QUEEN'S
UNIVERSITY
BELFAST

# DSA8002 COURSEWORK 2

Heart Disease Prediction

Author Name: Kulbushan Shah

| **Serial No:** | **Contents** | **Page No.** |
|---|---|---|
| **1.** | Introduction | 2 |
| **2.** | About data | 3 |
| **3.** | Manipulation of data | 4 |
| **4.** | Data analysis | 6 |
| 5. | Conclusion | 10 |

# 1. Introduction

There are several types of heart disease that tend to occur in humans based on certain type of heart condition. By considering the diagnostic test, carried out based on symptoms, about heart disease, a predication from the sample of population is elicited. There are fourteen different attributes from which are used to generate a diagnostic test for heart disease. Among which chest pain type and fasting blood sugar are commonly used in numeric test. This report is built by considering visualization and predictive modelling techniques used over the Cleveland heart disease dataset, to predict the presence or absence of heart disease based from the sample of data that is collected from the UCI repository (given link below).

The link of the dataset is secured here: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

# 2. About Data

## 2.1 What is data?

Data can be anything from which a data product can be built with certain analysis technique by determining the strength and weakness about an attribute belonging to an object. We live in the world that is drowning in the data and we can make use of this data to build data product: which we sometime call it as Website or Software. Example of which can be considered as a smart chip that is incorporated in student's id card. Collecting data to store on database can be initial approach to articulate a database design. These products can be used to solve our mundane task for which human minds are not build to accomplish it.

## 2.2 Why choose this data?

Heart disease data is used here to fuel predictive analysis, and visualisation techniques to establish, nearly, all major symptoms that human heart possess which may lead to certain type of heart disease. The analysis of the data is been based to accumulate the often-occurring symptoms which is result of individual life style.

## 2.3 How to get data?

The data here is acquired from the UCI repository whose solely purpose is use to articulate specific purpose. The source of the relevant data collected used to predict heart disease cause is secured in the Introduction section of this report.

# 3. Manipulation of data

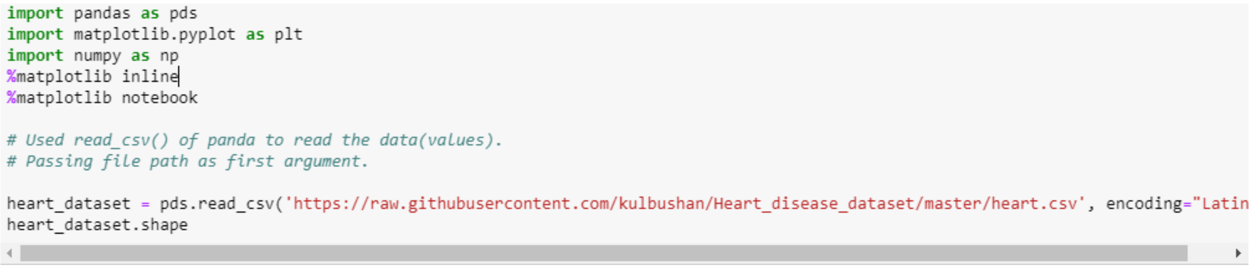
Manipulation techniques are implemented over the heart disease data (furnished here) to achieve relevant goals and to establish data product as per the intended request. Several manipulation methods such as reading data from the source file, extracting specific information, cleaning and articulating relevant data, processing the data to generate graphs are used with support of the in-built functionalities of Python. Few of which are furnished below:

3.1 **Reading the data**:

The read_csv() function of Pandas are used to read the data from the source file specified in the file path.

**Code Snippet:**

heart_dataset =
pds.read_csv('https://raw.githubusercontent.com/kulbushan/Heart_disease_dataset/master/heart.csv')



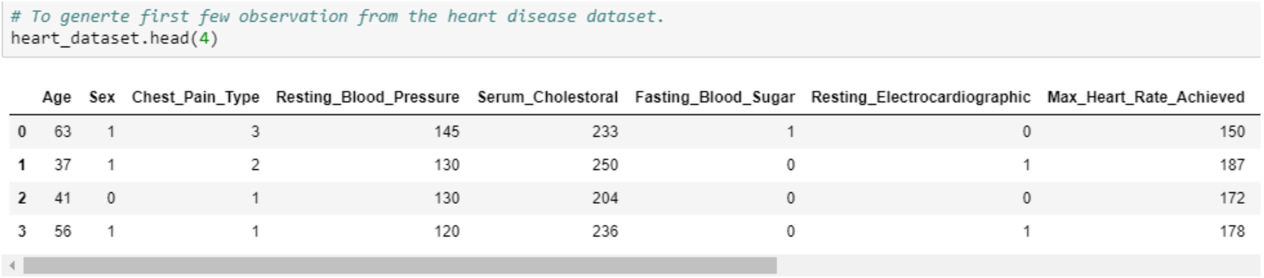
```
import pandas as pds
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline
%matplotlib notebook

# Used read_csv() of panda to read the data(values).
# Passing file path as first argument.

heart_dataset = pds.read_csv('https://raw.githubusercontent.com/kulbushan/Heart_disease_dataset/master/heart.csv', encoding="Latin
heart_dataset.shape
```

