

2-2-1 데이터 가공의 이해

Contents

1. 데이터 가공 프로세스
2. 데이터정제를 위한 기준 및 처리방법
3. 데이터 가공
4. 데이터 품질평가
5. 거래데이터 품질진단 및 절차
6. AI데이터 가공 사례

1. 데이터가공 프로세스

1.데이터 수집: 다양한 소스에서 데이터를 수집. 데이터의 종류, 형식, 출처 등을 파악하고, 수집 방법을 결정

2.데이터 정제: 수집한 데이터를 정리하고, 불필요한 정보를 제거. 데이터의 오류를 수정하고, 중복된 데이터를 제거

3.데이터 가공: 정제된 데이터를 분석 목적에 맞게 가공. 데이터를 분류하고, 예측 모델을 구축을 위한 작업

4.데이터 분석: 가공한 데이터를 분석하여 인사이트를 도출. 통계 분석, 머신러닝 등의 기술을 활용하여 데이터를 분석

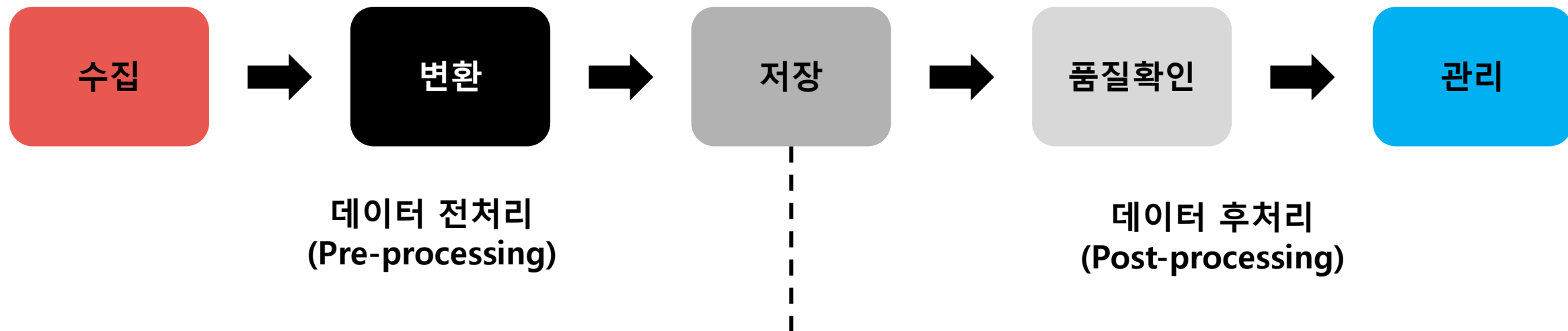
5.데이터 시각화: 분석 결과를 시각적으로 표현. 그래프, 차트, 대시보드 등을 활용하여 데이터를 시각화

6.데이터 활용: 분석결과 기반 의사결정, 새로운 정책에 활용

1. 데이터가공 프로세스

데이터 정제 및 가공

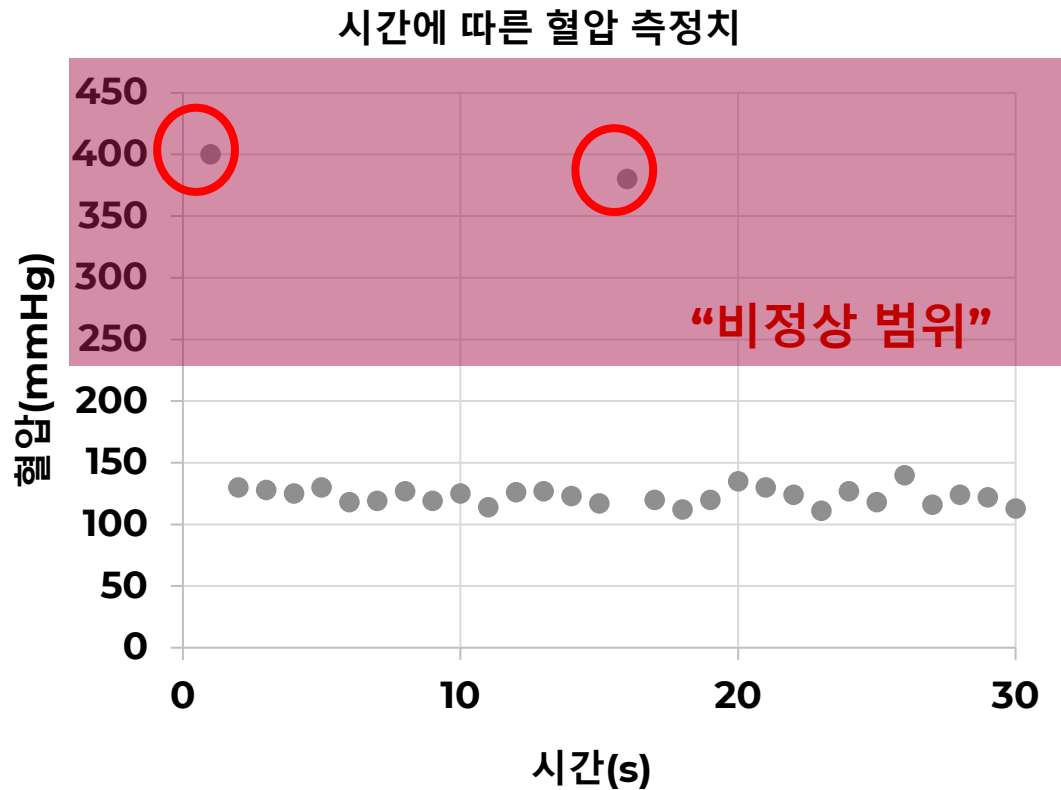
- 수집된 데이터로부터 분석에 필요한 데이터를 추출하고 통합
- 이상치, 결측치 등을 처리



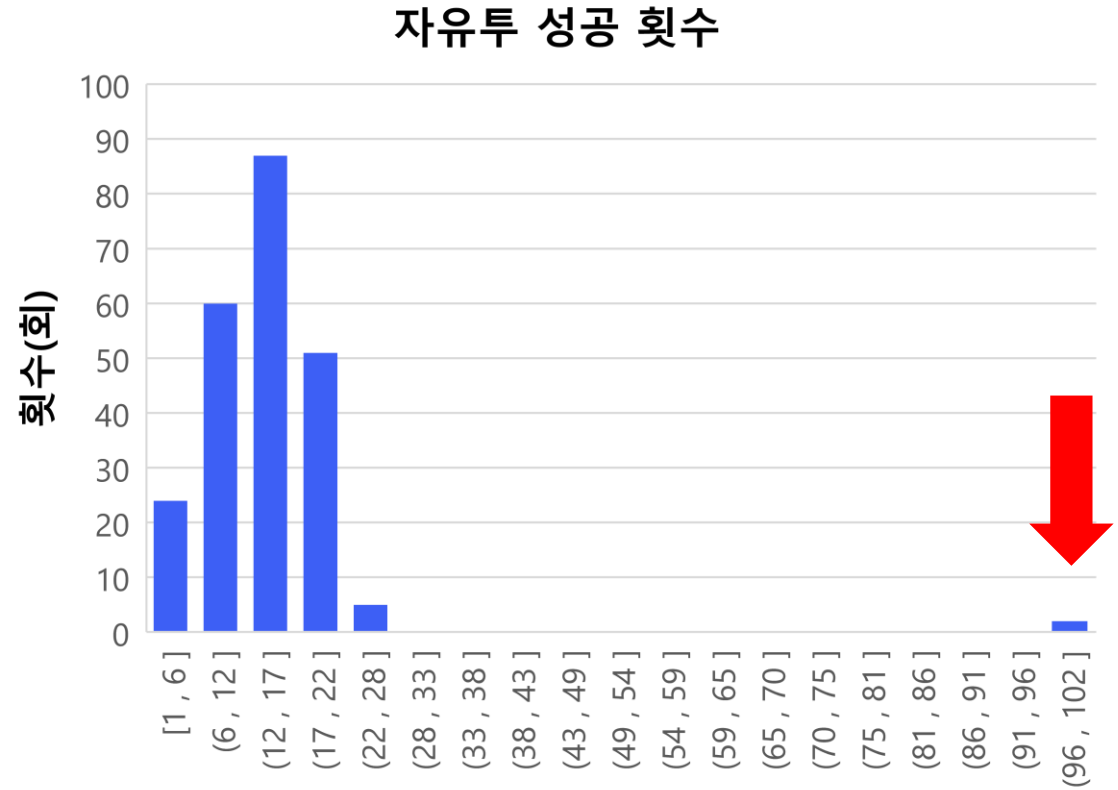
2.데이터 정제를 위한 기준 및 처리 방법

데이터 정제 기준

이상치(Outlier) : 모든 데이터가 상식적인 범위 안에 존재하는가



(건강보험공단 시범 DB의 혈압측정치)



(자유투 Free throw : 10-20 per game)

2.데이터 정제를 위한 기준 및 처리 방법

예시 1: 스포츠 데이터, NBA 자유투 데이터 포스트 시즌 2023-24

ESPN NBA Home Scores Schedule Standings Stats Teams Odds Fantasy Men's Basketball

