Data Mining: Exploring Data

Lecture Notes for Chapter 3

Introduction to Data Mining

by

Tan, Steinbach, Kumar

What is data exploration?

Özelliklerini daha iyi anlamak için veriler üzerinde ön araştırma yapma işi

- «Data exploration» temel motivasyonlar :
  - Ön işleme veya analiz için doğru aracı seçmeye yardımcı olma
  - İnsanların kalıpları/örüntüleri tanıma yeteneklerinden yararlanma
    - **İnsanlar veri analizi araçları tarafından yakalanmayan kalıpları tanıyabilir**

- Related to the area of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey
  - Seminal book is Exploratory Data Analysis by Tukey
  - A nice online introduction can be found in Chapter 1 of the NIST Engineering Statistics Handbook

  http://www.itl.nist.gov/div898/handbook/index.htm

Techniques Used In Data Exploration

- EDA'da, orijinal olarak Tukey tarafından tanımlandığı gibi
  - Odak noktası görselleştirme (*visualization*) idi
  - Kümeleme ve anormallik tespiti keşif teknikleri (*explatory techniques*) olarak görüldü
  - Veri madenciliğinde, kümeleme ve anormallik tespiti başlıca ilgi alanlarıdır ve sadece keşif amaçlı olarak düşünülmez

- In our discussion of data exploration, we focus on
  - Summary statistics
  - Visualization
  - Online Analytical Processing (OLAP)

Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Summary Statistics

- **Özet istatistikler**, verilerin özelliklerini özetleyen sayılardır
  - Özet özellikler arasında sıklık (**frequency**), konum (**location**) ve yayılma (**spread**) bulunur
    - Examples: location - mean
      spread - standard deviation
  - Özet istatistiklerin çoğu, veriler üzerinden **tek bir geçişte** (**in a single pass through the data**) hesaplanabilir.

Frequency and Mode

- Bir öznitelik değerinin sıklığı, **değerin veri kümesinde var olma yüzdesidir**.
  - Örneğin, "cinsiyet" özniteliği ve temsili bir insan popülasyonu verildiğinde, cinsiyet "kadın" yaklaşık %50 oranında ortaya çıkar.
- Bir özniteliğin modu (mode of an attribute) **en sık görülen öznitelik değeridir**
- Sıklık (frequency) ve mod kavramları tipik olarak **kategorik verilerle** kullanılır

Percentiles

- Sürekli veriler (***continuous data***) için, yüzdelik (*percentile*) kavramı daha kullanışlıdır.

Sıralı veya sürekli bir *x* özniteliği ve 0 ile 100 arasında bir *p* sayısı verildiğinde, *p*. yüzdelik dilim x_p , x'in **gözlemlenen değerlerinin %*p* 'sinden küçük olacak şekilde** bir x değeridir.

- Örneğin, 50. yüzdelik dilim $x_{50\%}$, x'in tüm değerlerinin %50'sinin ondan daha küçük olacağı değerdir.

Measures of Location: Mean and Median

- Ortalama (*mean*), bir nokta kümesinin konumunun en yaygın ölçüsüdür.
- Bununla birlikte, ortalama (mean), **uç değerlere (*outliers*) karşı çok hassastır**.
- Bu nedenle, medyan (*median*) veya kırpılmış ortalama da yaygın olarak kullanılır.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Location: Mean and Median

- **Trimmed mean (**kırpılmış ortalama**):**
  - 0 ile 100 arasında bir yüzde *p* belirlenir, verilerin **üst** ve **alt %(*p / 2*)** 'si atılır ve daha sonra ortalama normal şekilde hesaplanır.
  - Örnek
    - {1,2,3,4,5,90} değerler kümesini düşününüz.
    - What is the mean, median and the **trimmed mean** with *p*=40%?
  - Answer
    - mean=**17.5**
    - median=**3.5**
    - **trimmed mean**(40%)=**3.5**

Measures of Spread: Range and Variance

- Range, maksimum ve minimum arasındaki farktır.
- Varyans veya standart sapma, bir nokta kümesinin yayılmasının (*spread*) en yaygın ölçüsüdür.

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- Fakat, bu da uç değerlere duyarlıdır, bu nedenle **sıklıkla başka ölçüler** kullınılır.

absolute average deviation (AAD)

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

median absolute deviation (MAD)

$$\text{MAD}(x) = median\left(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\right)$$

interquartile range (IQR)

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

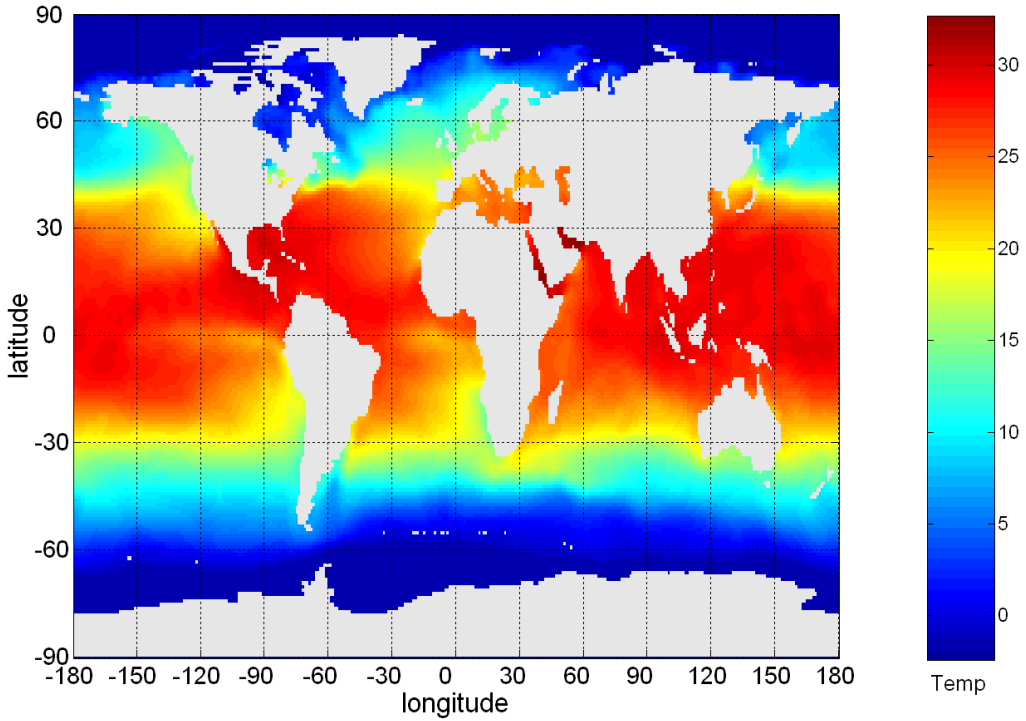
Visualization

Görselleştirme (*Visualization*) verilerin karakteristiklerinin ve veri öğeleri veya öznitelikler arasındaki ilişkilerin analiz edilebilmesi veya raporlanabilmesi için **verilerin görsel veya tablo biçiminde bir biçime dönüştürülmesidir**.

- Verilerin görselleştirilmesi, veri keşfi(*data exploration*) için en güçlü ve çekici tekniklerden biridir.
  - **İnsanlar görsel olarak sunulan** büyük miktarda bilgiyi analiz etme konusunda **gelişmiş bir beceriye** sahiptir.
  - Genel kalıpları ve eğilimleri tespit edebilir
  - Uç değerleri ve alışılmadık kalıpları tespit edebilir

Example: Sea Surface Temperature

- Aşağıda, Temmuz 1982 için Deniz Yüzeyi Sıcaklığı (SST) gösterilmektedir.
  - On binlerce veri noktası tek bir şekilde özetlenmiştir



Representation

- **Bilginin görsel bir formatla eşleştirilmesi**
- Veri nesneleri, öznitelikleri ve veri nesneleri arasındaki ilişkiler, noktalar, çizgiler, şekiller ve renkler gibi **grafiksel öğelere çevrilir.**
- Örnek:
  - Nesneler genellikle **noktalar** olarak temsil edilir
  - Öznitelik değerleri, noktaların **konumu** veya noktaların **özellikleri**, örn. **renk**, **boyut** ve **şekil** olarak gösterilebilir.
  - **Konum** bilgisi kullanılırsa, noktaların ilişkileri, yani gruplar oluşturup oluşturmadıkları veya bir noktanın uç değer olup olmadığı kolayca algılanır.

Arrangement

- Görsel öğelerin bir ekran içinde yerleşimidir
- Verileri anlamanın ne kadar kolay olduğu konusunda **büyük bir fark yaratabilir**
- Örnek:

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

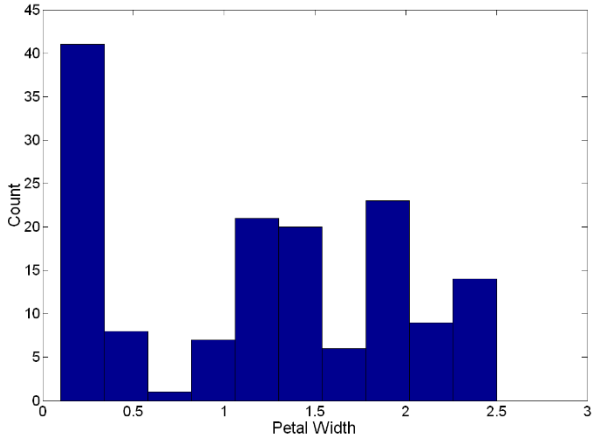
Satırların ve sütunların ilişkilerinin belirgin hale getirildiği altı tane ikili niteliğe (sütun) sahip dokuz nesneden (satır) oluşan bir tablo.

