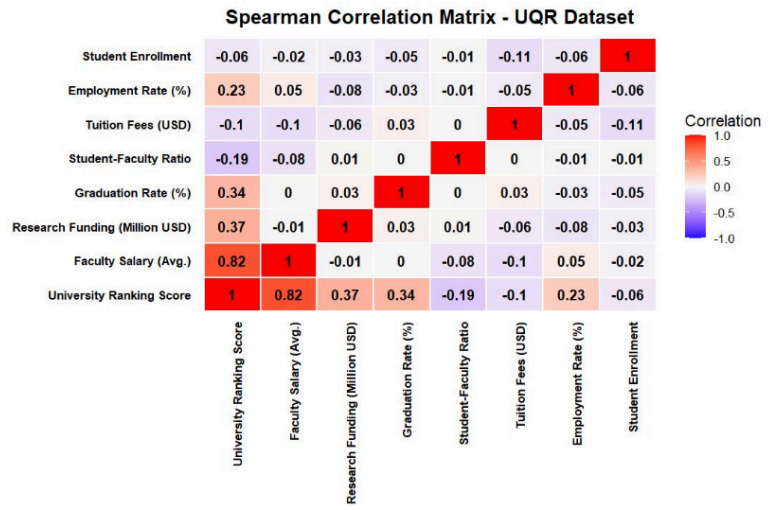


Figure 45 correlation matrix diagram

This analysis explored the correlation between University Ranking Score and various institutional variables using the non-parametric Spearman method. The heatmap revealed strong positive correlations for Faculty Salary ($\rho = 0.82$), Research Funding ($\rho = 0.37$), and Graduation Rate ($\rho = 0.34$), indicating their key role in improving rankings. Employment Rate also showed a positive association ($\rho = 0.23$). On the other hand, Student-Faculty Ratio had a weak negative correlation ($\rho = -0.19$), while Tuition Fees ($\rho = -0.10$) and Student Enrollment ($\rho = -0.06$) showed negligible effects. The findings suggest that Sri Lankan universities should prioritize improving faculty quality, research capacity, and student success to enhance global rankings, rather than focusing on tuition hikes or larger enrollments.

1.3 Correlation Analysis between Predictor Variables and University Ranking Score

To investigate the strength and direction of association between University Ranking Score and various institutional factors, a series of Pearson correlation tests and scatter plot visualizations were conducted. The analysis focuses on seven key numeric predictors.

1.3.1. Student Enrollment vs. University Ranking Score

```
# Clean column names
colnames(data) <- make.names(colnames(data))

# Check to confirm
print(colnames(data))  # look for "Faculty.Salary..Avg.."
```

```
# 1. Correlation: Student Enrollment vs. Ranking Score
cor.test(data$Student.Enrollment, data$University.Ranking.Score)

ggplot(data, aes(x = Student.Enrollment, y = University.Ranking.Score)) +
  geom_point(color = "blue") +
  labs(title = "Student Enrollment vs. University Ranking Score",
       x = "Student Enrollment", y = "University Ranking Score")
```

```
> # Clean column names
> colnames(data) <- make.names(colnames(data))
>
> # Check to confirm
> print(colnames(data))  # look for "Faculty.Salary..Avg.."
 [1] "Institution.Name"              "Institution.Type"
 [3] "Student.Enrollment"            "Faculty.Salary..Avg.."
 [5] "Research.Funding..Million.USD." "Graduation.Rate...."
 [7] "Student.Faculty.Ratio"         "Tuition.Fees..USD."
 [9] "Employment.Rate...."           "University.Ranking.Score"
[11] "RankGroup"
>
```

```
> # 1. Correlation: Student Enrollment vs. Ranking Score
> cor.test(data$Student.Enrollment, data$University.Ranking.Score)

        Pearson's product-moment correlation

data:  data$Student.Enrollment and data$University.Ranking.Score
t = -1.2747, df = 498, p-value = 0.203
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.14399637  0.03081885
sample estimates:
        cor
-0.05702586

>
> ggplot(data, aes(x = Student.Enrollment, y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   labs(title = "Student Enrollment vs. University Ranking Score",
+        x = "Student Enrollment", y = "University Ranking Score")
```

Figure 46 - student entraolement and ranking score correlation

The correlation coefficient (r = -0.057) suggests a very weak negative relationship between student enrollment and university ranking scores. The p-value (0.203) is greater than the standard alpha level of 0.05, indicating that the result is not statistically significant.

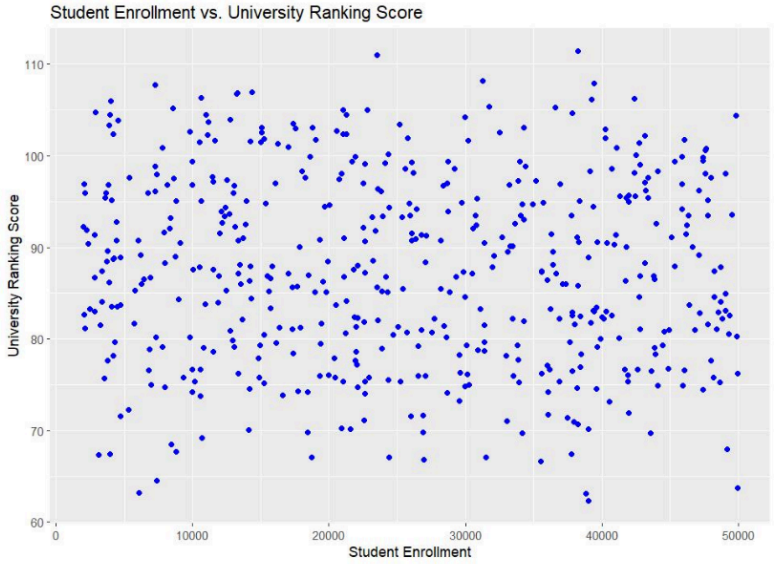


Figure 47 student enrollment and ranking score correlation

The scatter plot of Student Enrollment against University Ranking Score displayed a widely scattered distribution with no evident linear pattern. The corresponding Pearson correlation coefficient indicated a very weak and statistically insignificant relationship. This suggests that the number of enrolled students does not meaningfully impact a university's ranking performance. Therefore, institutional size alone may not be a reliable indicator of academic quality or reputation.

1.3.2.Faculty Salary (Avg.) vs. University Ranking Score

```
# 2. Correlation: Faculty Salary (Avg.) vs. Ranking Score
cor.test(data$Faculty.Salary..Avg.., data$University.Ranking.Score)

ggplot(data, aes(x = Faculty.Salary..Avg.., y = University.Ranking.Score)) +
  geom_point(color = "blue") +
  labs(title = "Faculty Salary vs. University Ranking Score",
       x = "Faculty Salary (Avg.)", y = "University Ranking Score")
```

```
> # 2. Correlation: Faculty Salary (Avg.) vs. Ranking Score
> cor.test(data$Faculty.Salary..Avg.., data$University.Ranking.Score)

        Pearson's product-moment correlation

data:  data$Faculty.Salary..Avg.. and data$University.Ranking.Score
t = 31.648, df = 498, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7858858 0.8444294
sample estimates:
      cor
0.8172554

>
> ggplot(data, aes(x = Faculty.Salary..Avg.., y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   labs(title = "Faculty Salary vs. University Ranking Score",
+        x = "Faculty Salary (Avg.)", y = "University Ranking Score")
```

Figure 48- faculty salary and ranking score correlation

The correlation coefficient (r = 0.817) indicates a strong positive relationship between faculty salary and university ranking score. The p-value is less than 0.05, which means the result is statistically significant. The confidence interval further confirms that the true correlation lies between 0.785858 and 0.844429, indicating a strong and consistent relationship between the two variables.

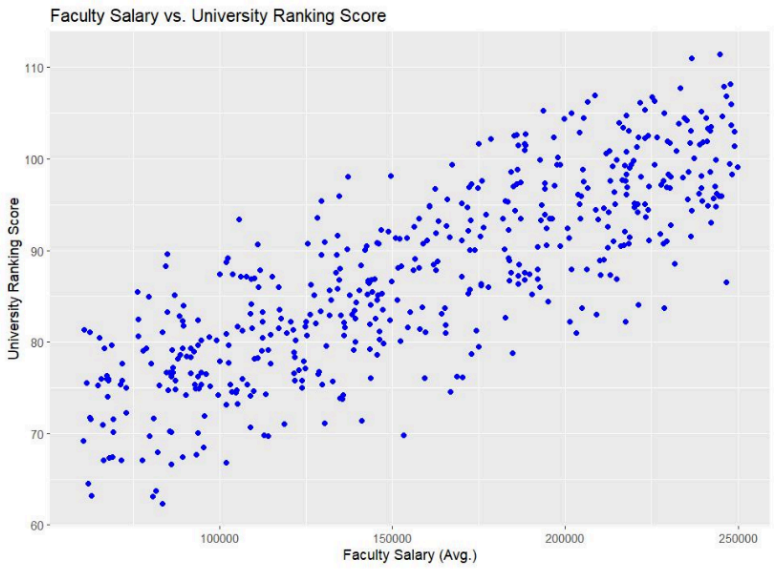


Figure 49- faculty salary and ranking score correlation

A strong positive relationship was identified between average faculty salary and University Ranking Score. The scatter plot exhibited a clear upward trend, and the Pearson correlation coefficient confirmed a statistically significant and strong correlation. This implies that institutions with higher faculty compensation tend to achieve better ranking outcomes, possibly due to their ability to attract and retain highly qualified academic staff, which contributes to research output and educational quality.

1.3.3.Research Funding (Million USD) vs. University Ranking Score

```
# 3. Correlation: Research Funding vs. Ranking Score
cor.test(data$Research.Funding..Million.USD., data$University.Ranking.Score)

ggplot(data, aes(x = Research.Funding..Million.USD., y = University.Ranking.Score)
  geom_point(color = "blue") +
  labs(title = "Research Funding vs. University Ranking Score",
       x = "Research Funding (Million USD)", y = "University Ranking Score")
```

```
> # 3. Correlation: Research Funding vs. Ranking Score
> cor.test(data$Research.Funding..Million.USD., data$University.Ranking.Score)

        Pearson's product-moment correlation

data:  data$Research.Funding..Million.USD. and data$University.Ranking.Score
t = 9.3039, df = 498, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3074976 0.4570786
sample estimates:
      cor
0.3848119

>
> ggplot(data, aes(x = Research.Funding..Million.USD., y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   labs(title = "Research Funding vs. University Ranking Score",
+        x = "Research Funding (Million USD)", y = "University Ranking Score")
```

Figure 50- research funding and ranking score correlation

The correlation coefficient (r = 0.385) indicates a moderate positive relationship between research funding and university ranking scores. The p-value is less than 0.05, which suggests that the result is statistically significant. The confidence interval indicates that the true correlation lies between 0.3075 and 0.4571, further supporting the moderate positive relationship. This result suggests that higher research funding is associated with better university ranking scores.

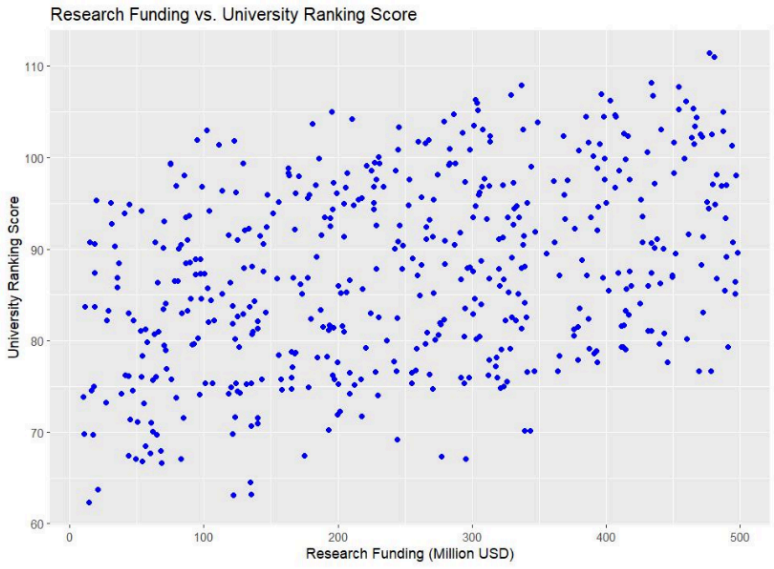


Figure 51- research funding and ranking score correlation

A moderate positive association was observed between research funding and university ranking. The scatter plot revealed a partial upward trend, while the correlation coefficient indicated a statistically significant, moderately strong relationship. This finding suggests that increased financial investment in research activities can enhance a university's ranking, potentially through improved academic visibility, innovation, and scholarly impact.

1.3.4. Graduation Rate (%) vs. University Ranking Score

```
389  # 4. Correlation: Graduation Rate vs. Ranking Score
390  cor.test(data$Graduation.Rate...., data$University.Ranking.Score)
391
392  ggplot(data, aes(x = Graduation.Rate...., y = University.Ranking.Score)) +
393    geom_point(color = "blue") +
394    labs(title = "Graduation Rate vs. University Ranking Score",
395         x = "Graduation Rate (%)", y = "University Ranking Score")
```

```
> # 4. Correlation: Graduation Rate vs. Ranking Score
> cor.test(data$Graduation.Rate...., data$University.Ranking.Score)

        Pearson's product-moment correlation

data:  data$Graduation.Rate.... and data$University.Ranking.Score
t = 8.3171, df = 498, p-value = 8.684e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2698033 0.4239391
sample estimates:
      cor
0.3492313

>
> ggplot(data, aes(x = Graduation.Rate...., y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   labs(title = "Graduation Rate vs. University Ranking Score",
+        x = "Graduation Rate (%)", y = "University Ranking Score")
```

Figure 52- graduation rate and ranking score correlation

The correlation coefficient (r = 0.349) suggests a moderate positive relationship between graduation rate and university ranking score. The p-value is highly significant (less than 0.05), confirming the relationship. The confidence interval shows that the true correlation is likely between 0.2698 and 0.4239.

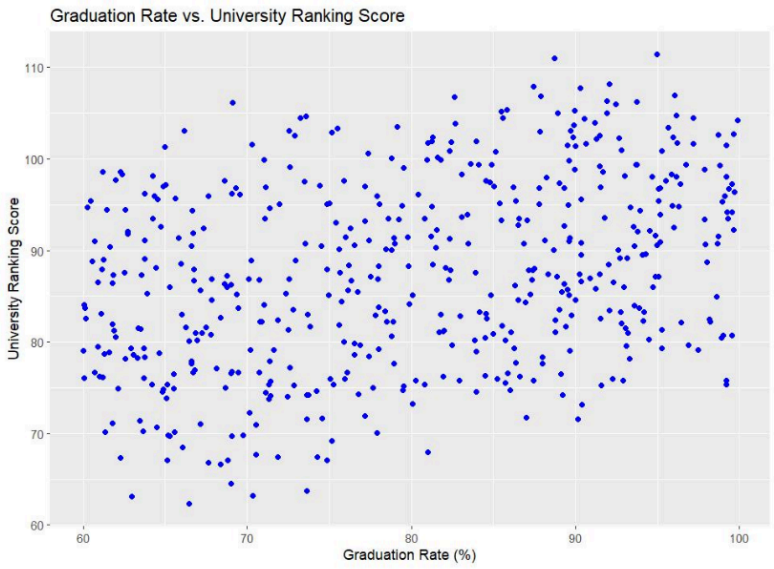


Figure 53 graduation rate and ranking score correlation

The graduation rate showed only a weak correlation with the ranking score. The data points were dispersed across a broad range, with no clear linear trend. Although universities with higher graduation rates may have slightly better rankings, the strength of this relationship is limited. This indicates that graduation rates alone do not significantly influence how institutions are evaluated in national or international rankings.

