Heart Disease Prediction

MATH 2319 Machine Learning Applied Project Phase I

Charles Galea (s3688570)

8 April 2018

Contents

1 Introduction

Heart disease is currently the leading cause of death across the globe. It is anticipated that the development of computation methods that can predict the presence of heart disease will significantly reduce heart disease caused mortalities while early detection could lead to substantial reduction in health care costs. Traditional statistical methods draw inferences from a limited number of variables obtained from experiments performed under controlled conditions. In contrast, Machine Learning methods can use a large number of often complex variables obtained from a variety of medical data banks to predict whether a patient has heart disease. Cardiovascular medicine generates a plethora of biomedical, clinical and operational data as a part of patient health care delivery, making this field ideal for the development and use of computational methods for predicting a patient has cardiac disease. Recent efforts to develop computational models capable of analysing and predicting whether a person has heart disease have shown great promise[1].

The objective of this project was to build classifiers to predict whether a person has cardiovascular disease. The data sets were obtained from the Cleveland Clinical Foundation, the Hungarian Institute of Cardiology (Budapest), the V.A Medical Center (Long Beach CA) and University Hospital Zurich (Switzerland). The overall database consists of medical test results and heart disease diagnoses for 920 individuals. The project was divided into two phases. Phase I focused on data pre-processing and exploration as covered in this report. Model building will be presented in Phase II of this project. Section 3 covers data pre-processing while in section 4 we explore each attribute and examine their inter-relationships. The results are summarized in the final section.

2 Data Set

Datasets (cleveland.data,[2] hungarian.data,[3] switzerland.data[4] and long-beach-va.data,[5] heart-disease.names) were obtained from the UCI Machine Learning Repository. The heart-disease.names file contains the details of attributes and variables. Each dataset contained 76 attributes but only 14 (including the target feature) were used in these analyses. The ‘goal’ field referred to the presence of heart disease in patients and was comprised of an integer value for 0 (no presence) to 4 (cardiac disease present). Names and social security numbers were removed and replaced with dummy values by the database administrators. In this study classifiers were built using the combined datasets and their performance was evaluated using cross-validation.

2.1 Target Feature

The response feature was ‘goal’ which is given as:

Diagnosis of heart disease (angiographic disease status). Designated as diameter narrowing in any major vessel.

$$\text{Diameter Narrowing} = \begin{cases} > 50\% & \text{Disease is Present} \\ \leq 50\% & \text{No Disease} \end{cases} \tag{1}$$

[1] Shameer K, Johnson KW, et al. Heart. 1-9, 2018.

[2] Detrano R, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation, CA, USA

[3] Janosi A, Hungarian Institute of Cardiology. Budapest, Hungary

[4] Steinbrunn W, University Hospital, Zurich, Switzerland

[5] Detrano R, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation, CA, USA

2.2 Descriptive Features

The variable description is produced here from the heart-disease.names file:

- Age: age in years
- Gender: gender (1 = male; 0 = female)
- Cp: chest pain type
  - Value 1: typical angina
  - Value 2: atypical angina
  - Value 3: non-anginal pain
  - Value 4: asymptomatic
- Trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- Chol: serum cholesterol in mg/dl
- Fbs: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- Restecg: resting electrocardiographic results
  - Value 0: normal
  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 005 mV)
  - Value 2: showing probable or definite left ventricular hypertropy by Estes criteria
- Thalach: maximum heart rate achieved in beats per minute (bpm)
- Exang: exercise induced angina (1 = yes; 0 = no)
- Oldpeak: ST depression induced by exercise relative to rest
- Slope: the slope of the peak exercise ST segment
  - Value 1: upsloping
  - Value 2: flat
  - Value 3: down-sloping
- Ca: number of major vessels (0-3) colored by fluoroscopy
- Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

The target feature has two classes and hence it is a binary classification problem. To reiterate, the goal is to predict **whether a person has heart disease**.

3 Data Pre-processing

3.1 Preliminaries

In this project, we used the following `R` packages.

```r
library(knitr)
library(mlr)
library(tidyverse)
library(GGally)
library(cowplot)
library(dplyr)
library(tidyr)
library(readr)
library(magrittr)
library(moments)
```

The combined datasets were read in treating the string values as characters. String columns were subsequently converted to factors (categorical) after data processing. For naming consistency with the data dictionary, we purposely skipped the headers and manually renamed the columns.

```
cleve <- read.csv('processed_cleveland2.csv', na = "?", stringsAsFactors = FALSE, header = FALSE)
hung  <- read.csv('processed_hungarian2.csv', na = "?", stringsAsFactors = FALSE, header = FALSE)
swiss  <- read.csv('processed_switzerland2.csv', na = "?", stringsAsFactors = FALSE, header = FALSE)
va  <- read.csv('processed_va2.csv', na = "?", stringsAsFactors = FALSE, header = FALSE)
full <- rbind(cleve, hung, swiss, va)

names(full) <- c('Age', 'Gender', 'CP', 'Trestbps', 'Chol', 'FBS', 'RestECG',
                 'Thalach', 'Exang', 'Oldpeak', 'Slope', 'CA', 'Thal', 'Goal')
```

3.2 Data Cleaning and Transformation

With `str` and `summarizeColumns` (Table 1), we noticed the following anomalies:

- The target feature, `Goal` had a cardinality of 5, which should be 2 since `Goal` was desiganted as the binary **target** feature.
- Ten of the 14 features contained missing values. Notably, the features Slope, CA and Thal had 309, 611 and 486 missing values, respectively.
- Trestbps (resting blood pressure) and Chol (serum cholestrol) contained several data entries with values of 0 which are not possible for these diagnostic tests.

```
str(full)
```

```
## 'data.frame':    920 obs. of  14 variables:
##  $ Age     : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ Gender  : chr  "male" "male" "male" "male" ...
##  $ CP      : chr  "typical angina" "asymptomatic" "asymptomatic" "non-anginal pain" ...
##  $ Trestbps: int  145 160 120 130 130 120 140 120 130 140 ...
##  $ Chol    : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ FBS     : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ RestECG : chr  "hypertropy" "hypertropy" "hypertropy" "normal" ...
##  $ Thalach : int  150 108 129 187 172 178 160 163 147 155 ...
##  $ Exang   : chr  "no" "yes" "yes" "no" ...
##  $ Oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ Slope   : chr  "downsloping" "flat" "flat" "downsloping" ...
##  $ CA      : int  0 3 2 0 0 0 2 0 1 0 ...
##  $ Thal    : chr  "fixed defect" "normal" "reversible defect" "normal" ...
##  $ Goal    : int  0 2 1 0 0 0 3 0 2 1 ...
```

```
summarizeColumns(full) %>% knitr::kable( caption =  'Feature Summary before Data Preprocessing')
```

Table 1: Feature Summary before Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
Age	integer	0	53.5108696	9.4246852	54.0	9.6369	28.0	77.0	0
Gender	character	0	NA	0.2108696	NA	NA	194.0	726.0	2
CP	character	0	NA	0.5945652	NA	NA	46.0	373.0	6
Trestbps	integer	59	132.1324042	19.0660695	130.0	14.8260	0.0	200.0	0
Chol	integer	30	199.1303371	110.7808104	223.0	68.1996	0.0	603.0	0
FBS	logical	90	NA	NA	NA	NA	138.0	692.0	2
RestECG	character	2	NA	NA	NA	NA	179.0	551.0	3
Thalach	integer	55	137.5456647	25.9262765	140.0	29.6520	60.0	202.0	0
Exang	character	55	NA	NA	NA	NA	337.0	528.0	2
Oldpeak	numeric	62	0.8787879	1.0912262	0.5	0.7413	-2.6	6.2	0
Slope	character	309	NA	NA	NA	NA	63.0	345.0	3

name	type	na	mean	disp	median	mad	min	max	nlevs
CA	integer	611	0.6763754	0.9356530	0.0	0.0000	0.0	3.0	0
Thal	character	486	NA	NA	NA	NA	46.0	196.0	3
Goal	integer	0	0.9956522	1.1426934	1.0	1.4826	0.0	4.0	0

Firstly, we removed the excessive white spaces for all character features.

```
full[, sapply(full, is.character)] <- sapply(full[, sapply(full, is.character)], trimws)
```

Secondly, a number of spelling error were detected and corrected in the CP feature of the dataset. Thirdly, several rows containing the variable ‘asymptomatic’ in multiple columns were removed.

```
sapply(full[sapply(full, is.character)], table)
```

```
## $Gender
## 
## female   male 
##    194    726 
## 
## $CP
## 
##     asymptomatic  atypical angina     aysmptomatic  non-angina pain 
##              373              174              123              101 
## non-anginal pain   typical angina 
##              103               46 
## 
## $RestECG
## 
##            hypertropy                normal ST-T wave abnormality 
##                   188                   551                   179 
## 
## $Exang
## 
##  no yes 
## 528 337 
## 
## $Slope
## 
## downsloping        flat   upsloping 
##          63         345         203 
## 
## $Thal
## 
##      fixed defect            normal reversible defect 
##                46               196               192
```

```
full$CP[full$CP == "non-angina pain"] <- "non-anginal pain"
full$CP[full$CP == "aysmptomatic"] <- "asymptomatic"
```

Fourthly, in the Goal column clinicians had graded patients as either having no heart disease (value of 0) or displaying various degrees of heart disease (values 1 to 4). We chose to group the data into 2 categories of ‘no heart disease’ (value of 0) and ‘displaying heart disease’ (value of 1) so it became binary.

It was noted that a higher proportion of people were diagnosed with heart disease (Table2). Therefore, we may require additional parameter-tuning in building models to cater for such an unbalanced class.

```
full$Goal[full$Goal == 2] <- 1
full$Goal[full$Goal == 3] <- 1
full$Goal[full$Goal == 4] <- 1
full$Disease <- factor(full$Goal, labels = c("No Disease", "Heart Disease"))
table(full$Goal) %>% kable(caption = 'Degree of Heart Disease')
```

Table 2: Degree of Heart Disease

Var1	Freq
0	411
1	509

We were concerned about the large number of missing values for the Slope (slope of the peak exercise ST segment), CA (number of major vessels) and Thal (β-Thalassemia cardiomyopathy) features. The number of missing datapoints represents 33% of the Slope, 66% of the CA and 53% of the Thal data. Since the percentage of missing data or the CA feature was greater than 60%, and was unlikely to add significant information, we decided to remove this feature from the dataset.

```
full <- full[,-12]
```

Inspection of the dataset showed that features associated with the Exercise Thread Mill Test (i.e. Trestbps, Thalach, Exang and Oldpeak) were missing for some patients indicating they did not undergo this test. Rows containing these missing values were removed since they represent less than 6% of the overall dataset (i.e. 55 out of 920 instances).

```
full <- full[!is.na(full$Trestbps),]
```

In addition, several rows were missing data for Chol, Thal and Slope. These rows were also removed as they comprise only 3% of the total dataset. Several additional rows of the Thal (β-Thalassemia) and Slope (slope of the peak exercise ST segment) feature columns that were missing values were replaced with 'unknown'. It was assumed that these patients were not tested for the inherited blood disorder β-Thalassemia. It also appeared that the slope of the peak exercise ST segment was not documented even though several of these patients were diagnosed with ST-T wave abnormalities from the electrocardiographic test.

```
full <- full[!is.na(full$Chol),]
full <- full[!is.na(full$RestECG),]
full <- full[!is.na(full$Oldpeak),]
full <- full[!is.na(full$FBS),]

full$Thal <- as.character(full$Thal)
full$Thal[is.na(full$Thal)] <- "unknown"

full$Slope <- as.character(full$Slope)
full$Slope[is.na(full$Slope)] <- "unknown"
```

We computed the level table for each character feature. The tables revealed:

- There were 649 males and only 185 females in the dataset.
- Only 37 patients exhibited typical angina (chest pain) symptoms even though 509 patients were diagnosed with cardiovascular disease.
- Only a small proportion of the dataset (i.e. 5%) exhibited typical angina symptoms while a significantly higher proportion displayed exercise induced angina (i.e. 61%).
- The majority (445 out of 740 patients) displayed normal electrocardiographic (ECG) results.

```r
sapply(full[sapply(full, is.character)], table)
```

```
## $Gender
##
## female   male
##    174    566
##
## $CP
##
##     asymptomatic  atypical angina non-anginal pain   typical angina
##              392              150              161               37
##
## $RestECG
##
##            hypertropy                normal ST-T wave abnormality
##                   175                   445                   120
##
## $Exang
##
##  no yes
## 444 296
##
## $Slope
##
## downsloping        flat     unknown   upsloping
##          48         310         209         173
##
## $Thal
##
##      fixed defect            normal reversible defect           unknown
##                39               187               174               340
```

Lastly, we converted all character features into factor.

```r
full[, sapply(full, is.character)] <- lapply( full[, sapply(full, is.character )], factor)
```

Table 3 presents the summary statistics after data preprocessing.

```r
summarizeColumns(full) %>% kable( caption = 'Feature Summary before Data Preprocessing' )
```

Table 3: Feature Summary before Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
Age	integer	0	53.0972973	9.4081267	54.0	10.3782	28	77.0	0
Gender	factor	0	NA	0.2351351	NA	NA	174	566.0	2
CP	factor	0	NA	0.4702703	NA	NA	37	392.0	4
Trestbps	integer	0	132.7540541	18.5812497	130.0	14.8260	0	200.0	0
Chol	integer	0	220.1364865	93.6145555	231.0	54.8562	0	603.0	0
FBS	logical	0	NA	0.1500000	NA	NA	111	629.0	2
RestECG	factor	0	NA	0.3986486	NA	NA	120	445.0	3
Thalach	integer	0	138.7445946	25.8460815	140.0	29.6520	60	202.0	0
Exang	factor	0	NA	0.4000000	NA	NA	296	444.0	2
Oldpeak	numeric	0	0.8943243	1.0871598	0.5	0.7413	-1	6.2	0
Slope	factor	0	NA	0.5810811	NA	NA	48	310.0	4
Thal	factor	0	NA	0.5405405	NA	NA	39	340.0	4

name	type	na	mean	disp	median	mad	min	max	nlevs
Goal	numeric	0	0.5175676	0.5000293	1.0	0.0000	0	1.0	0
Disease	factor	0	NA	0.4824324	NA	NA	357	383.0	2

4 Data Exploration

We explored the data for each feature individually and split them by the classes of target features. Then we proceeded to multivariate visualisation.