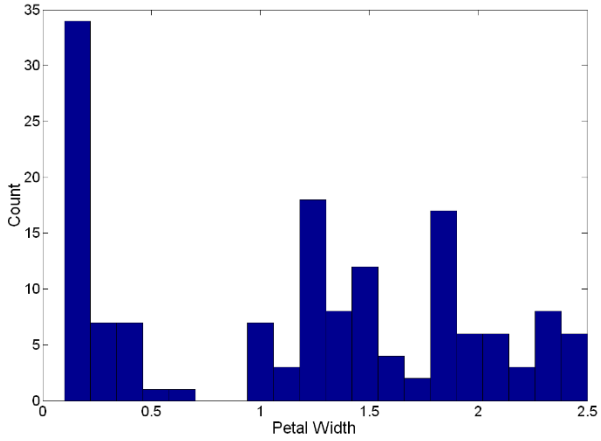
Selection

- **Belirli nesnelerin ve niteliklerin ortadan kaldırılması veya vurgulanmaması**
- Seçim (*selection*), özniteliklerin bir alt kümesinin seçilmesini içerebilir
  - Boyut azaltma (Dimensionality reduction), genellikle boyutların sayısını iki veya üçe düşürmek için kullanılır
  - Alternatif olarak, öznitelik çiftleri (pairs of attributes) düşünülebilir
- Seçim, ayrıca nesnelerin bir alt kümesini (a subset of objects) seçmeyi de içerebilir
  - Ekranın bir bölgesi yalnızca belirli sayıda nokta gösterebilir
  - Örnekleme yapılabilir, ancak seyrek alanlardaki noktalar korunmak istenir

Visualization Techniques: Histograms

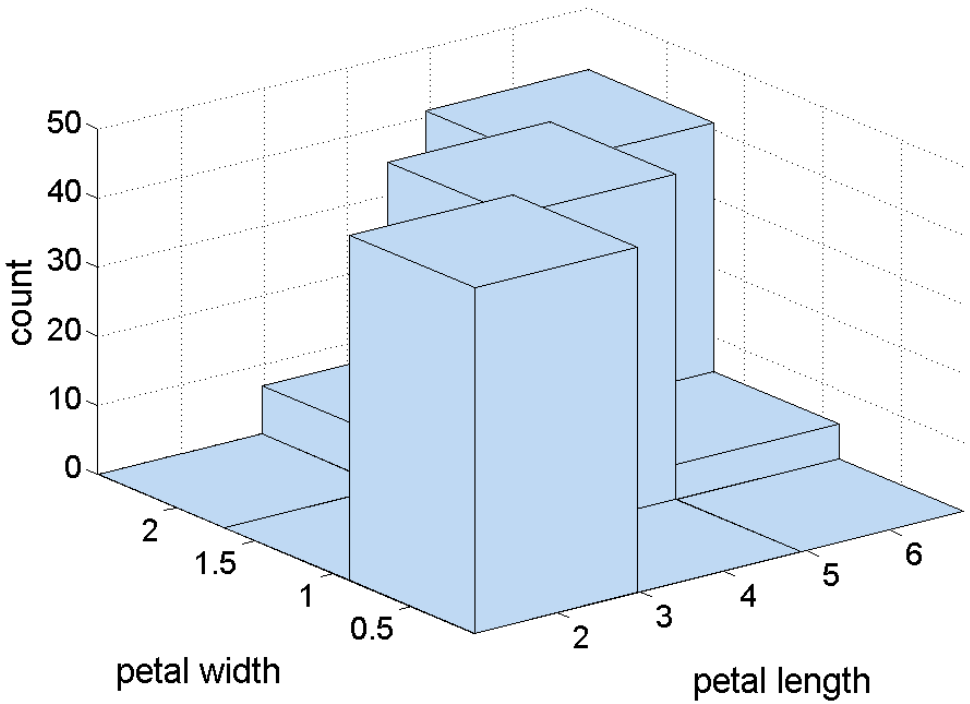
- Histogram
  - Genellikle **tek bir değişkenin değerlerinin dağılımını** gösterir.
  - **Değerler bölmelere (*bins*) dağıtılır** ve her bölmedeki nesnelerin sayısının **çubuk grafiği** gösterilir.
  - Her çubuğun yüksekliği nesnelerin sayısını gösterir
  - Histogramın şekli, bölme sayısına bağlıdır
- Example: Petal Width (10 and 20 bins, respectively)





Two-Dimensional Histograms

- **İki özniteliğin değerlerinin ortak dağılımını (*joint distribution*) gösterir**
- Example: petal width and petal length
  - What does this tell us?



İki boyutlu histogramlar, **iki özelliğin değerlerinin birlikte nasıl oluştuğuna ilişkin ilginç gerçekleri keşfetmek için kullanılabilirken**, görsel olarak daha karmaşıktır.

Çiçeklerin çoğu sadece üç bölmeye düşüyor—**köşegen boyunca olanlar.**

Tek boyutlu dağılımlara bakarak bunu görmek mümkün değil.

Pie Chart

- **Pie Chart (**Pasta Grafik**)**
  - histograma benzer, ancak **tipik olarak** nispeten az sayıda değere sahip **kategorik özelliklerle** kullanılır.
  - göreceli frekansı belirtmek için dairenin göreceli alanını kullanır.
  - teknik yayınlarda daha az sıklıkla kullanılır çünkü göreceli alanların boyutunun değerlendirilmesi zor olabilir

Her üç çiçek türünün de frekansı (sıklığı) aynı

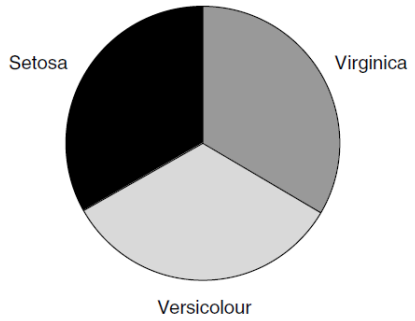
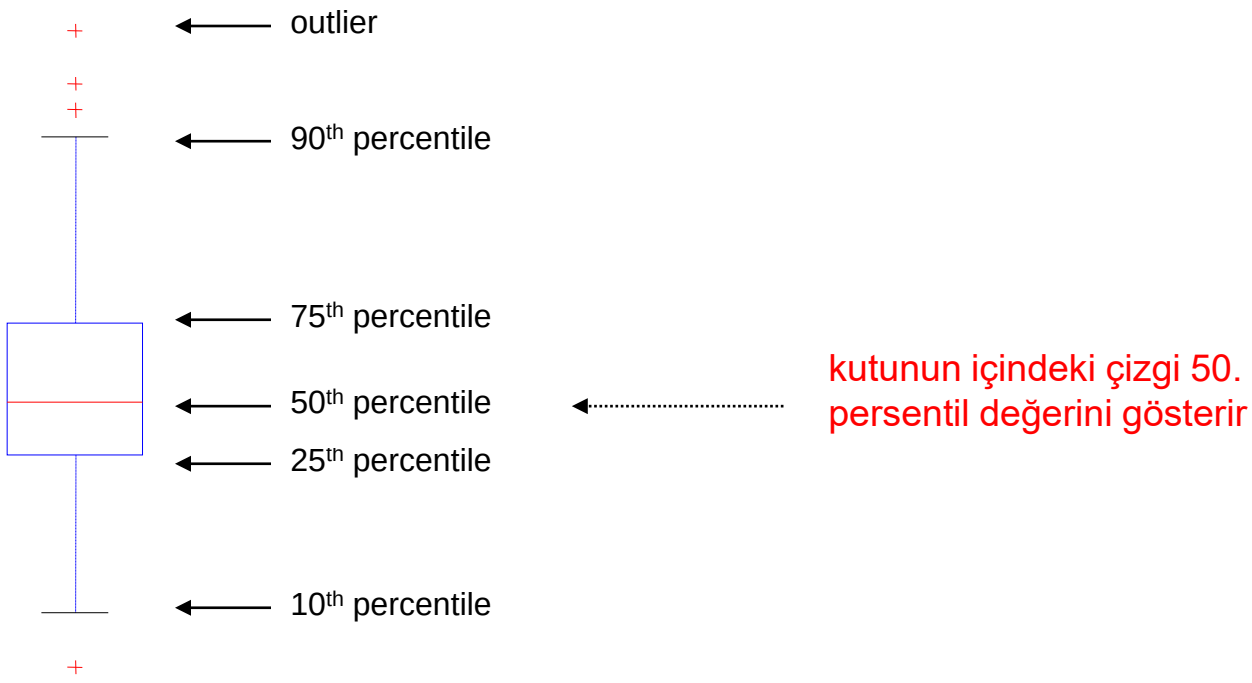


Figure 3.13. Distribution of the types of Iris flowers.

