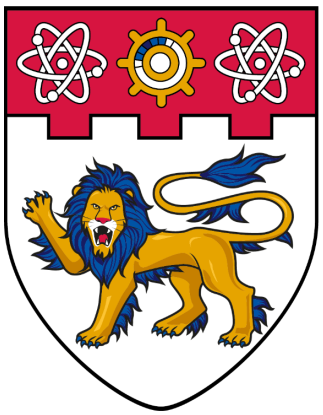


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BC2406 – Analytics I

Early Detection of Cardiovascular Diseases

Name	Matriculation Number
Ng Jun Long	U2110010D
Goh Xin Yi Rachel	U2111391F
Chow Yan Yu	U2011651J
Sherman Yeo	U2110985F
Tang Ming Wei	U2110861E
Nur Aisyah Bte Abdul Mutalib	U2110399H

Table of Contents

Executive Summary

This report seeks to examine and determine significant risk factors affecting cardiovascular diseases (CVD) as well as provide predictive models that allow National Heart Centre Singapore (NHCS) to detect early symptoms among individuals who may potentially be diagnosed with CVD. We believe that these models will help NHCS reduce unnecessary usage of healthcare resources brought about by delayed diagnoses of CVD, minimising NHCS' costs. Based on our current findings, we will also provide business recommendations for NHCS.

Dataset and Proposed Proof of Concept (POC)

The team has conducted in-depth exploratory data analysis (EDA) on our dataset and optimised it for our POC models. We came up with six different models throughout this project, utilising two algorithms:

1. **Logistic Regression:** A predictive model used to determine significant variables that affect the presence of heart disease in our patients, providing our clients and NHCS greater confidence in determining whether patients have heart diseases.
2. **Classification & Regression Tree:** An algorithm that generates a decision tree allowing us to identify key variables and to predict heart diseases.

These models will help us gain a better understanding of the variables that lead to heart diseases better and thus, empower NHCS in the road to CVD diagnosis among patients.

1. Business Understanding

1.1. Background on National Heart Centre Singapore (NHCS)

NHCS is a one-stop referral clinic for individuals with cardiovascular diseases (CVD) – providing comprehensive cardiac care ranging from preventive, diagnostic, therapeutic to rehabilitative services. NHCS handles over 120,000 outpatient consultations, 9,000 interventional and surgical procedures and 10,000 inpatients yearly (National Heart Centre Singapore, 2021).

1.2. Business Problem

Amongst the various long-term side effects brought about by COVID-19, one of the more prominent ones includes a substantial rise in the risk of cardiovascular disease (Sidik, 2022). This is an urgent cause of concern because CVD already accounts for 32% of all deaths in Singapore in 2021 (Singapore Heart Foundation, 2022).

The road to the diagnosis of CVD is often lengthy and convoluted. Hence, an early and accurate diagnosis of CVD is critical as it ensures rapid access to treatment, reducing the risk of long-term complications and preventing early deaths. Furthermore, the key problem in the delayed diagnoses of CVD is that patients are likely to suffer more symptoms, take more medications and tend to utilise more resources, which could be avoided with an early detection of CVD symptoms (Thompson & Yancy, 2004).

1.3. Current Approaches

A current solution implemented to detect symptoms of heart diseases is the use of electrocardiograms (ECG). ECG records the electrical signals from the heart and checks it against certain heart diseases. Although the accuracy rate of using ECG reaches nearly 98.5%, the use of ECG alone is insufficient as it can only detect three different types of heart diseases - Arrhythmias, Coronary Heart Disease and Heart Attacks (Auto, 2021). Heart diseases such as non-ST[1] segment elevation myocardial infarction (NSTEMI).

In addition, currently, some people who go to hospitals with symptoms of a heart attack, undergo a series of tests which take three to six hours or longer to determine if they have indeed suffered a heart attack (American Heart Association, 2018).

1.4. Opportunity Statement

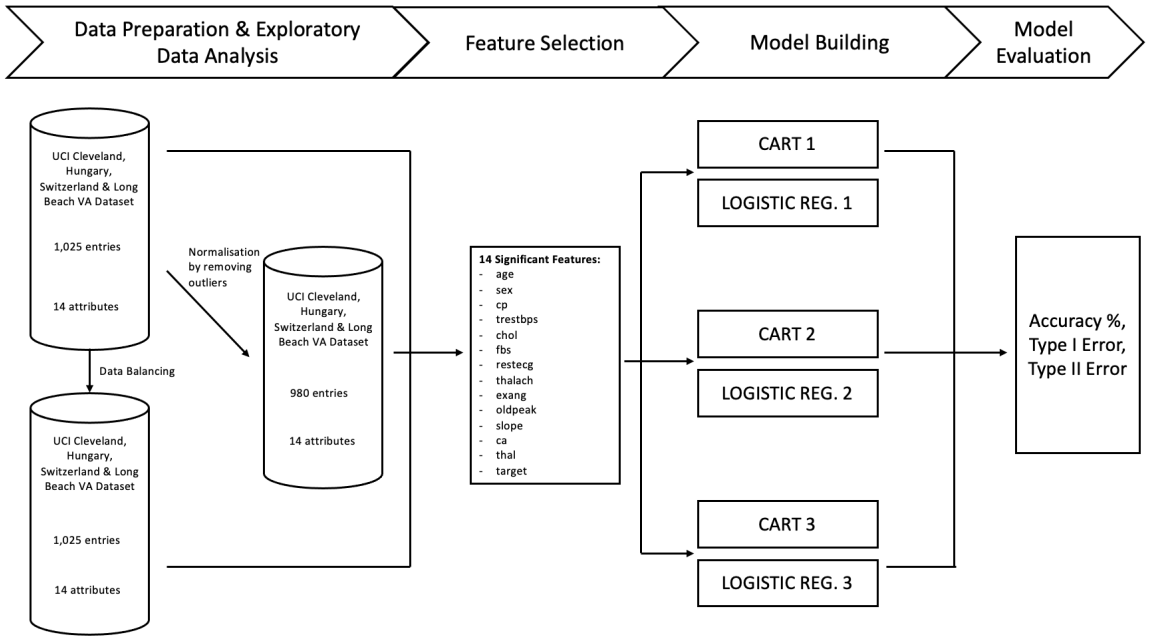
Our group aims to examine and determine significant risk factors of CVD as well as provide predictive models that allow NHCS and citizens to use predictive models to detect early symptoms among individuals who could be diagnosed with CVD. Based on our current findings, we will also provide business recommendations for NHCS. These results empower NHCS with the ability to predict heart diseases more accurately, reducing the resources spent on delayed diagnosis of CVD. This ultimately improve NHCS's cost efficiency.

---

[1] Non-ST-elevation myocardial infarction (NSTEMI) is **an acute ischemic event causing myocyte necrosis**. The initial ECG may show ischemic changes such as ST depressions, T-wave inversions, or transient ST elevations; however, it may also be normal or show nonspecific changes.

2. Methodology

Our hypothesis is that the presence of heart disease in an individual can be predicted through variables such as age, sex, type of chest pain, resting blood pressure, serum cholesterol, maximum heart rate achieved, presence of thalassemia, etc. To draw a conclusion for this, we use the following methodology (Figure 1), consisting of four main stages – data preparation and EDA, feature selection, model building as well as its evaluation.



3. Data Preparation

3.1. Data Planning

We gathered research on common risk factors of heart disease and heart attacks to ensure that the dataset we chose included variables related to heart disease. Our research shows that there are four conventional risk factors of coronary artery disease, which is the most common type of heart disease: cigarette smoking, diabetes (i.e., high fasting blood sugar), hyperlipidemia (i.e., high blood cholesterol), and hypertension (i.e., high blood pressure) (Khot et al., 2003).

Other major risk factors include: age, gender, family history, obesity, poor diet, physical inactivity and stress (Brown et al., 2022). The risk of heart disease increases with age, and males are at a higher risk of heart disease than females (Hajar, 2017). Although factors such as age, gender and family history cannot be modified or controlled, they can be useful in predicting heart disease.

With ‘the presence of heart disease’ being our categorical Y variable, we decided to use logistic regression and CA as our prediction models. We would then compare the accuracy of the two models in predicting the presence of heart disease.

3.2. Data Sources

Our dataset was taken from Kaggle, a well-known platform that hosts data science competitions and allows users to publish and download datasets, among other features. However, due to the fact that Kaggle is an open source platform, its data has to be scrutinised as the data could be poorly recorded due to a lack of recording standardisations.

3.3. Dataset

The dataset has the file name *heart.csv* and the dataset is a combination of four databases from Cleveland, Hungary, Switzerland, and Long Beach with patient records from 1988.

3.3.1. Y Variable

The Y variable we identified is binary and indicates the presence of heart disease in the patient, where ‘0’ = the absence of heart disease in the patient, and ‘1’ = the presence of heart disease in the patient.

3.3.2. X Variables

We took the rest of the variables in the dataset as X variables as they were all relevant predictors. They are: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, and thal. The data dictionary can be found in **Appendix A**. The dataset contains variables that cover three of the four conventional risk factors of heart disease, with the exception being smoking.

3.4. Data Cleaning

Data cleaning was carried out for all of the variables in the dataset. We considered the values "NA", "missing", "N/A", "", "m", "M", "na", and "." to be missing values and re-coded them to NA for easier data pre-processing.

The table below is a summary of the overall data cleaning process:

Variable	**Action**	**Reasoning**
Categorical variables	Data type changed to “factor” (**Appendix B**).	This factorises the data into a factored data type which allows R to better understand the nature of the data and support us in our analysis.
Continuous variables	Data type left as “integer” or “numeric”	For continuous variables, we decided to leave their class as their integer or numeric as their class is continuous in nature. We also checked that the continuous variables are not highly skewed, i.e., skewness not more than 3 (Kline, 2015).

	Created boxplots (**Appendix C**).	This would help us to check for potential outliers that may result in distortions in our findings from modelling.
All Variables	Checked for NA values (**Appendix D**).	This helps us find NA values that may either have to be replaced with the median or removed.
sex	Replaced 0 and 1 with “Female” and “Male” respectively.	This formats each of these categorical variables from integers into specific worded categories, making them easier to understand (Appendix E).
fbs	Replaced 0 and 1 with “False” and “True” respectively.	
restecg	Replaced 0, 1 and 2 with “Normal”, “ST-T Abnormality” and "Hypertrophy" respectively.	
cp	Replaced 0, 1, 2 and 3 with "Typical Anginal", "Atypical Anginal", "Non-anginal Anginal" and "Asymptomatic" respectively.	
slope	Replaced 0, 1 and 2 with “Upsloping”, “Flat” and “Downsloping” respectively .	
exang	Replaced 0 and 1 with “No” and “Yes” respectively.	
thal	Replaced 1, 2 and 3 with “Normal ”, “Fixed Defect” and “Reversible Defect” respectively.	

The dataset from Kaggle was relatively clean and there were no NA values, but outliers were found. Although outliers may skew the data, they may contain important information and

keeping these outliers could possibly reduce the bias of risk predictions. On the other hand, removing the outliers may improve the calibration of risk predictions (Moons et al., 2019). With these two arguments in mind, we decided to compare prediction models of a dataset with outliers and a dataset without outliers.

