제 5 회 L.POINT Big Data Competition

(category2vec 을 통한 소분류 군집 분석 및 새로운 제품 구매 유도 추천시스템 개발)

팀명 : 롯된놈들

김동규, 서원교, 신민용

Executive Summary

- 본 프로젝트는 고객별로 구매한 제품의 군집 분석을 시행하고 이를 바탕으로 고객과 군집의 연관성을 파악하고자 한다.

- 군집분석 결과를 바탕으로 고객이 해당 군집의 제품을 구매하는 고객인지 아닌지를 예측하며, 실제 영향력이 높은 변수를 생성하여 정확도를 높인다.

- 고객특성변수, 군집별 고객 특성 변수, 네이버 트렌드 변수를 통해 고객과 군집간의 연관성을 알 수 있는 일명 '온라인 군집 선호 지수'를 생성한다.

- 형성된 군집간의 거리에 따른 특성을 파악하고, 이를 활용하여 고객이 구매한 제품과 가까운 군집의 제품들을 추천해주는 시스템, 일명 '이건 어때'를 개발한다.

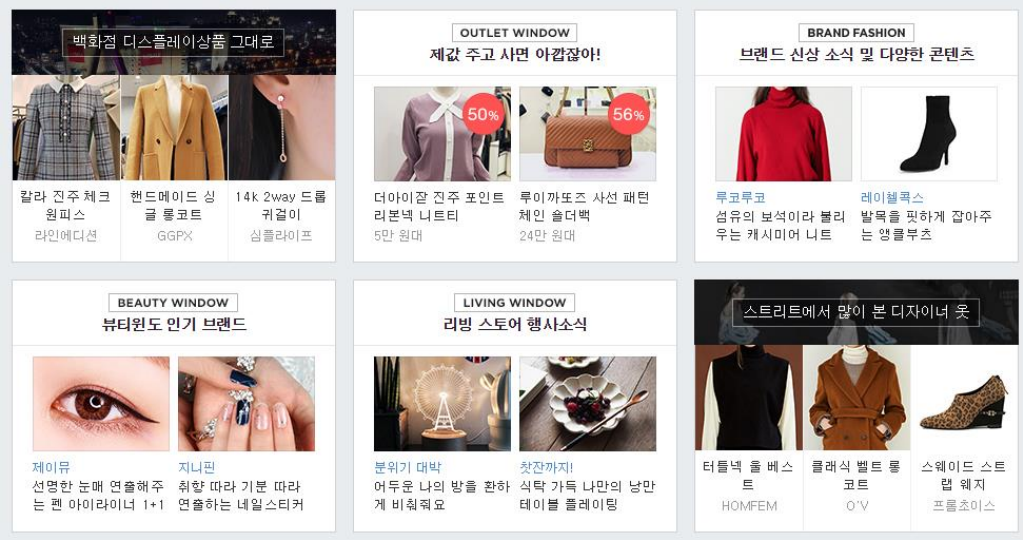
- 상품의 특성에 따른 군집 형성뿐만 아니라 고객 특성 맞춤 추천시스템 개발을 통해 고객에게 새로운 제품 구매를 유도한다.

목 차

1. 주제 선정 배경
2. 전체 프로세스
3. 데이터 수집 및 정제, 탐색
4. category2vec 활용 상품 군집화
5. 군집분석 및 지수 생성
6. 고객별 해당 군집 상품 구매 예측
7. 최종 서비스 제안
8. 결론

1. 주제 선정 배경

‘고객 입장에서’ 필요한 상품 추천의 필요성



<출처 : https://shopping.naver.com/ >

위는 최근에 들어서도 많은 고객들이 이용하는 ‘네이버 쇼핑’의 메인 중 일부이다. 언뜻 보기엔 별다른 문제가 없어 보이는 물품 추천 및 인기상품 목록이다. **다만, 중요한 것은 페이지에 접속한 ‘누구든’간에 위의 입력된 물품 목록만을 추천**한다는 점이다. 단적으로 본다면, 위의 추천들은 여성 이용자들에게는 제법 이목을 끌 수 있을지는 모른다. 단, 역으로 남성에게 있어서는 특별한 경우를 제외하고는 별다른 의미가 없을 거라 짐작할 수 있다. 심지어, 위의 화면은 남성인 작성자 본인의 네이버 계정으로 접속했을 때에 나타난 것이다.

인터넷 쇼핑몰에서는 대부분 ‘인기상품’들을 추천한다. ‘인기상품’은 그 자체로 이미 ‘인기가 있는’ 상품이기에, 추천하기에 더 없이 좋은 상품일 수 있다. 문제는 쉽게 생각해봐도 그 인기상품이 어느 누구에게든 필요한 것이라고 판단하기는 어렵다는 점이다.

<출처 : https://www.apple.com/kr >

간단한 예를 들어, 여기 말그대로 인기상품인 '에어팟'이 있다. 하지만, 그렇다고 모든 이용자들이 필요로 하지는 않는다. '에어팟'이 필요하려면, 일반적으로 '아이폰'을 이용한다는 가정이 앞서야한다. 물론 '에어팟'이 타 기업의 기종들과도 호환이 가능하다는 점을 고려한다 하더라도, 그 필요성에 대해서는 '아이폰'을 쓰지 않는 사람들은 크게 공감하지 못한다.

앞서 예를 들긴 했지만, 결국 '누가, 어떤 걸 필요로 하느냐?'에 대한 파악은 많은 고객들에 대해 정말로 아무것도 아는 것이 없다고 한다면 매우 어려운 문제다. 그러나, 여러 고객들의 수많은 구매정보들은 누적되고 있다. 그리고, 구매 데이터들에 대해 우리가 알고 있다면, 말그대로 이용하는 고객들에 맞추어 '필요할 만한 상품'들을 보다 적극적으로 추천해줄 수 있다고 여겼다.

'같이 산 것들'로 '필요할 것들'을 찾는다.

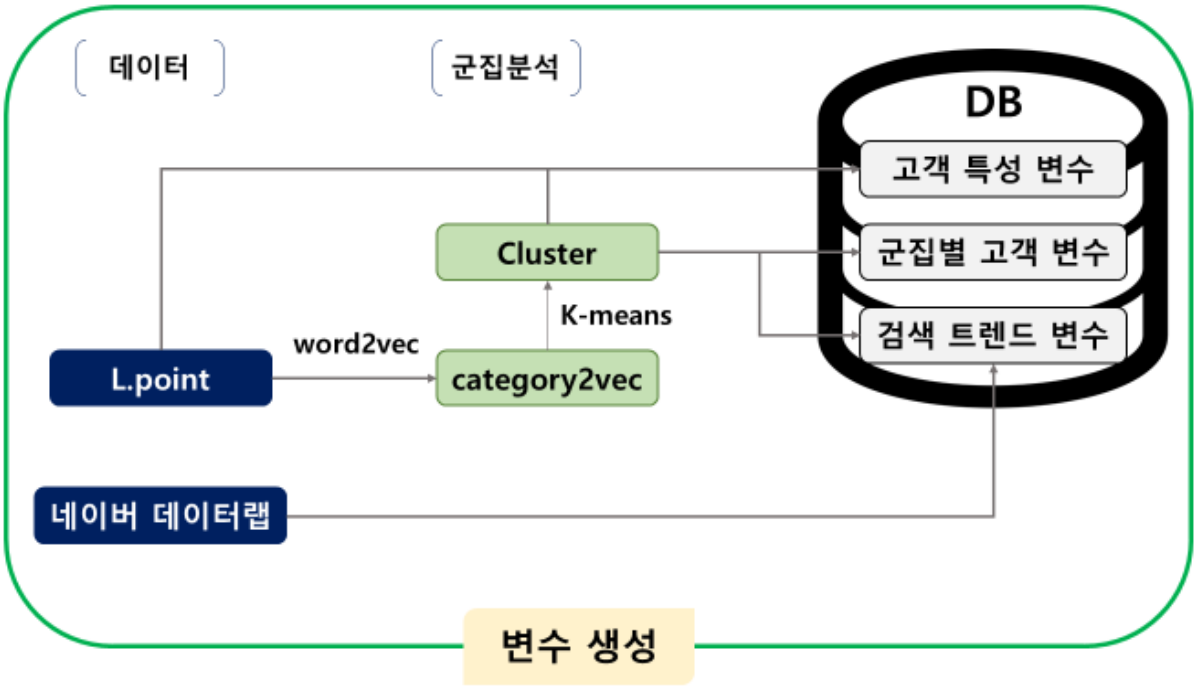
고객별로 구매하는 제품의 종류는 다양하며 그 조합은 무수히 많다. 어떤 고객은 전자기기를 선호해서 카메라, 태블릿 등의 제품을 구매할 것이고, 다른 고객은 어린 아이의 어머니로서 아동복이나 유아 장난감을 구매할 것이다. 수많은 조합의 특성을 파악할 수 있다면, 판매자는 이를 활용하여 효과적인 마케팅을 할 수 있을 것이다.

실제로 오프라인 매장의 경우, '장바구니 분석'을 통해 좋은 효과를 누린 사례가 적지 않다. 기저귀를 사러 오는 30~40 대 남성들이 맥주 6 캔을 함께 사는 것을 파악하고 두 상품을 가깝게 진열하거나, 편의점에서 동시에 자주 구매되는 음식 조합을 발견하고 이를 한 곳에 묶어 놓는 등이 이에 해당된다.

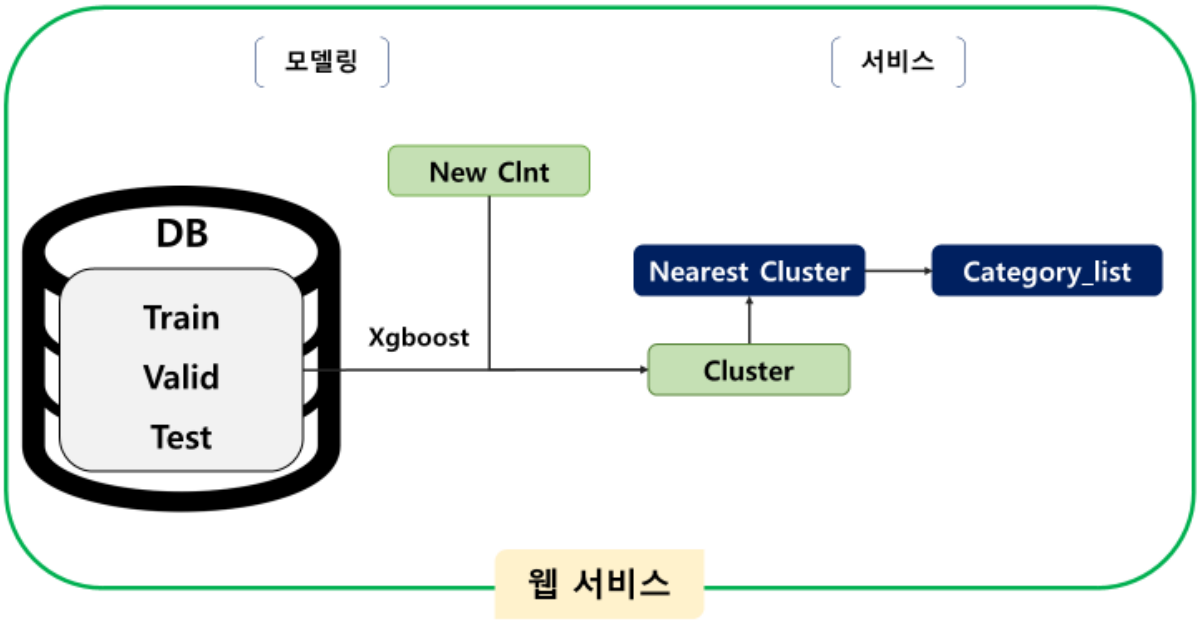
이러한 마케팅이 과연 오프라인에서만 통하는 방법일까? 온라인 구매 데이터를 통해서 역으로 그 사람의 특성 또는 선호 제품군을 파악할 수 있다. 카메라, 태블릿 등을 구매하는 고객들을 분석해 보면 전자기기를 선호하는 군집이라 할 수 있고, 아동복이나 장난감 등을 구매하는 고객들은 어린 아이가 있는 부모라고 생각 할 수 있다. **이에 착안하여 소분류의 구매 조합을 통해 새로운 군집 분석을 시행하고 이를 실제 마케팅에 적용하고자 한다.**

2. 전체 프로세스

(1) 군집분석, 변수 생성 OVERVIEW



(2) 수요 예측, 웹 서비스 OVERVIEW



3. 데이터 수집 및 정제, 탐색

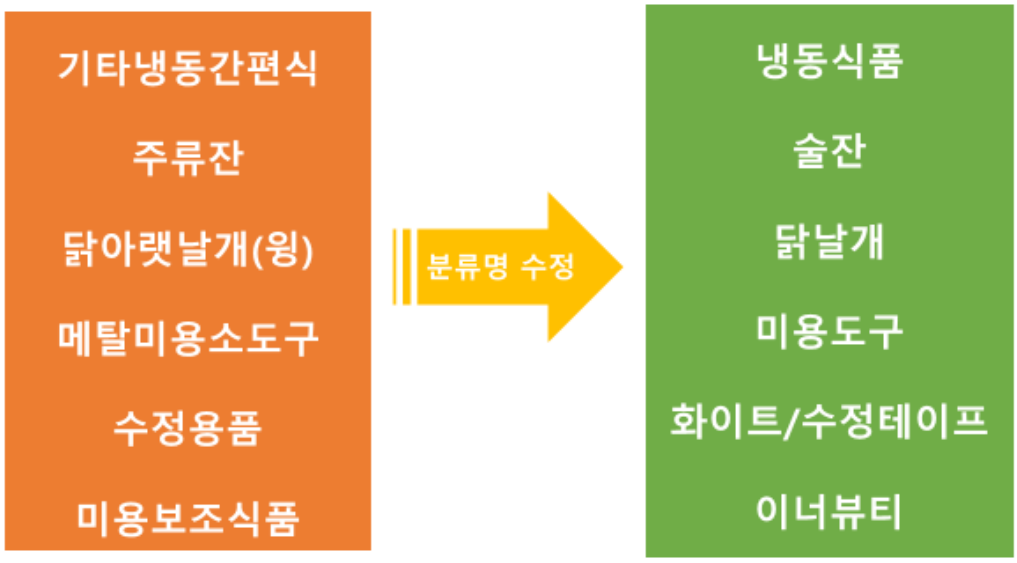
(1) 네이버 트렌드 API 를 이용한 월별 TREND 생성

< 외부데이터 출처 : https://datalab.naver.com/ >

	ctgr	period	ratio
0	욕실소품	2018-04-01	100.00000
1	욕실소품	2018-05-01	72.30538
2	욕실소품	2018-06-01	56.28742
3	욕실소품	2018-07-01	58.13373
4	욕실소품	2018-08-01	64.82035
5	욕실소품	2018-09-01	50.00000
6	선풍기	2018-04-01	16.57931
7	선풍기	2018-05-01	67.12119
8	선풍기	2018-06-01	97.03643
9	선풍기	2018-07-01	100.00000

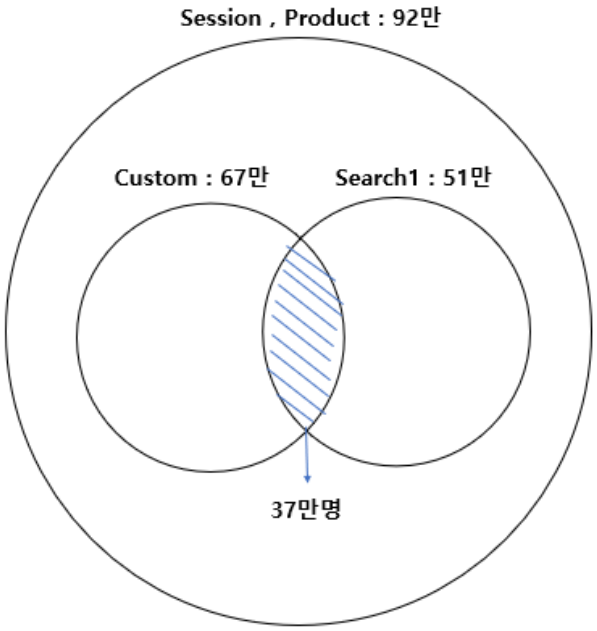
period	2단우산	3단우산	BB(파운데이션/컬쿠드류	DIY가구	DSLR	LCD	LED	OLED	PC부품	...
201804	100.00000	93.73051	100.00000	100.00000	79.51262	95.52197	85.91247	88.58240	92.02363	...
201805	88.61685	100.00000	86.12396	96.15797	86.20727	99.39736	84.44948	89.54059	100.00000	...
201806	82.03497	95.01887	67.09475	87.16020	79.92422	100.00000	78.62776	100.00000	71.93580	...
201807	93.89507	91.58050	61.17540	83.46690	100.00000	84.68384	69.70070	88.37843	98.37518	...
201808	72.20985	81.41309	63.90595	85.07807	91.95605	85.87057	100.00000	76.43905	97.78434	...
201809	61.78960	75.43082	61.19437	81.20300	71.62405	90.07046	72.79524	77.71505	88.92171	...

외부 변수를 얻기 위한 과정으로는 '네이버 개발자센터'에서 제공하는 데이터랩 API 를 이용했다. 기본적으로는 소분류명에 해당하는 CLAC3_NM 을 트렌드랩의 키워드로 입력하고, 그로부터 도출되는 트렌드 지표를 Melt / Reshape 과정을 거쳐 월별 Trend 지수라는 새로운 형태의 데이터프레임을 도출했다.



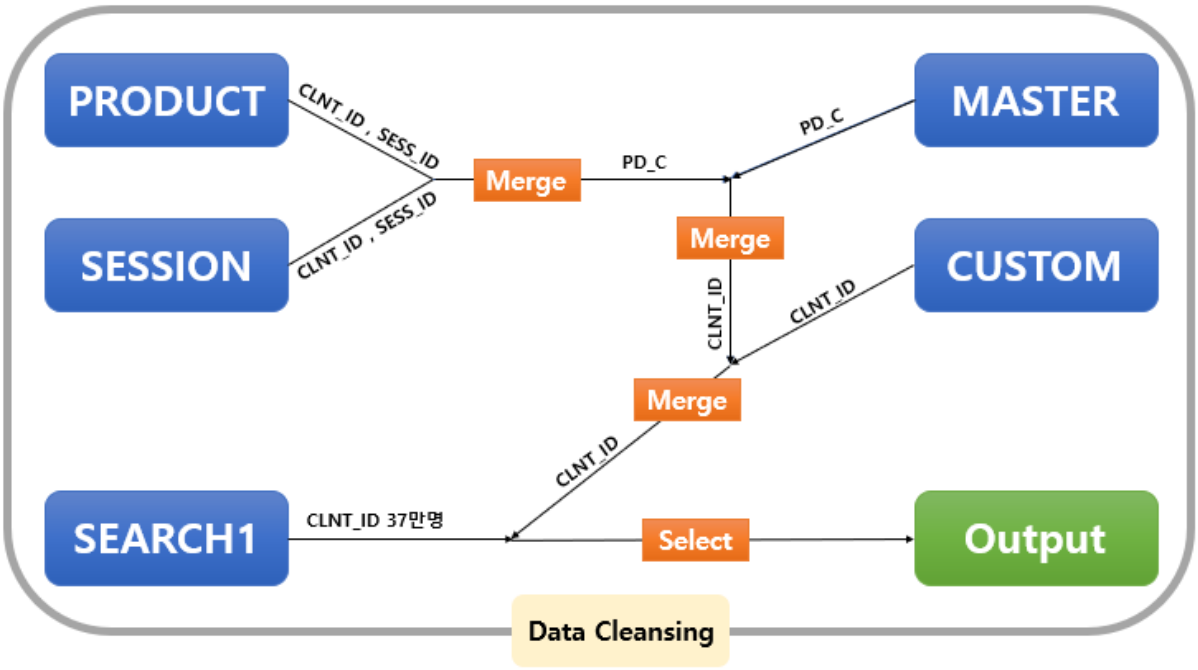
일부 소분류명들에 대해서는 원활하게 트렌드 지표가 되지 않는 경우가 발생했다. 이에 해당하는 대부분은 입력되어 있는 소분류명이 일상 언어와는 거리감이 있는 명칭으로 되어있는 경우였다. 때문에 일부 변수명들에 대해서 임의로 보다 일상적인 검색어 형태로 소분류명을 변환해줄 필요가 있었다.