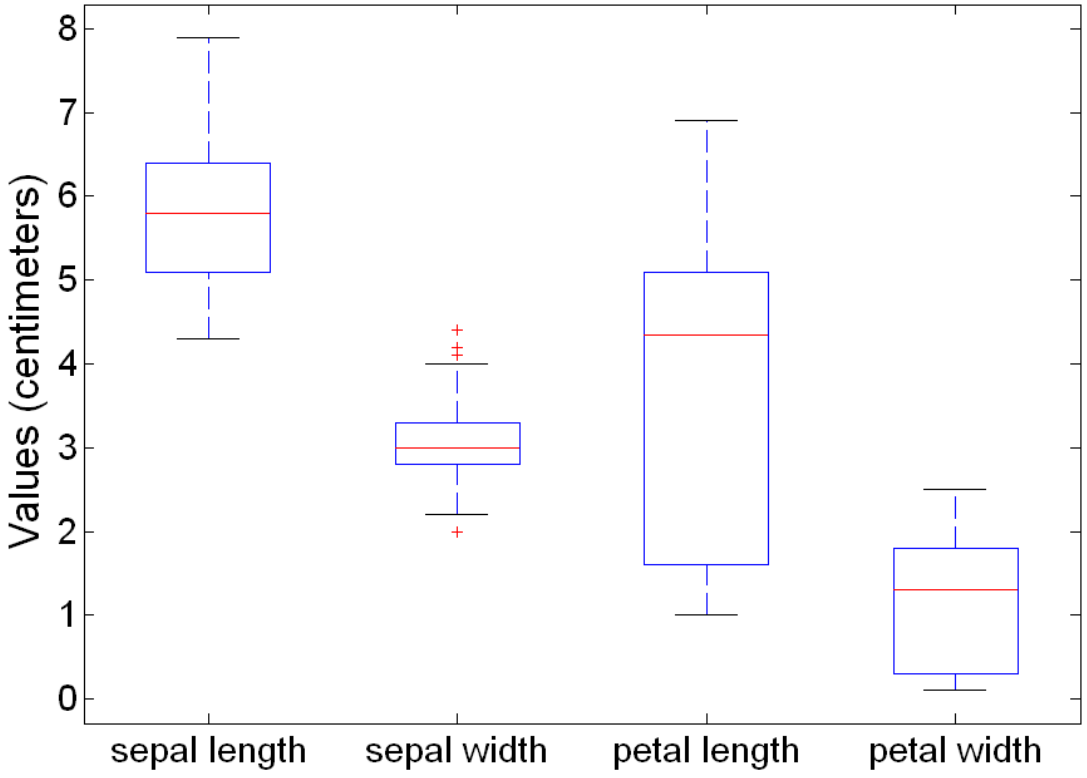
Visualization Techniques: Box Plots

- **Box Plots**
  - J. Tukey tarafından icat edilmiştir
  - Veri dağılımını göstermenin başka bir yolu
  - Aşağıdaki şekil bir kutu grafiğinin temel bölümünü göstermektedir



Example of Box Plots

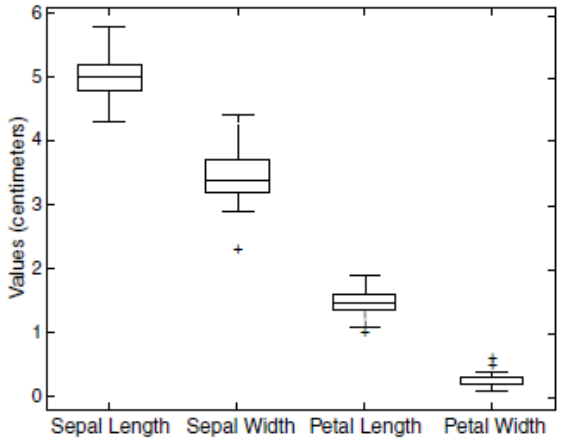
- Öznitelikleri karşılaştırmak için kutu grafikleri kullanılabilir



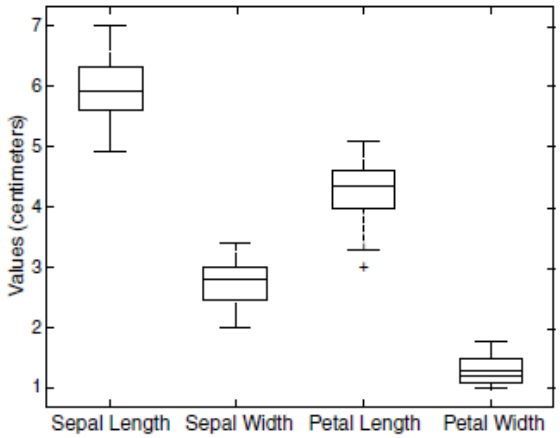
Box plot for Iris attributes

Example of Box Plots

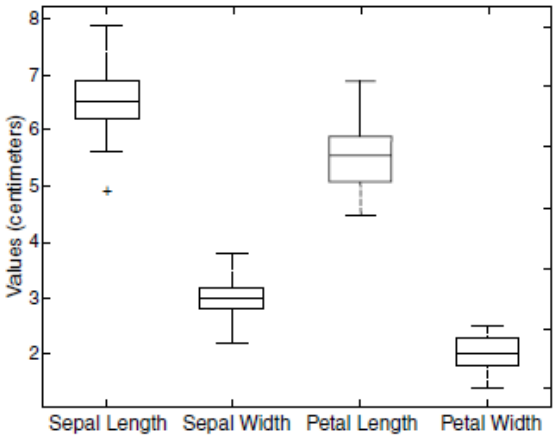
- Kutu grafikleri, **özniteliklerin farklı nesne sınıfları arasında nasıl değiştiğini** karşılaştırmak için de kullanılabilir.



(a) Setosa.



(b) Versicolour.



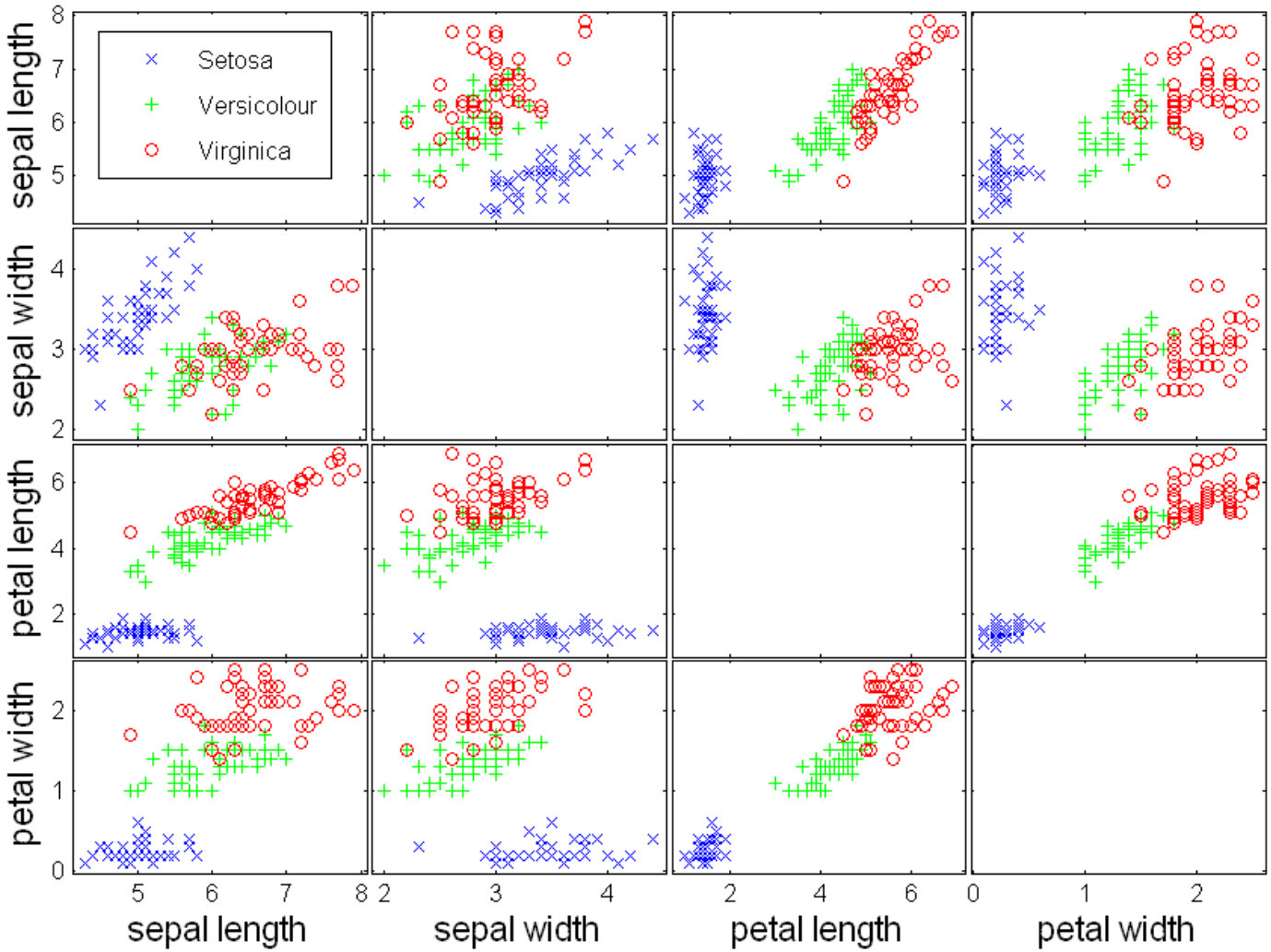
(c) Virginica.

