FINAL PROJECT

World Universities Rankings

2017 Analysis

SINDHUJA RAMINENI -108674023

KALYANI MEKA – 108516382

Table of Contents

[**EXECUTIVE SUMMARY:** 3](#_Toc532231145)

[**INTRODUCTION:** 3](#_Toc532231146)

[**DESCRIPTIVE STATISTICS:** 5](#_Toc532231147)

[**DATA ANALYSIS:** 6](#_Toc532231148)

[**EXPLARATORY DATA ANALYSIS:** 6](#_Toc532231149)

[**HISTOGRAMS:** 7](#_Toc532231150)

[**Observations:** 7](#_Toc532231151)

[**Further Analysis:** 8](#_Toc532231152)

[**CORRELATION VALUES:** 9](#_Toc532231153)

[**Missing Data Analysis:** 10](#_Toc532231154)

[**MODEL BUILDING:** 10](#_Toc532231155)

[**Model not selected:** 10](#_Toc532231156)

[**BEST MODEL:** 11](#_Toc532231157)

[**Graphs associated with the model:** 11](#_Toc532231158)

[**Errors independent of each other:** 12](#_Toc532231159)

[**Durbin-Watson test:** 12](#_Toc532231160)

[**STEP-WISE REGRESSION MODEL:** 13](#_Toc532231161)

[CONFIDENCE INTERVALS: 14](#_Toc532231162)

[ANOVA Table: 15](#_Toc532231163)

[AIC: 15](#_Toc532231164)

[**Cross Validation:** 16](#_Toc532231165)

[**PRINCIPLE COMPONENTS ANALYSIS:** 17](#_Toc532231166)

[**FACTOR ANALYSIS:** 18](#_Toc532231167)

[**Predictions for Ranking:** 19](#_Toc532231168)

[**Time-Series Forecasting:** 19](#_Toc532231169)

# **EXECUTIVE SUMMARY:**

This narrative tells us about the world-wide university rankings and how the students will be choosing the best universities based on the teaching, citations, research, income, international students, inter outlook range, total score, number of students, female to male ratio. On the overall analysis there have been so many problems which arose in selecting the universities for students to pursue their higher studies. Like for few of the universities which are top, are almost high in having all the above-mentioned criteria. Apart from that there are few universities whose rank is not that higher in terms of rank, but they are having good percentage of citations rating, teaching, research, inter outlook range and total score which are to be concentrated. We even separately categorized the universities present in Colorado. We took only the top 400 world wide ranked colleges in 2017 year and made these analyses.

# **INTRODUCTION:**

The dataset considered for the project is World University Rankings by Times Higher Education (THE) as it is one of the most influential international university ranking systems. This dataset contains 400 rankings of universities for the year 2017. Since the dataset also contains rankings with groups, we consider the first 200 rankings for the analysis. The source has been obtained from timeshighereducation.com. The complete dataset can be found at : <https://www.timeshighereducation.com/world-university-rankings/2016/world->[ranking#!/page/0/length/25/sort\_by/rank/sort\_order/asc/cols/stats](https://www.timeshighereducation.com/world-university-rankings/2016/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats)

We took few of the data from www.kaggle.com

The dataset contains two types of variables: quantitative and categorical.

The categorical variables in our data are:

|  |  |  |
| --- | --- | --- |
| University\_Name | Name Of University | String |
|  |  |  |
| Country | Country Of Each University | String |
|  |  |  |
| International\_Students | Percentage Of Students Who Are International | String |
|  |  |  |
| Female\_Male\_Ratio | Female Student To Male Student Ratio | String |
|  |  |  |

Similarly, the quantitative variables are:

|  |  |  |
| --- | --- | --- |
| **Column** | **Metadata** | **Datatype** |
|  |  |  |
| world\_rank | world rank for the university. Contains rank ranges and | Numeric |
|  | equal ranks (eg. =94 and 201-250) |  |
|  |  |  |
| Teaching | university score for teaching (the learning environment) | Numeric |
|  |  |  |
| international | university score international outlook (staff, students, | Numeric |
|  | research) |  |
|  |  |  |
| research | university score for research (volume, income and | Numeric |
|  | reputation) |  |
|  |  |  |
| citations | university score for citations (research influence) | Numeric |
|  |  |  |
| income | university score for industry income (knowledge transfer) | Numeric |
|  |  |  |
| total\_score | total score for university, used to determine rank | Numeric |
|  |  |  |
| num\_students | number of students at the university | Numeric |
|  |  |  |
| student\_staff\_ratio | Number of students divided by number of staff | Numeric |
|  |  |  |

In our analysis with regression, we try to answer the questions regarding the variables influencing the rank of each university in the world, other criteria. We train the first 150 records for the regression analysis and predicted the ranks for the remaining 50 records.

# **DESCRIPTIVE STATISTICS:**

Before doing the programming and building the model, we made the descriptive statistics to know how the data is and what all things we need to do in order to improve the model. We found that the data is good enough as the mean=median=mode and the kurtosis and skewness fall in their acceptable limit range.

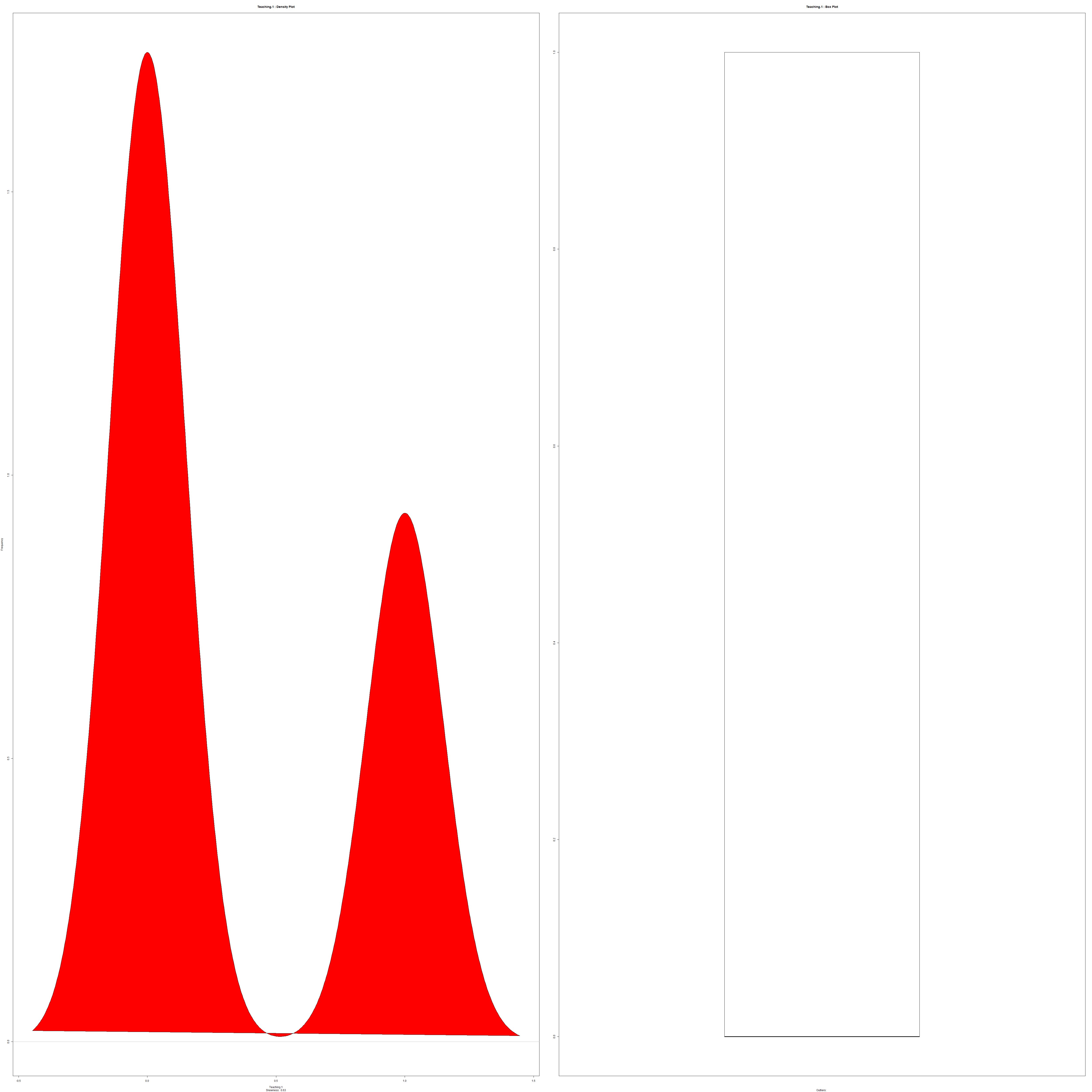
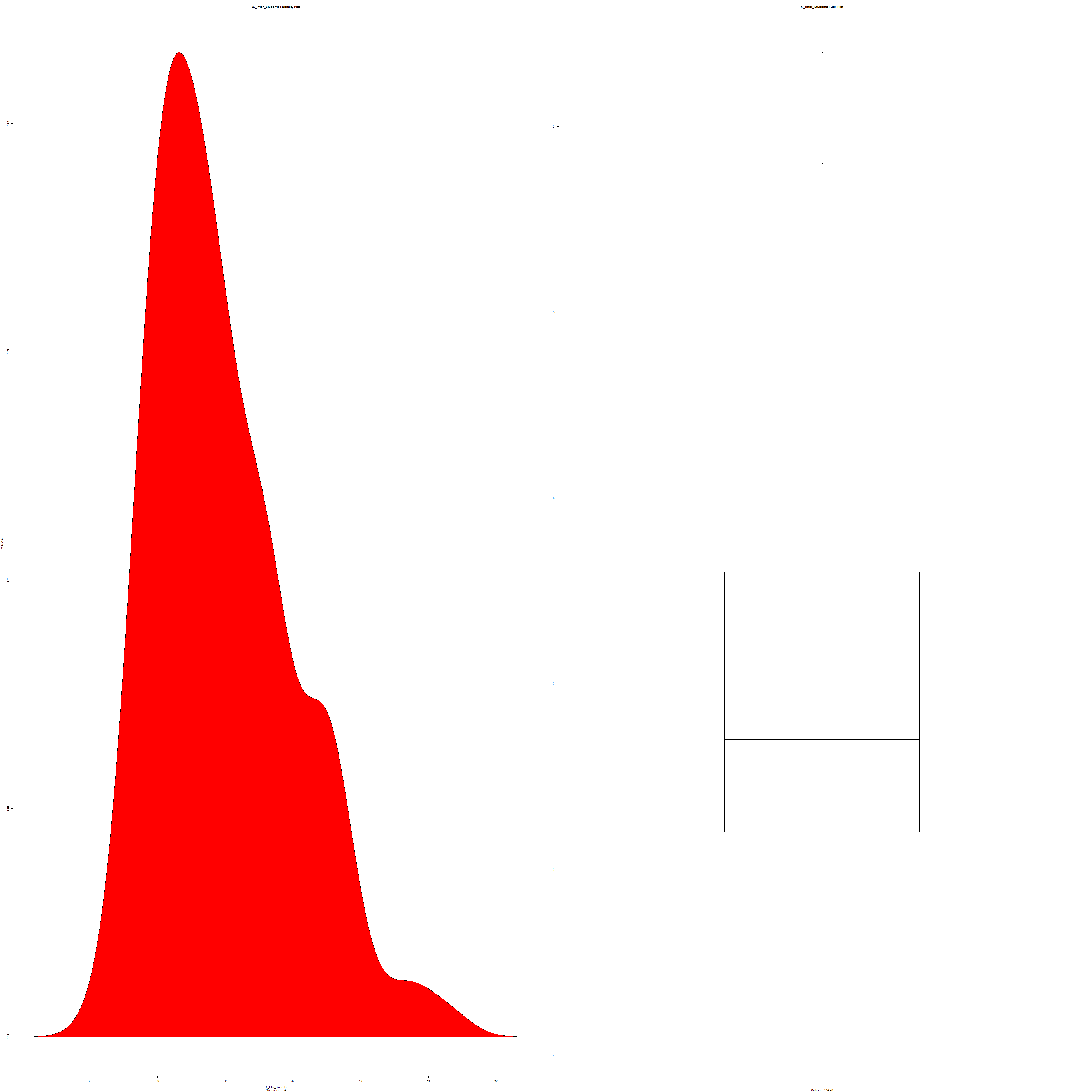
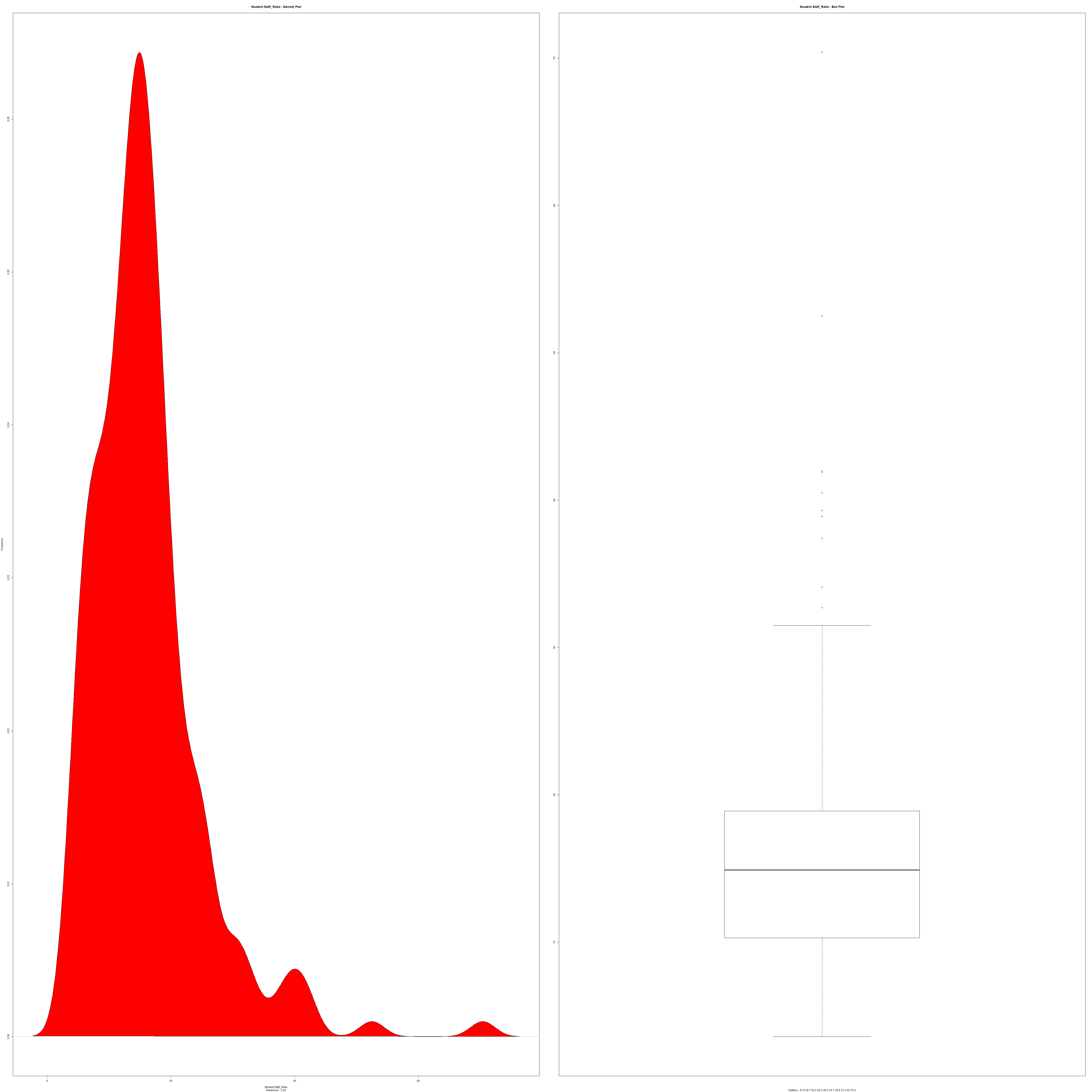
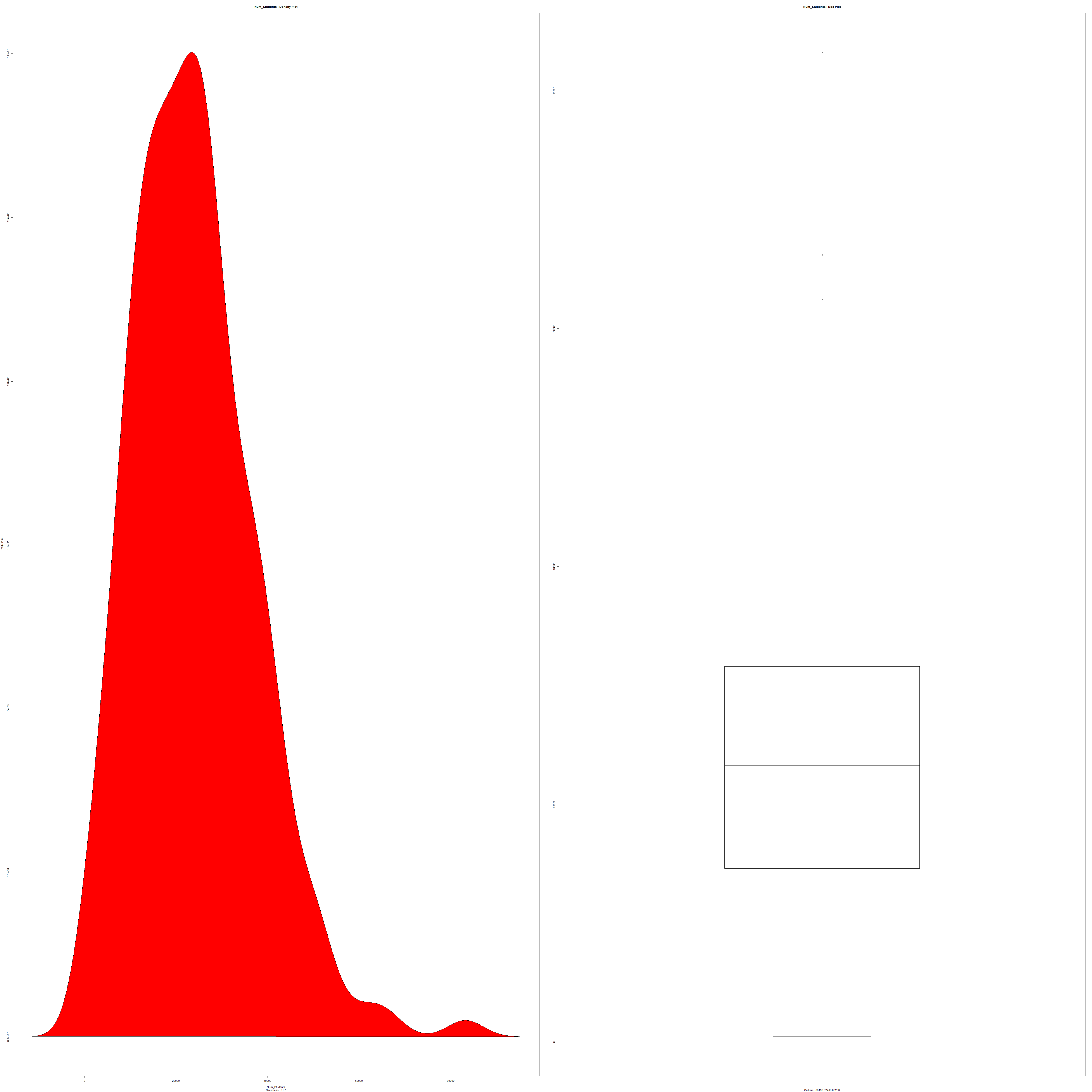
|  |  |
| --- | --- |
| World\_Rank | |
|  |  |
| Mean | 100.145 |
| Standard Error | 4.081521 |
| Median | 99.5 |
| Mode | 103 |
| Standard Deviation | 57.72142 |
| Sample Variance | 3331.763 |
| Kurtosis | -1.19659 |
| Skewness | 0.003135 |
| Range | 199 |
| Minimum | 1 |
| Maximum | 200 |
| Sum | 20029 |
| Count | 200 |

