# A Global Analysis of University Rankings Through Clustering

**Oh Tien Cheng**

**Abstract**

University rankings like the Times Higher Education World University Rankings are often used to assess universities around the world quantitatively, and identify top universities. These rankings are formed using factors such as the quality of research done at these universities, the quality of teaching, and amount of funding. Although they provide a way of assessing the quality of a universities, such rankings have been criticised as not being useful indicators of the actual quality of a university. In this paper, we attempt to segment universities on several university rankings based on their scores on these rankings. Principal Component Analysis is applied on the criterion scores of the universities to reduce the dimensionality of the data. A set of clustering algorithms are then applied on the data to attempt to find an optimal clustering of the universities.

## 1 Introduction

In this paper, we attempt to analyze and segment universities on several university rankings based on their scores on these rankings. Principal Component Analysis is applied on the criterion scores of the universities to reduce the dimensionality of the data. A set of clustering algorithms are then applied on the data to attempt to find a set of optimal clusterings of the universities.

### *1.1 University Rankings*

University rankings are a method for quantitatively assessing the general quality of a university through a number of factors. The concept of a university ranking first emerged in 1900 (4), when the publication *Where We Get Our Best Men*, was published in England, comparing universities based on the number of successful men who had studied there. The modern university ranking first became popularized in 1983, with the publication of *America's Best Colleges* by the US News and World Report (8), which provided statistics on universities in the United States, allowing for them to be compared quantitatively. Subsequently in 2004, and 2005, the release of the Academic Ranking of World Universities (ARWU) by Shanghai Jiao Tong University in China, and the Times Higher Education World University Rankings by the Times Higher Education magazine has led to university rankings becoming mainstream.

#### *1.1.1 Times Higher Education World University Rankings*

The Times Higher Education World University Ranking system has been described as one of the most influential university rankings in the world. The ranking system is based on a total score, which is based on a weighted sum of five overall indicators: income, international diversity, teaching, research, and citations. These five indicators can be further broken down into 13 indicators (3). A full breakdown of the ranking system can be found in table 1.

## 2 Related Works

As of the time of writing, there have been several works analysing the trends behind university rankings. In 2006, Steiner conducted an analysis of the Times Higher Education World University rankings, performing a Principal Component Analysis on it. He discovered that three principal components could adequately explain the scoring criterion behind the rankings, with the most important factor being academic performance, such as the quality of research done. In 2020, Selten et al. conducted an analysis on various university ranking systems and how they differ by regions using Principal Component Analysis, discovering that universities in English speaking regions are ahead of other universities in rankings, finding regional biases in these ranking systems.

## 3 Methodology

### *3.1 Data Set*

The data set used for this analysis are the compiled Times Higher Education World University rankings from the years 2011 to the year 2016, comprising six years of data, obtained from Kaggle. (9)

### *3.2 Data Cleaning*

The data set in it's raw form is unclean, and steps have to be taken to clean the data before it is able to be analysed. The data was first loaded into Python as a Pandas data frame. Pandas is a Python library for the analysis of tabular data (10). Data cleaning then was done using the Pandas library. The steps involved in cleaning the data are detailed in Table 2.

### *3.3 Exploratory Data Analysis*

Before we attempt our clustering analysis, we will first perform an exploratory data analysis on our data. Exploratory data analysis is an approach to data analysis that employs a number of different techniques to understand the underlying structure of data, discover important variables and their relationships, detect outliers and suggest suitable models for further analysis. (5)

### *3.4 Data Pre-processing*

Before we can analyse our data further, we will need to process our data to reduce the dimensionality of the data. To do this, we will first select the features we want to apply PCA on. We narrowed down the set of features to include:

- university_name
- teaching
- international

| Overall Indicator | Indicator | Percentage Weighting (%) |
|---|---|---|
| Income | Research income from industry per academic staff | 2.5 |
| International Diversity | Ratio of International to Domestic Staff | 3 |
| | Ratio of International to Domestic Students | 2 |
| Teaching | Reputational Survey | 15 |
| | PhDs awards per academic staff | 6 |
| | Undergrads admitted per academic staff | 4.5 |
| | Income per academic staff | 2.25 |
| | PhDs/undergraduate degrees awarded | 2.25 |
| Research | Reputational Survey | 19.5 |
| | Research income (scaled) | 5.25 |
| | Papers per research and academic staff | 4.5 |
| | Public research income/ total research income | 0.75 |
| Citations | Citation impact (normalised average citation per paper) | 32.5 |