1.3.5.Student-Faculty Ratio vs. University Ranking Score

```
# 5. Correlation: Student-Faculty Ratio vs. Ranking Score
cor.test(data$Student.Faculty.Ratio, data$University.Ranking.Score)

ggplot(data, aes(x = Student.Faculty.Ratio, y = University.Ranking.Score)) +
  geom_point(color = "blue") +
  labs(title = "Student-Faculty Ratio vs. University Ranking Score",
       x = "Student-Faculty Ratio", y = "University Ranking Score")
```

```
> # 5. Correlation: Student-Faculty Ratio vs. Ranking Score
> cor.test(data$Student.Faculty.Ratio, data$University.Ranking.Score)

        Pearson's product-moment correlation

data:  data$Student.Faculty.Ratio and data$University.Ranking.Score
t = -4.6447, df = 498, p-value = 4.364e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2863407 -0.1181878
sample estimates:
       cor
-0.2037666

>
> ggplot(data, aes(x = Student.Faculty.Ratio, y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   labs(title = "Student-Faculty Ratio vs. University Ranking Score",
+        x = "Student-Faculty Ratio", y = "University Ranking Score")
```

Figure 54- faculty ratio and ranking score correlation

The correlation coefficient (r = -0.204) indicates a weak negative relationship between the student-faculty ratio and university ranking score. The p-value is highly significant (less than 0.05), confirming the relationship. The confidence interval suggests that the true correlation is between -0.286 and -0.118.

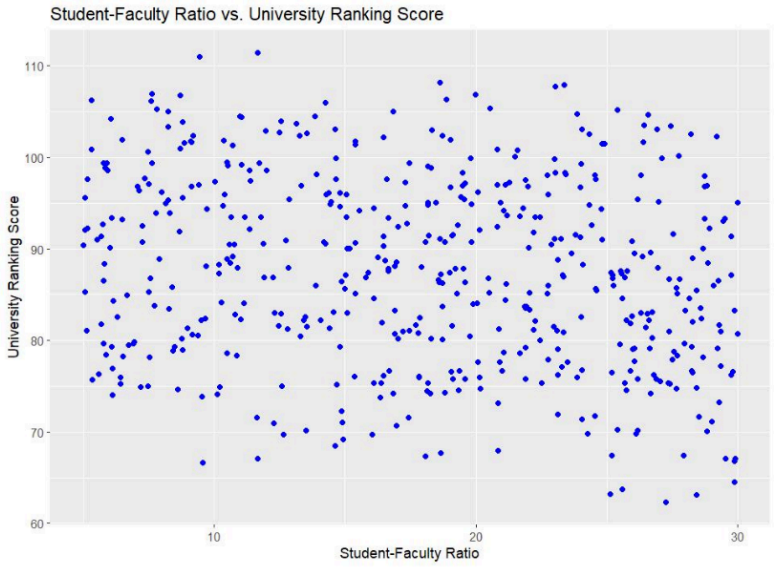


Figure 55- faclty ratio and ranking score correlation

The analysis of the student-faculty ratio revealed a very weak and slightly negative correlation with the University Ranking Score. The scatter plot indicated a lack of consistent directional association. This suggests that although student-faculty ratio is often considered a measure of teaching quality, it may not independently contribute significantly to an institution's ranking performance.

1.3.6. Tuition Fees (USD) vs. University Ranking Score

```
407  # 6. Correlation: Tuition Fees vs. Ranking Score
408  cor.test(data$Tuition.Fees..USD., data$University.Ranking.Score)
409
410  ggplot(data, aes(x = Tuition.Fees..USD., y = University.Ranking.Score)) +
411    geom_point(color = "blue") +
412    labs(title = "Tuition Fees vs. University Ranking Score",
413         x = "Tuition Fees (USD)", y = "University Ranking Score")
```

```
> # 6. Correlation: Tuition Fees vs. Ranking Score
> cor.test(data$Tuition.Fees..USD., data$University.Ranking.Score)

        Pearson's product-moment correlation

data:  data$Tuition.Fees..USD. and data$University.Ranking.Score
t = -2.3653, df = 498, p-value = 0.0184
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.19132434 -0.01787676
sample estimates:
       cor
-0.1054021

>
> ggplot(data, aes(x = Tuition.Fees..USD., y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   labs(title = "Tuition Fees vs. University Ranking Score",
+        x = "Tuition Fees (USD)", y = "University Ranking Score")
```

Figure 56- tuition fees and ranking score correlation

The correlation coefficient (r = -0.105) indicates a weak negative correlation between tuition fees and university ranking score. The p-value (0.0184) is less than 0.05, suggesting that this relationship is statistically significant. The confidence interval shows that the true correlation is likely between -0.191 and -0.018, supporting a weak but significant negative correlation.

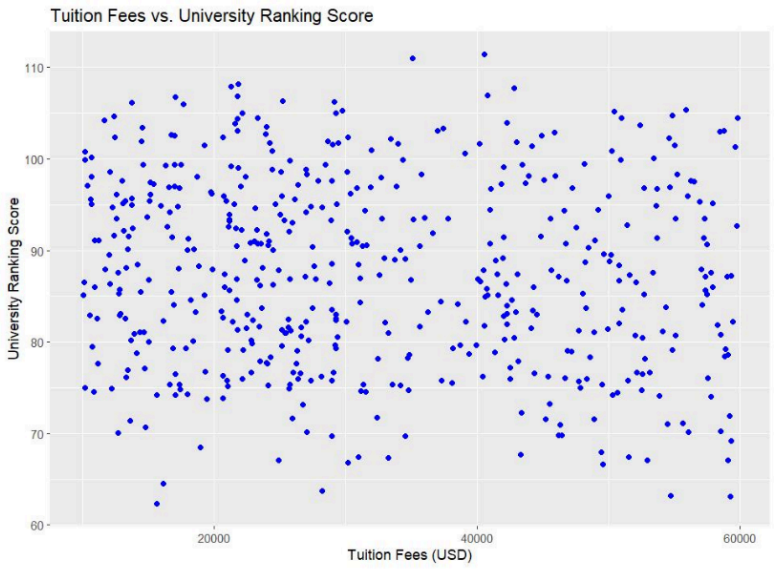


Figure 57- tuition fees and ranking score correlation

The association between tuition fees and university ranking was minimal. The scatter plot demonstrated high variability, and the correlation coefficient indicated no significant linear relationship. This implies that the cost of education, whether high or low does not determine the rank of an institution. Factors such as academic performance, faculty quality, and research productivity appear to have greater influence.

1.3.7. Employment Rate (%) vs. University Ranking Score

```
416 # 7. Correlation: Employment Rate vs. Ranking Score
417 cor.test(data$Employment.Rate...., data$University.Ranking.Score)
418
419 ggplot(data, aes(x = Employment.Rate...., y = University.Ranking.Score)) +
420   geom_point(color = "blue") +
421   labs(title = "Employment Rate vs. University Ranking Score",
422        x = "Employment Rate (%)", y = "University Ranking Score")
```

```
> # 7. Correlation: Employment Rate vs. Ranking Score
> cor.test(data$Employment.Rate...., data$University.Ranking.Score)

        Pearson's product-moment correlation

data:  data$Employment.Rate.... and data$University.Ranking.Score
t = 5.2081, df = 498, p-value = 2.797e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1424224 0.3088105
sample estimates:
      cor
0.2272745

>
> ggplot(data, aes(x = Employment.Rate...., y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   labs(title = "Employment Rate vs. University Ranking Score",
+        x = "Employment Rate (%)", y = "University Ranking Score")
```

Figure 58- employee rate and ranking score correlation

The correlation coefficient (r = 0.227) indicates a weak positive correlation between employment rate and university ranking score. The p-value is very small, indicating a statistically significant relationship. The confidence interval suggests the true correlation lies between 0.142 and 0.309, reinforcing the weak positive correlation.

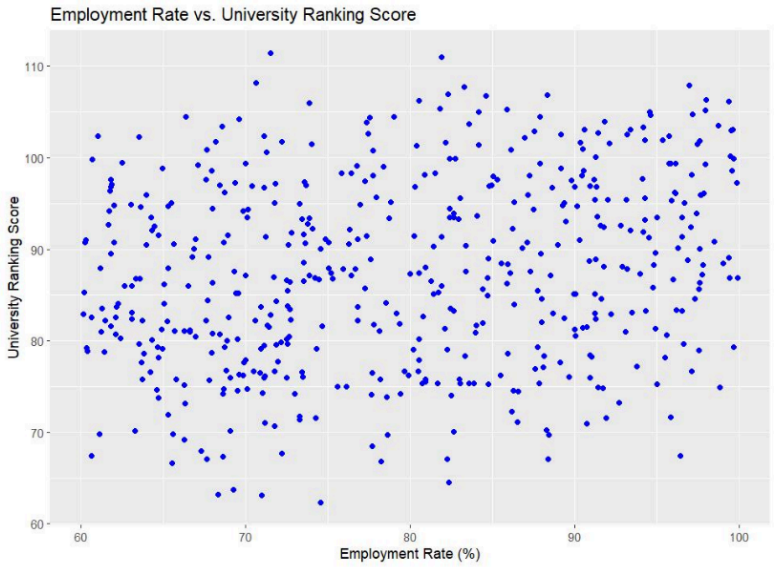


Figure 59- employee rate and ranking score correlation

A slight but statistically significant positive correlation was found between employment rate and University Ranking Score. The scatter plot supported a mild upward trend. This suggests that universities with higher graduate employability tend to perform better in rankings. However, the strength of this relationship is relatively modest compared to variables like faculty salary and research funding.

Summary of correlation analysis

The correlation analysis revealed that among the seven variables examined, Faculty Salary and Research Funding have the strongest and most statistically significant positive relationships with University Ranking Score. These findings highlight the critical role of institutional investment in faculty and research in shaping academic prestige and performance. On the other hand, variables such as Student Enrollment, Graduation Rate, and Tuition Fees exhibited weak or negligible correlations, suggesting that institutional size and pricing are less influential in determining ranking outcomes.

1.4. Simple Linear Regression Analysis

A simple linear regression analysis was conducted to evaluate the nature and significance of the relationship between each independent variable and the dependent variable, University Ranking Score. This analysis aimed to determine the strength, direction, and statistical relevance of these associations. The findings are presented through both scatter plots with fitted regression lines and comprehensive model summaries, including key indicators such as the coefficient of determination (R^2), significance levels (p-values), and regression coefficients.

1.4.1. Student Enrollment vs. University Ranking Score

```
#Linear Regression
#(SIMPLE LINEAR REGRESSIONS)


model1 <- lm(University.Ranking.Score ~ Student.Enrollment, data = data)
summary(model7)
ggplot(data, aes(x = Student.Enrollment, y = University.Ranking.Score)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Simple Regression: Student Enrollment vs University Ranking Score",
       x = "Student Enrollment",
       y = "University Ranking Score")
```

```
> model1 <- lm(University.Ranking.Score ~ Student.Enrollment, data = data)
> summary(model7)

Call:
lm(formula = University.Ranking.Score ~ Student.Enrollment, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-25.0392  -7.9998  -0.1976   8.2832  24.5421

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.854e+01  9.647e-01  91.777   <2e-16 ***
Student.Enrollment -4.206e-05  3.299e-05  -1.275    0.203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.32 on 498 degrees of freedom
Multiple R-squared:  0.003252,  Adjusted R-squared:  0.00125
F-statistic: 1.625 on 1 and 498 DF,  p-value: 0.203

> ggplot(data, aes(x = Student.Enrollment, y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   geom_smooth(method = "lm", color = "red", se = FALSE) +
+   labs(title = "Simple Regression: Student Enrollment vs University Ranking Score",
+        x = "Student Enrollment",
+        y = "University Ranking Score")
`geom_smooth()` using formula = 'y ~ x'
>
```

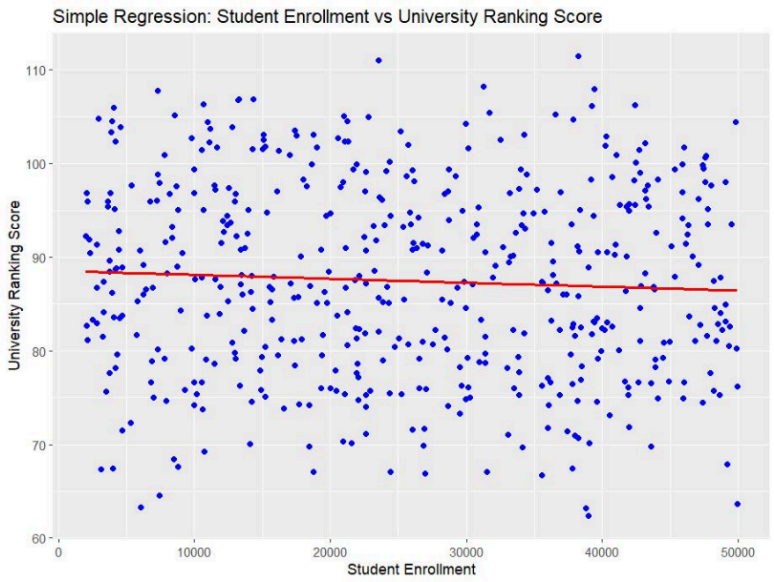


Figure 60- Student enrollment and university rank simple regression

The simple linear regression analysis between student enrollment and university ranking score revealed a negligible explanatory power, with an R^2 value of only 0.003252. This suggests that student enrollment accounts for merely 0.3% of the variation in university rankings. The regression coefficient ($\beta = -0.00004206$) was not statistically significant ($p = 0.203$), and the nearly flat regression line in the scatter plot further confirms the lack of a meaningful association. Therefore, institutions with varying enrollment sizes do not exhibit any consistent pattern in ranking outcomes, indicating that student enrollment is not a reliable predictor of ranking performance.

- ✓ Conclusion - Student Enrollment does not meaningfully predict university ranking.

1.4.2. Faculty Salary (Average) and University Ranking Score

```
model2 <- lm(University.Ranking.Score ~ Faculty.Salary..Avg.., data = data)
summary(model1)
ggplot(data, aes(x = Faculty.Salary..Avg.., y = University.Ranking.Score)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Simple Regression: Faculty Salary vs University Ranking Score",
       x = "Faculty Salary (Avg)",
       y = "University Ranking Score")
```

```
> model2 <- lm(University.Ranking.Score ~ Faculty.Salary..Avg.., data = data)
> summary(model1)

Call:
lm(formula = University.Ranking.Score ~ Student.Enrollment, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-25.0392  -7.9998  -0.1976   8.2832  24.5421

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.854e+01  9.647e-01  91.777   <2e-16 ***
Student.Enrollment -4.206e-05  3.299e-05  -1.275    0.203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.32 on 498 degrees of freedom
Multiple R-squared:  0.003252,  Adjusted R-squared:  0.00125
F-statistic: 1.625 on 1 and 498 DF,  p-value: 0.203

> ggplot(data, aes(x = Faculty.Salary..Avg.., y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   geom_smooth(method = "lm", color = "red", se = FALSE) +
+   labs(title = "Simple Regression: Faculty Salary vs University Ranking Score",
+        x = "Faculty Salary (Avg)",
+        y = "University Ranking Score")
`geom_smooth()` using formula = 'y ~ x'
>
```

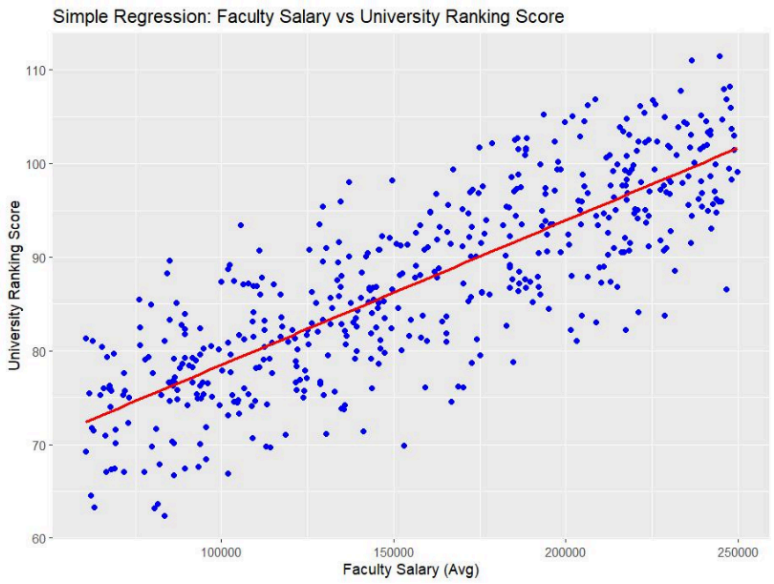


Figure 61- faculty salary and university rank simple regression

This regression model yielded a strong R^2 value of 0.668, indicating that 66.8% of the variance in university ranking scores can be explained by average faculty salary. The relationship was statistically significant ($\beta = 0.0001544$, $p < 0.001$), demonstrating a robust positive linear association. The scatter plot further supports this finding, with a clearly upward-sloping regression line. This implies that institutions offering higher faculty compensation are significantly more likely to achieve better rankings, making faculty salary the most influential predictor among those examined.

