STATS 500 - Homework 1 Solution set

Problem 1:

(a): binary and qualitative

(b): continuous and quantitative

(c): ordinal and qualitative

(d): continuous and quantitative

(e): discrete and quantitative

Problem 2:

First we load the data set:

```
library(faraway)
data(pima)
```

(a). The dimension of the data set is (n,p) = (#observations, #samples) =

```
dim(pima)
```

```
## [1] 768   9
```

(b). Before finding the numerical summaries, we categorize the variable 'test' (so R treats it appropriately as a categorical variable), and relabel its two levels (for better readability):

```
pima$test = factor(pima$test)
levels(pima$test) = c("Negative", "Positive")
```

Now we are ready to get the summaries:

```
summary(pima)
```

```
##     pregnant         glucose        diastolic         triceps     
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00  
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00  
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00  
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54  
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00  
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00  
##     insulin           bmi           diabetes           age       
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00  
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00  
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00  
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24  
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00  
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00  
##        test    
##  Negative:500  
##  Positive:268  
## 
```

```
##
##
##
```

(c).

```
pima$glucose[pima$glucose == 0] = NA
pima$diastolic[pima$diastolic == 0] = NA
pima$triceps[pima$triceps == 0] = NA
pima$insulin[pima$insulin == 0] = NA
pima$bmi[pima$bmi == 0] = NA
summary(pima)
```

```
##     pregnant         glucose        diastolic         triceps
##  Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :29.00
##  Mean   : 3.845   Mean   :121.7   Mean   : 72.41   Mean   :29.15
##  3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##                   NA's   :5       NA's   :35       NA's   :227
##     insulin            bmi           diabetes           age
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0780   Min.   :21.00
##  1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437   1st Qu.:24.00
##  Median :125.00   Median :32.30   Median :0.3725   Median :29.00
##  Mean   :155.55   Mean   :32.46   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##  NA's   :374      NA's   :11
##        test
##  Negative:500
##  Positive:268
##
##
##
##
##
```

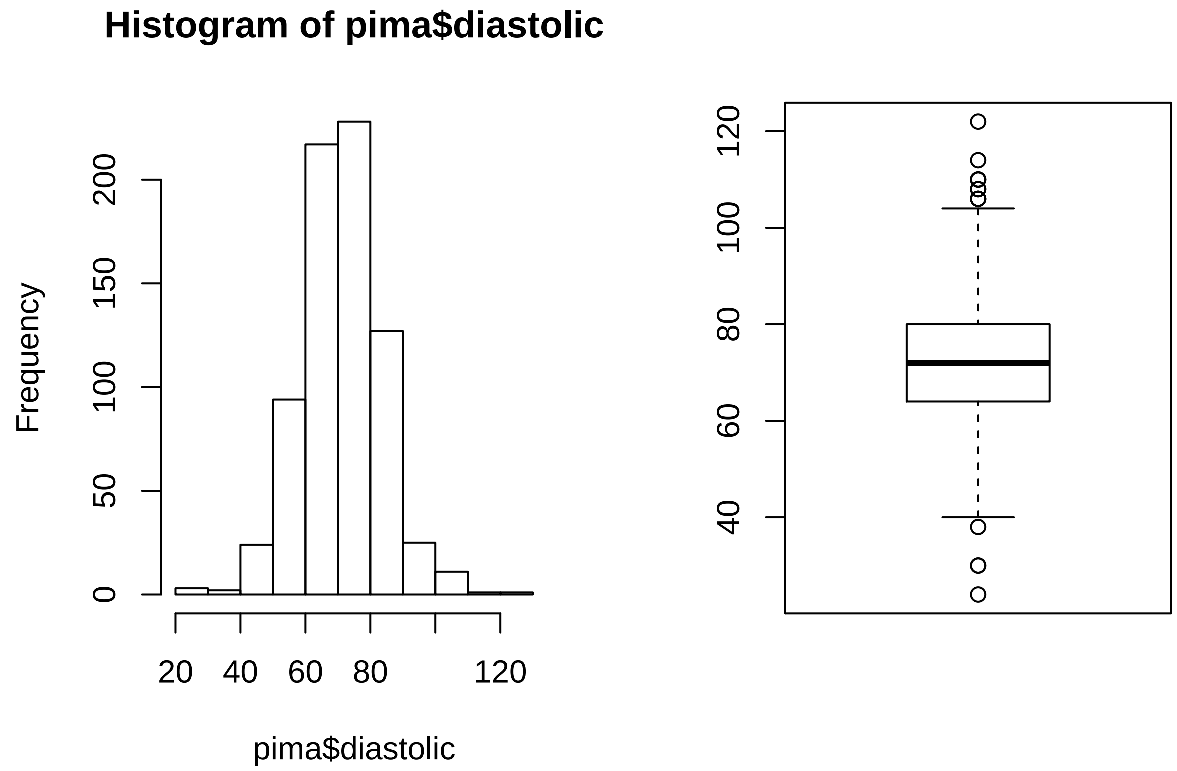
(d). The observation (row) with the highest diastolic blood presure is

```
which.max(pima$diastolic)
```

```
## [1] 107
```

