Aykırı Değer

Ali Valiyev

2024-04-16

Aykırı değer tespiti Box-Plot Yöntemi - İlk olarak "Rstatix" Paketini indiriyoruz

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.4     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
##
## The following object is masked from 'package:stats':
##
##     filter
```

```r
library(dplyr)
Dataset = ggplot2::diamonds
```

identify_outliers fonksiyonunu inceleyelim

```r
?identify_outliers
```

```
## starting httpd help server ... done
```

Box-Plot yöntemında 1. ve 3. çeyreklik değerler hesaba katılıyor. Fazla olan değerlerin aykırı değer olup olmadığını kontrol etmek için Q3+ 1.5*IQR IQR=Q3-Q1

```r
out = identify_outliers(Dataset["price"])
```

is.outlier - Aykırı değer olup olmadığını gösteriyor is.extreme - Extreme aykırı değer olup olmadığını gösteriyor

```r
names(out)
```

```
## [1] "price"      "is.outlier" "is.extreme"
```

minimun outlier değer:

```r
min(out$price)
```

```
## [1] 11886
```

minimun outlier değer:

```r
max(out$price)
```

```
## [1] 18823
```

yalnizca Extreme aykırı değerleri çıkarmak için:

```r
indexes = which(out$is.extreme == TRUE)
out[indexes, "price"]
```

```
## # A tibble: 120 x 1
##    price
##    <int>
##  1 18458
##  2 18462
##  3 18462
##  4 18468
##  5 18470
##  6 18472
##  7 18474
##  8 18475
##  9 18477
## 10 18480
## # i 110 more rows
```

extreme değerlerin sayısını bulma:

```
extreme = out[indexes, "price"]
length(extreme)
```

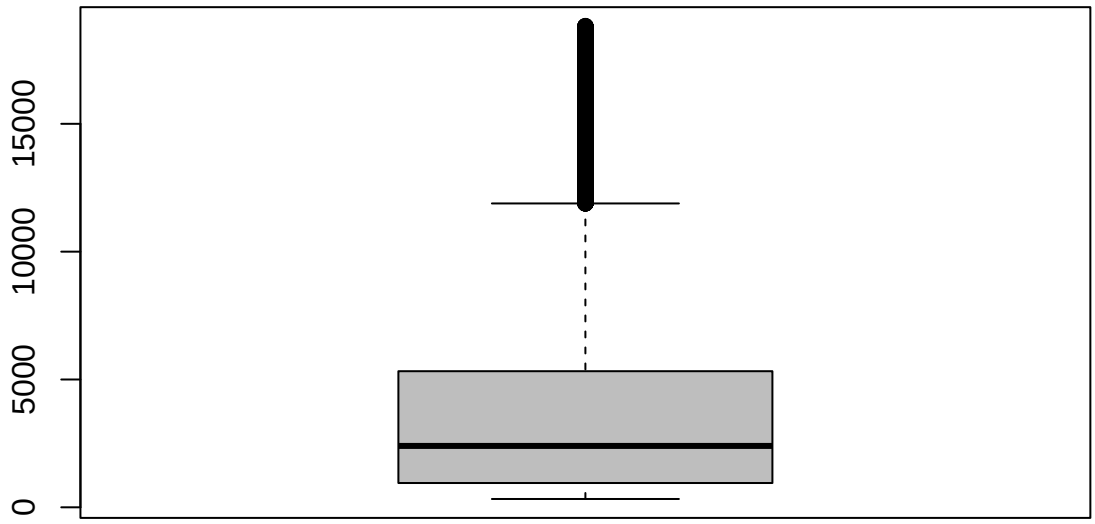
```
## [1] 1
```

outlier değerlerin sayısını bulma:

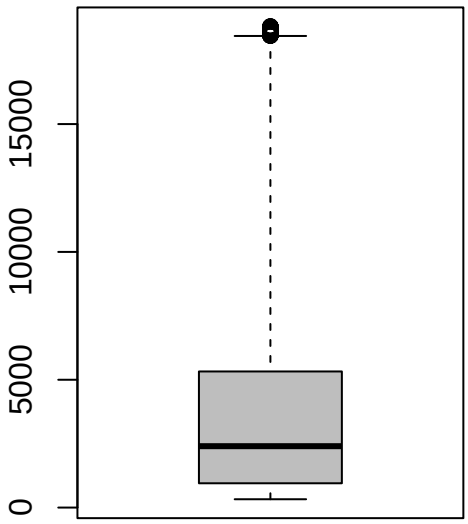
```
nrow(out)
```

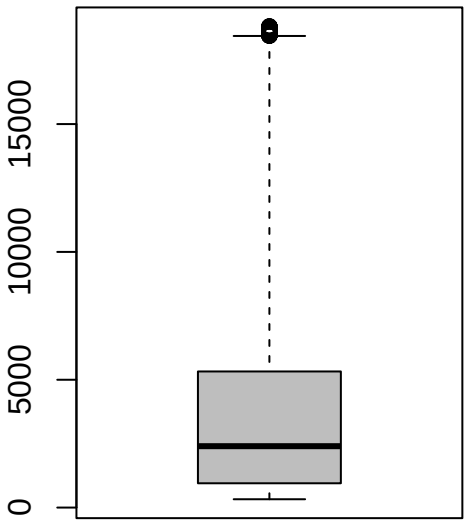
```
## [1] 3540
```

