DA5020 – Assignment 7

This assignment requires that you collect embedded data from a website and import them into R. You will scrape data from HTML using the rvest library. You will work with the following data that was obtained from Wikipedia that contains a list of countries by percentage of population living in poverty.

Question 1 — (50 points)

In this question, you will use rvest to parse the HTML and extract the tabular data on the "Percent of population living on less than $1.90, $3.20 and $5.50 a day" from the Wikipedia page.

1. (10 pts) Scrape the data from the webpage and extract the following fields: **Country**, < **$1.90, < $3.20, < $5.50**, **Year** and **Continent**. Prepare the data for analysis and ensure that the columns have meaningful names.

2. (10 pts) Calculate the mean and the standard deviation of the percent of the population living under $5.50 per day for each continent. Perform a comparative analysis (i.e. explanation) of the data from each continent.

3. (5 pts) What are the 10 countries with the highest percentage of the population having an income of less than $5.50 per day? Using a suitable chart, display the country name, the percentage and color-code by the Continent. Summarize your findings.

4. (5 pts) Explore the countries with the lowest percentage of the population having an income of less than $5.50 per day. What are the 5 countries with the lowest percentage, and how does the results compare to the other income groups (i.e. $1.90 and $3.20)?

5. (20 pts) Extract the data for any two continents of your choice. Fore each continent, visualize the percent of the population living on less than **$1.90**, **$3.20** and **$5.50** using box plots. Compare and contrast the results, while ensuring that you discuss the distribution, skew and any outliers that are evident.

Submission Details

- Your submission must contain two files: the .Rmd file, and 2) a knitted PDF or HTML. Name each respective file as follows: DA5020.A7.FirstName.LastName.Rmd and your PDF/HTML DA5020.A7.FirstName.LastName.{pdf,html}, where *FirstName.LastName* is your first and last name.

- The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it). Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.

- Not submitting a knitted PDF or HTML will result in reduction of 30 points.

- Not submitting the .Rmd file (or both) will result in a score of 0.

Useful Resources

- List of countries by percentage of population living in poverty

- Kingl, A. (Feb 7, 2018). Web Scraping in R: rvest Tutorial. Datacamp Community.

- Kaushik, S. (Mar 27, 2017). Beginner's Guide on Web Scraping in R (using rvest) with hands-on example. Analytics Vidhya.