Heart Disease Dataset Analysis

Group I

Data Exploration & Visualization

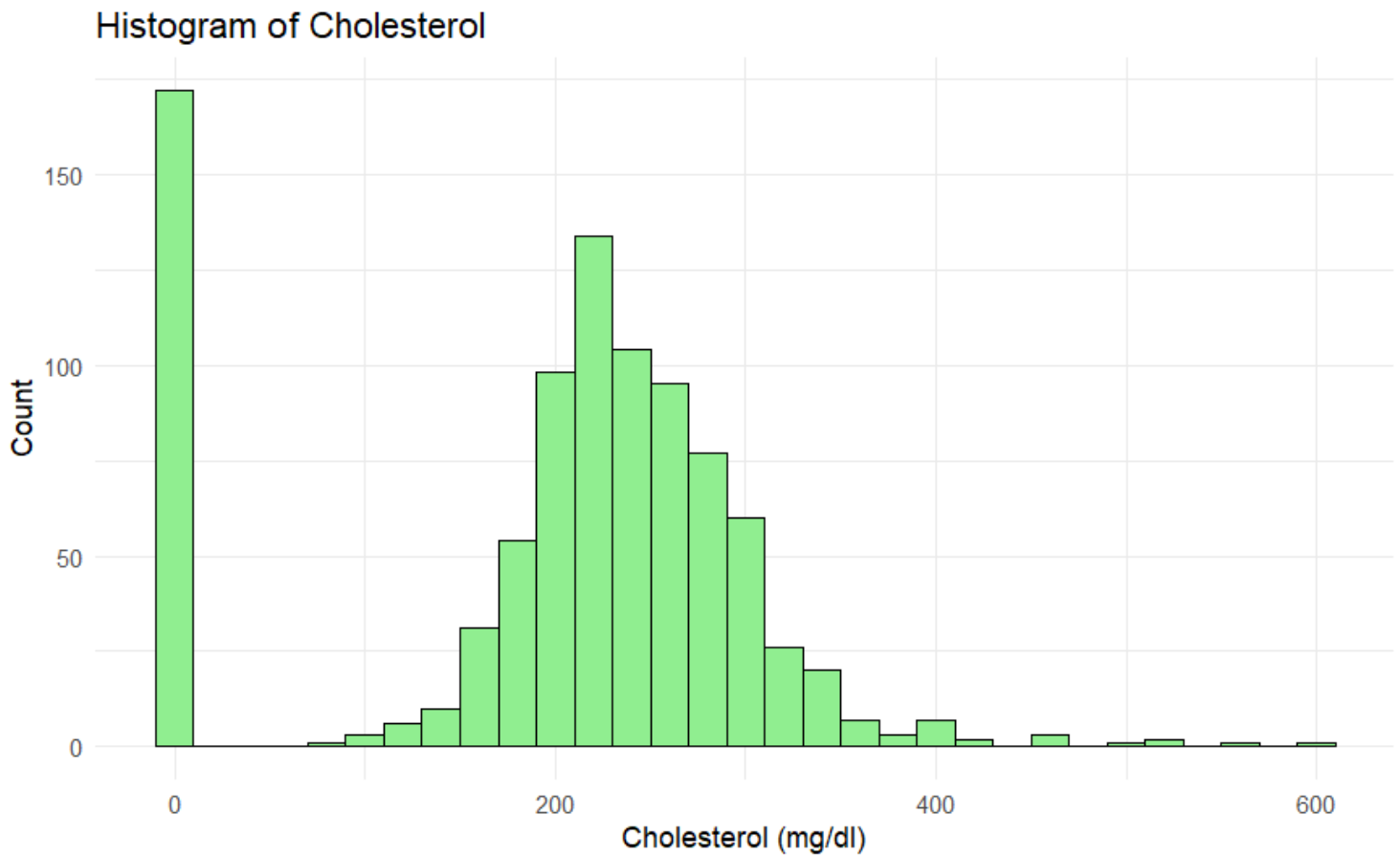
- 918 observations and 12 variables
- 10 variables are numeric and 2 are character

```
'data.frame':   918 obs. of  12 variables:
 $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
 $ Sex           : num  1 0 1 0 1 1 0 1 1 0 ...
 $ ChestPainType : chr  "ATA" "NAP" "ATA" "ASY" ...
 $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
 $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
 $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ RestingECG    : num  0 0 1 0 0 0 0 0 0 0 ...
 $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
 $ ExerciseAngina: num  0 0 0 1 0 0 0 0 1 0 ...
