الهيئة السعودية للتخصصات الصحية
Saudi Commission for Health Specialties

Comparing Universities' Performance In National Licensure Exam Using Ranking Categories Defined By Confidence Interval Approach.

June 2024

Contents

Introduction **3**
Confidence Interval Method 3
Advantages of Using Confidence Interval (CI) for Defining Ranking Categories 4
K-Mean Clustering Method 5
Methodology **7**
Sample of this report 7
Table 1. Descriptive statistics of total scores for SMLE across the year 7
Results **8**
Table 2. Criteria for each ranking category based on analysis (SMLE) 8
Table 3. Distribution of Universities across the defined ranking categories based on analysis 8
Categories Interpretation **9**
Figure 1. SMLE Distribution of Universities across the defined categories 9
Table 4. ANOVA Multiple Comparison Tests Results 10
Figure 2. Distribution of universities across the defined categories (Pass rate vs Overall Mean Score) 11
Figure 3. Distribution of universities across the defined categories based on Clustering Analysis ... 11
Potential Confounding Factors **13**
Conclusion **14**
References **15**
Appendix **16**
Appendix A. CI Estimate by University for SMLE 16
Appendix B. CI Estimate and Ranking Categories for the Entire Data SMLE 17

Introduction

This report aims to define ranking categories using machine learning-based K-Means clustering analysis along with confidence interval (CI) estimates methods based on Saudi Licensure Exams (SLEs) scores data for each Saudi University. Confidence intervals provide a range of values within which the true population parameter is likely to fall. By considering the mean scores, associated lower limit, and upper limit of mean scores related to each university, we established and validated ranking categories to compare universities' performance effectively. The K-means clustering algorithm helps group universities into distinct performance categories based on their overall mean scores, associated confidence intervals of the mean scores, and pass rates, allowing for the identification of distinct performance categories. These clusters were further validated using confidence intervals and analysis of variance(ANOVA) methods to ensure that the ranking categories were statistically significant and meaningful. This combination of clustering and confidence interval analysis provides a robust framework for comparing the performance of universities on the SLEs, ensuring that the rankings reflect true differences in performance levels.

The importance of using this methodology to define ranking categories lies in its ability to provide a clear, data-driven comparison of university performance. Recent studies underscore the critical role of cluster analysis in educational contexts. For instance, some research highlights the efficacy of using unsupervised machine learning techniques, such as K-Means clustering, to group universities based on performance metrics, which aids in setting clear institutional goals and strategies (Elbawab, 2020).

By categorizing universities into performance clusters, stakeholders can more easily identify strengths and areas for improvement based on statistically significant differences. These rankings can then be used to inform policy decisions, allocate resources, and develop targeted strategies to enhance educational outcomes across universities.

Confidence Interval Method

A CI is a statistical range that estimates the true value of a population parameter based on sample data. It provides a measure of the uncertainty surrounding an estimate. The total scores indicate the overall performance of the students from each university, while the pass rates reflect the proportion of students who successfully passed the SLEs. In this study, we have calculated CI for the mean performance of different universities. The lower limit represents the lower bound of the CI, while the upper limit represents the upper bound for 95% CI. The formula for the confidence interval for the sample mean is given below:

$$Point\ Estimate \pm (Critical\ Value)x(Standard\ Error)$$

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

where:

$\bar{x} = point\ estimate$

$s = sample\ standard\ deviation$

$t = critical\ value\ from\ t - distribution\ with\ n - 1\ degrees\ of\ freedom$

Note: t-distribution, rather than z-distribution, was used to obtain critical values since the sample data was used for analysis

Advantages of Using Confidence Interval (CI) for Defining Ranking Categories

Because universities have different numbers of students, it is important to define ranking categories based on CI rather than just relying on only the mean score or pass rate for the following reasons:

1. **Account for uncertainty:** By considering the CI, we can make a fairer comparison between universities by accounting for the uncertainty associated with the estimates. Relying exclusively on the mean score with different numbers of test takers can be misleading. A university with a high mean score but a small number of test takers may have a wider confidence interval, indicating a higher level of uncertainty around the estimate. In contrast, a university with a lower mean score but a larger number of test takers may have a narrower confidence interval, suggesting a more precise estimate.
2. **Statistical Significance:** CI provides insights into the statistical significance of the differences between universities and defined ranking categories. If the CIs of the two universities do not overlap or there is a significant drop in sample means, it suggests that there is a statistically significant difference between their mean scores. Additionally, an analysis of variance (ANOVA) is conducted in order to test the significance of the defined categories.
3. **Accounting for Variability:** CI takes into account the variability of the data. The mean score alone does not provide information about the spread or uncertainty associated with the estimate.

