Heart Disease Analysis Report

Sebastian Szelewicz

2024-06-07

Exploratory Data Analysis

```
# Generating summary statistics for all variables
summary_stats <- heart_data %>% summarise_all(funs(min, max, mean, median, sd, IQR))
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## ℹ Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
print(summary_stats)
```

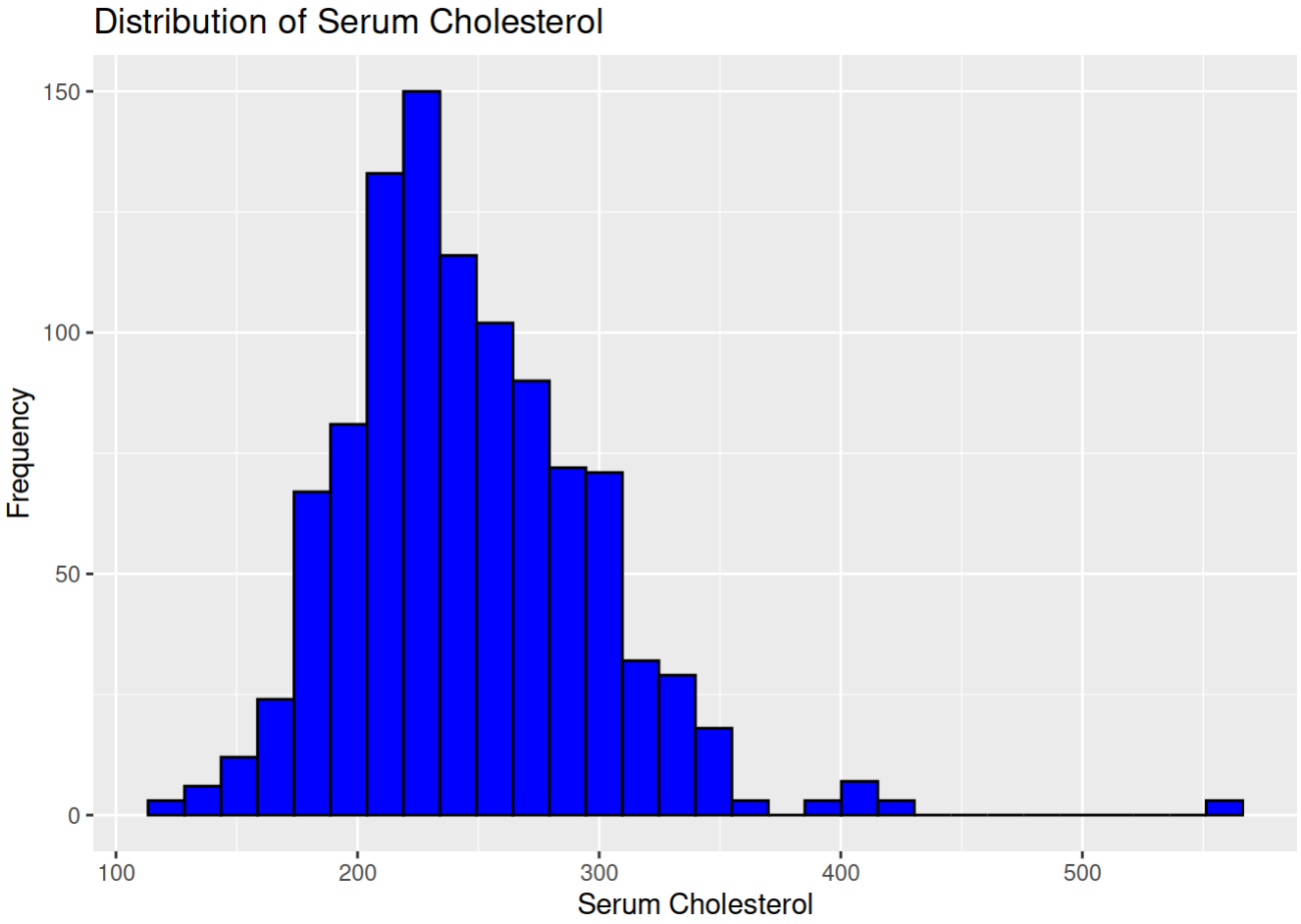
```
## # A tibble: 1 × 84
##   age_min sex_min cp_min trestbps_min chol_min fbs_min restecg_min thalach_min
##     <dbl>   <dbl>  <dbl>        <dbl>    <dbl>   <dbl>       <dbl>       <dbl>
## 1      29       0      0           94      126       0           0          71
## # ℹ 76 more variables: exang_min <dbl>, oldpeak_min <dbl>, slope_min <dbl>,
## #   ca_min <dbl>, thal_min <dbl>, target_min <dbl>, age_max <dbl>,
## #   sex_max <dbl>, cp_max <dbl>, trestbps_max <dbl>, chol_max <dbl>,
## #   fbs_max <dbl>, restecg_max <dbl>, thalach_max <dbl>, exang_max <dbl>,
## #   oldpeak_max <dbl>, slope_max <dbl>, ca_max <dbl>, thal_max <dbl>,
## #   target_max <dbl>, age_mean <dbl>, sex_mean <dbl>, cp_mean <dbl>,
## #   trestbps_mean <dbl>, chol_mean <dbl>, fbs_mean <dbl>, restecg_mean <dbl>, …
```

