Assignment_2

Chi-Yu, Chen

2018/7/14

1. Write a program to find the outliers using Box and Whisker's criterion discussed in the class.

```
outliers <- function(numbers) {
    q1 <- quantile(numbers, 1/4)
    q3 <- quantile(numbers, 3/4)
    iqr <- IQR(numbers)
    return(numbers[numbers <= (q1 - 1.5 * iqr) | numbers >= (q3 + 1.5 * iqr)])
}
x <- c(1,100,120,140,160,180,200,1000)
outliers(x)
```

```
## [1]    1 1000
```

2. Make two vectors:

```
X <- c(1,2,5,10,12)
Y <- c(2,5,1,0,12)
```

a. Find the values that are contained in both X and Y.

```
q2.a <- X[X%in%Y]
q2.a
```

```
## [1]  1  2  5 12
```

b. Find values that are in x but not in y and (vice versa).

```
q2.b.xny <- X[!(X%in%Y)]
q2.b.xny
```

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

```
q2.b.ynx <- Y[!(Y%in%X)]
q2.b.ynx
```

```
## [1] 0
```

c. Construct a vector that contains all values contained in either X or Y, and compare this vector to c(X,Y)

```
q2.c <- c(q2.a, q2.b.xny, q2.b.ynx)
q2.c
```

```
## [1]  1  2  5 12 10  0
```

```
q2.c.cmp <- c(X,Y)
q2.c.cmp
```

```
##  [1]  1  2  5 10 12  2  5  1  0 12
```

Load USArrests data set.

3. Which states has most and least assault, murder, and rape arrests.

```
data(USArrests)
q3.assault.most <- USArrests[USArrests$Assault==max(USArrests$Assault),]
row.names(q3.assault.most)
```

```
## [1] "North Carolina"
```

```
q3.assault.least <- USArrests[USArrests$Assault==min(USArrests$Assault),]
row.names(q3.assault.least)
```

```
## [1] "North Dakota"
```

```
q3.murder.most <- USArrests[USArrests$Murder==max(USArrests$Murder),]
row.names(q3.murder.most)
```

```
## [1] "Georgia"
```

```
q3.murder.least <- USArrests[USArrests$Murder==min(USArrests$Murder),]
row.names(q3.murder.least)
```

```
## [1] "North Dakota"
```

```
q3.rape.most <- USArrests[USArrests$Rape==max(USArrests$Rape),]
row.names(q3.rape.most)
```

```
## [1] "Nevada"
```

```
q3.rape.least <- USArrests[USArrests$Rape==min(USArrests$Rape),]
row.names(q3.rape.least)
```

```
## [1] "North Dakota"
```

4. Which states are in the bottom 25% of murder and in the top 25% of the murder.

```
q4.muder.q1 <- quantile(USArrests$Murder, 1/4)
q4.murder.lessthanq1 <- row.names(subset(USArrests, Murder < q4.muder.q1))
q4.murder.lessthanq1
```

```
##  [1] "Connecticut"   "Idaho"          "Iowa"           "Maine"
##  [5] "Minnesota"     "New Hampshire" "North Dakota"   "Rhode Island"
##  [9] "South Dakota"  "Utah"           "Vermont"        "Washington"
## [13] "Wisconsin"
```

```
q4.muder.q3 <- quantile(USArrests$Murder, 3/4)
q4.murder.higherthanq3 <- row.names(subset(USArrests, Murder > q4.muder.q3))
q4.murder.higherthanq3
```

```
##  [1] "Alabama"        "Florida"        "Georgia"        "Louisiana"
##  [5] "Maryland"       "Michigan"       "Mississippi"    "Nevada"
##  [9] "New Mexico"     "North Carolina" "South Carolina" "Tennessee"
## [13] "Texas"
```

5. The following function calculates the mean and standard deviation of a numeric vector

```
fn1 <- function(x){
    mean1 <- mean(x)
    sd1 <- sd(x)
    return(mean = mean1, sd = sd1)
}
```

Modify the function so that:

a. the default is to use rnorm to generate 30 random normal numbers, and return the standard deviation,

b. if there are missing values, the mean and standard deviation are calculated for the remaining values.

```
fn1 <- function(x=rnorm(30)) {
    mean1 <- mean(x, na.rm=TRUE)
    sd1 <- sd(x, na.rm=TRUE)
    return(c(mean=mean1, sd=sd1))
}
x <- c(1,7,2,3,NA,9,5,4,NA)
fn1(x)
```

```
##     mean       sd
## 4.428571 2.819997
```