

제 7회 롯데멤버스 빅데이터 경진대회


로티로리

김보현 유광열 윤성식

CONTENTS

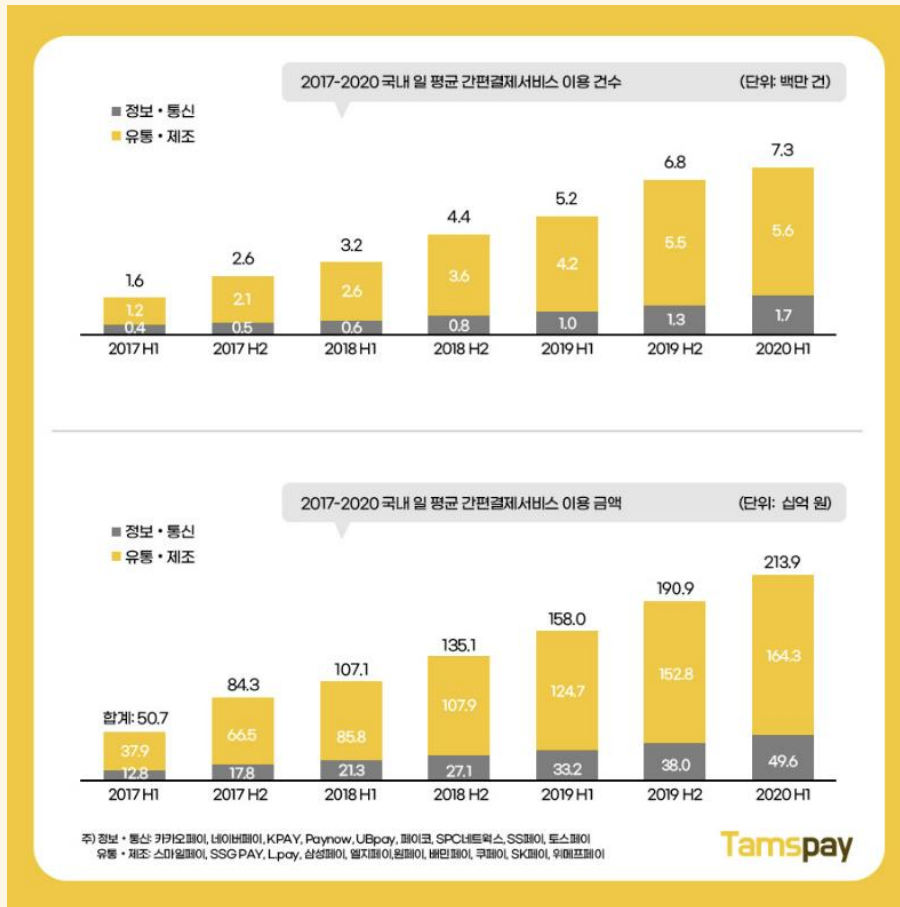


- 01 서론
- 02 데이터셋 설명
- 03 예측 모델
- 04 클러스터링
- 05 마케팅 제안 및 결론



01 서론

01 서론:개요



(2017 ~ 2020 국내 간편결제서비스 이용 추세)

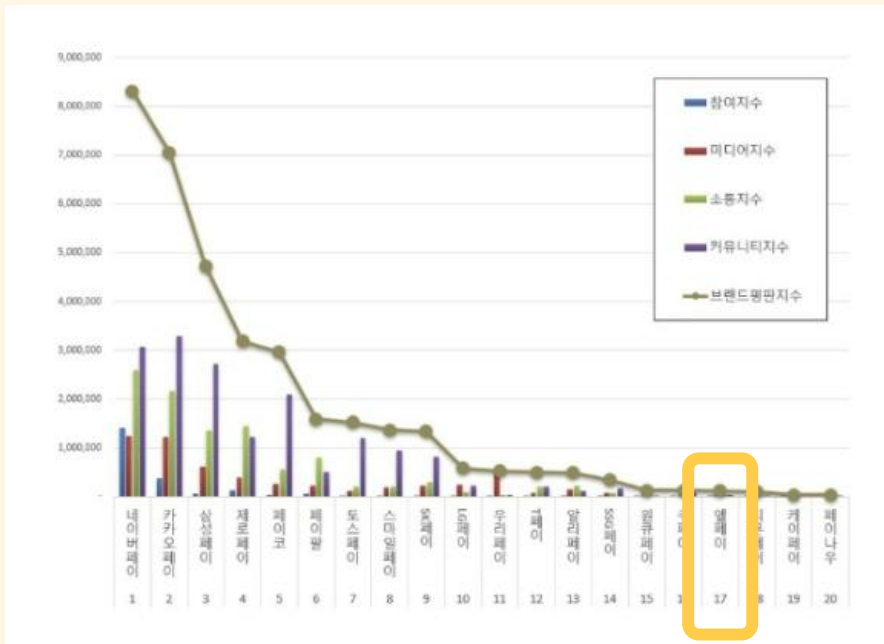
- 간편 결제 서비스¹ : 온라인과 오프라인 상거래에서 빠르고 간편하게 결제하는 전자 결제 서비스이며, 스마트폰, 스마트워치 등 기기에 저장된 생체 정보, 신용 카드 정보 등을 이용해 바로 결제되기 때문에 추가적인 인증 수단이 필요하지 않다.
- 현재 간편 결제 시장 추세² : 현재 여러 기업에서 간편 결제 시장에 뛰어들고 있으며, 기업 고유의 결제 시스템을 만들고 있다. 2017년부터 2020년까지의 간편 결제 서비스 이용 건수와 이용 금액은 꾸준히 증가하는 추세를 띠고 있음을 왼쪽 그림을 통해 확인할 수 있다.

01 서론:개요

간편 결제 서비스는 오프라인과 온라인 거래에서 모두 사용이 가능하지만, 주로 온라인 거래에서 자주 사용된다.

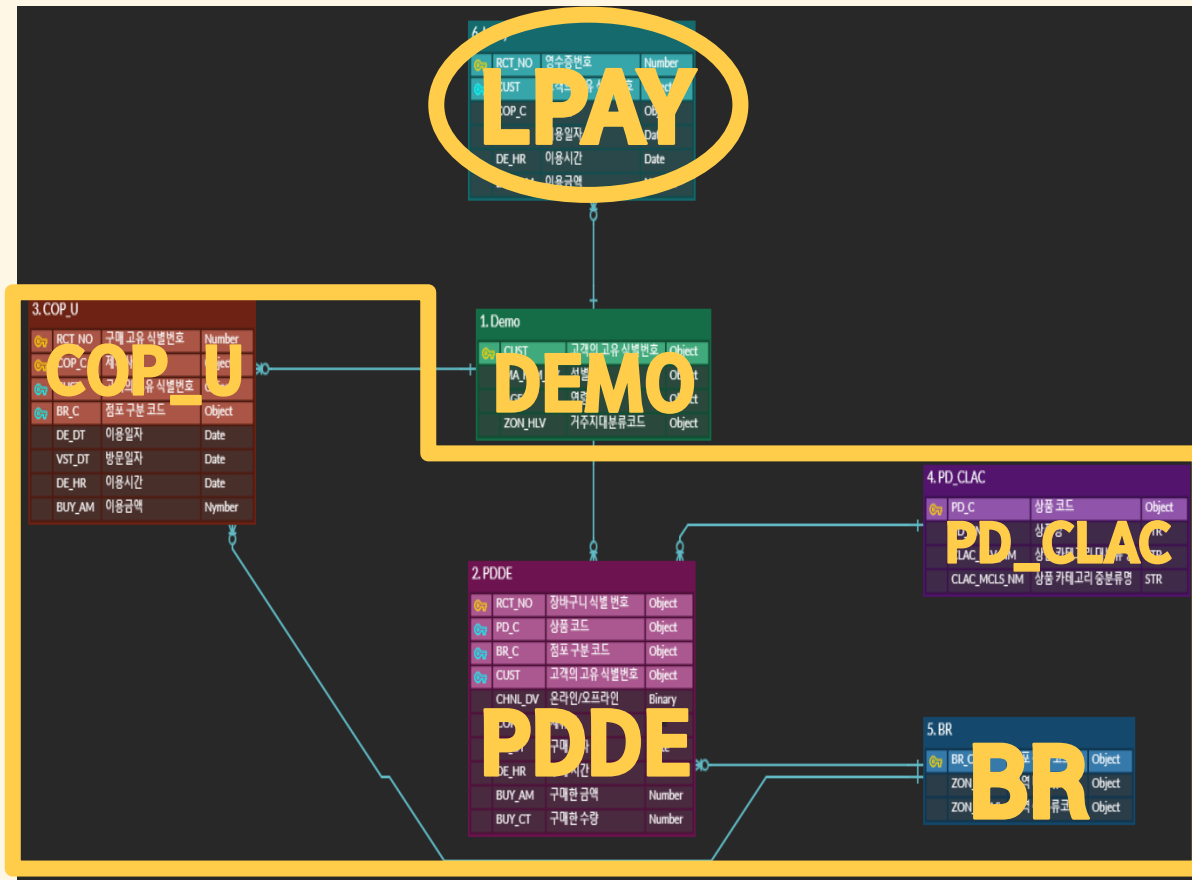
롯데는 다양한 종류의 유통사 및 제휴사가 존재하며, 전국에 다수의 지점을 보유하고 있다. 따라서, 간편 결제 서비스 시장에서 브랜드 평판 상위권을 차지하고 있는 네이버, 카카오와 같이 IT 기반 기업과는 다르게 오프라인에서의 거래가 활발할 수 있다는 장점이 존재한다.

오프라인에서의 강점이 돋보이는 롯데에서 제공하는 간편 결제 시스템인 L.PAY는 사용률이 저조할 수 밖에 없고, 실제로 브랜드 평판 17위로 낮은 순위를 보이고 있다.



(2022년 7월 기준)³

01 서론:SCHEMA



< SCHEMA >

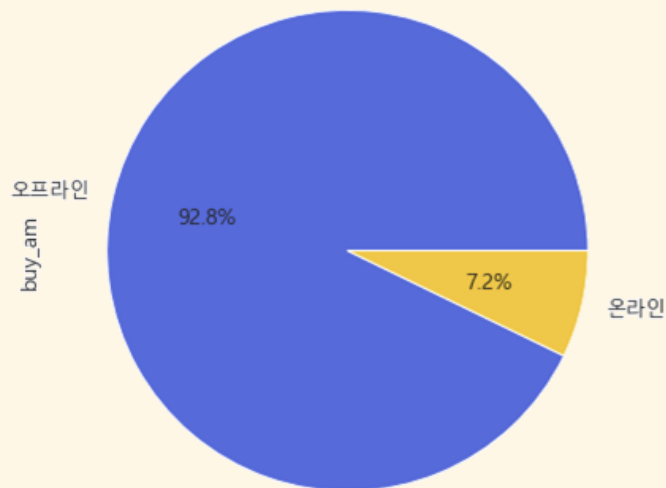
주어진 데이터를 구조화했을 때,
L.PAY 데이터는 Demo 데이터와만
병합이 가능할 것으로 판단했다.

PDDE, COP_U, PD_CLAC, BR
데이터들은 상호 의존적이라는
것을 알 수 있으며, 본 분석에서는
해당 데이터들을 조합하여 고객의
경험 데이터라고 칭하기로 했다.

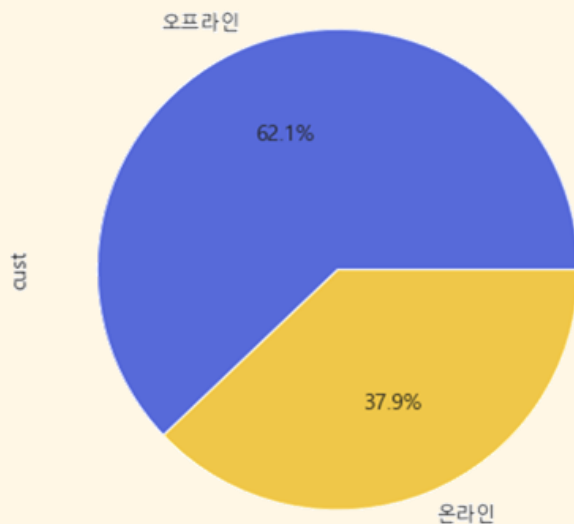
L.PAY 데이터와 고객 경험
데이터는 병합이 불가능하기
때문에 데이터를 따로 분석하였다.

01 서론:고객경험데이터분석

고객경험별 채널별 거래 총액



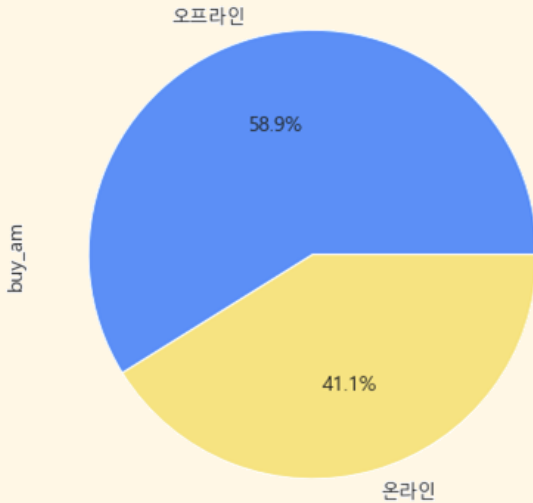
고객경험별 채널별 고객 방문 수



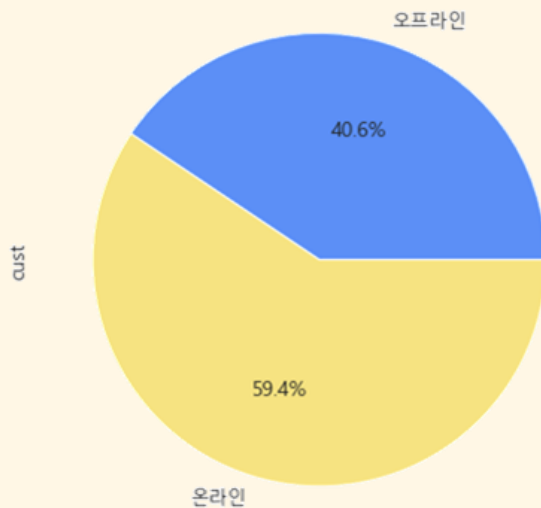
- 고객 경험 데이터를 통해 채널 별 거래 총액을 보았을 때, 92.8%의 거래가 오프라인에서 진행되었으며, 온라인 거래 총액보다 약 13배 많은 금액을 보였다.
- 채널 별 고객의 방문 수를 계산했을 때, 오프라인에서 방문한 고객의 비율은 전체의 62.1%로, 온라인에서 거래를 진행한 고객보다 1.6배 많은 방문 수를 보였다.
- 거래 총액과 방문 수 모두 오프라인 데이터가 온라인 데이터보다 많은 것을 볼 수 있는데, 이는 롯데 계열사 중 오프라인에서 구매할 수 있는 거래처가 많이 존재한다는 것을 보여주는 결과이다.