Table 1: Breakdown of Times Higher Education World University Ranking System

| Feature(s) | Problem | Data Cleaning Step |
|---|---|---|
| international, income, total_score, num_students, student_staff_ratio, international students, female_male_ratio | Missing data present as pre-filled value ("-"). | Replaced pre-filled value to indicate missing values, and dropped all missing values. |
| num_student | Comma used as decimal place, causing feature to be recognised as a string. | Commas replaced with periods as decimal points, and feature converted to a numerical feature. |
| international_students | Feature is in percentage format, causing it to be recognised as a string. | Percentage sign stripped from the feature, and converted to a float. |
| female_male_ratio | Feature represented as a ratio (e.g. 40:60), hence recognized as a string. | Converted to show the percentage of female students in the university. |

Table 2: Data Cleaning Steps

- research
- citations
- income
- num_students
- student_staff_ratio
- international_students
- female_male_ratio

The selected features are the five overall indicators which are used to calculate the total score, as well as additional statistics recorded by the Times Higher Education magazine.

The data set is then aggregated to get the average indicators for each university across the six years.

#### 3.4.1 Feature Scaling

We then perform feature scaling on our data. Feature scaling is a form of feature engineering that involves individually changing the scale of quantitative features, such that all features are on a similar scale. (16)

Feature scaling is required as our features are on different scales. This means that when we use principal component analysis to transform our data, features with a greater variance would end up being over-represented in the principal components (2).

To scale our data, we make use of the scikit-learn library (12), which includes the *StandardScaler* transformer to standardize our data. Standardization applies a transformation on our data, by subtracting each data point in a feature with its mean, and dividing by its standard deviation, as shown in the formula below:

$$X_{scaled} = \frac{x - mean(x)}{std(x)}$$

### 3.5 Principal Component Analysis

After performing feature scaling, Principal Component Analysis is performed on the data.

Principal Component Analysis (PCA) is an unsupervised learning approach for dimensionality reduction, invented by Karl

Pearson in 1901 (11). It does this by projecting high dimensional data into lower dimensions known as principal components. These principal components are uncorrelated linear combinations of variables which are ordered so that the first few principal components retain most of the variation present in all the original variables. (6)

In the scikit-learn library, the PCA process is implemented as the *PCA* transformer. The results of PCA are shown in Table 3.

#### 3.5.1 *Selection of the Number of Principal Components to Retain*

Once PCA is applied to our data, we then need to decide how many principal components to retain. We can do this by analysing the eigenvalues and explained variance of each principal component.

Upon analysis of our principal components, we decided to retain the first three principal components. This was done after using the Kaiser Criterion, which states that principle components with an eigenvalue greater than one should be retained. (7)

#### 3.5.2 *Interpretation of Principal Components*

The breakdown of loadings of the first three principal components are shown in the loading plot in figure 1 and table 4

| Variable | PC1 | PC2 | PC3 |
|---|---|---|---|
| teaching | 0.4741 | -0.2824 | 0.0478 |
| international | 0.3688 | 0.4810 | -0.0675 |
| research | 0.5050 | -0.2190 | 0.1056 |
| citations | 0.4180 | 0.0513 | 0.0872 |
| income | 0.2309 | -0.4191 | 0.1822 |
| num_students | -0.0246 | -0.0403 | 0.6799 |
| student_staff_ratio | -0.0833 | 0.1924 | 0.6441 |
| international_students | 0.3841 | 0.3934 | -0.1532 |
| female_male_ratio | 0.0140 | 0.5188 | 0.2018 |

Table 4: Loadings for the first three principal components

Principal Component 1 is a general measure of the academic performance of a university. It measures a weighted average of factors relating the research conducted at the university, the quality of teaching, the proportion of international staff and students, the citation impact of papers published and the amount of research income. The most dominant loadings are research (0.5050) and teaching (0.4741). In effect, universities with a strong reputation for academic research and teaching will score highly in PC1.

