Table of Contents

[Project Introduction 3](#_Toc100971481)

[Analysis Objectives 3](#_Toc100971482)

[Data Acquisition and Cleaning 4](#_Toc100971483)

[Python code for reading the dataset 4](#_Toc100971484)

[Python Code for cleaning the dataset 4](#_Toc100971485)

[Why Cleaning is needed for selected Dataset 4](#_Toc100971486)

[Screenshot of code 5](#_Toc100971487)

[Data and Exploratory Analysis 5](#_Toc100971488)

[Code for Question 1 5](#_Toc100971489)

[Screenshot for Question 1 6](#_Toc100971490)

[Code for Question 2 6](#_Toc100971491)

[Screenshot for Question 2 7](#_Toc100971492)

[Code for Question 3 8](#_Toc100971493)

[Screenshot for Question 3 8](#_Toc100971494)

[Code for Question 4 9](#_Toc100971495)

[Screenshot for Question 4 9](#_Toc100971496)

[Code for Question 5 9](#_Toc100971497)

[Screenshot for Question 5 10](#_Toc100971498)

[Code for Question 6 10](#_Toc100971499)

[Screenshot for Question 6 11](#_Toc100971500)

[Code for Question 7 11](#_Toc100971501)

[Screenshot for Question 7 11](#_Toc100971502)

[Data Visualization 11](#_Toc100971503)

[Code for Question 8 11](#_Toc100971504)

[Screenshot for Question 8 13](#_Toc100971505)

[Code for Question 9 14](#_Toc100971506)

[Screenshots for Question 9 14](#_Toc100971507)

[Code for Question 10 14](#_Toc100971508)

[Screenshot for Question 10 15](#_Toc100971509)

[Visualizations as separate Images 16](#_Toc100971510)

[Question 8 16](#_Toc100971511)

[Question 9 17](#_Toc100971512)

[Question 10 17](#_Toc100971513)

[Executive Summary 18](#_Toc100971514)

[Introduction 18](#_Toc100971515)

[Key Findings 18](#_Toc100971516)

[Solution 18](#_Toc100971517)

[References 19](#_Toc100971518)

# Project Introduction

The chosen dataset is taken from [kaggle/datasets](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset) and is related to the **heart disease analysis of a patient**. The **aim** of this project is to make different analysis on the heart disease dataset to explore the analysis of patients. There is an attribute ‘**target**’ which shows the presence of heart disease in a patient. Value **1** shows the disease, and the Value **0** shows no disease. The dataset holds **1025 rows** and **14 columns** initially. We will perform different operations on this dataset to make an analysis of this dataset.

This dataset was created by:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Following is the information regarding the columns in a dataset:

1. age: patient age.
2. sex: 0 is for female and 1 for female.
3. cp: It is the type of chest pain of a patient. It has 4 types of pains.
   1. Value 0: asymptomatic
   2. Value 1: atypical angina
   3. Value 2: non-anginal pain
   4. Value 3: typical angina
4. trestbps: It is the resting of patient blood pressure.
5. chol: Showing the serum cholesterol in mg/dl.
6. fbs: It is the blood sugar fasting > 120 mg/dl.
7. restecg: Shows resting electrocardiographic results. It has values 0, 1, and 2.
8. thalach: This is the maximum heart rate of a patient.
9. exang: This attribute is for exercise induced angina.
10. oldpeak: It is the ST depression induced relative to rest by exercise.
11. slope: It is the slope of the peak exercise ST segment.
12. ca: This shows the number of major vessels covered by fluoroscopy. It is numbered from 0-3.
13. thal: It is the thalassemia value where 0 shows the normal, 1 shows the fixed-defect, and 2 shows the reversable defect.
14. target: It is the disease status of patient. 0 shows no disease and 1 shows the disease.

