Final Project

Group 4

2024-04-30

Data Loading

```
# Load necessary library
library(readr)

# Load the dataset
data <- read_csv("/Users/cv/Downloads/heart.csv")

## Rows: 918 Columns: 12
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr (5): Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope
## dbl (7): Age, RestingBP, Cholesterol, FastingBS, MaxHR, Oldpeak,
HeartDisease
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

# Display the first few rows
head(data)

## # A tibble: 6 × 12
##     Age Sex   ChestPainType RestingBP Cholesterol FastingBS RestingECG
MaxHR
##   <dbl> <chr> <chr>             <dbl>       <dbl>     <dbl> <chr>
<dbl>
## 1    40 M     ATA                 140         289         0 Normal
172
## 2    49 F     NAP                 160         180         0 Normal
156
## 3    37 M     ATA                 130         283         0 ST
98
## 4    48 F     ASY                 138         214         0 Normal
108
## 5    54 M     NAP                 150         195         0 Normal
122
## 6    39 M     NAP                 120         339         0 Normal
170
```

```
## # ℹ 4 more variables: ExerciseAngina <chr>, Oldpeak <dbl>, ST_Slope <chr>,
## #   HeartDisease <dbl>
```

Dropping irrelvant columns which are not the part of our RQ

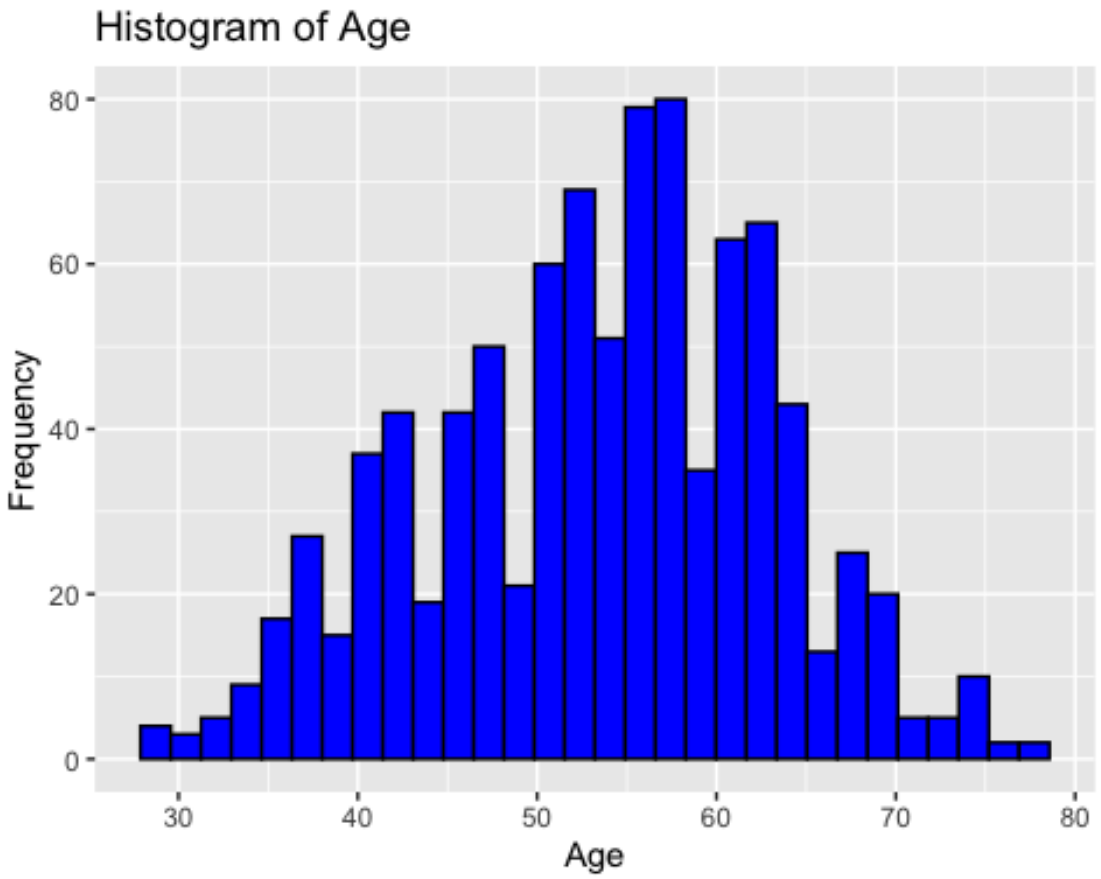
```
# Drop unnecessary columns
df <- data[, c("Age", "Sex", "Cholesterol", "RestingBP", "HeartDisease")]

# Check the structure of the cleaned dataset
head(df)
```

```
## # A tibble: 6 × 5
##     Age Sex   Cholesterol RestingBP HeartDisease
##   <dbl> <chr>       <dbl>     <dbl>        <dbl>
## 1    40 M             289       140            0
## 2    49 F             180       160            1
## 3    37 M             283       130            0
## 4    48 F             214       138            1
## 5    54 M             195       150            0
## 6    39 M             339       120            0
```

Visualisation to understand the data better

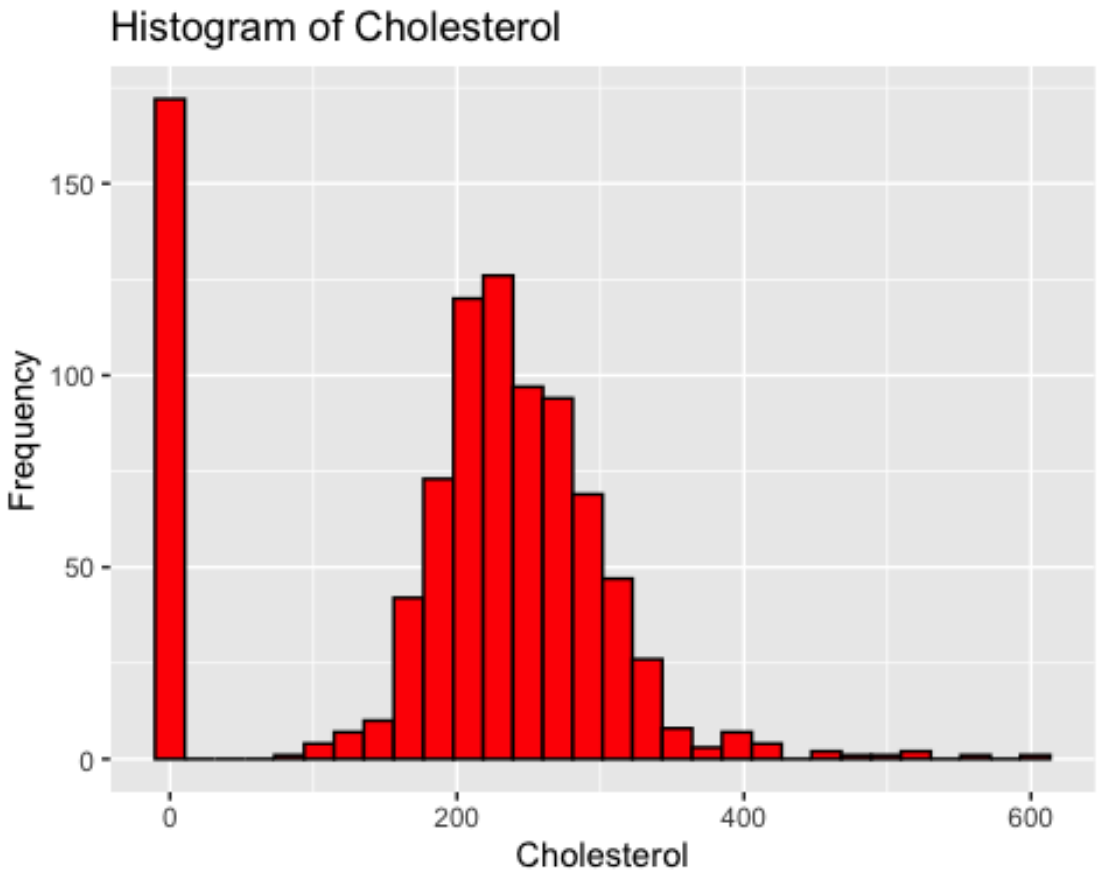
1. Age

```
library(ggplot2)

# Histogram for Age
ggplot(df, aes(x = Age)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  ggtitle("Histogram of Age") +
  xlab("Age") +
  ylab("Frequency")
```



2. Cholesterol

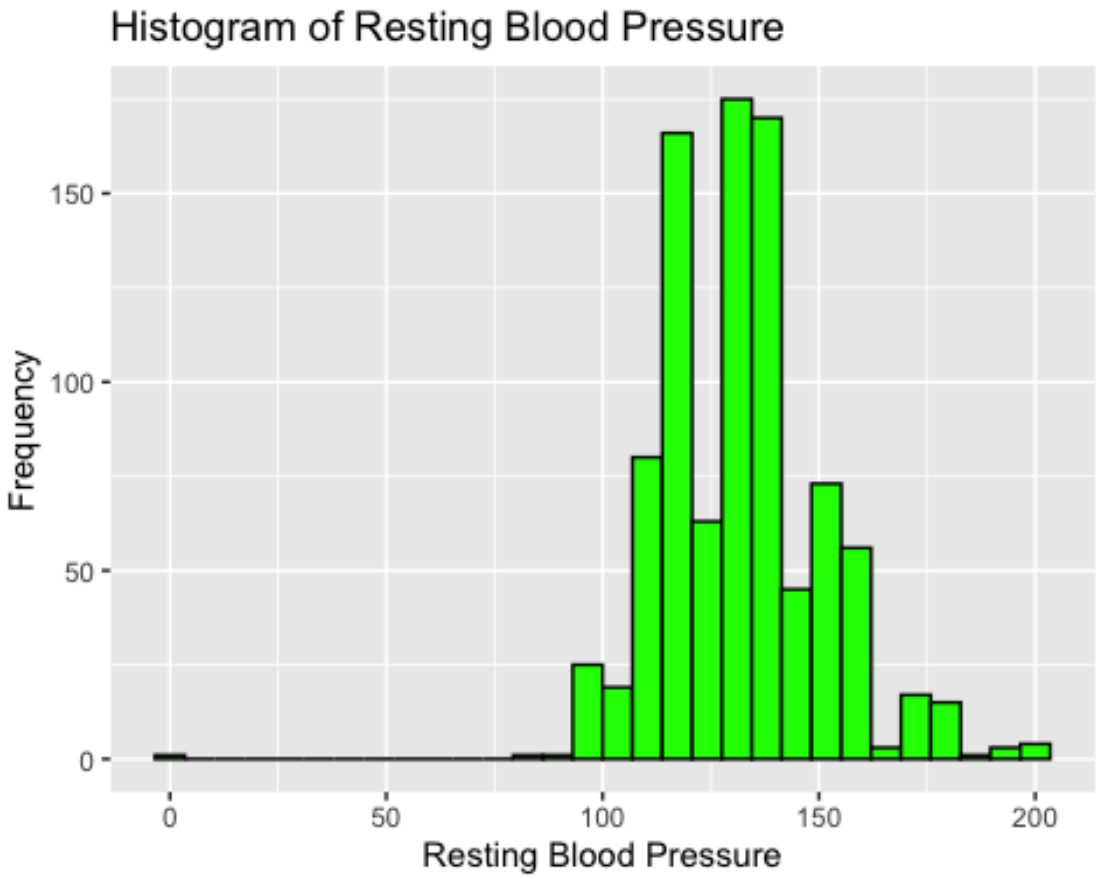
```r
# Histogram for Cholesterol
ggplot(df, aes(x = Cholesterol)) +
  geom_histogram(bins = 30, fill = "red", color = "black") +
  ggtitle("Histogram of Cholesterol") +
  xlab("Cholesterol") +
  ylab("Frequency")
```



-After looking at the histogram, it seems that we will have to handle zeroes in next step, as it seems there is data input error in dataset as 0 cholesterol is not possible.

