Quantum Heart Prediction

Elvira Khwatenge

2025-03-30

R Markdown

This is an analysis of a heart disease prediction dataset. The goal is to explore the data, build a predictive model, and evaluate its performance.

Data Loading and Preprocessing

First, we load the necessary libraries and read the data.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
data <- read.csv("Heart Prediction Quantum Dataset.csv")
head(data)
```

```
##   Age Gender BloodPressure Cholesterol HeartRate QuantumPatternFeature
## 1  68      1           105         191       107              8.362241
## 2  58      0            97         249        89              9.249002
## 3  44      0            93         190        82              7.942542
## 4  72      1            93         183       101              6.495155
## 5  37      0           145         166       103              7.653900
## 6  50      1           114         271        73              8.631604
##   HeartDisease
## 1            1
## 2            0
## 3             1
## 4             1
## 5             1
## 6             0
```

```
str(data)
```

```
## 'data.frame':    500 obs. of  7 variables:
##  $ Age                  : int  68 58 44 72 37 50 68 48 52 40 ...
##  $ Gender               : int  1 0 0 1 0 1 1 0 0 1 ...
##  $ BloodPressure        : int  105 97 93 93 145 114 156 156 116 121 ...
##  $ Cholesterol          : int  191 249 190 183 166 271 225 236 266 255 ...
##  $ HeartRate            : int  107 89 82 101 103 73 73 61 114 96 ...
##  $ QuantumPatternFeature: num  8.36 9.25 7.94 6.5 7.65 ...
##  $ HeartDisease         : int  1 0 1 1 1 0 1 0 0 0 ...
```

```
sum(is.na(data))
```

```
## [1] 0
```

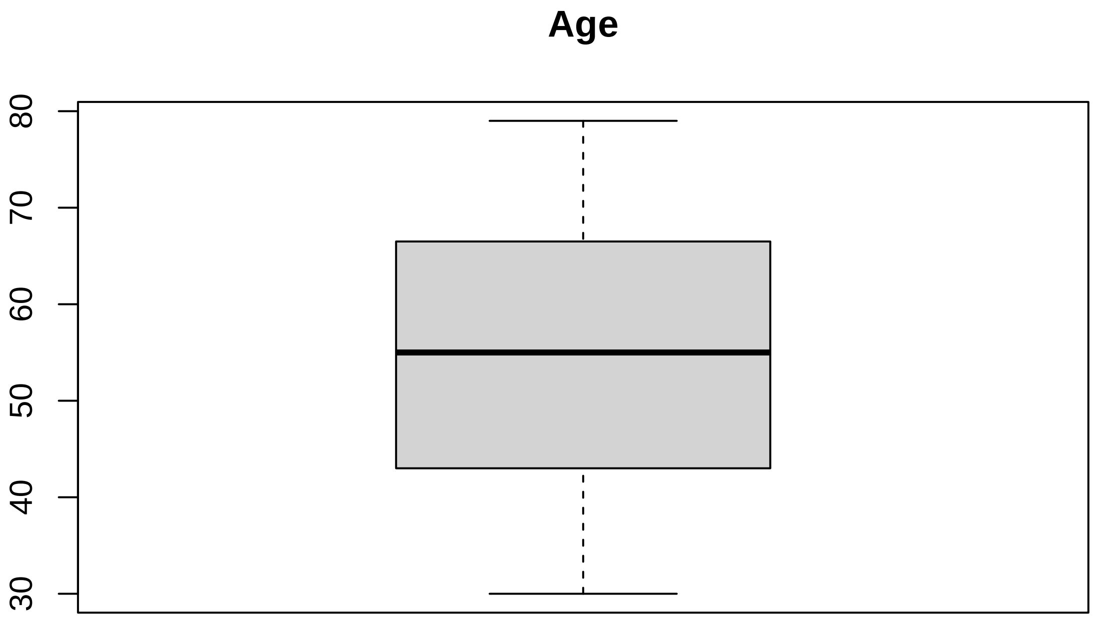
Data Processing

preprocess the data by converting categorical variables to factors, checking for outliers, and scaling numerical features.

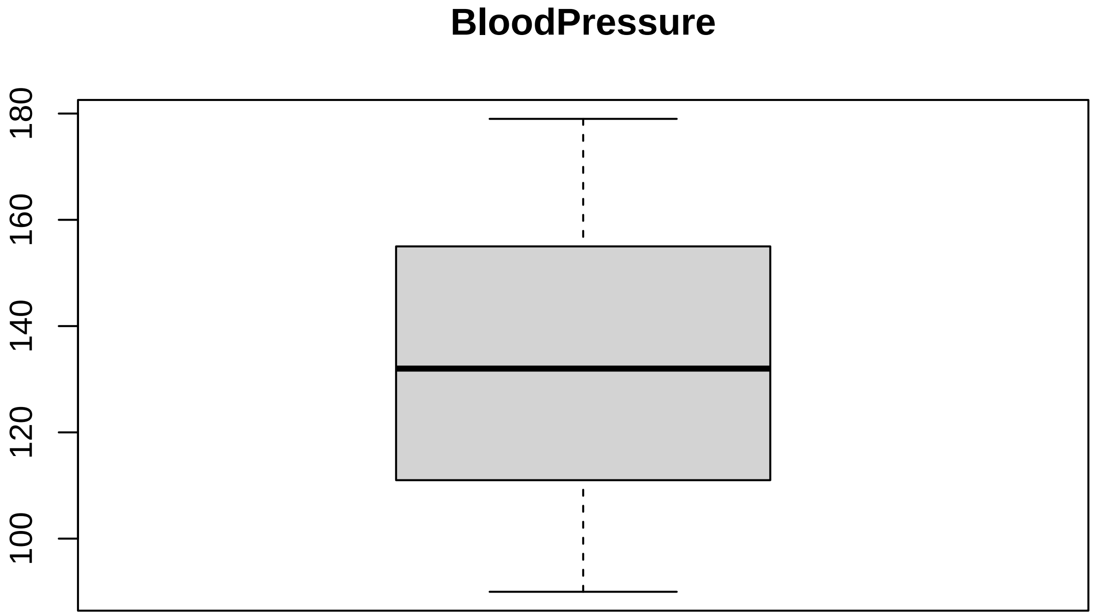
```
data$Gender <- as.factor(data$Gender)
data$HeartDisease <- as.factor(data$HeartDisease)

