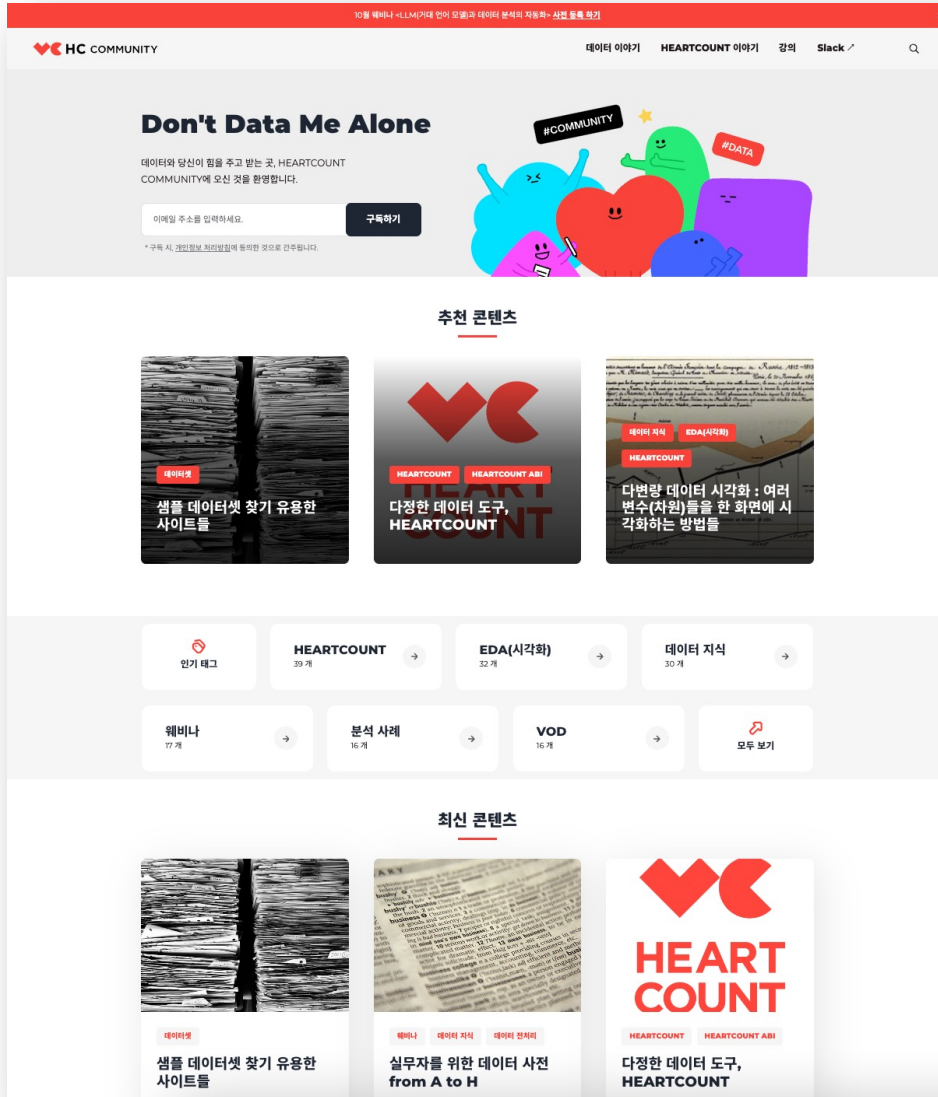


 HEARTCOUNT 10월 웨비나

LLM(거대 언어 모델)과 데이터 분석의 자동화

23.10.27 | IDK SQUARED 양승준



실무자를 위한
모든 데이터 콘텐츠가 한 곳에 !

- 데이터 강의
(데이터 리터러시, 통계학, SQL, 데이터 전처리 ..)
- 데이터 블로그·웨비나 VOD
- 하트카운트 활용 사례 & 실습 예제

지금 페이지를 방문하여 구독해보세요

1. 하트카운트 시작하기 - 데이터셋 연동
2. 시계열 데이터 분석
3. 여러 차원을 한 화면에 시각화하기
4. 상관관계와 분포 - 개별 레코드 수준에서의 시각화 기법들
5. 드릴다운과 트리맵 - 전체를 구성하는 개별 요소들 확인하기
6. 분석하기 좋은 데이터셋을 구성하는 요소들
7. 특정 행동을 보인 집단의 특성을 이해하기
8. 회귀 분석을 통해 지표 차이의 요인 이해하기
9. Causal Inference: 상관관계로 인과성에 대해 이야기하는 법
- 10. LLM(거대언어모델)과 데이터 분석의 자동화**
11. (tentative) 데이터 보고서 잘 쓰는 법
12. 월은 웨비나 쉽니다.

챗GPT와 데이터 분석의 자동화

Google

All Videos News Images Books More Tools

About 192,000 results (0.29 seconds)

Medium
<https://medium.com> > heartcount > 챗gpt와-데이터-...
챗GPT와 데이터 분석 업무의 자동화
 Apr 9, 2023 — 챗GPT와 데이터 분석 업무의 자동화 · 질문: 경영 지표의 급변동을 자동으로 감지하여 알려 주거나, · 분석: 지표를 최적화할 수 있는 패턴을 자동으로 ...