# Analysis Objectives

To analyze and visualize the data of chosen dataset, following are the questions we have defined:

1. Describe the attributes of chosen dataset using the statistical methods showing the statistical information, and the complete statistical chart for numerical attributes.
2. What percentage of patients have a heart disease in the age less than 45? Show their details.
3. Which type of chest pain is reported by the patients which have fixed defect in their heart? Is it all asymptomatic?
4. What are the average cholesterol values for both the males and females, which have diabetes, and which do not have diabetes?
5. Sort the dataset from older to younger patient and display it.
6. Show the maximum heart rate in ascending order of the females which have heart-disease.
7. Calculate the person’s average resting blood pressure against each resting electrocardiographic result group in ascending order.
8. Display the number of males and number of females with heart disease, and without heart disease.
9. Display the percentage of each type of chest pain of the patients who have heart disease using pie chart. Which type of pain is the most? Highlight that.
10. Display the cholesterol level range for the patients above age 50 of both with the heart disease, and without heart disease using scatter plot. Also display the average cholesterol level.

# Data Acquisition and Cleaning

## Python code for reading the dataset

# Importing necessary library

import pandas as pd

# Reading the chosen dataset

heart = pd.read\_csv('./Data/heart.csv')

# Displaying the dataset as a dataframe

heart

## Python Code for cleaning the dataset

# Cleaning the dataset

heart.drop\_duplicates(inplace=True, ignore\_index=True)

# Why Cleaning unnecessary data

print("Why Cleaning needed: There were 1025 records in the dataset initially. But the patients were not 1025 as dataset contained the data of one patient more than once. The cleaning of data was needed to manipulate the dataset which contain the details of different patients. We don't want to analyze single patient more than once that's why we do not need any duplicate values.")

# Displaying the cleaned dataset

heart

## Why Cleaning is needed for selected Dataset

There were 1025 records in the dataset initially. But the patients were not 1025 as dataset contained the data of one patient more than once. The cleaning of data was needed to manipulate the dataset which contain the details of different patients. We don't want to analyze single patient more than once that's why we do not need any duplicate values.

## Screenshot of code

A picture containing calendar

Description automatically generated

Text

Description automatically generated with low confidence

# Data and Exploratory Analysis

## Code for Question 1

# Providing the statistical information for each column as they are in numeric format

heart.describe()

# Providing the information about each attribute in a dataset

heart.info()

## Screenshot for Question 1

Text

Description automatically generated

Text

Description automatically generated

## Code for Question 2

# Storing the count of patients which have heart disease in the age less than 45.

disease\_less\_than\_45 = len(heart.loc[(heart['age'] < 45) & (heart['target'] == 1)])

# Storing total patients count

total\_patients = len(heart.loc[heart['target']==1])

# Calculating the percentage of patients which have heart disease in age less than 45.

percentage = (disease\_less\_than\_45 / total\_patients) \* 100

# Display the percentage of patients which have heart disease in an age < 45

print(f'Around {percentage.\_\_round\_\_()}% patients has a heart disease in a age less than 45. Following is the detail of them.')

# Showing the details of those patients

heart.loc[(heart['age'] < 45) & (heart['target'] == 1)]

## Screenshot for Question 2

Text

Description automatically generated

Table

Description automatically generated with low confidence

A picture containing text, meter, device, close

Description automatically generated

## Code for Question 3

# Display the types of chest pain in those patients

print(f"Type of chest pains in fixed defect patients: {set(heart[heart['thal']==1]['cp'])}\n"

       "It means that the patients with fixed defect are reporting all types of chest pain.")

# Storing the count of patients with fixed defect and asymptomatic pain

asymptomatic\_pain = len(heart.loc[(heart['thal']==1) & (heart['cp'] == 0)])

# Storing the count of total patients with fixed defect

total\_patients = len(heart[heart['thal']==1]['cp'])

# Calculating the ratio of patients with asymptomatic type of pain

asymptomatic\_ratio = (asymptomatic\_pain / total\_patients) \* 100

# Displaying the ratio

print(f"The ratio of asymptomatic pain in those patients is around {asymptomatic\_ratio.\_\_round\_\_()}%")

# 3.2: Check the patients which have heart disease and diabetes

heart\_disease\_with\_diabetes = (len(heart.loc[(heart['fbs'] == 1) & (heart['target']==1)]) / len(heart.loc[heart['target']==1])) \* 100

print(f"The heart disease patients with diabetes are around {heart\_disease\_with\_diabetes.\_\_round\_\_()}%")