```
#Box-Plot grafiği
boxplot(Dataset["price"], col="gray")
```



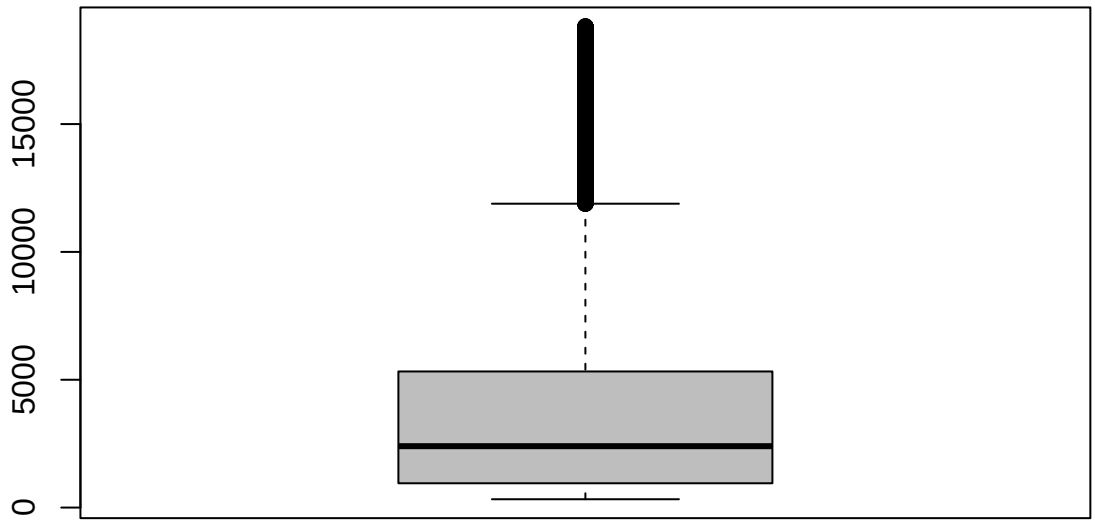

```
#Aynı veride 2 farklı box-plot grafiği
opar = par(mfrow = c(1,2))
boxplot(Dataset["price"], col = "gray" , range = 3)
boxplot(Dataset["price"], col = "gray" , range = 3)
```






```
par(opar)
```

```
#Aykırı değer listeleme:
bpx = boxplot(Dataset["price"], col="gray")
```




```
head(bpx$out,nL=10)
```

```
## [1] 11888 11888 11888 11897 11899 11899
```

```
#Aykırı değeri box-plot state ile bulma
bpstx = boxplot.stats(Dataset$price)
head(bpstx$stats)
```