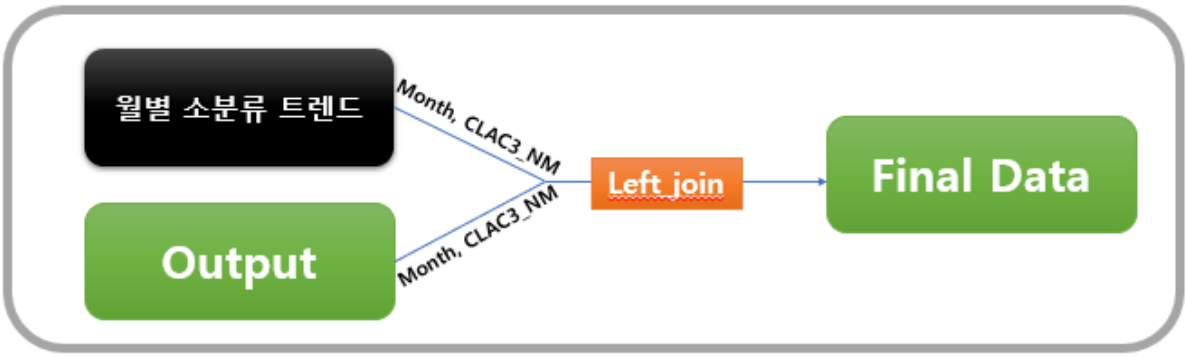
(2) 데이터별 특성 파악 후 데이터 정제



➔ 37 만명의 고객 데이터 사용

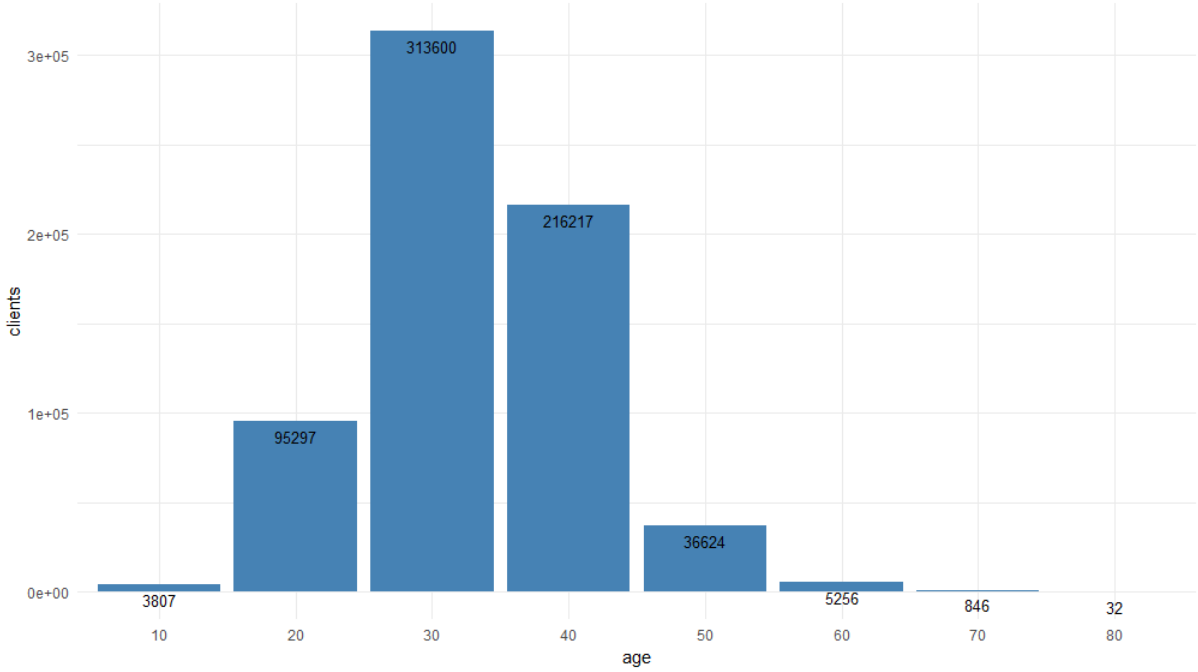
우리가 필요한 데이터로 정제하기 위해서 데이터별 고객의 분포를 파악





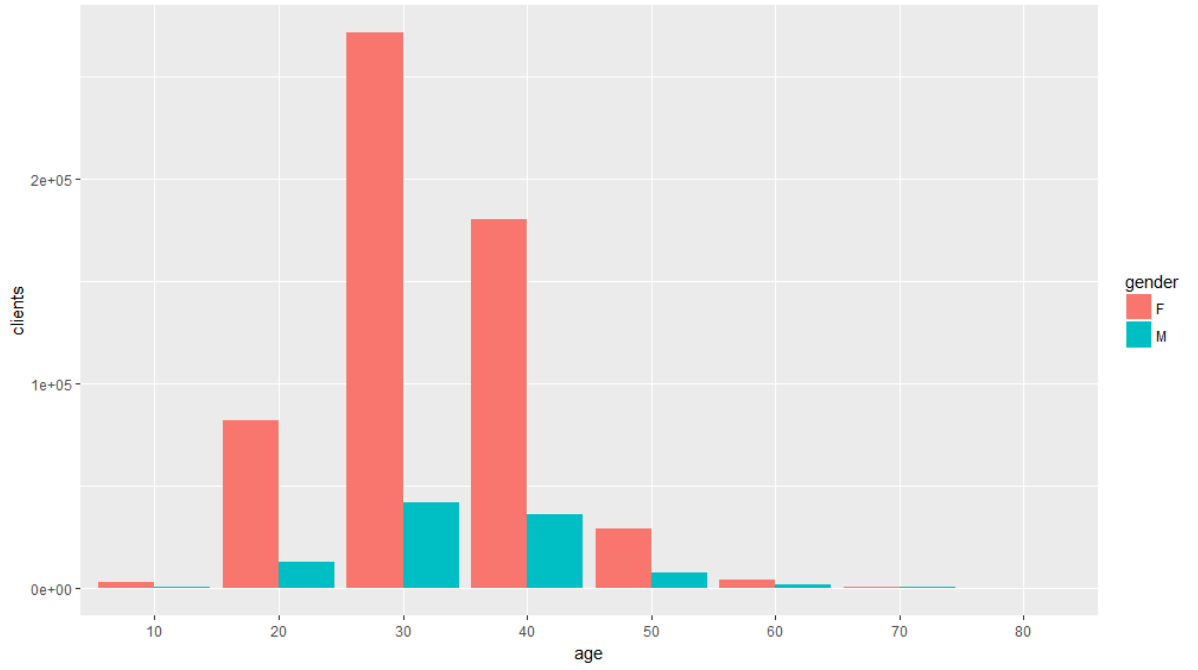
(3) 데이터 탐색을 위한 EDA

1. 고객수의 성별/연령대별 EDA



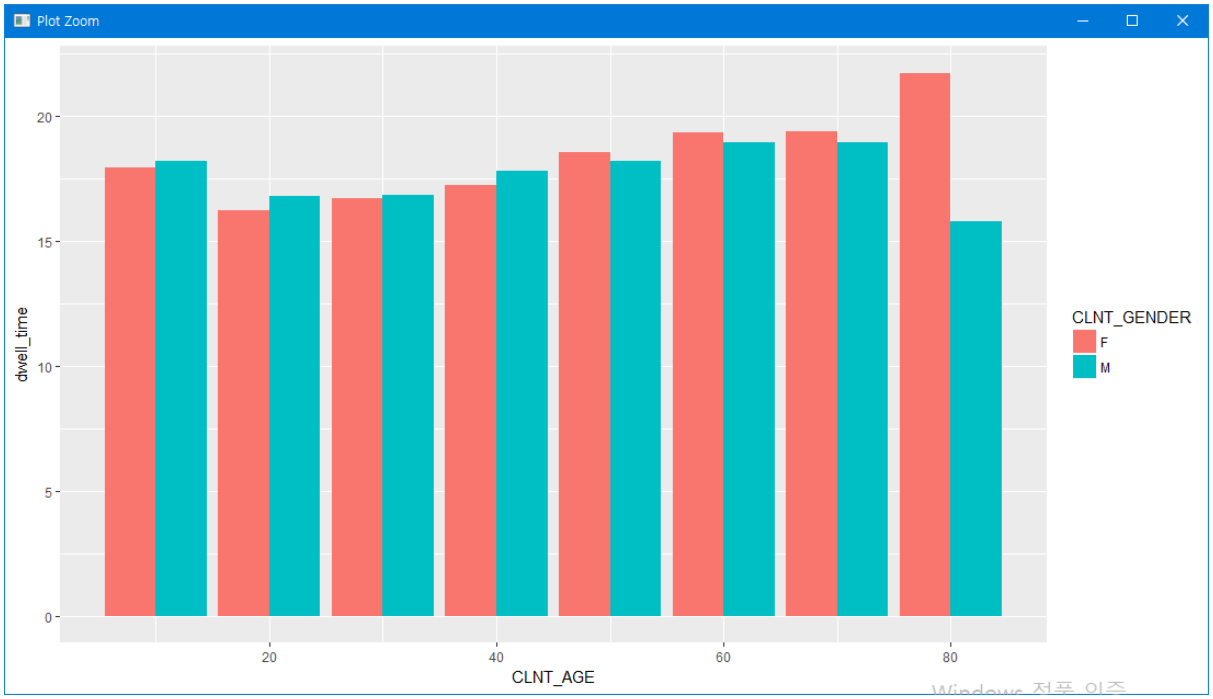
〈연령별 고객 수〉

고객의 주요 연령대는 30~40 대 인 것으로 확인됐다.



〈성별 연령별 고객 수〉

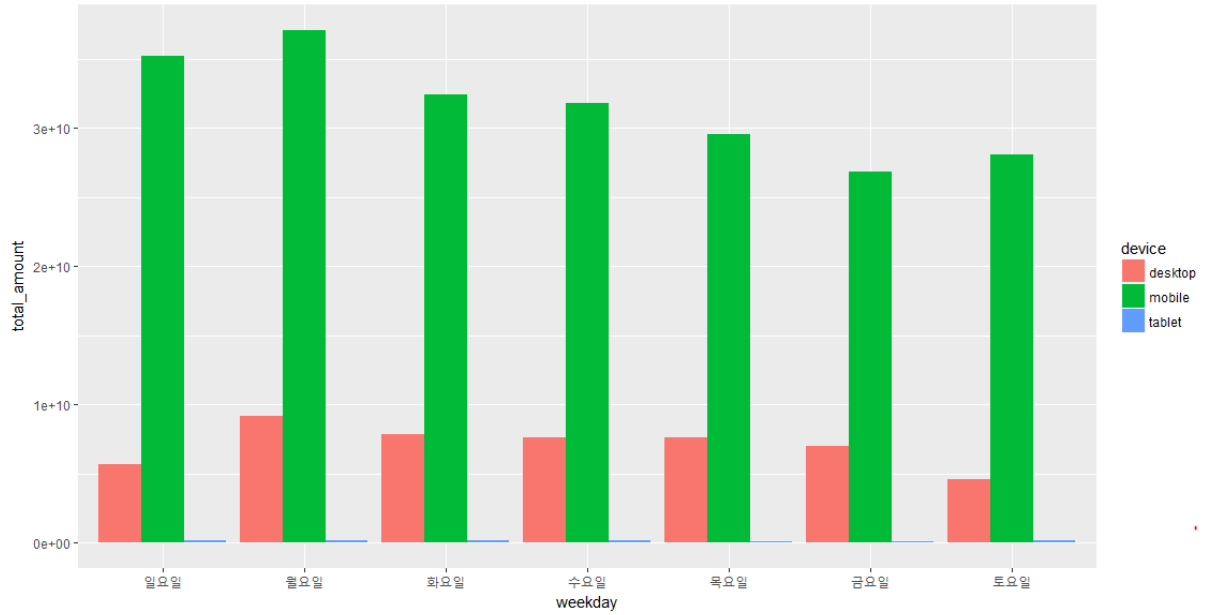
성별과 연령대를 같이 시각화를 해봤을 때 30~40 대의 여성이 주 고객층인 것을 확인됐다.



〈성별 연령별 페이지 평균 체류시간〉

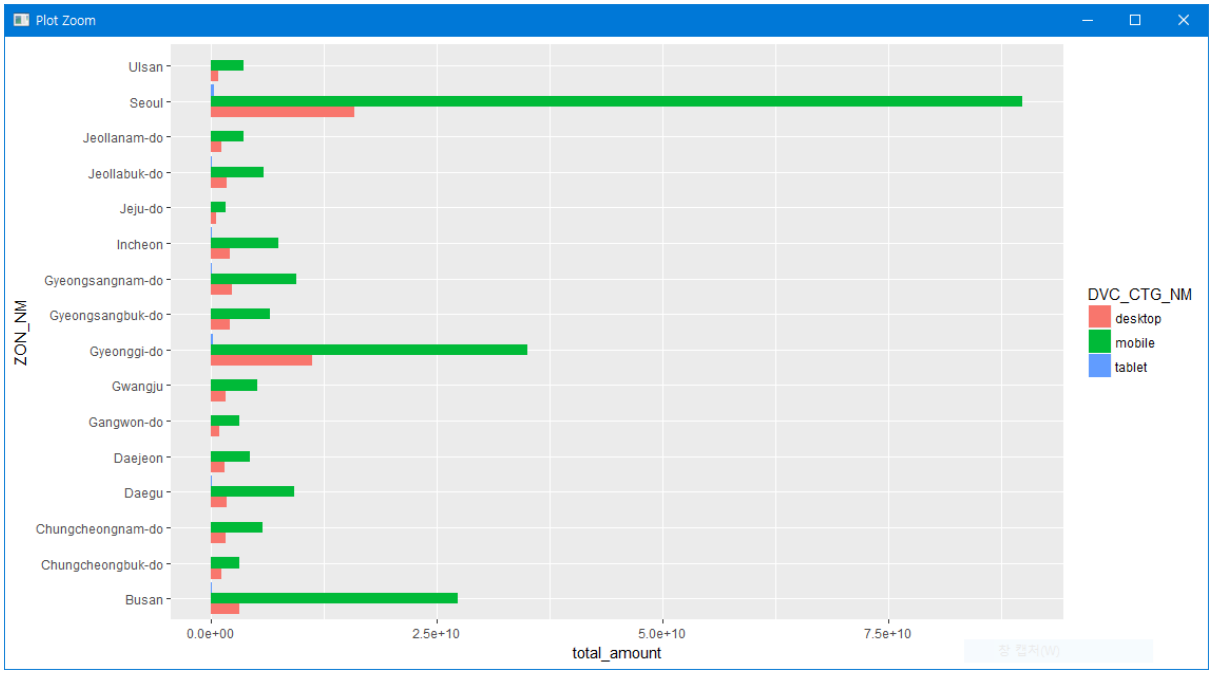
연령별로 남녀의 평균 체류시간은 비슷한 것으로 확인됐고 10 대를 제외하고 연령대가 높아질수록 페이지 평균 체류시간이 높아지는 것으로 확인됐다.

2. 기기별 요일/지역에 따른 EDA



〈요일별 기기별 구매총액〉

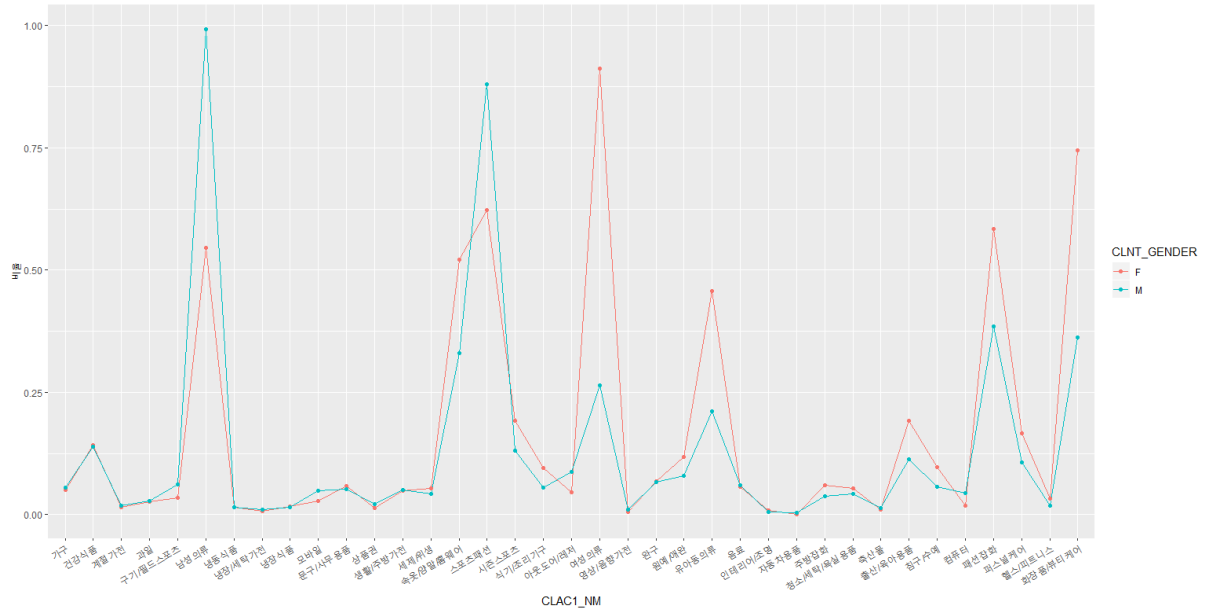
요일별로 봤을 때 월요일의 구매총액이 가장 높은 것으로 보아 고객들이 월요일에 주로 많은 구매를 하는 것으로 확인했다. 그리고 태블릿으로 구매하는 고객은 전혀 없으며 모바일로 구매하는 고객이 주를 이루는 것을 확인됐다.



〈지역별 기기별 구매총액 〉

지역별 구매총액이 높은 순으로 서울, 경기도, 부산 순으로 나타났으며 가장 구매총액이 낮은 지역은 제주도인 것으로 확인됐다.

