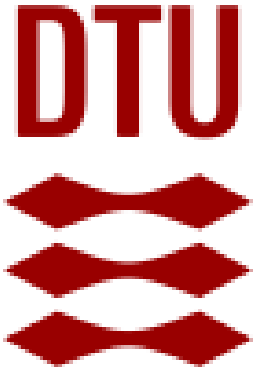


02450 - Introduction to Machine Learning and Data Mining

Project 1 Report

Group 9

Name	**Student Number**
FOO YAN RONG	S231651
KAI JIE, JARED LIANG	S231648

Contributions to Report

Report Section	**Contributions (Yan Rong - Jared)**
Description of Dataset	60% - 40%
Explanation of Attributes	40% - 60%
Data Visualisation and Principal Component Analysis	60% - 40%
Discussion	40% - 60%
Exam Questions	60% - 40%

1. Description of Dataset

1.1 Overview of Dataset

1.1.1 Title of Dataset: Heart Failure Prediction Dataset

This dataset is obtained from Kaggle and contains 918 observations and 12 attributes, comprising 11 clinical features used to predict the target variable, in this case, the occurrence of heart disease events. The 12 attributes are as follows: **Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope,** and **HeartDisease.** These attributes will be elaborated upon more in depth in the following section.

Reference to Dataset: https://www.kaggle.com/fedesoriano/heart-failure-prediction

1.1.2 Summary of Previous Analysis of Data

The dataset obtained from Kaggle was created from the combination of five independent datasets over 11 common features. The five datasets involved are as follows:

- Cleveland (303 observations)
- Hungary (294 observations)
- Switzerland (123 observations)
- Long Beach VA (200 observations)
- Stalog (Heart) Data Set (270 overvations)

The final dataset of 918 observations on Kaggle was obtained after removing the 272 duplicated observations from the five datasets..

1.2 Overall Problem of Interest and Research Plan

Heart Failure is a common event caused by Cardiovascular Diseases (CVDs), which is the number one leading cause of death globally. People with CVDs or who are at high cardiovascular risk can significantly benefit from early detection and swift intervention thereafter which would reduce chances of death and health complications.

This is where we hope to utilise machine learning models and classification and regression techniques to aid secondary prevention and screening efforts, should we be able to develop models either to identify potentially patients at risk of heart disease based on their clinical data, or predict clinical measurement values for certain attributes should they be unavailable for monitoring purposes. For the classification task, we aim to predict the **'HeartDisease'** class label, which is a discrete and nominal attribute, using a subset of the other 11 attributes in the dataset after data reduction and feature extraction methods, including Principal Component Analysis (PCA). For the regression task, we hope to predict continuous and ratio variables such as **'Age', 'RestingBP', 'Cholesterol'**, or **'MaxHR'**, and for each case we hope to build the regression model each using a subset of all other clinical attributes including 'HeartDisease' after data reduction and feature extraction methods.

To do so, we will first require preprocessing and transform our data appropriately. We plan to utilise the One-hot encoding method to convert categorical variables into a numerical format that can be fed into machine learning models subsequently. As for continuous attributes, we aim to use z-index standardisation to scale the attributes accordingly so that these attributes all have equal influence over the building of the model. (This will be elaborated on in subsequent sections)

