Heart Disease Detection Project

Jimmy Nguyen

Contents

The Data Science Methodology

1. Problem Understanding Phase
2. Data Preparation Phase
3. Exploratory Data Analysis Phase
4. Setup Phase
5. Modeling Phase
6. Evaluation Phase
7. Deployment Phase

Problem Understanding Phase

Project Objectives

The purpose of this project is to assist health-care providers and physicians to increase the number of accurate heart disease diagnosis.

The objectives of this analysis are as follows:

1. Learn about potential patients who may be at risk with heart disease. That is, learn the characteristics of those who have heart disease, as well as those who do not.
2. Develop an efficient and reliable method of identifying likely positive cases of heart disease, so that we may proceed with treatment and save lives. That is, develop a model or models that may identify likely positive cases of heart disease.

Translating These Objectives into a Data Science Problem

1. There are many ways to learn about patients with or without heart diseases.
    a. Use exploratory data analysis to analyze the relationship between variables by expressing visualizations such as observing an independent variable with the target variable overlain with information about positive or negative for heart disease.
    b. Use clustering to determine whether there are natural groupings among patients with either diagnosis for heart disease, for example, younger males vs. older males. Then, see if these clusters differ with respect to their diagnosis of heart disease.
    c. Use association rules to analyze useful relationships among the characteristics of patients diagnosed with heart disease.
2. Since the target variable in this data set has binary categorical values, this is treated using classification models and as a binary classification problem.
    a. Develope the best classification models, using the following algorithms:
        i. Decision Trees
        ii. Random Forests
        iii. Naive Bayes Classification
        iv. Neural Networks
        v. Logistic Regression.
    b. Evaluate each model based on the following predetermined model evlauation criteria:
        i. Misclassification costs
        ii. Precision
        iii. Sensitivity (Recall)
        iv. F-score (Harmonic Mean)

        Construct a table of the best models and their costs.

About the Data

The data set was originally provided by the University of California, Irvine repository and was hosted on Kaggle for the public. In total, this data set has only 303 rows and 14 columns. Thirteen of the columns contain information about each patient whereas the last column is the target variable of interest. The last column contains binary values of zeroes and ones indicating the presence of heart disease. Other columns behaved as independent variables such as the patients demographic information and measurements of their physical health. This data set contains both continuous and discrete values.

- The data set can be found here on Kaggle: https://www.kaggle.com/ronitf/heart-disease-uci

- The data set can be found here from UCI Repository: https://archive.ics.uci.edu/ml/datasets/heart+disease

Data Preparation Phase

Packages

- These are the packages used in this data report

```
library(flextable)
library(tidyverse)
library(pander)
library(plyr)
```

Load in Data

- The following code reads in the heart disease data-set.

R code

```
df <- read_csv("heart.csv")
```

Table 1: Heart Disease Data Set (continued below)

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
63	1	3	145	233	1	0	150	0
37	1	2	130	250	0	1	187	0
41	0	1	130	204	0	0	172	0
56	1	1	120	236	0	1	178	0
57	0	0	120	354	0	1	163	1
57	1	0	140	192	0	1	148	0

oldpeak	slope	ca	thal	target
2.3	0	0	1	1
3.5	0	0	2	1
1.4	2	0	2	1
0.8	2	0	2	1
0.6	2	0	2	1
0.4	1	0	1	1

Re-expressing Categorical Field Values

- The following chunk of code replaces numerical categorical values to their original categorical values as factors.

R code:

```
# mapped values to female and male
df$sex <- revalue(x = df$sex, replace = c("0" = "female",
"1" = "male"))

# mapped values for chest pain
df$cp <- revalue(x = df$cp, replace = c("0" = "typical angina",
                                        "1" = "atypical angina",
```

```
                                        "2" = "non-anginal pain",
                                        "3" = "asymptomatic"))

# mapped values for fastinb blood sugar
df$fbs <- revalue(x = df$fbs, replace = c("0"="false",
                                          "1" = "true"))

# mapped values for resting blood pressure
df$restecg <- revalue(x = df$restecg, replace = c("0" = "normal",
                                                  "1" = "abnormal",
                                                  "2" = "probable hypertrophy"))

# mapped values for exercised-induced angina
df$exang <- revalue(x = df$exang, replace = c("0" = "no",
                                              "1" = "yes"))

# mapped values for slope
df$slope <- revalue(x = df$slope, replace = c("0" = "upsloping",
                                              "1" = "flat",
                                              "2" = "downsloping"))
```

Handling Missing Data

- The following table displays each column as a row and the number of missing values in each column.

Table 3: Count of Missing Values - This table shows that there are no missing values in this data set.

Columns	Number of Missing Value
age	0
sex	0
cp	0
trestbps	0
chol	0
fbs	0
restecg	0
thalach	0
exang	0
oldpeak	0
slope	0
ca	0
target	0

Identifying Misclassifications

- The categorical variables in this data set are the following:

R code

```
cat_vars <- df[,names(df[,sapply(df, is.factor)])]
```

Table 4: Table of Categorical Variables (continued below)

sex	cp	fbs	restecg	exang	slope	ca
male	asymptomatic	true	normal	no	upsloping	0
male	non-anginal pain	false	abnormal	no	upsloping	0
female	atypical angina	false	normal	no	downsloping	0
male	atypical angina	false	abnormal	no	downsloping	0
female	typical angina	false	abnormal	yes	downsloping	0

target
1
1
1
1
1

- The following tables check the classes of all the categorical variables for validity and consistency.

Table 6: Classes in Sex

sex	total
female	96
male	207

Table 7: Classes in Chest Pain

cp	total
typical angina	143
atypical angina	50
non-anginal pain	87
asymptomatic	23

Table 8: Classes in Fasting Blood Sugar

fbs	total
false	258
true	45

Table 9: Classes in Resting Electrocardiographic Results

restecg	total
`normal`	147
`abnormal`	152
probable hypertrophy	4

Table 10: Classes in Exercise-Induced Angina

exang	total
no	204
yes	99

Table 11: Classes in Heart Disease Diagnosis

slope	total
upsloping	21
flat	140
downsloping	142

Check for Normality

Table 12: Continuous Variables

age	trestbps	chol	thalach	oldpeak
63	145	233	150	2.3
37	130	250	187	3.5
41	130	204	172	1.4
56	120	236	178	0.8
57	120	354	163	0.6