Principal Component 2 is a measure of how popular a university is internationally in contrast to the research income afforded to academic staff and the quality of research and teaching. The loadings for teaching, research and income are negative. The most dominant loadings are that for female to male ratio, international, and percentage of international students. A university that has a high international diversity, but poorer research and teaching quality will score highly in PC2.

Principal Component 3 is a measure of the student staff ratio and the number of students, in contrast to the percentage of international students in the university. Universities with a high number of mostly local students compared to the number of staff will score highly in PC3.

### 3.6 *Clustering*

Using the principal components, we will then attempt to cluster universities. We will utilize the following clustering algorithms, implemented in the *scikit-learn* library (12):

- K-Means Clustering
- Hierarchical Clustering

#### 3.6.1 *Silhouette Score*

To evaluate the quality of clustering, we will make use of the silhouette score. A silhouette score is a measure that of how similar a data point is to its cluster compared to other clusters. (13)It has a range of $[-1, 1]$, where values near 1 indicate a data point is far away from neighbouring clusters, while values close to 0 indicate a data point is close to the decision boundary between two clusters. A negative value indicates that the data point may be in the wrong cluster.

We will make use of the average silhouette score of each clustering obtained, and compare them to select the best clustering.

#### 3.6.2 *K-Means Clustering*

Before applying K-Means clustering, it is necessary to determine the number of clusters we want to obtain. This can be done through two means: an elbow plot, and by using silhouette scores.

An elbow plot displays the inertia (within cluster sum of squares) of the clustering. Inertia measures how internally coherent the clusters are and is given by the formula (1)

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (\|x_i - \mu|j\|^2)$$

As the number of clusters increases, the inertia of the clustering decreases. We can select the number of clusters by finding the "elbow" of the plot, when the decrease in the inertia when an additional cluster is added decreases, showing diminishing returns. The elbow plot is shown in figure 2

