CSC 604, Fall 2024, Assignments 8 and 9

handed out Nov. 11, due Nov. 18 at 11:59pm on Canvas

In this assignment, we are moving over to the area of Data Analysis. In particular, we will analyze more deeply the data set you encountered in Assignment 5. This time, we will use the full data set rather than the portion we used in Assignment 5. The dataset contains the medical records of patients who suffered from heart failure, collected during their follow-up period, where each patient profile has 13 clinical features. The database allows researchers to explore the incidence of death during the follow-up period of patients who suffered heart failure. In particular, it can help researchers associate clinical or demographic features with the chance of death after heart failure (True, this is not a cheerful data set! I could have found something a bit more fun!). This is a 2-part assignment covering both the data science concepts discussed in class such as the PANDA concept of a dataframe and its manipulation and the visualization of your findings using Matplot Lib. In particular, you will familiarize yourself with three types of graphs in the Matplot library of Python: Pie Charts, Bar Plots, and simple Bar Plots.

You are asked to work using Jupyter Notebooks instead of PyCharm or Spyder since I would like you to familiarize yourself with the type of IDE most commonly used when using Python in Data Science.

Here is the description of the assignment:

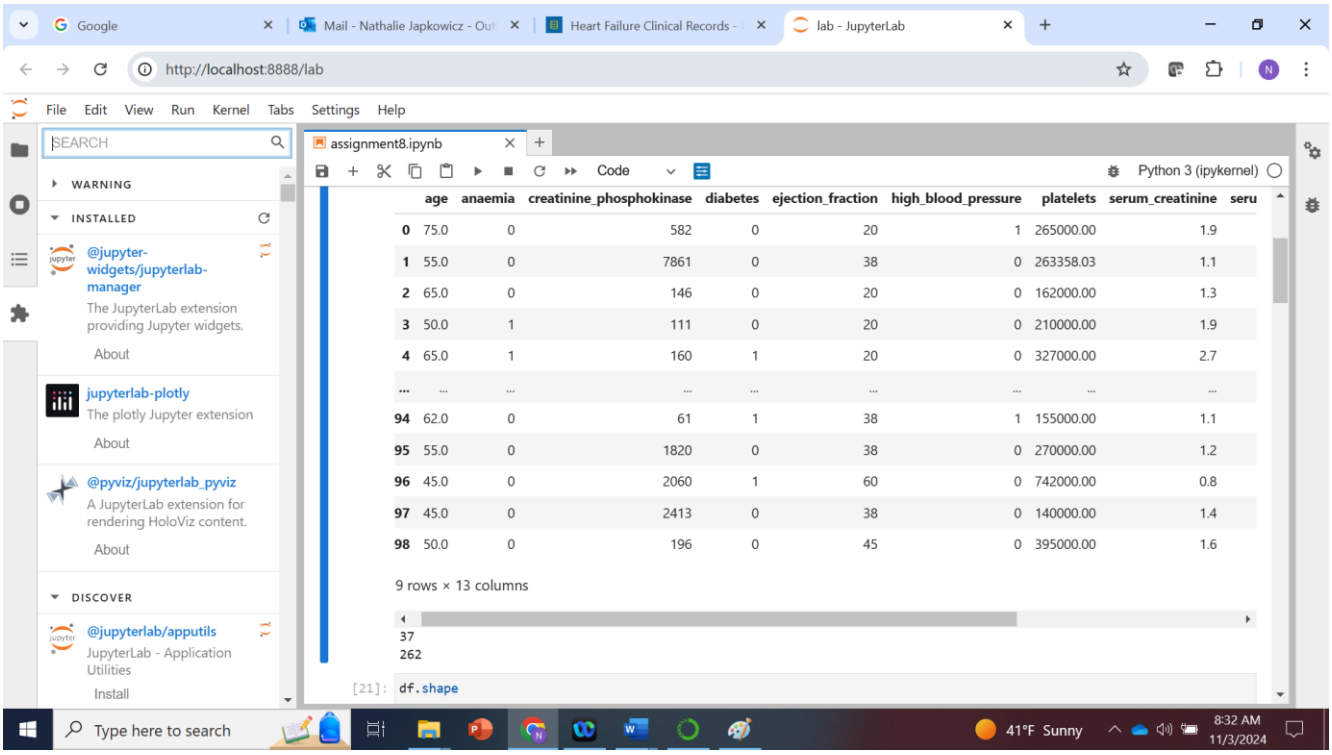
We will be working on the full Heart Failure Clinical Records data set already partially used in Assignment 5: https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records. Don't lose the link we will be also using the data set in the next assignment!

