Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

VERİ MADENCİLİĞİ (DATA MİNİNG)

Dr. Öğr. Üyesi Alper Talha Karadeniz

2025-2026

Hafta 2

Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

01

Veri Türleri

02

Sürekli ve Kesikli Veriler

03

Veri Kalitesi, Eksik Veriler, Gürültü ve Aykırı Değerler

04

Uygulama

Veri Nedir?

Veri madenciliğinde kullanılan veri kümesi, genellikle veri nesnelerinden (object/instance) ve özelliklerden (feature/attribute) oluşan bir koleksiyondur.

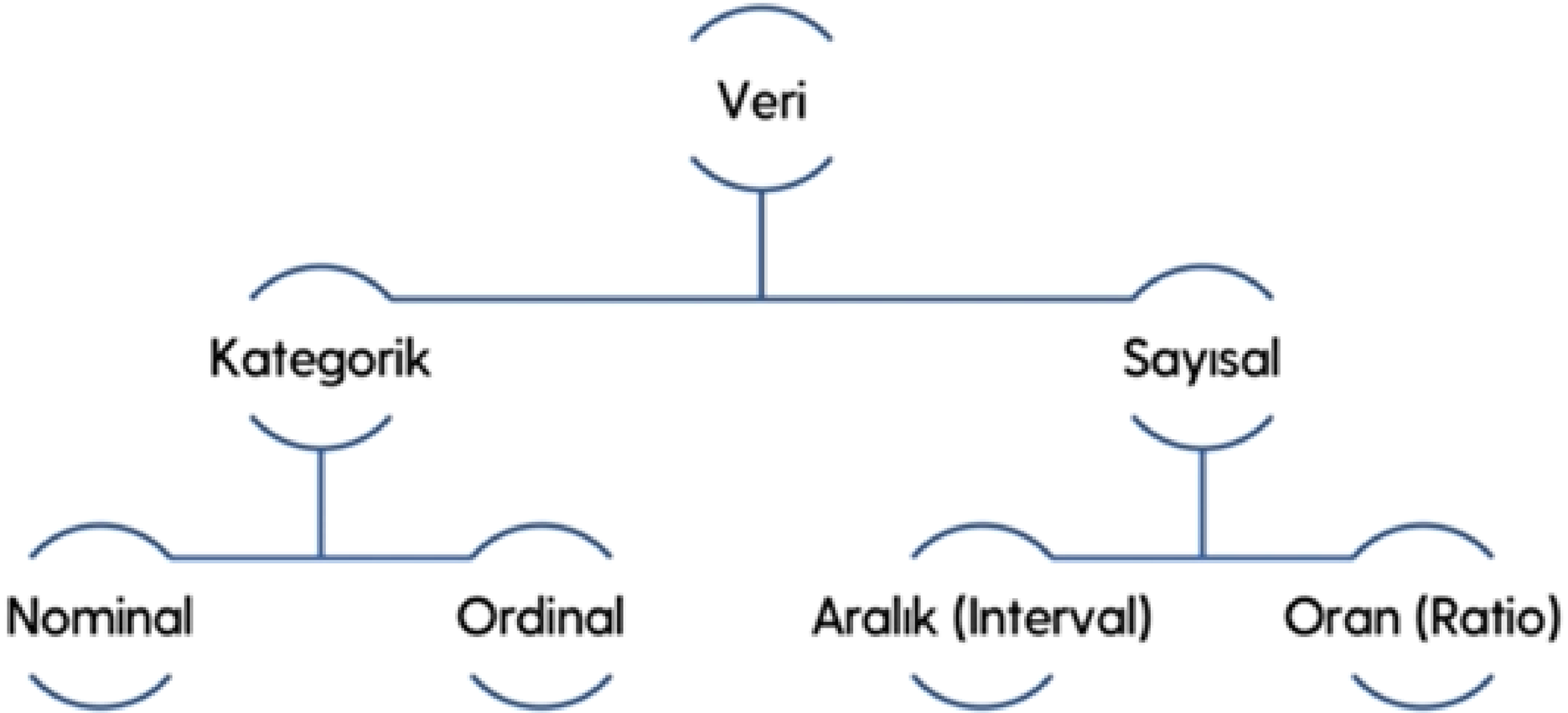
- **Veri nesnesi:** Gözlem birimidir. Her bir nesne, incelenen bir varlığı temsil eder (örneğin bir öğrenci, müşteri, ürün).
- **Özellik (feature):** Veri nesnesini tanımlayan niteliklerdir. Her nesne, bu özelliklere göre ayırt edilir (örneğin yaş, cinsiyet, fiyat, kategori).