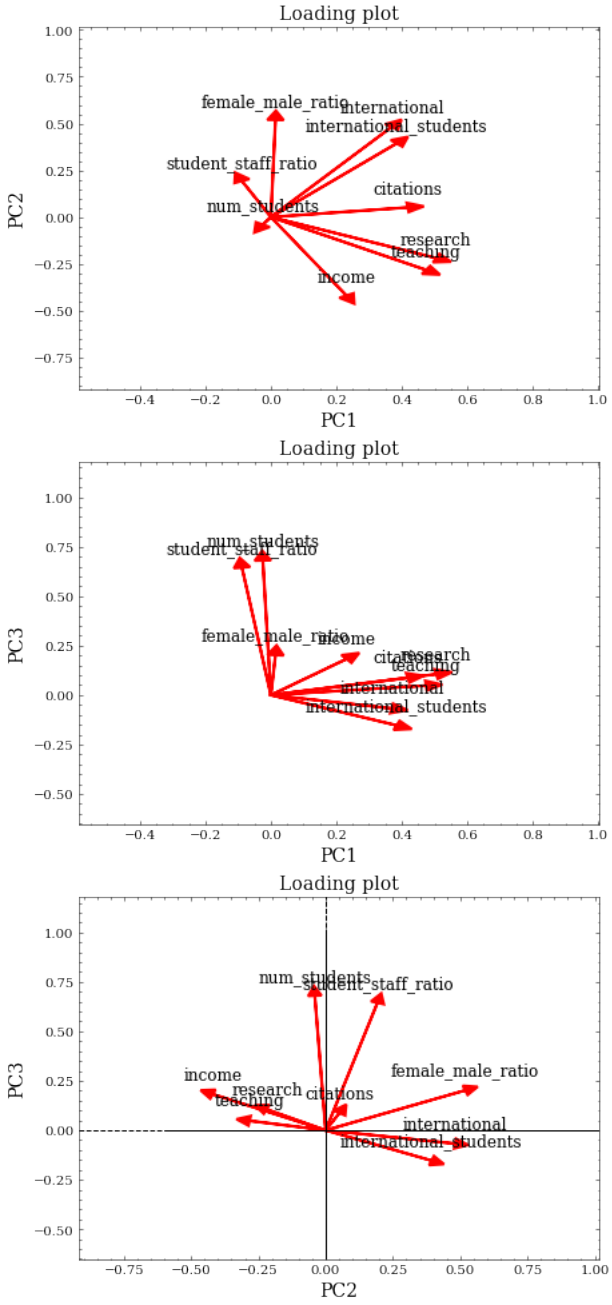


Figure 1: PCA Loading Plot

| Principal Component | Eigenvalue | Explained Variance (%) | Cumulative Explained Variance (%) |
|---|---|---|---|
| **1** | **3.0673** | **34.03** | **34.03** |
| **2** | **1.7194** | **19.08** | **53.11** |
| **3** | **1.3078** | **14.51** | **67.62** |
| 4 | 0.9827 | 10.90 | 78.52 |
| 5 | 0.7402 | 8.21 | 86.74 |
| 6 | 0.5450 | 6.05 | 92.78 |
| 7 | 0.4384 | 4.86 | 97.65 |
| 8 | 0.1403 | 1.56 | 99.2 |
| 9 | 0.0718 | 0.8 | 100 |

Table 3: Results of PCA

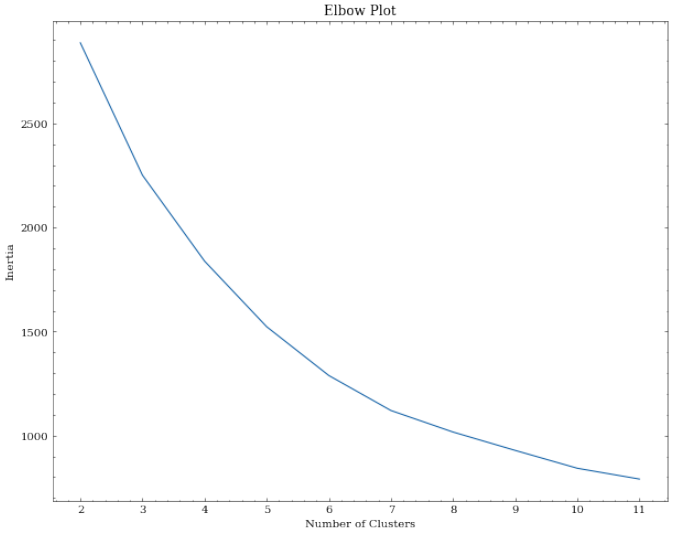


Figure 2: Elbow Plot

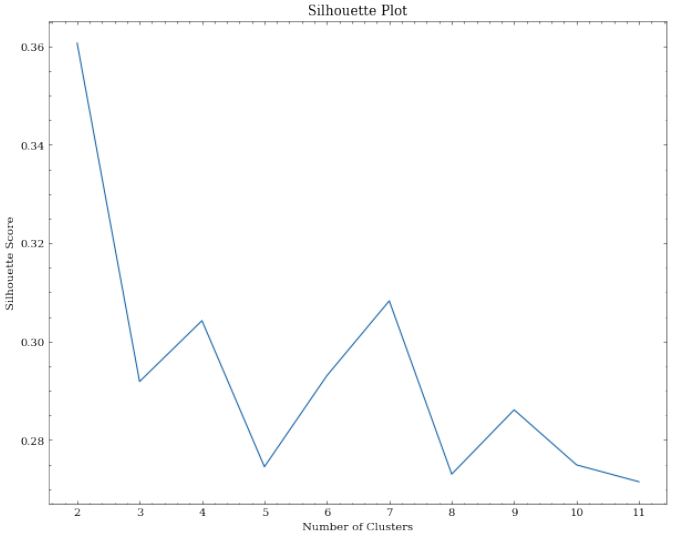


Figure 3: K-Means Silhouette Score Per Number of Clusters

From the elbow plot, the "elbow" is ambiguous for this data, but appears to be at 7, indicating 7 clusters is optimal. Since it is unclear what the optimal number of clusters is, we will calculate and plot out the average silhouette score of the K-Means clustering for the various number of clusters. The silhouette plot is shown in figure 3

The silhouette plot shows that two clusters gives the highest silhouette score (0.36), followed by 7 clusters (0.30). In this paper, we will choose to form 7 clusters. The resulting clustering is shown in figure 4, and the cluster centroids are shown in table 5.

