**Project 3: Association Rule Mining**

Weina Ma



Contents

[1. Abstract 3](#_Toc466930469)

[2. Data Preparation --Timesdata & cwurData 3](#_Toc466930470)

[2.1 Question of Interest 3](#_Toc466930471)

[2.2 Question of interest one-timesdata 4](#_Toc466930472)

[2.3 Question of interest two-cwurdata 5](#_Toc466930473)

[3. Modeling --Timesdata 6](#_Toc466930474)

[3.1 country 6](#_Toc466930475)

[3.1.1Create sets of frequent, closed and maximal itemsets 6](#_Toc466930476)

[3.1.2 Create sets of association rules 7](#_Toc466930477)

[3.1.3 Use filtering, sorting, tables, and visualization to discuss the found patterns 8](#_Toc466930478)

[3.2 Area-seven continents 10](#_Toc466930479)

[3.3 Subset-China 11](#_Toc466930480)

[4. Modeling -- cwurdata 12](#_Toc466930481)

[4.1 Create sets of frequent, closed and maximal itemsets 12](#_Toc466930482)

[4.2 Association rules 13](#_Toc466930483)

[5. Evaluation 14](#_Toc466930484)

[5.1 Three major findings 14](#_Toc466930485)

[5.2 Recommendations and advantages 15](#_Toc466930486)

[6. Acknowledgment 16](#_Toc466930487)

[7. References 16](#_Toc466930488)

# 1. Abstract

Due to the importance of university ranking in real world, this report is trying to figure out association rules between countries, rankings, research and some other features while we take no account of year and learn which factors have the greatest impact for university ranking improvement between 2014 and 2015. According to association rules, there are three major conclusions. First, we figure out ranking was influenced by academic work, such as research, citation and teaching. Second, different regions have different university culture. Third, improved publications, improved patents and improved broad impact are very important if universities want to improve the rank next year beside academic.

*Key words: University ranking, Association Rule*

# 2. Data Preparation --Timesdata & cwurData

## 

## 2.1 Question of Interest

In this report, we will use Timesdata and cwurData to build association rules. This report is interested in two things.

First is to study association rules between countries, rankings, research and some other features while we take no account of year.

We build 3 subsets of timesdata. Those 3 subsets are country, area and China.

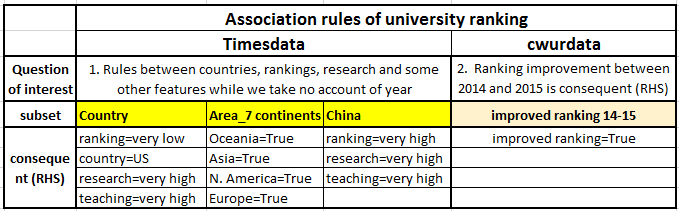
We use ranking/country/research/teaching in subset country as consequent (RHS) respectively.

We use Oceania/Asia/North America/Europe in subset Area as consequent (RHS) respectively.

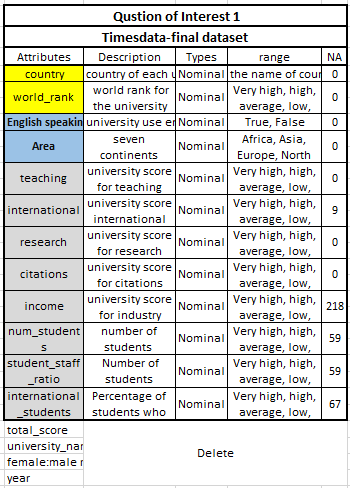
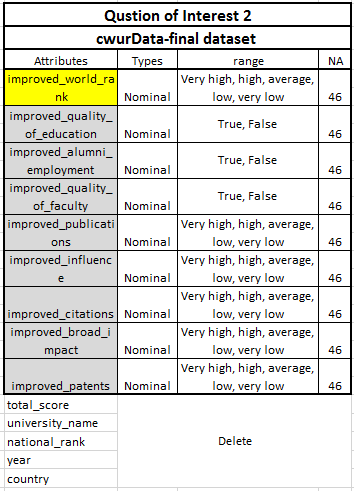
We use ranking/research/teaching in subset China as consequent (RHS) respectively.

Then we use other features to be antecedent (left-hand-side or LHS) during the process of mining rules.

Second is to learn which factors have the greatest impact for university ranking improvement between 2014 and 2015. So, we use improved ranking between 2014 and 2015 in cwurdata as consequent (RHS). And we use other features to be antecedent (left-hand-side or LHS) during the process of mining rules.



The final data set are shown below.

## 2.2 Question of interest one-timesdata

For the first question of interest, we use times data to build the association rules. The observations are 2603 after data cleaning.

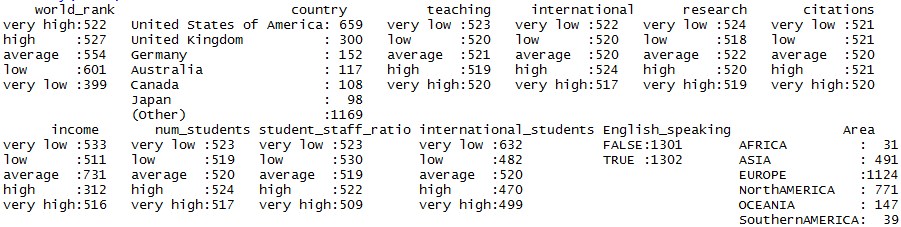


We import 2 new feature. One is binary feature English speaking. We use this variable to distinguish whether a university use English or not. We search the internet. We assign “True” if a specific country use English as official language. We get 1301 values. Accordingly, we assign “False” if a specific country does not use English as official language. We get 1302 values. The new variable is shown below.

The other feature is Area, which contains 7 continents, Africa, Asia, Europe, North America, Southern America and Oceania. The reason why we import this variable is some countries only have a few universities in the data set. Sometimes it is difficult to make some conclusions. So, using continents to distinguish universities is a good idea to analyze.

Then, we discretized continuous variables assigning frequent levels. (very high, high, average, low and very low).

Third, we get rid of useless and bad variables with lots of missing values, such as score, year, university names, female male ratio and we get our final data set for this question.



In addition, since area and country are highly correlated, thus, we should not put them in the same dataset to build association rules. So we will use these 2 features to set the rules respectively.

## 2.3 Question of interest two-cwurdata

Second question of interest 2 is to learn which factors have the greatest impact for university ranking improvement between 2014 and 2015.

Because we are interested in improvement between two years. So we would better choose data with equal length of year.

For cwurData, there are 1000 observations in 2014 and 2015. So it is a larger and better dataset if we do an improvement transformation from 2014 to 2015. That is to say, observations of 2014 minus observations of 2015 will not have too many missing values. Then we assign “True” to improvement and “False” to no improvement.

For Timesdata dataset, there are 400 observations each year. And we have tried to use times data to build association rules of improvement of 2014 and 2015. However, since the data set is smaller, which 400. After subtraction, there are only 370 observations. Also, the effect of association rules of times data set is bad after deleting missing values. So, we use cwurdata instead.

|  |  |
| --- | --- |
| cwurData | Timesdata |
|  |  |

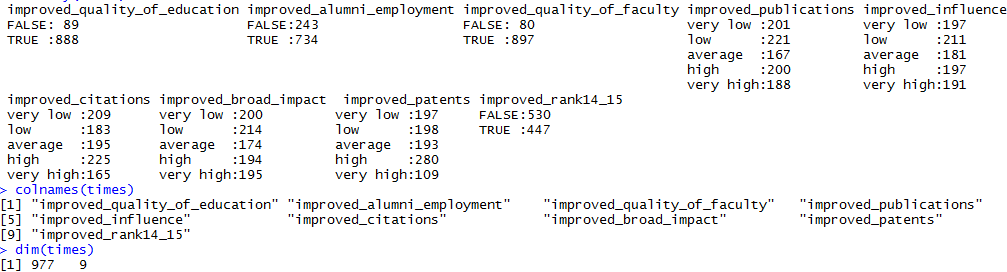
Next, we calculate the improvement of other features between 2014 and 2015.

