

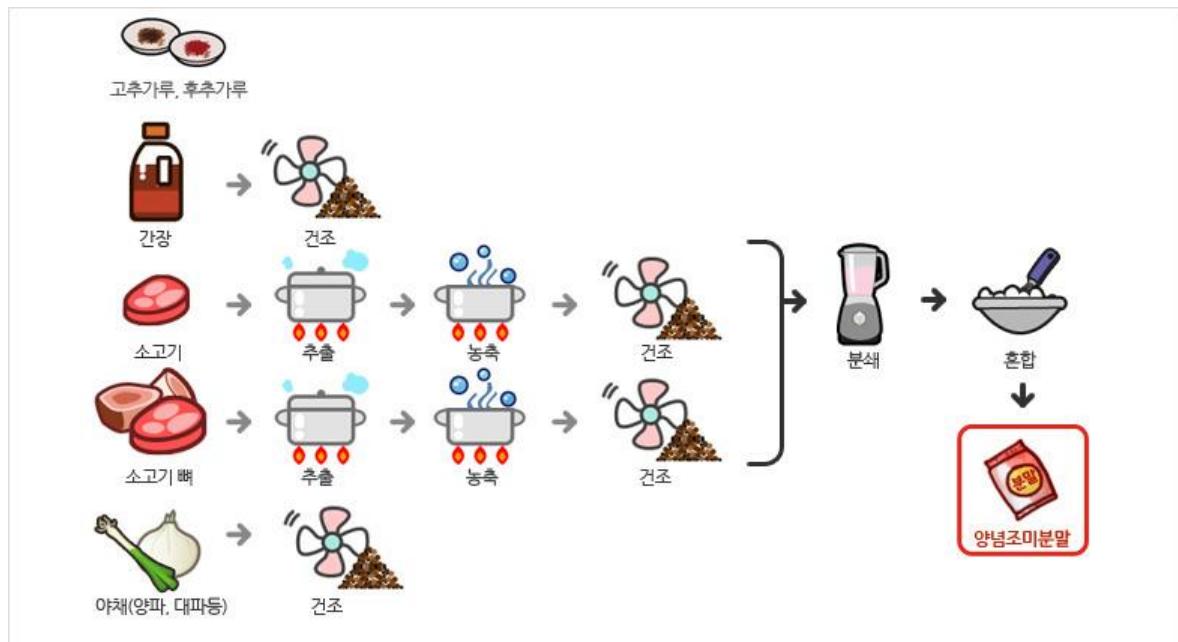


# 데이터의 품질 (Data Quality)

Hyerim Bae

Department of Industrial Engineering, Pusan National University

hrbae@pusan.ac.kr



[http://www.nongshim.com/ramyun/show\\_knowledge?groupCode=004&groupId=7](http://www.nongshim.com/ramyun/show_knowledge?groupCode=004&groupId=7)

# Contents

---

01

Data Quality

02

Data reduction

03

Data imputation

# 데이터 품질 향상 방법론

# 데이터 분석 절차

## • 절차를 정의하는 방식

### – SEMMA

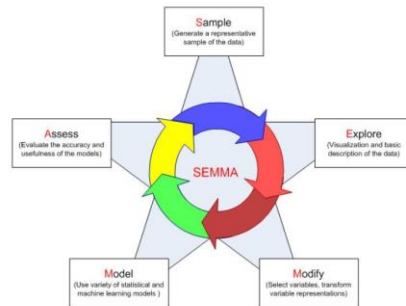
- Sampling(샘플링), Explore(탐색), Modify(수정)
- Modeling(모델링), Assessment(평가)

### – CRISP-DM:

- Business-understanding, Data-understanding, Data-preparation
- Modeling, evaluation, deployment

### – KDD

- Selection(추출), Preprocessing(전처리), Transformation(변환)
- Data mining, Interpretation(Evaluation)(해석 및 평가)



01

**데이터 수집**  
다양한 소스로부터  
데이터를 확인하여  
데이터를 수집

02

**데이터 처리**  
분석이 가능한 형태  
의 데이터로 변환, 결  
측데이터 복원, 데이  
터 품질 향상

03

**데이터 분석**  
데이터 분석 알고리  
즘을 적용하여 데이  
터로부터 원하는 결  
과를 얻어냄

04

**분석결과 활용**  
데이터 분석의 결과  
를 해당 비즈니스의  
가치로 전환하기 위  
한 활동을 수행

Garbage In Garbage Out ! = 좋지 않은 데이터가 입력되면 결과도 좋지 않다!



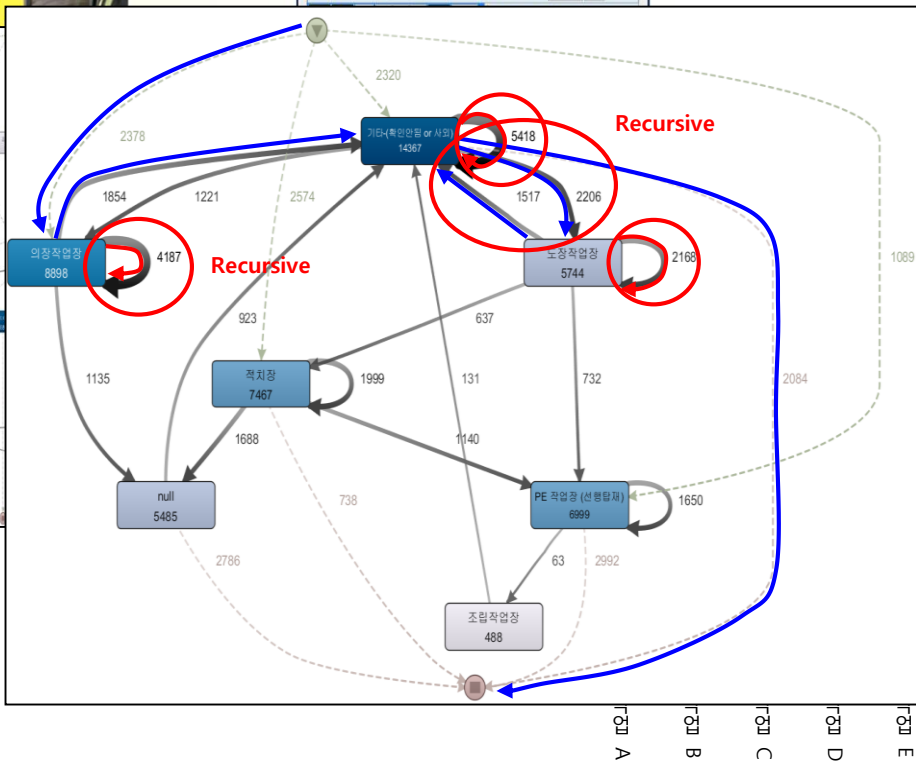
# Data imperfection(분석에 적합하지 않은 데이터)

- Missing data
  - 자료 누락
- Incorrect data
  - 잘못된 코드
- Imprecise data
  - 잘못된 측정 데이터
- Irrelevant data
  - 예측에 사용하기 힘든 자료

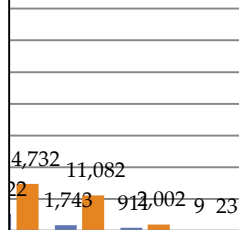
Table 1. Manifestation of quality issues in event log entities [6].

		Event log entities									
					Case		Activity			Event	
		Case	Event	Relationship	attrs.	Position	name	Timestamp	Resource	attrs.	
Event log quality issues	Missing data	I1	I2	I3	I4	I5	I6	I7	I8	I9	
	Incorrect data	I10	I11	I12	I13	I14	I15	I16	I17	I18	
	Imprecise data			I19	I20	I21	I22	I23	I24	I25	
	Irrelevant data	I26	I27								

