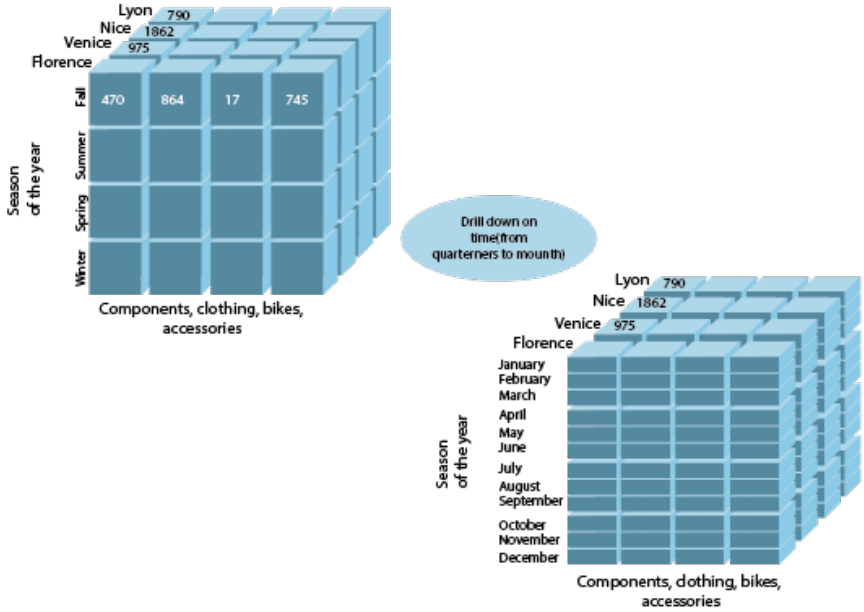
EDA 101 Series 3

Drill Down

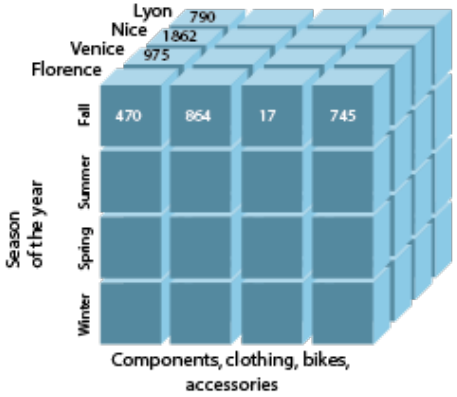
양 승 준 / sidney.yang@idk2.co.kr

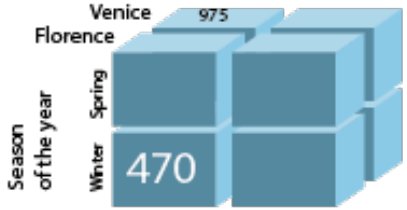
Definition: Drill-Down

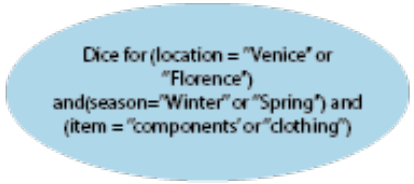
- "드릴다운"이란 상위 수준의 요약 정보를 하위 그룹으로 쪼개어 보다 세부적인 수준에서 지표(Metrics)를 들여다 보는 일
- 예) 계절별 매출 총액(Metrics)을 보다가, 월별 수준에서 보다 촘촘히 데이터를 보는 작업

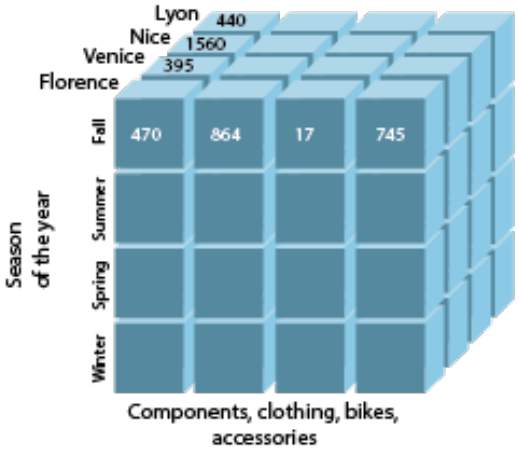


Slicing and Dicing









Simpson's Paradox

전체 지원자 합격률

	지원자 수	합격자 수	합격률
여자	1,000	150	15%
남자	1,000	250	25%

문과대 합격률

	지원자 수	합격자 수	합격률
여자	800	80	10%
남자	200	10	5%

이공대 합격률

	지원자 수	합격자 수	합격률
여자	200	70	35%
남자	800	240	30%

Simpson's Paradox

전체 지원자 합격률

	지원자 수	합격자 수	합격률
여자	1,000	150	15%
남자	1,000	250	25%

문과대 합격률

	지원자	합격자	합격률	합격률
여자	800	80	10%	9%
남자	200	10	5%	

이공대 합격률

	지원자	합격자	합격률	합격률
여자	200	70	35%	31%
남자	800	240	30%	

뭉뚱그린 수치는 현실을 왜곡할 수 있음
쪼개보는 일(Drill-Down by Diverse Dimensions)의 중요성

A, Ad-hoc Analysis

Ad-hoc Analysis
그때 그때의 질문에 대해 데이터로 답하는 일

A, Ad-hoc Analysis

Ad-hoc Analysis
그때 그때의 질문에 대해 데이터로 답하는 일

질문: 대시보드를 통해 답할 수 없는 모든 질문들

- 쉬운 질문, 몇개/얼마? ➔ Counting
- **어려운 질문, 왜? ➔ EDA**/고급분석

인사이트

- 패턴(사실) + 견해(해석)를 Narrative 형식으로 보고

A, Ad-hoc Analysis

어려운 질문, 왜? ➔ EDA/고급 분석

Q. 지표가 왜 변했나요?

➔ EDA: 범주 간 지표 차이 이해

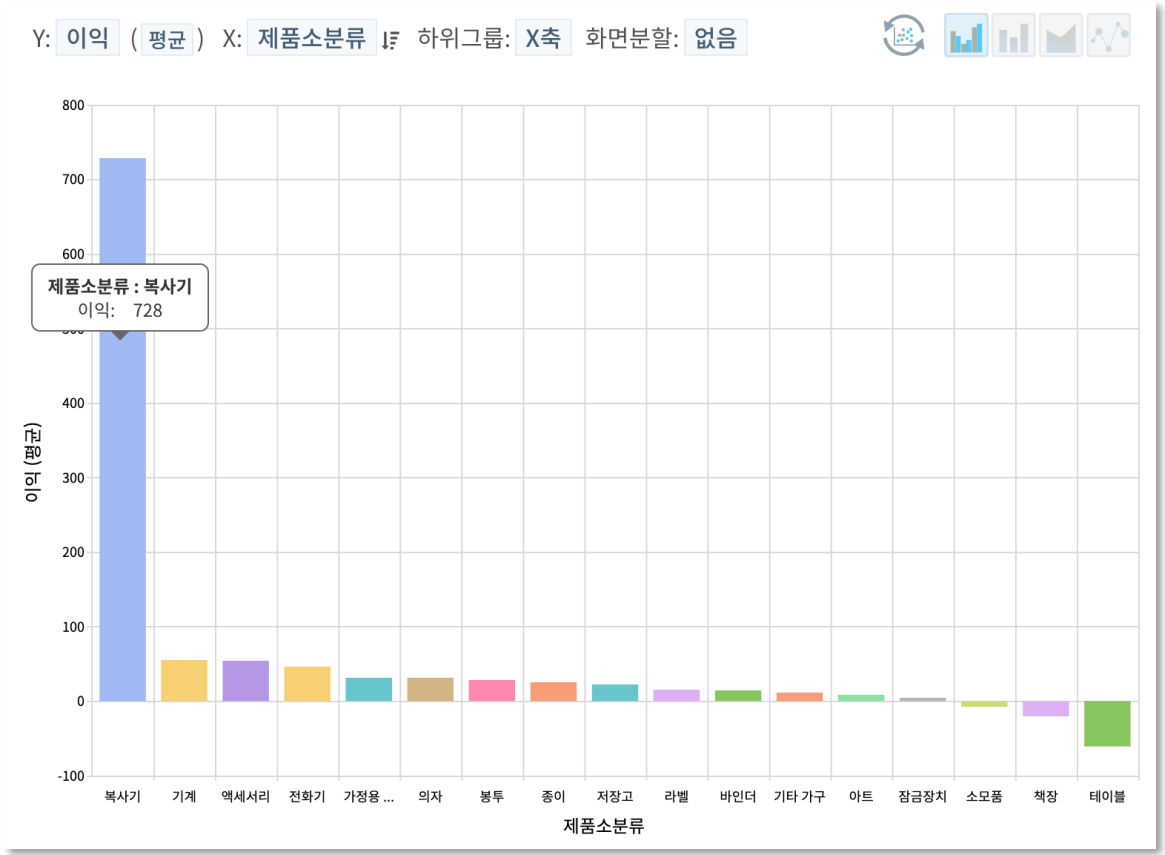
➔ EDA: 차이를 최적화할 (통제가능한) 대상/요인 찾기

A, Ad-hoc Analysis – “몇개?”말고 “왜?” 질문

범주 간 **지표 차이** 이해

➔

차이를 최적화할 통제가능한 대상과 요인 찾기



F, Feature Engineering – 숫자 변수 가공

	나이	나이_bin	나이_percentile
1	20	20~23	~20th (하위 20%) 3개의 레코드
2	24	24~27	
3	25	24~27	
4	29	28~31	~40th 3개의 레코드
5	33	31~34	
6	33	31~34	
7	39	38~41	~60th 3개의 레코드
8	40 (중앙값)	38~41	
9	41	38~41	
10	42	42~45	~80th 3개의 레코드
11	43	42~45	
12	43	42~45	
13	44	42~45	~100th (상위 20%) 3개의 레코드
14	51	50~53	
15	60	58~60	

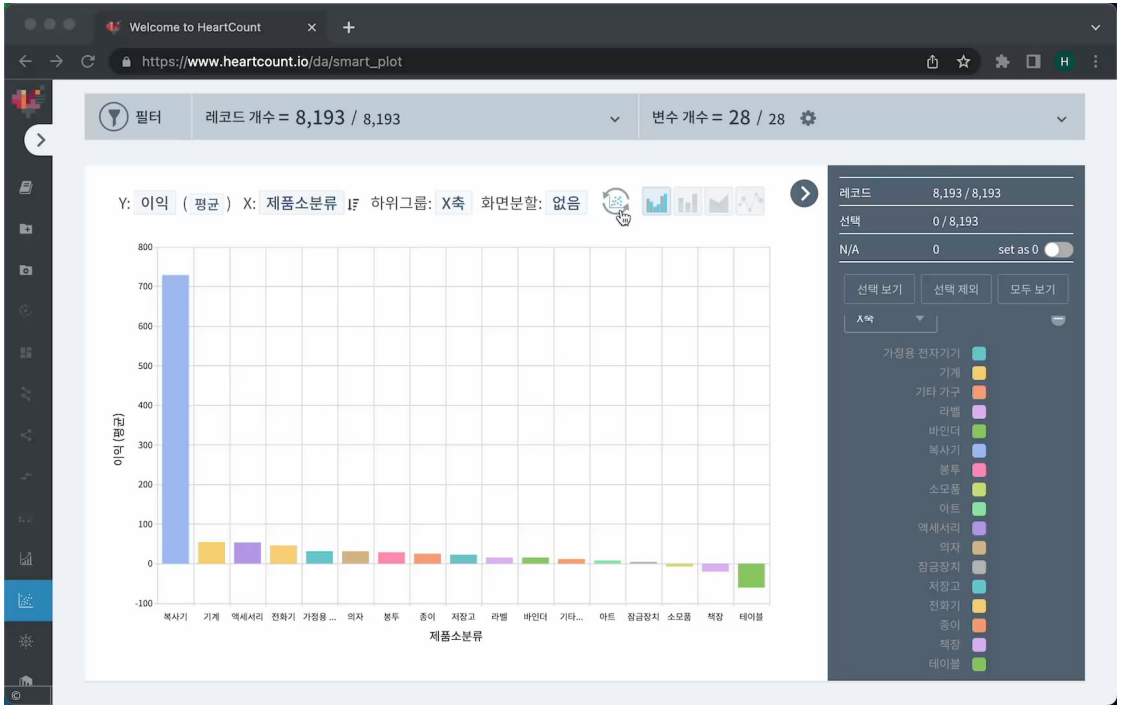
F, Feature Engineering – 숫자 변수 가공 (binning)

HEARTCOUNT

차이를 최적화할
통제가능한 대상과
요인 찾기

- 할인률 숫자 ➔
  할인률 구간(범주)
  변환

동영상 링크
https://www.loom.com/share/35ab326241344ac3a3f235e60bc1a61b



A, Ad-hoc Analysis – 빠르게 양질의 의사결정

Time-to-Insight: 그때 그때 질문에 바로 잘 답할 수 있으려면

- Curated Dataset for Ad-hoc Analysis
- KPI와 관련된 알 수 있는 모든 변수들을 모아 놓은 넓은(wide) 데이터셋

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
2023-5-7	16시	모바일	20세~24세	남자	육류	미상	0	0	12%	55963
2023-5-7	14시	PC	40세~44세	여자	주방용품	미상	123591	15	7%	0
2023-5-7	17시	모바일	20세~24세	여자	스킨용품	상품쿠폰	75827	13	12.5%	0
2023-5-7	13시	TV	40세~44세	여자	가구	미상	75509	4	7.4%	75500

A, Ad-hoc Analysis – 넓은 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
…										

실습 시간

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