



一品 데이터 품질 관리자 박삼이의 이커머스 데이터 품질 검사

B01 | 박나영, 이소희(발표자), 이정희(팀장), 이해원

목차



데이터 품질 검사의 필요성



이커머스 데이터 품질 검사 Part 1



이커머스 데이터 품질 검사 Part 2



부록

대시보드를 통한 데이터 품질 트래킹



문제점

의사결정 오류

고객 이해 부족

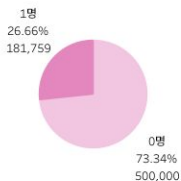
시장 트렌드 오해

데이터 품질 점검 대시보드

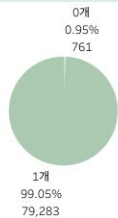
B01조 박삼이

세션별 유저수 2명 이상	user별 city수 2개 이상	일자별 user_id 없는 이벤트의 수	session 시간 30분 이상 session수	일자(user_id 없는 이벤트 수) 2022년 6월 30일
0	0	646	75,357	

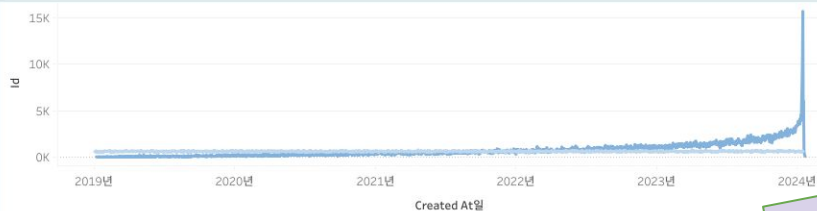
1. 세션별 유저수



2. 유저별 도시수

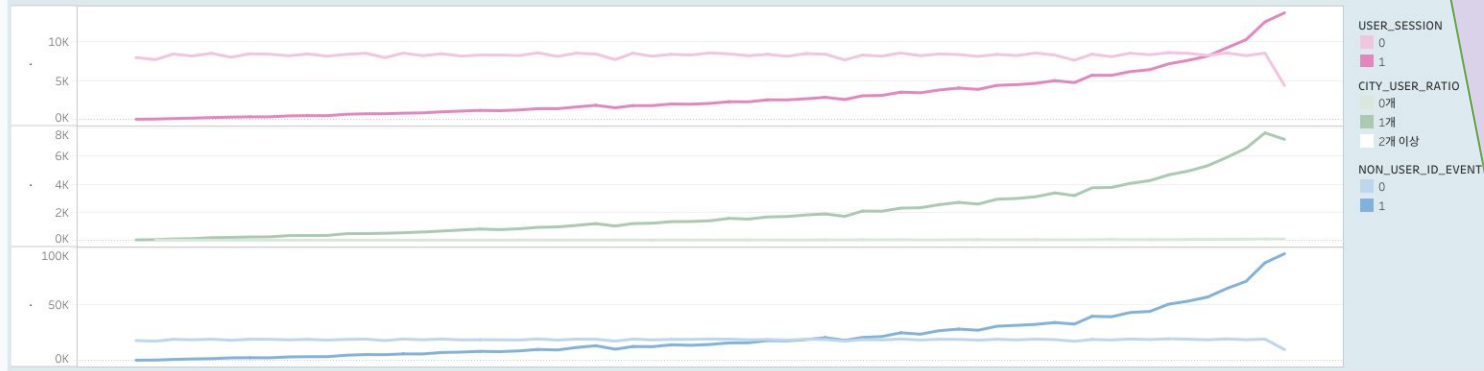


3. 일자별 user_id가 없는 이벤트 수



세션별 유저 수와
유저별 도시 수는
이상치가 없었음

4. 1-3 통합

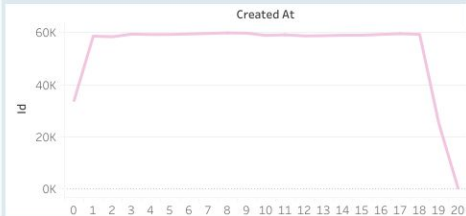


일자별 user id가
없는 이벤트는
500~700건
사이의
일관적인
값으로 확인

데이터 품질 점검 대시보드

B01조 박삼이

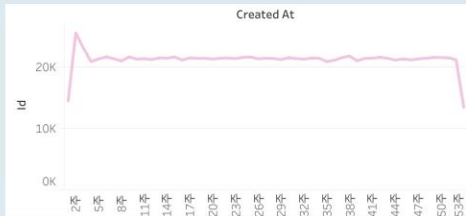
3. 시간대별 user_id가 없는 이벤트 수



3. 요일별 user_id가 없는 이벤트 수



3. 주별 user_id가 없는 이벤트 수



NON_USER_ID...

☒ 0☐ 1

Browser

☐ Chrome☐ Firefox☐ IE☐ Other☐ Safari

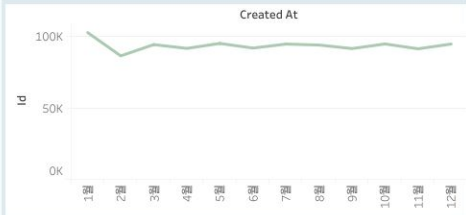
Traffic Source

☐ Adwords☐ Email☐ Facebook☐ Organic☐ YouTube

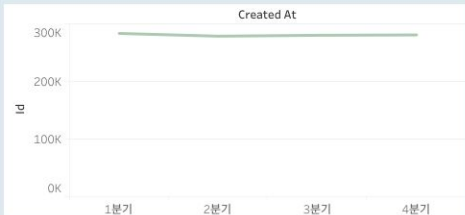
Event Type

☐ cancel☐ cart☐ department☐ product

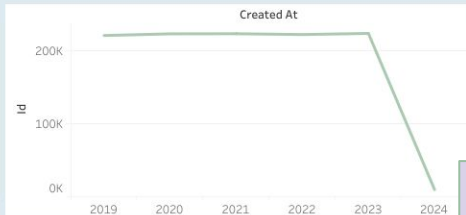
3. 월별 user_id가 없는 이벤트 수



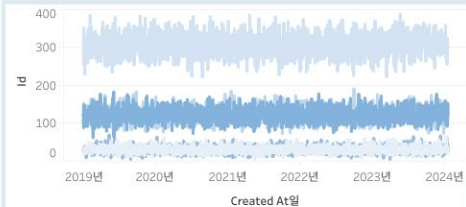
3. 분기별 user_id가 없는 이벤트 수



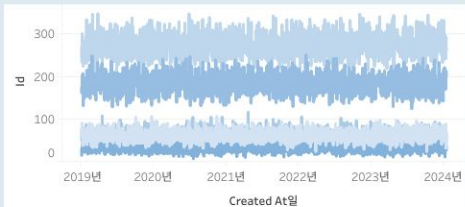
3. 연도별 user_id가 없는 이벤트 수



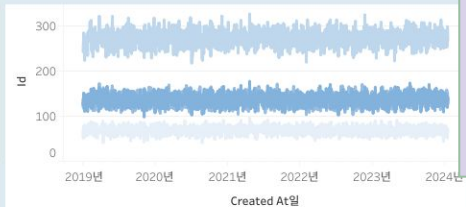
3. browser별 user_id가 없는 이벤트 수



3. traffic_source별 user_id가 없는 이벤트 수



3. event_type별 user_id가 없는 이벤트 수

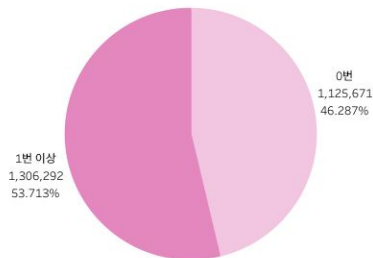


home, purchase 외
나머지 event type 모두
시스템 오류로 추정!
엔지니어팀 확인 요청

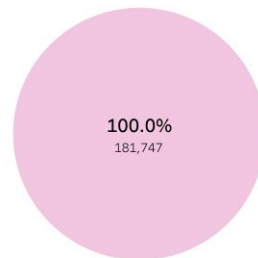
데이터 품질 점검 대시보드

B01조 박삼이

5. Event가 발생하지 않는 유저/상품 쌍 비율

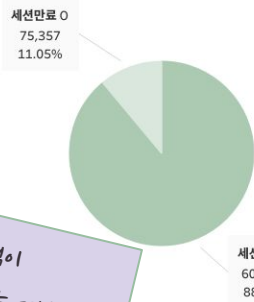


6. Event 시간 < Purchase 시간



세션 종료는
event_type이
purchase인
경우에만 발생

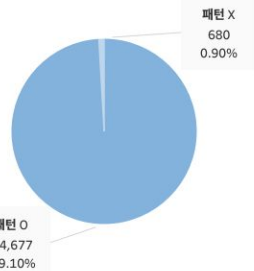
7. Session 시간 30분 이상인 session 수



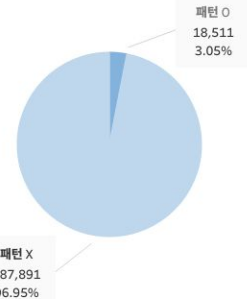
7. Session 시간 30분 이상인 경우 event type



8. 7번 항목 '참'값 중 event type 중복 비중



8. 7번 항목 '거짓'값 중 event type 중복 비중



세션 간격이
30분 이상 넘어가는 경우는
약 11%

30분 초과로 인한 세션 종료 원인 분석



패턴 발견

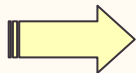
med:	session_id	created_at	event_type	product_id
0				
0	0002d566-70ed-44d8-8213-33942ffc0899	2021-10-04 22:13:42+00:00	department	NaN
1	0002d566-70ed-44d8-8213-33942ffc0899	2021-10-04 22:16:22+00:00	product	<u>6044.0</u>
2	0002d566-70ed-44d8-8213-33942ffc0899	2021-10-04 22:17:11+00:00	cart	NaN
3	0002d566-70ed-44d8-8213-33942ffc0899	2021-10-04 22:17:14+00:00	department	NaN
4	0002d566-70ed-44d8-8213-33942ffc0899	2021-10-04 22:17:52+00:00	product	<u>6044.0</u>
5	0002d566-70ed-44d8-8213-33942ffc0899	2021-10-04 22:20:43+00:00	cart	NaN
6	0002d566-70ed-44d8-8213-33942ffc0899	2021-10-07 22:22:26+00:00	purchase	NaN

17	00059a24-cd2e-4c6b-9292-2926d12f7fed	2023-03-30 04:53:53+00:00	department	NaN
18	00059a24-cd2e-4c6b-9292-2926d12f7fed	2023-03-30 04:55:28+00:00	product	<u>24647.0</u>
19	00059a24-cd2e-4c6b-9292-2926d12f7fed	2023-03-30 04:58:12+00:00	cart	NaN
20	00059a24-cd2e-4c6b-9292-2926d12f7fed	2023-03-30 04:59:08+00:00	department	NaN
21	00059a24-cd2e-4c6b-9292-2926d12f7fed	2023-03-30 05:00:35+00:00	product	<u>24647.0</u>
22	00059a24-cd2e-4c6b-9292-2926d12f7fed	2023-03-30 05:00:36+00:00	cart	NaN
23	00059a24-cd2e-4c6b-9292-2926d12f7fed	2023-03-30 05:01:32+00:00	department	NaN
24	00059a24-cd2e-4c6b-9292-2926d12f7fed	2023-03-30 05:01:53+00:00	product	<u>24647.0</u>
25	00059a24-cd2e-4c6b-9292-2926d12f7fed	2023-03-30 05:04:20+00:00	cart	NaN
26	00059a24-cd2e-4c6b-9292-2926d12f7fed	2023-03-31 05:05:25+00:00	purchase	NaN



이커머스 데이터 품질 검사 Part 2

패턴 발견



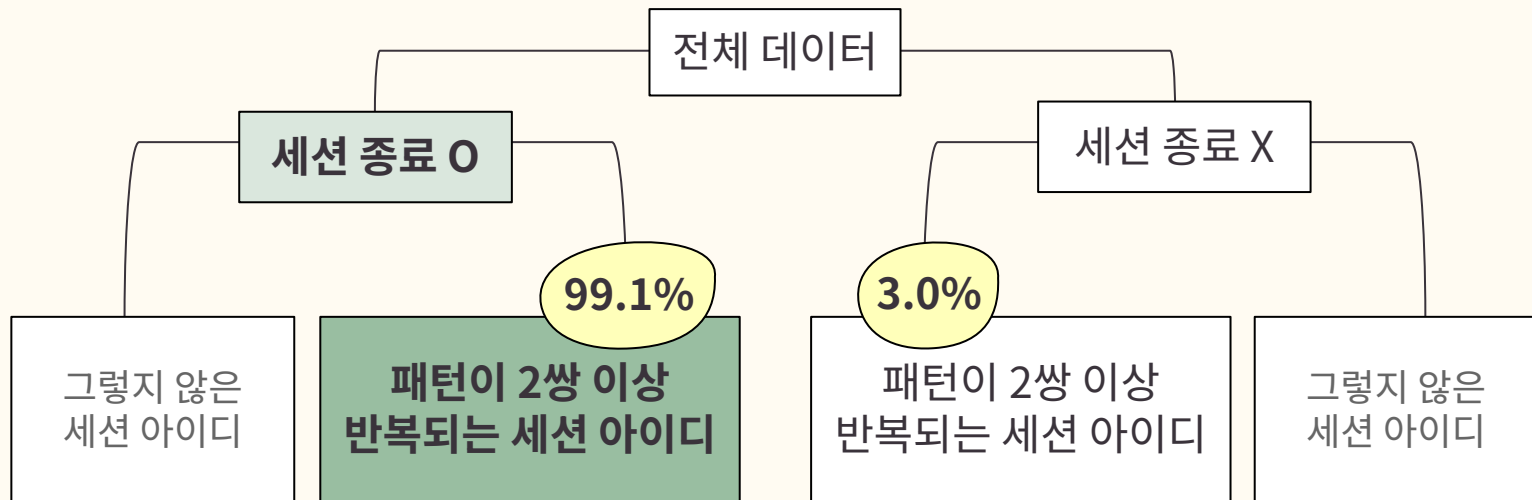
세션 종료 여부별 패턴 비율 비교

```
# 패턴이 2개 이상이고, 모든 product_id가 같은 session id 빈 리스트 만들어놓기
over2_pattern_same_product_sessions = []

# 패턴이 2개 이상인 세션 아이디 찾기
for session_id, i in final.groupby('session_id'):
    event_types = ''.join(i['event_type'])

    # 패턴이 2개 이상인 경우
    if event_types.count('departmentproductcart') >= 2:
        # product_id의 고유값 수가 1인 경우 (=product id가 모두 같음)
        product_id_countd = i[i['event_type'] == 'product']['product_id'].unique()
        if len(product_id_countd) == 1:
            over2_pattern_same_product_sessions.append(session_id)

# 해당하는 세션 아이디들의 데이터프레임 행을 뽑기
df_over2pattern_sameproduct = final[final['session_id'].isin(over2_pattern_same_product_sessions)]
```

용어 정리

세션 종료 O	세션 간격이 30분을 초과하여 세션이 종료된 경우	세션 종료 X	세션 간격이 30분을 초과한 경우가 없음
패턴	이벤트 타입 (department, product, cart) 1쌍		



이커머스 데이터 품질 검사 Part 2

event type이 purchase인 경우에만 세션 종료가 발생하는 원인

가설 1 시스템 오류

- 패턴(department, product, cart)이 한 번 이루어지는 데 약 4~5분 소요됨
- 이 패턴이 2번 이상 반복됨에 따라 session 간격이 30분을 넘었을 것

가설 2 소비자 행동

- 소비자가 구매를 고민하는 과정(cart에 넣었다 뺐다를 반복)의 시간이 session 시간에 포함
- event_type인 purchase에서의 session 간격이 30분을 넘었을 것

*현재 박삼이 팀이 가지고 있는 dataset 바탕으로 추가적인 확인이 불가했음

패턴이 2쌍 이상 로그 데이터로 입력됨



개발/엔지니어링 단계 재검토

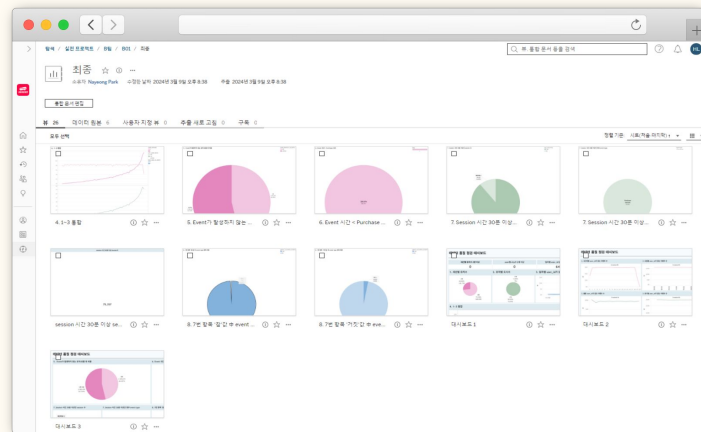
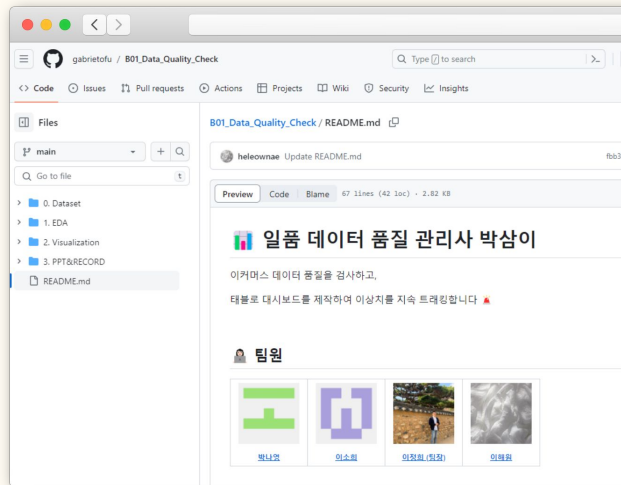
created_at	event_type	product_id	browser	traffic_source	city	date_diff
2021-10-04 22:13:42+00:00	department	NaN	Chrome	Adwords	Sidrolândia	0 days 00:00:00
2021-10-04 22:16:22+00:00	product	6044.0	Chrome	Adwords	Sidrolândia	0 days 00:02:40
2021-10-04 22:17:11+00:00	cart	NaN	Chrome	Adwords	Sidrolândia	0 days 00:00:49
2021-10-04 22:17:14+00:00	department	NaN	Chrome	Adwords	Sidrolândia	0 days 00:00:03
2021-10-04 22:17:52+00:00	product	6044.0	Chrome	Adwords	Sidrolândia	0 days 00:00:38
2021-10-04 22:20:43+00:00	cart	NaN	Chrome	Adwords	Sidrolândia	0 days 00:02:51
2021-10-07 22:22:26+00:00	purchase	NaN	Chrome	Adwords	Sidrolândia	3 days 00:01:43

고객의 구매 결정까지의 여정



제품, 마케팅, UI 등 다각도 고려

created_at	event_type	product_id	browser	traffic_source	city	date_diff	date_diff_seconds	boolean_time_diff_30
2021-10-04 22:13:42+00:00	department	NaN	Chrome	Adwords	Sidrolândia	0 days 00:00:00	0.0	False
2021-10-04 22:16:22+00:00	product	6044.0	Chrome	Adwords	Sidrolândia	0 days 00:02:40	160.0	False
2021-10-04 22:17:11+00:00	cart	NaN	Chrome	Adwords	Sidrolândia	0 days 00:00:49	49.0	False
2021-10-04 22:17:14+00:00	department	NaN	Chrome	Adwords	Sidrolândia	0 days 00:00:03	3.0	False
2021-10-04 22:17:52+00:00	product	6044.0	Chrome	Adwords	Sidrolândia	0 days 00:00:38	38.0	False
2021-10-04 22:20:43+00:00	cart	NaN	Chrome	Adwords	Sidrolândia	0 days 00:02:51	171.0	False
2021-10-07 22:22:26+00:00	purchase	NaN	Chrome	Adwords	Sidrolândia	3 days 00:01:43	259303.0	True



The background is a light cream color, decorated with several abstract geometric elements. In the top left, there is a cluster of yellow dots. To its right is a large yellow circle. Further right is a pink ring with blue diagonal lines passing through it. Below the yellow dots is a pink wavy line. On the left side, there is a blue curved shape. In the bottom left, there is a light green ring with pink diagonal lines passing through it. At the bottom center, there is a cluster of blue dots. On the right side, there is a yellow curved shape and a light green wavy line.

감사합니다.