# Data Quality의 중요성



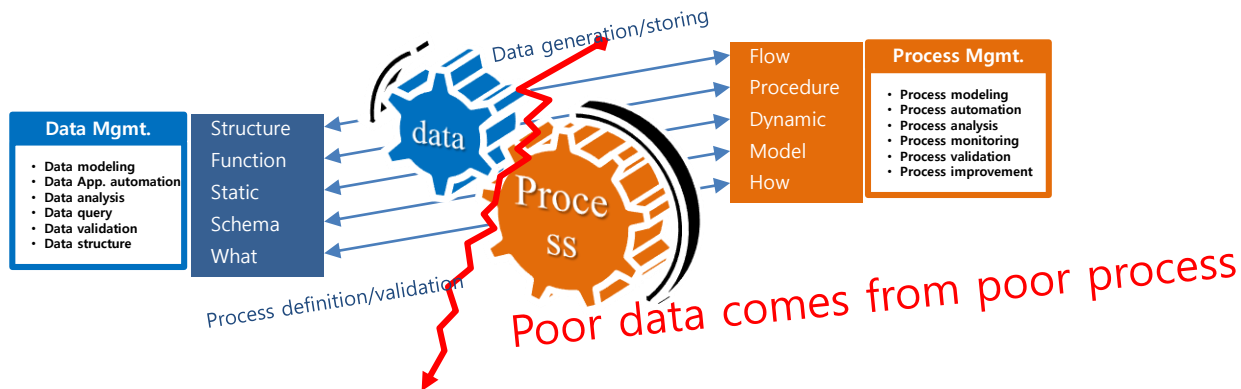
작업명 B



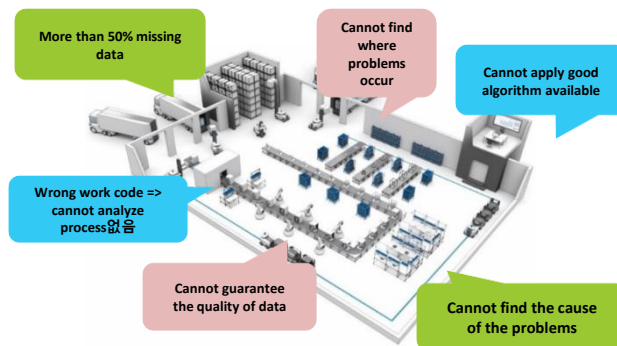
작업명 A  
작업명 B  
작업명 C  
작업명 D  
작업명 E  
작업명 F  
작업명 G  
작업명 H  
작업명 I  
작업명 J  
작업명 K  
작업명 L  
작업명 M  
작업명 N  
작업명 O  
작업명 P  
작업명 Q  
작업명 R  
작업명 S  
작업명 T  
작업명 U  
작업명 V  
작업명 W  
작업명 X  
작업명 Y  
작업명 Z



# Big-data vs. Big-process

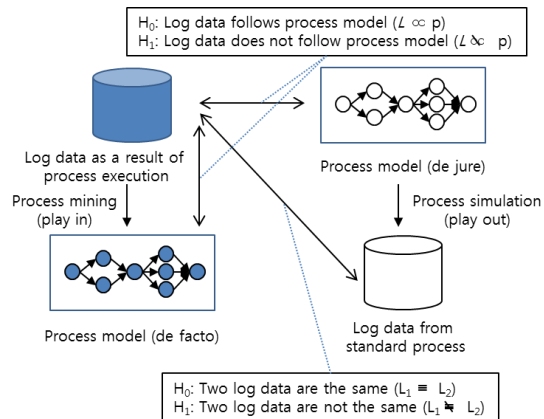
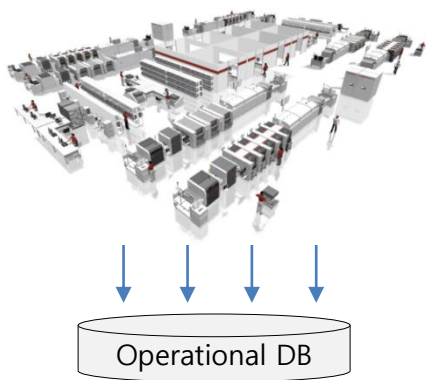


Data-Process Separation	
Undefined process	Missing data
Error in process definition	Data error
Unstandardized process	Low data quality
Process that are not digitalized	Low data value



# Data-Process compliance

- 만약 data-process separation을 없앨 수 있다면?



# 데이터 품질 요소

## 1. 일관성(Consistency) – 논리적 관계의 일관성이 있는지

- two similar IDs for two different employees (다른 두 직원에 대한 유사한 ID)
- a non-existent entry in another table (다른 테이블에 존재하지 않는 항목)

## 2. 정확성(Accuracy) – 어떠한 것의 실제 상태를 잘 나타내는지

- 이러한 데이터 기반의 계산은 실제 결과를 보여줌

## 3. 완전성(Completeness) – 필요한 모든 요소를 나타내는지

- 센서 데이터는 존재하나, 정확한 센서 위치에 대한 정보가 없다면 해당 데이터는 완전성이 떨어짐

## 4. 감사가능성(Auditability) – 유지 및 제어가 가능한지

- 데이터의 품질 감사는 더 좋은 품질의 데이터를 보장하는데 도움이 됨

## 5. 구조적 질서(Orderliness) – 구조적으로 잘 정리되어 있는지

- 오븐의 온도는 화씨로 측정되어야 하며, 음수 값을 가질 수 없음





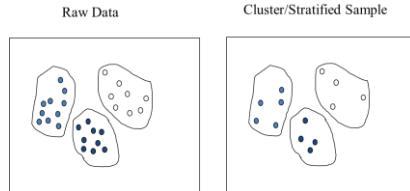
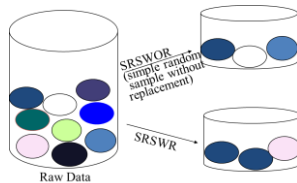
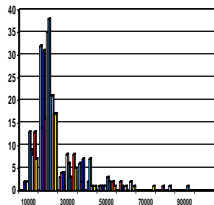
# 데이터 줄이기 (Data Reduction)

# 데이터 줄이기(Data reduction) 전략

- **Data reduction:** 기존 데이터 셋보다 훨씬 작은 크기의 데이터 셋을 활용하더라도 거의 동일한 분석 결과를 얻어낼 수 있을 때, Data reduction이 잘 수행되었다고 볼 수 있음
- 왜 데이터를 줄이는가?
  - 데이터베이스(또는 데이터웨어하우스)에는 수많은 양의 데이터가 존재함
  - 복잡하고 많은 수의 데이터를 분석하기 위해서는 많은 시간이 걸릴 수 있음
- 데이터 줄이기 전략
  - 차원 축소(Dimensionality reduction) eg. 중요하지 않은 성질(변수)은 제거
    - Wavelet transforms
    - Principal Components Analysis (PCA, 주성분분석)
    - Feature subset selection(변수 선택), feature creation(변수 생성)
  - 단순 데이터 수 줄이기
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - 데이터 압축

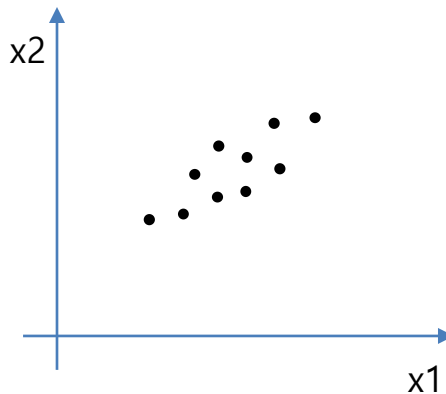
# 단순 데이터 수 줄이기(차원 축소의 개념과 다름)

- 기존 데이터 보다 더 좋은 표현방식을 선택하여 데이터의 크기를 줄임
- 매개변수를 활용한 방법(**Parametric methods**)
  - 데이터가 특정 모델에 적합하다고 가정하고, 모델에 맞는 매개변수를 추정함.
    - 추정한 모델의 매개변수만을 저장하여 데이터 대신 사용(이상치 제외)
  - Ex. 선형회귀 모델을 사용하여 데이터를 수식 형태로 표현 ( $y = ax+b$ )
- 매개변수를 활용하지 않은 방법(**Non-parametric methods**)
  - 모델을 가정하지 않음
  - Ex. 히스토그램, 클러스터링, 샘플링 등



# 차원 축소(Dimensional reduction)

- 차원 축소의 목적
  - 시각화(Visualization)
  - 노이즈 감소(Reduce noise)
  - 유용한 정보만 보존하기 위함
  - 데이터 분석에 필요한 시간적 / 공간적 복잡성 감소



# 주성분 분석(Principal Components Analysis, PCA)

**Goal:** 연속형 변수 집합의 크기를 줄임

상관관계가 높은 측정치들이 존재할 때 유리 => 변수 축소의 여지가 많다.