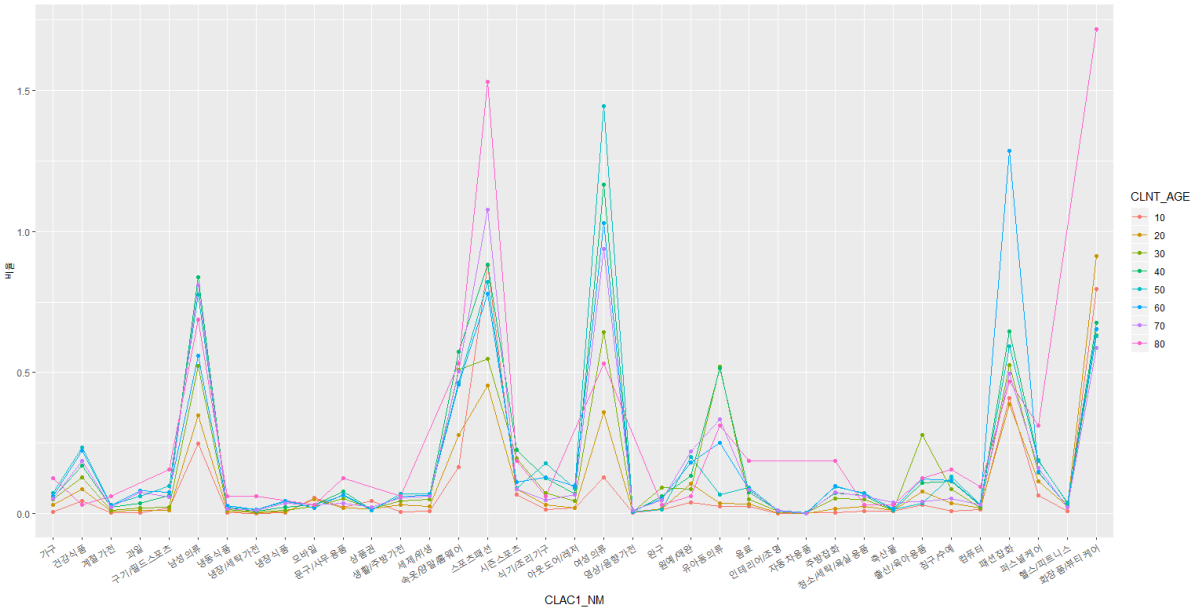
3. 대분류별 EDA



〈대분류별 성별 비율의 차이〉

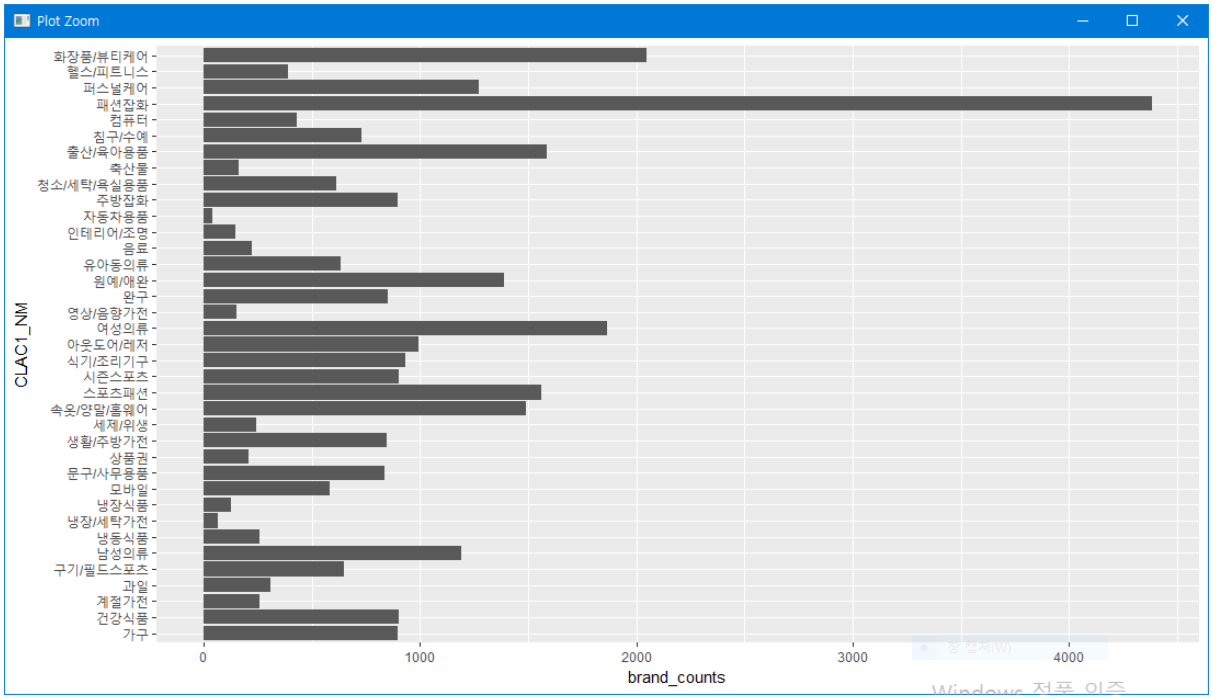
남성 고객과 여성 고객중 여성 고객이 압도적으로 많기 때문에 대분류별 성별 비율의 차이를 구하기 위해서는 비율로서 성별 고객수의 차이를 맞춰 줘야한다. 따라서 비율을 구할 때 비율 = (대분류를 구매한 남/여 의 수 / 전체 남/여 의 수) 로 맞춰주고 난 뒤 시각화한 결과이다. 상품을 구매함에 있어서 남/여의

차이는 존재하지만 그렇게 크게 나타나지는 않으며 대분류별 차지하는 남/여의 비율의 추세가 비슷한 분포를 띄는 것을 확인할 수 있다.



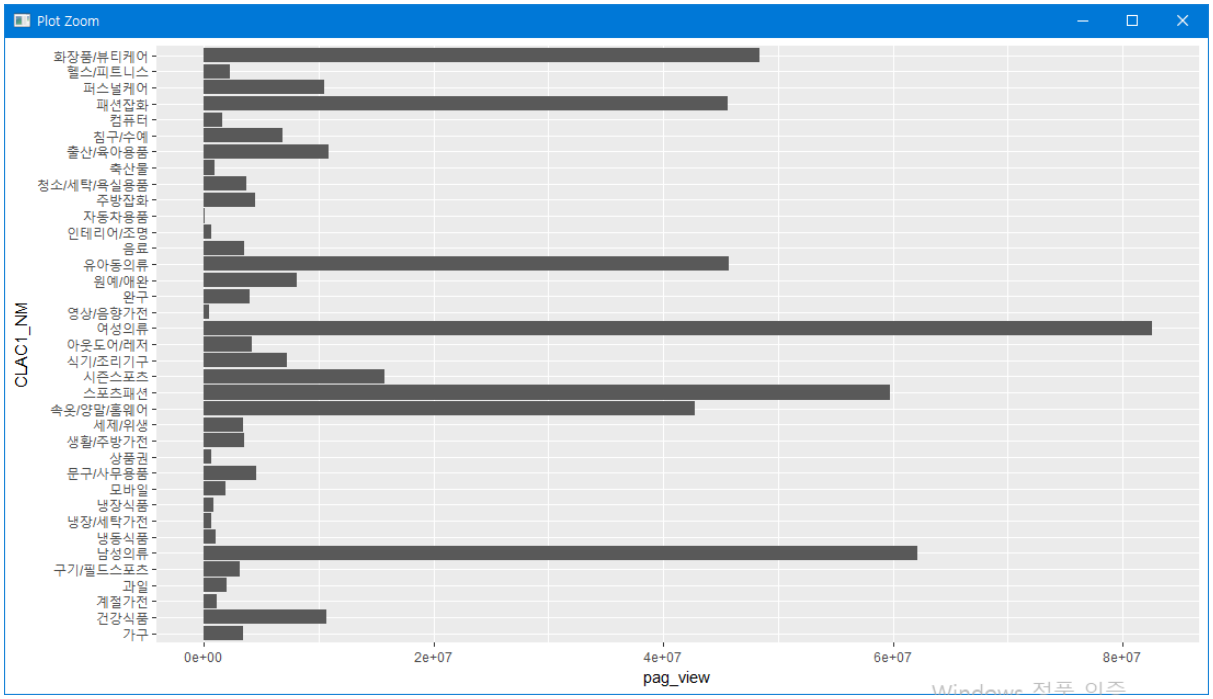
〈대분류별 연령별 비율 차이〉

비율을 구한 방식은 위와 같으며 대분류별 연령별로 차이는 존재하지만 그래프의 형태나 분포가 연령별로 비슷한 것을 확인할 수 있다.



〈대분류별 브랜드 수〉

대분류별 브랜드 수는 “패션잡화”, “화장품/뷰티케어”,”여성의류” 순으로 나타났으며 보통 의류가 속한 대분류들의 브랜드가 많은 것을 확인했다. 또한 가장 브랜드 수가 적은 대분류는 “자동차용품”,”냉장/세탁가전” 인 것을 확인할 수 있다.



〈대분류별 페이지뷰 합계〉

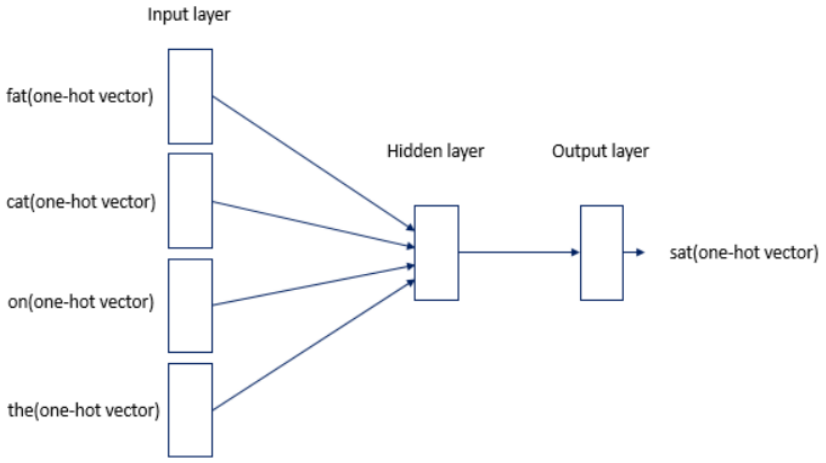
위와 마찬가지로 의류가 속한 대분류의 페이지뷰 합계가 대부분 높은 것을 확인 했다.

➔ 주어진 데이터가 어떻게 생긴지 파악하고 정제한 데이터로 여러 EDA 과정을 거치며 데이터에 대한 이해를 높였다. 또한 앞으로의 분석방향의 기초적인 지표로써 분석에 접근하기 전 워밍업을 통해서 더 나은 분석 결과를 제시할 수 있다.

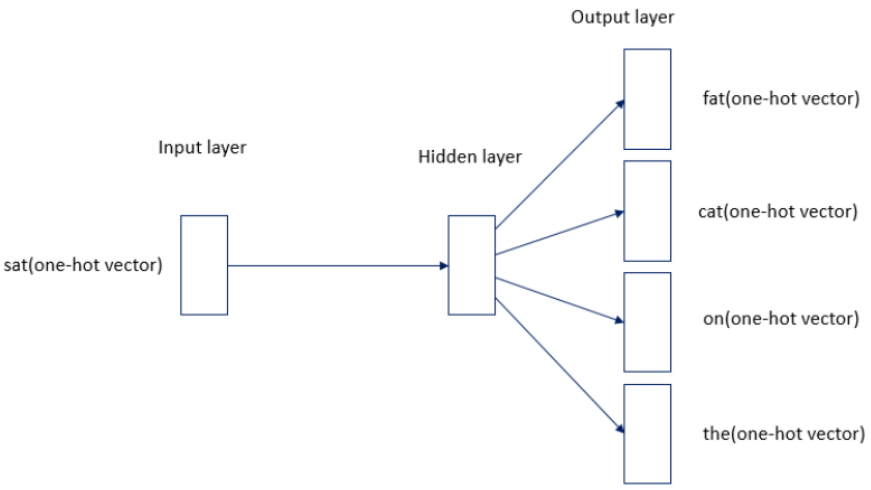
4. Category2vec 활용 군집분석

(1) 고객별 구매 제품(소분류) WORD2VEC 적용

Word2vec 기법은 NLP(자연어 처리) 분야에서 유용하게 쓰이는 알고리즘이다. 텍스트는 비정형 데이터이기 때문에 컴퓨터가 텍스트를 인지하기 위해서는 정형화가 필요하다. 원-핫 인코딩의 경우 단어 간 유사성을 계산할 수 없다는 단점이 있다. 단어 간 유사성을 고려하기 위해 단어의 의미를 벡터화 하는 것이 word2vec 의 핵심이다. Target 단어와 주변단어를 이용한 학습을 통해 단어를 벡터화 한 model 로써 주변에 많이 등장하는 단어가 유사성이 높으므로 함께 많이 등장하는 단어일수록 가까운 위치에 embedding 이 된다. Word2vec 에는 핵심단어를 통해 주변 단어를 예측하는 skip-gram 과 주변 단어를 통해 핵심 단어를 예측하는 cbow 가 있다.



〈 CBOW 의 뉴럴 네트워크 도식화 〉



〈 Skip-gram 의 뉴럴 네트워크 도식화 〉

〈출처 : https://wikidocs.net/22660 〉

'단어간의 유사성을 고려한다'라는 아이디어를 상품 구매 데이터에 적용해 보았다. 특정 상품과 동시 등장 하는 상품에는 유사성이 있을 것이라고 판단하고 이에 word2vec 을 적용한다면 유사한 특성을 갖는 상품일수록 가까운 위치에 embedding 될 것이다.

(1) 고객별 구매 상품의 소분류를 리스트 형식으로 재구성 한다.

A 고객 : ['요가/필라테스복', '에센스/세럼', '유아동스니커즈', '롤플리엥완구', …],

B 고객 : ['인삼가공식품', '여성원피스', '남성등산티셔츠', 여성런닝/트레이닝화', …]

…

(2) 구성한 소분류 리스트 형식에서 878 개의 소분류를 word2vec 을 통해 embedding 하여 category2vec 을 수행했다. 그 결과는 다음과 같다. (size = 50, window = 5, min_count = 0, skip-gram 적용)

```
w2v_ctgr.most_similar('드럼세탁기')
```

```
[('일반세탁기', 0.8349478840827942),
 ('양문형냉장고', 0.8270360231399536),
 ('뚜껑형김치냉장고', 0.8136491775512695),
 ('UHD', 0.8066824674606323),
 ('제습기', 0.7719041109085083),
 ('인덕션/가스레인지', 0.7710935473442078),
 ('일반형냉장고', 0.769775748252868 7),
 ('LED', 0.7383269667625427),
 ('오븐/전자레인지', 0.7108999490737915),
 ('스팀청소기', 0.6950793266296387)]
```

```
w2v_ctgr['드럼세탁기']
```

```
array([-0.3356509 ,  0.30415976,  0.45588037,  0.3934993 ,  0.12488841,
        0.38194773, -0.5414641 , -0.0914508 ,  0.18329015,  0.32376578,
        0.22474015,  0.3137486 ,  0.1596046 , -0.3477618 ,  0.27373183,
        0.03132236,  0.18943505,  0.2963021 , -0.2461694 , -0.03766029,
       -0.8108599 , -0.21801034, -0.4840352 ,  0.06306282,  0.03707084,
       -0.3065883 ,  0.16005425,  0.15600505,  0.18135636, -0.6668859 ,
       -0.13570362, -0.08206454,  0.70183194, -0.09853657, -0.17886744,
        0.2316252 ,  0.12810399,  0.43969887, -0.53095865, -0.2993681 ,
       -0.1553632 ,  0.07014281, -0.05745063, -0.65105414, -0.033983  ,
       -0.07665318, -0.13897672,  0.21875033,  0.04048109, -0.23704416],
      dtype=float32)
```

⇨ '드럼세탁기'와 유사도 비교 결과 가전제품과 유사한 것으로 보인다.

```
w2v_ctgr.most_similar('유아용기저귀')
```

```
[('유모차', 0.8580231666564941),
 ('젖병소독/건조용품', 0.8116005659103394),
 ('유아동침구세트', 0.7656912207603455),
 ('봉제인형', 0.7541308999061584),
 ('유아용카시트/매트', 0.7403544783592224),
 ('손싸개/발싸개', 0.7224968671798706),
 ('놀이방매트', 0.7218421697616577),
 ('기타유아안전용품', 0.7215864062309265),
 ('유아목욕용품', 0.7105535268783569),
 ('유아동의자', 0.7002077102661133)]
```

```
w2v_ctgr['유아용기저귀']
```

```
array([-0.4295905 , -0.4630228 ,  0.28183344, -0.4855301 ,  0.15694307,
       -0.09564234,  0.0352842 , -0.03075978,  0.17001054,  0.6068843 ,
        0.12169428,  0.32860583, -0.2883106 , -0.2013008 , -0.21400547,
        0.20374964,  0.35767025,  0.3433315 , -0.05241069,  0.71270347,
       -0.13687944, -0.80598587, -0.03084686, -0.7229786 ,  0.14962982,
       -0.41472012,  0.07813483,  0.18869066, -0.22313096, -0.02858695,
        0.44324937, -0.09535526, -0.19121158,  0.0144452 , -0.00277848,
        0.12612653,  0.13120559, -0.32104936, -0.3148508 , -0.5567928 ,
        0.12396365,  0.17694794,  0.34530404,  0.06376926, -0.14690839,
       -0.16555254, -0.73329365,  0.11216077,  0.08812086,  0.1778721 ],
      dtype=float32)
```

⇨ 유아 관련 제품들이 유사한 것으로 보인다.

```
w2v_ctgr.most_similar('고양이간식')
```

```
[('고양이캣타워/실내용품', 0.7145969867706299),
 ('애견식기/물병', 0.7081140279769897),
 ('고양이모래/배변용품', 0.703849196434021),
 ('애견주거/실내용품', 0.6885566711425781),
 ('고양이장난감', 0.6700047254562378),
 ('고양이사료', 0.635989785194397),
 ('애견사료', 0.631839394569397),
 ('애견장난감/훈련', 0.6286569237709045),
 ('고양이건강용품', 0.6133353114128113),
 ('고양이식기/급수', 0.6053150296211243)]
```

```
w2v_ctgr['고양이간식']
```

```
array([ 0.08618071,  0.3330157 ,  0.20590237, -0.19843762,  0.10846572,
        0.1777408 ,  0.02294319,  0.61077034, -0.04260405,  0.44562387,
        0.38952595,  0.72699213, -0.5681681 ,  0.02613336,  0.15724212,
       -0.6524382 ,  0.3498044 ,  0.43090844, -0.2770452 ,  0.37731037,
       -0.238187  , -0.04274313, -0.28643486, -0.01380714, -0.00267461,
       -0.37815267, -0.86158395,  0.46212447,  0.1729415 ,  0.07582544,
       -0.02086533,  0.22540154, -0.39833957, -0.23346503,  0.09509496,
       -0.00276615,  0.15812123,  0.64439696,  0.32308877,  0.10906809,
       -0.98041904, -0.3710912 , -0.00193002,  0.30319405,  0.12807219,
       -0.48796234,  0.07831641,  0.39114213, -0.01830657, -0.63391453],
      dtype=float32)
```

⇨ 고양이, 애완동물 관련 제품들이 유사한 것으로 보인다.

(2) CATEGORY2VEC 의 결과에 대해 K-MEANS 적용

K-means 알고리즘은 분리형 군집화 알고리즘 가운데 하나이다. 각 군집은 하나의 중심을 가지게 된다. 각 개체는 가장 가까운 중심에 할당되며, 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성한다. 여기서 사용자가 사전에 군집 수(k)를 지정해주어야 한다. K-means 진행 절차는 다음과 같다.

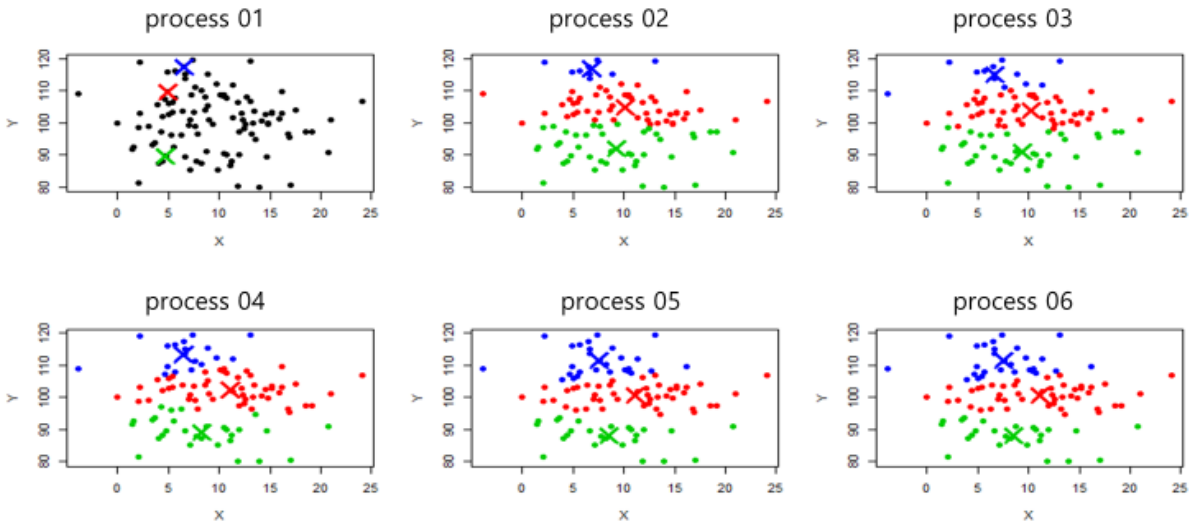
(1) 데이터 내 객체 중 임의로 K 개의 군집 중심점 설정

(2) 모든 객체에 대해 각 군집 중심점까지의 거리 계산

(3) 모든 객체를 가장 가까운 군집 중심점이 속한 군집으로 할당

(4) 각 군집의 중심점 재설정

(5) 군집의 중심점이 변경되지 않을 때까지 1~4 반복



〈 K-means 알고리즘 진행 절차 〉

〈출처 : http://www.learnbymarketing.com/methods/k-means-clustering/〉

군집 형성에 K-means 알고리즘을 사용한 이유는 거리 기반인 K-means 는 거리기반 알고리즘으로써 category2vec 으로 embedding 된 소분류간의 거리를 구해 군집 형성을 할 수 있기 때문이다.

