GİRİŞ

Eğitmen

Özgür YILDIRIM

linkedin.com/in/Ozgur-yldrm/

github.com/OzgurYldrm

- İstatistiksel metodlar
- Histogram, Scatter grafikleri
- Outlier, Box-Violin Plot
- ZScore
- Missing Data
- Scaling
- Encoding
- Feature Construction, Feature Selection

Sunum + Notebook + Homework → Github

Dataset

Tabular Data: Row ve column'lardan oluşan iki boyutlu tablo benzeri verilerdir.

Column: Sütun, genelde her bir sütun bir özelliği temsil eder.

Row: Satır, genelde her bir satır bir adet veriyi temsil eder

id	isim	yaş	gelir (10k $)
1	Ashish	25	22.5
2	Noam	40	30.0
3	Niki	50	41.75

Integer: Tam sayı veriler

Float: Ondalıklı sayılar

Object/Category: Sınıfsal veriler/Sayı olmayan veriler

integer ve float değerlerin sonundaki sayı kaç bit tutabileceğini gösterir. Yüksek bit daha geniş depolama demektir.

int32 → 2^{32} adet değer alabilir.

float64 > float32 (Hassasiyet-depolama)

NaN (Not a Number): Geçersiz, boş, yanlış değerleri ifade eder.

Depolama değer aralığının dışına düşen değerler nan veya +inf , -inf olarak ifade edilebilir

Mean-Median-Min-Max-Mode

Genellikle sayısal ifadeler için hesaplanır.

min →sayısal değeri en küçük olan eleman → min(1,2,3) , min(L)
argmin → Sayısal değeri en küçük olan elemanın bulunduğu index

max →sayısal değeri en büyük olan eleman → max(1,2,3) , max(L)
argmax → Sayısal değeri en büyük olan elemanın bulunduğu index

Mean → Aritmetik ortalama $\frac{\text{Elemanların toplamı}}{\text{Eleman sayısı}}$

Median → azalan veya artan listelerde ortadaki eleman 1,2,3,4,5 → Median = 3

count → Listedeki eleman sayısı
Mode → Bir set içerisinde en fazla tekrar eden değer

Standart Deviation
(Standart Sapma)

Bir veri setindeki verilerin ortalamadan ne kadar uzaklaştığını sayısal olarak ifade eder.

x	$(x\text{-mean})^2$
10	225
20	25
30	25
40	225

mean = 25

std = $(500/4)^{1/2}$ = 11.18

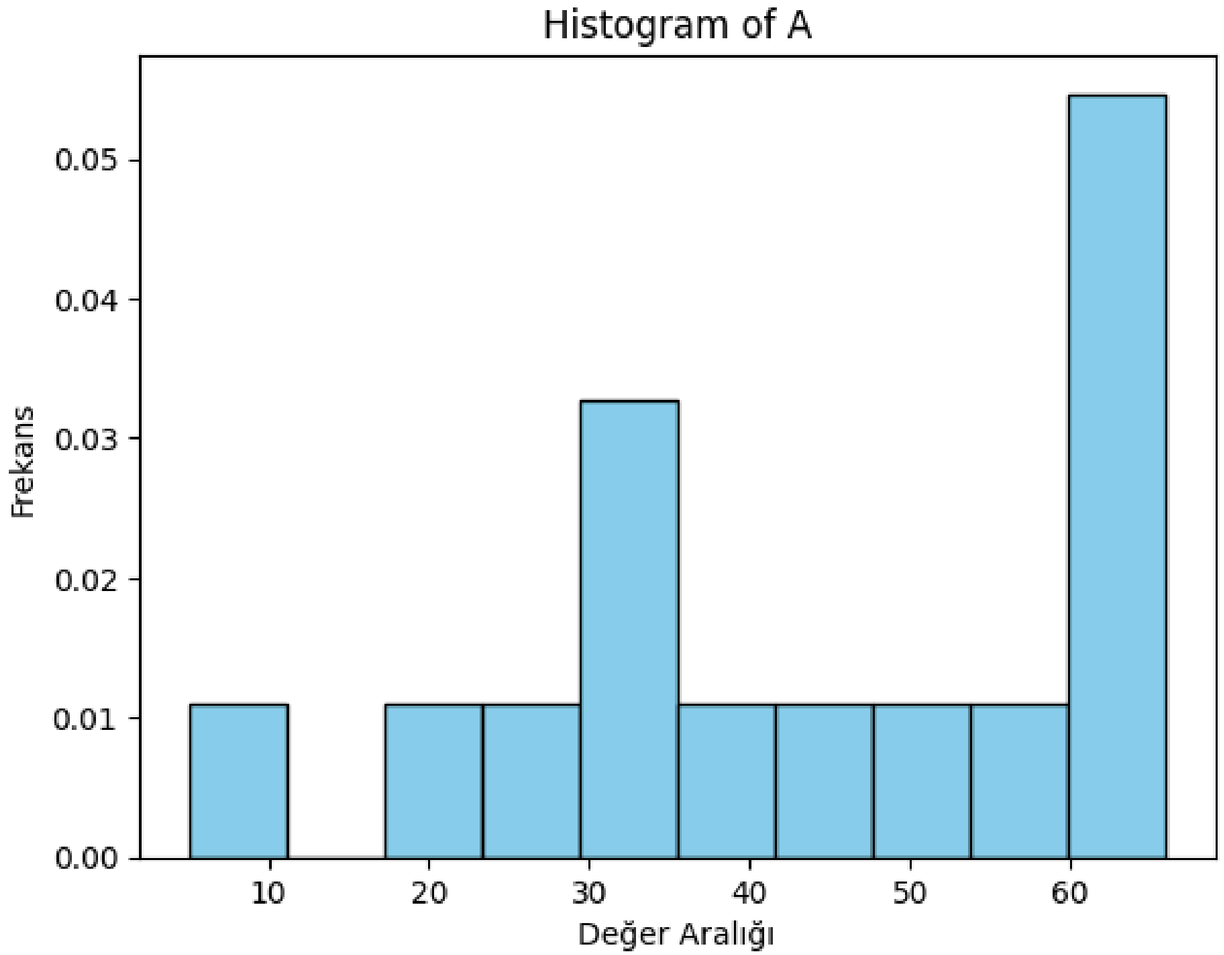
$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Variance (Varyans) : Standart sapmanın karesi, yayılım ölçüsüdür.

Histogram

continuous verilerin görselleştirilmesini sağlayan bir yöntem

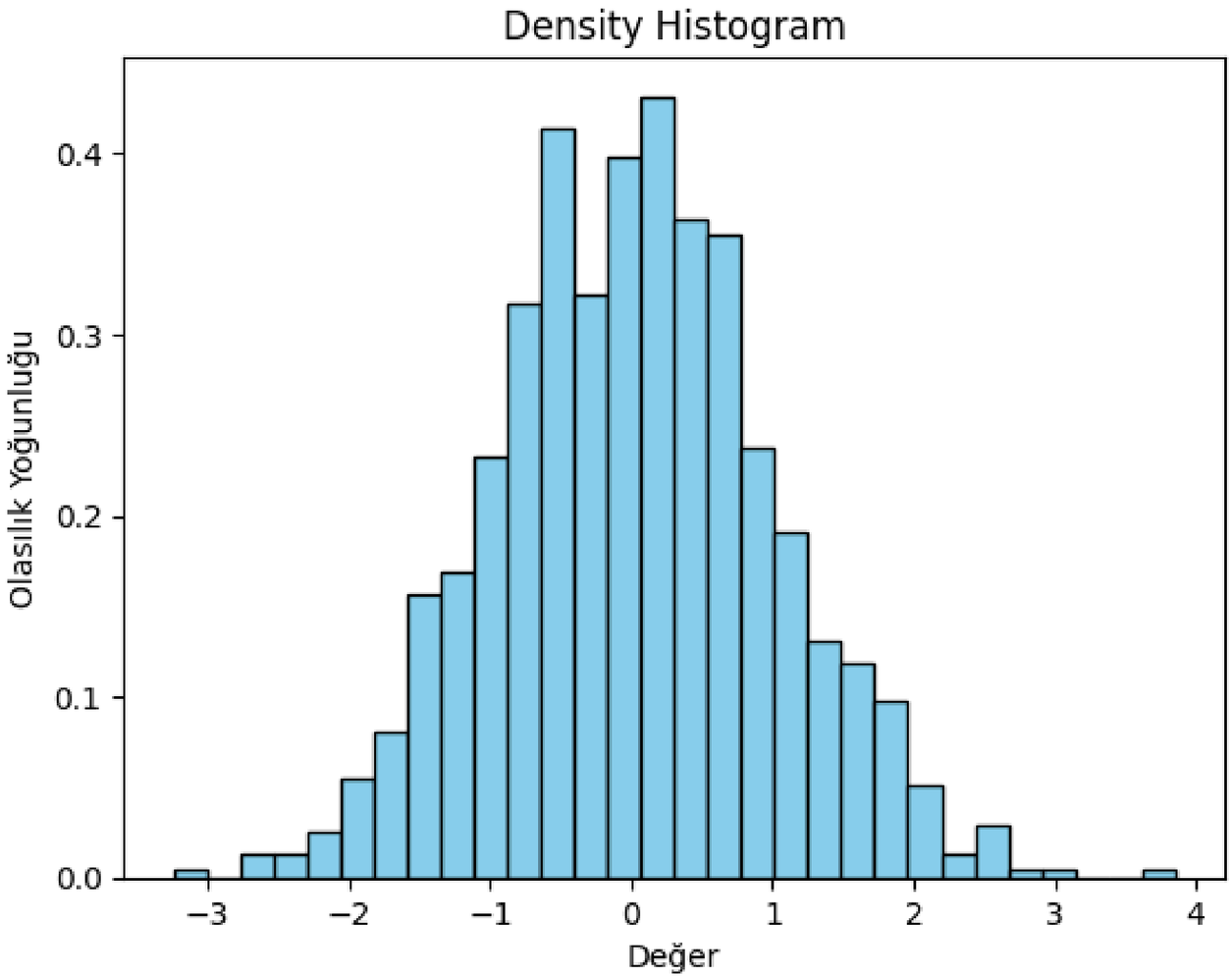
Veriler için belirli aralıklar (bins) belirlenir ve her aralığa kaç tane veri düştüğü sayılır.