summary(data)
```

```
##       Age        Gender  BloodPressure    Cholesterol      HeartRate     
##  Min.   :30.00   0:266   Min.   : 90.0   Min.   :150.0   Min.   : 60.00  
##  1st Qu.:43.00   1:234   1st Qu.:111.0   1st Qu.:183.8   1st Qu.: 73.00  
##  Median :55.00           Median :132.0   Median :221.0   Median : 89.00  
##  Mean   :54.86           Mean   :132.9   Mean   :221.5   Mean   : 88.77  
##  3rd Qu.:66.25           3rd Qu.:155.0   3rd Qu.:258.0   3rd Qu.:104.00  
##  Max.   :79.00           Max.   :179.0   Max.   :299.0   Max.   :119.00  
##  QuantumPatternFeature HeartDisease
##  Min.   : 6.165        0:200       
##  1st Qu.: 7.676        1:300       
##  Median : 8.323                    
##  Mean   : 8.317                    
##  3rd Qu.: 8.936                    
##  Max.   :10.785                    
```

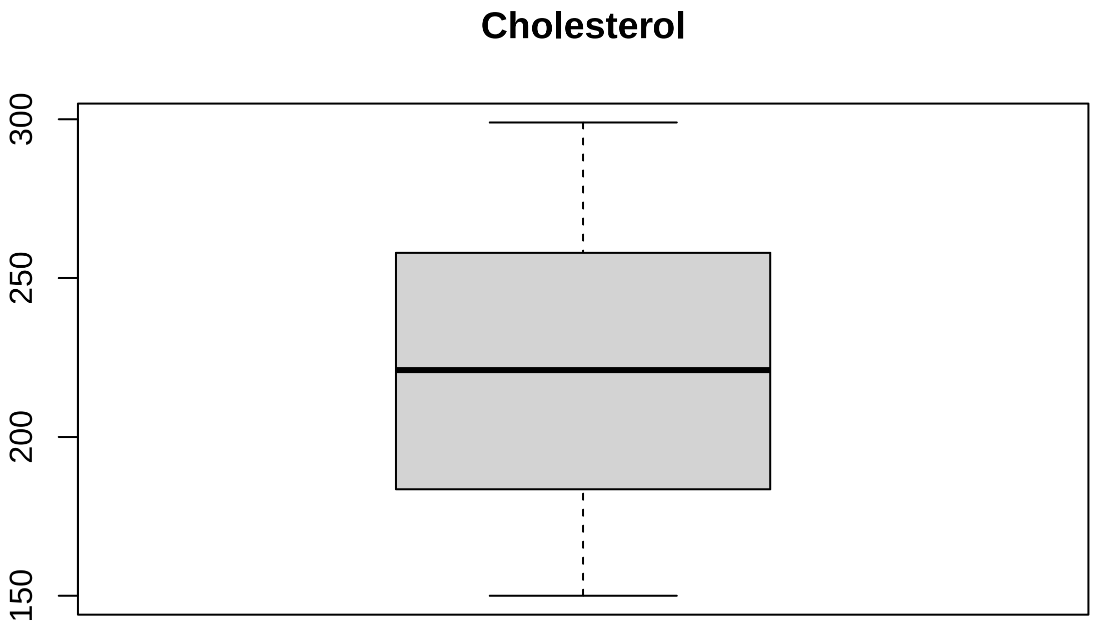
```
boxplot(data$Age, main="Age")
```



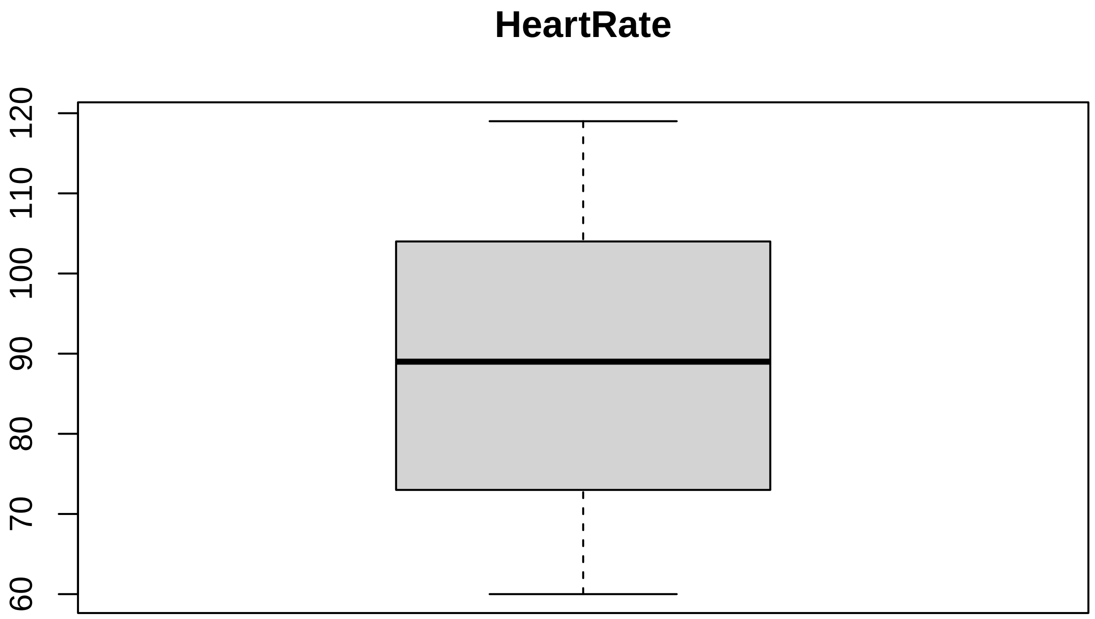
```r
boxplot(data$BloodPressure, main="BloodPressure")
```



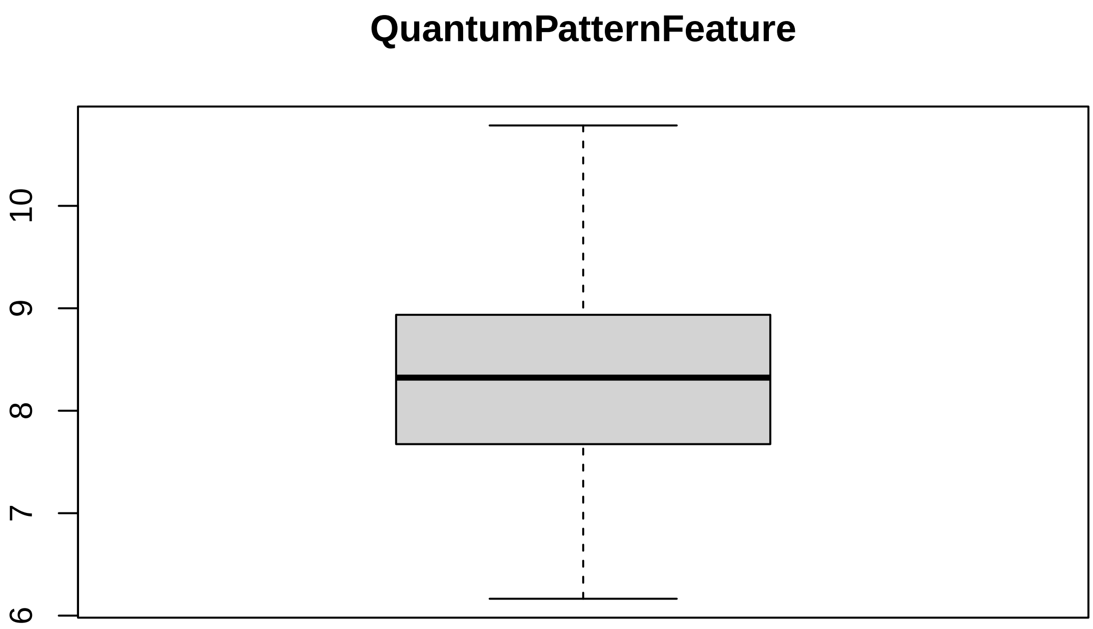
```r
boxplot(data$Cholesterol, main="Cholesterol")
```



