HEARTCOUNT

KOREA

LOGIN

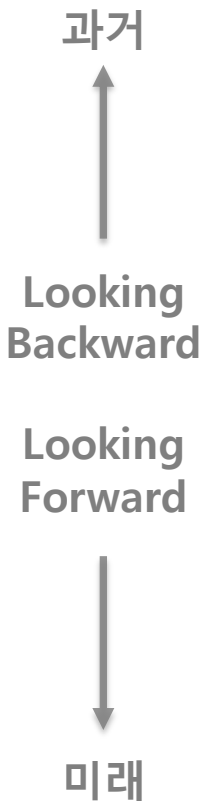
Module II-a
Data Understanding

아이디케이스퀘어드 양승준 / sidney.yang@idk2.co.kr
https://www.heartcount.io

데이터 분석 방법 (X와 Y)

분석하는 이유	• 궁금한 것(Y)을 데이터(X)로 더 잘 설명(예측) • X를 바꾸어서 Y를 개선하기 위해서
엑셀 (대쉬보드)	• 성과지표(Y)를 익숙한 관점(범주; X)으로 요약 • 과거에 대한 집계
데이터 시각화	• X와 Y를 점, 선, 크기, 색상으로 표현 (탐험분석) • X와 Y 사이의 패턴(관계) 시각적 발견; 가설 수립
통계	• 데이터의 특성과 모양 요약 (기술 분석) • 독립변수(통제가능; X)와 종속변수(Y) 간 가설 검증
기계학습	• 데이터 학습, Feature(X)로 Target(Y)을 예측·설명 • 의사결정 자동화 vs. 더 좋은 의사결정