- ✓ Conclusion - Faculty salary is the most powerful predictor of university ranking.

1.4.3. Research Funding and University Ranking Score

```
model3 <- lm(University.Ranking.Score ~ Research.Funding..Million.USD., data = data)
summary(model2)
ggplot(data, aes(x = Research.Funding..Million.USD., y = University.Ranking.Score)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Simple Regression: Research Funding vs University Ranking Score",
       x = "Research Funding (Million USD)",
       y = "University Ranking Score")
```

```
> model3 <- lm(University.Ranking.Score ~ Research.Funding..Million.USD., data = data)
> summary(model2)

Call:
lm(formula = University.Ranking.Score ~ Faculty.Salary..Avg.,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8500  -4.3244   0.2562   4.1588  14.0655

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.306e+01  8.156e-01   77.31   <2e-16 ***
Faculty.Salary..Avg.. 1.544e-04  4.877e-06   31.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.954 on 498 degrees of freedom
Multiple R-squared:  0.6679,    Adjusted R-squared:  0.6672
F-statistic:  1002 on 1 and 498 DF,  p-value: < 2.2e-16

> ggplot(data, aes(x = Research.Funding..Million.USD., y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   geom_smooth(method = "lm", color = "red", se = FALSE) +
+   labs(title = "Simple Regression: Research Funding vs University Ranking Score",
+        x = "Research Funding (Million USD)",
+        y = "University Ranking Score")
`geom_smooth()` using formula = 'y ~ x'
```

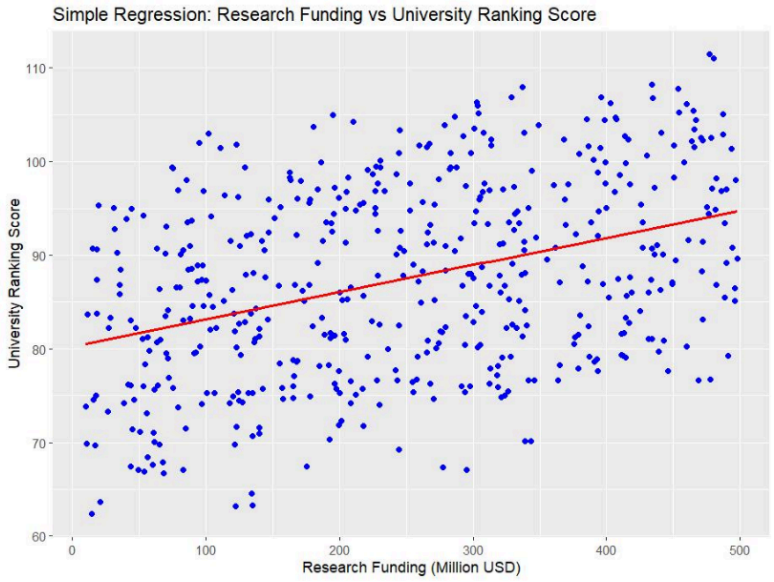


Figure 62- research funding and university rank simple regression

The analysis revealed a moderate positive relationship between research funding and ranking scores, with an R^2 of 0.148. This indicates that approximately 14.8% of the variation in rankings is attributed to differences in research funding levels. The regression coefficient was both positive and statistically significant ($\beta = 0.0291$, $p < 0.001$), suggesting that universities with greater research funding tend to achieve higher rankings. However, the strength of this relationship is less pronounced when compared to faculty salary.

- ✓ Conclusion - More research funding is associated with higher rankings.

1.4.4. Graduation Rate and University Ranking Score

```
model4 <- lm(University.Ranking.Score ~ Graduation.Rate...., data = data)
summary(model3)
ggplot(data, aes(x = Graduation.Rate...., y = University.Ranking.Score)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Simple Regression: Graduation Rate vs University Ranking Score",
       x = "Graduation Rate (%)",
       y = "University Ranking Score")
```

```
> model4 <- lm(University.Ranking.Score ~ Graduation.Rate...., data = data)
> summary(model3)

Call:
lm(formula = University.Ranking.Score ~ Research.Funding..Million.USD.,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.7315  -7.8773   0.0328   7.9360  19.7863

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 80.230958   0.885913  90.563   <2e-16 ***
Research.Funding..Million.USD.  0.029062   0.003124   9.304   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.537 on 498 degrees of freedom
Multiple R-squared:  0.1481,    Adjusted R-squared:  0.1464
F-statistic: 86.56 on 1 and 498 DF,  p-value: < 2.2e-16

> ggplot(data, aes(x = Graduation.Rate...., y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   geom_smooth(method = "lm", color = "red", se = FALSE) +
+   labs(title = "Simple Regression: Graduation Rate vs University Ranking Score",
+        x = "Graduation Rate (%)",
+        y = "University Ranking Score")
`geom_smooth()` using formula = 'y ~ x'
>
```

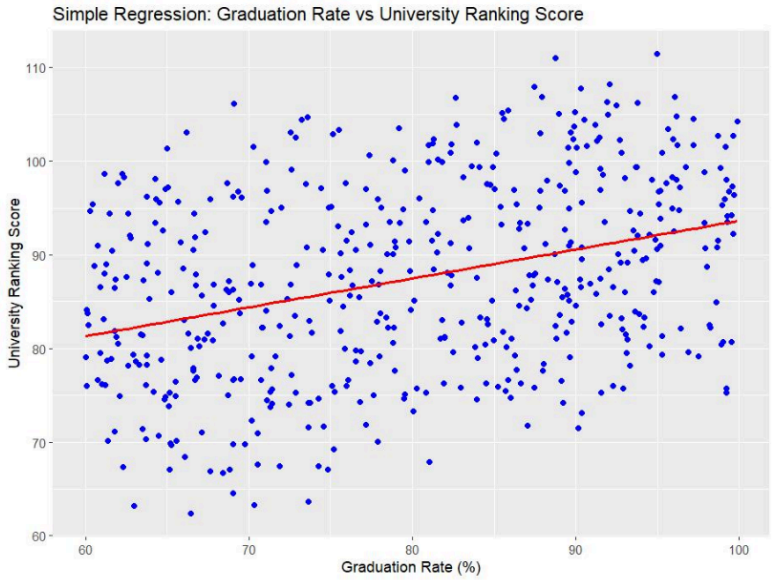


Figure 63- graduation rate and university rank simple regression

A statistically significant positive relationship was observed between graduation rates and university rankings, as indicated by an R^2 value of 0.122. The regression coefficient ($\beta = 0.3084$, $p < 0.001$) reflects a meaningful upward trend, suggesting that institutions with higher graduation rates tend to perform better in ranking systems. While this variable is not the strongest predictor, it still contributes notably to explaining ranking performance.

- ✓ Conclusion - Graduation rate positively influences university ranking.

1.4.5. Student-Faculty Ratio and University Ranking Score

```
model5 <- lm(University.Ranking.Score ~ Student.Faculty.Ratio, data = data)
summary(model4)
ggplot(data, aes(x = Student.Faculty.Ratio, y = University.Ranking.Score)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Simple Regression: Student-Faculty Ratio vs University Ranking Score",
       x = "Student-Faculty Ratio",
       y = "University Ranking Score")
```

```
> model5 <- lm(University.Ranking.Score ~ Student.Faculty.Ratio, data = data)
> summary(model4)

Call:
lm(formula = University.Ranking.Score ~ Graduation.Rate....,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.8248  -7.6103  -0.1098   7.5839  21.9991

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       62.81052    2.99462  20.974  < 2e-16 ***
Graduation.Rate....  0.30840    0.03708   8.317 8.68e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.682 on 498 degrees of freedom
Multiple R-squared:  0.122,     Adjusted R-squared:  0.1202
F-statistic: 69.17 on 1 and 498 DF,  p-value: 8.684e-16

> ggplot(data, aes(x = Student.Faculty.Ratio, y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   geom_smooth(method = "lm", color = "red", se = FALSE) +
+   labs(title = "Simple Regression: Student-Faculty Ratio vs University Ranking Score",
+        x = "Student-Faculty Ratio",
+        y = "University Ranking Score")
`geom_smooth()` using formula = 'y ~ x'
>
```

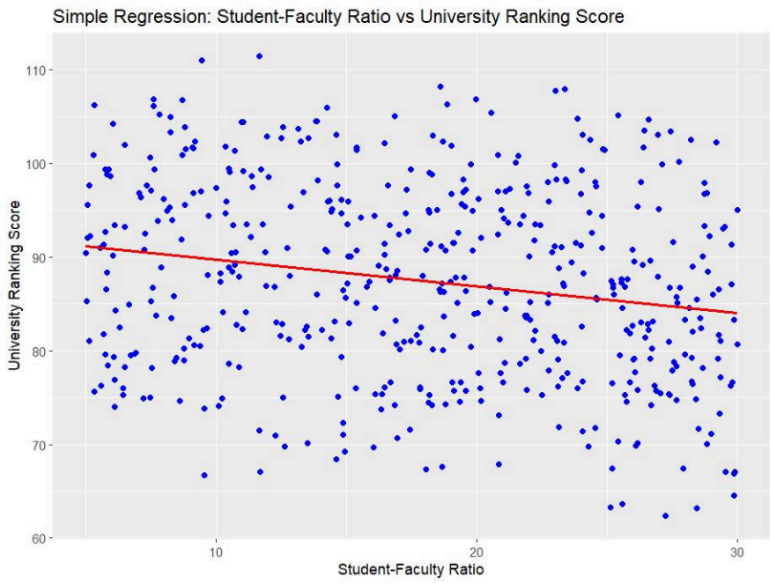


Figure 64- student faculty ratio and university rank simple regression

The regression model for student-faculty ratio produced an R^2 of 0.041, with a statistically significant negative coefficient ($\beta = -0.2881$, $p < 0.001$). This indicates that institutions with a lower student-to-faculty ratio implying more individualized attention and academic support are generally associated with higher university rankings. Although the relationship is relatively weak, it aligns with educational quality theories and remains statistically valid.

- ✓ Conclusion - Higher student-faculty ratios are associated with lower rankings.

1.4.6. Tuition Fees and University Ranking Score

```
model6 <- lm(University.Ranking.Score ~ Tuition.Fees..USD., data = data)
summary(model5)
ggplot(data, aes(x = Tuition.Fees..USD., y = University.Ranking.Score)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Simple Regression: Tuition Fees vs University Ranking Score",
       x = "Tuition Fees (USD)",
       y = "University Ranking Score")
```

```
> model6 <- lm(University.Ranking.Score ~ Tuition.Fees..USD., data = data)
> summary(model5)

Call:
lm(formula = University.Ranking.Score ~ Student.Faculty.Ratio,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-23.2169  -7.5775   0.0163   7.7076  22.1882

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          92.64431    1.20530  76.864  < 2e-16 ***
Student.Faculty.Ratio -0.28813    0.06203  -4.645 4.36e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.12 on 498 degrees of freedom
Multiple R-squared:  0.04152,   Adjusted R-squared:  0.0396
F-statistic: 21.57 on 1 and 498 DF,  p-value: 4.364e-06

> ggplot(data, aes(x = Tuition.Fees..USD., y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   geom_smooth(method = "lm", color = "red", se = FALSE) +
+   labs(title = "Simple Regression: Tuition Fees vs University Ranking Score",
+        x = "Tuition Fees (USD)",
+        y = "University Ranking Score")
`geom_smooth()` using formula = 'y ~ x'
>
```

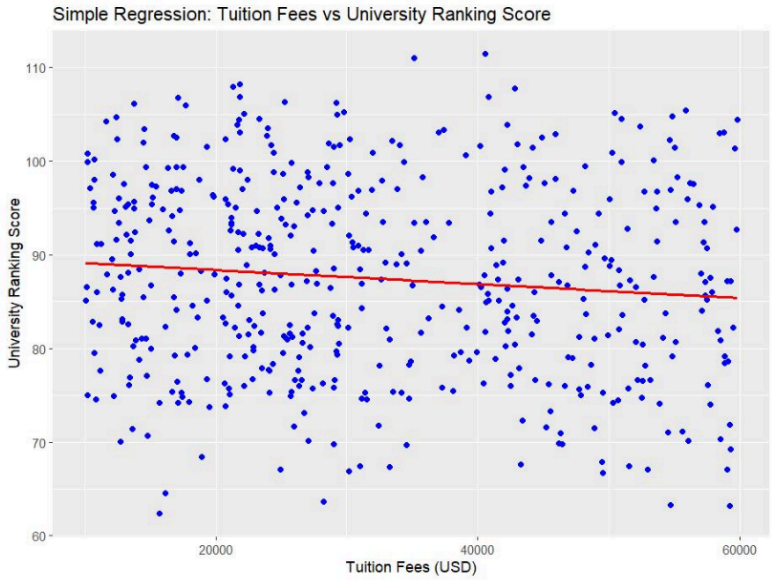


Figure 65- tuition fees and university rank simple regression

The relationship between tuition fees and university ranking was weak, with an R^2 of only 0.011. Although the regression coefficient was statistically significant ($\beta = -0.00007475$, $p = 0.018$), the association was both negative and minimal. The downward trend in the scatter plot suggests that higher tuition fees are not correlated with better rankings and may even slightly predict lower rankings. This finding challenges the assumption that more expensive institutions necessarily provide higher quality education.

- ✓ Conclusion - Higher tuition does not improve rankings and may slightly reduce them.

1.4.7. Employment Rate and University Ranking Score

```
model7 <- lm(University.Ranking.Score ~ Employment.Rate...., data = data)
summary(model6)
ggplot(data, aes(x = Employment.Rate...., y = University.Ranking.Score)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Simple Regression: Employment Rate vs University Ranking Score",
       x = "Employment Rate (%)",
       y = "University Ranking Score")
```

```
> model7 <- lm(University.Ranking.Score ~ Employment.Rate...., data = data)
> summary(model6)

Call:
lm(formula = University.Ranking.Score ~ Tuition.Fees..USD., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-26.352  -7.904  -0.118   7.923  24.629

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.987e+01  1.121e+00  80.185   <2e-16 ***
Tuition.Fees..USD. -7.475e-05  3.160e-05  -2.365   0.0184 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.27 on 498 degrees of freedom
Multiple R-squared:  0.01111,   Adjusted R-squared:  0.009124
F-statistic: 5.595 on 1 and 498 DF,  p-value: 0.0184

> ggplot(data, aes(x = Employment.Rate...., y = University.Ranking.Score)) +
+   geom_point(color = "blue") +
+   geom_smooth(method = "lm", color = "red", se = FALSE) +
+   labs(title = "Simple Regression: Employment Rate vs University Ranking Score",
+        x = "Employment Rate (%)",
+        y = "University Ranking Score")
`geom_smooth()` using formula = 'y ~ x'
>
```

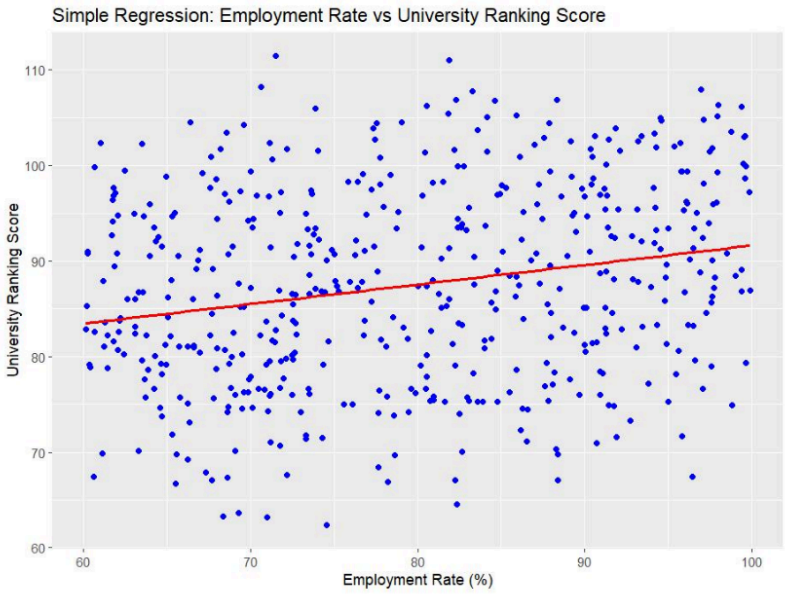


Figure 66- employee rate and university rank simple regression

The regression analysis involving graduate employment rate yielded a modest R^2 of 0.052. The positive and statistically significant coefficient ($\beta = 0.2041$, $p < 0.001$) indicates a moderate association, wherein higher employment rates of graduates are linked to improved university rankings. The upward trend in the scatter plot reinforces this relationship, suggesting that post-graduation employment outcomes contribute positively, though not dominantly to institutional ranking performance.