3. RestingBP

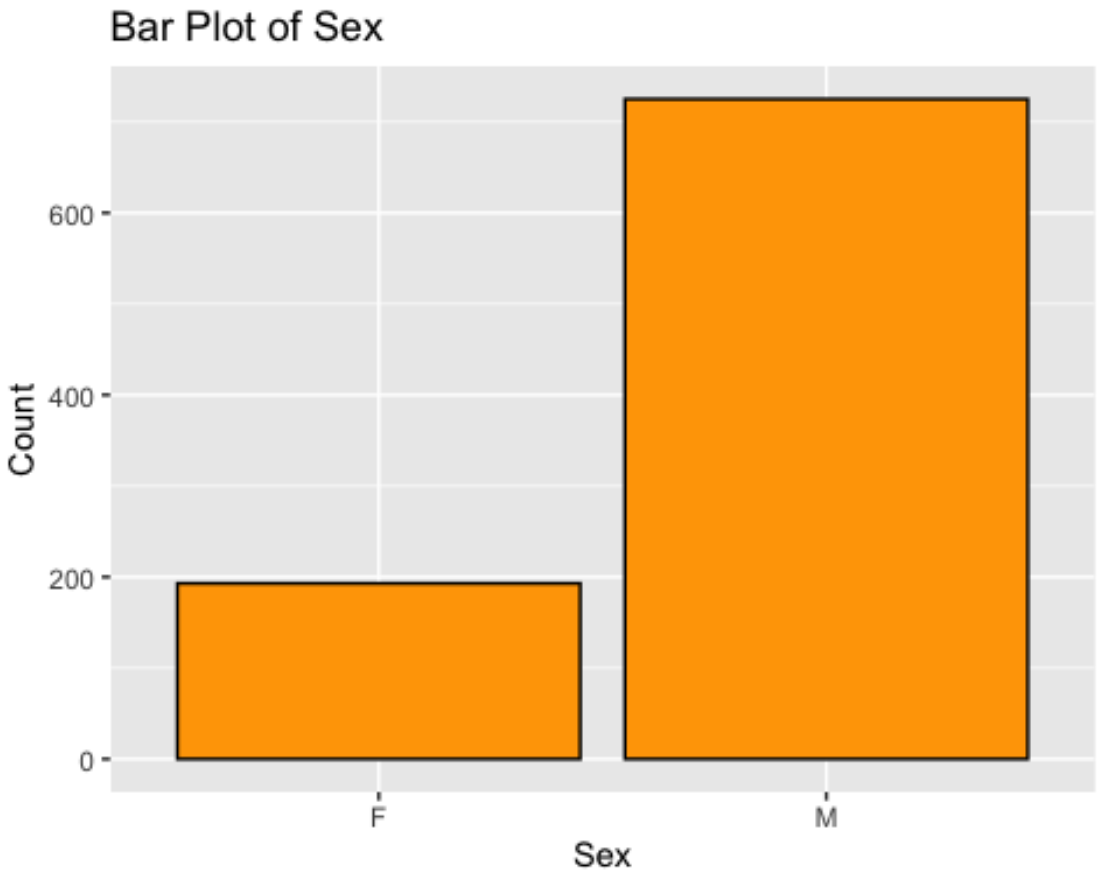
```
# Histogram for Resting Blood Pressure
ggplot(df, aes(x = RestingBP)) +
  geom_histogram(bins = 30, fill = "green", color = "black") +
  ggtitle("Histogram of Resting Blood Pressure") +
  xlab("Resting Blood Pressure") +
  ylab("Frequency")
```



-same can be said here at seem there is probably one 0 entry in RestingBP, we will handle it again in the next step

4. Sex

```
# Bar plot for Sex
ggplot(df, aes(x = Sex)) +
  geom_bar(fill = "orange", color = "black") +
  ggtitle("Bar Plot of Sex") +
  xlab("Sex") +
  ylab("Count")
```



Handling missing values and zeroes

```
# Replace zeroes with NA for RestingBP and Cholesterol
df$RestingBP[df$RestingBP == 0] <- NA
df$Cholesterol[df$Cholesterol == 0] <- NA

# Check for missing values and calculate replacements
summary(df)  # To see the distribution and identify missing values
```

```
##       Age            Sex             Cholesterol      RestingBP
##  Min.   :28.00   Length:918         Min.   : 85.0   Min.   : 80.0
##  1st Qu.:47.00   Class :character   1st Qu.:207.2   1st Qu.:120.0
##  Median :54.00   Mode  :character   Median :237.0   Median :130.0
##  Mean   :53.51                      Mean   :244.6   Mean   :132.5
##  3rd Qu.:60.00                      3rd Qu.:275.0   3rd Qu.:140.0
##  Max.   :77.00                      Max.   :603.0   Max.   :200.0
##                                     NA's   :172     NA's   :1
##   HeartDisease
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.5534
```

```
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

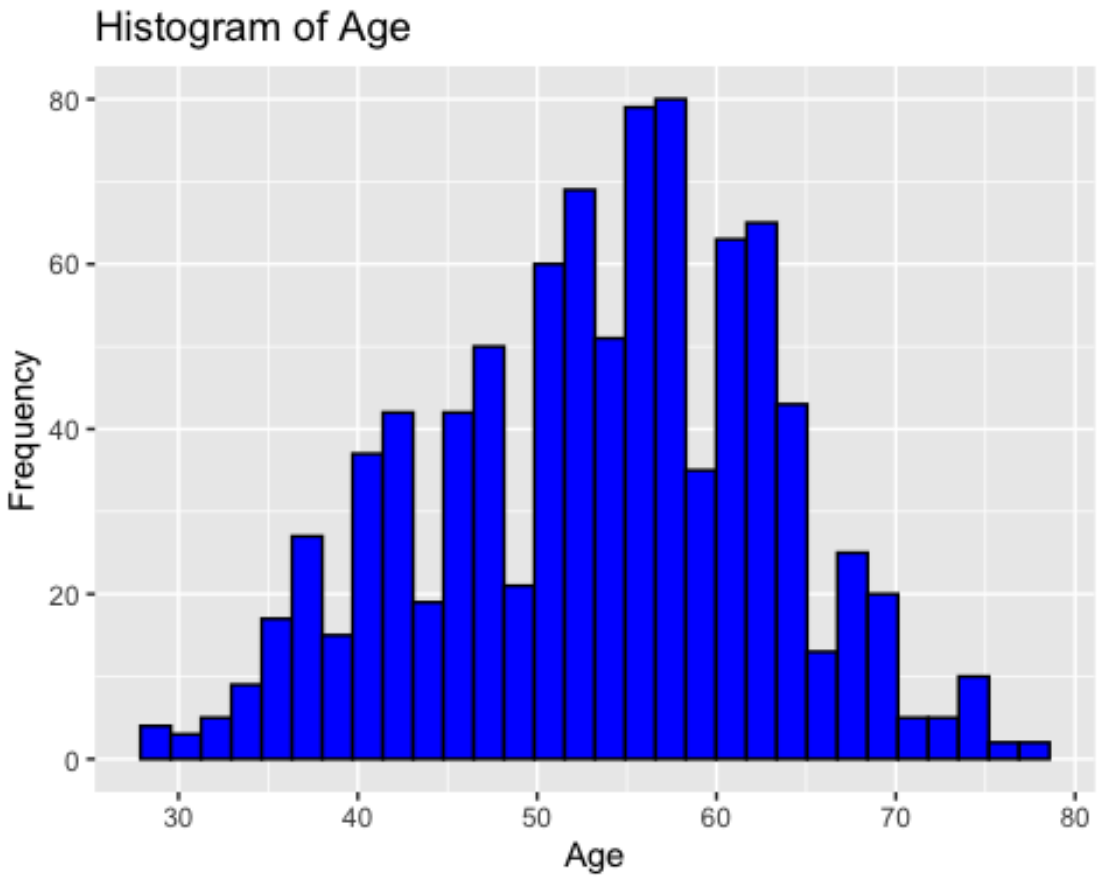
```r
# Impute missing values for Age using median
df$Age[is.na(df$Age)] <- median(df$Age, na.rm = TRUE)

# If Sex is a factor, replace missing values with the mode
# This assumes Sex is stored as a factor; adjust as necessary for your data
mode_sex <- names(which.max(table(df$Sex)))
df$Sex[is.na(df$Sex)] <- mode_sex

# Verify changes
summary(df)
```

```
##       Age            Sex             Cholesterol      RestingBP
##  Min.   :28.00   Length:918         Min.   : 85.0   Min.   : 80.0
##  1st Qu.:47.00   Class :character   1st Qu.:207.2   1st Qu.:120.0
##  Median :54.00   Mode  :character   Median :237.0   Median :130.0
##  Mean   :53.51                      Mean   :244.6   Mean   :132.5
##  3rd Qu.:60.00                      3rd Qu.:275.0   3rd Qu.:140.0
##  Max.   :77.00                      Max.   :603.0   Max.   :200.0
##                                     NA's   :172     NA's   :1
##   HeartDisease
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.5534
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

Visusalisation after handling missing values

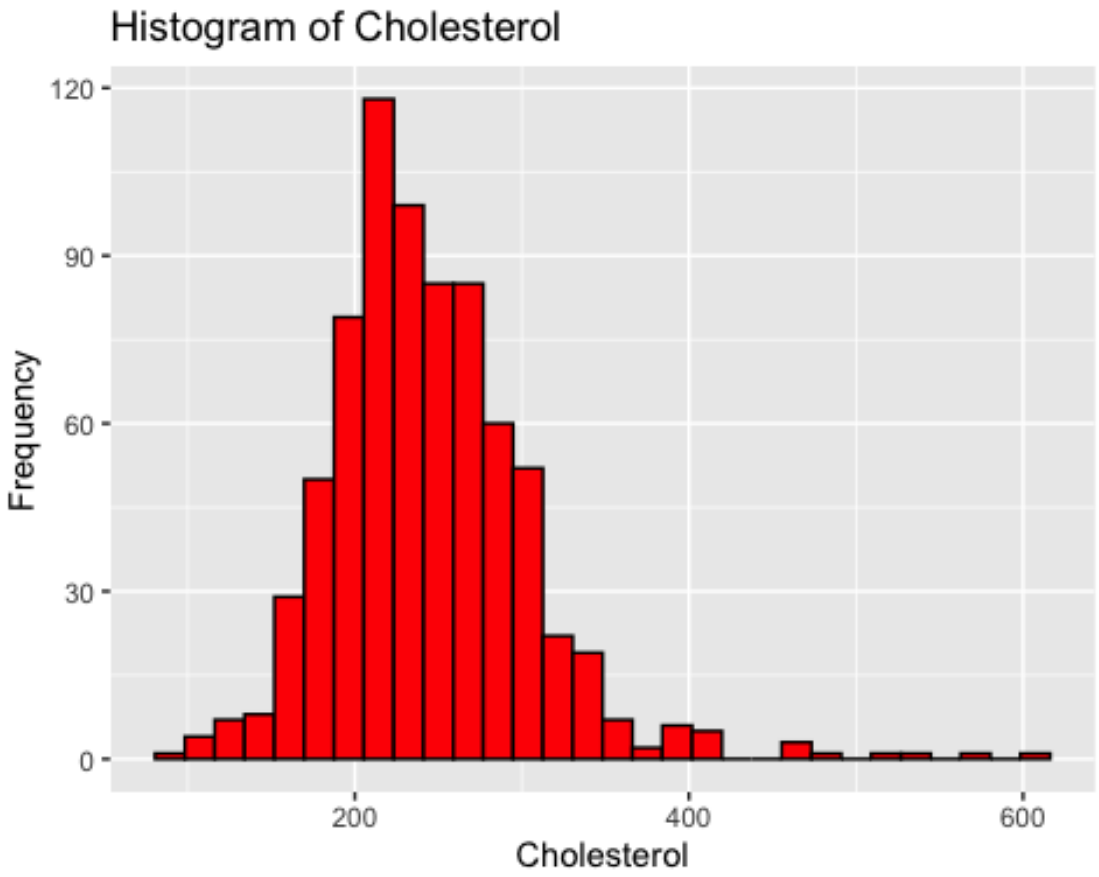
1. Age

```r
# Histogram for Age
ggplot(df, aes(x = Age)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  ggtitle("Histogram of Age") +
  xlab("Age") +
  ylab("Frequency")
```