4. Data Exploration

In order to have a better idea of the relationship between the variables, we have performed various plots between every X variable against the target categorical Y variable. We have also cross compared our X variables against other X variables to see if there are any interesting visual relations.

4.1. Categorical X variables against Categorical Y variable

Firstly, we used barcharts to show the relationship between the categorical X variables and the Y variable (**Appendix F**). We discovered that the proportion of females getting heart disease is higher than that of males (sex). We also found heart disease is most prominent in non-anginal anginal chest pain out of the types of chest pain (cp), in patients with fasting blood sugar > 120 mg/dl (fbs), ECG result of ST-T Abnormality out of the various ECG results (restecg) and patients with lack of exercise induced angina (exang). Heart disease appears most across 1 major vessel (ca) and appears the most frequent in fixed defects out of the defect types (thal). Lastly, it is also most frequent in downsloping peak exercise (slope).

After looking at the barchart, we found out that the variable 'thal' has a category '0'. As '0' does not represent any defect type, we removed records where thal == '0' and plotted another barchart to ensure the category '0' was removed (**Appendix F**).

4.2. Continuous X variables against Y variable

We used density plots to show the relationship between the continuous X variables and the Y variable (**Appendix G**). We have found that heart disease occurs most in patients between 50-60 (age), resting blood pressure between 120-135 bpm (trestbps), cholesterol level between 200-250 (chol), maximum heart rate between 150-175 bpm (thalach) and ST depression induced by exercise relative to rest of 0 (oldpeak).

4.3. Continuous X variables against continuous X variables

We created scatter plots with trend lines to determine the trends between every 2 continuous X variables (**Appendix H**). We also created a correlation matrix to determine the correlation between these variables (**Appendix I**). The table below briefly describes each set of variables:

Variables	Trend	Correlation coefficient
age vs trestbps	As age increases from, trestbps decreases slightly, then starts increasing when age is approximately 45. Once age	positive

	increases to 65 trestbps starts decreasing again	
age vs chol	As age increases, chol increases with decreasing gradient	positive
age vs thalach	As age increases, thalach decreases with very slightly increasing gradient	negative
age vs oldpeak	As age increases, oldpeak generally increases until age is approximately 62.5, after which it decreases	positive
trestbps vs chol	As trestbps increases, chol increases	positive
trestbps vs thalach	As trestbps increases, thalach remains generally constant	negative
trestbps vs oldpeak	As trestbps increases, oldpeak increases	positive
chol vs thalach	As chol increases, thalach increases, then decreases sharply, then increases again	negative
chol vs oldpeak	As chol increases, oldpeak decreases until chol is approximately 250, after which it increases	positive
thalach vs oldpeak	As thalach increases, oldpeak decreases	negative

4.4. Categorical X variables against continuous X variables

Next, we delved deeper into the relationship between the different types of our X variables. We created various violin plots that show the changes in density of each categorical X variable as each continuous X variable increases (**Appendix J**).

4.5. Categorical X variables against categorical X variables

To show the relationship between the categorical X variables, bar plots have been created. These bar plots mainly portray the relationship between sex and the other categorical X variables (**Appendix K**).

5. Model Training & Evaluation Methods

5.1. Train-Test Split

We performed a train-test split method based on a 70:30 train-test ratio. The split allows us to better estimate the performance of our predictive algorithms when they are used to predict

our categorical Y variable. The team has chosen to employ the 70:30 train-test ratio after empirical studies show that it yields the best result. (Gholamy, A., Kreinovich, V., & Kosheleva, O. , 2021).

5.2. Logistic Regression Model

```
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               2.816095   1.992361   1.413 0.157525
age                       0.044108   0.017328   2.545 0.010912 *
sexMale                  -2.255043   0.389678  -5.787 7.17e-09 ***
cpAtypical Anginal        1.050427   0.382588   2.746 0.006040 **
cpNon-anginal Anginal     1.873883   0.342780   5.467 4.58e-08 ***
cpAsymptomatic            2.083762   0.449073   4.640 3.48e-06 ***
trestbps                 -0.030286   0.007948  -3.810 0.000139 ***
chol                     -0.007989   0.002743  -2.912 0.003586 **
fbsTrue                   0.333013   0.386692   0.861 0.389136
restecgST-T Abnormality   0.194041   0.261479   0.742 0.458033
restecgHypertrophy       -1.903312   2.049780  -0.929 0.353125
thalach                   0.026566   0.007773   3.418 0.000631 ***
exangYes                 -0.838738   0.293615  -2.857 0.004282 **
oldpeak                  -0.205153   0.147801  -1.388 0.165126
slopeFlat                -0.818823   0.572303  -1.431 0.152501
slopeDownsloping          0.573312   0.611114   0.938 0.348171
ca1                      -2.236863   0.360777  -6.200 5.64e-10 ***
ca2                      -4.016805   0.544119  -7.382 1.56e-13 ***
ca3                      -2.203194   0.580849  -3.793 0.000149 ***
ca4                       1.609779   1.011180   1.592 0.111389
thalFixed Defect         -0.484292   0.536309  -0.903 0.366521
thalReversible Defect    -1.731899   0.526921  -3.287 0.001013 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 986.48  on 711  degrees of freedom
Residual deviance: 423.86  on 690  degrees of freedom
AIC: 467.86
```

Figure 2. Logistic Regression Model Output

Our team decided on a logistic regression due to the prediction of categorical variable Y. Logistic regression provides consistent and reliable results in identifying meaningful relationships between X variables and our target categorical Y variable. The logistic model will give us insights in identifying the most significant variables that contribute to a person having heart disease. The trained model will then be used to predict whether a person has heart disease.

5.2.1. Using AIC To Determine Optimal Logistic Model

```
(Intercept)            3.011451   1.936547   1.555 0.119931
age                    0.042583   0.017253   2.468 0.013580 *
sexMale               -2.231947   0.388461  -5.746 9.16e-09 ***
cpAtypical Anginal     1.082315   0.381524   2.837 0.004557 **
cpNon-anginal Anginal  1.966008   0.337932   5.818 5.96e-09 ***
cpAsymptomatic         2.114505   0.446393   4.737 2.17e-06 ***
trestbps              -0.030088   0.007641  -3.938 8.22e-05 ***
chol                  -0.008320   0.002666  -3.120 0.001806 **
thalach                0.027009   0.007797   3.464 0.000533 ***
exangYes              -0.806522   0.292702  -2.755 0.005861 **
oldpeak               -0.244232   0.144980  -1.685 0.092068 .
slopeFlat             -0.807493   0.557470  -1.448 0.147479
slopeDownsloping       0.569617   0.594922   0.957 0.338333
ca1                   -2.223993   0.351805  -6.322 2.59e-10 ***
ca2                   -3.888310   0.526989  -7.378 1.60e-13 ***
ca3                   -2.209274   0.566593  -3.899 9.65e-05 ***
ca4                    1.694960   1.016449   1.668 0.095409 .
thalFixed Defect      -0.531739   0.518631  -1.025 0.305234
thalReversible Defect -1.746288   0.508205  -3.436 0.000590 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 986.48  on 711  degrees of freedom
Residual deviance: 426.40  on 693  degrees of freedom
AIC: 464.4
```

Afterwhich, we used backward stepwise regression to reduce the number of insignificant variables in our dataset starting from the least significant variable. The backwards stepwise regression model allows us to discover only variables with high significance, this allows us to develop the optimal solution in using logistic regression to predict our categorical Y variable.

The AIC score is a good measure for the goodness of fit for the model and the lowest AIC score in the stepwise regression would indicate that the model is the best fit. Specifically, our team will choose the logistic regression model that produces the lowest AIC value.

5.2.2. Odds ratio and confidence interval of OR

```
> OR <- exp(coef(heart.log2))
> OR
          (Intercept)                   age               sexMale    cpAtypical Anginal
          20.31685010            1.04350291            0.10731927            2.95150407
cpNon-anginal Anginal        cpAsymptomatic              trestbps                  chol
           7.14211003            8.28548571            0.97036033            0.99171490
              thalach              exangYes               oldpeak             slopeFlat
           1.02737671            0.44640805            0.78330595            0.44597476
     slopeDownsloping                   ca1                   ca2                   ca3
           1.76758936            0.10817629            0.02047993            0.10978028
                  ca4      thalFixed Defect thalReversible Defect
           5.44642706            0.58758233            0.17442015
```

The odds ratio gives us a good representation on the impact that each X variable has on the Y variable. An Odd Ratio > 1 would indicate that the odds of Y increasing as the value of the X variable increases are favourable, and if Odd Ratio < 1, the odds of Y increasing as the value of the X-variable increases are less than the former. From this result, we can conclude that chest pain is the variable that has the largest impact on whether a person has heart disease or not. For example, a person with Non-anginal chest pain would increase the odds of having heart disease by a factor of 7.14.

```
> OR.CI
                              2.5 %       97.5 %
(Intercept)           0.476243716 963.97476665
age                   1.009103387   1.07988653
sexMale               0.048790335   0.22473360
cpAtypical Anginal    1.417413263   6.36302277
cpNon-anginal Anginal 3.733196084  14.09524339
cpAsymptomatic        3.513832077  20.32679739
trestbps              0.955583518   0.98469607
chol                  0.986543719   0.99697934
thalach               1.012218669   1.04374466
exangYes              0.250676929   0.79164829
oldpeak               0.585958342   1.03612865
slopeFlat             0.147244286   1.31794812
slopeDownsloping      0.538915958   5.59197092
ca1                   0.053268205   0.21229603
ca2                   0.006940369   0.05517874
ca3                   0.033880599   0.31595767
ca4                   0.796633688  42.94591370
thalFixed Defect      0.211848733   1.63433510
thalReversible Defect 0.063968210   0.47418737
```

The confidence interval of the odds ratio gives us an expected range for the true odds ratio for the population to fall within. For example, the table shows a 95 % confidence level that the odds ratio of a person with non anginal chest pain will fall between 3.733 and 14.095.

