**Project 4: Cluster Analysis**

**--world university ranking**

Weina Ma



Contents

[Abstract 2](#_Toc469204915)

[1. Data Preparation--Timesdata 2](#_Toc469204916)

[1.1 Question of Interest 2](#_Toc469204917)

[1.2 Data processing 4](#_Toc469204918)

[1.2.1 Outliers 4](#_Toc469204919)

[1.2.2 Scaling 5](#_Toc469204920)

[1.2.3 Clustering Tendency 5](#_Toc469204921)

[2. Modeling 6](#_Toc469204922)

[2.1 All features 6](#_Toc469204923)

[2.2 K-means Clusters—teaching, research 8](#_Toc469204924)

[2.2.1 Optimal Number of Clusters--Within Sum of Squares 8](#_Toc469204925)

[2.2.2 k-means Clustering 9](#_Toc469204926)

[2.2.3 Internal validation 9](#_Toc469204927)

[2.3 Hierarchical Clustering (Complete Link)—rank, research 10](#_Toc469204928)

[2.3.1 Optimal Number of Clusters--Average Silhouette Width 10](#_Toc469204929)

[2.4 Density-based Clustering—rank, teaching 11](#_Toc469204930)

[2.4.1 Density-based Clustering 12](#_Toc469204931)

[2.5 Partitioning Around Medoids(pam) & Hierarchical Clustering-- student staff ratio, international student 12](#_Toc469204932)

[2.5.1 Optimal Number of Clusters--Average Silhouette Width 12](#_Toc469204933)

[2.5.2 Partitioning Around Medoids(pam) 12](#_Toc469204934)

[2.5.3 Internal validation-- Silhouette plot 13](#_Toc469204935)

[2.5.4 Hierarchical Clustering 13](#_Toc469204936)

[2.6 Gaussian Mixture Models—income, citation 14](#_Toc469204937)

[2.7 External cross validation 15](#_Toc469204938)

[2.7.1 Rank vs "teaching","research" 15](#_Toc469204939)

[2.7.2 Continent vs “student\_staff\_ratio","international\_students" 15](#_Toc469204940)

[3. Evaluation [10 points] 16](#_Toc469204941)

[4. Acknowledgment 16](#_Toc469204942)

[5.References 16](#_Toc469204943)

# 

# Abstract

Due to the importance of university ranking in real world, this report is trying to perform cluster analysis using 5 methods, k-means, hierarchical clustering, Density-based Clustering, Partitioning Around Medoids(pam) and Gaussian Mixture Models. We use 3 different methods to determine a suitable number of clusters for each method, including Within Sum of Squares, Average Silhouette Width and Gap Statistic. We use internal validation and external measures to describe and compare the created clusters. Comparing different clustering algorithms, there are several conclusions. First, the performance of (rank, research and teaching) is better than the performance of (continent, student staff ratio, international students) Because the corrected rand is higher. That is to say the relationship between of (rank, research and teaching) is more straightforward and closer. Second, k mean is a better clustering algorithms hierarchical clustering for our dataset. Third, optimal k is harder to decide. Because different methods may give different optimal k for the same dataset.

*Key words: University ranking, clustering,*

# 1. Data Preparation--Timesdata

## 

## 1.1 Question of Interest

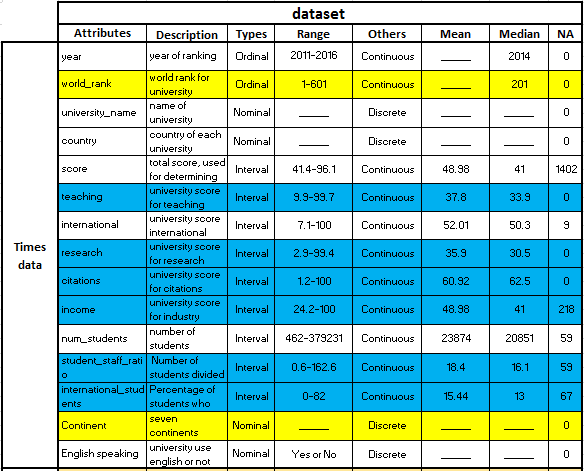
*• Describe which features you want to use for clustering and why.*

In this report, we would like to find groups of objects will be similar to one another and different from the objects in other groups. So we would like to use unsupervised learning, clustering, to study our question of interest.

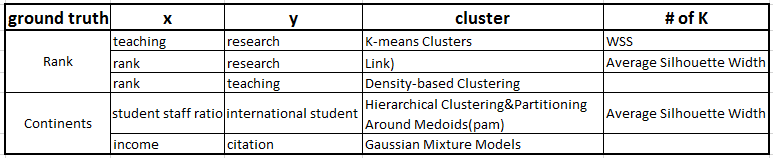
Based on project 2 and project 3, first, we can conclude that teaching and research are the most important features for ranking. Second, student staff ratio, international students number, income, citations are important if we want to distinguish in which continent does a university located. Therefore, we would like to use timesdata again to study these problems, because it has continent feature and all other features we need.

Based on table below, there are 15 features in total. We are interested in teaching, research, student staff ratio, international students number, income, citations most. So I highlight it with blue.