```
## [1]   326.0   950.0  2401.0  5324.5 11886.0
```

```
head(bpstx$n)
```

```
## [1] 53940
```

```
head(bpstx$conf)
```

```
## [1] 2371.24 2430.76
```

```
head(bpstx$out,100)
```

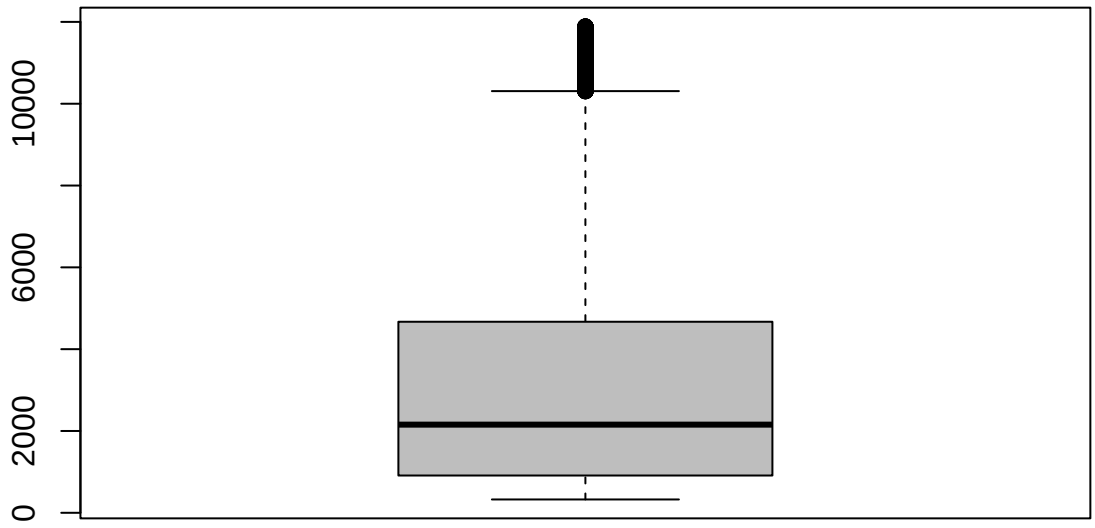
```
##   [1] 11888 11888 11888 11897 11899 11899 11901 11903 11904 11905 11906 11912
##  [13] 11913 11917 11917 11921 11922 11923 11923 11923 11924 11925 11926 11927
```

```
##  [25] 11927 11933 11934 11935 11939 11942 11943 11946 11946 11946 11946 11948
##  [37] 11948 11950 11951 11954 11955 11956 11957 11957 11957 11958 11962 11963
##  [49] 11965 11966 11966 11967 11968 11968 11969 11970 11971 11971 11973 11975
##  [61] 11975 11975 11976 11979 11982 11985 11986 11988 11988 11988 11988 11990
##  [73] 11998 11999 12000 12004 12005 12008 12009 12012 12013 12014 12016 12021
##  [85] 12028 12030 12030 12030 12030 12030 12030 12031 12032 12035 12036 12038
##  [97] 12044 12047 12047 12048
```

```
#Aykırı değerden temizlenmiş veri
adx = bpstx$out
cx = Dataset$price[-which(Dataset$price %in% adx)]
head(cx,100)
```

```
##   [1]  326  326  327  334  335  336  336  337  337  338  339  340  342  344  345
##  [16]  345  348  351  351  351  351  352  353  353  353  354  355  357  357  357
##  [31]  402  402  402  402  402  402  402  402  403  403  403  403  403  403  403
##  [46]  403  403  403  404  404  404  404  404  404  404  405  405  405  405  405
##  [61]  552  552  552  552  552  553  553  553  553  553  553  554  554  554  554
##  [76]  554  554  554  554  554  554  554  554  554  554  554  554  554  554  554
##  [91] 2757 2757 2757 2759 2759 2759 2759 2759 2760 2760
```

