Chapter. 01

시작에 앞서: 데이터 전처리는 왜 중요할까?

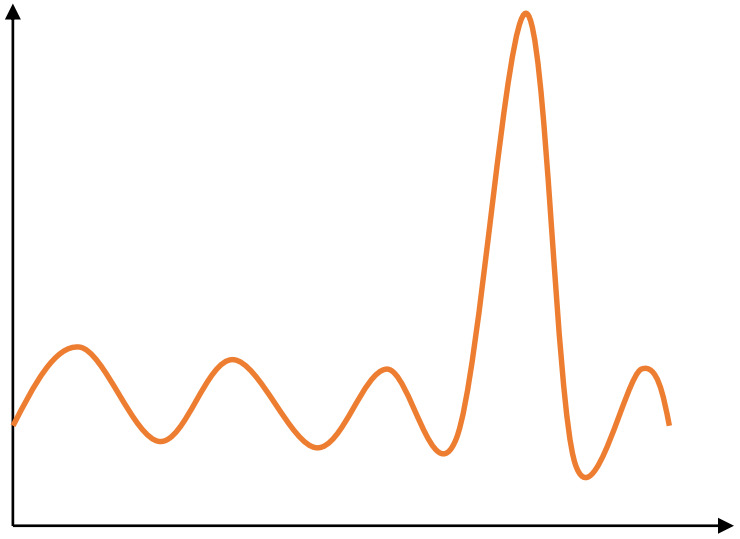
데이터 전처리의 중요성

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 현실 데이터

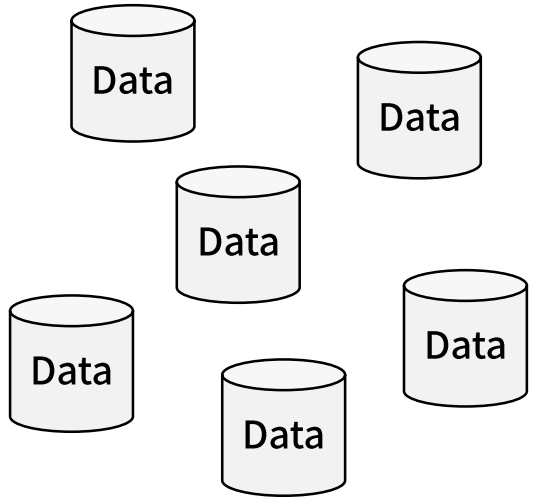
- 현실 데이터는 **분석 목적에 맞게 정리되어 있지 않아**, 데이터 분석 기법을 그대로 적용하기 어려움