 $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
 $ ST_Slope      : chr  "Up" "Flat" "Up" "Flat" ...
 $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

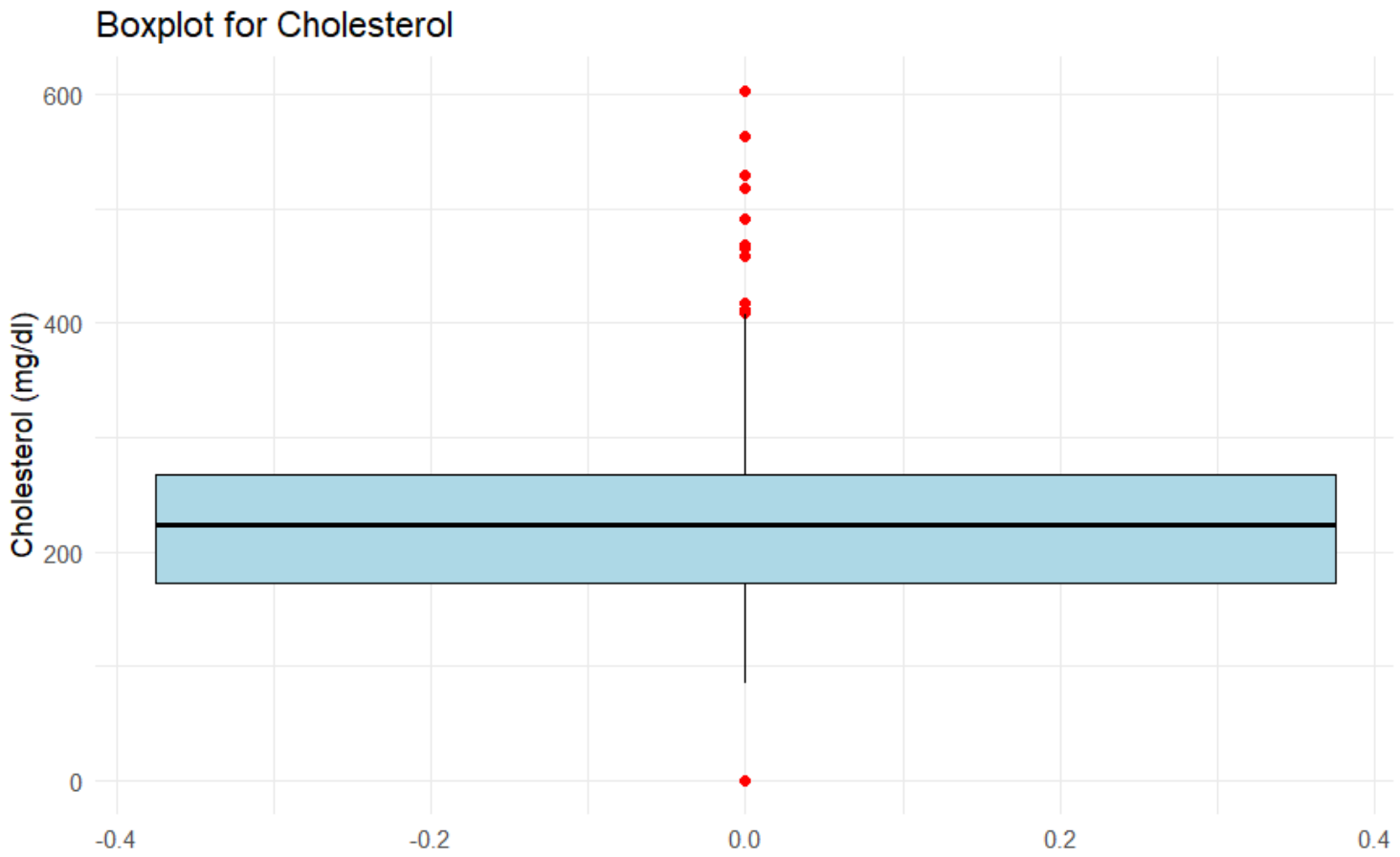
Data Exploration & Visualization

Variable	Definition	Mean	Std. Dev.	Min	Max
Age	Age of the patient in years	54.00	9.432	28.00	77.00
Sex	Gender of patient (female=0, male=1)	0.7898	0.4077	0.0	1.0
Chest Pain Type	Type of chest pain experienced: TA (Typical Angina), ATA(Atypical Angina), NAP (Non_Anginal Pain), ASY (Asympotomatic)				
Resting BP	The resting blood pressure of the patient, measured in mm Hg	132.4	18.514	0.0	200.0
