5CS037 – Concepts and Technologies of AI

Title: World University Rank 2023

Full Name: Aayush Bahadur Shahi

Student no: 2329565

Group: L5CG9

Lecturer: Mr. Siman Giri

Tutor: Mr. Basudeo Shrestha

Submission: 2023/12/24

Contents

1. Introduction

This report is all about world_university_rankings_2023. I got access to this dataset from our tutor Mr. Basudeo Shrestha sir. The size of this Dataset is 30433.There are all together 2341 rows and 13 columns. This dataset is the biggest and most varied one so far, covering 1,799 universities from 104 different countries. In total, over 680,000 data points were collected from 2,500 institutions. There are 13 columns. Analyzing Statistics and exploring data involves using the following methods:

- Pandas as pd
- Matplotlib.pyplot as plt
- Seaborn as sns
- Numpy as np

2. Data Cleaning and Summary Statistics.

1. Duplicates Values:

In this dataset there are altogether 29 duplicates values. Simply we drop all these 29 duplicates Values. After dropping all this duplicates value now there is zero duplicated values.

2. Null values

To identify null values within a dataset, we systematically go through every column, tallying and displaying the count of the null values for each. Simultaneously, we aggregate the overall count of missing values and subsequently present the total number of absent values in the dataset. If the null values are in the object column, we have drop it and display unknown, and if it is in numeric column for that we have calculated the median to fill the missing values. Then after all the missing values are filled with median.

3. Splitting Female: Male Ratio

The column Female: Male Ratio is transformed by splitting its values into two separate columns. We used the "str. split ()" method to for split. By dividing the Female: Male ratio column into two distinct columns the data become more accessible and easier to analyze as compare to what it was. After split we have converted them into float data types.

4. New column

We have created a new column named 'International Students '. In this column we have added a value by multiplying the two variables international student * no. of student. Because the International students were in percentage.

5. Dropping columns

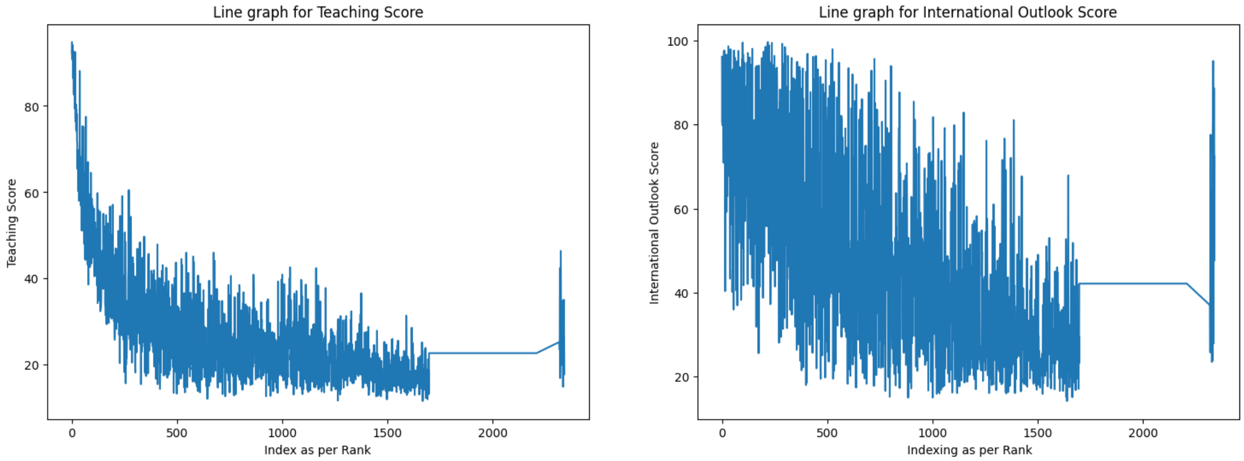
The original column Female: Male Ratio after splitting them we have drop it because we have already created an individual column for both of them. We also drop the original column international student because we have stored the values by multiplying in the new column.

2.1 Statistics Summary:

In summary for the numeric columns in the dataset. We can clearly see the all the columns count, mean, std, min, first, second, third quartile, Max. These statistics provide a comprehensive overview of the numeric attributes in the dataset. For the categorical column we have printed out the unique values and the most frequently occurring values.

