Veri nedir?

- Veri nedir?
  - Geometrik bir bakış açısı
    - Benzerlik
  - Olasılıksal bir bakış açısı
    - Yoğunluk
- Veri kalitesi
- Veri önişleme
  - Birleştirme
  - Örneklem
  - Veri küçültme
    - Temel bileşen analizi (Principal Component Analysis)
    - Çok boyutlu ölçekleme (Multidimensional Scaling)

Veri Nedir?

- Genel olarak veri *n*×*d* bir matris ile ifade edilir

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

  - *n* satır ve *d* sütun,
    - Satırlar örneklere

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})$$

    - Sütunlar özelliklere (değişkenlere) işaret eder.

$$X_j = (x_{1j}, x_{2j}, \cdots, x_{nj})$$

Veri Nedir?

- Örnek sayısı, *n*, verinin büyüklüğünü (size), öznitelik (özellik, değişken) sayısı, *d*, ise verinin boyutunu (boyutsallık-dimensionality)

d=1 –> Tek değişkenli analiz (univariate)

d=2 -> Çift değişkenli analiz (bivariate)

d>2 -> Çok değişkenli analiz (multivariate)

Iris veri seti

http://en.wikipedia.org/wiki/Iris_flower_data_set

	X_1 sepal length	X_2 sepal width	X_3 petal length	X_4 petal width	X_5 class
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Iris setosa

Iris versicolor

Iris virginica

Diğer tip veri setleri

- Her veri seti matris formatında olmayabilir
- Daha karmaşık veri setleri aşağıdaki öğeleri içerebilir
  - Kayıt tipi veri
    - İşlem/hareket (transaction)
  - Sıralı veri
    - DNA/Protein,
    - Metin,
    - Zaman serisi,
    - Resim,
    - Ses
    - Video, ve diğerleri…
  - Ağ (graph) tipi veri
    - World Wide Web
    - Molekül yapıları
- analiz için daha özelleşmiş teknikler gerektirir.

Öznitelik tipleri

- Alabileceği değer kümesine göre iki ana itpe ayrılır.
  - Kategorik öznitelikler
    - Nominal: Göz rengi, TC Kimlik No
    - Ordinal: Eğitim durumu, anket sorusu cevabı (kötü, orta, iyi)
  - Numerik öznitelikler
    - Aralık-ölçekli (Interval-scale): Sıcaklık
      - 10 ve 20 derece iki hava durumunu, dün bugünden iki kat soğuktu demeyiz.
    - Oran-ölçekli (Ratio-scaled): yaş
      - 20 yaşındaki bir insan 10 yaşındaki insandan iki kat daha yaşlıdır.

Farklı bir kategorizasyon
Kesikli ve Sürekli Öznitelikler

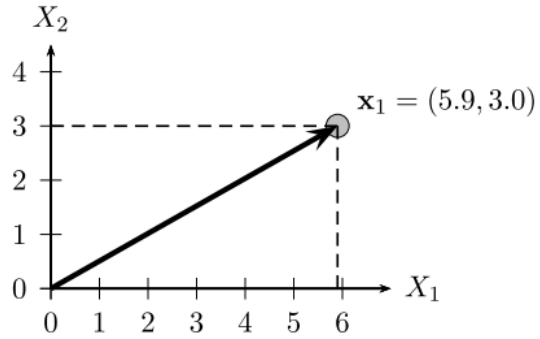
- Kesikli öznitelikler
  - Değer kümesi sayılabilen (countable) özniteliklerdir.
  - örnekler: posta kodu, paragrafta yer alan kelimeler, araba markaları
  - Çoğunlukla tam sayılar ile ifade edilir

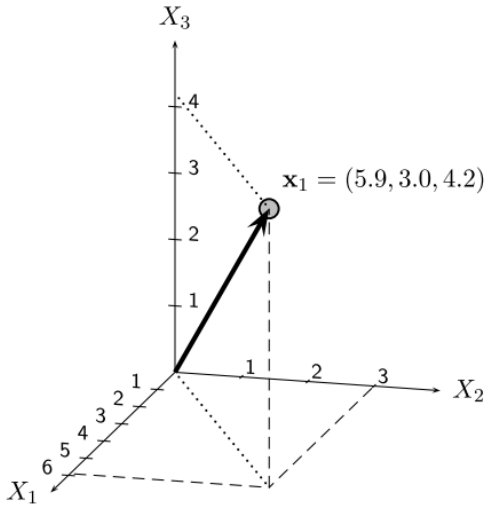
- Sürekli öznitelikler
  - Reel sayılar ile ifade edilir.
  - örnekler: sıcaklık, ağırlık, uzunluk
  - Ölçülebilen ve sonlu sayıda rakam ile ifade edilen özniteliklerdir.

Veri Analizine Yaklaşım
Geometrik bakış

- *d* tane öznitelik içeren *D* veri matrisinde tüm öznitelikler nümerik ise her satır d-boyutlu uzayda bir nokta olarak ifade edilebilir.