데이터 분석 주요기술

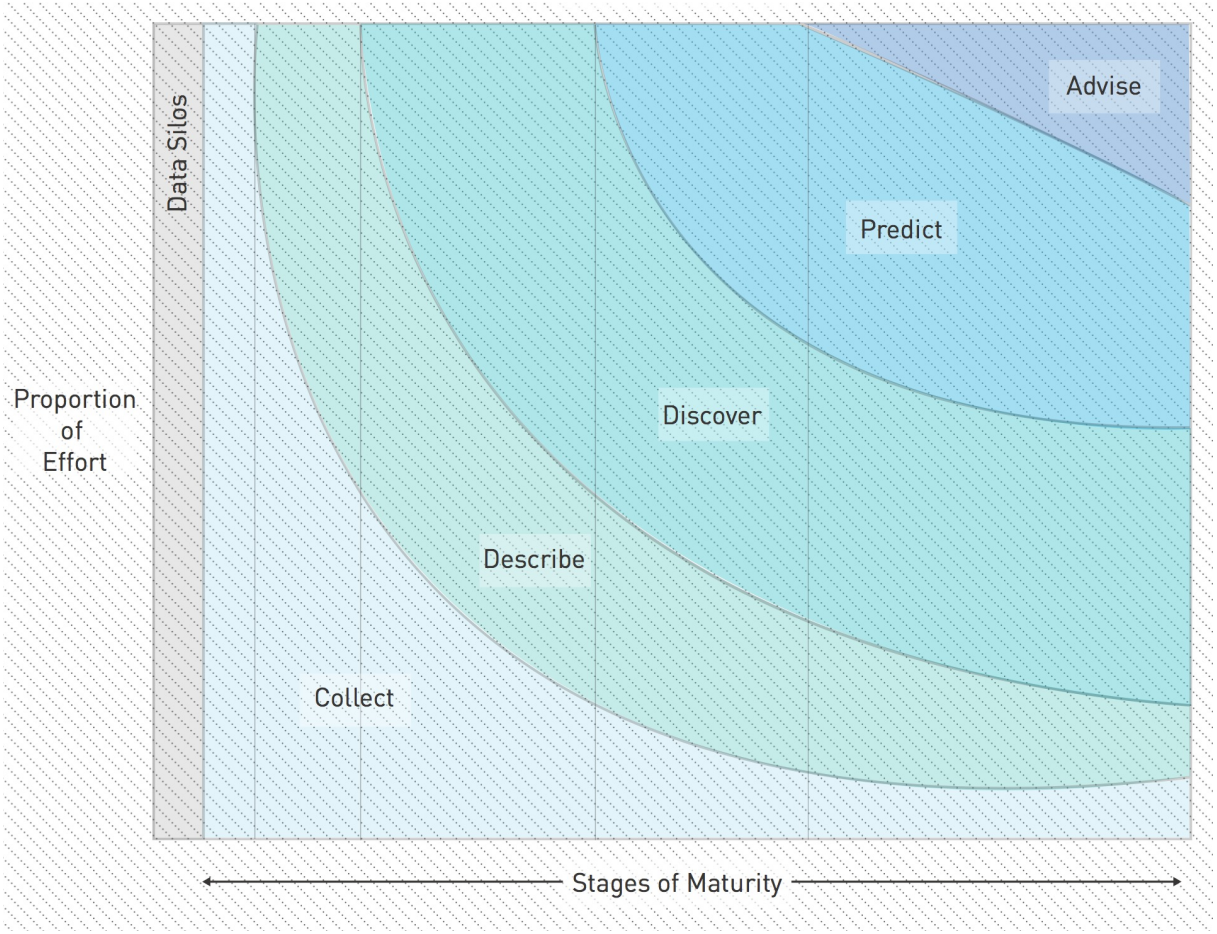


- **DESCRIBE (기술 분석)** - 엑셀
  - 데이터 특성과 모양을 (수치적으로) 요약
- **EXPLORE (탐험적 분석)** - 데이터 시각화 도구
  - 가설수립·데이터 감 잡기 위해 패턴 탐험
- **PREDICT/INFER (예측·추론 분석)** - 통계/ML
  - 패턴(모형)을 통해 주어진 문제를 예측·설명

Data Analysis Maturity Model

데이터 수집 → 데이터 기술(묘사) → 패턴 발견 → 예측 → 활용

우측으로 갈수록 성숙해진다기보다는 자기에게 필요한 단계를 잘 하면 됨



Source: Booz Allen Hamilton

EDA(Exploratory Data Analysis) = DESCRIBE (기술 분석) + EXPLORE (탐험 분석)

EDA, 데이터와 함께 떠나는 창의적 여행

- inspect data structure
- data quality
- summarize
- visualize data
- hypothesis generation
- != modeling

Source: Booz Allen Hamilton

데이터에 대해 사실적으로 묘사하는 법

Description
요약

변수의 대표값과
모양이 어떻나?

개별 변수(Y)의
통계값과 분포 확인

Comparison
비교

변수값의 차이가
어디서 얼마나 나나?

서로 다른 범주(X) 간
Y의 특성·모양 비교

Relationship
관계

변수(Y)의 변화와 관계를
갖는 다른 변수(X)는?

X와 Y 사이의
상관관계 파악

Analysis-Ready Dataset

분석하기 좋은 데이터셋

- 국가별로 1999/2000년에 결핵으로 사망한 환자수(Cases)와 전체인구(Population)를 정리한 데이터셋들
- 국가별 연도별 인구 10,000명당 결핵 사망률을 계산하기 가장 좋은 데이터는?

NOT SO GREAT

1

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

2

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

3

country	year	population
Afghanistan	1999	745 / 19987071
Afghanistan	2000	2666 / 20595360
Brazil	1999	37737 / 172006362
Brazil	2000	80488 / 174504898
China	1999	212258 / 1272915272
China	2000	213766 / 1280428583