2. Cholesterol

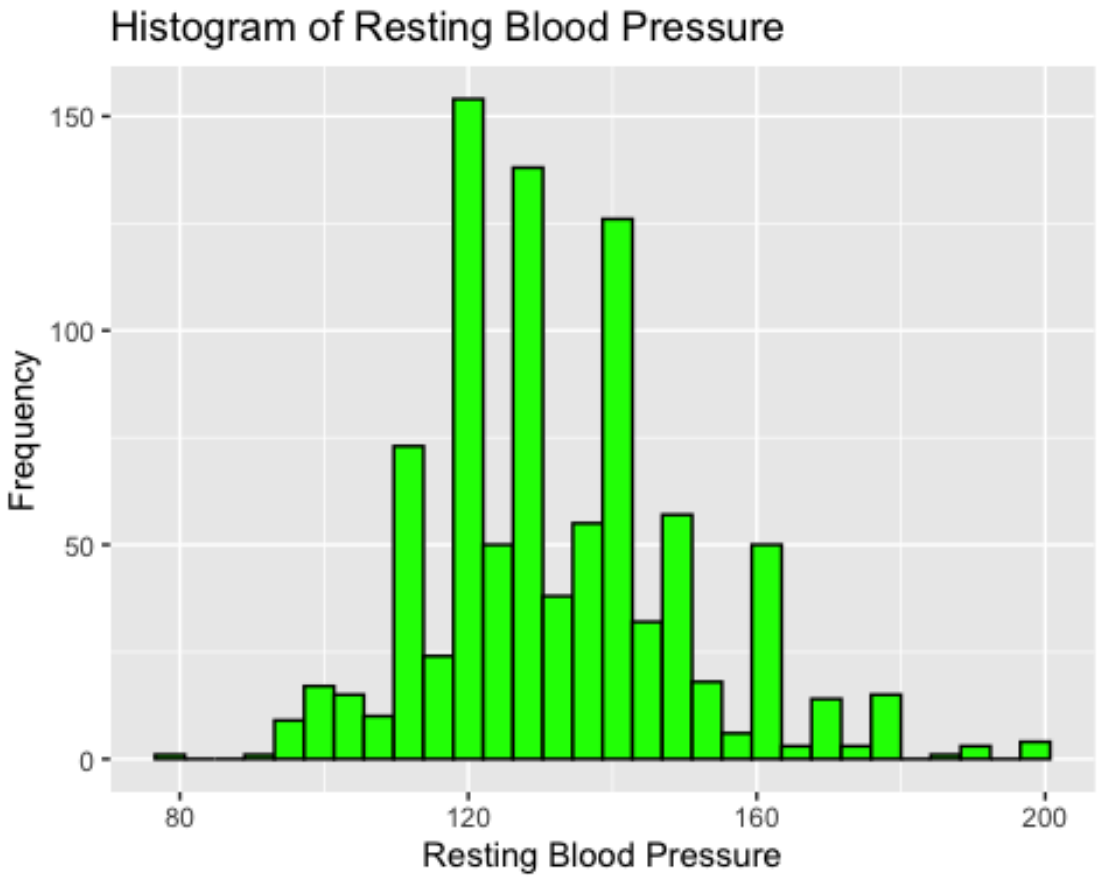
```
ggplot(df, aes(x = Cholesterol)) +
  geom_histogram(bins = 30, fill = "red", color = "black") +
  ggtitle("Histogram of Cholesterol") +
  xlab("Cholesterol") +
  ylab("Frequency")

## Warning: Removed 172 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



3.RestingBP

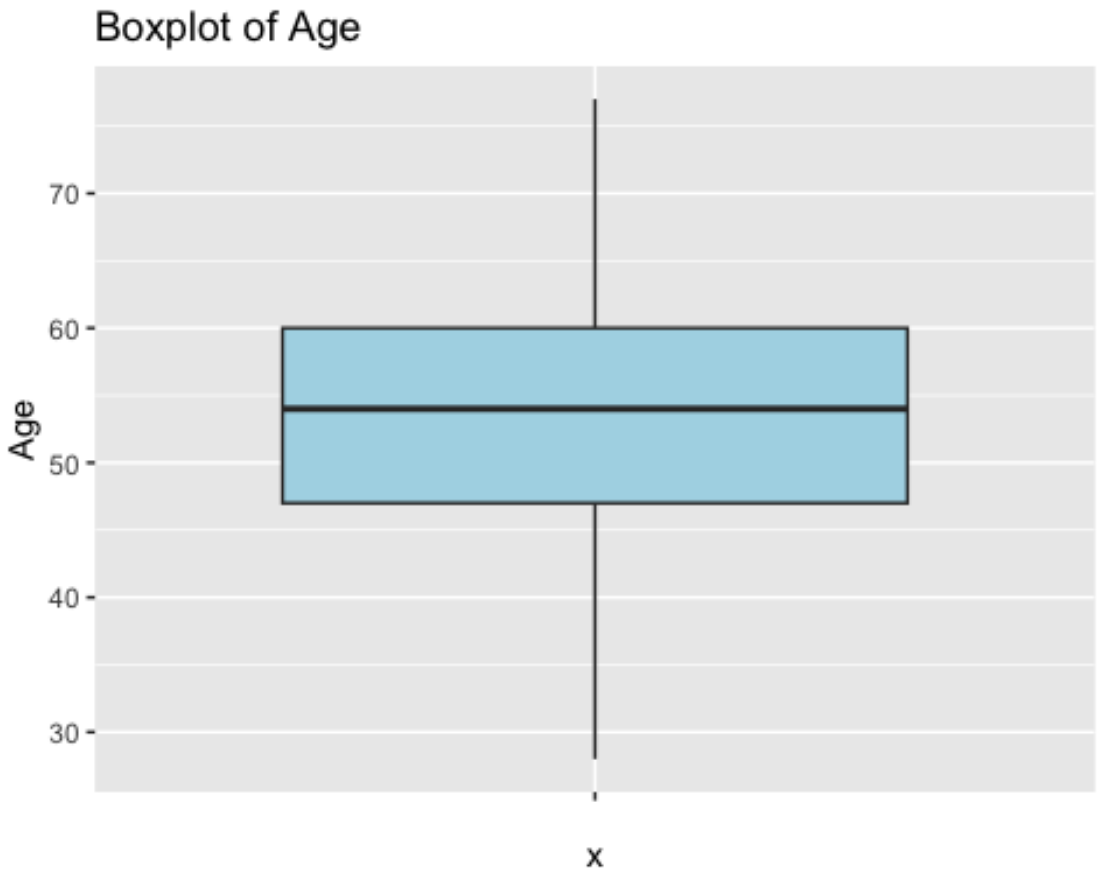
```
# Histogram for Resting Blood Pressure
ggplot(df, aes(x = RestingBP)) +
  geom_histogram(bins = 30, fill = "green", color = "black") +
  ggtitle("Histogram of Resting Blood Pressure") +
  xlab("Resting Blood Pressure") +
  ylab("Frequency")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```



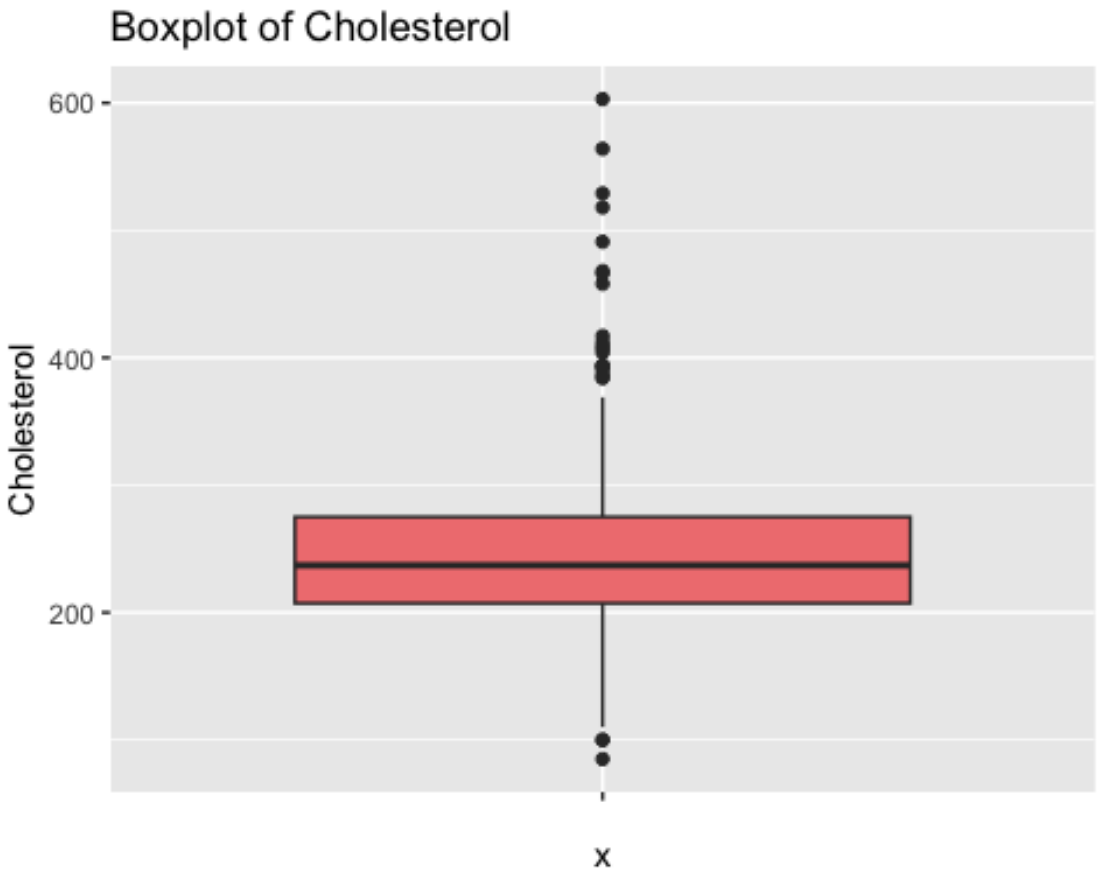
Outlier Detection

```
# Boxplot for Age
ggplot(df, aes(x = "", y = Age)) +
  geom_boxplot(fill = "lightblue") +
  ggtitle("Boxplot of Age")
```