5.2.3. Preventing Multicollinearity Using VIF

```
> vif(heart.log2)
              GVIF Df GVIF^(1/(2*Df))
age       1.551462  1        1.245577
sex       1.804997  1        1.343502
cp        1.894698  3        1.112389
trestbps  1.317556  1        1.147849
chol      1.309289  1        1.144242
thalach   1.554741  1        1.246892
exang     1.181324  1        1.086887
oldpeak   1.694344  1        1.301670
slope     2.071999  2        1.199768
ca        2.161284  4        1.101131
thal      1.432659  2        1.094046
```

To ensure that there is no multicollinearity within our logistic models. We conducted a Variance Inflation Factor (VIF) test. The general rule of thumb is that VIF should be less than 10 for continuous variables, and Generalised Variance Inflation Factor (GVIF) should be less than 2 for categorical variables. As the logistic model contains both continuous and categorical variables, the vif function in R returns GVIF, and we look at the third column, $GVIF^{(1/(2*Df))}$, to check that the VIF thresholds have not been crossed. Several discussions show that $GVIF^{(1/(2*Df))}$ is a better metric to measure the presence of multicollinearity (Alteryx, 2019). Hence we have decided to use that instead. Since the VIFs for all variables are less than 10, there is no multicollinearity and all the variables provide essential value to the model. Evaluating Model Performance

```
           log.predict.test
test.Actual   0   1
          0 125  24
          1  13 144
> #Overall accuracy on testset
> mean(log.predict.test == testset$target ) #87.9%
[1] 0.879085
```

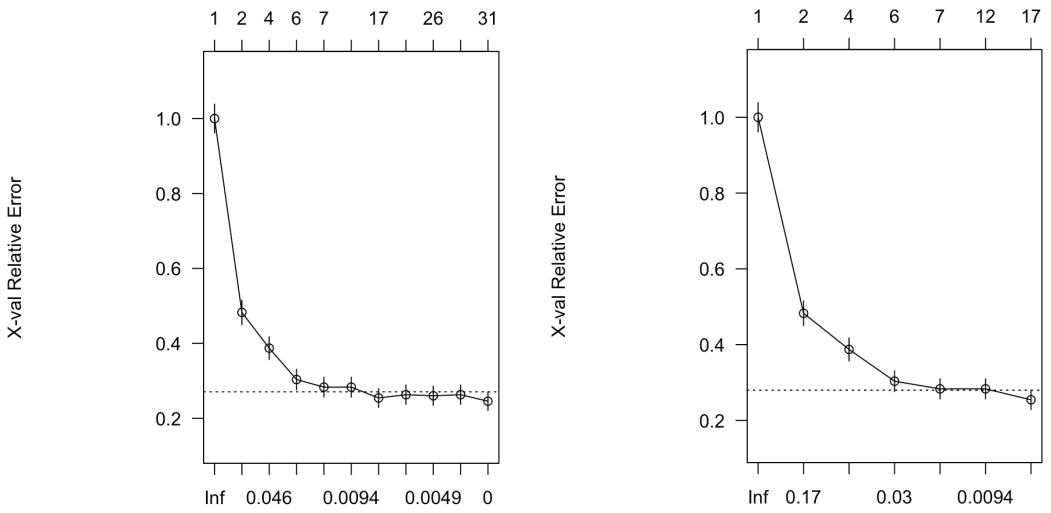
With the trained model using the trainset data, we proceed to test the accuracy of the model on the testset data. This will give us an accurate evaluation of the model. Our model shows a 87.9 % accuracy of predicting whether a person has heart disease with the current logistic regression model.

5.3. Classification and Regression Tree Model (CART)

The team has chosen to employ the use of a CART model as a supplementary analysis. CART produces a set of rules which are visualised as a binary decision tree. CART models start by identifying the root node with the highest significance in affecting our target Y variable. We are also able to adjust the minsplit value to prevent overfitting of our CART model. CART is also able to give us the ability to traverse the tree and evaluate each input based on the underlying decisions stated in the node of the tree.

5.3.1. Understanding CART Model

We start by growing our decision tree to the maximum, this is done so by setting the complexity parameter to 0. The complexity parameter imposes a penalty to the tree for having too many splits, by setting it to 0 there will be no penalty which allows us to obtain the maximum tree size possible. Next, we set the method to be “class” as our Y variable is categorical. We then set the minsplit to be = **15.** The min split parameter tells us the smallest number of observations in the parent note that could be split further. This value is thus determined by the size of your dataset and we decided that minsplit = 15 would give us the best tree that avoids overfitting or underfitting. The last step of the CART modelling is to prune the full sized tree to reduce the size of our decision tree, this is done so by eliminating nodes that do not provide significant classification power.

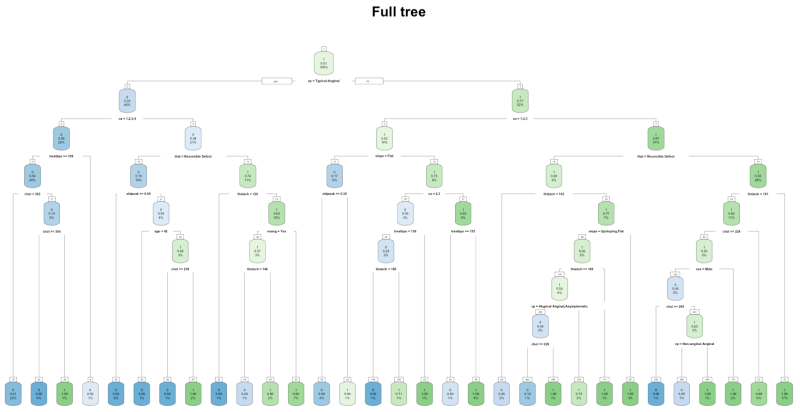


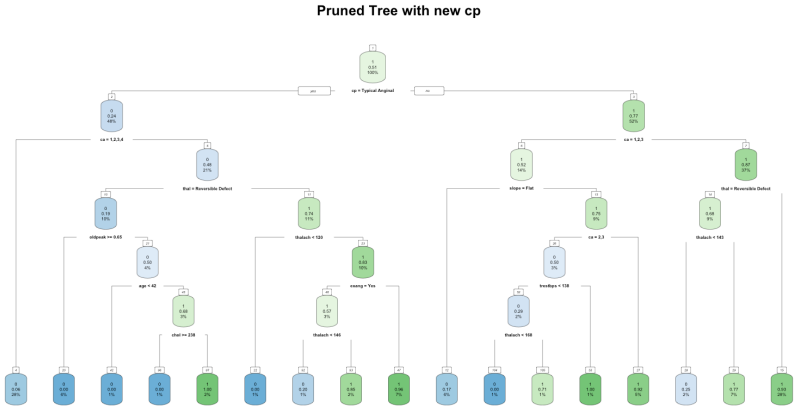
Before pruning

After pruning

An example of the plotcp graph is shown above

The plotcp graph gives us a visual representation of the cross validation results. The first graph shows that the 6th node will give us the most optimal tree where it first crosses below the line. From the second plot, we can see that there are only 7 nodes left after deleting the child nodes of a branch node.



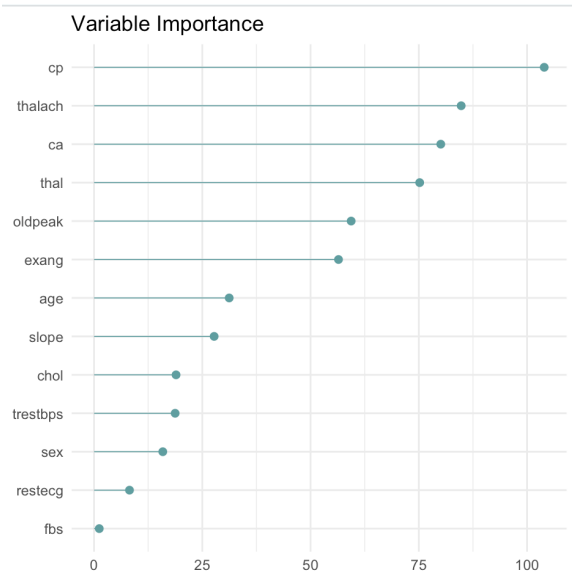


Visualisation of full tree vs pruned tree

As shown in the full tree, we can observe that the tree is larger. There is a possibility of overfitting since it is the full tree. After pruning the tree, we get a smaller tree where the final subsets are homogeneous in terms of the outcome variable.

```
> #Accuracy on test set
> cart.predict.test <- predict(m2.cart, newdata= testset, type = 'class')
> result.test <- table(Actual = testset$target, cart.predict.test, deparse.level = 2)
> mean(testset$target == cart.predict.test) #89.21%
[1] 0.8921569
```

After pruning the tree, we get our final model. Similar to logistic regression, we use the trained model using the trainset to test the accuracy on our testset. This cart model shows an 89.21% accuracy in predicting if a person has heart disease.



Variable importance plot

In the final step, we check for the variable importance of our X variables in predicting our target variable Y. Firstly, we checked for the variable importance and then placed it into a graphical representation using ggplot2. The results show the ranking of the X variables that are more significant in heart disease prediction in a descending order. In this case, cp represents type of chest pain, which tells us that the type of chest pain is the most important variable to determine if a patient has a heart disease using our CART model with an accuracy of 89.21%.

5.4. Comparison Of All Models

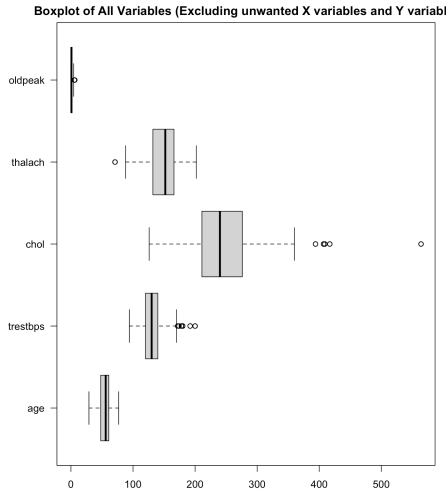
In order to decide the best model for heart disease prediction. We ran both logistic regression and CART in three different scenarios.

Scenario 1	CART and Logistic Regression **with** outliers
Scenario 2	CART and Logistic Regression **without** outliers
Scenario 3	CART and Logistic regression after **balancing**

Scenario 1 was showcased above (Sections 5.2 and 5.3) while explaining both models.

5.4.1. Scenario 2: Removing outliers

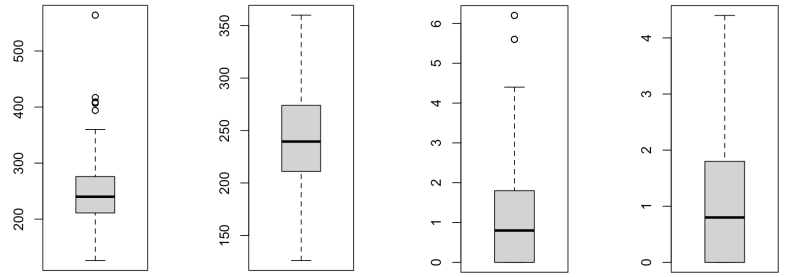
The purpose for our second model is to remove the outliers so as to see if the outliers are true natural outliers. True natural outliers provide significant statistical information and should not be removed. Thus, there is a need to do a comparison of the models.



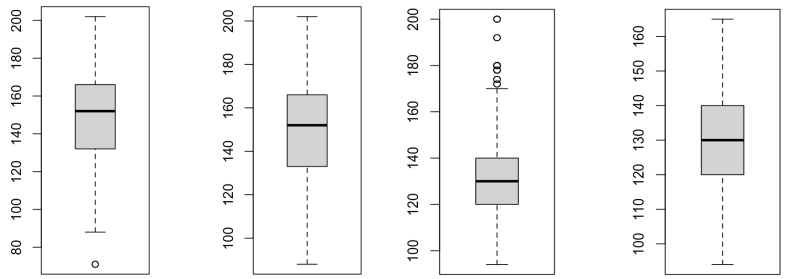
The box plot above shows us a graphical representation of all outliers.

```
> Q1_chol <- quantile(heart.dt$chol, .25)
> Q3_chol <- quantile(heart.dt$chol, .75)
> IQR <- IQR(heart.dt$chol)
> heart.dt2 <- subset(heart.dt, heart.dt$chol > (Q1_chol -1.5*IQR) & heart.dt$chol < (Q3_chol
```

We find the interquartile range, upper limit and lower limit, where data above the upper limit and lower limit of 1.5*IQR is considered to be an outlier.



