데이터 전처리 워크샵

kt

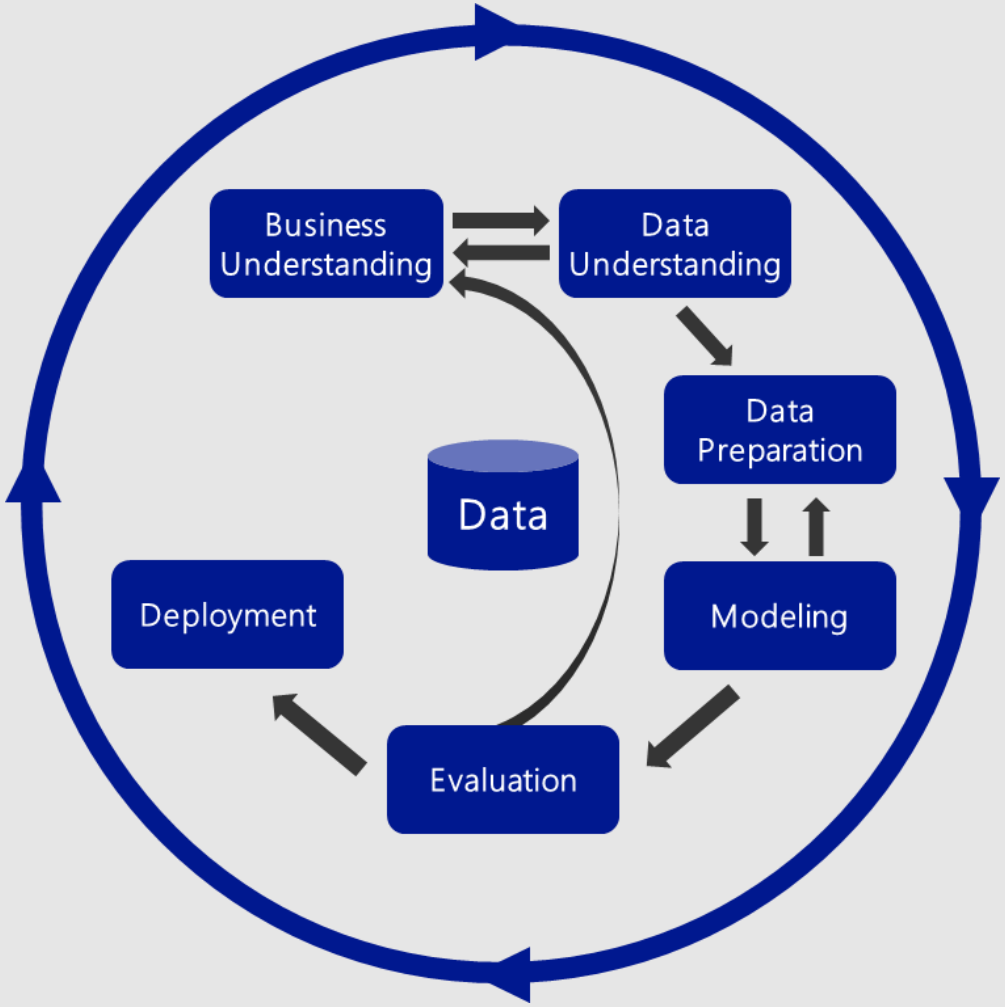
Data Insight

- ✓ 데이터 분석 표준 프로세스 CRISP-DM
- ✓ 분석을 위한 데이터 구조
- ✓ 실습① WHO 국가별 결핵 Case
- ✓ 추가변수
- ✓ 모델링을 위한 전처리
- ✓ 실습② 주가예측 전처리
- ✓ 종합실습③ 고객 이탈 예측 전처리

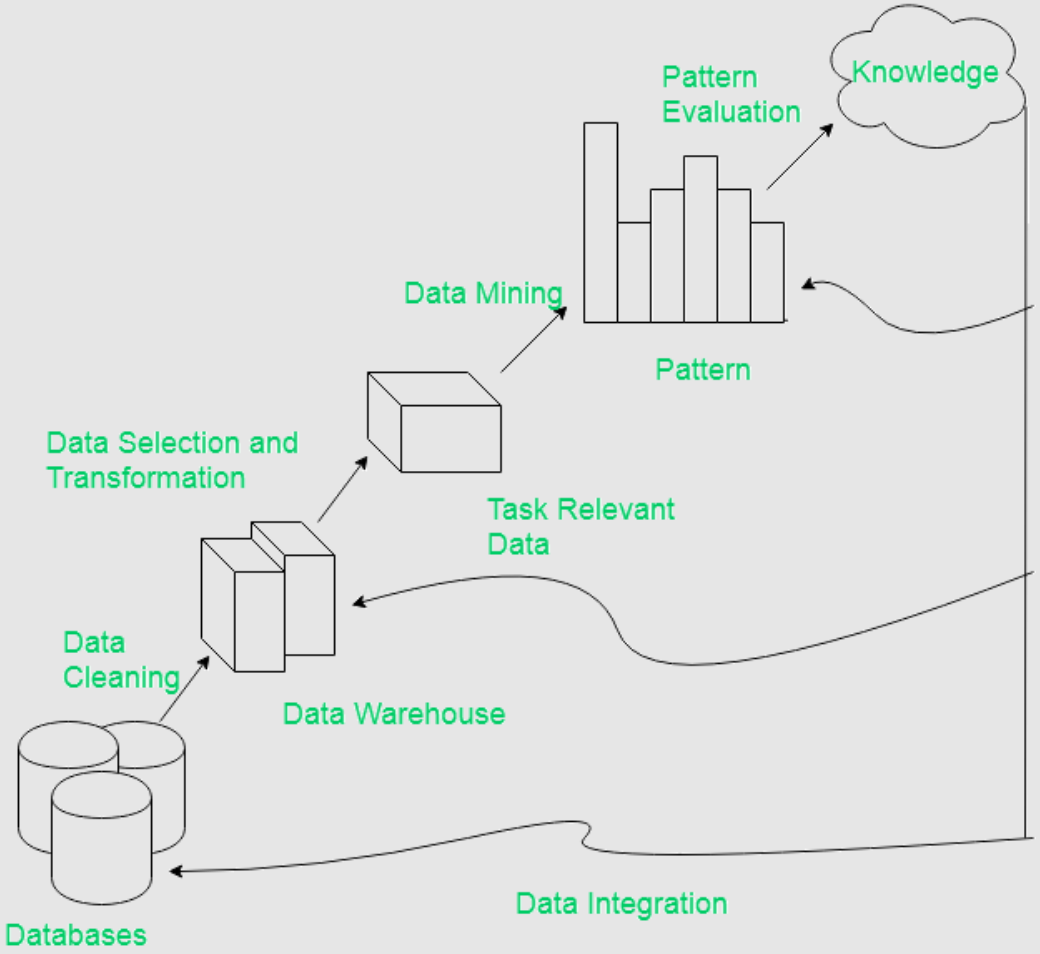
[데이터분석 표준 프로세스]