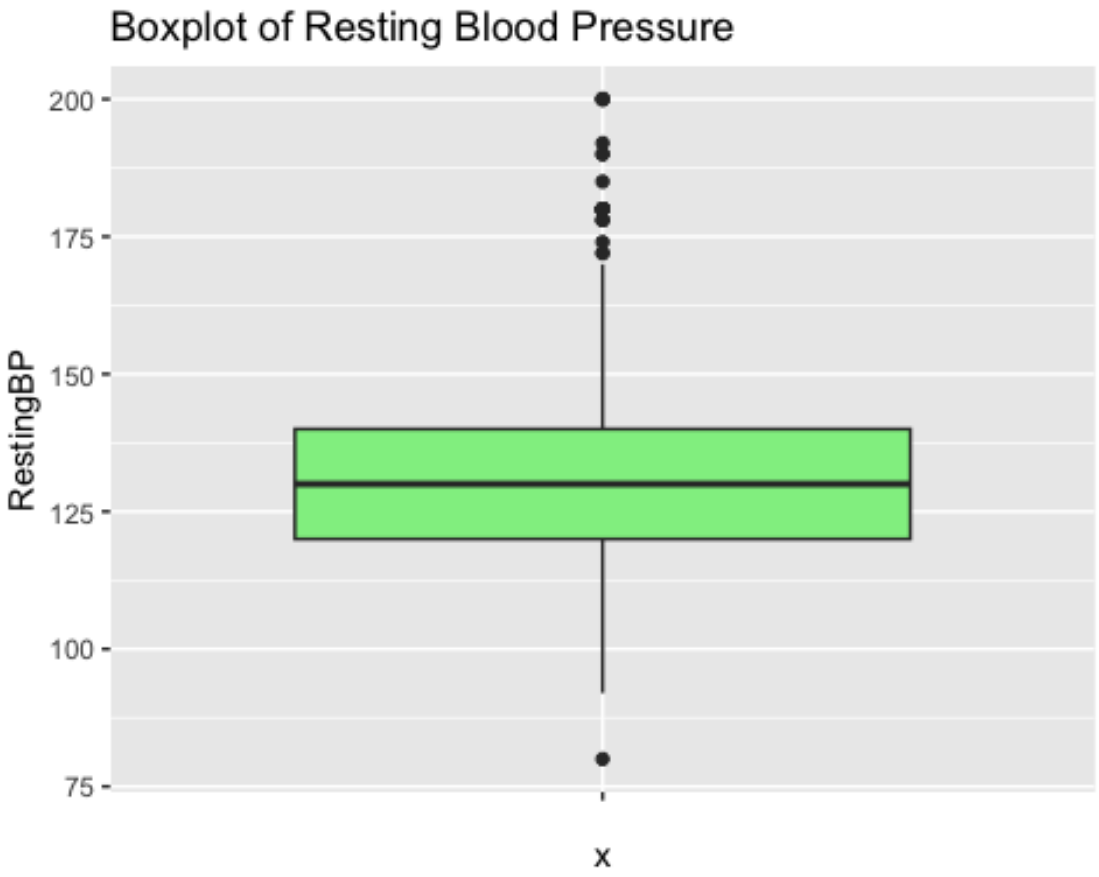
-The boxplot for age shows a distribution without noticeable outliers, indicating a fairly uniform spread of ages within the range, particularly between the mid-40s and mid-60s. The median age is close to the middle of the box, suggesting a symmetrical distribution around the central value. There's no clear indication of skewness or unusual values in the age data.

```
# Boxplot for Cholesterol
ggplot(df, aes(x = "", y = Cholesterol)) +
  geom_boxplot(fill = "lightcoral") +
  ggtitle("Boxplot of Cholesterol")

## Warning: Removed 172 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



-The boxplot of cholesterol levels indicates a median value around the center of the data range, with a relatively symmetric distribution of the middle 50%. There is a noticeable number of high outliers, suggesting some individuals have significantly higher cholesterol levels than the rest(which can be possible in a clinical context as some people might have high cholesterols than normal.

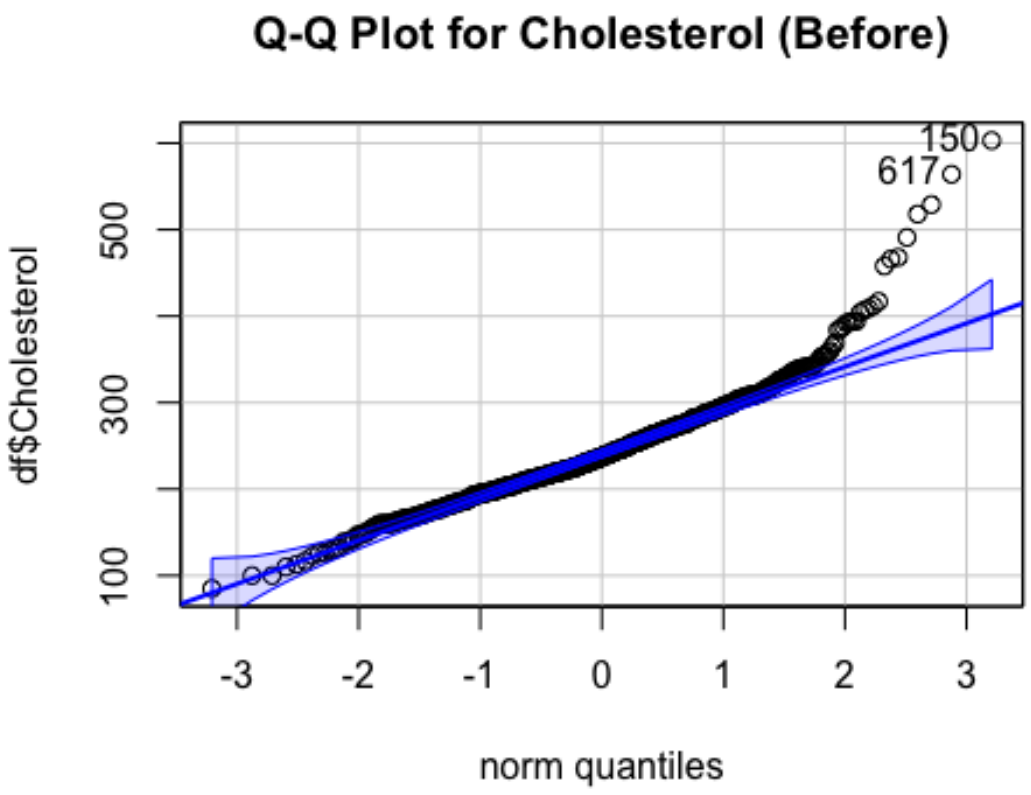
```
# Boxplot for Resting Blood Pressure
ggplot(df, aes(x = "", y = RestingBP)) +
  geom_boxplot(fill = "lightgreen") +
  ggtitle("Boxplot of Resting Blood Pressure")

## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_boxplot()`).
```



-The boxplot shows the distribution of resting blood pressure, with the majority of values clustered between the mid-120s and 140s, indicating a median near the center of this range. A few outliers are present above the upper whisker, suggesting occasional very high blood pressure readings. The distribution appears relatively symmetrical with no evident skewness.

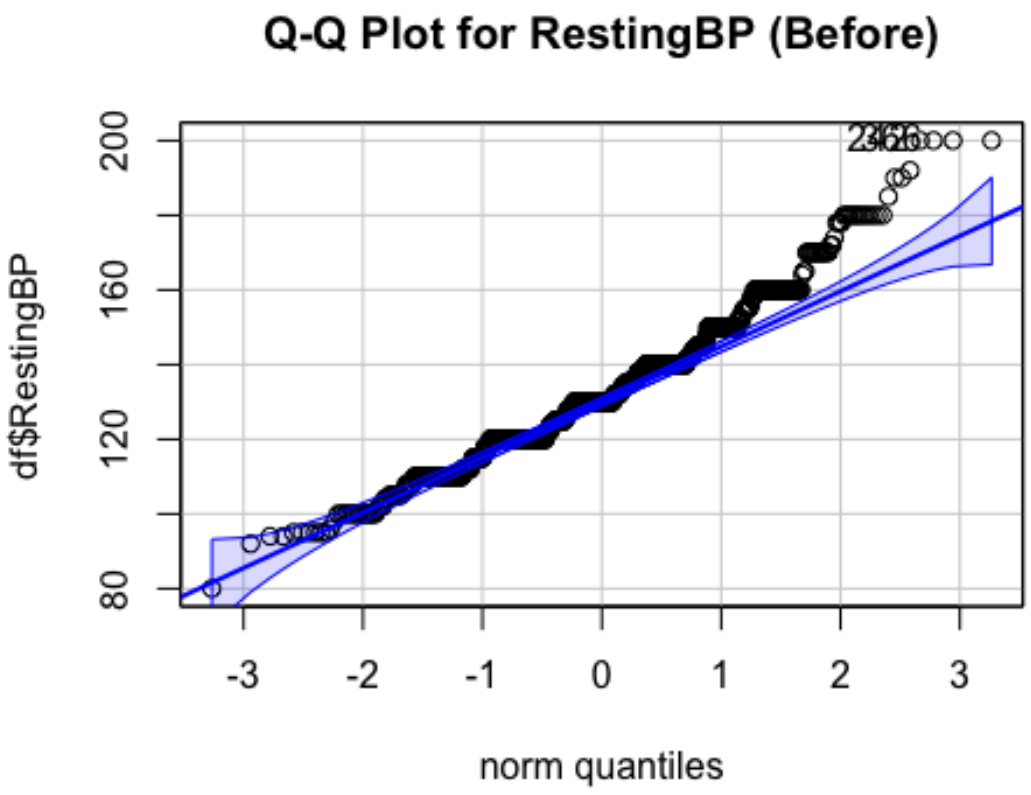
Q-Q Plot

```
# Install the 'car' package if it's not already installed
if (!require(car)) {
    install.packages("car", dependencies = TRUE)
}

## Loading required package: car

## Loading required package: carData

# Load the 'car' package
library(car)

# Q-Q plot for Cholesterol
qqPlot(df$Cholesterol, main = "Q-Q Plot for Cholesterol (Before)")
```




```
## [1] 150 617
```

```r
# Q-Q plot for RestingBP
qqPlot(df$RestingBP, main = "Q-Q Plot for RestingBP (Before)")
```




```
## [1] 242 366
```

Outlier removal

```
# Function to remove outliers based on the IQR
remove_outliers <- function(x) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = TRUE)
  H <- 1.5 * IQR(x, na.rm = TRUE)
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  return(y)
}

# Applying the function to Cholesterol and RestingBP
df$Cholesterol <- remove_outliers(df$Cholesterol)
df$RestingBP <- remove_outliers(df$RestingBP)

# Check the summary to see changes
summary(df$Cholesterol)
```

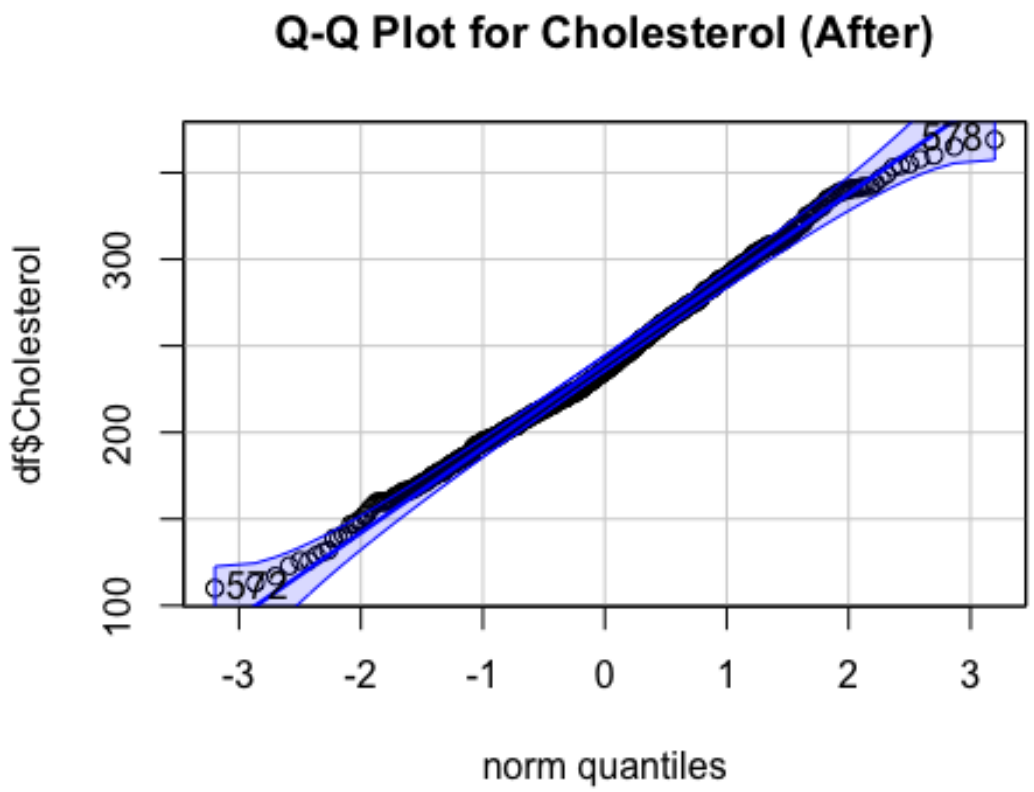
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   110.0   207.0   236.0   239.7   273.0   369.0     195
```

```r
summary(df$RestingBP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    92.0   120.0   130.0   131.1   140.0   170.0      28
```