The data set contains the records of patients represented according to the following variables:

- age: age of the patient (years)
- anaemia: decrease of red blood cells or hemoglobin (boolean)
- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
- diabetes: if the patient has diabetes (boolean)
- ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
- high blood pressure: if the patient has hypertension (boolean)
- platelets: platelets in the blood (kiloplatelets/mL) - sex: woman or man (binary)
- serum creatinine: level of serum creatinine in the blood (mg/dL)
- serum sodium: level of serum sodium in the blood (mEq/L)
- smoking: if the patient smokes or not (boolean)
- time: follow-up period (days)
- death event: if the patient died during the follow-up period (boolean)

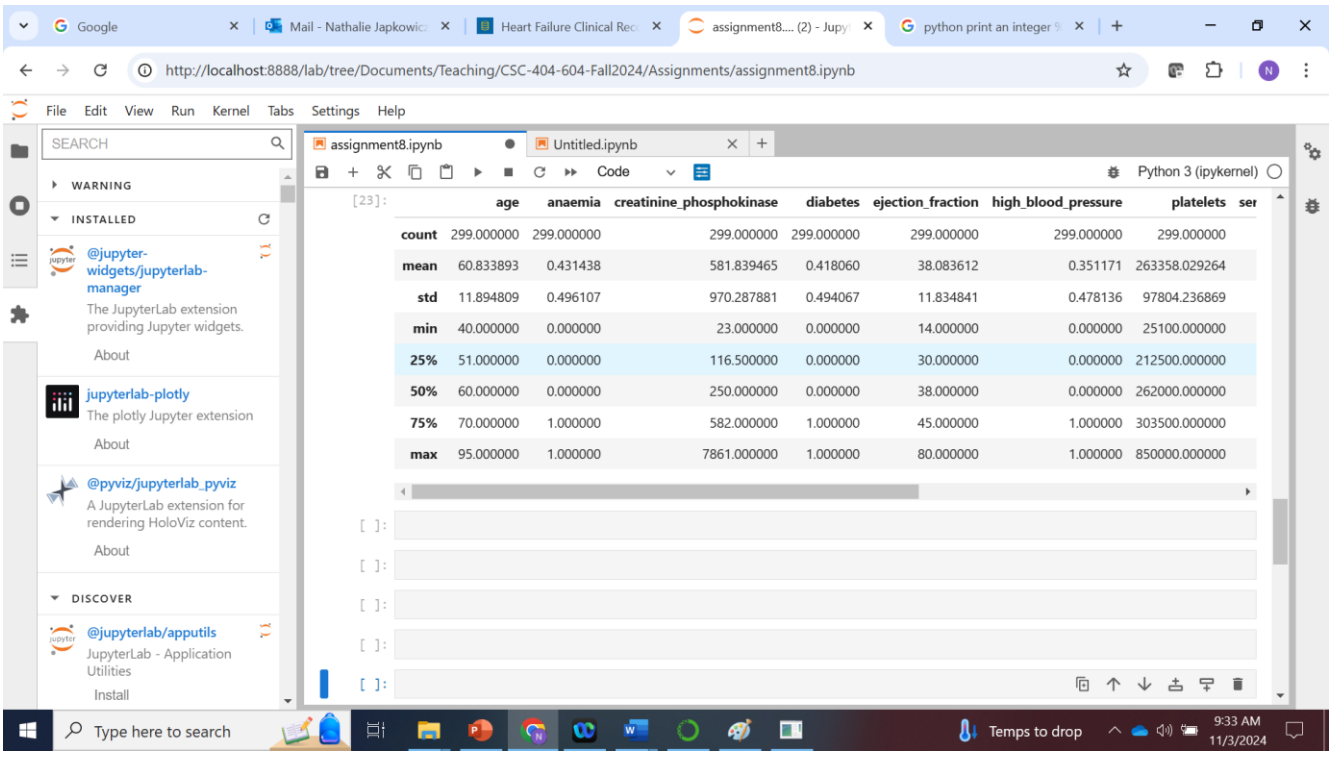
Part I: PANDAS

Your first task will be to read the file and display it.