# **DATA ANALYSIS:**

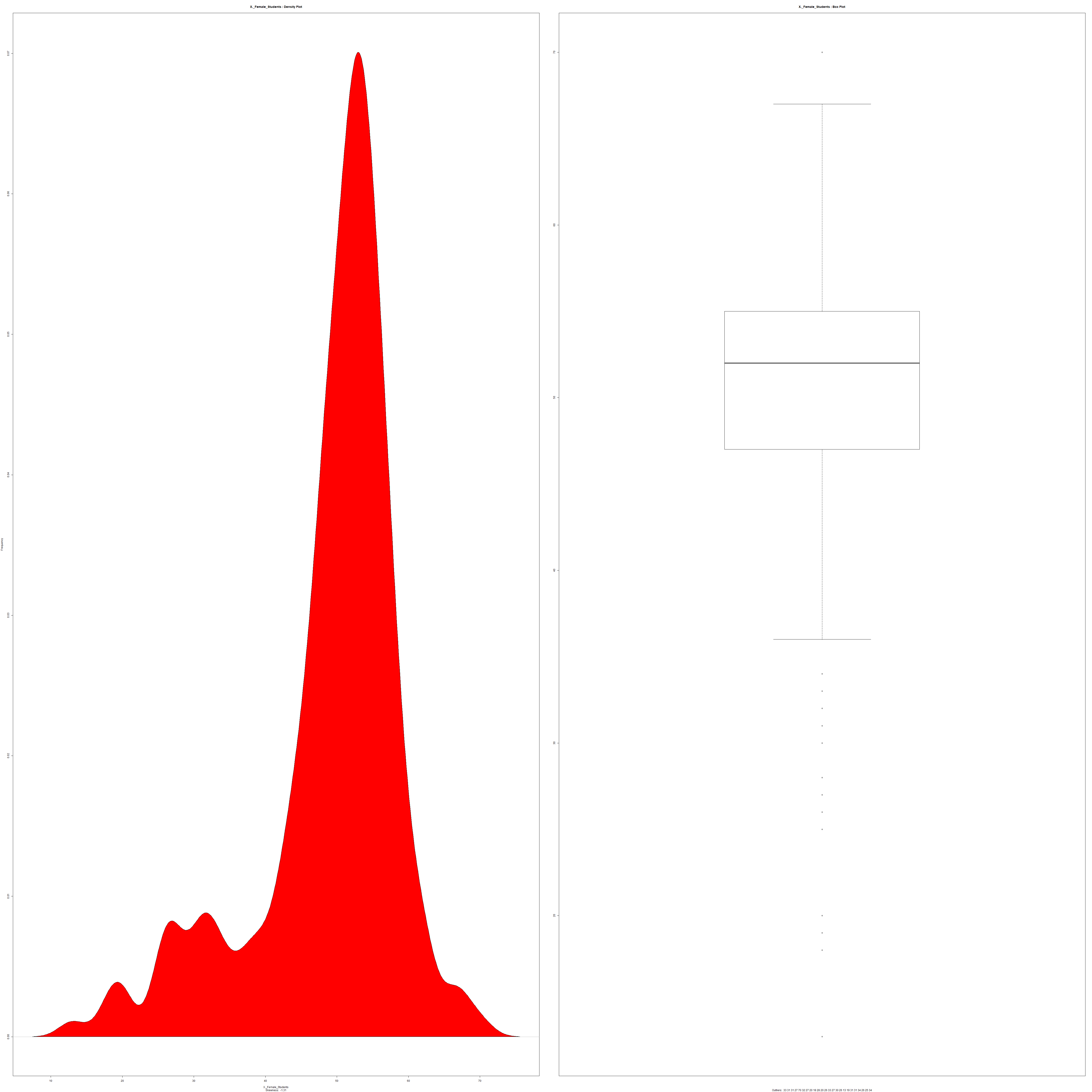
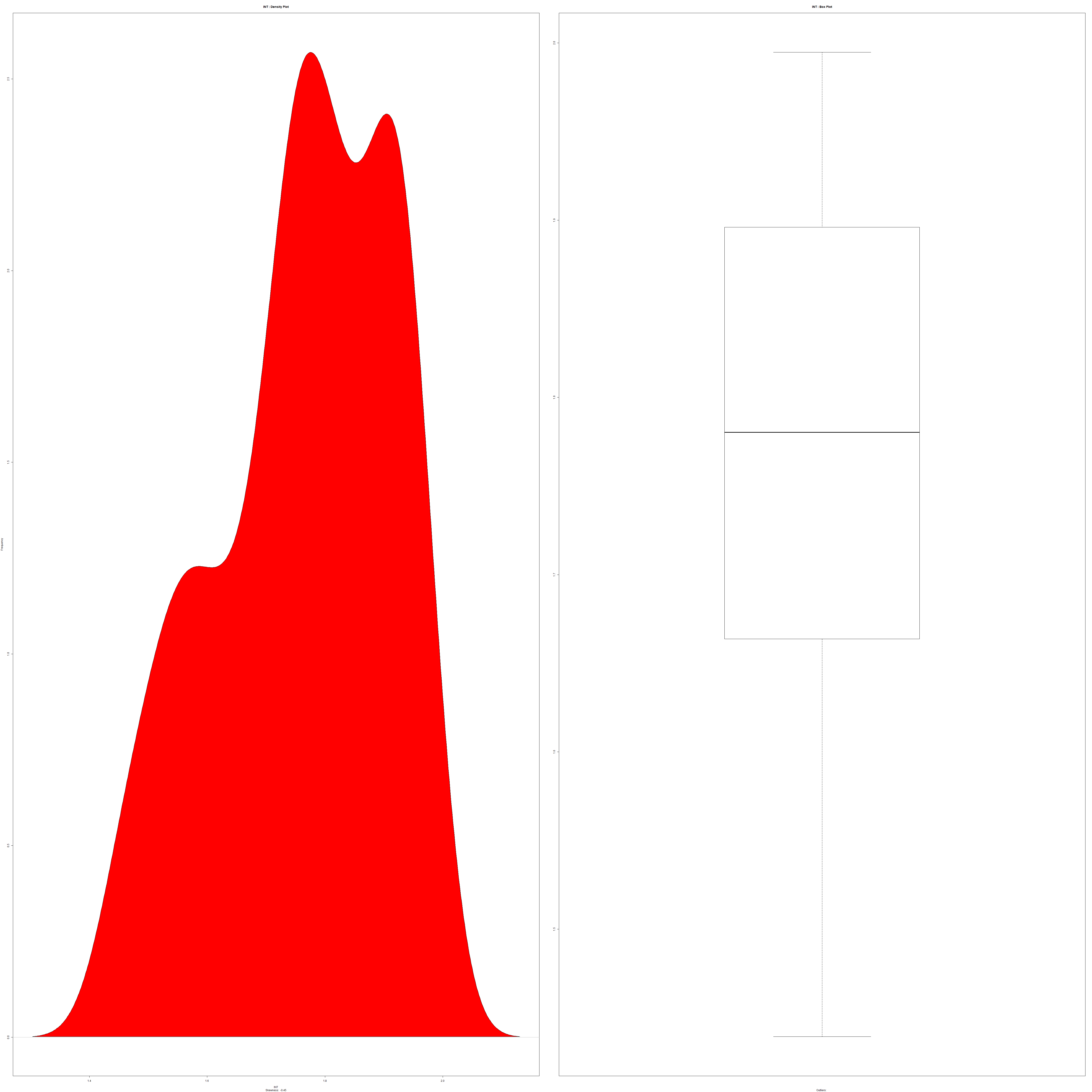
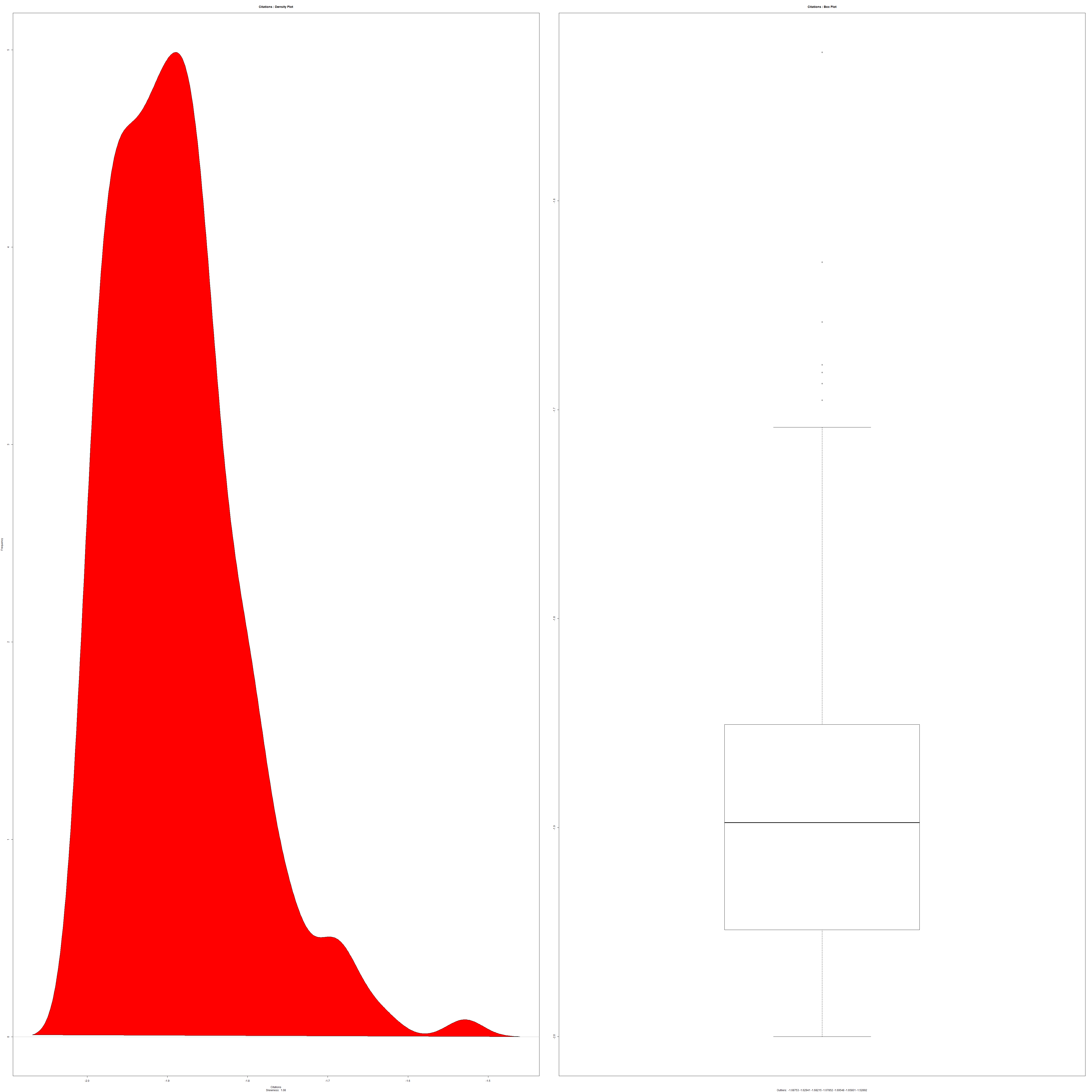
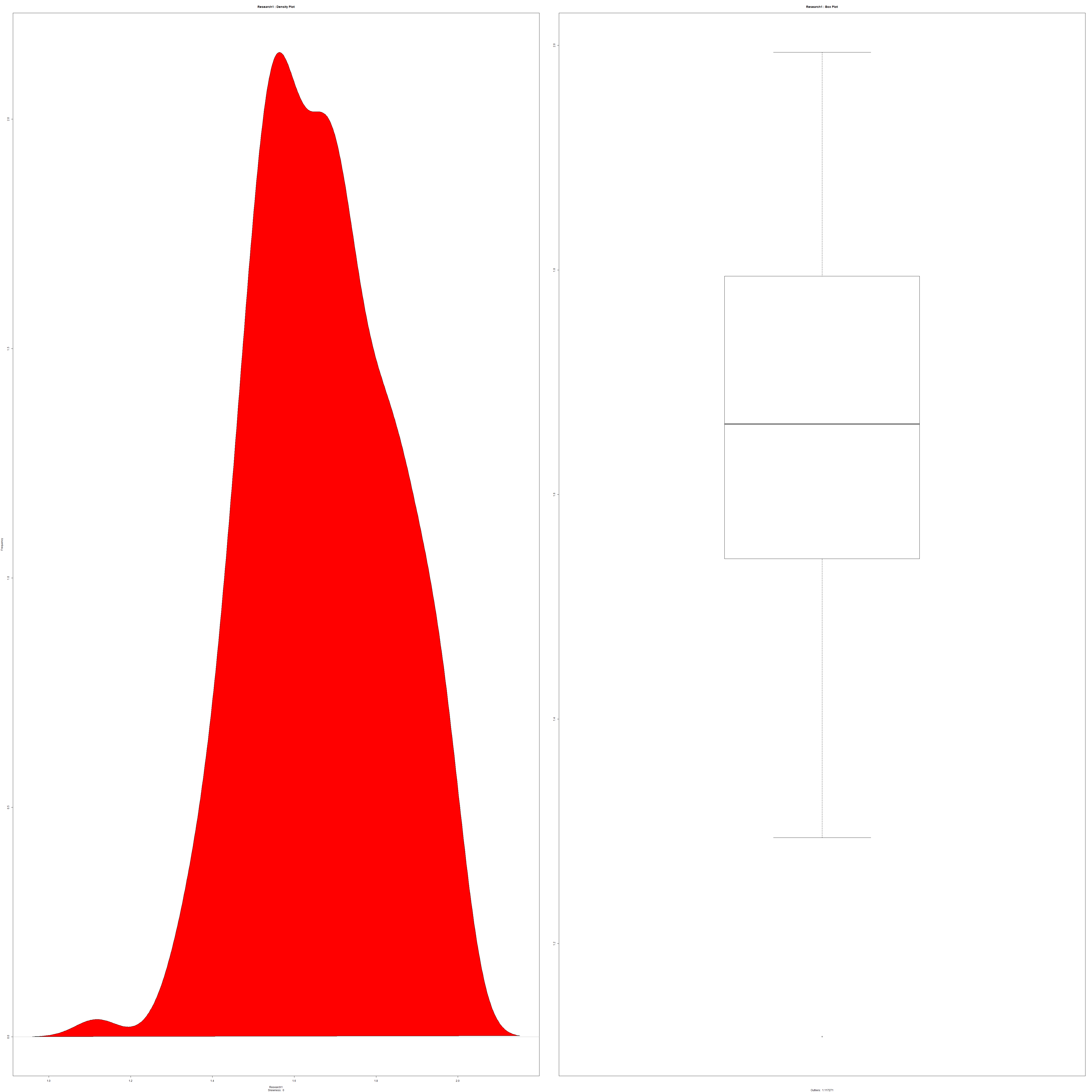
We tried several permutations and combinations to get the best model. We considered the top 200 world-wide ranks. And we splitted the into 2 halves and we cross validated the results. With one half we build the model and with the other half we validated the model. Data1 is used for the building the model and the analysis data2 is used to cross validate the results. We did several combinations for the tables. Like for