



# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

hafta-4

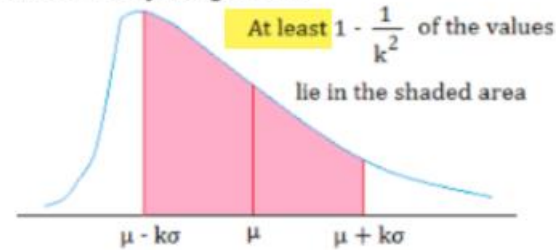
CEMİLE YILDIZÇAKAR

29.12.2020



# Chebychev Teoremi

Herhangi bir veri setinde, verilerin ortalamadan  $K$  standart sapma uzakta bulunma oranı  $1-1/K^2$  dir. Burada  $K$ , birden büyük pozitif sayıdır.



**$K=2$  ve  $K=3$  için;**

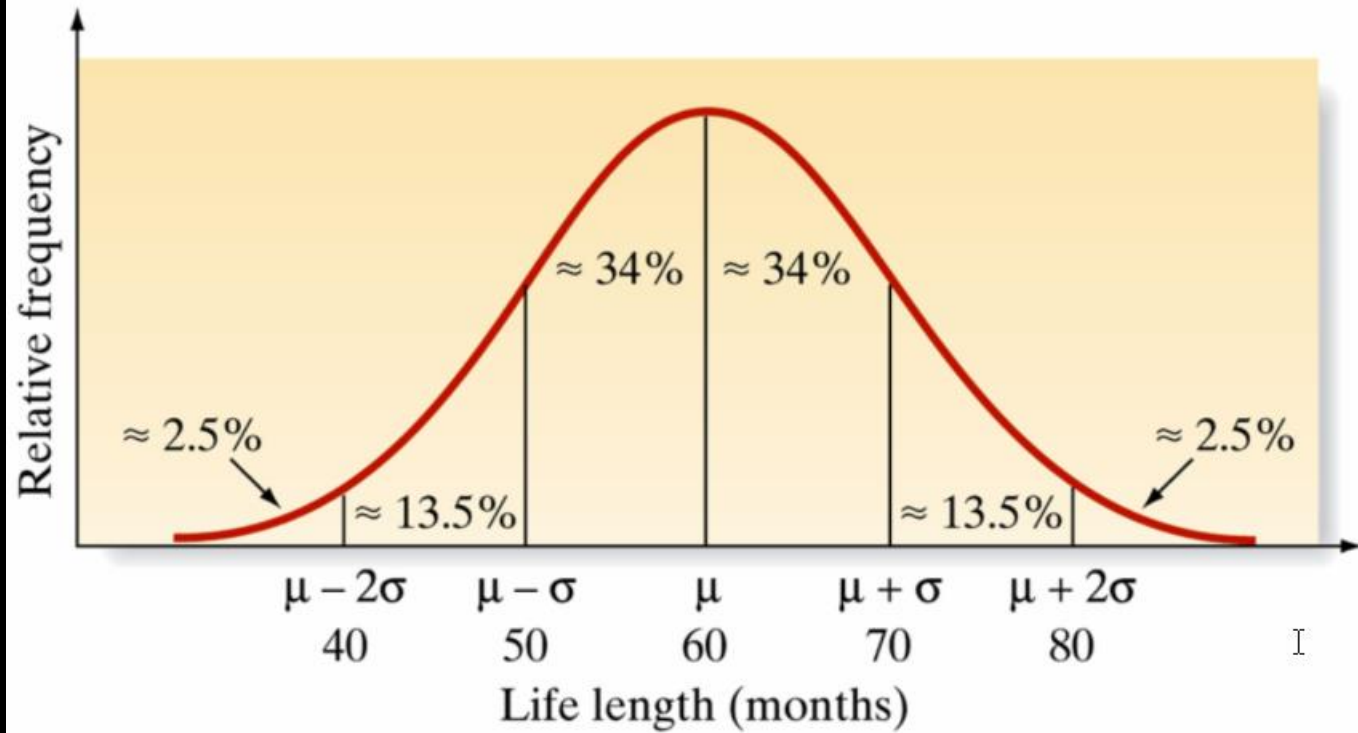
- Verilerin **en az**  $3/4$ ' ü (%75) ortalamanın 2 standart sapma uzağında bulunur.
- Verilerin **en az**  $8/9$ ' u (%89) ortalamanın 3 standart sapma uzağında bulunur.

