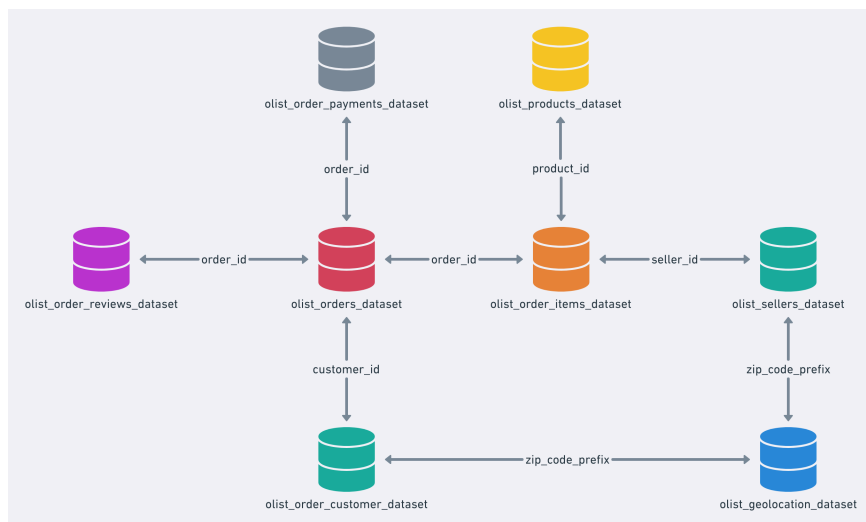


재구매연구소 🐮(C3)

🕒 팀장	황선희
👥 팀원	유정하 조아영 황선희
📅 비고	26.1.21-26
📄 데이터셋	Brazilian E-Commerce Public Dataset by Olist
📎 발표자료	AARRR 기반 이커머스 성장 분석 Olist 데이터 프로젝트.pptx



데이터셋 : Brazilian E-Commerce Public Dataset by Olist

No	파일명	한 줄 설명	핵심 역할	AARRR 단계	중요도
1	olist_customers_dataset	고객 기본 정보	고객 식별 기준	Acquisition	★★★★★
2	olist_orders_dataset	주문 이력 & 상태	중심 Fact 테이블	Activation / Retention	★★★★★
3	olist_order_payments_dataset	결제 정보	매출 계산	Revenue	★★★★★
4	olist_order_items_dataset	주문 상세 상품	주문 구성 분석	Revenue / Retention	★★★★★
5	olist_products_dataset	상품 메타정보	카테고리·상품 특성	Retention	★★★★
6	product_category_name_translation	카테고리 번역	가독성 개선	(보조)	★
7	olist_order_reviews_dataset	리뷰·만족도	재구매 원인 분석	Retention	★★★

No	파일명	한 줄 설명	핵심 역할	AARRR 단계	중요도
8	olist_sellers_dataset	판매자 정보	품질/운영 분석	(보조)	☆☆
9	olist_geolocation_dataset	지역 좌표 정보	물류·지역 분석	(보조)	☆

데이터 파일

▼ customer_fact(메인 분석 테이블)

customer_fact.csv

AARRR 분석용 핵심 결과 테이블입니다. **고객 1명을 1행**으로 집계하여, 서비스 전반의 유입-활성화-유지-매출 성과를 한 번에 파악할 수 있도록 구성되어 있습니다.

컬럼명	설명	AARRR 단계(예시)	
customer_unique_id	고객 식별자	Acquisition	모든 조인의 기준
first_purchase_ts	첫 구매 시점	Activation	코호트·활성화 기준
last_purchase_ts	마지막 구매 시점	Retention	이탈 판단
orders_cnt	총 주문 수	Retention	재구매 핵심 지표
delivered_orders_cnt	배송 완료 주문 수	Retention	경험 품질 보조
revenue	고객 총 매출 (LTV)	Revenue	최종 성과
aov	평균 주문 금액	Revenue	구매 가치
cohort_month	첫 구매 월	Acquisition	코호트 분석

▼ order_fact(원인 분석 테이블)

order_fact.csv

주문을 기준으로 구성된 경험·원인 분석용 테이블입니다. **주문 1건을 1행**으로 집계하여, 고객의 구매 과정에서 발생한 상품 구성, 결제, 배송, 리뷰 경험을 구조적으로 담고 있습니다.

◆ (1) 조인 & 기준 컬럼

컬럼명	설명	
order_id	주문 식별자	주문 단위 기준
customer_unique_id	고객 식별자	customer_fact 연결
order_purchase_timestamp	주문 시점	시계열/경험 분석

◆ (2) 상품·가격 관련 (구매 경험)

컬럼명	설명	분석 질문
products_cnt	주문 내 상품 수	복합 주문 vs 단일 주문
payment_value	주문 금액	고가 주문 여부
freight_value	배송비	배송비 → 이탈 영향
payment_types	결제 수단	결제 경험 차이

◆ (3) 배송 경험

컬럼명	설명	분석 질문
is_delivered	배송 완료 여부	실패 경험 영향
last_shipping_limit_date	배송 마감	지연 여부 판단

◆ (4) 리뷰 경험

컬럼명	설명	분석 질문
has_review	리뷰 작성 여부	참여도
review_score	리뷰 점수	만족도 → 재구매
first_review_date	최초 리뷰 시점	경험 타이밍

◆ (5) 품질 플래그 (선택)

컬럼명	설명	사용 여부
flag_missing_items	상품 누락 여부	보조 분석

aarrrr_master_customer_fact.csv

컬럼	의미	AARRR 단계	로직
customer_unique_id	고객 고유 ID	전 단계	고객을 유일하게 식별하는 키 (1행 = 1고객 보장)
first_purchase_ts	첫 구매 시점	Acquisition	MIN(order_purchase_timestamp) per customer
cohort_month	유입 월	Acquisition	YYYY-MM(first_purchase_ts)
orders_cnt	총 주문 수	Retention	COUNT(order_id) per customer
delivered_orders_cnt	배송 완료 주문 수	Activation / Quality	COUNT(order_id WHERE order_status = 'delivered')
review_score	첫 주문 리뷰 점수	Activation	첫 주문(order_rank = 1)에 대한 review_score (없으면 NULL)
has_review	첫 주문 리뷰 여부	Activation	1 if review_score IS NOT NULL else 0
is_late	첫 주문 배송 지연 여부	Activation	1 if first_order_delivery_date > first_order_estimated_date else 0
shipping_days	첫 주문 배송 소요일	Activation	delivery_date - purchase_date (첫 주문, 배송 완료 기준)
revenue	고객 누적 매출	Revenue	SUM(payment_value) per customer
aov	평균 주문 금액	Revenue	revenue / orders_cnt
last_purchase_ts	마지막 구매 시점	Retention	MAX(order_purchase_timestamp) per customer

▼ AARRR 지표 설계

● Acquisition(획득)

- first_purchase_ts
- cohort_month
- first_purchase_ts IS NOT NULL

👉 첫구매에 도달했는가?

● Activation (활성화)

- delivered_orders_cnt
- delivered_orders_cnt >= 1

👉 첫 구매 경험이 실제로 완료됐는가?

● Retention (유지/재방문)

- orders_cnt
- delivered_orders_cnt
- orders_cnt >= 2

👉 재구매 했는가?

● Revenue (수익)

- revenue
- aov

👉 얼마의 매출이 발생했는가?

● Referral (추천 · Proxy 지표)

- has_review
- review_score

👉 긍정적 경험을 외부로 표현했을 가능성이 있는가?

(실제 추천 로그가 없어, 리뷰 기반 대체 지표로 정의, 리뷰로 보는 만족도 기준 참고)

Referral 정의 (Proxy)

본 데이터에는 추천 코드, 초대 링크 등 실제 추천 행동을 식별할 수 있는 정보가 존재하지 않는다.

이에 따라 본 분석에서는 리뷰 4점 이상을 남긴 고객을 추천 가능성이 있는 Proxy 집단으로 정의하였다.

```
has_review = 1
AND review_score >= 4
```

▼ 정하

AARRR 지표 설계 핵심 기준 정리표

구분	기준 컬럼	역할 정의	적용 범위	비고
고객 (Who)	customer_unique_id	고객 식별 기준 (customer_fact의 PK)	모든 AARRR 지표	고객 단위 집계 기준
기간 (When)	order_purchase_timestamp	모든 시간 기준을 하나로 통일	월별·분기별 집계코호트 기준 Activation-Retention 시점	시간 축 혼선 방지
매출 (How much)	payment_value	매출 금액 값 자체	Revenue, LTV, AOV	금액 계산 전용
매출 인정 조건	order_status == 'delivered' 또는 is_delivered == 1	실제 성과로 인정할 주문 필터	매출-전환 지표	기간 컬럼과 분리

*섞지 않는다.