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id}) \in \mathbb{R}^d$$





Benzerlik – Farklılık (Similarity – Dissimilarity)

- Benzerlik
  - İki verinin ne kadar benzer olduğunun bir ölçütüdür.
  - Veriler benzediğinde büyük değerler alır.
  - Genellikle [0,1] aralığında ifade edilir.
- Farklılık.
  - İki verinin ne kadar farklı olduğunun bir ölçütüdür.
  - Veriler benzediğinde küçük değerler alır.
  - En düşük farklılık çoğunlukla sıfır ile ifade edilir.
  - Üst limit değişebilir.
- Benzerlik ya da farklılık kimi zaman Yakınlık (Proximity) olarak ifade edilir.

Basit öznitelikler için benzerlik/farklılık Öklit Uzaklık

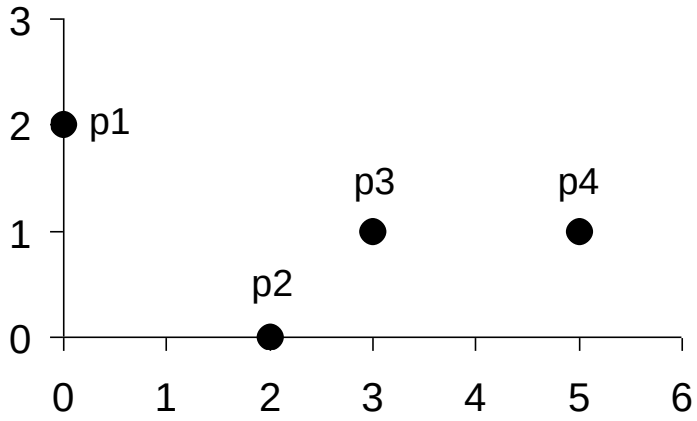
- Öklit uzaklık

$$dist = \sqrt{\sum_{k=1}^{t} (p_k - q_k)^2}$$

p ve q örneklerinin k öznitelik değerlerinin farklarının karesinin tüm öznitelikler (t tane) üzerinden toplanması ile bulunur.

- Öznitelik ölçekleri farklı olduğu durumda, standardizasyon gereklidir.

Basit öznitelikler için benzerlik/farklılık
Öklit Uzaklık



örnek	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Uzaklık Matrisi

(Distance Matrix)

Basit öznitelikler için benzerlik/farklılık Minkowski Uzaklık

- Minkowski Uzaklık genelleştirilmiş bir uzaklık ölçüsüdür

$$dist = (\sum_{k=1}^{t} | p_k - q_k |^r)^{\frac{1}{r}}$$

r parametresinin değerlerine göre uzaklık tanımı değişir.

Basit öznitelikler için benzerlik/farklılık Minkowski Uzaklık

- r = 1. Manhattan, L_1 norm
- r = 2. Öklit, L_2 norm
- $r \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - Öznitelikler arası farkların maximumu

Basit öznitelikler için benzerlik/farklılık
Minkowski Uzaklık

örnek	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Uzaklık matrisi

Veri Analizine Yaklaşım
Olasılıksal bakış

- Her öznitelik rassal bir değişkendir.
  - Her deneyin sonucunda belli bir kurala değer atayan bir fonksiyon.