4.1 Univariate Visualisation

4.1.1 Numerical Features

4.1.1.1 Age

The median age for patients examined in these studies was 54 (Fig. 1 and Table. 4) with the youngest and oldest being 28 and 77, respectively. Overall, individuals exhibiting cardiac disease had a higher median age of 57 compared to the non-diseased cohort which had a median of 51 (Table 5). The distribution of ages for the cardiac disease cohort was slightly skewed toward higher ages (skewness = 0.081) while the opposite was observed for the non-disease group (skewness = -0.325) (Fig. 1 and Tables 4 and 5). Therefore, age would be a predictive feature.

```
## Warning: Ignoring unknown parameters: fill
```

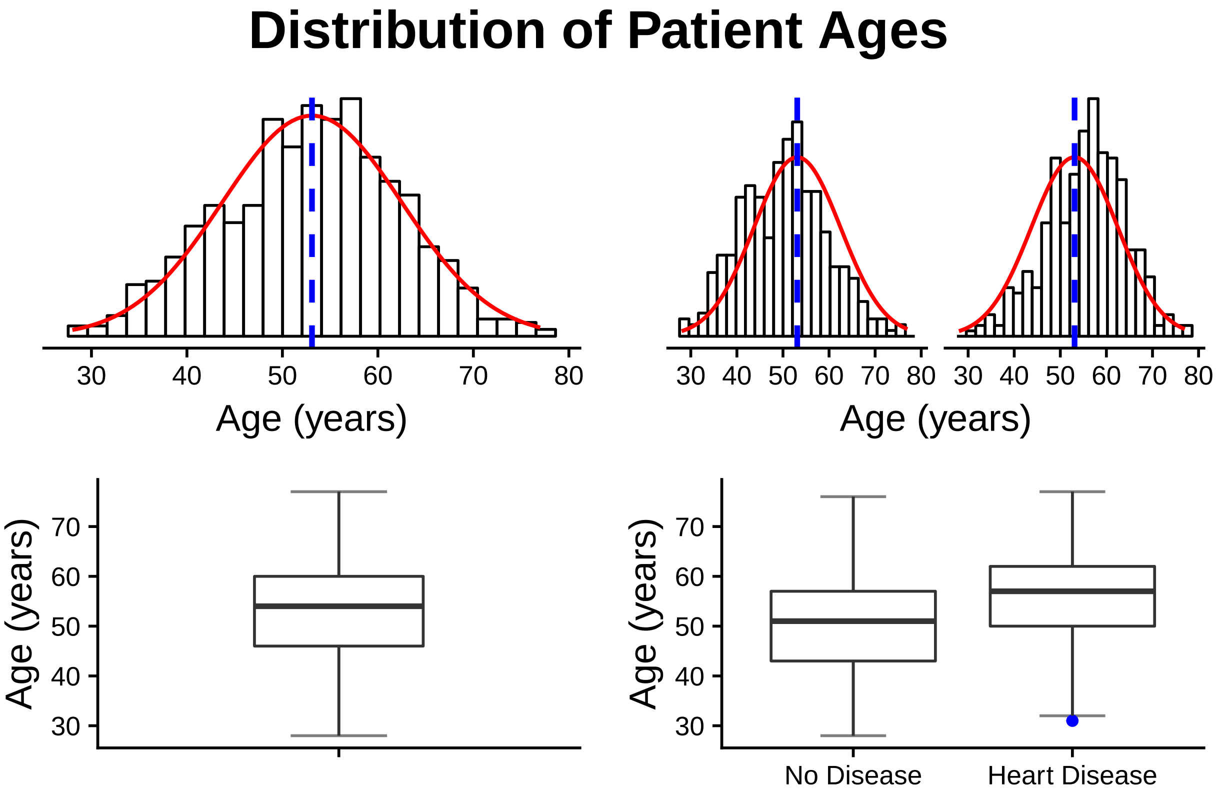

Figure 1. Distribution of patient ages. Histogram for all patients (top left) compared to those for diseased

and non-diseased patients (top right). The curve for the normal distribution (red line) and median value (blue dashed line) of all patients is overlayed on both figures and the median is shown as a blue dotted line. Boxplots for all patients (bottom left) compared to those for diseased and non-diseased patients (bottom right).

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

Table 4: Statistical parameters for patient age

Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
28	46	54	60	77	53.097	9.408	-0.161	740	-0.161

Table 5: Statistical parameters for ages of patients with and without cardiac disease

Disease	Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
No Disease	28	43	51	57	76	50.303	9.418	0.081	357	0
Heart Disease	31	50	57	62	77	55.702	8.630	-0.325	383	0

4.1.1.2 Resting Blood Pressure

The aggregated resting blood pressure for the entire cohort exhibited a median value of 130 which was similar to that for the diseased and non-diseased groups (i.e. 120 for both) (Fig. 2 and Table 6). The distribution of resting blood pressure values was slightly larger for the diseased compared to the non-diseased group (i.e. standard deviation of 18.7 and 16.6, respectively) (Fig. 2 and Tables 6 and 7). In addition, there was a slight difference in skewness between the diseased and non-diseased cohorts (skewness of 0.676 compared to 0.696 for diseased and non-diseased patients, respectively). Therefore, resting blood pressure was not a good predictive feature.

```
## Warning: Ignoring unknown parameters: fill
```

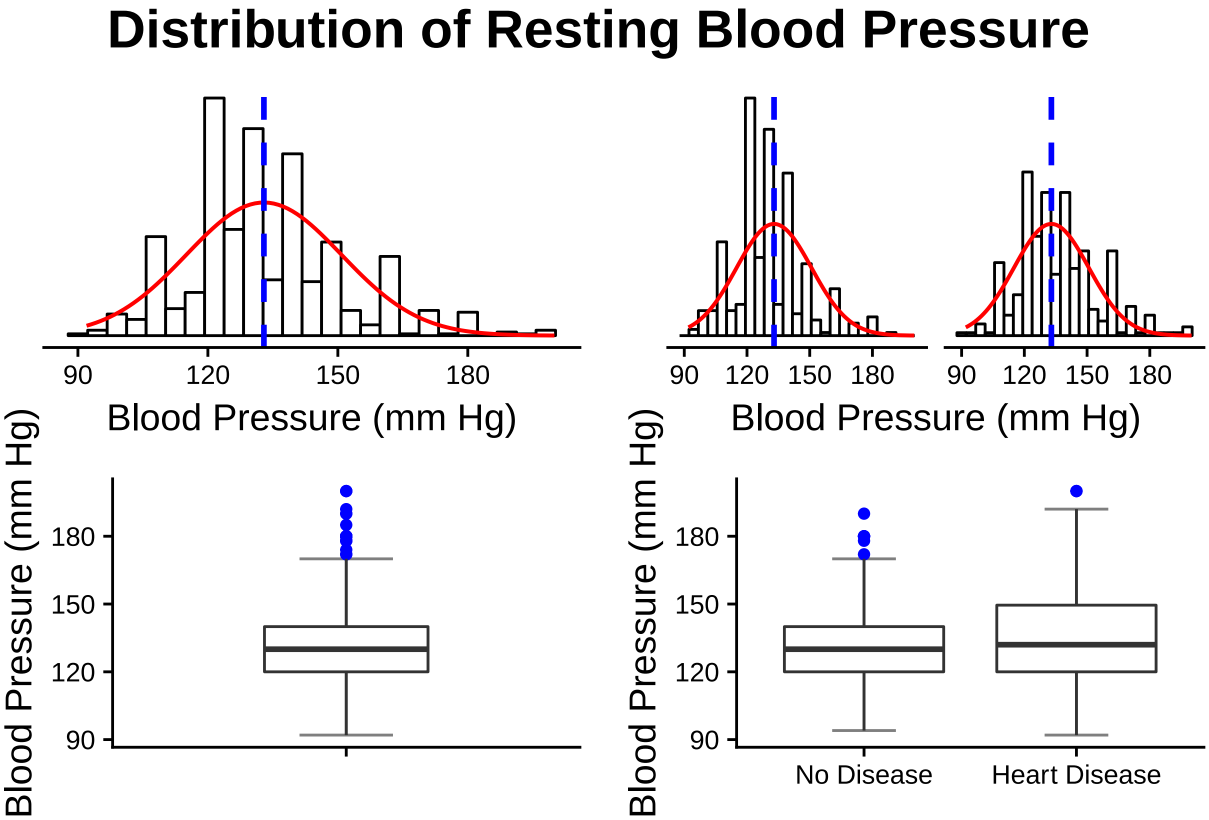

Figure 2. Distribution of patient resting blood pressure. Histogram for all patients (top left) compared to those for diseased and non-diseased patients (top right). The curve for the normal distribution (red line) and median value (blue dashed line) for all patients is overlayed on both figures. Boxplots for all patients (bottom left) compared to those for diseased and non-diseased patients (bottom right).

Table 6: Statistical parameters for patient Resting Blood Pressure

Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
92	120	130	140	200	132.934	17.939	0.718	739	0

Table 7: Statistical parameters for the Resting Blood Pressure of patients with and without cardiac disease

Disease	Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
No Disease	94	120	130	140.0	190	129.871	16.567	0.696	357	0
Heart Disease	92	120	132	149.5	200	135.796	18.706	0.676	382	0

4.1.1.3 Cholesterol Levels

Cholestrol levels for the non-disease cohort (median = 233 mg/dL) were lower compared to the diseased patients (median = 248 mg/dL) (Fig. 4 and Table 8). The histogram for the diseased patients also exhibited a slightly increased postive skewness (1.5 compared to 1.1) (Table 9). Both cohorts contained a number of individuals possessing very high cholestrol levels (outliers on the boxplots in Fig. 4) while there were more higher outliers in the diseased group.

```
## Warning: Ignoring unknown parameters: fill
```

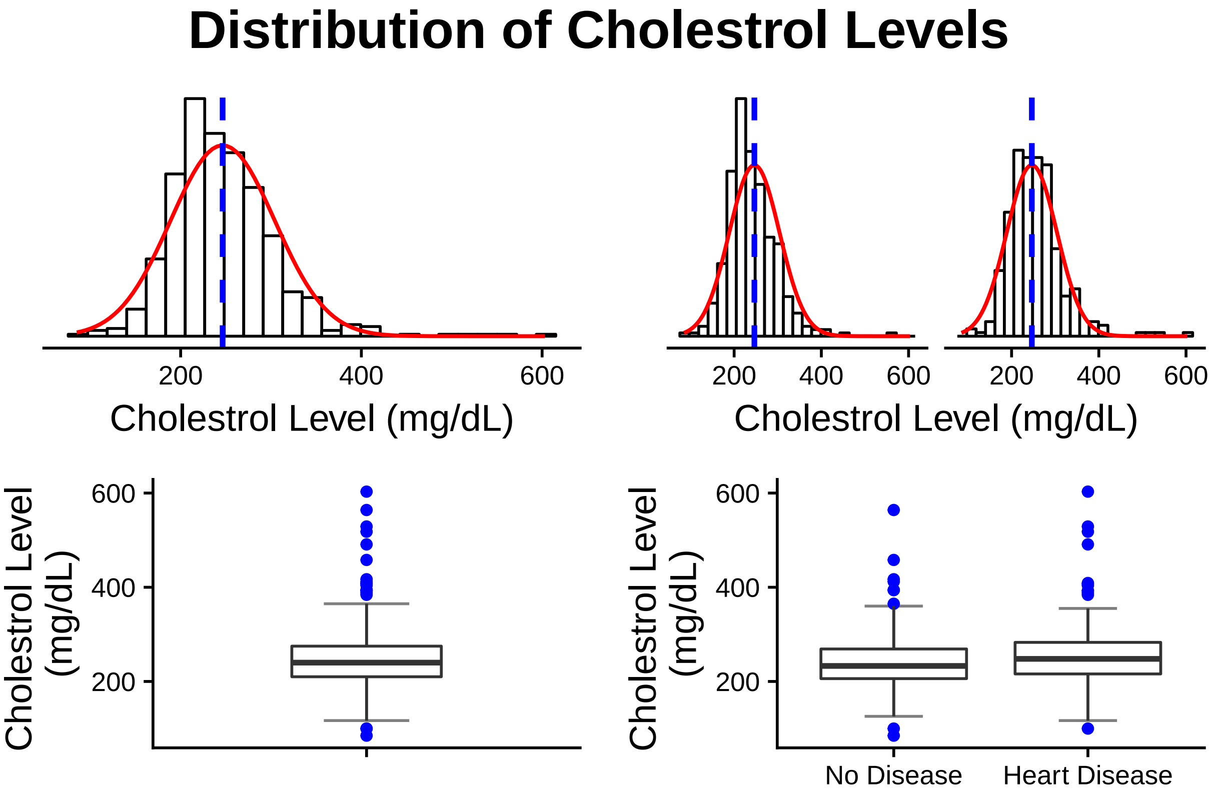

Figure 4. Distribution of patient cholestrol levels. Histogram for all patients (top left) compared to those for diseased and non-diseased patients (top right). The curve for the normal distribution (red line) and median value (blue dashed line) for all patients was overlayed on both figures. Boxplots for all patients (bottom left) compared to those for diseased and non-diseased patients (bottom right).