### *3.6.3 Hierarchical Clustering*

The second clustering algorithm we will utilize is Hierarchical Clustering. In our paper, we utilize the Ward linkage method. Before we attempt hierarchical clustering, it is also necessary to determine the optimal number of clusters. To do this, we will utilize a silhouette plot, as per K-Means. The resulting silhouette plot is shown in figure 5. From the silhouette scores, we see that two clusters gives us the highest silhouette score (0.33), followed by three to five clusters (0.24).

| Cluster | PC1 | PC2 | PC3 |
|---|---|---|---|
| 1 | -0.566523 | 0.412081 | 1.887642 |
| 2 | 1.045288 | 1.539183 | -0.371273 |
| 3 | -0.999915 | 0.505239 | -0.147379 |
| 4 | 4.148614 | -0.904672 | 0.187818 |
| 5 | 1.108621 | -0.995463 | -0.012676 |
| 6 | -0.904165 | -0.350776 | 10.297127 |
| 7 | -1.566273 | -1.385299 | -0.481503 |

Table 5: K-Means Centroids

## 4 Discussion

After evaluating the various cluster methods, we will select K-Means clustering with 7 clusters. We choose this as the silhouette score obtained is still higher than than a similar number of clusters for hierarchical clustering, and having more clusters allows for more analysis into the data.

### *4.1 Interpretation of Final Clustering*

The cluster centroids for our final clusterings are shown in table 5

#### *4.1.1 Cluster 1*

This cluster contains universities that score above average in PC3 and slightly above average in PC2, but score poorly in PC1. This means that universities in this cluster are not well known for their academic performance, but attract an above average amount of students, with some international staff and students in the university. This cluster appears to represent local public universities, which serve the general population.

#### *4.1.2 Cluster 2*

This cluster contains universities that a slightly above average in PC1 and PC2, but below average in PC3. This means these are universities with an above average academic record, containing a large population of international students and staff. Due to their below average student to staff ratio (PC3), these universities can provide more support to students, which could explain why they score highly in PC2, as international students could expect a higher quality of teaching as a result. Cluster 2 appears to be entirely dominated by the United Kingdom and Australia, having 209 and 71 universities in this cluster respectively.

#### *4.1.3 Cluster 3*

This cluster contains universities that score below poorly in PC1 and PC3, but score above average in PC2 (international diversity). These are universities that appear to have poor academic quality, and but still have international students coming there to study.

#### *4.1.4 Cluster 4*

This cluster contains universities that score extremely highly in PC1, but score very poorly in PC2. This cluster, which includes top ranking universities, such as MIT, Cambridge, groups universities that are well known for their academic excellence, have very little international diversity. This cluster is dominated by countries in the Western hemisphere, with only Hong Kong and Singapore having universities represented in this cluster.

#### *4.1.5 Cluster 5*

This cluster contains universities that score above average in PC1, but score poorly in PC2. Universities with a above average academic track record, but are very closed off to international students and staff. This cluster, which is dominated by American universities, seems to encompass many local technical universities, which are mostly for local students.

#### *4.1.6 Cluster 6*

This cluster contains universities that have a poor score for PC1 and PC2, but an extremely high score for PC3. This cluster contains universities from only three countries: Italy, Turkey and South Africa. These universities appear to be fairly small universities, with a low amount of staff.

#### *4.1.7 Cluster 7*

This cluster contains universities that score poorly in PC1, PC2 and PC3. These are universities that have a poor reputation, and hence receive few students locally and abroad, causing there to be a low student to staff ratio. This cluster is dominated by countries in the Eastern hemisphere like Japan, Taiwan, India, China and Iran. These are universities that are very insular, and meant mostly for local students.

