# Comprehensive Evaluation of University Ranking on PCA and Factor Analysis

Wang Yi, Wu Yicheng, Wang Yijie

May 29, 2023

## 1 Introduction

### 1.1 Background Introduction of the problem

University ranking is a process by which universities are evaluated and ranked according to a series of indicators. While university rankings may vary from institution to institution and country to country, they are usually evaluated based on factors such as academic reputation, teaching quality, research output, level of internationalization, faculty-student ratio, alumni employment and so on.

The university rankings can provide students with a reference for choosing a university, assess the academic level of a university, enhance competitiveness of universities and assist funders and donors in decision-making. Therefore, university rankings are of great significance to universities, students and society.

### 1.2 Dataset Description

We obtain the data from the QS World University Rankings website and use the data in 2023. And the indicators are as follows:

- Academic Reputation - the results are based on the responses to a survey distributed worldwide academics from a number of different sources. It not only illuminates the quality of the research, but the strength of the university in communicating that research, and the strength of the impact the research makes across the world. In the overall ranking, academic reputation carries the most weight at 40 %.
- Employer Reputation - the results are based on the responses to a survey distributed worldwide employers from a number of different sources. It is unique amongst current international evaluations in taking into consideration the important component of employment. In the overall ranking, employer reputation carries the second most weight at 10 %.
- Faculty/Student Ratio - the ratio of faculty members to students. It is a measure of the learning and teaching environment of the institution. In the overall ranking, faculty/student ratio carries the third most weight at 20 %.

- Citations per Faculty - the number of citations per faculty member. It is a measure of relative intensity and volume of research being done at an institute,taking into account institute size. In the overall ranking, citations per faculty carries the fourth most weight at 20 %.
- International Faculty Ratio - the proportion of faculty members that are international. It is a measure of the international diversity of the faculty. In the overall ranking, international faculty ratio carries the fifth most weight at 5 %.
- International Student Ratio - the proportion of students that are international. It is a measure of the international diversity of the student body. In the overall ranking, international student ratio carries the sixth most weight at 5 %.

# 2 Description of the methods

## 2.1 Principle Component Analysis

### 2.1.1 Introduction to Principal Component Analysis

Principal Component Analysis is a statistical dimension reduction method, and its basic idea is to use a few mutually independent components to reflect the vast majority of information of original variables. Suppose that $\mathbf{x} = (x_1, x_2, ..., x_p)'$ is a p−dimensional random vector with mean $\mu = (\mu_1, \mu_2, ..., \mu_p)'$ and covariance matrix $\Sigma = (\sigma_{ij})$, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\Sigma$, and $\mathbf{t_1}, \mathbf{t_2}, \cdots, \mathbf{t_p}$ be the corresponding eigenvectors of $\Sigma$, then the j-th population principal components of $\mathbf{x}$ x, denoted by $f_j$ , can be formulated as

$$f_j = \mathbf{t_j}'\mathbf{x}$$

, where $Var(f_j) = \lambda_j$ for j = 1, 2, ..., p. In the above definition, principal components $f_j$ 's are mutually uncorrelated. It is common practice to use the cumulative contribution rate $\sum_{j=1}^{m} \lambda_j / \sum_{j=1}^{p} \lambda_j$ to measure the degree to which the total variance is explained, where m represents the number of principal components. For a positive definite matrix, the size of its trace is often determined by a few large eigenvalues. Thus, only a few principal components are selected to make the cumulative contribution rate reach a relatively high level (e.g. 85%), so as to achieve the purpose of dimensionality reduction. Let $\mathbf{f} = (f_1, f_2, ..., f_p)'$ be the principal vector composed of all principal components. Then we can express it as

$$\mathbf{f} = \mathbf{T}'\mathbf{x}$$

in matrix form, where the coefficient matrix $\mathbf{T} = (t_1, t_2, ..., t_p)$ is a positive definite matrix.

### 2.1.2 Data Processing

Suppose that there are n universities to be evaluated, each university has p indexes, and the observed indexes of the i-th university are $x_{i1}, x_{i2}, \ldots, x_{ip}$, $i = 1, 2, \ldots, n$, then the observation

data matrix,$\mathbf{X} = (x_{ij})_{p\times n}$, is expressed as

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{11} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix}$$

Let $\bar{x_j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$, $j = 1, 2, \ldots, p$ be the sample mean of the j-th index of these n universities, $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X_j})^2$ be the sample variance. In order to ensure the unity of data magnitude, we standardize the original data by $z_{ij} = \frac{x_{ij}-\bar{x_j}}{\sqrt{s_j}}$. It is obvious that the covariance matrix of the standardized sample is exactly the sample correlation matrix of original data:

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

, where $r_{kl} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x_k})(x_{il} - \bar{x_l})/s_k s_l$, $(k = 1, 2, \ldots, p,\ l = 1, 2, \ldots, p)$ represents the correlation between the k-th and l-th samples with $r_{kk} = 1$.

