탐색적 데이터 분석

고려대학교 석 준 희

목차

- 탐색적 데이터 분석
- 데이터 요약
- 프로그래밍 실습

KOREA UNIVERSITY

탐색적 데이터 분석

기계학습의 절차

- 문제의 설정 → 종속변수 Y가 무엇인가?
- 데이터 수집 → 관련 데이터 수집 (X, Y)
  - (기존분석) 실험의 설계하고 수행하여 데이터 수집
  - (빅데이터분석) 이미 존재하는 DB에서 관련된 모든 데이터를 수집
- **탐색적 데이터 분석 (EDA, Exploratory data analysis)**
  - **데이터에 대해서 배우는 과정**
  - **시각화, 결측치, 이상치 탐색 등을 포함**
- 본격적 데이터 분석 (예측 모델)
  - 클린 데이터로부터 시작 (n = 100M, p = 10k)
  - 트레이닝 셋과 테스트 셋을 분리 (보통 시간순서에 따라)
  - 불필요한 종속변수 제거 (Feature selection)
  - 학습 모델 후보 선정 (EDA에 따라 4~5개정도 후보 선정)
  - (교차)검증 기법을 이용하여 모델 선정
  - 테스트셋을 이용하여 최종 성능 평가

- 완전히 새로운 데이터셋으로 다시 평가 (필드테스트)

KOREA UNIVERSITY

탐색적 데이터 분석 (EDA: Exploratory Data Analysis)

- 탐색적 데이터분석
  - 실제 본격적인 분석에 들어가기 전에 데이터를 살펴보는 과정
  - 주어진 데이터에 대한 감 혹은 일반적 이해를 목적으로 함
  - 주로 데이터 기반의 분석으로 다른 가정 없이 진행됨

- 탐색적 데이터 분석은 다음을 포함함
  - 데이터 분포의 확인 (단변량, 이변량)
    - 평균, 분산, 중간값 등등
  - 시각화
    - 히스토그램, 산점도, 박스플롯 등등
    - 차원 축소를 같이 사용하기도 함
  - 데이터 변형 (Feature engineering)
  - 이상치 탐색 (Outlier detection)
  - 결측치 처리 (Missing value handling)
  - 통계 분석 (Statistical analysis)

KOREA UNIVERSITY

데이터의 종류

- **정형 데이터**
  - 일반적으로 표현되는 숫자, 범주 등의 데이터
  - 각 변수가 고유의 특징을 갖고 있음
  - 데이터 행렬의 형태로 표현이 용이 (엑셀로 볼 수 있음)
  - 예: 성별, 판매량, 키, 주소 등
  - 통계, 기계학습, 딥러닝 등으로 분석 가능

- **비정형 데이터**
  - 정형이 아닌 다른 모든 형태의 데이터
  - 각 변수의 의미를 찾기가 어려움
  - 데이터를 정형의 형태로 변환하여 사용
  - 예: 이미지, 음성, 텍스트 등
  - 주로 딥러닝으로 분석

- 오랫동안 비정형 데이터는 분석이 어려웠지만 최근 딥러닝의 발전으로 분석이 가능해짐

KOREA UNIVERSITY

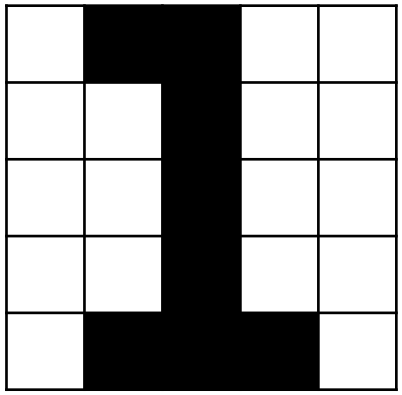
데이터의 종류

- 정형 데이터
  - 일반적으로 표현되는 숫자, 범주 등의 데이터

	A	B	C	D	E
1	환자ID	성별	나이	키	몸무게
2	1001	남	35	174	76
3	1002	여	29	160	54
4	1003	여	40	163	62
5	1004	남	22	182	91

$$\begin{bmatrix} 1 & 35 & 174 & 76 \\ 2 & 29 & 160 & 54 \\ 2 & 40 & 163 & 62 \\ 1 & 22 & 182 & 91 \end{bmatrix}$$

- 비정형 데이터
  - 정형이 아닌 다른 모든 형태의 데이터



$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

데이터 행렬

- 데이터 행렬
  - 일반적인 데이터의 표현
  - 행: 샘플, 표본, 개체
  - 열: 변수, 피처, 항목

- 변수의 수 (p): 데이터 차원
- 샘플의 수 (n)
  - 좁은 의미의 데이터의 크기
- 넓은 의미의 데이터의 크기
  - n x p

	X1	X2	X3	X4	X5	X6
s1	78.4	160.9	M	3	2	상
s2	70.4	167.8	M	2	1	상
s3	56.1	173.5	F	3	1	중
s4	58.8	166.1	F	3	1	하
s5	75.2	174.0	M	3	1	하
s6	63.2	160.0	F	1	2	중
s7	64.4	174.5	F	3	2	하
s8	59.9	179.2	F	2	3	상
s9	50.8	161.0	M	3	1	상
s10	60.7	169.4	M	1	3	하

- 날씬한 행렬 (n>>p): 일반적이고 분석이 쉬움
- 뚱뚱한 행렬 (n<p): 특수한 경우에 발생하고 분석이 어려움

- 차원의 저주
  - 데이터의 차원이 높아질 수록 분석이 매우 어려워 짐

데이터의 형태

- 수치형 데이터
  - 데이터가 숫자로 되어 있음
  - 연속형 데이터: 모든 실수값이 가능 (예: 키, 몸무게)
  - 이산형 데이터: 특정 값(보통은 정수)만 가능 (예: 문자메세지 수신 횟수)

- 범주형 데이터
  - 데이터가 범주(클래스, 팩터 등)로 되어 있음
  - 명목형 데이터: 순서가 없음 (예: 성별, 지역, 오류코드 등)
  - 순서형 데이터: 순서가 있음 (예: 상/중/하, 등급 등)

- 순서-범주형과 이산-수치형은 서로 다름
  - 상/하 vs. 2/1 은 수학적으로 유사하게 다룰 수 있음
  - 상/중/하 vs. 3/2/1 은 불가능
  - 일반적으로 서로 구별됨
    - 상과 하의 차이는 중과 하의 차이와 같지 않음

- 그 외에도 중도절단 데이터 등이 있음

KOREA UNIVERSITY

데이터의 형태

- 예시: 각 데이터의 형태는?
  - X4가 "자격증의 수"라면? "내신등급"이라면? "지역코드"라면?

	X1	X2	X3	X4	X5	X6
s1	78.4	160.9	M	3	2	상
s2	70.4	167.8	M	2	1	상
s3	56.1	173.5	F	3	1	중
s4	58.8	166.1	F	3	1	하
s5	75.2	174.0	M	3	1	하
s6	63.2	160.0	F	1	2	중
s7	64.4	174.5	F	3	2	하
s8	59.9	179.2	F	2	3	상
s9	50.8	161.0	M	3	1	상
s10	60.7	169.4	M	1	3	하

데이터의 형태에 따른 분석

- 수치형 데이터
  - 연속/이산 모두 동일하게 연속형 데이터로 분석

- 범주형 데이터
  - (기초단계) 명목/순서 모두 동일하게 명목형으로 분석
  - (고급단계) 순서형 데이터를 순서를 고려하는 방법론을 쓰거나 이산-수치형으로 변환하여 분석
    - 변수에 대한 이해가 필요

- 본 강의에서는
  - 연속-수치형 데이터
  - 명목-범주형 데이터
  - 로 나누어 살펴볼 예정

KOREA UNIVERSITY

수치형 데이터와 범주형 데이터의

- 어떻게 구분하나? 보면 안다!
- 변수가 너무 많아 일일이 볼 수 없다면?

- 보통은 R이나 Python에서 데이터를 읽는 함수가 고유의 알고리즘으로 판단

- 일반적인 판단의 절차
  - 값이 문자나 기호의 형태: 범주형
  - 숫자 형태인데 겹치는 값이 거의 없을 때: 수치형 데이터
    - 100개의 표본이 98개의 서로 다른 값을 갖는다
  - 숫자 형태인데 겹치는 값이 많은 때: 범주형 데이터에 대한 코드인지 확인
    - 100개의 표본이 1과 2 두 개의 값 만을 갖는다

KOREA UNIVERSITY

탐색적 데이터 분석

데이터 요약

수치형 데이터: 대표값

- 대표값: 많은 데이터를 요약하는 하나의 값
  - **평균** (Mean, Average)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - 중간값, 중앙값 (Median): 모든 값을 크기 순으로 일렬로 세웠을 때 가운데 있는 값
    - 표본이 짝수 개라면?
  - 최빈값 (Mode): 가장 많이 나타나는 값

- 예제
  - 2, 3, 5, 7, 7 의 데이터에서 평균, 중간값, 최빈값은?
  - 2, 3, 5, 7의 중간값은?

수치형 데이터: 분포의 표현

- 분포를 표현하는 값
  - **분산** (Variance) or **표준편차** (standard deviation)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  - 최대값(Max), 최소값(Min)
  - 백분위수 (Percentile, Quantile): 값들은 크기 순으로 일렬로 세웠을 때 하위 %에 해당하는 값
    - 25% Percentile (1st Quartile): 아래에서부터 25%에 해당하는 값
    - 50% Percentile (2nd Quartile): 아래에서부터 50%에 해당하는 값
    - 75% Percentile (3rd Quartile): 아래에서부터 75%에 해당하는 값
    - 0%, 100% Percentile: 최소값, 최대값

- 예제
  - 2, 3, 5, 7, 7 의 데이터에서 최대/최소, 각 percentile은?
  - 1, 2, 3의 데이터에서 분산은?

수치형 데이터: 분포의 표현

- 도수분포표 (Frequency Table)
  - 수치형 데이터의 개수를 구간별로 나누어 표시한 표

학생들의 키

(단위 : cm)

147 169 145 128
163 132 153 156
169 135 128 141
139 159 149 145
138 151 146 153

<정리되지 않은 자료>

학생들의 키

키 (cm)	학생 수 (명)
120 이상 ~ 130 미만	2
130 ~ 140	4
140 ~ 150	6
150 ~ 160	5
160 ~ 170	3
합 계	20

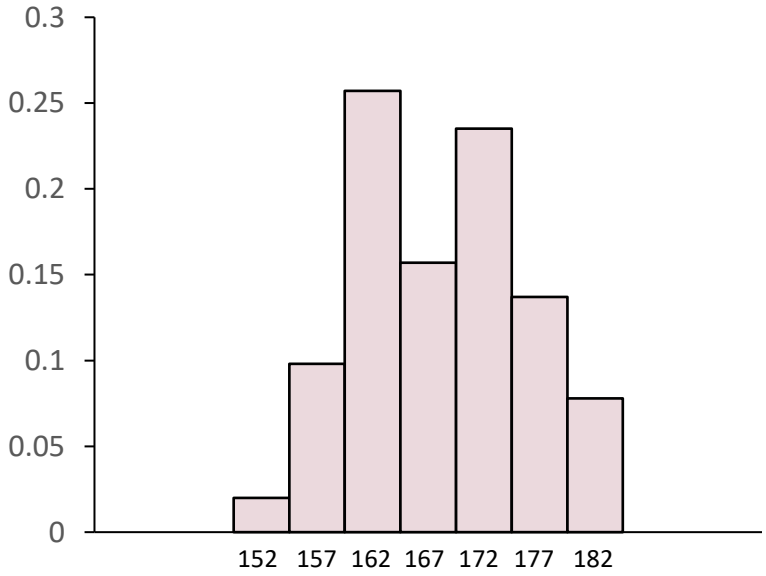
<도수분포표>

수치형 데이터: 시각화

- 히스토그램 (Histogram)
  - 연속형 변수의 도수분포표를 막대 그래프로 표현

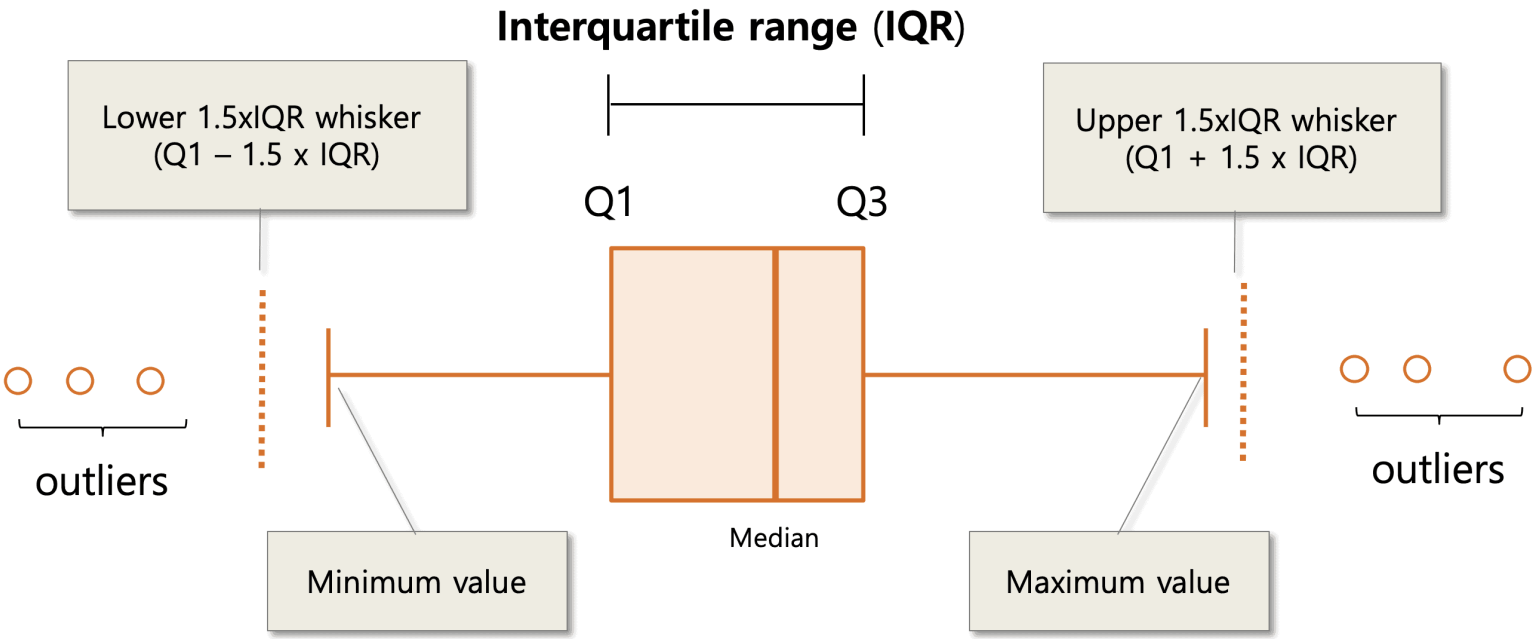
통계학과 신입생의 키에 관한 도수분포표

계급	계급구간(cm)	도수	상대 도수
1	149.5 ~ 154.5	1	0.020
2	154.5 ~ 159.5	5	0.098
3	159.5 ~ 164.5	14	0.257
4	164.5 ~ 169.5	8	0.157
5	169.5 ~ 174.5	12	0.235
6	174.5 ~ 179.5	7	0.137
7	179.5 ~ 184.5	4	0.078
합 계		51	1.000



수치형 데이터: 시각화

- 박스플롯 (Boxplot)
  - 연속형 변수의 분포를 박스의 형태로 표현하는 시각화
  - 다양한 퍼센타일과 이상치(outlier)들이 표현
  - 평균도 같이 표현되기도 함



https://help.ezbiocloud.net/box-plot/

범주형 데이터: 대표값

- 보통 최빈값을 사용

	최빈치 (mode)	중앙치 (median)	산술평균 (mean)
의미	가장 빈번하게 나타나는 값	자료를 크기 순으로 나열했을 때, 중앙에 위치하는 값	자료를 모두 더해서 자료의 개수로 나눈 값
특징	- 찾기 쉽다. - 포함하고 있는 정보가 적다. - 명목자료에 대해서는 최빈치가 대표치이다.	- 크기 순서가 변하지 않으면, 자료의 값이 변하더라도 변화하지 않는다. - 서열자료의 경우 평균을 사용할 수 없으므로 중앙치를 사용한다.	- 자료의 모든 값들을 고려하고 있으므로 정보가 풍부하다. - 일부 극단적인 값들에 크게 영향을 받는다. - 수학적인 연산에 의해 계산되므로 수학적인 조작이 용이하다.
예	유행하는 색상 인기있는 연예인	학급 석차의 중앙치	년간 평균 강수량 한달 교통사고량

범주형 데이터: 분포의 표현

- 도수분포표 (Frequency Table)
  - 범주형 데이터의 개수를 구간별로 나누어 표시한 표

혈액형	도수	상대 도수 (%)	각도 (◦)
A	22	36.67	132
B	20	33.33	120
AB	7	11.67	42
O	11	18.33	66
합	60	10.00	360

- 위 데이터에서 대표값은?

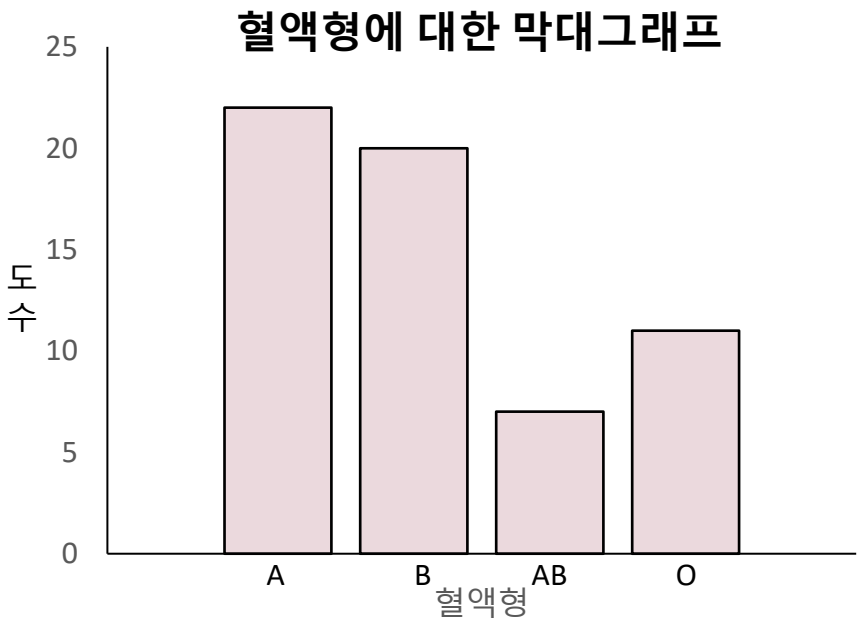
KOREA UNIVERSITY

범주형 데이터: 시각화

- 막대그래프 (Bar Plot):
  - 범주형 변수의 도수분포표는 일반 막대 그래프로 표현 가능

혈액형에 대한 도수분포표

혈액형	도수	상대 도수 (%)	각도 (◦)
A	22	36.67	132
B	20	33.33	120
AB	7	11.67	42
O	11	18.33	66
합	60	100.00	360



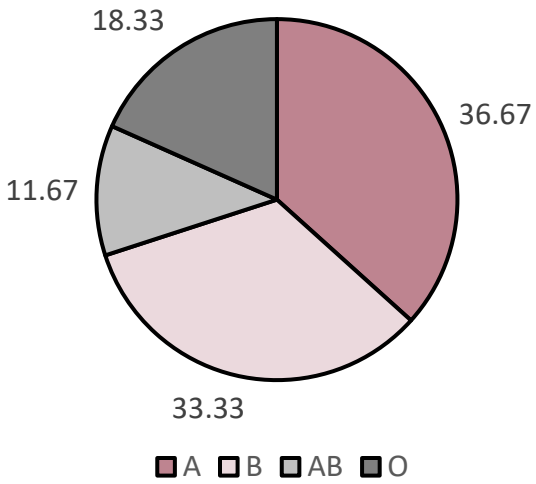
범주형 데이터: 시각화

- 원그래프 (Pie Chart)

혈액형에 대한 도수분포표

혈액형	도수	상대 도수 (%)	각도 (◦)
A	22	36.67	132
B	20	33.33	120
AB	7	11.67	42
O	11	18.33	66
합	60	100.00	360

혈액형에 대한 원형 그래프



이변량 데이터 요약

- 많은 경우 변수 자체보다는 변수와 변수 사이의 관계가 중요
  - 단변량 데이터 요약: 한 데이터에 대한 요약
  - 이변량 데이터 요약: 두 데이터 사이의 관계에 대한 요약
  - 다변량 요약: 2개 이상의 데이터 사이의 관계에 대한 요약

- 이변량 데이터 요약 (위: 시각화, 아래: 요약)

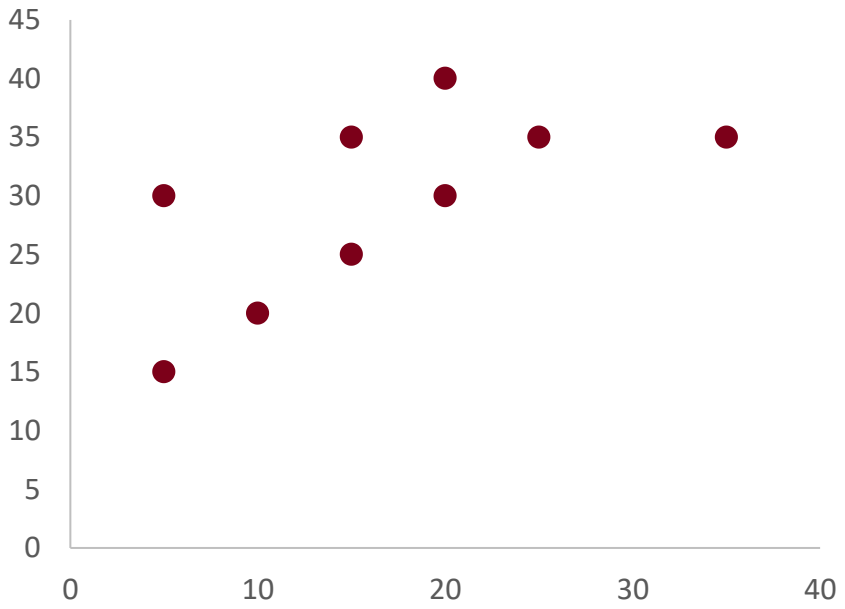
	수치형	범주형
수치형	산점도 상관계수	박스플롯 SMD
범주형	박스플롯 SMD	모자이크 플롯 Odd Ratio

- 이변량 데이터 요약의 목표
  - 두 변수 사이의 관련성을 확인하는 것
  - 두 변수가 관련이 있다는 것은 어떤 의미인가?

수치–수치: 시각화

- **산점도** (scatter plot): 두 변수의 값을 이차원 좌표 상에 점으로 표현
  - 두 수치형 변수 사이의 관계를 직관적으로 확인 가능
- **추세선** (trend line): 두 변수 사이의 관계를 선의 형태로 표현

Month	Series 1	Series 2
Sep	30	5
Aug	35	15
Jul	40	20
Jun	35	25
May	35	35
Apr	30	20
Mar	25	15
Feb	20	10
Jan	15	5



수치-수치: 요약

- 공분산 (covariance): 두 데이터가 얼마나 같이 변하는지 나타내는 값
  - 분산: 하나의 데이터가 변하는 정도를 나타내는 값
  - 데이터의 스케일에 따라 값이 달라짐. (e.g. 데이터가 10배되면 공분산도 10배)

$$r_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

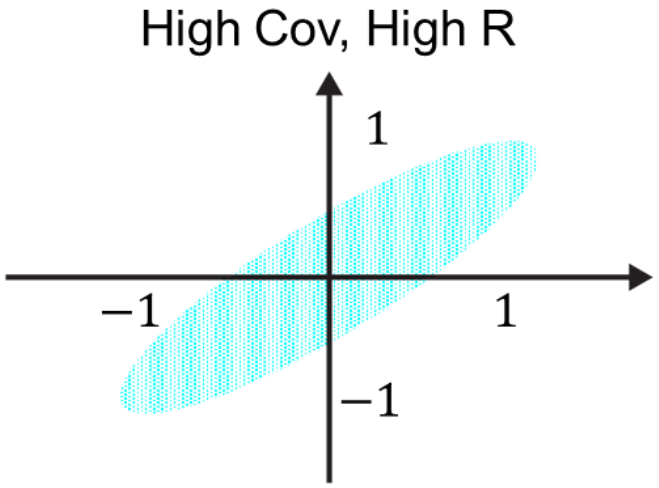
- **상관계수** (correlation coefficient)
  - 공분산의 분산으로 스케일한 값
  - -1과 1사이의 값으로 제한됨

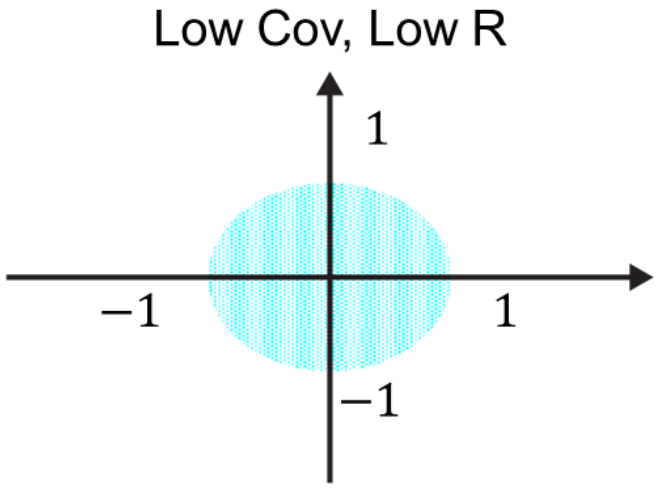
$$\rho_{XY} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

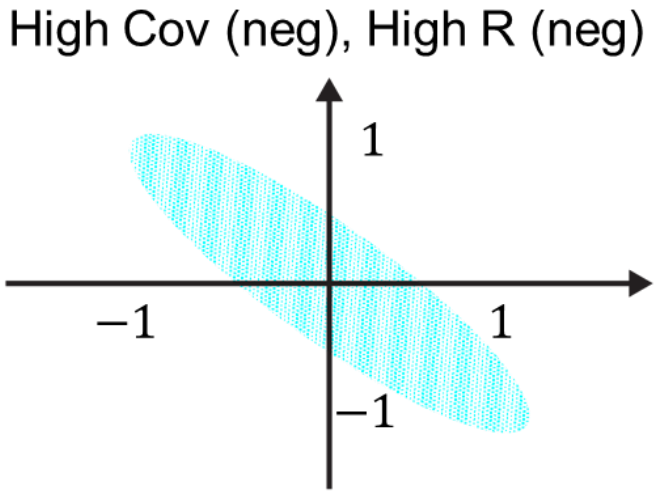
- 공분산 vs. 상관계수
  - Cov[Height(cm), Weight(kg)] vs. Cov[Height(m), Weight(g)]
  - 공분산은 다르지만 상관계수는 같음

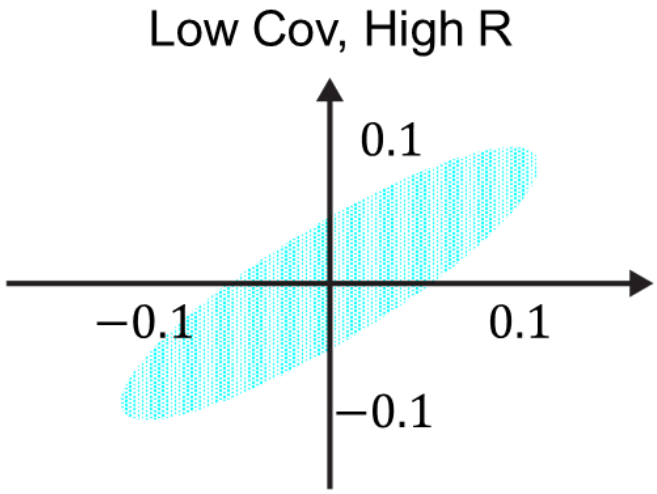
수치-수치: 요약

- 공분산은 스케일의 영향을 받지만 상관계수는 그렇지 않음









수치-범주: 데이터의 표현

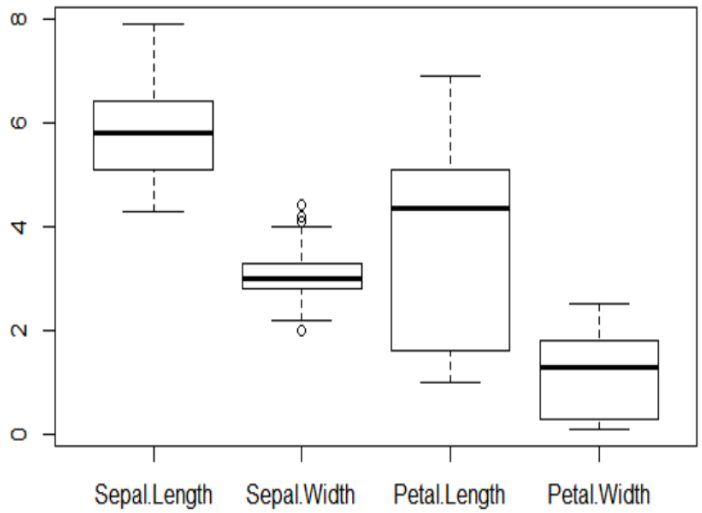
- Iris 데이터: Sepal Length vs. Species

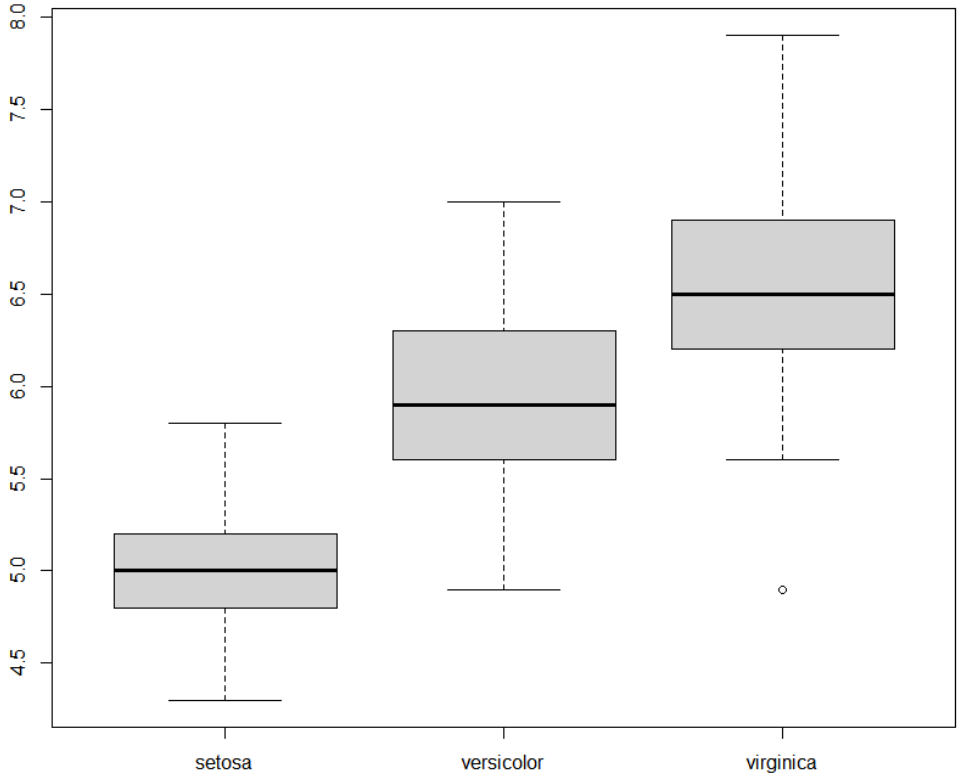
	A	B	C	D	E
1	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa

```
 Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

수치-범주: 시각화

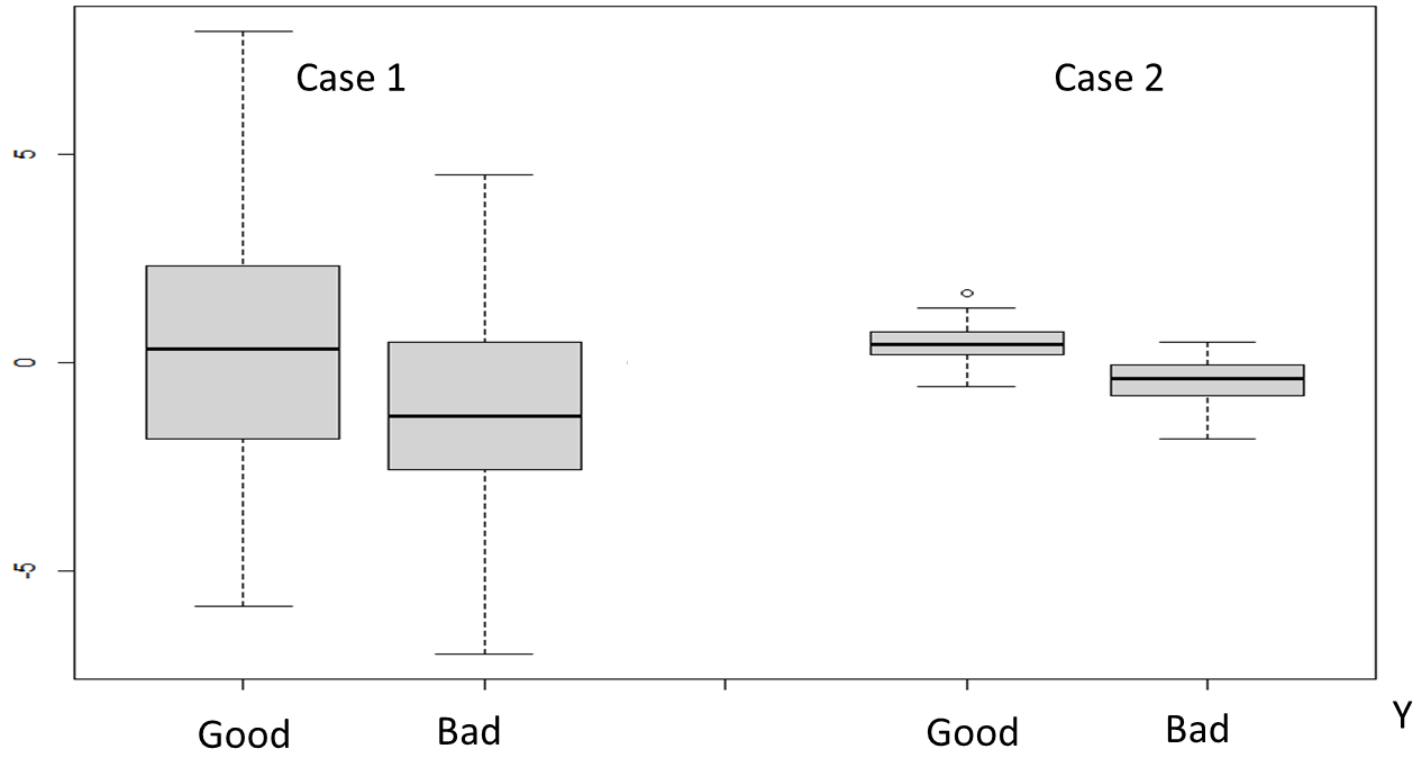
- 각 범주별로 수치형 데이터를 각각 시각화





수치-범주: 요약

- 범주형의 범주가 2개인 경우
  - 케이스 1: 평균의 차이 1.5
  - 케이스 2: 평균의 차이 0.5

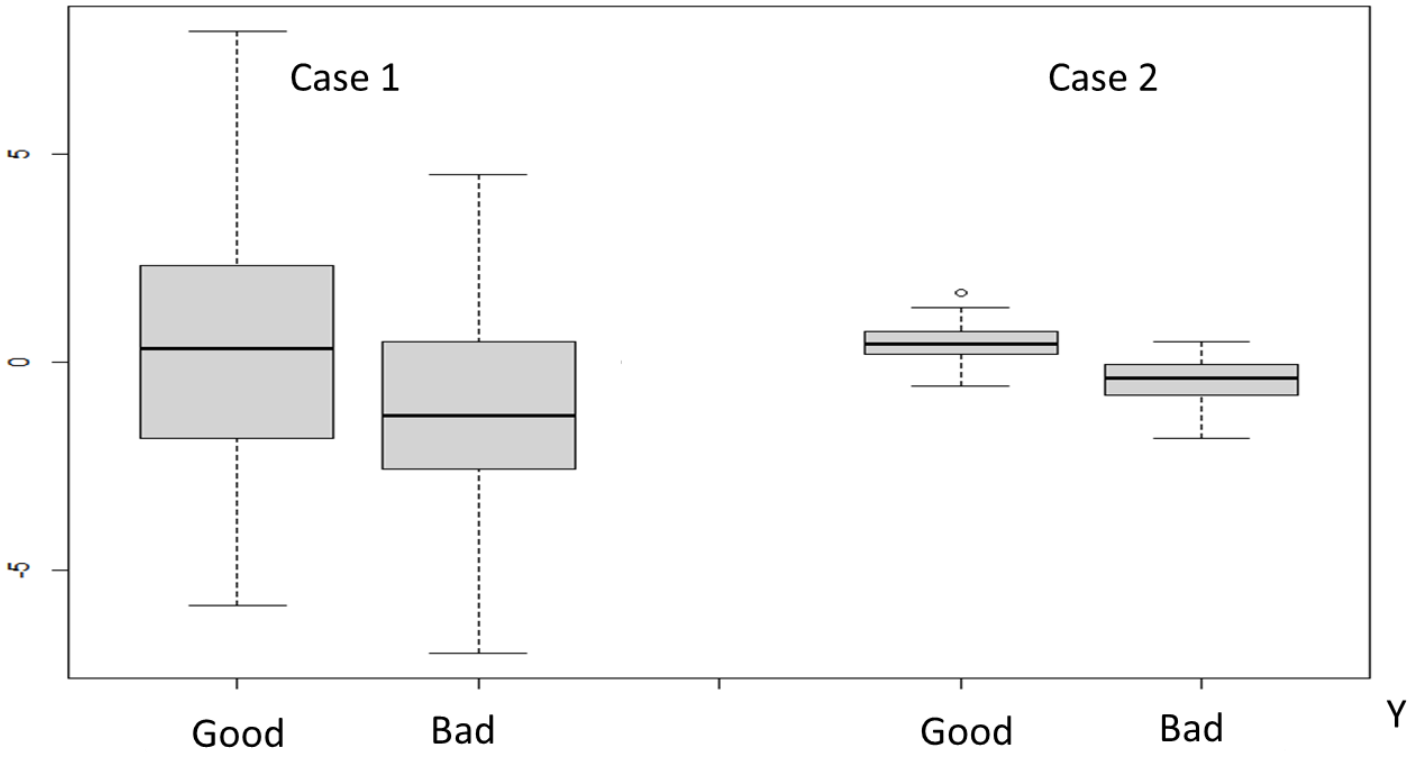


수치-범주: 요약

Standardized Mean Differences

- 범주형의 범주가 2개인 경우, 분산을 고려한 SMD (Cohen's d)값을 사용 가능

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s} \qquad s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$



범주-범주: 요약

- 두 개의 범주형 데이터에 대해서는 크로스 테이블을 이용해 요약

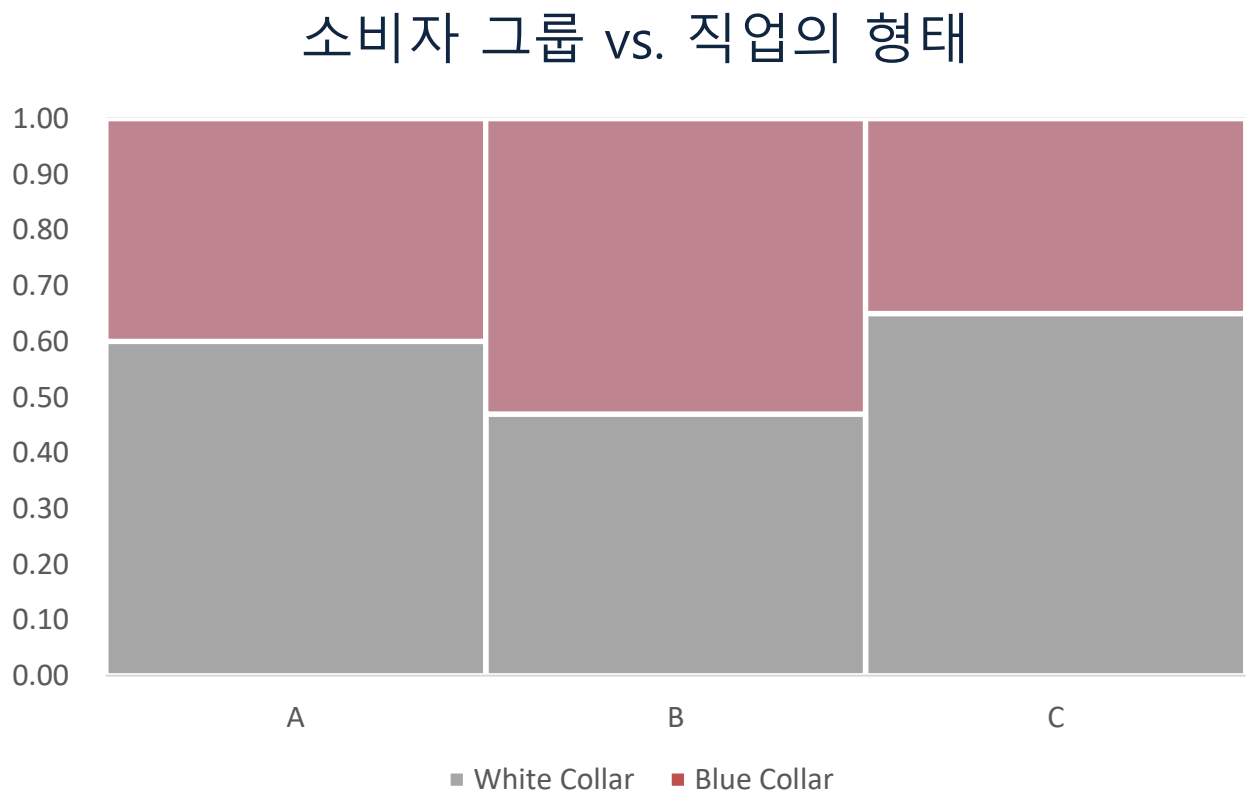
- 예제: 소비자 그룹 vs. 직업의 형태
  - X1: 소비 패턴에 따른 그룹 A, B, C
  - X2: 직업의 형태 (화이트/블루)

	A	B	C	Total
White collar	90	70	130	290
Blue collar	60	80	70	210
Total	150	150	200	500

KOREA UNIVERSITY

범주-범주: 시각화

- 모자이크 플롯: 두 범주형 데이터에 대한 시각화



범주-범주: 요약 (2x2의 경우)

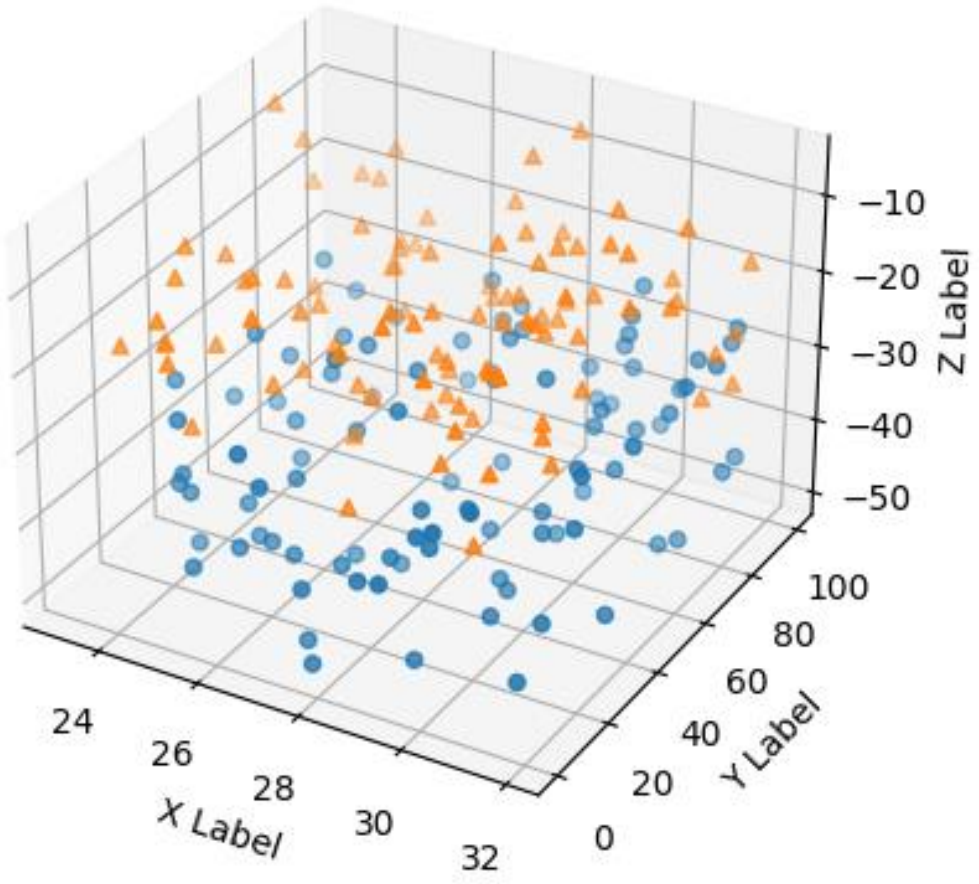
- 범주형 변수가 2개의 범주를 갖고 있을 때
  - **승수비** (odd ratio): "두 범주형 변수가 서로 관련이 있는가?"에 대한 수치적 값
  - 1에 가까울 수록 관련이 없고, 1에서 멀어질 수록 관련이 높음
  - Log(승수비), log odd ratio 가 일반적으로 사용

Above Average	Yes	No
Yes	A – Positives	C – Negatives
No	B - Negatives	D - Positives

$$\widehat{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

다변량 데이터 요약과 시각화

- 3차원 산점도 + 모양/색깔

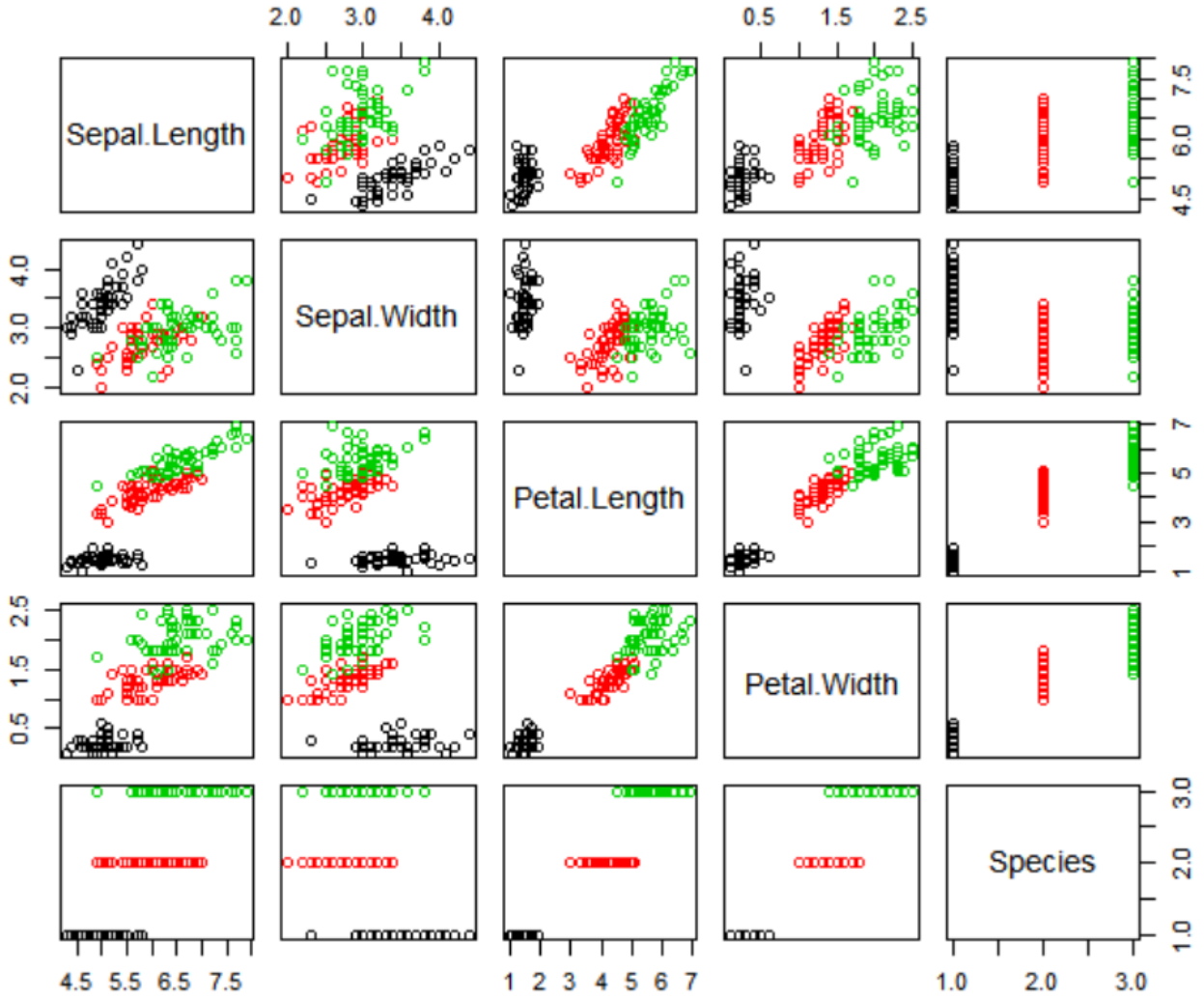


https://matplotlib.org/stable/gallery/mplot3d/scatter3d.html

KOREA UNIVERSITY

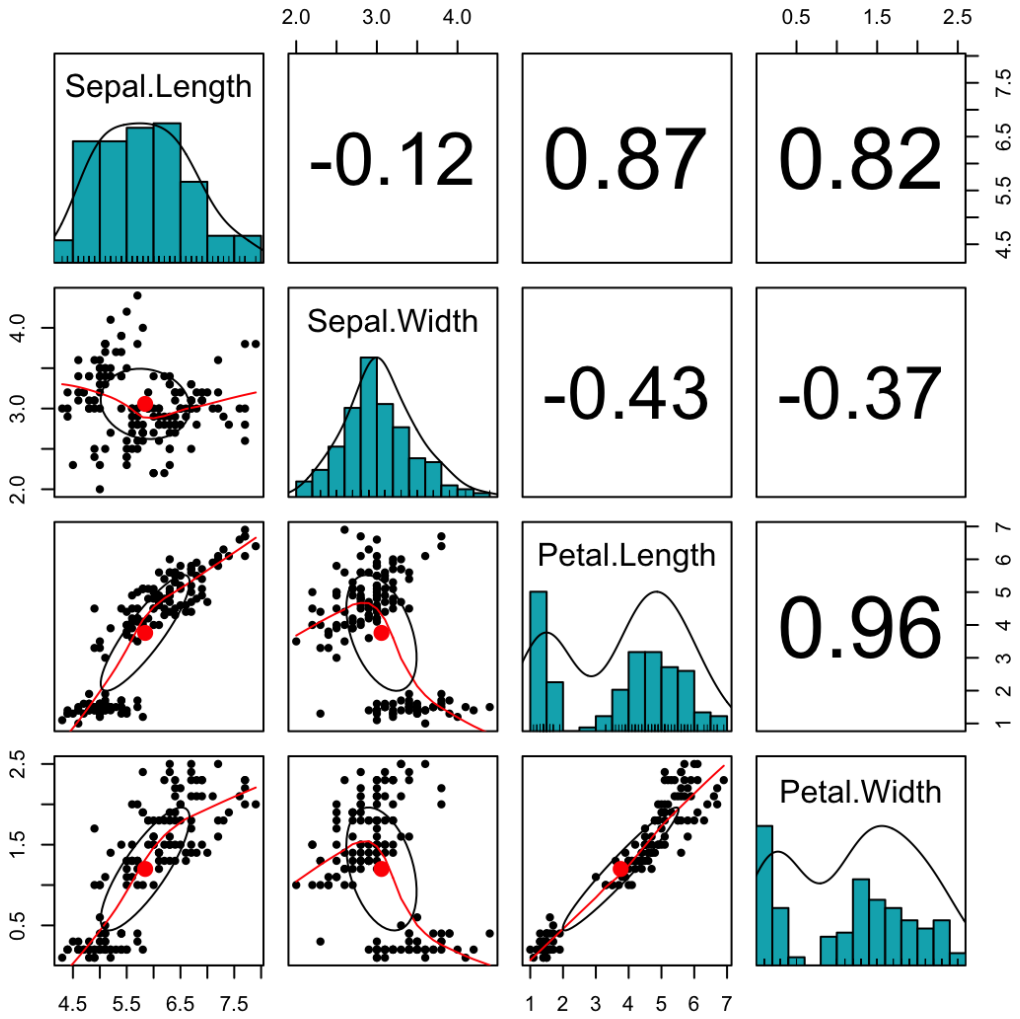
다변량 데이터 요약과 시각화

- Pairwise scatter plot



다변량 데이터 요약과 시각화

- Pairwise scatter plot



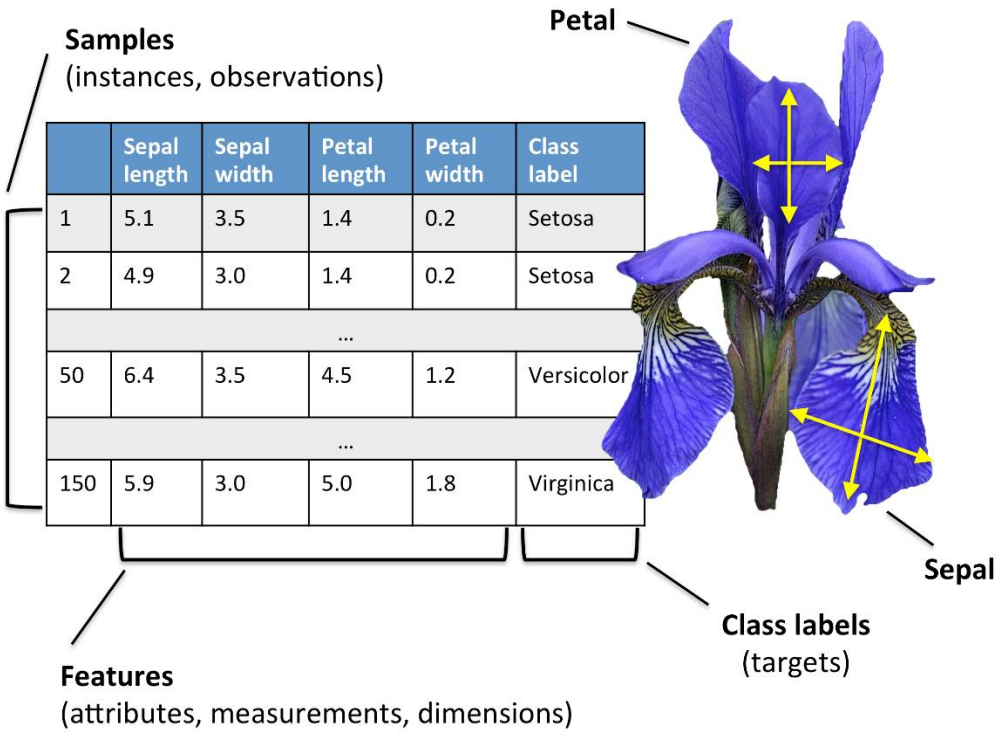
탐색적 데이터 분석

프로그래밍 실습

아이리스 데이터 셋

- 인공지능 교육에 사용되는 가장 기본적인 데이터 셋
- 꽃잎의 길이와 너비를 바탕으로 아이리스(붓꽃)의 품종을 예측

- 표본의 수: 150개
- 변수의 수: 4개
  - Sepal Length: 수치형
  - Sepal Width: 수치형
  - Petal Length: 수치형
  - Petal Width: 수치형
  - Species: 범주형



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

아이리스 데이터 셋

```
import pandas as pd
from sklearn.datasets import load_iris

X, y = load_iris(return_X_y=True,as_frame=True)
id2label = ['setosa','versicolor','virginica']
species = [id2label[i] for i in y]
data = X.copy()
data.columns = ['SepalLength','SepalWidth','PetalLength','PetalWidth']
data['Species'] = species
```

아이리스 데이터 셋

```
data
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

150 rows × 5 columns

데이터 요약

```
[ ]  # 데이터 크기
     data.shape
```

```
(150, 5)
```

```
     # 변수 정보
     data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   SepalLength  150 non-null    float64
 1   SepalWidth   150 non-null    float64
 2   PetalLength  150 non-null    float64
 3   PetalWidth   150 non-null    float64
 4   Species      150 non-null    object 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

데이터 요약

```
# 수치형 변수 대표값과 분포
data.describe()
```

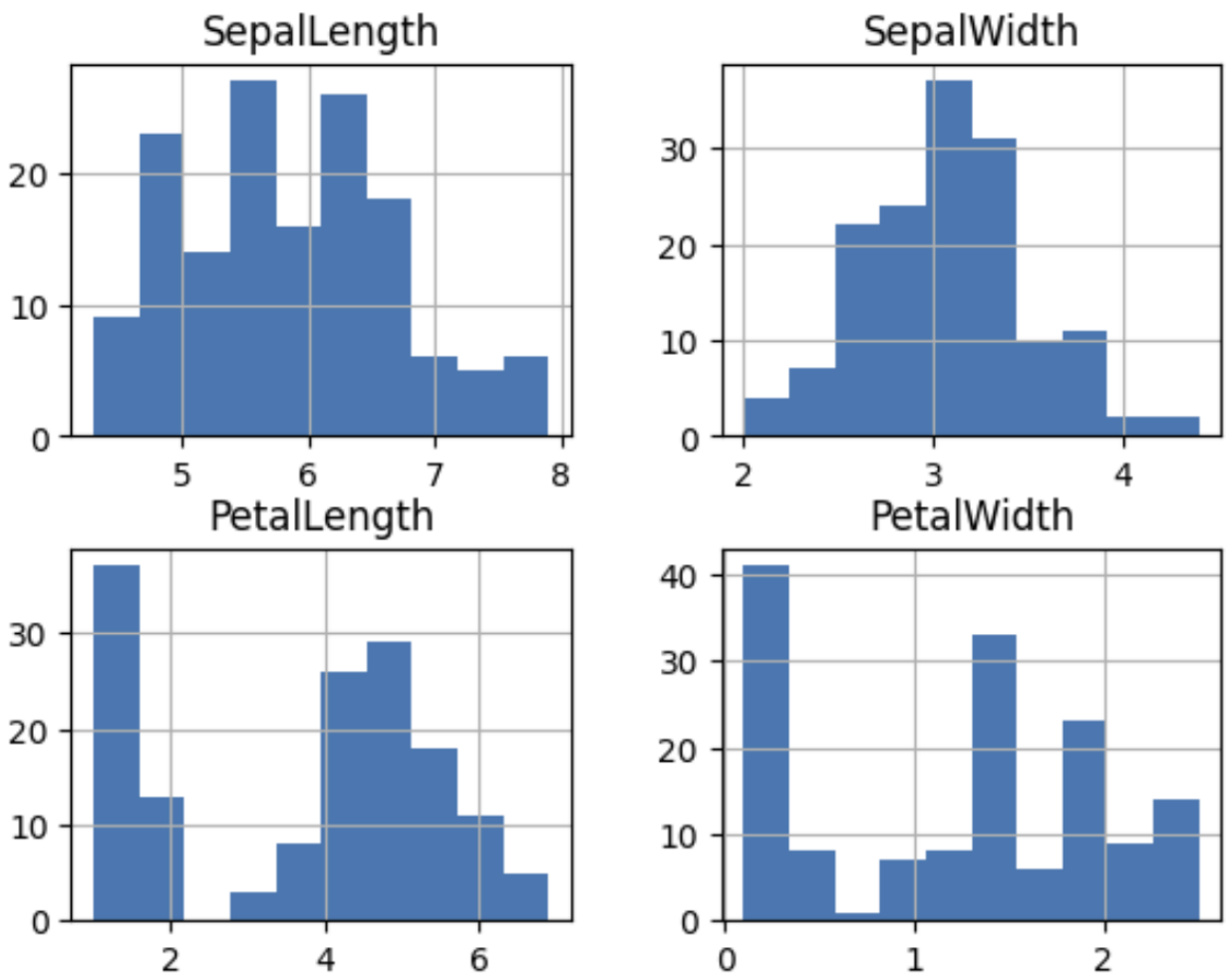
	SepalLength	SepalWidth	PetalLength	PetalWidth
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

```
[ ]  # 범주형 변수 도수분포
     data['Species'].value_counts()
```

```
setosa        50
versicolor    50
virginica     50
Name: Species, dtype: int64
```

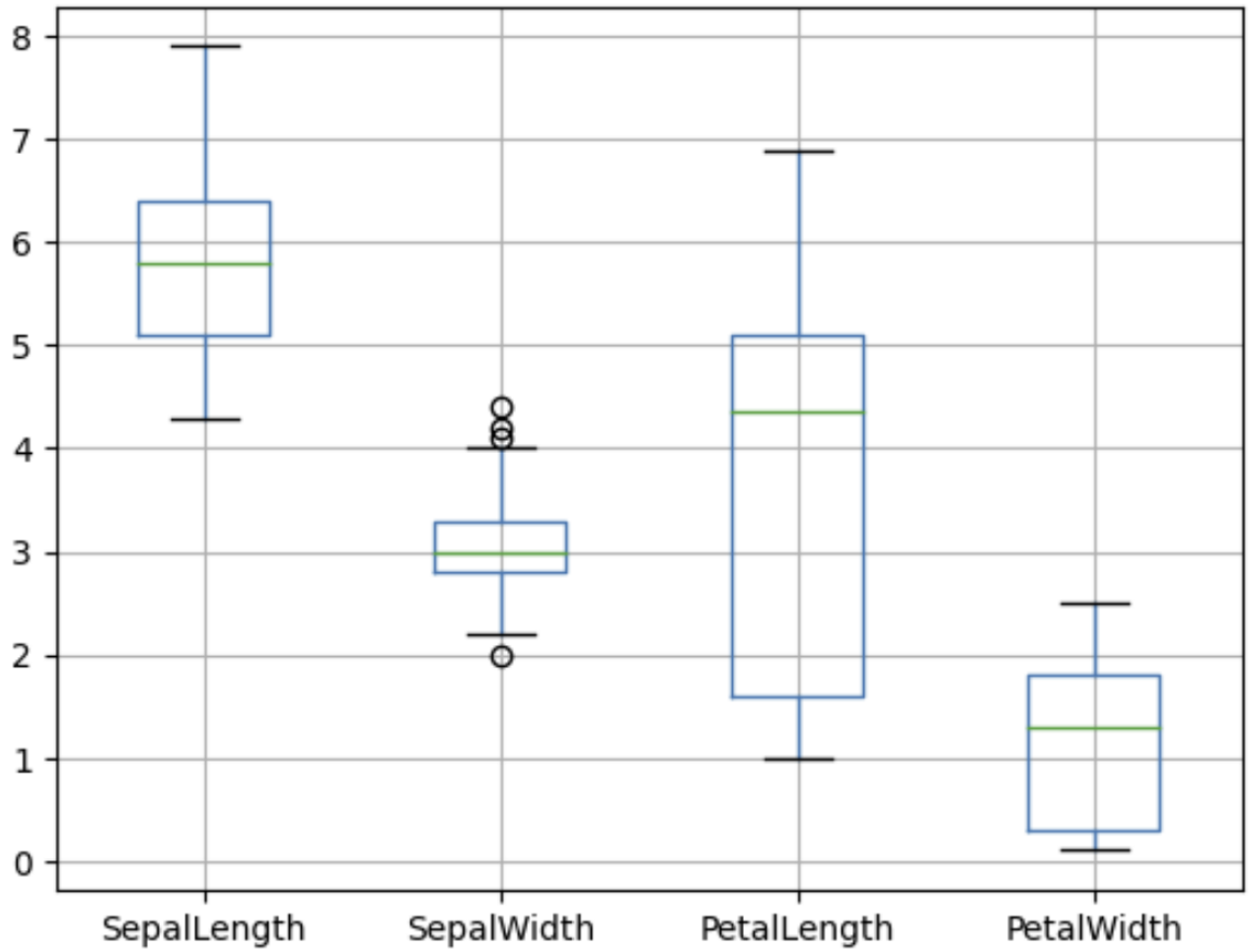
데이터 요약

```
# 히스토그램
data.hist()
```



데이터 요약

```
# 박스 플롯
data.boxplot()
```



이변량 데이터 분석 (수치-수치)

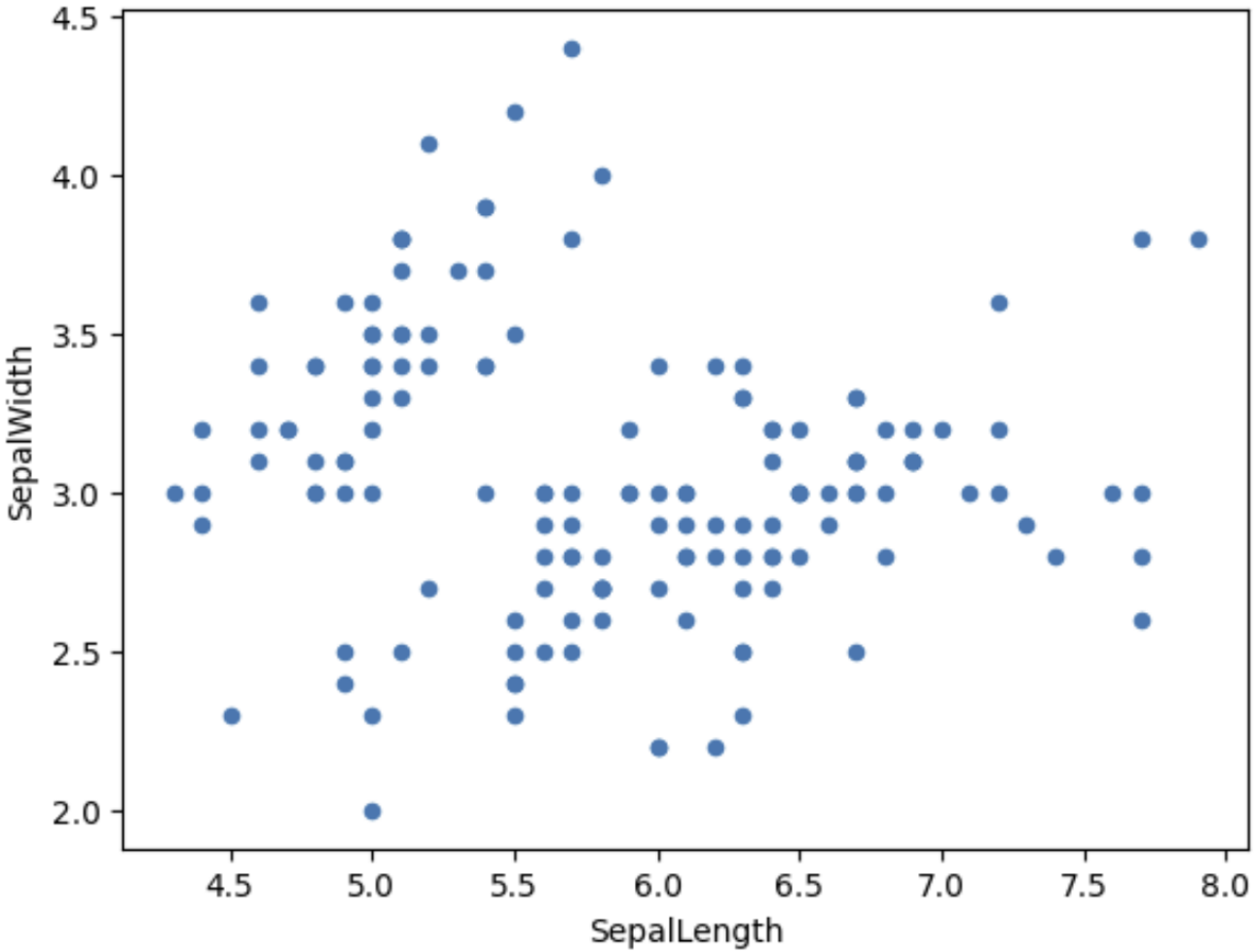
수치-수치

```
# 상관계수
data[['SepalLength','SepalWidth','PetalLength','PetalWidth']].corr()
```

	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	1.000000	-0.117570	0.871754	0.817941
SepalWidth	-0.117570	1.000000	-0.428440	-0.366126
PetalLength	0.871754	-0.428440	1.000000	0.962865
PetalWidth	0.817941	-0.366126	0.962865	1.000000

이변량 데이터 분석 (수치-수치)

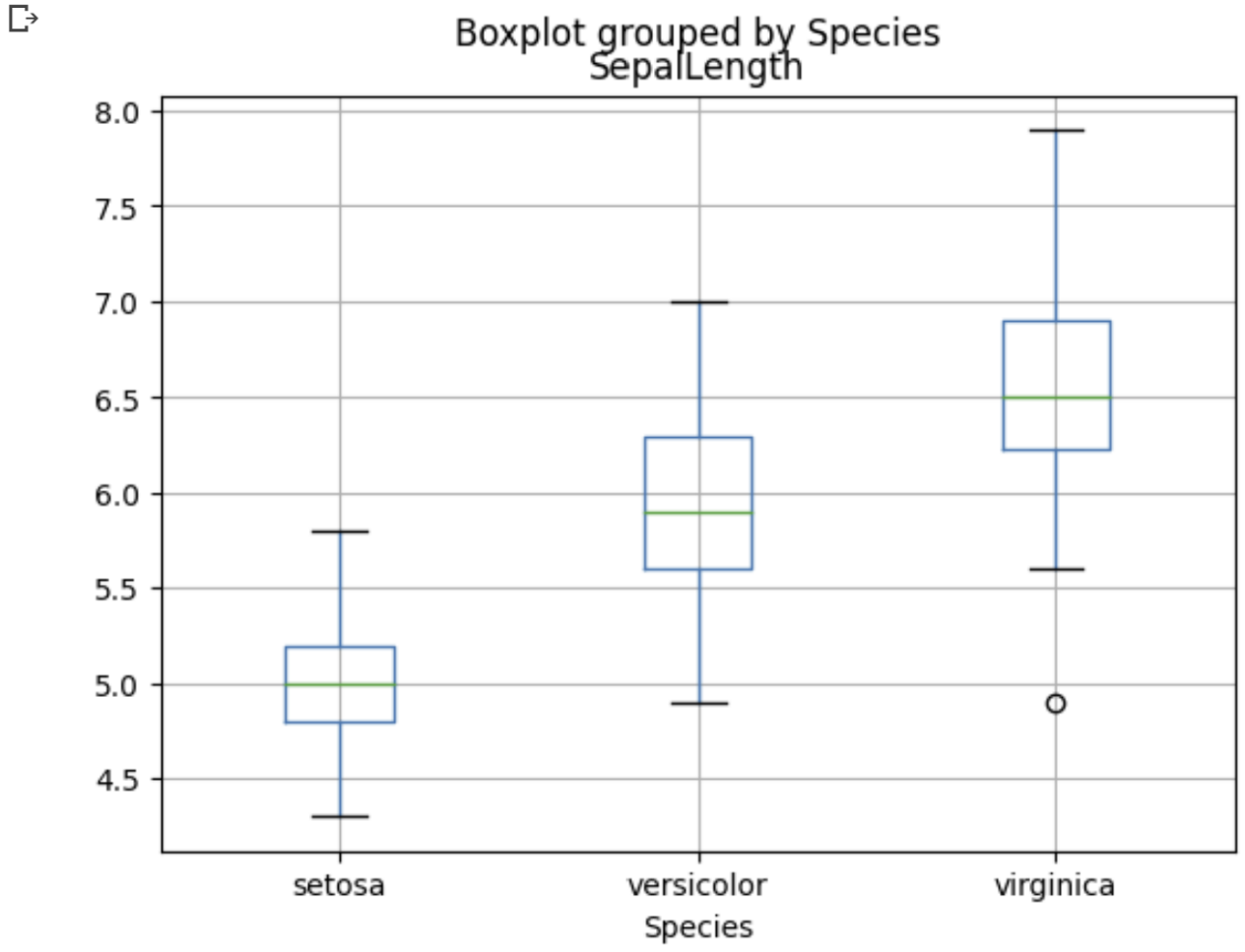
```
# 산점도
data.plot.scatter('SepalLength','SepalWidth')
```



이변량 데이터 분석 (수치-범주)

수치-범주

```
# 범주별 박스 플롯
data.boxplot('SepalLength',by='Species')
```



이변량 데이터 분석 (수치-범주)

```
[ ]  # 범주별 평균
     data.groupby('Species').mean()
```

Species	SepalLength	SepalWidth	PetalLength	PetalWidth
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

```
     # 범주별 분산
     data.groupby('Species').var()
```

Species	SepalLength	SepalWidth	PetalLength	PetalWidth
setosa	0.124249	0.143690	0.030159	0.011106
versicolor	0.266433	0.098469	0.220816	0.039106
virginica	0.404343	0.104004	0.304588	0.075433

이변량 데이터 분석 (수치-범주)

```
[ ]  # setosa와 versicolor 사이의 SMD
     import math
     d = abs(5.006 - 5.936)
     n1 = 50
     n2 = 50
     s = math.sqrt( ((n1-1)*0.124249 + (n2-1)*0.266433)/(n1+n2-2) )
     smd = d/s
     print(smd)
```

```
2.104196264251805
```

이변량 데이터 분석 (범주-범주)

범주-범주

```
# 현재 데이터에는 범주형 변수가 한 개이기 때문에 보기 어려움
# pd.crosstab(변수1, 변수2) 의 형태로 크로스테이블을 볼 수 있음
```

KOREA UNIVERSITY

감사합니다