Visualizing Distributions

Distribution of Serum Cholesterol

This histogram shows the distribution of serum cholesterol levels in the dataset. The x-axis represents the cholesterol values, and the y-axis shows the frequency or count of patients with each cholesterol level. The distribution appears to be roughly unimodal and right-skewed, with the highest frequency around 200-250 mg/dL. There is a long tail towards higher cholesterol values, indicating that some patients have very high cholesterol levels. The distribution tapers off at lower cholesterol levels, with few patients having extremely low cholesterol.

```
ggplot(heart_data, aes(x = chol)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  ggtitle("Distribution of Serum Cholesterol") +
  xlab("Serum Cholesterol") +
  ylab("Frequency")
```

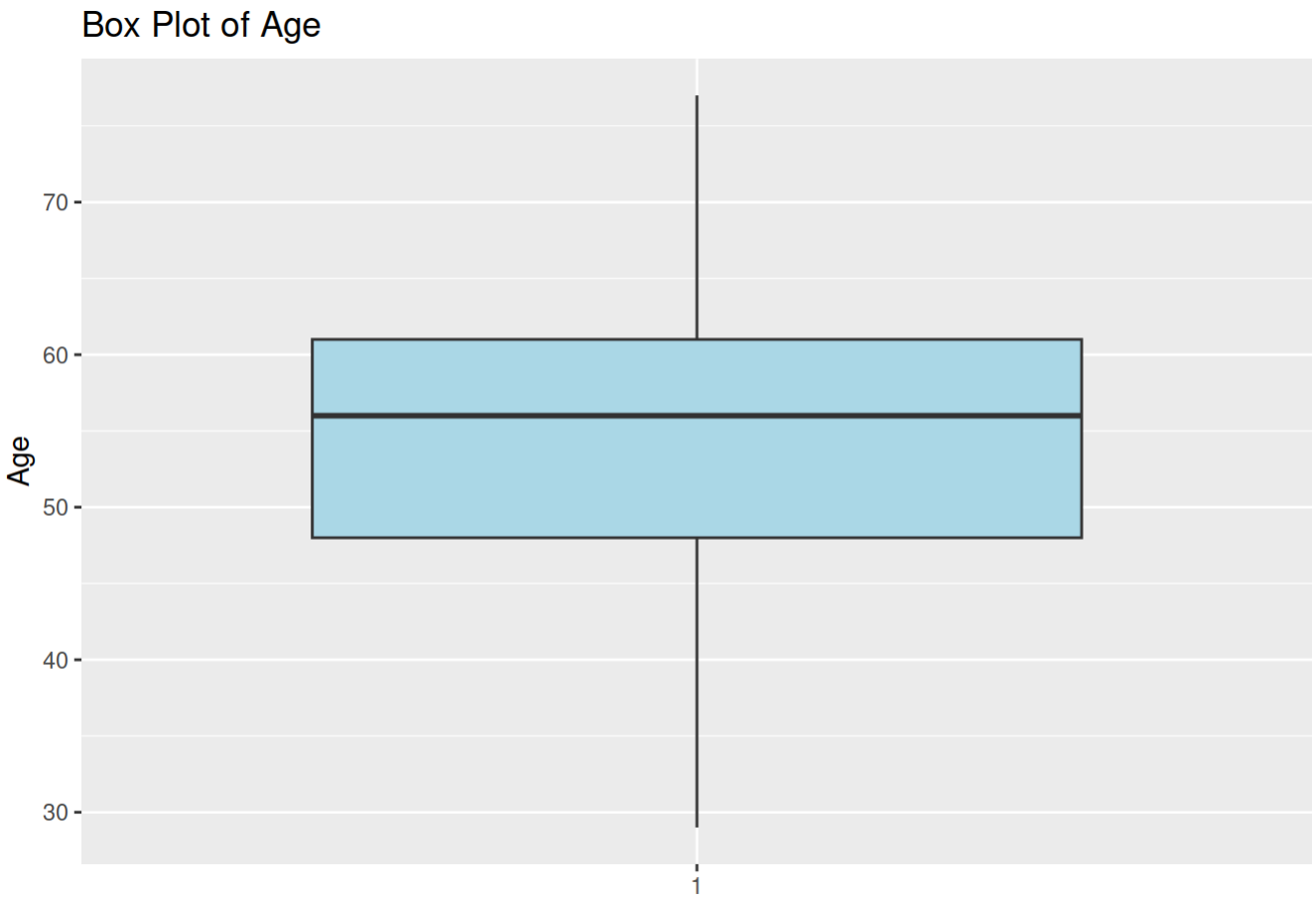


Box plot of age

This box plot visualizes the distribution of age in the dataset. The box represents the interquartile range (IQR), with the horizontal line inside the box indicating the median age. The whiskers extend to the minimum and maximum values, excluding outliers.

The box plot shows that the median age in the dataset is around 55-60 years old. The IQR is relatively narrow, suggesting that most patients are clustered around the median age. There are no extreme outliers in terms of age, as indicated by the absence of individual points beyond the whiskers

```
ggplot(heart_data, aes(y = age, x = factor(1))) +
  geom_boxplot(fill = "lightblue") +
  ggtitle("Box Plot of Age") +
  xlab("") +
  ylab("Age")
```



