aykırı gözlem

ilke

2022-08-06

1. aykırı gözlem mi?
2. iş bilgisini ve bazı teknikleri kullanarak aykırı gözlemleri belirle.

> tek değişken ise box plot kullan
>
> iki yya da daha fazla değişken ise kümeleme ya da ikişerli saçım grafiği ve kontur grafikleri
>
> aykırı gözlem mi yoksa yeni trend ve alışkanlığın habercısı mı?????

3.aykrıılar bulunduktan sonra

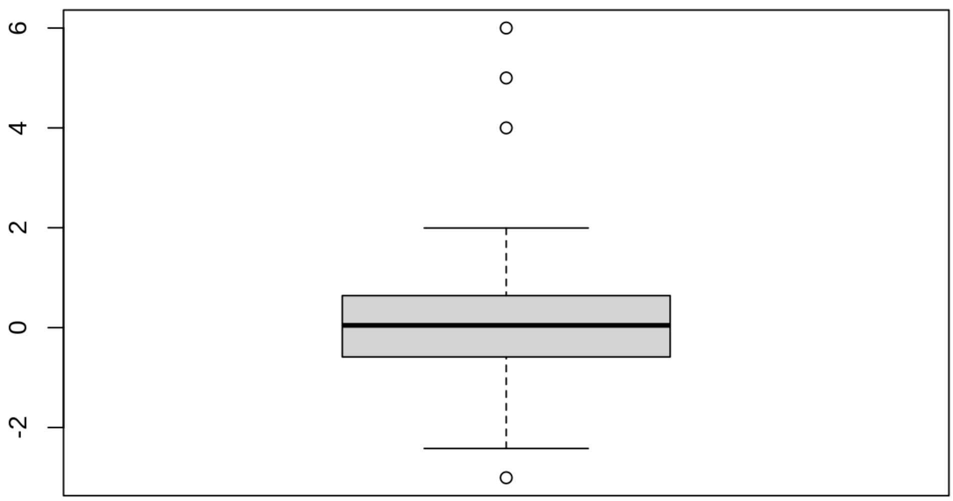
```
veri bol ise sil

değerli ise basit ya da tahmine dayalı değer ata
```

tek değişken aykırı

```
set.seed(54)
veri <- rnorm(100)
summary(veri)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -3.005748 -0.594959  0.001358 -0.059670  0.557480  1.994794
```

```
veri <- c(veri, c(4,5,6))
boxplot(veri)
```




```
boxplot.stats(veri)$out          #aykırı değerleri verir.
```

```
## [1] -3.005748  4.000000  5.000000  6.000000
```

```
which(veri %in% boxplot.stats(veri)$out)  #aykırı değerlerin indekslerini verir.
```

```
## [1]  98 101 102 103
```

iki veya daha fazla

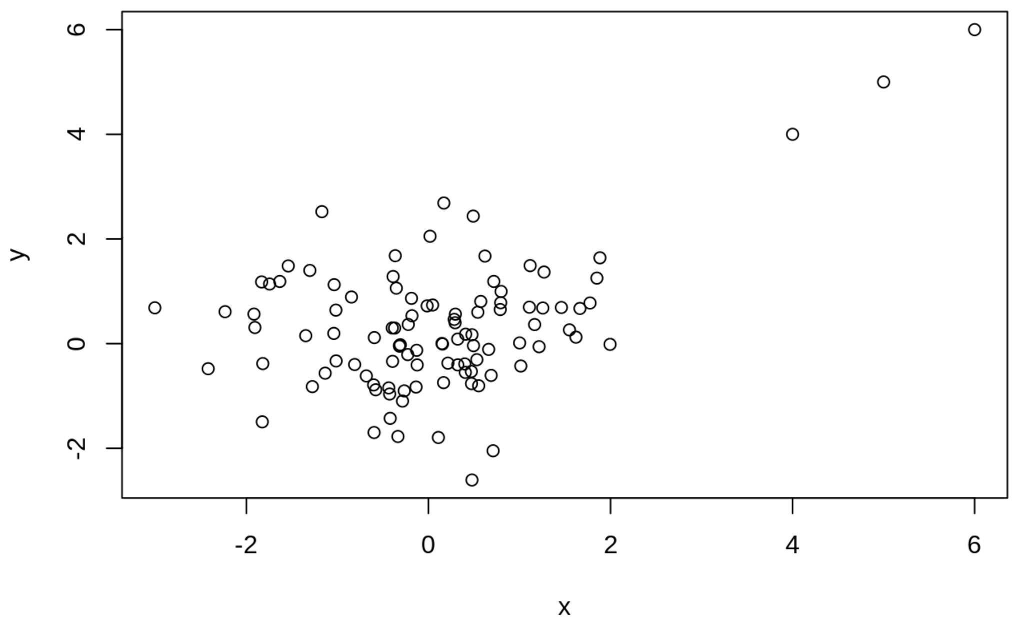
```
set.seed(54)
x <- rnorm(100)
x <- c(x, c(4,5,6))

set.seed(455)
y <- rnorm(100)
y  <- c(y, c(4,5,6))

df <- data.frame(x, y)

a <- which(df$x %in% boxplot.stats(df$x)$out)      #aykırı değerleri a'ya ata
b <- which(df$y %in% boxplot.stats(df$y)$out)

intersect(boxplot.stats(df$x)$out, boxplot.stats(df$y)$out)
```

```
## [1] 4 5 6
```

```
ortak <- intersect(a,b) #indeksler
ortak
```

```
## [1] 101 102 103
```

```
plot(df)          #genel bakış
```




```
#points(df[ortak_tum, ], col = "red", pch = "+", cex = 2.5)

#ortak_tum <- union(a,b)
```

