Module 3 Lab

For all lab 3 questions we will be using the heart.csv data provided with this lab. Along with the heart.csv data, I have provided a heart_data_dictionary.csv file that provides a description of each column. As you answer the lab questions, it may be beneficial to reference this data dictionary.

Subsetting data

1. Filter the heart data for all observations where the person is 50 years or older. How many observations are there?
2. Using the original heart data, filter for those observations that are male and 50 years or older. How many observations are there.
3. Using the original heart data, filter for those observations that are female, 50 years or younger, and have the disease (disease = 1). Select `chest_pain`, `chol`, and `max_hr` columns. How many rows and columns are in the resulting DataFrame?

Manipulating data

1. Are there any missing values in this data? If so, which columns? For these columns, fill the missing values with the value that appears most often (aka "mode"). This is a multi-step process and it would be worth reviewing the `.fillna()` docs (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.fillna.html).
2. Create a new column called `risk` that is equal to $\frac{age}{\text{res\_bp}+chol+\text{max\_hr}}$. What is the mean of this `risk` column?
3. Replace the values in the `rest_ecg` column so that:

    - normal = normal
    - left ventricular hypertrophy = lvh
    - ST-T wave abnormality = stt_wav_abn **Hint**: one of the original values may have an extra space at the end of the name! How many observations fall into each of the new `rest_ecg` categories?

Summarizing data

1. What is the mean resting blood pressure for males and females?
2. What is the mean and median cholesterol levels for males and females?
3. Which age group has the largest median cholesterol levels for males?
4. Compute mean `risk` value (the `risk` column was created in problem 2 of the "Manipulating data" section) for each age and sex. Which gender and age group has the highest average risk value?