Research, Citations and Inter\_Outlook tables we took logarithmic values. Used dummy variables for Teaching and % of female students’ tables for more than 50 – 1 and for less than 50 – 0. We made all these changes to keep the Skewness and Kurtosis in their ranges from -1 to 1 and -2 to 2.

# **EXPLARATORY DATA ANALYSIS:**

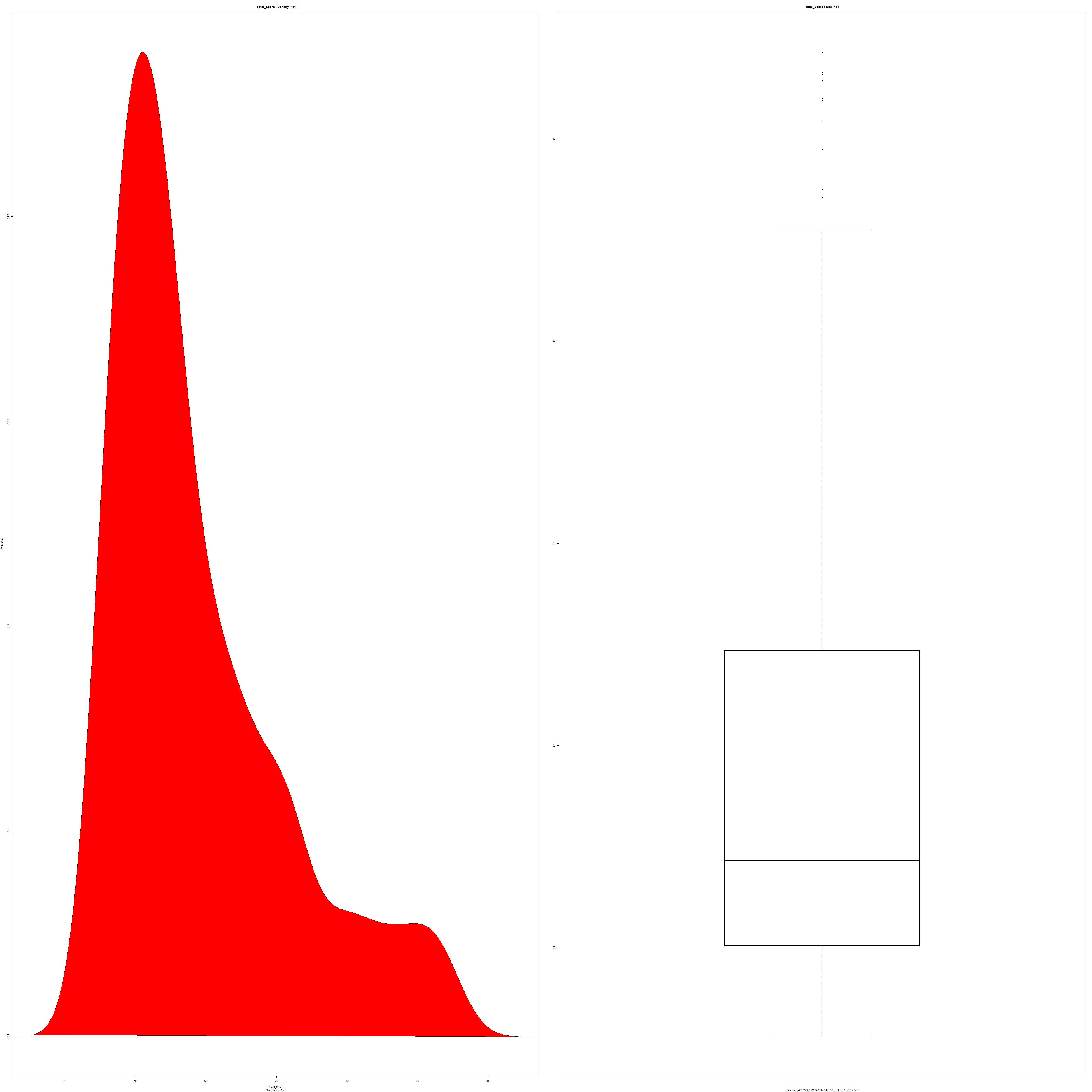
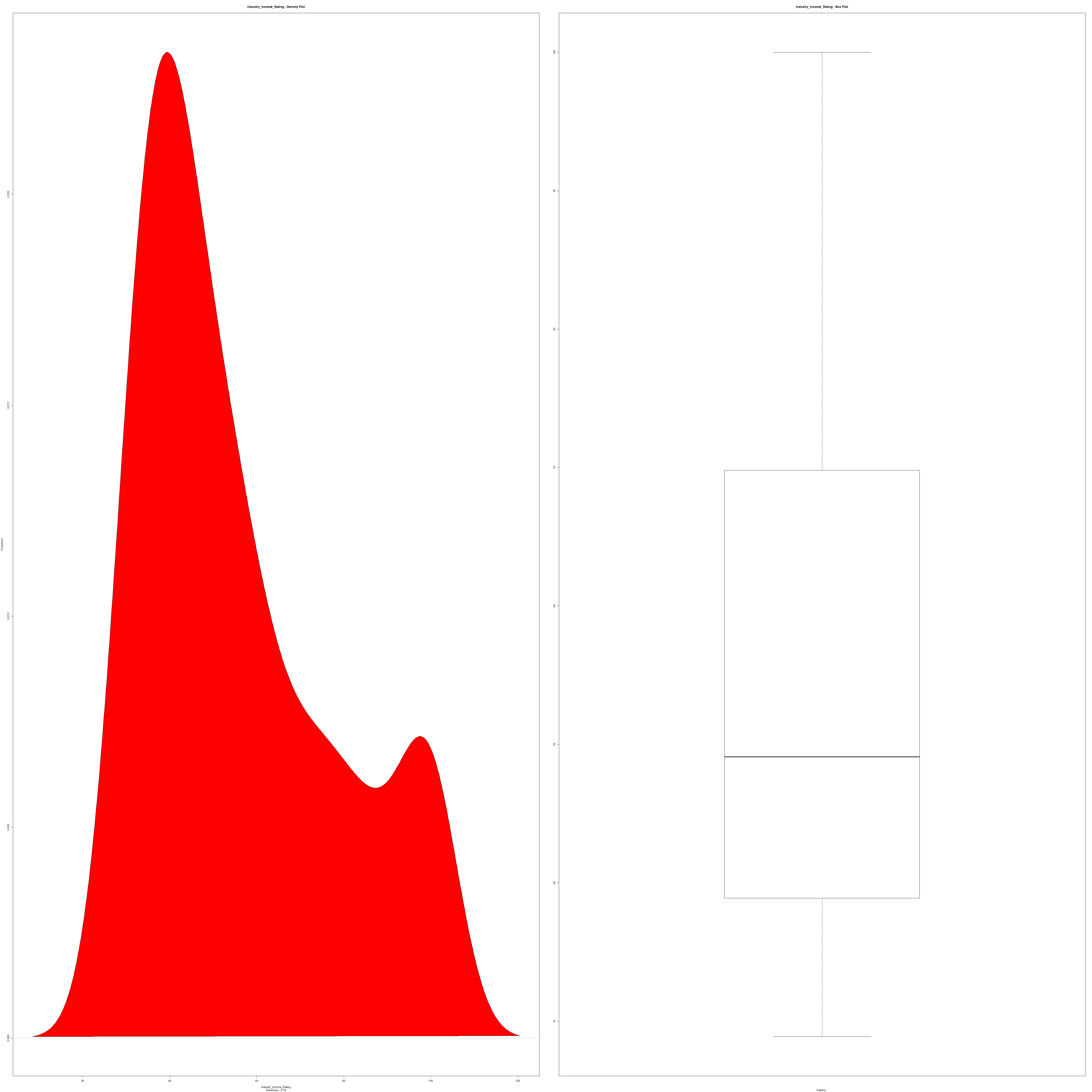
For the exploratory data analysis, we created density and box plots for the independent variables to check which of them can go into the best model. We got the below boxplots and density plots.