Since features of quality of education, alumni employment and quality of faculty are imbalanced, which means the majority of result is true. Thus, it causes inaccuracy of discretization. Therefore, we use “True” and “False” to represent the improvement of these 3 features.

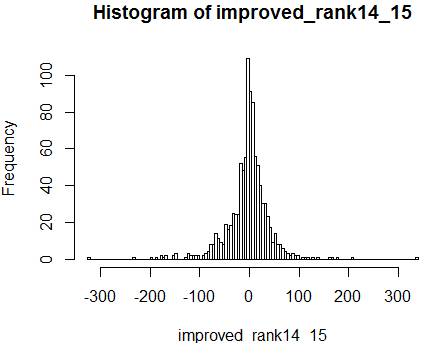
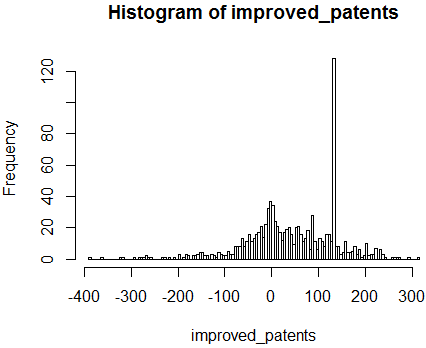
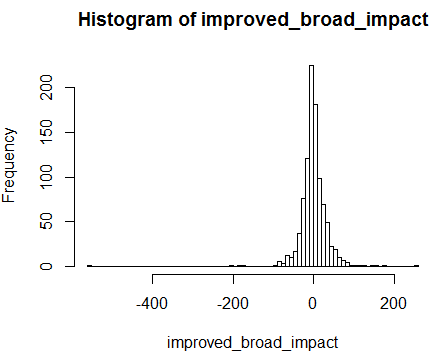
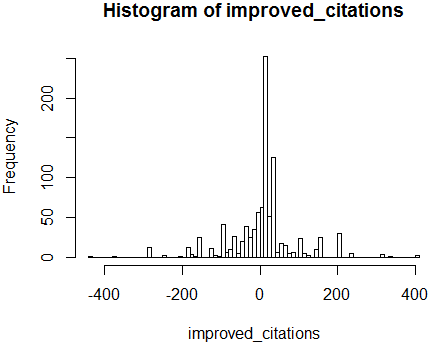
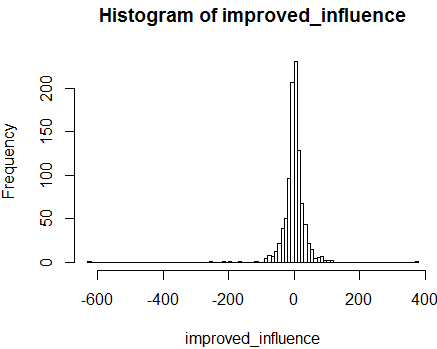
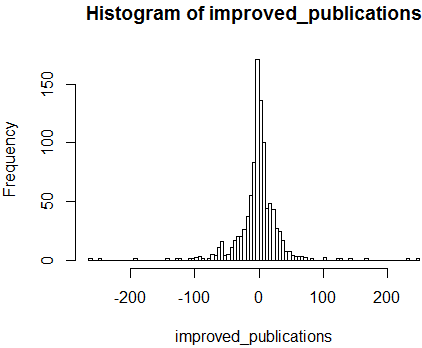
Then, we discretized continuous variables assigning frequent levels. (very high, high, average, low and very low).

In addition, our final data set is large after transformation. And missing values are 46, which is nice.

Last, we delete missing values and get the final data set with 977 observations.



After subtraction of all features, we can conclude that the transformation is reasonable because the features are normally distributed according to histograms below.



# 3. Modeling --Timesdata

## 3.1 country

We use ranking/country/research/teaching in subset country as consequent (RHS) respectively.

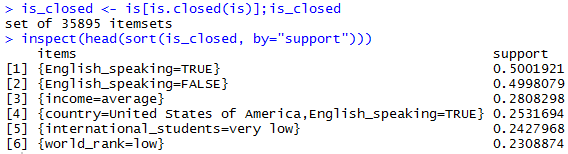
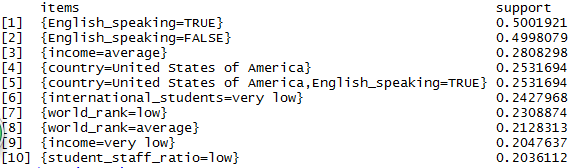
### 3.1.1Create sets of frequent, closed and maximal itemsets

For frequent dataset, there are 56800 itemsets. 

The most support item is English speaking(True and False), then the majority income of university is average. Also, most universities come from America and speak English, which is reasonable.

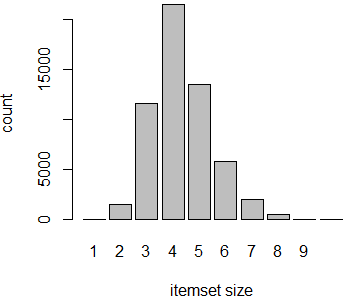
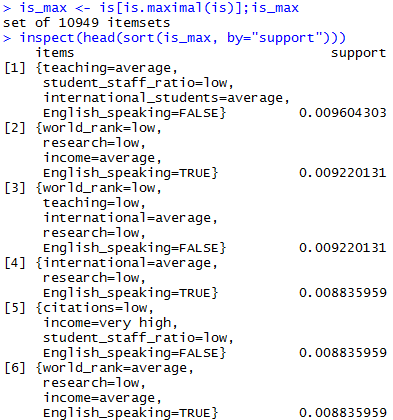
For closed dataset, there are 35895 itemsets.

For maximal dataset, there are 10949 itemsets.



From the barplot below, we can figure out that the majority rules contain 4 itemsets.

Also, barplot of frequent, closed and maximal shows our result is reasonable. Because frequent is the largest and maximal is the smallest dataset.

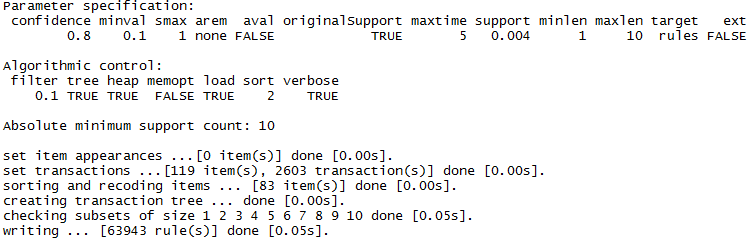
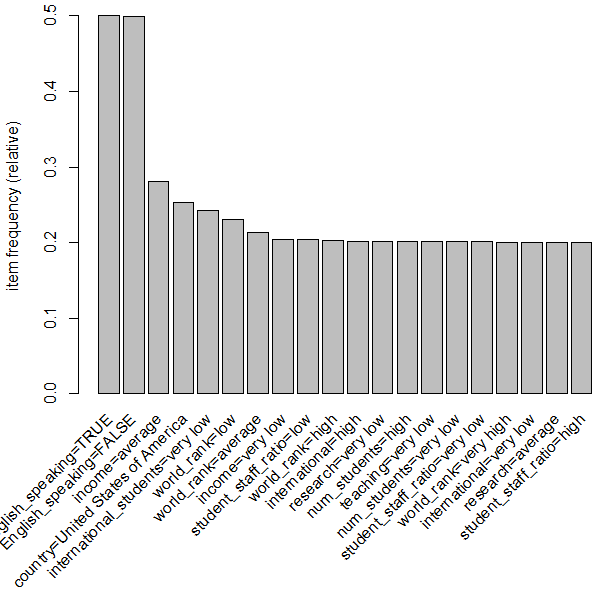
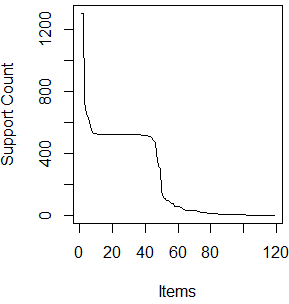