Table 8: Statistical parameters for patient cholestrol levels

Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
85	210	240	275	603	246.446	57.61	1.337	661	0

Table 9: Statistical parameters for the cholestrol levels of patients with and without cardiac disease

Disease	Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
No Disease	85	206	233	269	564	240.061	54.262	1.115	347	0
Heart Disease	100	216	248	283	603	253.503	60.402	1.487	314	0

4.1.1.4 Maximum Heart Rate

The histogram for the entire cohort exhibited a normal distribution (Fig. 5 and Table 10). Maximum heart rate was higher for the non-disease cohort (median = 135) compared to diseased patients (median = 112) (Fig. 5 and Table 11). However, the histogram for non-diseased patients was significantly more skewed toward older ages compared to the diseased group (i.e. skewness of -0.58 versus -0.026) (Table 11). It was anticipated that this feature should have high predictive power.

```
## Warning: Ignoring unknown parameters: fill
```

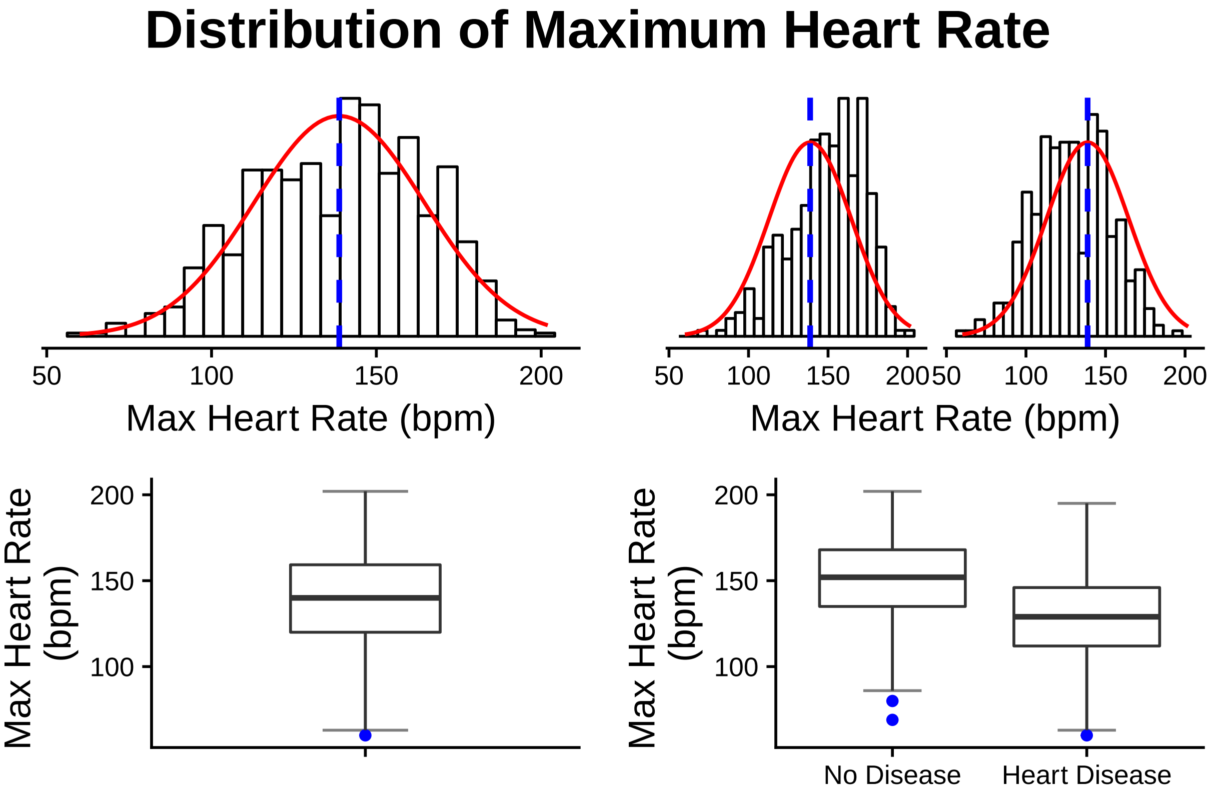

Figure 5. Distribution of patient maximum heart rates. Histogram for all patients (top left) compared to those for diseased and non-diseased patients (top right). The curve for the normal distribution (red line) and median value (blue dashed line) for all patients is overlayed on both figures. Boxplots for all patients (bottom left) compared to those for diseased and non-diseased patients (bottom right).

Table 10: Statistical parameters for patient Maximum Heart Rate

Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
60	120	140	159.25	202	138.745	25.846	-0.229	740	0

Table 11: Statistical parameters for the Maximum Heart Rate of patients with and without cardiac disease

Disease	Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
No Disease	69	135	152	168	202	149.291	23.644	-0.577	357	0
Heart Disease	60	112	129	146	195	128.914	23.885	-0.026	383	0

4.1.1.5 ST depression induced by exercise

The histogram for ST depression induced by exercise was heavily skewed due to the presence of a large number of low values (Fig. 6 and Tables 12 and 13). Removal of lower data values indicated that the histogram for the non-diseased cohort was significantly more skewed (skewness of 1.12 compared to 0.92) with differing median values (1.0 and 1.8, respectively) (Fig. 7 and Table 15). Suggesting that this feature should be highly predictive.

```
## Warning: Ignoring unknown parameters: fill
```

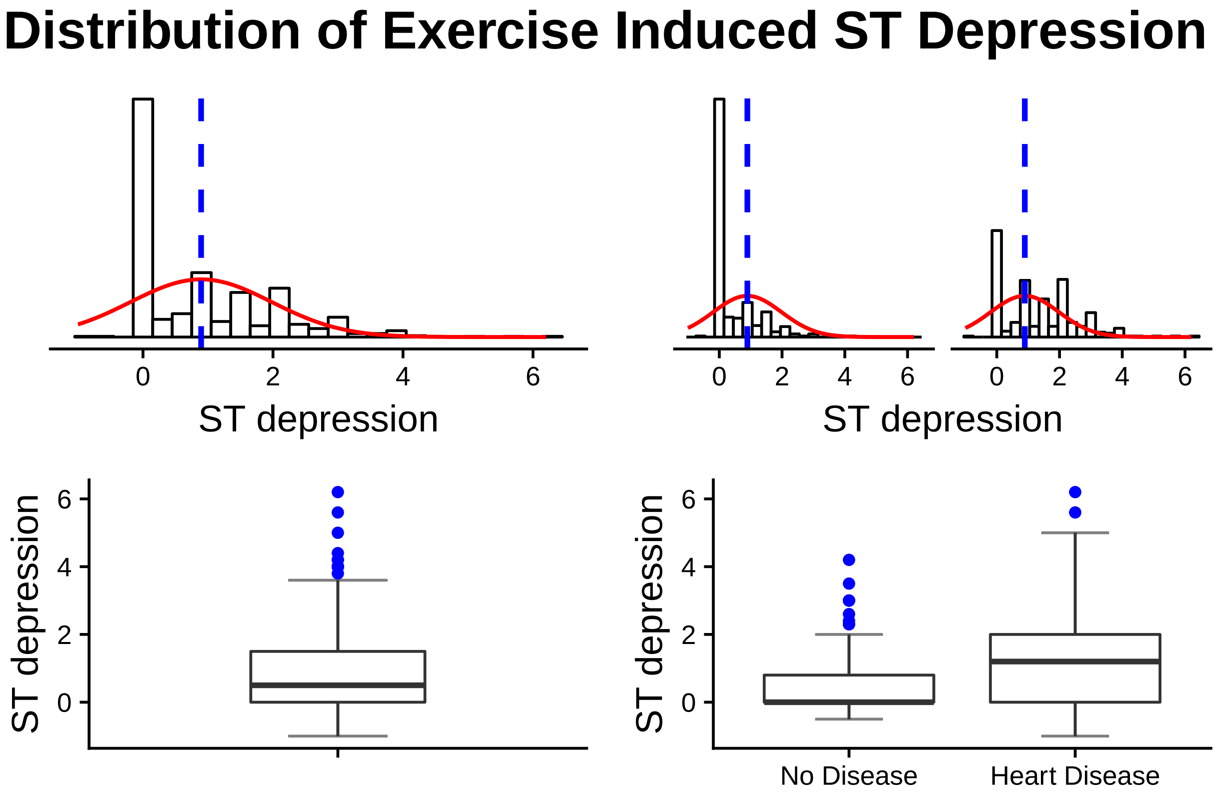

Figure 6. Distribution of patient ST depression induced by exercise. Histogram for all patients (top left) compared to those for diseased and non-diseased patients (top right). The curve for the normal distribution (red line) and median value (blue dashed line) for all patients were overlayed on both figures. Boxplots for all patients (bottom left) compared to those for diseased and non-diseased patients (bottom right).

Table 12: Statistical parameters for patient for ST Depression Induced by Exercise

Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
-1	0	0.5	1.5	6.2	0.894	1.087	1.204	740	0

Table 13: Statistical parameters for ST Depression Induced by Exercise for patients with and without cardiac disease

Disease	Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
No Disease	-0.5	0	0.0	0.8	4.2	0.425	0.712	1.879	357	0
Heart Disease	-1.0	0	1.2	2.0	6.2	1.332	1.190	0.701	383	0

```
## Warning: Ignoring unknown parameters: fill

## Warning: Removed 1 rows containing missing values (geom_bar).

## Warning: Removed 2 rows containing missing values (geom_bar).
```

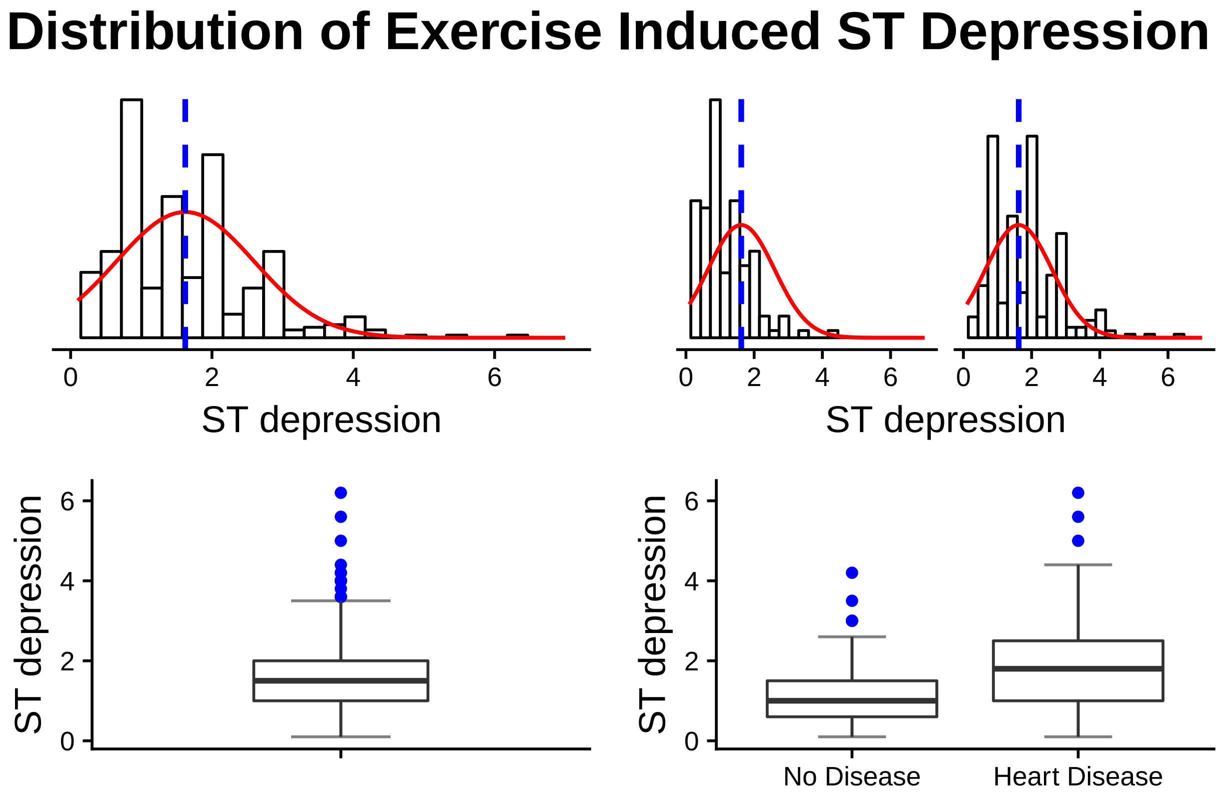

Figure 7. Distribution of patient ST depression induced by exercise after removing low values. Histogram for all patients (top left) compared to those for diseased and non-diseased patients (top right). The curve for the normal distribution (red line) and median (blue dashed line) for all patients is overlayed on both figures. Boxplots for all patients (bottom left) compared to those for diseased and non-diseased patients (bottom right).

Table 14: Statistical parameters for patient for ST Depression Induced by Exercise

Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
0.1	1	1.5	2	6.2	1.622	0.976	1.012	409	0

Table 15: Statistical parameters for ST Depression Induced by Exercise for patients with and without cardiac disease

Disease	Min	Q1	Median	Q3	Max	Mean	SD	SK	n	Missing
No Disease	0.1	0.6	1.0	1.5	4.2	1.145	0.731	1.119	133	0
Heart Disease	0.1	1.0	1.8	2.5	6.2	1.851	0.997	0.917	276	0

4.1.2 Categorical Features

For categorical features, we shall present the newly defined categorical features.

4.1.2.1 Gender

The cardiac disease cohort contained over 2-fold more males than females and a significantly higher proportion of males were diagnosed with cardiac disease compared to females (Fig. 8).