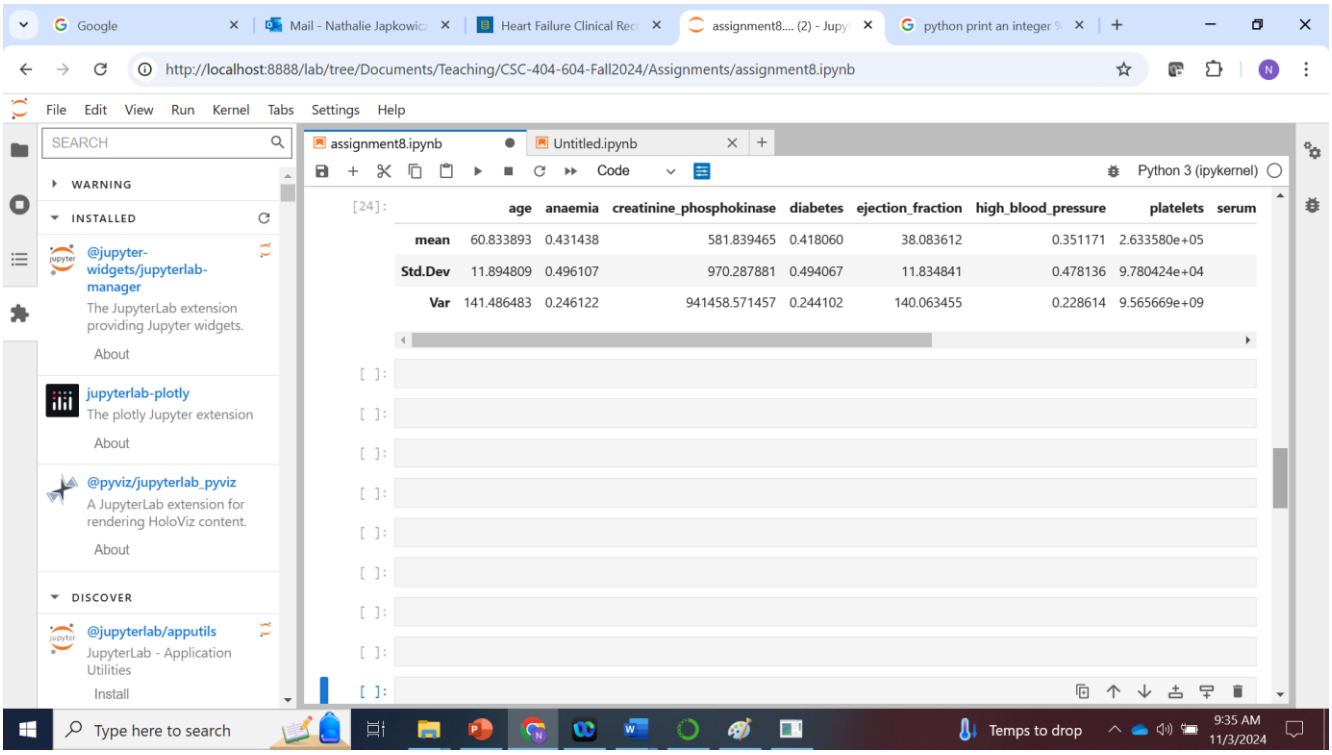


Please also display general information/statistics about each column of the data set in the following two ways, a) and b):

a)



b)



You must then find out and print the following quantities:

The number of people 45 or less in the database is A.
The number of people older than 45 in the database is B.
The youngest person in the database is C years old.
The oldest person in the database is D years old.

The number of people in the database who died following a heart failure event is E out of F patients.
The number of people in the database who survived following a heart failure event: G out of F patients.

In other words, the percentage of people who died following a heart failure event is H%.

The percentage of young people (<= 45) who died after a heart failure event is I%.
The percentage of old people (> 45) who died after a heart failure event is J%.
The percentage of smokers who died after a heart failure event is K%.