'A': [5,20,25,30,32,35,37,45,50,55,60,62,64,64,66]

- **Verinin simetrik olup olmadığı**
- **Uç değerler**
- **Yoğun bölgeler**

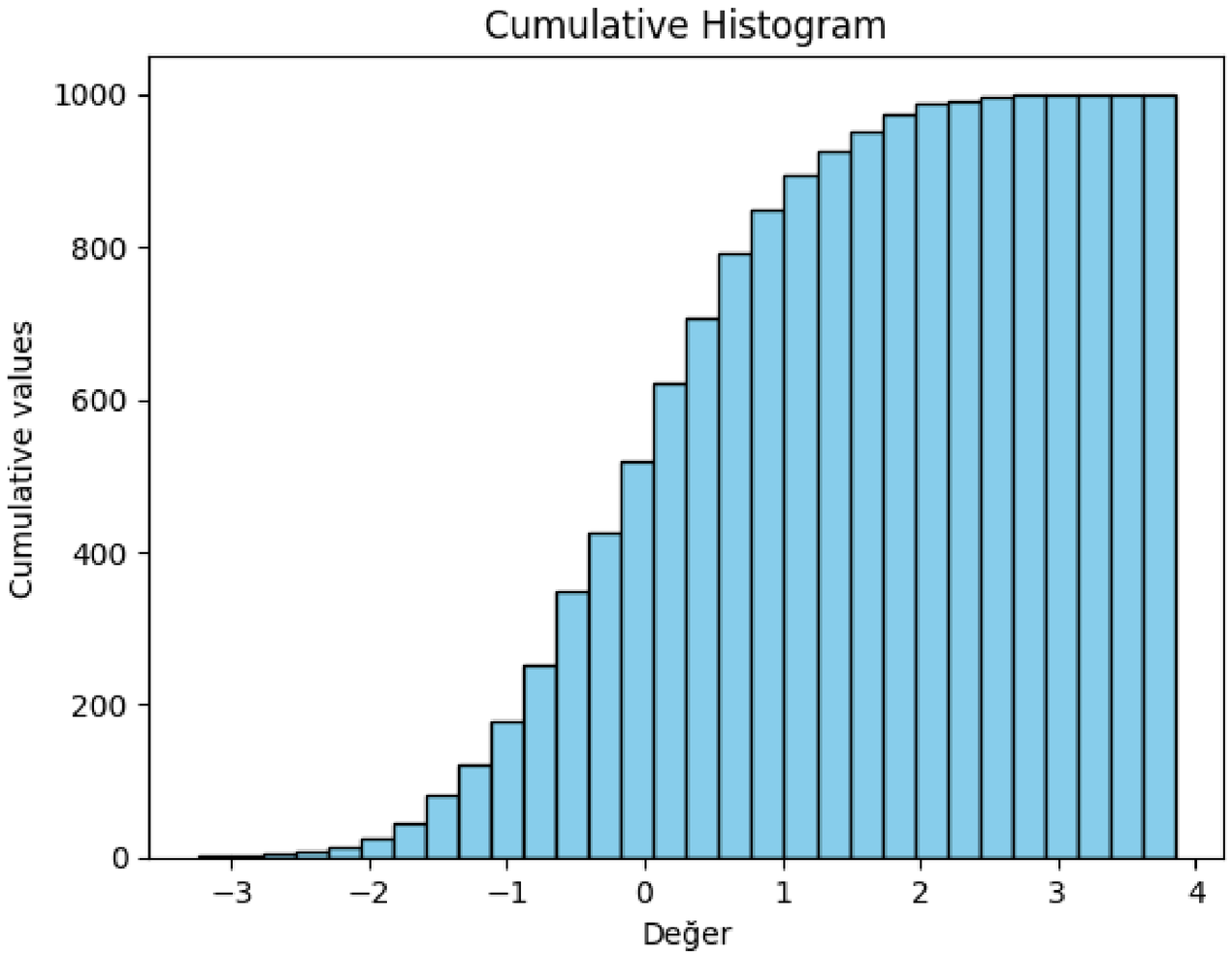
Histogram Types



Density Histogram

y ekseni frekans yerine olasılık

density = True

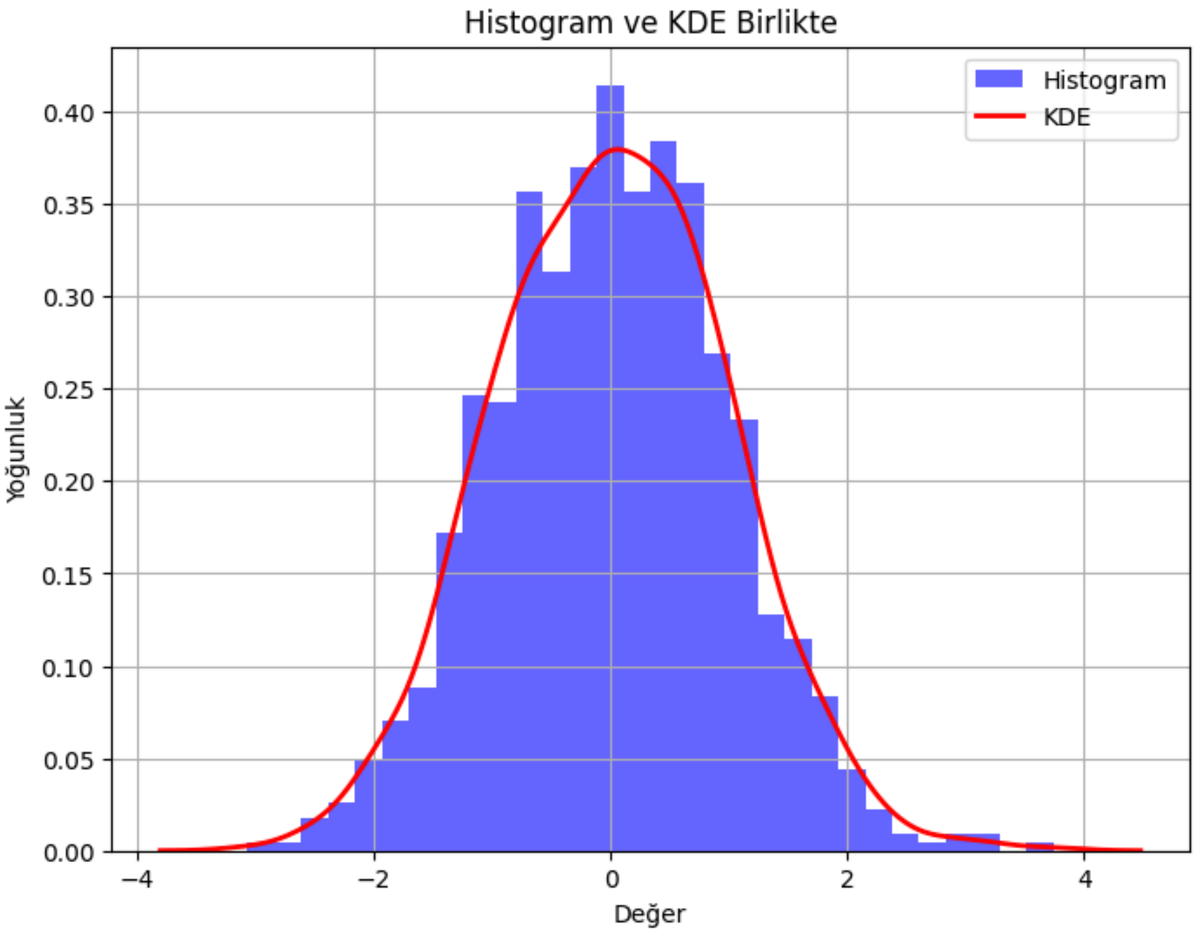


Cumulative Histogram

önceki bin değerlerini toplayarak ilerler

cumulative = True

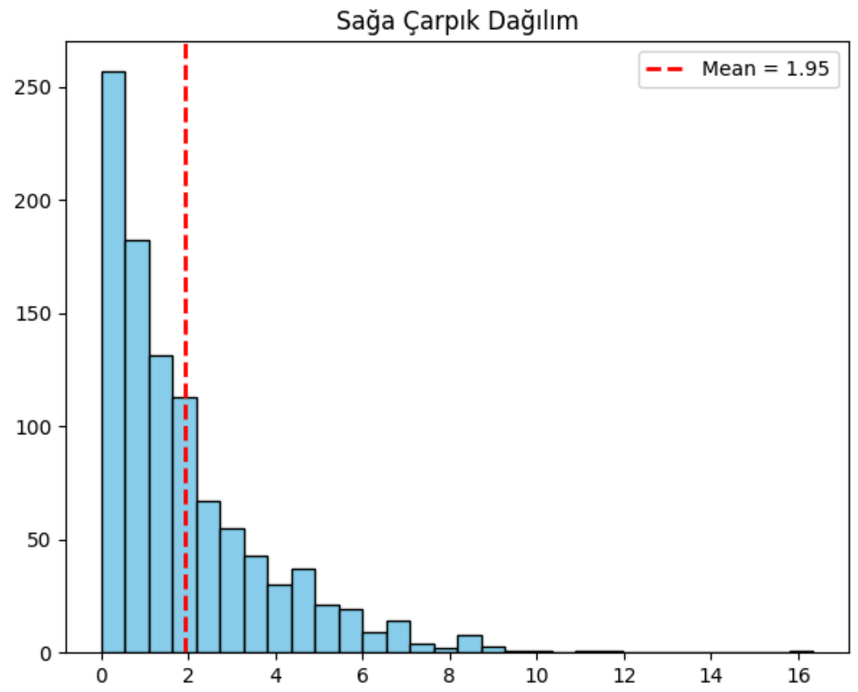
Histogram-KDE

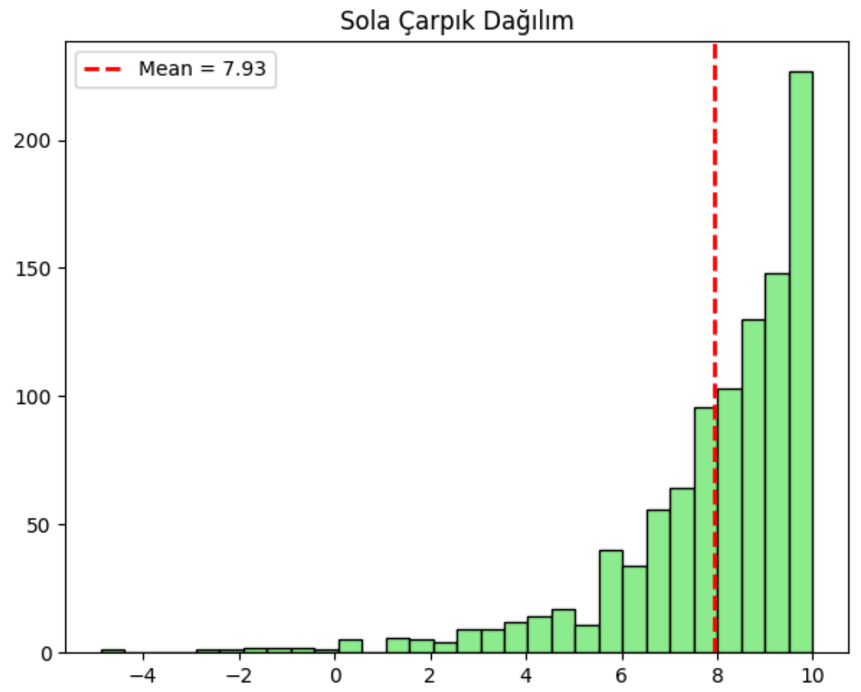


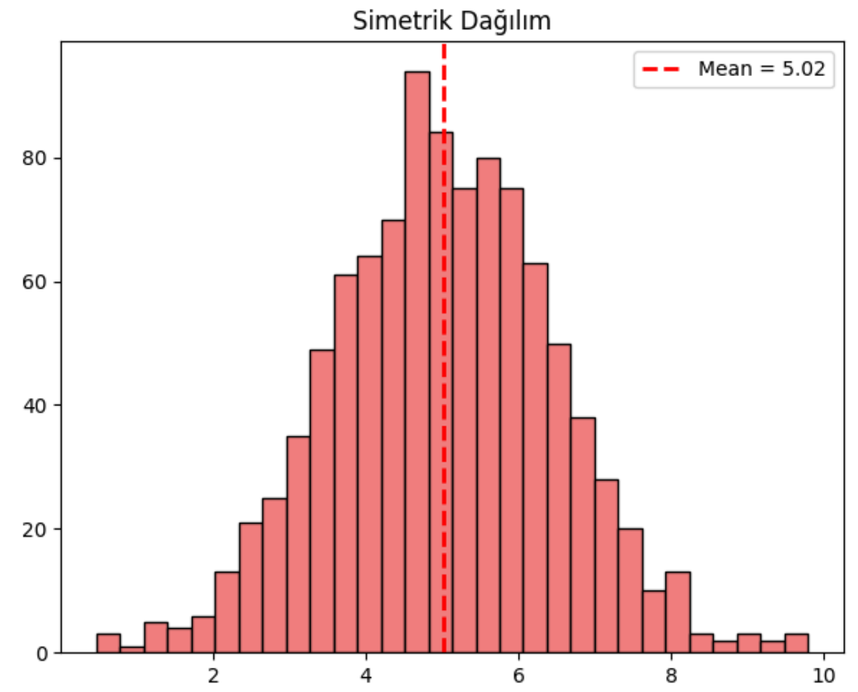
Kernel Density Estimation: Sayısal veri noktalarının altında yatan olasılık yoğunluk fonksiyonunu tahmin etmek için kullanılır.

