A Statistical Data Analysis

World University Rankings Dataset, *THE*

About the team

The Mean Squares, Section B

Anushka S. Hebbar
Chandradhar Rao
Ayush M. Kapasi
Chaitra B. Kayi

November 2020
Department of Computer Science and Engineering
PES University

The Dataset

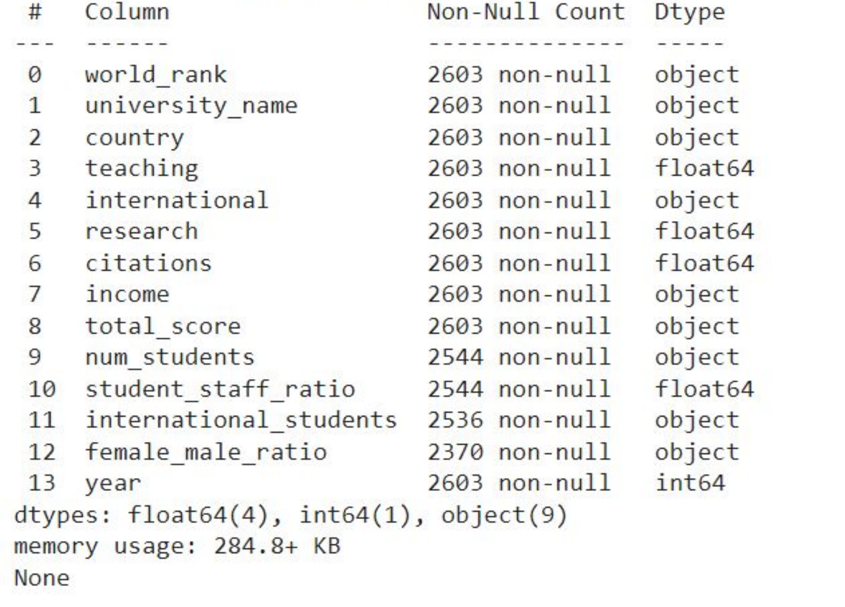
World University Rankings

Curated by The Times Higher Education (THE)

Global performance tables that judge research-intensive universities.

The rankings are calculated based on [illegible] facto[illegible]

These 5 factors are represented as 14 features (or columns) in the dataset.



```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   world_rank              2603 non-null   object
 1   university_name         2603 non-null   object
 2   country                 2603 non-null   object
 3   teaching                2603 non-null   float64
 4   international           2603 non-null   object
 5   research                2603 non-null   float64
 6   citations               2603 non-null   float64
 7   income                  2603 non-null   object
 8   total_score             2603 non-null   object
 9   num_students            2544 non-null   object
 10  student_staff_ratio     2544 non-null   float64
 11  international_students  2536 non-null   object
 12  female_male_ratio       2370 non-null   object
 13  year                    2603 non-null   int64
dtypes: float64(4), int64(1), object(9)
memory usage: 284.8+ KB
None
```

Libraries Used

- Pandas
    - Used for data manipulation
- Math
    - Provides access to mathematical functions
- Copy
    - Used to copy lists (shallow and deep copy)
- Numpy
    - Used for high level mathematical functions to find mean, roots etc.
- Matplotlib
    - Major plotting library used (for box plots, bar graphs etc)
- Plotly
    - Library used for plotting
- Scipy
    - Used for technical computing, eg-to find p value
- Seaborn
    - data visualisation

Data Wrangling

Processing num_students (to remove commas)

```
After processing 'num_students':

20152.0, 2243.0, 11074.0, 15596.0, 7929.0,
18812.0, 19919.0, 36186.0, 15060.0, 11751.0,
38206.0, 14221.0, 15128.0, 21424.0, 18178.0,
41786.0, 66198.0, 25055.0, 20376.0
```

```
Before processing ‘num_students:

20,152, 2,243, 11,074, 15,596, 7,929,
18,812, 19,919, 36,186, 15,060,
11,751, 38,206, 14,221, 15,128,
21,424, 18,178, 41,786, 66,198,
25,055, 20,376, 11,885
```

Processing international_students (to remove % symbols)

After processing 'international_students':

```
25.0, 27.0, 33.0, 22.0, 27.0, 34.0, 34.0,
15.0, 51.0, 20.0, 15.0, 21.0, 23.0, 19.0,
37.0, 16.0, 15.0, 28.0, 20.0, 35.0, 38.0,
46.0, 13.0, 17.0, 15.0, 10.0, 26.0, 4.0,
11.0, 25.0
```

Before processing international_students:

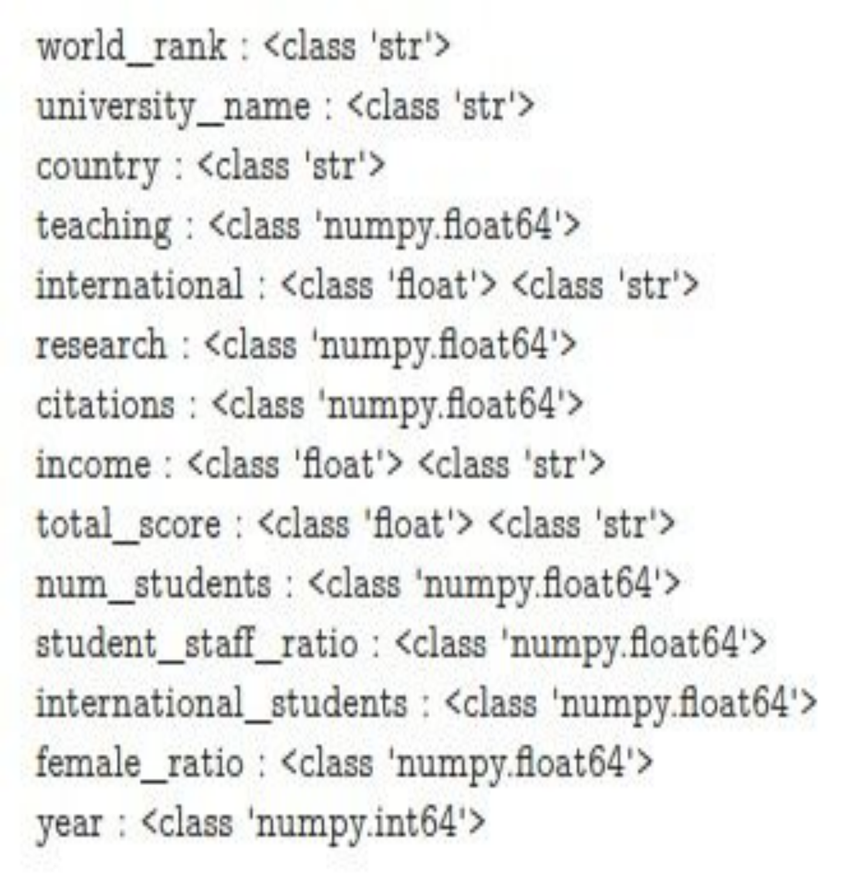
```
25%, 27%, 33%, 22%, 27%, 34%, 34%,
15%, 51%, 20%, 15%, 21%, 23%, 19%,
37%, 16%, 15%, 28%, 20%, 35%, 38%,
46%, 13%, 17%, 15%, 10%, 26%, 4%,
11%, 25%
```

Processing female_ratio

```
After processing 'female_ratio':

nan, 0.33, 0.37, 0.42, 0.45, 0.46,
0.46, 0.5, 0.37, 0.5, 0.52, 0.42, 0.5,
0.48, 0.31, 0.48, nan, nan, 0.51, 0.39
```