NBA Player free-throws Stats 2023-24

2023-24 Regular Season All Positions All NBA

RK	NAME	POS	GP	MIN	PTS	FGM	FGA	FG%	3PM
1	Stephen Curry GS	PG	74	32.7	26.4	8.8	19.5	45.0	4.8
2	Luka Doncic DAL	PG	70	37.5	33.9	11.5	23.6	48.7	4.1
3	CJ McCollum NO	SG	66	32.7	20.0	7.3	16.0	45.9	3.6
4	Donte DiVincenzo NY	SG	81	29.1	15.5	5.5	12.4	44.3	3.5
5	Klay Thompson GS	SG	77	29.7	17.9	6.4	14.7	43.2	3.5
6	Anfernee Simons POR	SG	46	34.4	22.6	7.8	18.2	43.0	3.4
7	Donovan Mitchell CLE	SG	55	35.3	26.6	9.1	19.8	46.2	3.3
8	Paul George LAC	F	74	33.8	22.6	7.9	16.7	47.1	3.3
9	Desmond Bane MEM	SG	42	34.4	23.7	8.6	18.5	46.4	3.3
10	Trae Young ATL	PG	54	36.0	25.7	8.0	18.7	43.0	3.2

게임당 자유투 점수 (points per game)

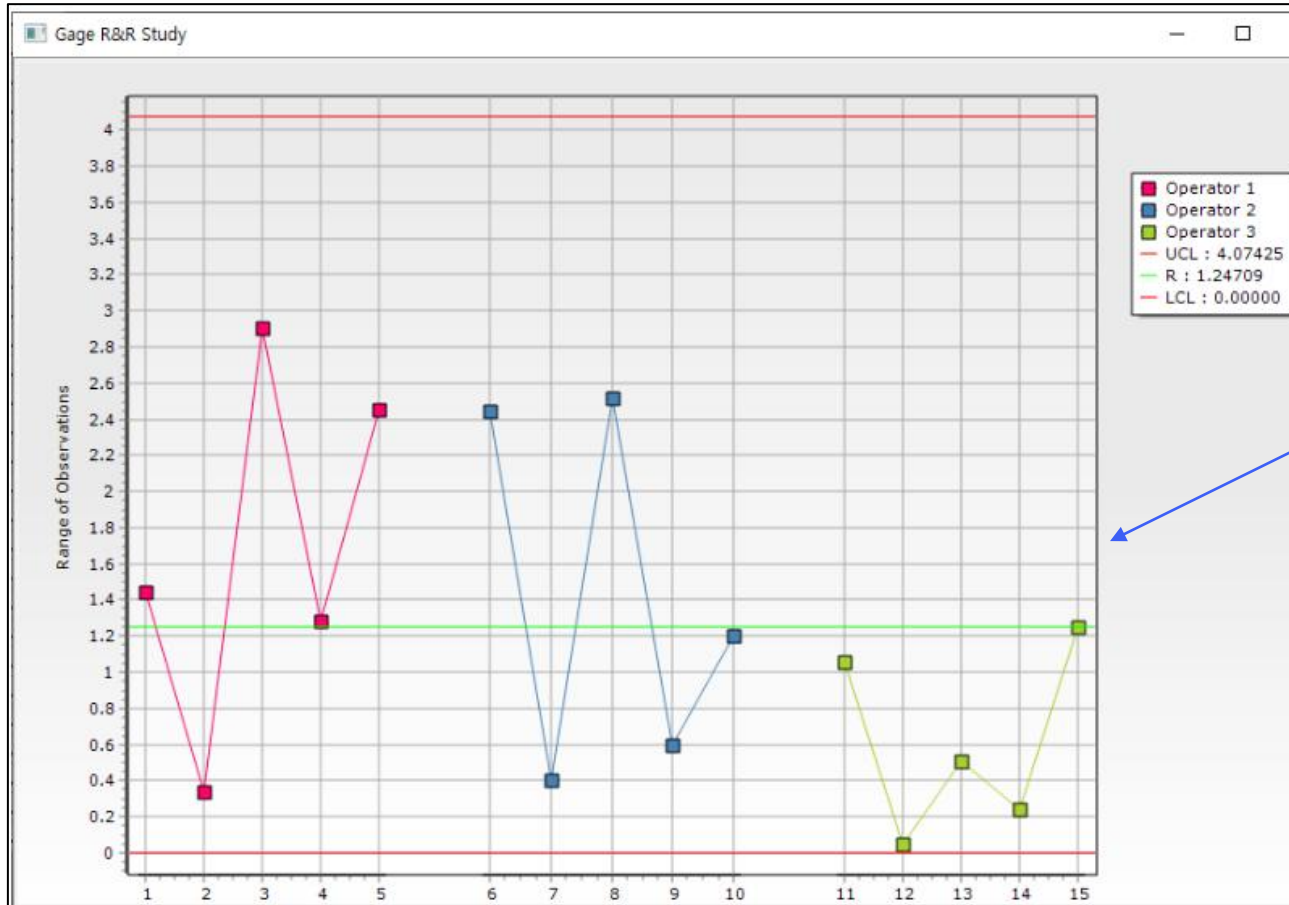
3점 슛 평균회수(게임당)



2.데이터 정제를 위한 기준 및 처리 방법

예시 2 : 제조업 측정데이터의 재현성 평가

Gage R & R(Repeatability and Reproducibility)



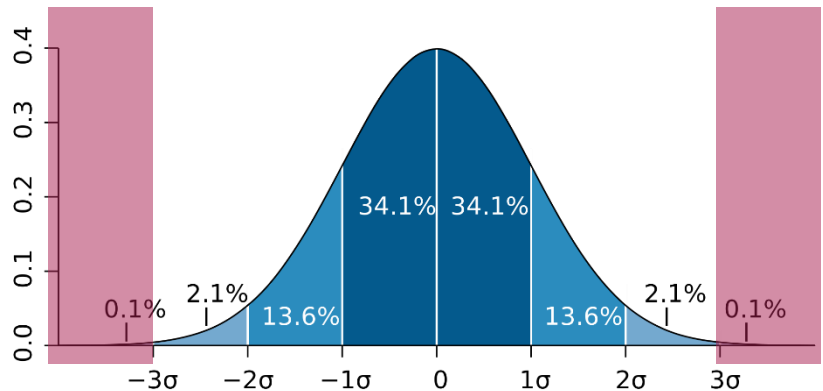
Repeatability(반복성) : 반복 측정에서 일관된 값이 측정되는지 (동일 측정자, 동일기계에 반복적 측정)

Reproducibility(재현성) : 측정자들이 동일 측정기를 사용하여 얻은 측정된 값들이 일관되게 나오는지 (다른 측정자(사람, 시점), 동일 기계 사용)

