Clustering을 통한 잠재고객 탐색 & 오프라인 팝업스토어 활용방안

Team : 롯데 감바스

CONTENTS

01 주제 선정 배경

Check point

잠재 고객 세분화

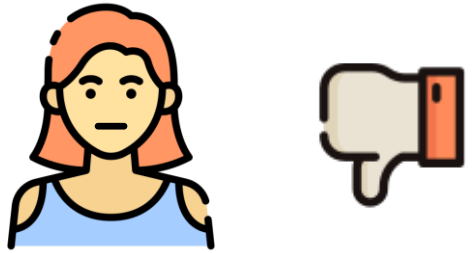
타겟 고객으로의 전환

타겟 고객 맞춤 상품 예측

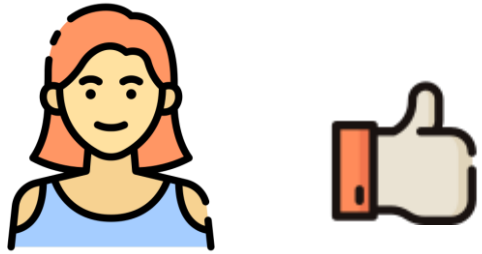
구매 고객으로의 전환

01 주제 선정 배경

Random



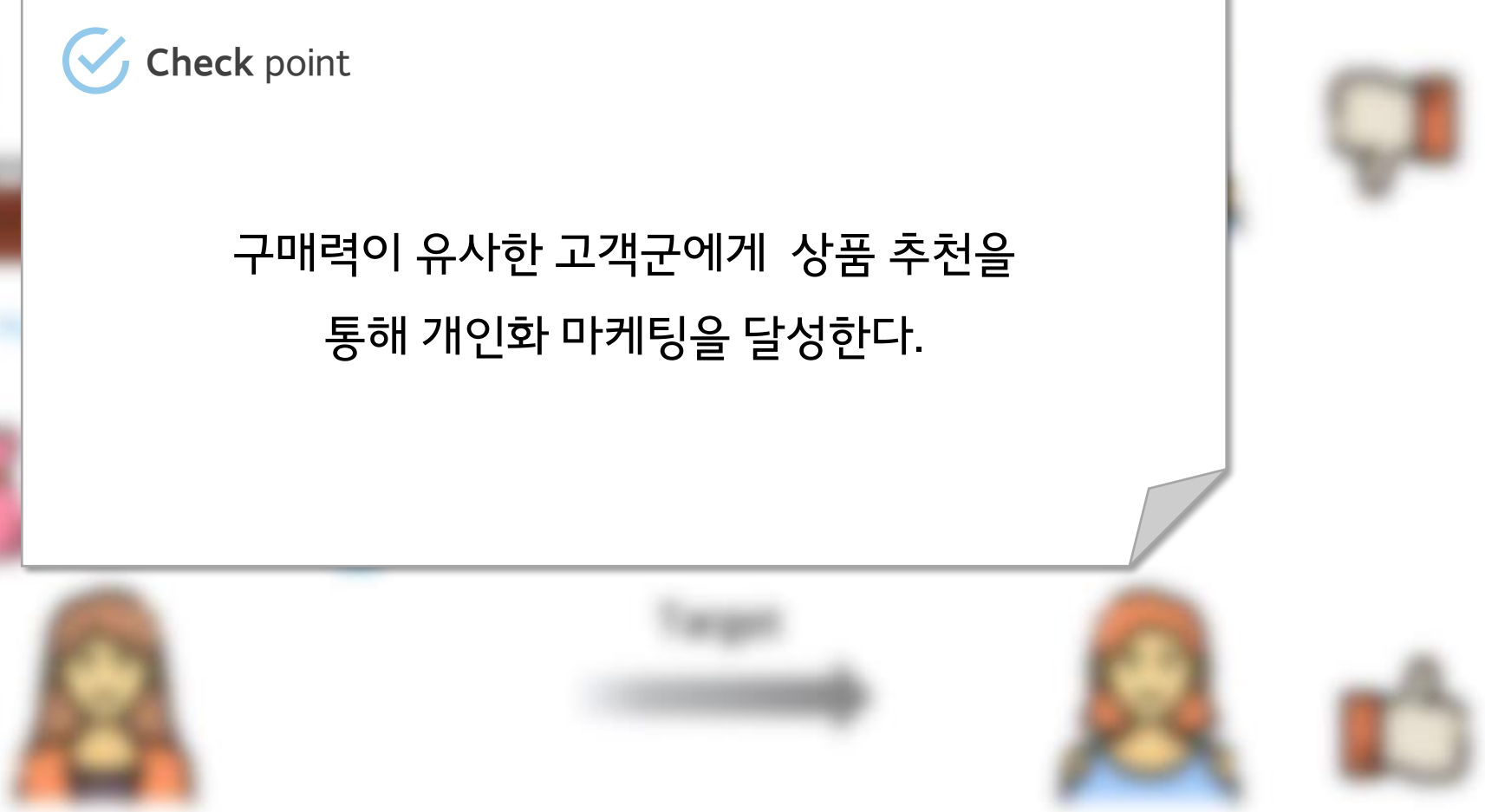
Target



Check point

구매력이 유사한 고객군에게 상품 추천을

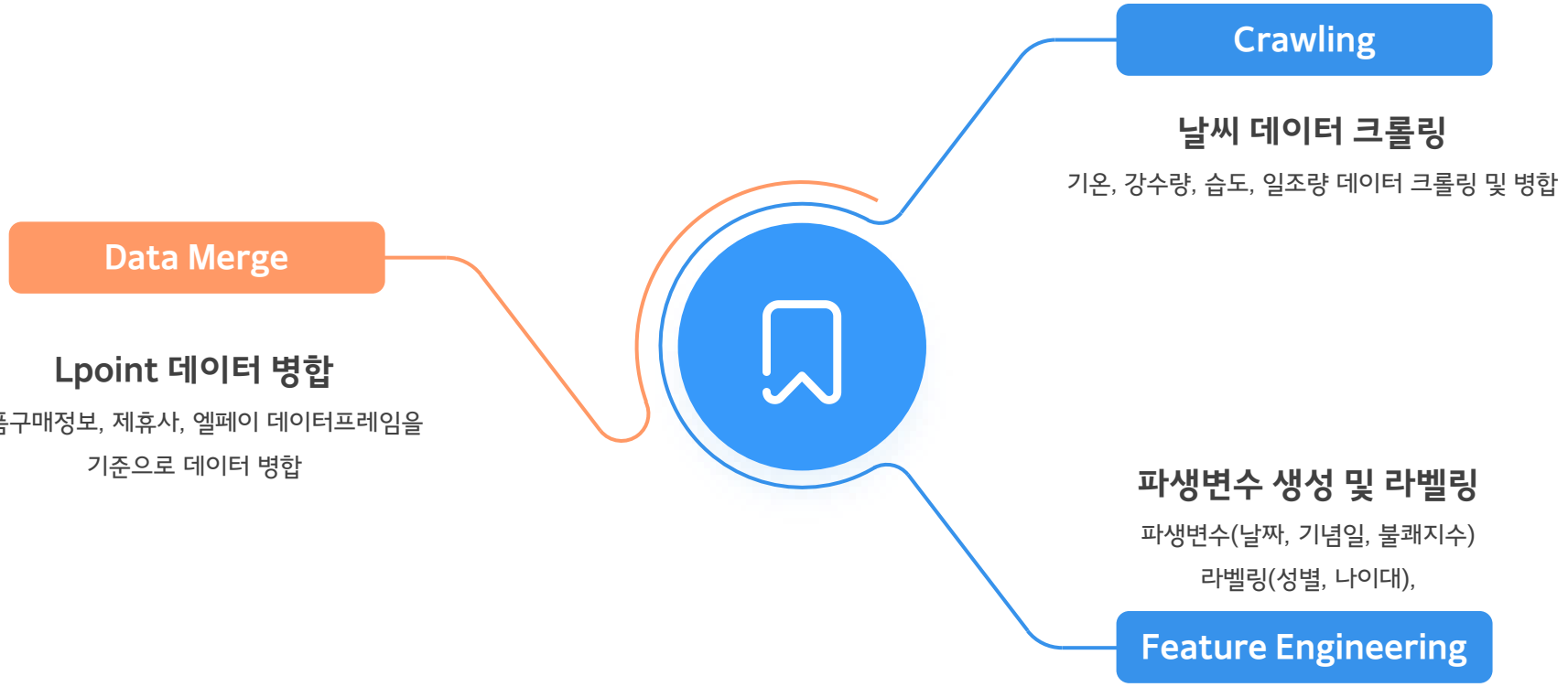
통해 개인화 마케팅을 달성한다.



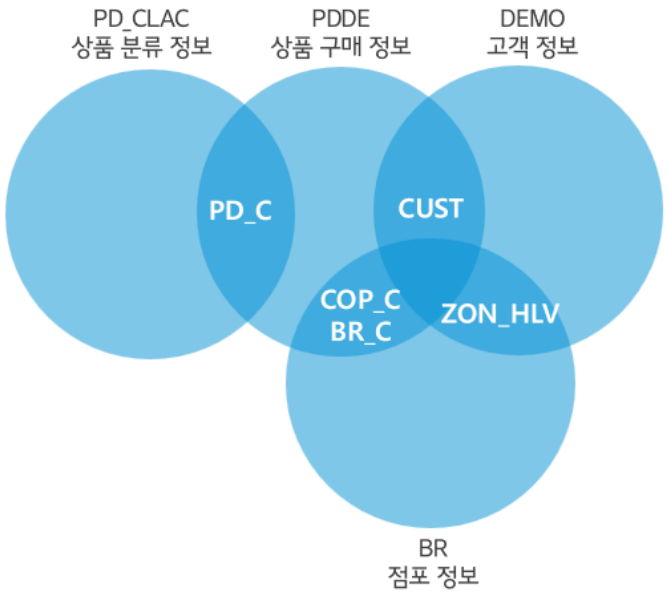
02 데이터 병합

02 데이터 병합

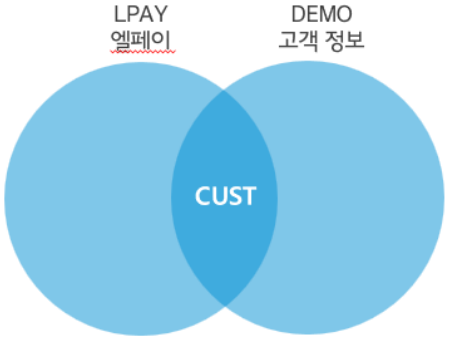
: 단계


Crawling
날씨 데이터 크롤링
기온, 강수량, 습도, 일조량 데이터 크롤링 및 병합
Data Merge
Lpoint 데이터 병합
상품구매정보, 제휴사, 엘페이 데이터프레임을
기준으로 데이터 병합
파생변수 생성 및 라벨링
파생변수(날짜, 기념일, 불쾌지수)
라벨링(성별, 나이대),
Feature Engineering

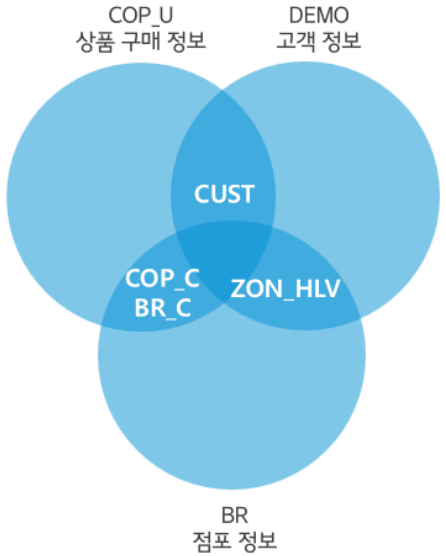
02 데이터 병합



pdde_final.csv



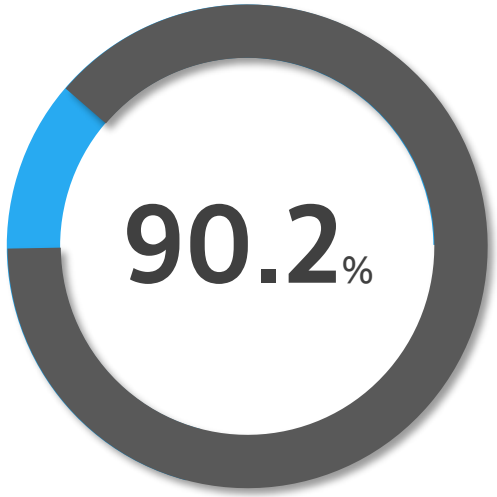
lpay_final.csv



cop_final.csv

03 EDA & 전처리

온/오프라인 비율



변수명	설명	예시
강수량(mm)	당일 날의 평균 강수량	3.497927
평균기온(℃)	당일 날의 평균 기온	13.391656
최저기온(℃)	당일 날의 최저 기온	8.654905
최고기온(℃)	당일 날의 최고 기온	18.835589
평균습도(%rh)	당일 날의 평균 습도	70.232968
최저습도(%rh)	당일 날의 최저 습도	24.424359
일조합(hr)	태양 광선이 지표에 닿는 시간	6.535048
일사합(MJ/m2)	태양의 광선이 지표에 닿는 양	14.652897

오프라인 데이터 Focusing

오프라인 구매에 영향을 줄 수 있는 기상 데이터 크롤링

출처 : 종관기상관측(ASOS)

03 EDA & 전처리

변수명	설명	예시
강수량(mm)	당일 날의 평균 강수량	3.497927
평균기온(℃)	당일 날의 평균 기온	13.391656
최고기온(℃)	당일 날의 최고 기온	18.835589
최저기온(℃)	당일 날의 최저 기온	8.654905
평균습도(%rh)	당일 날의 평균 습도	70.232968
최저습도(%rh)	당일 날의 최저 습도	24.424359
일조합(hr)	태양 광선이 지표에 닿는 시간	6.535048
일사합(MJ/m2)	태양의 광선이 지표에 닿는 양	14.652897

