Mid Term Project

Introduction to Data Science

Topic: Heart Attack Analysis & Prediction Dataset

Group Member:

MD.Rayhan Talukder ID:20-43764-2

Rowjatul Jannat ID:20-43976-2

Group Number : 19

Section: C

Dataset Description

A dataset for heart attack classification . Based on the dataset common health-related features, here's a brief description:

1. **Age:**
   - Numeric variable representing the age of individuals.
2. **Sex:**
   - Categorical variable indicating the gender (Male/Female).
3. **ChestPainType:**
   - Categorical variable describing the type of chest pain (e.g., ATA, NAP, ASY).
4. **RestingBP:**
   - Numeric variable representing the resting blood pressure.
5. **Cholesterol:**
   - Numeric variable indicating the cholesterol level.
6. **FastingBS:**
   - Binary variable (0/1) representing fasting blood sugar status.
7. **RestingECG:**
   - Categorical variable describing the resting electrocardiographic results.
8. **MaxHR:**
   - Numeric variable indicating the maximum heart rate achieved.
9. **ExerciseAngina:**
   - Categorical variable indicating the presence of exercise-induced angina (Y/N).
10. **Oldpeak:**
    - Numeric variable representing the ST depression induced by exercise relative to rest.
11. **ST_Slope:**
    - Categorical variable describing the slope of the peak exercise ST segment.
12. **HeartDisease:**
    - Binary variable (0/1) indicating the presence of heart disease.

Import the data set as csv and print the data set:

```
mydata <- read.csv("C:/Heart.csv", header = TRUE, sep = ",")
mydata
```

Output:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina
1	40	M	ATA	140	289	0	Normal	172	N
2	49	F	NAP	160	180	0	Normal	156	N
3	37	M	ATA	130	283	0	ST	98	N
4	NA	F	ASY	138	214	0	Normal	108	Y
5	54	M	NAP	-150	195	0	Normal	122	N
6	39	M	NAP	120	339	0	Normal	170	N
7	45	F	ATA	130	237	0	Normal	170	
8	54	M	ATA	110	208	0	Normal	142	N
9	37		ASY	140	207	0	Normal	130	Y
10	48	F	ATA	120	284	0	Normal	120	N
11	37	F	NAP	130	1000	0	Normal	142	N
12	58	M	ATA	136	1005	0	ST	99	Y
13	39	M	ATA	120	204	0	Normal	145	N
14	49	M	ASY	140	234	0	Normal	140	Y
15	42	M	NAP	115	211	0	ST	137	N
16	54	M	ATA	120	273	0	Normal	150	N
17	38	M	ASY	110	196	0	Normal	166	N
18	43	F	ATA	120	201	0	Normal	165	N
19	60	M	ASY	100	248	0	Normal	125	N
20	36	M	ATA	120	267	0	Normal	160	N
21	43	M	TA	100	223	0	Normal	142	N
22	44	M	ATA	120	184	0	Normal	142	N
23	49	F	ATA	124	201	0	Normal	164	N
24	NA	M	ATA	150	288	0	Normal	150	Y
25	40	M	NAP	130	215	0	Normal	138	N
26	36		NAP	130	209	0	Normal	178	N
27	53	M	ASY	124	260	0	ST	112	
28	52	M	ATA	120	284	0	Normal	118	N
29	53	F	ATA	113	468	0	Normal	127	N
30	51	M	ATA	125	188	0	Normal	145	N
31	53	M	NAP	145	518	0	Normal	130	N
32	56	M	NAP	130	167	0	Normal	114	N
33	NA	M	ASY	125	224	0	Normal	122	N
34	41	M	ASY	130	172	0	ST	130	N
35	43	F	ATA	150	186	0	Normal	154	N

Description :

Here is the code of import the dataset as csv file. It is the output of the dataset which is imported in RStudio.

To see the column name of the data set:
Code :

```
4  names(mydata)
5
```

Output:

```
 [1] "Age"            "Sex"            "ChestPainType"  "RestingBP"      "Cholesterol"
 [6] "FastingBS"      "RestingECG"     "MaxHR"          "ExerciseAngina" "Oldpeak"
[11] "ST_Slope"       "HeartDisease"
```

Description : In this code ,we can see the column name of the dataset. Here with this code can see the attributes names. The output of the name() function where we can see the attributes of the dataset.