- **Genellikle histogram plot ile beraber kullanılır.**
- **Histograma göre daha smooth (pürüzsüz) olan çizgi kullanılır.**

Skewness





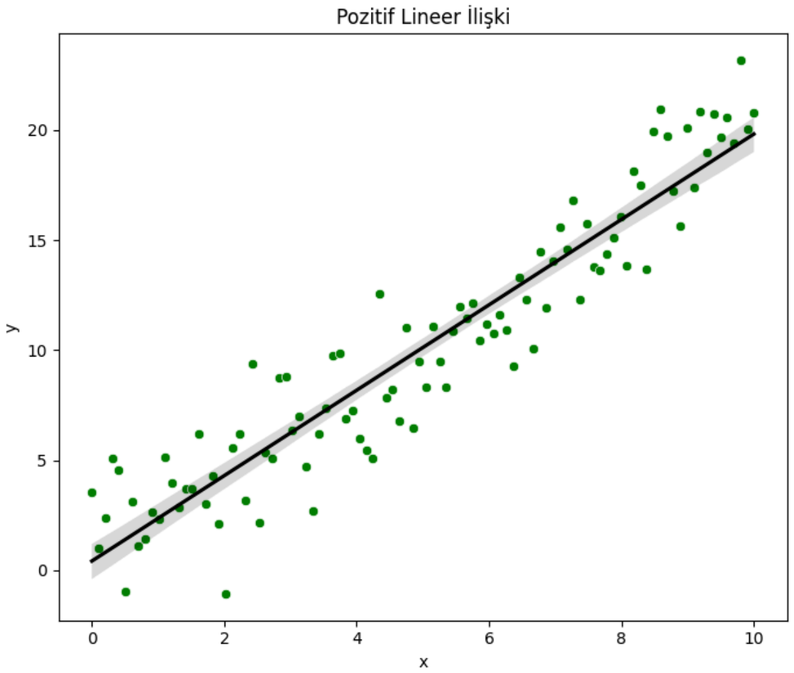


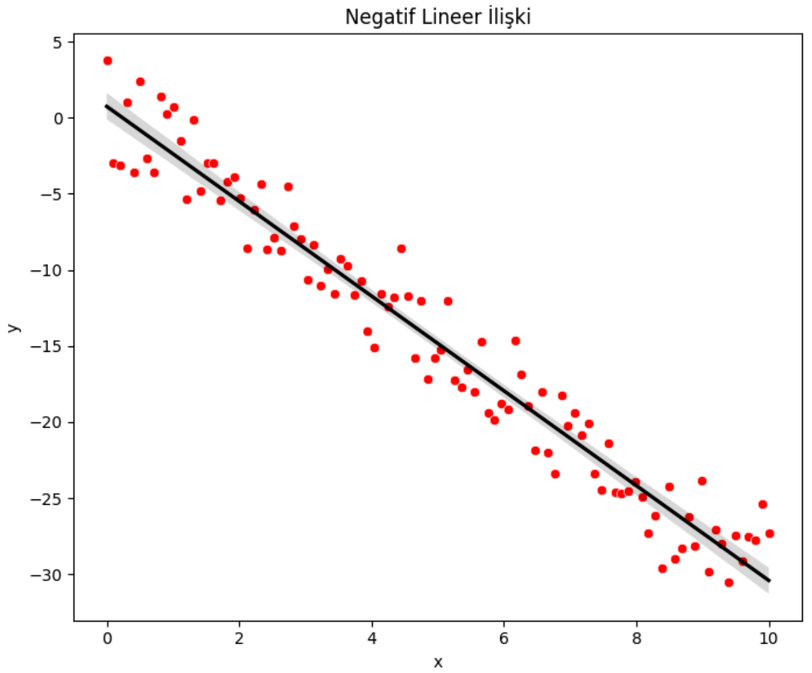
$$\text{skewness} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})}{(N-1)s^3}$$

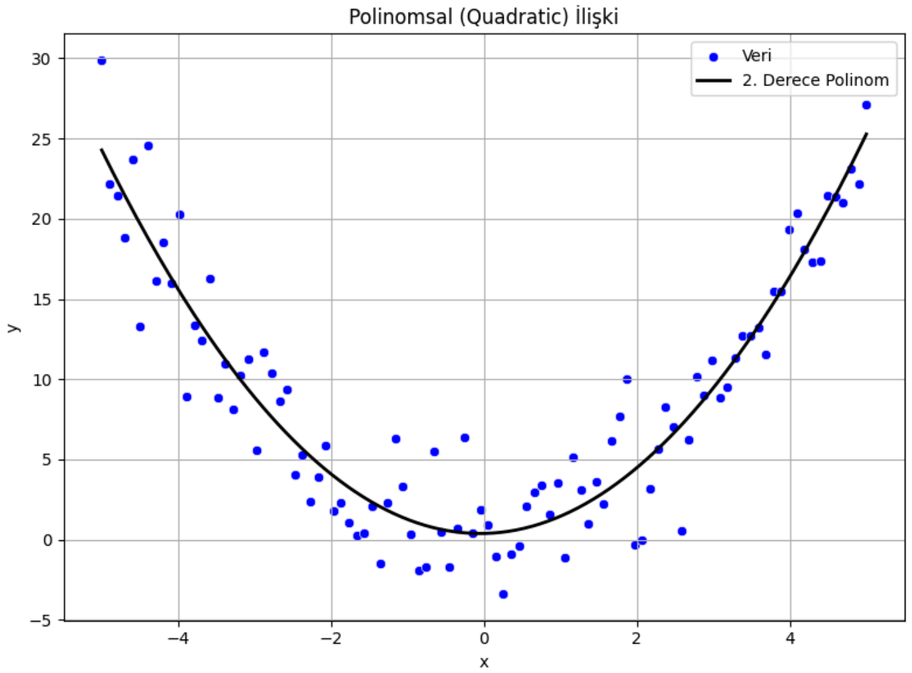
s : standart deviation
$\bar{x}$: mean of the distribution
N : number of observations of the sample

Scatter Plot - continuous

İki değişken arasındaki ilişkiyi incelemek için kullanılır





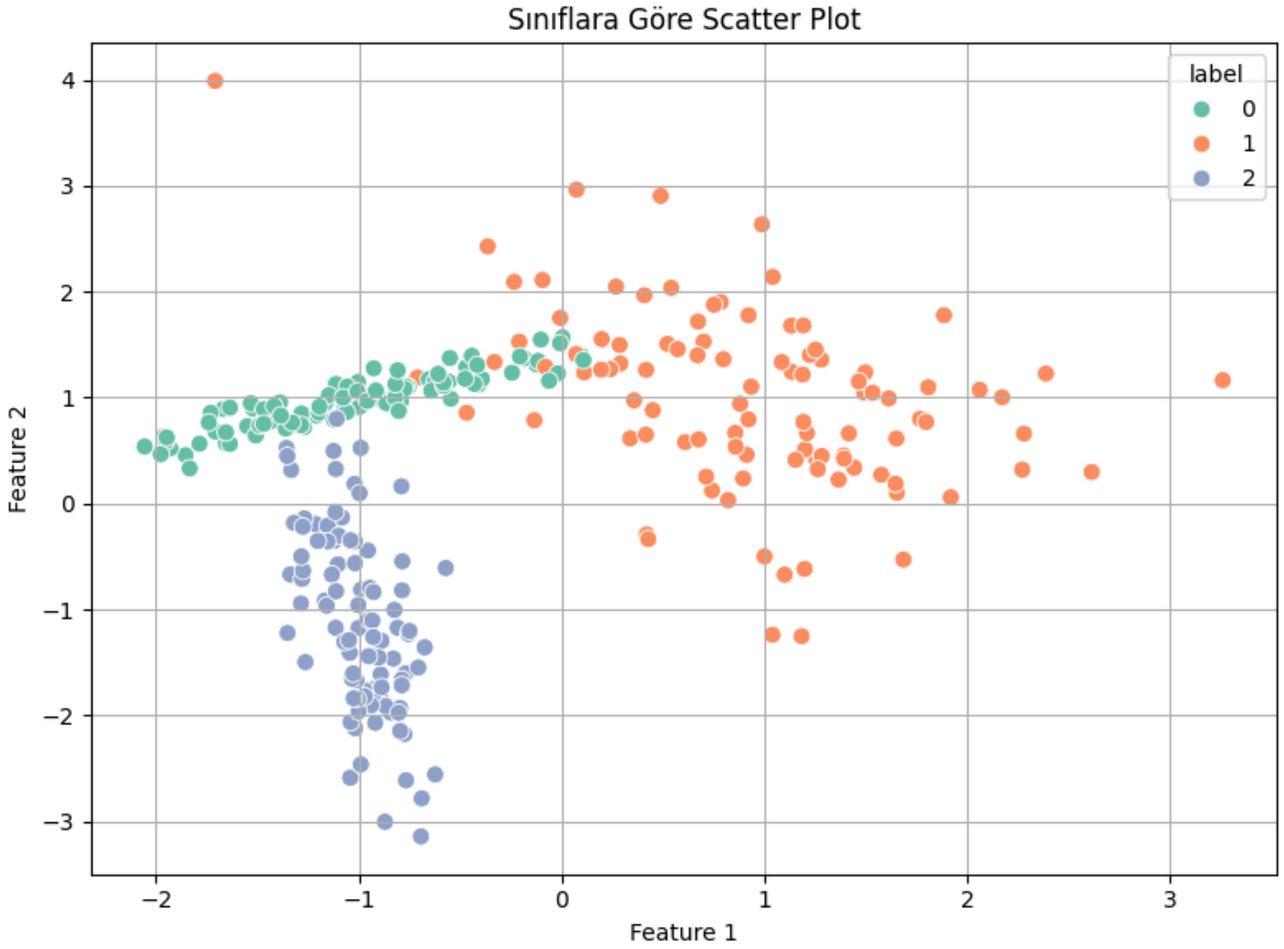


Her bir nokta bir veriyi ifade eder. Verilerin sahip oldukları iki feature kullanılarak bu grafikler çizilir

