Exercise 6.1 – Stephanie Ugwuanya

1. Data Source

The data I have chosen to explore is on World University Rankings between 2012 – 2015. This dataset explores university rankings worldwide but also the ranking within each country. Additionally, this dataset explores the quality of education, the rank for alumni employment, the quality of faculty, the rank of publications, the rank for influence and the number of students at the university.

This dataset is from Kaggle and the data has been collected externally from the sources below. We can see that these are reliable sources and data was most likely collected through surveys, research metrics, and statistical analyses:

- The Times Higher Education World University Ranking is widely regarded as one of the most influential and widely observed university measures. Founded in the United Kingdom in 2010.
- The Academic Ranking of World Universities, also known as the Shanghai Ranking, is an equally influential ranking. It was founded in China in 2003
- The Centre for World University Rankings, is a less well know listing that comes from Saudi Arabia, it was founded in 2012.

This data is relevant to this project as it meets the mandatory criteria. It consists of 2200 rows, has a mix of continuous and categoric data and contains a geographic component which is essential to this analysis.

2. Data Choice

Academia is an area I have always found interesting, and I have continued learning through structured education to this day. I thought it would be interesting to see how university rankings interact with other aspects of academia.

3. Data Profile

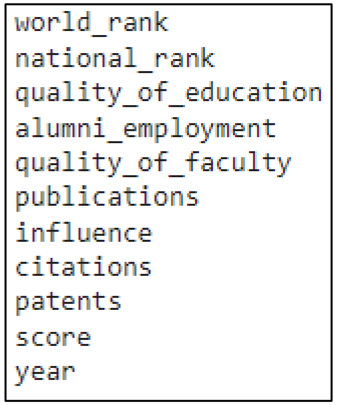
Date Shape before data cleaning: **(2200, 14)**

3.1 Finding Missing Values

There were 600 missing values in column broad_impact. I decided to remove this column altogether as it only had data for 2014 and 2015 and it could not be used to analyse all of the data.

3.2 Finding Columns with Mixed Data Types

The columns below have mixed data types, these were all corrected



```
world_rank
national_rank
quality_of_education
alumni_employment
quality_of_faculty
publications
influence
citations
patents
score
year
```

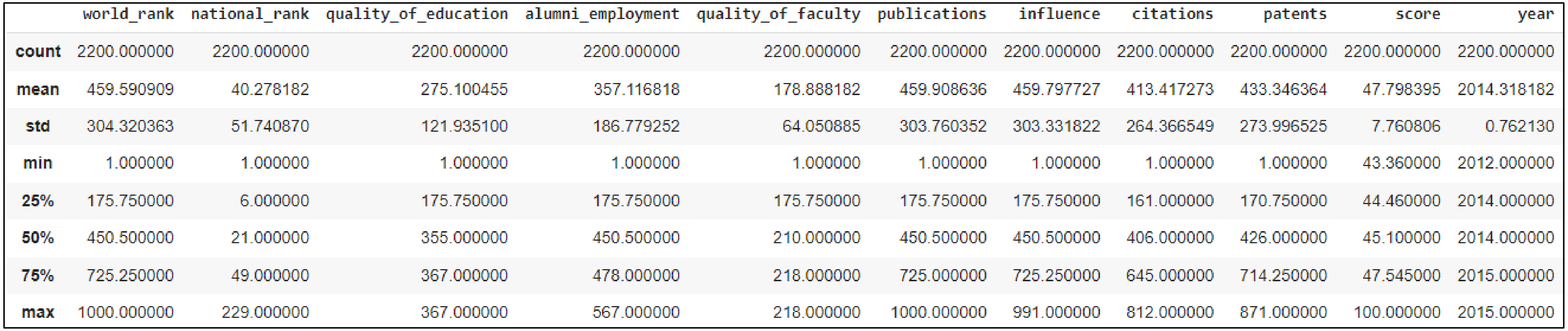
3.3. Check for consistencies in the columns

Columns checked were consistent.