01 서론 : L.PAY 데이터 분석

lpay 사용자들의 채널별 거래 총액

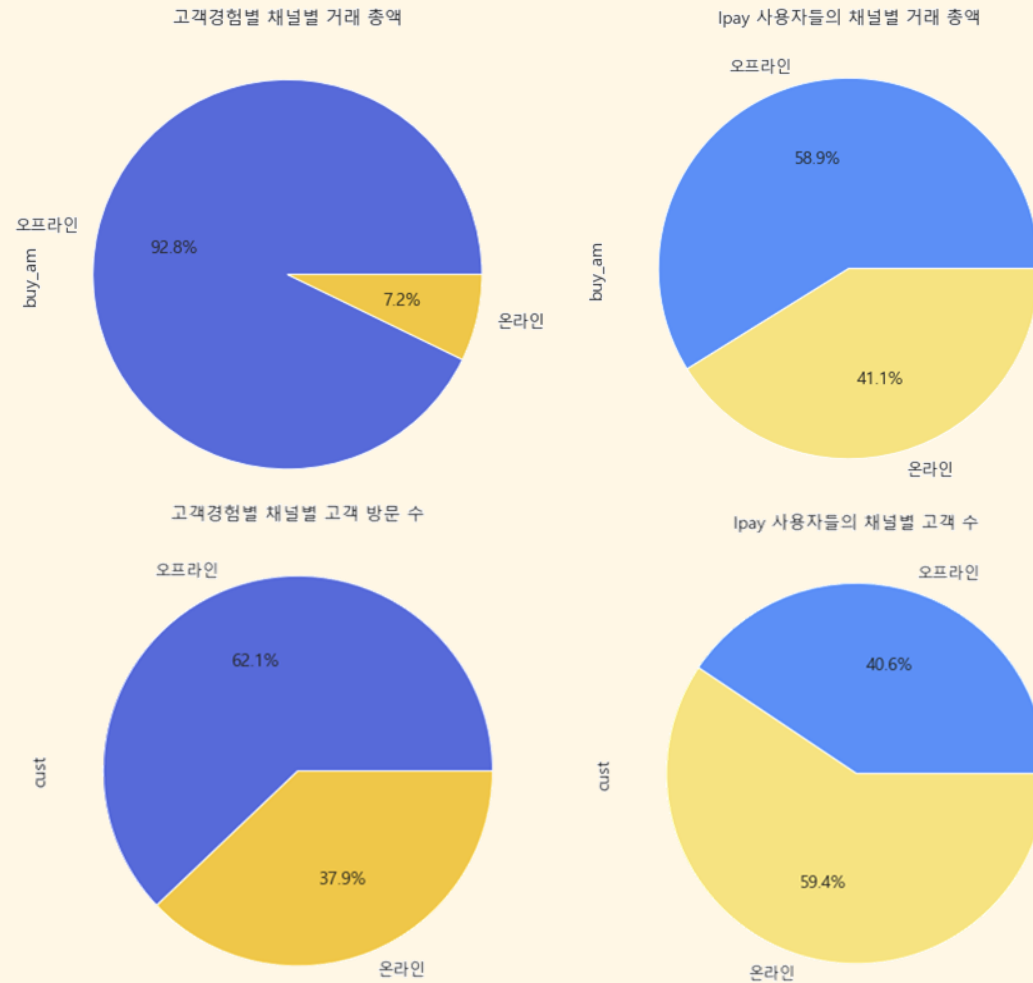


lpay 사용자들의 채널별 고객 수



- L.PAY 사용자들의 거래 총액을 보았을 때, 오프라인에서의 거래가 58.9%로 온라인보다 많다.
- L.PAY 사용자들의 고객의 수는 온라인이 59.4%로 오프라인보다 더 많은 고객 수가 존재한다는 것을 알 수 있다.
- 고객 경험 데이터에서의 압도적으로 많았던 오프라인 거래 총액과 비교할 때, L.PAY 사용자들의 오프라인 거래 총액의 비율은 훨씬 작다.
- 고객 경험 데이터에서의 고객 수는 오프라인이 더 많았지만 L.PAY 데이터에서는 온라인이 더 많은 것을 볼 수 있다.

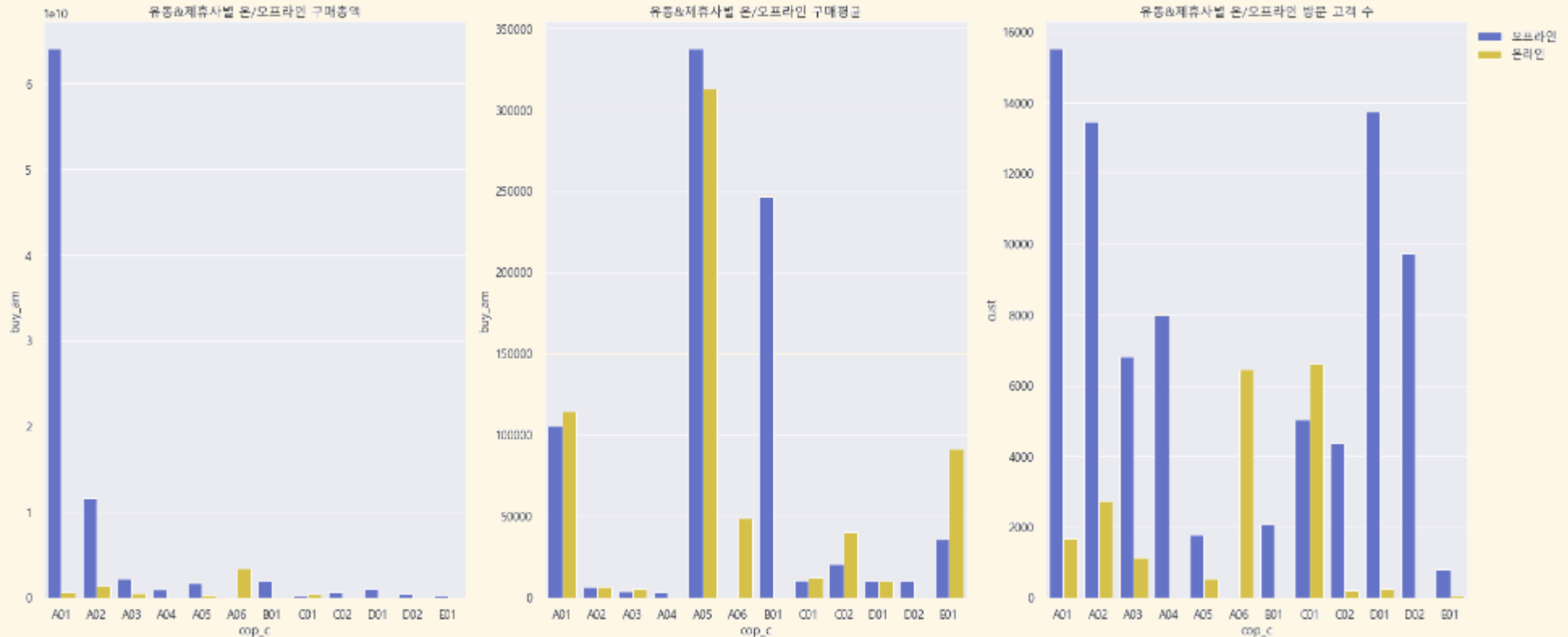
01 서론



위의 파이 도표를 통해, 채널 별로 비율의 차이가 많이 존재한다는 것을 알 수 있었다. 또한 앞서 확인한 스키마를 통해 분석하고자 했던, 고객 경험 데이터와 L.PAY 데이터를 분리해 보았을 때, 두 데이터가 서로 다른 특징을 가지고 있다는 것을 확인했다.

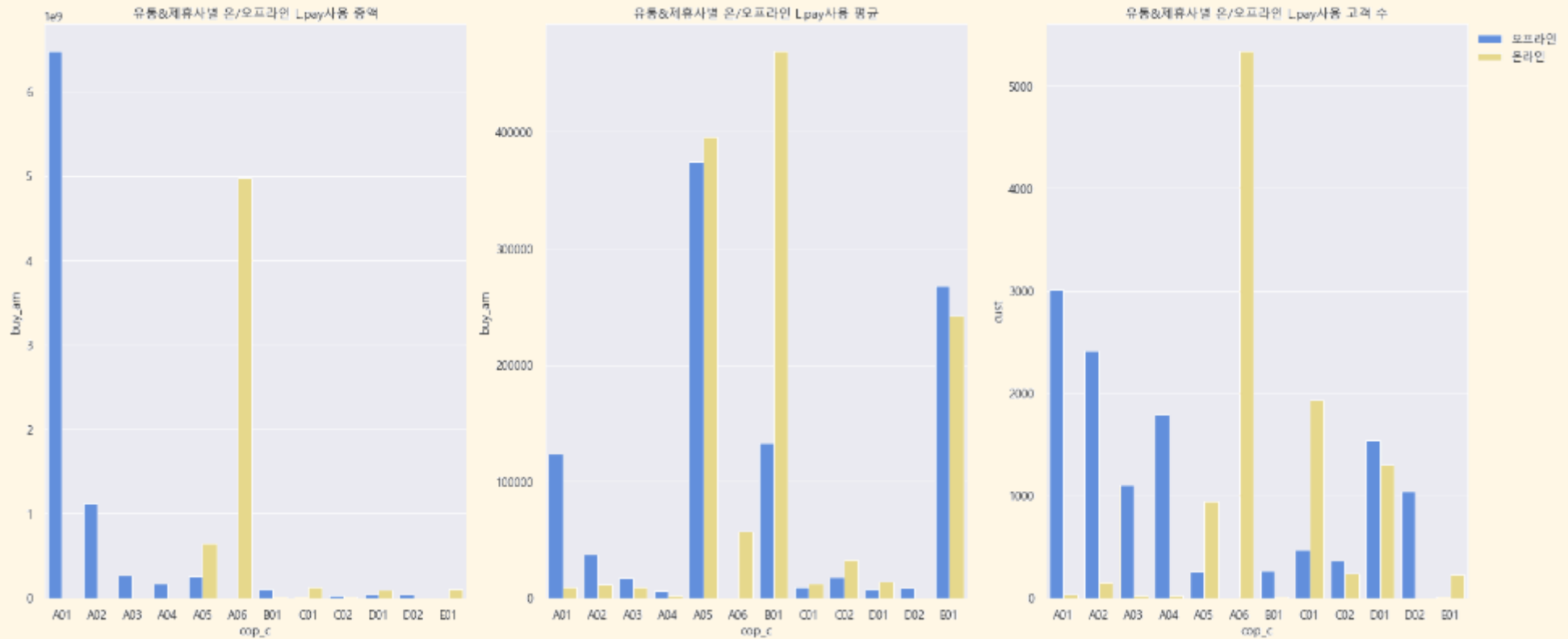
앞서 살펴본 고객 경험 데이터와 L.PAY 데이터의 채널 별 특징을 조금 더 자세히 살펴보기 위해 제휴사와 유통사 별 구매총액, 구매평균, 방문 고객 수 를 막대그래프로 시각화 하였다.

01 서론:제휴사별 데이터 분석



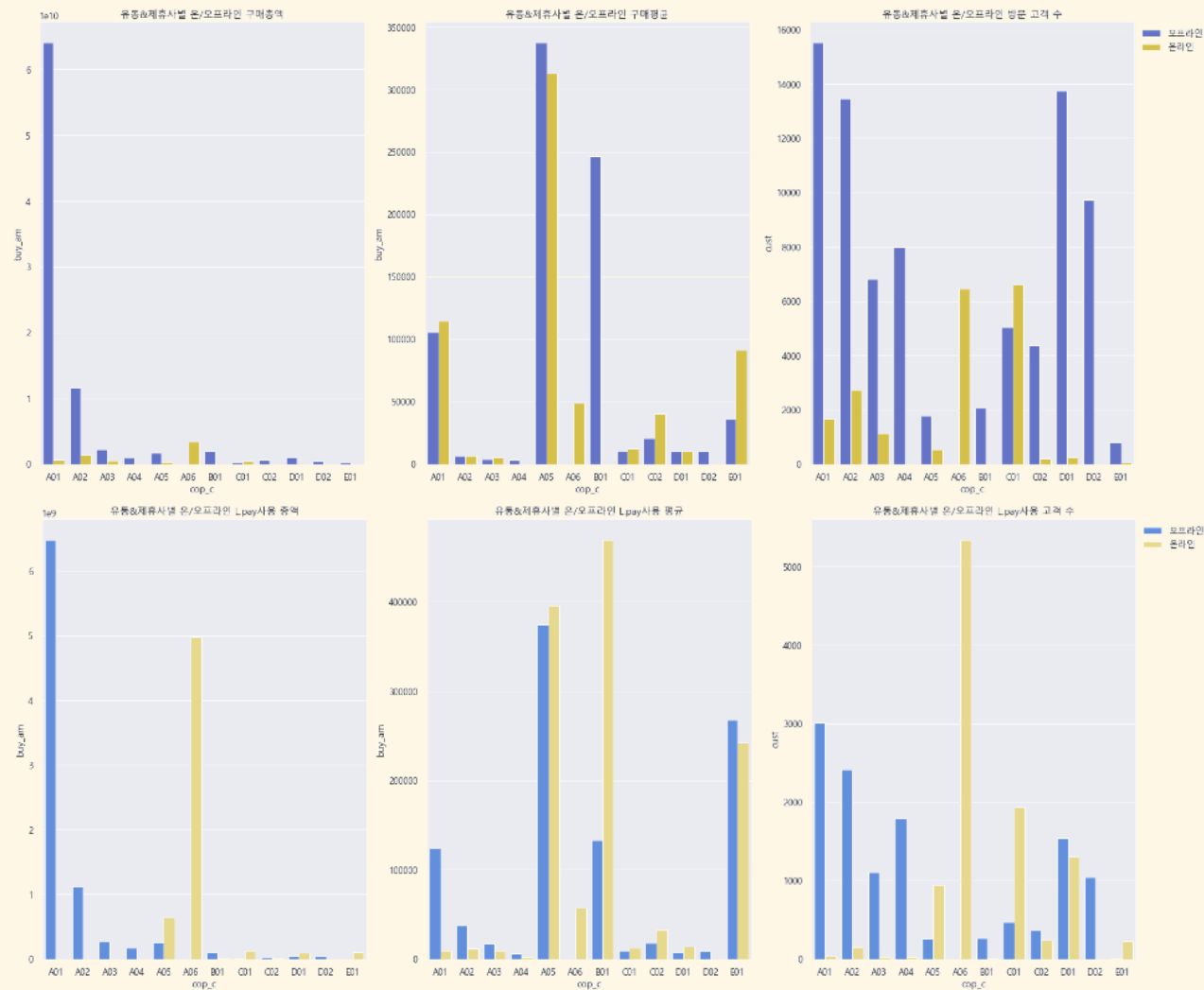
- 위의 그래프 3개를 봤을 때, **오프라인 채널**에서의 총 구매 금액, 평균 구매 금액, 방문 고객 수가 대부분 **높은 것**을 알 수 있다.
- A04, B01, D02 제휴사는 온라인에서의 거래가 발견되지 않았고, A06 제휴사는 오프라인에서의 거래가 발견되지 않았다.
- 총 구매 금액에서는 A01 제휴사가 다른 제휴사들에 비해 압도적으로 높지만 상품의 평균 구매금액은 3번째로 높은 수치로 나타난다.