데이터 마이닝 표준 프로세스

Cross-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

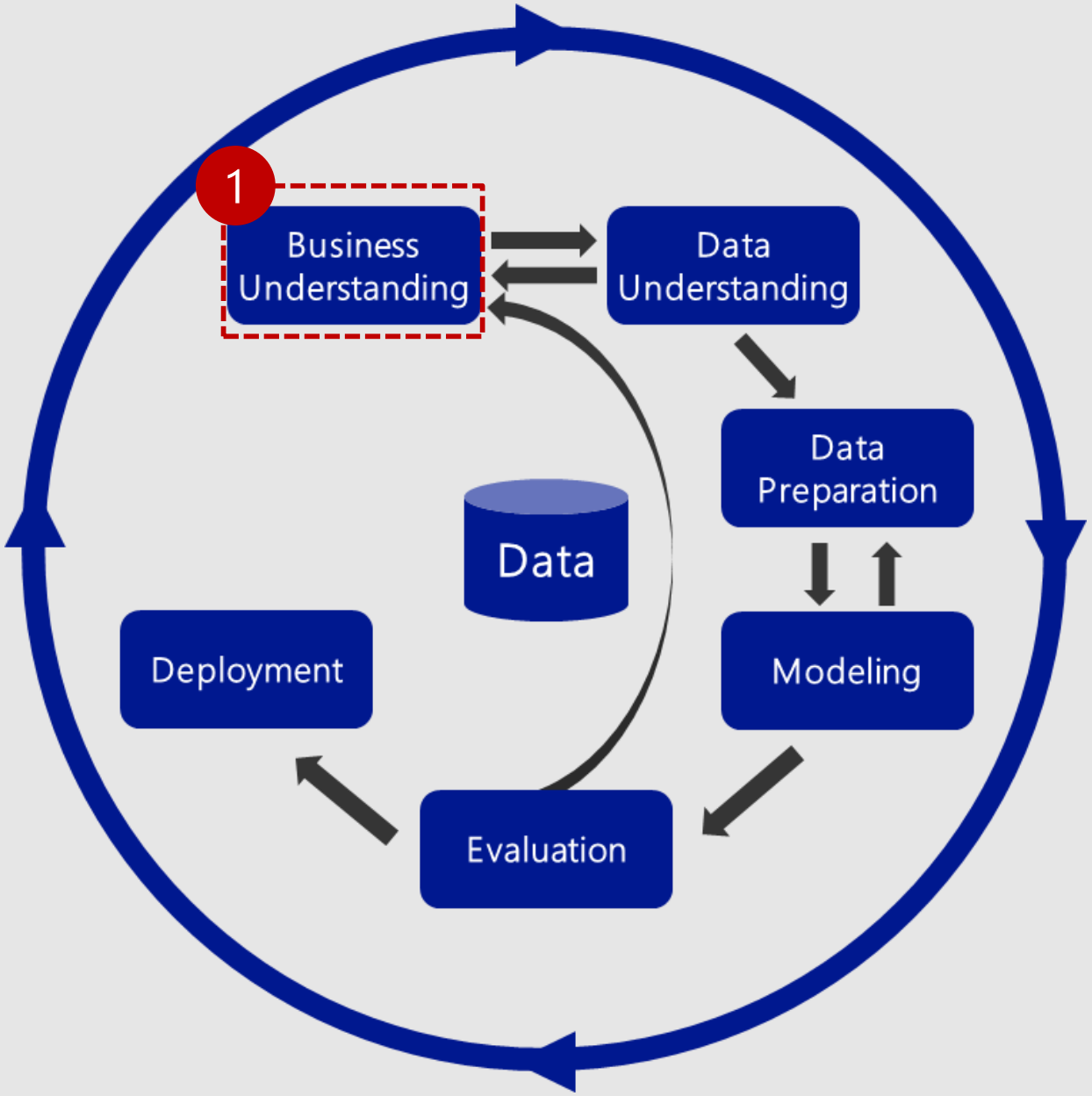


Knowledge **D**ata **D**iscovery



CRISP-DM

- ✓ 프로세스를 한번 거쳤음에도 문제가 해결되지 않을 수 있다
  ➔ 그렇다고 실패가 아님!
- ✓ 한번에 해결책을 찾지 못해도 데이터를 더 잘 이해하게 되는 계기가 됨
  ➔ 두번째 수행할 때는 더 많은 정보를 갖고 시작할 수 있음!



①Business Understanding

✓개요

- 잘 정의된 명확한 데이터분석 문제로 시작하는 프로젝트는 거의 없음.
- 문제를 파악해 가는 과정을 반복하면서 문제를 재정의하고 해결책을 정의하게 됨.

✓수행되는 내용

- 비즈니스 목표 검토
- 데이터 분석 목표 수립
- (초기)가설 수립

①Business Understanding

✓비즈니스 목표에서 데이터 분석 목표로…

비즈니스 관점	**목표**	올해 은행 대출 부서의 수익 1000억 달성 (작년 수익액 600억)
	방법	✓ 신용도 높은 사람의 대출 신청 승인 ✓ 신용도 낮은 사람의 대출 신청 거절
데이터 분석 관점	**문제정의**	대출 신청자들의 신용도를 예측할 수 있을까?
	목표	어느 정도 정확도로 예측할 수 있다면, ▪ 비즈니스 목표를 달성 할 수 있을까? ▪ 2년 이내 프로젝트 투자에 대한 BEP에 도달할 수 있을까?
	분석	▪ 분류문제 ▪ **신용도**에 영향을 미치는 **요인**은 무엇일까?

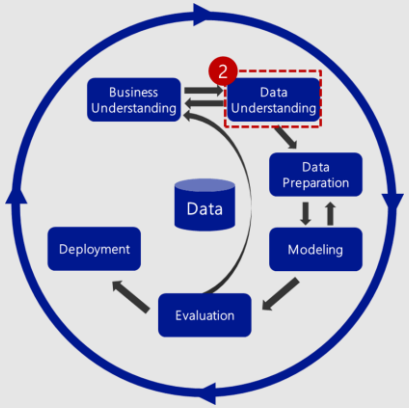
①Business Understanding

✓(초기)가설 수립

신용도에 영향을 미치는 **요인**은 무엇일까?

- 다양한 직무에 있는 사람들의 의견을 수렴할 필요가 있음.
- 데이터의 존재여부를 고려하지 말고 가설 도출.
- 초기 가설 수립 이후 데이터 탐색을 통해 가설을 구체화

②Data Understanding



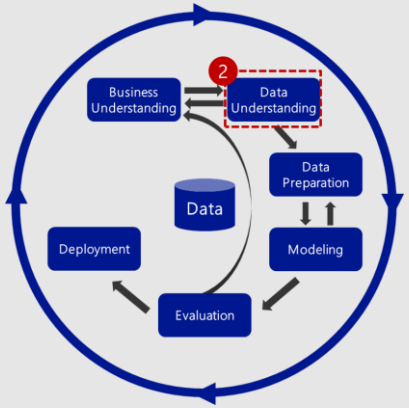
✓개요

- 데이터 : 문제의 해결책을 만드는 데 사용할 원자재
- **문제에 정확히 부합하는 데이터가 있는 경우는 거의 없음.**
- 데이터에 따라 데이터 취득 및 유지 비용이 다름.