GREAT

4

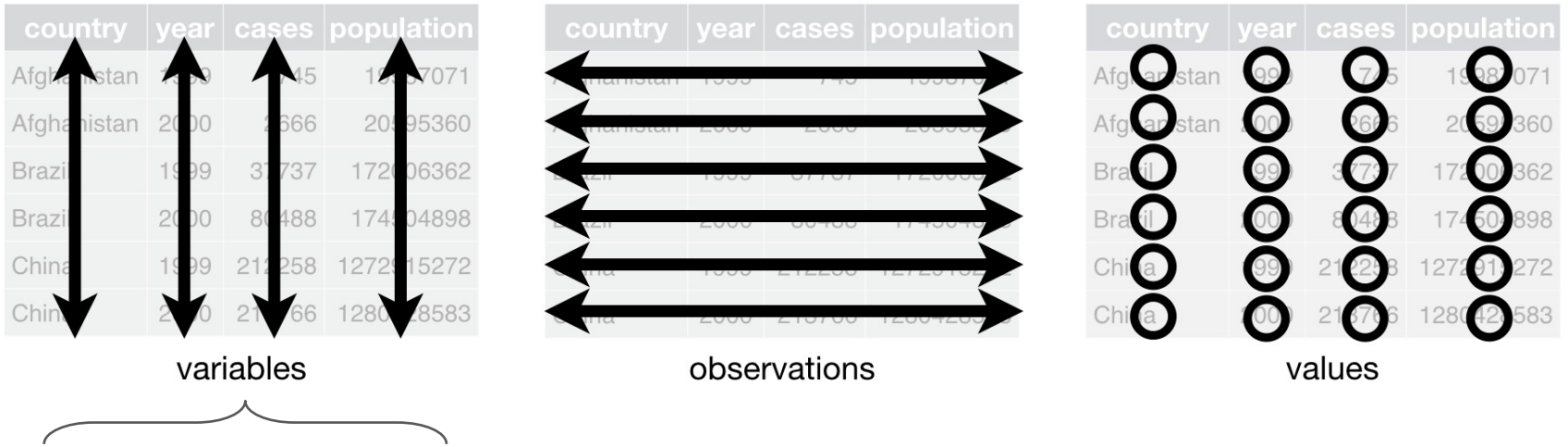
country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

rectangular data
data frame
data table
tidy dataset

Rectangular Dataset and Key Terms

분석하기 좋은 데이터셋

- **Dataset:** 값(Values)들의 집합으로 숫자 또는 범주로 구성
- **Values:** 변수(Variable)와 관측점(Observation)으로 구성
- **Variable:** 동일한 속성(나이, 매출)에 대한 측정값들로 행(Column)을 구성
- **Observation:** 동일한 대상(사람, 매장)에 대한 측정값들로 열(Row)를 구성



X
features
independent variables
input (variables)
predictor
attribute

Y
target
dependent variables
output (variable)
response

record
sample
Instance
case

Raw vs. Aggregated Dataset

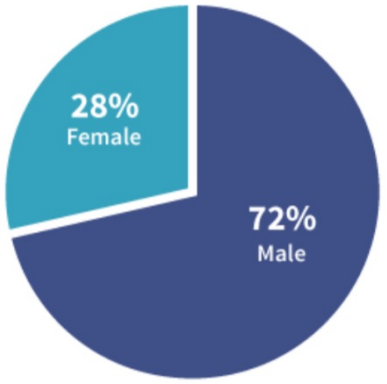
Raw Data

Name	Gender	Coffee
Bob Smith	M	Regular
Jane Doe	F	Regular
Dale Cooper	M	Mocha
Mary Brewer	F	Decaf
Betty Kona	F	Regular
John Java	M	Regular
Bill Bean	M	Regular
Jake Beatnik	M	Mocha
Bob Smith	M	Regular
Jane Doe	F	Regular
Dale Cooper	M	Mocha
Mary Brewer	F	Regular
John Java	M	Decaf
Bill Bean	M	Regular

Aggregated Data

Year	2000	2001	2002
Total sales	19,795	23,005	31,711
Male	12,534	16,452	19,362
Female	7,261	6,553	12,349
Regular	9,929	14,021	17,364
Decaf	6,744	6,833	10,201
Mocha	3,122	2,151	4,146

Q. 2001년 남녀 구매 비율은?

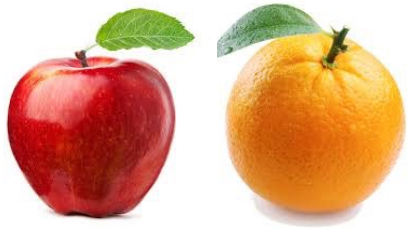


추가 질문
Q. 2001년 Regular Coffee 구매한 여자 고객수?
Q. 남자 고객이 선호하는 커피 종류는?

