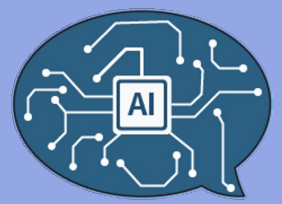
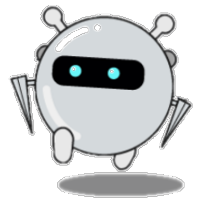


데이터 전처리 이론 1



진행자: 멘토 이창대

OUTTA

왜 데이터 전처리를 해야 할까?

컴퓨터가 데이터를 처리할 수 있도록 만들기!

OUTTA

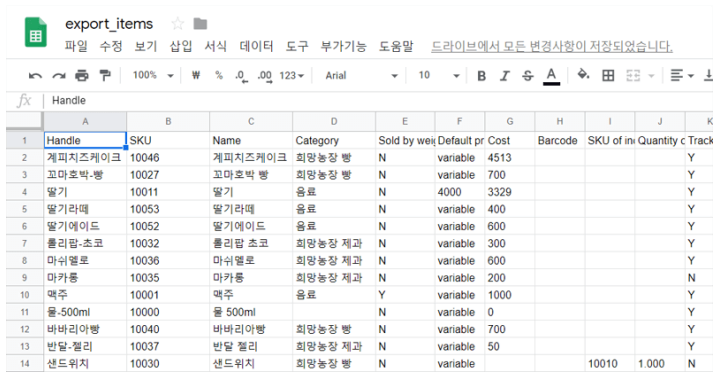
Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, I	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr	male		0	0	330877	8.4583		Q
7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mr	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercc	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, M	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, M	female	55	0	0	248706	16		S
17	0	3	Rice, Mast	male	2	4	1	382652	29.125		Q
18	1	2	Williams, M	male		0	0	244373	13		S
19	0	3	Vander Pla	female	31	1	0	345763	18		S
20	1	3	Masselmar	female		0	0	2649	7.225		C
21	0	2	Fynney, M	male	35	0	0	239865	26		S

목차

데이터 형태

정형 데이터

- 형태가 있음
- 연산이 가능
- 예) 엑셀, CSV



export_items

파일 수정 보기 삽입 서식 데이터 도구 부가기능 도움말 드라이브에서 모든 변경사항이 저장되었습니다.

Handle

	A	B	C	D	E	F	G	H	I	J	K
1	Handle	SKU	Name	Category	Sold by wei	Default pr	Cost	Barcode	SKU of in	Quantity c	Track
2	계피치즈케이크	10046	계피치즈케이크	희망농장 빵	N	variable	4513				Y
3	꼬마호박-빵	10027	꼬마호박 빵	희망농장 빵	N	variable	700				Y
4	딸기	10011	딸기	음료	N	4000	3329				Y
5	딸기라떼	10053	딸기라떼	음료	N	variable	400				Y
6	딸기에이드	10052	딸기에이드	음료	N	variable	600				Y
7	롤리팝-초코	10032	롤리팝 초코	희망농장 제과	N	variable	300				Y
8	마쉬멜로	10036	마쉬멜로	희망농장 제과	N	variable	600				Y
9	마카롱	10035	마카롱	희망농장 제과	N	variable	200				N
10	맥주	10001	맥주	음료	Y	variable	1000				Y
11	물-500ml	10000	물 500ml		N	variable	0				Y
12	바바리아빵	10040	바바리아빵	희망농장 빵	N	variable	700				Y
13	반달-젤리	10037	반달 젤리	희망농장 제과	N	variable	50				Y
14	샌드위치	10030	샌드위치	희망농장 빵	N	variable			10010	1.000	N

반정형 데이터

- 형태가 있음
- 연산이 불가능
- 예) HTML, XML

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML
<html>
    <head>
        <title>Example</title>
        <link href="screen.css" rel="sty
    </head>
    <body>
        <h1>
            <a href="/">Header</a>
        </h1>
        <ul id="nav">
            <li>
                <a href="one/">One</a>
            </li>
            <li>
                <a href="two/">Two</a>
            </li>
```

비정형 데이터

- 형태가 없음
- 연산이 불가능
- 예) 사진, 영상, 음성

OUTTA

데이터 형태 - 정형 데이터

Titanic dataset

CSV 형태

Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, I	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. '	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr	male		0	0	330877	8.4583		Q
7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mr	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercc	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, M	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, M	female	55	0	0	248706	16		S
17	0	3	Rice, Mast	male	2	4	1	382652	29.125		Q
18	1	2	Williams, M	male		0	0	244373	13		S
19	0	3	Vander Pla	female	31	1	0	345763	18		S
20	1	3	Masselmar	female		0	0	2649	7.225		C
21	0	2	Fynney, M	male	35	0	0	239865	26		S

데이터 형태 - 정형 데이터

Passenger	Survived	Pclass	Name	Sex	Age	SibSp
1	0	3	Braund, M	male	22	1
2	1	1	Cumings, I	female	38	1

승객 정보 | 생존 여부 | 좌석정보 | 이름 | 성별 | 나이 | 자녀수

Parch	Ticket	Fare	Cabin	Embarked
0	A/5 21171	7.25		S
0	PC 17599	71.2833	C85	C

함께 탑승한 가족 수 | 티켓 | 운임요금 | 좌석 | 출항항구

데이터 형태 - 반정형 데이터

온라인 박물관 정보

XML 형태

