

E-COMMERCE

경영전략제시

FINAL_PJT 1팀

#고지현 #김정빈 #박건우 #윤정옥 #정기중 #황지우



Table of Contents

#1 주제

- E-COMMERCE
- 매출 예측 & 고객 세분화
- E-COMMERCE 산업의 성장

#4 매출예측 모델

- 데이터 전처리
- 모델링
- 평가

#2 데이터 소개

- Brazilian E-Commerce Public Dataset by Olist

#3 EDA

- olist관련 8개 데이터셋 분석

#5 RFM&고객세분화

- 데이터 전처리
- 모델링
- 평가

#6 가치제공

- 경영전략제시(매출, 고객 타겟팅)

E-COMMERCE

매출 예측 & 고객 세분화

dataset에 포함된 컬럼들을 이용해
매출예측, 고객 세분화



매출전략제시
고객 타겟팅



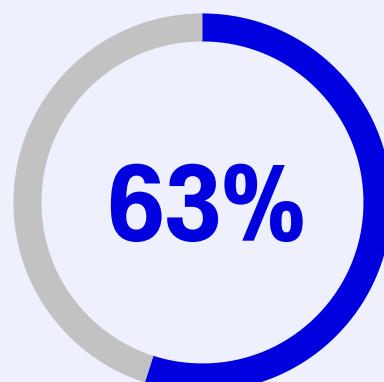
E-COMMERCE 성장

2025년 시장 규모

3조 4,533억

향후 5년간 250개 이상의 신약 출시

전자상거래 보급률

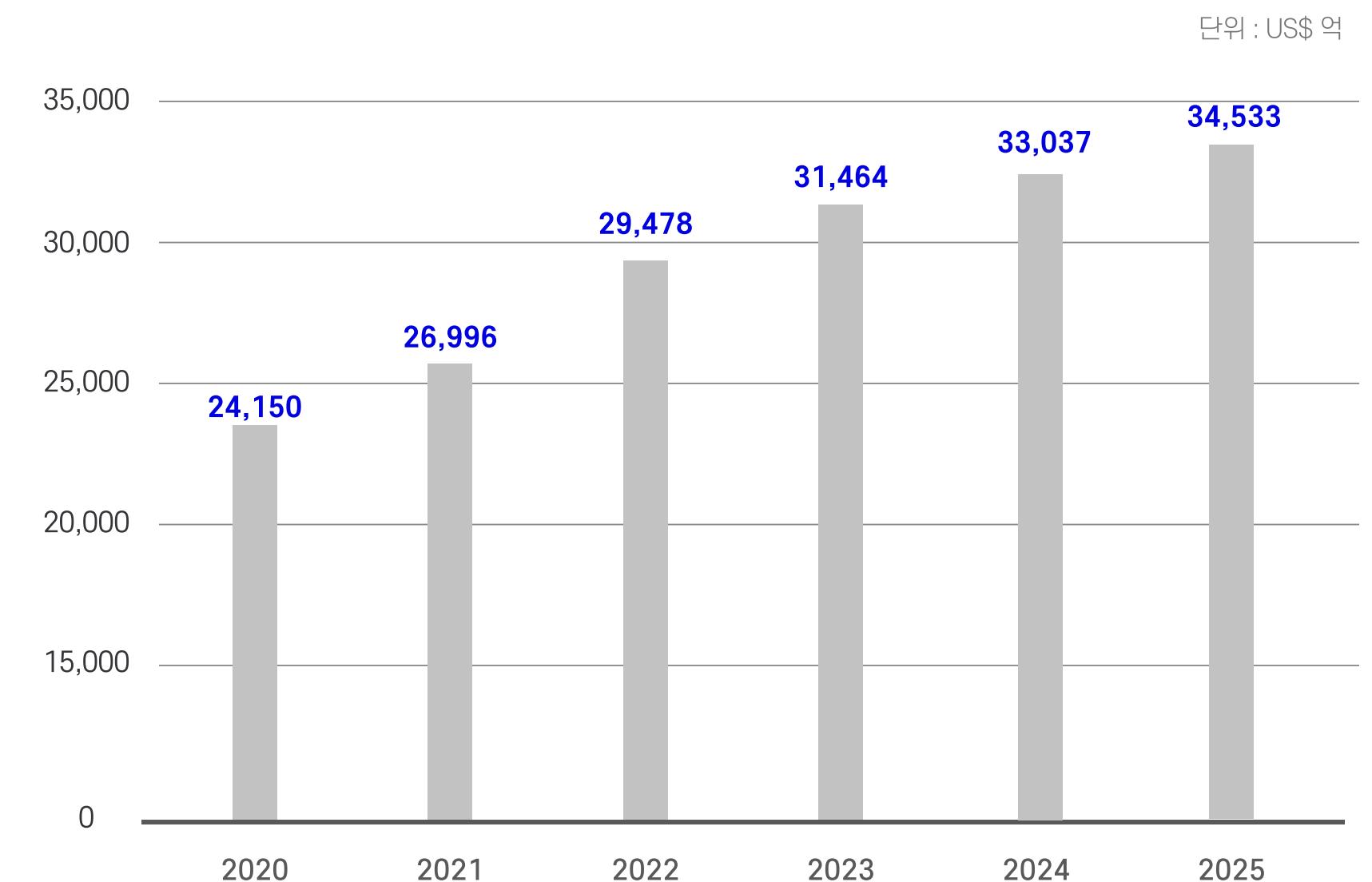


2025년 63.1%의 전자상거래 보급률 예상

연계산업 성장

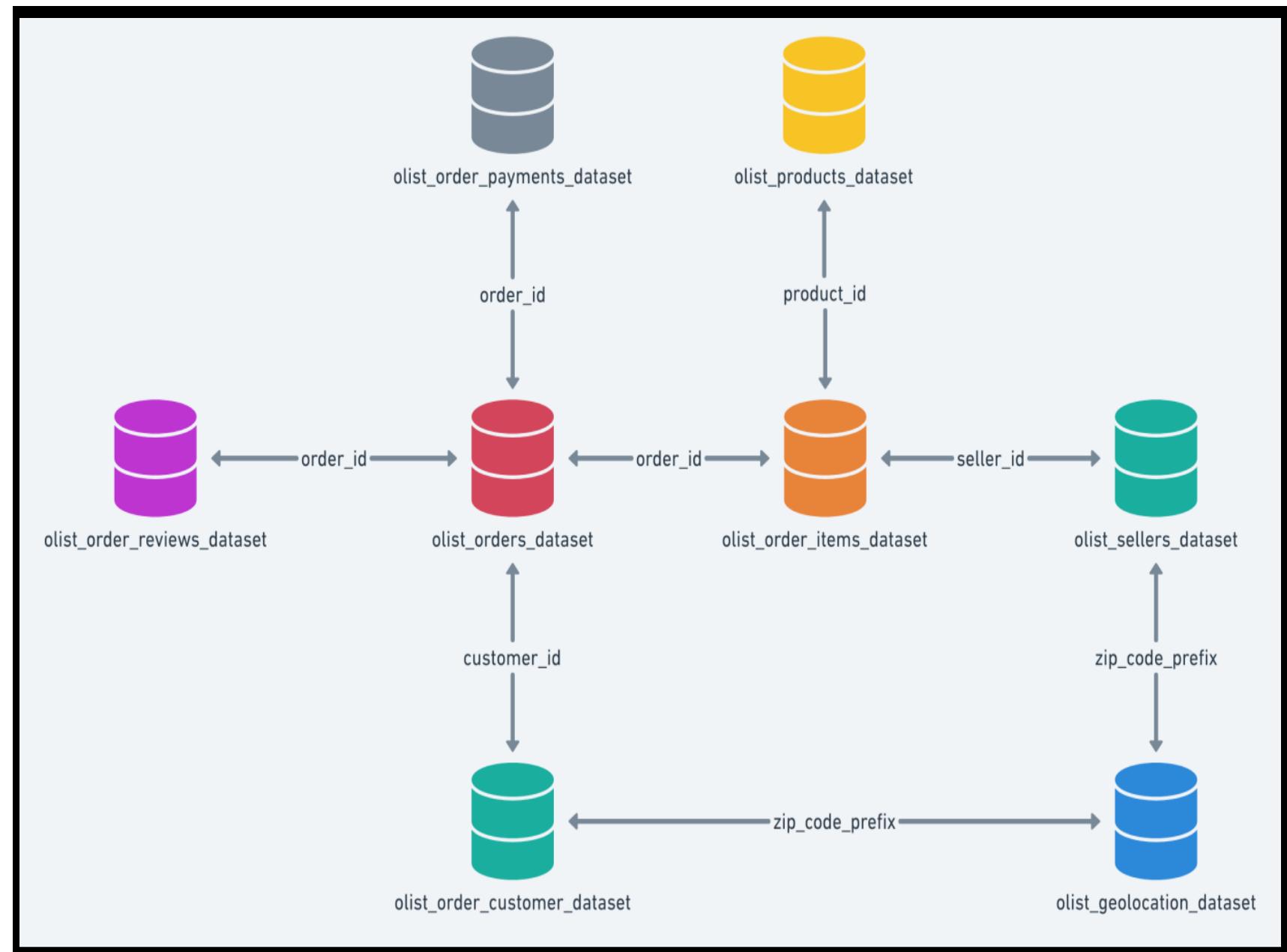
빅데이터

홍보를 위한 온라인 광고 시장, 인공지능
및 빅데이터 시장 성장 예상



출처 : 스탠더드

세계 전자상거래 매출



브라질 E-COMMERCE 사 OLIST 데이터

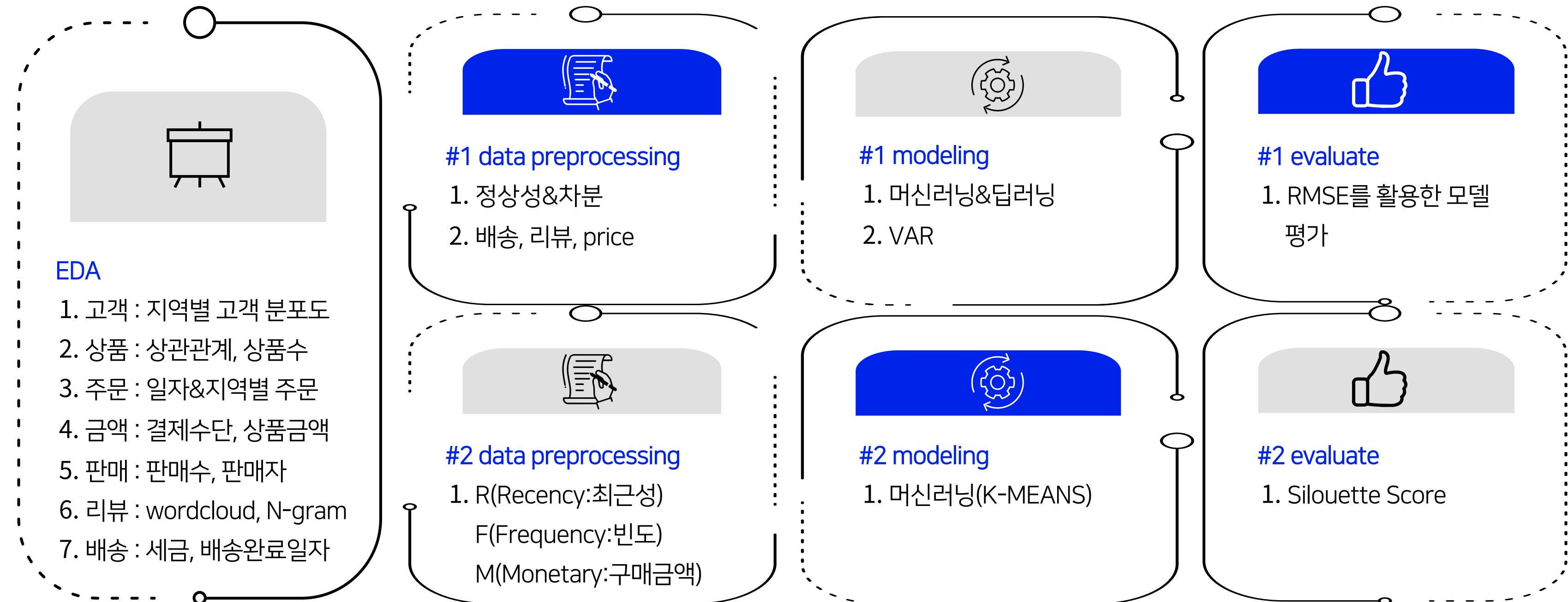
- 2016년~2018년 사이의 약 10만건의 주문 건

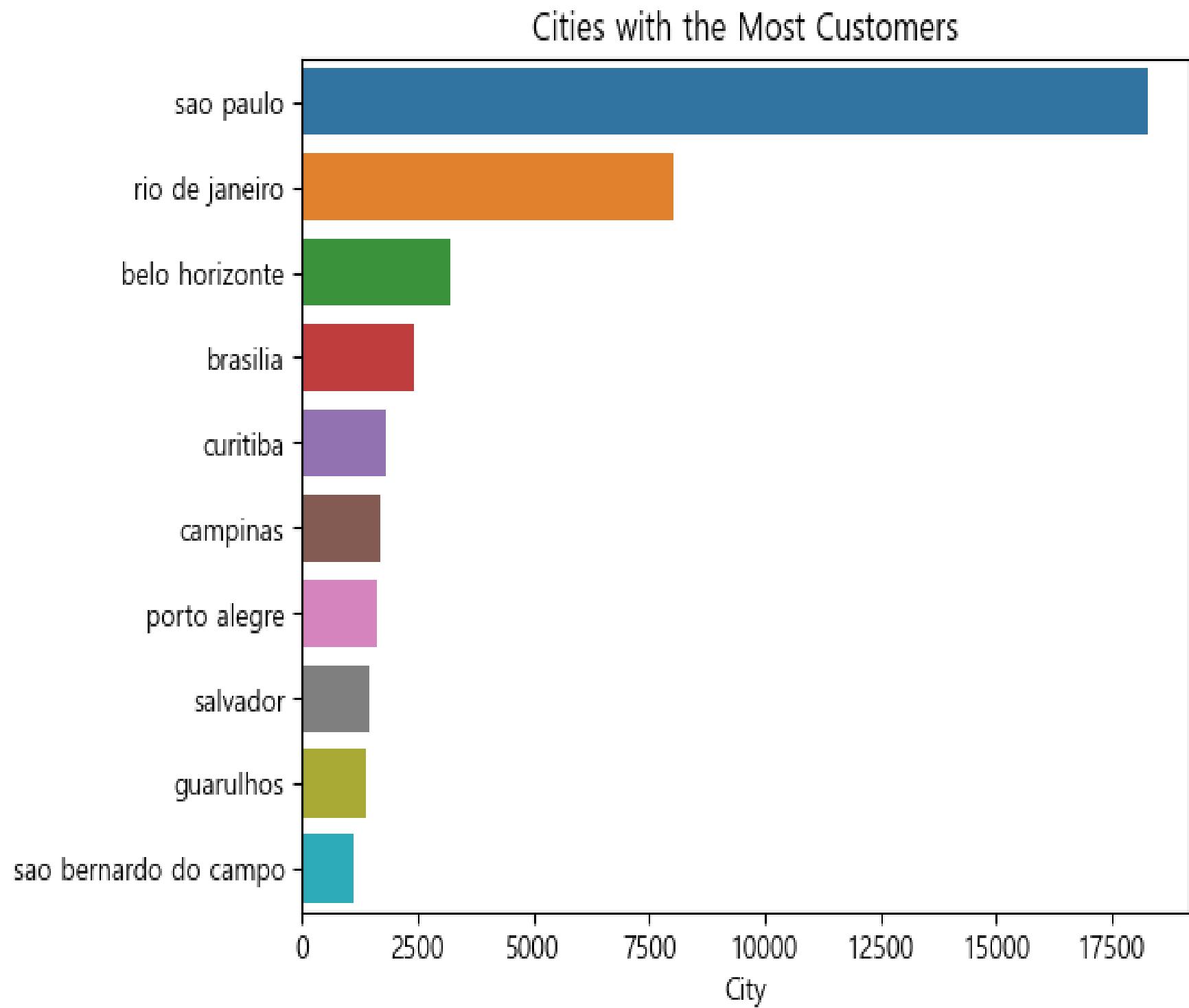
총 8개의 데이터셋 활용

- 고객, 지역, 주문, 배송, 결제, 리뷰, 상품, 판매로 구성

다양한 관점에서 분석

- 주문 상태, 가격, 배송, 위치, 리뷰 등 다양한 관점에서 데이터 분석 가능

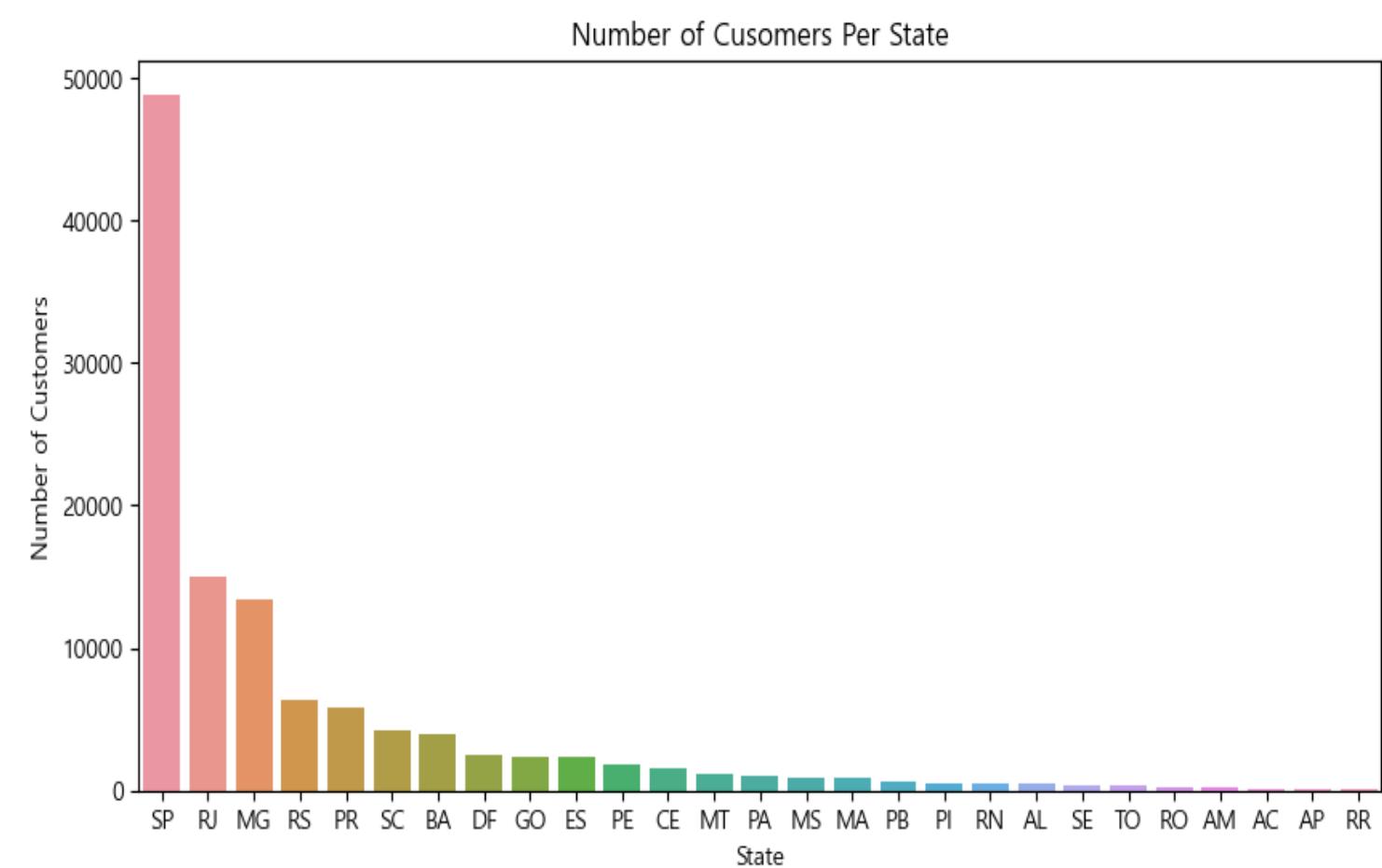




지역별 고객분포도



대부분의 고객들은 Sao paulo, rio de janeiro에 분포되고 있다.
그 중 특히 Sao paulo에 압도적으로 많은 고객층이 형성되어 있다.