From the above advantages, by considering CI, decision-makers can have a more comprehensive understanding of the performance of universities. It allows for a more informed decision-making process that takes into account both the mean score and the associated uncertainty.

K-Mean Clustering Method

K-Means Clustering is an unsupervised machine learning algorithm used to partition a dataset into K distinct, non-overlapping subsets (clusters). The goal is to minimize the within-cluster variance, which is the sum of squared distances between each data point and the centroid of its cluster. K-means cluster analysis method is utilized to determine the optimal number of clusters with Silhouette analysis (Wang, Franco-Penya, Kelleher, Pugh, & Ross, 2017) and to validate the defined clusters (Aldenderfer & Blashfield, 1984; Everitt, Landau, & Leese, 2001, Kaufman & Rousseeuw, 2009). When applying K-Means clustering to categorize universities based on their performance metrics, it is crucial to consider not only the average values but also the confidence intervals of these metrics. Confidence intervals provide a range within which the true mean is expected to lie with a certain probability (95%). Including CIs in clustering helps account for the uncertainty in the data, making the clusters more robust. Therefore, universities with similar means but different confidence intervals might reflect different levels of variability in their performance. Clustering with CIs helps to differentiate between consistently high-performing universities and those with more variability.

The K-means clustering method requires the following calculation steps:

1- Computing the Euclidean distance between each data point and centroid:

$$d(X_i, \mu_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - \mu_{jk})^2}$$

where μ_{jk} and x_{ik} are the *k-th* features of data point *i* and centroid *j*, respectively.

2- Assigning each data point to the cluster with the nearest centroid:

$$C_i = argmin_j d(X_i, \mu_j)$$

where the C_i is the cluster assignment for data point x_i.

3- Updating the centroid, the centroid of each cluster is recalculated using the following formula:

$$\mu_j = \frac{1}{|C_i|} \sqrt{\sum_{X_i e C_i} X_i}$$

where $|C_i|$ is the number of points in cluster j.

The importance of incorporating confidence intervals into K-means clustering can significantly enhance the robustness and reliability of the clusters formed. This method involves using confidence intervals to assess the stability of clusters by repeating the clustering process multiple times. This approach provides a measure of the stability and reliability of the clusters formed (de Jong et al., 2019; Li et al., 2023). Moreover, the clustering analysis considers multiple dimensions of performance, including mean scores, confidence intervals of the mean scores, and pass rates ensuring a multifaceted evaluation.

Methodology

For this research project, CI for the mean scores of the Saudi Medical Licensure Exam (SMLE) for each university is calculated for the entire data to assess the precision of the mean exam scores for each university (Altman & Bland, 2011). A chosen CI level of 95% is used for the analysis, which corresponds to a significant level of 0.05. Then, the results obtained from the entire data are compared to determine the best-ranking category intervals.

The k-means clustering analysis method is utilized to determine the optimal number of clusters with Silhouette analysis (Wang, Franco-Penya, Kelleher, Pugh, & Ross, 2017) and to validate the defined clusters (Aldenderfer & Blashfield, 1984; Everitt, Landau, & Leese, 2001, Kaufman & Rousseeuw, 2009). For this purpose, mean scores, confidence intervals of the mean scores, and pass-rate variables were used for clustering analyses. The analysis of variance (ANOVA) is performed to determine if there is a significant difference in exam scores among the defined ranking categories based on K-means clustering analysis results (Howell, 2012; Tabachnick & Fidell, 2013). Finally, the defined criterion based on cluster analyses for each category is applied to SMLE data to determine the ranking categories for each university.

Sample of this report

The descriptive statistics of SMLE total scores for the past academic years are reviewed. To capture a comprehensive performance, each year from May to June is outlined, and the past three years of data are chosen to be utilized to mitigate the impact of COVID-19 on exam results. Moreover, universities with less than 5 students were excluded from the analysis to exclude universities with very large uncertainty.

