UNCLEAN-RSCRIPT.R

mohil

2022-02-21

```
r = getOption("repos")
r["CRAN"]="http://cran.us.r-project.org"
options(repos=r)

#importing a dataset
df <- read.csv("C:/Users/mohil/OneDrive/Desktop/Studies/ALY-6000 Introduction to Analytics/Module 6/hea
summary(df)
```

```
##       Age            Sex            ChestPainType        RestingBP
##  Min.   :28.00   Length:918         Length:918         Min.   :  0.0
##  1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
##  Median :54.00   Mode  :character   Mode  :character   Median :130.0
##  Mean   :53.51                                         Mean   :132.4
##  3rd Qu.:60.00                                         3rd Qu.:140.0
##  Max.   :77.00                                         Max.   :200.0
##                                                        NA's   :23
##   Cholesterol      FastingBS       RestingECG            MaxHR
##  Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0
##  1st Qu.:172.8   1st Qu.:0.0000   Class :character   1st Qu.:120.0
##  Median :223.0   Median :0.0000   Mode  :character   Median :138.0
##  Mean   :198.2   Mean   :0.2331                      Mean   :136.8
##  3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0
##  Max.   :603.0   Max.   :1.0000                      Max.   :202.0
##  NA's   :10
##  ExerciseAOgiOa      Oldpeak          ST_Slope          HeartDisease
##  Min.   :0.0000   Min.   :-2.6000   Length:918         Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
##  Median :0.0000   Median : 0.6000   Mode  :character   Median :1.0000
##  Mean   :0.4041   Mean   : 0.8874                      Mean   :0.5534
##  3rd Qu.:1.0000   3rd Qu.: 1.5000                      3rd Qu.:1.0000
##  Max.   :1.0000   Max.   : 6.2000                      Max.   :1.0000
##
```

```
################################finding null values
is.null(df)
```

```
## [1] FALSE
```

```
a <- is.na(df$RestingBP)
sum(a)
```

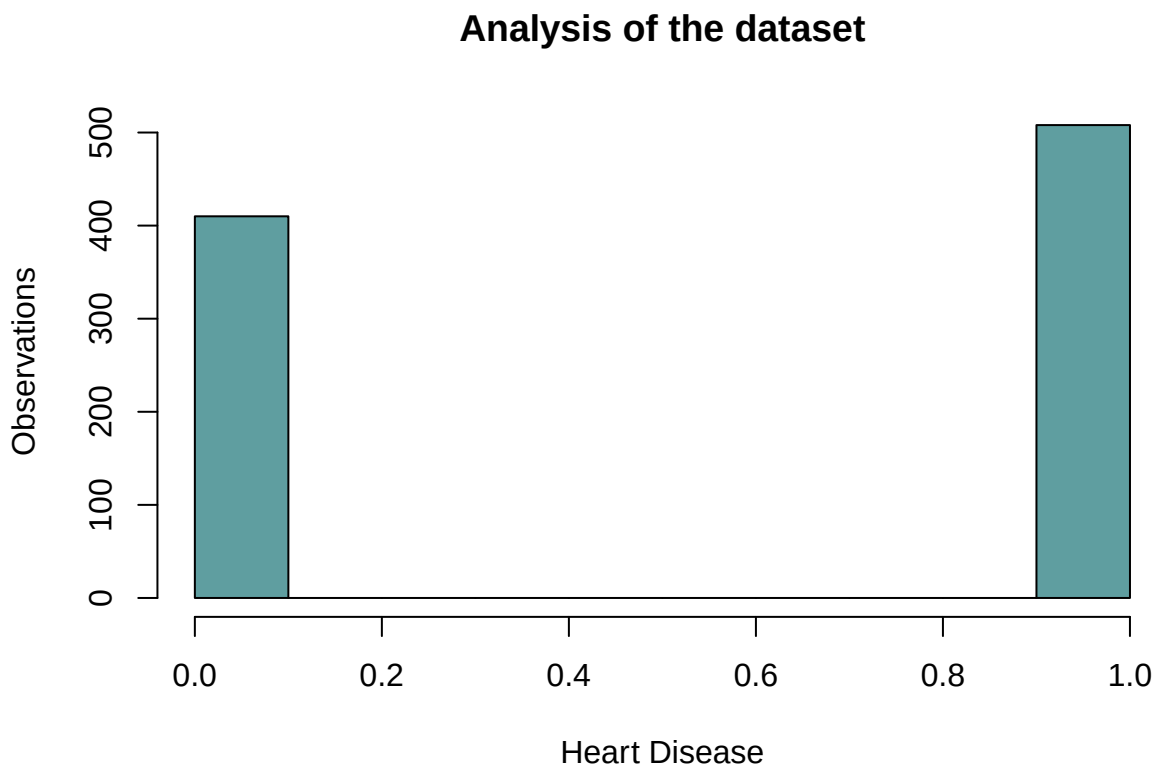
```
## [1] 23
```

```
b <- is.na(df$Cholesterol)
sum(b)
```

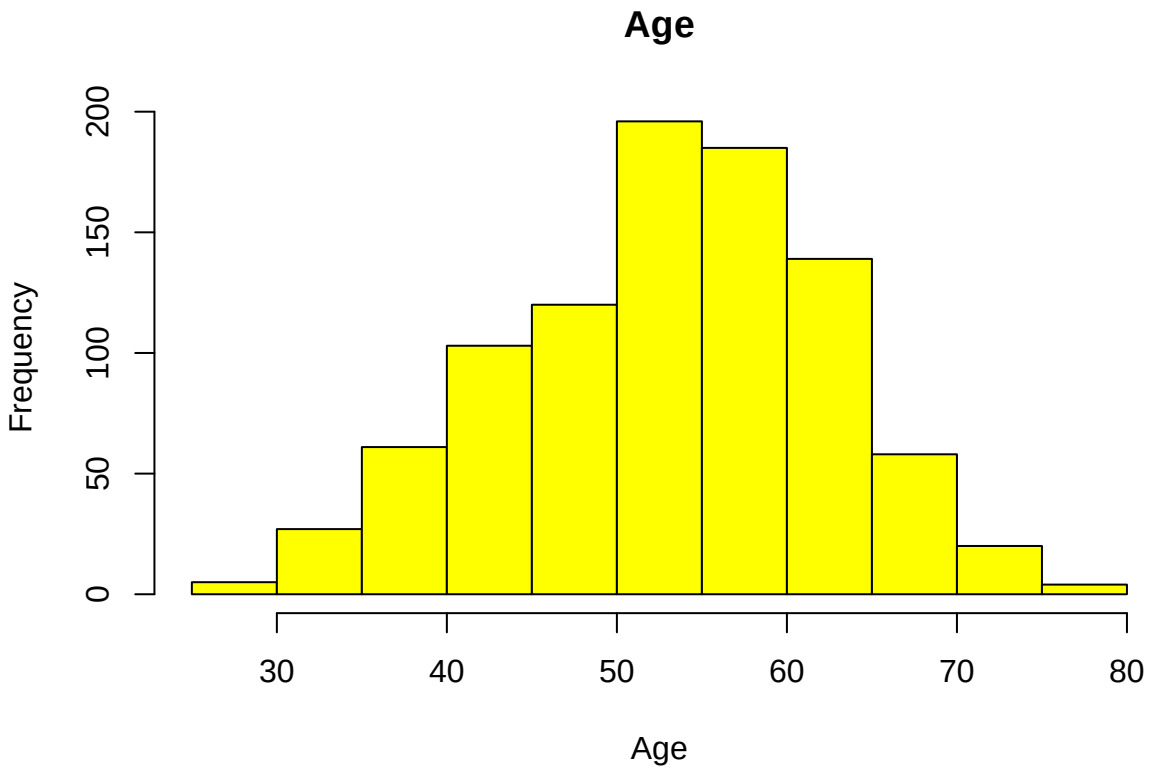
```
## [1] 10
```

```
################################data visualization on unclean data

hist(df$HeartDisease, main = "Analysis of the dataset",
     ylab = "Observations",xlab = "Heart Disease", col = "cadetblue")
```




```
hist(df$Age, main = "Age",
     xlab = "Age", ylab = "Frequency",
     col = "Yellow", freq = TRUE)
```




```
##############################missing value and null value treatment
#cholesterol
plot1 <- plot(df$Age ~ df$Cholesterol,
              main = "Cholesterol vs Age",
              ylab = "Age(yrs)",
              xlab = "Cholesterol(mm/dl)",
              pch = 20, col = 'maroon')
```


```
df$Cholesterol[df$Cholesterol==0]=NA
mean(df$Cholesterol)
```

```