CHOL OLDPEAK



THALACH TRESTBPS

We removed the outliers from 4 continuous variables, mainly, CHOL, OLDPEAK, THALACH, TRESTBPS. The figure above provides a graphical representation of the before and after comparison for each variable.

5.4.2. Scenario 3: Balancing the data

```
> summary(trainset2$target)
  0   1
324 357
> balancetrain_data = ovun.sample(target~., data=trainset2, method = "over", N = (2*357))$data

> summary(balancetrain_data$target)
  1   0
357 357
```

Firstly, we check for the summary of our Y variable to discover which level indicates the majority cases. In this case 1 is the majority. Afterwhich, we input 357 into ovun.sample to give us the final result where both levels in our categorical variable Y have an equal number (357) of results.

5.4.3. Logistic regression and CART comparison across the 3 scenarios

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              3.011451   1.936547   1.555 0.119931
age                      0.042583   0.017253   2.468 0.013580 *
sexMale                 -2.231947   0.388461  -5.746 9.16e-09 ***
cpAtypical Anginal       1.082315   0.381524   2.837 0.004557 **
cpNon-anginal Anginal    1.966008   0.337932   5.818 5.96e-09 ***
cpAsymptomatic           2.114505   0.446393   4.737 2.17e-06 ***
trestbps                -0.030088   0.007641  -3.938 8.22e-05 ***
chol                    -0.008320   0.002666  -3.120 0.001806 **
thalach                  0.027009   0.007797   3.464 0.000533 ***
exangYes                -0.806522   0.292702  -2.755 0.005861 **
oldpeak                 -0.244232   0.144980  -1.685 0.092068 .
slopeFlat               -0.807493   0.557470  -1.448 0.147479
slopeDownsloping         0.569617   0.594922   0.957 0.338333
ca1                     -2.223993   0.351805  -6.322 2.59e-10 ***
ca2                     -3.888310   0.526989  -7.378 1.60e-13 ***
ca3                     -2.209274   0.566593  -3.899 9.65e-05 ***
ca4                      1.694960   1.016449   1.668 0.095409 .
thalFixed Defect        -0.531739   0.518631  -1.025 0.305234
thalReversible Defect   -1.746288   0.508205  -3.436 0.000590 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 986.48  on 711  degrees of freedom
Residual deviance: 426.40  on 693  degrees of freedom
AIC: 464.4
```

With outliers

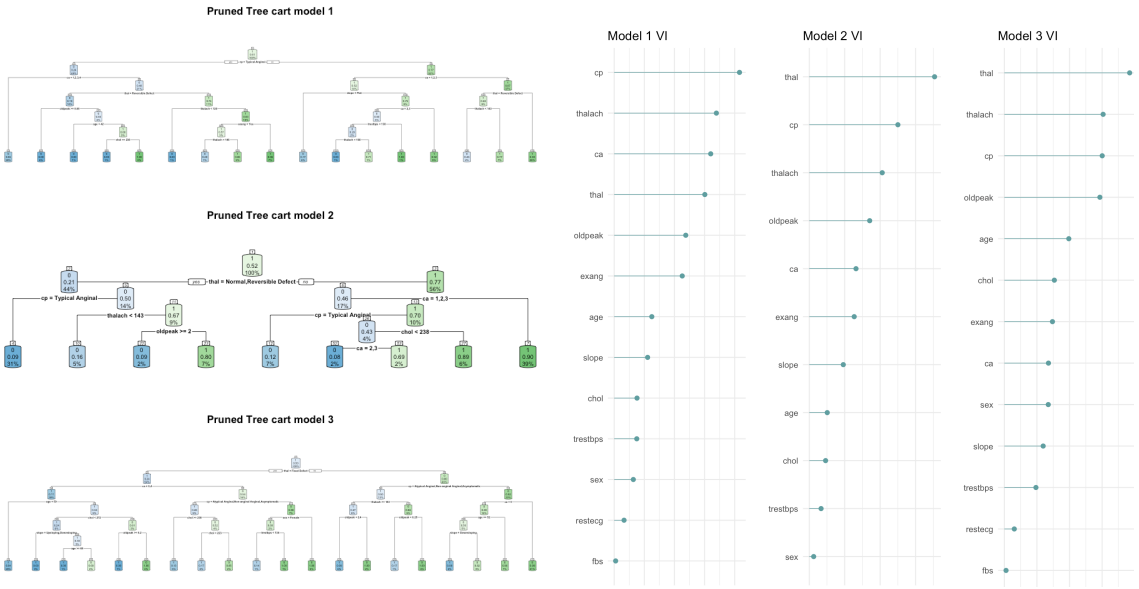
```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              0.530592   1.360183   0.390 0.696471
sexMale                 -1.930131   0.371923  -5.190 2.11e-07 ***
cpAtypical Anginal       1.020343   0.374482   2.725 0.006436 **
cpNon-anginal Anginal    2.200137   0.351677   6.256 3.95e-10 ***
cpAsymptomatic           1.811708   0.457495   3.960 7.49e-05 ***
chol                    -0.006976   0.002818  -2.475 0.013313 *
thalach                  0.023235   0.007305   3.181 0.001470 **
oldpeak                 -0.596357   0.167934  -3.551 0.000384 ***
slopeFlat                0.000274   0.555135   0.000 0.999606
slopeDownsloping         0.998727   0.585133   1.707 0.087852 .
ca1                     -2.376254   0.343967  -6.908 4.90e-12 ***
ca2                     -2.667153   0.485838  -5.490 4.02e-08 ***
ca3                     -3.329037   0.759995  -4.380 1.18e-05 ***
ca4                      1.307975   0.997520   1.311 0.189781
thalFixed Defect        -0.023819   0.533755  -0.045 0.964406
thalReversible Defect   -1.627066   0.523638  -3.107 0.001888 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 942.47  on 680  degrees of freedom
Residual deviance: 396.10  on 665  degrees of freedom
AIC: 428.1

Number of Fisher Scoring iterations: 6
```

Without outliers

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.311979   1.944516  -0.160 0.872534
age                      -0.034719   0.017115  -2.029 0.042499 *
sexMale                   1.702402   0.362005   4.703 2.57e-06 ***
cpAtypical Anginal       -0.753827   0.355753  -2.119 0.034093 *
cpNon-anginal Anginal    -2.209914   0.349981  -6.314 2.71e-10 ***
cpAsymptomatic           -1.873338   0.451967  -4.145 3.40e-05 ***
trestbps                  0.022654   0.009235   2.453 0.014162 *
chol                      0.008068   0.002836   2.844 0.004453 **
restecgST-T Abnormality  -0.663100   0.270441  -2.452 0.014210 *
restecgHypertrophy       -0.194782   2.728546  -0.071 0.943090
thalach                  -0.032183   0.007608  -4.230 2.34e-05 ***
oldpeak                   0.594822   0.164799   3.609 0.000307 ***
slopeFlat                 0.385764   0.624123   0.618 0.536516
slopeDownsloping         -0.517901   0.666164  -0.777 0.436901
ca1                       2.330012   0.343346   6.786 1.15e-11 ***
ca2                       2.409174   0.490556   4.911 9.06e-07 ***
ca3                       2.820450   0.731591   3.855 0.000116 ***
ca4                      -1.741179   1.097066  -1.587 0.112485
thalFixed Defect         -0.022160   0.545752  -0.041 0.967611
thalReversible Defect     1.807817   0.540674   3.344 0.000827 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 989.81  on 713  degrees of freedom
Residual deviance: 412.65  on 694  degrees of freedom
AIC: 452.65
```

Balanced without outliers

The logistic regression comparison of the three models alone is not representative of whether the model is good or bad, but we can make an analysis of the changes observed. From the logistic regression, we observed that the significance in the variables changes. For example, in the first scenario, age is slightly significant in contributing to the prediction of heart disease. After removing outliers, they are no longer considered to be significant in the model. Lastly in the third scenario, age reappears as slightly significant. This tells us that the outliers are statistically significant since it causes a change to the model. However, we are not able to determine whether the outlier is good or bad. We can also conclude from the comparison that the second scenario has the best fit model since its AIC value is the lowest amongst the three scenarios. The second best fit would be the third scenario, followed by the first scenario.



CART tree and variable importance

From the comparison across three scenarios. We can observe that after the removal of outliers, the tree got significantly smaller. The reason may be that removing outliers also removed certain statistical significance from the variable with outliers previously, making it less effective as a classifier. There could also be a case of underfitting for the second tree where there are not enough decisions to make an accurate prediction. When the model changes, we can also observe from the plot that there is a change to the important variables used to predict if a person has heart disease.

```
> accuracy_table
       Model Accuracy In Percentage TYPE 1 ERROR RATE TYPE 2 ERROR RATE
Log1                       87.91000          7.843137          4.248366
Log2                       85.27397          7.191781          7.534247
Log3                       85.95890          7.191781          6.849315
CART 1                     89.21569          7.516340          3.267974
CART 2                     85.61644          6.849315          7.534247
CART 3                     90.06849          5.479452          4.452055
```

5.4.4. Concluding which models is the best for NCHS

In evaluating our model's overall performance and application in detecting heart diseases. The team has decided to select the CART model 1 with the least Type 2 error rate. The prediction accuracy between CART 1 and the next highest accuracy is only a 0.8% difference. However the type 2 error is lower by more than 1% from the next lowest type 2 error rate.

In the case of a Type 1 error (i.e., predicting heart disease but the patient does not have heart disease), there would be no life threatening danger and the hospitals could do refunds for any medical treatment that the patient has received. However, Type 2 errors would be a case where a patient actually has heart disease but the hospital diagnosis shows no heart disease. Heart disease patients require immediate treatment and a delayed treatment could endanger the life of a patient. Furthermore, from a business point of view, it would be harder for the hospital to do any mitigation for Type 2 errors. In the case of Type 1 errors, hospitals would be able to use monetary compensation as mitigation. However, there will be severe backlash from the community of patients should NHCS's heart disease prediction model wrongly classify a patient as healthy when in fact, they have a heart disease. Hence, taking this into account, the team has decided to select a model that has the least Type 2 error.

6. Recommendations And Conclusion

Based on the most important variables highlighted by CART model 1 as seen below ranked in importance:

1) CP: Type of chest pain
2) Thalach: Maximum heart rate achieved
3) CA: Number of major blood vessels colored
4) Oldpeak: ST depression induced by exercise relative to rest
5) Exang: Exercise induced angina

6.1. NHCS To Develop & Utilise Singaporean Datasets

It is recommended that NHCS develops and utilises their own datasets when it comes to predicting heart diseases of Singaporeans. Whilst our model is able to predict heart diseases with a high accuracy, it is limited with regards to the amount of data it contains. For example, the current dataset does not contain any BMI values which is often found to be highly correlated with heart diseases. (Held et al., 2022) A lack of such data could prove to hinder our model's true accuracy in predicting heart diseases. Furthermore, research reveals that race plays a significant role in predicting heart disease. Hence it is imperative that NHCS develops its own datasets for Singaporeans. (Cleveland Clinic, 2022)

6.2. NHCS To Develop Real-Time Heart Disease Predictor To Reduce Wasted Resources

With regards to the team's business problem of reducing capacity strain for NHCS. We recommend that NHCS develop a front end web application that allows users to interact with our predictive algorithms. By doing this, NHCS can reduce the amount of unnecessary visits for unwarranted suspicions of heart diseases. Furthermore, these data can be sent directly to cardiologist and other cardiac care specialists that can tailor medical care according to the patient's specific input variables. For example, a user could use the app and select a specific value for chest pain. Doctors would then be able to narrow down the list of cardiac issues that the patient is currently experiencing, hence, this reduces the time and resources needed per patient which reduces the strain on the NHCS overall.