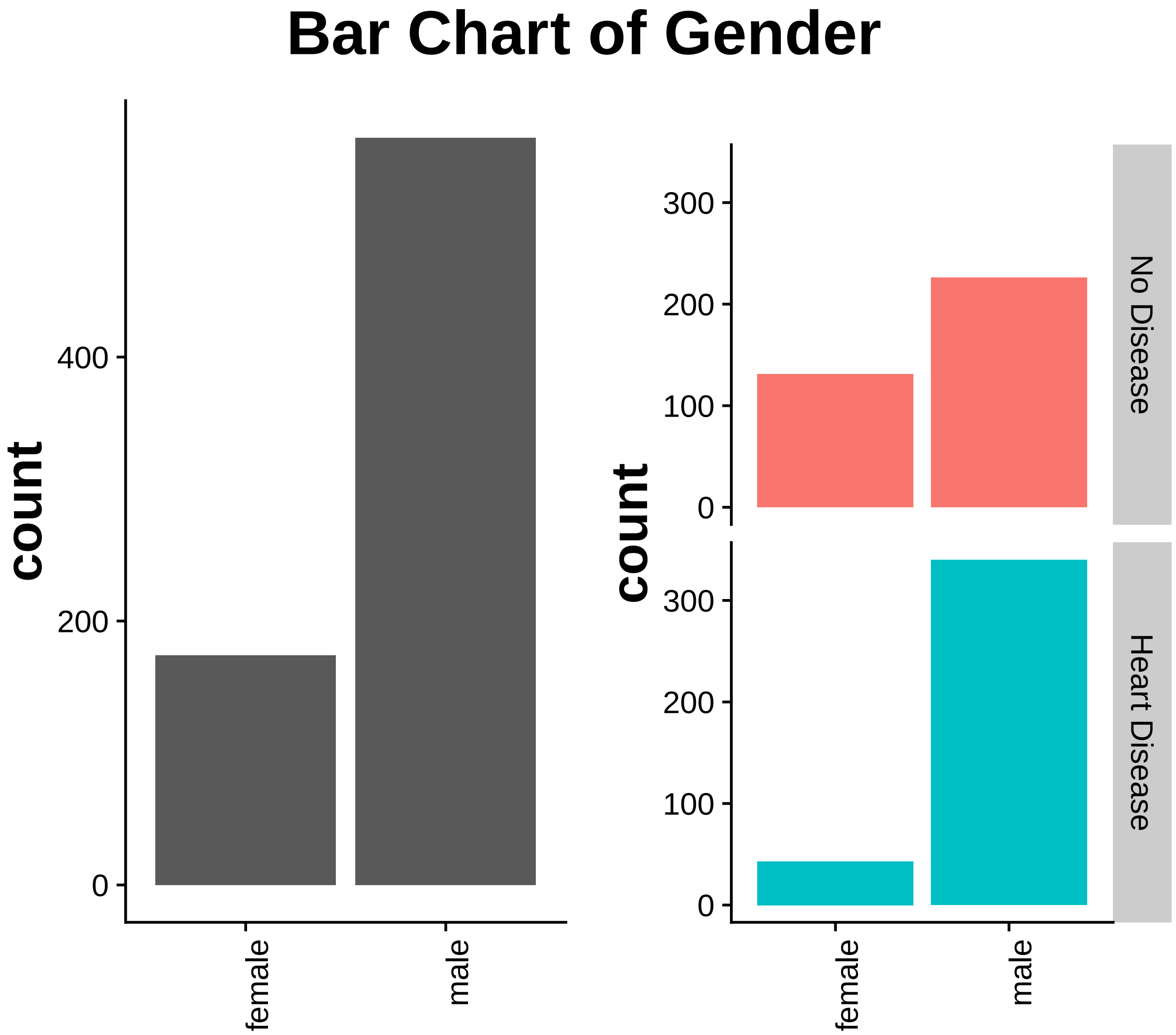

Figure 8. Patient gender. Bar chart of gender for all patients (left) compared to those for diseased and non-diseased patients (right).

4.1.2.2 Chest Pain Type

The majority of the cohort did not exhibit typical angina symptoms and there was little difference in the dataset when it was segregated by disease status (Fig. 9). Taken alone, the data suggest that angina pain was not a good predictor of the presence of disease. However, as we show later, this feature has strong predictive power when combined with the exercise induced angina feature.

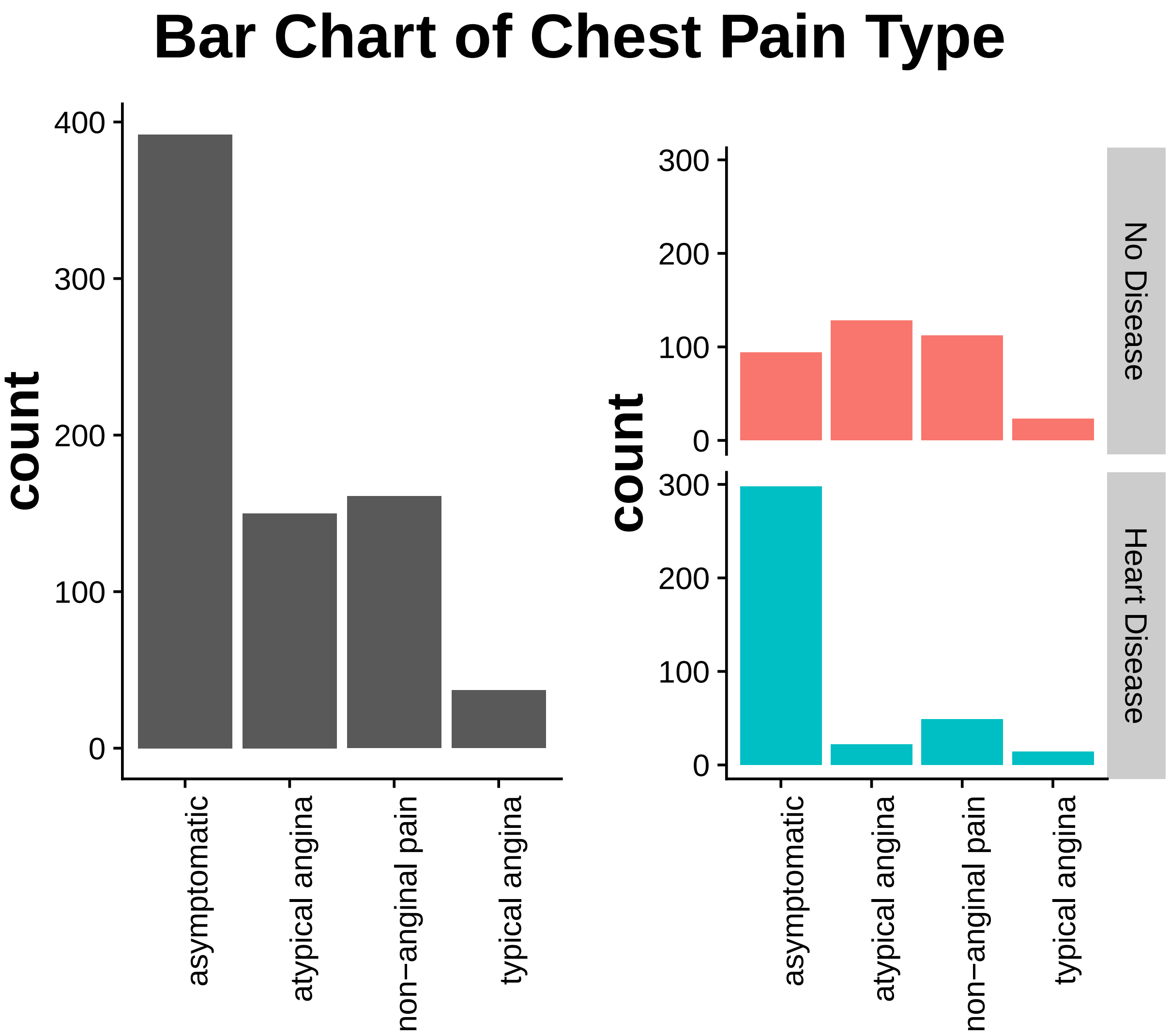

Figure 9. Chest pain type. Bar chart of chest pain type for all patients (left) compared to those for diseased and non-diseased patients (right).

4.1.2.3 Fasting Blood Sugar

Most individuals did not have fasting blood sugar levels greater than 120 mg/dL (Fig. 10). This did not change greatly when the data was divided based on the presence of disease although a slightly higher proportion of diseased patients exhibited higher levels of blood sugar.

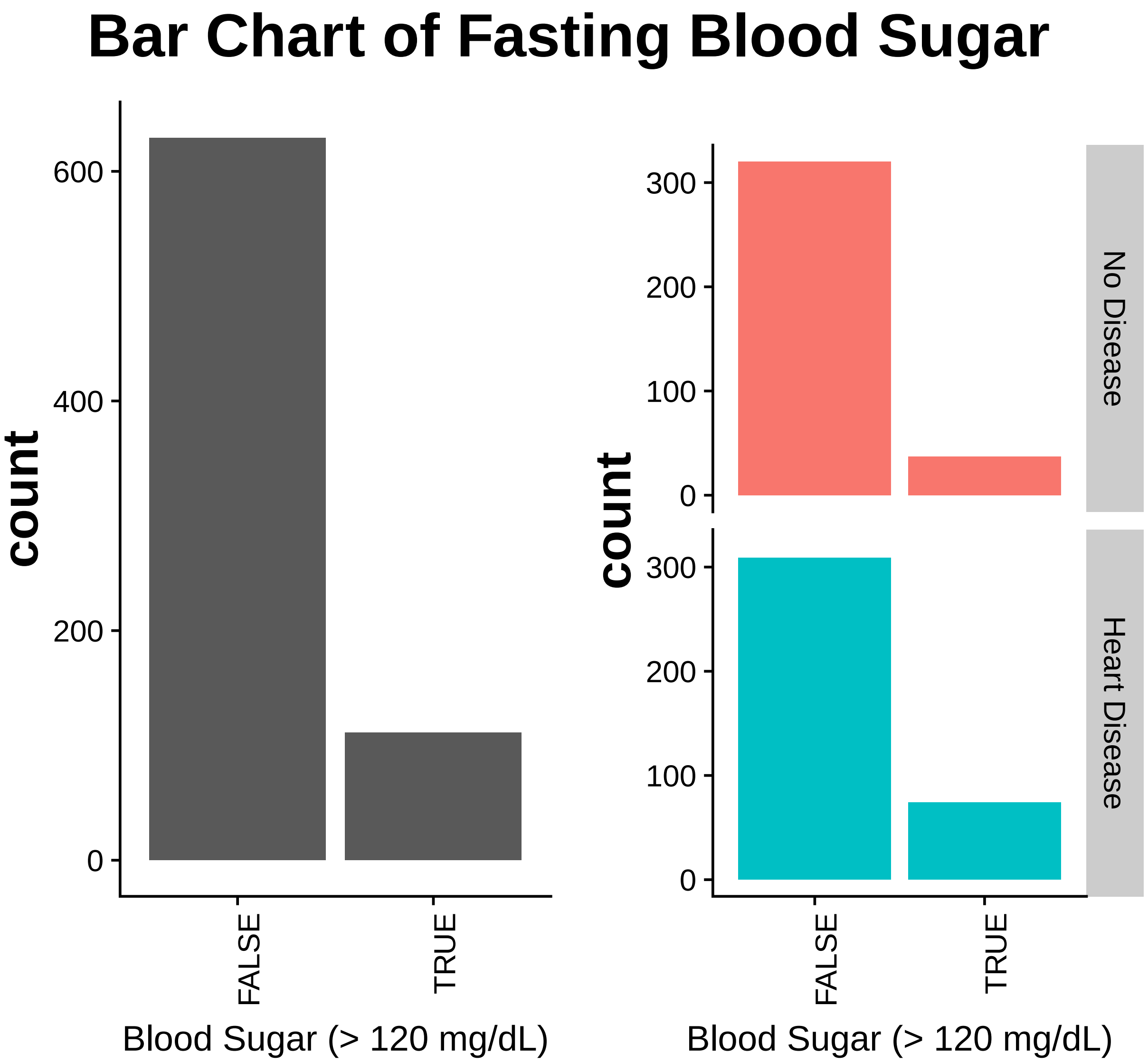

Figure 10. Fasting blood sugar. Bar chart of fasting blood sugar levels for all patients (left) compared to those for diseased and non-diseased patients (right).

4.1.2.4 Resting Electrocardiographic Results

Most patients exhibited normal resting electrocardiograhic results (Fig. 11). However, a higher proportion of diseased patients had abnormal ST wave patterns suggesting that this feature may contribute some predictive power.