Table 1. Descriptive statistics of total scores for SMLE across the year

Exam Year	N	Mean	Standard Deviation
2021-2022	3919	629.178	68.252
2022-2023	11358	651.779	69.944
2023-2024	2938	612.873	74.423

Based on the statistics provided in Table 1 for the SMLE, it is evident that there are noticeable fluctuations in the mean scores across the academic years. However, despite these differences, the variations do not significantly impact the integrity of the entire dataset. Therefore, despite these year-on-year variations, the overall data remains viable for a comprehensive analysis.

We suggest conducting these analyses every three-year cycle to evaluate the performance of each university with an adequate number of candidates and enough time to capture changes. This cyclical analysis will help in understanding long-term trends and making informed decisions to improve educational strategies and outcomes.

Results

The results of the analysis based on the provided methodology groups universities based on their overall performance metrics rather than fixed cut-scores, offering a more comprehensive view of how institutions perform relative to one another. Table 2. Present the ranking categories obtained A, B, C, D, and E for total mean scores. These categories provide a classification of the universities based on their confidence interval estimates and comparisons. Universities with higher mean scores are assigned to higher categories, indicating better performance.

Table 2. Criteria for each ranking category based on analysis (SMLE)

Category	Mean Score Range
Category A	670 and above
Category B	650 - 669
Category C	630 - 649
Category D	610 - 629
Category E	Below 610

Table 3. Distribution of Universities across the defined ranking categories based on analysis

	Entire Data (new)	
Ranking Category	Total Number	Percentage (%)
Category A	3	9.4%
Category B	7	21.9%
Category C	11	34.4%
Category D	6	18.8%
Category E	5	15.6%
Total	32	100.00%

Categories Interpretation

The interpretation of the proposed ranking categories is given below:

- **Category A:** Universities in this category have a mean score of 670 and above. Universities in this category are considered to be among the top-performing institutions in the period of conducting this report.
- **Category B:** Universities in this category have a mean score ranging from 650 to 669. Universities in this category exhibit above-average performance and are recognized for their strong performance in the exam.
- **Category C:** Universities in this category have a mean score ranging from 630 to 649. Universities in this category demonstrate satisfactory performance.
- **Category D:** Universities in this category have a mean score ranging from 610 to 629. Universities in this category exhibit a moderate level of performance indicating some areas of improvement. These universities may have room for enhancement in specific domains covered by the exam.
- **Category E:** Universities in this category have a mean score below 610. Universities in this category demonstrate lower performance among the universities.

These categories allow for a standardized comparison of university performances based on performance metrics used. They provide a clear understanding of the relative performance levels among different institutions, aiding in decision-making processes, such as resource allocation, funding, and policy changes, within the educational system.