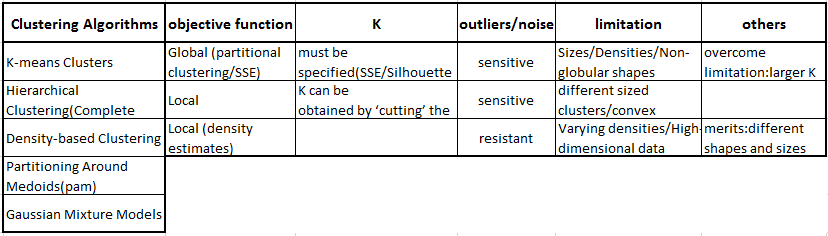
In addition, we want to use a feature as the ground truth (e.g., the continent or the rank) and perform external validation. So I highlight 2 ground truth, continent and world rank with yellow.



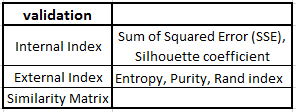
The table below represents our question of interest. Since we want to study relationship between features. We choose teaching, research, student staff ratio, international students number, income, citations and ranking as x or y. And corresponding clustering algorithms, and how to choose number of optimal clusters respectively.



The table below describe some attributes among different clustering algorithms.



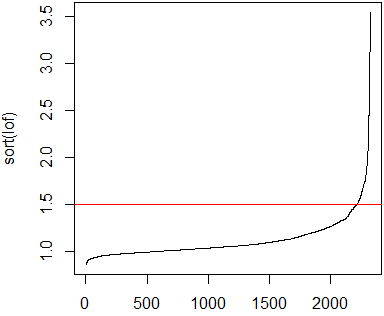
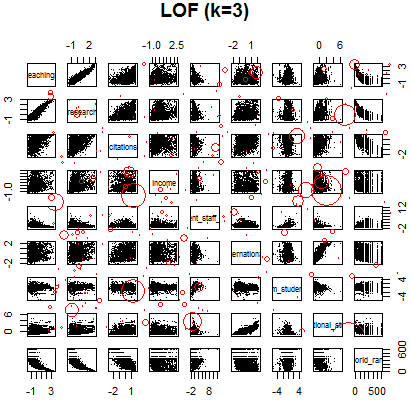
The table below describe main validation methods with their evaluation merics.



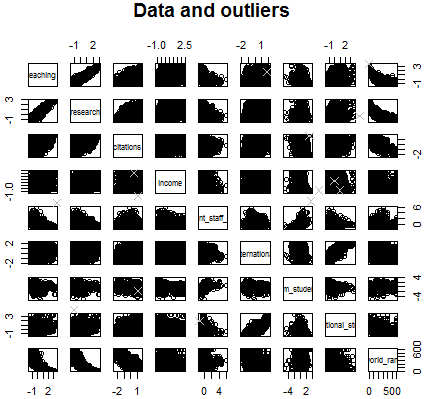
## 1.2 Data processing

### 1.2.1 Outliers

Based on figure below, we can conclude that there are some outliers in our dataset. So We decide to remove all outliers since some clustering algorithms are sensitive to outliers, and it may affect our result if we do nothing.



After getting rid of all outliers, we get a good dataset without outliers.





### 1.2.2 Scaling

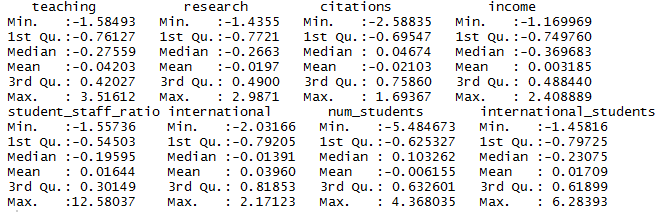
*What is the scale of measurement of the features and what are appropriate distance measures?*

Because some algorithms, like k means, don’t like missing values. So we omit the observations with those missing values. Then, we check if there are duplicated data in our dataset. Luckily, we don’t have duplicated data.

Next, we want to scale our dataset since some attribute with a large range may dominate the others for the distance calculation. Therefore, we scale each column in the data to zero mean and unit standard deviation (z-scores).

Finally, we get our final cleaned ad scaled dataset below.



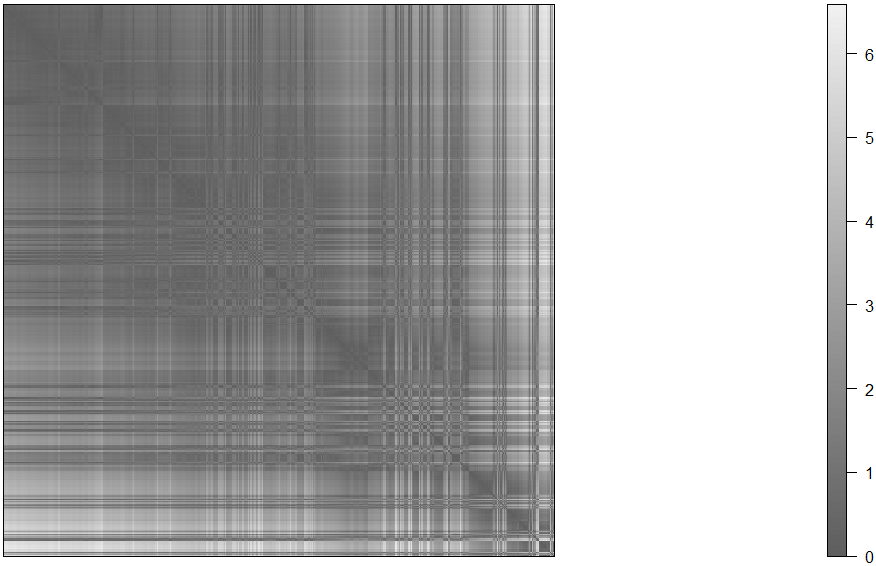
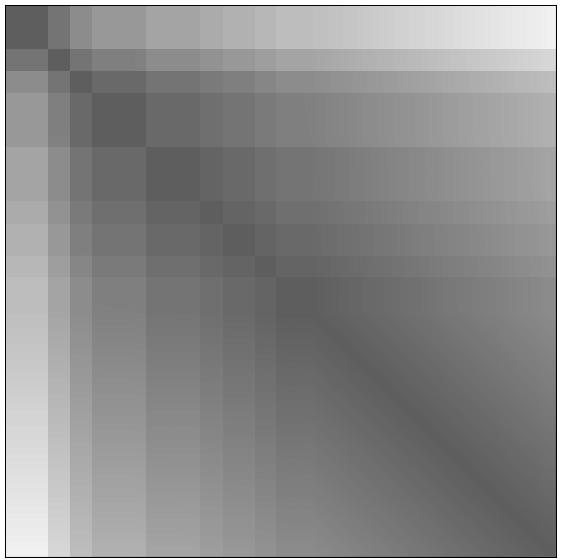


