QS Global University Rankings: A Data Science Journey

Yaman Youssef Shiha & Mohammad Ahmad Tellawi

Supervisor: Dr. Linda Mahmoudi

Lebanese University – Data Science Program

Academic Year: 2024–2025

Introduction & Motivation

Why Analyse QS Rankings?

•To understand what drives **ranking improvements or declines**.

•To identify **university clusters** based on shared characteristics.

•To forecast **future rank trends** and simulate potential changes.

•To help universities like ours improve their strategic planning and visibility.

Key Research Questions

- What factors most influence QS rankings?
- Can we detect meaningful groups of similar universities?
- Are rank movements predictable using data science methods?
- What actions can institutions take to climb in future rankings?

Dataset & Preprocessing

- Datasets used :
  - QS Rankings 2025–2026 (1,495 universities)
  - Historical QS rankings from 2004–2026

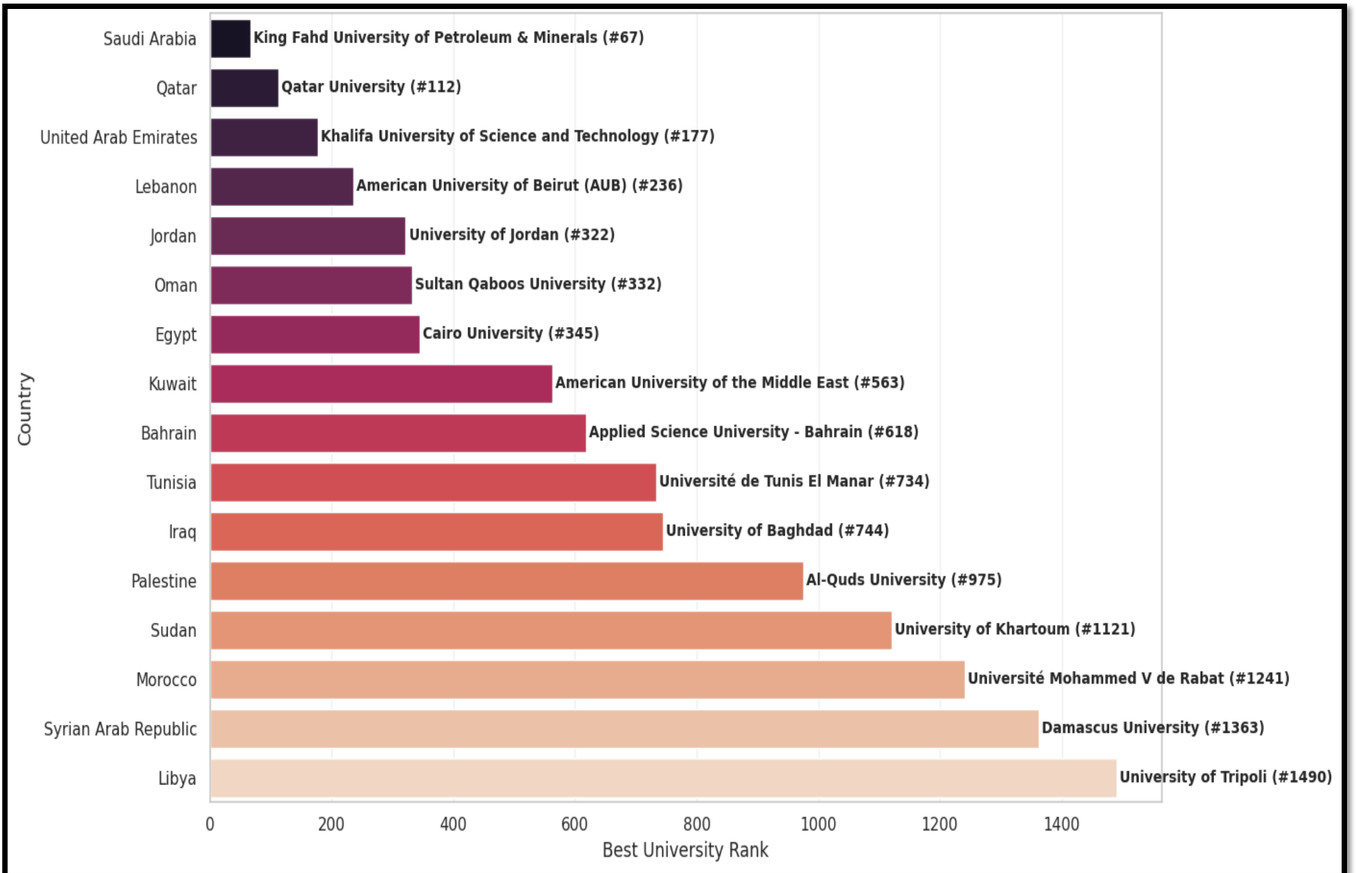
- Preprocessing :
  - Handling Missing values
  - Standardized university name