**The idea:** 연속형 변수 간에 존재하는 중복된 정보를 제거

=> “정보”라는 것은 변수의 분산 합으로 측정됨

**Final product:** 대부분의 정보(분산)를 표현할 수 있는 일부 변수를 선정

비교

연속형 변수가 아닌, 범주형 변수인 경우 => 대응분석(Correspondence analysis)



# 주성분 분석

## 주성분 분석의 수행 방식

- 기존 변수의 (가중된) 선형조합으로 이루어진 새로운 변수를 생성

Ex)  $x' = 0.5x_1 + 0.3x_2 + 0.1x_3$

- 기존 변수의 선형조합으로 만들어진 새로운 변수들은 서로 상관관계 (정보의 중복) 가 없으며, 그 중 일부에만 대부분의 원래 정보가 포함되어 있음
- 새롭게 만들어진 변수를 주성분(Principal components)이라 함

# 예제 – 시리얼(Cereals) 데이터 설명

Name: 시리얼의 이름

mfr: 제조사

type: 시리얼 타입(cold or hot)

calories: 칼로리

protein: 단백질 함유량(g)

fat: 지방 함유량(g)

sodium: 나트륨 함유량(mg)

fiber: 섬유질 함유량(g)

carbo: 복합탄수화물(g)

sugars: 당 함유량(g)

potass: 칼륨 함유량(g)

vitamins: 비타민 함유량

shelf: 진열대 높이

weight: 무게(oz)

cups: 1회 제공량

rating: 소비자 평가 점수

Name: name of cereal

mfr: manufacturer

type: cold or hot

calories: calories per serving

protein: grams

fat: grams

sodium: mg.

fiber: grams

Carbo(복합탄수화물): grams complex carbohydrates

sugars: grams

Potass(칼륨): mg.

vitamins: % FDA rec

shelf: display shelf (진열대 높이)

weight: oz. 1 serving

cups: in one serving

rating: consumer reports

name	mfr	type	calories	protein	...	rating
100%_Bran	N	C	70	4 ...		68
100%_Natural_Bran	Q	C	120	3 ...		34
All-Bran	K	C	70	4 ...		59
All-Bran_with_Extra_Fiber	K	C	50	4 ...		94
Almond_Delight	R	C	110	2 ...		34
Apple_Cinnamon_Cheerios	G	C	110	2 ...		30
Apple_Jacks	K	C	110	2 ...		33
Basic_4	G	C	130	3 ...		37
Bran_Chex	R	C	90	2 ...		49
Bran_Flakes	P	C	90	3 ...		53
Cap'n_Crunch	Q	C	120	1 ...		18
Cheerios	G	C	110	6 ...		51
Cinnamon_Toast_Crunch	G	C	120	1 ...		20

# 공분산과 상관계수(Covariance and correlation)

- 공분산

$$Cov(X, Y) = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

where

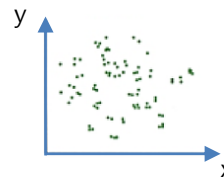
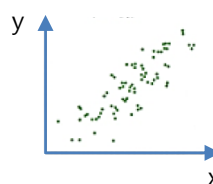
$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \text{ and } \bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$

are the means of  $X, Y$

- 상관 계수

$$\frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / (n-1)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)}}$$

$$Cor(\text{calories}, \text{ratings}) = \frac{-188.68}{\sqrt{379.63}\sqrt{197.2}} = -0.69$$



– 분산의 69%는 두 변수 간에 공유 됨

# 공분산 행렬

- 공분산 행렬(Covariance matrix)

$$C = \begin{pmatrix} \text{cov}(x,x) & \text{cov}(x,y) \\ \text{cov}(x,y) & \text{cov}(y,y) \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum (x_i - m_x)^2 & \frac{1}{n} \sum (x_i - m_x)(y_i - m_y) \\ \frac{1}{n} \sum (x_i - m_x)(y_i - m_y) & \frac{1}{n} \sum (y_i - m_y)^2 \end{pmatrix}$$

	calories	ratings
calories	379.63	-189.68
ratings	-189.68	197.32

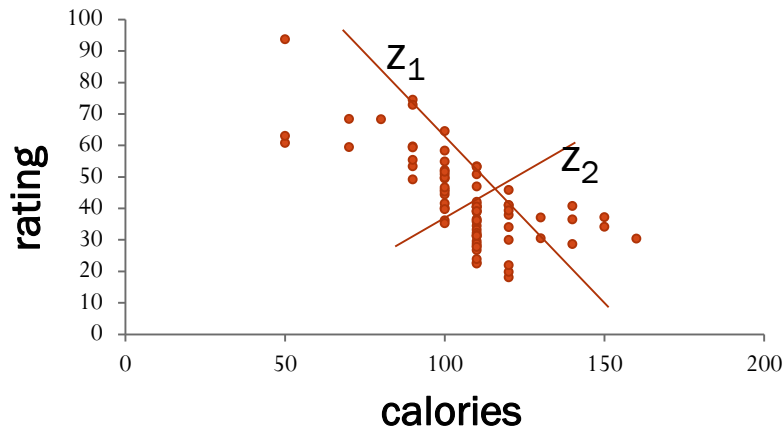
- 총 분산(“정보”)은 개별 변수 분산의 합과 같음: 총 분산 = 379.63 + 197.32
- 칼로리(Calories)는 총 분산(정보)의 66%를 차지함  
 $379.63 / (197.32 + 379.63) = 66\%$
- PCA
  - PCA는 가장 많은 분산을 설명하는 새로운 변수를 찾는 것을 목적으로 함

# 주성분(Principal Components)

$z_1$ 과  $z_2$ 는 두 개의 선형 조합을 나타냄

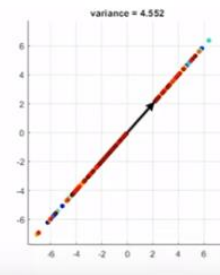
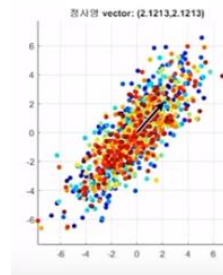
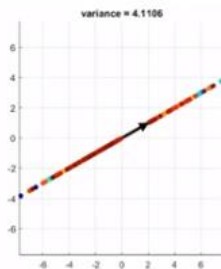
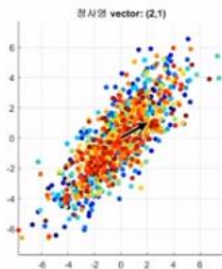
- $z_1$ 은 가장 높은 변동(분산)을 설명함 (값이 퍼진 정도)
- $z_2$ 은 가장 적은 변동(분산)을 설명함

$$z_i = a_{i1}(x_1 - \bar{x}_1) + a_{i2}(x_2 - \bar{x}_2) + \dots + a_{ip}(x_p - \bar{x}_p)$$



# 고유 벡터(Eigen vector)

- PCA는 변수간의 선형조합을 찾기 위함
  - 가장 많은 변동을 설명할 수 있어야 함
  - 고유 벡터(Eigen vector)
    - 선형 변환을 하더라도 동일한 방향을 갖는 0이 아닌 벡터
- PCA
  - 공분산 행렬의 고유 벡터에 데이터를 정사영 내린 것



# 주성분 분석의 일반화

$X_1, X_2, X_3, \dots, X_p$  : 기존에 존재하는  $p$ 개의 변수

$Z_1, Z_2, Z_3, \dots, Z_p$  : 기존  $p$ 개 변수의 가중 평균으로 구성된 새로운  $p$ 개의 변수

모든  $Z$  변수 쌍 간의 상관관계는 0 임

$Z$ 값이 갖는 분산 크기 순으로 정렬( $Z_1$ :분산이 가장 큼,  $Z_p$ : 분산이 가장 작음)

일반적으로, 처음 몇 개의  $Z$  변수에 대부분의 분산(정보)가 포함되어 있으므로 나머지  $Z$ 는 사용하지 않아도 됨

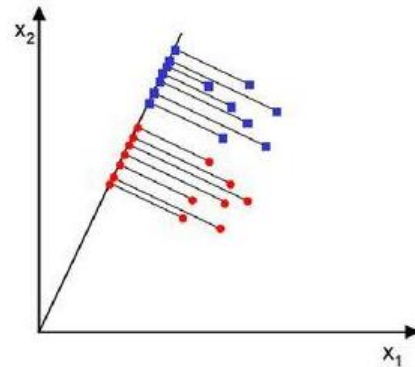
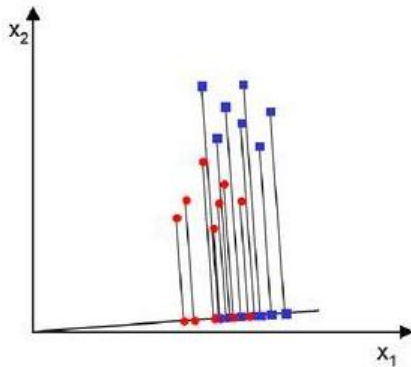
# 예제 – 데이터에 PCA 적용

- 초기 6개의 주성분 표현
- 2개의 주성분만으로 93%의 변동(정보)을 설명할 수 있음
- Note: data differ slightly from text(어떤 의미인지)