K=60 으로 했을 때 군집 분석 결과는 다음과 같다.

군집 번호	군집 내 소분류
1	가습기, 공기청정기, 공유기, 노트북, 로봇청소기, 물걸레청소기, 스팀청소기, 오븐 ...
2	모빌, 놀이방매트, 식기건조기, 손싸개/발싸개, 모유보관용품, 유아동침구, 젖병 ...
3	남성내의, 남성양말선물세트, 브래지어, 여성가운, 여성내의, 여성실내복, 여성팬티 ...
4	남성스킨케어세트, 데오도란트, 바디워시, 바디케어세트, 샴푸, 입욕제, 크림, ...
5	글루코사민, 기타영양제, 루테인, 어린이홍삼, 오메가3, 일반비타민,칼슘/미네랄, 한방음료
...	
60	그늘막/타프, 기타캠핑용품, 야외용돗자리, 오토캠핑용품, 캠핑취사, 캠핑침구, 텐트, ...

1 번 군집 : 집안 가전제품 또는 청소 도구 끼리 묶인 것으로 보인다.

2 번 군집 : 유아 용품, 유아동을 위한 소분류 끼리 묶인 것으로 보인다.

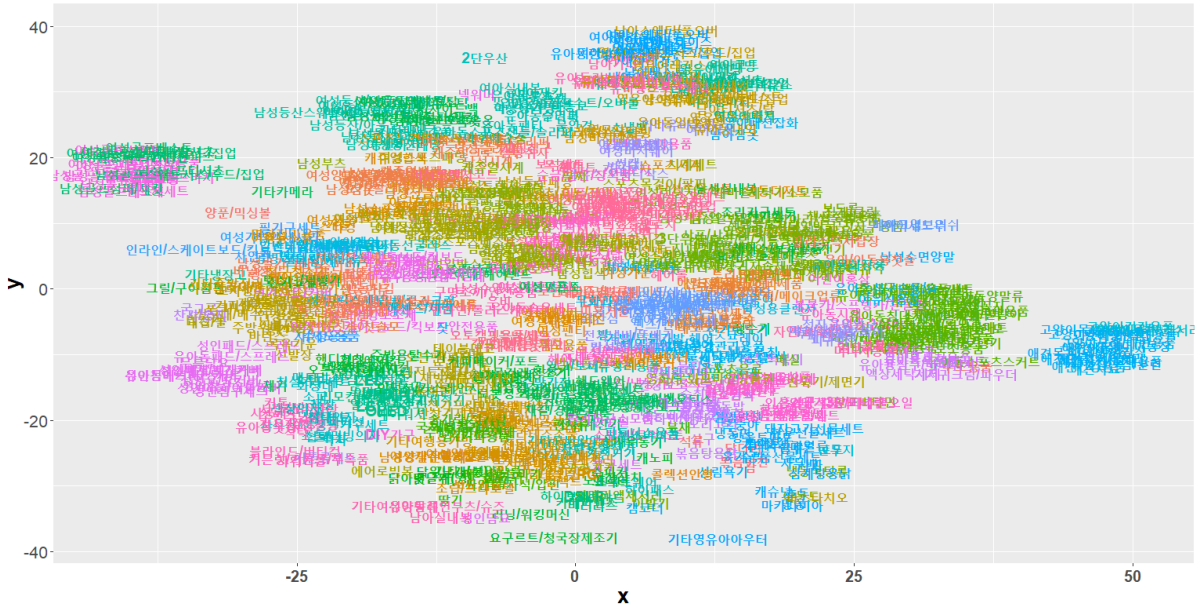
3 번 군집 : 여성이 남성을 위한 선물 또는 자신의 속옷, 내의로 묶인 것으로 보인다.

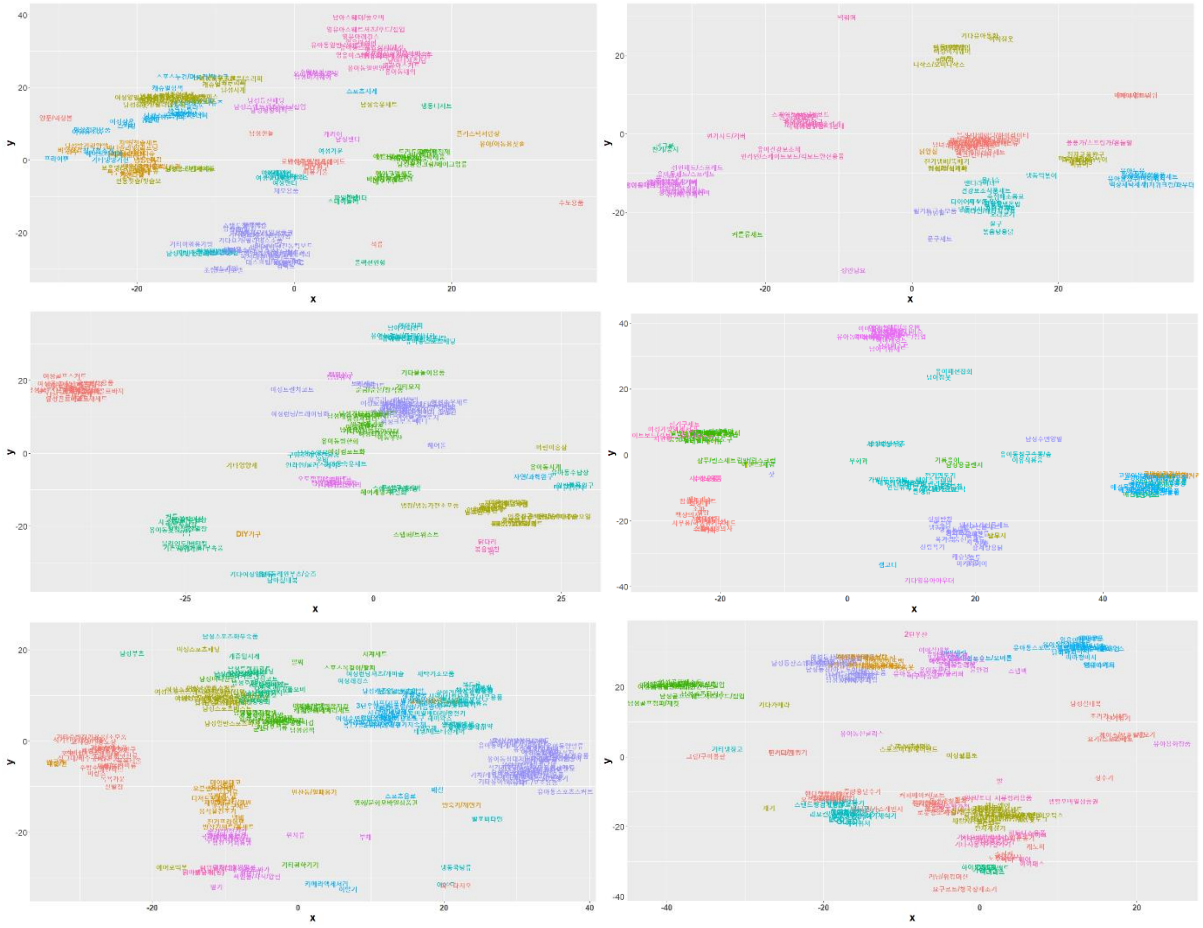
4 번 군집 : 욕실 용품 관련 소분류로 묶인 것으로 보인다.

5 번 군집 : 비타민, 영양제 등으로 묶인 것으로 보인다.

60 번 군집 : 캠핑 관련 소분류로 묶인 것으로 보인다.

소분류 군집분석 시각화

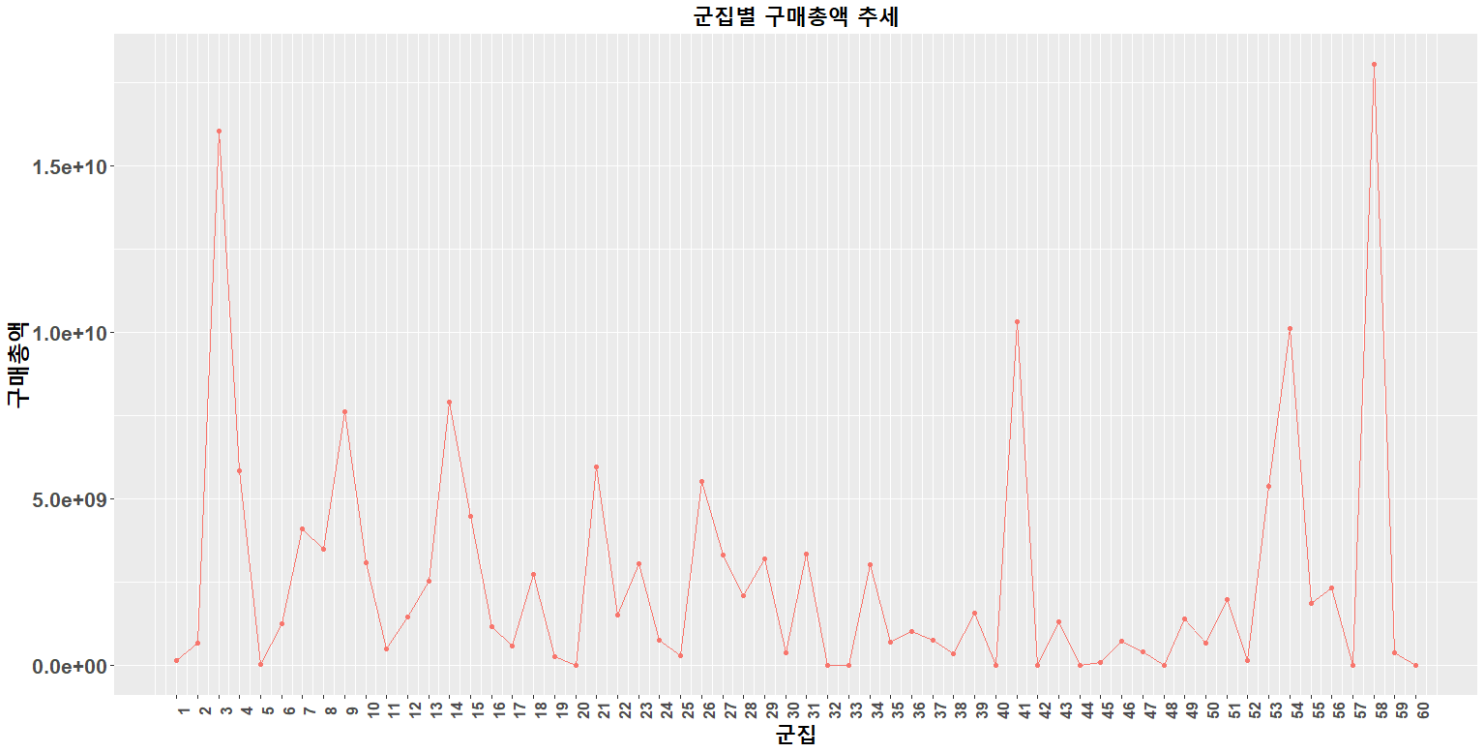




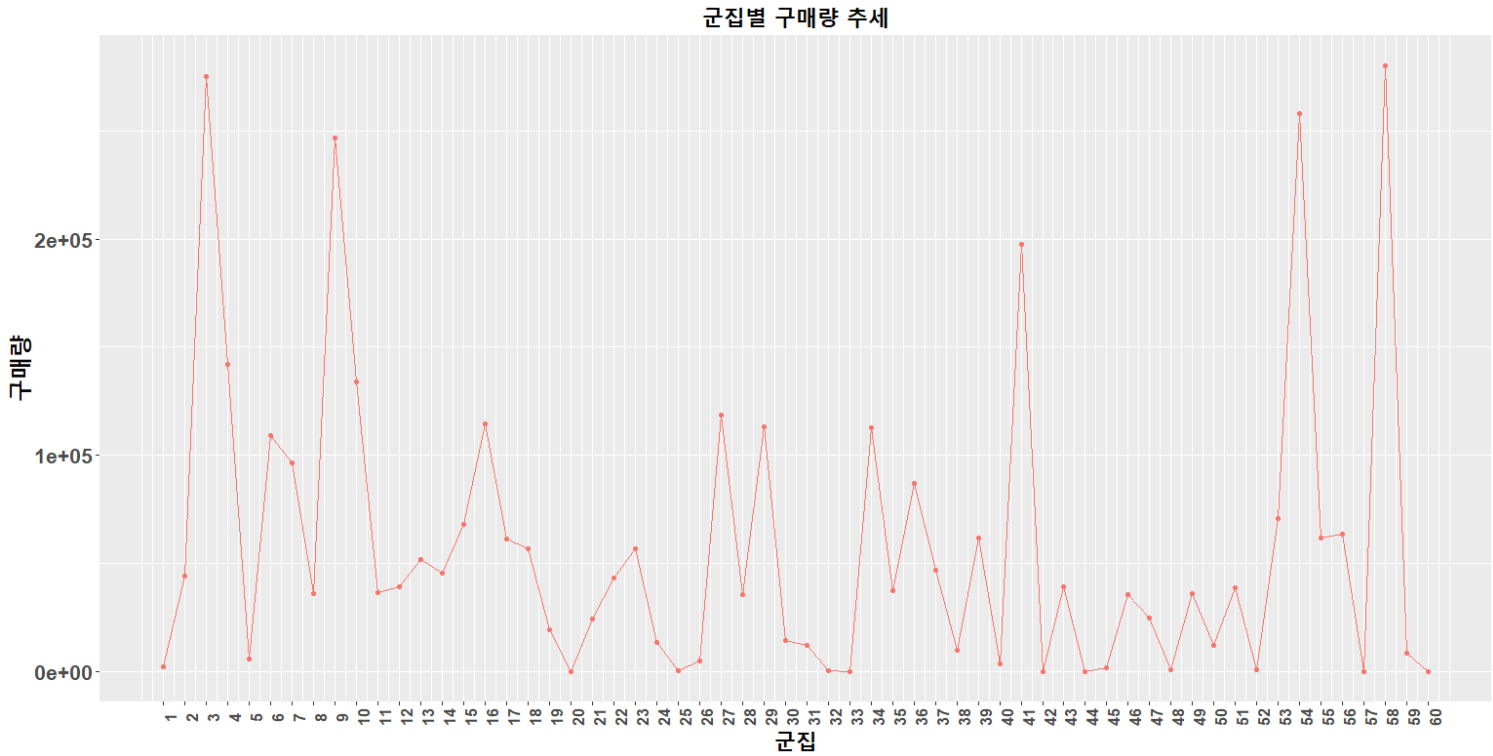
⇨

⇨ 군집분석 결과를 T-SNE 를 통해 2 차원으로 차원축소하여 표현한 결과, 군집 형성이 잘 되었음을 확인 할 수 있다. (맨 위 – 60 개 군집, 아래 6 개 – 각 10 개 군집)