Box plots of attributes by Iris species

Visualization Techniques: Scatter Plots

- Scatter plots
  - Özniteliklerin değerleri **konumu** belirler
  - En yaygın olan iki boyutlu dağılım (*scatter*) grafikleri, ancak **üç boyutlu dağılım grafikleri** olabilir
  - Genellikle, nesneleri temsil eden belirteçlerin (*markers*) boyutu, şekli ve rengi kullanılarak ek öznitelikler görüntülenebilir.
  - **Dağılım grafiği dizileri**, **birkaç öznitelik çiftinin** ilişkilerini kompakt bir şekilde özetleyebilmesi açısından kullanışlıdır.
    - Sonraki slayttaki örnek

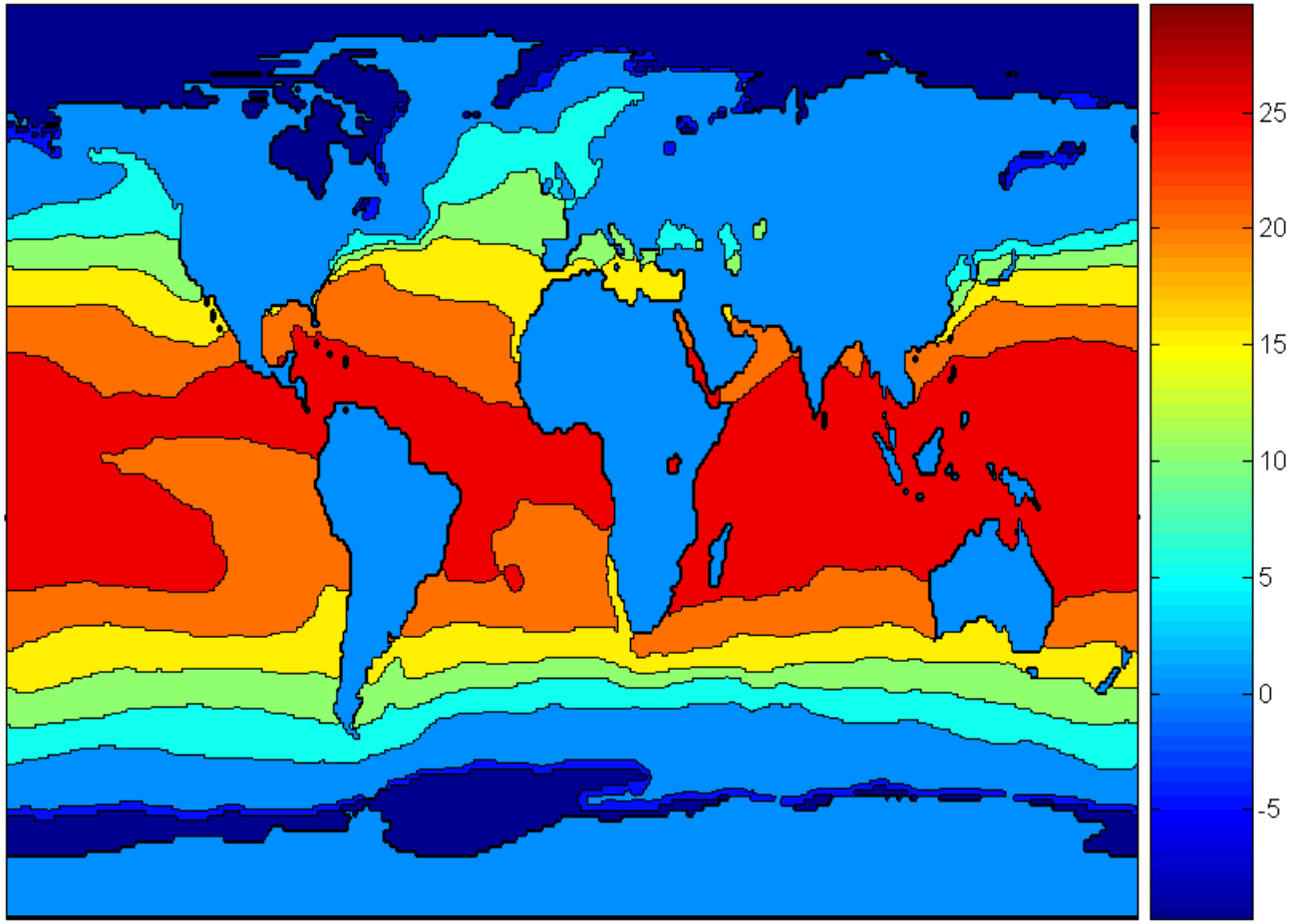
Scatter Plot Array of Iris Attributes



Visualization Techniques: Contour Plots

- Bazı üç boyutlu veriler için, iki öznitelik bir düzlemdeki bir konumu belirtirken, üçüncüsü sıcaklık veya yükselti gibi sürekli bir değere sahiptir.
- Contour plots
  - Uzamsal bir ızgarada (spatial grid) **sürekli bir öznitelik** ölçüldüğünde kullanışlıdır.
  - **Düzlemi benzer değerlere sahip bölgelere ayırırlar**
  - düzlemi, üçüncü özelliğin (sıcaklık, yükselti) değerlerinin yaklaşık olarak aynı olduğu ayrı bölgelere ayırır
  - En yaygın örnek, arazi konumlarının yükseltilerinin kontur haritalarıdır.
  - Ayrıca sıcaklık, yağış, hava basıncı vb. görüntülenebilir.
    - Deniz Yüzeyi Sıcaklığına (SST) bir örnek sonraki slaytta verilmiştir.

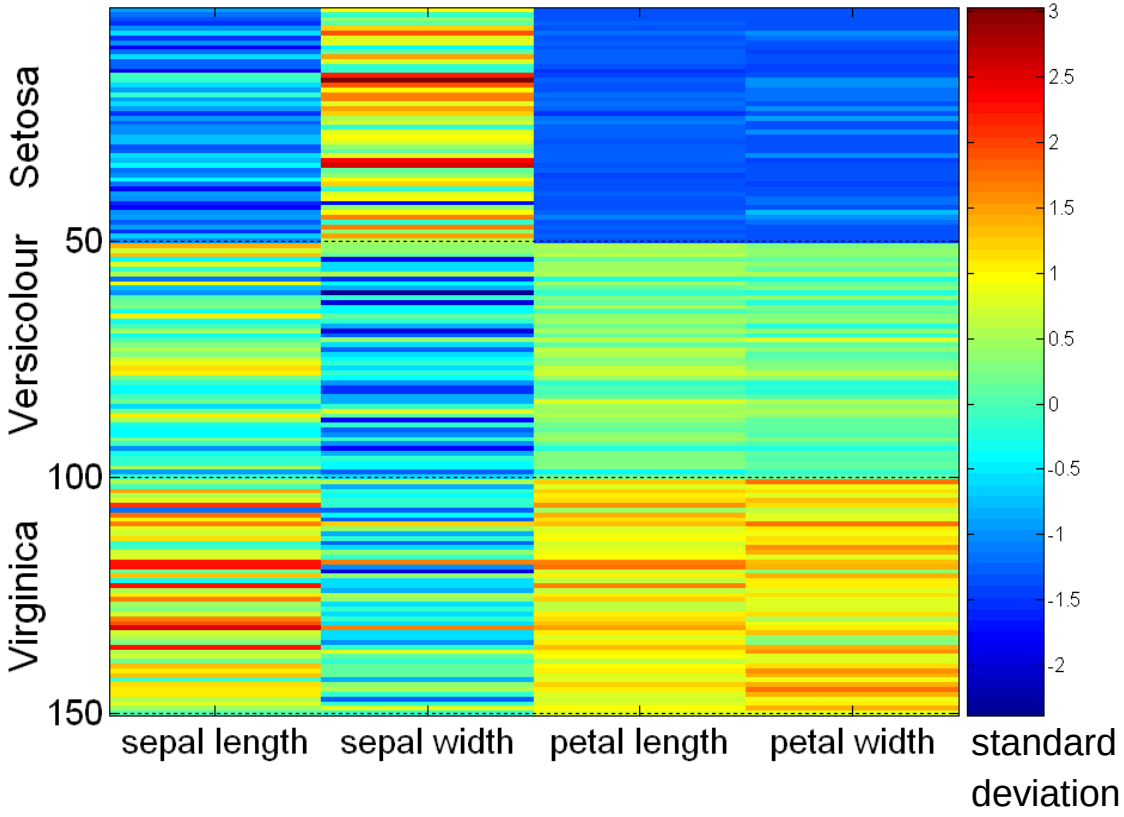
Contour Plot Example: SST Dec, 1998



Visualization Techniques: Matrix Plots

- Matrix plots
  - Veri matrisinin her girdisi görüntüdeki bir piksel ile ilişkilendirilerek bir veri matrisi bir görüntü olarak görselleştirilebilir.
  - **nesneler sınıfa göre sıralanır** (Sınıf etiketleri biliniyorsa)
    - böylece bir sınıfın tüm nesneleri bir arada olur
  - Tipik olarak, **bir özniteliğin grafiğe hakim olmasını önlemek için öznitelikler normalleştirilir**
    - Farklı özniteliklerin farklı aralıkları varsa öznitelikler genellikle sıfır ortalamaya (mean of zero) ve 1 standart sapmaya (standard deviation of 1) sahip olacak şekilde **standartlaştırılır**.
  - Benzerlik veya uzaklık matrislerinin grafikleri, nesneler arasındaki ilişkileri görselleştirmek için de yararlı olabilir.
  - Matris grafiklerinin örnekleri sonraki iki slaytta sunulmuştur.

