pimademo

2022-09-19

```
#Alex Khater
#Assignment 3
#DATS 6101
library(MASS)
```

```
pima <- data.frame(read.csv("pima.csv"))
head(pima)
```

```
##   npreg glu bp skin  bmi   ped age type ageGroup
## 1     6 148 72   35 33.6 0.627  50  Yes    41-50
## 2     1  85 66   29 26.6 0.351  31   No    31-40
## 3     1  89 66   23 28.1 0.167  21   No    21-30
## 4     3  78 50   32 31.0 0.248  26  Yes    21-30
## 5     2 197 70   45 30.5 0.158  53  Yes    51-60
## 6     5 166 72   19 25.8 0.587  51  Yes    51-60
```

```
# Exercise 1
summary(pima)
```

```
##      npreg             glu              bp              skin
##  Min.   : 0.000   Min.   : 65.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 96.0   1st Qu.: 64.00   1st Qu.:22.00
##  Median : 2.000   Median :112.0   Median : 72.00   Median :29.00
##  Mean   : 3.485   Mean   :119.3   Mean   : 71.65   Mean   :29.16
##  3rd Qu.: 5.000   3rd Qu.:136.2   3rd Qu.: 80.00   3rd Qu.:36.00
##  Max.   :17.000   Max.   :197.0   Max.   :110.00   Max.   :63.00
##       bmi             ped              age            type
##  Min.   :19.40   Min.   :0.0850   Min.   :21.00   Length:332
##  1st Qu.:28.18   1st Qu.:0.2660   1st Qu.:23.00   Class :character
##  Median :32.90   Median :0.4400   Median :27.00   Mode  :character
##  Mean   :33.24   Mean   :0.5284   Mean   :31.32
##  3rd Qu.:37.20   3rd Qu.:0.6793   3rd Qu.:37.00
##  Max.   :67.10   Max.   :2.4200   Max.   :81.00
##    ageGroup
##  Length:332
##  Class :character
##  Mode  :character
##
##
##
```

```
head(pima)
```

```
##   npreg glu bp skin  bmi   ped age type ageGroup
## 1     6 148 72   35 33.6 0.627  50  Yes    41-50
## 2     1  85 66   29 26.6 0.351  31   No    31-40
## 3     1  89 66   23 28.1 0.167  21   No    21-30
## 4     3  78 50   32 31.0 0.248  26  Yes    21-30
## 5     2 197 70   45 30.5 0.158  53  Yes    51-60
## 6     5 166 72   19 25.8 0.587  51  Yes    51-60
```

```
#Exercise 2
str(pima)
```