Makine-Süreç

Bilinmeyen olasılık
dağılımı
-tipi
-parametreleri

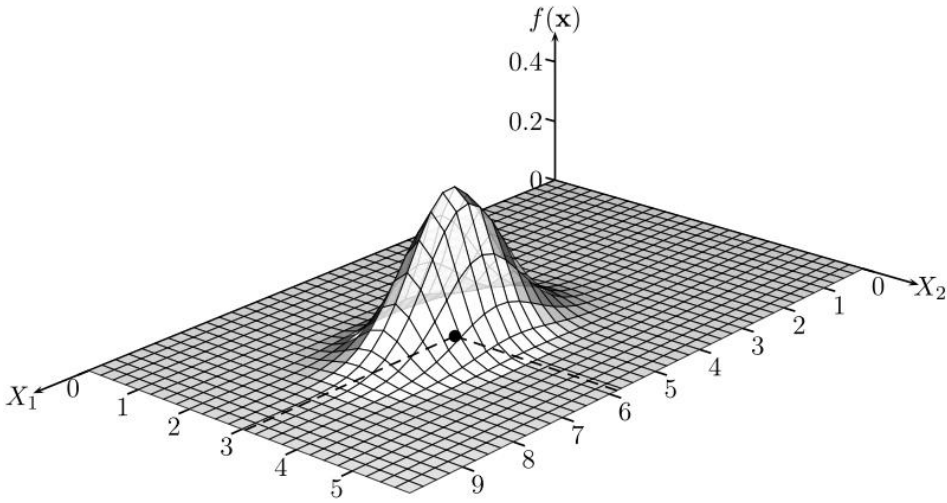
X

5.9	6.9	6.6	4.6	6.0	4.7	6.5	5.8	6.7	6.7	5.1	5.1	5.7	6.1	4.9
5.0	5.0	5.7	5.0	7.2	5.9	6.5	5.7	5.5	4.9	5.0	5.5	4.6	7.2	6.8
5.4	5.0	5.7	5.8	5.1	5.6	5.8	5.1	6.3	6.3	5.6	6.1	6.8	7.3	5.6
4.8	7.1	5.7	5.3	5.7	5.7	5.6	4.4	6.3	5.4	6.3	6.9	7.7	6.1	5.6
6.1	6.4	5.0	5.1	5.6	5.4	5.8	4.9	4.6	5.2	7.9	7.7	6.1	5.5	4.6
4.7	4.4	6.2	4.8	6.0	6.2	5.0	6.4	6.3	6.7	5.0	5.9	6.7	5.4	6.3
4.8	4.4	6.4	6.2	6.0	7.4	4.9	7.0	5.5	6.3	6.8	6.1	6.5	6.7	6.7
4.8	4.9	6.9	4.5	4.3	5.2	5.0	6.4	5.2	5.8	5.5	7.6	6.3	6.4	6.3
5.8	5.0	6.7	6.0	5.1	4.8	5.7	5.1	6.6	6.4	5.2	6.4	7.7	5.8	4.9
5.4	5.1	6.0	6.5	5.5	7.2	6.9	6.2	6.5	6.0	5.4	5.5	6.7	7.7	5.1

Iris verisetinde gövde uzunluğu,
n=150
- Sürekli bir rassal değişken

Veri Analizine Yaklaşım
Olasılıksal bakış

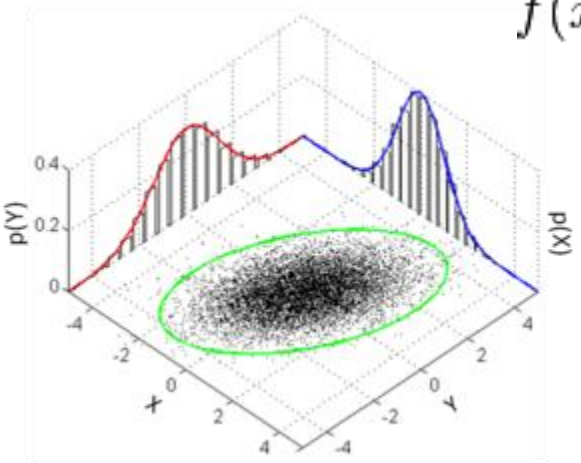
- Her özniteliği tek değişkenli bir rassal sayı olarak tanımlamak yerine, veri setimizin çok değişkenli bir rassal sayıdan oluştuğunu düşünebiliriz.
  - Zar atma
    - P(1,1)=1/36
    - P(1,2)=1/36
    - ….
  - Örnek
    - İki değişkenli (bivariate) Normal dağılım
      - Gaus (Gaussian) dağılım olarak da tanımlanır (bir ya da birden çok değişkenli rassal sayılar için)
  - Gerçek hayat problemleri çoğunlukla çok değişkenlidir.

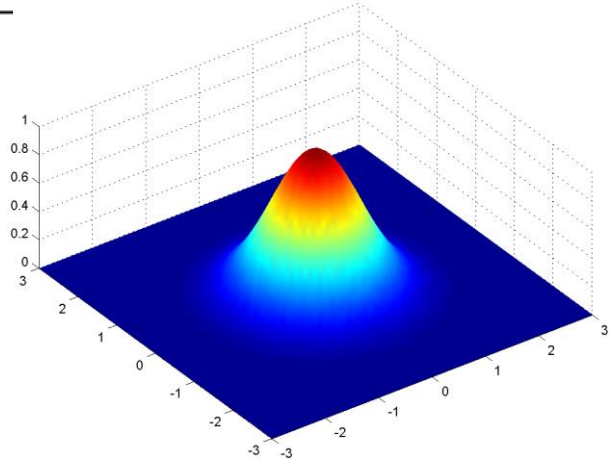


Olasılıksal bakış
Normal (Gaus) dağılım

- En çok kullanılan dağılımdır, neden?
  - Merkezi limit teoremi
    - Aynı tipte dağılıma sahip birbirinden bağımsız rassal değişkenlerden elde edilen sayıların ortalaması yaklaşık olarak Normal dağılım gösterir

