# VERİ ÖN İŞLEME

- Veri ön işleme Genel Bakış
- Aykırı Gözlem Analizi
- Eksik Gözlem Analizi
- Standartlaştırma
- Değişken Dönüşümleri

## Veri Ön işleme Genel Bakış

- **Veri temizleme**
  - → Gürültülü Veri
  - → Eksik Veri Analizi
  - → Aykırı Gözlem Analizi
- **Veri Standardizasyonu**
  - → 0-1 Dönüşümü
  - → Z-skoruna dönüştürme
  - → Logaritmik dönüşüm
- **Veri indirgeme**
  - → Gözlem Sayısının Azaltılması
  - → Değişken Sayısının Azaltılması
- **Değişken Dönüşümleri**
  - → Sürekli değişkenlerde dönüşümler
  - → Kategorik değişkenlerde dönüşümler
- **Değişken Mühendisliği**

#Veri ön işleme; modeller kurulmadan önce veri seti üzerinde yapılan birtakım düzeltme, eksik veriyi tamamlama, tekrarlanan verileri kaldırma, dönüştürme, bütünleştirme, temizleme, normalleştirme, boyut indirgeme vb. işlemlerdir.

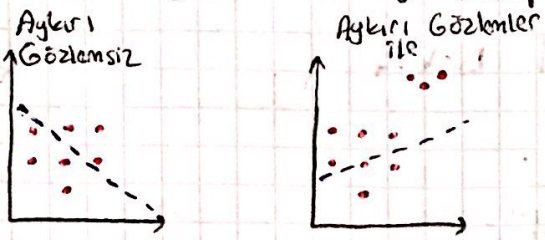
- Gürültülü Veri
  - Veri Kaynağına Bağlı Hatalar (anketler, veri tabanları, ara unsurlar)
  - Tutarsızlık (Cinsiyet = Erkek, Gebelik Durumu = 1, Kategori = Biberon, Fiyat = 900 bin TL, Vasıta Türü = Otomobil, Motor Gücü = 25 HP)
  - Kayıtlarda Çoklama

#Gürültülü veri; verinin içerisindeki gerçek yapının dışında ortaya çıkan, verinin taşıdığı bilgiyi bozan, karmaşıklaştıran, kararsızlığa sebep olabilecek olan veridir.

# Aykırı Gözlem Analizi

* Veride genel eğilimin oldukça dışına çıkan ya da diğer gözlemlerden oldukça farklı olan gözlemlere aykırı gözlem denir.
* Aykırılığı ifade eden nümerik değere aykırı değer denir.
* Aykırı değeri barındıran gözlem birimine aykırı gözlem denir.

* Genellenebilirlik kaygısı ile oluşturulan kural setlerini ya da fonksiyonları yanıltır. Yanlılığa sebep olur.




1. Sektör Bilgisi
   Örneğin bir ev fiyat tahmin modelinde 1000 metrekarelik evleri modellemeye almamak.
2. Standart Sapma Yaklaşımı
   Bir değişkenin ortalamasının üzerine aynı değişkenin standart sapması hesaplanarak eklenir. 1,2 ya da 3 standart sapma değeri ortalama üzerine eklenecek ortaya çıkan bu değer eşik değer olarak düşünülür. Bu değerden yukarıda ya da aşağıda olan değerler aykırı değer olarak tanımlanır.
3. Z-Skoru Yaklaşımı
   Standart sapma yöntemine benzer şekilde çalışır. Değişken standart normal dağılıma uyarlanır, yani standartlaştırılır. Sonrasında örneğin dağılımın sağından ve solundan +/- 2,5 değerine göre bir eşik değer konulur. Bu değerin üzerinde ya da altında olan değerler aykırı değer olarak işaretlenir.
4. Boxplot (interquartile range - IQR) Yöntemi
   En sık kullanılan yöntemlerden birisidir. Değişkenin değerleri küçükten büyüğe sıralanır. Çeyreklerine (yüzdeliklerine) yani Q1, Q3 değerlerine karşılık değerler üzerinden bir eşik değer hesaplanır ve bu eşik değere göre aykırı değer tanımı yapılır.

* Tek Değişkenli
  - → Box-Plot → Histogram
  - → Standart Sapma → Standart Normal Dağılım
* Çok Değişkenli - İstatistiksel Yöntemler
  - → Kümeleme Yöntemi
  - → İkişerli saçılım grafiği ve kontur grafikleri (%90)
  - → Kare Mahalanobis uzaklığı hesaplamak
  - → Genelleştirilmiş varyans oranı
* Çok Değişkenli - Diğer Yöntemler
  - → Derinlik Temelli Yaklaşımlar
  - → Sapma Temelli Yaklaşımlar
  - → Uzaklık Temelli Yaklaşımlar
  - → Yoğunluk Temelli Yaklaşımlar
  - → Yüksek Boyutlu Yaklaşımlar

# Eksik Veri Analizi

#İncelenen veri setindeki gözlemlerde eksiklik olması durumunu ifade etmektedir.

## Eksik Veri Adımları
→ Eksik verinin belirlenmesi
→ Yapısının Görsel teknikler ile incelenmesi
→ Eksikliğin rassallığının test edilmesi
→ Uygun yöntemler ile doldurulması

## Eksik Veriyi Direkt Silmenin Zararları
→ Veri setindeki eksikliğin yapısal bir eksiklik olup olmadığının bilinmesi gerekir.
→ NA her zaman eksiklik anlamına gelmez.
→ Bilgi Kaybı.

## Eksik Veri Türleri
### Tümüyle Raslantısal Kayıp
Diğer değişkenlerden ya da yapısal bir problemden kaynaklanmayan, tamamen rastgele oluşan gözlemler.
### Raslantısal Kayıp
Diğer değişkenlere bağlı olarak oluşabilen eksiklik türü.
### Raslantısal Olmayan Kayıp
Göz ardı edilemeyecek olan ve yapısal problemler ile ortaya çıkan eksiklik türü.

## Eksik Veri Rassallığının Testi
→ Bağımsız iki örneklem t testi
→ Korelasyon testi
→ Little'nin MCAR testi

## Eksik Veri Probleminin Giderilmesi
### Silme Yöntemleri
→ Gözlem ya da değişken silme yöntemi
→ Liste bazında silme yöntemi (Listwise Method)
→ Çiftler bazında silme yöntemi (Pairwise Method)

### Değer Atama Yöntemleri
→ Ortanca, Ortalama, Medyan
→ En benzer birime atama (hot deck)
→ Dış kaynaklı atama

### Tahmine Dayalı Yöntemler
→ Makine öğrenmesi
→ EM
→ Çoklu Atama Yöntemi