Homework 1

Instructions

- Insert the code and generate the figures you need to solve the problems using this notebook.

Problem - Analysis of heart disease data set

Cardiovascular diseases (CVDs), commonly known as heart disease, are the leading cause of death worldwide, accounting for 17.9 million deaths annually. Contributing factors to CVDs include hypertension, diabetes, overweight, and unhealthy lifestyles.

The dataset contains 14 features or attributes from 900 patients; however, published studies chose only 14 features that are relevant in predicting heart disease.

Below you can see the description of each column (this is often called meta data)

Medical Data Dictionary (Metadata)

Age

Sex

- Male: `1`
- Female: `0`

Chest Pain Type

- Value 1: Typical angina
- Value 2: Atypical angina
- Value 3: Non-anginal pain
- Value 4: Asymptomatic

Resting Blood Pressure

- In mm Hg on admission to the hospital

Serum Cholesterol

- In mg/dl

Fasting Blood Sugar

- (Fasting blood sugar > 120 mg/dl): `1` = True, `0` = False

Resting Electrocardiographic Results

- Value 0: Normal

- Value 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- Value 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria

Thalach

- Maximum heart rate achieved

Exercise Induced Angina

- `1` = Yes
- `0` = No

Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest.

Oldpeak

- ST depression induced by exercise relative to rest

Slope

- The slope of the peak exercise ST segment
  - Value 1: Upsloping
  - Value 2: Flat
  - Value 3: Downsloping

Vessels Colored by Flouroscopy

- Number of major vessels (0-3) colored by flouroscopy

Thalassemia

- A blood disorder called thalassemia
  - Value 3: Normal
  - Value 6: Fixed defect
  - Value 7: Reversable defect

Target

- `0` = No Heart Disease
- `1` = Heart Disease

1. Read the data into a pandas dataframe and assign it to a variable named `df`

```
## Place your code here
```

2. Print the first five rows of the data set.

```
## Place your code here
```

3. Print the last five rows of the data set. (Hint: There's a function similar to `pd.head` for it)

```
## Place your code here
```

4. Count the number of rows in the data and assign it to `n_rows` variable and print.

```
## Place your code here
```

5. Count the number of missing values in each variable of the data frame. Assing it the variable `missing_count` and print

```
## Place your code here
```

6. Calulculate the percentage of missing data in each variable and save it to the variable `missing_percentage`. Print it

```
## Place your code here
```

7. What are the two variables with the highest percentage of missing entries? What do you recommending doing about it?

Place your answer here as plain text

8. Calculate the percentage of men and women in the data set. Save it the to the variable `m_w_fraction` and print. Are the number of men and women in the experiment balanced?

```
## Place your code here
```

Discuss it here:

9. Plot the histogram of the colesterol variable(`chol`) variable using pandas with 20 bins. What can you observe from the histogram?

```
## Place your code here
```

Place your answer here as plain text

10. Make a scatter plot of the `age` with `chol` using pandas. What do you observe - Are there any visible patterns?

```
## Place your code here
```

Place your answer here as plain text