2. Explanation of Attributes

2.1 Age

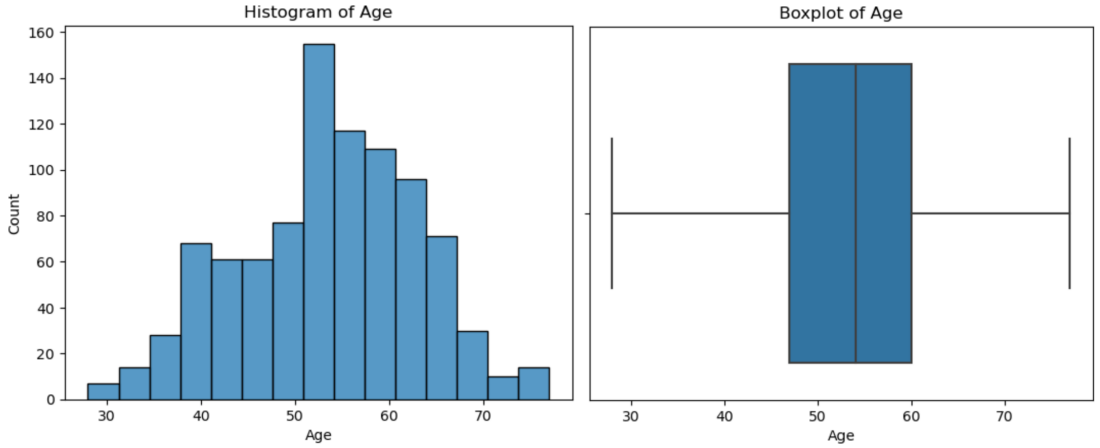


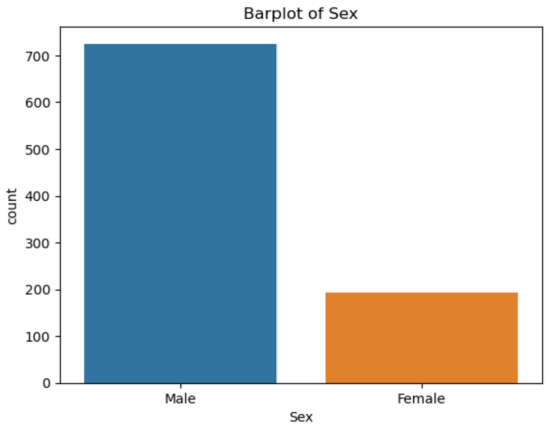
Figure 2.1: Top Left: Histogram of patient's age, Top Right: Boxplot of patient's age, Bottom Right: Summary Statistics

Age is a continuous and ratio variable. Age, measured in years, has a consistent interval between each year. Age also has a true zero point which represents the absence of age or the moment of birth.

There are no data issues with the variable age. There are also no outliers as seen from the boxplot. The variable seems to be normally distributed.

```
count    918.000000
mean      53.510893
std        9.432617
min       28.000000
25%       47.000000
50%       54.000000
75%       60.000000
max       77.000000
Name: Age, dtype: float64
```

2.2 Sex



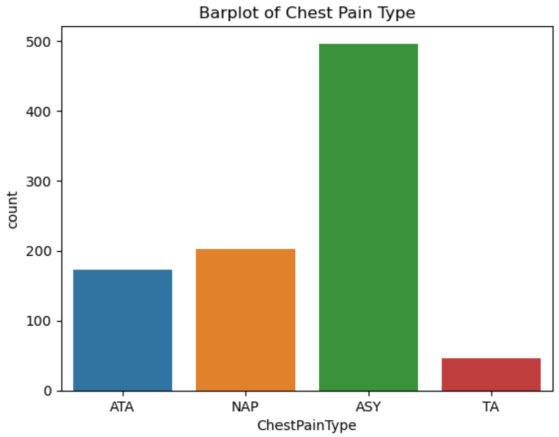

```
count     918
unique      2
top         M
freq      725
Name: Sex, dtype: object
```

Figure 2.2: Left: Barplot of patient's sex, Right: Summary Statistics

Sex is a discrete and nominal variable. There are 2 distinct values: "Male (M)" and "Female (F)". There is also no natural ordering.

There are no data issues with the variable age. There are no other unknown values. There seems to be significantly more male patients than female patients.

2.3 Chest Pain Type




```
count     918
unique      4
top       ASY
freq      496
Name: ChestPainType, dtype: object
```

Figure 2.3: Left: Barplot of patient's chest pain type, Right: Summary Statistics

ChestPainType is a discrete and nominal variable. There are 4 distinct values: "Typical Angina (TA)", "Atypical Angina (ATA)", "Non-Anginal Pain (NAP)" and "Asymptomatic (ASY)". There is also no natural ordering.

There are no data issues with the variable ChestPainType. There are no other unknown values. There seems to be significantly more patients with ASY compared to the other types of chest pain.