Num\_students Student\_staff ratio Inter students Teaching 1

Female students Inter\_outlook range Research 1 Citations



Industry\_income rating Total\_score

Among all these plots teaching1 – which is the dummy variables of teaching is having the exact bell-curve. So, in the model that should be definitely considered and after that num\_students are also having the good bell curve. So, these must be considered. Outliers are also minimum in these 2 plots.

## **HISTOGRAMS:**

The above histogram has been plotted to determine which of the world countries have the greatest number of universities in the top 200 universities. We clearly found that the United states is having the maximum number of universities in the top 200 universities ranking, followed by United Kingdom, Germany and Australia. So, this is the main reason most people come to United states to pursue their higher studies. Here there are several universities with same ranking. In the excel sheet it is categorized as the 100-150 and around 70 colleges are given. So, the count differs. This is because they have given the categorization based on the groups as well in few fields.

### **Observations:**

We separately plotted the top 15 colleges and bottom 10 colleges and found that there are few cases where we can see that the below 10 universities have greater values when compared to the top 15 universities. So, this is one more criterion, that needs to be concentrated along with the ranks in the selection as in the top universities as well all the independent variable is not high in number, few colleges are having few variables like teaching, income, citations and few are having research, number of students etc. So, the total score doesn’t depend only on a single variable but depends on the combination of variables.

TOP 15 UNIVERSITIES:

BOTTOM 10 UNIVERSITIES:

In the top 15 universities though California Institute of technology is world rank 1 college, but it is not having the top citations rating, when we see the Massachusetts Institute of Technology is having the 100% citations rating. In the bottom 10 colleges if we see Florida institute of technology is having the good citations rating. So, if the student is not having that good score to meet the ranks of the top-rated colleges, they may even join in the best colleges depending upon their scores by seeing the factors they need. Suppose if the student wants to continue his/her higher studies by doing PHD or research then FIT is a good option though its rank is around 200.

### **Further Analysis:**

For portraying the further analysis and as we are living in Denver, we took Colorado universities whose ranks are below 400.

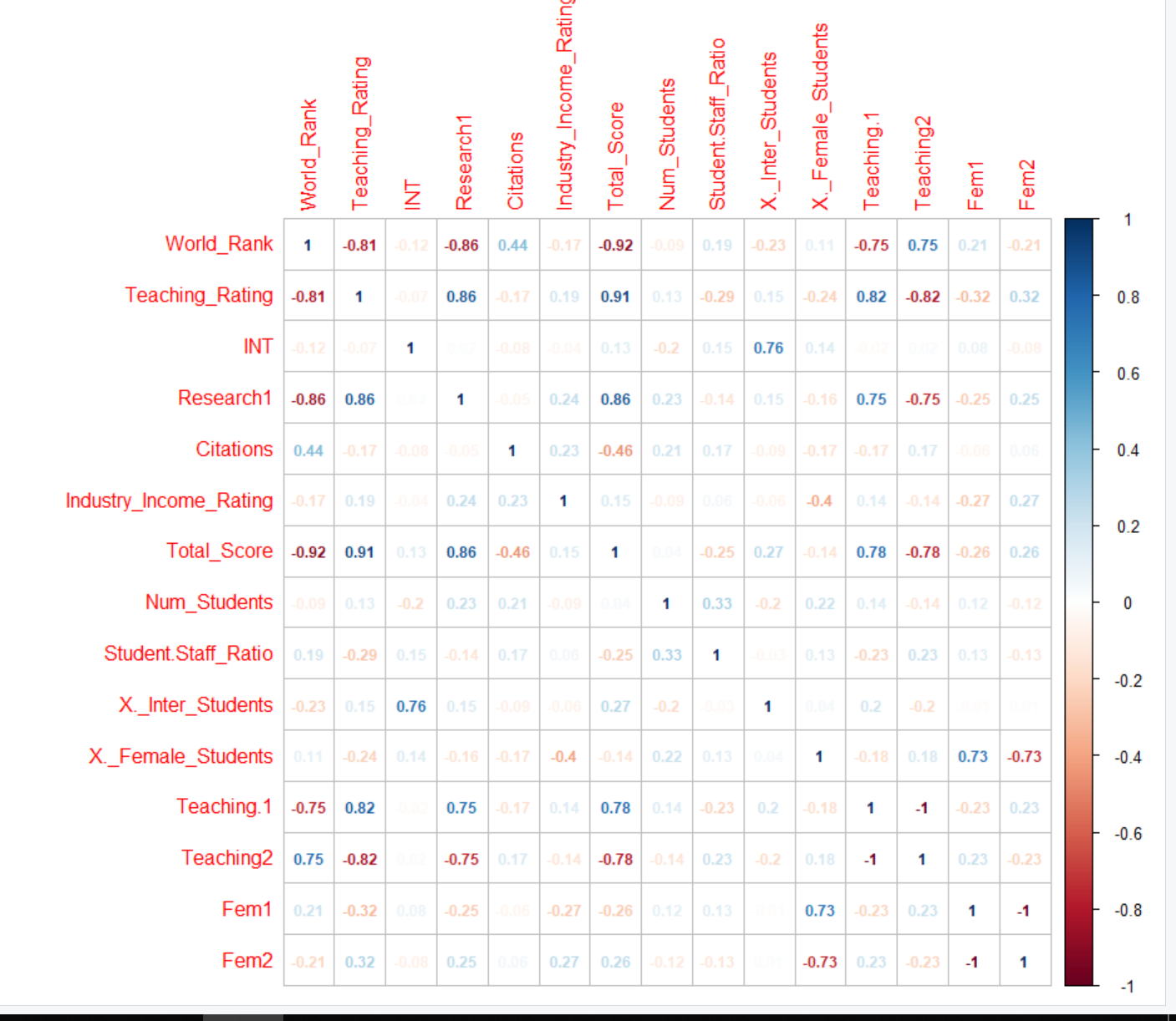
Here we can see that though University of Colorado Boulder is the top college in Colorado with world rank of 97, it is having many least factors when compared to the other universities in Colorado. If we look into the graph, we can see that Colorado School of mines whose world rank is 141 is having the top Inter\_outlook rating and Industry\_income\_rating. Teaching rating is the highest for University of Denver whose rank is around 300-350. UC Boulder is having the top citation rating with 97.4 but all the remaining universities like University of Colorado Denver, Colorado School of mines are also having the same rating. So, based on these factors we can conclude that Top ranking universities doesn’t mean that they have all the factors in greater amounts compared to other colleges. It is up to the students to investigate the characteristics and come in to the conclusion for their best desired college.

The students should design a plan to choose the university depending upon the Country, State, Area, International students count, female students count, teaching and other factors apart from the rank. There are few of the external factors that are to be considered like crime rate in that area, cost and the income that should be paid for that particular university which is selected and the area in which there are good number of companies etc. Some other factors like cost of living in that place should also be considered. This analysis is made to demonstrate our hep to others.

# **CORRELATION VALUES:**

To build the model, highly correlated values cannot go into the model. Only either one of them should be kept such that higher percentage of reduced R square is attained. From this correlation we can see that,

* **international\_students** are highly correlated to **international outlook**.
* The highest correlation is between **teaching** and **research**.
* **World\_rank** and **Total\_score** is each highly correlated to Teaching, research andcitations.



## **Missing Data Analysis:**

Initially when we got the data, there were several rows and columns with missing values. As the data is the open source, we went to different sites to get the missed data. And later all the missing places are fulfilled.

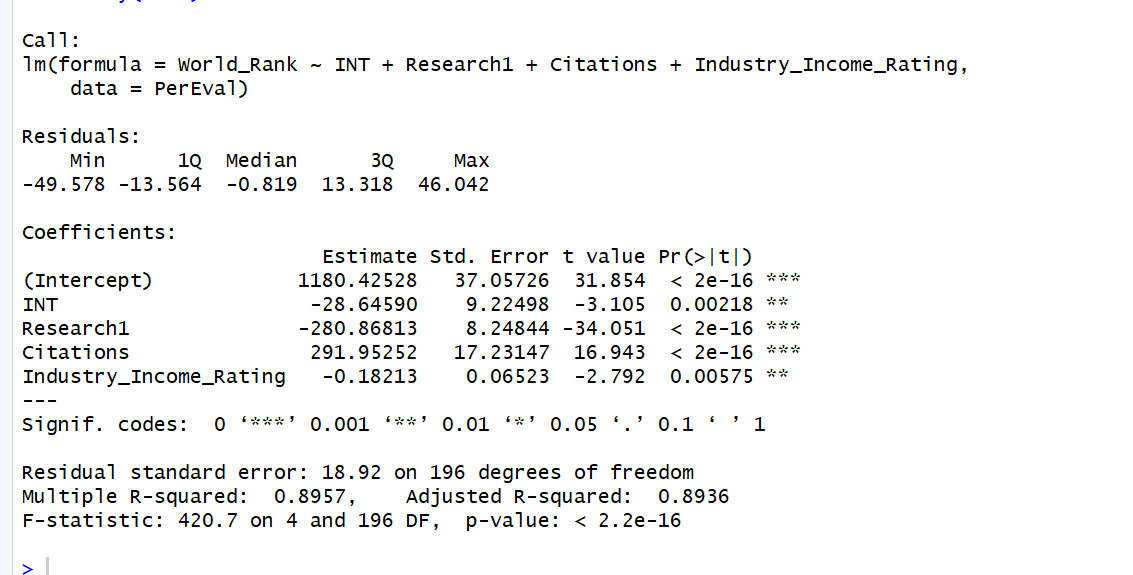
# **MODEL BUILDING:**

**Model not selected:** This is the model that we took with the factors present without any moderations. Only for research factor dummy variables are assigned. Greater than 50 is taken as 1 and less than 50 as 0. And for international outlook logarithmic values are given.

So, the multiple R square that we got is 89.5% and adjusted R square is 89.36%. We still wanted to increase the percentage.