```r
boxplot(data$HeartRate, main="HeartRate")
```



```r
boxplot(data$QuantumPatternFeature, main="QuantumPatternFeature")
```



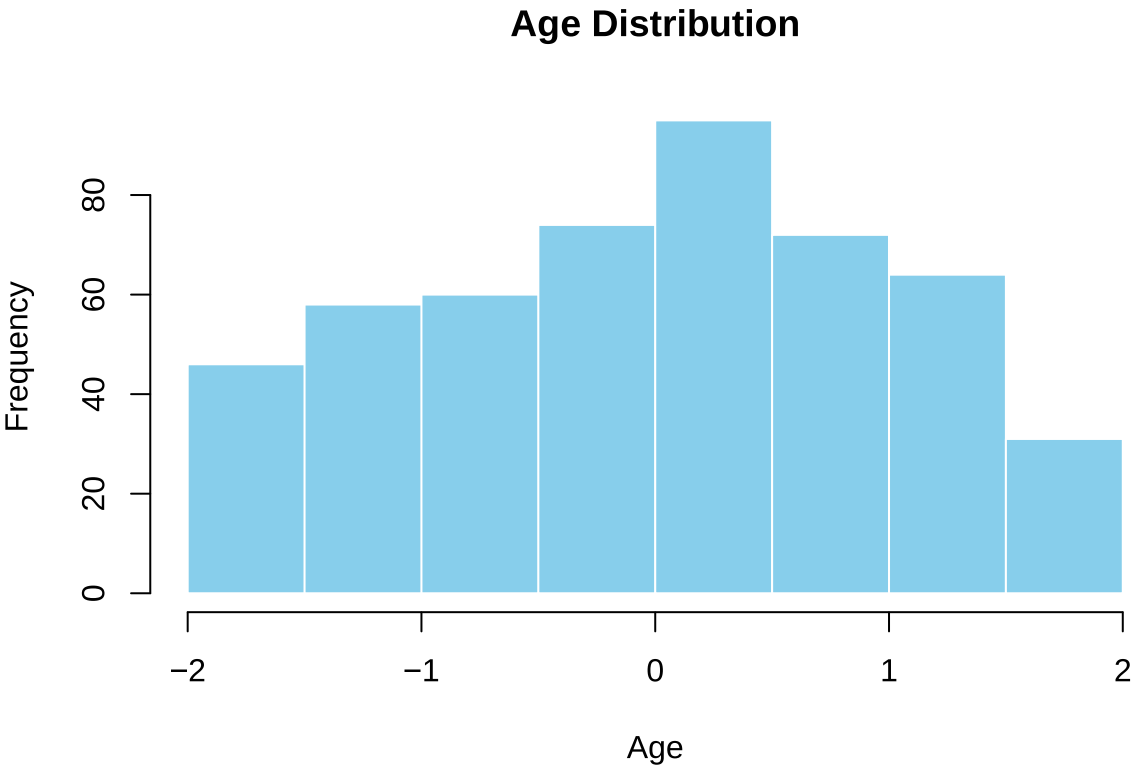
```
Q1 <- quantile(data$Cholesterol, 0.25)
Q3 <- quantile(data$Cholesterol, 0.75)
IQR <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

data <- data[data$Cholesterol >= lower_bound & data$Cholesterol <= upper_bound, ]

numerical_features <- c("Age", "BloodPressure", "Cholesterol", "HeartRate", "QuantumPatternFeature")
data[numerical_features] <- scale(data[numerical_features])