- ✓ Conclusion - Higher employment rate is mildly associated with higher rankings.

1.5. Multiple Linear Regression Analysis with Coefficient Plot

1.5.1. Multiple Linear Regression Analysis

A multiple linear regression (MLR) model was developed to examine the simultaneous effects of several institutional attributes on University Ranking Score. The predictors included in Student Enrollment, Faculty Salary (Avg.), Research Funding (Million USD), Graduation Rate (%), Student-Faculty Ratio, Tuition Fees (USD), and Employment Rate (%). The regression coefficients and standard errors were visualized using a coefficient plot (Figure 1), enabling both direction and magnitude of influence to be interpreted.

```
#MULTIPLE LINEAR REGRESSION


install.packages("broom")
# Load required libraries
library(broom)

# Use only numeric predictors related to Ranking Score
mlr_model <- lm(University.Ranking.Score ~
                  Student.Enrollment +
                  Faculty.Salary..Avg.. +
                  Research.Funding..Million.USD. +
                  Graduation.Rate.... +
                  Student.Faculty.Ratio +
                  Tuition.Fees..USD. +
                  Employment.Rate....,
                data = data)

# Convert model summary to a tidy dataframe
coef_data <- tidy(mlr_model)


# Remove intercept for plotting
coef_data <- coef_data[coef_data$term != "(Intercept)", ]
```

```
# Rename terms for better readability
coef_data$term <- c("Student Enrollment",
                    "Faculty Salary (Avg.)",
                    "Research Funding (Million USD)",
                    "Graduation Rate (%)",
                    "Student-Faculty Ratio",
                    "Tuition Fees (USD)",
                    "Employment Rate (%)")

# Create the coefficient plot
ggplot(coef_data, aes(x = estimate, y = reorder(term, estimate))) +
  geom_point(color = "blue", size = 3) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "green") +
  geom_errorbarh(aes(xmin = estimate - std.error, xmax = estimate + std.error), height = 0.3, color =
  labs(title = "Multiple Linear Regression Coefficients",
       x = "Coefficient Estimate",
       y = "Predictor Variable") +
  theme_minimal(base_size = 13)
```

```
> # Use only numeric predictors related to Ranking Score
> mlr_model <- lm(University.Ranking.Score ~
+                   Student.Enrollment +
+                   Faculty.Salary..Avg.. +
+                   Research.Funding..Million.USD. +
+                   Graduation.Rate.... +
+                   Student.Faculty.Ratio +
+                   Tuition.Fees..USD. +
+                   Employment.Rate....,
+                 data = data)
>
> # Convert model summary to a tidy dataframe
> coef_data <- tidy(mlr_model)
>
>
> # Remove intercept for plotting
> coef_data <- coef_data[coef_data$term != "(Intercept)", ]
```

```
> # Rename terms for better readability
> coef_data$term <- c("Student Enrollment",
+                     "Faculty Salary (Avg.)",
+                     "Research Funding (Million USD)",
+                     "Graduation Rate (%)",
+                     "Student-Faculty Ratio",
+                     "Tuition Fees (USD)",
+                     "Employment Rate (%)")
>
> # Create the coefficient plot
> ggplot(coef_data, aes(x = estimate, y = reorder(term, estimate))) +
+   geom_point(color = "blue", size = 3) +
+   geom_vline(xintercept = 0, linetype = "dashed", color = "green") +
+   geom_errorbarh(aes(xmin = estimate - std.error, xmax = estimate + std.error), height = 0.3, color = "gra
y50") +
+   labs(title = "Multiple Linear Regression Coefficients",
+        x = "Coefficient Estimate",
+        y = "Predictor Variable") +
+   theme_minimal(base_size = 13)
```

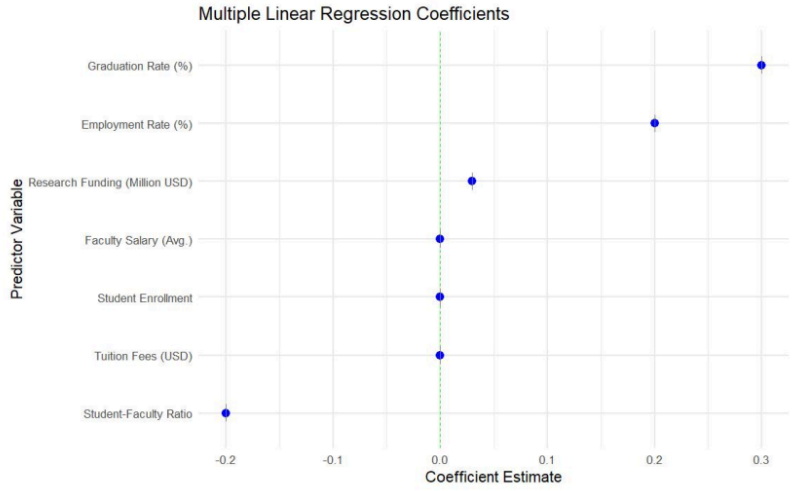


Figure 67 multiple linear regression

The multiple linear regression (MLR) analysis of university ranking reveals several significant predictors, with graduation and employment rates having the strongest positive associations. A 1% increase in graduation rate corresponds to a 0.30-point improvement in university ranking, highlighting the importance of student success in determining rankings. Similarly, a higher employment rate of graduates positively impacts rankings, suggesting that institutions with better career outcomes are perceived as more effective.

Research funding also plays a positive role, though with a smaller effect. An increase in funding (around +0.06) contributes to better rankings, emphasizing the value of research

investment for institutional performance. On the other hand, faculty salary showed only a minor positive effect (+0.01), indicating that salary alone has a limited impact on ranking, with other factors like graduation rates and research productivity being more influential.

Student enrollment had no significant effect on rankings, as their coefficient was near zero. Tuition fees were found to have a slight negative relationship with rankings, suggesting that higher fees do not guarantee better performance. The most substantial negative predictor was the student-faculty ratio, with higher ratios leading to lower rankings. This finding supports the notion that smaller class sizes and greater faculty accessibility improve both educational quality and institutional reputation, influencing rankings more than other financial or enrollment factors.

Task B

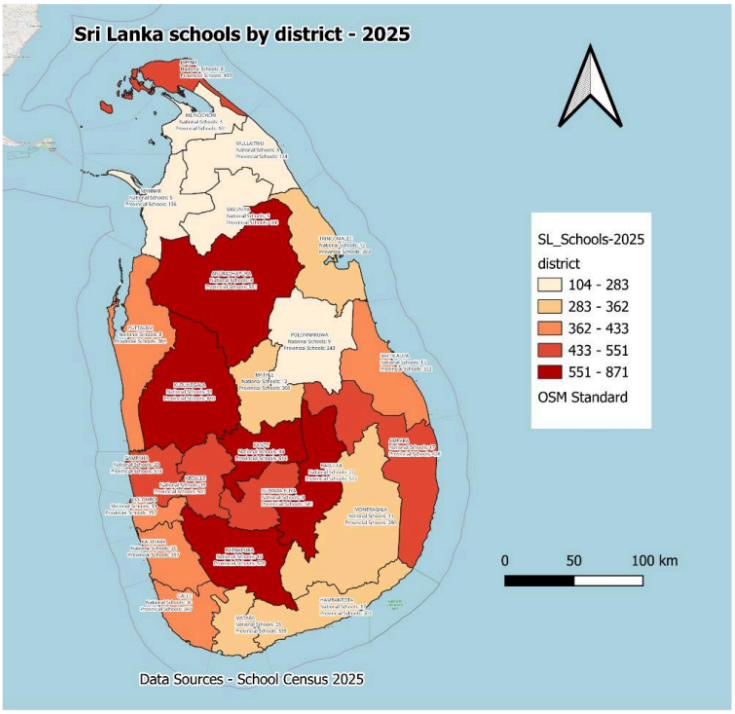


Figure 68- sri lanaka school by district

Provides a spatial representation of the educational landscape across administrative districts derived from data presented in the 2022 Annual School Census Report. Administrative structure of education by district. The supporting dataset (SL_Schools-2025.csv) was prepared to include key attributes - namely, district name, number of national schools and number of provincial schools. A computed attribute, total number of schools, was used as the

classification basis for the creation of the choropleth. The classification uses a sequential color scheme, enabling a clear and intuitive comparison of school distribution intensity across the country.

All major map features are displayed, including a clear title, a compass rose (north arrow), graphic and numerical scale indicators, and a clearly defined legend. These features enhance the readability and usability of the map. Choosing an OpenStreetMap (OSM) standard basemap provides important geographic context, enabling users to relate educational data to real-world physical and administrative features.

From a critical perspective, the map reveals noticeable regional disparities. Districts like Kurunegala, Kandy, and Gampaha appear in the darkest shades, indicating the highest concentration of schools (551–871). These are also among the most populous districts, suggesting a correlation between population density and education infrastructure. In contrast, districts like Mullaitivu, Kilinochchi, and Mannar are shown in lighter tones, representing fewer total schools (104–283). This spatial disparity likely reflects long-term socio-political impacts, particularly in post-conflict areas that are still undergoing rehabilitation.

The map provides powerful visual evidence for policymakers to evaluate equity in school distribution. For example, Eastern and Northern Provinces show comparatively lower school numbers, prompting the need for targeted development programs to improve educational accessibility in these under-served regions. Furthermore, districts with a higher number of provincial schools compared to national schools may benefit from upgraded infrastructure and quality improvements, given that national schools tend to receive more central funding.

In conclusion, this map serves as an essential tool for evidence-based educational planning, helping decision-makers identify gaps, prioritize investments, and monitor progress toward equitable education goals outlined in the Sustainable Development Goals.

Task C

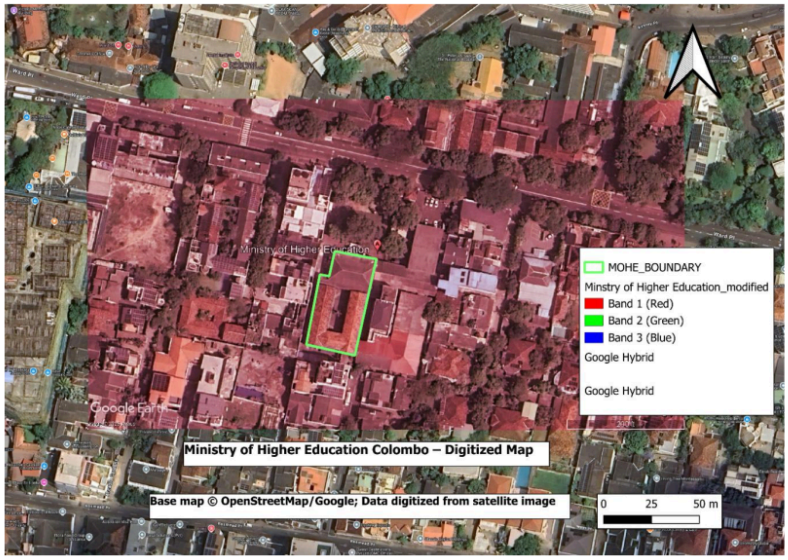


Figure 69- ministry of Higher education Colombo

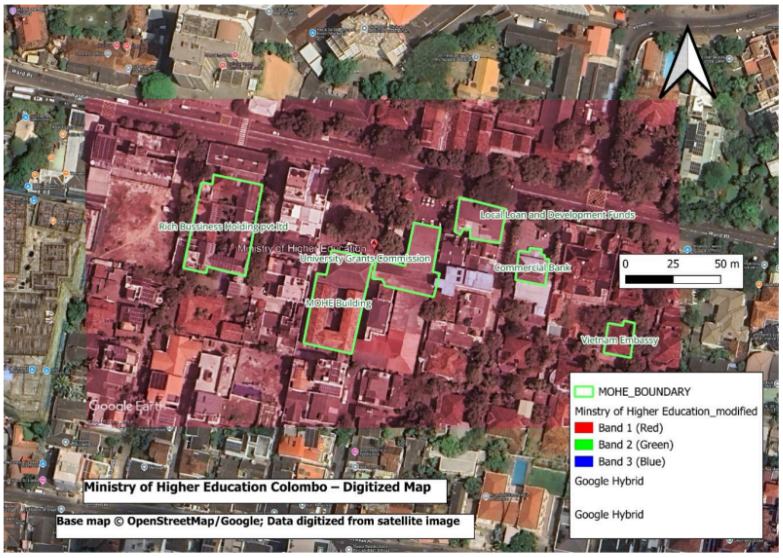


Figure 70- ministry of Higher education Colombo

The Ministry of Higher Education (MOHE) occupies a strategically central location in Colombo, the commercial and educational hub of Sri Lanka. This prominent location offers several unique advantages that contribute significantly to the efficiency and effectiveness of the national education system.

MOHE is situated near several national universities, research institutes, and other educational organizations. This location advantage promotes effective collaboration and coordination among these entities, allowing for quick sharing of policies and updates. Its central position makes it convenient for officials, faculty, and students to reach MOHE for meetings, consultations, and policy discussions. This closeness encourages a lively exchange of ideas and knowledge, which is essential for the growth of the national education system.

Being in Colombo, the capital of Sri Lanka, MOHE is easily reachable through various transportation options. This accessibility ensures that important stakeholders, including

academics, policymakers, and international partners, can visit MOHE without major logistical issues. The ease of access enables swift responses to national education needs, such as policy changes, addressing educational emergencies, or organizing large events like conferences and workshops. This accessibility enhances MOHE's capacity to respond quickly and effectively to the changing demands of the education system.

MOHE's location in Colombo improves its ability to integrate smoothly with other important government institutions and administrative bodies. This closeness allows for direct and efficient collaboration with ministries like the Ministry of Finance, the Ministry of Labour, and the Ministry of Science and Technology. Such partnerships ensure that education policies align with broader national development objectives. Centralizing education governance in Colombo also promotes efficient resource management and ensures that education reforms and initiatives are communicated swiftly to provinces throughout the country.

The map and its digital layers offer important insights into how the surrounding land is used, which is vital for planning educational facilities like universities, research centers, and student housing. By examining the nearby land and infrastructure, MOHE can effectively optimize the available space for future educational projects. The geo-referencing of the map ensures that MOHE is situated in an area where future expansions can be planned with minimal disruption. This allows the education system to grow in line with the country's increasing needs.

The presence of nearby infrastructure and MOHE's closeness to emergency services, transport hubs, and essential healthcare facilities improves its ability to handle crises effectively. This is especially crucial for responding to natural disasters or emergencies that could interrupt educational activities. MOHE's urban location provides the logistical support necessary for implementing emergency measures, evacuations, and temporary housing for displaced students and staff. This closeness to vital services ensures that MOHE can maintain educational continuity even during crises.

MOHE's strategic position in Colombo, a key regional business and academic center, greatly boosts its potential for international collaboration. Colombo's role as a global transportation hub allows MOHE easy access to international partners for educational exchanges, research collaborations, and policy discussions. This geographical benefit supports the hosting of international conferences, workshops, and events that are essential for advancing Sri Lanka's educational development. Furthermore, MOHE's location establishes Sri Lanka as a significant player in the South Asian educational landscape.