Annotating datasets:

```
mydata$Sex <- factor(mydata$Sex,
                     levels = c("M","F"),
                     labels = c(1,0))
mydata
```

Output:

```
> mydata$Sex <- factor(mydata$Sex,
+                      levels = c("M","F"),
+                      labels = c(1,0))
> mydata
   Age  Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR ExerciseAngina
1   40    1           ATA       140         289         0     Normal   172              N
2   49    0           NAP       160         180         0     Normal   156              N
3   37    1           ATA       130         283         0         ST    98              N
4   NA    0           ASY       138         214         0     Normal   108              Y
5   54    1           NAP      -150         195         0     Normal   122              N
6   39    1           NAP       120         339         0     Normal   170              N
7   45    0           ATA       130         237         0     Normal   170
8   54    1           ATA       110         208         0     Normal   142              N
9   37 <NA>           ASY       140         207         0     Normal   130              Y
10  48    0           ATA       120         284         0     Normal   120              N
11  37    0           NAP       130        1000         0     Normal   142              N
12  58    1           ATA       136        1005         0         ST    99              Y
13  39    1           ATA       120         204         0     Normal   145              N
14  49    1           ASY       140         234         0     Normal   140              Y
15  42    1           NAP       115         211         0         ST   137              N
16  54    1           ATA       120         273         0     Normal   150              N
17  38    1           ASY       110         196         0     Normal   166              N
18  43    0           ATA       120         201         0     Normal   165              N
19  60    1           ASY       100         248         0     Normal   125              N
20  36    1           ATA       120         267         0     Normal   160              N
21  43    1            TA       100         223         0     Normal   142              N
22  44    1           ATA       120         184         0     Normal   142              N
23  49    0           ATA       124         201         0     Normal   164              N
24  NA    1           ATA       150         288         0     Normal   150              Y
25  40    1           NAP       130         215         0     Normal   138              N
26  36 <NA>           NAP       130         209         0     Normal   178              N
27  53    1           ASY       124         260         0         ST   112
28  52    1           ATA       120         284         0     Normal   118              N
29  53    0           ATA       113         468         0     Normal   127              N
30  51    1           ATA       125         188         0     Normal   145              N
31  53    1           NAP       145         518         0     Normal   130              N
```

Description:

The sex column is converted from numeric (0 and 1) to a factor with labels "male" and "female".

Missing values:

Code:

```
14
15 missing_values <- colSums(is.na(mydata))
16 missing_values
```

Output:

```
> missing_values <- colSums(is.na(mydata))
> missing_values
          Age           Sex  ChestPainType      RestingBP    Cholesterol      FastingBS
            3             3              0              0              0              0
   RestingECG         MaxHR ExerciseAngina        Oldpeak       ST_Slope   HeartDisease
            0             0              0              0              0              0
>
```

Description: It finds the row numbers where gender and Age have missing values .

Removing missing value:

Code:

```
mydata <- na.omit(mydata)
mydata
```

Output:

```
> mydata <- na.omit(mydata)
> mydata
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina
1	40	1	ATA	140	289	0	Normal	172	N
2	49	0	NAP	160	180	0	Normal	156	N
3	37	1	ATA	130	283	0	ST	98	N
5	54	1	NAP	-150	195	0	Normal	122	N
6	39	1	NAP	120	339	0	Normal	170	N
7	45	0	ATA	130	237	0	Normal	170	
8	54	1	ATA	110	208	0	Normal	142	N
10	48	0	ATA	120	284	0	Normal	120	N
11	37	0	NAP	130	1000	0	Normal	142	N
12	58	1	ATA	136	1005	0	ST	99	Y
13	39	1	ATA	120	204	0	Normal	145	N
14	49	1	ASY	140	234	0	Normal	140	Y
15	42	1	NAP	115	211	0	ST	137	N
16	54	1	ATA	120	273	0	Normal	150	N
17	38	1	ASY	110	196	0	Normal	166	N
18	43	0	ATA	120	201	0	Normal	165	N
19	60	1	ASY	100	248	0	Normal	125	N
20	36	1	ATA	120	267	0	Normal	160	N
21	43	1	TA	100	223	0	Normal	142	N
22	44	1	ATA	120	184	0	Normal	142	N
23	49	0	ATA	124	201	0	Normal	164	N
25	40	1	NAP	130	215	0	Normal	138	N
27	53	1	ASY	124	260	0	ST	112	
28	52	1	ATA	120	284	0	Normal	118	N
29	53	0	ATA	113	468	0	Normal	127	N
30	51	1	ATA	125	188	0	Normal	145	N
31	53	1	NAP	145	518	0	Normal	130	N
32	56	1	NAP	130	167	0	Normal	114	N
34	41	1	ASY	130	172	0	ST	130	N
35	43	0	ATA	150	186	0	Normal	154	N
36	54	1	ATA	125	254	0	Normal	155	N
37	65	1	ASY	140	306	1	Normal	87	Y
38	41	0	ATA	110	250	0	ST	142	N

Description: it removes the row for missing value.

Summary of the structure of data set:
Code:

```
11  str(mydata)
```

Output:

```
> str(mydata)
'data.frame':   145 obs. of  12 variables:
 $ Age           : num  40 49 37 54 39 45 54 48 37 58 ...
 $ Sex           : Factor w/ 2 levels "1","0": 1 2 1 1 1 2 1 2 2 1 ...
 $ ChestPainType : chr  "ATA" "NAP" "ATA" "NAP" ...
 $ RestingBP     : num  140 160 130 -150 120 130 110 120 130 136 ...
 $ Cholesterol   : num  289 180 283 195 339 ...
 $ FastingBS     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ RestingECG    : chr  "Normal" "Normal" "ST" "Normal" ...
 $ MaxHR         : num  172 156 98 122 170 170 142 120 142 99 ...
 $ ExerciseAngina: chr  "N" "N" "N" "N" ...
 $ Oldpeak       : num  0 1 0 0 0 0 0 0 0 2 ...
 $ ST_Slope      : chr  "Up" "Flat" "Up" "Up" ...
 $ HeartDisease  : num  0 1 0 0 0 0 0 0 0 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:6] 4 9 24 26 33 40
  ..- attr(*, "names")= chr [1:6] "4" "9" "24" "26" ...
```

Description :
The structure of the dataset is displayed using str().

Finding mean ,medium, max:

Code:

```
13  summary(mydata)
```

Output:

```
> summary(mydata)
      Age          Sex     ChestPainType        RestingBP       Cholesterol
 Min.   : 32.00   1:102   Length:144         Min.   :-150.0   Min.   :  85.0
 1st Qu.: 42.75   0: 42   Class :character   1st Qu.: 120.0   1st Qu.: 203.5
 Median : 49.50           Mode  :character   Median : 130.0   Median : 241.0
 Mean   : 50.01                              Mean   : 128.9   Mean   : 259.5
 3rd Qu.: 54.00                              3rd Qu.: 140.0   3rd Qu.: 277.5
 Max.   :172.00                              Max.   : 190.0   Max.   :1005.0
   FastingBS        RestingECG            MaxHR       ExerciseAngina        Oldpeak
 Min.   :0.00000   Length:144         Min.   : 82.0   Length:144         Min.   :0.0000
 1st Qu.:0.00000   Class :character   1st Qu.:124.0   Class :character   1st Qu.:0.0000
 Median :0.00000   Mode  :character   Median :140.0   Mode  :character   Median :0.0000
 Mean   :0.09028                      Mean   :140.2                      Mean   :0.5556
 3rd Qu.:0.00000                      3rd Qu.:156.5                      3rd Qu.:1.0000
 Max.   :1.00000                      Max.   :190.0                      Max.   :4.0000
   ST_Slope          HeartDisease
 Length:144         Min.   :0.0000
 Class :character   1st Qu.:0.0000
 Mode  :character   Median :0.0000
                    Mean   :0.3681
                    3rd Qu.:1.0000
                    Max.   :1.0000
```

Description: Here is the code to see the descriptive Statistics. To see descriptive statistic, we use the summary() function. In the output here min, max, median, and mean are shown.

Categorical value check for mode:

Code:

```
sex_summary <- table(mydata$sex)
print(sex_summary)
```

Output:

```
  1   0
102  42
```

Code:

```
Age_summary <- table(mydata$Age)
print(Age_summary)
```

Output:

```
 32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
  2   1   3   2   2   3   3   8   4   5   3   9   4   4   5   2   5   7   4   3  11   5  24
 55  56  57  58  59  60  61  63  65  66 170 172
  3   5   2   4   2   1   1   1   3   1   1   1
```

Description: It counts the age and male female number for sex column.

Insert new row:

Code:

```
new_row <- data.frame(  Age = 25,  Sex = 1,  ChestPainType = "NAP",  RestingBP = 120,  Cholesterol = 200,  FastingBS = 0,  Res
mydata <- rbind(mydata, new_row)
mydata
```

Output:

```
 [ reached 'max' / getOption( max.print ) -- omitted 62 rows ]
> new_row <- data.frame(  Age = 25,  Sex = 1,  ChestPainType = "NAP",  RestingBP = 120,  Cholesterol = 200,  FastingBS = 0,  Restin
gECG = "Normal",  MaxHR = 150,  ExerciseAngina = "N",  Oldpeak = 0.5,  ST_Slope = "Flat",  HeartDisease = 0)
> mydata <- rbind(mydata, new_row)
> mydata
   Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR ExerciseAngina Oldpeak ST_Slope HeartDisease
1   40   1           ATA       140         289         0     Normal   172              N     0.0       Up            0
2   49   0           NAP       160         180         0     Normal   156              N     1.0     Flat            1
3   37   1           ATA       130         283         0         ST    98              N     0.0       Up            0
5   54   1           NAP      -150         195         0     Normal   122              N     0.0       Up            0
6   39   1           NAP       120         339         0     Normal   170              N     0.0       Up            0
7   45   0           ATA       130         237         0     Normal   170                    0.0       Up            0
8   54   1           ATA       110         208         0     Normal   142              N     0.0       Up            0
10  48   0           ATA       120         284         0     Normal   120              N     0.0       Up            0
11  37   0           NAP       130        1000         0     Normal   142              N     0.0       Up            0
12  58   1           ATA       136        1005         0         ST    99              Y     2.0     Flat            1
13  39   1           ATA       120         204         0     Normal   145              N     0.0       Up            0
14  49   1           ASY       140         234         0     Normal   140              Y     1.0     Flat            1
15  42   1           NAP       115         211         0         ST   137              N     0.0       Up            0
16  54   1           ATA       120         273         0     Normal   150              N     1.5     Flat            0
17  38   1           ASY       110         196         0     Normal   166              N     0.0     Flat            1
18  43   0           ATA       120         201         0     Normal   165              N     0.0       Up            0
19  60   1           ASY       100         248         0     Normal   125              N     1.0     Flat            1
20  36   1           ATA       120         267         0     Normal   160              N     3.0     Flat            1
21  43   1            TA       100         223         0     Normal   142              N     0.0       Up            0
22  44   1           ATA       120         184         0     Normal   142              N     1.0     Flat            0
23  49   0           ATA       124         201         0     Normal   164              N     0.0       Up            0
25  40   1           NAP       130         215         0     Normal   138              N     0.0       Up            0
27  53   1           ASY       124         260         0         ST   112                    3.0     Flat            0
28  52   1           ATA       120         284         0     Normal   118              N     0.0       Up            0
29  53   0           ATA       113         468         0     Normal   127              N     0.0       Up            0
30  51   1           ATA       125         188         0     Normal   145              N     0.0       Up            0
31  53   1           NAP       145         518         0     Normal   130              N     0.0     Flat            1
32  56   1           NAP       130         167         0     Normal   114              N     0.0       Up            0
34  41   1           ASY       130         172         0         ST   130              N     2.0     Flat            1
```

Description: Adding new values as a new row

Correcting invalid value:

Code:

```
27  median_age <- median(mydata$Age, na.rm = TRUE)
28  mydata$Age[mydata$Age > 100 | is.na(mydata$Age)] <- median_age
29  mydata
30
```

Output:

42	54	0	NAP	130	294	0	ST	100	Y	0.0	Flat	1
43	35	1	ATA	150	264	0	Normal	168	N	0.0	Up	0
44	52	1	NAP	140	259	0	ST	170	N	0.0	Up	0
45	43	1	ASY	120	175	0	Normal	120	Y	1.0	Flat	1
46	59	1	NAP	130	318	0	Normal	120	Y	1.0	Flat	0
47	54	1	ASY	120	223	0	Normal	168	N	0.0	Up	0
48	50	1	ATA	140	216	0	Normal	170	N	0.0	Up	0
49	36	1	NAP	112	340	0	Normal	184	N	1.0	Flat	0
50	41	1	ASY	110	289	0	Normal	170	N	0.0	Flat	1
51	49	1	ASY	130	233	0	Normal	121	Y	2.0	Flat	1
52	47	0	ASY	120	205	0	Normal	98	Y	2.0	Flat	1
53	45	1	ATA	140	224	1	Normal	122	N	0.0	Up	0
54	41	0	ATA	130	245	0	Normal	150	N	0.0	Up	0
55	52	0	ASY	130	180	0	Normal	140	Y	1.5	Flat	0
56	51	0	ATA	160	194	0	Normal	170	N	0.0	Up	0
57	49	1	ASY	120	270	0	Normal	153	Y	1.5	Flat	1
58	58	1	NAP	130	213	0	ST	140	N	0.0	Flat	1
59	54	1	ASY	150	365	0	ST	134	N	1.0	Up	0
60	52	1	ASY	112	342	0	ST	96	Y	1.0	Flat	1
61	49	1	ATA	100	253	0	Normal	174	N	0.0	Up	0
62	43	0	NAP	150	254	0	Normal	175	N	0.0	Up	0
63	45	1	ASY	140	224	0	Normal	144	N	0.0	Up	0
64	46	1	ASY	120	277	0	Normal	125	Y	1.0	Flat	1
65	50	0	ATA	110	202	0	Normal	145	N	0.0	Up	0
66	37	0	ATA	120	260	0	Normal	130	N	0.0	Up	0
67	45	0	ASY	132	297	0	Normal	144	N	0.0	Up	0
68	32	1	ATA	110	225	0	Normal	184	N	0.0	Up	0
69	52	1	ASY	160	246	0	ST	82	Y	4.0	Flat	1

Description: It corrects the value of age over 100.

Measure of central tendency:

Code:

```
mean_age <- mean(mydata$Age, na.rm = TRUE)
median_age <- median(mydata$Age, na.rm = TRUE)
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
mode_age <- Mode(mydata$Age)
cat("Mean Age:", mean_age, "\n")
cat("Median Age:", median_age, "\n")
cat("Mode Age:", mode_age, "\n")
```

Output:

```
> cat( Mean Age: ", mean_age,  "\n" )
Mean Age: 47.99315
> cat("Median Age:", median_age, "\n")
Median Age: 49
> cat("Mode Age:", mode_age, "\n")
Mode Age: 54
```

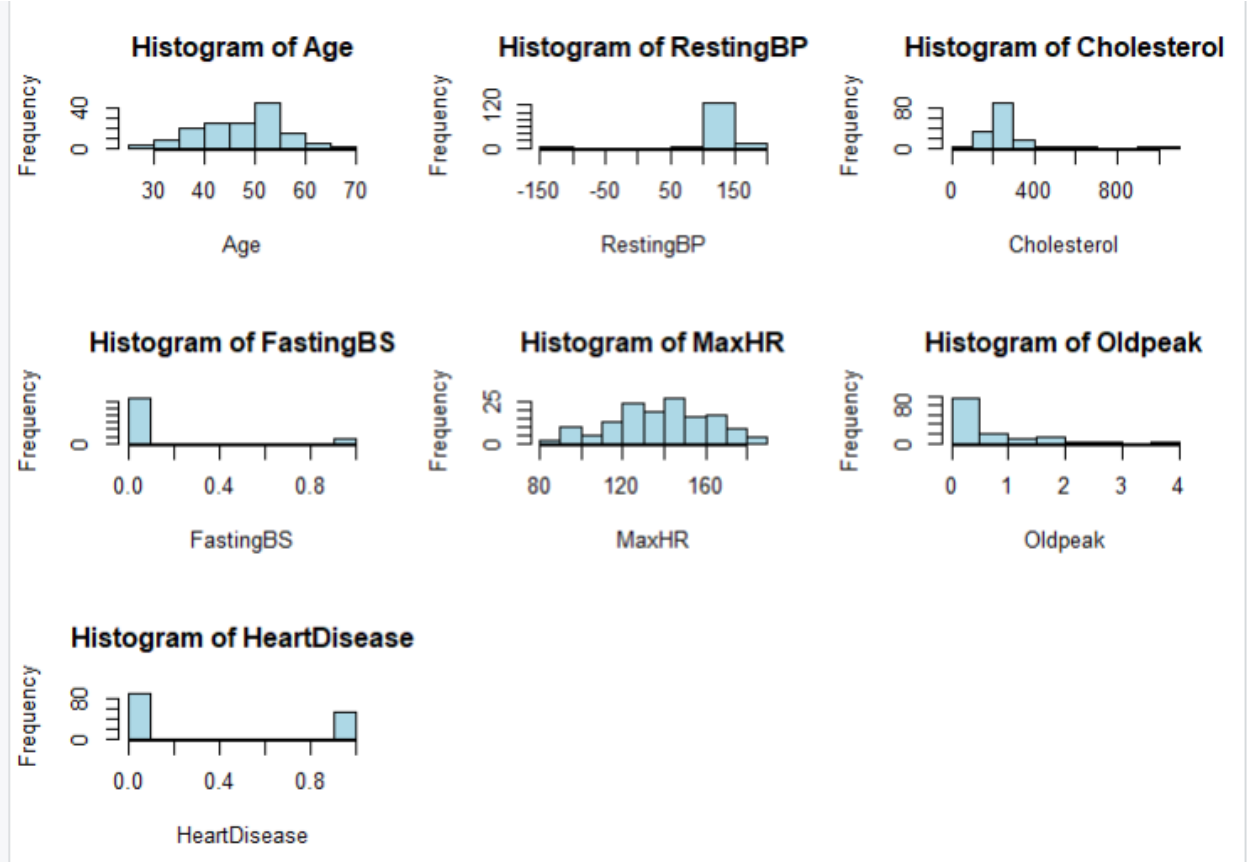
Description: It defines the single column details.

Histogram:

Code:

```
numerical_columns <- sapply(mydata, is.numeric)
par(mfrow = c(3, 3))
for (col in names(mydata)[numerical_columns]) {
hist(mydata[[col]], main = paste("Histogram of", col), xlab = col, col = "lightblue", border = "black")
  }
```

Output:



Description: Histograms for those available numerical column.

Bargraph:

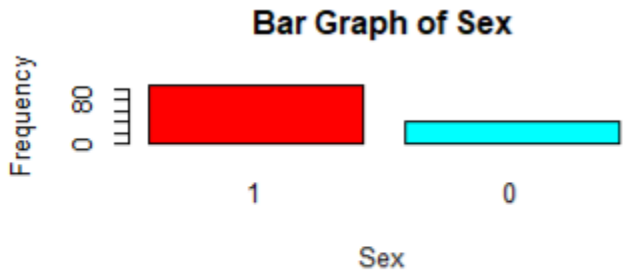
Code:

```r

categorical_columns <- sapply(mydata, is.factor)
par(mfrow = c(3, 3))
for (col in names(mydata)[categorical_columns]) {
category_counts <- table(mydata[[col]])
barplot(category_counts, main = paste("Bar Graph of", col), xlab = col, ylab = "Frequency", col = rainbow(length(category_counts)))
}


```

Output:



Description: It shows the frequency of a column.

Boxplot:

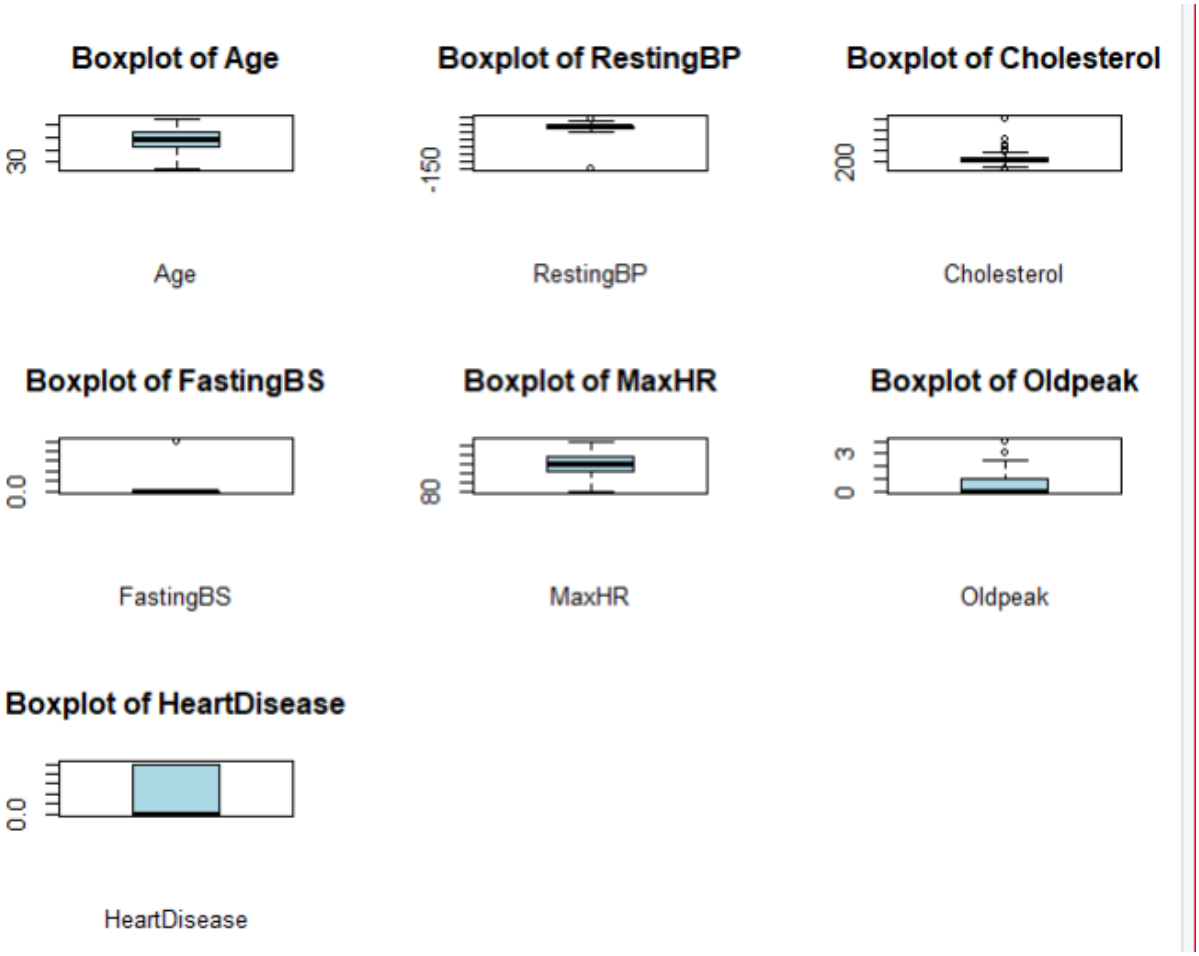
Code:

```r


numerical_columns <- sapply(mydata, is.numeric)
par(mfrow = c(3, 3))
for (col in names(mydata)[numerical_columns]) {
boxplot(mydata[[col]], main = paste("Boxplot of", col), xlab = col, col = "lightblue", border = "black")
}


```

Output:



Description: It shows the representation of the distribution of this dataset.