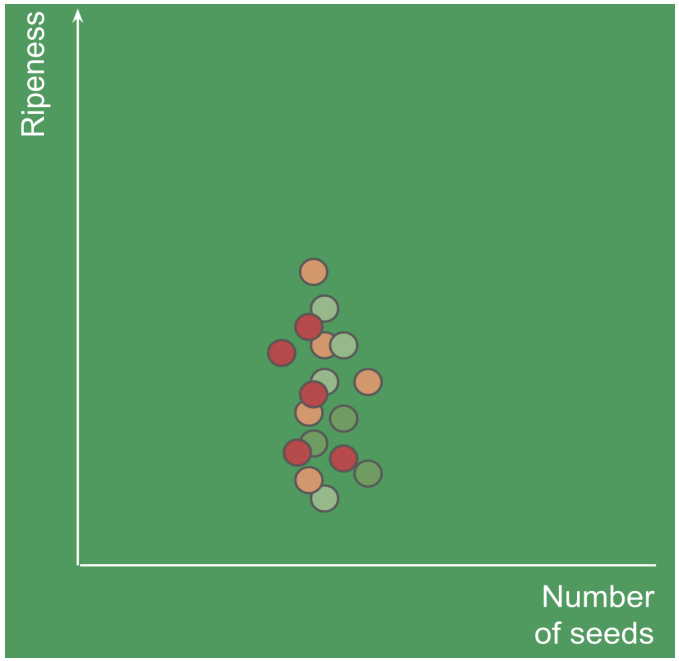
Raw Data: Zoom-in(새로운 질문) 가능

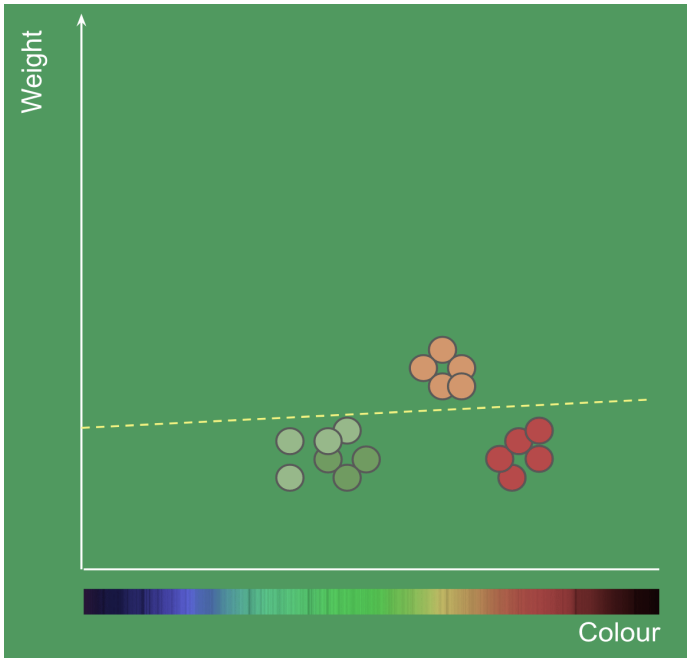
Features / Attributes / X

Feature: Y(Output/Target)를 설명하거나 분류(예측)하는데 사용되는 속성
<u>좋은 Feature</u>를 발굴하는 것이 참 중요함.



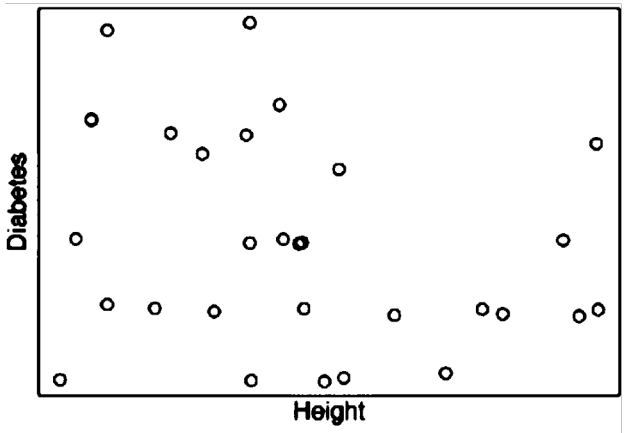
	Ripeness	# of Seeds	Weight (g)	Color	Fruit
	0.56	5	320	Orange	Orange
	0.61	6	280	Red	Apple