$$\mu \pm \sigma$$



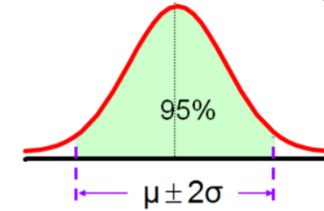
$$100[1 - (1/k^2)]\%$$

Cemile YILDIZÇAKAR

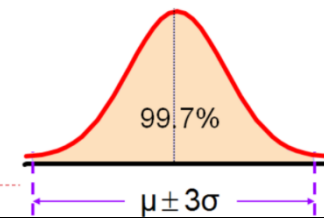


### Deneysel kural (The Empirical Rule)

□  $\mu \pm 2\sigma$  aralığında tüm verilerin %95'i yer alır.



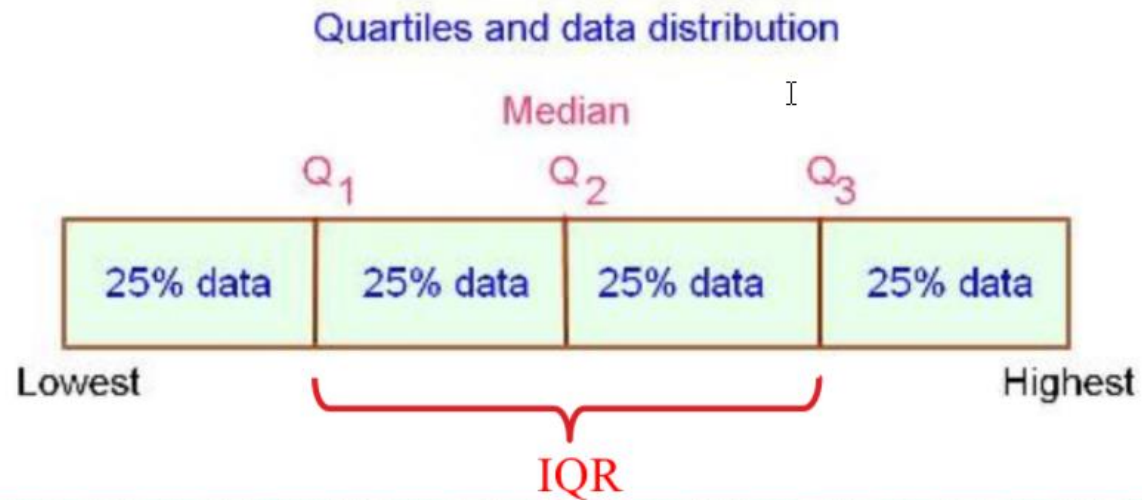
□  $\mu \pm 3\sigma$  aralığında tüm verilerin %99.7'si yer alır.



15.10.2019

Cemile YILDIZÇAKAR

A **quartile** divides a *sorted* data set into 4 equal parts, so that each part represents  $\frac{1}{4}$  of the data set