Cholesterol	Total cholesterol level in blood, measured in mg/dl	198.8	109384.000	0.0	603.0
Fasting BS	Fasting sugar level (means < 120mg/dl= 0, means > 120 mg/dl = 1)	0.2331	0.4320	0.0	1.0
Resting ECG	Results of the resting electrocardiogram: Normal, ST, LVH	0.1939	0.396	0.0	1.0
Max HR	The maximum heart rate achieved by the patient during exercise	136.8	25.460	60.0	202.0
Exercise Angina	Angina (chest pain) induced by exercise (No =0, Yes=1)	0.4041	0.491	0.0	1.0
Oldpeak	The ST depression induced by exercise relative to rest, measured in mm	0.8874	1.067	(-)2.6	6.00
ST Slope	The slope of the ST segment during exercise: Up (upsloping), Flat, Down (downsloping)				
Heart Disease	The target variable: 0 = no heart diesease, 1= heart disease)	0.5534	0.497	0.0	1.0

Data Exploration & Visualization



Data Exploration & Visualization

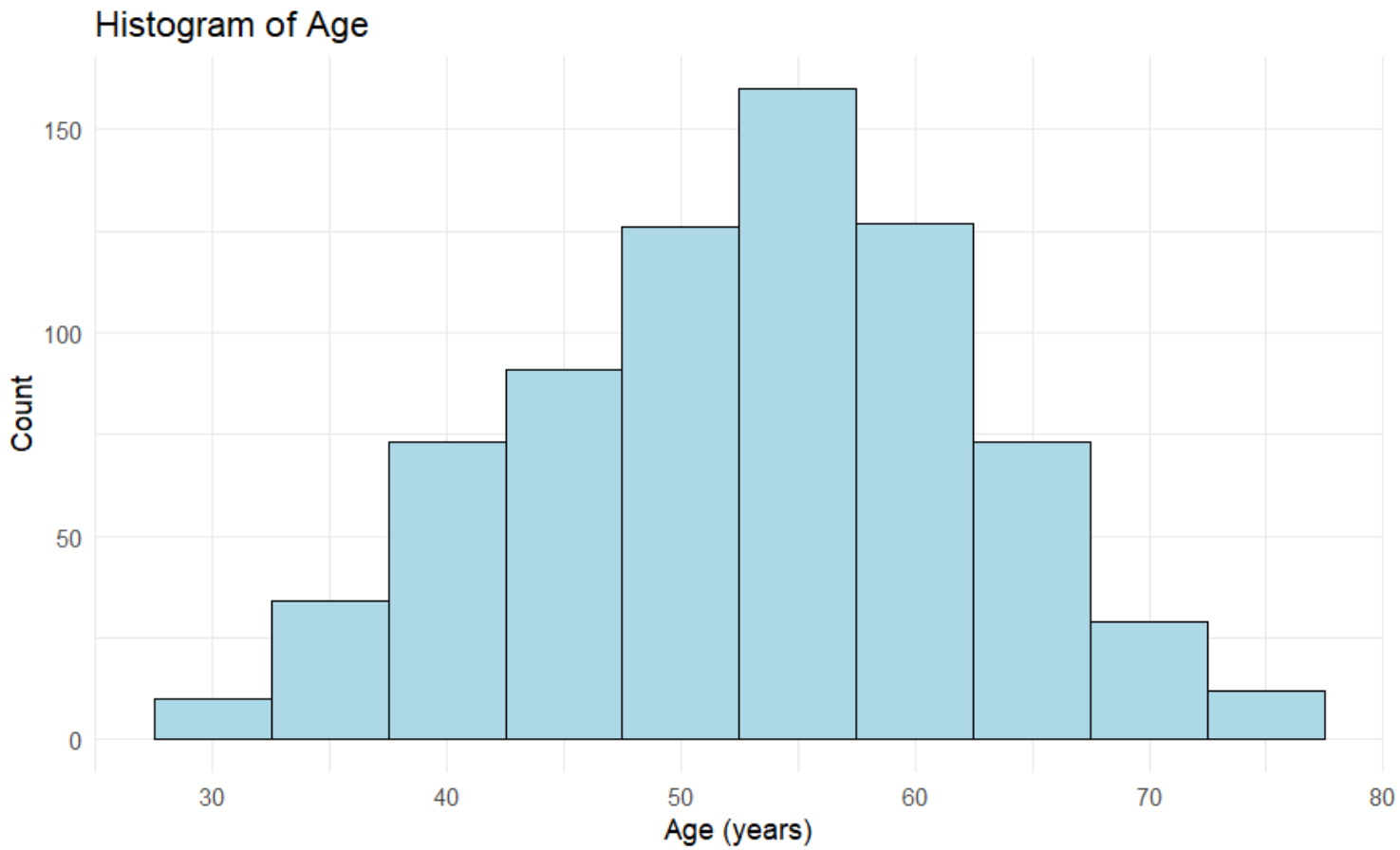


Data Exploration & Visualization

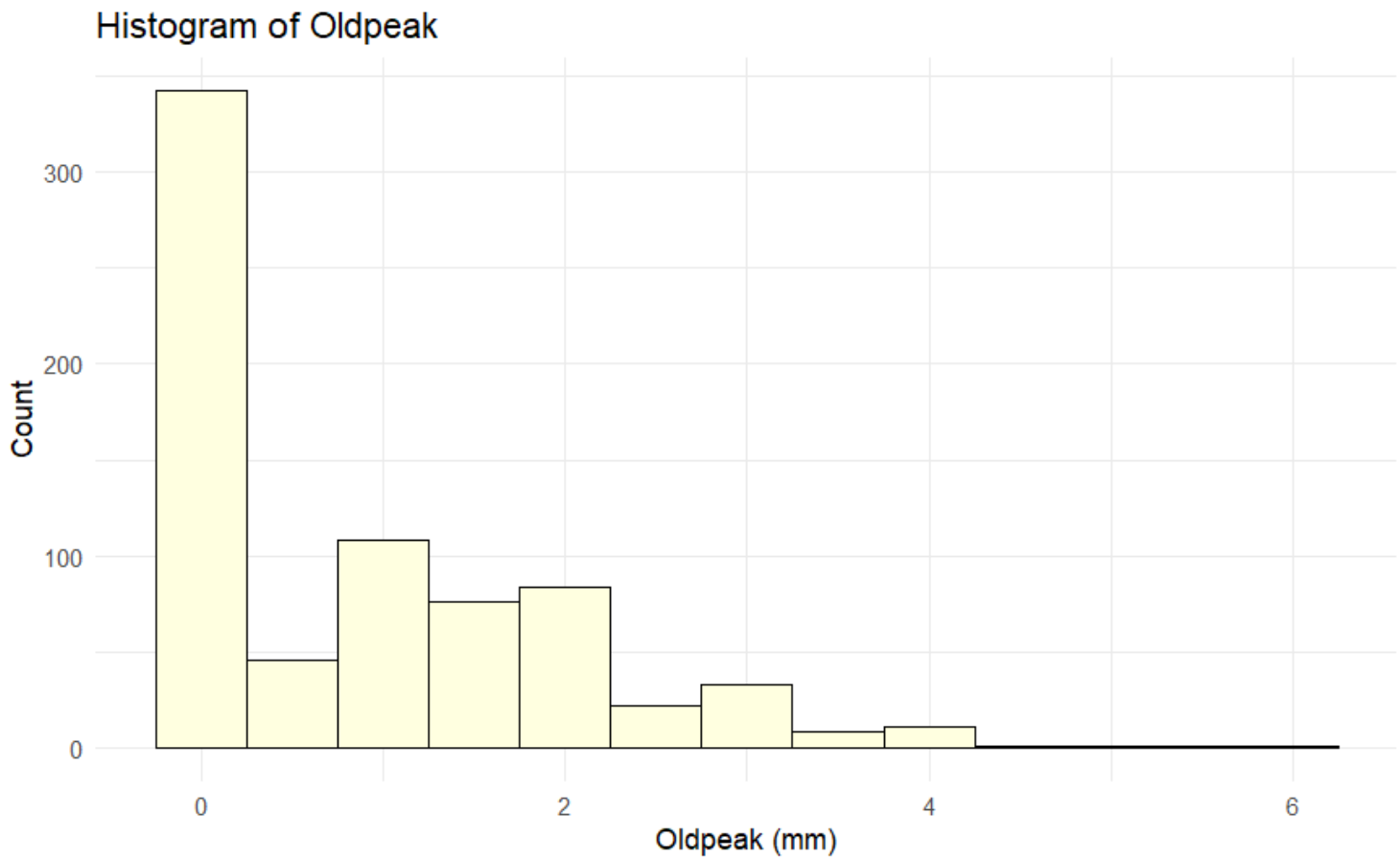
- After removing the outliers: 735 obervations and 12 variables

```
'data.frame':   735 obs. of  12 variables:
 $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
 $ Sex           : num  1 0 1 0 1 1 0 1 1 0 ...
