R assignment

The file **data_exercise.csv** contains data from a random sample of 200 individuals. The available variables are the estimated glomerular filtration rate (eGFR) (in milliliters of cleansed blood per minute per body surface (mL/min per 1.73 m^2)), a reliable indicator of kidneys' function, the age (in years), the serum creatinine levels (in mg/dl), the gender and the race.

Using R programming language, complete the following tasks:

A. Import the dataset into R. Generate a random positive integer **k** from the set $\{100, 101, ..., 180\}$. Then, pick a random subset of the dataset, of size **k**, and save it to an object named **data_sample**. Make sure your results are reproducible. **This will be the dataset you are going to work from now on.** Create a new variable in the dataset, named **log_eGFR**, as the natural logarithm of eGFR and drop the old variable from the dataset. Calculate descriptive statistics for all the variables (mean(sd)/median(IQR) for the continuous ones, n(%) for the categorical ones; you are advised to convert each categorical variable to a factor, with appropriate labels for each label).

B. Produce the following plots:

- a histogram of log_eGFR,
- a barplot of race with y-axis showing percentages,
- a scatterplot of log_eGFR to age,
- a scatterplot of log_eGFR to serum creatinine, with points colored according to gender variable.

For all the plots, provide titles, appropriate names for the axes and legends whererever is needed.

C. Construct a new variable in your dataset, named **cat_SC**, which takes the value of 1 when serum creatinine ≤ 0.8, the value of 2 when $0.8 <$ serum creatinine < 1 and the value of 3 when serum creatinine ≥ 1 (check `ifelse()` function). Provide the labels **low, normal, high** to each value respectively. Then, conduct the following statistical tests:

- a two-sample t-test to compare the levels of log_eGFR between males and females,
- a χ^2-test for the null hypothesis of independence between gender and race,
- a χ^2-test for the null hypothesis of independence between gender and cat_SC,
- a test for the null hypothesis of Pearson correlation coefficient between log_eGFR and serum creatinine being equal to 0 (obtain, also, the point estimate of Pearson correlation coefficient).

Comment briefly on the results.

The deliverables for the assignment include a file (word/pdf) with your results and comments/interpretations, plus your R code for all questions either in a table at the end of the file or in a separate R script (in this case, you need to submit 2 files - one with the presentation of results, interpretation, etc., and one for the script). Your code may contain comments wherever you find it useful. Optionally, instead of the above, you can submit an Rmarkdown file (html, pdf, word) that combines code with results, comments, etc. (introduction and guide to Rmarkdown can be found here and here).