최고 기온 — 최저 기온

“ ”

일교차 파생변수 생성

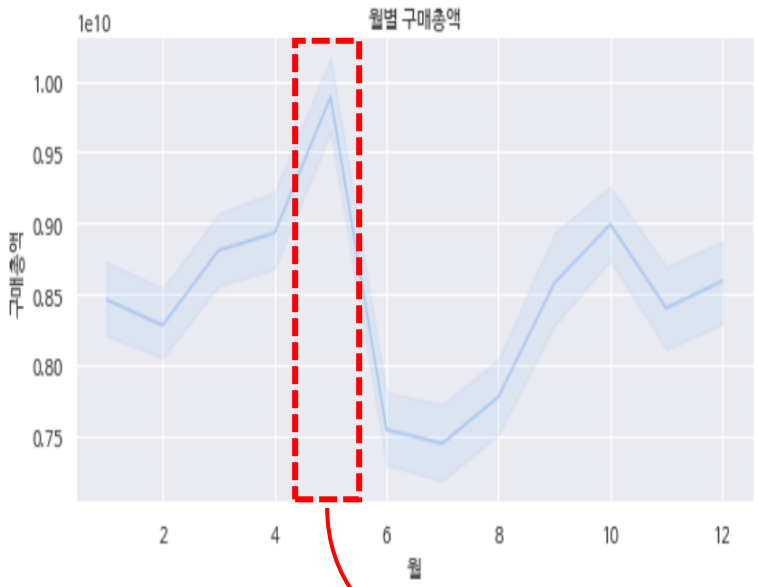
03 EDA & 전처리

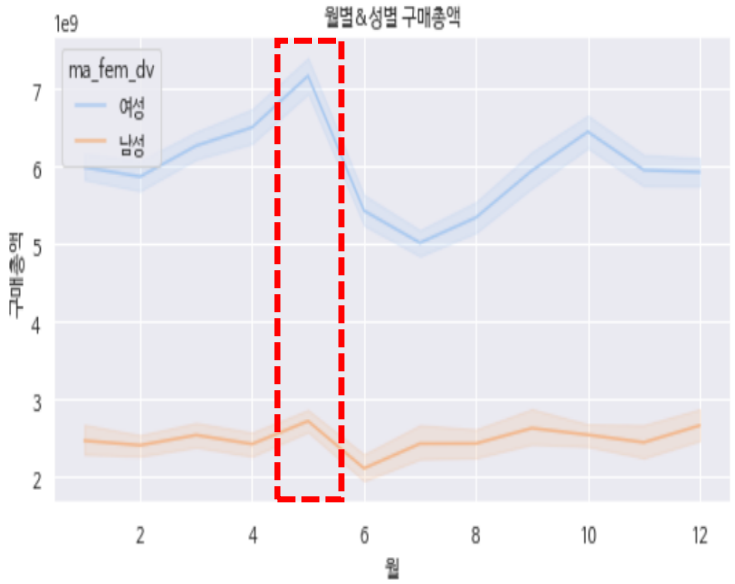
변수명	설명	예시
강수량(mm)	당일 날의 평균 강수량	3.497927
평균기온(℃)	당일 날의 평균 기온	13.391656
최고기온(℃)	당일 날의 최고 기온	18.835589
최저기온(℃)	당일 날의 최저 기온	8.654905
평균습도(%rh)	당일 날의 평균 습도	70.232968
최저습도(%rh)	당일 날의 최저 습도	24.424359
일조합(hr)	태양 광선이 지표에 닿는 시간	6.535048
일사합(MJ/m2)	태양의 광선이 지표에 닿는 양	14.652897

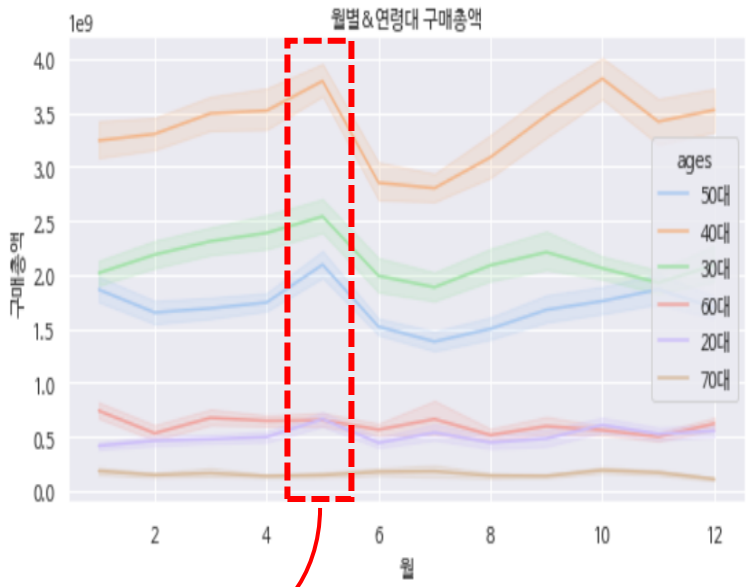
9/5 * 평균기온
-0.55 * (1-평균습도/100)
*(9/5*평균기온-26) + 32

"불쾌지수 파생변수 생성"

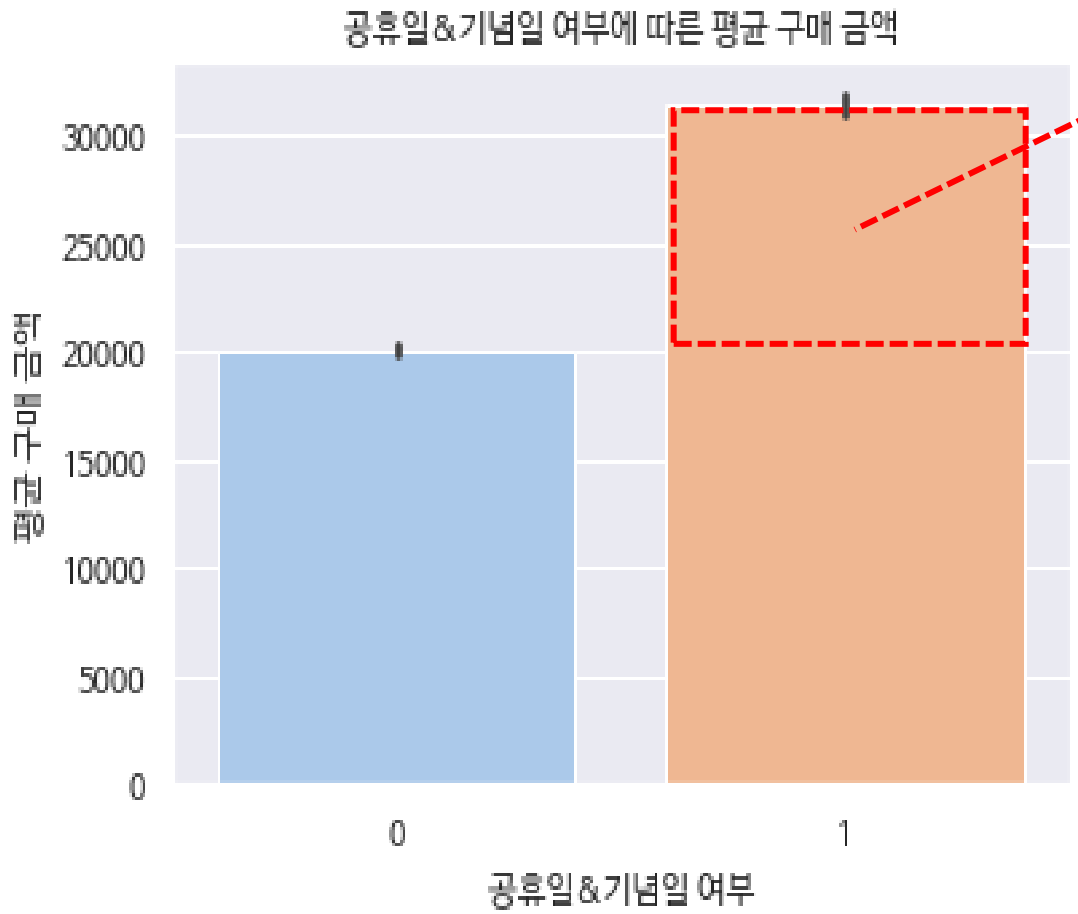
03 EDA & 전처리







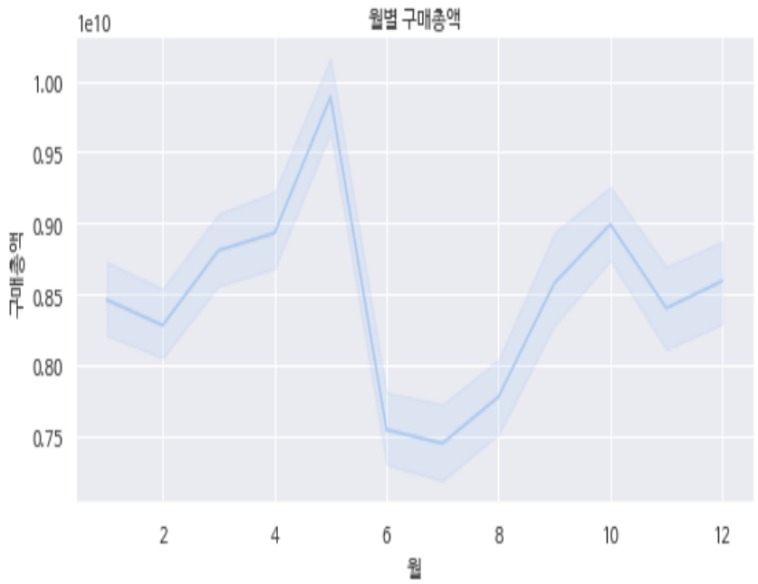
“ 공휴일 및 기념일이 많은 5월 달에 최고 매출액 기록
“ holiday_anniversary “ 공휴일 & 기념일 파생 변수 ”

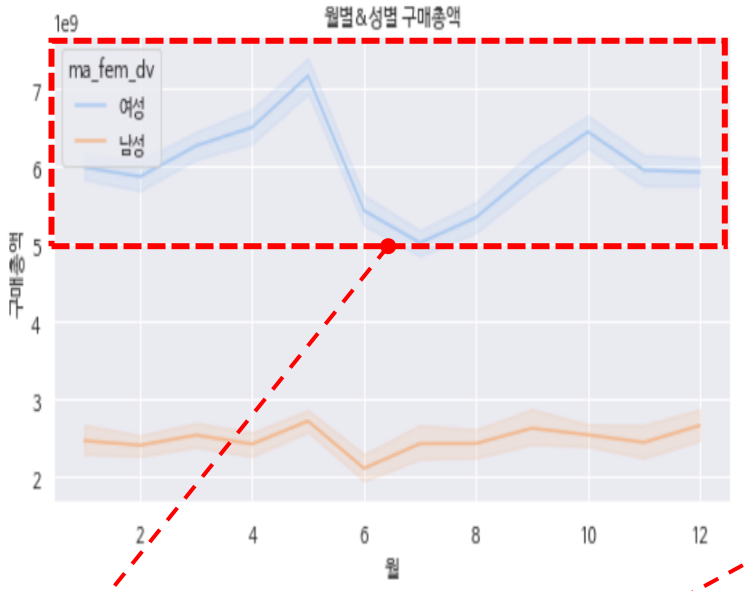


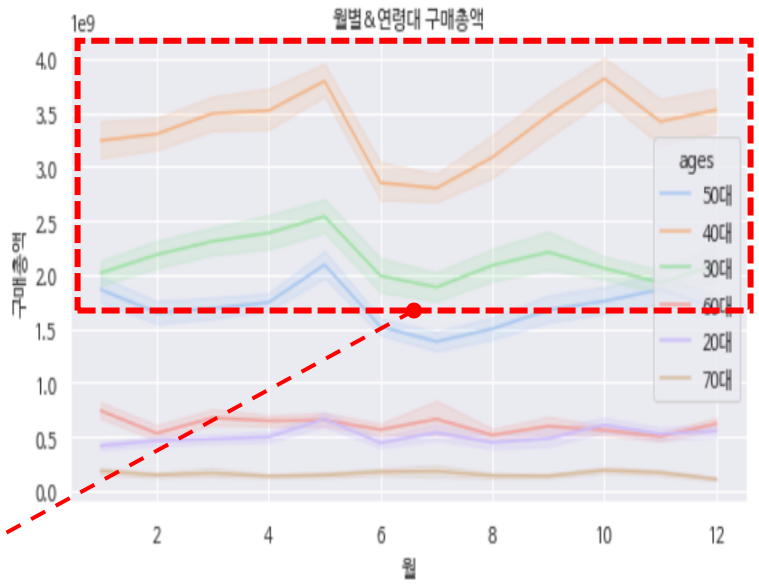
공휴일 & 기념일에 평균 지출 금액이 더 높다.

"공휴일 & 기념일에 고객 구매력 상승"

03 EDA & 전처리







“ 여성의 소비 지출액, 40대 & 30대의 소비지출액에 주목할 필요 ”

DATA LABELING

변수명	설명
ma_fem_dv	성별
ages	나이
zon_hlv	지역 코드
clac_hlv_nm	상품 대분류명

→ 라벨링

DATA DROP

- 2020년 데이터 제외 : 2021년 데이터만 사용하기 위함

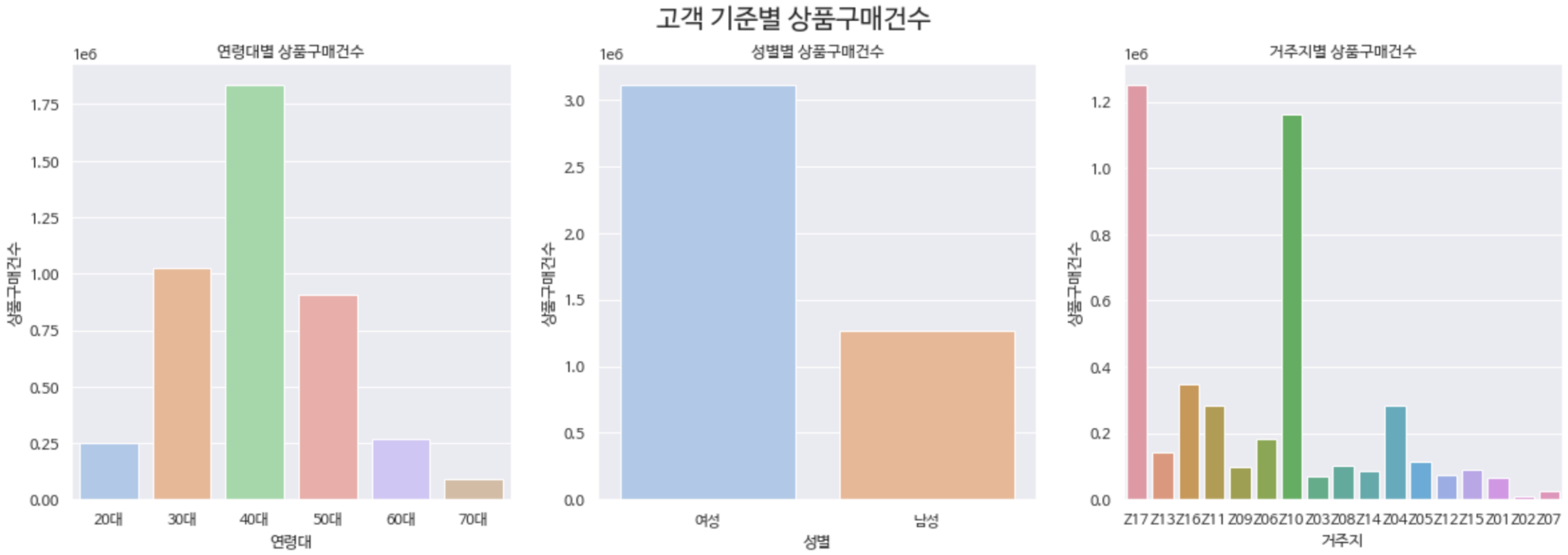
DATA SPLIT

de_dt →

변수명	예시
year	2021
month	6
요일	0~6까지 월~일 표현

03 EDA & 전처리

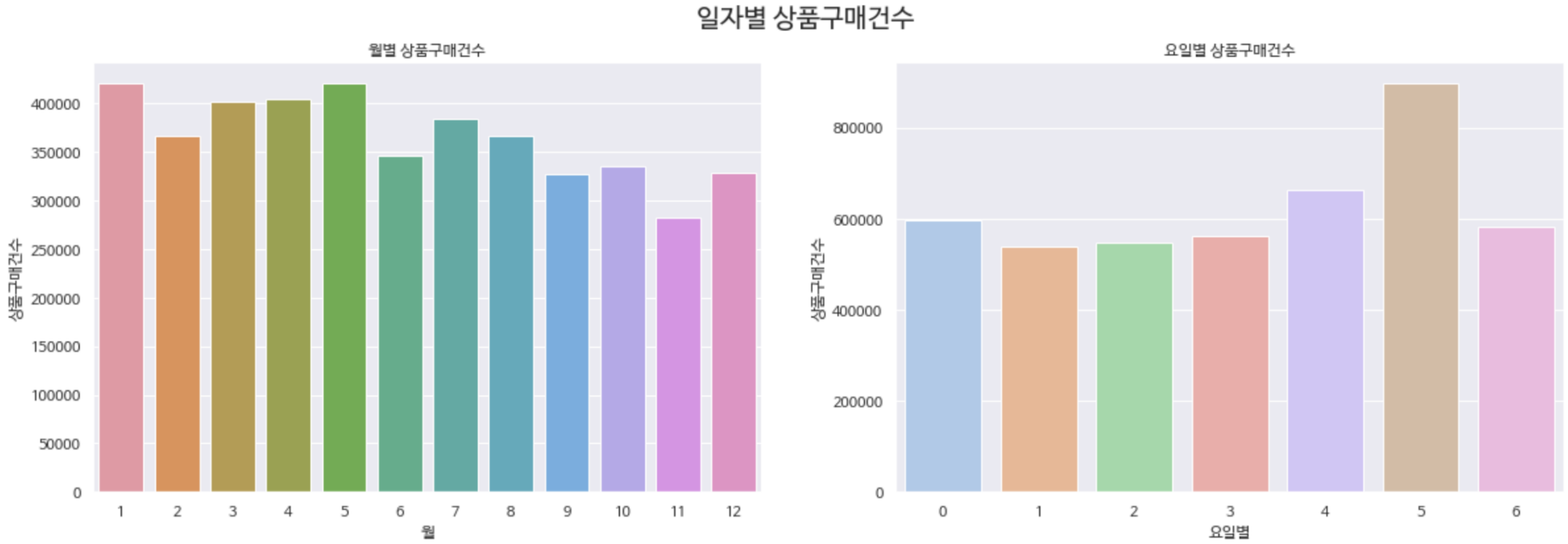
- 상품 구매 건수



- **상품 구매 건수에서 연령(30대, 40대), 여성의 높은 비율을 가지고 있음**
- **특정 지역(Z17, Z10)지역의 구매건수가 높았음**

03 EDA & 전처리

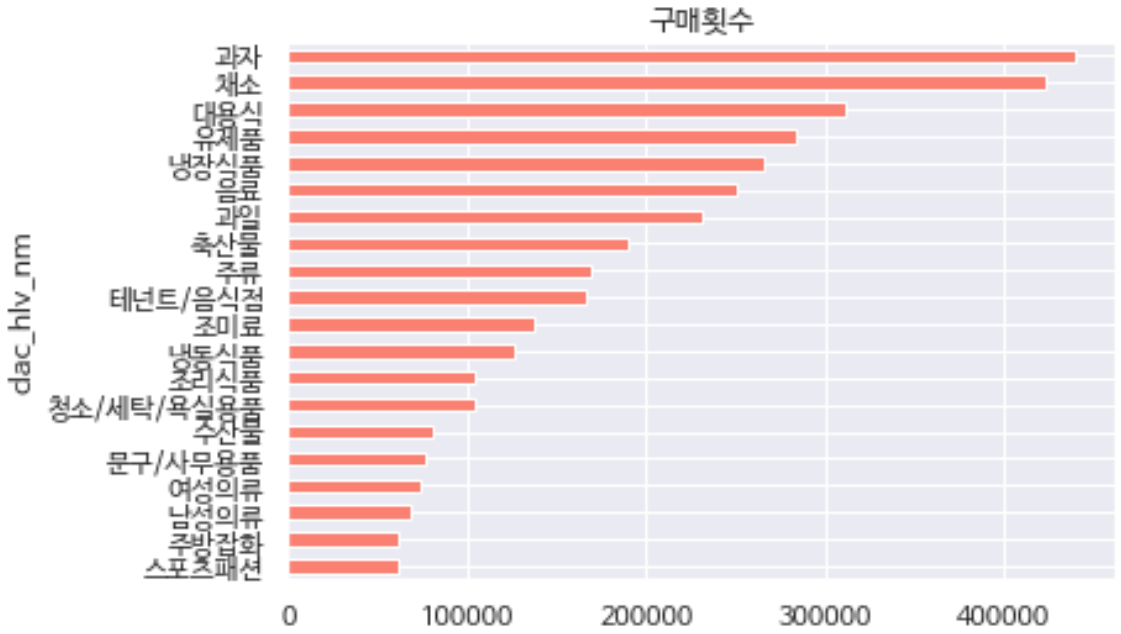
- 상품 구매 건수

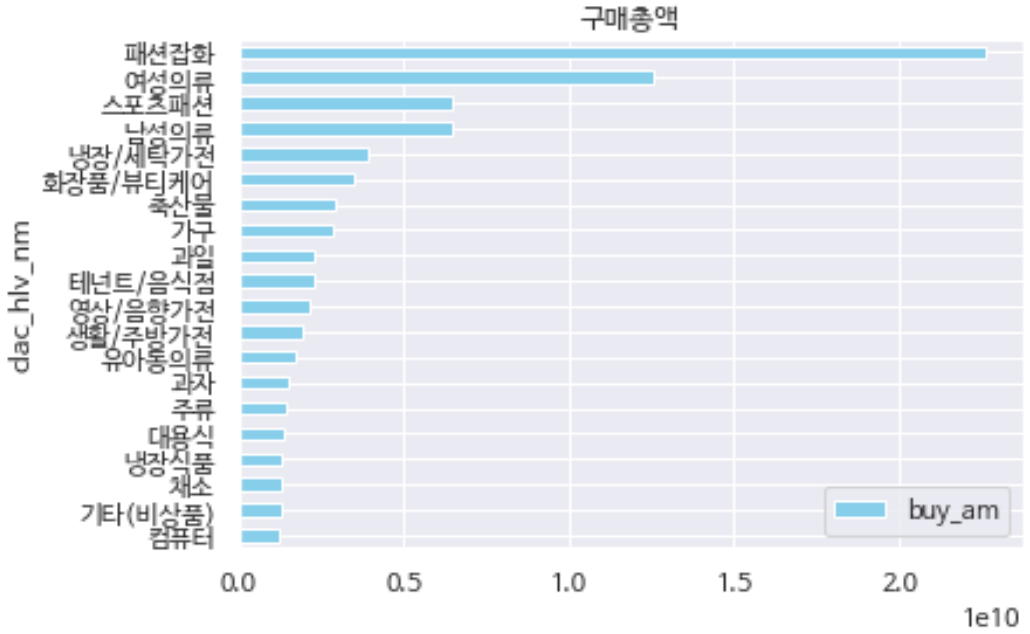


- **5월에 가장 많은 상품 구매가 이루어졌고, 11월에 저조한 구매 형태를 보임**
- **토요일에 상품 구매가 가장 높게 나타났고 화요일에 가장 낮게 나타남**

03 EDA & 전처리

- 상품 대분류 분석



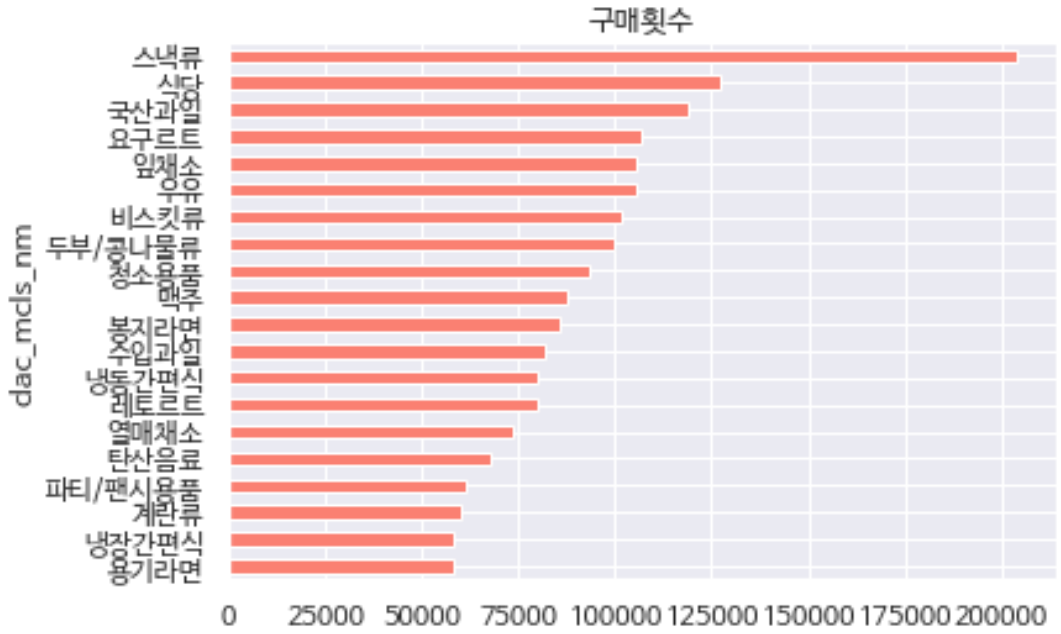


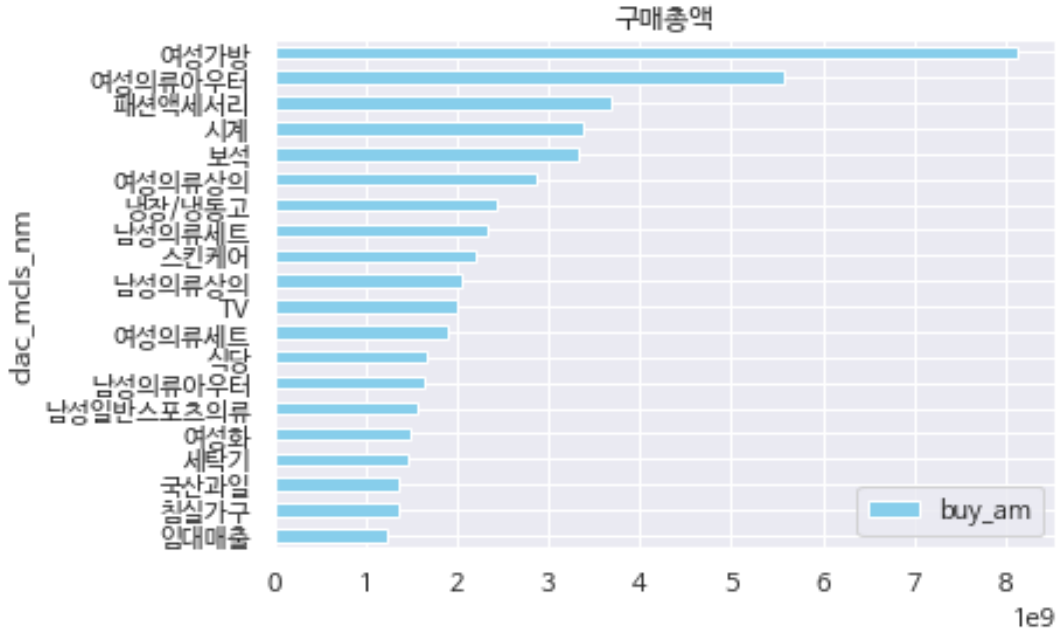
- **상품 대분류 구매건수는 과자, 채소, 대용식 등 식료품에 높은 비중을 보임**
- **구매총액을 보면 패션잡화, 여성의류에서 가장 높은 매출액이 나오는 것을 알 수 있음**

➔ **고객군의 '구매력'에 따라 상품군 추천이 가능할 것으로 보임.**

03 EDA & 전처리

- 상품 중분류 분석



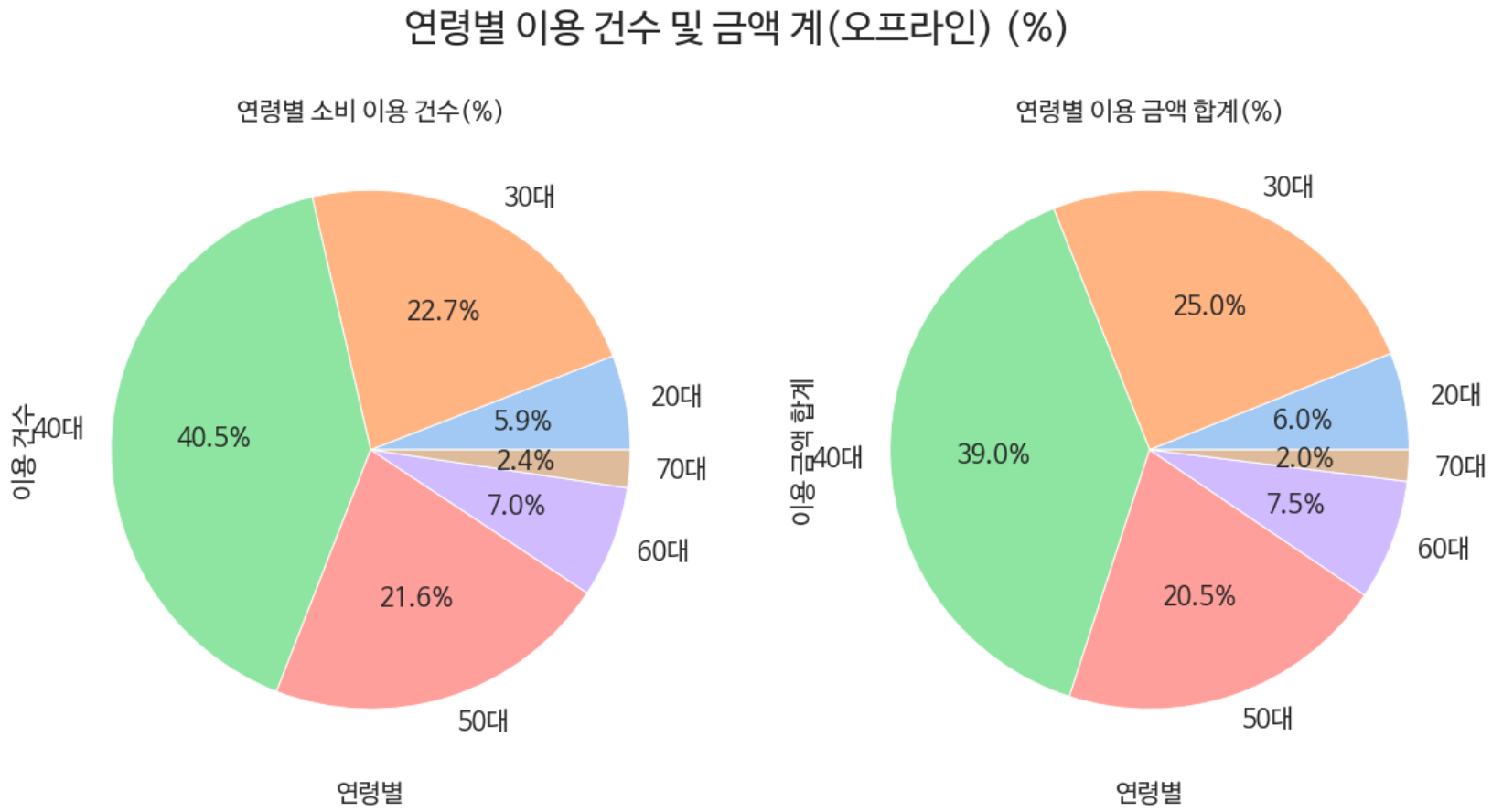


- **상품 중분류 구매건수는 스낵류, 식당, 국산 과일 등에 높은 비중을 보임**
- **구매총액을 보면 여성가방, 여성의류 아우터, 패션액세서리, 시계 등에서 높은 매출액을 기록**

➔ 잠재고객의 상품 대분류 및 구매력을 예측하면 관련 중분류 상품들을 추천 가능

03 EDA & 전처리

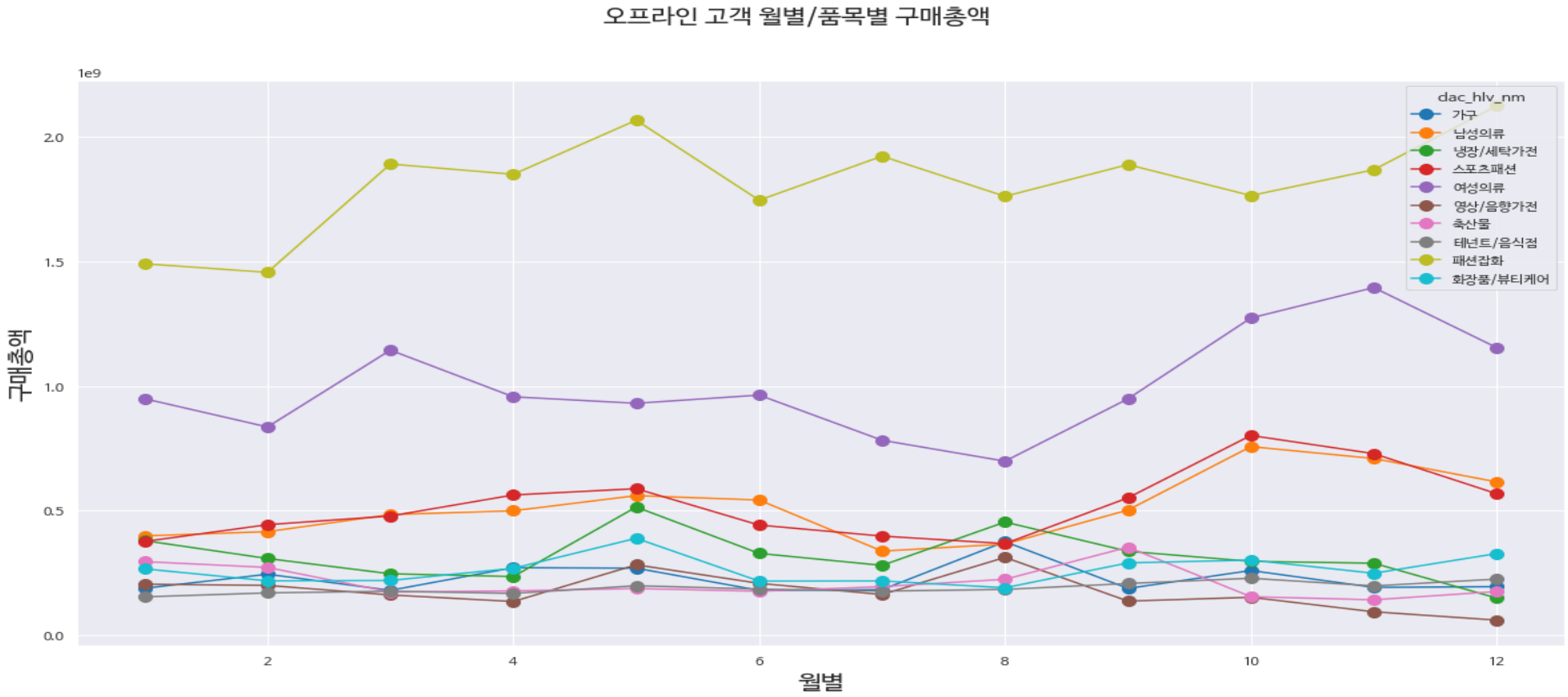
- 오프라인 고객군 연령대 분포



- **이용건수로는 40대가 가장 높은 비율로 나타났고, 30대, 50대 순으로 그 다음 이용건수를 나타냄**
- **이용금액 합계로는 40대가 마찬가지로 가장 높은 비율로 나타남**

03 EDA & 전처리

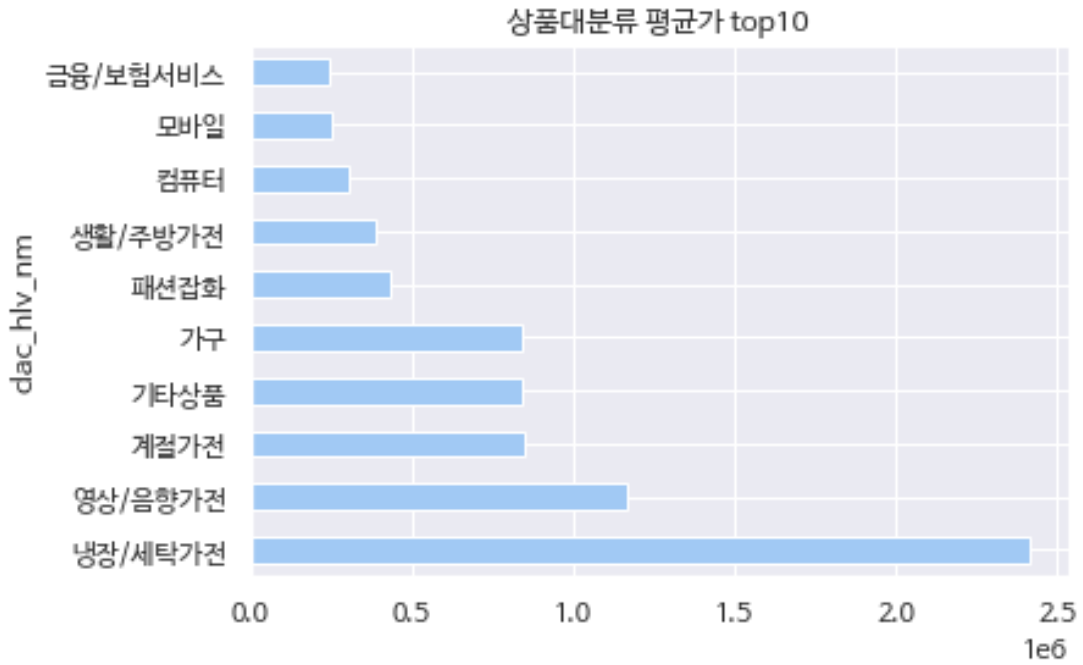
- 계절성 파악(오프라인 상품)

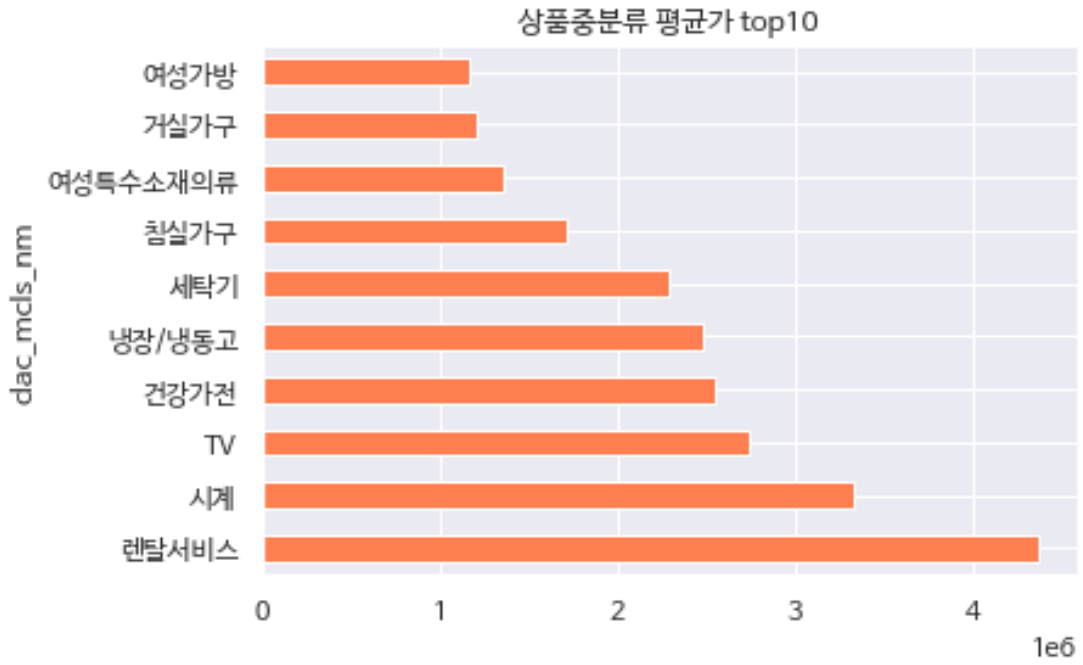


- **고객 구매 데이터(오프라인 고객)의 구매총액 결과 패션잡화, 여성의류, 스포츠패션의 비중이 높게 나타남을 알 수 있음**

03 EDA & 전처리

- 상품 평균가 파악





- **상품 대분류 결과 냉장/세탁가전, 영상/음향가전이 가장 높은 비율로 나타남**
- **상품 중분류 결과 렌탈서비스, 시계, TV등 고가의 상품이 높은 평균가로 나타남**

04 고객 분류 모델

04 고객 분류 모델

pdde_merged.csv

K-Means Clustering

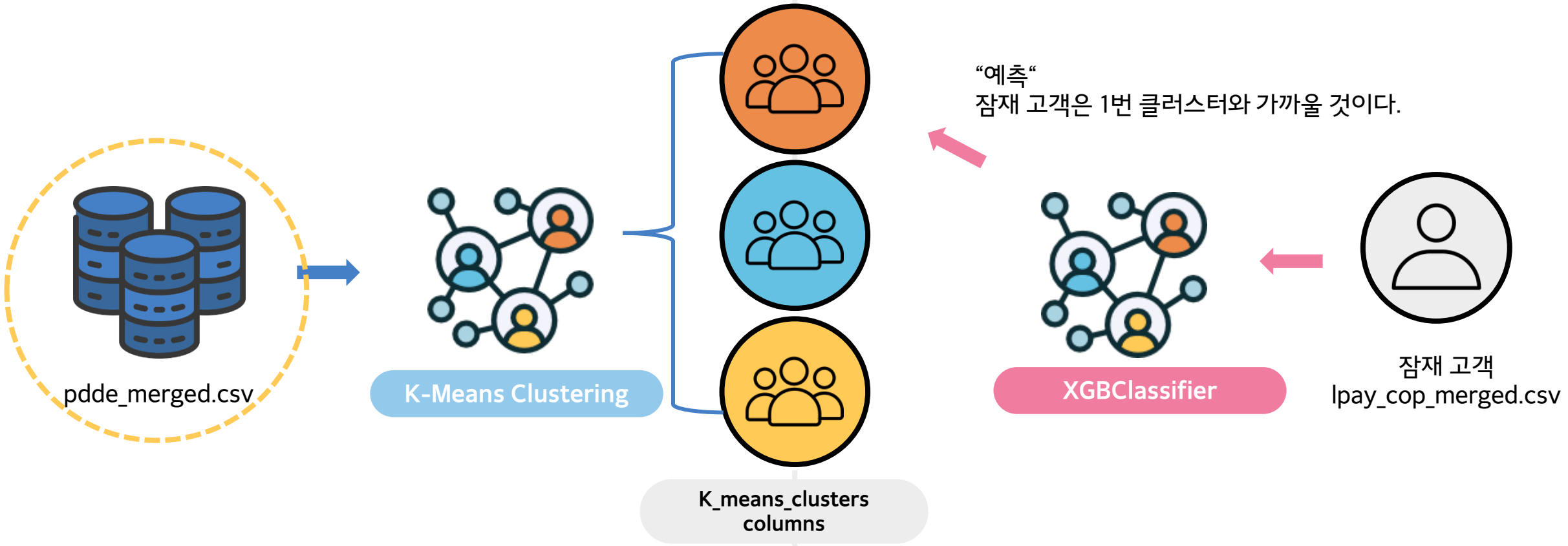
K_means_clusters
columns

"예측"
잠재 고객은 1번 클러스터와 가까울 것이다.

XGBClassifier

잠재 고객
lpay_cop_merged.csv

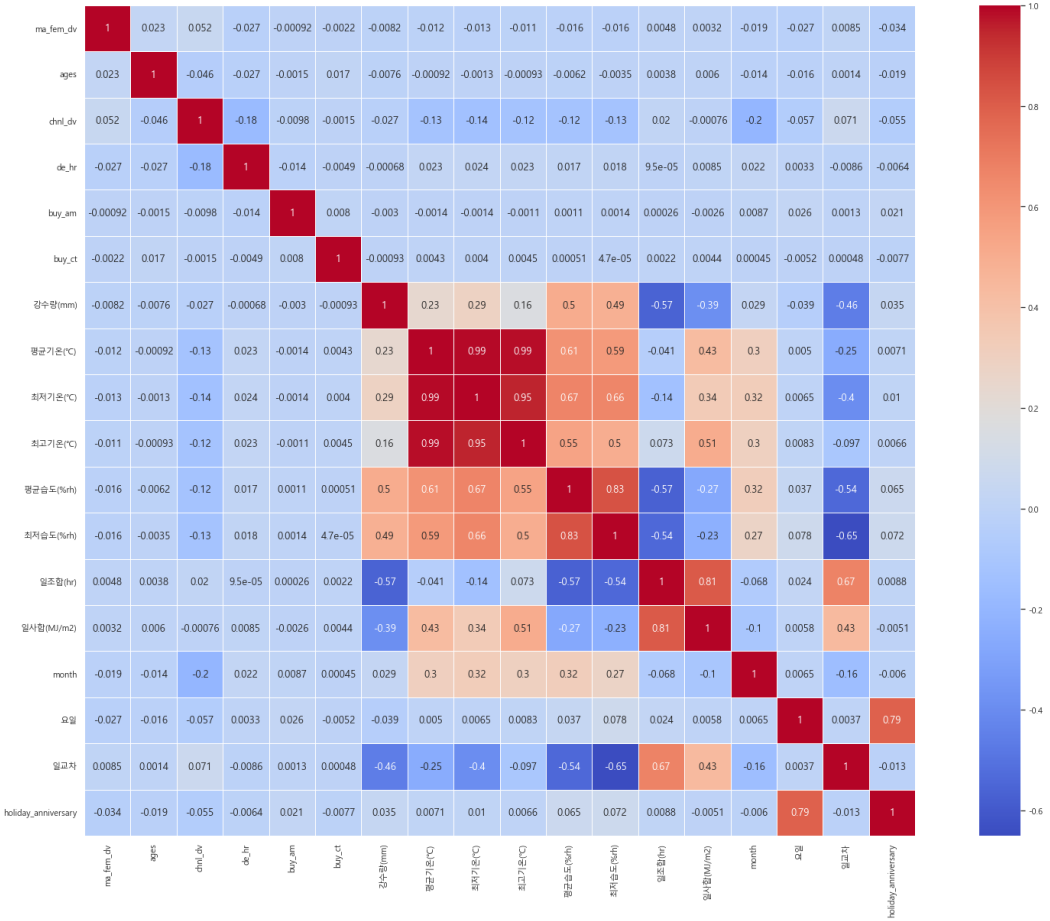
04 고객 분류 모델


pdde_merged.csv
K-Means Clustering
K_means_clusters
columns
"예측"
잠재 고객은 1번 클러스터와 가까울 것이다.
XGBClassifier
잠재 고객
lpay_cop_merged.csv

04 고객 분류 모델

STEP 01. 변수 선택, 전처리

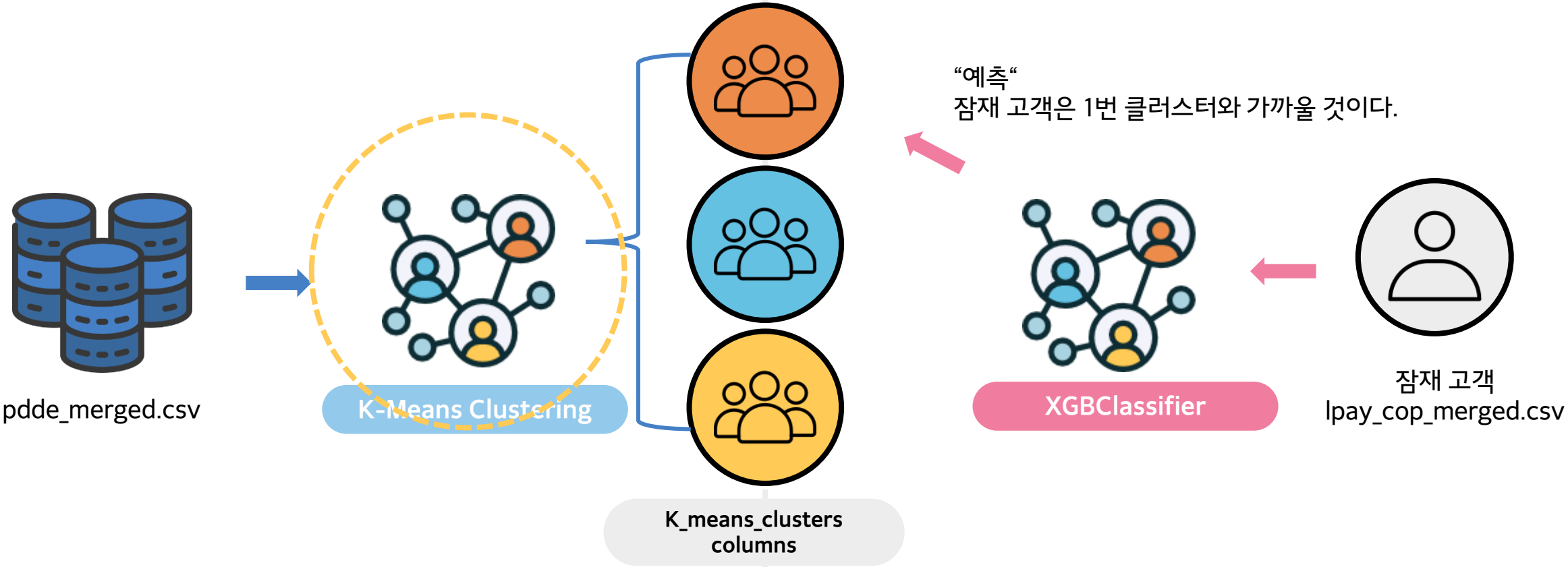
K-Means Clustering 사용을 위해 연속형 변수만 선택, 변수 제외



1)
상품 대분류에서 금융/보험서비스는
Row가 1개라 Drop

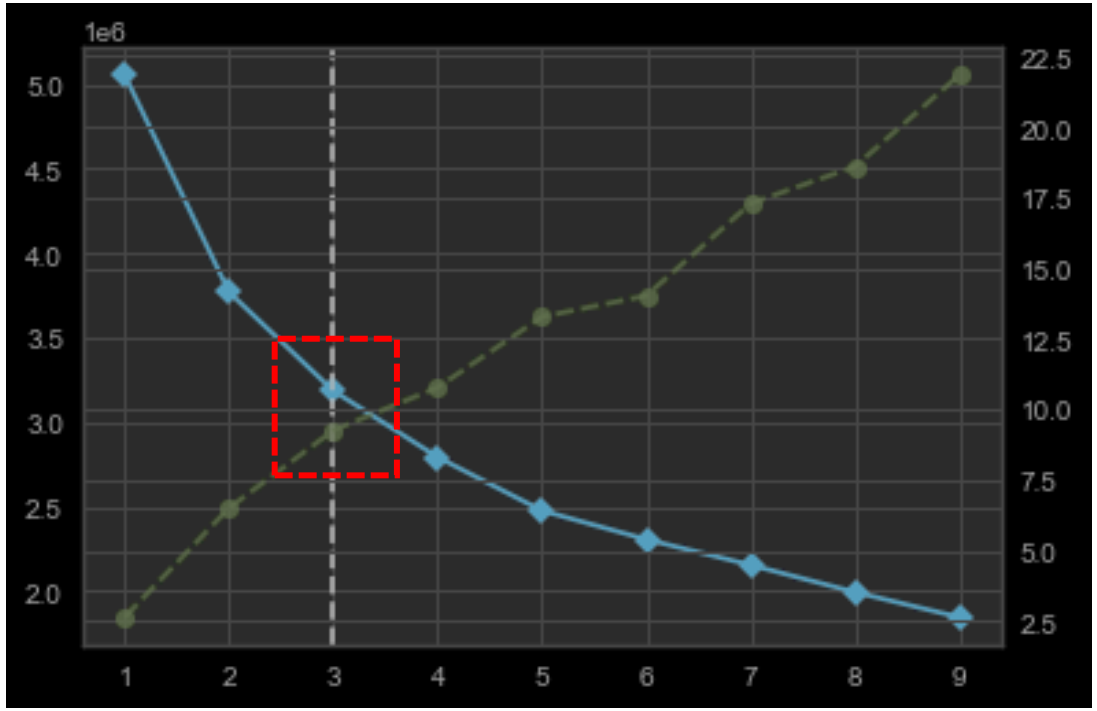
2)
K-Means Clustering 사용을 위해 범주형
변수 제거(AGE 제외)

04 고객 분류 모델


pdde_merged.csv
K-Means Clustering
K_means_clusters
columns
"예측"
잠재 고객은 1번 클러스터와 가까울 것이다.
XGBClassifier
잠재 고객
lpay_cop_merged.csv

STEP 02. 최적의 클러스터 개수 탐색

Silhouette Index & Score Index를 통한 최적의 클러스터 개수 탐색 및 시각화



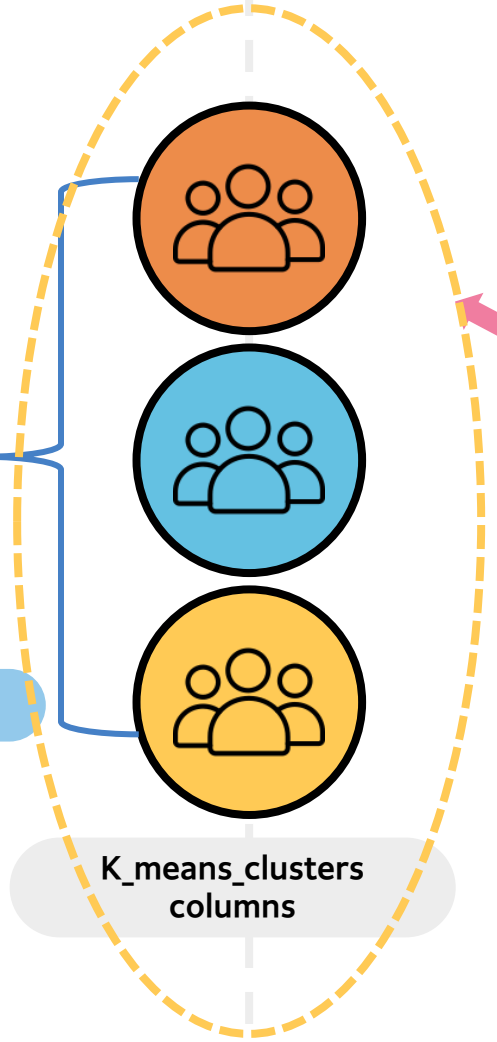
“ **Optimal Cluster** ”

K = 3

04 고객 분류 모델

pdde_merged.csv

K-Means Clustering



K_means_clusters
columns

"예측"
잠재 고객은 1번 클러스터와 가까울 것이다.

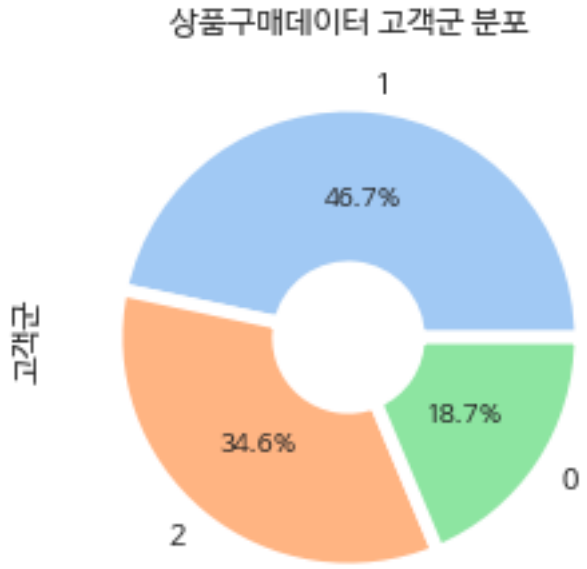
XGBClassifier

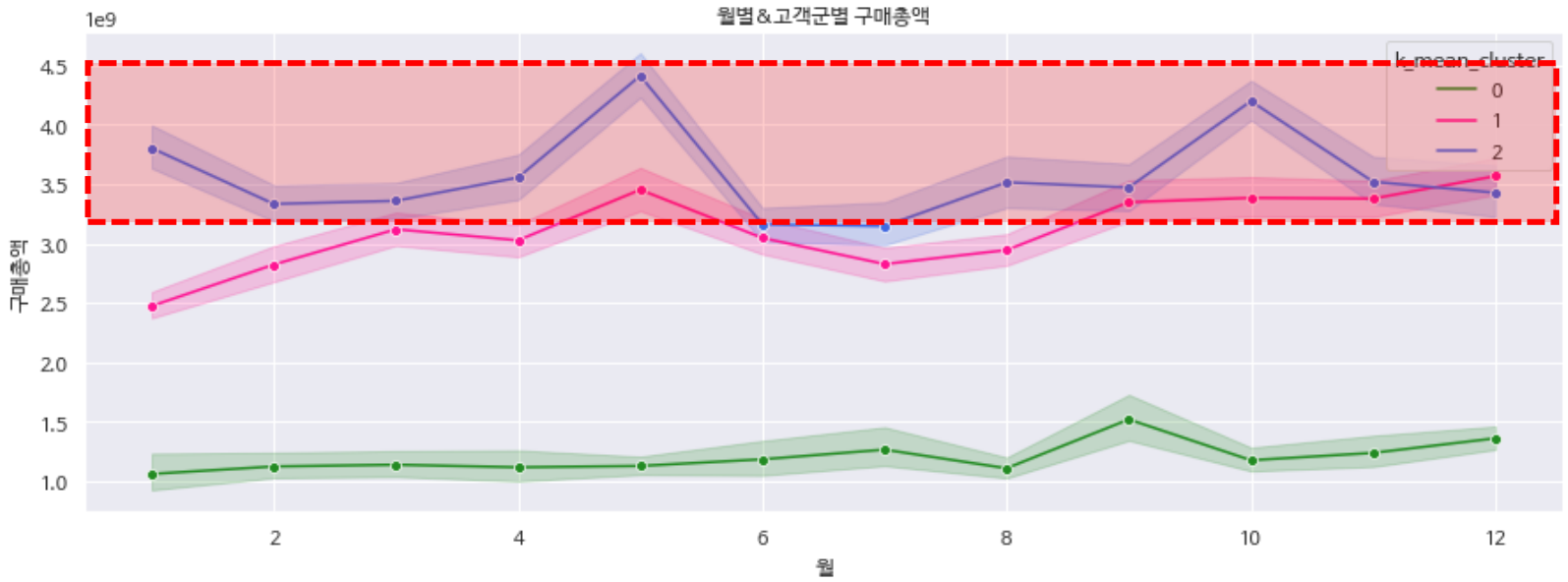
잠재 고객
lpay_cop_merged.csv

04 고객 분류 모델

STEP 03. 클러스터링 결과

고객군 분포, 고객군/월 별 총 구매 금액





- 고객군 분포 크기 : 1번 > 2번 > 0번
- 고객군 별 월간 총 구매 금액은 2번 그룹이 가장 높다.

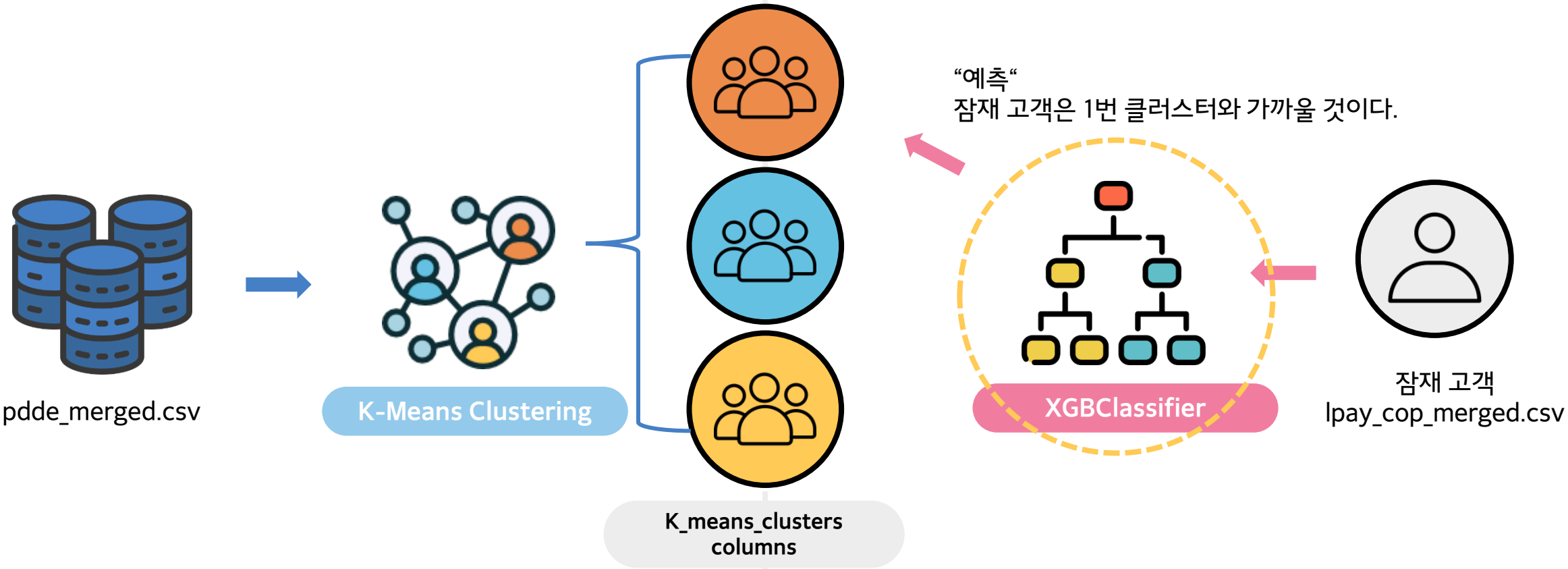
2번 고객군의 평균 소비 금액이 가장 높을 것으로 추측된다.

Check point

구매력이 큰 2번 고객군을 타겟
고객으로 설정한다.

05-1 고객 예측 모델

05-1 고객 예측 모델


pdde_merged.csv
K-Means Clustering
K_means_clusters
columns
"예측"
잠재 고객은 1번 클러스터와 가까울 것이다.
XGBClassifier
잠재 고객
lpay_cop_merged.csv

05-1 고객 예측 모델

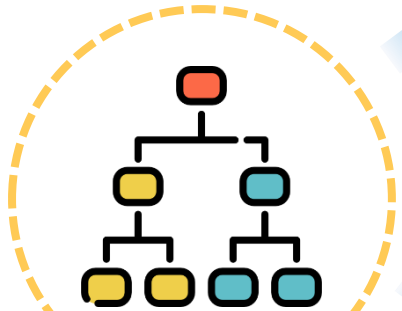
STEP 01. 모델링

PDDE 전처리 | XGBoost 모델링

PDDE

holiday_anniversary	k_mean_cluster	Clac_hlv_nm
0	0	0
1	2	2

엘포인트 (LPAY), 제휴사(COP) 고객 예측을 위해
고객 구매 데이터 (PDDE)의 clac_hlv_nm 지표를 삭제한다.



XGBClassifier

LPAY

고객 번호	불쾌지수
M629656521	53.18051
M216016456	55.3781

+

k_mean_cluster
0
2

COP

holiday_anniversary	불쾌지수
0	53.18051
1	55.3781

+

k_mean_cluster
0
2

Lpay, Cop 의 고객 데이터를 3개의 군집으로 예측한다.

05-1 고객 예측 모델

STEP 01. 데이터 전처리

Model_1_고객 분류 모델 결과 추가

Cop_latent

일교차	holiday_anniversary	불쾌지수
14.5	0	53.18051
13.7	1	55.3781
12.8	0	43.10208

일교차	holiday_anniversary	불쾌지수	k_mean_cluster
14.5	0	53.18051	0
13.7	1	55.3781	2
12.8	0	43.10208	0

Lpay_latent

일교차	불쾌지수	holiday_anniversary
7.9	78.4869	0
7.9	78.4869	0
7.9	78.4869	0

일교차	불쾌지수	holiday_anniversary	k_mean_cluster
7.9	78.4869	0	0
7.9	78.4869	0	1
7.9	78.4869	0	1

“고객 예측 모델“ 에서 예측된 고객군 데이터를 lpay & cop 에 변수로 추가한다.

STEP 02. 파라미터 설정

XGBClassifier 주요 파라미터 설정

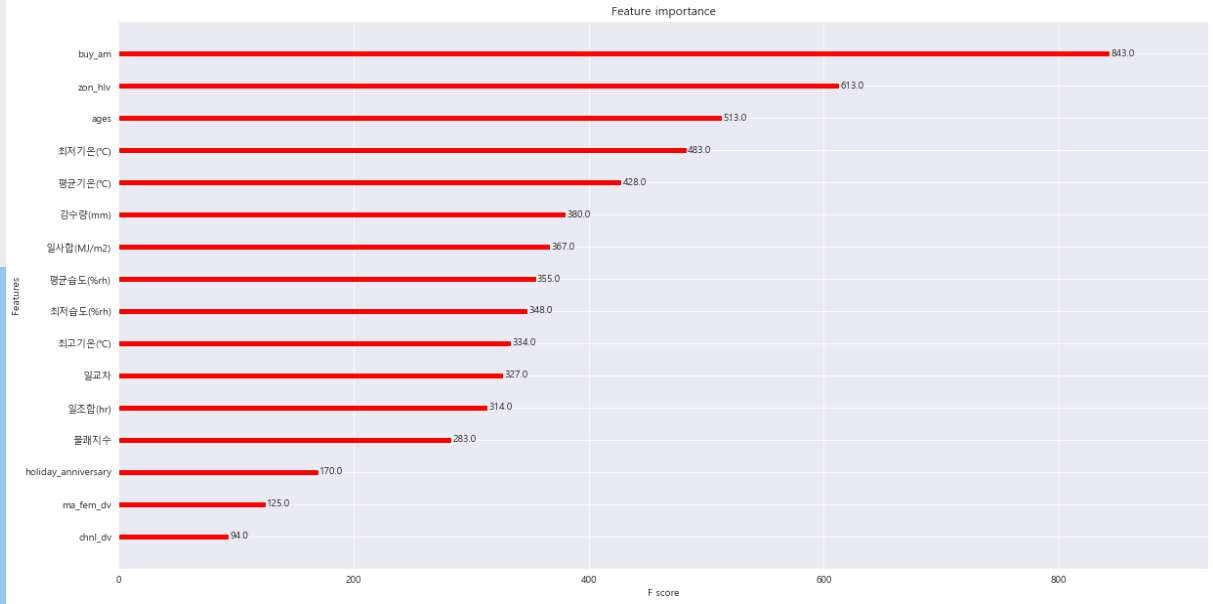
Parameter	Option	Explanation
booster	gbtree	트리계열
n_estimators	50	반복횟수
Learning_rate	0.05	Overfitting을 방지하기 위해 낮게 설정
Objective	'multi:softprob'	다중분류모델
Eval_metric	Mlogloss, auc	다중분류 평가지표
Early_stopping	100	Overfitting 방지
Max_Depth	8	최대 트리 깊이

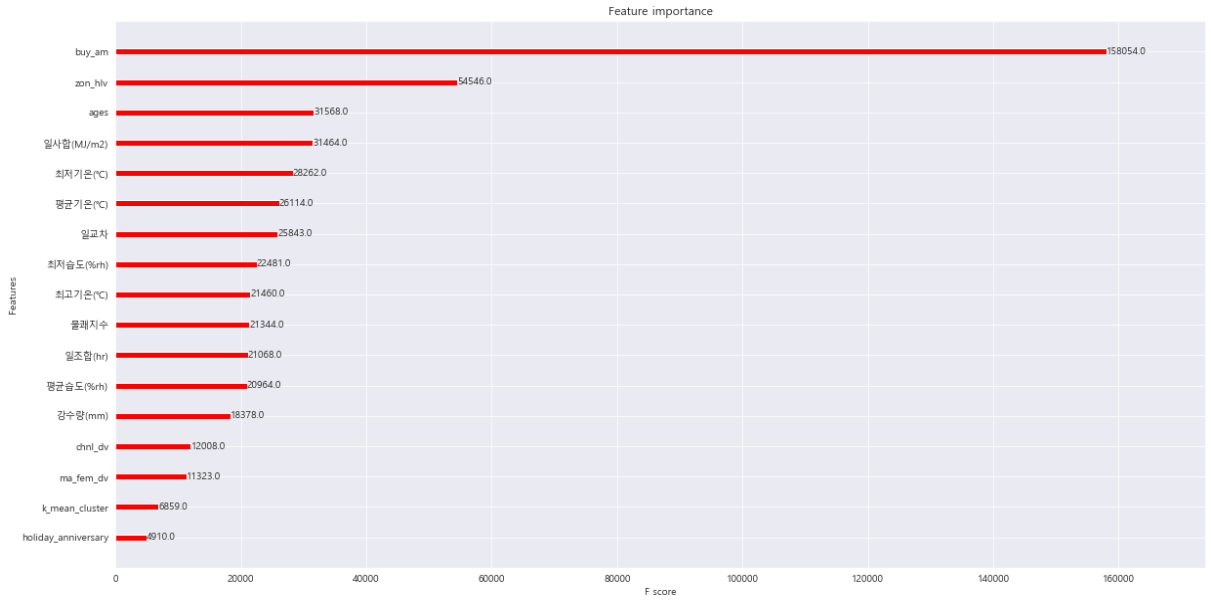
다중분류를 위하여 objective를 multi:softprob로 결정
고객 구매 정보 데이터의 과적합을 막기 위해 learning_rate, early_stopping 등 핵심지표 설정

05-1 고객 예측 모델

STEP 03. 고객 예측 모델 별 Feature Importance

lpay / cop 데이터의 고객군 예측을 위해 활용된 변수 및 F score로 표현한 변수 중요도

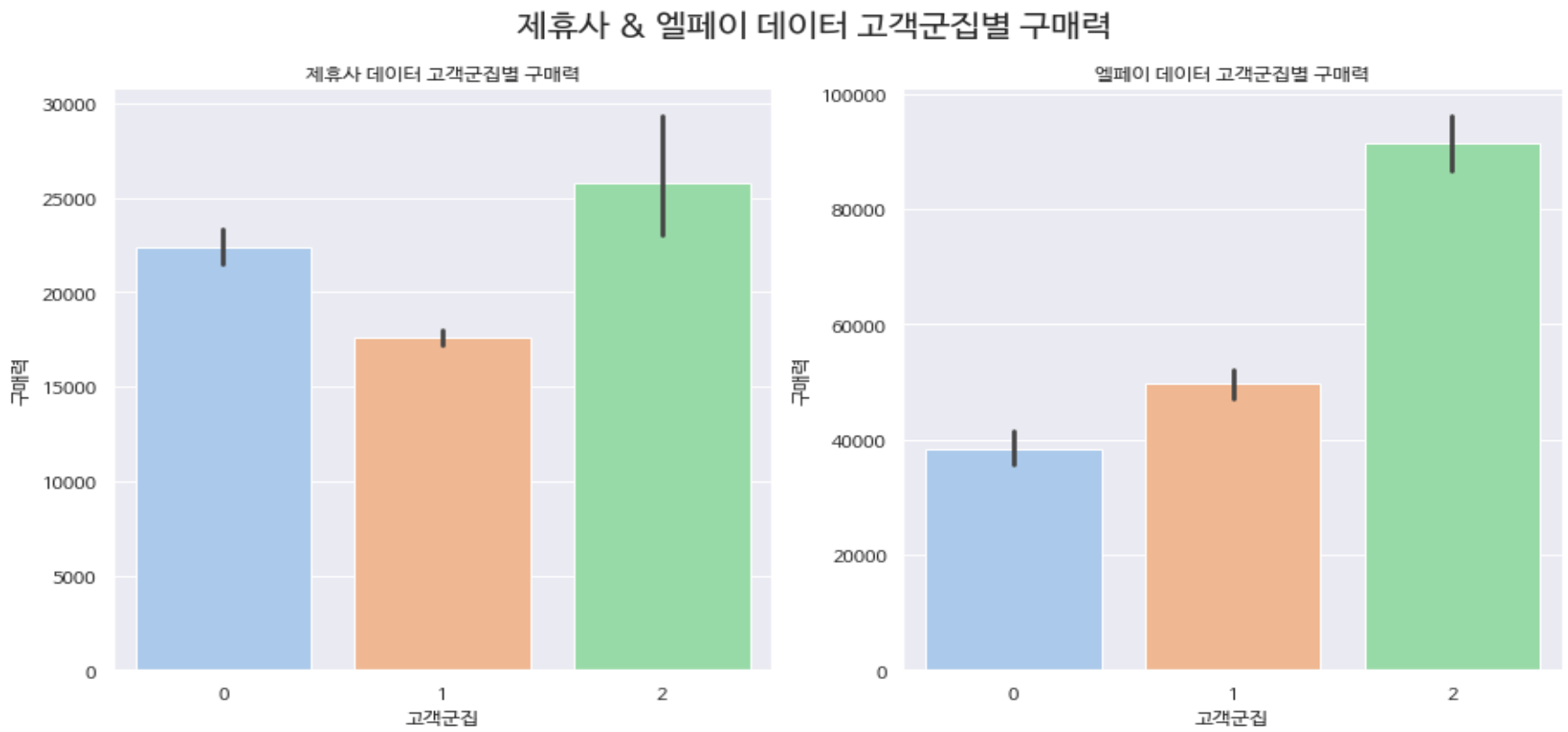




“ Lpay와 Cop 두 고객 데이터의 상위 3개 지표가 “ buy_am | zon_hlv | ages “ 으로 동일하다. ”

STEP 03. 고객군 예측 결과

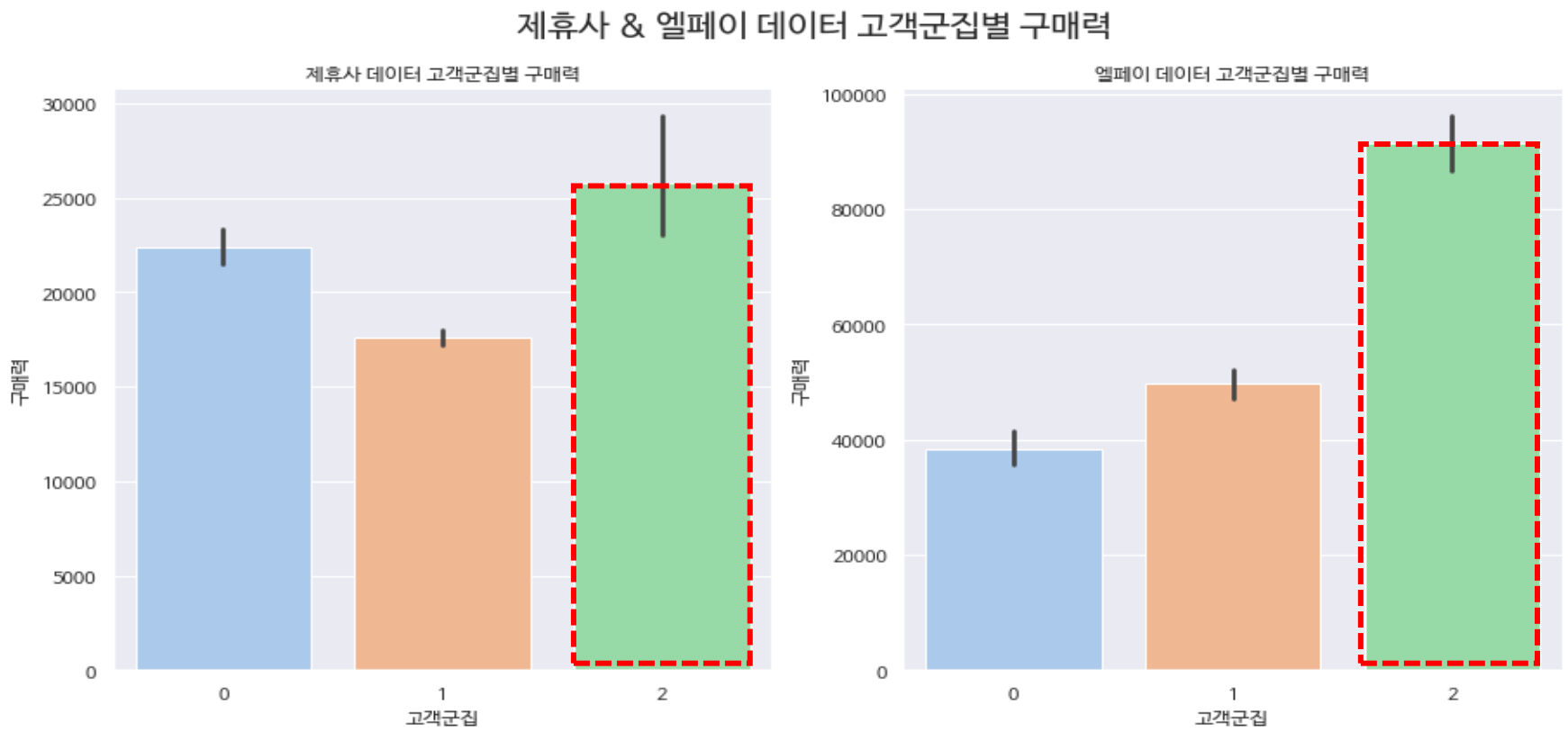
결과 분석/ 군집분석



➔ 엘페이, 제휴사 데이터에서도 상품구매데이터와 마찬가지로 2번 고객군집에서 높은 평균지출금액을 보였음

STEP 03. 고객군 예측 결과

결과 분석/ 군집분석

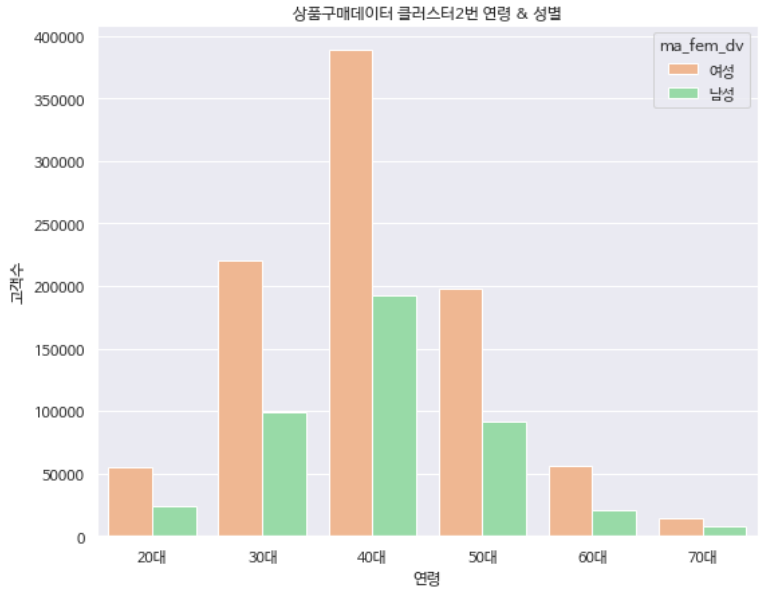


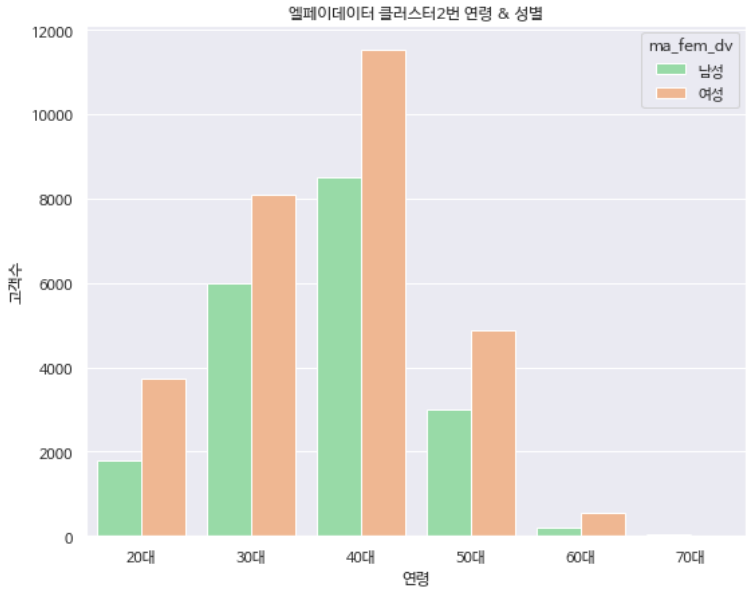
“ 오프라인 데이터 & 2번 고객군 데이터에 focusing ”

STEP 03. 고객군 예측 결과

결과 분석 / 군집 분석

데이터별 2번 고객군 특성 파악



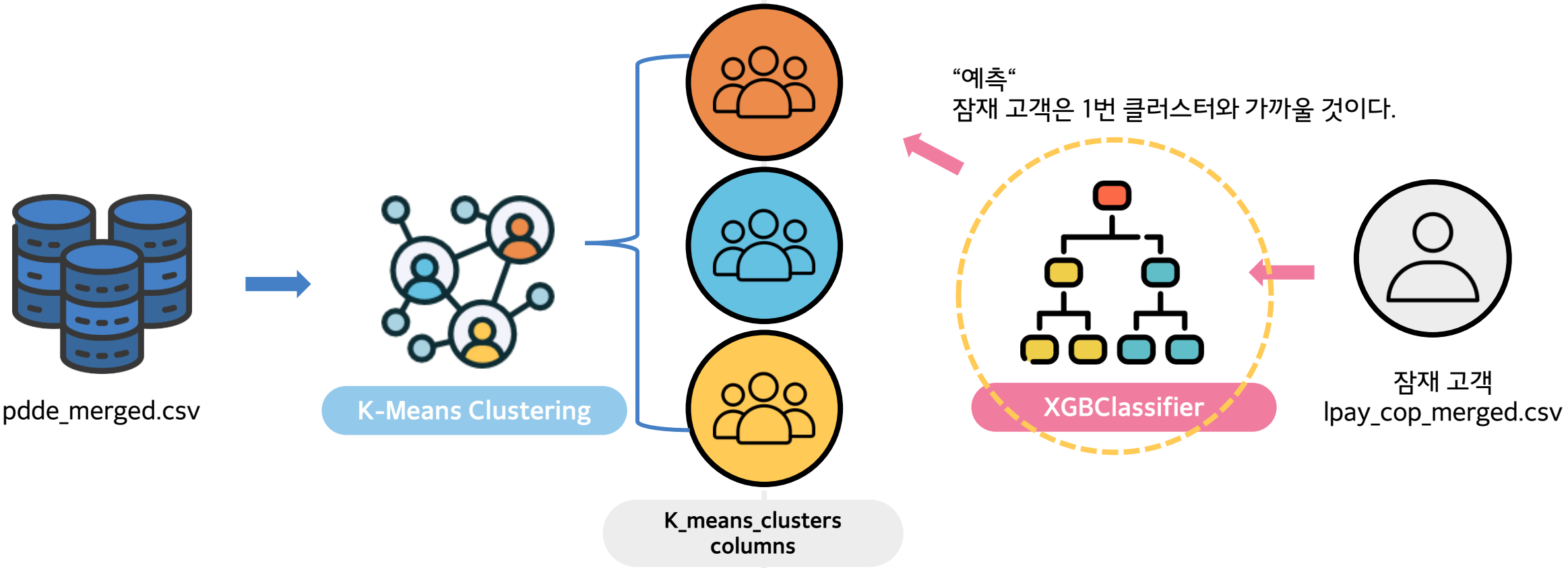


상품구매데이터, 제휴사 데이터, 엘페이 데이터 모두에서 비슷한 고객 특성을 보임.

30대, 40대에 많은 여성이 분포돼 있음을 확인할 수 있고, 남성 또한 높은 비율을 보이고 있음을 확인
➔ 앞선 자료와 비교했을 때, 30대, 40대 고객이 구매력이 높은 것을 추측할 수 있음

05-2 상품 추천 모델

05-2 상품 추천 모델


pdde_merged.csv
K-Means Clustering
K_means_clusters
columns
"예측"
잠재 고객은 1번 클러스터와 가까울 것이다.
XGBClassifier
잠재 고객
lpay_cop_merged.csv

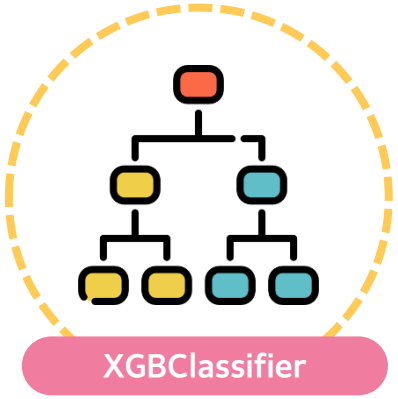
05-2 상품 추천 모델

STEP 01. XGBoost 모델링

데이터

PDDE

holiday_anniversary	clac_hlv_nm	k_mean_cluster
0	34	0
1	57	2



LPAY

고객 번호	불쾌지수	k_mean_cluster
M629656521	53.18051	0
M216016456	55.3781	2

+

clac_hlv_nm
2 (패션잡화)
1 (식료품)

LPAY

고객 번호	불쾌지수	k_mean_cluster
M629656521	53.18051	0
M216016456	55.3781	2

+

clac_hlv_nm
4 (가구)
39 (완구)

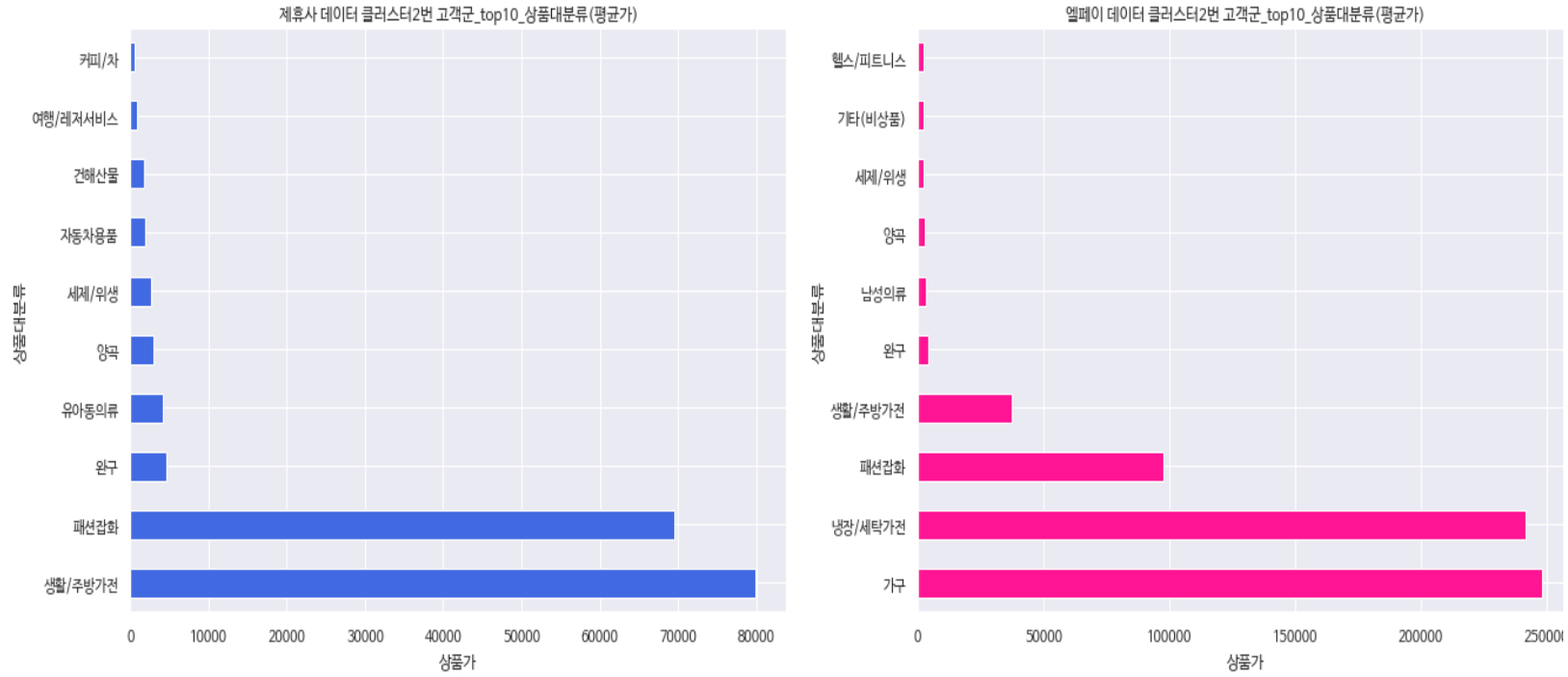
고객군 데이터를 추가한 LPAY / COP 데이터를 활용해 상품 추천을 진행한다.

학습 과정 및 데이터 전처리는 [고객 예측 모델] 과 동일하다.

05-2 상품 추천 모델

STEP 02. 상품 추천 결과

결과 분석 / 상품군 분석



제휴사, 엘페이별 상품을 추천한 후 평균 이용금액이 높은 공통된 상품들을 사용한다.
➔ 생활/주방가전, 패션잡화, 완구

06 마케팅 활용방안

방안 : 맞춤형 상품군을 기반으로 한 공간 마케팅 for MZ

구매력에 기반한
고객 군집 분류

고객군에 따른
맞춤형 상품 진열

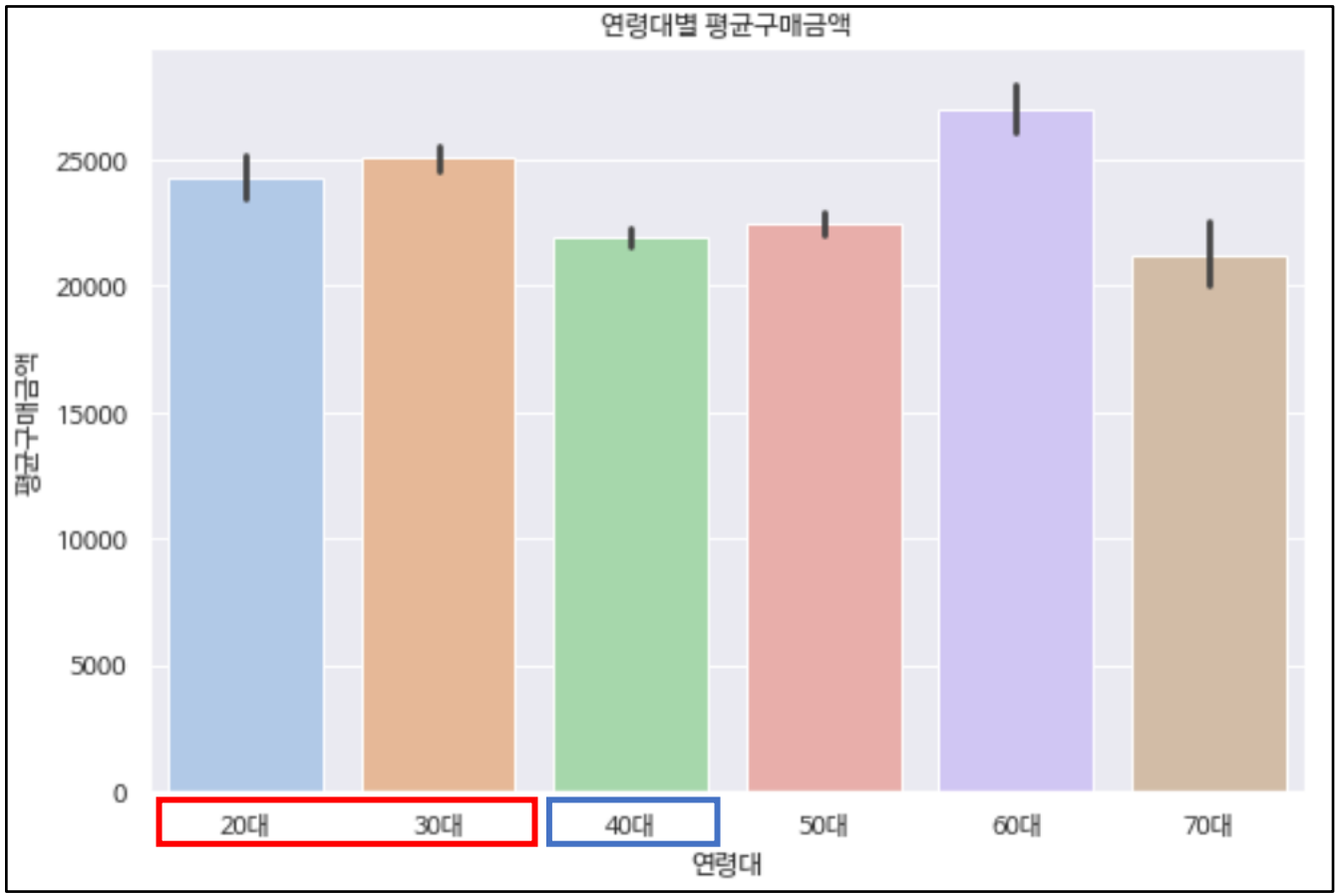
오프라인
공간경험 극대화

➔ 구매력에 기반해 **MZ 잠재 고객군에게 맞춤형 상품으로 구성된 공간을 제공할 수 있다.**

06 마케팅 활용방안

: 타겟층 - MZ세대

새로운 소비 권력으로 떠오르는 MZ 세대를 타겟팅한 마케팅 필요



40대 여성 고객은 총지출액이 가장 높지만 소비건수도 높다

이에 반해, MZ세대의 소비건수는 40대보다 훨씬 낮지만 '구매력'이 높다

이에 따라 MZ세대는 구매력이 높지만 아직 브랜드 충성도가 낮은 잠재고객으로 파악할 수 있다

06 마케팅 활용방안

: 기대효과

고객군 분류모델 + 상품군 추천모델 + 오프라인 공간

패션잡화

생활가전

완구

맞춤형 상품군으로 구성된 오프라인 팝업스토어 제공을 통해
MZ세대의 공간 경험 극대화 및 고객 충성도 확보

Thank You