Variable	1	2	3	4	5	6
calories	0.07624155	-0.01066097	0.61074823	-0.61706442	0.45754826	0.12601775
protein	-0.00146212	0.00873588	0.00050506	0.0019389	0.05533375	0.10379469
fat	-0.00013779	0.00271266	0.01596125	-0.02595884	-0.01839438	-0.12500292
sodium	0.98165619	0.12513085	-0.14073193	-0.00293341	0.01588042	0.02245871
fiber	-0.00479783	0.03077993	-0.01684542	0.02143376	0.00872434	0.271184
carbo	0.01486445	-0.01731863	0.01272501	0.02175146	0.35580006	-0.56089228
sugars	0.00398314	-0.00013545	0.09870714	-0.11555841	-0.29906386	0.62323487
potass	-0.119053	0.98861349	0.03619435	-0.042696	-0.04644227	-0.05091622
vitamins	0.10149482	0.01598651	0.7074821	0.69835609	-0.02556211	0.01341988
shelf	-0.00093911	0.00443601	0.01267395	0.00574066	-0.00823057	-0.05412053
weight	0.0005016	0.00098829	0.00369807	-0.0026621	0.00318591	0.00817035
cups	0.00047302	-0.00160279	0.00060208	0.00095916	0.00280366	-0.01087413
rating	-0.07615706	0.07254035	-0.30776858	0.33866307	0.75365263	0.41805118
Variance	7204.161133	4833.050293	498.4260864	357.2174377	72.47863007	4.33980322
Variance%	55.52834702	37.25226212	3.84177661	2.75336623	0.55865192	0.0334504
Cum%	55.52834702	92.78060913	96.62238312	99.37575531	99.93440247	99.96785736

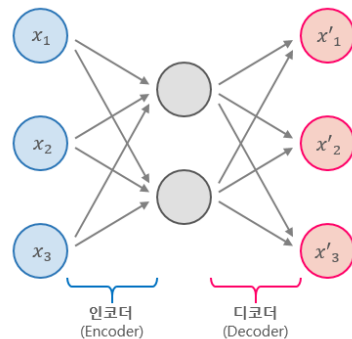


# LDA를 활용한 데이터 축소



# Neural Network를 활용한 데이터 축소: Autoencoder

- Autoencoder(오토인코더)의 목적
  - 데이터 축소



# 정리

---

- 데이터 축소(요약)은 데이터 탐색에 매우 중요함
- 데이터 축소는 데이터를 나타내는 수치(평균, 중앙값 등) 및 시각적 자료(히스토그램 등)이 포함됨
- 데이터 축소는 기존 데이터의 정보를 더 작은 하위 집합으로 압축하는 데 유용함
  - 범주형 변수의 경우, 유사한 범주를 결합함으로써 축소 가능
  - 주성분 분석은 기존의 연속형(수치형) 데이터 셋의 대부분 정보를 표현할 수 있는 더 적은 수의 변수로 데이터 셋을 변환함

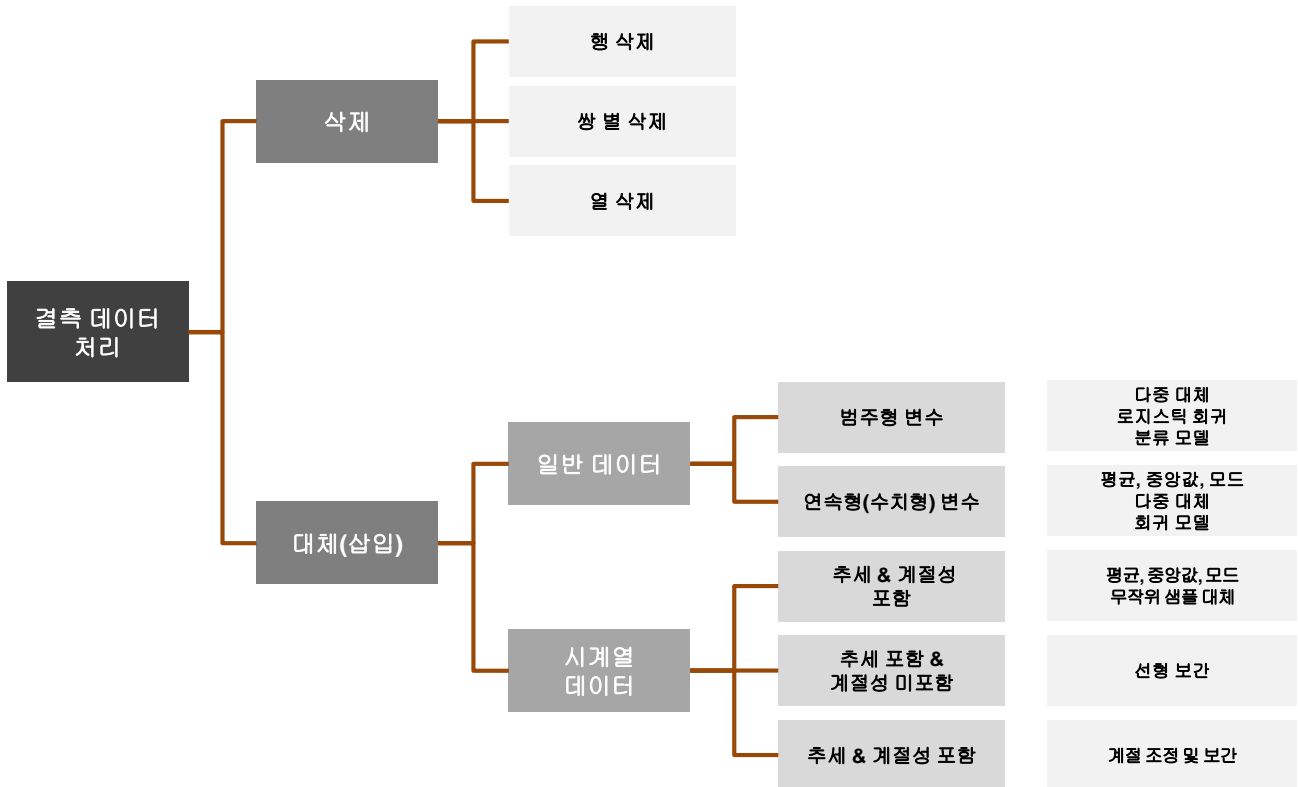


# 데이터 결측치 대체 (Data Imputation)

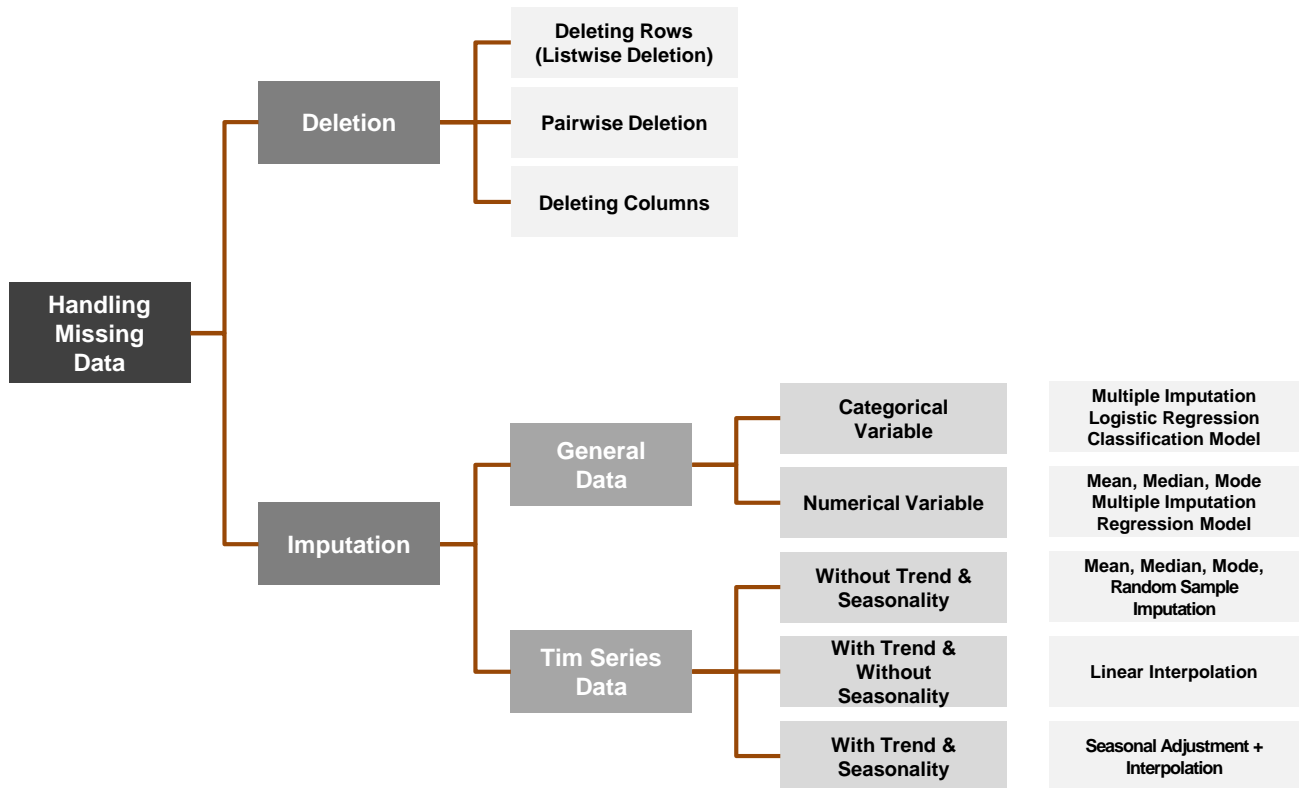
# 결측치에 대한 이해

- 데이터 정제 및 탐색 분석 시 가장 일반적인 문제 중 하나는 결측치를 처리하는 것
- 결측치는 크게 세 종류로 분류할 수 있음
  - 무작위 결측(Missing at Random, MAR) : 데이터가 누락되는 경향이 자료내 다른 변수와 관련
  - 완전 무작위 결측(Missing Completely at Random, MCAR) : 결측치 발생이 다른 변수와 전혀 관련이 없음
  - 무작위가 아닌 결측(Missing not at Random, MNAR) Missing여부가 해당 변수의 값에 의해서 결정

