UTM

UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING

SEMESTER 2 2023/2024

SECI1143 PROBABILITY & STATISTICAL DATA ANALYSIS

SECTION 02

PROJECT 2

QS World University Rankings for 2025 Data Set

LECTURER: DR. NOORFA HASZLINNA BINTI MUSTAFFA

GROUP 4

Student Name	Matric No.
LAM YOKE YU	A23CS0233
GOE JIE YING	A23CS0224
TEH RU QIAN	A23CS0191
TAN YI YA	A23CS0187

TABLE OF CONTENTS

1.0 INTRODUCTION **3**
2.0 DATA SET **4**
2.1 Description of Data Set 4
2.2 Statistical Test Analysis 5
3.0 DATA ANALYSIS **7**
3.1 Hypothesis 1 Sample Test 7
3.1.1 Academic Reputation 7
3.1.2 Employer Reputation 7
3.1.3 QS Overall Score 8
3.2 Correlation test 9
3.2.1 Correlation between Academic Reputation and Employer Reputation 9
3.2.2 Correlation between Academic Reputation and QS Overall Score 10
3.2.3 Correlation between Employer Reputation and QS Overall Score 11
3.3 Regression test 11
3.3.1 Academic Reputation vs Employer Reputation 12
3.3.2 Academic Reputation vs QS Overall Score 13
3.3.3 Employer Reputation vs QS Overall Score 14
3.4 Goodness of fit test 15
4.0 CONCLUSION **17**
5.0 APPENDIX **18**

1.0 INTRODUCTION

The landscape of global higher education is continually changing, with institutions aiming for excellence in teaching, research and social impact. The QS World University Rankings provide a thorough analysis of how well universities perform globally, highlighting both their strong points and potential areas for development. In order to investigate the factors impacting the rankings of the top 500 universities, this project will analyse the QS World Rankings dataset using a variety of statistical methods, such as one-sample hypothesis testing, correlation and regression analysis and goodness of fit test.

The dataset for this study is based on the QS World University Rankings for 2025, which includes detailed information on the world's top 500 institutions. Numerous metrics are included in this dataset, such as size, academic reputation, employer reputation, QS overall score and more.

The objectives of this project are:

- To apply and perform statistical test analyses on the QS World Rankings dataset focusing on the 500 universities.
- To determine if the selected variables from the dataset are dependent on each other, thereby identifying significant factors contributing to the ranking of these top institutions.

By limiting our analysis to the top 500 universities, we aim to gain a deeper understanding of the distinguishing factors and performance metrics that set these universities apart. In the following sections, we will delve into the statistical methodologies and analyses performed to achieve these objectives.

Presentation and Reflection link: https://www.youtube.com/watch?v=iC5TdS7HTh0

2.0 DATA SET

2.1 Description of Data Set

The data set we have chosen is collected from online sources. It consists of information of QS world University Rankings for the year 2024 and 2025. This secondary data set consists of 16 different variables, and 500 data of different universities. The details of each variable are shown as below:

Variable	Description	Level of Measurement	Type of Data
2025 Rank	University's QS rank in the year 2025.	Ordinal	Discrete
2024 Rank	University's QS rank in the year 2024.	Ordinal	Discrete
Institution Name	Name of the university.	Nominal	Categorical
Location	Country code where the institution is located.	Nominal	Categorical
Location Full	Full name of the country where the institution is located.	Nominal	Categorical
Size	Size of the institution. (S, M, L, XL)	Nominal	Categorical
Academic Reputation	Score of the university's academic reputation.	Ratio	Continuous
Employer Reputation	Score of the university's employer reputation.	Ratio	Continuous
Faculty Student	Ratio of faculty members to students, indicating the level of personal attention students may receive.	Ratio	Continuous
Citation Per Faculty	Measure of the impact of research produced by faculty members.	Ratio	Continuous
International Faculty	Percentage of faculty members who are international.	Ratio	Continuous
International Students	Percentage of international students.	Ratio	Continuous
International Research Network	Score of the institution’s international research collaborations.	Ratio	Continuous
Employment Outcomes	Employment rate of graduates.	Ratio	Continuous