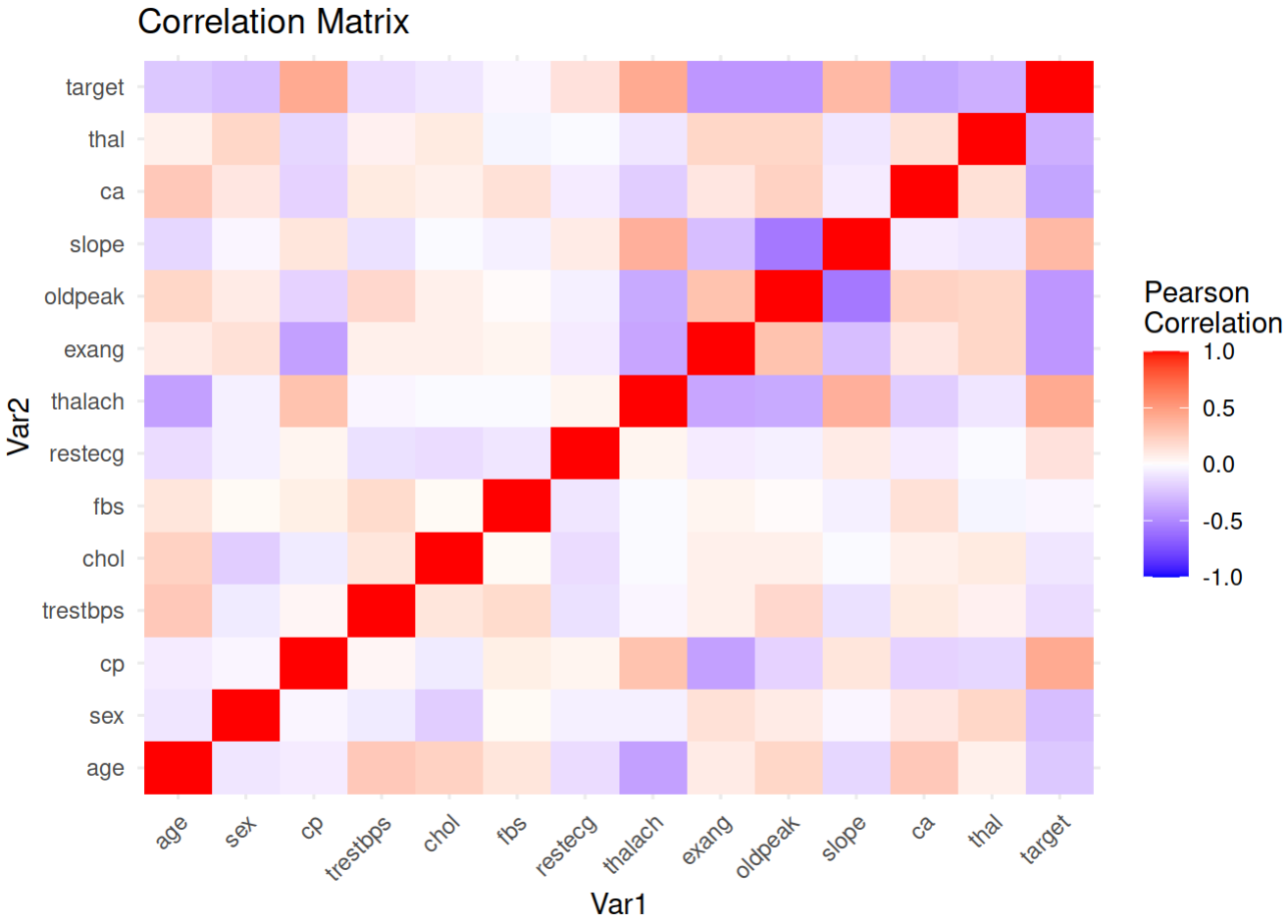
Correlation Analysis

This correlation matrix displays the Pearson correlation coefficients between different variables in the dataset. The color scale ranges from dark blue (strong negative correlation) to dark red (strong positive correlation), with lighter shades indicating weaker correlations. Some notable observations from the correlation matrix: - The target variable (likely indicating heart disease presence) has a strong positive correlation with age, cholesterol (chol), and resting blood pressure (trestbps), among others. - Age is positively correlated with cholesterol and resting blood pressure, suggesting that older individuals tend to have higher values for these variables. - Cholesterol and resting blood pressure are also positively correlated with each other. - Some variables, such as sex and slope, have weak or negligible correlations with most other variables

```
# Calculate correlation matrix
cor_matrix <- cor(heart_data[, sapply(heart_data, is.numeric)])  # select only numeric c
olumns

# Use melt from reshape2 to convert the correlation matrix for ggplot2
cor_matrix_melted <- melt(cor_matrix)

# Visualize the correlation matrix using ggplot2
ggplot(cor_matrix_melted, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("Correlation Matrix")
```