2.4 Resting Blood Pressure

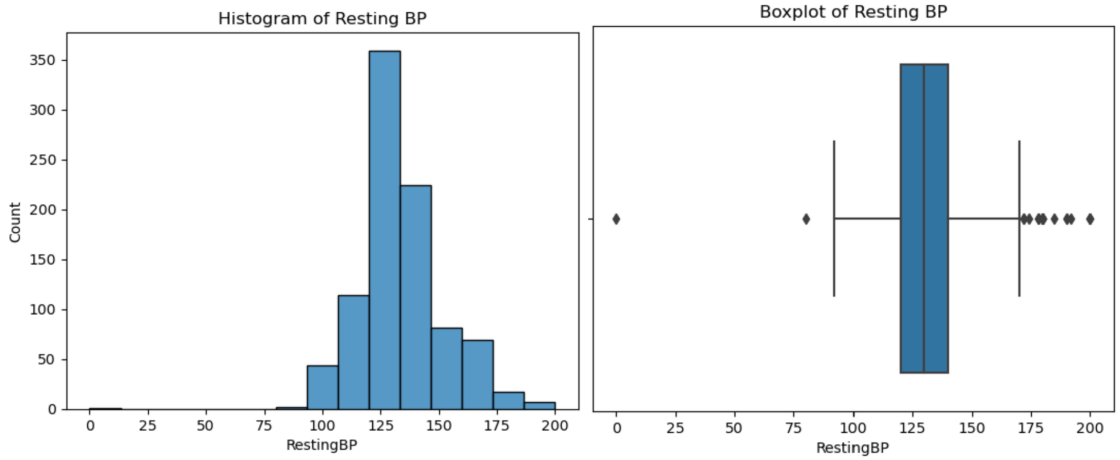


Figure 2.4 Top Left: Histogram of patient's resting blood pressure, Top Right: Boxplot of patient's resting blood pressure, Bottom Right: Summary Statistics

RestingBP is a continuous and ratio variable. It has a consistent interval between values. Blood pressure has a natural zero since 0 mm HG indicates a lack of blood pressure in the arteries.

```
count    918.000000
mean     132.396514
std       18.514154
min        0.000000
25%      120.000000
50%      130.000000
75%      140.000000
max      200.000000
Name: RestingBP, dtype: float64
```

There are some data issues with the RestingBP variable. A resting blood pressure of close to 0 mm HG is very unlikely and would typically not be compatible with life. Although there are several outliers as seen from the right side of the boxplot, these data points are still legitimate and relevant observations. The variable seems to be mostly normally distributed.

2.5 Cholesterol

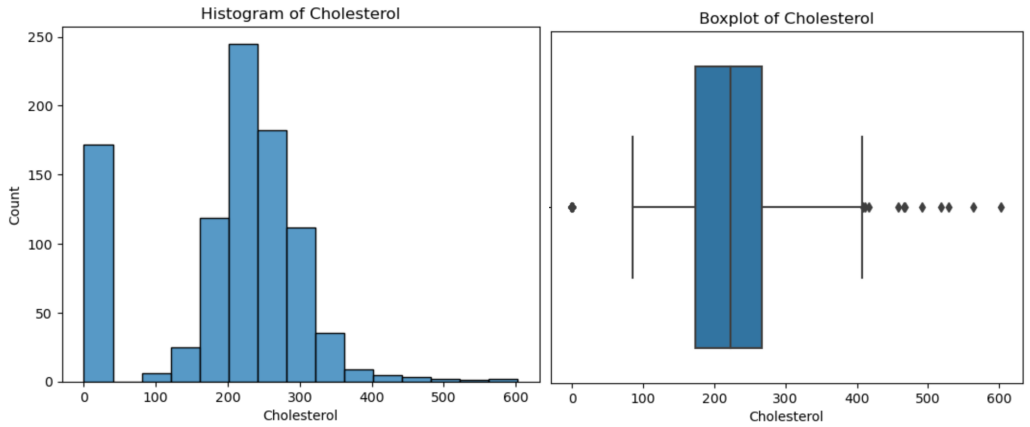


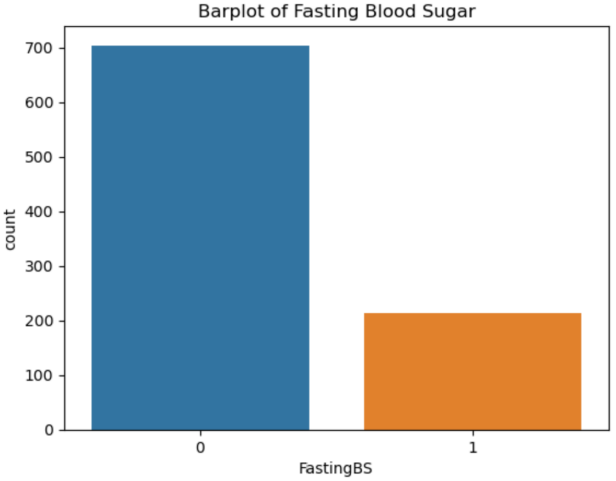
Figure 2.5: Top Left: Histogram of patient's cholesterol, Top Right: Boxplot of patient's cholesterol. Bottom Right: Summary Statistics

Cholesterol is a continuous and ratio variable. It has a consistent interval between values. Cholesterol has a natural zero since 0 mm/dL indicates the absence of cholesterol in a patient's body.

