# 

# **BC2406 – ANALYTICS l: VISUAL AND PREDICTIVE TECHNIQUES**

# **GROUP PROJECT REPORT**

# **A Holistic Approach to Heart Disease Referrals using Machine Learning Paradigms**

# 

# **Seminar Class 3, Team 1 (Tuesday - 6.30PM)**

# *Prepared for: Prof. Josephine Zhou*

# 

|  |  |
| --- | --- |
| **Name** | **Matriculation Number** |
| Javier Tan | U2121055B |
| David Tey Bo Yuan | U2121810J |
| Eric Wu Jiaqing | U2021696D |
| Syed Saad Muntasirul Islam | U2121273K |
| Shermaine Ng Jinglin | U2121773F |

# 

# Table of Contents

[**Executive Summary**](#_btm17f3xnynq)4

[**Introduction**](#_y1bnbs9btowt)5

[Business Opportunity](#_ijue8wu6um86) 5

[Current Solutions](#_v3ek25t6psjj) 5

[Justification for Business Problem](#_y822urnvm26n) 6

[Our Hypothesis](#_uacc3zqxxj6d) 7

[Proposed Solution](#_riywhmqus4gu) 8

[**Data, Methodology & Results**](#_ly2pkw17fdnq)8

[Dataset](#_kklye2rcv4js) 8

[Exploratory Data Analysis](#_qoiy4cds0uav) 9

[Preprocessing](#_maujdyap57v0) 11

[Clustering](#_t78uvahecox) 11

[Logistic Regression](#_s949d61t824p) 11

[Binary Predictions](#_bwgi71hn3rdx) 12

[Multi-Class Predictions](#_m6dqqcmbs8j0) 13

[CART](#_b36k4oj3fcmt) 15

[Binary Predictions](#_e937gml2i859) 15

[Multi-Class Predictions](#_v5zwwhhu8rkn) 16

[**Evaluation**](#_j88th0wqmgex)17

[Evaluation of Our Models](#_ubjgm0il9od6) 17

[Binary Predictions](#_chh0z3qkpndb) 17

[Multi-Class Predictions](#_sxjmcbt4zww) 19

[Comparison Against Current Solutions](#_qdp03zld57ys) 19

[Areas of Improvement](#_gpl2y91tsvec) 21

[Recommended Implementation](#_t04i3gucqhe4) 21

[Referral Process](#_s22ba8am3gfo) 22

[Prototype](#_rjwit9kmxt02) 22

[**Conclusion**](#_fwfspfof3j9f)24

[**References**](#_v2sbsulf93t4)24

[**Appendix A - Data Dictionary**](#_chfggsyhnp3s)27

[**Appendix B - Univariate Distributions**](#_v00wbo9pbcug)28

[**Appendix C - Bivariate Distributions & Correlations**](#_1qaq7nn36fy4)37

[**Appendix D - K-Means Clustering & Hierarchical Clustering**](#_ro6w9p5dbcq2)46

[Preprocessing](#_tsfu951mwep4) 46

[K-Means Clustering](#_j5vmlo99mcbw) 46

[Hierarchical Clustering](#_ldi0lp24s9j8) 49

[**Appendix E - Model Visualisation & Evaluation**](#_elt5joq90wkk)51

[Logistic Regression](#_4jrww4sqttap) 51

[Model with Original Dataset](#_icavmcymtq1h) 51

[Model with K-Means Clusters](#_76bbt67hokwe) 51

[Model with Hierarchical Clusters](#_oc360alfhkie) 52

[Binary Model](#_cfvrhokazb0w) 52

[Binary Prediction](#_o1lxcuxcyyhn) 53

[Multi-Class Prediction](#_j4gxq42rjnso) 54

[CART](#_mqolgj81ha9m) 55

[Model with Original Dataset](#_a20e0wbte5d4) 55

[Model with K-Means Clusters](#_d0uyo4t0rwif) 56

[Model with Hierarchical Clusters](#_futa628bxs0h) 57

[Binary Model](#_w8lu5edzg8jj) 58

[Binary Prediction](#_npvsxim8ybw6) 59

[Multi-Class Prediction](#_udafr2dhp264) 59

[**Appendix F - Estimation on reduction of number of outpatients**](#_s9g0z7j40khp)61

# 

# Executive Summary

As the leading government heart disease healthcare agency in Singapore, NHCS is where patients go for their cardiovascular health needs. However, this had led to GPs around the country over-referring their patients to NHCS, causing staff to be severely overworked and on the brink of burn-out. This has also caused unreasonably long waiting times for patients, as well as cost NHCS a significant amount of resources.

With the expected number of elderly in Singapore to double by 2030, where one in four Singaporeans will be over the age of 65 (US Trade, 2022). The future demand for cardiovascular healthcare will surge as cardiovascular disease remains the most common cause of death of older adults (Jaul, E., Barron, J., 2017). NHCS will be responsible for handling this wave of increasing demand. Simply relying on existing methods to gauge whether a patient requires specialist care is insufficient, unsustainable, and a waste of resources; the current solutions to assist patient referral for GPs such as the Framingham Risk Score, are not very effective as they predict based on 10-year risk, instead of immediate risk.

To better equip NHCS for the increased patient needs and further digitalising itself to align with Singapore government’s smart nation initiative, we propose implementing new Machine Learning Models in complement with the currently used Framingham Risk Score to improve the referral accuracy of patients from General Practitioners (GPs) to NHCS. This new proposed approach reduces the number of false positive patients referred to NHCS by predicting immediate risk, and enables a new system of referral priority whereby more urgent patients can gain access to doctors in time.

Using the UCI heart disease data set with over 800 detailed patient records, our machine learning models, Logistic Regression and Classification And Regression Trees (CART), achieved an initial prediction accuracy rate of 86% on referral accuracy and 57% on priority referral accuracy.

NHCS could capitalise on its closely connected network of GPs in Singapore and include the use of our solution as part of their standard operating procedures. This new approach can be implemented onto a website to allow easy access to the GPs. A simple user interface with our analytics model integrated into the backend makes the use of our solution easy and scalable.

Should our solution be adopted, its accuracy can be increased even further with more patient data available for prediction, catering to the Singapore population. Allowing for improved efficiency, lightening of workload for overworked doctors and operational cost savings. Our initial estimates suggest a 64% reduction in unnecessary outpatient consultations, amounting to almost SGD 3.25 million in savings.

# Introduction

## Business Opportunity

The emergence of interest around machine learning these past few years has opened up a whole new range of possible applications of such technology into older industries, including the healthcare industry. The healthcare industry is forever evolving and researchers are constantly on the lookout for new technologies to solve decade old problems, however, there are countless issues that need solving, and research that needs funding. Unfortunately this has meant that in many cases, healthcare professionals have yet to tap into such modern technology and have to rely on old and possibly outdated solutions.

Although the 21st century is often called the “Data Age” for its abundance of data, healthcare professionals may not have been able to fully utilise this treasure trove of valuable information due to the time it takes to manually analyse and interpret trends in the data. This is where machine learning comes in; the ability of machine learning to analyse vast amounts of data, discover hidden trends, and make predictive models is unmatched by any human researcher or analyst. Our physical and dietary behaviour has also changed over time with the rise of processed foods and remote work, meaning that population health trends are also likely to evolve over time. These attributes combined allow Machine learning to maintain predictive accuracy much better than traditional trend-based analysis. As heart disease cases have been on the rise (MOH, 2021) there is a need to use analytical methods to find more efficient solutions.

This brings us to National Heart Centre Singapore (NHCS). As one of the leading centres for cardiovascular care in the region, there is immense demand for their services. They handle thousands of cases a year, with the vast majority of them being outpatient consultations.

Although it is possible to book appointments with one of their specialists through NHCS website, the average person would not have enough medical knowledge to know which type of specialist to visit, thus, their main channel for appointments is through referrals by General Practitioners (GP).

## Current Solutions

To assist GPs in knowing when to refer a patient to NHCS for heart disease, NHCS has made available a simple excel calculator to allow GPs to calculate their patient’s Framingham Risk Score. The Framingham Risk score provides GPs with a patient’s estimated 10-year risk of developing heart disease (Fig 1).

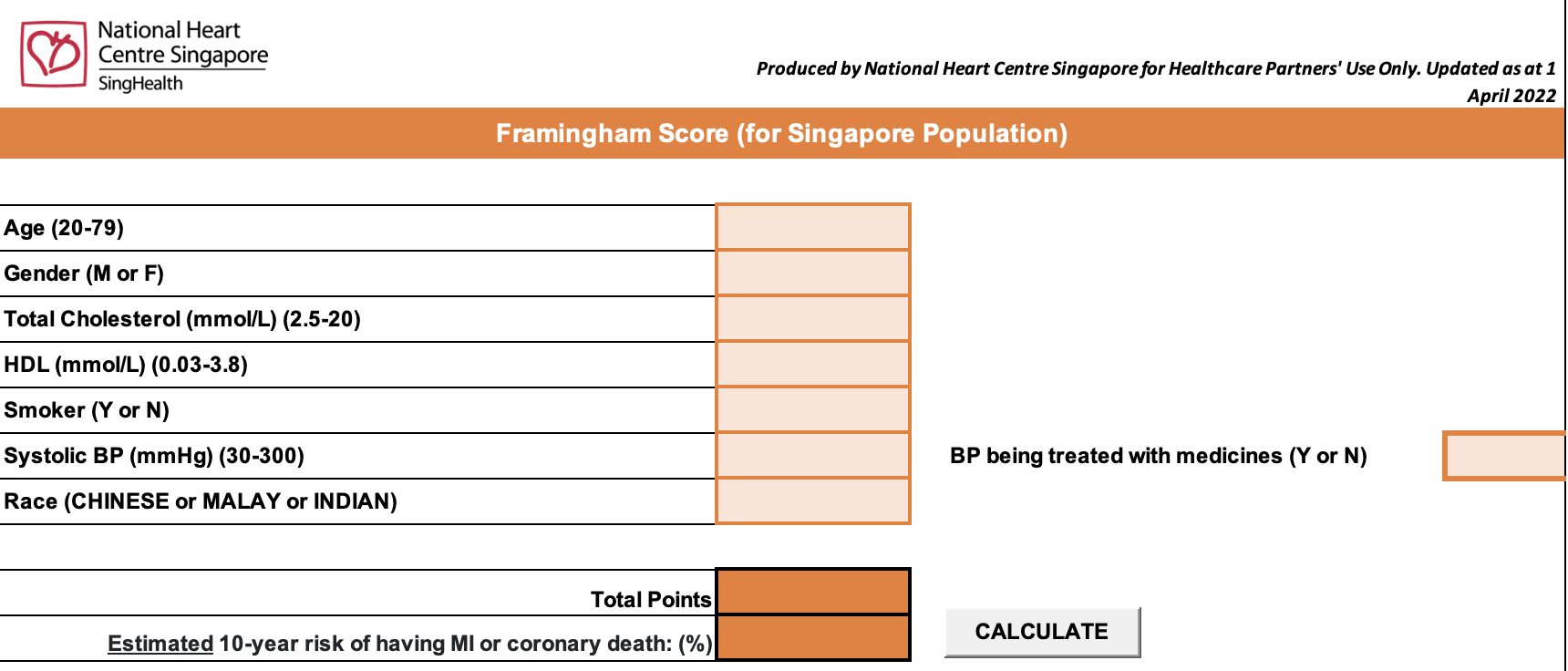


Fig 1.Framingham Risk Score Calculator from NHCS

The simplified current flow for referrals to NHCS by GPs is shown below (Fig 2):

Patient visits GP

Framingham Risk Score is calculated

Patient is referred to NHCS

If Score exceeds threshold

Fig 2.Current NHCS referral flow

Patient’s then either go for various tests or visit the specialist directly.

Moreover, modern predictive methods currently in the industry such as Euro-SCORE and QRISK® that were developed to assist healthcare workers in diagnosing patients have the same issue as the Framingham Risk Score as they only predict the 10-year risk. Hence, the issue is that there is no current solution to accurately gauge current risk and whether to refer a patient to a specialist.

## Justification for Business Problem

NHCS relies on the Framingham Risk Score, however, it is greatly limited as the calculations are based only on data obtained in the Framingham Heart Study from decades ago, and it was only last updated in 2008. Many have also claimed that the scoring system is arbitrary, and it is also difficult to objectively evaluate its accuracy due to several biases being introduced over the timespan of 10 years. Similarly, research results reveal a lack of predictive accuracy in risk assessment tools and hence, raise an urgent need to relook and fine-tune current risk assessment techniques (Blaha, 2015).

Studies have shown that such calculators can greatly overestimate the risk of a person developing heart disease (Brindle, 2003), and NHCS also found that there is an excessive amount of unnecessary referrals, where patients did not actually need to visit a specialist (Lai, 2021). This implies that the Framingham Risk Score causes unnecessary referrals to NHCS which will put unnecessary strain on the healthcare workers and result in excessive waiting times for appointments for heart specialists. The current waiting time to see a heart specialist under subsidised care at the National Heart Centre (NHC) can take up to 4 - 6 weeks, and they do not accept walk-ins (Tan, 2020).

By relying on the Framingham Risk Score, NHCS is casting a wide net that would allow them to detect all patients at risk of developing heart disease over the span of 10 years. However, it does not measure the immediate risk of a patient developing heart disease, and does not indicate whether a patient requires immediate specialist attention from NHCS. Therefore, patients with relatively high Framingham Risk Scores may not benefit from being referred to NHCS as their immediate risk level is low and they only require preventive care to reduce their future risk of developing heart disease.

Singapore’s healthcare workers are also severely overworked (CNA, 2022) and NHCS reportedly handles over 120,000 outpatient consultations yearly - several times more than their surgical and interventional patients, and inpatients combined. This shows that a significant portion of their resources are dedicated to caring for their outpatients that were referred by their GPs to NHCS.

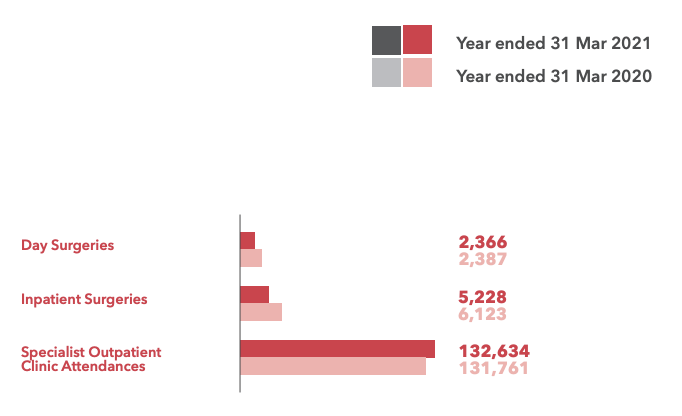


Fig 3.NHCS Annual report 2020-2021

It is clear that there is great stress being placed on NHCS and its staff, and that a large portion of their limited resources are being used up in outpatient services. This is a source of significant inefficiency as more time spent on non-urgent cases means that other patients with more critical needs may not get the care and attention they need, and that there are fewer resources left for such patients.

The inefficient allocation of resources stemming from the inaccuracy and outdated Framingham Risk Score poses an urgent need for us to find a solution to this business problem.

## Our Hypothesis

We can observe from the state of the industry and NHCS that the over-referral rate is high and the current tools are not looking at the correct metric to predict on. This presents an opportunity for us to step in to come up with a solution to reduce the over-referral rate by looking at underlying factors which contribute to this problem.

With an abundance of data in the world of today, we are able to capitalise on machine learning to aid us in the development of our solution, which is to come up with a model that is able to predict the immediate risk level of developing heart disease on top of the 10-year risk. We hypothesise that our solution will be able to assess patients and determine whether to refer a patient more accurately by not only predicting on immediate risk, but also by incorporating different machine learning models and looking at a wider scope of predictors.

Therefore, machine learning will aid us in coming up with our proposed solution. Through its ability to categorise data provided, and observe trends and patterns, we can gain insights which will be helpful in solving our business problem.

## Proposed Solution

Our solution aims to tackle this inefficiency by enabling GPs to more accurately detect a person’s immediate risk level of developing heart disease, on top of their 10-year risk. By looking at their immediate risk, GPs would be able to make more informed and accurate decisions to refer the patients to heart specialists. This would also allow them to provide better preventive care to the other patients who do not need urgent help from specialists. Furthermore, GPs would be able to provide more timely and accurate referrals to NHCS which would not only lessen the burden on NHCS staffs and reduce wastage of resources, but also reduce the long appointment times to see heart specialists and allow those who urgently need the care to receive it as promptly as possible.

We would do so by introducing a more sophisticated and accurate risk measurement model using machine learning that is able to predict whether a patient is at immediate risk of developing heart disease, and if so, the stage of heart disease they are at risk of developing. This would complement the Framingham Risk Score that only predicts a person’s long-term risk of heart disease, and provide a more holistic view on a person’s risk of heart disease, allowing GPs to make more informed decisions on whether a patient requires specialist attention from NHCS or just preventive care.

The second part of our solution consists of a priority referral system. Knowing a patient’s immediate risk level would also allow GPs to indicate priority levels to their referrals. Patients that are predicted to have more serious stages of heart disease would be given higher priority when being referred to NHCS such that they get appointments sooner and receive the treatment they need as soon as possible. This would ensure a fair and efficient allocation of NHCS’ resources such that patients with the most urgent needs are treated first, while low-risk patients can be attended to once other patients have been treated as they do not require the treatment as urgently.