Variable	Description	Level of Measurement	Type of Data
Sustainability	The university's commitment to sustainable practices and environmental responsibility.	Ratio	Continuous
QS Overall Score	Overall score combining various metrics to rank the institution.	Ratio	Continuous

2.2 Statistical Test Analysis

Selected Variable	Objectives	Test Analysis and Expected Outcome
Academic Reputation	To determine if the Academic Reputation score is significantly higher than the assumed average.	**Analysis:** Hypothesis 1 Sample Test **Expected Outcome:** By conducting a one-sample t-test, we compare the mean Academic Reputation score with the assumed average of 40. If the p-value is significant (typically $p < 0.05$), the null hypothesis is rejected, suggesting the Academic Reputation score is significantly higher than the assumed average.
Employer Reputation	To determine if the Employer Reputation score is significantly higher than the assumed average.	**Analysis:** Hypothesis 1 Sample Test **Expected Outcome:** By conducting a one-sample t-test, we compare the mean Employer Reputation score with the assumed average of 40. If the p-value is significant (typically $p < 0.05$), the null hypothesis is rejected, suggesting the Employer Reputation score is significantly higher than the assumed average.
QS Overall Score	To determine if the QS Overall Score is significantly higher than the assumed average.	**Analysis:** Hypothesis 1 Sample Test **Expected Outcome:** By conducting a one-sample t-test, we compare the mean QS Overall Score with the assumed average of 40. If the p-value is significant (typically $p < 0.05$), the null hypothesis is rejected, suggesting the QS Overall Score is significantly higher than the assumed average.

Selected Variable	Objectives	Test Analysis and Expected Outcome
Academic Reputation, Employer Reputation	To examine the relationship between Academic Reputation and Employer Reputation.	**Analysis:** Correlation Analysis, Regression Test **Expected Outcome:** By calculating the correlation coefficient and performing regression analysis, we can determine the strength and direction of the relationship between Academic Reputation and Employer Reputation. A positive and statistically significant correlation would indicate that as Academic Reputation increases, Employer Reputation tends to increase as well.
Academic Reputation, QS Overall Score	To examine the relationship between Academic Reputation and QS Overall Score.	**Analysis:** Correlation Analysis, Regression Test **Expected Outcome:** Through correlation analysis and regression testing, we can evaluate how well Academic Reputation predicts the QS Overall Score. A significant relationship would suggest that Academic Reputation is a good predictor of QS Overall Score.
Employer Reputation, QS Overall Score	To examine the relationship between Employer Reputation and QS Overall Score.	**Analysis:** Correlation Analysis, Regression Test **Expected Outcome:** Similarly, we will use correlation analysis and regression testing to assess the relationship between Employer Reputation and QS Overall Score. A significant relationship would suggest that Employer Reputation is a strong predictor of QS Overall Score.
University Sizes (L, M, S, XL)	To determine if the sizes of universities are equally proportioned.	**Analysis:** Goodness of Fit Test **Expected Outcome:** By using a one-way contingency table at a 0.05 significance level, we aim to determine whether the observed distribution of university sizes aligns with the assumption of equal proportions. If the test statistic exceeds the critical value, we reject the null hypothesis, suggesting that the sizes are not equally proportioned.

3.0 DATA ANALYSIS

3.1 Hypothesis 1 Sample Test

Since our sample are universities of the top 500 universities according to the QS ranking. We are interested to know if their academic reputation, employer reputation and QS overall score are above average. By assuming the average score of all universities is 40 and using a significance level of 0.05 to test if the academic reputation, employer reputation and QS overall score of the 500 top universities are above average.

3.1.1 Academic Reputation

$$H_0: \mu = 40$$

$$H_1: \mu > 40$$

```
> hypothesisTesting(academicReputation, 40, 0.05)
[1] "Sample mean:  42.0486"
[1] "Sample Standard Deviation:  26.8962986263835"
[1] "Test statistics:  1.70313726893729"
[1] "Critical value:  1.64791298405971"
[1] "P-value:  0.0445825312374856"
```

Based on the output from R, we know that:

Mean academic reputation of the sample, $\bar{x} = 42.0486$

Academic reputation standard deviation of the sample, $s = 26.896$

Test statistics, $t_0 = 1.703$

Critical value, $t_{0.05,\,499} = 1.648$

P-value $= 0.045$

Since $t_0 = 1.703 > t_{0.05,\,499} = 1.648$, H_0 is rejected. There is sufficient evidence to conclude that the mean academic reputation of the top 500 universities is greater than 40 at the significance level of 0.05.

