Smart Certificate - Case study - R for Data Science

*** Considered an updated dataset from the Forbes Website, as I could not access JLC Lab properly ***

*** The data and the answers might differ a bit, because the data was updated ***

Instructions

Please Note: Access for graded assignment is restricted to 1 attempt, no additional attempt will be provided for these assignments.

Use the dataset named **The World's Most Valuable Brands List - Forbes.html**, placed at **C\DS Full stack\Assignments\Graded Assignment\ R for Data Science** this is a dataset about the most valuable brands across globe. Suppose you work in the data journalism division of a media house. You are tasked to help a team of content marketers, in extracting and cleaning data from this html page. The data that we are interested in, is the table inside this html page that details about the brand, its rank, valuation, revenue, industry etc. With this context in mind, answer the questions asked below.

Questions & Answers

1. (a) Use readHTMLTable() function from XML library in R to read this html file inside R. If you use class() function on this object, the output is _____________

(Hint: Just mention output, without any console details, make sure you do include quotation marks, if your output shows quotes)

Answer:

data.frame

1. (b) How many elements are inside the object you just read in?

Answer:

109   8

1. (c) When you convert the relevant html table into a dataframe, how many columns do you obtain for this dataframe

Answer:

8

2. Now that you have the relevant data in the form of a dataframe, answer the following questions, keeping in mind you have the correct dataframe with relevant data

(a) In this dataframe, is there a column called “Brand_Revenue”?

Answer:

False (*Brand Revenue is present, without the _*)

2. (b) The data type of “Brand Value”, column of this dataframe is numeric.

Answer:

False (chr)

2. (c) How many unique values are there in the column “Industry”?

Answer:

18

2.(d) How many rows in the data for the column Industry, has a value “Automotive”?

Answer:

12

3. Once you explore the dataframe, containing, relevant information, its time, that you start cleaning this dataframe. Keeping this in mind, answer the following questions

(a) For the column, called “Company Advertising” you can see that some observations have units in Millions of dollars, while some are recorded in Billions of dollars. Count how many observations in this column are recorded in Millions of dollars.

Answer:

23

3. (b) In this table, there are many rows, across many columns, where, the values are missing, but aren’t being treated as NA values. Take an appropriate action to resolve this issue., and fill in the blanks:

Number of NA values in Company Advertising column is__________

Answer:

9

3. (c) After dropping the rows with missing values (be careful, there are some columns in your data with missing values across all, rows, drop these columns, before you drop rows with missing values), we end up with……………………………… rows in the dataframe.

Answer:

21.52

3. (d) Now, coming back to the column, talking about company advertisement, there are some rows in the data where the observations are made in the units of millions of dollars. You need to normalize this data and make sure all the rows in this column are measuring data in billions of dollars. After you normalize and clean this column, the average value of this column is (round your answer to 2 decimal places using round() function)?

Answer:

182.63