### 1.2.3 Clustering Tendency

Because most clustering algorithms will always produce a clustering, if we need. So we would like to check cluster tendency first since it can prevent from creating clusters without a cluster structure.

The first figures below represent clustering tendency of teaching and world rank. The second figures below represent clustering tendency of teaching and research.

It looks like they do have several clusters existing inside. Thus, we can proceed our next step, modeling.

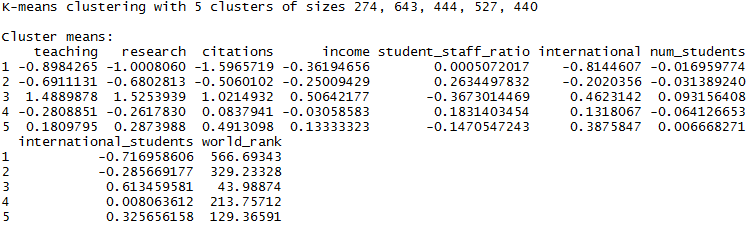


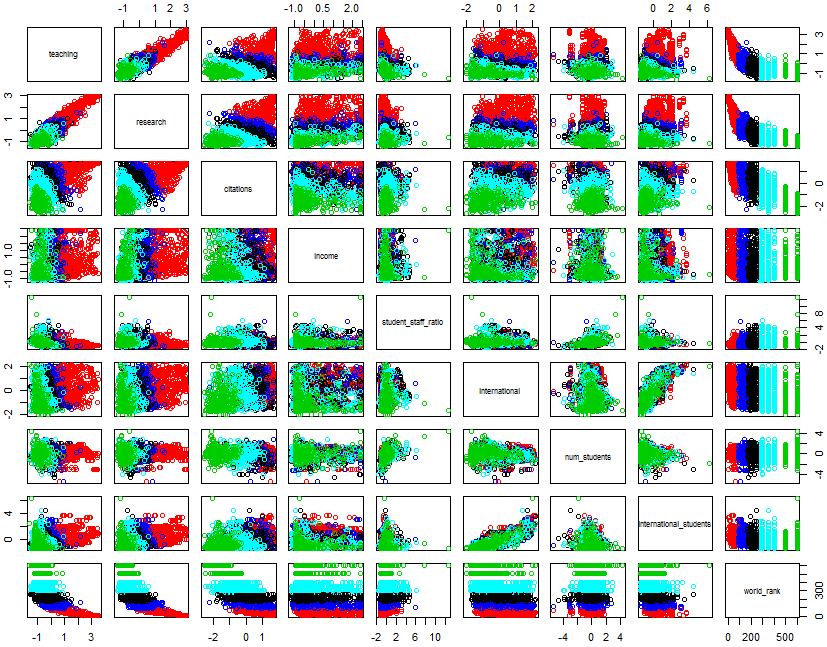
# 2. Modeling

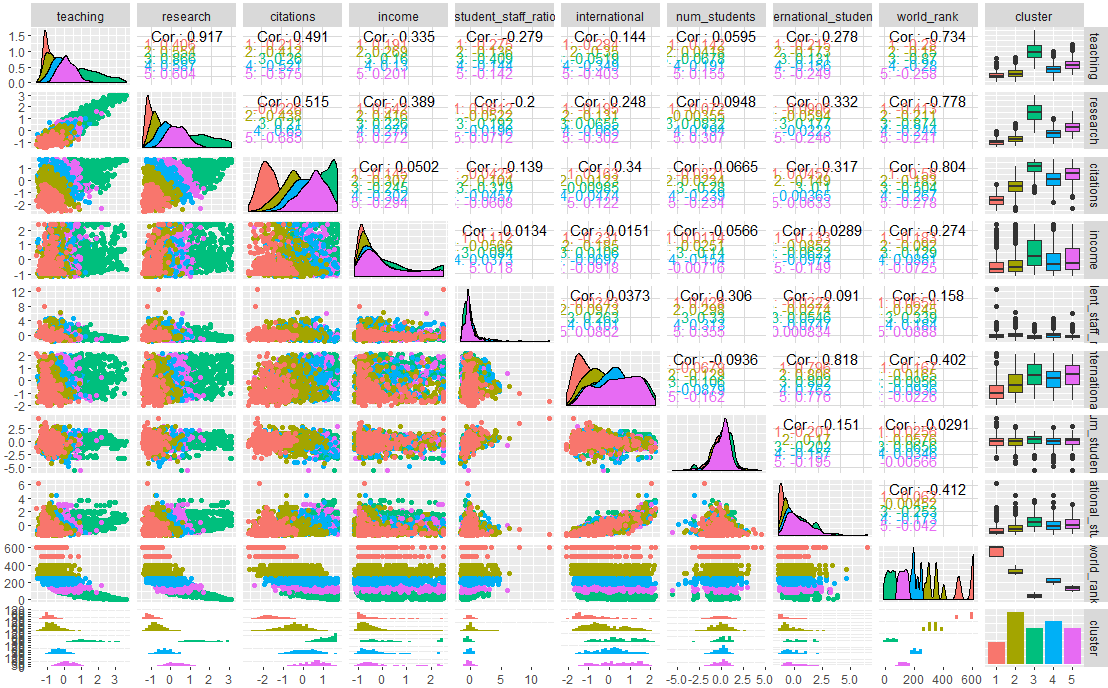
## 2.1 All features

Although we already have several questions of interest that we would like to learn, still it is beneficial to create clusters among all numeric features. We choose to use k means to create models.

Based on figure below, we can see the roughly distribution and relationship between features. Some features are linear correlated while others don’t.