Appendices

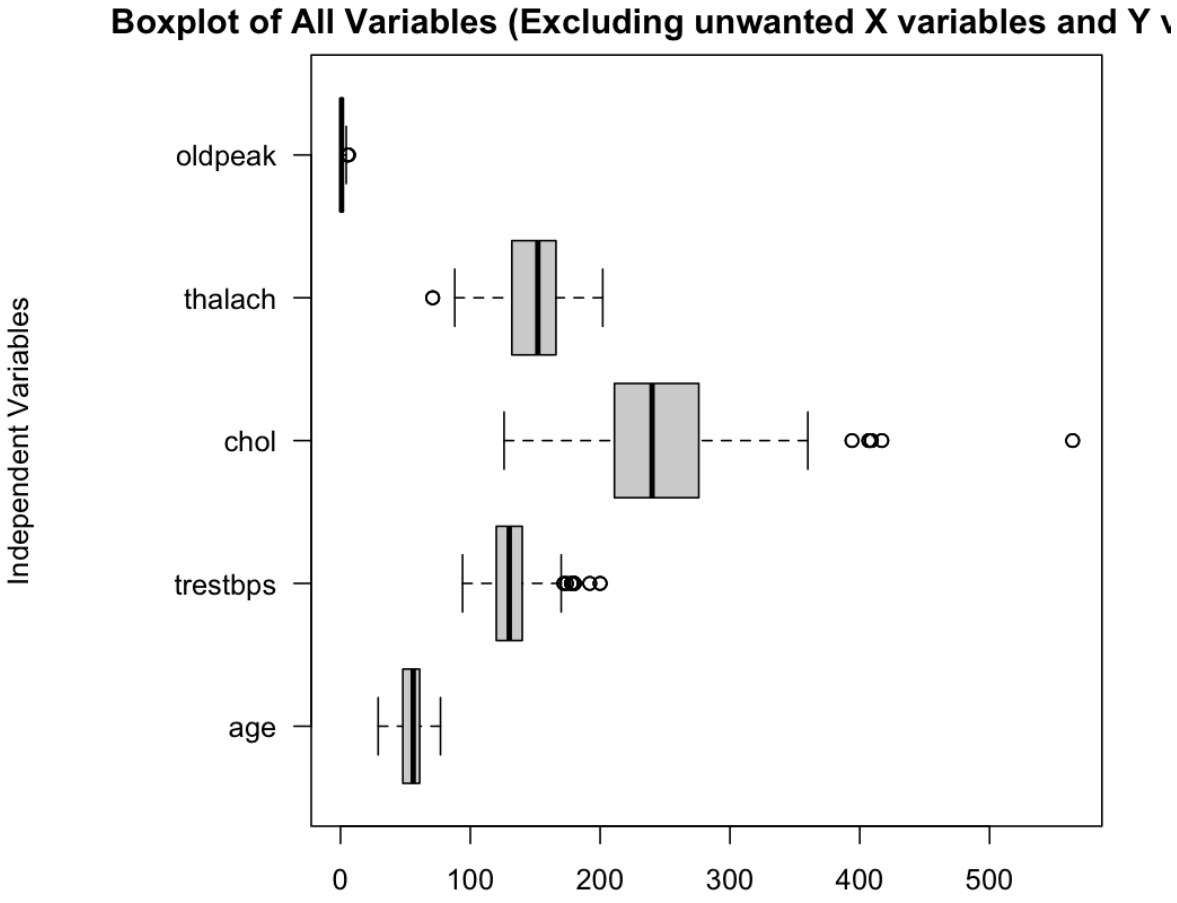
Appendix A: Data Dictionary

Dataset 1. heart.csv		
Name	**Type**	**Description**
age	Continuous	Age in years
sex	Categorical	Sex
cp	Categorical	Type of chest pain
trestbps	Continuous	Resting blood pressure (in mm Hg on admission to the hospital)
chol	Continuous	Serum cholesterol in mg/dl
fbs	Categorical	Fasting blood sugar > 120 mg/dl
restecg	Categorical	Resting electrocardiographic results
thalach	Continuous	Maximum heart rate achieved
exang	Categorical	Exercise induced angina
oldpeak	Continuous	ST depression induced by exercise relative to rest
slope	Categorical	Slope of the peak exercise ST segment
ca	Categorical	Number of major vessels coloured by fluoroscopy
thal	Categorical	An inherited blood disorder called thalassemia
target	Categorical	0 = Heart disease is absent 1 = Heart disease is present

Appendix B: Factorising Data

```
## Setting the appropriate datatype for each variable ----
heart.dt$sex <- as.factor(heart.dt$sex)
heart.dt$cp <- as.factor(heart.dt$cp)
heart.dt$fbs <- as.factor(heart.dt$fbs)
heart.dt$restecg <- as.factor(heart.dt$restecg)
heart.dt$exang <- as.factor(heart.dt$exang)
heart.dt$thal <- as.factor(heart.dt$thal)
heart.dt$ca <- as.factor(heart.dt$ca)
heart.dt$slope <- as.factor(heart.dt$slope)
heart.dt$target <- as.factor(heart.dt$target)
```

Appendix C: Box Plot of Continuous Variables



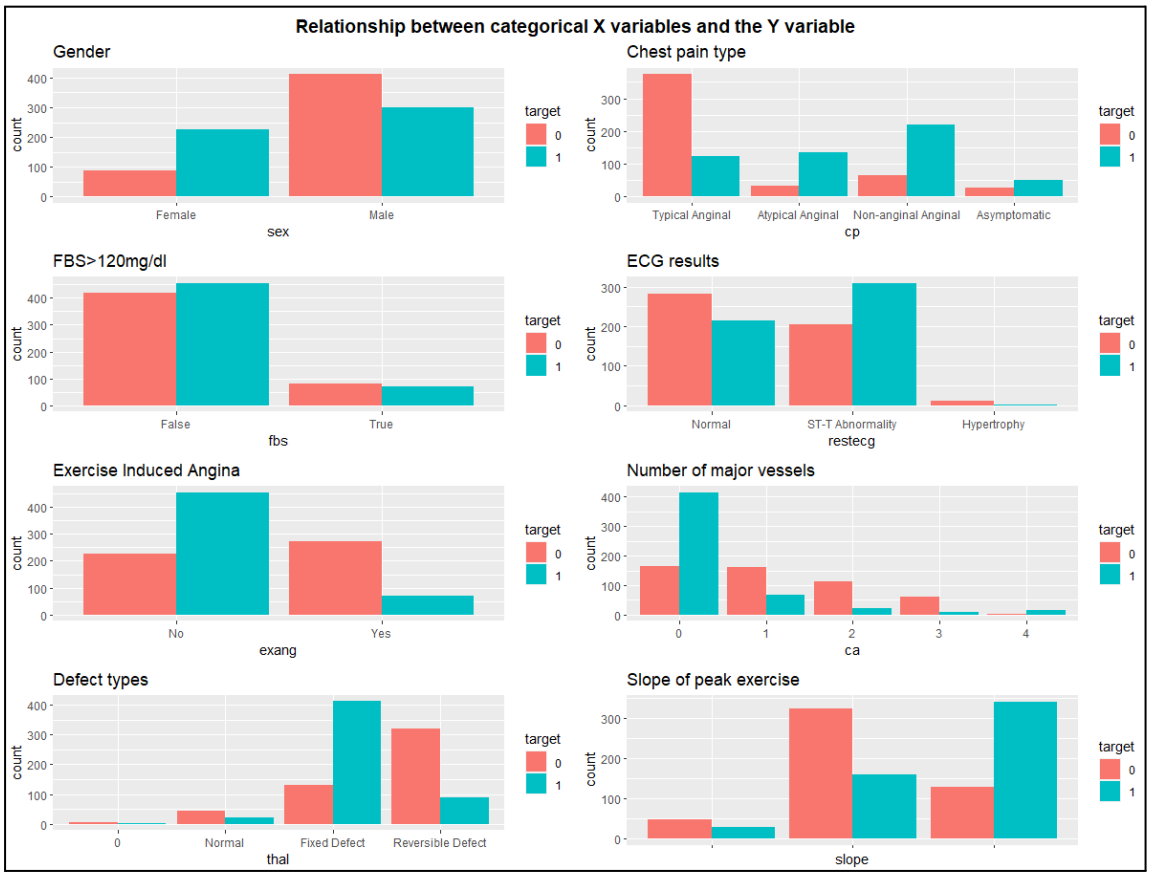
Appendix D: Checking For NA Values

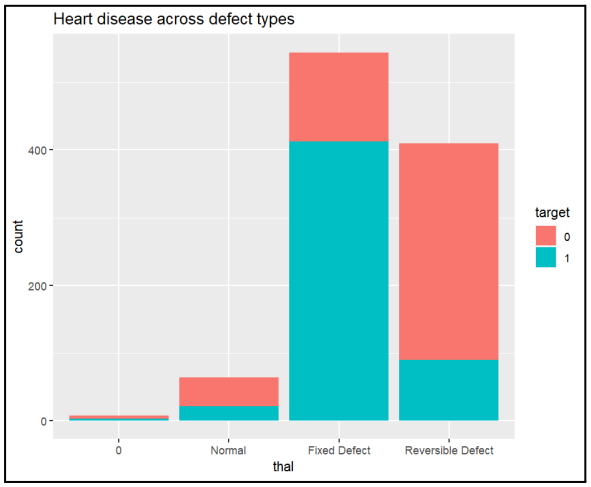
```
> sum(is.na(heart.dt)) # 0 NA values
[1] 0
```

