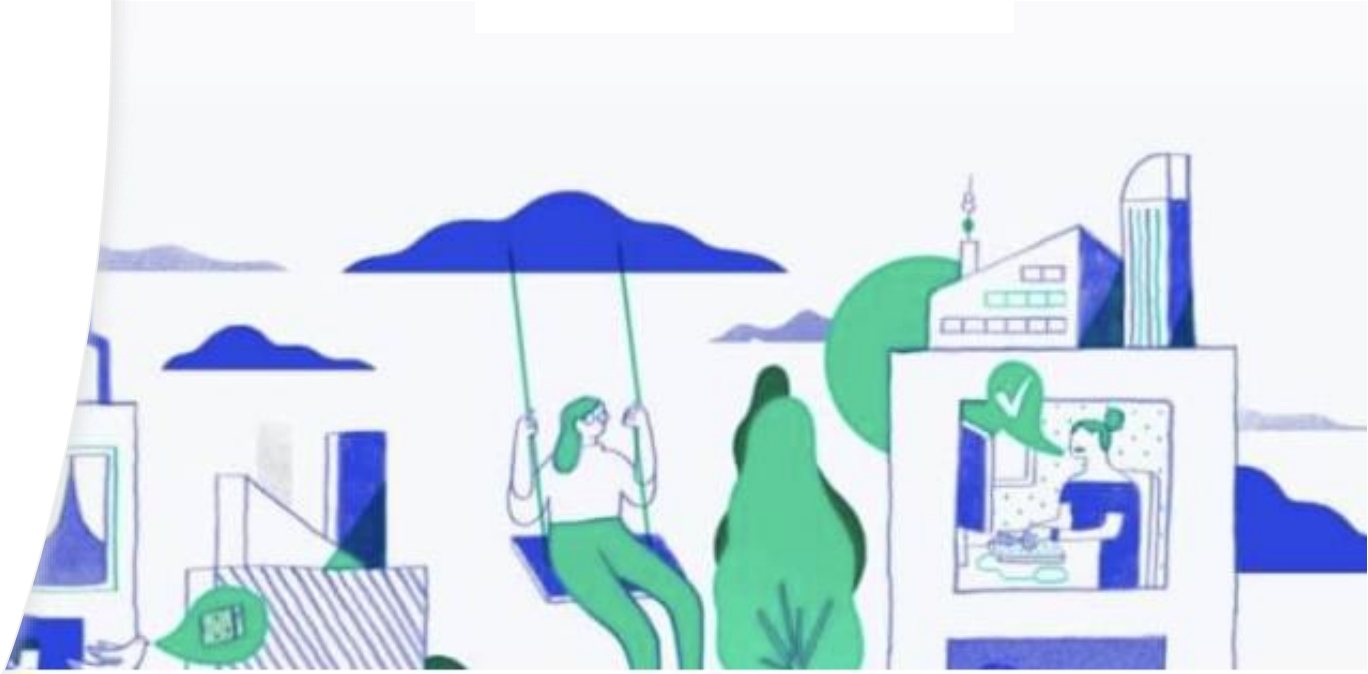
브라질
이커머스 플랫폼
고객만족 분석

Business Analysis & Customer Satisfaction prediction

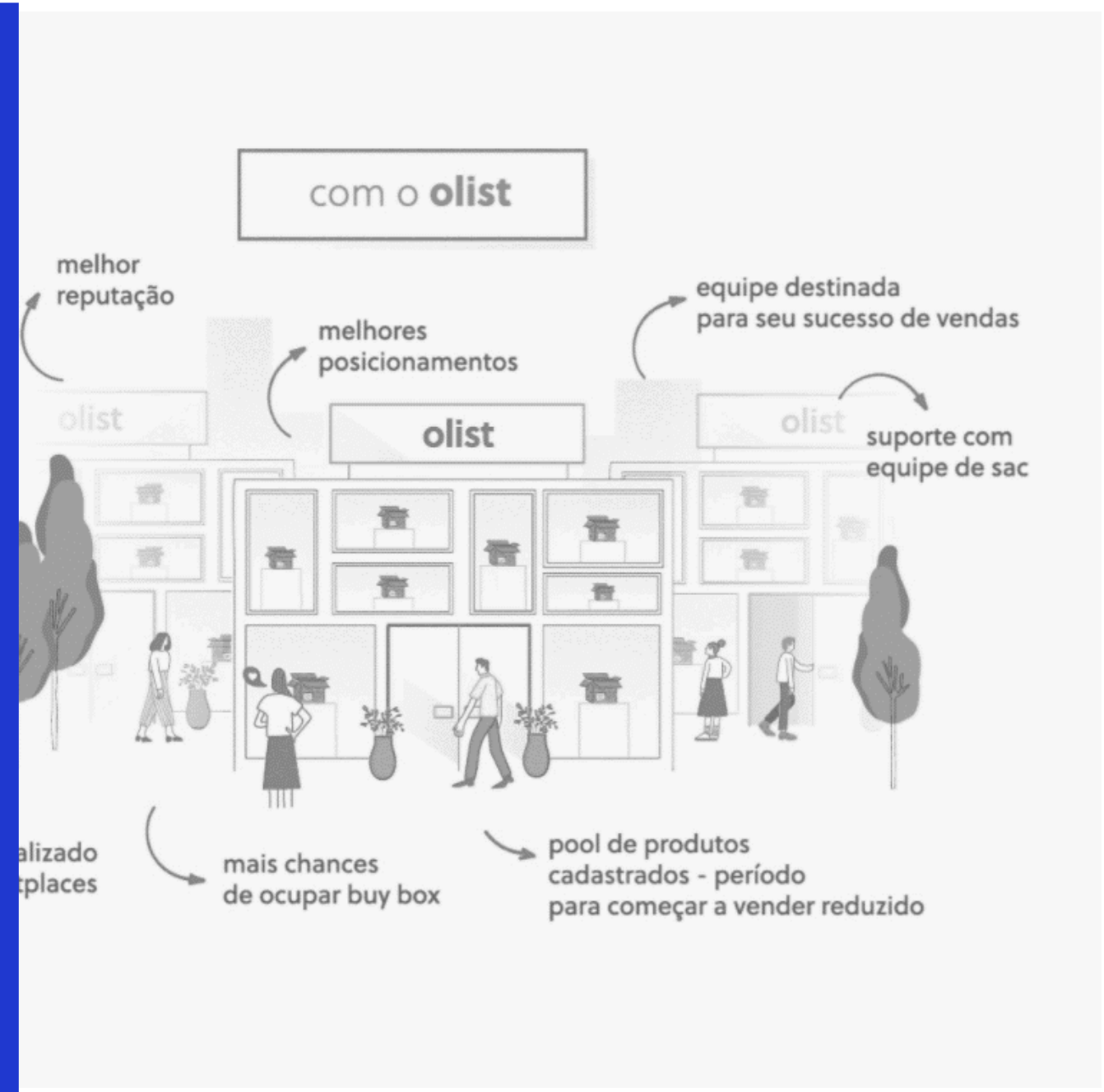
olist



2021. 3. 25

BRO (브라질넛트오일)

CONTENTS



데이터셋 선정 배경

PROJECT OVERVIEW

01

- **프로젝트 방향성** : 실제 기업 데이터 분석을 통해 비즈니스 기회 요소를 찾고, 개선 방향을 제안해보자
- **관심산업 및 분야** : Retail / E-commerce / 고객분석
- **최종 선정 데이터셋** : Brazilian E-Commerce Public Dataset by Olist (https://www.kaggle.com/olistbr/brazilian-ecommerce)

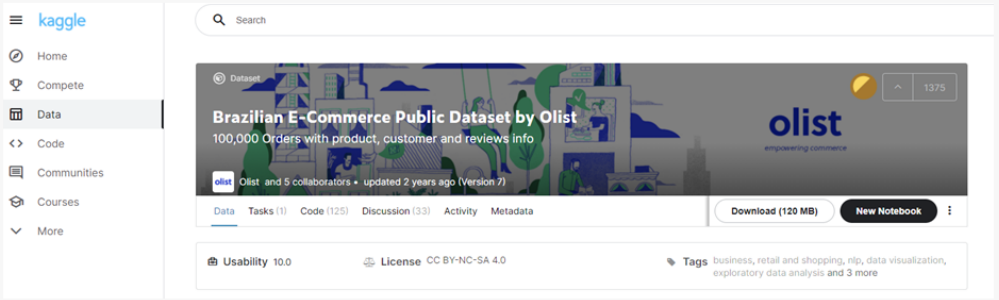
데이터 선정 배경

실제 기업 Data 활용

비즈니스 문제 정의 및 인사이트 도출

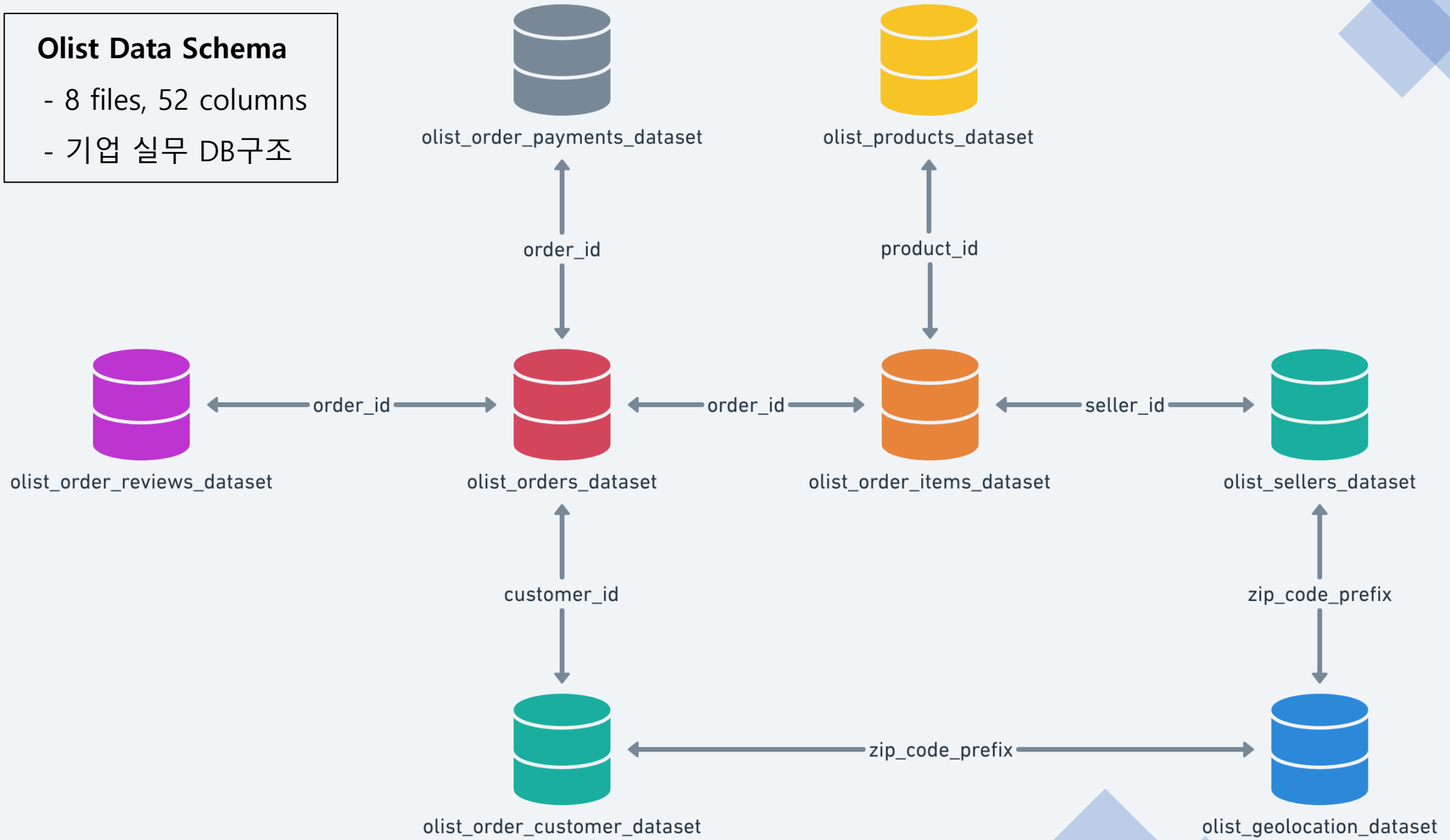
Retail/E-commerce/ 고객분석

Kaggle



Olist Data Schema

- 8 files, 52 columns
- 기업 실무 DB구조



기업 소개

OLIST OVERVIEW

02

Olist?