Bu yapı sayesinde algoritmalar, nesneler arasındaki desenleri ve ilişkileri analiz edebilir.

Özellikler

Nesneler

	A	B	C
1	**Product**	**Quantity**	**Total Sales**
2	Hoodie	30	1770
3	T-shirt	50	1450
4	T-shirt	45	1305
5	Sweater	100	8900
6	Hoodie	120	7080
7	Hoodie	155	9145
8	T-shirt	80	2320
9	Sweater	90	8010
10	T-shirt	60	2940
11	Sweater	80	7120
12	Sweater	120	10680


Veri
Kategorik
Sayısal
Nominal
Ordinal
Aralık (Interval)
Oran (Ratio)

Veri Türleri - Kategorik

Nominal Veri Türü:

Nominal veriler, yalnızca kategorilere veya gruplara ayrılmış, sayısal anlam taşımayan, yalnızca isimlendirme veya etiketleme için kullan verilerdir. Bu tür verilerde sınıflar arasında herhangi bir sıralama veya derecelendirme yoktur. Bu tür veriler, sayılarla ifade edilse bile sayılar yalnızca birer etiket görevi görür; matematiksel anlam taşımaz.

Örnekler:

- Cinsiyet: Erkek, Kadın
- Kan grubu: A, B, AB, 0
- Renkler: Kırmızı, Mavi, Yeşil
- Ülke isimleri: Türkiye, Fransa, Japonya

Ordinal Veri Türü:

Ordinal veriler, belirli bir sıraya göre düzenlenmiş kategorilerdir. Bu veri türünde sıralama vardır ancak kategoriler arasındaki farkların ne kadar olduğu bilinmez. Yani bir kategori diğerinden "daha fazla" veya "daha az" olabilir, ama ne kadar daha fazla/az olduğu belli değildir.

Örnekler:

- Memnuniyet düzeyi: Çok memnun, Memnun, Kararsız, Memnun değil, Hiç memnun değil
- Eğitim seviyesi: İlkokul < Ortaokul < Lise < Üniversite
- Yarış sıralaması: 1.'lik, 2.'lik, 3.'lük
- Aciliyet düzeyi: Düşük, Orta, Yüksek

Veri Türleri - Sayısal

Interval (Aralık) Veri Türü:

Interval veri türü, veriler arasında eşit aralıklar bulunan, ancak mutlak sıfır noktası olmayan veri türüdür. Yani bu verilerle toplama ve çıkarma yapılabilir, ancak oran hesaplaması (çarpma ve bölme) yapılamaz çünkü "sıfır" eksiklik anlamına gelmez.Sıfır noktası keyfidir.

Örnekler:

- Sıcaklık (Celsius veya Fahrenheit): 30°C, 40°C. 0°C mutlak sıfır değildir. 40°C, 20°C'nin iki katı sıcaklık anlamına gelmez.
- Tarih: 1990, 2000 yılları arasında 10 yıl vardır ama "2000 yılı, 1000 yılın iki katı değildir".

Ratio (Oranlı) Veri Türü:

Ratio veri türü, sayısal değerlerin eşit aralıklı olduğu ve aynı zamanda mutlak bir sıfır noktası içeren veri türüdür. Bu sayede hem toplama-çıkarma hem de çarpma-bölme işlemleri yapılabilir. En güçlü ve en anlamlı ölçüm düzeyidir. Tüm matematiksel işlemler (toplama, çıkarma, çarpma, bölme) geçerlidir.

