

Retail Membership

고객-상품 micro segmentation을 통한 최적 데이터 마케팅

2018. 12. 15

TEAM 1

김정규 노현호 신호준 안중호 전해진

INDEX

1. 분석의 배경과 데이터 구성
2. 가설/EDA를 통한 분석 과제 구체화
3. 행태-시간 기반 고객 군집
4. 상품 기반 추천 모델과 성능 검증
5. 결론 및 향후 과제

[1]

분석의 배경과 데이터 구성

분석 배경 및 과제 설정

전반적으로 내수 화장품 업체의 매출이 감소하는 가운데
온라인/소셜 마케팅 중심 기업들이 성장하며 기존 고객 매출은 하락
본 분석의 목적이 된 A 브랜드의 경우 기존 회원 매출이 13.1% 감소



2018년 6월 기준, 이전 30개월 회원 데이터에서 구매 행동에 영향을 주는 요인 검출
고객 세그멘테이션을 파악하여 재유입 유도, 이탈 방지 등 매출 향상 전략 수립

분석을 위한 데이터 테이블 생성

16년 1월 ~ 18년 6월, 30개월 기간에 대한 구매내역, 회원정보 수집 및 상품분류 통합
총 660만건의 데이터를 확보하여 분석에 필요한 기본 테이블 생성

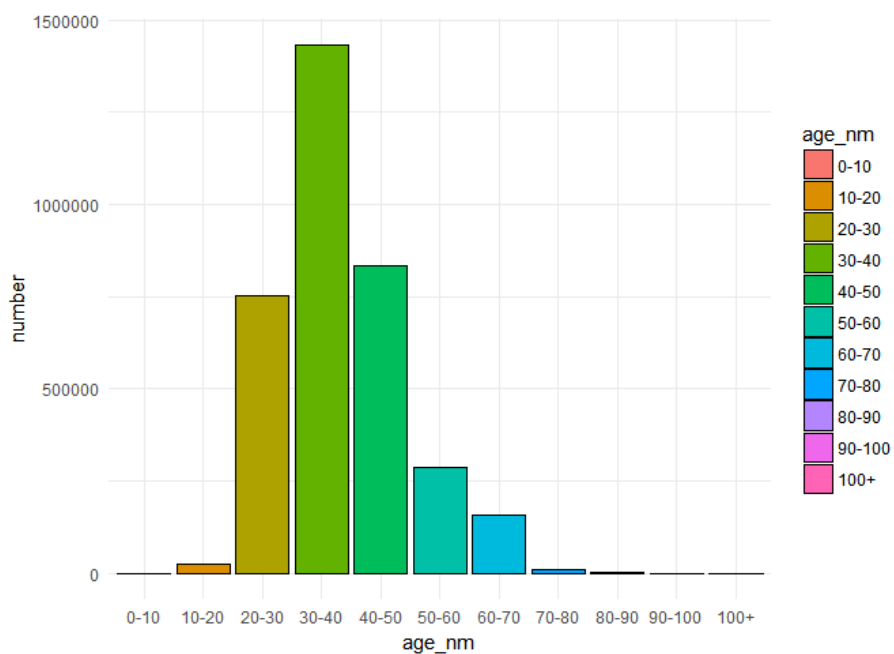
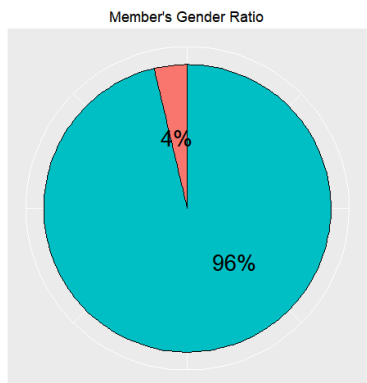
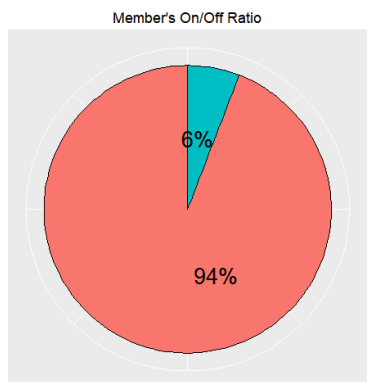
구분	영문명	한글명	상세 설명
상품구매 (Product)	custid	고객ID	■ 회원가입된 고객 ID 실제값은 아님, 영수증번호에 매겨진 구분 식별 번호를 근거로 매겨진 가상의 고객별 Unique ID 값
	date	결제일자	■ 결제완료일자 / 기간 : 2016.01 ~ 2018.06.30
	on_off	오프라인/온라인	■ 구매한 상품을 온라인 매장 구매인지, 오프라인 매장 구매인지 채널별로 나눔
	prod_code	상품코드	■ 구매한 상품의 코드(최소단위)
	amt	상품구매매출	■ 구매한 상품 총 매출
	qty	상품구매수량	■ 구매한 상품의 수량
회원 (Custom)	custid	고객ID	■ 회원가입된 고객 ID 실제값은 아님, 영수증번호에 매겨진 구분 식별 번호를 근거로 매겨진 가상의 고객별 Unique ID 값
	grade	등급	■ 고객등급정보
	sex	성별	■ 성별정보 [남자 : M / 여자 : F]
	age	연령대	■ 연령대 정보 [10대 : 10 / 20대 : 20 / 30대 : 30 ...]
상품분류 (Division)	prod_code	상품코드	■ 구매한 상품의 코드(최소단위)
	prod_nm	상품명	■ 구매한 상품의 상품명
	cate	상품 대분류	■ 구매한 상품의 속성중 가장 크게 구분한 대분류 ex) 스킨 / 헤어 / 네일케어 / 메이크업 / 액세서리
	cate_ftn	상품 중분류	■ 구매한 상품의 속성중 기능에 의해 구분한 중분류 ex) 클렌징 / 에센스 / 스크럽 / 마스크
	cate_line	상품 라인분류	■ 구매한 상품의 향기나 원료에 따라 분류한 카테고리 ex) 바나나향 / 딸기향 / 레몬향 / 올리브오일

[2]

가설/EDA를 통한 분석 과제 구체화

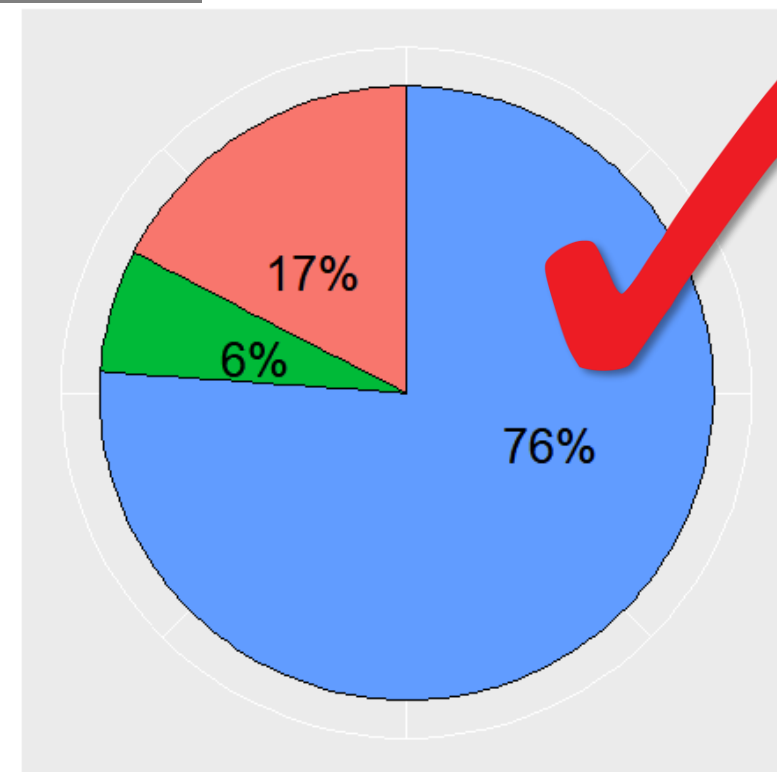
분석 방향 설정

단순 “누적매출액”에 의거한 기존 3단계 멤버십제도 개편 필요
구매 증가/감소 요인 파악 실시, 고객등급 재분류 및 세분화를 통한 추천모델 생성



기존 고객등급

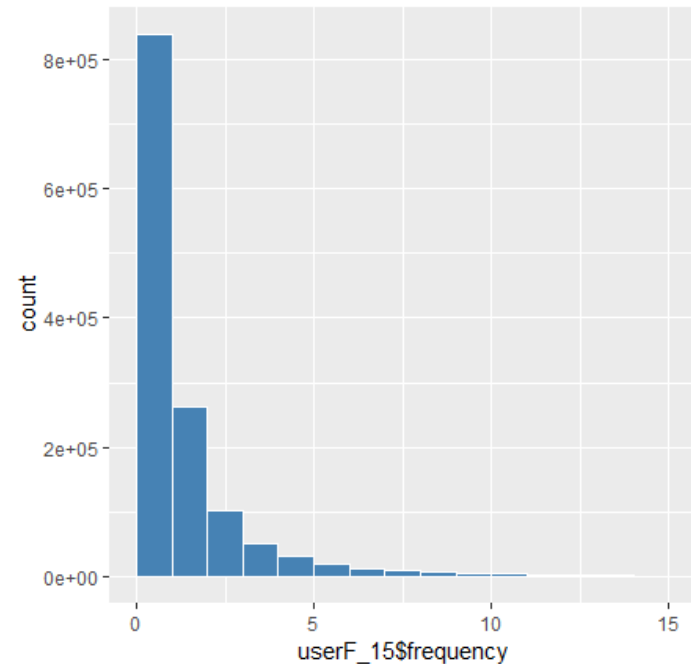
Membership Grades Ratio



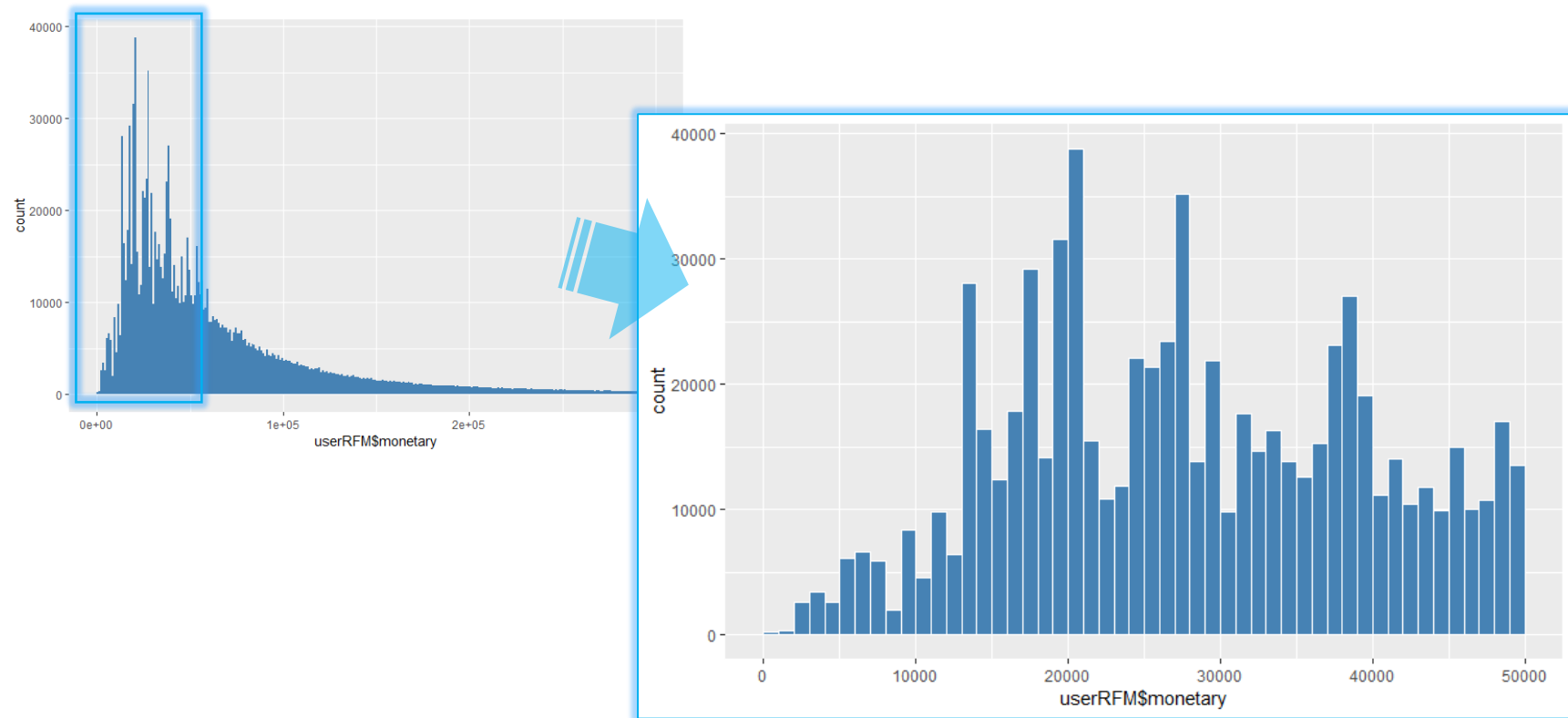
이상치 제거

1년에 7번 구매가 최고 등급 유지 조건, 30개월 15회 이상 구매건 이상치로 제외 (평균 고객은 연 2회 구매)
100만원 이상은 기업특판/개인특판 주문건에 해당, 고객분석에 적합한 주문이 아니므로 제외

총 15회 이상 방문건 제외 (일기준)

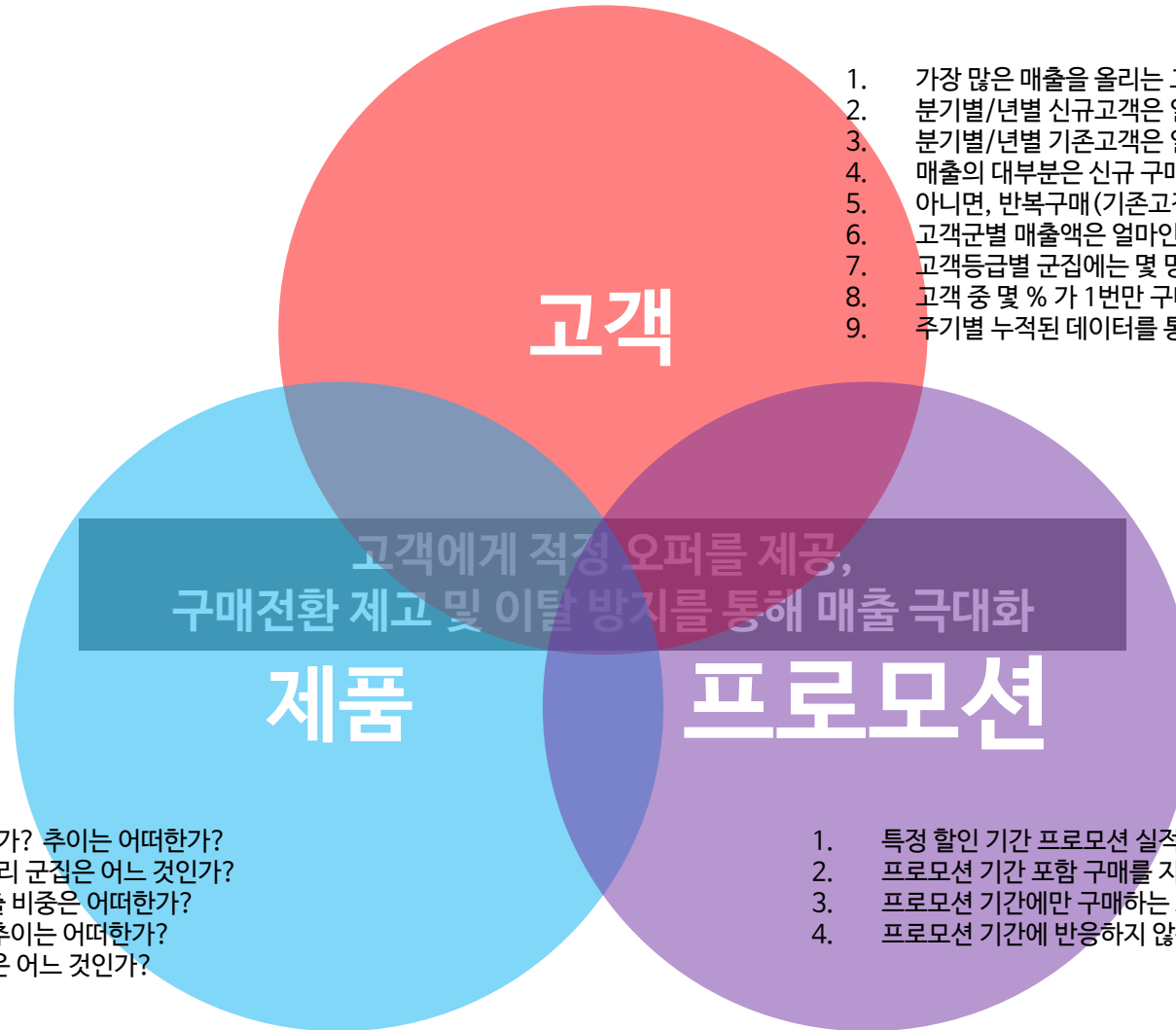


회당 100만원 이상 구매건 제외 (일기준)



#고객들은 2만원에서 2만1천원의 범위의 지출을 가장 많이 하였음.
#고객들은 2만7천원에서 2만8천원의 범위의 지출을 두 번째로 많이 하였음.
#고객들은 2만9천원에서 3만원의 범위의 지출을 세 번째로 많이 하였음.

분석을 위한 비즈니스 가설



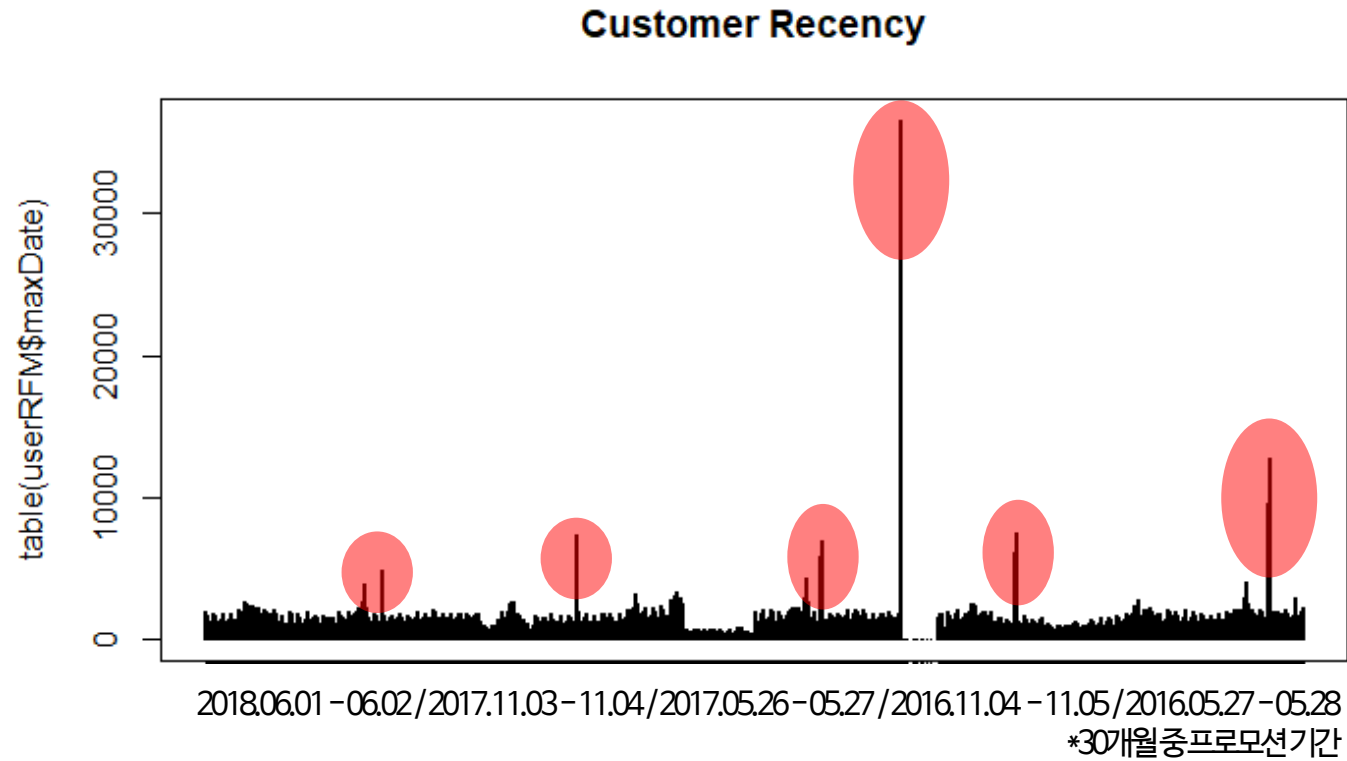
1. 가장 많은 매출을 올리는 고객등급(누적된 고객등급 패턴 반영)은 무엇인가?
2. 분기별/년별 신규고객은 얼마인가?
3. 분기별/년별 기존고객은 얼마인가?
4. 매출의 대부분은 신규 구매자로부터 발생하는가?
5. 아니면, 반복구매(기존고객)로부터 발생하는가?
6. 고객군별 매출액은 얼마인가? 추세는 어떠한가?
7. 고객등급별 군집에는 몇 명의 고객이 있는가? 매출 기여도는 어떠한가?
8. 고객 중 몇 % 가 1번만 구매한 고객인가?
9. 주기별 누적된 데이터를 통해 고객의 등급변화 확인

1. 제품-카테고리 군집에는 몇 명이 있는가? 추이는 어떠한가?
2. 가장 많은 매출을 올리는 제품-카테고리 군집은 어느 것인가?
3. 카테고리별(스킨, 네일케어, 헤어) 매출 비중은 어떠한가?
4. 각 라인별 군집에는 몇 명이 있는가? 추이는 어떠한가?
5. 가장 많은 매출을 올리는 라인별 군집은 어느 것인가?

1. 특정 할인 기간 프로모션 실적을 기준으로 등급변화 구매 빈도 연관성 확인
2. 프로모션 기간 포함 구매를 지속하는 고객
3. 프로모션 기간에만 구매하는 고객
4. 프로모션 기간에 반응하지 않는 고객

프로모션 효과 진단

프로모션시 평균 6배 매출 증대효과
고객재분류 모델에 프로모션 연관파생변수 도입



프로모션 파생변수 추출

30개월간 총 5회 실시한 프로모션을 기준으로 고객 행동 분석
변수 추출 및 EDA, 향후 프로모션 연관 고객관리 마케팅에 반영



프로모션기간 포함
구매를 지속하는 고객



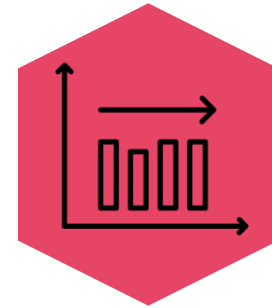
세일 외기간에도 지속적으로 구매되는 제품을
전략제품으로 추천
세일기간 소액 구매 고객 66%가 상시 방문
세일 정보 대신 추천 제품 정보 제공



세일기간에만 구매하는 고객
(가격혜택에 민감)



매세일진행시 재유입 유도가능
매세일기간 모니터링 및 관리 필요
세일정보빠짐없이 E-mail/앱푸시/SMS/TM



세일에 반응하지 않는 고객
(필요시에만 자연구매)



평상시 구매 3회 이상이지만
세일기간 구매 없는 고객
보다 강력한 고객접촉을 통해 구매 횟수 확대

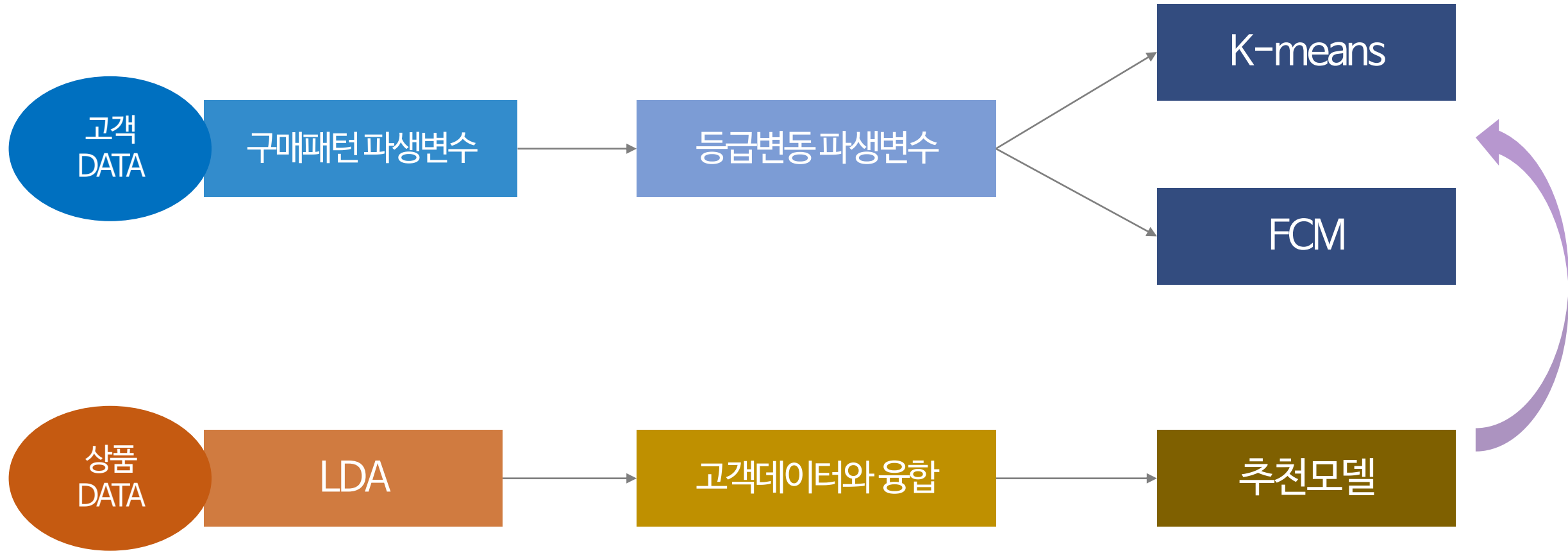
Bottom-Up 이탈 고객 판단

- 1) 세일 기간에만 구매하는 고객: 가격 혜택에 민감하므로 별도 관리
- 2) 세일에 반응하지 않는 고객: 필요시 자연 구매하므로 별도 관리
- 3) *1년간 구매하지 않는 고객: 이후 자연 구매 발생, 이탈로 정의하기 어려움



이탈을 **작위적으로 정의하기 보다는 실질적 이탈 파악이 필요
데이터에 기반, 우수/유지/이탈위험/이탈 고객군을 분류하여 세분화 전략 수립

분석 시나리오 구체화

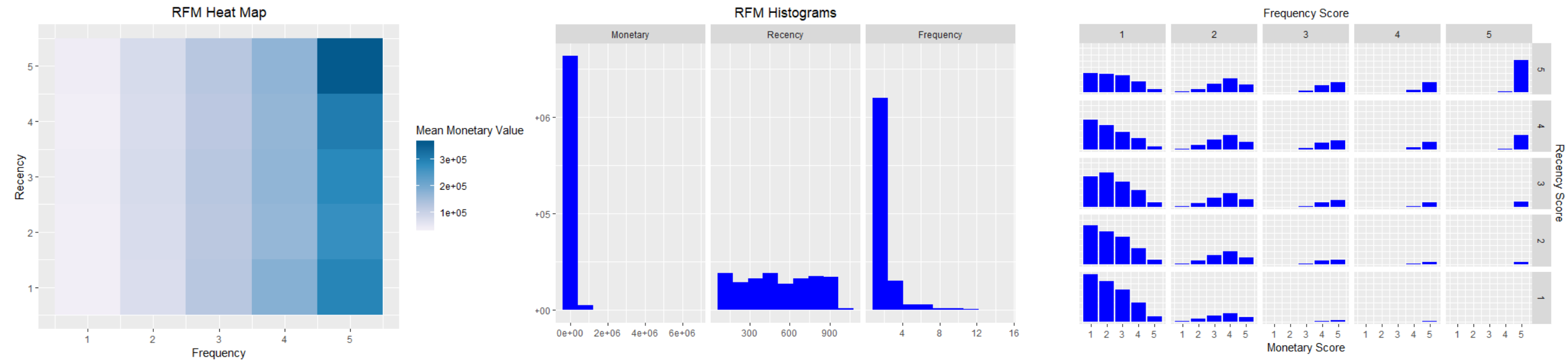


[3]

행태-시간 기반 고객 군집분석

구매패턴 파생변수 도출

최근구매(Recency), 구매빈도(Frequency), 구매액(Monetary)기준 5분위 등급으로 지수화
F의 영향력이 크므로 0.8을 4등급으로 규정, 0.05 구간으로 재정의 (최고등급은 최상위 5%)



*F4점 미만인 경우는 R이 크게 상관없으나 F가 5점이상이 되면서 R이 높을수록 M도 함께 높아지는 경향성 발견

등급변동 파생변수 제작

비즈니스 관점에서 고객 등급 관리가 최우선
Micro-clustering을 위해 등급 변동에 대한 파생변수 정밀 제작

upcnt

고객의 등급이 올라간 횟수

down_cnt

고객의 등급이 내려간 횟수

g_mean

고객의 등급의 평균

start_grade

고객의 시작 등급

end_grade

고객의 마지막 등급

grade_sd

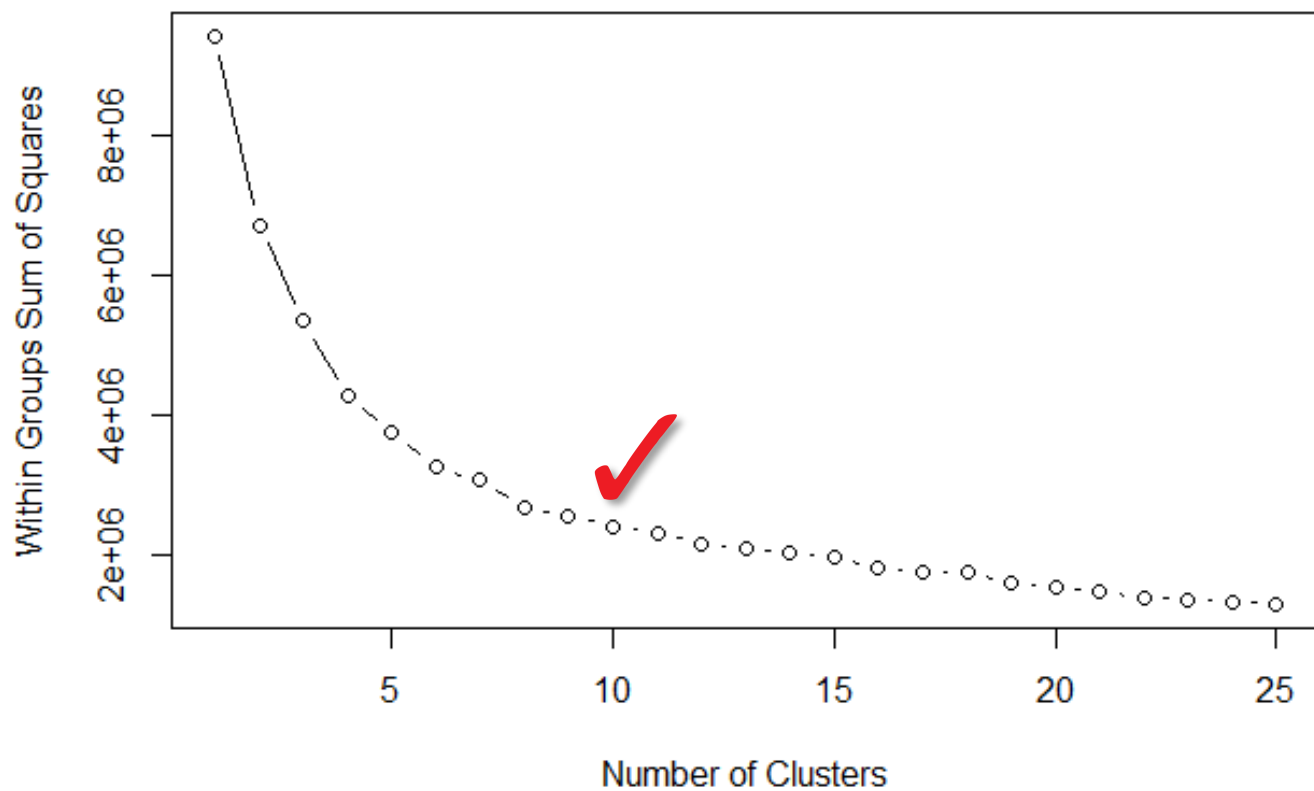
고객 등급 변화의 표준편차

e_cnt

연속적인 E 등급의 갯수

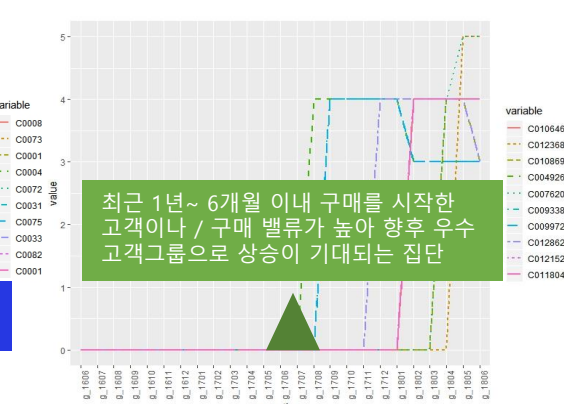
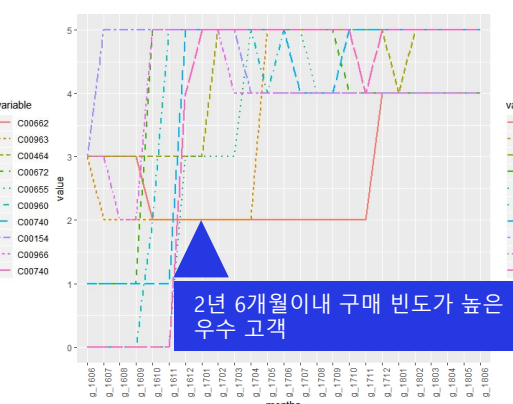
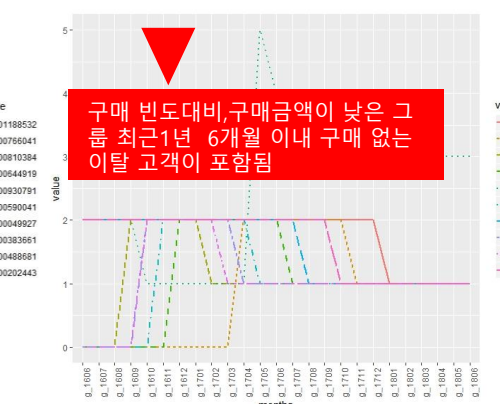
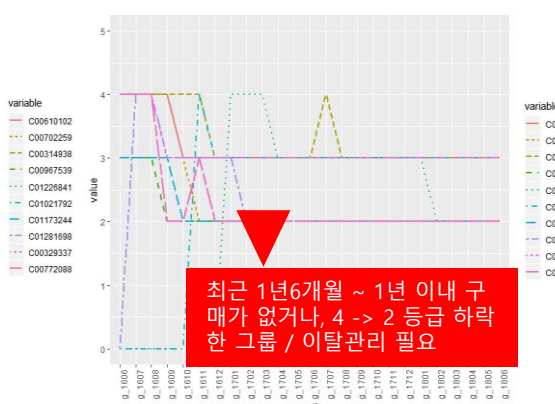
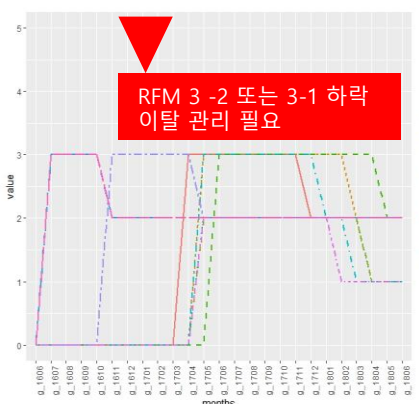
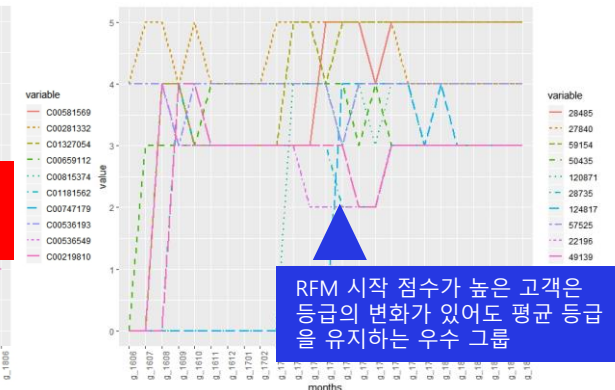
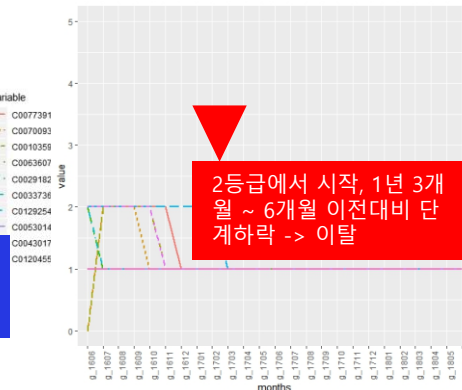
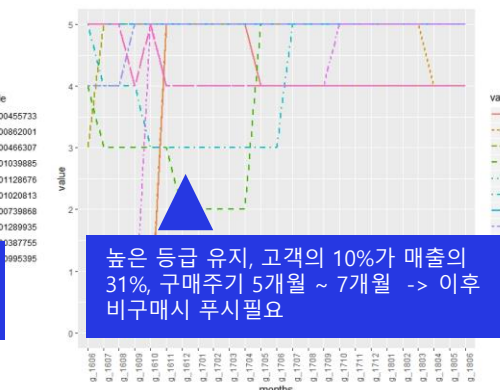
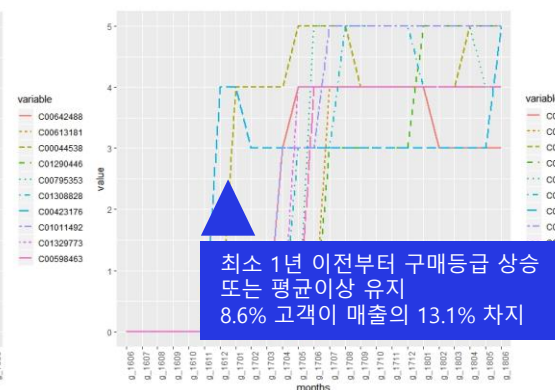
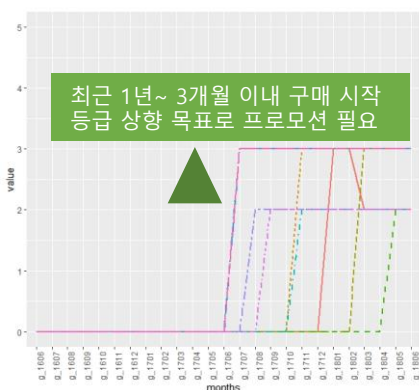
K-means 고객 군집 알고리즘

Elbow Method를 통해 K를 10으로 설정, 확연히 구분되어지는 경향 발견
K가 10개 이상인 경우 유사한 변화패턴 등장하여 군집 수 증가에 대한 실익이 없음



K-means 고객 군집 결과

각각 우수, 기대, 위험 고객집단으로 판단되는 총 10개의 군집 생성
이탈의 경우, 형태가 다른 4가지 패턴을 보이므로 세분화 전략 수립 가능



K-means 고객 군집 정량 분석

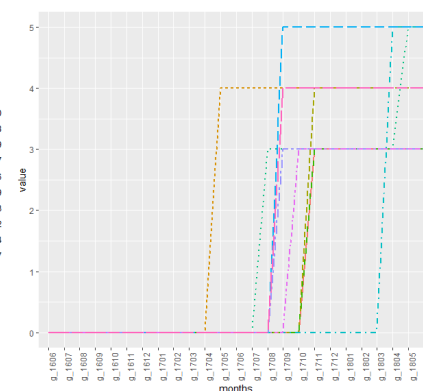
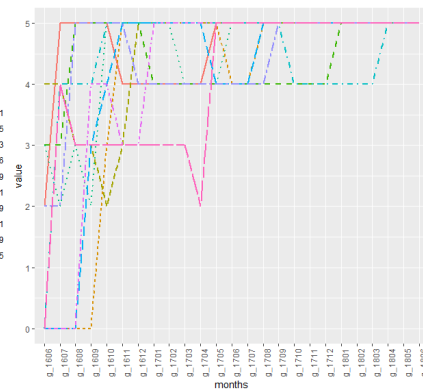
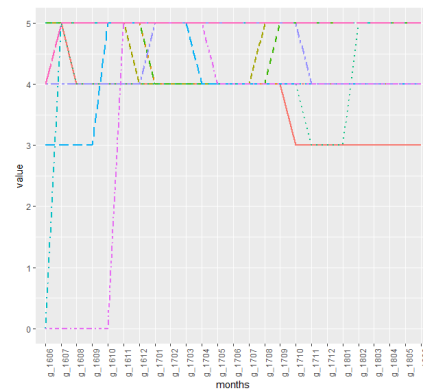
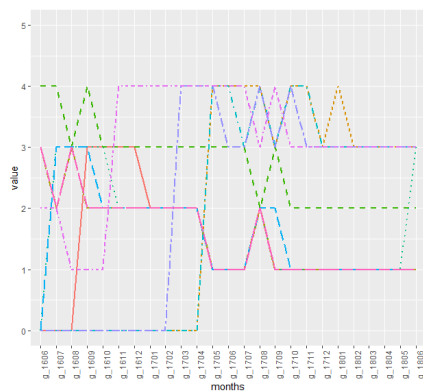
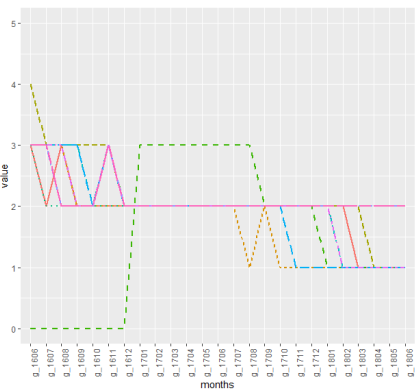
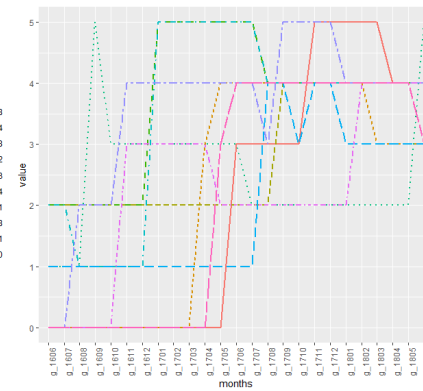
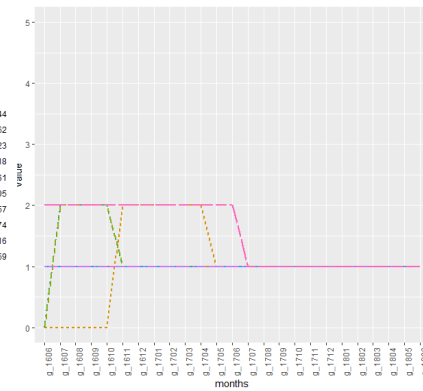
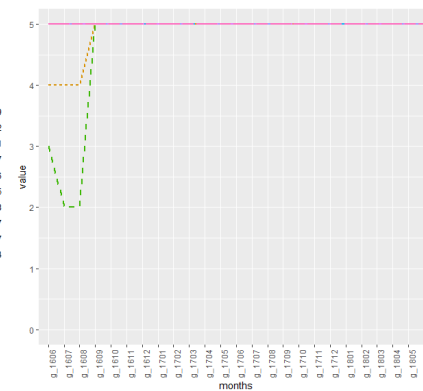
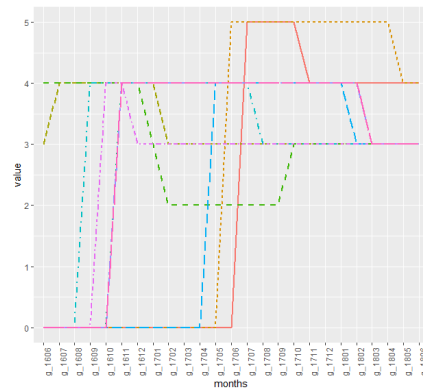
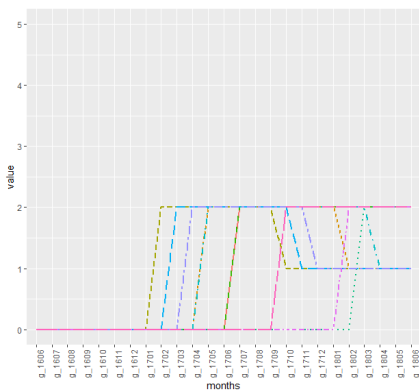
고객등급을 총 10개로 분류하여 매출비중과 객단가 파악
고객의 로열티 강화 및 이탈 방지 전략 수립에 활용

평가	군집	고객비중	매출비중	LTV, ₩
최우수	3	10.1%	31.6%	246,978
우수고객	9	9.1%	15.1%	131,103
우수가망고객	10	9.5%	9.9%	81,882
우수고객	2	8.6%	13.1%	120,076
우수고객	5	10.8%	9.1%	66,858
모니터필요	1	10.8%	3.2%	23,206
이탈관리필요	4	10.1%	2.6%	19,970
이탈관리필요	6	9.5%	3.3%	27,513
이탈관리필요	8	7.7%	2.2%	22,435
이탈고객	7	13.7%	9.9%	57,102

*기간 30개월, LTV : LIFE TIME VALUE (객단가 * 구매횟수) ₩78,962

비교/검증 : fuzzy c-means 분석

K-means와 유사한 군집이 도출되어 신뢰성을 확인함
계산량(3.26sec /67hours) 문제로 k-means를 주요 알고리즘으로 선택



Who,
When,

What?



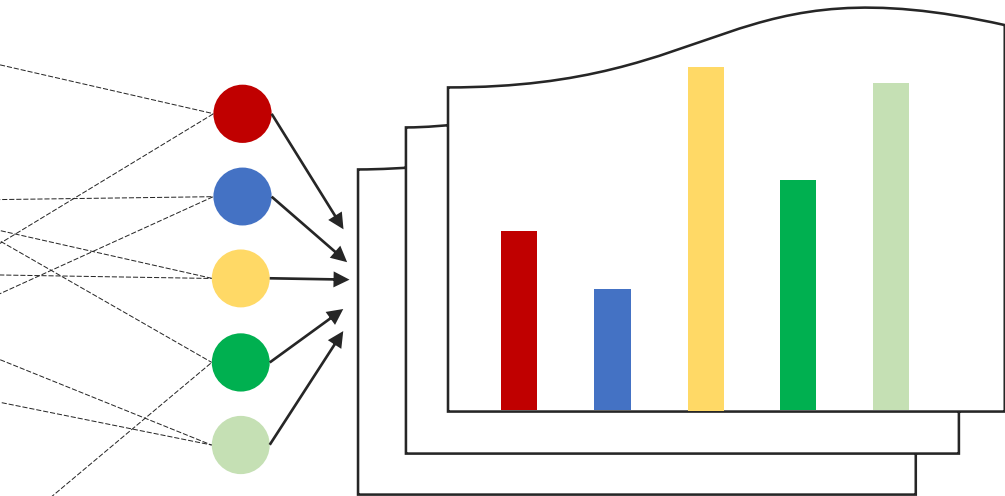
[4]

상품 기반 군집 알고리즘과 추천 모델

구매 상품별 고객 군집

품목에 대한 확률분포인 “토픽모형”을 활용, 마이크로세그멘테이션 실시
해당 토픽에서만 특별히 확률이 큰 상품을 주요 상품으로 *판단

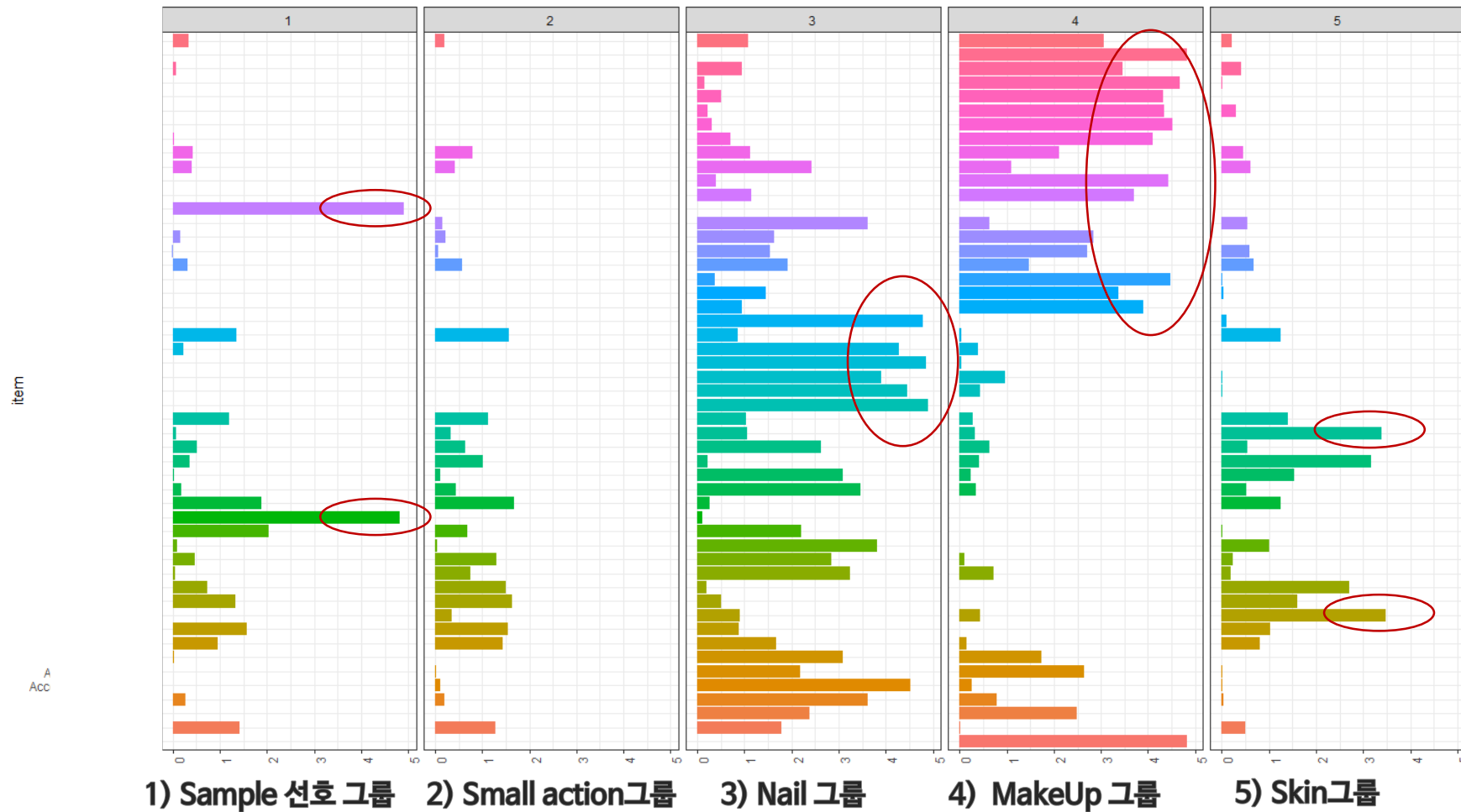
- C00000001 - XXXXXXXXXXXX, AAAAAA, CCCCCC
- C00000002 - VVVVVV, WWWWWW, TTTTTT
- C00000003 - EEEEE, UUUUUU, HHHHHHHHHHHH
- C00000004 - TTTTTTTTTTTTTT, UUUU, BBBB
- C00000005 - NNNNN, PPPPP, CCCCCC
- C00000006 - KKKKKKKKK, NNNNNNNNN, GGGGGGG, VVVVVVVVV
- C00000007 - TTTTTTTTTT
- C00000008 - JJJJJJJJJJJJJJJJJJJJJ, CCCCCCCCCCCCCC, SSSSSSSSS
- ...
- C01344296 - DDDDDDDDDDD, XXXXXXXXXXXX
- C01344297 - VVVVVVVVVVV, RRRRRRRRRRRRRRRR, HHHHHHHHHHHH



*토픽에서 상품의 확률을 쇼핑 자료 전체에서 해당 상품이 차지하는 비율로 나눈 값 기준
(김용대, 토픽모형을 이용한 빅데이터 기반 마이크로세그멘테이션 방법론 연구)

LDA 군집 결과

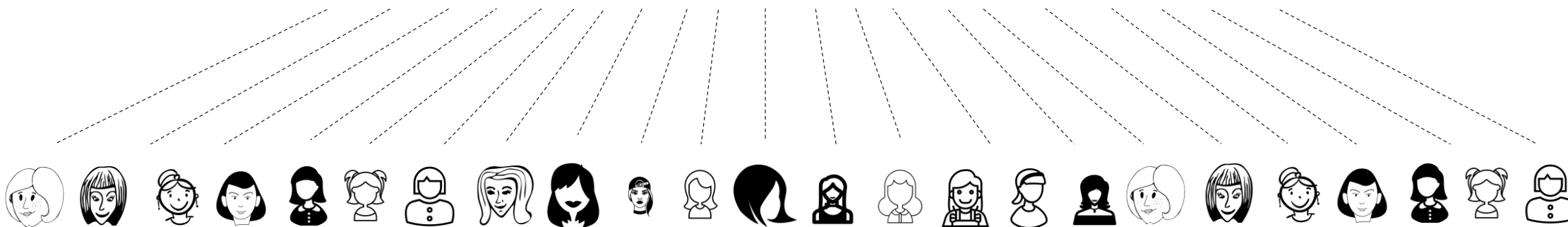
뚜렷이 구분되는 5가지의 상품군 도출, 각 그룹을 비즈니스 관점에서 재정의



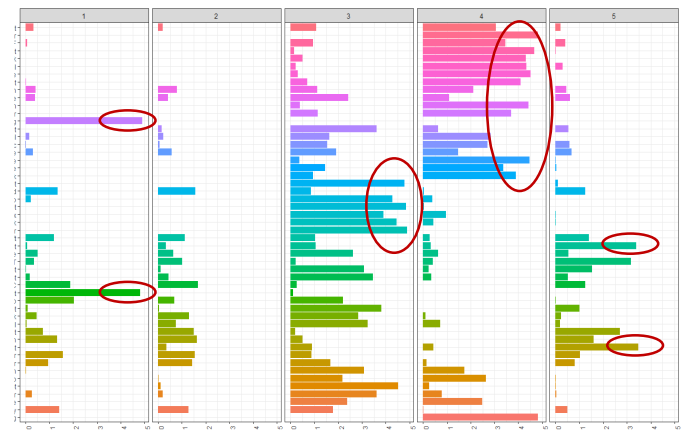
LDA (상품 데이터를 군집) + 고객 데이터와 융합



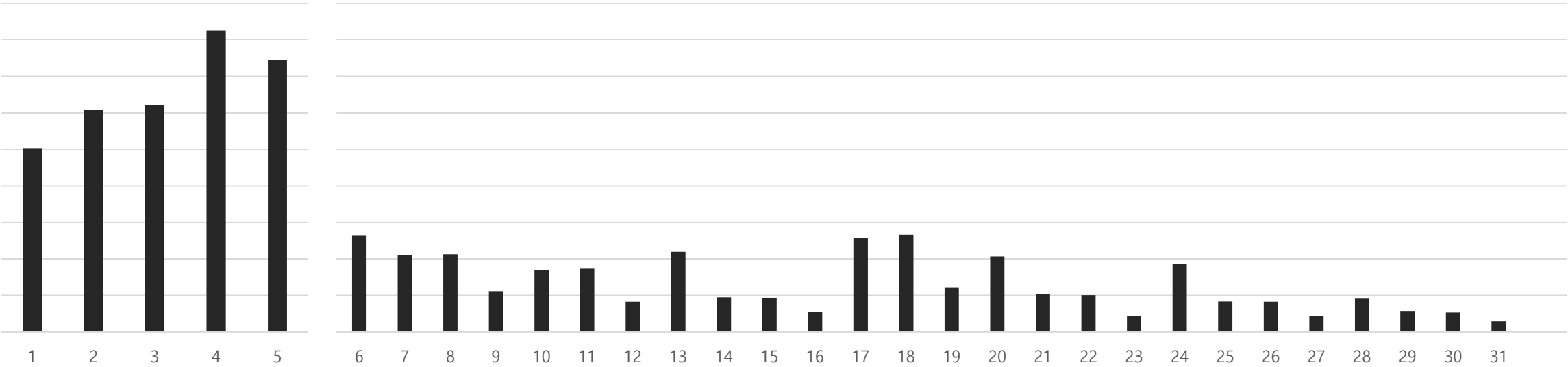
1) Sample 선호그룹 2) Small action그룹 3) Nail 그룹 4) MakeUp그룹 5) Skin그룹



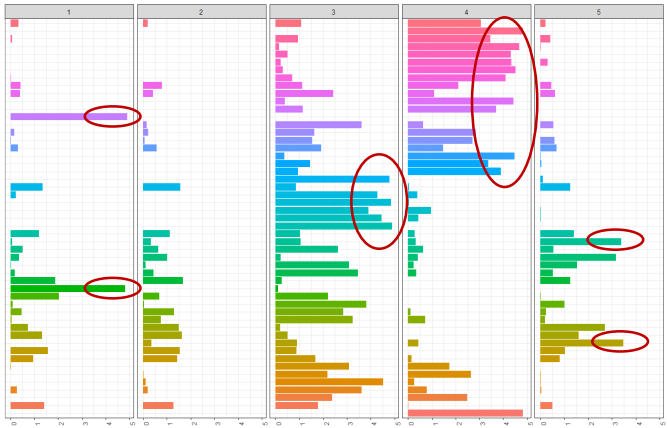
LDA 군집 고객 분포도



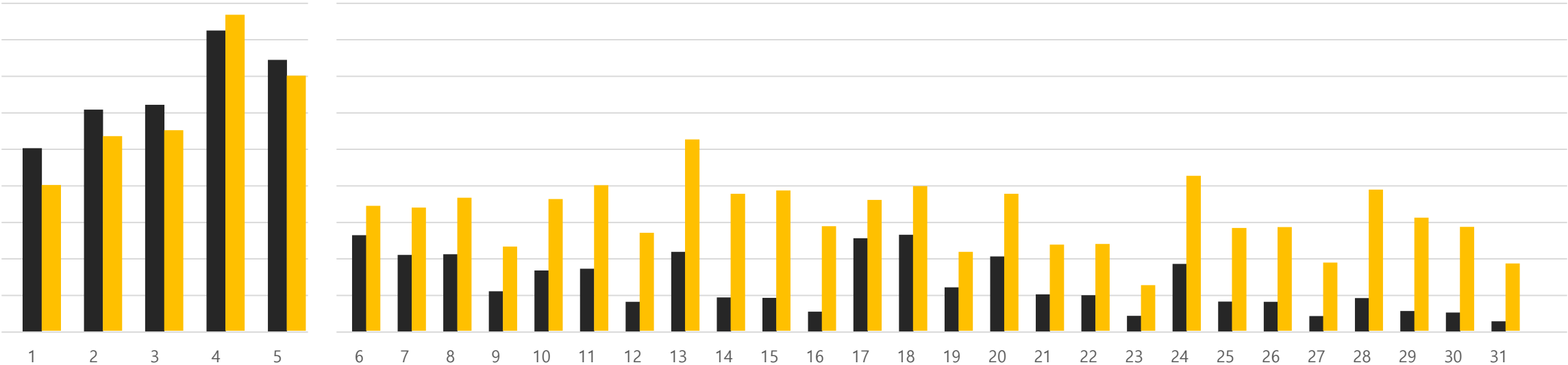
2⁵ - 1 = 31개의 집단.
1~5 집단 고객 인원: 66만
6~31 집단 고객 인원: 68만



LDA 군집 고객매출 분포도



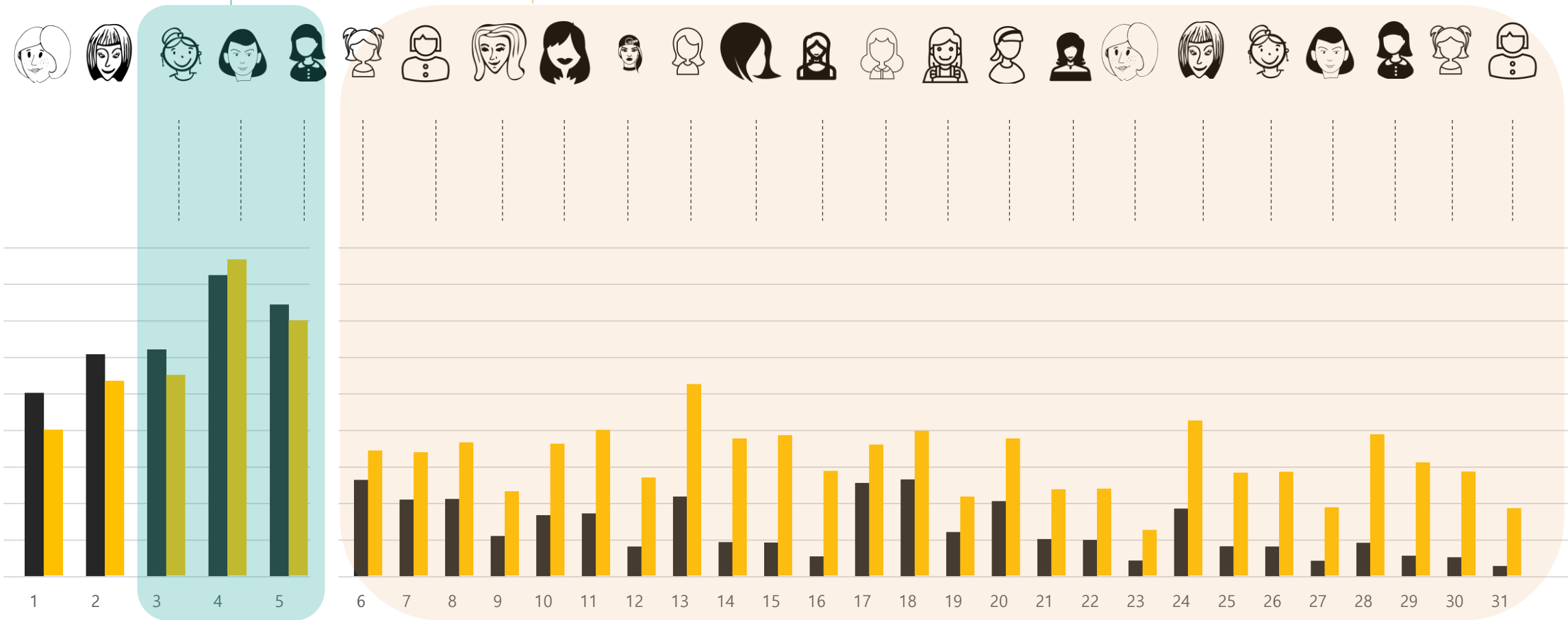
*교차카테고리구매자들이인원대비구매매출이 훨씬 높다.



LDA 군집 고객의 특성

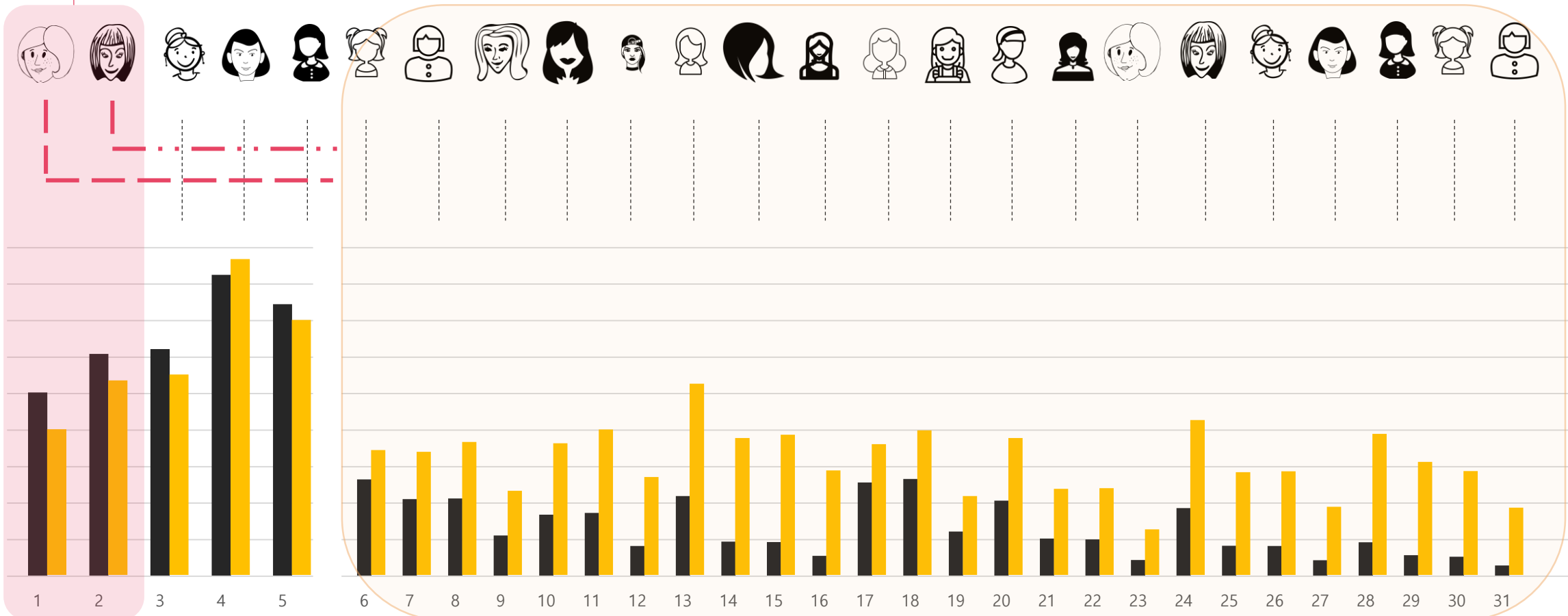
LDA3~5집단은 자기추향 뚜렷 (헤어면 헤어, 네일이면, 네일향수면 향수) A브랜드에서 본인의 구매 품목만 선택

LDA6~31집단은 인원 대비 매출 높음, 본인추향의 제품만 선택하며 A브랜드 제품 이해도가 높음.



LDA 군집 고객의 특성

1~2: 샘플링을 구매하거나 소극적인 신규 고객 집단
샘플링하는 제품의 카테고리에 대한 추가 탐색 - 크로스 셀링 카테고리 추천 필요



추천모델 검증 프로세스

〉 기존 추천 알고리즘



- 고객 별 의 전체 구매품목을 기준으로 Random으로 Train, Test 데이터 Split.
- 시간 개념 적용되지 않음.

〉 개선 추천 및 검증 알고리즘



2 years (2016.01 ~ 2017.12)

6 month (2018.01 ~ 2018.06)

- 고객 구매 정보를 시간순으로 정렬 2016.01 ~ 2018.06 .
- 추천모델 구현 후 2016 ~ 2017년 구매 고객 데이터 기반으로 추천 리스트 구성, 실제 2018년 동일 고객 구매 정보와 일치율 비교.

성능 기반 추천모델 결정

〉 추천리스트 기반으로 가장 정확도 높은 모델 확인 (Precision, recall, F-score)

UserBase, Cosine 유사도

Precision : 0.0405 , recall : 0.0297 , F-score : 0.0343

UserBase, Pearson 유사도

Precision : 0.0405 , recall : 0.0297 , F-score : 0.0343

ItemBase, Cosine 유사도

Precision : 0.00 , recall : 0.00 , F-score : 0.00

ItemBase, Pearson 유사도

Precision : 0.00 , recall : 0.00 , F-score : 0.00

Binary UserBase, Jaccard 유사도

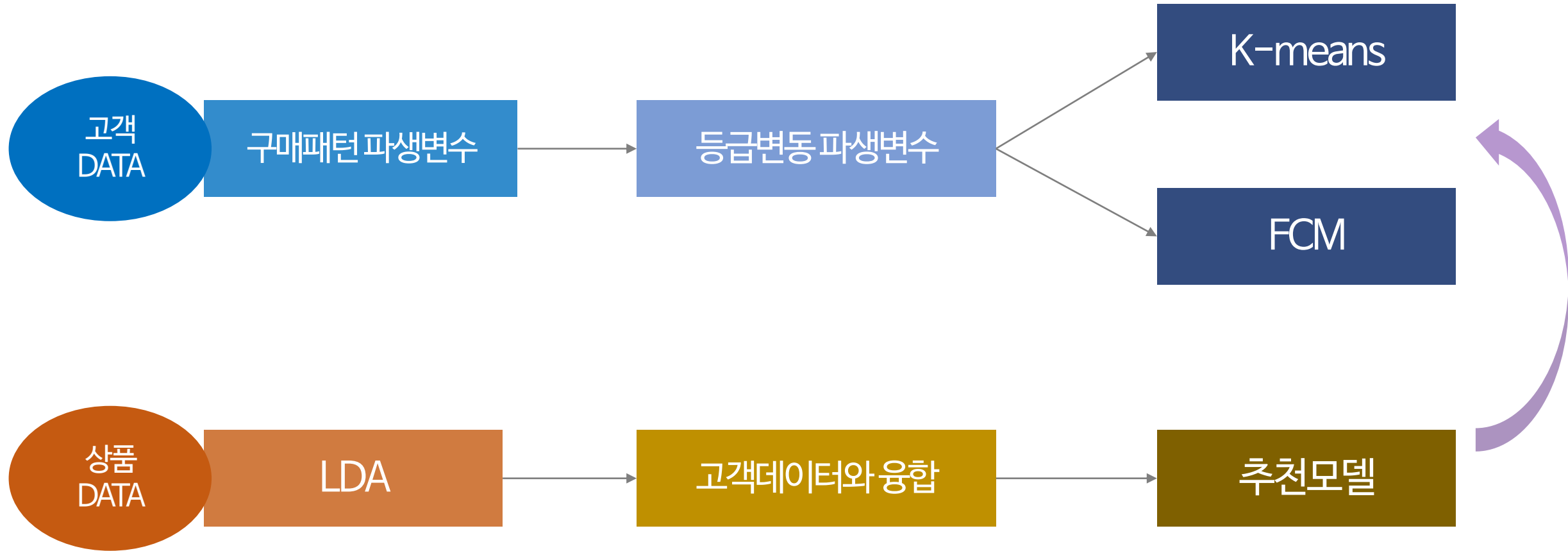
Precision : 0.1425 , recall : 0.1044 , F-score : 0.1205

Binary ItemBase, Jaccard 유사도

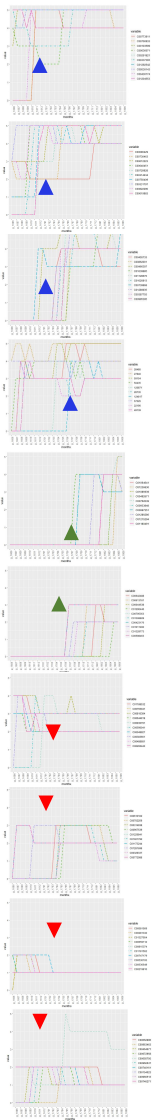
Precision : 0.0468 , recall : 0.0343 , F-score : 0.0396

* 평균 알고리즘의 precision은 0.05

타겟별 최적 시점에 추천상품 push



고객집단별 구매주기 파악, 실시간 세분화 마케팅



평가	군집	고객비중	매출비중	LTV, ₩	유형	고객관리
최우수고객	3	10.1%	31.6%	246,978	최장기간 상위등급유지	*매출기여도 31.6% 집단 년간세일참여비중이높음, 최우수 등급상황및유지하기위한WIP 프로모션필요, 개인구매주기, 선호에따른 지속적인제품추천
우수고객	9	9.1%	15.1%	131,103	구매빈도가가장우수	*매출기여도 28.2% 집단 최우수 고객집단으로 상황도기위한추천서비스 적용
우수고객	2	8.6%	13.1%	120,076	최근 1년이내 구매, 객단가 상위	
우수기망고객	5	10.8%	9.1%	66,858	등급변화가빈번, 상위등급유지	
우수기망고객	10	9.5%	9.9%	81,882	최근 8~3개월이내 구매, 객단가 상위	
우수기망고객	1	10.8%	3.2%	23,206	최근 12~3개월이내 구매시작	*매출기여도 22.2% 집단 등급하락시점이전개인푸시필요 개인별구매주기파악, 구매 객단가상승에필요한쿠폰발급
이탈관리필요	7	13.7%	9.9%	57,102	최근 18개월~12개월이내 구매가없거나, RFM4->2등급하락	*매출 18.5% / 고객모수 41.0% 고객별이탈제품및이탈주기에따른해택마련, 12개월이전구매전 환유도필요, 함께구매단가상승위한전략필요 (추천아이템을강력한할인율로 적용하여 정보제공, 복귀유도)
이탈관리필요	6	9.5%	3.3%	27,513	12개월중 1회구매빈도, RFM2등급하락	
이탈고객	4	10.1%	2.6%	19,970	RFM2등급에서시작, 1년3개월~6개월이전대비하락 ->이탈	
이탈고객	8	7.7%	2.2%	22,435	18개월이내 구매없음	

*기간 30개월 평균고객당객단가 ₩78,962

[5]

결론 및 향후 과제

프로젝트의 의미와 향후 과제



본 프로젝트에서는 2016 ~ 2017년 구매 고객 데이터 기반 추천 리스트를 2018년 상반기 동일 고객 구매 정보와 비교하여 성능 확인
현실적 한계로 인해 '대과거'의 데이터를 통해 과거를 예측한 결과, 실제 비즈니스에 활용하기 어렵다는 한계 존재



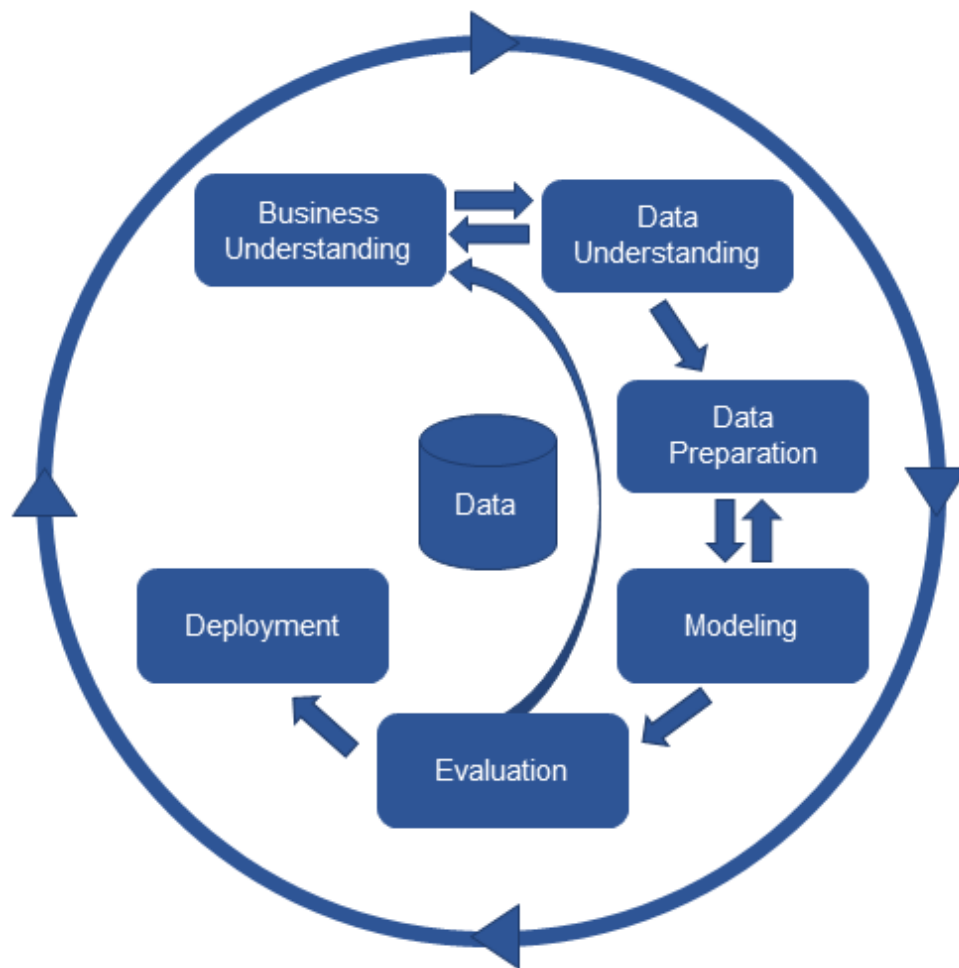
Next Step : 성능이 검증된 이번 알고리즘을 과거 30개월 데이터에 사용, 기 수집된 올 7~ 12월분 데이터와 비교해 성능 개선



Next Step 2 : 위의 향상된 알고리즘을 전체 36개월 데이터에 재적용하여 내년 1~ 6월을 예측, 선제적 마케팅 전략 수립 및 성능 재검증

보다 향상된 예측 알고리즘

연령대별, 성별, 카테고리별, 이용채널, 이용지역 등 변수 추가
외부데이터 도입 등을 통해 모델 성능 향상, 지속적비즈니스개선구조 확립



A top-down view of a business meeting around a wooden table. Several people are seated around the table, their hands and arms visible. They are working with various documents and charts. One document features a pie chart with four segments in shades of blue. Another shows a bar chart with five bars of increasing height. A third document has a line graph with a red line showing an upward trend. There are also some smaller charts and a notebook with a list of items. Several white disposable coffee cups are scattered on the table. A smartphone and a tablet are also visible. The overall atmosphere is professional and collaborative.

Thank You!