### 3.1.2 Create sets of association rules

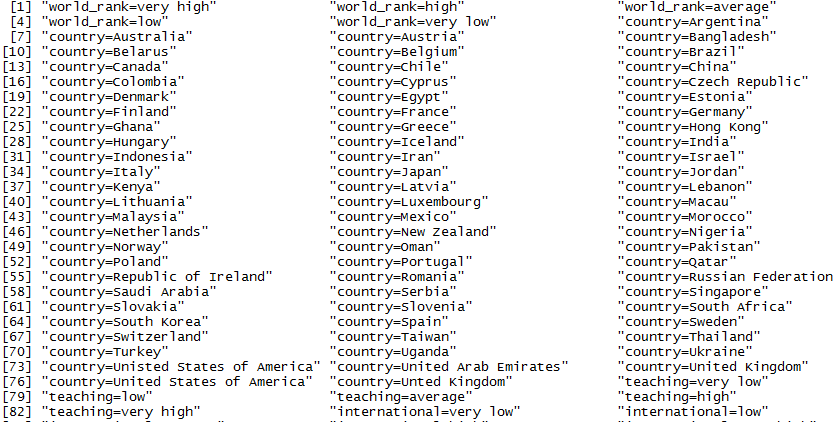
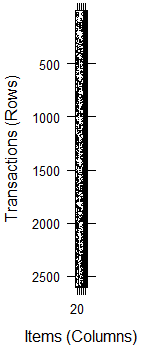
We create association rules of subset country. Then we get 63943 rules. 

We can realize from the frequency plot below that half universities use English to teach and half not. the majority income of university is average. Also, most universities come from America. The number of international students, income, student and staff ratio of most universities are low.

The curve below describes the trend of support.

We know that the black place represents the university has the corresponding values while the white place represents not. And the first half of the columns are different countries. So, the output makes sense because each university correspond to one country.



The confidence is an estimation of condition probability. The conditional probability that rhs variable happens under the condition of lhs happens. Confidence is 1 means the rule is always right in the dataset.

Confidence of 0.917 represents that if the population of research is very low, citation is very low, international students is very low happened, there are 91.7 percent chance that research is very low, citation is very low, international students is very low and world rank is very low.

That is to say confidence of 0.917 represents that in the population of research is very low, citation is very low, international students is very low, there are 91.7 percent world rank is very low.

Support means in all kinds of combination, there are 5.15 percent chance that research is very low, citation is very low, international students is very low, world rank is very low happen. So support is less useful than confidence and lift in analyzing this table. But it is useful to describe the frequent items.

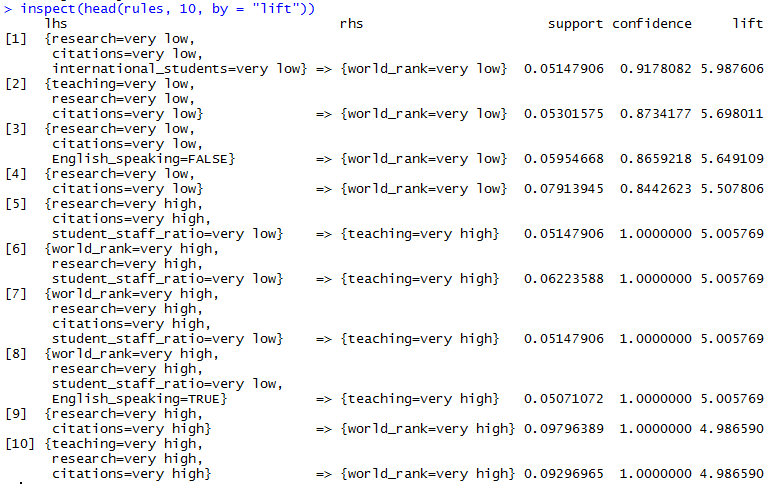
Lift is a measurement that take statistical dependence into account for rule.

If lift < 1, then lhs and rhs are negatively associated.

If lift > 1, then lhs and rhs are positively associated.

The larger the lift is, the more positive the relationship are.

So we sort the lift, lift of first observation is 5.9876. So the group of research is very low, citation is very low, international students is very low and world rank is very low are positively associated. It means if group of research is very low, citation is very low, international students is very low happened, it may cause word rank is very low happen.



### 3.1.3 Use filtering, sorting, tables, and visualization to discuss the found patterns

Because we want to study association rules between countries, rankings, research and some other features while we take no account of year.

We use ranking/country/research/teaching in subset country as consequent (RHS) respectively.

Then we use other features to be antecedent (left-hand-side or LHS) during the process of mining rules.

Then we sort the lift and get the tables below.

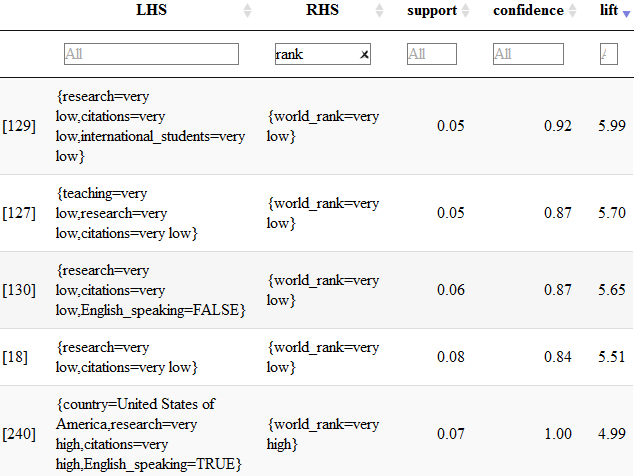
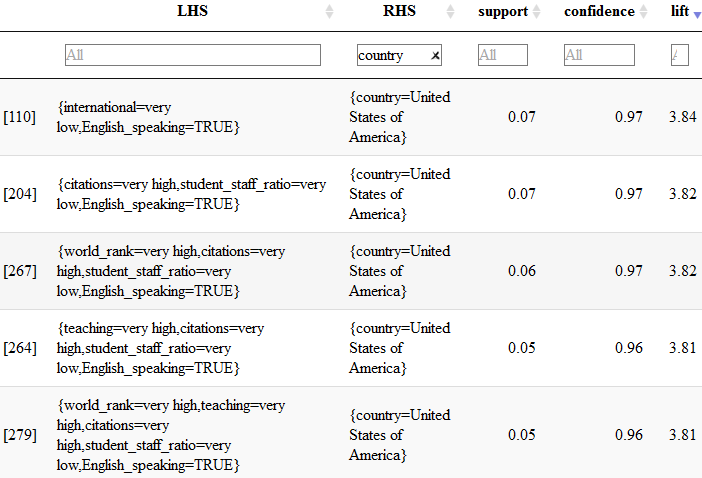
Lift is a measurement that take statistical dependence into account for rule.