## [1] NA
```

```
mean(df$Cholesterol, na.rm = TRUE)
```

```
## [1] 244.5299
```

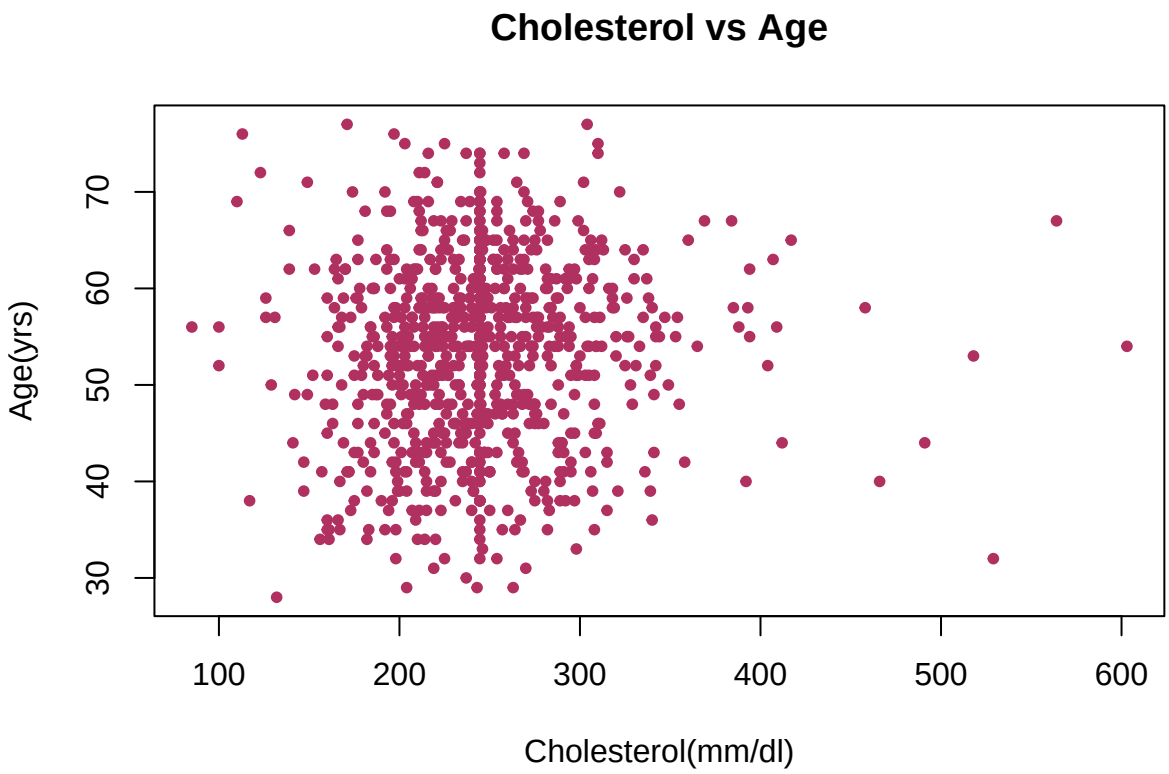
```
which(is.na(df$Cholesterol))
```

```
##   [1]   5  29  73 111 142 190 246 286 294 295 296 297 298 299 300 301 302 303
##  [19] 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321
##  [37] 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339
##  [55] 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357
##  [73] 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375
##  [91] 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393
## [109] 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411
## [127] 412 413 414 415 416 422 424 425 428 429 430 431 435 436 437 438 439 440
## [145] 441 442 443 447 450 451 452 454 456 457 458 459 460 462 464 465 467 468
## [163] 471 472 473 475 476 478 480 481 482 484 485 493 509 515 516 519 536 537
## [181] 551 585
```

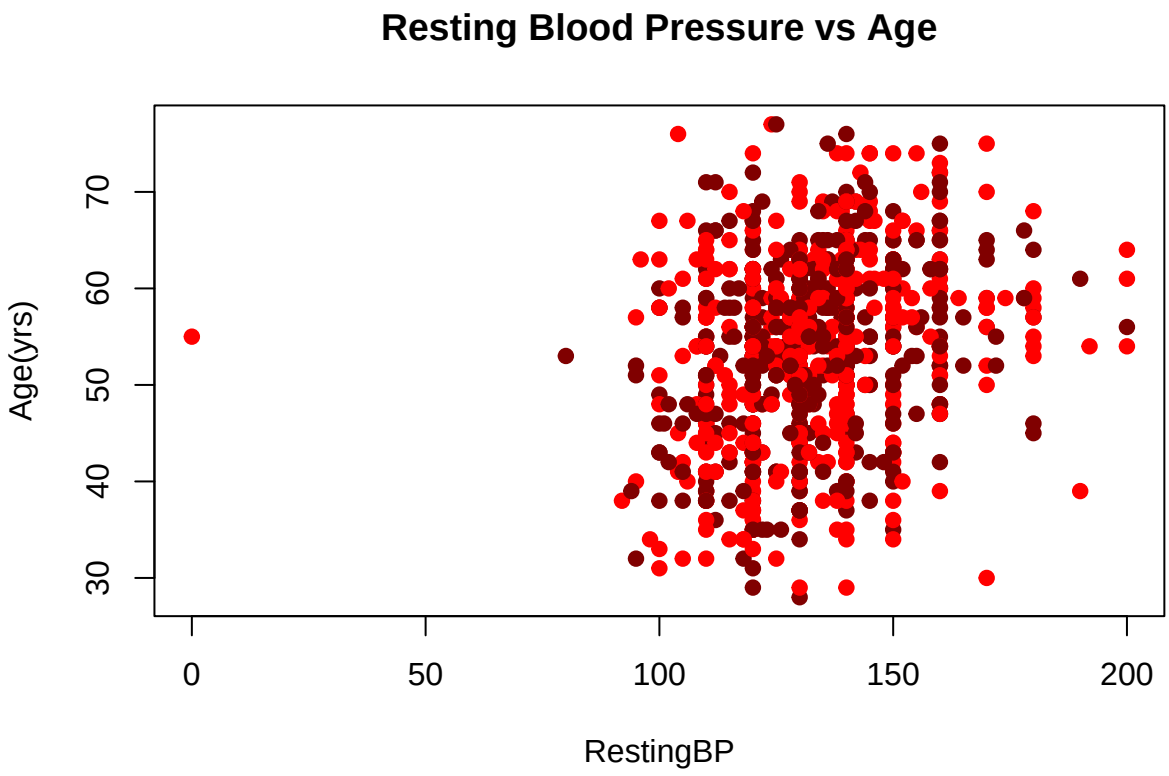
```
df$Cholesterol[is.na(df$Cholesterol)] <- mean(df$Cholesterol, na.rm = TRUE)
summary(df)
```

```
##       Age            Sex            ChestPainType        RestingBP
##  Min.   :28.00   Length:918         Length:918         Min.   :  0.0
##  1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
##  Median :54.00   Mode  :character   Mode  :character   Median :130.0
##  Mean   :53.51                                         Mean   :132.4
##  3rd Qu.:60.00                                         3rd Qu.:140.0
##  Max.   :77.00                                         Max.   :200.0
##                                                        NA's   :23
##   Cholesterol      FastingBS       RestingECG           MaxHR
##  Min.   : 85.0   Min.   :0.0000   Length:918         Min.   : 60.0
##  1st Qu.:215.0   1st Qu.:0.0000   Class :character   1st Qu.:120.0
##  Median :244.5   Median :0.0000   Mode  :character   Median :138.0
##  Mean   :244.5   Mean   :0.2331                      Mean   :136.8
##  3rd Qu.:266.0   3rd Qu.:0.0000                      3rd Qu.:156.0
##  Max.   :603.0   Max.   :1.0000                      Max.   :202.0
##
##  ExerciseAOgiOa      Oldpeak          ST_Slope          HeartDisease
##  Min.   :0.0000   Min.   :-2.6000   Length:918         Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
##  Median :0.0000   Median : 0.6000   Mode  :character   Median :1.0000
##  Mean   :0.4041   Mean   : 0.8874                      Mean   :0.5534
##  3rd Qu.:1.0000   3rd Qu.: 1.5000                      3rd Qu.:1.0000
##  Max.   :1.0000   Max.   : 6.2000                      Max.   :1.0000
##
```

```
plot1 <- plot(df$Age ~ df$Cholesterol,
              main = "Cholesterol vs Age",
              ylab = "Age(yrs)",
              xlab = "Cholesterol(mm/dl)",
              pch = 20, col = 'maroon')