노이즈가 포함된 데이터

X_1	X_2	X_3	X_4
NaN	0.7	1.2	3.5
0.2	5.2	NaN	4.1
2.6	5.8	NaN	4.4

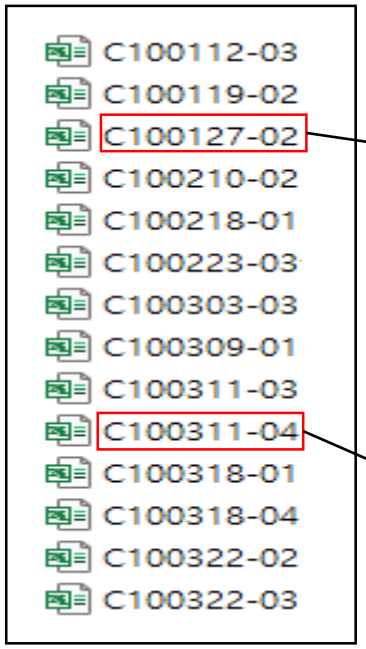
결측이 포함된 데이터



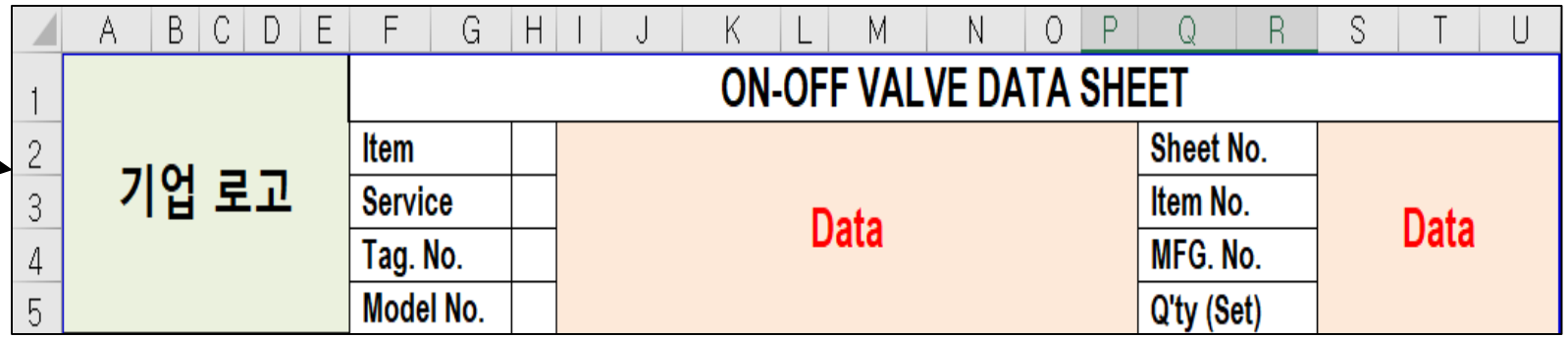
파편화된 데이터

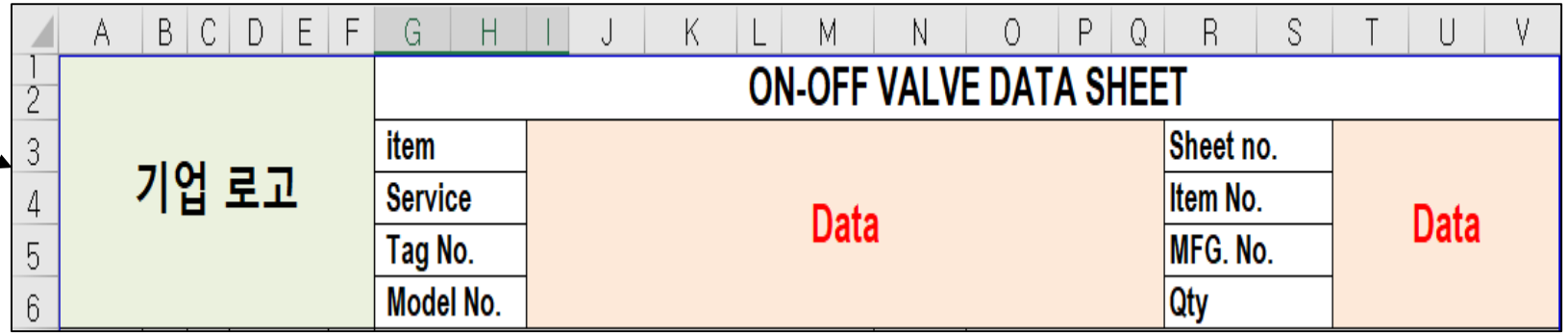
I 현실 데이터 사례: P 밸브사 – 수요 예측

- 2010년부터 2019년까지의 주문서 데이터를 바탕으로 월별 수요를 예측하는 프로젝트



약 5천개의 주문서 데이터
(파일명, 폴더 등이 정리되어 있지 않음)





포맷이 다른 주문서: 통일되지 않은 셀 위치와 제목 셀 이름

(예: Service의 위치, item – Item, Q'ty (Set) – Qty)