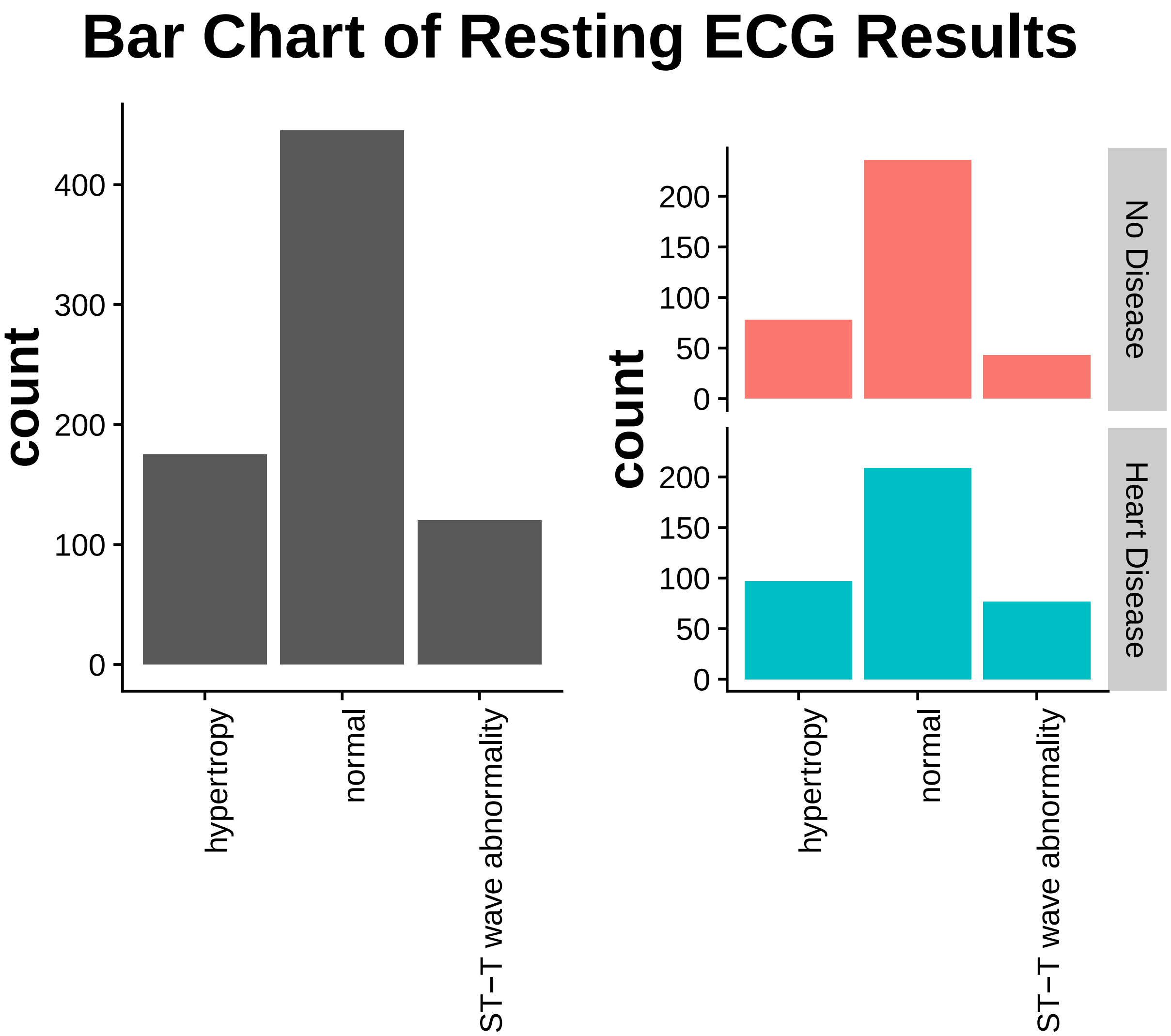

Figure 11. Resting ECG results. Bar chart of resting ECG results for all patients (left) compared to those for diseased and non-diseased patients (right).

4.1.2.5 Exercise Induced Angina

Significantly more patients in the diseased cohort displayed exercise induced angina (Fig. 12). This feature should be strongly predictive.

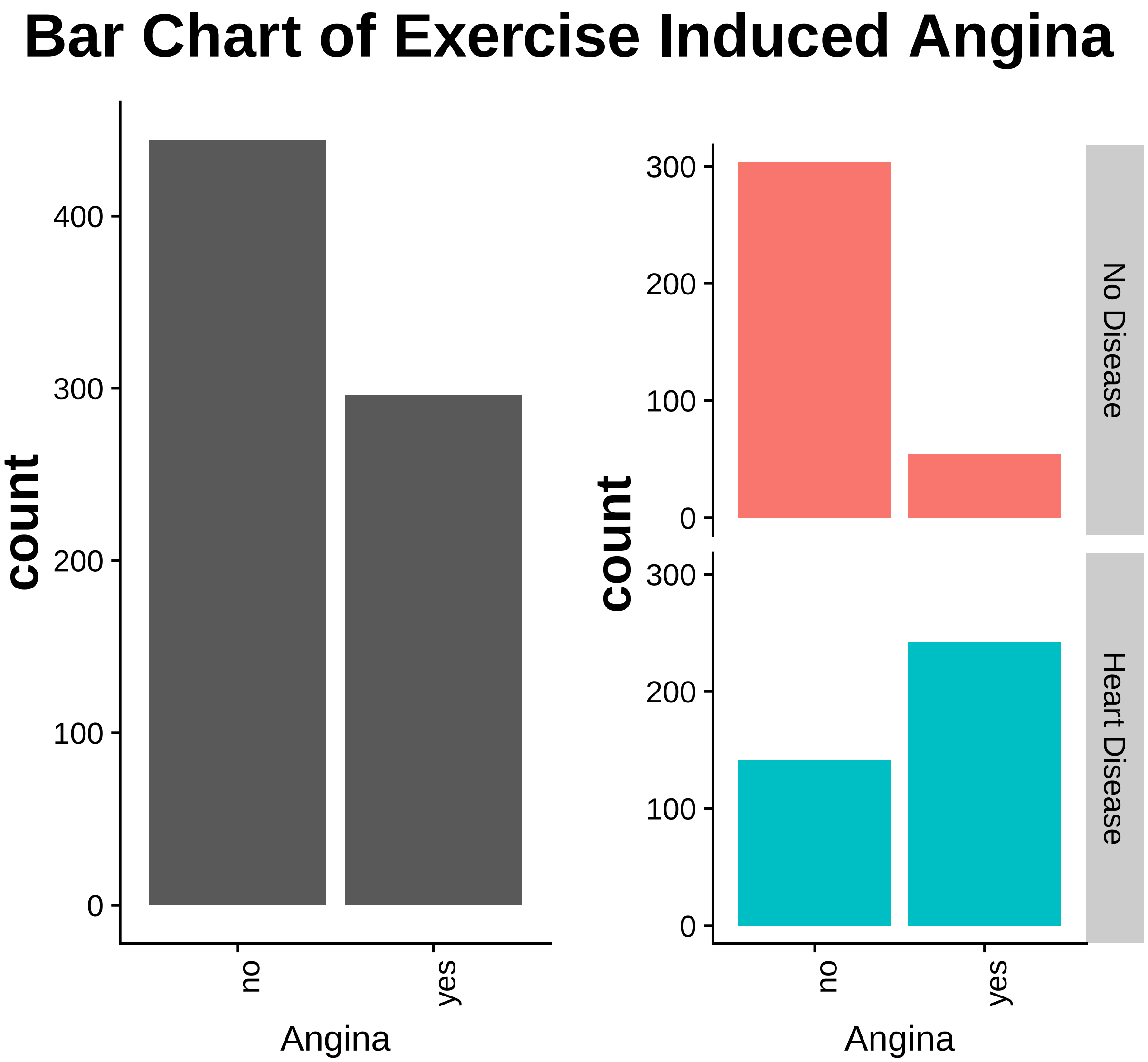

Figure 12. Exercise induced angina. Bar chart of exercise induced angina for all patients (left) compared to those for diseased and non-diseased patients (right).

4.1.2.6 Slope of Peak Exercise ST Segment

The slope of the peak exercise ST segment differed between the non-disease and diseased cohorts with the majority of cardiac disease patients exhibiting a flat ST slope (Fig. 13).

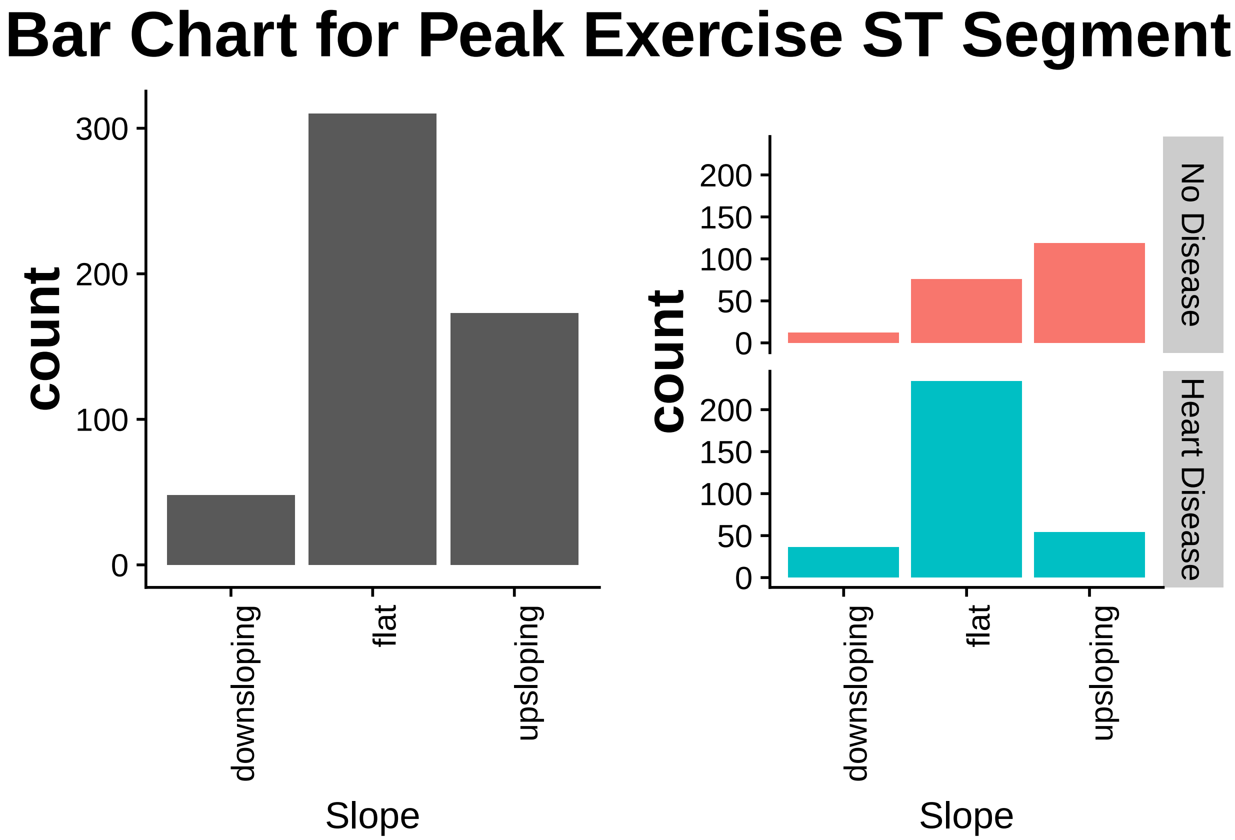

Figure 13. Slope of Peak Exercise ST Segment. Bar chart of slope of Peak Exercise ST Segment for all patients (left) compared to those for diseased and non-diseased patients (right).

4.1.2.7 β-Thalassemia Cardiomyopathy

Patients diagnosed with a reversible β-Thalassemia defect were more prevelant in the cardiac disease cohort while most non-disease patients exhibited a normal phenotype (Fig. 14). This difference may contribute to the prediction of heart disease.

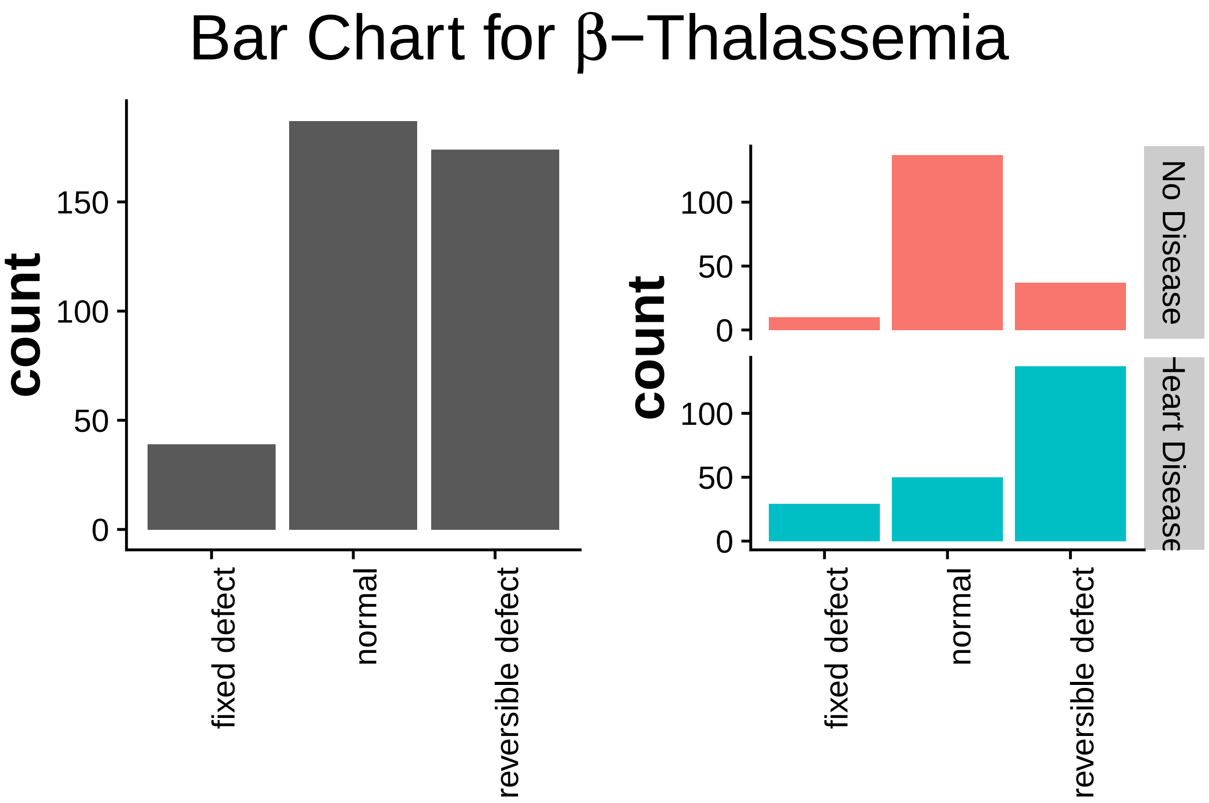

Figure 14. β-Thalassemia Cardiomyopathy . Bar chart for β-Thalassemia Cardiomyopathy patients (left) compared to those for diseased and non-diseased patients (right).