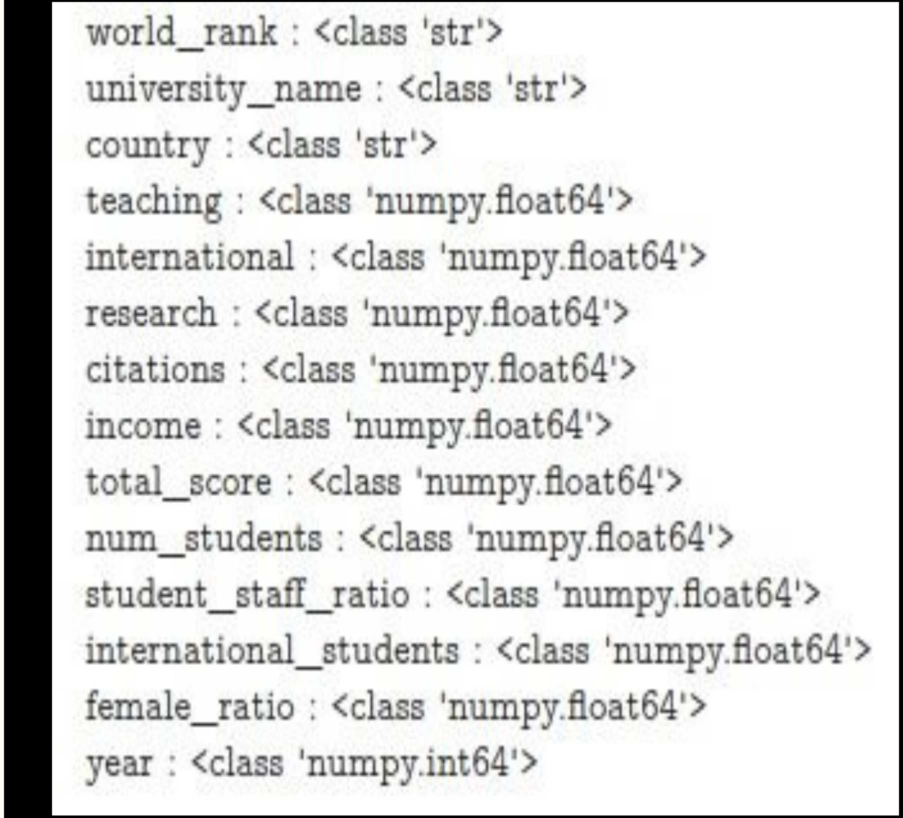
```
Before processing female_ratio:

nan, 33 : 67, 37 : 63, 42 : 58, 45 :
55, 46 : 54, 46 : 54, 50 : 50, 37 :
63, 50 : 50, 52 : 48, 42 : 58, 50 :
50, 48 : 52, 31 : 69, 48 : 52, nan,
nan, 51 : 49, 39 : 61
```

Before Data Type Appropriation



```
world_rank : <class 'str'>
university_name : <class 'str'>
country : <class 'str'>
teaching : <class 'numpy.float64'>
international : <class 'float'> <class 'str'>
research : <class 'numpy.float64'>
citations : <class 'numpy.float64'>
income : <class 'float'> <class 'str'>
total_score : <class 'float'> <class 'str'>
num_students : <class 'numpy.float64'>
student_staff_ratio : <class 'numpy.float64'>
international_students : <class 'numpy.float64'>
female_ratio : <class 'numpy.float64'>
year : <class 'numpy.int64'>
```

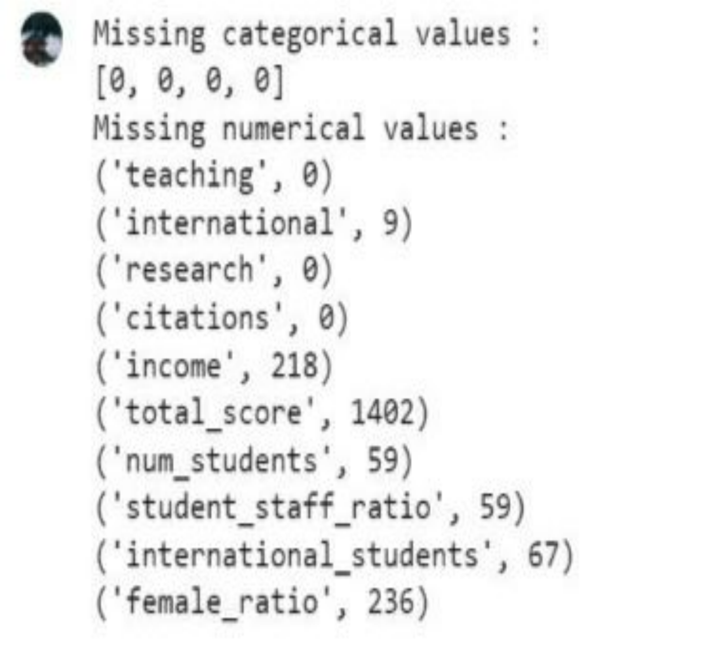
After Data Type Appropriation



```
world_rank : <class 'str'>
university_name : <class 'str'>
country : <class 'str'>
teaching : <class 'numpy.float64'>
international : <class 'numpy.float64'>
research : <class 'numpy.float64'>
citations : <class 'numpy.float64'>
income : <class 'numpy.float64'>
total_score : <class 'numpy.float64'>
num_students : <class 'numpy.float64'>
student_staff_ratio : <class 'numpy.float64'>
international_students : <class 'numpy.float64'>
female_ratio : <class 'numpy.float64'>
year : <class 'numpy.int64'>
```