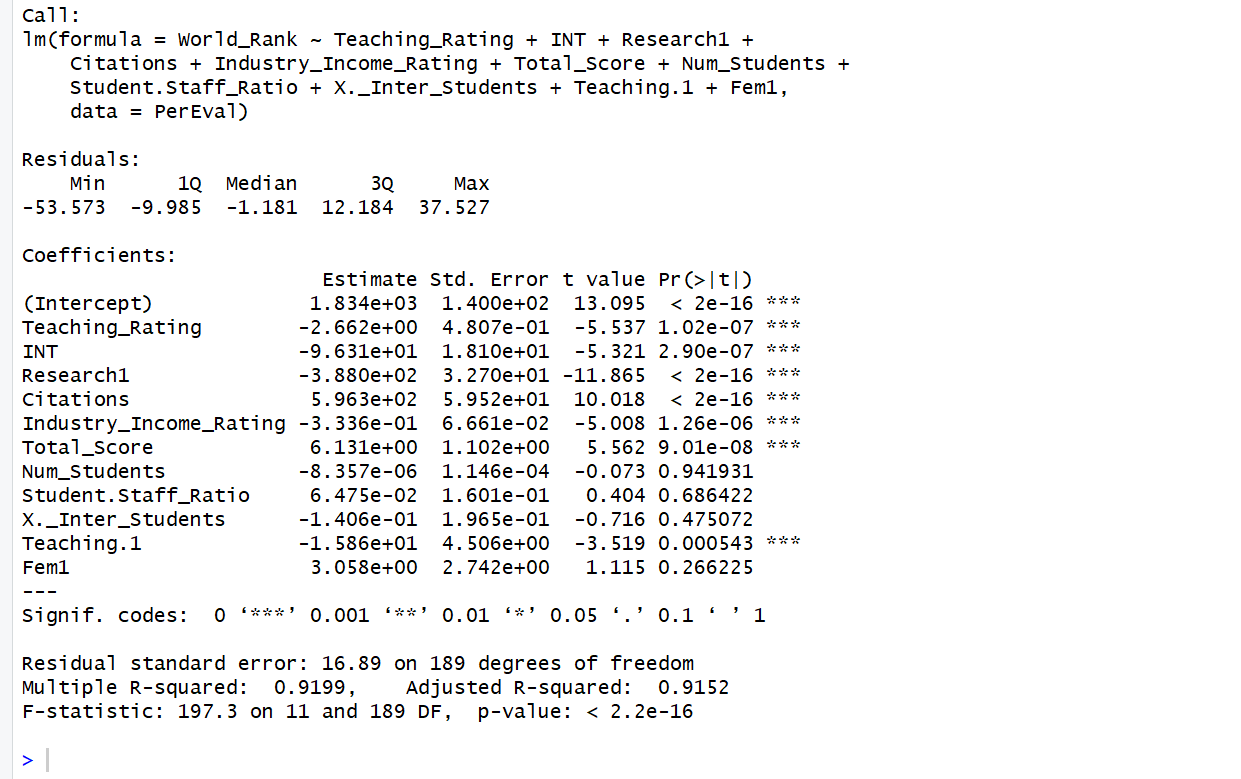
Adjusted R square shows the proportion of variation in Y explained by all X variables adjusted for the number of Xvariables used.

By using the dummy variables, it assumes the slopes associated with numerical independent variables do not change with the value for the categorical variable.



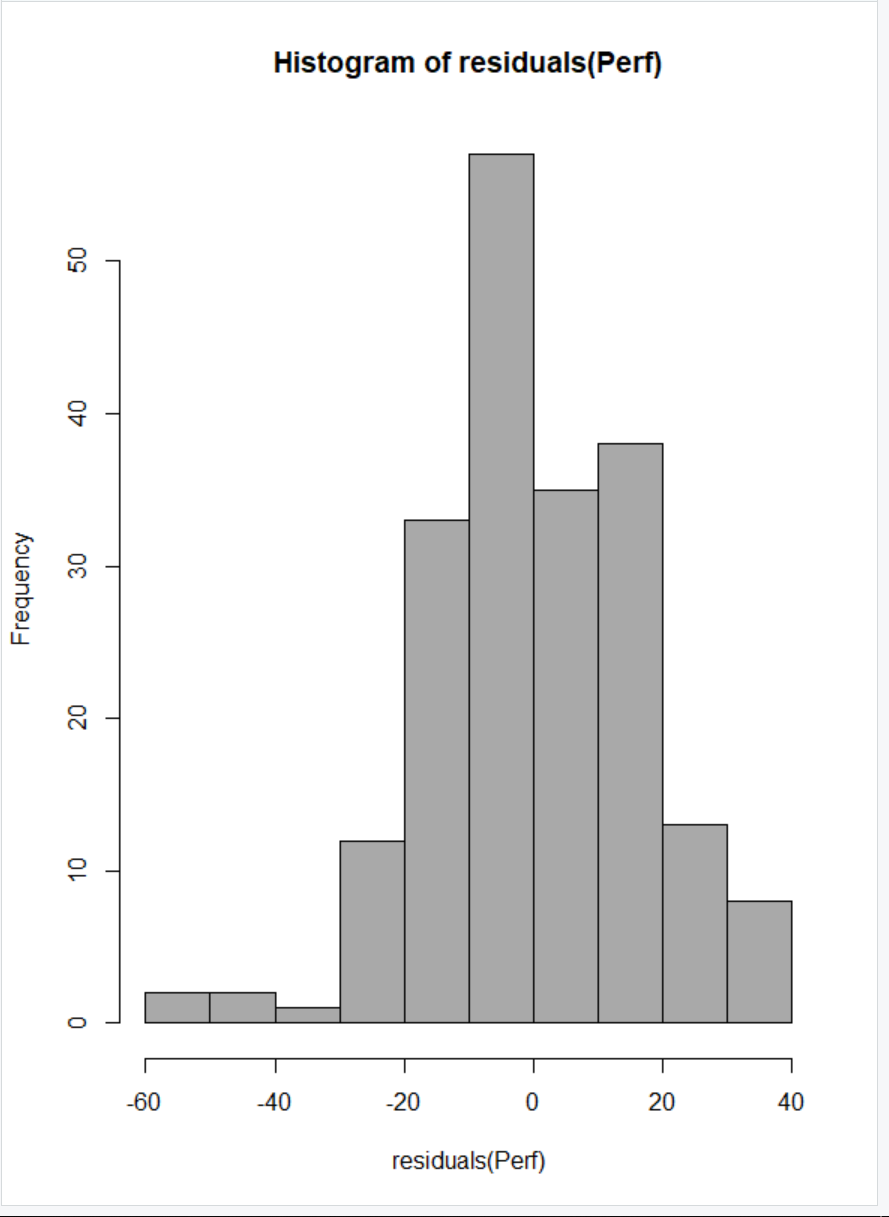
## **BEST MODEL:**

We still improved the model by taking dummy variables for teaching and for female to male ratio by taking the dummy variables. For the international students’ logarithmic values are taken. By this we got multiple R square as 92% and adjusted R -square of 91.5%. And we can see that the variables are significant. With this we could predict the values and the predictions that we got were almost equal to the ranks that we took for 2017.

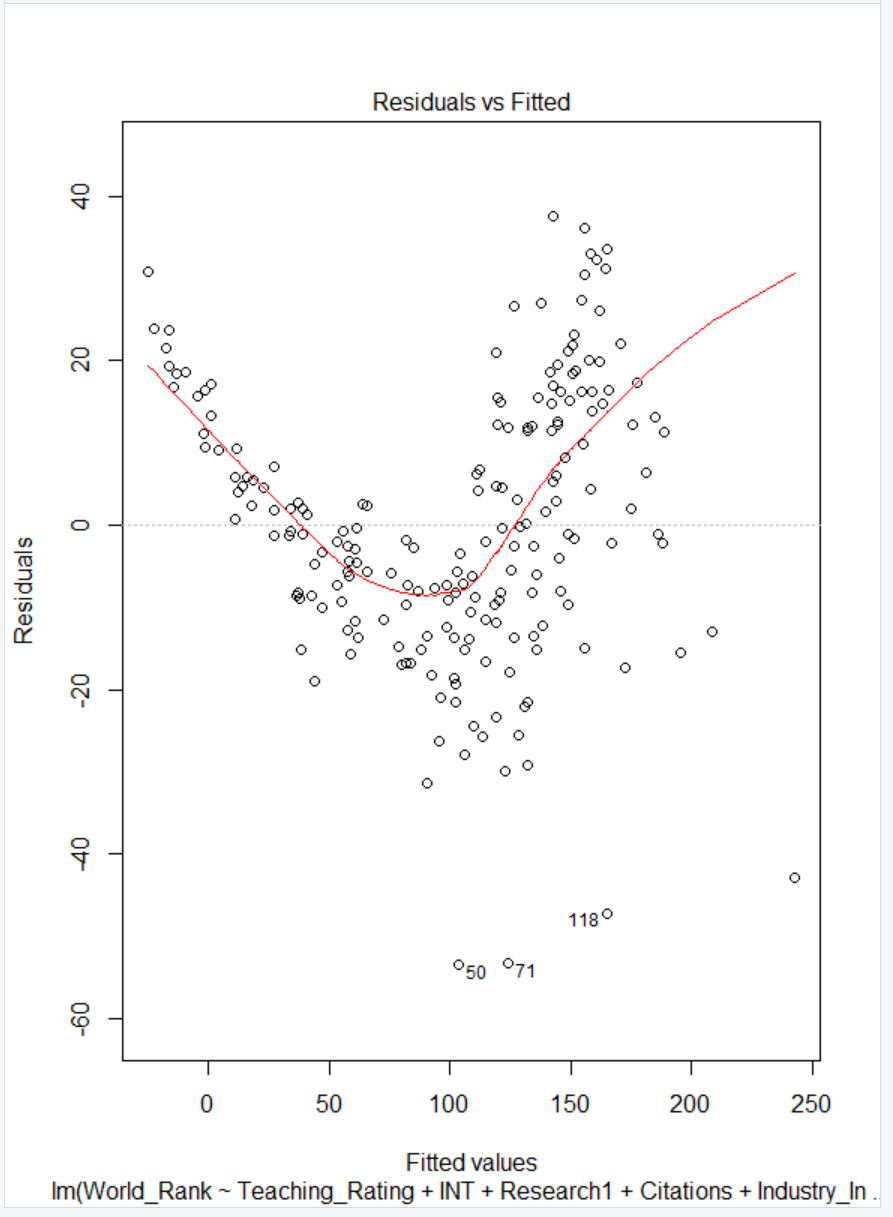
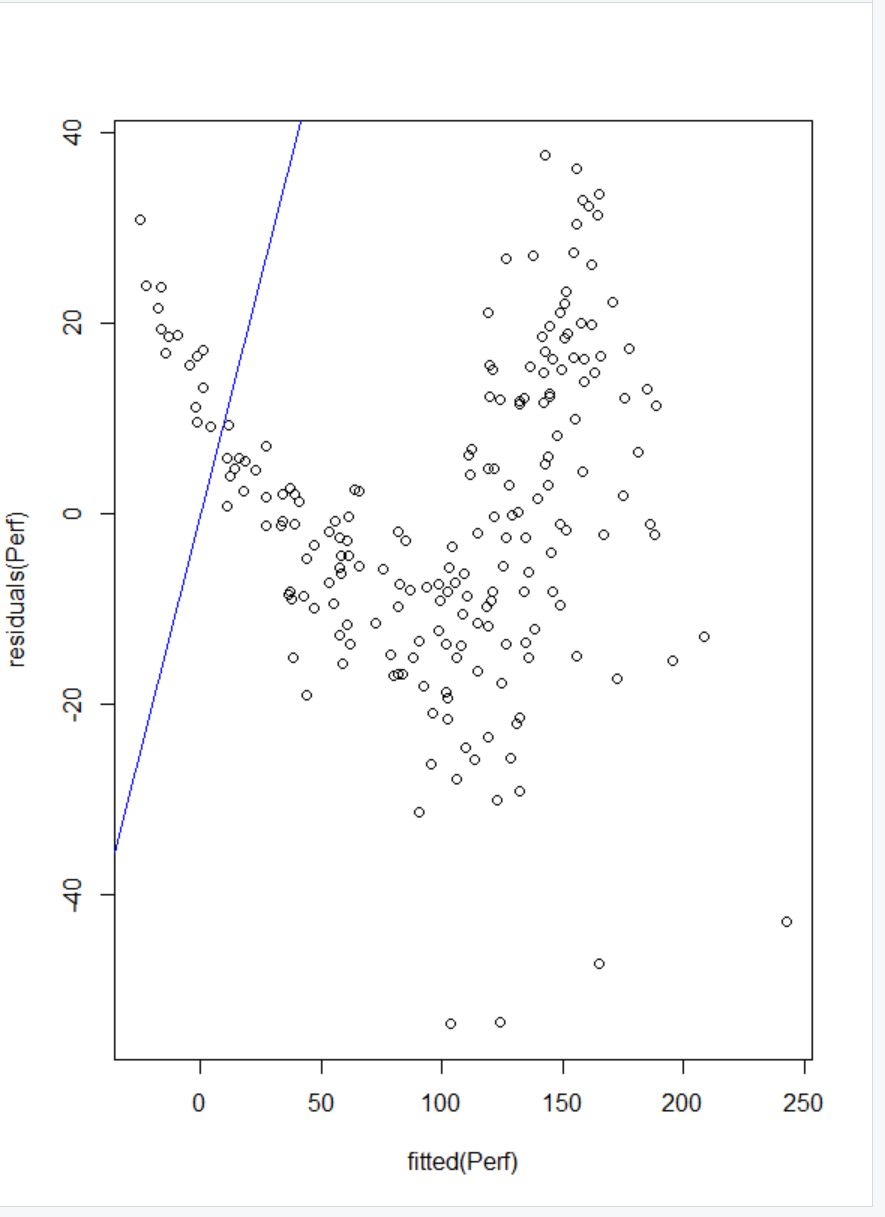


### **Graphs associated with the model:**

The normality plot indicates that most of the points lie on the imaginary line. Except for the outliers at the end. But those outliers have no significant effect on the model. The histogram is almost symmetric, but it is slightly skewed towards right.