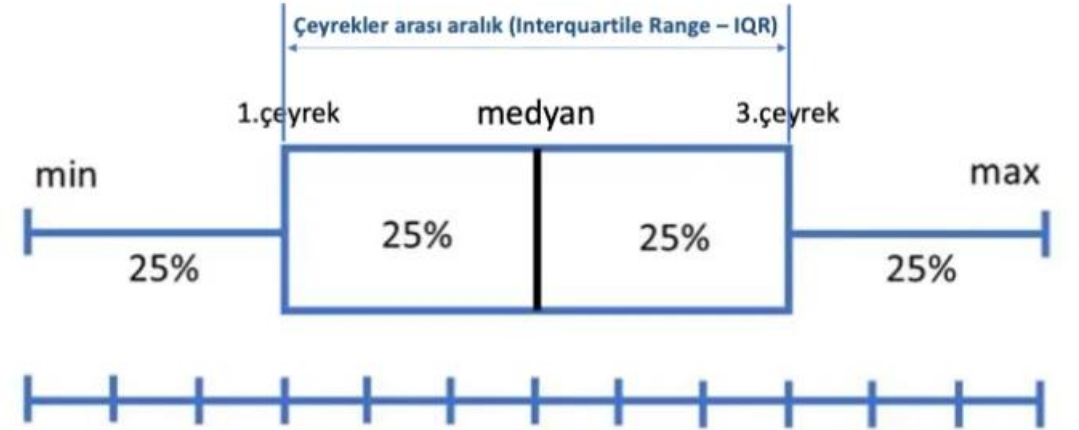


Cemile YILDIZÇAKAR

# Kutu grafiđi (Box Plot)

*Bir kutu grafiđi (Boxplot), veri eyreklerini (veya yzdelikleri) ve ortalamaları grntleiyerek sayısal verilerin ve deđiřkenliđin grsel olarak dađılımını gstermek iin kullanılır. Veri analizinde sıklıkla kullanılan bir grafik trdr.*

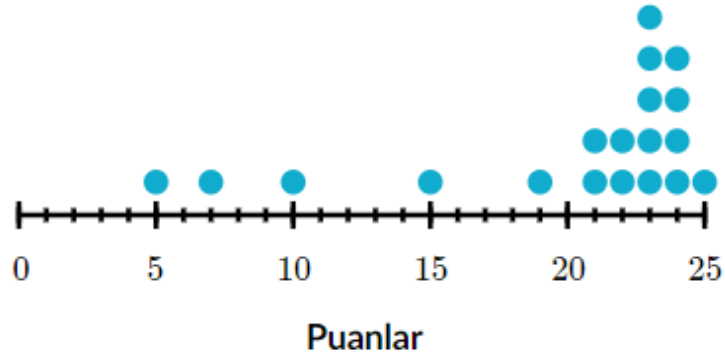
Ařađıdaki grnt, mkemmel bir normal dađılım olan verileri temsil eder ve ođu kutu grafiđinin bu simetriye uymaz.