head(data)
```

```
##          Age Gender BloodPressure Cholesterol   HeartRate QuantumPatternFeature
## 1  0.9176386      1    -1.0550933  -0.6953369  1.04689080            0.04875145
## 2  0.2190708      0    -1.3579112   0.6269431  0.01343493            1.01301086
## 3 -0.7589240      0    -1.5093202  -0.7181348 -0.38846458           -0.40762679
## 4  1.1970657      1    -1.5093202  -0.8777203  0.70240551           -1.98150828
## 5 -1.2479214      0     0.4589963  -1.2652852  0.81723394           -0.72149462
## 6 -0.3397833      1    -0.7144232   1.1284976 -0.90519252            0.34165582
##   HeartDisease
## 1            1
## 2            0
## 3            1
## 4            1
## 5            1
## 6            0
```

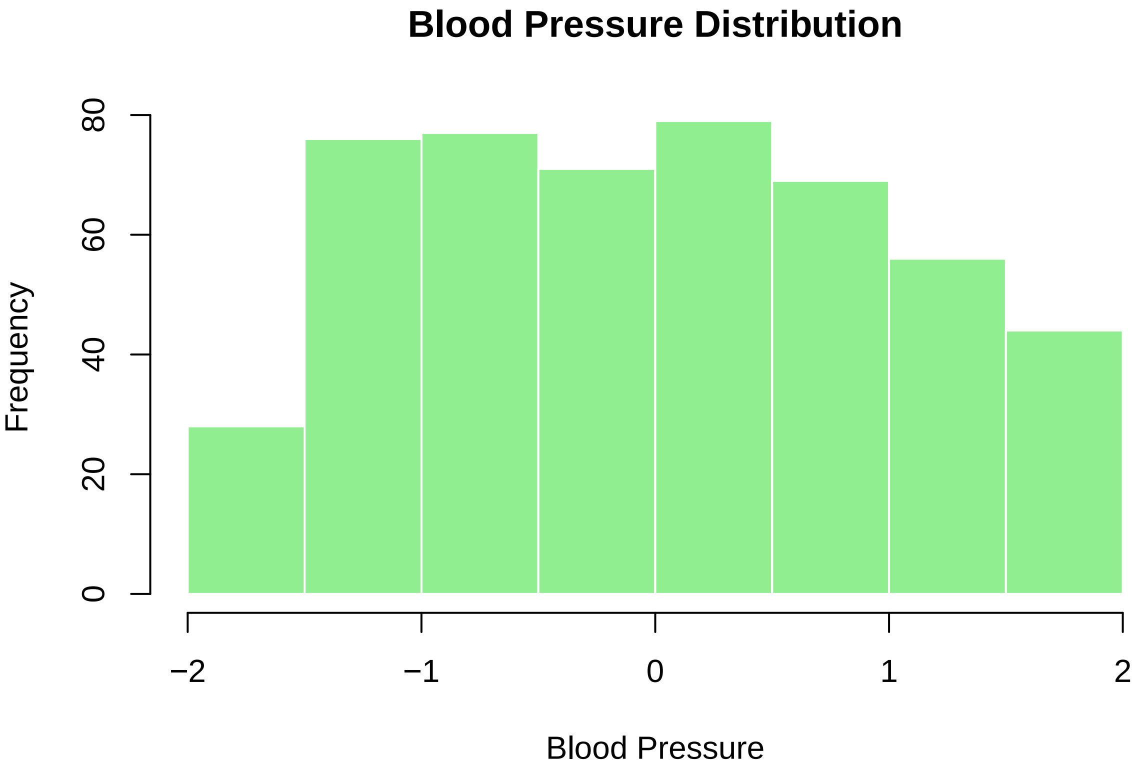
###Exploratory Data Analysis (EDA)