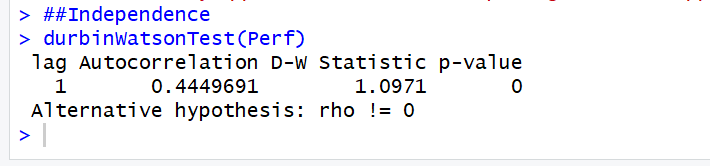
 

**Errors independent of each other:** To see this we need to plot residuals Vs fitted values. We can see that the points are randomly distributed. We can also see the residuals tend to fluctuate up and down in a cyclical pattern. Durbin-Watson test can also be conducted to determine the independence between the variables.

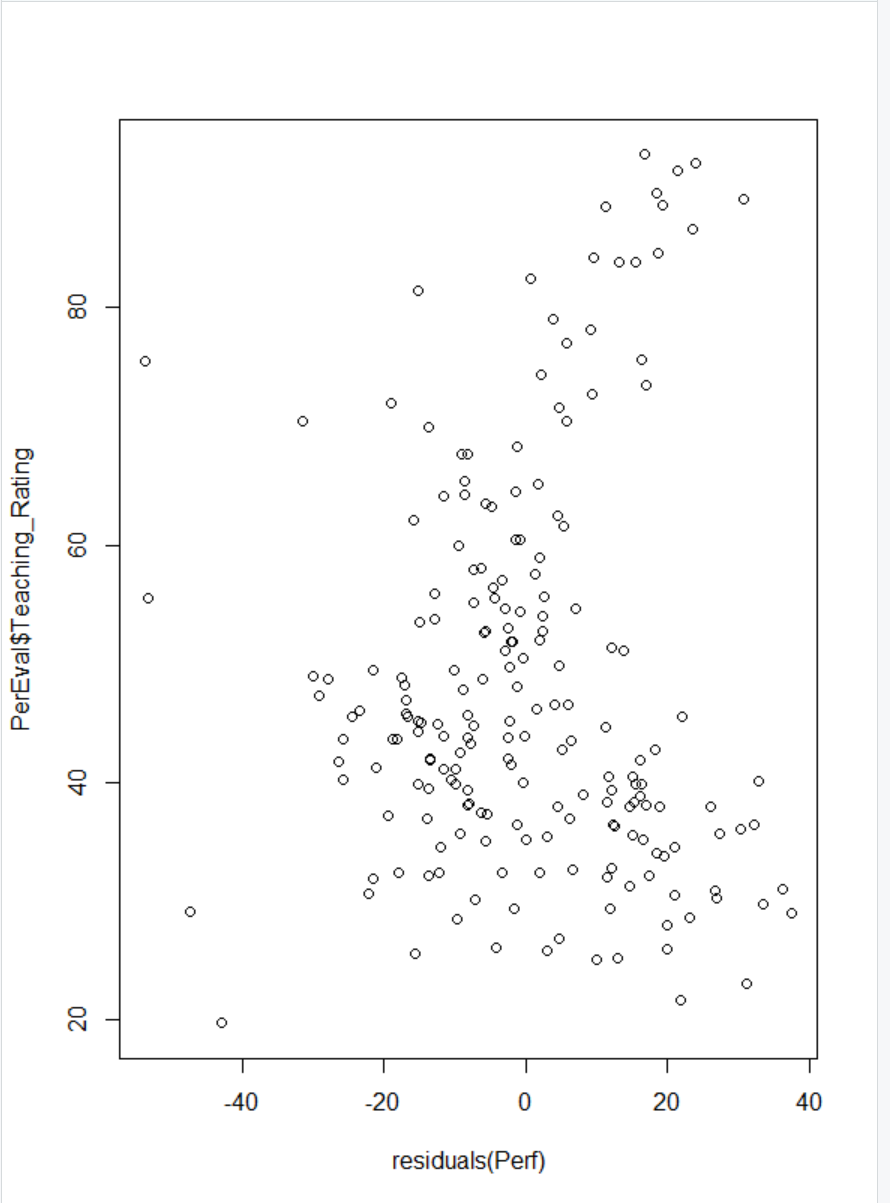
 

### **Durbin-Watson test:**

This is done to know about the independence between the variables and to find the auto-correlation. Over here if the D-W statistic is greater than 2 then that is indicating no positive autocorrelation. So, the value is less than 2 that means positive auto-correlation occurs.



To check errors, have constant variance or not: Plot residuals Vs X value – which is dependent value and check for the variance. We can see that there is approximately same amount of variation in the residuals at each value of Xi. There is a violation of assumptions, we should extract information from the residuals, and put back into the model, because we want to information in the model not in the residuals. Residuals Vs time is plotted to investigate independency of residuals.

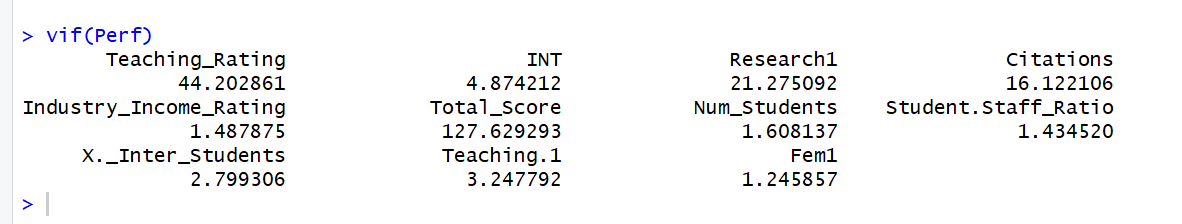
 

# **STEP-WISE REGRESSION MODEL:**

We adopted Step-wise regression model, to build up the model. One dependent variable should be used changing the independent variables to get the best fit model. Using at least one independent variable predicting the values for dependent variable is called Regression. Then we checked for VIF and ANOVA table.

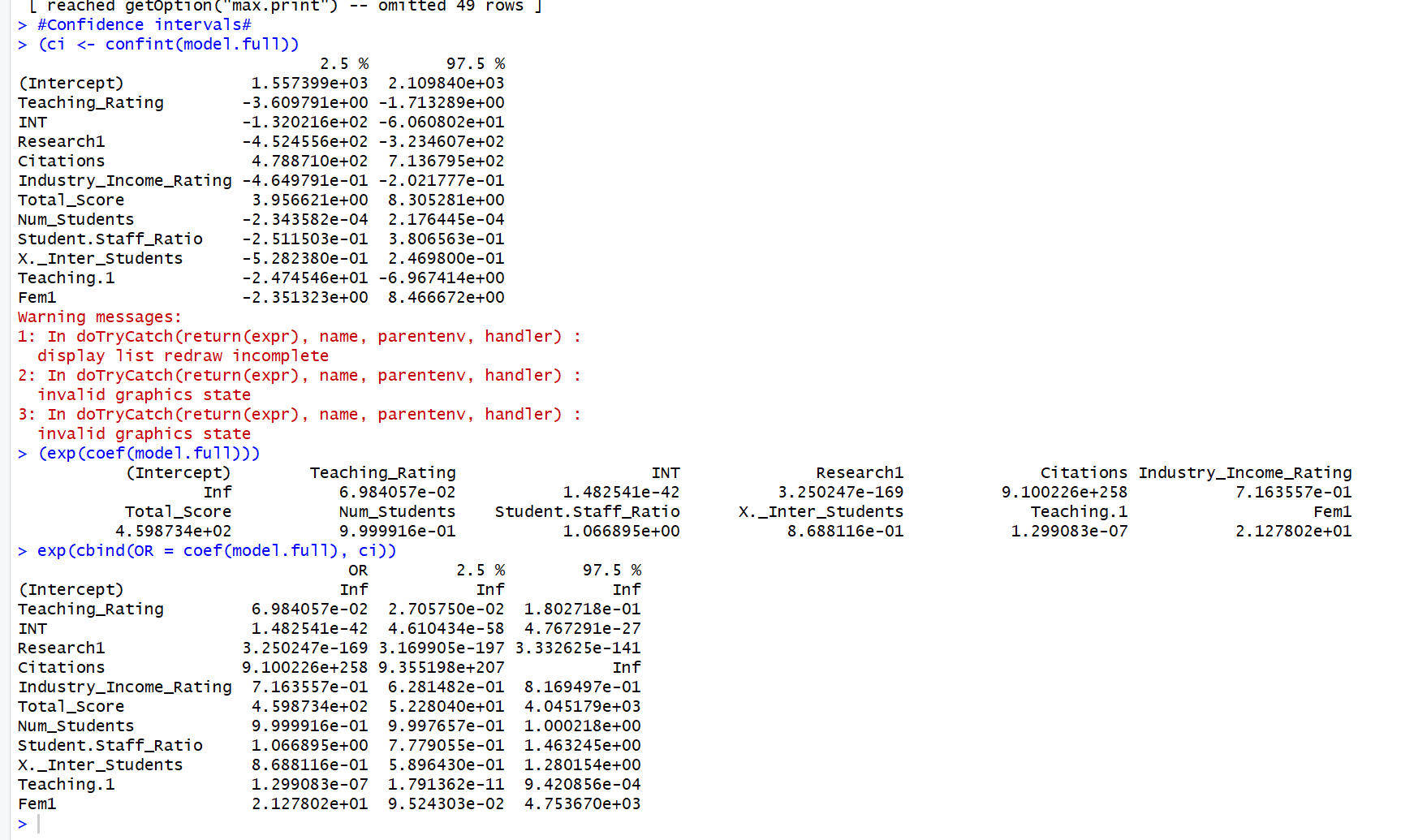
There can be several things that can be considered in Step-wise approach like, forward search, backward search, forward=backward search. We took forward=backward search where we took the variables which can go into the model and depending upon the R2 that we got we deleted those variables.

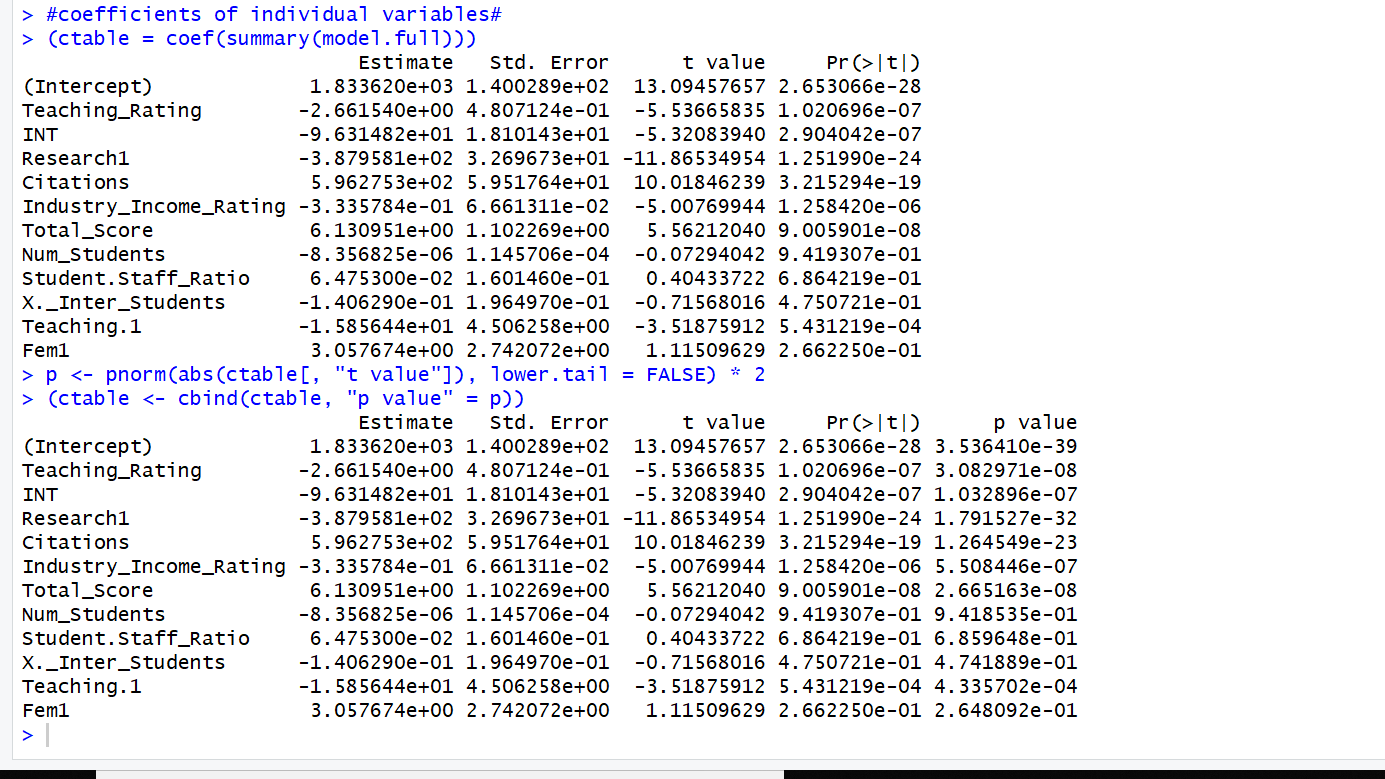
VIF:(Variance Inflation factor): This is used to check the collinearity and to decide what are the variables that can go into the model. If the inflation factor is greater than 5 then only those variables should be selected and the values lesser than 5 should be rejected.