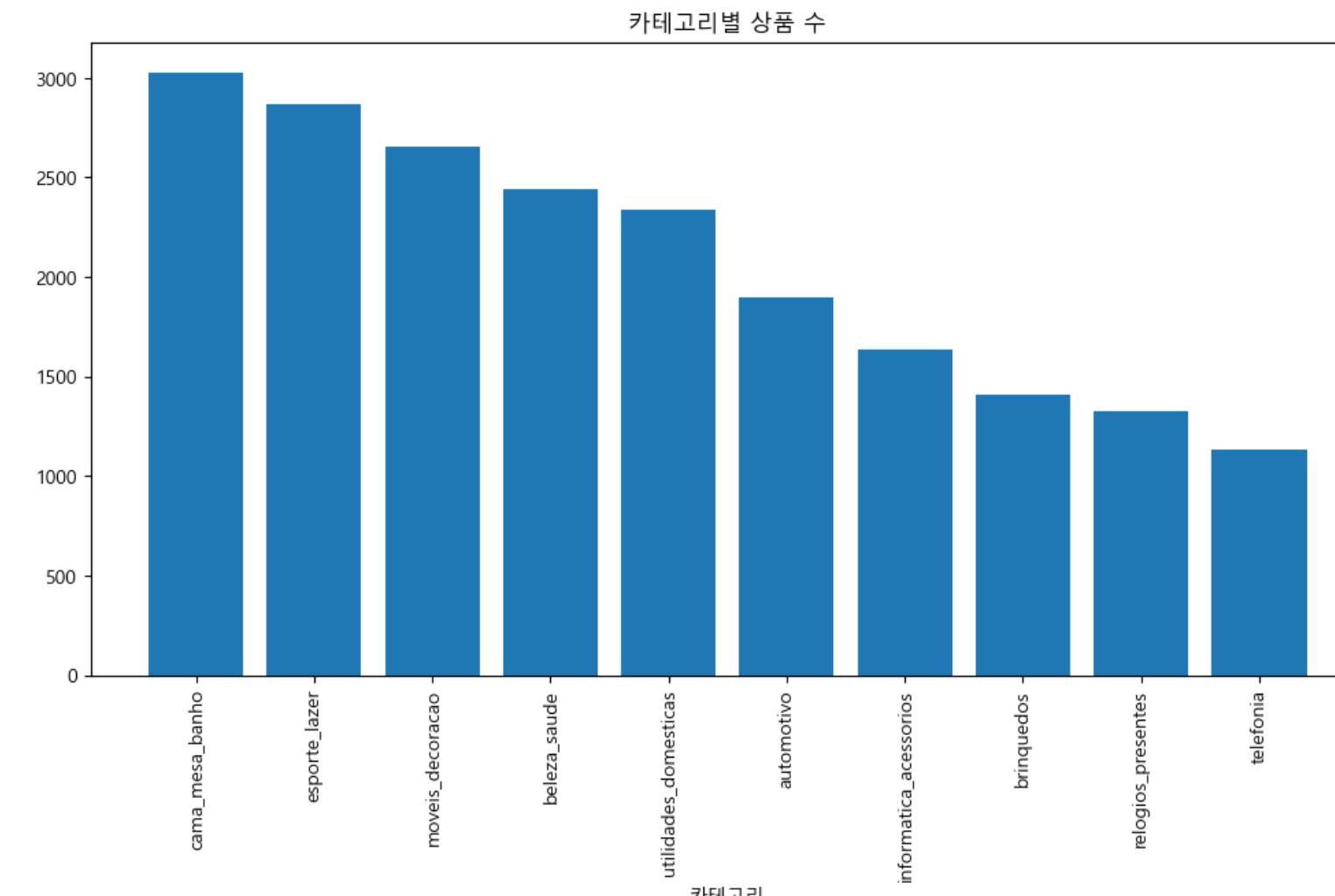
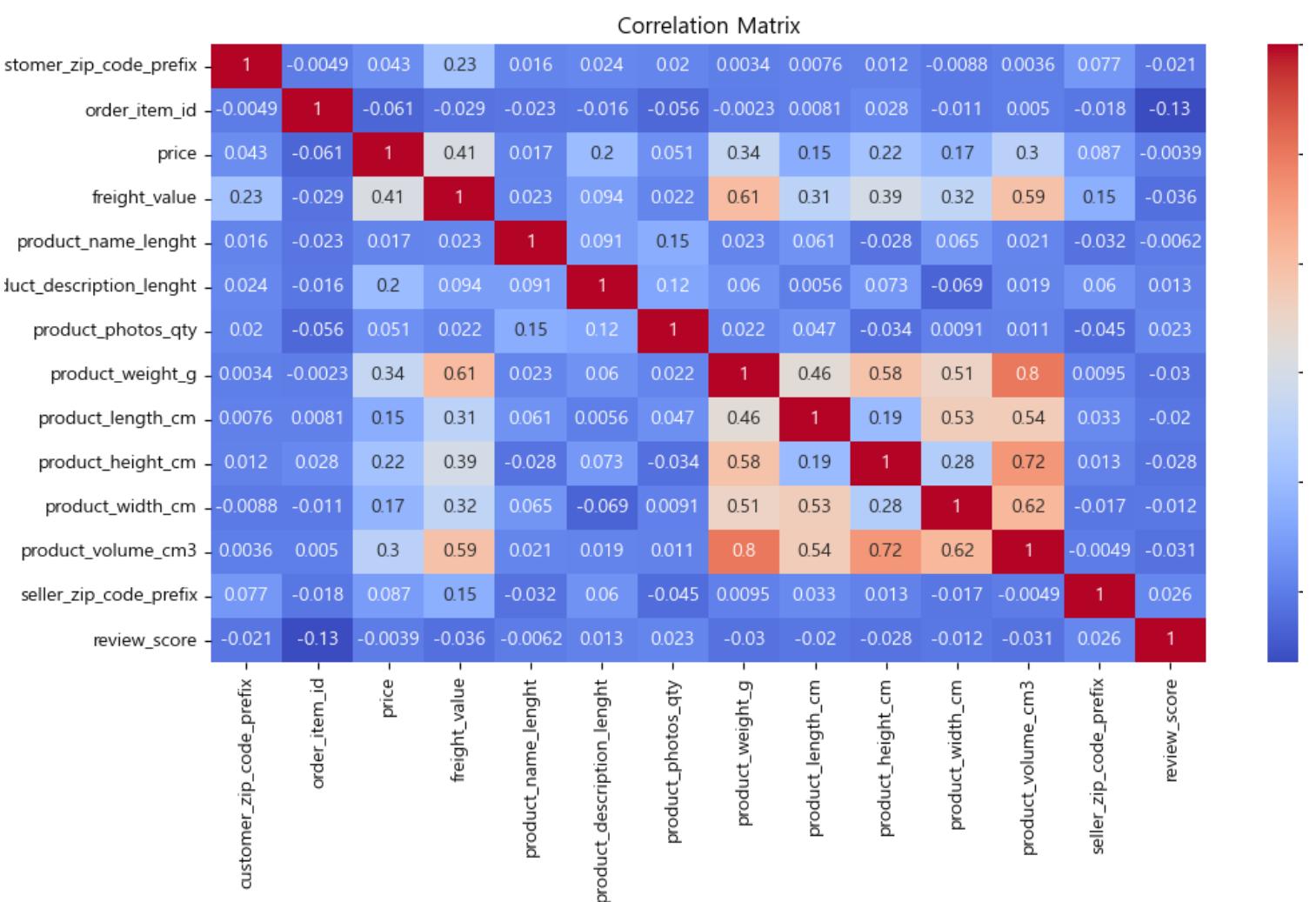


FEATURE

상관관계

무게, 길이, 높이처럼 부피와 관련된 feature들은 높은 상관관계를 가진다.

배송비는 부피, 무게와 높은 상관관계를 가진다.



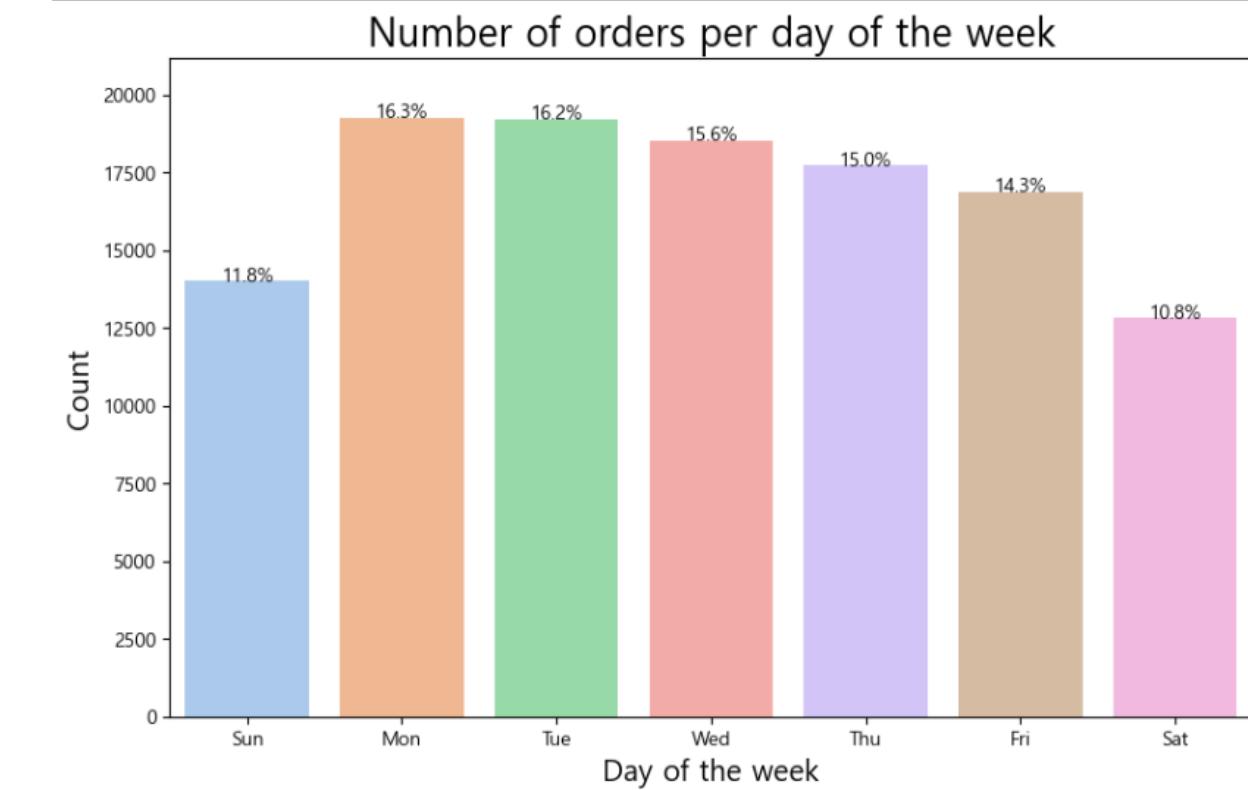
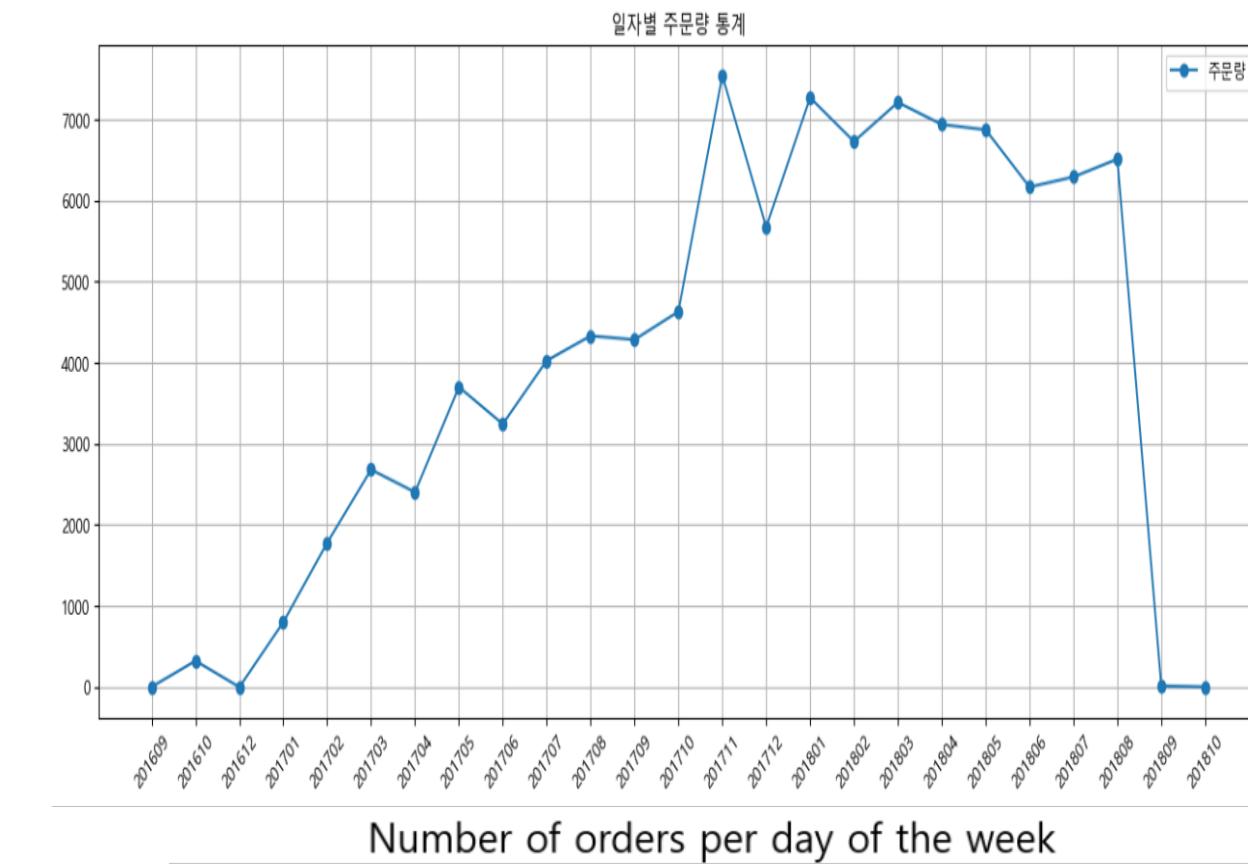
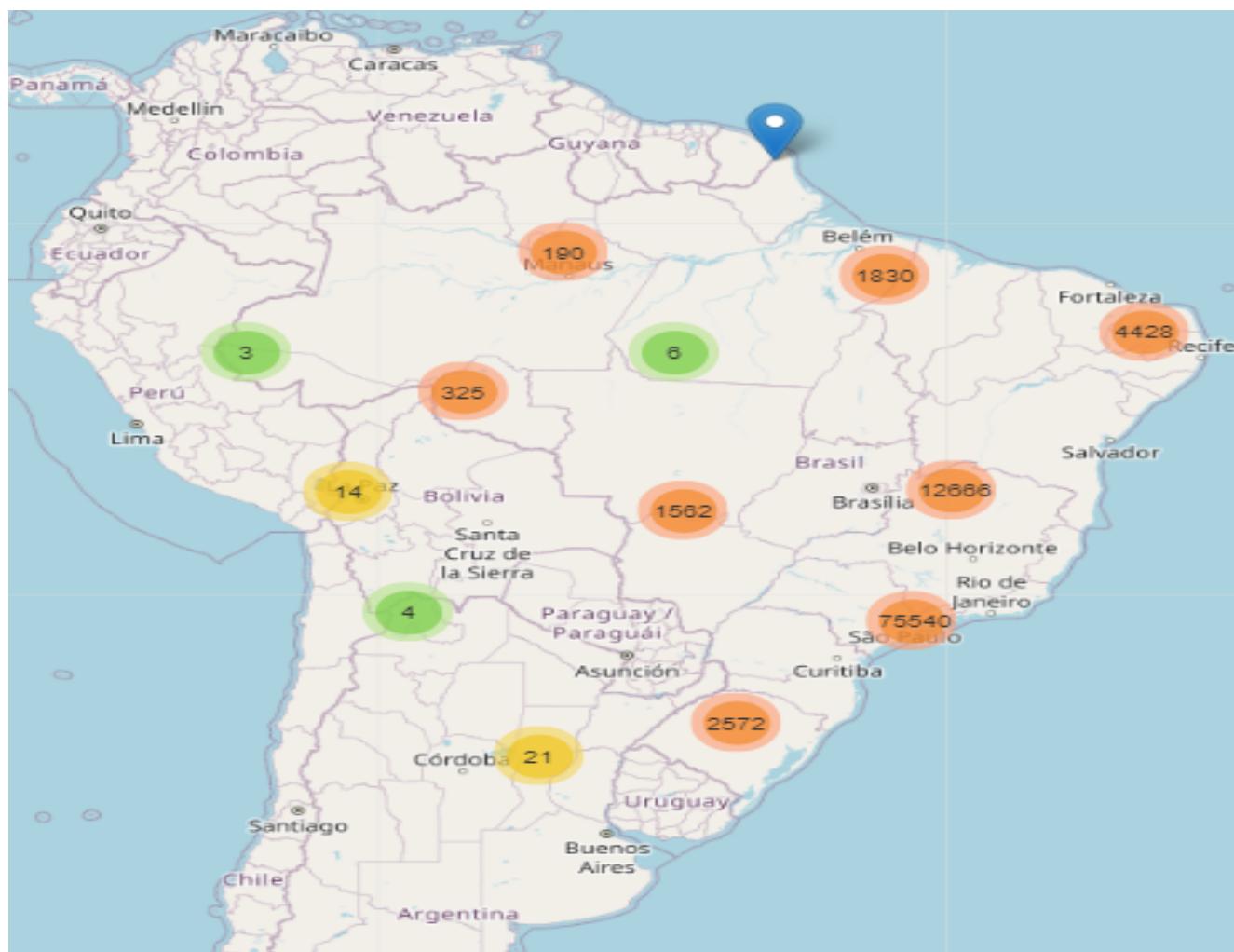
카테고리별 상품수 TOP10

cama_mesa_banho(침대테이블목욕), esporte_lazer(스포츠레저)
순으로 많은 상품을 보유했다.

일자&지역에 따른 주문

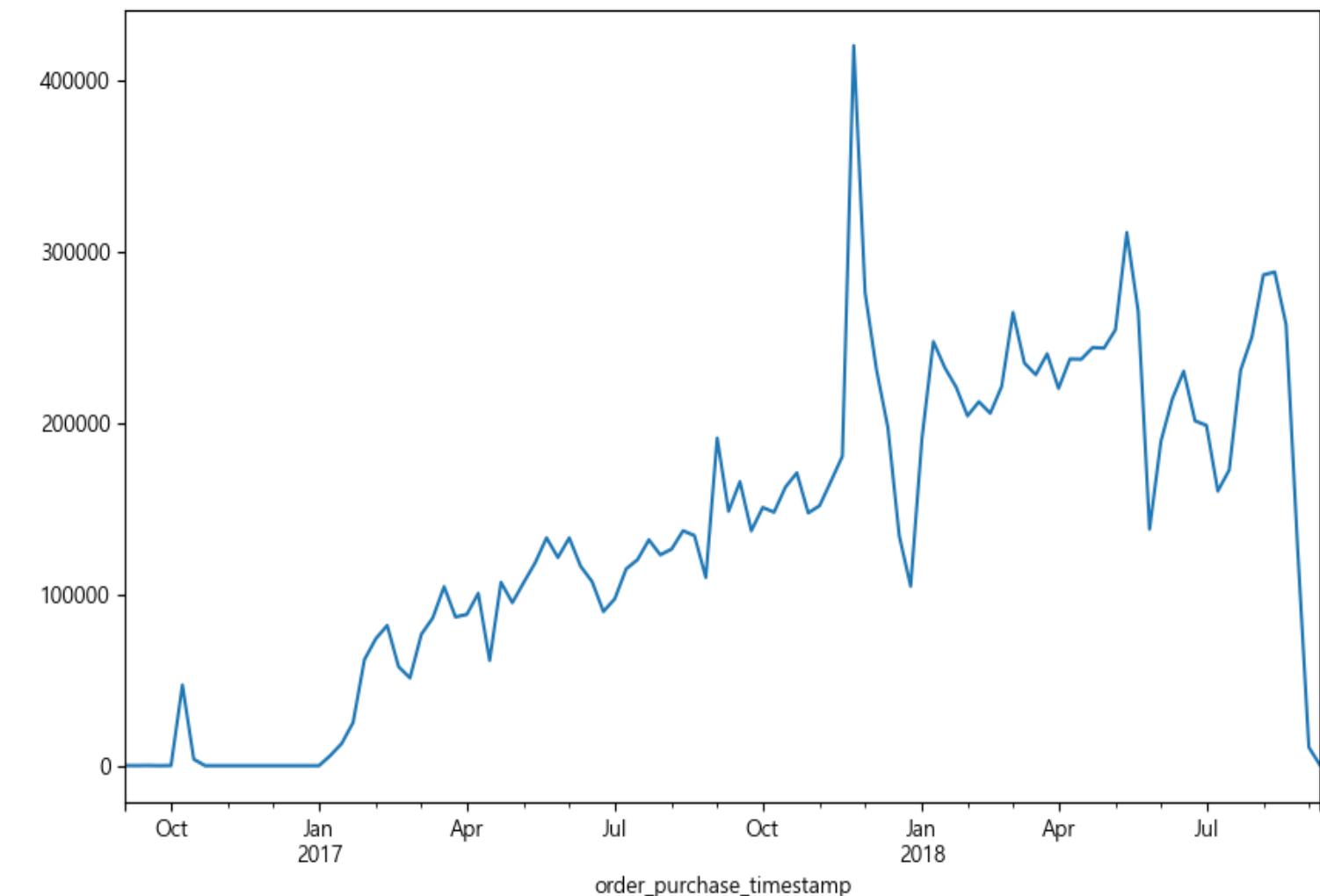
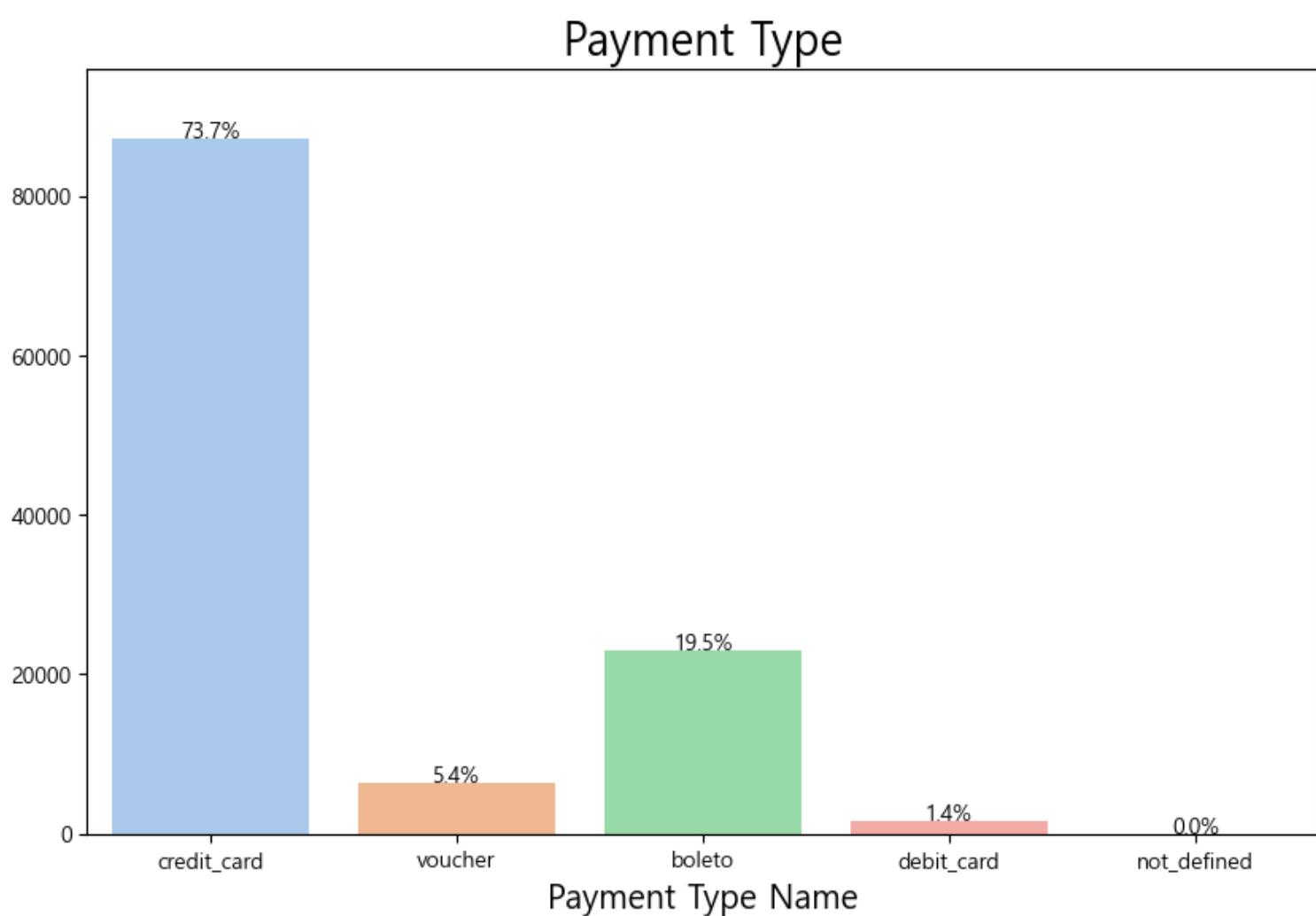
전체적인 주문량은 남동지역에 고르게 분포되어있고 상파울루가 상대적으로 높은 주문량을 보여준다.

2017년11월 ~ 2018년8월에 주문이 가장 많고, 이후 급락한다.
주말보다 평일에 더 많은 주문을 했다.



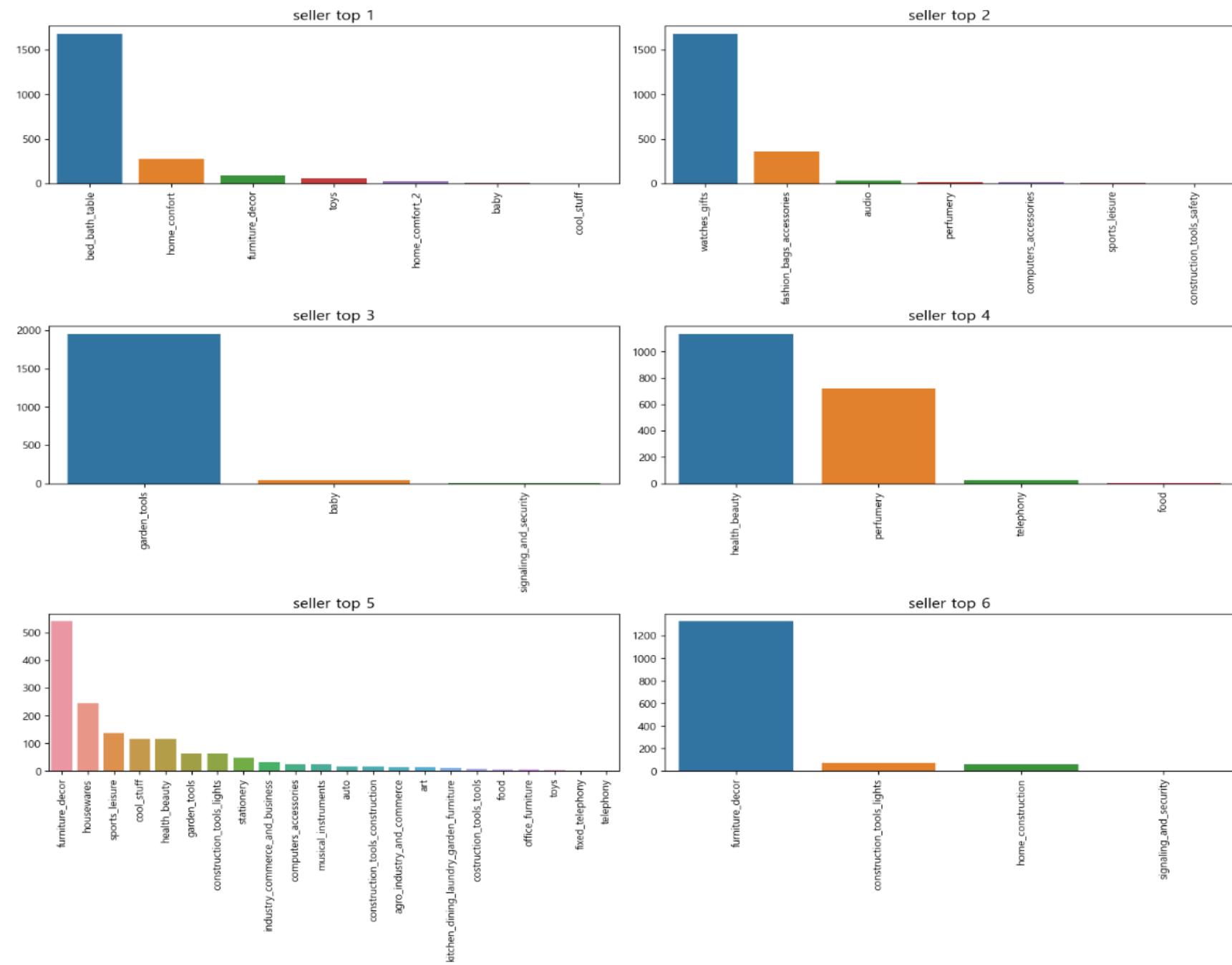
결제수단

모든 판매건의 결제수단 중 credit card가 73% 이상으로 가장 일반적인 결제 수단이며 다음으로 boleto(19.5%), voucher(5.4%) 등이 사용됐다.



일자별 상품금액

주문수가 가장 많았던 2017년 11월에 금액이 가장 높았고 그 이후로 하락세를 보여준다.



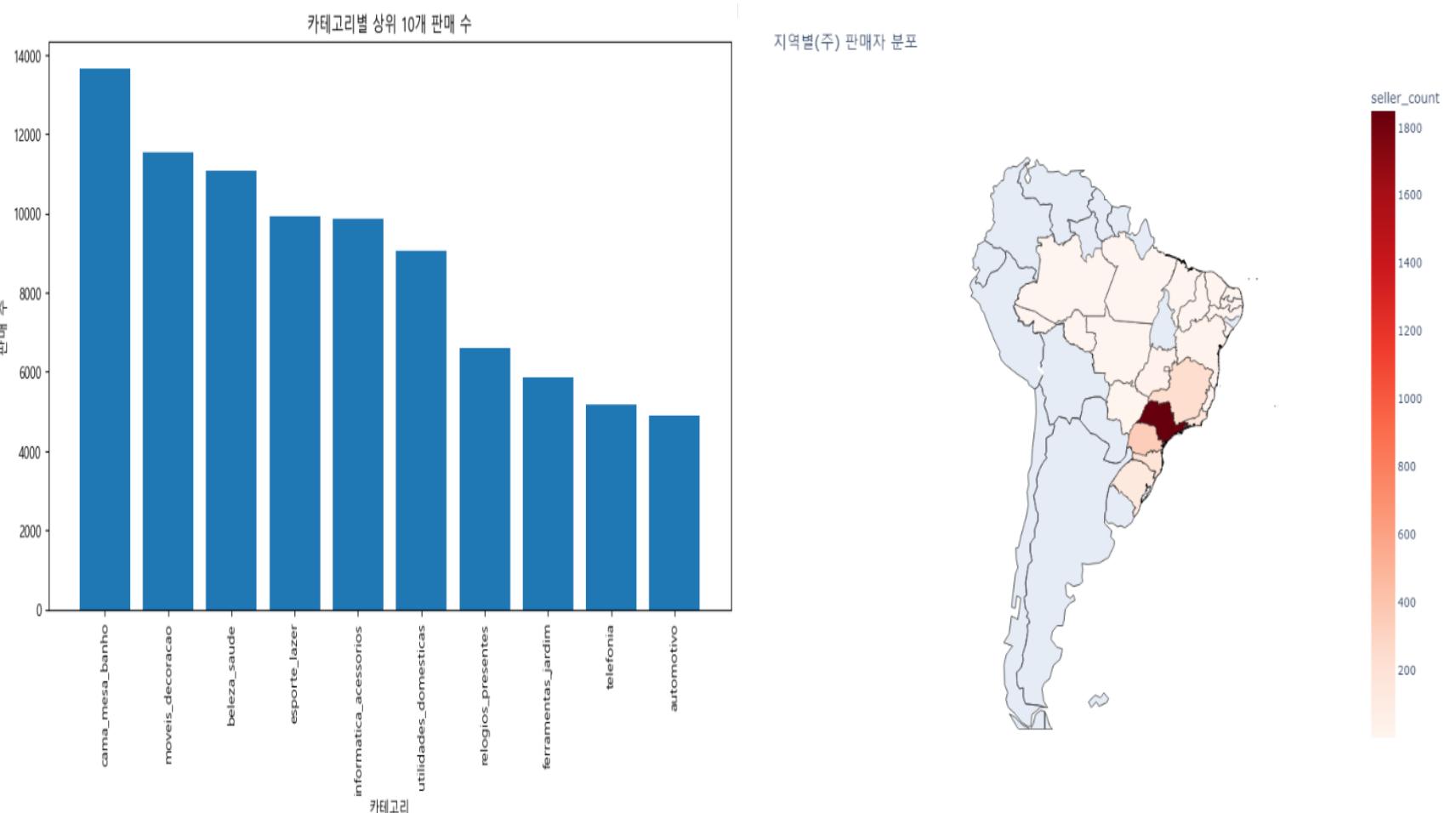
TOP6 판매자 상품 카테고리

seller top 6 까지 1~2가지 카테고리가 높은 비중을 차지했다.

seller top 5는 판매 비중이 고루 분포되어 있다.

판매수 & 판매자 분포

cama_mesa_banho(침대테이블목욕)가 가장 많은 판매를 기록했다.
상파울루가 주문량이 가장 많았고, 역시 판매자도 가장 많았다.

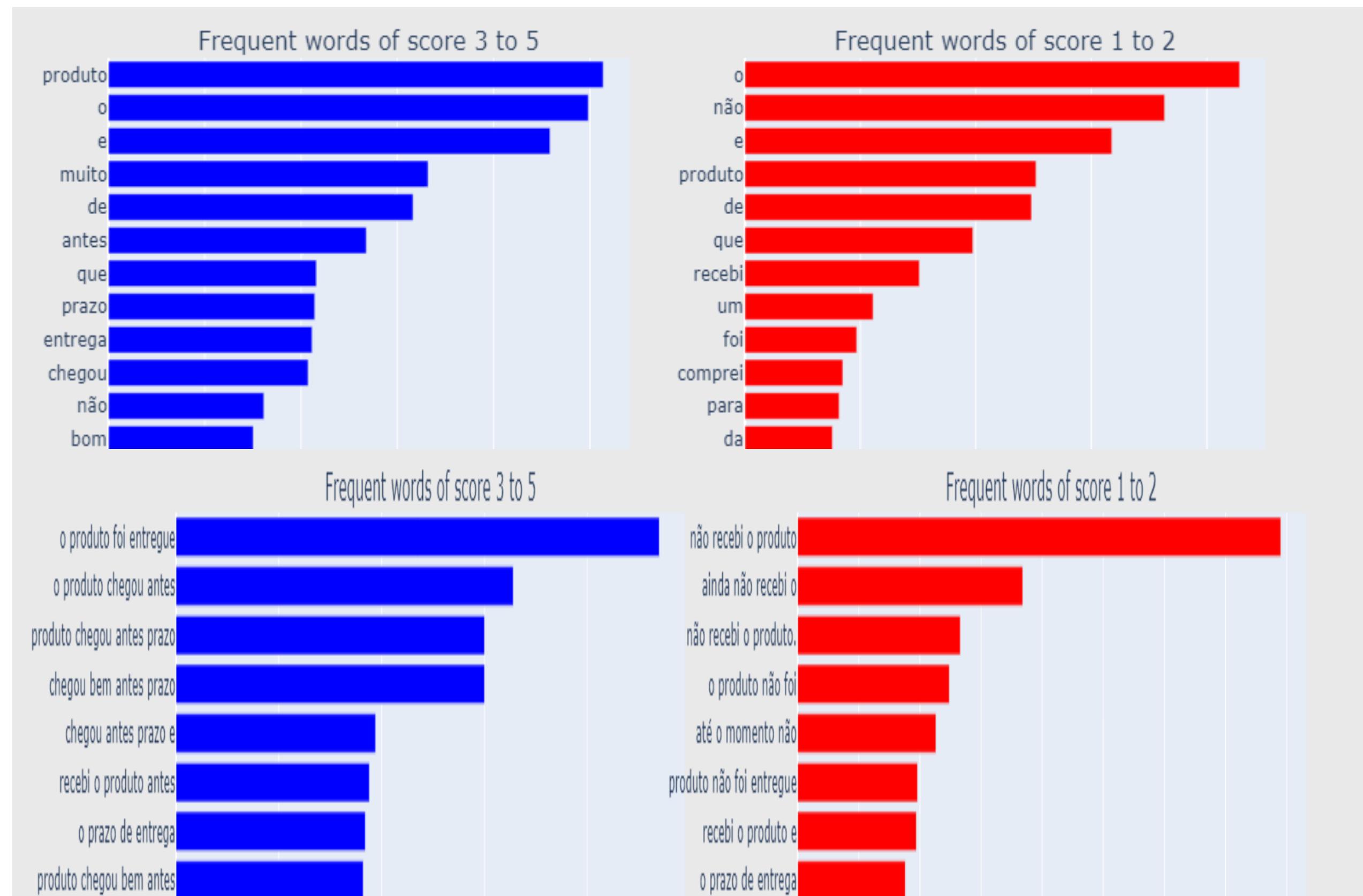
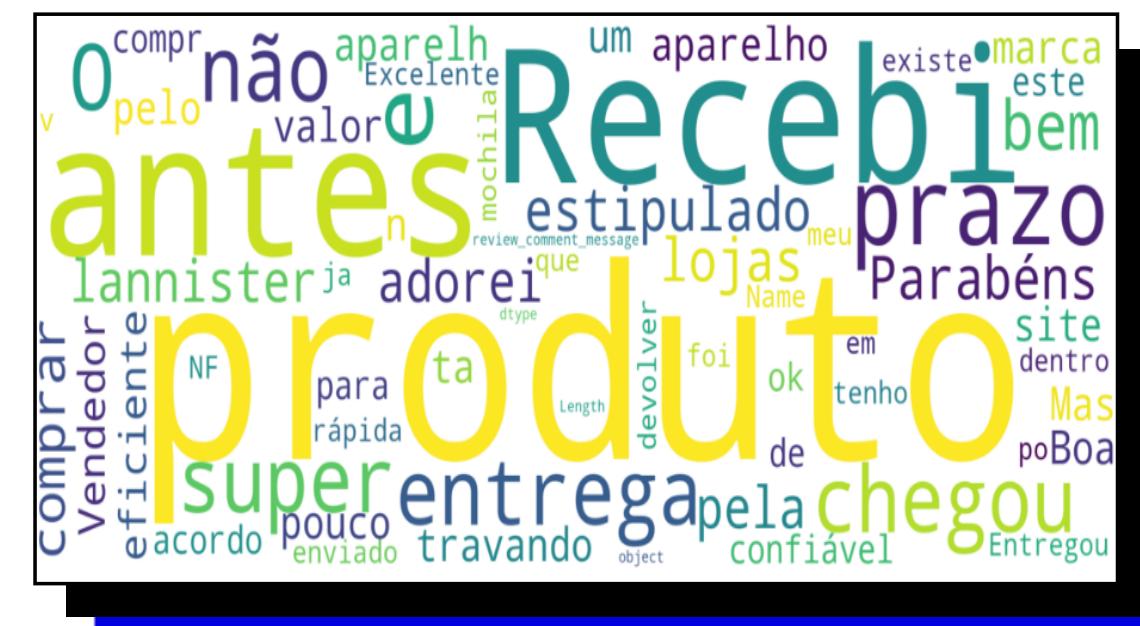


WORDCLOUD & N-GRAM

1-gram : produto(제품), muito(매우)
nao(아니오), que(무엇)

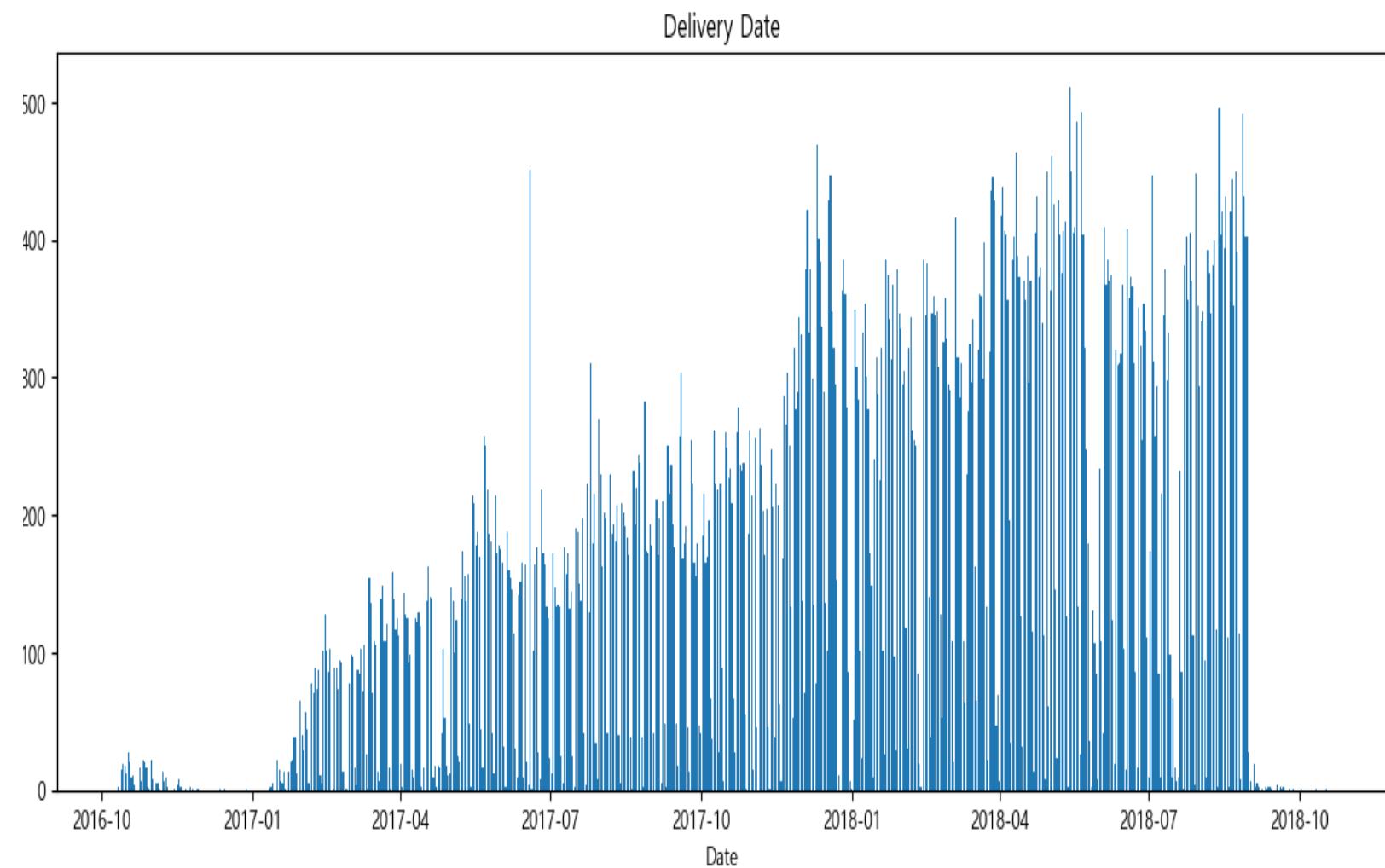
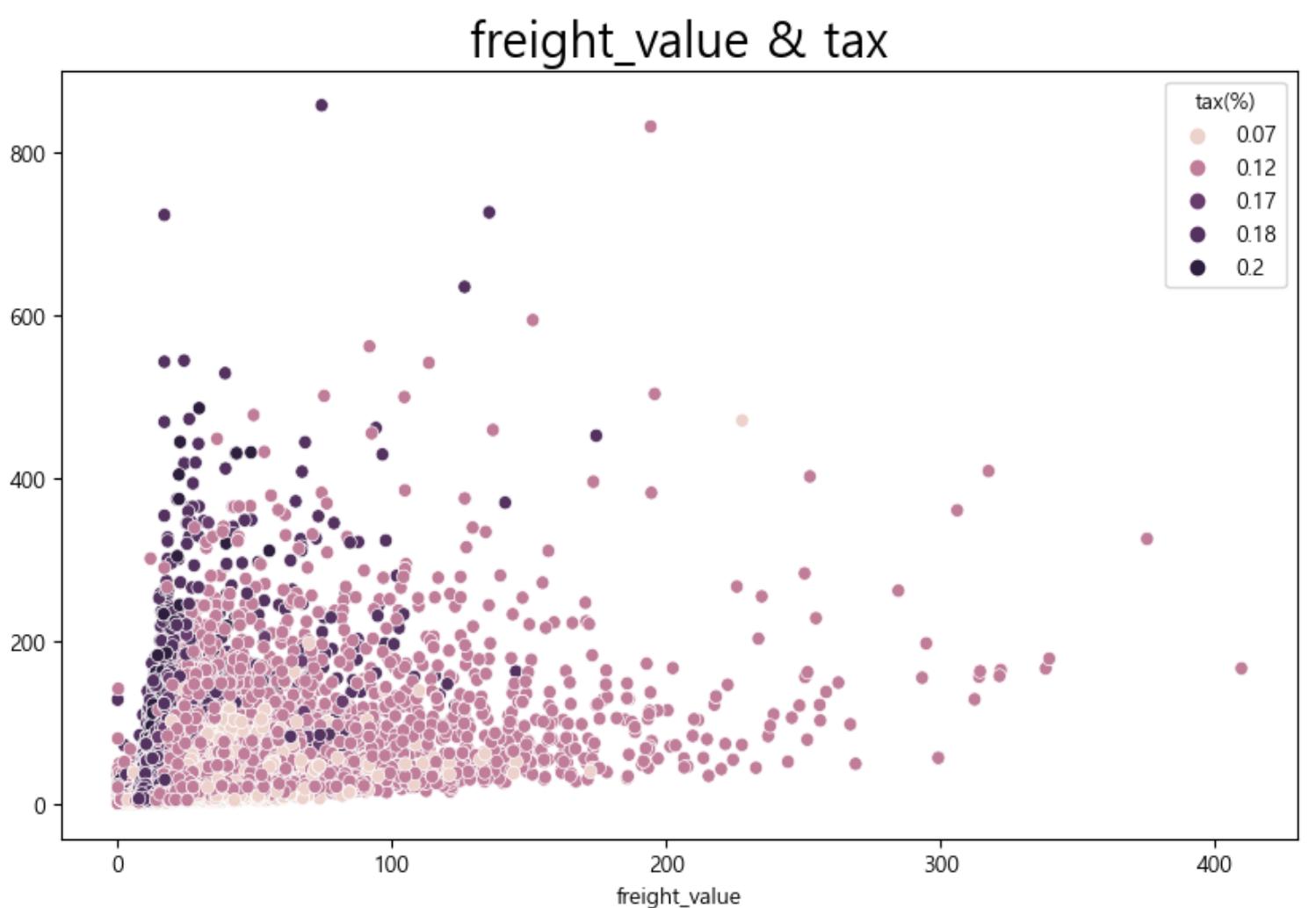
4-gram : o produto foi entrgue(제품이 배송되었습니다), nao recebi o produto (상품을 받지 못했다)

wordcloud : produto(제품), antes(전에)



배송비와 세금

브라질에서 우편화물의 경우 과세 대상 금액에 대해 세금 부과
과세율이 높을수록 배송비가 낮은 경향이 있다.



배송완료 일자

주문수가 가장 많았던 2017년 11월에 오히려 배송완료 상태가 많지
않았고, 그 다음달인 12월부터 배송완료 상태가 급증한다.



매출 예측 모델

delivery_score
3.0
4.0
4.0
3.0
3.0

3.0
3.0
2.0
4.0
4.0

#DELIVERY_SCORE

실제배송일/배송예정일로 delivery_Point_row 생성
배송기간과 예상배송일 기준 등급화
배송관련feature들로 모델링

pred1_values	total_price	price_score	total_score	total_score_scaled
0.639977	88.92	3.0	59.906764	0.017881
0.639977	700.89	5.0	453.553548	0.148713
0.995973	363.96	5.0	367.494178	0.120110
0.661137	44.08	2.0	31.142897	0.008321
0.985685	40.01	2.0	41.437237	0.011742
...
0.639977	67.62	3.0	46.275251	0.013350
0.639977	56.78	3.0	39.337900	0.011045
0.985131	155.63	4.0	157.315932	0.050256
0.639977	528.69	5.0	343.349492	0.112086
0.639977	36.77	2.0	25.531958	0.006456

#REVIEW_SCORE

제품가격+배송비로 total_price 생성
생성된 total_price 사분위수 기준 등급화
reviews_dataset 활용해 모델링
위에서 구한 수치들을 통해 total_score
생성 후 정규화

order_purchase_timestamp	price
2016-09-25	0.00
2016-10-02	100.00
2016-10-09	14264.35
2016-10-16	765.18
2016-10-23	0.00
...	...
2018-08-12	118139.64
2018-08-19	122883.55
2018-08-26	72605.00
2018-09-02	9078.99
2018-09-09	145.00

#PRICE

주(week)별 매출 집계

매출예측모델링

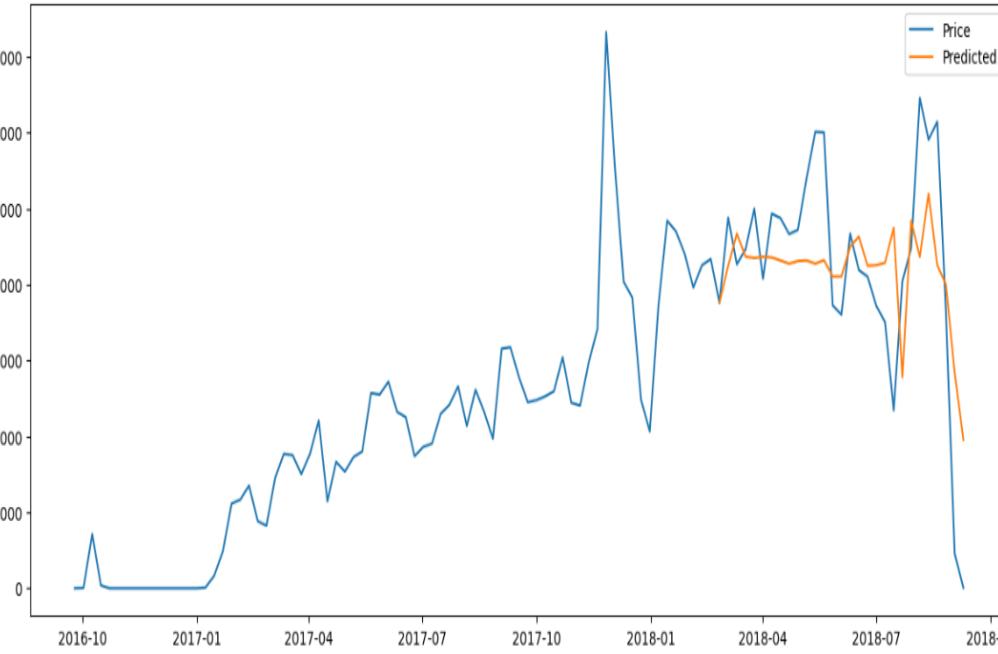
- 단일특성(Price) XGBRegressor
- 다중특성(Price, review, delivery)

XGBRegressor

- 단일특성(Price) 딥러닝(LSTM)
- 다중특성(Price, review, delivery)

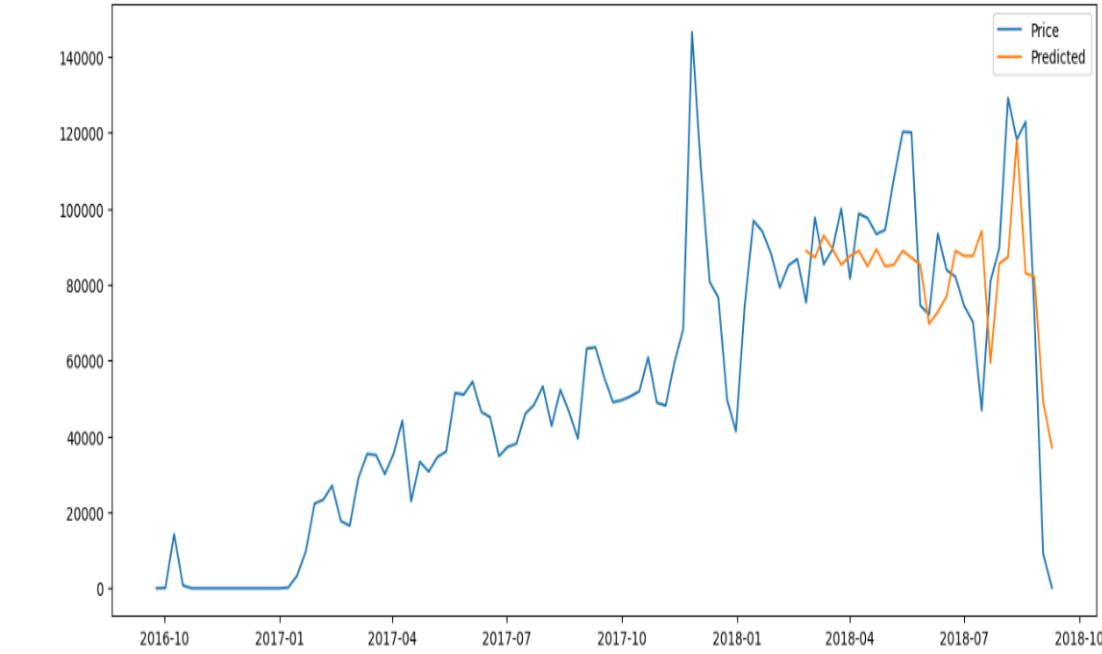
딥러닝(LSTM)

#단일 XGBRegressor



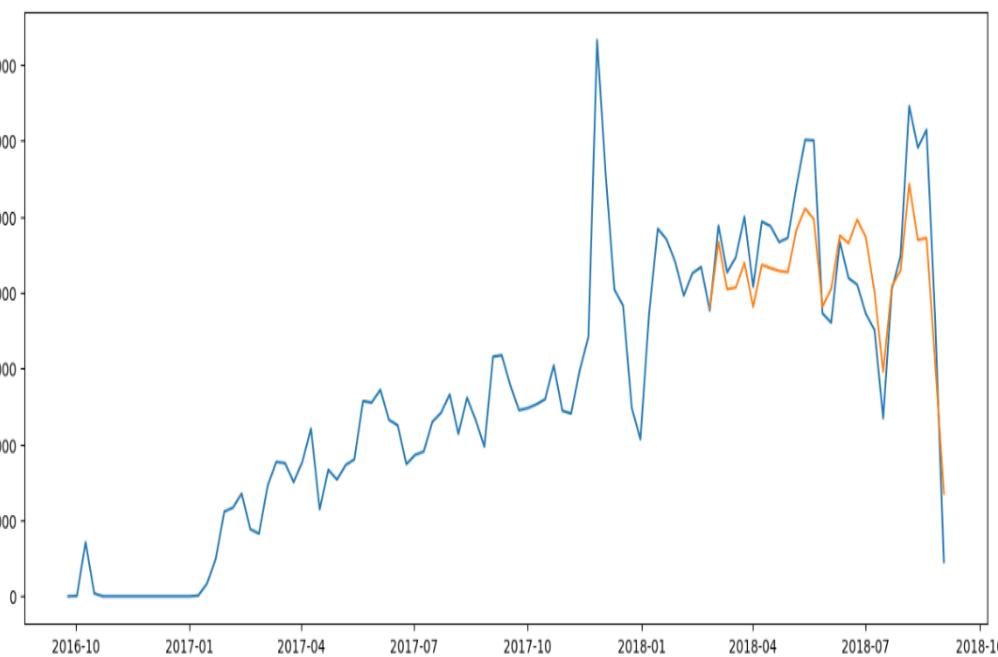
RMSE : 22335.00

#다중 XGBRegressor



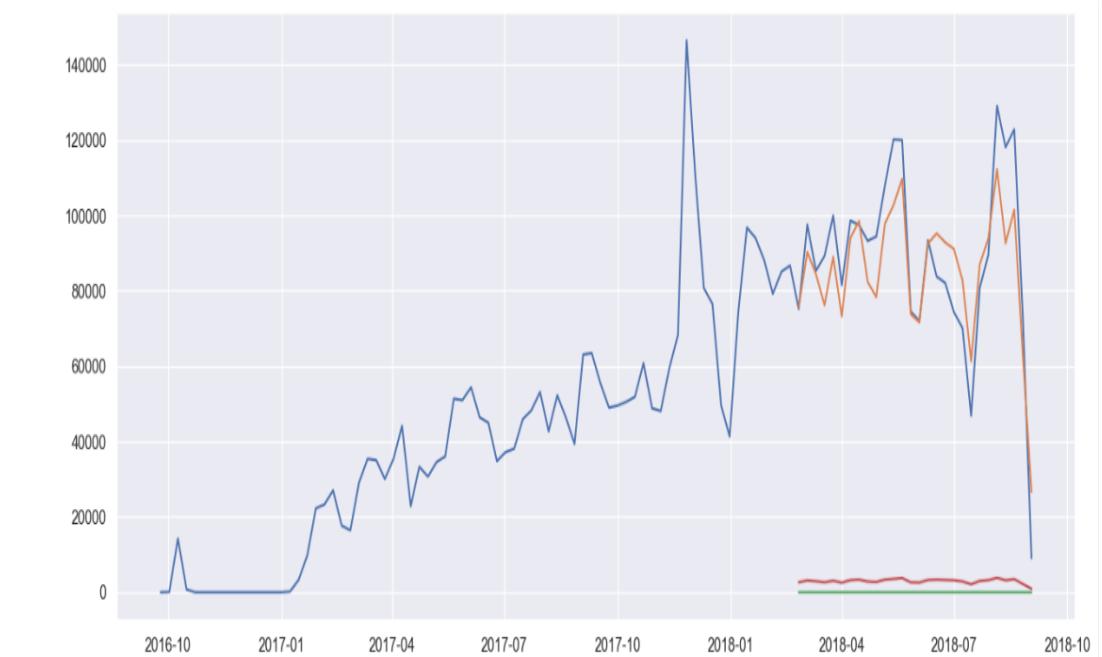
RMSE : 21446.46

#단일 LSTM



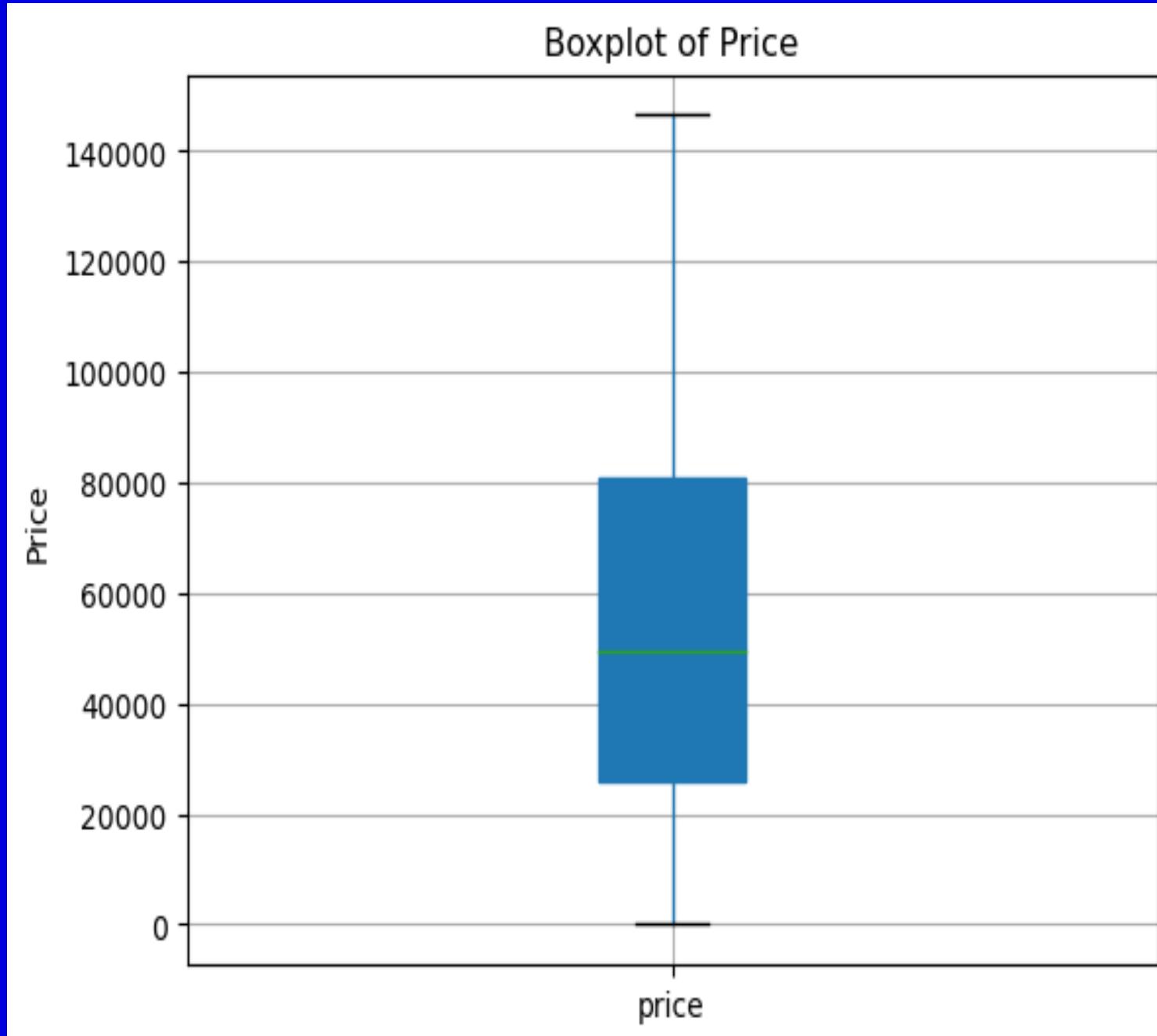
RMSE : 14825.18

#다중 LSTM



RMSE : 10086.48

#'PRICE' BOX-PLOT



성능이 가장 좋았던
LSTM(10086.48)을 통해
다음주 매출을 Predict한 결과 261.53\$ 산출



RFM& 고객세분화

RFM SCORE

가중치를 부여하지 않고 RFM_SCORE 생성

각 컬럼 outlier 제거

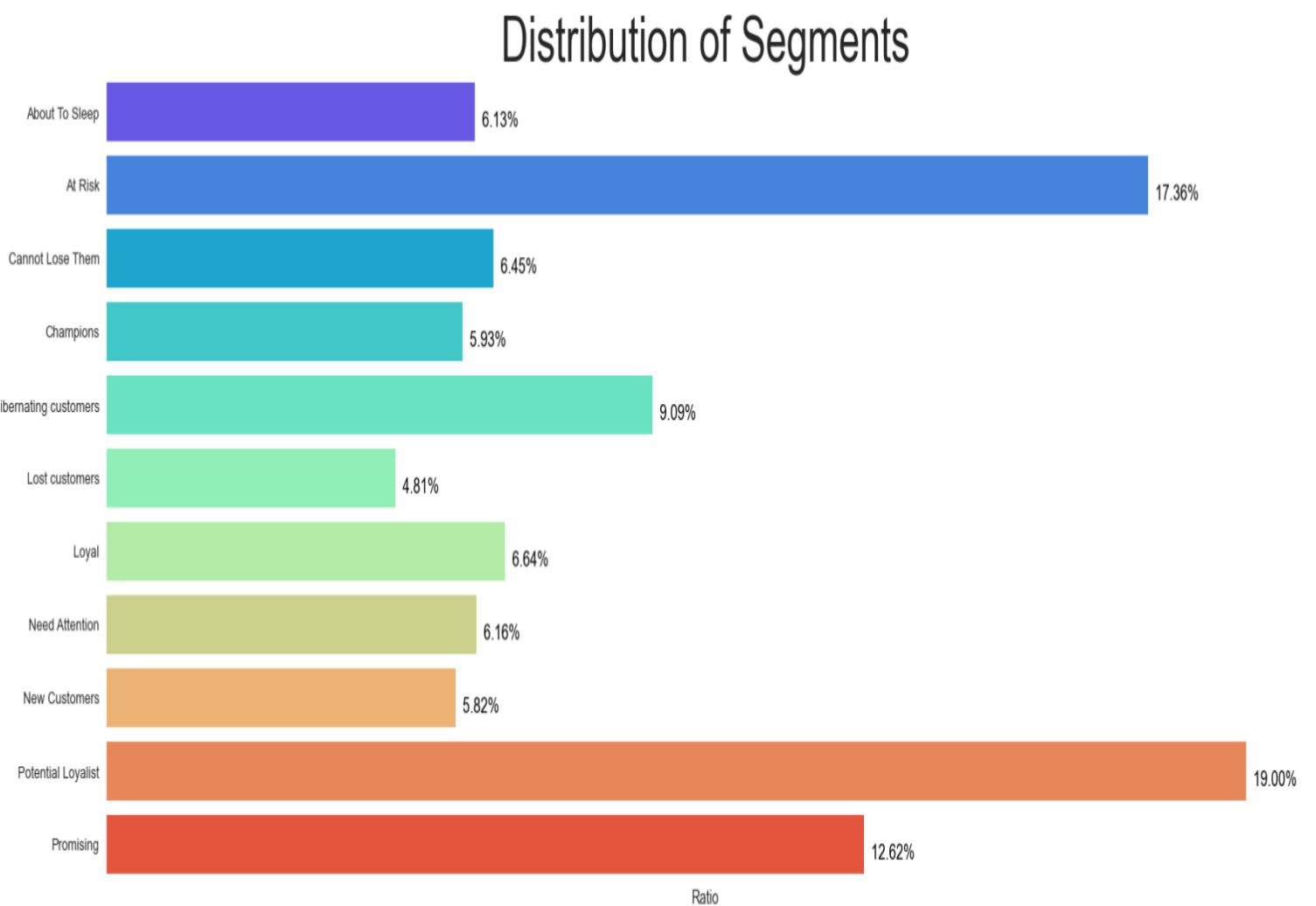
Recency	Frequency	Monetary	recency_score	frequency_score	monetary_score	rfm_score
110	1	141.90	4	1	4	414
113	1	27.19	4	1	1	411
287	1	196.89	2	1	5	215
145	1	166.98	4	1	4	414
130	1	35.38	4	1	1	411
...
516	1	32.42	1	5	1	151
32	1	35.36	5	5	1	551
63	1	89.19	5	5	3	553
361	1	55.00	2	5	2	252
13	1	63.42	5	5	2	552



SEGMENTATION

RFM 각 Score의 segmentation 진행

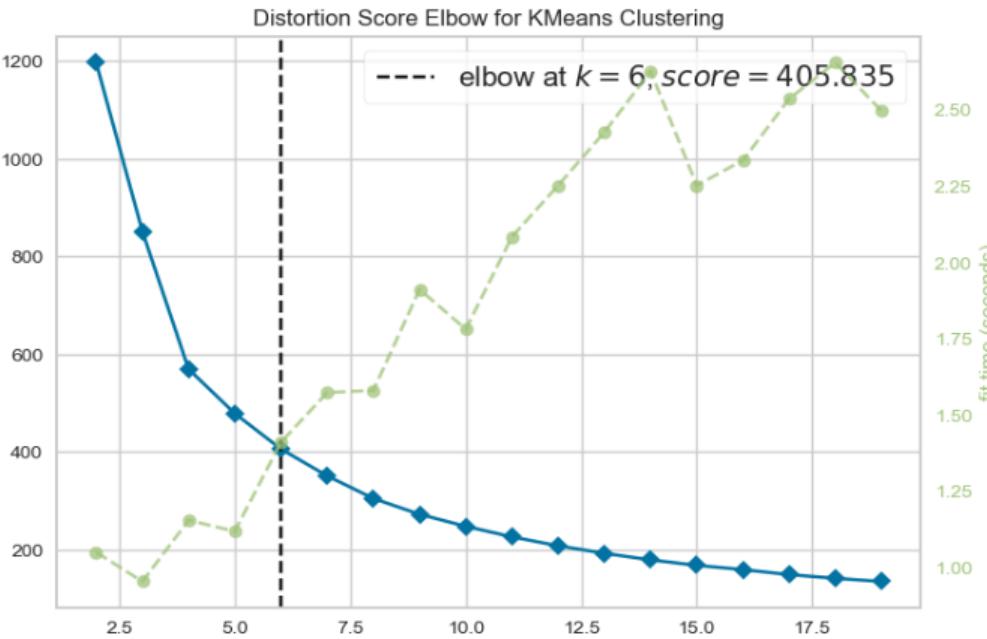
OLIST 내 고객 segment 별 비중 확인 후 고객 등급화



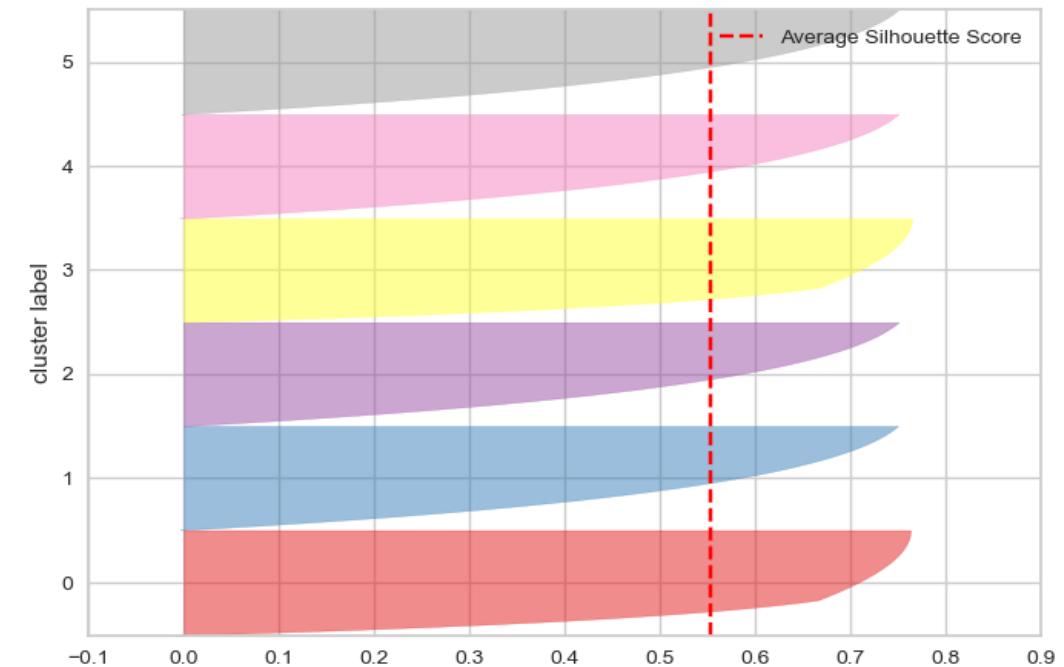
RFM_SCORE & 세분화

- 앞서 구한 rfm_score -> KMeans
- Silhouette score를 통해 평가
- 가중치를 부여한 rfm_score 산출
- 새로구한 rfm_score 별로 고객 세분화

#KMeans



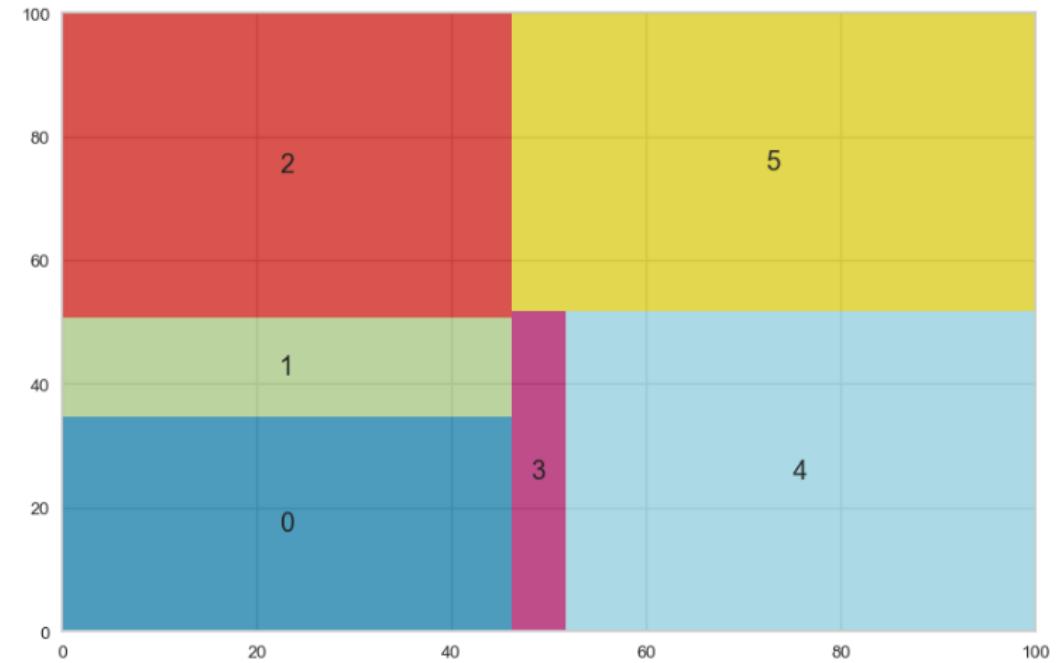
#Silhouette score



#RFM_SCORE

recency	frequency	monetary	Cluster	RFM_SCORE
0.165478	0.0	0.123513	5	73.789009
0.169757	0.0	0.016040	5	37.983479
0.417974	0.0	0.175033	2	136.572287
0.215407	0.0	0.147010	4	90.733106
0.194009	0.0	0.023713	4	44.940808
...
0.744650	0.0	0.020940	0	142.680169
0.054208	0.0	0.023694	5	19.880177
0.098431	0.0	0.074128	5	44.967890
0.523538	0.0	0.042095	2	110.252626
0.027104	0.0	0.049984	5	23.968969

#SEGMENTATION



가치 제공

: 해당 프로젝트를 통해 제공하는 가치



경영전략제시

상품매출을 예상을 통해 매출전략 제시

RFM으로 고객을 Segmentation함으로서 고객 타겟팅

THANK YOU

FINAL_PJT 1팀

#고지현 #김정빈 #박건우 #윤정옥 #정기중 #황지우