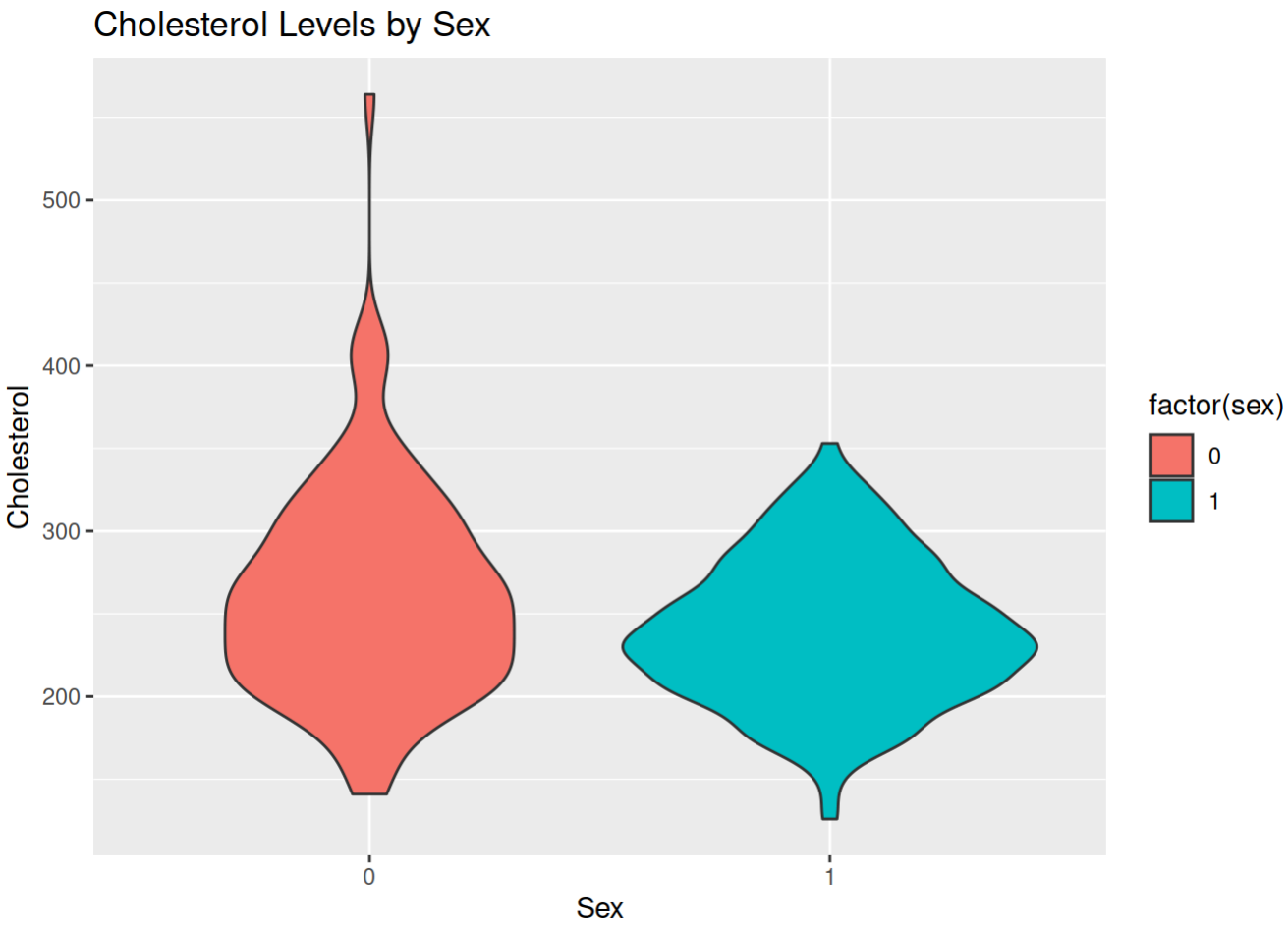
Group Comparisons

Cholesterol Levels by Sex

This violin plot shows the distribution of cholesterol levels for males (0) and females (1) in the dataset. The shape of the violin represents the probability density, with wider sections indicating higher concentrations of data points. The violin plot reveals that males tend to have higher cholesterol levels than females, as indicated by the wider

and more right-skewed distribution for males. The peak of the distribution for females is slightly lower than that for males, suggesting that females generally have lower cholesterol levels in this dataset

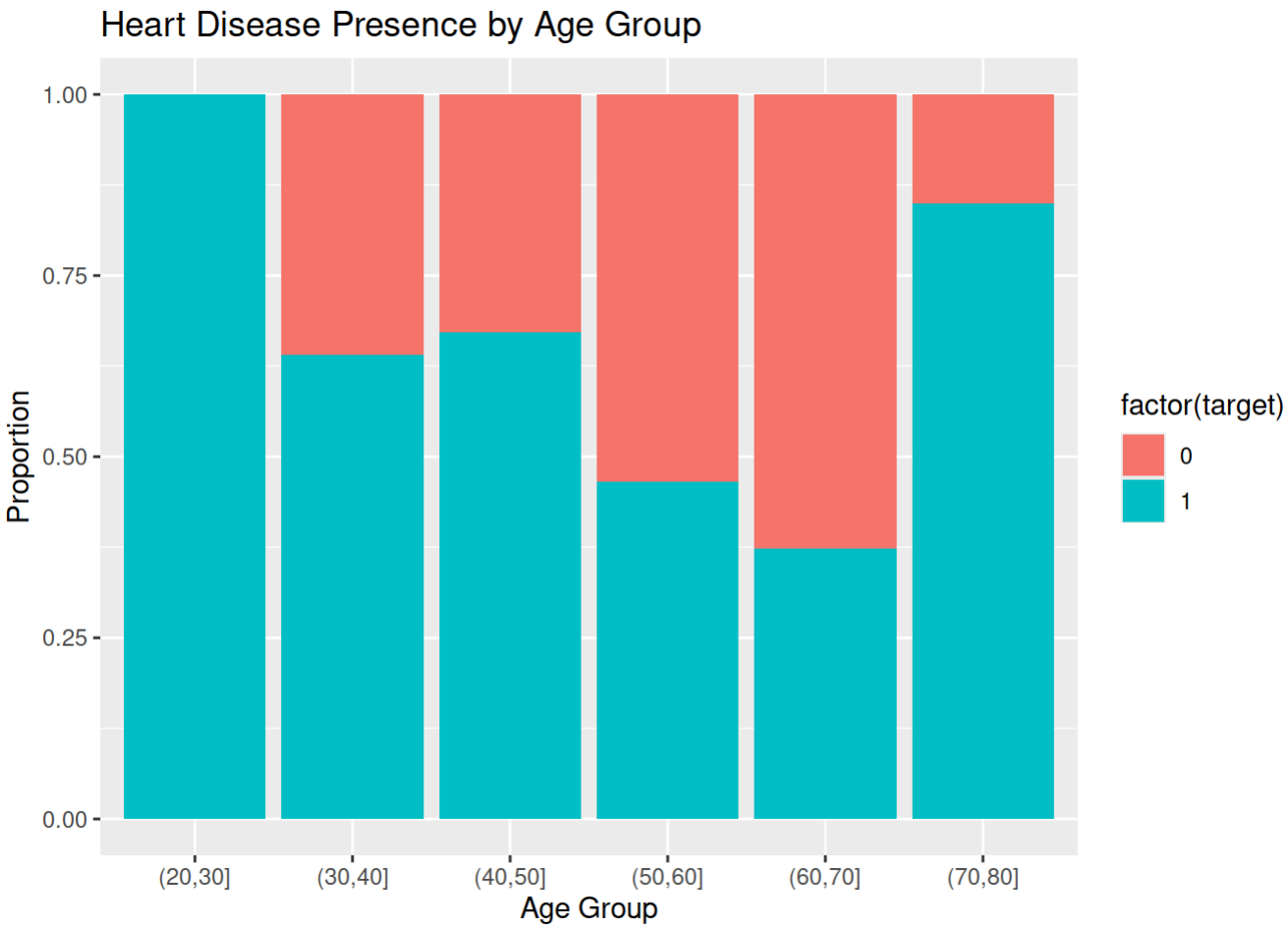
```
ggplot(heart_data, aes(x = factor(sex), y = chol, fill = factor(sex))) +
  geom_violin() +
  labs(title = "Cholesterol Levels by Sex", x = "Sex", y = "Cholesterol")
```



