Homework 1

Task 1: Attribute Description

The attributes that exist in the dataset include the following:

person ID - This is a number used to identify the patient record and holds no intrinsic informational value, thus this is a **nominal** attribute.

Age - A number to indicate the age (in years) of the recorded patient. We can create ratios between patient ages, i.e. Patient 1 is twice the age of Patient 2, thus this a **numeric ratio** attribute

Gender - A letter (M or F) to indicate the gender of the patient. This is a binary **nominal** attribute.

chest pain type - There is no ordering in the type of chest pain, that is, chest pain Type 1 is no less or greater pain than chest pain Type 2 (or 3 and 4) and vice versa. It is solely an indicator of the existence (or non-existence) of a type of chest pain. This is a **nominal** attribute.

resting blood pressure - A measure of the pressure in the arteries when the heart rests between beats (also called the diastolic pressure). This is a **numeric ratio** attribute as there is a natural zero and differences are meaningful.

serum cholesterol in mg/dl - An indication of the risk of heart disease; this is a measure of the amount of HDL and LDL (high/low-density lipoprotein) and triglycerides (blood fat combined with cholesterol) in the blood. There exists low, normal, and high levels of cholesterol, but typically never zero. This attribute can be **ordinal** and/or **numeric ratio** because we can order by low, normal, and high cholesterol or use the cholesterol values for differences

fasting blood sugar > 120 mg/dl - A binary attribute (Yes/No) that indicates whether the patient's blood sugar level after not eating was above the normal/moderate level. A fasting blood sugar level above 120 mg/dl is a warning sign of diabetes and other issues, i.e. the body cannot process glucose properly with/without food in the body. This is a **nominal** attribute.

resting electrocardiographic results - An indication of any abnormalities in the heart after processing its electrical signals. There can be no abnormalities, hypertrophy, wave abnormalities, etc. These abnormalities have no specific ordering, making this attribute **nominal**.

maximum heart rate achieved - The highest heart rate achieved after a period of exercise. Typically this decreases as a person ages. This is an integer attribute with a natural zero and a comparison like Patient 1 had twice the maximum heart rate as Patient 2 would be meaningful, so this is a **numeric ratio** attribute.

exercise induced angina - This attribute is an indication of whether a patient records feeling pain in the chest (angina) after a period of exercise; it is a binary value (Yes/No) making it a **nominal** attribute.

oldpeak = ST depression induced by exercise relative to rest - An integer representing how far below baseline a patient's ST segment sits in ECG results *after* exercise. It is considered a reliable finding for the diagnosis of CAD. A natural zero does not exist for this attribute because the baseline is arbitrary; this is a **numeric interval** attribute.

the slope of the peak exercise ST segment - This attribute describes the orientation of the slope in one of three ways: 0 - downslope, 1 - flat, 2 - upslope. There is no ordering to these orientations, they only mean to describe the ST segment, making this attribute **nominal**

number of major vessels (0-3) colored by flourosopy - An integer indicating how many of the vessels were marked; An absolute zero exists so this is a **numeric ratio** attribute

thal: 3 = normal; 6 = fixed defect; 7 = reversable defect - An indication of the existence of a blood disorder/defect. This attribute identifies whether a patient has thalassemia making this attribute **nominal**

Has heart disease? - The target value for a prediction model. Identifies whether a patient has heart disease or not and is binary (Yes/No) making this attribute **nominal**

Task 2: Proximity Selection

The *numerical* attributes consist of: Age, resting BP, serum cholesterol in mg/dl, maximum heart rate achieved, oldpeak = ST depression induced by exercise relative to rest, number of major vessels (0-3) colored by flourosopy

For these attributes I would use the Euclidean distance as a proximity measurement. All except one attribute (oldpeak) are numeric ratio attributes with natural zero's; the differences in the magnitudes of these attributes is meaningful. Using the cosine similarity or correlation would describe relationships (linear for correlation) in the data. There is no scaling or translation required in the data and the Euclidean distance would effectively calculate the distance between patients. The dimensionality is also not too high for Euclidean distance calculations.