Appendix E: Renaming Columns

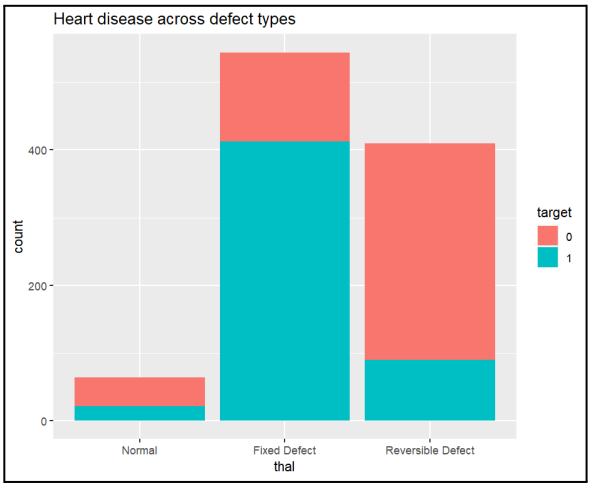
```r
## Formatting the dataset ----
heart.dt[heart.dt$sex == 0,]$sex <- "Female"
heart.dt[heart.dt$sex == 1,]$sex <- "Male"

heart.dt[heart.dt$fbs == 0,]$fbs <- "False"
heart.dt[heart.dt$fbs == 1,]$fbs <- "True"

heart.dt[heart.dt$restecg == 0,]$restecg <- "Normal"
heart.dt[heart.dt$restecg == 1,]$restecg <- "ST-T Abnormality"
heart.dt[heart.dt$restecg == 2,]$restecg <- "Hypertrophy"

heart.dt[heart.dt$cp == 0,]$cp <- "Typical Anginal"
heart.dt[heart.dt$cp == 1,]$cp <- "Atypical Anginal"
heart.dt[heart.dt$cp == 2,]$cp <- "Non-anginal Anginal"
heart.dt[heart.dt$cp == 3,]$cp <- "Asymptomatic"

heart.dt[heart.dt$slope == 0,]$slope <- "Upsloping"
heart.dt[heart.dt$slope == 1,]$slope <- "Flat"
heart.dt[heart.dt$slope == 2,]$slope <- "Downsloping"

heart.dt[heart.dt$exang == 0,]$exang <- "No"
heart.dt[heart.dt$exang == 1,]$exang <- "Yes"

heart.dt[heart.dt$thal == 1,]$thal <- "Normal"
heart.dt[heart.dt$thal == 2,]$thal <- "Fixed Defect"
heart.dt[heart.dt$thal == 3,]$thal <- "Reversible Defect"
```

Appendix F: Relationship Between Categorical X Variables & Categorical Y Variable



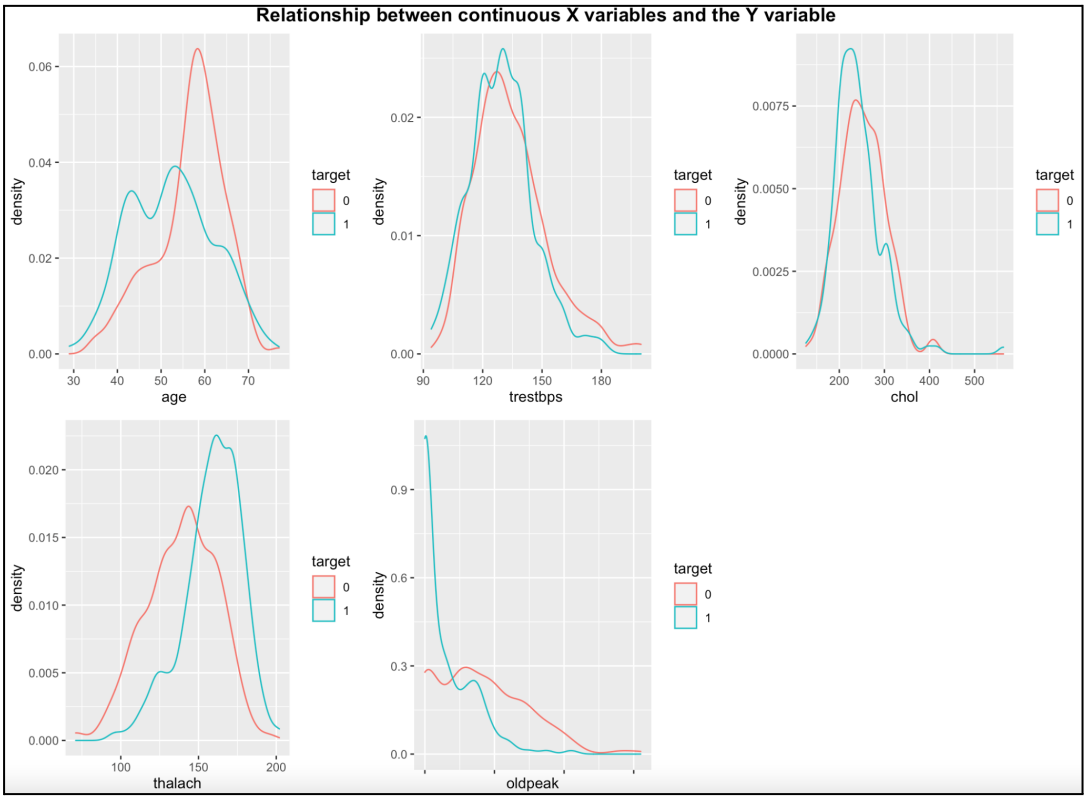


Before removing thal == '0'

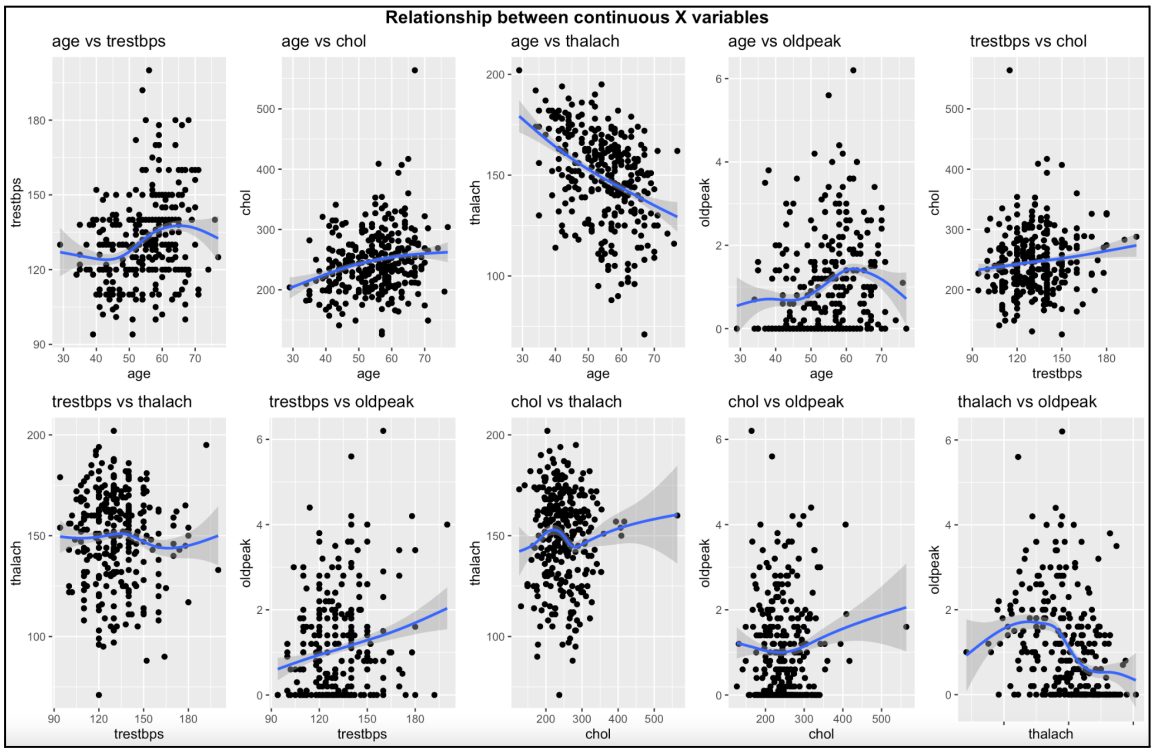


After removing thal == '0'

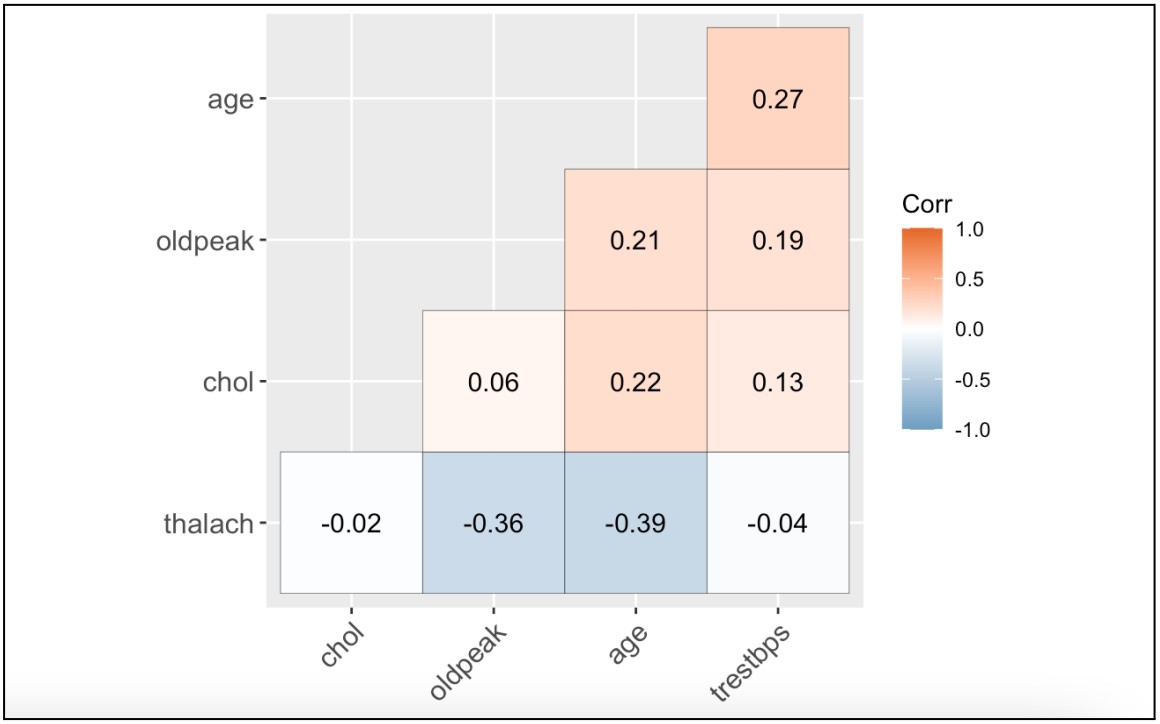
Appendix G: Relationship Between Continuous X Variables and Categorical Y Variable



