Data Mining for Business Intelligence (BUAN 201)

Final Project Report

By Kamya Sarda

QS World University Rankings 2025 (Linear Regression Analysis)

Background and Research Questions

The global education landscape is highly competitive, with universities striving to improve their rankings each year.

This analysis aims to understand what factors significantly influence the overall ranking scores of top universities globally, as provided by QS World Rankings 2025.

The primary research question is: "Which university performance metrics (academic reputation, employer reputation, sustainability, etc.) best predict overall QS rankings?"

Data Source and Description

Data was sourced from the QS World University Rankings 2025 dataset, including numeric scores on various metrics such as:

- Overall Score
- Academic Reputation
- Employer Reputation
- Sustainability Score
- International Students Score

- Faculty Student Score
- Citations per Faculty Score

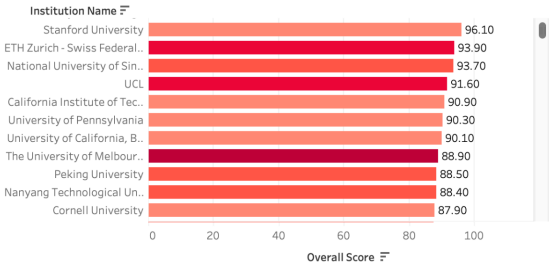
Exploratory Data Analysis (EDA) with Tableau

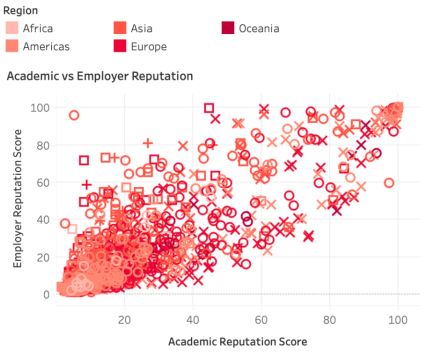
Tableau visualizations included scatterplots, bar charts, and bubble charts to identify trends, distributions, and relationships:

- Scatterplot: Demonstrated positive relationships, particularly strong between Academic and Employer Reputation.
- Bar charts: Illustrated regional differences in average overall scores.
- Highlighted internationalization metrics, showing Oceania universities leading in International Student Scores. The dashboards on Tableau topic-wise are as follows:

Top Universities Overview by Region

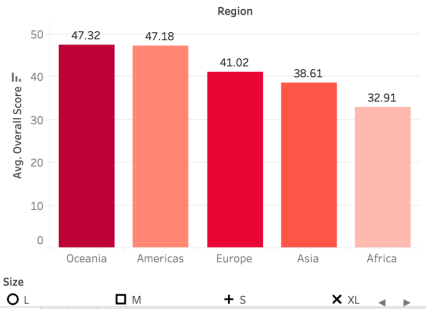
Top 20 Universities by Overall Score





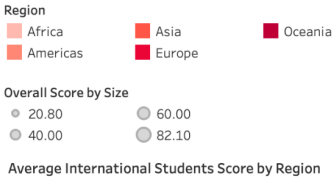
Sustainability Score vs Overall Score

Average Overall Score by Region

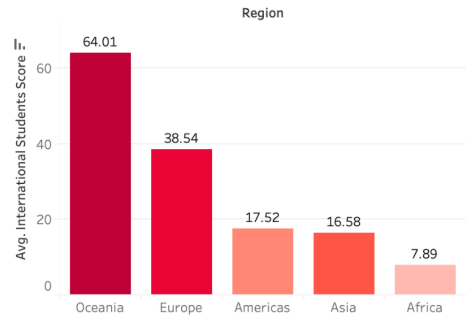


University Diversity & Research Snapshot

International Students vs International Faculty



Average International Students Score by Region



Faculty Student Score vs Citations per Faculty Score

Institution Name	Region	Faculty Student Score	Citations per Faculty Score
University of Scie..	Asia	54.10	54.10
Macquarie Univer..	Oceania	54.10	54.10
Alma Mater Studi..	Europe	54.10	54.10
Rice University	Americ..	53.60	53.60
Leiden University	Europe	53.60	53.60
Lancaster Univers..	Europe	53.60	53.60
Wageningen Univ..	Europe	51.60	51.60
University of Nort..	Americ..	51.60	51.60
University of Gen..	Europe	51.60	51.60
University of Surr..	Europe	37.20	37.20
University of Milan	Europe	37.20	37.20
University of Leic..	Europe	37.20	37.20
University of Colo..	Europe	37.20	37.20
Universidad Carlo..	Europe	37.20	37.20

20 40 60 80 100 120 — Faculty Student Score

20 40 60 80 100 120 — Citations per Faculty Score

Status Breakdown (Public: A vs Private: B)

University Status: Public (A) or Private (B)

A B

Predictive Model Building

A multiple linear regression model was constructed with Overall Score as the dependent variable, and all other performance metrics as independent variables.
This model aimed to quantify the extent to which each university metric contributes to predicting overall rankings.