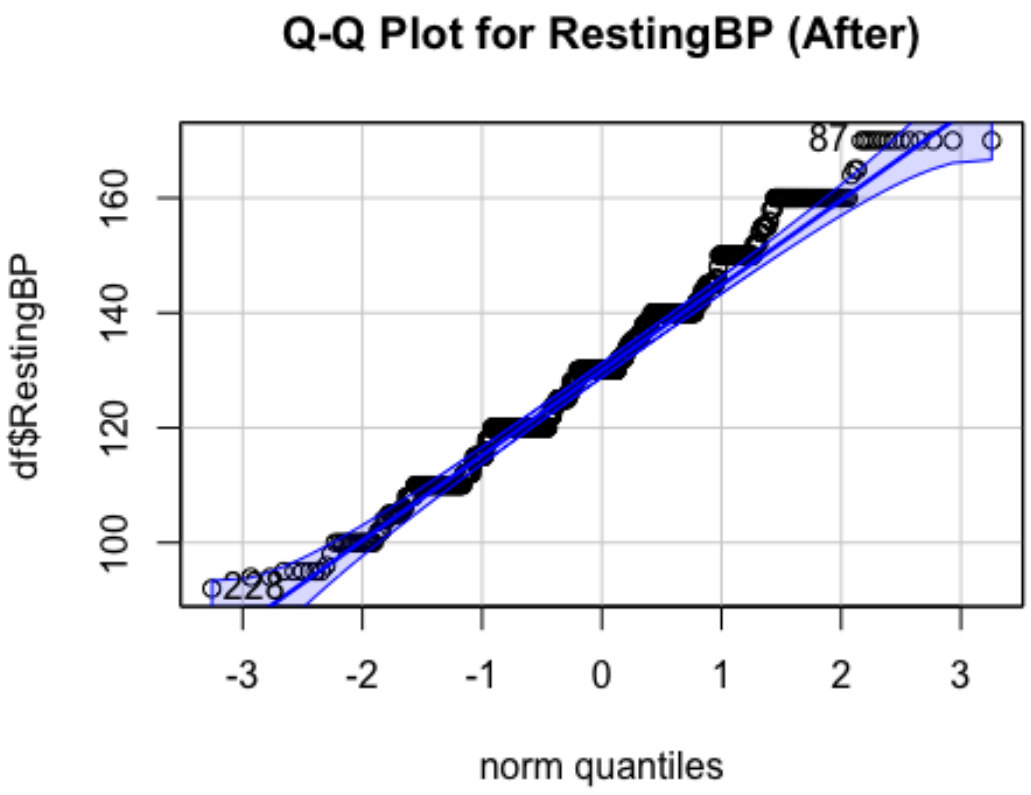
Q-Q plot after

```r
# Q-Q plot for Cholesterol
qqPlot(df$Cholesterol, main = "Q-Q Plot for Cholesterol (After)")
```




```
## [1] 572 578
```

```r
# Q-Q plot for RestingBP
qqPlot(df$RestingBP, main = "Q-Q Plot for RestingBP (After)")
```




```
## [1] 228  87
```

Interpretatio of both Q-Q plots in terms of normality of Cholestrol,RestingBP and Age data

-Cholesterol Before Outlier Removal:

The initial Q-Q plot for cholesterol showed a pronounced departure from normality at higher quantiles. In a clinical setting, this suggests a subset of patients with significantly high cholesterol levels, which might be expected due to conditions like hyperlipidemia. These values, while statistically outliers, are clinically meaningful and should be assessed carefully before any removal.

-Cholesterol After Outlier Removal:

The post-removal Q-Q plot still shows some deviation from normality, but the extremes are less pronounced. This implies that the most extreme cholesterol values have been reduced, likely retaining those within a plausible range for clinical scenarios. It's important that these remaining higher values are not automatically treated as outliers but rather as possible indications of patients with serious hypercholesterolemia.

For Resting Blood Pressure and Age, the same principles apply

Spearman Corelational analysis

-Why spearman over pearson

Non-Normality: Spearman's rank correlation does not require the data to be normally distributed, while Pearson's correlation assumes normality.

Outliers: Spearman's correlation is less affected by outliers since it uses ranks instead of raw data values.

Non-Linear Relationships: Spearman's correlation can detect monotonic relationships, linear or non-linear, while Pearson's correlation is limited to linear relationships.

```
# Load necessary library
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'
```

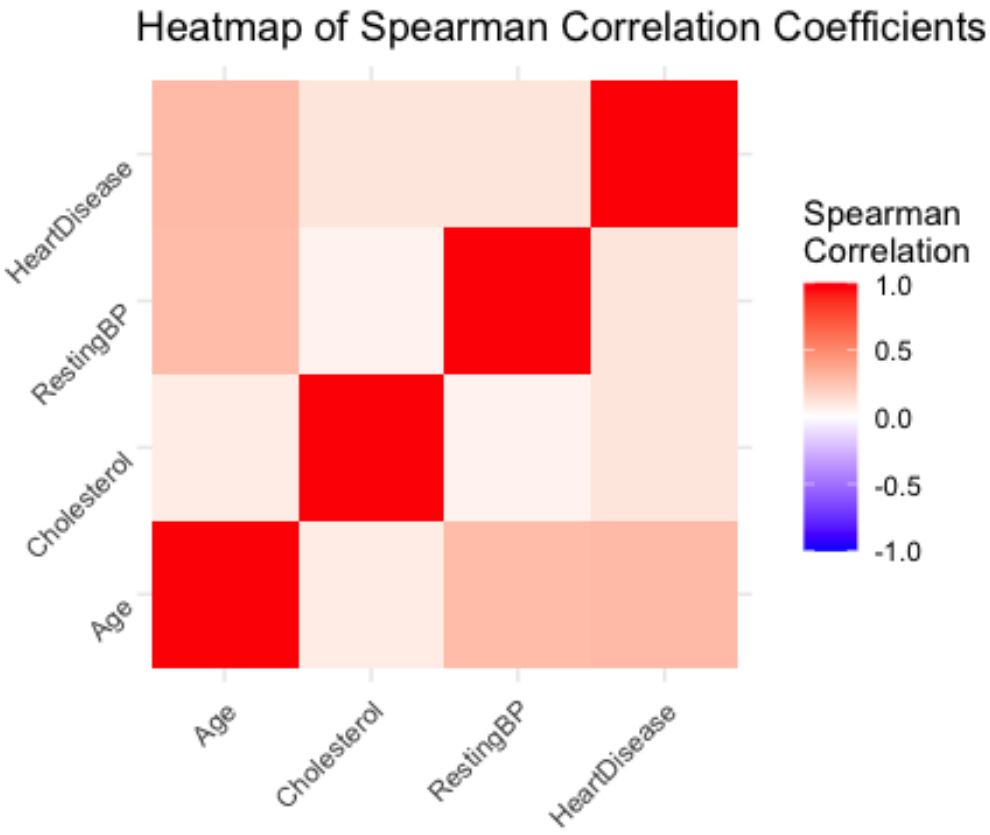
```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
# Calculate Spearman correlation coefficients
correlation_matrix <- rcorr(as.matrix(df[,c("Age", "Cholesterol",
"RestingBP", "HeartDisease")]), type="spearman")

# View the correlation matrix
correlation_matrix
```

```
##                Age Cholesterol RestingBP HeartDisease
## Age           1.00        0.08      0.28         0.29
## Cholesterol   0.08        1.00      0.05         0.11
## RestingBP     0.28        0.05      1.00         0.11
## HeartDisease  0.29        0.11      0.11         1.00
##
## n
##               Age Cholesterol RestingBP HeartDisease
## Age           918         723       890          918
## Cholesterol   723         723       704          723
## RestingBP     890         704       890          890
## HeartDisease  918         723       890          918
##
## P
##               Age    Cholesterol RestingBP HeartDisease
## Age                  0.0254      0.0000    0.0000
## Cholesterol   0.0254             0.1953    0.0036
## RestingBP     0.0000 0.1953                0.0010
## HeartDisease  0.0000 0.0036      0.0010
```

Visualising the corealtaion matrix

```r
# First, ensure the matrix is in a proper format
library(reshape2)
cor_mat <- matrix(c(1, 0.08, 0.28, 0.29,
                    0.08, 1, 0.05, 0.11,
                    0.28, 0.05, 1, 0.11,
                    0.29, 0.11, 0.11, 1),
                  nrow = 4, byrow = TRUE)
colnames(cor_mat) <- rownames(cor_mat) <- c("Age", "Cholesterol",
"RestingBP", "HeartDisease")

# Melt the matrix for plotting
corr_melt <- melt(cor_mat)
names(corr_melt) <- c("Variable1", "Variable2", "Value")

# Plotting using ggplot2
ggplot(data = corr_melt, aes(Variable1, Variable2, fill = Value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint =
0, limit = c(-1, 1), space = "Lab", name="Spearman\nCorrelation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1),
        axis.text.y = element_text(angle = 45, vjust = 1, hjust=1)) +
  labs(x = '', y = '', title = 'Heatmap of Spearman Correlation
Coefficients') +
  coord_fixed()
```



The correlation matrix and the heatmap together indicate the strength and significance of relationships between age, cholesterol, resting blood pressure, and the presence of heart disease:

1. **Age and Heart Disease**: The strongest correlation is between age and heart disease ($\rho = 0.29$), and the heatmap reflects this with a more intense red color. The p-value (< 0.0001) confirms that this relationship is statistically significant. This suggests a clear trend where the risk of heart disease increases with age.

2. **Cholesterol and Heart Disease**: Cholesterol has a positive correlation with heart disease ($\rho = 0.11$), but it's weaker than that of age. The heatmap reflects this with a lighter red shade. The significance ($p = 0.0036$) indicates that higher cholesterol levels are indeed associated with heart disease, though the effect is not as strong as that of age.

3. **RestingBP and Heart Disease**: Similarly, resting blood pressure has a weak but statistically significant positive correlation with heart disease ($\rho = 0.11$), mirrored by the heatmap with a color intensity similar to that of cholesterol. The significance ($p = 0.0010$) suggests a relationship, though it's not a dominant one.

4. **Age and RestingBP**: Age and resting blood pressure also show a moderate positive correlation ($\rho = 0.28$), indicating that as people get older, their blood pressure tends

to be higher. This is consistent with medical knowledge and is also statistically significant ($p < 0.0001$).

The heatmap serves as a visual summary, illustrating these relationships. Darker shades of red indicate stronger and more significant correlations. The consistent red tones across the variables when correlated with heart disease underscore the point that these factors are positively associated with the risk of developing heart disease, albeit to varying degrees.

The visual and statistical evidence combined provide a compelling narrative: as individuals age, their risk for heart disease increases, and this risk is further influenced by physiological measures such as cholesterol and blood pressure. These findings reinforce the importance of monitoring and managing these factors in clinical settings to mitigate heart disease risk.

Chi-Square test for Sex

-Null Hypothesis (H0): There is no association between sex and the presence of heart disease. -Alternative Hypothesis (H1): There is an association between sex and the presence of heart disease.

```
# Assuming 'Sex' is a categorical variable in your dataframe 'df' and
'HeartDisease' is coded as 0 or 1.
# You may need to factor the 'Sex' variable if it's not already a factor.

# Ensure 'Sex' is a factor
df$Sex <- as.factor(df$Sex)

# Create a contingency table of counts for Sex and HeartDisease
sex_disease_table <- table(df$Sex, df$HeartDisease)

# Perform the Chi-Square Test
chi_square_result <- chisq.test(sex_disease_table)

# Print the result
print(chi_square_result)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sex_disease_table
## X-squared = 84.145, df = 1, p-value < 2.2e-16
```

Interpretation:

With such a significant p-value, we can confidently reject the null hypothesis, concluding that there is a statistically significant association between sex and the presence of heart disease in this dataset. This means that the distribution of heart disease is not the same for different sexes; one sex may have a higher or lower rate of heart disease than the other.

Mann-Whitney U Test:

Hypotheses

Age

Null Hypothesis (H0): There is no difference in the median age of individuals with and without heart disease. Alternative Hypothesis (H1): There is a difference in the median age of individuals with and without heart disease.

Cholesterol

Null Hypothesis (H0): There is no difference in the median Cholesterol levels of individuals with and without heart disease. Alternative Hypothesis (H1): There is a difference in the median Cholesterol levels of individuals with and without heart disease.