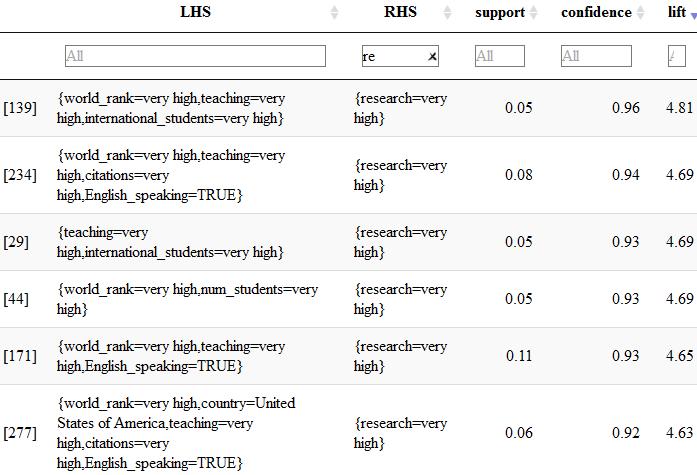
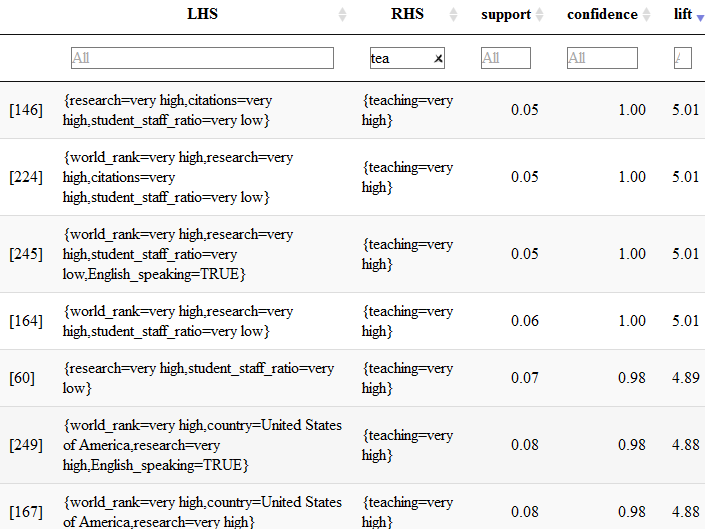
If lift < 1, then lhs and rhs are negatively associated. If lift > 1, then lhs and rhs are positively associated.

The larger the lift is, the more positive the relationship are.

For the first table, Lift of first observation is 5.99. So the group of research is very low, citation is very low, international students is very low and world ran is very low are positively associated. It means if group of research is very low, citation is very low, international students is very low happened, it may cause word ran is very low happen.

For the second table, lift of first observation is 3.84. So the group of international is very low, English speaking is True and country is US are positively associated. It means if group of international is very low, English speaking is True happened, and country is US will happen.

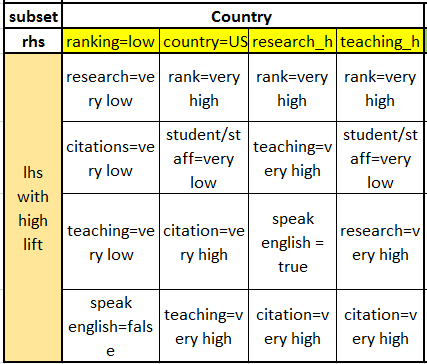
And We summarize the output and count the frequency of created items of the rules. We get the table below.

If research=very low, citations=very low, teaching=very low, speak English=false, then the ranking=low.

If rank=very high, student/staff=very low, citation=very high, teaching=very high, then the country=US.

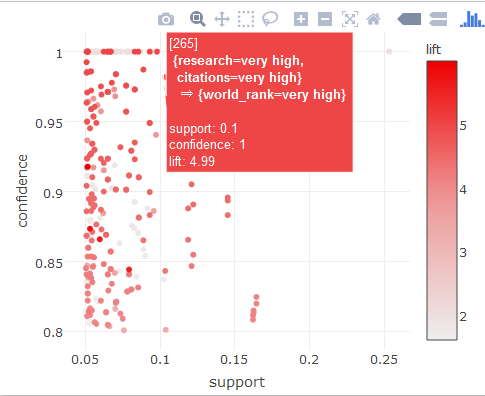
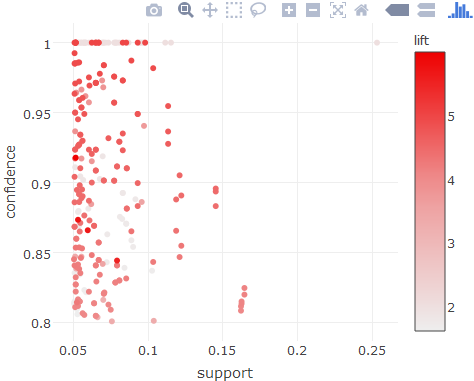
If rank=very high, teaching=very high, speak English = true, citation=very high, then the research=very high.

If rank=very high, student/staff=very low, research=very high, citation=very high, then the teaching=very high.



Then we visualize the rules. And choose one of the dark red with confidence equal to 1.

We find that if research is very high, citation is very high, the world rank is usually very high.



Next, we add phi and gini index to evaluate the association rules.

The phi coefficient is a measure of the degree of association between two binary variables. This measure is similar to the correlation coefficient in its interpretation.

There is general rule of thumb for correlation coefficients and we can use the same rule for the Phi coefficient.

-1.0 to -0.7 strong negative association. -0.7 to -0.3 weak negative association.

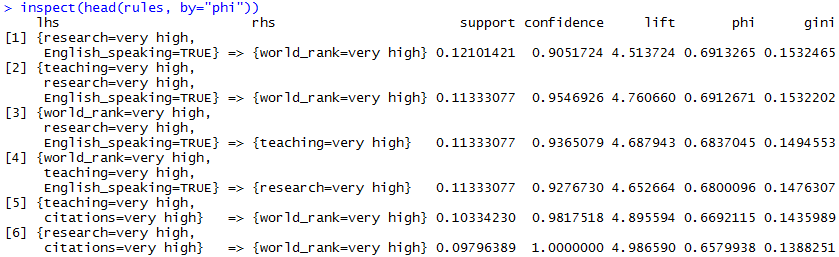
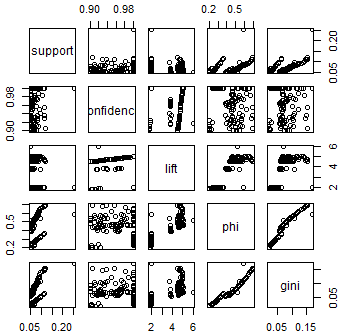
-0.3 to +0.3 little or no association. +0.3 to +0.7 weak positive association.

+0.7 to +1.0 strong positive association

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

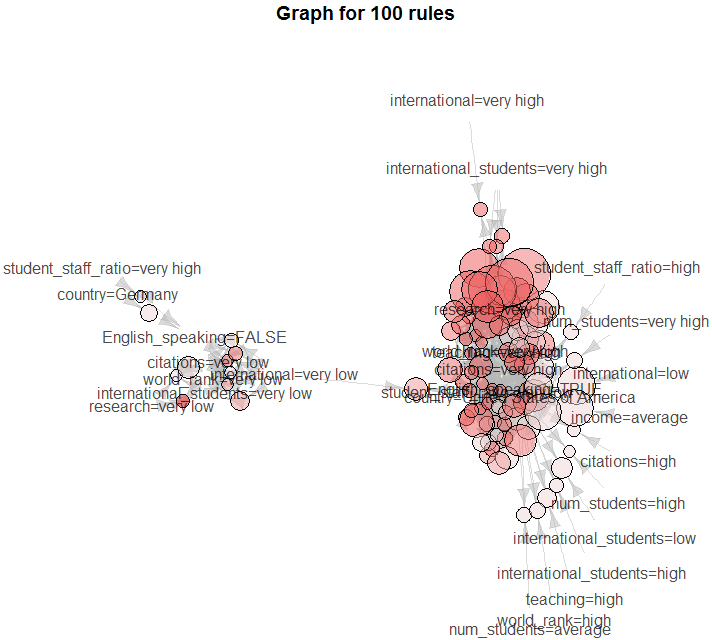
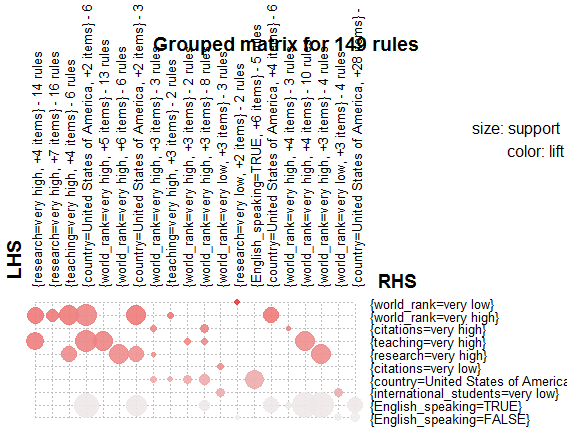
Since gini index estimate the misclassification rate. So the smaller the better.