# 결측 데이터 처리 방법



# 결측 데이터 처리 방법



# 결측 데이터 처리 방법

- 행 삭제(Listwise deletion)
  - 결측 데이터를 처리하는 가장 일반적인 방법은 결측 데이터를 생략하고 나머지 데이터를 분석하는 것
  - 이러한 접근 방식을 complete case (or available case) analysis or list-wise deletion이라 함

	Mobile Package	Download Speed	Data Limit Usage
y <sub>1</sub>	Fast	157	80%
y <sub>2</sub>	Lite	99	70%
y <sub>3</sub>	Fast	167	10%
y <sub>4</sub>	Fast	NA	80%
y <sub>5</sub>	Lite	76	70%
y <sub>6</sub>	Fast	155	10%
y <sub>7</sub>	NA	NA	95%
y <sub>8</sub>	Lite	76	77%
y <sub>9</sub>	Fast	180	NA



	Mobile Package	Download Speed	Data Limit Usage
y <sub>1</sub>	Fast	157	80%
y <sub>2</sub>	Lite	99	70%
y <sub>3</sub>	Fast	167	10%
y <sub>5</sub>	Lite	76	70%
y <sub>6</sub>	Fast	155	10%
y <sub>8</sub>	Lite	76	77%



# 결측 데이터 처리 방법

- 쌍 별 삭제(available-case analysis)
  - 결측 된 관측치만 무시하고, 존재하는 변수에 대해서는 분석을 수행
  - 데이터 셋의 다른 곳에 누락된 데이터가 존재할 경우, 기존 값 사용
  - 쌍 별 삭제는 관찰된 모든 데이터를 사용하므로 행 삭제보다 더 많은 정보를 보존함

	Mobile Package	Download Speed	Data Limit Usage
y <sub>1</sub>	Fast	157	80%
y <sub>2</sub>	Lite	99	70%
y <sub>3</sub>	Fast	167	10%
y <sub>4</sub>	Fast	NA	80%
y <sub>5</sub>	Lite	76	70%
y <sub>6</sub>	Fast	155	10%
y <sub>7</sub>	NA	NA	95%
y <sub>8</sub>	Lite	76	77%
y <sub>9</sub>	Fast	180	NA



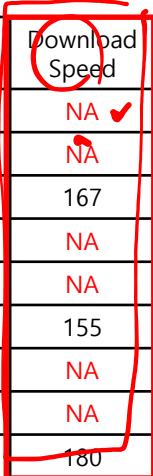
	Mobile Package	Download Speed	Data Limit Usage
y <sub>1</sub>	Fast	157	80%
y <sub>2</sub>	Lite	99	70%
y <sub>3</sub>	Fast	167	10%
y <sub>4</sub>	Fast		80%
y <sub>5</sub>	Lite	76	70%
y <sub>6</sub>	Fast	155	10%
y <sub>7</sub>			95%
y <sub>8</sub>	Lite	76	77%
y <sub>9</sub>	Fast	180	

# 결측 데이터 처리 방법

- 열 삭제(available-variable analysis)

- 특정 변수에 대해 결측 데이터가 너무 많을 경우, 해당 변수 자체를 삭제할 수 있음
- 열 삭제는 최후의 선택 방법이며, 열을 삭제하더라도 모델 성능이 향상되는지 확인해야 함

	Mobile Package	Download Speed	Data Limit Usage
y <sub>1</sub>	Fast	NA ✓	80%
y <sub>2</sub>	Lite	NA	70%
y <sub>3</sub>	Fast	167	10%
y <sub>4</sub>	Fast	NA	80%
y <sub>5</sub>	Lite	NA	70%
y <sub>6</sub>	Fast	155	10%
y <sub>7</sub>	Fast	NA	95%
y <sub>8</sub>	Lite	NA	77%
y <sub>9</sub>	Fast	180	80%



	Mobile Package	Data Limit Usage
y <sub>1</sub>	Fast	80%
y <sub>2</sub>	Lite	70%
y <sub>3</sub>	Fast	10%
y <sub>4</sub>	Fast	80%
y <sub>5</sub>	Lite	70%
y <sub>6</sub>	Fast	10%
y <sub>7</sub>	Fast	95%
y <sub>8</sub>	Lite	77%
y <sub>9</sub>	Fast	80%

# 결측 데이터 처리 방법

- 단순 대체(평균, 중앙값, 모드 활용)
    - 결측 데이터를 통계적 추정치로 대체
      - 평균, 중앙값, 모드 등을 활용하여 대체
- Ex) Mean = 130, Median = 155, Mode = 200

	Mobile Package	Download Speed	Data Limit Usage
y <sub>1</sub>	Fast	157	80%
y <sub>2</sub>	Lite	99	70%
y <sub>3</sub>	Fast	167	10%
y <sub>4</sub>	Fast	NA	80%
y <sub>5</sub>	Lite	76	70%
y <sub>6</sub>	Fast	155	10%
y <sub>7</sub>	Fast	NA	95%
y <sub>8</sub>	Lite	76	77%
y <sub>9</sub>	Fast	180	80%



	SI with Mean	SI with Median	SI with Mode
y <sub>1</sub>	157	157	157
y <sub>2</sub>	99	99	99
y <sub>3</sub>	167	167	167
y <sub>4</sub>	130	155	76
y <sub>5</sub>	76	76	76
y <sub>6</sub>	155	155	155
y <sub>7</sub>	130	155	76
y <sub>8</sub>	76	76	76
y <sub>9</sub>	180	180	180

# 결측 데이터 처리 방법

- 시계열 데이터에 활용가능한 방법론(LOCF, NOCB and Linear Interpolation)
  - Last Observation Carried Forward(LOCF)
    - 시계열 데이터인 경우, 가장 일반적으로 사용되는 방법
    - 결측치 존재 시, 마지막으로 관찰 된 값으로 대체

	Date	Download Speed	Data Limit Usage
y <sub>1</sub>	1-MAR	157	80%
y <sub>2</sub>	2-MAR	99	70%
y <sub>3</sub>	3-MAR	167	10%
y <sub>4</sub>	4-MAR	NA	80%
y <sub>5</sub>	5-MAR	76	70%
y <sub>6</sub>	6-MAR	155	10%
y <sub>7</sub>	7-MAR	NA	95%
y <sub>8</sub>	8-MAR	76	77%
y <sub>9</sub>	9-MAR	NA	80%



	Date	Download Speed	Data Limit Usage
y <sub>1</sub>	1-MAR	157	80%
y <sub>2</sub>	2-MAR	99	70%
y <sub>3</sub>	3-MAR	167	10%
y <sub>4</sub>	4-MAR	167	80%
y <sub>5</sub>	5-MAR	76	70%
y <sub>6</sub>	6-MAR	155	10%
y <sub>7</sub>	7-MAR	155	95%
y <sub>8</sub>	8-MAR	76	77%
y <sub>9</sub>	9-MAR	76	80%

# 결측 데이터 처리 방법

- 시계열 데이터에 활용가능한 방법론
  - Next Observation Carried Backward(NOCB)
    - LOCF와 유사한 방식
    - 결측치 발견 이후 첫번째 관측 값을 활용하여 결측치 대체

	Date	Download Speed	Data Limit Usage
y <sub>1</sub>	1-MAR	157	80%
y <sub>2</sub>	2-MAR	99	70%
y <sub>3</sub>	3-MAR	167	10%
y <sub>4</sub>	4-MAR	NA	80%
y <sub>5</sub>	5-MAR	76	70%
y <sub>6</sub>	6-MAR	NA	10%
y <sub>7</sub>	7-MAR	NA	95%
y <sub>8</sub>	8-MAR	76	77%
y <sub>9</sub>	9-MAR	180	80%