Örnekler:

- Uzunluk: 0 cm gerçekten hiç uzunluk yok demektir. 10 cm, 5 cm'nin iki katıdır.
- Ağırlık: 0 kg = yok. 6 kg, 3 kg'nin iki katıdır.
- Yaş: 0 yaş = doğmamış. 20 yaş, 10 yaşın iki katıdır.
- Sıcaklık (Kelvin): Mutlak sıfır noktası var.

Sürekli ve Kesikli DEğişkenler

Kesikli (Süreksiz) Değişken:

- Sayılabilir , genellikle tam sayı değerler alır.
- İki değer arasında başka (ara) değer yoktur.
- Sayma yoluyla elde edilir. Sayılabilir.
- Veri noktaları birbirinden ayrı ve belirgindir.
- İkili (Binary) gösterim bu gösterimin bir çeşididir.
- **Örnekler:** Öğrenci sayısı, araba sayısı, telefon çağrısı sayısı.

Sürekli Değişken:

- Belirli bir aralıkta sonsuz sayıda değer alabilir.
- Genellikle küsuratlı değerler ile ifade edilir.
- Ölçüm yoluyla elde edilir.
- Değerler kesintiszdir.
- **Örnekler:** Boy uzunluğu, ağırlık, sıcaklık, zaman.

Veri Analizinde Kullanılan Temel Ölçüler

Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

1-Merkezî Eğilim Ölçüleri

Merkezî eğilim ölçüleri, bir veri setindeki değerlerin topluca hangi noktada yoğunlaştığını, yani ortalama eğilimin hangi değere yakın olduğunu gösterir.

- **Aritmetik Ortalama:** Tüm verilerin toplanıp, veri sayısına bölünmesiyle elde edilir. Verilerin genel eğilimini gösterir.
- **Medyan (Ortanca):** Sıralı bir veri setinde ortada kalan değerdir. Aykırı değerlerden etkilenmez.
- **Mod (Tepe Değeri):** En sık görülen değerdir. Özellikle nominal verilerde kullanılır.