So we sort the data by phi. It shows research=very high, English speaking=True are correlated with world rank= very high.

Also, we could visualize the association rules that we created.

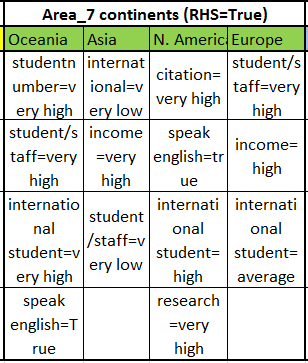
We can conclude that the relationship between lhs and rhs. For instance, if research is very high (14 rules), teaching is very high (6 rules), country is US (6 rules), then the world rank is very high.



## 3.2 Area-seven continents

We use Oceania/Asia/North America/Europe in subset Area as consequent (RHS) respectively. We create association rules of subset continents. Then we get 74868 rules.

And We summarize the output and count the frequency of created items of the rules. We get the table below.

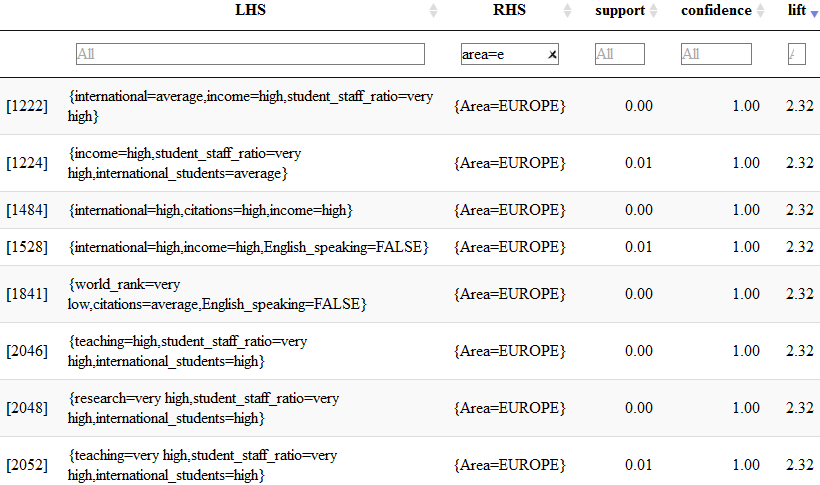
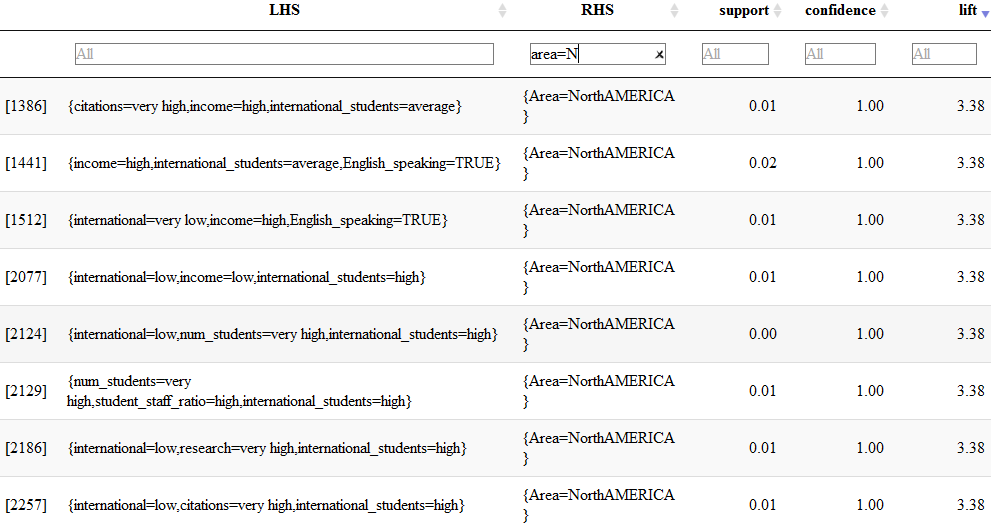
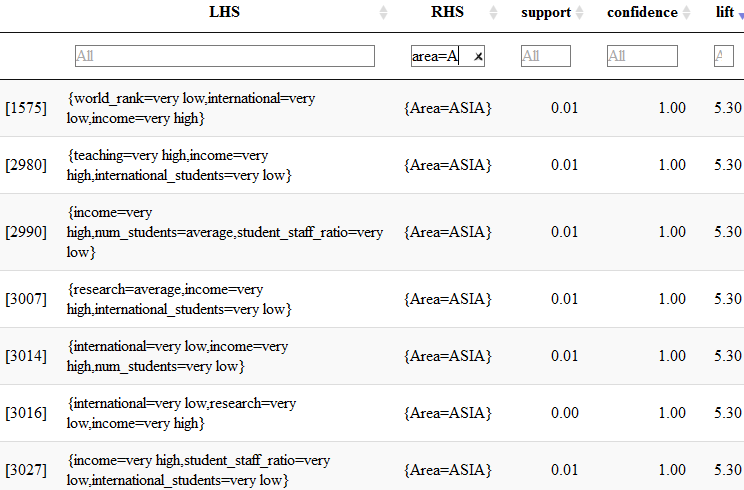
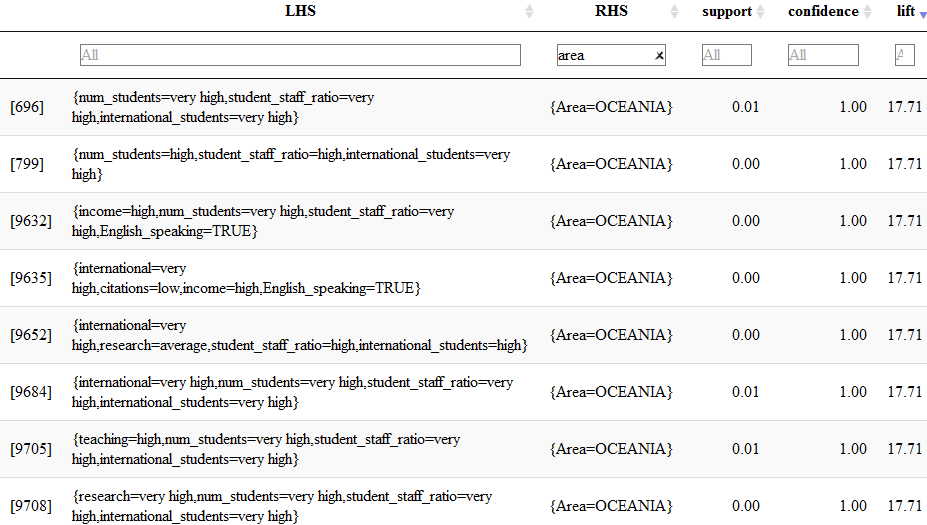


If student number=very high, student/staff=very high, international student=very high, speak English=True, then the country=Oceania.