	Date	Download Speed	Data Limit Usage
y <sub>1</sub>	1-MAR	157	80%
y <sub>2</sub>	2-MAR	99	70%
y <sub>3</sub>	3-MAR	167	10%
y <sub>4</sub>	4-MAR	76	80%
y <sub>5</sub>	5-MAR	76	70%
y <sub>6</sub>	6-MAR	76	10%
y <sub>7</sub>	7-MAR	76	95%
y <sub>8</sub>	8-MAR	76	77%
y <sub>9</sub>	9-MAR	180	80%

# 결측 데이터 처리 방법

- 시계열 데이터에 활용가능한 방법론
  - 선형 보간법
    - 함수를 데이터에 맞게 조정하고, 해당 함수를 사용하여 누락된 데이터 대체
    - 가장 간단한 방식의 보간법은 결측 데이터의 앞/뒤 값 사이의 평균을 만드는 선형 보간법

	Date	Download Speed	Data Limit Usage
y <sub>1</sub>	1-MAR	157	80%
y <sub>2</sub>	2-MAR	99	70%
y <sub>3</sub>	3-MAR	167	10%
y <sub>4</sub>	4-MAR	NA	80%
y <sub>5</sub>	5-MAR	76	70%
y <sub>6</sub>	6-MAR	NA	10%
y <sub>7</sub>	7-MAR	150	95%
y <sub>8</sub>	8-MAR	76	77%
y <sub>9</sub>	9-MAR	180	80%



	Date	Download Speed	Data Limit Usage
y <sub>1</sub>	1-MAR	157	80%
y <sub>2</sub>	2-MAR	99	70%
y <sub>3</sub>	3-MAR	167	10%
y <sub>4</sub>	4-MAR	121.5	80%
y <sub>5</sub>	5-MAR	76	70%
y <sub>6</sub>	6-MAR	113	10%
y <sub>7</sub>	7-MAR	150	95%
y <sub>8</sub>	8-MAR	76	77%
y <sub>9</sub>	9-MAR	180	80%

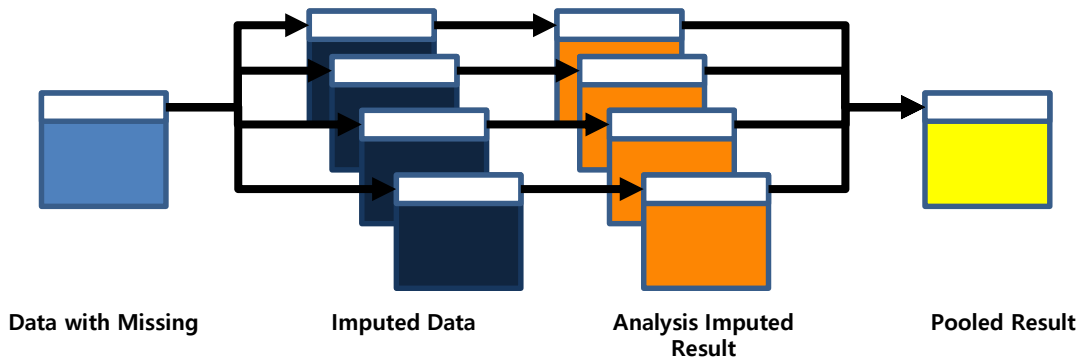
$$(167+76)/2 = 121.5$$

$$(76+150)/2 = 113$$

# 결측 데이터 처리 방법

- 다중 대체

- 다중 대체(MI)는 결측치를 처리하기 위한 통계적 기법
- 다중 대체에는 아래의 3가지 구성요소가 포함 됨
  - 결측치 생성: 관측된 데이터의 분포를 활용하여 결측치에 대한 그럴듯한 값을 추정
  - 대체 및 분석: 결측치는 여러 추정 값으로 대체되며, 각각의 경우에 대해 분석 수행
  - 풀링(Pooling): 추정값을 결합하여 매개변수 추정값 도출



- MI technique has various methods depending on how data is imputed, and there are various imputation method such as MICE (Multiple Imputation by Chain Equation), Random Forest Imputation, KNN Imputation, Expectation-Maximization Imputation.

# 결측 데이터 처리 방법

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M



# 결측 데이터 처리 방법

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Simple  
Imputation  
using  
mean

Age	Income	Gender
33	12.771	F
18	12,000	F
25.5	13,542	M

# 결측 데이터 처리 방법

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Simple  
Imputation  
using mean

Age	Income	Gender
33	12.771	F
18	12,000	F
25.5	13,542	M

Age back  
to NA

Age	Income	Gender
33	12.771	F
18	12,000	F
NA	13,542	M

# 결측 데이터 처리 방법

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Simple  
Imputation  
using mean

Age	Income	Gender
33	12.771	F
18	12,000	F
25.5	13,542	M

Age back  
to NA

Age	Income	Gender
33	12.771	F
18	12,000	F
NA	13,542	M

Regression  
 $\text{Age} \sim \text{Income} + \text{Gender}$

Predict Age

Age	Income	Gender
33	12.771	F
18	12,000	F
NA	13,542	M

# 결측 데이터 처리 방법

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Simple  
Imputation  
using mean

Age	Income	Gender
33	12.771	F
18	12,000	F
25.5	13,542	M

Age back  
to NA

Age	Income	Gender
33	12.771	F
18	12,000	F
NA	13,542	M

Regression  
Age ~ Income + Gender

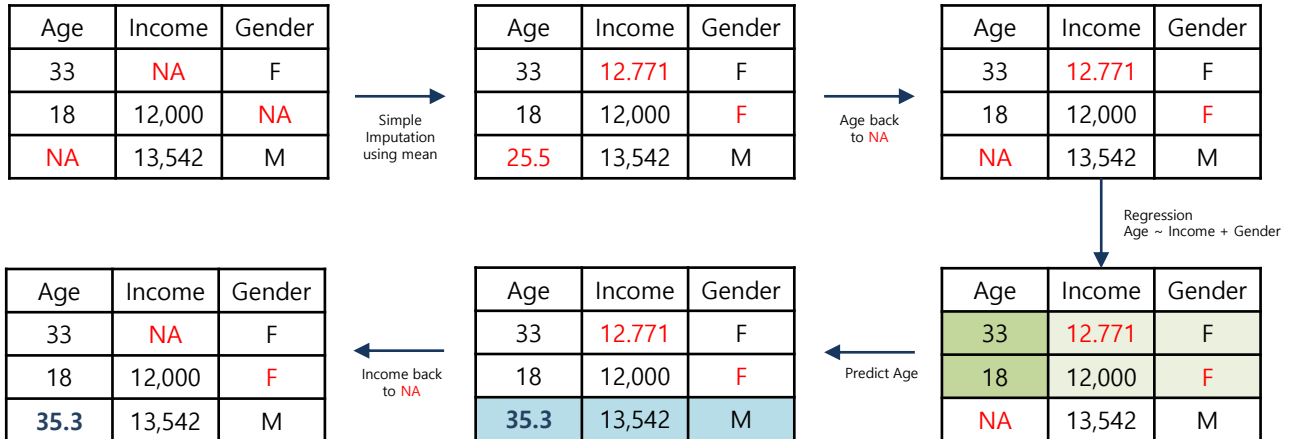
Predict Age

Age	Income	Gender
33	12.771	F
18	12,000	F
35.3	13,542	M

Age	Income	Gender
33	12.771	F
18	12,000	F
NA	13,542	M

# 결측 데이터 처리 방법

Understanding for MICE – Single Iteration



# 결측 데이터 처리 방법

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Simple  
Imputation  
using mean

Age	Income	Gender
33	12,771	F
18	12,000	F
25.5	13,542	M

Age back  
to NA

Age	Income	Gender
33	12,771	F
18	12,000	F
NA	13,542	M

Regression  
Age ~ Income + Gender

Age	Income	Gender
33	12,771	F
18	12,000	F
NA	13,542	M

Predict Age

Age	Income	Gender
33	12,771	F
18	12,000	F
35.3	13,542	M

Income back  
to NA

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	M

Regression  
Income ~ Age + Gender

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	M

# 결측 데이터 처리 방법

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Simple  
Imputation  
using mean

Age	Income	Gender
33	12,771	F
18	12,000	F
25.5	13,542	M

Age back  
to NA

Age	Income	Gender
33	12,771	F
18	12,000	F
NA	13,542	M

Regression  
Age ~ Income + Gender

Age	Income	Gender
33	12,771	F
18	12,000	F
NA	13,542	M

Predict Age

Age	Income	Gender
33	12,771	F
18	12,000	F
35.3	13,542	M

Income back  
to NA

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	M

Regression  
Income ~ Age + Gender

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	M

Predict  
Income

Age	Income	Gender
33	13,103	F
18	12,000	F
35.3	13,542	M

# 결측 데이터 처리 방법

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Simple Imputation using mean