01 서론:제휴사별 데이터 분석



- L.PAY 사용 총 구매금액에서도 A01 제휴사가 가장 높은 수치를 기록했지만 온라인 이용 고객만 존재하는 A06 제휴사가 두번째로 높은 수치를 보이고 있다.
- A06 제휴사는 L.PAY 온라인 이용 고객만 존재하지만, 모든 제휴사를 통틀어서 L.PAY 사용 고객 수가 가장 많은 것을 확인할 수 있다.
- E01 제휴사는 L.PAY 평균 구매 금액이 높게 나타나지만, L.PAY 총 사용 금액과 L.PAY 총 고객 수는 낮은 수치를 보인다.

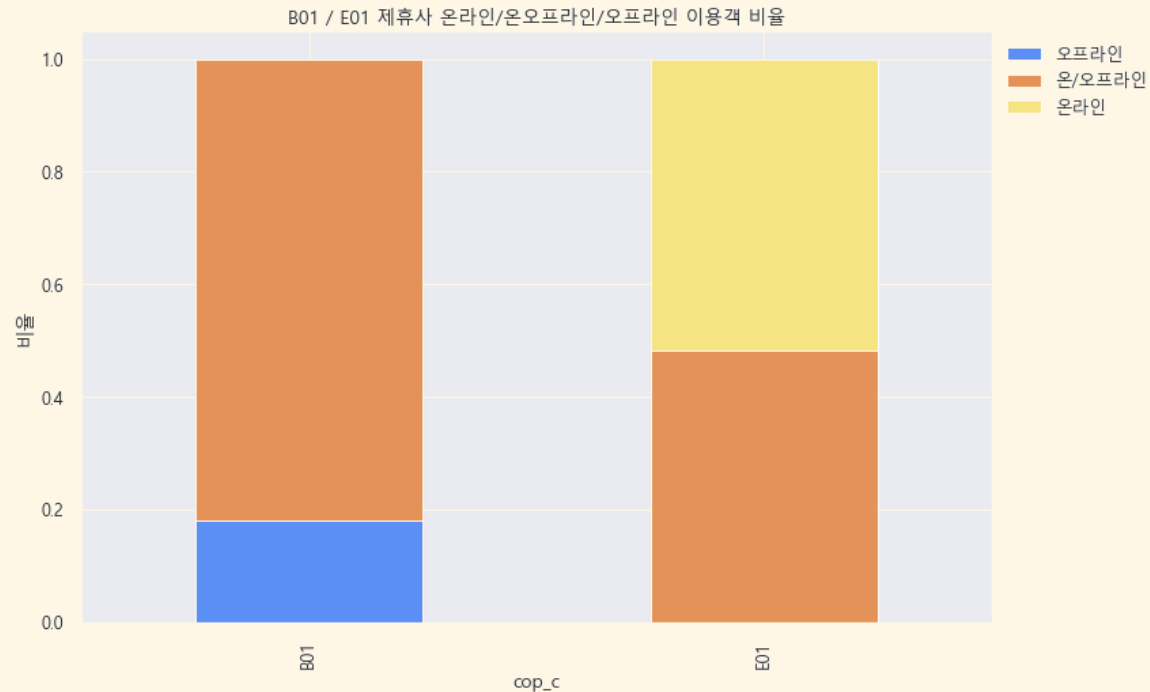
01 서론



E01 제휴사의 경우에는 고객 경험 데이터에서 총액이나 방문자 수는 오프라인이 많지만, L.PAY의 경우 총액이나 방문자 수가 온라인이 더 많음을 확인할 수 있었다.

B01 제휴사의 경우 오프라인 고객 경험 데이터만 존재하는 데도 L.PAY는 온라인에서 사용하는 고객이 존재한다. 고객 경험 데이터에 오프라인 이용만 존재한다고 해서, 온라인에서 L.PAY를 사용하지 않는다고 단정지을 수 없다.

01 서론



앞서 특이점이 있는 제휴사였던 B01과 E01을 오프라인과 온라인으로 이분화 시킨 것이 아닌 온/오프라인 혼용 채널까지 포함하여 위 차트는 B01과 E01 제휴사의 채널 별 이용 비율을 나타낸다.

이 차트를 통해, 고객 중에 채널 간 경험에 상관 없이 L.PAY를 오프라인, 온라인을 동시에 사용하는 고객이 존재하기 때문에 이를 4가지로 구분 해야 하는 필요성이 존재한다.

⇒ L.PAY 사용자 (오프라인에서만 사용하는 고객 / 온라인에서만 사용하는 고객 / 혼용 고객)

⇒ L.PAY 미사용자

01 서론:마무리

앞에서 데이터를 분석한 결과, L.PAY 사용자를 4가지 유형으로 분류하여 개인화 마케팅을 진행해야 한다는 판단을 내렸다.

L.PAY 데이터에서 고객별로 4가지 유형으로 분류하고, 이를 구분할 수 있는 특징들을 만들어 모델링을 진행하여, 앞으로 고객이 L.PAY를 거래에 활용할 것인지, 활용하지 않을 것인지를 예측하고, 더 나아가 L.PAY를 거래에 활용할 경우 오프라인에서 사용할지, 온라인에서 사용할지, 혼용해서 사용할지를 예측한다.

이렇게 분류된 고객들을 클러스터링을 진행하여 비슷한 유형의 군집으로 나누고, 이 군집별로 특성을 나누어 각 군집에 대해 소비자가 상품을 선택하도록 유도하는 개별화된 넛지 마케팅⁴을 시행하고자 한다.



02 데이터셋 설명

02 데이터셋 설명: 데이터셋

※ 내부 데이터

- 고객 경험 데이터에 L.PAY 사용 데이터의 고객과 사용 여부만 본 후, Labeling을 진행했다.
- PDDE(유통사)와 COP_U(제휴사)의 관점으로 나누어 데이터를 설명할 수 있는 Feature를 생성하였다.

※ 외부 데이터

- 기온, 강수량, 대기지수, 습도 데이터를 활용했으며 기상청에서 제공하는 데이터 셋이다.
- 공휴일 데이터를 활용했으며 직접 수집을 진행해 생성한 데이터 셋이다.

02 데이터셋 설명 : features

PDDE 관점

- 상품 중분류 구매 다양성
- 주구매 상품
- 채널 다양성
- 채널별 사용 횟수
- 최다 사용 채널
- 유통사 이용 다양성
- 주 사용 유통사
- 구매 상품 다양성

COP_U 관점

- 채널 다양성
- 채널별 사용 횟수
- 제휴사 이용 다양성
- 주 사용 제휴사
- 구매 금액이 가장 큰 제휴사

02 데이터셋 설명 : features

BR(점포) 관점

- 주구매 지점
- 주구매 지역

L.PAY 관점

- 오프라인 핫 타임 구매수량
- 오프라인 핫 타임 구매건수
- 온라인 핫 타임 구매수량
- 온라인 핫 타임 구매건수

외부데이터 관점

- 평균 기온
- 평균 습도
- 평균 강수량
- 평균 대기지수
- 공휴일 때의 이용 횟수

02 데이터셋 설명 : features

통합 관점

- 평균 구매액
- 최대 구매액
- 내점 일수
- 내점 당 구매액
- 내점 당 구매건수
- 월초, 월중반, 월말별 구매 평균 금액
- 월초, 월중만, 월말별 구매 건수
- 구매 주기
- 주말 구매 비율
- 요일별 평균 구매 금액
- 요일별 평균 구매 건수
- 가장 구매를 많이 한 일자
- 가장 구매를 많이 한 시간
- 가장 구매를 많이 한 요일
- 가격이 가장 비싼 상품
- 전체 유통사+제휴사별 이용 횟수
- 유통사+제휴사 총 이용 수
- 거래 횟수
- RFM Score
- 구매 수량

각 세부 데이터 별로 관점을 나눠 Feature를 생성했다. 결측치들은 Numerical Feature의 경우 0으로, Categorical Feature의 경우에는 '정보없음'으로 처리하여 진행하였다.

02 데이터셋 설명 : train test split

본 분석에서는 현재의 고객이 다음에도 L.PAY를 사용하는지 여부와, 어떤 채널에서 사용하는지에 대한 여부를 모두 학습해야 하고, 신규 고객이 들어왔을 때에도 L.PAY 사용 여부와 채널을 판단해야 한다. 그렇기 때문에, 모든 고객들은 거래별로 독립적이라는 가정 하에 분석을 진행하였다.

과거의 데이터를 통해 현재(미래)의 데이터를 예측을 해야 하기 때문에 시간 흐름에 따라 train set과 test set을 나누었다. 주어진 데이터는 2021년 1년치 데이터이기 때문에 계절을 기준으로 4분기로 쪼개어서, 3분기치 데이터를 학습시키고, 남은 마지막 분기 데이터를 예측하고자 했다.

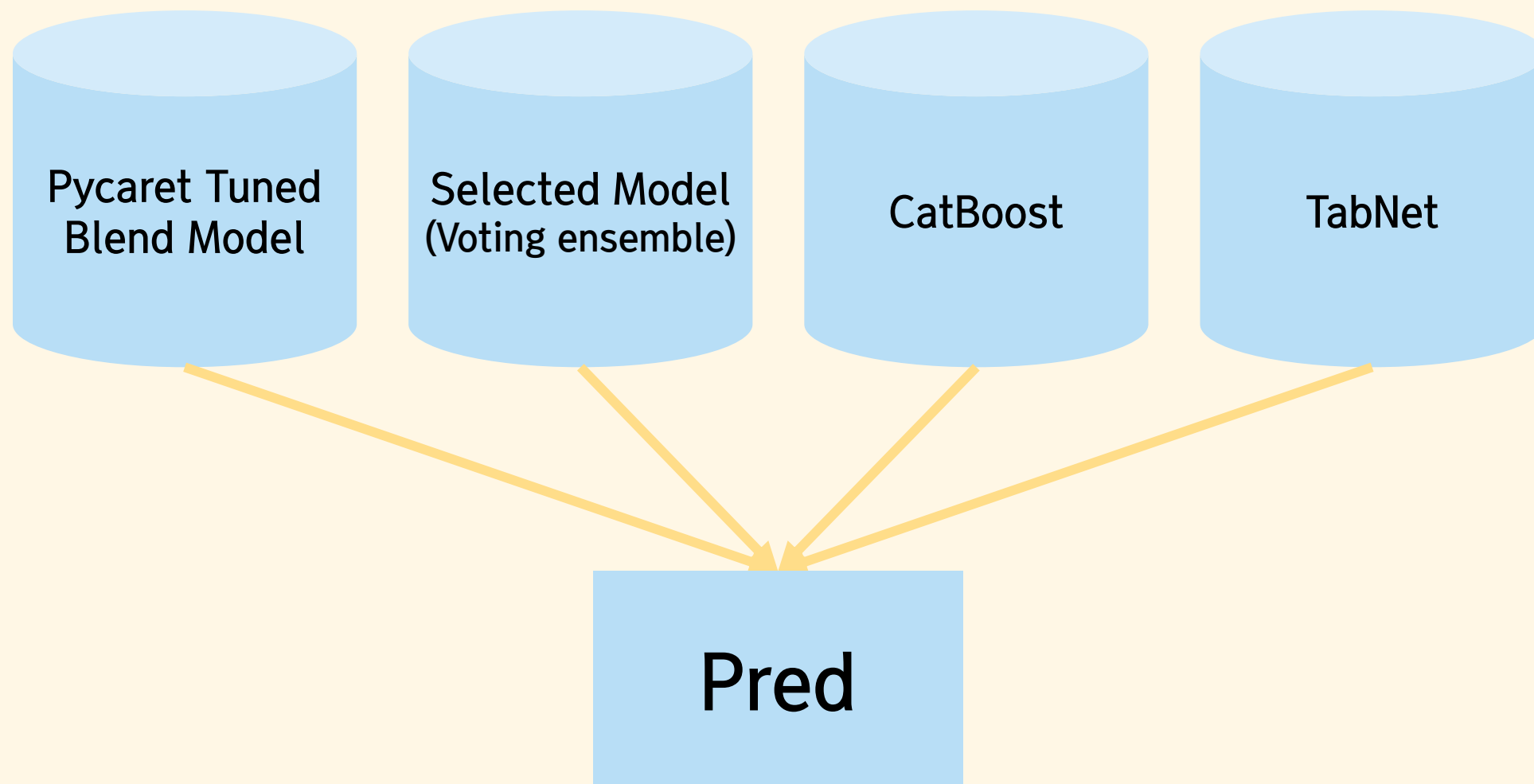
하지만 실험에 다양성을 주기 위해 봄+여름(1+2 분기) 데이터를 train, 가을(3 분기) 데이터를 test, 여름+가을(2+3 분기) 데이터를 train, 겨울(4분기) 데이터를 test로 놓고 실험을 진행하였다.

분기별로 고객의 unique 값이 비슷하기 때문에 학습 시킬 때의 데이터 양의 문제가 존재하기 때문에 test 데이터 셋은 label의 비율을 유지하며 20%만 사용하였다.



03 예측 모델

03 예측 모델



4가지의 서로 다른 모델을 Averaging Ensemble 기법을 활용해 예측값 도출했다.