I 현실 데이터 사례: M 공업사 – 설비 비가동 시점 예측

- 제조 설비가 언제 비가동 상태로 접어들 것인지를 예측하는 프로젝트

이종 데이터의 혼합

16진수로 구성된 리스트

CollectedData	WorkshopId	EquipCode	OpCode	StatusCode	Version	dType	Model	Time	Code	Type	Data	SensorCount	rev	DataCnt	Id
오후 12:00:02	1730			0	20190404	SSS	AD	1.55E+12		EC	048"|"0048"|"0047"|"0046	1	1	100	1
오후 12:00:02	1730			0	20190404	SSS	AD	1.55E+12		TP	["0150"]	1	1		2
오후 12:00:02	1730			0	20190404	SSS	AD	1.55E+12		WL	["010E"]	1	1		3
오후 12:00:02	1730			0	20190404	CNC	AD	1.55E+12					1		
오후 12:00:02	1730			0	20190404	SSS	AD	1.55E+12		HL	5a2"|"0553"|"0105"|"02c	1	1	10	4
오후 12:00:02	1730			0	20190404	CNC	AD	1.55E+12					1		
오후 12:00:02	1730			0	20190404	CNC	AD	1.55E+12					1		
오후 12:00:03	1730			0	20190404	SSS	AD	1.55E+12		EC	048"|"0049"|"0049"|"004	1	1	100	1
오후 12:00:03	1730			0	20190404	SSS	AD	1.55E+12		TP	["014B"]	1	1		2
오후 12:00:03	1730			0	20190404	SSS	AD	1.55E+12		WL	["010E"]	1	1		3
오후 12:00:03	1730			0	20190404	SSS	AD	1.55E+12		HL	054b"|"05b0"|"062c"|"063	1	1	10	4
오후 12:00:02	1730			0	20190404	CNC	AD	1.55E+12					1		
오후 12:00:03	1730			0	20190404	CNC	AD	1.55E+12					1		
오후 12:00:03	1730			0	20190404	CNC	AD	1.55E+12	[]	opMsg			1		
오후 12:00:03	1730			0	20190404	CNC	AD	1.55E+12	[]	alarm			1		
오후 12:00:02	1730			0	20190404	CNC	AD	1.55E+12					1		
오후 12:00:02	1730			0	20190404	CNC	AD	1.55E+12					1		

비정상적인 시간 순서

기록되지 않은 알람

결측 다수 존재

I 현실 데이터 사례: K 홈쇼핑 –매출 예측을 통한 방송 편성 최적화

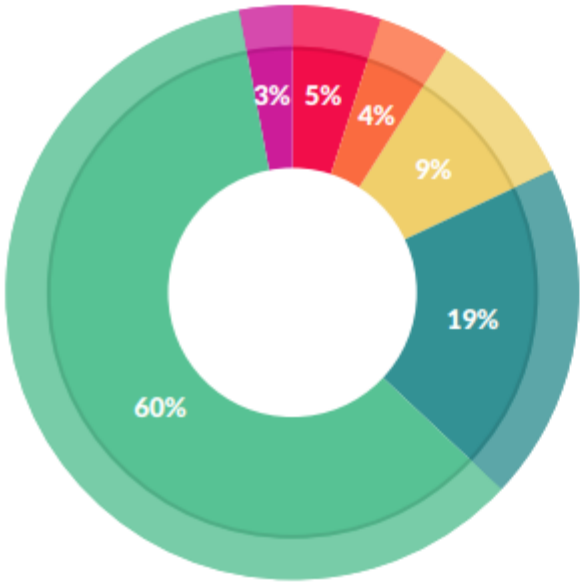
- 기존 홈쇼핑 방송 기록을 바탕으로 상품별 매출을 예측하고, 매출을 최대화하는 방향으로 방송을 편성하는 프로젝트

방송편성ID	방송일자	요일명	방송시간대역값	상품ID	상품명	품대분류코	품대분류	품중분류코	품중분류	품소분류코	품소분류	품세분류코	품세분류	성별구분코드	연령대역값	주문인원수	주문수량
1000066290	20171001	일	51	376546	드 노블클래식	20	패션잡화	5	명품	11	품악세사리	99	기타잡화	1	45	2	2
1000066290	20171001	일	51	376546	드 노블클래식	20	패션잡화	5	명품	11	품악세사리	99	기타잡화	1	50	3	3
1000066290	20171001	일	51	376546	드 노블클래식	20	패션잡화	5	명품	11	품악세사리	99	기타잡화	1	55	1	1
1000066290	20171001	일	51	376546	드 노블클래식	20	패션잡화	5	명품	11	품악세사리	99	기타잡화	1	65	1	1
1000066290	20171001	일	51	376546	드 노블클래식	20	패션잡화	5	명품	11	품악세사리	99	기타잡화	2	30	1	1
1000066290	20171001	일	51	376546	드 노블클래식	20	패션잡화	5	명품	11	품악세사리	99	기타잡화	2	35	2	2
1000066290	20171001	일	51	376546	드 노블클래식	20	패션잡화	5	명품	11	품악세사리	99	기타잡화	2	40	1	1
1000066290	20171001	일	51	376546	드 노블클래식	20	패션잡화	5	명품	11	품악세사리	99	기타잡화	2	45	1	1
1000066290	20171001	일	51	376546	드 노블클래식	20	패션잡화	5	명품	11	품악세사리	99	기타잡화	2	50	4	4
1000066290	20171001	일	51	376546	드 노블클래식	20	패션잡화	5	명품	11	품악세사리	99	기타잡화	2	55	4	4

