Kutu-Grafik Gösterimi ve Veri Kümelerinin Karşılaştırılması

Kutu-grafikleri (Box-plotlar) veri kümeleri arasındaki benzerlik ya da farklılıkları görmek için kullanılır.

Kutu-grafiği kullanılarak veri kümesinin

1. Konumu
2. Yayılımı
3. Çarpıklığı
4. Kuyruk uzunluğu
5. Aykırı değerleri

tespit edilmektedir.

Örnek: Aşağıdaki veri setinde Türkiye'de nüfusu en çok olan 16 şehrin 2018 yılına ait nüfusları verilmektedir.

No	Şehir	Nüfus(×10000)
16	İstanbul	1506
15	Ankara	550
14	İzmir	432
13	Bursa	299
12	Antalya	243
11	Adana	222
10	Konya	221
9	Şanlıurfa	204
8	Gaziantep	203
7	Kocaeli	191
6	Mersin	181
5	Diyarbakır	173
4	Hatay	161
3	Manisa	143
2	Kayseri	139
1	Samsun	134

16 şehir için nüfuslar küçükten büyüğe doğru sıralandığında

134 139 143 161 173 181 191 203 204 221 222 243 299 432 550 1506

olarak elde edilir.

Şehirlerin bu sıralamaya göre elde edilen sıra numaraları tablonun ilk sütununda verilmiştir.

Bu veri setindeki toplam gözlem sayısının

$$n = 16$$

olduğu görülür.

Medyanın derinliği aşağıdaki eşitlik kullanılarak

$$Medyanın\ derinliği = \frac{16 + 1}{2} = 8.5$$

olarak hesaplanır. Buradan,

$$Medyan = \frac{X_{(8)} + X_{(9)}}{2} = \frac{203 + 204}{2} = 203.5$$

olarak bulunur. Medyanın derinliği kullanılarak dörtlüğün derinliği

$$Dörtlüğün\ derinliği = \frac{[8.5] + 1}{2} = 4.5$$

olarak hesaplanır. Buradan, alt dörtlük ve üst dörtlük sırasıyla

$$Alt\ dörtlük = \frac{X_{(4)} + X_{(5)}}{2} = \frac{161 + 173}{2} = 167$$
$$Üst\ dörtlük = \frac{X_{(12)} + X_{(13)}}{2} = \frac{243 + 299}{2} = 271$$

olarak bulunur.

Veri setindeki en küçük ve en büyük gözlemler olarak tanımlanan uç değerlerin ise sırasıyla

$$X_{(1)} = 134 \quad ve \quad X_{(16)} = 1506$$

olduğu görünür.

Bu değerler kullanılarak, 5 değerli özet

#16

M	8.5
F	4.5
	1

203.5	
167	271
134	1506

olarak elde edilir.

Bu veri seti için dörtlüğün yayılımı

$$d_F = 271 - 167 = 104$$

olarak hesaplanır. Veri setinde aykırı değer olup olmadığını belirlemek için alt ve üst kesim noktaları sırasıyla,

$$c_L = F_L - 1.5d_F = 167 - (1.5)104 = 11$$
$$c_U = F_U + 1.5d_F = 271 + (1.5)104 = 427$$

olarak elde edilir.

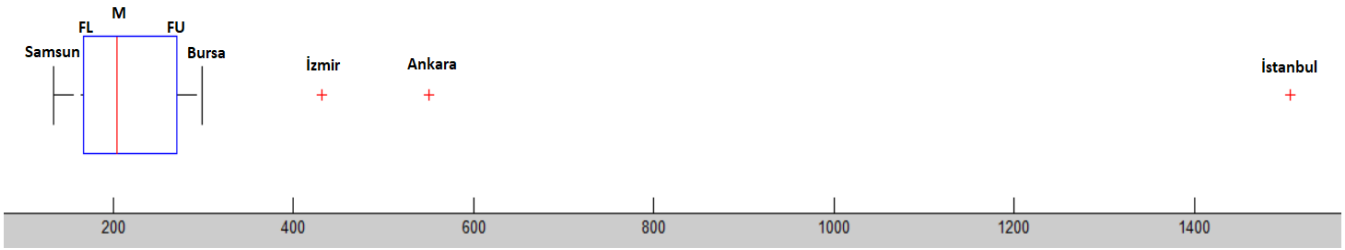
Bu veri setinde nüfusu 110.000'den az olan şehir bulunmamaktadır. Ancak, İstanbul, Ankara ve İzmir şehirlerinin nüfusları 4.270.000'den fazla olduğu için bu şehirler nüfus bakımından aykırı değer olarak belirlenir.

Kutu-grafiğinin çizimi

1. Öncelikle alt dörtlükle başlayan ve üst dörtlükle biten bir kutu çizilir.
2. Bu kutu içerisinde medyanın yeri belirlenir.

3. F_L'den sola doğru, F_U'dan sağa doğru aykırı değer olmayan gözlem değerine kadar çizgi çizilir.
4. Kesim noktalarının dışında kalan aykırı değerler işaretlenir.