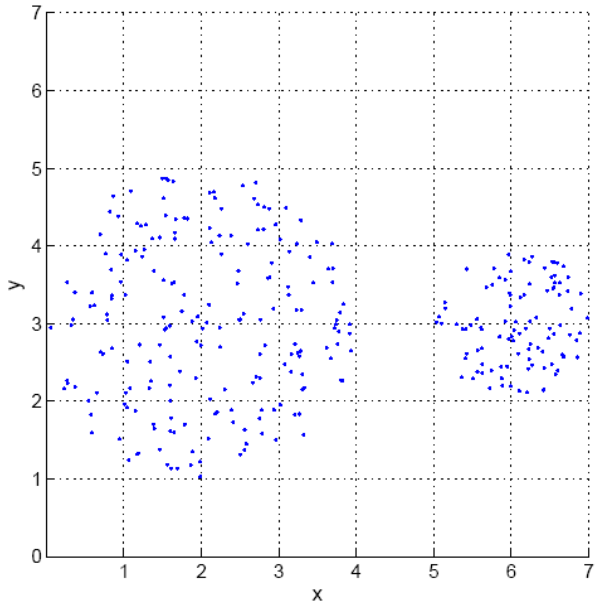
$$f(x, \mu, \sigma) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$





Olasılıksal bakış
Öklit dağılım

- Eğer dağılımı bilmiyorsak
  - Uzayı eşit aralıklara bölü noktaları sayarak dağılımı ifade edebiliriz.

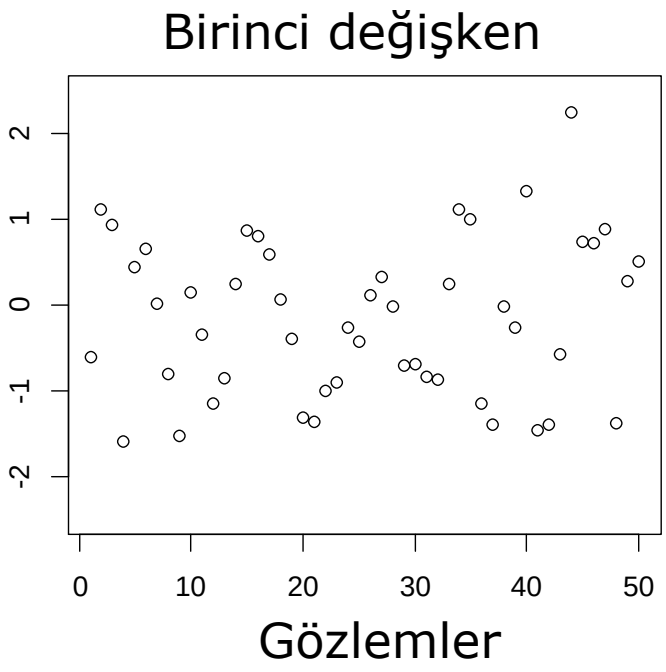


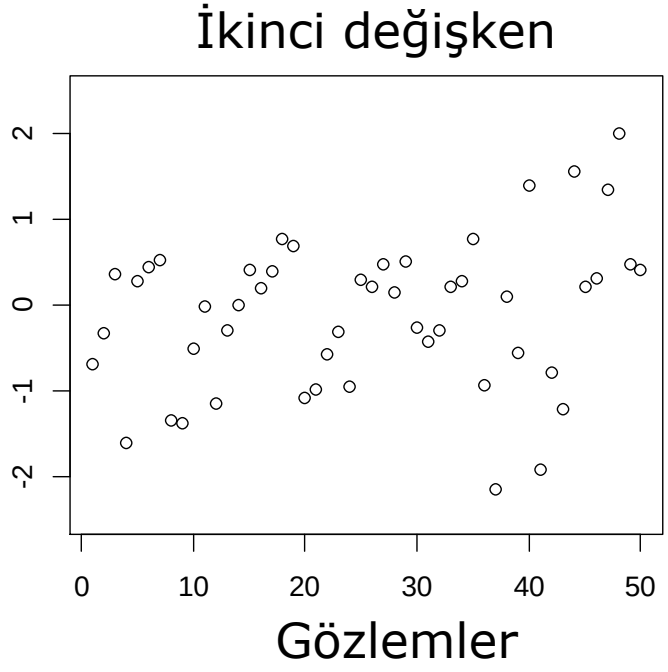
0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Olasılıksal bakış