Veri Analizinde Kullanılan Temel Ölçüler

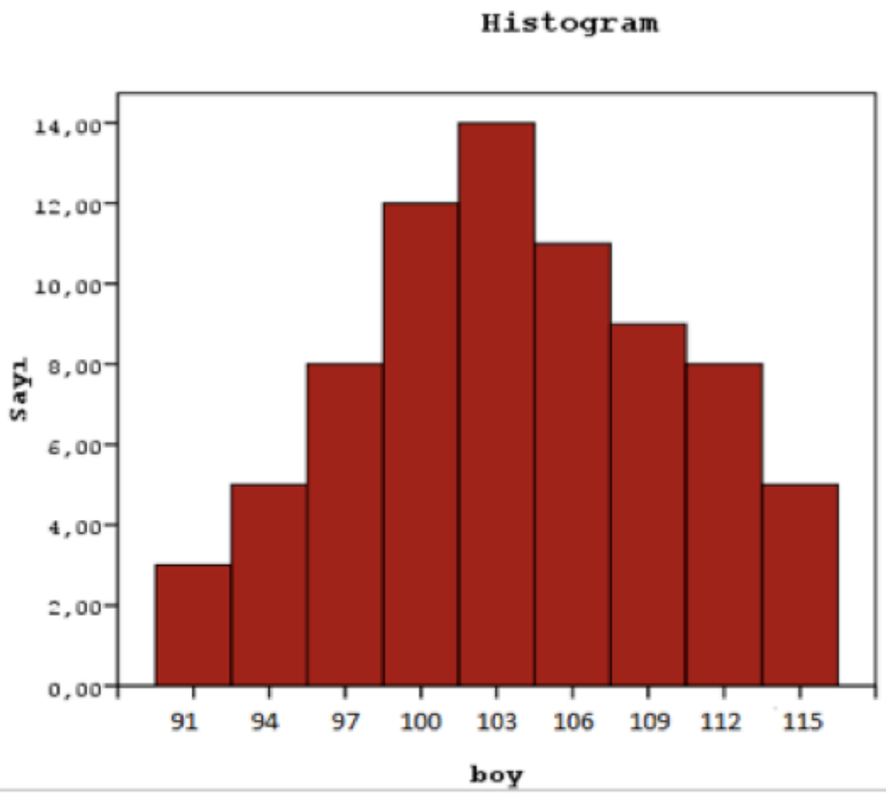
2-Dağılım Ölçüleri

Dağılım ölçüleri, verilerin ne kadar yayıldığını ya da birbirine ne kadar uzak olduğunu gösterir. Merkezî eğilim ölçüleri tek başına yeterli olmadığında, veri dağılımını anlamak için kullanılır.

- **Varyans:** Verilerin ortalamaya göre ne kadar sapma gösterdiğini gösteren ölçüdür.
- **Standart Sapma:** Varyansın kareköküdür. Verilerin ortalamaya olan ortalama uzaklığını verir.
- **Değişim Katsayısı:** Standart sapmanın ortalamaya oranıdır, farklı ölçeklerdeki veri setlerini karşılaştırmak için kullanılır.
- **Çeyrekler Açıklığı (IQR):** Üst ve alt çeyrekler arasındaki farktır. Veri setindeki orta %50'lik kısmın yayılımını gösterir

