Task 6.1: Sourcing Open Data
Samuel Callender

- **Data Source:** The data is available from an open-source platform, Kaggle.com, via the following link: World University Rankings. This data set was chosen because I have a genuine interest in higher education, specifically, the factors that propel a university to a higher world ranking. The main objective is to perform an analysis that could aid prospective university students deciding on where they should attend college and why.

- **Data Collection and Contents:** Ranking universities is a difficult, political, and controversial practice. There are hundreds of different national and international university ranking systems, many of which disagree with each other. This dataset contains three global university rankings from very different places. The rankings include *Times Higher Education World University Ranking*, *Academic Ranking of World Universities*, and *The Center for Word University Rankings*. The CWUR (Center for World University Rankings) data samplings do not come from surveys and university data submissions. Instead, the rankings rely more on outcome-based samplings.

- **Data Consistency and Cleaning:**
  - Shape of the Data: 2200 Rows and 14 Columns.
  - While there is some missing data, there is not enough to affect the analysis.
  - No duplicates found in the data.
  - Overall, a clean data set.

```
In [6]: df.describe()
```

Out[6]:

	world_rank	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	influence	citations	broad_impact	patent
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2000.000000	2200.00000
mean	459.590909	40.278182	275.100455	357.116818	178.888182	459.908636	459.797727	413.417273	496.699500	433.34636
std	304.320363	51.740870	121.935100	186.779252	64.050885	303.760352	303.331822	264.366549	286.919755	273.99652
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.00000
25%	175.750000	6.000000	175.750000	175.750000	175.750000	175.750000	175.750000	161.000000	250.500000	170.75000
50%	450.500000	21.000000	355.000000	450.500000	210.000000	450.500000	450.500000	406.000000	496.000000	426.00000
75%	725.250000	49.000000	367.000000	478.000000	218.000000	725.000000	725.250000	645.000000	741.000000	714.25000
max	1000.000000	229.000000	367.000000	567.000000	218.000000	1000.000000	991.000000	812.000000	1000.000000	871.00000

```
In [7]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2200 entries, 0 to 2199
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   world_rank            2200 non-null   int64
 1   institution           2200 non-null   object
 2   country               2200 non-null   object
 3   national_rank         2200 non-null   int64
 4   quality_of_education  2200 non-null   int64
 5   alumni_employment     2200 non-null   int64
 6   quality_of_faculty    2200 non-null   int64
 7   publications          2200 non-null   int64
 8   influence             2200 non-null   int64
 9   citations             2200 non-null   int64
 10  broad_impact          2000 non-null   float64
 11  patents               2200 non-null   int64
 12  score                 2200 non-null   float64
 13  year                  2200 non-null   int64
dtypes: float64(2), int64(10), object(2)
memory usage: 240.8+ KB
```

- **Data Profile**

COLUMN	DESCRIPTION	TIME VARIANT/ INVARIANT	DATA TYPE
world_rank	World Rank for the University (1-1000)	INVARIANT	QUANTITATIVE
institution	Name of the University	INVARIANT	QUALITATIVE
country	Country for each University	INVARIANT	QUALITATIVE
national_rank	Rank of the University within its' Country (1-229)	INVARIANT	QUANTITATIVE
quality_of_education	Rank for Quality of Education (1-367)	INVARIANT	QUANTITATIVE
alumni_employment	Rank for Alumni Employment (1-567)	INVARIANT	QUANTITATIVE
quality_of_faculty	Rank for Quality of Faculty (1-218)	INVARIANT	QUANTITATIVE
publications	Rank for Publications (1-1000)	INVARIANT	QUANTITATIVE
influence	Rank for Influence (1-991)	INVARIANT	QUANTITATIVE
citations	Rank of the University for high number of Citations (1-812)	INVARIANT	QUANTITATIVE
broad_impact	Rank for the University with high Broad Impact (1-1000)	INVARIANT	QUANTITATIVE
patents	Rank based on the Number of Patents (1-871)	INVARIANT	QUANTITATIVE
score	General Score of the University (43.4-100)	INVARIANT	QUANTITATIVE
year	Time frame when the Criteria was analyzed	VARIANT	QUANTITATIVE

- **Limitation and Ethics**
  - While the data that comes from these rankings is limited to three, *Times Higher Education World University Ranking*, *Academic Ranking of World Universities*, and *The Center for Word University Rankings,* this does provide a more diverse collection of sources versus only one primary source. It should be noted that *Times Higher Education World University Ranking* has been criticized for its commercialization and for undermining non-English-instructing institutions. Also, *Academic Ranking of World Universities* has been criticized for focusing on raw research power and for undermining humanities and quality of instruction. There is the potential for some bias, though limited.
  - One limitation is that the data being analyzed is only for the time frame 2012-2015. The limited time frame and data that is a few years old are limitations.
  - There is no personal information in this data set, therefore there is no risk for personal information to be exposed.

- **Questions to Explore**
    - Which countries have the most ranked universities?
    - What are the highest ranked universities broken down by country?
    - What are the top or main qualities of the highest ranked universities?
        - Alumni faculty?
        - Quality of Education?
        - Quality of the faculty?
        - Number of Publications?
    - Do the university rankings drastically change over time?