✓수행되는 내용

- 데이터 원본 식별 및 취득
- 데이터 탐색 : EDA, CDA

②Data Understanding



✓데이터 원본 식별 및 취득

- (초기)가설에서 도출된 데이터의 원본을 확인

있는 데이터

없는 데이터

가용한 데이터

그대로 사용가능

가공해야 사용가능

취득 가능

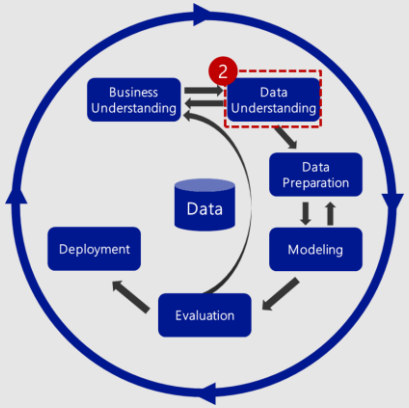
취득 불가능

Data Pipeline
Data Environment
검토 및 과제 수행

취득 비용 산정
과제 수행

정보의 의미 정의
→ 정보 분할
→ 최대한 가용한 데이터 영역으로 만들기

②Data Understanding



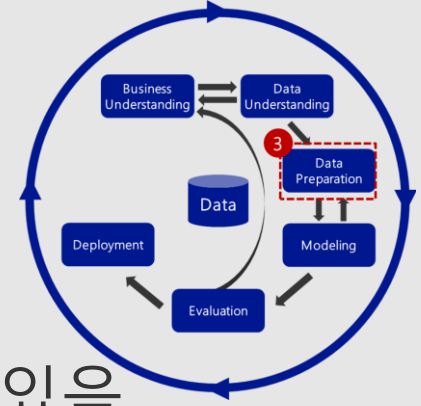
✓ 데이터 탐색 : EDA, CDA

- 데이터를 탐색하는 두 가지 방법

데이터 통계량	데이터 시각화
분할표(Contingency Table) MIN, MAX, SUM, MEAN Quartile …	Histogram, Box plot, Density plot Bar plot, Pie chart Scatter plot …

- EDA (Exploratory Data Analysis)
  - 개별 데이터의 분포, 가설이 맞는지 파악
  - NA, 이상치 파악
- CDA (Confirmatory Data Analysis)
  - 탐색으로 파악하기 애매한 정보는 통계적 분석 도구(가설 검정) 사용

③ Data Preparation



✓ 개요
- 데이터 분석을 위해 특정 조건에 맞는 데이터 유형과 구조가 있음
- 더 좋은 결과를 얻을 수 있도록 데이터의 형태를 조작하고 변환하는 과정 필요.

✓ 수행되는 내용
- 데이터 정제
- 추가 변수(Feature Engineering)

✓ 결과물 : **하나의 잘 정리/정제된 데이터프레임(테이블)**

분석할 수 있는 데이터?