Kovaryans

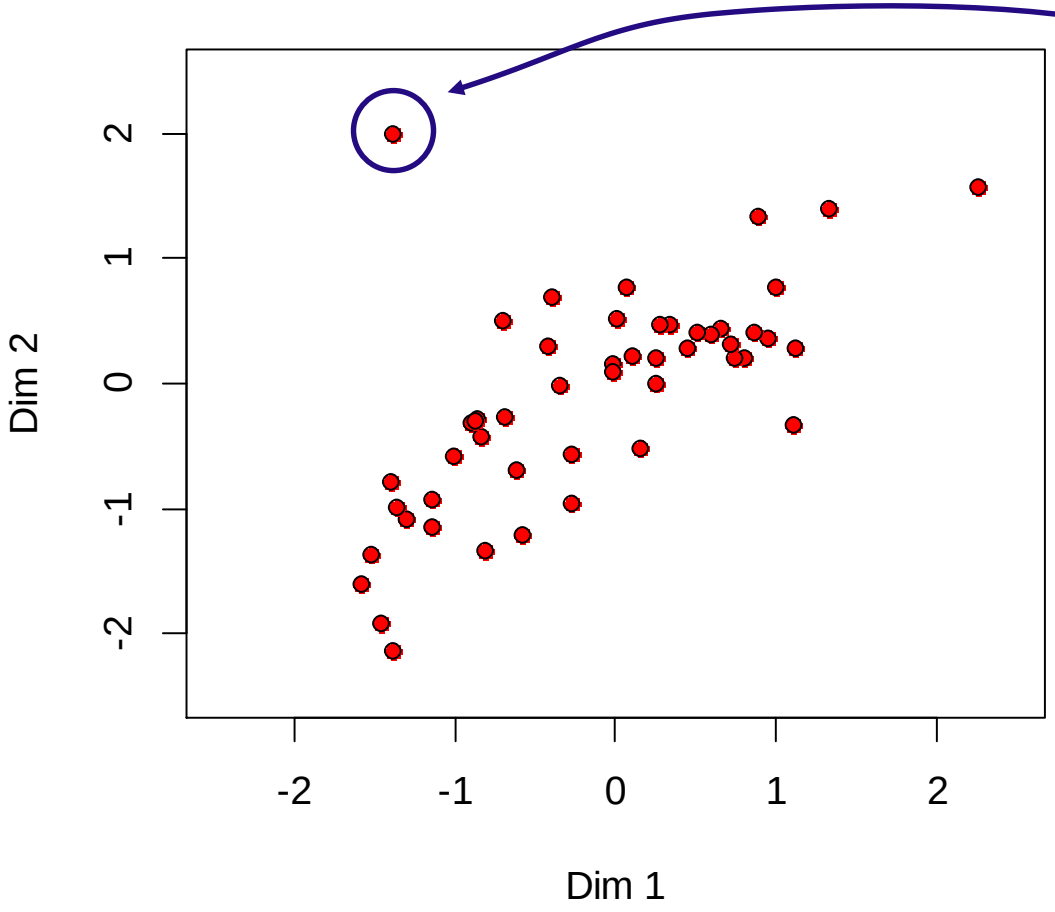
- İki rassal değişkenin birlikte değişiminin ölüsüdür.
  - Örneğin 2 boyutlu Normal dağılım izleyen bir 50 örnekli bir veri setimiz olsun.





Olasılıksal bakış
Kovaryans

- İki boyut birlikte çizildiğinde



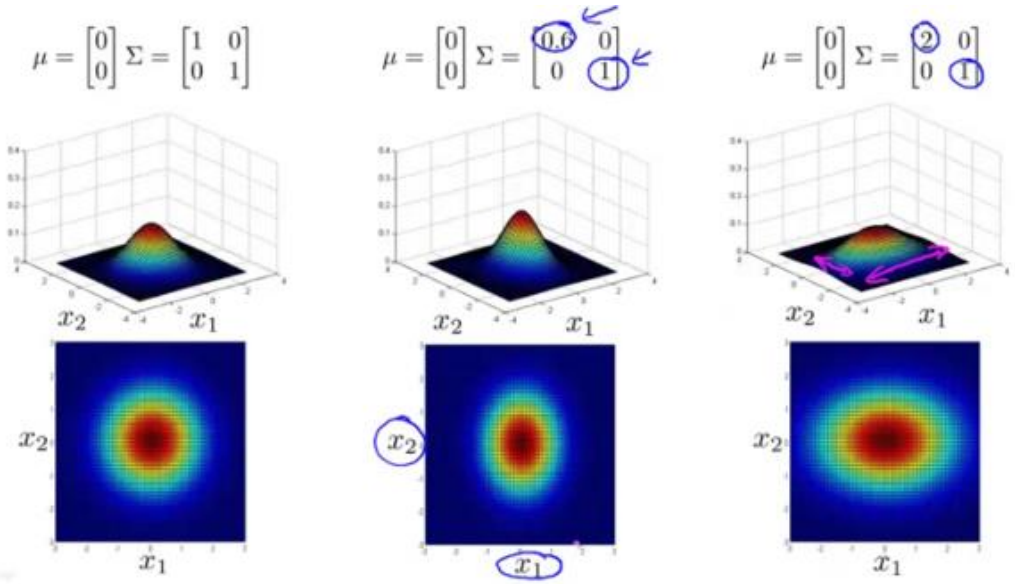
Problem, gürültü ya da aykırı davranış?

İlişkilerin modellenmesi önemlidir.