### 2.1.3 Estimation of Coefficient Matrix

Starting from the correlation matrix $\mathbf{R}$, we solve the characteristic equation $|\lambda\mathbf{I} - \mathbf{R}| = 0$ to obtain the eigenvalues and their corresponding eigenvectors, denoted as $\hat{\lambda_1} \geq \hat{\lambda_2} \geq \cdots \geq \hat{\lambda_p}$, $\hat{\mathbf{t_1}}, \hat{\mathbf{t_2}}, \ldots \hat{\mathbf{t_p}}$ respectively. Then, the j-th sample principal components $\hat{f_j}$ has the following expression

$$\hat{f_j} = \hat{\mathbf{t_j}}'\mathbf{z}$$

, for $j = 1, 2, \ldots, p$), where $\mathbf{z}$ is the normalized vector of each component. And the sample principal vector $\hat{\mathbf{f}} = (\hat{f_1}, \hat{f_2}, \ldots, \hat{f_p})'$ is expressed as

$$\hat{\mathbf{f}} = \hat{\mathbf{T}}'\mathbf{z}$$

, where $\hat{\mathbf{T}} = (\widehat{\mathbf{t_1}}, \widehat{\mathbf{t_2}}, \ldots, \widehat{\mathbf{t_p}})$ is the estimated coefficient matrix.

### 2.1.4 Comprehensive Evaluation Strategy

Given the predefined threshold $\alpha_0$ , when the cumulative contribution rate $\sum_{j=1}^{m}\hat{\lambda_j}/\sum_{j=1}^{p}\hat{\lambda_j} \geq \alpha_0$, m is selected as the number of principal components. Based on the first m sample principal components, we construct a comprehensive evaluation function as

$$F_z = \sum_{l=1}^{m}(\frac{\lambda_l}{\kappa})\hat{f_l} = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_p x_p$$

with $\kappa = \lambda_1 + \lambda_2 + \cdots + \lambda_m$, where $\frac{\lambda_l}{\kappa}$ represents the contribution rate of the l-th principal component $\hat{f_l}$ and $\omega_j$ is the original weight of index $x_j$ . Note we need to further normalize the

weights to satisfy $\sum_{j=1}^{p} \omega_j = 1$. Finally, the comprehensive score (C-Score) of target university is then calculated as

$$C - Score = \sum_{j=1}^{p} \omega_j y_j$$

where $y_j$ is the score of the j-th evaluation index in the target university according to the scoring standards, whose value is between 0 and 5. The scoring standards are determined by how each university ranks in individual categories, with 5 points awarded to those in the top 20%, 4 points to those in the 20%-40% range, and so on, with only 1 point awarded to those in the bottom 20%.

## 2.2 Factor Analysis

### 2.2.1 Factor Extraction

We use the factor analysis to extract the common factors from the indicators. The factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. For example, it is possible that variations in six observed variables mainly reflect the variations in two unobserved variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors, plus "error" terms. Factor analysis aims to find independent latent variables. Models with various degrees of independency can be studied.

### 2.2.2 Calculate the score of each university

We use the factor score to represent the score of each university. The factor score is calculated by the following formula:

$$\text{socre}_j = \sum_{i=1}^{m} \frac{v_i}{k} \times Z_{ij}$$

where $v_i = \sum_{i=1}^{p} c_{ij}^2$ is the variance explained by the $i$ the factor, $c_{ij}$ is the $ij$-entry of the loading matrix after rotation, $k = v_1 + v_2 + \cdots + v_m$ and $Z_{ij}$ is the value of the $i$ the factor of the $j$ the university.

The score essentially is the weighted mean of the common factor values and the weight is the variance explained by the factor(or the contribution rate).

## 2.3 Decision Tree

### 2.3.1 Introduction to Decision Tree

A decision tree is a machine learning algorithm used for solving classification and regression problems. It creates a tree-like structure to represent the decision-making process and makes predictions about the target variable based on the input features.

The construction of a decision tree starts with a root node, which represents the initial decision. By evaluating different values of features, the decision tree divides the dataset into

different subsets, each corresponding to a branch or child node. This splitting process is based on selecting the best feature and its corresponding split point to maximize the purity of each subset.

In a decision tree, purity is measured using impurity metrics. Common impurity metrics include Gini impurity and information gain (or gain ratio). Gini impurity measures the probability of misclassifying samples within a subset, while information gain is based on the concept of information entropy and measures the disorder of samples within a subset.

The construction process of a decision tree is repeated recursively until a stopping criterion is met, such as reaching the maximum depth, having a minimum number of samples in a node, or achieving impurity below a predefined threshold.

### 2.3.2 Calculation of Multiple Index Weights by Decision Tree

The weights or importance of variables in a decision tree are typically calculated based on their usage within the tree. Here are two common methods for computing variable weights:

1. Information Gain or Gain Ratio: These metrics are used to select the best splitting points and features in the decision tree. During the construction process, when selecting a splitting point, the importance of a feature is measured by calculating the information gain or gain ratio. Information gain measures how much the uncertainty of the target variable decreases after selecting that feature. Gain ratio normalizes the information gain, taking into account the number of branches for a feature. Higher information gain or gain ratio indicates a greater contribution of the feature to the prediction target, implying a higher weight.

2. Frequency of feature usage: Another approach to calculating variable weights is based on the frequency of feature usage within the decision tree. Once the decision tree is constructed, the count or proportion of each feature used at the splitting nodes can be tallied. Features with higher frequencies can be considered more important since they are used in more decision-making processes within the tree, indicating a larger impact on the prediction outcome.

## 2.4 Cluster Analysis - K-means

### 2.4.1 Introduction of K-means

K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. K-means can be conduct by the following steps:

1. Select $k$ points as the initial cluster centers.

2. Assign each point to the cluster with the nearest center.

3. Update the cluster centers by taking the average of the assigned points.

4. Repeat step 2 and 3 until the cluster centers do not change.

#### 2.4.2 How to choose K

We use the elbow method to select the number of clusters. The elbow method is a method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset. If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best. The "elbow" is the point of inflection on the curve.