```xml
<?xml version="1.0" encoding='UTF-8' ?>
<!DOCTYPE DATA_RESULT [
  <!ELEMENT _ (DATA_RECORD*)>
  <!ELEMENT DATA_RECORD (SUBJECT?,HITS?,REGIST_DATE?,MEKE_DATE?,MOVIE_URL?,DATA_TYPE?,CATEGORY_NAME?)+>
  <!ELEMENT SUBJECT (#PCDATA)>
  <!ELEMENT HITS (#PCDATA)>
  <!ELEMENT REGIST_DATE (#PCDATA)>
  <!ELEMENT MEKE_DATE (#PCDATA)>
  <!ELEMENT MOVIE_URL (#PCDATA)>
  <!ELEMENT DATA_TYPE (#PCDATA)>
  <!ELEMENT CATEGORY_NAME (#PCDATA)>
]>
<DATA_RESULT>
  <DATA_RECORD>
    <SUBJECT>대한민국의 시간이 흐르는 곳, 대한민국역사박물관</SUBJECT>
    <HITS>748</HITS>
    <REGIST_DATE>2022-09-07</REGIST_DATE>
    <MEKE_DATE>2022-09-07</MEKE_DATE>
    <MOVIE_URL>http://www.much.go.kr/museum/onlinemuseum/detail.do?nttId=3412</MOVIE_URL>
    <DATA_TYPE>동영상</DATA_TYPE>
    <CATEGORY_NAME>박물관 홍보영상</CATEGORY_NAME>
  </DATA_RECORD>
  <DATA_RECORD>
    <SUBJECT>[Travel with Seoul-Mates!] Instagram-worthy spots in Seoul, South Korea</SUBJECT>
    <HITS>384</HITS>
    <REGIST_DATE>2022-08-31</REGIST_DATE>
    <MEKE_DATE>2022-08-31</MEKE_DATE>
    <MOVIE_URL>http://www.much.go.kr/museum/onlinemuseum/detail.do?nttId=3406</MOVIE_URL>
    <DATA_TYPE>유튜브</DATA_TYPE>
    <CATEGORY_NAME>박물관 홍보영상</CATEGORY_NAME>
  </DATA_RECORD>
  <DATA_RECORD>
    <SUBJECT>와... 이게 되네...? OO도 기증할 수 있다고?</SUBJECT>
    <HITS>200</HITS>
    <REGIST_DATE>2022-08-31</REGIST_DATE>
    <MEKE_DATE>2022-08-31</MEKE_DATE>
    <MOVIE_URL>http://www.much.go.kr/museum/onlinemuseum/detail.do?nttId=3405</MOVIE_URL>
```

OUTTA

데이터 형태 - 반정형 데이터