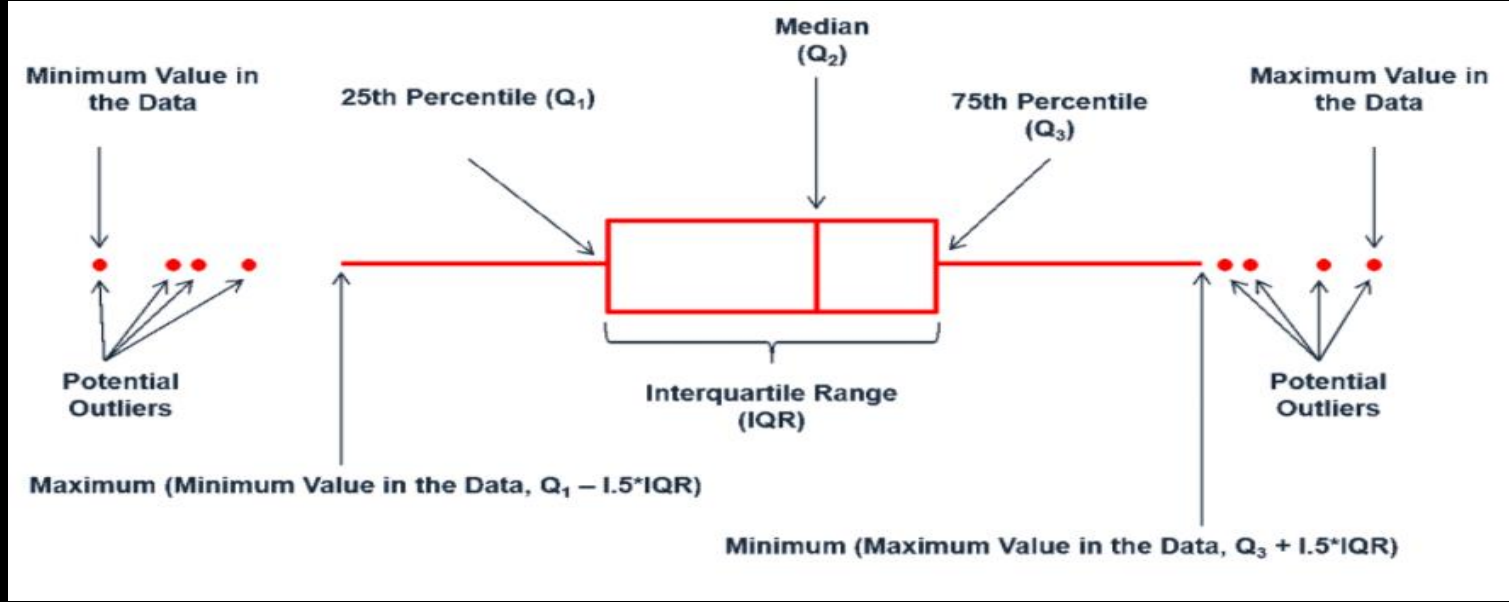


# KUTU GRAFİĞİ (BOXPLOT) NEREDE VE NASIL KULLANILIR?

- *Kutu grafiği, verilerdeki değerlerin nasıl yayıldığıнын iyi bir göstergesidir. Kutu grafikleri bir histograma göre ilkel gibi görünse de, birçok grup veya veri kümesi arasındaki dağılımları karşılaştırırken yararlıdır.*
- *Kutu ne kadar uzun olursa veri o kadar dağılmış olur. Kutu ne kadar küçük olursa veri o kadar az dağılmış olur.*

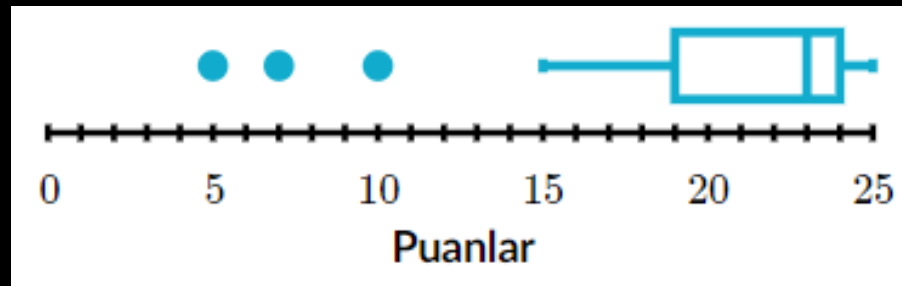
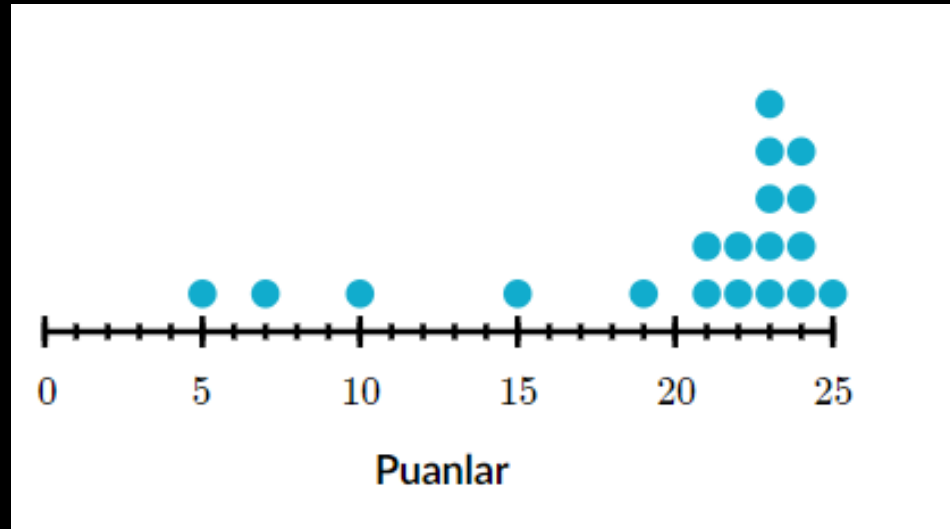
Aşağıdaki dağılım, 19 aday için bir sürücü sınavının puanlarını gösterir. Kaç aykırı değer görüyorsunuz?