2.데이터 정제를 위한 기준 및 처리 방법

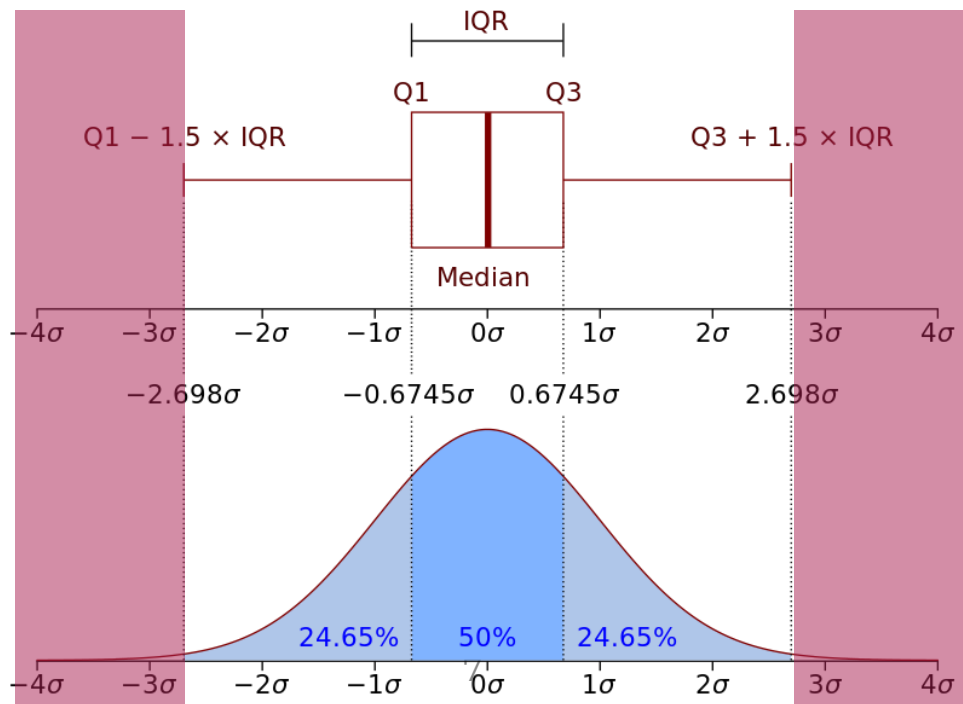
데이터 정제 기준

- 3-시그마 규칙



- 1.5 IQR 규칙

Inter Quartile Range
= $3Q(75\%) - 1Q(25\%)$



* 이상치로 볼 수 있음

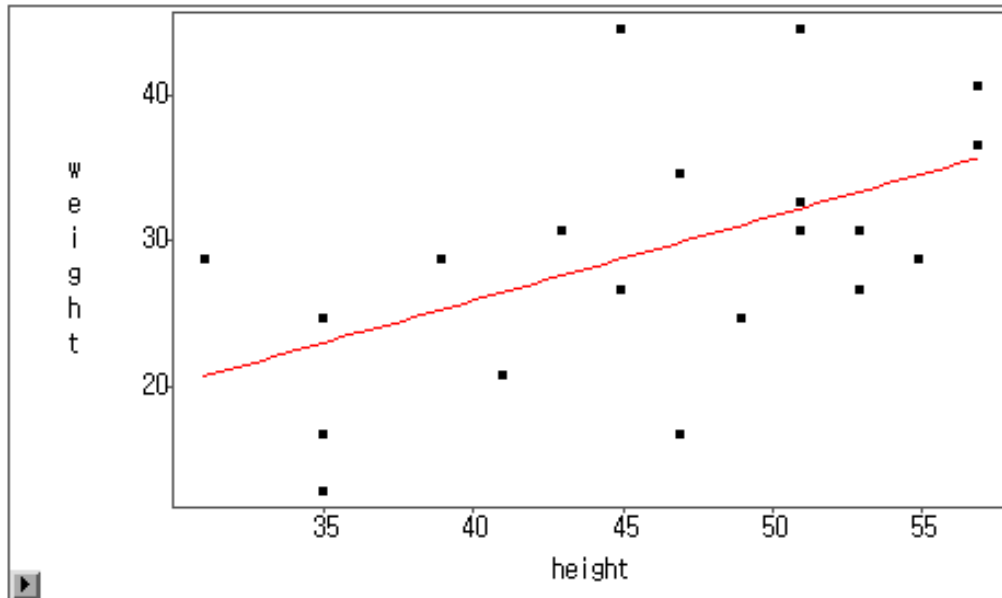
2. 데이터 정제를 위한 기준 및 처리 방법

예시 3 : 이상치 및 오류 데이터

통계치의 왜곡 - Monkey 데이터

Model Equation
weight = 2.7356 + 0.5797 height

상관계수 : 0.53



Summary of Fit

Mean of Response	29.4000	R-Square	0.2775
Root MSE	7.5730	Adj R-Sq	0.2374



ID	height	weight
1	55	29
2	45	27
3	35	17
4	39	29
5	53	31
6	41	21
7	51	31
8	35	13
9	57	37
10	57	41
11	45	45
12	47	35
13	35	25
14	49	25
15	43	31
16	51	33
17	31	29
18	53	27
19	47	17
20	51	45

2. 데이터 정제를 위한 기준 및 처리 방법

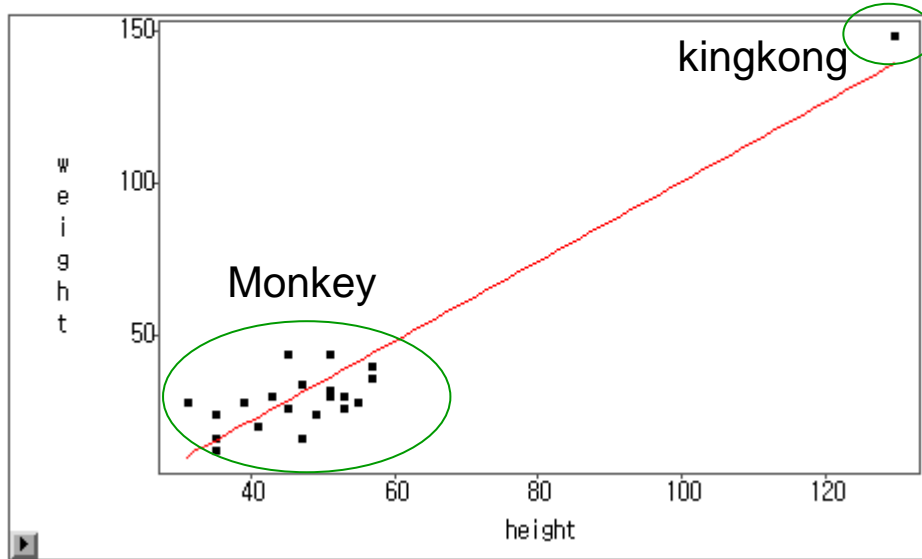
예시 3 : 이상치 및 오류 데이터

Monkey 데이터 + Kingkong 한마리

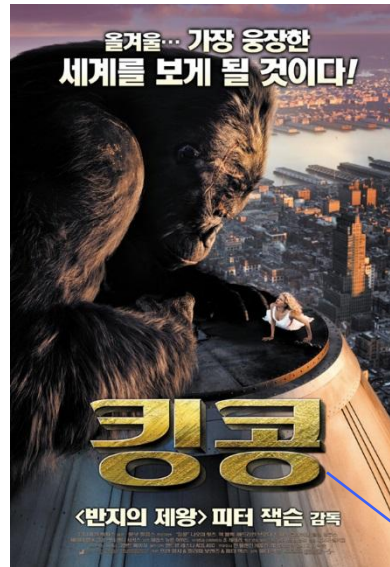
Model Equation
weight = - 30.2495 + 1.3078 height

상관계수 : 0.94

상관계수 : 0.53



Summary of Fit
Mean of Response 35.1429 R-Square 0.8843
Root MSE 9.6461 Adj R-Sq 0.8782



ID	height	weight
1	55	29
2	45	27
3	35	17
4	39	29
5	53	31
6	41	21
7	51	31
8	35	13
9	57	37
10	57	41
11	45	45
12	47	35
13	35	25
14	49	25
15	43	31
16	51	33
17	31	29
18	53	27
19	47	17
20	51	45
21	130	150

2. 데이터 정제를 위한 기준 및 처리 방법

데이터 정제 : 이상치 및 오류 데이터

ID	height	weight
1	55	29
2	45	27
3	35	17
4	39	29
5	53	31
6	41	21
7	51	31
8	35	NaN
9	57	37
10	57	41
11	45	45
12	47	NaN
13	35	25
14	49	25
15	43	31
16	51	33
17	31	29
18	53	27
19	47	17
20	51	45
21	130	150

결측치(Missing Value)

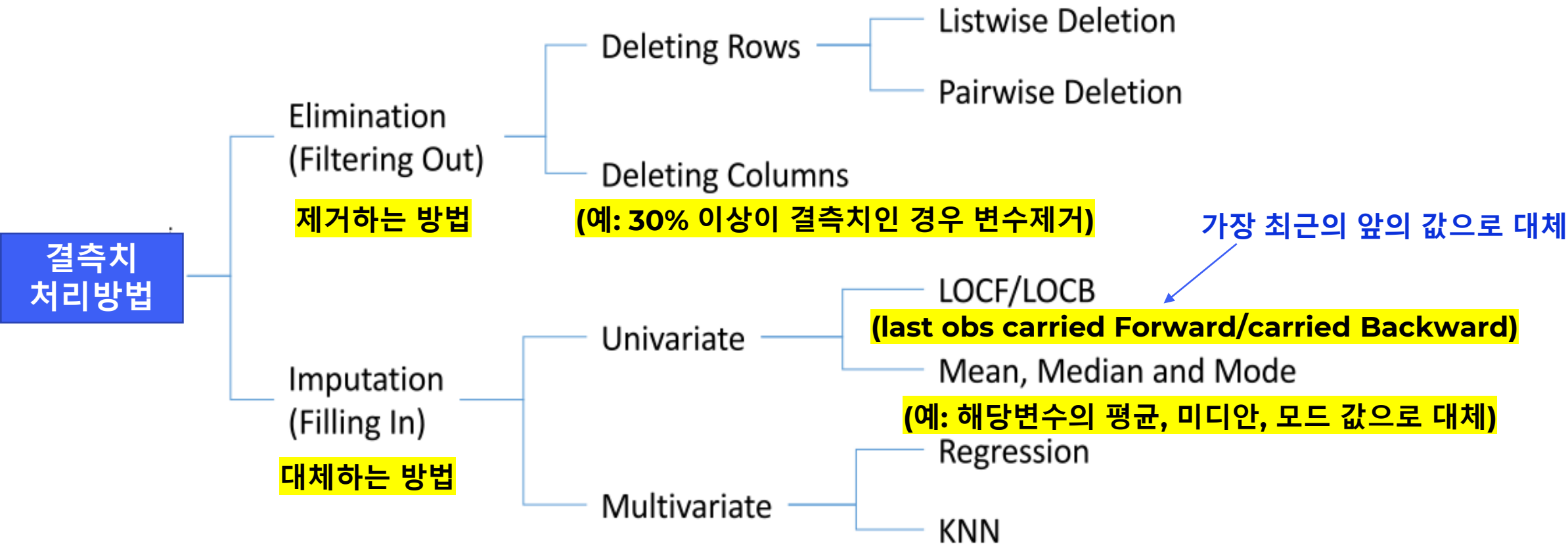
- 입력이 누락된 값
- 무시하거나 적절한 값으로 대체

이상치(Outlier)

- 이상치 발생 원인 파악
- 삭제, 대체, 변환 등을 통해 처리

2. 데이터 정제를 위한 기준 및 처리 방법

이상치 및 오류 데이터의 처리방법



2. 데이터 정제를 위한 기준 및 처리 방법

예시 : 결측치 처리방법 (LOCF 예시)

환자의 혈압데이터 (일별 측정 – 시계열 데이터로 볼수 있음)

Day	Blood Pressure
Day 1	120
Day 2	125
Day 3	Missing
Day 4	Missing
Day 5	130



Day	Blood Pressure (LOCF)
Day 1	120
Day 2	125
Day 3	125
Day 4	125
Day 5	130

수축기혈압

2. 데이터 정제를 위한 기준 및 처리 방법

데이터 유형별 오류 형태

문자열

사람별 거주 도시

ID	Name	City
1	Jay	Washington
2	Susan	washington
3	Lee	NA
4	Michael	Argentina
5	Park	33
6	Max	Londin
7	Jay	Washington