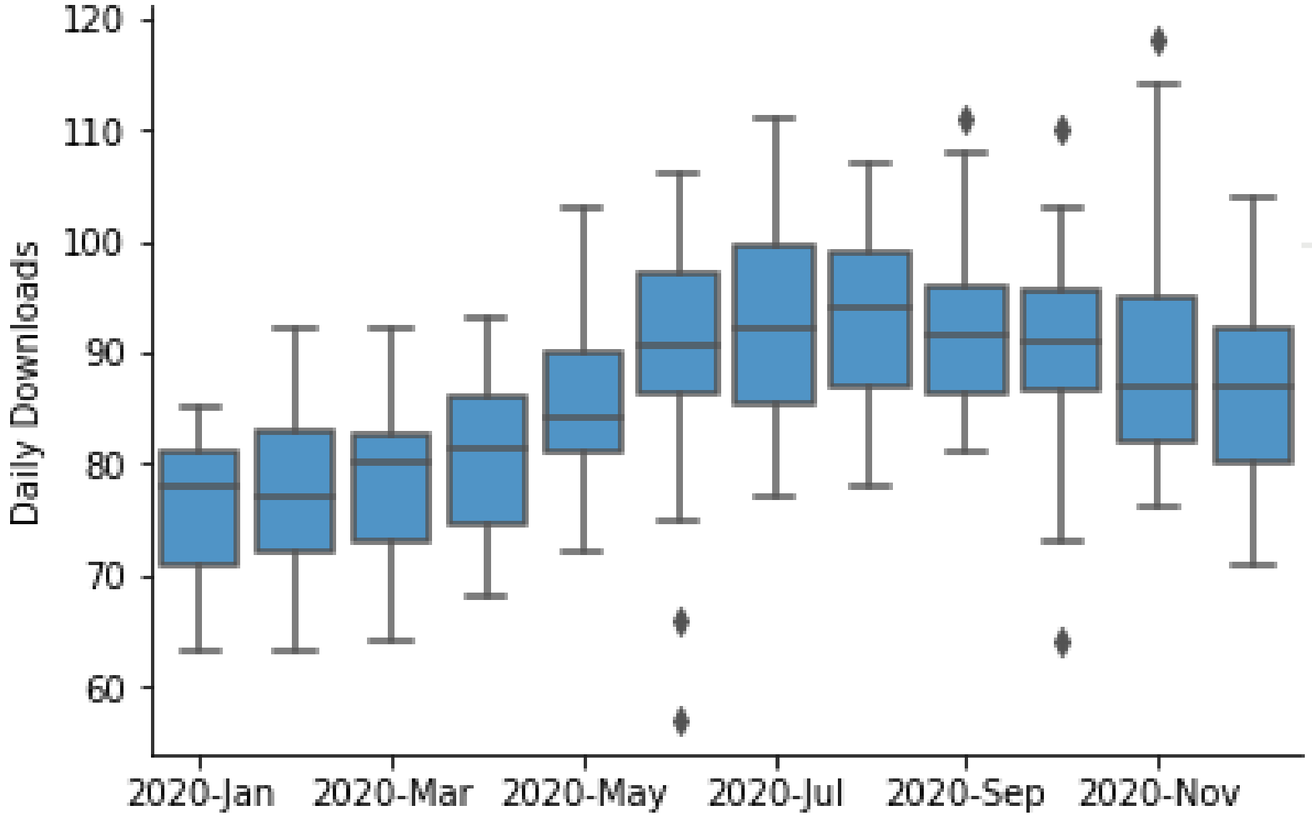
Histogram (Kutup) Grafikleri

Boy Uzunluğu	Sayı	Yüzde(%)
90-92	3	4.0
93-95	5	6.7
96-98	8	10.7
99-101	12	16.0
102-104	14	18.7
105-107	11	14.7
108-110	9	12.0
111-113	8	10.7
114-116	5	6.7
Toplam	75	100.0



Histogram grafiği ile verinin frekans dağılımı, çarpıklık vgibi özelliklerini görebiliriz.

Kutu Grafikleri (Box Plots)



Kutu grafiği ile verinin medyan, çeyrekler, aykırı değerler ini görsel olarak inceleyebiliriz.

Veri Kalitesi Nedir?

Veri kalitesi, bir verinin belirli bir amaca uygunluk derecesini ifade eder. Yani veri ne kadar doğru, eksiksiz, güncel ve tutarlıysa kalitesi o kadar yüksektir.

Zayıf veri kümesi verilerin analizi aşamasında gürültülü , eksik , tekrarlı , yanlış veriler gibi sorunlar doğurabilir.

Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

Veri Kalitesi Etkileyen Başlıca Faktörler

Doğruluk	→	Veri gerçeği ne kadar yansıtıyor?
Tutarlılık	→	Veriler birbiriyle uyumlu mu?
Güncellik	→	Veri ne kadar yeni?
Eksiksizlik	→	Tüm gerekli bilgiler mevcut mu?
Erişilebilirlik	→	Veri kolay ulaşılabilir mi?
Anlaşılırlık	→	Veri açık ve yorumlanabilir mi?

Eksik Veri (Missing Data)

Veri analizinde sık karşılaşılan sorunlardan biri, bazı gözlemlerde belirli değişkenlerin değerlerinin bulunmamasıdır. Bu durum “eksik veri” olarak adlandırılır. Eksik veriler, araştırmanın güvenilirliğini ve geçerliliğini ciddi şekilde etkileyebilir; çünkü istatistiksel analizlerin çoğu tam veri kümeleri üzerine kurulmuştur. Eksik veriler, rastgele ya da sistematik nedenlerle ortaya çıkabilir ve bu doğrultuda üç farklı kategoriye ayrılır:

Rastgele Eksik (MCAR)
Koşullu Eksik (MAR)
Rastgele Olmayan Eksik (MNAR)

Eksik verilerle başa çıkma yöntemleri:

- Silme (Listwise / Pairwise Deletion)
- Ortalama ile doldurma
- Regresyon tahmini

İleri düzey yöntemler (Multiple Imputation, EM algorithm)

Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

Rastgele Eksik (MCAR):

Eksiklik tamamen Rastgeledir

Koşullu Eksik (MAR):

Eksiklik başka gözlemlerle ilişkilidir.

