الهيئة السعودية للتخصصات الصحية
Saudi Commission for Health Specialties

Comparing Universities' Performance In National Licensure Exam Using Ranking Categories.

June 2025

Contents

Introduction

University rankings have emerged as a key tool for institutional performance assessment, guiding prospective students in making learning choices, and shaping policy formulation in higher education systems (Hazelkorn, 2015). In the health education setting, the outcomes of licensing examinations are key quality and effectiveness indicators of academic distinction programs. The Saudi Licensure Examination (SLE) is a standard comparison measure for health care graduates' preparedness to practice in Saudi Arabia (Saudi Commission for Health Specialties (SCFHS, 2024). SLE's scores provide a quantitative and objective value for the performance comparison of health care schools across the country. These assessments typically impact an institution's reputation, financial resource allocation, and capacity to attract a high-caliber student population and successful teaching staff (Marginson & van der Wende, 2007).

The aim of this report is to define ranking categories using machine learning-based K-Means clustering analysis and using confidence interval (CI) estimations based on Saudi Licensure Exams (SLEs) scores data for every Saudi University. CI provides a range of values within which the population parameter will likely fall. Taking into consideration mean scores, lower limit, and upper limit of mean scores representing each university, we established and cross-validated ranking categories for comparisons of university performance effectively. K-means clustering algorithm allows universities to be grouped into separate classes of performance in terms of their overall mean scores, corresponding mean score CI, and pass rates so that separate performance classes can be identified. These clusters were then validated with the use of CI and analysis of variance (ANOVA) methods to ensure that the ranking categories were not just numerical but significant. This convergence of cluster analysis and CI analysis provides a robust foundation for the comparison of university performance on the SLEs while ensuring that the rankings reflect real differences in levels of performance.

The importance of using this method in ranking category designation lies in its ability to provide a clear, data-driven comparison of university performance. Recent studies have focused on the importance of cluster analysis in the academic setting. Certain studies have named the effectiveness of using unsupervised machine learning techniques, such as K-Means clustering, in clustering universities based on performance metrics, which allows for establishing open institutional strategies and goals (Elbawab, 2022).

By clustering universities into performance groups, stakeholders can more effectively identify areas of excellence as well as areas of intervention through statistically significant differences. Such rankings can, in turn, be used to inform policy-making, guide resource allocation, as well as support development of 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 of the 95% CI. The formula for the CI 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 using CI rather than just relying on 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 CI, 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 CI, 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 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 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 CIs of these metrics. CIs 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 CIs 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 CI into K-means clustering can significantly enhance the robustness and reliability of the clusters formed. This method involves using CIs 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, CIs of the mean scores, and pass rates ensuring a multifaceted evaluation.

Methodology

In this report, confidence intervals (CIs) for the mean SPLE scores of each university were calculated using the full dataset to assess the precision of the universities' average exam scores (Altman & Bland, 2011). A chosen CI level of 95% is used for analysis, which is equivalent to a significant level of 0.05.

K-means clustering analysis method is used to determine the appropriate number of clusters by applying Silhouette analysis (Wang, Franco-Penya, Kelleher, Pugh, & Ross, 2017) and to validate the discovered clusters (Aldenderfer & Blashfield, 1984; Everitt, Landau, & Leese, 2001, Kaufman & Rousseeuw, 2009). Mean scores, mean score confidence intervals, and pass-rate variables were utilized to carry out clustering analyses. Analysis of variance (ANOVA) is performed to determine if a significant difference exists in exam results across the established ranking categories from the K-means clustering analysis results (Howell, 2012; Tabachnick & Fidell, 2013). Finally, the criterion established using cluster analyses to each category is employed to select the ranking categories for each university from SPLE data.

Sample of this report

The descriptive statistics of SPLE total scores for the past academic years are reviewed. To capture comprehensive performance, each year from June to May is outlined, and the past three years of data are chosen to be utilized to evaluate the cumulative performance of the universities over the last three years. The report included only test takers who took the exam between June 2022 and May 2025 and who graduated or are in their internship year during the same period.