실무자를 위한 데이터 사전

← → <https://community.heartcount.io/ko/dh--webinar-data-fromatoh/>

웨비나 다시 보기

목차

- 웨비나 다시 보기
- 발표 자료 다운로드
- 요약 블로그
- 후기 이벤트



실무자를 위한 데이터 사전 A to H

Ad-hoc Analysis, BI(The end of BI), ChatGPT, Data Newsletter, Everyone is an Analyst, Feature engineering, Granularity, HeartCount

실무자를 위한 데이터 사전

2023. 06. 09 PM 3:00 - 4:30

Watch on YouTube

I. 2015~

AUGMENTED
ANALYTICS

- 실무자, 패턴 자동 발견
- 개별 레코드 수준 시각화



II. 2020~

B-TO-THE-A

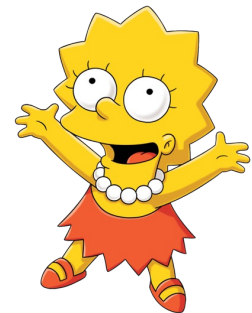
- BI to the Analytics
 - 분석 워크플로우 통합
 - 지표 추가 질문 답변
 - 데이터 접근

데이터에서 질문에 대한
답을 찾는 사용자 경험

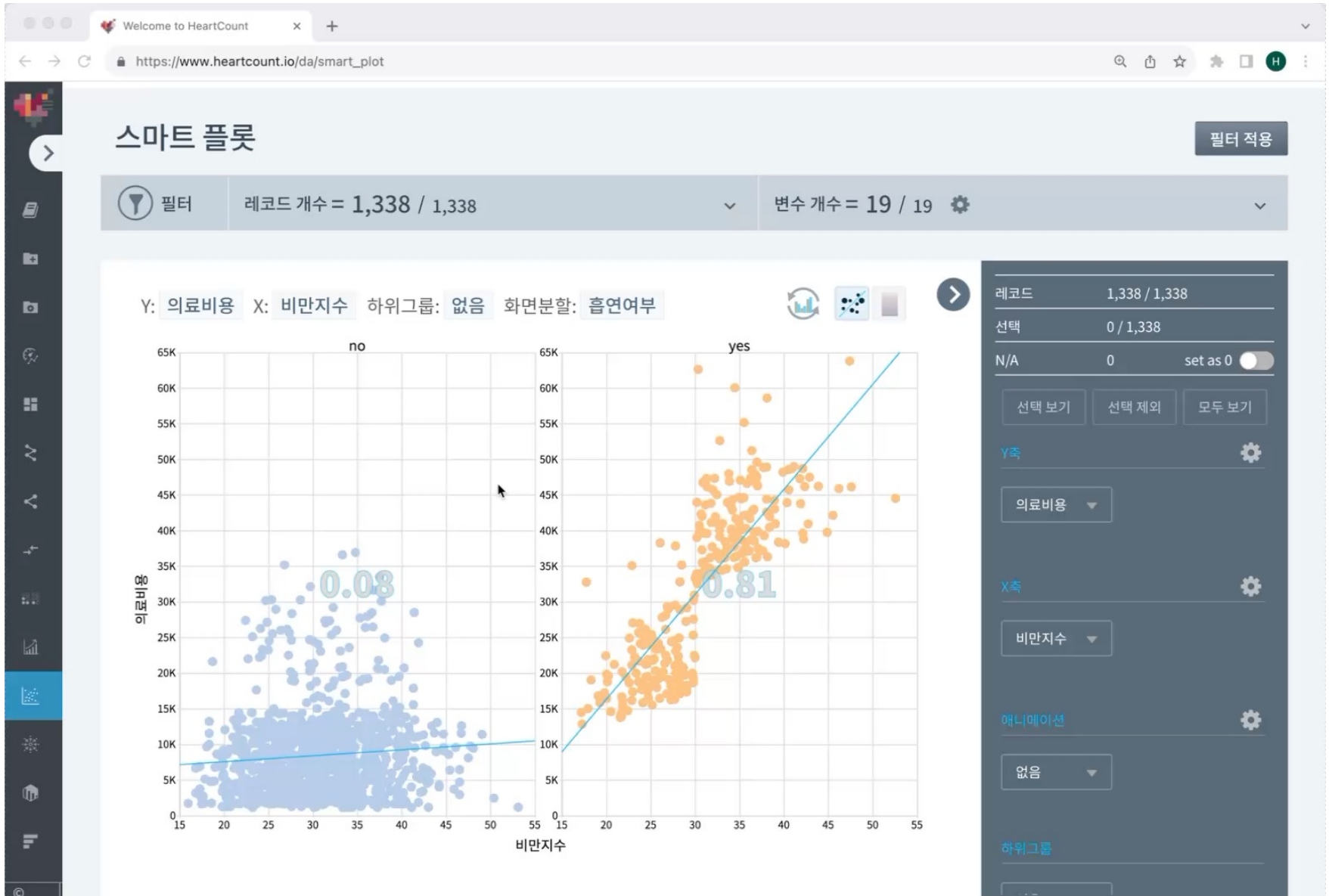
III. 2023~

DATA
EXPERIENCE

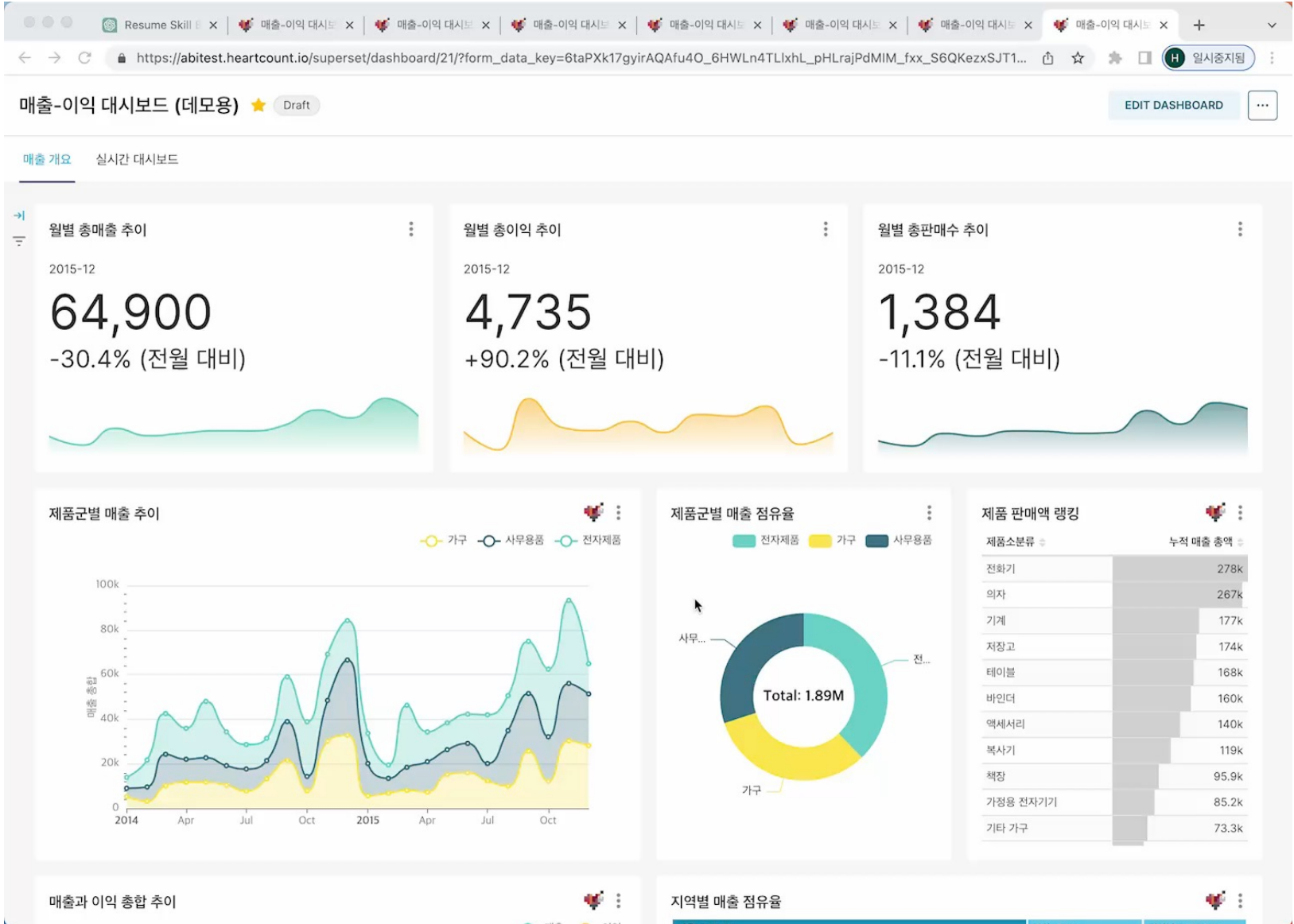
- 데이터 질문-답변 자동화
 - Text-to-SQL
 - Dialogue
 - Signal