Observing Normality in: *age*

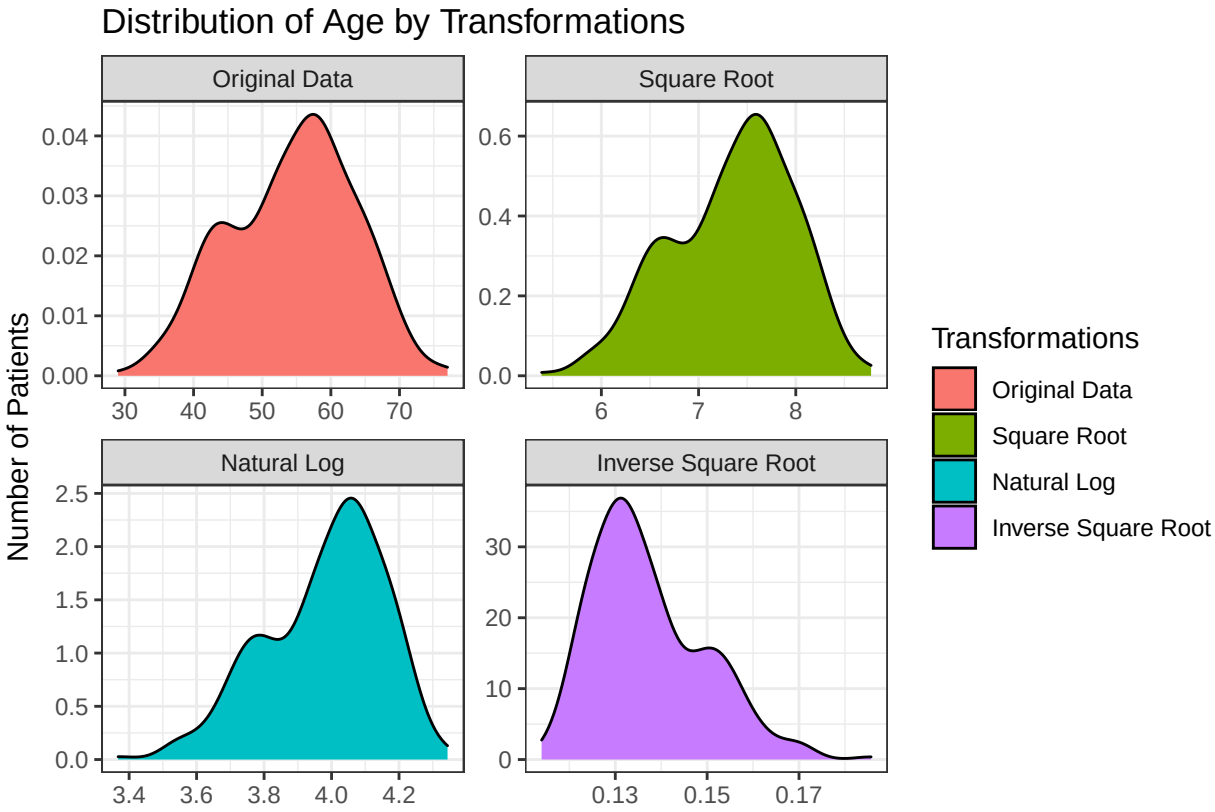
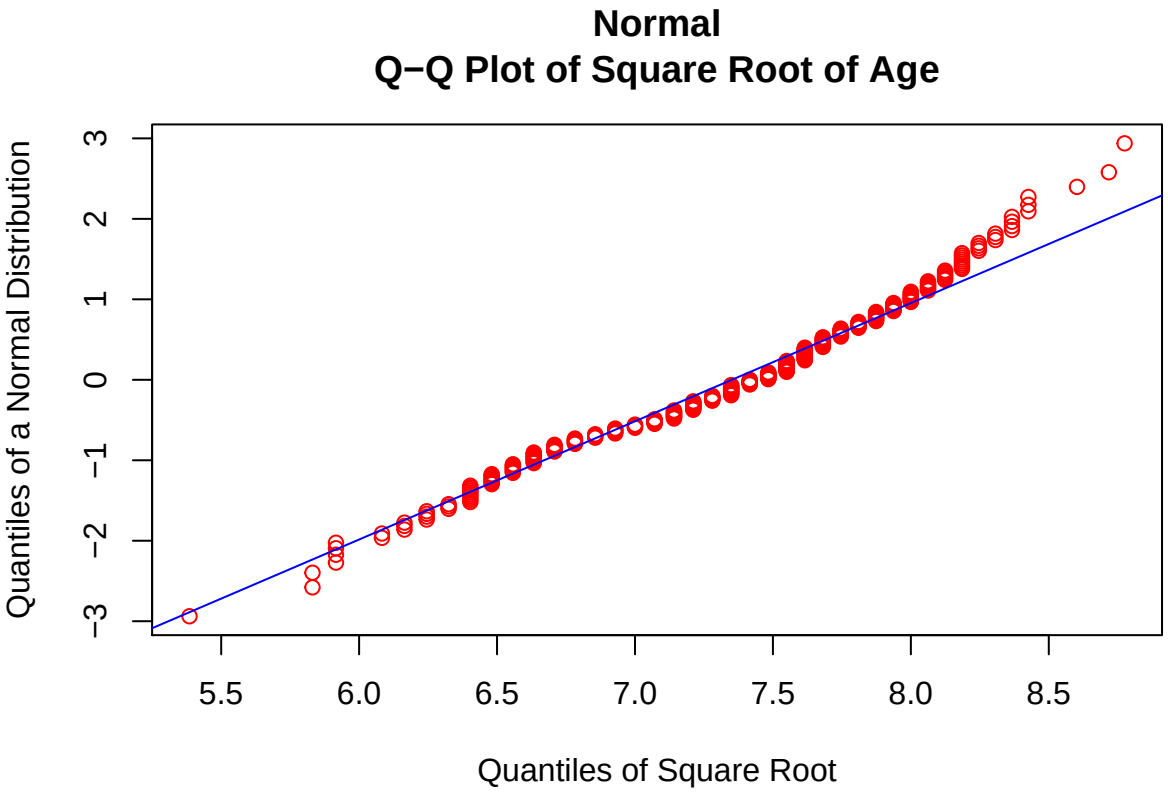


Table 13: Skewness by Transformation: 'Age'

Transformations	Skewness
Original Data Skewness	-0.2093
Square Root of Data Skewness	-0.3323
Natural Log of Data Skewness	-0.4514
Inverse Square Root of Data Skewness	0.5644