3.Aykırı gözlem problemini çözmek

3.1 silme

```
set.seed(54)
x <- rnorm(100)
x <- c(x, c(4,5,6))

set.seed(455)
y <- rnorm(100)
y  <- c(y, c(4,5,6))

df <- data.frame(x, y)


ortak_tum <- union(a,b)
df[-ortak_tum,]
```

```
##             x            y
## 1    1.88379189  1.640216449
## 2    0.49458450 -0.038685759
## 3   -0.36465169  1.679672272
## 4    1.62062305  0.122970294
## 5    1.16632493  0.364185572
## 6   -1.03942740  0.194995684
## 7   -0.01455964  0.721258707
## 8   -1.17047647  2.521680839
## 9    1.77494017  0.776505340
## 10   0.78859615  0.651529124
## 11   1.66401163  0.671787351
## 12   0.47801682  0.169779768
## 13  -0.22171347  0.366136827
## 15  -2.23401654  0.610089343
## 16  -1.30288209  1.398774155
## 17   0.29292466  0.399088697
## 18  -0.31703132 -0.048940098
## 19  -0.81071149 -0.399323496
## 20   1.11726681  1.490754687
## 21  -1.81983295 -0.381600011
## 22   0.39944853 -0.388606822
## 23  -0.13606239 -0.829091562
## 24   1.00166983  0.013298826
## 25   0.66140161 -0.110158299
## 26   1.10883164  0.696482046
## 27  -0.26595629 -0.902886735
## 28  -1.74510970  1.139108571
## 29   0.16606626 -0.745332764
## 30  -0.84554552  0.890738702
## 31   0.53168490 -0.306845460
## 32   0.49181625  2.436394366
## 33   0.62061269  1.672896510
## 34   1.26985635  1.367375342
## 35  -0.30936626 -0.021613474
## 36   0.47289420 -0.764055374
## 37   1.85027380  1.252042804
## 38  -1.54000026  1.484873124
## 39   0.68864833 -0.606909774
## 40  -0.68249373 -0.617530732
## 41  -1.83243914  1.179212703
## 42  -0.42021630 -1.427790370
## 43   0.79358126  0.781419796
## 45  -0.18633158  0.865345035
## 46  -1.90651465  0.307903875
## 47  -0.22748100 -0.209337328
## 48  -0.39793928  0.299944587
## 49  -0.18124660  0.531780839
## 50  -0.12948997 -0.127533101
## 51  -1.01671443  0.640326190
## 52  -0.33526559 -1.774308054
## 53  -0.57933184 -0.881894717
## 54   0.04562350  0.737628413
## 55   1.21554735 -0.058443256
## 56   0.21389184 -0.372732960
## 57   0.55180719 -0.805415042
## 58   0.15433916 -0.007754222
## 59   0.32143044  0.086854230
## 60  -0.43602756 -0.846072225
## 61   0.10909174 -1.793648281
## 62   0.71008151 -2.047481212
## 63  -0.38838482  1.282677178
## 64  -1.34789142  0.151669272
## 65   1.54890116  0.263024666
## 66   1.01445562 -0.427395121
## 67  -1.91645710  0.564354800
## 68   0.28328974  0.463288067
## 69  -1.03562955  1.126592357
## 70   0.01727577  2.052020082
## 71  -0.37172844  0.297515863
## 72   0.54220295  0.600250324
## 73  -1.13470295 -0.564671613
## 74  -0.28499945 -1.097823643
## 75  -0.59419355  0.115292040
## 76  -1.01451370 -0.331118174
## 77   0.47129741 -0.532081567
## 78   1.25604353  0.680524404
## 79   0.39418180  0.338699077
```

```
## 79  -0.39418180 -0.338699077
## 80  -1.63296094  1.188155748
## 81  -0.35296616  1.059874624
## 82   0.40490390 -0.549166078
## 83  -1.82547586 -1.495025057
## 84  -0.60197549 -0.788883820
## 85  -1.27538942 -0.821693260
## 86   0.40850565  0.180785192
## 87   0.14876981  0.003407964
## 88  -0.59725443 -1.698231187
## 89   0.71774281  1.188530280
## 90   0.79791837  0.996993879
## 91  -0.12402052 -0.407495694
## 92   0.32232735 -0.405742012
## 93   0.57449690  0.805932325
## 94   1.46038154  0.690659591
## 95  -2.41977112 -0.479418087
## 96  -0.42633465 -0.962895479
## 97  -0.31933342 -0.037879250
## 99   1.99479423 -0.014222721
## 100  0.29559711  0.563528679
```

```
summary(df[-ortak_tum,])
```

```
##        x                  y
##  Min.   :-2.41977   Min.   :-2.0475
##  1st Qu.:-0.59419   1st Qu.:-0.4075
##  Median :-0.01456   Median : 0.1517
##  Mean   :-0.03720   Mean   : 0.1750
##  3rd Qu.: 0.57450   3rd Qu.: 0.7376
##  Max.   : 1.99479   Max.   : 2.5217
```