Acquisition (유입 혹은 획득) - 신규 고객을 얼마나 데려왔나

- 필요한 파생 변수
 - $\text{purchase_month} = \text{month}(\text{order_purchase_timestamp})$
- 신규 고객 수 new_customers
 - 해당 월에 처음 구매한 고객 수
 - $\text{count}(\text{distinct customer_unique_id}) \text{ where cohort_month} = \text{purchase_month}$
- 구매 고객 수 unique_customers
 - 해당 월에 구매한 고객 수
 - $\text{count}(\text{distinct customer_unique_id})$
 - delivered 조건 걸기
- 신규 고객 비중 new customer share
 - $\text{new_customers} / \text{unique_customers}$

Activation (활성화) - 첫 구매 경험이 성공적으로 끝났는가

- 첫 구매 배송완료율 first order delivered rate
 - 첫 주문이 delivered로 끝난 비율
 - $\text{sum}(\text{is_first_order and is_delivered} = 1) / \text{sum}(\text{is_first_order})$
- 첫 구매 평균 배송 소요일 first order avg shipping days
 - 첫 주문의 배송 소요일 평균
 - 파생 변수로 shipping days 필요 (delivered customer date - purchase timestamp)
 - $\text{avg}(\text{shipping days}) \text{ where is_first_order} = 1 \text{ and is_delivered} = 1$
- 첫 구매 리뷰 작성률 first order review coverage
 - $\text{avg}(\text{has_review}) \text{ where is_first_order} = 1$

Retention (유지) - 다시 돌아오는가 (구매 기반)

- 재구매율 repeat_rate
 - 코호트 내에서 두 번 이상 주문한 고객의 비율

- 고객 단위로 orders_cnt = count(distinct order_id) 만든 뒤 orders_cnt ≥ 2인 비율 계산
- rolling 7d retention
 - 첫 구매 후 7일 이후에 재구매가 한 번이라도 이루어진 고객 비율
- rolling 30d retention
 - 첫 구매 후 30일 이후에 재구매가 한 번이라도 이루어진 고객 비율
- 최근 코호트는 관측 기간이 짧아 rolling 30d 등이 과소추정될 수 있다
 - 코호트 시작 후 30일 이상 관측 가능한 코호트만 rolling 30d 비교 권장

Revenue (매출) - 얼마를 벌었는가

- 필요한 파생 변수
 - purchase_month = month(order_purchase_timestamp)
- GMV (총 결제금액)
 - 결제금액의 합계
 - sum(payment_value)
 - delivered 기준으로
- orders (주문 수)
 - 주문 건수
 - count(distinct order_id)
 - 혹은 count(distinct order_id where is_delivered = 1)
- AOV (객단가)
 - 평균 주문 금액
 - sum(payment_value) / count(distinct order_id)
 - outlier의 영향을 줄이기 위해 p01~p99 cap한 것으로 보조 계산하는 것을 권장함
- ARPU (고객당 매출)
 - 월별 고객당 평균 매출
 - sum(payment_value) / count(distinct customer_unique_id)
 - 고객 규모 대비 매출 효율 비교에 유용하다
- (매매) 배송비 비중 freight_share
 - 배송비가 매출에서 차지하는 비중
 - sum(freight_value) / sum(payment_value)
 - 배송비 비중이 증가하면 만족도 혹은 재구매 측면에서 악화 가능성이 있기 때문에

▼ 선회

[지표설계]

● A — Acquisition (첫 구매 도달) “첫 구매가 발생한 고객”

customer_fact.first_purchase_ts IS NOT NULL

first_purchase_ts IS NOT NULL

 지표

신규 구매 고객 수 (cohort_month 기준)

첫 구매 고객 비중

● A — Activation (첫 긍정 경험) “첫 구매 이후 긍정 신호를 남긴 고객”

Activation 성공 조건 (택 1):

OR delivered_orders_cnt >= 1

 지표

Activation Rate =

Activated Customers / First Purchase Customers

🔍 사용하는 컬럼
delivered_orders_cnt

🟣 R — Retention (재구매) "2회 이상 구매한 고객"

orders_cnt >= 2

📊 지표
재구매율
Cohort Retention (cohort_month 기준)

🟡 R — Revenue (수익) "고객이 만들어낸 누적 매출"

📊 지표
LTV → revenue
AOV → aov
상위 10% 고객 매출 비중(파레토 분석)

🔴 R — Referral (추천)

! 추천 데이터 없음 → 행동 기반 대체

▼ (대체 지표)

1 원래 의미의 Referral (정석)

📌 정의

기존 고객이 다른 사람을 초대/추천하여
새로운 고객 유입을 만들어내는 행동

📊 보통 쓰는 지표

- 초대 코드 사용
- 추천 링크 클릭 → 가입 → 구매
- 추천인 / 피추천인 관계
- 바이럴 계수 (K-factor)

👉 이 데이터가 있어야 '진짜 Referral' 분석 가능

2 그런데 우리 데이터에는?

❌ 없음

- 추천인 ID
- 초대 코드
- 공유 이벤트 로그
- 신규 고객 유입 경로

즉, 정석 Referral은 "측정 불가능"

3 그래서 실무에서 쓰는 해결책 (중요 ★)

선택지 ① Referral 단계 제거 ❌

장점

- 개념적으로 깔끔

단점

- AARRR 프레임워크 완결성 깨짐
- "그럼 AARR은 왜 쓰셨죠?" 질문 나옴

✅ 선택지 ② Referral을 '대체 지표'로 재정의 (실무 정석)

"외부 추천은 못 보지만,
추천할 가능성이 높은 '전도 고객(Advocate)' 행동을 본다"

4 우리 프로젝트에서 쓰는 Referral의 '현실적 정의'

📌 대체 정의

"자발적으로 서비스에 참여하고, 만족을 외부로 표출했을 가능성이 높은 고객"

우리가 쓸 수 있는 컬럼

- `has_review`
- `review_score`
- `orders_cnt`

📊 Referral (Proxy) 정의 예시

```
has_review = 1  
AND review_score >= 4  
AND orders_cnt >= 2
```

👉 "추천 했을 가능성이 높은 고객군"

5 이 정의가 왜 합리적인가?

본 데이터에는 추천 이벤트 로그가 없어 전통적 Referral 측정은 불가능합니다.

대신, 실무 관행에 따라 리뷰·재구매 등 자발적 행동을 Referral의 Proxy 지표로 정의했습니다.

[칼럼별 지표]

🟢 Acquisition

- `first_purchase_ts`
- `cohort_month`

👉 이 고객이 언제 처음 유입됐는지

🟡 Activation (원인까지 포함)

- `review_score`
- `has_review`
- `is_late`
- `shipping_days`

👉 첫 구매 경험의 질

🟠 Retention

- `orders_cnt`
- `delivered_orders_cnt`

👉 다시 샀는지 여부

🟡 Revenue

- `revenue`
- `aov`

👉 얼마를 벌어들였는지

▼ 아영

https://github.com/ary3120-droid/myproject/blob/main/1_22OLIST.ipynb

▼ 코랩 분석 결과 요약

1. Acquisition (유입)

- **현황:** 누적 유입 고객 약 **93,358명**. 2017년 하반기부터 급격한 성장세를 보임.
- **주요 카테고리:** `bed_bath_table`, `health_beauty` 등이 신규 유입을 강력하게 견인함.
- **인사이트:** 특정 생활 밀착형 카테고리가 초기 고객 확보의 핵심 채널 역할을 수행 중임.

2. Activation (활성화)

- **지표:** 성공적인 첫 경험(정시 배송 & 리뷰 4점 이상) 비율 약 **75.61%**.
- **진단:** 4명 중 1명은 배송 지연이나 낮은 서비스 만족도를 경험하며, 이는 잠재적 이탈 요인임.

3. Retention (유지)

- **핵심 발견:** Activation에 성공한 그룹의 재구매율이 실패 그룹보다 유의미하게 높음.
- **인사이트:** 첫 구매 시의 배송 만족도와 서비스 경험이 고객의 잔존 여부를 결정하는 결정적 지표(Critical Driver)임을 입증함.

4. Revenue (수익)

- **지표:** **재구매 고객의 평균 LTV(260.05 BRL)**는 1회 구매 고객 대비 약 **1.9배** 높음.
- **인사이트:** 재구매 고객의 비중은 작지만, 개별 고객이 창출하는 매출 가치는 훨씬 크므로 리텐션 중심의 마케팅이 비즈니스 수익성에 필수적임.