온라인 판매를 원하는 **셀러**가 브라질 주요 **마켓플레이스에 입점하고 판매부터 배송까지 전 과정을 간편하게 관리**할 수 있는 기능을 제공하는 **통합관리 솔루션 업체**

► 설립연도

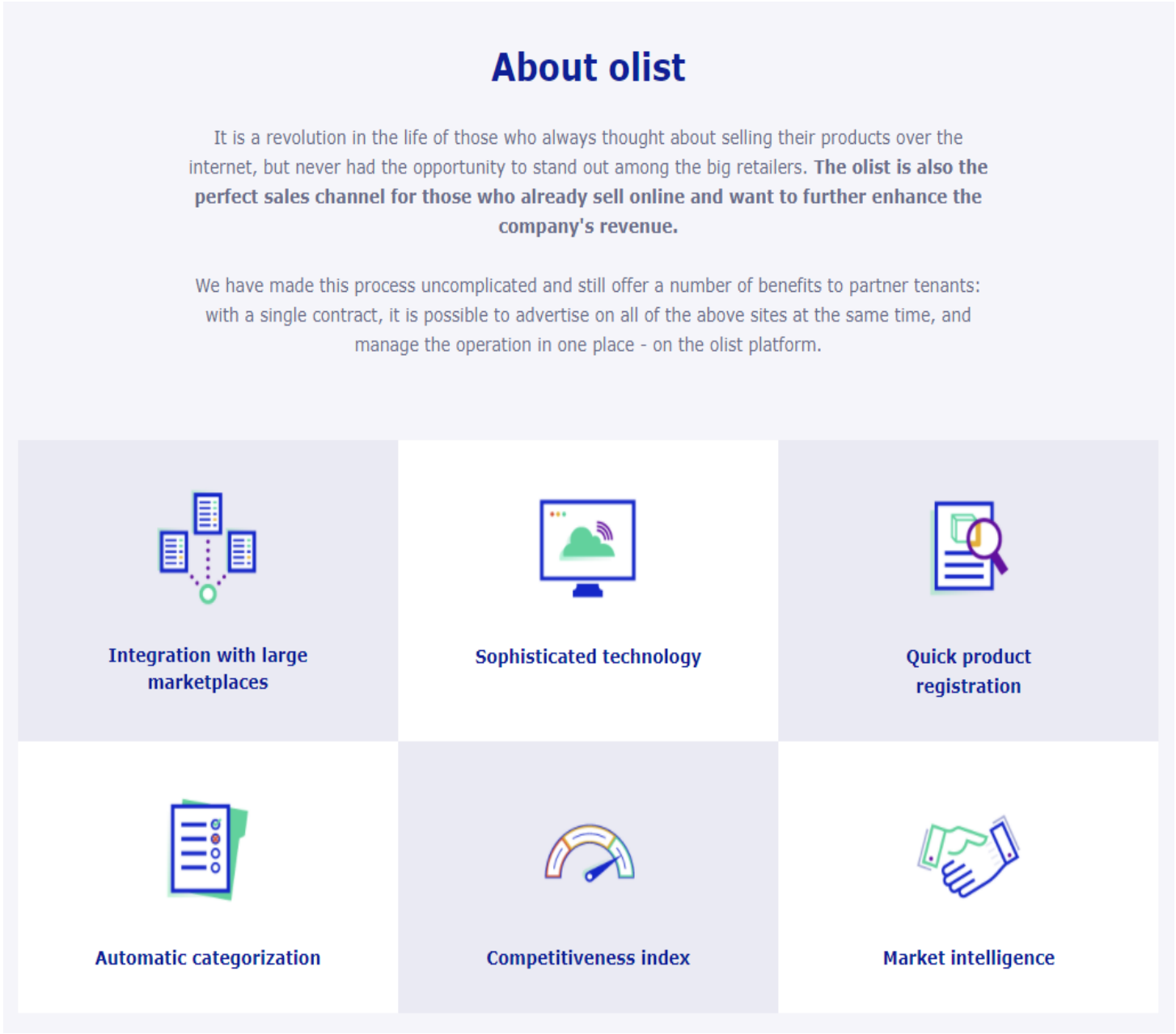
- 2015년

► 제휴 마켓플레이스

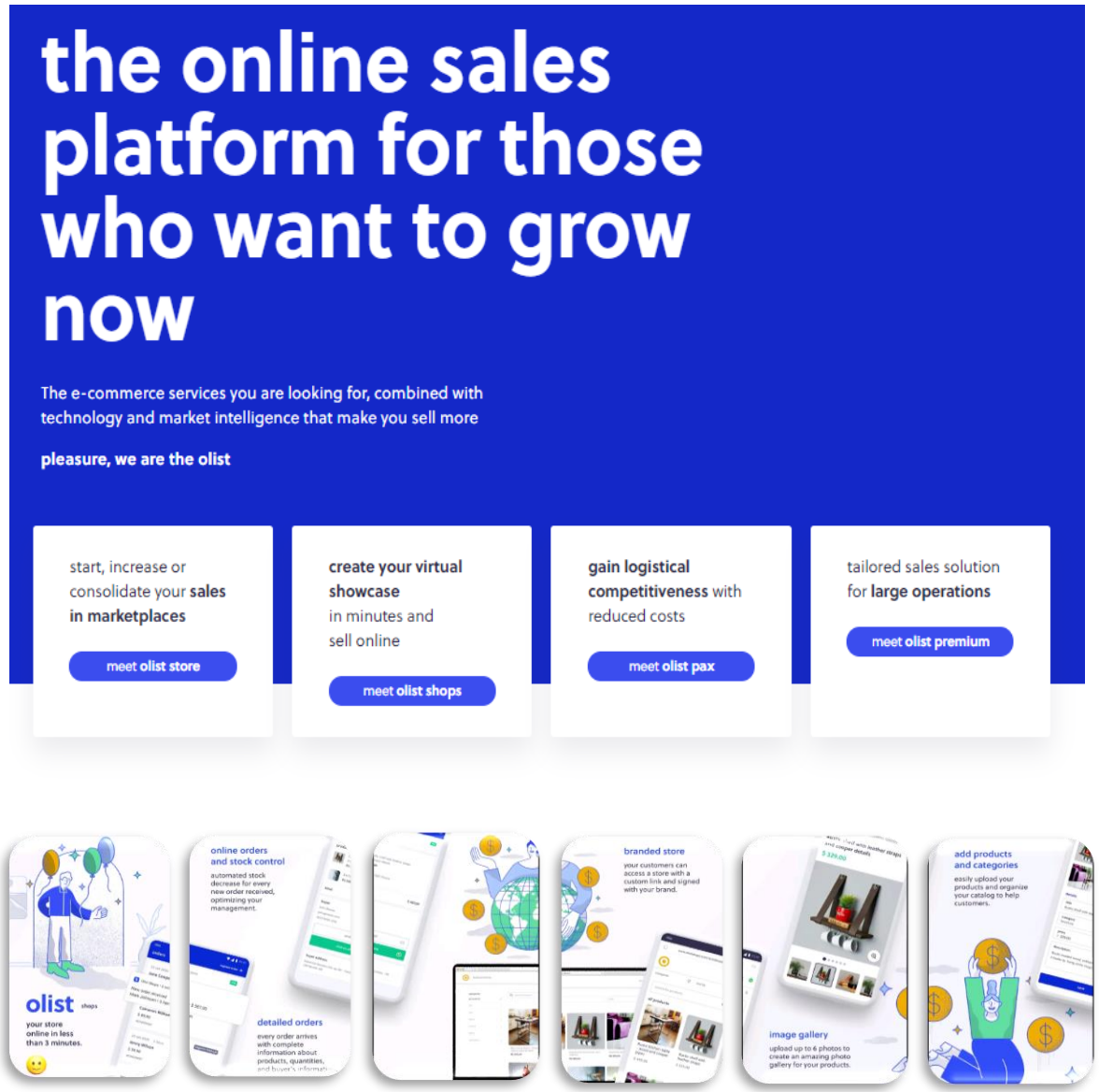
- Amazon, Mercado Livre, Americanas 등

► 제공 서비스

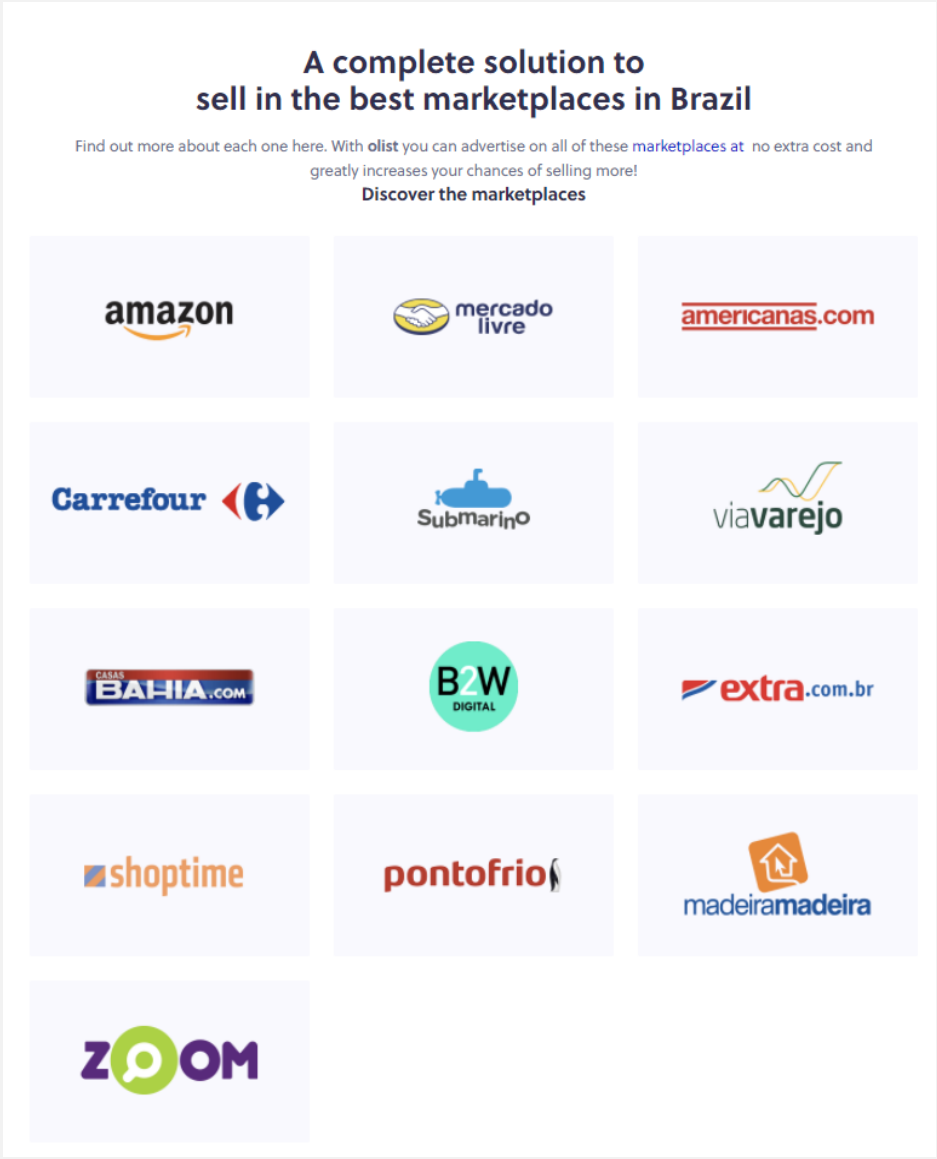
- 마켓플레이스별 주문, 결제, 배송, CS 통합관리
- 제품 홍보 콘텐츠(카탈로그) 컨설팅
- 최적 가격 컨설팅
- 공식 물류 파트너인 브라질 우체국을 통해 15% 배송비용 절감 및 배송기간 단축



Olist 서비스 상세 페이지



Olist 통해 입점 가능한 쇼핑몰



수익 구조

1) 제품 판매 건당 중개 수수료
2) 플랫폼 월 사용료
3) 멤버쉽 가입비

→ 제품을 구매하는 일반 고객이 아닌 **셀러로부터 매출이 발생**하는 구조

OLIST KEY BUSINESS AGENDA

1) 신규 셀러 유치

2) 기존 셀러 멤버십 업그레이드

3) 기존 셀러 거래규모↑

lite

R $ 29.90 / month
R $ 29.90 cost of membership *

21% commission per product
+ shipping fee
understand

Create your account

what's included?

- maximum registration of 30 products
- preconfigured ads
- competitiveness analysis tool
- financial reconciliation of sales channels

recommended

pro

R $ 249.90 / month
R $ 349 cost of membership *

19% commission per product
+ shipping fee
understand

Create your account

what's included?

- unlimited product registration
- preconfigured ads
- competitiveness analysis tool
- financial reconciliation of sales channels
- sale without barcode *
- order pickup *
- access to sales consultancy and on-call service

premium

request contact and learn more

olist's enterprise solution for large companies

talk to a consultant

what's included?

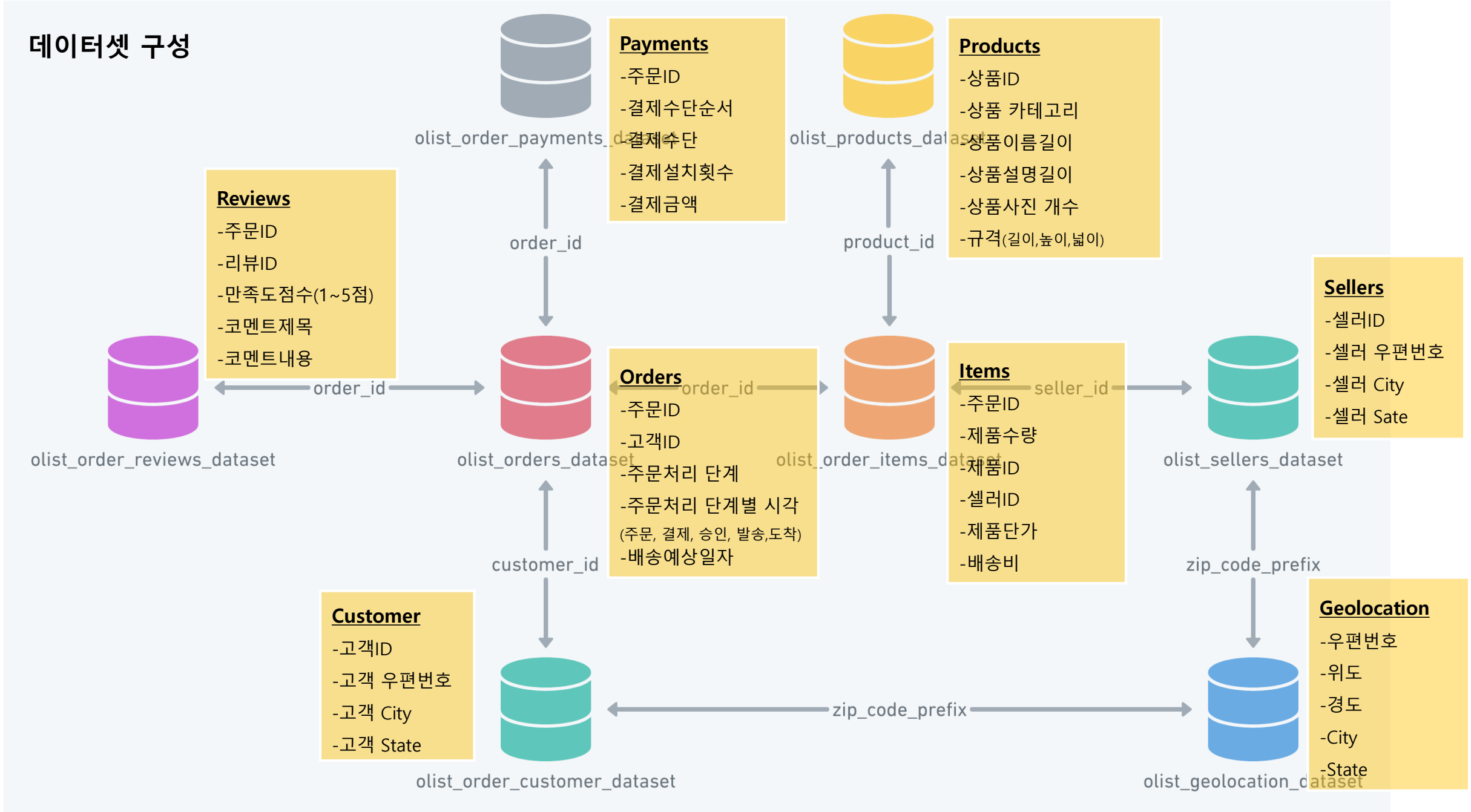
- account Manager
- customizable solutions
- Market intelligence
- increase in marketshare and brand presence
- competitiveness analysis tool
- participation of promotional campaigns in channels and marketplaces
- catalog curation and ad optimization
- financial reconciliation of sales channels
- order pickup *