# Data, Methodology & Results

## Dataset

The data set we used is the UCI heart disease dataset, taken from the UCI Data Repository. The dataset consists of 16 variables including the ID and target variable “num”. The remaining 14 are composed of various variables related to cardiovascular health such as the chest pain and resting blood pressure. The link to the dataset can be found in the [References](#_v2sbsulf93t4). The full data dictionary can be found in [Appendix A](#_v2sbsulf93t4).

The following variables were chosen to be used as the input to our models:

* age
* sex
* cp (Chest Pain)
* trestbps (resting blood pressure)
* fbs (fasting blood sugar)
* restecg (resting ecg results)
* thalch (max heart rate)
* exang (exercised induced angina)
* oldpeak (ST depression induced by exercise relative to rest)
* slope (the slope of the peak exercise ST segment)
* thal (Thalassemia)
* ca (number of major vessels (0-3) colored by fluoroscopy)

## Exploratory Data Analysis

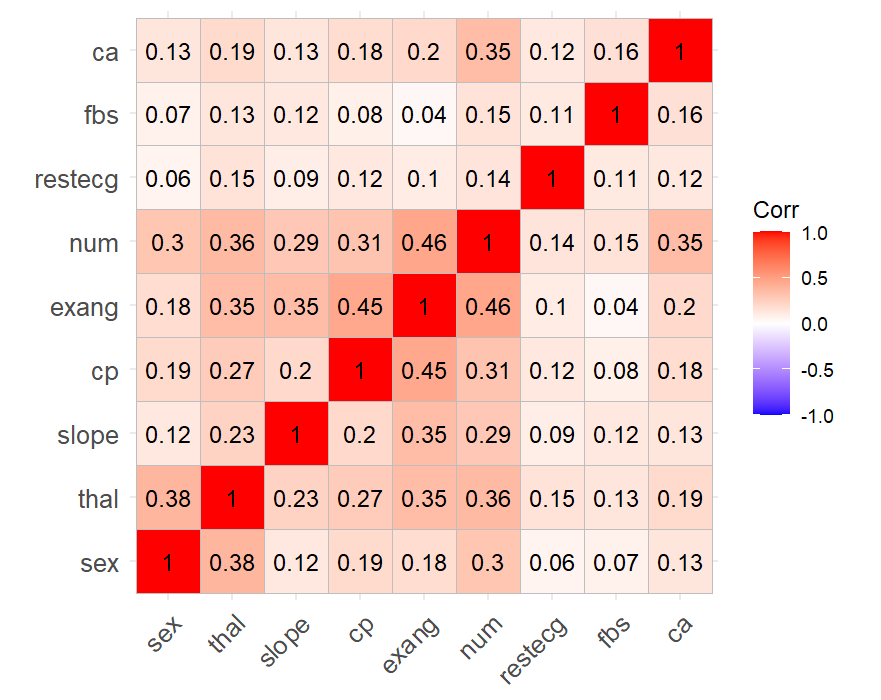


Fig 4.Correlation matrix of categorical variables

For the categorical variables in our dataset, we measured the Cramer’s V value of each pair of categorical variables, which measures the association between the two variables and gives an output from 0 to 1 (Fig 4). A higher value means that the two variables are more highly associated or correlated, thus suggesting that the independent variable has more predictive power.

From the correlation matrix, we can see that the dependent variable “num” has a moderate association with 6 of the independent variables (> 0.3).

There is also some multicollinearity between the independent variables:

* “exang” and “cp”
* “thal” and “sex”
* “exang” and “slope”
* “exang” and “thal”

Despite this, we decided not to drop any of the independent variables due to multicollinearity as the associations are still moderate in magnitude.

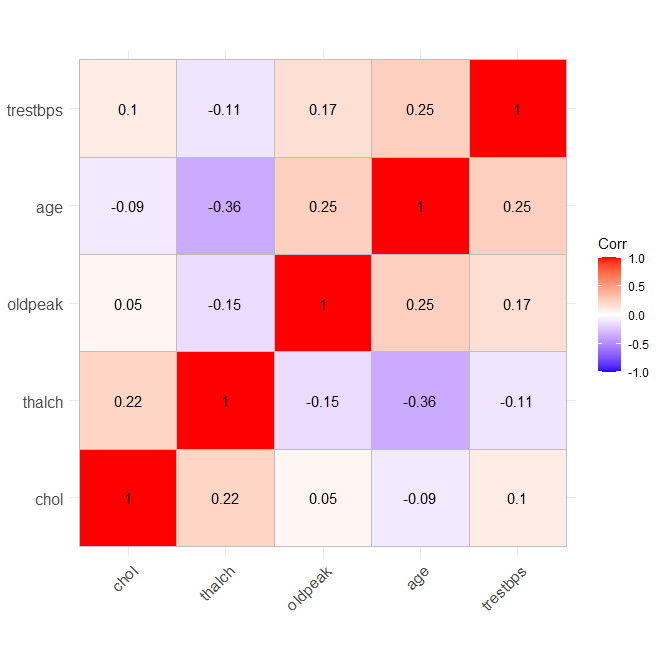


Fig 5.Correlation matrix of continuous variables

For the continuous variables, the correlation matrix (Fig 5) shows little multicollinearity as all of them have an absolute value of less than 0.3 except for “age” and “thalch”.

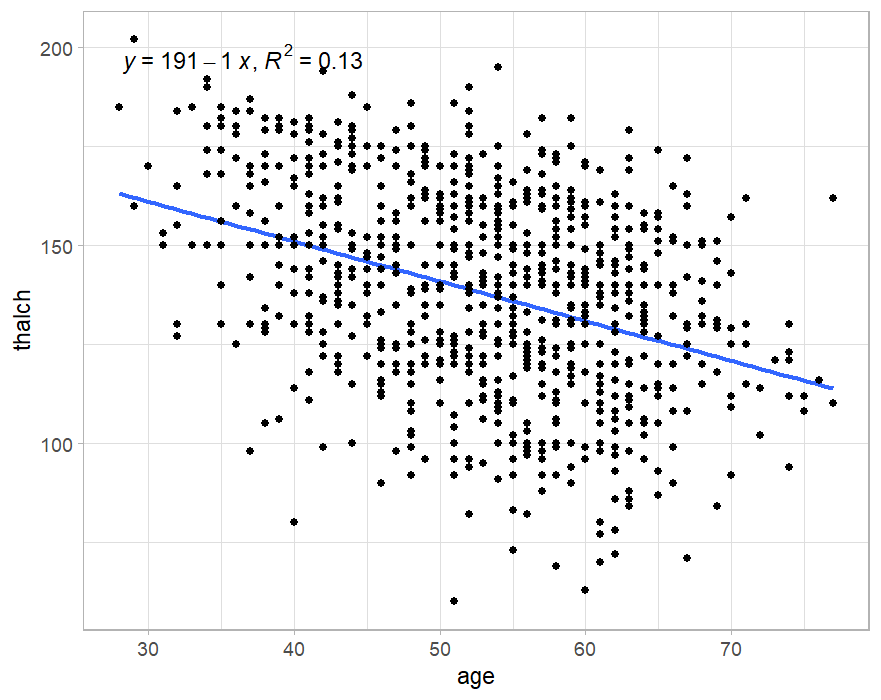


Fig 6.Thalch v Age

By looking more into “thalch” and “age” (Fig 6), we found that while there is a negative correlation between the two variables, the R-squared remains quite low at 0.13. Therefore, we would not be removing any continuous variables.

Further data exploration was done on the relationship between num and the continuous variables and the relevant graphs can be found under [Appendix C](#_1qaq7nn36fy4).

## Preprocessing

Before moving on to the models, we had to first preprocess our data to take care of missing values. Since CART models are able to handle missing data through the use of surrogates, we did not need to remove the missing data when training our CART models. As such, we split the dataset into two, in which we kept all the rows in the CART dataset, and removed rows with missing values for the logistic regression dataset.

Removing the missing values did not prove to be too big an issue as our dataset was relatively large, thus ensuring that we had enough data to train our models. The missing data also coincided within different columns, allowing us to eliminate most of the missing values by dropping the same rows.

Most importantly, as our model is intended to be used by GPs without access to specialised testing equipment, we decided to keep only those input variables that could easily be collected by GPs. Therefore, we decided to drop the column “ca” which represents the number of major vessels which requires fluoroscopy to determine.

In order to prepare our dataset for clustering analysis, we first did one-hot encoding for the original categorical variables. This includes the missing values, which are treated as 0 for the entries that are missing. We also normalised our data to scale the numerical variables between 0 and 1. As distances between points are calculated using L2 norm in clustering analysis, we had to ensure that the different columns were similar in magnitude to ensure equal weightage between the various columns. A more technical explanation can be found in [Appendix D](#_ro6w9p5dbcq2).

## Clustering

Clustering is an unsupervised machine learning model that can help identify groups or clusters within our dataset by looking at the similarities and differences between patients. By labelling each patient with a learned cluster, we may be able to more accurately predict the patient’s stage of heart disease.

For our clustering analysis, we used both K-means clustering as well as hierarchical clustering. It was found that the optimal number of clusters was 10 for K-means, and 8 for hierarchical. As mentioned previously, a technical explanation can be found in [Appendix D](#_ro6w9p5dbcq2).

## Logistic Regression

One of the chosen predictive models is logistic regression as it allows us to predict outcomes with more than 2 classes, such as “num”. The process was as follows:

* Train-Test split - The data must be split into 2 subsets known as the train and test sets in the ratio 80:20. The model will only be “trained” on the train set and tested using the test set. This is to ensure that unbiased model evaluation can be conducted as the model would have never encountered the test set, ensuring that the tests are more accurate and representative of its true accuracy. Stratified random sampling was used to split the data into the respective sets. This method ensures that the proportion of occurrences of each level of “num” was roughly equal in both sets.
* We also fit 3 separate models, 1 using the original dataset, 1 using the dataset along with the K-means cluster data, and 1 using the dataset along with the hierarchical cluster data. This is to see if the clusters allow the model to more accurately classify the data.
* We also fit an additional model on a purely binary target variable, to see if the model is able to learn better and classify between low and high risk of heart disease more accurately, the P-values of the most statistically significant variables can be found in [Appendix E](#_elt5joq90wkk).

Since “num” has 5 levels from 0-4, in order to test the binary accuracy between low immediate risk vs high immediate risk, all the levels from 1-4 were changed to 1. By doing so, we have 2 outcome results that affect a GPs decision. First, the binary outcome will indicate whether a patient has high immediate risk and if they need to be referred to NHCS. If they are at high risk, the GP can then do the multi-class prediction to determine which stage of heart disease they are at risk of and their priority level when referring. Therefore, 2 accuracy measures are required to assess the 2 stages of predictions.

Once the models have been fit and used to predict values on the test set, we can generate a confusion matrix of the value predicted by the model against the true reference value in the test set. This also gives us the predictive accuracy of the model.

### 

### 

### Binary Predictions

Below are the results of each of the models in binary predictions.

0 - Low immediate risk

1 - High immediate risk

|  |  |
| --- | --- |
| Default model with original dataset | Model including K-means clusters |
| Model including hierarchical clusters | Model trained on binary dependent variable |

From the above table, we see that all the models had an accuracy of at least 79.5%, with the highest being the default model with the original dataset at 82.2% binary accuracy.

### Multi-Class Predictions

Below are the results of each model in multi-class predictions.

|  |  |
| --- | --- |
|  | Default model with original dataset |
|  | Model including K-means clusters |
|  | Model including hierarchical clusters |

The above table shows that each of the models have multi-class prediction accuracy of at least 53.4%, with the model including the hierarchical clusters having the highest accuracy at 57.5%.

## CART

The second chosen model is CART - Classification and Regression Trees. CART is ideal for this situation as not only can it handle missing values - allowing GPs to decide if certain variables are even necessary, they are also easy to understand and interpret. This would allow GPs to see the reasoning behind the predictions, they can then use their own professional judgement to further evaluate the patient according to the outcome of the model. The process was as follows:

* Train-Test split - the dataset was split in the ratio 80:20 using stratified sampling.
* 3 models were trained for the 3 different variations of data.
* An additional model was also trained on a purely binary dataset to see if the model is able to classify between low and high risk of heart disease more accurately
* The decision tree was then grown to its max by using all of the variables
* The tree then had to be pruned to find the simplest tree that was able to classify accurately.

Similar to logistic regression, the model can also be used for both multi-class predictions as well as binary predictions.

### Binary Predictions

Below are the results of each of the models in binary predictions.

0 - Low immediate risk

1 - High immediate risk

|  |  |
| --- | --- |
| Default model with original dataset | Model including K-means clusters |
| Model including hierarchical clusters | Model trained on binary dependent variable |

From the above table, we see that all the models had an accuracy of at least 74.6%, with the highest being the purely binary model with an accuracy of 86.8%.

### Multi-Class Predictions

|  |  |
| --- | --- |
|  | Default model with original dataset |
|  | Model including K-means clusters |
|  | Model including hierarchical clusters |

The above table shows that each of the models have multi-class prediction accuracy of at least 52.73%, with the default model having the highest accuracy at 57.0%.

# Evaluation

## Evaluation of Our Models

We shall first take a look at the models that we have produced and compare them to determine which are the better models.

### Binary Predictions

|  |  |
| --- | --- |
| CART Model trained on binary dependent variable | Default Logistic Regression model with original dataset |

|  |  |  |
| --- | --- | --- |
| Model | False Negative Rate | False Positive Rate |
| CART |  |  |
| Logistic Regression |  |  |

We first take a look at the binary predictive accuracies of both CART and Logistic Regression, we can see that the best CART model is able to produce a higher predictive accuracy of 86.8% as compared to that of the best Logistic Regression model with 82.2%.

We can also see that the CART model has the lowest false negative rate of 10.4%, while the Logistic Regression model has the lowest false positive rate of 2.86%.

By comparing the metrics above, we can see that CART is the better model in terms of predictive accuracy. In particular, the false-positive rate is an important metric to draw our attention to. In our context, the false-positive rate represents the proportion of incorrect referrals when the patients do not actually need to consult a heart specialist. Hence, a lower false-positive error rate would aid us in our goal to achieve a solution of achieving more accurate referrals and reducing the over-referral rate. However, although CART has a higher false positive rate than the Logistic Regression model, we believe that CART is still the better model as it also has a much lower false negative rate. This is important as misclassifying someone that does have risk of heart disease as having low risk could have life-or-death consequences. Therefore, the [Binary CART](#_w8lu5edzg8jj) achieves a good balance of false positives and false negatives.

CART is also much more suitable due to its simplicity and explainability, allowing GPs to clearly see the reasoning behind the prediction, and override it if they deem necessary. It is also resistant to missing values, allowing GPs to decide if certain tests need to be done.

### Multi-Class Predictions

|  |  |
| --- | --- |
| Default CART model with original dataset | Logistic Regression Model including hierarchical clusters |

In terms of multi-class predictions, we can see that the best Logistic Regression model has a slightly higher predictive accuracy of 57.53% as compared to that of the best CART model with 56.97%.

## Comparison Against Current Solutions

Looking at our best CART model with a binary predictive accuracy of 86.8%, we are able to more accurately determine whether a patient should be referred to NHCS. This shows that the CART model is very accurate in determining whether a patient is at immediate risk of developing heart disease, and that such analysis can produce very accurate results. As such, the number of unnecessary referrals to NHCS would be cut down, relieving a great amount of stress from the overworked healthcare workers and freeing up resources to benefit those that need them most.

In our introduction, we have already mentioned that NHCS found that seven in 10 people referred to it by the nine SingHealth Polyclinics did not actually need to see a specialist (Lai, 2021). From this statistic, we can say that NHCS referral system has a 70% false-positive error rate in our context. We compare that to the false-positive error rate with our best, which has a 16.7% false-positive error rate.

In comparison to current efforts to improve their referral system, NHCS is trying to streamline their process and reduce the number of visits to get their results (Lai, 2021). While their efforts have indeed improved efficiency of their process, our solution targets the problem from a different angle. NHCS’s streamlined process enables each patient to reduce their number of visits individually. However, patients that did not actually require the referrals are still getting referred to NHCS. Therefore, through tackling the issue of over-referrals, we can further compound the efficiency gains from their streamlined process.

Our solution not only achieves better predictive accuracy and a reduced false-positive error rate, but also enables a priority referral system which will help to ensure that patients receive treatment and consultation in order of urgency. This would greatly help patients to receive timely treatment and prevent dire consequences. We can take a look at the healthcare system in England for illustration, where patients suffering from heart disease are enduring long waits for treatment and may die as a result (Campbell, 2022). From this illustration, we can see how a priority system would be vital in saving the lives of many, where the order of treating patients can help to determine if one lives or dies.

