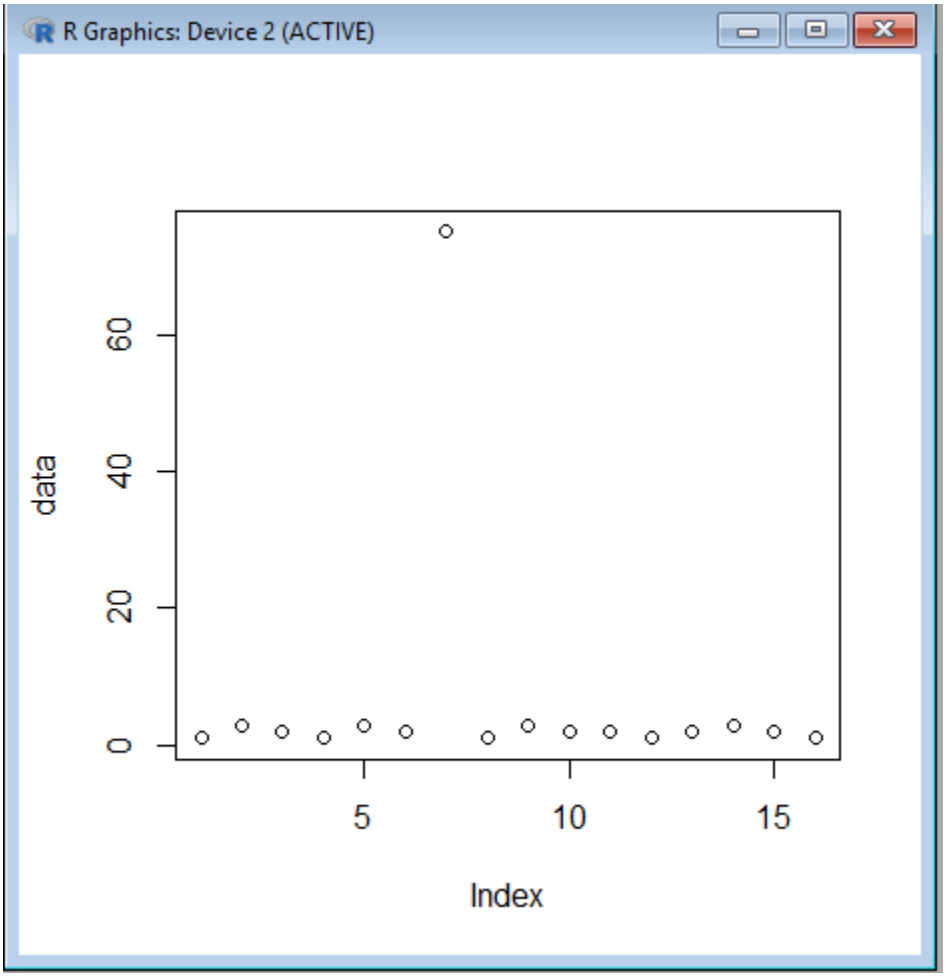
Assignment 1 - due at 02:00 AM ET, TUESDAY, OCTOBER 30, 2018

1. Which of the following points from each of the following sets of points below is an outlier? Why?(3 marks) (This is exercise 1 from chapter1 of the textbook)