4.2 Multivariate Visualisation

4.2.1 Gender, Chest Pain Type and Resting ECG

The following visual shows that most individuals diagnosed with cardiac disease were males with typical signs of angina pain and an abnormal ST wave pattern (Fig. 15). A significant poportion of males with heart disease and abnormal ST wave patterns were asymptomatic for angina pain, indicating that angina symptoms alone were not highly predictive for heart disease. This was further supported by the fact that a high proportion of diseased male patients that were asymptomatic for angina exhibited hypertropy (i.e. enlarged heart) as diagnosed by ECG analysis. It should be noted that many males in the heart disease cohort exhibiting normal ECG ST wave patterns were either asymptomatic for angina pain or displayed typical angina pain symptoms.

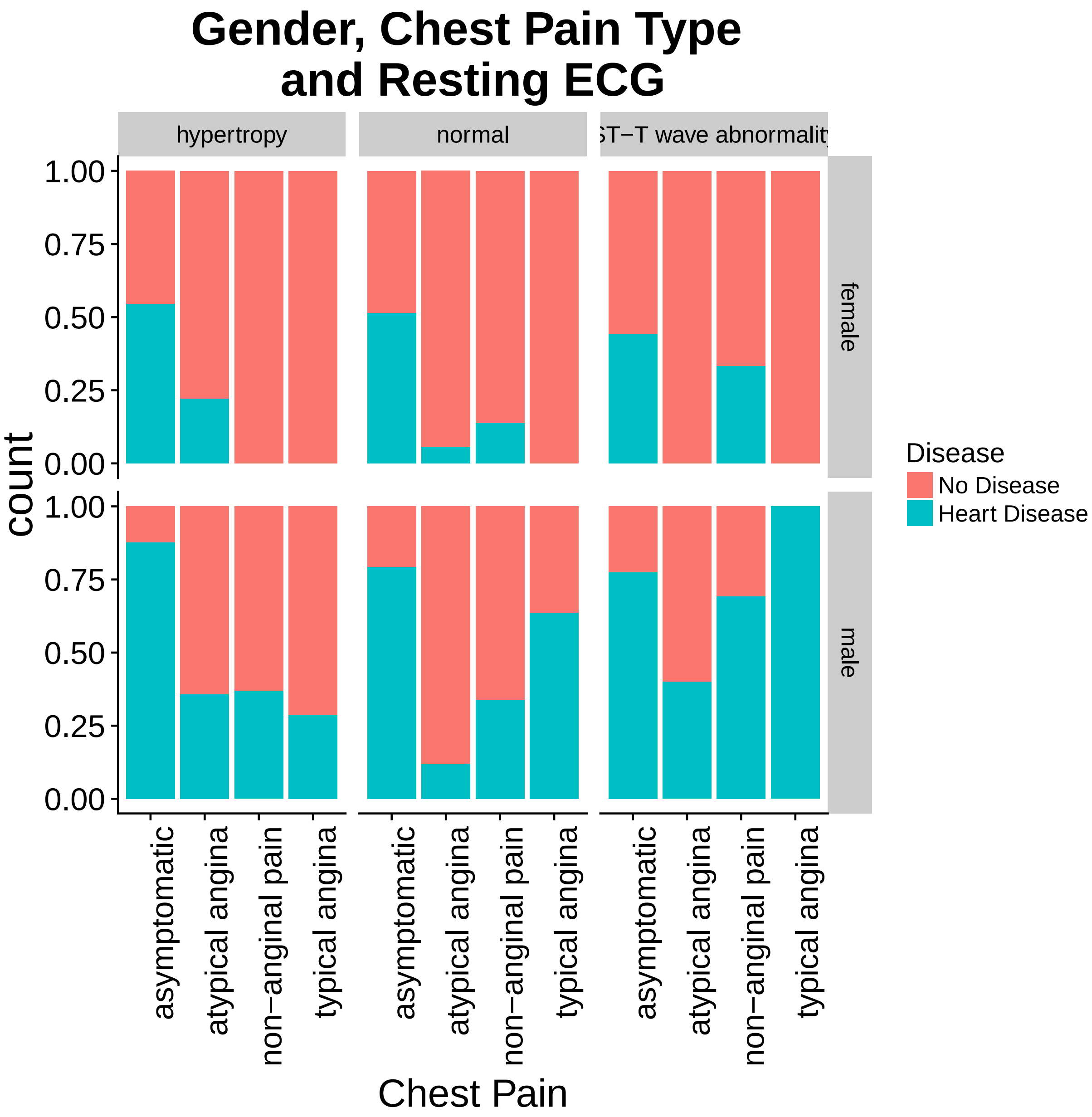

Figure 15. Comparison of gender, chest pain type and resting ECG . Proportional bar chart for diseased and non-diseased patients.

4.2.2 Gender, Chest Pain Type and Exercise Induced Angina

Although the majority of males and females diagnosed with cardiac disease were asymptomatic for angina pain, they did exhibit exercise induced angina pain (Fig. 16). This added further support to the use of the exercise induced angina pain feature for the prediction of heart disease.

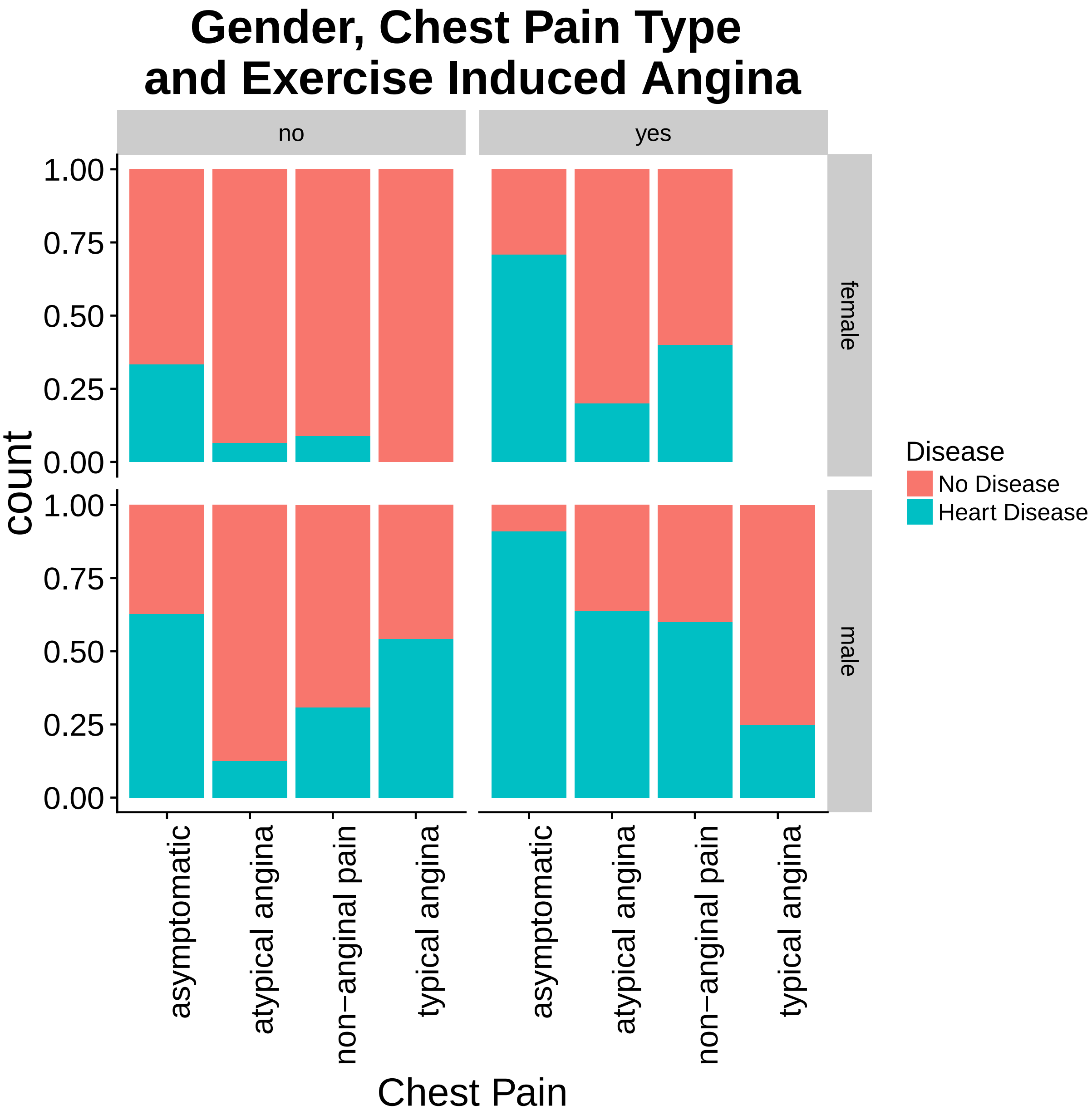

Figure 16. Comparison of gender, chest pain type and excerise induced angina . Proportional bar chart for diseased and non-diseased patients.

4.2.3 Gender, Chest Pain Type and β-Thalassemia

Most female and male patients diagnosed with β-Thalassemia, or the reversible phenotype, and displaying signs of cardiac disease did not exhibit non-stress induced angina pain (i.e. asymptomatic) (Fig. 17).

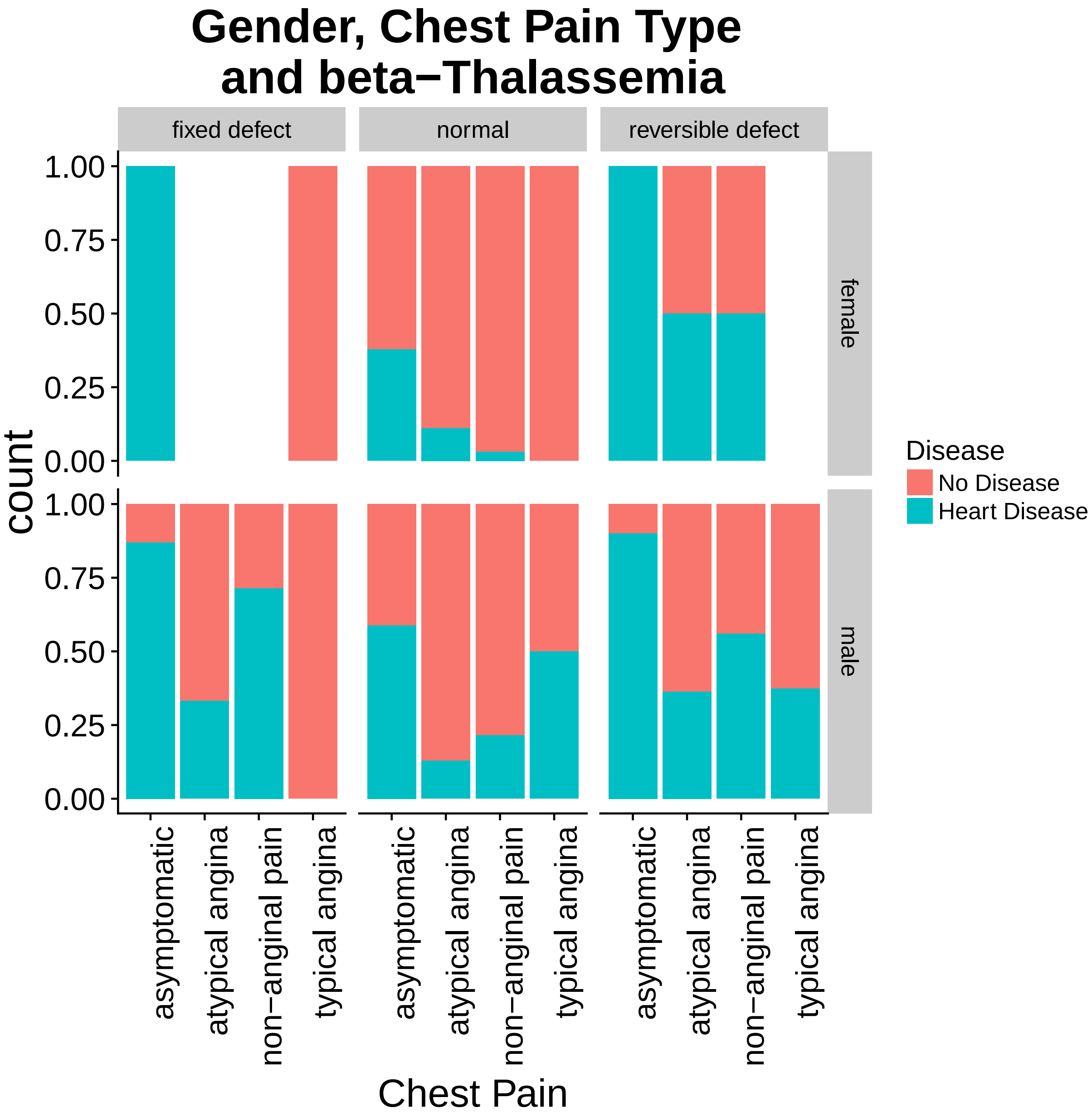

Figure 17. Comparison of gender, chest pain type and β Thalassemia. Proportional bar chart for diseased and non-diseased patients.

4.2.4 Gender, Chest Pain and Age

The *stacked* histograms revealed that there was little difference in the age distribution of female and male patients diagnosed with heart disease and either asymptomatic or experiencing non-stress induced angina pain (Fig. 18).

