HOUSE OF INNOVATION

PRACTICE 5: MORE ADVANCED CLEANING

7316 - INTRODUCTION TO DATA ANALYSIS WITH R

In this assignment, you will create your own dataset by scraping a website, manipulating strings, merging the dataset with another dataset, and writing your own routine to manipulate the data.

1. Go to the webpage https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_area. It provides land and water areas per country. Save the URL as a string.
2. Use `read_html()` and `html_table()` from the `rvest` package to extract the table with the reviews (use the header option).
3. View the structure of the table.
4. `df` contains a list of 4 objects. The table containing the data we are interested in is the second object. Extract it from the list.
5. Drop the first and the last columns of the table. They contain no relevant information.
6. The column `Totalin km2 (mi2)` contains the area in square km, and in parenthesis, the area in square miles. Separate the information into 2 columns. Make sure to use the appropriate names for each column.
7. The total areas (km^2 and mi^2) are recorded as `character`. Convert them to `integer`.
8. Create a function to do the same for `Landin` and `Waterin`: split the areas in km^2 and mi^2, convert the results into integers, and return it to new columns in the `data.frame`.
9. Create a loop that runs your function on the two variables.
10. The variables `Country / dependency` and `%water` contain special characters in their name. Rename them to follow the guidelines of Module 1.
11. Ensure the variable `Country / dependency` contains no space upfront or at the end of the string.
12. For all the country names, make sure they are written consistently, remove multiple spaces, and replace invisible characters with blanks, if any. Convert the country name to uppercase.
13. The United Kingdom appears multiple times in the table. Once as a whole, and each of the islands separately. Keep only the whole UK in the table. Remove the information for the islands.
14. The table we scraped does not contain the country population, but it is available in this other table: https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population. Get the information from there. Keep only the country name and country population as columns.

15. Merge the two tables into one. Ensure that the table results contain all the information of both tables.
16. What problem do you encounter in the merging? What would you need to do to solve it?
17. Create a new variable: the population density per country (population / total area in km^2).
18. Create a histogram of the population density.