Moreover, universities with fewer than 5 students were excluded from the analysis to exclude universities with very large uncertainty. Additionally, outlier analysis was conducted to minimize the impact of the outliers on the results. Based on the outlier analysis, 19 cases out of 11428 were detected as outliers and excluded from the analysis.

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

Exam Year	N	Mean	Standard Deviation
2022-2023	3226	543.248	71.407
2023-2024	4085	542.984	74.587
2024-2025	4098	559.337	82.220

Table 1 presents the descriptive statistics of total scores on the SPLE for three academic years, with an increase in mean scores in the last two years. Standard deviations have also increased, though these variations are within acceptable limits and do not compromise the integrity or comparability of the data set. The increasing trend in mean scores can portray candidate performance improvement or education quality enhancement over the years.

It is suggested to conduct these analyses every year using the last three academic year data to evaluate the overall performance of each university with an adequate number of candidates and enough time to capture changes and observe cumulative performance of the universities within the last three years. This 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 Cluster analysis 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 lower limit and upper limit of ranking categories obtained A, B, C, D, and E for total mean scores. These categories provide a classification of the universities based on multiple metrics. Universities with higher mean scores are assigned to higher categories, indicating better performance. Categories are defined based on relative performances of the universities using K-means cluster analysis. The following table provides the lower limit and upper limits of clustered universities.

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

Category	Mean Score Range
Category A	590 and above
Category B	560 - 589
Category C	515 - 559
Category D	499 - 514
Category E	Below 499

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

	Entire Data (new)	
Ranking Category	Total Number	Percentage (%)
Category A	8	26.7%
Category B	8	26.7%
Category C	5	16.7%
Category D	6	20.0%
Category E	3	10.0%
Total	30	100.00%

Table 3 demonstrates the distribution of 30 universities across five traditional ranking categories based on SPLE scores. The majority of the institutions are in the higher-ranking categories, of which Category A holds 26.7% and Category B 26.7%, indicating an upward trend in institutional performance. Category C in the middle has 16.7%, while the lower categories D and E have 20% and 10%, respectively. A total of 53.4% of the universities fall in Categories A and B, and 30% fall in Categories D and E, showing an even better split towards enhanced performance. This shows that a large percentage of institutions are being ranked higher, even though more effort must be continued to help those ranked lower and push system-wide progress.

Categories Interpretation

The interpretation of the proposed ranking categories is given below:

- **Category A:** Universities in this category have a mean score of 590 or more. They are considered to be among the top-performing institutions during the period of conducting this report.
- **Category B:** Universities in this category have a mean score ranging from 589 to 560. They exhibit above-average performance in the exam.
- **Category C:** Universities in this category have a mean score ranging from 559 to 515 and they demonstrate satisfactory performance.

- **Category D:** Universities in this category have a mean score ranging from 514 to 499 and they 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 499, and they 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.