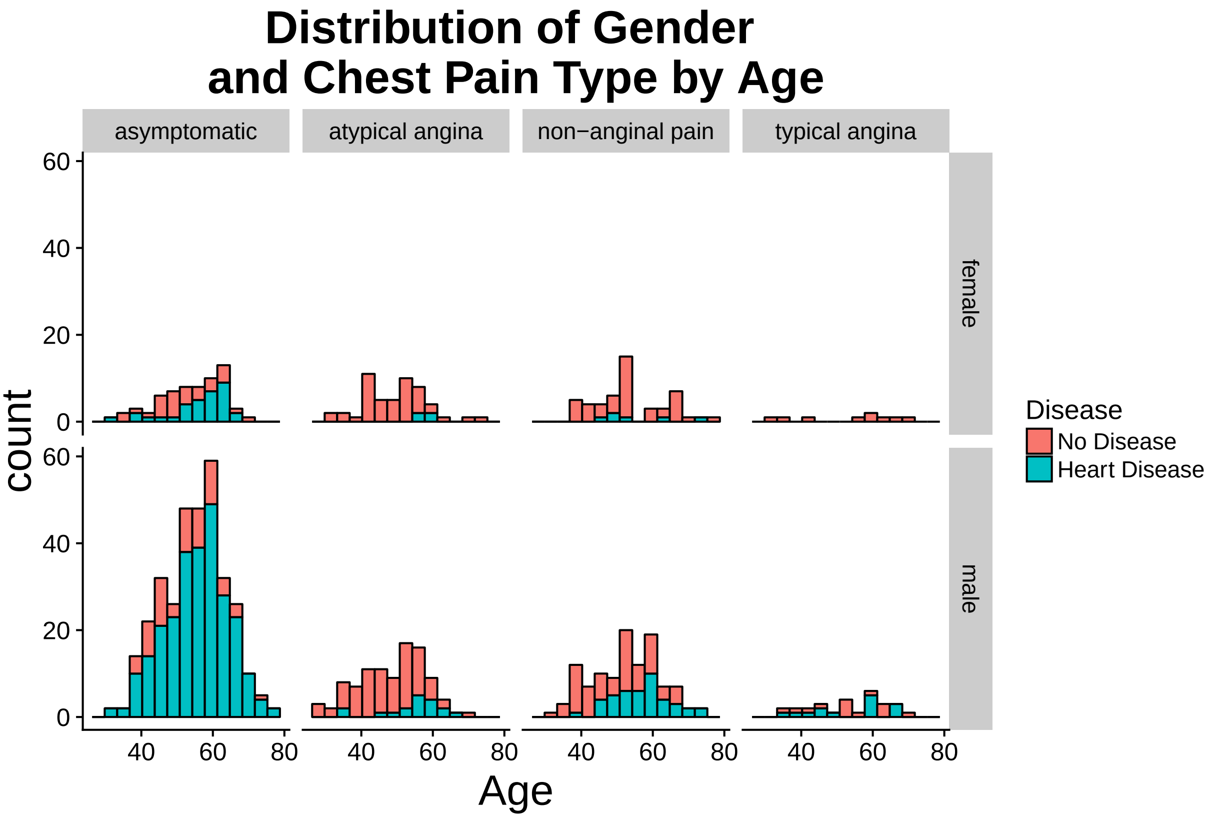

Figure 18. Comparison of gender, chest pain type and age. Proportional bar chart for diseased and non-diseased patients.

4.2.5 Gender, Resting ECG and Age

The distribution of gender and resting ECG results with age mostly did not highlight any significant difference between the various patient cohorts (Fig. 19). There was a higher proportion of male individuals with normal ECG results exhibiting no signs of disease in the younger age group.

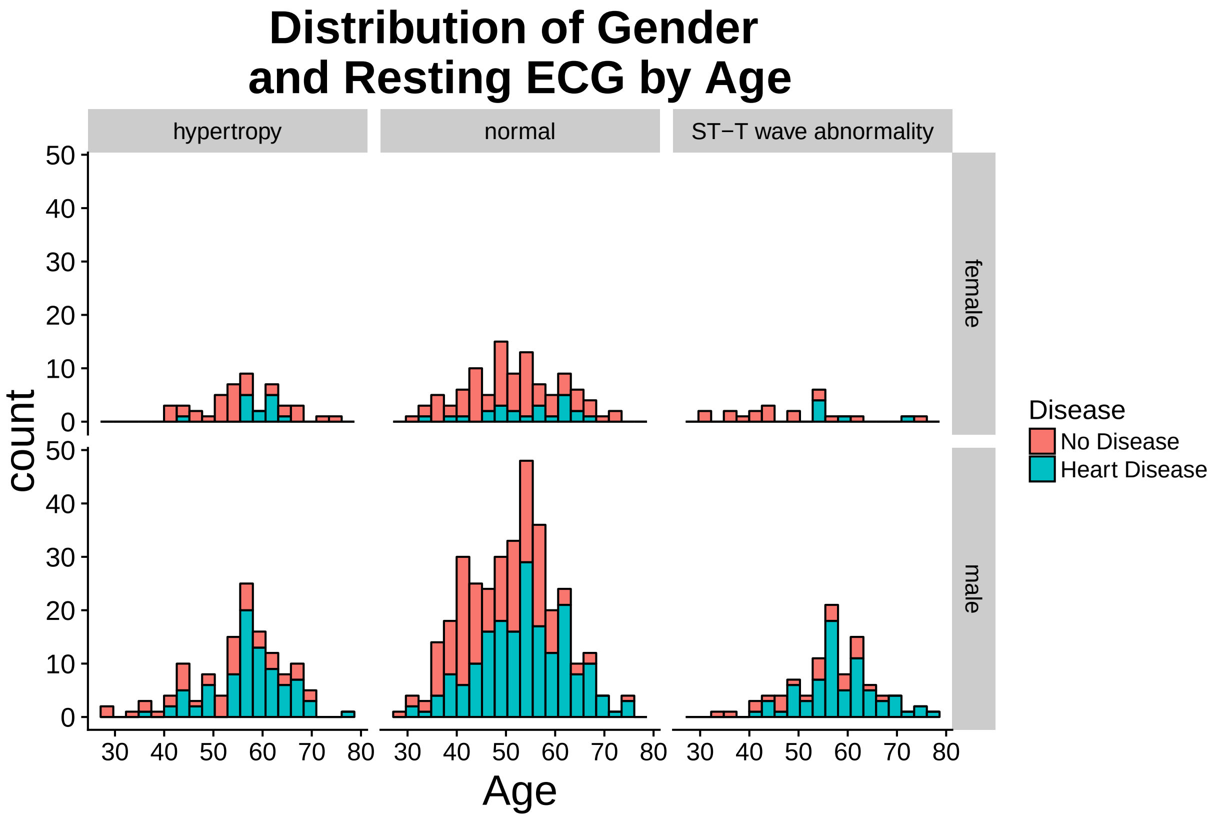

Figure 19. Comparison of gender, resting ECG and age. Proportional bar chart for diseased and non-diseased patients.

4.2.6 Gender, Slope of Peak Exercise ST Segment and Age

Examination of the distribution of gender and ECG ST peak wave segment slope with age did not reveal any significant differences between these features (Fig. 20). Overall, a higher proportion of cardiac disease patients (males and females) exhibited a upsloping wave pattern but the age distribution of these individuals did not dramatically differ.

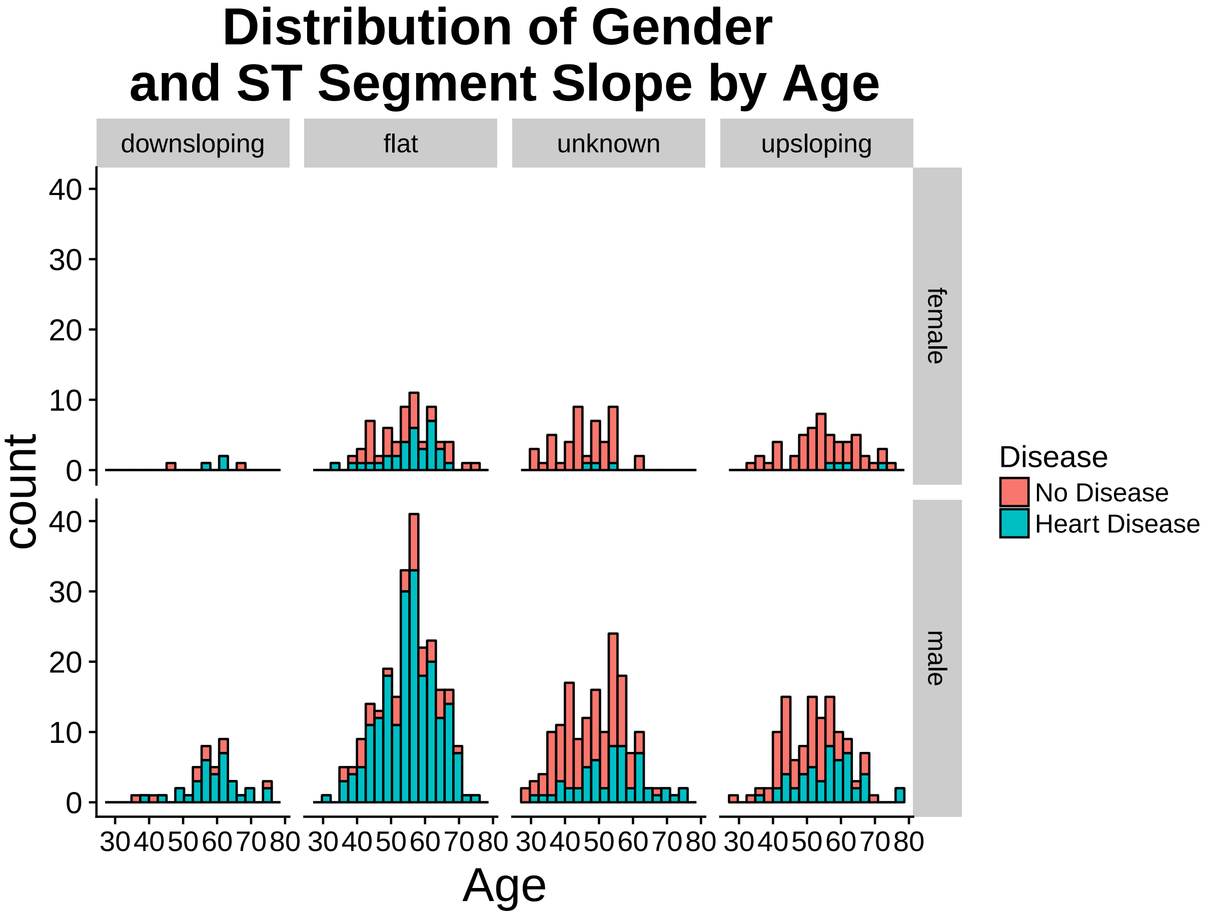

Figure 20. Comparison of gender, ST segment slope and age. Proportional bar chart for diseased and non-diseased patients.

4.2.7 Gender β-Thalassemia Cardiomyopathy and Age

Examination of the distribution of gender for β-Thalassemia cardiomyopathy suffers with age indicated little difference for individuals with a fixed or reversible phenotype (Fig. 21). However, most patients exhibiting a normal β-Thalassemia phenotype did not have heart disease and those that did tended to be in a higher age group. There were far fewer females diagnosed with β-Thalassemia cardiomyopathy in the patient cohort.

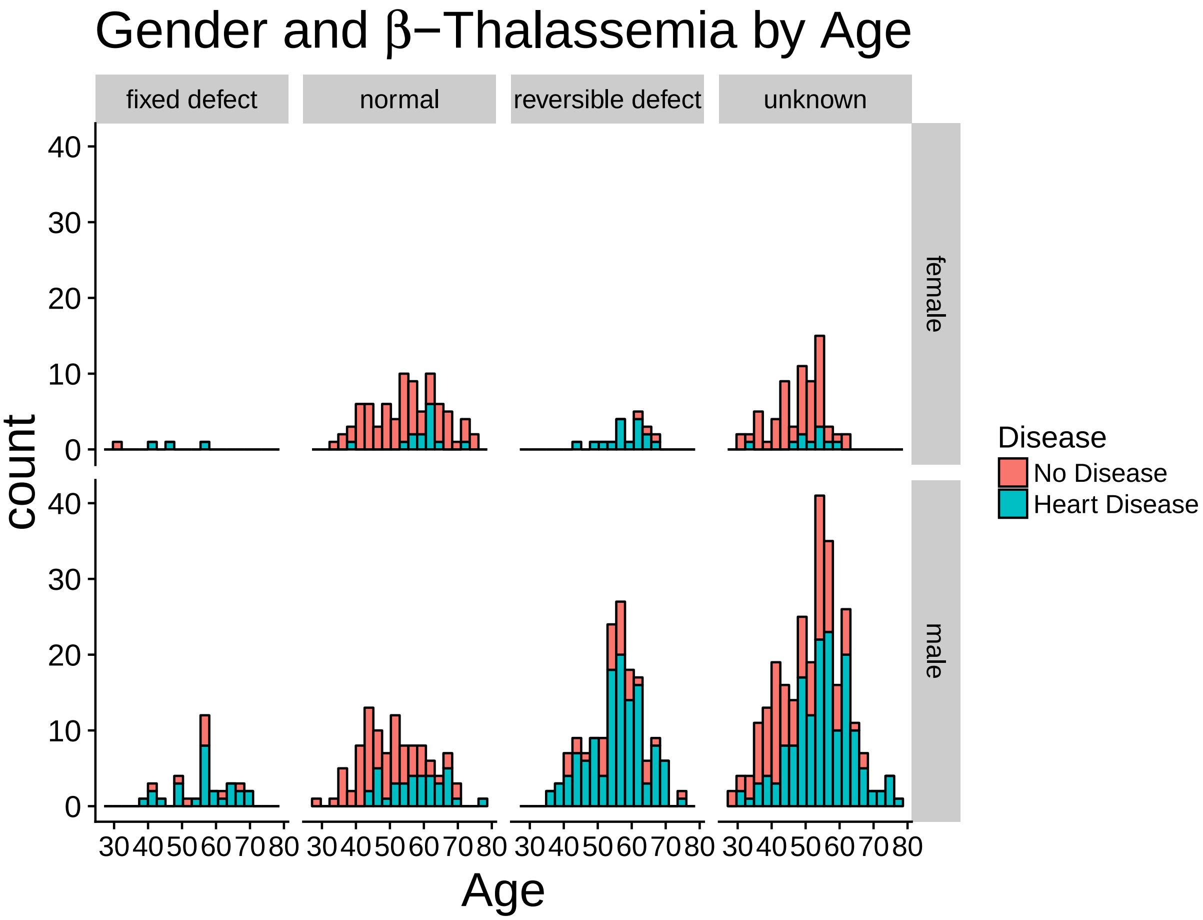

Figure 21. Comparison of gender, β-Thalassemia cardiomyopathy and age. Proportional bar chart for diseased and non-diseased patients.