Heart Disease Presence by Age Group

- In the youngest age group (20-30), all patients are free of heart disease. The proportion of patients with heart disease increases with age, peaking in the 50-60 age group.
- In the 60-70 age group, the proportion of patients with heart disease is slightly lower than in the 50-60 age group.
- The oldest age group (70-80) shows a high proportion of patients with heart disease, but not as high as the 50-60 age group.
- This visualization indicates that heart disease prevalence generally increases with age, with a notable peak in the 50-60 age group.

```
heart_data$age_group <- cut(heart_data$age, breaks = seq(20, 80, by = 10))
ggplot(heart_data, aes(x = age_group, fill = factor(target))) +  # Assuming 'target' is
the heart disease variable
  geom_bar(position = "fill") +
  labs(title = "Heart Disease Presence by Age Group", x = "Age Group", y = "Proportion")
```



Feature Engineering

Interaction Terms

```
# Interaction between Age and Max Heart Rate
heart_data$age_thalach_interaction <- heart_data$age * heart_data$thalach
```

Binning Continuous Variables

```
heart_data$age_category <- cut(heart_data$age, breaks=c(29, 40, 55, 120), labels=c("youn
g", "middle-aged", "elderly"), include.lowest=TRUE)
```

Polynomial Features

```
heart_data$age_squared <- heart_data$age^2
heart_data$oldpeak_squared <- heart_data$oldpeak^2
```

Creating Dummy Variables

```
heart_data <- cbind(heart_data, model.matrix(~age_category - 1, data=heart_data))
```

