2019

Hierarchical Cluster Analysis: A New Type of Ranking Criteria Based on ARWU Ranking Data

Zhengshuo Li
Technological University Dublin

Hierarchical cluster analysis: a new type of ranking criteria based on ARWU ranking data

Student Name

Zhengshuo Li

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

2019

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: **Zhengshuo Li**

Date: **29/12/2018**

ABSTRACT

The advent of big data leads to many applications of Machine Learning techniques. University rankings is one of the applicable domains, which is currently playing a crucial role in the assessment of the universities' performance. Currently, the rankings are usually carried out by some authoritative ranking institutions by means of weighting techniques and the results are conveyed in numerical rankings. Three of the most famous university ranking institutions have been introduced from a technical perspective. However, these institutions have been proven to be subjective in relation to their data selection and weighting method. Data used in this research is gathered from one of the most known ranking institutions: the ARWU ranking which consists of six indicators namely the numbers of: Alumni wining Nobel Prize, highly cited researchers, papers published in Nature &Science, papers indexed in Science and the per capita performance. ARWU ranks universities based on their overall score derived from weighting the aforementioned indicators. Because of the unrepresentativeness of the ARWU data, which has a huge influence on international ranking, this paper proposed a new type of ranking based on hierarchical clustering analysis which is a type of unsupervised learning technique. The agreement between existing ARWU ranking and cluster analysis is verified in this research. Hierarchical clustering method applied on the same indicators as ARWU can be considered as an alternative way to rank universities, which can supply the rankings from different perspectives.

Key words: *Big data; Machine Learning; University rankings; Unsupervised learning; Hierarchical cluster analysis; ARWU world university rankings*

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to all the people who helped and supported me while writing the MSC dissertation.

Firstly, I would like to express my appreciation and thanks to my supervisor Tamara Matthews who patiently guided me on my dissertation.

To all the DIT staffs, for their support on giving such a comfortable and convenient learning environment.

To my family, for their encouragement and finical supporting during the last two years.

To my classmates, for their patience to help me with any problems existing in learning.

And finally, to my friend and classmate Runzuo Yang for his endless encouragement and belief in helping me finish my Master degree.

TABLE OF CONTENTS

ABSTRACT 3

Chpater 1 – Introduction 11

1.1 Background 11

1.2 Research Objectives 11

1.3 Research Hypotheses 12

1.4 Research Methodologies 12

1.5 Scope and Limitations 13

1.6 Document Outline 13

Chapter 2 - Literature review 14

2.1 Background and application domain 14

2.1.1 University ranking institutions 14

2.1.2 Cumulative table & Criticism 16

2.2 Other methodologies 16

2.2.1 Conclusion table 18

2.3 Method review 19

2.3.1 Pearson’s correlation 19

2.3.2 Linear regression 19

2.3.3 Principle component analysis (PCA) 20

2.3.4 Clustering methods 20

2.3.5 Partitioning clustering method 21

2.3.6 Hierarchical clustering 21

Chapter 3 - Design and methodology 23

3.1 Exploratory analysis 23

3.1.1 Data pre-processing 23

3.2 Feature selection 25

3.2.1 Pearson's correlation 25
3.2.2 Linear regression 26
3.2.3 Principal component analysis (PCA) 26
3.3 Feature supplement 27
3.5 Design procedure and metrics 27
3.6 Validation 28
3.6.1 Internal validation 29
3.6.2 Design of validation for hierarchical clusters 29
Chapter 4 – Implementation, Results, evaluation and discussion 32
4.1 Data processing 32
4.1.1 Missing data 32
4.1.2 Aggregation 34
4.2 Feature supplement 35
4.3 Feature selection 36
4.3.1 Pearson correlation 36
4.3.2 Linear regression 38
4.3.3 Principal component analysis (PCA) 40
4.4 Implementation of multiple clustering method 46
4.4.1 Accessing the clustering tendency 46
4.4.2 Choosing the optimal number of clusters 48
4.4.3 Generating clusters and internal validation 50
4.4.4 Sub-clusters 54
4.4.5 Result of all the clusters 58
4.4.6 Discussion & analysis 60
Chapter 5 – Conclusion 66
5.1 Research Overview 66

5.2 Problem Definition..... 66
5.3 Design/Experimentation, Evaluation & Results 66
5.4 Contributions and impact..... 67
5.5 Future Work & recommendations..... 67
References..... 68

TABLE OF FIGURES

(Figure 2.1: Example of a dendrogram) 22

(Figure 3.1: Choosing optimal number of hierarchical cluster) 30

(Figure 4.1: Visualization of Pearson correlation results) 38

(Figure 4.2: Variables' importance) 39

(Figure 4.3: Result from scree plot of explained variance) 41

(Figure 4.4: The spread of eigenvalues in each component) 42

(Figure 4.5: correlation circle plot) 43

(Figure 4.6: Cos^2 of each variable) 44

(Figure 4.7: Contribution of variables in component 1&2) 45

(Figure 4.8: VAT of PCA data) 47

(Figure 4.9: The set of possible clustering numbers by dendrogram) 51

(Figure 4.10: Dendrogram of hierarchical cluster) 52

(Figure 4.11: K-means clustering plot) 53

(Figure 4.12: The spread of hierarchical clusters) 54

(Figure 4.13: Dendrogram of sub-clusters of cluster 7) 55

(Figure 4.14: Dendrogram of sub-clusters in cluster 8) 55

(Figure 4.15: Sub-clusters of cluster 7) 57

(Figure 4.16: Sub-clusters of cluster 8) 57

(Figure 4.17: Performance of clusters in each indicator) 62

(Figure 4.18: Performance of each cluster in all the indicators) 63

TABLE OF TABLES

(Table 2.1: Criteria and indicators of ARWU) 15

(Table 2.2: Areas and indicators of THE) 15

(Table 2.3: Metric of QS) 16

(Table 2.4: Cumulative table of ranking methods) 16

(Table 2.5: Cumulative table of university ranking methods) 19

(Table 3.1: Overview of ARWU data) 24

(Table 3.2: Summarization of ARWU data) 24

(Table 3.3: Aggregation method for ARWU data) 25

(Table 3.4: Linkage methods) 28

(Table 4.1: Missing numerical data report) 32

(Table 4.2: Missing values in the ARWU data) 32

(Table 4.3: Imputed ARWU data) 34

(Table 4.4: Aggregated ARWU data) 35

(Table 4.5: Supplemented ARWU data) 36

(Table 4.6: Pearson correlation coefficient table) 37

(Table 4.7: Probability Value table) 37

(Table 4.8: Evaluation of regression model) 39

(Table 4.9: Fitness of the variables) 39

(Table 4.10: Relationship between eigenvalue and variance) 42

(Table 4.11: Contribution of the variables in principle components) 45

(Table 4.12: Linkage cophenetic correlation coefficient) 49

(Table 4.13: Internal validation of H cluster by Silhouette width) 51

(Table 4.14: Internal validation of H cluster by Dunn index) 51

(Table 4.15: Internal validation of K-means by Silhouette width) 52

(Table 4.16: Internal validation of K-means by Dunn index) 52

(Table 4.17: Hopkins statistics of cluster 7 and 8) 55
(Table 4.18: Silhouette width of sub-clusters of cluster 7)..... 56
(Table 4.19: Dunn index of sub-clusters of cluster 7) 56
(Table 4.20: Silhouette width of sub-clusters of cluster 8)..... 56
(Table 4.21: Dunn index of sub-clusters of cluster 8) 56
(Table 4.22: University of cluster A)..... 58
(Table 4.23: Universities of cluster B) 58
(Table 4.24: Universities of cluster C) 58
(Table 4.25: Universities of cluster D) 58
(Table 4.26: Universities of cluster E)..... 58
(Table 4.27: Universities of cluster F)..... 59
(Table 4.28: Universities of sub-cluster GA) 59
(Table 4.29: Universities of sub-cluster GB)..... 59
(Table 4.30: Universities of sub-cluster GC)..... 59
(Table 4.31: Universities of sub-cluster GD) 59
(Table 4.32: Universities of sub-cluster GE)..... 59
(Table 4.33: Universities of sub-cluster HA) 59
(Table 4.34: Universities of sub-cluster HB)..... 60
(Table 4.35: Universities of sub-cluster HC)..... 60
(Table 4.36: Clusters' ranking in all the majors)..... 64
(Table 4.37: Majors' ranking in each cluster) 64
(Table 4.38: The original world rank of universities in each cluster) 65

CHPATER 1 – INTRODUCTION

1.1 Background

Rob, Lawrence and Davis (2015) reported that the university rankings of research performance currently play an important role in national policies. The universities with good reputation have dedicated resources to assess areas such as the acknowledgement of education level, the improvement of knowledge and technologies, the publication of research papers and the achievement of their alumni (Olcay and Bulu, 2017). Therefore, the rankings of universities will significant influence on both educational institutions' social situation and prestige.

However, there are still some problems in relation to university ranking that should be of concern. Davis (2016) pointed out that some faculty members doubted whether the ranking result made by some ranking institutions is exactly measured according to the data they collected, and the author also argued that university rankings cannot take responsibility of conveying understandable ranking criteria without improving technical ranking methods. An investigation of two famous ranking institutions (ARWU and THE) conducted by (Saisana, d'Hombres and Saltelli, 2011) leads to the result that these institutions were influenced by politics issues while they are ranking universities. The authors also indicate that although these rankings are specific, they still do not necessarily contribute to satisfying the requirements from students and educational policy makers.

1.2 Research Objectives

Instead of generating specific numerical ranking for each university, this paper aims to use hierarchical clustering technique to create new rankless clustering of universities based on ARWU ranking data where each criterion may include several numbers of universities. Therefore, the research question can be formulated as follow:

"Can the improved rankless clusters be found for universities from the ARWU data based on hierarchical clustering method, which is no longer linear but structured based on clustering criteria that can reveal the major strength of each group?"

1.3 Research Hypotheses

The goal of this research is that the new type ranking criteria can be built successfully, which will lead to the hypothesis of this research:

Null hypothesis: The better rankless clusters cannot be found for universities from the ARWU data based on hierarchical clustering method, which is no longer linear but structured based on clustering criteria that can reveal the major strength of each group.

Alternative hypothesis: The better rankless clusters can be found for universities from the ARWU data based on hierarchical clustering method, which is no longer linear but structured based on clustering criteria that can reveal the major strength of each group.

1.4 Research Methodologies

The data used in this research is gathered from ARWU official website (http://www.shanghairanking.com/), which is one of the most authoritative ranking institution around the world. In this research, the top 500 universities' ranking data around the world over the past 10 years (2008 – 2018) was gathered for analysis. The core method used in this research is the hierarchical clustering technique which is used to investigate the inside patterns from the ARWU ranking data and these patterns will be presented as clusters.

The advent of the Big Data era has generated a lot of unlabeled data to be organized into useful clusters, which is called Unsupervised Learning (Rubio, Palomo, & Francisco, 2018). Unsupervised learning is a Machine Learning approach, which is able to learn latent patterns from ordinary datasets. Clustering algorithms have been divided into two major categories, hierarchical and partitional clustering. The former results in nested clusters, and the latter results in non-nested clusters (Kim, Kohane, & Lucila, 2002). Hierarchical clustering aims to ranking observations into clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster (Sinharay, 2010). Therefore, in this research, the new type of university ranking criteria will be created based on the hierarchical clustering technique.

1.5 Scope and Limitations

The scope of this paper is to create clusters based on relevant educational features included in aggregated ARWU university ranking data over the last ten years. Consequently, these clusters will be used to create criteria for the rank of university and the clusters will be marked to letters in alphabetical order.

The limitation of this paper is that as the hierarchical clustering technique is used to create ranking criteria for universities in this paper, these universities will be grouped into clusters so that these clusters of universities cannot be directly compared with the linear ranking result created by ARWU.

1.6 Document Outline

The remaining of the paper is organized as follows. Chapter 2 reviewed what ranking techniques were used by three famous ranking institutions and some state-of-art implemented for ranking universities, in chapter 3, the process of experimental design of hierarchical clustering model including evaluation will be described along with relevant methodologies. Chapter 4 initially describes the implementation of the experiment and evaluation based on chapter 3, and then the result of them will be demonstrated followed by analysis and discussion. A conclusion summarizes the paper and highlight future works. An overall conclusion and self-assessment of the related work will be described in chapter 5, which importantly contributes to future work presented at the end of this dissertation.

CHAPTER 2 - LITERATURE REVIEW

This chapter will comprehensively introduce how the relevant techniques contribute to university rankings. There are two parts contained in this chapter, the first one describes the background and methodology used in the three most authoritative world ranking institutions – ARWU, THE and QS. A critical assessment of the ranking methods used by these institutions will be conveyed. In the second part of this section an overview of other state-of-art is presented.

2.1 Background and application domain

2.1.1 University ranking institutions

The initial worldwide university ranking institution was published in 2009 by the Academic Ranking for World Universities also known as ARWU, which is currently operated by Shanghai Ranking Consultancy (Pavel, 2015). According to a paper published by Saisana, Hombres and Saltelli (2011), ARWU ranks universities from four different criteria which are: (1) Quality of Education, (2) Quality of Faculty, (3) Research Output, (4) Per Capita Performance. These criteria are composed of six indicators (**Table 2.1** in page 15). The ranking is computed based on total score of each university. **Equation 2.1** demonstrates how the total score is calculated using above criteria:

$$total\_score = 0.1 * alumni + 0.2 * award + 0.2 * hici + 0.2 * ns + 0.2 * pub + 0.1 * pcp$$

(Equation 2.1)

By using the qualification weighting method assigning a weight to each indicator, a given total score of a university will directly determine the specific rank (100 is the maximum score)

Times Higher Education, which is also well known as THE, is another famous world university ranking institution (from United Kingdom). Similar as to ARWU, THE compiles university rankings from five main areas and all these areas contain their own indicators (**Table 2.2** in page 15). However, THE uses different methodology which is a type of Z-scoring to measure the proportion of each indicator unlike ARWU. By the sum of the proportion of indicators based on each area, the final proportion of each area can be computed and the total score can be known from the following **Equation 2.2**:

*total_score= 0.3*teaching + 0.3*research + 0.3*citations + 0.075*international_outlook + 0.025*industry_income* （Equation 2.2）

Criteria	Indicator	Code	Weight
Quality of Education	Alumni of an institution winning Nobel Prizes and Fields Medals	Alumni	10%
Quality of Faculty	Staff of an institution winning Nobel Prizes and Fields Medals	Award	20%
	Highly cited researchers in 21 broad subject categories	HiCi	20%
Research Output	Papers published in Nature and Science	N&S	20%
	Papers indexed in Science Citation Index-expanded and Social Science Citation Index	PUB	20%
Per Capita Performance	Per capita academic performance of an institution	PCP	10%

(Table 2.1: Criteria and indicators of ARWU)

Areas	Indicators	Proportion
Teaching	Reputation survey; Staff-to-student ratio; Doctorate-to-bachelor's ration; Doctorates-awarded-to-academic-staff ratio; Institutional income;	30%
Research	Reputation Survey; Research income; Research productivity;	30%
Citations	Research influence	30%
International outlook	Proportion of international students; Proportion of international staff; International collaboration;	7.5%
Industry income	Knowledge transfer	2.5%

(Table 2.2: Areas and indicators of THE)

The QS World University Ranking is regarded as one of the three authoritative ranking institutions, along with ARWU and THE. Different from the two institution above, QS only build six metrics without revealing indicators in each of them (**Table 2.3**). The overall score of universities can be computed by the following **Equation 2.3**:

*overall_score = 0.4 * academic_reputation + 0.1 * employer_reputation + 0.2 * faculty_student_ratio + 0.2 * citations_per_faculty + 0.1 * international_faculty & student_ratio* (Equation 2.3)

Metrics	Source	Weighting
Academic reputation	International academic survey	40%
Employer reputation	Graduate employers survey	10%
Student and faculty ratio	Measure from teaching commitment	20%
Citations per faculty	Measure from research impact	20%
International faculty & student ratio	Measure from the diversity of the academic staff and student community	10%

(Table 2.3: Metric of QS)

2.1.2 Cumulative table & Criticism

However, although ARWU, THE and QS are the most professional and authoritative ranking institutions, they have some shortcomings. ARWU is criticized for putting specific attention on universities which published more scientific journals. Therefore, this quantitative methodology has its drawback. THE and QS are widely trusted because of the diverse and rigorous consideration of indicators, as it is favouring universities with more publications (such criteria could be debatable). Nevertheless, the global survey conducted by them can cause subjective bias on the final raking result (Pavel, 2015). This is because although the source of the data these three institutions used can be proved as objective, the method they used to weight their indicators has never been revealed and all of them share with the same problem that one or two of their indicators are weighted much more than the other indicators. Therefore, it is reasonable to believe that they are subjectively focusing on some area in relation to university rankings.

Study	Ranking method	Number of Universities	Number of features
ARWU	Weighting	1000	4
THE	Z-Score	1200+	5
QS	Weighting	1000	5

(Table 2.4: Cumulative table of ranking methods)

2.2 Other methodologies

Tabassum et al. (2017) improved several algorithms based on gathering data from the result of QS university ranking, these algorithms including Feature selection, Outlier detection and ranking score of universities which enables create specific ranks for each university. An experiment was operated through these algorithms and training and test datasets developed previously. The ROC curve was illustrated so that the performance of experiment can be demonstrated by the relationship between Accuracy and Deviation.