Task 3: Proximity Selection

Since we are calculating the Euclidean distance for each individual numeric attribute for each pair the formula

$$d(x, y) = \sqrt{\sum_{1}^{n} (x_i - y_i)^2}$$

can be simplified to

$$d_i(x, y) = |x_i - y_i|$$

For each nominal attribute, for each pair we can use a simple matching coefficient where the dissimilarity is 0 for matches and 1 for non-matches.

The columns for step 1 include person 1 ID, person 2 ID, and the numerical attributes followed by the nominal attributes. The entries for a row correspond to the pair (Person1_ID, Person2_ID) and each attribute column is the dissimilarity for this attribute for the specified pair. Each row is a new pair.

For step 2, there are three columns Person1_ID, Person2_ID, and Overall Dissimilarity. Each row is a new pair.

The pair with smallest dissimilarity is PersonID 42 and PersonID 54 with a score of 11.200
The pair with the highest dissimilarity is PersonID 1 and PersonID 60 with a score of 502.400

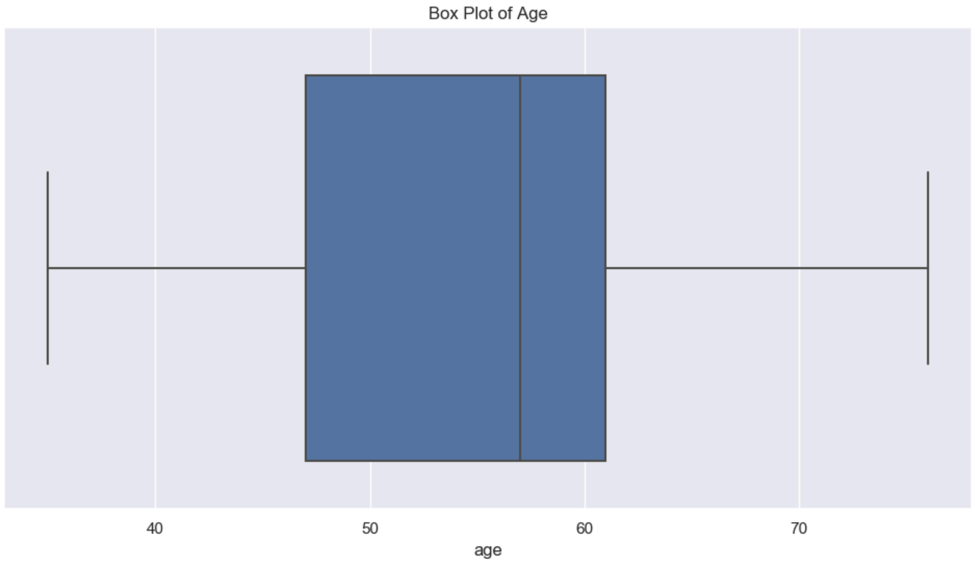
Task 4: Summary Statistics

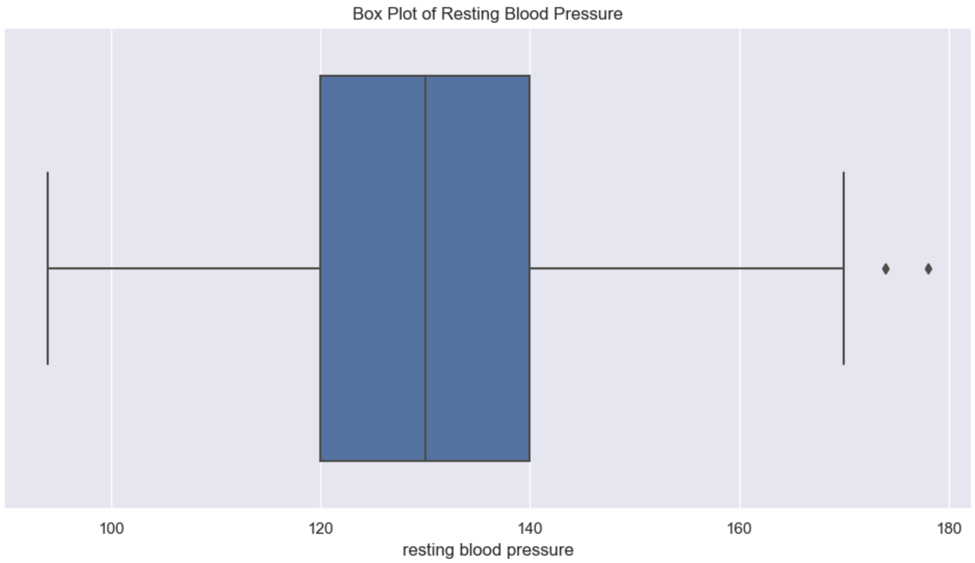
The mean, standard deviation, and 5-number summary of each attribute is listed below:

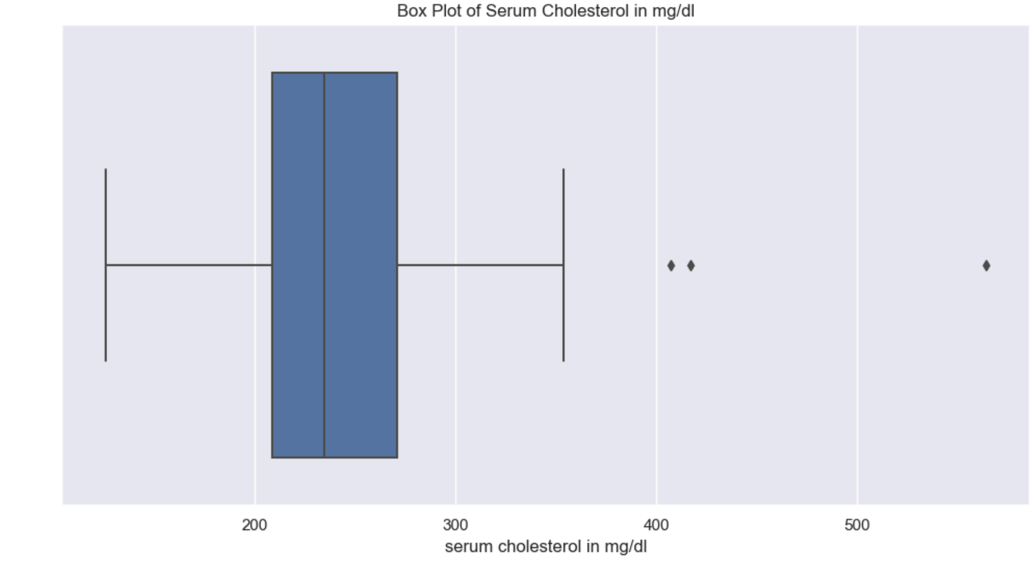
Attribute	Mean	Std. Deviation	Minimum	Q1	Median	Q3	Maximum
Age	54.84	9.248	35	47	57	61	76

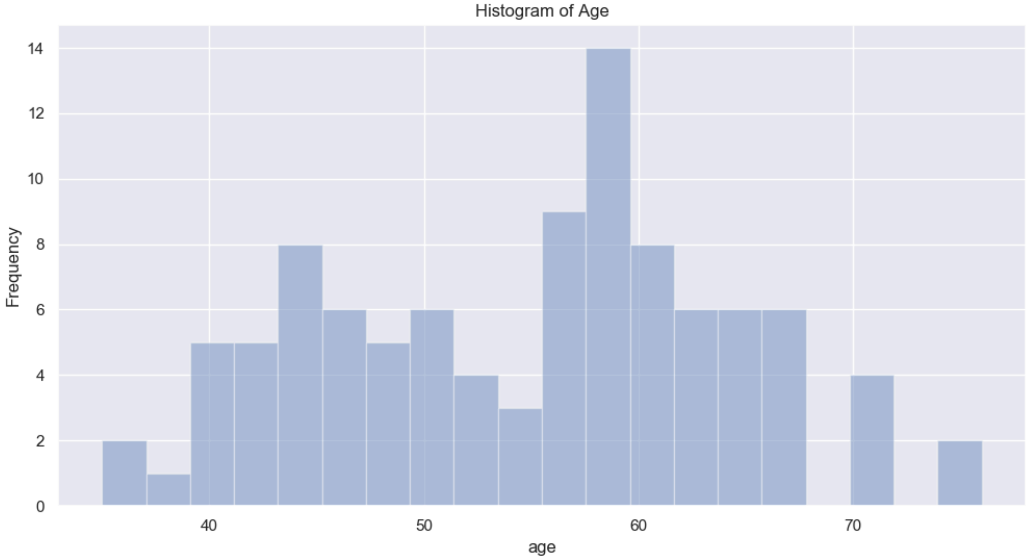
Resting blood pressure	130.27	16.509	94	120	130	140	178
Serum cholesterol (mg/dl)	247.01	58.992	126	209	234.5	270.75	564
Maximum heart rate	147.57	22.874	96	129.75	149.5	165.75	186
Oldpeak = ST depression induced by exercise relative to rest	0.902	1.079	0	0	.45	1.525	4.2