Observing Normality in: ***trestbps***

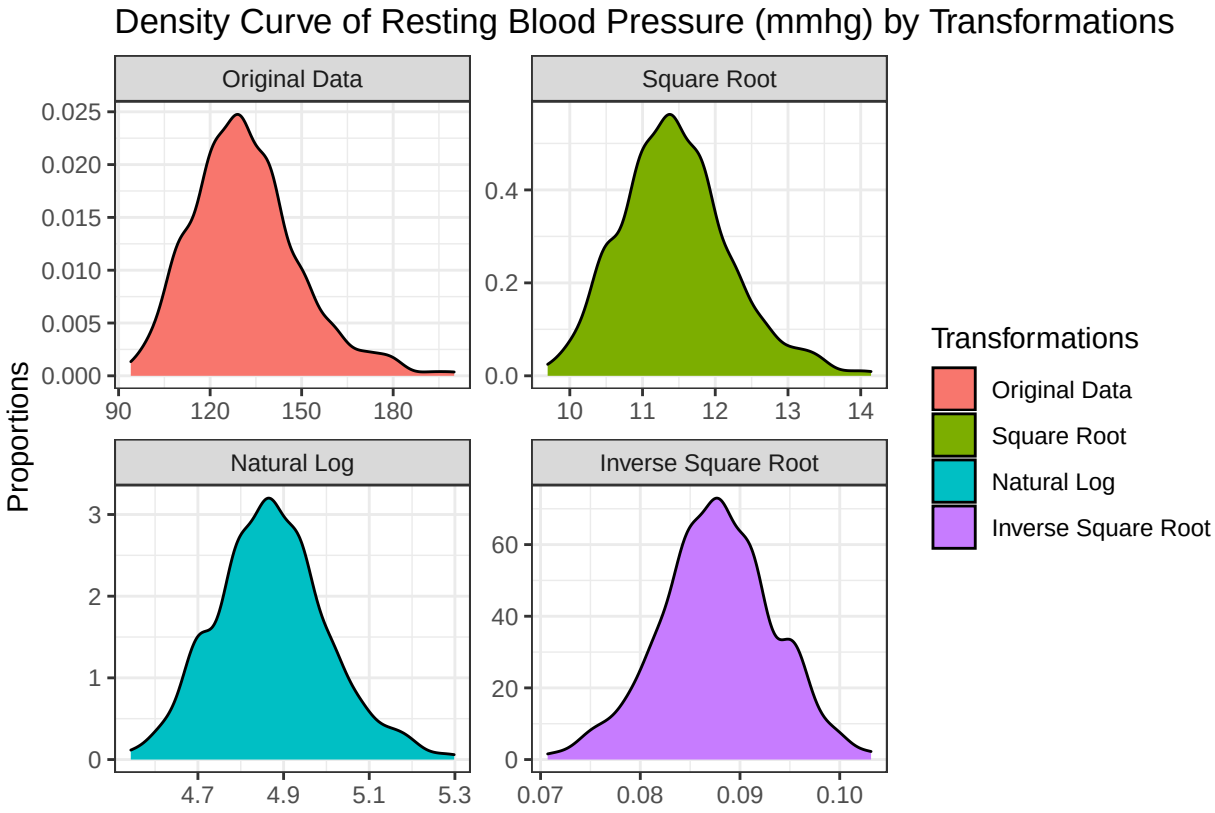
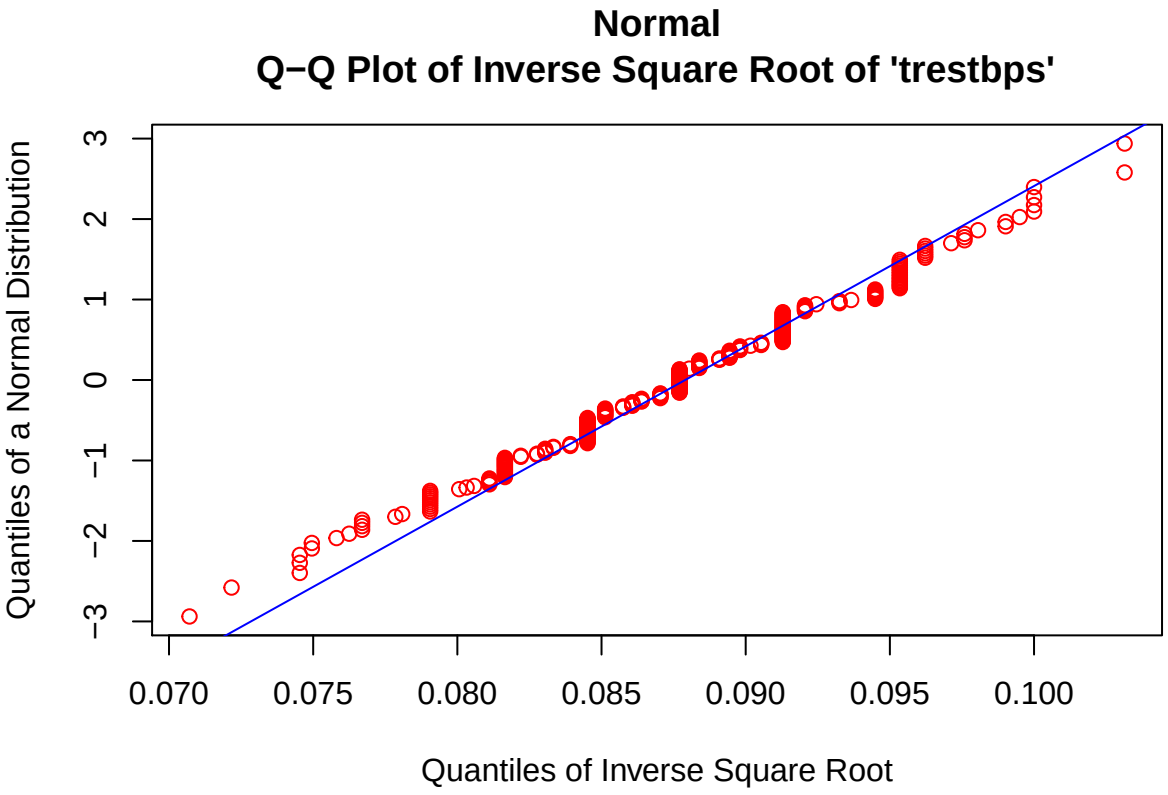


Table 14: Skewness by Transformation: 'trestbps'

Transformations	Skewness
Original Data Skewness	0.2778
Square Root of Data Skewness	0.1846
Natural Log of Data Skewness	0.08898
Inverse Square Root of Data Skewness	0.00782