Olasılıksal bakış
Çok değişkenli Normal dağılım

- Tek değişkenli normal dağılımın parametreleri ortalama ve standart sapmadır.
- Çok değişkenli Normal dağılım ortalama vektörü ve kovaryans matrisi ile tanımlanır.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$



Olasılıksal bakış
Çok değişkenli Normal dağılım

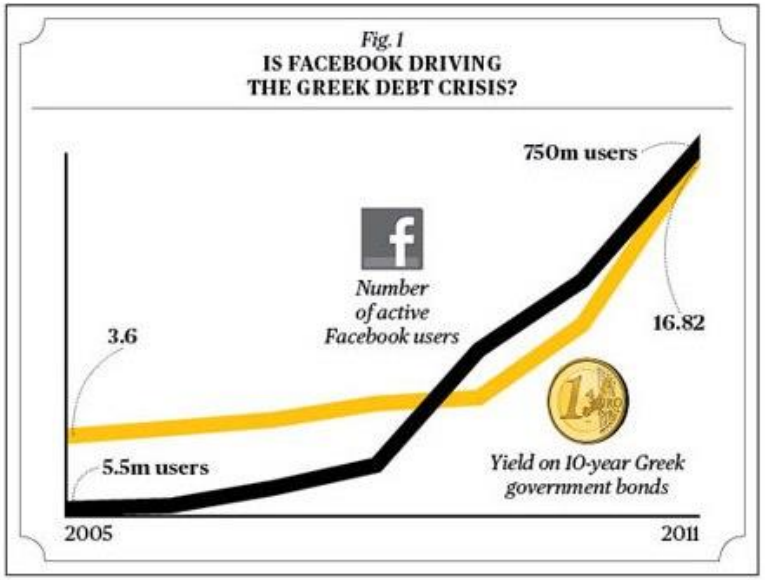
- R örnekleri

Korelasyon

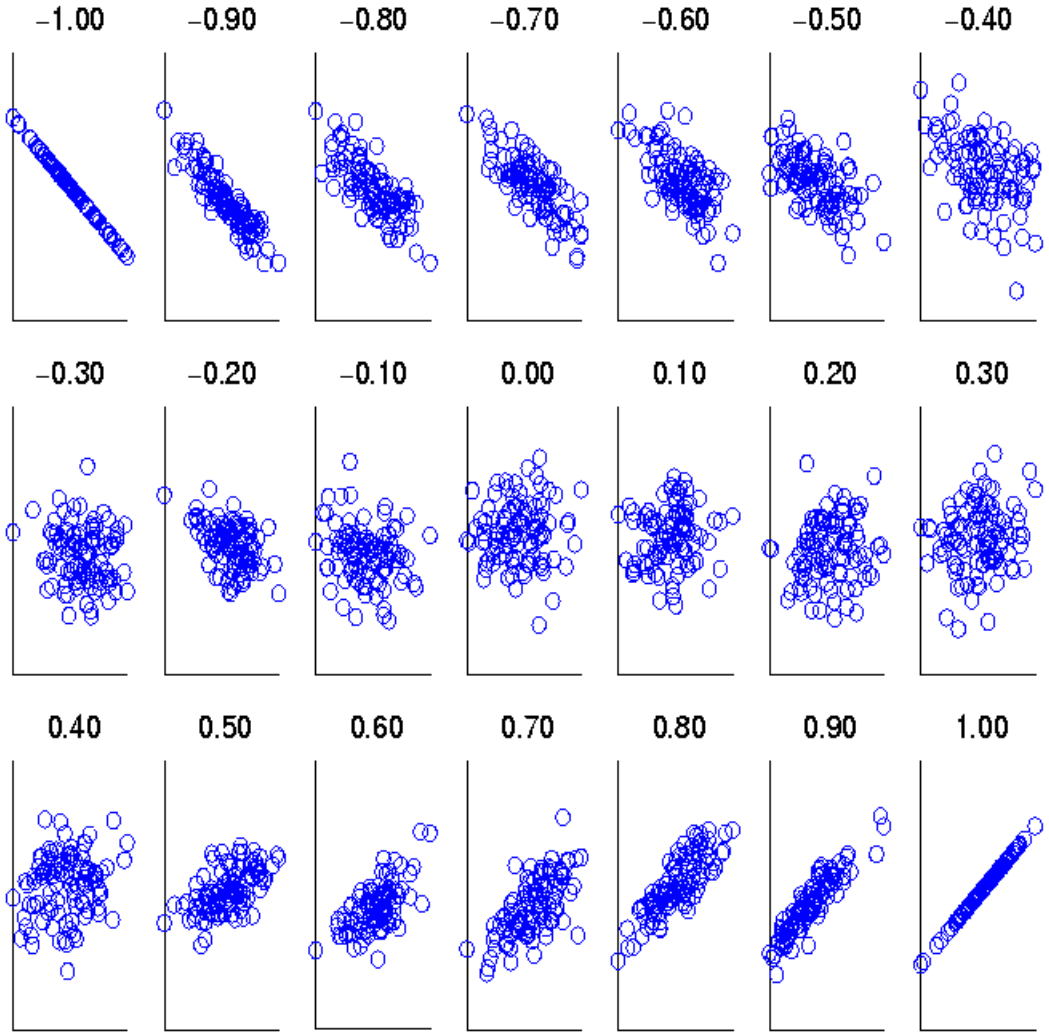
- İki örnek arası lineer ilişkiyi modeller

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$



Görsel olarak korelasyon



İlişkinin değeri –1 ile 1 arası değişir.