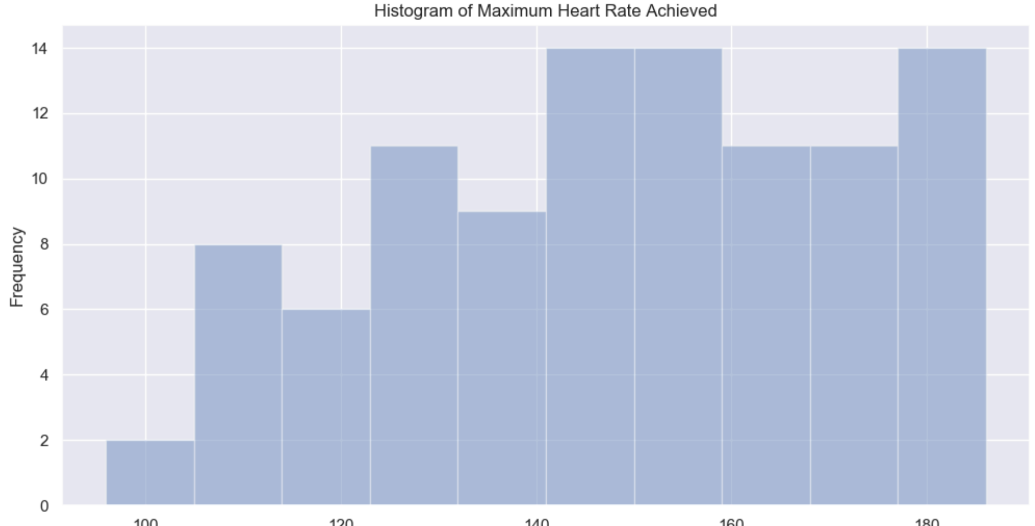
Task 5: Charts

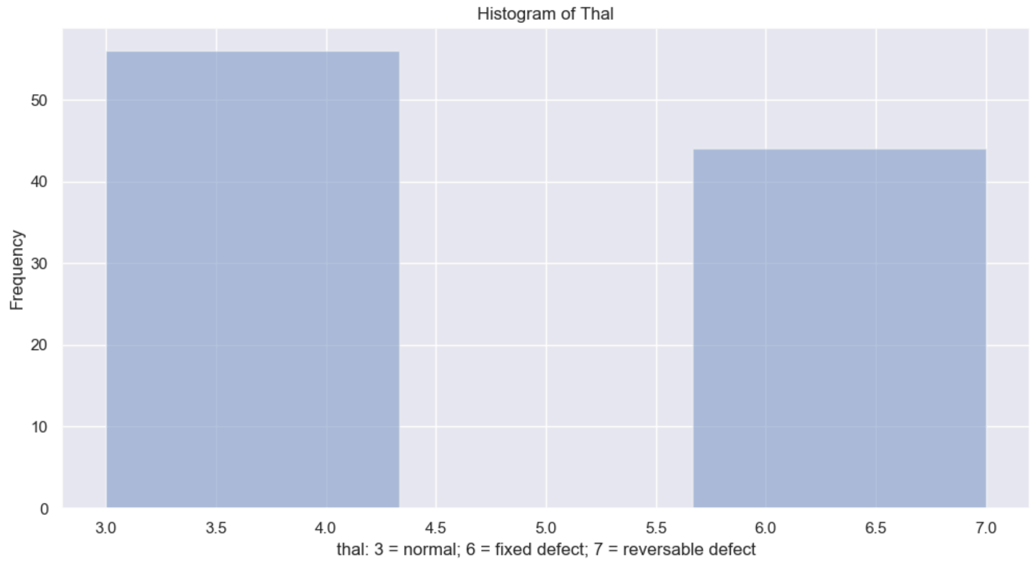


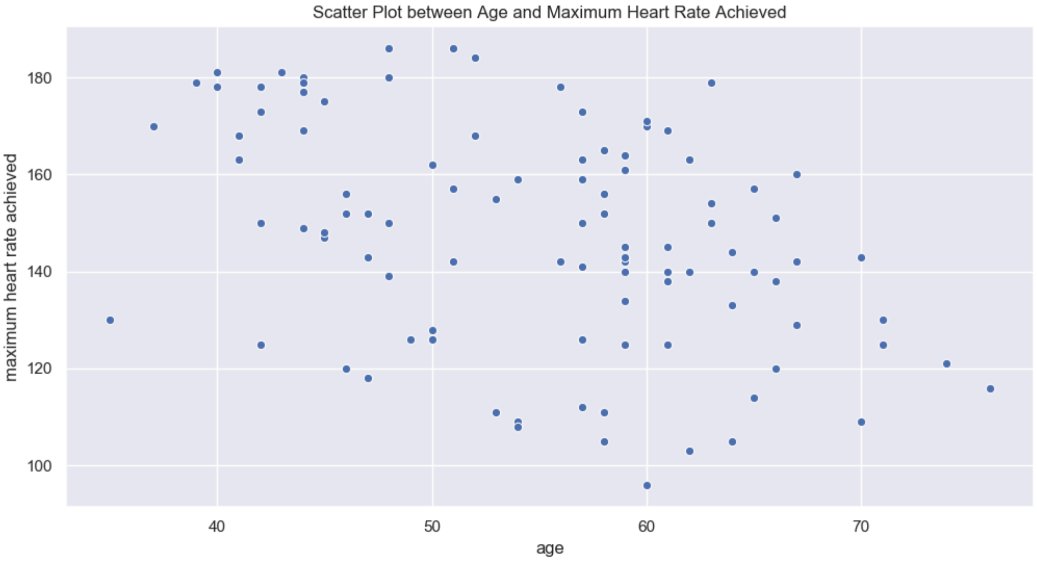


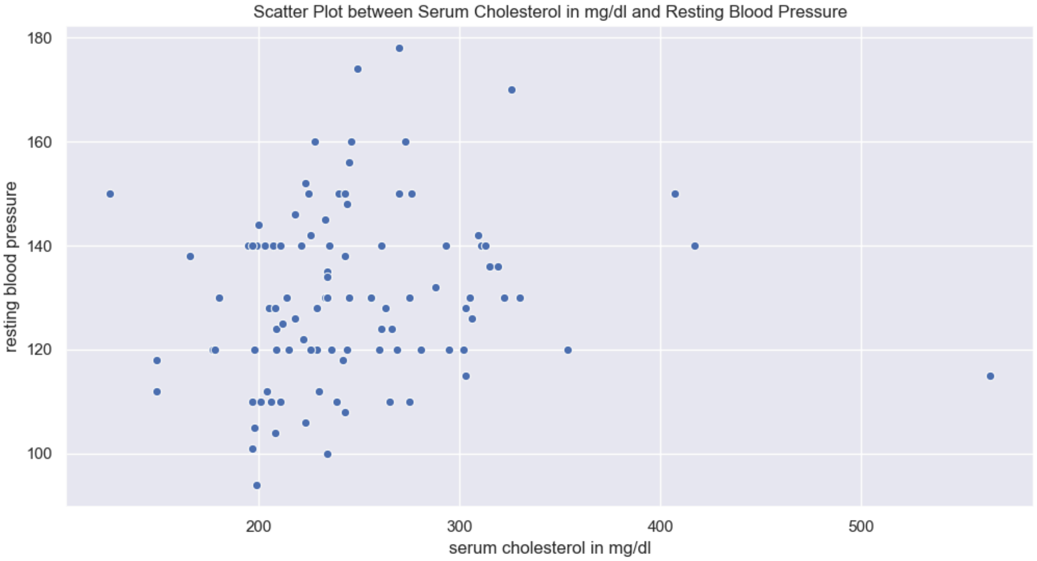

Box Plot of Serum Cholesterol in mg/dl
200
300
400
500
serum cholesterol in mg/dl

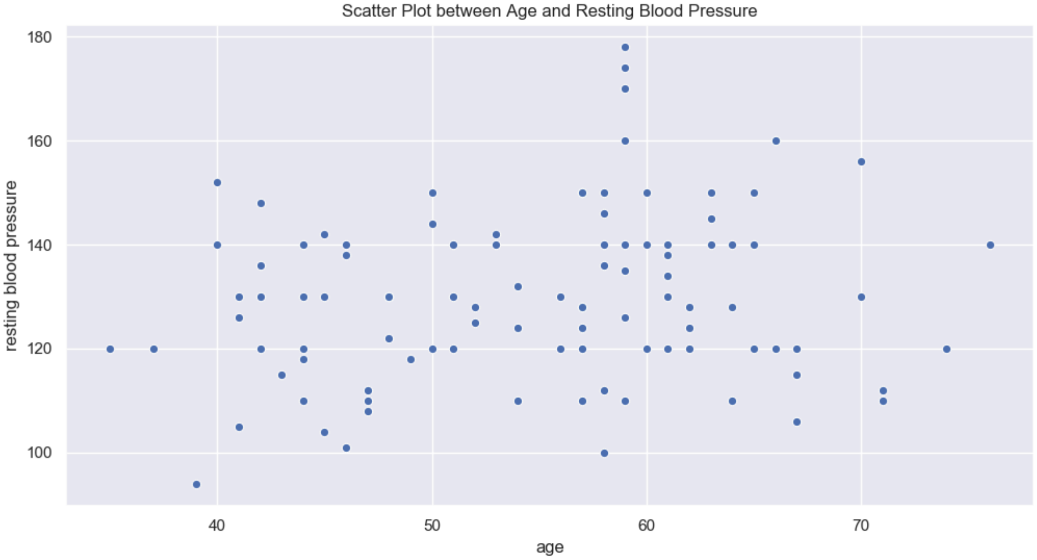

Histogram of Age
Frequency
0
2
4
6
8
10
12
14
40
50
60
70
age


Histogram of Maximum Heart Rate Achieved
Frequency
0
2
4
6
8
10
12
14
100
120
140
160
180


Histogram of Thal
Frequency
0
10
20
30
40
50
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
thal: 3 = normal; 6 = fixed defect; 7 = reversable defect


Scatter Plot between Age and Maximum Heart Rate Achieved
maximum heart rate achieved
100
120
140
160
180
40
50
60
70
age


Scatter Plot between Serum Cholesterol in mg/dl and Resting Blood Pressure
180
160
140
120
100
resting blood pressure
200
300
400
500
serum cholesterol in mg/dl


Scatter Plot between Age and Resting Blood Pressure
180
160
140
120
100
resting blood pressure
40
50
60
70
age

Task 6: Tools and Languages

For computation and processing, the Python language was used alongside the Pandas, Numpy, and Itertools libraries. These libraries and language made it extremely easy to process the data and apply computations to it, especially for Task 3. Had I used another language, the code would not have been as straightforward, especially dealing with files and large data. The computation is quick and even with nested loops, the code is efficient. Excel would have been possible to use but the speed of Pandas is unmatched for large datasets.

For plots, I used the Python language alongside the Matplotlib, Seaborn, and Pandas libraries. The libraries automatically visualize data in whatever specified plot and can be passed multiple types of data. These libraries are all easily integrated with each other, for instance Pandas and Numpy or Seaborn and Matplotlib, and are easily readable.