Observing Normality in: *chol*

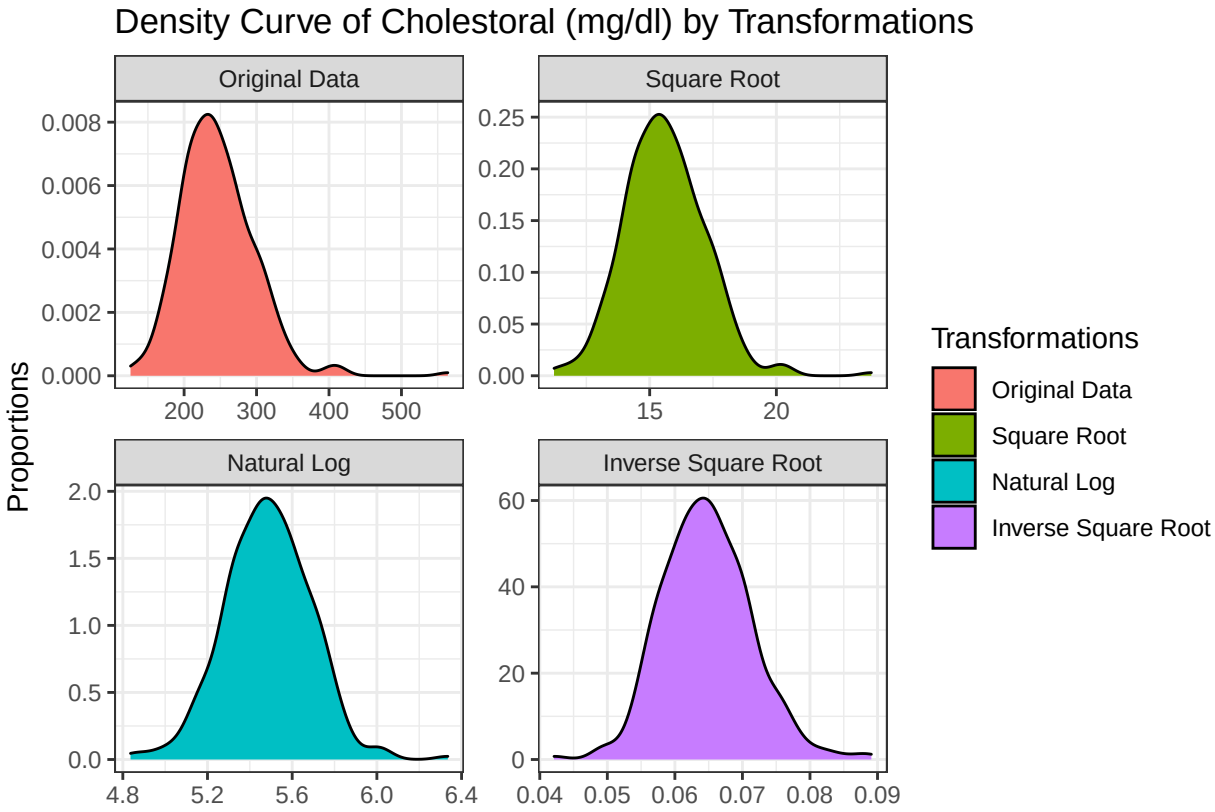
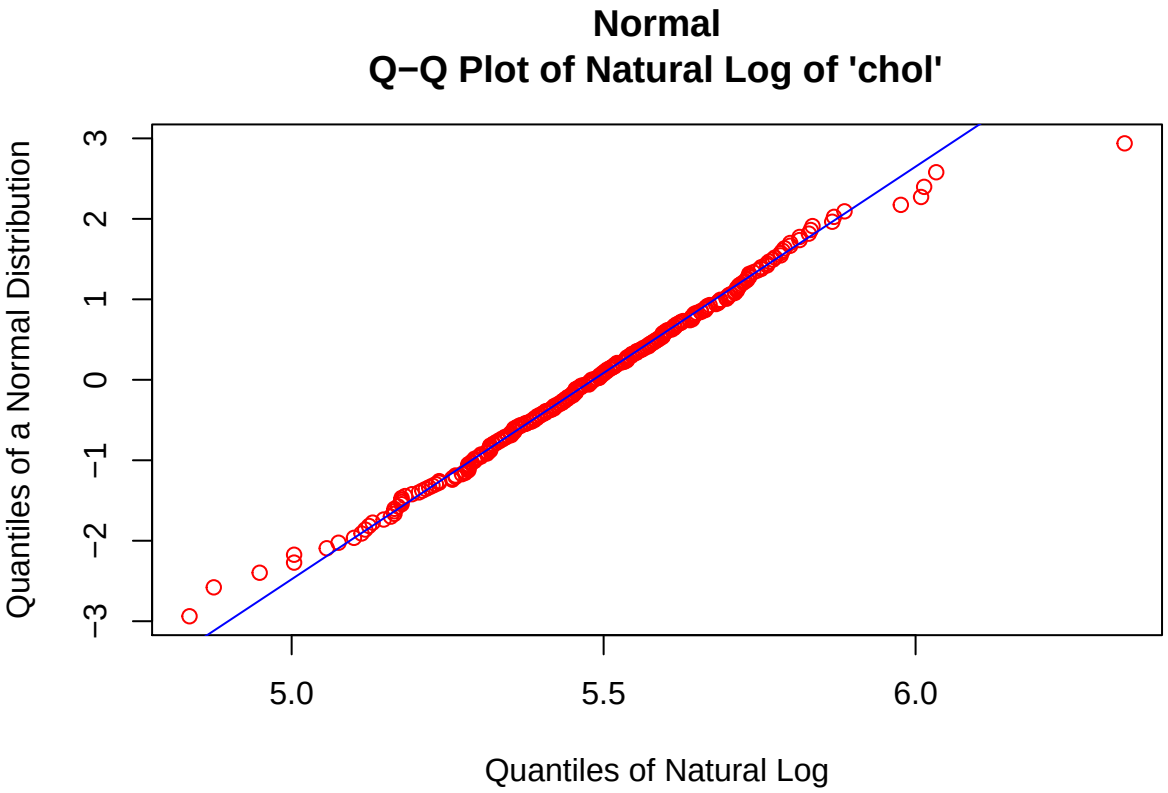


Table 15: Skewness by Transformation: 'chol'

Transformations	Skewness
Original Data Skewness	0.3626
Square Root of Data Skewness	0.2216
Natural Log of Data Skewness	0.07175
Inverse Square Root of Data Skewness	0.0801



Observing Normality in: *thalach*

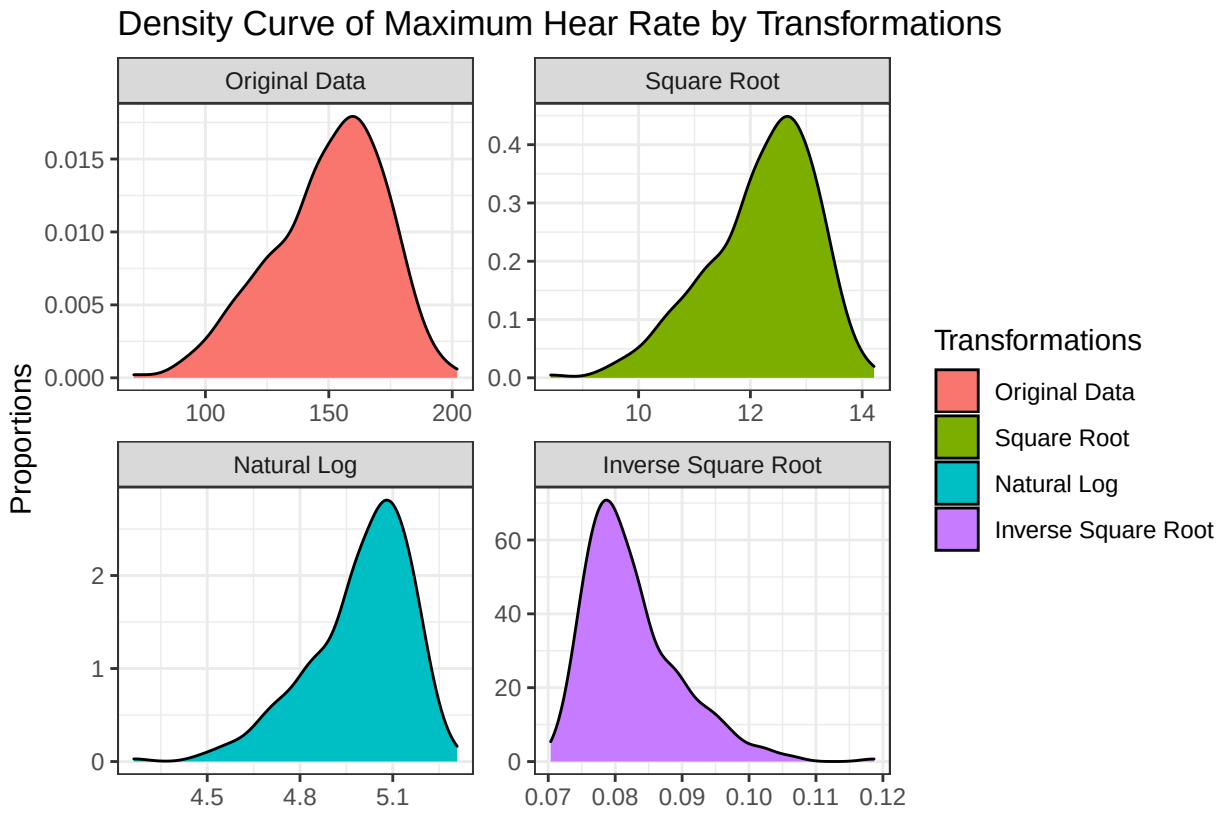
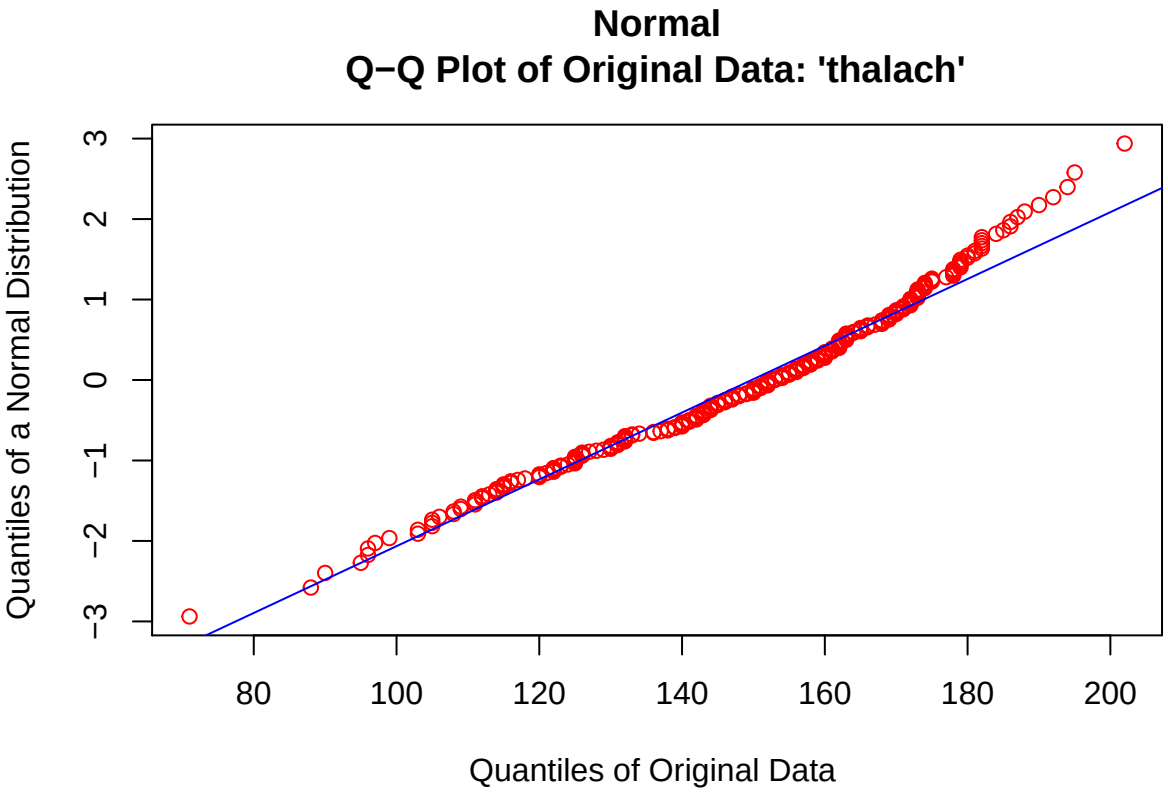