There are some data issues with the Cholesterol variable. There are values close to or equal to 0 mm/dL, and such low values are very unlikely and would typically not be compatible with life since humans have cholesterol naturally present in their bodies. Hence, these observations may not be accurate. Excluding 0 values, the variable seems to be normally distributed.

```
count    918.000000
mean     198.799564
std      109.384145
min        0.000000
25%      173.250000
50%      223.000000
75%      267.000000
max      603.000000
Name: Cholesterol, dtype: float64
```

2.6 Fasting Blood Sugar



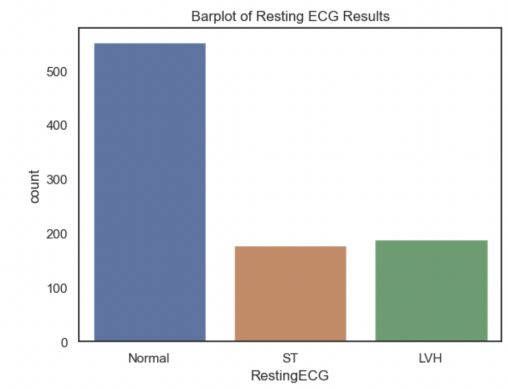

```
count       918
unique        2
top       False
freq        704
Name: FastingBS, dtype: object
```

Figure 2.6: Left: Barplot of patient with fasting blood sugar, Right: Summary Statistics

ChestPainType is a discrete and nominal binary variable. There are only 2 distinct values: "False (0)" and "True (1)". There is no natural ordering.

There are no data issues with the variable FastingBS. There are no other unknown values present in this variable. There seems to be significantly more patients with no fasting blood sugar compared to those who have.

2.7 Resting ECG Results




```
count        918
unique         3
top       Normal
freq         552
Name: RestingECG, dtype: object
```

Figure 2.7: Left: Barplot of patient's resting ECG results, Right: Summary Statistics

RestingECG is a discrete and nominal variable. There are 3 distinct values: "Normal", "ST" and "LVH". There is no natural ordering.

There are no data issues with the variable RestingECG. There are no other unknown values present in this variable. There seems to be significantly more patients with normal resting ECG results compared to those who do not.

2.8 Maximum Heart Rate Achieved

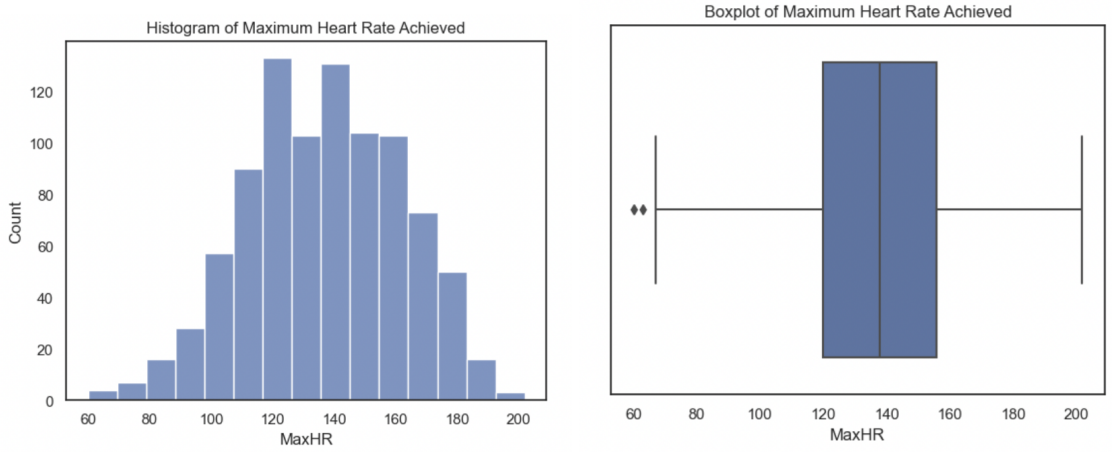


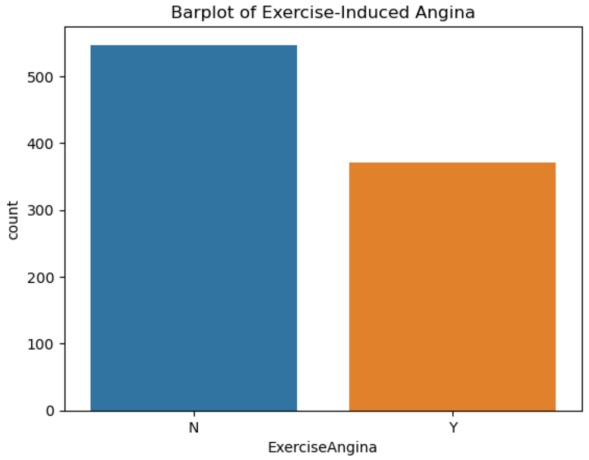
Figure 2.8: Top Left: Histogram of patient's maximum heart rate, Top Right: Boxplot of patient's maximum heart rate, Bottom Right: Summary Statistics

MaxHR is a continuous and ratio variable. It has a consistent interval between values. MaxHR has a natural zero since 0 indicates the absence of heartbeats in a patient's body.