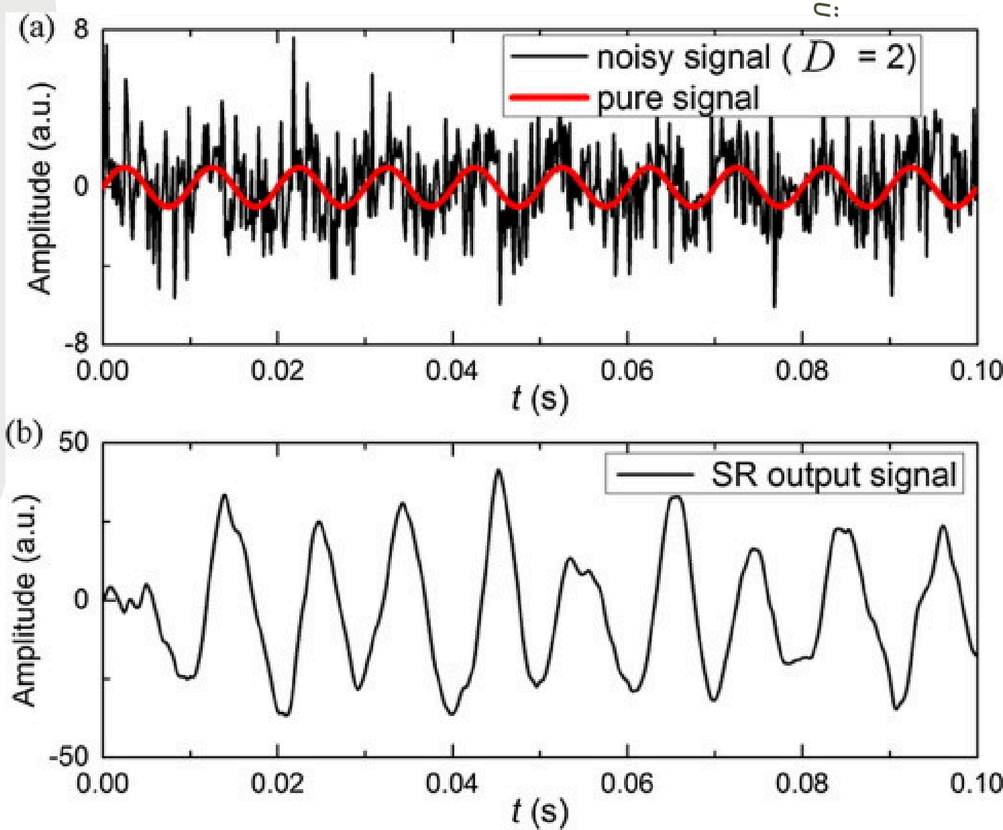
Rastgele Olmayan Eksik (MNAR):

Eksikliğin nedeni gözlemin kendisinden kaynaklanır.

Gürültülü Veri (Değer)

Gürültülü veri, veri setindeki gerçek bilginin üzerine binen rastgele, düzensiz ve genellikle anlamsız sapmaları ifade eder. Bu tür sapmalar çoğunlukla ölçüm hatalarından, veri girişindeki tutarsızlıklardan ya da çevresel faktörlerin etkilerinden kaynaklanır.

Gürültü, verilerin analiz edilmesini zorlaştırabilir çünkü gerçek desenleri veya eğilimleri gizleyebilir. Bu nedenle, gürültülü veriler genellikle ön işleme (preprocessing) adımında tespit edilir ve filtreleme, düzeltme ya da çıkarma yöntemleriyle analizden önce temizlenir.