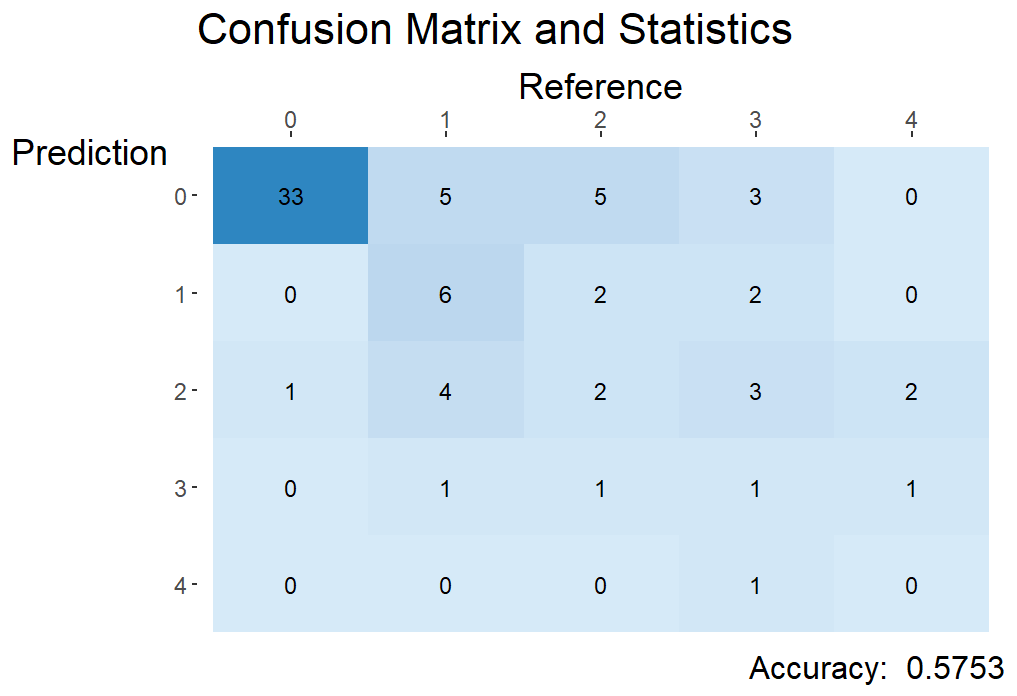


Fig 7. Logistic Regression model including hierarchical clusters

Although the raw number of 57.53% (Fig 7) may seem to be low, we have to factor in that these are multi-class predictions, which are harder to predict and would naturally achieve lower predictive accuracies as compared to binary predictions.

Additionally, we need to take into account the fact that the model is attempting to predict heart disease stages using suboptimal data. These data are typically considered insufficient in standard diagnostic procedures in the medical literature. For example, tests such as CT coronary angiography, Coronary artery calcium score and Cardiac stress imaging can only be conducted in a specialist clinic (HealthHub, 2022). While the accuracy may not be optimal, the value that our models add is to provide a best guess at the severity of heart disease with data from non-invasive procedures. Therefore, we deem that our accuracies for multi-class prediction as acceptable.

Given that most GPs would have some form of information to make an informed decision, we should not be comparing our model to a purely random guess. However, they would base their decision off their experience and intuition rather than a model. It is crucial to take note that the current models that are available in the industry right now do not provide any form of metric to aid the GPs in indicating a form of priority on their referrals, making ours the first of its kind.

## Areas of Improvement

One limitation of our models is their false negative rate, which is the rate at which patients that are at immediate risk of developing heart disease are instead classified as not having immediate risk. Our best model which is the CART model has a false negative rate of 10.4%. However, this is due to the inherent limitations in this project, such as the lack of enough relevant data as well as our lack of domain expertise.

We can further improve our methodology by collecting and training our models on local data. This would ensure that our model captures local patterns thus increasing the accuracy for the actual population that it is serving. We can also explore variables that are outside the original dataset. This may include other medical data such as personal medical history and family history, or can even include alternative data such as smoking and drinking patterns and exercise frequency. In general, the more independent variables with explanatory power, the higher the accuracy of the model.

Finally, we can explore machine learning models beyond logistic regression and CART. By using more sophisticated models such as SVM, Naive Bayes, LightGBM and Random Forest algorithms, we can potentially capture non-linear relationships in the data and achieve a higher accuracy (Karthick et al., 2022). However, we need to take into account the increased model complexity and thus the lower model transparency and explainability as we use more sophisticated models for our predictions.

With access to the right resources, we believe that the models can be improved to be significantly more accurate, with much more nuanced predictions than the stage of heart disease, allowing for even more detailed analysis of a patient's health and tailored healthcare to suit their needs.

# 

# Recommended Implementation

## Referral Process

Patient visits GP

Framingham Risk Score is calculated

If Score exceeds threshold

Immediate risk is predicted

Referral to NHCS

Other preventive care measures

High immediate risk

Low immediate risk

Stage 4

Stage 3

Stage 2

Stage 1

Priority level

Fig 8. Flowchart of new referral process

Fig 8 shows our recommended approach on how to best implement our analytics solutions to achieve more accurate referrals as well as prioritise patients that are in need of the most urgent care.

NHCS should capitalise on its closely connected network of GPs in Singapore and include the use of our solution as part of their standard operating procedures. Implementation of this solution will be very feasible as GPs look to NHCS for the latest guidelines on dealing with cardiovascular health (NHCS, 2022).

Firstly, when the patient visits the GP, their Framingham Risk Score is calculated. This is to estimate their 10 year risk of developing heart disease. If their risk score exceeds the threshold, the GP would then conduct further tests as deemed necessary, in order to predict the patient’s immediate risk through our models.

If the patient has a Framing Risk Score exceeding the threshold but has low immediate risk according to our models, then the patient should not be referred to NHCS. Instead, the patient should be advised on taking preventive care measures instead such as working towards a healthier lifestyle. As these patients have high 10 year risk but low immediate risk, the best course of action is through preventive measures rather than diagnostic tests, as symptoms of heart disease may not be prevalent yet. This would help reduce unnecessary referrals to NHCS.

However, if the patient has a high immediate risk, they would be referred to NHCS for further diagnosis. The patient’s heart disease stage would also be predicted to gauge the urgency of their referral. This would help sort out patients that are at higher risks, allowing them to consult with the specialists ahead of the lower risk patients.

## Prototype

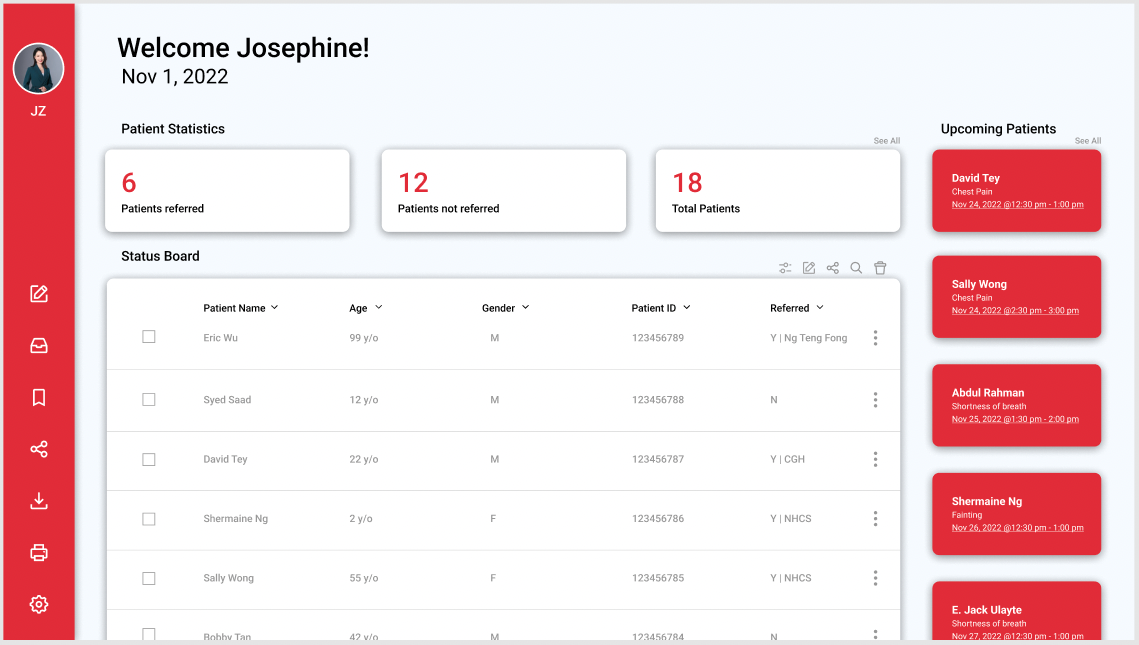


Fig 9. UI of Home Screen

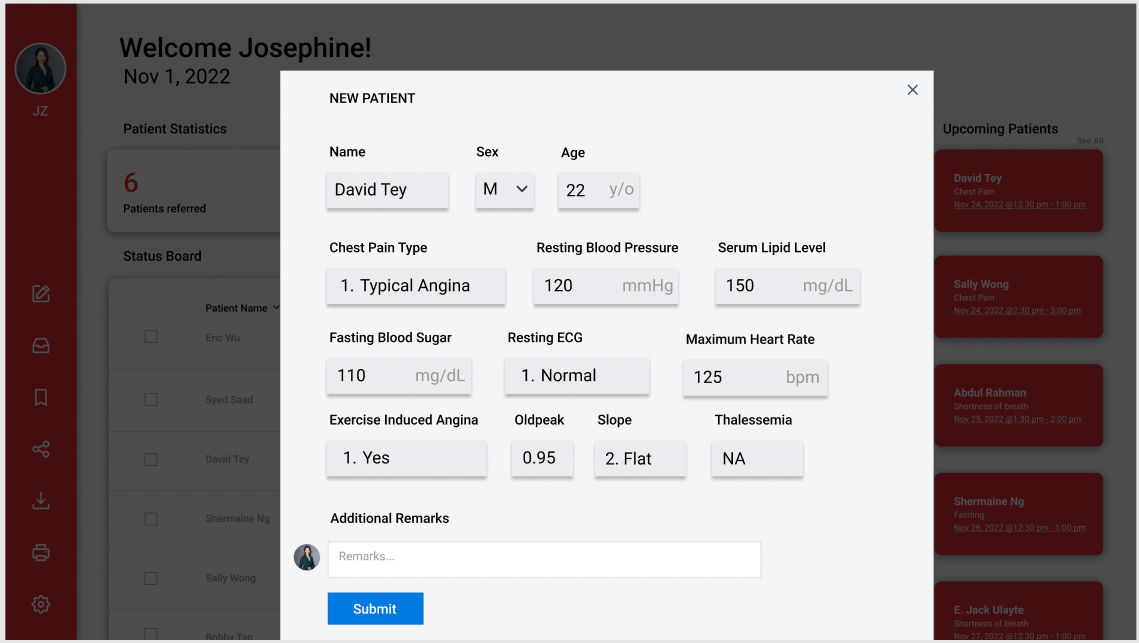


Fig 10. UI of new patient popup

Adoption of this solution will be relatively simple as it is accessible and easy to pick up. Our solution will be made available on the web and GPs can just access the website from their work computer. GPs can just log in and create a patient profile along with the patient data obtained from his preliminary tests and receive a binary decision of whether to refer the patient or not. Should the decision be to refer, there will also be a referral priority that is reflected based on our multi-class prediction to give the specialist a gauge of the urgency of the referral. Patient data will be stored in a centralised database for tracking of patients and to receive updates on the patients’ status.

# Conclusion

We believe that our proposed solution is accurate, easy to implement, cost-efficient, and provides a high return on investment. By combining the Framingham Risk Score with the patient’s immediate risk level, GPs would be able to more accurately refer patients to NHCS. This would significantly reduce the number of simple outpatient consultations that NHCS needs to handle, thereby reducing the strain on their manpower and wastage of resources.

Looking at the figures from Mar 2021 (Fig 3), according to our estimates (refer to [Appendix F](#_s9g0z7j40khp)) our best model would have been able to cut down the number of patients referred to NHCS from 132,634 to 47,768, a total reduction of 84866 referral, representing a decrease of 64.0%. As each trip costs a patient approximately $38 dollars (Lai, 2021) this will amount to a total expenditure reduction of $3,224, 908 dollars.

This would very effectively allow NHCS to achieve favourable business outcomes, as the reduction of over-referrals would significantly save on operating cost, while improving their staff’s morale through reduction of overwork. With this newfound efficiency, the quality of care and service provided to their other patient’s would also greatly increase, boosting NHCS’ reputation as the leading centre for cardiovascular health in the region.

We believe that our preliminary research would lay the foundation for NHCS and other specialist centres in Singapore to replicate and improve on our work. With more research into this with more sophisticated models, the predictive accuracy can be significantly improved as well. This could also allow the use of machine learning models in other healthcare applications.

Additionally, since NHCS is a government organisation, reducing costs on unnecessary consultations would free up its budget for other health programs which would further benefit the Singapore population.

# References

US Trade. (2022). Singapore's aging population. International Trade Administration | Trade.gov. Retrieved October 29, 2022, from https://www.trade.gov/market-intelligence/singapores-aging-population#:~:text=As%20the%20number%20of%20seniors,over%20the%20age%20of%2065

Jaul, E., &amp; Barron, J. (2017, December 11). Age-related diseases and clinical and public health implications for the 85 years old and over population. Frontiers in public health. Retrieved October 29, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5732407/

*Screening for heart disease*. HealthHub. (2022). Retrieved October 29, 2022, from https://www.healthhub.sg/live-healthy/16/screening\_heart\_disease

Lai, L. (2021, November). *News*. HealthXchange. Retrieved October 24, 2022, from https://www.healthxchange.sg/news/new-initiative-nhcs-shp-saves-heart-patients-time-money

Tan, J. (2020, September 18). *Seeing a cardiologist in Singapore: The Complete Guide (2021)*. Human. Retrieved October 24, 2022, from https://www.human.com.sg/the-complete-guide-to-seeing-a-cardiologist-in-singapore

Blaha, M. B. (2015, February 16). *Most clinical 'calculators' over-estimate heart attack risk - 02/16/2015*. Johns Hopkins Medicine, based in Baltimore, Maryland. Retrieved October 24, 2022, from https://www.hopkinsmedicine.org/news/media/releases/most\_clinical\_calculators\_over\_estimate\_heart\_attack\_risk

Campbell, D. (2022, June 16). *Almost 100,000 facing excessive wait for serious cardiac care in England*. The Guardian. Retrieved October 25, 2022, from https://www.theguardian.com/society/2022/jun/16/record-6-5-million-people-waiting-for-nhs-hospital-treatment-in-england

*Commentary: Healthcare workers are at their breaking point but most don't want to quit*, *CNA*. Available at: https://www.channelnewsasia.com/commentary/frontline-workers-burnout-covid-19-quit-leave-hospital-doctor-nurse-health-2553806 (Accessed: October 29, 2022).

NHCS. (2022). *Healthcare partners*. Healthcare Partners – GP | National Heart Centre Singapore. Retrieved October 25, 2022, from https://www.nhcs.com.sg/education-training/healthcare-partners

MOH, (2021)<https://www.myheart.org.sg/health/heart-disease-statistics/#:~:text=Cardiovascular%20disease%20accounted%20for%2032,to%20heart%20diseases%20or%20stroke>

https://www.singhealth.com.sg/about-singhealth/newsroom/Documents/[Web%20Version]%20SingHealth%20DukeNUS%20Annual%20Report%2020-21.pdf

Lim, B (2022) <https://www.channelnewsasia.com/commentary/frontline-workers-burnout-covid-19-quit-leave-hospital-doctor-nurse-health-2553806>

Karthick, K., Aruna, S. K., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Thelkar, A. R. (2022). Implementation of a heart disease risk prediction model using machine learning. *Computational and Mathematical Methods in Medicine*, *2022*, 1–14. https://doi.org/10.1155/2022/6517716

Dataset: <https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data>

QRISK: <https://qrisk.org/three/>

Euro-Score: https://www.mdcalc.com/calc/10179/european-system-cardiac-operative-risk-evaluation-euroscore-ii

