No collaboration.

Task 1

1. person ID: A unique number assigned to each person in the study. It's nominal because it's just a label without any numerical meaning or order.
2. Age: How old the person is in years. It's numeric ratio because we can perform all mathematical operations with it, and zero age has a meaningful interpretation.
3. gender: The biological sex of the person. It's nominal because it categorizes people without implying any order or numerical value.
4. chest pain type: The specific kind of chest discomfort experienced. It's nominal because the different types are distinct categories without a natural order.
5. resting blood pressure: The force of blood in arteries when a person is at rest. It's numeric ratio because it allows all mathematical operations and has a meaningful zero point.
6. serum cholesterol: The amount of a certain fat-like substance in the blood. It's numeric ratio because all mathematical operations are possible and zero cholesterol is meaningful.
7. fasting blood sugar > 120 mg/dl: Whether blood sugar exceeds a specific level when not eating. It's nominal because it's a binary classification.
8. resting electrocardiographic results: The outcome of a heart activity test while resting. It's nominal because it categorizes results without implying any order.
9. maximum heart rate achieved: The highest number of heartbeats per minute during exercise. It's numeric ratio because all mathematical operations are meaningful and zero has significance.
10. exercise induced angina: Whether physical activity causes chest pain. It's nominal because it's a binary classification.
11. oldpeak: The extent of a specific change in a heart test during exercise. It's numeric ratio because it allows all mathematical operations and has a meaningful zero point meaning no change.
12. slope of the peak exercise ST segment: The direction of a particular measurement in a heart test during peak exercise. It's ordinal because the categories have a meaningful order of intensity, 1, 2, 3 generally means Upsloping, flat, and down sloping.
13. number of major vessels colored by fluoroscopy: The count of main heart blood vessels showing blockage in an imaging test. It's numeric ratio because it allows all mathematical operations and zero is meaningful meaning no vessels colored.
14. thal: A measure related to blood flow in the heart. It could be ordinal if the categories imply increasing severity, or nominal if they're viewed as distinct types without inherent order.
15. Has heart disease?: Whether the person is diagnosed with heart disease. It's nominal because it's a binary classification.

Task 2

Cosine similarity and correlation are invariant to scaling, while Euclidean distance is not. Correlation and Euclidean are invariant to translation while cosine similarity is not. Cosine similarity is range from (-1, 1), correlation is also range from (-1, 1). -1 means perfectly different, 0 means no relationship, 1 means perfectly the same. Euclidean distance is range from (0, infinity). 0 means perfectly the same, while higher value means larger difference between two data instances.

Cosine similarity is calculated when x and y are two vectors (e.g., term-frequency vectors), then cos(x, y) = (x •y) /(||x|| ||y||). Correlation is calculated by covariance of x and y over std of x times std of y, cor = cov(x,v)/(std(x)\*std(y)). Euclidean distance is calculated by d(x,y)=sqr((x1​−y1​)^2+(x2​−y2​)^2+⋯+(xn​−yn​)^2).

XXX\_proximity.csv stores the result, XXX can be replaced be cosine\_similarity, correlation, and euclidean\_distance that determines which calculation is stored in the corresponding csv. The rows and the columns are responding to each line of data’s ID. Each line of data is numbered from 1 to 100. Each cell is the corresponding intersection of the rows and columns data instances’ proximity calculations.

Task 3

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Standard Deviation** | **Minimum** | **Q1** | **Median** | **Q3** | **Maximum** |
| **age** | 54.84 | 9.25 | 35.0 | 47.0 | 57.0 | 61.0 | 76.0 |
| **resting blood pressure** | 130.27 | 16.51 | 94.0 | 120.0 | 130.0 | 140.0 | 178.0 |
| **serum cholesterol in mg/dl** | 247.01 | 58.99 | 126.0 | 209.0 | 234.5 | 270.75 | 564.0 |
| **maximum heart rate achieved** | 147.57 | 22.87 | 96.0 | 129.75 | 149.5 | 165.75 | 186.0 |
| **oldpeak = ST depression induced by exercise relative to rest** | 0.90 | 1.08 | 0.0 | 0.0 | 0.45 | 1.525 | 4.2 |
| **number of major vessels (0-3) colored by flourosopy** | 0.63 | 0.90 | 0.0 | 0.0 | 0.0 | 1.0 | 3.0 |

Task 4A chart of blood pressure

Description automatically generated

1. This box plot gives the distribution of resting blood pressure across different age groups. Median resting blood pressure tends to increase with age before 60 and then decrease after 60.

A chart of different colored rectangular shapes

Description automatically generated

1. This box plot gives the distribution of serum cholesterol levels based on the number of major vessels colored by fluoroscopy. Median cholesterol levels are slightly higher in individuals with more vessels colored.