4.3 Regression Analysis

4.3.1 Age versus Resting Blood Pressure

Regression analysis of resting blood pressure versus age for the cohort indicated little significant difference between those displaying no signs of disease and individuals diagnosed with cardiac disease (Fig. 22). There was a slight postive correlation between age and resting blood pressure for both cohorts of patients.

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing missing values (geom_point).
```

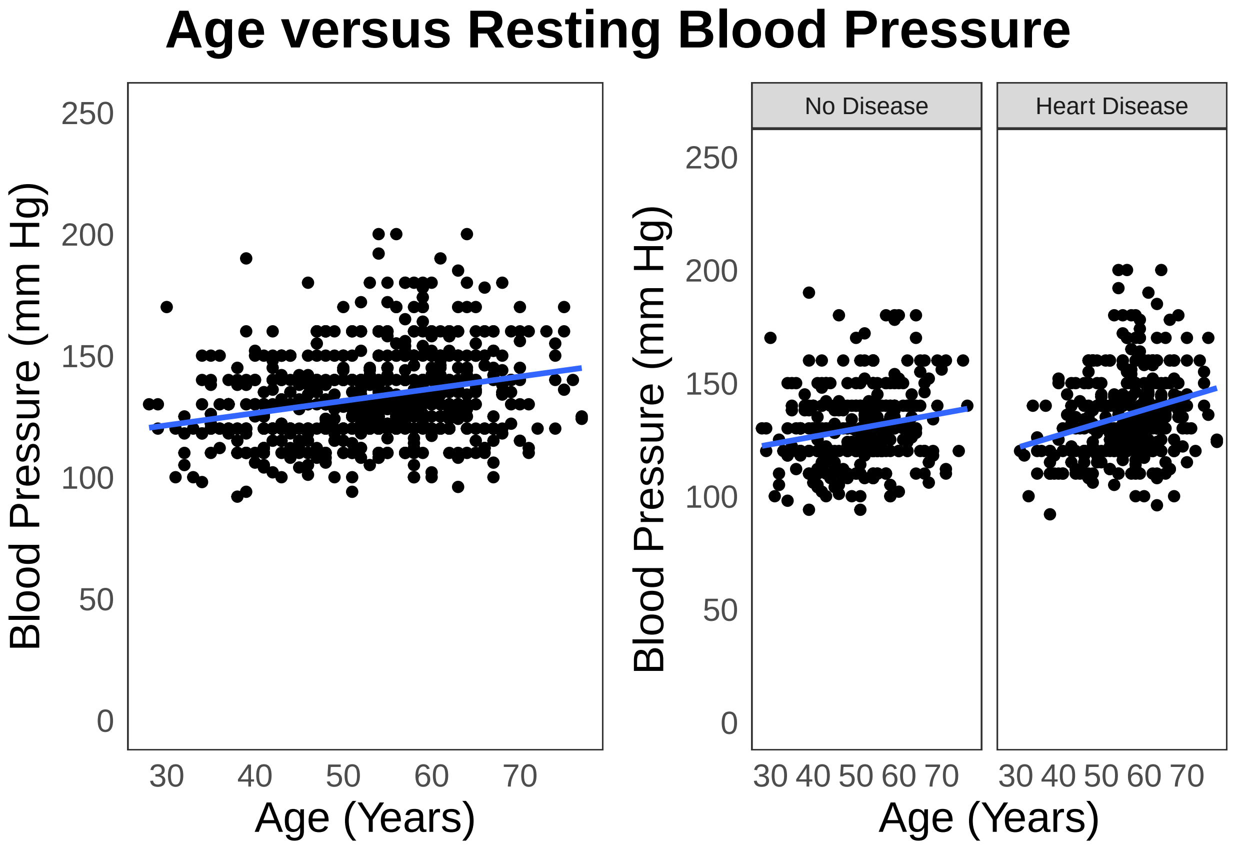

Figure 22. Comparison of age and resting blood pressure. Regression analysis for diseased and non-diseased patients.

```
## [[1]]
##
## Call:
## lm(formula = Age ~ Trestbps, data = .)
##
## Coefficients:
## (Intercept)     Trestbps
##     35.9303       0.1107
##
##
## [[2]]
##
## Call:
## lm(formula = Age ~ Trestbps, data = .)
##
## Coefficients:
## (Intercept)     Trestbps
##     41.2740       0.1065
```

4.3.2 Age versus Cholestrol Level

Linear regression analysis of age versus cholestrol levels demonstrated a positive correlation (slope = 0.032) for the no-diseased group but a negative correlation (slope = -0.009) for the diseased cohort (fig. 23). However, several outliers will need to be removed in future studies to determine their effect upon the regression analysis.

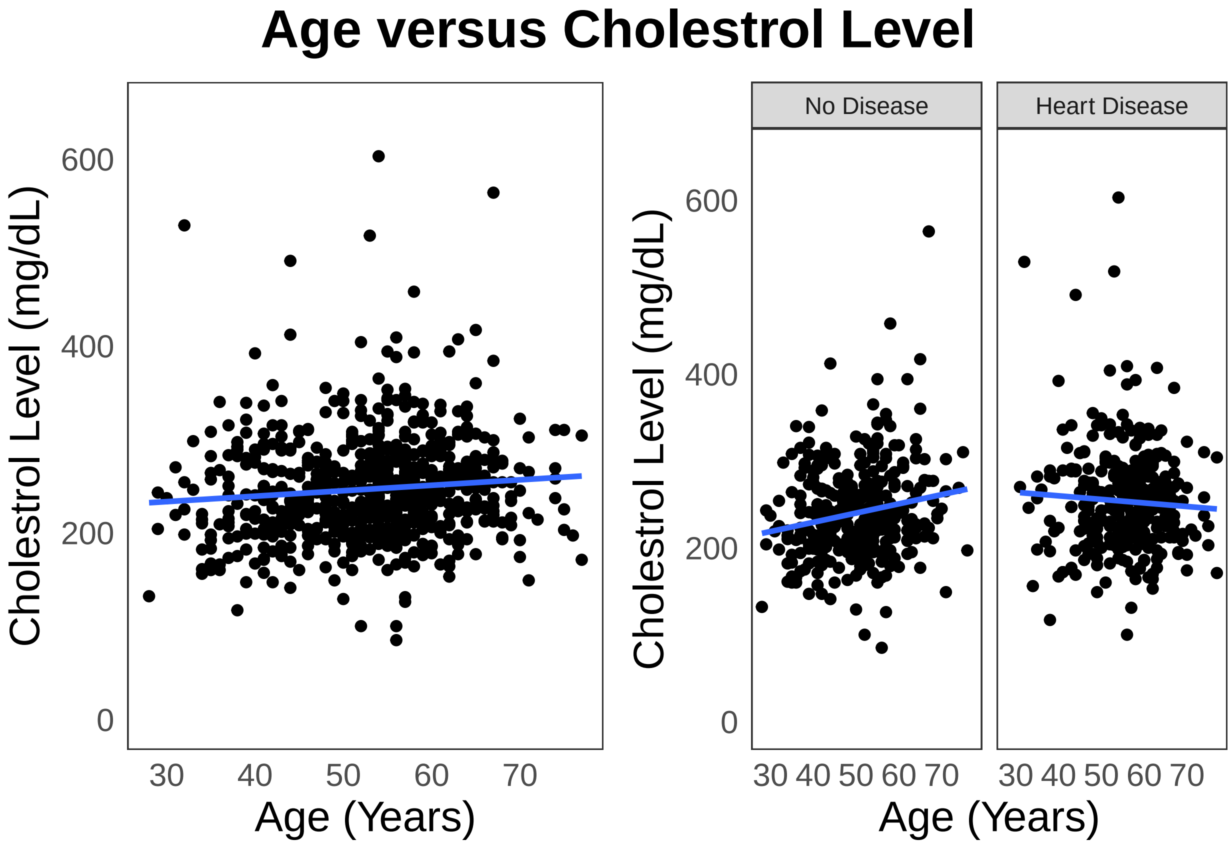

Figure 23. Comparison of age and cholestrol levels. Regression analysis for diseased and non-diseased patients.

```
## [[1]]
##
## Call:
## lm(formula = Age ~ Chol, data = .)
##
## Coefficients:
## (Intercept)         Chol
##    42.45566      0.03177
##
##
## [[2]]
##
## Call:
## lm(formula = Age ~ Chol, data = .)
##
## Coefficients:
## (Intercept)         Chol
##   57.574152    -0.008546
```

4.3.3 Age versus Maximum Heart Rate

Regression analysis of maximum heart rate versus age showed little difference between the overall cohort and diseased and non-diseased individuals (Fig. 24).

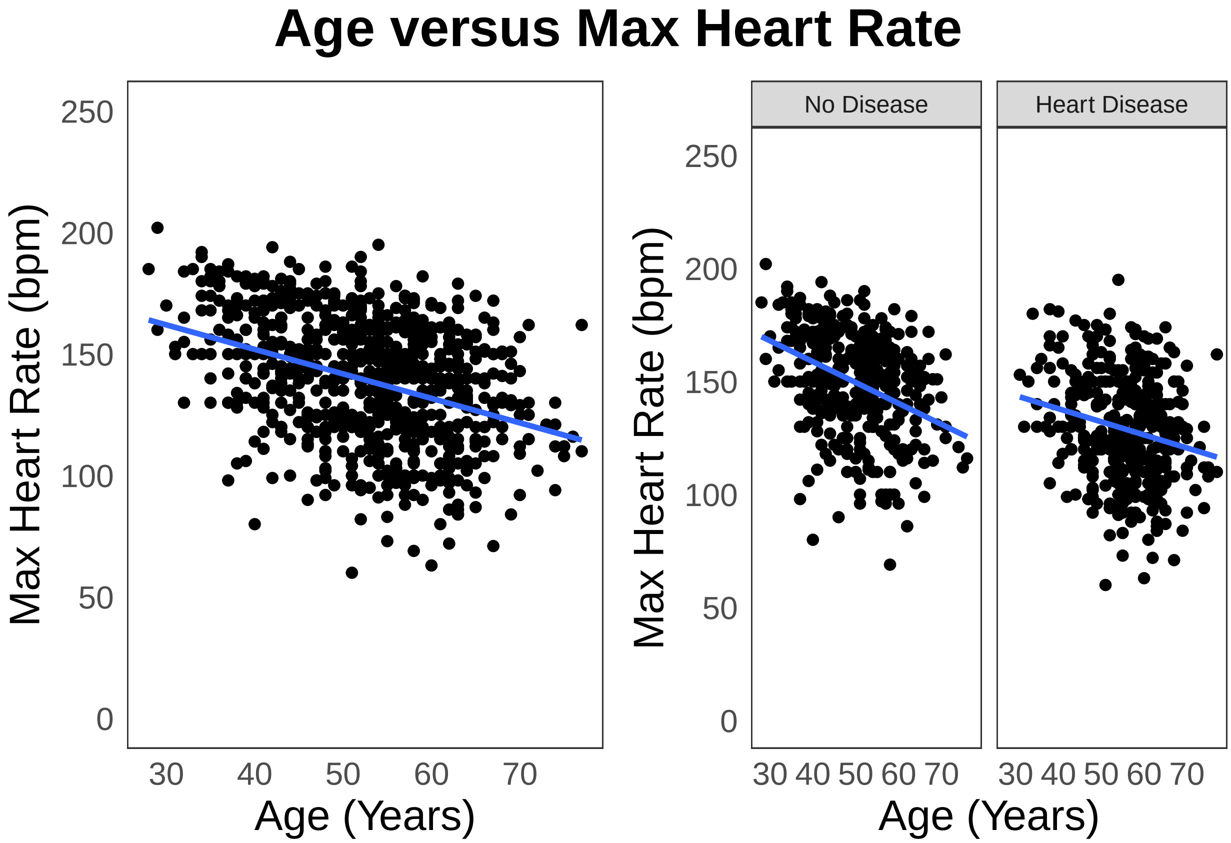

Figure 24. Comparison of age and maximum heart rate. Regression analysis for diseased and non-diseased patients.

```
## [[1]]
##
## Call:
## lm(formula = Age ~ Thalach, data = .)
##
## Coefficients:
## (Intercept)      Thalach
##     72.1712      -0.1465
##
##
## [[2]]
##
## Call:
## lm(formula = Age ~ Thalach, data = .)
##
## Coefficients:
## (Intercept)      Thalach
##    65.44445     -0.07557
```

5 Summary

The CA (number of major vessels) feature was removed due to the large proportion ($> 60\%$) of missing values. Rows containing missing values in multiple columns were also removed since they comprised only a small proportion of the dataset (i.e. 6%). Exploration of the data indicated that patient age, cholestrol level, maximum heart rate, ST peak depression induced by exercise and slope of the peak exercise ST segment, gender and exercise induced angina were possible useful features for predicting the presence of cardiac disease. Electrocardiographic results and β-Thalassemia diagnosis were also found to have a potentially minor predictive power.