형태 불일치

결측치

유형 불일치

유형 불일치

오타

데이터 중복

범주형

학생별 학점

ID	Name	Grade
1	Jay	D
2	Susan	A
3	Lee	NA
4	Michael	A
5	Park	8
6	Max	C
7	Jay	D

결측치

유형 불일치

데이터 중복

수치형

학생별 키

ID	Name	Height
1	Jay	178
2	Susan	170
3	Lee	NA
4	Michael	-10
5	Park	input
6	Max	180
7	Jay	178

결측치

모순된 데이터

유형 불일치

데이터 중복

3. 데이터 가공

분석을 위한 데이터 가공

- 모델에 적용하기 위한 데이터 가공
- 결측치 처리, 이상치 처리, 데이터 형식 변환 등 데이터 전처리
- 변수 선택, 차원 축소, 파생변수 생성 등을 통해 추가 가공

관리를 위한 데이터 가공

- 데이터를 저장, 검색, 접근, 백업 및 보관의 효율화
- 데이터의 품질을 안정적으로 유지 및 개선
- 데이터 정규화, 중복 제거, 보안을 위한 접근 제어, 데이터 버전 관리 등

3. 데이터 가공

데이터가공 : 분석용 데이터 구축

서울시립과학관 관람객 현황(2017-2023)

서울시립과학관 2017년 개관 이래 2023년 9월말 현재까지의 관람객 현황에 대한 데이터로 연도, 과학관명, 관람객 인원 등의 항목을 제공합니다.

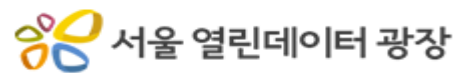
	년월	과학관명	관람객 인원
0	2017년 05월	서울시립과학관	31620
1	2017년 06월	서울시립과학관	18827
2	2017년 07월	서울시립과학관	16734
3	2017년 08월	서울시립과학관	24244
4	2017년 09월	서울시립과학관	14189
5	2017년 10월	서울시립과학관	18881
6	2017년 11월	서울시립과학관	17077
7	2017년 12월	서울시립과학관	29013
8	2018년 01월	서울시립과학관	19647
9	2018년 02월	서울시립과학관	12010

•
•
•



- 결측치와 이상치가 존재하는가?
- 존재한다면 처리를 어떻게 할 것인가?
- 데이터를 시각화 해본다면?
- 분기별/연도별 관람객 인원은?

...



3. 데이터 가공

데이터가공 : 분석용 데이터 구축

```
# 관람객 인원이 수치형 데이터인지 확인
is_numeric_dtype(df['관람객 인원'])
```

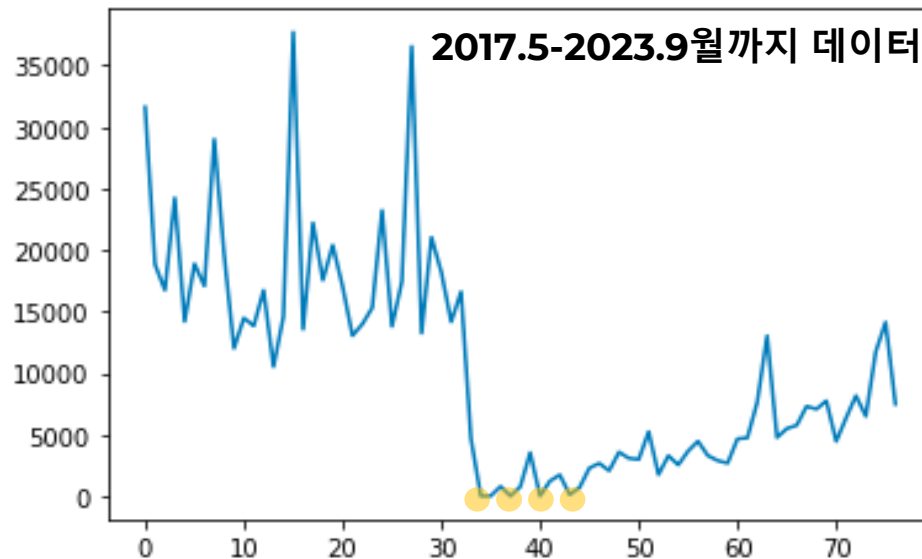
✓ 0.0s

False

2019년 12월	서울시립과학관	14165
2020년 01월	서울시립과학관	16607
2020년 02월	서울시립과학관	4676
2020년 03월	서울시립과학관	휴관
2020년 04월	서울시립과학관	휴관
2020년 05월	서울시립과학관	778
2020년 06월	서울시립과학관	휴관
2020년 07월	서울시립과학관	755
2020년 08월	서울시립과학관	3528
2020년 09월	서울시립과학관	휴관
2020년 10월	서울시립과학관	1216
2020년 11월	서울시립과학관	1723
2020년 12월	서울시립과학관	108
2021년 01월	서울시립과학관	684
2021년 02월	서울시립과학관	2277

정상 관람객수

	년월	과학관명	관람객 인원
34	2020년 03월	서울시립과학관	휴관
35	2020년 04월	서울시립과학관	휴관
37	2020년 06월	서울시립과학관	휴관
40	2020년 09월	서울시립과학관	휴관



3. 데이터 가공

서울시립과학관 관람객 현황(2017-2023)

년월	과학관명	관람객 인원
2019년 11월	서울시립과학관	18202
2019년 12월	서울시립과학관	14165
2020년 01월	서울시립과학관	16607
2020년 02월	서울시립과학관	4676
2020년 03월	서울시립과학관	휴관
2020년 04월	서울시립과학관	휴관
2020년 05월	서울시립과학관	778
2020년 06월	서울시립과학관	휴관
2020년 07월	서울시립과학관	755
2020년 08월	서울시립과학관	3528
2020년 09월	서울시립과학관	휴관
2020년 10월	서울시립과학관	1216
2020년 11월	서울시립과학관	1723
2020년 12월	서울시립과학관	108
2021년 01월	서울시립과학관	684
2021년 02월	서울시립과학관	2277
2021년 03월	서울시립과학관	2657
2021년 04월	서울시립과학관	2055
2021년 05월	서울시립과학관	3571
2021년 06월	서울시립과학관	3063

이더로 연도, 과학관명, 관람객 인원 등의 항목을 제공합니다.

- (1) 휴관은 관람객 인원 0으로 처리
혹은
- (2) 휴관 (월)은 제외?

제1급 법정감염병 지정기간

2020년 1월 20일부터 2022년 4월 24일까지

질문1 : 2020년의 월평균 관람객수는?

$$\text{Sum}(20.1-20.12)/8=3673$$

$$\text{Sum}(20.1-20.12)/12=2449$$

질문2: 코로나 기간중 2020년의 월평균 관람객수는?

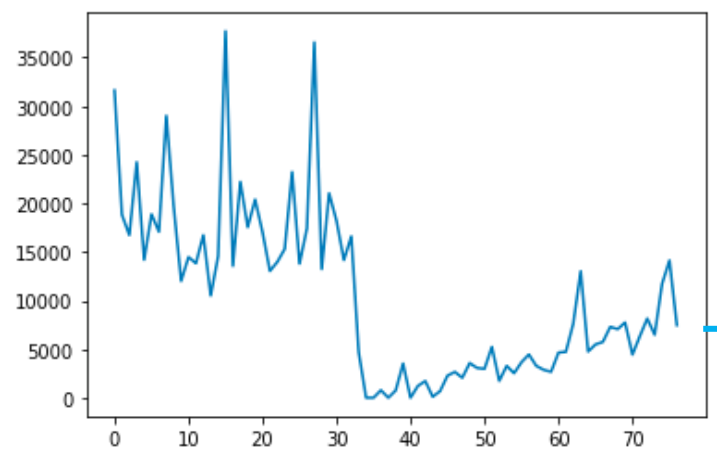
(2020.2월-2020.12월까지로 산정)

3. 데이터 가공



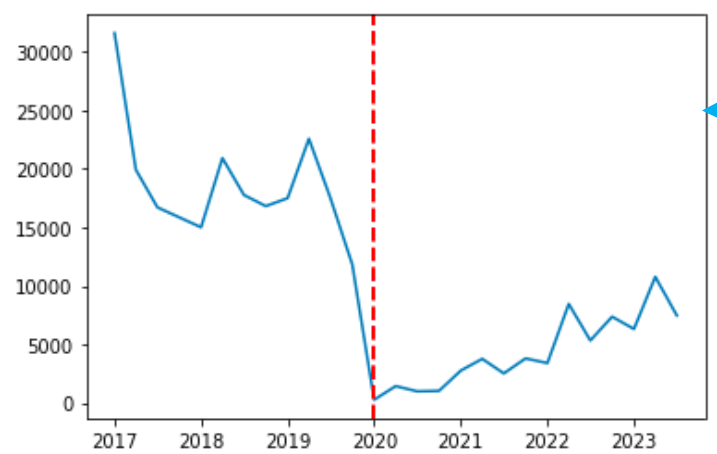
데이터가공 : 분석용 데이터 구축

제1급 법정감염병 지정기간
2020년 1/20 -2022년 4/24



1차 통계량 - 직접 얻어진 통계량

분기별 평균 관람객 수



2차 통계량(가공통계) - 일정한 연산을 가하여 얻어진 통계

	분기	분기별 관람객 인원
0	2017Q1	31620
1	2017Q2	19935
2	2017Q3	16715
3	2018Q1	15010
4	2018Q2	20929
5	2018Q3	17774
6	2018Q4	16823
7	2019Q1	17502
8	2019Q2	22588
9	2019Q3	17491
10	2019Q4	11816
11	2020Q1	259
12	2020Q2	1427
13	2020Q3	979

