Final Project Step 2

DSC 520 Week 10

Blandon Lee

02/18/2022

Data importing and cleaning steps are explained in the text and follow a logical process. Outline your data preparation and cleansing steps.

For the project I will be importing the cleaned data to help answer the questions I am interested in. The questions that I have outlines in step one are listed below.

- What impacts the ranking the most?
- Which country has the highest ranking?
- Which establishment has been consistently ranked?
- Does a better ranking affect the quality of education?
- How does rankings translate to student employment?
- Does male to female ratios impact rankings and or education quality?

To gain useful insights into the proposed questions I have narrowed down the main variables that I plan to use from each dataset.

Main Variables:

- Score
- Institution
- Citations
- World_rank

I also want to explore other variables in the datasets that may be beneficial in the project and help to answer the questions.

Additional Variables:

- quality_of_education,
- teaching,
- research.

Cleaning Data

1. Ensure all overlapping variables in the datasets share the same spellings and or characters.

2. Address missing data in the datasets.
    a) na.omit was used
    b) NA variables were not a concern because they appeared in data that will not impact the outcome. So, removing them didn't present a concern.

With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

Dataset 1: CWUR.2020

```
str(CWUR.2020)
'data.frame':  2000 obs. of  9 variables:
 $ world_rank          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ institution         : chr  "Harvard University" "Massachusetts Institute of Tech"..
 $ country             : chr  "USA" "USA" "USA" "United Kingdom" ...
 $ national_rank       : int  1 2 3 1 2 4 5 6 7 8 ...
 $ quality_of_education: chr  "2" "1" "9" "4" ...
 $ alumni_employment   : chr  "1" "10" "3" "19" ...
 $ quality_of_faculty  : chr  "1" "2" "3" "5" ...
 $ research_performance: chr  "1" "5" "2" "11" ...
 $ score               : num  100 96.7 95.2 94.1 93.3 92.6 92 91.6 91.1 90.7 ...
```

Dataset 2: timesData

```
str(timesData)
'data.frame':  2603 obs. of  14 variables:
 $ world_rank            : chr  "1" "2" "3" "4" ...
 $ institution           : chr  "Harvard University" "California Institute of Tech"...
 $ country               : chr  "United States of America" "United States of America"...
 $ teaching              : num  99.7 97.7 97.8 98.3 90.9 90.5 88.2 84.2 89.2 92.1 ...
 $ International_Outlook : chr  "72.4" "54.6" "82.3" "29.5" ...
 $ research              : num  98.7 98 91.4 98.1 95.4 94.1 93.9 99.3 94.5 89.7 ...
 $ citations             : num  98.8 99.9 99.9 99.2 99.9 94 95.1 97.8 88.3 91.5 ...
 $ income                : chr  "34.5" "83.7" "87.5" "64.3" ...
 $ score                 : chr  "96.1" "96.0" "95.6" "94.3" ...
 $ num_students          : chr  "20,152" "2,243" "11,074" "15,596" ...
 $ student_staff_ratio   : num  8.9 6.9 9 7.8 8.4 11.8 11.6 16.4 11.7 4.4 ...
 $ international_students: chr  "25%" "27%" "33%" "22%" ...
 $ female_male_ratio     : chr  "" "33 : 67" "37 : 63" "42 : 58" ...
 $ year                  : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
```

Dataset 3: University_world_rank_Data

```
str(University_world_rank_Data)
'data.frame':  2200 obs. of  13 variables:
 $ world_rank          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ institution         : chr  "Harvard University" "Massachusetts Institute of Tech"...
 $ country             : chr  "USA" "USA" "USA" "United Kingdom" ...
 $ national_rank       : int  1 2 3 1 4 5 2 6 7 8 ...
 $ quality_of_education: int  7 9 17 10 2 8 13 14 23 16 ...
 $ alumni_employment   : int  9 17 11 24 29 14 28 31 21 52 ...
 $ quality_of_faculty  : int  1 3 5 4 7 2 9 12 10 6 ...
 $ publications        : int  1 12 4 16 37 53 15 14 13 6 ...
 $ influence           : int  1 4 2 16 22 33 13 6 12 5 ...
 $ citations           : int  1 4 2 11 22 26 19 15 14 3 ...
 $ patents             : int  5 1 15 50 18 101 26 66 5 16 ...
 $ score               : num  100 91.7 89.5 86.2 85.2 ...
 $ year                : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
```

Dataset 4: world.university.ranking.2020

```
str(world.university.rankning.2020)
'data.frame':  1396 obs. of  15 variables:
 $ world_rank             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ institution            : chr  "University of Oxford" "California Institute of Tech"...
 $ country                : chr  "United Kingdom" "United States" "United Kingdom"...
 $ Number_students        : chr  "20,664" "2,240" "18,978" "16,135" ...
 $ Numb_students_per_Staff: num  11.2 6.4 10.9 7.3 8.6 8.1 9.2 5.4 5.7 11.7 ...
 $ international_students : chr  "41%" "30%" "37%" "23%" ...
 $ Percentage_Female      : chr  "46%" "34%" "47%" "43%" ...
 $ Percentage_Male        : chr  "54%" "66%" "53%" "57%" ...
 $ teaching               : num  90.5 92.1 91.4 92.8 90.5 90.3 89.2 92 89.1 84.5 ...
 $ research               : num  99.6 97.2 98.7 96.4 92.4 96.3 98.6 94.8 91.4 87.6 ...
 $ citations              : num  98.4 97.9 95.8 99.9 99.5 98.8 99.1 97.3 96.7 97 ...
 $ Industry_Income        : num  65.5 88 59.3 66.2 86.9 58.6 47.3 52.4 52.7 69.9 ...
```

What do you not know how to do right now that you need to learn to import and cleanup your dataset?

Two variables (country and institution) present a challenge for me. What I need to figure out is converting words to numbers to use a logistic regression. I can go into Excel and do it manually, but that is quite time consuming. However, if it comes down to it, I plan to change the United States to a 1 and all other countries to a 0 and something similar for institutions.

Discuss how you plan to uncover new information in the data that is not self-evident.

With the variables (country and institution) being categorical and not measurable I have to find a way to compare those with other variables. This would allow me additional insights into actual locations education quality and not be as focused on ranking. This would be a good area for a multiple regression to compare more than one variable in the dataset.

What are different ways you could look at this data to answer the questions you want to answer?

The best approach I believe will be visual representation and or summary statistics. This will better convey the story and answer the questions because where one falls short the other is there to push forward.

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

I had planned to join the datasets together, but that is working out to be a difficult task and one that I will need to figure out prior to deadline. I found that some of the variables are the same, but the information they represent are different. I managed to merge the datasets into two, but I was initially aiming for one. I will be using them for the summary statistics and visuals and then kind of compare them.

How could you summarize your data to answer key questions?

My plan is to use the summary statistics to answer key questions, but also use tables and graphs/plots. I will then be able to gather the information and place them in a reader friendly format.

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend, if necessary, scales are appropriate, appropriate geoms used, etc.).

- neuromas scatterplots
- histograms
- regression lines
- maybe other if I feel they better display the information

What do you not know how to do right now that you need to learn to answer your questions?

Merging all the datasets into one would be quite useful. In my mind working with one data set will be easier to analyze. I feel that it me come back to bite me, but time will tell.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Using some nearest neighbor and or k-means clustering would be useful, but I have not really decided yet. I think that checking for accuracy would be quite important and is something I would like top incorporate.