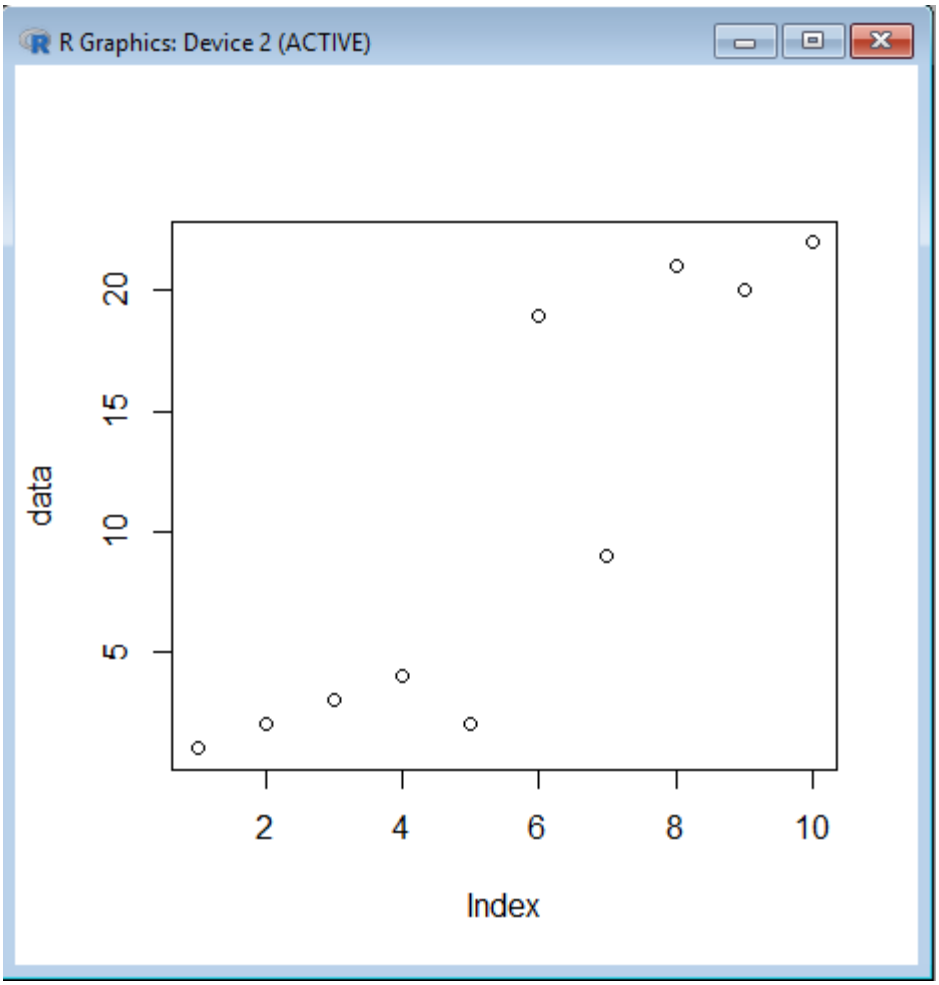
{ 1, 3, 2, 1, 3, 2, 75, 1, 3, 2, 2, 1, 2, 3, 2, 1 }

75 is an outlier. Because the median of the set is 2, 75 is the only one far away from 2.