Analysis & Visualization

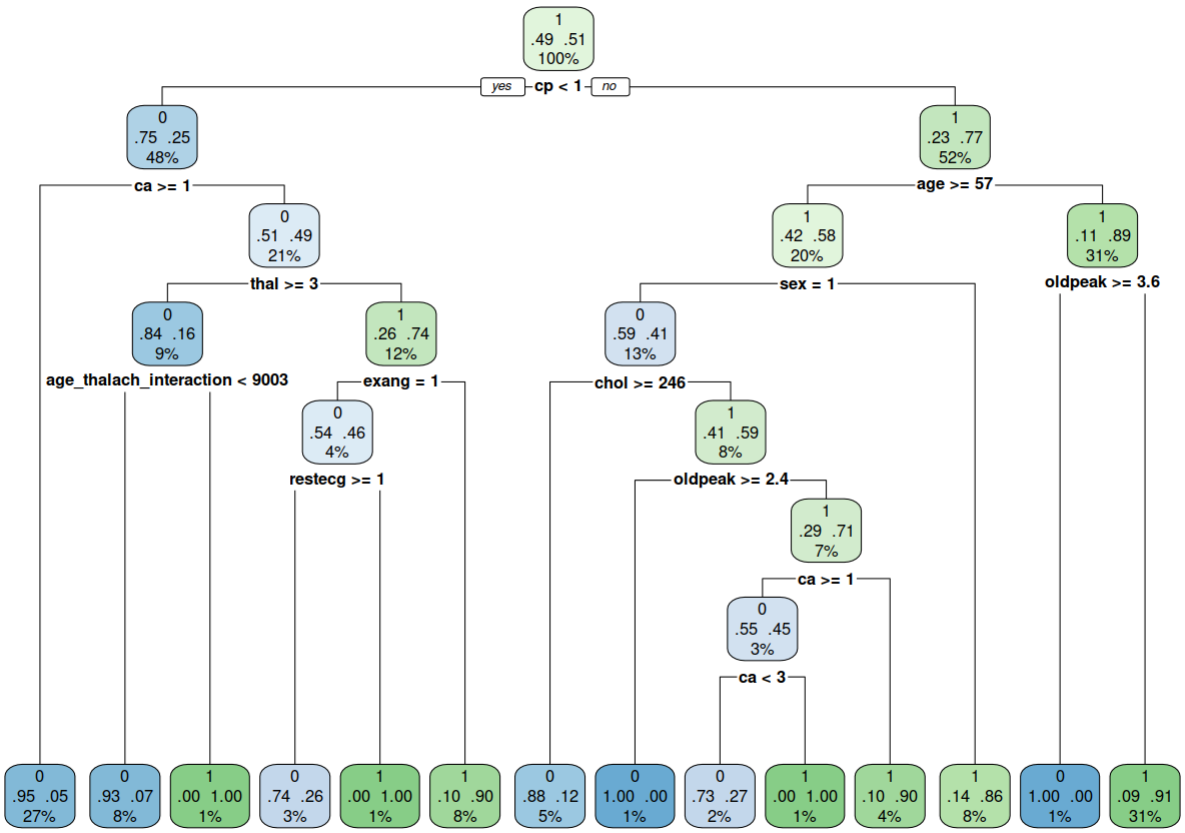
Decision Tree Model

- The root node splits based on the feature “cp” (chest pain type).
- Subsequent splits are based on features such as “ca” (number of major vessels colored by fluoroscopy), “thal” (thalassemia), “age”, “sex”, “chol” (cholesterol), and “oldpeak” (ST depression induced by exercise relative to rest).
- The leaf nodes at the bottom show the final predictions, with the proportion of patients with and without heart disease.

This decision tree provides insights into the importance of various features in predicting heart disease. For example, chest pain type, number of major vessels, and thalassemia are key features in the model.

```
# Building the decision tree model
# Assume 'target' is your outcome variable and you are using all other columns as predic
tors
tree_model <- rpart(target ~ ., data=heart_data, method="class")
```

```
# Visualize the decision tree
rpart.plot(tree_model, extra=104)  # Provides a detailed visual of the tree
```



Advanced Visualizations

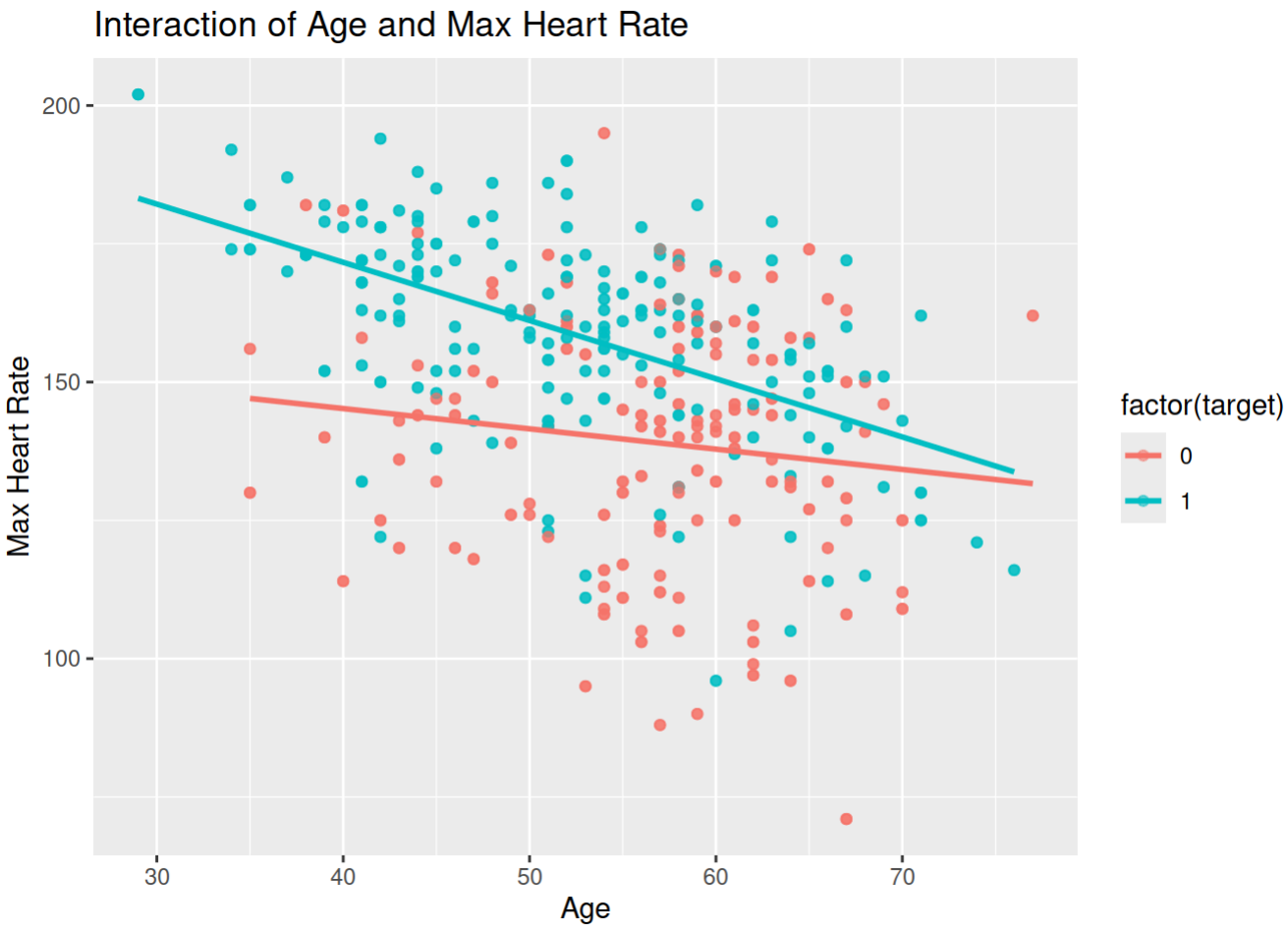
Interaction of Age and Max Heart Rate

- There is a negative correlation between age and maximum heart rate, indicating that older patients tend to have lower maximum heart rates.
- Patients without heart disease (0) generally have higher maximum heart rates compared to those with heart disease (1).
- The regression lines for both groups show a downward trend, with the line for patients with heart disease being lower.