3.1.2 Employer Reputation

$$H_0: \mu = 40$$

$$H_1: \mu > 40$$

```
> hypothesisTesting(employerReputation, 40, 0.05)
[1] "Sample mean:  40.656"
[1] "Sample Standard Deviation:  28.9605478308219"
[1] "Test statistics:  0.506503054365125"
[1] "Critical value:  1.64791298405971"
[1] "P-value:  0.306363630741559"
```

Based on the output from R, we know that:

Mean employer reputation of the sample, $\bar{x} = 40.656$

Employer reputation standard deviation of the sample, $s = 28.961$

Test statistics, $t_0 = 0.507$

Critical value, $t_{0.05,\,499} = 1.648$

P-value $= 0.306$

Since $t_0 = 0.507 < t_{0.05,\,499} = 1.648$, H_0 is not rejected. There is insufficient evidence to conclude that the mean employer reputation of the top 500 universities is greater than 40 at the significance level of 0.05.

3.1.3 QS Overall Score

$$H_0: \mu = 40$$

$$H_1: \mu > 40$$

```
> hypothesisTesting(overallScore, 40, 0.05)
[1] "Sample mean:  45.7004"
[1] "Sample Standard Deviation:  18.3164490130838"
[1] "Test statistics:  6.95903550400777"
[1] "Critical value:  1.64791298405971"
[1] "P-value:  5.41009236148759e-12"
```

Based on the output from R, we know that:

Mean QS overall score of the sample, $\bar{x} = 45.700$

QS overall score standard deviation of the sample, $s = 18.316$

Test statistics, $t_0 = 6.959$

Critical value, $t_{0.05,\,499} = 1.648$

P-value $= 5.41 \times 10^{-12}$

Since $t_0 = -5.249 < t_{0.05,\,499} = 1.648$, H_0 is rejected. There is sufficient evidence to conclude that the mean QS overall score of the top 500 universities is greater than 40 at the significance level of 0.05.

3.2 Correlation test

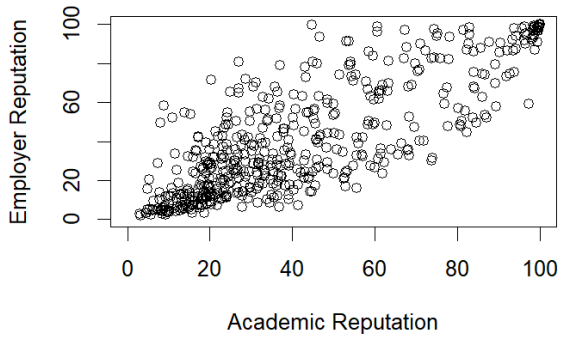
In this section, we will examine the relationships between various metrics associated with the top 500 universities according to the QS World University Rankings. We aim to explore the correlations between the QS Overall Score and other key metrics to understand how they are interrelated. The metrics under consideration are Academic Reputation, Employer Reputation, and the QS Overall Score. This analysis will be conducted using correlation coefficients and hypothesis testing at a significance level of 0.05 to determine the strength and direction of the associations.

3.2.1 Correlation between Academic Reputation and Employer Reputation

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

```
> correlationTesting(academicReputation, employerReputation, "Academic Reputation", "Employer Reputation", 0.05)
[1] "Correlation = 0.806487992488982"
[1] "t-statistic = 30.4397497920233"
[1] "Degrees of freedom = 498"
[1] "Critical value at alpha 0.05 = 1.96473898296729"
```