 $ ChestPainType : chr  "ATA" "NAP" "ATA" "ASY" ...
 $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
 $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
 $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ RestingECG    : num  0 0 1 0 0 0 0 0 0 0 ...
 $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
 $ ExerciseAngina: num  0 0 0 1 0 0 0 0 1 0 ...
 $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
 $ ST_Slope      : chr  "Up" "Flat" "Up" "Flat" ...
 $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

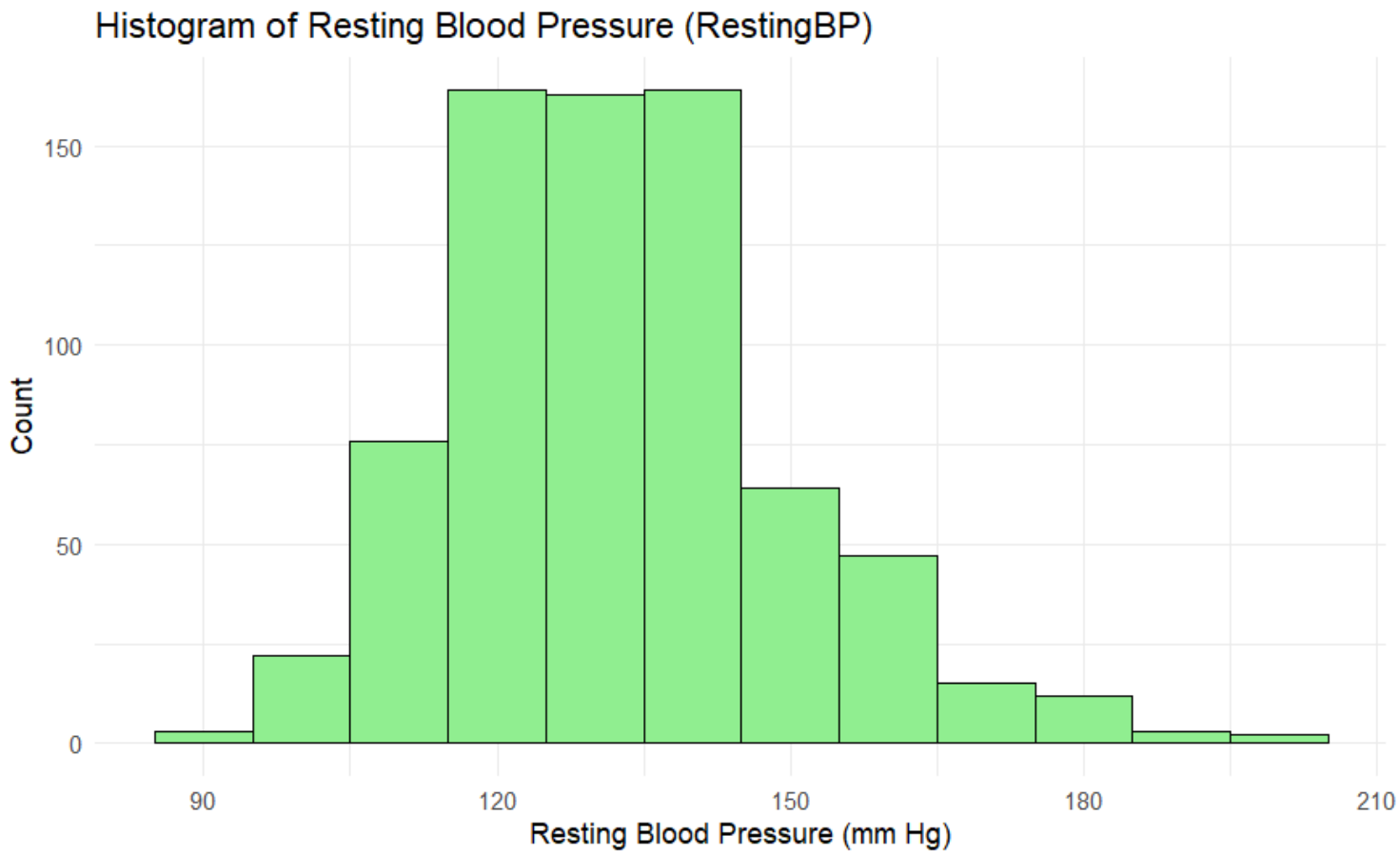
Data Exploration & Visualization



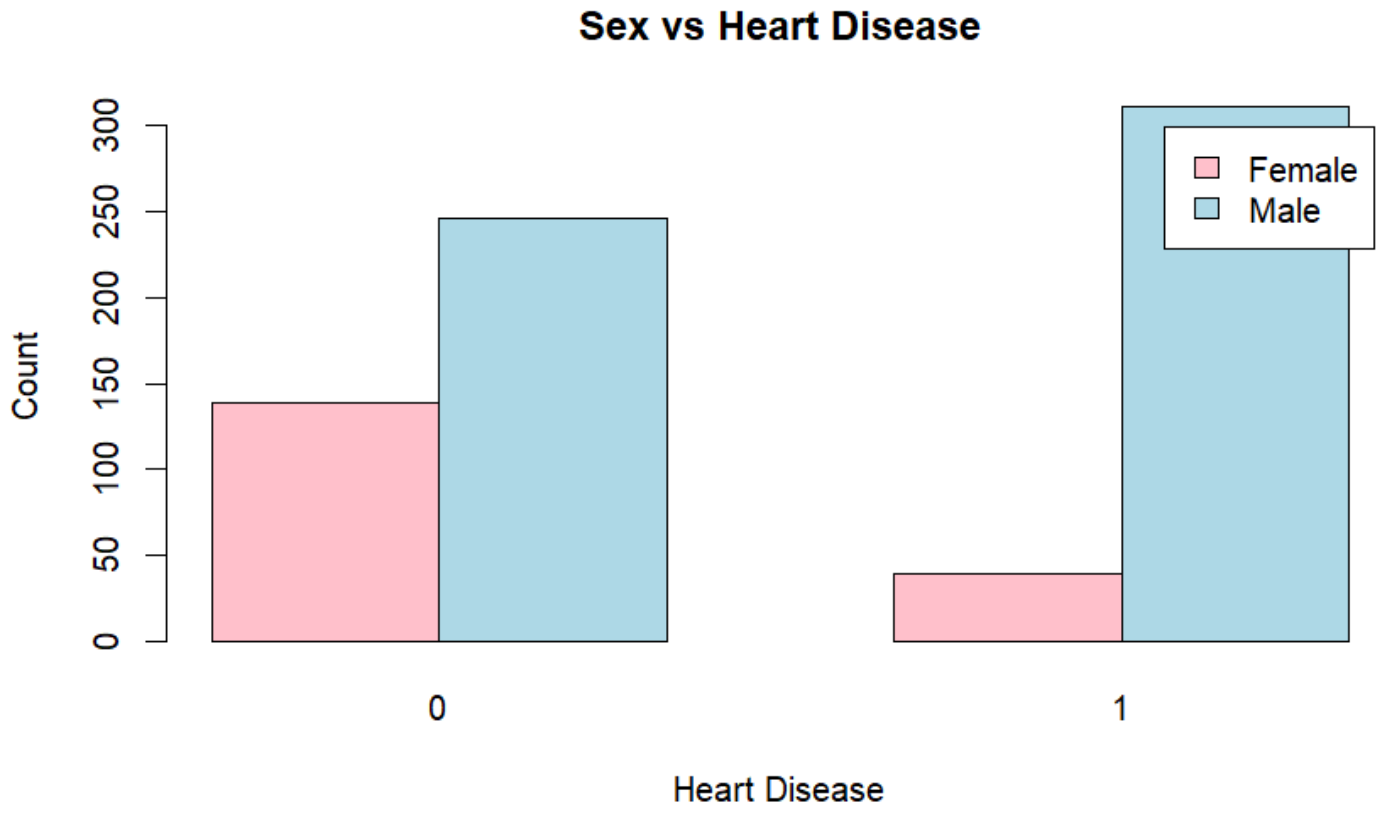
Data Exploration & Visualization



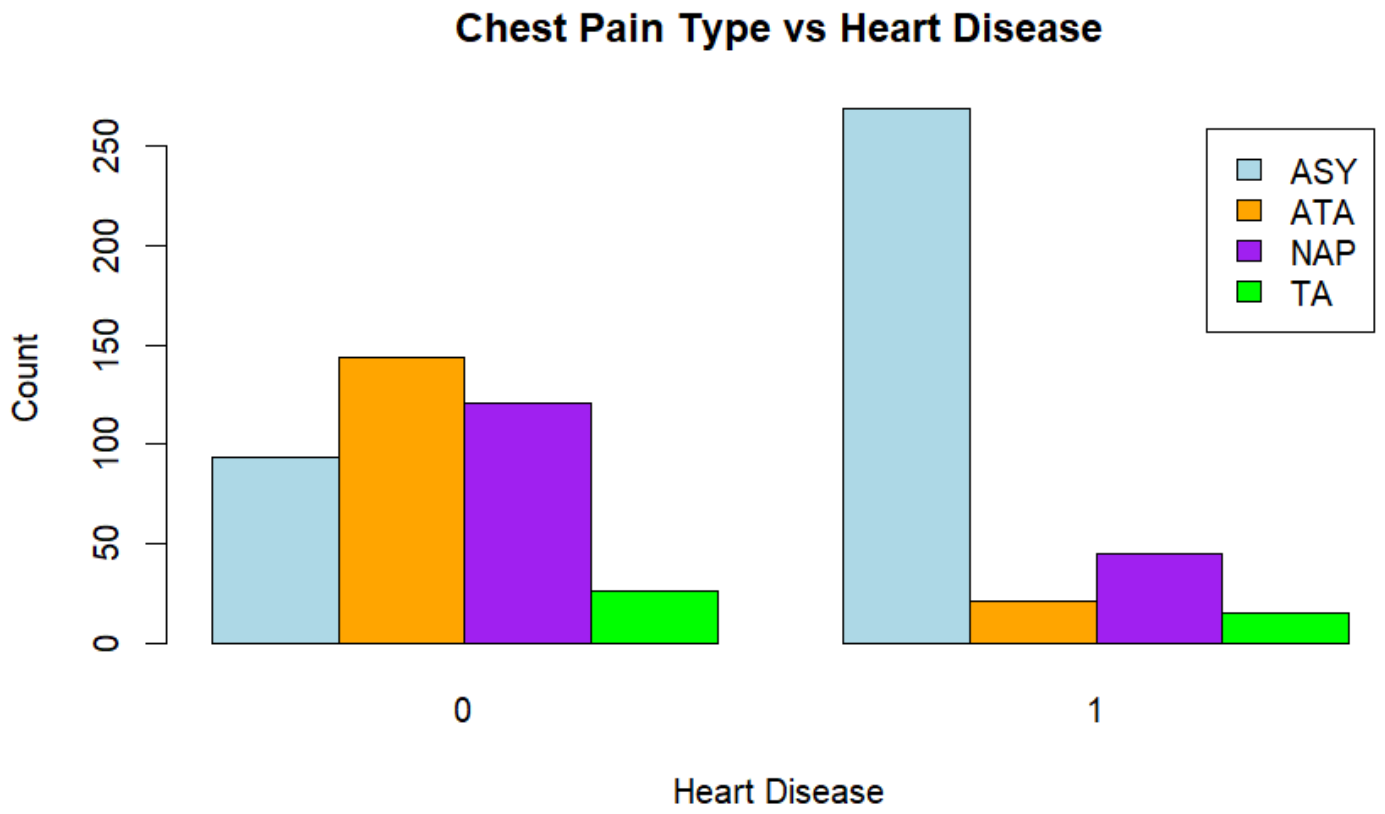
Data Exploration & Visualization



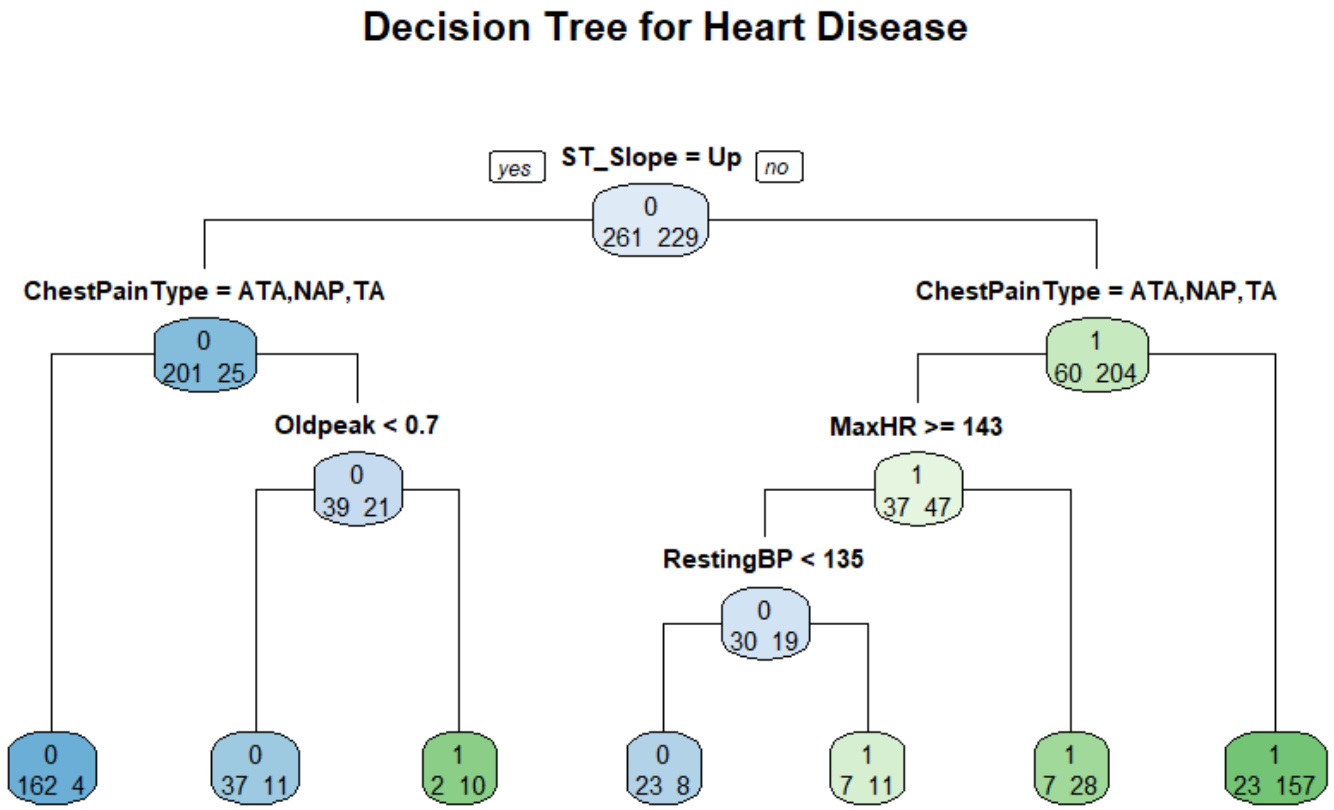