## 2.2 K-means Clusters—teaching, research

*• Perform cluster analysis using several methods (at least k-means and hierarchical clustering)*

*using different feature subsets.*

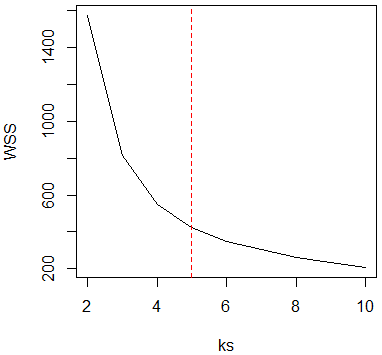
We would like to use k means to create model between teaching and research.

*• How did you determine a suitable number of clusters for each method?*

### 2.2.1 Optimal Number of Clusters--Within Sum of Squares

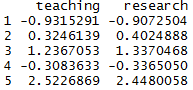
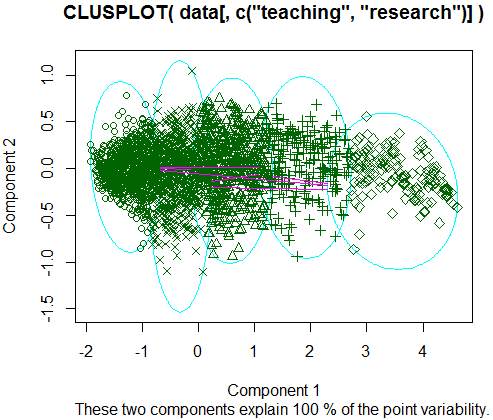
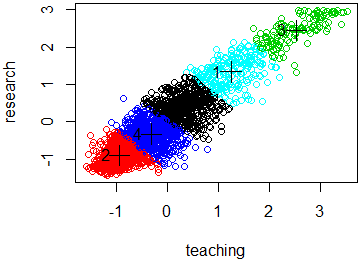
We would like to choose the optimal number of clusters before creating models. In this case, we use within sum of squares. Because different methods may give different optimal number of clusters. Although we have tried several methods to get the optimal k, we won’t use all of it. We would pick one method to use.

Based on figure below, we find that the optimal k is between 5 and 6. We choose 5 as k.

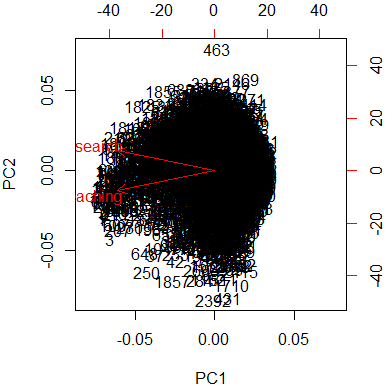
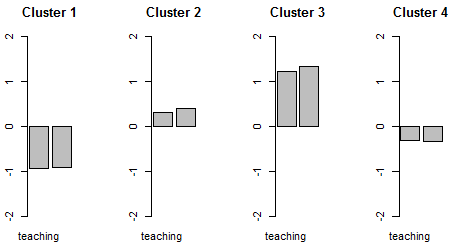