```
## ’data.frame’:    332 obs. of  9 variables:
##  $ npreg   : int  6 1 1 3 2 5 0 1 3 9 ...
##  $ glu     : int  148 85 89 78 197 166 118 103 126 119 ...
##  $ bp      : int  72 66 66 50 70 72 84 30 88 80 ...
##  $ skin    : int  35 29 23 32 45 19 47 38 41 35 ...
##  $ bmi     : num  33.6 26.6 28.1 31 30.5 25.8 45.8 43.3 39.3 29 ...
##  $ ped     : num  0.627 0.351 0.167 0.248 0.158 0.587 0.551 0.183 0.704 0.263 ...
##  $ age     : int  50 31 21 26 53 51 31 33 27 29 ...
##  $ type    : chr  "Yes" "No" "No" "Yes" ...
##  $ ageGroup: chr  "41-50" "31-40" "21-30" "21-30" ...
```

```
#Exercise 3
names(pima)
```

```
## [1] "npreg"    "glu"      "bp"       "skin"     "bmi"      "ped"      "age"
## [8] "type"     "ageGroup"
```

```
#Exercise 4
#bmi stats
mean(pima$bmi)
```

```
## [1] 33.23976
```

```
median(pima$bmi)
```

```
## [1] 32.9
```

```
max(pima$bmi)
```

```
## [1] 67.1
```

```
min(pima$bmi)
```

```
## [1] 19.4
```

```
range(pima$bmi)
```

```
## [1] 19.4 67.1
```

```
nrow(pima)
```

```
## [1] 332
```

```
#age stats
mean(pima$age)
```

```
## [1] 31.31627
```

```
median(pima$age)
```

```
## [1] 27
```

```
max(pima$age)
```

```
## [1] 81
```

```
min(pima$age)
```

```
## [1] 21
```

```
range(pima$age)
```

```
## [1] 21 81
```

```
nrow(pima)
```

```
## [1] 332
```

```
#Exercise 5
#This data set entirely consists of women so the number of rows (subjects) will tell us
nrow(pima)
```

```
## [1] 332
```

```
#Exercise 6
pima[1:5, 1:4]
```

```