```
#Aykırı değerden temizlenmiş veri grafiği
bpx1 = boxplot(cx, col="gray")
```



```
head(bpx1$out,100)
```

```
##   [1] 10309 10309 10311 10312 10313 10313 10314 10314 10314 10314 10316 10316
##  [13] 10317 10317 10317 10319 10321 10327 10329 10330 10330 10331 10331 10331
##  [25] 10332 10333 10333 10333 10333 10333 10335 10336 10337 10338 10338 10338
##  [37] 10339 10340 10340 10341 10341 10341 10342 10342 10342 10345 10346 10346
##  [49] 10349 10349 10349 10349 10350 10350 10350 10351 10351 10351 10352 10353
##  [61] 10356 10356 10357 10357 10359 10362 10362 10362 10365 10367 10367 10367
##  [73] 10368 10371 10372 10374 10377 10378 10378 10378 10380 10384 10384 10387
##  [85] 10388 10388 10389 10389 10392 10395 10396 10396 10396 10398 10399 10401
##  [97] 10401 10406 10407 10409
```

"Z", "T" ve ve "ChiSquare" Skorlarına göre aykırı değer kontrolü: Böyle bir işlem için scores fonskıyonunu kullanıyoruz Type - Aykırı değerin skorunun hangi tipde olacağını belirtiyoruz prop - Hangi kısımdan sonrası aykırı değer olarak hesaplanıcak

```
library(outliers)
head(scores(Dataset$price, type = "z" , prob = 0.6 ),100)
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

TRUE-lar Aykırı değerleri bize bildiriyor.

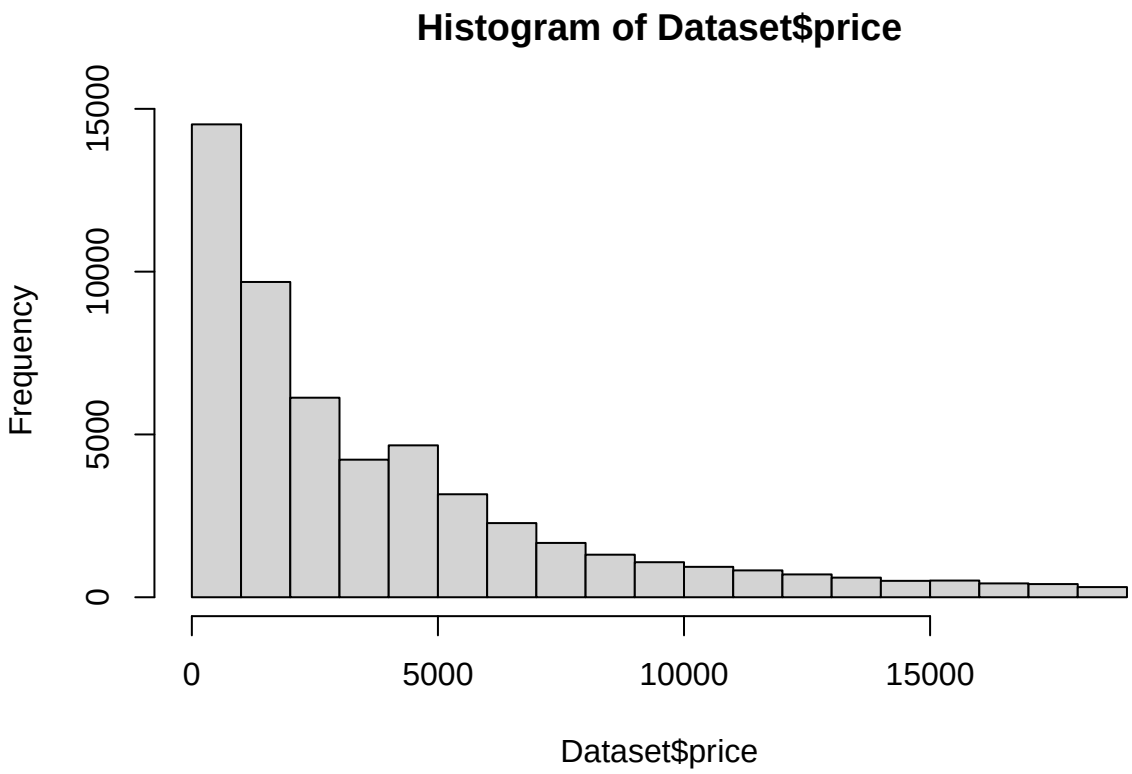
Değeri görmek için:

```
out = scores(Dataset$price, type = "z" , prob = 0.6 )
value = which(out == TRUE)
head(Dataset$price[value])
```

```
## [1] 326 326 327 334 335 336
```