Based on the output from R, correlation,

Correlation, $r = 0.806$

Test-statistic, $t_0 = 30.440$

Degree of freedom $= 498$

Critical value $t_{0.05,\,498} = 1.965$

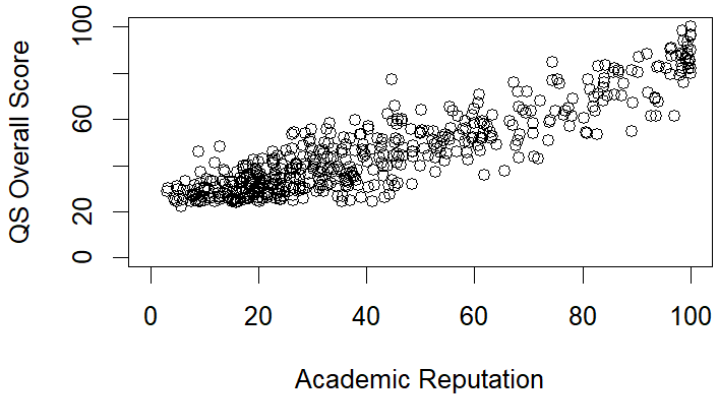
Since $t_0 = 30.440 > t_{0.05,\,498} = 1.965$, H_0 is rejected. There is sufficient evidence to conclude that there exists a correlation between Academic Reputation and Employer Reputation in the top 500 universities according to the QS World University Rankings.

3.2.2 Correlation between Academic Reputation and QS Overall Score

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

```
> correlationTesting(academicReputation, QSoverallScore, "Academic Reputation", "QS Overall Score", 0.05)
[1] "Correlation = 0.895234438246447"
[1] "t-statistic = 44.8343325883364"
[1] "Degrees of freedom = 498"
[1] "Critical value at alpha 0.05 = 1.96473898296729"
```



Based on the output from R, correlation,

Correlation, r = 0.895

Test-statistic, $t_0 = 44.834$

Degree of freedom = 498

Critical value $t_{0.05,\,498} = 1.965$

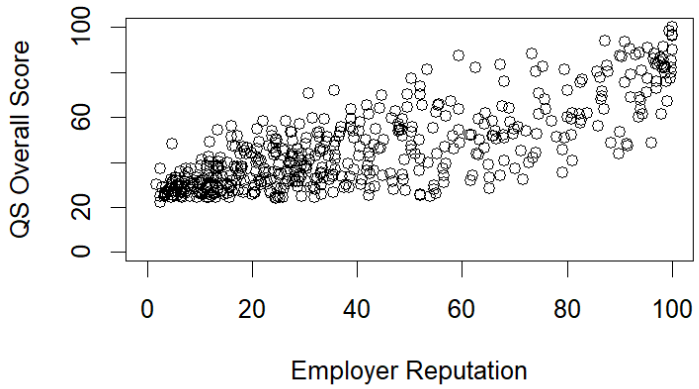
Since $t_0 = 44.8340 > t_{0.05,\,498} = 1.965$, H_0 is rejected. There is sufficient evidence to conclude that there exists a correlation between Academic Reputation and QS Overall Score in the top 500 universities according to the QS World University Rankings.

3.2.3 Correlation between Employer Reputation and QS Overall Score

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

```
> correlationTesting(employerReputation, QSoverallScore, "Employer Reputation", "QS Overall Score", 0.05)
[1] "Correlation = 0.784115646061569"
[1] "t-statistic = 28.1950394162909"
[1] "Degrees of freedom = 498"
[1] "Critical value at alpha 0.05 = 1.96473898296729"
```



Based on the output from R, correlation,

Correlation, $r = 0.784$

Test-statistic, $t_0 = 28.195$

Degree of freedom $= 498$

Critical value $t_{0.05,\,498} = 1.965$

Since $t_0 = 28.195 > t_{0.05,\,498} = 1.965$, H_0 is rejected. There is sufficient evidence to conclude that there exists a correlation between Employer Reputation and QS Overall Score in the top 500 universities according to the QS World University Rankings.

3.3 Regression test

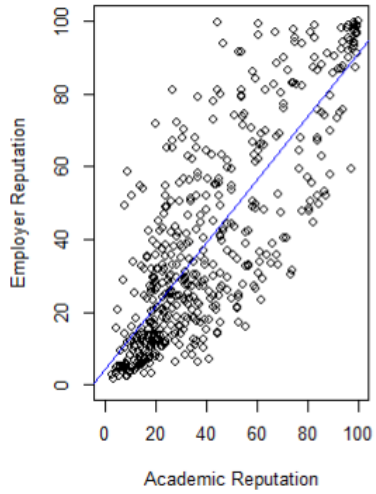
In this section, we explore the relationships between key metrics associated with the top 500 universities according to the QS World University Rankings. Our analysis focuses on understanding these relationships through regression analysis, examining how Academic Reputation and Employer Reputation predict the QS Overall Score. By using a significance level of 0.05, we assess the strength and direction of these relationships.