However, this experiment only leads to several plots which show the performance rather than giving specific accuracy of the model. To summarize, there are some interesting features in the experiment such as the Outlier detection and Feature Selection methods. Nevertheless, the evaluation method should be improved.

(Ivančević and Luković, 2018) completed a research on whether collecting open data from the Ministry of Education, Science, and Technological Development of Serbia can generate new indicators in order to rank Serbia's universities nationally. Each of the indicators is created based on investigating similar features that both occurred in authoritative university ranking institutes and open data collections. The performance of these indicators was measured by absolute value, i.e., higher values represent higher performance. The result of prediction demonstrated that only one university of Serbia was ranked incorrectly. However, considering that only seven universities participated in this study, which is not a representative sample. Moreover, the evaluation of feature selection is not implemented. As mentioned in the article, authors as well as argued that the presented method may not fit all the other universities and is significant in their own research only.

A theoretical framework for a field based ranking system depending on all management departments that belongs to Turkish Higher Education System was proposed by (Alma, Coşkun and Övendireli, 2018). By gathering data from survey, they grouped all the indicators into 6 components and each of these components is assigned by a total weight based on the single weight from each indicator. As a result, total weights of components directly determinate how important these components are in relation to management department and ranking universities. This paper points out an interesting idea based on doing field research as a method of feature selection which is more likely to contribute to rank universities. However, this conceptual framework was not used in a real-life university ranking experiment as the data collected was not representative and it is only a theoretical framework which needs to be tested by experiments.

Cinzia, Bonaccorsib and LéopoldSimarc (2015) argued that university rankings are currently influenced by policy and media. As a result, instead of focusing on seeking drawbacks from exiting university ranking techniques, they created a new method to address the possible problems that may occur in university rankings. The method can

be divided into four parts: (1) Reduce Monodimensionality. (2) Generate new estimators that are sensitive to extreme values and outliers. (3) Use directional conditional efficiency analysis to solve dependence problems on university size and subject mix. (4) Compute technical efficiency indicators based on an explicit input–output structure to address the lack of consideration of the input–output structure. As the result of addressing all these limitations, it has been proven that data integration is more likely to develop current ranking methodologies.

Rebeka, Damjan and Peter (2009) carried out a new university ranking table by improving current existing university ranking tables based on investigating ranking methodologies and indicators selection. In addition to educational and research indicators, environmental ones are considerably focused as well. For the creation of new ranking table, features are initially selected based on sustainability model. As introduced in this paper, sustainability is able to organize features into economic, social and environmental dimensions. Consequently, these dimensions are grouped into three types of indicators listed above. During the process of weighting indicators, AHP (Analytic Hierarchy Process) model was used to reflect the importance of each indicators so as to assign weight to them. To evaluate the consistency of AHP model, judgmental matrix is formed and leads to a consistency ratio R_c. The results show all the matrix is acceptable based on the baseline ($R_c < 1$) which represents the evaluation of it is satisfactory consistent. The aggregation of those weights is used to calculate the rank of universities and the new university ranking table can be formed.

2.2.1 Conclusion table

In order to make a clear comparison between these methodologies, **Table 2.5** demonstrates some relevant issues related to section 2.2.1.

In conclusion, although these methods are able to generating university rankings and have been proven that their result are convincing and reliable, there are still several problems and drawbacks remaining. Some of those methods are based on conducting Machine Learning models such as regression model, but all the results are shown as linear rankings of university. As aforementioned, it is difficult to specifically ranking universities because there are too many issues needed to take into account. It starts to be subjective when narrowing down these issues into several specific areas such as

alumni's working condition, papers published or highly cited research. As a result, rather than using these kinds of 'narrowed' data to generating linear universities rankings, this paper aims to build a new type of university ranking criteria by using hierarchical clustering technique.

Countries	Ranking method	Number of universities	Number of features	Reference
Global	Machine Learning	1000	5	Tabassum et al. (2017)
Serbia	Qualify features	7	10	Ivančević and Luković, (2018)
Turkey	Weighting	1000	6	Alma, Coşkun and Övendireli (2018)
USA & UK	Weighting	35	3	Rebeka, Damjan and Peter (2009)
Europe	Reduce existing limitation	313	8	Cinzia, Bonaccorsib and LéopoldSimarc (2015)

(Table 2.5: Cumulative table of university ranking methods)

2.3 Method review

This section reviewed the methodologies used for later experiments, the design processes and implementation of them are presented in Chapter 3 and Chapter 4, respectively.

2.3.1 Pearson's correlation

It has been proved by (Puth, Neuhäuser, & Ruxton, 2015) that Pearson's correlation can effectively reflect the linear correlations between two variables and even can discard the assumption that the distribution of variables has to be normal.

2.3.2 Linear regression

Besides being implemented as a supervised learning technique, linear regression can be used as an important feature selection method. To acquire the interactive correlations between variables, a regression model will analyse the main information by focusing on the inside pattern of data and eventually grasp the main features (Gan et al., 2018). Therefore, as feature selection is the goal of this section, implementing a regression model for getting the importance of each feature is needed.

2.3.3 Principle component analysis (PCA)

"The central idea of principal component analysis is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set." from Jolliffe (2002).

(Cangelosi and Goriely, 2007) summarized the twelve main methods which can indicate how many components should be retained during the process of analyzing PCA result. They indicated that before starting the experiment, log transformations and normalization of data should be included in data processing as these two methods are helpful for formatting data on the same scale. They also stated that Guttman-Kaiser rule is one of the most common way to choose the optimal number of components, which is selecting those components with eigenvalues higher than the average eigenvalue. Rencher (1998) pointed that if the variables in data are highly correlated, changing the average rule mentioned above to a cut-off of 70% of the average value will perform better. Therefore, as aforementioned, here the cut-off rate is set up to 70% of average eigenvalue because most of the numeric variables in the aggregated ARWU data are highly correlated.

2.3.4 Clustering methods

Clustering techniques are extensively used because they are best at analysing the intrinsic information from data without requiring labels (unsupervised methods). Accordingly, as an inductive Machine Learning task, the clustering method can analyse the input data when the predefined target variable in the data is lacking. Unlike any classification or regression method, the evaluation of clustering model cannot be tested after the clusters are generated because of the lack of target value (actual result). Nonetheless, rather than evaluating the clusters, the performance of the clusters built can be measured and this measurement is also called validation. Clustering is executed based on investigating the similarity or the dissimilarity of each observation. Similarity and dissimilarity can be computed by a few distance metrics which are executed based on subjective determination by users.

2.3.5 Partitioning clustering method

Partitioning clustering approach refers to the process that "partitions the data points into k clusters such that the data points within a cluster are more similar to each other than data points in different clusters" (Zhao, Han and Pan, 2010).

K-means is one of the widest used partitioning clustering methods which can divide all the observations from a dataset in to k (a random selected number) subsets so these subsets are used as k clusters (Nisha et al, 2015). Accordingly, similar as hierarchical clustering algorithm, k – the number of clusters generated depends on the users' inductive bias i.e. the requirement of number of clusters in real-world experiments. As k is determined by the users, so the number of (k) clusters will not be changed. The major theory of this method is to focus on reducing the sum of dissimilarities between pairs of observations. The first step of this method is to arbitrarily split the dataset into k clusters by specifying k-central points which are observations or data points taking the k lowest mean or median dissimilarity among all the other observations in a dataset. Once the central points are computed, all the other observations will be assigned to the nearest k-clusters based on the distance between each observation and clusters. The last step of this approach is that after assigning all the observations to k-clusters, new central points and assignments will be repeated until the best similarities of all the observations in the dataset are acquired (Chouhan and Chauhan, 2014).

To summarize, K-means is conducted by performing the following steps:

a) Set up K (arbitrary) number of clusters based on requirement.

b) Finding central points of each cluster

c) Assigning observations to their nearest clusters

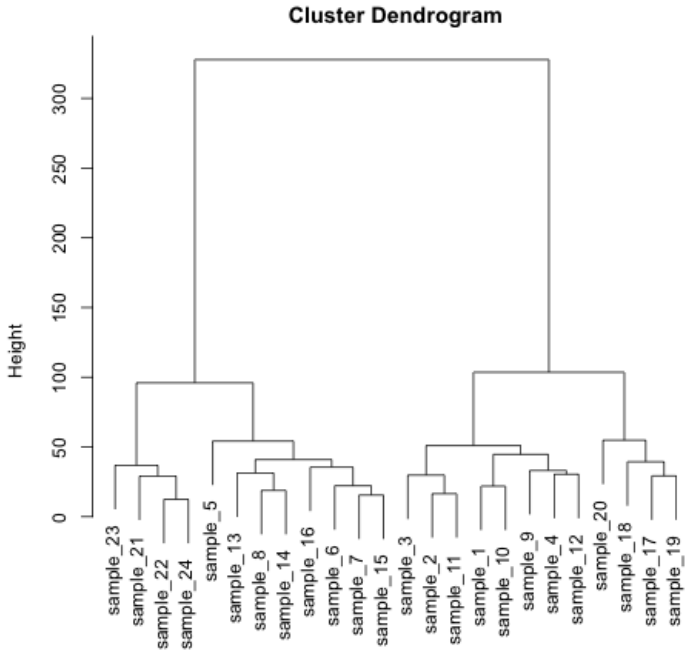
d) Repeat step a-b.

2.3.6 Hierarchical clustering

There are two main categories of clustering techniques: Hierarchical clustering and Partitioning clustering. Hierarchical clustering is formed by representing all the clusters in a hierarchy or tree which are commonly demonstrated from dendrograms. All the clusters can be specified by means of spotting the nodes from each tree and elements in

each cluster can be located at the bottom of each tree or hierarchy. There are two types of hierarchical clustering methods performed as agglomerative and divisive, respectively (Nisha and Kaur, 2015). The Divisive hierarchical clustering method also known as "divisive cluster analysis" offers a "top down" method in the inverse sequence in contrast with agglomerative clustering's "bottom up" clustering technique (Nazari et al., 2015).

Divisive cluster analysis also referred as DIANA initially merges all the observations into one cluster, and then iteratively splits observations from the initial cluster into two new clusters based on the dissimilarity of each observation. The splitting stops when a threshold value is satisfied. Agglomerative clustering method performed in a reverse way from the divisive clustering method. Clusters implemented by agglomerative clustering model will be such that will initially regard every observation as one cluster preliminarily and merge these clusters based on the similarity of each other. In order to demonstrate the process and result of hierarchical clustering method, a dendrogram is demonstrated in **Figure 2.1**.



(Figure 2.1: Example of a dendrogram)

CHAPTER 3 - DESIGN AND METHODOLOGY

This chapter of the paper aims will be presented in two parts. A specific exploration and analysis of ARWU university ranking data is going to be conveyed initially. Although ARWU ranking methods are controversial because of its subjective weighting technique, Docampo (2011) reported that ARWU is the only institution concentrating on research output and does not contain subjective data as it will not try to assess the quality at measurable research output. In order to lead a holistic overview of design, a detailed introduction of hierarchical clustering technique and how it was used to generate a new type of ranking will be demonstrated.

3.1 Exploratory analysis

3.1.1 Data pre-processing

The data was collected from Kaggle which is a raw extraction of ARWU's official website and contains worldwide university rankings from 2008 to 2018. The main features of data have been presented in **Table 3.1**. Besides, a comprehensive summarization of the data is shown (**Table 3.2**) below. As it is shown in **Table 3.2**, the data is comprised of 5504 entries and 11 features with 3 categorical and 7 numerical features. Except all the categorical data, all the rest of numeric ones include missing values and it will cause many problems during the process of analysis. In order to keep a representative sample, some of missing values will be replaced with the average values in each of feature. Rather than removing these missing values directly, replacing them with average values is more reasonable as there are other useful values such as university name existing in the same row where the missing value is. Therefore, simply removing missing values by each row containing them will ruin the representativeness of the data. However, the feature total_score is made of excessive missing values (4403 NAs) unlike any other features that merely have maximum 20 missing values. Comparing to the correct result as it should be, it is impossible to use average values to take the place of missing values in total_score column, because all the average values will be explicitly same which is completely wrong in this case as the total score is used to create the world rankings, so it is certainly wrong that over 3000 universities have the same average total score. However, the total score of each university can still be computed based on the **Equation 2.1** derived from ARWU's methodology.

world_rank	university_name	national_rank	total_score	alumni	award	hici	ns	pub	pcp	Year
1	Harvard University	1	100	100	100	100	100	100	79.6	2018
2	Stanford University	2	75.6	44.5	88.5	76.6	78.6	76.5	56	2018
3	University of Cambridge	1	71.8	82.3	95.4	56.7	57.6	70.9	59.5	2018
4	Massachusetts Institute of Technology (MIT)	3	69.9	70.9	83.6	52.5	71.4	64.4	70.3	2018
5	University of California, Berkeley	4	68.3	65.6	78.4	61.3	67.8	65.1	58.2	2018
6	Princeton University	5	61	55.8	97.9	44.9	47.1	44.2	73.3	2018

(Table 3.1: Overview of ARWU data)

	non-missing	missing	missing percent	mode	Mean	min	max
total_score	1101	4403	80	336	36.39	23.8	100
alumni	5504	0	0	349	8.26	0	100
award	5504	0	0	306	7.26	0	100
hici	5504	0	0	434	15.98	0	100
ns	5479	25	0.45	539	15.37	0	100
pub	5504	0	0	638	38.69	7.3	100
pcp	5504	0	0	427	21.6	8.3	100
year	5504	0	0	11	2013	2008	2018
world_rank	5504	0	0	Null	Null	Null	Null
university_name	5504	0	0	Null	Null	Null	Null
National_rank	5504	0	0	Null	Null	Null	Null

(Table 3.2: Summarization of ARWU data)

In order to create university ranking criteria by means of hierarchical clusters, an aggregation for the ARWU data should be processed. It is because hierarchical clustering works by measuring the distance between data points, so that the result of hierarchical clustering cannot specify each data point without giving a unique identification of it. As the ARWU dataset used in this paper contains worldwide university rankings over 10 years, there are vast numbers of repetitive university name from different years. Therefore, the data was aggregated based on the names of universities and the rest of indicators were computed as the mean over 10 years, so that

it allows hierarchical clustering to acquire specific identification of each data point from its unique university name. Another precondition of this aggregation is that the year column in the data should be removed because of the reason listed above and it is not statistically significant to average years. **Table 3.3** demonstrates the aggregated method used for each indicator in the data along with their criteria.

Criteria	Aggregate method	Indicators
University name	Group	university_name
Quality of education	Average	Alumni
Quality of Faculty	Average	Award
	Average	HiCi
Research Output	Average	N&S
	Average	PUB
Per Capita Performance	Average	PCP

(Table 3.3: Aggregation method for ARWU data)

3.2 Feature selection

3.2.1 Pearson's correlation

Examine the multiple correlations between numeric variables is a quite important step before conducting feature selection. The method used to examine correlations is Pearson's correlation based on the assumption that a Linear relationship exist in ARWU numeric variables. In order to test the significance of correlations, a p-value was set up to 0.01 in the assumption. To specifically investigate the Pearson correlation coefficient between variables, a correlation coefficient table need to be calculated. Moreover, a visualization of the Pearson correlation coefficient, the P value, and the histogram of each variable is required to be demonstrated so as to have an intuitive overview of how variables are correlated. The Pearson correlation coefficients (r) indicates three different types of the extents of how variables are correlated. In this case, the Pearson correlation coefficients (r) indicates that, except variable "national_rank", all the variables are at least moderately (absolute r value between 0.3 and 0.7) correlated with other ones, and most of them are strongly (absolute r value between 0.7 and 1.0) correlated. In contrast, there are some weak correlations (absolute r value between 0.0 and 0.3) in relation to variable national_rank. Furthermore, as the p-value is visualized (Each significance level is associated to a symbol: p-values (0, 0.001, 0.01, 0.05, 0.1, 1) <=> symbols("***", "**", "*", ".", " "), all the correlations are statistically significant (with p-value = 0.001)

except the relationship between hici and national_rank (with p-value = 0.05) but is also significant (at 0.01 level).

To summarize, considering that all the relationships between numeric variables except national_rank in ARWU are moderately and strongly correlated and can be proved to be statistically significant (all the p-values < 0.1), the national_rank variable needs to be discarded in the following work.

3.2.2 Linear regression

The interpretation of the multiple regression model corresponds to several phases. The first step of interpreting the model is to examine F-statistic and the relevant P-value, if these two values are lower than .05, the model can be considered as it has statistical significance, so that the model is acceptable to be used for further analysis. According to the correlation coefficients in the model, whether a variable in the model is significant can be verified.

Further inspection is required to validate the regression model, because it is crucial to assess the performance of the model as it can affect the final result of the importance of each feature. Therefore, the validation of the model is an important way to estimate the model and needed to be implemented in the following work. Cross-validation is one of the most popular validation techniques to measure the performance of model, it focuses on measuring whether the model performed as expected during the process of generating models. Cross-validation is a popular validation method which can effectively estimate the average prediction error. The fundamental methodology of cross-validation is to repeat the process of building a model while specifying a same length fold of data and use this left-out fold as testdata in each time of repeat. Moreover, the testdata chosen in cross-validation in each time should be same size but with completely different content comparing with other testdata in different times. Cross-validation will eventually summarize the results given by each result derived from these testdata.