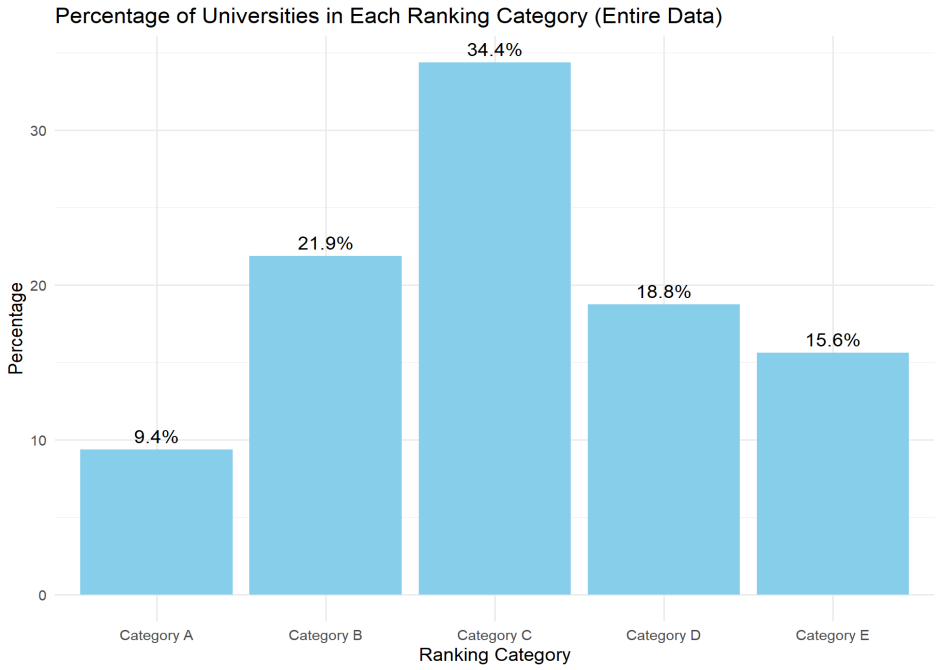


Figure 1. SMLE Distribution of Universities across the defined categories

After determining the ranking categories, the ANOVA is performed to determine if there is a significant difference in exam scores (dependent variable) among the ranking categories (independent variable). The ANOVA results of multiple comparisons based on the least significant difference (LSD) method (Sauder, & DeMars, 2019) show a significant difference between the following Ranking Categories in mean scores, suggesting that the different categories have a significant impact on the performance of the students with a very low p-value of .000 ($p < .05$).

Table 4. ANOVA Multiple Comparison Tests Results

		Mean Difference			95% Confidence Interval	
(I) Rank	(J) Rank	(I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Category A	Category B	16.981*	5.655	.000	5.376	28.586
Category B	Category C	22.152*	4.159	.000	14.022	30.284
Category C	Category D	17.033*	4.413	.000	8.498	25.568
Category D	Category E	26.808*	4.963	.000	16.625	36.991

*. The mean difference is significant at the 0.05 level.

Based on the Silhouette Scores, it appears that the dataset could be best clustered into 5 clusters since Cluster 5 has the highest Silhouette Score with a lower number of clusters, indicating the strongest clustering structure. These results are aligned with the confidence interval results. On the other hand, alignment between the confidence-interval based method and clustering analysis method appears to be greater than 82%. Figure 2 and 3 depict the results of a clustering analysis by showing how universities are grouped across different defined categories. It provides a snapshot of the distribution and aid in understanding the university ranking categories based on certain criteria.

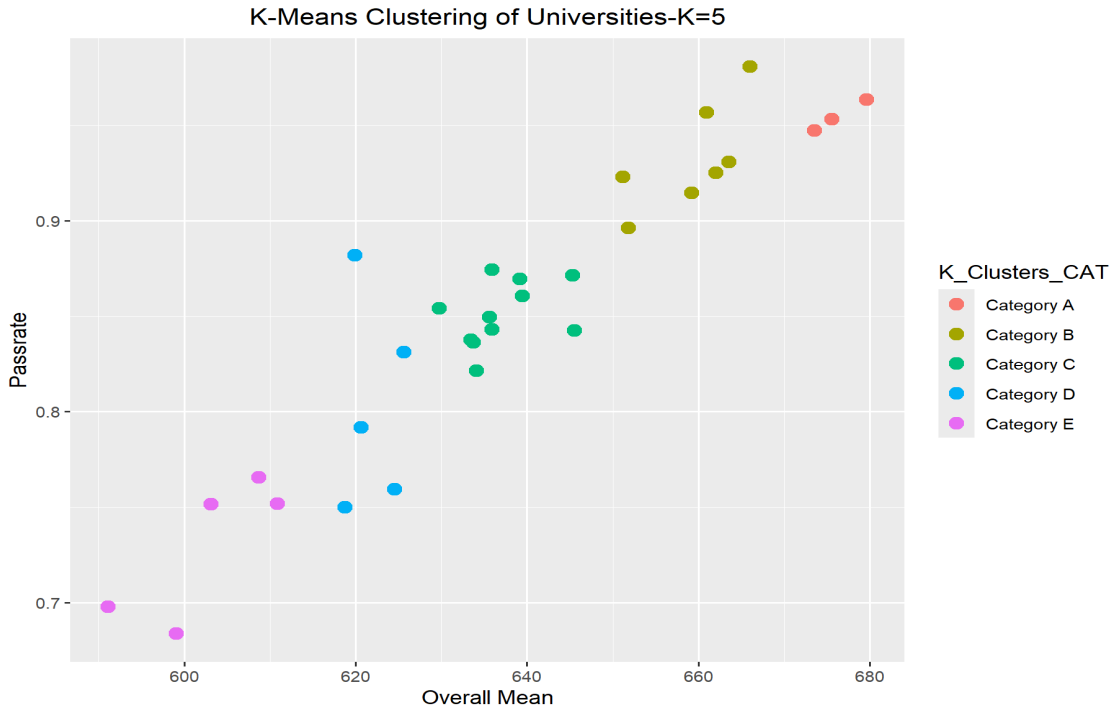


Figure 2. Distribution of universities across the defined categories (Pass rate vs Overall Mean Score)