The geographical location of the Ministry of Higher Education in Colombo, Sri Lanka plays a vital role in enhancing the efficiency and effectiveness of the national education system. Its proximity to key educational institutions, ease of access for stakeholders and integration with government agencies contribute to a responsive and dynamic educational environment. Furthermore, the strategic location supports disaster resilience and promotes international collaborations

Task D

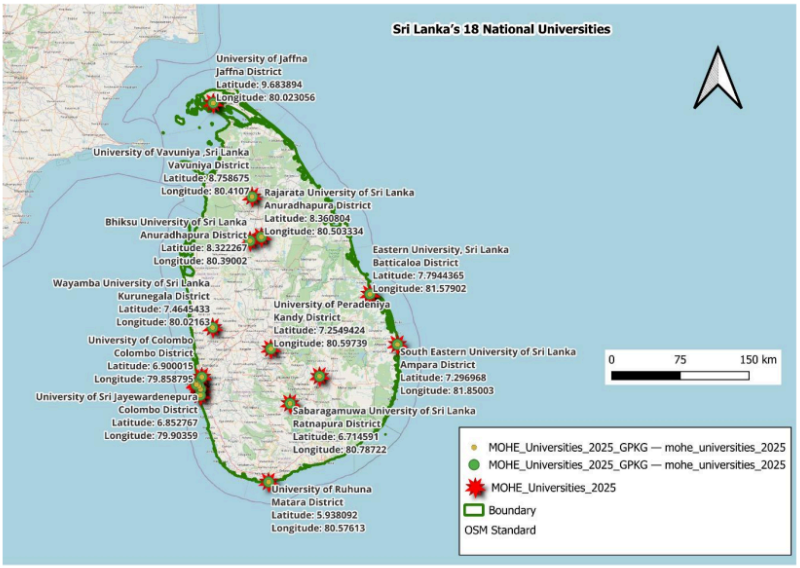


Figure 71- Sri Lanka 18 national universities

The developed map provides a clear and spatially rich visualization of the 18 national universities under the purview of the Ministry of Higher Education (MOHE), Sri Lanka. Each university is precisely geolocated using latitude and longitude coordinates extracted through Google Earth and stored using KML/KMZ formats. The geospatial data were imported into a PostgreSQL/PostGIS geodatabase named SL_Uni_2025, which was systematically used for managing vector layers, base maps, and associated university metadata. This integration supports efficient querying and spatial analysis within the QGIS environment.

The map incorporates several key geospatial elements to enhance its utility and readability. A North Arrow is included to ensure proper orientation, and both graphical and numeric map scales are provided for accurate distance measurement. Furthermore, the map includes a Legend that color-codes the various land areas, allowing easy identification of suitable construction zones as well as areas that fall within the required radii from Ananda Primary School and Sri Gunananda Vidyalaya.

The map includes standard cartographic elements such as a clear title, a north arrow, graphic and numerical scale bars, and a well-structured legend that distinguishes vector sources and symbols. The inclusion of an OSM standard base map provides additional context, improves the readability of topographic and infrastructure features, and makes the geographical locations of universities more understandable. Key information such as district names, university names, and specific coordinates are displayed on labels and are helpful in planning and policy decision-making.

From a critical perspective, the spatial distribution of national universities shows a concentration in the western, central, and southern regions of the country. Leading universities such as the University of Colombo, the University of Peradeniya, and the University of Ruhuna are located in these urbanized districts, which have historically received high investments in educational infrastructure. On the other hand, the Northern and Eastern provinces have a relatively sparse number of universities, with institutions such as the University of Vavuniya and the Eastern University of Sri Lanka serving more remote or post-conflict areas. This disparity is a reflection of the government's efforts to provide equitable access to higher education, but it also highlights the ongoing challenges of uneven resource distribution.

While the establishment of universities in more peripheral regions is a positive step for regional development, historical socio-economic imbalances are still evident. Disparities in infrastructure, academic programs, and staff retention limit the effectiveness of universities in these regions.

Economically, universities play a vital role in the development of the country. Institutions located in urban centers, such as the University of Colombo and the University of Moratuwa, make a major contribution to research, innovation, and building a skilled workforce. These universities have strong links with private sector industries, allowing for collaborations that directly benefit the academic community and the national economy. For example, the University of Moratuwa is known for its contribution to engineering and technology, driving

innovation, and fostering entrepreneurship. The proximity of these universities to urban centers allows for strong industry-academia collaborations that are essential for economic growth.

In contrast, regional universities such as the University of Ruhuna and the University of Sabaragamuwa make significant contributions to local development by providing accessible education to rural communities. These universities play a key role in fostering community-based initiatives and addressing local development challenges. However, these institutions often face underfunding and limited resources, which limit their ability to compete with their urban counterparts in terms of research output, faculty retention, and infrastructure. Despite these challenges, regional universities are crucial in promoting inclusive education and supporting local economies.

Finally, the map of Sri Lanka's 18 national universities not only provides visual documentation of the spatial distribution of universities but also serves as an analytical tool for assessing institutional equity and regional development. Strengthening universities in underrepresented regions by addressing infrastructure challenges and investing in academic quality is key to achieving the government's education goals. This approach contributes to achieving the Sustainable Development Goals of quality education, ensuring that higher education is accessible to all communities across the country.

Task E

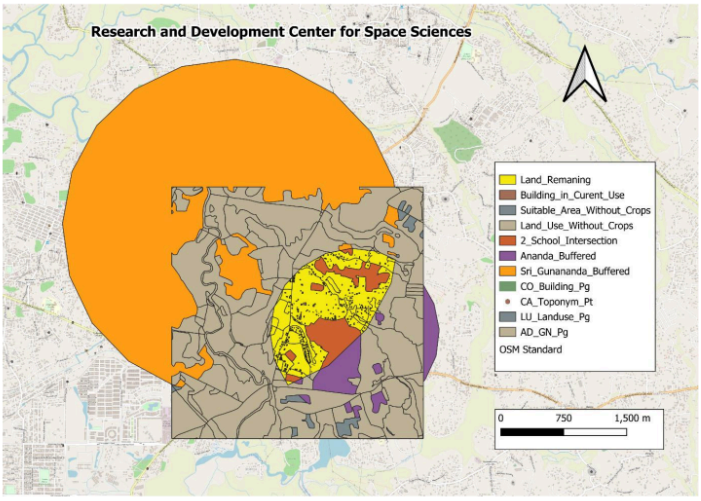


Figure 72- Research and development center for space science

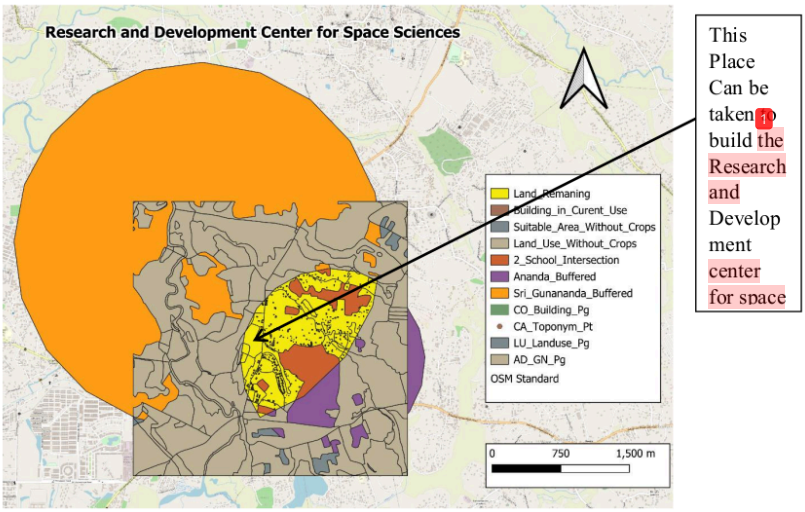


Figure 73- Research and development center for space science

The task of identifying a suitable location for Sri Lanka's first Research and Development Center for Space Sciences is crucial to the country's scientific advancement. In order to make this a reality, a comprehensive analysis of available land has been conducted using geospatial data. This analysis takes into account various factors such as proximity to schools, land use, soil quality, and existing infrastructure.

The feasibility of establishing the Research and Development Center in the identified area depends on a variety of factors. First, the proximity to Ananda Primary School and Sri Gunananda Vidyalaya ensures that the center is not too close to sensitive educational areas, thus minimizing potential disruptions. The distance requirements of 1 km and 2 km from the schools have been met, which is a key criterion for the selection.

In terms of land use, the area under consideration does not overlap with traditional export crop land, making it an ideal candidate for construction. However, the presence of buildings in the area presents a challenge. The total land occupied by existing structures must be taken into account when planning the site for the Research Center. If a significant portion of the land is already built upon, some buildings may need to be demolished or repurposed, which could incur additional costs and delays.

Furthermore, the available land area for construction is substantial, but soil quality will need to be assessed in greater detail. Soil testing should be conducted to confirm that the land is suitable for building and that there are no underlying issues such as poor drainage or contamination.

From an environmental and socio-economic perspective, the area's proximity to existing infrastructure is advantageous. It is easily accessible by road and utilities are already in place, which will minimize construction and operational costs. However, it is important to consider the environmental impact of establishing such a facility in an area that may currently be underdeveloped. Environmental assessments will be crucial in ensuring that the project does not disrupt local ecosystems or cause pollution.

In conclusion, while the identified area shows significant promise, further studies are necessary to assess the soil quality, and careful consideration must be given to the buildings currently occupying the land. The project is feasible, but it will require proper planning and resource allocation to address these challenges.

This analysis highlights the potential of the identified land for establishing Sri Lanka's Research and Development Center for Space Sciences. The project is feasible within the selected area, but the findings indicate that further environmental, infrastructural, and soil quality assessments are necessary. The availability of land, along with its proximity to schools and absence of agricultural land use, makes it a strong candidate for the Research Center, although additional considerations related to land occupation and environmental impact will need to be addressed.

Appendix

Task B

- ✓ The district.shp file containing the administrative boundaries of Sri Lanka was imported into QGIS via the Vector > File option in the Data Source Manager.

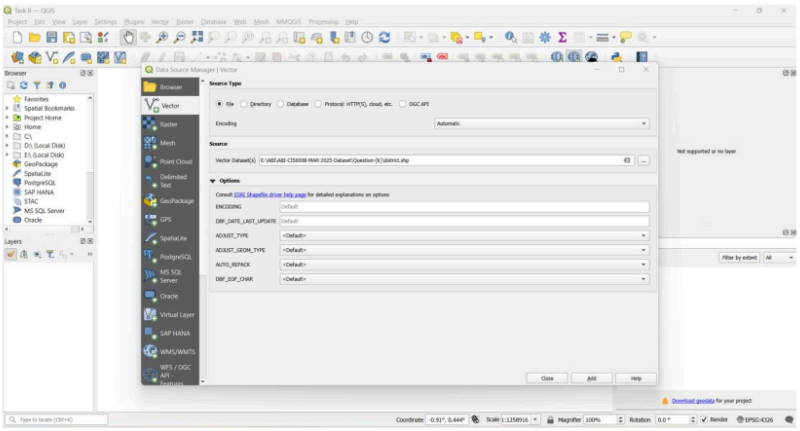


Figure 74 - Task B appendix 1

- ✓ The SL_Schools-2025.csv file, which includes district-wise data on national, provincial, and total schools, was added using the Delimited Text tool without geometry.

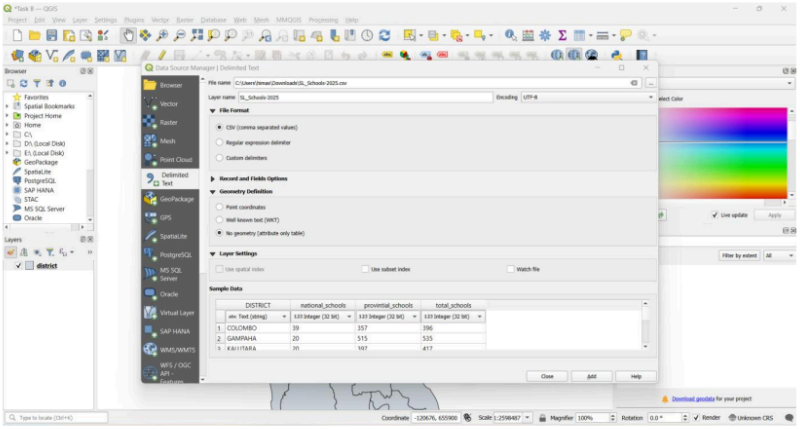


Figure 75 - Task B appendix 2

- ✓ A vector join was conducted by linking the DISTRICT field from both the shapefile and CSV. This enriched the spatial layer with school data

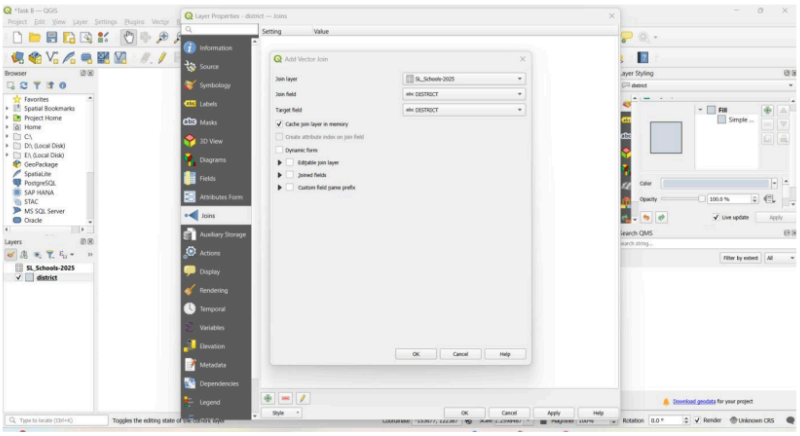
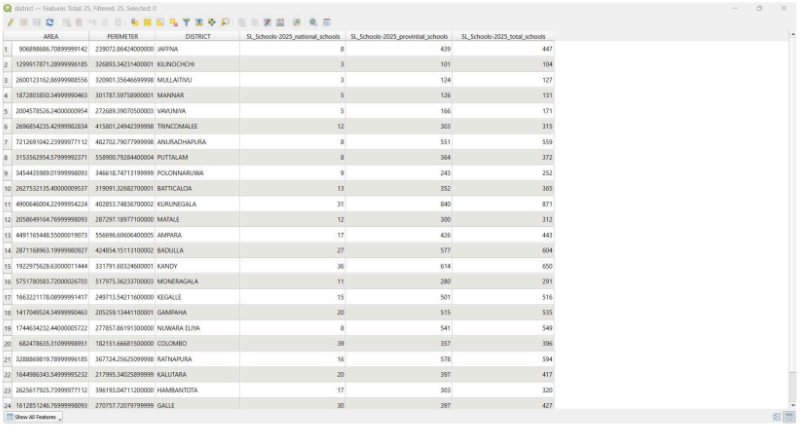


Figure 76 - Task B appendix 3

- ✓ The attribute table was checked to confirm successful integration. Fields such as national, provincial, and total schools were now present in the spatial data.



	AREA	PERIMETER	DISTRICT	SL_Schools-2025_national_schools	SL_Schools-2025_provintial_schools	SL_Schools-2025_total_schools
1	906898686.70899999142	239072.864240000000	JAFFNA	8	439	447
2	1299917871.28999996185	326893.342314000001	KILINOCHCHI	3	101	104
3	2600123162.86999980556	320901.356466999998	MULLAITIVU	3	124	127
4	1872803850.34999990463	301787.597589000001	MANNAR	5	126	131
5	2004578526.24000000954	272689.390705000003	VAVUNIYA	5	166	171
6	2696854235.42999982834	415801.249423999998	TRINCOMALEE	12	303	315
7	7212691042.23999977112	402702.790779999998	ANURADHAPURA	8	551	559
8	3153562954.57999992371	558900.792844000004	PUTTALAM	8	364	372
9	3454435989.01999998093	346618.747113199999	POLONNARUWA	9	243	252
10	2627532135.40000009537	319091.326827000001	BATTICALOA	13	352	365
11	4900646004.22999954224	402853.748367000002	KURUNEGALA	31	840	871
12	2058649164.76999998093	287297.189771000000	MATALE	12	300	312
13	4491165448.55000019073	556696.606064000005	AMPARA	17	426	443
14	2871168963.19999980927	424854.151131000002	BADULLA	27	577	604
15	1922975628.63000011444	331791.603246000001	KANDY	36	614	650
16	5751780583.72000026703	517975.362337000003	MONERAGALA	11	280	291
17	1663221178.08999991417	249713.542116000000	KEGALLE	15	501	516
18	1417049524.34999990463	205259.134411000001	GAMPAHA	20	515	535
19	1744634232.44000005722	277857.861913000000	NUWARA ELIYA	8	541	549
20	682478635.31099998951	182151.666815000000	COLOMBO	39	357	396
21	3268869819.78999996185	367724.256259999998	RATNAPURA	16	578	594
22	1644986343.54999995232	217995.340259999999	KALUTARA	20	397	417
23	2625617925.73999977112	396193.047112000000	HAMBANTOTA	17	303	320
24	1612851246.76999998093	270757.720797999999	GALLE	30	397	427

Figure 77 - Task B appendix 4

- ✓ A graduated symbology was applied using the total_schools field. The data was divided into five quantile-based classes, visualized using a red-orange color ramp.