3.2.3 Principal component analysis (PCA)

Due to the fact that most of the variables in the aggregated ARWU data are strongly correlated since the variable “national_rank” was removed, it is appropriate to implement PCA to the data and it also has been widely used in feature selection. It is

because PCA is more convenient to summarize and visualize multiple inter-correlated quantitative variables, and the goal of PCA is to reduce multidimensional data with remaining most information of data which is beneficial to removing redundant features and improving the speed of building clustering models.

In this paper, PCA was used depending on the assumption that whether a lower dimensional feature set computed or gathered from original aggregated ARWU data can represent an acceptable percentage of population variance according to original ones. The aim of PCA in the following work is to combine the internal patterns of correlated variables from the data into a new dimension reducing the dimensionality of the original data accounts. In the nest step of this work the PCA analysis will be performed through: reliability testing, result presentation and result interpretation.

3.3 Feature supplement

In the original ARWU data, the country name of each university is demonstrated as its national flag. However, these flags are designed to be functional buttons i.e. a national ranking classifier which enable candidates to view national university rankings in every country in the ARWU data. Accordingly, it is impossible to capture the country column from the data because buttons on website cannot be directly transformed to vectors or strings that can be stored into data. In order to remain the complexity of the data in this case, the information of country name was scraped and loaded into the aggregated ARWU data by implementing a Python package called Beautifulsoup which allows users to get access to HTML files and gather data from them.

3.5 Design procedure and metrics

This section will discuss the processes of conducting the experiment for building new type criteria of university ranking. The different methods used in creating hierarchical clusters and what types of clustering method are required to make a comparison will be demonstrated.

Various types of distance metrics are usually used to measure the similarity or proximity when conducting a clustering experiment. A Distance measurement indicates the internal patterns showing how close some data points are to other ones. A shorter distance commonly means that there is a stronger similarity between observations and

these can be considered to join into clusters. The most common used distance metrics are: Euclidean distance, Squared Euclidian distance, Manhattan distance, Maximum distance, Mahalanobis distance. The Maximum distance metric has been used in this experiment, and the reason for choosing this metric has been explained in section 4.4.2.

The Maximum distance metric also known as Chebyshev distance measures distance by given two points p and q as the following Equation 3.1:

$$D_{Chebyshew}(p,q) = \max(|p_i - p_q|) \qquad \text{Equation 3.1}$$

As aforementioned, agglomerative approach splits data points initially to clusters in pairs. The linkage method is the criterion of determination of pairwise distance between observations. There are four types of linkage criteria that are widely used: complete-linkage clustering, average-linkage clustering, single-linkage clustering and Ward linkage clustering. **Table 3.4** specifically introduces how these linkage criteria work in a dataset. Ward method was chosen to implement as the linkage criterion in this experiment. The reason for this will be discussed in section 4.4.2.

Assuming that there are clusters r and s, and each observation is denoted by n_r and n_s, the linkage methods can be implemented as follows:

Linkage method	Formulation
Complete linkage:	$dc(r, s) = min\{d(x_{rw}, x_{sy})\}$
Average linkage:	$dc(r, s) = max\{d(x_{rw}, x_{sy})\}$
Single linkage:	$dc(r, s) = \frac{1}{n_r n_s}\sum_{w=1}^{nr} d(x_{rw}, x_{sy}) \sum_{y=1}^{ns} d(x_{rw}, x_{sy})$
Ward's linkage:	$dc(r, s) = \|\| x_{rw} - x_{sy} \|\|^2$

(Table 3.4: Linkage methods)

3.6 Validation

In general, there are two types of validation methods for measuring the fitness of clustering models: internal validation and external validation. In this case, internal validation of clusters will be implemented as the validation method of this clustering experiment because the process of implementing external validation requires comparison between the clustering result and the known, correctly labelled target variable (Zerabi and Meshoul, 2017).

3.6.1 Internal validation

The purpose of clustering objects is to make all the objects in the same cluster be as similar as possible and objects in different clusters as distinctive as possible (Liu et al., 2013). Therefore, the internal validation focus on two criteria:

a) Compactness

The compactness of a cluster is a measurement of to what extent the objects in the same cluster are close to each other. The way of measuring the compactness of a cluster is its variance where lower variance indicates better compactness.

b) Separation

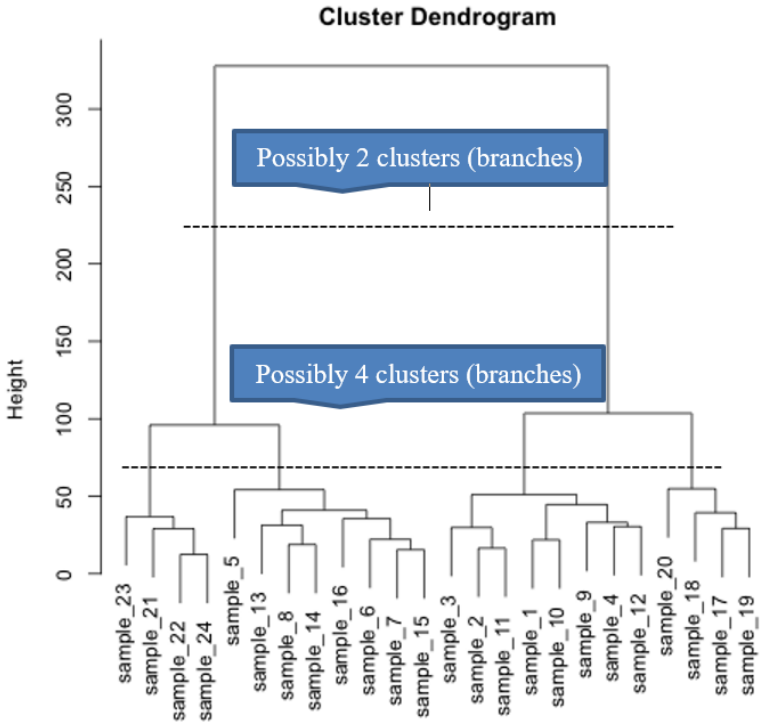
The separation of clusters is the extent of the distinctiveness of clusters, i.e., how well the clusters are separated. The measurement of the separation of clusters depends on pairwise distances, pairwise minimum distance, density and so on.

3.6.2 Design of validation for hierarchical clusters

The design of validation is carried out by two main steps:

a) Initially, generating a dendrogram to approximately identify the possible number of clusters based on counting the number of branches which are widely separated from each other. One of the advantages of hierarchical clustering technique is that the dendrogram can intuitively allow users to specify the probable number of clusters by looking at the well separated branches. **Figure 3.2** demonstrates how to choose appropriated clustering numbers with the help of a dendrogram.

As is shown from **Figure 3.1**, the head of dendrogram tree is always linked as clearly separated, so it is convenient to acquire the possible optimal number of clusters by cutting these kinds of branches.



(Figure 3.1: Choosing optimal number of hierarchical cluster)

b) Once the set of optimal numbers of clusters have been chosen, an internal validation is required to be executed. As mentioned in section 3.4.1, the main criteria of the internal validation are checking the compactness and separation of clusters. Therefore, this experiment offers two types of validation of the clustering result, (the Silhouette index and Dunn index), which are both performing well on measuring the compactness and separation of clusters.

Silhouette width

The Silhouette width measures to what extent an observation is close to its own cluster (compactness) comparing to other clusters (separation). The range of Silhouette width is [-1,1], strongly positive values indicate that the observations inside a cluster are well compact and all the clusters are well separated, and vice versa. A positive Silhouette is more likely to be considered as clusters are well separated (Cichosz, 2015).

For a given observation x from cluster d related to a dissimilarity measure σ on dataset S, the Silhouette can be expressed as the following **Equation 3.2 – 3.4**:

$$WS_{\delta,s}(x,d') = \frac{min_{d'}\,\Delta_{\delta,s}(x,d') - \Delta_{\delta,s}(x)}{max\{min_{d'}\,\Delta_{\delta,s}(x,d'), \Delta_{\delta,s}(x)\}}$$ Equation 3.2

where Δσ,s(x,d') is the average dissimilarity between x and all the observations from another cluster d':

$$\Delta_{\delta,s}(x,d') = \frac{1}{|s^{d'}|}\sum_{x \in s^{d'}} \delta(x,x')$$ Equation 3.3

and Δ(δ,s) (x) represents the average dissimilarity between x and other observation in its same cluster d.

$$\Delta_{\delta,s}(x) = \frac{1}{|s^{d}|-1}\sum_{x \in s^{d}-(x)} \delta(x,x')$$ Equation 3.4

Dunn index

The Dunn index measures the ratio of the minimum separation of clusters and the maximum compactness in each cluster (Cichosz, 2015). Therefore, higher values of Dunn index are preferred as value of the separation of clusters is ideally considered to be more and the value of compactness of a cluster is preferred to be less. The formulation of Dunn index is related to dissimilarity measure σ, in D = {D_1,…,D_n} clusters and dataset S as **Equation 3.5** :

$$\mathrm{Dunn}_{\delta,s} = \frac{\min\{\mathrm{Sep}_{\delta,s}\,(\mathrm{D_i},\mathrm{D_j})\}}{\max\{\mathrm{Comp}_{\delta,s}\,(\mathrm{D_n})\}}$$ Equation 3.5

CHAPTER 4 – IMPLEMENTATION, RESULTS, EVALUATION AND DISCUSSION

This chapter will firstly discuss the experiment undertaken such as package chosen and function and argument set up. The second part of this chapter aims to lead to specific interpretation of experiment results which are mostly demonstrated by tables and figures so as to be straightforward. All the processes or preparation of experiment except feature supplement were implemented based on programming language R.

4.1 Data processing

4.1.1 Missing data

Imputing missing data is an important part in data pre-processing due to the fact that many machine learning algorithms built in R cannot be implemented with missing data, which will result to an information lost if the missing data is only simply removed. In this case, the initial report of the missing data is shown in **Table 4.1,** which is generated with the help of package **dataQualityR** in R. For the reason that all the categorical data in the ARWU data does not contain any missing values, so that **Table 4.1** merely presents the data quality report of numerical data. **Table 4.2** leads to the presentation of the missing values included in the ARWU data.

	missing	missing percent	Mean	min	max
total_score	4403	80	36.39	23.8	100
alumni	0	0	8.26	0	100
award	0	0	7.26	0	100
hici	0	0	15.98	0	100
ns	25	0.45	15.37	0	100
pub	0	0	38.69	7.3	100
pcp	0	0	21.6	8.3	100
year	0	0	2013	2008	2018

(Table 4.1: Missing numerical data report)

world_rank	university_name	national_rank	total_score	alumni	award	hici	ns	pub	pcp	year
151-200	London School of Economics and Political Science	17-21	NA	23.8	16.1	0	NA	31.3	27.7	2018
201-300	University of Toulouse 1	14-Sep	NA	0	29.4	9.6	NA	11.6	32.2	2018
401-500	Stockholm School of Economics	11-Oct	NA	0	16.1	0	NA	11.2	42.4	2018
151-200	London School of Economics and Political Science	18-20	NA	23.8	16.1	10.9	NA	29.7	28.1	2017
301-400	University of Toulouse 1	15-17	NA	0	29.4	0	NA	11.1	30.8	2017

(Table 4.2: Missing values in the ARWU data)

As it is shown in **Table 4.1** and **Table 4.2**, the "total_score" and ns are the two variables that consist of missing values. Especially in the "total_score" column, 80 percent of the data is missing so that it is necessary to impute the missing data in several ways rather than directly removing these rows that include missing values, which will cause the loss of 80% of ARWU data.

Due to the fact that the "total_score" column contains various of missing data derived from the original ARWU data, it is certain that the mean value of this column cannot be used to replace the missing value. However, the accurate value of the "total_score" can be obtained base on the ARWU's equation shown in chapter 2. The final result can be regarded as the replacement of the missing value in the "total_score" column by using **ifelse()** function as follow:

```
Result %>%

0.1 * ARWU$alumni + 0.2 * ARWU$award + 0.2 * ARWU$hici + 0.2 * ARWU$ns
+ 0.2 * ARWU$pub + 0.1 * ARWU$pcp

ARWU$total_score%>%

ifelse(is.na(ARWU$total_score), result , ARWU$total_score)
```

This code primarily functions such that all the missing values from "total_score" are replaced by the value from the variable *result* which is derived from the weighting equation by ARWU.

While imputing the variable *ns* in this case, it is decided to use the average value of *ns* to replace the missing values. This is because the missing values (which are missing at random) in the variable *ns* are not too many and this experiment is conducted based on multiple variables, so mean imputation for *ns* is acceptable as it contributes to full size of the ARWU data and does not generate huge bias to the data. With the following code, the imputation of the variable *ns* will be achieved. The coding logic of implementing the imputation is same as the last one.

```
ARWU$ns %>%

ifelse(is.na(ARWU$ns), round(mean(ARWU$ns,na.rm = T),1),ARWU$ns)
```

As a result, the imputed ARWU data is shown in **Table 4.3.**

world_rank	university_name	national_rank	total_score	alumni	award	hici	ns	pub	pcp	year
151-200	London School of Economics and Political Science	17-21	98	23.8	16.1	0	15.4	31.3	27.7	2018
201-300	University of Toulouse 1	14-Sep	74.1	0	29.4	9.6	15.4	11.6	32.2	2018
401-500	Stockholm School of Economics	11-Oct	70.3	0	16.1	0	15.4	11.2	42.4	2018
151-200	London School of Economics and Political Science	18-20	68.5	23.8	16.1	10.9	15.4	29.7	28.1	2017
301-400	University of Toulouse 1	15-17	66.9	0	29.4	0	15.4	11.1	30.8	2017
401-500	Stockholm School of Economics	11	59.7	0	16.1	0	15.4	9.6	41.2	2017

(Table 4.3: Imputed ARWU data)

4.1.2 Aggregation

An aggregation of currently supplementary ARWU data needs to be conducted due to the fact that this experiment aims to create new university ranking criteria over the last 10 years (from 2008 to 2018) so that it is reasonable to aggregate the ARWU data by each school over years. Before conducting the aggregation, the variables "total_score", "world_rank" should be provisionally removed because aggregating the ARWU data may cause the same world rankings or total score of universities. However, the result of the "total_score" and "world_rank" variables will be computed again after the aggregation of the ARWU data by using the same coding aforementioned in section 4.1.1. To achieve this goal, a built-in function called **aggregate()** from R should be used which simply supplies ways for users to aggregate data using several types of arithmetic methods. The implementation of is demonstrated below:

```
Aggarwu %>%

aggregate(x=finalarwu2[,c(1,9)],by=list(finalarwu2$university_name),
FUN=mean)
```

This code addresses the aggregation problem by grouping the ARWU data by the names of universities with computing the mean of all the variables in the data. The variable year is permanently removed as the aggregated year makes no sense and is not needed in the experiment. The aggregated data is presented in **Table 4.4** in page 35**.**

It is worth to notice that variable *world_rank* and *national_rank* will not be aggregated at this step, because the *world_rank* variable is derived from the variable *total_score* and will be acquired based on the aggregated *total_score*. However, it is not possible to

calculate variable *national_rank* after the data being aggregated because it is derived from the countries of the universities in the data. Therefore, it is reasonable to supplement the countries in to the aggregated ARWU data.

university_name	total_ score	alumni	award	hici	ns	pub	pcp
Harvard University	100	100	100	100	100	100	74.76
Stanford University	73.55	41.16	83.05	84.67	71.17	71.41	55.13
University of California Berkeley	70.62	66.78	78.88	66.98	68.8	68.31	55.45
Massachusetts Institute of Technology (MIT)	70.53	69.81	81.45	62.55	70.76	61.77	64.68
University of Cambridge	69.99	82.99	95.22	53.56	55.52	66.71	57.19

(Table 4.4: Aggregated ARWU data)

While the feature “Country” introduced in section 4.2 is valuable for understanding the location of universities, this can also introduce a specific bias since clusters may form around this already known parameter. This feature will not be included in the analysis at this stage.

4.2 Feature supplement

In this step of the experiment, the country name of each university is considered to be supplemented in to the aggregated ARWU data for future analysis. The purpose of it is to classify the countries during the process of recomputing the *national_rank* variable. Unlike another process in the experiment, the programming language Python will be used because of its convenience and quick speed of web crawling. The use of package **Beautifulsoup** from Python allows user to extract content as information from HTML webpage. By inspecting the source of ARWU ranking web page, the country name of each university is attached in the **title** argument within the <**a**> element. There are three major procedures when gathering the data:

a) Using package **Request** and function **url()** inside it to access the content of ARWU ranking webpage while typing in the web address into **url()** function.

b) Conducting **find.all()** function from package **Beautifulsoup** in order to get access to every element existing the webpage. By setting argument as “a” inside **find.all()** function, all the <**a**> element will be extracted.

c) For the purpose of acquiring country name from the **title** argument in each <**a**> element, function **get()** from package **Beautifulsoup** is used. In this case, the

code **get("title")** is used which will specify the content i.e. country name from the **title** argument.

d) The last step is to run a **for loop** to repeat step 2 and 3 until all the country names are gathered.