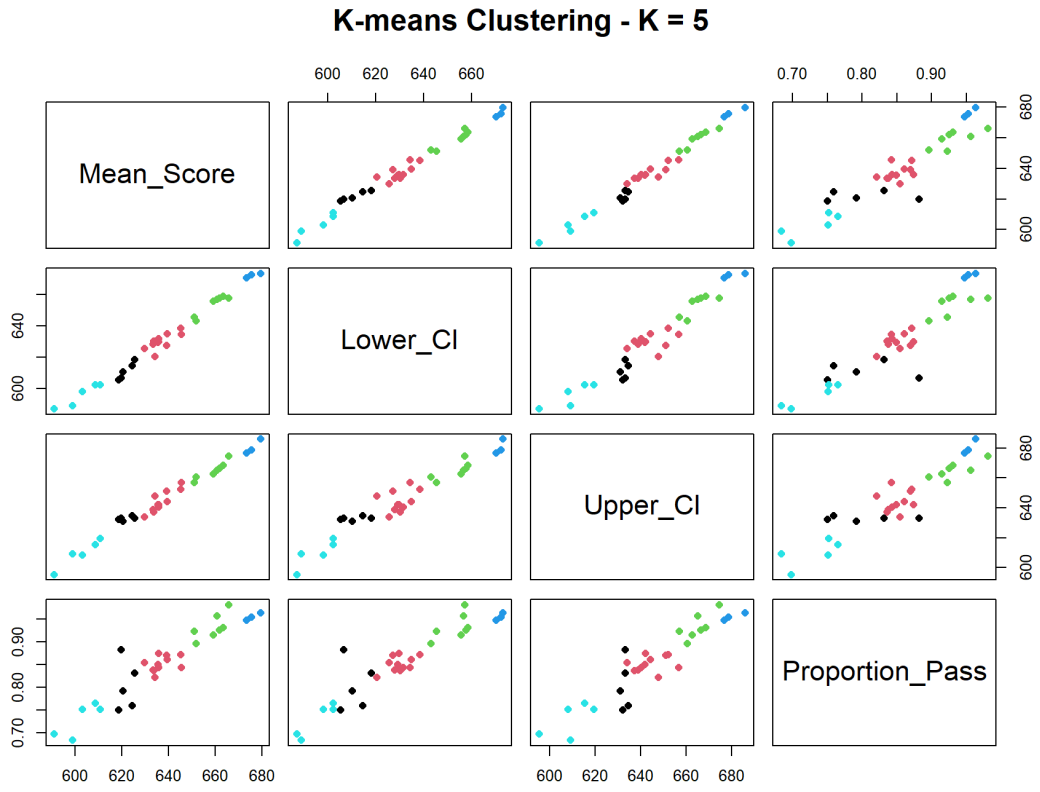


Figure 3. Distribution of universities across the defined categories based on Clustering Analysis

Considering the results provided in this report, there are several arguments to support using total score-based ranking criteria, supplemented by confidence intervals, alongside pass-rate-based criteria in clustering analysis methods:

Comprehensive Assessment: Total score-based ranking criteria provide a comprehensive assessment of university performance. By integrating confidence intervals of total scores and pass rates in k-means clustering, we can capture the actual scores achieved by students more accurately, allowing for better differentiation between universities. This method captures the variability in scores within each university and encourages the pursuit of excellence and the attainment of high-performance levels.

Enhanced Accuracy and Reliability: Confidence intervals offer a statistical measure of the precision of the scores. By incorporating these intervals into k-means clustering, the accuracy and reliability of the clustering results are significantly increased. This approach ensures that the clusters formed reflect the true performance trends of the universities, accounting for the variability and uncertainty in the data.