There are no data issues with the variable MaxHR. Values lie within 60 and 202 which are all valid values for the variable. The variable seems to be normally distributed.

```
count    918.000000
mean     136.809368
std       25.460334
min       60.000000
25%      120.000000
50%      138.000000
75%      156.000000
max      202.000000
Name: MaxHR, dtype: float64
```

2.9 Exercise-Induced Angina




```
count     918
unique      2
top         N
freq      547
Name: ExerciseAngina, dtype: object
```

Figure 2.9: Left: Barplot of patient with exercise-induced angina, Right: Summary Statistics

ExerciseAngina is a discrete and nominal binary variable. There are only 2 distinct values: "Yes (Y)" and "No (Y)". There is no natural ordering.

There are no data issues with the variable ExerciseAngina. There are no other unknown values present in this variable. There seems to be more patients with no exercise-induced angina compared to those who do.

2.10 Oldpeak

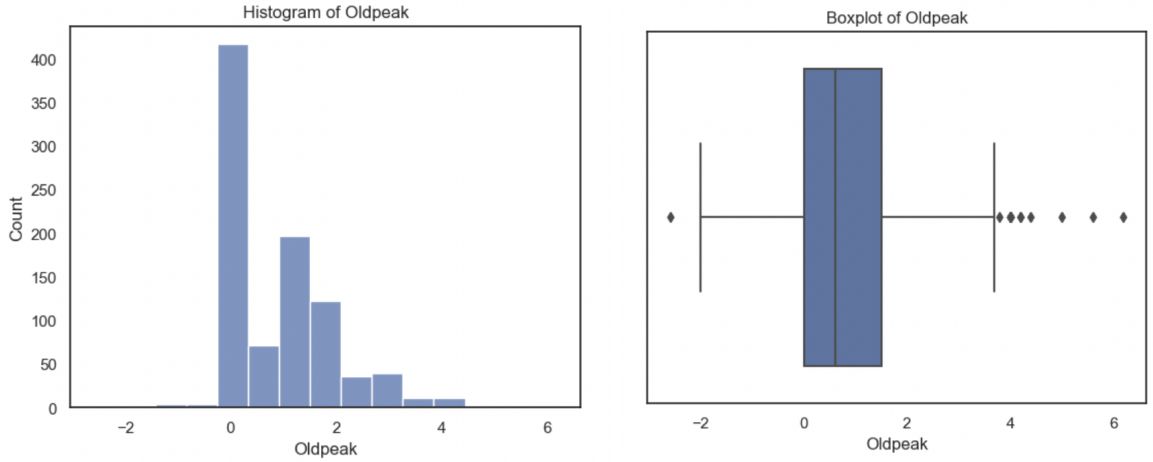


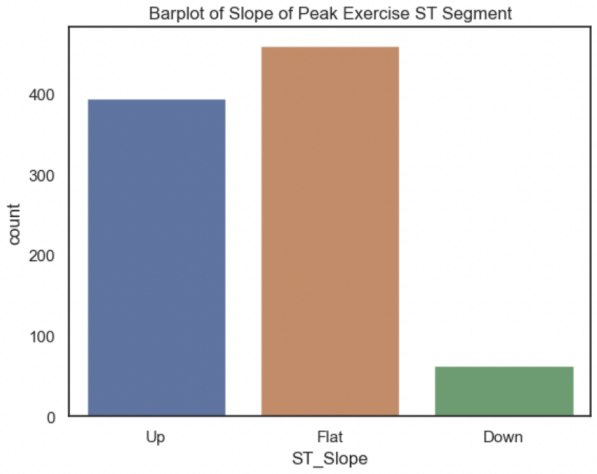
Figure 2.10: Left: Histogram of patient's old peak, Right: Boxplot of patient's old peak, Bottom Right: Summary Statistics

```
count    918.000000
mean       0.887364
std        1.066570
min       -2.600000
25%        0.000000
50%        0.600000
75%        1.500000
max        6.200000
Name: Oldpeak, dtype: float64
```

Oldpeak is a continuous and interval variable. Oldpeak quantifies the amount of depression of the ST segment from the baseline. It is likely measured in millimetres, representing the vertical distance between the baseline and the lowest point of the ST segment during a specific phase of the cardiac cycle. It has a consistent interval between values. Oldpeak does not have a true zero point. A value of 0 would indicate 0 depression. However, Oldpeak contains negative values which indicate an elevation or upward deviation.