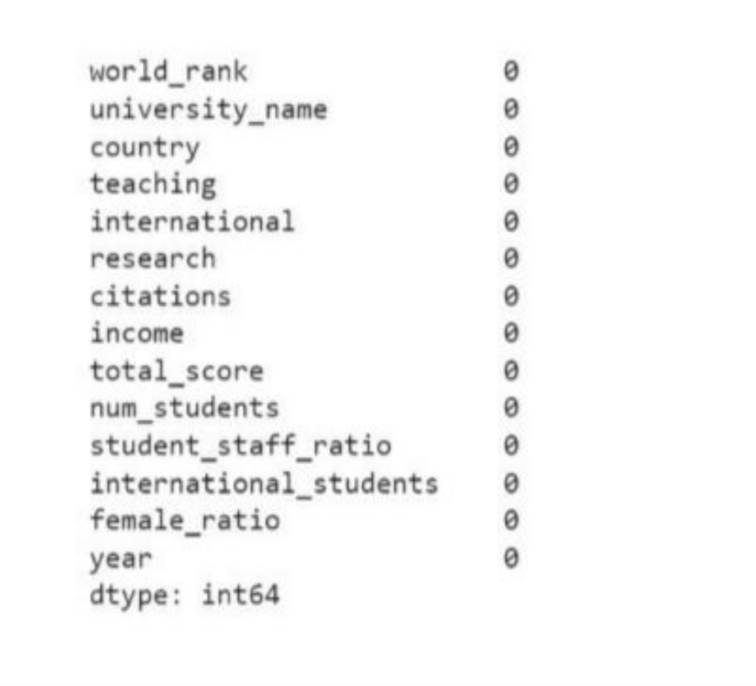
Handling Missing Values

Before



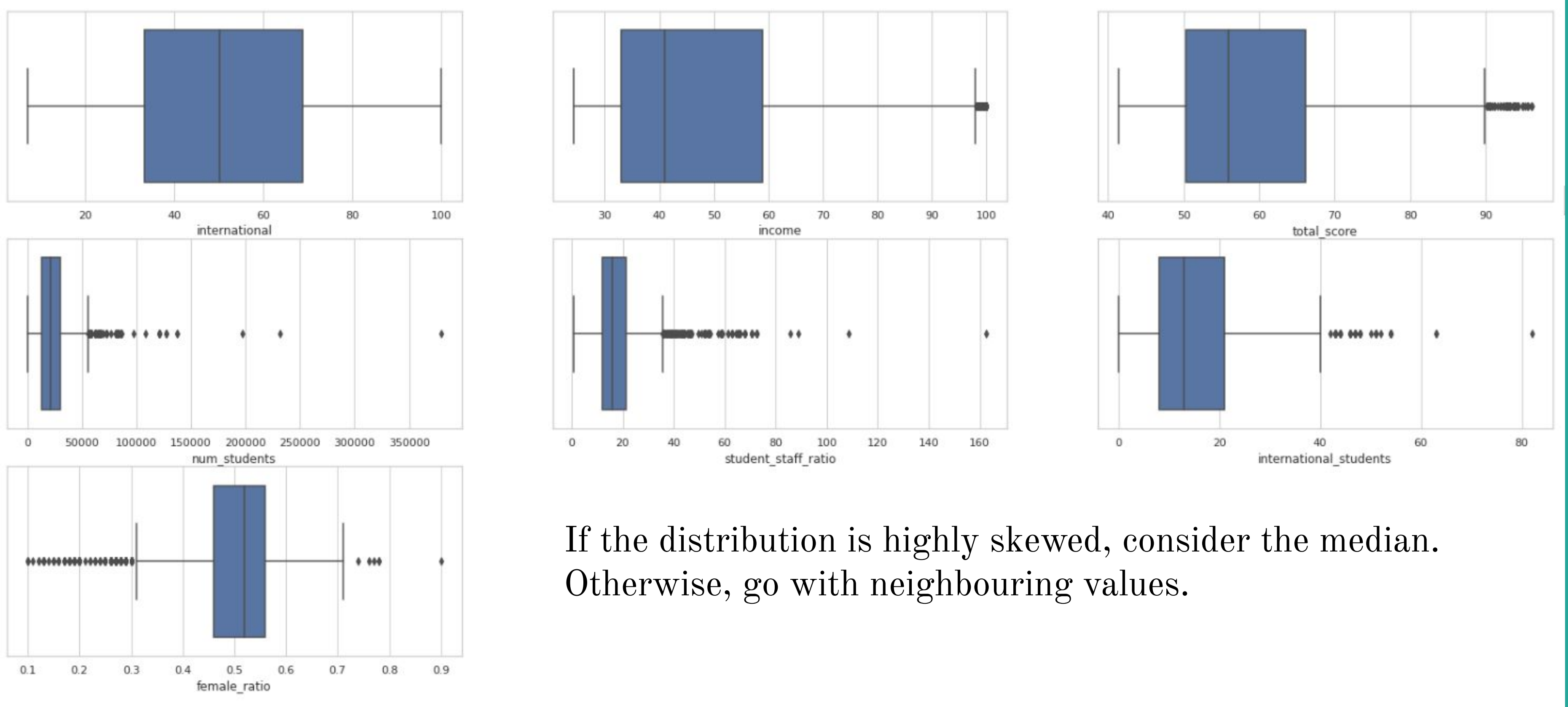
```
Missing categorical values :
[0, 0, 0, 0]
Missing numerical values :
('teaching', 0)
('international', 9)
('research', 0)
('citations', 0)
('income', 218)
('total_score', 1402)
('num_students', 59)
('student_staff_ratio', 59)
('international_students', 67)
('female_ratio', 236)
```

After



```
world_rank                0
university_name           0
country                   0
teaching                  0
international             0
research                  0
citations                 0
income                    0
total_score               0
num_students              0
student_staff_ratio       0
international_students    0
female_ratio              0
year                      0
dtype: int64
```

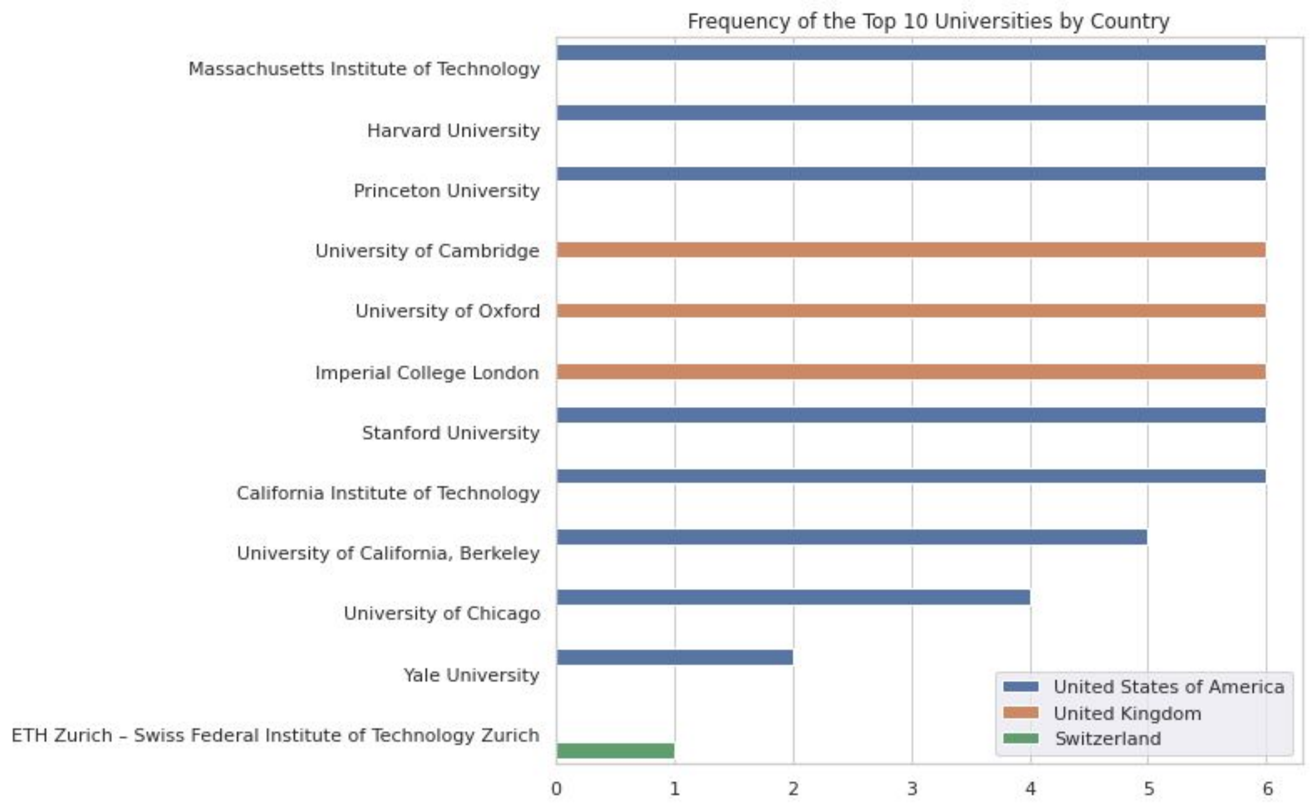
Determine the measure of central tendency with which to replace missing values in numerical variables.