3. 데이터 가공

- 정형/비정형 데이터
- 특정시기별 구분 (코로나기간/일반기간)
- ‘휴관’- 숫자가 아닌 데이터 처리 =>
- 분기별 관람객 분석 (파생변수 생성 -> 가공데이터 생성)



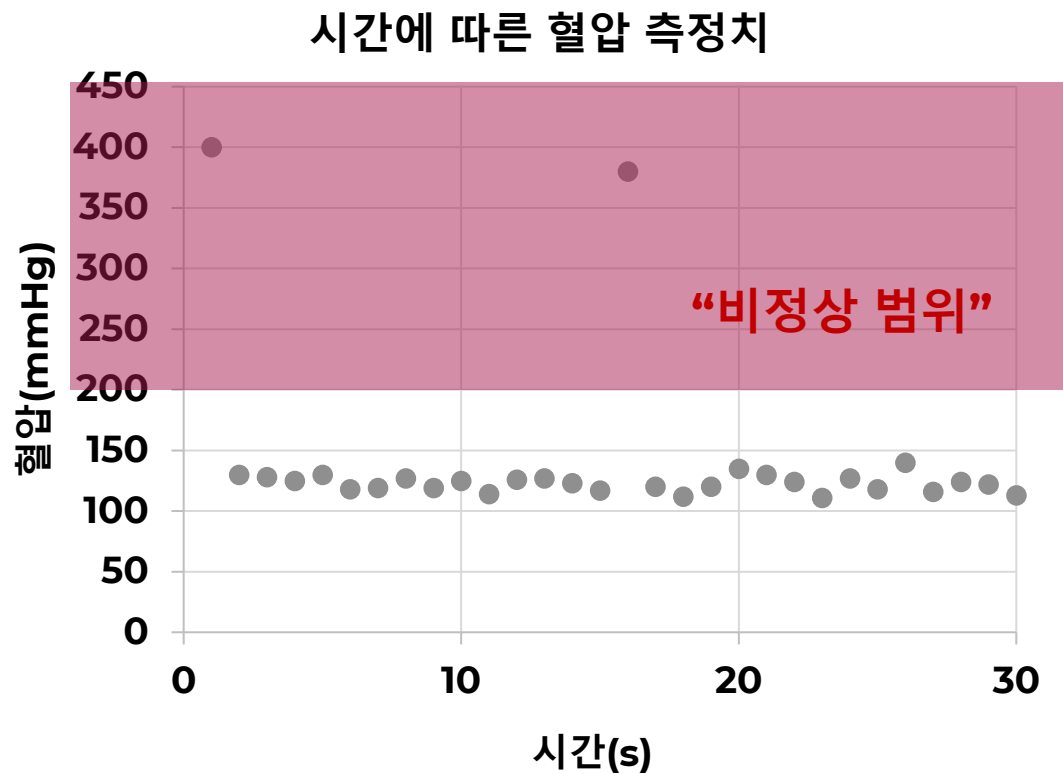
목적?

4. 데이터 품질평가

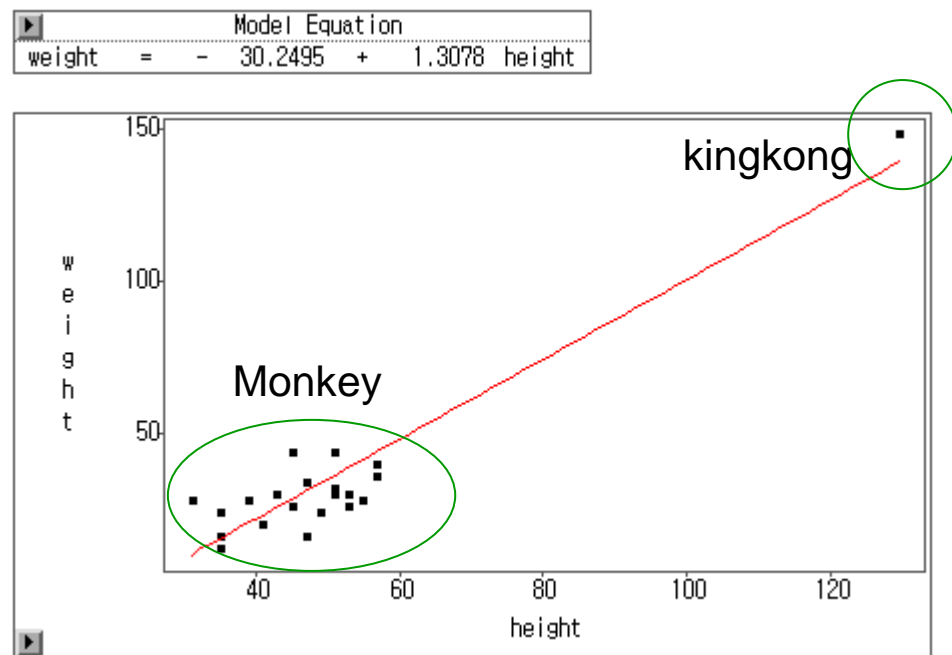
- 완전성 (completeness) : 데이터가 필요한 모든 정보를 갖고 있는지 여부
- 정확성 (Accuracy): 데이터가 실제를 정확하게 반영 여부. 결측치, 오류 또는 부정확한 값 여부
- 일관성 (Consistency): 데이터가 동일한 형식으로 표현되고 일관된 값으로 저장
- 유효성 (Validity): 데이터가 정확한 형식과 범위 내.
- 신뢰성 (Reliability): 데이터 수집방법의 신뢰성
- 시간성 (Timeliness): 데이터가 실시간 정보가 반영되는지. 업데이트 여부
- 가용성 (Accessibility): 데이터에 쉽게 접근 가능성, 사용자가 필요할 때 사용 가능한가

4. 데이터 품질평가

데이터 품질 : 유효성 예시 (혈압)



데이터 품질 : 정확성 예시 (데이터 오류)



4. 데이터 품질평가

- 데이터 신뢰성 확보
- 데이터 구매-수요자 간 **품질 증명**
- 품질 관리를 통한 원활한 데이터의 활용



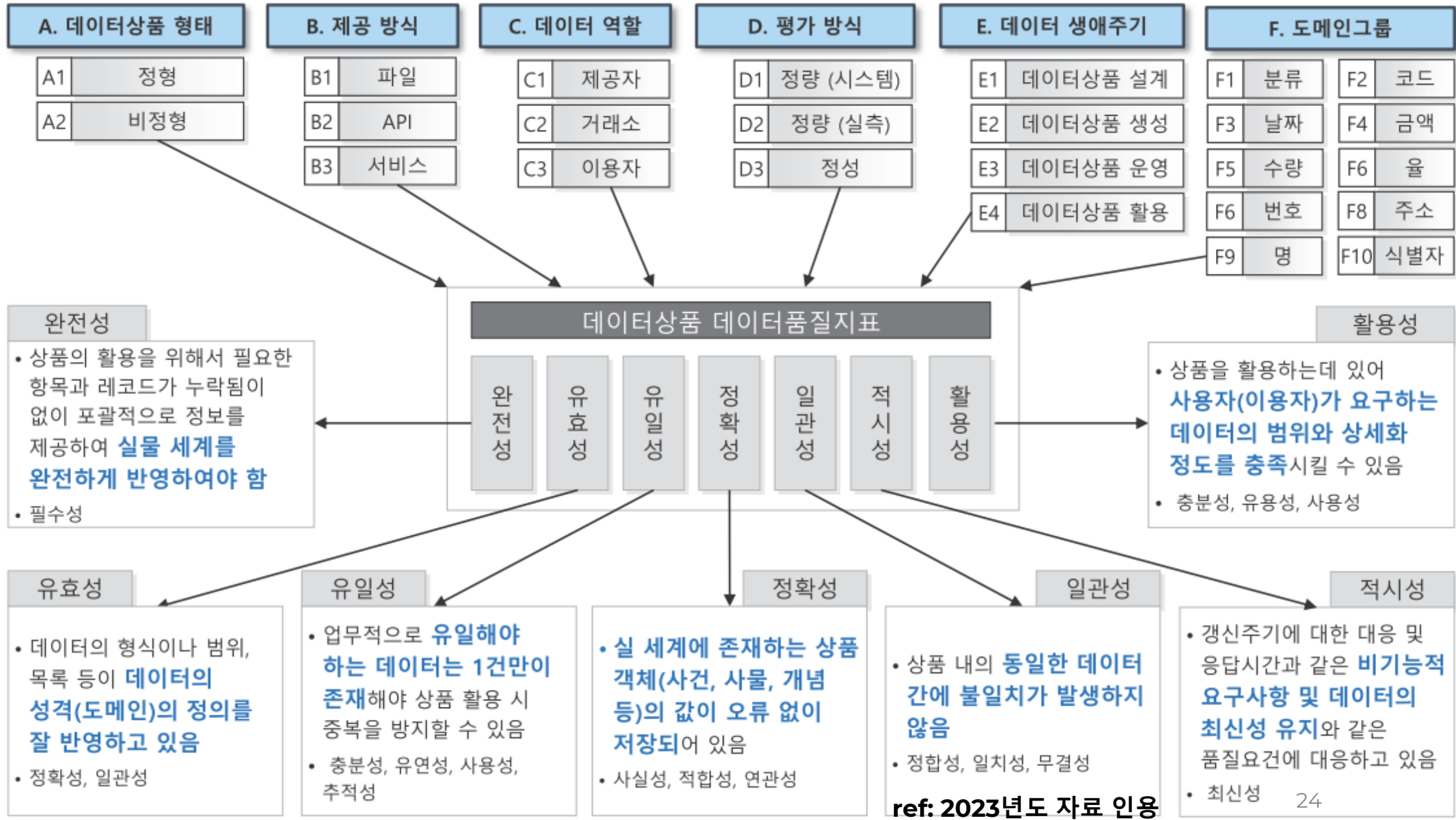
4. 데이터 품질평가

한국데이터거래소 (<https://kdx.kr/main>) : 민간 데이터거래소 (무료/ 유료)