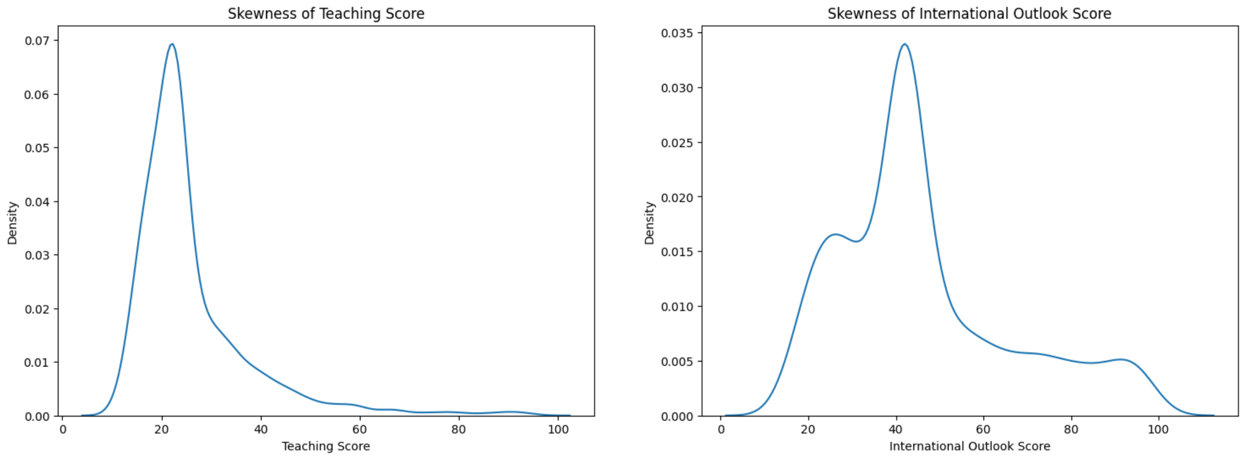
3. Data Explorations and Visualization

Univariate Analysis



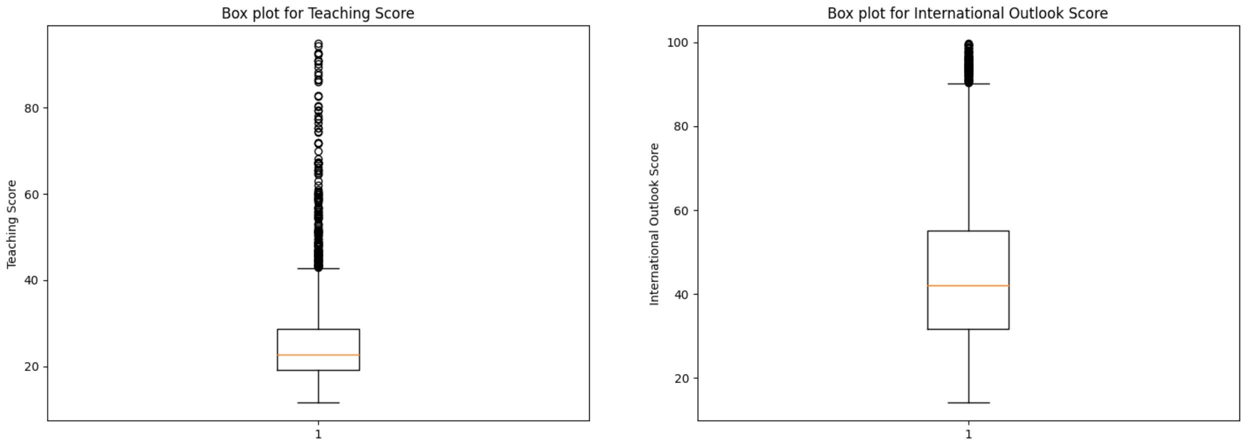
1. Line graphs

The visual output provides insights into how 'Teaching score' and international Outlook Score' change in the relation to the university rankings.