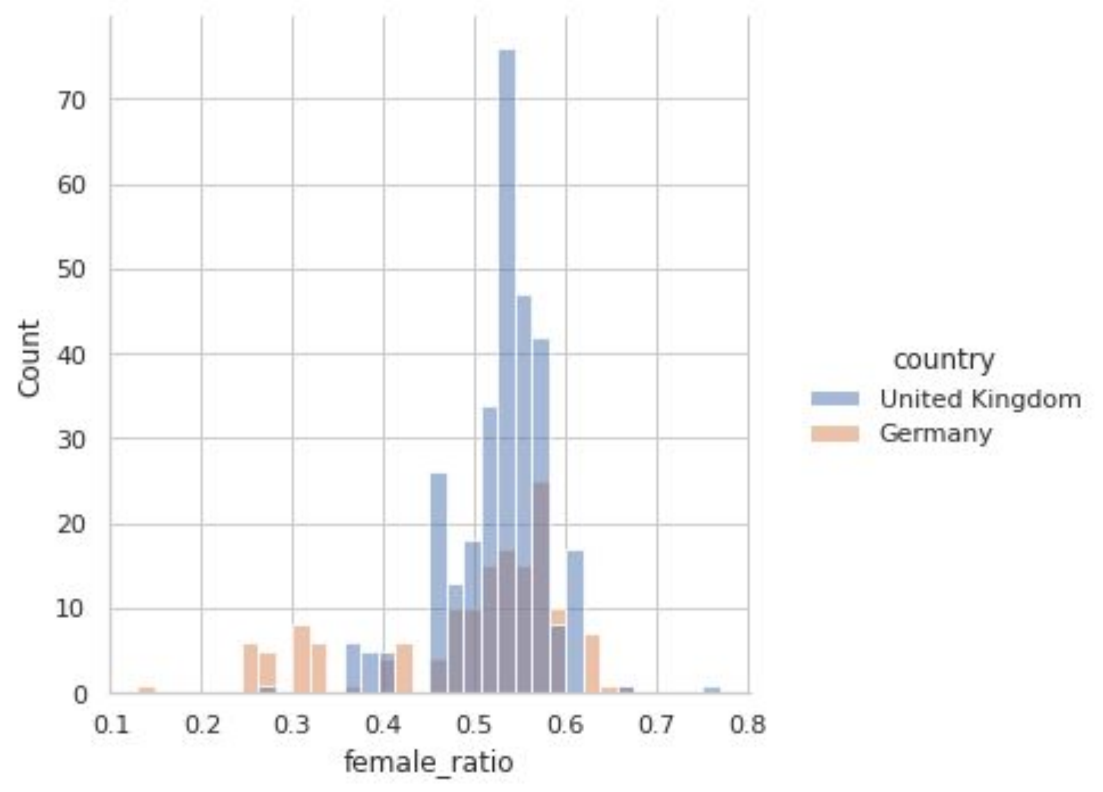
If the distribution is highly skewed, consider the median. Otherwise, go with neighbouring values.

Data Visualization

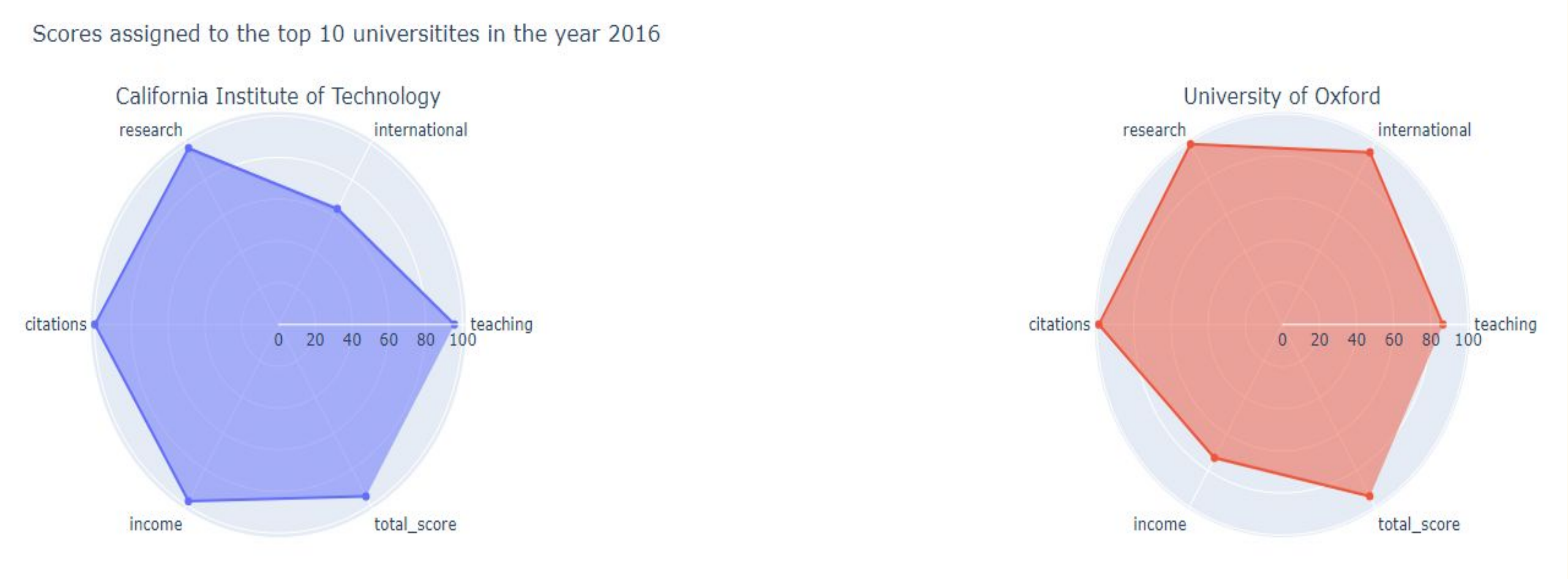
Which countries around the world feature in the top 10 universities and how many times?



How do the ratios of female students in universities located in Germany and the UK in the year 2011 differ?



Scores assigned to the top 10 universities in the year 2016 using spyder / radar Plots

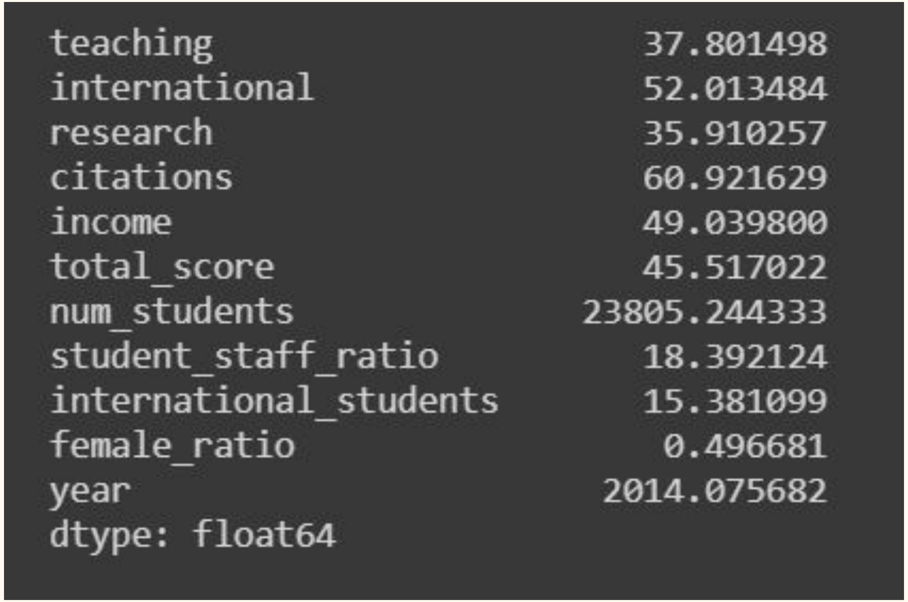


Ranked # 1

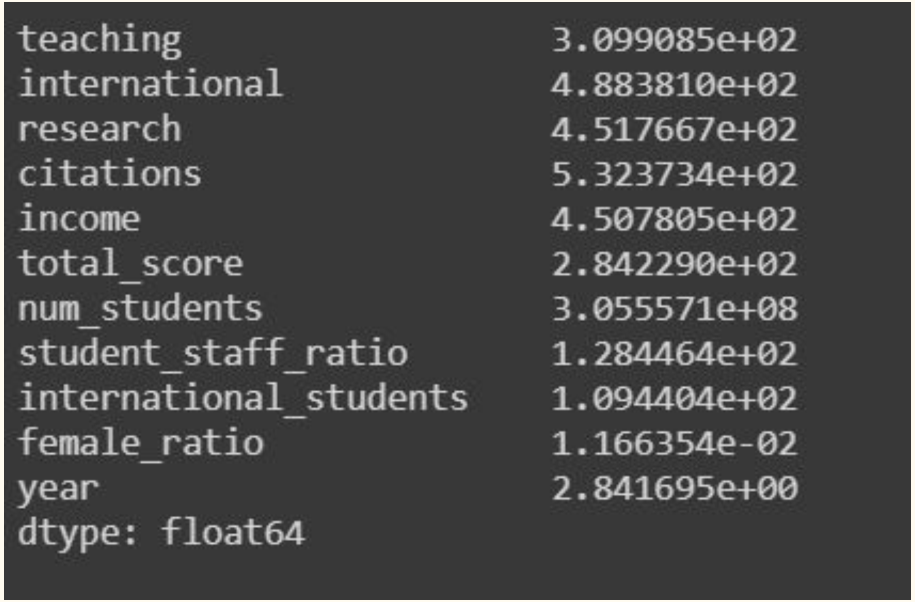
Ranked # 2

Normalization and Standardization

Before normalization, the mean and variance of the columns were :