This visualization highlights the relationship between age, maximum heart rate, and heart disease status, suggesting that lower maximum heart rates in older patients may be associated with heart disease.

```
ggplot(heart_data, aes(x=age, y=thalach, color=factor(target))) +
  geom_point(alpha=0.5) +
  geom_smooth(method="lm", se=FALSE) +
  labs(title="Interaction of Age and Max Heart Rate", x="Age", y="Max Heart Rate")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



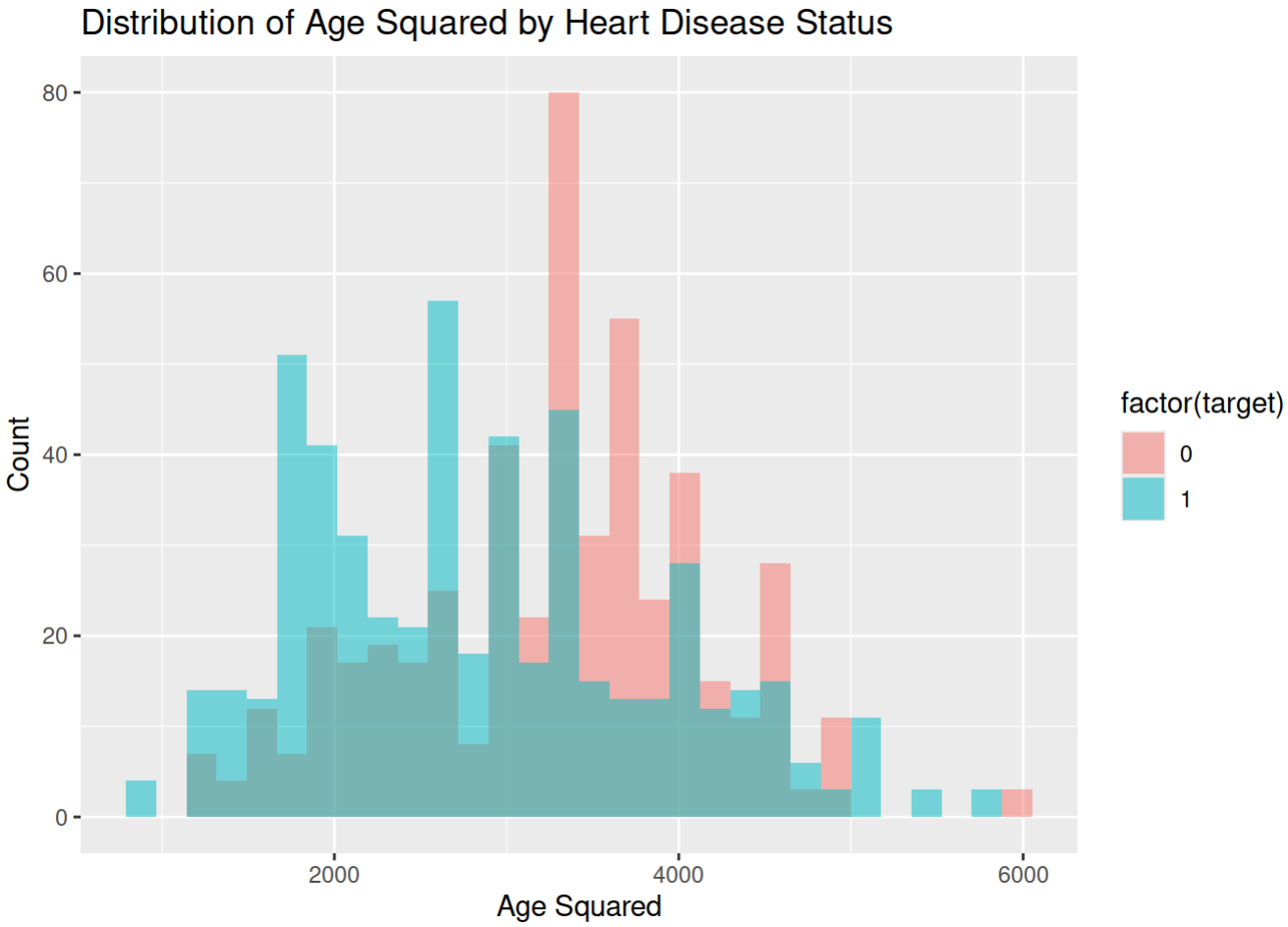
Distribution of Age Squared by Heart Disease Status

- The distribution for patients without heart disease (0) is more spread out, with peaks at lower values of age squared.

- The distribution for patients with heart disease (1) shows a higher concentration at higher values of age squared.
- There is a noticeable overlap between the two distributions, but patients with heart disease tend to have higher age squared values.

This visualization suggests that higher age squared values (indicating older age) are more common among patients with heart disease, reinforcing the trend observed in the age group analysis.