Gürültülü Veri Örnekleri

Sensör Verilerinde Gürültü

Rüzgar, titreşim veya elektromanyetik girişim nedeniyle sensör okumalarının dalgalanması

Ses Tanıma Sistemlerinde Gürültü

Konuşma sırasında arka planda araba kornası, insan sesi ya da müzik gibi seslerin mikrofona gelmesi

Görüntü İşlemede Gürültü

Kamera ile çekilen bir fotoğrafta, ışığın yetersiz olması sonucu oluşan “karlı” görüntü

Anket ve Form Verilerinde Gürültü

Katılımcının rastgele, tutarsız cevaplar vermesi veya zorunlu alanların "asdf" gibi geçersiz veriyle doldurulması

Finansal Verilerde Gürültü

Hisse senedi fiyatlarında bir gün içinde ani, anlamsız sıçramalar (çok küçük zaman aralıklarındaki dalgalanmalar)

Zaman Serilerinde Gürültü

GPS verilerinde birkaç saniyelik tutarsız koordinatlar

Aykırı Veri

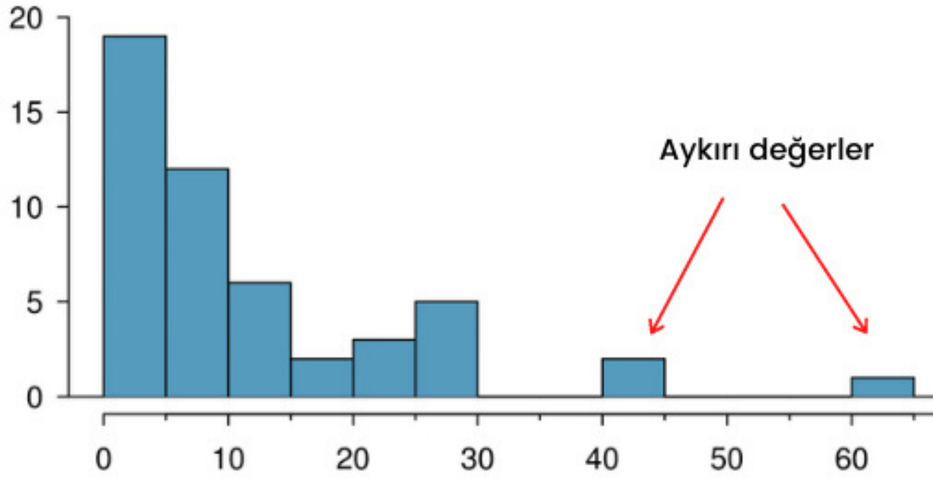
Aykırı veri, bir veri kümesinde yer alan ve diğer gözlemlerden belirgin şekilde farklı olan verilerdir. Bu veriler genellikle veri kümesindeki dağılımın dışında yer alır ve istatistiksel analizlerde ortalama, varyans gibi ölçütleri bozabilir. Aykırı değerler; veri giriş hatası, ölçüm sorunu veya gerçekten istisnai bir durumun göstergesi olabilir.

Aykırı verilerin tespiti için **Z-skoru**, **IQR yöntemi (çeyrekler arası açıklık)**, **boxplot (kutu grafiği)** gibi yöntemler kullanılır. Aykırı veriler silinmeden önce mutlaka analiz edilmeli; verinin hatalı mı yoksa özel bir durumu mu temsil ettiği anlaşılmalıdır.

- Bir sınıfta notlar genelde 60-90 aralığında iken, bir öğrencinin 5 alması veya 100 alması
- 10 yaşındaki çocukların boyu ortalama 140 cm iken birinin boyu 200 cm çıkması
- Bir araç 50 km/s hızla giderken, bir anda 300 km/s gösteren konum verisi

gibi veriler aykırı veriye örnek verilebilir.

Samsun Üniversitesi
Yazılım Mühendisliği Bölümü



Tekrarlı veriler

Tekrarlı veriler, aynı veri kümesinde birden fazla kez yer alan aynı ya da neredeyse aynı kayıtlardır. Genellikle veri girişi hataları, sistem birleştirme süreçleri veya eksik kimlik doğrulama nedeniyle oluşur.