If international=very low, income=very high, student/staff=very low, then the country=Asia.

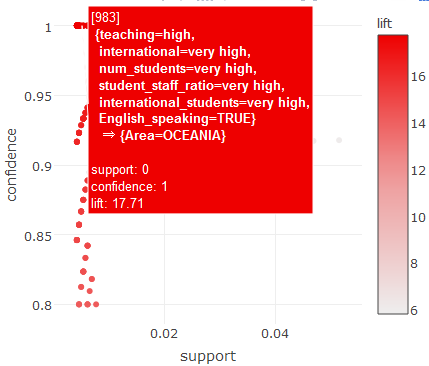
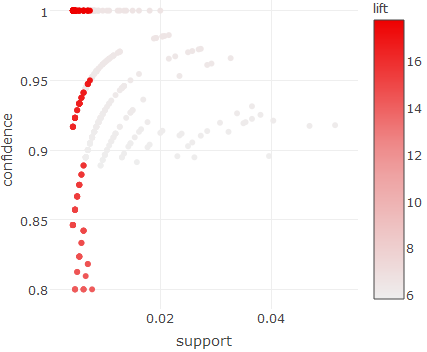
If citation=very high, speak English=true, international student=high, research=very high, then the country=N. America.

If student/staff=very high, income=high, international student=average, then the country=Europe.



Then we visualize the rules. And choose one of the dark red with confidence equal to 1.

We find that if teaching=very high, student/staff=very high, international student=very high, speak English=True, number of student is very high, then the country=Oceania.



## 3.3 Subset-China

We use ranking/research/teaching in subset China as consequent (RHS) respectively.

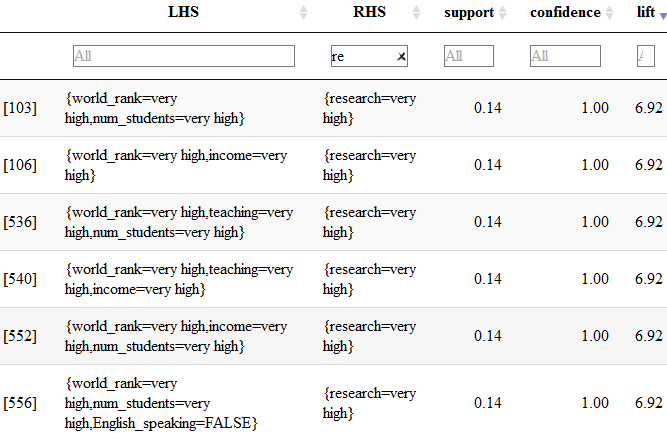
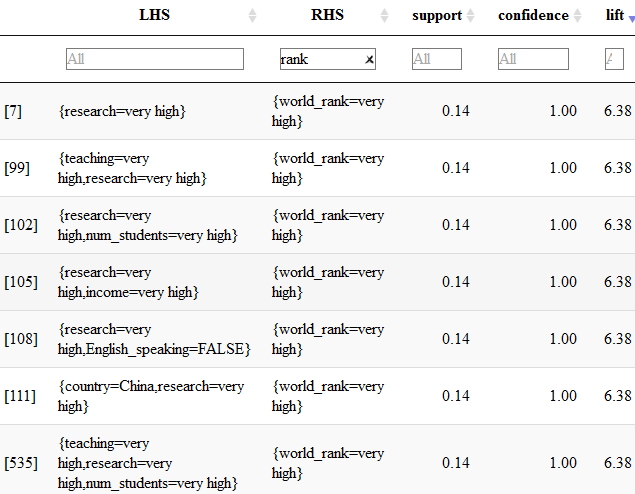
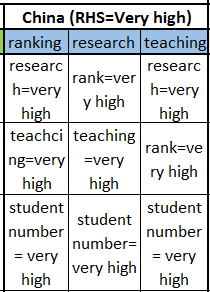
We create association rules of subset continents. Then we get 2540 rules. 

And We summarize the output and count the frequency of created items of the rules. We get the table below.

In China, if research=very high, teaching=very high, student number= very high, then the ranking=very high.

In China, if rank=very high, teaching=very high, student number= very high, then the research=very high.

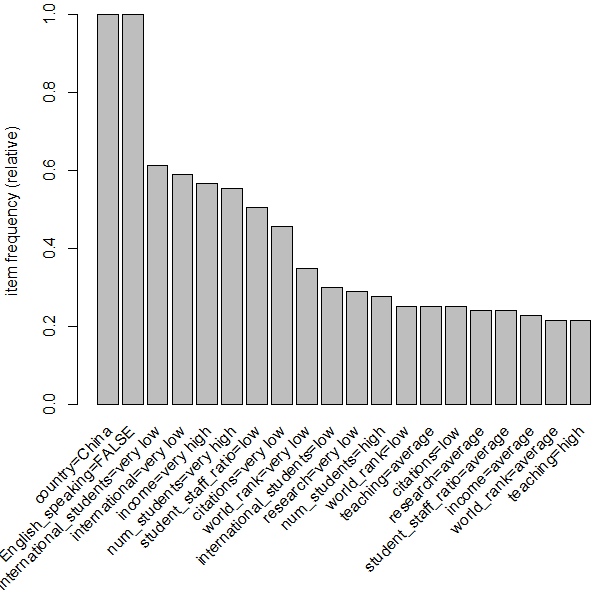
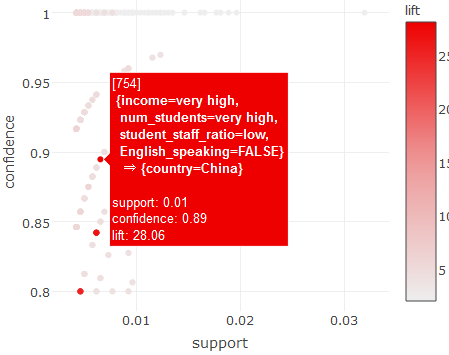
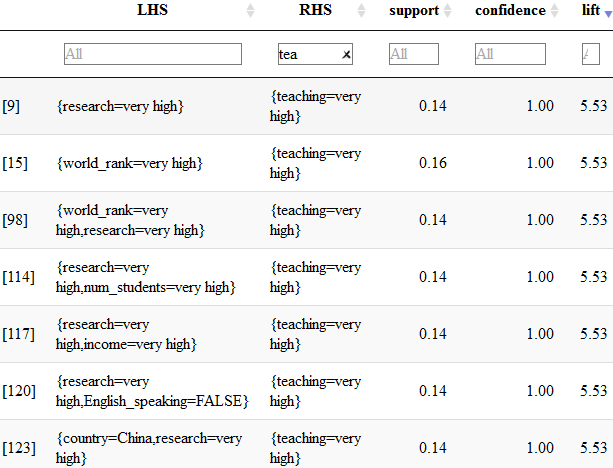
In China, if research=very high, rank=very high, student number= very high, then the teaching=very high.



Then we visualize the rules. And choose one of the dark red with confidence equal to 1.

We find that if income is very high, student/staff is very low, speak English is false, number of student is very high, then the country is China, which makes sense.

We can realize from the frequency plot below that international students is very low, income is very high, number of students is very high, students staff ratio is very low in China.



# 4. Modeling -- cwurdata

Second question of interest is to learn which factors have the greatest impact for university ranking improvement between 2014 and 2015.