Detailed Performance Evaluation: Pass-rate-based criteria may not capture the full range of performance and can be limited in evaluating the abilities and knowledge levels of students. Total scores, supplemented by confidence intervals, provide a more detailed and nuanced evaluation of student performance, highlighting not just whether students pass but how well they perform.

Therefore, total score-based ranking criteria, determined by the confidence interval method, were used for this report to offer a more comprehensive, precise, and aligned approach for assessing universities' performance. This integration ensures that the analysis is statistically robust and reflective of the true academic capabilities of the institutions.

Potential Confounding Factors

When defining ranking categories, it is important to ensure that other relevant factors that may influence Exam scores are taken into account. Some potential confounding factors to consider in this analysis could include:

1. **Student demographics:** Variations in student demographics such as age, gender, socioeconomic status, or cultural background could affect Exam scores.
2. **Teaching quality and resources:** Differences in teaching quality, resources, faculty qualifications, and student-to-teacher ratios among universities can impact Exam scores.
3. **Student aptitude and preparation:** Variation in students' prior academic performance, aptitude, and preparation for the Exam may affect their scores.
4. **Curriculum and course offerings:** Universities may have different curricula, course offerings, or areas of specialization, which can influence Exam scores.
5. **The number of exam takers:** The number of exam takers can impact the margin of error for universities with a low sample size. Universities with fewer exam takers may have wider confidence intervals, resulting in less precise estimations of their performance.
6. **First attempt analysis**: mean scores were calculated twice for the entire dataset and the first attempt dataset. The comparison of the results obtained from these two datasets showed similar findings in the period of this report. Therefore, the first attempt results were excluded from the report because the entire dataset is more comprehensive and reflects the performance of the universities more accurately.
7. **Quality control process**: A rigorous QC process helps identify and address data anomalies, errors, and inconsistencies, minimizing biases and enhancing the accuracy of the results. By verifying data accuracy, assessing quality indicators, and ensuring consistency, researchers can enhance the robustness of their analysis and accurately reflect the relationship between ranking categories and exam scores.

Conclusion

In this report a comprehensive analysis is achieved by using K-Means clustering to categorize universities into distinct clusters based on their performance metrics. This categorization highlights patterns and similarities among universities, facilitating a deeper understanding of the data. Clustering analysis not only helps in identifying high-performing and low-performing groups but also reveals intermediate categories, providing a nuanced view of university performance (Elbawad, 2022).

Methodologically, the incorporation of confidence intervals into the clustering process strengthens the evaluation by ensuring that the rankings reflect both the central tendency and the variability in the data. Ranking categories determined by K-Means clustering provide an objective method to classify universities, reducing subjective biases in assessment and ensuring a standardized evaluation framework. This helps in differentiating between universities with similar mean scores but different levels of performance consistency.

The implementation of ANOVA stands as a cornerstone in our evaluation, allowing us to ascertain if substantial differences are present among the scores across different university categories. The value of ANOVA lies not just in confirming our findings but also in highlighting specific areas where disparities might exist. Furthermore, enriching our analytical approach, cluster analysis serves as an instrumental tool. The techniques and principles of cluster analysis are essential in understanding patterns within datasets (Everitt, Landau, Leese, & Stahl, 2011), presenting a structured way to uncover and understand prevalent patterns and validate the results of the proposed method.

By considering the mean, lower limit, and upper limit of each university's performance, we ensure a fair comparison and a clearer picture of university effectiveness. This methodology contributes to a more comprehensive and reliable assessment of universities' effectiveness in preparing students for the SMLE, rather than just relying solely on pass rates (Cumming, 2014; Kline, 2013).

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills, CA: Sage.

Altman, D. G., & Bland, J. M. (2011). How to obtain the confidence interval from a P value. *BMJ*, 343, d2090.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7-29.

de Jong, J., Emon, M. A., Wu, P., Karki, R., Sood, M., Godard, P., Ahmad, A., Vrooman, H., Hofmann-Apitius, M., & Fröhlich, H. (2019). Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience*, *10*(1), Article giaa028. https://doi.org/10.1093/gigascience/giaa028

Elbawab, R. (2022). University Rankings and Goals: A Cluster Analysis. *Economies*, *10*, 209. https://doi.org/10.3390/economies10090209

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis*. London, UK: Arnold.