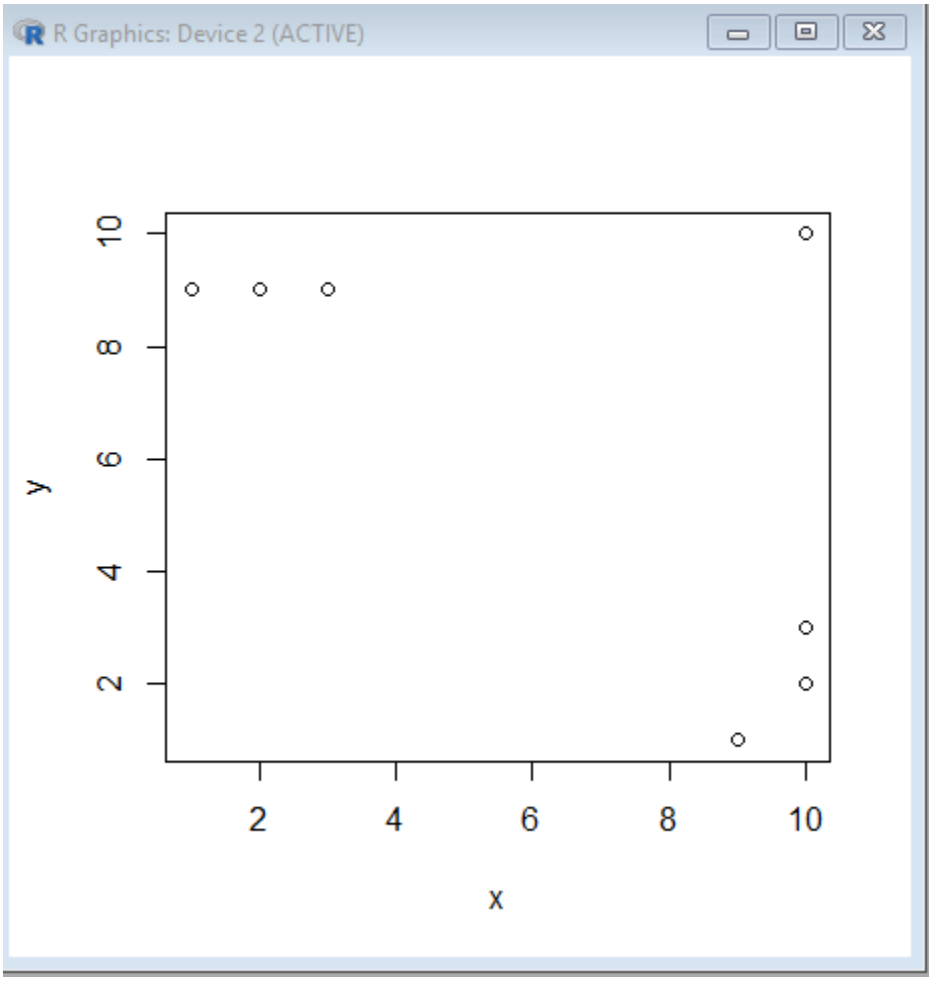
{ 1, 2, 3, 4, 2, 19, 9, 21, 20, 22 }

9 is an outlier. Because there are 2 groups of data {1, 2, 3, 4, 2} and {19, 21, 20, 22 }, but 9 is far away from the 2 clusters.



{ (1, 9), (2, 9), (3, 9), (10, 10), (10, 3), (9, 1), (10, 2) }

(10, 10) is an outlier. Because there are 2 groups of 2D data{(1, 9), (2, 9), (3, 9)} and {(10, 3), (9, 1), (10, 2) }, but (10, 10) is far away from the 2 clusters.



2. Apply the Z-value test to each of the cases above. For the 2-dimensional case, apply the Z-value test to the individual dimensions. Do you discover the correct outliers? (4 marks) (This is exercise 4 from chapter1 of the textbook)

> data<-c(1, 3, 2, 1, 3, 2, 75, 1, 3, 2, 2, 1, 2, 3, 2, 1)

> (data-mean(data))/sd(data)

[1] -0.3008265 -0.1914351 -0.2461308 -0.3008265 -0.1914351 -0.2461308 3.7466579 -0.3008265

[9] -0.1914351 -0.2461308 -0.2461308 -0.3008265 -0.2461308 -0.1914351 -0.2461308 -0.3008265

> scale(data, center=TRUE, scale=TRUE)

[,1]

[1,] -0.3008265

[2,] -0.1914351

```
 [3,] -0.2461308
 [4,] -0.3008265
 [5,] -0.1914351
 [6,] -0.2461308
 [7,]  3.7466579
 [8,] -0.3008265
 [9,] -0.1914351
[10,] -0.2461308
[11,] -0.2461308
[12,] -0.3008265
[13,] -0.2461308
[14,] -0.1914351
[15,] -0.2461308
[16,] -0.3008265
attr(,"scaled:center")
[1] 6.5
attr(,"scaled:scale")
[1] 18.28296
# obviouly, 75 is the outlier with z-value 3.7466579
> data<-c(1, 2, 3, 4, 2, 19, 9, 21, 20, 22)
> (data-mean(data))/sd(data)
 [1] -1.0255551 -0.9152804 -0.8050056 -0.6947309 -0.9152804  0.9593903 -0.1433572  1.1799397
 [9]  1.0696650  1.2902145
> scale(data, center=TRUE, scale=TRUE)
           [,1]
 [1,] -1.0255551
 [2,] -0.9152804
 [3,] -0.8050056
```

```
 [4,] -0.6947309
 [5,] -0.9152804
 [6,]  0.9593903
 [7,] -0.1433572
 [8,]  1.1799397
 [9,]  1.0696650
[10,]  1.2902145
attr(,"scaled:center")
[1] 10.3
attr(,"scaled:scale")
[1] 9.06826
# Here 9 is the outlier with z-value -0.1433572  that is away from the two groups.
>
> x<-c(1, 2, 3, 10, 10, 9, 10)
> (x-mean(x))/sd(x)
[1] -1.2932853 -1.0550485 -0.8168118  0.8508456  0.8508456  0.6126088  0.8508456
> scale(x, center=TRUE, scale=TRUE)
          [,1]
[1,] -1.2932853
[2,] -1.0550485
[3,] -0.8168118
[4,]  0.8508456
[5,]  0.8508456
[6,]  0.6126088
[7,]  0.8508456
attr(,"scaled:center")
```