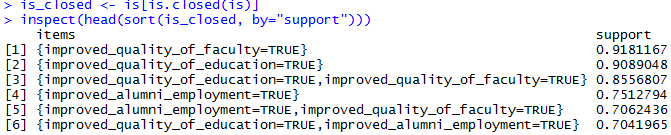
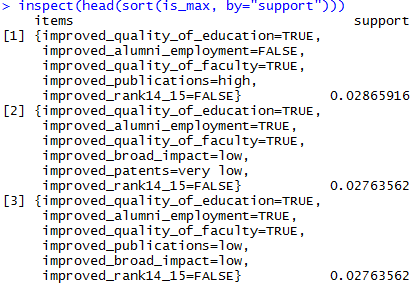
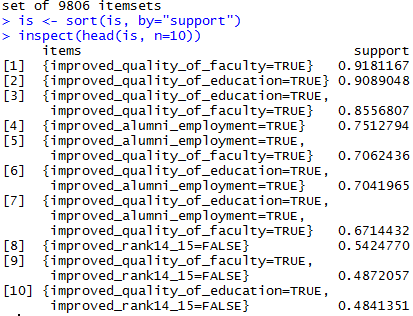
## 4.1 Create sets of frequent, closed and maximal itemsets

We create association rules of cwurdata.

For frequent dataset, there are 9806 itemsets.

The most support item is improved quality of faculty =True, improved education=True, improved alumni employment=True.

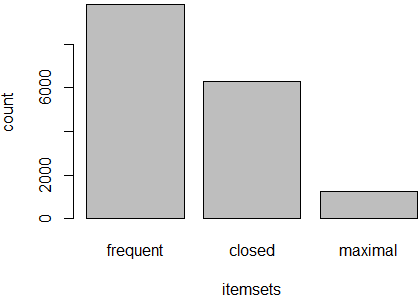
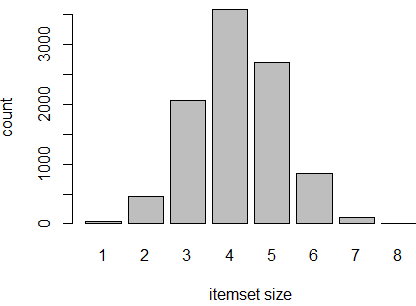
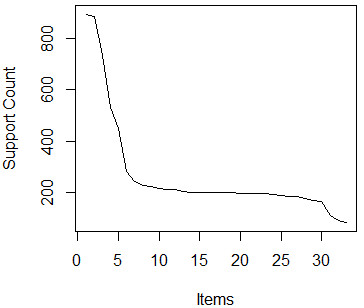
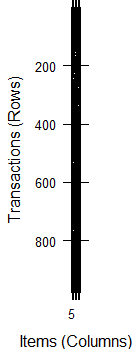
Closed dataset and maximal dataset are similar.



We know that the black place represents the university has the corresponding values while the white place represents not. Because we divide the country into 7 continents. So lots of blank disappeared. The curve below describes the trend of support.

From the barplot below, we can figure out that the majority rules contain 4 itemsets.

Also, barplot of frequent, closed and maximal shows our result is reasonable. Because frequent is the largest and maximal is the smallest dataset.



## 4.2 Association rules

Create sets of association rules. [5 points]

Use filtering, sorting, tables, and visualization to discuss the found patterns. What do the

patterns mean and how are they useful? [20 points]

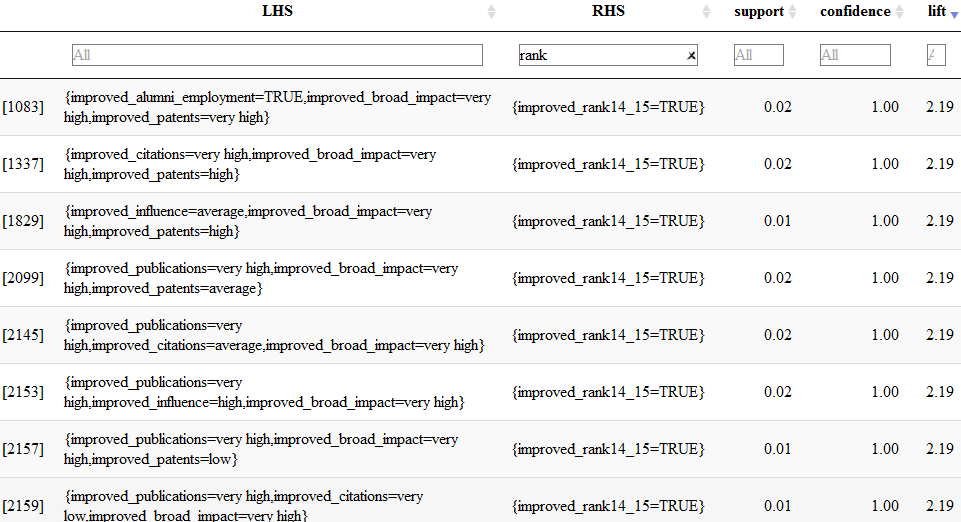
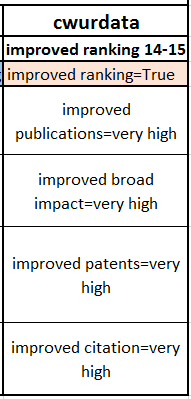
We use improved ranking between 2014 and 2015 in cwurdata as consequent (RHS). And we use other features to be antecedent (left-hand-side or LHS) during the process of mining rules.

We create 11854 association rules of cwurdata. 

Then we sort the lift and get the tables below.

And We summarize the output and count the frequency of created items of the rules. We get the table below.

If improved publications=very high, improved broad impact=very high, improved patents=very high, improved citation=very high, improved ranking=True.

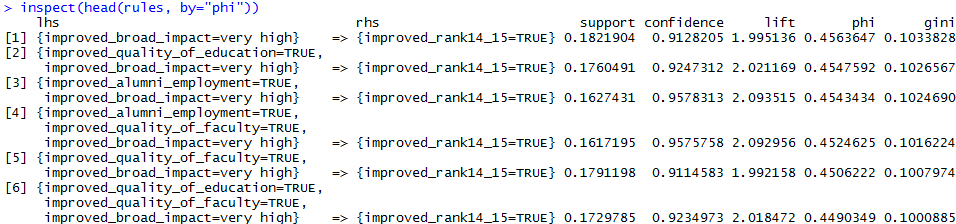
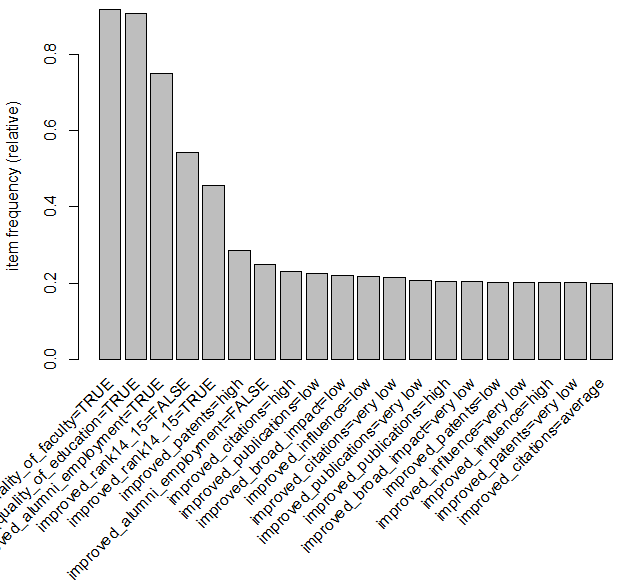


We can realize from the frequency plot below that the most frequent item is improved quality of faculty =True, improved education=True, improved alumni employment=True.

Next, we add phi and gini index to evaluate the association rules.