And we use the silhouette score to evaluate the clustering result. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). And it can be calculated by the following formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average distance between $i$ and all other data within the same cluster, $b(i)$ is the average distance between $i$ and all other data points in the next nearest cluster. The silhouette ranges from $-1$ to $+1$, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

# 3 Empirical Analysis

## 3.1 Data Acquisition and Description

We divided universities into Top 50, Top 50 100 and After Top 100, and then carried out a simple exploratory analysis (Figure 7).

## 3.2 Analysis steps and results

### 3.2.1 Through Principal Component Analysis

We choose the first 240 universities from the dataset QS World University Rankings which rates the universities' performance in 2023 on a number of dimensions, including academic reputation, employer reputation, faculty student ratio, citations per faculty, international faculty ratio, international students ratio, international research network and employment outcome. We use the first six dimensions as our indicators and standardize the original score data. Then we conduct principal component analysis and calculate the eigenvalues and the corresponding cumulative contribution rate of the correlation matrix. The results are shown in Table 1 and the scree plot is shown in Figure 3.

It can be seen from Table 1: that the cumulative contribution rate of the first four principal components has reached 90%, so we can use the first four principal components to analyze the original six indicators.

From Table 2, we can see that the first principal component has a large loading (in absolute value) on International Student Ratio and International Faculty Ratio. According to the meaning of these indexes, we interpret the first principal component as 'international diversity'.The second

principal component has a large loading on Academic Reputation and Employer Reputation, which are then interpreted as ' overall reputation' . The third principal component has a great Faculty/Student Ratio, which mainly reflects the learning and teaching environment. The fourth principal component has a considerable loading on Citations per Faculty. Therefore, we interpret the fourth principal component as 'overall research level'.

Through the comprehensive evaluation strategy, we can get the weight of each index for the comprehensive score of universities. See Table 3 for details:

Finally, the comprehensive score of each fund based on the evaluation index can be calculated by the following formula:

$$C - Score = \sum_{j=1}^{4} \omega_j y_j$$

, where $w_j$' s values are given in Table 3, and $y_i$ is the score of the university's ith evaluation index, whose value is between 0 and 5. Then, based on the comprehensive scoring and scoring standards, we get a new ranking of the 240 universities, which is shown in Table 4.

From this ranking, we can see that the results obtained through the principal component analysis are not satisfactory, which may be because the weight of the calculated part is negative, thus affecting the effect of the final score. We tried to improve it in a few ways.

### 3.2.2 Factor Analysis Results

First we use the first 240 universities and standardize each variables. And then do the barlett inference and the p-value is extremely small($10^{-70}$), which means that there are correlations among the variables. Next, we need to examine how many factors should be extracted. Here we just do a spectral decomposition of the sample variance matrix and plot the eigenvalues in Figure 4. And we find that the first four eigenvalues can explain large part of the total variance.

So we choose the first four factors. Then we use the principal component method to extract the factors. The loadings in the Table 5. The communalities in the Table 6. And we can see all the communalities approach to 1 which means that the common factor can explain every variable very well. The total variance explained by the four factors is 90.5%.

Then We explain it according to the bigger value of each factor. First factor lands big values on ifr and isr, which means that the first factor is about the international diversity. Second factor have big values on ar and er, which means that the overall reputation. Third factor have big values on fsr, which means that the learning and teaching environment. Fourth factor have big values on cpf, which means the overall research level considering the institute size.

Next we calculate the score of each university using method described in 2.2.2. Based on the new score, we can get the new rank of the universities. We can see the new rank is different from the original rank. The stable institutions(the absolute value of fluctuation is less than 2) are listed on the Table 7.

### 3.2.3 K-means Results

First we use the elbow method to select the number of clusters, here we plot the elbow curve in the Figure 6. We can see that the elbow is at 5, so we choose 5 as the number of clusters.

Then we first cluster the data with the original variables and examine the variable score of cluster center of each group, the result is shown in the Figure 7.a but it is hard to explain the difference, this is because the variables are not independent. So we use the factor score and the PCA score to cluster the data and the result is shown in the Figure 7.b and Figure 7.c. We can see that the difference is more obvious. To start with, we analyze and summarize the results obtained by using factor score. As we can see, the third group has high score on all the factors, and it contains almost top 50 of the original ranking. We can define them as "Comprehensive strong school"

The fourth group has high score on learning and teaching environment and overall research level and relative low score on overall reputation and international diversity, which means that those institutions have strong academic strength and their learning atmosphere is very good, but they need to expand their international visibility, we can define it as "Small minority strong school". For example, the Southern University of Science and Technology is a young university, so it is not well known internationally, but it has a very good learning atmosphere and the overall research level is very high.

And the second group has very high score on international diversity and relative low score on other factors especially learning and teaching environment, which means that those institutions accept a wide range of foreign students and their learning is relatively easy. These institutions are mainly from the UK and Australia. We can define these institutions as "Relaxed International School".

The fifth group has high score on the overall reputation and learning/teaching environment but low score on international diversity and overall research level, which means that their standard of teaching is very high and the students who go out from them are very good. It may be because of location and admissions policy, they have relatively low international diversity and have limited resources to conduct research. And they are mostly from germany, and as we all know, German schools are known for strict teaching. So we can define these kind of institution as "Local Elite School".

The first group has high score on overall reputation and overall research level which means they are well known for their research such as UT Austin, UIUC, and GaTech. We can define these kind of institution as "Research Elite School".

On the other hand, When we observe the results obtained through the principal component analysis, we find that the first group has outstanding performance in all the four principal components. Even if it is slightly inferior in the teaching environment, it can still reflect its strong comprehensive strength, so we name it "all-round strong School".