인기 AI 데이터

 <p>K 패션 학습용 데이터셋</p> <p>인공지능 2023.08.04 업데이트</p> <p>URBAN UNION 어반유니온</p> <p>AI 학습용 패션 데이터셋(샘플)</p> <p>K패션 표준화를 위한 인공지능 학습용 패션 데이터</p> <p>무료</p>	 <p>2020 AI 해커톤 활용 대용량 동영상 콘텐츠 AI 학습데이터의 ...</p> <p>인공지능 2020.10.30 업데이트</p> <p>MBN MBN</p> <p>2020 인공지능 학습용 데이터 해커톤에 사용되는 대용량 동영상 데이터의 객체...</p> <p>무료</p>	 <p>방송 카테고리별 + 인터뷰 텍스트 데이터</p> <p>인공지능 2022.11.10 업데이트</p> <p>KDX KDX한국데이터거래소</p> <p>방송 카테고리별 + 인터뷰 텍스트 데이터</p> <p>매일방송 프로그램 카테고리별 + 인터뷰로 분류한 텍스트 데이터</p> <p>22,000,000원</p>	 <p>방송 인터뷰 텍스트 데이터</p> <p>인공지능 2022.11.10 업데이트</p> <p>KDX KDX한국데이터거래소</p> <p>방송 인터뷰 텍스트 데이터</p> <p>매일방송 프로그램, 인터뷰로 분류한 텍스트 데이터</p> <p>8,800,000원</p>
--	---	---	---

4. 데이터 품질평가



5. 거래데이터 품질진단 및 절차

- 공공데이터, 마이데이터의 품질 진단은 데이터수집, 데이터정제, 데이터분석, 품질평가, 피드백 및 개선과 지속적인 관리 과정을 통해 가능함



5. 거래데이터 품질진단 및 절차

- 마이데이터는 금융거래소와 한국데이터거래소

금융데이터거래소 : <https://www.findatamall.or.kr/>



5. 거래데이터 품질진단 및 절차

한국데이터거래소 (<https://kdx.kr/main>) : 민간 데이터거래소

KDX 한국데이터거래소


데이터 마켓 AI 비즈니스 서비스 커뮤니티 기업회원 사이트 데이터 바우처

KDX를 통해 다양한 기업들이 데이터를 판매합니다.

Logos displayed on the page include: MBN, KDX 한국데이터거래소, CJ 올리브네트웍스, KB 캐피탈, KDX SK telecom Academy Data, SK telecom, Kyung Dong City Gas, dun & bradstreet, NICE 디앤알, NH농협은행, Samsung Card, 신한카드, SUMMERCE PLATFORM, URBAN UNION, LG CNS, LG U+, Welcome, WPS, GS 리테일, KCB.

5. 거래데이터 품질진단 및 절차


• 공공데이터 – 지역 공공데이터 포탈



Busan is good
부산이 좋다
Busan Open Data Portal

데이터 부산통계 맞춤형 데이터 인포그래픽 이용안내

“ 부산지역 공공데이터가 필요하신가요? ”



부산시 미세먼지
변화 한눈에보기

부산시 미세먼지 변화 분석 인포그래픽

인포그래픽 바로가기

인기 데이터 | 최신업데이트 데이터

- 01 부산광역시 남구_인구현황 >
- 02 부산광역시_강서구_사회조사결과 정보 >
- 03 부산광역시 해운대구_물가관리 >
- 04 부산광역시_기장군_주민등록 인구현황 >
- 05 부산교통공사_부산도시철도 열차내 안내방송 >

정부공공데이터포털
국가에서 보유하고 있는 데이터
통합포털

LOD
부산 영화·영상·관광 Linked
Open Data

공공데이터지도
지도 기반의 부산 공공데이터·
안전식생활·자동심장충격기·
상권 정보

인허가업종데이터
지방자치에서 인허가 하는 업
종별 데이터 개방

교육

국토관리

공공행정

재정금융

산업고용

사회복지

식품건강

문화관광

보건의료

재난안전

교통물류

환경기상

과학기술

























농축수산

법률

5. 거래데이터 품질진단 및 절차

- 금융관련 AI플랫폼

마이데이터 사업 : 종합금융 플랫폼

 KB증권 마이데이터 KB증권   M-able을 통해 제공되는 KB증권의 마이데이터 서비스...	 마이데이터 서비스 주식회사 아이엠뱅크   iM뱅크에서 편하게 나의 자산 관리	 캐시노트 주식회사 한국신용...   사장님을 위한 경영관리 서비스, 캐시노트	 모니모 자산관리 삼성카드주식회사   삼성금융네트웍스 통합 플랫폼 모니모에서 제공되는 마이데...
 하나Pay 하나카드   내 자산, 지출 분석부터 주변 핫플레이스까지 한 눈에	 한국투자 한국투자증권 (주)   한국투자증권 대표 MTS 한국투자입니다.	 KB Pay 마이데이터 KB국민카드   부자되기 첫 걸음! KB Pay로 쉽게 관리하는 내 ...	 나이스지킴이 나이스평가정보(주)   나를 가장 잘 아는 마이데이터, 나이스지킴이

마이데이터 종합포털 : <https://www.mydatacenter.or.kr:3441/myd/mydsvc/sub2.do>

5. 거래데이터 품질진단 및 절차

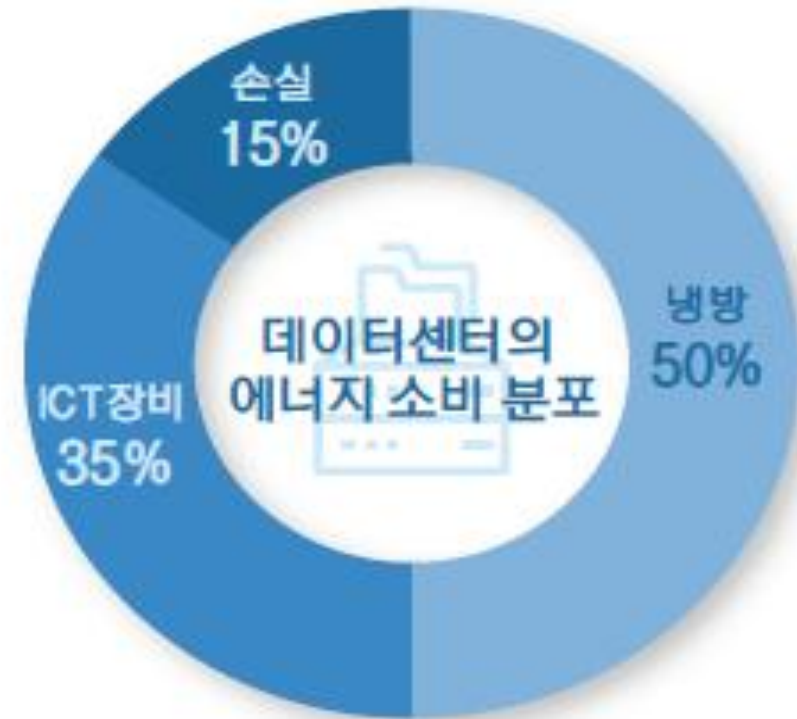


데이터와 환경, ESG : Environmental, Social, Governance)

“ 데이터센터 ”