- Feature Engineering
  - Rank Improvement = 2025 Rank – 2026 Rank
  - Rank Trend = Improved / Declined / Stable

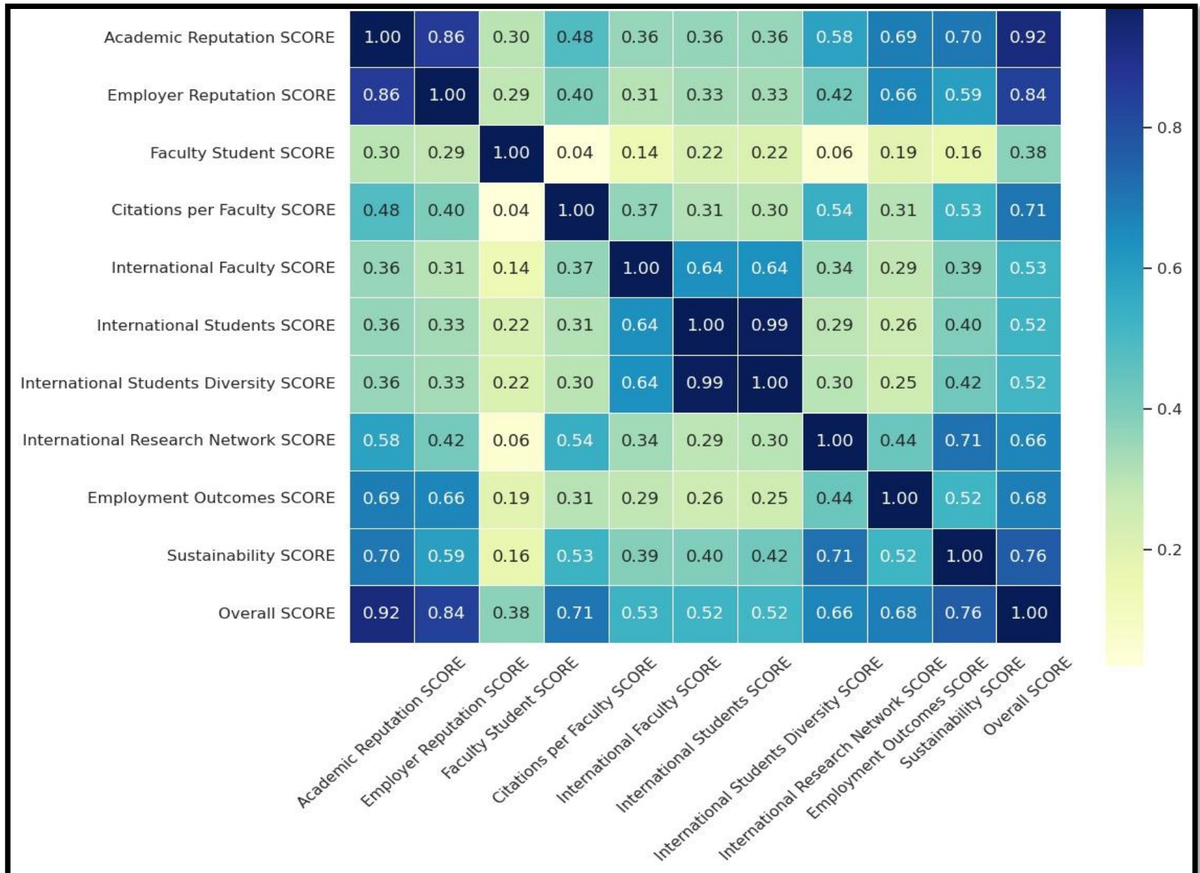
- Normalization & Modelling prep:
  - Z-score normalization for fair clustering
  - Clustering → used performance scores only
  - Classification → used 2025 scores to predict 2026 trend

Exploratory Data Analysis:

- Top Arabic Countries by Best University Rank



- Academic Reputation is the strongest driver of overall score



Clustering Methodology :

•**K-Means:**

•Objective: minimize within-cluster variance (WCSS).

