COMP809 – Data Exploration

Lab 1 – Task 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 publically 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, behavioral 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").

Work in the following:

1. Import the data into Python.
2. Classify the variables (nominal, ordinal, discrete, continuous).
3. Extract and print the variable's names in python.
4. How many observations are?
5. Is there any NA value? If so, present a table with the number of NA values per attribute.
6. Define a new data frame that does not include the NA values. How many observations are in this new data set?
7. Check that *age* does not contain errors.

8. Check that *total cholesterol* does not contain errors.
9. Calculate some descriptive statistics for *total cholesterol.*
10. Generate a histogram for *total cholesterol.* Comment on the distribution.
11. How many males and female are in the data set?
12. Calculate the mean of the *total cholesterol,* per gender.
13. Generate the boxplots for the *cigarettes per day* for males and females. Compare the distributions. Hint: you can use the boxplot from seaborn python package.
14. How can you visually study the relationship between *total cholesterol* and *systolic blood pressure*?