Howell, D. C. (2012). *Statistical methods for psychology*. Cengage Learning.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: An introduction to cluster analysis (Vol. 344). *John Wiley & Sons*.

Kline, R. B. (2013). *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences* (2nd ed.). Washington DC: American Psychological Association.

Li, B., Peng, C., You, Z., Zhang, X., & Zhang, S. (2023). Single-cell RNA-sequencing data clustering using variational graph attention auto-encoder with self-supervised learning. *Briefings in Bioinformatics*, *24*(6), Article bbad383. https://doi.org/10.1093/bib/bbad383

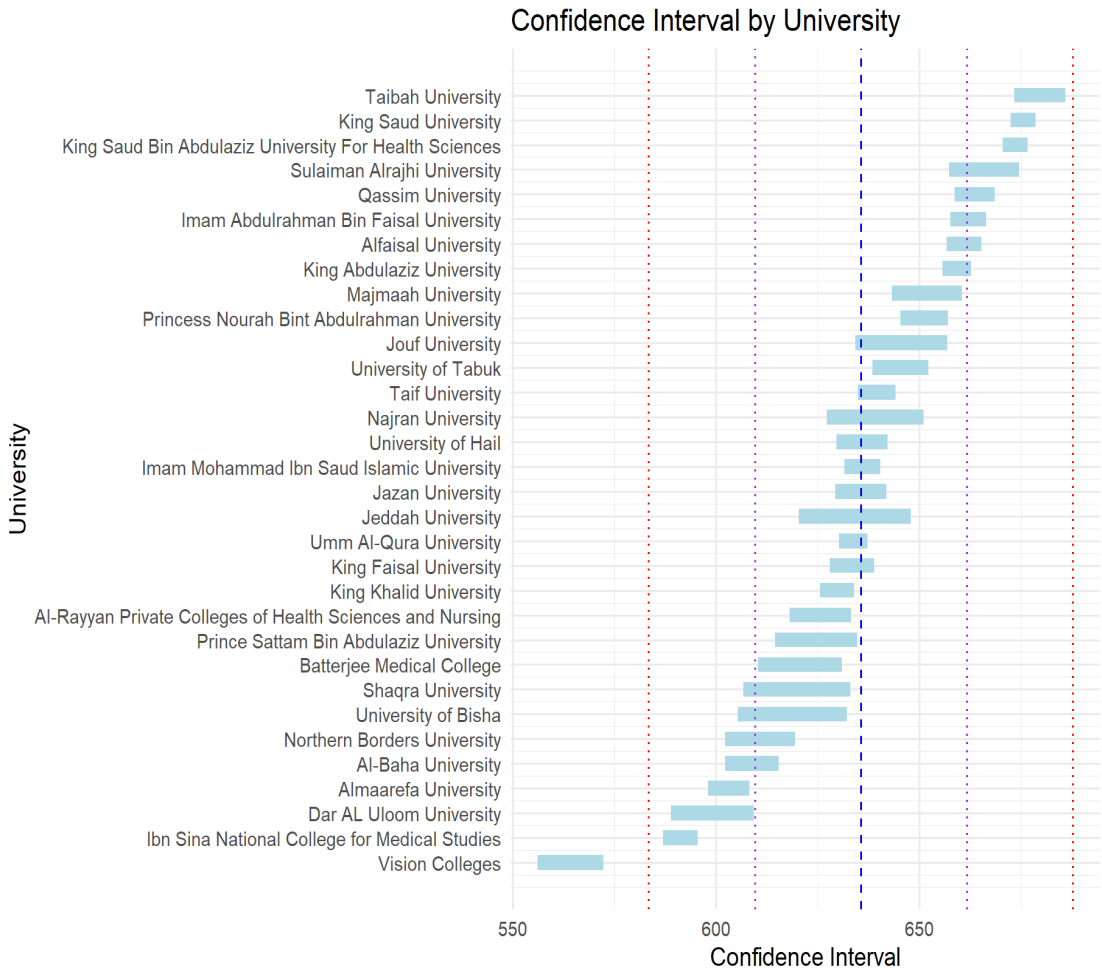
Sauder, D. C., & DeMars, C. E. (2019). An updated recommendation for multiple comparisons. *Advances in Methods and Practices in Psychological Science*, *2*(1), 26–44. https://doi.org/10.1177/2515245918808784

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Boston, MA: Pearson.