Yaygın kullanılan bir kural, eğer bir veri noktası  $1,5 \cdot \text{ÇA}$  üçüncü çeyreğin üstündeyse veya birinci çeyreğin altındaysa, bu veri noktasının bir aykırı değer olduğunu söyler. Farklı şekilde söylersek, düşük aykırı değerler  $Q_1 - 1,5 \cdot \text{ÇA}$ 'nın altındadır ve yüksek aykırı değerler  $Q_3 + 1,5 \cdot \text{ÇA}$ 'nın üstündedir.





### 1- Simetrik – Normal :

Bir sürecin, normal dağılıma uygun olması verilerin ortalama etrafında homojen dağıldığını gösterir.

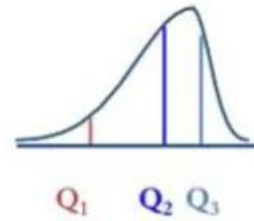
### 2-Eğik – Pozitif :

Verilerinizin alt limite yakın. Bu verilerin, üretilen bir ürünün bir ölçüsü olduğunu kabul edelim. 100 cm nominal değer beklentisi olan bir ölçü var. 0,1 cm de toleransınız olsun. Yani 99 – 101 cm aralığında ürettiğiniz her ürün kabul edilecektir. Ancak yapılan ölçümler sonucu toplanan veriler çoğunlukla 99 cm ve ona yakın değerler (99,4; 99,5; v.b.) çıkıyorsa üretim prosesiniz alt limite yakın üretim yapıyor demektir. İlerleyen dönemde

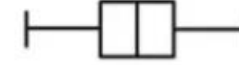
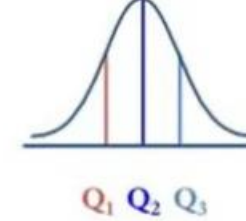
### 3-Eğik – Negatif :

Verilerinizin üst limite yakın olması demektir. prosesinizin kontrol dışına çıkıp belirtilen alt limit değerini

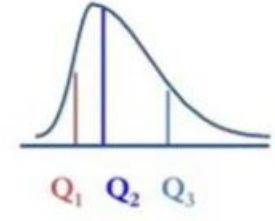
Eğik - Negatif



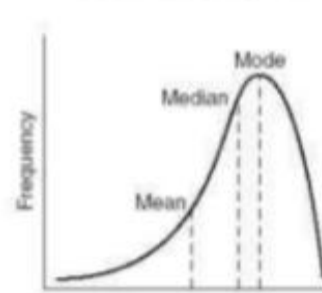
Simetrik - Normal



Eğik - Pozitif

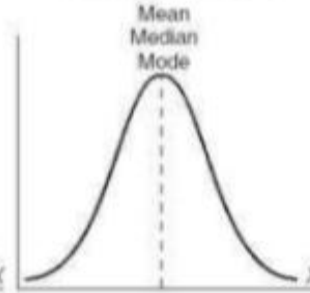


(a) Negatively skewed



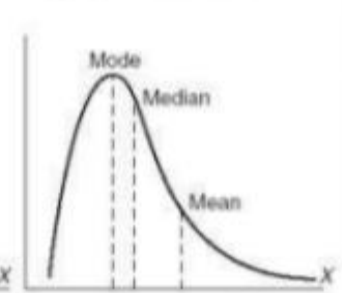
Negative direction

(b) Normal (no skew)



The normal curve represents a perfectly symmetrical distribution

(c) Positively skewed



Positive direction

# Standartlaştırma (z-skoru)

- Standartlaştırma → Her bir değişken değerinden, ortalamanın farkının alınması ve elde edilen farkın standart sapmaya bölünmesidir.
- Böylece ham veriler standart verilere dönüştürülerek, ölçü birimi farklılığı ortadan kaldırılmış olur.