```
<DATA_RESULT>
  <DATA_RECORD>
    <SUBJECT>대한민국의 시간이 흐르는 곳, 대한민국역사박물관</SUBJECT>
    <HITS>748</HITS>
    <REGIST_DATE>2022-09-07</REGIST_DATE>
    <MEKE_DATE>2022-09-07</MEKE_DATE>
    <MOVIE_URL>http://www.much.go.kr/museum/onlinemuseum/detail.do?nttId=3412</MOVIE_URL>
    <DATA_TYPE>동영상</DATA_TYPE>
    <CATEGORY_NAME>박물관 홍보영상</CATEGORY_NAME>
```

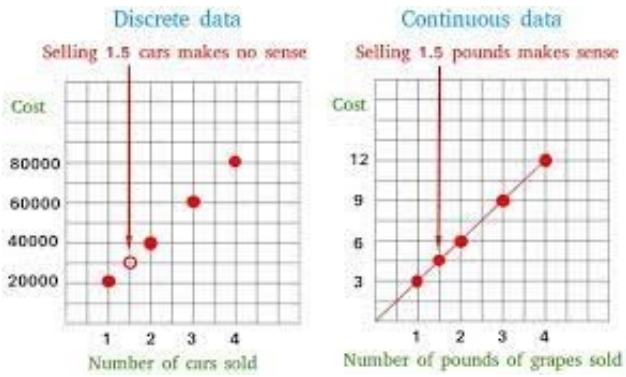
순번	한글명	영문명
1	제목	SUBJECT
2	조회수	HITS
3	등록일	REGIST_DATE
4	제작일	MEKE_DATE
5	영상경로	MOVIE_URL
6	자료유형	DATA_TYPE
7	카테고리명	CATEGORY_NAME

OUTTA

데이터 분류

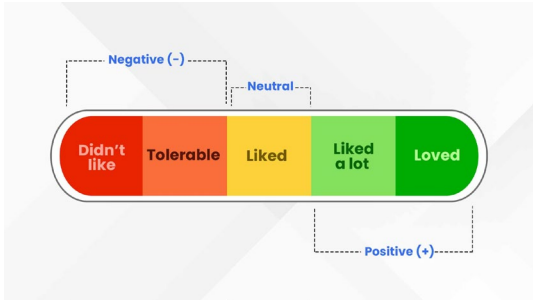
수치 데이터
(양적 데이터)

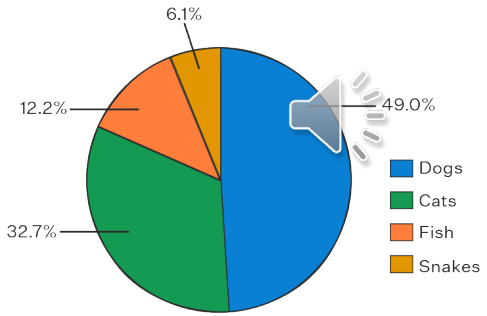
- 연속형 데이터
  예) 몸무게, 매출액
- 이산형 데이터
  예) 판매 수량, 나이



범주형 데이터
(질적 데이터)

- 순서형 데이터
  예) 설문지(좋다, 나쁘다, 매우 좋다), 학점(A, B, B+)
- 명목형 데이터
  예) 성별, 거주지

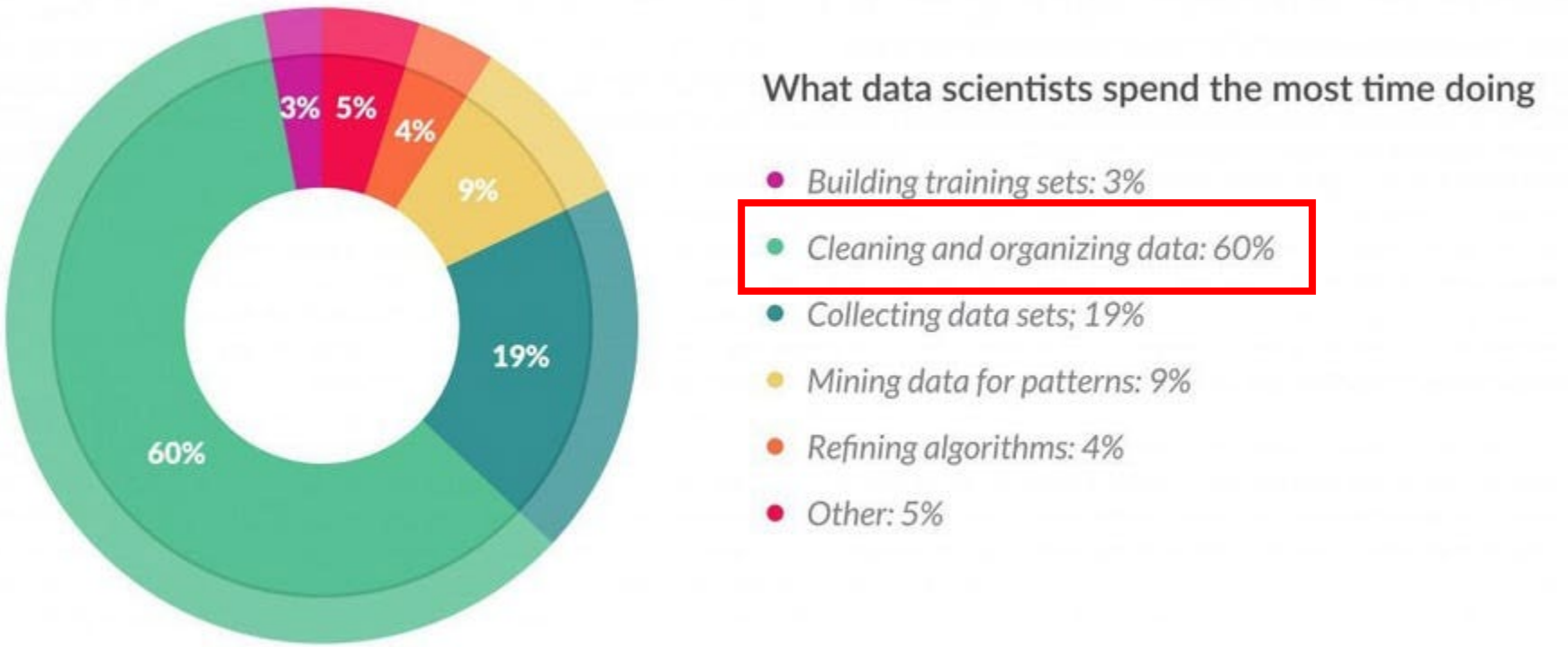




OUTTA

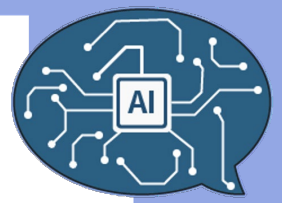
데이터 전처리 개요

대부분의 데이터 사이언티스트가 전처리에 시간을 많이 쏟는다고 응답

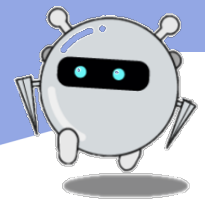


출처: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=123b5f906f63

OUTTA



감사합니다



OUTTA