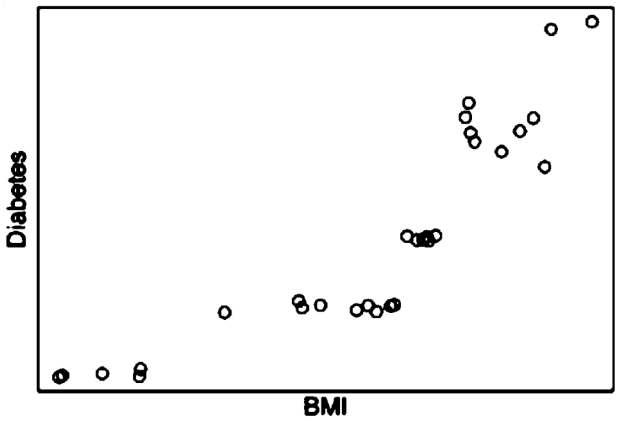
Feature Engineering: From Raw Variable to Derived Variable

Y를 더 잘 설명하거나 분류(예측)할 수 있도록
기존 변수를 창의적으로 가공하여 새로운 변수를 만드는 일



당뇨병 위험도와 상관관계가 높은 변수

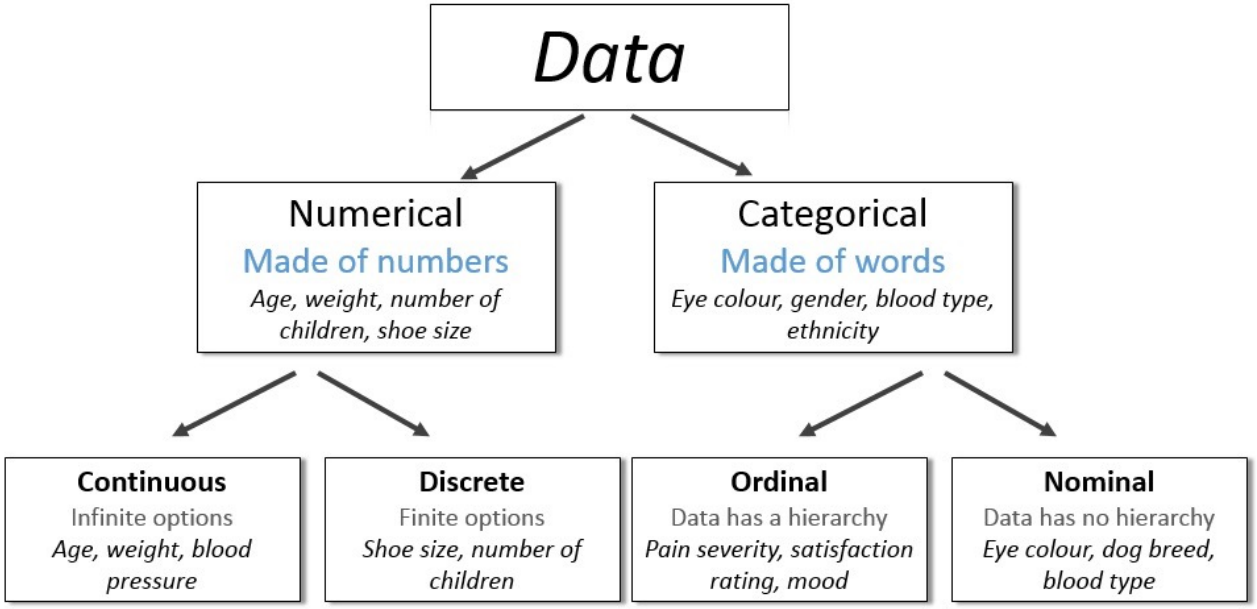
- 같은 몸무게라도 비만도는 키에 좌우됨
- 비만도를 더 잘 반영할 수 있는(키와 몸무게의 상호작용을 잡아낼 수 있는) 새로운 변수 가공
- ***BMI(Body Mass Index) = kg/m²**