Shape of Cleaned Data: **(2200, 13)**

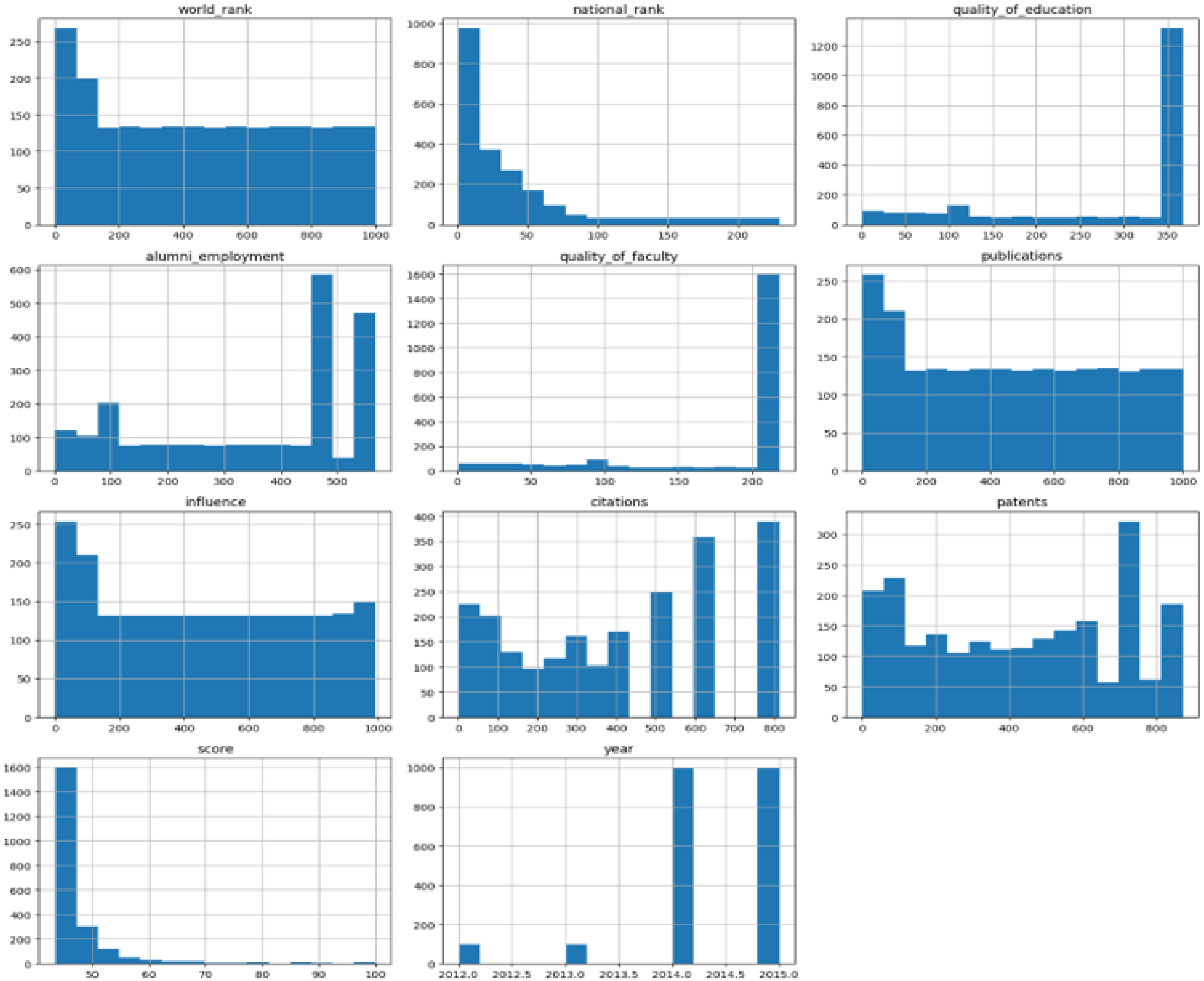
4.Descriptive Analysis

Below are the basic statistics of the cleaned data.



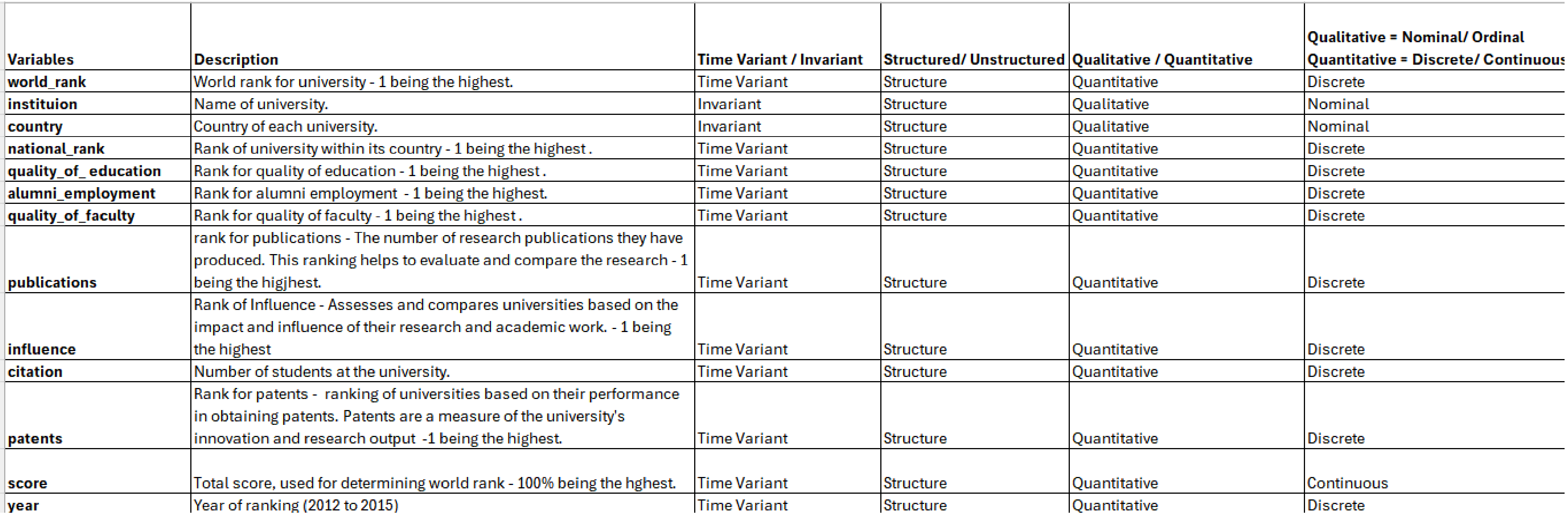
	world_rank	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	influence	citations	patents	score	year
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000
mean	459.590909	40.278182	275.100455	357.116818	178.888182	459.908636	459.797727	413.417273	433.346364	47.798395	2014.318182
std	304.320363	51.740870	121.935100	186.779252	64.050885	303.760352	303.331822	264.366549	273.996525	7.760806	0.762130
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	43.360000	2012.000000
25%	175.750000	6.000000	175.750000	175.750000	175.750000	175.750000	175.750000	161.000000	170.750000	44.460000	2014.000000
50%	450.500000	21.000000	355.000000	450.500000	210.000000	450.500000	450.500000	406.000000	426.000000	45.100000	2014.000000
75%	725.250000	49.000000	367.000000	478.000000	218.000000	725.000000	725.250000	645.000000	714.250000	47.545000	2015.000000
max	1000.000000	229.000000	367.000000	567.000000	218.000000	1000.000000	991.000000	812.000000	871.000000	100.000000	2015.000000