The second and third groups contain most of the mid-ranking schools, most of which excel in one or two categories but lag behind or are less dominant in others. For example, the second group has a very high reputation and outstanding teaching environment, indicating that they have gained recognition and attention by virtue of their good teaching level and learning atmosphere, but due to some limitations, they can not make progress and breakthrough in scientific research or international, we name them "education strong school". The third group takes advantage of international diversity and teaching environment, but lags behind in research atmosphere and

overall reputation, indicating that they pay attention to student education and international exchange. They are a kind of open and global university, so they can be called "liberal education school".

While the fourth and fifth groups have obvious advantages and disadvantages, they are particularly outstanding in one aspect, but lag behind significantly in other factors. Specifically, the fourth group is far ahead in international diversity, which indicates that these schools attach great importance to the introduction of foreign teachers and the enrollment of foreign students. We name these schools as "international development universities". The fifth group performs well in the overall research level, but neglects the development of other aspects, which reflects their enthusiasm and pursuit for scientific research. We call these schools "high-level research universities".

### 3.2.4 Through Decision Tree

We apply the decision tree algorithm to our data set. First, we used the QS official ranking as the training label and the first 6 scores as independent variables for decision tree fitting of the top 240 university data sets. We get the weights of each variable in the Table 8.

We use the weight $\omega_i$ of the decision tree to get the final result by weighted summation of each score:

$$C_{DT} - Score = \sum_{i=1}^{6} \omega_i X_i$$

Here $X_i$ is i-th score for each university. We can rank the new scores $C_{DT}$ to get the ranking using the weight generated by the decision tree in Table 9.

Of course, the decision tree has a problem, it is easy to overfit. Of course, we can control it by controlling the depth of the decision tree, but adjusting these hyperparameters is also a complicated process.

Next, universities ranked from 0 to 20 are used as the test set, and universities ranked from 20 to 240 are used as the training set to fit the new decision tree. The final results are as Table 10.

We can see that the results are significantly worse than before. This is what happens when an overfit occurs.

- In order to reduce the occurrence of overfitting, we can extract the important information in the original data and discard the noise. As we know, PCA can reduce the dimension of data to extract important information in the original data, and then we will use the data after PCA dimension reduction to fit the decision tree to get the final ranking.

The weights of each principal component are shown in the Table 11 and the results are shown in Figure 8.

We were surprised to find that the final results were completely consistent with the QS official rankings, which showed that PCA effectively extracted effective information from the original data set and discarded useless noise.

# 4 Visualization

## 4.1 Web Data Analysis Platform

We also made a simple web version of the data analysis platform so that users can more easily analyze the data and get results. We conducted exploratory analysis, PCA and factor analysis on the data in the platform and visualized the results (Figure 7).

# 5 Conclusion

1. Principal component analysis can effectively reduce the dimension of the original data, so as to facilitate data visualization and processing. In addition, major features can be extracted from the original data to reduce the influence of redundant information. Moreover, by eigenvalue decomposition of the covariance matrix, the calculation of large-scale data can be transformed into the calculation of a few feature vectors, so as to improve the calculation efficiency. However, the results obtained through principal component analysis are not explanatory enough: the meaning of the interpretation of principal component is generally somewhat fuzzy, not as clear and exact as the meaning of the original variable, which is the price that has to be paid in the process of variable dimension reduction. Principal components extracted from PCA may require additional analysis and interpretation to reach a conclusion, which is often a difficult and even unattainable process.

2. Factor Analysis can actually simplify the variable, it can help reveal the underlying structure and relationships among variables. The factor loading matrix and factor plot can visually display the correlations and groupings of variables.Based on the factor loading matrix, original variables can be transformed into factor scores, allowing for the reconstruction and simplification of variables. However, there are some limitations. For example, Multiple interpretations: Factor analysis may have multiple interpretations. Different factor extraction methods and rotation methods can yield different factor structures, requiring careful selection and interpretation. Also, determining the interpretation of latent factors is not always intuitive and consistent. It requires subjective judgment and theoretical support. Therefore, there may be subjective biases in naming and interpreting the factors.

3. Decision tree seems to have a good performance in multi-index evaluation tasks like this, but its problem is that it needs artificial label data to fit, if there is no label data, the decision tree method will not be used. In addition, decision trees are prone to overfitting problems, and PCA can reduce the occurrence of overfitting to some extent.

# Appendix

# A Tables

## A.1 Principal Component Analysis Result Tables

| Original Variables | Eigenvalue | Contribution Rate | Cumulative Contribution Rate |
|---|---|---|---|
| Academic Reputation | 1.9335 | 0.3222 | 0.3222 |
| Employer Reputation | 1.7456 | 0.2909 | 0.6132 |
| Faculty/Student Ratio | 0.9946 | 0.1658 | 0.7789 |
| Citations per Faculty | 0.7561 | 0.1260 | 0.9050 |
| International Faculty Ratio | 0.0521 | -0.1340 | 0.9570 |
| International Student Ratio | 0.0430 | 0.1722 | 1.0000 |

Table 1: Cumulative Contribution Rate

| Original Variables | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Academic Reputation | -0.5353 | 0.7173 | 0.2202 | -0.0098 |
| Employer Reputation | -0.5298 | 0.7325 | 0.1200 | 0.1785 |
| Faculty/Student Ratio | -0.1162 | 0.3706 | -0.8120 | -0.4356 |
| Citations per Faculty | -0.5160 | -0.3068 | 0.4296 | -0.6663 |
| International Faculty Ratio | -0.6648 | -0.6000 | -0.1895 | 0.1334 |
| International Student Ratio | -0.8028 | -0.3212 | -0.2277 | 0.2696 |