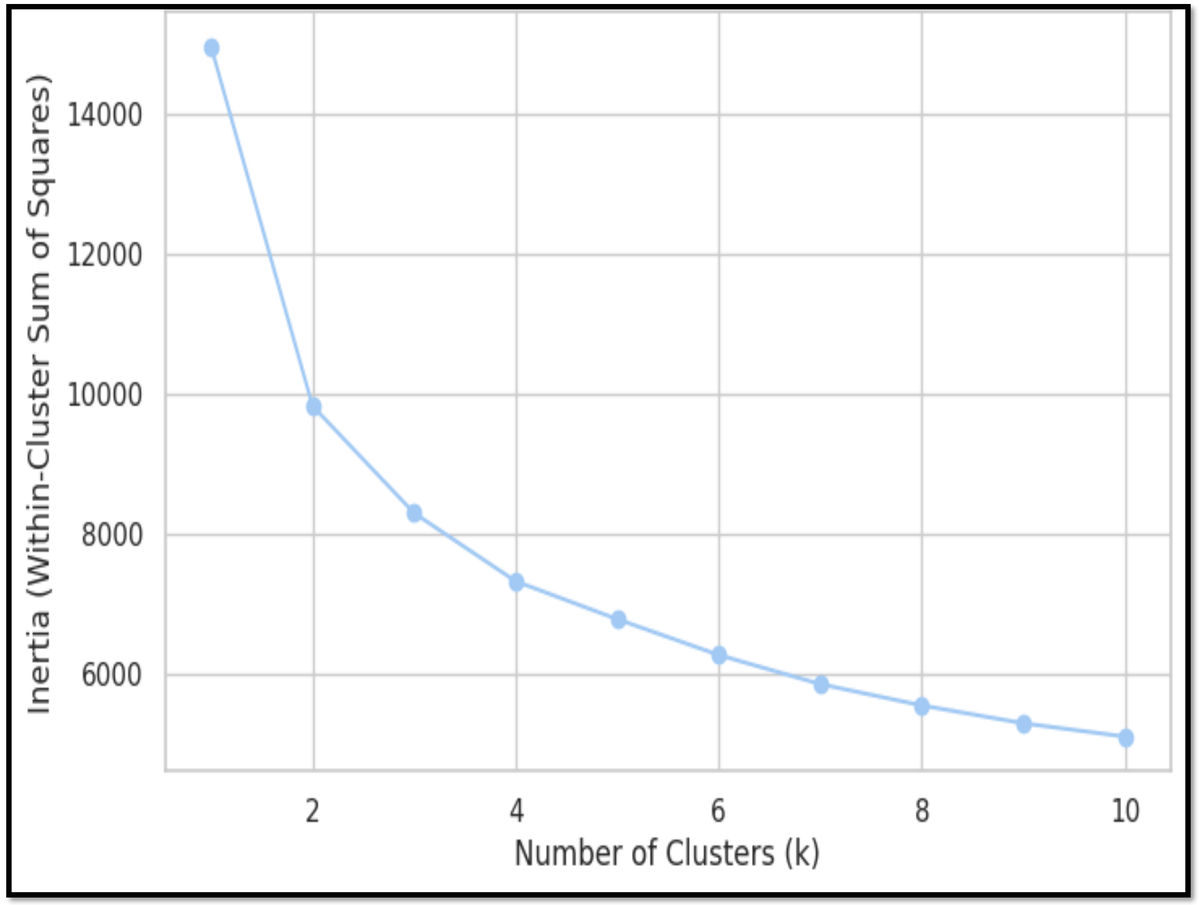
•Parameter tuning via **Elbow Method** and **Silhouette Score**.