Age	Income	Gender
33	12,771	F
18	12,000	F
25.5	13,542	M

Age back to NA

Age	Income	Gender
33	12,771	F
18	12,000	F
NA	13,542	M

Regression  
Age ~ Income + Gender

Age	Income	Gender
33	12,771	F
18	12,000	F
NA	13,542	M

Predict Age

Age	Income	Gender
33	12,771	F
18	12,000	F
35.3	13,542	M

Income back to NA

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	M

Regression  
Income ~ Age + Gender

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	M

Predict Income

Age	Income	Gender
33	13,103	F
18	12,000	F
35.3	13,542	M

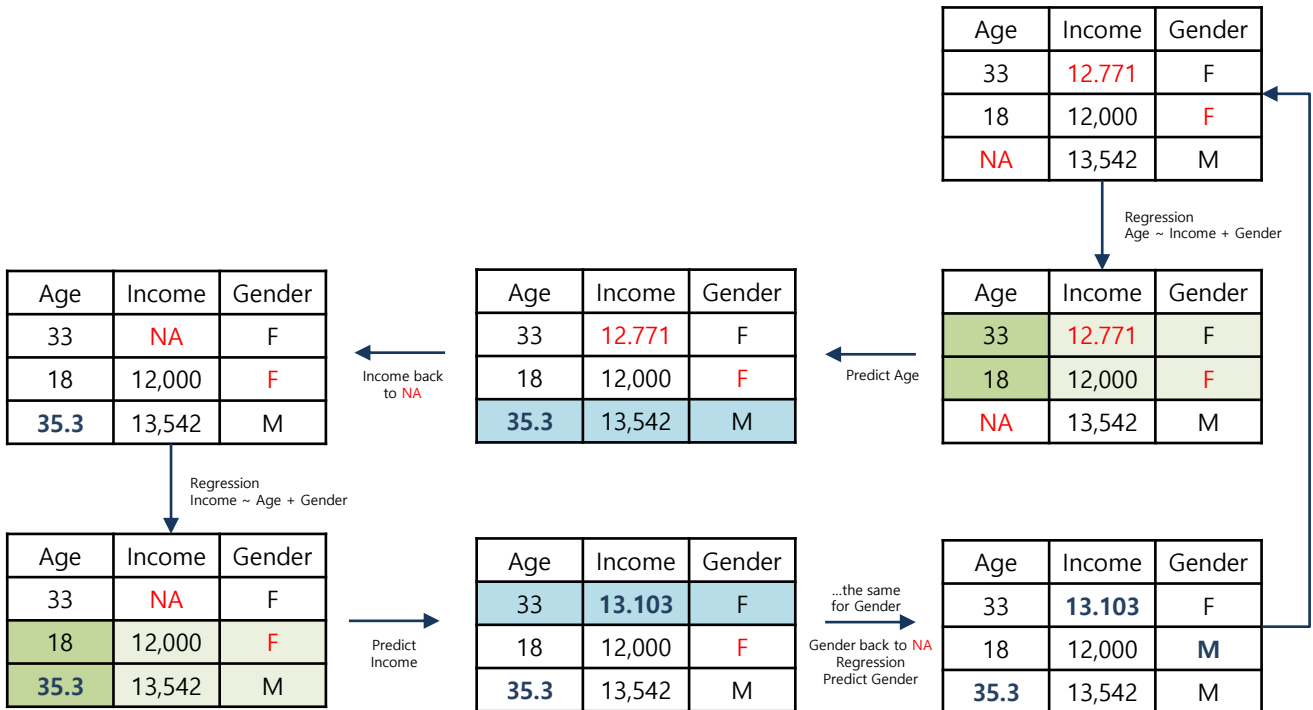
...the same for Gender  
Gender back to NA  
Regression  
Predict Gender

Age	Income	Gender
33	13,103	F
18	12,000	M
35.3	13,542	M



# 결측 데이터 처리 방법

Understanding for MICE – Single Iteration



# 특별한 유형의 데이터를 처리하기 위한 다중 대체 방법

- 이벤트 로그 복구를 위한 Likelihood 기반 다중 대체

	Case	Activity	Resource	Part Desc.		Start Time	End Time
$e_1$	Case <sub>1</sub>	Turning & Milling	Machine 4	Cable Head		2012-01-29 23:24	2012-01-30 05:43
$e_2$	Case <sub>1</sub>	Turning & Milling	Machine 4	Cable Head		2012-01-30 05:44	2012-01-30 06:42
$e_3$	Case <sub>1</sub>	Turning & Milling	Machine 4	Cable Head		2012-01-30 06:59	2012-01-30 07:21
$e_4$	Case <sub>1</sub>	Turning & Milling	Machine 4	Cable Head		2012-01-30 07:21	2012-01-30 10:58
$e_5$	Case <sub>1</sub>	Turning & Milling Q.C	Quality Check 1	Cable Head		2012-01-31 13:20	2012-01-31 14:50
$e_6$	Case <sub>1</sub>	Laser Marking	Machine 7	Cable Head		2012-02-01 08:18	2012-02-01 08:27
$e_7$	Case <sub>1</sub>	Lapping	Machine 1	Cable Head		2012-02-14 00:00	2012-02-14 01:15
$e_8$	Case <sub>1</sub>	Lapping	Machine 1	Cable Head		2012-02-14 00:00	2012-02-14 01:15
$e_9$	Case <sub>1</sub>	Lapping	Machine 1	Cable Head		2012-02-14 09:05	2012-02-14 10:20
$e_{10}$	Case <sub>1</sub>	Lapping	Machine 1	Cable Head	...	2012-02-14 09:05	2012-02-14 09:38
$e_{11}$	Case <sub>1</sub>	Round Grinding	Machine 3	Cable Head		2012-02-14 09:13	2012-02-14 13:37
$e_{12}$	Case <sub>1</sub>	Round Grinding	Machine 3	Cable Head		2012-02-14 13:37	2012-02-14 15:27
$e_{13}$	Case <sub>1</sub>	Final Inspection Q.C.	Quality Check 1	Cable Head		2012-02-16 06:59	2012-02-16 07:59
$e_{14}$	Case <sub>1</sub>	Final Inspection Q.C.	Quality Check 1	Cable Head		2012-02-16 12:11	2012-02-16 16:12
$e_{15}$	Case <sub>1</sub>	Final Inspection Q.C.	Quality Check 1	Cable Head		2012-02-16 12:43	2012-02-16 13:58
$e_{16}$	Case <sub>1</sub>	Packing	Packing	Cable Head		2012-02-17 00:00	2012-02-17 01:00
$e_{17}$	Case <sub>2</sub>	Turning & Milling	Machine 9	Spur Gear		2012-01-17 07:01	2012-01-17 11:05
$e_{18}$	Case <sub>2</sub>	Turning Q.C.	Quality Check 1	Spur Gear		2012-01-17 11:06	2012-01-17 11:15
$e_{19}$	Case <sub>2</sub>	Turning & Milling	Machine 9	Spur Gear		2012-01-17 19:24	2012-01-17 20:01
$e_{20}$	Case <sub>2</sub>	Turning & Milling	Machine 9	Spur Gear		2012-01-17 20:01	2012-01-17 23:43
...	...	...	...	...		...	...

# 특별한 유형의 데이터를 처리하기 위한 다중 대체 방법

- 이벤트 로그 복구를 위한 Likelihood 기반 다중 대체

	Mobile Package	Download Speed	Data Limit Usage
$y_1$	NA	157	80%
$y_2$	Lite	99	NA
$y_3$	Fast	167	10%
$y_4$	Fast	NA	80%

Each included in a case are dependent.

General Data Set with missing

VS

	Case	Activity	Resource	Part Desc.
$e_1$	Case	NA	Machine 4	Cable Head
$e_2$	Case	Turning & Milling	NA	Cable Head
$e_3$	Case	Turning & Milling	Machine 4	NA
$e_4$	Case	Turning & Milling	Machine 4	Cable Head
$e_5$	Case	NA	NA	Cable Head
...	...	...	...	...

Start Time	End Time
2012-01-29 23:24	2012-01-30 05:43
2012-01-30 05:44	2012-01-30 06:42
2012-01-30 06:59	2012-01-30 07:21
2012-01-30 07:21	2012-01-30 10:58
2012-01-31 13:20	2012-01-31 14:50
...	...

Observation included in a case are dependent.