RestingBP

Null Hypothesis (H0): There is no difference in the median RestingBP of individuals with and without heart disease. Alternative Hypothesis (H1): There is a difference in the median RestingBP of individuals with and without heart disease.

```
# Mann-Whitney U Test for Age
age_result <- wilcox.test(Age ~ HeartDisease, data = df, exact = FALSE)
cat("Mann-Whitney U Test for Age:\n")

## Mann-Whitney U Test for Age:

print(age_result)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Age by HeartDisease
## W = 69138, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0

# Mann-Whitney U Test for Cholesterol
cholesterol_result <- wilcox.test(Cholesterol ~ HeartDisease, data = df,
exact = FALSE)
cat("Mann-Whitney U Test for Cholesterol:\n")

## Mann-Whitney U Test for Cholesterol:

print(cholesterol_result)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Cholesterol by HeartDisease
```

```
## W = 57019, p-value = 0.003654
## alternative hypothesis: true location shift is not equal to 0

# Mann-Whitney U Test for RestingBP
bp_result <- wilcox.test(RestingBP ~ HeartDisease, data = df, exact = FALSE)
cat("Mann-Whitney U Test for Resting Blood Pressure:\n")

## Mann-Whitney U Test for Resting Blood Pressure:

print(bp_result)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  RestingBP by HeartDisease
## W = 85543, p-value = 0.001043
## alternative hypothesis: true location shift is not equal to 0
```

Interpretation :

Mann-Whitney U Test for Age:

Statistic (W = 69138): This value represents the sum of the ranks for the observations from one of the groups (either with or without heart disease), suggesting significant differences in age distribution between the groups. P-value (< 2.2e-16): This extremely low p-value indicates a statistically significant difference in the median age of individuals with and without heart disease. Given that the p-value is far less than 0.05, we reject the null hypothesis that there is no difference in the median ages between the two groups. This implies that age is a significant factor in the presence of heart disease, with either younger or older individuals showing higher prevalence, depending on the data distribution.

Mann-Whitney U Test for Cholesterol:

Statistic (W = 57019): This statistic also represents the sum of ranks for one group, showing how cholesterol levels are ranked between groups. P-value (0.003654): The p-value here suggests that the difference in median cholesterol levels between those with and without heart disease is statistically significant. Since the p-value is less than 0.05, we reject the null hypothesis, indicating that cholesterol levels differ significantly between individuals with and without heart disease, suggesting a potential role of cholesterol in heart disease risk.

Mann-Whitney U Test for Resting Blood Pressure:

Statistic (W = 85543): The high rank sum indicates a significant difference in the distribution of resting blood pressure values between the groups. P-value (0.001043): Similar to the other tests, this low p-value leads us to reject the null hypothesis of no difference in median resting blood pressures between individuals with and without heart disease. This indicates that resting blood pressure is a significant factor in heart disease, likely contributing to or associated with the condition.

#Logistic Regression

```r
# Load necessary libraries
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:Hmisc':
##
##     src, summarize

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(stats)

# Assume 'df' is your dataset name and it's already been read into R

# Prepare the data: ensure factors are correctly specified
df$Sex <- as.factor(df$Sex)
df$HeartDisease <- as.factor(df$HeartDisease)  # Make sure HeartDisease is a
factor for logistic regression

# Base model with only Age
model1 <- glm(HeartDisease ~ Age, family = binomial(link = "logit"), data =
df)
summary(model1)

##
## Call:
## glm(formula = HeartDisease ~ Age, family = binomial(link = "logit"),
##     data = df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.21308    0.42001  -7.650 2.01e-14 ***
## Age          0.06434    0.00780   8.248  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1262.1  on 917  degrees of freedom
## Residual deviance: 1186.7  on 916  degrees of freedom
## AIC: 1190.7
##
## Number of Fisher Scoring iterations: 4
```

```
# Add Cholesterol
model2 <- glm(HeartDisease ~ Age + Cholesterol, family = binomial(link =
"logit"), data = df)
summary(model2)
```

```
##
## Call:
## glm(formula = HeartDisease ~ Age + Cholesterol, family = binomial(link =
"logit"),
##     data = df)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.767074   0.624689  -7.631 2.33e-14 ***
## Age          0.072197   0.009018   8.006 1.19e-15 ***
## Cholesterol  0.003476   0.001665   2.088   0.0368 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1000.40  on 722  degrees of freedom
## Residual deviance:  920.49  on 720  degrees of freedom
##   (195 observations deleted due to missingness)
## AIC: 926.49
##
## Number of Fisher Scoring iterations: 4
```

```
# Add Sex
model3 <- glm(HeartDisease ~ Age + Cholesterol + Sex, family = binomial(link
= "logit"), data = df)
summary(model3)
```

```
##
## Call:
## glm(formula = HeartDisease ~ Age + Cholesterol + Sex, family =
binomial(link = "logit"),
##     data = df)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.850918   0.742240  -9.230  < 2e-16 ***
## Age          0.075802   0.009516   7.966 1.64e-15 ***
## Cholesterol  0.005634   0.001774   3.175   0.0015 **
```

```
## SexM         1.745945   0.222795   7.837 4.63e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1000.40  on 722  degrees of freedom
## Residual deviance:  847.36  on 719  degrees of freedom
##   (195 observations deleted due to missingness)
## AIC: 855.36
##
## Number of Fisher Scoring iterations: 4
```

```
# Add RestingBP
model4 <- glm(HeartDisease ~ Age + Cholesterol + Sex + RestingBP, family =
binomial(link = "logit"), data = df)
summary(model4)
```

```
##
## Call:
## glm(formula = HeartDisease ~ Age + Cholesterol + Sex + RestingBP,
##     family = binomial(link = "logit"), data = df)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.104980   0.992707  -8.165 3.23e-16 ***
## Age          0.069915   0.009816   7.122 1.06e-12 ***
## Cholesterol  0.005439   0.001810   3.006  0.00265 **
## SexM         1.808267   0.231753   7.803 6.07e-15 ***
## RestingBP    0.011767   0.005788   2.033  0.04207 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 973.68  on 703  degrees of freedom
## Residual deviance: 817.37  on 699  degrees of freedom
##   (214 observations deleted due to missingness)
## AIC: 827.37
##
## Number of Fisher Scoring iterations: 4
```

Interpretation:

These results provide a detailed insight into how adding each variable incrementally affects the logistic regression model for predicting heart disease. Let's interpret each model's output:

Model 1: HeartDisease ~ Age

- **Intercept and Coefficients**: The model estimates that younger individuals have a lower log-odds of having heart disease, with age significantly increasing the log-odds (coefficient = 0.06434).
- **Statistical Significance**: Both the intercept and the age coefficient are highly significant ($p < 0.0001$), indicating a strong relationship between age and heart disease.
- **Model Fit**: The AIC of 1190.7 suggests the goodness of fit, with a lower AIC indicating a potentially better model.

Model 2: HeartDisease ~ Age + Cholesterol

- **Additional Variable**: Adding cholesterol slightly increases the coefficient for age and introduces a new coefficient for cholesterol, which is also statistically significant ($p = 0.0368$), although weaker than age.
- **Model Fit**: The AIC decreases to 926.49, indicating an improvement in model fit with the addition of cholesterol.

Model 3: HeartDisease ~ Age + Cholesterol + Sex

- **Sex as a Predictor**: The inclusion of sex significantly improves the model. The coefficient for male sex (SexM) is highly significant ($p < 2e\text{-}16$) and positive, indicating a higher likelihood of heart disease in males compared to females.
- **Model Fit**: The AIC further reduces to 855.36, showing a better fit.

Model 4: HeartDisease ~ Age + Cholesterol + Sex + RestingBP

- **Further Improvement**: Adding RestingBP introduces another significant predictor ($p = 0.04207$). The coefficient for RestingBP is positive, suggesting that higher blood pressure is associated with an increased likelihood of heart disease.
- **Model Fit**: The lowest AIC among all models (827.37) suggests that this model fits the data best among the ones considered.

Interpretation and Insights

1. **Progressive Enhancement**: Each additional variable has contributed to a better model fit, as indicated by decreasing AIC values.
2. **Significance of Predictors**:
   - **Age**: Continues to be a strong predictor across all models.
   - **Cholesterol**: Its significance increases slightly when combined with other factors.
   - **Sex**: Shows a strong effect, with males having higher odds of heart disease.
   - **RestingBP**: Also an important predictor, reinforcing common medical knowledge about the risks associated with high blood pressure.
3. **Clinical Implications**: These findings highlight the importance of considering multiple factors in assessing heart disease risk. The significant impact of sex and age suggests targeted interventions might be beneficial.

4. **Model Utility**: The final model could be used for risk stratification and preventive health strategies in clinical settings.

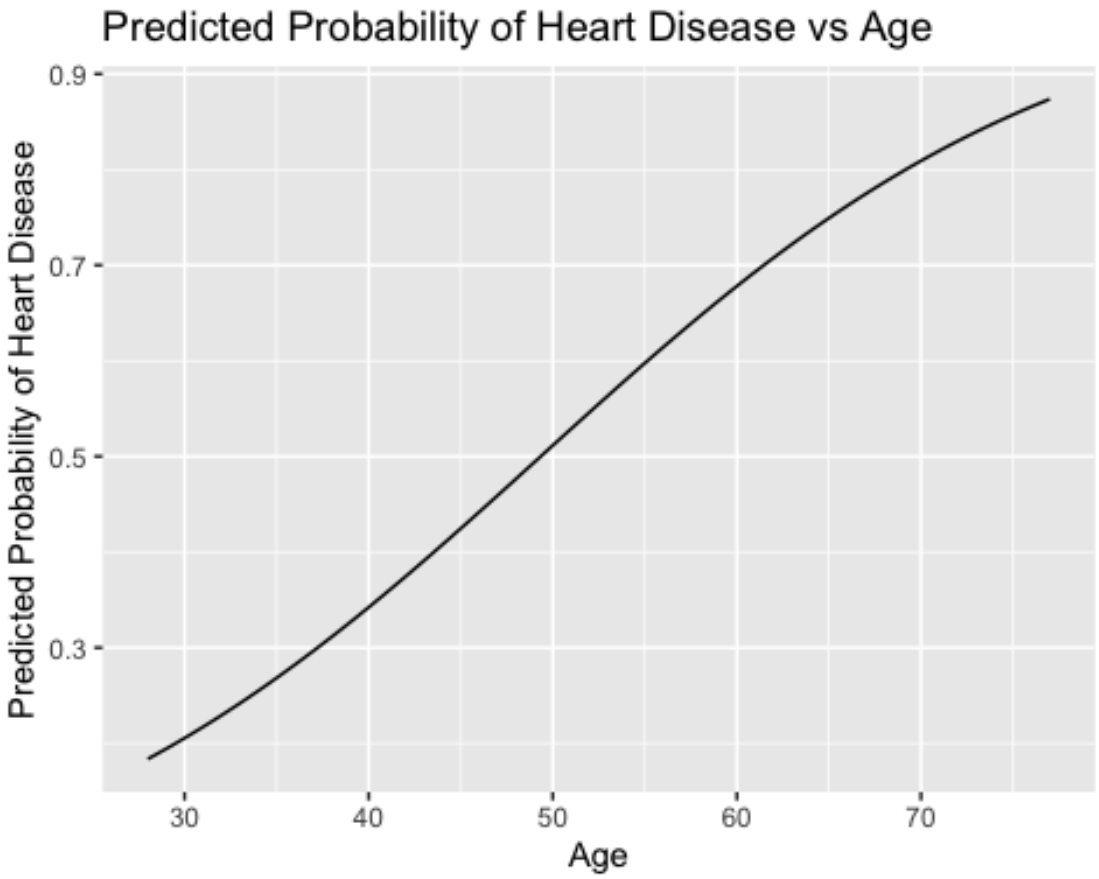
Fisher Scoring Iterations: In each model, the logistic regression uses Fisher Scoring iterations to estimate the model coefficients. The number of iterations required to achieve convergence gives us insight into the complexity and stability of the model's estimation process.