perform EDA to understand the relationships between variables.

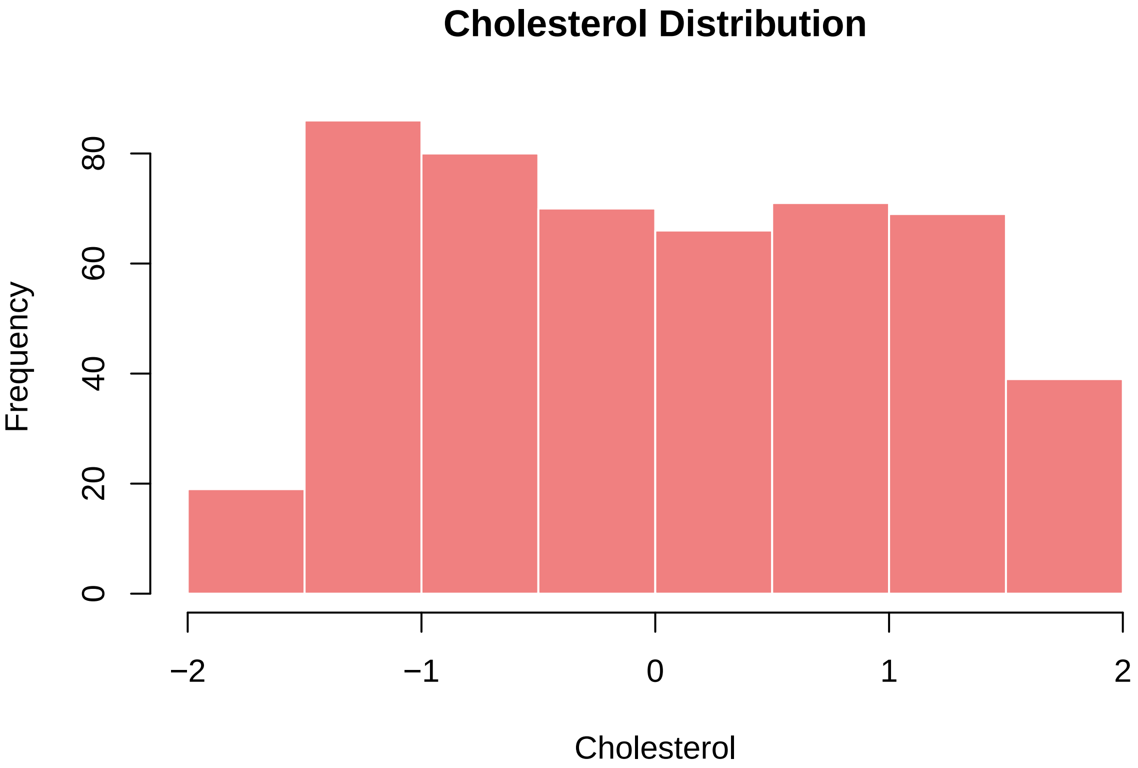
```
hist(data$Age, main="Age Distribution", xlab="Age", col="skyblue", border="white")
```



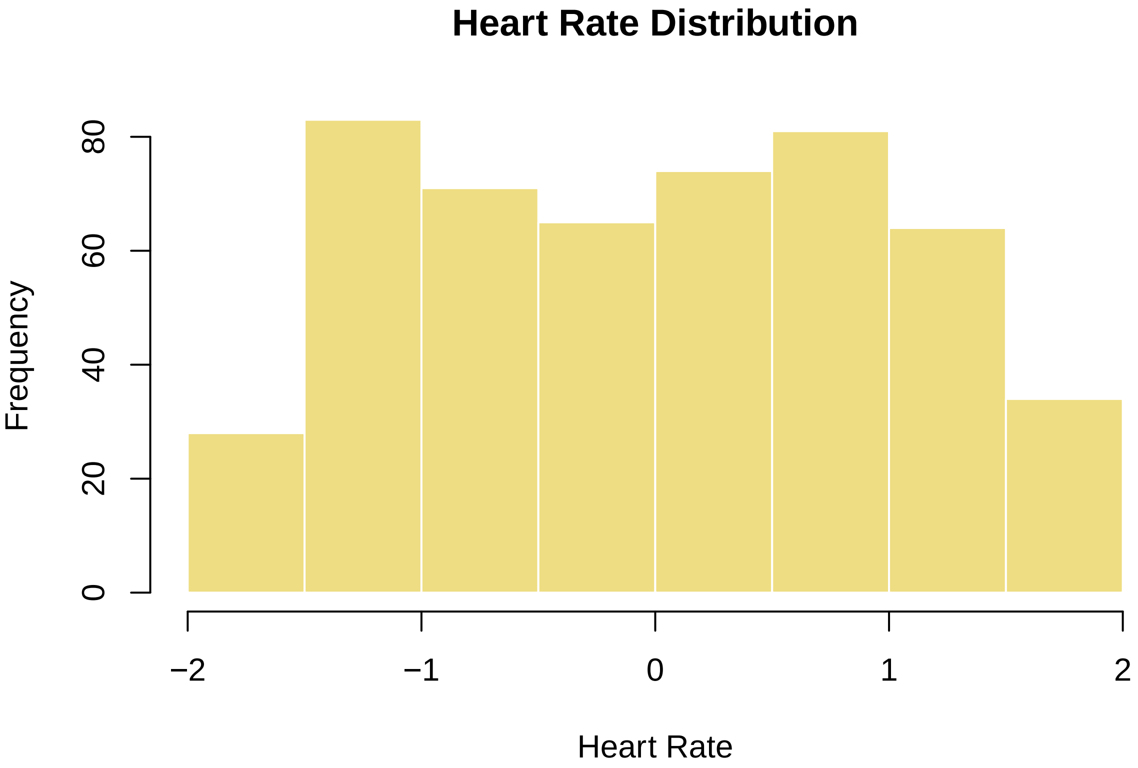
```
hist(data$BloodPressure, main="Blood Pressure Distribution", xlab="Blood Pressure", col="lightgreen", bo
```



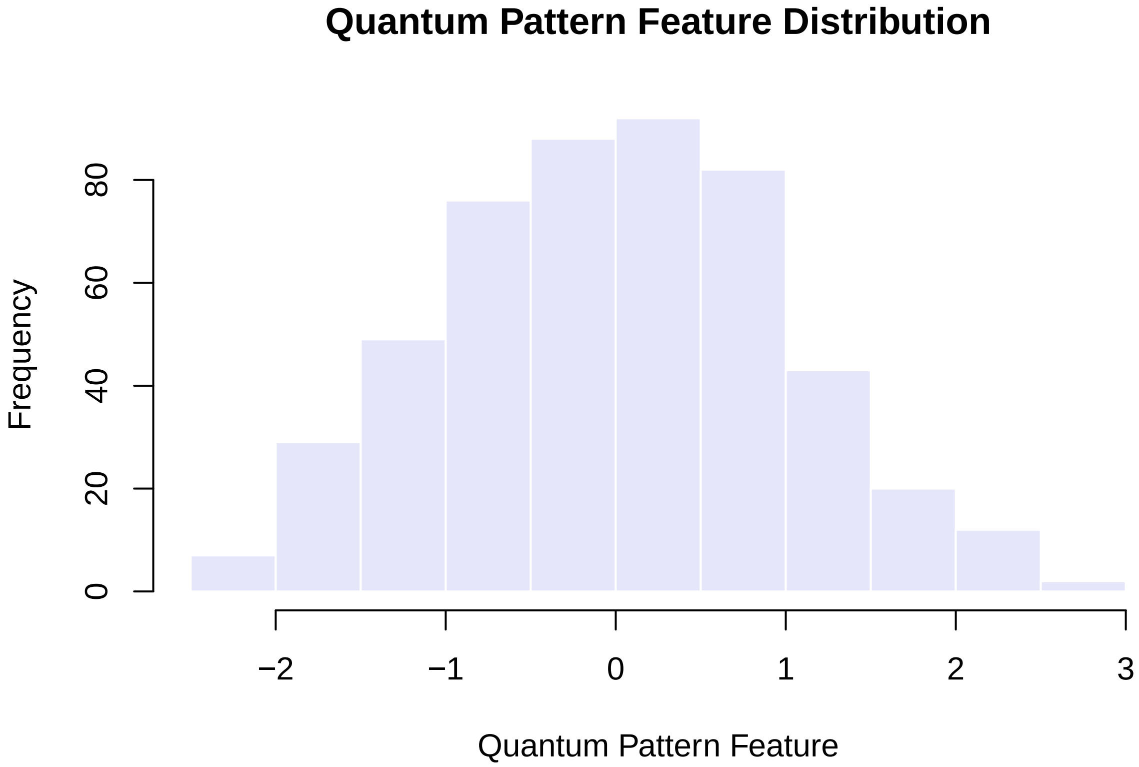
```r
hist(data$Cholesterol, main="Cholesterol Distribution", xlab="Cholesterol", col="lightcoral", border="wh
```



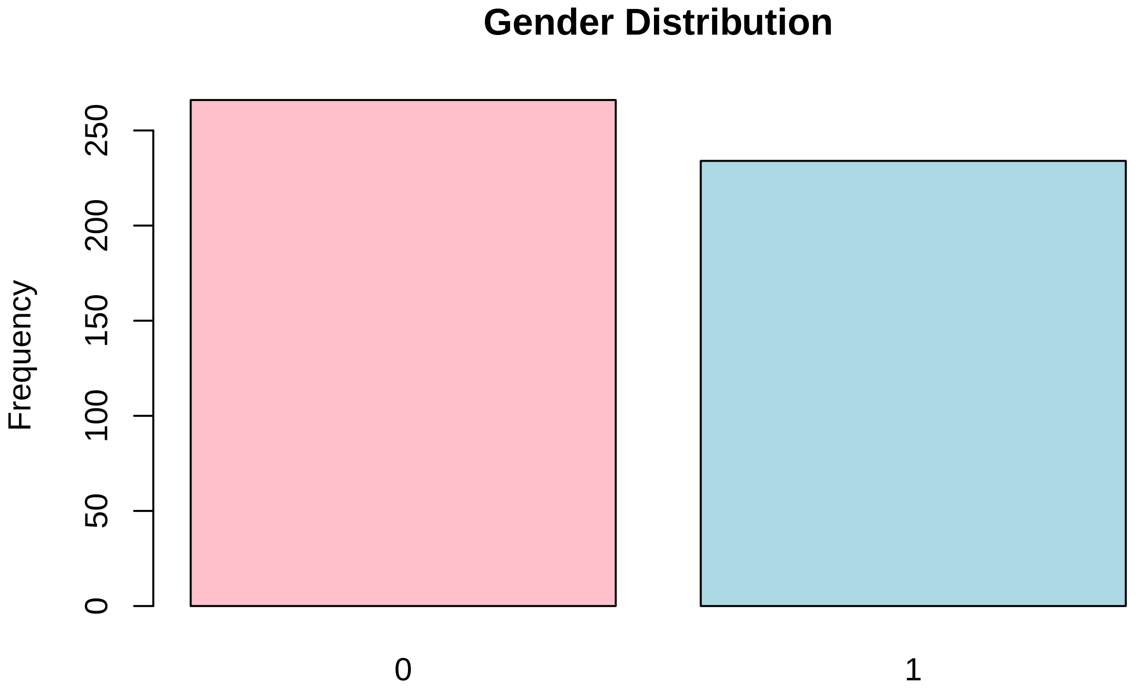
```r
hist(data$HeartRate, main="Heart Rate Distribution", xlab="Heart Rate", col="lightgoldenrod", border="wh
```



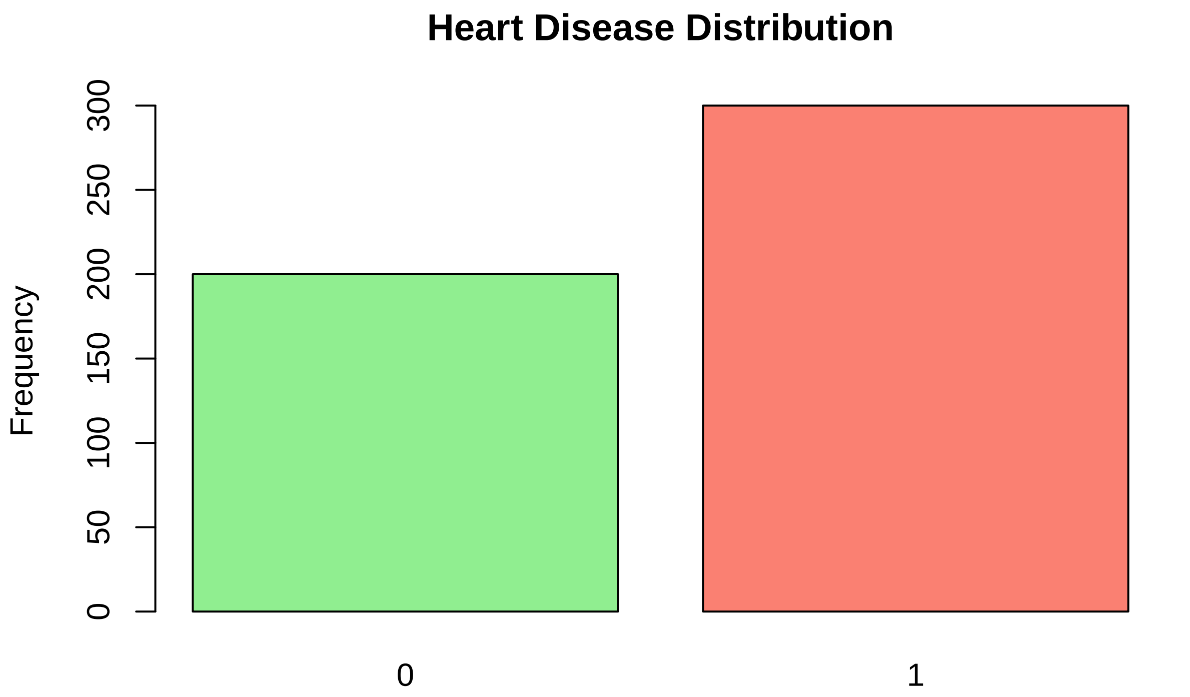
```r
hist(data$QuantumPatternFeature, main="Quantum Pattern Feature Distribution", xlab="Quantum Pattern Fea
```



