

BIG DATA ANALYTICS

WEEK-06 | Exploring Data

**Yonsei University
Jungwon Seo**

Data 탐색이란?

- 데이터의 특성을 더 잘 이해하기 위한 데이터의 예비 탐색
- 데이터 탐색의 주요 동기
 - 전처리 또는 분석에 적합한 도구를 선택하도록 지원
 - 인간의 능력을 활용하여 패턴 인식이 가능하게 하기 위함
 - 인간은 데이터 분석 도구로 캡처되지 않은 패턴을 인식 가능
 - 극단적인 예: 말투 파악
- 탐색적 데이터 분석(EDA)
 - 통계학자 John Tukey에 의해 창시
 - <https://www.itl.nist.gov/div898/handbook/index.htm>

데이터 탐색에 사용되는 기술

- Tukey의 정의에 의하면 EDA는
 - 시각화에 중점을 둠
 - 클러스터링 및 이상 감지하는 탐색 기술로 간주
 - 데이터 마이닝에서 클러스터링 및 이상 감지하는 단순한 탐색이상의 주요 관심 영역
- 주요 내용
 - 요약 통계
 - 시각화
 - 온라인 분석 처리 (OLAP)

아이리스 샘플 데이터 세트

- 많은 탐색 데이터 기술이 Iris Plant 데이터 세트로 설명가능
 - UCI Machine Learning Repository에서 획득 가능
 - <http://www.ics.uci.edu/~mlearn/MLRepository.html>
 - 통계 학자 Douglas Fisher에 의해 정립
 - 세 가지 꽃 종류 (클래스) :
 - Setosa
 - Virginica
 - Versicolour
 - 4 가지 (non-class) 속성
 - 분리 폭(width)과 길이(length)
 - 꽃잎 폭(width)과 길이(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

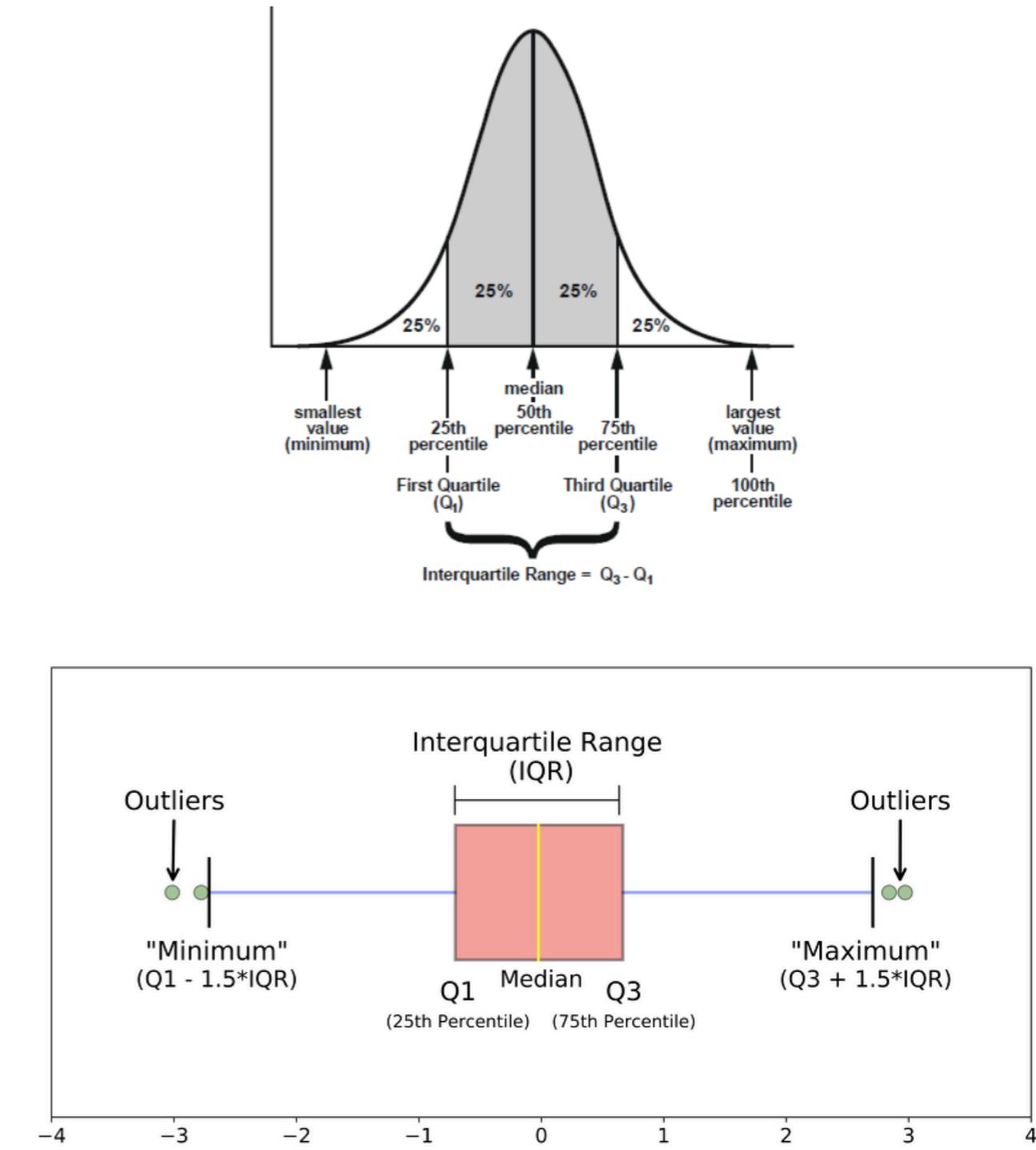
- 요약 통계는 데이터의 속성을 요약하는 숫자들
 - 요약 된 속성에는 빈도, 위치 및 분포가 포함됩니다
 - 예 : 위치-평균, 분포-표준 편차
- 대부분의 요약 통계는 한 번에 계산 가능

빈도수와 최빈값

- 속성 값의 빈도는 데이터 세트에서 값이 발생하는 시간의 백분율
 - 성별이라는 속성에서 여성이라는 값은 50% 발생함
- 최빈값: 특정 속성에서 가장 빈번하게 나타나는 값
 - 한국의 거주하는 사람들의 국적
 - 최빈값?
- 빈도수와 최빈값은 주로 범주형(categorical) 데이터에 사용됨
 - 키 같은 경우를 예를들면, 최빈값보다 평균이 더 적합 (170.1, 170.2, 170.3 ...)

