Working with data: homework

Certificate Medical Data Science

October 31, 2020

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

- A PDF document that you produced.
- The GitHub link where the Rmd file, which reproduces the submitted PDF file, is stored.

Deadline

January 6, 2021.

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

Task

Write an R Markdown report that treats the following issues.

1. Compute a new variable `BMI` and create an overview table for the variable `BMI` for both `cardio` groups.
2. How does the systolic blood pressure and the `BMI` correlate? Is there a difference between the two classes of cardiovascular disease?
3. Answer the same question for the diastolic blood pressure.
4. 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`.
5. How is `age` distributed in the different categories of `cardio`? Display `age` in years.
6. Create a plot that show the distribution of `age` for both types of `gender` and both types of `cardio`.
7. Extend this plot by taking the different types of glucose into account.
8. 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 to illustrate your results.
Use the `tidyverse` packages to create your report.
Work within a private GitHub repository.