## 5 Conclusions

In conclusion, we observe from the cluster results that top ranking universities appear to be concentrated in the fourth cluster, while low ranking universities appear to be in the seventh cluster, with the main difference between them being a contrast of scores in PC1, measuring the general academic quality of the school, followed by how open they are to international staff. We also note a disparity in the countries represented in these two clusters, where top ranking universities in cluster four are almost all from Western countries, while those in cluster seven are almost all from Eastern universities. This supports earlier works (14), which find that such ranking systems tend to be biased towards English speaking regions.

## References

[1] 2.3. Clustering — scikit-learn 0.24.2 documentation. URL https://scikit-learn.org/stable/modules/clustering.html.

[2] Importance of Feature Scaling — scikit-learn 0.24.2 documentation. URL https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html.

[3] P. Baty. THE unveils broad, rigorous new rankings methodology, June 2010. URL https://www.timeshighereducation.com/news/the-unveils-broad-rigorous-new-rankings-methodology/411907.article.

[4] B. Hammarfelt, S. de Rijcke, and P. Wouters. From Eminent Men to Excellent Universities: University Rankings as Calculative Devices. *Minerva*, 55(4):391–411, Jan. 2017. ISSN 1573-1871. doi: 10.1007/s11024-017-9329-x. URL https://europepmc.org/articles/PMC5686281.

[5] H. Hinterberger. *Exploratory Data Analysis*, pages 1080–1080. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_1384. URL https://doi.org/10.1007/978-0-387-39940-9_1384.

[6] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition, 2002. ISBN 978-0-387-95442-4. doi: 10.1007/b98835. URL https://www.springer.com/gp/book/9780387954424.

[7] H. F. Kaiser. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, 20(1):141–151, Apr. 1960. ISSN 0013-1644. doi: 10.1177/001316446002000116. URL https://doi.org/10.1177/001316446002000116. Publisher: SAGE Publications Inc.

[8] P. T. M. Marope, P. J. Wells, E. Hazelkorn, and Unesco. *Rankings and accountability in higher education: uses and misuses*. 2013. ISBN 978-92-3-001156-7. OCLC: 903071414.

[9] M. O'Neill. World University Rankings. URL https://kaggle.com/mylesoneill/world-university-rankings.

[10] T. pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. URL https://doi.org/10.5281/zenodo.3509134.

[11] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. Nov. 1901. doi: 10.1080/14786440109462720. URL https://zenodo.org/record/1430636.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, Nov. 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

[14] F. Selten, C. Neylon, C.-K. Huang, and P. Groth. A Longitudinal Analysis of University Rankings. *arXiv:1908.10632 [physics]*, Jan. 2020. URL http://arxiv.org/abs/1908.10632. arXiv: 1908.10632.

[15] J. E. Steiner. World University Rankings - A Principal Component Analysis. *arXiv:physics/0605252*, May 2006. URL http://arxiv.org/abs/physics/0605252. arXiv: physics/0605252.

[16] A. Zheng and A. Casari. Feature Engineering for Machine Learning. In *Feature Engineering for Machine Learning*, pages 29–31. O'Reilly Media, Inc., Apr. 2018. URL https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/.