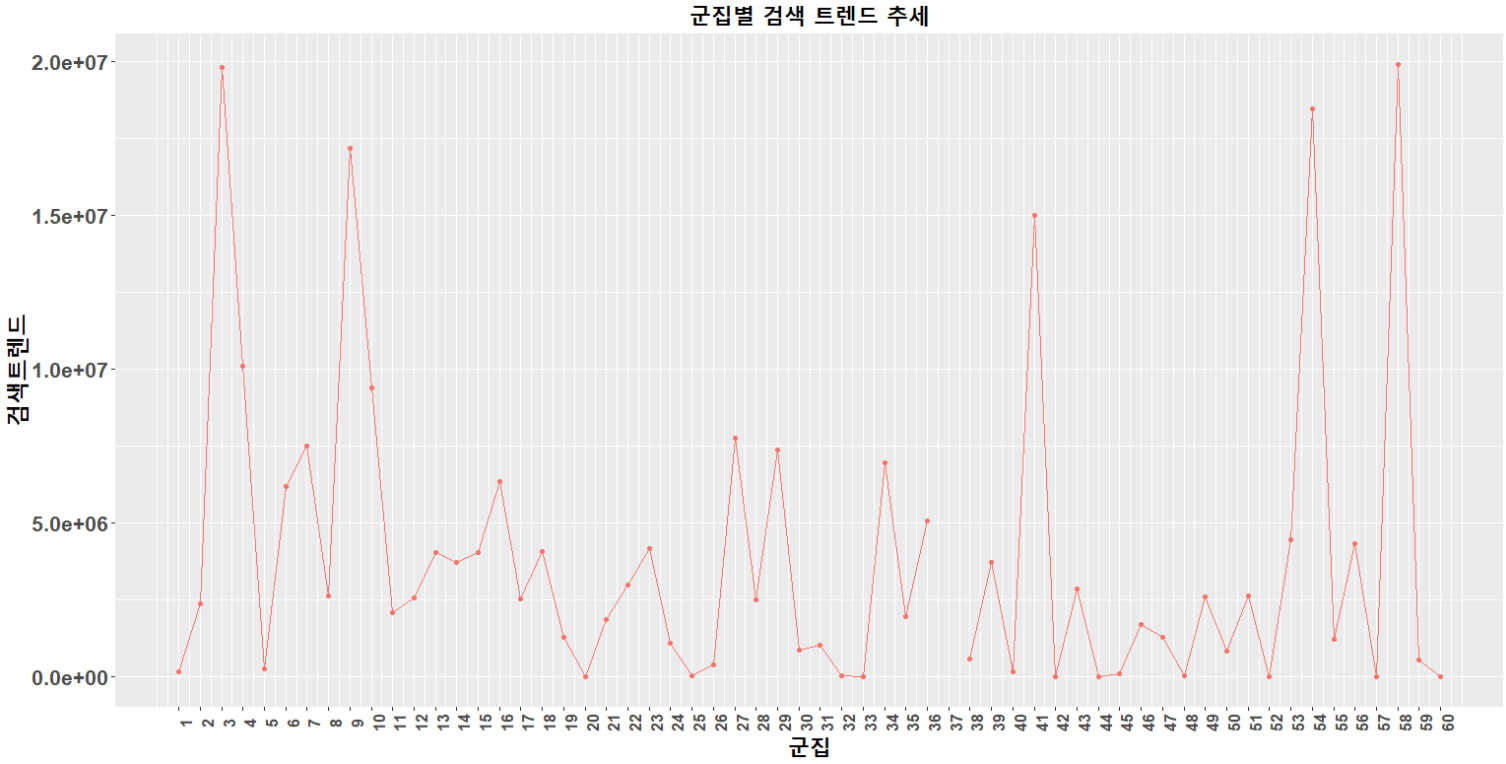
(3) 군집과 고객의 관계



군집별 고객의 구매 총액



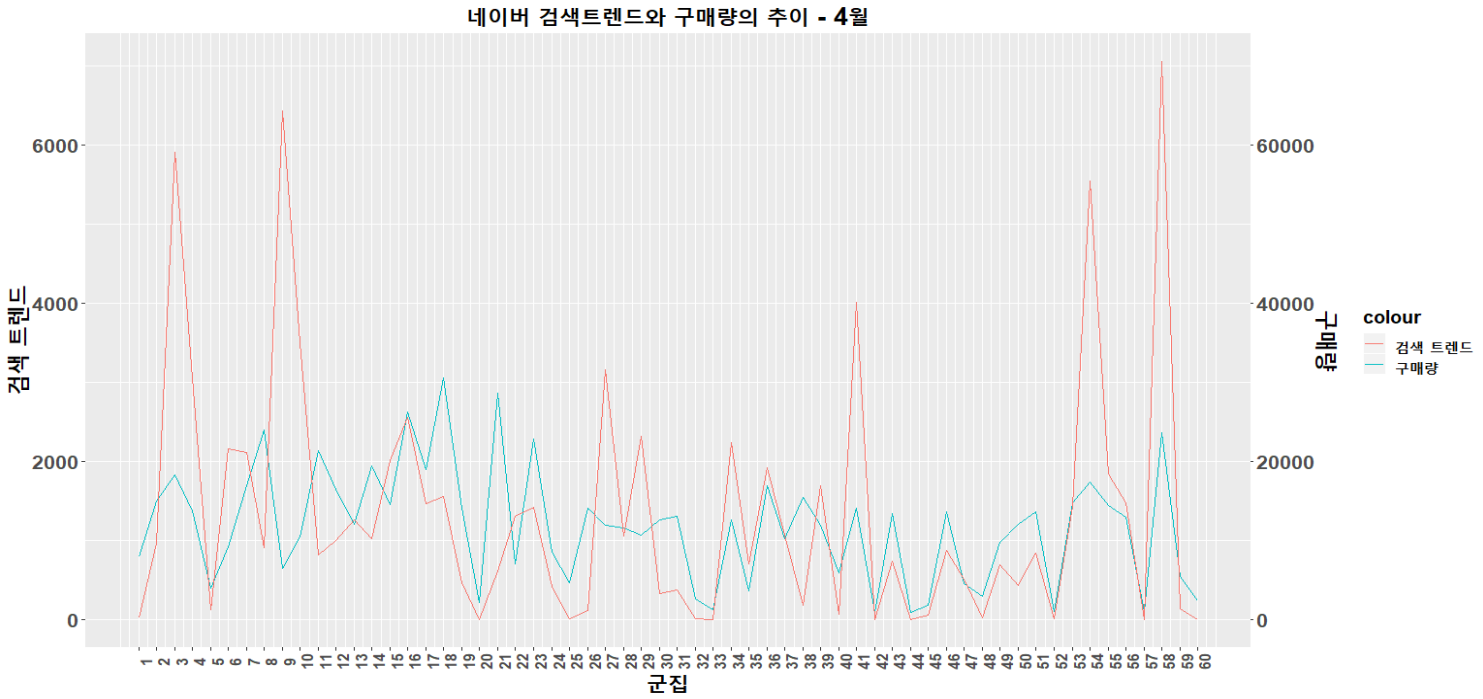
군집별 고객의 구매량



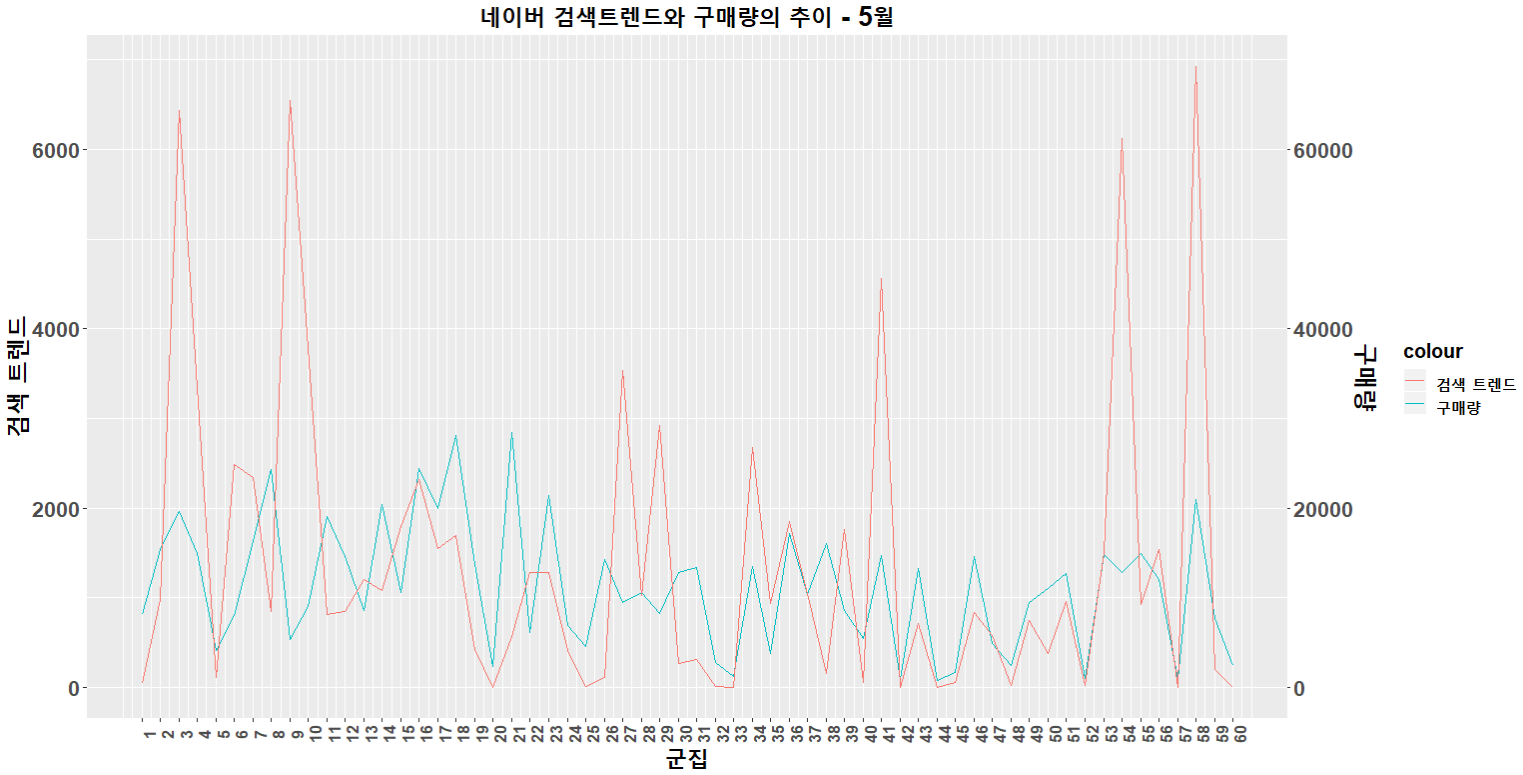
군집별 고객의 검색 트렌드(내부데이터)

고객의 특성을 군집별로 살펴본 결과 군집별로 비슷한 그래프를 보이고 있다. 3,9,41,54,58 번 군집이 높이 솟아 있는 형태이며 그 외의 모습에서도 비슷한 형태를 보이고 있다. 이를 통해 군집화가 잘 이루어져 있다고 볼 수 있다.