### 2.2.2 k-means Clustering

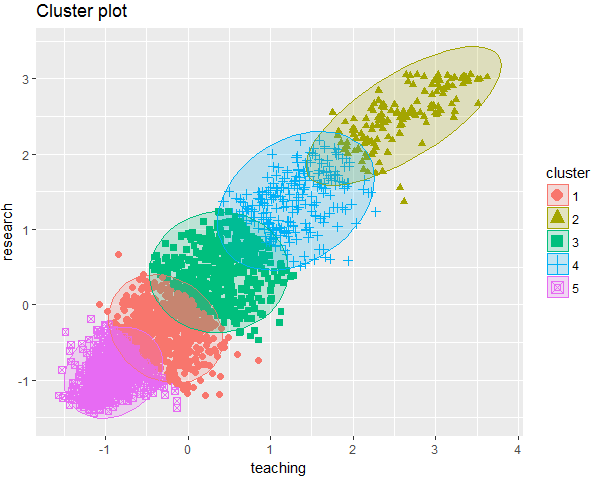
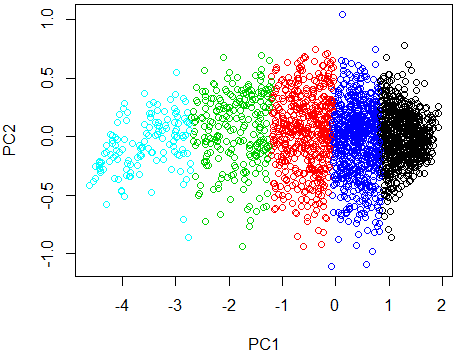
Then we build the models with 5 components. The figure below shows the relationship between teaching and research. The performance is good based on visual check. Also, we get parameters for each cluster.



Because pca and help us reveal some clusters, so we create principal components analysis. And we find that teaching and research are worthy to proceed the clustering.



Also, the cluster distribution of two components is similar to original data.



### 2.2.3 Internal validation

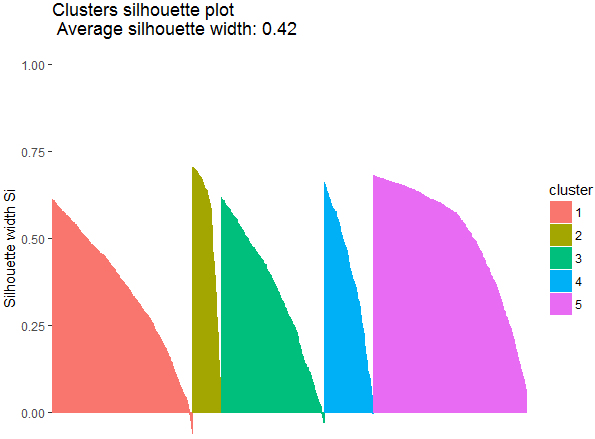
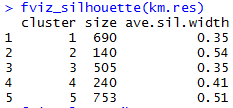
*• Use internal validation measures to describe and compare the clusterings and the clusters (some*

*visual methods would be good).*

We use Silhouette plot and coefficient to evaluate the models.

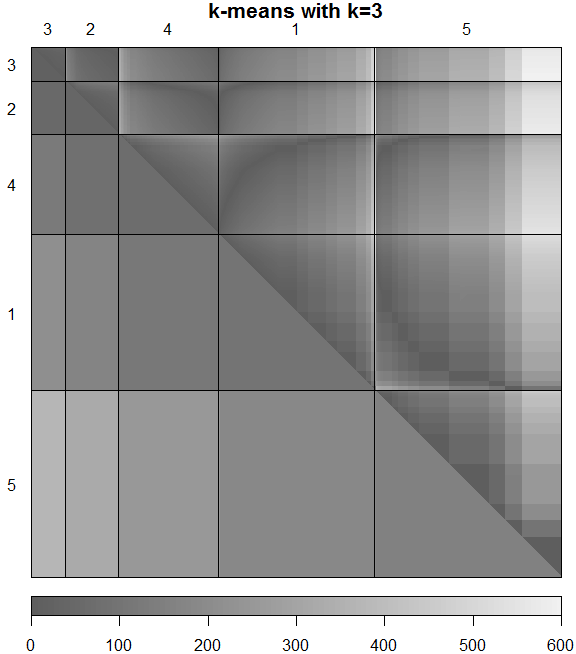
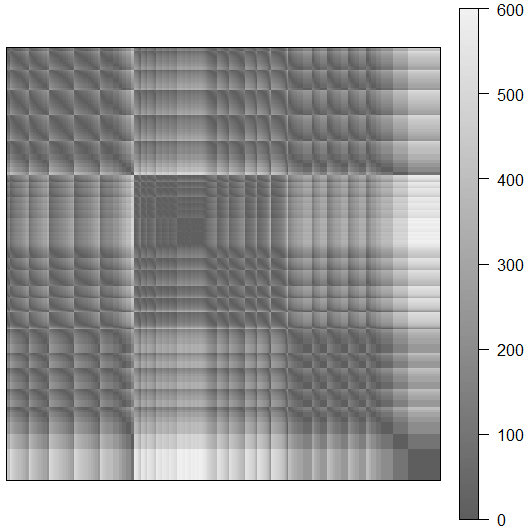
**2.2.3.1 Silhouette plot**

Because another Silhouette plot cannot display correctly since our n is too large. So, I choose another Silhouette plot. The result shows that the performance of each group is good. Although there are some negative silhouette values, but quite a few. And average silhouette width is around 0.35 to 0.5. Not very high, but it is good.

**2.2.3.2 Visualize the Distance Matrix**

This dissplot which rearranges clusters, adds cluster labels, and shows average dissimilarity in the lower half of the plot. The figure looks good. There are 5 clusters based on the right figure.



## 2.3 Hierarchical Clustering (Complete Link)—rank, research

*• Perform cluster analysis using several methods (at least k-means and hierarchical clustering)*

*using different feature subsets.*

Then, we use hierarchical clustering to create model between rank and research.