✓**연속형**
- 숫자
- 날짜

✓**예**
- 주문일
- 판매량
- 금액
- 나이

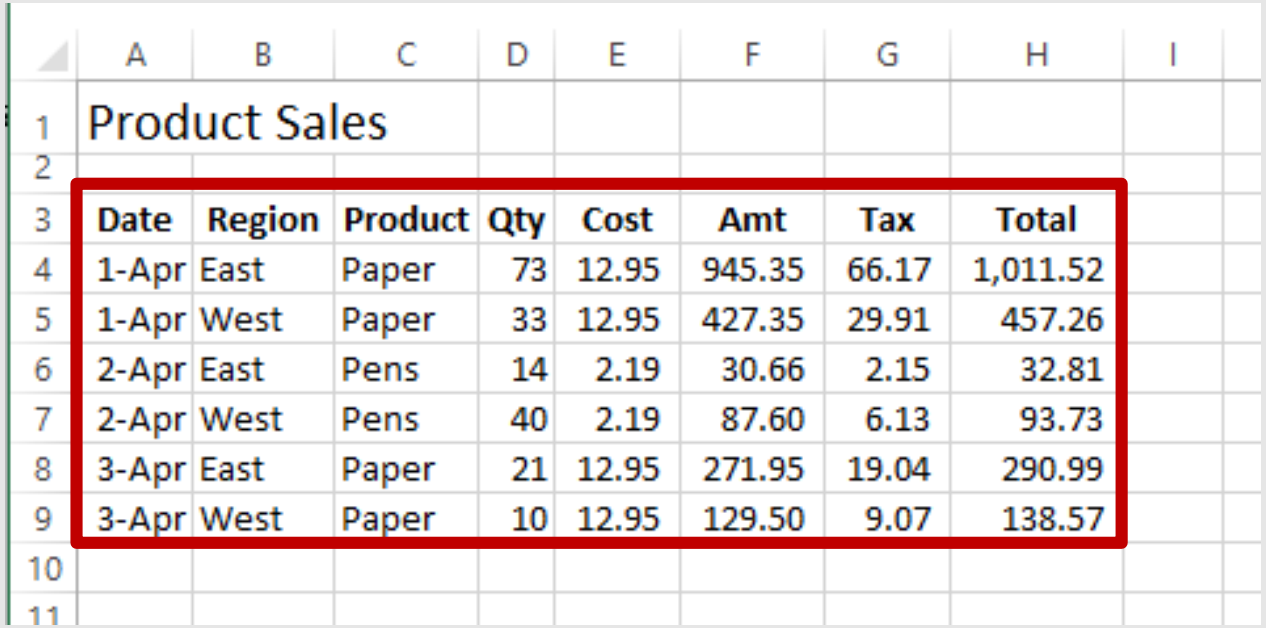
✓ **범주형**
- 순서형, 명목형, 이항형

✓ **예**
- 상품카테고리
- 성별
- 고객
- 지역
- 연령대

분석을 위한 데이터 구조

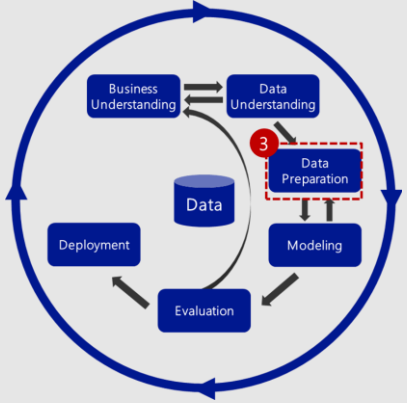
✓ Table **형태**

① 데이터가 아래로 쌓이는 구조
② 열(column) 성격에 맞는 데이터가 입력되어야 함
③ 열(column) 마다 데이터 형식이 있음



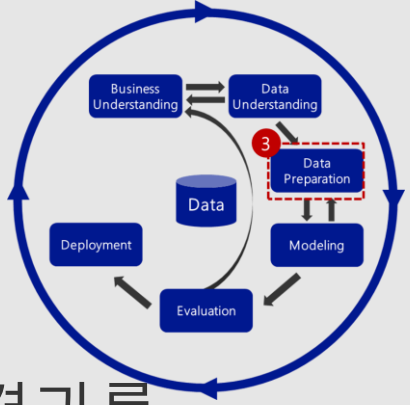
	A	B	C	D	E	F	G	H	I
1	Product Sales								
2									
3	**Date**	**Region**	**Product**	**Qty**	**Cost**	**Amt**	**Tax**	**Total**	
4	1-Apr	East	Paper	73	12.95	945.35	66.17	1,011.52	
5	1-Apr	West	Paper	33	12.95	427.35	29.91	457.26	
6	2-Apr	East	Pens	14	2.19	30.66	2.15	32.81	
7	2-Apr	West	Pens	40	2.19	87.60	6.13	93.73	
8	3-Apr	East	Paper	21	12.95	271.95	19.04	290.99	
9	3-Apr	West	Paper	10	12.95	129.50	9.07	138.57	
10									
11									

③ Data Preparation



- ✓ 데이터 정제
  - 잘못된 데이터 정제
  - 결측치(NA) 식별 및 조치
    - 중요한 요인에 결측치가 존재한다면 반드시 조치해야 한다.
    - 예 : 옷을 추천하는데, 고객의 나이나 성별에 결측치가 존재한다면, 옷을 추천하기 곤란.
  - 이상치 식별 및 조치
    - 잘못된 값
    - 값 자체는 정상이나 다른 값들의 분포에 비해 치우친 값
    - 이러한 값은 데이터 분석 시 잘못된 결과를 얻게 하는 원인이 됩니다.

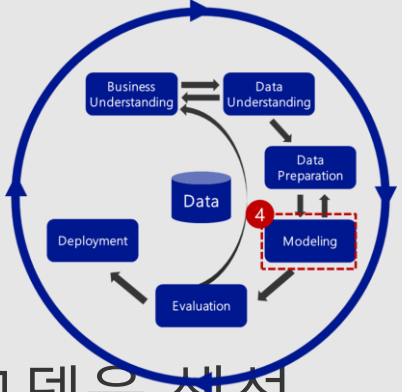
③ Data Preparation



✓추가변수(Feature Engineering)

- 기존에 저장된 데이터를 그대로 사용해서는 제대로 된 예측 결과를 얻기 어렵다.
- 데이터베이스에 데이터를 저장하는 방식
  - 트랜잭션 발생 순으로 저장 ➔ 저장된 데이터 자체가 비즈니스의 Insight가 되지 못함.
- 비즈니스의 경험 + 데이터 분석을 통해 인사이트를 발견하고, 이를 담아내는 정보가 필요
- 사례
  - 페이스북 고객 중 **가입 후 10일 이내 7명의 친구를 사귄 사람**은 그렇지 않은 사람보다 잔존율이 훨씬 높다!
  - 음주 습관에 대한 분석 : age 변수를 이용해서 age >= 20 ➔ 음주가능연령
  - 아파트가격 분석 : 방 수 >=4 & 화장실수 >=2 ➔ Premium

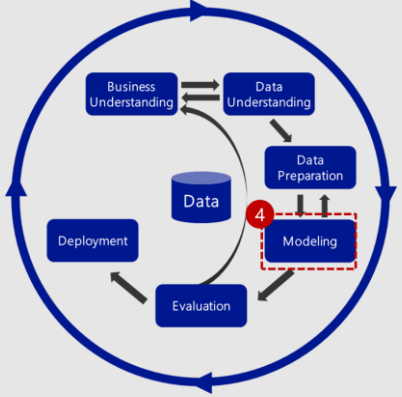
④ Modeling



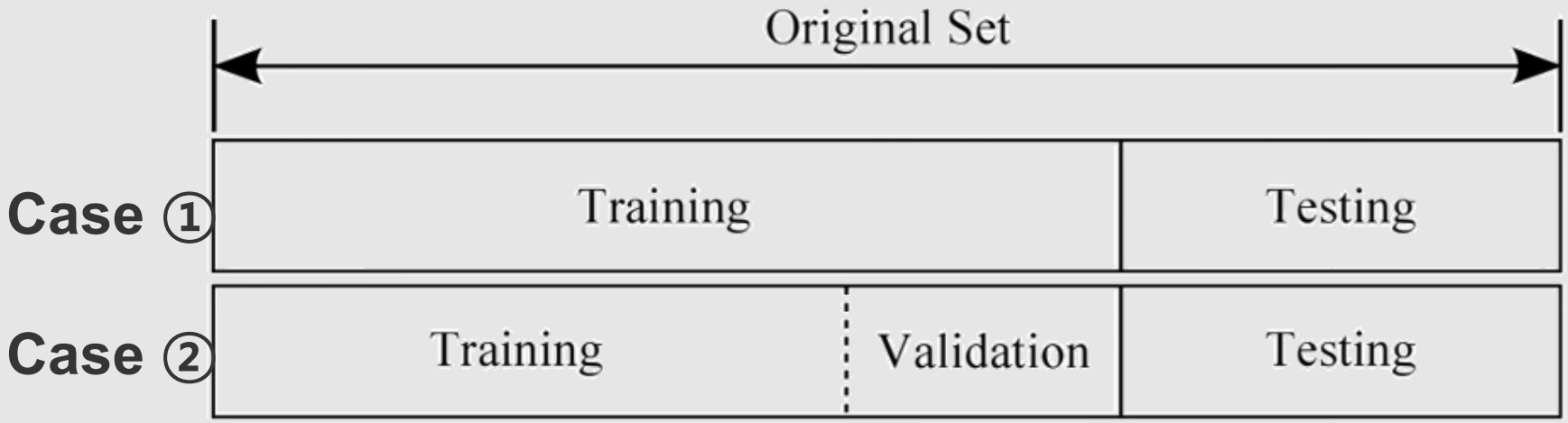
✓ 개요
- 중요 변수들을 선택하고, 적절한 알고리즘을 적용하여 예측 모델을 생성
- 생성된 모델을 평가

✓ 수행되는 내용
- 데이터셋 분리
- 중요 변수 선정
- 머신러닝 알고리즘 적용하여 모델 생성
- 모델 테스트

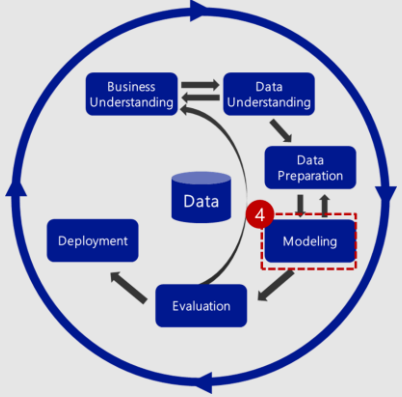
④ Modeling



- ✓ 데이터 셋 분리
  - Case ① : 학습할 때
    - Train Set : 알고리즘을 이용해서 모델을 생성
    - Test Set : 모델 성능 검증
  - Case ② : 실전에서 주로 사용
    - Validation Set : 모델 성능 검증
    - Test Set : 모델 최종 평가