Veri kalitesine etki eden faktörler

- Doğruluk
  - Veri girişinde yapılan hatalar
- Bütünlük
  - Kayıp veri
- Teklik (Uniqueness)
  - Aynı verinin birden fazla kaydı
- Güncellik (Timeliness)
  - Vakti geçmiş işe yaramayan veri
- Tutarlılık
  - Verinin kendisiyle çeliştiği durumlar

Veri analizindeki en önemli aşama önişleme aşamasıdır.

- Teknikler: Örnekleme, Boyut küçültme, Değişken seçimi.
- Zorlu bir aşamadır ama genelde en önemli aşamalardan biridir.

Veri kalitesi
Kirli veri

- Kirli veri ne demek?
  - Eksik veri
    - kayıp (missing) değerler
  - Tutarsız veri
    - (farklı kodlama, imkansız değerler)
  - Gürültülü veri
    - (hatalı girilmiş değerler)

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

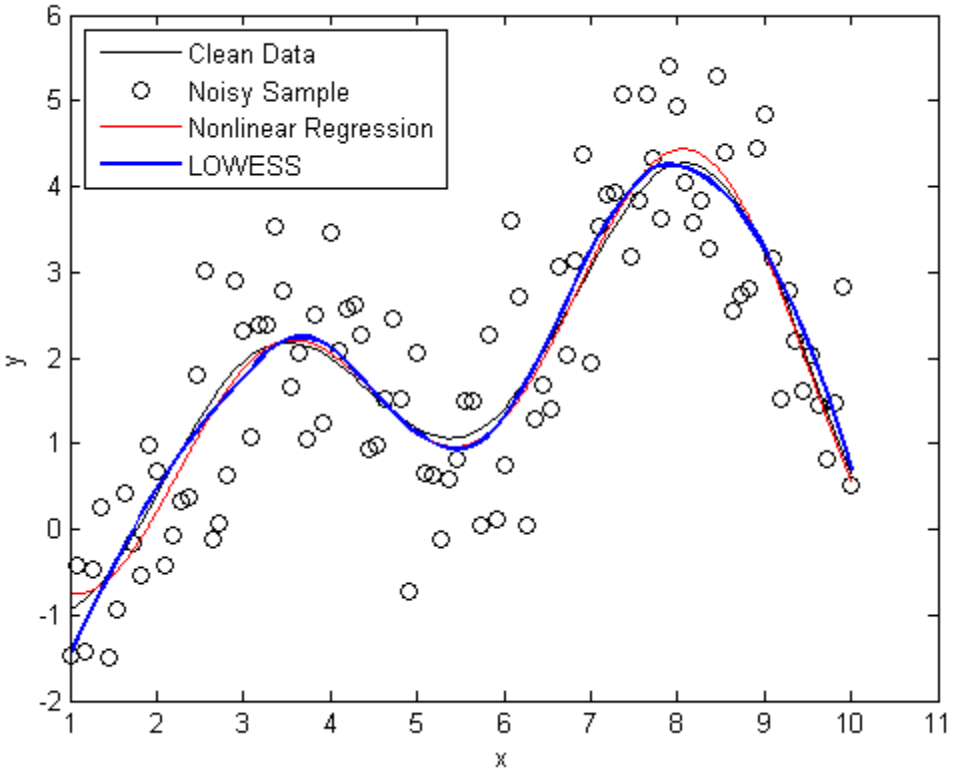
Veri kalitesi
Eksik veri

- Kayıp değerler
  - Bilginin toplanamaması
    (yaşını ya da kilosunu söylemek istemeyen kişi)
  - Bazı bilgilerin olmaması
    (örneğin çocuklar yıllık geliri yoktur)

- Kayıp değerler ile çalışmak
  - Veriyi atmak
  - **Kayıp değeri tahmin etmek**
  - Analiz sırasında eksik veriyi gözardı etmek
  - Olası tüm değerleri kayıp veriyi doldurup, sonuçları karşılaştırmak