백분위수 (Percentiles)

- 연속 데이터에는 백분위수의 개념이 더 유용
- 순서형 또는 연속형 속성 x 와 0에서 100 사이의 숫자 p 를 가 주어졌을 때, p 번째 백분위 x 의 값은 X_p , 즉 관측값의 $p\%$ 가 X_p 보다 작음
- 예를 들어, x 의 50번째 백분위수 $X_{50\%}$ 는 x 의 모든 값의 50%가 $X_{50\%}$ 보다 작음



Measures of Location: Mean and Median

- 평균값은 점 집합의 위치를 나타내는 가장 일반적인 측정값
- 그러나 평균값은 특이치에 매우 민감
- 따라서 중위값이나 다듬은(trimmed) 평균이 일반적으로 사용됨

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

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

Measures of Spread: Range and Variance

- 범위는 최대값과 최소값의 차이
- 분산 또는 표준 편차는 점 집합의 분포에 대한 가장 일반적인 측정 값

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

- 그러나 이 또한 특이치에 민감하기 때문에 다른 방법을 사용하는 경우가 많음

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

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

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

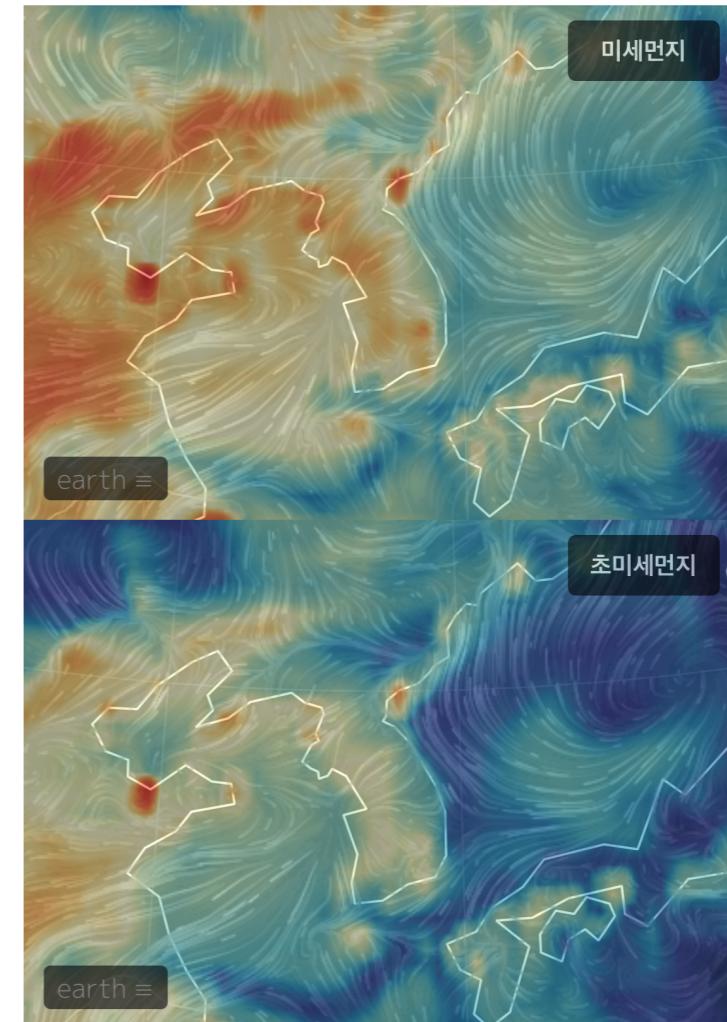
Visualization

- 시각화는 데이터의 특성과 데이터 항목이나 속성 간의 관계를 분석하거나 보고할 수 있도록 데이터를 그림 또는 표 형식으로 변환하는 것
- 데이터의 시각화는 데이터 탐색을 위한 가장 강력하고 설득력 있는 기술 중 하나
 - 인간은 시각적으로 제시되는 대량의 정보를 분석할 수 있는 잘 발달된 능력을 가지고 있음
 - 일반적인 패턴과 트렌드를 감지할 수 있음
 - 특이치와 특이한 패턴을 감지할 수 있음

시각화 예시

국가	확진자	사망자	완치	사망 (%)	완치 (%)	발생률*
1 미국 🇺🇸	1,212,900 (+20,960)	69,921 (+1,206)	188,068 (+9,397)	5.8	15.5	3,664
2 스페인 🇪🇸	248,301	25,428	151,633	10.2	61.1	5,311
3 이탈리아 🇮🇹	211,938 (+1,221)	29,079 (+195)	82,879 (+1,225)	13.7	39.1	3,505
4 영국 🇬🇧	190,584 (+3,985)	28,734 (+288)	344	15.1	0.2	2,807
5 독일 🇩🇪	166,152 (+407)	6,993 (+127)	135,100 (+2,400)	4.2	81.3	1,983
6 러시아 🇷🇺	145,268	1,356	18,095	0.9	12.5	995
7 프랑스 🇫🇷	131,863 (+576)	25,201 (+306)	51,371 (+587)	19.1	39.0	2,020
8 터키 🇹🇷	127,659 (+1,614)	3,461 (+64)	68,166 (+5,015)	2.7	53.4	1,514
9 브라질 🇧🇷	108,620 (+6,794)	7,367 (+316)	45,815 (+2,824)	6.8	42.2	511
10 이란 🇮🇷	98,647	6,277	79,379	6.4	80.5	1,174
11 중국 🇨🇳	82,881 (+1)	4,633	77,944	5.6	94.0	58
12 캐나다 🇨🇦	60,772 (+1,298)	3,854 (+172)	26,017 (+1,109)	6.3	42.8	1,610
13 벨기에 🇧🇪	50,267	7,924	12,378	15.8	24.6	4,337
14 페루 🇵🇪	47,372 (+1,444)	1,344 (+58)	14,427 (+877)	2.8	30.5	1,437

널스쿨



05월 05일
15:00

미세미세 지도 안양대연구소 일본기상청 널스쿨

Representation

- 정보를 시각적 형식에 매핑하는 방법
- 데이터 객체, 그 속성, 데이터 객체 간의 관계를 점, 선, 모양, 색상 등의 그래픽 요소로 변환
- 예:
 - 객체는 종종 점으로 표현
 - 속성 값은 점의 위치 또는 점의 특성(예: 색상, 크기 및 모양)으로 표현 가능
 - 위치를 사용하는 경우, 점의 관계, 즉 점들이 그룹을 형성하는지, 점이 특이치인지 쉽게 인식 가능

Arrangement

- 시각적 요소를 하는 방법
- 단순한 배치차이가 데이터의 이해도를 크게 좌우
- 예 :

	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

Selection

- 특정 대상과 속성의 제거 또는 강조
- 속성의 하위 집합을 선택
 - 차원 축소는 종종 차원 수를 2 개 또는 3 개로 줄이기 위해 사용
 - 또는 속성 쌍을 고려
- 객체의 하위 집합을 선택
 - 데이터의 패턴은 유지하며 샘플링



8000 points



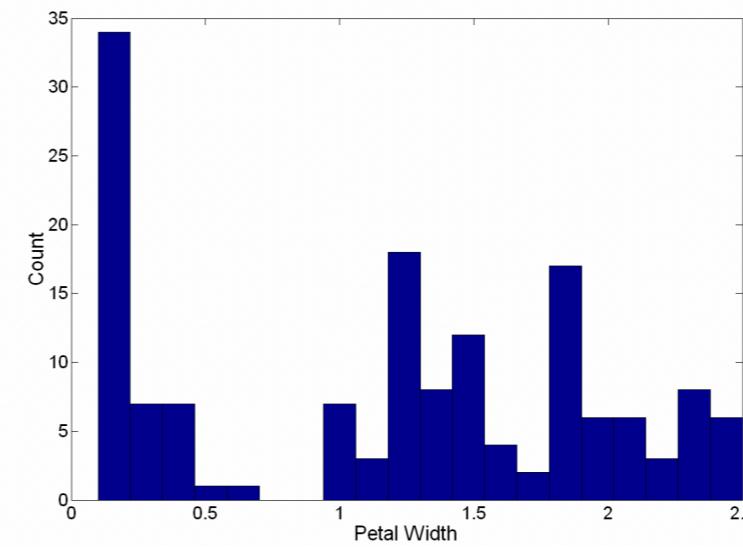
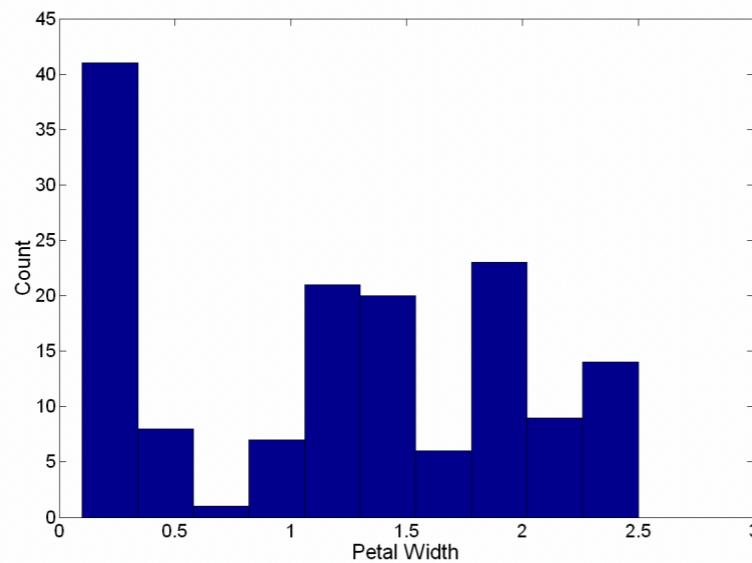
2000 Points



500 Points

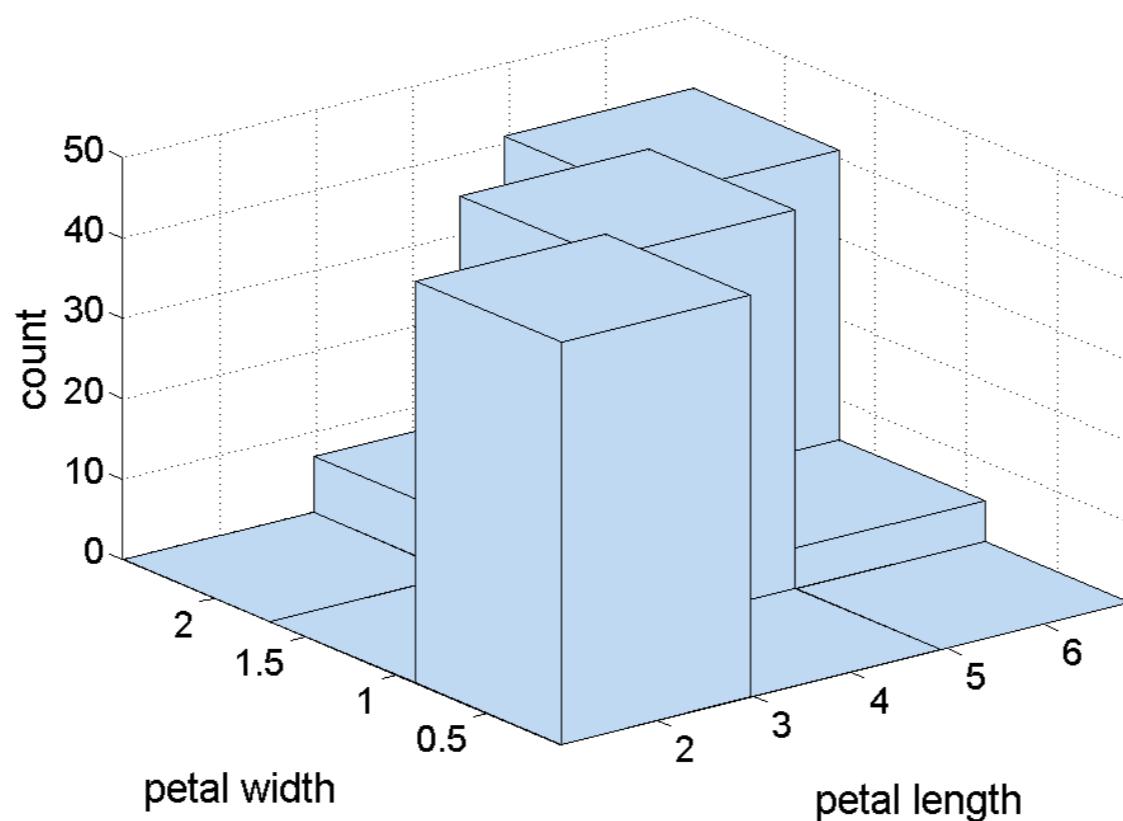
시각화 기법: Histograms

- 히스토그램
 - 일반적으로 단일 변수의 값 분포를 표시함
 - 값을 빈(bin)으로 나누고 각 빈에 있는 개체 수에 대한 막대 그래프를 표시
 - 각 막대의 높이는 객체의 수를 표현
 - 히스토그램의 모양은 빈의 수에 따라 달라짐
- 예: 꽃잎 너비(bin 10개, bin 20개)



Two-Dimensional Histograms

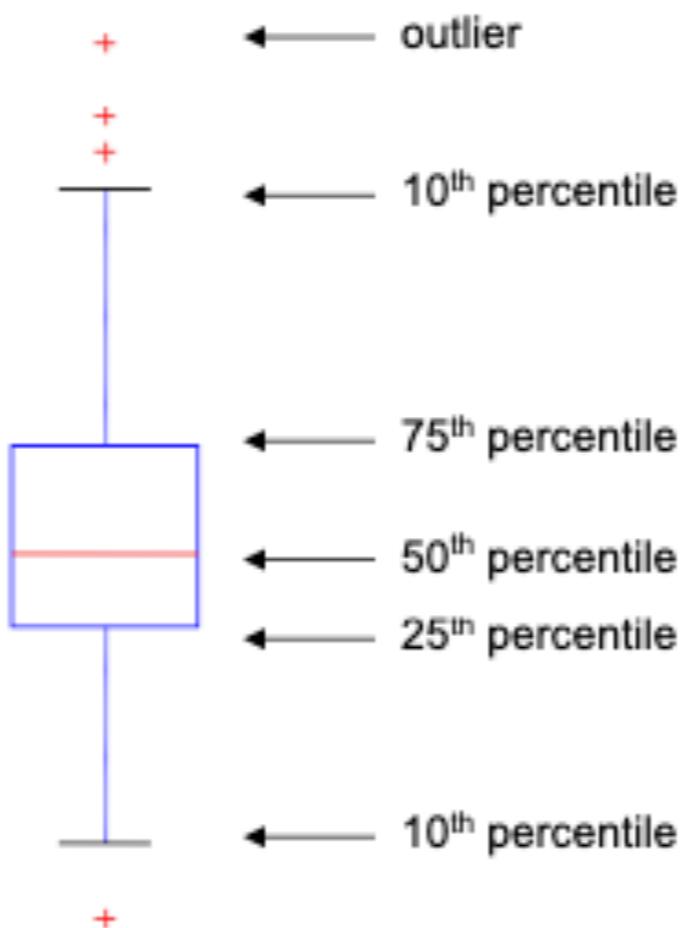
- 두 속성의 값의 공동 분포 표시
- 예제: 꽃잎 폭 및 꽃잎 길이
- 아래 히스토그램을 통해 무엇을 알수 있을까?



시각화 기법: Box Plots

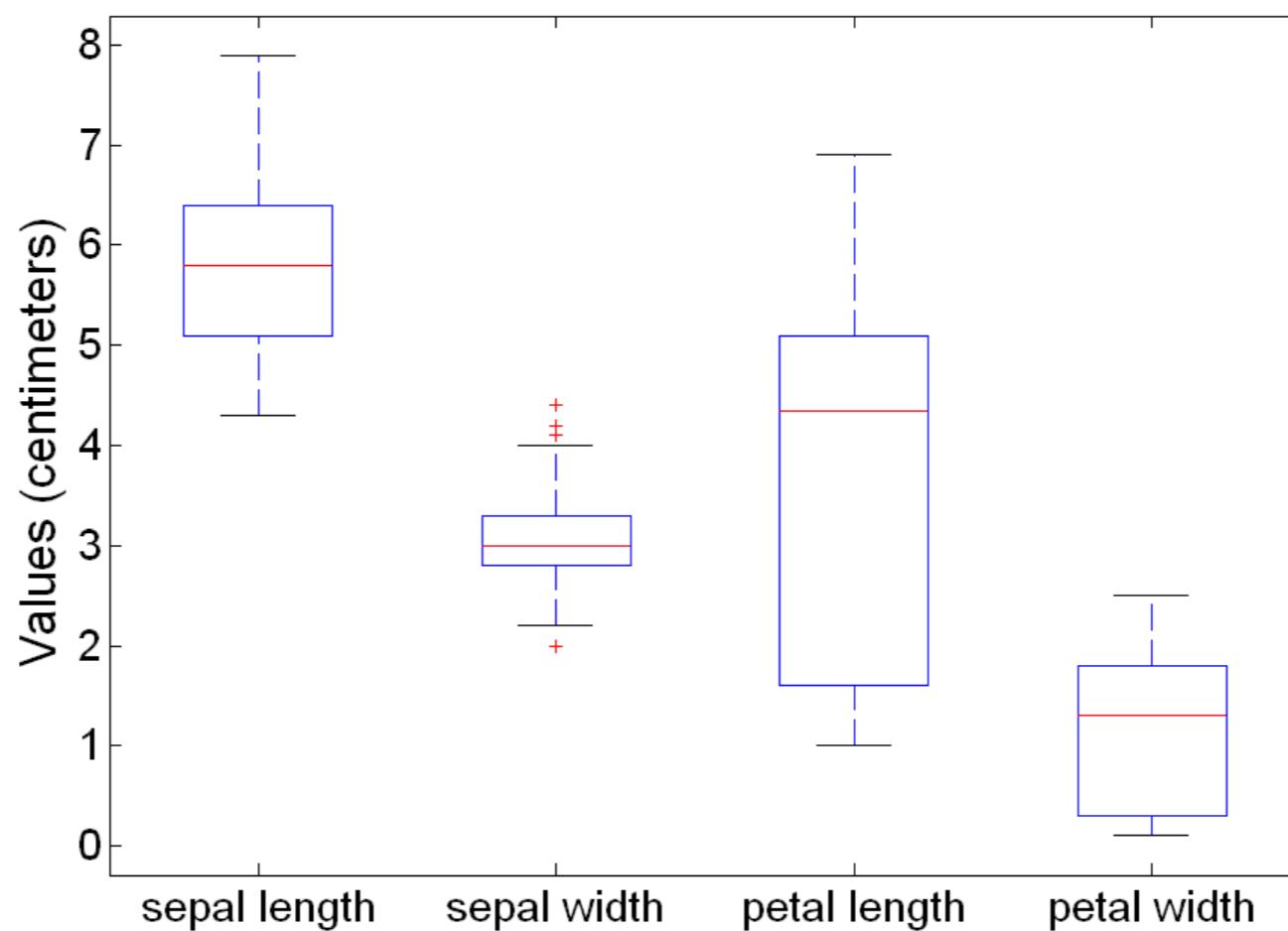
- Box Plots

- J. Tukey에 의해 개발
- 데이터 분포를 표시하는 또 다른 방법



Box Plot 예제

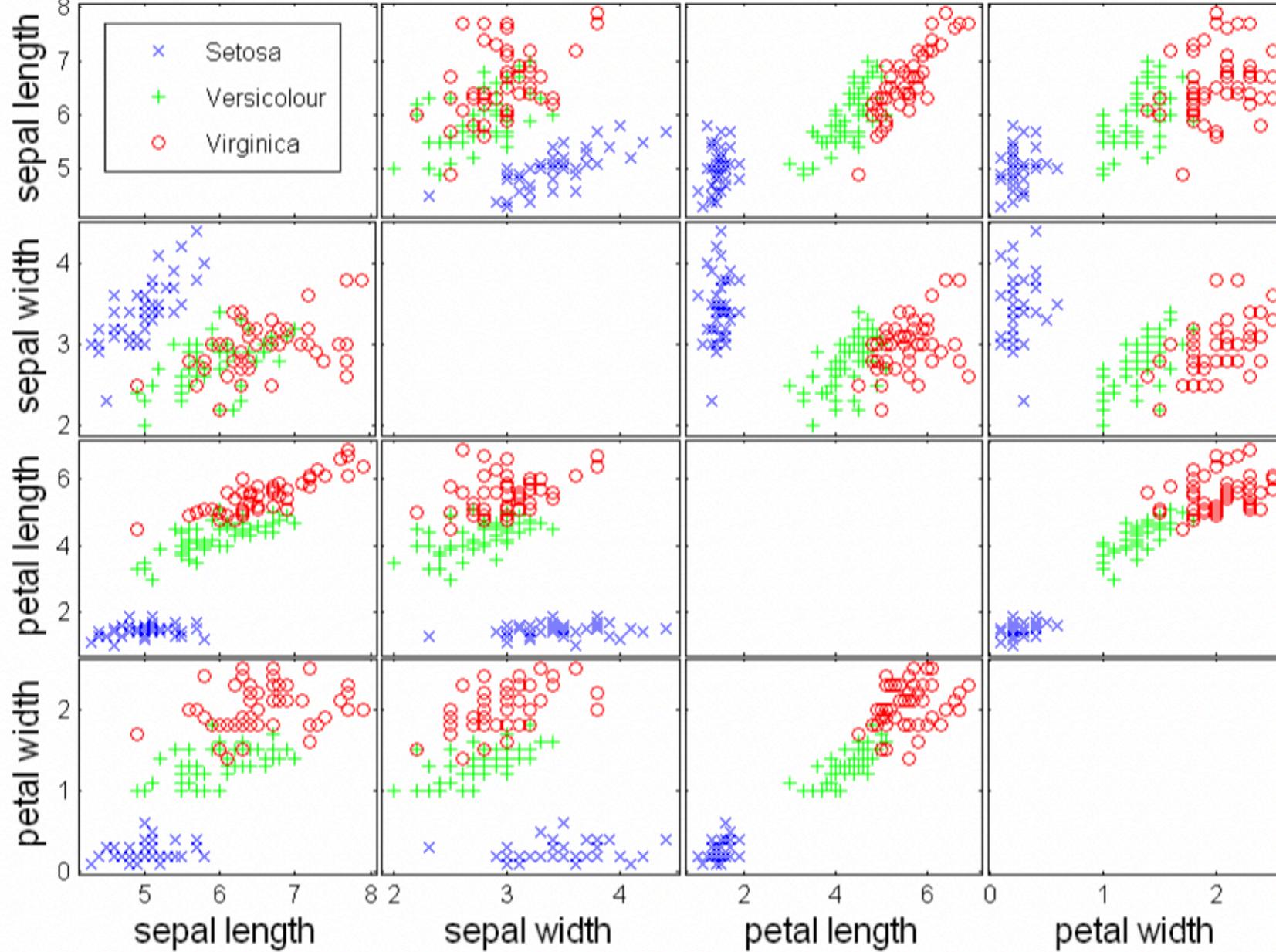
- Box Plot은 속성들을 비교할 때 사용 될 수 있다.



시각화 기법: Scatter Plots

- 속성 값이 위치를 결정함
- 2차원 scatter plot이 가장 일반적이지만 3차원도 가능
 - 객체를 나타내는 마커의 크기, 모양 및 색상을 사용하여 추가 속성을 표시
- Scatter plot 배열은 여러 쌍의 속성의 관계를 요약 가능

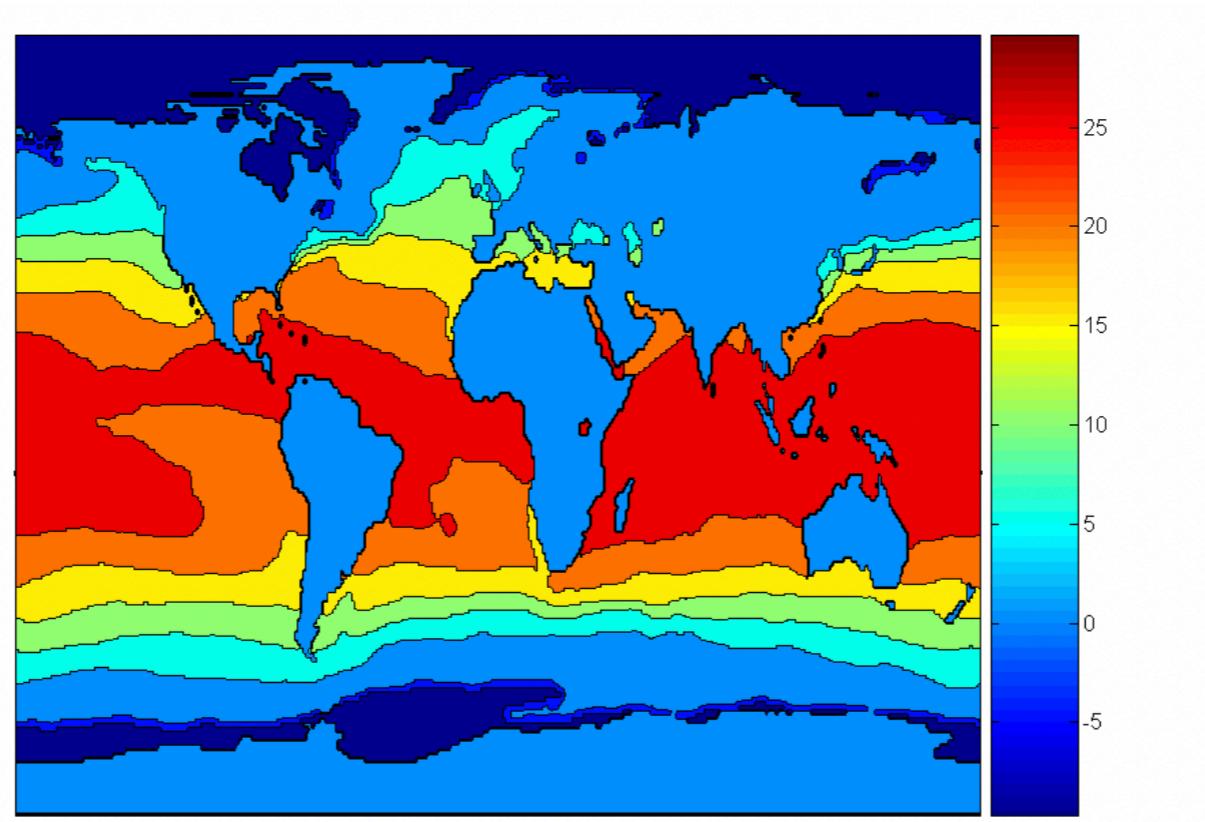
Scatter Plot 예제



시각화 기법: Contour Plot

- **Contour Plot**

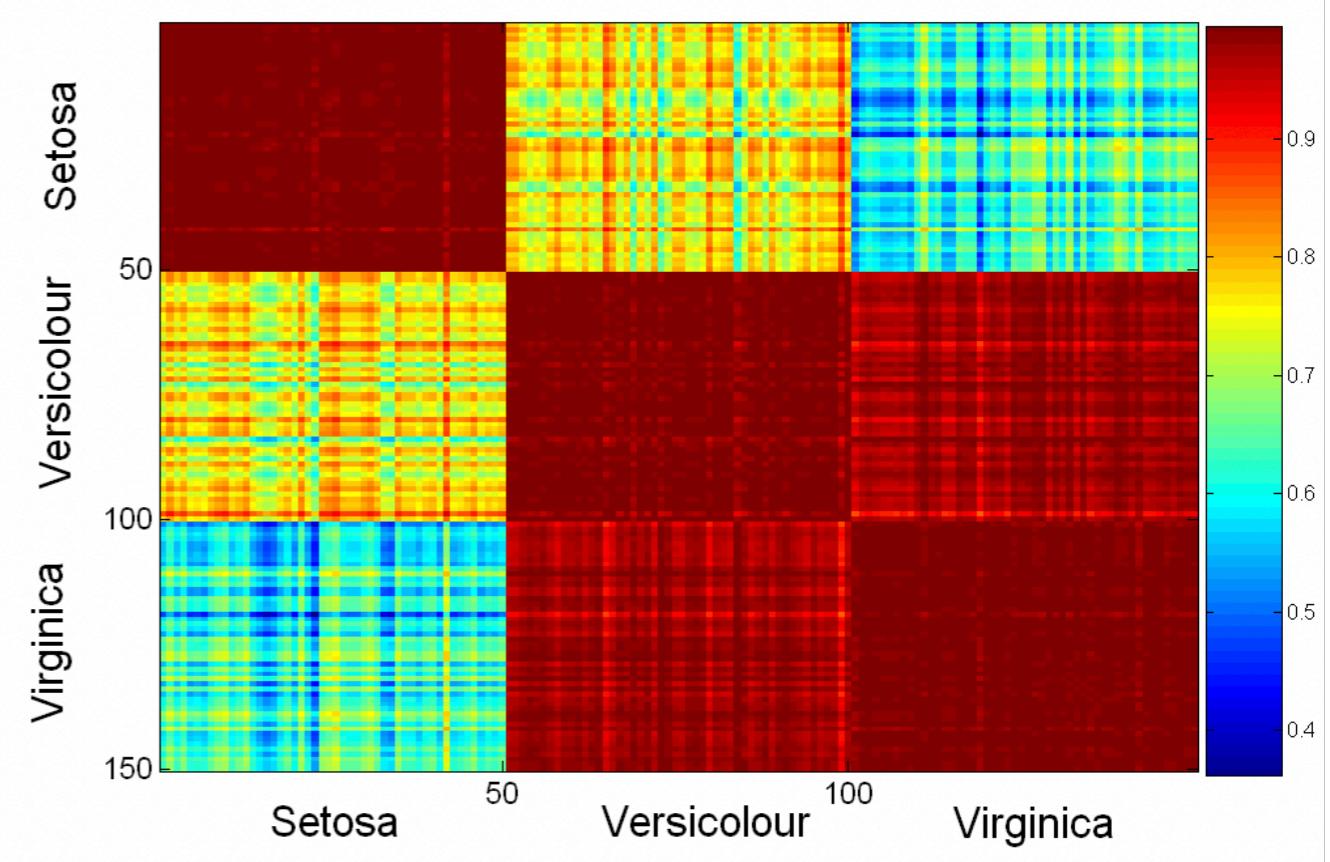
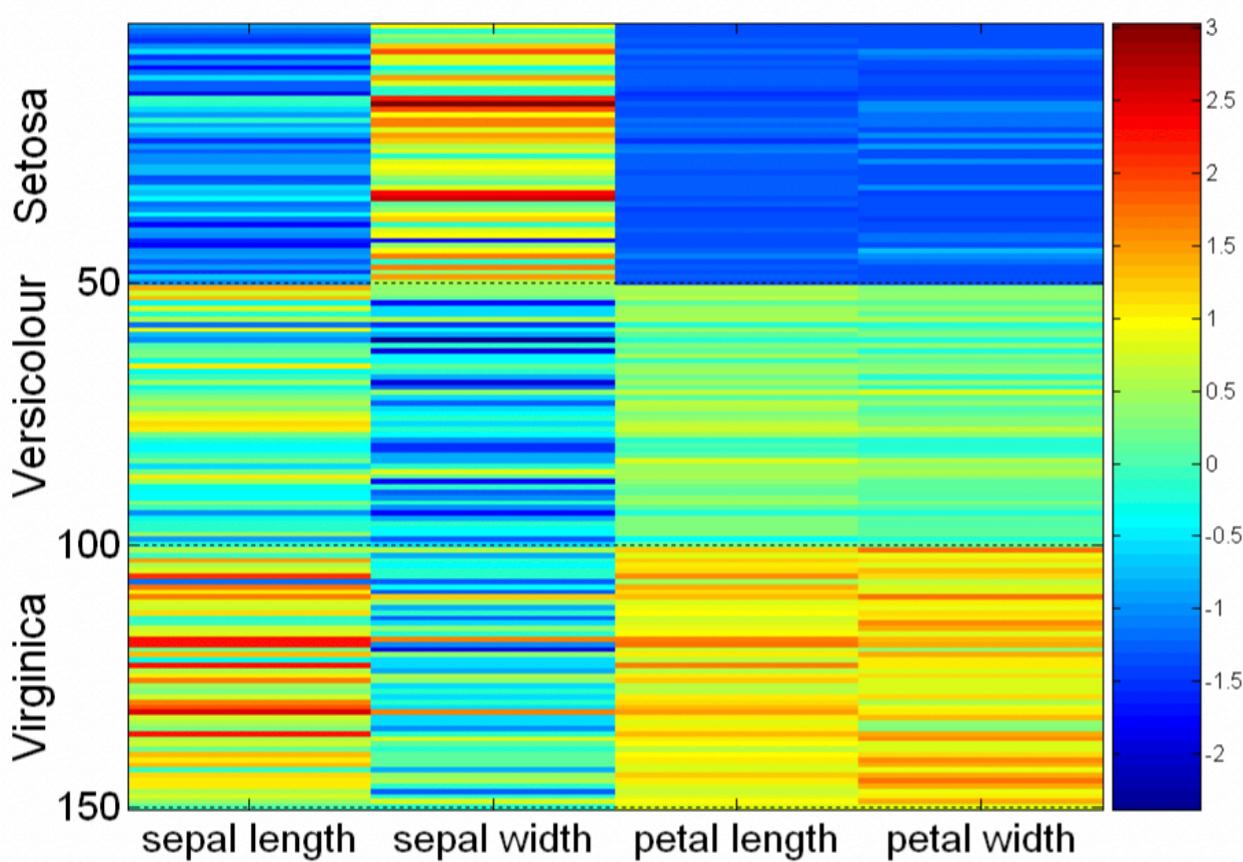
- 연속 그리드가 공간 그리드에서 측정 될 때 유용
- 평면을 유사한 값의 영역으로 분할
- 영역 내의 같은 값을 갖는 점들을 연결하여 등고선을 형성
- 온도, 강우, 기압 등을 표시 가능



시각화 기법: Matrix Plots

- Matrix Plot

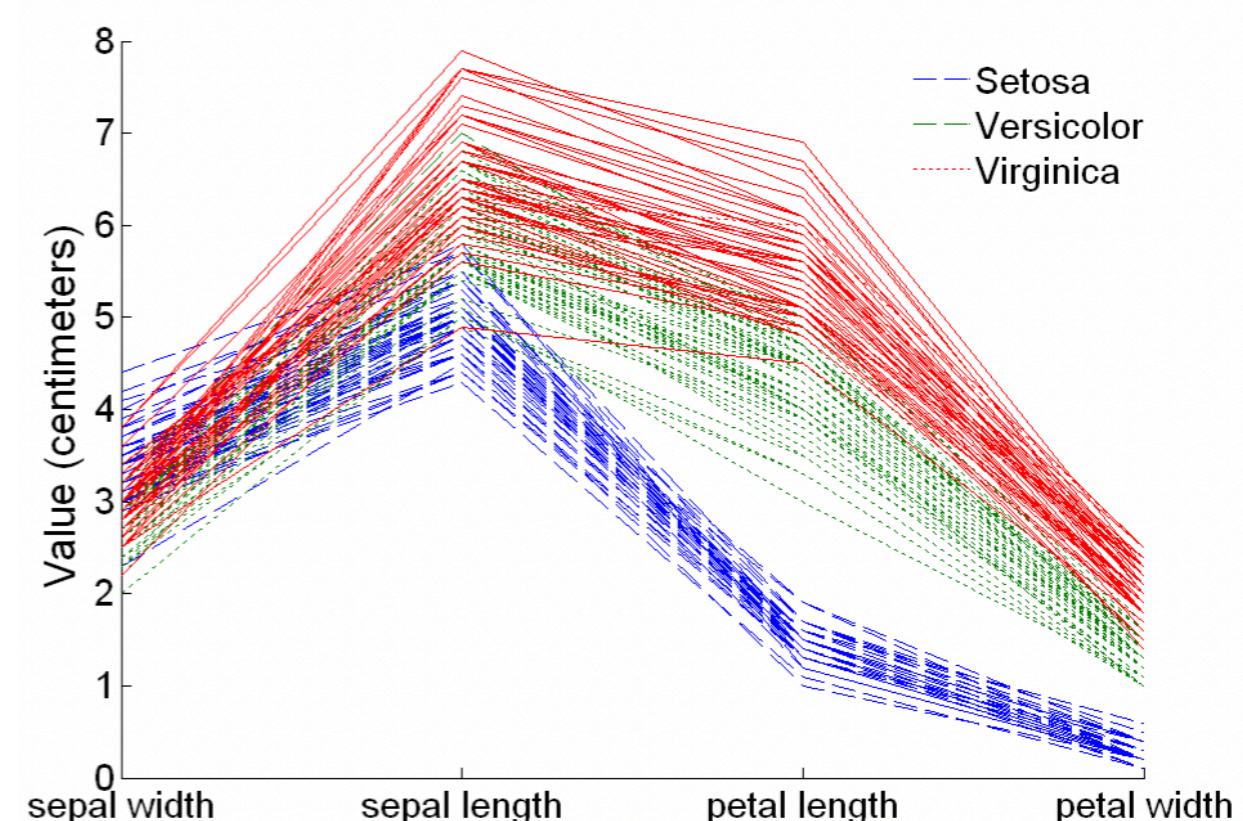
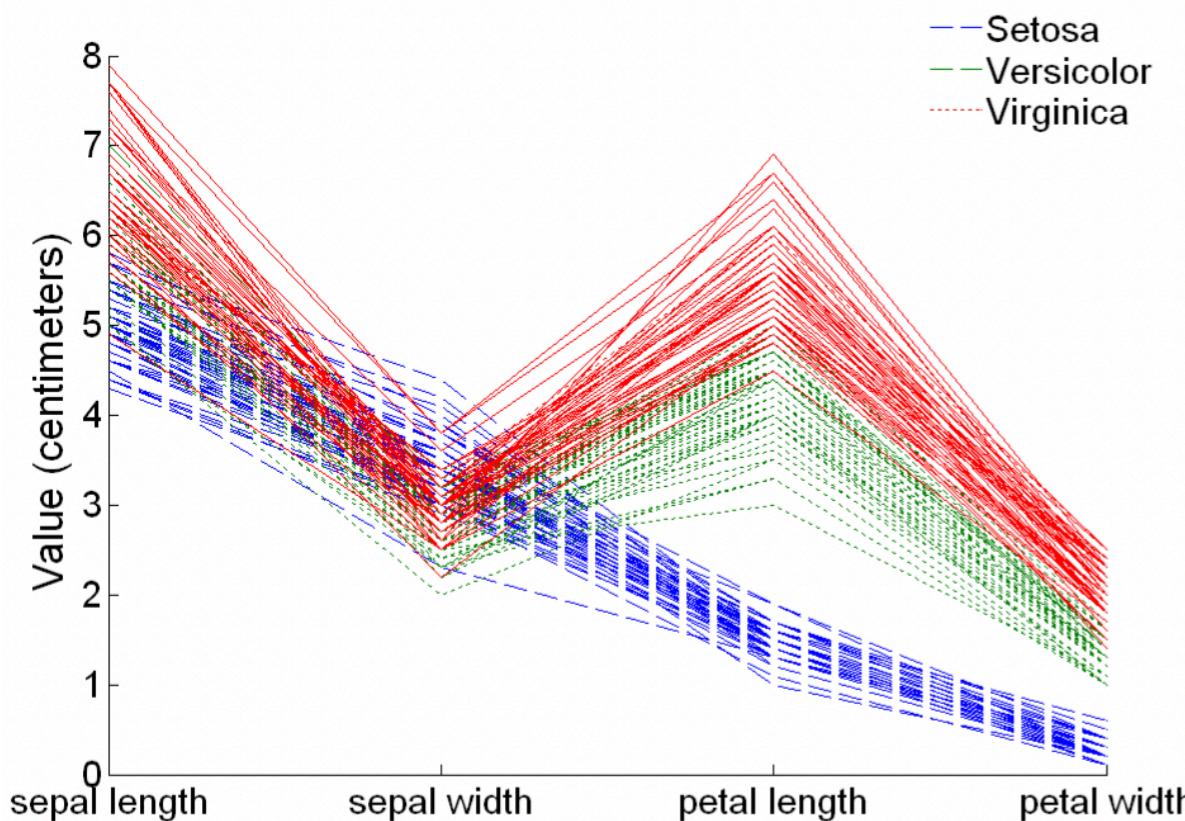
- 데이터 매트릭스를 표현 가능
- 클래스에 따라 객체가 정렬 되어있을 때 유용
- 일반적으로 한 특성(attribute)이 플롯을 지배하지 않도록 정규화(normalization)과정이 포함
- 유사성 (similarity) 또는 거리 (distance) 행렬의 플롯은 객체 간의 관계를 시각화하는 데 유용



시각화 기법: Parallel Coordinates

- **병렬 좌표**

- 고차원 데이터의 속성 값을 플로팅하는데 사용
- 직각 축을 사용하는 대신 평행 축 세트를 사용
- 각 객체의 속성 값은 각 해당 좌표축에 점으로 표시되고 점은 선으로 연결
- 따라서 각 객체는 선으로 표시
- 일부 속성에 대해서는 다른 클래스여도 모여있는 경우가 있음
- 그룹을 볼 때 속성의 순서가 중요



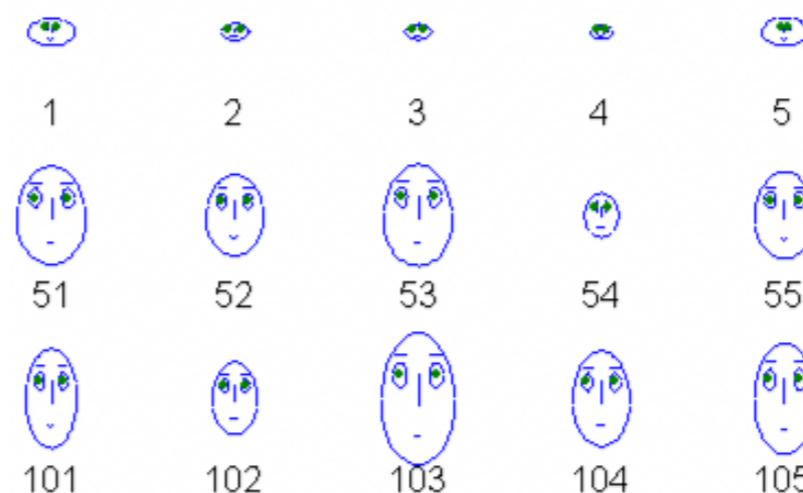
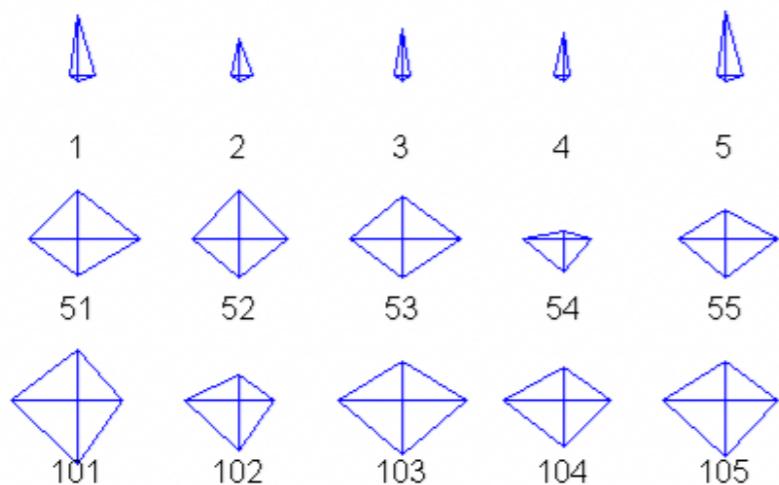
Other Visualization Techniques

- Star Plot

- 평행 좌표에 대한 유사한 접근 방식이지만 중심점에서 축이 방사
- 객체 값을 연결하는 선은 다각형입니다.

- Chernoff Plot

- Herman Chernoff가 만든 접근 방식
- 이 방법은 각 속성을 얼굴의 특성과 연결
- 각 속성의 값은 해당 얼굴 특성의 모양을 결정
- 사람의 얼굴을 구별하는 능력에 의존



E.O.D