The supplemented ARWU data is shown in **Table 4.5** below.

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank
USA	Harvard University	100	100	100	100	100	100	74.76	1	1
USA	Stanford University	73.55	41.16	83.05	84.67	71.17	71.41	55.13	2	2
USA	University of California Berkeley	70.62	66.78	78.88	66.98	68.8	68.31	55.45	3	3
USA	Massachusetts Institute of Technology (MIT)	70.53	69.81	81.45	62.55	70.76	61.77	64.68	4	4
UK	University of Cambridge	69.99	82.99	95.22	53.56	55.52	66.71	57.19	1	5

(Table 4.5: Supplemented ARWU data)

4.3 Feature selection

From this step, several packages will be imported from the community of R for experimental use. Feature selection is important in Machine Leaning, it enables users to find which variables are efficient and effective to contribute to a Machine Learning experiment. The supplement variable *country* will not be implemented in this step as it is a categorical variable and only used for formulating variable *national_rank* the advantages of it are: reduce the dimensionality of a data; reduce the chance of overfitting for a model; reduce the complexity of a model.

4.3.1 Pearson correlation

The major manner of conveying the result of Pearson correlation from aggregated ARWU data is Pearson coefficient table and data visualization. A Pearson correlation coefficient table can give an overview of to what extent that the variables are correlated with each other. Package **Hmisc** supplies a function called **rcorr** which can generate a Pearson coefficient table (**Table 4.6)** and a probability value table (**Table 4.7**) while setting up the argument **type** to Pearson. By analysing the probability table, whether the correlations are significant can be decided. In this case, the significant level is set up to .05, which means all the probability values derived from Pearson correlation analysis are considered as statistically significant if they are smaller than .05.

	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank
total_score	1.00	0.80	0.84	0.87	0.93	0.76	0.78	-0.24	-0.78
alumni	0.80	1.00	0.76	0.58	0.70	0.49	0.64	-0.19	-0.55
award	0.84	0.76	1.00	0.62	0.73	0.43	0.71	-0.21	-0.54
hici	0.87	0.58	0.62	1.00	0.83	0.62	0.63	-0.08	-0.69
ns	0.93	0.70	0.73	0.83	1.00	0.67	0.71	-0.15	-0.73
pub	0.76	0.49	0.43	0.62	0.67	1.00	0.46	-0.35	-0.75
pcp	0.78	0.64	0.71	0.63	0.71	0.46	1.00	-0.29	-0.61
national_rank	-0.24	-0.19	-0.21	-0.08	-0.15	-0.35	-0.29	1.00	0.29
world_rank	-0.78	-0.55	-0.54	-0.69	-0.73	-0.75	-0.61	0.29	1.00

(Table 4.6: Pearson correlation coefficient table)

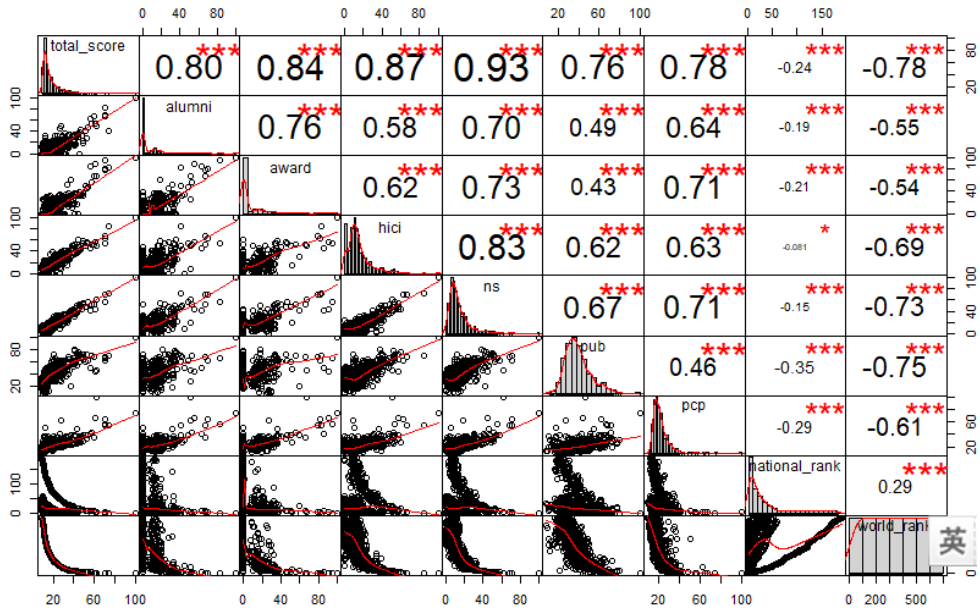
Table 4.6 shows that the variable "national_rank" is the sole variable in the aggregated ARWU data that is poorly correlated with other numeric variables with most of its Pearson coefficient are smaller than 0.3.

	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank
total_score	NA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
alumni	0.000	NA	0.000	0.000	0.000	0.000	0.000	0.000	0.000
award	0.000	0.000	NA	0.000	0.000	0.000	0.000	0.000	0.000
hici	0.000	0.000	0.000	NA	0.000	0.000	0.000	0.032	0.000
ns	0.000	0.000	0.000	0.000	NA	0.000	0.000	0.000	0.000
pub	0.000	0.000	0.000	0.000	0.000	NA	0.000	0.000	0.000
pcp	0.000	0.000	0.000	0.000	0.000	0.000	NA	0.000	0.000
national_rank	0.000	0.000	0.000	0.032	0.000	0.000	0.000	NA	0.000
world_rank	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	NA

(Table 4.7: Probability Value table)

Table 4.7 indicates that all the probability values are considered to be statistically significant as all of them are smaller than 0.05.

All the Pearson correlation coefficients, histogram of each variable, bivariate scatter plots and significant level can be demonstrated in one plot based on a package called **PerformanceAnalytics** in R. In spite of using multiple tables to show all the result, **chart.correlation()** function in the aforementioned package supplies a more holistic and intuitive review of the correlation part in the experiment. In order to meet the basic implementing condition of this function, the input data should merely contain numeric data. As a result, the variable "university_name" is temporally removed so that all the variables left are numeric which contributes to the form of the final visualization of them. The results are shown in **Figure 4.1**.



(Figure 4.1: Visualization of Pearson correlation results)

4.3.2 Linear regression

In this case, multiple linear regression offers a scenario to see to what extent that other variables contribute to world rank. Multiple linear regression models can be carried out with the help of a package in R called **caret.** Using the embedded function in **caret** called **train()** is able to create models by executing multiple well-known Machine Learning algorithms. The configuration of **train()** in this case can be represented by its 2 important arguments: **method** and **traincontrol**. To satisfy the precondition of **train()** function, i.e. a variable needs to be regarded as target variable so as to proceed to this regression model. Therefore, as the variable "world_rank" is the target variable in the experiment, the representation of it is a special formula in **train**(): **`world_rank ~ total_score+ alumni+ pcp+ pub+hici+ns+award+national_rank`**. The argument **method** was configured as "glm" which indicates that a multiple linear regression will be implemented. Furthermore, to improve the accuracy of this model, the **traincontrol** argument, which can efficiently supply comprehensive validation method to different models, was deployed to validate the model by utilizing repeated cross validation. The evaluation and this model and the fitness of the variables in this model are shown in **Table 4.8** and **Table 4.9**.

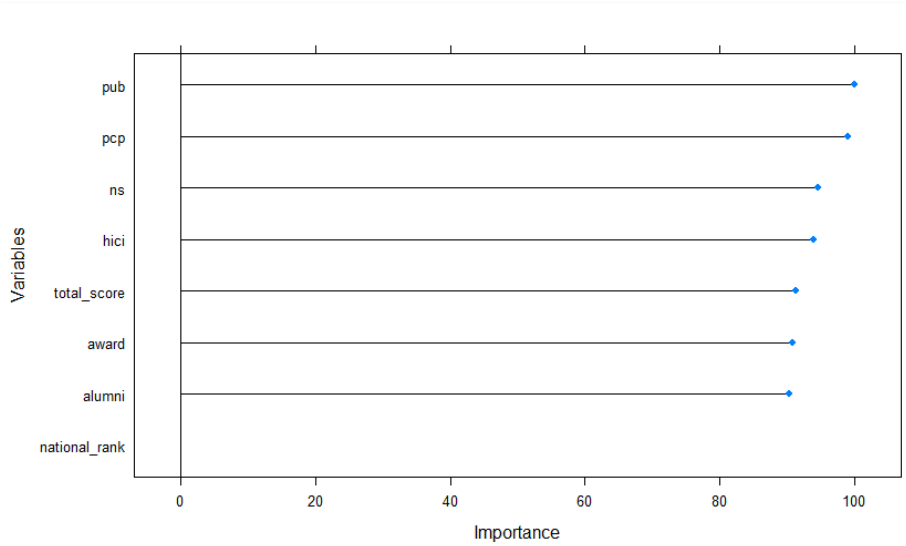
RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
94.99819	0.865406	75.41028	43.8886	0.090754	24.96944

(Table 4.8: Evaluation of regression model)

	Std. Error	t value	Pr(>|t|)
total_score	32.068	13.009	0.000
alumni	3.384	-12.887	0.000
pcp	3.298	-13.949	0.000
pub	6.585	-14.059	0.000
hici	6.709	-13.325	0.000
ns	6.612	-13.404	0.000
award	6.721	-12.944	0.000
national_rank	0.093	1.906	0.057

(Table 4.9: Fitness of the variables)

To test the importance of each variable, **caret** provides a function in terms of **varimp()** which can extract the internal information from models created by **train()** function. As a consequence, the **varimp()** function was used to analysis the multiple linear regression model merely after it was created. The final step of executing the model is to visualize it for the purpose of comprehensive understanding (see **Figure 4.2)**. The system built-in function namely **plot()** is capable of generating visualization from the result of **varimp().**



(Figure 4.2: Variables' importance)

To summarize, all the results from Pearson correlation coefficient table and the variable importance of regression model indicate that the variable "national_rank" in the aggregated ARWU data has poor correlation with other variables and make an extremely

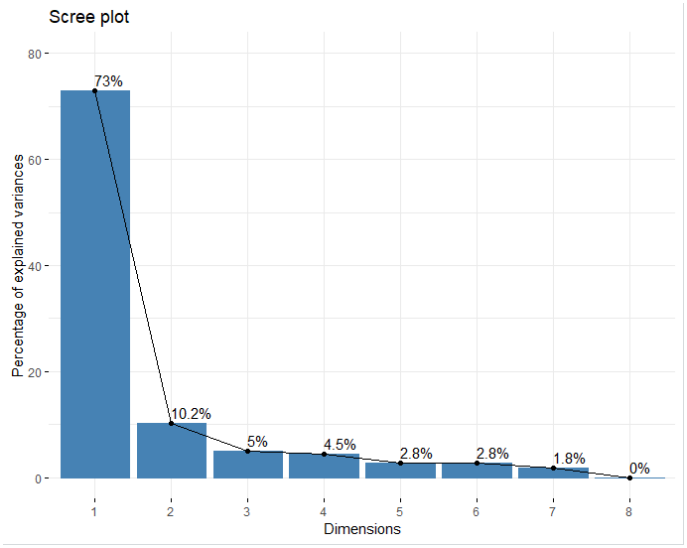
low contribution to the original ranks of universities made by ARWU. As a result, variable national rank will be abandoned for the rest of this experiment.

4.3.3 Principal component analysis (PCA)

Considering that PCA is one of the most widely used method to reduce dimensionality of data, there are many packages that enable users to conduct a PCA analysis. Package **ade4** and **factoextra** are used to convey PCA, the advantage of this choice is that the combination of these 2 packages contribute to specific and versatile visualization of the result of PCA. The function called **dudi.pca()** in package **ade4** is used to generate PCA so as to reduce the dimensionality of the aggregated ARWU data. In order to supply an intuitive way to understand the result from **dudi.pca()**, this part of experiment is designed to be demonstrated by data visualization with the help of package **factoextra.** The following content will specifically introduce all the functions used in experiment to visualize PCA result from **dudi.pca()**:

a) fviz_screeplot is used to draw a scree plot which can give an initial view the cumulative explained variance given by the components generated after conducting the PCA to aggregated ARWU data. It is can also offer a quick view which is helpful for making a basic decision on how many components should be considered in the model. By setting ncp = 8 which means the numbers of components needed to be analyzed in this experiment is 8. All the bars in the chart are connected by line segments which carries out a brief view of the tendency of cumulative variance explained by components.

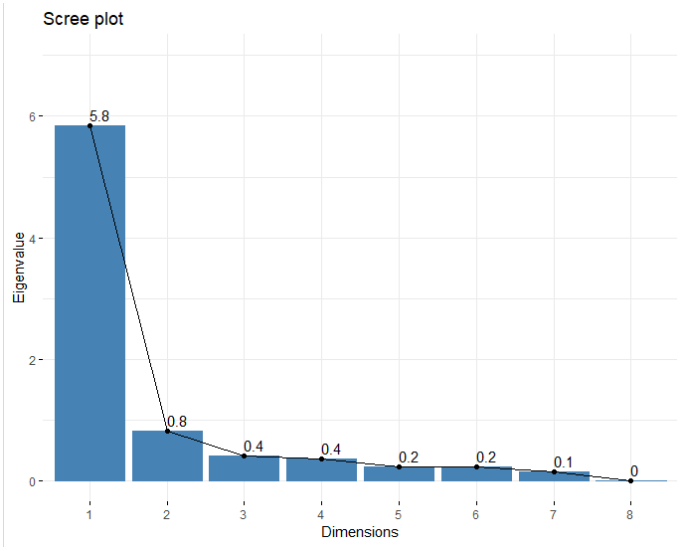
   Figure 4.3 shows the scree plot derived from the result of PCA of the aggregated ARWU data. As it is shown **Figure 4.3**, the scree plot intuitively demonstrates that the first 2 components explained the most variance in relation to the original variance, because there is not a considerable change of variance explained after the second component occurs.



(Figure 4.3: Result from scree plot of explained variance)

b) Another important issue of determining the proper number of components when implementing a PCA is analyze the eigenvalues of components. Eigenvalues are the values of projection index in each direction of project, they mainly reflect to what extent that each component explains the original information of data (Bezdek and Hathaway, 2002). Therefore, eigenvalues are one of the important indexes contributing to specifying the optimal number of principle components

Although there is not a unique standard which is able to specify the correct cut-off number of components, some scientific metric still support that eigenvalues make important contribution to the final determination. fviz_eig() can conveniently visualize eigenvalues by bar chart, simply setup ylim(0,100) can make the limitation of y axis of the bar chart from 0 to 100, which leads to a better perspective to analysis and view the result. **Table 4.10** shows the eigenvalues and cumulative variance explained after applying PCA to the aggregated ARWU data. As shown in **Table 4.10**, the percentage of cumulative variance represented by Component 1 and 2 is over 80% and both of their eigenvalues are higher than the 70% of average eigenvalue 0.7 (average eigenvalue is 1) while other components failed to meet the requirement. It is generated by the function get_enginvalue(), these three types of values are important elements in the process of choosing components.



(Figure 4.4: The spread of eigenvalues in each component)

Figure 4.4 demonstrates the scree plot of PCA which indicates the condition of how much variance is explained by the result of PCA compares to the original data. The other scree plot conveys the details of the account of eigenvalue in each principle component.

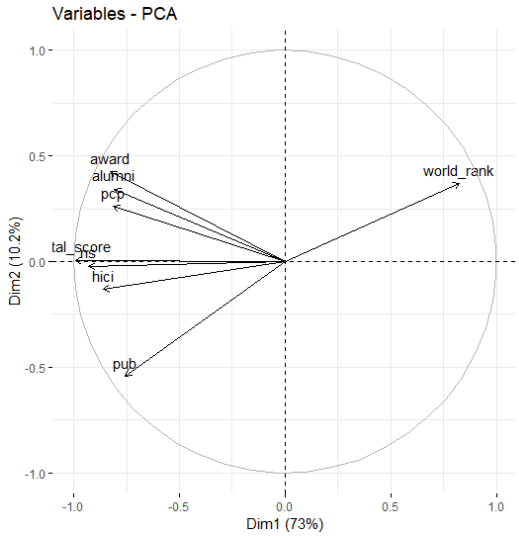
Eigenvalue	Variance percent	Cumulative variance percent
5.841	73.014	73.014
0.815	10.194	83.208
0.399	4.983	88.191
0.360	4.498	92.689
0.221	2.763	95.452
0.220	2.752	98.204
0.144	1.795	99.999
0.000	0.001	100.000

(Table 4.10: Relationship between eigenvalue and variance)

Table specifically shows the information in relation to the accurate eigenvalue of each principle component, the percentage of variance explained by each principle component and the percentage of cumulative variance explained by principle components.