3.2. Aykiri Gozlemlerin Ortalama ile Doldurulmasi

```
set.seed(54)
x <- rnorm(100)
x <- c(x, c(4,5,6))

set.seed(455)
y <- rnorm(100)
y  <- c(y, c(4,5,6))

df <- data.frame(x, y)

a <- which(df$x %in% boxplot.stats(df$x)$out)
b <- which(df$y %in% boxplot.stats(df$y)$out)

df[a, ]$x
```

```
## [1] -3.005748  4.000000  5.000000  6.000000
```

```
df[a, ]$x <- mean(df$x)
summary(df$x)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.41977 -0.50768  0.08770 -0.02534  0.54701  1.99479
```

3.3 Aykiri Gozlemlerin Baskilanmasi

```
#3.çeyrek değer yazdırılır.

set.seed(54)
x <- rnorm(100)
x <- c(x, c(4,5,6))

set.seed(455)
y <- rnorm(100)
y  <- c(y, c(4,5,6))

df <- data.frame(x, y)

a <- which(df$x %in% boxplot.stats(df$x)$out)
b <- which(df$y %in% boxplot.stats(df$y)$out)

summary(df$x)    #5.indeks 3.çeyrek değer olduğundan
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.00575 -0.58676  0.04562  0.08770  0.64101  6.00000
```

```
summary(df$x)[5]   #bu şekilde 3.çeyrek değere erişebiliriz
```

```
##   3rd Qu.
## 0.6410072
```

```
df[a, ]$x<- summary(df$x)[5]
#veya fivenum() ile atayabiliriz.
df[a, ]$x <- fivenum(df$x)[4]
```

3.4. Aykiri Gozlemlerin Tahminle Doldurulmasi

```
set.seed(54)
x <- rnorm(100)
x <- c(x, c(4,5,6))

set.seed(455)
y <- rnorm(100)
y  <- c(y, c(4,5,6))

df <- data.frame(x, y)

a <- which(df$x %in% boxplot.stats(df$x)$out)
b <- which(df$y %in% boxplot.stats(df$y)$out)

ortak_tum <- union(a,b)

df[ortak_tum,]
```

```
##              x          y
## 98  -3.0057480  0.6845525
## 101  4.0000000  4.0000000
## 102  5.0000000  5.0000000
## 103  6.0000000  6.0000000
## 14   0.4780069 -2.6059174
## 44   0.1691997  2.6865560
```

```
df[a, ]$x <- NA
df[b, ]$y <- NA
summary(df)
```

```
##        x                  y
##  Min.   :-2.41977   Min.   :-2.0475
##  1st Qu.:-0.58676   1st Qu.:-0.4071
##  Median : 0.01728   Median : 0.1607
##  Mean   :-0.02991   Mean   : 0.1802
##  3rd Qu.: 0.56315   3rd Qu.: 0.7335
##  Max.   : 1.99479   Max.   : 2.5217
##  NA's   :4          NA's   :5
```

```
df_all <- data.frame(df, rnorm(103), rnorm(103),rnorm(103))
#rf_data <- missForest(df_all)

#rf<- rf_data$ximp
#summary(rf)
```