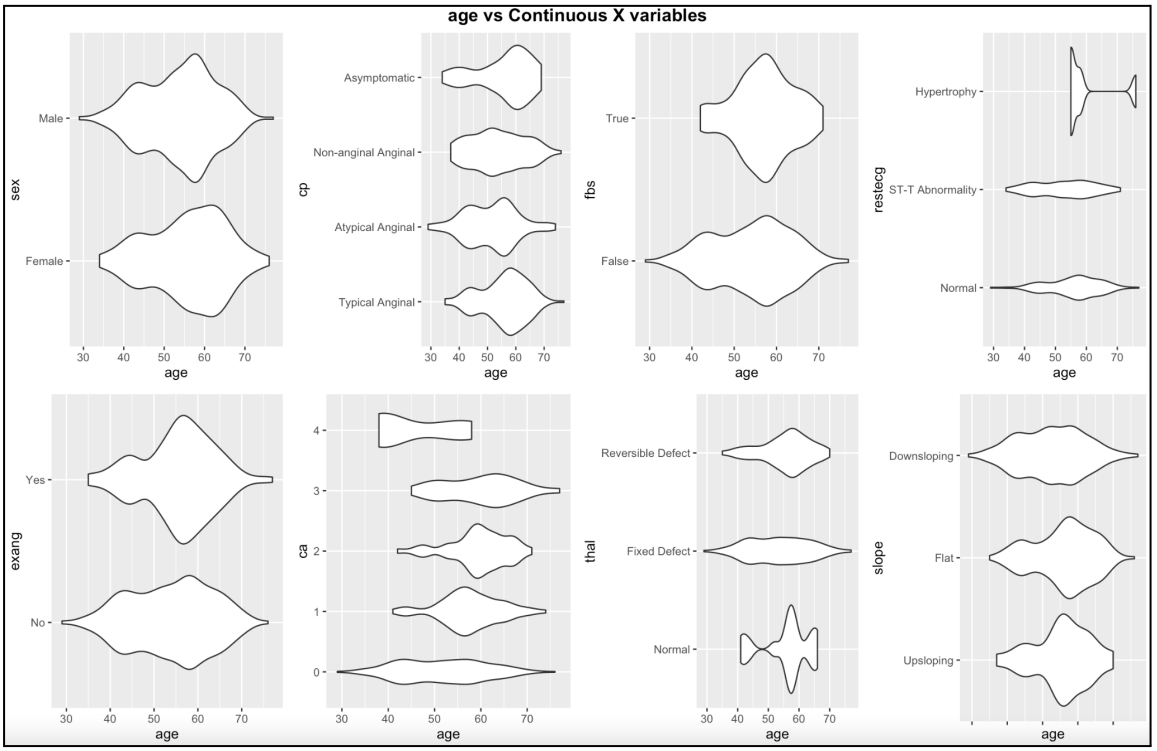
Appendix H: Relationship Between Continuous X Variables

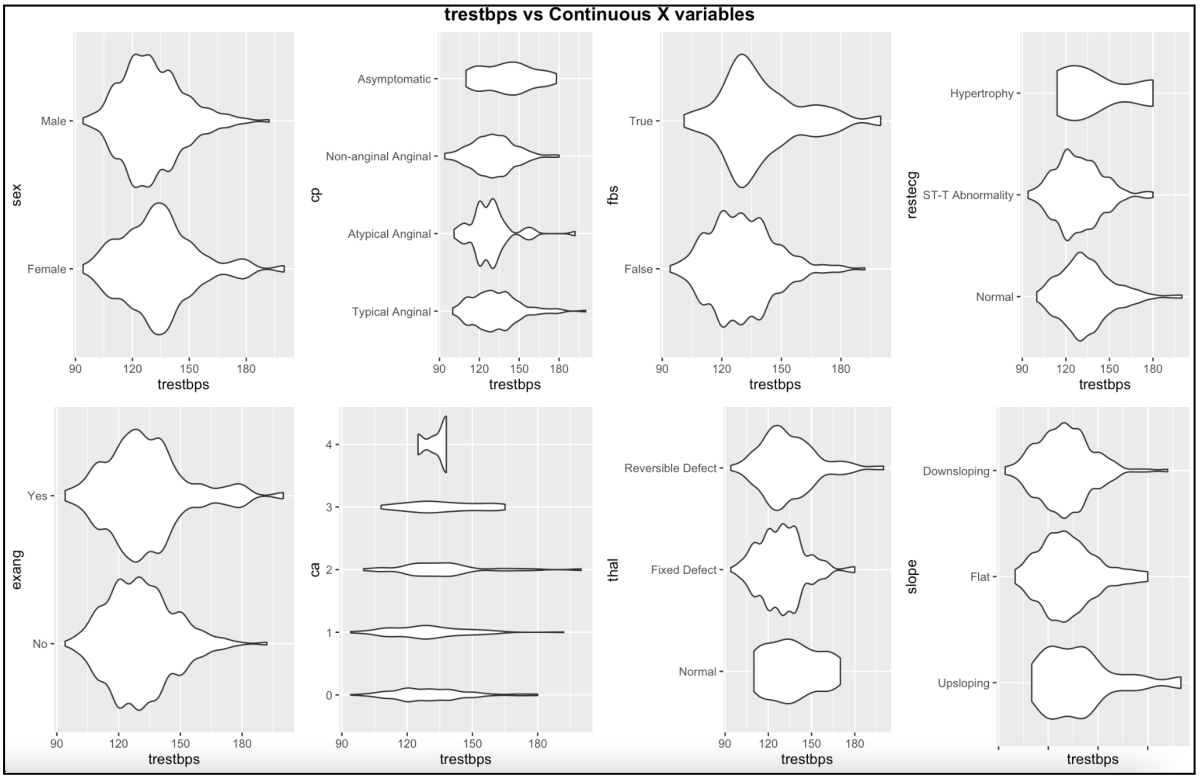


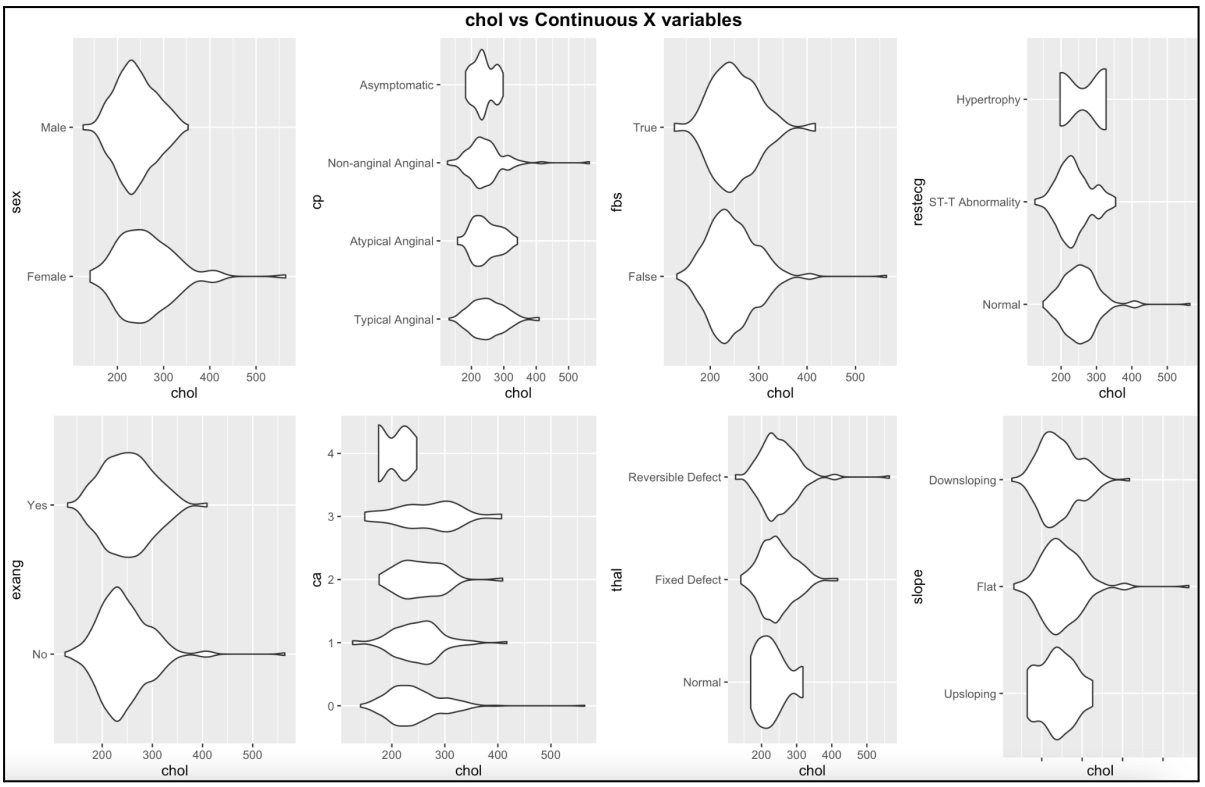
Appendix I: Correlation Graph Between Variables