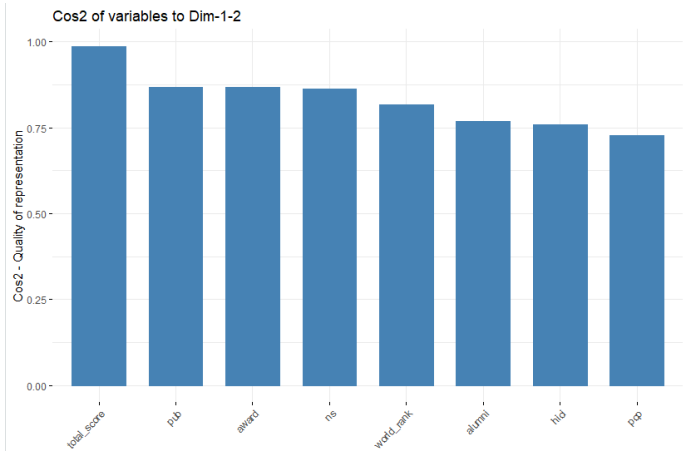
c) During the process of analyzing the result of PCA, how the variables affect the principal component is important. This effect can be derived from the correlation between principle components and variables, as the correlation is used as the coordinates of the variable on principle components. Using fviz_pca_var()

function can visualize it as a correlation circle which is also known as variable correlation plots. The interpretation of this correlation circle is as following principles: variables grouped together are positively correlated; variables projected on opposites are negatively correlated; the length of each arrow i.e. the distance between the center of the circle and each variable indicates the quality of the variables on the factor map, therefore, the longer an arrow means a better representation of a variable. The correlation circle is shown in **Figure 4.5**.



(Figure 4.5: correlation circle plot)

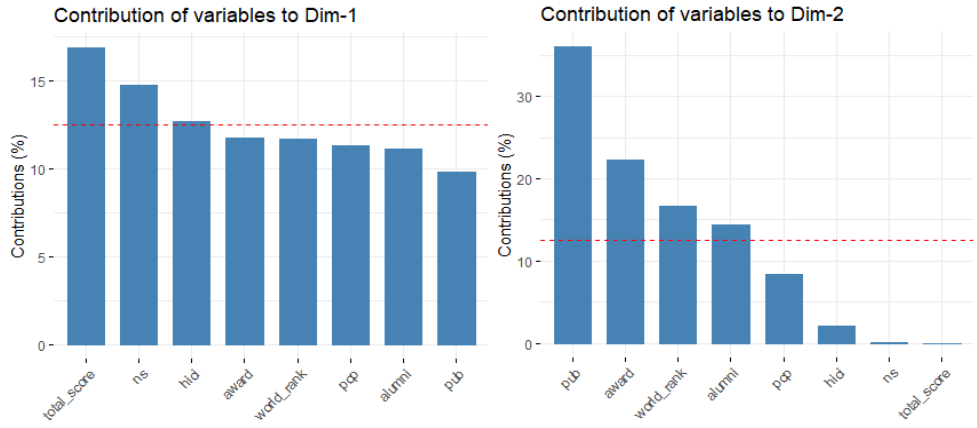
d) The quality of the representation of the variables can also be obtained from the variation plots by calculating the square cosine (cos^2) of each variable (arrow in the circle). The result of it can be interpreted based on the highest squared cosine value represent the best representation of the variable on the principal component i.e. the variable in the correlation variable plot is positioned close to the circumference of the correlation circle. To visualize it, function **fviz_cos2** can plot this table to a bar chart with ease so as to make a comparison between them. Argument placed in this function are **choice** and **axes**, which are set up to "var" and 1:2 in this experiment, respectively. As a result, the condition of representation of variables between principle component 1 and 2 will be demonstrated to analysis. The visualization of cos^2 of each variable in the correlation plot has been shown in **Figure 4.6.**



(Figure 4.6: Cos^2 of each variable)

e) For the purpose of investigating the contributing relationship between each variable and components. **fviz_contrib()** function can plot this type of relationship into bar charts. The **axis** argument can used to choose the correct identification of a component so that those relationships can be seen. For instance, if the **axes** argument is set up to 1, how the variables make contributions to the first component will be shown. Commonly, the first (i.e., principle component 1) and the second (i.e., principle component 2) component are able to explain the most of variability of the data and variables on these two components are correlated with them. Therefore, the axes argument is set up to 1 and 2 for future analysis. If the variables are not correlated with the rest of principle components i.e., the contribution is extremely low on these components, those components will be considered to be dropped. The cumulative contribution of the variables on the components 1 and 2 has been demonstrated on **Figure 4.7.**

Figure 4.7 shows the variables' contribution on the two components; the red dash line in the figure is the reference line which is formed based on the average contribution and corresponds to the expected value of contribution a variable should achieve.



(Figure 4.7: Contribution of variables in component 1&2)

Table 4.11 presents the contribution of all the variables in both principle components, as aforementioned principle component 1 explained the most variance of the original ARWU data, and the variables in **Table 4.11** make uniform contributions which indicates that these variables are equivalently important for the most of AWRU data.

	PC.1	PC.2
total_score	16.90234	0.004924
alumni	11.14366	14.45162
award	11.72537	22.27979
hici	12.69558	2.090128
ns	14.75016	0.056938
pub	9.804988	36.08712
pcp	11.30911	8.319891
world_rank	11.66879	16.70959

(Table 4.11: Contribution of the variables in principle components)

In conclusion, all the figures and tables in this section shows that all the result from PCA such as variables' contribution, the variance explained and the spread of eigenvalues on principle components indicates that the variables have good fit on those principle components and the first two components are sufficient enough to represent the information of the aggregated ARWU data. As a result, variable *total_score, pub, award* and *ns* are the most representative features according to PCA. The result of implementing PCA is to use the features included in component one and two to build the clustering models.

4.4 Implementation of multiple clustering method

The goal of this section is to generated two types of clusters (hierarchical and K-means) based on the result derived from PCA in section 4.3. The process of this section can be divided in to 4 steps as follow:

a) Accessing the clustering tendency.

b) Determining the optimal number of clusters

c) Generating two types of clustering results based on step b

d) Validating these clustering results derived from step c

e) Choosing the better clustering algorithms.

4.4.1 Accessing the clustering tendency

Accessing the clustering tendency is a crucial part before implementing, which can decide whether a data has the possibility of being clustered. Considering the fact that clustering algorithms impose a clustering process to a dataset even the dataset is uniformly distributed which means that there is no cluster can be presented in the dataset. In order to fix this problem, an assessment of clustering tendency for the PCA result is required so as to determine whether the data can be meaningfully clustered. There are two main ways to access the clustering tendency of a dataset: visual clustering tendency (VAT) method and statistical method (Hopkins statistic).

VAT is a widely used as it can measure the dissimilarity of a data by presenting it pair wisely as a square digital image (Bezdek and Hathaway, 2002). Assuming that there is a set of observations O**,** which $O= \{O_1,\ldots,O_n\}$ and the pair wise dissimilarity can be computed as $\mathbf{R= [R_{XY}]}$, where R_{XY} is commonly a distance indicating the pair wise dissimilarity between observation o_x and o_y, for $1 \leq x, y \leq n$.

Hopkins statistics is used to measure clustering tendency which belongs to sparse sampling test and also performs as a statistical hypothesis test. By conducting Hopkins statistic to a dataset, the returned value H is helpful for checking whether the dataset is uniformly distributed. The process of forming the Hopkins statistic as follows:

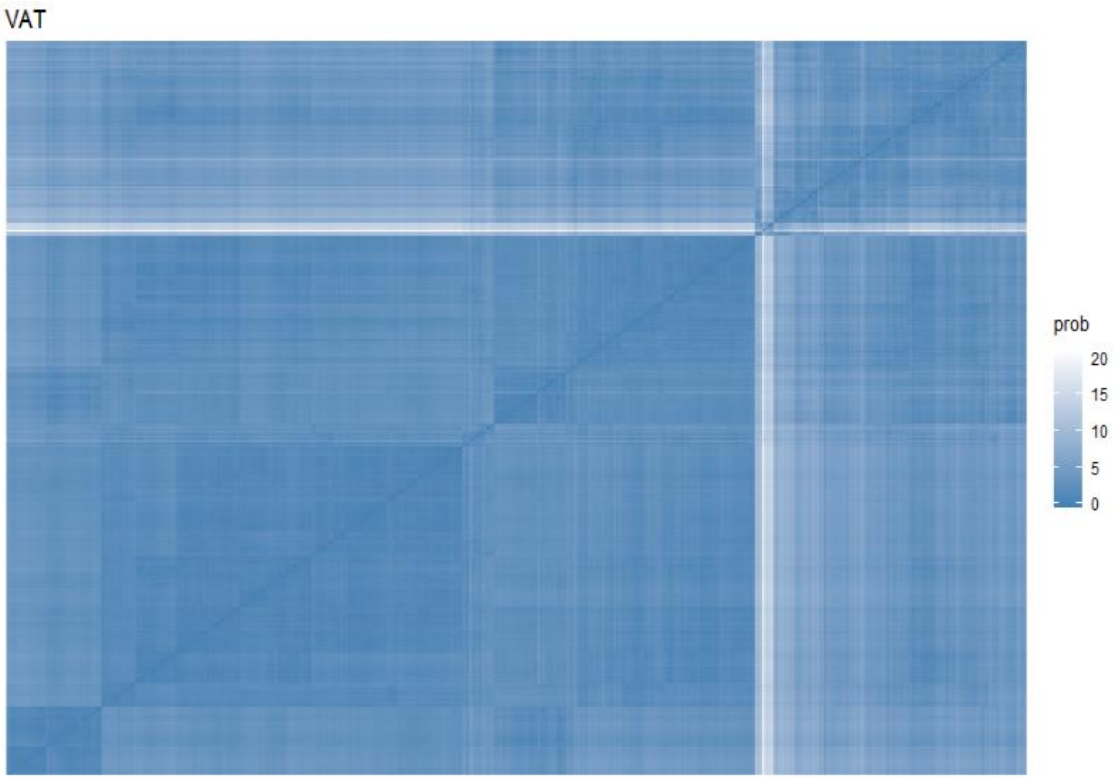
a) Randomly selecting n points from dimension D, for each p in p_i ($1 \leq i \leq n$), find the minimum distance between p_i and v ($v \in D$) and set x_i as the minimum distance where x_i= min(dist (p_i, v)).

b) Randomly selecting n points from dimension D, for each q in q_i ($1 \leq i \leq n$), find the minimum distance between q_i and v ($v \in D$) and set y_i as the minimum distance where y_i= min(dist (q_i, v)).

c) Then the formulation of the Hopkins statistic is as **Equation 4.1**:

$$H = \frac{\sum_{i=1}^{n} yi}{\sum_{i=1}^{n} xi + \sum_{i=1}^{n} yi}$$ Equation 4.1

When the H is over than .5, it indicates that the clusters can be generated in the dataset which is not uniformly distributed and has statistical significance.



(Figure 4.8: VAT of PCA data)

Function **get_clus_tendency()** in package **factoextra** can implement Hopkins statistics and VAT simultaneously and storage the result into a list. In this case, the Hopkins

statistic (H) is .92 which is higher than .5 so that the result of PCA can be used to generate clusters and the result of VAT is shown in **Figure 4.8**.

The interpretation from **Figure 4.8** is that the latent clusters are presented in blocks with consecutive order and as heavier a block is, the higher similarity is in this cluster. The white lines in **Figure 4.8** indicates the data can be well separated into clusters. Therefore, if some distinctive blocks can be found in a VAT plot, there is usually a clustering tendency.

4.4.2 Choosing the optimal number of clusters

The determination of the optimal number of clusters is a necessary part when conducting a clustering experiment. However, there is no evident definition of how to choose a correct number of clusters.

In this case, rather than subjectively setting up a cut-off number of selecting clusters, function Nbclust() is used to determine the optimal number of clusters. The implementation of Nbclust() is to run loops for the number of hierarchical clusters based on configurating the distance metric, linkage method and internal validation methods on each turn. The best result of implementing internal validation methods on each number of clusters contributes to the optimal number.

In order to decide the method and metric, the cophenetic correlation coefficient will be used to justify the result. The cophenetic correlation coefficient is the measurement of the degree of fitness of a dendrogram preserving the pairwise distance (Gopal and Shitan, 2015), a higher cophenetic correlation coefficient indicates a better linkage result while forming hierarchical clusters. Therefore, the cophenetic correlation coefficient can be used to determine the distance metric and linkage method implemented in this clustering experiment.

The formulation of cophenetic correlation coefficient c is as follow:

1. Assuming there is a set of observations **X** where $X\{X_1,\ldots X_n\}$
2. Assuming that there is a dendrogram **T** which is generated based on **X,** where $T= T\{T_1,\ldots T_n\}$
3. Set x(i, j) equals to the distance between X_i and X_n

4. Set t(i, j) equal s to the distance between points T_i and T_j. This distance is the height of the node where T_i and T_j initially joint.

5. Then, the cophenetic correlation coefficient ***c*** can be calculated by **Equation 4.2**

$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}}.$$

Equation 4.2

Where $\bar{x}$ is the average of x(i, j) and $\bar{t}$ is the average of t(i, j)

The result of possible combinations of distance metric and linkage method and their cophenetic correlation coefficient is shown in **Table 4.12.**

	average	single	complete	ward
Euclidean	0.991780	0.979975	0.992983	0.997998
Manhattan	0.987442	0.976663	0.992456	0.997834
Canberra	0.938702	0.906312	0.944833	0.995333
Minkowski	0.991780	0.979975	0.992983	0.997998
Maximum	0.988856	0.981973	0.992867	0.998101

(Table 4.12: Linkage cophenetic correlation coefficient)

As shown in **Table 4.12**, the highest value of cophenetic correlation coefficient between a distance metric and a linkage method is .998 with the combination of Maximum distance metric and Ward method. As a result, these 2 elements are set up into function **Nbclust()**, and the final implementation of it is as follows:

```
NbClust(finalpca2, distance = "maximum", min.nc = 2, max.nc
= 10, method = "ward.D2",index = 'dunn')
```

In this case, the **index** argument is identified as “dunn” which means the validation of this clustering experiment is using Dunn index and the result of optimal number returns 8. The **min.nc** and **max.nc** determines the range of a number of the clusters required to be generated and are set up to 2 and 10. As a result, the **Nbclust()** proposes that the optimal clustering number of the hierarchical clustering model is 8 and the K-means is 2. However, the result from **Nbclust()** is not 100 percent trustful, so a specific internal validation of this clustering implementation is needed and presented in section 4.4.2.

4.4.3 Generating clusters and internal validation

Both of the two clustering techniques hierarchical and K-means clustering technique will be implemented in this experiment. It is because these two clustering methods are the most popular and it is interesting to making comparison between them so as to see the best result. This way, the final result will be more representative and reliable.

Package **cluster** and **factoextra** are used in this experiment, which are used as generate multiple clustering results and visualize the results of these two clustering methods, respectively. Function **eclust()** derived from package **cluster** is used to create clusters. The configuration of **eclust()** depends on the distance metric and linkage method which are important to create clusters, the details of this function are shown as below:

```
eclust %>%

(data=finalpca,FUNcluster = c("hclust",”kmeans”) ,hc_metric
= "maximum", hc_method = "ward")
```

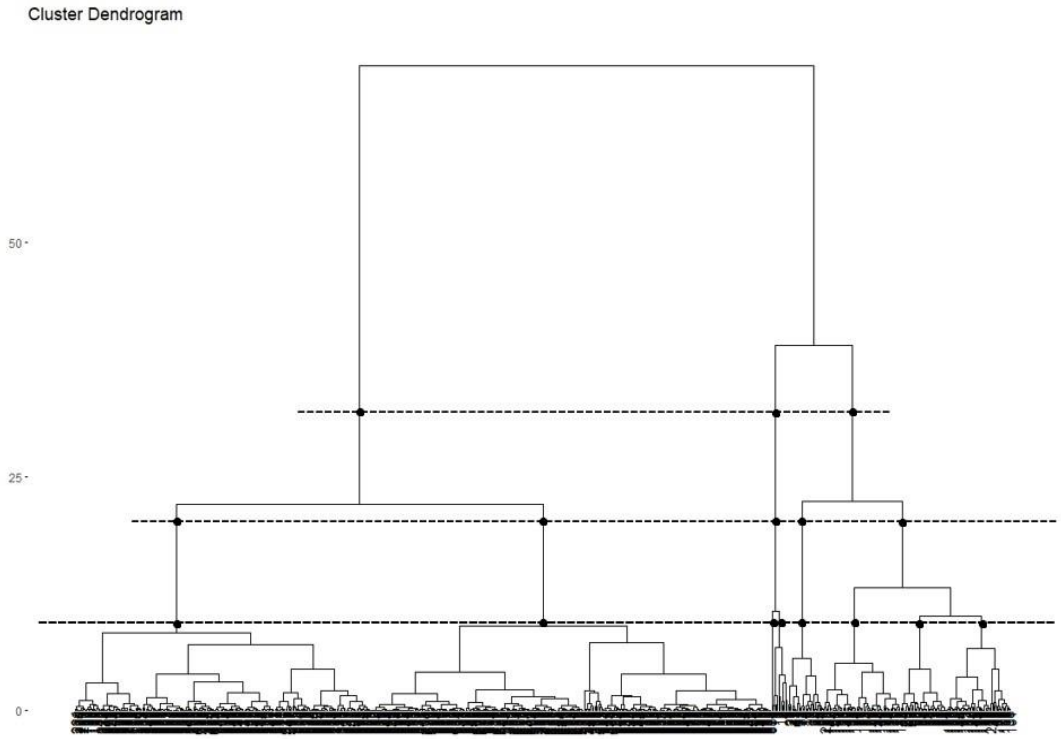
The content of argument **FUNcluster** is nominated by the name of clustering methods that need to be implemented. In this case, it is set up to “hclust” and “kmeans” in each implementation, which indicates that hierarchical clustering and K-means clustering methods will be conducted in the experiment. Argument **hc_metric** enables users to define the distance metric to hierarchical clustering methods, “maximum” is used as the distance metric for the agglomerative and divisive clustering method. When conducting a hierarchical clustering method, the linkage method needs to be defined. Therefore, the argument **hc_method** is designed for the configuration of linkage method and is assigned to Ward method in the experiment. The reason for the choices of the distance metric and the linkage method was explained in section 4.4.3.