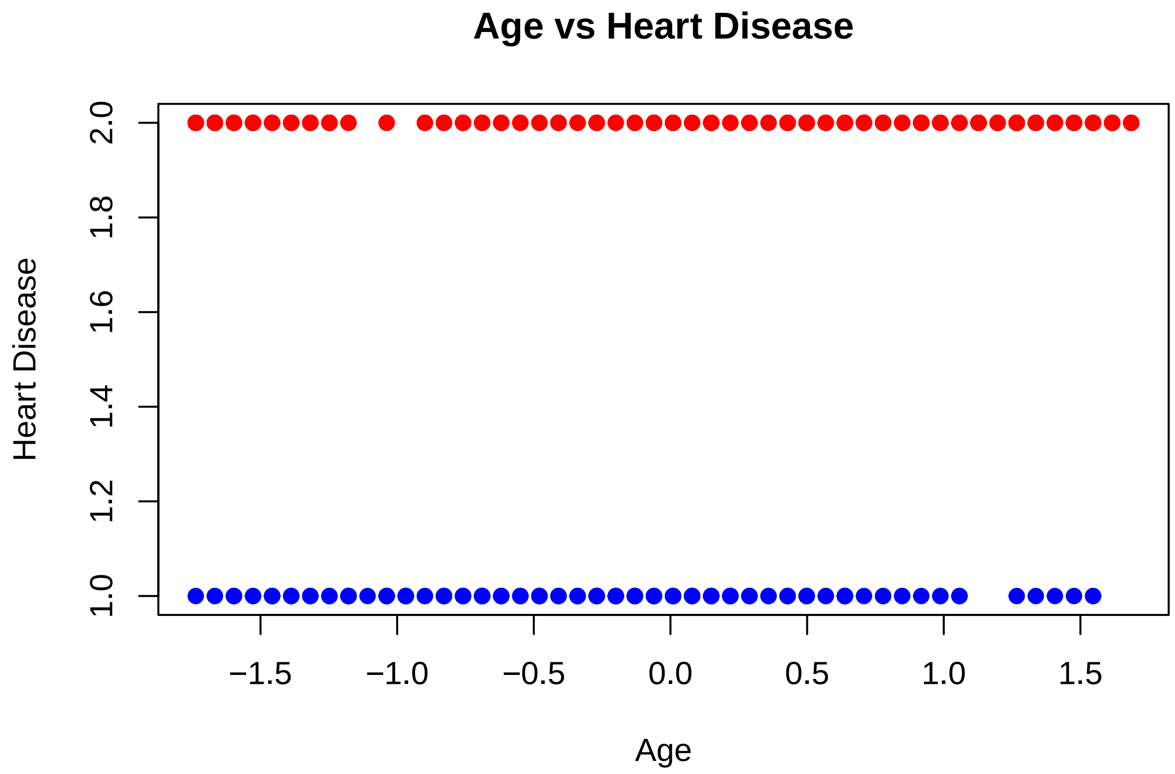
```r
# Bar plots with colors
plot(data$Gender, main="Gender Distribution", ylab="Frequency", col=c("pink", "lightblue"))
```



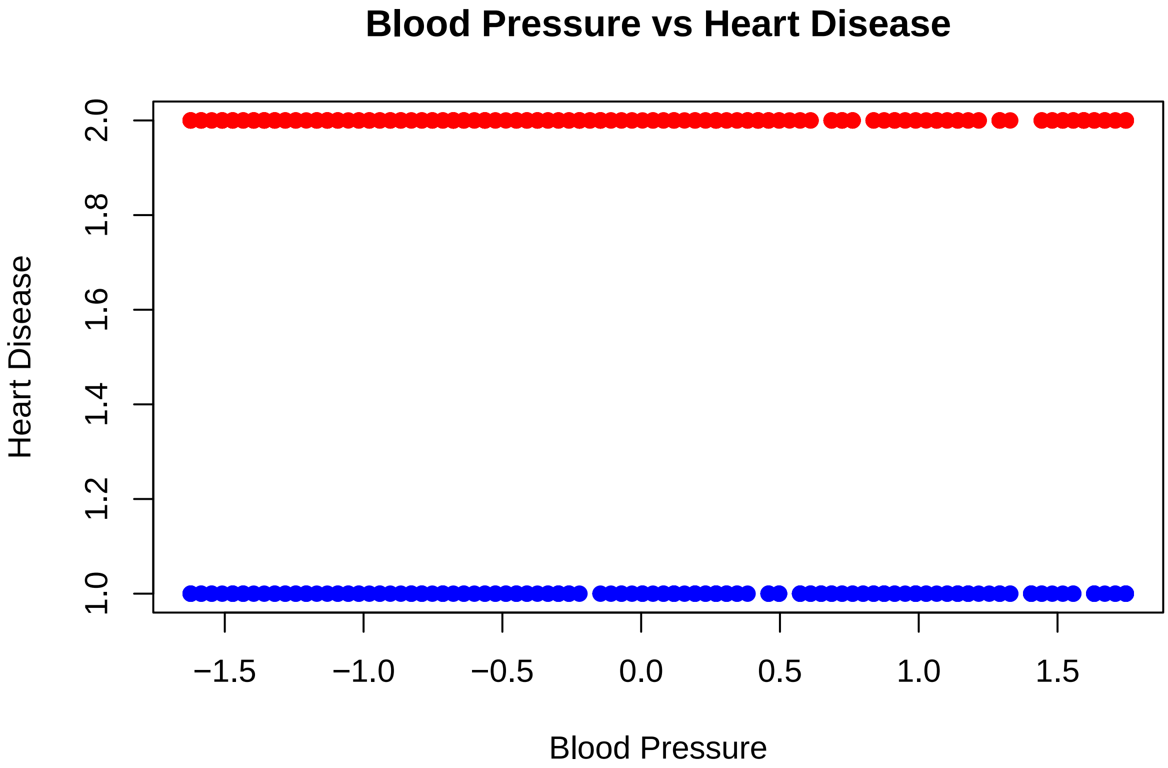
```
plot(data$HeartDisease, main="Heart Disease Distribution", ylab="Frequency", col=c("lightgreen", "salmon
```



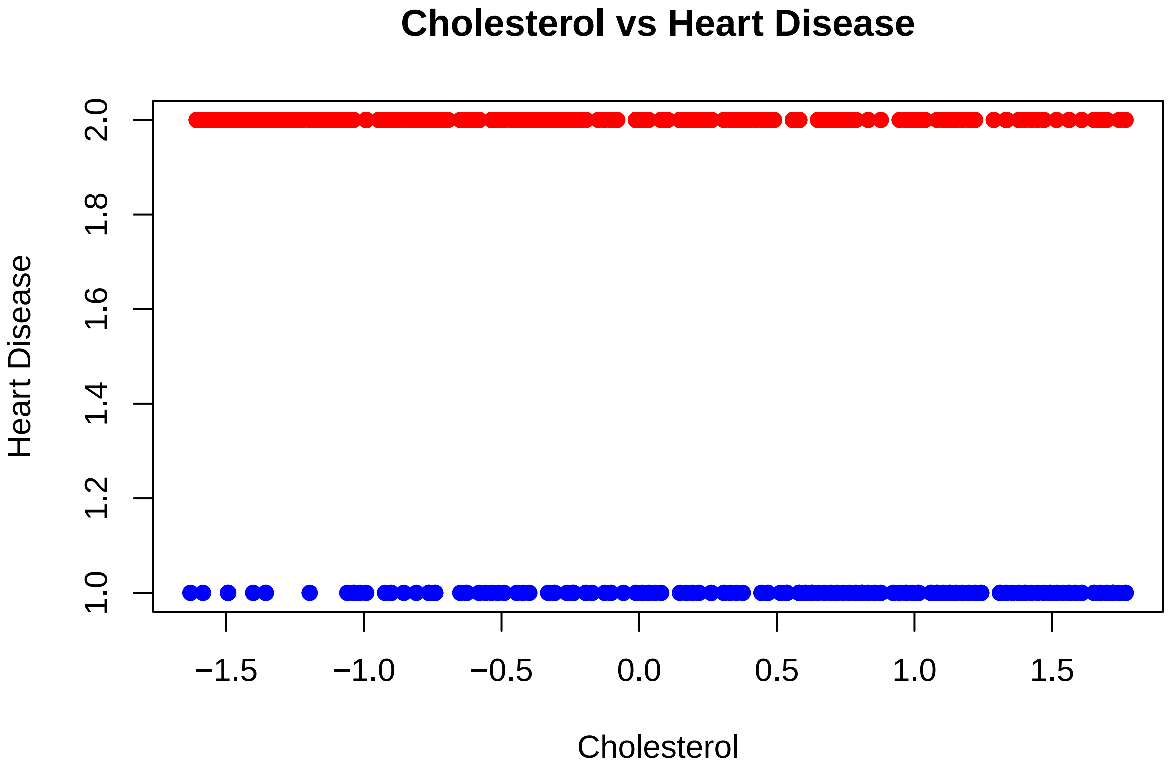
```
# Scatter plots with colors
plot(data$Age, data$HeartDisease, main="Age vs Heart Disease", xlab="Age", ylab="Heart Disease", col=ifel
```



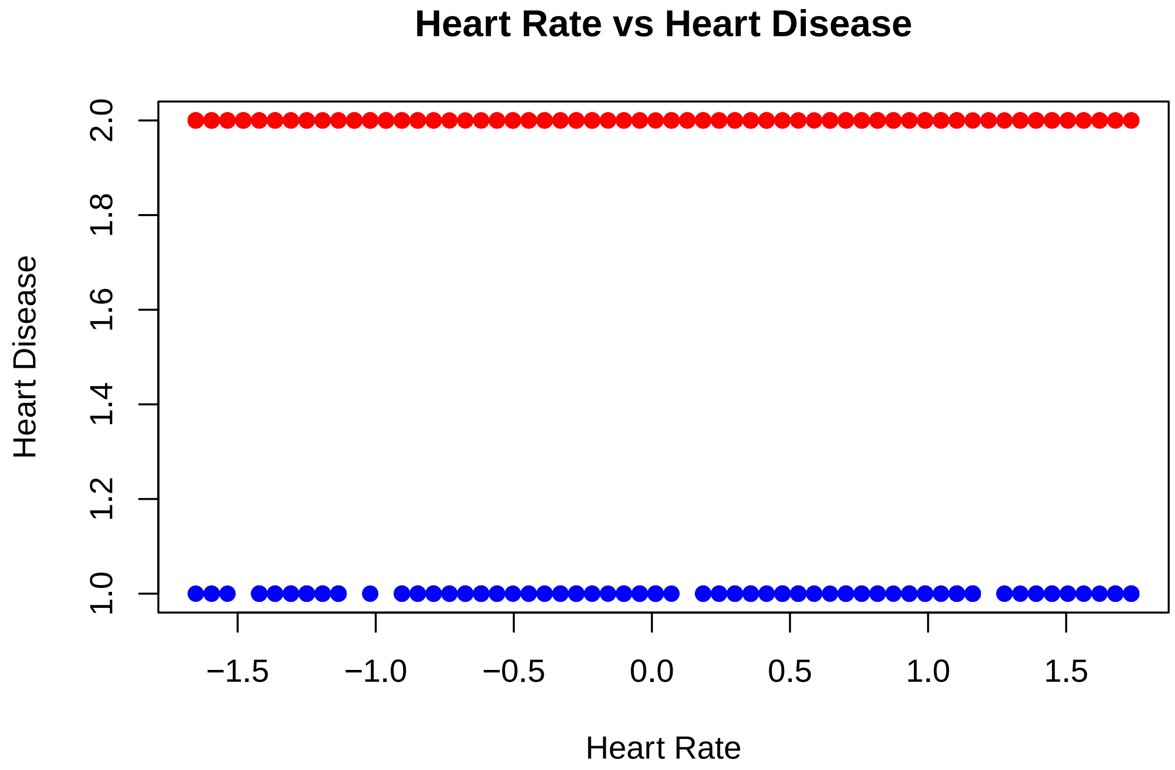
```
plot(data$BloodPressure, data$HeartDisease, main="Blood Pressure vs Heart Disease", xlab="Blood Pressur
```



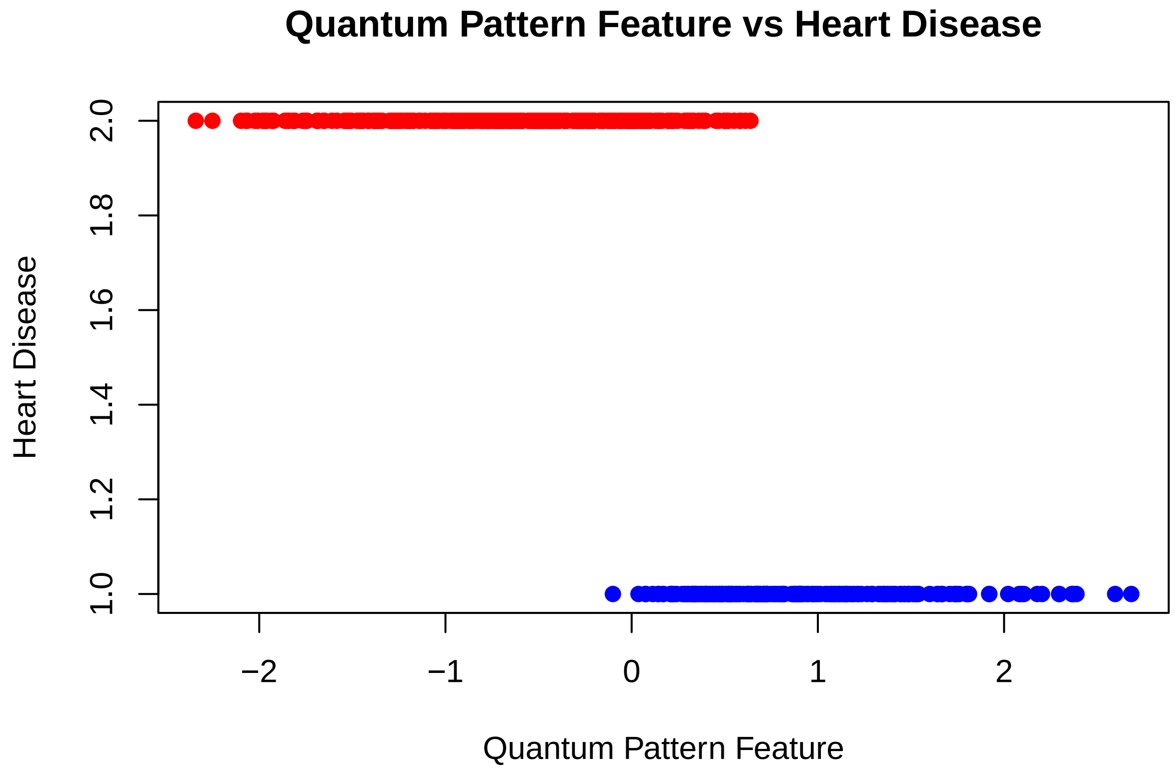
```
plot(data$Cholesterol, data$HeartDisease, main="Cholesterol vs Heart Disease", xlab="Cholesterol", ylab=
```