- Depolama israfı
- Analizlerde çarpıtılmış sonuçlar

gibi sorunlara yol açabilir.

Veri Birleştirme (Data Integration)

Veri birleştirme, birden fazla kaynaktan gelen verilerin, tek ve tutarlı bir veri kümesinde birleştirilmesi işlemidir. Farklı veri formatları, sistemler veya dosyalar arasındaki veriler uyumlu hale getirilir.Veri analitiği, raporlama ve makine öğrenmesi için sağlam temel sağlar.

UYGULAMA

İris Veri Seti I

Kod Parçası

```python
import pandas as pd

# Ana veri seti
df1 = pd.DataFrame({
    'id': [1, 2, 3],
    'ad': ['Ahmet', 'Ayşe', 'Mehmet']
})

# Ek veri seti
df2 = pd.DataFrame({
    'id': [1, 2, 3],
    'şehir': ['İstanbul', 'Ankara', 'İzmir']
})

# Birleştirme (id üzerinden)
birlesik_df = pd.merge(df1, df2, on="id", how="left")

print("📌 Birleştirilmiş Veri:")
print(birlesik_df)

```

Kod Çıktısı

```
📌 Birleştirilmiş Veri:
   id      ad     şehir
0   1   Ahmet  İstanbul
1   2    Ayşe    Ankara
2   3  Mehmet     İzmir
```

UYGULAMA

İris Veri Seti II

Kod Parçası

```python
import pandas as pd
import numpy as np

# Örnek veri seti
df = pd.DataFrame({
    'ad': ['Ahmet', 'Ayşe', 'Mehmet'],
    'şehir': ['İstanbul', np.nan, 'İzmir']
})

print("📌 Boş Değerler Öncesi:")
print(df)

# Boş değerleri "Bilinmiyor" ile doldurma
df['şehir'] = df['şehir'].fillna('Bilinmiyor')

print("\n📌 Boş Değerler Doldurulduktan Sonra:")
print(df)
```

Kod Çıktısı

```
📌 Boş Değerler Öncesi:
      ad     şehir
0  Ahmet  İstanbul
1   Ayşe       NaN
2 Mehmet     İzmir

📌 Boş Değerler Doldurulduktan Sonra:
      ad       şehir
0  Ahmet    İstanbul
1   Ayşe  Bilinmiyor
2 Mehmet       İzmir
```

UYGULAMA

İris Veri Seti III

Kod Parçası

```python
import pandas as pd
import numpy as np
from sklearn.datasets import load_iris

# 1️⃣ Iris veri setini yükle
iris = load_iris()
df1 = pd.DataFrame(iris.data, columns=iris.feature_names)

# 2️⃣ Etiket sütununu ekle
df1['target'] = iris.target

# 3️⃣ İkinci bir veri seti oluştur (bazı eksik veriler var)
df2 = pd.DataFrame({
    'region': ['Europe', 'Asia', None, 'America', 'Africa'] * 30,
    'year': [2020, 2021, 2022, None, 2023] * 30
})

# 4️⃣ Veri birleştirme (yanyana)
df = pd.concat([df1, df2], axis=1)


# 6️⃣ Eksik verileri doldurma
df['region'] = df['region'].fillna('Unknown')
df['year'] = df['year'].fillna(df['year'].mean())

# 7️⃣ Son hali görüntüle
print("\nİlk 5 satır:")
print(df.head(5))

```

Kod Çıktısı

```
İlk 5 satır:
   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)  target   region    year
0                5.1               3.5                1.4               0.2       0   Europe  2020.0
1                4.9               3.0                1.4               0.2       0     Asia  2021.0
2                4.7               3.2                1.3               0.2       0  Unknown  2022.0
3                4.6               3.1                1.5               0.2       0  America  2021.5
4                5.0               3.6                1.4               0.2       0   Africa  2023.0
```

Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

VERİ MADENCİLİĞİ (DATA MİNİNG)

Dr. Öğr. Üyesi Alper Talha Karadeniz

2025-2026