3.2 **Extracting specific information**:

The head() function generates the first fours observation(five by default) of the heart disease dataset.

**Code Snippet**:

heart_dataset.head(4)



```
# To generte first few observation from the heart disease dataset.
heart_dataset.head(4)
```

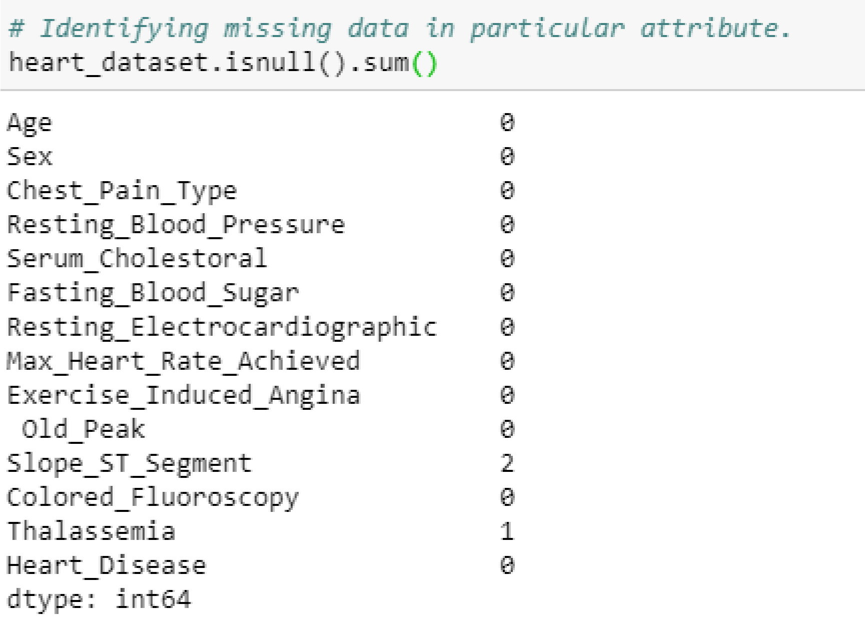
| | Age | Sex | Chest_Pain_Type | Resting_Blood_Pressure | Serum_Cholestoral | Fasting_Blood_Sugar | Resting_Electrocardiographic | Max_Heart_Rate_Achieved |
|---|---|---|---|---|---|---|---|---|
| 0 | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 |
| 1 | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 |
| 2 | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 |
| 3 | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 |

3.3 **Cleaning data**:

Cleaning data is essentially to erase badly formatted data and to build unique, relevant that can used to accomplish well design software.

Firstly, here, in heart disease dataset, few irrelevant data are identified.

**<u>Code snippet</u>:**
heart_dataset.isnull().sum()

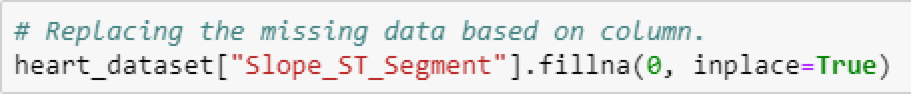


```
# Identifying missing data in particular attribute.
heart_dataset.isnull().sum()
```

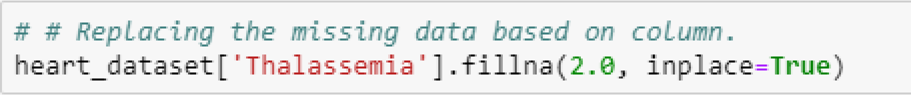
```
Age                               0
Sex                               0
Chest_Pain_Type                   0
Resting_Blood_Pressure            0
Serum_Cholestoral                 0
Fasting_Blood_Sugar               0
Resting_Electrocardiographic      0
Max_Heart_Rate_Achieved           0
Exercise_Induced_Angina           0
 Old_Peak                         0
Slope_ST_Segment                  2
Colored_Fluoroscopy               0
Thalassemia                       1
Heart_Disease                     0
dtype: int64
```

Lastly, the badly formatted data, or missing data or irrelevant data are erased and filled with certain data.