```




```
#restingbp
plot2 <- plot(df$Age ~ df$RestingBP,
              main = "Resting Blood Pressure vs Age",
              xlab = "RestingBP",
              ylab = "Age(yrs)",
              pch = 19,
              col = rgb((1:2)/2,0,0))
```




```
df$RestingBP[df$RestingBP==0]=NA
mean(df$RestingBP)
```

```
## [1] NA
```

```
mean(df$RestingBP, na.rm = TRUE)
```

```
## [1] 132.5738
```

```
which(is.na(df$RestingBP))
```

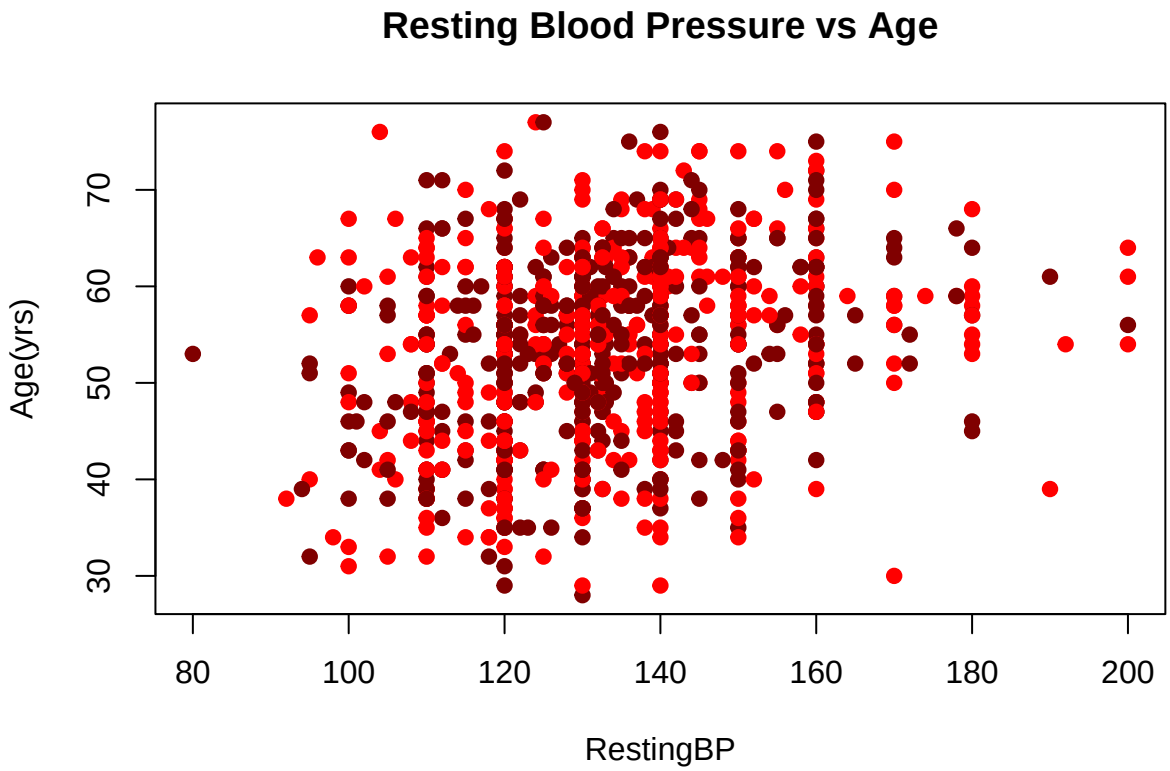
```
##  [1]   2  13  52 121 189 236 303 340 372 373 374 415 450 468 573 673 769 803 820
## [20] 833 867 892 910 917
```

```
df$RestingBP[is.na(df$RestingBP)] <- mean(df$RestingBP, na.rm = TRUE)
summary(df)
```

```
##       Age            Sex            ChestPainType        RestingBP
##  Min.   :28.00   Length:918         Length:918         Min.   : 80.0
##  1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
##  Median :54.00   Mode  :character   Mode  :character   Median :130.0
##  Mean   :53.51                                         Mean   :132.6
##  3rd Qu.:60.00                                         3rd Qu.:140.0
```

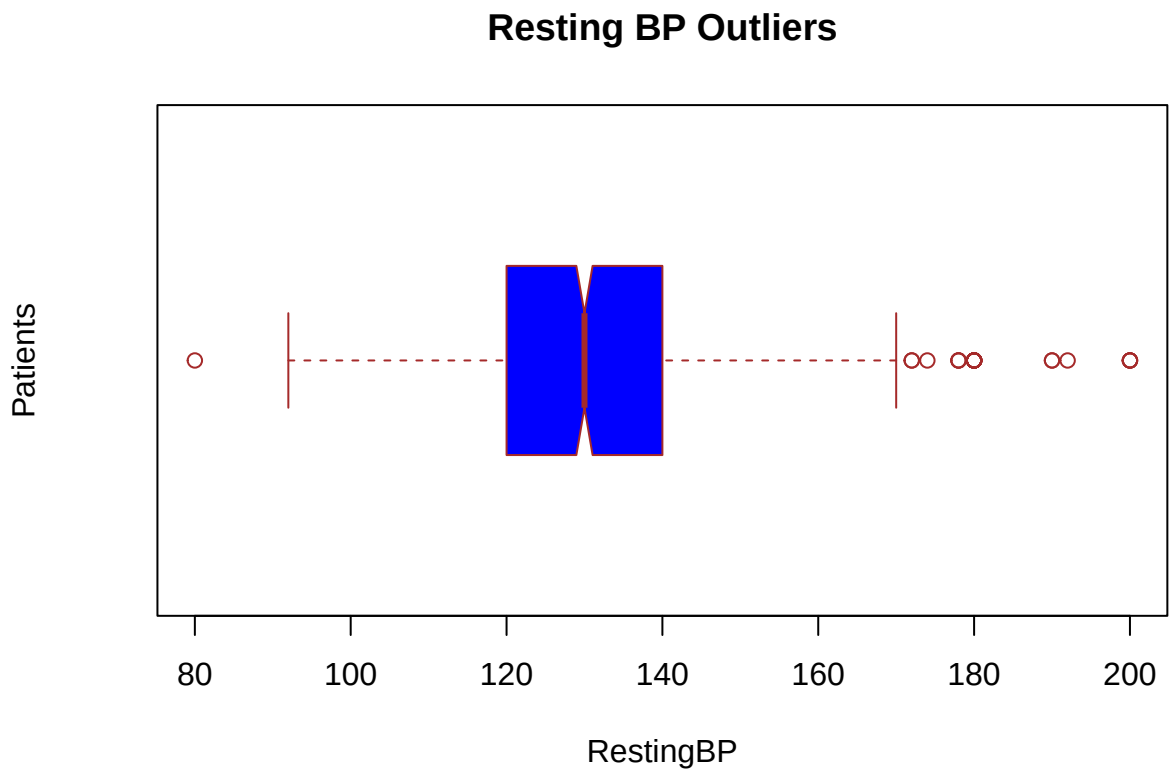
```
##  Max.   :77.00                                          Max.   :200.0
##   Cholesterol      FastingBS       RestingECG            MaxHR
##  Min.   : 85.0   Min.   :0.0000   Length:918         Min.   : 60.0
##  1st Qu.:215.0   1st Qu.:0.0000   Class :character   1st Qu.:120.0
##  Median :244.5   Median :0.0000   Mode  :character   Median :138.0
##  Mean   :244.5   Mean   :0.2331                      Mean   :136.8
##  3rd Qu.:266.0   3rd Qu.:0.0000                      3rd Qu.:156.0
##  Max.   :603.0   Max.   :1.0000                      Max.   :202.0
##  ExerciseAOgiOa      Oldpeak          ST_Slope          HeartDisease
##  Min.   :0.0000   Min.   :-2.6000   Length:918         Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
##  Median :0.0000   Median : 0.6000   Mode  :character   Median :1.0000
##  Mean   :0.4041   Mean   : 0.8874                      Mean   :0.5534
##  3rd Qu.:1.0000   3rd Qu.: 1.5000                      3rd Qu.:1.0000
##  Max.   :1.0000   Max.   : 6.2000                      Max.   :1.0000
```

```r
plot2 <- plot(df$Age ~ df$RestingBP,
              main = "Resting Blood Pressure vs Age",
              xlab = "RestingBP",
              ylab = "Age(yrs)",
              pch = 19,
              col = rgb((1:2)/2,0,0))
```




```r
################################outlier treatment
```

```r
#for restingBP
boxplot(df$RestingBP, main = "Resting BP Outliers", horizontal = TRUE,
        xlab = 'RestingBP', ylab = 'Patients', col = "blue",
        border = "brown", notch = TRUE)
```



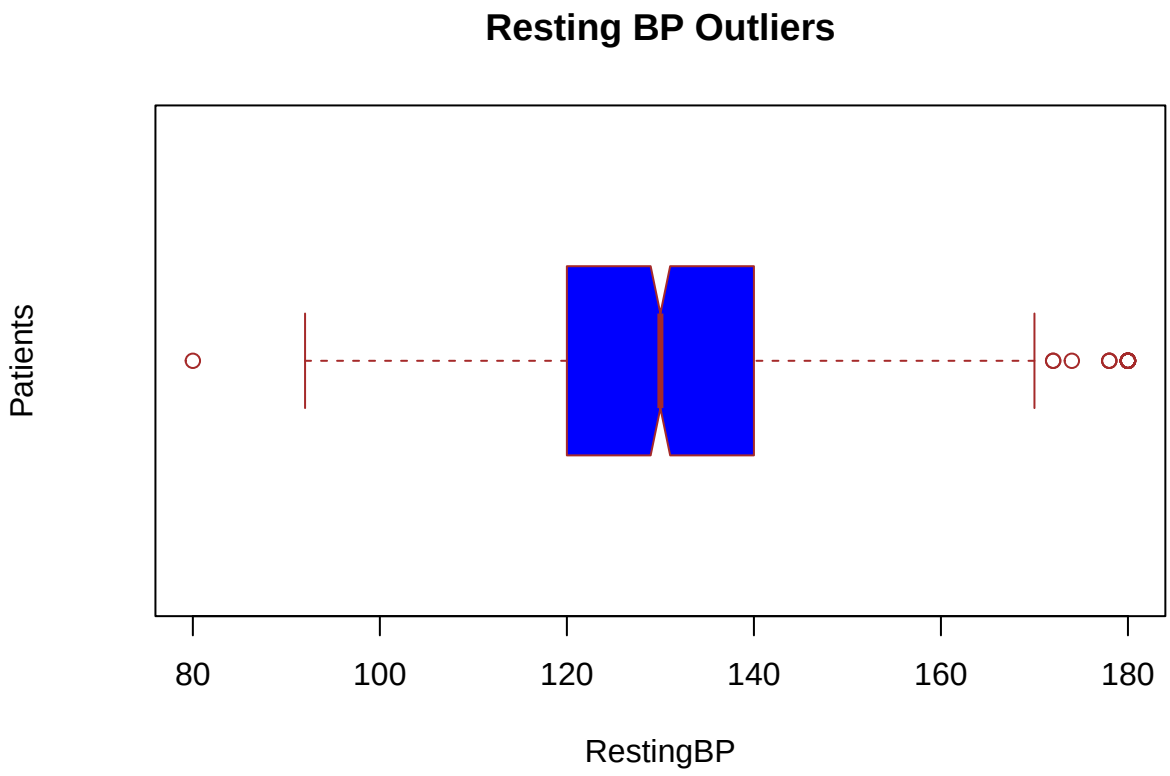

```r
quantile(df$RestingBP,0.99)
```

