EDA 101

Tidy Data Framework

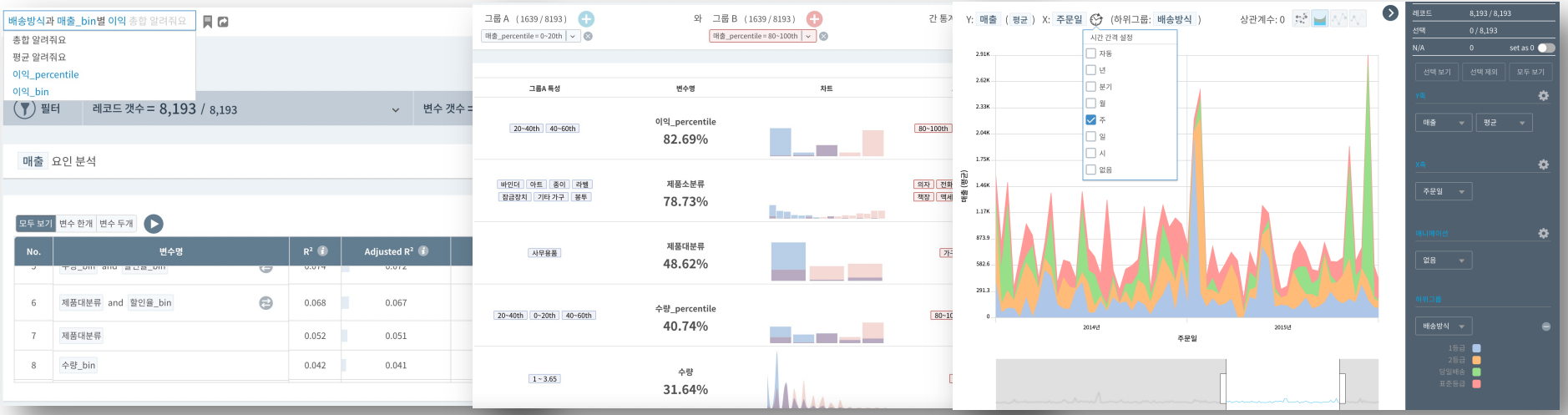
양 승 준 / sidney.yang@idk2.co.kr

다정한 데이터 도구, HEARTCOUNT

- HEARTCOUNT(하트카운트)는 비전문가도 쉽게 엑셀 데이터셋을 업로드하여 시각화하고 분석할 수 있는 SaaS 솔루션입니다.
- Google 계정만 있다면, 홈페이지에서 바로 사용을 시작할 수 있어요!
- 지금 시작하기 : https://www.heartcount.io/

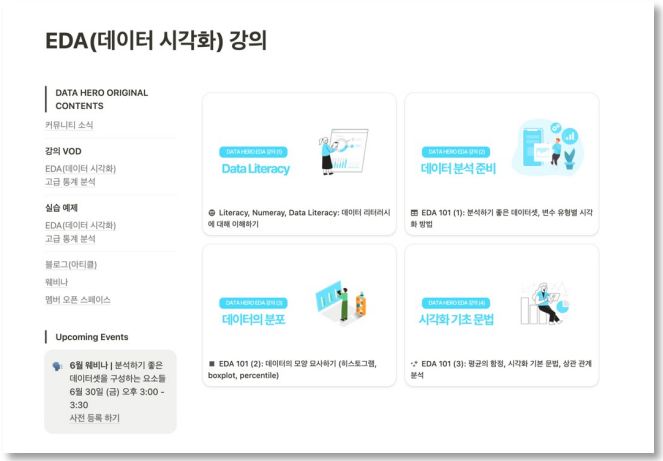
다정한 데이터 도구, HEARTCOUNT

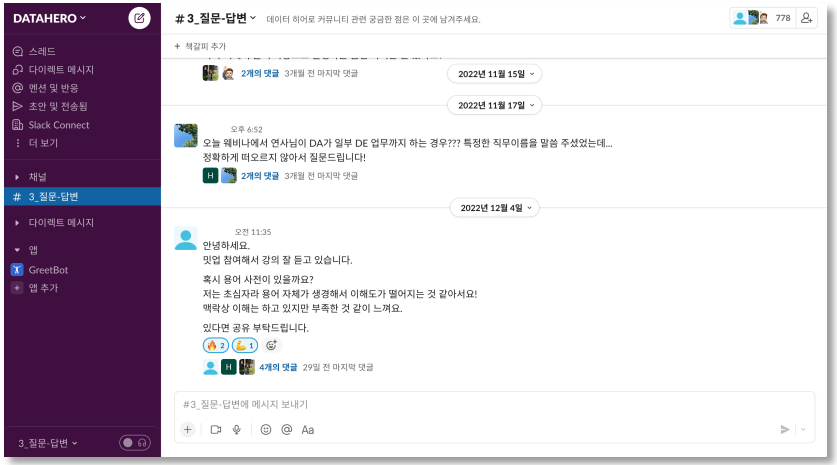
- 특장점으로는 '개별 레코드 수준의 시각화, '파생 변수 자동 생성', '패턴 자동 발견', '자연어 검색 + 설명' 등이 있습니다.



학습하고 소통하는 공간, DATA HERO

- 데이터의 기초부터 실전까지, 전용 페이지에서 무료로 학습 가능
- 하트카운트팀은 물론 다양한 실무자들과의 소통 공간
- 다양한 집중 교육 캠프, 오프라인 밋업 등 이벤트





오늘 다룰 내용들

1. 분석하기 좋은 데이터셋(Analytics-Ready/Tidy Dataset)의 구조와 모양
2. wide 와 long 형식의 데이터셋과 둘 간의 변환
3. 데이터셋을 구성하는 변수(칼럼) 유형, 변수 유형 변경하기, 유형에 따른 EDA
4. 주어진 데이터셋으로 알 수 있는 것(패턴)과 알 수 없는 것

What is Tidy Dataset?

http://vita.had.co.nz/papers/tidy-data.pdf

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

Tidy Dataset vs. Messy Dataset

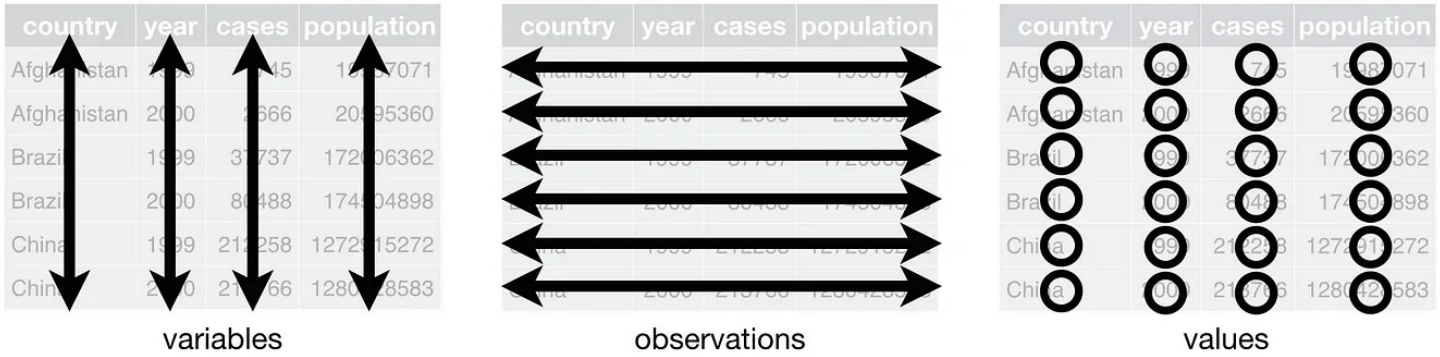
- "Happy families are all alike; every unhappy family is unhappy in its own way" - Leo Tolstoy

- "Tidy(Clean) data are all alike; every messy dataset is messy in its own way" – Hadley Wickham

What is Tidy Dataset?

Tidy datasets(분석하기 좋은 데이터셋)은 데이터의 물리적 구조와 그 의미를 연결하는 표준을 제시해 줌.

- 구조: 데이터의 형식과 모양; 대부분의 데이터셋은 사각형의 모양을 하고 있으며 행(row)과 열(columns)로 구성됨
- 의미(Semantics): 데이터셋은 숫자(quantitative) 또는 문자열(qualitative)로 표현된 값(Value)의 집합으로 다음 두가지에 속하게 됨:
  - Variables: 모든 분석 단위(unit)에 대해 측정한 동일한 속성값 (키, 온도, 매출 등)
  - Observations: 동일 분석 단위(unit)에 대해 측정한 모든 측정값들의 집합



Structure

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Table 1: Typical presentation dataset.

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Table 2: The same data as in Table 1 but structured differently.

대부분의 데이터셋은 행과 열로
구성된
사각형의 테이블 구조를 하고 있음

Structure and Semantics

사람, 트리트먼트 종류, 결과 이렇게 세개의
변수(Variable)를 사람과 트리트먼트 종류를 관측의
단위(key, unit)로 해서 측정한 결과를 잘 담은 데이터셋

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Table 1: Typical presentation dataset.

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

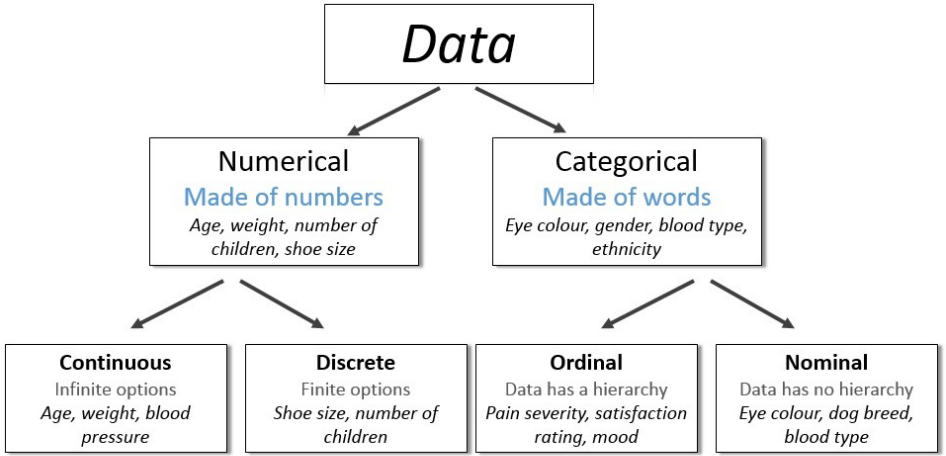
Table 2: The same data as in Table 1 but structured differently.

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Data Type (변수의 종류)

숫자형(Quantitative)과 범주형(Qualitative)

분석이란 숫자와 숫자 사이의 연관성,
범주 간 숫자의 차이를 이해하는 것



- 숫자형 자료는 이산형(discrete)이나 연속형(continuous)으로 나뉨
- 범주형 자료는 명목형(nominal)이나 순서형(ordinal)으로 나뉨

Structure and Semantics

- 서로 다른 treatment 효과 사이의 상관관계를 알고 싶다면?
- 서로 다른 treatment 간 효과의 차이를 알고 싶다면?

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

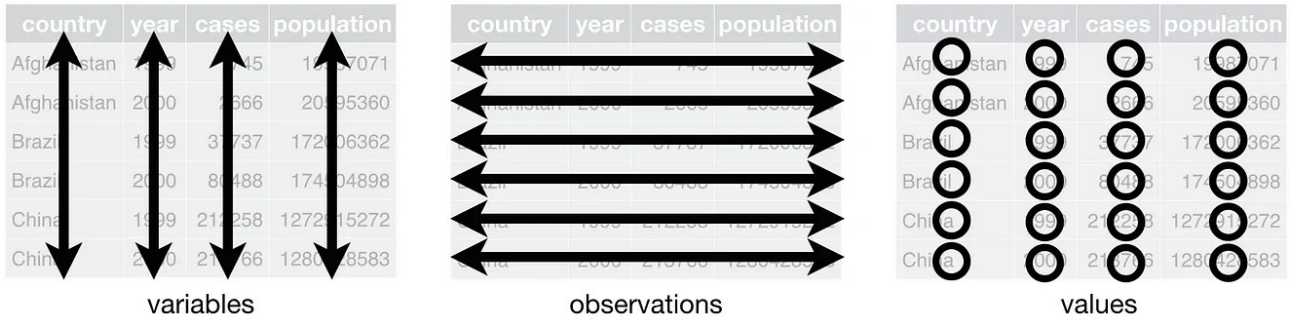
Table 1: Typical presentation dataset.

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Tidy Dataset

데이터셋의 의미와 구조를 잘 연결한 것

- Variable: 동일한 속성(나이, 매출)에 대한 측정값들로 행을 구성
- Observation: 분석의 단위(사람, 사건, 매출 등)에 대한 측정값들로 열을 구성



X
features
independent variables
input (variables)
predictor
attribute

Y
target/label
dependent variables
output (variable)
response

record
sample
Instance
case

Tidy Dataset?

- 국가별로 1999년과 2000년에 결핵으로 사망한 환자수(Cases)와 전체인구 (Population)를 정리한 데이터셋들
- 국가별 연도별 인구 10,000명당 결핵 사망률을 계산하기 가장 좋은 데이터는?

1

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

2

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

3

country	year	population
Afghanistan	1999	745 / 19987071
Afghanistan	2000	2666 / 20595360
Brazil	1999	37737 / 172006362
Brazil	2000	80488 / 174504898
China	1999	212258 / 1272915272
China	2000	213766 / 1280428583

4

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

EDA (Descriptive Data Analysis)

데이터에 대해 사실적으로 묘사하는 법

Description
요약

변수의 대표값과
모양이 어떻나?

개별 변수(Y)의
통계값과 분포 확인

Comparison
비교

변수값의 차이가
어디서 얼마나
나나?

서로 다른 범주(X)
간 Y의 특성·모양
비교

Relationship
관계

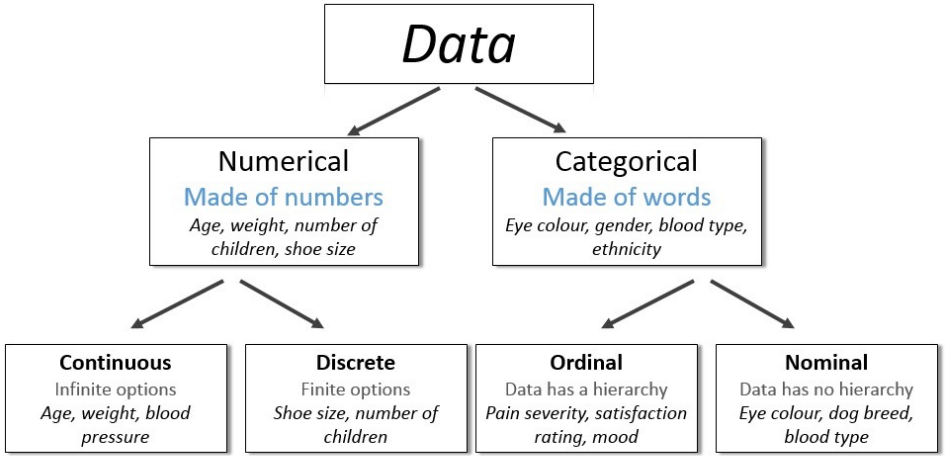
두 숫자 변수(X, Y)
간 관계가 어떻나?

X와 Y 사이의
상관관계 파악

Data Type (변수의 종류)

숫자형(Quantitative)과 범주형(Qualitative)

분석이란 숫자와 숫자 사이의 연관성,
숫자의 차이를 가져오는 범주를 발견하는 것



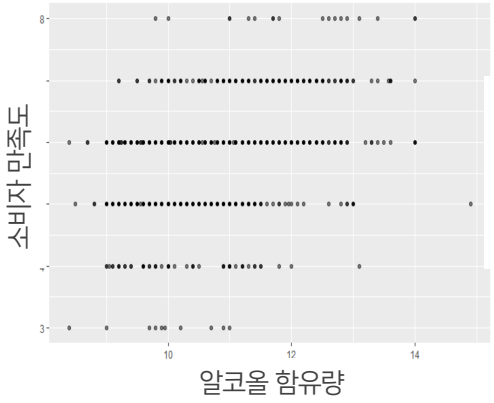
- 숫자형 자료는 이산형(discrete)이나 연속형(continuous)으로 나뉨
- 범주형 자료는 명목형(nominal)이나 순서형(ordinal)으로 나뉨

Data Type에 따른 시각화 방법

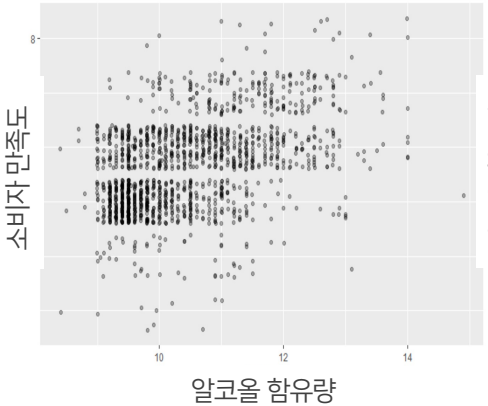
변수 유형에 따라 분석 방법과 효과적 시각화 방법이 달라짐

Alcohol(%): 와인 알코올 함량, Quality: 소비자가 매긴 점수

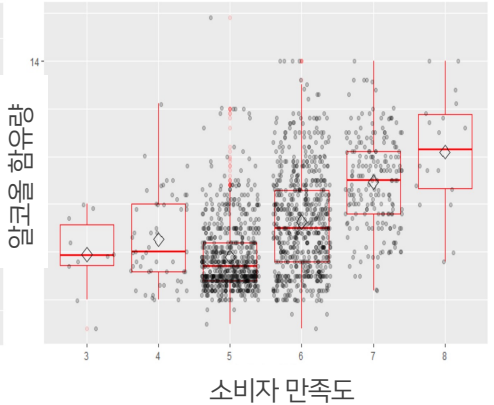
숫자 x 숫자 = Scatterplot
Overplotting (점이 겹침)!



Jittering 기법으로 Noise 추가
Jittering (인위적으로 퍼지게)!



소비자 만족도를 범주로 처리
Boxplotting (분포 시각화)!



순서형(Ordinal) 변수는 범주(Category)로 다루는 게 좋다!

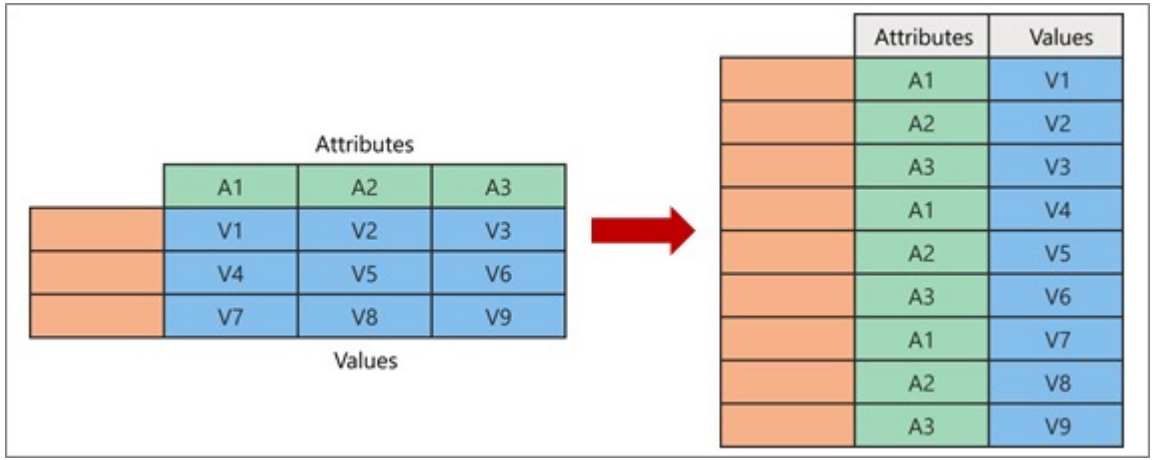
From wide to long (melting)

칼럼 제목이 변수가 아니라 변수값임

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

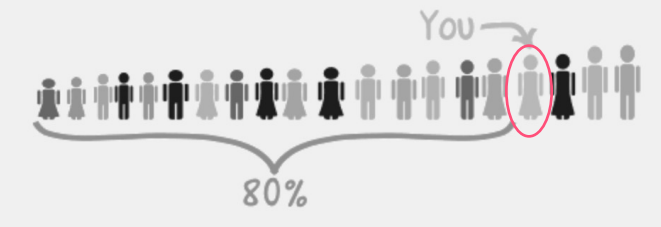
From wide to long



- Spreadsheet: Unpivot
  - https://docs.google.com/spreadsheets/d/1mhovD7d_7lh0rljU7Zl_ZkDzlHa8zq6XTkTEG63VpMg/edit#gid=0
- Python/R: Melting
  - https://chat.openai.com/share/cc83ddb9-5985-4b94-b100-a4bf8453ae2e

숫자에서 범주로: Percentile

Percentile: 전체 관측값들의 분포를 고려했을 때 특정값의 상대적 위치



내 키가 185cm로 20명 중
네번째로 키가 크다면

185cm = 80th Percentile

내 밑으로 80%가 있다!

Score [정렬된 점수]	Percentile Rank	Quartile [사분위]
29	8th	Q1, 1사분위 (최하위 25%)
32	17th	
38	25th	
41	33th	Q2, 2사분위 (차하위 25%)
53	42th	
54	50th	
55	58th	Q3, 3사분위 (차상위 25%)
74	67th	
93	75th	
99	83th	Q4, 4사분위 (최상위 25%)
134	92th	
209	100th	

숫자에서 범주로: Percentile

Who Are Our Best Customer?

- RFM: Recency, Frequency, and Monetary Value
- Decile: 10분위수 (1: 최하위 10%; 10: 최상위 10%)

RECENCY
The *freshness* of customer activity.

FREQUENCY
The *frequency* of customer transactions.

MONETARY
The *willingness* to spend.

	Frequency	monetary_value	Recency
CustomerID			
12346.0	1	77183.60	325
12747.0	103	4196.01	2
12748.0	4596	33719.73	0
12749.0	199	4090.88	3
12820.0	59	942.34	3

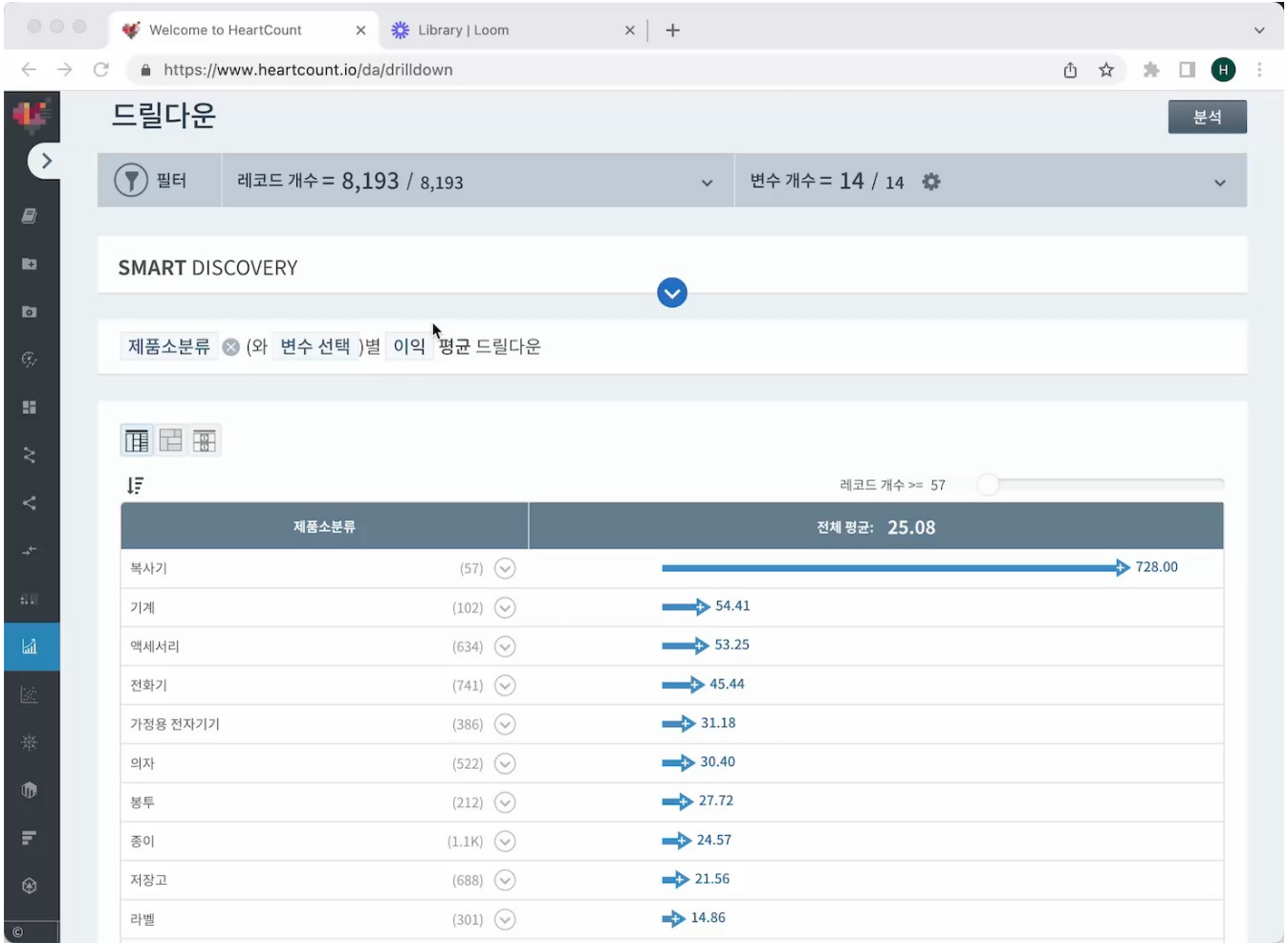
CustomerID	Recency Decile	Frequency Decile	Monetary Decile
727783	1	1	1
729689	1	1	1
834275	1	1	2
215474	1	2	1
911756	1	2	2
...			
671990	9	10	10
579843	10	9	10
562266	10	10	10

숫자에서 범주로: Percentile

HEARTCOUNT

고(저)성과 레코드가 높은 집단 찾기

- 이익 숫자 ➔ 이익 Percentile (범주) 변환



데이터셋으로 알 수 있는 것 (Insight)

데이터의 넓이(사실)와 경험의 깊이(견해)

데이터의 넓이

- 패턴: 데이터셋에 담긴 단어와 숫자로 만들 수 있는 최선의 문장
- 22~23시, TV 채널로 주문한 40~44세 여성의 전자제품 취소율이 40%로 높았다.

경험의 깊이

- 왜 취소율이 높았나? 어떻게? ➔ 해석과 판단력의 영역

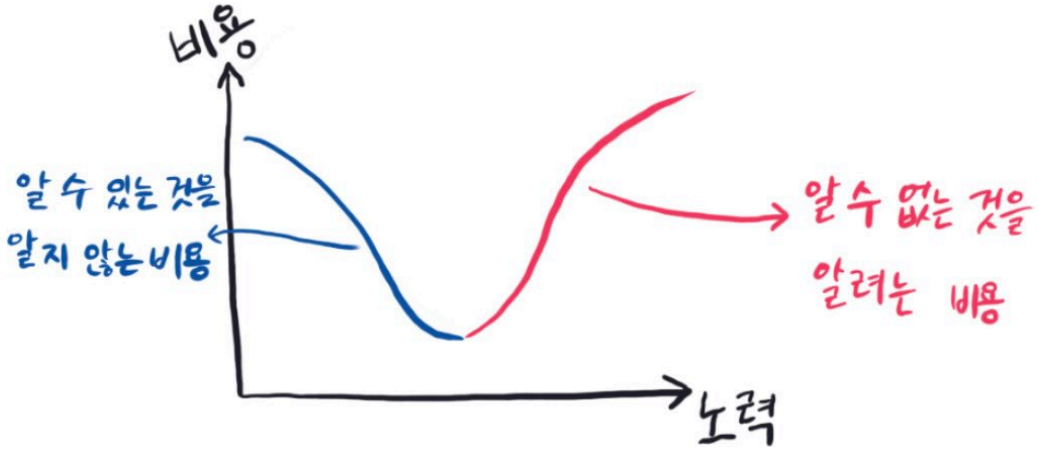
홈쇼핑 주문내역 취소율 데이터셋

범주							지표			
주문일	주문 시간대	주문 채널	연령대	성별	상품분류	이벤트 유형	순주문 금액	순주문 수량	취소율	취소 금액
2023-5-7	22시	TV	40세~44세	여자	전자제품	상품쿠폰	350000	23	38%	0
…										

데이터셋으로 알 수 있는 것 (Insight)

질문에 대한 정량적 답변 빠르게 구하기

- 알 수 없는 걸 알려고 하지 않기
- 완벽한 정보를 가지고 의사결정 못 함



실습 시간

EDA 도구: https://www.heartcount.io/login