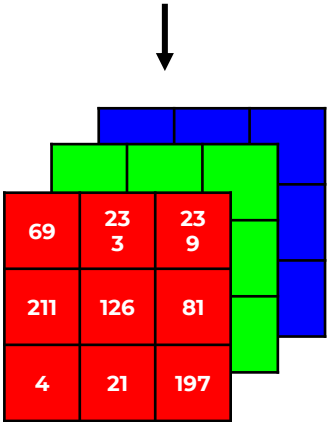
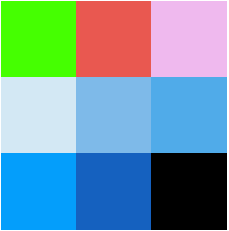
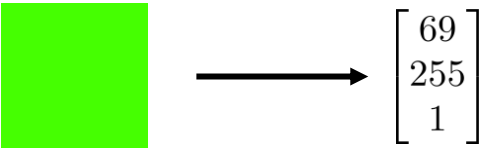
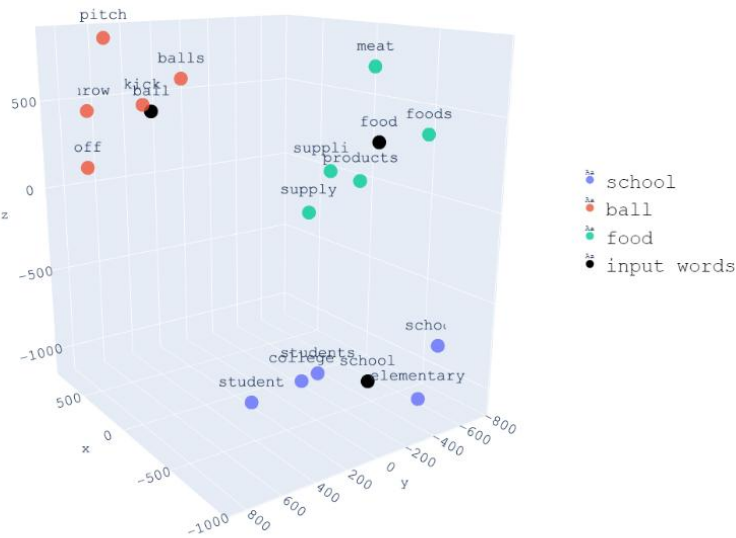
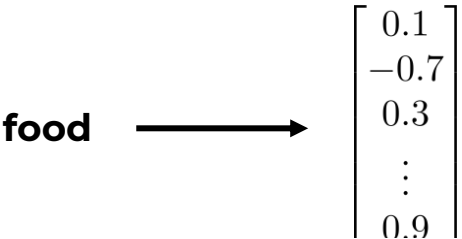
클라우드, 온라인 게임, VOD, 소셜네트워크(SNS)로 주고받는 메시지와 자료 등 각종 온라인 데이터를 저장하고 전송하는 시설



[자료=에너지경제연구원]

6. AI데이터 가공 사례

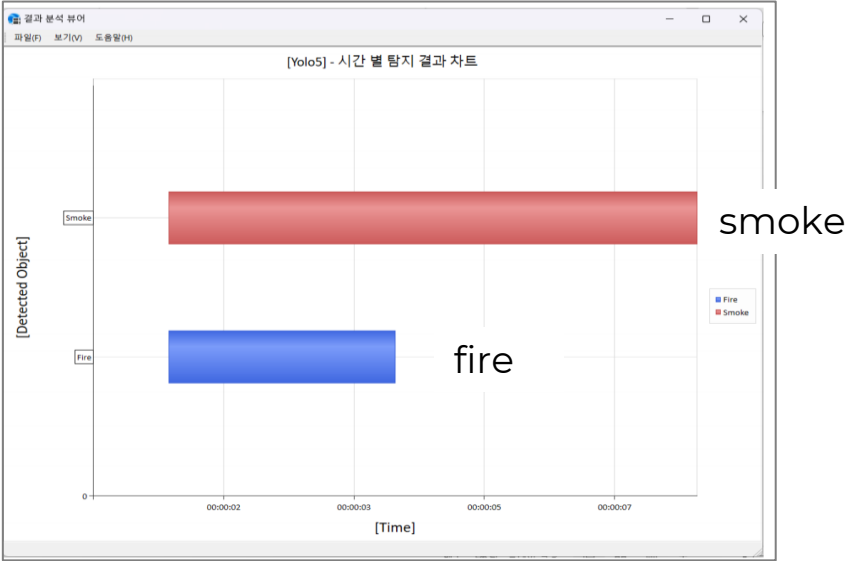
- 비정형 데이터(텍스트, 이미지, 오디오 등)를 정형 데이터로 변환하는 작업



6. AI데이터 가공 사례

(1) 이미지(영상)데이터 분석을 활용한 위험감지 경보 시스템 (산업체)

시간별 모니터링 데이터

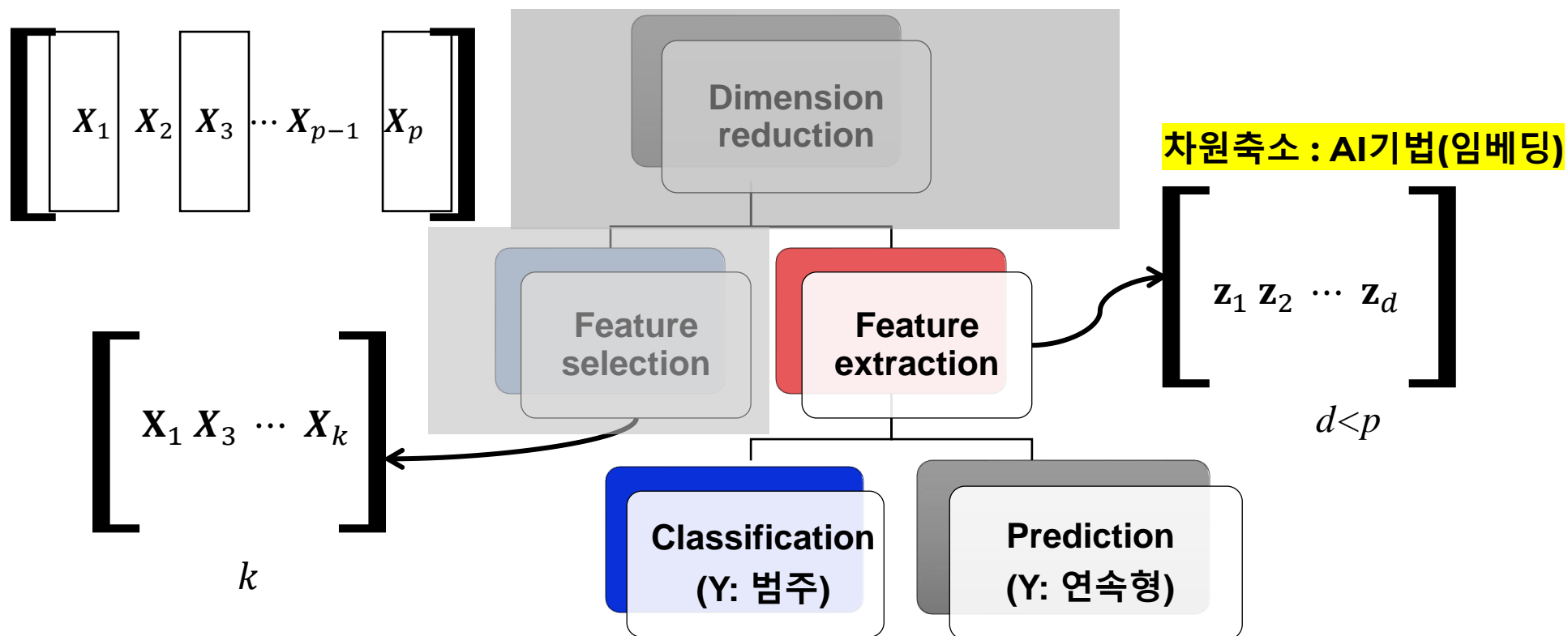


로그창

	strTime	nFPos	nldx	nTotal	nLeft	nTop	nRight	nBottom	nLabel	strClass
7	00:00:02	60	2	2	36	158	389	274	1	Smoke
8	00:00:02	65	1	4	285	180	354	311	0	Fire
9	00:00:02	65	2	4	691	265	878	501	0	Fire
10	00:00:02	65	3	4	33	160	274	269	1	Smoke
11	00:00:02	65	4	4	293	183	352	251	0	Fire
12	00:00:02	70	1	5	715	157	909	434	0	Fire