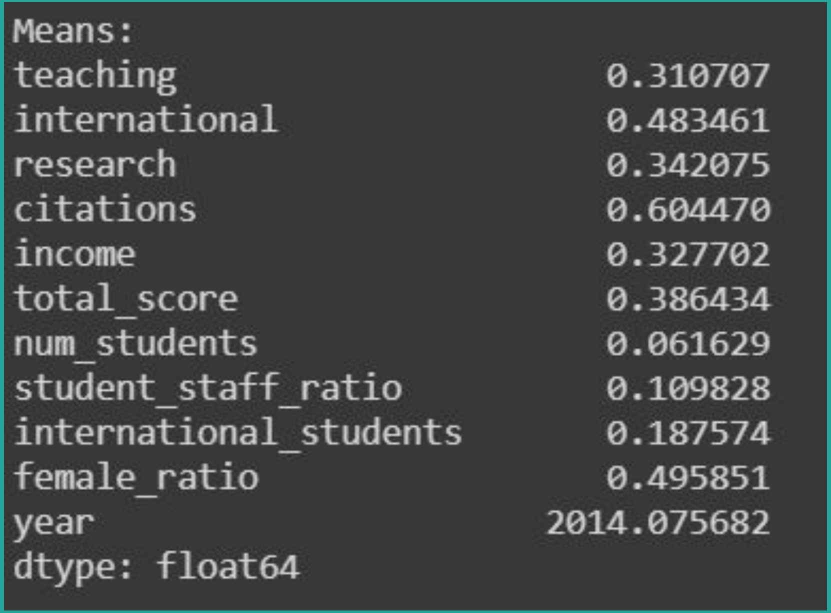


```
teaching                       37.801498
international                  52.013484
research                       35.910257
citations                      60.921629
income                         49.039800
total_score                    45.517022
num_students                23805.244333
student_staff_ratio            18.392124
international_students         15.381099
female_ratio                    0.496681
year                         2014.075682
dtype: float64
```

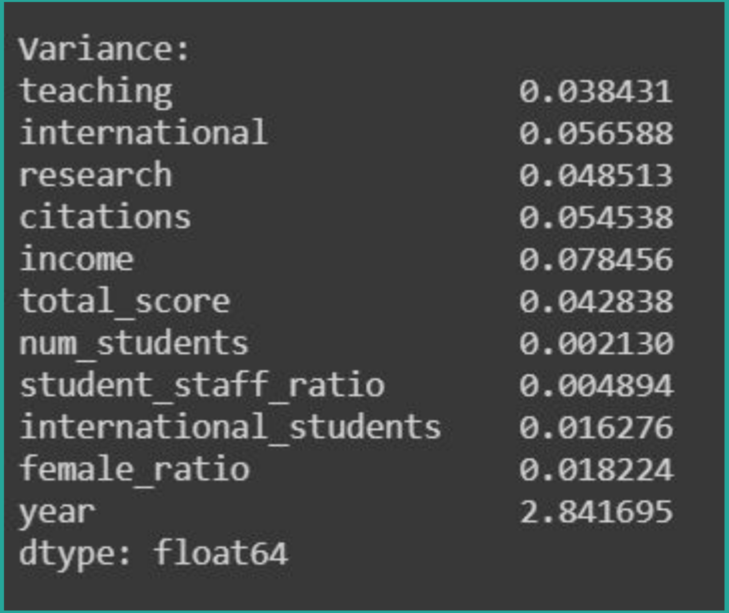


```
teaching                  3.099085e+02
international             4.883810e+02
research                  4.517667e+02
citations                 5.323734e+02
income                    4.507805e+02
total_score               2.842290e+02
num_students              3.055571e+08
student_staff_ratio       1.284464e+02
international_students    1.094404e+02
female_ratio              1.166354e-02
year                      2.841695e+00
dtype: float64
```

After normalization, the mean and variance of the data columns were:

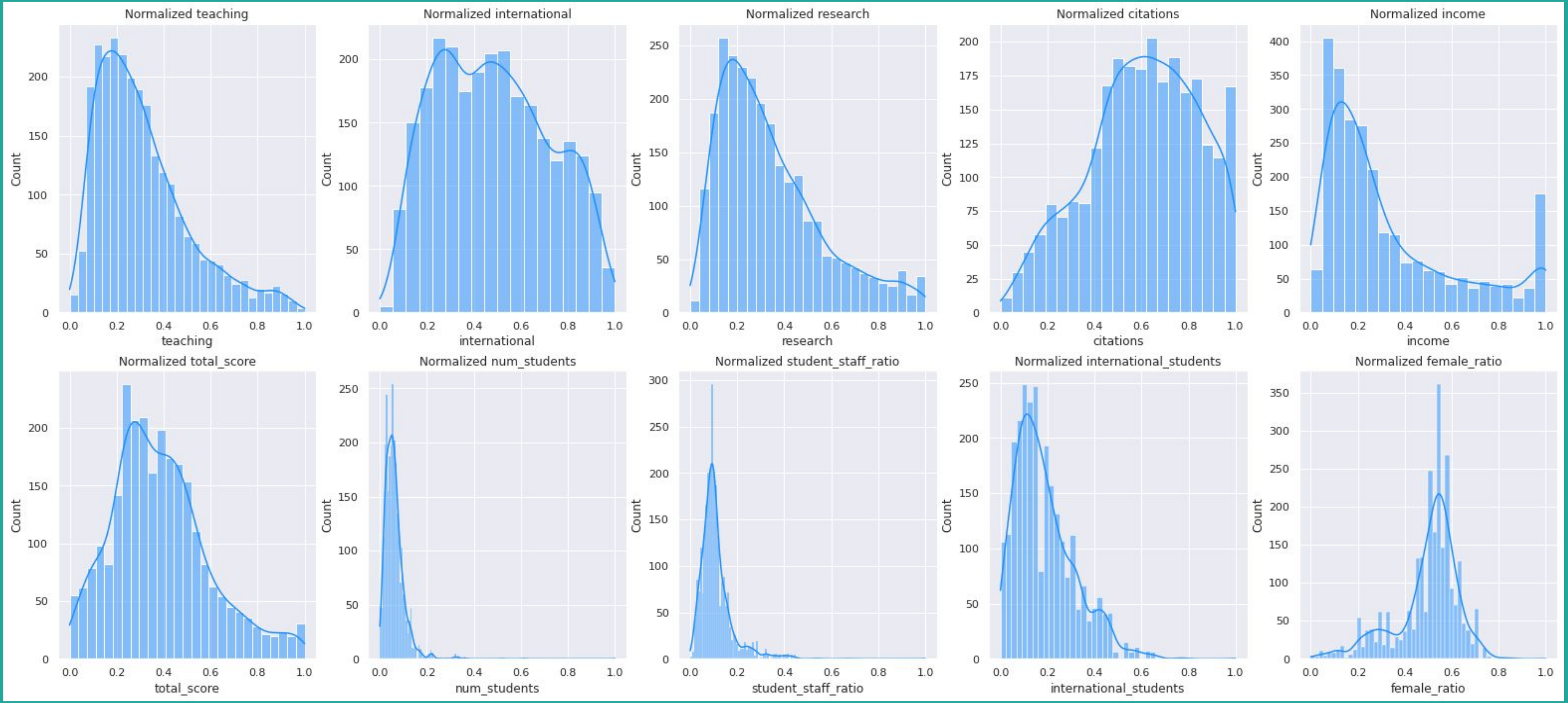


```
Means:
teaching                     0.310707
international                0.483461
research                     0.342075
citations                    0.604470
income                       0.327702
total_score                  0.386434
num_students                 0.061629
student_staff_ratio          0.109828
international_students       0.187574
female_ratio                 0.495851
year                      2014.075682
dtype: float64
```