```
## 99%
## 180
```

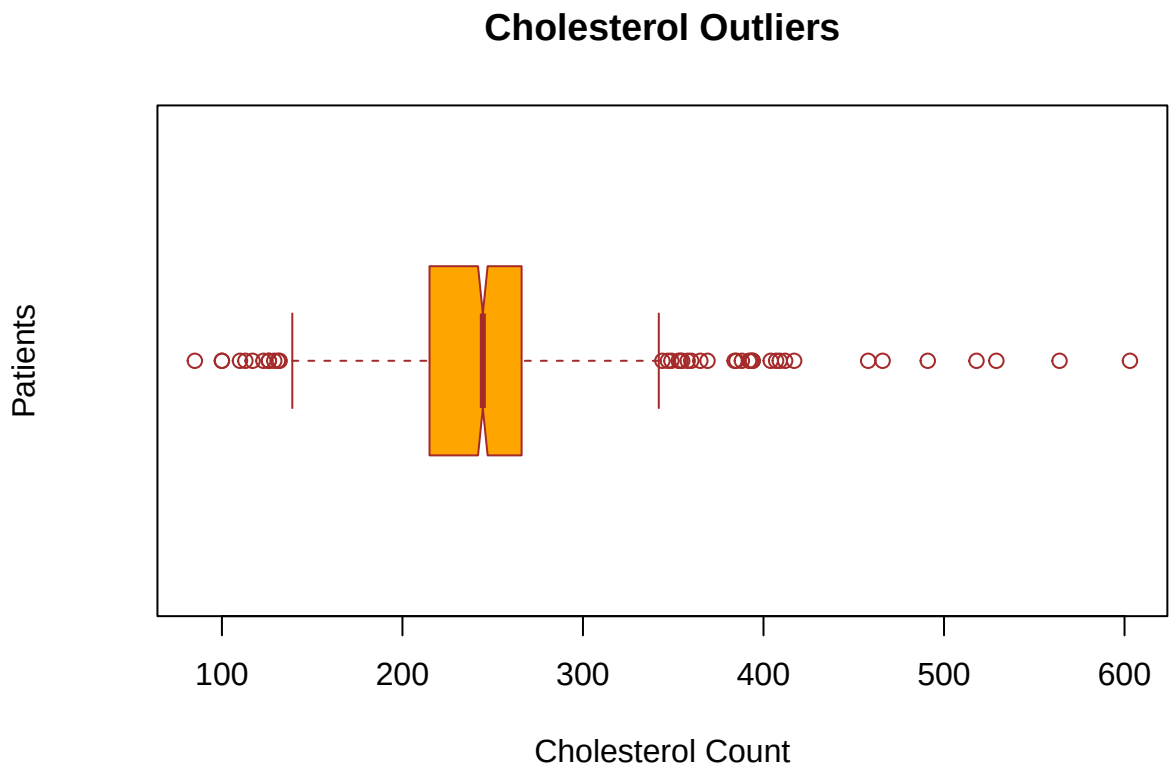
```r
uv= 1*quantile(df$RestingBP,0.99)
df$RestingBP[df$RestingBP > uv] <- uv
summary(df$RestingBP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    80.0   120.0   130.0   132.5   140.0   180.0
```

```r
boxplot(df$RestingBP, main = "Resting BP Outliers", horizontal = TRUE,
        xlab = 'RestingBP', ylab = 'Patients', col = "blue",
        border = "brown", notch = TRUE)
```




```
#for cholesterol
boxplot(df$Cholesterol, main = "Cholesterol Outliers", horizontal = TRUE,
       xlab = 'Cholesterol Count', ylab = 'Patients', col = "orange",
       border = "brown", notch = TRUE)
```




```
quantile(df$Cholesterol,0.99)
```

```
##    99%
## 408.66
```

```
uv= 0.75*quantile(df$Cholesterol,0.99)
df$Cholesterol[df$Cholesterol > uv] <- uv
summary(df$Cholesterol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    85.0   215.0   244.5   240.2   266.0   306.5
```

```
quantile(df$Cholesterol,0.01)
```

```
##     1%
## 129.34
```

```
lv = 1*quantile(df$Cholesterol,0.01)
df$Cholesterol[df$Cholesterol < lv] <- lv
summary(df)
```

```
##       Age            Sex            ChestPainType        RestingBP
##  Min.   :28.00   Length:918         Length:918         Min.   : 80.0
```

```