④ Modeling

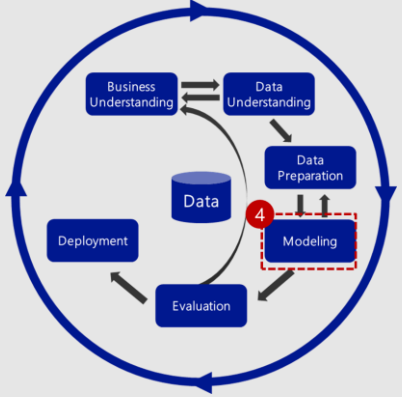


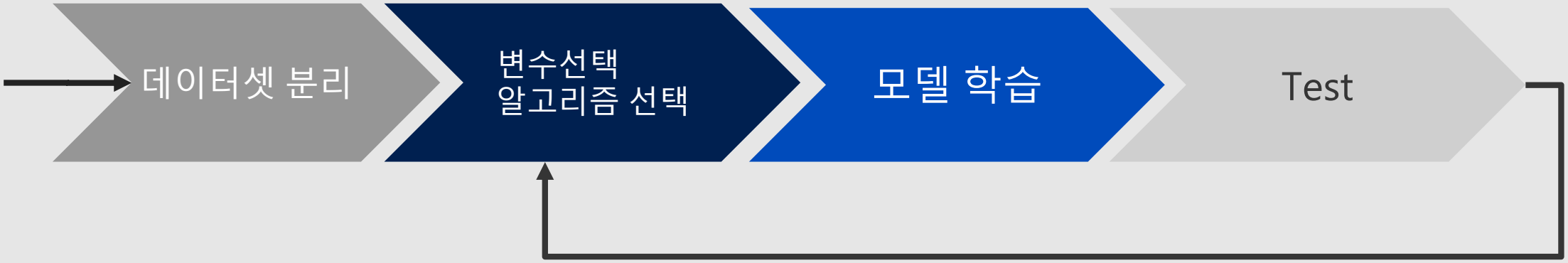
✓머신러닝 알고리즘

Supervised Learning	Unsupervised Learning
지도학습, 감독학습	비지도학습, 비감독학습, 자율학습
Label이 있다.	Label이 없다.
Regression, Logistic Regression , SVM, KNN, Decision Tree , Neural Net, Random Forest 등	Clustering : K-Means, DBSCAN 등

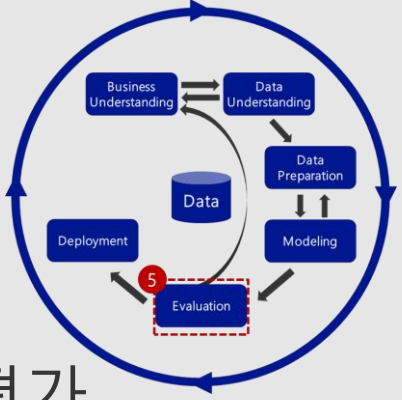
④ Modeling

✓모델 생성





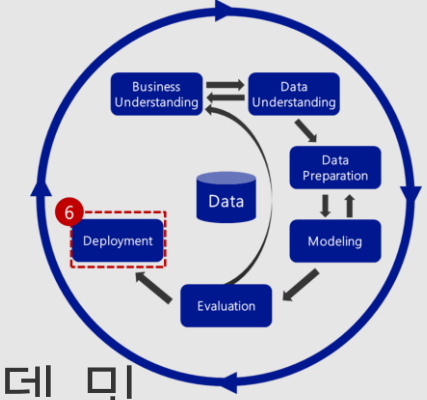
⑤ Evaluation



✓ 개요
- 모델에 대한 데이터 분석 목표와 비즈니스 목표달성에 대한 평가
- 모델과 데이터에서 추출한 패턴이 진정한 규칙성을 갖고 있는지, 단지 특정 예제 데이터에서만 볼 수 있는 특이한 성질은 아닌지 확인
- 비즈니스 목표에 부합되는지 보장

✓ 수행되는 내용
- 모델에 대한 최종평가 : Test Set 이용
- 비즈니스 기대가치 평가

⑥ Deployment



✓ 개요

- 프로젝트 결과물 최종 확정: 프로덕션 환경의 파이프라인, 모델 및 배포가 고객 목표를 충족하는지 확인
- 운영시스템에서 품질(성능 목표) 유지 기준을 정하고, 모니터링 계획을 수립