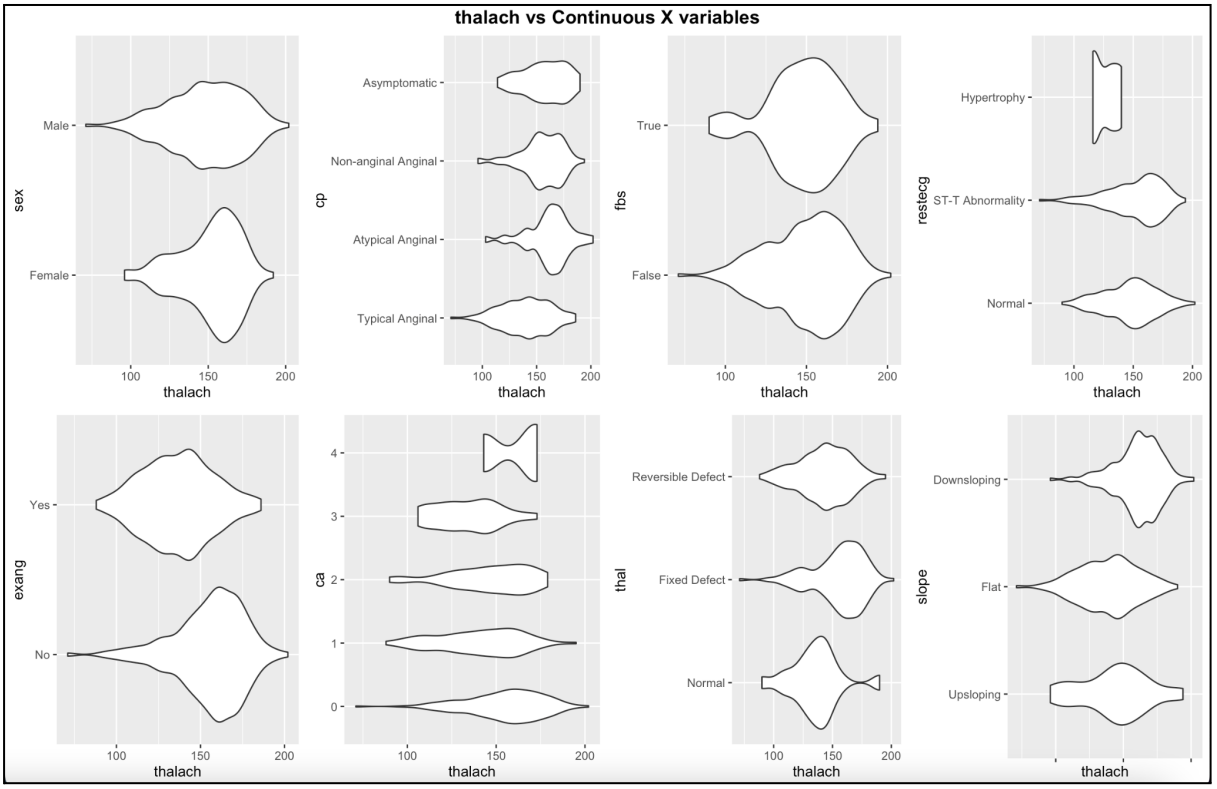


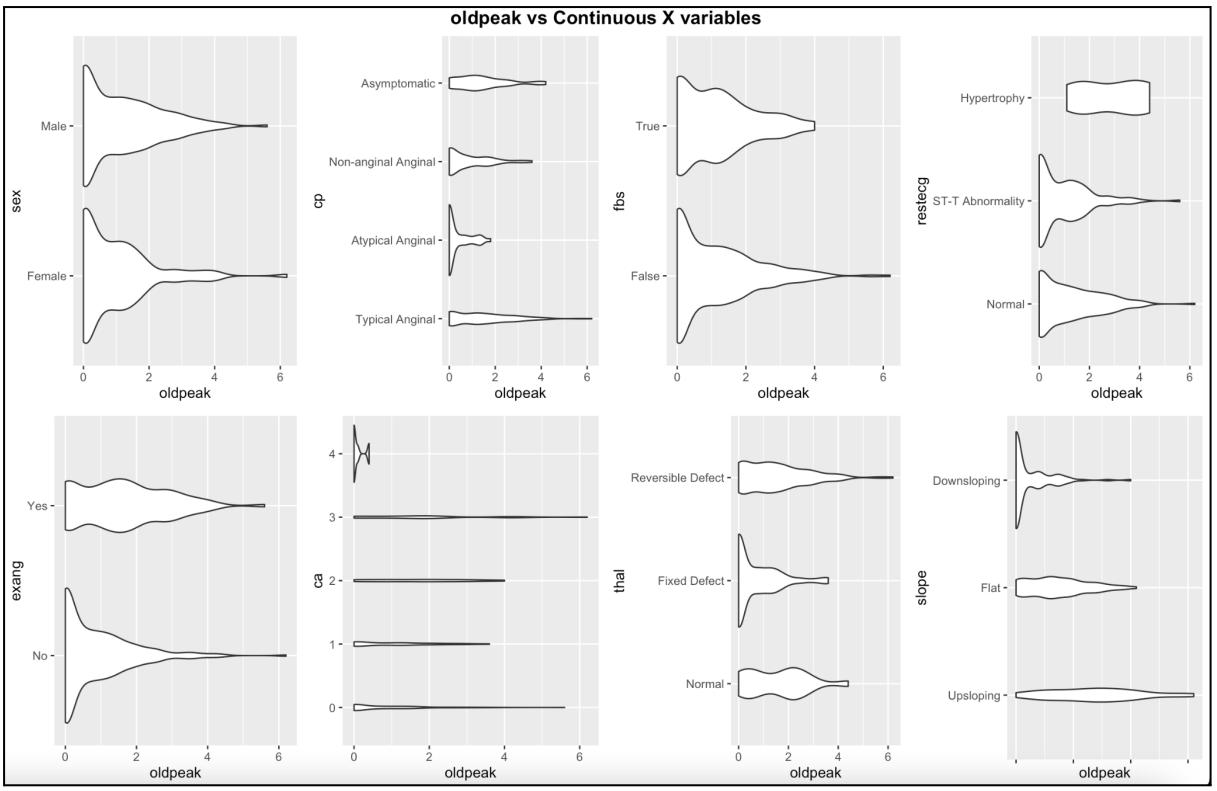
Appendix J: Relationship Between Different Types of X Variables



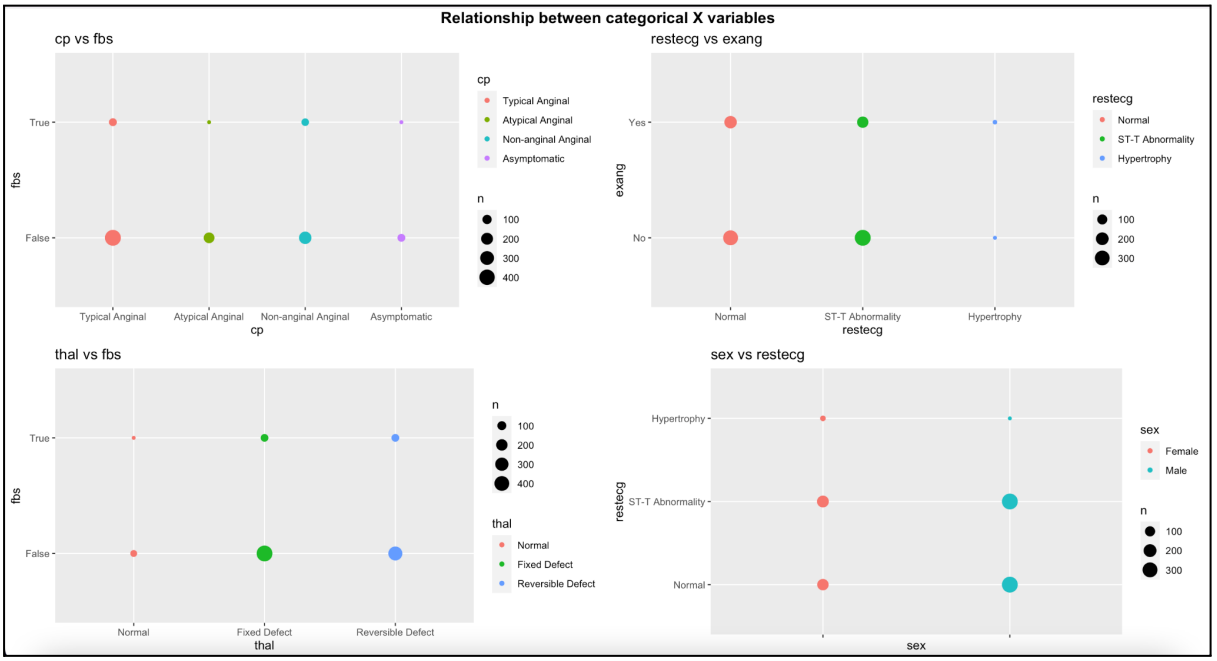



chol vs Continuous X variables
sex
Male
Female
cp
Asymptomatic
Non-anginal Anginal
Atypical Anginal
Typical Anginal
fbs
True
False
restecg
Hypertrophy
ST-T Abnormality
Normal
exang
Yes
No
ca
thal
Reversible Defect
Fixed Defect
Normal
slope
Downsloping
Flat
Upsloping
200 300 400 500
chol


thalach vs Continuous X variables
sex
Male
Female
cp
Asymptomatic
Non-anginal Anginal
Atypical Anginal
Typical Anginal
fbs
True
False
restecg
Hypertrophy
ST-T Abnormality
Normal
exang
Yes
No
ca
thal
Reversible Defect
Fixed Defect
Normal
slope
Downsloping
Flat
Upsloping
100 150 200
thalach


oldpeak vs Continuous X variables
sex
Male
Female
cp
Asymptomatic
Non-anginal Anginal
Atypical Anginal
Typical Anginal
fbs
True
False
restecg
Hypertrophy
ST-T Abnormality
Normal
exang
Yes
No
ca
thal
Reversible Defect
Fixed Defect
Normal
slope
Downsloping
Flat
Upsloping
oldpeak

Appendix K: Relationship Between Different Categorical X Variables



Appendix L: Overall Comparison Of Model Accuracy, Type 1 Error and Type 2 Error

```
> accuracy_table
       Model Accuracy In Percentage TYPE 1 ERROR RATE TYPE 2 ERROR RATE
Log1                       87.91000          7.843137          4.248366
Log2                       85.27397          7.191781          7.534247
Log3                       85.95890          7.191781          6.849315
CART 1                     89.21569          7.516340          3.267974
CART 2                     85.61644          6.849315          7.534247
CART 3                     90.06849          5.479452          4.452055
```

References

American Heart Association News. (2021, August 16). *New blood test is better, faster at diagnosing a heart attack*. Heart. https://www.heart.org/en/news/2018/08/06/new-blood-test-is-better-faster-at-diagnosing-a-heart-attack

Auto, H. (2021, June 10). *Singapore researchers invent new AI tool that could speed up diagnosis of heart disease*. The Straits Times. Retrieved October 29, 2022, from https://www.straitstimes.com/singapore/singapore-team-invents-new-ai-tool-which-could-speed-up-diagnosis-of-heart-disease

Brown JC, Gerhardt TE, Kwon E. (2022, June 5). *Risk Factors For Coronary Artery Disease.* National Library of Medicine. Retrieved October 29, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK554410/

Cedars-Sinai Medical Center. (2022, October 21). *COVID-19 Surges Linked to Spike in Heart Attacks*. Covid-19 surges linked to spike in heart attacks. Retrieved October 28, 2022, from https://www.cedars-sinai.org/newsroom/covid-19-surges-linked-to-spike-in-heart-attacks/

Cleland, J. G. F., Swedberg, K., & Poole-Wilson, P. A. (1998, August 1). *Successes and failures of current treatment of heart failure*. The Lancet. Retrieved October 29, 2022, from https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(98)90015-0/fulltext#secd16321343e780

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. In *Departmental Technical Reports (CS)* (No. 1209). https://scholarworks.utep.edu/cgi/viewcontent.cgi?article=2202&context=cs_techrep

Hajar, R. (2017). Risk factors for coronary artery disease: Historical perspectives. *Heart Views*, *18*(3), 109. https://doi.org/10.4103/heartviews.heartviews_106_17

Heart Disease Statistics. (2022). Singapore Heart Foundation. Retrieved October 29, 2022, from https://www.myheart.org.sg/health/heart-disease-statistics/

Khot, U. N. (2003). Prevalence of Conventional Risk Factors in Patients With Coronary Heart Disease. *JAMA*, *290*(7), 898. https://doi.org/10.1001/jama.290.7.898

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition*. Guilford Publications.

Moons, K. G., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Annals of Internal Medicine*, *170*(1), W1. https://doi.org/10.7326/m18-1377

National Heart Centre Singapore. (2022, April 14). *Overview – National Heart Centre Singapore*. https://www.nhcs.com.sg/about-us/

Sidik, S. M. (2022, February 10). *Heart-disease risk soars after COVID — even with a mild case*. Nature. https://www.nature.com/articles/d41586-022-00403-0?error=cookies_not_supported&code=54b36871-6d02-4b85-8a7e-5e884933ff2f

Singh, S. (2021, July 27). *An emphasis on the minimization of false negatives/false positives in binary classification*. Medium. Retrieved October 23, 2022, from https://medium.com/@Sanskriti.Singh/an-emphasis-on-the-minimization-of-false-negatives-false-positives-in-binary-classification-9c22f3f9f73#:~:text=To%20minimize%20the%20number%20of,optimally%20reaches%20a%20global%20minimum.

Thompson, B. S., & Yancy, C. W. (2004, August 1). *Immediate vs delayed diagnosis of heart failure: Is there a difference in outcomes? results of a harris interactive® patient survey*. Journal of Cardiac Failure. Retrieved October 28, 2022, from https://www.onlinejcf.com/article/S1071-9164(04)00533-0/fulltext

StackOverflow. *Which variance inflation factor should I be using: GVIF or $\text{GVIF}^{1/(2\cdot\text{df})}$?* (2013, September 22). Cross Validated. https://stats.stackexchange.com/questions/70679/which-variance-inflation-factor-should-i-be-using-textgvif-or-textgvif