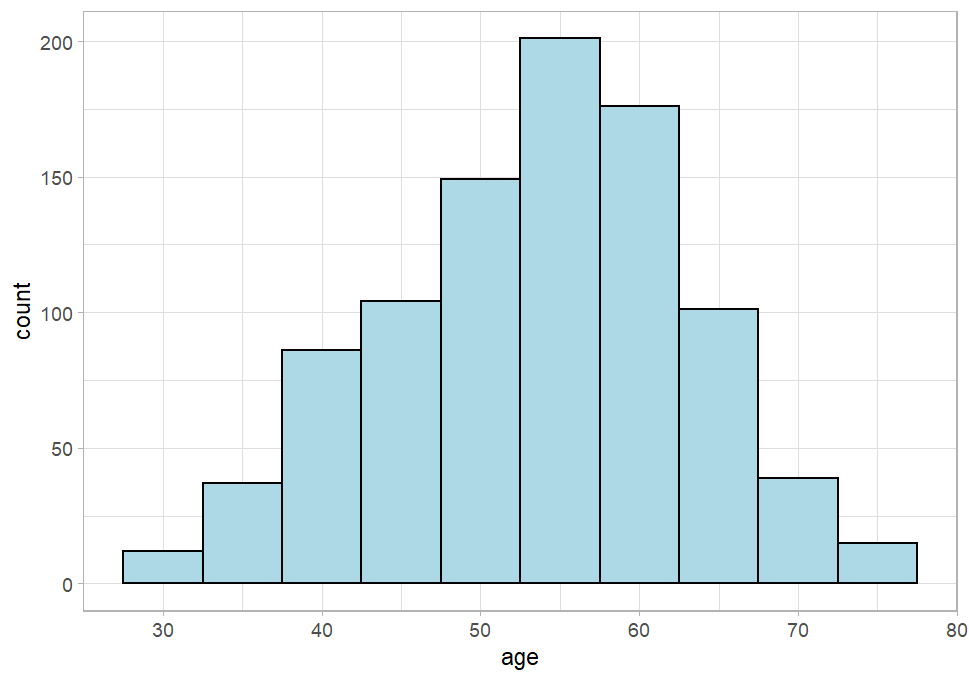
# 

# Appendix A - Data Dictionary

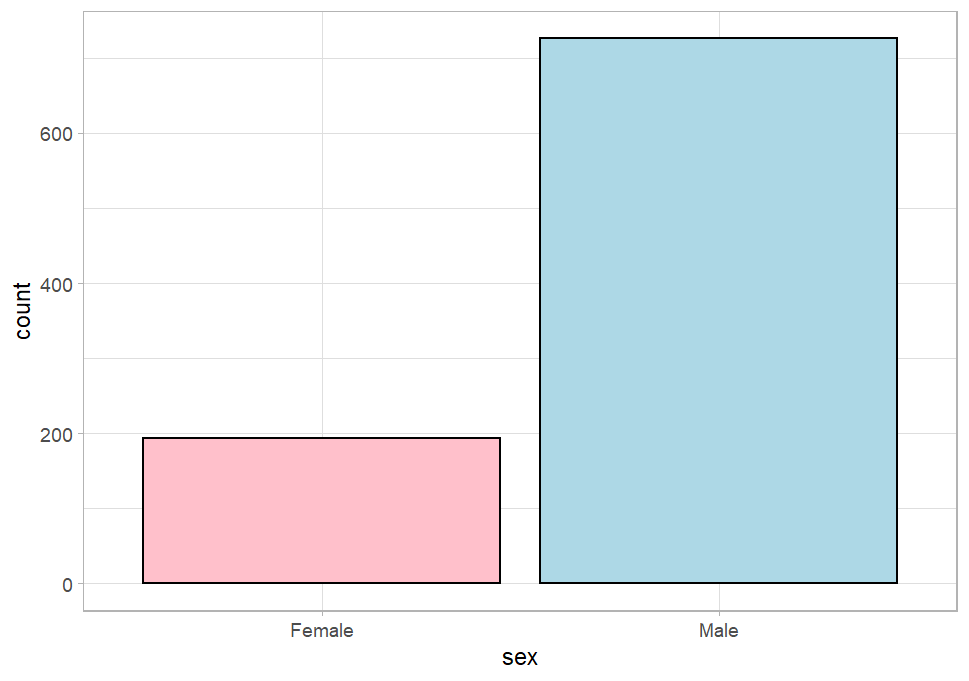
|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Data Type** |
| Age | Age in years | Numerical |
| Sex | Male & Female | Categorical |
| CP | Chest pain: typical angina, atypical angina, non-anginal, asymptomatic | Categorical |
| trestbps | Resting blood pressure (in mmHg at entry to the health center) | Numerical |
| chol | Serum lipid level in mg/dL | Numerical |
| fbs | TRUE if the fasting blood sugar level > 120 mg/dL; else FALSE | Categorical |
| restecg | Resting ECG results: normal, st-t abnormality; lventricular, hypertrophy | Categorical |
| thalach | Maximum heart rate achieved | Numerical |
| exang | Exercise induced angina (TRUE/FALSE) | Categorical |
| oldpeak | ST depression induced by exercise relative to rest | Categorical |
| slope | The slope of the peak exercise ST segment: upsloping, flat, downsloping | Categorical |
| thal | Thalassemia: normal, fixed defect, reversible defect | Categorical |
| Num | Predicted attribute | Categorical |

# Appendix B - Univariate Distributions

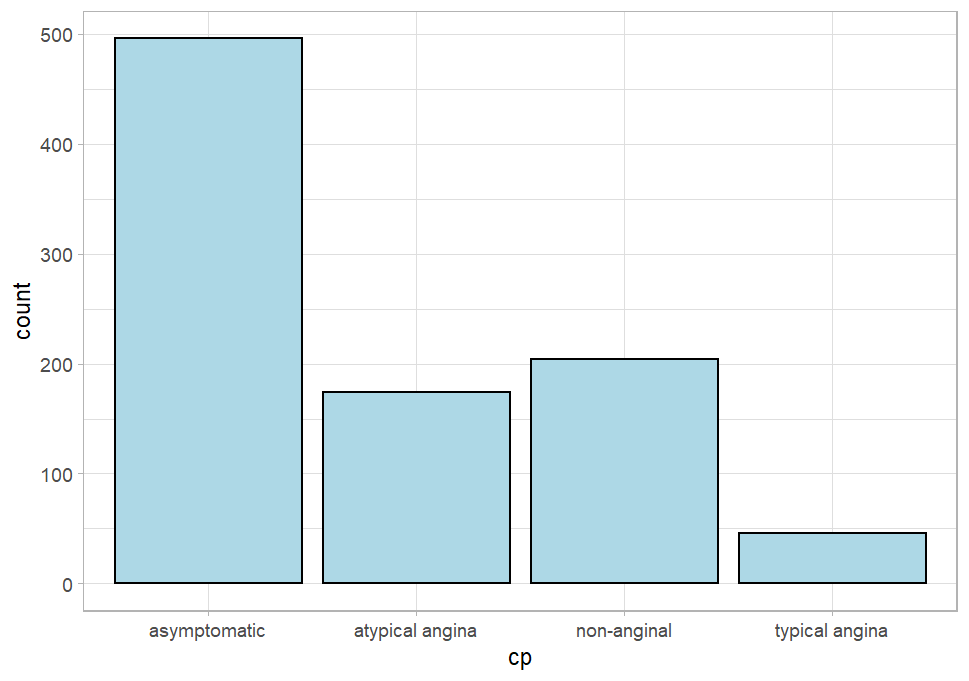
Age Distribution

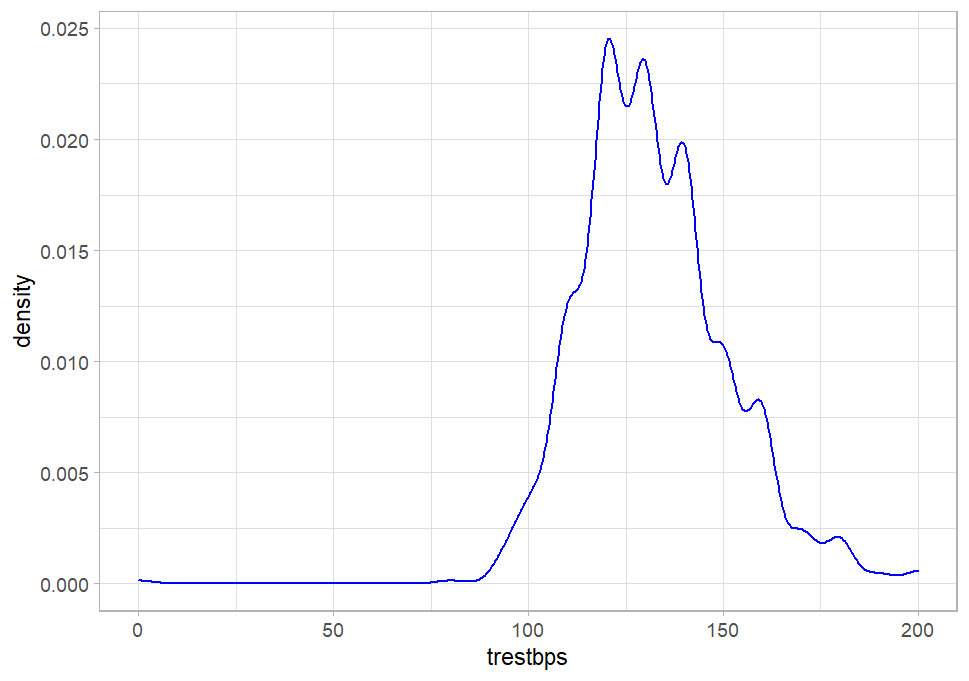


Gender Distribution

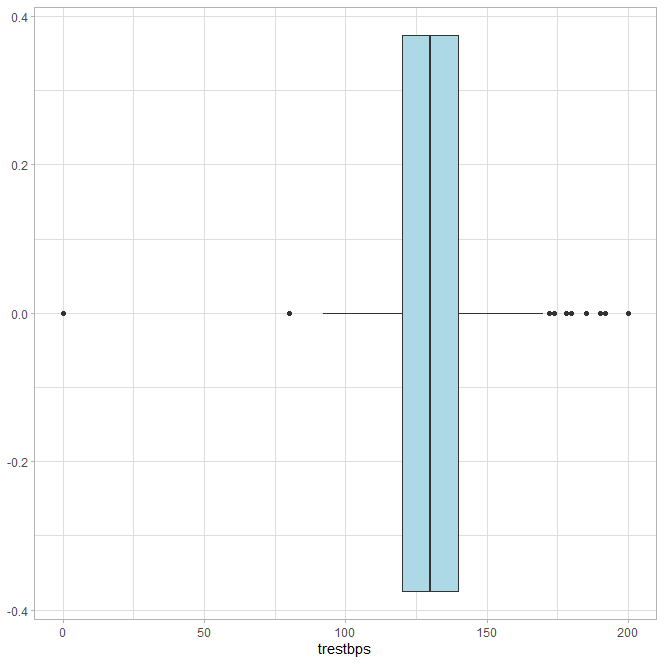


Chest Pain Type

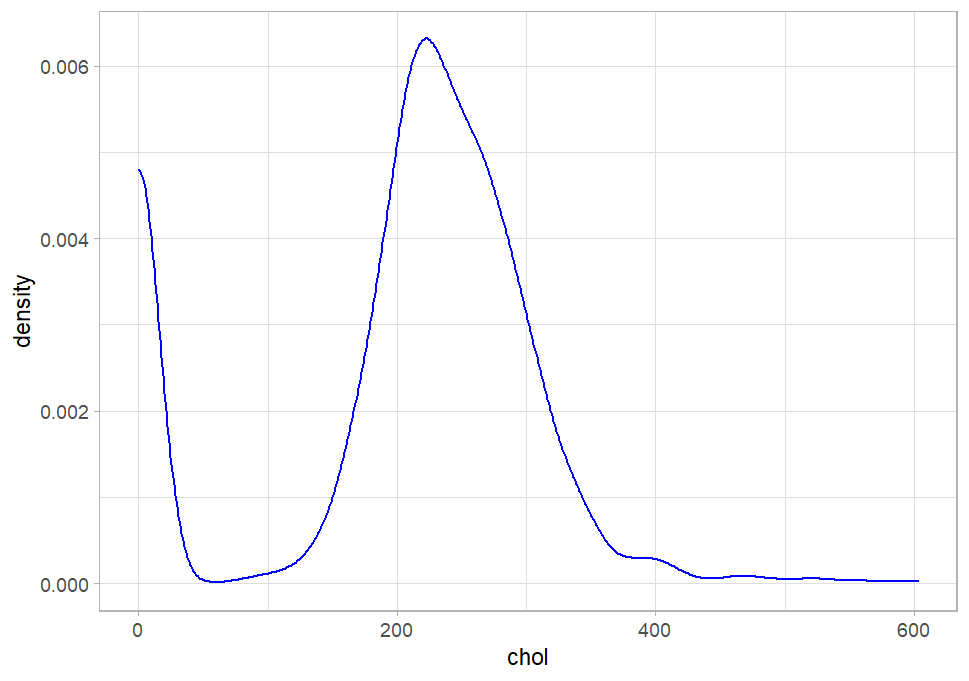


Distribution of Resting Blood Pressure 

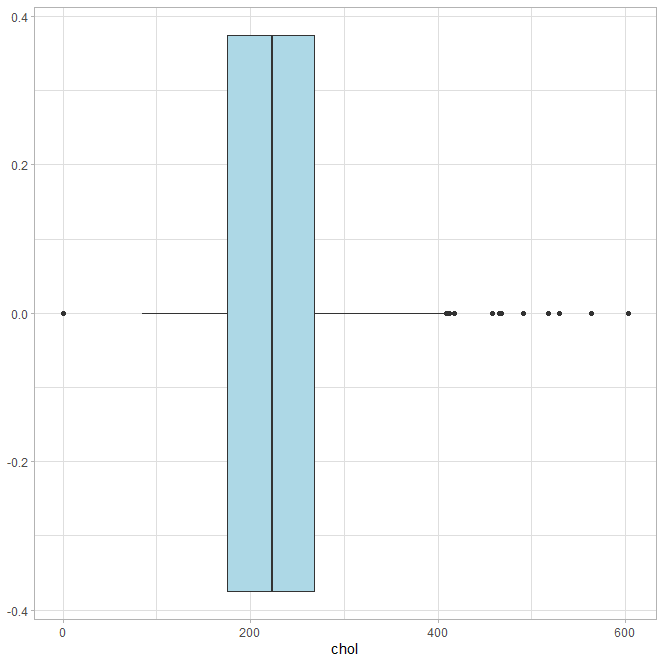
Boxplot of Resting Blood Pressure



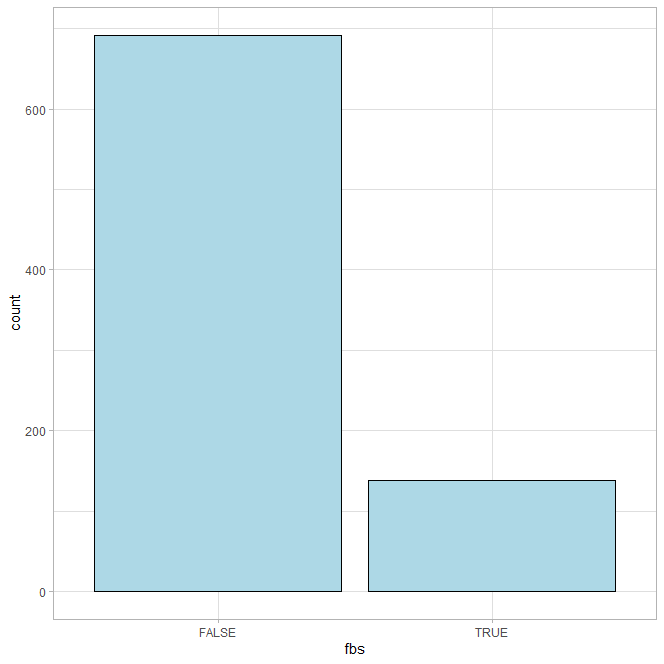
Density plot of (serum cholesterol in mg/dl)



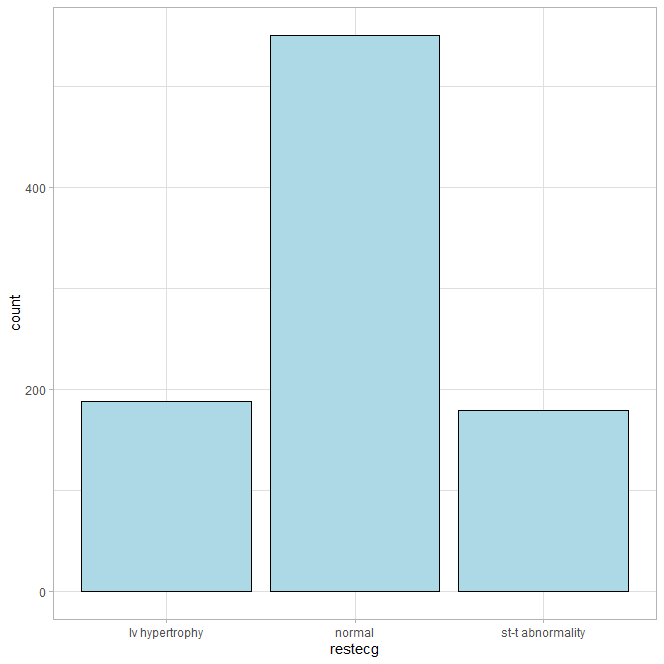
Boxplot of (serum cholesterol in mg/dl)



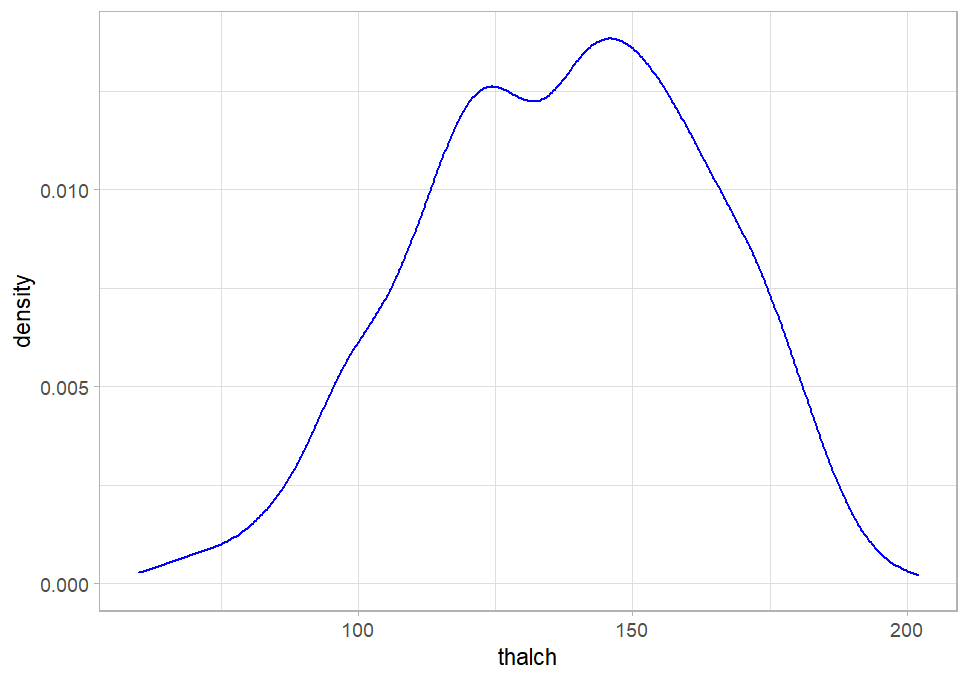
Barplot of fbs (if fasting blood sugar > 120 mg/dl)