Visualization of the Iris Data Matrix

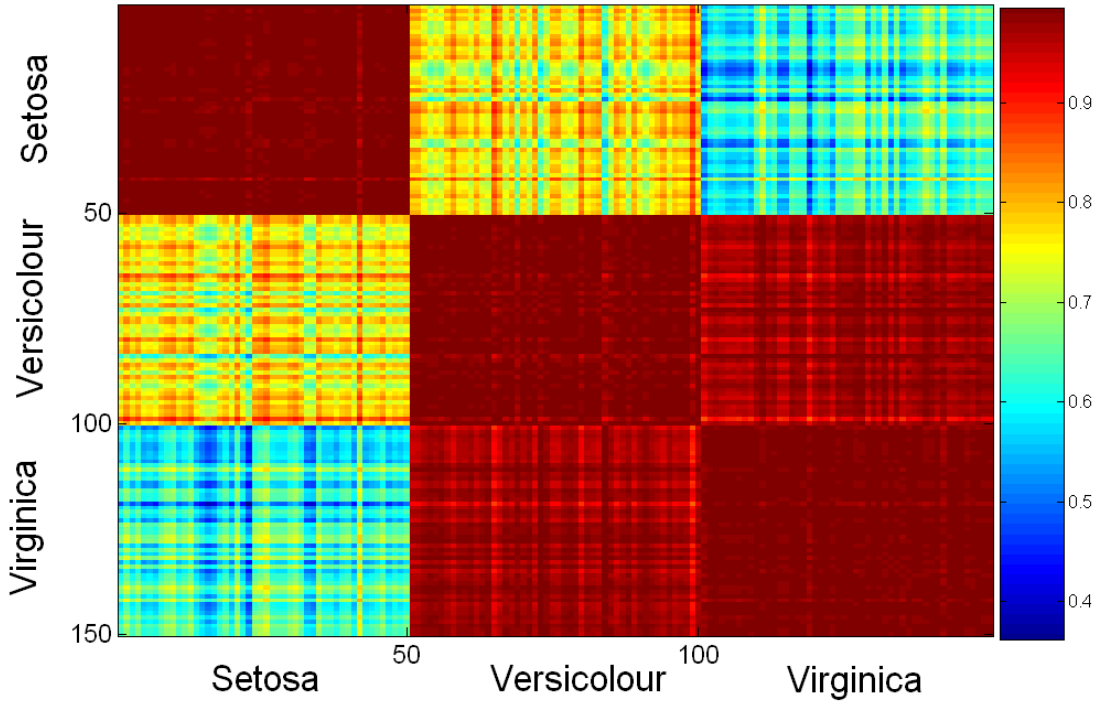


Sütunların ortalama 0 ve standart sapma 1 olacak şekilde standardize edildiği **Iris data matrix** grafiği

İlk 50 sıra **Setosa**, sonraki 50 **Versicolour** ve son 50 **Virginica** türünden Iris çiçeklerini temsil eder.

Setosa çiçeklerinin taç yaprağı (*petal*) genişliği ve uzunluğu **ortalamanın çok altındadır**, Versicolour çiçekleri ise **ortalama** taç genişliği ve **uzunluğuna** sahiptir. Virginica çiçeklerinin taç yaprağı genişliği ve uzunluğu **ortalamanın üzerinde**.

Visualization of the Iris Correlation Matrix



Plot of the **Iris correlation matrix**.

Bir dizi veri nesnesi için yakınlık matrisinin grafiğinde yapı (*structure*) aramak da yararlı olabilir.

Yine, **benzerlik matrisinin satırlarını ve sütunlarını** (sınıf etiketleri bilindiğinde), bir sınıftaki tüm nesnelerin bir arada olması için **sıralamak** yararlıdır.

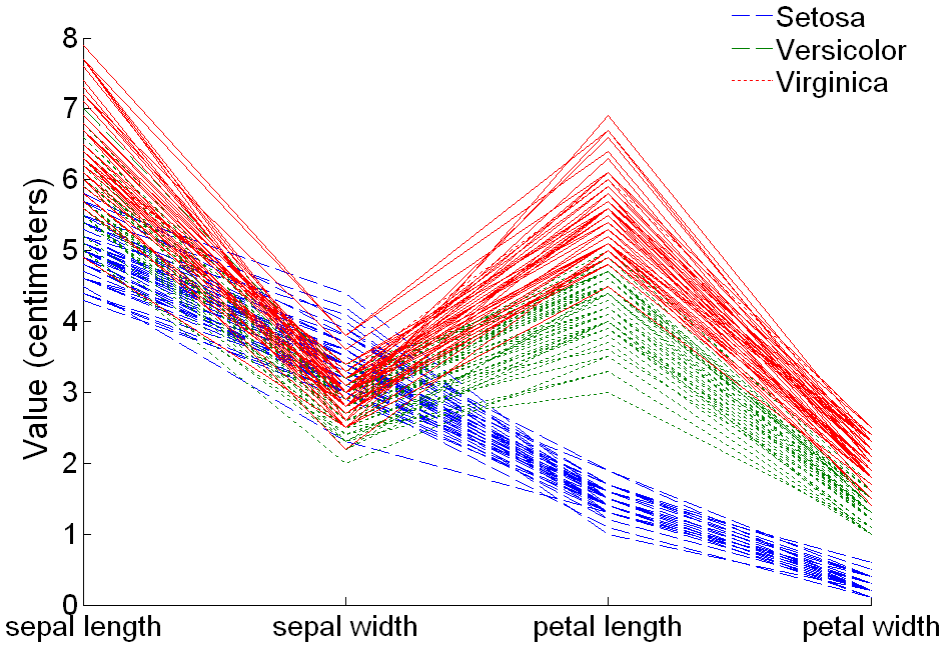
Bu, **her bir sınıfın bağlılığını** ve diğer sınıflardan ayrılığının **görsel bir değerlendirmesine** izin verir.

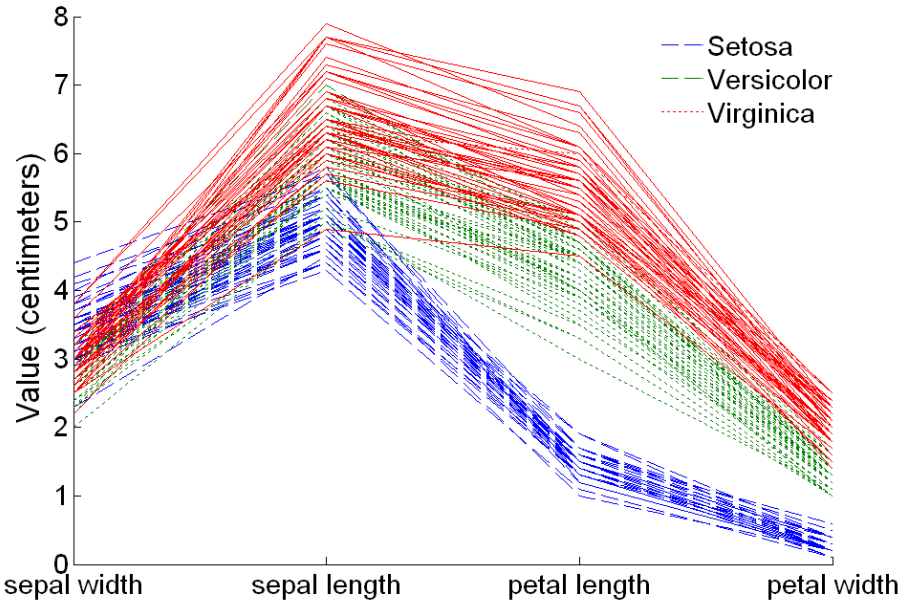
Her gruptaki çiçekler birbirine en çok benziyor, ancak Versicolour ve Virginica Setosa'dan çok birbirine benziyor.