$$z = \frac{x_i - \mu}{\sigma}$$

# Z-skoru

- *z-skoru veri setindeki gözlemlerin ortalamaya olan uzaklıklarını gösteren bir ölçüdür.*
- *z-skor pozitif ya da negatif olabilir.*
- *Bir diğer ifade ile; istatistikte, bir gözlemin z-skoru (veya standart skor), popülasyon ortalamasının üstünde veya altında standart sapma sayısıdır.*
- *Yine bir başka ifade ile, Z skoru yardımıyla elinizde bulunan örnek kümedeki sayısal verilerin, ortalamanın ne kadar altında ya da üstünde olduğunu görebilirsiniz.*
- *Bir z-skoru hesaplamak için popülasyon (hesaplanamıyor ise örneklemin) ortalamasını ve popülasyon standart sapmasını bilmelisiniz.*

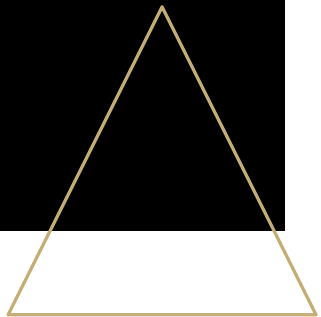
- Birbirinden farklı ölçü birimlerinin karşılaştırılmasında kullanılır.
- Z-score bütün veri yığınlarındaki birimlerin, ortak bir birim aralığına yığılmasını sağlar

$$z = \frac{x_i - \mu}{\sigma}$$

Formülde yer alan  $x(i)$  bizim gözlem değerimizdir.



Aşağıdakilere benzer soruları cevaplamak için bir z-puan görselleştirme oluşturabilirsiniz:

- Değerlerin yüzde kaçı belirli bir değerin altına düşüyor?
  - Hangi değerler olağanüstü sayılabilir? Örneğin, bir IQ testinde hangi puanlar ilk yüzde beşi temsil ediyor?
  - Bir dağıtımın diğerine karşı göreceli toplamı nedir? Örneğin, Michael ortalama bir erkekten daha uzun ve Emily ortalama bir kadın daha uzun, ancak cinsiyetleri arasında nispeten daha uzun kim?
- 

## Örnek

- Ayşe analiz arasından 80, istatistik arasından ise 68 almıştır.
- Analiz sınavında sınıf ortalaması 83, standart sapma ise 10'dur.
- İstatistik sınavında sınıf ortalaması 62, standart sapma ise 6'dır.
- Buna göre Ayşe hangi sınavda daha yüksek performans göstermiştir?

İpucu:


- Negatif bir z skoru, incelenen veri ortalamadan az demektir.
- Pozitif bir z skoru, incelenen veri ortalamadan çok demektir.

## Çözüm

$$\square Z_A = (80-83)/10 = -0.3$$

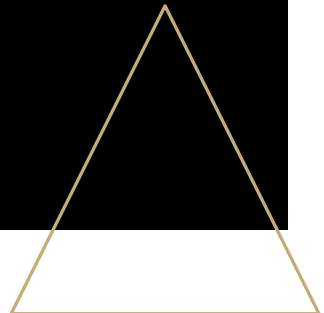
$$\square Z_I = (68-62)/6 = 1$$

- Ayşe sınıf arkadaşlarına göre istatistik sınavında daha başarılı olmuştur.