비즈니스 분석

BUSINESS ANALYSIS

03

데이터셋 구성

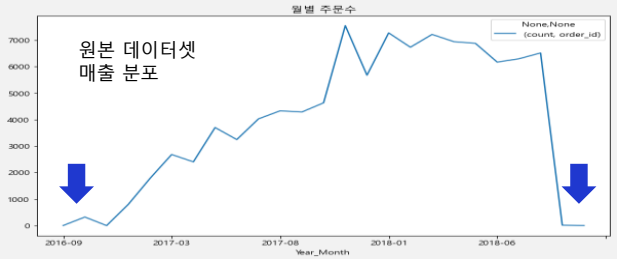

Payments
-주문ID
-결제수단순서
-결제수단
-결제설치횟수
-결제금액
olist_order_payments_dataset
Products
-상품ID
-상품 카테고리
-상품이름길이
-상품설명길이
-상품사진 개수
-규격(길이,높이,넓이)
olist_products_dataset
Reviews
-주문ID
-리뷰ID
-만족도점수(1~5점)
-코멘트제목
-코멘트내용
order_id
product_id
olist_order_reviews_dataset
Orders
-주문ID
-고객ID
-주문처리 단계
-주문처리 단계별 시각
(주문, 결제, 승인, 발송,도착)
-배송예상일자
olist_orders_dataset
order_id
Items
-주문ID
-제품수량
-제품ID
-셀러ID
-제품단가
-배송비
olist_order_items_dataset
seller_id
Sellers
-셀러ID
-셀러 우편번호
-셀러 City
-셀러 Sate
olist_sellers_dataset
customer_id
zip_code_prefix
Customer
-고객ID
-고객 우편번호
-고객 City
-고객 State
olist_order_customer_dataset
zip_code_prefix
Geolocation
-우편번호
-위도
-경도
-City
-State
olist_geolocation_dataset

비즈니스 분석 목적

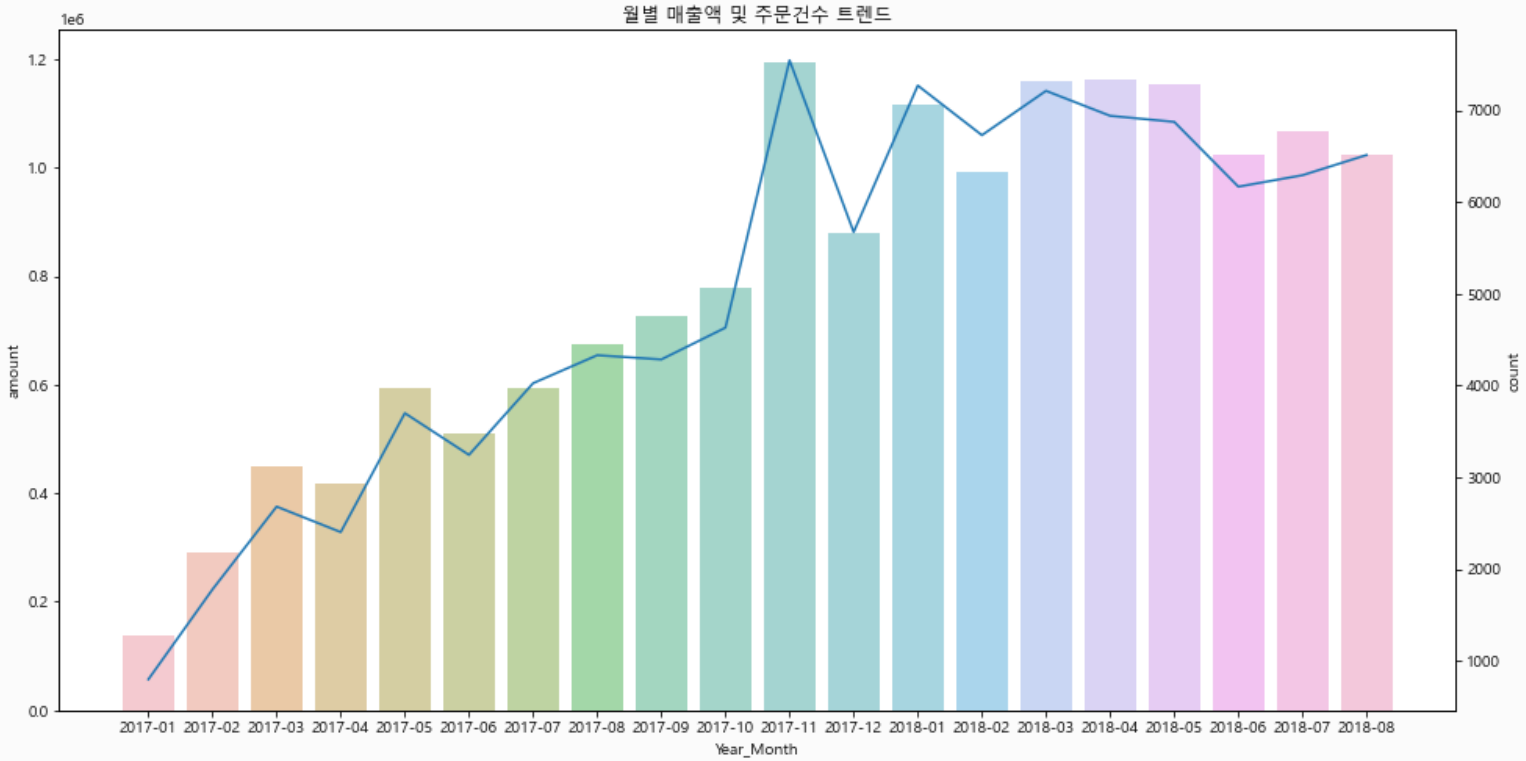
- 다각도의 데이터 EDA를 통해 비즈니스 현황을 파악하고 주요 사업 이슈를 찾고자 함

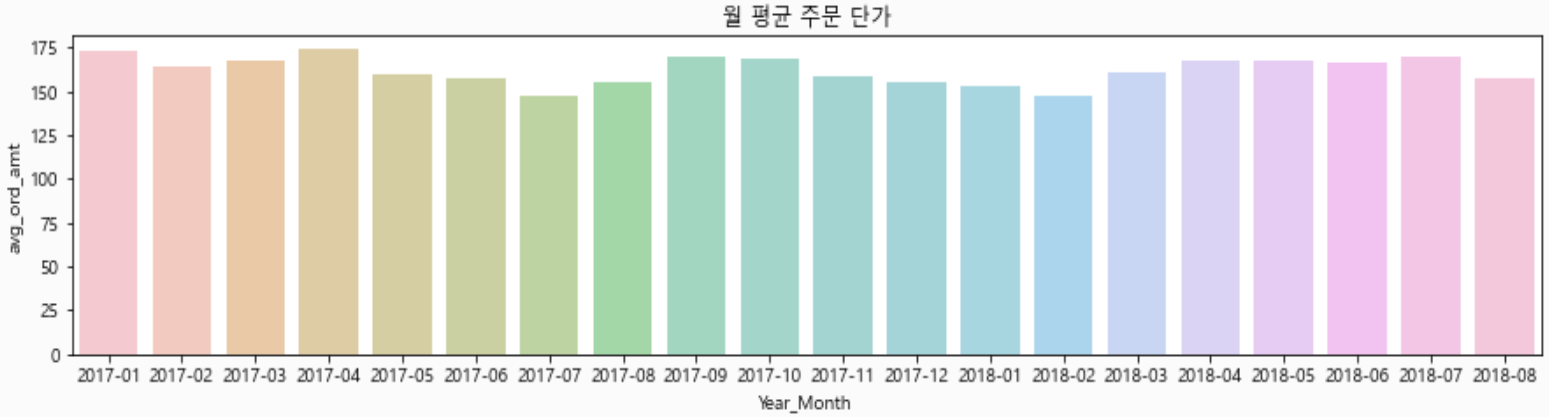
Sales

- **매출발생 기간**
  - 2016년 9월~2018년 12월
- **데이터셋 특이사항**
  - 주문 데이터가 존재하는 전체 기간 중 첫 3개월, 마지막 1개월은 주문량이 현저하게 낮아 데이터 수집에 이슈가 있었을 것으로 판단
  - → **2017년 1월~2018년 8월로 분석기간 조정**



- **매출 추이**
  - 월 매출액 및 주문건수 상승, 월 평균 주문단가 일정
  - → 거래량 증가를 통한 매출 상승

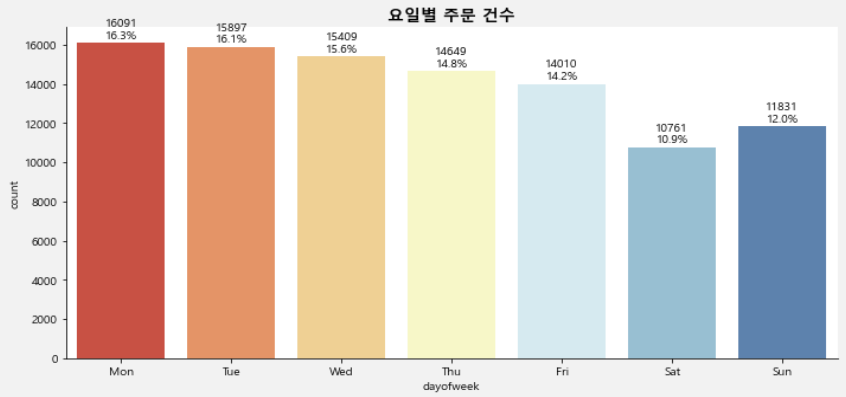


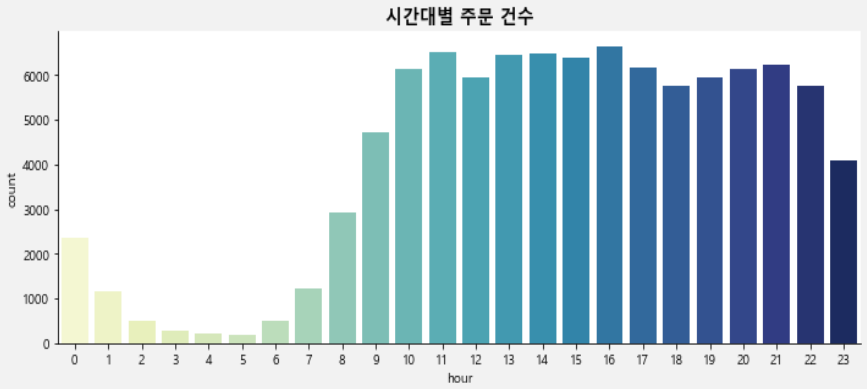


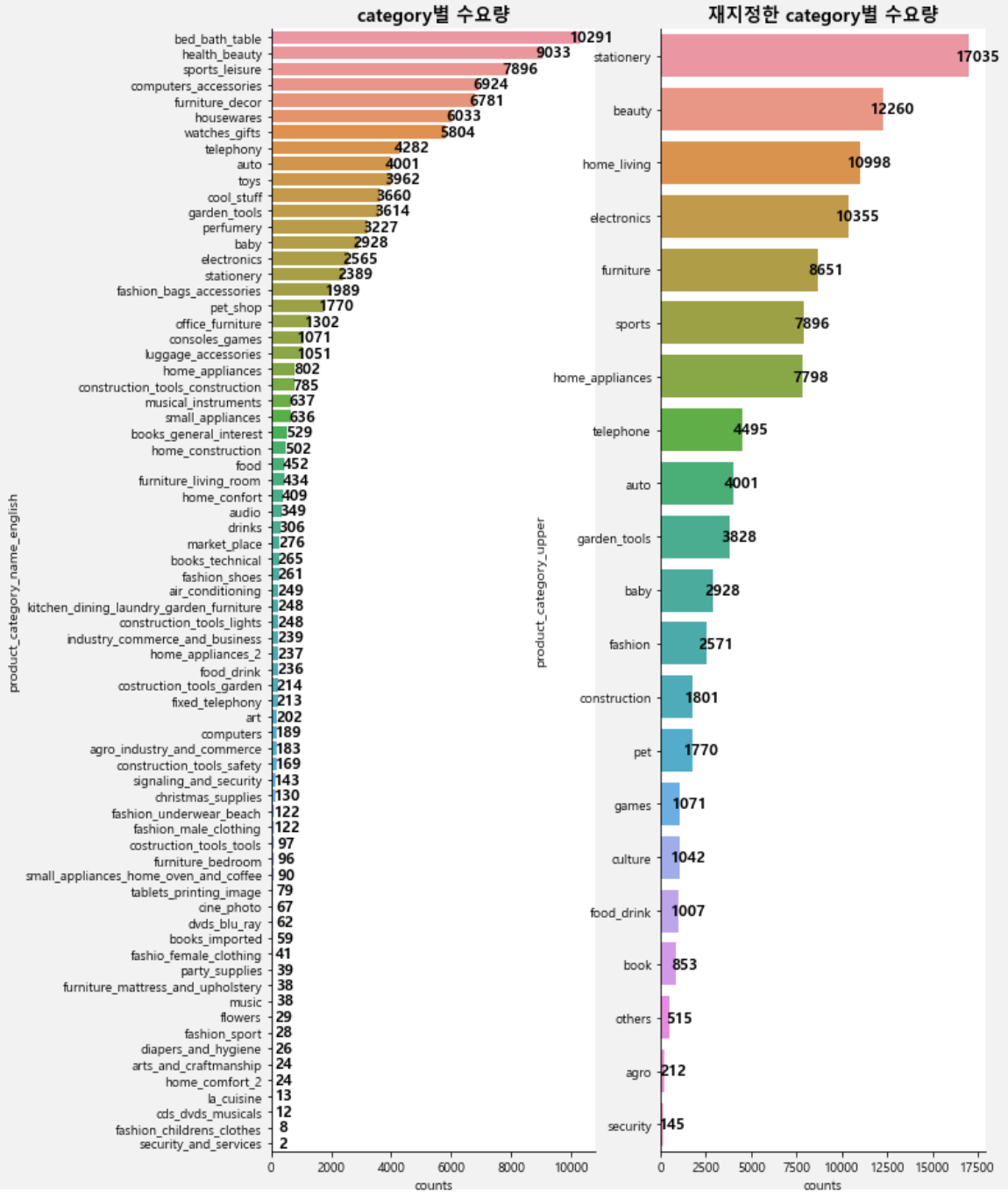
Sales Analysis (플랫폼 기준)

- **주문 트렌드**
  - 주말보다는 평일에 주문이 많이 발생
  - 오전 10시 ~밤 10시 사이 큰 편차 없이 일정한 주문량

- **제품 카테고리**
  - 중복된 카테고리명이 많아 카테고리명 재지정 (76→21개로 조정)
  - 문구 > 뷰티 > 리빙 > 가전 >가구 순 많은 주문 발생





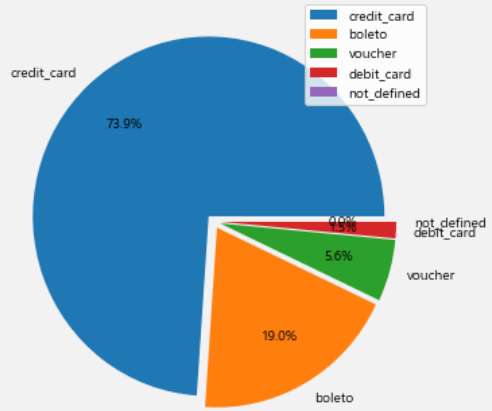


Operations

▪ 결제수단

- 신용카드 74% > 볼레토* 19% > 바우처 5.6% > 직불카드 1.5%

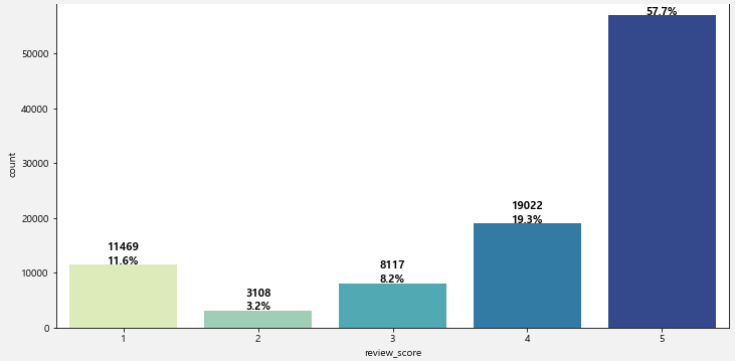
(* 볼레토 : 브라질 은행 결제수단)



▪ 고객 만족도 분포

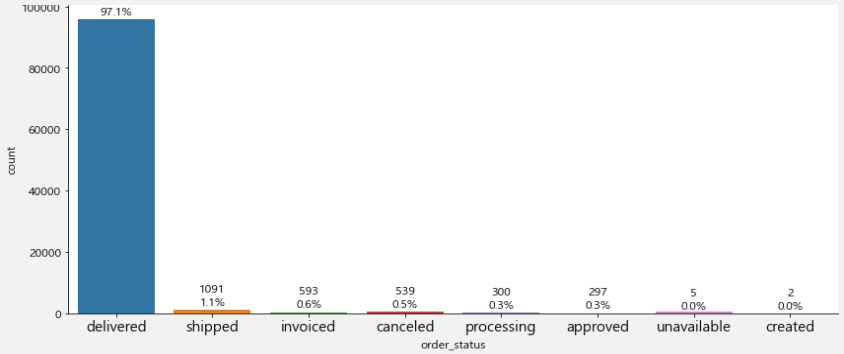
- 매우 만족(5) 57.5% / 만족(4) 19.3% / 보통(3) 8.2%

불만족(2) 3.2% / 매우 불만족(1) 11.6%



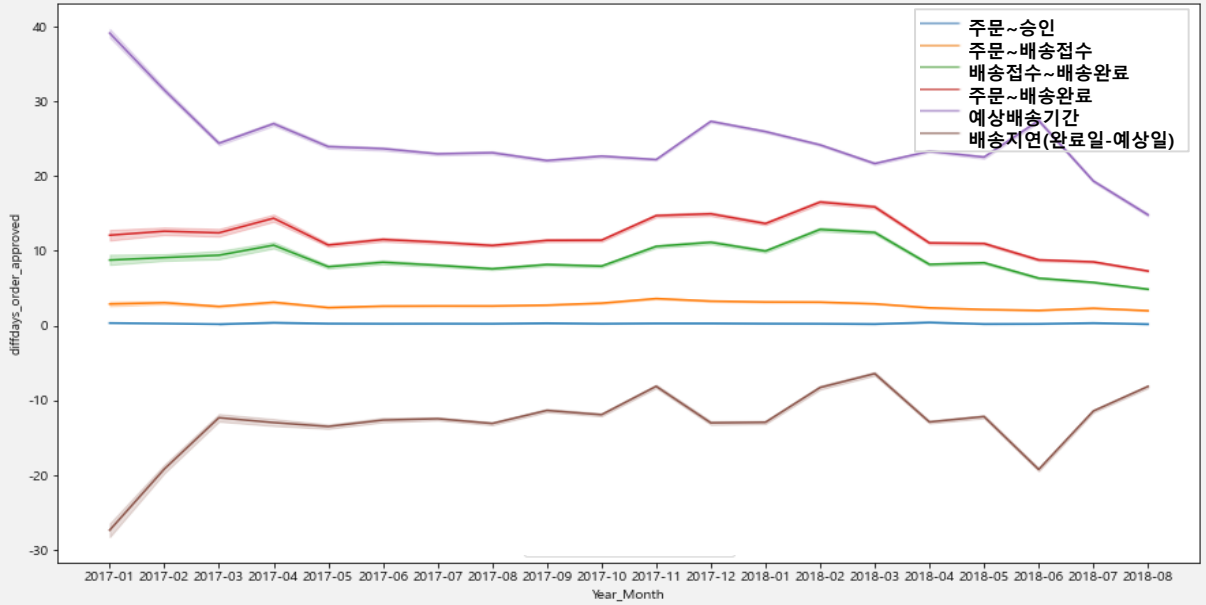
▪ 주문 처리상태

- delivered, shipped, invoiced, canceled, processing, approved, unavailable, created로 분류
- 97.1%가 주문완료(delivered) 상태



▪ 배송 단계별 평균 소요일수

- 주문~배송접수 : 3일 / 주문~배송완료: 12일 / 예상 배송 소요기간: 23일

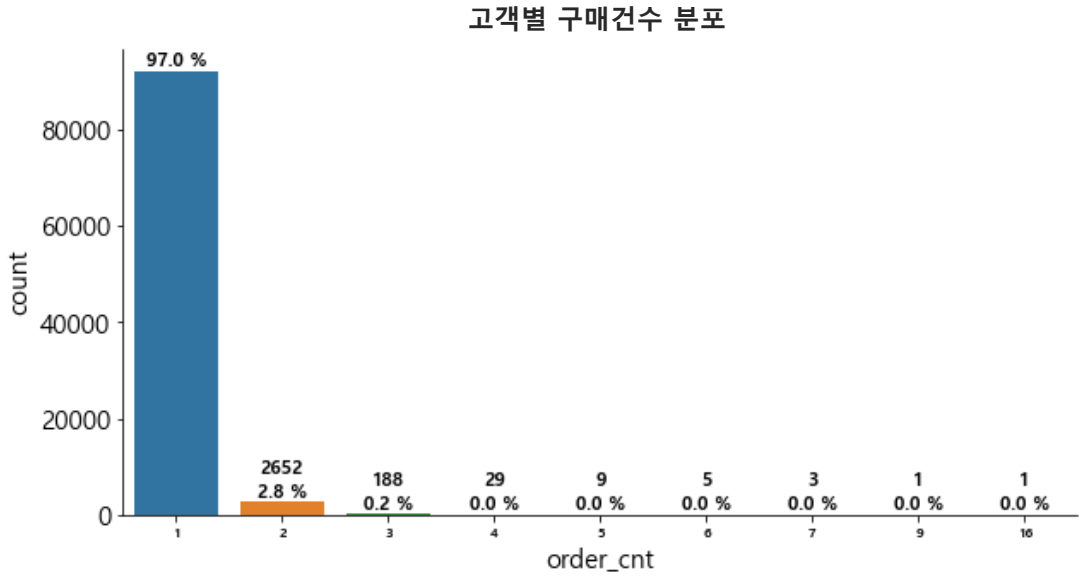


Customers & Sellers

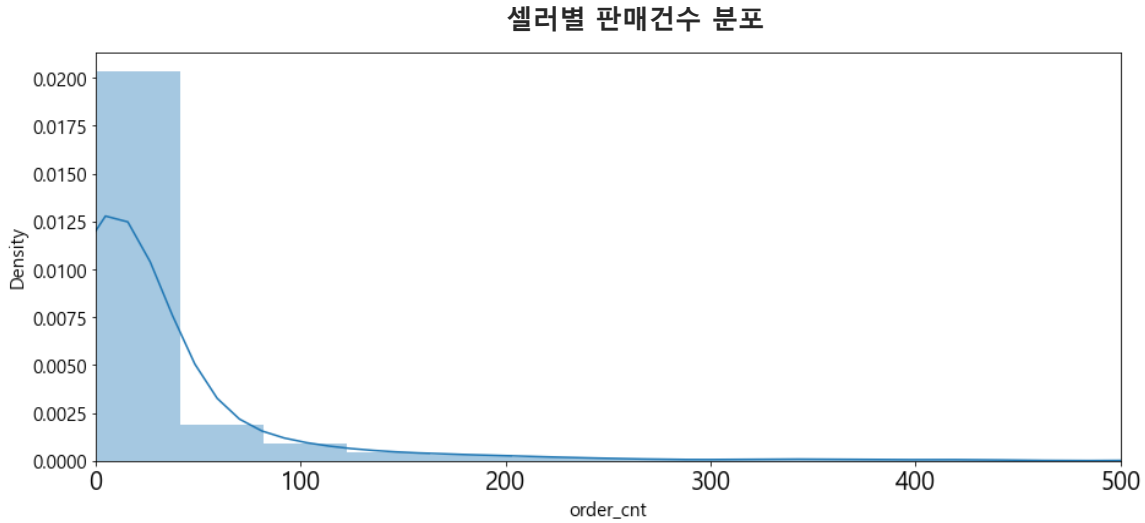
- **Customer**
  - 94,989명 중 약 97%는 구매 내역이 1회
  - Customer retention 관련해서는 분석이 어려울 것으로 판단

- **Seller**
  - 3,029명 중 약 18%는 판매 횟수가 1회
  - 판매 건수 분포는 평균 37건, 중간값 8건, 최대값 2025건
  - 중앙값과 최대값의 편차가 크므로 판매실적이 우수한 셀러들이 존재하는 것으로 보임

Customer	
count	94,989
mean	1.03
std	0.21
min	1
25%	1
50%	1
75%	1
max	16



Seller	
count	3,029
mean	36.89
std	119.99
min	1
25%	2
50%	8
75%	25
max	2025

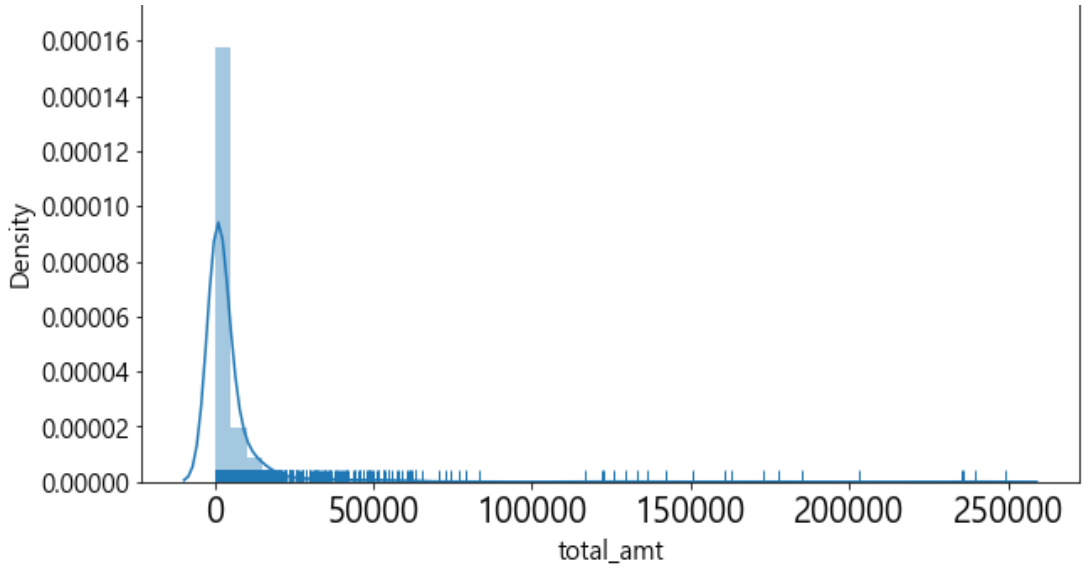


Sales Analysis (셀러 기준)

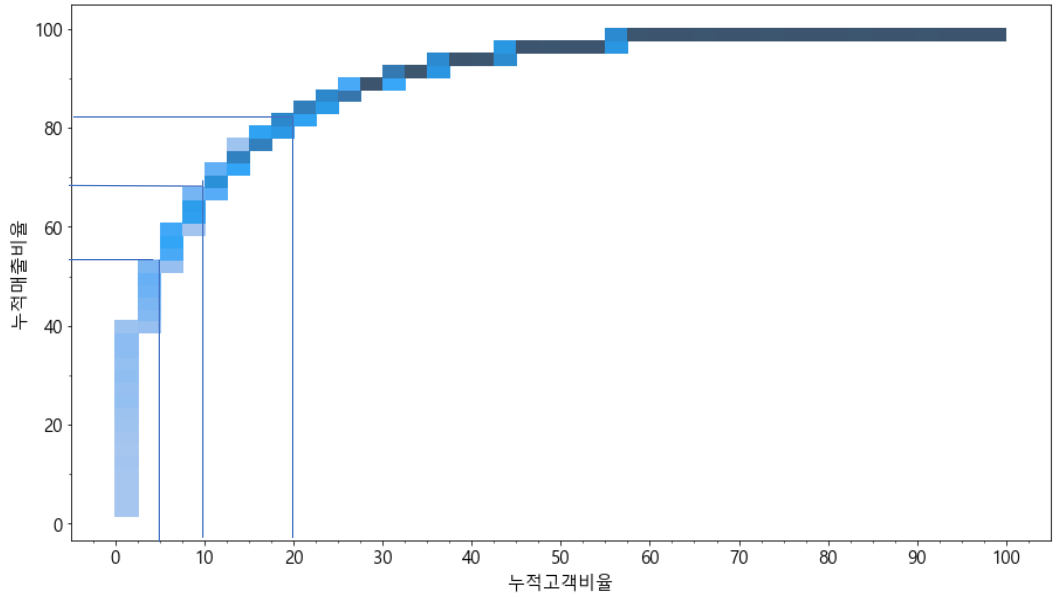
- **상위 셀러 매출 비중**
  - 상위 4%의 Seller가 전체 매출의 50% 차지
  - 상위 10%의 Seller가 전체 매출의 70% 차지
  - 상위 20%의 Seller가 전체 매출의 80% 차지

Seller_Sales	
count	3,029
mean	5,178
std	16,000
min	12
25%	281
50%	1,015
75%	4,060
max	249,393

셀러별 매출총액 분포

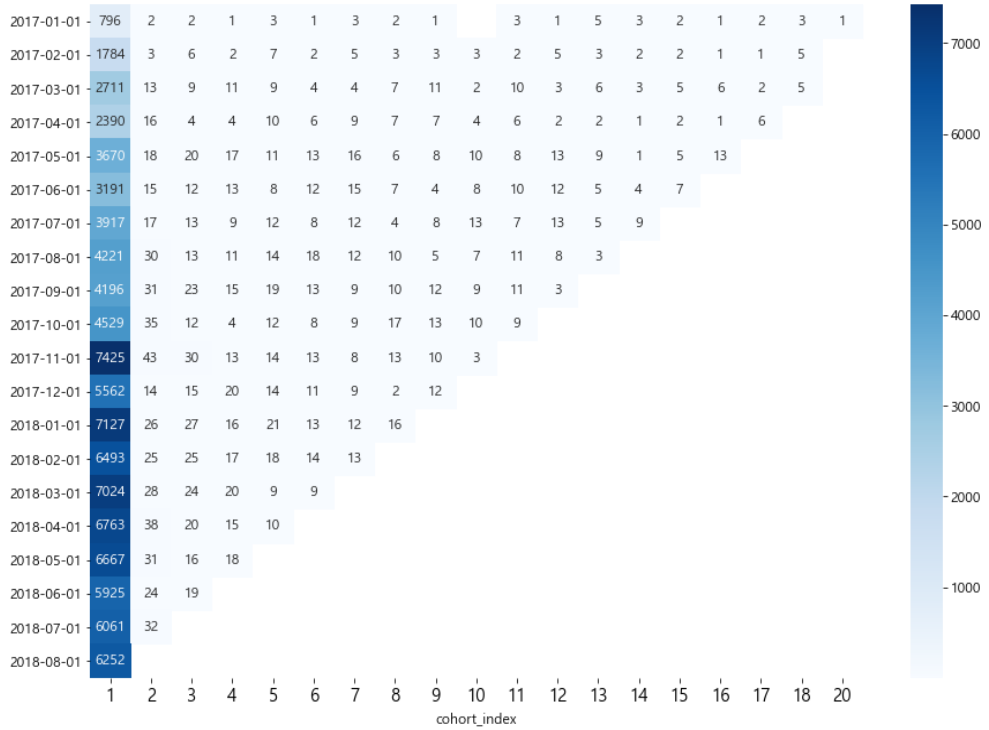


셀러 매출 Pareto 차트

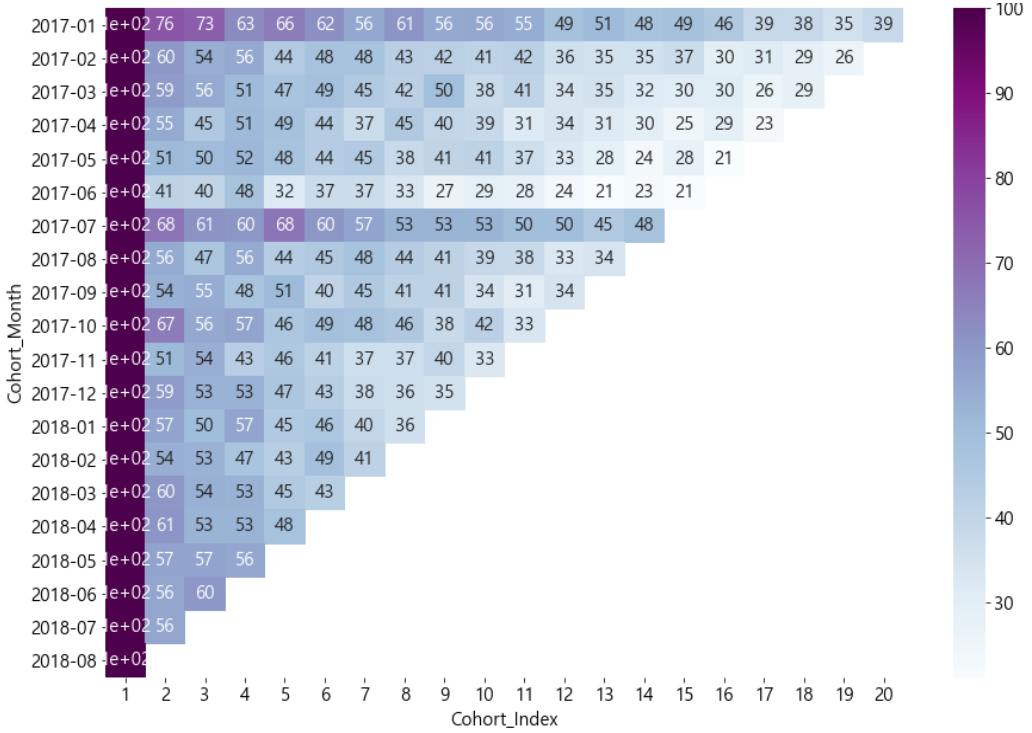


Cohort 분석

- **Customer_첫 구매 월 대비 재구매 건수**
  - 매월 신규 고객은 늘어나지만 재구매율 저조



- **Seller_첫 판매 월 대비 잔존율(%)**
  - 입점 후 셀러의 잔존율 약 30%대 수준



KEY FINDINGS FROM EDA

- Olist의 주 고객인 'Seller' 데이터로 분석 진행
  ('Customer' 데이터의 경우 1회만 구매한 계정이 대다수로, 분석에 customer 특성을 활용하기는 어려울 것으로 판단)
- Olist의 주 매출은 셀러에게서 발생하므로, 셀러의 판매 빈도를 높이고 우수 셀러를 육성하기 위한 방안 모색이 필요 할 것

분석질문 설정

SETTING AGENDA

04

셀러 판매 활성화를 위한 인사이트를 어떻게 찾을 수 있을까?

- 신규셀러 유입을 위한 Olist 광고홍보전략 ➔ 광고/유입경로 데이터 부재로 불가
- 판매율 높은 마켓플레이스 영업전략 ➔ 판매 채널 기록 부재로 불가
- 셀러별 매출 성과와 멤버십 등급 상관성 증명 통한 멤버십 업그레이드 유도 ➔ 멤버십 정보 부재로 불가

고객 리뷰 분석을 통해 **서비스 개선 요소**를 찾고
고객만족도 향상에 초점을 맞춰 플랫폼을 운영하면
고객 재방문이 늘고, **셀러 판매량**도 증가하지 않을까?

고객 만족도에 주요한 영향을 미치는 요인을 발견하고
서비스 개선을 위한 액션플랜을 도출하기 위해
고객 만족도를 예측하는 모델을 만들어보자!

Customer Satisfaction

피쳐 엔지니어링

FEATURE ENGINEERING

05

Review Comment

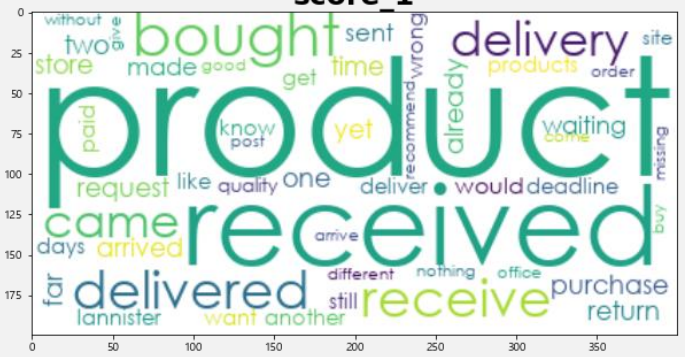
- 고객이 주문 건에 대해 만족/불만족 했던 주요 원인 파악을 위해 Review Comment 단어 빈도분석 진행 (포르투갈어 → 영어 번역 구글 API 이용)

- 전반적으로 '**Product**'와 '**Delivery'**(receive, came, arrived 등)에 대한 언급이 많음

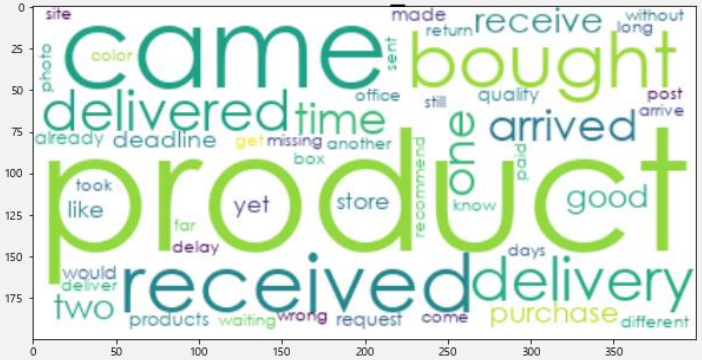
➔ **고객 만족도 예측을 위해 제품 및 배송 관련 피쳐를 중점적으로 확인**

불만족

score_1



score_2



score_3

만족

score_5

score_4

Feature Selection

- 각 피쳐별로 고객 만족도와의 관계를 시각화로 확인하여 예측 모델에 사용할 최종 변수 선정
- 리뷰가 없는 주문에 대한 만족도 예측에도 활용할 수 있는 변수 선정

변수	분류	내용	비고	변수 종류
review_score_binary	목표변수	review_score 1~3점은 '1', 4~5점은 '0'으로 분류	불만족: 1, 만족: 0	Cat
item_nb	제품 관련	주문제품 수량		Num
product_description_length		제품 설명 텍스트 길이		Num
product_photos_qty		제품 사진 개수		Num
total_payment		총 결제 금액	제품 가격 + 배송비 총액	Num
category		제품 카테고리	(76개->21개 재분류)	Cat
freight_value		배송비		Num
freight_value_rate		total_payment 중 freight_value 비율		Num
customer_state	배송 관련	배송 도착 지역		Cat
seller_state		배송 출발 지역		Cat
deivery_error		배송완료일이 예상 배송일을 지났는지 여부		Num
delivery_preparation		배송 준비 기간(배송접수일 – 주문 접수일)	셀러 주관	Num
delivery_periods		배송 기간(배송완료일 - 배송접수일)	택배사 주관	Num
delivery_delay_rolling		지연일수 (실제 도착일 – 예상 도착일) 평균	셀러별 과거 3개월 판매 내역을 집계	Num
delivery_preparation_rolling		준비일수 (배송접수일 – 주문접수일) 평균		Num
delivery_error_rolling		'delivery_error' 누적 횟수		Num
delivery_error_rate_rolling		'delivery_error' 비율		Num
cancel_rolling	주문 취소	셀러_주문 취소 건수 평균		Num
canceled_rate_rolling		셀러_주문 취소 건 비율		Num
comment_nb_cumsum	셀러 평점	셀러_리뷰 코멘트 글자 수 평균	해당 주문 이전 까지 누적 셀러 만족도 집계	Num
review_score_binary_cummean		셀러_Review_score_binary 평균		Num

Feature Eng. 변수

Product

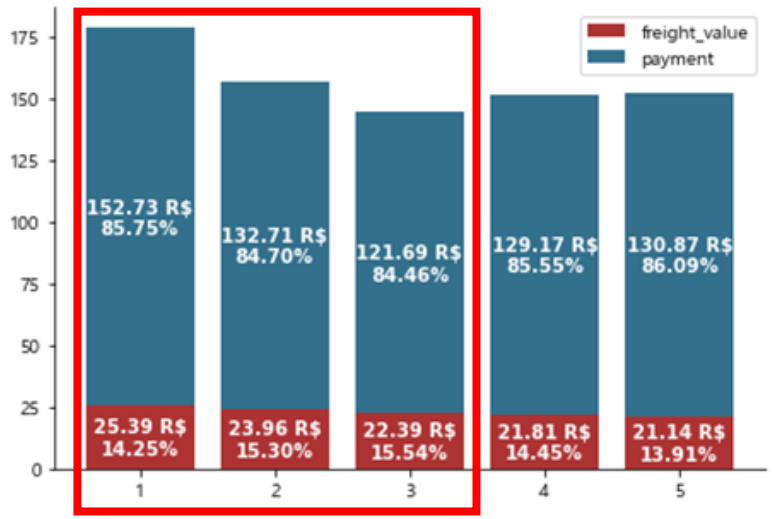
- **가격 및 주문제품 개수**
  - **구매금액이 높을수록 기대치가 높아지는 부분이 만족도에 영향을 미치지 않을까?**

  ➔ 주문 금액 및 주문 제품 개수와 만족도는 반비례하는 경향이 있는 것으로 확인됨

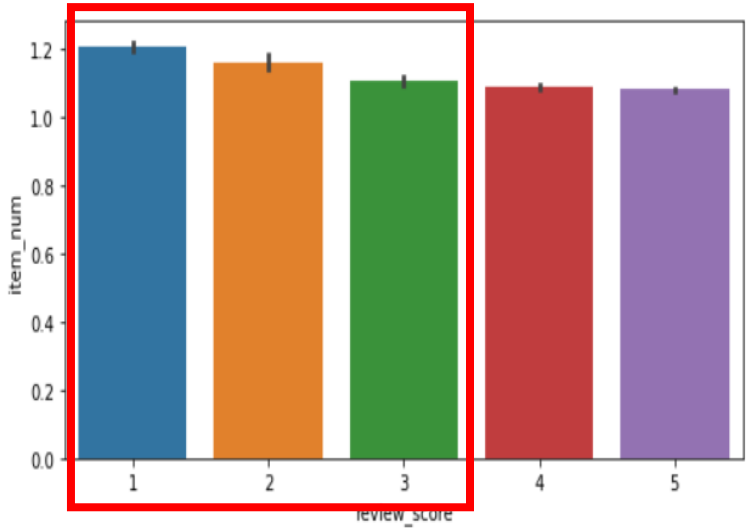
- **제품관련 정보**
  - **제품 정보(설명 문구 및 사진)가 자세할수록 판매에 정성을 쏟는 셀러일 것이고, 고객의 기대치와 실제 제품의 Gap이 작아 만족도가 높지 않을까?**

  ➔ 그래프상 유의미한 차이는 보이지 않지만 만족도 예측 시 유효한 피쳐가 될 수도 있을 것으로 판단

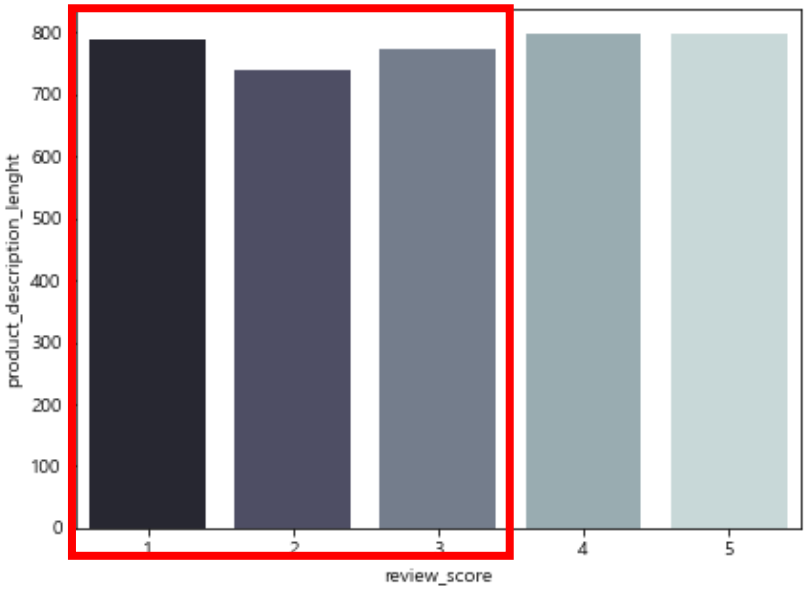
리뷰스코어별 제품 가격 및 배송비 비율



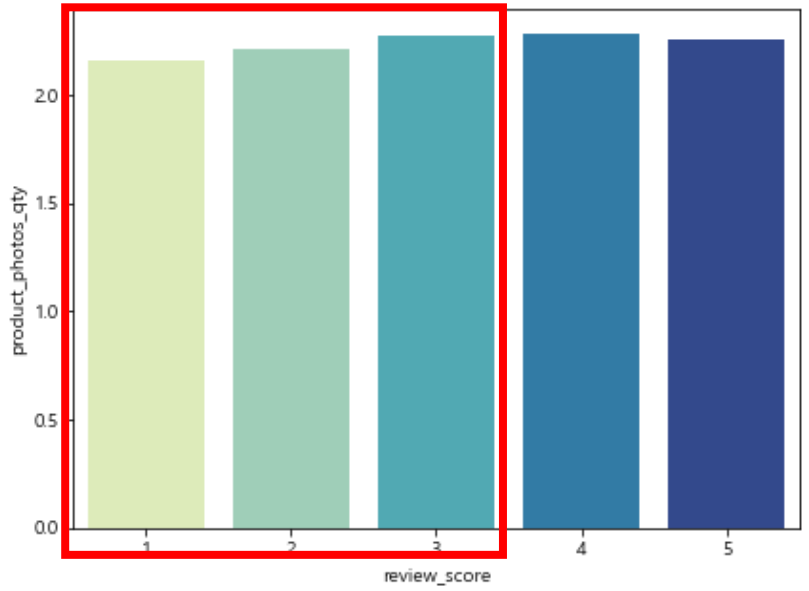
리뷰스코어별 주문 제품 개수



리뷰 스코어별 제품 설명 길이 평균



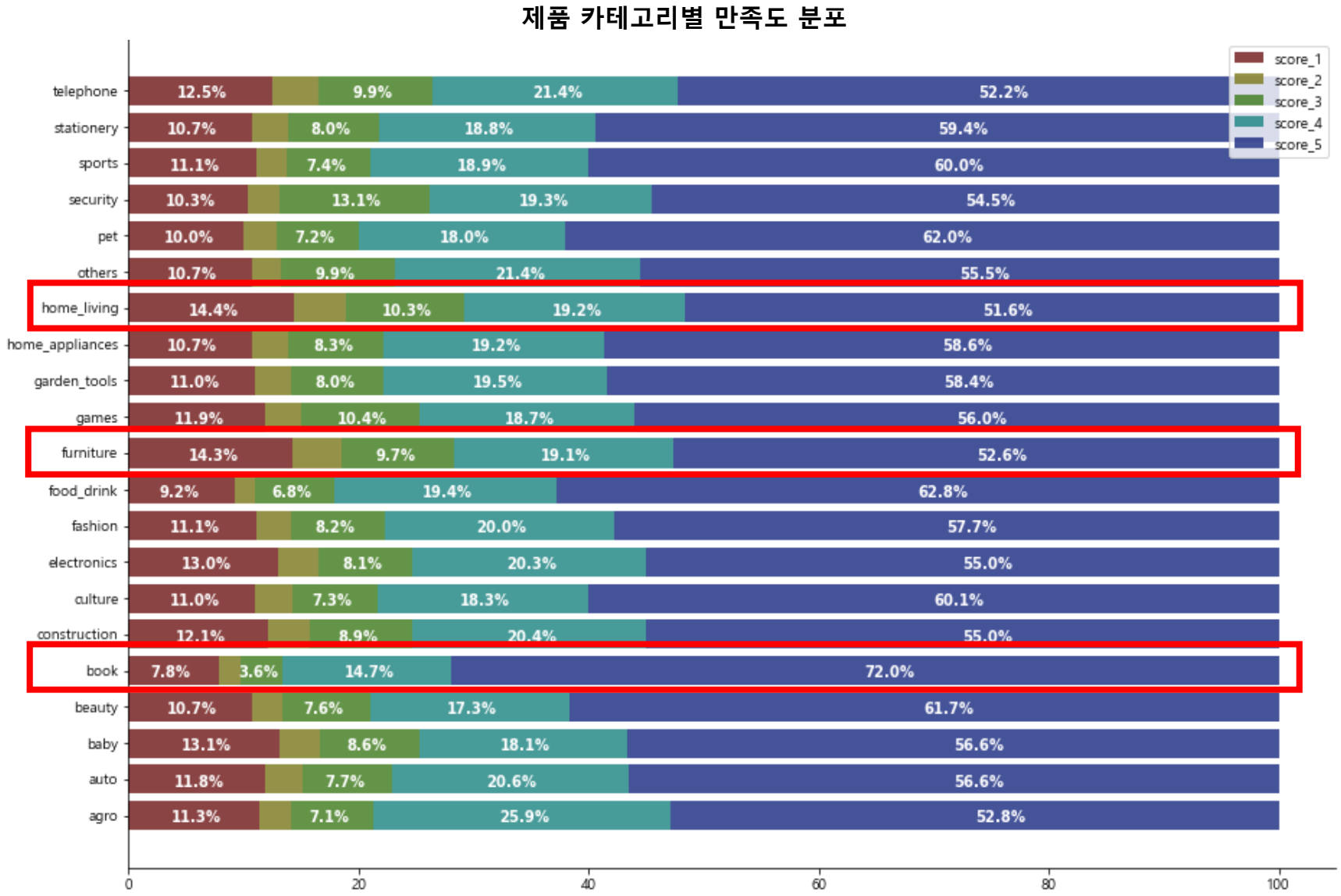
리뷰스코어별 제품 사진 개수 평균



Product

- **제품 카테고리**
  - **제품간 품질 편차가 큰 카테고리와 적은 카테고리의 만족도 분포는 다르지 않을까?**
    (ex. 비교적 균일한 퀄리티의 book vs 다양한 퀄리티의 제품이 존재할 수 있는 home_living, furniture)

➔ book 카테고리는 리뷰스코어 4~5점, home_living, furniture 카테고리는 1~3점 비율이 상대적으로 높은 것으로 나타남



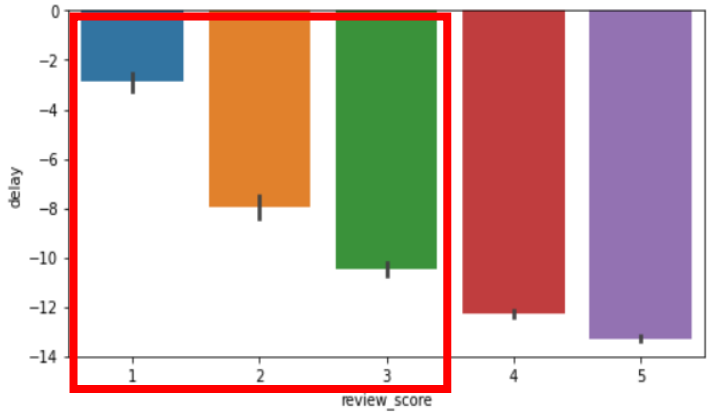
Delivery

- **배송 지연 및 소요일**
  - **배송기간이 길수록, 예상일 보다 늦어질 수록 만족도가 낮아지지 않을까?**

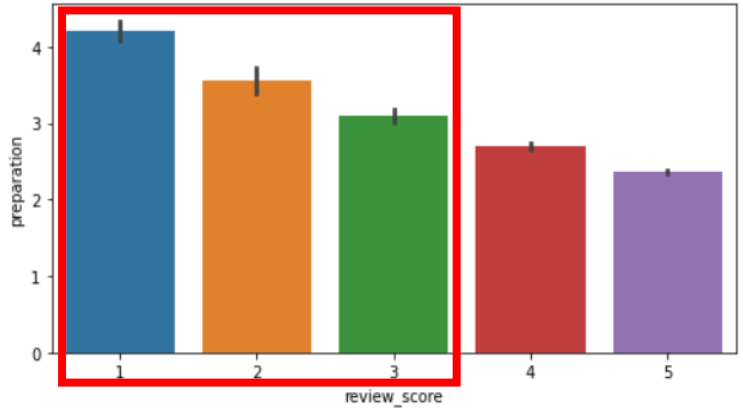
➔ 배송지연일수, 배송준비일수, 배송일수 모두 만족도와 반비례하는 경향 확인 (배송지연일수의 경우 평균적으로 예상일보다 빠르게 배송되어 음수로 나타남)

리뷰스코어별 배송지연일수, 배송준비일수, 배송소요일수

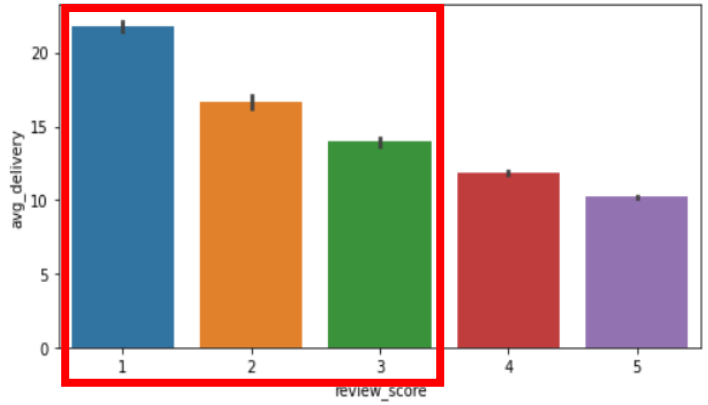
* **배송지연일수** : 실제 도착일 – 예상 도착일
(예정일 대비 며칠이나 빨리/늦게 도착했는지)



* **배송준비일수** : 배송 접수일 – 주문일
(주문 후 셀러가 제품을 배송접수하기까지 며칠 소요됐는지)



* **배송소요일수** : 실제 도착일 – 주문 접수일
(고객이 주문 후 제품을 받기까지 며칠 소요됐는지)



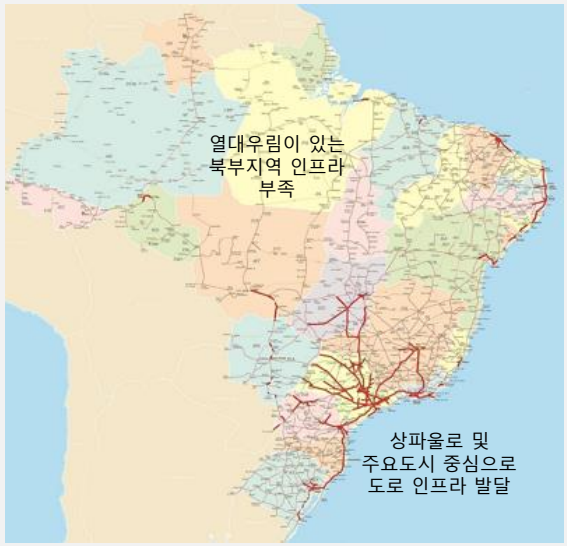
<이슈>
- 리뷰는 제품 수령시점 혹은 예상 배송일이 지난 시점부터 남길 수 있어, **고객이 리뷰를 남긴 시점에 제품을 받지 못한 상태인 경우 배송지연일수, 배송준비일수, 배송소요일수가 만족도에 정확히 반영되지 않음**
- 위 피쳐를 고객 만족 예측에 직접 사용하는 것은 부적절할 것으로 판단

<대안>
- **해당 주문 건 셀러에 대한 과거 3개월 배송지연일수, 배송준비일수 평균을 계산하여 피쳐로 활용. 셀러별로 서비스 속도 경향성이 있을 것으로 추측**
- '배송 소요일' 자체는 물류 인프라 및 특수상황에 큰 영향을 받을 수밖에 없는 피쳐로, 각 셀러의 특성으로 활용하기에는 부적절할 것으로 판단하여 미사용

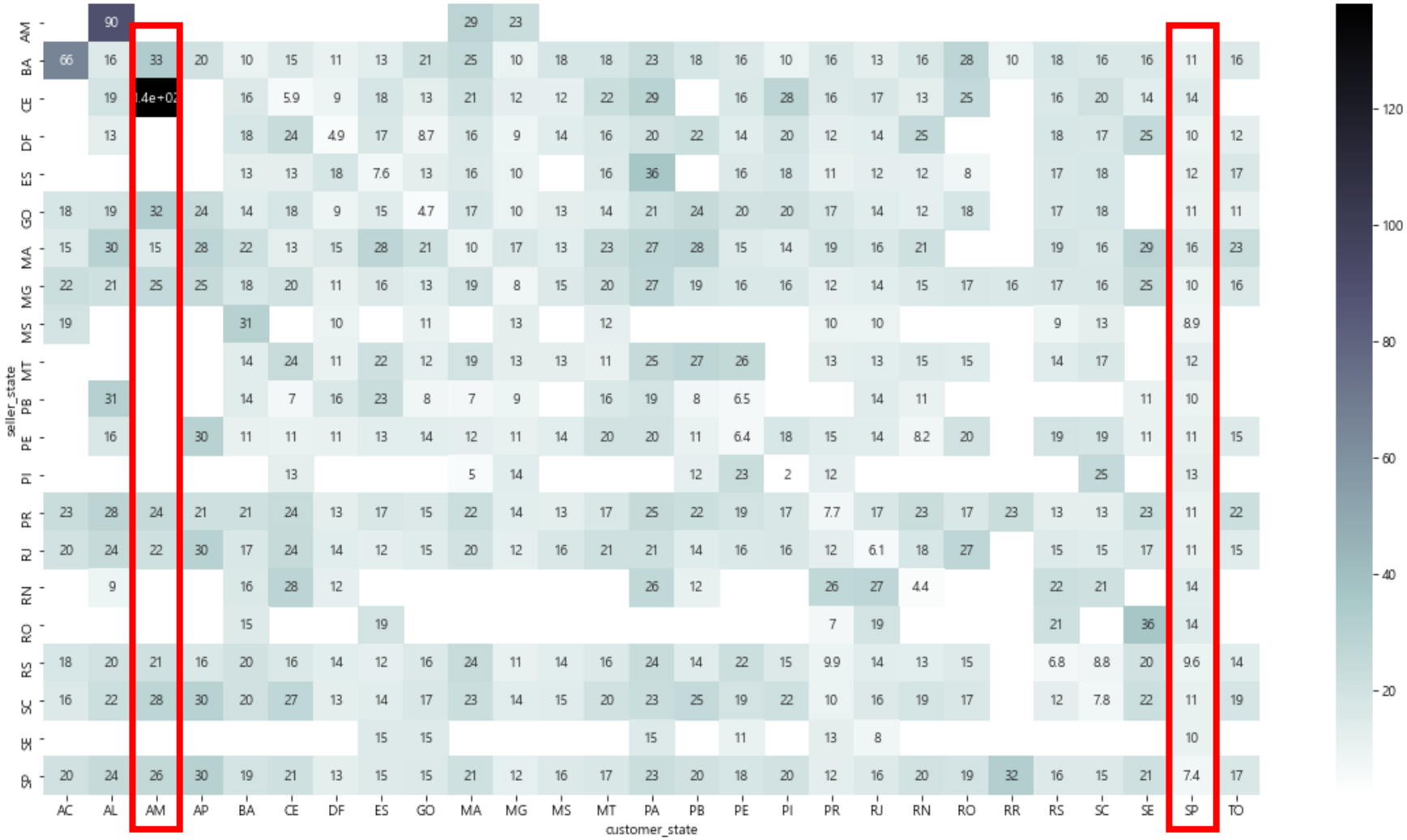
Delivery

- **셀러 및 주문자 거주지**
- **도로가 발달하지 않은 지역은 상대적으로 배송이 지연되는 경우가 많지 않을까?**

➔ 브라질에는 안정적인 물류 인프라를 갖추지 못한 state가 많으며, 이에 따른 배송 기간 증가는 고객 만족도에 영향을 미칠 것으로 가정



셀러 및 주문자 거주지별 배송일 평균



Review Score

- **셀러 만족도 평균**
- **만족도 평균이 낮은 셀러에게서 구입한 경우 불만족할 확률이 높아지지 않을까?**

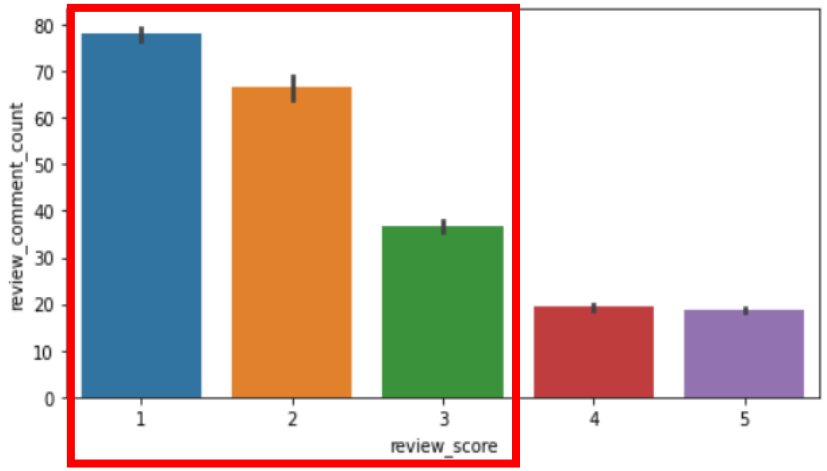
- **코멘트 길이**
- **고객이 주문에 대해 불만족한 경우 코멘트를 더 길게 남기지 않을까?**

➔ 만족도가 낮았던 주문은 리뷰 코멘트 글자 수가 많은 경향

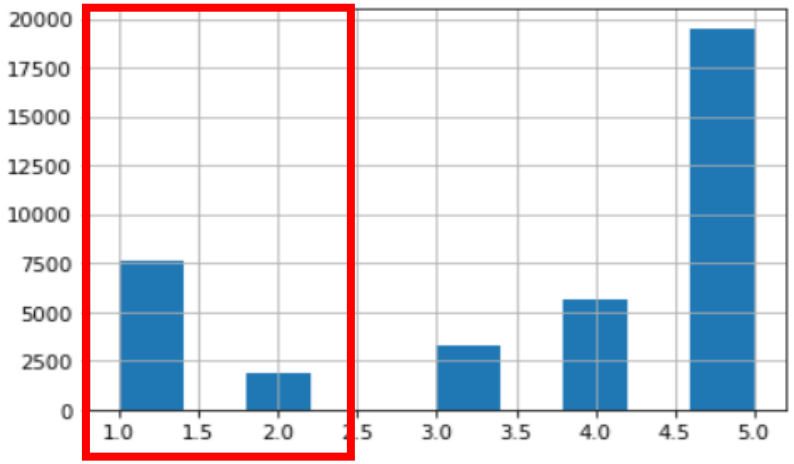
➔ 코멘트를 남긴 경우와 남기지 않은 경우 비교시 남긴 경우에 불만족 건이 더 많음을 확인

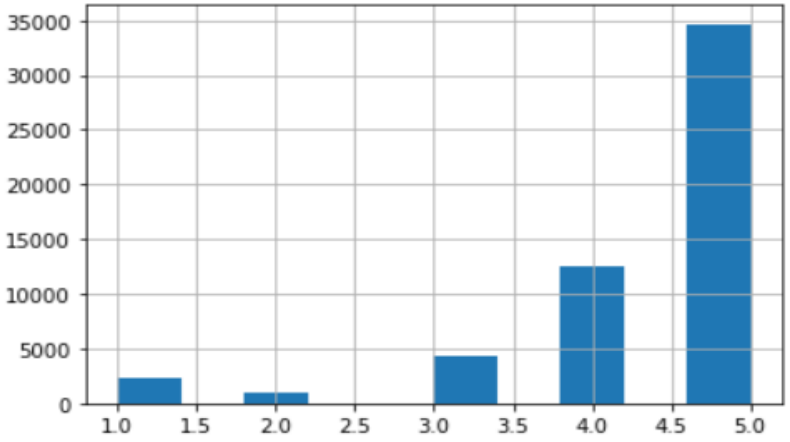
➔ 해당 주문 이전 시점까지 해당 셀러가 받은 리뷰 스코어와 코멘트길이의 평균을 피쳐로 선정

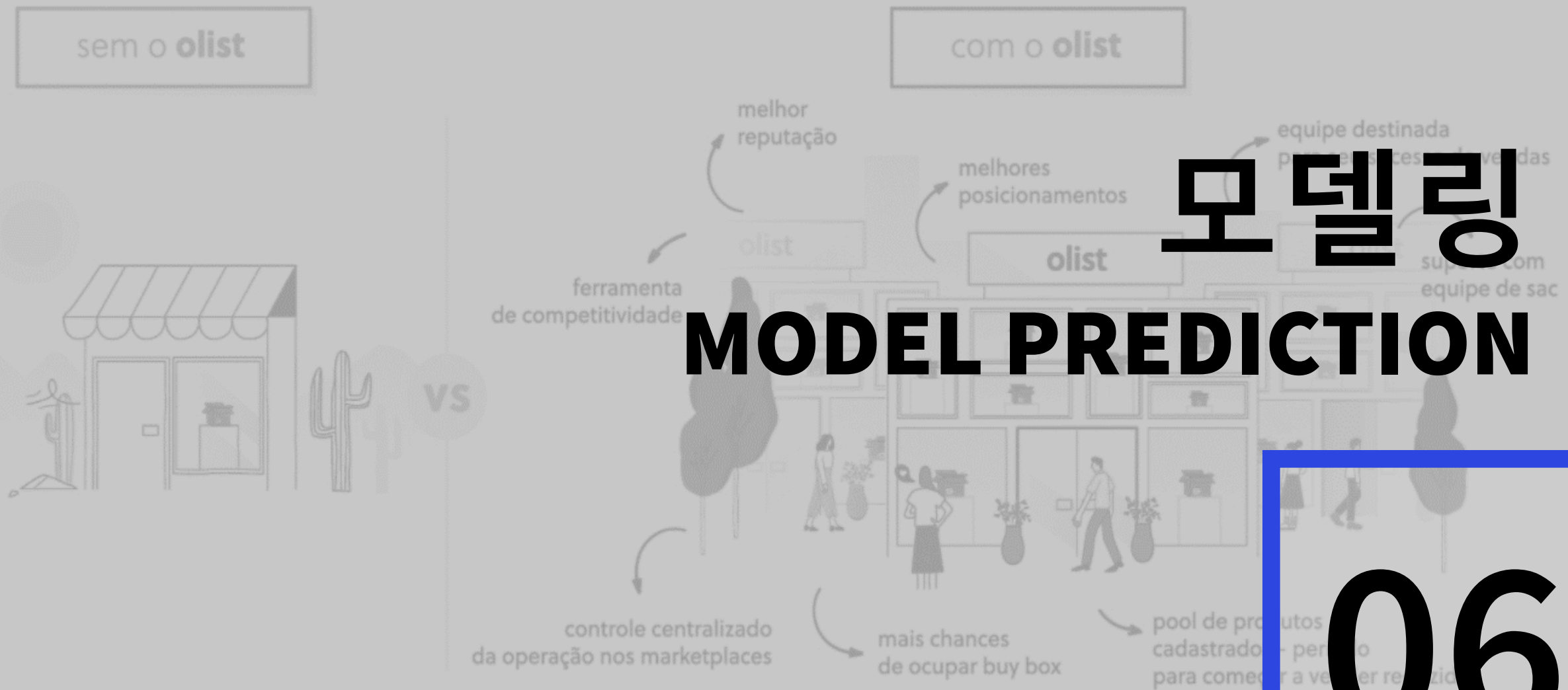
리뷰 스코어별 코멘트 길이



코멘트를 남겼을 때 만족도 분표 (좌) vs 코멘트를 남기지 않았을 때(우) 만족도 분포 비교







모델링

MODEL PREDICTION

06

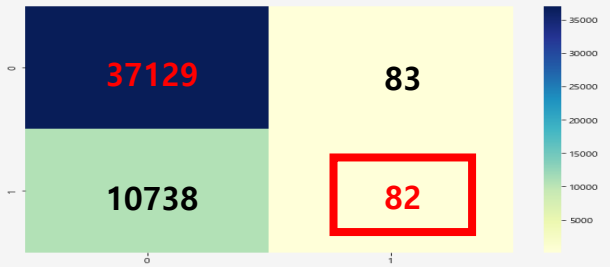
모델 적합 (LightGBM)

- 결측치가 허용되는 분류모델 XGB, LightGBM 중 더 우수한 성능을 보인 **LightGBM** 최종 사용
- **실제 불만족한 고객(리뷰 스코어 1~3점)을 모델이 얼마나 분류해냈는지**가 중요한 문제이므로 **recall**을 평가 metric으로 설정하고, 피쳐 엔지니어링, 모델 튜닝을 통해 개선

Base model (기본 변수만 사용)

	precision	recall	f1-score	support
0	0.78	1.00	0.87	37212
1	0.50	0.01	0.01	10820

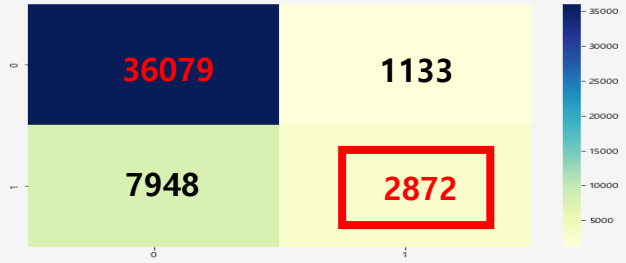
변수
item_nb
product_description_length
product_photos_qty
total_payment
category
customer_state
seller_state



(Feature Eng. 변수 추가)

	precision	recall	f1-score	support
0	0.82	0.97	0.89	37212
1	0.72	0.27	0.39	10820

Feature Eng. 변수	
delivery_error_rolling	freight_value
delivery_error_rate_rolling	freight_value_rate
delivery_delay_rolling	deivery_error
canceled_rate_rolling	delivery_preparation
comment_nb_cumsum	delivery_periods
review_score_binary_cummean	cancel_rolling
delivery_preparation_rolling	

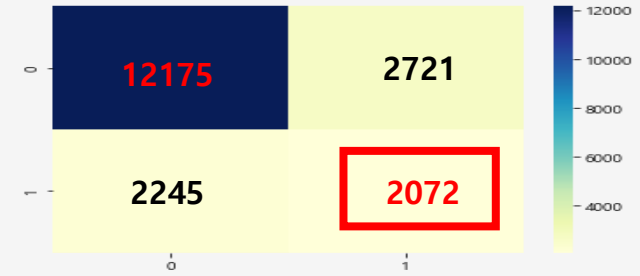


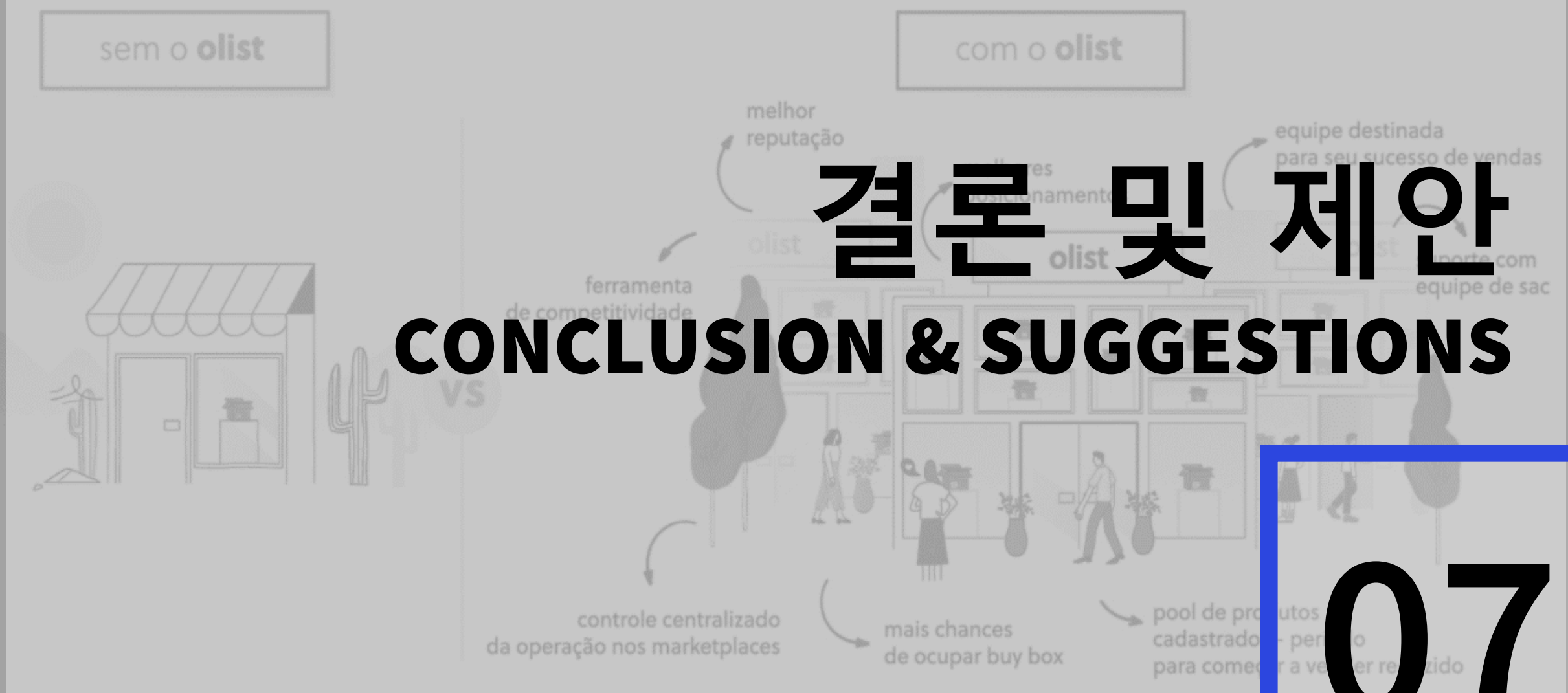
Tuned Model (모델 튜닝 후)

	precision	recall	f1-score	support
0	0.84	0.82	0.83	14896
1	0.43	0.48	0.45	4317

1. **Test set 비율 :** 0.2
2. **OverSampling**
   - 0.25 / 0.75 ➔ 0.5 / 0.5
3. **Permutation Importance** 기준 컬럼 선택
4. **하이퍼 파라미터 조정**
   - learning rate➔ 0.185
   - n_estimators➔ 35
   - max_depth➔ 5
   - boosting_types➔ gbdt
   - num_leaves➔ 30

4. Threshold 조정 ➔ 0.5





결론 및 제안

CONCLUSION & SUGGESTIONS

07

결론 및 제안사항

1) Logistics

- **물류 관련 변수와 고객 만족도 관계**
  - 고객이 제품을 주문한 후 받기까지 걸린 전체 시간이 길어질수록 고객 만족도는 감소하는 경향
  - 물류 인프라가 낙후한 지역에서 발생한 주문 건에 대한 만족도가 낮은 경향
  - 배송비 금액이 높을수록 만족도가 낮은 경향

➔ **제안사항 : 물류 파트너 및 인프라 확보**

- 기존 우체국 외에, 새로운 물류 파트너 및 지역별 물류 거점을 확보하여 배송 프로세스를 개선하고 배송기간을 단축할 것을 제안
- Olist는 실제로 물류에 대한 투자를 확대중이며, 작년 하반기 물류사 PAX를 인수하였음
- 기존 5개 물류허브에 더해 올해 30개를 추가로 오픈하여 다양한 도시에서의 배송 커버할 예정

BUSINESS

SoftBank-backed Olist acquires technology and logistics startup PAX

It is the startup's second M&A in less than a month; Olist wants to expand cross-docking activities by six times

With the acquisition, Olist plans to close 2021 with 30 new hubs in the main cities of the country and become the pioneer in collecting daily orders. Ferraz explained that with a logistics and convenience service Olist can deliver in one to three days, when before the product would take seven to 10 days deadline.

기사 출처 : https://labsnews.com/en/news/business/brazil-based-olist-buys-technology-and-logistics-startup-pax

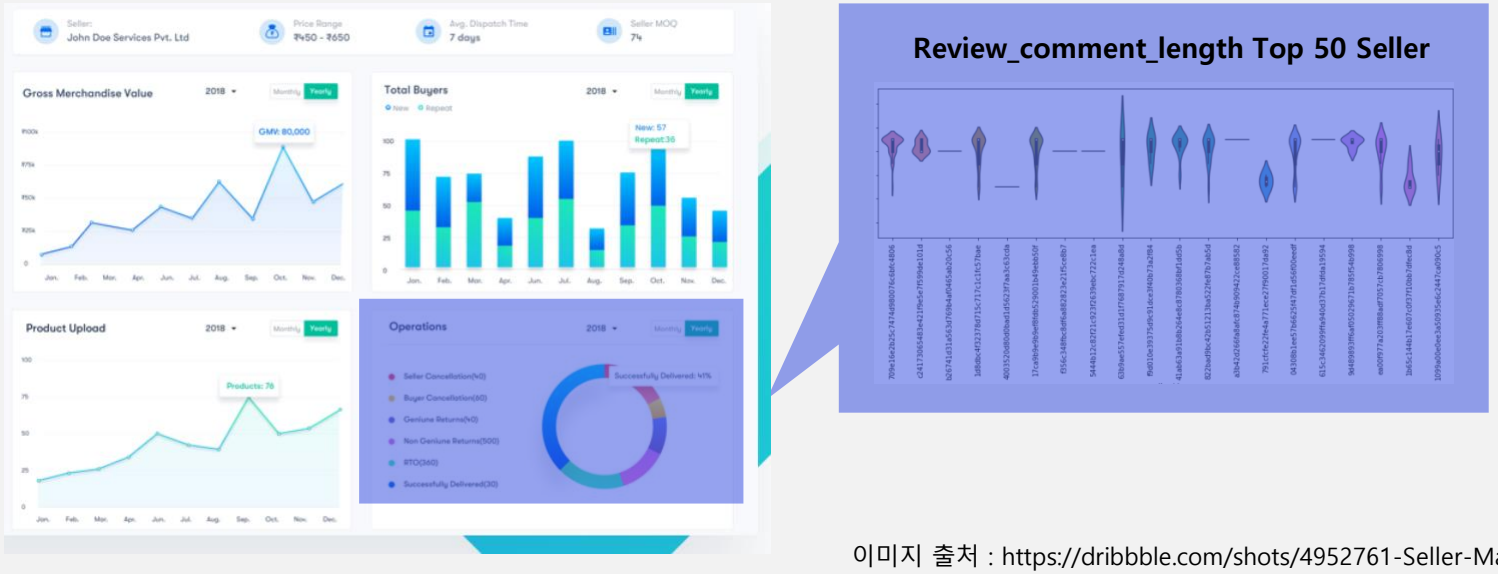
결론 및 제안사항

2) Seller Care

- **셀러 서비스 수준과 고객 만족도 관계**
  - 해당 셀러가 그동안 판매했던 제품의 준비기간 평균이 길수록 다음 고객의 만족도 감소하는 경향
  - 해당 셀러가 그동안 받았던 리뷰 스코어가 낮을수록 다음 고객의 만족도 감소하는 경향

➔ **제안사항 : 셀러 서비스 지표 관리**

- 셀러별로 만족도 평점, 리뷰 코멘트 글자수 등의 고객 만족도지표와 제품 배송 준비 기간 등 서비스 상태를 알 수 있는 지표를 추적하고 **서비스 개선이 필요한 셀러를 탐지 및 관리하여 불만족 재발생을 예방**



이미지 출처 : https://dribbble.com/shots/4952761-Seller-Management-Dashboard-for-Admin-Use

분석 한계점

1) 불충분한 판매 데이터

- 주어진 데이터셋에서 **판매 건수가 1건인 셀러**는 전체 3,095명 중 563명으로 **18.2%**를 차지
- 당초 셀러를 클러스터링 하여 그룹을 나누고, 고객만족도 예측에 해당 분류를 사용하고자 하였으나
  각 셀러의 특징을 구분하기 위한 데이터가 충분하지 못해 클러스터링을 활용하지 못하고 개별 입력변수로 대체
  (판매 건수가 적은 셀러는 직전 3개월의 매출, 판매건수, 배송준비일수 등에 결측치 발생)

2) 리뷰 데이터 구분

- 1건의 주문에 2종류 이상의 제품, 2명 이상의 셀러가 존재하는 경우 **정확히 어떤 셀러, 어떤 제품에 대한 만족도인지 구별할 수 없어 분석에서 제외하고 진행**
- 제품과 판매자 단위의 리뷰 스코어 기록이 가능하다면 데이터 손실 없이 분석을 진행할 수 있을 것으로 보임

3) 마케팅 활용 방안

- 주어진 데이터셋은 주문내역, 결제내역, 리뷰에 대한 정보로, **마케팅 제안을 하기위해 필요한 정보 부족**
  (셀러 유입경로, 멤버십 등급, 판매채널 등)
- 마케팅 데이터와 연계한다면 보다 풍부한 사업적 분석 및 수익성 제고를 위한 제안이 가능할 것으로 보임

모델 개선 방향

1) 새로운 모델 생성 및 앙상블

- 금번 분석에서는 **결측치를 남겨둔 경우와 대체한 경우(NA ➔ 99999) 모두 실험**을 진행하고자 **XGB, LightGBM** 모델 사용
- 다른 Tree-based 모델을 생성하여 앙상블 진행시 추가로 성능을 향상시킬 수 있을 것으로 보임

2) Feature Engineering

- Olist에서 제공된 데이터셋 종류가 다양해 분석기간 중 지속적으로 유의미한 변수를 발견할 수 있었으나,
  추가적으로 리뷰 스코어와 상관관계가 있는 새로운 변수를 찾을 경우 모델 성능 개선 가능

3) Hyperparameter Tuning

- 상기 과정을 수행 후 파라미터를 최적화할 경우 추가적인 성능 개선이 가능할 것으로 보임

OBRIGADO!