Visualization Techniques: Parallel Coordinates

- Paralel Koordinatlar
  - **Yüksek boyutlu verilerin öznitelik değerlerini** çizmek için kullanılır
  - Dikey eksenler kullanmak yerine bir dizi paralel eksen kullanılır
  - Her **nesnenin öznitelik değerleri**, karşılık gelen her koordinat ekseninde bir nokta olarak çizilir ve noktalar bir çizgi ile bağlanır.
  - Böylece**, her nesne bir çizgi olarak temsil edilir**
  - Çoğu zaman, belli bir nesne sınıfını temsil eden çizgiler, en azından bazı öznitelikler için **birlikte gruplanır**.
  - Özniteliklerin sıralanması, bu tür gruplamaları görmede önemlidir

Parallel Coordinates Plots for Iris Data

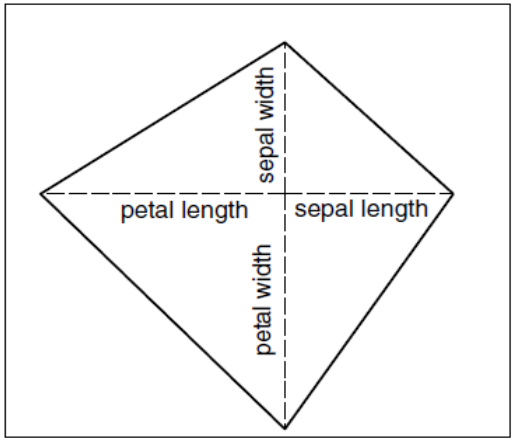




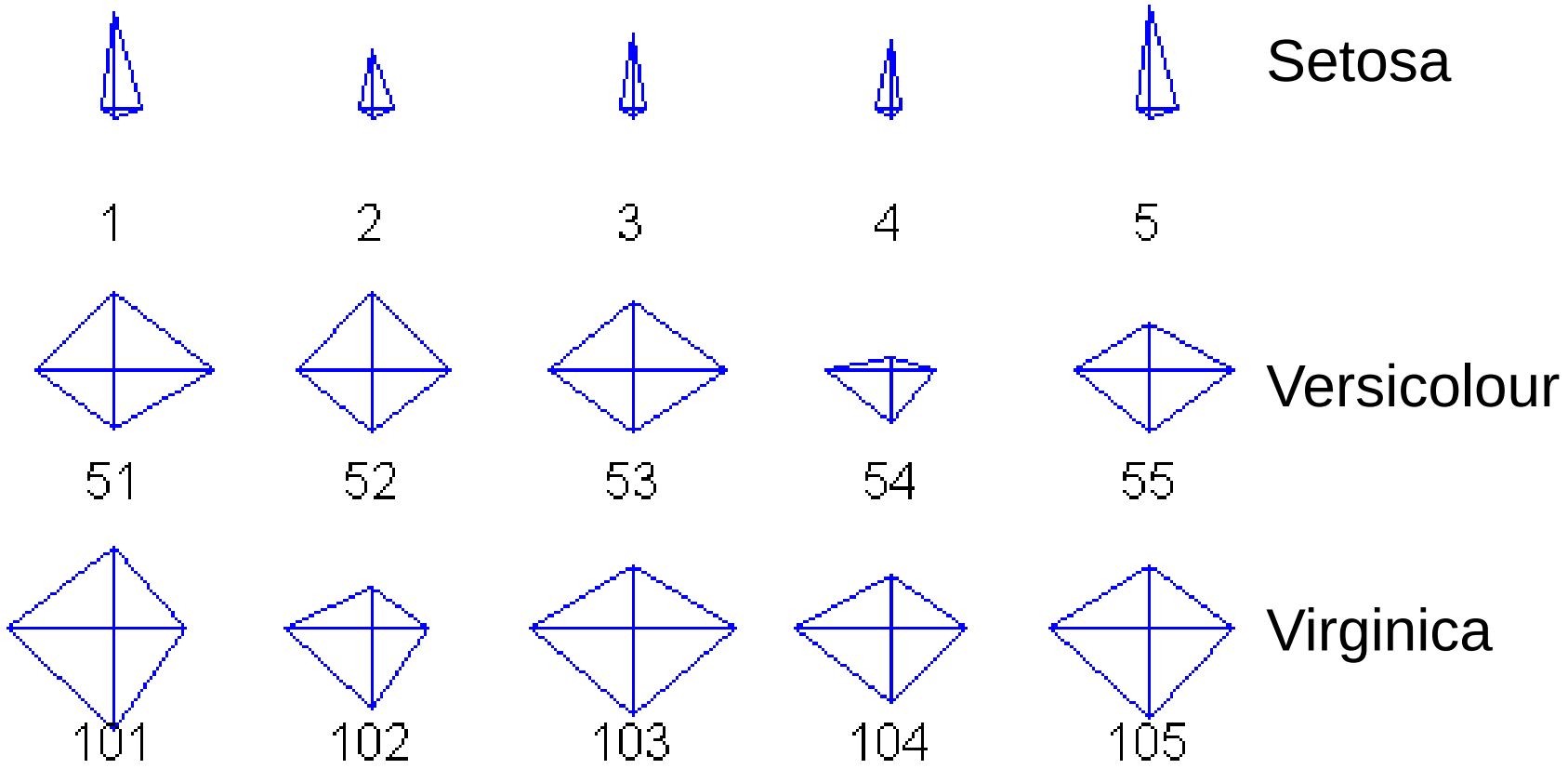
Other Visualization Techniques

- **Star Plots**
  - Paralel koordinatlara benzer yaklaşım, ancak **eksenler merkezi bir noktadan yayılır**
  - Bu teknik, her özellik için bir eksen kullanır.
  - Tipik olarak, tüm öznitelik değerleri [0,1] aralığına eşlenir.
  - Bir nesnenin değerlerini birleştiren çizgi bir **çokgendir**

Iris veri kümesinin 150. çiçeğinin yıldız koordinatları grafiği



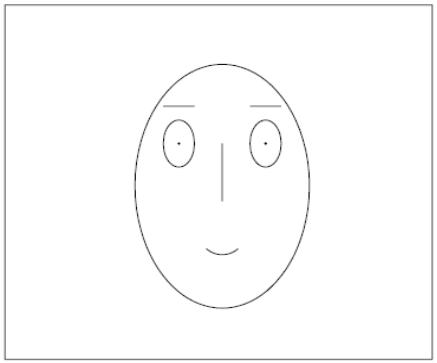
Star Plots for Iris Data



Other Visualization Techniques

- **Chernoff Faces**
  - Herman Chernoff tarafından oluşturulan yaklaşım
  - Bu yaklaşım, her bir özelliği yüzün bir özelliğiyle ilişkilendirir.
  - Her özelliğin değerleri, **karşılık gelen yüz karakteristiğinin görünümünü belirler**.
  - **Her nesne ayrı bir yüz** olur
  - **İnsanın yüzleri ayırt etme yeteneğine** dayanır

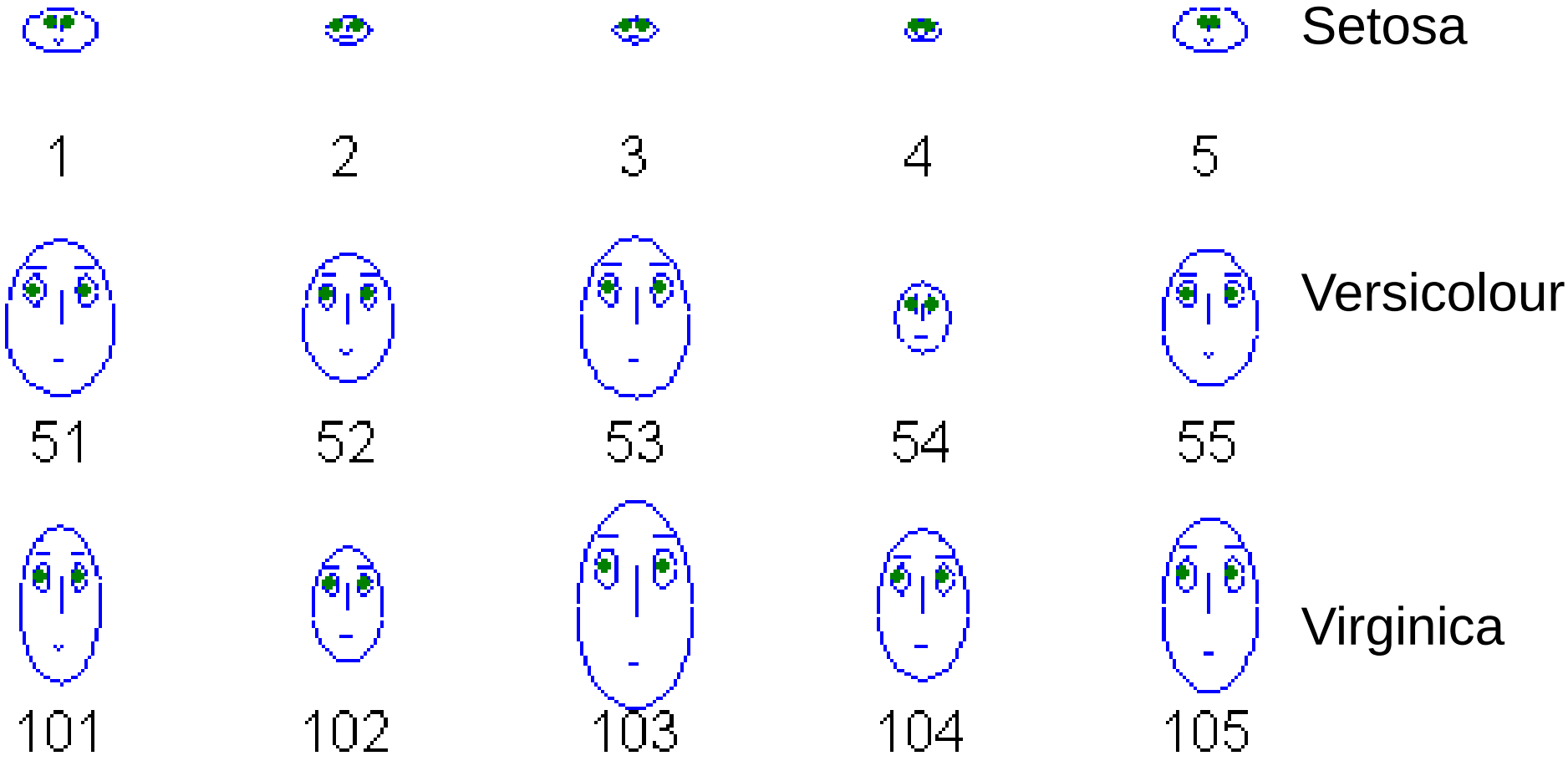
Iris veri kümesinin 150. çiçeğinin Chernoff yüzü



Data Feature	Facial Feature
sepal length	size of face
sepal width	forehead/jaw relative arc length
petal length	shape of forehead
petal width	shape of jaw

Gözler arası genişlik ve ağız uzunluğu gibi yüzün diğer özelliklerine **varsayılan değerler** verilmiştir.