03 예측 모델

Pycaret Blend Model

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.7201	0.8563	0.4590	0.6840	0.6919	0.4181	0.4282	26.9490
gbc	Gradient Boosting Classifier	0.7189	0.8560	0.4549	0.6822	0.6895	0.4118	0.4233	18.2640
lightgbm	Light Gradient Boosting Machine	0.7155	0.8511	0.4519	0.6776	0.6866	0.4075	0.4178	1.9050
rf	Random Forest Classifier	0.7141	0.8488	0.4147	0.6698	0.6683	0.3680	0.3928	1.0430
xgboost	Extreme Gradient Boosting	0.7143	0.8458	0.4474	0.6764	0.6842	0.4026	0.4136	21.4460
et	Extra Trees Classifier	0.7090	0.8421	0.4010	0.6684	0.6587	0.3466	0.3757	1.3950
lda	Linear Discriminant Analysis	0.6948	0.8273	0.4544	0.6632	0.6740	0.3799	0.3856	2.0260
ada	Ada Boost Classifier	0.7065	0.8188	0.4608	0.6736	0.6843	0.4010	0.4076	1.4870
lr	Logistic Regression	0.6669	0.7636	0.3066	0.5869	0.5755	0.1524	0.2135	13.5690
knn	K Neighbors Classifier	0.6040	0.6814	0.3477	0.5704	0.5842	0.2072	0.2095	3.0880
dt	Decision Tree Classifier	0.6141	0.6640	0.4075	0.6171	0.6154	0.2844	0.2845	0.7680
nb	Naive Bayes	0.5007	0.6636	0.3272	0.5590	0.4892	0.1469	0.1607	0.1630
qda	Quadratic Discriminant Analysis	0.1081	0.5045	0.2590	0.4778	0.0931	0.0054	0.0129	0.7970
dummy	Dummy Classifier	0.6464	0.5000	0.2500	0.4178	0.5076	0.0000	0.0000	0.1450
svm	SVM - Linear Kernel	0.5393	0.0000	0.3131	0.5590	0.4778	0.1408	0.1807	1.3530
ridge	Ridge Classifier	0.6886	0.0000	0.3755	0.6312	0.6342	0.2936	0.3248	0.6510

03 예측 모델



Pycaret Blend Model

- 적은 코드로 머신러닝 워크 플로우를 자동화하는 라이브러리이다.
- 전처리, 모델 선택, 파라미터 튜닝 작업을 자동화해준다.
- AUC를 기준으로 상위 5개를 가지고 Blend를 진행해준다.
- Pycaret 자체로 Tuning을 진행해준다.
- 본 실험에서는 GBM, LGBM, RF, XGB, CAT를 사용해 Blend한 모델을 사용해 예측을 진행했다.

- **평가지표 AUC**

- ROC 커브 아랫부분의 면적이며 다중 분류에서 사용할 경우, 각 클래스별 AUC 값의 평균으로 계산함
- Scikit-learn의 Roc_auc_score에서 다중 분류를 지원해줌
- One-versus-rest 방식을 사용함

03 예측 모델

**Selected
Model
(Voting
ensemble)**

- Auto-ML 기법에서 사용한 AUC가 높은 모델과, 트리 계열 모델인 ExtraTrees Classifier 모델을 선정하여 성능을 파악했다.
- 성능이 높은 모델을 선정해 하이퍼 파라미터 Tuning을 진행했다.
- Tuning 후 성능이 높은 모델끼리 Voting Ensemble을 진행해 최종 값을 도출한다.
- 범주형 변수를 One-Hot 인코딩 방법을 통해 인코딩을 진행해준 후, PCA를 사용해 전처리를 해준 후 사용하였다.
- ExtraTrees Classifier
 - 랜덤 포레스트와 유사하지만 더 극단적으로 랜덤하게 만든 모델
 - 붓스트랩을 하지 않고 전체 원 데이터를 그대로 가져다 쓰며, 배깅을 사용하지 않고, 비복원 추출을 하는 방식
- LightGBM Classifier (AUC 높은 모델)
 - Gradient Boosting 프레임 워크로 트리 기반 학습 알고리즘
 - 트리 기반의 모델이 수평적으로 확장되는 반면, 수직적으로 확장되는 특징이 존재하며 시간이 단축됨

03 예측 모델

A blue cylinder with the text "Catboost" inside.

Catboost

- Pycaret에서 성능이 가장 높게 나왔던 CatBoost를 앞선 두 모델에 사용하지 않고 따로 이질적으로 사용한다.
- XGB와 비슷하게 Level-wise로 트리를 만들어나가는 알고리즘이다.
- 기존 부스팅 모델과 다르게 일부의 데이터를 가지고만 잔차 계산을 하고 모델을 만든 뒤 예측값을 사용한다는 차별점이 존재한다.

A blue cylinder with the text "TabNet" inside.

TabNet

- 고성능이고 해석이 가능한 딥러닝 정형 데이터 네트워크로 전통적인 머신러닝 모델이 아니라는 점에서 차별점이 존재한다.
- 범주형 변수는 라벨 인코딩을 사용하여 전처리를 수행했고, 그 외 전처리와 Feature Selection은 모델 내에서 자동으로 수행된다.

03 예측 모델: 성능

1, 2 분기 학습 3분기 예측

	Pycaret	Voting Model	Catboost	TabNet
AUC	0.78	0.80	0.80	0.76
Ensemble				
AUC	0.80			

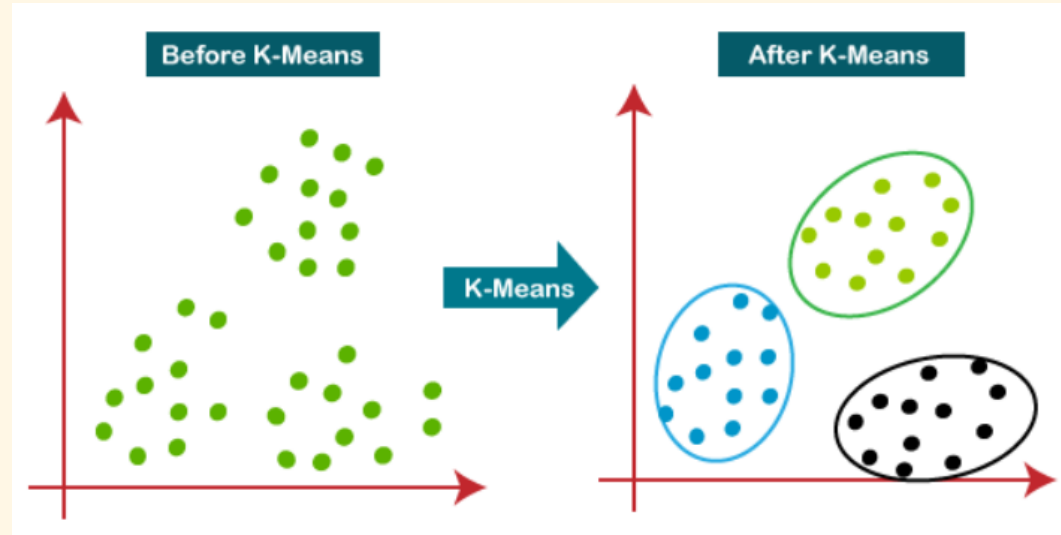
2, 3 분기 학습 4분기 예측

	Pycaret	Voting Model	Catboost	TabNet
AUC	0.79	0.78	0.79	0.75
Ensemble				
AUC	0.79			



04 클러스터링

04 클러스터링



고객별 세분화를 위해서는 가장 널리 알려진 방법인 클러스터링 기법을 활용한다. 클러스터링 기법은 집단 간 정보와 분류 규칙 없이도 개체들의 다양한 특성 관계를 기반으로 이들을 유사 집단으로 분류할 수 있기 때문에 주어진 데이터들 사이에서 의미 있는 자료 구조를 찾을 수 있다.

따라서 개체 별 특성을 통해 클러스터를 형성하고 라벨링 함으로써 기존에는 파악할 수 없었던 소비자들의 특징을 파악하고 개인화된 마케팅을 제시할 수 있을 것이라 예상한다.

클러스터링 시 가장 중요한 요소는 클러스터링을 진행할 변수를 지정하는 것이다. 개인의 경험에 기반한 효과적인 클러스터링 및 세그멘테이션을 위해 앞선 분석에서 사용한 Classification 모델에서 사용된 Feature들 중 4가지 관점에 맞추어 사용할 변수를 선택하였다.

04 클러스터링

	사용 Features	설명
유통사 경험 관련 관점	<ul style="list-style-type: none"> - sum_use_pdde - mpd_nunique - Favorite_pd - most_expensive - 'A01','A02','A03','A04','A05','A06' 	<ul style="list-style-type: none"> - 유통사 제휴사 총 이용수 - 상품 구매 다양성 - 중분류 주구매 상품 - 구매상품 중 비싼 상품 - 해당 유통사 이용 수
제휴사 경험 관련 관점	<ul style="list-style-type: none"> - cop_u_cop_c_most_use - sum_use_cop_u - 'B01','C01','C02','D01','D02','E01' 	<ul style="list-style-type: none"> - 주 사용 제휴사 - 제휴사 총 이용수 - 해당 제휴사 이용 수
날씨 관련 관점	<ul style="list-style-type: none"> - mean_temperature - mean_humidity - mean_rainfall - mean_atmosphere 	<ul style="list-style-type: none"> - 온도 평균 - 습도 평균 - 강수량 평균 - 대기지표 평균
날짜 관련 관점	<ul style="list-style-type: none"> - Sun_count ~ - Sat_count - most_buying_weekday 	<ul style="list-style-type: none"> - 월요일 ~ 일요일 구매 횟수 - 가장 많이 이용한 날짜

유통사 경험 관련 관점, 제휴사 경험 관련 관점, 날씨, 날짜 관련 관점 4가지 관점의 30개 Features에 대해 클러스터링을 진행하였다.

04 클러스터링

알고리즘을 통해 클러스터링을 진행하기 이전에 관점 별 Features에 대해 분포를 살펴보고자 하였다.

3차원 이상의 변수들에 대해 시각화를 진행할 수는 없기 때문에 **고차원 분포를 시각화하기 위한 차원축소 방법을 활용**하였다.

대표적인 차원 축소 방법인 PCA, TSNE, UMAP의 개념 및 특징은 다음과 같다.

	PCA	TSNE	UMAP
개 념	<ul style="list-style-type: none">- 분산이 최대인 축을 찾고 이 축과 직교하며 분산이 최대인 두번째 축을 찾아 투영	<ul style="list-style-type: none">- 고차원 벡터의 유사성이 저차원에서도 유사도록 보존- T분포를 사용한 거리 및 유사도 계산	<ul style="list-style-type: none">- Global Structure를 더 잘 보존- TSNE보다 탄탄한 이론적 배경
특 징	<ul style="list-style-type: none">- 선형 방식으로 정사영- 군집 데이터들이 뭉개짐	<ul style="list-style-type: none">- 데이터 개수가 n개라면 연산량이 n제곱만큼 늘어남	<ul style="list-style-type: none">- 빠른 속도- Embedding 차원 크기에 대한 제한이 없음

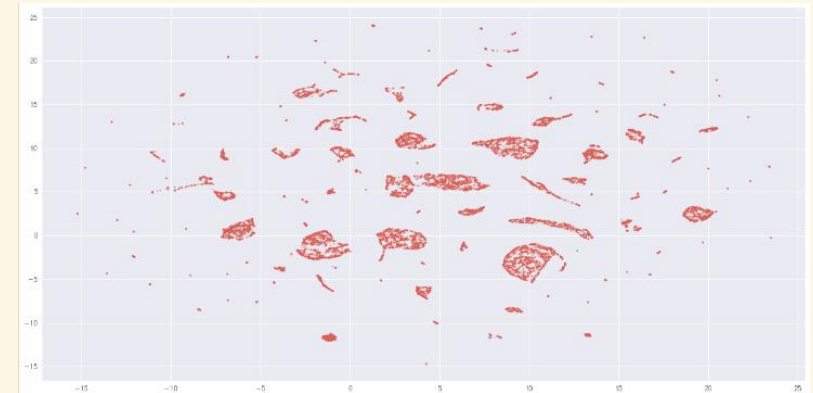
이와 같은 특징을 지닌 3가지 차원축소 방법 중 Global Structure를 잘 보존하면서 많은 양의 데이터에 대해 빠른 속도로 시각화를 진행할 수 있는 **UMAP 방법을 사용**하였다.

04 클러스터링

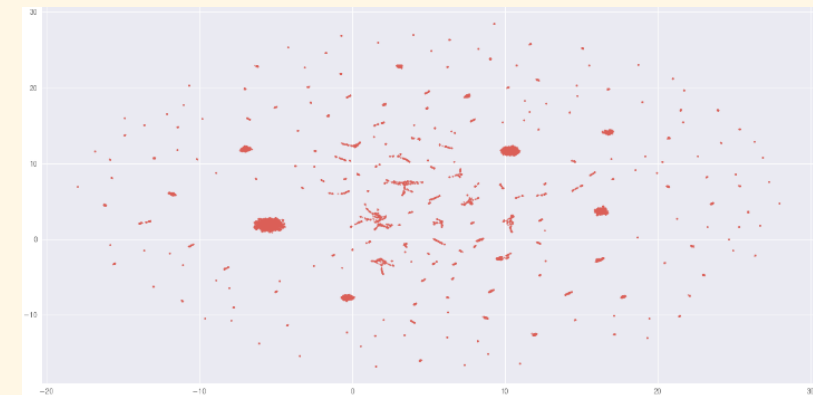
UMAP을 통해 선정한 Feature들의 분포를 2차원 Scatter Plot으로 시각화 한 결과이다. 위 그림부터 유통사 경험 관련, 제휴사 경험 관련, 날짜 관련 관점이다.

Clustering을 진행하기 전이지만 각 개체들이 밀집된 지역들이 어느정도 식별되는 것으로 보인다.

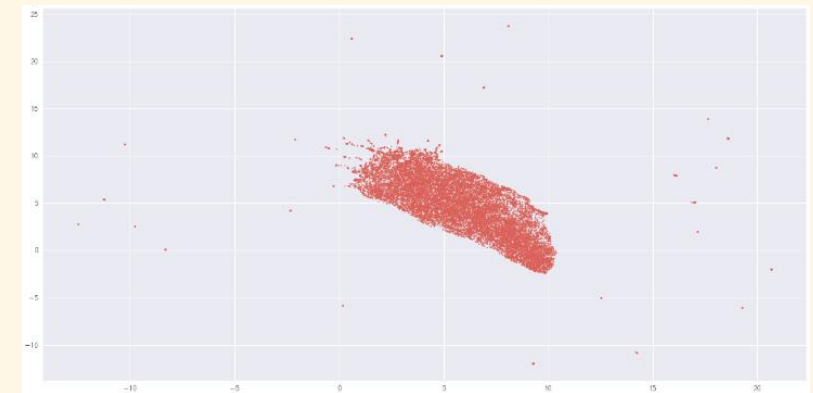
Clustering을 진행한 후 해당 분포를 통해 Clustering이 적절히 되었는지를 육안으로 파악하는 과정을 거칠 예정이다.



<PDDE_UMAP>

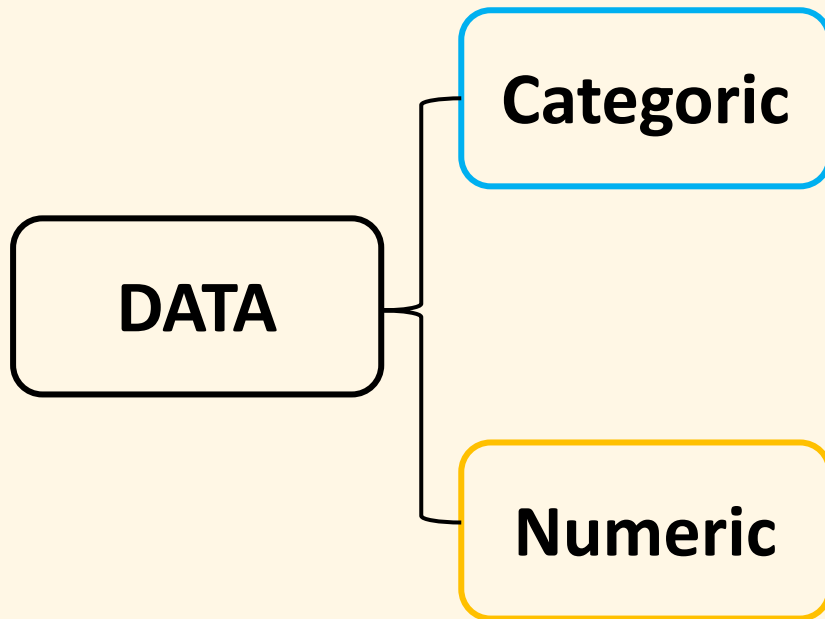


<COP_UMAP>



<WEEK_UMAP>

04 클러스터링



$$d(X_i, Q_l) = \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c)$$

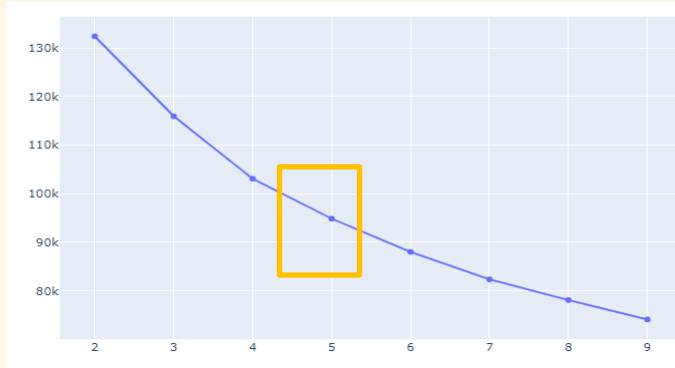
〈K-Prototype 수식〉

클러스터링을 위한 알고리즘으로는 거리기반의 알고리즘인 K-Prototype 알고리즘을 사용했다. K-Prototype은 K-Means와 K-Modes의 개념을 동시에 활용해 연속형과 범주형 자료를 동시에 활용할 수 있는 클러스터링 방식이다.

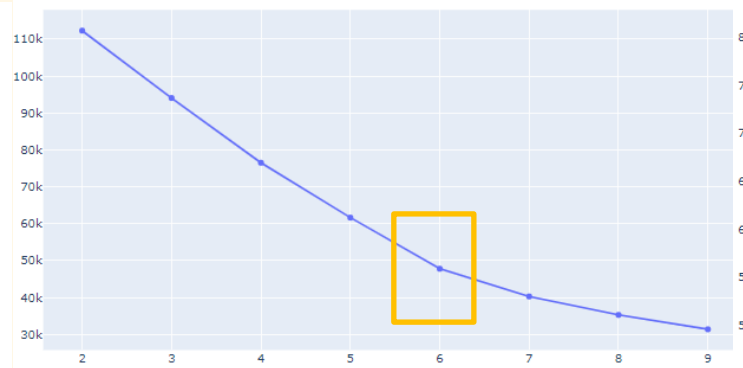
좌측의 수식을 통해 거리가 계산되며, 연속형 자료는 유클리디안 거리를 구하고, 범주형 자료는 비유사도를 구한 다음, 비유사도에 가중치를 부여하여 둘을 합한 것으로 정의한다.

따라서, 연속형 변수와 범주형 변수를 동시에 사용한 **유통사 경험 관련 관점**, **제휴사 경험 관련 관점**, **날짜 관련 관점**에 대해서는 **K-Prototype 방식의 Clustering**을 활용해 Cluster를 구분하였다.

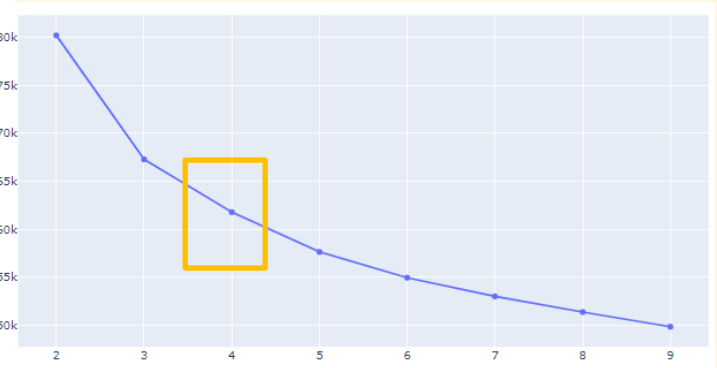
04 클러스터링



〈PDDE_scee plot〉



〈COP_U_scee plot〉



〈WEEK_scee plot〉

K-Prototype의 중요 파라미터인 K의 개수를 정하기 위해 scree plot을 그려, 그래프가 꺾이는 부분에서 K의 개수를 설정했다.

유통사 경험 관련 관점의 경우 Scree Plot이 완만하게 도출되어 중심값인 5를 K의 개수로 설정했다.

제휴사 경험 관련 관점의 경우 전반적으로 6에서 그래프가 완만하게 꺾이는 것을 확인하였다.

날짜 관련 관점의 경우 4에서 그래프가 완만해지는 경향이 있다.

유통사 경험 관련 관점 K = 5

제휴사 경험 관련 관점 K = 6

날짜 관련 관점 K = 4

04 클러스터링

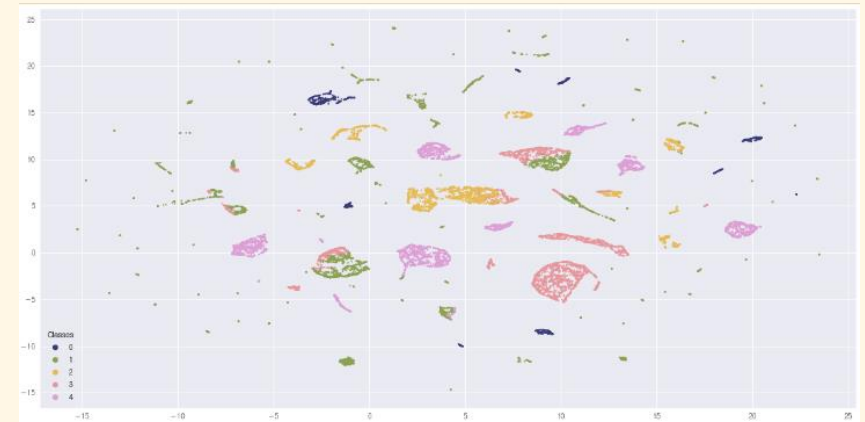
각 관점별로 K-Prototypes를 사용해 Clustering을 진행한 결과를 분포에 적용하여 Cluster가 적절히 형성되었는지 확인하였다.

PDDE의 경우 적절한 개체별로 적절한 밀집도를 가지고 있는 분포가 보이고 Cluster들의 거리도 적당히 분할되어 있는 것으로 보인다.

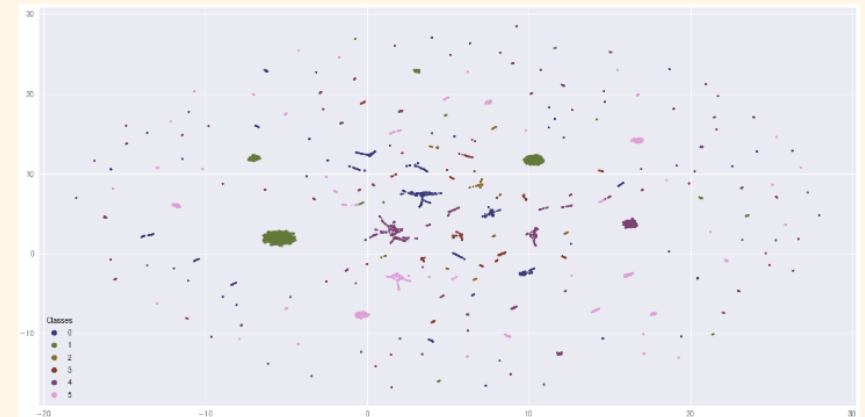
COP_U의 경우 개체들의 밀집도가 상당히 낮으나 뭉쳐 있는 개체들에 대해서는 Cluster가 잘 된 것으로 보인다.

WEEK의 경우 개체들의 밀집도가 굉장히 높은 것으로 보이고 Cluster별 구역을 눈으로 식별할 수 있을 만큼 Cluster가 적절히 나뉘어 있음을 확인했다.

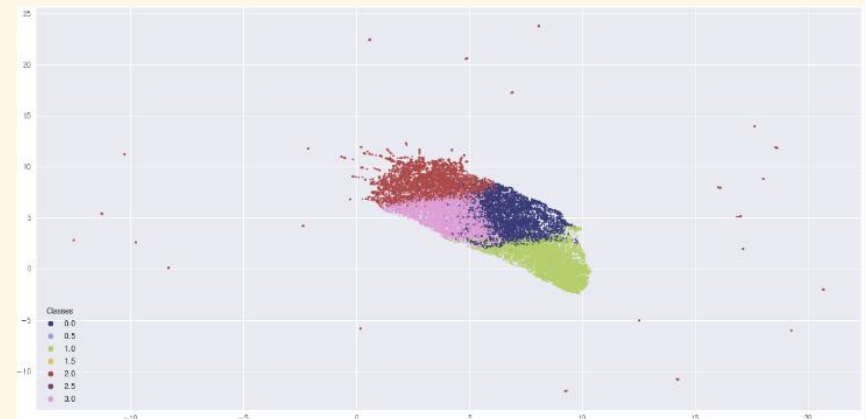
이로써 세 관점 모두 Clustering이 적절히 되었다고 생각한다.



〈PDDE_UMAP with Cluster〉

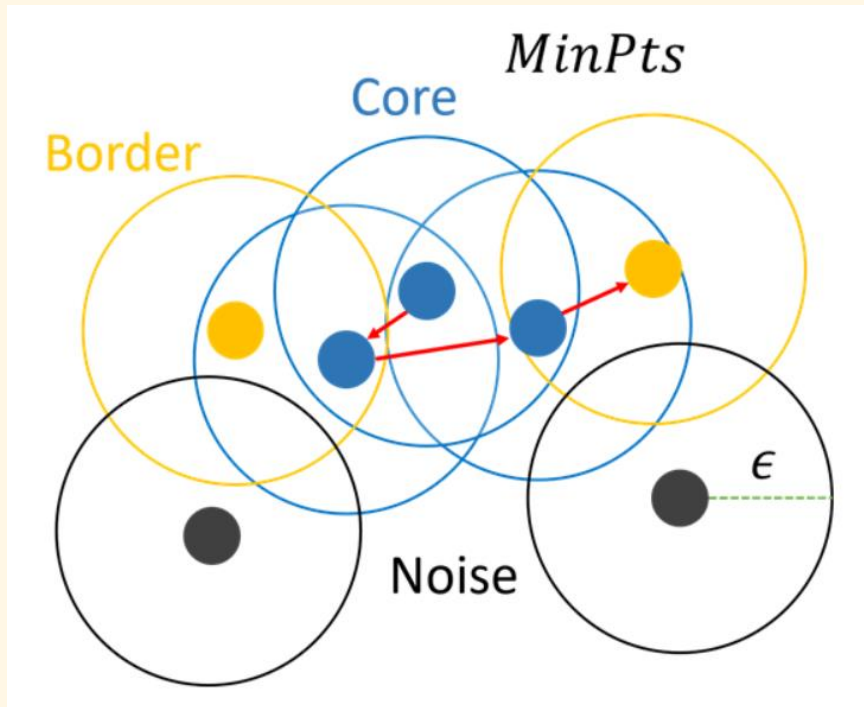


〈COP_U_UMAP with Cluster〉



〈WEEK_UMAP with Cluster〉

04 클러스터링 : DBSCAN



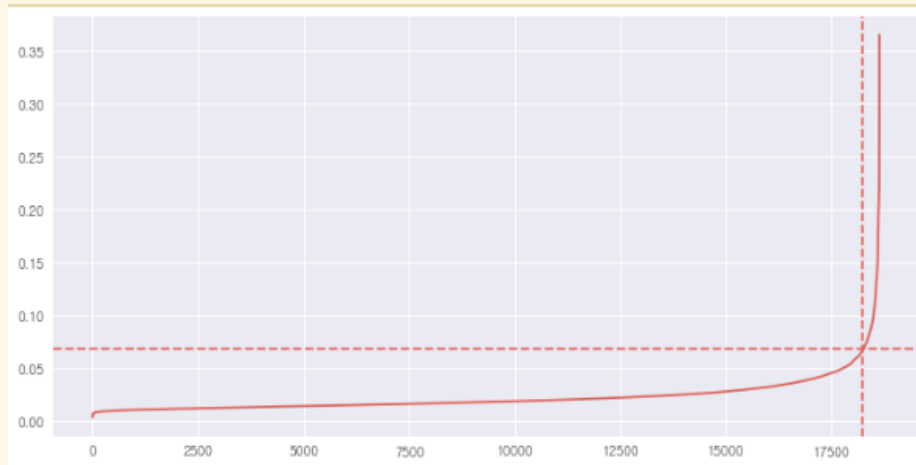
4가지 관점 중 남은 하나인 날씨관련 관점에 대해서는 밀도기반 Clustering 방법 중 하나인 DBSCAN 방법을 사용하였다.

DBSCAN은 클러스터 개수를 미리 지정할 필요가 없으며 이상치를 효과적으로 제외할 수 있는 알고리즘이다.

이를 위해서는 K-means 방식과는 다르게 반경의 크기(ϵ)와 최소 군집의 크기($MinPts$)를 결정해야 한다.

따라서, 시각화를 통해 ϵ 크기를 선정하고 이를 통해 Clustering을 진행했다.

04 클러스터링 : DBSCAN



〈eps 설정을 위한 역 Screeplot〉

- 군집 개수: 이상치 군집 포함 4개
- 실루엣 계수: 0.5154
(0 ~ 1 사이의 스코어로 높을 수록 좋음)

앞서 언급한 DBSCAN의 중요 파라미터인 `eps`와 `min_samples`을 설정하기 위해 역 Screeplot을 그려 값이 급격히 변동하는 부분을 표시하였고 y축에 해당하는 값을 `eps`값으로 설정하고 `min_samples`는 실루엣 계수를 확인하며 휴리스틱하게 설정하였다.

최종적으로 실루엣 계수 0.5154 하에서 군집개수는 이상치 군집을 포함하여 4개, `eps`는 0.068, `min_samples`는 6이라는 값을 얻을 수 있었다.

04 클러스터링 : 고객별 Tag 및 Segmentation

이렇게 구한 관점 별 클러스터를 통해 고객별로 Tag를 달고 세분화를 진행하였다.

유통사 경험 관련 관점 5개 Cluster, 제휴사 경험 관련 관점 6개 Cluster
날짜 관련 관점 4개 Cluster, 날씨 관련 관점 4개 Cluster

이론상으로 480개의 세그먼트를 나누어 고객을 다각도의 관점에서 바라보고 마케팅 방안을 제시할 수 있다.

뿐만 아니라 앞서 구축한 채널 별 L.PAY 사용 여부를 예측하는 모델과 함께 사용하면 훨씬 더 세분화된 마케팅 방안을 제시할 수 있다.

관점 별 클러스터에 대한 설명을 위해 Strip Plot, Box Plot, Bar Plot등을 그리며 시각화를 진행하였으며 실제 EDA 결과는 Appendix를 통해 확인할 수 있다. 해석 시에 사용한 유통사 및 제휴사 관련 내용은 추정을 통해 발견하였다.

(ex. A01 = 백화점, B01 = 롯데 호텔) 이를 통해 파악한 특징은 다음 표와 같다.

04 클러스터링 : 고객별 Tag 및 Segmentation

관점	클러스터 번호	설명	TAG
유통사 경험 관련 관점	Cluster 0	<ul style="list-style-type: none"> - 컴퓨터 주변기기 / 주방가전 등 자주 산 목록 상위권 - A05(하이마트) 쇼핑 존재 	가전 관심군
	Cluster 1	<ul style="list-style-type: none"> - 의류세트 등 구매 수 상위권 - A01(백화점) 쇼핑 비율이 높고 고가 물건 중 패션잡화가 다수 존재 	의류 및 패션 관심군
	Cluster 2	<ul style="list-style-type: none"> - 청소, 생활, 건강식품 등 구매 상위권 - 스포츠, 레저, 헬스 등이 고가물건 상위권에 존재 	건강 및 스포츠 레저 관심군
	Cluster 3	<ul style="list-style-type: none"> - 유아동 의류/ 완구 등이 고가물건 상위권 존재 - A02(마트 or 슈퍼)쇼핑 비율이 높음 	케어 및 아동제품 관심군
	Cluster 4	<ul style="list-style-type: none"> - 가구 / 침구 물품 등이 고가물건 상위권 존재 - A02(마트 or 슈퍼)쇼핑 비율이 높음 	생활 및 하우스 라이프 관심군
제휴사 경험 관련 관점	Cluster 0	<ul style="list-style-type: none"> - C02(롯데월드)와 D01(롯데리아) 이용률이 높은 군집 	여가생활 관심군
	Cluster 1	<ul style="list-style-type: none"> - 제휴사 관련 경험이 잘 없는 군집 - 이용시에 롯데 C01(시네마)나 롯데리아 이용 	제휴사 관심 부족군
	Cluster 2	<ul style="list-style-type: none"> - 제휴사 관련 다양성이 높은 군집 - 렌트 이용이 높은 군집 	렌트 및 제휴사 다양군
	Cluster 3	<ul style="list-style-type: none"> - B01(롯데 호텔)과 D02(엔제리너스) 이용이 높은 군집 	호텔 관심군
	Cluster 4	<ul style="list-style-type: none"> - D01, D02, C01 를 이용한 군집 - F&B와 엔터쪽에 집중 	문화생활 관심군
	Cluster 5	<ul style="list-style-type: none"> - 다양성이 가장 적은 군집 - 롯데리아와 롯데시네마 이용 내역 존재 	F&B 관심군

04 클러스터링 : 고객별 Tag 및 Segmentation

관점	클러스터 번호	설명	TAG
날씨 관련 관점	Cluster -1	- 날씨에 대한 영향을 파악하기 쉽지 않은 군집	-
	Cluster 0	- 비가 오지 않을 때 주로 쇼핑하는 군집	맑은날 선호군
	Cluster 1	- 기온이 상대적으로 낮고 비가 와도 쇼핑하는 군집	강수량 비영향군
	Cluster 2	- 온도, 습도, 강수에는 쇼핑을 하나 미세먼지가 높으면 하지 않는 군집	미세먼지 악영향군
날짜 관련 관점	Cluster 0	주중 구매비율이 주말보다 높은 군집	주중 선호군
	Cluster 1	다른 날의 구매보다 토요일의 구매 수가 높은 군집	토요일 선호군
	Cluster 2	평일과 주말 구매 수가 비슷한 군집	-
	Cluster 3	주말 구매비율이 월등히 높은 군집	주말 선호군

<Cluster 별 Tagging 예시>

	cust	cluster_pdde	clusters_cop	clusters_week	clusters_weather
542	M657144966	의류 및 패션 관심군	여가생활 관심군	주말 선호군	맑은날 선호군
543	M795653145	가전 관심군	여가생활 관심군	주말 선호군	맑은날 선호군
544	M280112960	생활 및 하우스 라이프 관심군	제휴사 관심 부족군	토요일 선호군	맑은날 선호군
545	M310879031	건강 및 스포츠 레저 관심군	F&B 관심군	주말 선호군	맑은날 선호군
546	M896369103	의류 및 패션 관심군	여가생활 관심군	주말 선호군	맑은날 선호군



05 마케팅 제안 및 결론

05 마케팅 제안 및 결론

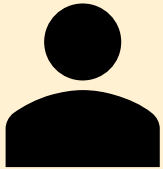
앞서 구축한 채널 별 L.PAY 이용 여부 예측 모델과 Clustering 모델을 통해 생성한 고객별 Tag를 이용하여 각 고객에 대해 L.PAY 사용을 넋지 할 수 있는 마케팅 방안을 제시하고자 한다.

넋지 마케팅⁵을 시행한 기업에서는 장바구니 고객 공략, SMS 대신 넋지를 사용한 마케팅, 타임세일 등을 도입한 3주 만에 구매 전환율 20배 증가, 매출 44.7% 증가라는 큰 성과를 보였다. 이처럼 소비자가 다음 분기에 어떠한 채널을 통해 L.PAY를 사용할지 혹은 L.PAY를 사용하지 않을 지에 대해 예측하고 사용여부와 채널에 따라 소비자에게 접근하는 방식을 채택한 후 소비자의 태그에 따라 구매 경험을 제안하여 성과를 내보고자 한다.

05 마케팅 제안 및 결론:마케팅프로세스

〈step1〉

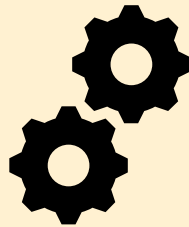
Input
Data



L.PAY와 연동되지 않은
고객 경험 데이터
(구매 경험 or 제휴사
이용 경험)

〈step2〉

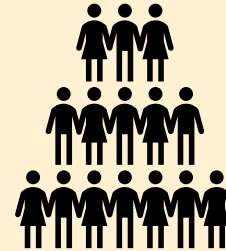
Classification
Model



해당 고객 경험 데이터를
Input으로 사용해
L.PAY를 이용할지 아닐지,
이용한다면 어떠한
채널에서 이용할지 예측

〈step3〉

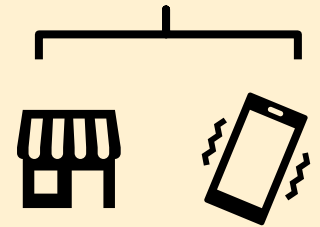
Clustering
Model



step 1의 고객 경험
데이터를 Input으로
사용해 고객 경험에 대한
tag를 삽입

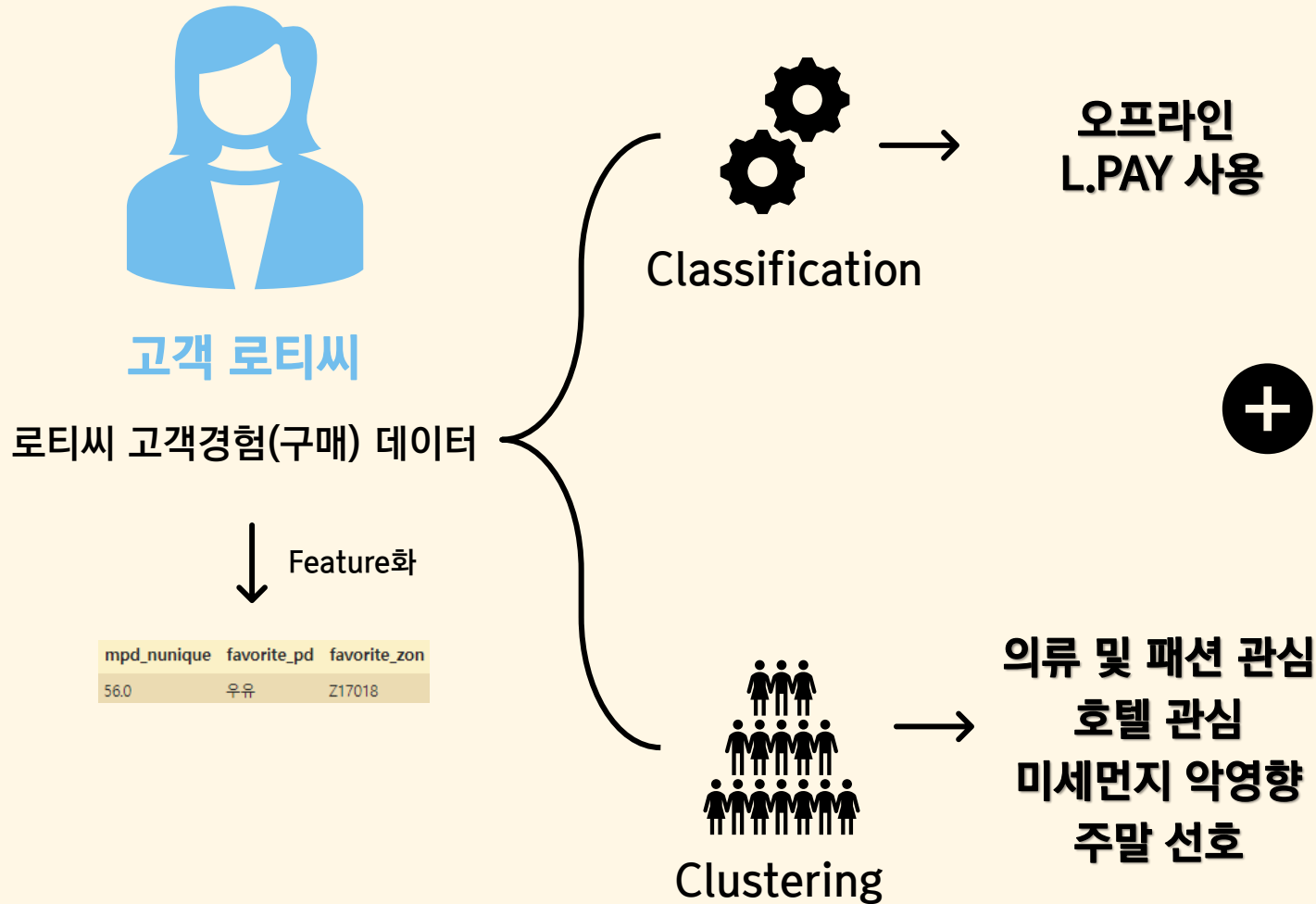
〈step4〉

Nudge
Marketing



예측된 L.PAY 사용 여부와
채널을 통해 온라인 이면
온라인 / 오프라인이면
오프라인 / 온오프라인을
동시에 사용한다면
옴니채널로 고객에게
L.PAY 사용을 넋지

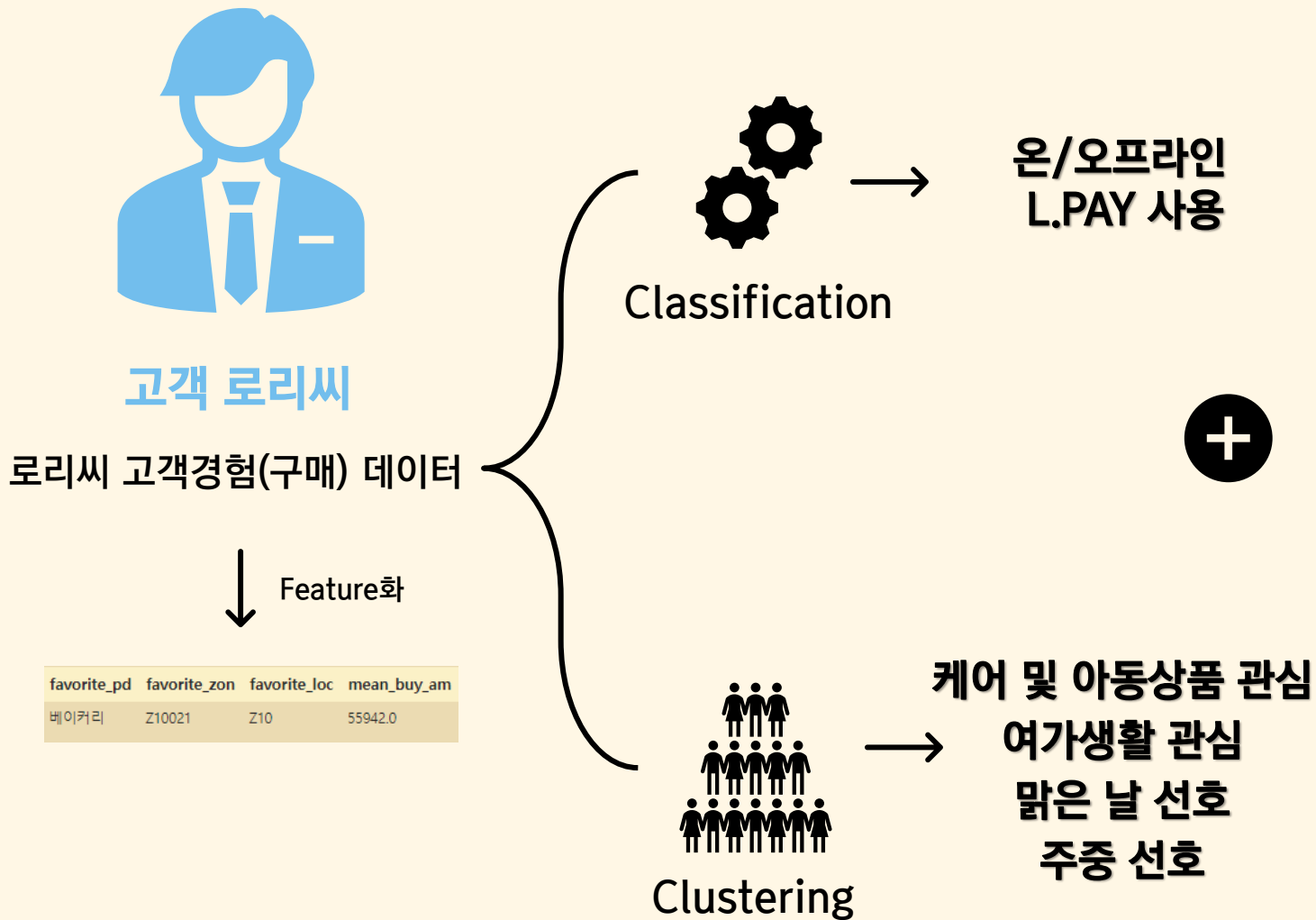
05 마케팅 제안 및 결론:마케팅예시



고객 로티씨의 경우 예측 모델을 확인한 결과 오프라인에서 L.PAY를 사용한다고 예측되었다.

미세먼지가 없는 주말에 백화점 의류 쇼핑몰에서 L.PAY를 사용할 시에 적용할 수 있는 할인쿠폰을 제공하거나 L.PAY 사용 시 뷰가 좋은 호텔 룸을 배정받을 수 있도록 L.PAY 사용을 위한 넛지 마케팅을 진행할 수 있다.

05 마케팅 제안 및 결론:마케팅예시



고객 로리씨의 경우 예측 모델을 확인한 결과 온/오프라인에서 L.PAY를 사용한다고 예측되었다. 해당고객은 온/오프라인에서 L.PAY를 일관적으로 사용할 수 있도록 옴니채널 마케팅이 필요해 보인다.

예를 들어 L.PAY를 사용해서 온라인에 여가생활 티켓을 구매할 시 아동할인을 적용할 수 있고, L.PAY를 사용해 구매한 티켓 소지 시 여가활동 장소에 소재한 점포 이용 시 추가적인 혜택을 부여할 수 있다.

05 마케팅 제안 및 결론 :모델의장점및기대효과

채널 별 예측

온라인, 오프라인 뿐 아니라 온/오프라인 채널을 예측하고 고객의 경험에 맞추어 마케팅을 실시하기 때문에 옴니채널 마케팅 등 다양한 관점의 접근이 가능하고, 이를 통해 L.PAY 사용율을 늘리는데 다양하게 기여할 수 있을 것이라 예상한다.

관점 별 Tagging

관점 별 Tagging 방법을 선택하여 Clustering 방법을 사용 했음에도 개인별로 더 세분화된 마케팅을 제공할 수 있다.

범용성 확보

고객 구매경험에 기반한 데이터를 사용했기 때문에 같은 형식의 구매경험 데이터를 확보할 수 있다면 잠재 고객에 대해 L.PAY 사용 여부 및 채널을 예측할 수 있다.

참고 자료

[목차 그림 출처] <https://www.ktnews.com/news/articleView.html?idxno=120491>

[1] 간편 결제 서비스

<https://terms.naver.com/entry.naver?docId=3596828&cid=42346&categoryId=42346>

[2] 2017 ~ 2020 국내 간편결제서비스 이용 추세

<https://blog.naver.com/tamspay/222377301859>

[3] L.PAY 브랜드 평판 순위

http://www.thebigdata.co.kr/view.php?ud=20220729084826513907d270612f_23

[4] 넛지 마케팅

<https://terms.naver.com/entry.naver?docId=300312&cid=43665&categoryId=43665>

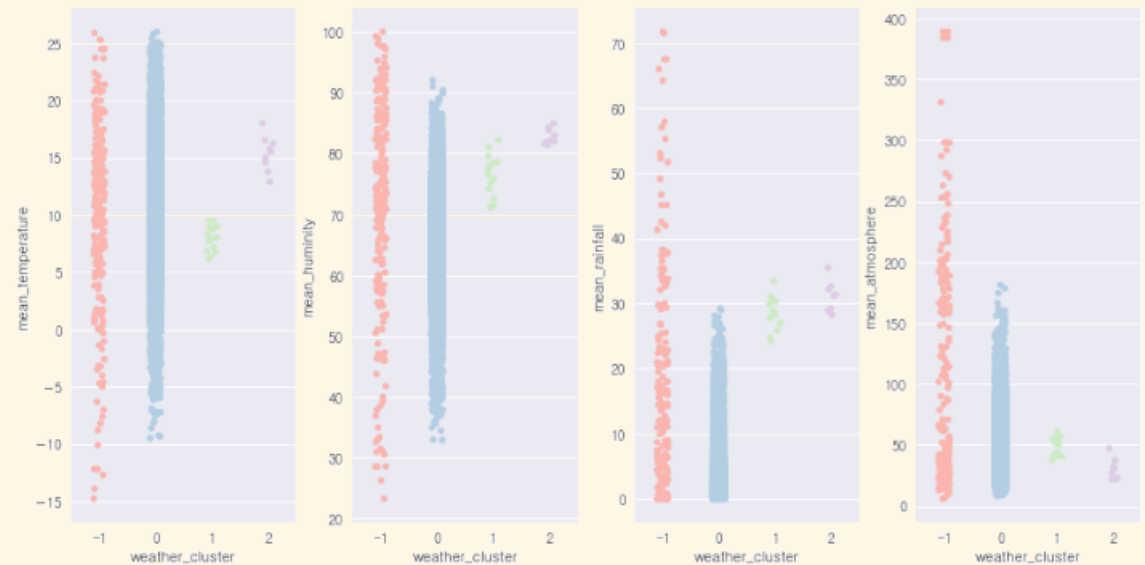
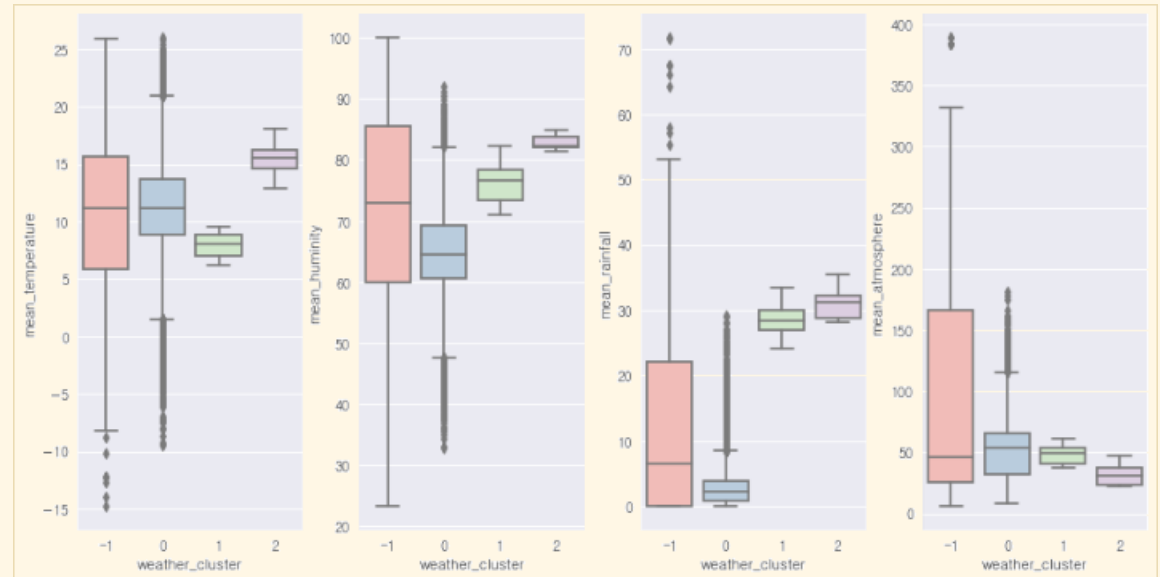
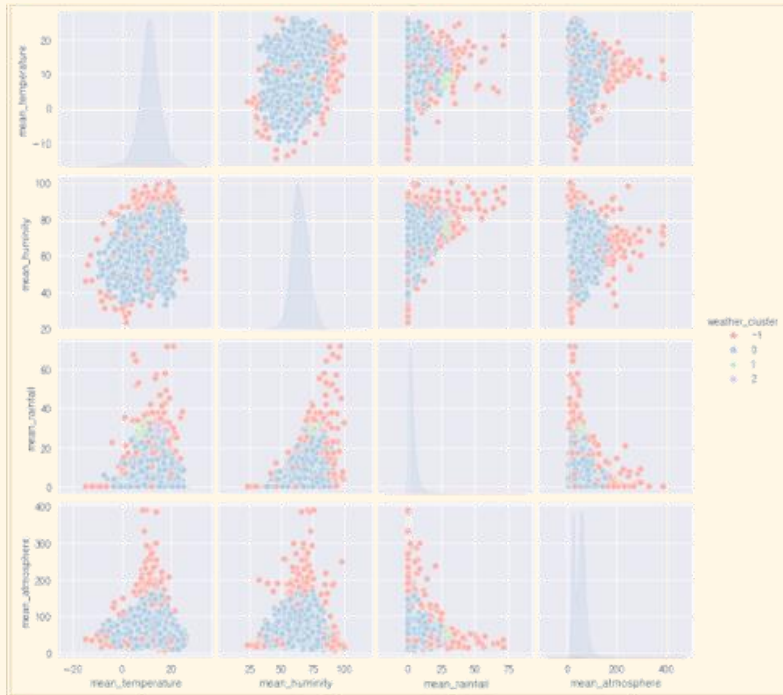
[5] 넛지 마케팅 예시 <https://channel.io/ko/blog/nudge-ec>

[결론 그림 출처]

<https://adventure.lotteworld.com/kor/enjoy/performance/character/contentsid/398/index.do>

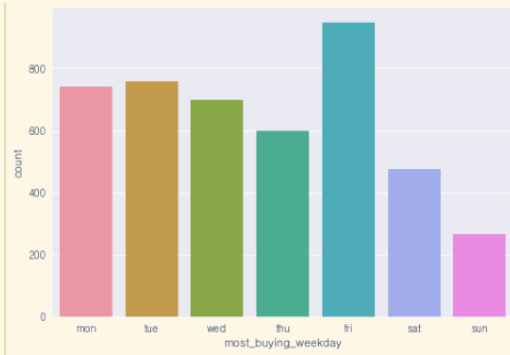
Appendix : 관점별 태그 생성을 위한 시각화

〈날씨 관점〉

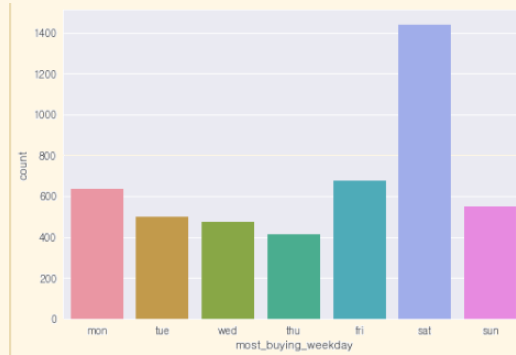


Appendix : 관점별 태그 생성을 위한 시각화

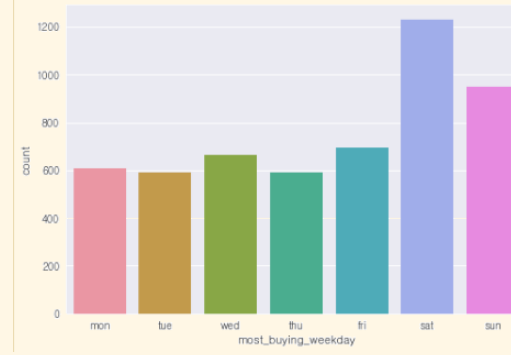
〈날짜 관점〉



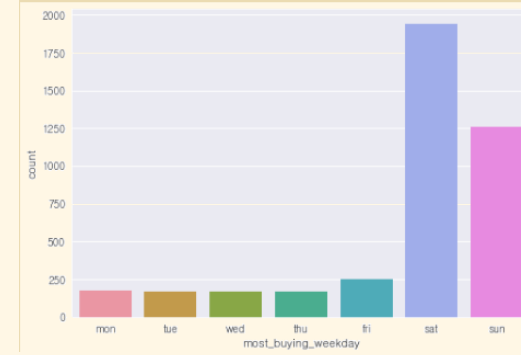
〈0번 Cluster〉



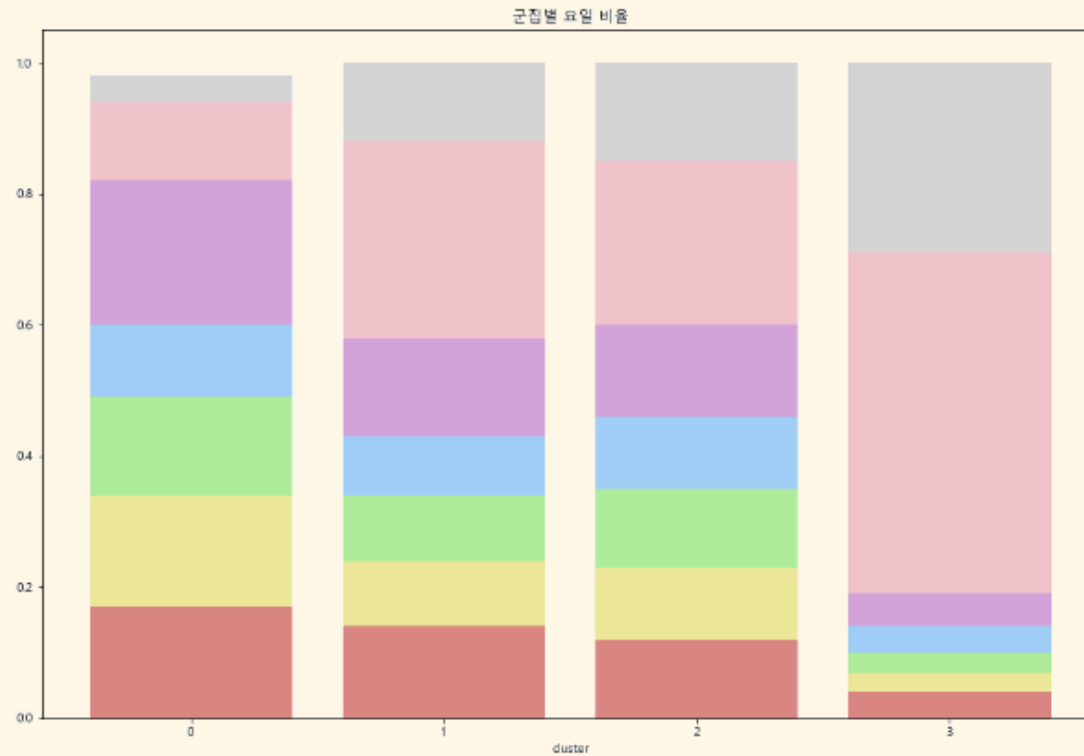
〈1번 Cluster〉



〈2번 Cluster〉



〈3번 Cluster〉



Appendix : 관점별 태그 생성을 위한 시각화

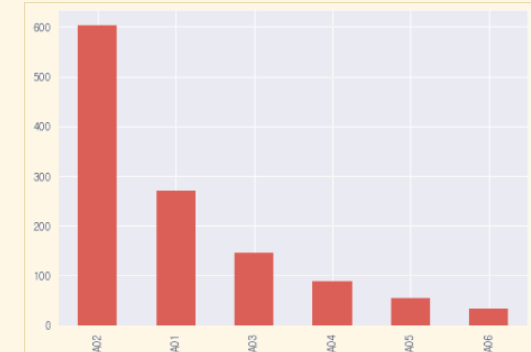
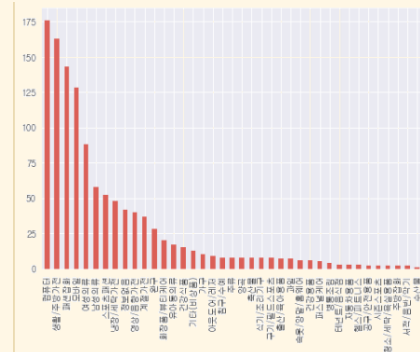
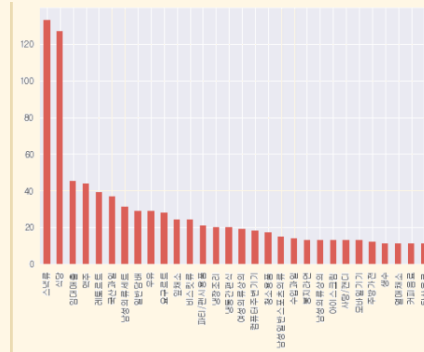
〈유통사 경험 관점〉

가장 많이 산 물품

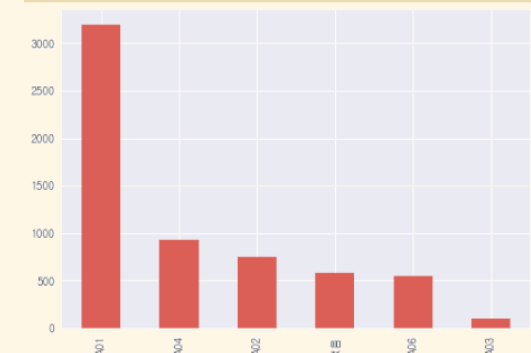
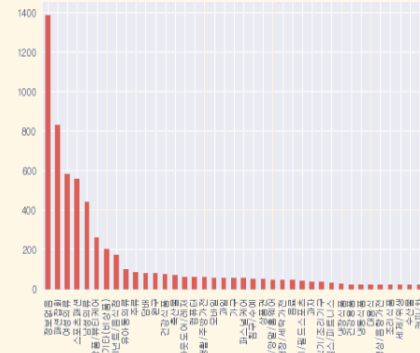
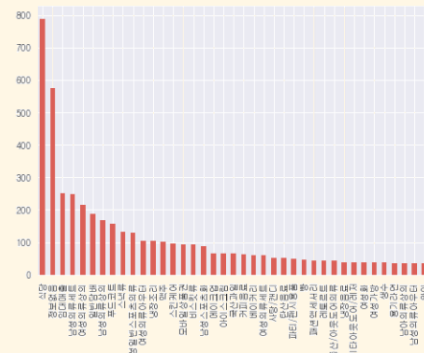
산 것 중 비싼 물품

가장 이용한 유통사

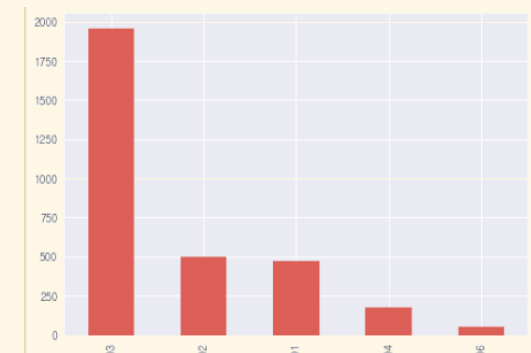
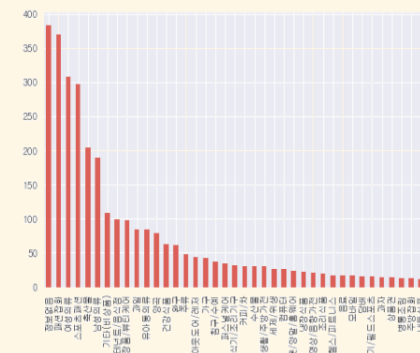
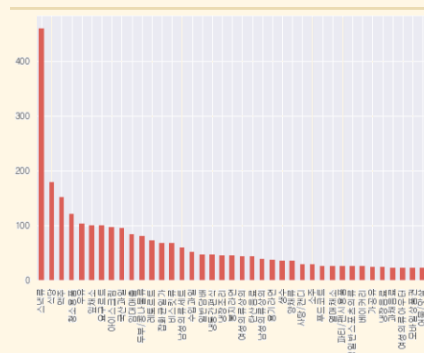
〈0번 Cluster〉



〈1번 Cluster〉



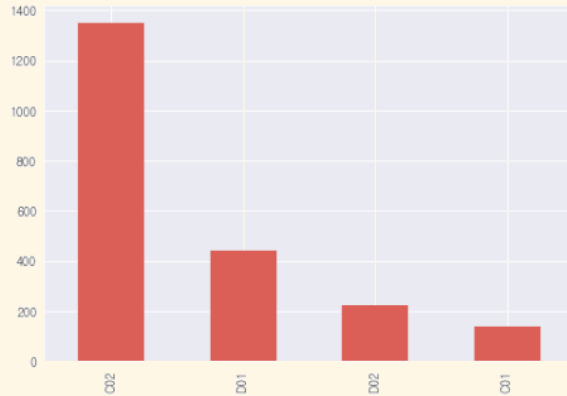
〈2번 Cluster〉



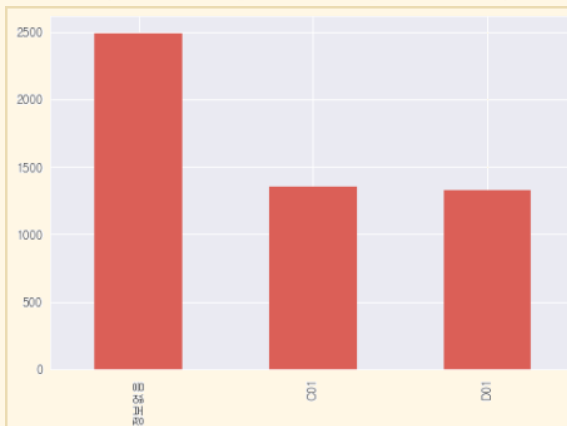
Appendix : 관점별 태그 생성을 위한 시각화

〈제휴사 경험 관점 : 유통사 이용 수〉

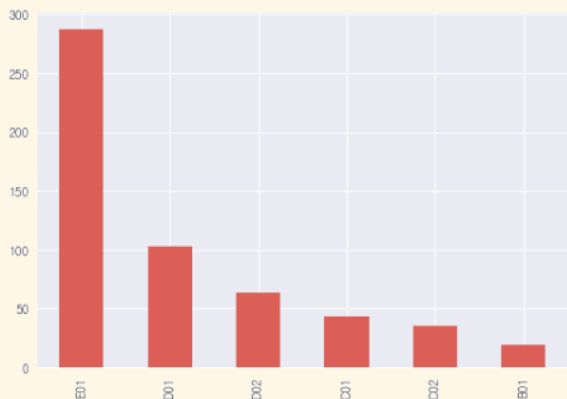
〈0번 Cluster〉



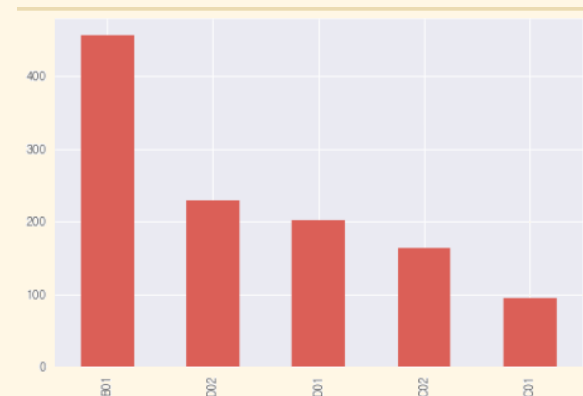
〈1번 Cluster〉



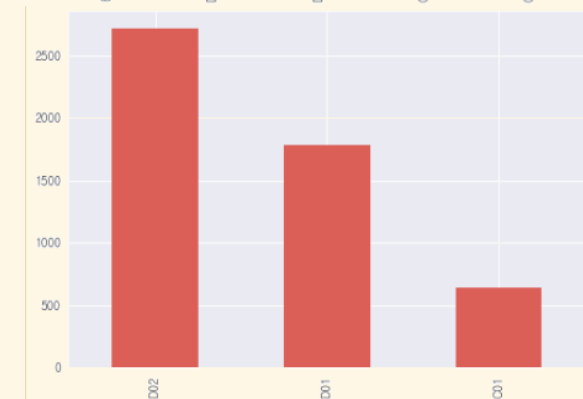
〈2번 Cluster〉



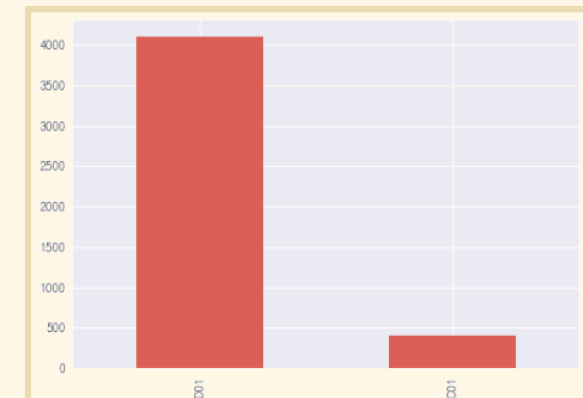
〈3번 Cluster〉



〈4번 Cluster〉

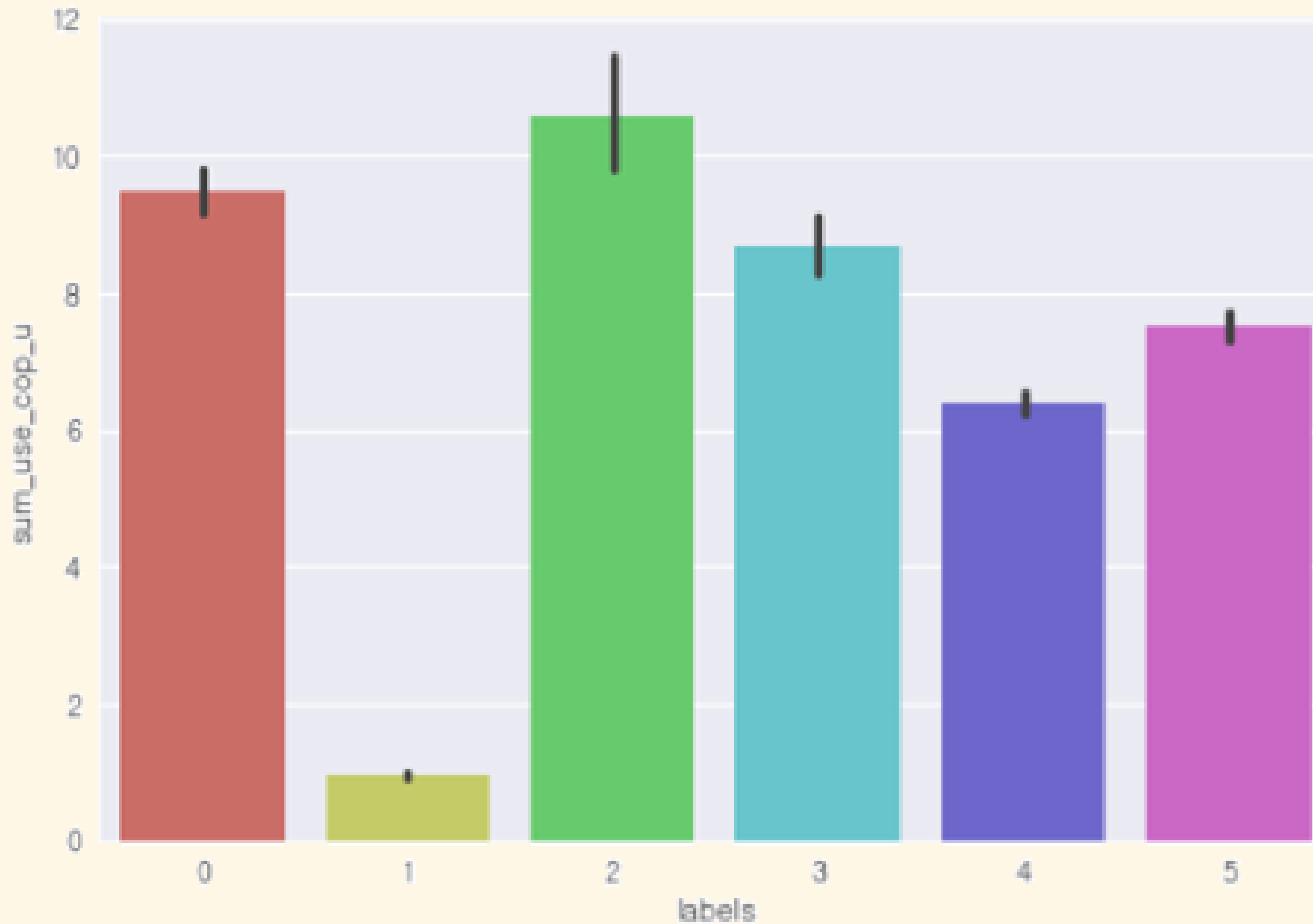


〈5번 Cluster〉



Appendix : 관점별 태그 생성을 위한 시각화

〈제휴사 경험 관점 : 클러스터 별 제휴사 이용 수〉





감사합니다

Thank You