Bu grafikler kullanılarak

- **Korelasyon ve ilişki**
- **Veri yoğunluğu**
- **Outlier**

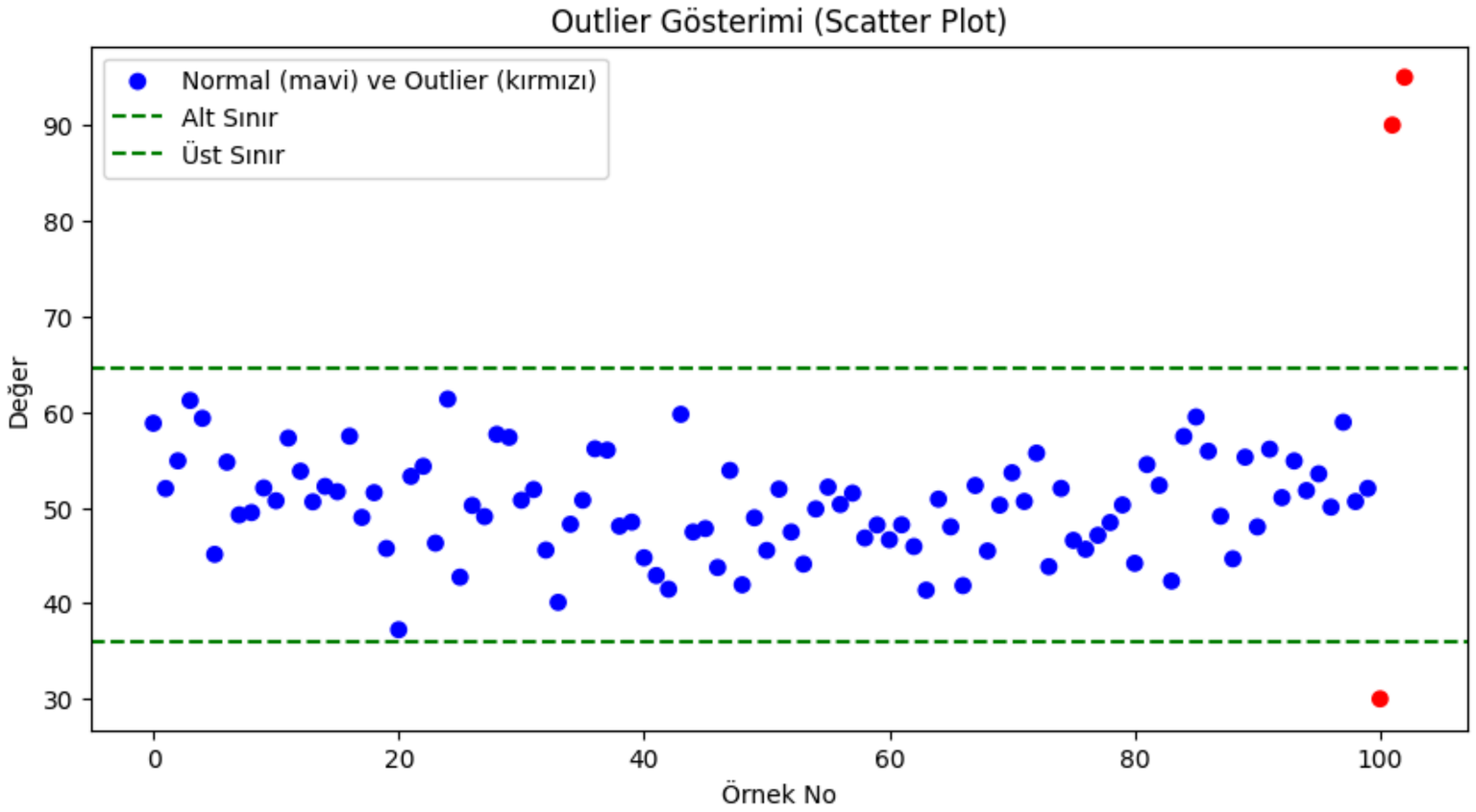
Scatter Plot - Class



Hedef sınıfların kümelenme durumu bu grafikler ile incelenebilir

Outlier

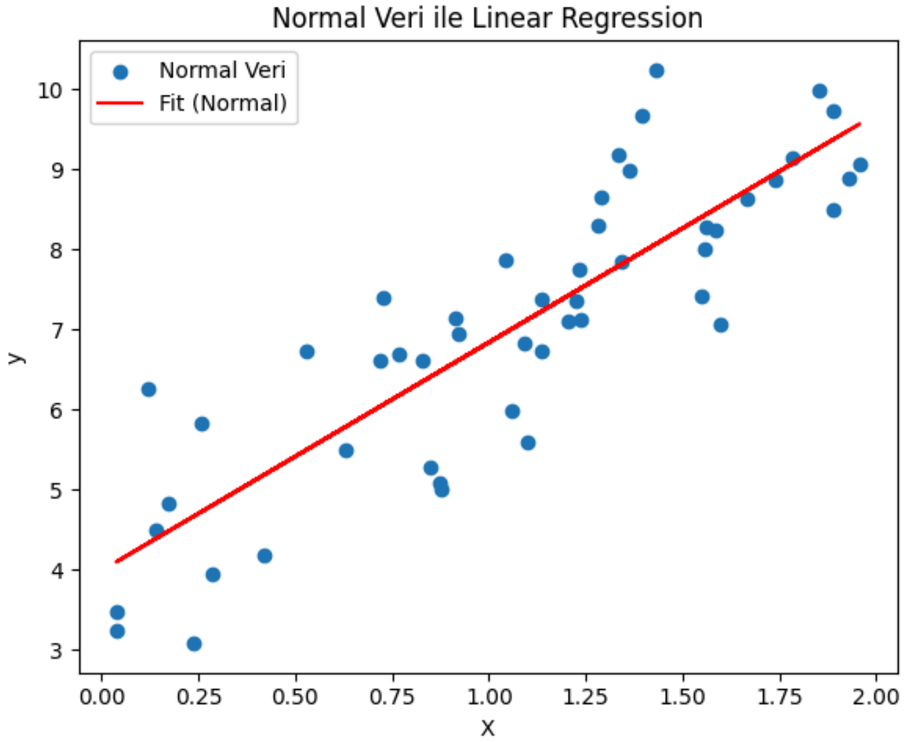
Outlier: Diğer verilere göre uçta kalan aykırı verilerdir. Normal dağılımın dışına düşerler

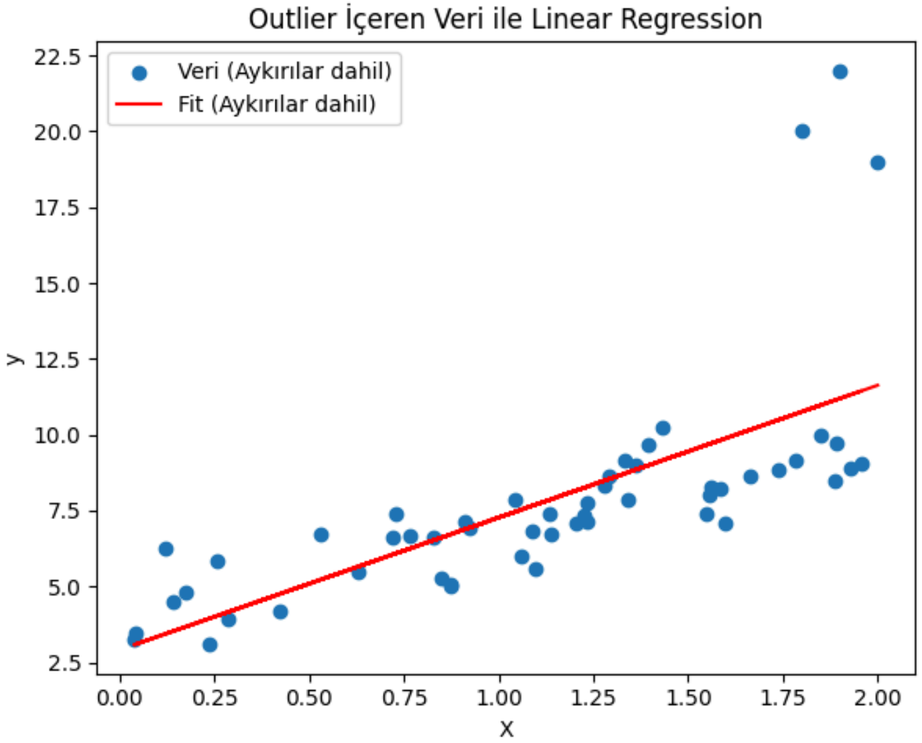


Outlier değerler

- Ölçüm hatalarından
- Veri ön işlemedeki hatalardan
- Bazen de doğal olarak

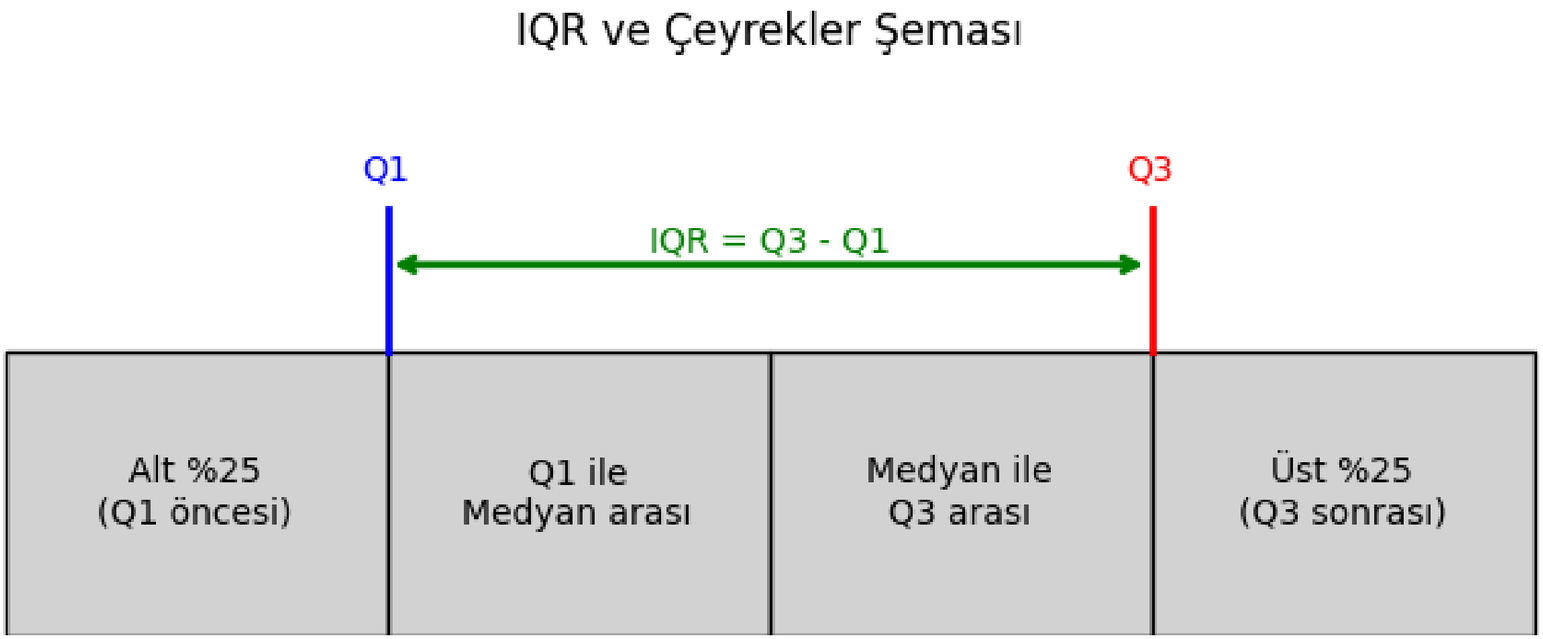
Outlier Etkileri





- Linear Regression, logistic regression gibi algoritmaların yanlış öğrenmesine sebep olabilirler. Uç değerleri öğrenmeye çalışırken verilerin yoğun olduğu yerden sapabilir. Bir çeşit overfitting oluşur
- Decision Tree modelleri outlier'lara karşı daha dayanıklı olsa da hedef (target) değişkende outlier değerler modellerin kötü sonuç üretmesine neden olur.
- Outlier değerler her zaman veri setinden silinmek zorunda olmayabilir.