Model 1: HeartDisease ~ Age Fisher Scoring Iterations: The model required 4 iterations to converge. This indicates that the model with just age was straightforward and converged quickly, suggesting that the relationship between age and heart disease is direct and strong without requiring extensive computational adjustments. Model 2: HeartDisease ~ Age + Cholesterol Fisher Scoring Iterations: Again, only 4 iterations were required. The quick convergence, even with the addition of cholesterol, suggests that this model remains stable and efficient in parameter estimation. Model 3: HeartDisease ~ Age + Cholesterol + Sex Fisher Scoring Iterations: This model also converged in 4 iterations. The consistent number of iterations across models indicates that the estimation process is robust, even as more predictors are added. The inclusion of sex, while significantly improving model accuracy (as shown by AIC reduction), did not complicate the convergence process. Model 4: HeartDisease ~ Age + Cholesterol + Sex + RestingBP Fisher Scoring Iterations: The final and most comprehensive model still required only 4 iterations to converge. This is indicative of a well-specified model where the additional variables, although increasing the complexity, are well integrated into the model without causing instability or issues in reaching an optimal solution.

Viusalisation:

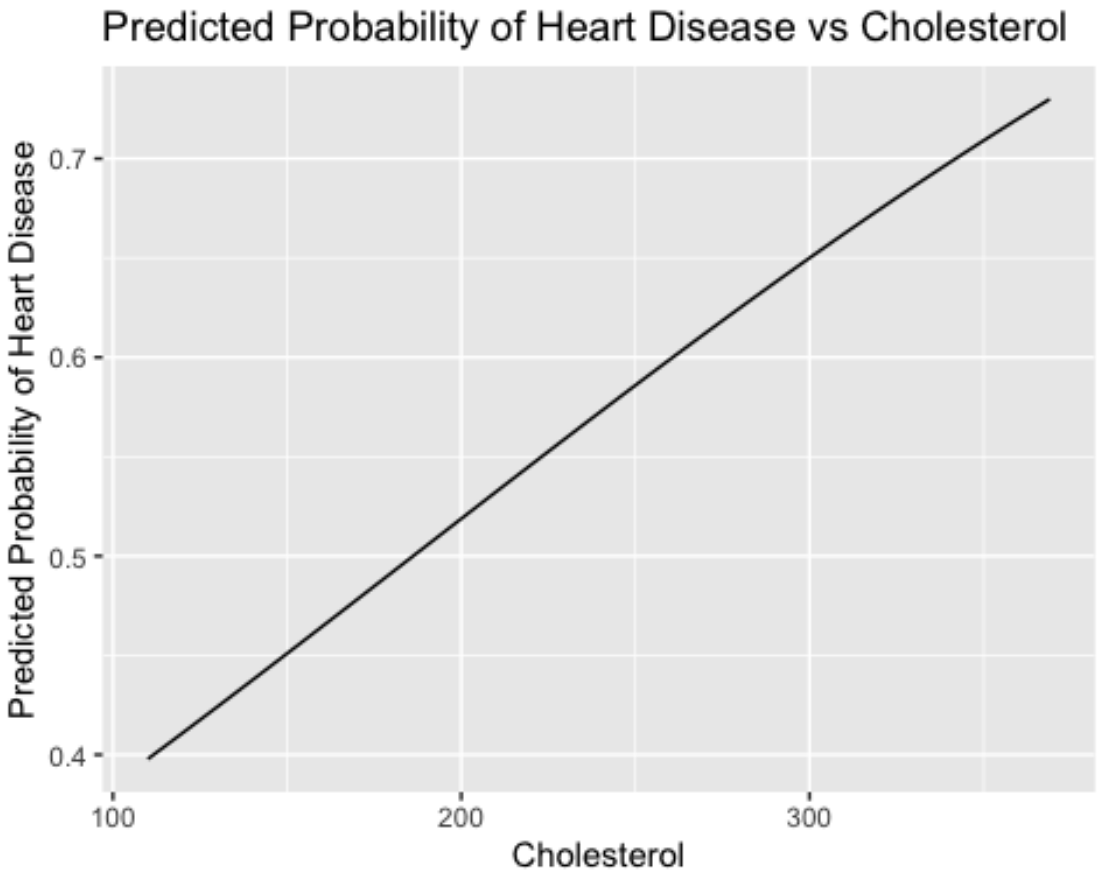
1. Predicted Probabilities Plot for age

```
# Data frame for prediction
pred_data <- with(df, data.frame(Age = seq(min(Age), max(Age), length.out =
100),
                                 Cholesterol = mean(Cholesterol, na.rm =
TRUE),
                                 Sex = factor("M", levels = levels(Sex)),
                                 RestingBP = mean(RestingBP, na.rm = TRUE)))

# Predict probabilities
pred_data$HeartDisease_Prob <- predict(model4, newdata = pred_data, type =
"response")

# Plotting
ggplot(pred_data, aes(x = Age, y = HeartDisease_Prob)) +
  geom_line() +
  labs(title = "Predicted Probability of Heart Disease vs Age",
       x = "Age",
       y = "Predicted Probability of Heart Disease")
```



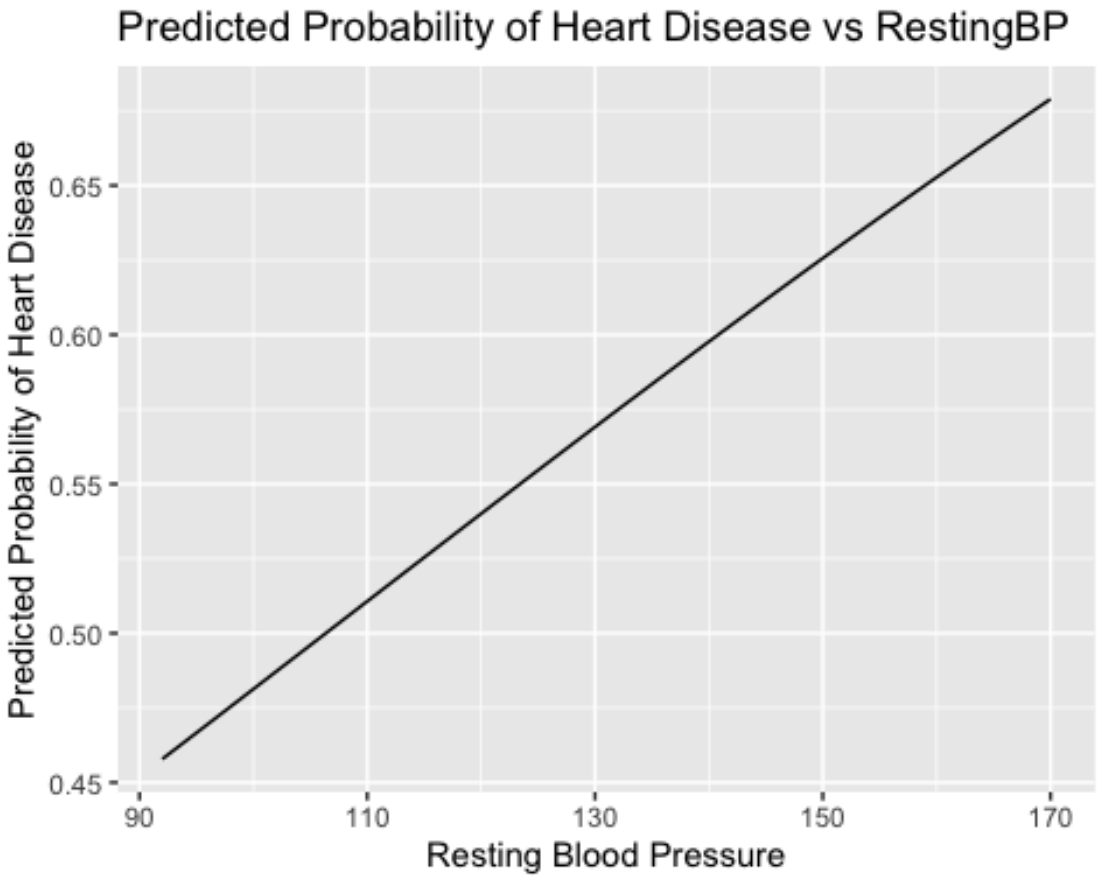
2. Predicted Probabilities Plot for Cholesterol

```
pred_data_chol <- with(df, data.frame(Age = mean(Age, na.rm = TRUE),
                                      Cholesterol = seq(min(Cholesterol,
na.rm = TRUE), max(Cholesterol, na.rm = TRUE), length.out = 100),
                                      Sex =
factor(ifelse(mean(as.numeric(Sex)) > 0.5, "M", "F"), levels = levels(Sex)),
                                      RestingBP = mean(RestingBP, na.rm =
TRUE)))

# Predict probabilities for Cholesterol
pred_data_chol$HeartDisease_Prob <- predict(model4, newdata = pred_data_chol,
type = "response")

# Plotting for Cholesterol
ggplot(pred_data_chol, aes(x = Cholesterol, y = HeartDisease_Prob)) +
  geom_line() +
  labs(title = "Predicted Probability of Heart Disease vs Cholesterol",
       x = "Cholesterol",
       y = "Predicted Probability of Heart Disease")
```



3. Predicted Probabilities Plot for RestingBP

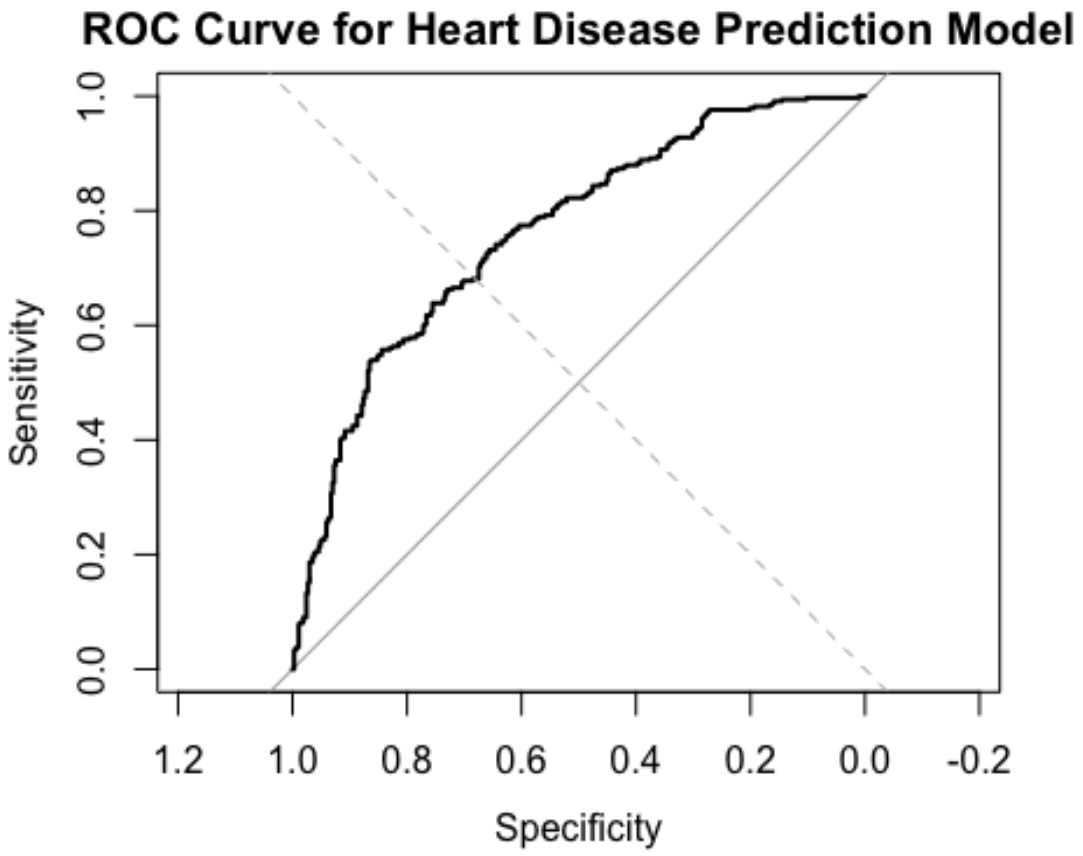
```
# Create a data frame for RestingBP prediction
pred_data_bp <- with(df, data.frame(Age = mean(Age, na.rm = TRUE),
                                    Cholesterol = mean(Cholesterol, na.rm =
TRUE),
                                    Sex = factor(ifelse(mean(as.numeric(Sex))
> 0.5, "M", "F"), levels = levels(Sex)),
                                    RestingBP = seq(min(RestingBP, na.rm =
TRUE), max(RestingBP, na.rm = TRUE), length.out = 100)))

# Predict probabilities for RestingBP
pred_data_bp$HeartDisease_Prob <- predict(model4, newdata = pred_data_bp,
type = "response")

# Plotting for RestingBP
ggplot(pred_data_bp, aes(x = RestingBP, y = HeartDisease_Prob)) +
  geom_line() +
  labs(title = "Predicted Probability of Heart Disease vs RestingBP",
       x = "Resting Blood Pressure",
       y = "Predicted Probability of Heart Disease")
```