4.1 Frequency Tables



Frequency Tables were generated to explore the skew of data.

5. Data Descriptions and Data Types



Variables	Description	Time Variant / Invariant	Structured/ Unstructured	Qualitative / Quantitative	Qualitative = Nominal/ Ordinal Quantitative = Discrete/ Continuous
world_rank	World rank for university - 1 being the highest.	Time Variant	Structure	Quantitative	Discrete
instituion	Name of university.	Invariant	Structure	Qualitative	Nominal
country	Country of each university.	Invariant	Structure	Qualitative	Nominal
national_rank	Rank of university within its country - 1 being the highest .	Time Variant	Structure	Quantitative	Discrete
quality_of_ education	Rank for quality of education - 1 being the highest .	Time Variant	Structure	Quantitative	Discrete
alumni_employment	Rank for alumni employment - 1 being the highest.	Time Variant	Structure	Quantitative	Discrete
quality_of_faculty	Rank for quality of faculty - 1 being the highest .	Time Variant	Structure	Quantitative	Discrete
publications	rank for publications - The number of research publications they have produced. This ranking helps to evaluate and compare the research - 1 being the higjhest.	Time Variant	Structure	Quantitative	Discrete
influence	Rank of Influence - Assesses and compares universities based on the impact and influence of their research and academic work. - 1 being the highest	Time Variant	Structure	Quantitative	Discrete
citation	Number of students at the university.	Time Variant	Structure	Quantitative	Discrete
patents	Rank for patents - ranking of universities based on their performance in obtaining patents. Patents are a measure of the university's innovation and research output -1 being the highest.	Time Variant	Structure	Quantitative	Discrete
score	Total score, used for determining world rank - 100% being the hghest.	Time Variant	Structure	Quantitative	Continuous
year	Year of ranking (2012 to 2015)	Time Variant	Structure	Quantitative	Discrete

I will be creating flags and grouping data to create profiles e.g. grouping countries by continent and university rankings.

6. Limitations and Ethics

The database on universities rankings was compiled by a number of reputable sources however there are a number of bias that may be present in the data due to how it was collected. These have already been outlined on the data card on Kaggle.

6.1 Bias

- **Commercialisation Bias:**
  Times Higher Education (THE): The ranking system has been criticised for its commercialisation meaning that the rankings may prioritising universities that align with their business interest, potentially skewing results.

- **Regional Bias**:
  Shanghai Ranking (ARWU): This ranking, founded in China, might have inherent biases favouring institutions with strong research outputs in certain scientific disciplines, possibly reflecting regional priorities in research and education.

- **Data Collection and Self-Reporting Bias:**
  As outlined earlier it likely some of this data was collected through surveys. Universities may provide data that is more positive to improve their reputation. Differences in data collection methods and the reliability of self-reported data can introduce inconsistencies and biases in the rankings.

6.2 Limitations/ Ethical Considerations.

One of the main limitations discovered from the description analysis is there is much more information on universities in 2014 and 2015 than 2012 and 2013 meaning that we may not be able to see the change in some universities' ratings over time as data is not complete for all universities. However, The Data Card has been very transparent in presenting the sources used to collect this data and has been very open on possible bias with the data. There is also equality as there are a wide range of universities included from across the world. Additionally, there is data privacy as no PII data has been included.

Questions

There are some focus areas to the question generated - university rankings, quality of education and universities relationship with innovation/ quality of education.

Do universities with a higher-ranking produce student that are more likely to obtain alumni employment?

Does the number of students at a university effect the quality of education? Does the quality of education vary around the world?

Do universities that invest more in innovation produce graduates that are more employable?

Does the ranking of universities vary over time?