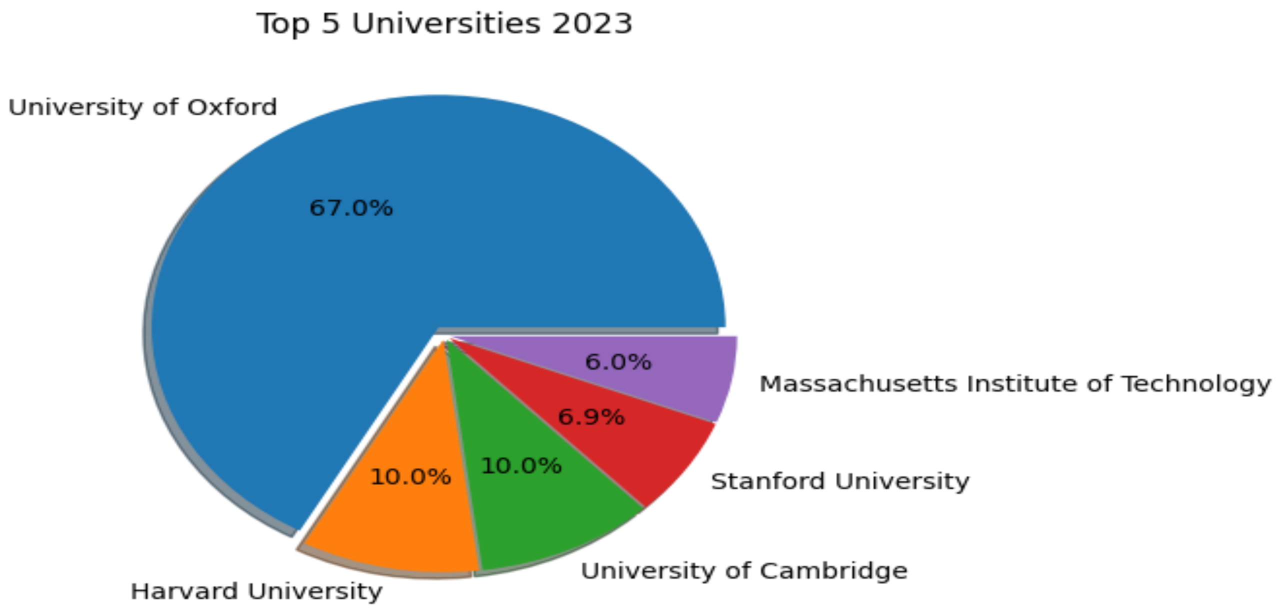
2.Skewness

This is a skewness diagram of the ‘Teaching Score’ and ‘International Outlook Score ‘in our university dataset. A positive skewness suggests that the data is stretched to the right, with a longer tail on the positive side. This implies that the mean is greater that the median, and the mode may be situated to the left of the median. Here the international outlook score has two modes.



3. Box plot

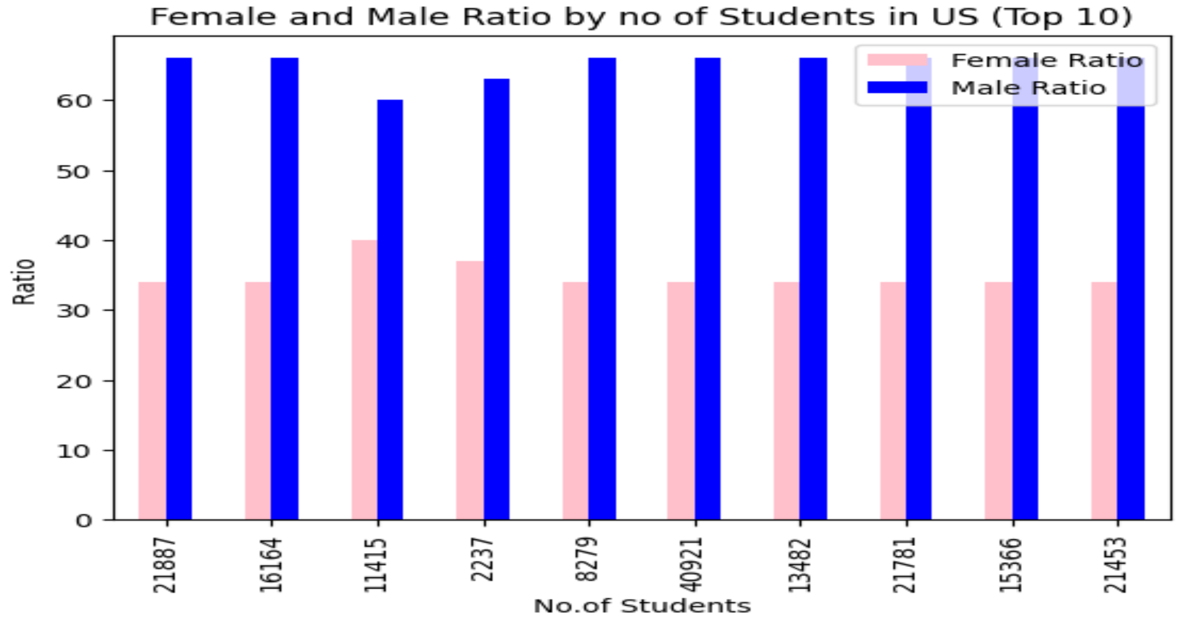
These box plots provide a visual representation of the distribution of ‘Teaching Score’ and ‘International Outlook Score,’ highlighting any potential outliers in the data.