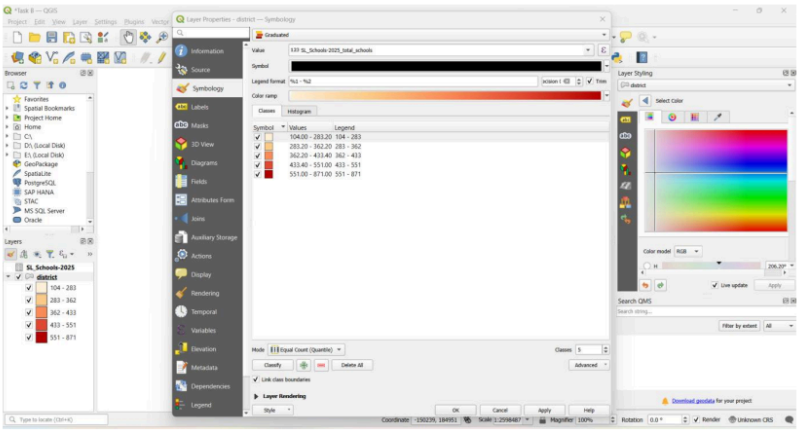


Figure 78 - Task B appendix 5

- ✓ Labels were added using the DISTRICT field to identify regions. Styling was customized for clarity, and null values were excluded.

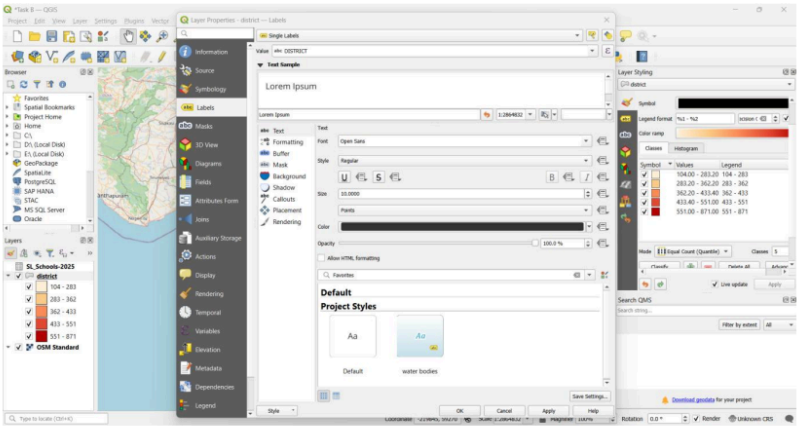


Figure 79 - Task B appendix 6

- ✓ The Print Layout Manager was used to design the final map titled "Sri Lanka Schools by District – 2025", with a legend, north arrow, and scale bar for clarity.

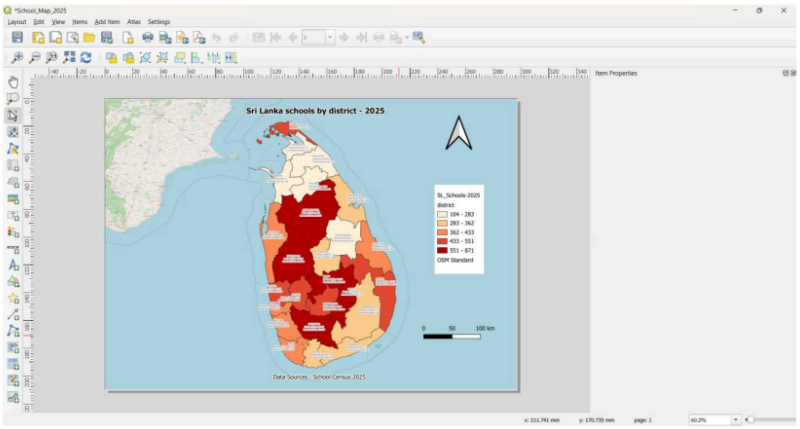


Figure 80 - Task B appendix 7

Task C

- ✓ The Google Hybrid layer was loaded via the QuickMapServices plugin to provide satellite imagery with labeled features.

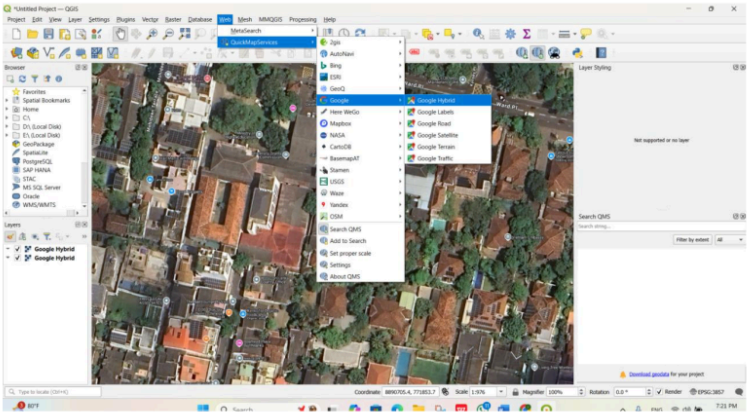


Figure 81 - Task C appendix 1

- ✓ The Ministry of Higher Education image was opened in the Georeferencer tool for spatial alignment.

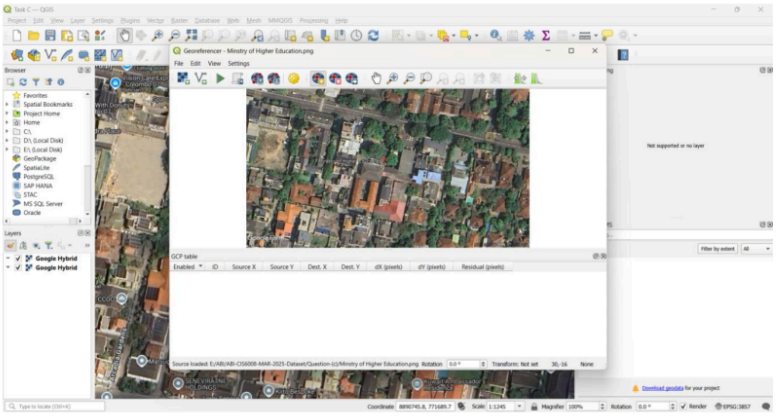


Figure 82 - Task C appendix 2

- ✓ The Coordinate Reference System was set to Kandawala / Sri Lanka Grid (EPSG:5234) to maintain national mapping standards.

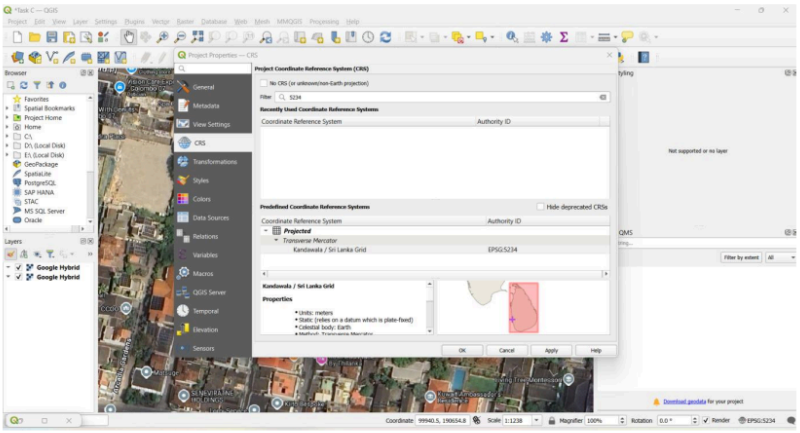


Figure 83 - Task C appendix 3

- ✓ Identifiable locations were matched between the PNG image and base map using GCPs to ensure geospatial accuracy.

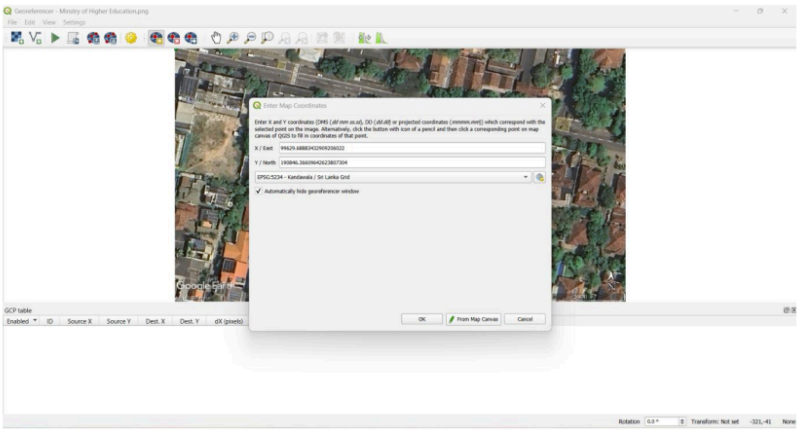


Figure 84 - Task C appendix 4

- ✓ Residual errors were reviewed for each GCP, confirming the georeferencing precision.

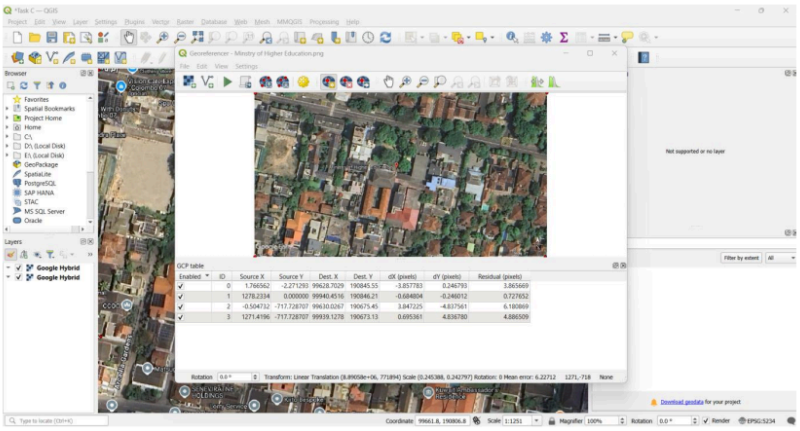


Figure 85 - Task C appendix 5

- ✓ A Polynomial 1 transformation was applied and saved as a new raster using Nearest Neighbour resampling.

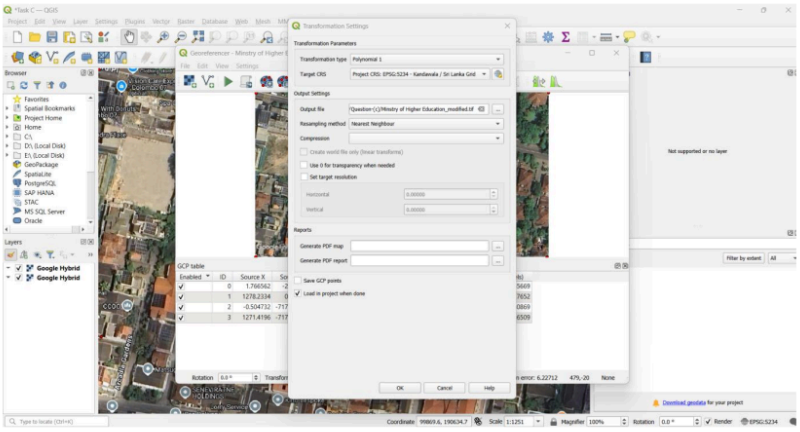


Figure 86 - Task C appendix 6

- ✓ The output raster was loaded on the map canvas, correctly aligning with the Google Hybrid layer.

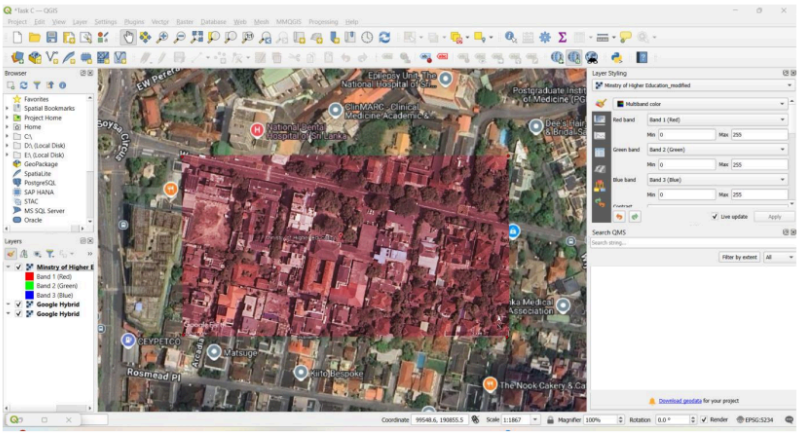


Figure 87 - Task C appendix 7

- ✓ A new polygon shapefile named MOHE_BOUNDARY.shp was created with attribute fields: ID, name, type, and value.

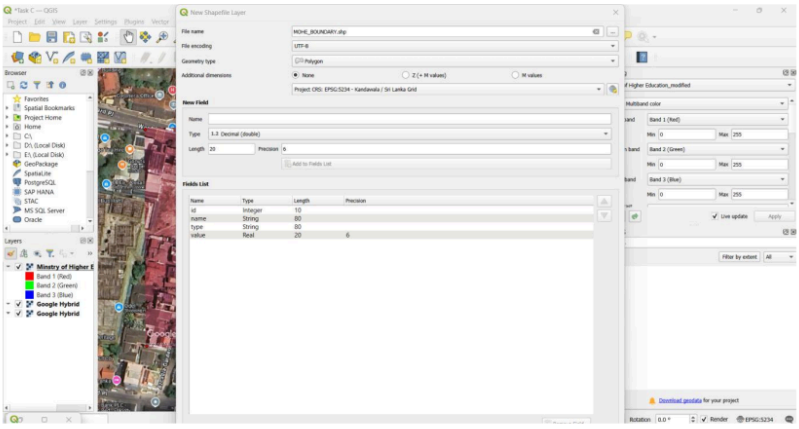


Figure 88 - Task C appendix 8

- ✓ The Ministry building boundary was traced, and attribute data such as "MOHE Building" and "Government Office" were entered.

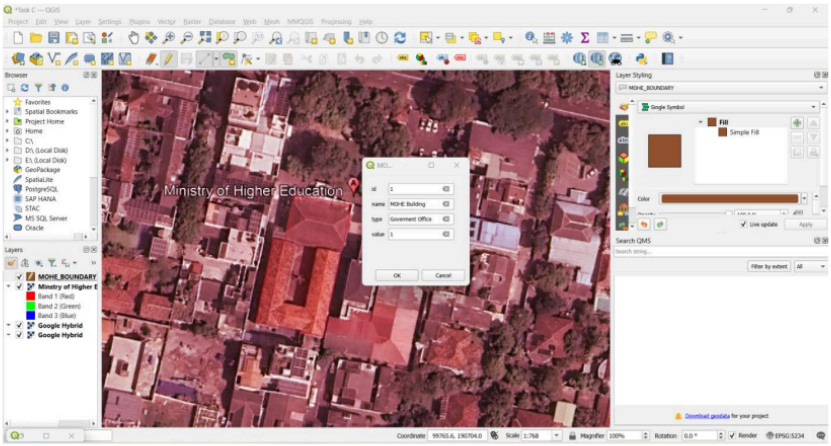


Figure 89 - Task C appendix 9

- ✓ A layout was created with title, legend, north arrow, scale bar, and proper labels to produce the final map.