Here, the use of fillna() method is acquired to format the data according to the need.



```
# Replacing the missing data based on column.
heart_dataset["Slope_ST_Segment"].fillna(0, inplace=True)
```



```
# # Replacing the missing data based on column.
heart_dataset['Thalassemia'].fillna(2.0, inplace=True)
```

# 4. Data analysis

After the performing certain manipulation method (mentioned above), which are half a part of the data analysis, over the data we will dive into some of the essential concept of data analysis which we call as Exploratory data analysis and Communication.

**Exploratory data analysis**:

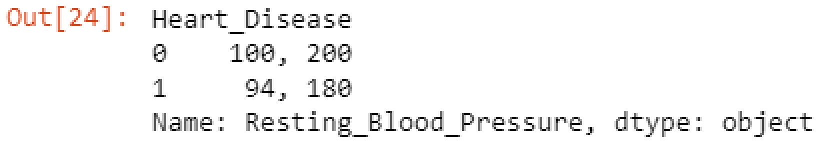
Once the data has been cleaned, exploring the data become more important as the acquire a desire pattern that can be invaluable for understandable the data. However, it may still require some iterative process to improvise the data by exploring some technique which we can as descriptive analysis. Once among other task to achieve descriptive) analysis technique is grouping the data.

Here in the heart disease dataset, groupby() function is used to predict certain attribute based on which heart disease report can be concluded.

Below is the code snippet of groupby() used in predicting heart disease assignment.

```
heart_dataset.groupby(["Heart_Disease"])["Resting_Blood_Pressure"].min().astype(str)+',’+
heart_dataset.groupby(["Heart_Disease"])["Resting_Blood_Pressure"].max().astype(str)
```

which results in below output, from which we can predict the min and max Resting Blood Pressure of an individual from the sample population.

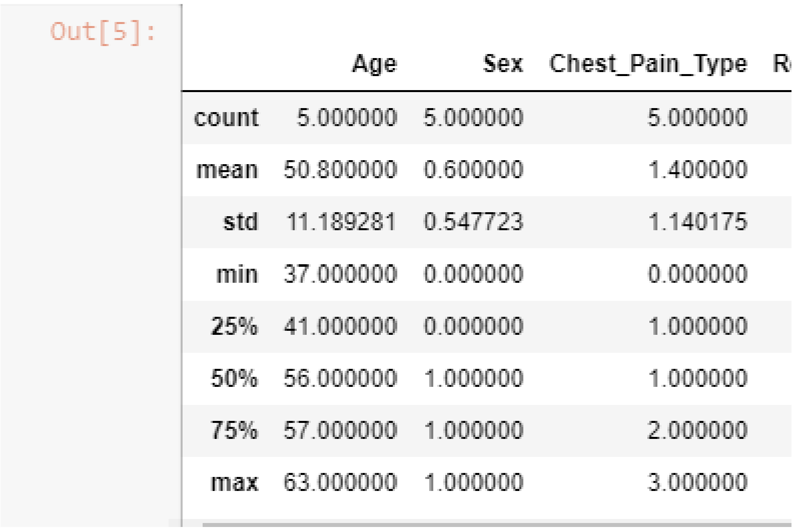


```
Out[24]: Heart_Disease
         0    100, 200
         1     94, 180
         Name: Resting_Blood_Pressure, dtype: object
```

Below describe() method used to understand the attribute used in predicting the heart disease. Which will result in generating some statistical results; count, mean, standard devation, min, max, 25th percentile, 50th percentile, and 75th percentile.

heart_dataset.head().describe(include=[np.number])