6. AI데이터 가공 사례

(2) AI 활용 데이터 가공 (임베딩을 포함한 딥러닝 기술 활용) 사례

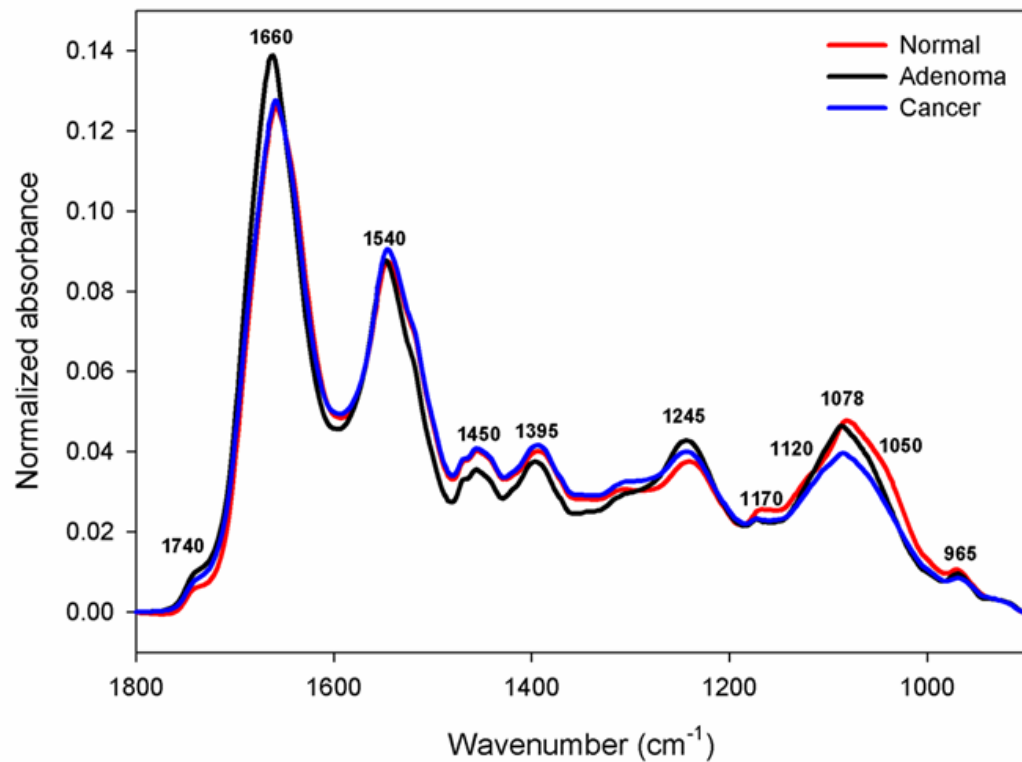


2. AI 데이터 분석단계

(2) AI 활용 데이터 가공 (임베딩을 포함한 딥러닝 기술 활용) 사례

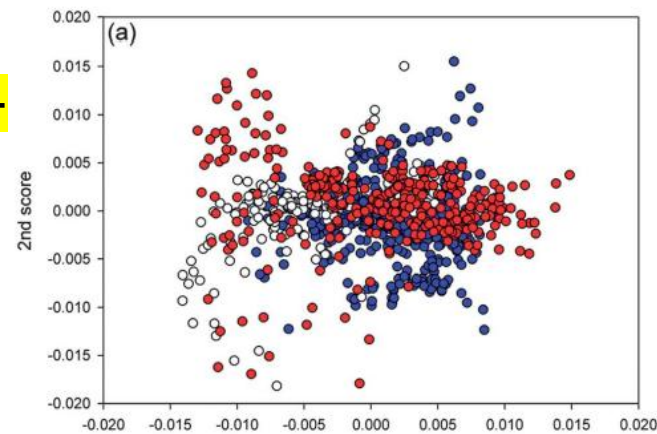
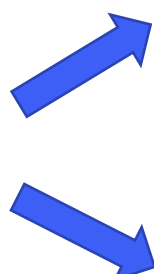
- 대장암(정상/전이단계/비정상) 분류

흡광도 데이터 (근적외선 데이터)

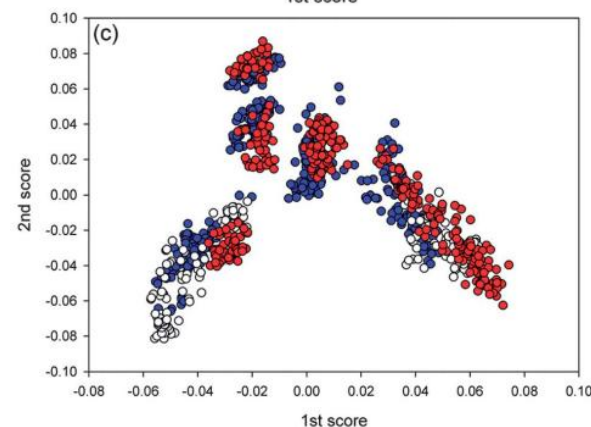


colon cancer(대장암)

AI기법활용
임베딩기법 적용



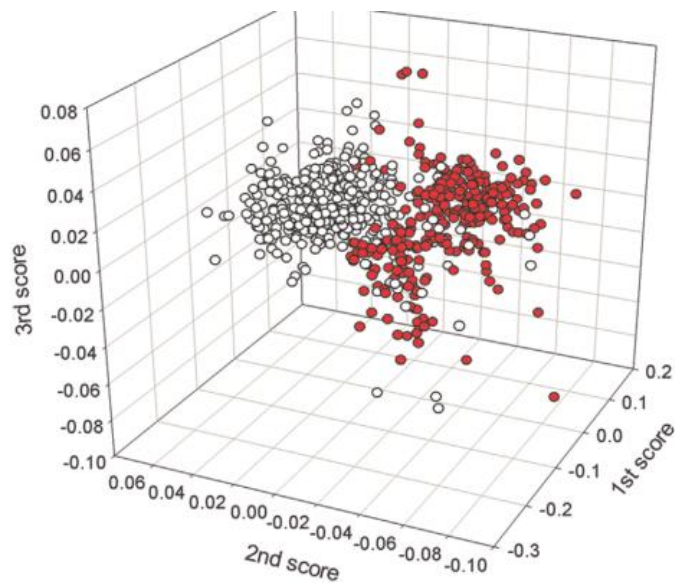
주성분분석



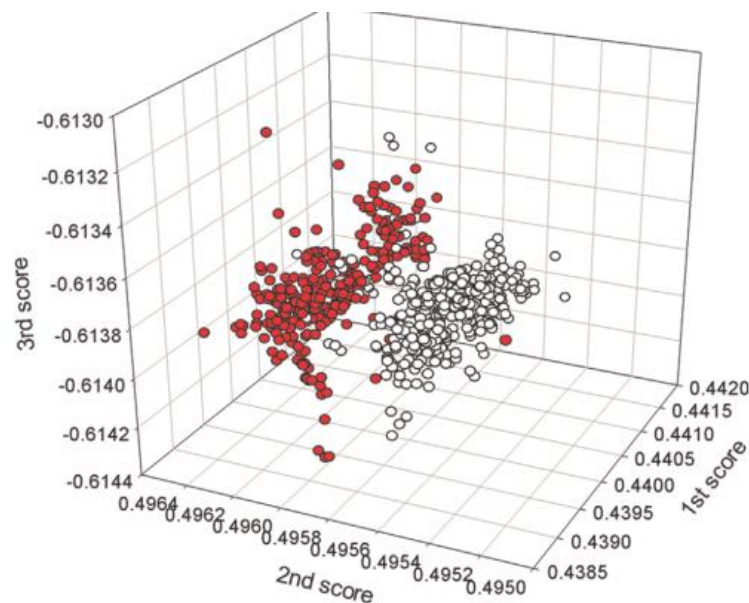
인접보존기법

2. AI 데이터 분석단계

(3) AI 활용 데이터 가공 - 농산물 원산지 분류



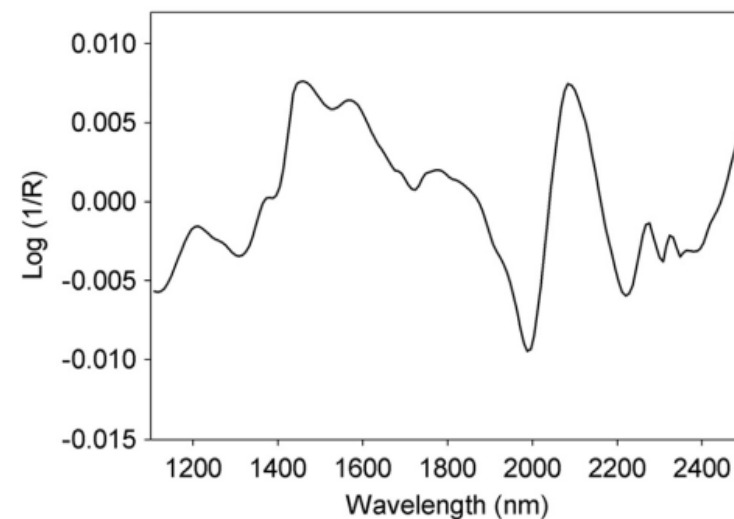
주성분분석 : 특징1, 특징2, 특징3



인접보전기법 : 특징1, 특징2, 특징3



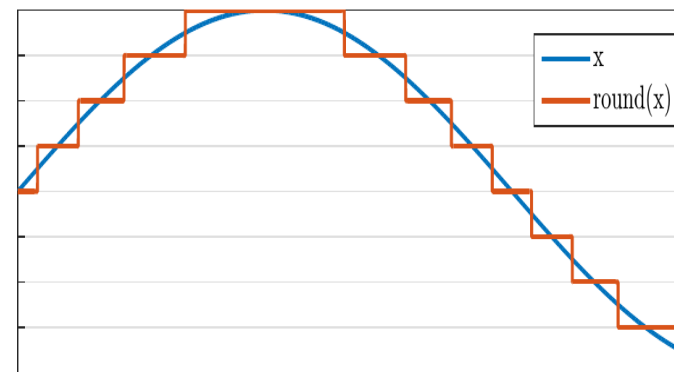
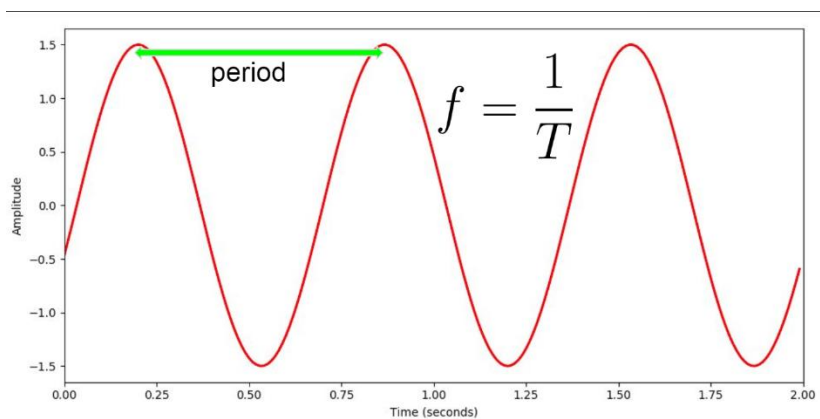
원래의 데이터차원



Ref: *Talanta*, 144, 960–968 (2015)

6. AI데이터 가공 사례

- 오디오 데이터 : Analog Digital Conversion을 거쳐 이산 벡터를 생성-> 딥러닝 모델 적용



<https://hyunlee103.tistory.com/54>

감사합니다