Wang, F., Franco-Penya, HH., Kelleher, J.D., Pugh, J., Ross, R. (2017). *An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity*. In: Perner, P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2017. Lecture Notes in Computer Science, vol 10358. Springer, Cham. https://doi.org/10.1007/978-3-319-62416-7_21

Appendix

Appendix A. CI Estimate by University for SMLE



The above appendix illustrates the confidence intervals for universities, ordered by their mean values in descending order. Each university's name is presented on the y-axis, while the confidence interval, bounded by the lower and upper limits, is denoted by a horizontal blue bar along the x-axis. The overall mean of all the university means is marked with a dashed blue vertical line. Additionally, dotted vertical lines in purple highlight one standard deviation above and below this overall mean, providing context for dispersion. A further two standard deviations above and below the overall mean are demarcated with dotted red vertical lines, giving a broader perspective on the variability of the means in relation to the collective average. The graph allows for a direct comparison of each university's mean value relative to others and offers insights into the dispersion and consistency of the data, particularly in relation to the overall mean and standard deviations.

Appendix B. CI Estimate and Ranking Categories for the Entire Data SMLE

University	N*	Mean	Margin of Error	Confidence Interval		Ranking Category
				Lower Limit	Upper Limit	
Taibah University	356	679.63	6.33	673.30	685.96	Category A
King Saud University	1346	675.55	3.09	672.46	678.64	Category A
King Saud Bin Abdulaziz University For Health Sciences	1439	673.53	3.06	670.47	676.59	Category A
Sulaiman Alrajhi University	155	665.99	8.61	657.37	674.60	Category B
Qassim University	591	663.55	4.95	658.61	668.50	Category B
Imam Abdulrahman Bin Faisal University	733	662.05	4.38	657.67	666.43	Category B
Alfaisal University	553	660.95	4.26	656.69	665.21	Category B
King Abdulaziz University	1370	659.19	3.45	655.74	662.65	Category B
Majmaah University	241	651.84	8.64	643.20	660.49	Category B
Princess Nourah Bint Abdulrahman University	389	651.22	5.85	645.36	657.07	Category B
Jouf University	184	645.57	11.28	634.29	656.84	Category C
University of Tabuk	412	645.34	6.88	638.46	652.22	Category C
Taif University	903	639.50	4.69	634.80	644.19	Category C
Najran University	115	639.15	11.90	627.25	651.05	Category C
University of Hail	422	635.94	6.29	629.65	642.22	Category C
Imam Mohammad Ibn Saud Islamic University	1044	635.93	4.42	631.50	640.35	Category C
Jazan University	505	635.65	6.30	629.34	641.95	Category C
Jeddah University	112	634.14	13.75	620.39	647.89	Category C
Umm Al-Qura University	1685	633.72	3.50	630.22	637.22	Category C
King Faisal University	727	633.46	5.44	628.02	638.90	Category C
King Khalid University	1029	629.73	4.21	625.52	633.94	Category C
Al-Rayyan Private Colleges of Health Sciences and Nursing	314	625.68	7.54	618.13	633.22	Category D
Prince Sattam Bin Abdulaziz University	216	624.58	10.08	614.51	634.66	Category D
Batterjee Medical College	168	620.64	10.36	610.28	631.01	Category D
Shaqra University	110	619.89	13.21	606.68	633.10	Category D
University of Bisha	144	618.75	13.43	605.32	632.18	Category D
Northern Borders University	294	610.88	8.60	602.28	619.48	Category D
Al-Baha University	435	608.74	6.60	602.14	615.35	Category E
Almaarefa University	696	603.14	5.06	598.08	608.20	Category E
Dar AL Uloom University	174	599.07	10.25	588.82	609.32	Category E
Ibn Sina National College for Medical Studies	1046	591.16	4.23	586.93	595.39	Category E
Vision Colleges	303	564.20	8.08	556.11	572.28	Category E

*N represents the total number of attempts in the licensure exam during the measurement years identified in this report.

الهيئة السعودية للتخصصات الصحية
Saudi Commission for Health Specialties

مركز الاتصال 920019393 | @SchsOrg

www.scfhs.org.sa