개별 레코드 수준 시각화, 패턴 자동 발견/자연어 설명



대시보드(BI)는 FAQ 용 - 수많은 추가 질문들 답변 X



B-to-the-A

BI와 Analytics이 유기적으로 결합된 세계 최초의 데이터 활용 도구

숲도 보고

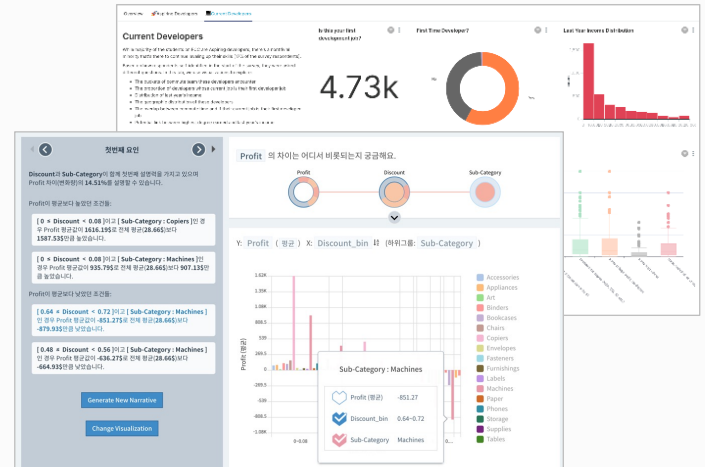


BI(대시보드) 도구로
지표(Metrics) 현황/추이 확인

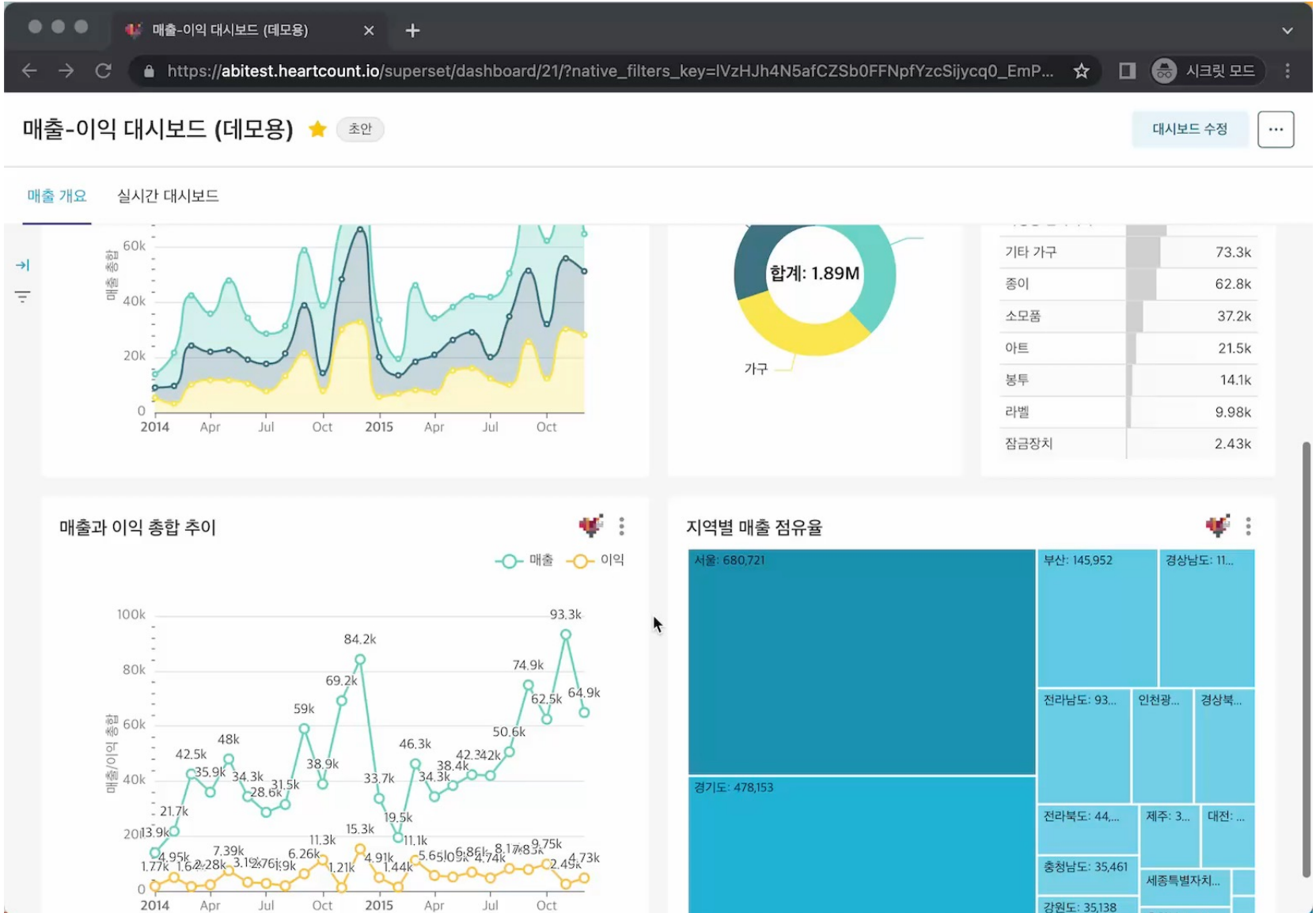
나뭇잎도 보고



Analytics 도구로 그때 그때 질문에
대한 정량적 답변 즉시 확인



B-to-the-A: 맥락 담긴 데이터셋으로 그때 그때 질문에 딱딱 답변





좋은 한데,

- SQL 배우도 데이터 추출 어려운데
- 데이터가 있어도 뭘 물어봐야 좋을지 모르겠네
- 주요 변화를 누가 알려줄 수 있다면

데이터 관련 질문-답변 자동화 Automating ad-hoc Question Answering

그때 그때 질문에
대한 답변

몇 개, 몇 명, 얼마? ▶ **Text-to-SQL**

왜 차이가 나니/하락했니, 어떻게 하면 좋겠니? ▶ **Dialogue**


지표에 뭐 이상한 거 없었니? ▶ **Signal**

내 삶이 바뀌는
데이터 경험

데이터 질문에 대한 답변하는 일의 생산성

Time-to-Insight ▶ **생각(Opinion) 빼고 모든 걸 자동화**

몇 개, 몇 명, 얼마? ▶ Text-to-SQL

질문	Counting	Why, How		
	Question-Answer	Questions	Q-Chart	C-Narrative
일	질문에 바로 정량적 답변	데이터에 질문	질문을 차트로 표현	차트에서 정량적 사실 도출
노력	데이터 부탁, 추출	분석적 추론	도구 학습	정량적 해석
	*LLM 자연어 SQL 변환	자체 알고리즘 다음 질문 제시	자체 알고리즘 질문에 상응하는 차트 표현	자체 알고리즘 차트에서 정량적 사실 도출

10월 5일 10:20



제품군별 지난 달 매출 총액,
고객수를 알려주실래요?

```
select
  product_category,
  state,

  sum(sales) as total_sales_amount,
  count(distinct customer_id) as number_of_customers

from sales
where date_trunc('year', sold_at) = 2021
group by 1, 2
```

product_category	state	year	total_sales_amount	number_of_customers
brewing supplies	NY	2021	612	27
brewing supplies	VT	2021	172	9
coffee beans	NY	2021	575	27
coffee beans	VT	2021	256	12
merchandise	NY	2021	415	16
merchandise	VT	2021	325	14

10월 5일 11:20



앗... 제품군 x 지역으로 해서
다시 알려주실래요?

10월 7일 12:20



일자별로 다시 부탁드립니다
될까요?

SQL 한 번 배워보시면..



The screenshot shows the HeartCount ABI 2.0 SQL Lab interface. The browser address bar displays `https://abi.nhwel.com/superset/sqlab/`. The page header includes navigation links for '대시보드', '차트', 'SQL Lab', '데이터', and '분석 서비스로 이동'. The main interface features a sidebar on the left with a database selection dropdown set to 'postgresql PostgreSQL_postgres', a schema dropdown set to 'mart', and a table selection dropdown containing 'order', 'pay', and 'user_count'. Below these are schema details for 'user_count' and 'pay'. The 'user_count' table has columns: '기준일' (DATE), '회원구분' (VARCHAR), '회사구분' (VARCHAR), '회사명' (VARCHAR), '주문자 성별' (VARCHAR), '주문자 연령대' (VARCHAR), '회원수' (INTEGER), and '가입수' (INTEGER). The 'pay' table has columns: '배치 통합일' (TIMESTAMP WITH TIME ZONE), '복지관 주문번호' (BIGINT), 'ASP 주문번호' (VARCHAR), '서비스타입' (VARCHAR), '기준일 타입' (VARCHAR), '기준일' (DATE), '주문일' (TIMESTAMP WITH TIME ZONE), and '취소일' (TIMESTAMP WITH TIME ZONE).

The central area is a text-to-SQL query editor with a text input field containing the placeholder text 'Enter a question and we'll answer it with a SQL.' and a 'GENERATE QUERY' button. Below the input field is a large empty text area for the generated SQL. At the bottom of the editor, there is a '실행하기' button, a '제한 행 수: 1 000' dropdown, a green timer showing '00:00:00.00', a '다른 이름으로 저장' button, and a '링크 복사' button. The bottom of the interface shows a '결과' tab and a '쿼리 실행 이력' tab.

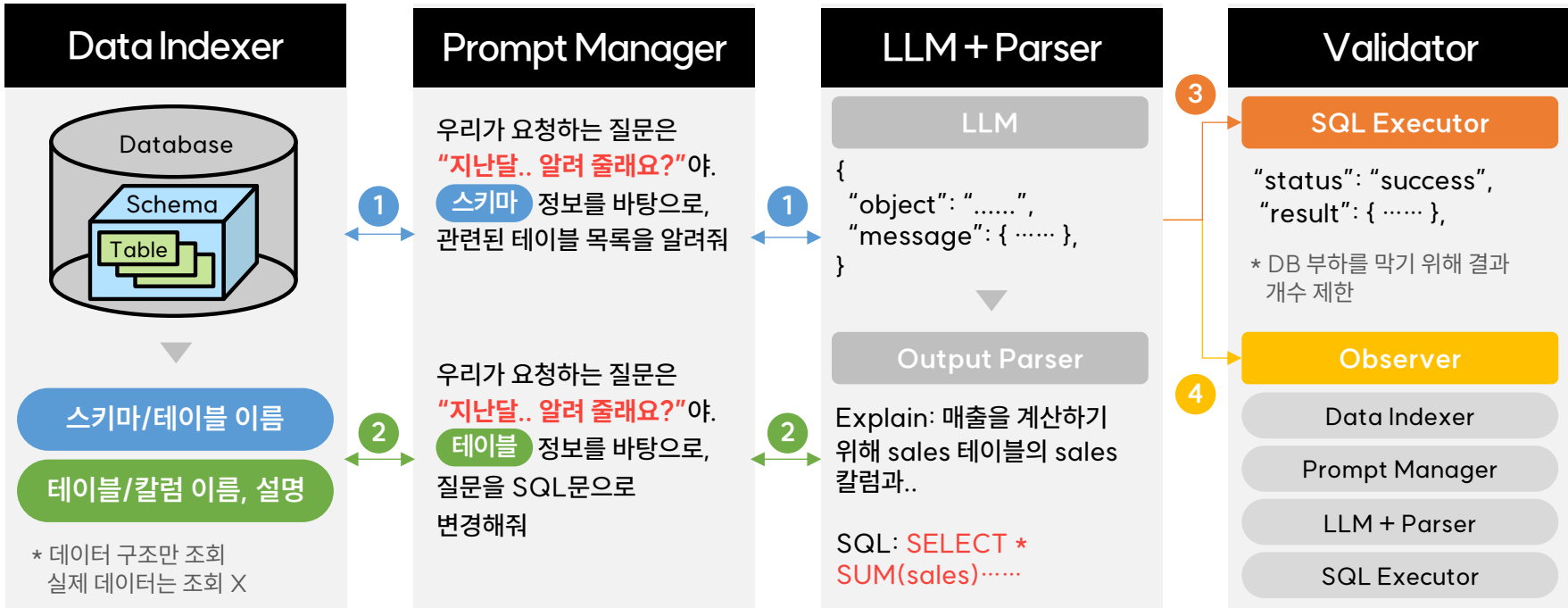


지난달 제품군별
매출 총액과 고객수를
일자별로 알려 주세요?

- 1 질문 관련 테이블 목록 추출
- 2 테이블/칼럼 정보 구조화, SQL 생성
- 3 SQL 실행 여부 확인 후 화면 표시
- 4 잘 안되면, 1~3을 반복

```
SELECT sum(sales) as
Total_sales
      , sum(profit) as Total_profit
FROM sales_table
WHERE year =
YEAR(CURDATE())
GROUP BY month;
```

Total_sales	Total_profit
100,000	50,000



왜 차이가 나니/하락했니, 어떻게 하면 좋겠니? ▶ Dialogue

질문	Counting	Why, How		
	Question-Answer	Questions	Question-Chart	Chart-Narrative
일	질문에 답변	데이터에 질문	질문을 차트로 표현	차트에서 정량적 사실 도출
노력	데이터 부탁, 추출	분석적 추론	도구 학습	정량적 해석
	*LLM 자연어 SQL 변환	자체 알고리즘 다음 질문 제시	자체 알고리즘 질문에 상응하는 차트 표현	자체 알고리즘 차트에서 정량적 사실 도출

LLM vs. Algorithm

	LLM	Algorithm
방식	통계 모형	로직
정확도		
확장성		
프로세스	-	(선형) 절차

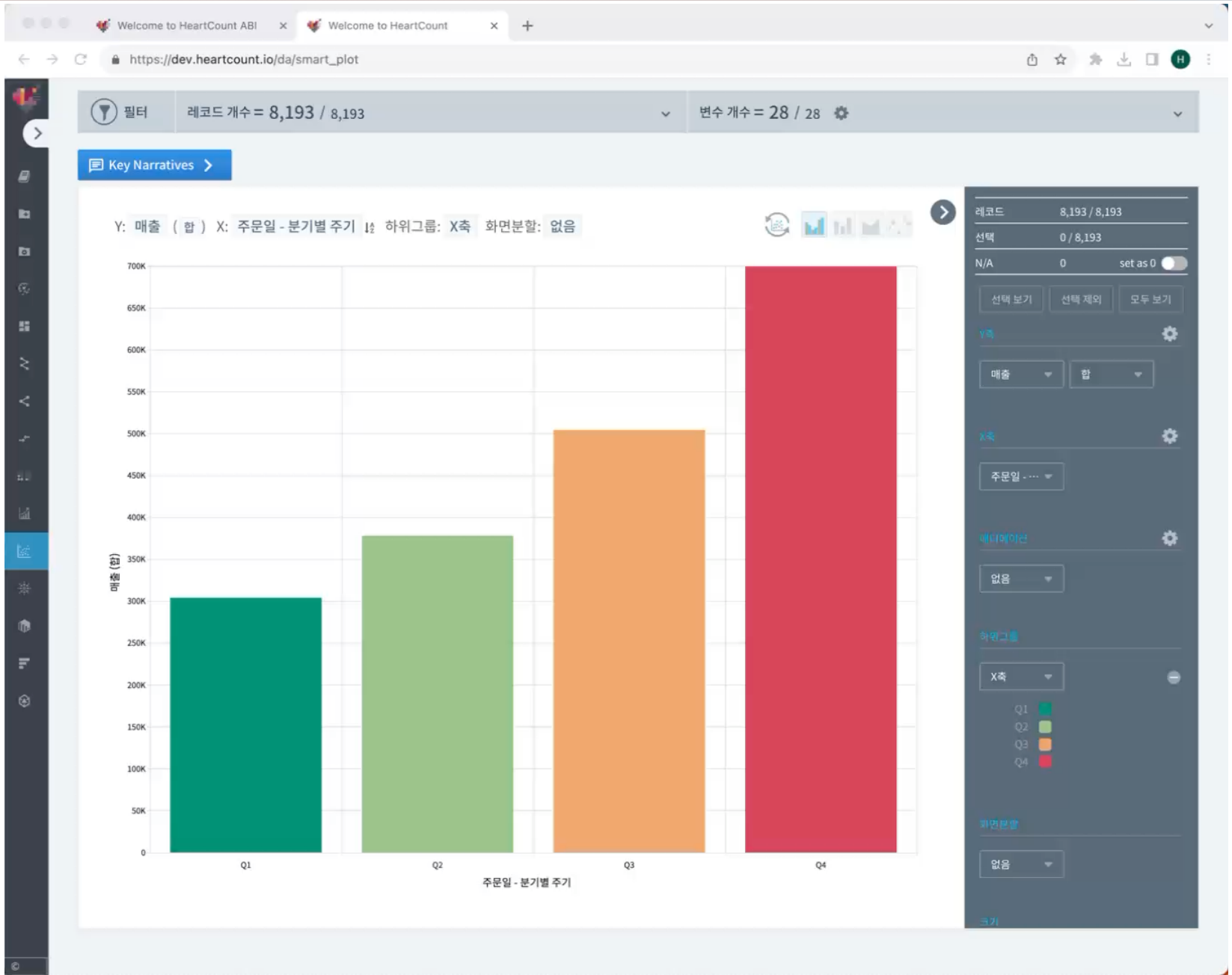
SOCRATES



Socratic Dialogue (소크라테스 대화법)

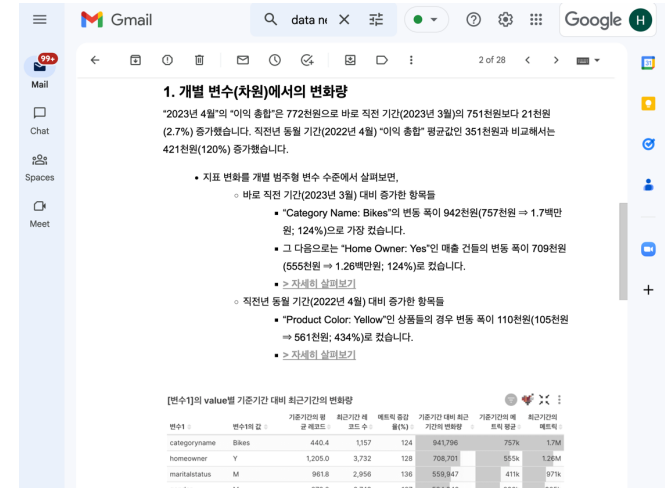
당신이 사실에 기반한 좋은 의견을 형성할 수 있으려면

- 분석적 추론, 데이터 문법에 익숙하지 않은 사람에게
- 좋은 의견을 구성할 수 있도록 정확한 사실, 새로운 관점 제공
- 미처 질문할 수 없었던 좋은 질문을 대신 물어 주는 것

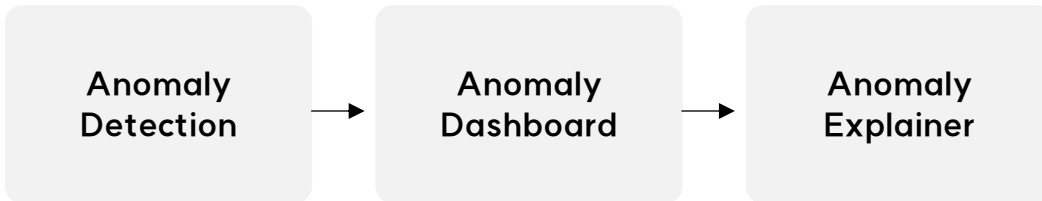


지표에 뭐 이상한 거 없었니? ▶ Signal

특정 Metrics 변화 요인에 대해 정기적으로 찾아가는 뉴스레터



모든 Metrics에 대해 이상변화 감지 그 주요 요인 설명



You received a **Signal**
Jan 30 at 6:58pm

[제품분류=의류] 주문총액이
2020-01-30일에 34%
급증하였습니다.
• 평균주문액: 9% ↔
• 주문건수: 31%. ↗



Between 01-01-2020 and
01-30-2020

[Explore](#)

누가 물어보기 전에 특정 Metrics 변화 요인에 대해 정기적으로 찾아가는 뉴스레터

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

누가 물어보기 전에 모든 Metrics에 대해 이상변화 감지 그 주요 요인 설명

범주별 손해액(kpi) 증감량 및 증감을 확인하기

이전_기간	현재_기간	요약	상세 대시보드	kpi_증감량	kpi_증감율	이전_레코드수	현재_레코드수
22년 11월	22년 12월	[사고_대분류: 질병]일 때 변동폭은 +28.01억원 (261.37 → 289.38억원, +10.7%) 입니다.	바로가기	2,801,191,700	10.7%	71,977	78,879
22년 11월	22년 12월	[담보_대분류: 건강]일 때 변동폭은 +21.04억원 (200.63 → 221.67억원, +10.5%) 입니다.	바로가기	2,103,863,600	10.5%	73,501	80,692
22년 11월	22년 12월	[담당_본부_id: 담당본부_A]일 때 변동폭은 +20.21억원 (145.10 → 165.31억원, +13.9%) 입니다.	바로가기	2,021,093,600	13.9%	35,424	38,966
22년 11월	22년 12월	[담보_중분류: 의료비]일 때 변동폭은 +19.72억원 (184.07 → 203.80억원, +10.7%) 입니다.	바로가기	1,972,220,500	10.7%	67,574	74,134
22년 11월	22년 12월	[담보_대분류: 질병]일 때 변동폭은 +9.99억원 (91.46 → 101.45억원, +10.9%) 입니다.	바로가기	999,248,640	10.9%	7,630	8,269
22년 11월	22년 12월	[담보_대분류: 상해]일 때 변동폭은 +8.85억원 (57.40 → 66.25억원, +15.4%) 입니다.	바로가기	885,409,000	15.4%	7,545	8,507
22년 11월	22년 12월	[사고_중분류: 기타질환]일 때 변동폭은 +8.46억원 (61.17 → 69.63억원, +13.8%) 입니다.	바로가기	845,754,240	13.8%	27,799	30,528
22년 11월	22년 12월	[사고_대분류: 상해]일 때 변동폭은 +8.04억원 (61.28 → 69.32억원, +13.1%) 입니다.	바로가기	804,291,140	13.1%	10,995	12,187
22년 11월	22년 12월	[사고_중분류: 일상생활중]일 때 변동폭은 +8.04억원 (61.28 → 69.32억원, +13.1%) 입니다.	바로가기	804,022,900	13.1%	10,995	12,186
22년 11월	22년 12월	[사고_중분류: 암]일 때 변동폭은 +5.79억원 (61.72 → 67.51억원, +9.4%) 입니다.	바로가기	578,508,100	9.4%	3,005	3,258
22년 11월	22년 12월	[담보_중분류: 진단비]일 때 변동폭은 +5.29억원 (55.97 → 61.25억원, +9.4%) 입니다.	바로가기	528,600,830	9.4%	3,724	4,147
22년 11월	22년 12월	[사고_대분류: 교통]일 때 변동폭은 +5.28억원 (34.52 → 39.80억원, +15.3%) 입니다.	바로가기	527,860,200	15.3%	5,974	6,726
22년 11월	22년 12월	[담당_본부_id: 담당본부_J]일 때 변동폭은 +5.04억원 (30.71 → 35.75억원, +16.4%) 입니다.	바로가기	504,133,980	16.4%	7,203	7,911
22년 11월	22년 12월	[담보_중분류: 후유장애]일 때 변동폭은 +4.85억원 (11.82 → 16.66억원, +41.0%) 입니다.	바로가기	484,840,380	41.0%	112	156
22년 11월	22년 12월	[담당_본부_id: 담당본부_D]일 때 변동폭은 +4.45억원 (31.63 → 36.09억원, +14.1%) 입니다.	바로가기	445,416,860	14.1%	8,233	9,065
22년 11월	22년 12월	[담보_중분류: 사망]일 때 변동폭은 +4.21억원 (15.24 → 19.45억원, +27.6%) 입니다.	바로가기	420,751,620	27.6%	48	58
22년 11월	22년 12월	[담당_본부_id: 담당본부_E]일 때 변동폭은 +3.30억원 (26.87 → 30.17억원, +12.3%) 입니다.	바로가기	329,832,100	12.3%	6,216	6,822
22년 11월	22년 12월	[담보_중분류: 수술비]일 때 변동폭은 +3.09억원 (36.11 → 39.20억원, +8.6%) 입니다.	바로가기	309,297,730	8.6%	6,049	6,699
22년 11월	22년 12월	[사고_중분류: 심질환]일 때 변동폭은 +2.94억원 (9.59 → 12.53억원, +30.7%) 입니다.	바로가기	294,170,600	30.7%	1,216	1,313
22년 11월	22년 12월	[사고_중분류: 운전중]일 때 변동폭은 +2.93억원 (27.27 → 30.20억원, +10.7%) 입니다.	바로가기	292,586,900	10.7%	4,580	5,165
22년 11월	22년 12월	[담당_본부_id: 담당본부_B]일 때 변동폭은 +2.72억원 (31.33 → 34.06억원, +8.7%) 입니다.	바로가기	272,299,580	8.7%	7,374	8,133

감사합니다