**Figure 1**:



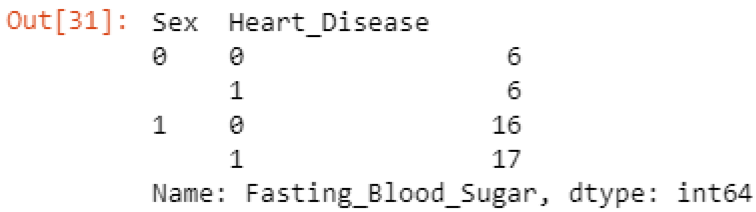
Out[5]:

| | Age | Sex | Chest_Pain_Type | R |
|---|---|---|---|---|
| count | 5.000000 | 5.000000 | 5.000000 | |
| mean | 50.800000 | 0.600000 | 1.400000 | |
| std | 11.189281 | 0.547723 | 1.140175 | |
| min | 37.000000 | 0.000000 | 0.000000 | |
| 25% | 41.000000 | 0.000000 | 1.000000 | |
| 50% | 56.000000 | 1.000000 | 1.000000 | |
| 75% | 57.000000 | 1.000000 | 2.000000 | |
| max | 63.000000 | 1.000000 | 3.000000 | |

Code snippet used to establish observation. Grouping data based on gender and heart disease status(presence or absence): **Observation 1**:

heart_dataset.groupby(['Sex', 'Heart_Disease'])['Fasting_Blood_Sugar'].sum()

**Figure 2**:

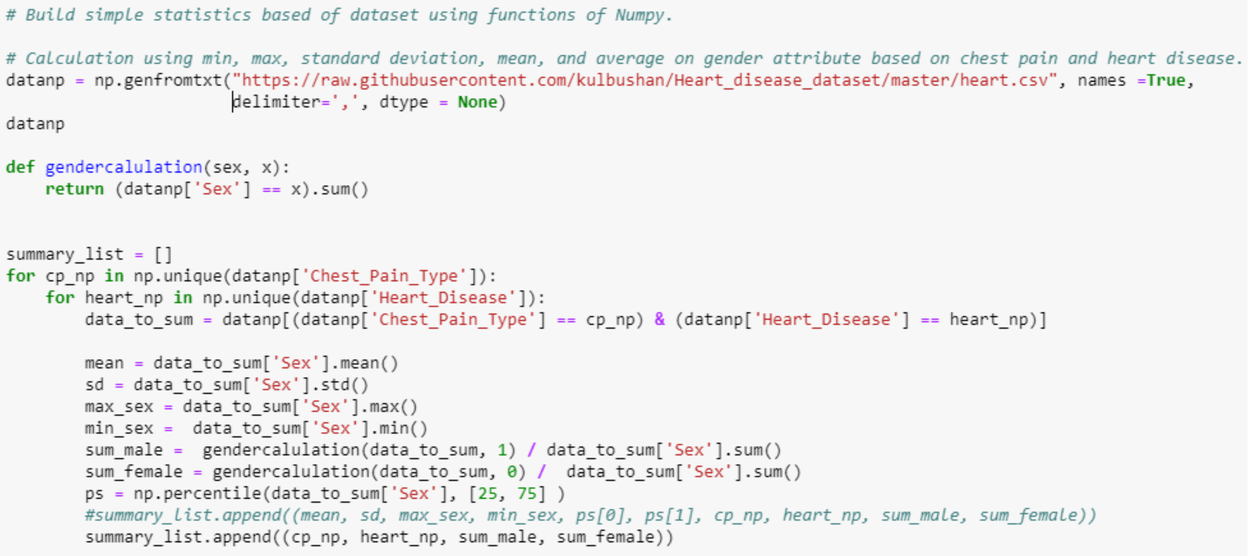


```
Out[31]: Sex  Heart_Disease
         0    0                 6
              1                 6
         1    0                16
              1                17
         Name: Fasting_Blood_Sugar, dtype: int64
```

Code snippet used to establish observation. Statistic methods used to predict heart disease comparison between male and female (Figure 3 illustrate the code snipped below): **Observation 2**:

print((str(summary_list[7:]).strip('[]')))