```
plot(data$HeartRate, data$HeartDisease, main="Heart Rate vs Heart Disease", xlab="Heart Rate", ylab="Hea
```



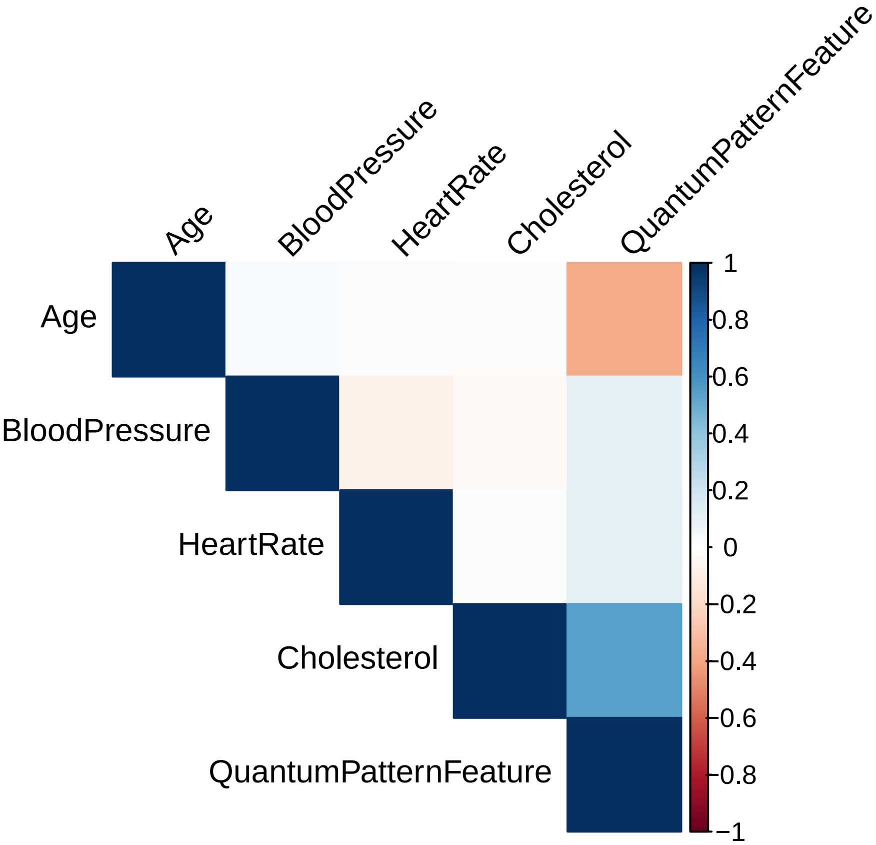
```
plot(data$QuantumPatternFeature, data$HeartDisease, main="Quantum Pattern Feature vs Heart Disease", xla
```