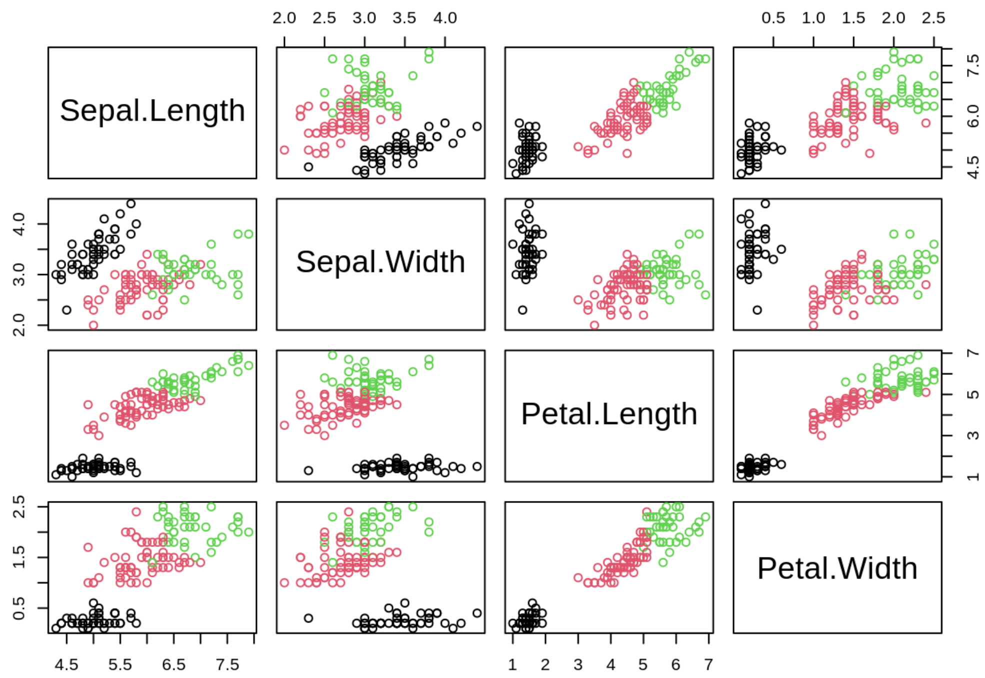
4. çok değişkenli aykırı gözlem

4.1 local outlier factor

4. Kumeleme Yontemi Ile Aykiri Gozlem Analizi

4.1 Aykiri Gozlemlere Erismek

```
#4,5 kümeye ayır. hiçbir yere ayrılamayan aykırı denebilir
#kmeansin içinde kümeleri ve merkezleri vardır
df <- iris[,1:4]

k_ort <- kmeans(df, centers = 3)
plot(df, col = k_ort$cluster)
```




```
k_ort$centers
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.006000    3.428000     1.462000    0.246000
## 2     5.901613    2.748387     4.393548    1.433871
## 3     6.850000    3.073684     5.742105    2.071053
```

```
k_ort$cluster
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3 2 3 3 3 3
## [112] 3 3 2 2 3 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 3 3 3 3 2 3 3 3 2 3 3 3 2 3
## [149] 3 2
```

```
merkez_df <- k_ort$centers[k_ort$cluster,]


uzakliklar <- sqrt(rowSums(df, merkez_df)^2)

aykirilar <- order(uzakliklar, decreasing = T)[1:10]
print(df[aykirilar,])
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width
## 118          7.7         3.8          6.7         2.2
## 132          7.9         3.8          6.4         2.0
## 119          7.7         2.6          6.9         2.3
## 110          7.2         3.6          6.1         2.5
## 106          7.6         3.0          6.6         2.1
## 123          7.7         2.8          6.7         2.0
## 136          7.7         3.0          6.1         2.3
## 108          7.3         2.9          6.3         1.8
## 126          7.2         3.2          6.0         1.8
## 131          7.4         2.8          6.1         1.9
```