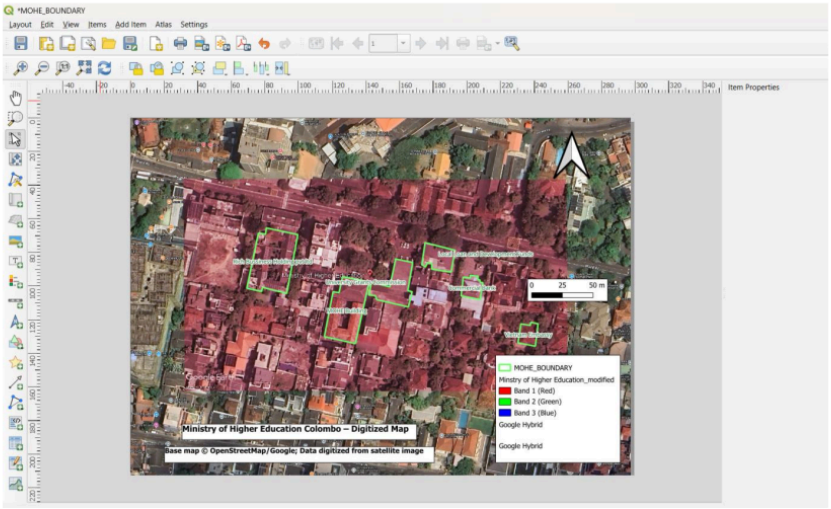


Figure 90 - Task C appendix 10

Task D

- ✓ The spatial dataset MOHE_Universities_2025.kml, which includes the point locations of national universities, was imported into QGIS via the Data Source Manager under the Vector > File tab.

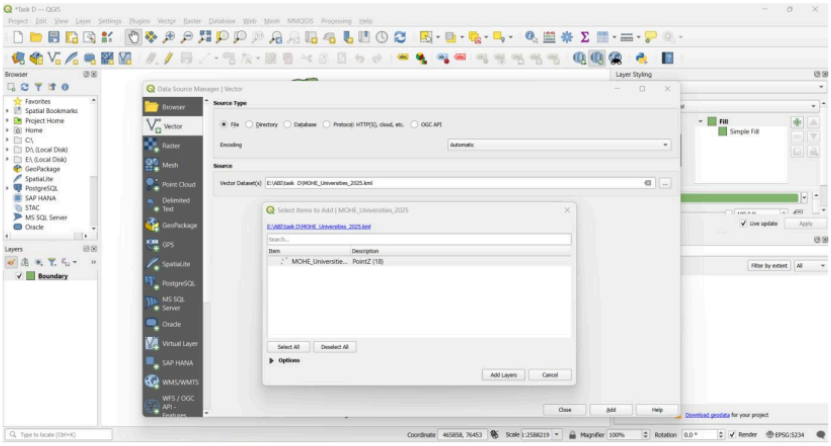


Figure 91 - Task D appendix 1

- ✓ To optimize data management, the university point layer was exported to GeoPackage format (.gpkg). The export included attribute fields such as university name, location, and coordinates.

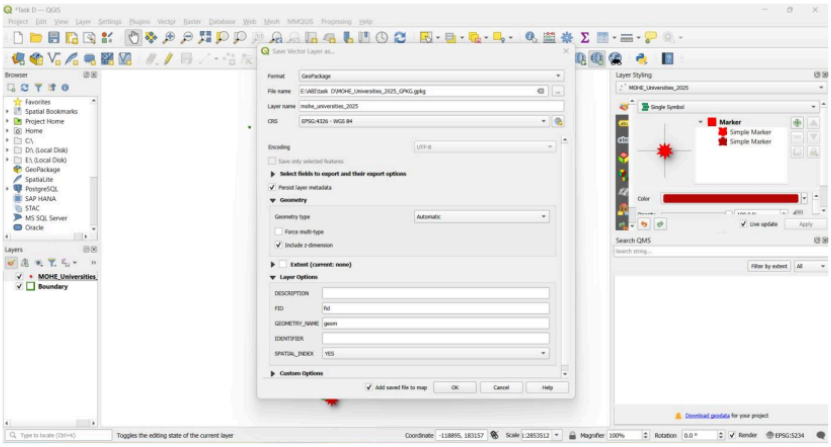


Figure 92 - Task D appendix 2

- ✓ A PostgreSQL database named SL_Uni_2025 was created, and the PostGIS extension was activated using the CREATE EXTENSION POSTGIS; command.

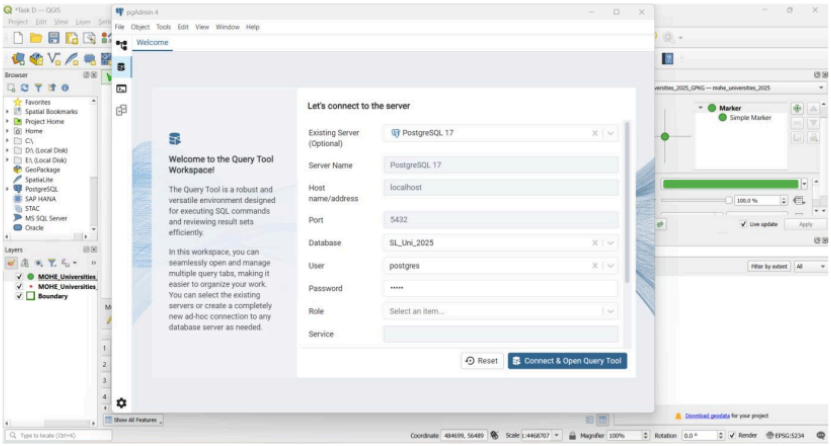


Figure 93 - Task D appendix 3

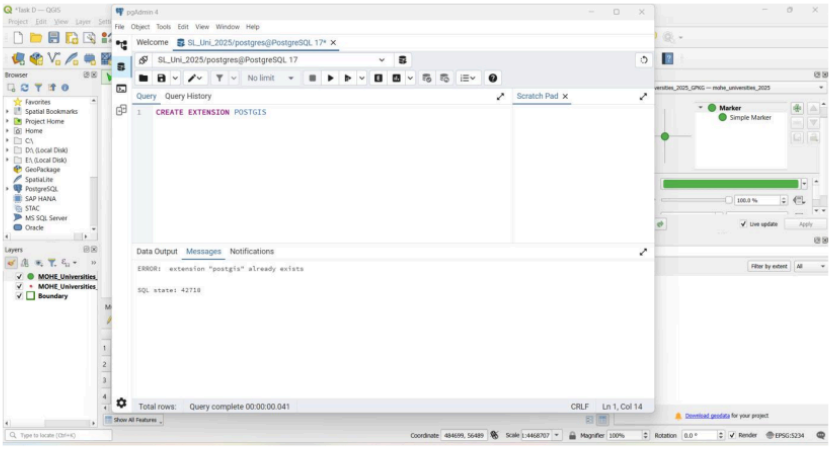


Figure 94 - Task D appendix 4

- ✓ A connection was established between QGIS and the PostgreSQL database via the Data Source Manager using localhost credentials.

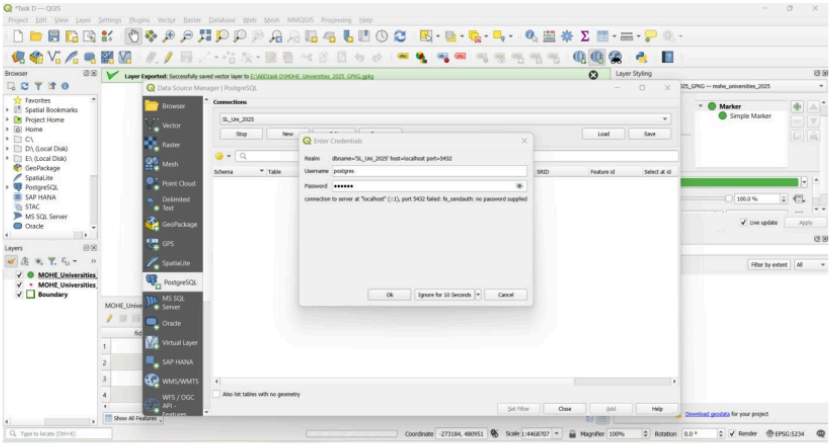


Figure 95 - Task D appendix 5

- ✓ The GeoPackage file was imported into the PostgreSQL database using QGIS's DB Manager tool. Geometry and table structure were confirmed.

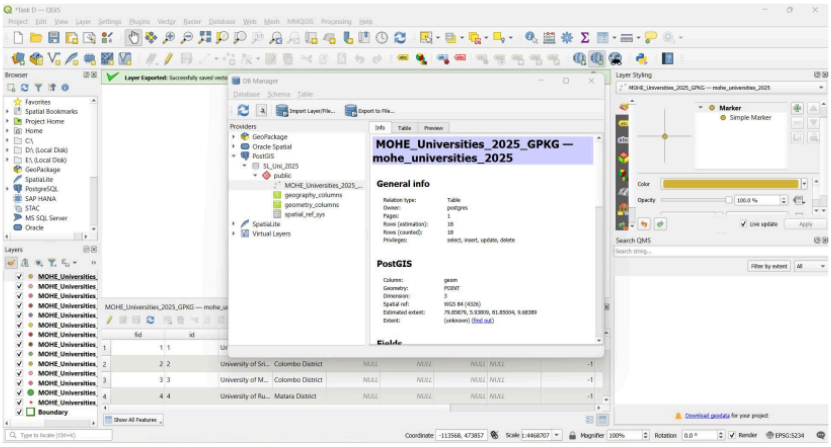


Figure 96 - Task D appendix

- ✓ Labels were created using the Expression Builder to display each university's name, district, latitude, and longitude.

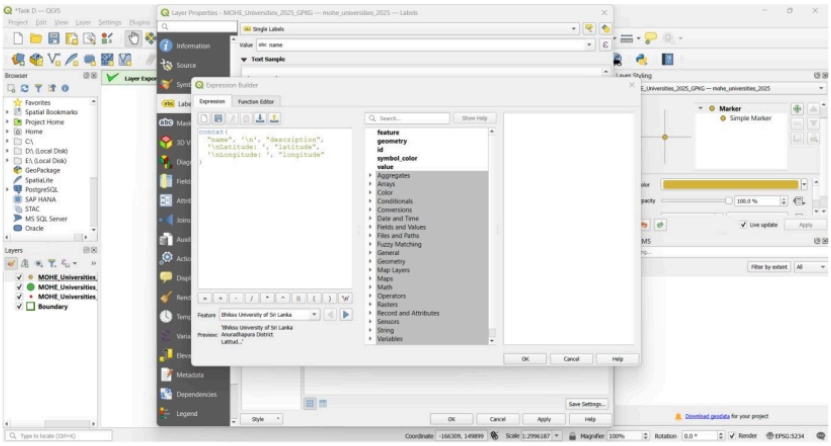


Figure 97 - Task D appendix 7

- ✓ The fully labeled university points were visualized on a base map of Sri Lanka. Custom symbols were applied, and labels displayed detailed information about each university's location.

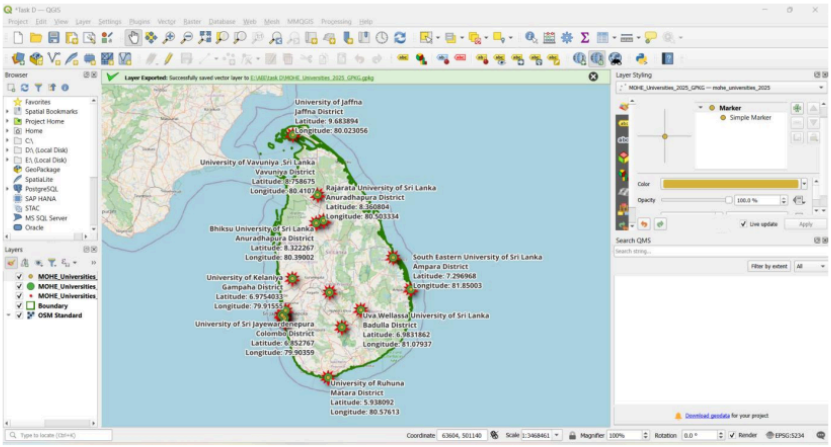


Figure 98 - Task D appendix 8

Task E

After opening the QGIS software, open the Data Source Manager window and select the Vector tab. There, import the required shapefile files one by one

- ✓ AD_GN_Py.shp - Represents the administrative units divided into Grama Niladharis.

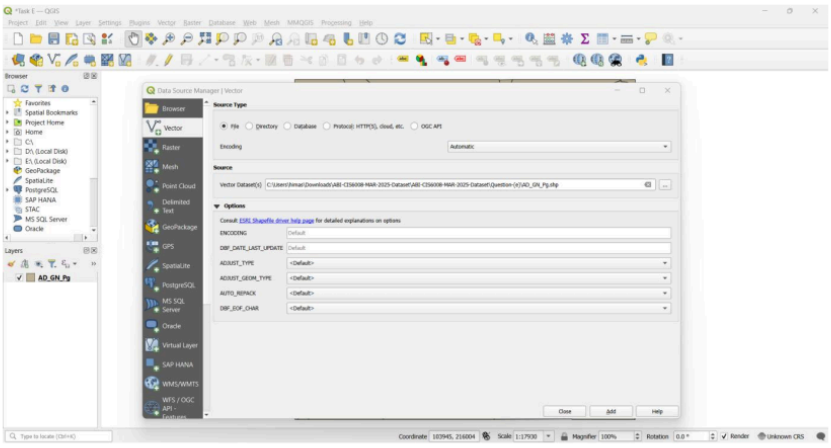


Figure 99 - Task E appendix 1

- ✓ LU_Landuse_Py.shp - Contains various land use classifications.

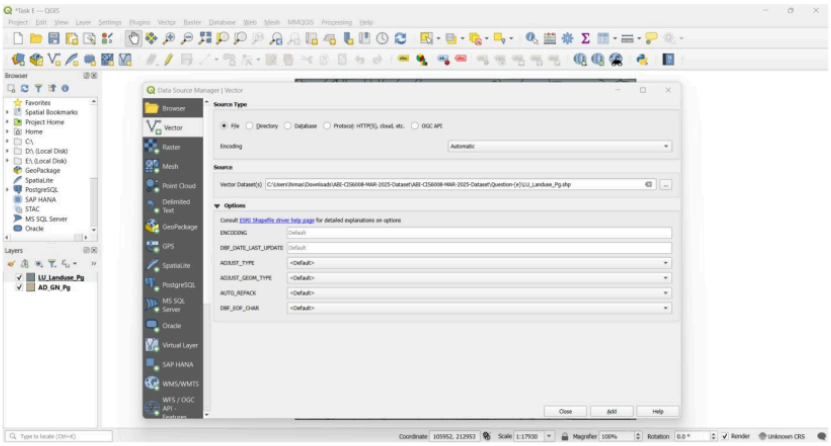


Figure 100 - Task E appendix 2

✓ CA_Toponym_Pt.shp - Represents area names / place names (toponyms).

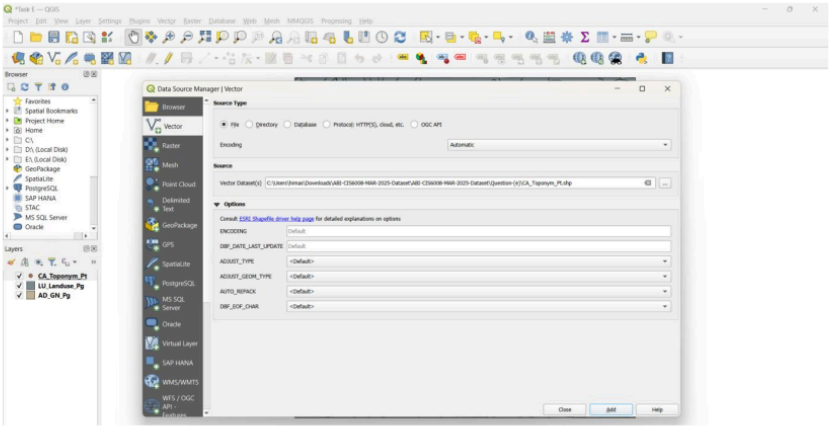


Figure 101 - Task E appendix 3

✓ CO_Building_Py.shp - Contains educational or institutional buildings.