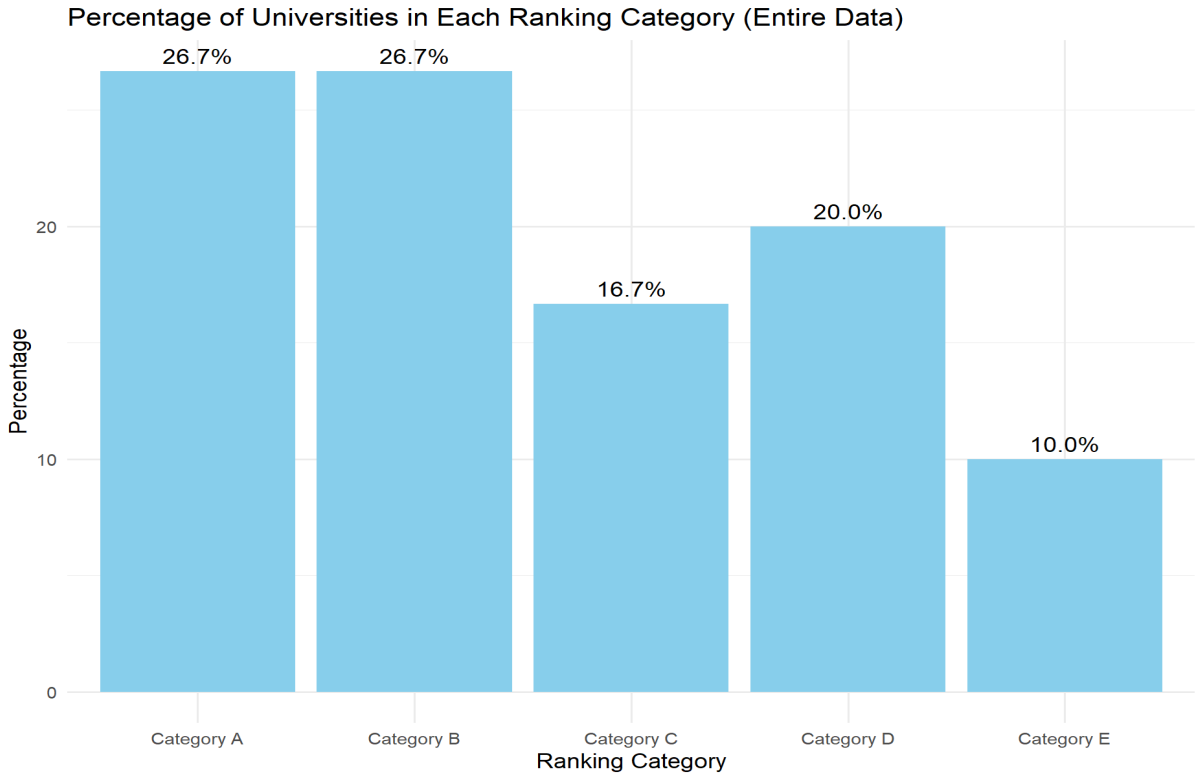


Figure 1. SPLE Distribution of Universities across the defined categories

After determining the ranking categories, the ANOVA is performed to determine if there is a significant difference in exam scores (dependent variable) among the two corresponding 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			

(I) Rank	(J) Rank	(I-J)	Std. Error	t-ration	Sig.
Category A	Category B	30.473*	4.283	7.115	0.000
Category B	Category C	42.883*	4.884	8.781	0.000
Category C	Category D	22.917*	5.187	4.418	0.000
Category D	Category E	32.128*	6.057	5.304	0.000

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

Based on the Silhouette Scores, the data was best clustered into 2 clusters, while also 5 clusters had relatively high Silhouette scores that signifies strong structure of clustering. Also, classification accuracy of cluster analysis with 5 clusters is about 0.95 indicating high validity of clustering results with 5 clusters. On the other hand, due to the high number of samples, university mean scores previously had relatively narrower confidence intervals. Figures 2 and 3 display the result of a clustering analysis by showing how the universities are grouped in different specified categories. It provides a summary of distribution and assists in understanding the categories of universities according to some stated criteria.

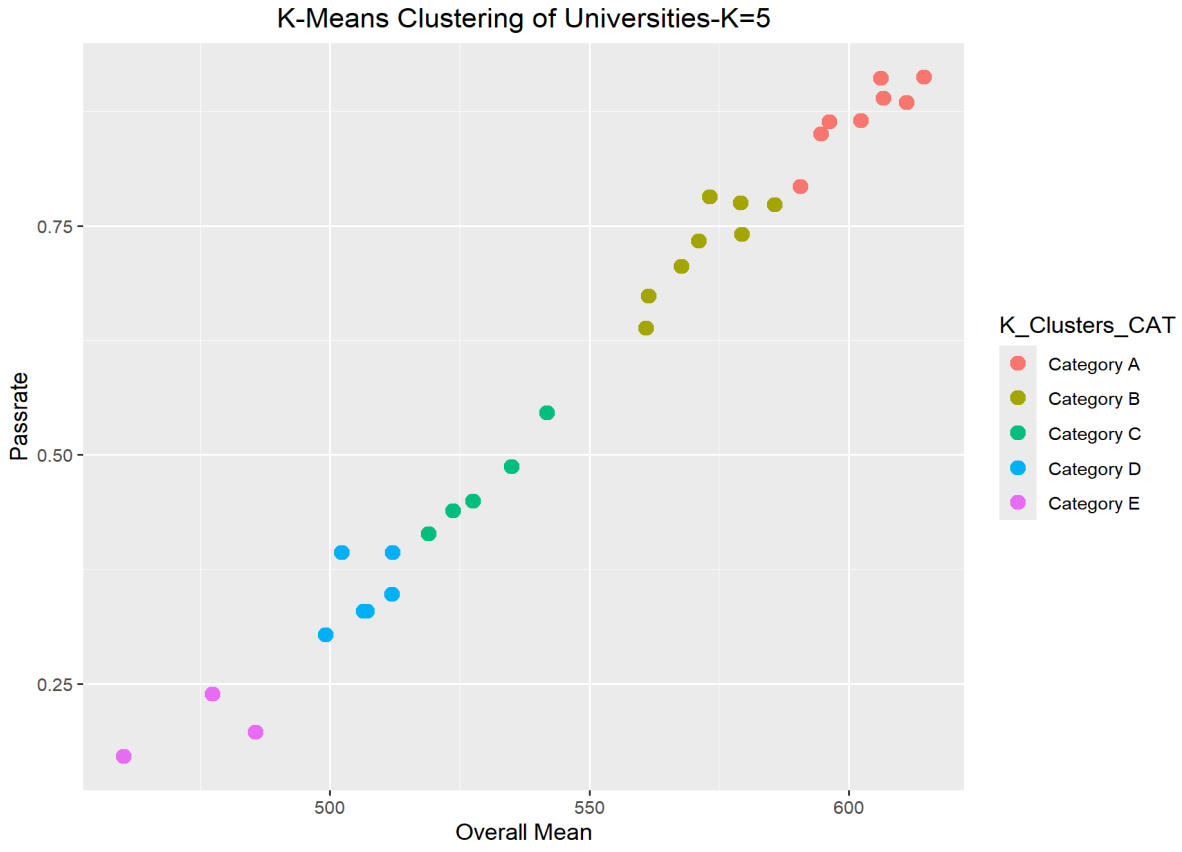


Figure 2. Distribution of universities across the defined categories (Pass rate vs Overall Mean Score)