##   npreg glu bp skin
## 1     6 148 72   35
## 2     1  85 66   29
## 3     1  89 66   23
## 4     3  78 50   32
## 5     2 197 70   45
```

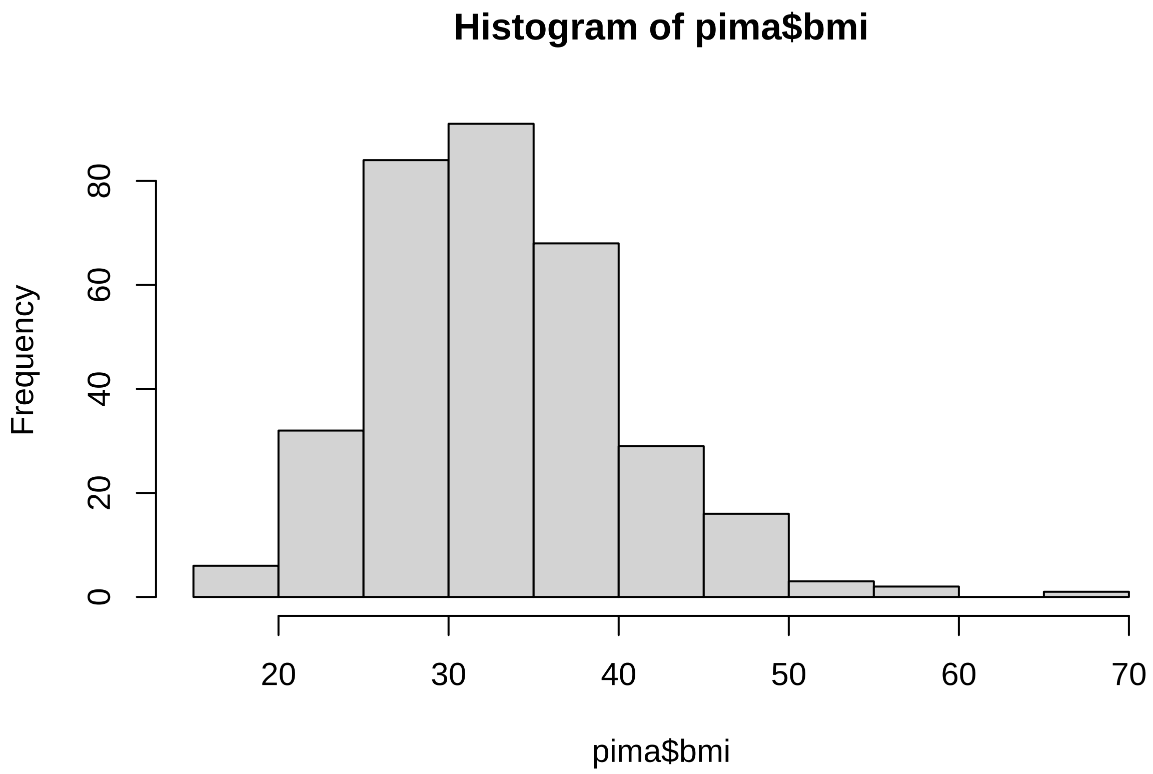
```r
#Exercise 7
which(pima$bmi>50)
```

```
## [1]  55  57  79 107 198 292
```

```r
#Exercise 8
#The "Yes" column corresponds to the number of subjects with Diabetes according to WHO guidelines. It i
table(pima$type)
```

```
##
##  No Yes
## 223 109
```

```r
#Exercise 9
hist(pima$bmi)
```

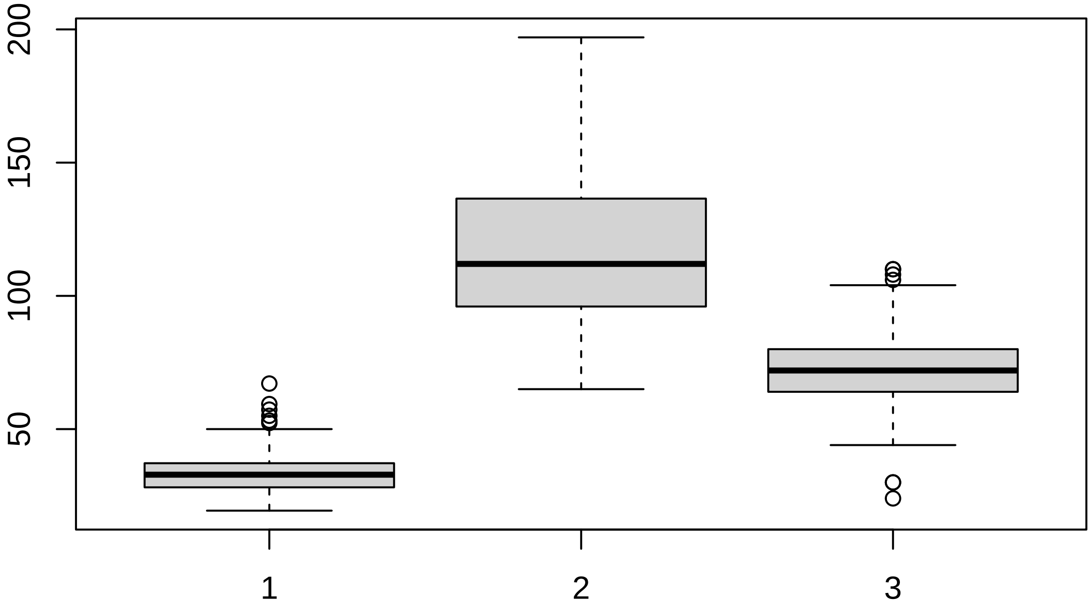


```r
#Exercise 10
mean(pima$bmi)- median(pima$bmi)
```

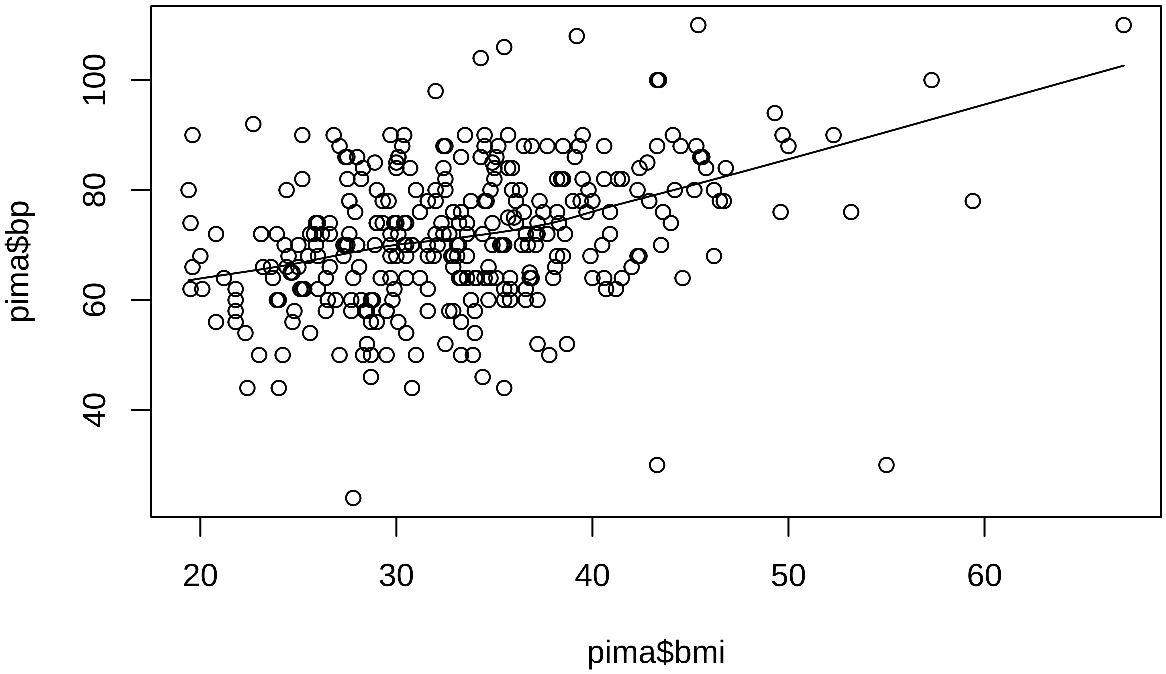
```
## [1] 0.339759
```

```
#Extra Stuff

boxplot(pima$bmi, pima$glu, pima$bp)
```



```
scatter.smooth(x=pima$bmi, y=pima$bp)
```



```
cor(pima$bmi, pima$bp)
```

```
## [1] 0.3381926
```

```
linearMod <- lm(bmi ~ bp, data=pima)  # build linear regression model on full data
print(linearMod)
```

```
##
## Call:
## lm(formula = bmi ~ bp, data = pima)
##
## Coefficients:
## (Intercept)           bp
##     19.4512       0.1924
```

```
summary(linearMod)
```

```
##
## Call:
## lm(formula = bmi ~ bp, data = pima)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1702  -4.9814  -0.6262   4.1627  29.7758
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.45116    2.14548   9.066  < 2e-16 ***
## bp           0.19243    0.02948   6.528 2.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.864 on 330 degrees of freedom
## Multiple R-squared:  0.1144, Adjusted R-squared:  0.1117
## F-statistic: 42.62 on 1 and 330 DF,  p-value: 2.511e-10
```