There are no data issues with the variable Oldpeak. There seems to be some outliers with values 4-6mm as seen in the boxplot. However, these are likely to be valid observations. The variable does not seem to be normally distributed and seems right-skewed instead.

2.11 Slope of Peak Exercise ST Segment




```
count      918
unique       3
top       Flat
freq       460
Name: ST_Slope, dtype: object
```

Figure 2.11: Left: Barplot of patient's slope of peak exercise ST segment, Right: Summary Statistics

ST_Slope is a discrete and ordinal variable. There are 3 distinct values: "Up", "Flat" and "Down" which likely have a natural ordering. An upsloping pattern indicates heart rate increases with increased blood flow to the heart muscle. A flat line suggests that the ST segment remains relatively level without significant change to the blood supply. Lastly, a downsloping pattern likely suggests insufficient blood supply to the heart.

There are no data issues with the variable ST_Slope. There are no unknown values present in this variable. There seems to be more patients with upsloping patterns or flat slopes compared to patients who have downsloping patterns.

2.12 Heart Disease


```
count    918.000000
mean       0.553377
std        0.497414
min        0.000000
25%        0.000000
50%        1.000000
75%        1.000000
max        1.000000
Name: HeartDisease, dtype: float64
```

Figure 2.12: Left: Barplot of patient with heart disease, Right: Summary Statistics

HeartDisease is a discrete and nominal binary variable. This is also the target variable that shows if a patient has a heart disease or not. There is no natural ordering.

There are no data issues with the variable HeartDisease.

3. Data Visualisation and Principal Component Analysis

3.1 Correlation between Continuous Feature Variables

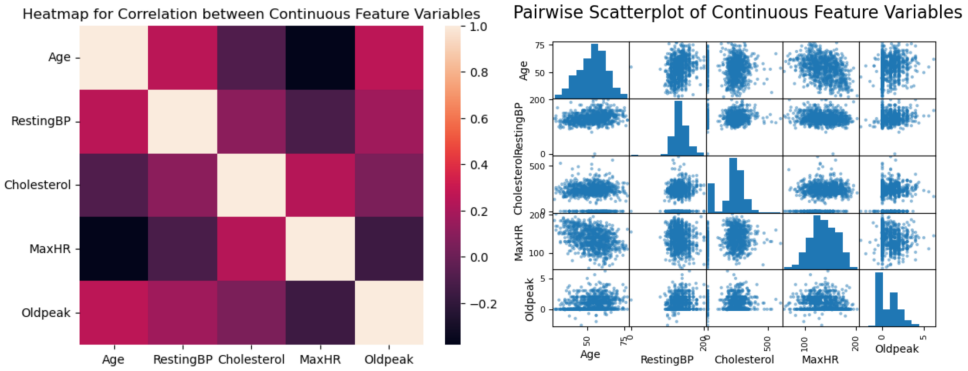


Figure 3.1: Left: Heatmap and pairwise scatter plot for correlation between continuous feature variables

The correlation between continuous feature variables are displayed above in the heatmap and pairwise scatterplots. Judging by the plots, the variables do not seem to have a strong correlation between each other.

3.2 One-hot Encoding for Categorical Variables

One-hot encoding is performed to convert categorical data into a numerical format that can be fed into machine learning models. To do this, n-1 dummy variables were created for each categorical variable with n representing the number of unique categories in each variable. A value of 0 is given for the absence of that category and a value of 1 is given for the presence of that category. For example, the sex variable has 2 distinct values: M or F. A dummy variable is created, Sex_M, where a value of 1 indicates that the patient is a male and a value of 0 indicates that the patient is a female. The same steps are performed for all the categorical variables, namely: Sex, ChestPainType, RestingECG, ExerciseAngina and ST_Slope. Note that the FastingBS variable is already expressed as a binary variable. Hence, the categorical variables are replaced by the following dummy variables.

Sex_M	ChestPainType_TA	ChestPainType_ATA	ChestPainType_NAP	RestingECG_N	RestingECG_ST	ExerciseAngina_Y	ST_Slope_U	ST_Slope_F
1	0	1	0	1	0	0	1	0
0	0	0	1	1	0	0	0	1
1	0	1	0	0	1	0	1	0
0	0	0	0	1	0	1	0	1
1	0	0	1	1	0	0	1	0
...	...	...	...	...	...	...	...	...
1	1	0	0	1	0	0	0	1
1	0	0	0	1	0	0	0	1
1	0	0	0	1	0	1	0	1
0	0	1	0	0	0	0	0	1
1	0	0	1	1	0	0	1	0