Table 16: Skewness by Transformation: 'thalach'

Transformations	Skewness
Original Data Skewness	-0.4392
Square Root of Data Skewness	-0.5414
Natural Log of Data Skewness	-0.637
Inverse Square Root of Data Skewness	0.7237



Observing Normality in: *thalach*

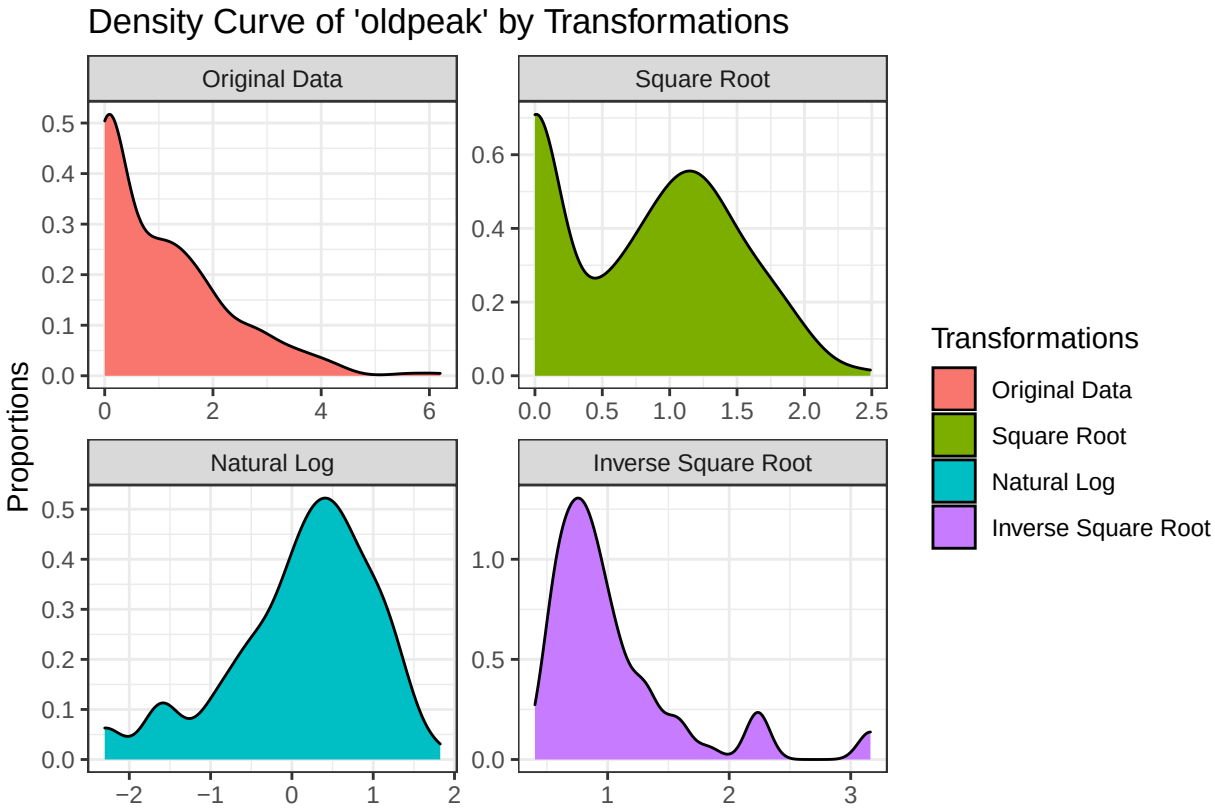
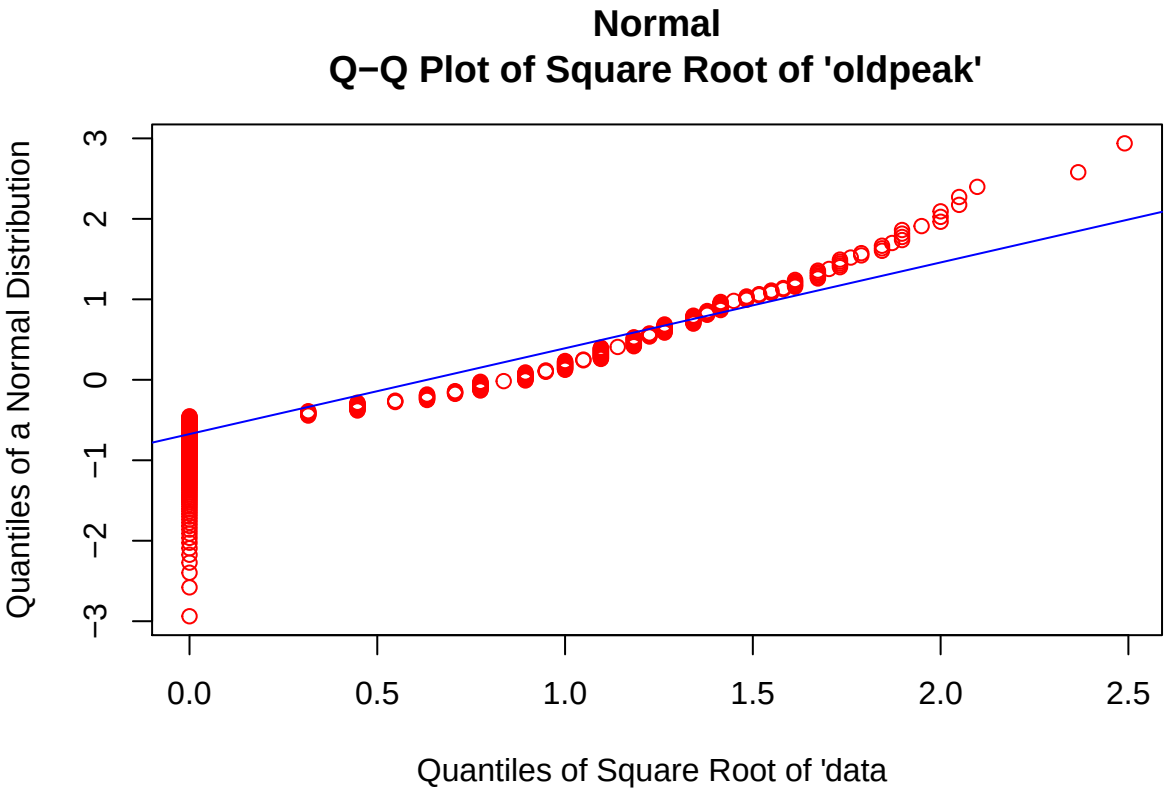