Chernoff Faces for Iris Data



OLAP

- **On-Line Analytical Processing** (OLAP) ilişkisel veritabanının babası olarak bilinen Edgar Frank Codd tarafından önerildi.
- İlişkisel veritabanları verileri tablolara koyarken, **OLAP çok boyutlu bir dizi temsili kullanır**.
  - Verilerin bu tarz temsil edilmesi daha önce istatistik ve diğer alanlarda olmuştur.
- Böyle bir veri temsiliyle daha kolay hale gelen bir dizi veri analizi ve veri keşfi işlemi vardır.

Creating a Multidimensional Array

- Tablo verilerinin çok boyutlu bir diziye dönüştürülmesinde iki temel adım.
  - İlk olarak, hangi özniteliklerin **boyutlar** olacağını ve hangi özniteliğin değerleri çok boyutlu dizide girişler (*entry*) olarak görünen **hedef öznitelik (***target attribute***)** olacağını belirleyin.
    - **Boyut** olarak kullanılan öznitelikler **ayrık değerlere sahip olmalıdır**
    - **Hedef değer tipik olarak bir sayı veya sürekli bir değerdi**r, örneğin bir öğenin maliyeti
  - İkinci olarak, (hedef özniteliğin) değerlerini veya o girdiye karşılık gelen öznitelik değerlerine sahip tüm nesnelerin sayısını toplayarak **çok boyutlu dizideki her girdinin değerini** bulun.

Example: Iris data

- Özniteliklerin (petal length, petal width, and species type) **çok boyutlu bir diziye nasıl dönüştürülebileceği:**
  - *İlk olarak,* petal width ve petal length'i *kategorik değerlere sahip olacak şekilde ayrıklaştırırız: low, medium, ve high*
  - Aşağıdaki tabloyu elde ederiz - count özniteliğine dikkat edin

Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

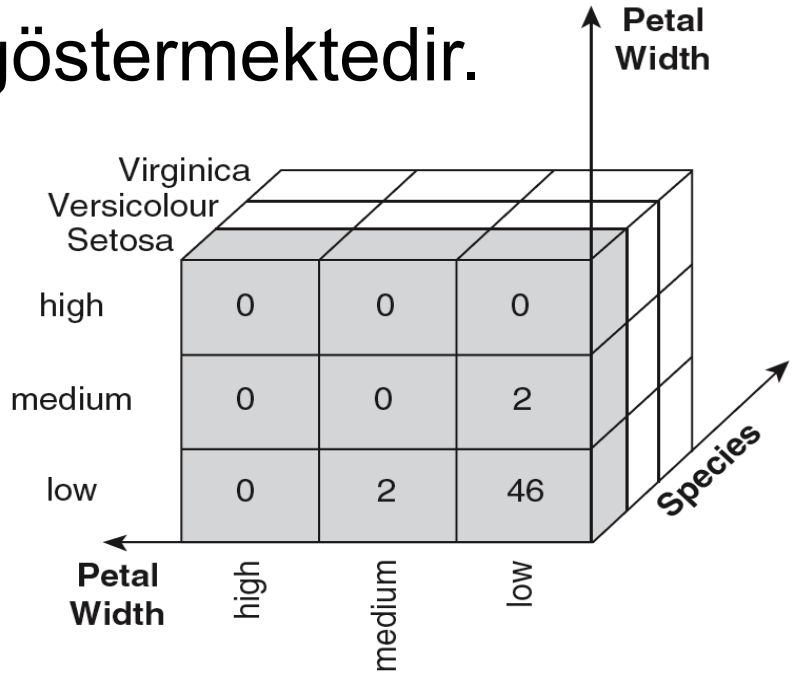
Discretization

- Category boundaries for **petal width**
  - low → [0, 0.75)
  - medium → [0.75,1.75)
  - high →[1.75, ∞)
- Category boundaries for **petal length**
  - low → [0, 2.5)
  - medium → [2.5, 5)
  - high → [5, ∞)

Example: Iris data (continued)

- **Petal width**, **petal length** ve **species type**'ın her benzersiz demeti (tuple), dizinin (array) bir öğesini tanımlar.
- Bu elemana **karşılık gelen sayı değeri atanır**.
- Yandaki şekil sonucu göstermektedir.
- Belirtilmemiş tüm demetler 0'dır.

(*All non-specified tuples are 0.*)



A multidimensional data representation for the Iris data set

Example: Iris data (continued)

- Çok boyutlu dizinin dilimleri aşağıdaki çapraz tablolarla (cross-tabulations) gösterilmiştir.
- Bu tablolar bize ne anlatıyor?

Length \ Width	low	medium	high
low	46	2	0
medium	2	0	0
high	0	0	0

Cross-tabulation of flowers according to petal length and width for flowers of the **Setosa species**.

Length \ Width	low	medium	high
low	0	0	0
medium	0	43	3
high	0	2	2

Cross-tabulation of flowers according to petal length and width for flowers of the **Versicolour species**.

Length \ Width	low	medium	high
low	0	0	0
medium	0	0	3
high	0	3	44

Cross-tabulation of flowers according to petal length and width for flowers of the **Virginica species**.

Bu tablolar, her Iris **türünün petal uzunluğu ve genişliğinin farklı bir değer kombinasyonu ile karakterize olduğunu göstermektedir.**

Setosa çiçekleri düşük genişlik ve uzunluktadır, **Versicolour** çiçekleri orta genişlik ve uzunluktadır ve **Virginica** çiçekleri yüksek genişlik ve uzunluktadır.

OLAP Operations: Data Cube

- Bir OLAP'ın temel işlemi bir veri küpünün (**data cube**) oluşmasıdır
- Veri küpü, verilerin tüm olası toplamlarla (*aggregates*) birlikte çok boyutlu bir temsilidir.
- Olası tüm toplamlar derken, boyutların uygun bir alt kümesini seçerek ve kalan tüm boyutların toplamını alarak sonuçlanan toplamaları kastediyoruz.
- Örneğin, Iris verilerinin **species type** **boyutunu seçersek** ve diğer tüm boyutların toplamını alırsak, sonuç, her biri her bir türün çiçek sayısını veren üç entry'li tek boyutlu bir girdi olacaktır.

Data Cube Example

- Çeşitli tarihlerde bir dizi şirket mağazasında ürünlerin satışını kaydeden bir veri kümesi düşünün.

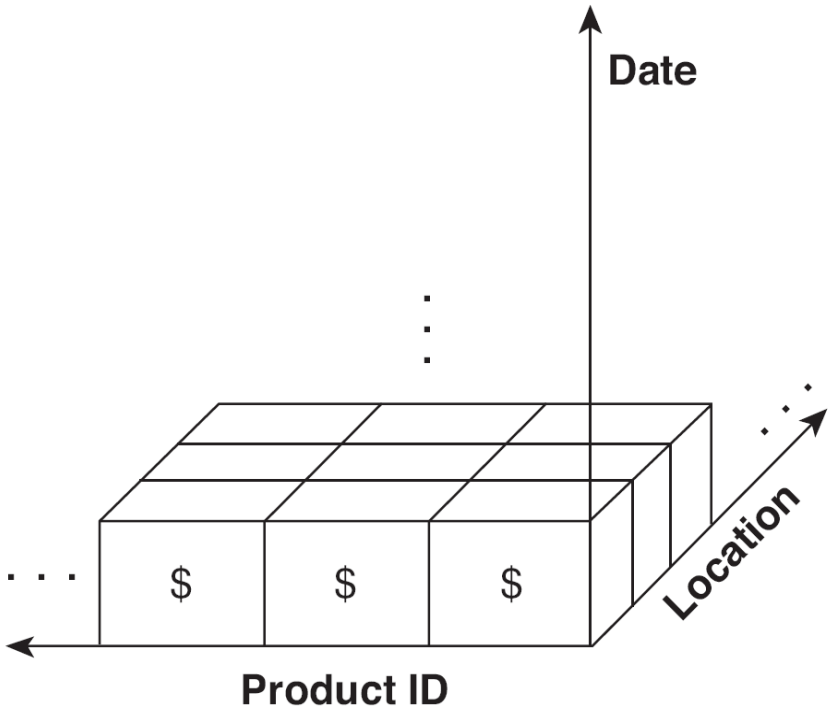
Table 3.11. Sales revenue of products (in dollars) for various locations and times.