### 

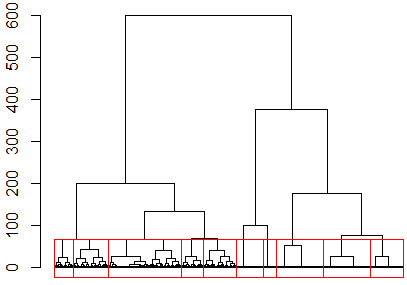
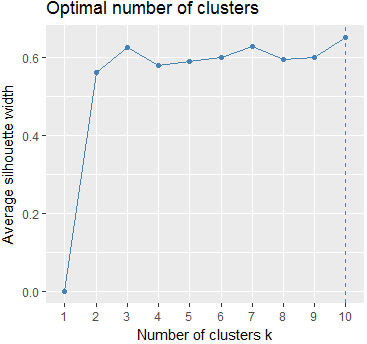
### 2.3.1 Optimal Number of Clusters--Average Silhouette Width

*• How did you determine a suitable number of clusters for each method?*

In this case, we use average silhouette width to find the optimal k.

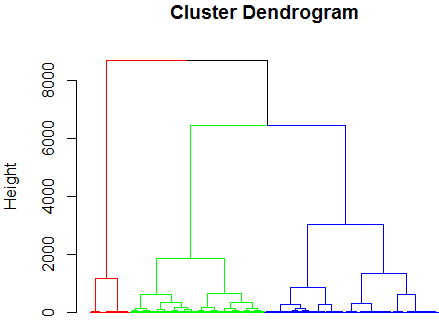
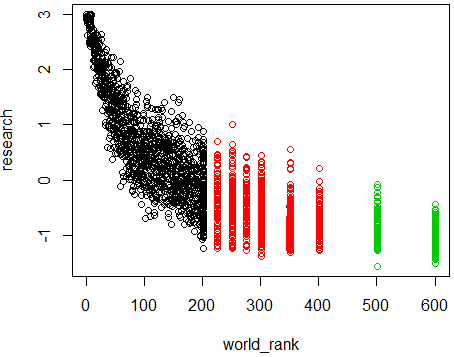
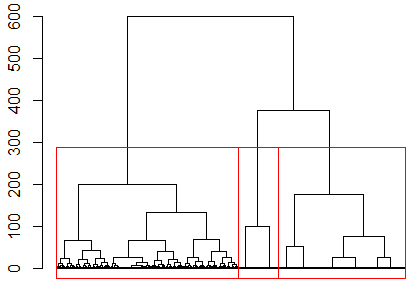
It shows the optimal k is 10. However, we find that 3 is also a good cluster. So, we proceed the model with different optimal k, 3 and 10.

First, we choose 10 as k and create clusters. The hierarchical clustering result are shown below. 10 clusters look good. But we want to try k=3.



Second, we choose 3 as k and create clusters. The hierarchical clustering result are shown below. 3 clusters work better according to visual check. It is more straightforward and neat if k=3. Also, it could represent most information. And the cluster is good based on scatter plot.

Therefore, we choose model with k=3 as final model.



## 2.4 Density-based Clustering—rank, teaching

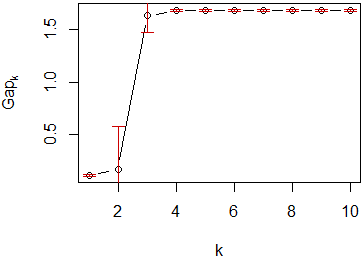
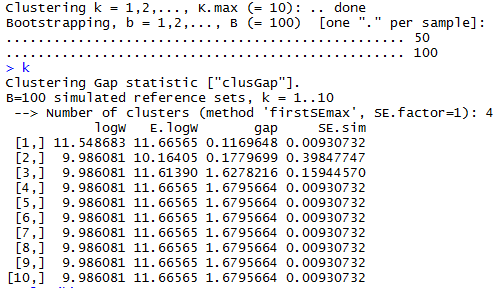
*• Perform cluster analysis using several methods (at least k-means and hierarchical clustering)*

*using different feature subsets.*

Then, we use density based clustering to create model between rank and teaching.

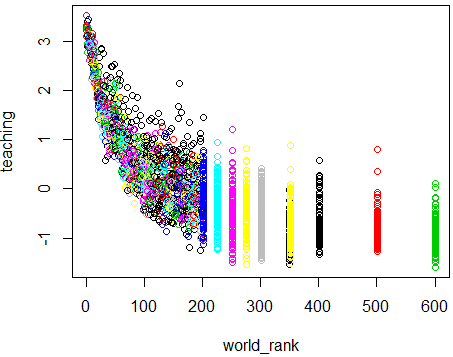
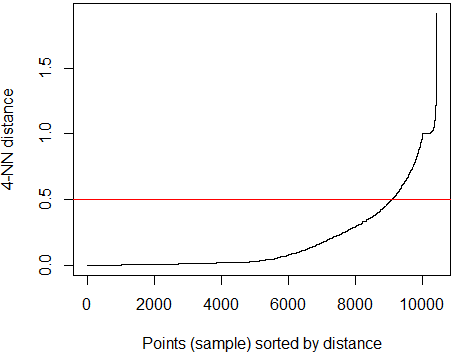
*• How did you determine a suitable number of clusters for each method?*

We use DBSCAN to conduct the model. And DBSCAN does not require one to specify the number of clusters in the data. Therefore, we don’t need this step. However, we want to see which recommendation number of k is if we want to use k means or hierarchical clustering. So we use gap statistic to determine number of k. The table and figure below show that k =4. And we may use it in k means or hierarchical clustering.