Table 17: Skewness by Transformation: 'oldpeak'

Transformations	Skewness
Original Data Skewness	0.6191
Square Root of Data Skewness	-0.5193
Natural Log of Data Skewness	NA
Inverse Square Root of Data Skewness	NA



Identify Outliers

Data Set Ready for EDA

Table 18: Heart Disease Data Set (continued below)

age	sex	cp	trestbps	chol	fbs	restecg
63	male	asymptomatic	145	233	true	normal
37	male	non-anginal pain	130	250	false	abnormal
41	female	atypical angina	130	204	false	normal
56	male	atypical angina	120	236	false	abnormal
57	female	typical angina	120	354	false	abnormal
57	male	typical angina	140	192	false	abnormal

thalach	exang	oldpeak	slope	ca	target
150	no	2.3	upsloping	0	1
187	no	3.5	upsloping	0	1
172	no	1.4	downsloping	0	1
178	no	0.8	downsloping	0	1
163	yes	0.6	downsloping	0	1
148	no	0.4	flat	0	1

Note. Information about patients who were or were not diagnosed with heart disease"

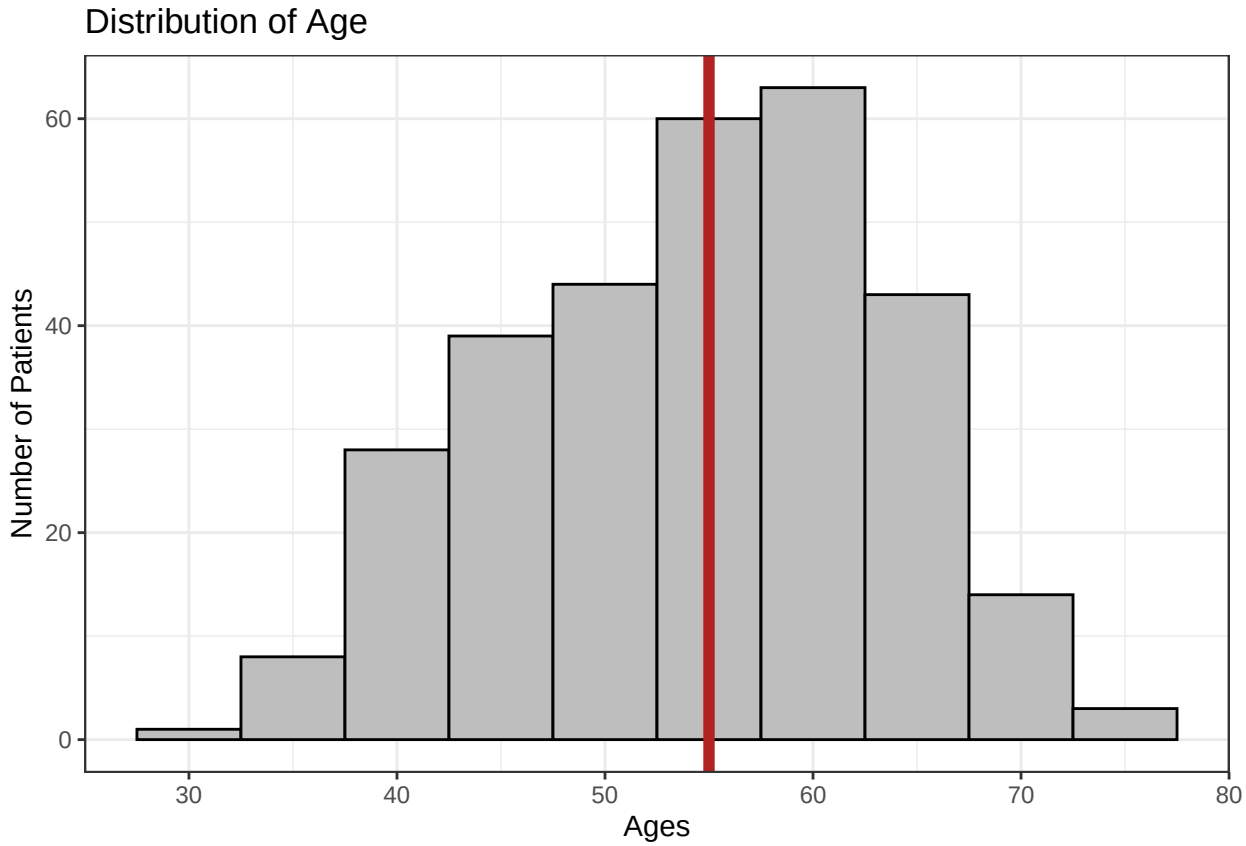
Exploratory Data Analysis

Univariate Analysis

- The following section will examine each variable at a time through summarization and visualizations

Variable: *Age*

- The following histogram looks at the distribution of ages.



- The following table describes the range of ages.

Table 20: Range of Ages Table

Youngest	Median	Oldest
29	55	77

- The following table describes the frequency and relative frequency of age groups.

Table 21: Frequency Distribution Table by Age Groups

Age_Groups	Frequency	Relative_Frequency
(20,29]	1	0.003
(30,39]	15	0.05
(39,40]	3	0.01
(40,49]	69	0.228
(49,50]	7	0.023
(50,59]	118	0.389

Age_Groups	Frequency	Relative_Frequency
(59,60]	11	0.036
(60,69]	69	0.228
(69,70]	4	0.013
(70,79]	6	0.02

Note. This table shows the frequency and relative frequency of age groups. The age group with the highest number of patients are in their 50s.

Variable: *Sex*

- The following bar graph describes the frequency of male and female patients in this data-set.