**Figure 3**:



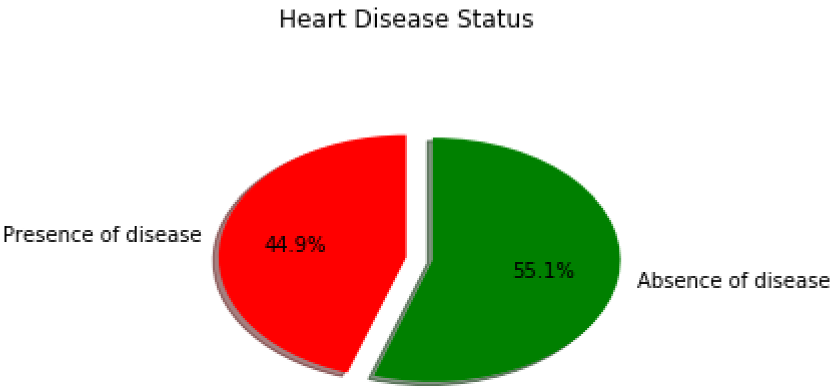
```
# Build simple statistics based of dataset using functions of Numpy.

# Calculation using min, max, standard deviation, mean, and average on gender attribute based on chest pain and heart disease.
datanp = np.genfromtxt("https://raw.githubusercontent.com/kulbushan/Heart_disease_dataset/master/heart.csv", names =True,
                       delimiter=',', dtype = None)
datanp

def gendercalulation(sex, x):
    return (datanp['Sex'] == x).sum()


summary_list = []
for cp_np in np.unique(datanp['Chest_Pain_Type']):
    for heart_np in np.unique(datanp['Heart_Disease']):
        data_to_sum = datanp[(datanp['Chest_Pain_Type'] == cp_np) & (datanp['Heart_Disease'] == heart_np)]

        mean = data_to_sum['Sex'].mean()
        sd = data_to_sum['Sex'].std()
        max_sex = data_to_sum['Sex'].max()
        min_sex =  data_to_sum['Sex'].min()
        sum_male =  gendercalulation(data_to_sum, 1) / data_to_sum['Sex'].sum()
        sum_female = gendercalulation(data_to_sum, 0) /  data_to_sum['Sex'].sum()
        ps = np.percentile(data_to_sum['Sex'], [25, 75] )
        #summary_list.append((mean, sd, max_sex, min_sex, ps[0], ps[1], cp_np, heart_np, sum_male, sum_female))
        summary_list.append((cp_np, heart_np, sum_male, sum_female))
```

**Communication**:

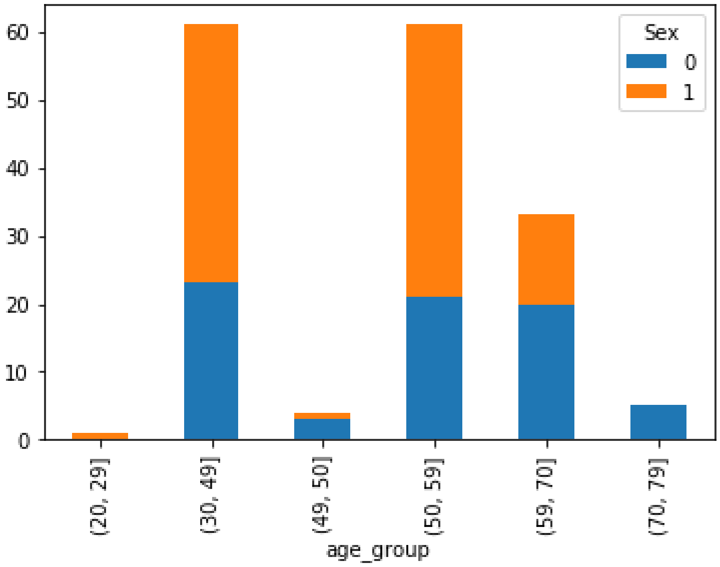
While statistic being a primary source finding descriptive analysis, likewise, communicating it also becomes more important on the other hand. Though, it is also considered as an iterative process for analysing the data.

Communication the data gives more insights on what the dataset is possessed of. Communication will also be helpful in just present the important numbers that should be solely focused on during certain part of analysis. The communication is presented using visualisation method (i.e., graph, pie chats).

Here, in the heart disease dataset, the bar graph and pie chart are been used. This shows that the number of samples with presence (red highlighted) and absence (green highlighted) of heart disease from the dataset.



Next, the below stacked graph, manifest the total number of heart disease sample on y-axis corresponding to age group along with gender on x – axis.



# 5. Conclusion

From the observation 1 (highlighted in bold on page 7) we can predict that fasting blood sugar in male's with presence (fbs > 120 mg/dL) and absence (fbs < 120 mg/dL) and found that they are essentially identical. Additionally, unlike than male's, there are 1 percent of female's with higher number of milligrams per decilitre of fasting blood tend to have heart disease.

From the observation 2 (highlighted in bold on page 7), manifests that, in contrast to females, the data manifest that there is seventeen percent on an average male, with non-anginal chest pain which resulted to cause heart disease.

Considering visualization and predictive modelling techniques on given Cleveland heart disease dataset, a conclusion can be drawn, that, individual with gender as male, chest pain type > 0, 30 < age < 69 , fasting blood sugar < 120 mg/dl are likely to result in presence of heart disease.