Örneğin ; Bir matematik sınav sonuçlarının olduğu veri setimiz olduğunu düşünelim. Bu sınav sonucunda ortalamanın ( $\mu$ ) 60 olduğu ve standart sapmanın ( $\sigma$ ) ise 10 olduğu tespit edilmiştir. Eğer 49 ve altında puan alan oranını bulmak istersek standardizasyon işlemi sonrası z-puan tablosunu kullanabiliriz.

$$Z = \frac{49 - 60}{10} = -1.1$$

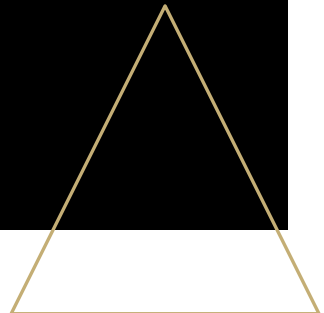




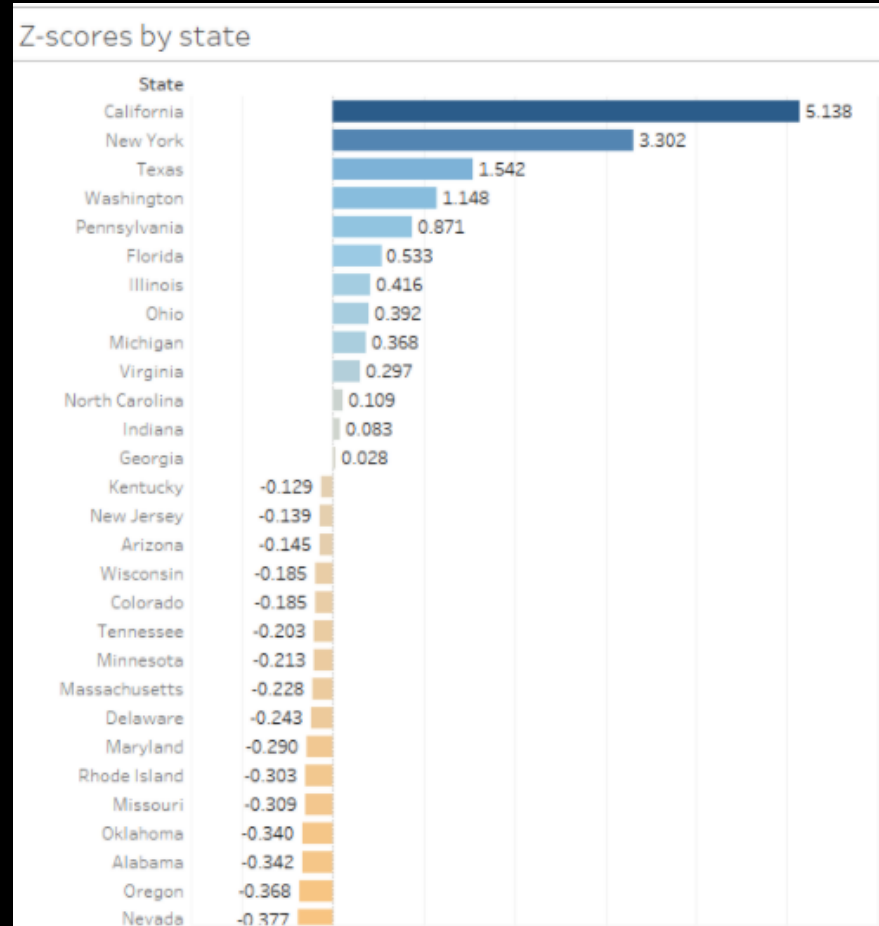


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379

z puanımız standardizasyon işlemi sonucunda -1.1 olarak bulundu. z -puan tablosuna bakıldığında toplam popülasyonun %13.57'sinin 49 ve daha altında puan aldığı tespit edilmiştir.



Genel bir kural olarak, -1.96'dan düşük veya 1.96'dan daha yüksek olan z-skorumları alıřılmadıđ ve ilginç olarak değerdendirilir. Yani, bunlar istatistiksel açıdan belirgin aykırı değerdlerdir.



California ve New York'un ikisinin de z skorumları 1,96'dan büyüktür.

Buradan Kaliforniya ve New York'un

diğerd eyaletlere kıyasla çok daha yüksek ortalama satıřlar elde ettiđine karar verebilirsiniz.

# Teşekkür Ederim



## LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



## Email

[yildizcakar.cemile@gmail.com](mailto:yildizcakar.cemile@gmail.com)

Cemile YILDIZÇAKAR

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB