

## CHAPTER 01

# Understanding the data

**01** What is a feature?

**02** Type of feature

**03** Putting data into the model

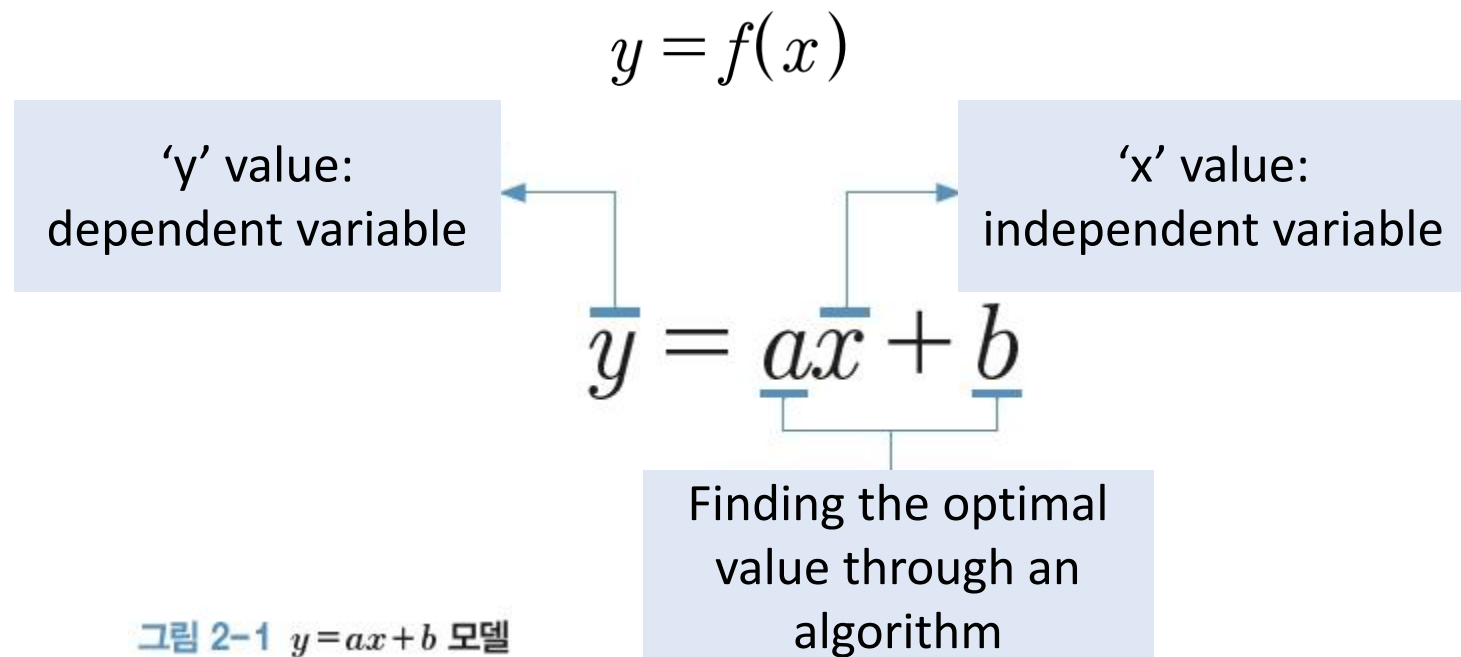
**01**

**What is a feature?**

# 01 What is a feature?

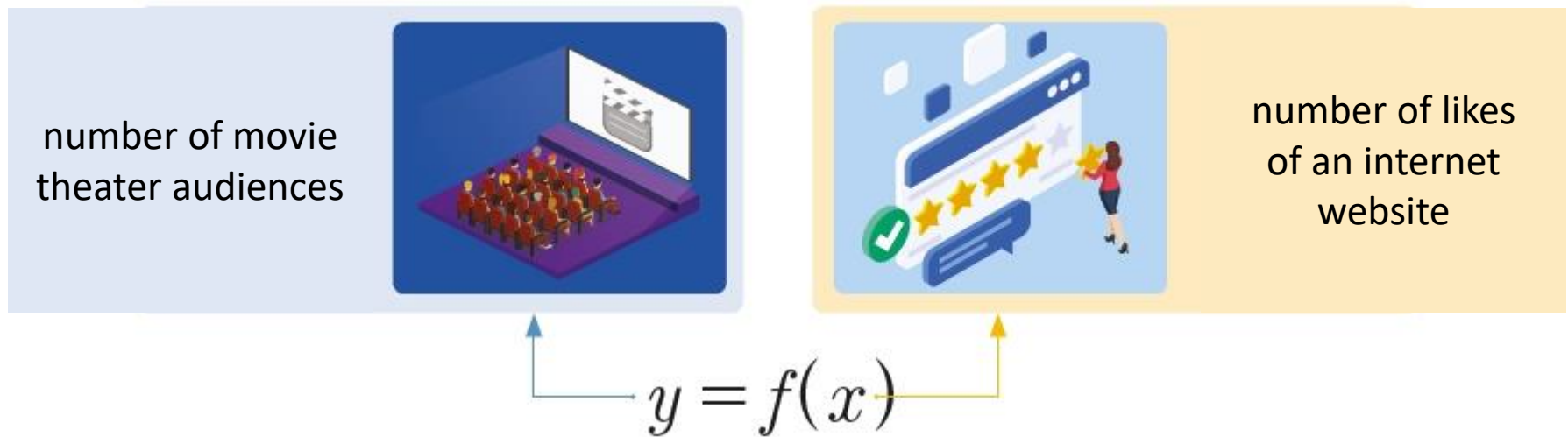
## 1. The concept of a feature

- Feature: means 'characteristic' or 'variable'
- Data has the greatest influence on building the model.
- The model is expressed as a function.



# 01 What is a feature?

- If the number of movie theater audiences (  $x$  ) can be predicted by the number of likes (  $y$  ) of an internet website, then the  $y$  value can simply be predicted with one kind of  $x$  value.



# 01 What is a feature?

- Boston House Price dataset

13  $x$  variables

$y$  variable

[01] CRIM	자치시(town)별 1인당 범죄율
[02] ZN	25,000 평방피트를 초과하는 거주지역의 비율
[03] INDUS	비소매상업지역이 점유하고 있는 토지의 비율
[04] CHAS	찰스강에 대한 더미변수(강의 경계에 위치한 경우는 1, 아니면 0)
[05] NOX	10ppm 당 농축 일산화질소
[06] RM	주택 1가구당 평균 방의 개수
[07] AGE	1940년 이전에 건축된 소유주택의 비율
[08] DIS	5개의 보스턴 직업센터까지의 접근성 지수
[09] RAD	방사형 도로까지의 접근성 지수
[10] TAX	10,000달러 당 재산세율
[11] PTRATIO	자치시(town)별 학생/교사 비율
[12] B	$1000(B_k - 0.63)^2$ , 여기서 $B_k$ 는 자치시별 흑인의 비율을 말함
[13] LSTAT	모집단의 하위계층의 비율(%)
[14] MEDV	본인 소유의 주택가격(중앙값) (단위 : \$1,000)

# 01 What is a feature?

- A model of how values such as crime rate, number of rooms, and property tax rate affect house price ( $y$ )
- Expressed by linear combination of 13 different independent variables as  $x_n$  and weight  $\beta_n$

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_{13} x_{13} + \beta_0 \times 1$$

- In machine learning, the independent variable  $x$  is called a feature.

**[TIP]** Among the above expressions,  $\beta$  is a commonly used expression in statistics. In general, machine learning uses a lot of  $w$  meaning weight.

# 01 What is a feature?

## 2. notation of features

- data table: a representation of data in a table
- A feature corresponds to one column name in the data table.
- data instance: one piece of data
  - Also called tuple, observation
  - line by line in excel

ID <int>	Gender <chr>	Grade <int>	Horoscope <chr>	Subject <chr>	IntExt <chr>	OptPest <chr>	ScreenTime <dbl>	Sleep <dbl>	PhysActive <int>
1	male	4	Scorpio	Math	Extravert	Optimist	1	7	10
2	female	4	Capricorn	Gym	Extravert	Optimist	1	8	5
3	male	4	Taurus	Math	Introvert	Optimist	4	9	22
4	male	4	Aquarius	Math	Don't Know	Don't Know	3	9	9
5	male	4	Scorpio	Gym	Don't Know	Don't Know	1	9	10
6	male	4	Pisces	Gym	Extravert	Optimist	2	9	20
7	male	3	Scorpio	Art	Introvert	Optimist	1	11	4
8	male	6	Taurus	Math	Extravert	Optimist	4	9	12
9	male	6	Aries	Gym	Introvert	Pessimist	6	8	4
10	male	6	Pisces	Math	Introvert	Don't Know	3	9	12

1-10 of 10 rows | 1-10 of 17 columns

# 01 What is a feature?

$$x^{(1)} = \begin{bmatrix} 1 \\ 0.00632 \\ 18 \\ 2.31 \\ 0.538 \\ \vdots \\ 24 \end{bmatrix} \quad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_{13} \end{bmatrix}$$

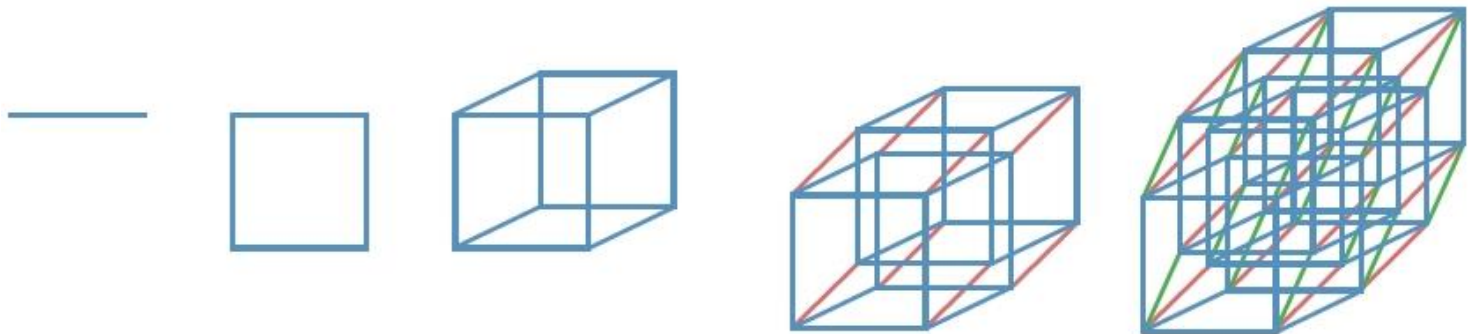
- One row is expressed as a column vector, and a weight is also expressed as a column vector.
- In  $x_j^i$ ,  $i$  is the order of data in the entire data table and  $j$  is the order of features.



# 01 What is a feature?

## 3. Curse of Dimensionality

- In actual machine learning, the number of features increases significantly, and accordingly, it is necessary to obtain a lot of data to improve model performance.
- As the dimension increases, we become unable to imagine or express it.



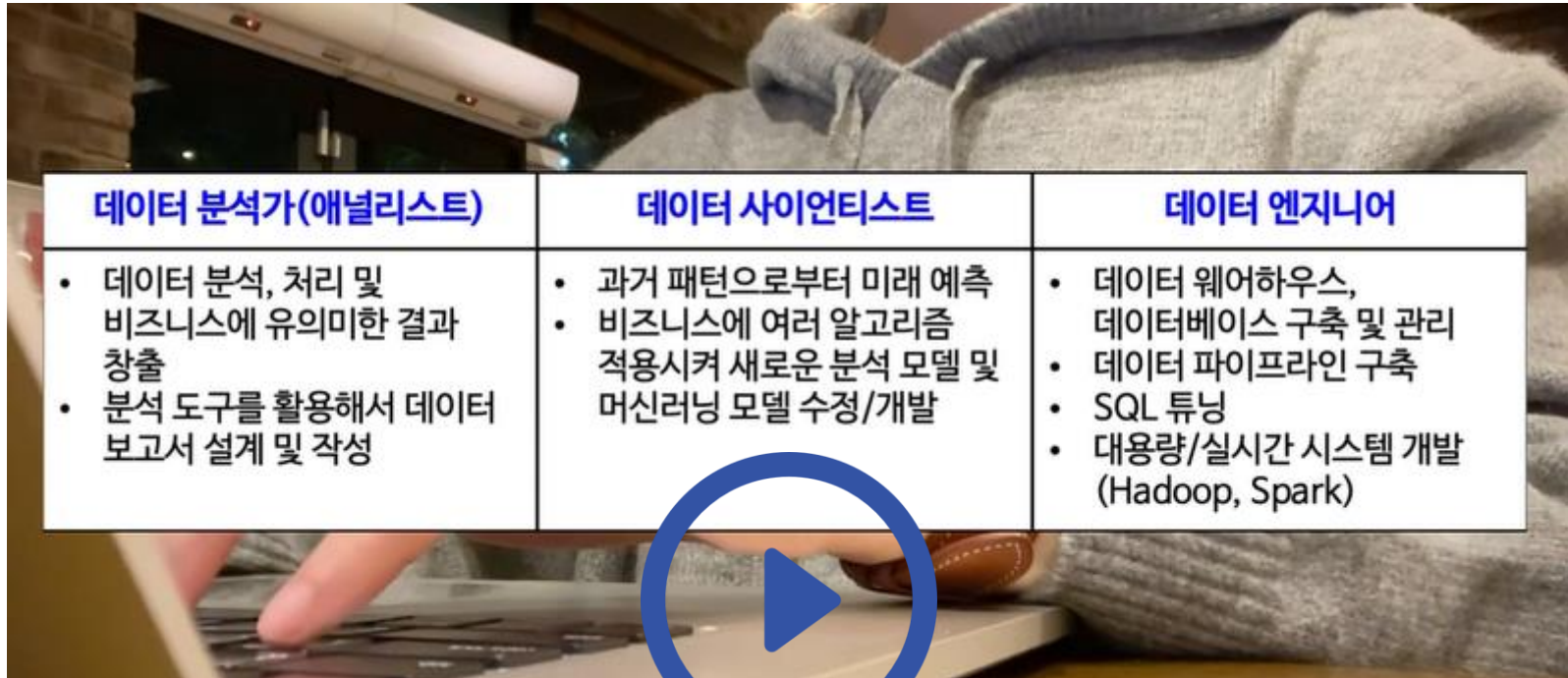
# 01 What is a feature?

## ■ 차원의 저주(Curse of Dimensionality) :

변수가 늘어남에 따라 차원이 커지면서 분석을 위한 최소한의 필요 데이터 건수가 늘어나면서 예측이 불안정해지는 문제를 말한다. 쉽게 생각해서 2차원 A4용지 위에 딱 차도록 구슬을 올려놓을 때와, A4용지 상자 안에 구슬을 딱 채워 넣을 때 필요한 구슬의 차이라고 보면 된다. 여기에서 차원은 데이터셋의 변수와 같은 의미다. 차원이란 공간 내에 있는 점의 위치를 나타내기 위해 필요한 축의 개수를 뜻한다.

일반적으로 한 개의 변수(차원)에 30건의 데이터가 필요하다. 따라서 만약 사용되는 변수가 20개면  $20 \times 30 = 600$ 건의 데이터가 최소한으로 필요하다. 이는 최소한이므로 정교한 모델을 위해서는 보다 많은 데이터를 확보해야 한다. 이처럼 변수가 늘어날수록 과적합(Overfitting)의 위험성이 증가한다. 또한 상관관계가 높은 변수로 인한 다중공선성 문제도 발생할 수 있어 많은 주의가 필요하다.

- 희박한 벡터 생성 : 벡터 공간에 너무 많은 0이 포함된 형태로 값이 없는 벡터들이 증가하여 모델 전체의 정확도를 떨어뜨림
- 데이터 처리 속도와 메모리 공간 문제 : 샘플 데이터가 너무 많아져 문제 발생



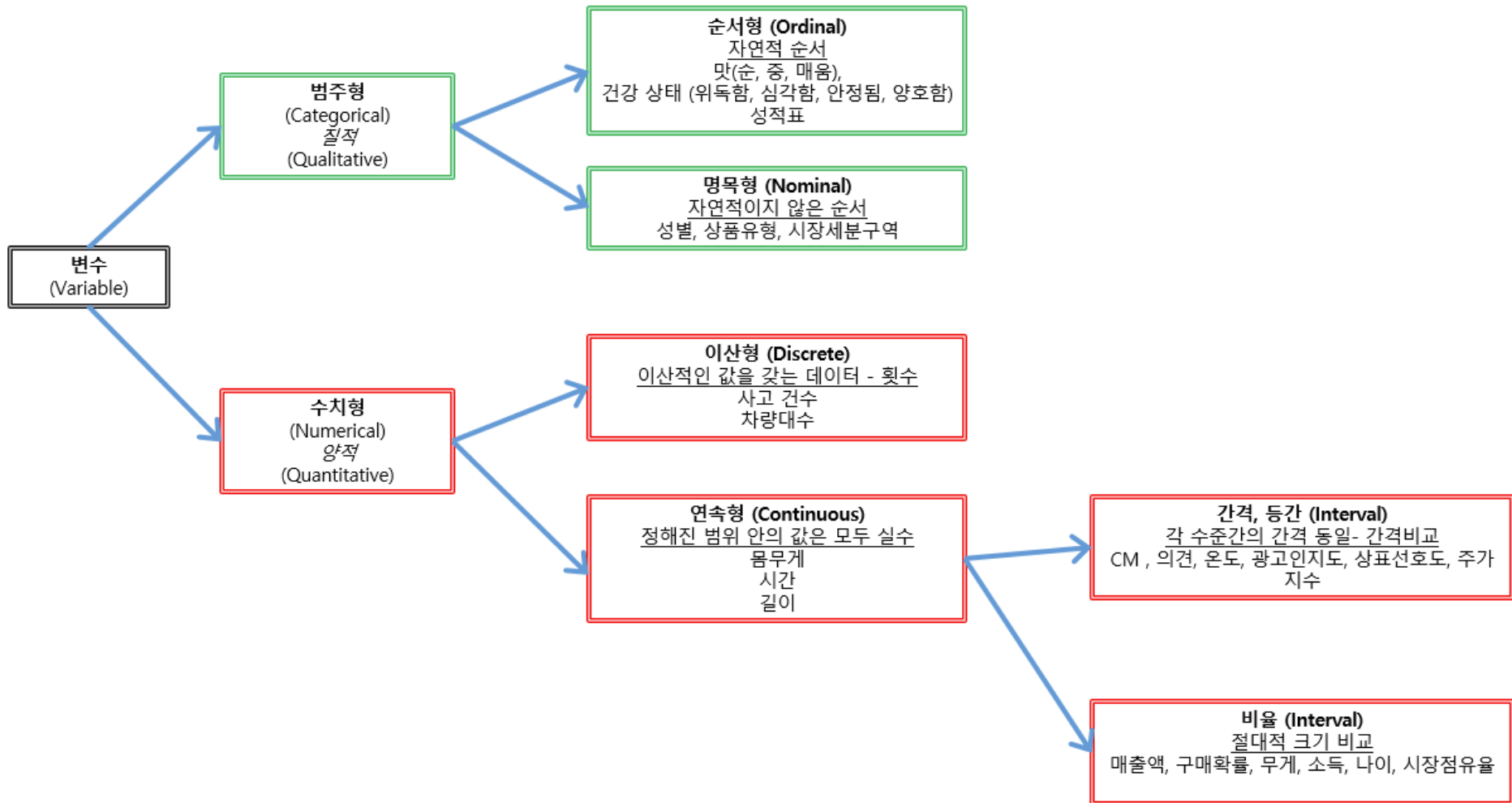
데이터 분석가(애널리스트)	데이터 사이언티스트	데이터 엔지니어
<ul style="list-style-type: none"><li>• 데이터 분석, 처리 및 비즈니스에 유의미한 결과 창출</li><li>• 분석 도구를 활용해서 데이터 보고서 설계 및 작성</li></ul>	<ul style="list-style-type: none"><li>• 과거 패턴으로부터 미래 예측</li><li>• 비즈니스에 여러 알고리즘 적용시켜 새로운 분석 모델 및 머신러닝 모델 수정/개발</li></ul>	<ul style="list-style-type: none"><li>• 데이터 웨어하우스, 데이터베이스 구축 및 관리</li><li>• 데이터 파이프라인 구축</li><li>• SQL 튜닝</li><li>• 대용량/실시간 시스템 개발 (Hadoop, Spark)</li></ul>

**02**

# **피쳐의 종류**

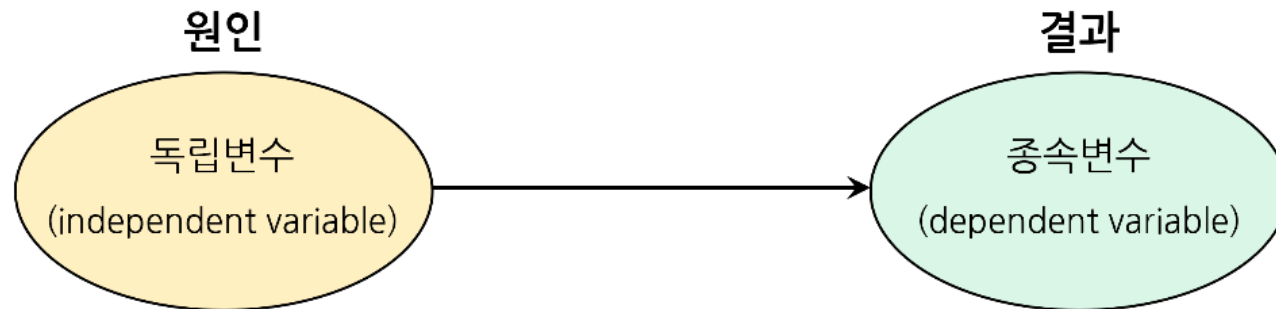
# 02 Type of feature

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT.MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	0



# 02 Type of feature

## 1. independent and dependent variable



= 설명변수 (explanatory variable)  
= 입력변수 (input variable)  
= 예측변수 (predictor variable)  
= 조작 변수 (manipulated variable)  
= 특징 (feature)

= 반응변수 (response variable)  
= 출력변수 (outcome variable)  
= 피예측변수 (predicted variable)  
= 측정변수 (measured variable)  
= 표적변수 (target variable)

# 02 Type of feature

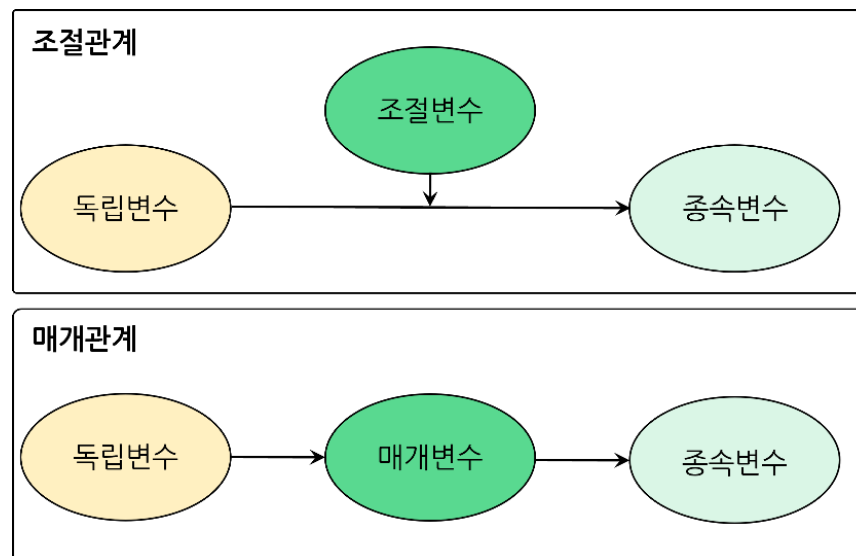
## 2. Variable Relationship Types

- **Causal relationship:** This is the basic relationship between the independent variable and the dependent variable. A variable has a causative effect on another variable.
- **Correlational relationship:** A relationship in which a relationship exists between variables. It is a superordinate concept of Causal relationship.
- **Spurious relationship:** A relationship in which there is a correlation between variables, but the correlation is caused by another variable.

# 02 Type of feature

## 2. Variable Relationship Types

- **Moderating relationship:** A relationship that has an indeterminate effect between the independent variable and the dependent variable.
- **Mediational relationship:** A relationship between the independent variable and the dependent variable in which the parameter transmits the influence.





# 02 Type of feature

## 3. Type of scale

- **Nominal scale:** It is a scale created for the purpose of classifying the properties or categories of the research object.
- **Ordinal scale:** It is a scale that measures the ordinal relationship between the objects by measuring the attribute size of the object to be investigated.



# 02 Type of feature

## 3. Type of scale

- **Interval scale:** In addition to the information possessed by the ranking scale, it also has information that can compare the difference in the 'relative size' of the properties of the survey subject.
- **Ratio scale:** Includes information on the relative size between objects and ratio information through absolute standards.

종류		포함 정보			
질적척도	명목척도	범주			
	서열척도	범주	순서		
양적척도	등간척도	범주	순서	상대적 크기	
	비율척도	범주	순서	상대적 크기	절대적 크기

**03**

# **데이터를 모델에 대입하기**

# 03 Putting data into the model

## 1. Basic terminology for data tables

- **Data table:** A data set in the form of a table that can be checked when data is loaded using pandas.
- **Feature:** field or column in Excel, **attribute** in database
- **Instance:** One set of data for the same object. An entire data collection of all features for a single object. **Rows** in excel, **tuples** in database. And **observations**.

# 03 Putting data into the model

data table, sample

데이터 테이블, 샘플

Attributes, Fields,  
Features, Columns

속성, 필드, 피쳐, 열

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.9	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	48.9	4.9671	2	242.0	17.8	396.9	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.9	5.33	36.2
5	0.02985	0.0	2.18	0	0.458	6.43	58.7	6.0622	3	222.0	18.7	394.12	5.21	28.7
6	0.08829	12.5	2.18	0	0.524	6.012	66.6	5.5605	5	311.0	15.2	395.6	12.43	22.9
7	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311.0	15.2	396.9	19.15	27.1
8	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311.0	15.2	386.63	29.93	16.5
9	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311.0	15.2	386.71	17.1	18.9
10	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311.0	15.2	392.52	20.45	15.0

인스턴스, 튜플, 행

Instances,  
Tuples, Rows

피쳐 벡터

Feature vector

데이터

Data

# 03 Putting data into the model

<https://bit.ly/3rhZOzb>

```
In [1]: import pandas as pd                # (1) pandas 모듈 호출

data_url = 'https://archive.ics.uci.edu/ml/machine-learning-
databases/housing/housing.data' # (2) 데이터 URL을 변수 data_url에 넣기
df_data = pd.read_csv(data_url, sep='\s+',
    header = None) # (3) csv 데이터 로드
df_data.columns = ['CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE',
    'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT', 'MEDV'] # (4) 데이터
의 열 이름 지정
df_data.head()                # (5) 데이터 출력
```

Out  
[1]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.9	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	48.9	4.9671	2	242.0	17.8	396.9	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.9	5.33	36.2

# 03 Putting data into the model

## 2.2 Apply formulas to data

$$y = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \dots + w_{13}x_{13} + w_0 \times 1$$

- Transform the formula because the data is in the form of a matrix rather than a vector

$$y = w_1x_1 + w_2x_2 + \dots + w_{13}x_{13} + w_0x_0 = \sum_{j=0}^{13} w_jx_j = W^T X$$

- Multiplication of two vectors can be expressed as the dot product of a vector

## 03 Putting data into the model

```
In [2]: data_url = 'https://archive.ics.uci.edu/ml/machine-learning-
databases/housing/housing.data'
df_data = pd.read_csv(data_url, sep='Ws+', header = None)
df_data.columns = ['CRIM','ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE',
'DIS', 'RAD', 'TAX', 'PTRATIO', 'B', 'LSTAT', 'MEDV']
df_data.head()          # 이전 코드와 동일한 코드
                        (pandas 모듈은 위에서 호출 완료)
import numpy as np      # (1) numpy 모듈 호출
df_data['weight_0'] = 1  # (2) weight 0 값 추가
df_data = df_data.drop("MEDV", axis=1) # (3) Y 값 제거
df_matrix = df_data.values
                        # (4) 행렬(Matrix) 데이터로 변환하기
weight_vector = np.random.random_sample((14, 1))
                        # (5) 가중치 w 생성
df_matrix.dot(weight_vector) # (6) 내적 연산 실행 결과 출력
```



## 03 Putting data into the model

```
Out [2]: array([[236.92351326],  
                [235.27213482],  
                [217.75053128],  
                [198.01232971],  
                [206.41387943],  
                [209.7193267 ],  
                [248.05381244],  
                [276.10651715],  
                [278.84025893],  
                [264.92706025],  
                [274.07244333],  
                [263.68862219],  
                [222.49940062],  
                [237.02203146],  
                [253.5816465 ]],
```