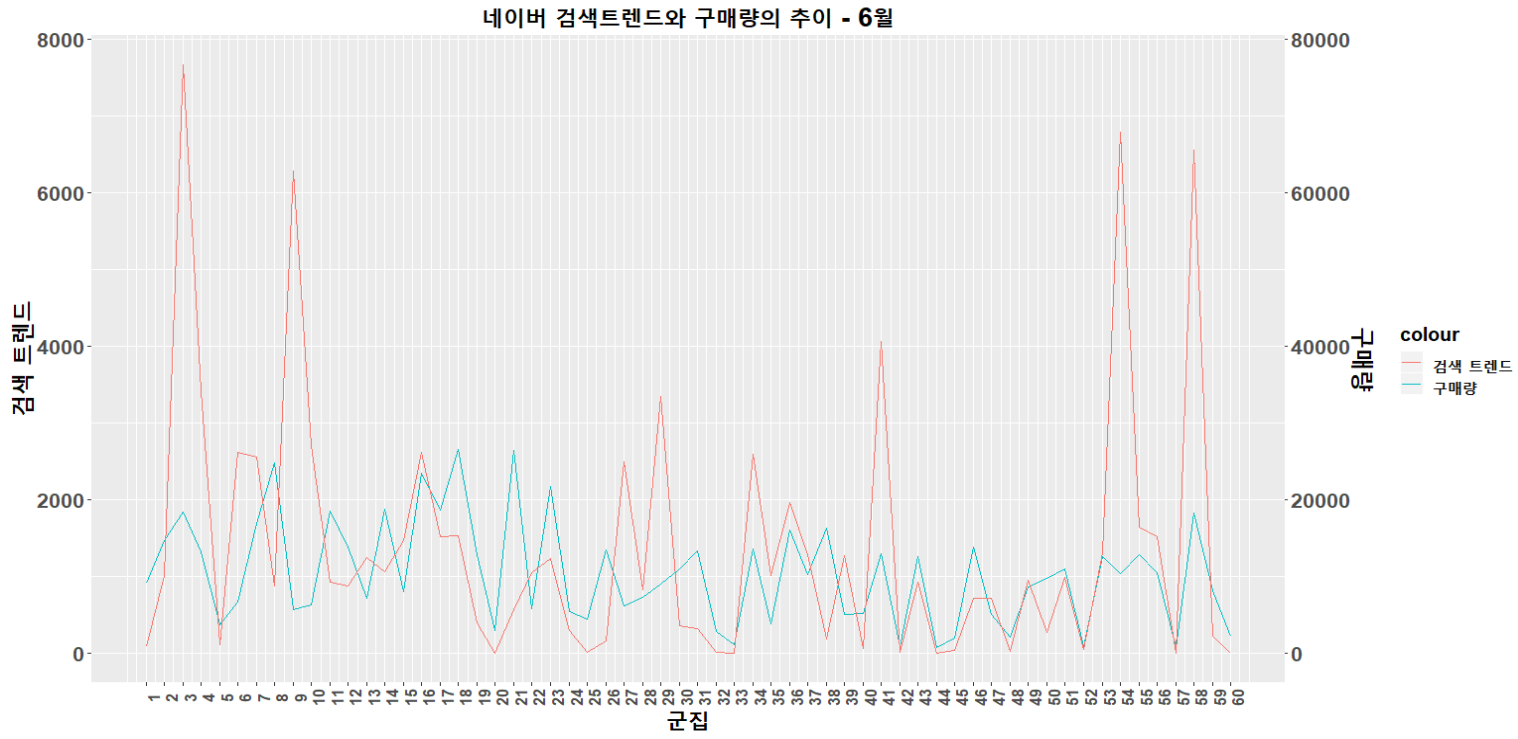
(4) 군집과 검색어 트렌드와 관계



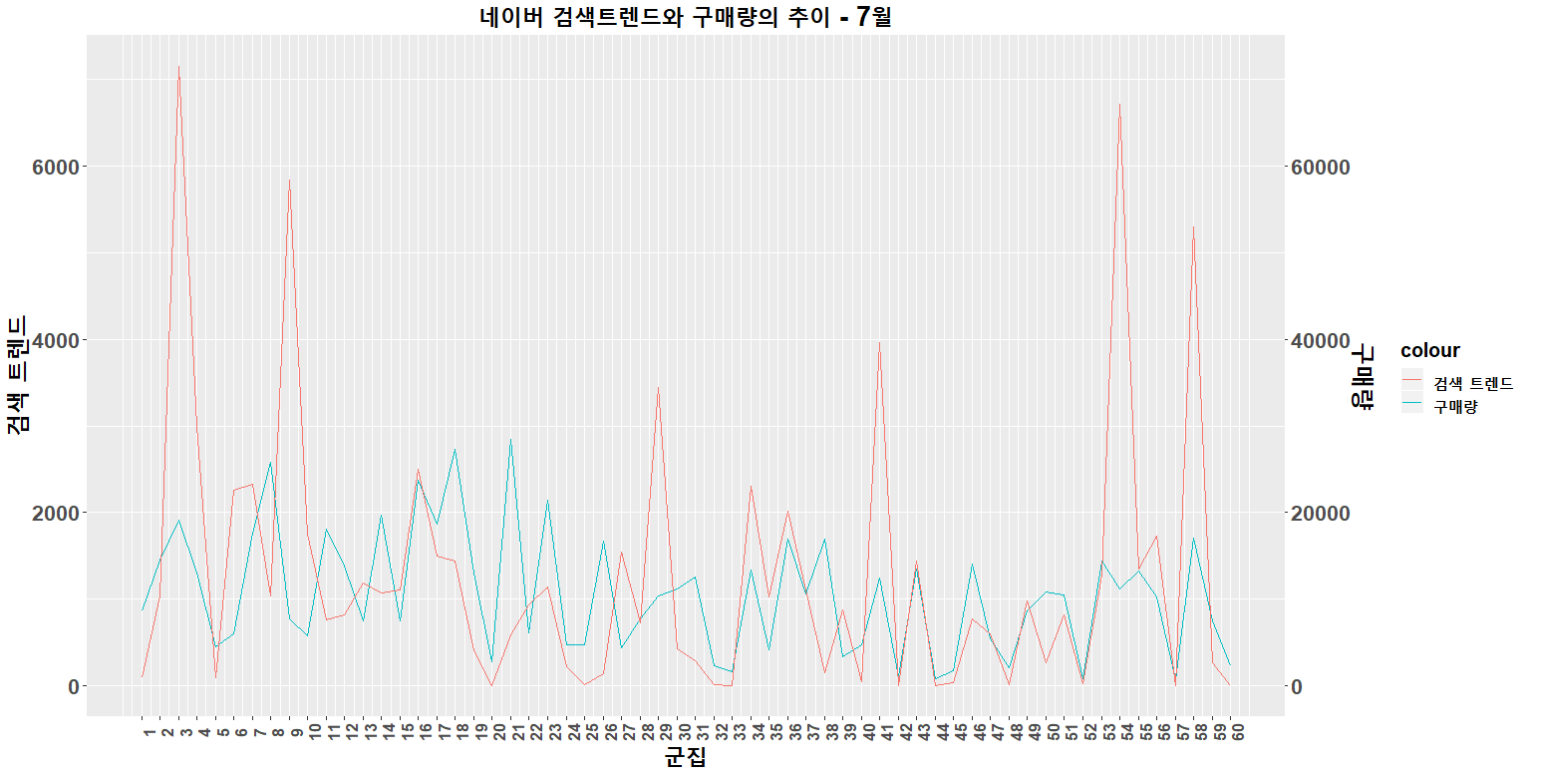
상관계수 0.50213



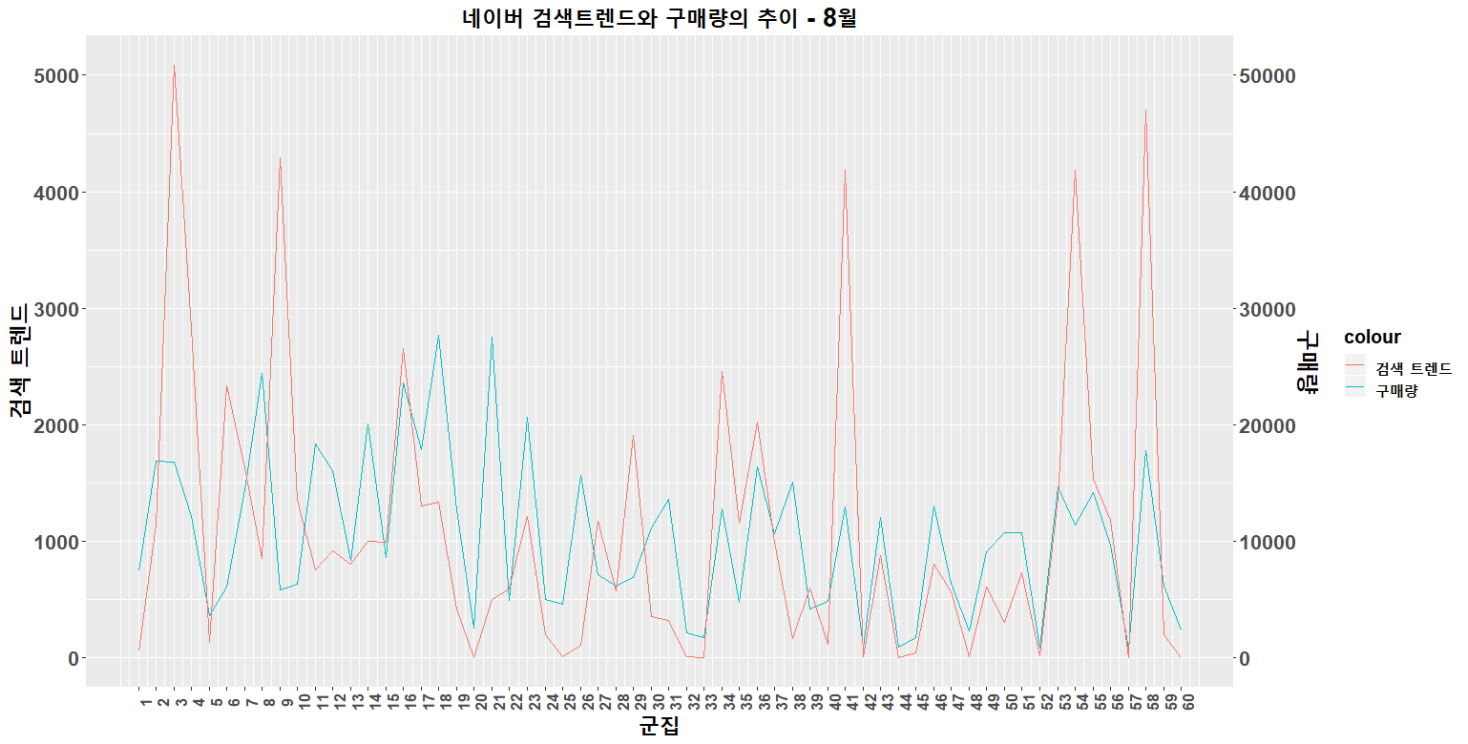
상관계수 0.36597



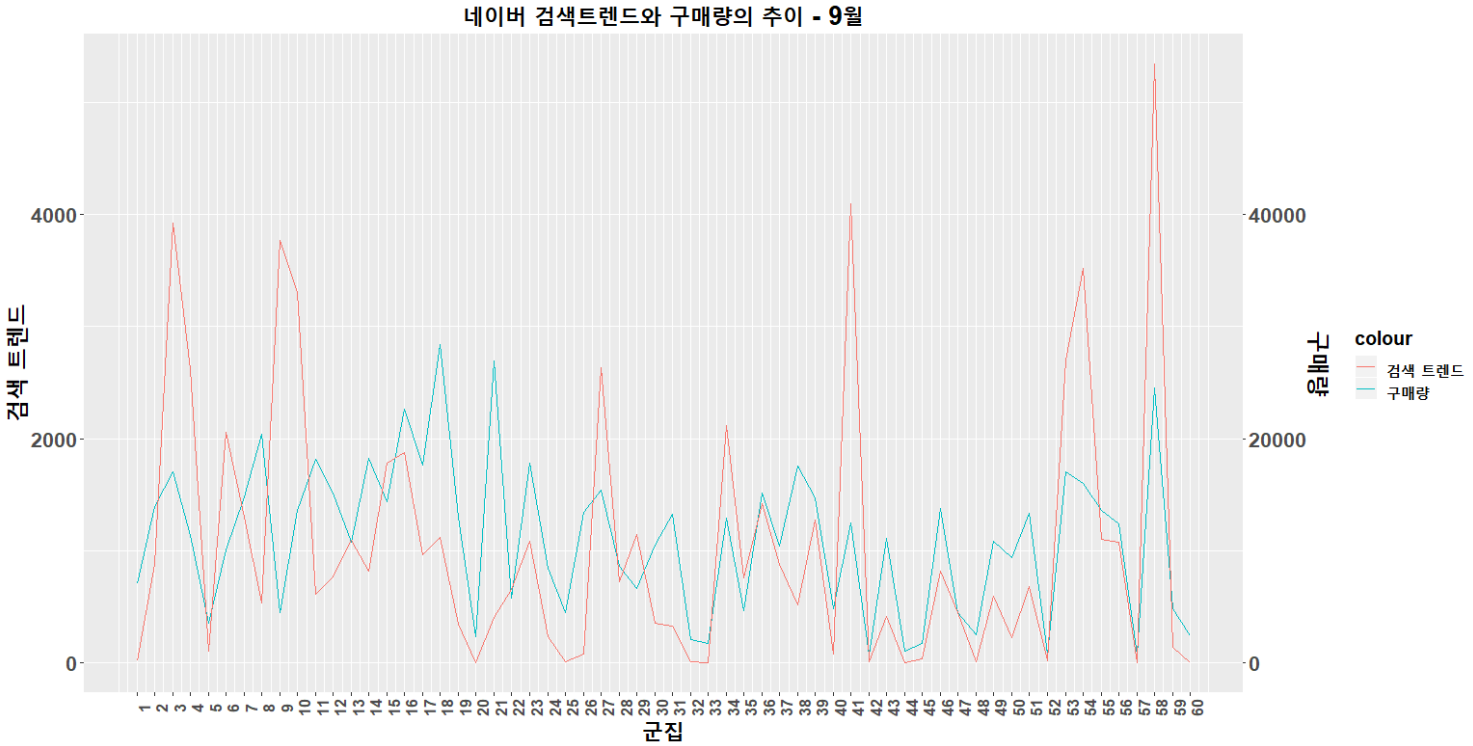
상관계수 0.38918



상관계수 0.44337



상관계수 0.48236

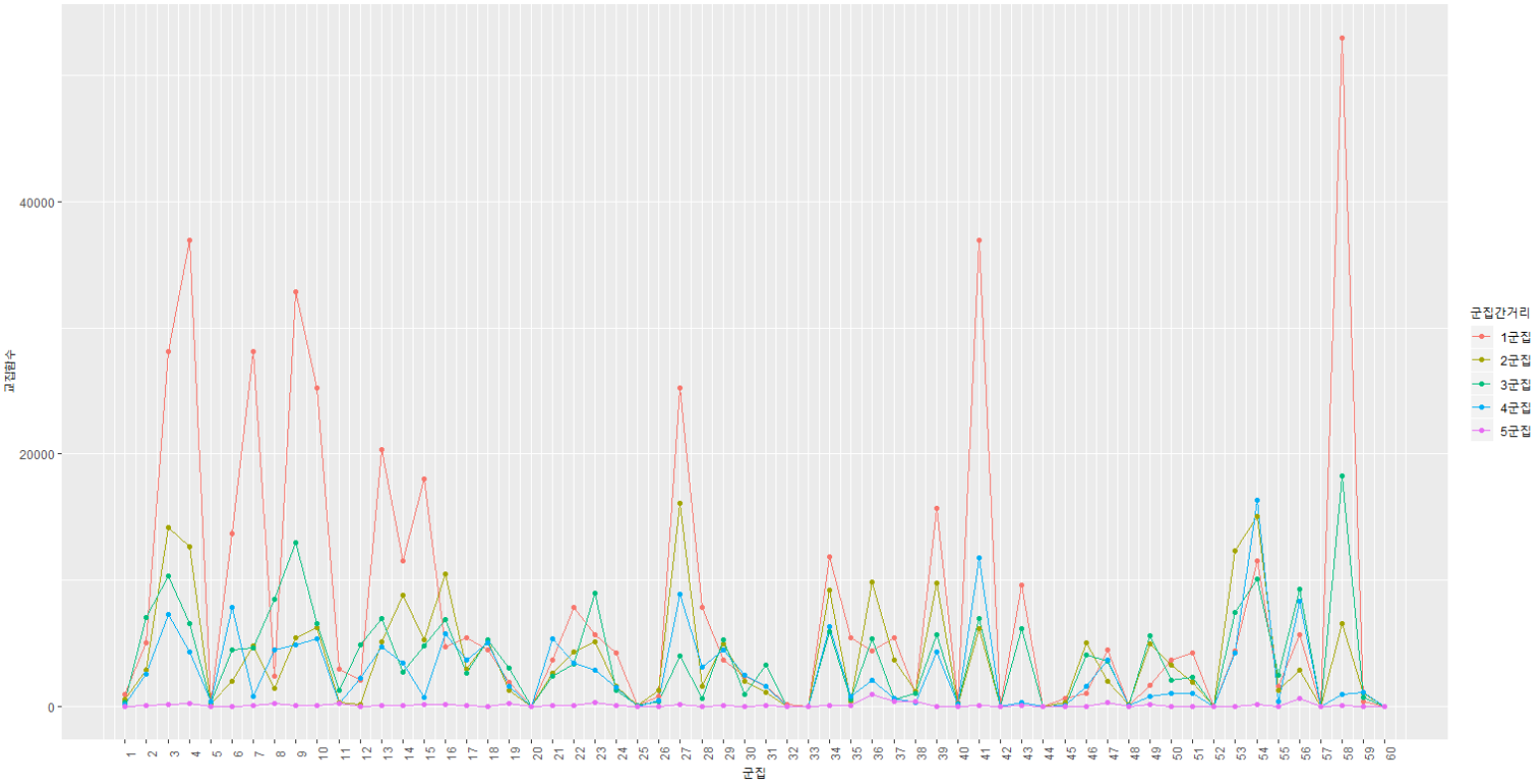


상관계수 0.49528

네이버 트렌드와 군집간의 관계를 살펴본 결과 0.36~0.49 정도의 다소 유의미한 상관계수 값을 보였고 이는 실제로 해당 고객이 제품 구매 여부에 좋은 변수가 될 수 있을 것이다.

(5) 해당 군집과 거리가 가장 가까운 군집

지금까지 898 개의 소분류 군집화와 이를 분석했다. Word2vec 과 K-means 는 거리기반 이기 때문에 군집간의 거리를 분석했다. 여기서 한 가지 '해당 군집의 제품을 구매한 사람은 거리가 가장 가까운 군집의 제품을 구매 할 가능성이 높다'라는 가설을 세웠다. 이를 토대로 시각화 한 결과는 다음과 같다.



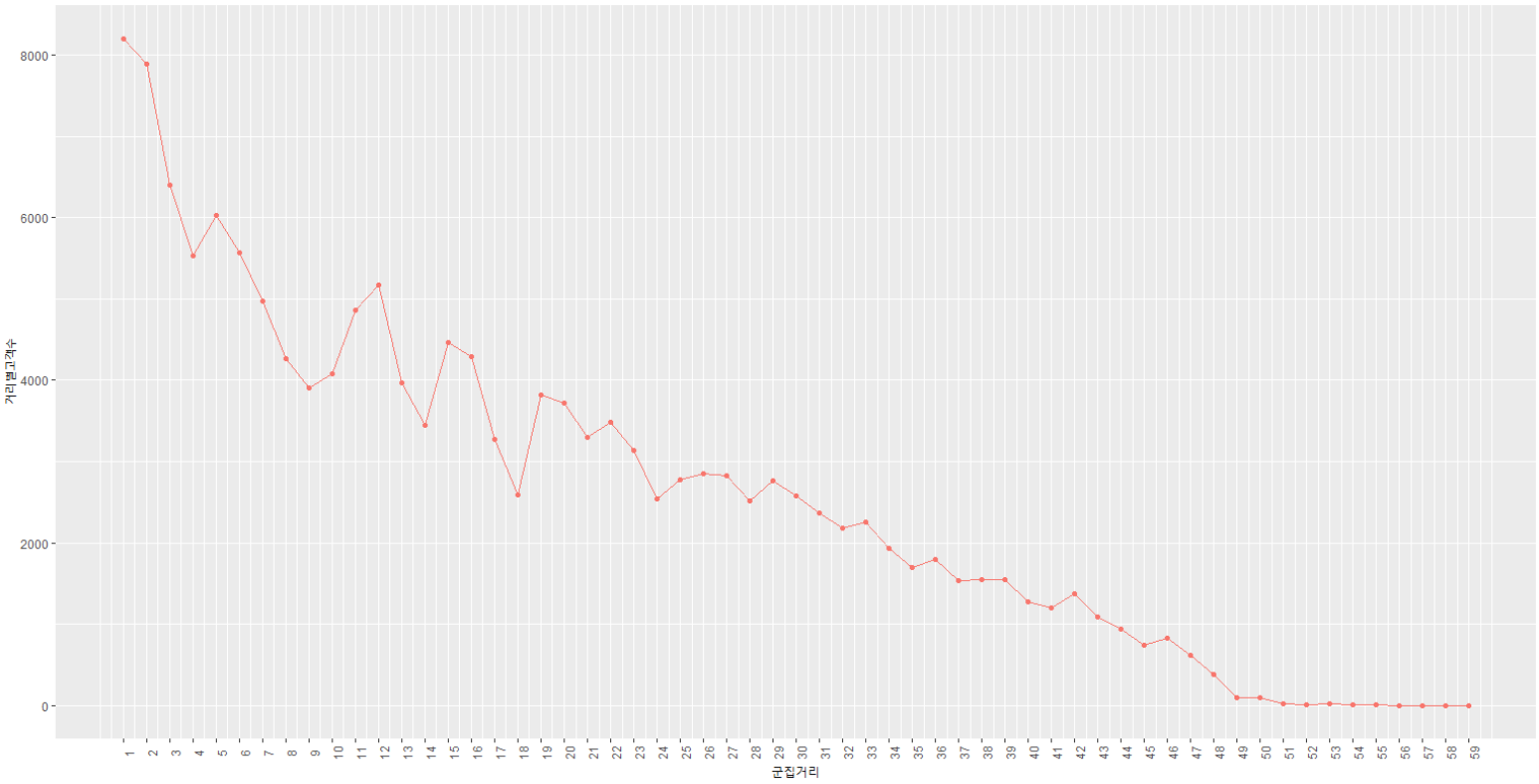
위의 그래프는 x 축의 해당 군집 기준으로 1,10,20,30,40 군집 만큼 먼 군집과 해당 군집의 교집합(두 군집 모두 구매 한적이 있는 사람) 의 수를 시각화 한 것이다. 대체적으로 군집간의 거리가 가까울수록 두 군집 모두 구매하는 사람의 수가 많아 짐을 알 수 있고, 군집간의 거리가 멀수록 교집합의 수가 매우 작아짐을 알 수 있다.

추가적으로 이 후 나오는 분석에는 60 개의 군집 중에서 14 번, 27 번 군집에 대해 분석 및 에측 할 예정이다. 14 번과 27 의 경우 가장 가까운 군집과 그다음 군집의 차이가 크지도, 작지도 않다.

14 번 군집 : "군모, 기타광학기기, 기타보석류, 남성머니클립, 남성서류가방, 남성일반지갑, 남성카드/명함지갑, 남성클러치백, 남성힙색, 노트북가방, 스포츠목걸이/팔찌, 시계세트, 여성백팩 ,여성선글라스, 여성숄더백, 여성일반지갑, 여성카드/명함지갑, 여성클러치백 ,열쇠고리, 영화/문화모바일상품권 ,캐쥬얼숄더백, 팔찌, 패션액세서리세트"

27 번 군집 : "남아바지, 남아셔츠, 남아청바지, 여아남방셔츠, 여아스커트, 여아재킷, 여아청바지, 여아코트, 여아티셔츠/탑, 여아패딩, 영유아가디건, 영유아바지, 영유아블라우스, 영유아재킷, 영유아점프수트/오버롤, 영유아패딩, 유아동스니커즈, 유아동스포츠스웨트셔츠/후드/집업"

다음 그래프는 군집간의 거리에 따른 고객 수를 나타낸 것이다. 거리가 멀어 질수록 고객수가 낮아지고 있으며 가장 가까울 경우 가장 많은 고객수를 보이고 있다.



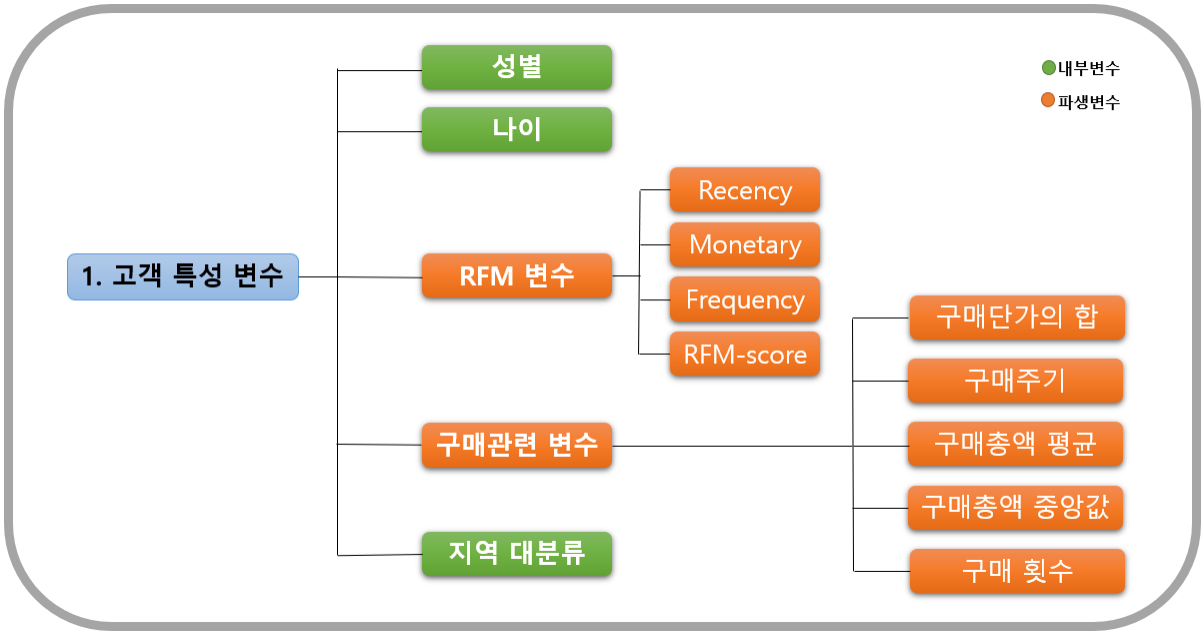
이러한 결과를 토대로 '해당 군집의 제품을 구매한 고객은 가장 가까운 군집의 제품을 구매 할 것이다.'의 가설을 뒷받침 해준다.

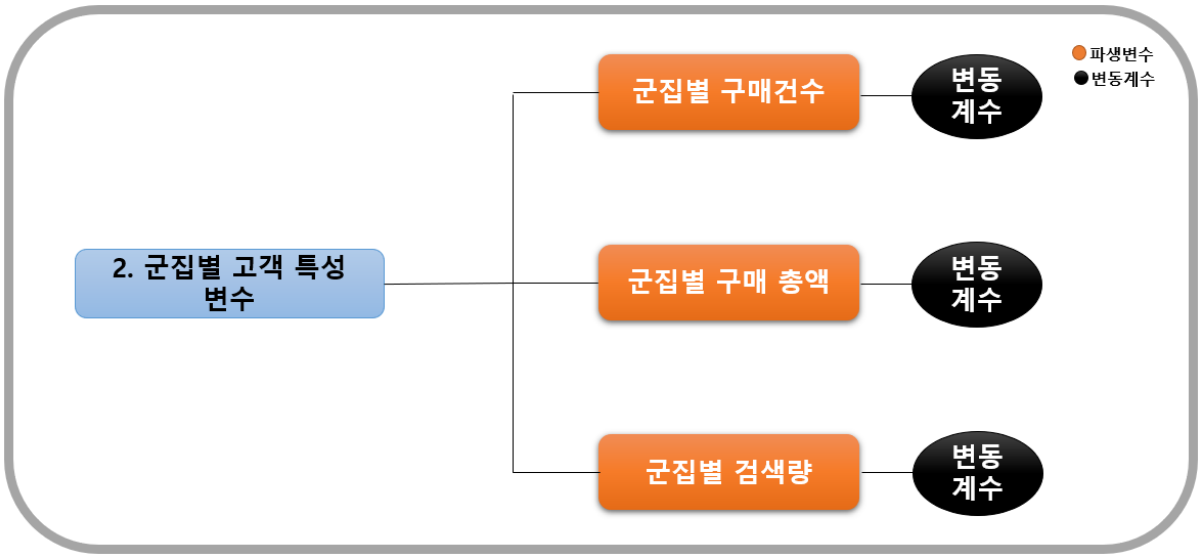
우리는 이러한 시각화 결과를 바탕으로 인사이트 및 서비스를 제안하고자 한다.