Table 2: Loading Matrix

| Original Variables | Final Weight |
|---|---|
| Academic Reputation | -0.0914 |
| Employer Reputation | -0.1115 |
| Faculty/Student Ratio | 0.1897 |
| Citations per Faculty | 0.2807 |
| International Faculty Ratio | 0.3945 |
| International Student Ratio | 0.3379 |

Table 3: Weight Result

| Old Rank | Institution | New Rank |
|---|---|---|
| 54 | City University of Hong Kong | 1 |
| 16 | EPFL | 2 |
| 104 | Technical University of Denmark | 3 |
| 136 | University of Basel | 4 |
| 127 | University of Geneva | 5 |
| 2 | University of Cambridge | 6 |
| 100 | Rice University | 7 |
| 9 | ETH Zurich - Swiss Federal Institute of Technology | 8 |
| 48 | Institut Polytechnique de Paris | 9 |
| 6 | California Institute of Technology (Caltech) | 10 |

Table 4: New ranking

## A.2 Factor Analysis Tables

| Original Variables | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Academic Reputation | -0.0209 | 0.9100 | 0.0709 | 0.1267 |
| Employer Reputation | 0.0593 | 0.9225 | 0.0623 | -0.0710 |
| Faculty/Student Ratio | 0.0018 | 0.0994 | 0.9933 | -0.0582 |
| Citations per Faculty | 0.1902 | 0.0465 | -0.0611 | 0.9731 |
| International Faculty Ratio | 0.8935 | -0.1340 | -0.0264 | 0.1969 |
| International Student Ratio | 0.9158 | 0.1722 | 0.0265 | 0.0563 |
| Accumulative Contribution Rate | 0.2795 | 0.5693 | 0.7361 | 0.9050 |

Table 5: Loading Matrix

| Original Variables | Commualitiy |
|---|---|
| Academic Reputation | 0.8495 |
| Employer Reputation | 0.8634 |
| Faculty/Student Ratio | 0.9999 |
| Citations per Faculty | 0.9889 |
| International Faculty Ratio | 0.8557 |
| International Student Ratio | 0.8723 |

Table 6: Communalities

| Rank | Institution |
|---|---|
| 1 | Massachusetts Institute of Technology (MIT) |
| 2 | University of Cambridge |
| 4 | University of Oxford |
| 6 | California Institute of Technology (Caltech) |
| 9 | ETH Zurich - Swiss Federal Institute of Techno... |
| 11 | National University of Singapore (NUS) |
| 13 | University of Pennsylvania |
| 18 | Yale University |
| 19 | Nanyang Technological University, Singapore (NTU) |
| 22 | Columbia University |
| 58 | University of Amsterdam |
| 63 | Brown University |
| 67 | Universidad de Buenos Aires (UBA) |
| 75 | Lomonosov Moscow State University |
| 84 | University of Zurich |
| 119 | Washington University in St. Louis |
| 160 | King Fahd University of Petroleum & Minerals |
| 199 | Vanderbilt University |
| 224 | The Hebrew University of Jerusalem |
| 227 | Southern University of Science and Technology |
| 231 | Gadjah Mada University |

Table 7: Stable Institutions

### A.3 Decision Tree Tables

| Original Variables | Decision Tree Weight |
|---|---|
| Academic Reputation | 0.18828451882845187 |
| Employer Reputation | 0.17573221757322174 |
| Faculty/Student Ratio | 0.17573221757322174 |
| Citations per Faculty | 0.14644351464435146 |
| International Faculty Ratio | 0.15481171548117156 |
| International Student Ratio | 0.17154811715481172 |

Table 8: Decision Tree Weight

| Old Rank | Institution | New Rank |
|---|---|---|
| 1 | Massachusetts Institute of Technology (MIT) | 1 |
| 2 | University of Cambridge | 2 |
| 3 | Stanford University | 3 |
| 4 | University of Oxford | 4 |
| 5 | Harvard University | 5 |
| 8 | UCL | 6 |
| 7 | Imperial College London | 7 |
| 6 | California Institute of Technology (Caltech) | 8 |
| 10 | University of Chicago | 9 |
| 14 | The University of Edinburgh | 10 |

Table 9: Decision Tree New ranking

| Old Rank | Institution | New Rank |
|---|---|---|
| 1 | Massachusetts Institute of Technology (MIT) | 1 |
| 2 | University of Cambridge | 2 |
| 4 | University of Oxford | 3 |
| 7 | Imperial College London | 4 |
| 8 | UCL | 5 |
| 6 | California Institute of Technology (Caltech) | 6 |
| 9 | ETH Zurich - Swiss Federal Institute of Techno... | 7 |
| 3 | Stanford University | 8 |
| 15 | The University of Edinburgh | 9 |
| 5 | Harvard University | 10 |

Table 10: Decision Tree New ranking for Split Test Data

| Original Variables | Decision Tree Weight |
|---|---|
| Component 1 | 0.20502092050209203 |
| Component 2 | 0.27615062761506276 |
| Component 3 | 0.24686192468619245 |
| Component 4 | 0.27196652719665276 |