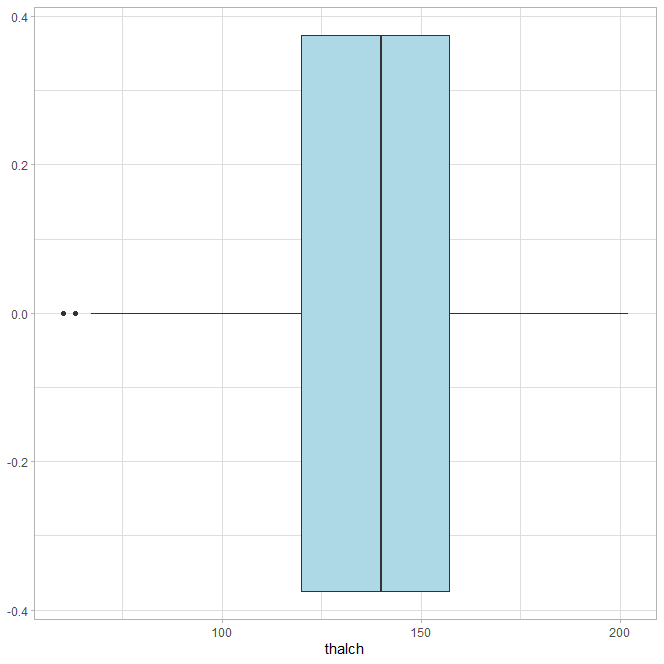
Barplot of restecg (resting electrocardiographic results)



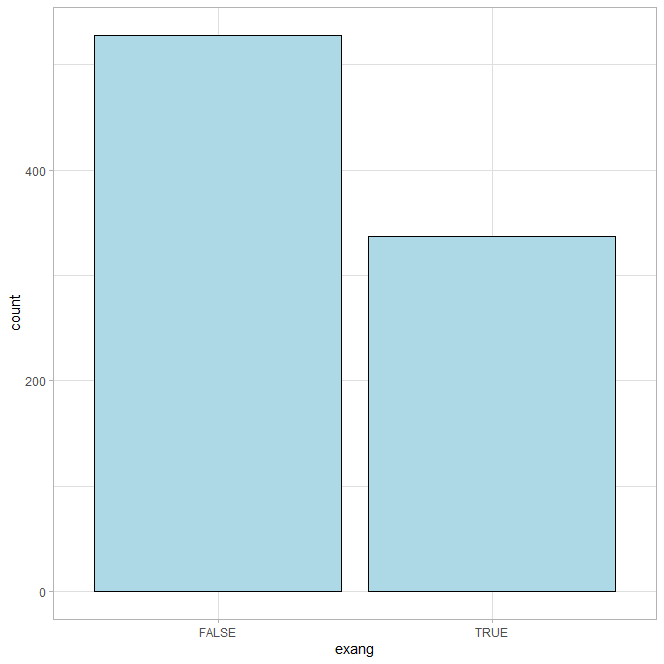
Density plot of thalch (maximum heart rate achieved)



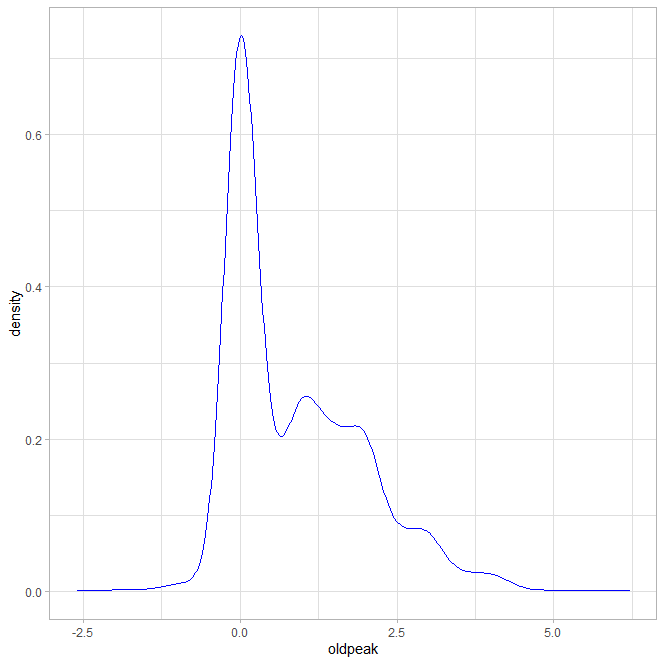
Boxplot of thalch (maximum heart rate achieved)



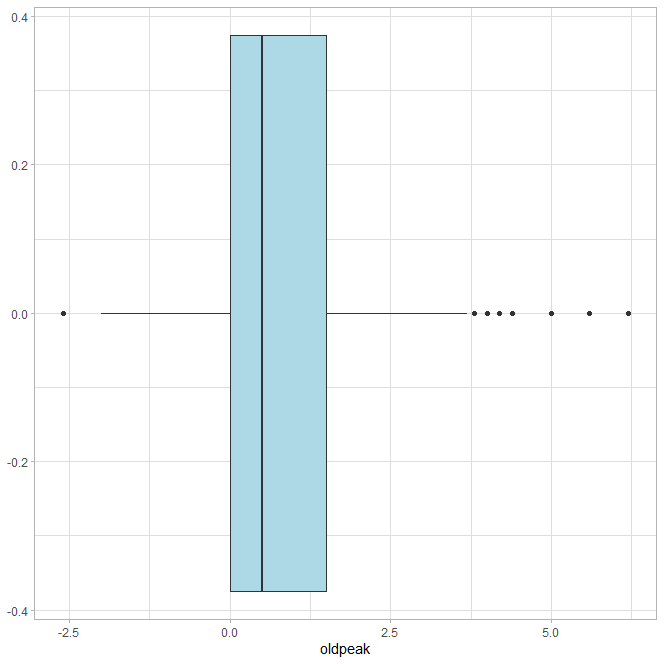
Barplot of exang (exercise-induced angina (True/ False))



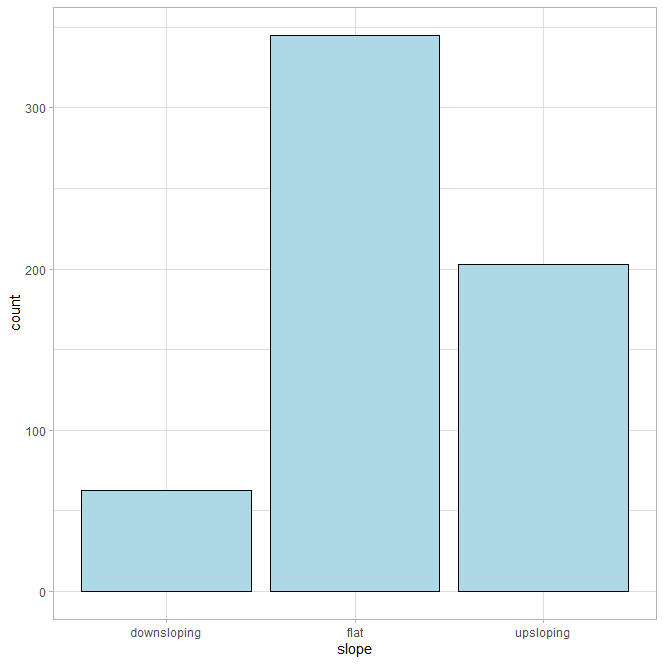
Density plot of oldpeak: ST depression induced by exercise relative to rest



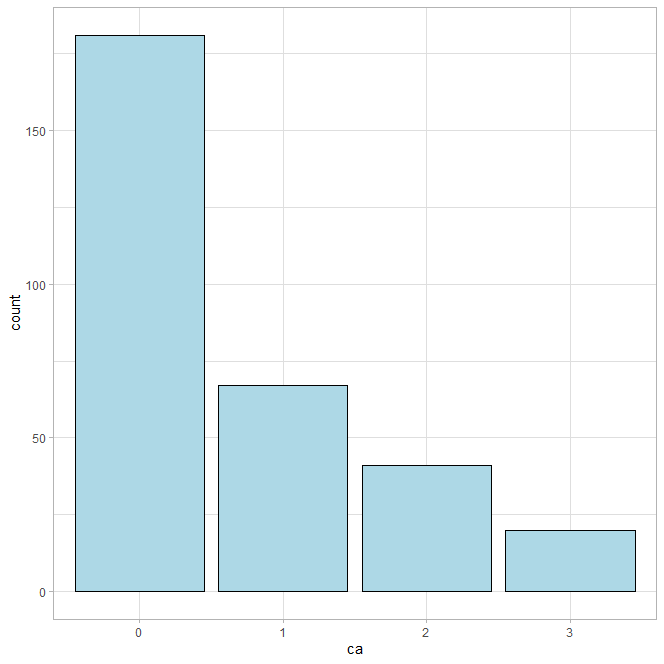
Boxplot of oldpeak: ST depression induced by exercise relative to rest



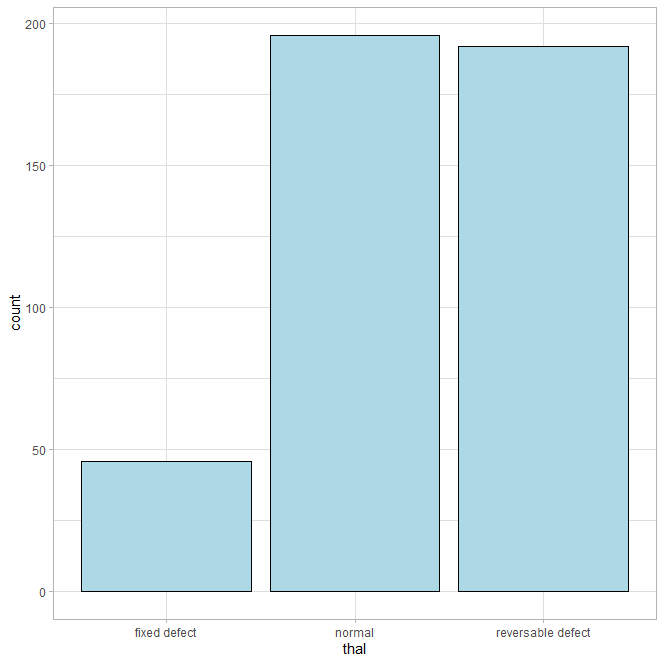
Barplot of slope : slope of the peak exercise ST segment



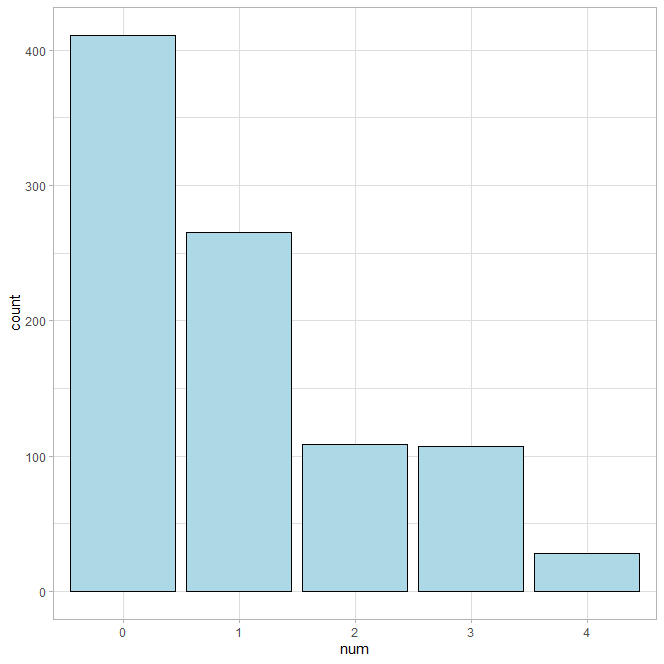
Barplot of ca: number of major vessels (0-3) colored by fluoroscopy



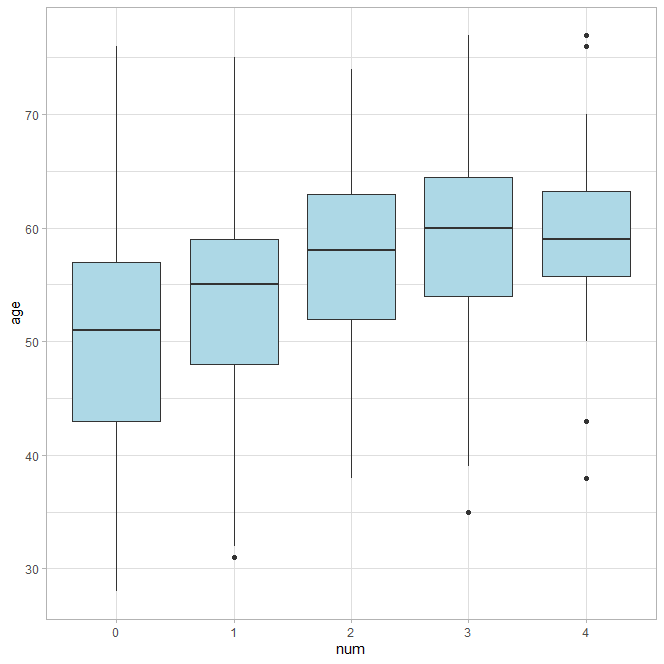
Barplot of Thalassemia : [normal; fixed defect; reversible defect]



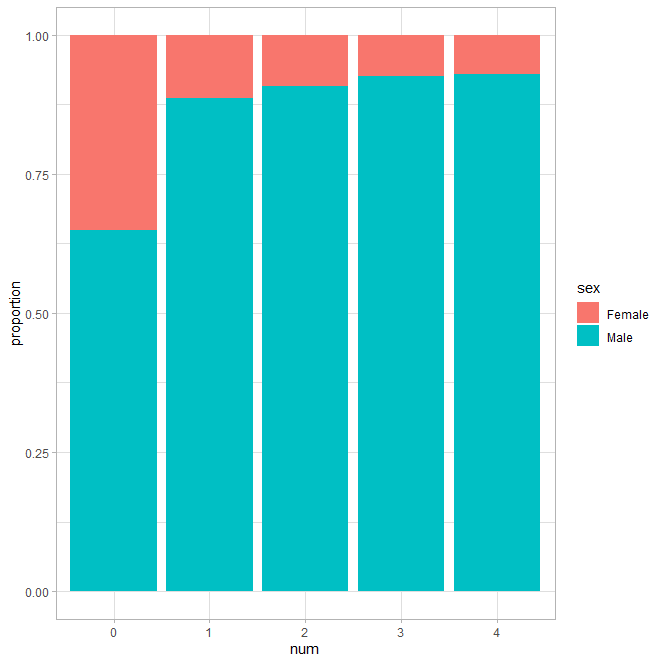
Barplot of Num:



Boxplot of age against num



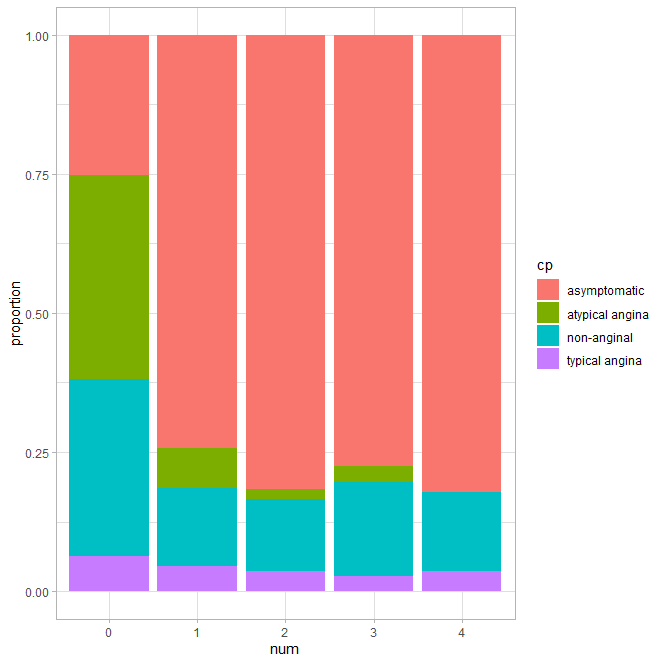
Distribution of num against sex



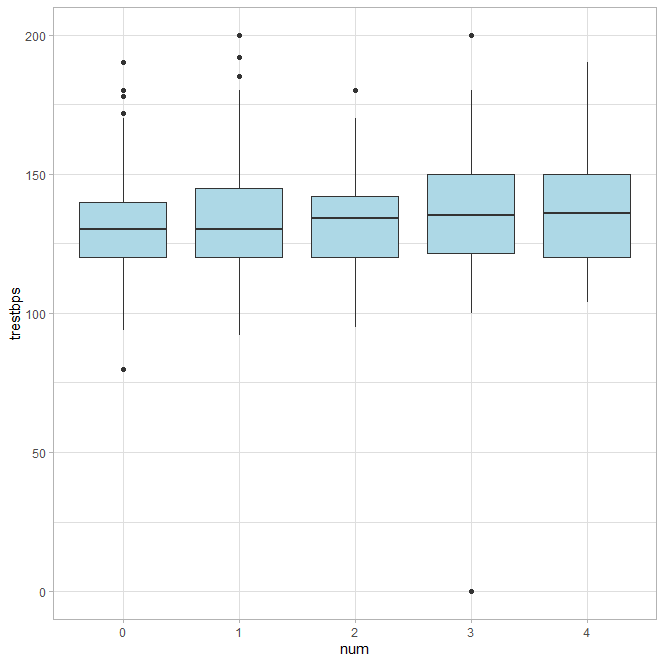
# 

# Appendix C - Bivariate Distributions & Correlations

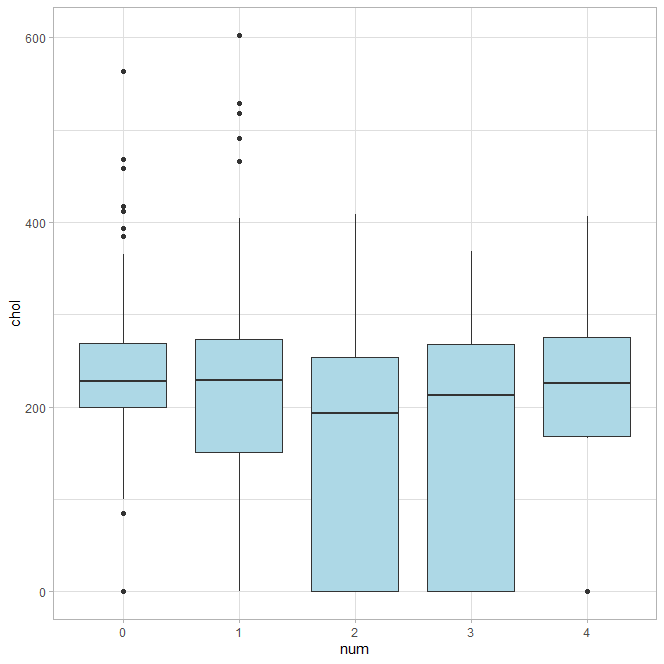
Stacked bar chart of cp against num



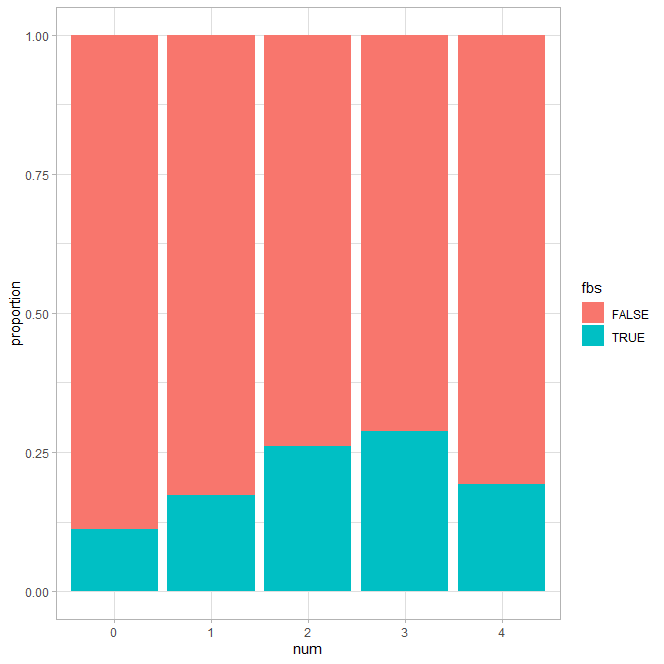
Boxplot of trestbps: resting blood pressure against num



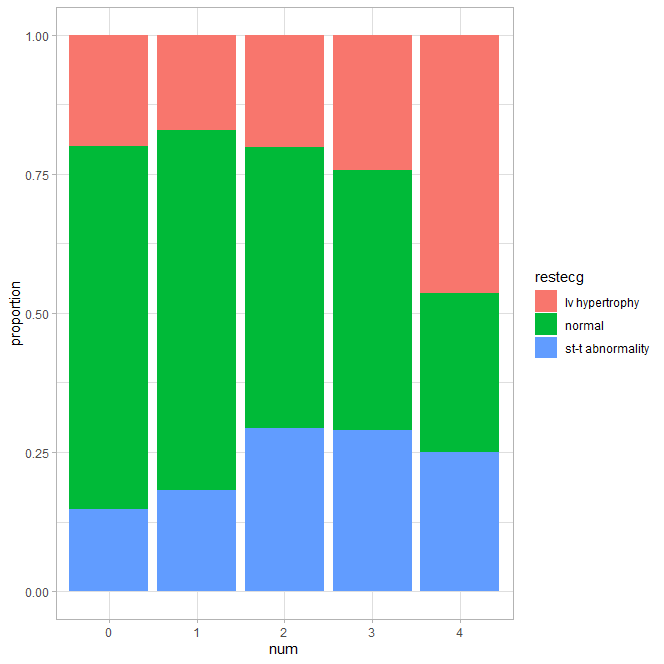
Boxplot of chol (serum cholesterol in mg/dl)l against num



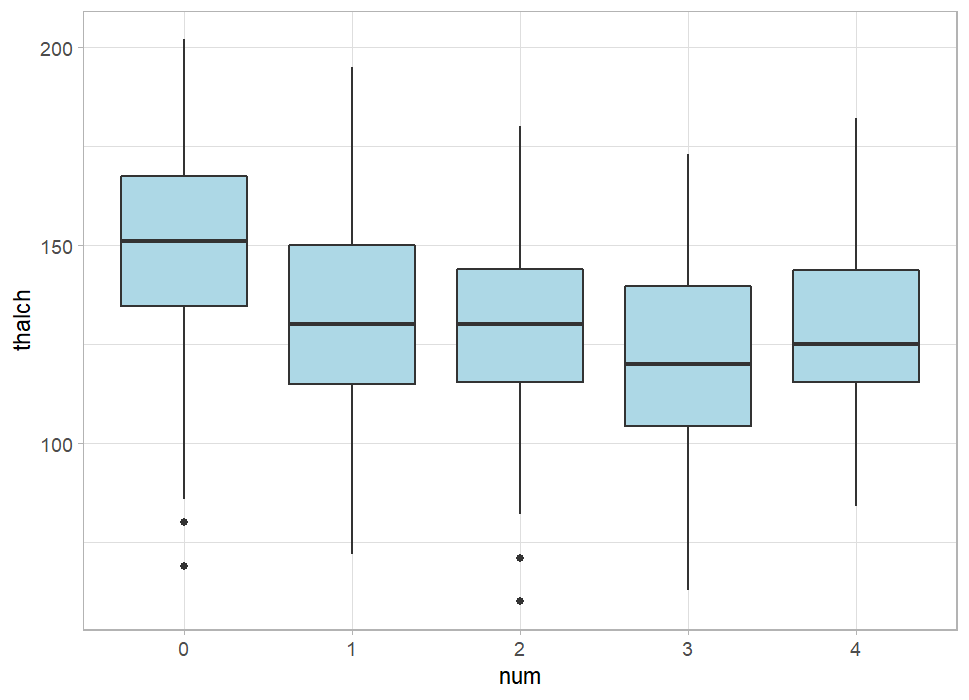
Stacked bar chart of fbs against num



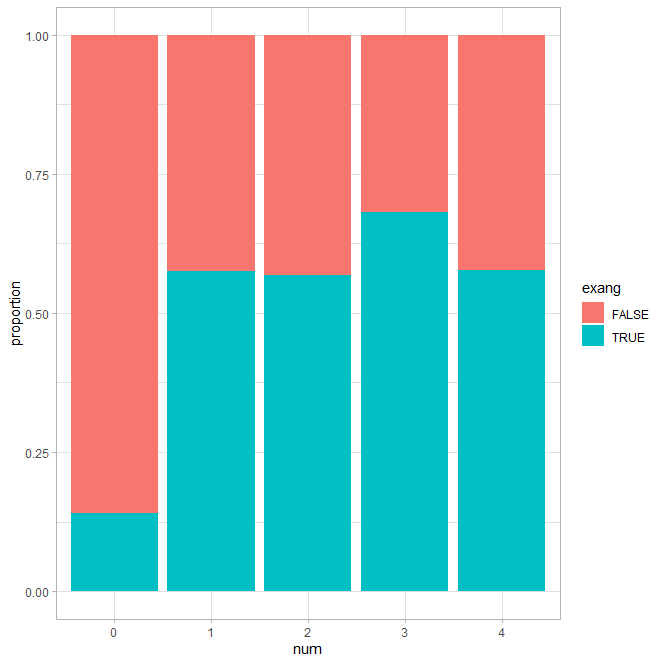
Stacked bar chart of restecg(resting electrocardiographic results) against num



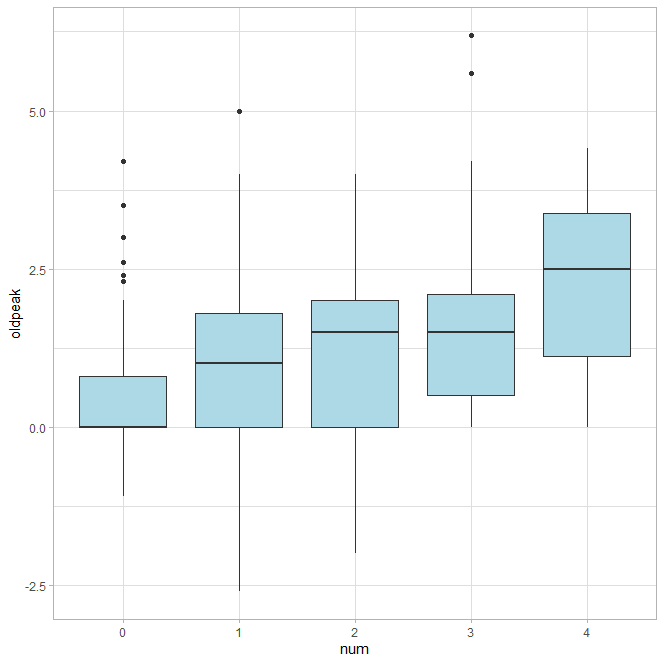
Boxplot of thal(Thalassemia ) against num



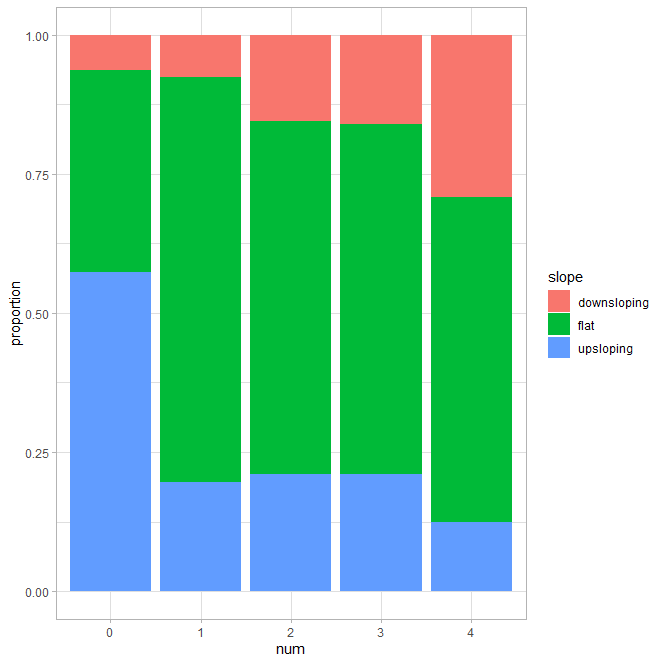
Stacked bar chart of exang(exercise-induced angina ) against num



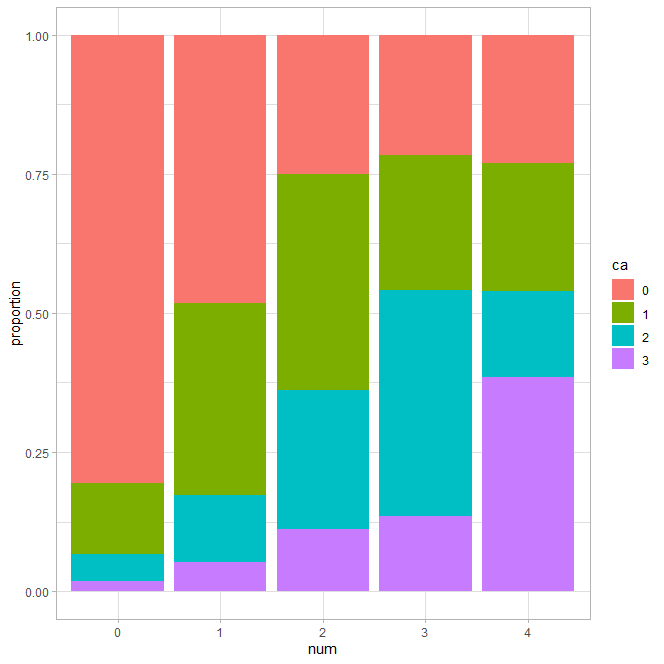
Boxplot of oldpeak ( ST depression induced by exercise relative to rest) against num



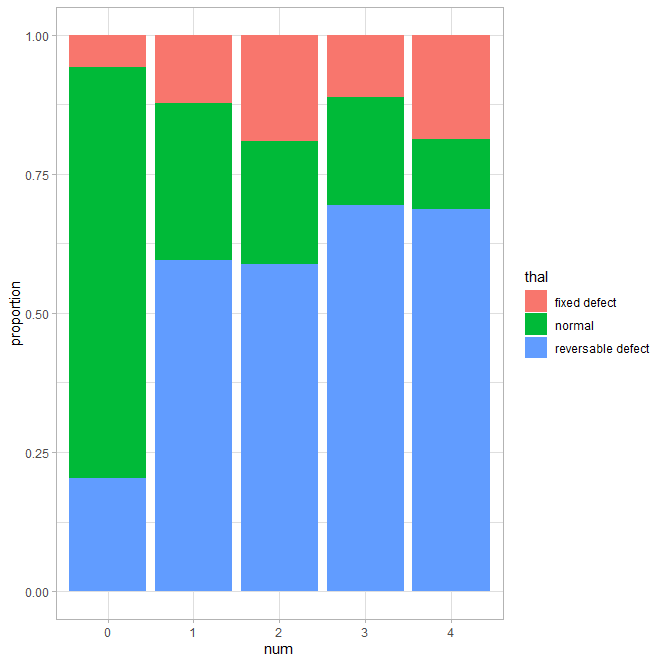
Stacked bar chart of slope against num



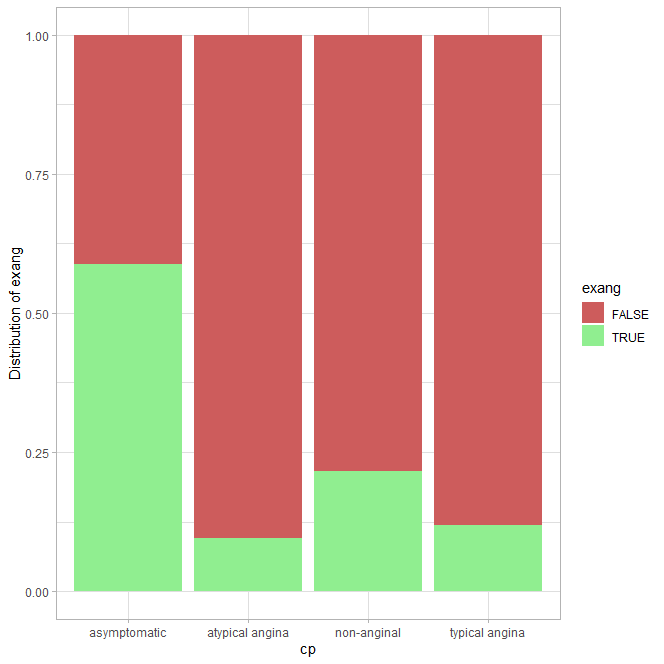
Stacked bar chart of ca (number of major vessels (0-3) colored by fluoroscopy) against num



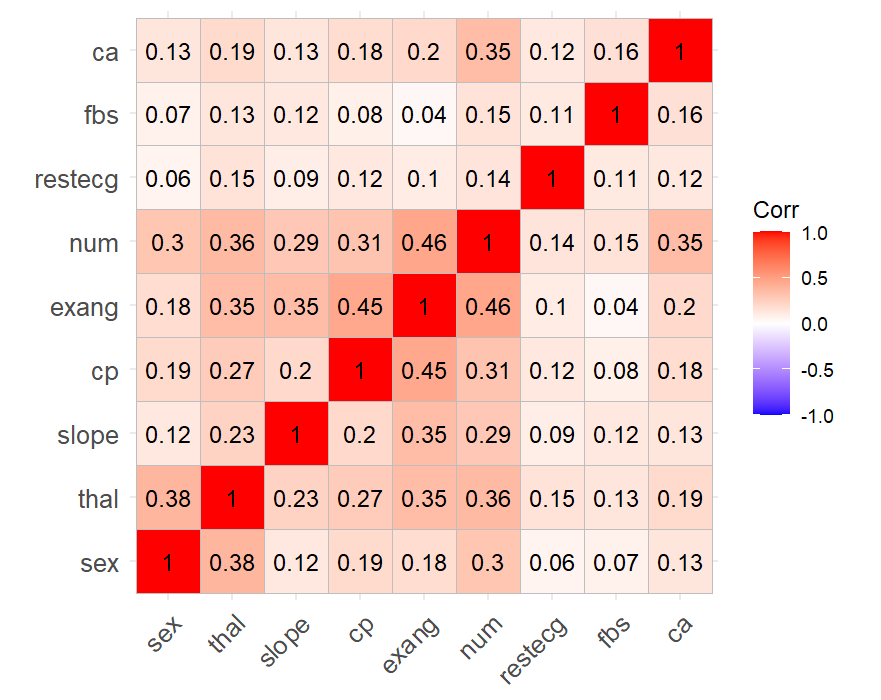
Stacked bar chart of thal (Thalassemia ) against num