4. Pie Chart

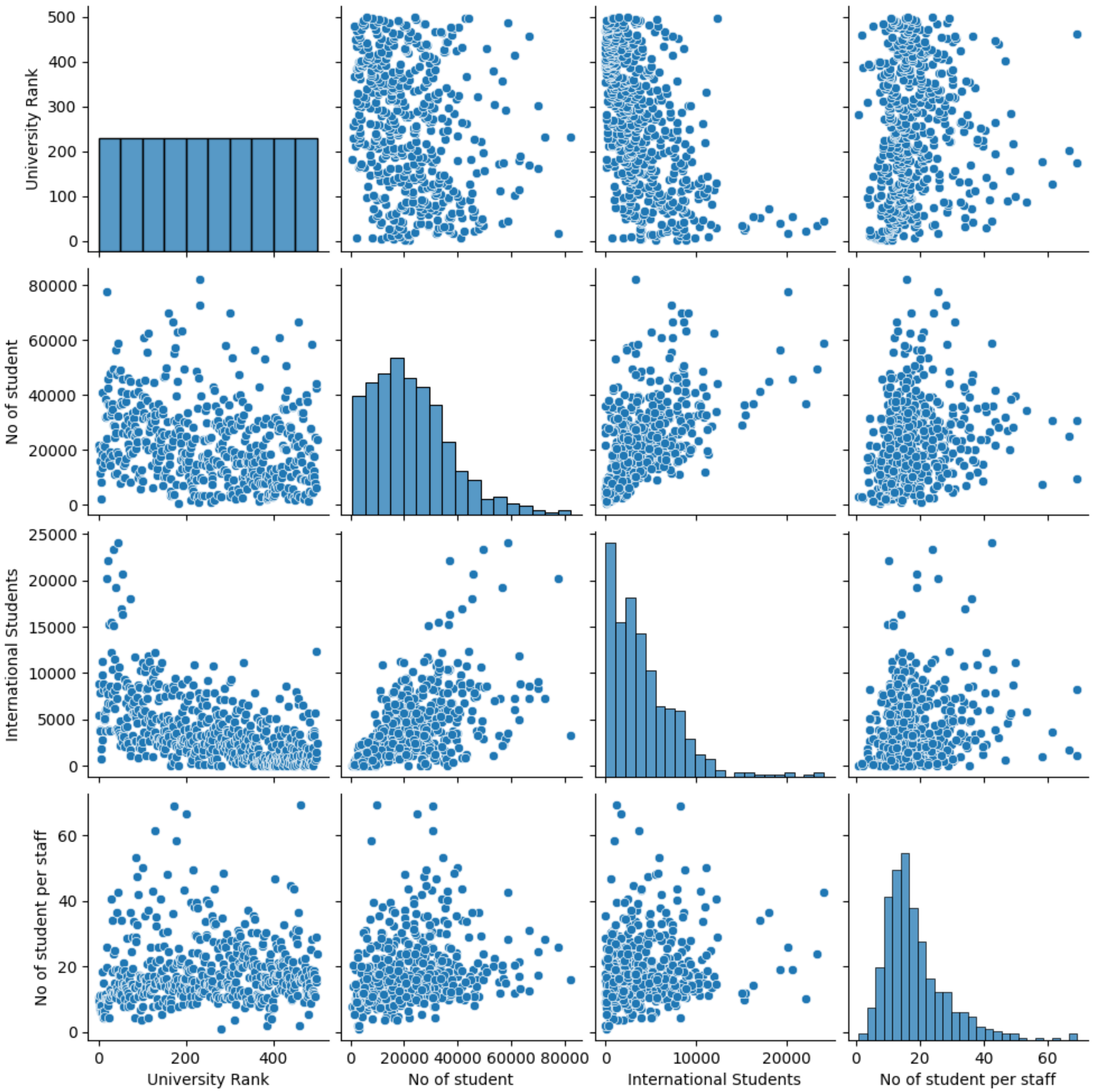
We have created a pie chart showing the distribution of ‘Industry Income Score’ among top 5 universities in 2023. Each slice of the pie represents a university, and the slice indicates how much of the total ‘Industry Income Score’ that university contributes to the top 5. This chart gives a quick visual overview of how industry income is distributed among these top universities.