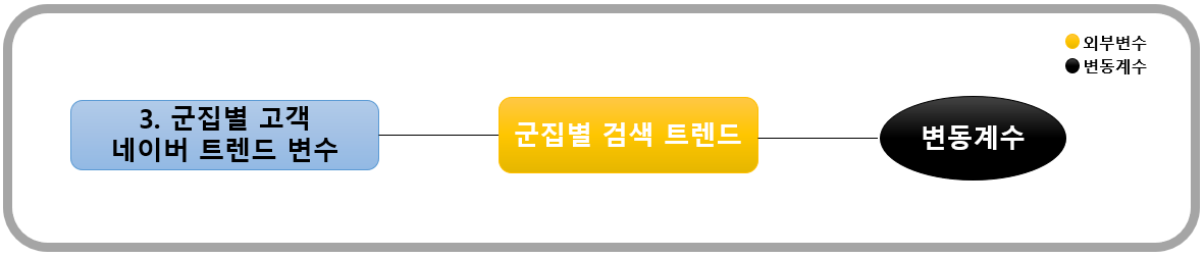
1. 새로운 고객이 해당 군집의 제품을 살 것인지 사지 않을 것인지 예측 하는 구매 트렌드 예측
2. 해당 군집과 가장 가까운 군집을 보여주며 구매 유도 하는 추천 시스템 개발

5. 변수생성 및 지수개발

(1) 변수별 프로세스





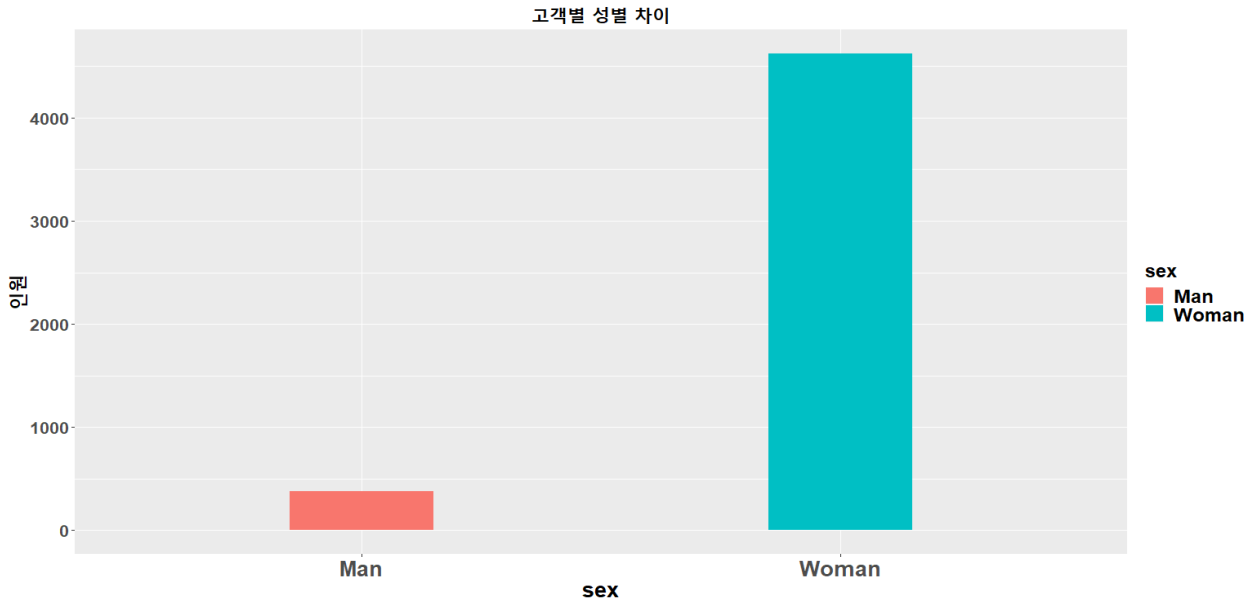


(2) 27 번 군집에 대한 변수별 분석

***27 번 군집에 대해서 Random Sample 로 뽑음.**

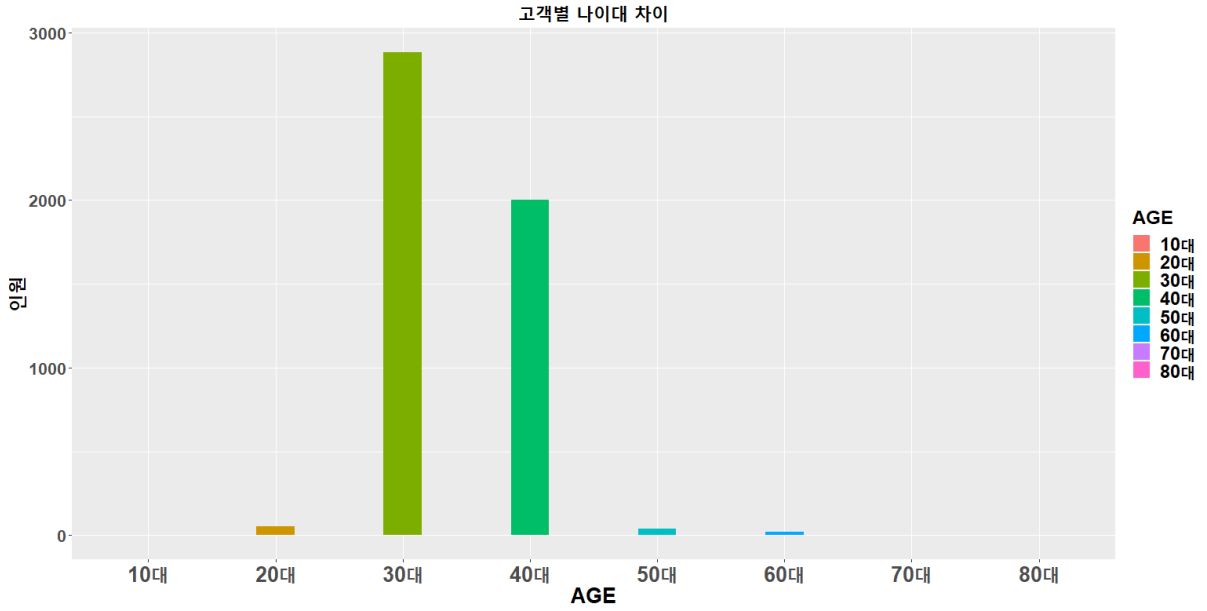
1. 고객 특성 변수

- 성별 : 해당 군집내의 고객별 성별을 뜻한다



대부분 여성 고객임을 알 수 있다.

- 나이 : 해당 군집내의 고객별 나이를 뜻한다



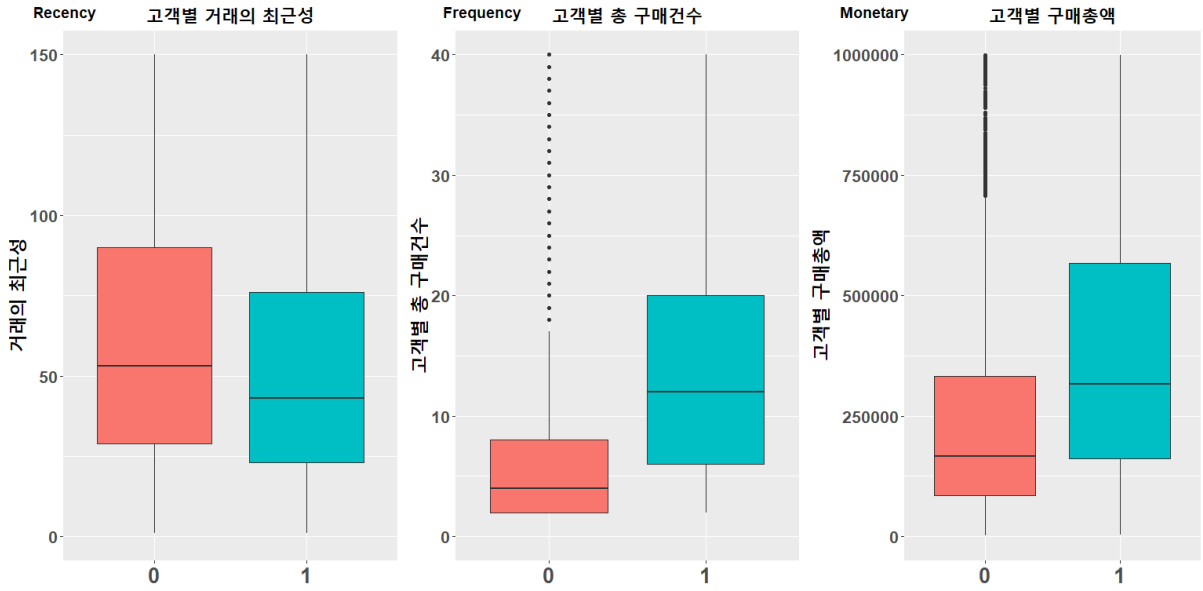
대부분 30/40 대 고객임을 알 수 있다.

- 지역 대분류 : 해당 군집내의 고객별 주요 지역 대분류를 뜻한다

-RFM 변수 : 가치 있는 고객을 추출하여 이를 기준으로 고객을 분류할 수 있는 지표로써 주로 구매 가능성이 높은 고객을 선정하기 위한 데이터 분석 방법으로 사용한다.

1. Recency : 거래의 최근성(최근 구매 후 경과 일수 - 본 분석에서는 2018/10/01 을 기준)
2. Frequency : 고객별 총 구매 건수(PD_BUY_CT 의 합)
3. Monetary : 고객별 구매 총액(PD_BUY_CT * PD_BUY_AM 의 합)

0 : 군집내 상품 비구매 고객 , 1: 군집내 상품 구매 고객



거래의 최근성은 수치가 낮을수록 기준이 되는 시점과 가까운 즉, 가장 최근에 구매한 데이터이므로 수치가 낮을수록 더 많은 정보를 반영하기 때문에 군집내 상품을 구매하는 고객일수록 낮은 추이를 보인다.

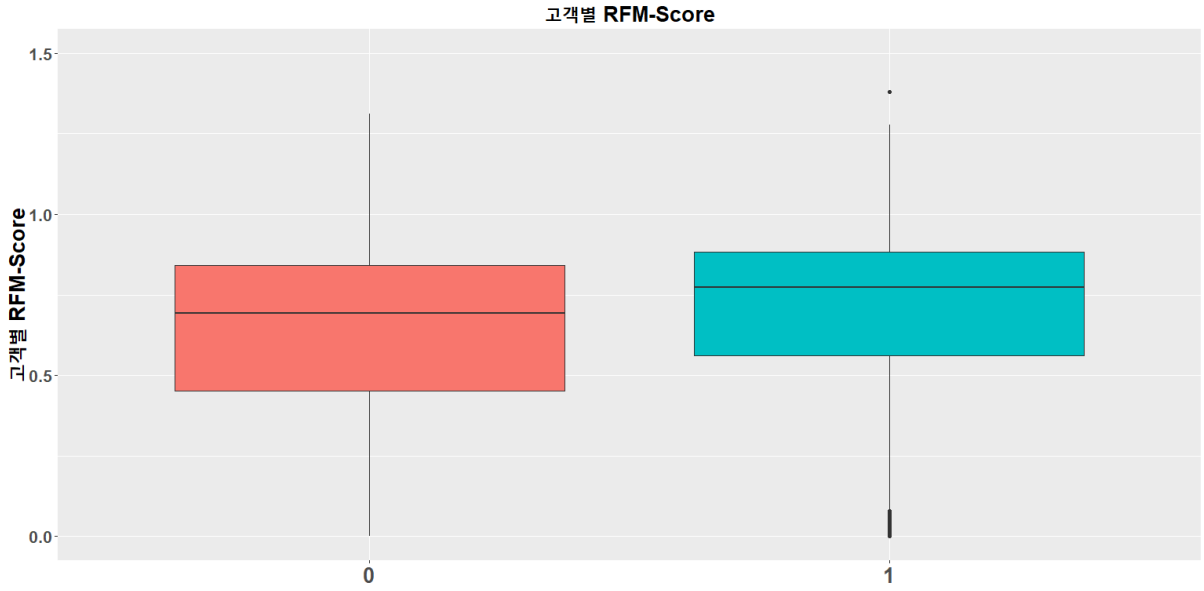
고객별 총 구매건수는 많이 구매하는 고객일수록 더 많은 정보를 반영하기 때문에 군집내 상품을 구매하는 고객일수록 높은 추이를 보인다.

고객별 구매총액도 마찬가지로 군집내 상품을 구매하는 고객일수록 높은 추이를 보인다.

따라서 고객별로 위의 세가지 변수(Recency,Frequency,Monetary)를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.

4. RFM-Score : 1,2,3 번의 변수를 이용한 고객별 점수

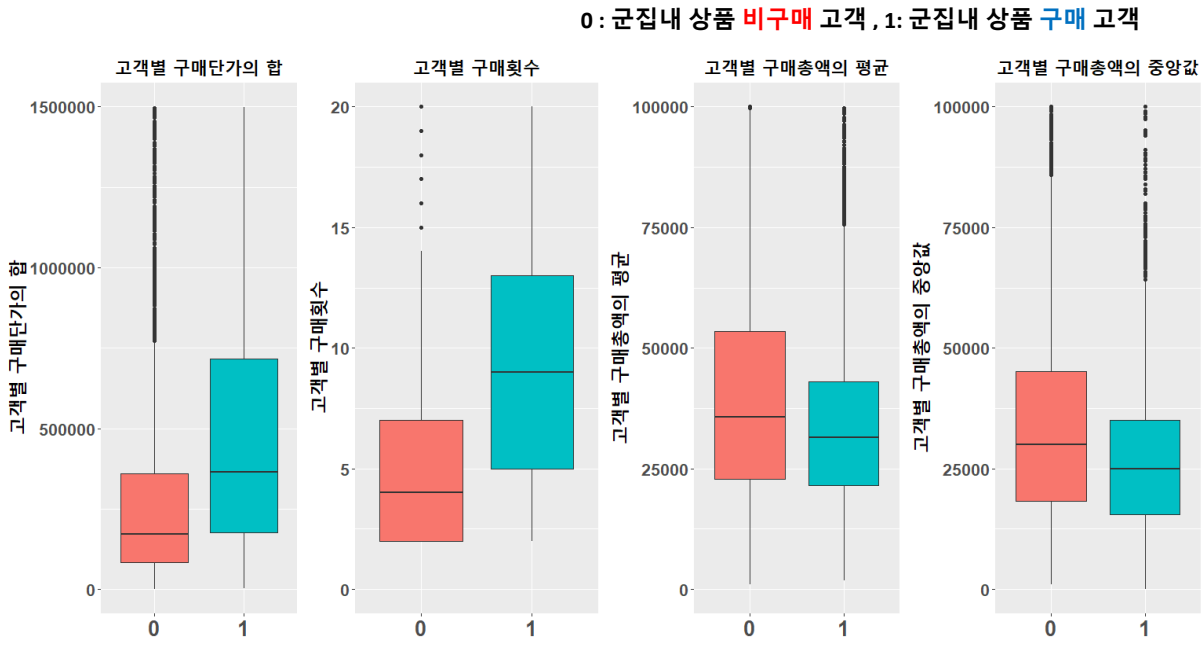
```
(1- (recency-min(recency))/(max(recency)-min(recency))) +
(frequency-min(frequency))/(max(frequency)-min(frequency)) +
(monetary-min(monetary))/(max(monetary-min(monetary))
```



해당 고객이 군집내 상품의 구매여부를 파악하기 위한 지수로써 Recency, Frequency, Monetary 변수를 사용하여 RFM-Score 를 생성하여 확인해본 결과 군집내 상품을 구매하는 고객일수록 RFM-Score 가 높은 추이를 보였다. 따라서 고객의 구매여부를 분류하기 위한 적절한 변수임을 나타낸다.

-구매 관련 변수 : 해당 군집내의 고객별 구매에 관련된 변수를 뜻한다.

1. 구매단가의 합 : 고객이 구매한 물품 단가들의 합을 뜻한다.
2. 구매총액 평균 : 고객별 구매 총액의 평균을 뜻한다.
3. 구매총액 중앙값 : 고객별 구매 총액의 중간값을 뜻한다.
4. 구매 횟수 : 고객별 구매건수가 아닌 구매 횟수를 뜻한다.



고객별 구매단가의 합과 고객별 구매횟수는 군집내 상품을 구매하는 고객과 구매하지 않는 고객을 분류하기 위한 변수로써, 군집내 상품을 구매하는 고객일수록 높은 추이를 보인다. 하지만 구매 총액의 평균과 중앙값은 군집내 상품을 구매하지 않는 고객일수록 높은 추이를 보이며 설명은 앞과 비슷하다

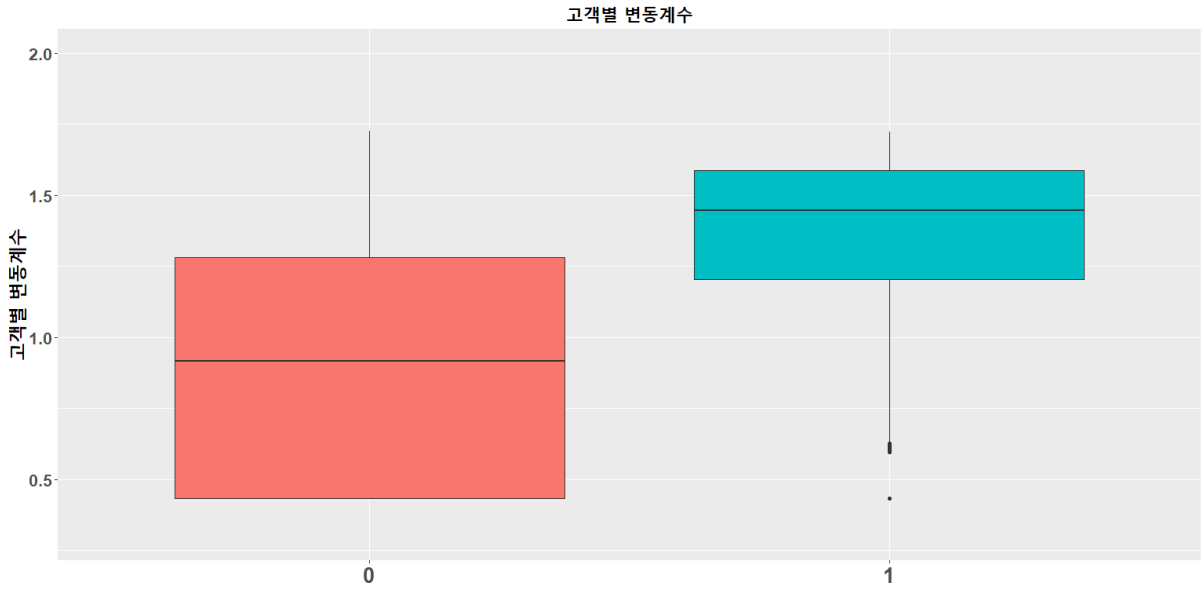
#변동계수란?

서로 다른 특성을 가지거나 데이터별 다른 속성을 가질 때 그리고 측정하는 척도가 다를 때 효율적이고 효과적으로 비교할 수 있는 값으로써 기존에 표준편차로만 비교하고 분석할 수 없는 한계를 보완해주는 역할을 한다. 변동계수는 (표준편차(sd) / 평균(mean))으로 구한다 우리는 고객별로 군집들에 대한 특성이나 속성이 각기 다르기 때문에 고객별로 군집에 대한 변동계수를 구하여 효율적으로 비교하기 위해 사용

ex)

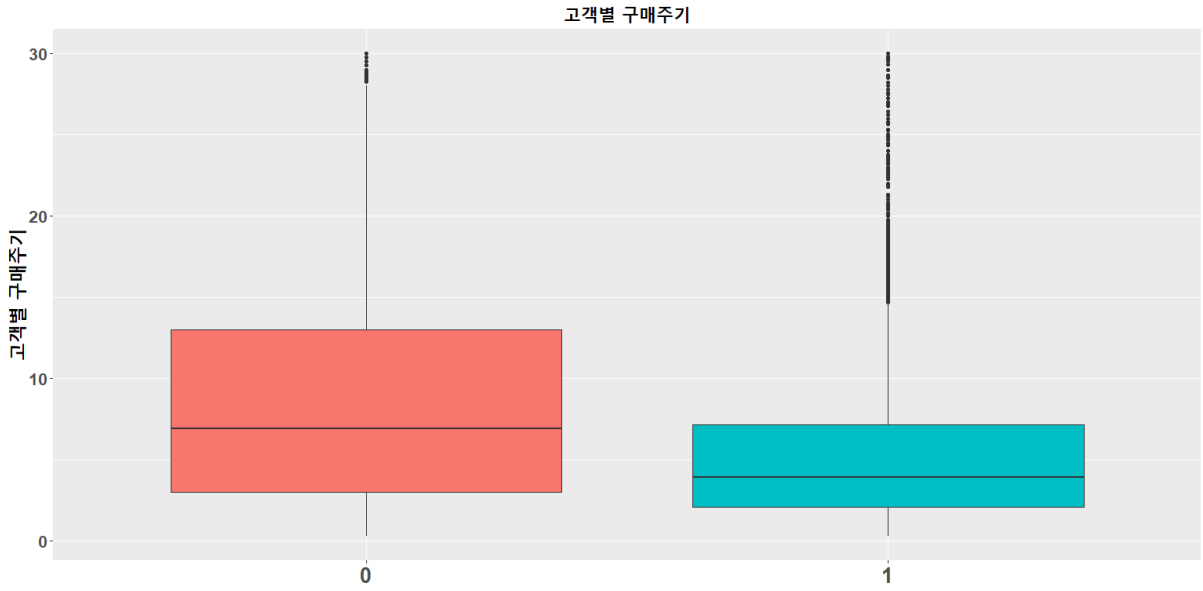
```
ser_cv<-function(x){
  sd(x)/(sum(x)/59)
}
ser_cv_f<-apply(clugun3_ser[,-1],1,ser_cv)
```

5. 구매 총액에 대한 변동계수(CV) : Monetary, 구매총액 평균, 구매총액 중앙값의 변수를 이용한 고객별 변동계수



구매총액과 관련된 변수들(Monetary, 구매총액 평균, 구매총액 중앙값)을 이용하여 변동계수를 만들어 확인해본 결과 군집내 상품을 구매하는 고객일수록 변동계수가 높은 추이를 보였다. **따라서 고객별 구매총액에 대한 변동계수를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.**

6. 구매주기 : 고객이 온라인으로 물품을 구매하는 주기를 뜻한다.