4. ROC curve

```
# Load the pROC package
library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

# Make sure to use the same data for prediction as was used for fitting the
model
complete_cases <- complete.cases(df$HeartDisease, df$Age, df$Cholesterol,
df$Sex, df$RestingBP)
df_complete <- df[complete_cases, ]
predicted_probabilities <- predict(model4, newdata = df_complete, type =
"response")

# Create the ROC object using only the complete cases
roc_obj <- roc(response = df_complete$HeartDisease, predictor =
predicted_probabilities)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

# Plot the ROC curve
plot(roc_obj, main="ROC Curve for Heart Disease Prediction Model")
abline(a=0, b=1, col="gray", lty=2)  # Diagonal line for reference
```




```
# Calculate the AUC (Area Under the Curve)
auc(roc_obj)

## Area under the curve: 0.7619
```

Interpretation:

Combining the insights from the ROC curve and its metrics along with the clinical implications of the model's predictive capabilities provides a comprehensive interpretation of how well the model performs in predicting heart disease. Here is a detailed combined interpretation, along with an explanation of sensitivity and specificity in the context of this project:

Combined Interpretation:

ROC Curve and AUC Metrics:

- **High AUC (0.7619)**: This indicates a good level of discrimination by the model. It suggests that there is approximately a 76.19% chance that the model can correctly distinguish between a patient with heart disease and one without, based on the predictors used (Age, Cholesterol, Sex, and RestingBP).
- **Shape and Location of the Curve**: The ROC curve's significant bow towards the upper left corner signifies that the model successfully maximizes the true positive rate (sensitivity) while minimizing the false positive rate (1-specificity). This shape is indicative of a strong predictive model.
- **Optimal Threshold**: The point nearest to the top left corner of the ROC curve represents the optimal balance between sensitivity and specificity. Choosing this threshold allows for the best trade-off, minimizing both false negatives and false positives.

Sensitivity and Specificity:

- **Sensitivity (True Positive Rate)**: In this project, sensitivity refers to the model's ability to correctly identify patients who do have heart disease. A high sensitivity means that the model is effective at catching most of the true cases of heart disease, thereby reducing the risk of false negatives (i.e., failing to identify a patient with heart disease).
- **Specificity (True Negative Rate)**: Specificity, on the other hand, measures the model's ability to correctly identify patients who do not have heart disease. High specificity indicates that the model is good at avoiding false positives (i.e., incorrectly diagnosing heart disease in patients who are actually disease-free).

Clinical Implications:

- **Practical Utility**: The ROC curve and AUC value support the model's utility in clinical settings, suggesting that it can be a reliable tool for screening patients for heart disease risk.
- **Adjustment of Decision Threshold**: Depending on the clinical context, adjusting the decision threshold might be necessary. For example, in a screening scenario where missing a case of heart disease could be detrimental, a lower threshold might be preferred to increase sensitivity, even at the expense of increasing false positives.
- **Validation and Reliability**: While the model shows promise, further validation with different populations or prospective clinical data is recommended to confirm its reliability and generalizability.

Conclusion for ROC:

The model developed in this project shows strong predictive performance, as evidenced by its ROC curve and AUC. It effectively balances sensitivity and specificity, making it a valuable tool for predicting heart disease. The optimal threshold identified from the ROC curve should guide clinical decision-making, ensuring that the model's use is tailored to

specific clinical needs and contexts. Further testing and validation are advised to bolster confidence in its applicability across diverse patient groups and clinical settings.

Conclusion and Findings

Based on the analyses conducted throughout this project using logistic regression and other statistical methods, the following conclusions and findings have been drawn regarding the predictors of heart disease:

Key Findings:

1. **Age as a Predictor**:
   - **Age** is a significant predictor of heart disease, with older individuals showing a higher likelihood of developing heart disease. The positive coefficient in the logistic regression model indicates that as age increases, so does the probability of having heart disease.
2. **Influence of Cholesterol**:
   - **Cholesterol** levels also play a crucial role, although their impact is less pronounced than age. Higher cholesterol levels are associated with an increased risk of heart disease, as indicated by the positive coefficient in the model.
3. **Sex Differences**:
   - **Sex** significantly influences the risk of heart disease, with males having a higher likelihood of heart disease compared to females. This effect was consistently strong across different models, highlighting sex as a critical factor in heart disease risk assessment.
4. **Resting Blood Pressure**:
   - **Resting Blood Pressure** adds additional risk, with higher values increasing the likelihood of heart disease. This variable, while contributing less than age and sex, still shows a statistically significant association with heart disease.
5. **Model Performance**:
   - The logistic regression model that included all four predictors (age, cholesterol, sex, and resting blood pressure) demonstrated the best fit, as indicated by the lowest Akaike Information Criterion (AIC). This suggests that a combination of these factors provides a robust framework for predicting heart disease.

Conclusions:

- The stepwise approach to adding variables to the logistic regression model revealed that each factor contributes uniquely to the prediction of heart disease, with age and sex being particularly strong predictors. This approach helped in understanding the individual and combined effects of each predictor on the probability of heart disease.

- The visualization of model coefficients and predicted probabilities helped in interpreting the logistic regression model, providing clear insights into how each predictor affects the likelihood of heart disease. These visual representations are invaluable for communicating statistical findings to a non-technical audience.
- The findings from this project underscore the importance of considering multiple physiological and demographic factors in assessing the risk of heart disease. This holistic approach to risk assessment can aid healthcare providers in identifying high-risk individuals for targeted prevention strategies.

Clinical Implications:

- **Targeted Screening**: Older adults, particularly males with high cholesterol and elevated blood pressure, should be prioritized in screening programs for heart disease.
- **Preventive Measures**: Lifestyle interventions aimed at reducing cholesterol and blood pressure can be crucial in lowering heart disease risk, especially in high-risk demographics.
- **Personalized Medicine**: The insights from the model can be used to tailor treatments and interventions based on individual risk profiles, enhancing the effectiveness of clinical care.

Future Research:

- **Data Enrichment**: Including additional variables such as diet, physical activity, smoking status, and genetic factors could further improve the model's accuracy.
- **Advanced Modelling Techniques**: Exploring machine learning techniques like random forests or support vector machines could provide new insights and potentially higher predictive accuracy.
- **Longitudinal Analysis**: A follow-up study with a longitudinal design could examine the causality and changes in heart disease risk factors over time, providing a dynamic perspective on risk prediction.

This project has successfully demonstrated the utility of logistic regression in understanding and predicting heart disease risk based on clinical and demographic factors, providing a solid foundation for further research and clinical application.

Certainly, reflecting on the limitations of the analyses conducted in this project is crucial for understanding the scope and implications of the findings. Here are the limitations relevant to the heart disease prediction project:

Limitations:

1. **Data Completeness and Quality**:
    - **Missing Data**: The dataset had missing values, particularly in variables like Cholesterol and RestingBP, which required handling that could influence the

analysis outcomes. Missing data treatment methods (e.g., imputation) might introduce biases or distort the true relationships.
    - **Measurement Errors**: If any variables were inaccurately measured or reported, this could affect the reliability of the findings.
2. **Model Assumptions**:
    - **Linearity Assumption in Logistic Regression**: Logistic regression assumes a linear relationship between the logit of the outcome and each predictor. If this assumption doesn't hold—for instance, if the effect of cholesterol is not linear—the model might not capture the true nature of the relationship.
    - **Independence of Observations**: The logistic model assumes that the observations are independent of each other. If there is clustering in the data (e.g., patients from the same family or community), this could violate the assumption and affect the model's performance.
3. **Sample Bias**:
    - **Representativeness**: If the sample is not representative of the general population (e.g., overrepresentation of a certain age group or sex), the results may not generalize well to other groups or populations.
4. **Confounding Variables**:
    - **Unmeasured Confounders**: There may be other factors that influence heart disease risk, such as genetic predispositions, lifestyle factors like smoking and diet, or socio-economic status, which were not included in the model. The omission of these variables can lead to omitted variable bias, where the effects of included variables are over or underestimated.
5. **Statistical Power and Effect Size**:
    - **Small Effect Sizes**: While some predictors were statistically significant, their actual impact on heart disease risk (effect size) might be small and of limited clinical relevance.
    - **Multiple Comparisons**: The project involved multiple tests and model comparisons without adjusting for multiple comparisons, which increases the risk of Type I errors (false positives).

Addressing Limitations:

- **Advanced Statistical Techniques**: Using techniques like regularization might help address overfitting and improve model generalization. Furthermore, methods like generalized additive models (GAMs) could be used to relax the linearity assumption.
- **Expanded Data Collection**: Future studies could include more diverse participants and additional relevant variables to enhance the comprehensiveness and applicability of the model.
- **Sensitivity Analyses**: Performing sensitivity analyses could help understand the impact of missing data and measurement errors. It would also be useful to test the robustness of the results to different methodological assumptions.

Project Conclusion:

Despite these limitations, the project provides valuable insights into the factors contributing to heart disease risk and highlights the potential of logistic regression in clinical prediction settings. Future research should aim to address these limitations by incorporating more comprehensive data, exploring more complex models, and validating findings across different populations to enhance the predictive power and reliability of the risk assessment tools developed.

Q.Why did you employ a non paramteric test at first in spearman corealtio and then logistic regression?

In the context of this project, employing both Spearman correlation and logistic regression is indeed viable and provides complementary insights. The rationale behind using both methods can be outlined as follows:

Spearman Correlation Rationale:

- **Distribution of Variables**: Spearman correlation is a non-parametric measure of rank correlation. It is used when the assumptions for Pearson correlation (like normal distribution of variables) are not met. In our case, variables such as Cholesterol and RestingBP were not normally distributed, making Spearman the appropriate choice for initial analysis.
- **Detecting Monotonic Relationships**: Spearman correlation can detect any monotonic relationship, not just linear ones. This means it is good for an initial assessment of whether an increase or decrease in one variable generally associates with an increase or decrease in another variable, regardless of the exact nature of their relationship.

Logistic Regression Rationale:

- **Binary Outcome**: Logistic regression is specifically designed for cases where the response variable is binary (e.g., presence or absence of heart disease). It models the probability of the outcome as a function of the predictor variables.
- **Multiple Variables**: Unlike correlation, which typically examines the relationship between two variables, logistic regression can incorporate multiple predictors simultaneously. This allows for the adjustment of each predictor's effect on the outcome for the presence of other variables.
- **Non-Linearity**: Logistic regression can handle non-linear relationships through the logit transformation. The probability of the outcome is modeled as a logistic function of the predictors, which is inherently non-linear.

Integration of Both Methods:

- **Preliminary Analysis**: Starting with Spearman correlation provides a simple, straightforward assessment of potential relationships without making assumptions about data distribution or linearity. It can guide which variables are worth investigating further.

- **Building Complexity**: After identifying which variables have a monotonic relationship with the outcome using Spearman correlation, logistic regression is used to build a more complex and realistic model of the relationships, taking into account multiple variables and their interactions.
- **Confirming Predictive Power**: While Spearman correlation may suggest a relationship exists, logistic regression helps confirm whether these relationships hold when other factors are controlled for, thus assessing the predictive power of each variable.

Conclusion and Justification:

The project's approach is justified because it starts by identifying potential predictors through a non-parametric method that is robust to certain data irregularities. It then transitions to a more complex, multivariate method that is appropriate for the binary outcome of interest and can take into account the influence of multiple predictors together. This two-step approach ensures that non-normally distributed variables are initially assessed in a way that does not violate statistical assumptions, and that the final predictive model is suitable for binary classification, which is the ultimate goal of the project. This methodological rigor enhances the reliability and validity of the findings.