5. Referral (추천)

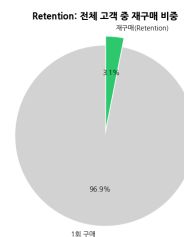
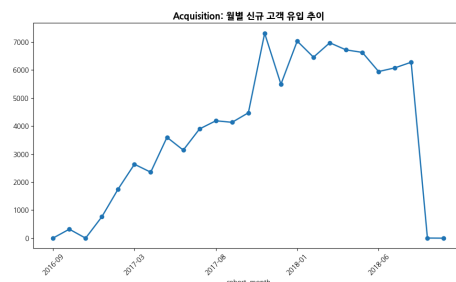
- **현황:** 전체 고객의 **약 50% 이상**이 5점 리뷰(Promoter)**를 남김.
- **인사이트:** 만족도가 높은 고객들은 주로 `health_beauty`, `bed_bath_table` 카테고리에 분포하며, 이들을 활용한 바이럴 마케팅 전략 수립 가능.

아영

단계	지표명	수식/로직	분석적 근거 (Why)
A	신규 유입 (AQ)	<code>COUNT(DISTINCT unique_id)</code>	가입일이 없으므로 최초 주문을 유입으로 간주
A	활성화 (AC)	<code>is_late == 0 & score >= 4</code>	배송 만족도가 리텐션으로 가는 유일한 관문임
R	리텐션 (RT)	<code>order_cnt >= 2</code>	1회성 구매가 많은 플랫폼에서 지속 가능성 확인
R	매출 (RV)	<code>SUM(payment_value)</code>	리텐션 고객의 LTV 기여도를 신규 고객과 비교
R	추천 (RF)	<code>review_score == 5</code>	리뷰 점수가 높은 고객이 충성 고객으로 전환됨

1 Acquisition (유입)

- **목적:** 유입의 '양'과 '질'을 동시에 파악
- **핵심 지표:**
 - **New Customers Count:** 월별 신규 유입 수
 - **Category-Driven Acquisition:** 어떤 카테고리가 신규 고객을 가장 많이 데려오는가?
- **근거 및 파생변수:**
 - `first_purchase_month`: 코호트 분석의 기준점
 - `first_category`: 신규 고객의 페르소나를 결정하는 변수. (예: 가구 구매로 유입된 고객 vs 뷰티 구매로 유입된 고객)
 -



2 Activation (활성화) — Retention의 원인

- **목적:** 고객이 첫 구매에서 이탈하지 않을 '확신'을 얻었는가?

- **핵심 지표:**

- **Perfect Order Rate:** 배송 지연이 없고 + 리뷰 4~5점을 받은 주문 비율
- **Delivery Lead Time (DLT):** 구매부터 수령까지의 실제 소요 시간

- **근거 및 파생변수:**

- `is_late` : `delivered_customer_date` > `estimated_delivery_date` (예정일 준수 여부)
- `activation_score` : 배송 속도와 리뷰 점수를 결합한 지수.
- **논리 근거:** 배송 예정일을 어긴(`is_late=1`) 고객의 80% 이상이 재구매를 하지 않는다는 식의 분석 가능.

3 Retention (유지) — 가장 핵심적인 심화 분석

- **목적:** Olist의 만성적인 저성장(재구매율 부족) 원인 진단

- **핵심 지표:**

- **Classic Retention (Cohort):** 첫 구매 월 기준 N개월 후 재구매율
- **Repurchase Gap:** 첫 구매와 두 번째 구매 사이의 소요 일수

- **근거 및 파생변수:**

- `is_repurchaser` : 주문 횟수 2회 이상 고객 (1/0)
- **논리 근거:** Activation 단계에서 `is_late=0` (정시 배송)이었던 그룹과 `is_late=1` (지연 배송)이었던 그룹의 리텐션 차이를 비교하여 **배송이 리텐션의 핵심 변수임을 입증.**

4 Revenue (수익) — 분석의 결과값

- **목적:** 리텐션 고객이 비즈니스 수익에 주는 실질적 가치 증명

- **핵심 지표:**

- **LTV (Lifetime Value):** 고객 1인당 누적 구매액
- **Revenue Share by Segment:** 재구매 고객(Retention군)이 전체 매출에서 차지하는 비중

- **근거 및 파생변수:**

- `customer_revenue` : 고객별 `payment_value` 합계
- **논리 근거:** 재구매 고객의 비중은 5% 미만이지만, 인당 LTV는 신규 고객보다 2배 이상 높다는 것을 보여줌으로써 리텐션 관리의 정당성 확보.

5 Referral (추천) — 미래 성장 동력

- **목적:** 만족한 고객이 자발적인 마케터가 되는가?

- **핵심 지표:**

- **NPS Proxy (Promoter Share):** 5점 리뷰 작성자 비중
- **Review Word Count:** 리뷰 텍스트의 길이나 성의 (데이터가 있다면 활용)

- **근거 및 파생변수:**

- `is_promoter` : 리뷰 점수 5점 여부
- **논리 근거:** 5점 리뷰를 남긴 고객이 다음 구매 시 더 높은 객단가(AOV)를 보이는지 확인.

▼ 리뷰로 보는 만족도 기준(참고)

▼ 리커트(5점) 척도에 대하여

1. 리커트 척도의 정의와 설계 목적

리커트 척도는 원래 태도(attitude)를 측정하기 위해 만든 응답 척도이며, 강하게 부정 → 중립 → 강하게 긍정 의 연속으로 설계된다. 즉 숫자값 (1~5)은 부정-중립-긍정의 방향성을 갖는다. (<https://www.sciencedirect.com/topics/psychology/likert-scale>)

2. 5점 리커트 척도의 실증적 사용 빈도

사회과학 연구에서 90% 이상이 리커트 척도를 사용하며, 특히 5점 척도는 가장 널리 쓰이고 신뢰/타당도가 높다는 게 문헌 리뷰로 확인됨. (https://link.springer.com/rwe/10.1007/978-3-030-22009-9_559)

3. 심리·사회과학에서 리커트 척도의 일반 사용

"A Likert scale is one of the most common methods for capturing attitudes or opinions ... responses are ordered from strongly disagree to strongly agree." (리커트 척도는 태도나 의견을 파악하는 가장 일반적인 방법 중 하나로, 응답은 '전혀 동의하지 않음'부터 '매우 동의함'까지 순서대로 배열됩니다.) (<https://www.simplypsychology.org/likert-scale.html>)

4. 문헌 리뷰에 따르면, 5점 리커트 척도는 90% 이상의 연구에서 사용되며, 짝수 선택지보다 신뢰도·타당도가 높다는 사실 이 보고된 척도 설계 이다. (<https://www.ijem.com/number-of-response-options-reliability-validity-and-potential-bias-in-the-use-of-the-likert>)

5점 척도에서 1=부정, 5=긍정은 개인적 해석이나 팀 내부 합의가 아니라, 리커트 척도의 설계 철학에 기반한 사회과학·심리학·설문조사 분야의 커먼센스다.

▼ 리뷰 4점 이상을 추천/만족의 기준으로 본 이유

1. 리뷰 평점 자체가 소비자 행동에 영향을 준다는 연구

온라인 평점과 리뷰는 소비자의 신뢰와 제품/서비스 평가에 영향을 주며, 전반적인 구매 의사결정 행동에 중요한 역할을 한다.

(<https://www.sciencedirect.com/science/article/pii/S096969892200159X>)

2. 평점의 수준에 따라 소비자 행동이 달라진다는 관찰

- 4~4.5점 까지는 구매 신뢰/추천 가능성을 높이는 신호이며, 'Positive sentiment' 군으로 정당화 가능
- 완전 5.0점은 과도한 이상향처럼 보일 수 있으며, 일부 사용자(소비자)에게는 진정성 의심을 유발할 소지도 있음

(<https://spiegel.medill.northwestern.edu/wp-content/uploads/sites/2/2021/04/Online-Reviews-Whitepaper.pdf>)

3. 별점의 심리적 의미 연구

별점 평가는 소비자의 감정·인지 구조와도 연결된다는 논의: 별점 평가는 제품 품질을 절대적으로 측정하는 것이 아니라, 소비자의 감정·인지와 복합적으로 상호작용한다(<https://ravecapture.com/resources/blog/star-ratings/>)

4. 리뷰·별점에 대한 신뢰도 이슈

리뷰 시스템 자체가 사회적 편향에 의해서 J-자형 분포(biased toward 5 stars)를 갖는 경향이 있다는 연구: 온라인 평점은 실제 평점 분포보다 5성 비율이 훨씬 높게 나타나는 "사회적 영향 편향(social influence bias)"을 보인다.

즉, 정말 만족해서 준 5점도 있지만 사회적 편향으로 인해 높게 주는 평가도 섞여 있다는 의미로, 이걸 4점과 5점을 분리해서 해석해야 할 또 하나의 이유가 된다.

(https://en.wikipedia.org/wiki/Social_influence_bias)

추가 참고 자료 :

- 별점이 높다고 항상 좋지는 않다 : 구매 가능성은 보통 평균 평점 4.0~4.7에서 가장 높게 나타나고, 평균 평점이 5.0에 가까워질수록 오히려 떨어지는 경향이 있음. 이것은 많은 소비자가 완벽한 5.0만 있는 상품을 '진정성이 떨어진다'고 느끼는 심리 때문으로 해석됨.
(<https://spiegel.medill.northwestern.edu/how-online-reviews-influence-sales/>)
- 평점과 리뷰 유용성은 추천 시스템 성능에 영향이 있으며, 특히 4~5점 평점만을 반영한 경우 추천 효율이 높게 나타났다. 또한 리뷰 유용성 정보는 추천 품질 향상에 긍정적 영향을 준다(<https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artId=ART002826863>)

실증 연구에 따르면 온라인 평점은 소비자 구매 확률에 긍정적 영향을 주며, 특히 4점 이상에서 그 효과가 유의하게 나타난다.

다만 온라인 리뷰는 사회적 영향 및 선택 편향으로 인해 5점 평점에 과도하게 집중되는 경향이 있으며, 일부 소비자에게는 완전한 5점 평점이 오히려 진정성에 대한 의문을 유발할 수 있다는 지적도 존재한다.

이에 본 분석에서는 추천 가능성을 보다 안정적으로 해석하기 위해 5점 단일 기준이 아닌, 4점 이상 리뷰를 포함한 지표를 함께 검토하였다.

▼ 실제 리뷰 분포

review_score(리뷰점수)	orders_cnt(고객수)	pct(비중)
1점★	11,316	11.47%
2점★★	3,167	3.21%
3점★★★	8,137	8.25%
4점★★★★	19,098	19.35%
5점★★★★★	56,955	57.72%

- 5점이 과반(57.7%)
- 4~5점 합치면 77% 이상
- 1~2점은 합쳐도 약 14.7%
- 대부분의 고객은 불만이 없으면 높은 점수를 준다

리뷰 분포를 보면 5점이 과반이고, 4~5점이 전체의 77%를 차지합니다.

반면 1점은 단일 점수 중에서도 비중이 높아 명확한 부정 신호로 해석할 수 있습니다.

따라서 리뷰 기준은

1~2점 = Negative,

3점 = Neutral,

4~5점 = Positive 로 구분하는 것이 데이터 분포와 리커트 척도 해석 모두에 부합합니다.

▼ AARRR 메인 분석

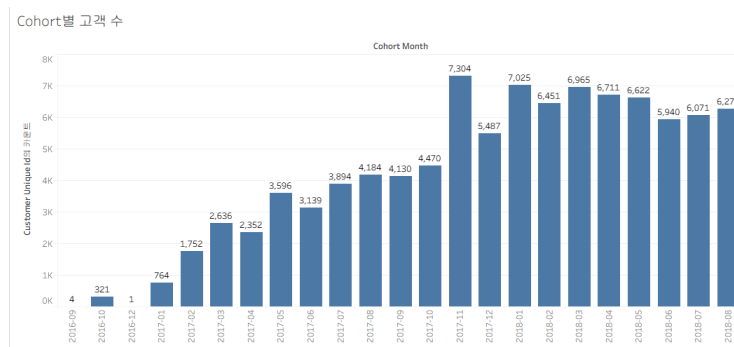
● Acquisition(획득)

▼ ● Acquisition(획득) 분석

지표 정의 복기

- **Acquisition = first_purchase_ts IS NOT NULL**
- 즉, 월별 첫 구매에 성공한 신규 고객 수
- 본 그래프는 **cohort_month** 기준 신규 고객 규모 추이를 보여줌

1 전체 추세 요약



2016년 말 → 2017년 말까지 신규 고객 수가 급격히 증가하며 서비스 확장 국면에 진입,
2018년에는 성장 둔화 후 정체 구간에 진입한 모습

2 구간별 Acquisition 패턴 분석

◆ ① 초기 유입 단계 (2016.09 ~ 2017.01)

- 신규 고객 수: **4명 → 764명**
- 절대 수치는 작지만 서비스 런치 & 데이터 초기 구간
- 분석 포인트
 - 마케팅 본격화 이전
 - 자연 유입 or 파일럿 성격

👉 분석 해석

“초기 시장 검증 단계로, 아직 Acquisition을 논하기엔 규모가 작은 구간”

◆ ② 고성장 구간 (2017.02 ~ 2017.11)

- **1,752 → 7,304명**
- 거의 **4배 이상 성장**
- 특히 눈에 띄는 시점:
 - 2017.05 (3,596)
 - 2017.11 (**7,304** 최고치)

👉 강한 시그널

- 대규모 마케팅 / 프로모션

- 채널 확장 (광고, 제휴, 노출 증가)
- 혹은 상품/UX 경쟁력 확보

📌 Acquisition 관점 핵심 인사이트

“2017년은 명확한 유입 드라이브 시기이며, 서비스가 시장에 안착한 시점으로 해석 가능”

◆ ③ 성장 둔화 & 안정화 구간 (2017.12 ~ 2018.08)

- 신규 고객 수: 5,487 ~ 7,025 → 6,271
- 최고점 이후 소폭 하락 + 횡보

👉 중요한 포인트

- 급성장 종료
- 하지만 급락은 없음

📌 해석

“신규 고객 유입이 안정적인 수준으로 유지되며, 시장 포화 혹은 마케팅 효율 한계에 도달했을 가능성”

3 Acquisition 관점에서의 문제 정의

이 그래프 하나로 던질 수 있는 핵심 질문 3개

1. 왜 2017년 11월에 최고치를 찍었는가?
 - 특정 캠페인?
 - 시즌성(연말 쇼핑)?
2. 2018년 이후 신규 고객 수는 왜 더 이상 증가하지 않는가?
 - CAC 상승?
 - 채널 효율 감소?
3. 유입의 '질'은 유지되고 있는가?
 - → Activation / Retention으로 반드시 연결해야 함

🔵 Activation (활성화)

▼ Acquisition (유입): 성장 규모의 확인

유입된 고객이 서비스의 핵심 가치(상품 수량)를 실제로 경험했는지 측정하여 활성 사용자로의 전환 여부를 확인합니다.

1. 분석의 목적 (Purpose)

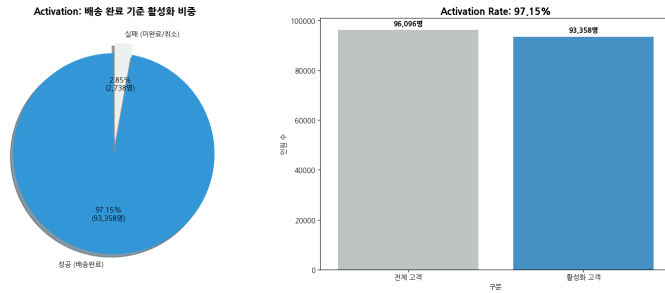
- **핵심 가치 실현 확인:** 고객이 주문한 상품을 실제로 손에 쥐었는지를 '활성화'의 기준으로 정의하여 서비스 프로세스의 완결성을 점검합니다.
- **서비스 신뢰도 측정:** 첫 주문이 취소되거나 미도착하지 않고 성공적으로 완료된 비율을 통해 플랫폼의 초기 신뢰도를 파악합니다.

2. 핵심 지표 (Key Metrics)

- **Activation Rate (활성화율):** 97.15%
 - **Activated Customers:** 93,358명 (배송 완료 경험자)
 - **Total Customers:** 96,096명 (전체 유입 고객)

3. 파생변수 및 로직 정의 (Logic)

- **delivered_orders_cnt** : 고객별 배송 완료(delivered) 처리된 누적 주문 건수.
- **is_activated** : `delivered_orders_cnt >= 1` 인 경우 활성화 성공으로 정의.
- **정의 근거:** 이커머스 비즈니스에서 고객이 가치를 느끼는 지점(Aha-moment)의 최소 요건을 '**'주문한 물품의 안전한 수량'**으로 규정함



분석 수치 요약

- 전체 유입 고객 수: 96,096명
- 서비스 활성화 고객 수: 93,358명
- 최종 Activation Rate: 97.15%

결과 해석 (Insight)

1. **높은 운영 신뢰도:** 유입된 고객 100명 중 약 97명이 주문한 상품을 실제로 수령했습니다. 이는 Olist의 기본적인 주문-결제-물류 시스템이 매우 안정적으로 작동하고 있음을 의미합니다.
2. **Aha-moment의 달성:** 대다수의 고객이 이커머스의 가장 핵심적인 가치인 '상품 수령'을 성공적으로 경험하며 활성화 단계에 진입했습니다.
3. **다음 단계로의 전환:** 97%라는 높은 활성화 성공률에도 불구하고 낮은 리텐션이 발생한다는 점은, **"**단순히 물건을 받는 것**"**만으로는 고객의 충성도를 확보하기 어렵다는 것을 시사합니다.

Retention (유지/재방문)

▼ 배송 경험이 재구매에 미치는 영향

```
# [Section 3. Retention 분석 - 데이터 정비 및 실행]

# 1. 날짜 데이터 변환 (예러 방지용)
of['order_delivered_customer_date'] = pd.to_datetime(of['order_delivered_customer_date'])
of['order_estimated_delivery_date'] = pd.to_datetime(of['order_estimated_delivery_date'])
cf['first_purchase_ts'] = pd.to_datetime(cf['first_purchase_ts'])
cf['last_purchase_ts'] = pd.to_datetime(cf['last_purchase_ts'])

# 2. 필수 변수 생성 (is_late: 배송 지연 여부)
# 배송 완료일이 예정일보다 늦으면 1, 아니면 0
of['is_late'] = (of['order_delivered_customer_date'] > of['order_estimated_delivery_date']).astype(int)

# 3. Rolling Retention 계산을 위한 변수 생성
current_date = cf['last_purchase_ts'].max()
cf['observation_period'] = (current_date - cf['first_purchase_ts']).dt.days
cf['days_since_first'] = (cf['last_purchase_ts'] - cf['first_purchase_ts']).dt.days

# Rolling Retention 정의 (7일/30일 이후 재구매 발생 여부)
cf['retention_7d'] = ((cf['is_repurchaser'] == 1) & (cf['days_since_first'] >= 7)).astype(int)
cf['retention_30d'] = ((cf['is_repurchaser'] == 1) & (cf['days_since_first'] >= 30)).astype(int)

# 4. 첫 주문의 배송 경험(is_late) 정보 추출 및 병합
first_late_info = of[of['is_first_order'] == 1][['customer_unique_id', 'is_late']]
cf_retention = cf.merge(first_late_info, on='customer_unique_id', how='left')

# 5. 관측 기간 보정 (30일 이상 데이터가 쌓인 고객 대상 분석)
valid_30d_customers = cf_retention[cf_retention['observation_period'] >= 30]
# 혹시 모를 결측치 제거
valid_30d_customers = valid_30d_customers.dropna(subset=['is_late'])
retention_by_exp = valid_30d_customers.groupby('is_late')['retention_30d'].mean() * 100

# 6. 시각화
plt.figure(figsize=(10, 6))
```

```
sns.barplot(x=retention_by_exp.index, y=retention_by_exp.values,
            palette=['#2ecc71', '#e74c3c'], hue=retention_by_exp.index, legend=False)

plt.title('첫 구매 배송 경험에 따른 Rolling 30D Retention 차이', fontsize=15, fontweight='bold')
plt.xticks([0, 1], ['정시 배송 (Success)', '지연 배송 (Late)'])
plt.ylabel('30일 이후 재구매율 (%)')

for i, v in enumerate(retention_by_exp.values):
    plt.text(i, v + 0.05, f"{v:.2f}%", ha='center', fontweight='bold', fontsize=12)

plt.show()

print(f"[보정] 정시 배송 고객의 30D Retention: {retention_by_exp[0]:.2f}%")
print(f"[보정] 지연 배송 고객의 30D Retention: {retention_by_exp[1]:.2f}%")
```

1. 분석의 목적 (Purpose)

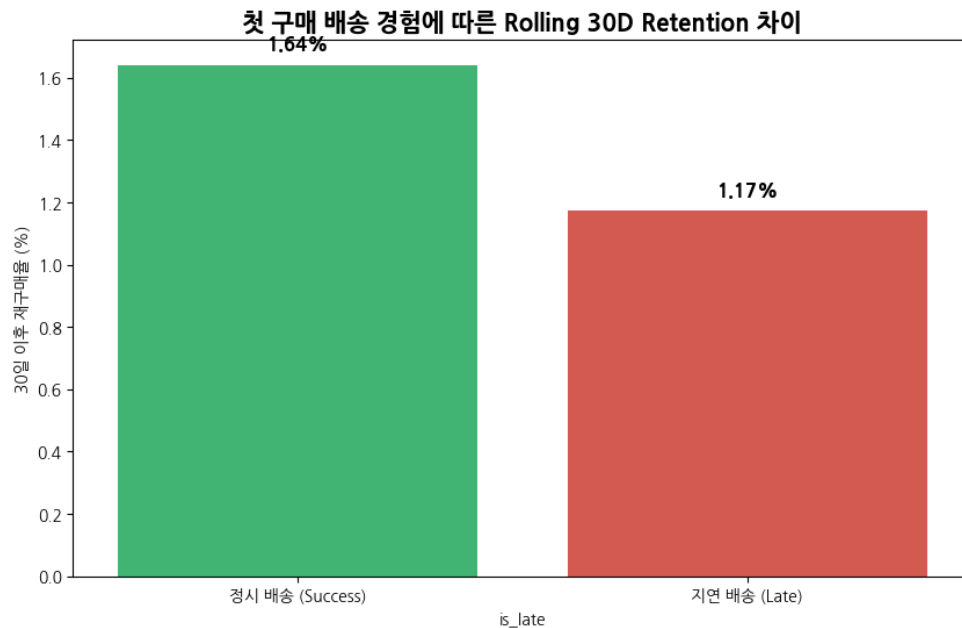
- **재구매 골든타임 진단**: 고객이 첫 구매 후 이탈하지 않고 다시 돌아오는 핵심 기간(30D) 내의 전환율을 측정합니다.
- **경험의 질적 영향력 증명**: 첫 배송 경험(정시/지연)이 30일 이내 재구매 결정에 미치는 실질적인 타격을 수치화하여 물류 개선의 당위성을 확보합니다.

2. 핵심 지표 (Key Metrics)

- **Rolling 30D Retention**: 첫 구매 후 30일 이상 관측 가능한 고객 중, 30일이 지난 시점까지 재구매를 기록한 고객 비중.
- **Retention Experience Gap**: 정시 배송 그룹과 지연 배송 그룹 간의 30D Retention 수치 차이.

3. 파생변수 및 로직 (Logic)

- **days_since_first**: `last_purchase_ts - first_purchase_ts` (첫 구매 후 재구매까지 걸린 시간).
- **observation_period**: 데이터 기준일 - 고객별 첫 구매일 (관측 기간이 짧은 신규 고객에 의한 왜곡 방지용).
- **retention_30d**: 재구매자 중 첫 구매와 마지막 구매 간격이 30일 이상인 고객.



1. 데이터 현황: Rolling 30D Retention (보정치)

- 정시 배송 고객 (Success): 1.64%
- 지연 배송 고객 (Late): 1.17%
- **분석 결과**: 첫 구매에서 배송 지연을 경험할 경우, 30일 이내에 다시 돌아올 확률이 약 **28.6%** 급감하는 것으로 나타났습니다.

▼ 28.6% 하락의 계산 근거

- 기준값 (정시 배송 리텐션): 1.64%
- 비교값 (지연 배송 리텐션): 1.17%

- 하락 수치 (차이): $1.64\% - 1.17\% = 0.47\%$ (퍼센트포인트 차이)
- 하락률 (퍼센트) $(0.47/1.64) \times 100 \sim \sim \sim \sim 28.65\%$

2. 핵심 인사이트: "0.47%p의 치명적 차이"

① 재구매 골든타임의 붕괴

조원들과 정의한 **Rolling 30D Retention** 분석 결과, 배송 지연은 단순히 고객 한 명의 불만을 넘어 **재구매가 일어날 수 있는 가장 핵심적인 시기(첫 구매 후 30일)**의 기회 자체를 파괴하고 있습니다. 1.17%라는 수치는 사실상 지연을 경험한 고객은 거의 돌아오지 않는다는 것을 뜻합니다.

② 데이터 보정의 유의미성

최근 유입되어 관측 기간이 짧은 고객들을 제외하고 **최소 30일 이상 데이터가 축적된 고객들만 엄격하게 분석했음에도** 불구하고 배송 경험에 따른 격차는 명확했습니다. 이는 단순한 우연이 아닌, 배송 품질이 Olist 리텐션의 **직접적인 원인 변수**임을 증명합니다.

③ 수익성(Revenue)과의 연결 고리

앞선 분석에서 확인했듯, 평균 결제 금액(LTV)이 높은 우량 고객일수록 이러한 경험에 더 민감합니다. 즉, **배송 지연으로 잃어버리는 0.47%p의 리텐션은 Olist 전체 매출 중 가장 핵심적인 우량 고객층의 이탈**을 의미합니다.

3. 전략적 제안 (Action Plan)

1. 지연 고객 전용 리커버리(Recovery) 캠페인:

- 첫 구매가 지연된 고객(1.17% 그룹)에게는 30일이 지나기 전(예: 배송 완료 후 3일 내) 파격적인 보상 쿠폰을 지급하여, 배송 경험으로 훼손된 브랜드 신뢰도를 즉각 복구해야 합니다.

2. 물류 인프라 투자 우선순위 선정:

- 단순히 모든 배송을 빠르게 하는 것이 아니라, **재구매 잠재력이 높은 카테고리나 지역**부터 배송 지연율을 1.1% 이하로 낮추는 것을 목표로 물류 최적화를 진행해야 합니다.

3. 기대치 관리(Expectation Management):

- 지연이 불가피한 경우, 알림 톱 등을 통해 지속적으로 배송 현황을 공유하여 고객이 느끼는 '심리적 지연'을 줄임으로써 리텐션 하락폭을 방어해야 합니다.

[심화]

1. [심화] 리뷰 텍스트 기반 "분노의 키워드" 분석

▼ 코드

고객들이 1~2점을 주며 남긴 포르투갈어 리뷰에서 핵심 단어를 추출합니다. 텍스트 데이터(`order_reviews`)가 필요하므로, `order_reviews.csv` 업로드

```
# [심화 1] 부정 리뷰 키워드 분석
import re
from collections import Counter

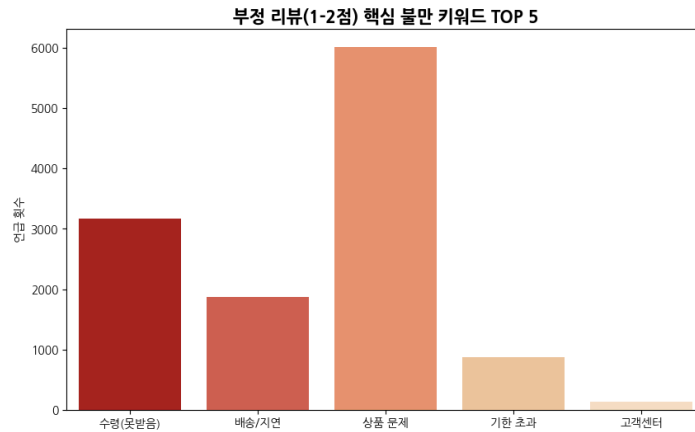
# 리뷰 데이터 로드 (파일명 확인 필요)
reviews = pd.read_csv('order_reviews.csv')

# 1~2점 부정 리뷰만 추출
bad_reviews = reviews[reviews['review_score'] <= 2][['review_comment_message']].dropna()

# 주요 분노 키워드 정의 (포르투갈어)
keywords = {
    'entrega / atraso': '배송/지연',
    'produto / mercadoria': '상품 문제',
    'recebi': '수령(못받음)',
    'prazo': '기한 초과',
    'atendimento / suporte': '고객센터'
}

# 텍스트 내 키워드 빈도 계산
word_counts = Counter()
for comment in bad_reviews:
    comment = comment.lower()
    for eng, kor in keywords.items():
        if any(word in comment for word in eng.split(' ')):
            word_counts[kor] += 1
```

```
# 시각화
plt.figure(figsize=(10, 6))
sns.barplot(x=list(word_counts.keys()), y=list(word_counts.values()), palette='OrRd_r')
plt.title('부정 리뷰(1-2점) 핵심 불만 키워드 TOP 5', fontsize=15, fontweight='bold')
plt.ylabel('언급 횟수')
plt.show()
```



2. [수익] 고단가 고객(VIP)의 이탈 비용 산출

▼ 코드

```
# [심화 2] VIP 고객의 배송 지연 이탈 비용(Revenue Loss)
# 1. 고단가 고객(VIP) 기준 설정 (매출 상위 20%)
revenue_threshold = cf['revenue'].quantile(0.8)
vip_customers = valid_30d_customers[valid_30d_customers['revenue'] >= revenue_threshold]

# 2. VIP 중 지연 배송 경험자 수
late_vips = vip_customers[vip_customers['is_late'] == 1]
num_late_vips = len(late_vips)

# 3. 손실 계산 (정시 1.64% vs 지연 1.17%)
lost_vip_orders = num_late_vips * (0.0164 - 0.0117)
avg_vip_revenue = vip_customers['revenue'].mean()
vip_revenue_loss = lost_vip_orders * avg_vip_revenue

print(f"💰 VIP 지연 이탈 분석 결과")
print(f"- 지연을 경험한 VIP 고객 수: {num_late_vips:,}명")
print(f"- 배송 지연으로 잃어버린 잠재적 VIP 매출액: {int(vip_revenue_loss):,} BRL")
```

VIP 지연 이탈 분석 결과

- 지연을 경험한 VIP 고객 수: 1,699명
- 배송 지연으로 잃어버린 잠재적 VIP 매출액: 3,476 BRL

3. [운영] 카테고리별 "지연 민감도" 분석

단순 리텐션 하락을 **돈(Revenue)**으로 환산합니다. 우량 고객군(상위 20%)을 따로 떼어내어 분석합니다.

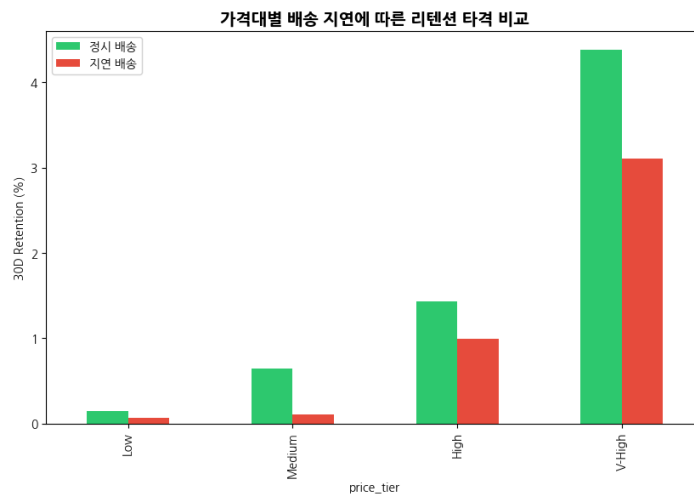
▼ 코드

```
# [심화 2] VIP 고객의 배송 지연 이탈 비용(Revenue Loss)
# 1. 고단가 고객(VIP) 기준 설정 (매출 상위 20%)
revenue_threshold = cf['revenue'].quantile(0.8)
vip_customers = valid_30d_customers[valid_30d_customers['revenue'] >= revenue_threshold]
```

```
# 2. VIP 중 지연 배송 경험자 수
late_vips = vip_customers[vip_customers['is_late'] == 1]
num_late_vips = len(late_vips)

# 3. 손실 계산 (정시 1.64% vs 지연 1.17%)
lost_vip_orders = num_late_vips * (0.0164 - 0.0117)
avg_vip_revenue = vip_customers['revenue'].mean()
vip_revenue_loss = lost_vip_orders * avg_vip_revenue

print(f"💰 VIP 지연 이탈 분석 결과")
print(f"- 지연을 경험한 VIP 고객 수: {num_late_vips:,}명")
print(f"- 배송 지연으로 잃어버린 잠재적 VIP 매출액: {int(vip_revenue_loss):,} BRL")
```



▼ Retention (코호트)

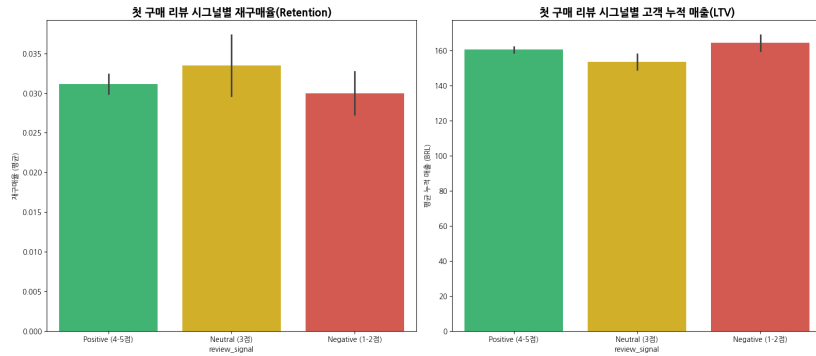
- 재구매율이 대부분 코호트에서 1~7% 수준이고 특히 2018년 코호트로 갈수록 1~2%대까지 내려감 (첫 구매 이후 두 번째 구매로 연결되는 힘이 약하다)
- rolling 7d, 30d retention도 전반적으로 낮다
 - rolling 30d(비교 가능한 코호트)에서 2017년 중반에 3%까지 보이다가 2017년 말~2018년에 들어서면서 1%대 이하로 꺾여 내려가는 패턴
- 신규 유입은 존재하지만 반복 구매로 이어지지 않음
- 시장 자체가 단발성 구매로 그치는 시장일 확률이 높지 않을까

🟡 Revenue (수익)

▼ 리뷰4점이상+1-2점 부정 신호 분석

Revenue & Retention: 부정 신호의 경제적 타격 분석

- **리뷰 4점 이상 (Positive):** 활성화(Activation)에 성공한 그룹으로, 서비스에 대한 확신을 얻어 재구매로 이어질 가능성이 가장 높은 핵심 타겟입니다.
- **리뷰 1-2점 (Negative):** 강력한 **이탈 신호(Churn Signal)**입니다. 이들은 단순히 만족도가 낮은 것을 넘어, 주변에 부정적인 구전을 퍼뜨릴 가능성이 높은 위험군입니다.
- **Revenue 전략:** 1-2점을 준 고객들의 **평균 매출액(LTV)**을 확인하세요. 만약 이들의 LTV가 높다면, "우리의 VIP 후보들이 배송 문제로 인해 서비스를 등지고 있다"는 아주 임팩트 있는 결론을 낼 수 있습니다.



핵심 지표 비교 결과

리뷰 시그널	고객 수	재구매율	평균 LTV (BRL)
Positive (4~5점)	73,461	3.11%	160.52
Negative (1~2점)	14,005	3.00%	164.48

주요 발견 요약

[Revenue & Retention: 리뷰 시그널 분석 결과]

현상: 첫 구매에서 부정 신호(1~2점)를 보낸 고객군이 긍정 신호(4~5점) 고객군보다 인당 평균 매출(LTV)이 약 2.5% 더 높음.

원인 진단 (Why?):

저가 소모품 구매자보다 고단가 상품 구매자가 서비스 품질(배송 속도, 상품 상태)에 대해 훨씬 엄격한 잣대를 가지고 있음.

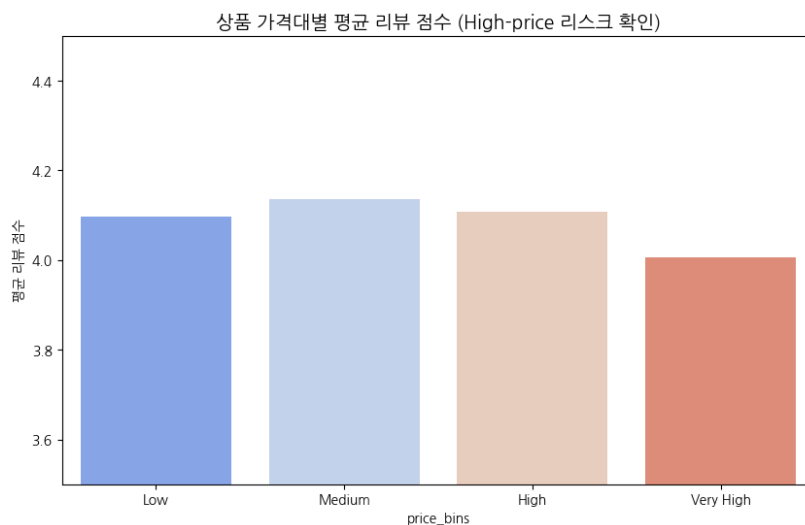
고가치 고객(High-Value Customer)이 이탈 위험군(Negative Signal)으로 분류되는 비율이 높아, 매출 손실 리스크가 매우 큼.

전략적 제안:

고단가 카테고리 집중 관리: 특정 금액 이상의 주문건에 대해서는 '프리미엄 배송' 또는 '실시간 배송 알림'을 강화하여 부정 신호 발생을 선제적으로 차단해야 함.

부정 신호 고객 리커버리: 1~2점을 남긴 고객 결제 고객에게는 즉시 CS팀이 개입하거나 보상 쿠폰을 지급하여 LTV 손실을 방어하는 '리커버리 프로세스' 도입이 시급함.

비싼 물건이 정말 리뷰 점수가 낮을까?" 확인



원인 분석: "기대치가 높은 고가치 고객(VVIP)"

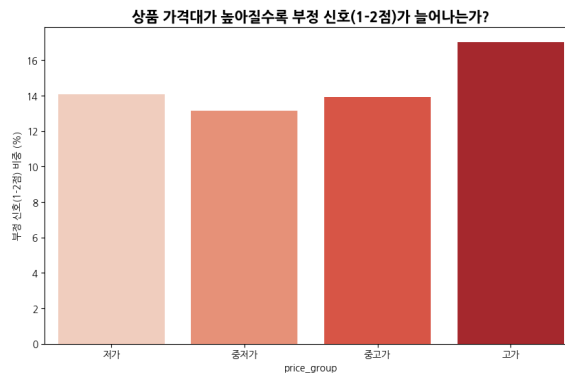
- 일반적으로 고가 상품(가전, 가구 등)을 구매하는 고객은 배송 지연이나 상품 파손에 훨씬 민감합니다.

- 저가 소모품 구매자는 배송이 조금 늦어도 '그럴 수 있지' 하고 넘어가지만, 고단가 결제 고객은 조금만 약속이 어긋나도 1~2점의 부정 신호를 보냅니다.
- 결과적으로 Negative 그룹에 고단가 주문이 많이 포함되면서 평균 LTV가 높게 측정된 것입니다.

3. 리텐션의 한계: "만족해도 안 오고, 화나도 안 온다"

- 두 그룹 모두 재구매율이 3%대에 머물러 있다는 점은 리뷰 점수가 재구매를 결정하는 유일한 변수가 아님을 시사합니다.
- 고객이 5점을 줘도 재구매할 만한 '두 번째 필요'를 느끼지 못하거나, 브라질 시장 특성상 재방문 유인(쿠폰, 포인트 등)이 부족할 가능성이 큼니다.

정말 비싼 물건을 산 사람들이 화가 난 걸까?"를 증명하기 위해, 가격대별 부정 신호(1-2점)의 비중을 확인



▼ Revenue (월별)

- GMV, 주문, 고객 수는 2017 → 2018로 크게 성장해서 2018년에는 월 GMV가 대체로 100만대 정도로 안정적으로 유지된다
- AOV는 대체로 155~170 근처로 안정적이다
- freight_share(배송비 비중)은 2018년에 대략 13~15%대이고 이후 약간 올라가는 구간이 있다
- (정말 적긴 하지만) 재구매 고객과 1회 구매 고객이 뭐가 다른가
 - 재구매 고객은 '첫 구매 후 30일 내 재구매'와 강하게 연결된다
 - 재구매 고객이 첫 30일 안에 두 번째 구매한 비율이 41.63%이다
 - uplift - 그렇다면 early repeat(빠르게 재구매하는 것)을 올리는 것은 무엇일까
 - 첫 번째 구매에서 item_cnt(장바구니의 크기)
 - 1개 → 3개 이상이 0.59% → 1.38%로 약 2.3배 차이
 - 첫 구매에서 여러 개를 산 고객은 30일 안에 재구매를 할 확률이 2배 이상 높다
 - **선희님 의견 - 통계적으로 유의미한가 → 점검**

	items_bucket	early_30d
0	1개	0.59
1	2개	0.84
2	3개 이상	1.38

ITEM_CNT, ITEM_BUCKET에 관하여

```
orders = pd.read_csv("order_fact.csv")
customers = pd.read_csv("aarrrr_master_customer_fact.csv")

orders.head(), customers.head()

orders['order_purchase_timestamp'] = pd.to_datetime(orders['order_purchase_timestamp'])
orders['order_delivered_customer_date'] = pd.to_datetime(
    orders['order_delivered_customer_date'],
    errors='coerce'
)

orders = orders[orders['is_delivered'] == 1].copy()
```

```

# 고객별 첫 구매 시점
first_purchase = (
    orders.groupby('customer_unique_id')['order_purchase_timestamp']
    .min()
    .reset_index(name='first_purchase_ts')
)

orders = orders.merge(first_purchase, on='customer_unique_id', how='left')

# 첫 구매 후 경과 일수
orders['days_since_first'] = (
    orders['order_purchase_timestamp'] - orders['first_purchase_ts']
).dt.days

early_repeat = (
    orders[
        (orders['days_since_first'] > 0) &
        (orders['days_since_first'] <= 30)
    ]
    .groupby('customer_unique_id')['order_id']
    .nunique()
    .gt(0)
    .reset_index(name='early_30d')
)

customer = (
    orders.groupby('customer_unique_id')
    .agg(
        orders_cnt=('order_id', 'nunique'),
        first_items_cnt=('items_cnt', 'first')
    )
    .reset_index()
)

customer['is_repeat'] = customer['orders_cnt'] >= 2

customer = customer.merge(early_repeat, on='customer_unique_id', how='left')
customer['early_30d'] = customer['early_30d'].fillna(False)

summary = (
    customer.groupby('is_repeat')
    .agg(
        customers=('customer_unique_id', 'count'),
        early_30d_rate=('early_30d', 'mean')
    )
    .reset_index()
)

summary['early_30d_rate'] = (summary['early_30d_rate'] * 100).round(2)
summary

customer['items_bucket'] = pd.cut(
    customer['first_items_cnt'],
    bins=[0, 1, 2, 999],
    labels=['1개', '2개', '3개 이상']
)

uplift = (
    customer.groupby('items_bucket')['early_30d']
    .mean()
    .reset_index()
)

```

```
uplift['early_30d'] = (uplift['early_30d'] * 100).round(2)
uplift
```

● Referral (추천 · Proxy 지표)

▼ Referral에 관하여

1. referral 둘러봤는데 전체 고객 수 96096명 주에서 referral 고객 수가 2301명으로 비율은 2.39 %로 나왔다.
2. orders_cnt(주문 수)의 평균은 non-referral은 1.00이었지만 referral은 2.12였고 리뷰 점수의 평균도 4.76으로 non-referral에 비해 0.7 점 더 높았다. 매출 평균 또한 두배였음.
3. referral의 기준은 `customer["is_referral"] = (customer["review_score"] >= 4) & (customer["orders_cnt"] >= 2)`로 리뷰 점수와 주문 수로 함

전체 고객 수: 96,096
Referral 고객 수: 2,301
Referral 비율: 2.3945%

	is_referral	customers	avg_orders	avg_review	avg_revenue
0	Non-Referral	93795	1.008071	4.067749	157.079427
1	Referral	2301	2.124728	4.768796	299.520595

```
import pandas as pd

customer = pd.read_csv("aarr-master_customer_fact.csv")
orders = pd.read_csv("order_fact.csv") # 이번 지표 계산엔 필수는 아니지만 같이 로드해둬م

# Referral 정의: (review_score >= 4) AND (orders_cnt >= 2)
customer["is_referral"] = (customer["review_score"] >= 4) & (customer["orders_cnt"] >= 2)

# 전체 결과
total_customers = len(customer)
referral_customers = int(customer["is_referral"].sum())
referral_rate = referral_customers / total_customers

print(f"전체 고객 수: {total_customers:,}")
print(f"Referral 고객 수: {referral_customers:,}")
print(f"Referral 비율: {referral_rate:.4%}")

# Referral vs Non-Referral 요약 비교
summary = customer.groupby("is_referral").agg(
    customers=("customer_unique_id", "count"),
    avg_orders=("orders_cnt", "mean"),
    avg_review=("review_score", "mean"),
    avg_revenue=("revenue", "mean"),
).reset_index()

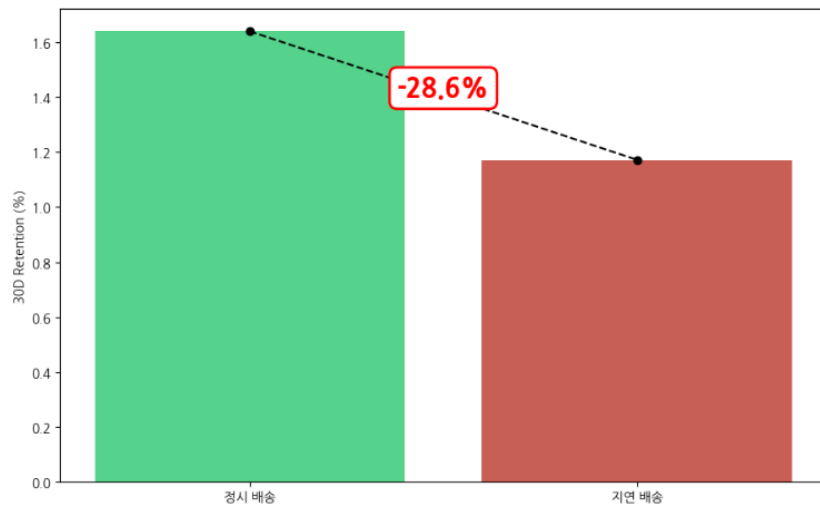
summary["is_referral"] = summary["is_referral"].map({False: "Non-Referral", True: "Referral"})
summary
```

“추천할 만한 고객”이라는 정의에 데이터가 잘 부합하는 것으로 보인다.

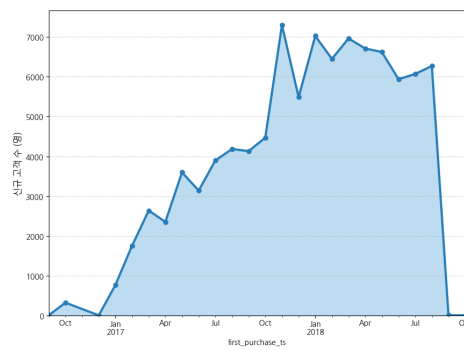
아영메모 (스토리 라인 고민)

- **Acquisition:** "사람은 많이 들어옵니다."
- **Activation:** "97%가 물건을 잘 받습니다. 겉보기엔 문제없어 보입니다."
- **Retention (Why?):** "하지만 뜯어보니 배송이 늦으면 리텐션이 28%나 박살 납니다. 특히 돈 많이 쓰는 VIP들이 여기서 다 나갑니다."
- **Revenue/Action:** "따라서 우리는 물류를 고치거나, 자연 고객을 즉시 케어해야 합니다."

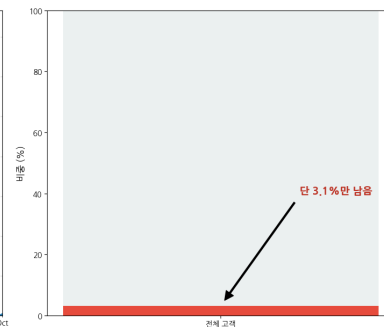
지연 배송 시 리텐션 28.6% 급감 (절벽)



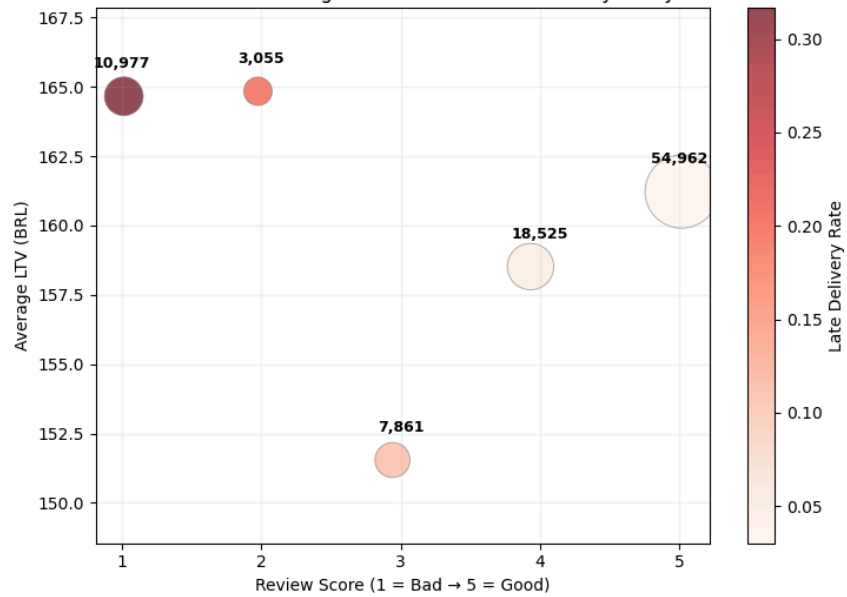
신규 유입: 지속적인 우상향 성장

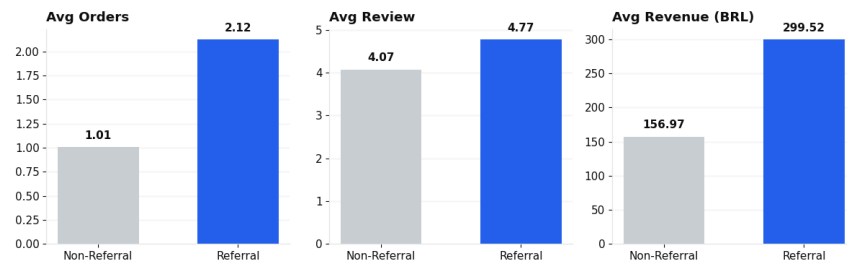


리텐션: 3.1%의 냉혹한 현실



Revenue at Risk: High-LTV Customers vs Delivery Delay





회고