Working with data: homework

Certificate Medical Data Science

November 11, 2022

Dataset

The dataset data_cardio.csv [1] has 70,000 rows and the following columns:

Variable	short name	scale
Age	age	int (days)
Height	height	int (cm)
Weight	weight	float (kg)
Gender	gender	categorical code
Systolic blood pressure	ap_hi	int
Diastolic blood pressure	ap_lo	int
Cholesterol	cholesterol	1: normal, 2: above normal, 3: well above normal
Glucose	gluc	1: normal, 2: above normal, 3: well above normal
Smoking	smoke	binary
Alcohol intake	alco	binary
Physical activity	active	binary
Cardiovascular disease	cardio	binary (absent or present)

The main research question is whether the variable `cardio` can be explained by the other ones.

Submission

- Send the PDF (or HTML) document that you produced by R Markdown by mail.
- Invite vey@imbi.uni-heidelberg.de to your private repository for the submission.

Deadline

January 08, 2023.

[1] source: `https://www.kaggle.com/sulianova/cardiovascular-disease-dataset`

Tasks

Write an R Markdown report that treats the following tasks, whereby each task is presented as separate chapter starting on a new page.

1. Check the continuous variables for outliers and remove implausible values (in your discretion).
2. Convert a new variable `BMI` and create a summary table for the variable `BMI` for both `cardio` groups.
3. How does the systolic blood pressure and the `BMI` correlate to each other? Is there any difference between the two classes of cardiovascular disease?
4. Answer the same question for the diastolic blood pressure.
5. Repeat the two tasks before by restricting to patients whose respective blood pressure is below the 95% quantile threshold of the respective blood pressure and whose `BMI` is below the 95% quantile of `BMI`.
6. How is `age` distributed in the different categories of `cardio`? Display `age` in years.
7. Create a plot that show the distribution of `age` for both types of `gender` and both types of `cardio`.
8. Extend this plot by taking the different types of `glucose` into account.
9. Further risk factors for a cardiovascular disease may be smoking, alcohol, and insufficient physical activity. Create an overview table of how these three parameters are distributed between the two types of `cardio` and compare all three with a χ^2-test, respectively. Draw a conclusion about which of these parameters may be risk factors for cardiovascular diseases.

Choose appropriate tables and plots for illustration and describe your results in a very few sentences. (Use the `tidyverse` packages to write your report. Hint: The `kableExtra` package generates awesome tables.)
Include your code with comments, suppress messages but show warnings and errors. Work within a private GitHub repository.