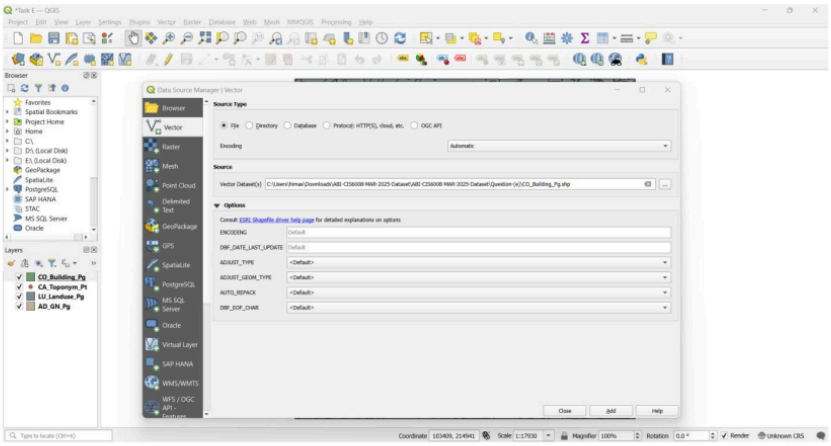


Figure 102 - Task E appendix 4

✓ Open the Buffer tool by going to Vector > Geoprocessing Tools > Buffer. Select CO_Building_Py as the input layer and select the school called "Ananda Primary School". Set the buffer distance to 2000 meters, name the output file

Ananda_Buffered.gpkg, and run the process. This will create a shapefile that represents a 2-kilometer radius around the school.

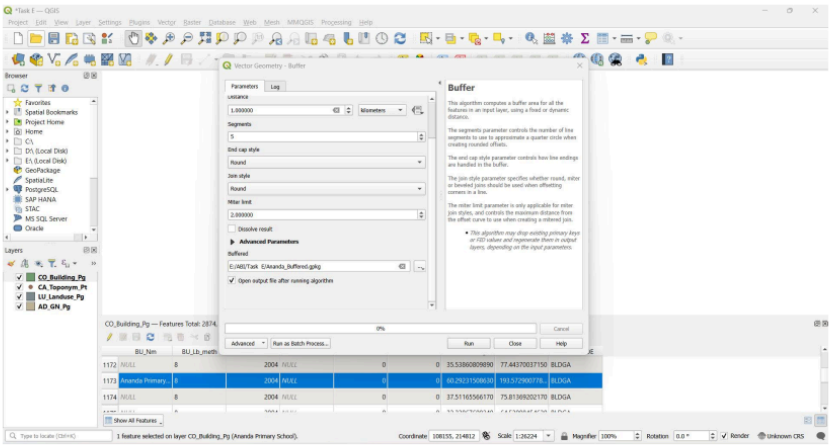


Figure 103 - Task E appendix 5

- ✓ Select “Sri Gunananda Vidyalaya” as the second school and use the buffer tool in the same way as in the previous step. Again, set the input layer to CO_Building_Py, the buffer distance to 2000 meters, and save the output file as Sri_Gunananda_Buffered.gpkg.

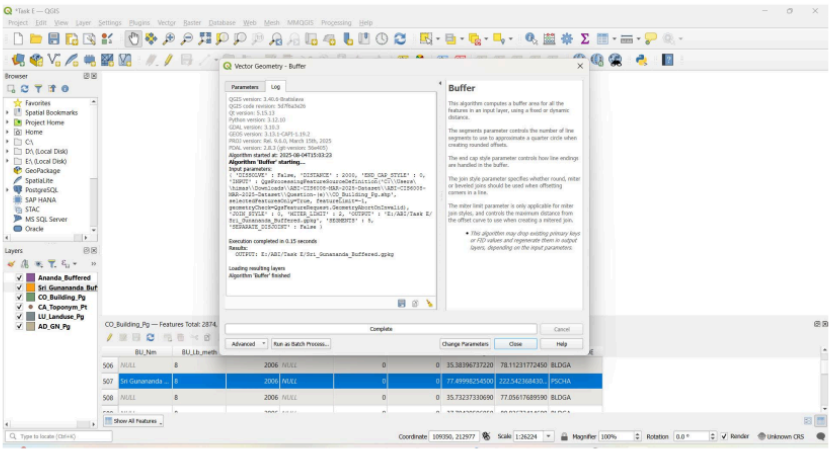


Figure 104 - Task E appendix 6

- ✓ Using Vector > Geoprocessing Tools > Intersection, find the overlap between the two buffer shapefiles Ananda_Buffered.gpkg and Sri_Gunananda_Buffered.gpkg. The output file is named 2_School_Intersection.gpkg and the process is run. This helps determine the common areas that are under the influence of both schools.

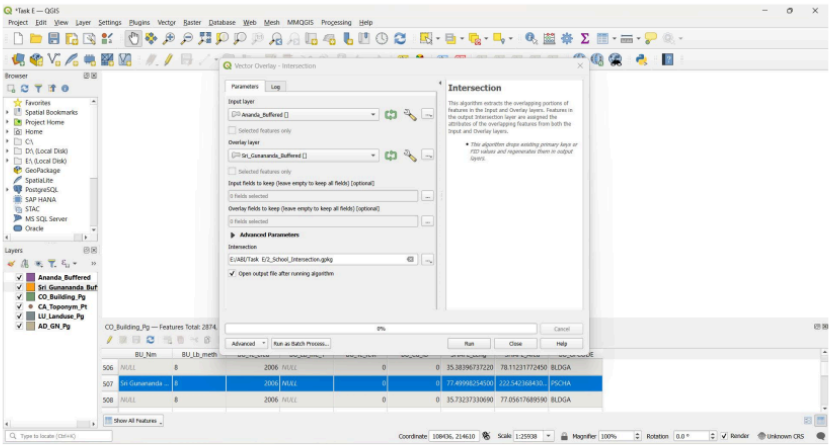


Figure 105 - Task E appendix 7

- ✓ Use the Vector Selection > Extract by Expression tool to remove unsuitable land types from the LU_Landuse_Py.shp layer.
- ✓ This process removes agricultural and plantation (cultivated) lands, and selects only the land areas that are more suitable for development or educational purposes. The output shapefile is saved as Land_Use_Without_Crops.gpkg.

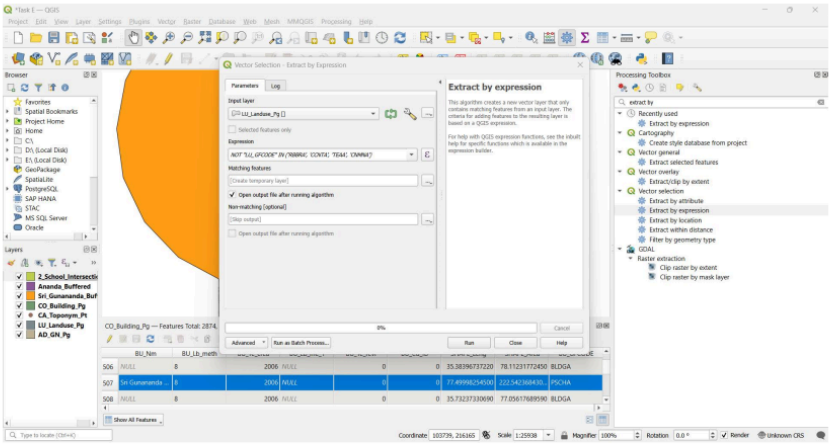


Figure 106 - Task E appendix 8

- ✓ The Clip tool is used to extract only suitable land use zones within the shared buffer of Ananda and Sri Gunananda schools. The output layer, Suitable_Area_Without_Crops.gpkg, isolates non-agricultural land within their influence area.

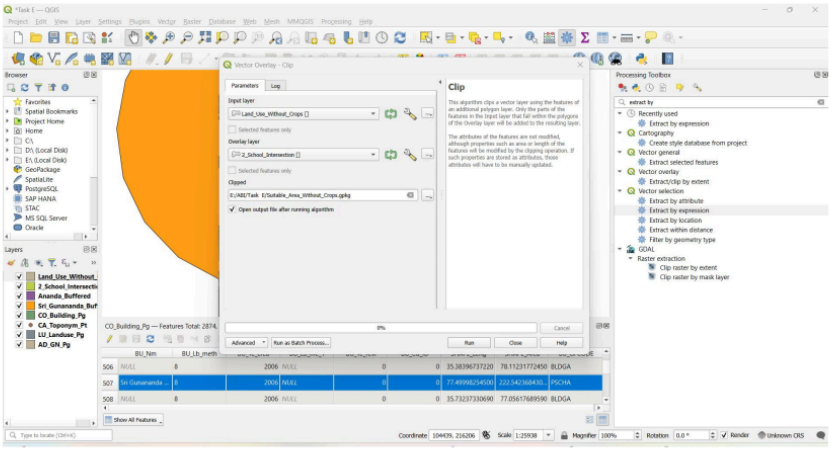


Figure 107 - Task E appendix 9

- ✓ Identifies educational buildings located within the suitable land zone. The output, Buildings_in_Current_Use.gpkg, highlights existing institutions in optimal locations for planning.

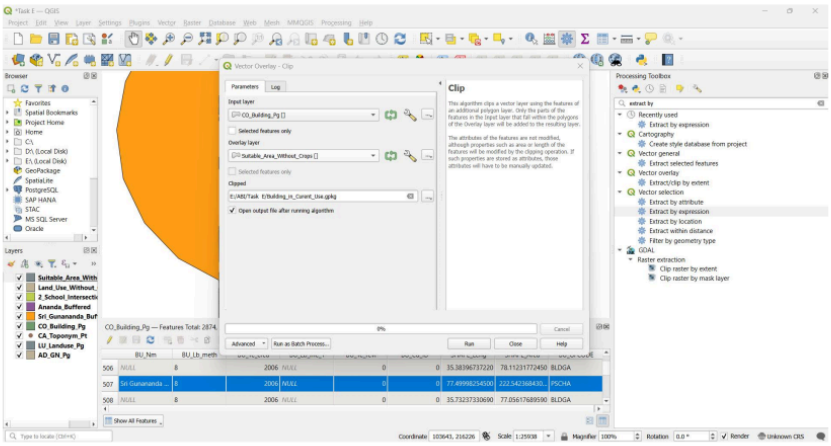


Figure 108 - Task E appendix 10

- ✓ Using the Difference tool, the land already occupied by buildings is removed from the suitable area. The output, Land_Remaining.gpkg, shows the land that is still available for development.

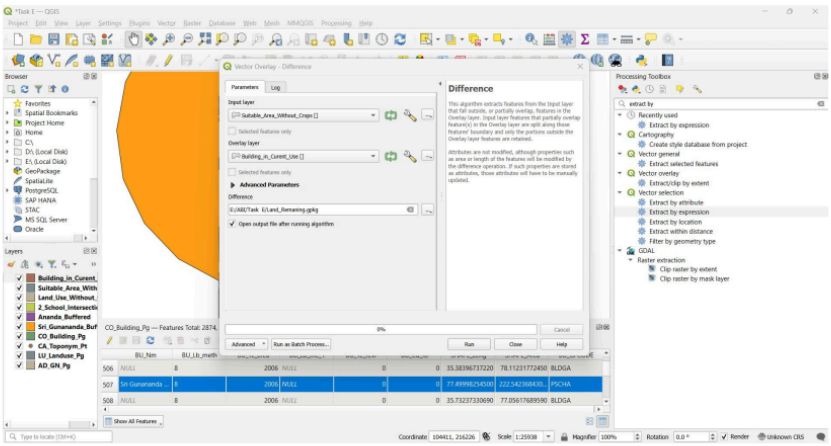


Figure 109 - Task E appendix

- ✓ The CRS is set to EPSG:5234 – Kandawala / Sri Lanka grid, ensuring accurate alignment and spatial measurements during all operations.

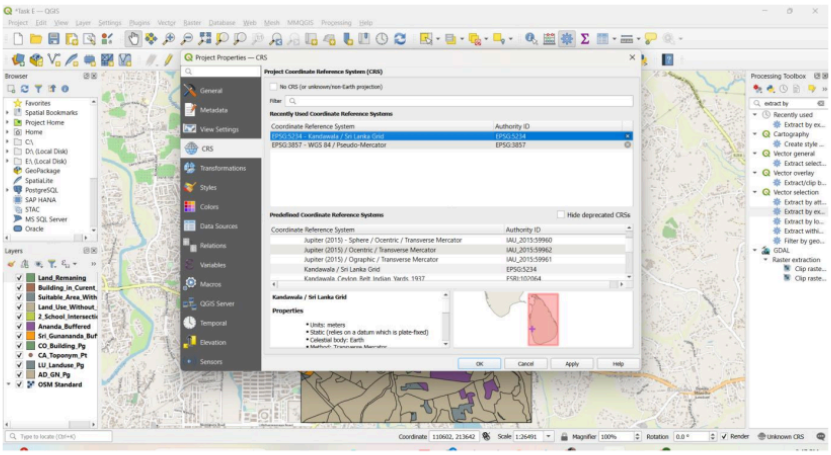


Figure 110 - Task E appendix 12

- ✓ The final map shows all processed layers – buffers, intersections, filtered land, buildings and remaining land – providing clear visibility for decision-making in education and urban development.

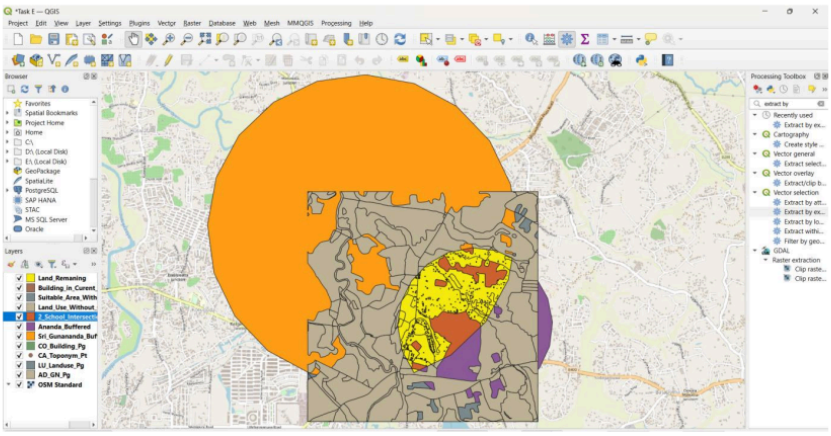


Figure 111 - Task E appendix 13

References

Geeks, f. G., 2025. *Geeks for Geeks.* [Online]
Available at: https://www.geeksforgeeks.org/r-language/shapiro-wilk-test-in-r-programming/
[Accessed 8 Augest 2025].

Pannell, R., 2022. *LEANSCAPE.* [Online]
Available at: https://leanscape.io/an-introduction-to-the-anderson-darling-normality-test/
[Accessed 08 Augest 2025].

Zaiontz, C., 2025. *real-statistics.com.* [Online]
Available at: https://real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/lilliefors-test-normality/
[Accessed 08 Augest 2025].

st20258024-CIS6008(1).pdf

ORIGINALITY REPORT

36%	4%	4%	35%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Wales Institute, Cardiff Student Paper	34%
2	Submitted to ESoft Metro Campus, Sri Lanka Student Paper	1%
3	link.springer.com Internet Source	<1%
4	Peter K. Dunn. "Scientific Research and Methodology - An Introduction to Quantitative Research and Statistics", CRC Press, 2025 Publication	<1%
5	Submitted to University of Kent at Canterbury Student Paper	<1%
6	Submitted to De Montfort University Student Paper	<1%
7	Ali E. M. Jarghon, Nyoman Anita Damayanti, Inge Dhamanti, Hari Basuki Notobroto, Atik Choirul Hidajah, Anas M. M. Awad. "Mapping Vulnerability to Potential Crisis Events in Surabaya City: A GIS-Based Approach", F1000Research, 2024 Publication	<1%
8	Walker, Jan, Almond, Palo. "EBOOK: Interpreting Statistical Findings: A Guide For	<1%

	Health Professionals And Students", EBOOK: Interpreting Statistical Findings: A Guide For Health Professionals And Students, 2010 Publication	
9	docslib.org Internet Source	<1 %
10	mpm2019.eu Internet Source	<1 %
11	Sai Kiran Oruganti, Dimitrios A Karras, Srinesh Singh Thakur, Janapati Krishna Chaithanya, Sukanya Metta, Amit Lathigara. "Digital Transformation and Sustainability of Business", CRC Press, 2025 Publication	<1 %
12	Rohan Paul, Swapnil Mishra, Jitendra Khatti. "Role of Artificial Intelligence (AI) Techniques in Tunnel Engineering—A Scientific Review", Indian Geotechnical Journal, 2025 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		