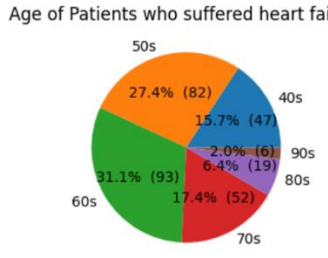
The number of men in the database is L.
The number of men survivors in the database is M or N%.
The number of women in the database is O.
The number of women survivors in the database is P or Q%.

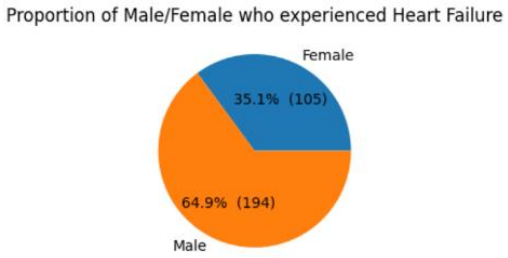
Letters A-Q above need to be replaced by the correct numbers. If the number is a decimal, please display two decimal places. If it is an integer, please display it as such, without decimals.

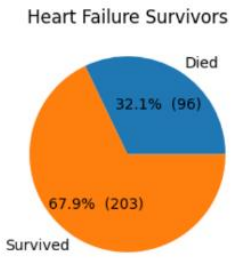
In addition, if you are interested in additional statistics about this data set, please go ahead and include them.

PART II: Matplot Lib

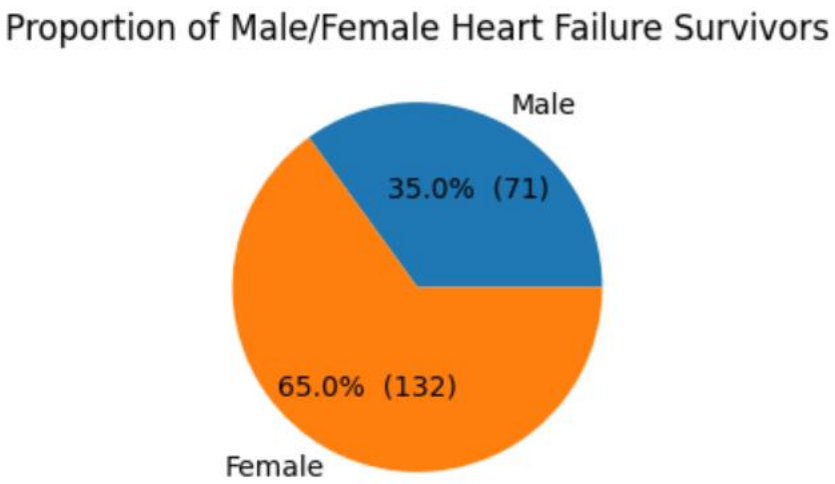
Your first task will be to show pie charts to capture the proportions of age, sex, and survivors in the Heart Failure data set. The age groups considered should be 40s (< 50), 50s (>= 50; < 60), 60s (>= 60; < 70), 70s (>= 70; < 80), 80s (>= 80; < 90), 90s (>= 90).



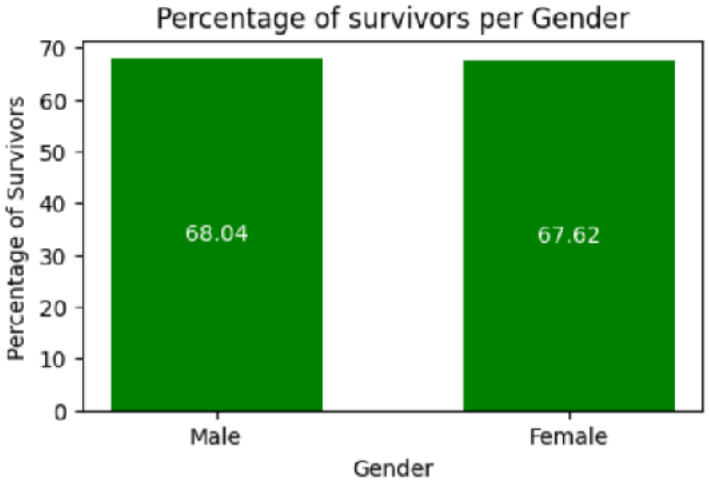




Another Pie chart will capture the proportion of male to female Heart Failure survivors:



Next, you are asked to draw a bar graph showing the percentage of survivors per gender:



Finally, you will plot a graph tracing the percentage of Heart Failure survivors per age group.