## CONFIDENCE INTERVALS:

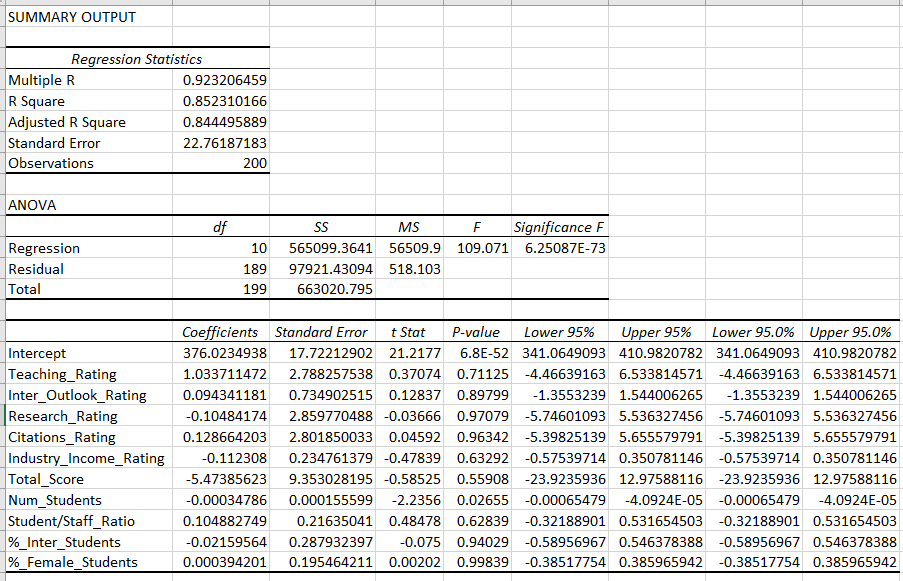
The table below tells the coefficients of all independent variables and ranks.





## ANOVA Table:

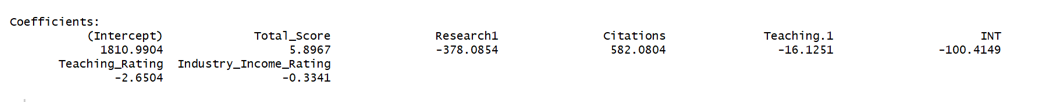
It is used to determine multiple coefficient in Excel. It shows the t-test, partial f-test results, Adjusted R-square, multiple R-square, statistic components of F are also determined with degrees of freedom.

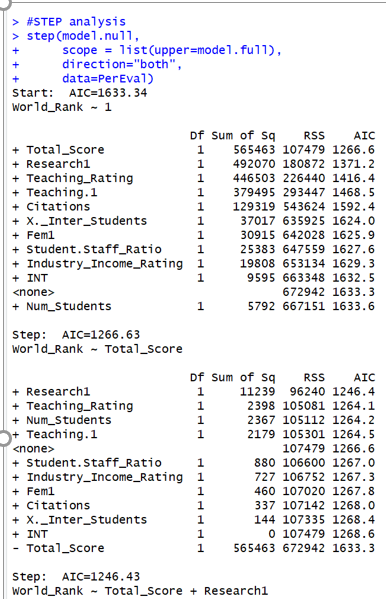
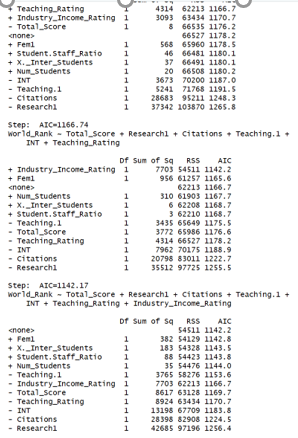


## AIC:

The Akaike information criterion (**AIC**) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data. **AIC** estimates the quality of each model, relative to each of the other models. Thus, **AIC** provides a means for model selection. Generic function calculating Akaike's ‘An Information Criterion’ for one or several fitted model objects for which a log-likelihood value can be obtained, according to the formula −2log-likelihood+knpar, where npar represents the number of parameters in the fitted model, and k=2 for the usual AIC, or k=log(n) (n being the number of observations) for the so-called BIC or SBC (Schwarz's Bayesian criterion). The variables need to be selected with less AIC.

Initially we calculated AIC factor for all the variables, we got the AUC value as very large. Then we started to add the variables. Then we observed that the AIC factor kept on decreasing. So, for the final model which we took we got the AIC factor small when compared to the individual factor analysis. Hence, we kept that model.



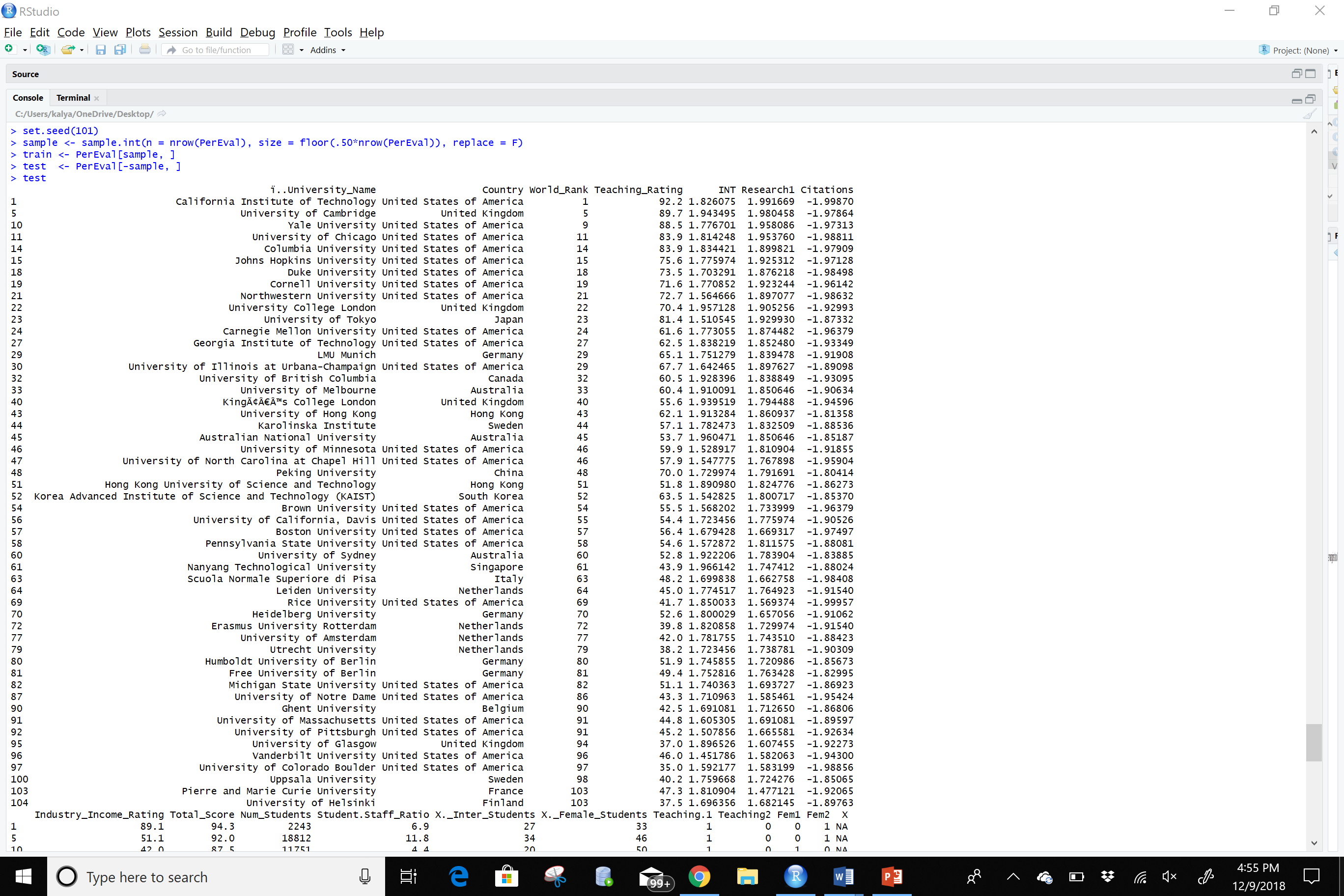
 

# **Cross Validation:**

Initially we took 2 data sets. First data set is for training data – data1 to build up the model and the second data set is to test the model – data2 whether this model would be useful for some other data for other years and whether this model serves good.

We took leave-k out cross validation approach where in we need to remove the observations fit the model, find the prediction errors. This should be done many times such that sum of squares of prediction errors is attained.

We have observed the cross validated data for the complete university ranking. From the sample data, we have observed the train and tested data sets which appear to be normal and has good performance.