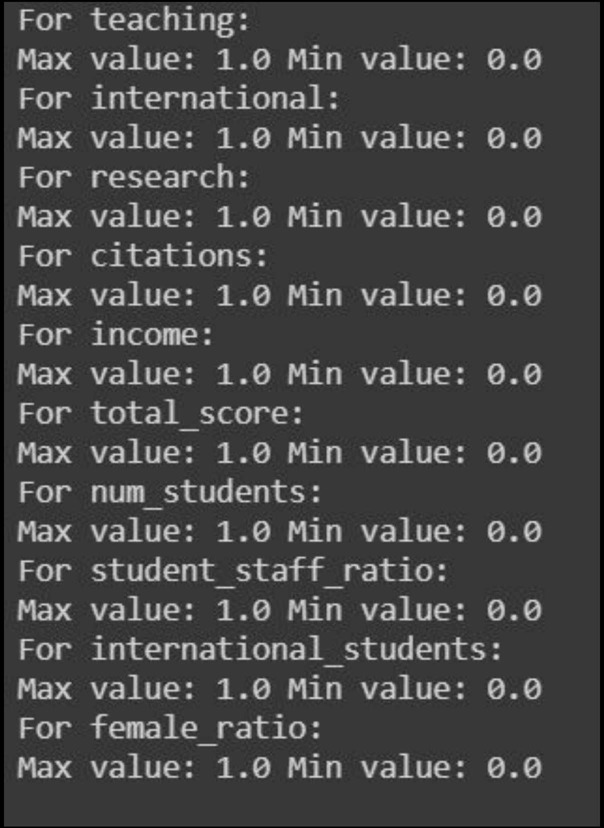
```
Variance:
teaching                  0.038431
international             0.056588
research                  0.048513
citations                 0.054538
income                    0.078456
total_score               0.042838
num_students              0.002130
student_staff_ratio       0.004894
international_students    0.016276
female_ratio              0.018224
year                      2.841695
dtype: float64
```

The normalized data set represented through histograms



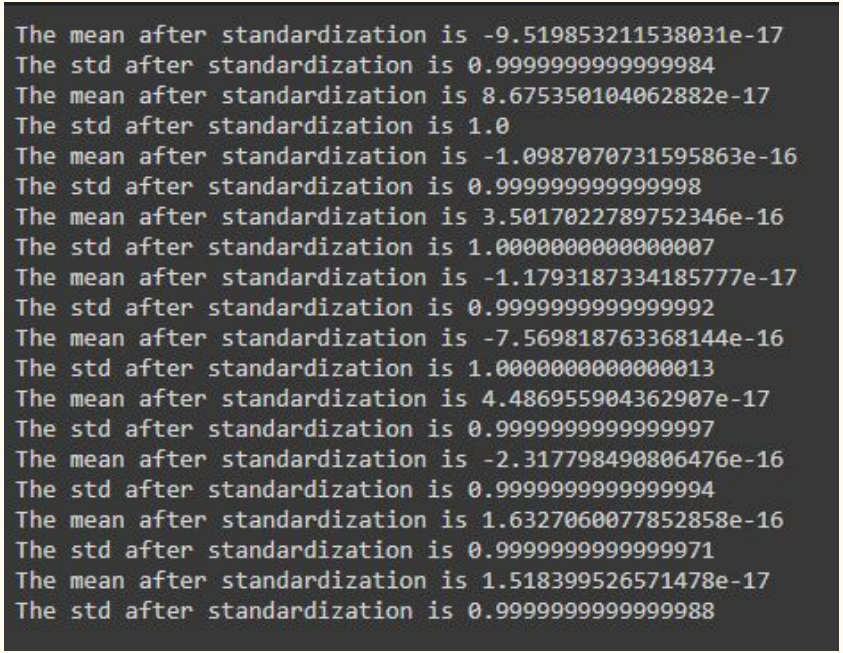
After normalizing

We observe that numeric values in the data set have a common scale between 0 to 1.



```
For teaching:
Max value: 1.0 Min value: 0.0
For international:
Max value: 1.0 Min value: 0.0
For research:
Max value: 1.0 Min value: 0.0
For citations:
Max value: 1.0 Min value: 0.0
For income:
Max value: 1.0 Min value: 0.0
For total_score:
Max value: 1.0 Min value: 0.0
For num_students:
Max value: 1.0 Min value: 0.0
For student_staff_ratio:
Max value: 1.0 Min value: 0.0
For international_students:
Max value: 1.0 Min value: 0.0
For female_ratio:
Max value: 1.0 Min value: 0.0
```

After standardizing



```
The mean after standardization is -9.519853211538031e-17
The std after standardization is 0.9999999999999984
The mean after standardization is 8.675350104062882e-17
The std after standardization is 1.0
The mean after standardization is -1.0987070731595863e-16
The std after standardization is 0.9999999999999998
The mean after standardization is 3.5017022789752346e-16
The std after standardization is 1.0000000000000007
The mean after standardization is -1.1793187334185777e-17
The std after standardization is 0.9999999999999992
The mean after standardization is -7.569818763368144e-16
The std after standardization is 1.0000000000000013
The mean after standardization is 4.486955904362907e-17
The std after standardization is 0.9999999999999997
The mean after standardization is -2.317798490806476e-16
The std after standardization is 0.9999999999999994
The mean after standardization is 1.6327060077852858e-16
The std after standardization is 0.9999999999999971
The mean after standardization is 1.518399526571478e-17
The std after standardization is 0.9999999999999988
```

We observe the mean and standard deviation of the numeric columns.

Hypothesis Testing

Is the average percentage of international students in UK universities higher than that in the US?

Our Hypotheses

Null Hypothesis

There is no difference between the average percentage of US and UK university students who are of international origin over the years.

$$H_o : \mu_x = \mu_y$$

Alternative Hypothesis

The average percentage of UK university students who are of international origin over the years, is higher than in the US.

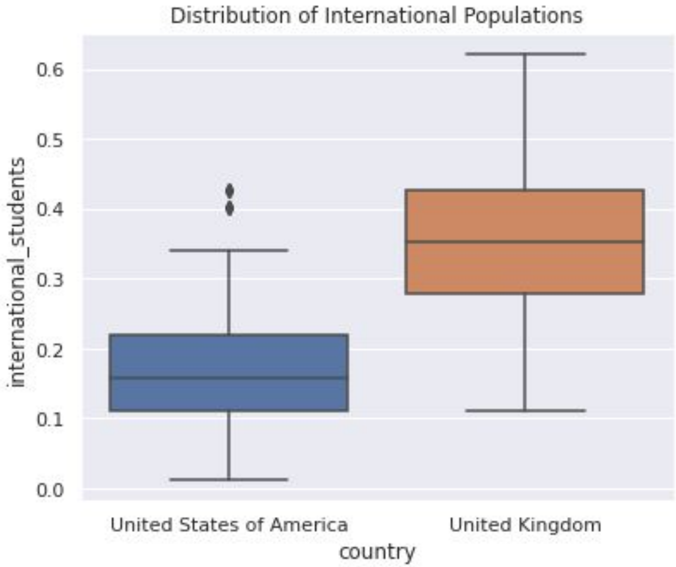
$$H_1 : \mu_x > \mu_y$$

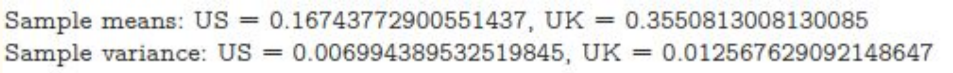
x : *Set of UK international students percentages*

y : *Set of US international students percentages*

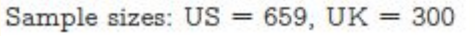
Statistical Testing

- Sample sizes are different
- Sample variances are different
- Unequal variances and sample sizes
- Sample sizes > 30. Implies approximately normal distributions