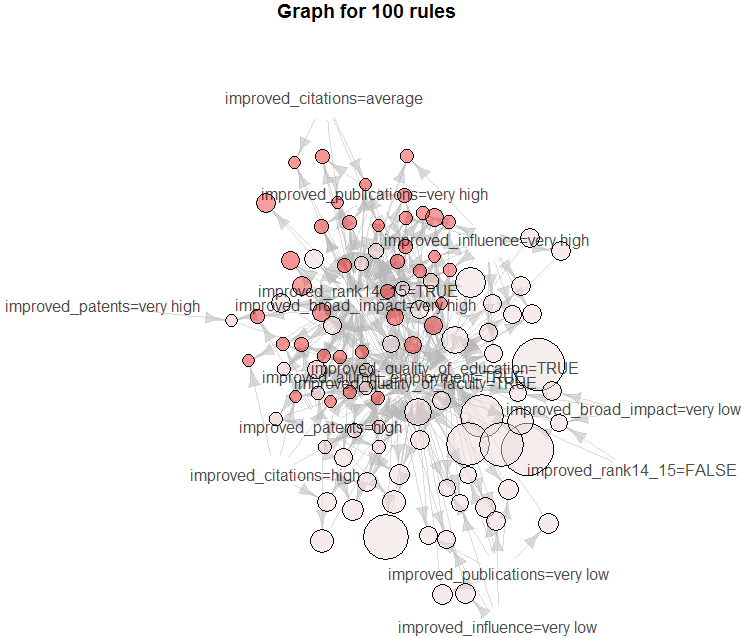
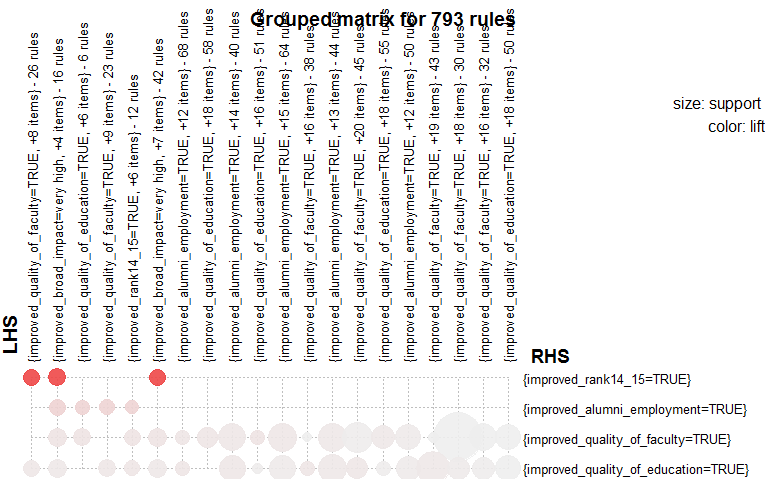
The phi coefficient is a measure of the degree of association between two binary variables.

So we sort the data by phi. It shows improved broad impact=very high, improved quality of education=True are correlated with improved rank 2014-2015=True.



Also, we could visualize the association rules that we created.

We can conclude that the relationship between lhs and rhs. For instance, improved broad impact=very high(16 rules, improved quality of faculty = True(26 rules), then improved rank 2014-2015=True.



# 5. Evaluation

## 5.1 Three major findings

Do they support what you have found in the other two projects? [10 points]

We summarize the output and count the frequency of created items of the rules. We get the table below. And there are 3 major things that I found.

First, I found that ranking was influenced by academic work, like research, citation and teaching.

Second, different regions have different university culture. For instance, the student and staff ratio of Asia is low while it is high in Oceania and Europe. The reason why it happened is the population structure. There are some developing countries in Asia, they want to develop the education. But there are too many students. So the staff resources are not enough. On the contrary, there are lots of educated and aging people in Oceania and Europe with few newborn. So, the students and staff ratio is low. Also, lots of good universities are in the US.

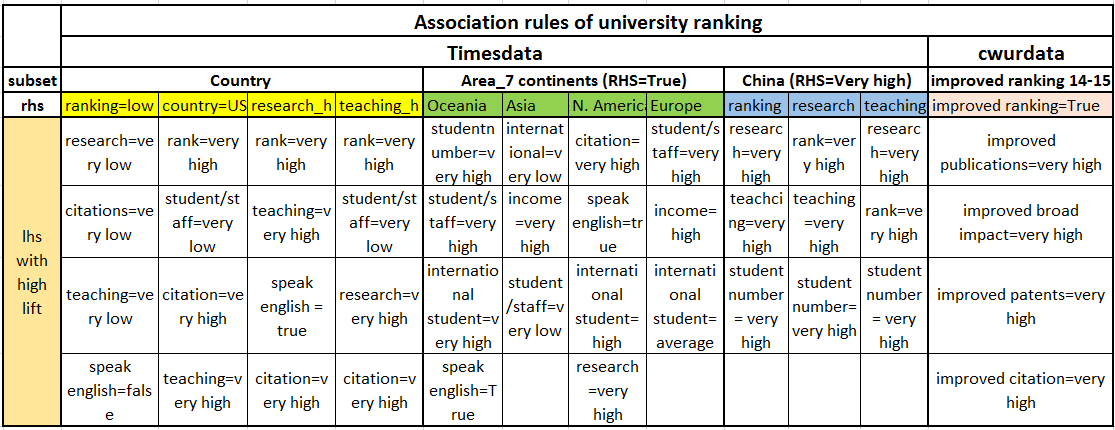
Third, improved publications, improved patents and improved broad impact are also very important if universities want to improve the rank next year beside academic.

My three findings support what I have found in the other 2 projects.

First, in the second project, I found that the most important features are Improved\_teaching and Improved\_research in timesdata. So What I have found that ranking was influenced by academic work, like research, citation and teaching in project 3 support the idea that most important features for improved ranking are Improved\_teaching and Improved\_research.

Second, I found that lots of good universities are in the US. In the first project, I found the same result.

Third, I found improved publications are one important feature for rank improvement in project3, which support what I have found Improved\_pub and award\_X2014 are the most important features in project2.



## 5.2 Recommendations and advantages

Findings:

If research=very low, citations=very low, teaching=very low, speak English=false, then the ranking=low.

If rank=very high, student/staff=very low, citation=very high, teaching=very high, then the country=US.

If rank=very high, teaching=very high, speak English = true, citation=very high, then the research=very high.

If rank=very high, student/staff=very low, research=very high, citation=very high, then the teaching=very high.

If student number=very high, student/staff=very high, international student=very high, speak English=True, then the country=Oceania.

If international=very low, income=very high, student/staff=very low, then the country=Asia.

If citation=very high, speak English=true, international student=high, research=very high, then the country=N. America.

If student/staff=very high, income=high, international student=average, then the country=Europe.

In China, if research=very high, teaching=very high, student number= very high, then the ranking=very high.

In China, if rank=very high, teaching=very high, student number= very high, then the research=very high.

In China, if research=very high, rank=very high, student number= very high, then the teaching=very high.

If improved publications=very high, improved broad impact=very high, improved patents=very high, improved citation=very high, improved ranking=True.

If the administrator of an university wants to improve the rank next year. Academic work is the most important thing that they should consider. Academic work includes research, citation, teaching. Academic work should get more funds.

For students and parents, if they want to study in a good university. Region is the most important thing that they should consider. Because different areas have different university culture. For example, there are lots of students in Asia. If you want to communicate with students more, maybe you can think about study in Asia. While in Europe, there are lots of staff. If you prefer communicate with professors more, maybe you could study in Europe.

Also, the results are useful. For students and parents, they can choose a well-matched university for them. It is a big thing, it may affect their life. For company and administration of universities, the association rules are straightforward and easy to understand. Also, the visualization is beautiful. They can use it to do the strategy for next year.

# 6. Acknowledgment

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

# 7. References

Source:

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>

Image of university ranking

1. https://www.timeshighereducation.com/world-university-rankings/2016/world-ranking#!/page/0/length/25/sort\_by/rank\_label/sort\_order/asc/cols/rank\_only