Bivariate



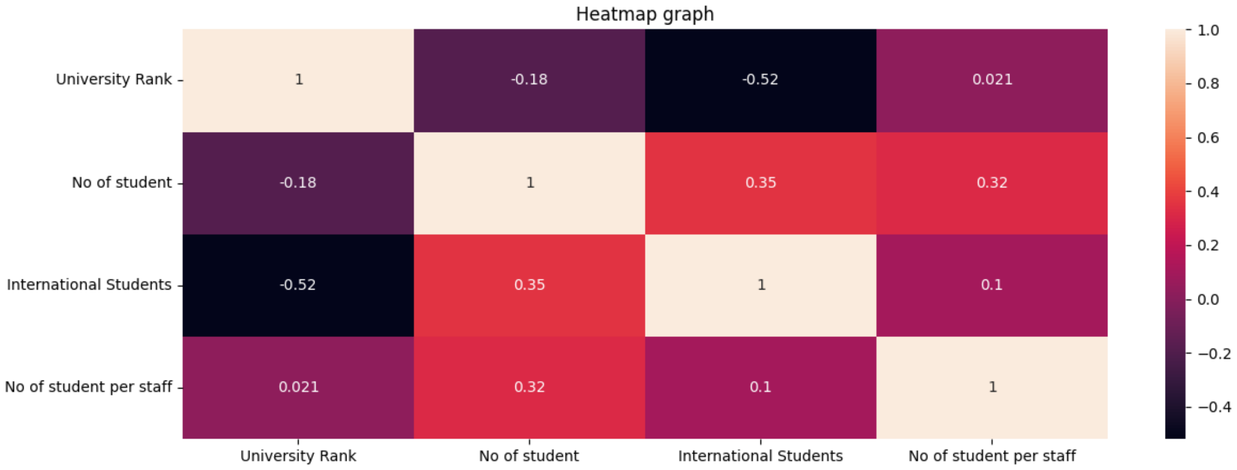
5. Bar Diagram

We have created a bar chart showing the female and male ratios for the top 10 rows of universities in the United States, based on the no of students. In the figure pink shows the girls ratio and blue shows the male ratio. This information can be useful for understanding the gender distribution in the top 10 universities in the United States.



6. Pair plot

Here we have generated a scatterplot matrix (pair plot) to visualize the relationship and correlations between ‘University Rank’, ‘No of Student per staff’, International Students’, and ‘No of student’ for the first 500 rows of the ‘uni_df_numeric’ Data frame. Each scatter plot shows the relationship between two variables, and diagonal plots shows the distribution of individual variables. This type of plot is useful for quickly finding out if there’s a relationship or trend between different pieces of information.



7.Heatmap

We create a heatmap that visually represent the correlation between the selected numerical variables. In the heatmap, each square indicates how strongly and in what direction two variables are related. This type of visualization is helpful for understanding the relationships between different features in the dataset.

Conclusion

In summary, our exploration of the 2023 world university rankings dataset involved effective data cleaning, statistical summaries, and insightful visualizations. We addressed duplicates and missing values, automated rankings, and enhanced data clarity. Visualizations uncovered trends in teaching and international scores, gender ratios, and industry income distribution.