Interquantile Range (IQR)



Veri seti küçükten büyüğe sıralanınca
First Quantile (Q1) = %25. nokta
Third Quantile (Q3) = %75. nokta

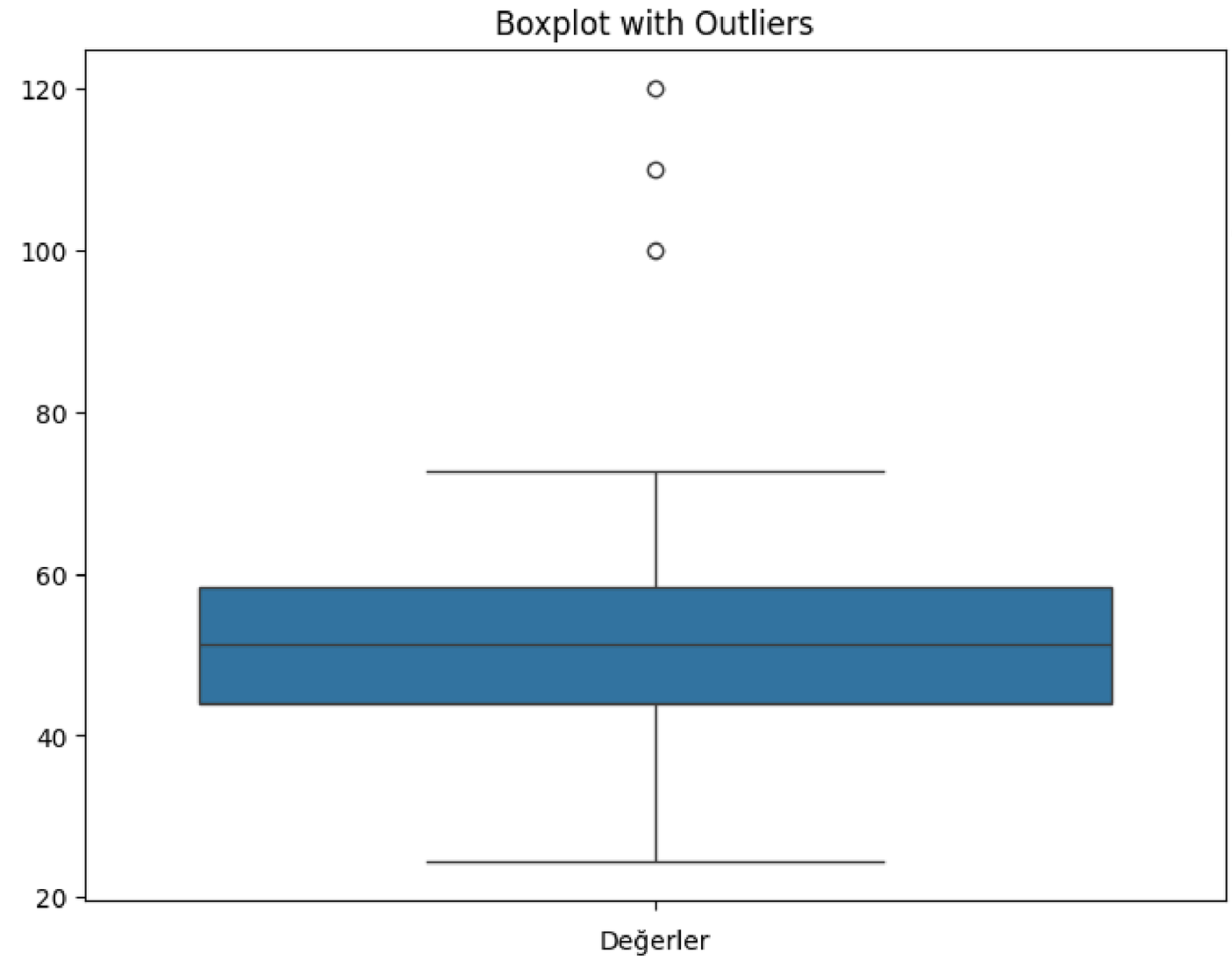
Alt sınır = Q1 - 1.5 * IQR
Üst Sınır= Q3 + 1.5 * IQR

Bu sınırlar dışında kalan değerler outlier olarak kabul edilebilir

Q1 = 50
Q3 = 150 → **IQR = 100** → **Alt Sınır = -100**
Üst Sınır = 300 → **-100'den küçük ve 300'den büyük değerler outlier**

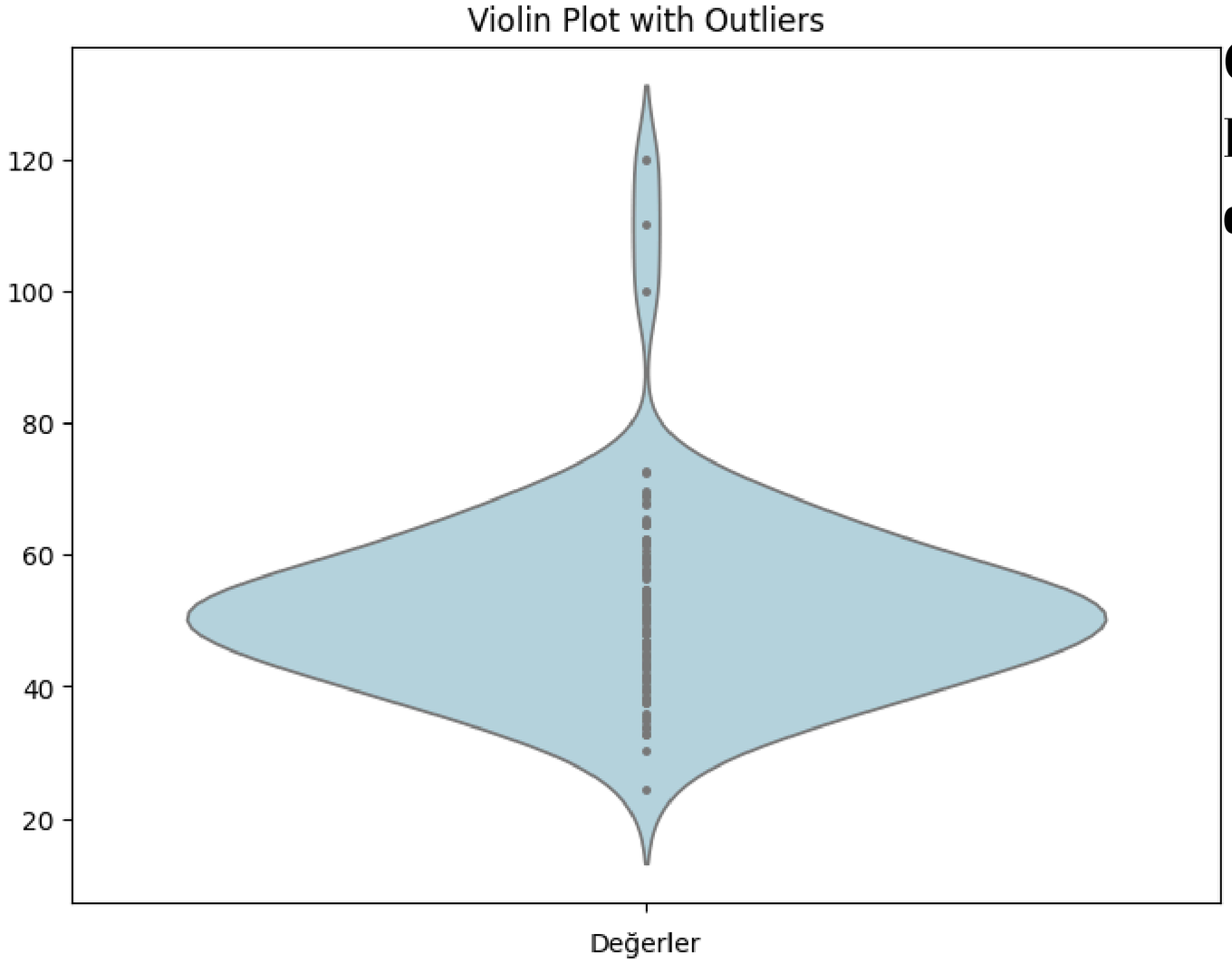
***1.5 yerine 3 gibi daha büyük çarpanlar kullanılması sadece daha aykırı verilerin seçilmesini sağlar**

Box Plot



Outlier görselleştirmek için kullanılır. Whisker dışında kalan noktalar outlier'ları ifade eder

Violin Plot



Outlier'ları görselleştirmek için kullanılır. Box plot ile farkı, Violin Plot veri dağılımının yoğunluğunu da gösterir.

*IQR Hesaplama Yöntemleri

10
12
13
15
18
19
25
100

Tukey's Hinge

median = 16.5

üst kısım → [18,19,25,100] —median→ Q3 = 22

alt kısım → [10,12,13,15] → Q1 = 12.5

Interpolation

Q1 index→ %25 * (n+1) = 2.25 → 12 + 0.25*(13-12) = 12.25

Q3 index→ %75 * (n+1) = 6.75 → 19 + 0.75*(25-19) = 23.5

Z-score

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

x : data
μ : mean
σ : standart deviation

Bir verinin, ortalamaya göre ne kadar standart sapma uzaklıkta olduğunu gösteren istatistiksel bir ölçüdür.

20
25
30
18
50

Mean = 28.6
std = 11.48

—— x=50 için ⟶ **(50 - 28.6) / 11.48 = 1.86**

Eğer
abs(z) > 3 ise outlier

Modified Z-score

$$MAD = \text{median}(|X_i - \text{median}(X)|)$$

MAD : Median Absolute Deviation

$$\text{Modified } z = 0.6745 \cdot \frac{x - \text{medyan}}{MAD}$$

Küçük veri setleri için daha uygundur.