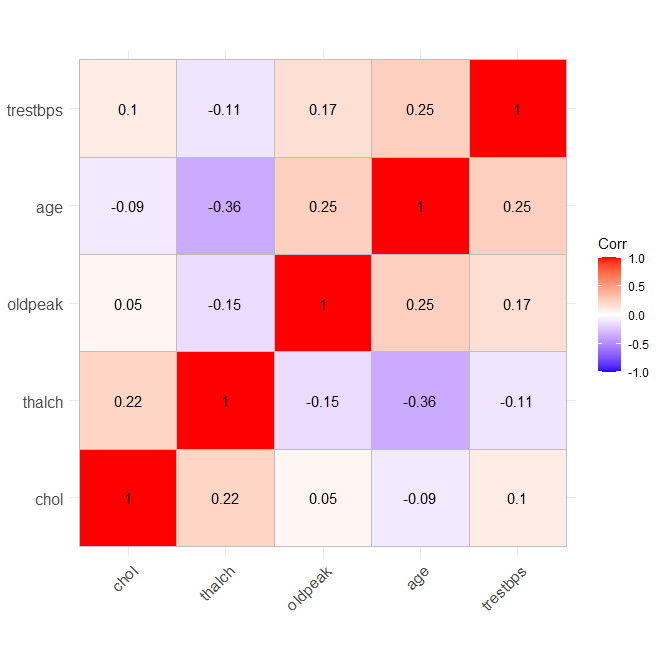
Stacked bar chart of exang against cp



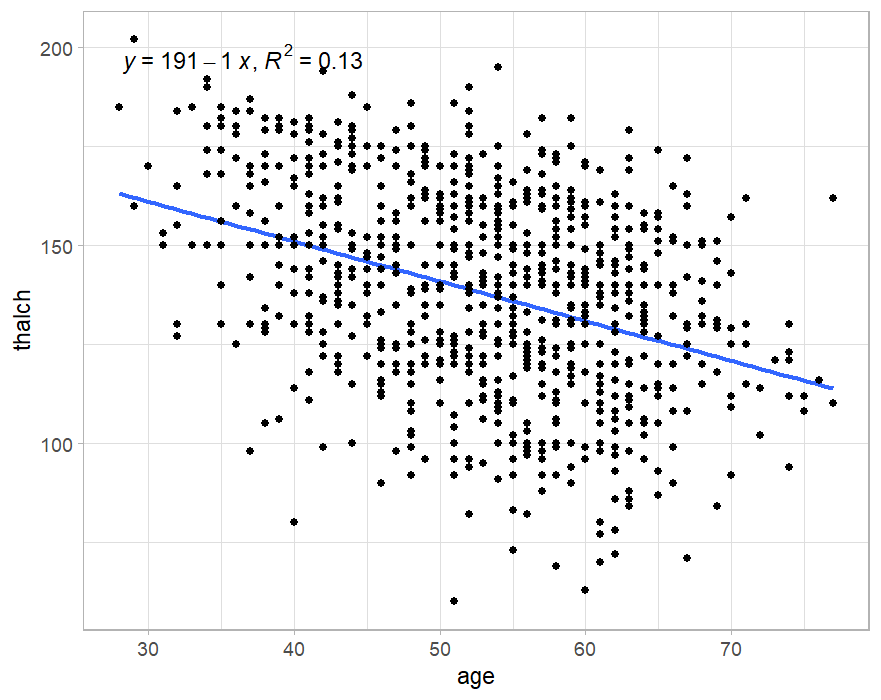
Correlation matrix of categorical variables



Correlation matrix of continuous variables



Scatterplot of age against thalch



# 

# 

# 

# 

# Appendix D - K-Means Clustering & Hierarchical Clustering

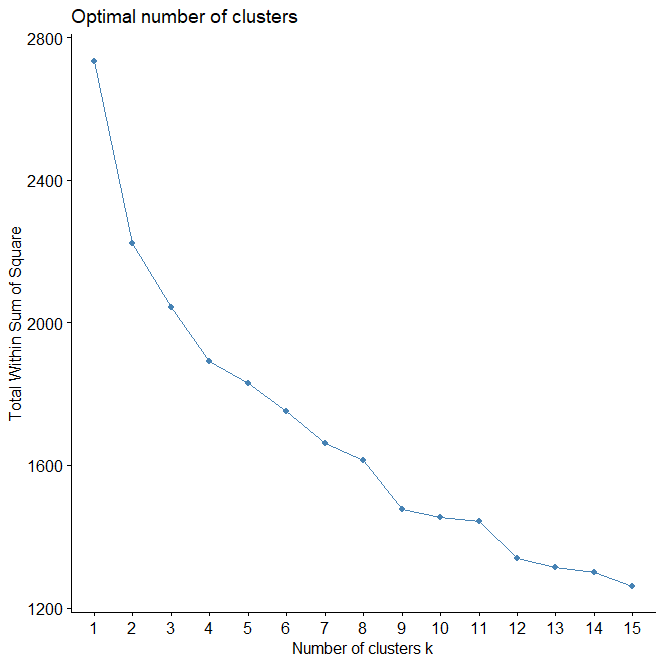
## Preprocessing

In order to prepare our dataset for clustering analysis, we first did one-hot encoding for the original categorical variables. This includes the missing values, which are treated as 0 for the entries that are missing. We also normalised our data to scale the numerical variables between 0 and 1. As distances between points are calculated using L2 norm in clustering analysis, we had to ensure that the different columns were similar in magnitude to ensure equal weightage between the various columns.

As the L2 norm sums the square of each independent variable, an independent variable that has a bigger range would naturally add up to a disproportionate amount of the L2 norm, thus drowning out other variables, especially categorical variables that only range from 0 to 1. Min-max normalisation would help mitigate this issue by scaling all variables to between 0 and 1.

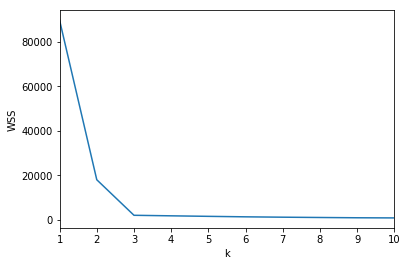
## K-Means Clustering

Total Within Sum of Squares vs Number of Clusters

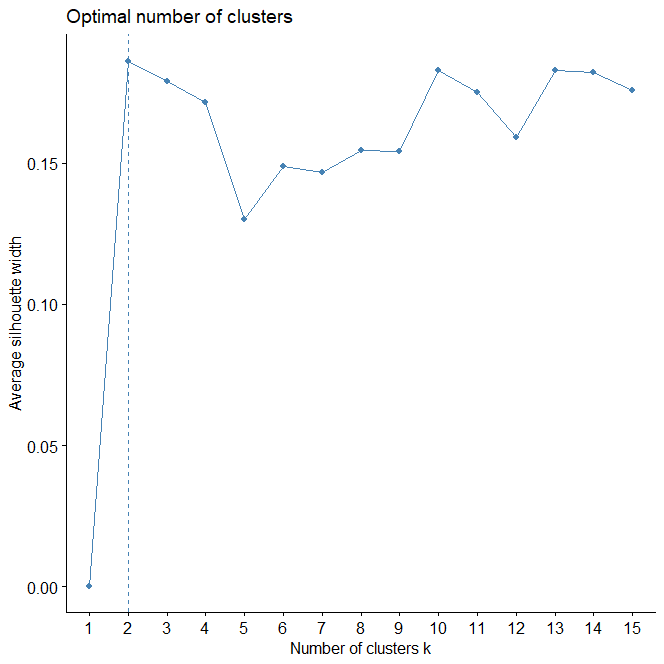


To find the optimal number of clusters, we first plotted the total within sum of squares (WSS) metric against the number of clusters k from 1 to 15. The WSS metric first calculates the centre of each cluster, then calculates the squared distance of each point from its cluster centre. It then adds up all the squared distances. Typically, the plot would look something like an elbow, where the WSS would decrease rapidly until a certain number of clusters, before becoming flat or decreasing very slightly. The kink in the graph typically represents a good or natural amount of clusters as the total distance from cluster centres is no longer decreasing even as you add more clusters.

Figure from <https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb>



Unfortunately, this pattern was not observed in our plot and we had to use the silhouette method to find the optimal number of clusters.



The silhouette method compares how similar a point is to its own cluster to how dissimilar a point is to other clusters. Thus, a higher silhouette width means that the difference between similarity to its own cluster and dissimilarity to other clusters is bigger, representing a better clustering of data.

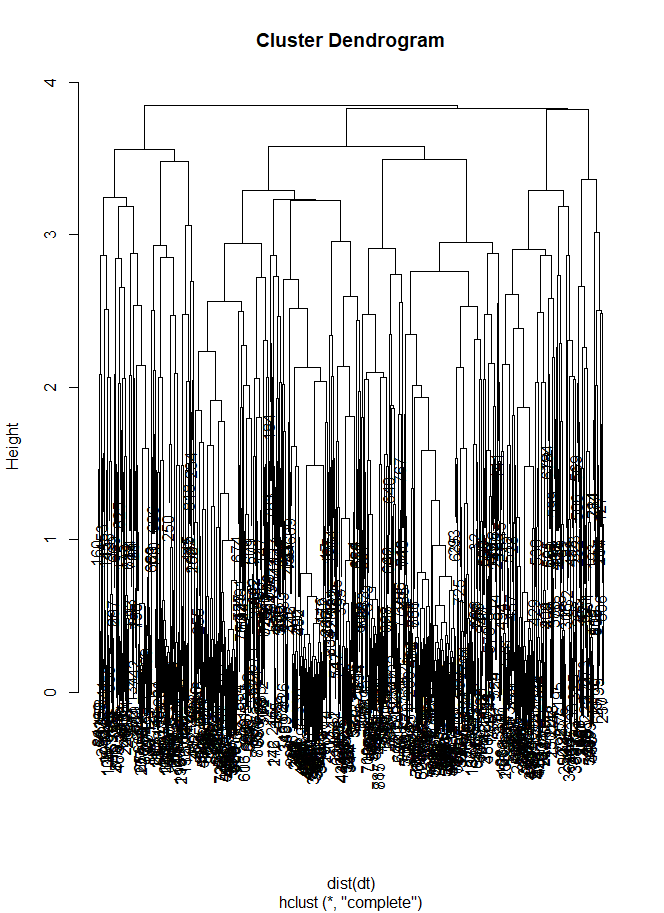
While the silhouette width peaked at k = 2, it bottomed at k = 5 before increasing again to similar levels at k = 10. As the difference was quite small, we opted for k = 10 as there may

potentially be smaller clusters within the 2 main clusters that are useful for model prediction.

## Hierarchical Clustering

We also used hierarchical clustering to cluster our dataset.

Plot of hierarchical cluster



Hierarchical clustering starts off with the minimum distance between two data points and slowly increases this distance or “height”. As the height increases, points that are spaced closer together, with a distance that is less than the height would be merged together to form a new cluster. This process would continue until the height increases enough such that all points are included together as one single cluster.

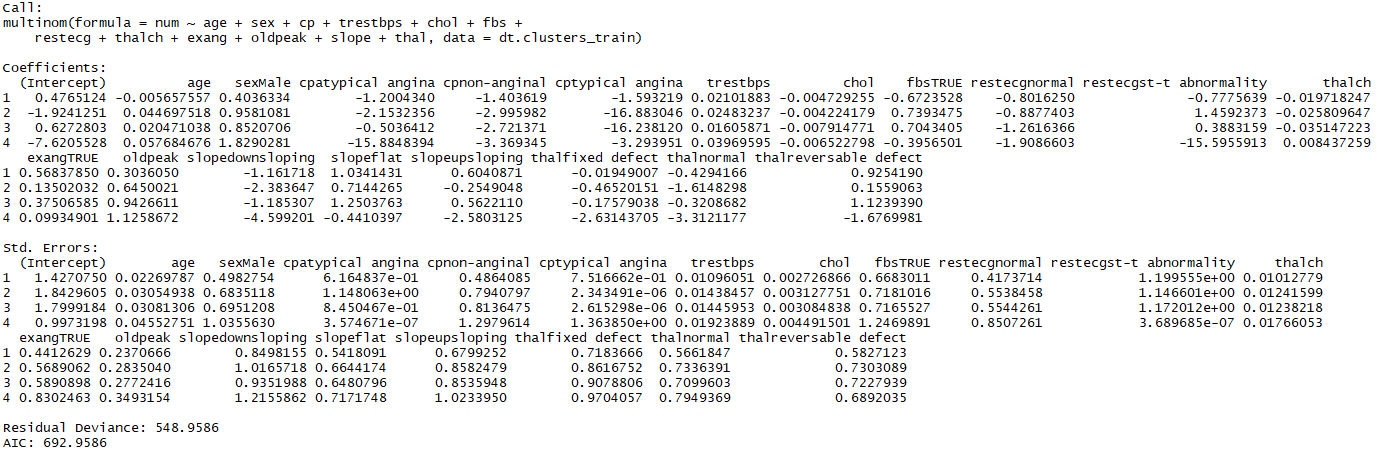
This clustering effect can be visualised using the cluster dendrogram, and the optimal number of clusters can be chosen by looking at the largest vertical distance between any 2 splits. According to the dendrogram, this seems to be at k = 4 or k = 8, and for similar reasons to our K-means clustering, we picked k = 8 for our hierarchical clustering and subsequently cut the tree at k = 8.

# 

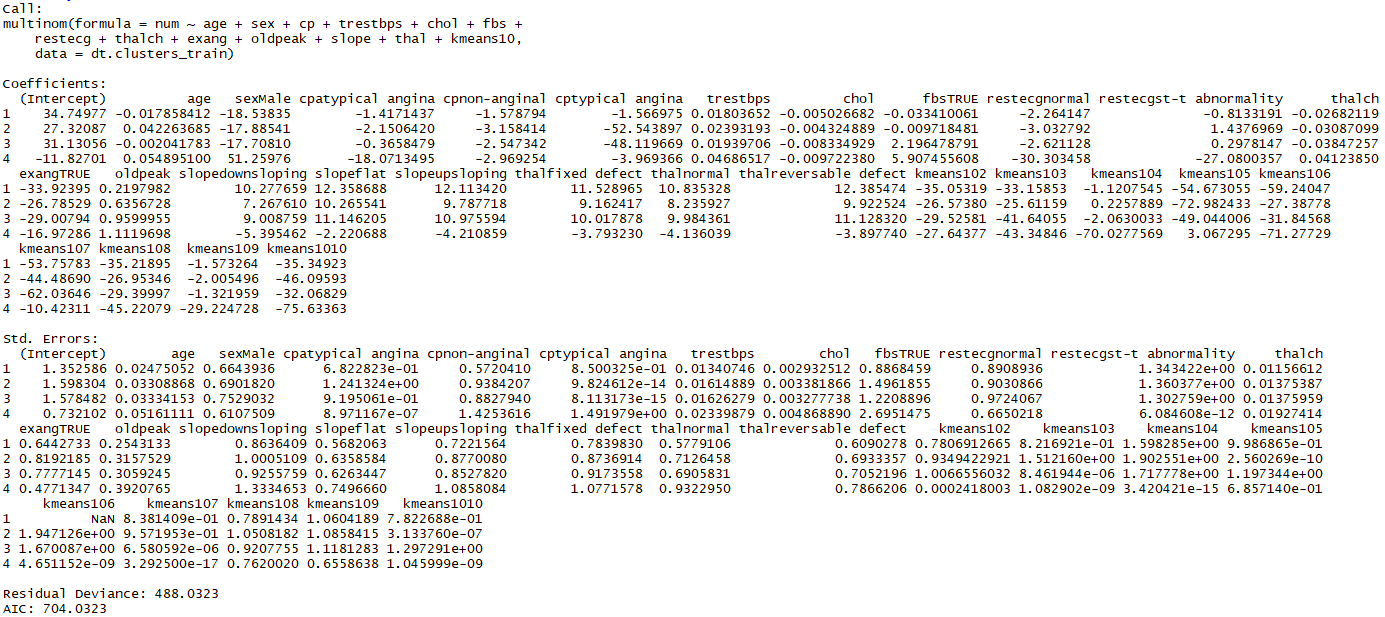
# Appendix E - Model Visualisation & Evaluation