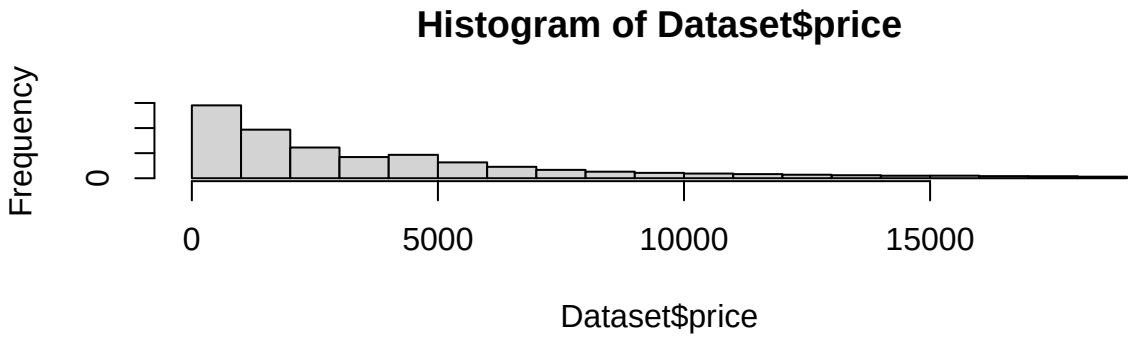
en küçük değer olarak 326 bulunmuş. Bu yüzden de 326 bizim minimum Aykırı değerimizdir. Histogram olarak inceleme

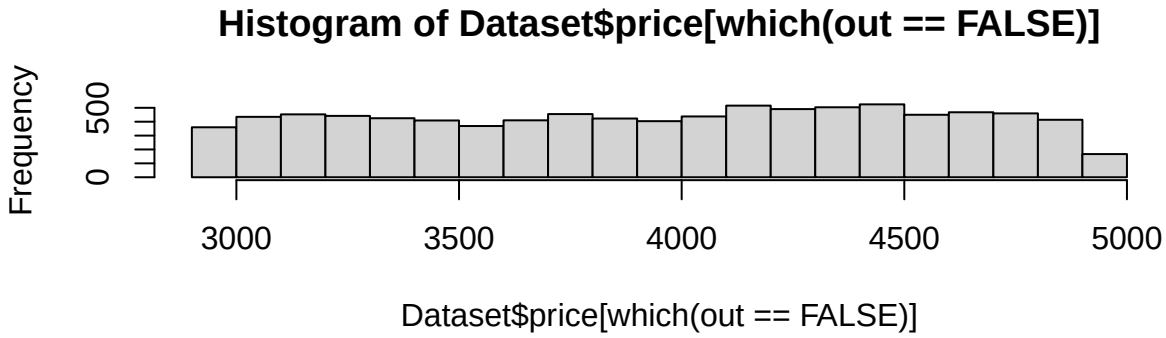
```
hist.default(Dataset$price)
```



326-dan sonra gelen değerler aykırı değer olarak sayılmış

```
#Altlı Üstlü Histogram Grafiği
out = scores(Dataset$price, type = "z" , prob = 0.6 )
par(mfrow = c(2,1))
hist(Dataset$price)
hist(Dataset$price[ which(out == FALSE)])
```





T dağılıma göre :

```
head(scores(Dataset$price, type = "t" , prob = 0.6 ),100)
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Değeri görmek için:

```
out = scores(Dataset$price, type = "t" , prob = 0.6 )
value = which(out == TRUE)
head(Dataset$price[value])
```

```
## [1] 326 326 327 334 335 336
```

Chi Square dağılına göre :

```
head(scores(Dataset$price, type = "chisq" , prob = 0.6 ),100)
```

```
##   [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [25]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [37]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [49]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [61]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [73]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [85]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE
```

Değeri görmek için:

```
out = scores(Dataset$price, type = "chisq" , prob = 0.6 )
value = which(out == TRUE)
head(Dataset$price[value])
```

```
## [1] 326 326 327 334 335 336
```