THE

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column Name | Description | Data Type | Notes | Changes  Considered | Column Changes Made |
| world\_rank | World rank for university | Object | After 200, rankings change to groupings. Ex. 201-225 | Consider removing rankings below 200. Change data type to integer. |  |
| university\_name | Name of university | Object | Some university names had special characters that were converted to symbols. | Need to match university names to other data sets and remove special characters |  |
| country | Name of country | Object | Need to change to match json geo file data. |  | Does not exist in ARWU. Need to create column to match. |
| teaching | University score the learning environment | Float64 | Is to one decimal place. No missing values |  | THE\_teaching |
| international | University score for international outlook including staff, students, and research | object | Is to one decimal place. A – is used for missing values. Nine missing values | Find and replace – with nothing so can be checked for missing values. Change object type to float64. Evaluate missing values. | THE\_international |
| research | University score for research based on volume, income, and reputation | Float64 | Is to one decimal place. No missing values |  | THE\_research |
| citations | University score for citations showing the influence of its research | Float64 | Is to one decimal place. No missing values |  | THE\_citations |
| income | University score for industry income and knowledge transfer | Object | Is to one decimal place. A – is used for missing values. 218 missing values. | Find and replace – with nothing so can be checked for missing values. Change object type to float64. Evaluate missing values. | THE\_income |
| total\_score | Score used to determine ranking | Object | Is to one decimal place. A – is used for missing values. 1,402 missing values. | Find and replace – with nothing so can be checked for missing values. Change object type to float64. Evaluate missing values. |  |
| num\_students | Number of students at the university | Object | Numeral. There are 59 missing values | Change object type to int. Need to remove comma in numbers. Evaluate missing values | No match in other database. Not used to calculate score. Delete column. |
| student\_staff\_ratio | Number of students divided by number of staff | Float64 | Is to one decimal place. 59 missing values. Same as missing student numbers. | Evaluate missing values. | No match in other database. Not used to calculate score. Delete column. |
| international\_students | Percentage of students who are international | Object | A percentage with no decimal places. 67 missing values | Change percentages to decimals. Change object type to int. Evaluate missing values. | No match in other database. Not used to calculate score. Delete column. |
| female\_male\_ratio | Female student to male student ratio | Object | This column was converted to date and time by Excel upon downloading. This column is not needed for analysis as it is not part of the other data sets. | Delete column. | No match in other database. Delete column. |
| year | Calendar year | Int64 | 2011-2016 | Change object to date. | Match  Keep same |

CWU

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column Name | Description | Data Type | Notes | Changes Considered | Column Changes Made |
| world\_rank | World rank for university | Int64 | After 100 only have data for two years 2014 and 2015 | Consider removing rankings below 100. Change data type to integer. |  |
| institution | Name of university | Object | Some university names had special characters that were converted to symbols. | Am waiting to change special character until deciding which data to keep. | university\_name |
| country | Name of country | Object | Looks clean |  |  |
| national\_rank | Rank of university within its country | Int64 | Is a numeral with no decimal places. No missing data. |  |  |
| quality\_of\_education | Rank for quality of education | Int64 | Is a numeral with no decimal places. No missing data. |  | CWU\_quality\_education |
| alumni\_employment | Rank for alumni employment | Int64 | Is a numeral with no decimal places. No missing data. |  | CWU\_alumni\_employment |
| quality\_of\_faculty | Rank for quality of faculty | Int64 | Is a numeral with no decimal places. No missing data. |  | CWU\_quality\_of\_faculty |
| publications | Rank for publications | Int64 | Is a numeral with no decimal places. No missing data. |  | CWU\_publications |
| influence | Rank for influence | Int64 | Is a numeral with no decimal places. No missing data. |  | CWU\_influence |
| citations | Rank for number of citations | Int64 | Is a numeral with no decimal places. No missing data. |  | CWU\_citations |
| broad\_impact | Rank for broad impact. | Float64 | Only available for 2014 and 2015. 200 missing. | Evaluate missing values. | CWU\_broad\_impact |
| patents | Rank for patents | Int64 | Is a numeral with no decimal places. No missing data. |  | CWU\_patents |
| score | Total score used to determine ranking | Float64 | Is a numeral with two decimal places. No missing data. |  | total\_score |
| year | Calendar year | Int64 | 2012-2015 | Change object to date. |  |

ARWU

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column Name | Description | Data Type | Notes | Changes to consider | Column Changes Made |
| world\_rank | World rank for university | Object | Contains rank ranges and equal ranks. Starting with 101 uses groupings.Ex. 101-152 | Consider removing rankings below 100. Change data type to integer. |  |
| university\_name | Name of university | Object | Some university names had special characters that were converted to symbols. One missing value. | Consider how to match after deciding which data to keep. Evaluate missing value. | One row had all variable missing. Row was deleted. University names had to be standardized across all datasets. |
| national\_rank | Rank of university within its country | Object | Number with no decimal places. One missing value. | Evaluate missing value. Rows that were aggregated based on world ranking were all aggregated here. Some of the aggregations overlap. There was also an error on input that place the aggregations beginning with 1-12 into date format. | One row had all variable missing. Row was deleted. Used formula TEXT(C1630, "M-D") to convert back to correct format and then pasted special. |
| total\_score | Score used to determine rank | Float64 | Is to one decimal place. There are 3,796 missing values. | Extreme number of missing values. |  |
| alumni | Based on number of alumni of an institution winning nobel prize or field medals. | Float64 | Is to one decimal place. One missing value. | Evaluate missing value. | ARWU\_alumni\_prizes |
| award | Score based on the number of staff winning nobel prizes in physics, chemistry, medicine, and economics and field medals in math. | Float64 | Is to one decimal place. Two missing values. | Evaluate missing value. | ARWU\_awards |
| hici | Score based on the number of Highly Cited Researchers selected by Thomson Rueters | Float64 | Is to one decimal place. Two missing values. | Evaluate missing values. | ARWU\_highly\_cited |
| ns | Based on numbers of papers published in nature and science | Float64 | Is to one decimal place. 22 missing values. | Evaluate missing values. | ARWU\_papers\_published |
| pub | Score based on number of papers indexed in the Science Citation Index-Expanded and Social Science Citation Index | Float64 | Is to one decimal place. Two missing values. | Evaluate missing values. | ARWU\_papers\_indexed |
| pcp | Weighed scores of the five indicators divided by the number of full time academic staff | Float64 | Is to one decimal place. Two missing values. | Evaluate missing values. | Remove column. There is no comparable data in other datasets. |
| year | Calendar year | Int64 | 2005-2015 | Convert to date. |  |