COMP809 Lab 2

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardiovascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. The dataset framingham.csv is publicly available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. Each variable considered in this studied is a potential risk factor. There are both demographic, behavioural, and medical risk factors. The variables are the following:

Demographic:

- sex: male or female (registered as male, i.e., 1 if the person is male, 0 otherwise).
- age: age of the patient.

Behavioural current:

- Education: education level, being 1 the lowest level.
- Smoker: whether or not the patient is a current smoker.
- cigsPerDay: the number of cigarettes that the person smoked on average in one day.

Medical (history):

- BPMeds: whether or not the patient was on blood pressure medication.
- prevalentStroke: whether or not the patient had previously had a stroke.
- prevalentHyp: whether or not the patient was hypertensive.
- diabetes: whether or not the patient had diabetes.

Medical(current):

- totChol: total cholesterol level.
- sysBP: systolic blood pressure.
- diaBP: diastolic blood pressure.
- BMI: Body Mass Index.
- heartRate: heart rate.
- glucose: glucose level.

Predict variable (desired target):

- 10 year risk of coronary heart disease CHD (“1”, means “Yes”, “0” means “No”).

Your task is:

1. Create a dataset with the numerical variables.
2. Generate a scatter plot for *total cholesterol* and *systolic blood pressure*. Comment on it. Hint use scatter from matplotlib.
3. Calculate the correlation between the variables in question 2. Comment on it.

4. Generate a scatter plot for *diastolic blood pressure* and *systolic blood pressure.* Calculate the correlation coefficient. Comment on your findings.
5. How many pair of variables have at most a weak linear relationship?
6. Are the means of the average number of cigarettes smoked per day different for males and female? Define the corresponding hypotheses and test them. Comment on your findings.
7. Are *prevalent stroke* and *sex* not independent? Define the corresponding hypotheses and test them. Comment on your findings.
8. If 10 year risk of coronary heart disease is the target variable, i.e., we want to explain this variable as function of the rest of variables, do we have enough evidence to remove the variable sex from the analysis? Justify your answer.
9. Calculate the principal components for the continuous variables.
    a. Is it necessary the standardisation of the variables?
    b. How many principal components are enough to represent the original variables? Comment on it.
    c. Plot the first 2 components according to the gender. Comment on it.