A graph of different colored squares

Description automatically generated

1. This box plot gives the distribution of oldpeak values across different age groups. Median oldpeak values tend to be higher in older age groups. There are more outliers in 41-60 age groups, indicating higher ST depression in some individuals. A graph of a number of vessels colored

   Description automatically generated
2. This histogram displays the frequency distribution of the number of major vessels colored by fluoroscopy. Most Common Value is 0 vessels colored, indicating many individuals have no major vessels affected. Fewer individuals have higher numbers of vessels colored.

A graph with blue squares

Description automatically generated

1. This histogram gives frequency of oldpeak values. Many individuals have an oldpeak of 0. The right skewness indicates most individuals have low ST depression values.

A graph of a graph

Description automatically generated with medium confidence

1. This histogram gives the frequency of different values of maximum heart rates achieved during exercise. Values spread from approximately 90 to 190 bpm. A roughly left skewness reflects a relatively high tendency of higher maximum heart rate in the dataset.

A graph of blood pressure

Description automatically generated

1. This scatter plot gives the relationship between age and resting blood pressure. There is aslight increase in resting blood pressure with age.

A graph of blue dots

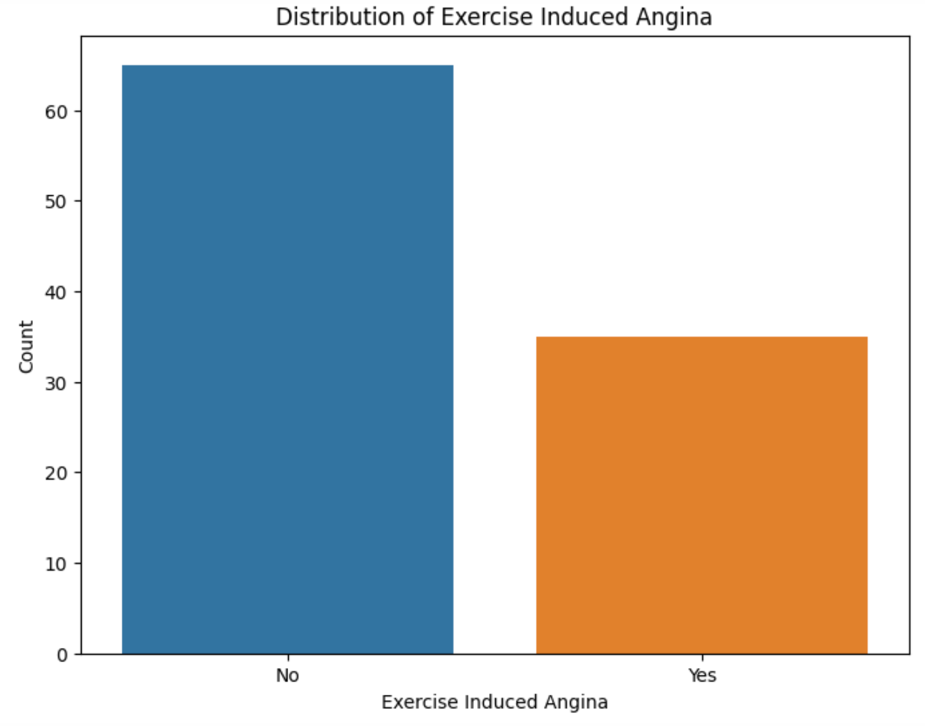
Description automatically generated

1. This scatter plot gives the relationship between oldpeak values and maximum heart rate achieved. Higher oldpeak values tend to correspond with lower maximum heart rates.

A graph of different colored dots

Description automatically generated with medium confidence

1. This scatter plot gives how serum cholesterol levels vary with age, with heart disease status highlighted. Cholesterol levels are widely distributed across ages.There is no clear trend. Individuals with heart disease are present across all cholesterol levels.



1. This bar chart gives how many individuals experienced exercise-induced angina. Majority of the dataset do not experience angina during exercise.

Task 5

I initially tried to use Excel to calculate the three proximity calculations. The pro of using Excel is that it is easy to get hands-on and start. It does not require extra setup of the environment, and the formulas are easy to understand as we directly input mathematical formulas into cells. However, the con is also obvious: the constructing process is tedious and requires a lot of manual work. It causes a lot of human errors. Thus, I switched to Python libraries like Matplotlib, NumPy, and Pandas. The con of this approach is that I need to set up everything before coding, and I need to read documentation on how to use certain package methods. The pros include many aspects, such as high efficiency on large and complex datasets, and it is less likely to make human errors since everything is done by code.