- 불필요한 값이 너무 많아 데이터가 매우 크고, 프로젝트 목표에 부합하지 않는 구조의 데이터임

I 데이터 분석에 소요되는 시간

- 모든 데이터 분석 프로젝트에서 **데이터 전처리는 필수**적인 과정이며, 많은 분석가들이 데이터 전처리에 가장 많은 시간을 투입함



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

전체 분석 시간의 **79%**를 데이터 준비에 사용하므로,
전처리 역량이 분석 시간을 줄이는데 가장 중요

source: CrowdFlower survey, 2016

데이터 전처리의 주요 효과

- **(효율적인) 분석**을 가능하게 해준다.
- 불필요한 정보를 제거함으로써 **인사이트를 얻는데** 도움이 된다.
- 머신러닝 **모델의 성능을 향상**시킨다.

Chapter. 01

시작에 앞서: 데이터 전처리는 왜 중요할까?

데이터 전처리를 잘하는 방법

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 경험을 쌓아라

- 데이터 전처리 역량을 쌓는 가장 확실한 방법은 **좋은 경험**을 많이 쌓는 것이다
- 대다수의 경험은 “이런 방법을 쓰니까 시간만 날리더라. 다른 방법이 제일 좋았다”라는 **레퍼런스**가 됨
- 강사 사례: 학부 3학년 때의 뉴스 토픽 모델링 (크롤링한 뉴스에서 단어 추출 및 변환하기)

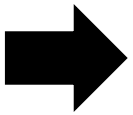
> **도쿄**에서는 **천여 명**을 **검사**했는데 **250명**이나 **확진 판정**을 받았습니다.

I 결과를 생각하라

- 전처리의 결과인 **전처리된 데이터의 구조**를 미리 생각해야 **불필요한 피드백 루프**를 막을 수 있음
- (예시) 구매 기록을 바탕으로 추천 시스템 구현하기

일자	회원 ID	구매 물품
2020.07.10	001	{A, B, C}
2020.07.10	002	{D}
2020.07.10	003	{A, B}
2020.07.11	001	{A, D, E}
2020.07.11	004	{B, C, D}
2020.07.11	005	{B, D}

원 데이터



회원 ID	A	B	C	D	E
001	2	1	1	1	1
002	0	0	0	1	0
003	1	1	0	0	0
004	0	1	1	1	0
005	0	1	0	1	0

전처리된 데이터

I 처리 과정을 생각하라

- 원 데이터를 결과 데이터로 바꾸기 위한 과정을 **단계별로 정의**해야 함

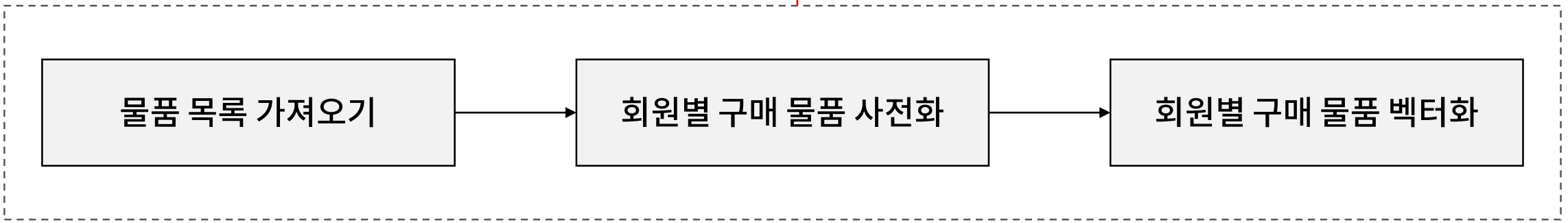
- (예시) 구매 기록을 바탕으로 추천 시스템 구현하기

일자	회원 ID	구매 물품
2020.07.10	001	{A, B, C}
2020.07.10	002	{D}
2020.07.10	003	{A, B}
2020.07.11	001	{A, D, E}
2020.07.11	004	{B, C, D}
2020.07.11	005	{B, D}