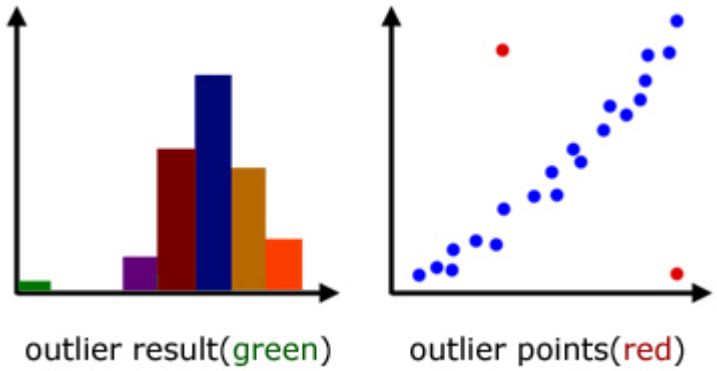
Veri kalitesi
Gürültülü Veri

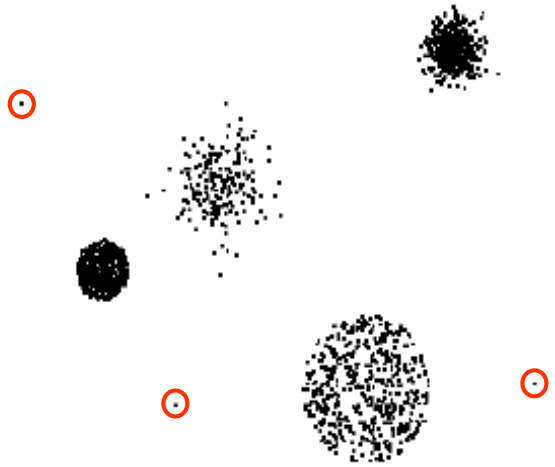
- Asıl değerlerin değişim göstermesine gürültü (noise) denir.
  - Örnek: Telefonda insanın sesi



Veri kalitesi
Aykırı veri

- Aykırı veri diğer verilerden dağılım ya da uzaklık anlamında farklılık gösterir.





Veri önişleme

- Veri temizleme
  - Kayıp değerleri doldurma, gürültü azaltma, aykırı veriyi ayıklama
- Veri dönüştürme
  - Standardizasyon, normalizasyon
- Veri azaltma
  - Bilgi kaybını en aza indirecek şekilde veriyi azaltma

Birleştirme (Aggregation)

- Birden çok özniteliği tek öznitelik olarak ya da birden çok örneği tek örnek olarak ifade etmek

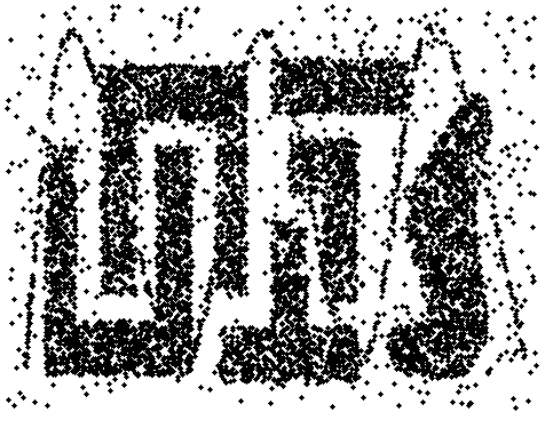
- Amaç
  - Veri küçültmek
    - öznitelik ya da örnek sayısını azaltmak
  - Ölçek değiştirme
    - Şehirleri bölgeler ya da ülkeler cinsinden ifade etmek
  - Daha 'stabil' veri
    - Birleştirilen veri geellikle daha az varyansa sahiptir.

Örnekleme

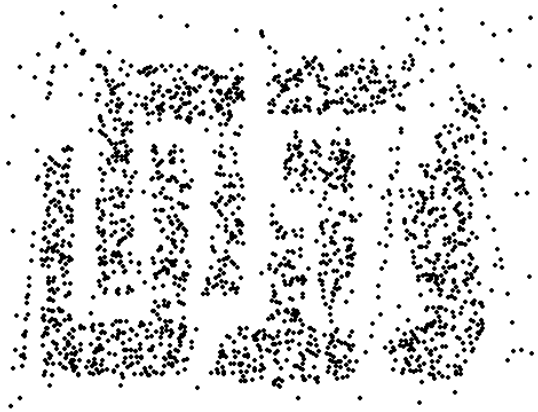
- Veri seçimi için en çok kullanılan yöntemdir.
  - Çoğunlukla ön analiz için tercih edilir.
  - Bazı yöntemler farklı örneklemler üzerinde çalışıp, farklı modelleri birleştirir.

- İstatistikçiler örneklem üzerinde çalışır çünkü tüm veriyi elde etmek masraflı olabilir. Ayrıca tüm veriyle çalışmak hesaplama zamanı açısından sorun yaratabilir.

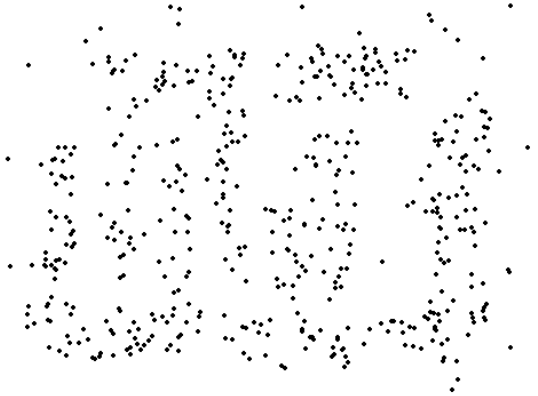
Örneklem Büyüklüğü



8000 gözlem



2000 Gözlem



500 Gözlem

Örneklem Büyüklüğü

- 10 grubun her birinden en az bir tane örnek alabilmek için ne kadar büyük bir örneklem gerekli?