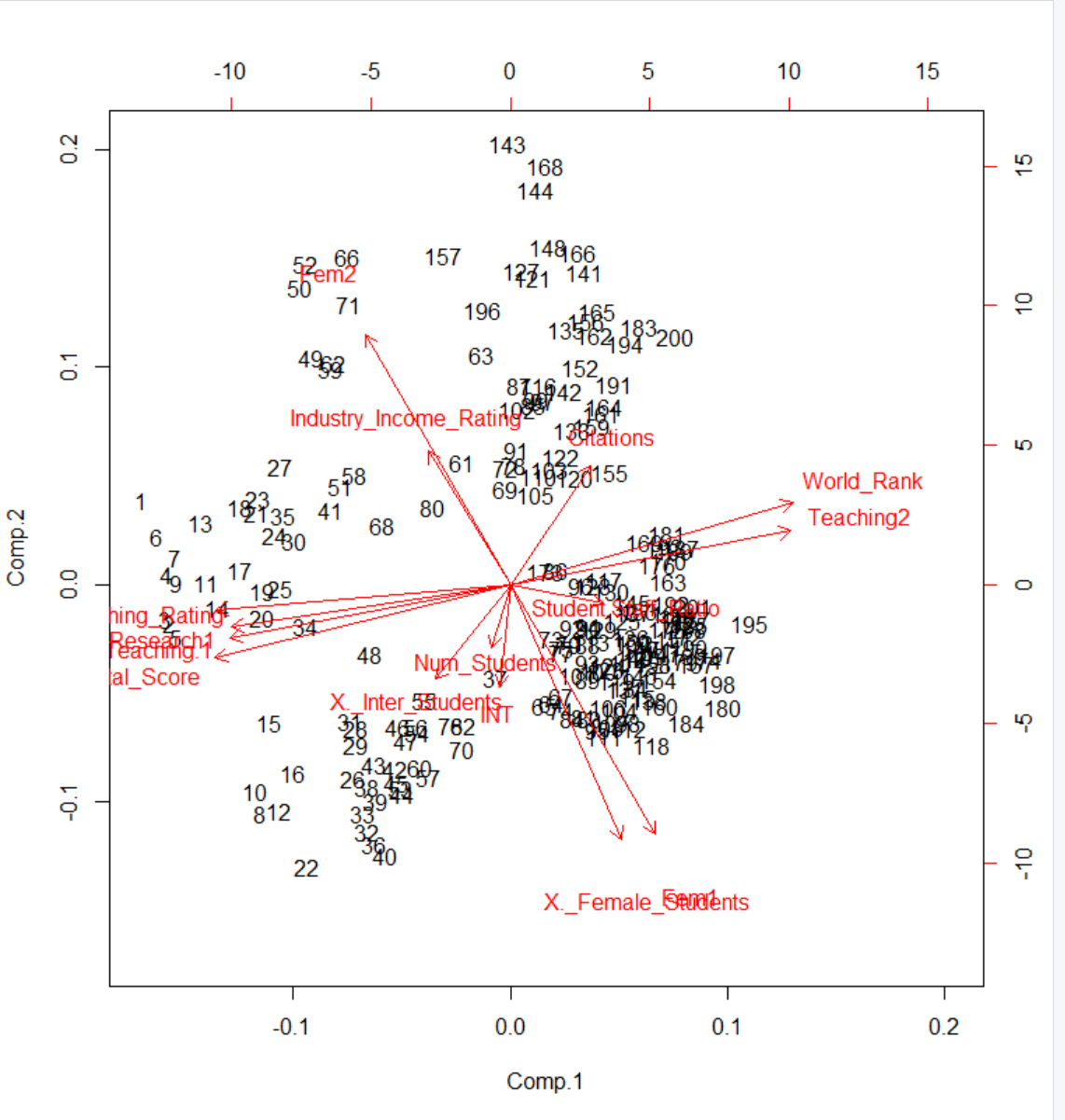


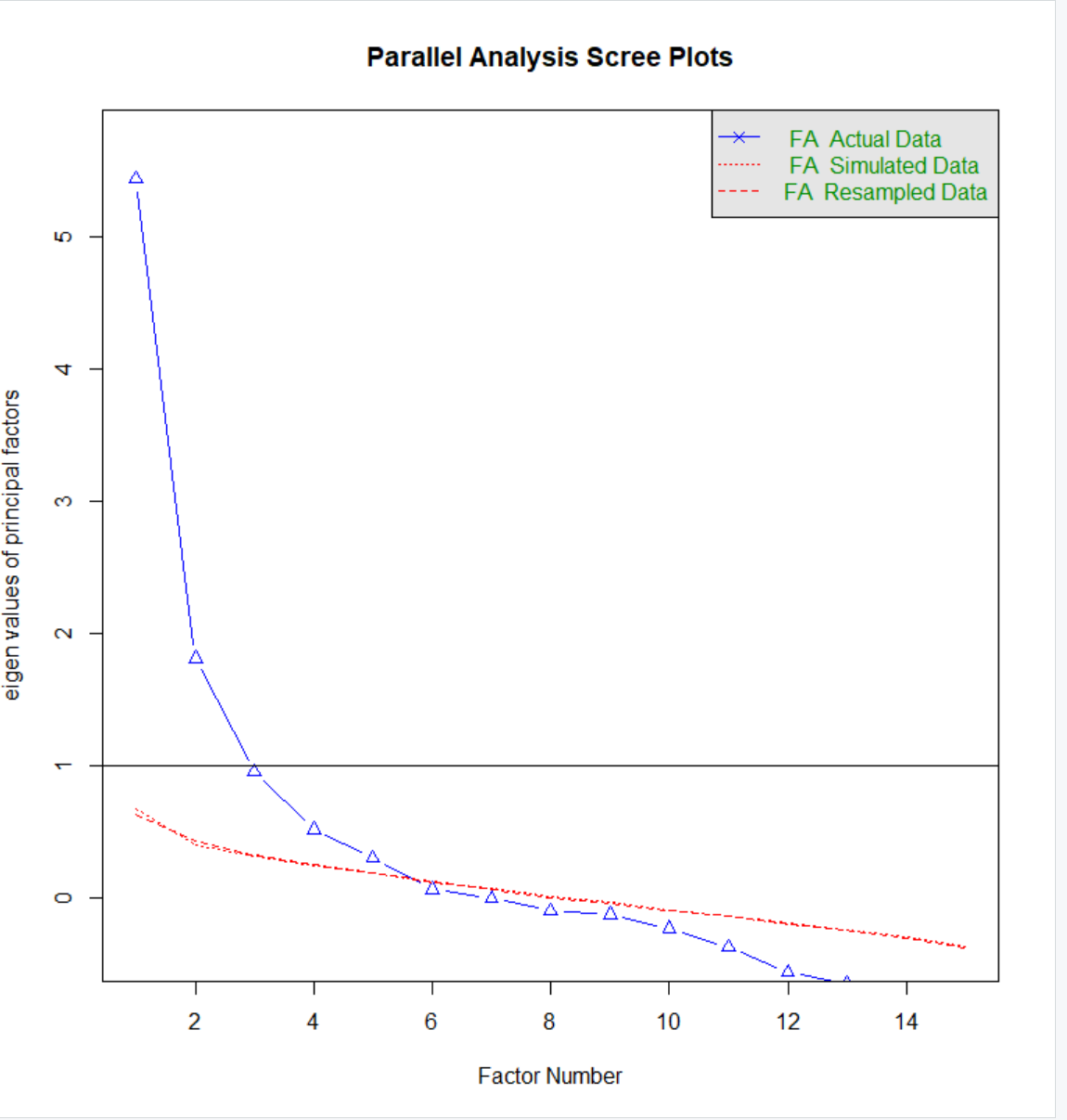
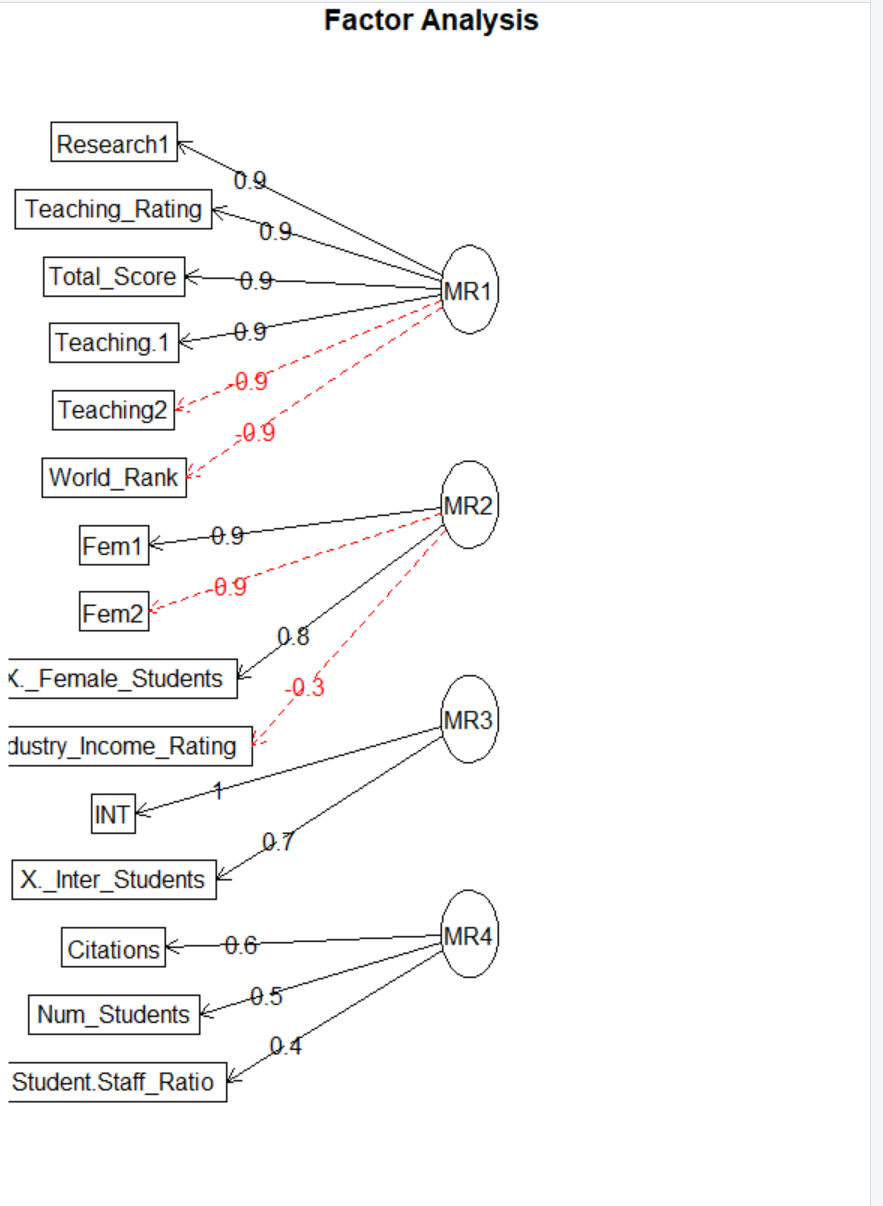
# **PRINCIPLE COMPONENTS ANALYSIS:**

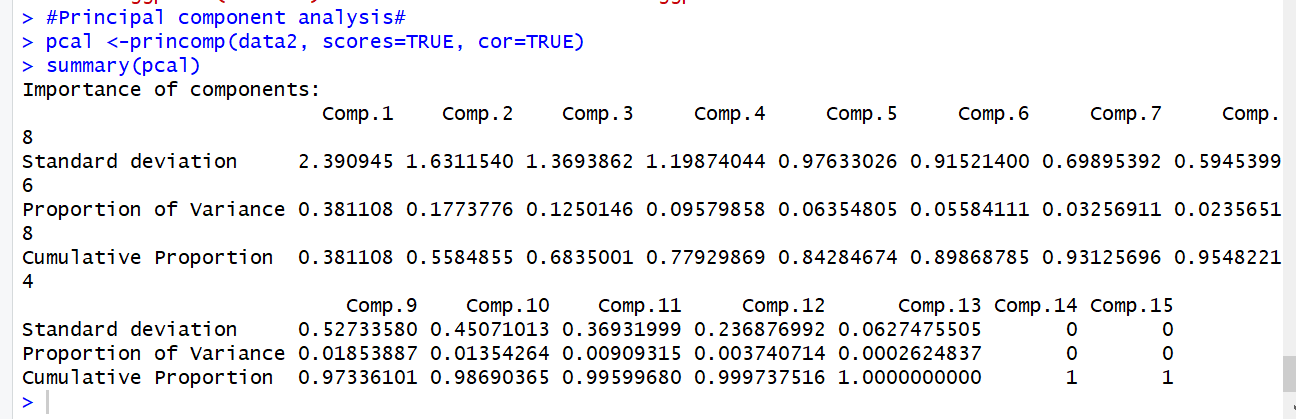
To have a proper analysis on variance between independent variables we have used principal component analysis. The data sets have been transformed into several components based on the variables and we can observe the standard deviation, proportion of variance and cumulative variance. Plot and the biplot can depict the graphical representation of these component variance as observed.

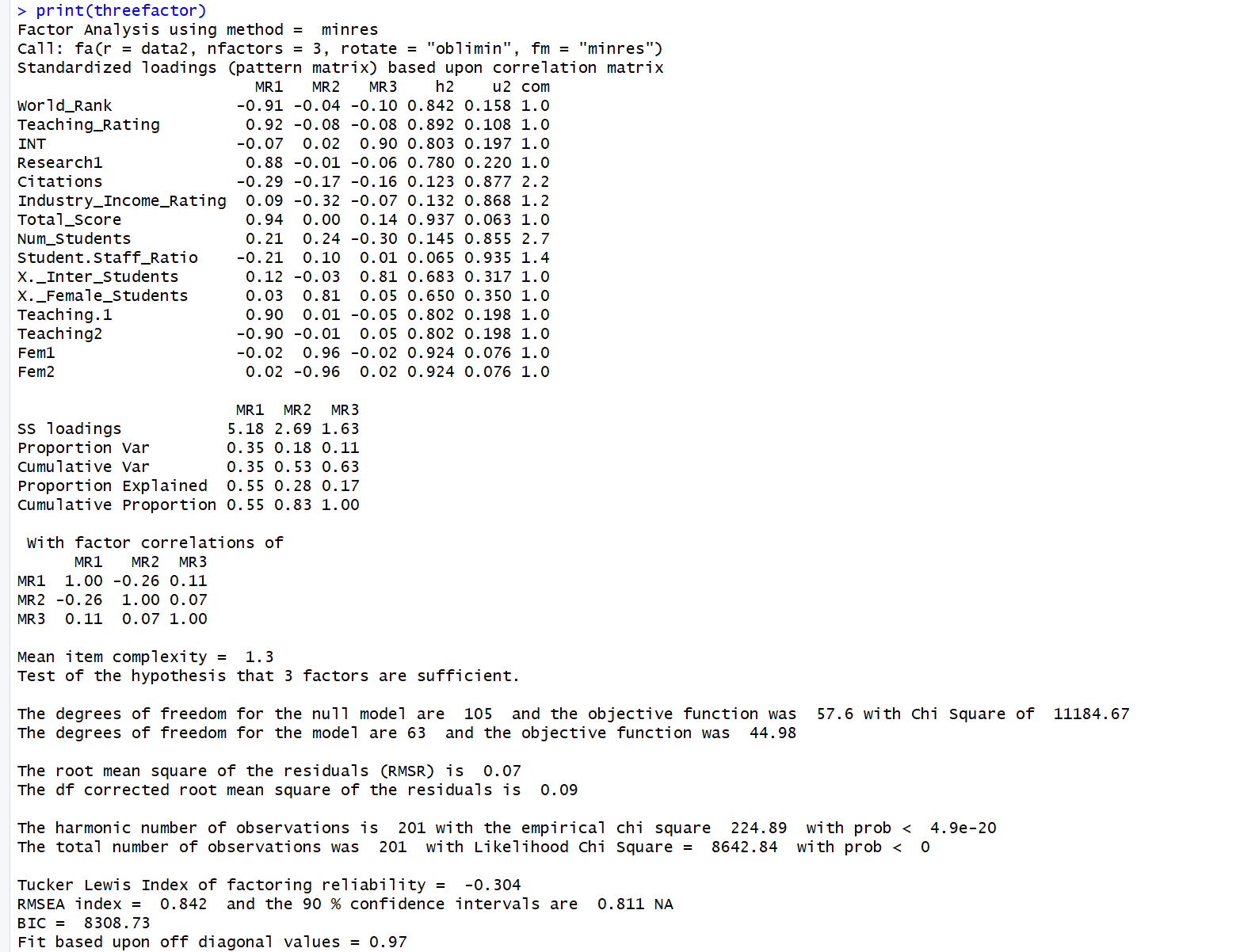
# **FACTOR ANALYSIS:**

We have used other libraries like psych and GPA rotation. We have done factor analysis both three and four factors respectively. Each individual component has been divided within four factors and joined a branch. And for these factors the variance is observed from the loadings.

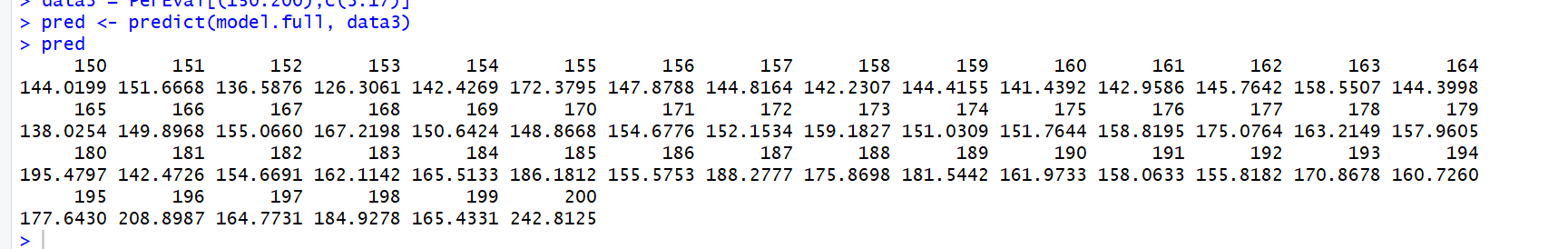
 





# **Predictions for Ranking:**

We predicted the ranks in R as well as in Excel. In excel we used the time-series forecasting and predicted the values. The values that we got were almost equal. We predicted the values for the last 50 universities.



# **Time-Series Forecasting:**

We took the previous years data through various websites apart from 2017. Then using linear forecasting predicted the values for remaining years. Using those values by adopting moving averages method we predicted the Colorado Universities ranking for 2018, 2019 and 2020.

