Mini Project: Vector Subsetting and Manipulation

Shahryar Minhas

2/2/2021

Guidelines

This homework is due on **Feb 9** by 11:59pm. You may work with up to two other people, each team member is expected to contribute equally to the assignment. You are not allowed to share your code or work with members from any other team. If I find that members from different teams collaborated, everyone on both teams will receive a zero for the assignment. You are allowed to consult any online resource. If you do so, please include a link to the website that you used in your script. If you use code from my lectures, there's no need to include a citation.

Please remember that this homework counts as one of the mini-projects for the class.

Top of your script

You may want to start your script with the following lines of code. `setwd` is only necessary if the files you are inputting into your script are located in a different directory than your script.

Style Guide

Take a look at the style guide posted here for how you should format your code: http://adv-r.had.co.nz/Style.html (http://adv-r.had.co.nz/Style.html).

Data

Data for the assignment is stored in `miniProject_vector.rda`. As with the data for the vectorTest exercise there are a variety of named vectors included in this assignment.

```
load('miniProject_vector.rda')
```

Questions

Q1

- Create a new vector named `us_profile` that consists of the 2000 United States (note that R is case sensitive) values for United States on life expectancy (lifeExpect_2000), population (population_2000), and GDP per capita (gdpCap_2000). Once created name the elements of `us_profile` by the respective variable from which they were subsetted (i.e., 'lifeExpect', 'population', and 'gdpCap').

- List the countries that had a GDP per capita greater than the 90th percentile in both 1980 and 2000.

- List the countries that had an estimated life expectancy less than or equal to the 10th percentile in 1960, 1980, and 2000.

- List the countries that had a GDP per capita greater than the 90th percentile in 1960 **but not** 2000. The `setdiff` function will be useful here. Unlike `intersect` note that the order in which you input vectors into `setdiff` matters.

Q2

- Using the `sort` function list the **ten** countries with the highest GDP per capita scores in 2000. Countries with higher scores should be at the top.

- How many countries had life expectancy scores in the top 10 and GPD per capita scores in the top 10 in 2000.

- Create a new vector named `population_2000_no_outliers` that turns all the elements of `population_2000` greater than the 95th percentile to NA.

Q3

- Create a new vector named `gdpCap_2000_quantile_class` that converts `gdpCap_2000` to the following four category level variable:
    - all values in the 0-25th%ile to 1
    - all values in the 26th-50th%ile to 2
    - all values in the 51st-75th%ile to 3
    - all values in the 76th%ile to 4
- Provide a frequency table for the classes you created in `gdpCap_2000_quantile_class` .