원 데이터

회원 ID	A	B	C	D	E
001	2	1	1	1	1
002	0	0	0	1	0
003	1	1	0	0	0
004	0	1	1	1	0
005	0	1	0	1	0

전처리된 데이터

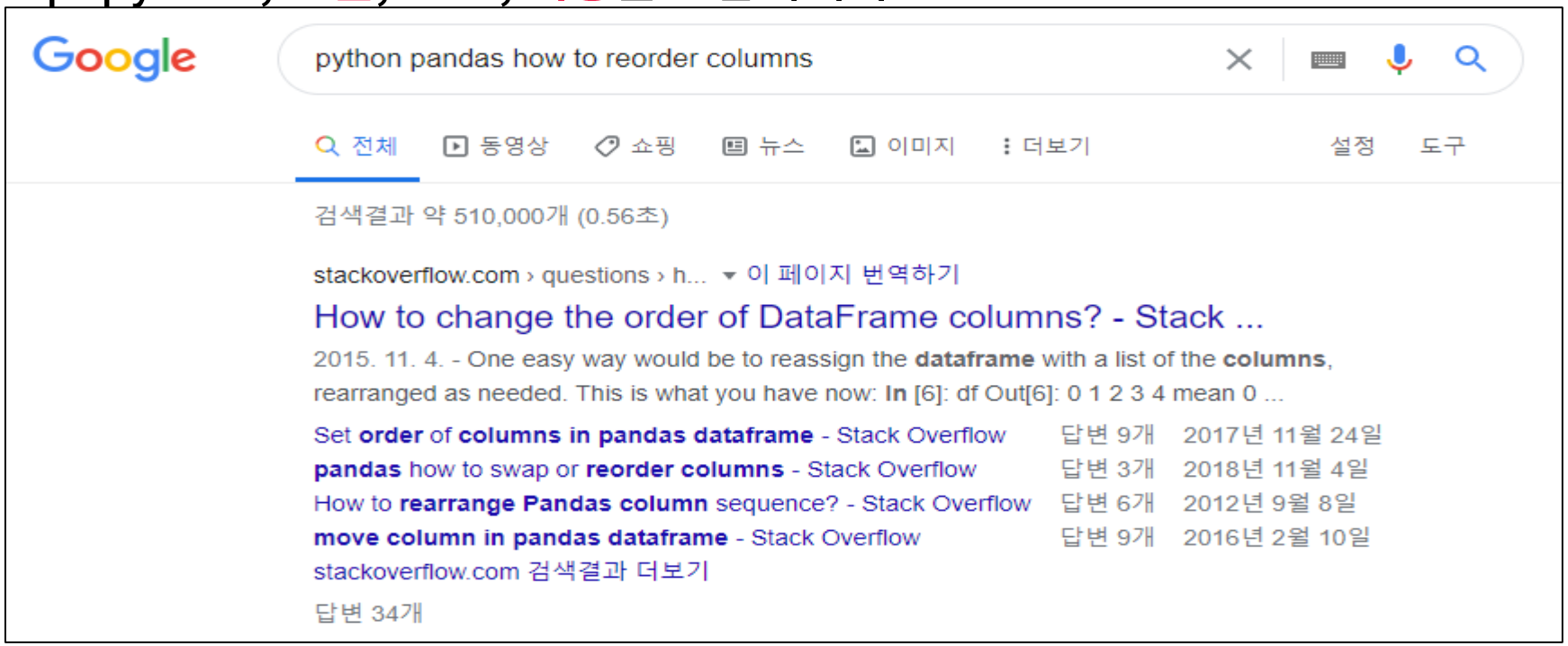


I 검색 능력을 쌓아라

- 본 강의에서 대표적인 데이터 전처리 문제와 해법을 다루지만, 데이터 전처리는 **데이터 by 데이터**가 매우 심하기 때문에 본 강의 내용을 그대로 적용하지 못하는 상황이 반드시 발생할 수 있음

- 따라서 본 강의를 통해 쌓은 데이터 전처리에 대한 기초적인 역량을 십분 발휘하기 위해서는, 필요한 전처리를 수행하는 방법을 검색하는 역량을 쌓아야 함 (feat. 구글님은 모든 것을 알고 계십니다)

- 검색 쿼리 Tip: python, **모듈**, **how**, **내용**을 포함시켜라



Chapter.

시작에 앞서: 데이터 전처리는 왜 중요할까?

데이터 전처리의 중요성

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승