3.3.1 Academic Reputation vs Employer Reputation

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

```
[1] "T-test Results:"
[1] "Test statistic (t): 30.440"
[1] "Degrees of freedom: 498"
[1] "Critical t-value (alpha = 0.05 ): 1.9647389830"
```



Based on the output from R,

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-37.842 -10.755  -2.796   9.378  56.642

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.14164    1.42356   2.909  0.00378 **
x            0.86838    0.02853  30.440  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.14 on 498 degrees of freedom
Multiple R-squared:  0.6504,    Adjusted R-squared:  0.6497
F-statistic: 926.6 on 1 and 498 DF,  p-value: < 2.2e-16
```

Estimated of the regression intercept, $b_0 = 4.142$

Estimated of the regression slope, $b_1 = 0.868$

Estimator of the standard error of the slope, $s_{b1} = 0.0285$

Standard error of the residuals, $s_\epsilon = 17.14$

Degree of Freedom, $df = 498$

Coefficient of Determination, $R^2 = 0.6504$

Estimated Regression Model, $\widehat{y_i} = 4.142 + 0.868x$

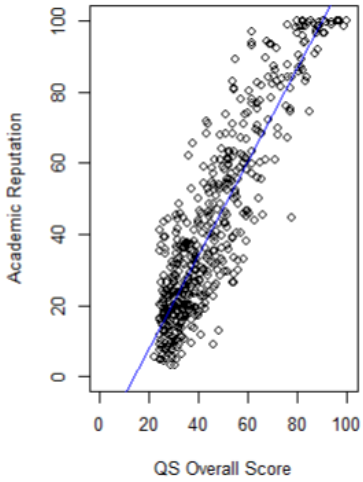
Since $t_0 = 30.440 > t_{0.05,\,498} = 1.965$, H_0 is rejected. There is sufficient evidence to suggest a statistically significant relationship between Academic Reputation and Employer Reputation.

3.3.2 Academic Reputation vs QS Overall Score

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

```
[1] "T-test Results:"
[1] "Test statistic (t): 44.834"
[1] "Degrees of freedom: 498"
[1] "Critical t-value (alpha = 0.05 ): 1.9647389830"
```



Based on the output from R,

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-39.152  -8.235  -0.396   7.316  34.926

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.02837    1.44339  -12.49   <2e-16 ***
x             1.31458    0.02932   44.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12 on 498 degrees of freedom
Multiple R-squared:  0.8014,    Adjusted R-squared:  0.801
F-statistic:  2010 on 1 and 498 DF,  p-value: < 2.2e-16
```

Estimated of the regression intercept, $b_0 = -18.028$

Estimated of the regression slope, $b_1 = 1.315$

Estimator of the standard error of the slope, $s_{b1} = 0.02932$

Standard error of the residuals, $s_\epsilon = 12.00$

Degree of Freedom, $df = 498$

Coefficient of Determination, $R^2 = 0.8014$

Estimated Regression Model, $\widehat{y_i} = -18.028 + 1.315x$

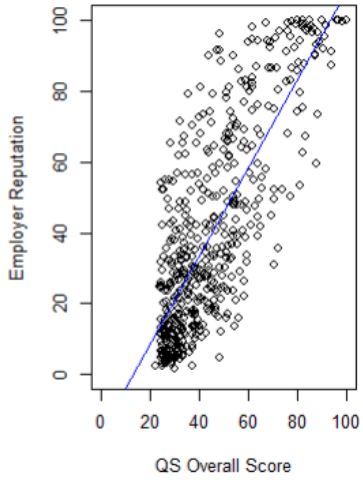
Since $t_0 = 44.834 > t_{0.05,498} = 1.965$, H_0 is rejected. There is sufficient evidence to suggest a statistically significant relationship between Academic Reputation and QS Overall Score.