```
# Select numeric columns and handle missing values
numerical_features <- c("Age", "BloodPressure", "Cholesterol", "HeartRate", "QuantumPatternFeature")
numeric_data <- data[, numerical_features]
numeric_data <- na.omit(numeric_data)

# Remove columns with zero variance (if any)
numeric_data <- numeric_data[, apply(numeric_data, 2, var) != 0]

# Compute correlation matrix
correlation_matrix <- cor(numeric_data, use = "complete.obs")

# Visualize with corrplot
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
corrplot(correlation_matrix, method = "color", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



Model Building

Build a logistic regression model to predict heart disease

```
set.seed(123)
train_index <- createDataPartition(data$HeartDisease, p = 0.7, list = FALSE)
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
```

```r
# Train the logistic regression model
model <- glm(HeartDisease ~ ., data = train_data, family = "binomial")

# Summarize the model
summary(model)
```

```
##
## Call:
## glm(formula = HeartDisease ~ ., family = "binomial", data = train_data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.7752     0.4391   4.043 5.27e-05 ***
## Age                    -0.2477     0.2749  -0.901    0.368
## Gender1                 0.2093     0.5015   0.417    0.676
## BloodPressure          -0.1250     0.2349  -0.532    0.595
## Cholesterol             0.1128     0.2723   0.414    0.679
## HeartRate              -0.2070     0.2441  -0.848    0.397
## QuantumPatternFeature  -7.6702     1.0599  -7.237 4.60e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 471.11  on 349  degrees of freedom
## Residual deviance: 110.27  on 343  degrees of freedom
## AIC: 124.27
##
## Number of Fisher Scoring iterations: 8
```

```r
# Make predictions on the test data
predictions <- predict(model, newdata = test_data, type = "response")
predicted_classes <- ifelse(predictions > 0.5, 1, 0)

# Evaluate the model
confusionMatrix(as.factor(predicted_classes), test_data$HeartDisease)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 58  8
##          1  2 82
##
##                Accuracy : 0.9333
##                  95% CI : (0.8808, 0.9676)
##     No Information Rate : 0.6
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8634
##
##  Mcnemar's Test P-Value : 0.1138
##
##             Sensitivity : 0.9667
##             Specificity : 0.9111
##          Pos Pred Value : 0.8788
##          Neg Pred Value : 0.9762
##              Prevalence : 0.4000
##          Detection Rate : 0.3867
##    Detection Prevalence : 0.4400
##       Balanced Accuracy : 0.9389
##
##        'Positive' Class : 0
##
```

Discussion and Interpretation

##Demographic and Health Indicators

The Age Distribution histogram shows a relatively normal distribution of ages, with a slight right skew suggesting that while the majority of the population is middle-aged, there's a notable presence of older individuals in the dataset. The blood pressure histogram reveals a bimodal distribution, indicating two distinct peaks. This could suggest two subgroups within the population, possibly representing normal and high blood pressure groups. The cholesterol levels show a right-skewed distribution, implying that while most individuals have cholesterol levels within a normal range, there's a significant subset with higher cholesterol levels. The heart rate histogram appears normally distributed, suggesting that most individuals in the dataset have heart rates within the expected range. Quantum Pattern Feature Distribution feature shows a multimodal distribution, which could indicate distinct subgroups or patterns within the population that might be relevant to heart disease prediction.

Gender and Heart Disease Distribution The gender bar plot shows the proportion of males to females in the dataset for understanding potential gender-based differences in heart disease risk. Heart Disease Distribution illustrates the prevalence of heart disease in the sample. The relative sizes of the bars provide insight into the overall heart disease burden in the population studied.

Relationships Between Variables and Heart Disease 8-12. These scatter plots (Age, Blood Pressure, Cholesterol, Heart Rate, and Quantum Pattern Feature vs Heart Disease) use color coding (red for heart disease, blue for no heart disease) to visualize the relationship between each variable and heart disease occurrence. Patterns or clusters in these plots can reveal important risk factors or predictive indicators for heart disease.

Correlation Matrix heatmap visualizes the correlations between numerical features. Stronger correlations (darker colors) suggest potential multicollinearity in predictors, which is important for model interpretation and feature selection.

Model Performance

The logistic regression model's performance, as indicated by the confusion matrix, provides insights into its predictive accuracy, sensitivity, and specificity. This information is crucial for assessing the model's reliability in predicting heart disease.

Implications For Academia these visualizations and the model provide a foundation for further research into heart disease predictors, especially the novel Quantum Pattern Feature. The multivariate nature of heart disease risk is clearly illustrated, encouraging interdisciplinary approaches.

For Industry the insights gained can guide the development of more sophisticated diagnostic tools and personalized health monitoring devices. The Quantum Pattern Feature, in particular, might represent a new frontier in health tech innovation.

For Policy Makers the distributions of key health indicators and their relationships with heart disease can inform public health strategies. For instance, targeted interventions for high-risk groups identified through these analyses could be prioritized.