Yukarıdaki veri seti için bu adımlar izlenerek elde edilen kutu-grafiği aşağıdaki gibi elde edilir.



Box-plotlar veri kümeleri arasındaki benzerlik ve farklılıkları görmek için kullanılır.

Yorum:

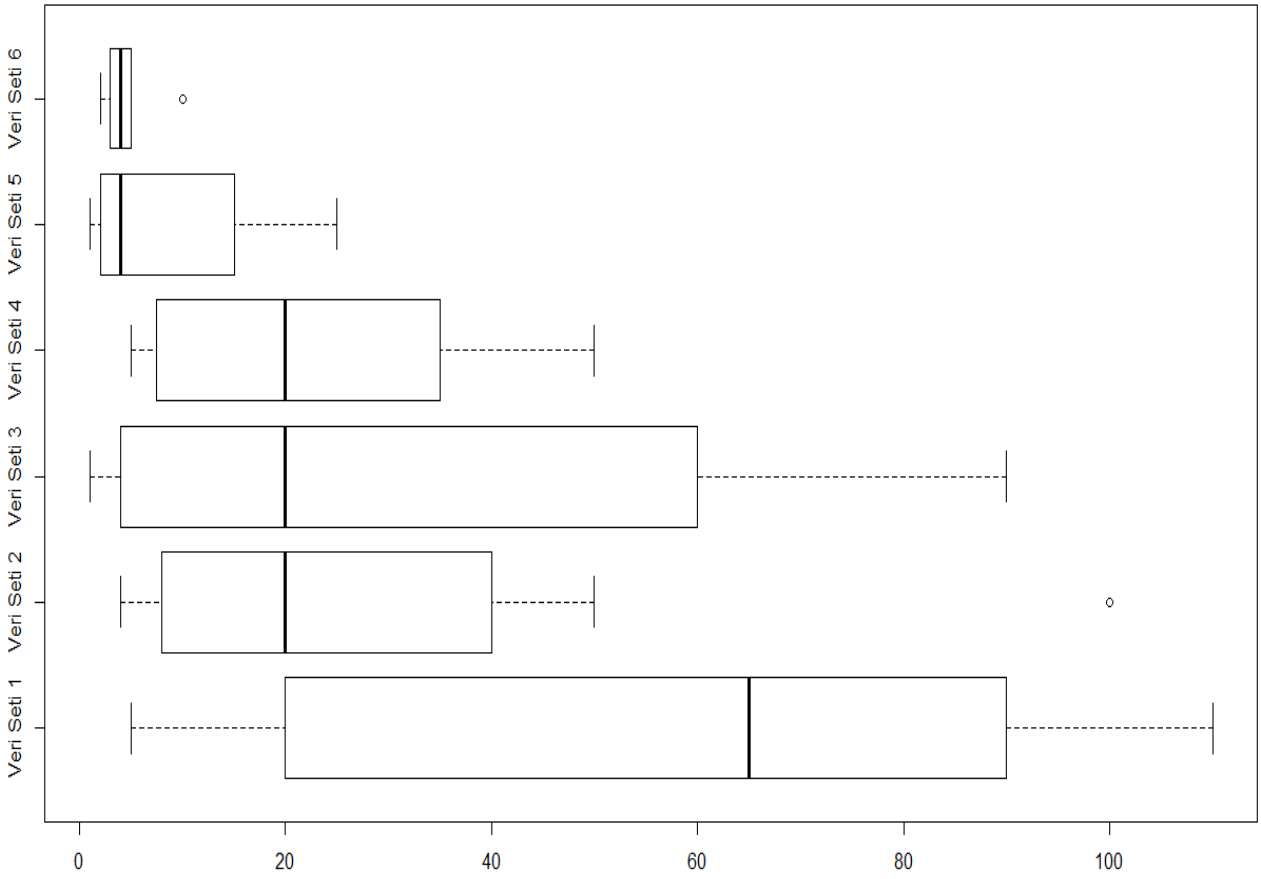
- Bu veri setinde İstanbul, Ankara ve İzmir nüfus bakımından aykırı değerdir.
- Veri setinin dağılımı sağ çarpıktır

Örnek: Harf değerleri konusunda verilen örnek için 5 değerli özetler elde edilmişti. Aşağıdaki tabloda bu özet değerleri aykırı değerlerle beraber gösterilmiştir.

Özet tablo:

Veri kümesi	**M**	**F_L**	**F_U**	**d_F**	**Aykırı değer**
1	65	20	90	70	-
2	20	8	40	32	100
3	20	4	60	56	-
4	20	7.5	35	27.5	-
5	4	2	15	13	-
6	4	3	5	2	10

Bu değerler kullanılarak elde edilen kutu-grafikleri aşağıda verilmiştir.



Yorum:

- 1. veri setinin yayılımı en büyük, 6. veri setinin yayılımı en küçüktür, dolayısıyla en heterojen veri seti 1 en homojen veri seti 6 dır.
- 1. veri setinin dağılımı negatif çarpık, 2., 3. ve 4. veri setlerinin dağılımı ise pozitif çarpıktır.
- 2. ve 6. veri setlerinde aykırı değerler vardır.