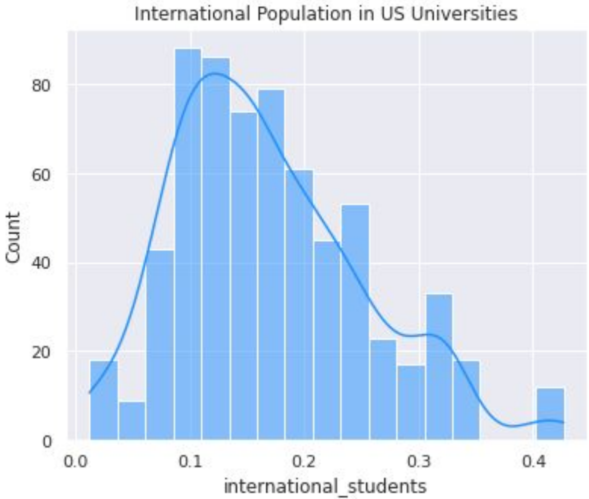


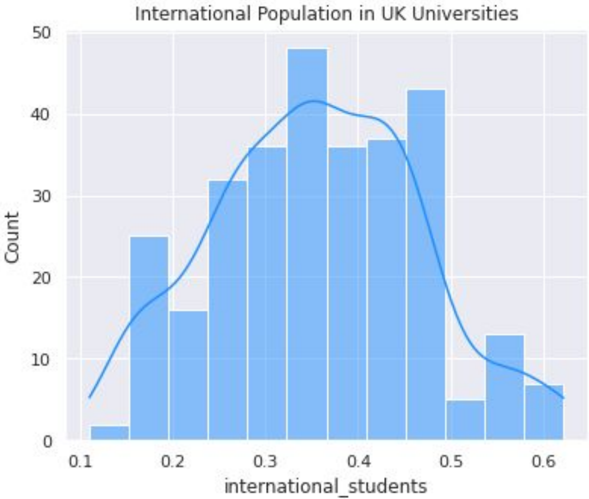


```
Sample means: US = 0.16743772900551437, UK = 0.3550813008130085
Sample variance: US = 0.006994389532519845, UK = 0.012567629092148647
```



```
Sample sizes: US = 659, UK = 300
```



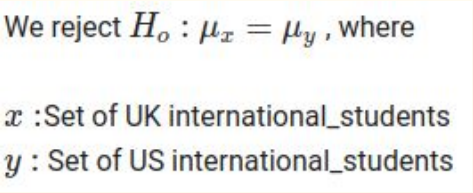


Student’s t-test cannot be predictably used.

Therefore, we use the **Welch t-test**

Observation and Conclusion

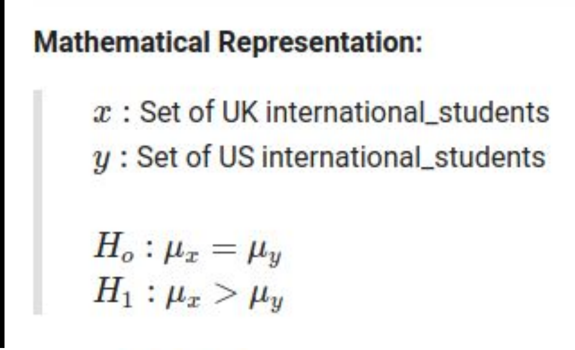
Obtained p value lesser than the considered statistical significance coefficient (alpha = 0.005)



We reject $H_o : \mu_x = \mu_y$, where

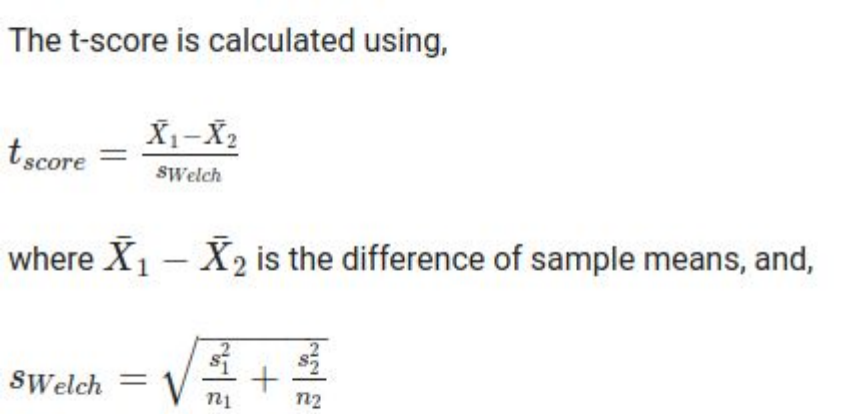
x :Set of UK international_students

y : Set of US international_students



Mathematical Representation:

> x : Set of UK international_students
>
> y : Set of US international_students
>
> $H_o : \mu_x = \mu_y$
>
> $H_1 : \mu_x > \mu_y$



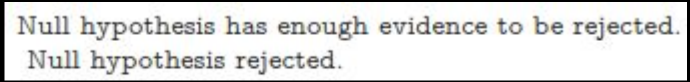
The t-score is calculated using,

$$t_{score} = \frac{\bar{X}_1 - \bar{X}_2}{s_{Welch}}$$

where $\bar{X}_1 - \bar{X}_2$ is the difference of sample means, and,

$$s_{Welch} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

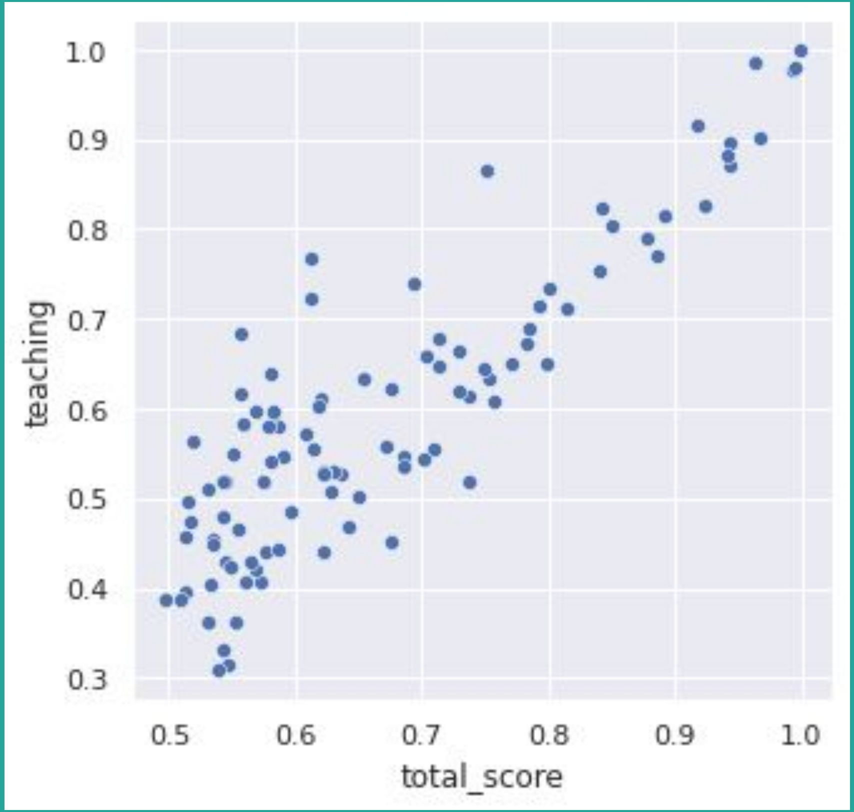
$p < 0.005$



Null hypothesis has enough evidence to be rejected.
Null hypothesis rejected.