Figure 4.9 demonstrates the original dendrogram with all the possible clustering numbers pointed. The possible set of numbers are 3,5 and 8 as they represent those branches which are widely separated from other ones in the same horizon line.

The result of hierarchical cluster and K-means cluster are illustrated by package **factoextra** as dendrogram and clustering plot in **Figure 4.10** (page 52) and **Figure 4.11** (page 53), respectively. To have a basic overview of these clustering results, the spread

of hierarchical clustering result in each clustering experiment is shown in **Figure 4.12**, which intuitively demonstrates the amount of universities in each cluster with the number on the top

For the purpose of evaluating these 2 types of clustering models and choosing optimal clustering numbers, an internal validation of these clustering results will be presented in the next part, the Silhouette width and Dunn index are the validation method mentioned in section 3.4.2. The aim of the internal validation is to specifically investigate the quality of these 2 types of clusters.



(Figure 4.9: The set of possible clustering numbers by dendrogram)

3 clusters	5 clusters	8 clusters
0.5946	0.4113	0.4131

(Table 4.13: Internal validation of H cluster by Silhouette width)

3 clusters	5 clusters	8 clusters
0.0143	0.0068	**0.0157**

(Table 4.14: Internal validation of H cluster by Dunn index)

Table 4.13 indicates that all the possible numbers of hierarchical clusters have a positive silhouette. Therefore, all of them can be considered as an optimal choice. As there is no specific range of metric for Dunn index, so 8 clusters will be the optimal number of

cluster because a higher Dunn index indicates a better performance of clusters and 8 clusters have the highest Dunn index showed in **Table 4.14**.

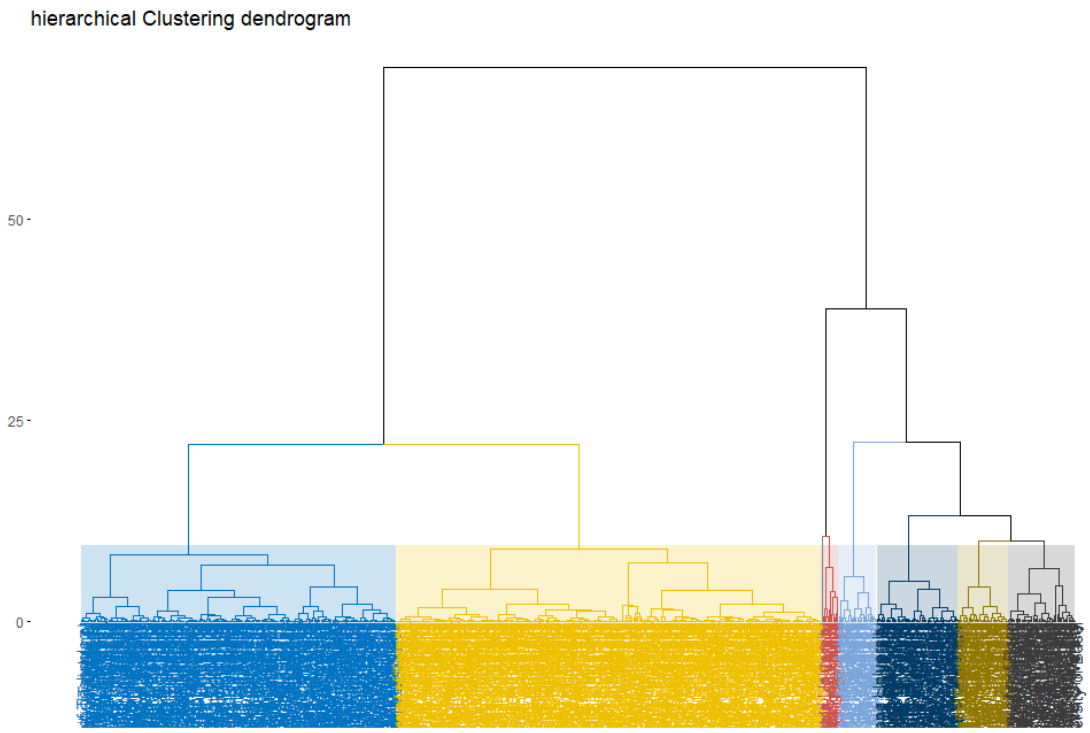
2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	7 clusters	8 clusters	9 clusters	10 clusters
0.6444	0.5421	0.4826	0.4347	0.4038	0.3884	0.3827	0.3792	0.3285

(Table 4.15: Internal validation of K-means by Silhouette width)

2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	7 clusters	8 clusters	9 clusters	10 clusters
0.0066	0.0038	0.0034	0.0056	0.0036	0.0041	0.0034	0.0056	0.0045

(Table 4.16: Internal validation of K-means by Dunn index)

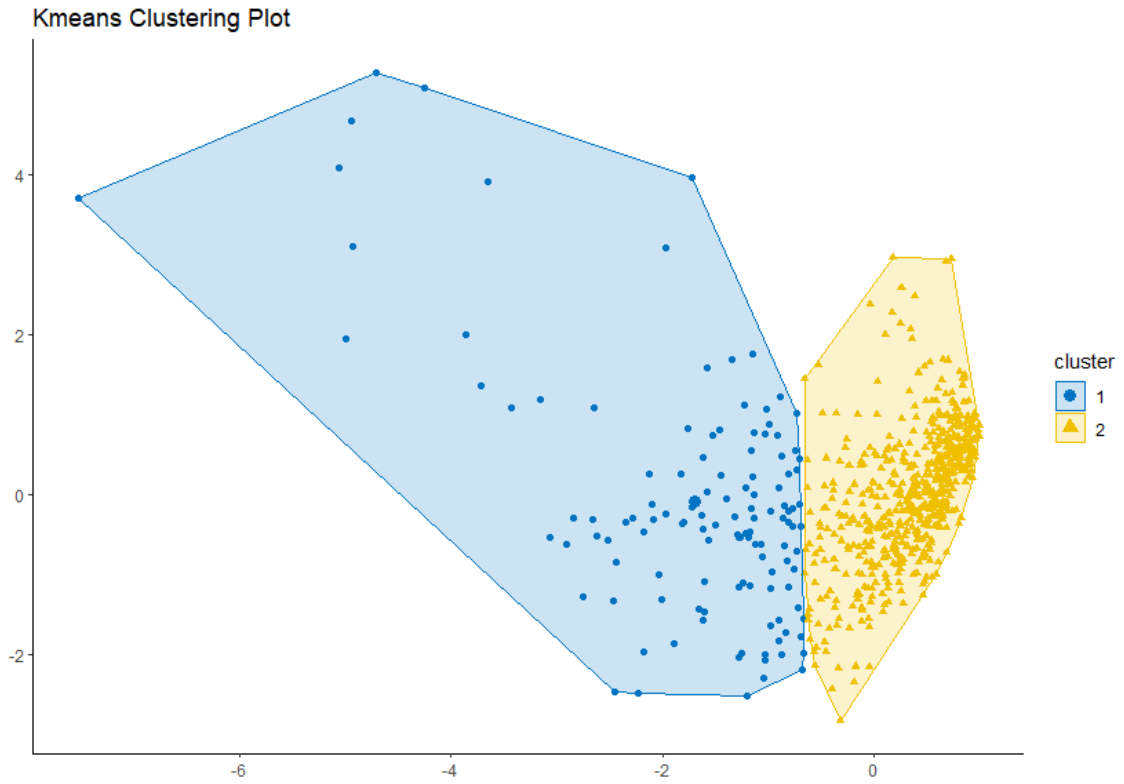
Due to the reason that K-means clustering technique cannot generate dendrogram so that the optimal number of clusters can only be subjectively decided. Therefore, conducting internal validation for K-means is beneficial for finding the optimal number of K-means clustering. **Table 4.15** and **Table 4.16** offer the validating results from multiple possible number of K-means cluster which leads to that 2 clusters have better performance than other ones, according to all the Silhouette width is positive and 2 clusters have the highest Dunn index 0.0066.



(Figure 4.10: Dendrogram of hierarchical cluster)

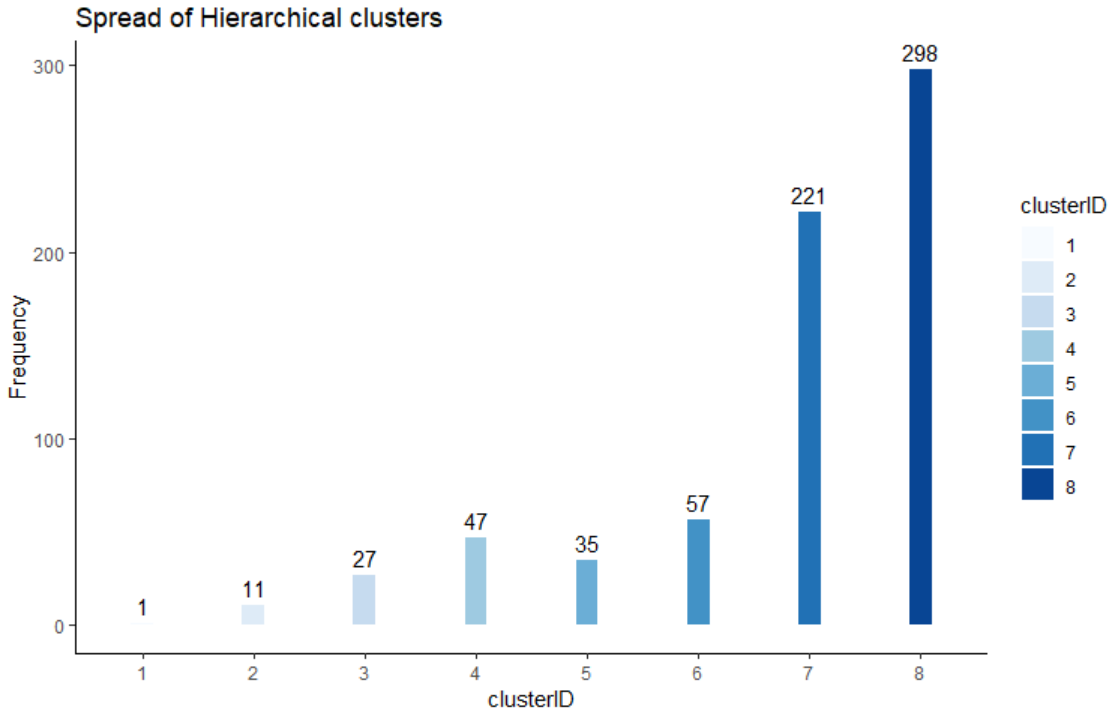
As is shown in **Figure 4.10**, the nodes from dendrogram are cut off at 7 indicating that there are 7 clusters generated. All the clusters in this dendrogram are distinguished by different colours and blocks. The end of nodes are the names of universities but considering that there are over 500 universities in this experiment so that it is hard to demonstrate them in one plot because of overlapping. However, the source of this dendrogram is stored in the result of **eclust()** as a list. Therefore, the detail of these clusters can still be extracted.

Figure 4.11 demonstrate the result of K-means clustering, there are 2 clusters generated by K-means which indicates that the universities can optimally be classified into 2 groups. However, merely 2 clusters derived from K-means is not satisfactory with the requirement of this case. It is because there are over 500 universities required to be clustered as different groups (clusters), 2 groups are too simple to be interpreted. For example, if these 2 clusters are labelled as "good university" and "bad university", it will make a low contribution for accessing the real educational level of a school as there are only two types of universities – good and bad.



(Figure 4.11: K-means clustering plot)

The investigation of **Figure 4.12** presents the number of observations in each cluster and both cluster 7 and cluster 8 contains over 200 observations i.e. 200 universities. Considering the goal of this dissertation is to find better ranking criteria of universities, a cluster consisting of over 200 universities is not considered an acceptable criterion in this case as it is too hard to be interpreted specifically and cannot reflect more accurate condition of universities. Therefore, a decision has been made that cluster 7 and cluster 8 need to be implemented into hierarchical clustering method again so as to generate sub-clusters for them which can lead to more specific criteria.



(Figure 4.12: The spread of hierarchical clusters)

4.4.4 Sub-clusters

The steps of creating sub-clusters are same as the steps listed in section 4.4.2: the measurement of clustering tendency for the data in cluster 7 and cluster 8 is needed so as to see whether the data has latent clusters; then, the two dendrograms will be presented so that the possible set of clustering numbers can be identified; next, the internal validation of the possible number of sub-clusters will be conducted based on Silhouette width and Dunn index, which can determine the optimal number of sub-clusters; lastly, the generation of sub-clusters for cluster 7 and 8 will be demonstrated.

a) The clustering tendency of cluster 7 and cluster 8

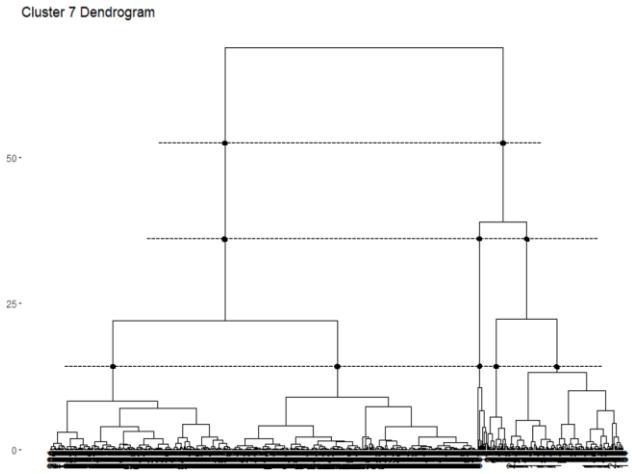
Table 4.17 demonstrated the Hopkins statistics (H) of cluster 7 and cluster 8, both of the H of them are greater than .5 with H = 0.71 and H = 0.79, respectively. Therefore, the data of these 2 clusters has latent clustering patterns.

Hopkins statistics of cluster 7	Hopkins statistics of cluster 8
0.71	0.79

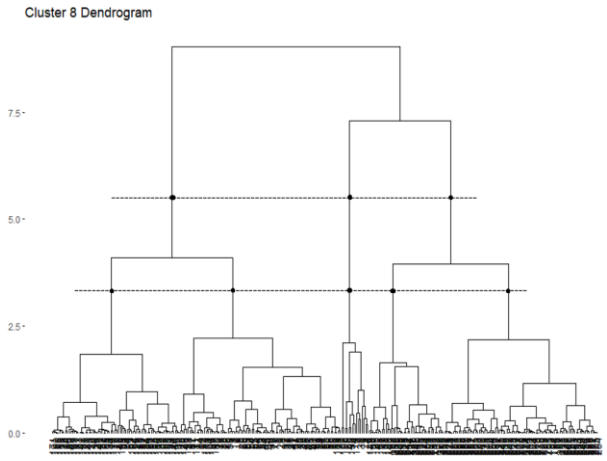
(Table 4.17: Hopkins statistics of cluster 7 and 8)

b) Dendrograms of sub-clusters for cluster 7 and cluster 8

Figure 4.13 and **Figure 4.14** show the dendrograms derived from the data inside cluster 7 and cluster 8.



(Figure 4.13: Dendrogram of sub-clusters of cluster 7)



(Figure 4.14: Dendrogram of sub-clusters in cluster 8)

As are shown, the latent number of sun-clusters in cluster 7 is 2, 3 and 5, in cluster 8 is 3 and 5. Accordingly, the internal validation will be conducted to these 2 sets of possible clustering numbers.

c) Internal validation of sub-clusters

The internal validation using Silhouette width and Dunn index is implemented on them in order to finding the optimal number of their sub-clusters. The range of possible number of sub-clusters is listed above. The result has been shown in **Table 4.18**, **Table 4.19**, **Table 4.20** and **Table 4.21**.