20
25
30
18
50

Medyan = 25
MAD = 5

—— x=50 için ⟶ **0.6745 * (50 - 25) / 5= 3.37**

50 → outlier

Handling Outliers

Outlier ile karşılaşınca

- Outlier olan satırları sil
- Outlier değerleri sınır değerler ile değiştir.
- Kendi haline bırak

Missing Data

id	isim	yaş	gelir (10k $)
1		25	22.5
2	Noam		
3	Niki	50	41.75

- Bir satırdaki bir çok veri boş ise satır silinebilir. (Drop Row)
- Bir sütundaki bir çok veri boş ise sütun silinebilir (Drop Column).
- Boş sayısal veriler mean ile doldurulabilir (Filling with mean).
- Boş sayısal veriler median ile doldurulabilir(Filling with median).
- Boş sayısal ve kategorik veriler mode ile doldurulabilir. (Filling with mode)
- Boş sayısal ve kategorik veriler sabit bir sayı veya class ile doldurulabilir.(Filling with constant)

Group By

Meslek	Maaş
Mühendis	
Doktor	10000
Mühendis	11000
Öğretmen	12000
Doktor	
Öğretmen	
Mühendis	13000

Group By Meslek

Mühendis	
Mühendis	11000
Mühendis	13000

Boş maaş = 12000

Doktor	10000
Doktor	

Boş maaş = 10000

Öğretmen	12000
Öğretmen	

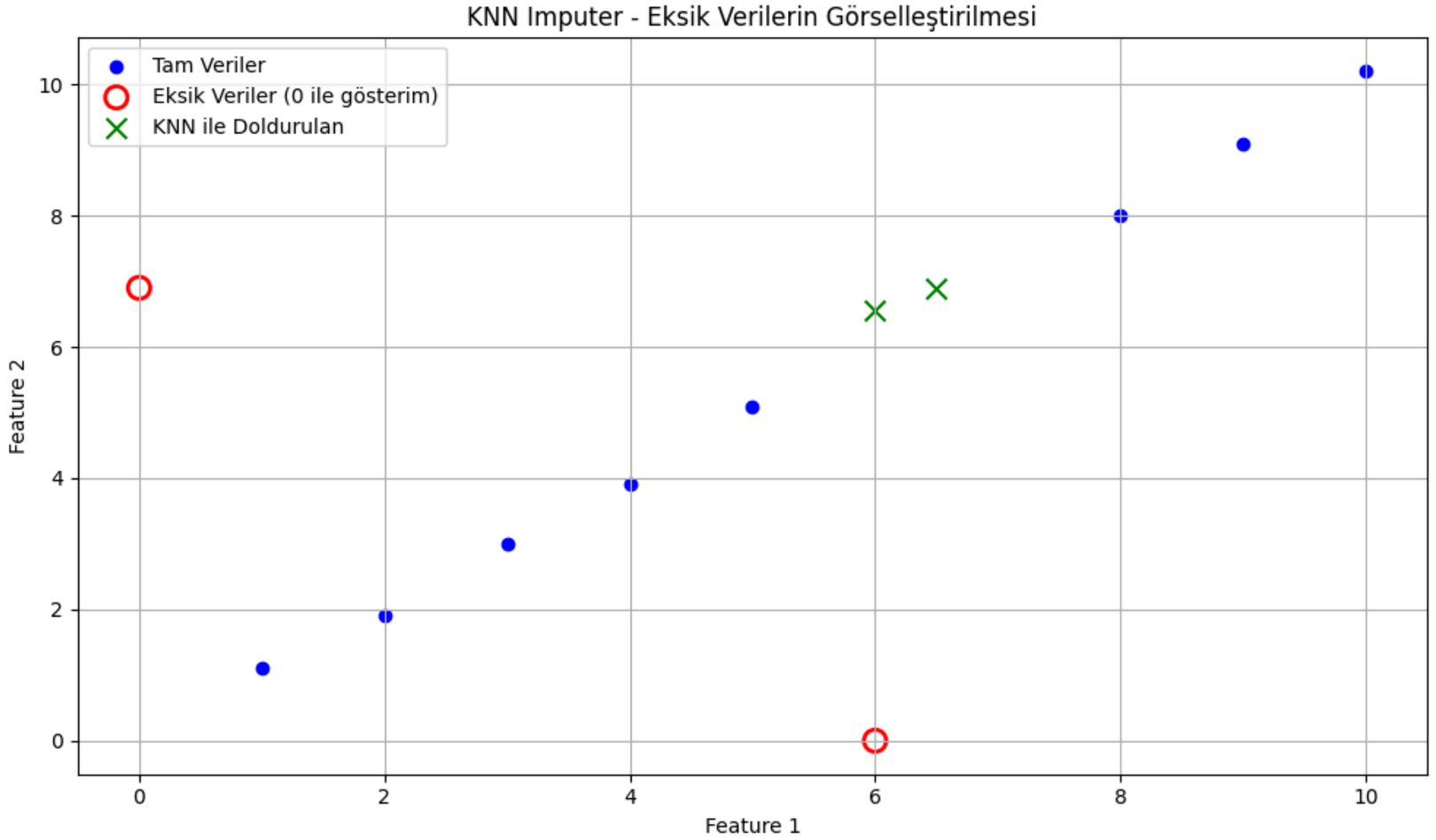
Boş maaş = 12000

Belirli bir(kaç) sütuna göre gruplama yapıldıktan sonra her boş veri ait olduğu grubun

- mean
- mode
- median

değerleri ile doldurulabilir

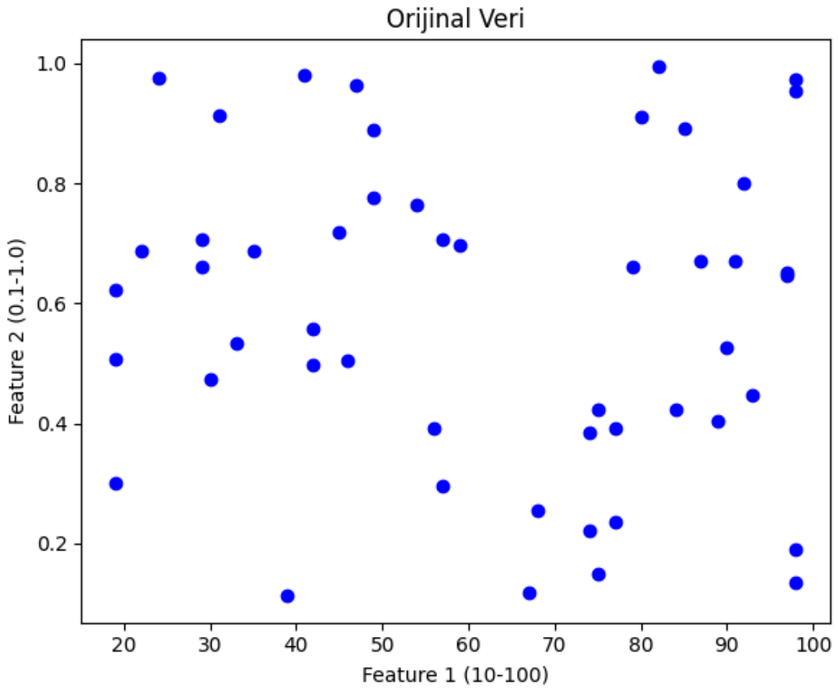
KNN Imputer

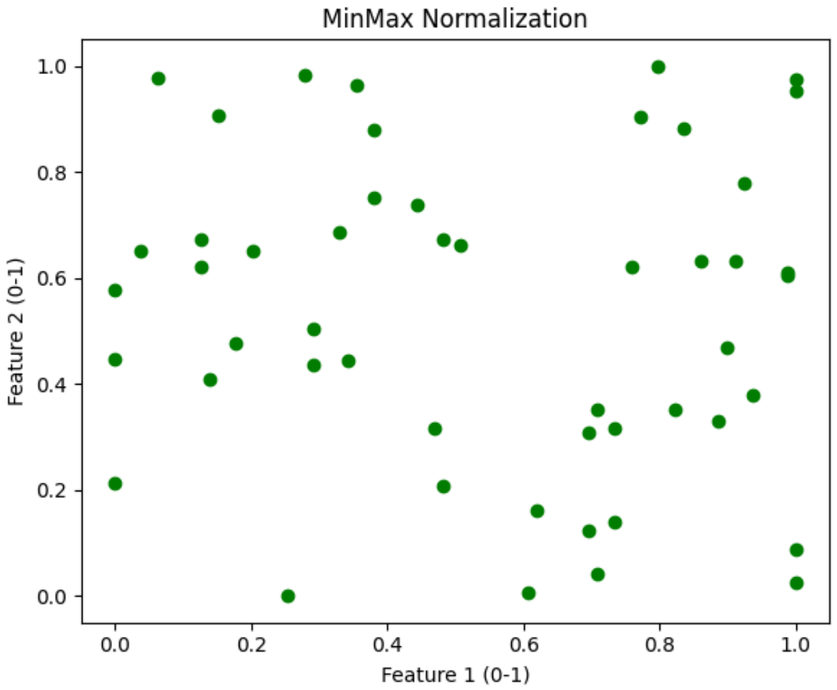


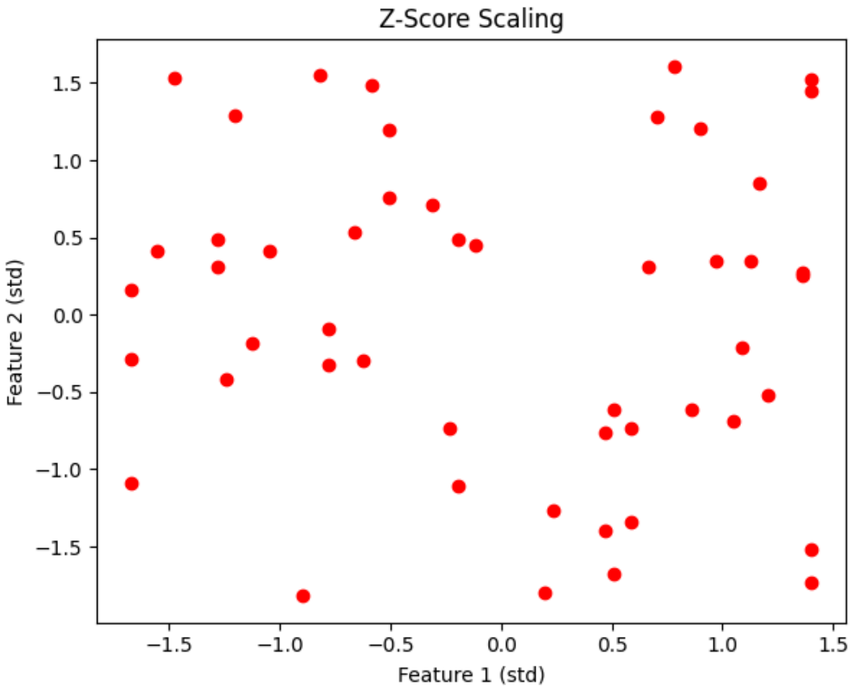
İçerisinde missing data olan satır için

- missing olmayan verilere bakarak diğer veriler ile öklid mesafesini ölçer
- mesafelere göre en yakın k komşuyu seçer
- bu komşuların (ağırlıklı) ortalamalarını kullanarak boş kısmı doldurur

Feature Scale







!!Eksen değer aralıkları farklı

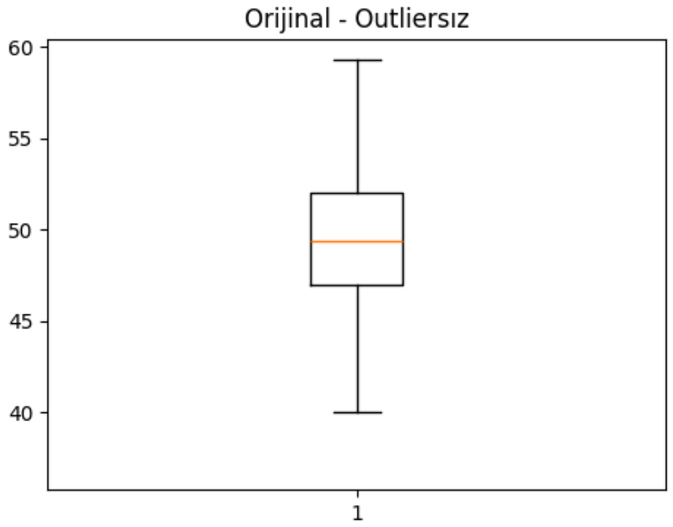
- Algoritmanın öğrenmek için kullandığı özelliklerin farklı aralıklarda (0-1 ve 0-100) olması algoritmanın öğrenmesini geciktirir ve yanlış öğrenmesine sebep olabilir.
- KNN,SVM,Linear ve Logistic Regression ve NN aralık farklılıklarından etkilenebilir.
- Normalization ve Standartization veri aralıklarını aynı ya da benzer yapmayı amaçlar.