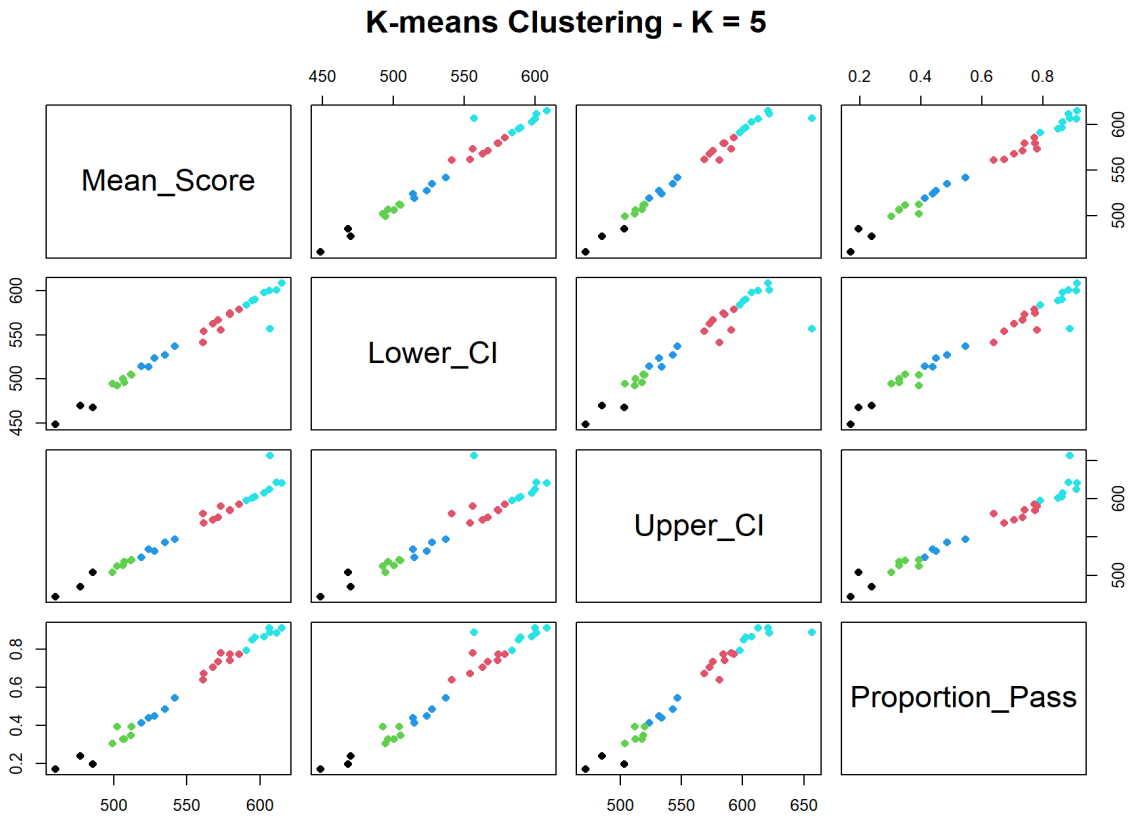


Figure 3. Distribution of universities across the defined categories based on Clustering Analysis

According to the results shown in this report, there are a series of reasons why total score-based rank criteria, supplemented with confidence intervals, are to be applied together with pass-rate-based criteria in clustering analysis methods:

Comprehensive Assessment: Total score-based ranking criteria enable an overall measurement of university performance. With the inclusion of total score confidence intervals and pass rates in K-means clustering, we can measure more precisely the actual scores achieved by students, which allows us to distinguish between universities more meaningfully. The strategy identifies the difference in scores among each university and encourages seeking excellence and accomplishment at high performance levels.

Enhanced Accuracy and Reliability: Confidence intervals offer a statistical approximation of the accuracy of the scores. The addition of the confidence intervals to k-means clustering ensures the accuracy and reliability of the clustering are very high. The use of this method guarantees that the resulting clusters reflect the actual performance trends of the universities, considering the uncertainty and variability of the data.

Detailed Performance Evaluation: Pass-rate based measurements do not indicate the whole range of performance and are poor in evaluating the strengths and the level of knowledge among the students. Total scores, supported by confidence intervals, provide a better detailed and more precise evaluation of student

performance, not just showing whether or not students pass but also with what performance.

Therefore, score-based ranking metrics for the overall score were used for this report to offer a more inclusive, precise, and harmonized mechanism for assessing universities' performance. The inclusion ensures that the examination is statistically valid and reflective of the institutions' actual academic capabilities.

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 utilizing K-Means clustering to group universities into various clusters based on their performance metrics. This grouping achieves trends and similarities between universities and provides a clear insight into the information. Clustering analysis not only helps identifying top-performing and low-performing classes but also identifies intermediate classes, providing a more intricate insight into university performance (Elbawad, 2022).