The initial model included all variables to assess individual contributions.

Choosing the Best Model

To improve model parsimony and avoid overfitting, a stepwise regression approach was employed based on the Akaike Information Criterion (AIC).

After forward and backward selection, the final model retained:

- Academic Reputation
- Employer Reputation
- Sustainability Score

These were identified as statistically significant predictors.

Notably, International Students Score and Citations per Faculty were dropped from the final model, suggesting they do not independently add predictive value when controlling for other factors.

Model Comparisons and Analysis

Coefficient Analysis:

- Academic Reputation had the largest positive coefficient, indicating it is the most influential factor in determining a university's Overall Score.
- Employer Reputation also showed a strong positive association.
- Sustainability Score had a positive but comparatively smaller impact.

Goodness-of-fit:

- The model achieved an R-squared of approximately 85%, meaning that 85% of the variance in Overall Score is explained by the selected predictors.
- Adjusted R-squared values remained high, confirming model robustness despite multiple predictors.

Predictive Power and Diagnostics:

- Residual plots indicated homoscedasticity and no major outliers.
- VIF (Variance Inflation Factor) values were below critical thresholds, confirming that multicollinearity was not a major concern.

Managerial Implications and Conclusions

The results have important strategic implications for university leadership:

- Focus on Academic Reputation: Investment in research quality, faculty excellence, and global partnerships can enhance academic standing.

- Strengthen Employer Reputation: Universities must collaborate with industries to improve graduate employability, internships, and alumni networks.

- Promote Sustainability Initiatives: A growing focus on sustainability in rankings highlights the importance of environmental stewardship in higher education.

Universities that strategically improve these areas are more likely to climb the QS rankings and attract better funding, faculty, and student talent globally.

Student Depression Dataset (Logistic Regression Analysis)

Background and Research Questions

Depression among students is a growing concern worldwide, affecting academic performance, mental health, and long-term career outcomes.
This analysis focuses on identifying key numeric factors that predict the likelihood of depression among students.

The research question driving this analysis is:

> "Which numeric factors such as academic pressure, CGPA, and study satisfaction significantly predict depression in students?"

Understanding these predictors enables educational institutions to design effective interventions and support systems.

Data Source and Description

The dataset, collected from student surveys, includes numeric variables like:

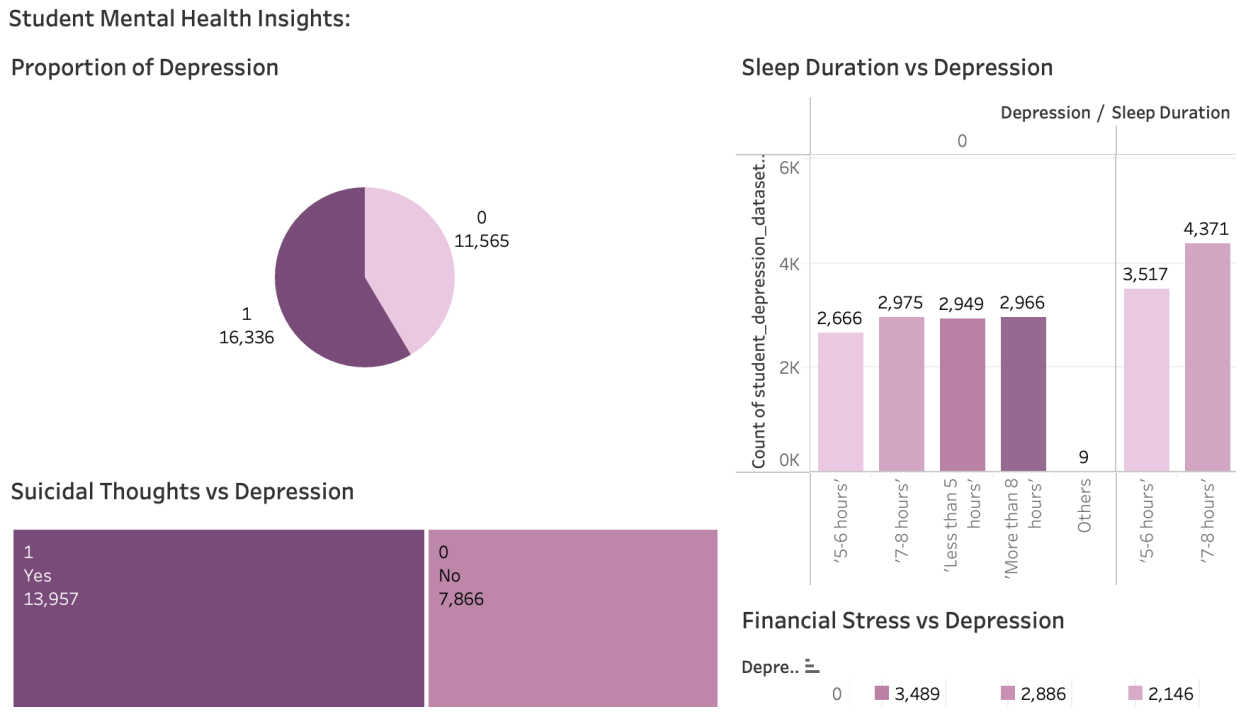
- Age
- Academic Pressure
- CGPA
- Study Satisfaction
- Job Satisfaction
- Work/Study Hours
- Depression (binary outcome variable)