## Screenshot for Question 3

A screenshot of a computer

Description automatically generated with medium confidence

## Code for Question 4

# Using a dataframe to display the average cholesterol values grouped by sex and fbs

average\_cholesterol\_values = pd.DataFrame(heart.groupby(['fbs', 'sex'])['chol'].mean())

# Displaying the average cholesterol values

print("The average cholesterol values for those patients are displayed below:")

average\_cholesterol\_values

## Screenshot for Question 4

Text

Description automatically generated

## Code for Question 5

# Displaying the dataset in reverse order with respect to age

print("The dataset in reverse order with respect to age is shown below:")

heart.sort\_values(by=['age'], ascending=False, ignore\_index=True)

## Screenshot for Question 5

Text

Description automatically generated with low confidence

## Code for Question 6

# Using the sorting and filtering to display the maximum heart rate of females who have heart-disease

max\_heart\_rate = pd.DataFrame(list(heart.loc[(heart['sex'] == 0) & (heart['target']==1)].sort\_values(by='thalach')['thalach']), columns=['Max Heart Rate of Females with Heart Disease'])

# Displaying the max heart rate in ascending order

print("Showing the maximum heart rate in increasing order of the females who have heart disease below:")

max\_heart\_rate

## Screenshot for Question 6

Text

Description automatically generated

## Code for Question 7

# Average blood pressure grouped by resting electrocardiographic result

print("Average Resting Blood Pressure grouped by resting electrocardiographic result is shown below:")

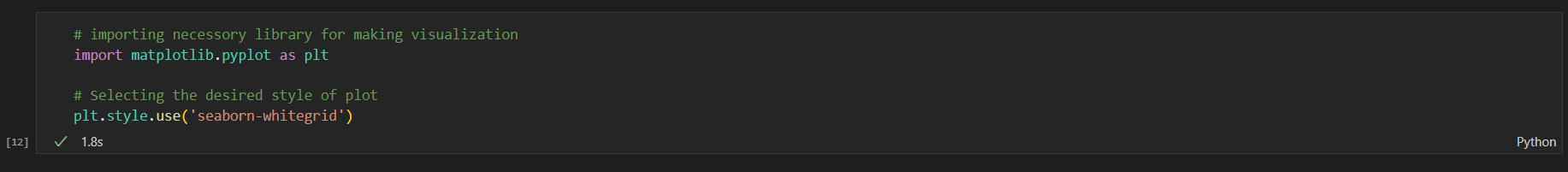
pd.DataFrame(heart.sort\_values(by='restecg').groupby('restecg')['trestbps'].mean())

## Screenshot for Question 7

Text

Description automatically generated

# Data Visualization



## Code for Question 8

# using necessory library to get the evenly spaced values

import numpy as np

# labels to be displayed on x-axis

labels = ['Not Heart Disease', 'Heart Disease']

# length of males with heart disease and without heart disease

male\_values = [len(heart.loc[(heart['sex'] == 1) & (heart['target']==0)]), len(heart.loc[(heart['sex'] == 1) & (heart['target']==1)])]

# length of females with heart disease and without heart disease

female\_values = [len(heart.loc[(heart['sex'] == 0) & (heart['target']==0)]), len(heart.loc[(heart['sex'] == 0) & (heart['target']==1)])]

# Specifying the width of bars

width = 0.35

# Creating the evenly spaced numbers for plotting the bars and labels on x-axis

x = np.arange(len(labels))

# Creating a matplotlib figure

fig, ax = plt.subplots()

# Making a bar plot for males

male\_plot = ax.bar(x - width/2, male\_values, width, label='Male')

# Making a bar plot for females

female\_plot = ax.bar(x + width/2, female\_values, width, label='Female')

# Customizing the plot

ax.set(title="Male and Female with and without Heart-Disease",

       ylabel='Count')

ax.set\_xticks(x, labels)

ax.legend()

ax.bar\_label(male\_plot, padding=3)

ax.bar\_label(female\_plot, padding=3)

# Displaying the plot

plt.show()

## Screenshot for Question 8

Text

Description automatically generated

Chart, bar chart

Description automatically generated

## Code for Question 9

# labels for each type of chest pain

labels = 'asymptomatic', 'atypical angina', 'non-anginal pain', 'typical angina'