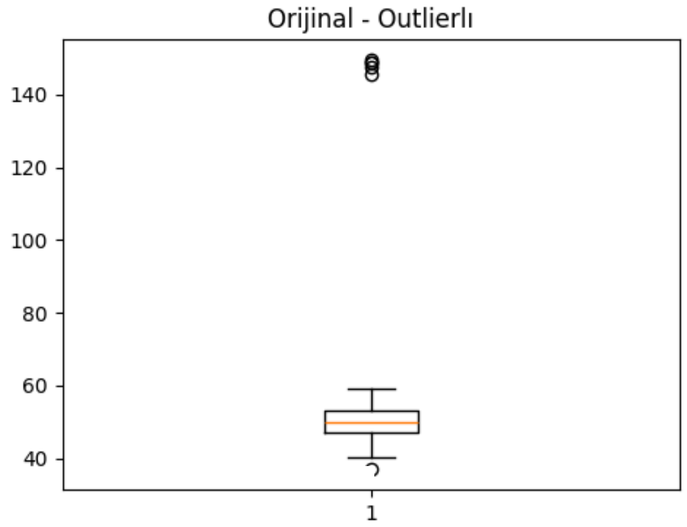
Feature Scale

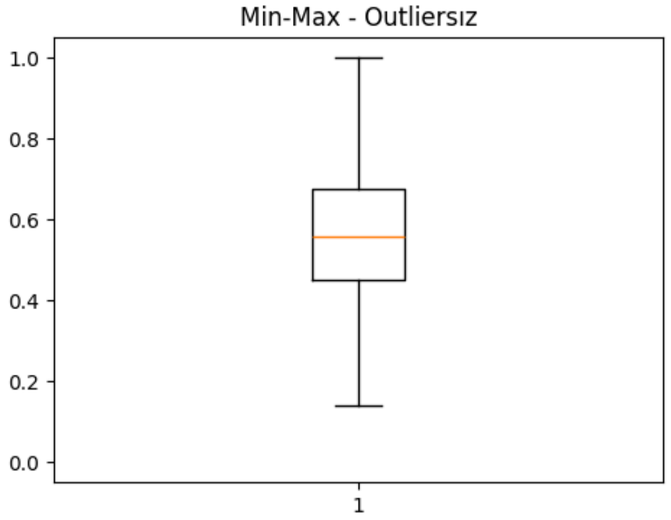
Normalization	Standardization
Sabit bir aralığa koyar (0-1 gibi)	Mean değerini 0 std değerini 1 yapmayı amaçlar. Aralığı sabit değildir.
Outlier'lar etkiler (genellikle)	Outlier'lar son std ve mean etkilemez Outlier'lar veri dağılımını etkiler
Ölçekleri farklı ve min-max önemli olan verilerde kullanılır	Veri dağılımı "Gauss Distribution"a benzerse

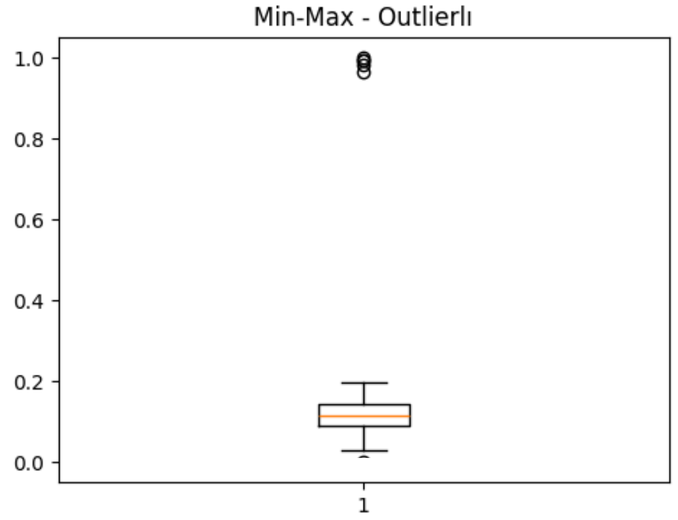
Literatürde iki terim de birbirini yerine çokça kullanılır. Kesin net bir ayrım yok. İkisi de Scaling türü diyebiliriz.

Scaling Nasıl görünüyor

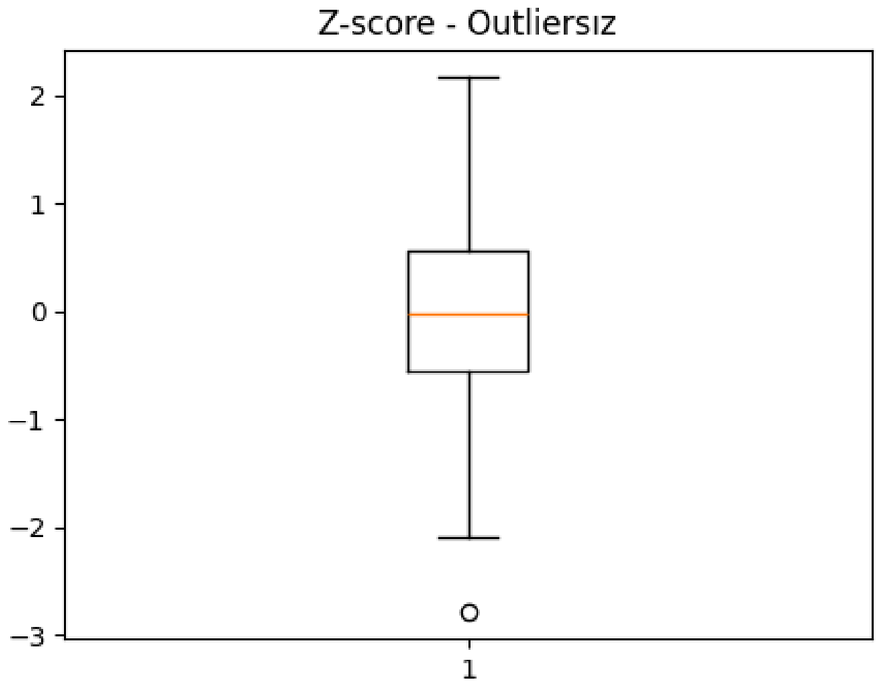


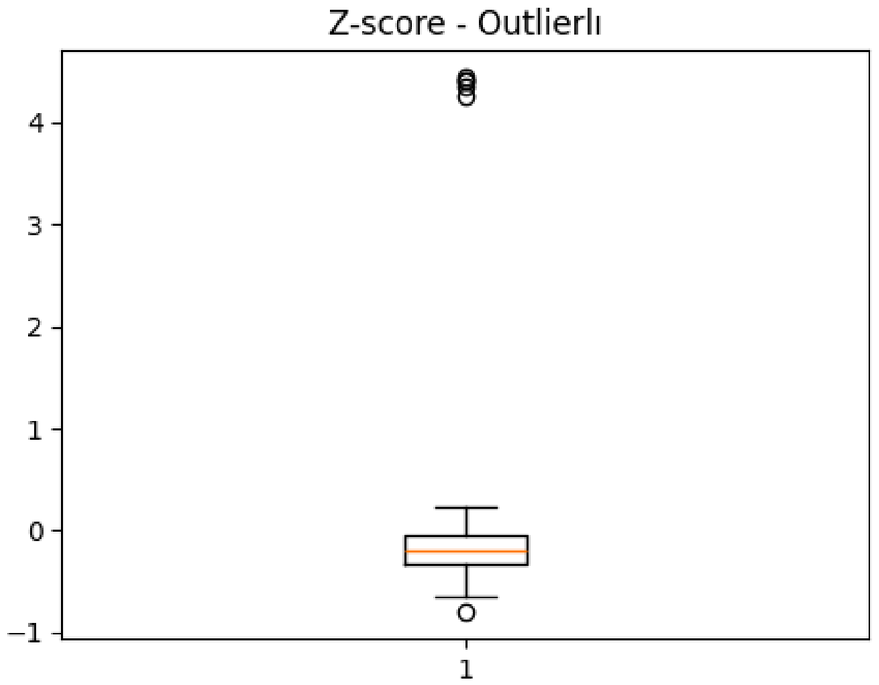






Scaling Nasıl görünüyor





Min-Max Scaling

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Normalization olarak da adlandırılabilir.
Veri dağılımını 0-1 arasına çeker.
Outlier'lara karşı duyarlıdır.

Before Scaling	After Scaling
0	0
20	0.2
40	0.4
60	0.6
80	0.8
100	1.0

$X_{max} - X_{min} = 100-0 = 100$

For 0

(0-0)/100 = 0

For 60

(60-0)/100 = 0.6

For 100

(100-0)/100 = 1.0

MaxAbs Scaling

Before	After
-10	-0.2
0	0
10	0.2
50	1

$|X_{max}| = 50$

$$X_{scaled} = \frac{X}{max(|X|)}$$

[-1,1] aralığına çeker.
negatif değerleri korur.

Z-score Scaling

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

x : data
μ : mean
σ : standart deviation

Standardization olarak da adlandırılır.

Outlier'lara karşı duyarlı değildir.

mean=0 , std = 1 yapar.

Gauss dağılımı gibiyse uygulanır.

40	-1.35
50	-0.45
60	0.45
70	1.35

std = 11.1 **std = 1**

mean = 55 **mean = 0**

Robust Scaling

$$X_{scaled} = \frac{X - median}{IQR}$$

Right Skew ya da left skew veri dağılımlarında tercih edilebilir.
Outlier'lara dayanıklıdır.

40	-1.0
50	-0.33
60	0.33
70	1.0

median = 55

IQR = 62.5-47.5 = 15

Encoding

- AI/ML algoritmaları sayısal değerler ile çalışır. Sayısal olmayan değerlerin sayısal değerlere çevrilme çabası vardır.
- Categorical verilerin sayısal değerlere çevrilmesi işlemine “encoding” denir.
- Encoding işleminin feature’lara uygun yapılması önemlidir. Bu yüzden feature sütunları bu işlemden önce incelenmeli bilgi sahibi olunmalıdır.

Unique Value: Bir sütun içerisindeki her bir eşsiz değere denir. Bir sütunda kaç tane unique value olduğu hem veri analizi için hem de encoding işlemi için önemlidir.

- Unique value ile uğraşırken izlenebilecek en iyi yöntem veri setini detaylıca inceleyip her bir unique değerin anlamını bilmektir. Ancak her zaman mümkün değil.

Ankara
İstanbul
İzmir
Ankara

Unique Value = 3

Kırmızı
Yeşil
Mor
Mavi

Unique Value = 4

positive
negative
positive
positive

Unique Value = 2

positive
negative
positie
positive

Unique Value = 3

Yazım hatası gibi durumlar yeni bir unique value gibi görünebilir. Encoding işlemine geçmeden önce bu yanlışlıkların düzeltilmesi iyidir.

Label Encoding

Red	→	0
Green		1
Red		0
Blue		2

- Her bir unique değer için bir sayı atanır. Atanan sayıların mantıksal bir sıralaması yoktur.
- Bu encoding işlemi ile yanlış yorumlanabilen durumlar ortaya çıkabilir. Örneğin 2>1 ancak blue>green ?
- En basit encoding yöntemidir.

Ordinal Encoding

AA	→	2
BA		1
BB		0
AA		2

- Her bir unique değere bir sayı atanır. Atanan sayılar mantıksal bir hiyerarşi oluşturur.
- Sütun içerisinde çok fazla unique değer varsa bu işlemin yükü fazla olabilir. 1000 adet unique değerin her birini mantıksal sıralamaya koymak hem yanıltıcı hem de zor olabilir.