Methodologically, incorporating confidence intervals during the clustering process reinforces the evaluation with the assurance that the rankings on both central tendency and variability in data are accomplished. Ranking categories defined through K-Means clustering provide an objective means for classifying universities, exempt from subjectively biased assessment and inclusive of a uniform evaluation framework. This allows for discrimination among universities with similar mean scores but different consistency of performance.

The use of ANOVA is a cornerstone in our analysis which allows us to ascertain if differences are significant between the scores in different categories of universities. ANOVA's biggest strength lies not just in substantiating our findings but in also leading us towards specific places where differences might exist. Apart from filling out our analytical framework, cluster analysis is also a very important tool. The theory and technique of cluster analysis are the basis of pattern understanding within datasets (Everitt, Landau, Leese, & Stahl, 2001), presenting an organized way of finding and understanding general patterns and assessing the results of the supposed approach.

By considering the mean, lower bound, and upper bound of the performance of every university, we are making a comparable examination and having a clearer notion regarding university effectiveness. This method assists towards more valid and comprehensive assessment of the effectiveness of universities in preparing students for the SPLE, rather than solely employing 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

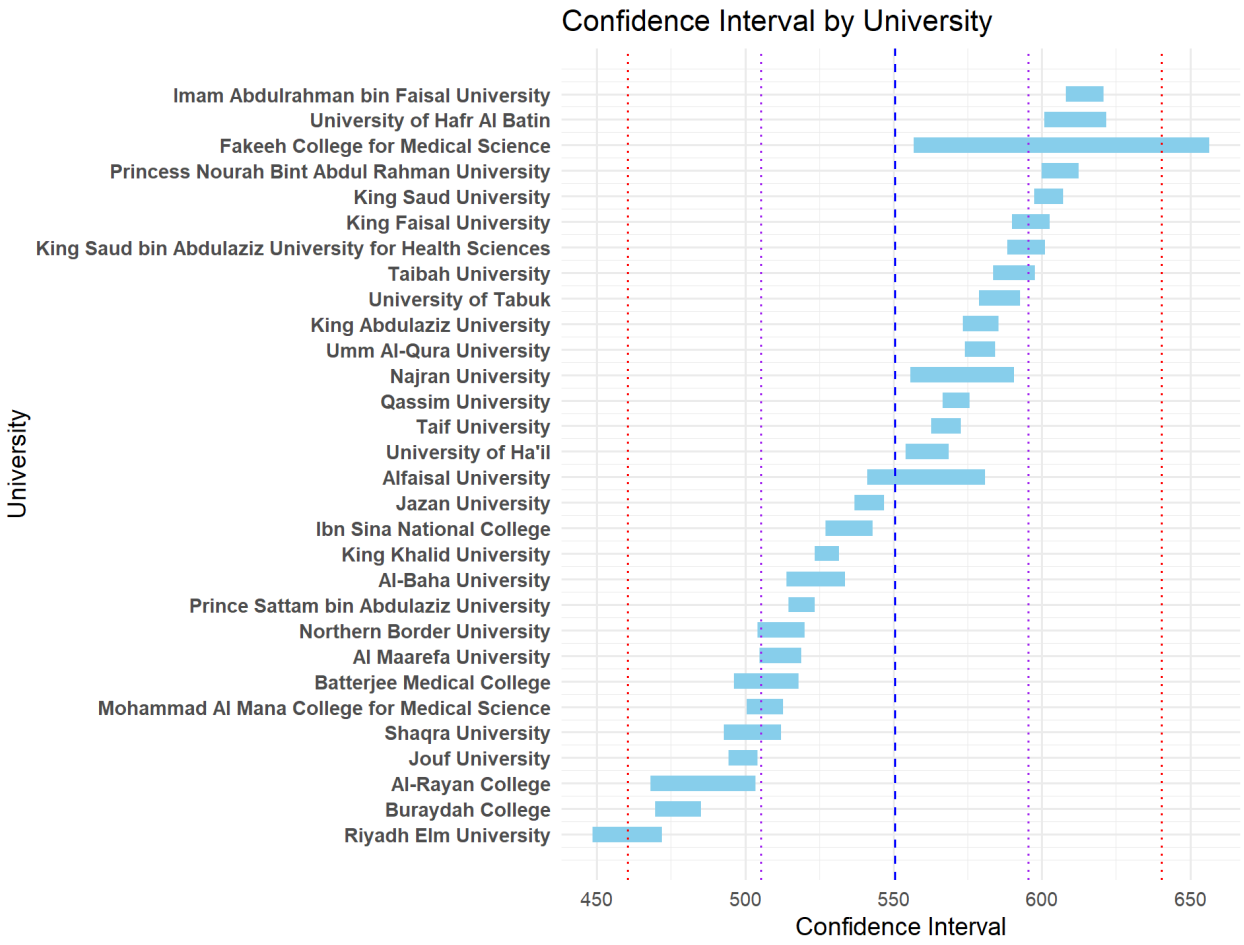
Sauder, D. C., & DeMars, C. E. (2019). An updated recommendation for multiple comparisons. *Advances in Methods and Practices in Psychological Science, 2*(1), 26–44. https://doi.org/10.1177/2515245918808784