## Logistic Regression

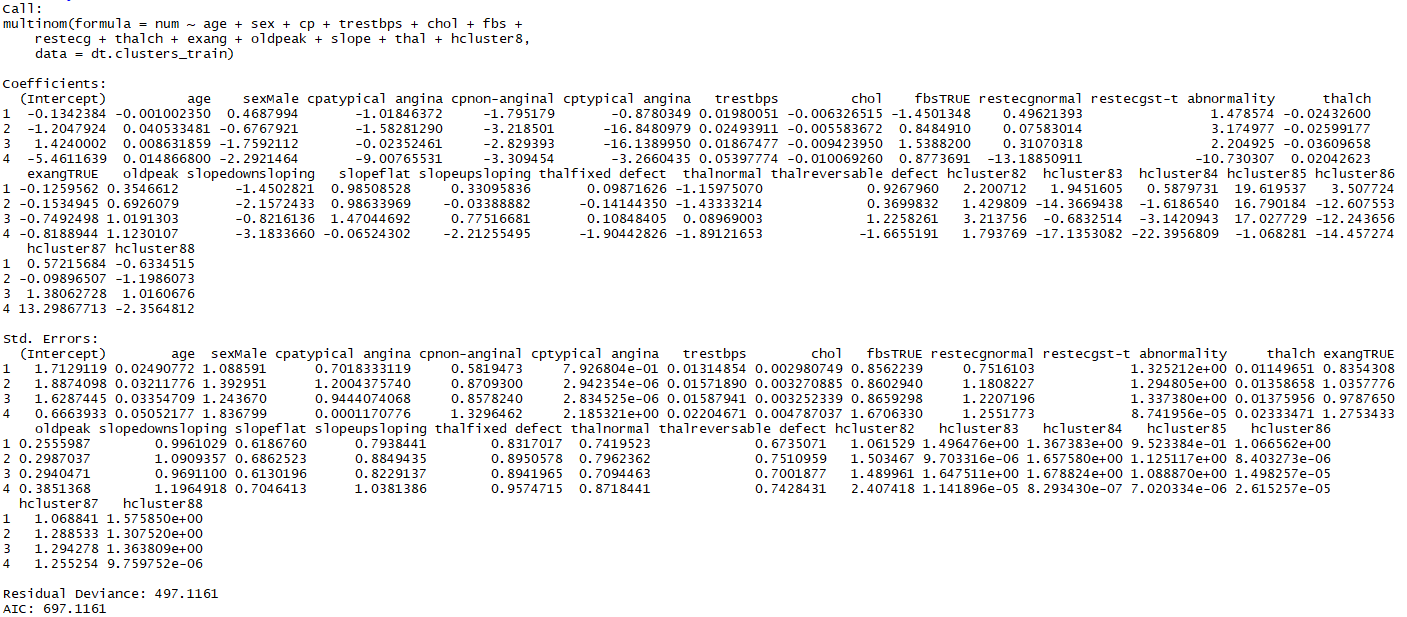
### Model with Original Dataset



### Model with K-Means Clusters

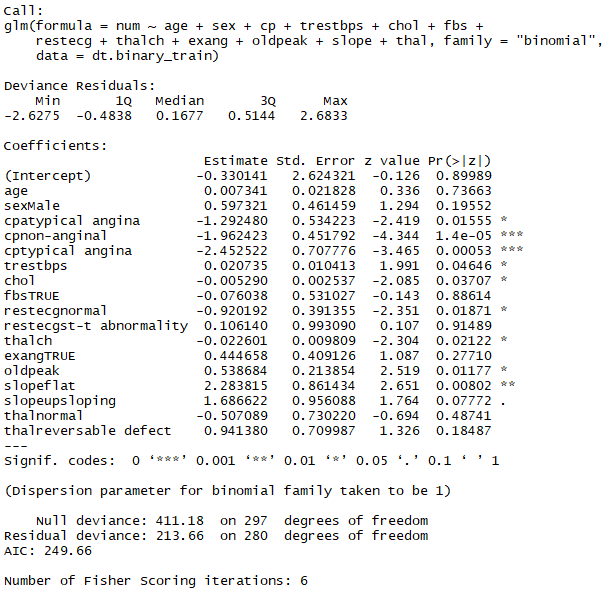


### Model with Hierarchical Clusters

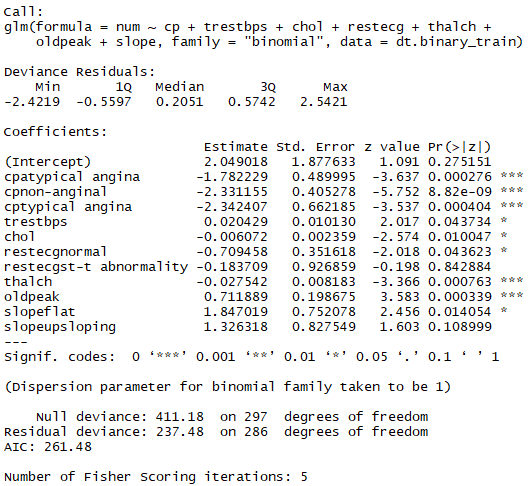


### Binary Model

Original binary model



Binary model after removing non significant variables



The variables age, sex, fbs, exang, and thal were not statistically significant in the original model. Therefore, they were removed and the new logistic regression model was generated.

### Binary Prediction

|  |  |
| --- | --- |
| Default model with original dataset | Model including K-means clusters |
| Model including hierarchical clusters | Model trained on binary dependent variable |

### Multi-Class Prediction

|  |  |
| --- | --- |
|  | Default model with original dataset |
|  | Model including K-means clusters |
|  | Model including hierarchical clusters |

## CART

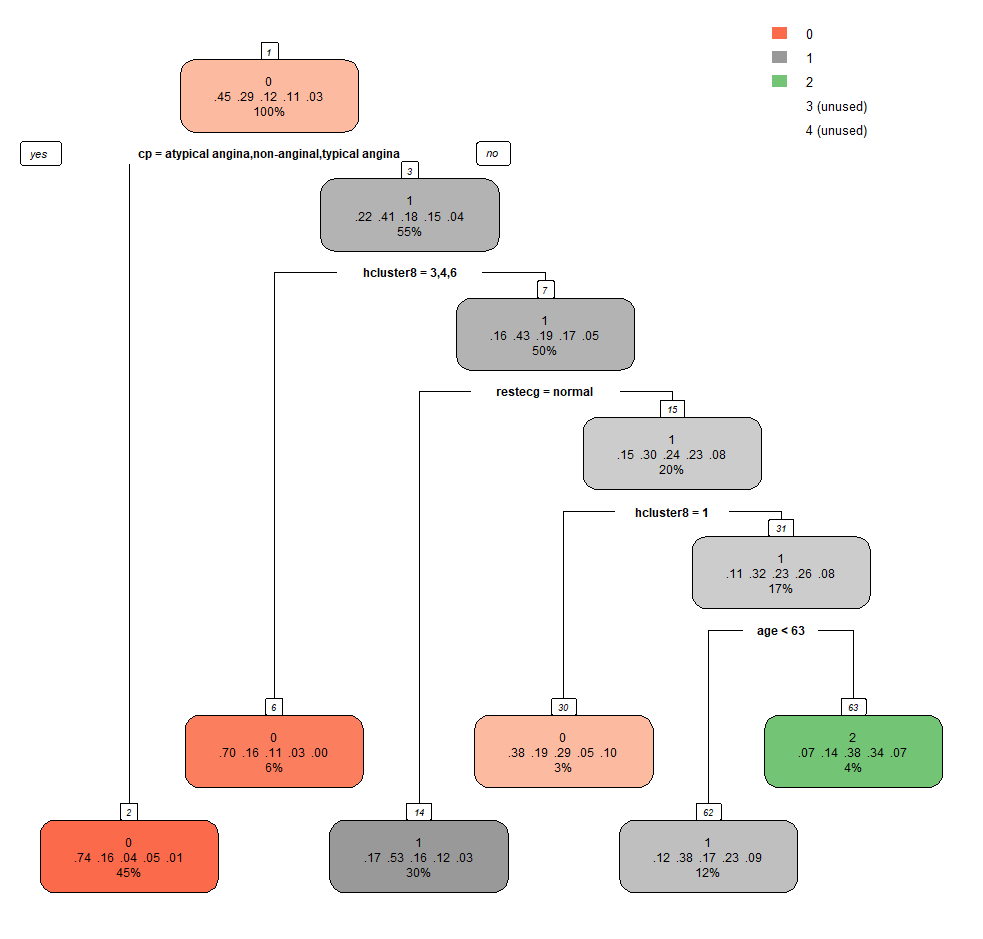
### Model with Original Dataset

|  |
| --- |
| node), split, n, loss, yval, (yprob)  \* denotes terminal node   1) root 660 363 0 (0.45 0.29 0.12 0.11 0.029)   2) cp=atypical angina,non-anginal,typical angina 296 78 0 (0.74 0.16 0.044 0.054 0.01) \*  3) cp=asymptomatic 364 216 1 (0.22 0.41 0.18 0.15 0.044)   6) exang=FALSE 153 98 0 (0.36 0.31 0.19 0.092 0.046)   12) chol>=42.5 104 53 0 (0.49 0.29 0.12 0.067 0.038)   24) thal=normal 36 8 0 (0.78 0.14 0.083 0 0) \*  25) thal=fixed defect,reversable defect 68 43 1 (0.34 0.37 0.13 0.1 0.059)   50) oldpeak< 0.65 43 22 0 (0.49 0.37 0.12 0.023 0)   100) age< 54.5 24 11 1 (0.46 0.54 0 0 0)   200) trestbps>=127.5 12 4 0 (0.67 0.33 0 0 0) \*  201) trestbps< 127.5 12 3 1 (0.25 0.75 0 0 0) \*  101) age>=54.5 19 9 0 (0.53 0.16 0.26 0.053 0) \*  51) oldpeak>=0.65 25 16 1 (0.08 0.36 0.16 0.24 0.16)   102) restecg=normal 13 5 1 (0 0.62 0.15 0.23 0) \*  103) restecg=lv hypertrophy,st-t abnormality 12 8 4 (0.17 0.083 0.17 0.25 0.33) \*  13) chol< 42.5 49 31 1 (0.082 0.37 0.35 0.14 0.061)   26) thalch>=123 26 13 2 (0.038 0.38 0.5 0.077 0) \*  27) thalch< 123 23 15 1 (0.13 0.35 0.17 0.22 0.13) \*  7) exang=TRUE 211 111 1 (0.11 0.47 0.18 0.19 0.043)   14) age< 62.5 169 78 1 (0.12 0.54 0.16 0.14 0.036) \*  15) age>=62.5 42 25 3 (0.071 0.21 0.24 0.4 0.071)   30) trestbps< 121 8 3 1 (0 0.62 0.12 0.25 0) \*  31) trestbps>=121 34 19 3 (0.088 0.12 0.26 0.44 0.088) \* |

### Model with K-Means Clusters

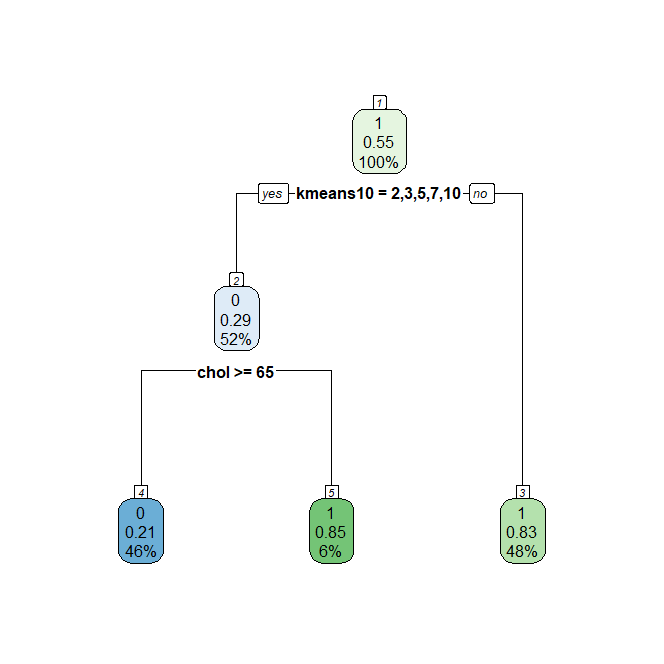
|  |
| --- |
| node), split, n, loss, yval, (yprob)  \* denotes terminal node   1) root 660 363 0 (0.45 0.29 0.12 0.11 0.029)   2) kmeans10=2,3,5,7,10 344 100 0 (0.71 0.18 0.058 0.035 0.015)   4) chol>=42.5 297 60 0 (0.8 0.15 0.02 0.02 0.0067) \*  5) chol< 42.5 47 30 1 (0.15 0.36 0.3 0.13 0.064) \*  3) kmeans10=1,4,6,8,9 316 185 1 (0.17 0.41 0.19 0.19 0.044)   6) cp=atypical angina,non-anginal,typical angina 73 43 0 (0.41 0.29 0.11 0.15 0.041)   12) thalch>=135.5 35 13 0 (0.63 0.23 0.029 0.057 0.057) \*  13) thalch< 135.5 38 25 1 (0.21 0.34 0.18 0.24 0.026)   26) age< 60 23 13 1 (0.22 0.43 0.26 0.043 0.043) \*  27) age>=60 15 7 3 (0.2 0.2 0.067 0.53 0) \*  7) cp=asymptomatic 243 133 1 (0.095 0.45 0.21 0.2 0.045)   14) kmeans10=1,4 138 59 1 (0.08 0.57 0.14 0.17 0.036)   28) age< 62.5 113 40 1 (0.08 0.65 0.14 0.11 0.027) \*  29) age>=62.5 25 14 3 (0.08 0.24 0.16 0.44 0.08) \*  15) kmeans10=6,8,9 105 74 1 (0.11 0.3 0.3 0.24 0.057)   30) chol>=206.5 59 38 1 (0.17 0.36 0.24 0.17 0.068) \*  31) chol< 206.5 46 29 2 (0.043 0.22 0.37 0.33 0.043)   62) thal=reversable defect 39 22 2 (0.051 0.23 0.44 0.28 0) \*  63) thal=fixed defect,normal 7 3 3 (0 0.14 0 0.57 0.29) \* |

### Model with Hierarchical Clusters



|  |
| --- |
| node), split, n, loss, yval, (yprob)  \* denotes terminal node   1) root 660 363 0 (0.45 0.29 0.12 0.11 0.029)   2) cp=atypical angina,non-anginal,typical angina 296 78 0 (0.74 0.16 0.044 0.054 0.01) \*  3) cp=asymptomatic 364 216 1 (0.22 0.41 0.18 0.15 0.044)   6) hcluster8=3,4,6 37 11 0 (0.7 0.16 0.11 0.027 0) \*  7) hcluster8=1,2,5,7,8 327 185 1 (0.16 0.43 0.19 0.17 0.049)   14) restecg=normal 196 93 1 (0.17 0.53 0.16 0.12 0.026) \*  15) restecg=lv hypertrophy,st-t abnormality 131 92 1 (0.15 0.3 0.24 0.23 0.084)   30) hcluster8=1 21 13 0 (0.38 0.19 0.29 0.048 0.095) \*  31) hcluster8=2,5,8 110 75 1 (0.11 0.32 0.23 0.26 0.082)   62) age< 62.5 81 50 1 (0.12 0.38 0.17 0.23 0.086) \*  63) age>=62.5 29 18 2 (0.069 0.14 0.38 0.34 0.069) \* |

### Binary Model



|  |
| --- |
| node), split, n, loss, yval, (yprob)  \* denotes terminal node  1) root 704 317 1 (0.4502841 0.5497159)   2) kmeans10=2,3,5,7,10 363 104 0 (0.7134986 0.2865014)   4) chol>=64.5 322 69 0 (0.7857143 0.2142857) \*  5) chol< 64.5 41 6 1 (0.1463415 0.8536585) \*  3) kmeans10=1,4,6,8,9 341 58 1 (0.1700880 0.8299120) \* |

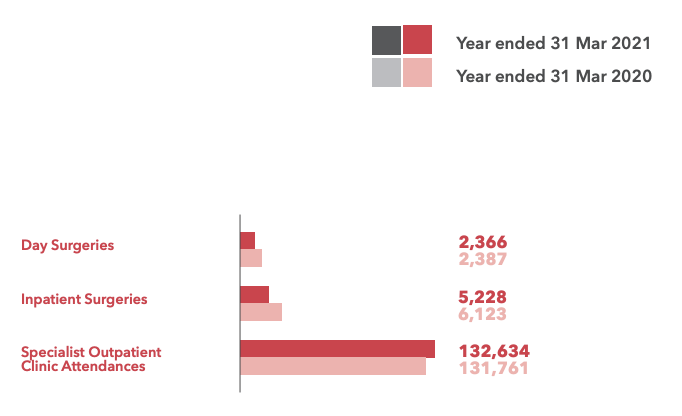
### Binary Prediction

|  |  |
| --- | --- |
| Default model with original dataset | Model including K-means clusters |
| Model including hierarchical clusters | Model trained on binary dependent variable |

### Multi-Class Prediction

|  |  |
| --- | --- |
|  | Default model with original dataset |
|  | Model including K-means clusters |
|  | Model including hierarchical clusters |

# Appendix F - Estimation on reduction of number of outpatients



|  |
| --- |
| Mar 2021 - 132634 outpatients  Assuming a 70% over referral rate (Lai, 2021) - 30% of patients in actual need of specialist care  30% of 132634 = 39790.2  Using our best model, CART trained on the binary dependent variable, we have a false positive rate of 16.7% - 83.3% of patients in actual need of specialist care.  Therefore, 39790.2 represents 83.3%, total number of outpatients referred using our model = patients referrals  reduction in patients referrals  Total referrals reduced = referrals  Total patient cost saving = |