Product ID	Location	Date	Revenue
⋮	⋮	⋮	⋮
1	Minneapolis	Oct. 18, 2004	$250
1	Chicago	Oct. 18, 2004	$79
⋮	⋮	⋮	
1	Paris	Oct. 18, 2004	301
⋮	⋮	⋮	⋮
27	Minneapolis	Oct. 18, 2004	$2,321
27	Chicago	Oct. 18, 2004	$3,278
⋮	⋮	⋮	
27	Paris	Oct. 18, 2004	$1,325
⋮	⋮	⋮	⋮

The dimensions of the multidimensional representation are the ***product ID***, ***location***, **and** ***date*** attributes, while the target attribute is the ***revenue***.

Data Cube Example

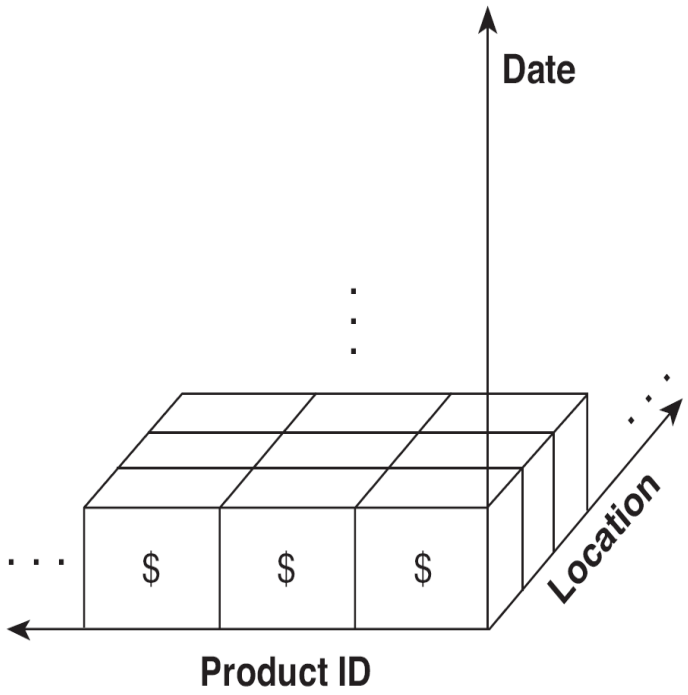
- Bu veriler 3 boyutlu bir dizi olarak gösterilebilir



Multidimensional data representation for sales data.

Data Cube Example

- There are 3 two-dimensional aggregates,
  3 one-dimensional aggregates,
  and 1 zero-dimensional
  aggregate (the overall total)



Data Cube Example (continued)

- Tablo, çeşitli tarih (date) ve ürün (product) kombinasyonları için **tüm konumların toplamının (summing over all locations)** sonucunu gösterir.

Table 3.12. Totals that result from summing over all locations for a fixed time and product.

product ID	date: Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004
1	$1,001	$987	...	$891
⋮	⋮			⋮
27	$10,265	$10,225	...	$9,325
⋮	⋮			⋮

Basit olması için, tüm tarihlerin bir yıl içinde olduğunu varsayalım. Yılda 365 gün ve 1000 ürün varsa, Tablo 3.12'de her ürün-veri çifti için bir tane olmak üzere **365.000 girdi (toplam)** vardır.

We could also specify
- the store location and date and **sum over products**, or
- the location and product and **sum over all dates**.

Data Cube

- Verilerin çok boyutlu temsili (multidimensional representation), tüm olası toplamlarla (aggregates) birlikte **veri küpü (data cube)** olarak bilinir.
- İsmine rağmen, her boyutun büyüklüğünün (öznitelik değerlerinin sayısı) **eşit olması gerekmez**.
- Ayrıca, bir veri küpünün üçten fazla veya daha az boyutu olabilir.
- Daha da önemlisi, bir veri küpü, istatistiksel terminolojide çapraz tablo (**cross-tabulation**) olarak bilinen şeyin bir genellemesidir

Data Cube Example (continued)

- Aşağıdaki şekildeki tablo, iki boyutlu toplamalardan (*two dimensional aggregates*) birini, iki tane tek boyutlu toplamayı (*one-dimensional aggregates*) ve genel toplamı (*overall total*) gösterir.

Table 3.13. Table 3.12 with marginal totals.

product ID	date: Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
1	$1,001	$987	...	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	...	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	...	$631,221	$227,352,127

These totals are the result of **further summing over** either **dates** or **products**.

OLAP Operations: Slicing and Dicing

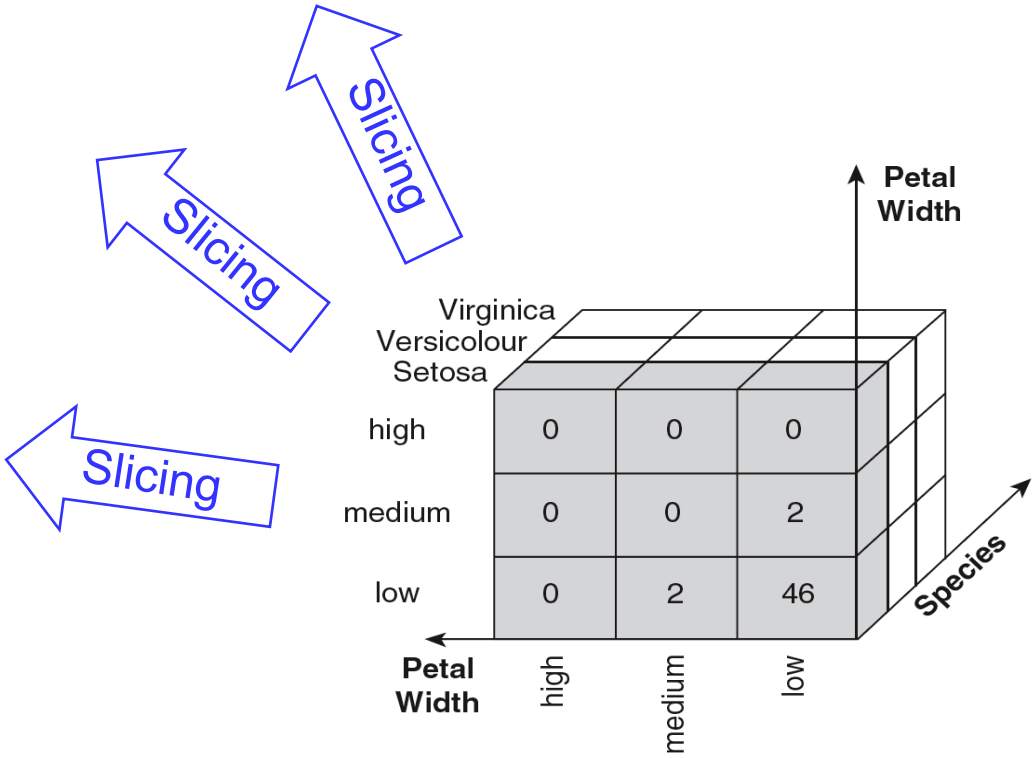
- **Slicing**, bir veya daha fazla boyut için belirli bir değer belirterek **tüm çok boyutlu diziden bir hücre grubu seçmektir.**
- **Dicing**, bir öznitelik değerleri aralığı belirleyerek **bir hücre alt kümesini seçmeyi** içerir.
  - Bu, **tüm diziden bir alt dizi** tanımlamaya eşdeğerdir.
- Uygulamada, her iki işleme de bazı boyutlarda birleştirme (*aggregation*) eşlik edebilir.

Slicing operation

Length	Width: low	Width: medium	Width: high
low	46	2	0
medium	2	0	0
high	0	0	0

Length	Width: low	Width: medium	Width: high
low	0	0	0
medium	0	43	3
high	0	2	2

Length	Width: low	Width: medium	Width: high
low	0	0	0
medium	0	0	3
high	0	3	44



OLAP Operations: Roll-up and Drill-down

- Öznitelik değerleri genellikle **hiyerarşik bir yapıya** sahiptir.
  - Her tarih; bir yıl, ay ve hafta ile ilişkilendirilir.
  - Bir konum; kıta, ülke, eyalet (il vb.) ve şehir ile ilişkilidir.
  - Ürünler; giyim, elektronik ve mobilya gibi çeşitli kategorilere ayrılabilir.
- Bu kategorilerin genellikle iç içe geçtiğini ve bir ağaç (*tree*) veya kafes (*lattice*) oluşturduğunu unutmayın.
  - Bir yıl, günleri içeren ayları içerir
  - Bir ülke, eyaletleri içerir ve onlar da şehirleri içerir.

OLAP Operations: Roll-up and Drill-down

- Bu hiyerarşik yapı, **roll-up** (yuvarlama) ve **drill-down** (detaya inme) işlemlerine imkan tanır.
  - Satış verileri için, satışları bir aydaki tüm tarihlerdekileri toplayarak birleştirebiliriz. (roll up)
  - Tersine, zaman boyutunun aylara bölündüğü verilerin bir görünümü verildiğinde, aylık satış toplamlarını detayına inerek günlük satış toplamlarına geçebiliriz. (drill down)
    - Elbette **bu**, temel satış verilerinin günlük ayrıntı düzeyinde (*daily granularity*) mevcut olmasını **gerektirir**.
  - Aynı şekilde, konum (location) veya ürün numarası (product ID) özelliklerinde roll up veya drill-down yapılabilir.