[1] 6.428571

attr(,"scaled:scale")

[1] 4.197505

> y<-c(9, 9, 9, 10, 3, 1, 2)

> (y-mean(y))/sd(y)

[1] 0.7262730 0.7262730 0.7262730 0.9804686 -0.7989003 -1.3072915 -1.0530959

> scale(y, center=TRUE, scale=TRUE)

          [,1]

[1,] 0.7262730

[2,] 0.7262730

[3,] 0.7262730

[4,] 0.9804686

[5,] -0.7989003

[6,] -1.3072915

[7,] -1.0530959

attr(,"scaled:center")

[1] 6.142857

attr(,"scaled:scale")

[1] 3.933979

>

it is hard to get the 2D point outlier if we use z-test on individual dimensions.

3. Give any one distinction between noise and anomalies in data (1 mark)

Anomalies need to be unusual in aan interesting way. The sematic distinction between noise and anomalies is based on analyst interest**.**

4. Suppose you have a classifier with Precision=25 and Recall=50. If there are 1000 positive cases and 10,000 negative cases in the sample, how many true and false positives will this classifier return? (Hint: clearly understand the formulae on page 33 of the textbook) (2 marks)

recall = 0.5 = TP/positive case, So true positive = 0.5*1000 = 500.

precision = 0.25 = TP/predicted positive, So total predicted positive = 500/0.25 = 2000, false positive = 2000-500 = 1500.

Download the Arrhythmia data set from the UCI Machine Learning Repository. Implement (in R) a 20-nearest neighbor classifier which classifies the majority class as the primary label. Use a 3:1 ratio of costs between the normal class, and any other minority class. Determine the overall accuracy and the cost-weighted accuracy. Note that the outcome variable is the last one in the dataset. (This is the first part of exercise 1 from chapter 7 of the textbook)

```
> data<-read.csv("arrhythmia.data", header=FALSE, na.strings="?")

>

> for(col in 1:ncol(data)){

+ numMissing = nrow(subset(data,is.na(data[,col])))

+ if(numMissing > 0){

+ print(paste(col,":", numMissing)}}

Error: unexpected '}' in:

"if(numMissing > 0){

print(paste(col,":", numMissing)}"

> for(col in 1:ncol(data)){

+ numMissing = nrow(subset(data,is.na(data[,col])))

+ if(numMissing > 0){

+ print(paste(col,":", numMissing))}}

[1] "11 : 8"

[1] "12 : 22"

[1] "13 : 1"

[1] "14 : 376"

[1] "15 : 1"

> data<-data[,-14]
```

```
> data<-na.omit(data)
> dim(data)
[1] 420 279
> anyNA(data)
[1] FALSE
> set.seed(0)
> intrain<-createDataPartition(y=data[,ncol(data)], p=0.7, list=FALSE)
> training<-data[intrain,]
> testing<-data[-intrain,]
> training[,ncol(training)] = factor(training[,ncol(training)] )
> trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
> knn<-train(V280 ~., data = training, method = "knn",
+  trControl=trctrl,
+  preProcess = c("center", "scale"),
+  tuneLength = 10)
> test_pred <- predict(knn, newdata = testing)
```