### 2.4.1 Density-based Clustering

Then we build the models with 170 components. The figure below shows the relationship between teaching and world rank. The performance is not good based on visual check.





## 2.5 Partitioning Around Medoids(pam) & Hierarchical Clustering-- student staff ratio, international student

*• Perform cluster analysis using several methods (at least k-means and hierarchical clustering)*

*using different feature subsets.*

Then, we pam clustering and hierarchical clustering to create model between students staff ratio and international student.

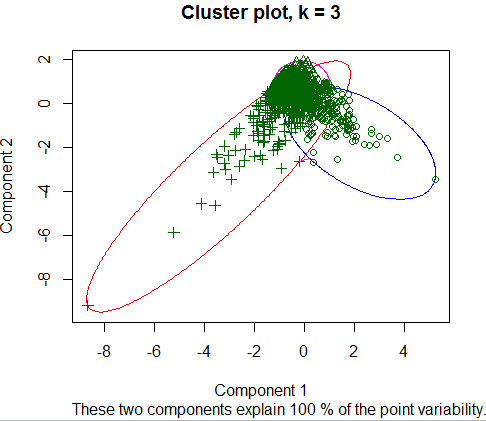
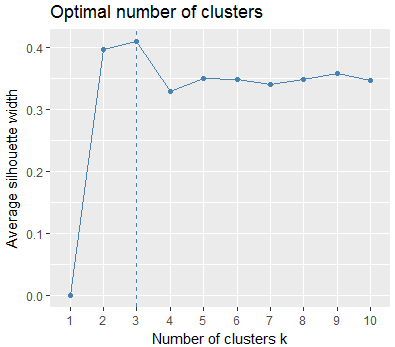
### 2.5.1 Optimal Number of Clusters--Average Silhouette Width

*• How did you determine a suitable number of clusters for each method?*

Based on average silhouette width, we find that 3 is the optimal number of clusters. So k=3.

### 2.5.2 Partitioning Around Medoids(pam)

Then we build the model based on k=3. We get 2 cluster plots. Based on cluster plot, we can tell that this clustering algorithm is not good or bad, we have to see internal validation.

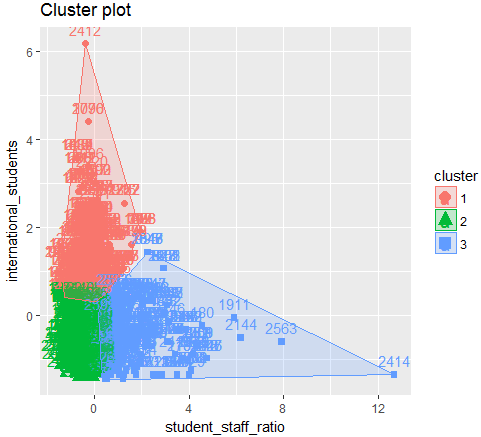


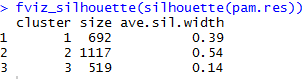
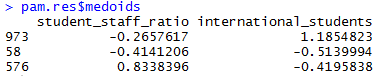
### 2.5.3 Internal validation-- Silhouette plot

*• Use internal validation measures to describe and compare the clusterings and the clusters (some*

*visual methods would be good).*

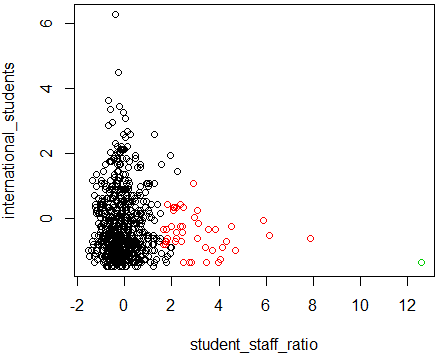
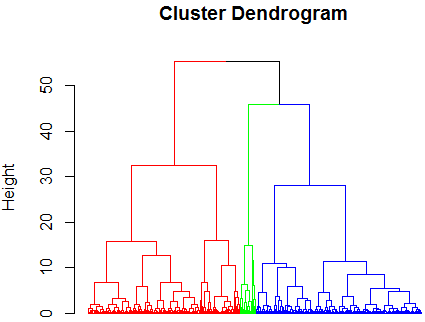
Because another Silhouette plot cannot display correctly since our n is too large. So, I choose another Silhouette plot. The result shows that some group perform good while others don’t. There are some negative silhouette values. And average silhouette width is around 0.14 to 0.5. It seems cluster 2 is better than other 2 clusters.

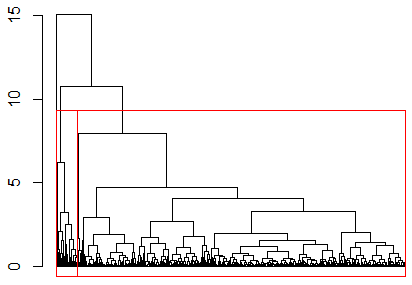




### 2.5.4 Hierarchical Clustering

Because the performance of DBSCAN in this case is not so good, so we would like to use hierarchical clustering to create clusters again. The results shown as below. The performance is also not good. Based on scatter plot and cluster dendrogram.