Table 11: Decision Tree Weight After PCA

# B Graphs

## B.1 EDA

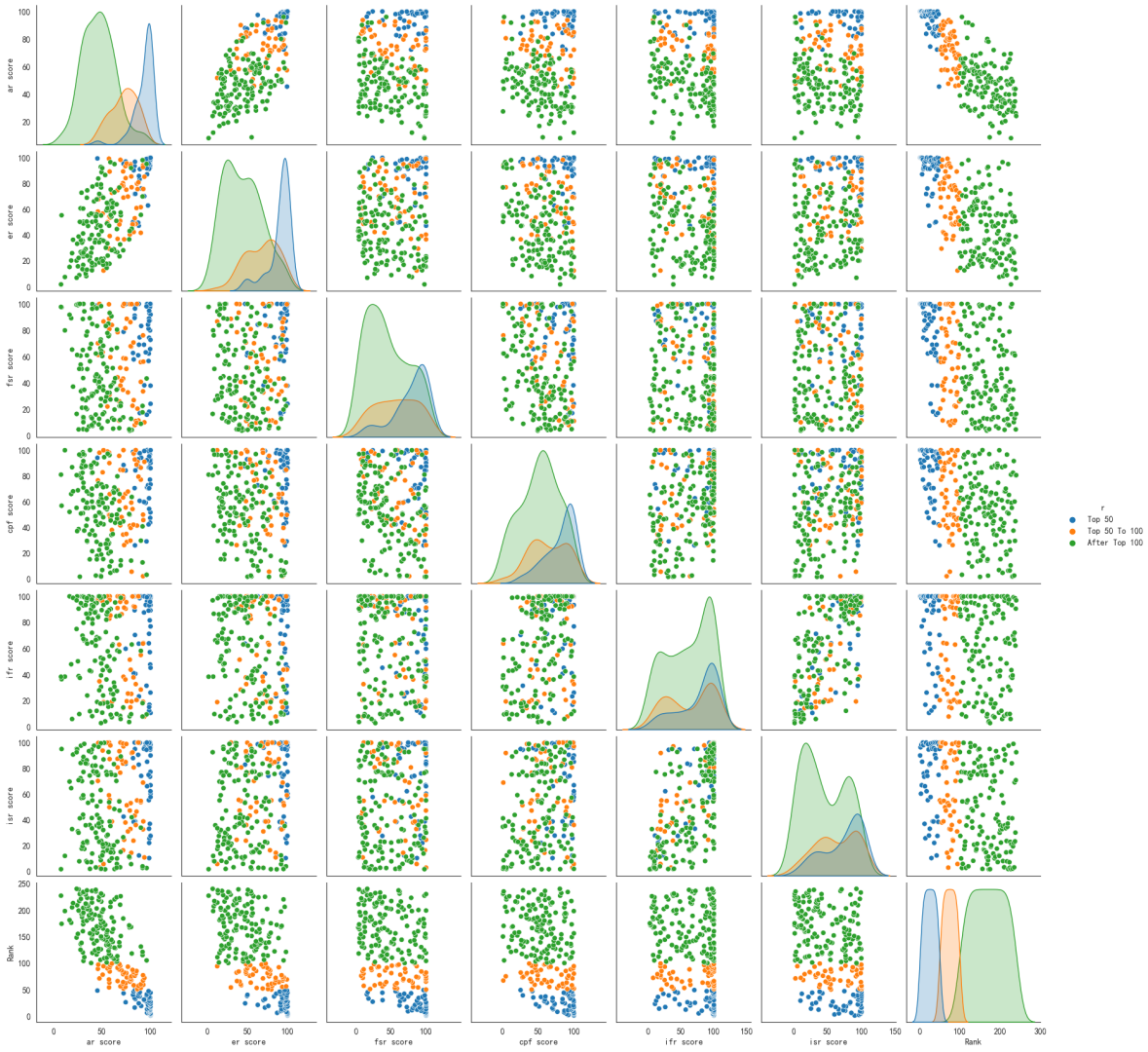



Figure 1: EDA

## B.2 Visualization Graphs

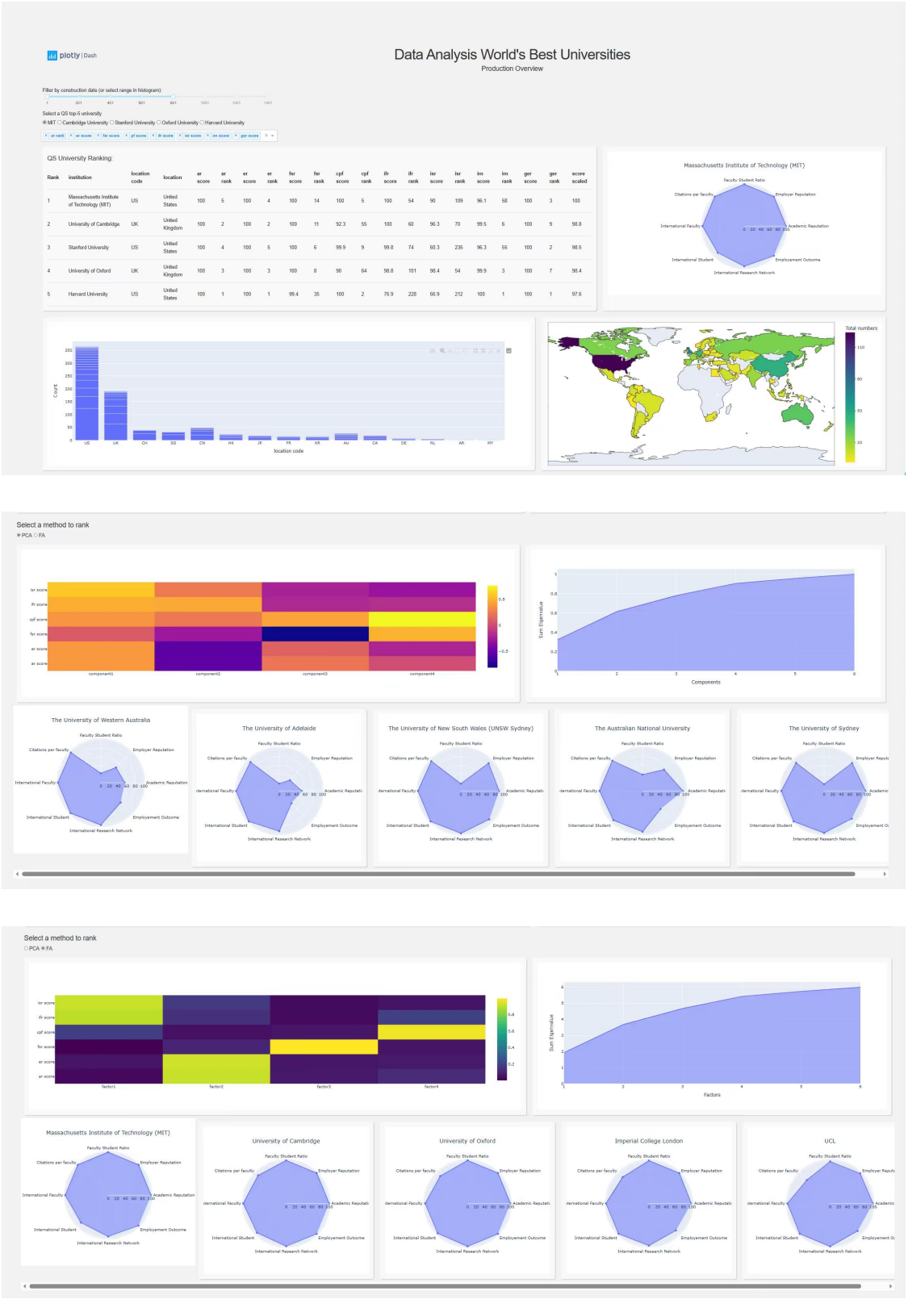



Figure 2: Web Data Analysis Platform

### B.3 Principal Component Analysis Graphs

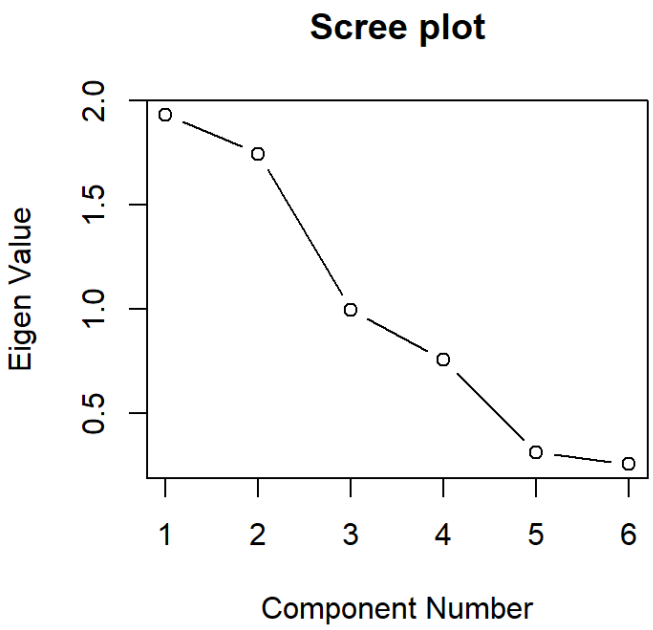



Figure 3: Scree Plot

### B.4 Factor Analysis Graphs

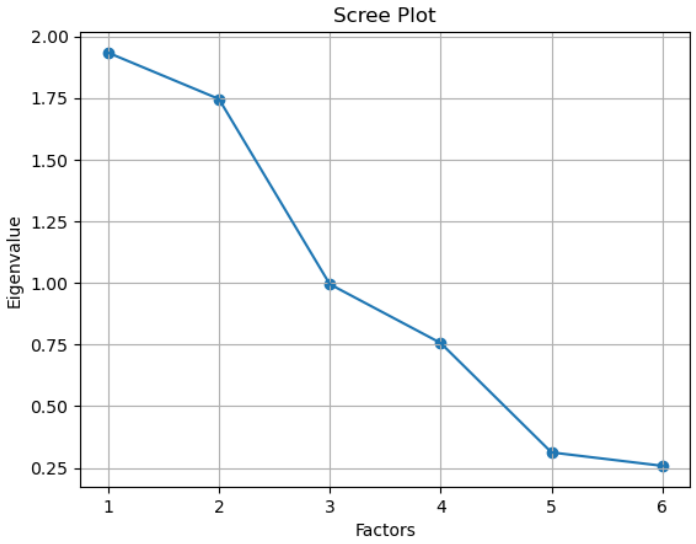



Figure 4: Scree Plot

## B.5 Decision Tree Graphs

| | Component1 | Component2 | Component3 | Component4 | Rank | institution | PCA_DT_Scores |
|---|---|---|---|---|---|---|---|
| **0** | -0.692704 | -0.422391 | 0.002504 | 0.023206 | 1 | Massachusetts Institute of Technology (MIT) | -0.191788 |