##  1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
##  Median :54.00   Mode  :character   Mode  :character   Median :130.0
##  Mean   :53.51                                         Mean   :132.5
##  3rd Qu.:60.00                                         3rd Qu.:140.0
##  Max.   :77.00                                         Max.   :180.0
##   Cholesterol      FastingBS       RestingECG            MaxHR
##  Min.   :129.3   Min.   :0.0000   Length:918         Min.   : 60.0
##  1st Qu.:215.0   1st Qu.:0.0000   Class :character   1st Qu.:120.0
##  Median :244.5   Median :0.0000   Mode  :character   Median :138.0
##  Mean   :240.4   Mean   :0.2331                      Mean   :136.8
##  3rd Qu.:266.0   3rd Qu.:0.0000                      3rd Qu.:156.0
##  Max.   :306.5   Max.   :1.0000                      Max.   :202.0
##  ExerciseAOgiOa      Oldpeak          ST_Slope          HeartDisease
##  Min.   :0.0000   Min.   :-2.6000   Length:918         Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
##  Median :0.0000   Median : 0.6000   Mode  :character   Median :1.0000
##  Mean   :0.4041   Mean   : 0.8874                      Mean   :0.5534
##  3rd Qu.:1.0000   3rd Qu.: 1.5000                      3rd Qu.:1.0000
##  Max.   :1.0000   Max.   : 6.2000                      Max.   :1.0000
```

```
boxplot(df$Cholesterol, main = "Cholesterol Outliers", horizontal = TRUE,
        xlab = 'Cholesterol Count', ylab = 'Patients', col = "orange",
        border = "brown", notch = TRUE)
```