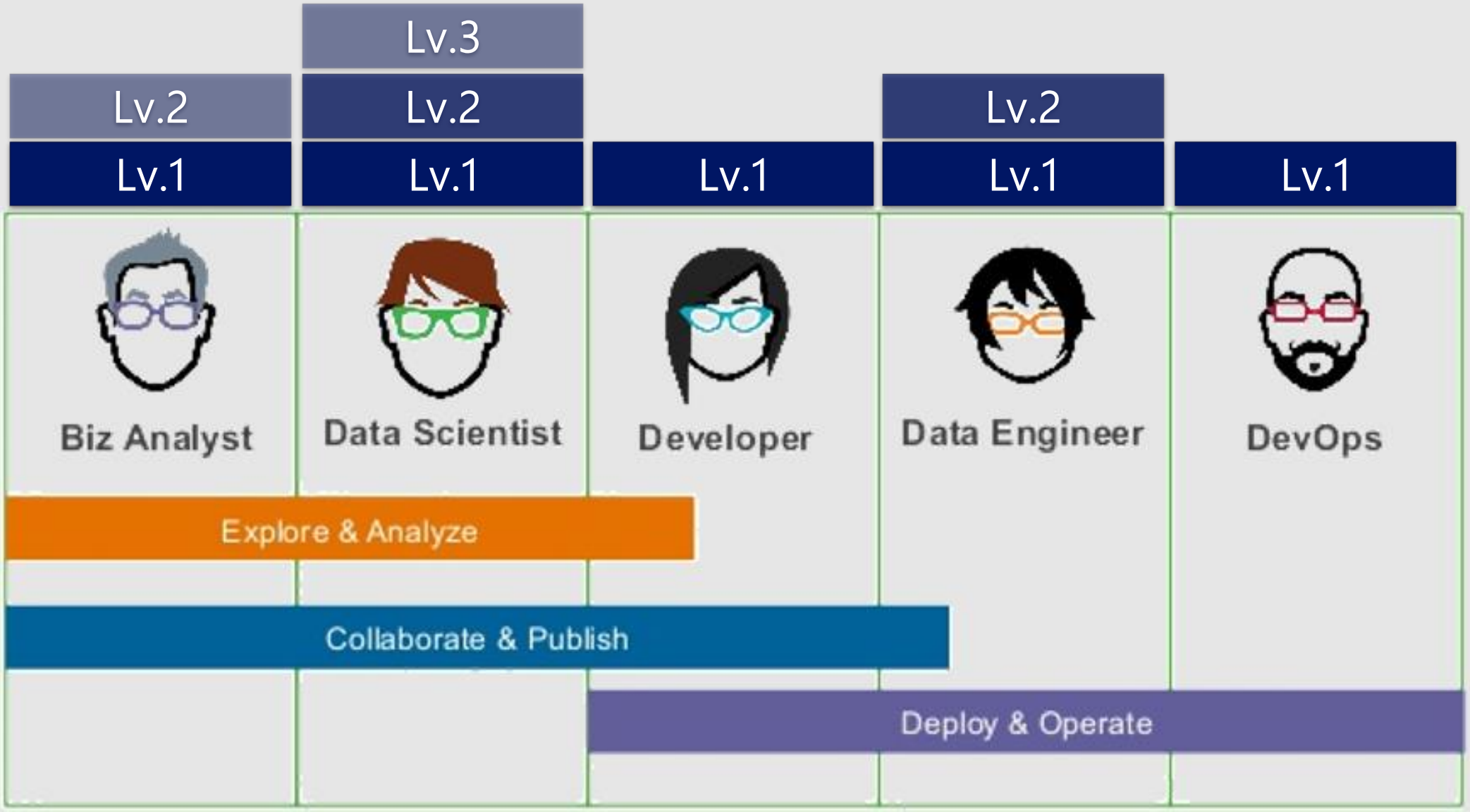
✓ 수행되는 내용

- 시스템 유효성 검사: 배포된 모델과 이 고객 요구 사항을 충족 하는지 확인
- 프로젝트 이전 : 운영환경으로 배포

코드 구조

00 환경준비	10 데이터이해	20 데이터준비	30 모델링
01. import	11. 둘러보기	21. 변수 정리	31. import
02. read_csv	12. 기초통계량	22. NA처리	32. 모델선언
	13. 탐색하기	23. Dummy variable	33. 모델링(학습)
		24. Scaling	34. 예측
		25. Feature Engineering	35. 평가
		26. Data Split	
		27. DataFrame to Numpy	

이 모든 일을 혼자서 다 할 수 없다!



[실습]분류문제 : 엑셀로 분석하기

❖1912년 4월 15일 타이타닉호가 영국 퀸즈타운에서 뉴욕으로 항해하는 중 침몰하였습니다.

❖이 사고로 승객과 승무원 **2,224명 중 1,502명**이 목숨을 잃게 되었습니다.

❖사망자가 이렇게 많이 발생한 이유 중 하나는 **구명보트의 수가 승선한 사람의 수에 비해 부족**했기 때문입니다.

❖부서진 배의 조각을 붙들고 운 좋게 살아난 사람도 있지만, 생존자 대부분은 구명보트를 타고 살아나게 됩니다. **과연 어떤 사람들이 구명보트를 타게 되었을까요?**

[실습]분류문제 : 엑셀로 분석하기

- ✓ 800**명의 데이터를 제공합니다.**
  (“**엑셀로분석_가설수립 및 예측**.xlsx”)
  - Train시트 : 640명, 생존여부 정보 O
  - test 시트 : 160명, 생존여부 정보 X

[분석을 위한 데이터 구조]

데이터 전처리란?

✓데이터를 분석 가능한 형태로 만드는 작업

- 1단계 : **비즈니스 관점**

  ① **결측치**(NA), **이상치 데이터를 처리하고**

  ② **필요한 변수가 충분히 도출된**(Feature Engineering)

  - 가설로 부터 변수 도출하여 생성
  - 기존 변수를 가공하거나, 변수들 간의 조합으로 새로운 변수 생성

  하나의 데이터프레임(테이블) 형태 ➔ 분석 가능한 형태

- 2단계 : **기술적 관점**

  ① dummy variable

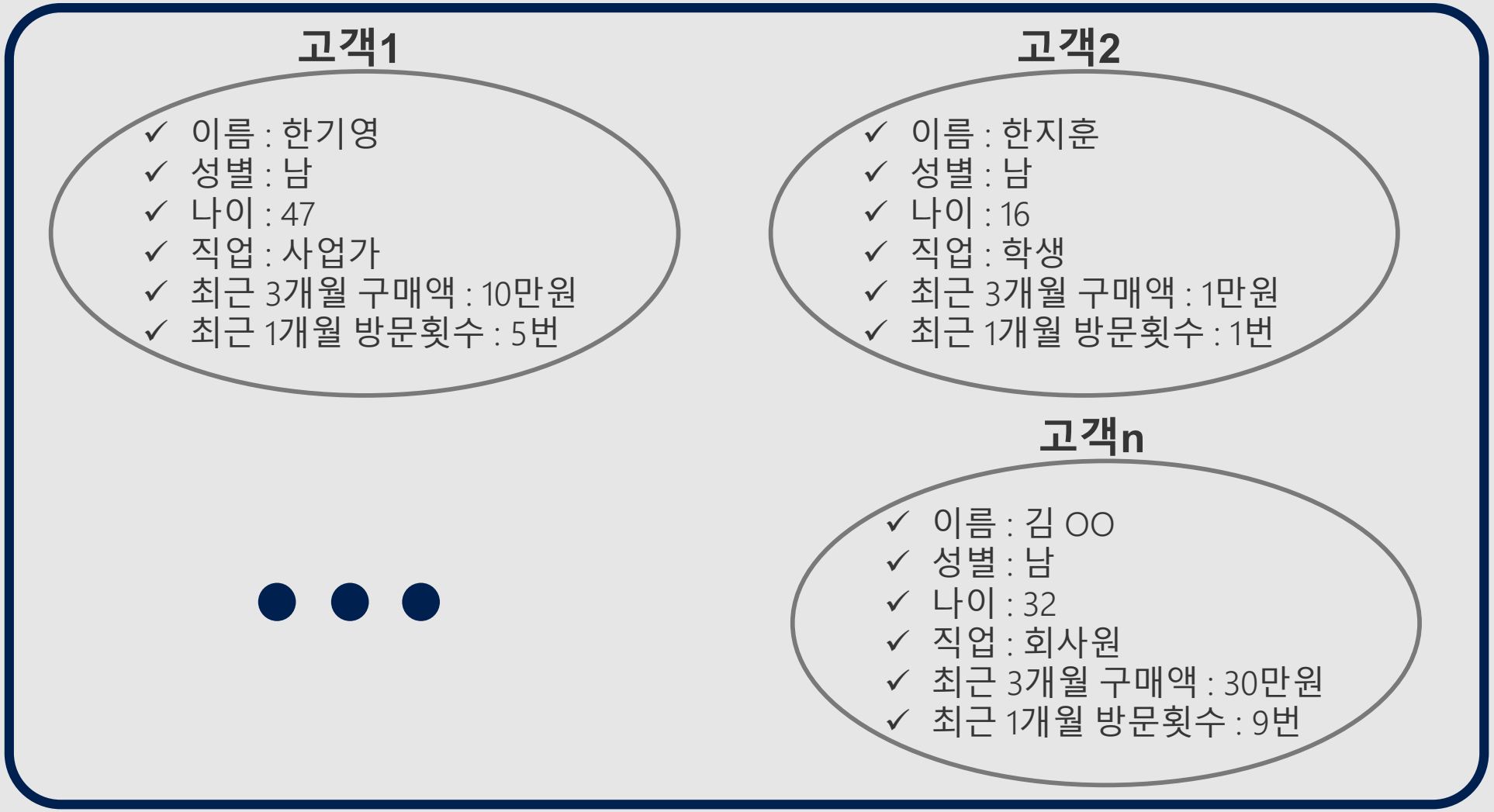
  ② Scaling

  ③ Data Split

  ④ (option) Dataframe to Array(Matrix)