| **1** | -0.715346 | -0.377334 | 0.090753 | 0.002643 | 2 | University of Cambridge | -0.492286 |
| **2** | -0.498925 | -0.470384 | -0.102268 | 0.222290 | 3 | Stanford University | -0.746095 |
| **3** | -0.714207 | -0.366801 | 0.118443 | -0.013926 | 4 | University of Oxford | -1.041935 |
| **4** | -0.393410 | -0.521557 | -0.055404 | 0.023305 | 5 | Harvard University | -1.243356 |
| **5** | -0.614001 | -0.422371 | -0.040773 | 0.027429 | 6 | California Institute of Technology (Caltech) | -1.540716 |
| **6** | -0.699859 | -0.250035 | 0.097517 | -0.033688 | 7 | Imperial College London | -1.871116 |
| **7** | -0.670101 | -0.222084 | 0.184406 | -0.011549 | 8 | UCL | -2.178273 |
| **8** | -0.727651 | -0.330109 | -0.004059 | -0.047440 | 9 | ETH Zurich - Swiss Federal Institute of Techno... | -2.367187 |
| **9** | -0.419554 | -0.420314 | 0.062795 | 0.010725 | 10 | University of Chicago | -2.653481 |

Figure 5: New Rank after PCA + Decision Tree

## B.6 Clustering Graphs

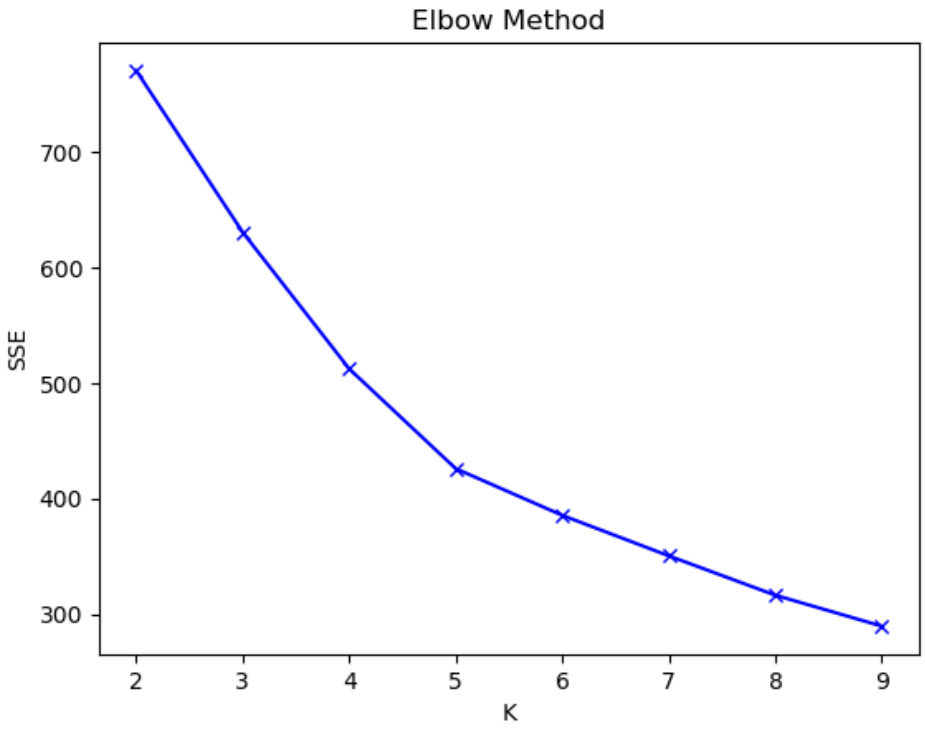
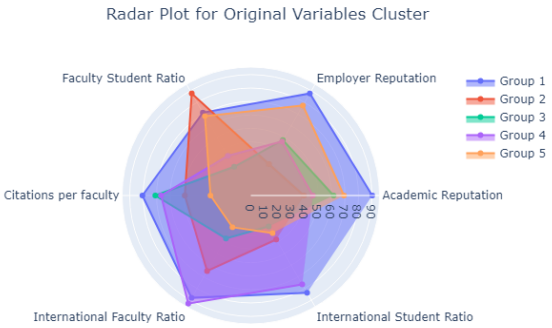
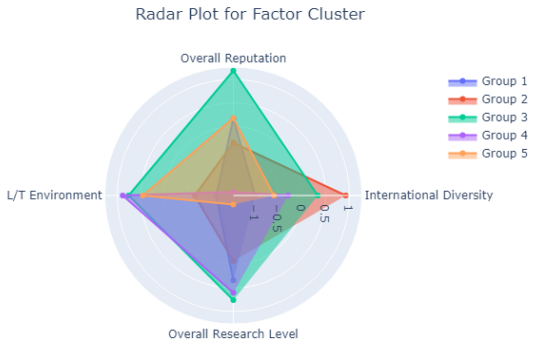


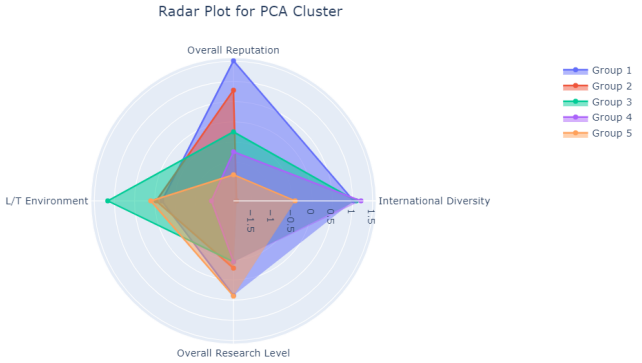

Figure 6: Elbow Curve




(a) Using Original Variables




(b) Using Factor




(c) Using PCA

Figure 7: Clustering Result