## 2.6 Gaussian Mixture Models—income, citation

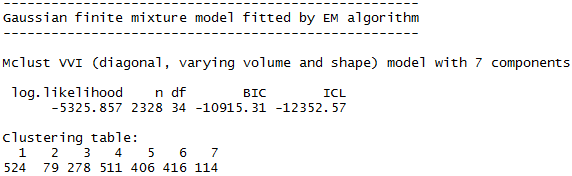
*• Perform cluster analysis using several methods (at least k-means and hierarchical clustering)*

*using different feature subsets.*

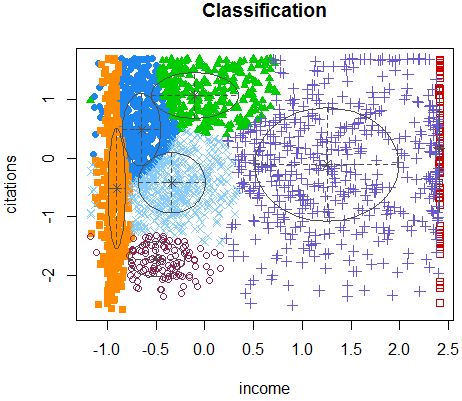
Last, we use Gaussian mixture model to evaluate relationship between income and citation.

*• How did you determine a suitable number of clusters for each method?*

Gaussian mixture model tell this model contain 7 components.



Based on figure below, we can conclude that this clustering algorithm is suitable in this case. The clustering is reasonable according to visual check.



## 2.7 External cross validation

*• Can you use a feature as the ground truth (e.g., the continent or the rank) and perform external*

*validation?*

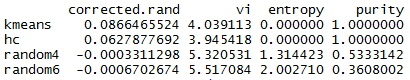
We would like to use rank and continent as ground truth to perform external validation and evaluate the performance of our models.

### 2.7.1 Rank vs "teaching","research"

Based on table below, there are 2 conclusions.

First, we can conclude that the performance of our 2 models is good because corrected. Rand is much higher than random cases.

Second, k means is better than hierarchical clustering because the corrected. Rand is higher.

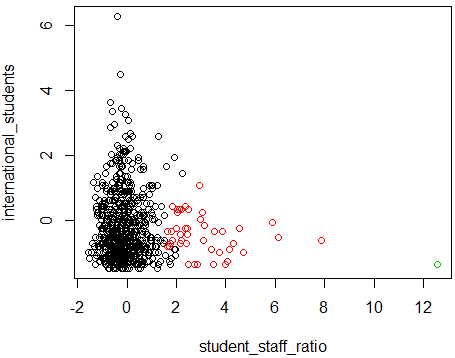
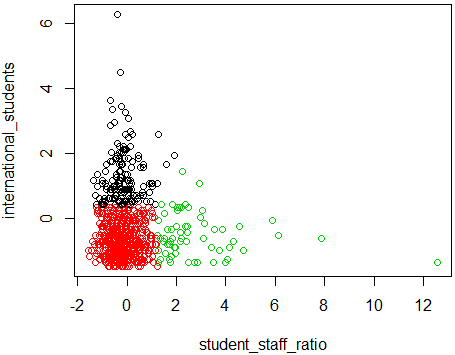


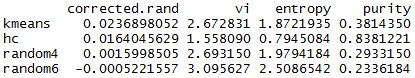
### 2.7.2 Continent vs “student\_staff\_ratio","international\_students"

Based on table below, there are 2 conclusions.

First, we can conclude that the performance of our 2 models is not so good because corrected. Rand is close to random cases. But is still a little bit higher. Second, k means is better than hierarchical clustering because the corrected. Rand is higher.

Also, the figure in the left is the clusters created by k means. The figure in the right is clusters created by hierarchical clustering. Based on visual check, we can also conclude that the performance of k means is better than hierarchical clustering in this case.





# 3. Evaluation [10 points]

*• Describe your results. What findings are the most interesting?*

Based on the models above, there are several conclusions.

First, the performance of (rank, research and teaching) is better than the performance of (continent, student staff ratio, international students) Because the corrected rand is higher. That is to say the relationship between of (rank, research and teaching) is more straightforward and closer. If you want to evaluate the rank is high or median, research and teaching are suitable features that we could use. On the contrary, if you want to evaluate which continent a university is located, student staff ratio and international student may not be a very good index to reveal some rule.

Second, k mean is a better clustering algorithms hierarchical clustering for our dataset. Third, optimal k is harder to decide. Because different methods may give different optimal k for the same dataset.

# 4. Acknowledgment

Specially thanks for [Dr. Michael Hahsler](http://michael.hahsler.net/)’s great help and mentoring on this project

# 5.References

Source: http://michael.hahsler.net/SMU/EMIS7332/

1. ARWU (Shanghai Ranking) Methodology:

<http://www.shanghairanking.com/ARWUMethodology2016.Html>

1. WUR (Times Ranking) Methodology:

https://www.timeshighereducation.com/news/rankingmethodology2016

1. Dataset

https://www.kaggle.com/mylesoneill/world-university-rankings

1. R code

<http://michael.hahsler.net/SMU/EMIS7332/>

1. R code

<https://emudrak.github.io/2015-01-15-cornell/lessons/R/03-data-frames.html>

1. R package

<http://stackoverflow.com/questions/17376939/problems-when-trying-to-load-a-package-in-r-due-to-rjava>

1. Continents

<https://en.wikipedia.org/wiki/List_of_territorial_entities_where_English_is_an_official_language>

1. Continents

<https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population>

1. Word cloud

<https://tagul.com/cloud/3>

1. http://www.sthda.com/english/wiki/determining-the-optimal-number-of-clusters-3-must-known-methods-unsupervised-machine-learning