E-Commerce 데이터 분석

이중훈

DATA INTRODUTION

문제 정의 및 활용

문제 정의

1) Olist 데이터 분석을 통해 이커머스의 매출 감소, 지역별 판매 불균형, 시간대별 매출 집중, 제품 카테고리 간 매출 편차 등 주요 비즈니스 문제를 파악.

2) RFM 분석을 통해 고객을 세분화하고, 각 세그먼트의 행동 및 매출 기여도를 분석하여, 충성 고객 유지 및 이탈 고객 회복이 필요.

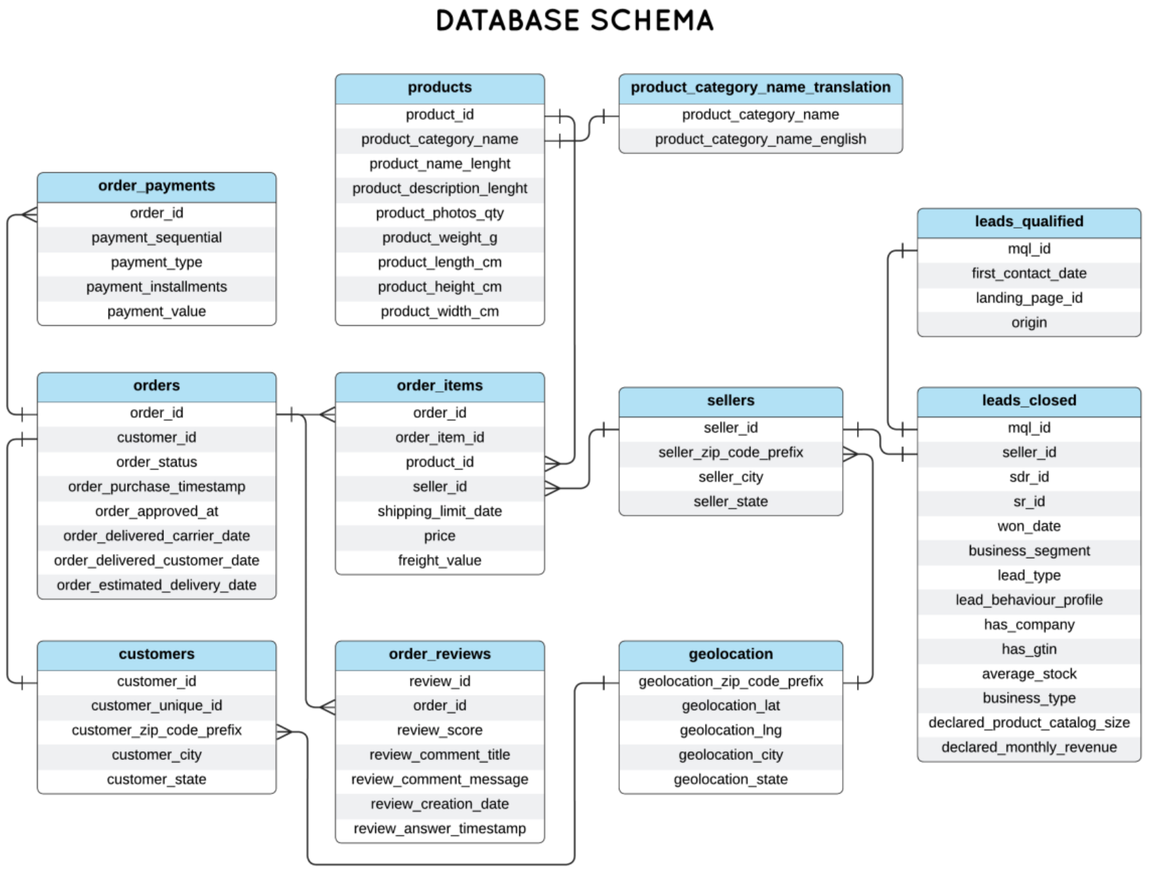
활용 목표

데이터 기반으로 고객 행동과 매출 성과를 분석하여, 판매 성과의 주요 요인을 파악하고 의사결정을 지원하는 인사이트를 도출하여 고객 유지, 매출 증대, 및 운영 효율성 개선을 달성.

DATA INTRODUTION

데이터 소개

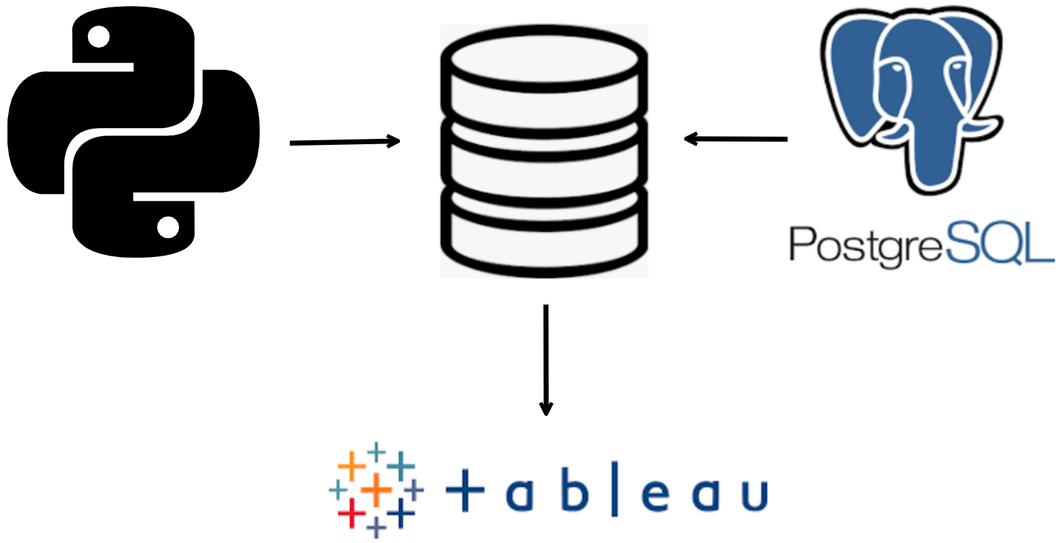
활용 데이터 셋은 Kaggle에 있는 Olist E-Commerce 데이터 셋으로 브라질 전자상거래 데이터입니다.
2016년부터 2018년까지 Olist에서 이루어진 약 10만 건의 주문 정보가 포함된 데이터로 데이터 스키마는 다음과 같습니다.



ANALYSTIC METHOD

분석 방법

결측치 및 오류값은 파이썬으로 전처리한 후 DB에 데이터를 적재하고 SQL문으로 분석에 필요한 지표 및 수치 데이터를 가공.
태블로와 연동하여 시각화 및 대시보드 생성 진행.



ANALYSTIC METHOD

분석 목적

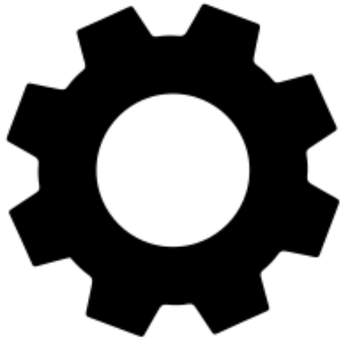
개요 :
Olist 데이터를 분석하는 가장 중요한 목적 중
하나는 고객 행동을 이해하는 것입니다.

목적 :
고객의 구매 패턴 및 행동 이해.
고객 세그먼트 정의 및 맞춤형 마케팅 전략 개발.

활용 지표 :
- 지역별 구매금액, 구매건수, 구매유저수, 건당 구매 금액 추세
- 특정 시간대나 요일별 구매 트렌드
- 고객의 행동 및 가치를 평가할 수 있는 지표
- RFM 분석을 통한 고객 세분화 지표

DATA PREPROCESSING

데이터 전처리



01 중복값 제거 및 주요 변수 생성

데이터 예시

order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_
e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:5
e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:5
e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-10-04 19:5

is_revenue	order_status	customer_unique_id
y	shipped	1008
	invoiced	300
	delivered	94488
	approved	2
n	unavailable	6
	processing	284
	canceled	428

중복값 제거

customer_id와 order_id의 조합은 고유해야하므로 중복제거.
데이터 수 : 115609 -> 96516

order_status를 통해 매출 포함여부에 따라 변수 생성

1. 고객의 주문이 실제 매출로 이어진 상태

상세 설명
approved (승인): 주문이 승인되어 매출로 연결될 가능성이 높음.
delivered (배송 완료): 고객에게 상품이 전달되었으므로 확정 매출로 간주됨.
invoiced (송장 발행): 결제가 완료되고 송장이 발행된 상태. 대부분 매출로 기록됨.
shipped (배송 중): 매출로 집계할 가능성이 있음(특히 배송 시점에 매출로 기록하는 경우).

2. 주문이 최종 결제에 포함되지 않거나, 매출이 취소된 상태.

상세설명
canceled (취소): 고객 요청 또는 기타 사유로 주문이 취소되어 매출에서 제외됨.
created (생성): 주문이 생성되었지만 승인이나 결제가 진행되지 않은 초기 단계. 매출 확정 아님.
processing (처리 중): 주문 처리 단계로, 아직 매출 확정 전 상태.
unavailable (처리 불가): 주문이 불가능하거나 실패한 상태로 매출에 포함되지 않음.

최종 데이터 : 96516개

02 중복값 제거 및 주요 변수 생성

HAVERSINE 공식을 사용한 거리 계산

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)$$

$$c = 2 \cdot \operatorname{atan2}\left(\sqrt{a}, \sqrt{1-a}\right)$$

$$d = R \cdot c$$

데이터 예시

customer_zip_code	customer_city	seller_zip_code	seller_city	customer_lat	customer_lng	seller_lat	seller_lng	distance_km
7,197	guarulhos	31,310	belo horizonte	-23.4602659008	-46.5155054656	-19.873399788	-43.9823259355	477.0389630893
30,830	belo horizonte	31,310	belo horizonte	-19.9025390571	-44.0032852095	-19.873399788	-43.9823259355	3.9117099637
9,895	sao bernardo do campo	31,310	belo horizonte	-23.7044274467	-46.5701770548	-19.873399788	-43.9823259355	502.8139954909
44,059	feira de santana	31,310	belo horizonte	-12.2042864655	-38.945165069	-19.873399788	-43.9823259355	1,008.1804121297
30,535	belo horizonte	31,310	belo horizonte	-19.9238846267	-43.9914588157	-19.873399788	-43.9823259355	5.6942942788
22,730	rio de janeiro	31,310	belo horizonte	-22.9200024032	-43.3695755323	-19.873399788	-43.9823259355	344.6534330459
85,812	cascavel	31,310	belo horizonte	-24.9493591573	-53.4479202135	-19.873399788	-43.9823259355	1,124.3250959148
22,251	rio de janeiro	31,310	belo horizonte	-22.9453439401	-43.185010682	-19.873399788	-43.9823259355	351.4120737166
65,980	carolina	31,310	belo horizonte	-7.3317411256	-47.4686511282	-19.873399788	-43.9823259355	1,444.3412236863
5,024	sao paulo	31,310	belo horizonte	-23.5301415778	-46.6854093556	-19.873399788	-43.9823259355	493.2348420069
81,312	curitiba	31,310	belo horizonte	-25.5042277727	-49.3094262727	-19.873399788	-43.9823259355	830.8409587351

구매자와 판매자간 거리 변수를 생성

수식 설명

- φ_1, φ_2: 두 지점의 위도 (라디안 단위로 변환 필요)
- λ_1, λ_2: 두 지점의 경도 (라디안 단위로 변환 필요)
- $\Delta\varphi = \varphi_2 - \varphi_1$: 위도의 차이
- $\Delta\lambda = \lambda_2 - \lambda_1$: 경도의 차이
- R: 지구 반지름 (평균값: 6371 km)

geolocation테이블에 zip_code,state,city,경도,위도 등의 정보가 포함되어있어 거래데이터에 매핑시켜 구매자와 판매자 간의 거리(Km)를 계산하여 distance_km 변수를 생성함.

```
def haversine(lat1, lon1, lat2, lon2):
    R = 6371.0  # 지구 반지름 (km 단위)

    # 위도와 경도를 라디안으로 변환
    lat1_rad, lon1_rad = map(math.radians, [lat1, lon1])
    lat2_rad, lon2_rad = map(math.radians, [lat2, lon2])

    # 위도 및 경도 차이 계산
    dlat = lat2_rad - lat1_rad
    dlon = lon2_rad - lon1_rad

    # Haversine 공식 적용
    a = math.sin(dlat / 2)**2 + math.cos(lat1_rad) * math.cos(lat2_rad) * math.sin(dlon / 2)**2
    c = 2 * math.atan2(math.sqrt(a), math.sqrt(1 - a))
    distance = R * c

    return distance
```

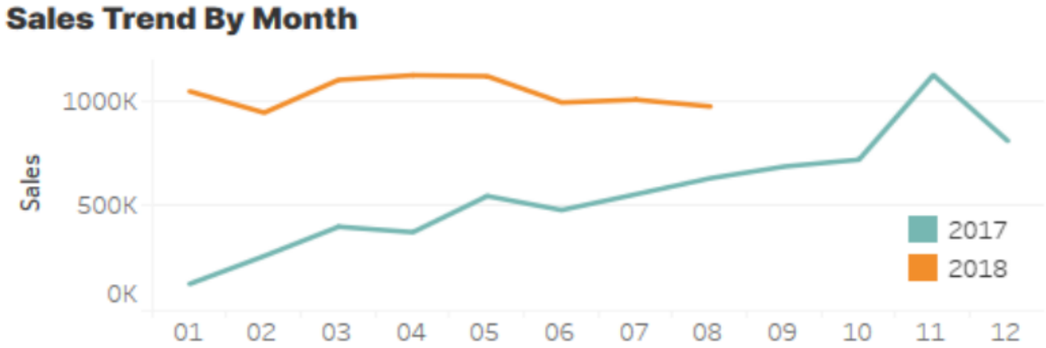
COUNTRY ANALYSIS

성과 지표 분석
문제 파악 및 방안

목적
특정 기간 동안 전기간 대비 성과 비교와 성장 또는 감소 요인을 분석
카테고리별 판매 데이터를 기반으로 시장 트렌드 및 수요 파악
프로세스 개선을 통한 서비스 사용자의 만족도 상승

대시보드 링크 :

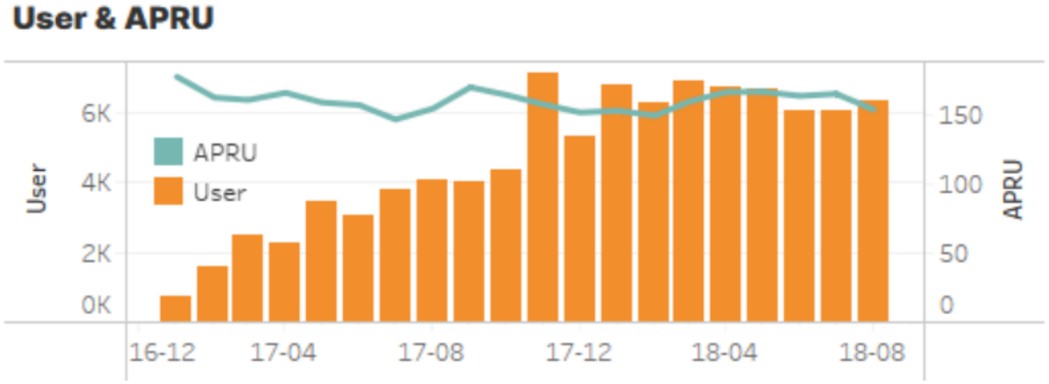
01 기간별 지표 추이



월별 매출 추이

2017년은 11월까지 꾸준히 매출이 상승하다가 12월에 소폭 감소함.(11월에 블랙프라이데이가 포함되어 매출이 급상승한것으로 나타남.)

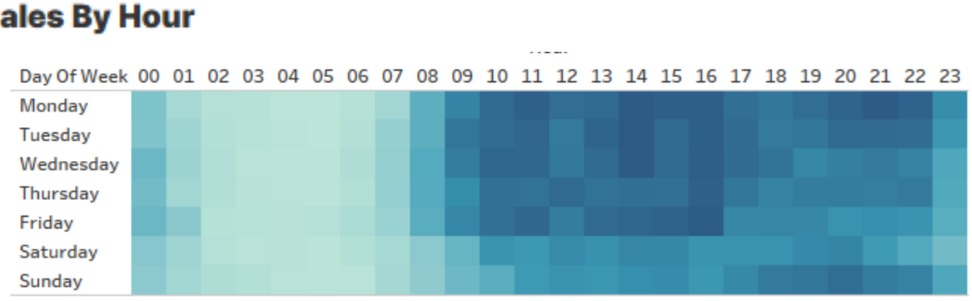
2018년은 전체적으로 매출이 2017년 보다 높지만 2분기에 상승하다가 3분기들어 감소하는 추세를 보임.



유저수 X APRU별 매출

유저수는 꾸준히 상승 추세지만, APRU가 감소 추세를 보이는 것으로 보아 신규 유저의 대부분이 저수익 고객일 가능성이 있음.

단순히 유저 수를 늘리는 것보다, 높은 지출을 하는 유저의 비율을 늘리는 전략이 필요.
(ARPU를 늘리기 위해 고가 상품이나 업셀링(업그레이드 구매) 기회를 제공할 수 있어야함.)



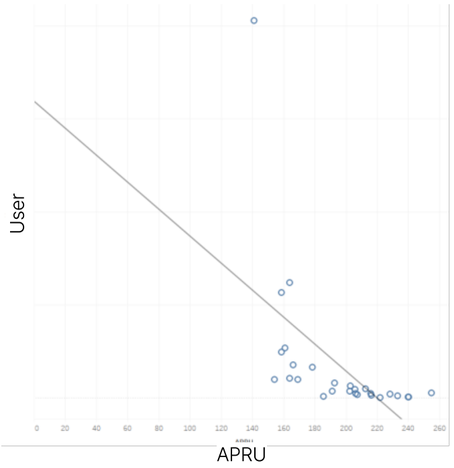
요일 X 시간대별 매출

전체적인 매출은 월~금의 매출이 토,일 매출보다 대체로 높았음.
평일에는 약 10시부터 16시까지 매출이 가장 높은 것으로 나타남.
주말에는 약 18시부터 22시까지 매출이 가장 높은것으로 나타남

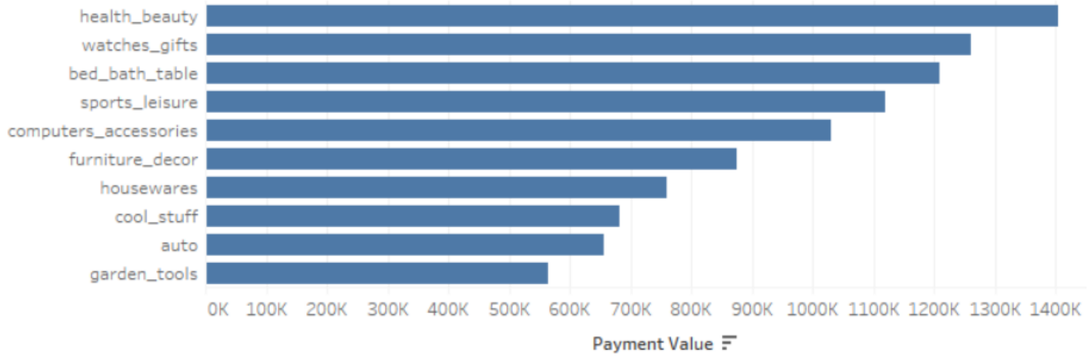
02 주별 매출 및 카테고리 분석

Sales By State

User&APRU



Top 10 Sales Category



주별 매출

주별 매출은 상파울로주가 5,655,665헤알로 가장 높았고, 호라이마주가 9,647헤알로 가장 낮았음.

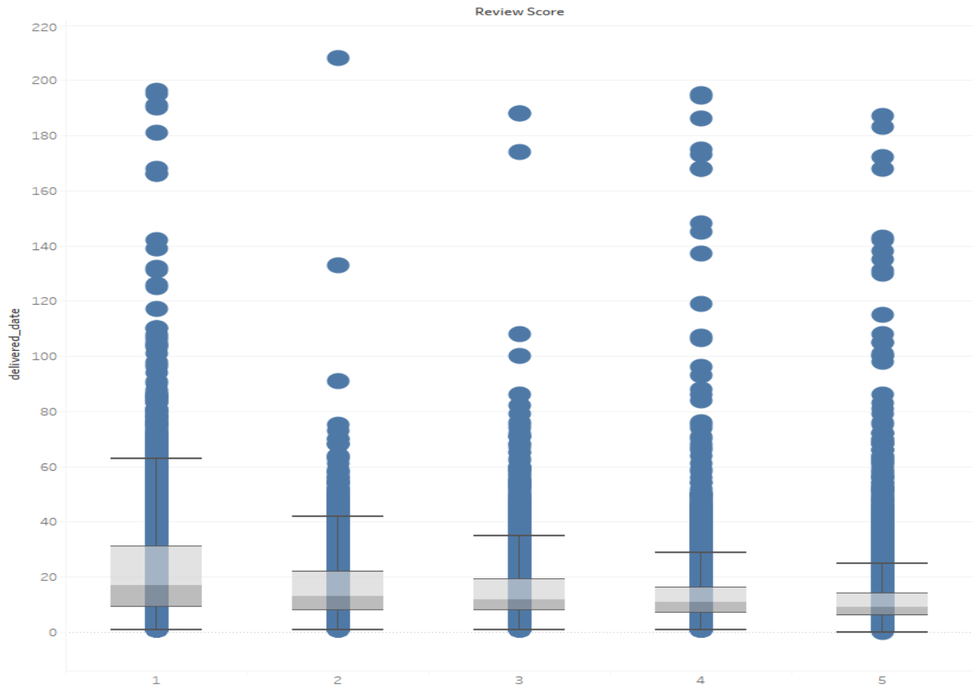
그러나 상파울로주는 APRU가 가장 낮아 고객당 구매 금액이 140.5헤알로 가장 낮은 것으로 나타남.
APRU가 가장 높은 주는 파라이바주로 254.5헤알.

카테고리별 매출

Health & Beauty 카테고리가 가장 높은 매출을 기록하고 있으며, 소비자들에게 가장 인기 있는 제품군임을 나타냅니다.
Watches & Gifts와 Bed, Bath & Table 카테고리가 그 뒤를 이어 높은 매출을 차지하고 있으며, 선물용 제품이나 생활용품의 수요가 강세를 보이고 있음을 보여줍니다.

상위 카테고리의 매출 기여도를 보면, 개인적인 소비(뷰티, 패션)와 생활 필수품(침구 및 가구)에 대한 수요가 강하다는 것을 확인할 수 있습니다.

03 문제 정의 및 파악



리뷰점수	1점	2점	3점	4점	5점
평균 배송기간	20.9일	16.1일	13.8일	11.8일	10.2일

리뷰 점수별 배송 기간의 박스플롯

각 리뷰 점수(1~5점)에 대한 배송 기간의 분포가 박스플롯으로 표시하였음.

점수별 중앙값 → 1점 : 9일 / 2점 : 10일 / 3점 : 12일 / 4점 : 13일 / 5점 : 16일
리뷰 점수가 낮을수록 배송 기간의 중앙값과 상한선이 높은 경향을 보이며,
점수가 높아질수록 배송 기간의 중앙값이 낮아지는 경향을 보입니다.

1점에서는 배송 기간이 평균 20.9일로 가장 길었으며, 데이터의 분산도 컸음.
5점에서는 배송 기간이 평균 10.2일로 가장 짧으며 데이터 분산도 작은 것으로 나타남.

또한 리뷰점수와 배송기간의 상관관계는 0.33으로 배송기간이 리뷰점수에 어느정도 영향을 미친다고 볼 수 있음.

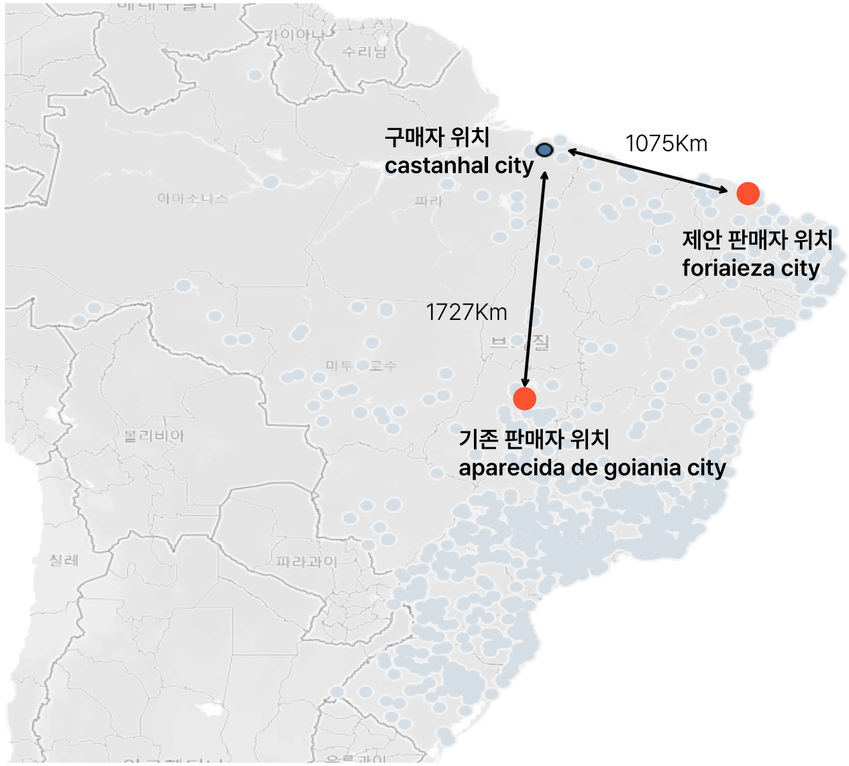
결론 및 활용 방안

배송 지연의 부정적인 영향 : 1점과 2점에서 배송 기간이 길거나 변동 폭이 큰 점을 보면, 배송 지연이 고객 불만족으로 이어지는 가능성이 있습니다.

배송 기간 단축은 리뷰 점수를 개선할 수 있는 중요한 요인으로 보입니다. 배송기간이 길어지는 원인을 파악하여 문제를 개선하는데 초점을 둠.

03 문제점 개선 방안

구매자와 판매자 간 거리 지도



기존 castahal의 baby카테고리 배송지 및 배송거리

seller_state	seller_city	geolocation_lat_y	geolocation_lng_y	cust_lat	cust_lng	distance_km
GO	aparecida de goiania	-16.766552	-49.339375	-1.29866	-47.898252	1727.169212

제안 castahal의 baby카테고리 배송지 및 배송거리

seller_state	seller_city	geolocation_lat_y	geolocation_lng_y	cust_lat	cust_lng	distance_km
CE	fortaleza	-3.789944	-38.550589	-1.29866	-47.898252	1074.622360

배송기간과 배송거리의 관계 및

배송거리와 배송기간의 상관관계는 0.39로 배송기간이 길어지는 원인을 배송거리라고 보고, 결국 고객 만족도를 낮추는 원인이 비효율적인 배송거리라고 가정했음.

지도 예시의 구매자는 Pará주의 catanhal city에서 baby카테고리의 상품을 구매함. 하지만 배송거리가 1727km로 배송거리가 매우 긴 것으로 판단하였음.
그러므로 같은 baby카테고리를 판매하는 모든 판매자들과의 거리를 계산해서 가장 가까운 거리에 있는 판매자를 탐색하였음.

그 결과, Ceará주의 forialeza city의 baby 카테고리 판매자가 1075km로 가장 가까운 거리에 있는 판매자로 계산되었음.

결론 및 문제점 개선방안

비교적 가까운 위치에 같은 카테고리의 판매자가 있으나, 멀리 있는 판매자로 부터 상품이 배송되는 경우가 있었고, 이는 배송기간을 길게 하여 고객 만족도(리뷰 점수)를 낮추는 원인이 된다고 판단함. (해당 데이터는 상품명없이 카테고리만 존재하여 부정확할 수 도 있음.)

배송 시간 최적화: 현재 배송 경로가 비효율적일 수 있으므로, 각 지역별 판매자-구매자 간 배송 루트 최적화를 통해 배송거리를 추가로 단축하고, 이를 기반으로 배송 시간 단축과 물류 비용 절감을 동시에 달성하여 고객 만족도를 극대화할 수 있습니다.

CUSTOMER ANALYSIS

고객 세분화 분석

목적
각 고객 그룹에 맞춤형 캠페인을 제공함으로써 마케팅 비용을 최적화
Recency가 낮은 고객을 식별하여 이탈 원인을 파악해 고객 이탈을 방지
세분화된 데이터를 활용하여 고객 가치를 극대화할 수 있는 전략을 수립

대시보드 링크 :

01 지표 및 세그먼트 분류 기준

지표&점수	1점	2점	3점	4점	5점
최근성 (Recency)	180일 초과	91~180일	31~90일	16~30일	0~15일
구매빈도 (Frequency)	1회	2회	3회	4회	5회 이상
구매금액 (Monetary)	하위 25%	25%~40%	40%~70%	70%~90%	상위 10%

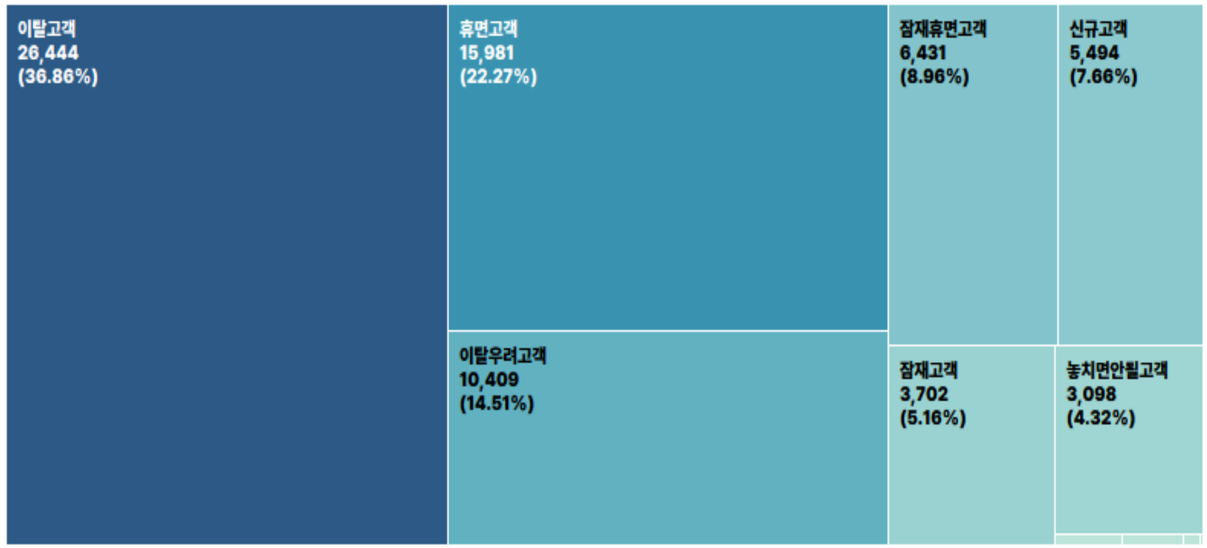
최근성은 2018/08/29
(2017/09~2018/08 중 마지막 구매일 기준)

Scores	Segment (Kor)
555, 545, 455, 445	챔피언 고객
554, 544, 454, 553, 543, 533, 443, 433, 355, 345, 335	충성 고객
551, 552, 541, 532, 531, 431, 432, 423, 352, 351	잠재적 충성 고객
521, 515, 522, 422, 421	신규 고객
525, 524, 523, 515, 514, 415, 315, 314, 313	잠재 고객
534, 424, 344, 334, 324, 323	관심 필요한 고객
312, 232, 331, 231, 222, 223	잠재 휴면 고객
215, 254, 245, 244, 234, 224	이탈 방지 필요 고객
311, 325, 243, 242, 141, 142, 113, 124	이탈 위험 고객
223, 222, 123, 122, 212, 211	휴면 고객
111, 112, 121, 131, 141, 151	이탈 고객

02 고객 세그먼트 결과

지표&점수	1점	2점	3점	4점	5점
최근성 (Recency)	46.76%	27.43%	16.19%	5.86%	3.76%
구매빈도 (Frequency)	97.53%	2.31%	0.13%	0.02%	0.01%
구매금액 (Monetary)	25.01%	14.99%	30.01%	19.99%	10%

RFM Segment



지표별 비율 해석

최근성(Recency)

1점 고객이 46.76%로 가장 높은 비중을 차지하며, 이는 많은 고객이 최근 구매 활동이 없음을 의미합니다.

상위 점수(4~5점)의 고객 비율은 각각 5.86%, 3.76%로 매우 낮아, 최근 구매를 유도하는 캠페인이 필요합니다.

구매빈도(Frequency)

대부분의 고객이 1점(97.53%)으로, 구매 활동이 거의 없는 고객이 많습니다.

상위 빈도 고객(4~5점)의 비중은 0.02% 이하로, 충성 고객 확보가 필요합니다.

구매금액(Monetary)

3점(30.01%)이 가장 많은 비중을 차지하며, 중간 수준의 구매력이 있는 고객이 다수입니다.

1점(25.01%) 및 2점(14.99%) 비율도 높아, 상위 고객군으로 전환하기 위한 업셀링 전략이 요구됩니다.

RFM 점수변환 후 세그먼트 결과

이탈 고객(36.86%)이 가장 큰 비중을 차지하며, 이들의 재참여를 유도하는 전략이 필요함.

휴면 고객(22.27%)과 이탈 우려 고객(14.51%) 또한 큰 비중을 차지해 주요 관리 대상이며, 이들을 재활성화하기 위한 맞춤형 캠페인이 요구됨.

잠재 고객(5.16%)과 신규 고객(7.68%)은 구매 증대 전략으로 활성화를 유도하여 상위 고객군으로 전환이 필요합니다.

03 최종 세그먼트별 전략 제안

Segment	특징	목표	전략
우수 고객	최근 구매, 자주 구매하며 높은 금액을 소비하는 핵심 고객	충성도 유지 및 브랜드 홍보 촉진	SNS 및 추천 프로그램, 멤버십 혜택 강화, 리뷰 요청 및 보상 제공
충성 고객	자주 구매하고, 높은 금액을 소비하며 브랜드 충성도가 높은 고객	유지 및 구매 확대 유도	업셀링 및 크로스셀링 제안, 친구 초대 프로그램, 정기구독 상품 홍보
잠재적 충성 고객	최근 구매 이력이 있으며, 잠재적으로 충성 고객으로 전환 가능성이 있는 고객	자주 구매하도록 유도	재구매 추가 혜택 제공, 구매이력 기반 제품 추천, 멤버십 가입 유도 캠페인
신규 고객	최근 첫 구매를 한 고객	충성 고객으로 전환	웰컴 이메일 캠페인, 다음 구매 할인코드 제공, SNS 팔로우 추가혜택
잠재 고객	구매 빈도가 낮지만 구매 가능성이 있는 고객	구매 행동을 유도	첫 구매 할인 제공, 한정시간 프로모션, 방문기록 활용 리타겟팅
관심 필요한 고객	구매 이력이 있으나 관심과 추가적인 관리를 필요로 하는 고객	재방문 유도 및 브랜드 관심 증대	할인 및 제품 체험행사 초대, 포인트 소멸 알림, 재구매 쿠폰 발송
잠재 휴면 고객	구매 빈도가 점차 감소하여 휴면 상태로 전환될 가능성이 있는 고객	재참여 유도	재구매 혜택 캠페인, 고객 복귀 프로모션, 무료 배송 프로모션
이탈 방지 필요 고객	이탈 가능성이 있어 구매 복귀를 유도해야 하는 고객	이탈 방지 및 재활성화	복귀 할인 제공, 포인트 추가적립 캠페인, 최저가 보상제
이탈 위험 고객	이탈 위험이 높으며, 브랜드와의 연결성이 약화된 고객	브랜드와의 연결성 회복	재구매 쿠폰 제공, 재방문 유도 이메일, 소규모 상품 추첨
휴면 고객	장기간 활동하지 않아 관계 회복이 필요한 고객	관계 회복 및 구매 재활성화	과거 관심상품 광고노출, 휴면 복귀 캠페인, 이탈이유 설문조사 후 상품제공
이탈 고객	이미 이탈한 고객	고객 복귀 유도	복귀 이벤트 초대, 신제품 수시 정보제공, 이탈이유 설문조사 후 상품제공