Figure 3.2: Dummy variables replacing categorical variables

3.3 Standardization of Variables

Principal Component Analysis (PCA) focuses on variance maximizing. Hence, it is important for features to be on a similar scale. This is to ensure that features with a large range and higher magnitude do not necessarily have a larger influence over the model due to their larger raw variance value. However, it is clear that the features are on different scales. For example, the "Age" variable has values ranging from about 30 to 80 while the “Cholesterol” variable has values ranging from about 100 to 600 (excluding 0 values). Hence, the values are standardised using the z-index standardisation with the below formula:

$$Z = \frac{x - \mu}{\sigma}$$

where z is the Z-index, x is the data point, μ is the variable mean and σ is the variable standard deviation.

3.4 Principal Component Analysis (PCA) Results

Principal Component Analysis (PCA) is performed to reduce the number of features (dimensions) in a dataset while preserving the most important information or variance. PCA is performed by using the singular value decomposition method and the following plot (Fig 3.3) is produced, showcasing the amount of variance explained by each principal component individually and cumulatively. PCA was performed using all feature variables, including continuous and categorical variables after performing one-hot encoding, giving a total of 14 features. The reason for including the categorical variables is to allow PCA to consider the relationships and patterns involving those variables. This can help uncover hidden structures and interactions in the data, considering that these variables may be important in determining if a patient has heart disease.

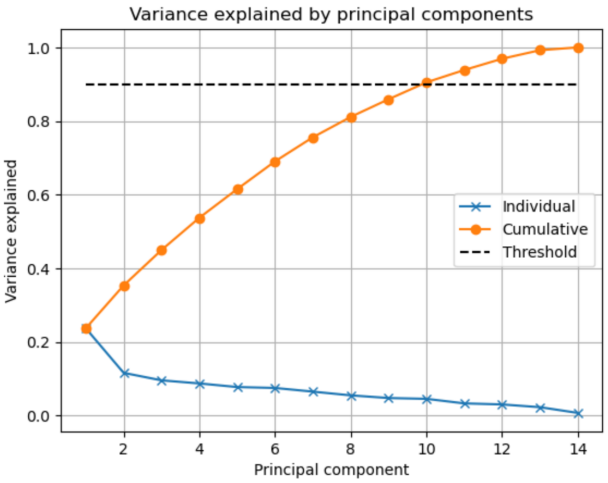


Figure 3.3: Graph showcasing the amount of variance explained by principal components individually and cumulatively

The above graph describes the amount of variation explained as a function of the number of PCA components we include. From the graph, it takes about **10 principal components** to explain about **90% of the variance** in the dataset.

3.5 Principal Directions of Considered PCA Components

The first two principal components are considered for this exercise. The component coefficients against the feature variables are plotted in Figure 3.4. Considering principal component 2 (PC2), it is observed that normal RestingECG (RestingECG_N) has a large negative coefficient while RestingECG with a value of ST (RestingECG_ST) has a large positive coefficient. This PC makes it easy to differentiate patients with normal ECG results from patients with ST results for their ECG.

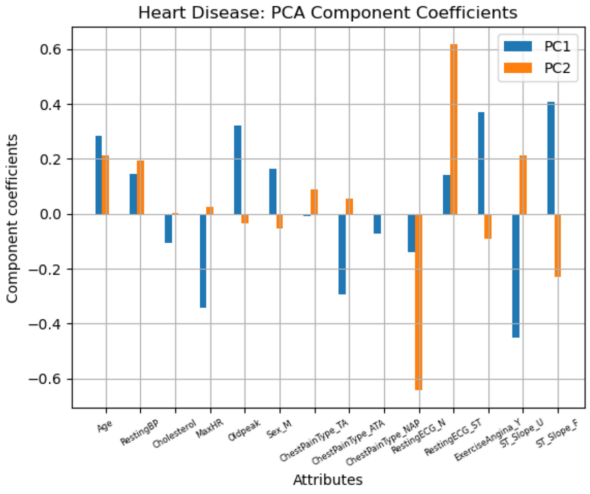
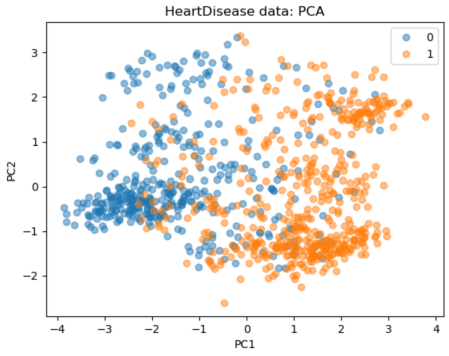