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 SPLE



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 SPLE

University	N*	Mean	Margin of Error	Confidence Interval		Ranking Category
				Lower Limit	Upper Limit	
Imam Abdulrahman bin Faisal University	296	614.53	6.35	608.18	620.87	Category A
University of Hafr Al Batin	130	611.23	10.48	600.75	621.71	Category A
Fakeeh College for Medical Science	9	606.67	49.87	556.79	656.54	Category A
Princess Nourah Bint Abdul Rahman University	235	606.21	6.31	599.91	612.52	Category A
King Saud University	612	602.42	4.89	597.53	607.31	Category A
King Faisal University	307	596.27	6.36	589.91	602.64	Category A
King Saud bin Abdulaziz University for Health Sciences	294	594.71	6.33	588.39	601.04	Category A
Taibah University	342	590.68	7.01	583.67	597.70	Category A
University of Tabuk	286	585.74	7.01	578.73	592.75	Category B
King Abdulaziz University	512	579.38	6.01	573.37	585.39	Category B
Umm Al-Qura University	604	579.25	5.12	574.13	584.37	Category B
Najran University	64	573.22	17.47	555.75	590.69	Category B
Qassim University	742	571.13	4.59	566.54	575.73	Category B
Taif University	621	567.75	5.05	562.70	572.80	Category B
University of Ha'il	361	561.45	7.25	554.19	568.70	Category B
Alfaisal University	47	561.02	19.93	541.10	580.95	Category B
Jazan University	777	541.87	5.05	536.82	546.92	Category C
Ibn Sina National College	263	535.04	7.97	527.08	543.01	Category C
King Khalid University	1214	527.63	4.08	523.55	531.71	Category C
Al-Baha University	269	523.82	9.83	513.99	533.65	Category C
Prince Sattam bin Abdulaziz University	849	519.06	4.42	514.64	523.48	Category C
Northern Border University	361	512.12	7.88	504.23	520.00	Category D
Al Maarefa University	279	511.96	7.01	504.94	518.97	Category D
Batterjee Medical College	149	507.13	10.91	496.21	518.04	Category D
Mohammed Al-Mana College for Medical Sciences	371	506.57	6.12	500.45	512.68	Category D
Shaqra University	341	502.38	9.64	492.74	512.02	Category D
Jouf University	547	499.27	4.88	494.38	504.15	Category D
Al-Rayan College	61	485.74	17.73	468.01	503.47	Category E
Buraydah College	302	477.35	7.66	469.69	485.01	Category E
Riyadh Elm University	164	460.23	11.63	448.60	471.86	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