Frequency Encoding

Red	→	2	or	0.4
Green		2		0.4
Green		2		0.4
Blue		1		0.2
Red		2		0.4

- Her bir unique değerin sütun içerisinde kaç kere bulunduğunu baz alır.
- Normalize edilerek ya da edilmeden kullanılabilir.
- *Feature engineering yöntemlerinden biri olarak da sayılabilir çünkü tree based modellerin bu yöntemden faydalanma ihtimali yüksektir.

One-Hot Encoding

Red
Green
Green
Blue
Red

is_red	is_green	is_blue
1	0	0
0	1	0
0	1	0
0	0	1
1	0	1

- Her bir unique value için yeni bir sütun oluşturulur. Bu sütunda unique değere karşılık gelen satırlara 1 geri kalanlara 0 yazılır.
- Her bir unique value için algoritmanın öğrenmesi sağlanır.
- Çok fazla unique value varsa bu işlem maliyetli olur.

Target Encoding

	Target
Cat	1
Dog	0
Rat	0
Dog	1
Cat	1
Rat	0

mean for cat = 2/2 =1

⟶ mean for dog = ½ = 0.5

mean for rat = 0/2 = 0

1
0.5
0
0.5
1
0

- Her bir unique değerin target colum'a bakılarak ortalaması alınır. Bu ortalama değeri unique değerin yerine yazılır.
- Bu yöntem encoding içerisinde target column (y) kullandığı için data leakage içerebilir, overfittinge sebep olabilir.

Ordered Target Encoding
CatBoost Encoder

CatBoost: unbiased boosting with categorical features

Target encoding durumundaki data leakage'ı önlemek için encoding'i sıralı bir şekilde yapmaktadır.

Mantığı garip olsa da çalışıyor!

Feature Construction
Yeni özellik üretimi

Mevcut Özellikleri kullanarak yeni özellikler üretmek.

En * Boy = Alan

BMI = kilo / boy^2

Birden fazla özelliğin toplanması, çarpılması vb.

Metindeki kelime sayısı, frekansı vb.

Time series özellikleri

Verilerin normalize edilmiş, standartlaştırılmış, transform uygulanmış halleri de yeni özellik olarak eklenebilir.

Threshold işlemi (Yaşı 18'den küçük ise 0 büyük ise 1)

Yeni özellik üretim işlemi genellikle domain-specific (Çalışılan alana özgü) bir işlemdir.

Yeni özellik algoritmanın öğrenme gücünü artırabileceği gibi çok fazla özellik eklemek öğrenme sürecini olumsuz etkileyebilir.

Hangi yeni özellikleri üreteceğini bilmiyor musun → LLM'e sor. İşi ne cevaplasın.

Feature Selection

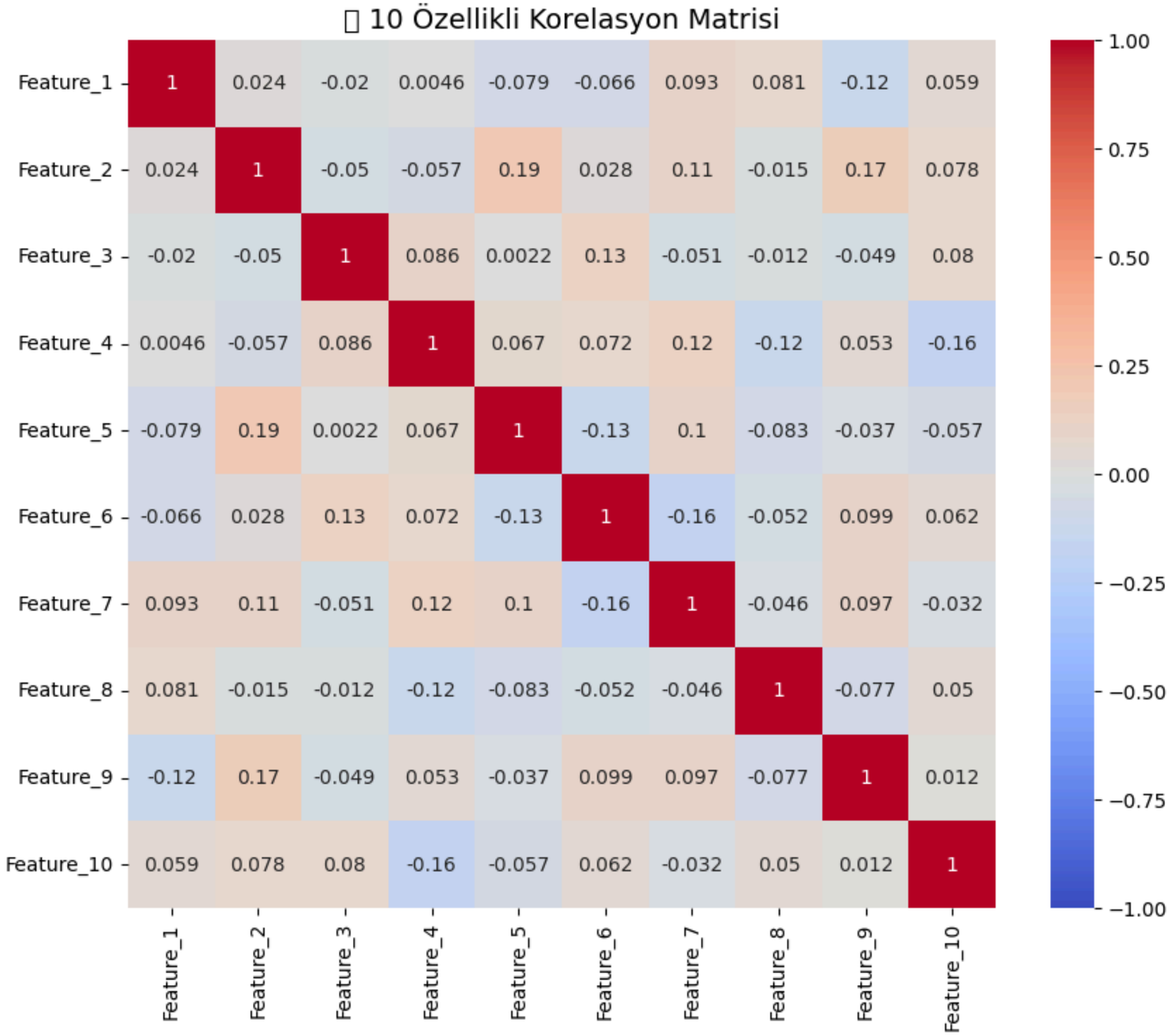
Eldeki feature'ların anlamlı ve önemlilerin seçilmesidir.

- **Çok fazla feature olması overfitting'e sebep olabilir. (Aslında olmayan kuralları uydurup ezberleme)**
- **Çok fazla özellik = fazla eğitim süresi**
- **Daha az özellik = Daha iyi interpretability**

Curse of Dimensionality → Çok fazla boyutun (özelliğin) modelin öğrenme performansına zarar verdiğini öne süren bir kavram.

- **Curse of dimensionality'den kurtulmak için**

Correlation Matrix



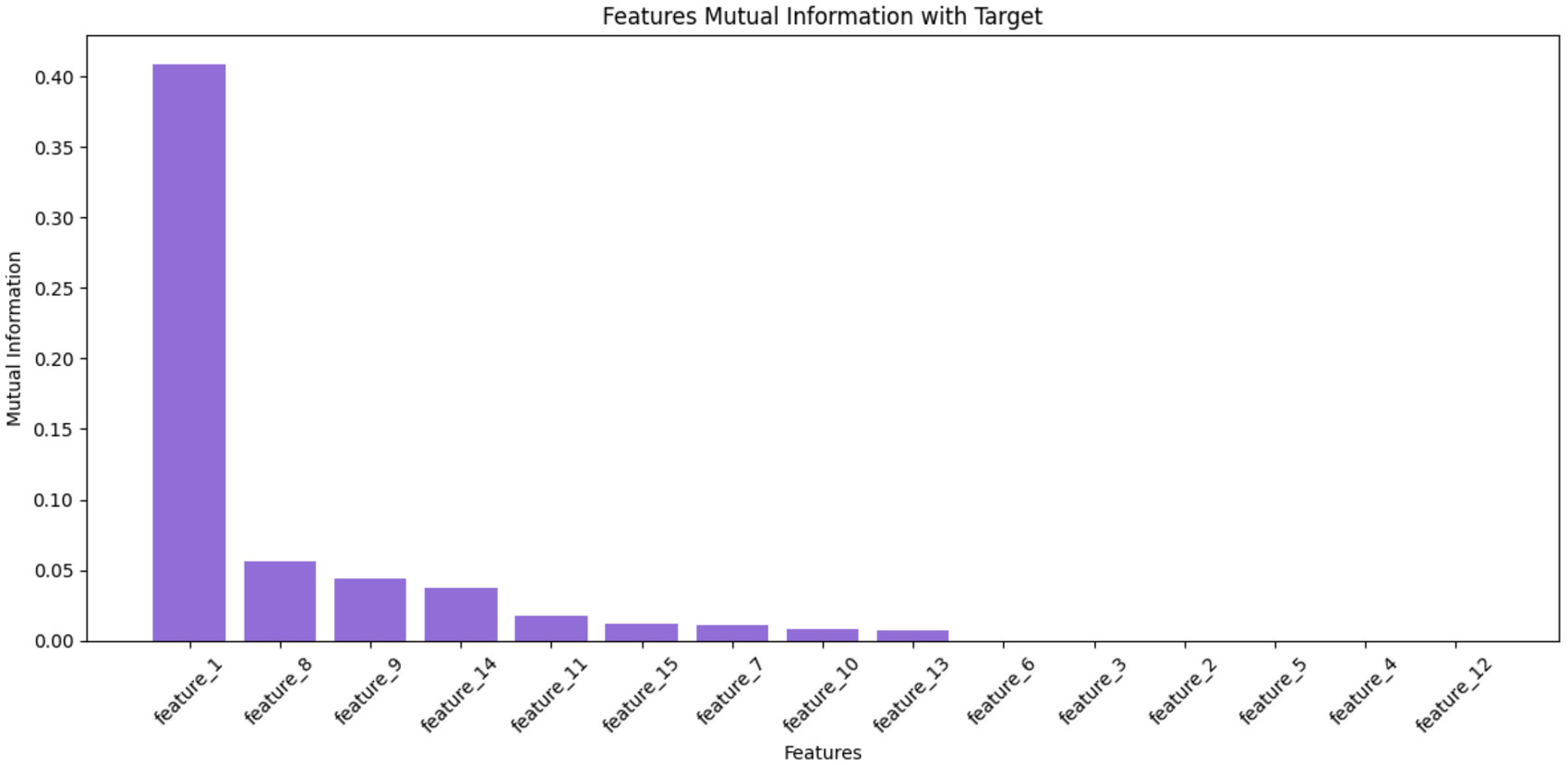
İki sayısal özelliğin birbirine göre nasıl değiştiğini gösterir.

- **-1 → negatif fazla ilişki**
- **0 → İlişki yok**
- **+1 → pozitif fazla ilişki**

Özelliklerin Target Column ile korelasyon göstermesi model için önemli özellikler olduğunu gösterir.

Sadece linear ilişkiyi ölçer. (Pearson)

Mutual Information



- Entropy kullnarak bir özelliğin başka bir özellik ile ilişkisini ölçer.
- Doğrusal olmayan ilişkileri de ölçer.
- *Çok fazla veri varsa hesaplaması uzun sürebilir