Event Log Structure with missing

# 특별한 유형의 데이터를 처리하기 위한 다중 대체 방법

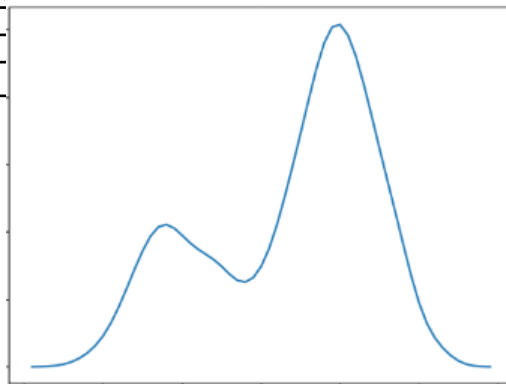
	Case	Activity
$e_1$	Case <sub>1</sub>	Turning & Milling
$e_2$	Case <sub>1</sub>	Turning & Milling
$e_3$	Case <sub>1</sub>	Turning & Milling
$e_4$	Case <sub>1</sub>	Turning & Milling
$e_5$	Case <sub>1</sub>	Turning & Milling Q.C
$e_6$	Case <sub>1</sub>	Laser Marking
$e_7$	Case <sub>1</sub>	Lapping
$e_8$	Case <sub>1</sub>	Lapping
$e_9$	Case <sub>1</sub>	Lapping
$e_{10}$	Case <sub>1</sub>	Lapping
$e_{11}$	Case <sub>1</sub>	Round Grinding
$e_{12}$	Case <sub>1</sub>	Round Grinding
$e_{13}$	Case <sub>1</sub>	Final Inspection Q.C.
$e_{14}$	Case <sub>1</sub>	Final Inspection Q.C.
$e_{15}$	Case <sub>1</sub>	Final Inspection Q.C.
$e_{16}$	Case <sub>1</sub>	Packing
...	...	...

Event

	Prior Event	Current Event	Posterior Event
$ec_1$	Start	Turning & Milling	Turning & Milling
$ec_2$	Turning & Milling	Turning & Milling	Turning & Milling
$ec_3$	Turning & Milling	Turning & Milling	Turning & Milling
$ec_4$	Turning & Milling	Turning & Milling	Turning & Milling Q.C
$ec_5$	Turning & Milling	Turning & Milling Q.C	Laser Marking
$ec_6$	Turning & Milling Q.C	Laser Marking	Lapping
$ec_7$	Laser Marking	Lapping	Lapping
$ec_8$	Lapping	Lapping	Lapping
$ec_9$	Lapping	Lapping	Lapping
$ec_{10}$	Lapping	Lapping	Round Grinding
$ec_{11}$	Lapping	Round Grinding	Round Grinding
$ec_{12}$	Round Grinding	Round Grinding	Final Inspection Q.C.
$ec_{13}$	Round Grinding	Final Inspection Q.C.	Final Inspection Q.C.
$ec_{14}$	Final Inspection Q.C.	Final Inspection Q.C.	Final Inspection Q.C.
$ec_{15}$	Final Inspection Q.C.	Final Inspection Q.C.	Packing
$ec_{16}$	Final Inspection Q.C.	Packing	End
...	...	...	...

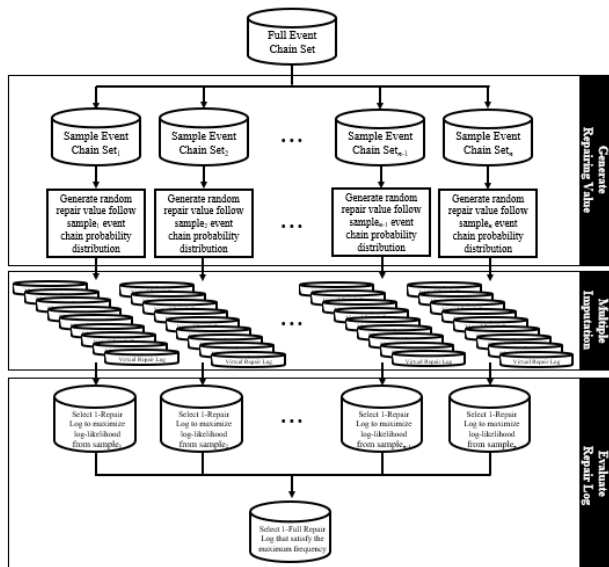
Event Chain

Fitting Target  
Distribution Using  
Event Chain



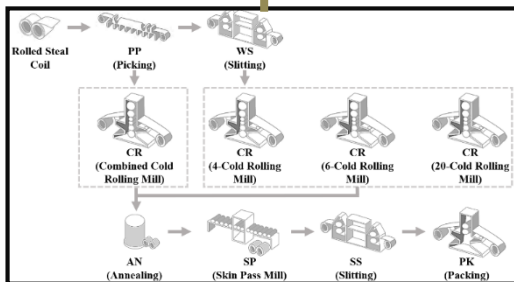
# 특별한 유형의 데이터를 처리하기 위한 다중 대체 방법

- We developed the concept of an event chain that can reflect sequential information contained in one case for dealing with missing events.
- By converting the events included in the event into an event chain, and replacing the event value that can approximate the distribution of the event chain instead of missing, we developed a restoration method that can restore the restored event log to the original log.



# Multiple Imputation Method for handling special types of data

- Likelihood based Multiple Imputation using Event chain for Repairing Event Log
  - Case Study: Korea Steel Company Event Log Data with MIEC(Multiple Imputation by Chained Equations, Expectation-Maximizing Imputation, Random Forest Imputation, K Nearest Neighbor Imputation)



COIL_NO	PRC_CD	PRC_CD1	THK	WDT	WGT	SDT	EDT	PLNPRC_CD	ORD_NO	DRTCOIL_NO	Machine
15KM12191111A	PP21	PP	5.99	1132	22900	01/04/2016 16:1600	01/04/2016 16:3800	PP21092	KD16090327	15KM1219111	PP2
15KM12191111A	CR21	RC	3.5	1132	22900	01/05/2016 10:1000	01/05/2016 10:5000	PP21092	KD16090327	15KM1219111	CR2
15KM12191111A	RC11	RC	3.5	1132	22900	01/05/2016 10:4500	01/05/2016 11:1000	PP21092	KD16090327	15KM1219111	RC1
15KM12191111A	WS31	WS	3.5	106	1717	01/07/2016 16:4000	01/07/2016 17:5000	PP21092	KD16090327	15KM1219111	WS3
15KM12191111A	PK41	PK	3.5	106	1736	01/07/2016 17:5500	01/07/2016 17:5500	PP21092	KD16090327	15KM1219111	PK4
15KM12191111A	PR11	PR	3.5	106	1736	01/07/2016 17:3000	01/07/2016 17:3000	PP21092	KD16090327	15KM1219111	PR1
15KM12191112A	PP21	PP	5.99	1132	22900	01/04/2016 16:1600	01/04/2016 16:3800	PP21092	KD16090327	15KM1219111	PP2
15KM12191112A	CR21	RC	3.5	1132	22900	01/05/2016 10:1000	01/05/2016 10:5000	PP21092	KD16090327	15KM1219111	CR2
15KM12191112A	RC11	RC	3.5	1132	22900	01/05/2016 10:4500	01/05/2016 11:1000	PP21092	KD16090327	15KM1219111	RC1
15KM12191112A	WS31	WS	3.5	106	1717	01/07/2016 16:4000	01/07/2016 17:5000	PP21092	KD16090327	15KM1219111	WS3
15KM12191112A	PK41	PK	3.5	106	1736	01/07/2016 17:5500	01/07/2016 17:5500	PP21092	KD16090327	15KM1219111	PK4
15KM12191112A	PR11	PR	3.5	106	1736	01/07/2016 17:3000	01/07/2016 17:3000	PP21092	KD16090327	15KM1219111	PR1
15KM12191113A	PP21	PP	5.99	1132	22900	01/04/2016 16:1600	01/04/2016 16:3800	PP21092	KD16090327	15KM1219111	PP2
15KM12191113A	CR21	RC	3.5	1132	22900	01/05/2016 10:1000	01/05/2016 10:5000	PP21092	KD16090327	15KM1219111	CR2
15KM12191113A	RC11	RC	3.5	1132	22900	01/05/2016 10:4500	01/05/2016 11:1000	PP21092	KD16090327	15KM1219111	RC1
15KM12191113A	WS31	WS	3.5	106	1717	01/07/2016 16:4000	01/07/2016 17:5000	PP21092	KD16090327	15KM1219111	WS3
15KM12191113A	PK41	PK	3.5	106	1743	01/07/2016 17:5400	01/07/2016 17:5400	PP21092	KD16090327	15KM1219111	PK4
15KM12191113A	PR11	PR	3.5	106	1743	01/07/2016 17:3000	01/07/2016 17:3000	PP21092	KD16090327	15KM1219111	PR1
15KM12191113B	PP21	PP	5.99	1132	22900	01/04/2016 16:1600	01/04/2016 16:3800	PP21092	KD16080281	15KM1219111	PP2

Missing Rate	MICE	EMBI	RFI	KNNI	MIEC
5%	48.9%	51.3%	60.1%	44.4%	93.2%
10%	44.3%	45.5%	59.8%	42.1%	91.0%
15%	34.7%	36.9%	50.3%	40.3%	88.7%
20%	28.5%	30.1%	47.7%	32.2%	80.4%

Performance Event Imputation method using MIEC (Korea Steel Company Event log)