Figure 3.4: Graph on component coefficients against feature variables for PC1 and PC2

3.6 Data Projected onto the Principal Components (PC1 and PC2)




```
PC2:
[ 0.21393113  0.19398884  0.00238733  0.02493241 -0.03364273 -0.05464801
  0.0904244   0.05560666 -0.0052394  -0.64252912  0.61787847 -0.08939433
  0.21190349 -0.23032197]


Patient with normal ECG (positive value for RestingECG_N and negative value for RestingECG_ST)
[-1.43235901  0.41068502  0.82462075  1.38217478 -0.83197888  0.51567132
 -0.22955354  2.07404613 -0.53254747  0.81383119 -0.49018212 -0.82310761
  1.15004709 -1.00163503]
...and its projection onto PC2
-0.3709593987033945


Patient with ST ECG (negative value for RestingECG_N and positive value for RestingECG_ST)
[-1.75040438 -0.12944227  0.7697682  -1.5243071  -0.83197888  0.51567132
 -0.22955354  2.07404613 -0.53254747 -1.22741754  2.03783579 -0.82310761
  1.15004709 -1.00163503]
...and its projection onto PC2
2.25719514804321
```

Figure 3.5: Left: Graph showcasing the projection of data onto PC1 and PC2, Right: projected value of data on PC2

Considering a patient with normal ECG results, where RestingECG_N =1 and RestingECG_ST = 0 (before standardization), the projected data onto PC2 will likely be negative. A patient with ST results for their ECG, where RestingECG_N = 0 and RestingECG_ST = 1 (before standardization) will likely have the projected data onto PC2 be positive. Refer to Fig 3.5 (Right).

4. Discussion on Findings

From our exploratory data analysis, multicollinearity is unlikely an issue given that the feature variables have a low degree of correlation between each other. There is also no class imbalance problem as the proportion of patients with heart disease is comparable to the proportion of patients with no heart disease as seen from the barplot.

Prior to running PCA, we have standardised our attributes to be on a similar scale before applying them to the models. This is especially important for models which we will implement in Project 2, which include logistic regression and neural networks that use gradient descent, as well as Support Vector Machine which runs on a distance-based algorithm, hence it is also affected by the range of features.

From the PCA, it is clear that each principal component does not explain much of the variance individually. The primary source of variation in the dataset is likely not well-aligned with the direction of the principal components. We also require a relatively huge proportion of the principal components (10 out of 14) to explain over 90% of the variance, and each principal component still explains a substantial new amount of variance.

Overall, our primary machine learning modelling aim seems feasible based on our analysis and visualisations. Our continuous variables are not strongly correlated to one another and are mostly normally distributed, there are no unknown values in our data and the rare outlier values obtained are mostly still valid and medically possible measurements. Through our visualisations, we have also managed to identify certain relationships between the attributes and the principal components, as well as when it comes to projecting the data onto the principal components as well. Given that we have both discrete and nominal target variables for classification and continuous and ratio variables for regression to predict, we are optimistic in fulfilling our primary machine learning modelling aim.

EXAM QUESTIONS

Question 1: Option D. Time of day is an interval variable because it is expressed in 30-mins interval. Hence, the difference in time can be measured via addition/subtraction of these intervals. Traffic lights is a ratio variable because 0 means there were no broken traffic lights. Likewise, running over is a ratio variable since 0 means there were 0 run over accidents. Congestion level is an ordinal variable since a value of high indicates greater higher levels of congestion compared to a value of low.

Question 2: Option A. Using the p-norm formula, If $p = \infty$, then the p-norm distance between the 2 vectors is the maximum absolute difference between the elements of two vectors. In other words, $||x14 - x18||\infty = \max(|x14_i - x18_i|)$ for all i. max(26-19, 0-0, 2-0, 0-0, 0-0, 0-0, 0-0) = 7. Hence, $||x14 - x18||\infty = 7.0$.

Question 3: Option A. The explained variance is measured by the sum of squares of selected singular values over the total sum of squares of all the singular values. Using this formula and the singular values from the diagonal entries of matrix S, we realise that only option A correctly describes the proportion of variance explained.

Question 4: Option D. Each row of the matrix V is associated with a variable with the first row being associated with variable x1, second row being associated with variable x2, etc. Each column corresponds to a principal component with the first column being the first principal component and second column being the second principal component, etc. For each option, we take the principal component of interest, a column vector, and find the dot product with a vector with values for each variable in each option. If a value is not specified for a variable in the option, we assume them to have the value of 0 in the vector when performing dot product. We then use the result of the dot product to estimate if it is likely a positive or negative value to answer the question. Only option D is consistent using this method.