2 sub-clusters	3 sub-clusters	5 sub-clusters
0.3341	0.3213	0.2963

(Table 4.18: Silhouette width of sub-clusters of cluster 7)

2 sub-clusters	3 sub-clusters	5 sub-clusters
0.0405	0.0457	0.0569

(Table 4.19: Dunn index of sub-clusters of cluster 7)

3 sub-clusters	5 sub-clusters
0.4066	0.3502

(Table 4.20: Silhouette width of sub-clusters of cluster 8)

3 sub-clusters	5 sub-clusters
0.0284	0.0284

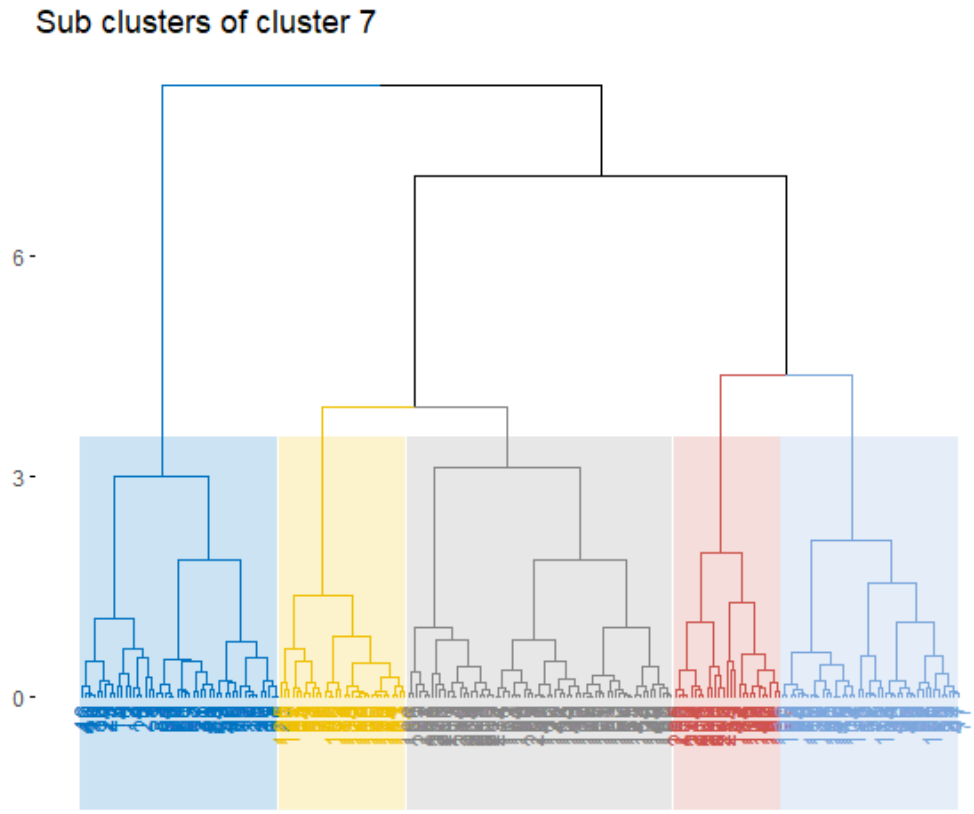
(Table 4.21: Dunn index of sub-clusters of cluster 8)

Table 4.18 and **Table 4.19** indicates that the optimal number of sub-clusters in cluster 7 is 5 with the greatest Dunn index .0569 and acceptable Silhouette width .2963.

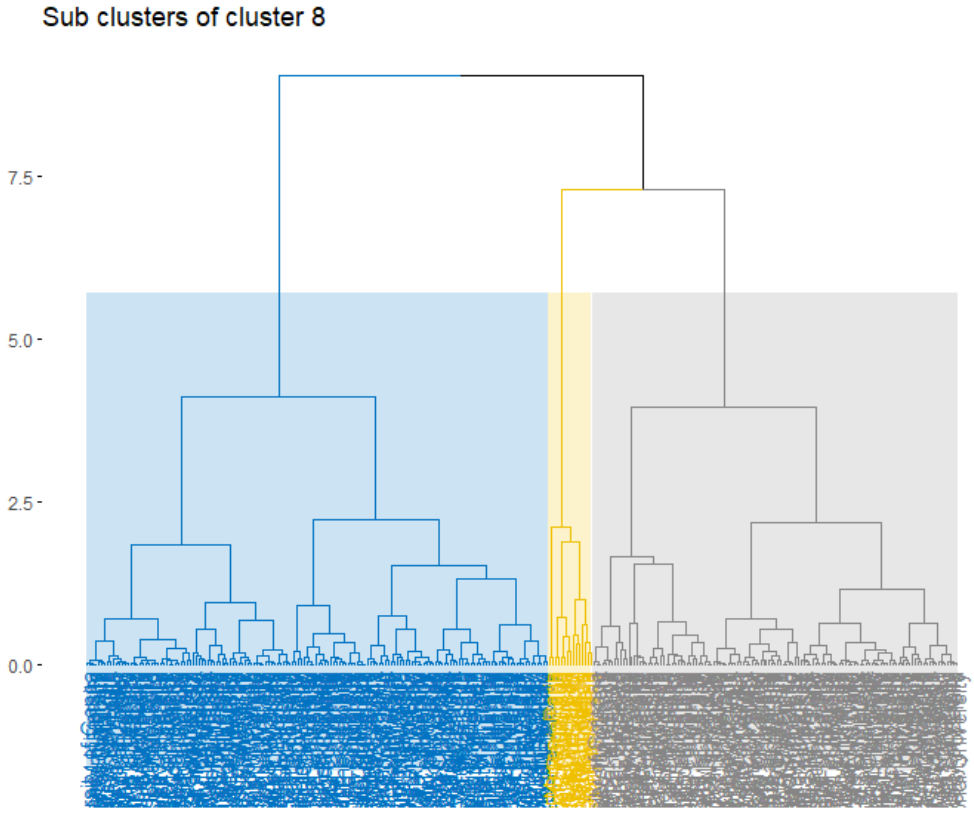
Table 4.20 and **Table 4.21** indicates that the optimal number of sub-clusters in cluster 8 is 3 with the equivalent Dunn index .0284 and greater Silhouette width .4066.

d) Generation of sub-clusters

The sub-clusters of cluster 7 and cluster 8 have been shown in **Figure 4.15** and **Figure 4.16.** The analysis and interpretation of all the clusters and sub-clusters will be presented in Chapter 5.



(Figure 4.15: Sub-clusters of cluster 7)



(Figure 4.16: Sub-clusters of cluster 8)

4.4.5 Result of all the clusters

Due to the fact that clusters generated do not have intrinsic order which indicates that the meaning of clusters have to be defined by creator of them. In order to avoid thinking that the clusters made in Chapter 4 have numeric order, the clusters are renamed in alphabetic order from A to H representing the cluster 1 to 8 and the sub-clusters for the cluster 7 and 8 are renamed from GA to GE and HA to HC, respectively.

Part of final result including all the clusters and sub clusters assigned to universities is shown below:

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
USA	Harvard University	100	100	100	100	100	100	74.76	1	1	A

(Table 4.22: University of cluster A)

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
USA	Stanford University	73.55	41.16	83.05	84.67	71.17	71.41	55.13	2	2	B
USA	University of California Berkeley	70.62	66.78	78.88	66.98	68.8	68.31	55.45	3	3	B
USA	Massachusetts Institute of Technology (MIT)	70.53	69.81	81.45	62.55	70.76	61.77	64.68	4	4	B

(Table 4.23: Universities of cluster B)

country	university_name	total_score	Alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
USA	University of California Los Angeles	52.07	28	45.06	57.02	48.83	72.7	32.48	10	12	C
USA	University of California San Diego	49.37	18.04	35.51	57.82	54.56	64.59	38.34	12	14	C
USA	University of Washington	48.55	23.06	32.81	53.78	51.14	72.83	29.09	13	15	C

(Table 4.24: Universities of cluster C)

country	university_name	total_score	Alumni	award	hici	ns	pub	pcp	national_rank	world_rank	cluster ID
USA	Rockefeller University	36.97	18.13	59.6	29.59	42.2	20.47	38.49	27	36	D
France	University of Paris Sud (Paris 11)	33.77	31.37	52.09	15.65	19.27	48.63	26.72	4	48	D
USA	Carnegie Mellon University	29.26	32.31	32.05	28.47	14.84	34.65	32.8	41	66	D

(Table 4.25: Universities of cluster D)

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
USA	University of Colorado at Boulder	36.34	13.37	32.58	36.51	39.05	45.61	33.25	28	37	E
France	Sorbonne University	36.1	33.3	27.1	25.3	29.9	64.7	26.2	1	38	E
France	Pierre and Marie Curie University Paris 6	35.2	34.05	27.4	25.8	30	59.6	23.85	2	40	E

(Table 4.26: Universities of cluster E)

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
USA	University of Pittsburgh	30.58	20.95	0	41.9	23.37	61.88	21.68	38	62	F
USA	Pennsylvania State University University Park	30.27	11.04	0	41.85	33.54	55.02	23.23	39	63	F
USA	University of California Davis	30.1	0	0	41.08	32.47	60.12	26.12	40	64	F

(Table 4.27: Universities of cluster F)

country	university_name	total_score	Alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
USA	Mount Sinai School of Medicine	17.05	0	0	19.82	20.57	35.47	18.92	87	222	GA
Japan	Kyushu University	16.96	0	0	13.55	15.1	46.8	18.8	8	223	GA
Netherlands	Delft University of Technology	16.93	12.28	0	9.45	19.39	39.16	20.97	10	224	GA

(Table 4.28: Universities of sub-cluster GA)

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
Italy	University of Roma La Sapienza	20.81	13.21	14.6	9.5	14.29	51.39	15.29	1	148	GB
Germany	University of Erlangen Nuremberg	16.5	0	0	17.15	14.53	40.22	21.24	18	237	GB
UK	Durham University	16.4	0	0	20.46	15.46	35.1	22.14	22	240	GB

(Table 4.29: Universities of sub-cluster GB)

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
France	University Grenoble Alpes	19.1	0	14.7	0	24.05	46.65	20.4	10	178	GC
Austria	University of Vienna	18.45	15.12	0	15.03	20.71	37.22	23.33	1	190	GC
Italy	University of Milan	17.99	16.62	0	13.79	11.32	46.06	20.92	4	196	GC

(Table 4.30: Universities of sub-cluster GC)

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
Austria	University of Innsbruck	14.94	0	9.96	11.56	16.95	24.95	22.54	3	282	GD
Italy	University of Florence	14.87	0	0	14.92	11.4	39.16	17.68	8	286	GD
USA	University of South Florida	14.64	0	0	15.34	9.92	39.35	17.17	109	293	GD

(Table 4.31: Universities of sub-cluster GD)

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
UK	Cardiff University	20.53	0	16.3	14.65	18.81	41.72	21.99	15	153	GE
USA	Oregon State University	19.86	11.36	0	23.82	22.34	35.62	23.66	73	168	GE
Japan	Hokkaido University	19.65	8.63	11.86	11.69	15.45	45.33	19.33	7	171	GE

(Table 4.32: Universities of sub-cluster GE)

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
USA	State University of New York at Albany	12	0	0	17.18	6.2	27.73	17.67	132	415	HA
China	Capital Medical University	11.55	0	0	0	1.77	39.92	31.98	32	448	HA
China	Wuhan University of Technology	11.47	0	0	19.37	1.57	30.8	11.27	34	460	HA

(Table 4.33: Universities of sub-cluster HA)

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
France	Toulouse School of Economics	13.25	0	29.8	9.45	7.7	10.05	18.2	20	349	HB
France	Paris Dauphine University (Paris 9)	13.03	20.4	26.45	0.9	0	13.9	27.05	22	360	HB
France	ESPCI ParisTech	13.02	7.28	18.72	0	11.76	15.34	31.1	23	362	HB

(Table 4.34: Universities of sub-cluster HB)

country	university_name	total_score	alumni	award	hici	ns	pub	pcp	national_rank	world_rank	clusterID
France	University of Toulouse 1	15.35	0	29.4	4.8	15.4	11.35	31.5	14	269	HC
USA	Medical University of South Carolina	12.29	0	0	13.1	5.77	30.46	24.28	128	401	HC
China-Taiwan	China Medical University	12.15	0	0	15.88	1.8	31.05	23.95	6	407	HC

(Table 4.35: Universities of sub-cluster HC)

It is worth to notice that the universities in the same cluster or sub-cluster presented in above tables are not ranked, which means they are equivalent in the cluster and sorted based on alphabetic order.

4.4.6 Discussion & analysis

This section focuses on discussing and analyzing the result of hierarchical clusters generated by ARWU data. Multiple boxplots will be used in this case in order to reflect the performance of different clusters from several aspects where the ARWU data includes.

As mentioned before, although the data gathered from ARWU is objective (Docampo, 2011), the university rankings from ARWU are still doubtful according to its unrepresentativeness of data and subjective ranking method. Therefore, rather than using unrepresentative data to specifically rank universities, the new ranking criteria (clusters) have been created based on hierarchical clustering approach and are shown below by boxplots. The range of a boxplot starts with minimum value of a variable and sequentially contain the other information through first quartile (Q1), median, third quartile (Q3), it will eventually end with the maximum value of a variable while drawing a “box” where the range is from Q1 to Q3 of a variable. Therefore, boxplot is useful to show the distribution of the numerical variables in a data, which allows readers to intuitively make a comparison between variables and see the trend of how variables are performed.

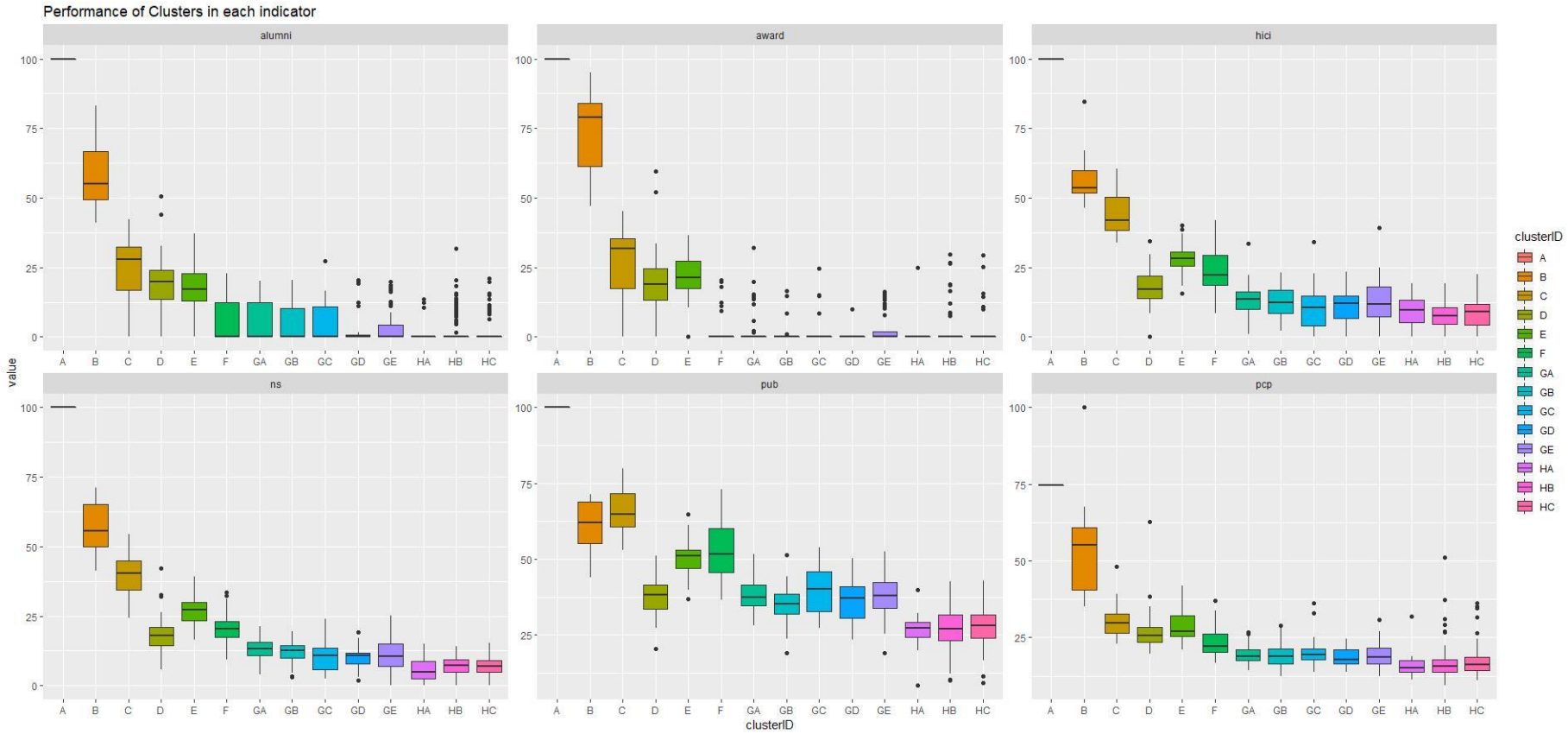
Figure 4.17 in page 62 demonstrates the rank of all the clusters between each variable in the ARWU data, which carries out the result of how the clusters of universities

perform in each variable which can be interpreted as relevant university ranking domain selected by ARWU.
In another view, **Figure 4.18** in page 63 shows the rank of all the variables in each cluster, which is helpful for knowing which university ranking domain is performed as the best in each cluster of university.