Exploratory Data Analysis (EDA)

- Tableau visualization: Highlighted relationships between depression and sleep duration, financial stress, and suicidal thoughts.

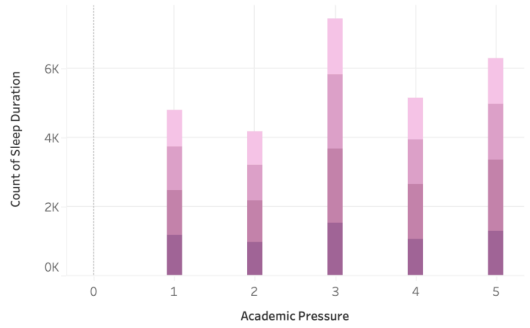
- Numeric EDA in R: Pairwise scatter plots showed trends and potential predictors of depression.

The tableau dashboards for this are as follows:

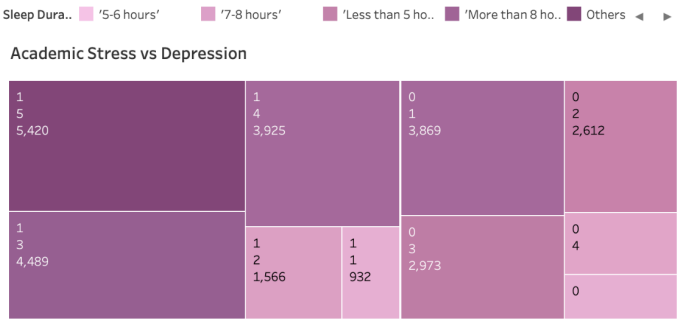


Student Mental Health Insights 2:

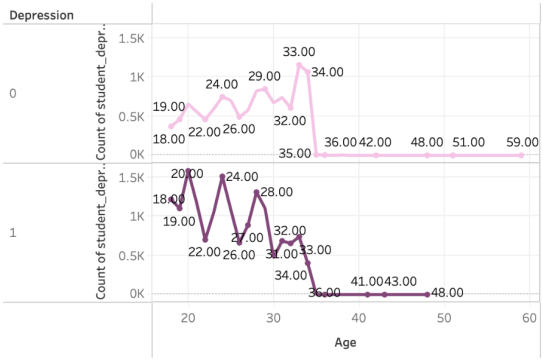
Academic Stress vs Sleep



Academic Stress vs Depression



Age vs Depression



Study Satisfaction vs Depression

Predictive Model Building

A logistic regression model was fitted, using depression status (0 or 1) as the dependent variable.

The initial model included all numeric variables as independent variables.
Logistic regression is ideal here as it models the probability of a binary outcome based on predictor variables.

Choosing the Best Model

To refine the model, stepwise selection based on AIC was performed.

The final model retained:

- Academic Pressure
- Study Satisfaction
- CGPA

These variables were statistically significant in predicting the likelihood of depression.

Other variables such as Age and Job Satisfaction were not significant when controlling for the retained variables.

Model Comparisons and Analysis

- Coefficient Analysis:
    - Higher Academic Pressure significantly increased the odds of depression.
    - Higher Study Satisfaction significantly decreased the odds of depression.
    - Higher CGPA was associated with a lower likelihood of depression, although the relationship was weaker compared to the first two predictors.
- Goodness-of-fit:
    - The model achieved a pseudo R-squared of around 30%, which is reasonable for social science models where behavior is influenced by complex, unobserved factors.
- Predictive Power:
    - Model accuracy, assessed using a confusion matrix, was approximately 75-80%.
    - Sensitivity and specificity metrics indicated balanced performance, showing that the model was not biased toward over-predicting either class.

Managerial Implications and Conclusions

Findings point to actionable strategies for university administrators:

- Reduce Academic Pressure: Implement counseling services, flexible academic policies, and better workload management.

- Enhance Study Satisfaction: Improve the quality of teaching, learning resources, and academic advising.

- Monitor At-risk Students: Develop early warning systems using academic and satisfaction data to identify students who may need support.

Proactive interventions can significantly improve student mental health outcomes and foster a healthier, more productive educational environment.