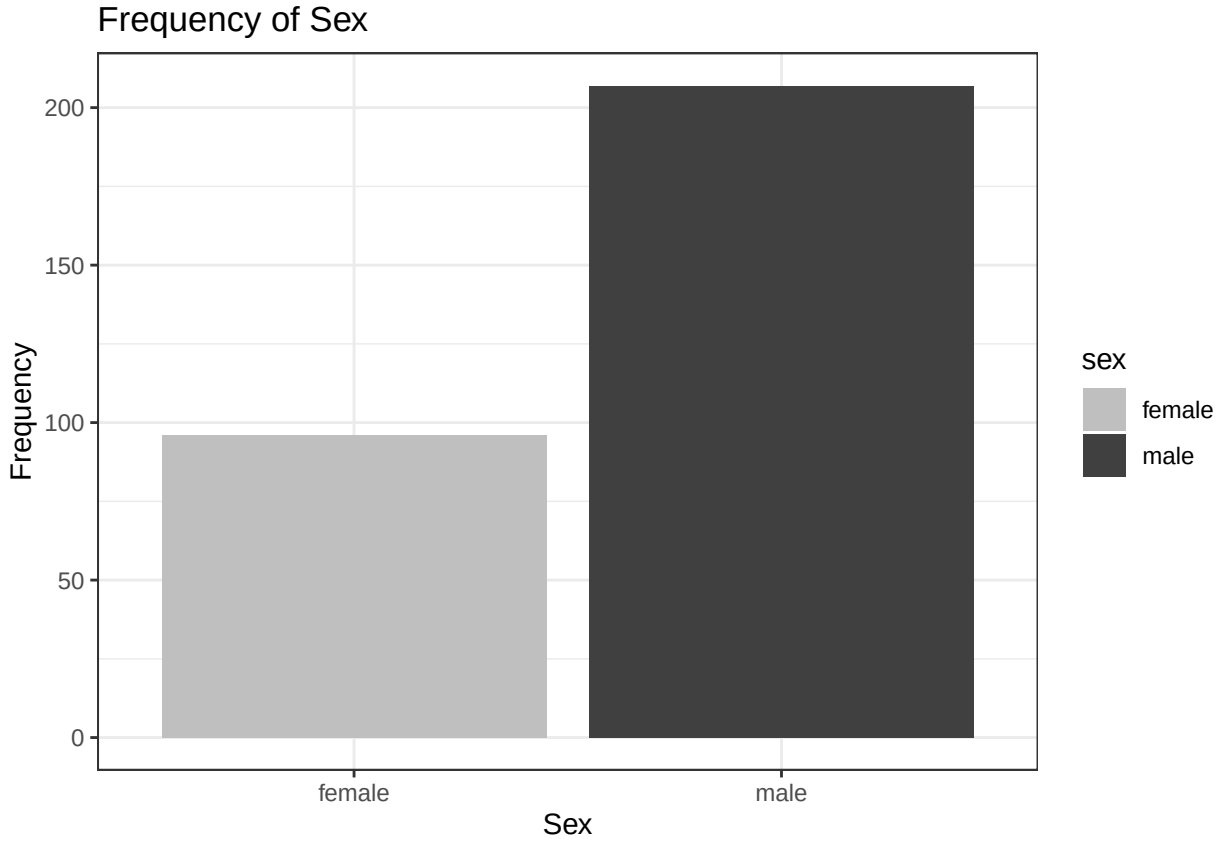


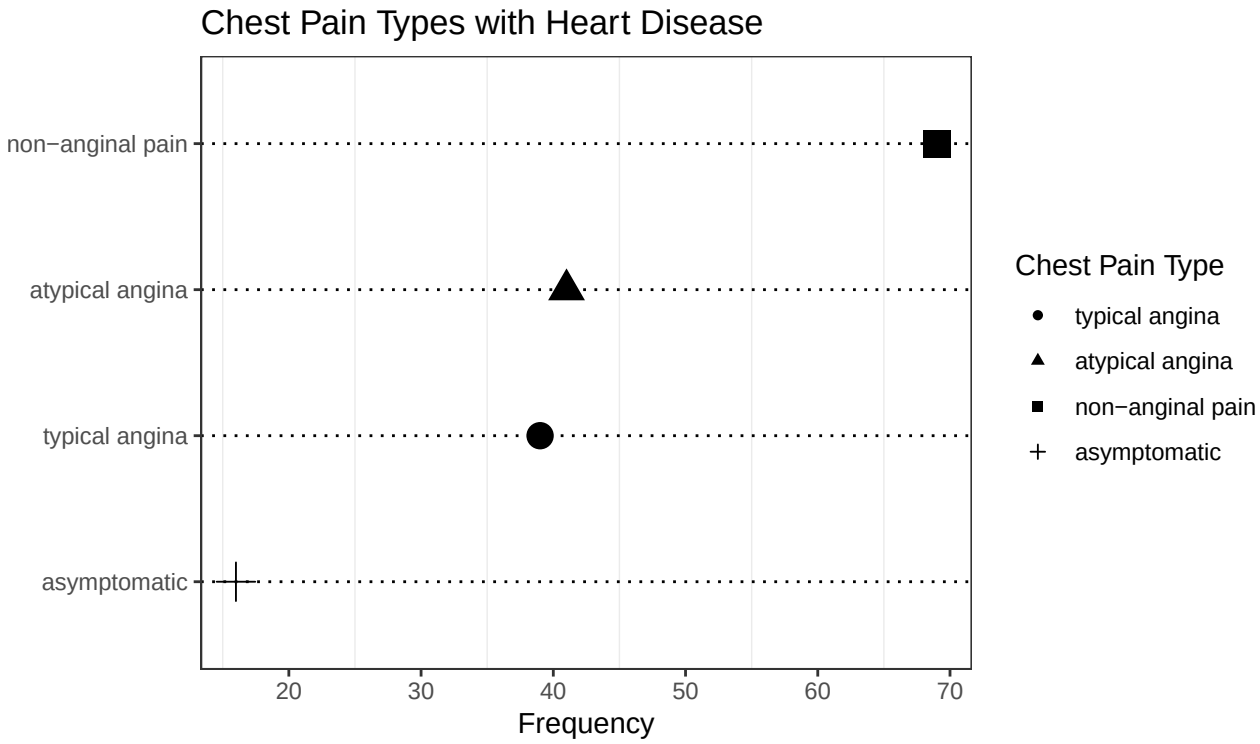
Table 22: Contingency Table for Sex

Sex	Frequency
female	96
male	207
total	303

Note. This is a contingency table that describes the frequency of male and female patients in this data-set

Variable: *Chest Pain*

- The following dot chart looks at the frequency of patients who have heart disease by chest pain type



- The following table looks at the frequency and relative frequency distribution of chest pain types by those who do have the heart disease.