분석 가능한 데이터

✓ 고객 이탈 예측



분석 가능한 데이터 - 데이터프레임

분석 가능한 데이터

✓ Dataframe 사례

- 고객 이탈 데이터프레임

고객ID	성별	나이	최근1개월 구매액	최근 1개월 방문횟수	이탈여부
					0
					1

- 이미지 분류 : MNIST

숫자	p1	p2	p3	p4	p5	...	p781	p782	p783	p784
4	0	0	0	0	34	...	167	250	0	0

[NA(Nan) 다루기]

결측치, 이상치를 어떻게 다룰 것인가?

✓ 데이터 분석 전에 **반드시** 결측치와 이상치를 처리해줘야 한다.

구분	① 제거	② 대체
이상치	▪ 학습 데이터가 많고 ▪ 결측치가 존재하는 변수가 중요하지 않을 때 ▪ 향후 운영환경에서 결측치가 발생되지 않을때	▪ 특정 값(max, min, mode, mean 등) ▪ 시계열 데이터 : 비슷한 시기의 데이터 ▪ 비즈니스 의미에 맞는 값 ▪ 값을 추정(예측)해서 대체 (KNN Imputation)
결측치		

결측치, 이상치를 어떻게 다룰 것인가?

③ 데이터셋을 분리한다. (대체할 방법이 없고, 중요한 변수라면)

고객ID	이름	성별	나이	구매액
1	###	남	10	1000
2	@@@	여	24	2200
3	$$$	여	54	3400
4	%%%	남	38	5000
5	&&&	NA	18	1000
6	***	NA	22	2200

고객ID	이름	성별	나이	구매액
1	###	남	10	1000
2	@@@	여	24	2200
3	$$$	여	54	3400
4	%%%	남	38	5000

모델1

고객ID	이름	나이	구매액
5	&&&	18	1000
6	***	22	2200

모델2

[실습 1 : WHO 국가별 결핵 Case]

데이터 설명

✓ country , iso2 및 iso3 는 국가를 중복해서 지정하는 세 개의 변수

✓ 다른 모든 열(예: new_sp_m014)은 변수가 아니라 값이다.

- 처음 세 글자 : 결핵 사례가, 새로운 사례인지 과거 사례인지를 나타낸다.
- 그 다음 두 글자는 다음의 결핵의 유형을 기술한다.
  - rel :재발
  - ep : 폐 외 (extrapulmonary) 결핵
  - sn : 폐 얼룩으로 보이지 않는 폐결핵 (smear negative)
  - sp : 폐 얼룩으로 보이는 폐결핵(smear positive)
- 여섯 번째 글자는 결핵 환자의 성별을 나타낸다. 남성(m), 여성(f)
- 나머지 숫자는 연령대를 나타낸다.
  - 014 : 0-14세
  - 1524 : 15-24세
  - 2534 : 25-34세
  - 3544 : 35-44세
  - 4554 : 45-54세
  - 5564 : 55-64세
  - 65 : 65세 이상

목표

- ✓ 모든 열은 변수가 되도록 한다.
- ✓ 중복된 열은 제거
- ✓ 모든 행은 관측치가 되도록 한다.
- ✓ 모든 값은 단일값이어야 한다.

추가 변수 만들기

추가변수 개요

- ✓ **중요 값을 기준으로 변수 만들기**
  - 음주 습관에 대한 분석 : age 변수를 이용해서 age >= 20 ➔ 음주가능연령
  - 아파트가격 분석 : 방 수 >=4 & 화장실수 >=2 ➔ Premium
  - 유통 판매분석 : 명절여부, 주요이벤트여부

- ✓ **복수의 변수로부터 도출하기**
  - 일교차 = 일최고기온 – 일최저기온

- ✓ **시계열 데이터의 과거 데이터 계산**
  - 주가 데이터 : 최근 7일 이동평균값

- ✓ Dummy variable
  - 범주형 데이터를 명시적인 숫자로 변형

추가변수 개요

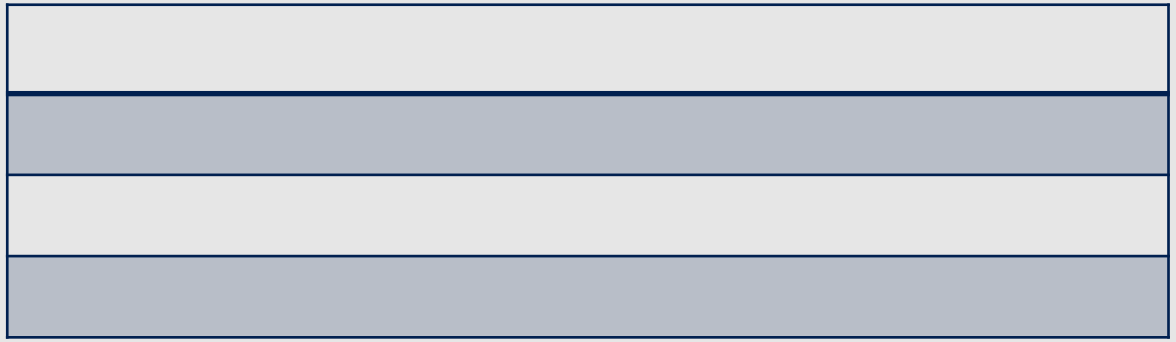
- ✓ 추가변수는 왜 필요한가?
  - 주어진 데이터는 비즈니스의 인사이트를 그대로 담고 있지 않다.
  - 있는 데이터로는 부족하다.
  - 도출된 가설을 데이터 구조로 만들어야 분석할 수 있다.
    - 예 : 신규 가입한 후 10일 이내 7명의 친구를 사귄 사람은, 그렇지 않은 사람보다 잔존율이 더 높을 것이다.

집계를 이용한 변수 추가

✓ 신규 가입한 후 10일 이내 7명의 친구를 사귄 사람은, 그렇지 않은 사람보다 잔존율이 더 높을 것이다.