# the chest pain percentage for each type

sizes = list(heart[heart['target']==1]['cp'].value\_counts(normalize=True) \* 100)

# highlighting the maximum percentage type

explode = (0.1, 0, 0, 0)

# making the figure and axes

fig, ax = plt.subplots()

# making a pie chart

pie = ax.pie(sizes, explode=explode, labels=labels, shadow=True, autopct='%1.1f%%', startangle=90)

# Customizing the plot

ax.set(title="Ratio of Chest Pain types in Patients with Heart Disease");

## Screenshots for Question 9

Text

Description automatically generated

## Code for Question 10

# getting the patients older than 50 years

over\_50 = heart[heart['age'] > 50]

# Creating a figure and axes

fig, ax = plt.subplots(figsize=(10, 6))

# Plotting the data using axes

scatter = ax.scatter(x = over\_50['age'],

                     y = over\_50['chol'],

                     c = over\_50['target'],

                     cmap='winter')

# Customizing the plot

ax.set(title = 'Heart Disease and Cholestrol',

       xlabel = 'Age',

       ylabel = 'Cholestrol')

ax.set\_xlim([50, 80])

ax.set\_ylim([50, 600])

# Adding the legend

ax.legend(\*scatter.legend\_elements(), title = 'Target')

# Adding a horizontal line for mean value

ax.axhline(over\_50['chol'].mean(),

           linestyle='--');

## Screenshot for Question 10

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

## Visualizations as separate Images

### Question 8

Chart, bar chart

Description automatically generated

### Question 9

Chart, pie chart

Description automatically generated

### Question 10

Chart, scatter chart

Description automatically generated

# Executive Summary

## Introduction

This project is related to the heart disease of the patient. The dataset is manipulated, and the analysis is done to check the symptoms, conditions, and behavior of different attributes that contribute to making sound decisions regarding the patients and the disease. Skill in useful libraries like pandas and matplotlib is used to get helpful information. We have patients of different ages whose data has been taken as input, and the goal state is determined. Thirteen additional attributes of the patient helped to analyze the heart disease prediction.

## Key Findings

The dataset was read successfully and stored in the pandas' data frame using the pandas' library. Then different analyses were made by following the specific criteria in each task. The team was able to sort the patient details, filter the patients based on some condition(s), and successfully group the patients into different categories. We were also able to visualize the parties regarding the patients as we have made a chart showing the males and females with and without the disease. Some exciting information related to the chest pain types of the patients has been visualized using the pie chart. We also used the scatter plot to keep track of the patients' cholesterol levels concerning their age, and the action is differentiating the patients with heart disease and without heart disease using legends. We used the mean () to collect the average value of the different attributes at different places to make quick decisions. We can now make some valuable recommendations based on these findings.

## Solution

From our key findings, we have concluded that a relatively more minor number of patients have heart disease before their 50s. So, heart disease has been reported after 50 in most cases. We have seen that the person said an issue with the minimum age of 29 and a maximum of 77 in our dataset, which means that before the age of 30, there is an exceptional case that the person would have heart disease. Also, many people who do not have pain in the chest have normal thalassemia value. Otherwise, if a person has any chest pain, they will also have some problem with thalassemia value. But we have seen that more than 50% of the person with the fixed defect has asymptomatic chest pain. Our findings show that females have less cholesterol value if they don't have diabetes. It increases in the case of having diabetes. The male's cholesterol usually doesn't fluctuate much with diabetes. We have also seen that the ratio of heart disease in males is more than the disease in females. Our analysis shows that the cholesterol value of the patients is between 200 and 300 for most of the patients. This is regardless of the disease in a patient.

Our analysis has provided us with different helpful information on which we have made some predictions and the decisions about different categories of patients. We are now able to make some strong recommendations like if a person with age less than 30 is having some issue in heart or breathing etc., there is a rare chance that they would have heart disease. And the probability in females becomes even less. We can also recommend that if a person has heart disease, there is less chance that they would have diabetes as well. Different other recommendations can be made because the information from the dataset justifies them. The data analysis has helped a lot in analyzing the dataset to make some good decisions.

# References

<https://archive.ics.uci.edu/ml/datasets/heart%2BDisease>

<https://pandas.pydata.org/docs/index.html>

<https://matplotlib.org/>