```
ggplot(heart_data, aes(x=age_squared, fill=factor(target))) +
  geom_histogram(position="identity", alpha=0.5, bins=30) +
  labs(title="Distribution of Age Squared by Heart Disease Status", x="Age Squared", y
="Count")
```



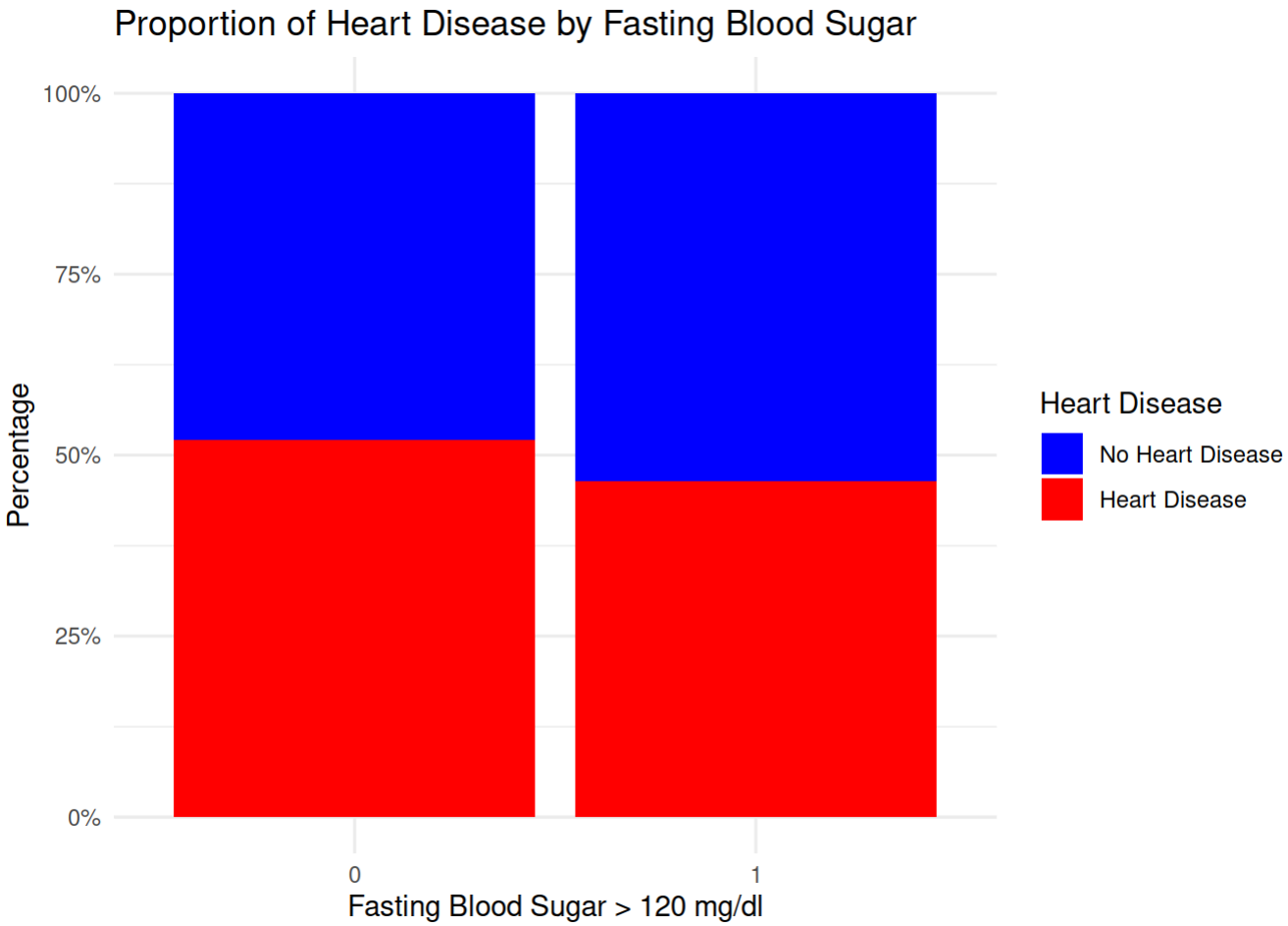
Risk Factor Analysis

Proportion of Heart Disease by Fasting Blood Sugar

- For patients with fasting blood sugar levels less than or equal to 120 mg/dL (0), the proportion of patients with heart disease (red) is slightly less than those without heart disease (blue).
- For patients with fasting blood sugar levels greater than 120 mg/dL (1), the proportion of patients with heart disease is higher compared to those without heart disease.

This visualization suggests that higher fasting blood sugar levels are associated with an increased proportion of heart disease.

```
ggplot(heart_data, aes(x = factor(fbs), fill = factor(target))) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Proportion of Heart Disease by Fasting Blood Sugar",
       x = "Fasting Blood Sugar > 120 mg/dl",
       y = "Percentage",
       fill = "Heart Disease") +
  scale_fill_manual(values = c("0" = "blue", "1" = "red"),
                    labels = c("No Heart Disease", "Heart Disease")) +
  theme_minimal()
```



Probability of Heart Disease Across Age

- This line plot shows the probability of heart disease across different ages.
- The x-axis represents age, and the y-axis represents the probability of heart disease.
- The plot indicates fluctuations in the probability of heart disease across different ages.
- There are peaks and troughs, suggesting that the risk of heart disease varies significantly with age.
- The probability is generally higher in older age groups, but there are notable variations.

```r
# Grouping data by age and calculating the probability of heart disease
age_data <- heart_data %>%
  group_by(age) %>%
  summarise(HeartDisease = mean(target == 1), .groups = 'drop')

# Creating the point plot
ggplot(age_data, aes(x = age, y = HeartDisease)) +
  geom_point() +
  geom_line(group=1, colour="red") +
  labs(title = "Probability of Heart Disease Across Age",
       x = "Age",
       y = "Probability of Heart Disease") +
  theme_minimal()
```