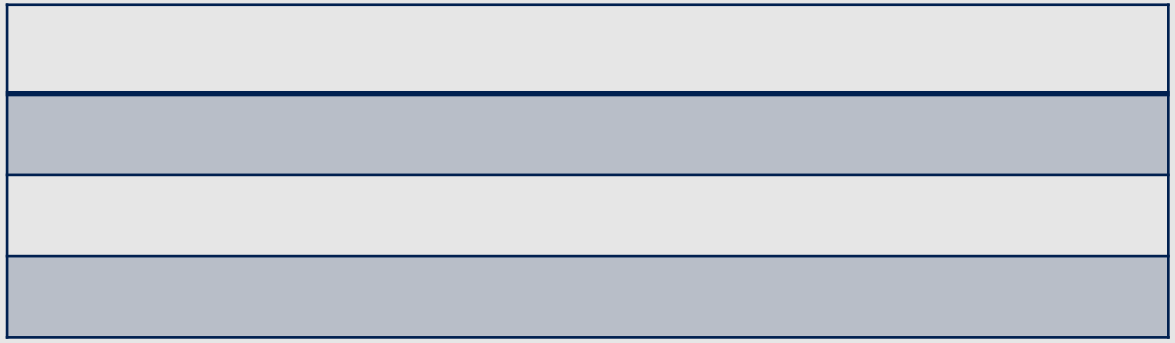
회원ID	성별	가입일
Aaa	F	2020.03.30
Bbb	M	2020.04.12
ccc	F	2020.02.10

회원ID	날짜	새친구 ID
Aaa	2020.04.10	Ddd
Aaa	2020.04.11	Edd
Ccc	2020.02.10	Fff
Bbb	2020.04.13	ccc
Ccc	2020.02.11	Ggg
Ccc	2020.04.13	Bbb



집계를 이용한 변수 추가

- ✓ 신규 가입한 후 10일 이내 7명의 친구를 사귄 사람은, 그렇지 않은 사람보다 잔존율이 더 높을 것이다.
  - 변수를 추가하는 것도 중요하지만,
  - Label을 만드는 것이 더 중요. (이를 Labeling이라 한다) ➔ 잔존 여부 칼럼.



시계열 데이터의 변수 추가

- ✓ 날짜는 분석대상인가?
- ✓ 날짜로부터 추출 가능한 요소(추가변수)
  - Year, Month, Day, Quarter(혹은 Season)
  - Weekday, Week number, Working day
  - 월초, 월말, 연초, 연말
  - 명절 : 구정, 추석, 크리스마스시즌
  - 특별한 기간 : 김장철, Black Friday, 光棍节, 삼삼Day(이때 삼겹살 특별히 잘 팔리지 않음!)

시계열 데이터의 변수 추가

- ✓ 시계열 데이터 다루기
  - Time lag
  - Rolling average
- ✓ 주식 : 오늘 종가에 영향을 미치는 요인은?

날짜	종가						
2020-03-03	?						
2020-03-02	25,000						
2020-02-28	24,500						
2020-02-27	24,700						

모델링을 위한 데이터 준비

[Dummy Variable & Scaling]

Dummy variable

✓ 범주형 입력변수는 **가변수화** 하여 사용해야 한다. (0, 1로)

- 성별

$$\mathrm{m}ale = \begin{cases} 1, & if\ x = male \\ 0, & otherwise \end{cases}$$

- 계절은?

계절
봄
여름
봄
가을
겨울

봄	여름	가을	겨울
1	0	0	0
0	1	0	0
1	0	0	0
0	0	1	0
0	0	0	1

Scaling

- ✓ 값의 범위를 맞춰 주기 위해서 변수를 변환
- ✓ 방법 1 : Normalization
  - 모든 변수의 범위를 0~1로 변환
  - 입력변수 X가 [a, b] 범위라면(a=min, b=max)

$$X_{norm} = \frac{x - a}{b - a}$$

Scaling

✓ 방법2 : Standardization

- 모든 변수의 값을, 평균 = 0, 표준편차 = 1 인 값의 분포로 변환
- 그렇다고 표준 정규분포가 되는 것은 아님!

$$X_z = \frac{x - mean}{std}$$

[실습 2 : 주가예측 전처리]

데이터 설명

- ✓ 뉴욕거래소에 상장된 KT 주가의 일별 데이터입니다.
- ✓ Open ~ Adj Close 값의 단위 : 달러
- ✓ Volume은 거래량을 의미

	Date	Open	High	Low	Close	Adj Close	Volume
2	2014-09-09	17.160000	17.240000	17.070000	17.110001	15.917809	198700
3	2014-09-10	17.080000	17.200001	16.959999	17.200001	16.001539	297900
4	2014-09-11	17.020000	17.070000	16.959999	17.020000	15.834081	166500
5	2014-09-12	17.250000	17.309999	17.219999	17.260000	16.057358	255000
6	2014-09-15	17.240000	17.260000	17.139999	17.200001	16.001539	346200
7	2014-09-16	17.049999	17.160000	16.990000	17.139999	15.945719	332900
8	2014-09-17	17.230000	17.240000	17.020000	17.080000	15.889898	317800
9	2014-09-18	16.830000	16.830000	16.730000	16.740000	15.573590	298000
10	2014-09-19	17.030001	17.100000	17.000000	17.010000	15.824777	271900
11	2014-09-22	16.980000	17.049999	16.930000	16.959999	15.778260	163600

목표

- ✓ 우리는 KT 주가를 예측하는 모델을 만들고자 합니다.
  - 오늘 장이 마감된 이후 내일 주가 예측

- ✓ 이런 모델링을 위한 전처리를 수행하시오
  - 추가변수를 최소한 10개 이상 만드시오.

[종합실습 : 고객 이탈 예측 전처리]

데이터 설명

✓ Customer, Product, Sales 세 테이블이 주어집니다.

Products

	ProductID	ProductName	Category	SubCategory
0	p1052661	새우깡	간식	과자
1	p1054261	고구마스틱	간식	과자
2	p1097821	짱구	간식	과자
3	p1097831	감자칩	간식	과자

Customers

	CustomerID	RegisterDate	Address	Gender	BirthYear	Addr1	Addr2
0	c328222	2014-09-25	강원 원주시 늘품로	F	1960	강원도	원주시
1	c281448	2013-06-18	강원 원주시 치악로	F	1974	강원도	원주시
2	c038336	2003-10-10	강원 춘천시 서부대성로	F	1968	강원도	춘천시
3	c084237	2007-03-09	강원도 강릉시 연곡면 황어대길	F	1982	강원도	강릉시
4	c162600	2010-06-14	강원도 속초시 농공단지길	F	1978	강원도	속초시

Sales

	OrderID	Seq	OrderDate	ProductID	Qty	Amt	CustomerID
0	107	2	2016-01-02	p1036481	2	2100	c150417
1	69	1	2016-01-02	p1152861	1	1091	c212716
2	69	7	2016-01-02	p1013161	1	2600	c212716
3	69	8	2016-01-02	p1005771	1	1650	c212716

목표

- ✓ Labeling : 고객의 이탈여부에 해당하는 Target변수를 생성 하시오.
- ✓ 고객 이탈여부에 영향을 줄 요인을 도출하시오.(최소 10개)
- ✓ 요인들을 Feature로 추가하여 데이터셋을 완성하시오.