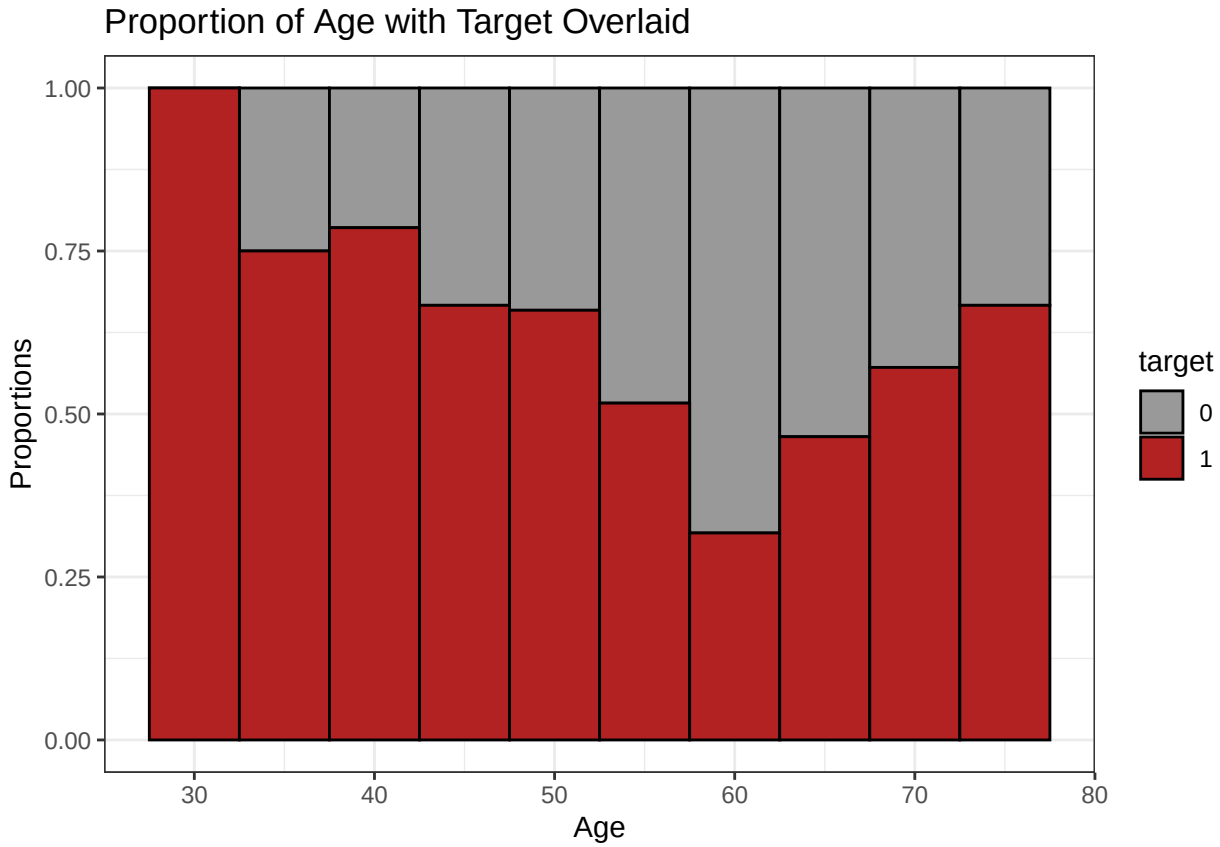
Table 23: Frequency Distribution Table by Chest Pain Type with Heart Disease *Note.* This distribution tables describes that the chest pain type of non-anginal pain has the highest proportion of patients in this data set who also has the heart disease.

Chest Pain Type	Frequency	Relative Frequency
typical angina	39	0.129
atypical angina	41	0.135
non-anginal pain	69	0.228
asymptomatic	16	0.053

Summarization and Visualization of Multivariate Relationships

Exploring Variables: *Age* and *Target*

- The following histogram visualizes the proportions of age groups that has heart disease.



Note. This histogram shows that patients under the age of 50 are more frequent to have a heart disease than those over the age of 50.

- The following creates a table showing the frequency and relative frequency of age groups that has heart disease.

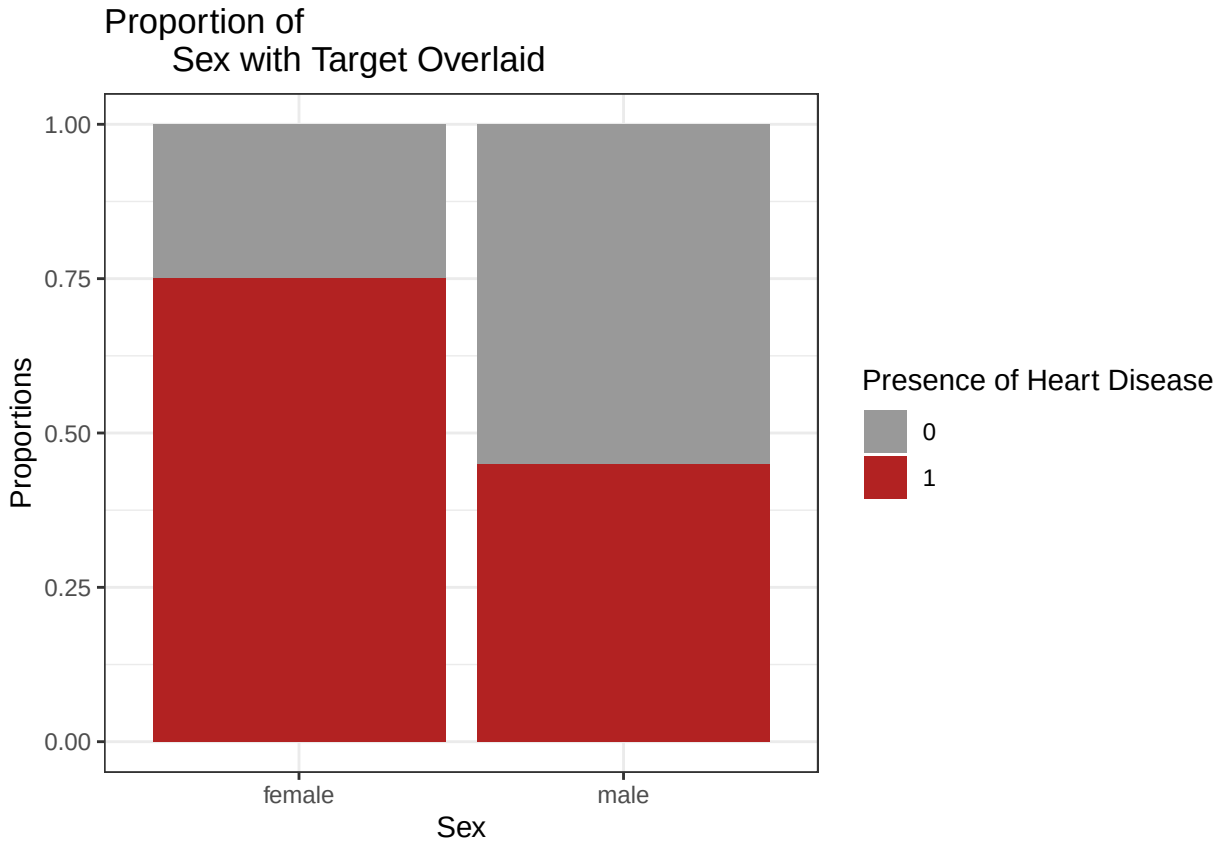
Table 24: Frequency Distribution Table by Age Groups who has Heart Disease

Age_Groups	Frequency	Relative_Frequency
(20,29]	1	0.003
(30,39]	11	0.036
(39,40]	1	0.003
(40,49]	49	0.162
(49,50]	4	0.013
(50,59]	61	0.201
(59,60]	3	0.01
(60,69]	29	0.096
(69,70]	1	0.003
(70,79]	5	0.017

Note. This table shows the frequency and relative frequency of age groups that has heart disease. The age group that has heart disease with the highest number of patients are in their 50s.

Exploring Variables: *Sex* and *Target*

- The following bar graph compare the proportions of sex with the presence of heart disease.



Note. Proportion of Patients by Sex with Diagnosis Overlaid

Table 25: Contingency Table for Sex with Target Classes

	female	male
0	0.25	0.55
1	0.75	0.45

Note. This table summarizes the proportion of sex by heart disease diagnosis. There is a higher proportion of female patients who have heart disease than male patients.