*발명한 사람의 이름을 따서
Quetelet Index라고도 함

숫자형(Quantitative)과 범주형(Qualitative)

분석: 숫자와 숫자 사이의 연관성, 숫자의 차이를 가져오는 범주를 발견하는 것

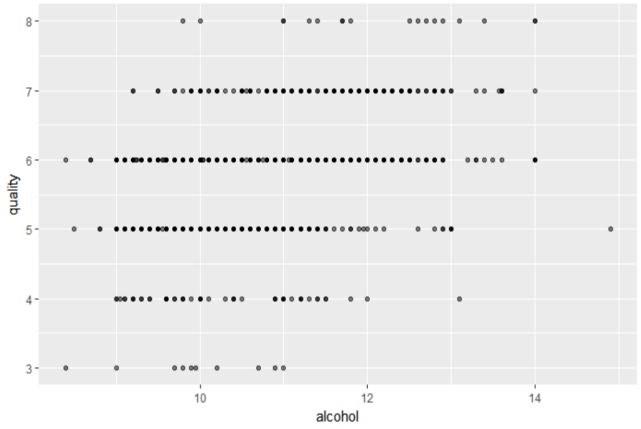


- **숫자**형 자료는 **이산형(discrete)**이나 **연속형(continuous)**으로 나뉨
- **범주**형 자료는 **명목형(nominal)**이나 **순서형(ordinal)**으로 나뉨

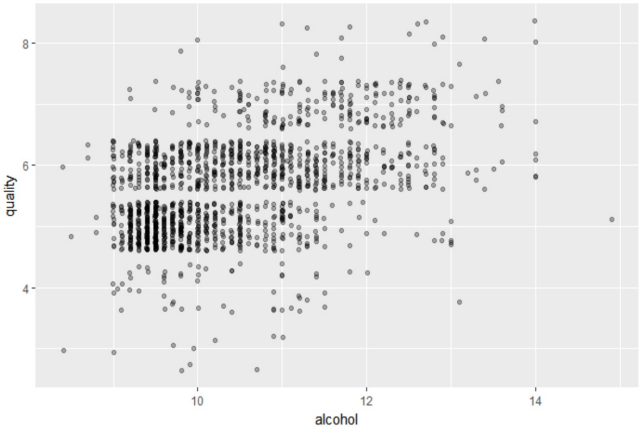
변수 유형에 따라 분석 방법과 효과적 시각화 방법이 달라짐

Alcohol(%): 와인 알코올 함량, Quality: 소비자가 매긴 점수

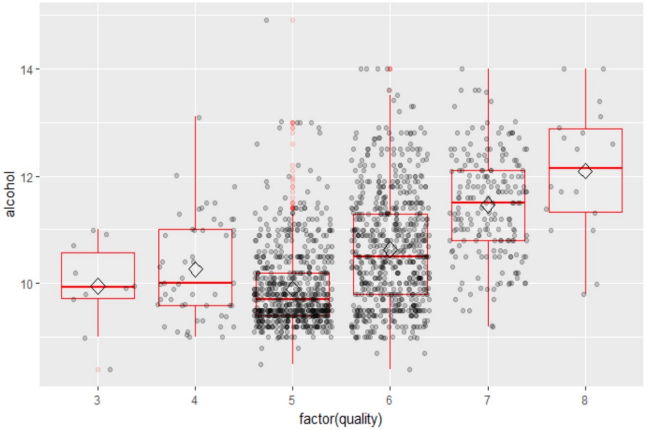
숫자 x 숫자 = Scatterplot
Overplotting(점이 겹침)!



Jittering 기법으로 Noise 추가
Jittering(인위적으로 퍼지게)!



Quality를 범주로 처리
Boxplotting(분포 시각화)!



순서형(Ordinal) 변수는 범주(Category)로 다루는 게 좋다!