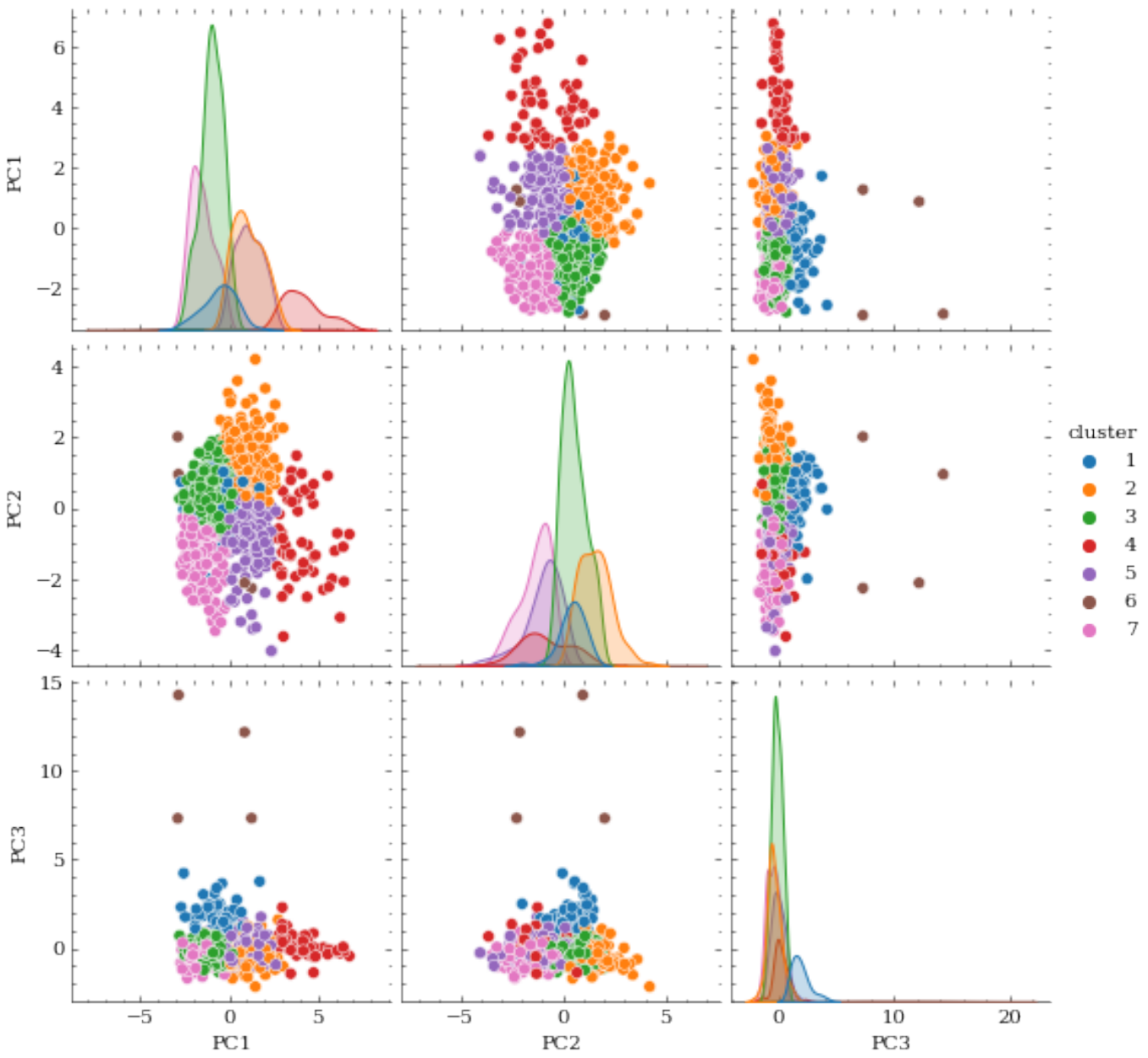


Figure 4: K-Means Clustering (7 Clusters)

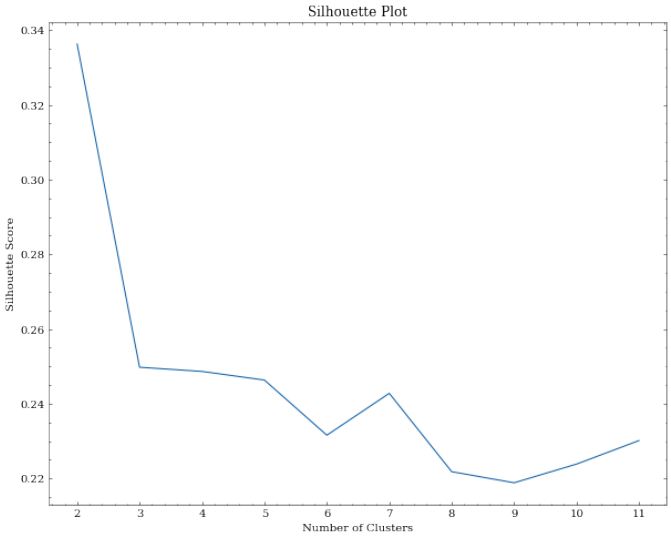


Figure 5: Hierarchical Clustering Silhouette Score Per Number of Clusters