3.3.3 Employer Reputation vs QS Overall Score

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

```
[1] "T-test Results:"
[1] "Test statistic (t): 28.195"
[1] "Degrees of freedom: 498"
[1] "Critical t-value (alpha = 0.05 ): 1.9647389830"
```



Based on the output from R,

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-40.850 -12.911  -3.507  11.434  51.973

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.00257    2.16462  -7.393 6.13e-13 ***
x             1.23978    0.04397  28.195  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.99 on 498 degrees of freedom
Multiple R-squared:  0.6148,    Adjusted R-squared:  0.6141
F-statistic:   795 on 1 and 498 DF,  p-value: < 2.2e-16
```

Estimated of the regression intercept, $b_0 = -16.003$

Estimated of the regression slope, $b_1 = 1.240$

Estimator of the standard error of the slope, $s_{b1} = 0.04397$

Standard error of the residuals, $s_\epsilon = 17.99$

Degree of Freedom, $df = 498$

Coefficient of Determination, $R^2 = 0.6148$

Estimated Regression Model, $\widehat{y_i} = -16.003 + 1.240x$

Since $t_0 = 28.195 > t_{0.05,498} = 1.965$, H_0 is rejected. There is sufficient evidence to suggest a statistically significant relationship between Employer Reputation and QS Overall Score.

3.4 Goodness of fit test

This research investigates the size distribution of the top 500 universities in the QS World University Rankings 2025. There are four size categories among these universities: S, M, L and XL. We assume that the sizes of all universities are equally proportioned. By using a one-way contingency table at a 0.05 significance level, we aim to determine whether the observed distribution of university sizes aligns with this assumption.

Below is the table showing the observed frequencies of university sizes for the top 500 universities:

```
> print(noSize)

  L   M   S  XL
238  77  24 161
```

$$H_0: p_L = p_M = p_S = p_{XL} = \frac{1}{4}$$

H_1: At least one $p_i \neq \frac{1}{4}$

```
> goodnessOfFitTest(degrees_of_freedom, critical_value, output)
[1] "Degree Of Freedom: 3"
[1] "Critical Value: 7.81472790325118"

        Chi-squared test for given probabilities

data:  noSize
X-squared = 212.56, df = 3, p-value < 2.2e-16
```

By using RStudio, we know that:

Degree of Freedom, df = 3

Critical Value, $X^2_{3,\,0.05} = 7.815$

Test Statistic, $X^2 = 212.56$

Since $X^2 = 212.56 > X^2_{3,\,0.05} = 7.815$, H_0 is rejected. There is sufficient evidence to conclude that the proportions of university sizes different significantly.

4.0 CONCLUSION

It has been a rewarding experience to apply the knowledge learned in class to a real world case scenario, especially on the QS World University Ranking, which is relevant to students, lecturers and universities.

Our analysis centres around two key metrics which are the academic reputation and employer reputation which contribute 30% and 15% respectively to the QS Overall Score. The one-sample hypothesis allowed us to assess whether the Academic Reputation, Employer Reputation and QS Overall Score were significantly above an assumed average of 40, which as a result, there is sufficient evidence that the mean Academic Reputation and and the QS Overall Score of the top 500 universities are above the assumed average at a significance level of 95%. The correlation and regression test shows that there linear relationship does exist between each of the three metrics. Other than that, we are also interested to know the distribution of the university sizes. The goodness-of-fit test highlights that university sizes are not equally proportioned.

The most compelling finding was the significant influence of Academic Reputation on Employer Reputation and OS Overall Score. This shows the importance of academic excellence in shaping a university's overall standing and reputation. Besides, with the regression model, we can use the Academic Reputation to predict both Employer Reputation and QS Overall Score. However, we also understand that there are various other factors that contribute to the QS Overall Score and each metrics as well have their very own scoring criteria.

As a whole, exploring metrics like Academic Reputation and Employer Reputation among the top 500 universities provided valuable insights into factors that drive their global standing. Nevertheless, it is also important to acknowledge that our findings may not fully represent the entire university population, as our study was limited to this specific sample. This limitation encourages further exploration into broader trends and variations among universities worldwide.

5.0 APPENDIX

Data Source:
https://www.kaggle.com/datasets/darrylljk/worlds-best-universities-qs-rankings-2025/data