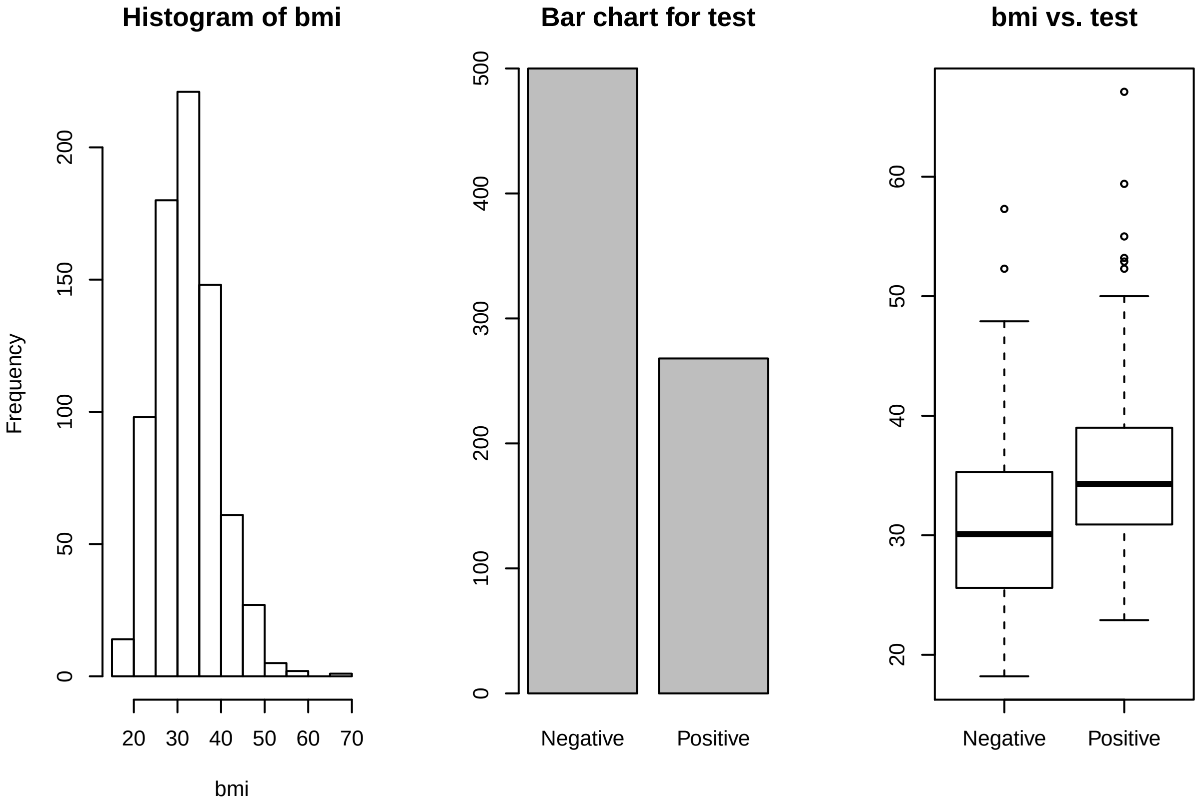
(e). Graphical summaries for ‘diastolic’:

```
par(mfrow = c(1,2))
hist(pima$diastolic)
boxplot(pima$diastolic, rm.na = TRUE)
```



(f).

```r
par(mfrow = c(1,3))
hist(pima$bmi, main = "Histogram of bmi", xlab = "bmi")
plot(pima$test, main = "Bar chart for test")
plot(pima$test, pima$bmi, main = "bmi vs. test")
```


Histogram of bmi
Frequency
0
50
100
150
200
20
30
40
50
60
70
bmi
Bar chart for test
0
100
200
300
400
500
Negative
Positive
bmi vs. test
20
30
40
50
60
Negative
Positive