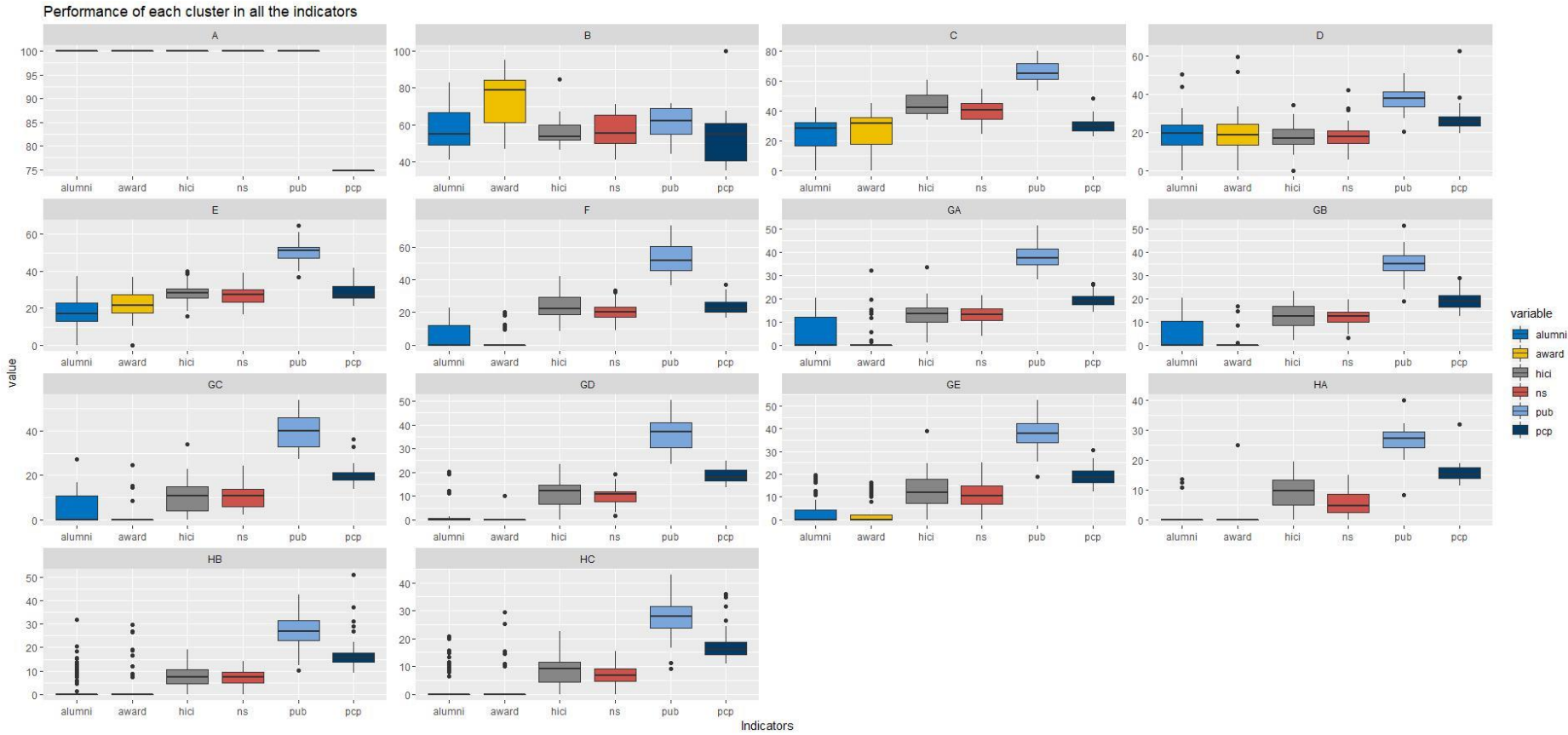
Table 4.36 and **Table 4.37** in page 64 are the numerical result derived from **Figure 4.17** and **Figure 4.18,** respectively. Both of them shows the clear ranks based on the 2 aspects discussed above. **Table 4.38** in page 65 presents all the world ranks of universities in each cluster derived from the aggregated ARWU data.

According to **Table 4.36**, **Table 4.37** and **Table 4.38**, there are different perspectives and result between ARWU rankings and the result of this research. For instance, most of universities in cluster B are ranked higher than the universities in cluster C, but the result of this research shows that universities in cluster C perform better than cluster B in variable *pub* (papers published in Nature & Science). Therefore, assuming that there are some students who want to choose universities which are best at Nature & Science, what ARWU can suggest for these students are only specific numeric rank of each university without leading to a perspective on which universities are better at this area. Similarly, by looking at the ARWU ranking result, these students may intend to think the universities in cluster B are overwhelmingly better than universities in cluster C in any area as these schools are ranked higher in cluster B. In another point of view, the result of this research also supplies more choices of universities which ARWU cannot give. Because universities are clustered into groups with ranked in different majors and domains, it is more convenient for international students to look at the universities from different countries but with similar properties.

In conclusion, this research carried out a different perspective of universities rankings. Comparing with the result of ARWU rankings, this research not only focuses on broadening the perspective of choosing universities but also emphasizing the advantages of similar universities.



(Figure 4.17: Performance of clusters in each indicator)



(Figure 4.18: Performance of each cluster in all the indicators)

	Variable					
clusterID	alumni	award	hici	ns	pub	pcp
A	1	1	1	1	1	1
B	2	2	2	2	3	2
C	3	3	3	3	2	3
D	4	5	6	6	7	5
E	5	4	4	4	5	4
F	6	6	5	5	4	6
GA	6	6	7	7	9	8
GB	6	6	8	8	11	9
GC	6	6	11	9	6	7
GD	6	6	9	10	10	11
GE	6	6	10	11	8	10
HA	6	6	12	14	13	14
HB	6	6	14	12	14	13
HC	6	6	13	13	12	12

(Table 4.36: Clusters' ranking in all the majors)

	ClusterID													
variable	A	B	C	D	E	F	GA	GB	GC	GD	GE	HA	HB	HC
alumni	1	5	6	3	6	5	5	5	5	5	5	5	5	5
award	1	1	4	4	5	5	5	5	5	5	5	5	5	5
hici	1	6	2	6	2	2	3	4	4	3	3	3	3	3
ns	1	3	3	5	3	4	4	3	3	4	4	4	4	4
pub	1	2	1	1	1	1	1	1	1	1	1	1	1	1
pcp	2	4	5	2	4	3	2	2	2	2	2	2	2	2

(Table 4.37: Majors' ranking in each cluster)

clusterID	Original world rank in aggregated ARWU data
A	1
B	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13
C	12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 39, 43, 45, 47
D	36, 48, 66, 72, 78, 79, 85, 86, 87, 88, 93, 97, 100, 101, 102, 104, 106, 107, 108, 109, 112, 113, 115, 119, 120, 121, 122, 125, 134, 135, 136, 140, 144, 157, 158, 162, 163, 164, 165, 166, 173, 174, 179, 182, 188, 192, 218
E	37, 38, 40, 41, 42, 44, 46, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 65, 67, 68, 69, 70, 71, 73, 74, 75, 77, 82, 83, 84, 89, 95
F	62, 63, 64, 76, 80, 81, 90, 91, 92, 94, 96, 98, 99, 103, 105, 110, 111, 114, 116, 117, 118, 123, 124, 126, 127, 128, 129, 130, 131, 132, 133, 137, 138, 139, 141, 142, 143, 145, 146, 147, 149, 150, 151, 152, 154, 155, 156, 159, 160, 161, 167, 169, 170, 177, 180, 186, 201
GA	222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 238, 239, 241, 242, 243, 244, 245, 247, 248, 249, 250, 251, 252, 254, 256, 258, 260, 262, 263, 265, 266, 267, 270, 271, 275, 276, 278, 279, 280, 281, 283, 284, 287, 292
GB	148, 237, 240, 246, 253, 257, 259, 261, 264, 268, 272, 273, 277, 285, 288, 289, 290, 291, 294, 296, 298, 301, 302, 303, 307, 309, 310, 313, 315, 316, 318, 319, 321, 322, 325, 326, 330, 331, 333, 336, 338, 339, 342, 352, 361, 369, 480
GC	178, 190, 196, 197, 200, 205, 206, 211, 214, 216, 219, 220, 255, 274, 297, 305, 314, 324, 340, 356, 359, 363, 367, 371, 372, 376, 411, 418, 426, 441
GD	282, 286, 293, 295, 299, 300, 304, 306, 308, 311, 312, 317, 323, 327, 328, 329, 332, 334, 335, 337, 341, 343, 344, 345, 346, 347, 348, 350, 351, 354, 355
GE	153, 168, 171, 172, 175, 176, 181, 183, 184, 185, 187, 189, 191, 193, 194, 195, 198, 199, 202, 203, 204, 207, 208, 209, 210, 212, 213, 215, 217, 221, 320, 353, 357, 358, 364, 365, 366, 368, 370, 373, 375, 377, 378, 379, 381, 384, 386, 389, 391, 392, 393, 395, 397, 398, 402, 404, 406, 409, 410, 412, 416, 420, 430, 442
HA	415, 448, 460, 496, 597, 598, 599, 600, 601, 610, 614, 625, 628, 636, 661
HB	349, 360, 362, 374, 380, 382, 383, 385, 387, 388, 390, 394, 396, 399, 400, 403, 405, 408, 413, 414, 417, 419, 421, 422, 423, 424, 427, 428, 429, 431, 432, 433, 434, 435, 436, 437, 438, 439, 444, 445, 447, 449, 450, 451, 452, 453, 456, 457, 458, 461, 462, 463, 464, 465, 466, 467, 468, 469, 471, 472, 473, 474, 475, 476, 477, 478, 479, 482, 486, 487, 489, 602, 603, 604, 605, 606, 607, 608, 609, 611, 612, 613, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 626, 627, 629, 630, 631, 632, 633, 634, 635, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 685, 686, 687, 688, 689, 690, 692, 693, 694, 696, 697
HC	269, 401, 407, 425, 440, 443, 446, 454, 455, 459, 470, 481, 483, 484, 485, 488, 490, 491, 492, 493, 494, 495, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 684, 691, 695

(Table 4.38: The original world rank of universities in each cluster)

CHAPTER 5 – CONCLUSION

5.1 Research Overview

The objective of this research is to create a new type of university ranking criteria, namely hierarchical cluster analysis by using ten years ARWU ranking data from 2008 to 2018. This new type ranking criteria proposes a substitutional way for ranking universities comparing with conventional linear ranking method and supplies a ranking result in a different perspective. The resulting dendrogram and the boxplot of the presentation of clustering rank have been presented in section 4.4.6.

5.2 Problem Definition

Due to the reason that current authoritative ranking institutions have fatal drawbacks when ranking universities such as data unrepresentativeness and subjective bias on gathering data, it is reasonable to propose a new type of university ranking criteria by clustering universities into groups based on investigating inside patterns from the ARWU data.

The result of the experiment presented in chapter 4 addresses the research question and rejects the Null hypothesis, because the approach successfully achieving the goal of this research: “the better ranking criteria have been found for universities from the ARWU data based on hierarchical clustering method”.

5.3 Design/Experimentation, Evaluation & Results

In order to comprehensively understand the ARWU data, the investigation of data including exploratory analysis, feature selection and feature supplement have been scientifically implemented. The clustering experiment is correctly carried out by using hierarchical clustering technique. The evaluation of this experiment indicates that all the clusters and sub-clusters built in this experiment are statistically significant because of being evaluated Silhouette width and Dunn index. The result of this experiment shows that there are fourteen clusters created (14 university clusters having similar performance inside of each), the full result of clusters of each university has been shown in chapter 4.

5.4 Contributions and impact

This research proposed a new type of ranking criteria based on cluster analysis. The advantage of cluster analysis is that the discrepancies of subjective ranking method such as weighting and arbitrary data selection can be avoided by the use of this technique. This research not only revealed that the traditional and reputational ranking institutions are giving doubtful result, but also offered an alternative clustering measure of university rankings.

5.5 Future Work & recommendations

The future work will focus on the two aspects as follows:

a) As ranking universities are correlated to many properties of universities such as the location of universities, the language speaking of universities or big/small city/country where a university resides in. It is important to keep the representativeness of a data, as it will contribute to the more accurate result universities rankings. Therefore, collecting more data in relation to universities is recommended in the future.

b) Including the feature "Country" at later steps in the analysis (at secondary clustering) may bring more intricate connections within clusters and may reveal unforeseen similarities between university ranks from different countries.

c) Due to the fact that this research conducted hierarchical cluster analysis on the ARWU data successfully, it is reasonable to implement the same analysis on the data from other authoritative ranking institutions, which is helpful for investigating the reliability of these institutions and techniques used by these institutions.

REFERENCES

Alma, B., Coşkun, E., & Övendireli, E. (2016). University Ranking Systems and Proposal of a Theoretical Framework for Ranking of Turkish Universities: A Case of Management Departments. *Procedia - Social And Behavioral Sciences*, *235*, 128-138. doi:10.1016/j.sbspro.2016.11.008

Cangelosi, R., & Goriely, A. (2007). *Biology Direct*, *2*(1), 2. doi: 10.1186/1745-6150-2-2

Chouhan, R., & Chauhan, A. (2014). An Ameliorated Partitioning Clustering Algorithm. *2014 International Conference On Computational Intelligence And Communication Networks*. doi: 10.1109/cicn.2014.119

Cichosz, P. (2015). *Data mining algorithms*. Chichester: John Wiley & Sons Inc.

Davis, M. (2016). Can College Rankings Be Believed?. *She Ji: The Journal Of Design, Economics, And Innovation*, *2*(3), 215-230. doi:10.1016/j.sheji.2016.11.002

Docampo, D. (2011). Adjusted sum of institutional scores as an indicator of the presence of university systems in the ARWU ranking. *Scientometrics*, *90*(2), 701-713. doi: 10.1007/s11192-011-0490-y

Gan, J., Wen, G., Yu, H., Zheng, W., & Lei, C. (2018). Supervised feature selection by self-paced learning regression. *Pattern Recognition Letters*. doi:10.1016/j.patrec.2018.08.029

Hathaway, R., & Bezdek, J. (2002). Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm. *Pattern Recognition Letters*, *23*(1-3), 151-160. doi:10.1016/s0167-8655(01)00115-5

Ivančević, V., & Luković, I. (2018). National university rankings based on open data: A case study from Serbia. *Procedia Computer Science*, *126*, 1516-1525. doi:10.1016/j.procs.2018.08.124

Jolliffe, I. (2002). *Principal component analysis*. New York: Springer.

Kim, J., Kohane, I., & Ohno-Machado, L. (2002). Visualization and evaluation of clusters for exploratory analysis of gene expression data. *Journal Of Biomedical Informatics*, *35*(1), 25-36. doi:10.1016/s1532-0464(02)00001-1

Law, R., Fong, L., & Fong, D. (2015). How useful are university rankings in tourism?. *Annals Of Tourism Research*, *54*, 219-221. doi:10.1016/j.annals.2015.06.005

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S. (2013). Understanding and Enhancement of Internal Clustering Validation Measures. *IEEE Transactions On Cybernetics*, *43*(3), 982-994. doi: 10.1109/tsmcb.2012.2220543

López-Rubio, E., Palomo, E., & Ortega-Zamorano, F. (2018). Unsupervised learning by cluster

Lukman, R., Krajnc, D., & Glavič, P. (2010). University ranking using research, educational and environmental indicators. *Journal Of Cleaner Production*, *18*(7), 619-628. doi:10.1016/j.jclepro.2009.09.015

Nazari, Z., Kang, D., Asharif, M., Sung, Y., & Ogawa, S. (2015). A new hierarchical clustering algorithm. *2015 International Conference On Intelligent Informatics And Biomedical Sciences (ICIIBMS)*. doi: 10.1109/iciibms.2015.7439517

Nisha, & Kaur, P. (2015). Cluster quality based performance evaluation of hierarchical clustering method. *2015 1St International Conference On Next Generation Computing Technologies (NGCT)*. doi: 10.1109/ngct.2015.7375201

Olcay, G., & Bulu, M. (2017). Is measuring the knowledge creation of universities possible?: A review of university rankings. *Technological Forecasting And Social Change*, *123*, 153-160. doi:10.1016/j.techfore.2016.03.029

Pavel, A. (2015). Global University Rankings - A Comparative Analysis. *Procedia Economics And Finance*, *26*, 54-63. doi:10.1016/s2212-5671(15)00838-2

Puth, M., Neuhäuser, M., & Ruxton, G. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, *102*, 77-84. doi:10.1016/j.anbehav.2015.01.010

Rencher, A. (1988). On the Use of Correlations to Interpret Canonical Functions. *Biometrika*, *75*(2), 363. doi: 10.2307/2336185

Saisana, M., d'Hombres, B., & Saltelli, A. (2011). Rickety numbers: Volatility of university rankings and policy implications. *Research Policy*, *40*(1), 165-177. doi:10.1016/j.respol.2010.09.003

Sinharay, S. (2010). An Overview of Statistics in Education. *International Encyclopedia Of Education*, 1-11. doi:10.1016/b978-0-08-044894-7.01719-x

Tabassum, A., Hasan, M., Ahmed, S., Tasmin, R., Abdullah, D., & Musharrat, T. (2017). University ranking prediction system by analyzing influential global performance indicators. *2017 9Th International Conference On Knowledge And Smart Technology (KST)*. doi: 10.1109/kst.2017.7886119

Zerabi, S., & Meshoul, S. (2017). External clustering validation in big data context. *2017 3Rd International Conference Of Cloud Computing Technologies And Applications (Cloudtech)*. doi: 10.1109/cloudtech.2017.8284735

Zhao, H., Han, Q., & Pan, H. (2010). A Hierarchical Clustering Algorithm Based on Grid Partition. *2010 International Conference On Multimedia Communications*. doi: 10.1109/mediacom.2010.46