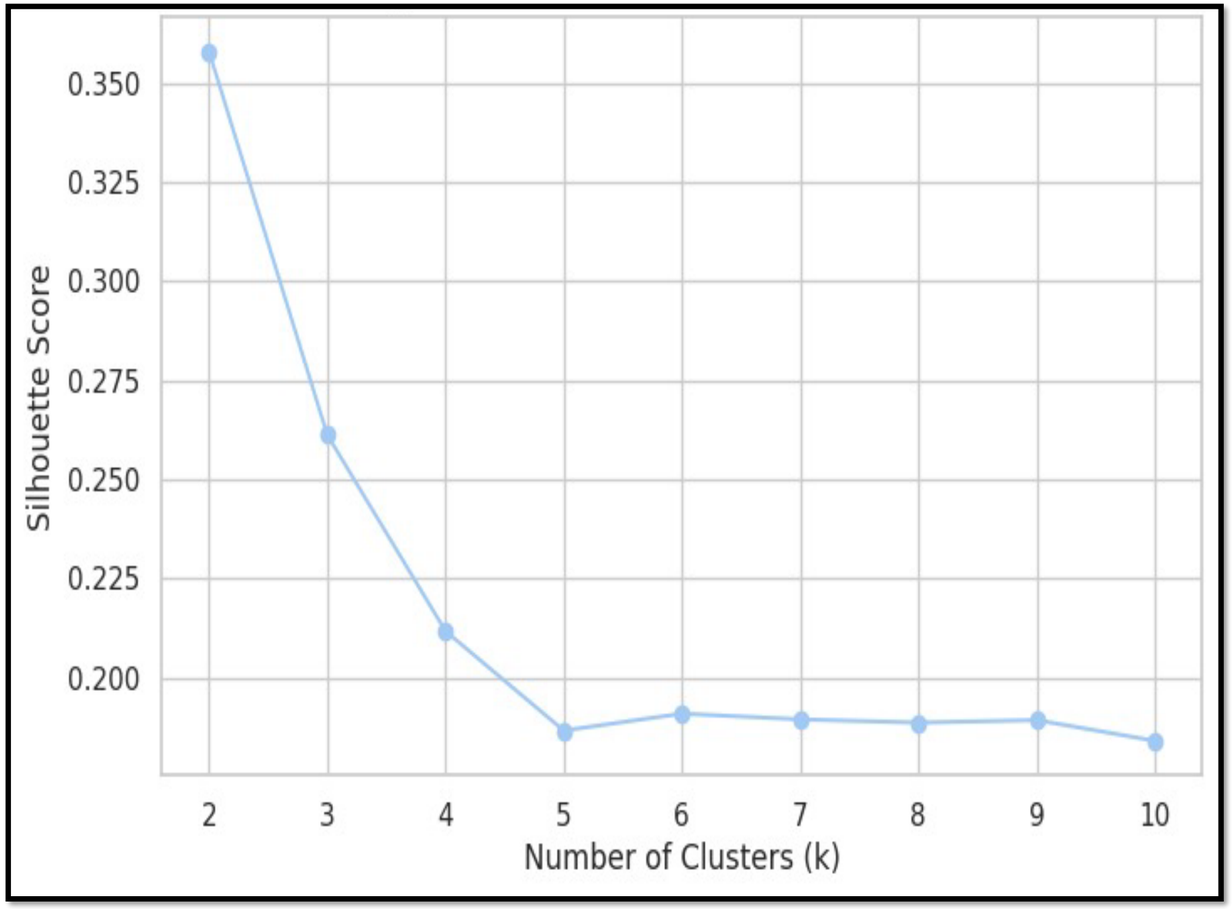
•Formula: $\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$

DBSCAN:

- Density-based clustering, handles outliers.
- Parameters: (neighbourhood radius), min points to form a cluster
- Input: z-score normalized QS performance scores.

Clustering Methodology 2:





Use Elbow to narrow down the range of k values.

Use Silhouette Score to pick the best one from that range.

Time-Series Forecasting:

- Used historical QS rank data (2004–2026) for trend prediction.
- Forecasting model: **Facebook Prophet**.
- Model captures trend + seasonality: $y(t) = g(t) + s(t) + h(t) + \epsilon_t$

  G(t): trend function

  S(t): seasonal effects

  H(t): regime shifts (QS methodology change)
- tested ETS and Linear Regression.
- Evaluated model accuracy using MAE and RMSE.

Classification

Hidden Champion Classifier

•Compares predicted vs. actual performance
•Hidden Champion: Under-ranked
•Overhyped: Over-ranked
•Aligned: Fairly ranked

Regional Tier Classifier:

- Predicts if a university is **Top/Mid/Lower** tier within its region
- Based on 2025 indicators and clustering output
- Helps identify **regional excellence patterns**

Classification Equations:

Logistic Regression Formula (Multinomial Case)

- Used to estimate probability of university belonging to each class (Improved, Declined, Stable): $P(y = k \mid \mathbf{x}) = \frac{\exp(\beta_k^T \mathbf{x})}{\sum_{j=1}^{K} \exp(\beta_j^T \mathbf{x})}$

Random Forest Classifier

- Ensemble of decision trees built on random subsets of data and features
- Captures non-linear relationships and interactions
- Feature Importance (Gini Impurity): $\text{Gini}(t) = 1 - \sum_{i=1}^{C} p_i^2$

XGBoost Classifier

- Boosted tree-based algorithm with high predictive power
- Automatically handles missing values and outliers
- Captures complex feature interactions via boosting

Streamlit App:

Unified dashboard for:

- EDA filters
- Clustering visualizations (radar, PCA)
- Classification results (hidden champions)
- Score calculator & forecasts

Conclusion&Recommendation:

- Academic Reputation is the key driver
- Internationalization helps rank improvement
- Institutions should:

Invest in research (citations)

Enhance global partnerships

Monitor trend indicators continuously

THANK YOU