Correlation

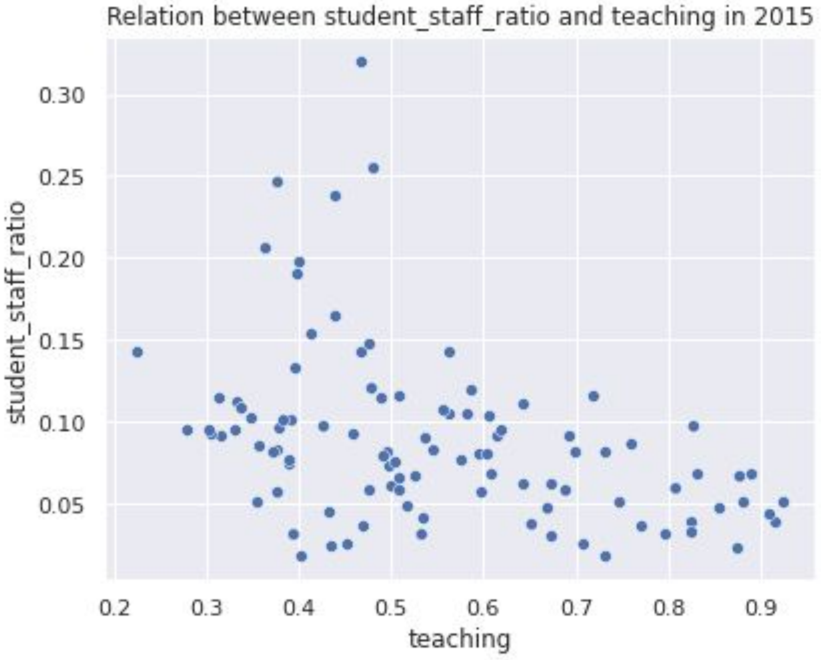
What is the correlation between teaching and total_score?

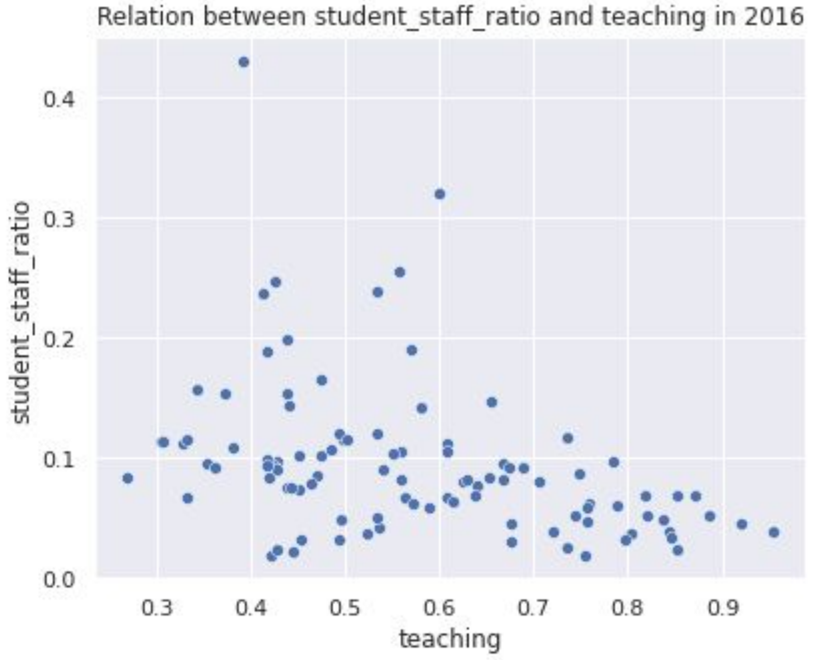


Positive

What is the correlation between student_staff_ratio and international students?

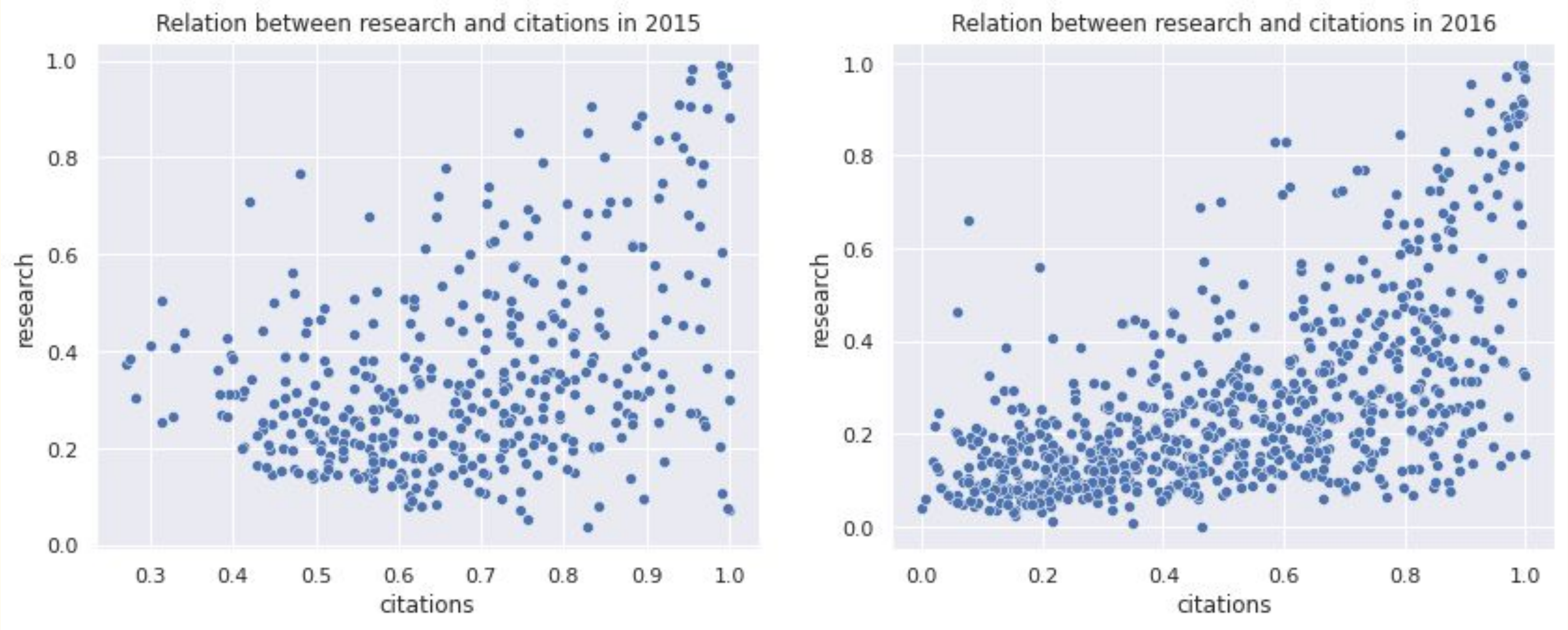
Negative





What is the correlation between research and citation in the year 2015 and 2016?

Positive



References

The Pitfalls of Data Normalization :

https://radiant-brushlands-42789.herokuapp.com/towardsdatascience.com/pitfalls-of-data-normalization-bf05d65f1f4c

How Low Can A p - Value Go? :

https://stats.stackexchange.com/questions/11812/sanity-check-how-low-can-a-p-value-go

Welch t - Testing Notes :

https://ocw.mit.edu/resources/res-6-009-how-to-process-analyze-and-visualize-data-january-iap-2012/lectures-and-labs/MITRES_6_009IAP12_lab3a.pdf