Data Exploration & Visualization



Data Exploration & Visualization



Decision Tree



Trainings Accuracy: 87.35 % , Test Accuracy: 83.67 %

Intepretation Decision Tree

- True Positive: 84.08 %
- False Positive: 15.92 %
- False Negative: 9.39 %

Logistic Regression

```
Call:
glm(formula = HeartDisease ~ ., family = "binomial", data = data_train)

Coefficients:
                 Estimate Std. Error z value    Pr(>|z|)
(Intercept)      -6.470697   2.432405  -2.660     0.00781 **
Age               0.040249   0.019466   2.068     0.03867 *
Sex               2.216785   0.440984   5.027 0.000000498 ***
ChestPainTypeATA -2.209324   0.506848  -4.359 0.000013069 ***
ChestPainTypeNAP -1.779697   0.409999  -4.341 0.000014201 ***
ChestPainTypeTA  -1.335044   0.695406  -1.920     0.05488 .
RestingBP         0.008839   0.009783   0.903     0.36630
Cholesterol       0.007214   0.003441   2.097     0.03603 *
FastingBS         0.337619   0.440739   0.766     0.44366
RestingECG       -0.356278   0.462916  -0.770     0.44151
MaxHR             0.003850   0.008122   0.474     0.63550
ExerciseAngina    0.552189   0.360911   1.530     0.12602
Oldpeak           0.516968   0.189287   2.731     0.00631 **
ST_SlopeFlat      0.500044   0.720163   0.694     0.48746
ST_SlopeUp       -2.200385   0.767971  -2.865     0.00417 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Trainings Accuracy: 87.07 % , Test Accuracy: 87.07 %

Intepretation Logistic Regression

- **Age:** For each additional year of age, the odds of having heart disease increase about 4.1 %
- **Sex:** Being male (compared to female) increases the odds of heart disease by a factor of 9.18
- **Chest Pain Type:**

→ Having typical angina (ATA) decreases the odds of heart disease by approximately 89% compared to asymptomatic individuals

→ Having non-anginal pain (NAP) decreases the odds of heart disease by approximately 83% compared to asymptomatic individuals

Intepretation Logistic Regression

- **Cholesterol:** For every unit increase in cholesterol, the odds of heart disease increase about 0.7%
- **Oldpeak**: Each unit increase in Oldpeak (ST depression induced by exercise) increases the odds of heart disease by 68%
- **ST Slope:** An upsloping ST segment reduces the odds of heart disease by about 89%

Sources

- https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction