

실무자를 위한 데이터 사전

from A to H

하트카운트 양승준

데이터 사전 A-H



A: Ad-hoc Analysis

B: BI (The end of BI)

C: ChatGPT and Data

D: Data Newsletter

E: Everyone is an Analyst

F: Feature Engineering

G: Granularity

H: HEARTCOUNT

- ④ **A:** Ad-hoc Analysis - 데이터에서 질문-답을 찾는 경험이 바뀌어야 한다.
- ② **B:** BI (The end of BI) - BI로 대표되는 데이터 도구가 재발명돼야 한다.
- ③ **C:** ChatGPT and Data - 챗GPT 같은 기술도 도움이 되겠지만
- ⑦ **D:** Data Newsletter - 데이터가 소비자에게 먼저 찾아가는 일
- ① **E:** Everyone is an Analyst - 모든 실무자들이 분석가가 되려면
- ⑤ **F:** Feature Engineering - 패턴 발견에 도움이 되게 변수 가공하는 일
- ⑥ **G:** Granularity - 집계 데이터 대신 해상도가 높은 데이터를 다루는 일
- ⑧ **H:** HEARTCOUNT - 다정한 데이터 도구 하트카운트가 함께 할게요!

A: Ad-hoc Analysis

A: Ad-hoc Analysis

B: BI (The end of BI)

C: ChatGPT and Data

D: Data Newsletter

E: Everyone is an Analyst

F: Feature Engineering

G: Granularity

H: HEARTCOUNT

A, Ad-hoc Analysis

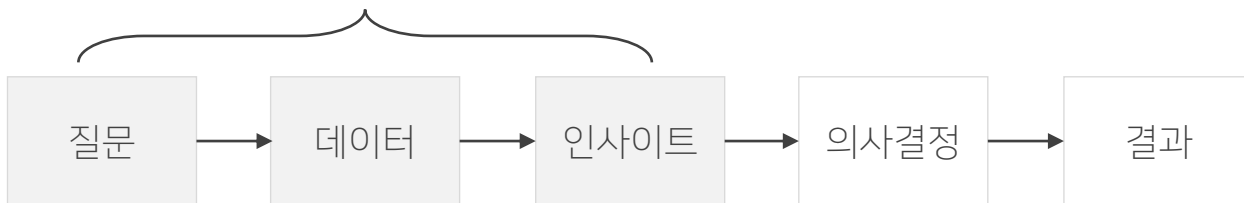
Ad-hoc Analysis
그때 그때의 질문에 대해 데이터로 답하는 일



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

어려운 질문, 왜? → EDA/고급 분석

Q. 지표가 왜 변했나요?

→ EDA: 범주 간 지표 차이 이해

→ EDA: 차이를 최적화할 통제/타게팅 할 수 있는 대상/요인 찾기

A, Ad-hoc Analysis

EDA(Exploratory Data Analysis) =
DESCRIBE (기술 분석) + EXPLORE (탐험 분석)



- 데이터 구조/모양
- 데이터 품질
- 요약
- 시각화
- 가설 생성/검증
- != 모델링

범주 간 지표 차이 이해



보르헤스, 중국 백과사전 동물 분류

- 황제에게 속하는 동물
- 향료로 처리하여 방부 보존된 동물
- 길잃은 개
- 인어
- 미세한 낙타털 붓으로 그릴 수 있는 동물
- 멀리서 볼 때 파리같이 보이는 동물

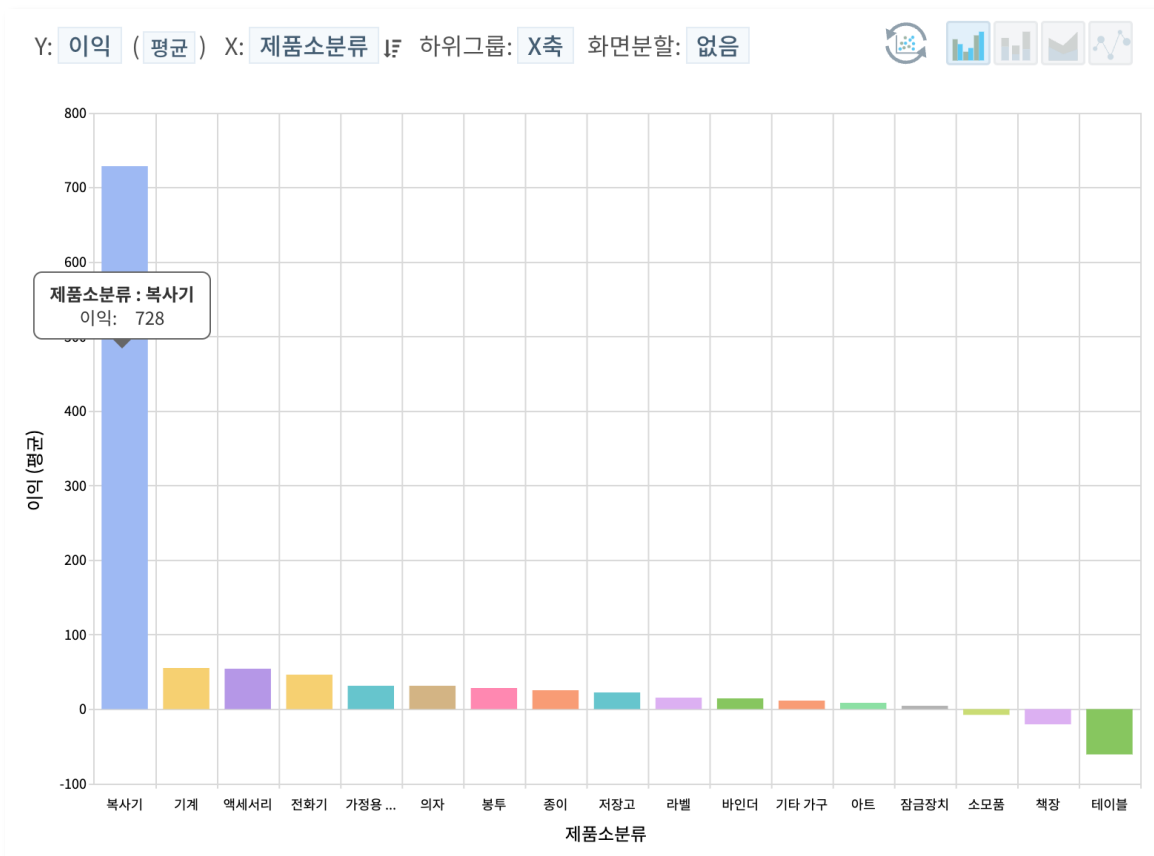
차이에 대해 말하려면 비교 대상이 존재해야 하고
범주는 개별 레코드를 관습 또는 필요에 따라 분류하는 기준

A, Ad-hoc Analysis - "몇개?" 말고 "왜?" 질문

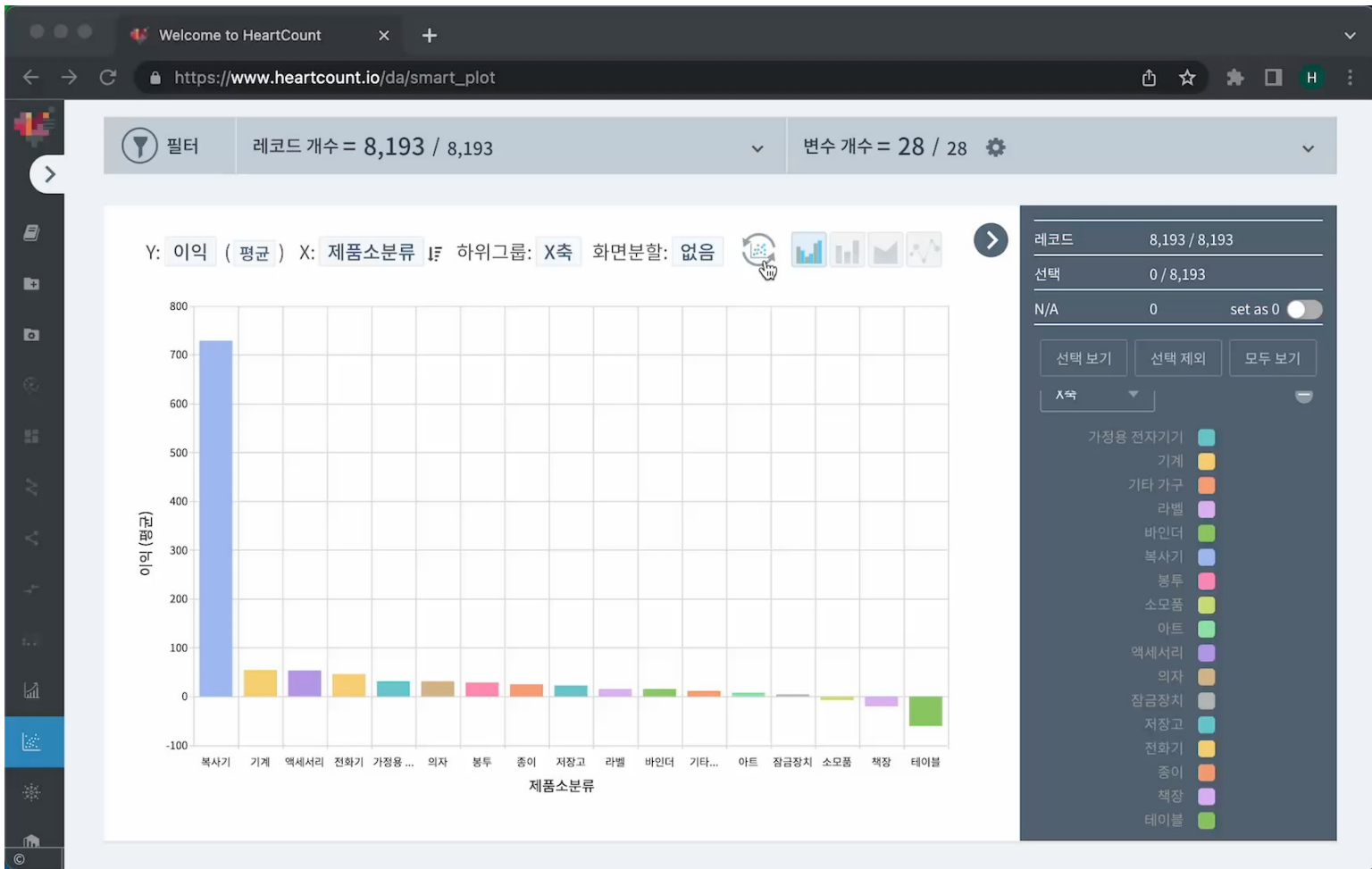
범주 간 지표 차이 이해



차이를 최적화할 통제가능한
대상과 요인 찾기



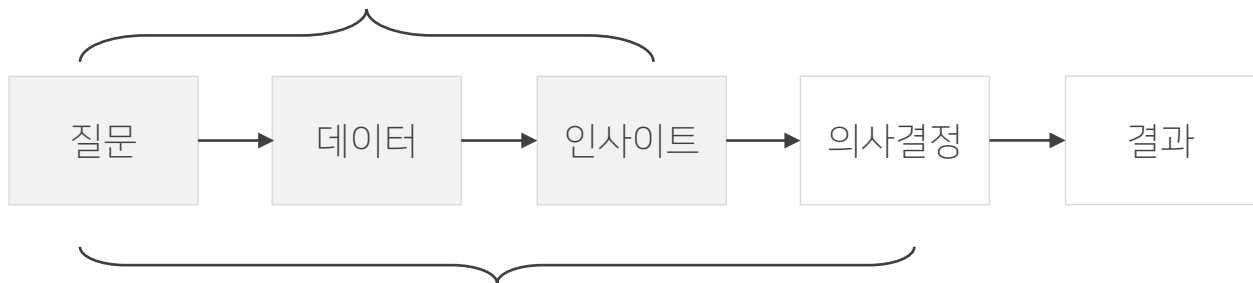
범주 간 지표 차이 이해 → 차이를 최적화할 통제가능한 대상과 요인 찾기



A, Ad-hoc Analysis

Ad-hoc Analysis

그때 그때의 질문에 대해 데이터로 답하는 일



Ad-hoc Analysis is to Support Ad-hoc Decision

A, Ad-hoc Analysis is to Support Ad-hoc Decision

Ad-hoc Decision: One-Way-Door vs. Two-Way-Door Decision

Two-Way-Door



One-Way-Door

- 되돌릴 수 없는 결정
- 최대한 천천히 신중하게

Two-Way-Door

- 결과에 따라 되돌릴 수 있는 결정
- 빠르게 양질의 결정; 반복 실험
- 일어나지 않은 일은 무엇이든 될 수 있다

A, Ad-hoc Analysis – 빠르게 양질의 의사결정

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

- Curated(/Governed) 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
...										

B: BI (The end of BI)

A: Ad-hoc Analysis

B: BI (The end of BI)

C: ChatGPT and Data

D: Data Newsletter

E: Everyone is an Analyst

F: Feature Engineering

G: Granularity

H: HEARTCOUNT

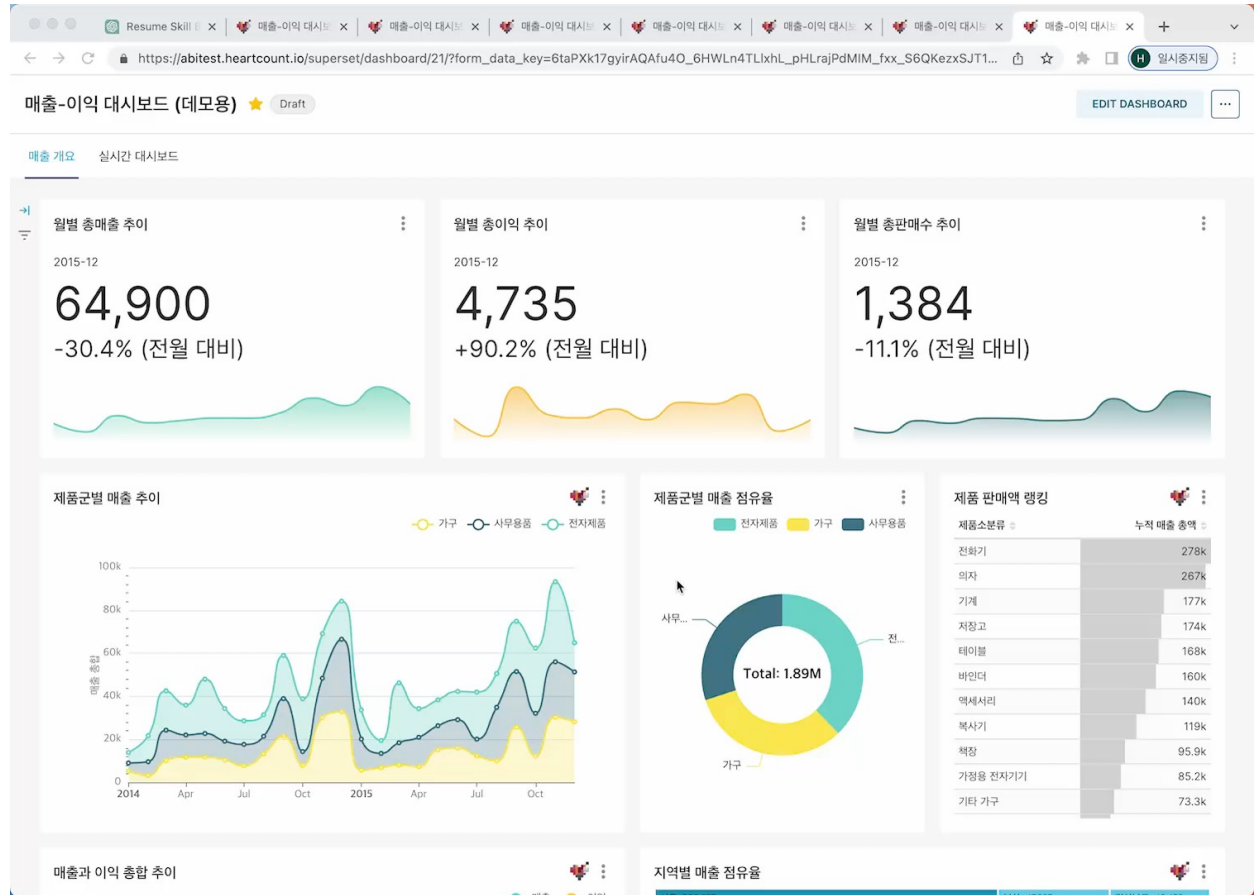
B, (The End of) Business Intelligence

대시보드 - FAQ

- 이번달 매출 얼마야?
- 제품군별 ranking?

추가 질문

- 지역별 x 제품군별 ranking?
- 수많은 질문들



Comic Book Guy's Channel x +

http://www.mytube.com/comicbookguychannel

GROUNDHOG DAY



Self-Serve BI

→ 대시보드를 똑딱 O

→ 질문에 대한 답 똑딱 X

Groundhog Day

→ 데이터 팀, ad-hoc 질문에
답하기 위해 어제도 오늘도
데이터 추출

B, (The End of) Business Intelligence

지표와 관련된 질문에 똑딱 답할 수 있으려면,

- 타인(데이터 팀)의 친절함과 노력에 의지하지 않고
- 질문이 생기면 넓은 데이터셋으로 바로 EDA 할 수 있어야 함

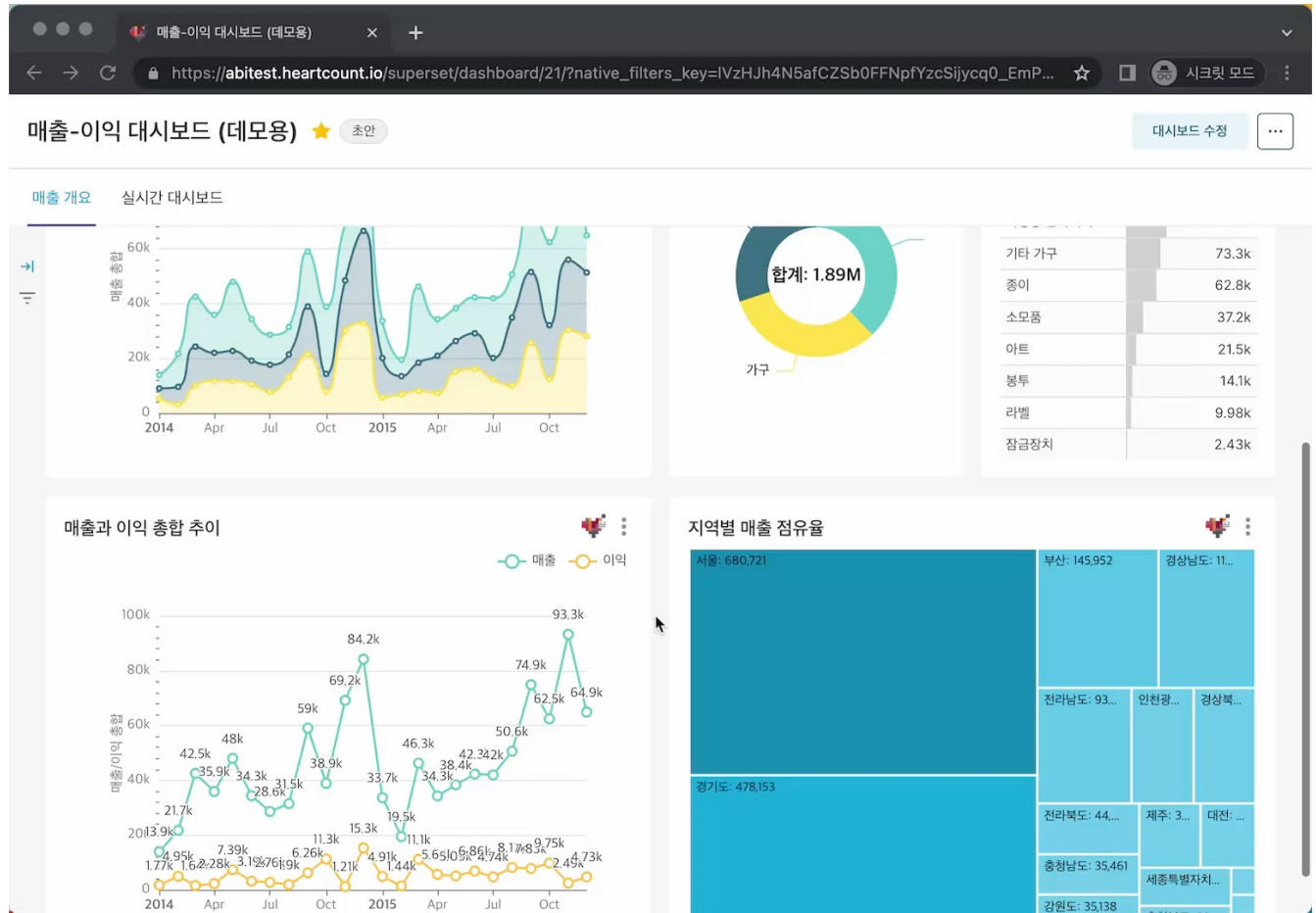




HEARTCOUNT

지역 x 제품군 제품 ranking

- 맥락담긴 넓은 데이터셋으로
그때 그때 질문에 똑딱 답변



C: ChatGPT and Data

A: Ad-hoc Analysis

B: BI (The end of BI)

C: ChatGPT and Data

D: Data Newsletter

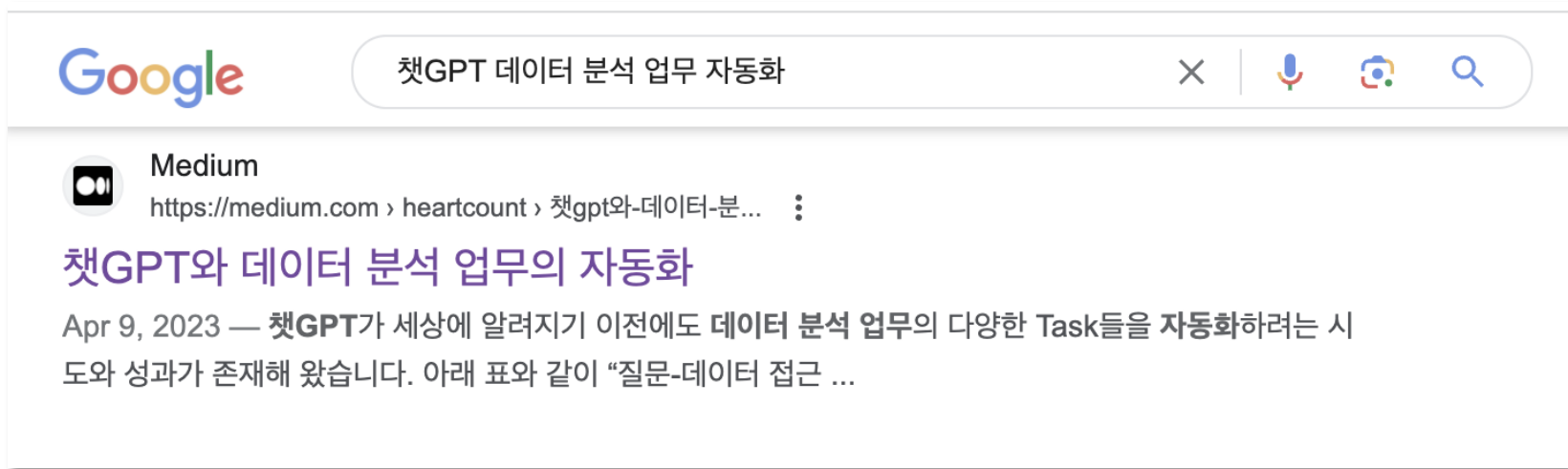
E: Everyone is Analyst

F: Feature Engineering

G: Granularity

H: HEARTCOUNT









C, ChatGPT - 데이터 분석 업무 자동화



The image shows a Google search interface. The search bar contains the text "챗GPT 데이터 분석 업무 자동화". Below the search bar, a search result is displayed for a Medium article. The article title is "챗GPT와 데이터 분석 업무의 자동화", dated "Apr 9, 2023". The snippet of the article text reads: "챗GPT가 세상에 알려지기 이전에도 데이터 분석 업무의 다양한 Task들을 자동화하려는 시도와 성과가 존재해 왔습니다. 아래 표와 같이 "질문-데이터 접근 ...".

C, ChatGPT - 데이터 분석/활용 업무 자동화

People Problem vs. Technical Problem



	질문	데이터 접근	분석 (인사이트 찾기)	보고서 작성
사람	 <p>지난 달 매출이 왜 갑자기 늘었죠?</p>	 <p>데이터 추출 - 월별 매출 - 웹 트래픽 데이터 편성</p>	 <p>계절적 패턴? 유입경로별? 제품/고객별?</p>	 <p>인사이트/견해로 구성된 보고서</p>
챗GPT		 <p>SQL 쿼리 작성</p>	 <p>Descriptive Analytics</p>	

Data Analyst's Block 질문이 없어 힘든 적 있나요?



H # 다음과 같은 칼럼으로 구성된 데이터셋이 있을 때, 분석 주제에 대한 아이디어를 줄 수 있겠니?
채용경로 성별 나이 근속기간 출신대학 대학구분 학점 학력 입사년도 직군 팀구분 퇴직
구분 매니저여부 잔업시간 성과점수 미사용.휴가일수 연봉 사업장위치 본인.인지능력 본인.인내
력 본인.실행력 매니저.비전 매니저.소통 매니저.실행력 직원만족도



이 데이터셋을 기반으로 다양한 분석 주제를 제안할 수 있습니다. 아래에 몇 가지 예시를 제시해 드리겠습니 다.  

1. 직원들의 성별, 나이, 근속기간, 학력 등과 연봉 간의 상관관계 분석
2. 다양한 직군과 팀 구분에 따른 직원들의 성과점수와 직원 만족도 분석
3. 입사년도별 채용 경로(공채, 산학 등)와 퇴직 여부 간의 관계 분석
4. 사업장 위치에 따른 직원들의 잔업시간, 미사용 휴가일수, 연봉 등 비교 분석
5. 매니저의 비전, 소통, 실행력과 직원들의 직원만족도, 성과점수 간의 상관관계 분석
6. 본인의 인지능력, 인내력, 실행력과 성과점수, 직원만족도 간의 상관관계 분석
7. 대학 구분(국립, 사립, 해외)과 학점에 따른 직원들의 성과점수, 연봉, 직원만족도 비교 분석

이러한 주제들을 통해 직원들의 특성, 성과, 만족도 등에 영향을 미치는 요인들을 파악하고, 이를 바탕으로 인사 관리, 교육 및 훈련, 경력 개발 등의 정책을 개선할 수 있습니다.

C, ChatGPT - 데이터 접근/확보

(아직) 복잡한 schema를 이해할 순 없지만, 현업이 SQL을 배우고 작성하는 데 실질적 도움이 될 수 있음

block3
1

지역 x 제품군 top-5 제품?

- Descriptive Analytics

Postgres SQL 테이블과 속성이 다음과 같습니다:

- # 테이블명 public.minhee_test
- # 주문일 TIMESTAMP WITHOUT TIME ZONE
- # 고객분류 TEXT
- # 시군구 TEXT
- # 지역 TEXT
- # 제품대분류 TEXT
- # 제품소분류 TEXT
- # 매출 DOUBLE PRECISION
- # 수량 BIGINT
- # 할인율 DOUBLE PRECISION
- # 이익 DOUBLE PRECISION
- #

위에 제공된 스키마를 사용해서 제품대분류와 지역별로 매출 총액 기준으로 최상위 5개 제품소분류를 뽑는 SQL 만들어주세요.

주어진 스키마를 사용하여 제품대분류와 지역별로 매출 총액 기준으로 최상위 5개 제품소분류를 뽑는 SQL 쿼리는 다음과 같습니다:

```
sql
WITH total_revenue_by_minor_category AS (
SELECT
    제품대분류,
```

Send a message.

ChatGPT may produce inaccurate information about people, places, or facts. [ChatGPT Mar 23 Version](#)

C, ChatGPT - 데이터 접근/확보



Text-to-SQL is coming to HEARTCOUNT in OCT 2023

HEARTCOUNT

AI Assist Retrieve all users who have CET listed as their timezone label, along with the channels they've created

```
1 SELECT * FROM main.users WHERE tz_label = 'Central European Time'
2
3 SELECT u.name, c.name FROM main.users AS u JOIN main.channels AS c ON u.id = c.creator WHERE u.tz_label = 'Central European Time'
```

RUN LIMIT: 1 000 00:00:01.19 SAVE COPY LINK

RESULTS QUERY HISTORY PREVIEW: 'USERS'

CREATE CHART DOWNLOAD TO CSV COPY TO CLIPBOARD Filter results

170 rows returned

color	deleted	has_2fa	id	is_admin	is_app_user	is_bot	is_owner	is_primary_owner	is_restricted	is_ultra_restricted
d1707d	NULL	NULL	UFP6V5W4E	NULL	NULL	NULL	NULL	NULL	NULL	NULL
a72f79	NULL	NULL	U01APT5FRQ	NULL	NULL	NULL	NULL	NULL	NULL	NULL
7d414c	NULL	NULL	U01AQLR42P5	NULL	NULL	NULL	NULL	NULL	NULL	NULL
4bbe2e	NULL	NULL	U01AS62LRFF	NULL	NULL	NULL	NULL	NULL	NULL	NULL
3c989f	NULL	NULL	U01AS95NNE9	NULL	NULL	NULL	NULL	NULL	NULL	NULL
674b1b	NULL	NULL	U01AVFZ4BU2	NULL	NULL	NULL	NULL	NULL	NULL	NULL
8f4a2b	NULL	NULL	U01AZLMAHL6	NULL	NULL	NULL	NULL	NULL	NULL	NULL

C, ChatGPT - 데이터 분석

Table 형식의 데이터셋에 대한 통계적 분석 못 함

- 테이블 크기 제한
- 평균, 총합, 상관관계, 드릴다운 등 못 함.



To find the total medical cost, we need to sum up the values in the 'medical.cost' column. Here is the total medical cost for the given data:

Total medical cost: 2,284,174.30635

H

correct answer is 2,587,831.437

The previous model used in this conversation is unavailable. We've switched you to the latest default model



I apologize for the error in my previous response. I see that I made a mistake in my calculation. The correct answer for the total medical cost for all individuals in the dataset is \$2,537,386.44.

- 연산할 레코드 개수가 커지면 덧셈조차 정확히 못 함.
- 왼편은 200개 레코드에 대한 총합을 계산해 달라고 요청한 결과.

H

so, is it fair to say that you are not able to do data analysis of actual dataset I provide?



You are correct. As an AI language model, I am unable to directly access or analyze datasets that you provide. My responses are based on my pre-existing knowledge and training, which does not include direct access to any specific datasets you might mention.

For actual data analysis, you would need to use statistical software or programming languages such as R, Python, or Excel to manipulate and analyze the data. If you have questions about how to perform certain analyses or interpret results, I would be happy to try to help based on my general knowledge of data analysis and statistics.

- 자기는 데이터셋 분석을 못하니 데이터 분석을 위해서는 R, Python, Excel 같은 소프트웨어 사용하라 함

ChatGPT plugins

We've implemented initial support for plugins in ChatGPT. Plugins are tools designed specifically for language models with safety as a core principle, and help ChatGPT access up-to-date information, run computations, or use third-party services.

[Join plugins waitlist](#)

[Read documentation](#) ↗

Code Interpreter 플러그인 (Python Interpreter)

- 손가락만 놀리면 열정적 초급 개발자가 노력 봉사합니다.
- *access to a very eager junior programmer working at the speed of your fingertips*
- 애쓰지 않는 프로그래밍이 가능해져서 모두가 프로그래머의 희열을 누릴 수 있습니다.
- *make completely new workflows effortless and efficient, open the benefits of programming to new audiences.*
- 아래 작업에 특히 유용합니다.
 - 수학 문제 풀기
 - 데이터 분석과 시각화

C, ChatGPT - 데이터 분석: Code Interpreter 플러그인

데이터셋: music.csv

- 아티스트 이름, 곡명, 발매년도와 같은 기본 정보와
- 곡 길이, 템포(BPM) 등의 콘텐츠 속성 정보로 구성

J

now let's do the following analysis:

- let's exclude year 0
- let's consider only years with over 100 songs

from that set of years, let's find a quantity that is the most positively correlated with time?



Finished working

Show work ▾

The quantity that is most positively correlated with time (song.year) in the filtered dataset is 'song.loudness', with a correlation coefficient of approximately 0.202.

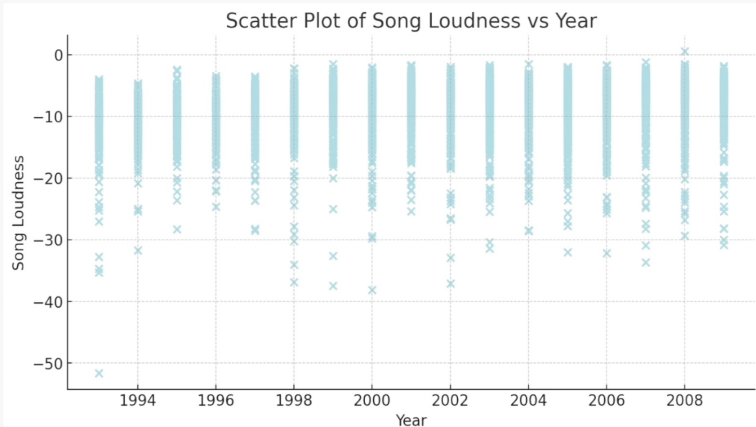
C, ChatGPT - 데이터 분석: Code Interpreter 플러그인

J can you plot that?



Finished working

Show work



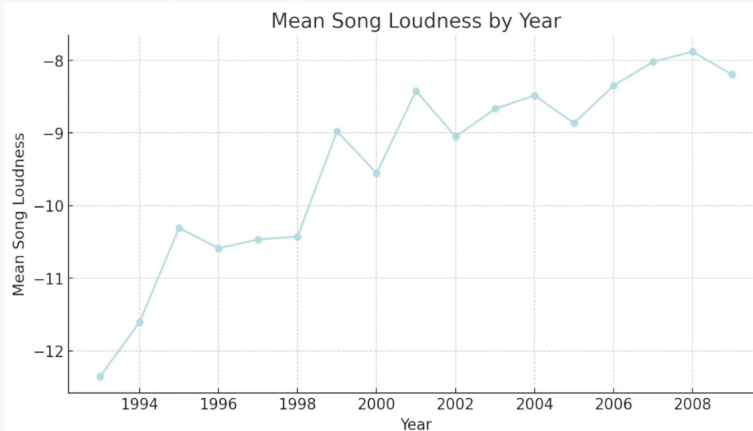
Here is a scatter plot of song loudness vs year in the filtered dataset. The plot shows the relationship between the loudness of songs and the year they were released, with data points representing individual songs.

J make an aggregated plot

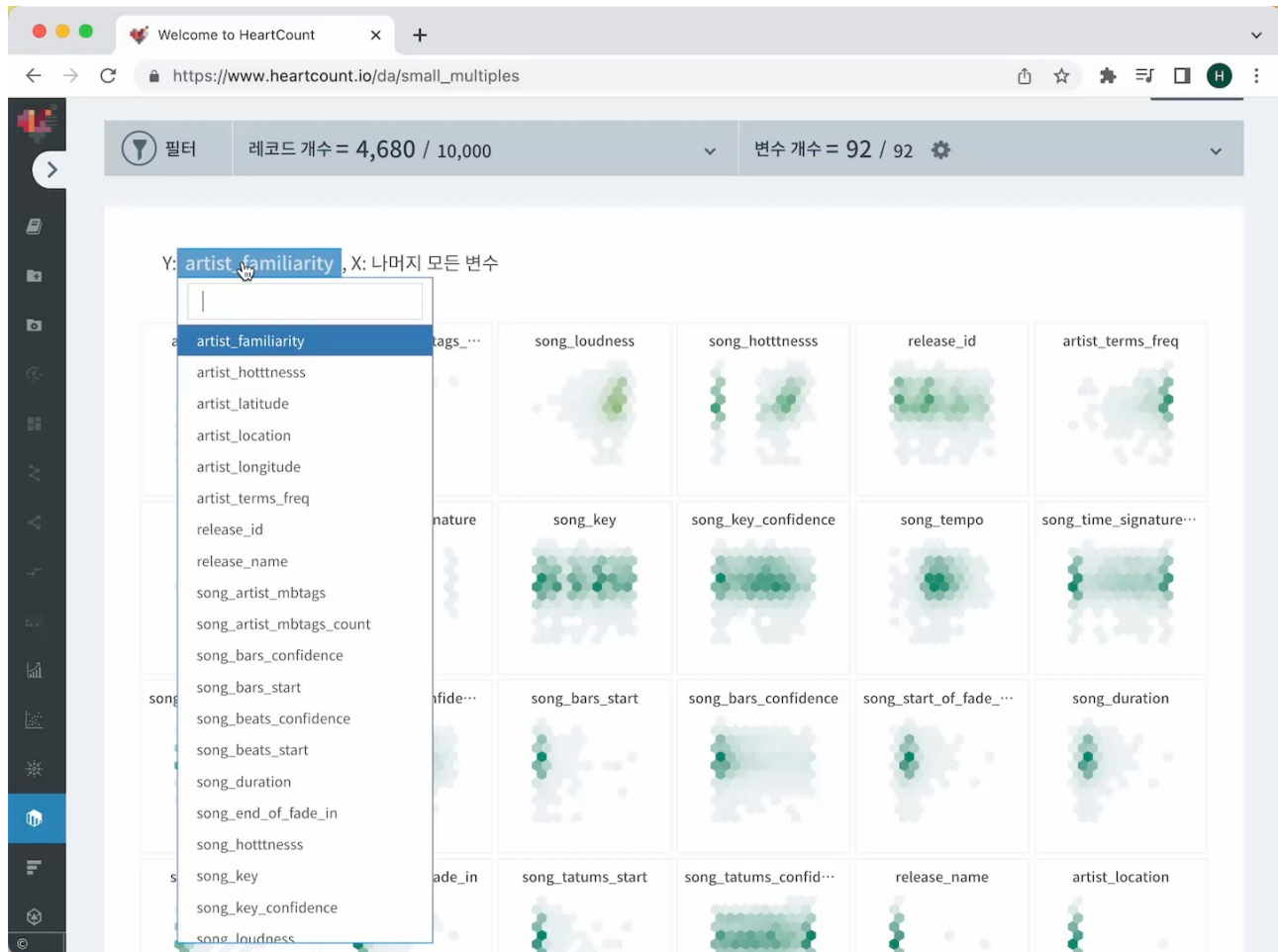


Finished working

Show work



Here is a plot of the mean song loudness by year in the filtered dataset. The plot shows the trend of the average loudness of songs over time, with data points representing the mean loudness for each year.



C, ChatGPT - 데이터 분석: MS 365 Copilot

Q. 이번 분기 사업 결과를 가지고 세가지 주요 트렌드를 알려주세요.

Country	Customer	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales	COGS
France	VanArsdel	Biodegradable Cleaning Products	Low	4345	\$10	\$352	\$1,529,440	\$15,294	\$1,514,146	\$
USA	VanArsdel	Biodegradable Cleaning Products	Low	4245	\$10	\$352	\$1,494,240	\$14,942	\$1,479,298	\$
Australia	VanArsdel	Biodegradable Cleaning Products	None	4124	\$10	\$352	\$1,451,648	\$0	\$1,451,648	\$
Mexico	VanArsdel	Biodegradable Cleaning Products	Medium	3954	\$10	\$352	\$1,391,808	\$83,508	\$1,308,300	\$
France	Contoso Pharmaceuticals	Biodegradable Cleaning Products	Low	4952	\$10	\$127	\$628,904	\$6,289	\$622,615	\$
France	Contoso Pharmaceuticals	Recycled Can	Low	4903	\$5	\$127	\$622,681	\$6,227	\$616,454	\$
Canada	Proseware	Natural Personal Care Products	Low	1954	\$270	\$302	\$590,108	\$5,901	\$584,207	\$
Canada	Proseware	Biodegradable Cleaning Products	High	1990	\$220	\$302	\$600,980	\$84,137	\$516,843	\$
USA	VanArsdel	Compostable Food Packaging	Low	1254	\$3	\$380	\$476,520	\$4,765	\$471,755	\$
USA	Proseware	Natural Personal Care Products	Medium	1465	\$224	\$302	\$442,430	\$26,546	\$415,884	\$
USA	Proseware	Sustainable Snack Packaging	Medium	1410	\$120	\$302	\$425,820	\$25,549	\$400,271	\$
Australia	Proseware	Natural Personal Care Products	Medium	1369	\$224	\$302	\$413,438	\$24,806	\$388,632	\$
USA	VanArsdel	Compostable Food Packaging	Medium	1052	\$3	\$380	\$399,760	\$23,986	\$375,774	\$
Canada	VanArsdel	Compostable Food Packaging	Low	980	\$312	\$380	\$372,400	\$3,724	\$368,676	\$
Australia	Proseware	Sustainable Snack Packaging	Medium	1298	\$120	\$302	\$391,996	\$23,520	\$368,476	\$
Australia	VanArsdel	Compostable Food Packaging	None	954	\$3	\$380	\$362,520	\$0	\$362,520	\$
Canada	Contoso Pharmaceuticals	Biodegradable Cleaning Products	Low	2785	\$110	\$127	\$353,695	\$3,537	\$350,158	\$

Key trends:

- Sales are trending up for all customers, except Proseware which is down this quarter.
- Manufacturing costs are down this quarter after an increase in Q3.
- Low discount products made up nearly half of total sales last month.

Explain



숫자(매출, 제조 원가)에 대해 분기와 범주(고객, 할인률 밴드 등)로 피벗 테이블을 만든 후, 해당 테이블을 언어 모델이 설명(Descriptive Analytics)하는 방식으로 작동하리라 추정

C, ChatGPT - 데이터 분석: MS 365 Copilot

File Home Insert Formulas Data Review View Automate Help

Aptos (Body) 11 B General

Proseware Sales growth by Product					
Product	Q1-22	Q2-22	Q3-22	Q4-22	Q1-23
Biodegradable Cleaning Products	3%	4%	6%	10%	5%
Compostable Food Packaging	9%	11%	15%	5%	13%
Reusable Containers	7%	10%	11%	6%	-9%
Reusable Water Bottles	15%	18%	18%	10%	17%
Natural Personal Care Products	10%	14%	9%	5%	33%
Sustainable Snack Packaging	7%	11%	6%	15%	8%
Total Growth	8%	11%	10%	7%	1%

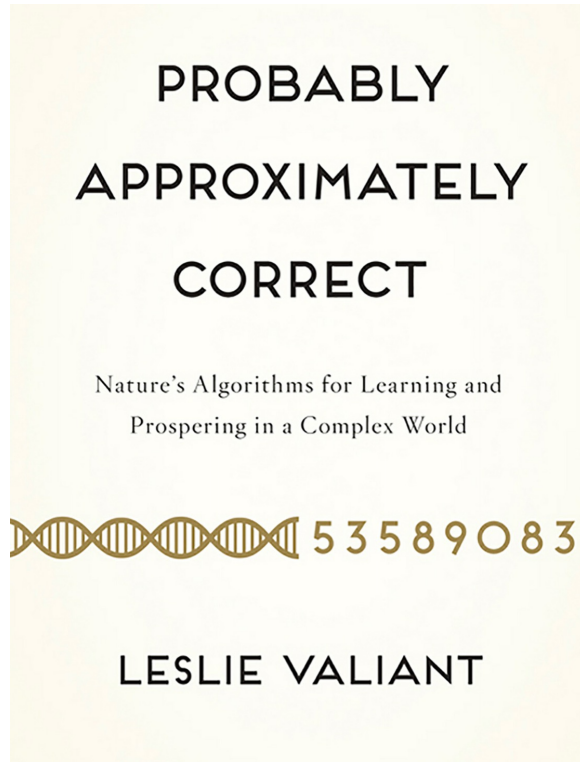
Proseware Sales growth by Country					
Country	Q1-22	Q2-22	Q3-22	Q4-22	Q1-23
Australia	12%	17%	16%	13%	3%
Canada	5%	8%	7%	5%	1%

“Proseware”의 매출 변화를 드릴다운 해줄래요?

Copilot

- half of total sales last month.
 - Explain
- Show me a breakdown of Proseware sales growth
- Here's a new sheet with a breakdown of Proseware sales over time. Remember to check for accuracy.
 - Undo
 - Explain
- Change from quarters to months
- Add a chart of Sales by Discount Band
- Help me visualize what contributed to the decline in sales growth

< > ≡ Data Summary +



확률적 판단

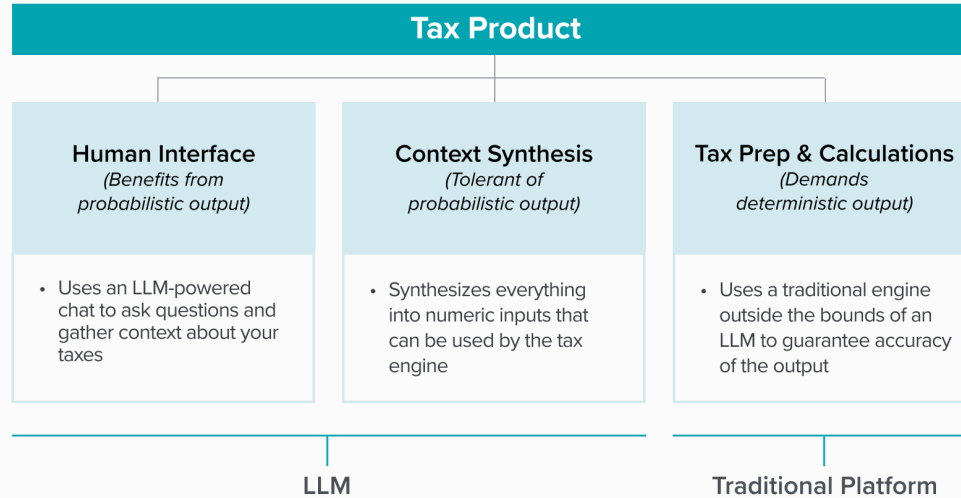
- Faced with any task, a computer can either be programmed to handle it, or it can be made to learn how to handle it, or perhaps a combination.
- Compared with the possibilities of creating a faultless program for a task, the learning solution will always have this essential weakness.
- Learning has a statistical aspect and for that reason cannot be made error-free.

C, ChatGPT for a Probabilistic Product

Probabilistic Products

- non-deterministic and often emergent attributes.
- 예측하기 힘들다가 좋은 의미로 쓰일 수 있는 경우

Sample Probabilistic Product: An AI Tax-Filing Assistant



Source: xxxxx



C, ChatGPT와 실무자들의 데이터 분석/활용 자동화

	질문	데이터 접근	분석 (인사이트 찾기)	보고서 작성
사람	 <p>현실 이해 개선 의지</p>	 <p>데이터 큐레이션 편성</p>	 <p>가설 수립 유용한 패턴 발견</p>	 <p>패턴+견해로 구성된 보고서</p>
챗GPT		 <p>SQL 쿼리 작성</p>	 <p>Descriptive Analytics 차트 해석</p>	 <p>문법/번역/스타일</p>

D: Data Newsletter

A: Ad-hoc Analysis

B: BI (The end of BI)

C: ChatGPT and Data

D: Data Newsletter

E: Everyone is an Analyst

F: Feature Engineering

G: Granularity

H: HEARTCOUNT

D, Data Newsletter

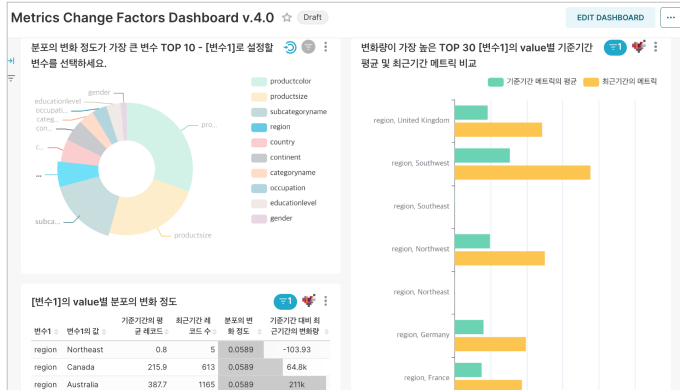
- 대시보드나 리포트는 pull 방식의 데이터 활용
- 정기적/이벤트 발생 시 데이터 뉴스레터를 발송(push)

Metrics Change Contributor Algorithm
범주 간 지표 차이 이해 →
차이를 최적화할 통제가능한 대상/요인 찾기

Narrative
Engine

“2023년 4월”의 “Sales 총합”은 전년
동월(2022.4) 6,000만원과 비교하여
1,000만원(12%) 증가했습니다.

감소 원인을 “Sub-Category” 수준에서
살펴보면, “의자”의 변동 폭이 1,500만원에서
750만원(50% 감소)으로 가장 컸고, 그
다음으로는 “Tables”이...



HTML
newsletter

1. 개별 변수(차원)에서의 변화량

“2023년 4월”의 “이익 총합”은 772천원으로 바로 직전 기간(2023년 3월)의 751천원보다 21천원 (2.7%) 증가했습니다. 직전년 동월 기간(2022년 4월) “이익 총합” 평균값인 351천원과 비교해서는 421천원(120%) 증가했습니다.

- 지표 변화를 개별 범주형 변수 수준에서 살펴보면,
 - 바로 직전 기간(2023년 3월) 대비 증가한 항목들
 - “Category Name: Bikes”의 변동 폭이 942천원(757천원 ⇒ 1.7백만원; 124%)으로 가장 컸습니다.
 - 그 다음으로는 “Home Owner: Yes”인 매출 건들의 변동 폭이 709천원 (555천원 ⇒ 1.26백만원; 124%)로 컸습니다.
 - > [자세히 살펴보기](#)
 - 직전년 동월 기간(2022년 4월) 대비 증가한 항목들
 - “Product Color: Yellow”인 상품들의 경우 변동 폭이 110천원(105천원 ⇒ 561천원; 434%)로 컸습니다.
 - > [자세히 살펴보기](#)

[변수1]의 value별 기준기간 대비 최근기간의 변화량

변수1	변수1의 값	기준기간의 평균 레코드	최근기간 레코드 수	메트릭 증감율(%)	기준기간 대비 최근기간의 변화량	기준기간의 메트릭 평균	최근기간의 메트릭
categoryname	Bikes	440.4	1,157	124	941,796	757k	1.7M
homeowner	Y	1,205.0	3,732	128	708,701	555k	1.26M
maritalstatus	M	961.8	2,956	136	559,947	411k	971k
gender	M	872.2	2,340	127	524,340	390k	695k

Gmail

data n

Google

99+

Mail

Chat

Spaces

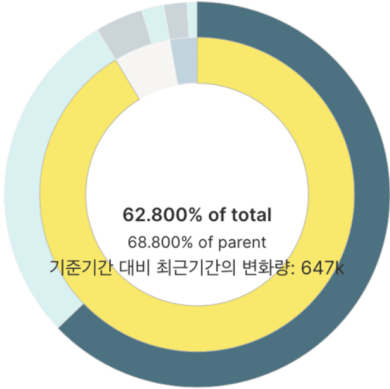
Meet

2. 변수 두개의 조합 관점에서의 변화량

- 지표 변화를 변수 두개의 조합으로 “이의 총합”의 변화량을 살펴 본 경우,
 - 바로 직전 기간(2023년 3월) 대비 증가한 항목들
 - “Category Name: Bikes”이고 "Home Owner: Y"인 경우 변동 폭이 647천원(527천원 ⇒ 1.17백만원; 123%)으로 가장 컸습니다.
 - > [자세히 살펴보기](#)

메트릭 변화에 가장 크게 기여한 [변수1]과 [변수2]의 조합 TOP 10

categoryname, Bikes homeowner, Y 62.800%



62.800% of total
68.800% of parent
기준기간 대비 최근기간의 변화량: 647k

관련 대시보드 이동

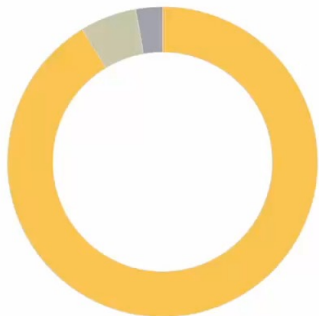
Metrics Change Factors Dashboard v.3.0

☆ 초안

대시보드 수정

메트릭값 변화량 기준 개별 범주 분포 변화 정도 기준

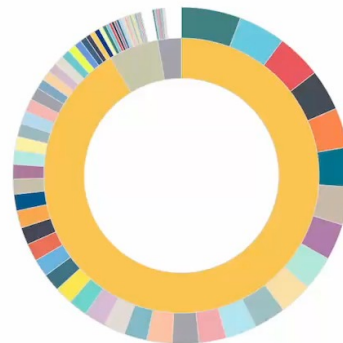
메트릭 변화에 가장 크게 기여한 [변수1]의 value TOP 10



[변수1]의 value별 기준기간 대비 최근기간의 변화량

변수1	변수1의 값	기준기간의 평균 레코드 수	최근기간 레코드 수	메트릭 증감률 (%)	기준기간 대비 최근기간의 변화량
categoryname Bikes	440.4	1,157	124	941,796	
categoryname Accessories	1,043.8	3,337	214	60,598	
categoryname Clothing	261.2	936	254	28,550	

메트릭 변화에 가장 크게 기여한 [변수1]과 [변수2]의 조합 TOP 10



[변수1]과 [변수2] 값들의 조합에 따른 기준기간 대비 최근기간의 메트릭 변화량

변수1	변수1의 값	변수2	변수2의 값	기준기간의 평균 레코드 수	최근기간 레코드 수	증감률 (%)	기준기간 대비 최근기간의 변화량	기준기간의 메트릭 평균	최근기간의 메트릭
categoryname Bikes	hoseowner	Y		305.2	788	123	647,497	527k	1.17M
categoryname Bikes	maritalstatus	M		227.1	586	131	510,576	389k	900k
categoryname Bikes	gender	M		215.9	581	132	489,657	370k	860k
categoryname Bikes	gender	F		221.9	570	117	448,185	381k	830k
categoryname Bikes	subcategoryname	Mountain Bikes		147.5	429	159	431,238	272k	703k

E: Everyone is an Analyst

A: Ad-hoc Analysis

B: BI (The end of BI)

C: ChatGPT and Data

D: Data Newsletter

E: Everyone is an Analyst

F: Feature Engineering

G: Granularity

H: HEARTCOUNT

F: Feature Engineering

A: Ad-hoc Analysis

B: BI (The end of BI)

C: ChatGPT and Data

D: Data Newsletter

E: Everyone is an Analyst

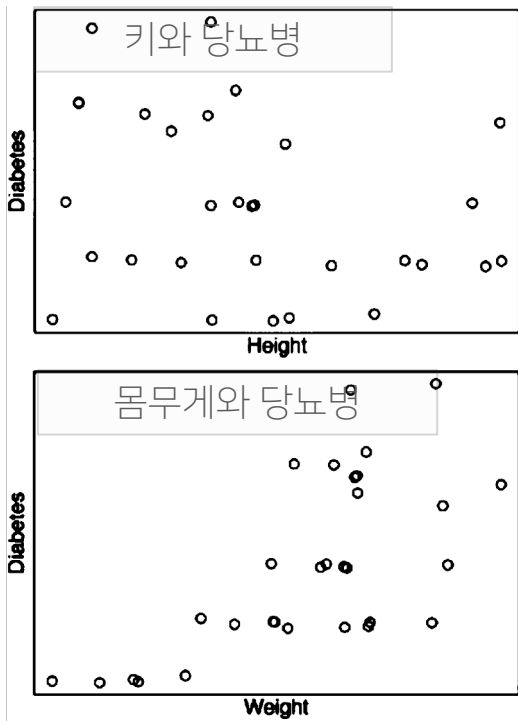
F: Feature Engineering

G: Granularity

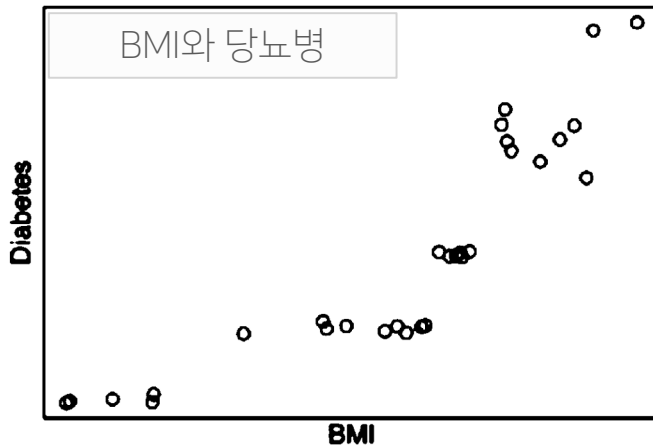
H: HEARTCOUNT

F, Feature Engineering

Feature Engineering: From Raw Variable to Derived Variable
Y를 더 잘 이해하고 설명하기 위해 기존 변수를 창의적으로 가공하는 일



- 비만도 더 잘 설명할 수 있는 새로운 변수 가공
- *BMI(Body Mass Index) = kg/m^2



*발명한 사람의 이름을 따서
Quetelet Index라고도 함

F, Feature Engineering

Ad-hoc Analysis(EDA) 관점에서 중요한 작업 두가지
숫자를 범주로
날짜를 범주로

F, Feature Engineering - 숫자 변수 가공

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

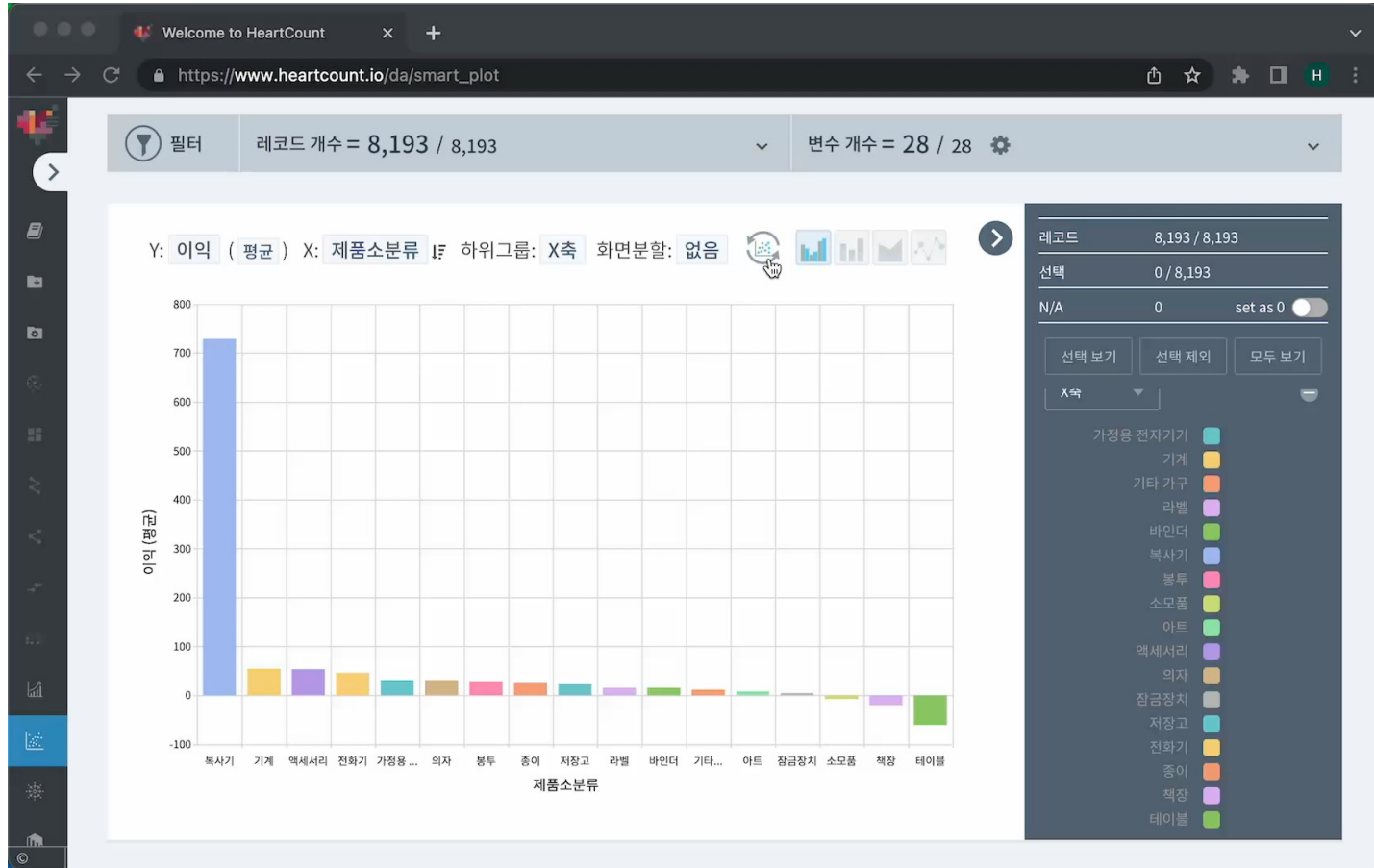
F, Feature Engineering - 숫자 변수 가공 (binning)



HEARTCOUNT

차이를 최적화할
통제가능한 대상과
요인 찾기

- 할인률 숫자 →
할인률 구간(범주)
변환



F, Feature Engineering - 숫자 변수 가공 (percentile)



HEARTCOUNT

고성과 집단과
저성과 집단 찾기

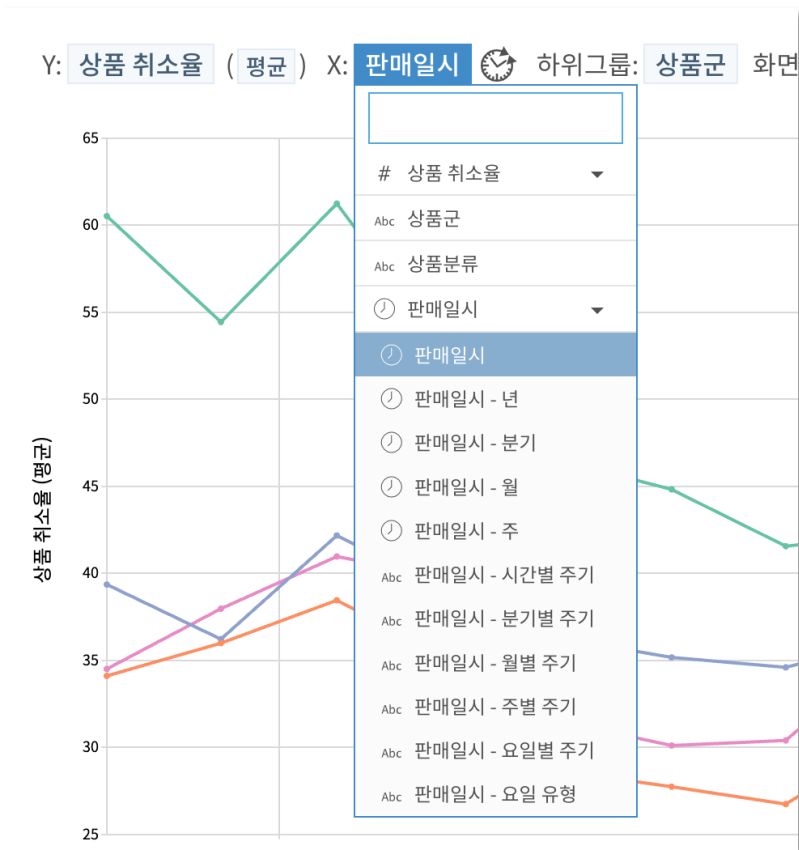
- 이익 숫자 →
이익 Percentile
(범주) 변환

The screenshot shows the HEARTCOUNT dashboard interface. At the top, there's a navigation bar with the title '드릴다운' (Drilldown) and a '분석' (Analysis) button. Below this, a filter section shows '레코드 개수 = 8,193 / 8,193' and '변수 개수 = 14 / 14'. The main content area is titled 'SMART DISCOVERY' and displays a search query: '제품소분류 (와 변수 선택) 별 이익 평균 드릴다운'. Below the query, there are view options (table, grid, chart) and a slider for '레코드 개수 >= 57'. The main data is presented in a table with columns for '제품소분류' (Product Category) and '전체 평균: 25.08' (Overall Average: 25.08). The table lists various product categories with their respective record counts and profit averages, along with a visual bar chart for each row.

제품소분류	전체 평균: 25.08
복사기 (57)	728.00
기계 (102)	54.41
액세서리 (634)	53.25
전화기 (741)	45.44
가정용 전자기기 (386)	31.18
의자 (522)	30.40
봉투 (212)	27.72
종이 (1.1K)	24.57
저장고 (688)	21.56
라벨 (301)	14.86

F, Feature Engineering - 시간 변수 가공

판매일시	상품 취소율	상품군	상품분류
2020-11-01 01:00	37.34	잡화	가방/핸드백
2020-11-01 01:00	36.28	보석	손목시계
2020-11-01 09:00	31.49	트렌드의류	코트
2020-11-01 09:00	20.05	트렌드의류	바지
2020-11-01 13:00	46.38	잡화	가방/핸드백
2020-11-01 13:00	37.08	잡화	가방/핸드백
2020-11-01 13:00	77.78	잡화	가방/핸드백
2020-11-01 13:00	41.18	잡화	가방/핸드백
2020-11-01 13:00	64.29	잡화	가방/핸드백
2020-11-01 13:00	27.27	잡화	머플러/스카프/솔
2020-11-01 23:00	37.5	의류	재킷
2020-11-01 23:00	34.12	의류	바지
2020-11-02 08:00	30.49	의류	다운/패딩
2020-11-02 08:00	28.21	의류	다운/패딩
2020-11-02 08:00	31.4	의류	니트/스웨터
2020-11-02 08:00	30.25	의류	니트/스웨터
2020-11-02 08:00	37.67	의류	다운/패딩
2020-11-02 08:00	48.74	의류	코트
2020-11-02 08:00	60.32	의류	코트
2020-11-02 10:00	27.18	잡화	가방/핸드백



익스플레이너

분석

필터
레코드 개수 = 3,119 / 3,119
변수 개수 = 12 / 13

◀
◀
두번째 요인
▶
▶

상품군과 판매일시 - 월별 주기가 함께 두번째 설명력을 가지고 있으며 상품 취소율 차이(변화량)의 22.19%를 설명할 수 있습니다.

상품 취소율이 평균보다 높았던 조건들:

- [상품군 : 보석]이고 [판매일시 - 월별 주기 : 1월]인 경우 상품 취소율 평균값이 61.2로 전체 평균(36.75)보다 24.45만큼 높았습니다.
- [상품군 : 보석]이고 [판매일시 - 월별 주기 : 11월]인 경우 상품 취소율 평균값이 60.49로 전체 평균(36.75)보다 23.74만큼 높았습니다.

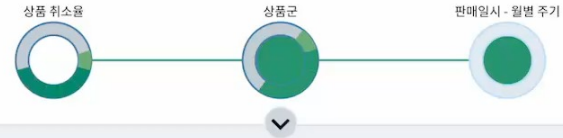
상품 취소율이 평균보다 낮았던 조건들:

- [상품군 : 의류]이고 [판매일시 - 월별 주기 : 10월]인 경우 상품 취소율 평균값이 25.94로 전체 평균(36.75)보다 -10.81만큼 낮았습니다.
- [상품군 : 의류]이고 [판매일시 - 월별 주기 : 5월]인 경우 상품 취소율 평균값이 26.71로 전체 평균(36.75)보다 -10.04만큼 낮았습니다.

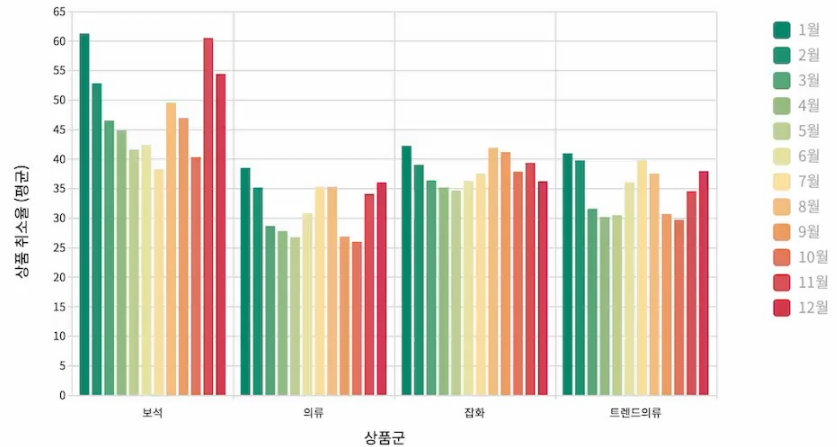
[Generate New Narrative](#)

[Change Visualization](#)

상품 취소율 의 차이는 어디서 비롯되는지 궁금해요.



Y: 상품 취소율 (평균) X: 상품군 하위그룹: 판매일시 - 월별 주기 화면분할: 없음



G: Granularity

A: Ad-hoc Analysis

B: BI (The end of BI)

C: ChatGPT and Data

D: Data Newsletter

E: Everyone is an Analyst

F: Feature Engineering

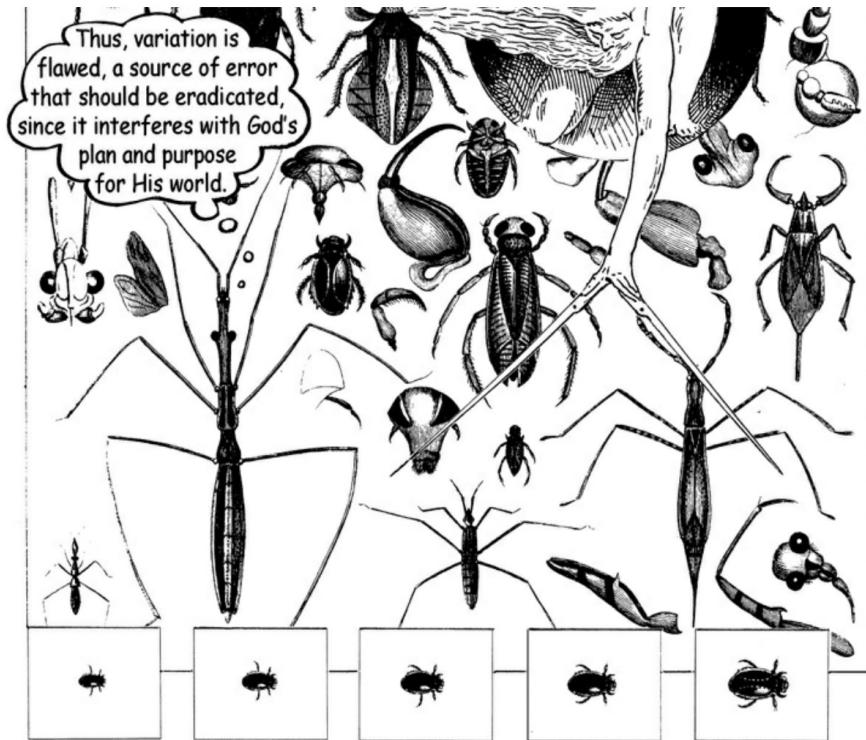
G: Granularity

H: HEARTCOUNT

G, Granularity – 평균의 문제

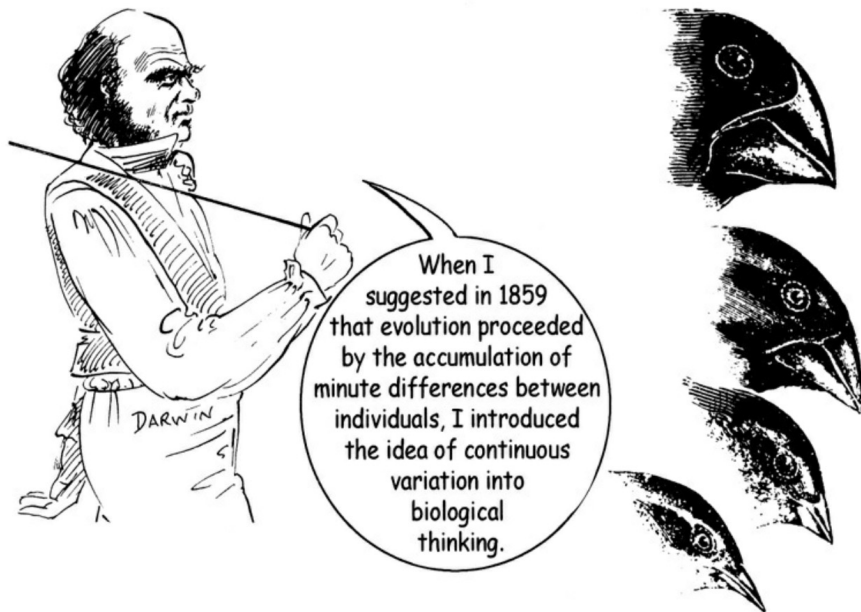
초기의 통계학

- 결정론적 세계관에 바탕을 둔 이데아/본질의 추구
- 평균값이 대상이 보유한 이상적인 속성(Idealized Mean)
- Variation(차이)은 제거해야 할 오류



다윈의 등장

- Type/Essence(본질) → Variation(차이)
- 차이(변이)의 점진적 누적에 의해 진화가 이루어진다는 발견
- 개별 개체에 존재하는 차이(변이)에 관심을 갖기 시작



G, Granularity - 평균의 문제

The Problems with Average: Not Robust!



G, Granularity – 요약값 vs. 개별 레코드(Granular) 시각화

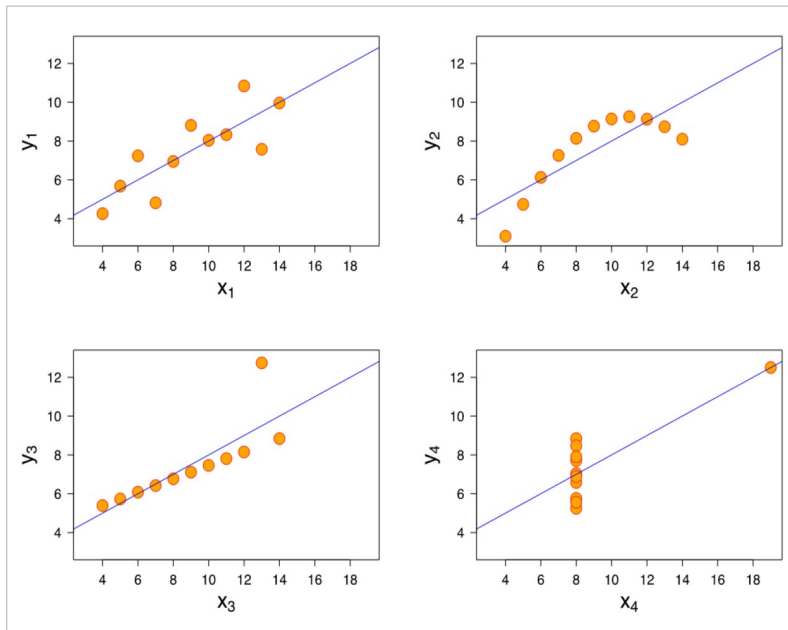
평균: 평균적 이해; 개별 레코드 시각화: 차이에 대한 이해

동일한 평균, 분산, 상관계수

	I		II		III		IV	
	X	Y	X	Y	X	Y	X	Y
평균	9	7.5	9	7.5	9	7.5	9	7.5
분산	11	4.1	11	4.1	11	4.1	11	4.1
상관계수	0.82		0.82		0.82		0.82	

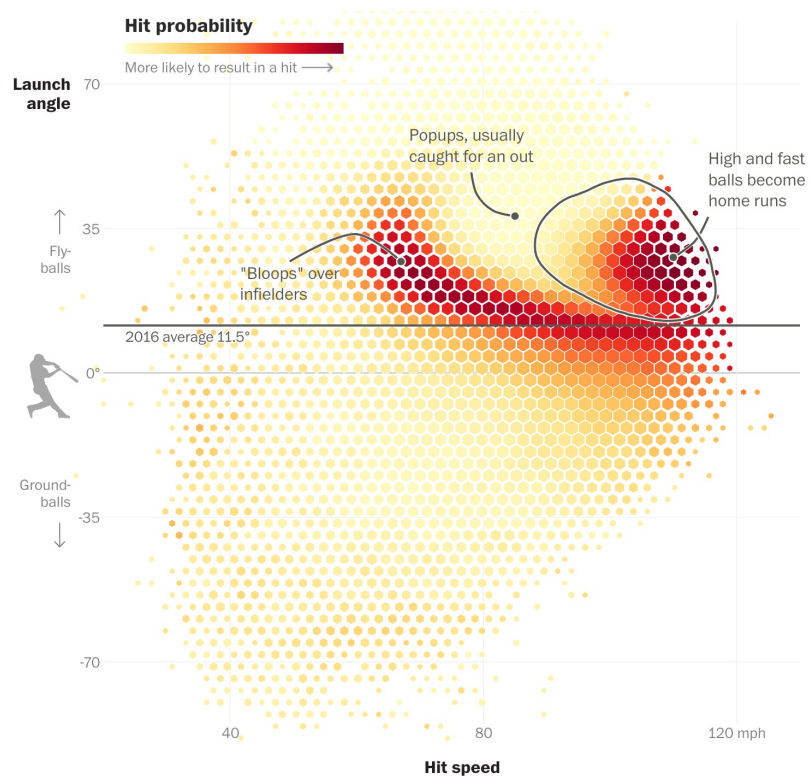
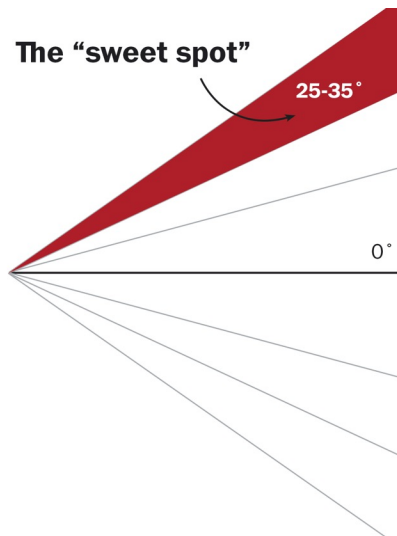
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

시각화를 통해 현실의 복잡성이 드러남



G, Granularity – 홈런을 찾아서

Analysts pinpointed the range of 25-35 degrees as the sweet spot for home runs, when paired with an exit velocity of 95 mph or greater.



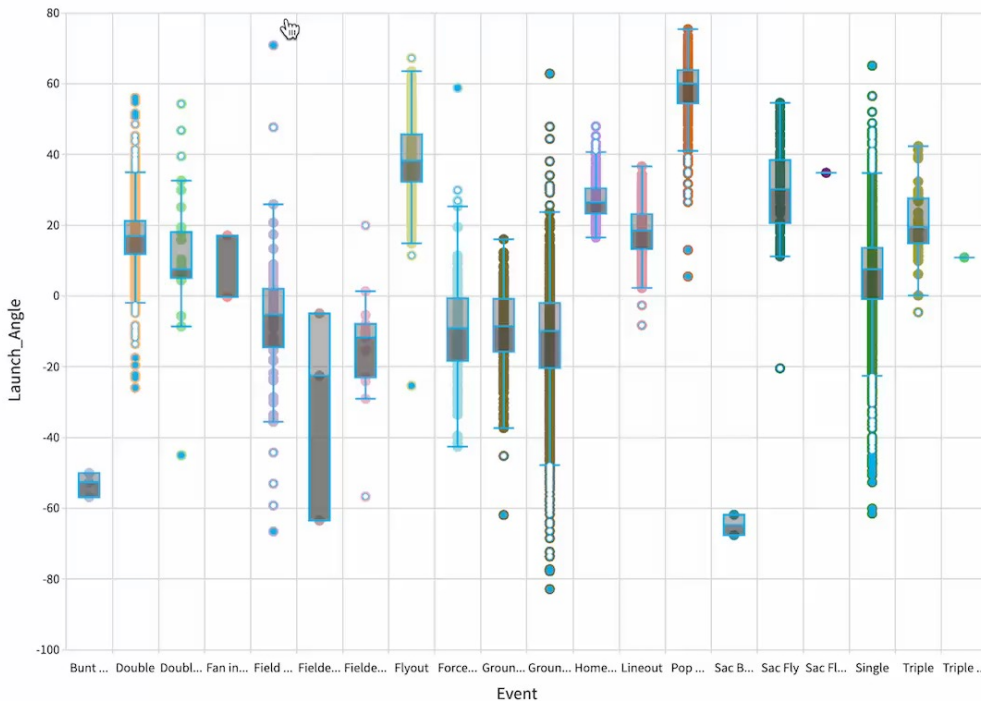
범주(이벤트) 간
차이 이해
→
범주(이벤트) 내
차이 이해

스마트 플롯

필터 적용

필터
레코드 개수 = 7,296 / 7,296
변수 개수 = 18 / 18

Y: Launch_Angle X: Event 하위그룹: 없음 화면분할: 없음



레코드 7,296 / 7,296
 선택 0 / 7,296
 N/A 0 set as 0

 Event

- Bunt Groundout ■
- Double ■
- Double Play ■
- Fan interference ■
- Field Error ■
- Fielders Choice ■
- Fielders Choice Out ■
- Flyout ■
- Forceout ■
- Grounded Into DP ■
- Groundout ■
- Home Run ■
- Lineout ■
- Pop Out ■
- Sac Bunt ■
- Sac Fly ■
- Sac Fly DP ■
- Single ■
- Triple ■
- Triple Play ■

H: HEARTCOUNT

A: Ad-hoc Analysis

B: BI (The end of BI)

C: ChatGPT and Data

D: Data Newsletter

E: Everyone is an Analyst

F: Feature Engineering

G: Granularity

H: HEARTCOUNT

H, HEARTCOUNT

데이터로 조직의 운명이 바뀌려면..

보통의 사람들이 애쓰지 않고 데이터에서 가치를 찾을 수 있는 도구
Effortless Analytics! Everyone's an Analyst!

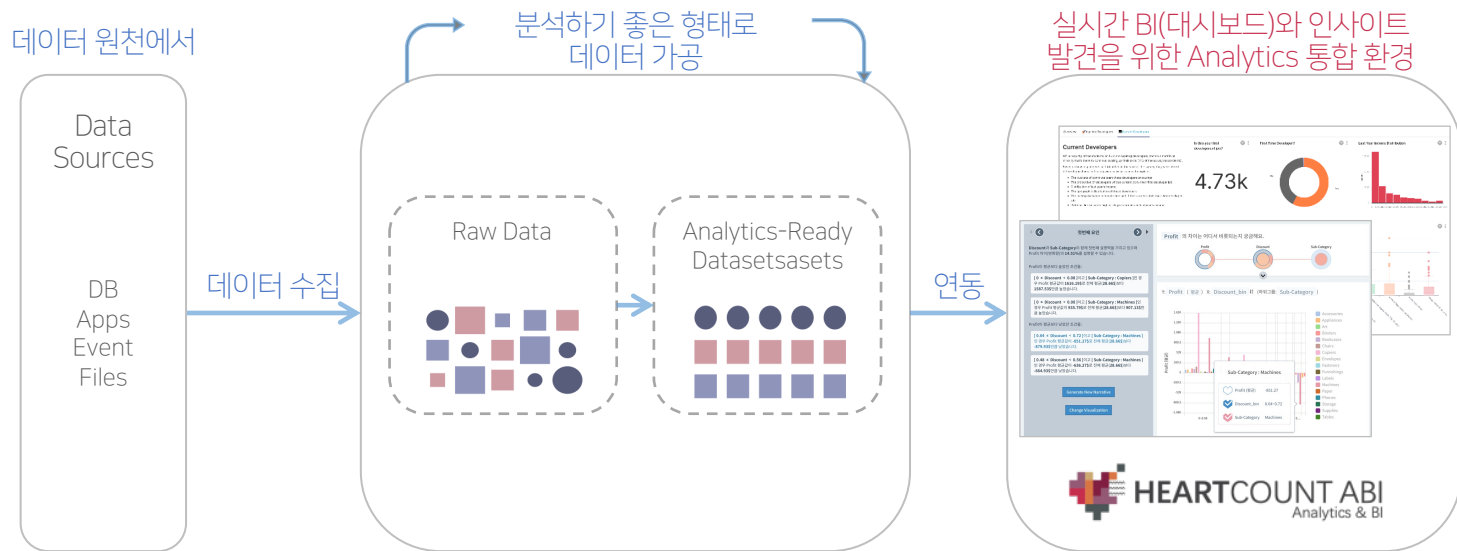
데이터와 도구와 당신이 서로 힘을 주고 받을 수 있도록,
HEARTCOUNT는 최선을 다하겠습니다.

- ④ **A:** Ad-hoc Analysis - 데이터에서 질문-답을 찾는 경험이 바뀌어야 한다.
- ② **B:** BI (The end of BI) - BI로 대표되는 데이터 도구가 재발명돼야 한다.
- ③ **C:** ChatGPT and Data - 챗GPT 같은 기술도 도움이 되겠지만
- ⑦ **D:** Data Newsletter - 데이터가 소비자에게 먼저 찾아가는 일
- ① **E:** Everyone is an Analyst - 모든 실무자들이 분석가가 되려면
- ⑤ **F:** Feature Engineering - 패턴 발견에 도움이 되게 변수 가공하는 일
- ⑥ **G:** Granularity - 집계 데이터 대신 해상도가 높은 데이터를 다루는 일
- ⑧ **H:** HEARTCOUNT - 다정한 데이터 도구 하트카운트가 함께 할게요!

H, HEARTCOUNT

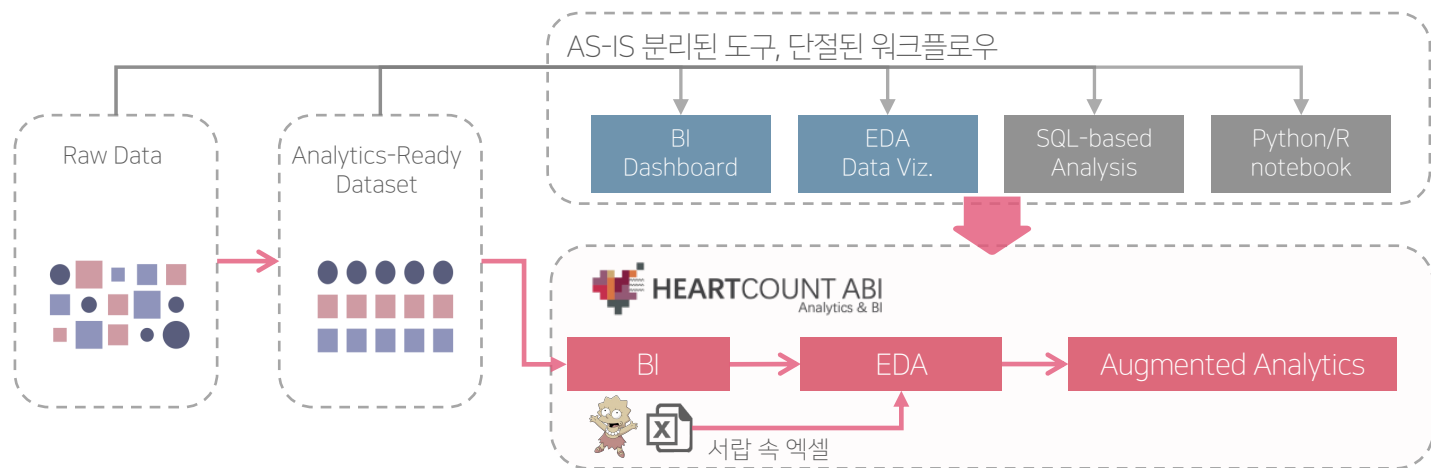
BI와 Analytics이 유기적으로 결합된 세계 최초의 데이터 활용 도구

- 숲도 보고: BI(대시보드) 도구로 지표 현황/추이 확인; 질문과 관련된 데이터셋 바로 접근
- 나무도 보고: Analytics 도구로 왜? 어떻게?에 대한 질문에 대한 정량적 답변 즉시 확인



H, HEARTCOUNT

하나의 도구로, 데이터 접근부터 인사이트 발견까지

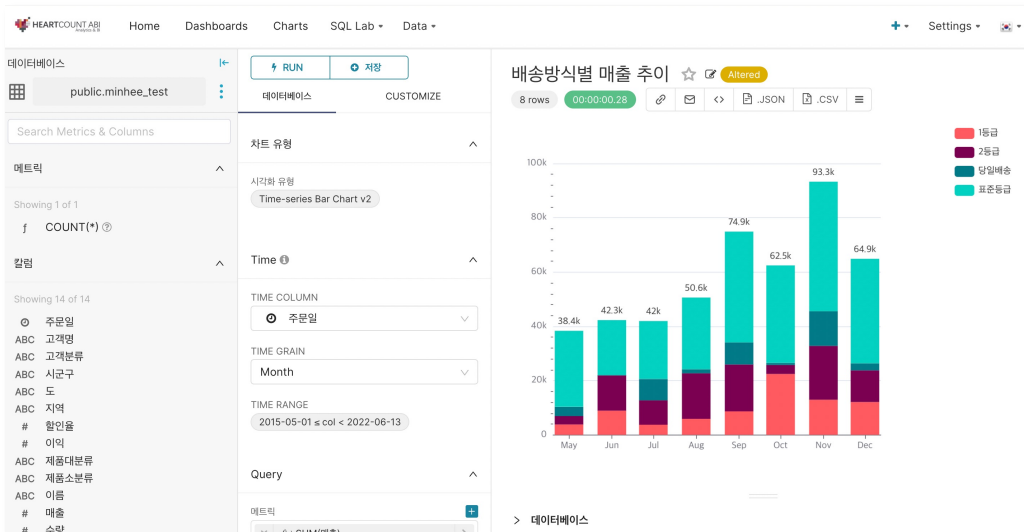


	Pain Points	HEARTCOUNT ABI
데이터 팀	<ul style="list-style-type: none"> • 잦은 데이터 추출 요청; 지원 지연 	<ul style="list-style-type: none"> • 미리 편성된 데이터셋 준비 • 고급 사용자를 위한 SQL editor
현업	<ul style="list-style-type: none"> • 수동 분석, 빠른 결과 도출 • R/Python 어려움 	<ul style="list-style-type: none"> • 질문 관련 데이터로 바로 EDA • 빠르지 않은 패턴 자동 발견

H, HEARTCOUNT

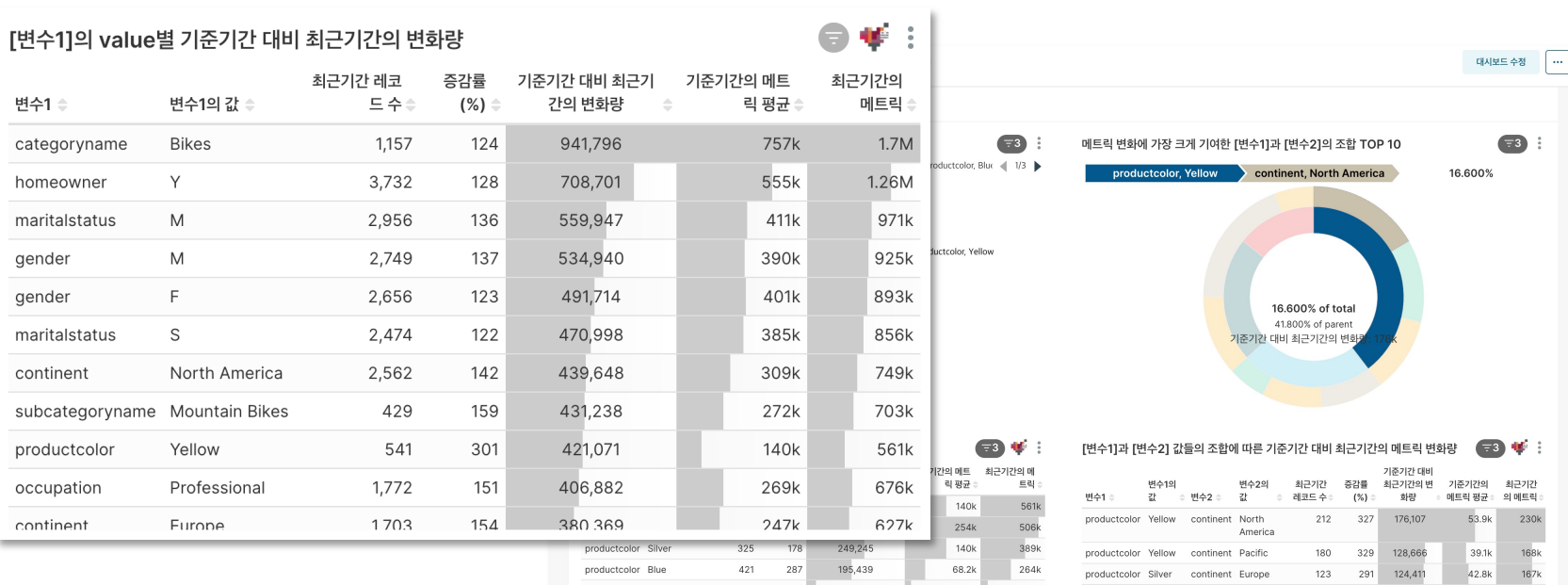
세상의 모든 DB, DW와 연동 가능

- RBAC(Role Based Access Control)로 역할에 따른 접근 통제
- 실무자들이 코딩 없이 Self-Service로 대시보드 똑딱 구현
- 실무자들이 질문과 관련된 데이터셋 추출, Analytics 연계 분석



지표 변화 요인 자동 발견

- 매출 총합, 전환률 등 주요 경영지표가 변했을 때 해당 변화에 가장 큰 기여를 한 요인을 자체 알고리즘으로 자동 발견
- 지표 변화의 주요 요인들을 변화에 미친 절대적 기여도와 통상적인 기여도 대비 급변한 정도, 두 가지 기준으로 제공



H, HEARTCOUNT

분석하지 말고, 질문하세요!

애쓰지 않고도 What, Why, and How 질문에 대한 뻔하지 않고 쓸모있는 답변 자동 발견

데이터셋

차트 맥락을 담고 있는
데이터셋 자동 연동

변수 자동 가공

분위수(Percentile)
월, 요일 등 변수 자동 가
공



What_대시보드&드릴다운
KPI 차이를 최대화하는 조건 자동 발견



Why_요인 분석
KPI 차이를 만드는 요인들 이해



Why_비교 분석
임의의 두 집단 간 특성 차이 이해



How_마이크로 세그멘테이션
KPI 개선을 위한 실행 전략 발견



Explore_시각화
변수들 관계 시각적 탐험·가설 생성

데이터 히어로 커뮤니티

- 6,000명 이상이 수료한 현업/실무자를 위해 구성된 데이터 활용 교육
- 본인 지식/경험 수준, 스케줄에 따른 Self-Paced Learning
- 실무 현장에서의 데이터 활용에 초점을 맞춘 시각화, 통계 분석 이론과 실습

“

실제 데이터를 가지고 다양한 분석을 할 수 있어서 좋았습니다.

머신러닝을 공부하면서 EDA에 대해 막히거나 막막했던 것들이 있었는데 그런 점을 해소할 수 있어서 좋았습니다.

데이터 초보자도 진행할 수 있게 강사가 함께 있는 게 좋았습니다!

데이터 분석에 대한 전반적인 개요와 실질적으로 어떻게 분석을 하고 해석하는지에 대한 큰 덩어리를 볼 수 있어서 좋았다.

데이터 분석에 관한 여러가지 관점과 이론들을 많이 배운 것도 물론 좋았지만 하트카운트라는 실무에서 사용하는 분석 툴을 처음으로 만져보고 이용하는 것이 좋았다.

확실히 내가 직접 코드로 인사이트를 찾으며 한줄한줄 코딩하는 것보다 더 적은 노력으로 더 시각적이고 훌륭한 분석이 가능했다.

”