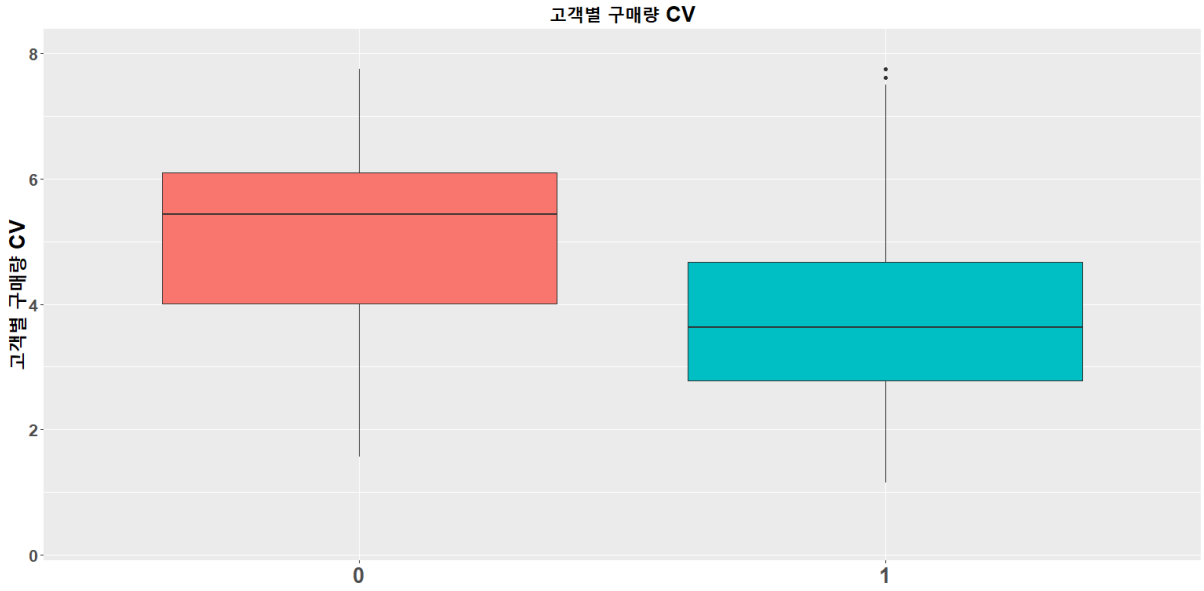


고객별 구매주기를 살펴본 결과 군집내 상품을 구매하지 않는 고객일수록 구매주기가 길다. 이를 통해서 군집내 상품을 구매하는데 있어서 구매주기가 짧으면 많은 정보를 반영하고 있는 것을 알 수 있으며 **또한 고객별 구매주기를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.**

2. 군집별 고객 특성 변수

-군집별 구매건수 : 고객별로 1~60 까지 총 60 개의 변수가 있으며 이는 1~60 까지의 군집에 대하여 고객의 군집화된 상품에 대한 구매건수를 뜻한다.

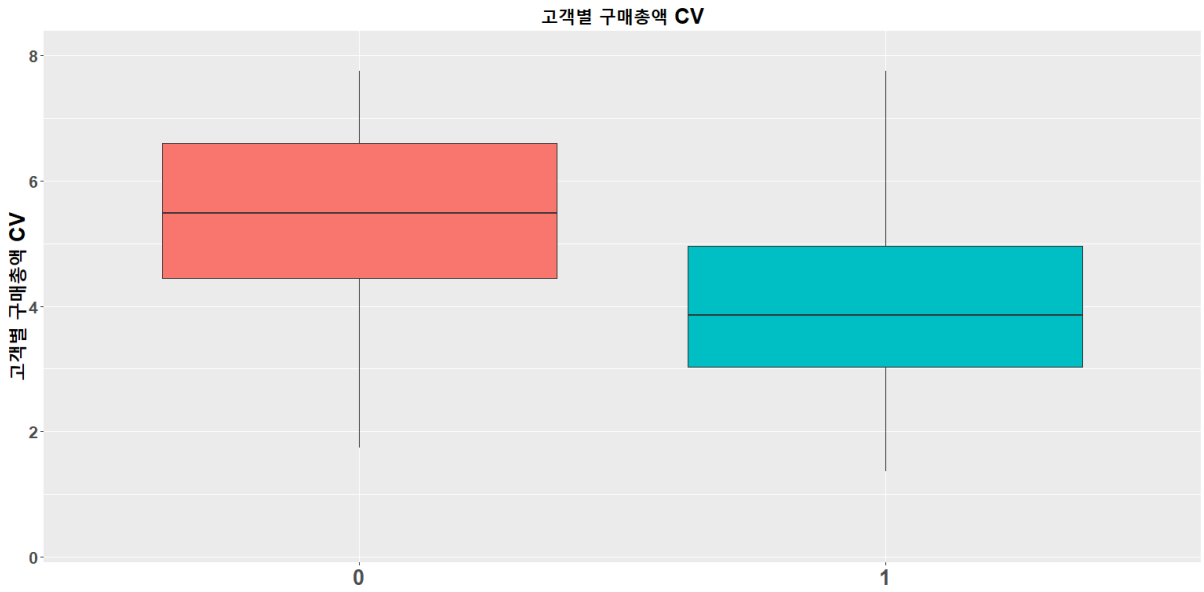
1. 군집별 구매건수의 변동계수(CV)



고객별 군집별 구매건수 변수들(군집 1~ 군집 60)을 이용하여 변동계수를 만들어 확인해본 결과 군집내 상품을 구매하지 않는 고객일수록 변동계수가 높은 추이를 보였다. **따라서 고객별 군집별 구매건수에 대한 변동계수를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.**

-군집별 구매총액 : 고객별로 1~60 까지 총 60 개의 변수가 있으며 이는 1~60 까지의 군집에 대하여 고객의 군집화된 상품에 대한 구매총액을 뜻한다.

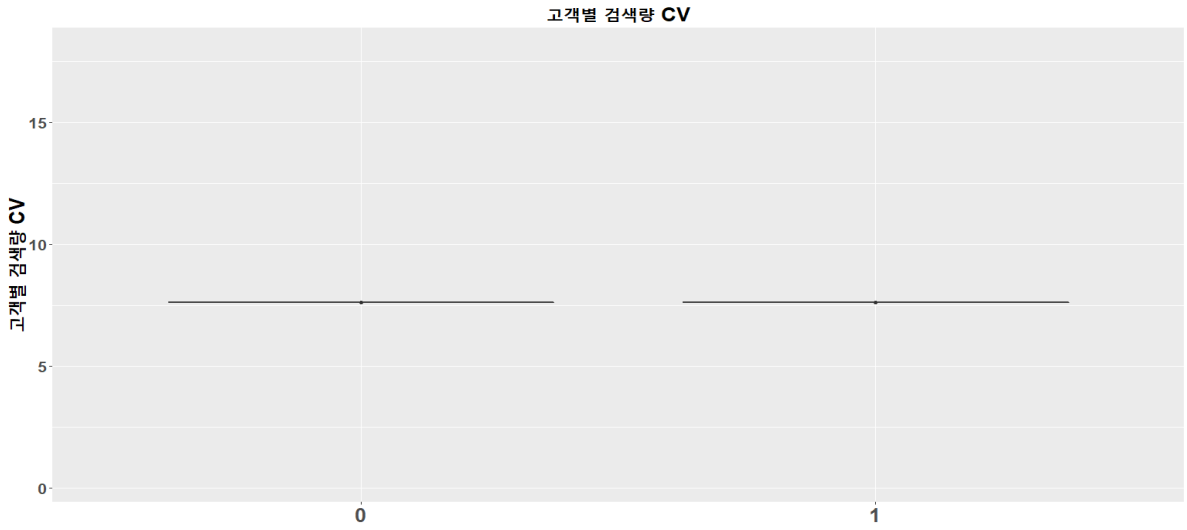
1. 군집별 구매총액의 변동계수(CV)



고객별 군집별 구매총액 변수들(군집 1~ 군집 60)을 이용하여 변동계수를 만들어 확인해본 결과 군집내 상품을 구매하지 않는 고객일수록 변동계수가 높은 추이를 보였다. **따라서 고객별 군집별 구매총액에 대한 변동계수를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.**

-군집별 검색량 : 고객별로 1~60 까지 총 60 개의 변수가 있으며 이는 1~60 까지의 군집에 대하여 고객의 군집화된 상품에 대한 검색량의 합를 뜻한다.

1. 군집별 검색량의 변동계수(CV)

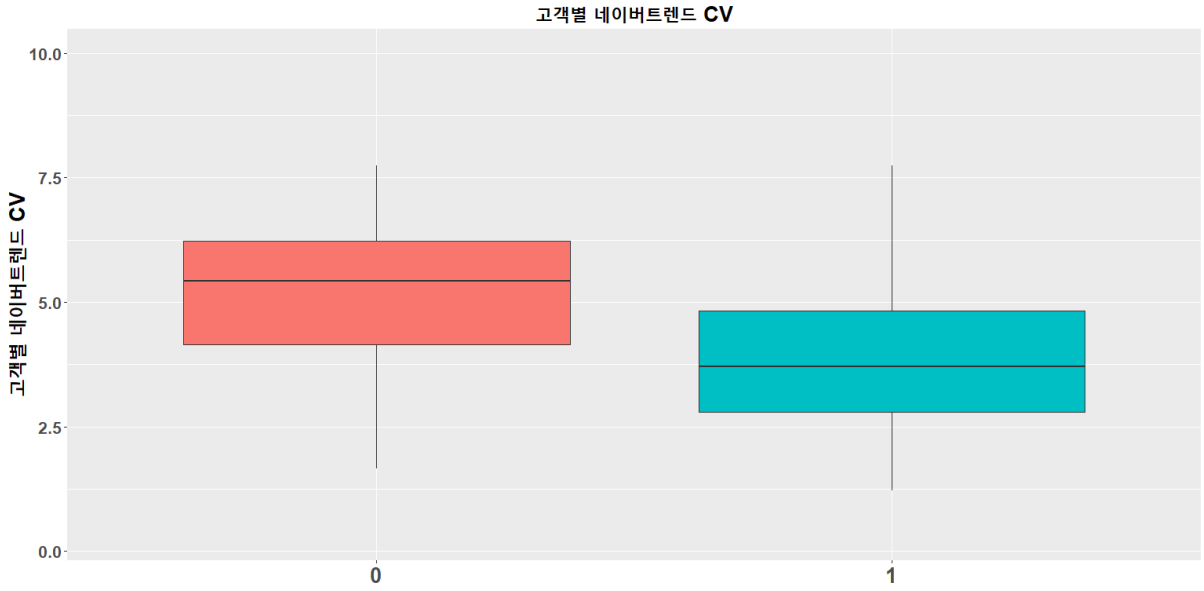


고객별 군집별 검색량의 변동계수를 확인해본 결과 어느 한쪽이 높은 추이를 보이거나 분포가 제대로 형성되지 않은 것을 확인할 수 있다. **따라서 고객별 군집별 검색량에 대한 변동계수를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 없다.**

2. 군집별 고객 네이버 트렌드 변수

-군집별 검색트렌드(Naver Trend) : 고객별로 1~60 까지 총 60 개의 변수가 있으며 이는 1~60 까지의 군집에 대하여 고객의 군집화된 상품에 대한 네이버 검색 트렌드의 합을 뜻한다.

1. 군집별 검색량의 변동계수(CV)



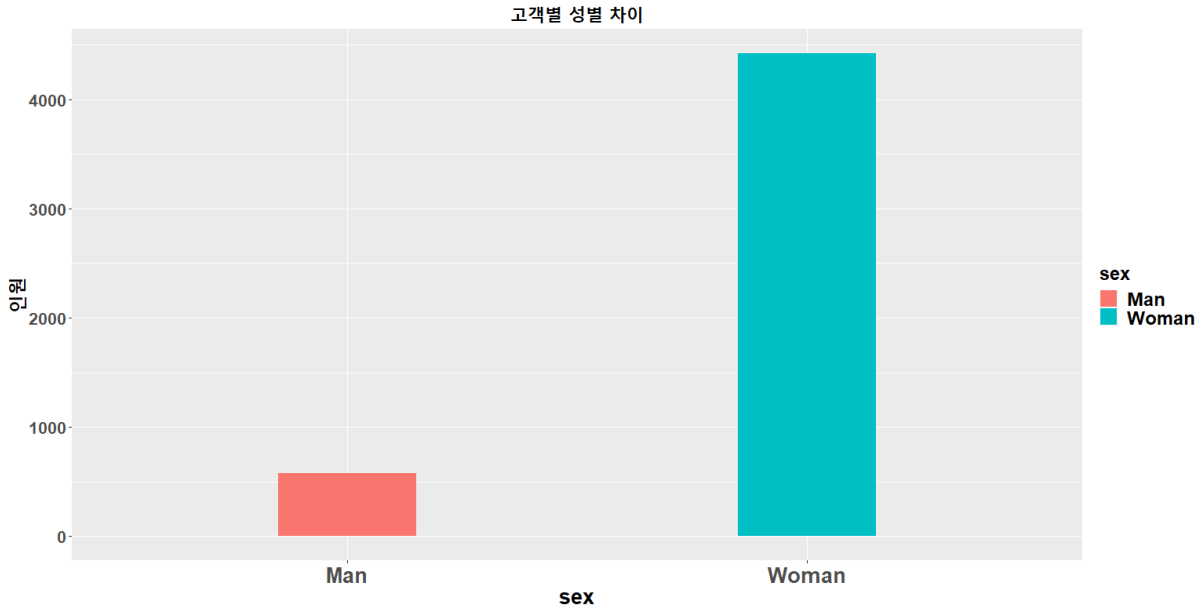
고객별 군집별 네이버 검색 트렌드 변수들(군집 1~ 군집 60)을 이용하여 변동계수를 만들어 확인해본 결과 군집내 상품을 구매하지 않는 고객일수록 변동계수가 높은 추이를 보였다. **따라서 고객별 군집별 네이버 검색 트렌드에 대한 변동계수를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.**

(2) 14 번 군집에 대한 변수별 분석

***14 번 군집에 대해서 Random Sample 로 뽑음.**

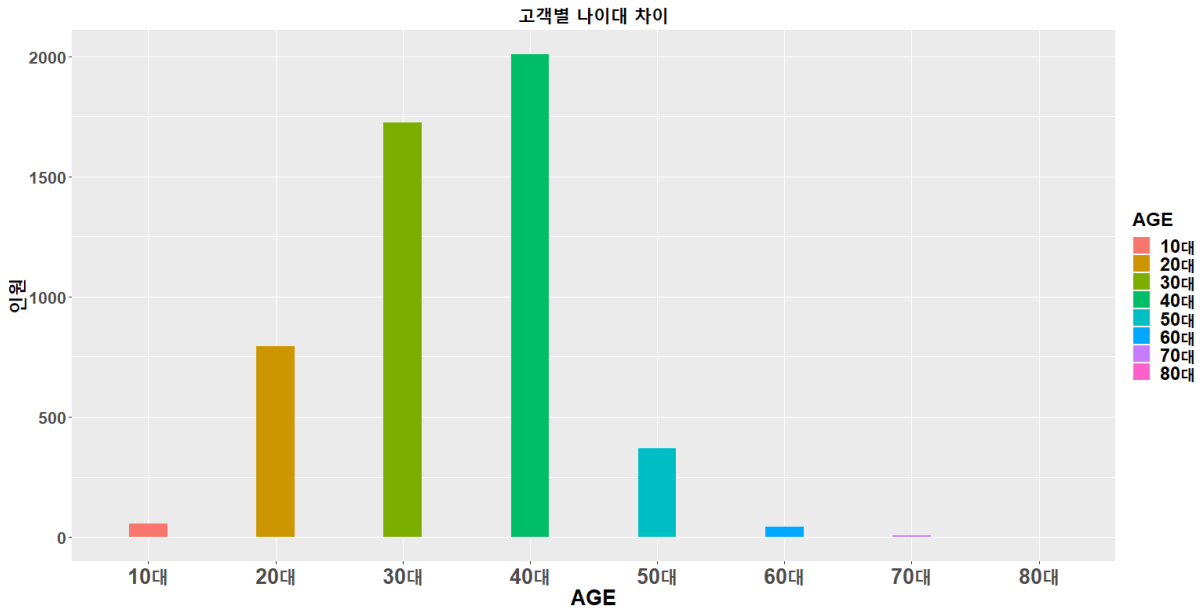
1. 고객 특성 변수

- 성별 : 해당 군집내의 고객별 성별을 뜻한다



대부분 여성 고객임을 알 수 있다.

- 나이 : 해당 군집내의 고객별 나이를 뜻한다



대부분 30/40 대 고객이지만 14 번 군집에서는 20 대와 50 대의 분포도 확인할 수 있다.

- 지역 대분류 : 해당 군집내의 고객별 주요 지역 대분류를 뜻한다

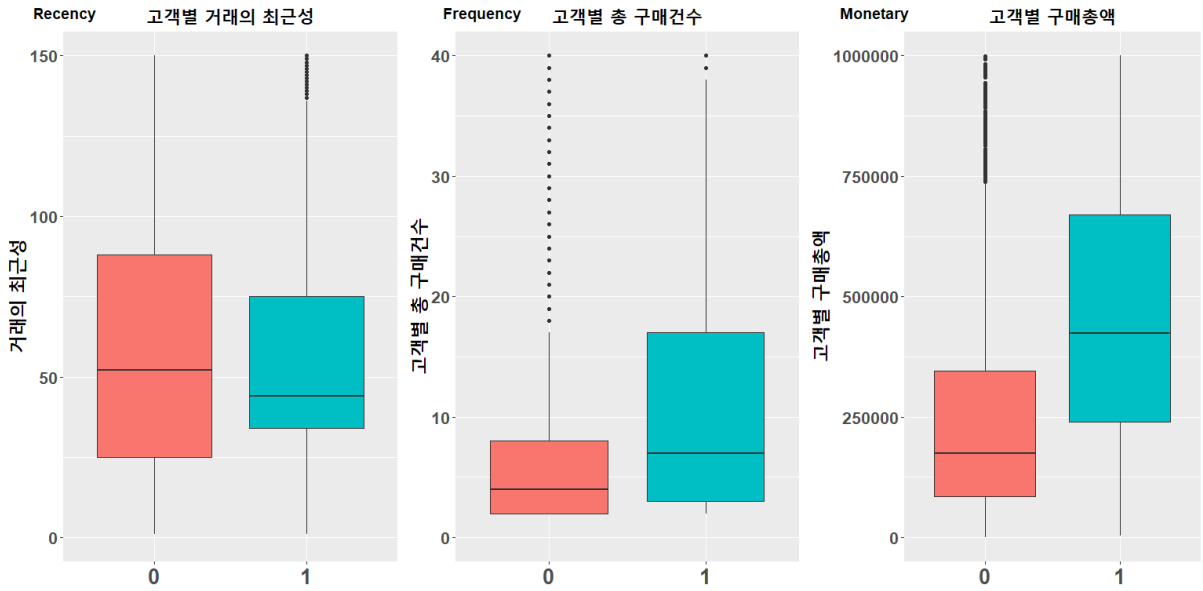
-RFM 변수 : 가치 있는 고객을 추출하여 이를 기준으로 고객을 분류할 수 있는 지표로써 주로 구매 가능성이 높은 고객을 선정하기 위한 데이터 분석 방법으로 사용한다.

1. Recency : 거래의 최근성(최근 구매 후 경과 일수 - 본 분석에서는 2018/10/01 을 기준)

2. Frequency : 고객별 총 구매 건수(PD_BUY_CT 의 합)

3. Monetary : 고객별 구매 총액(PD_BUY_CT * PD_BUY_AM 의 합)

0 : 군집내 상품 비구매 고객 , 1: 군집내 상품 구매 고객



27 번 군집과 비슷한 추세를 나타내는 것을 볼 수 있다. 마찬가지로 거래의 최근성은 수치가 낮을수록 기준이 되는 시점과 가까운 즉, 가장 최근에 구매한 데이터이므로 수치가 낮을수록 더 많은 정보를 반영하기 때문에 군집내 상품을 구매하는 고객일수록 낮은 추이를 보인다.

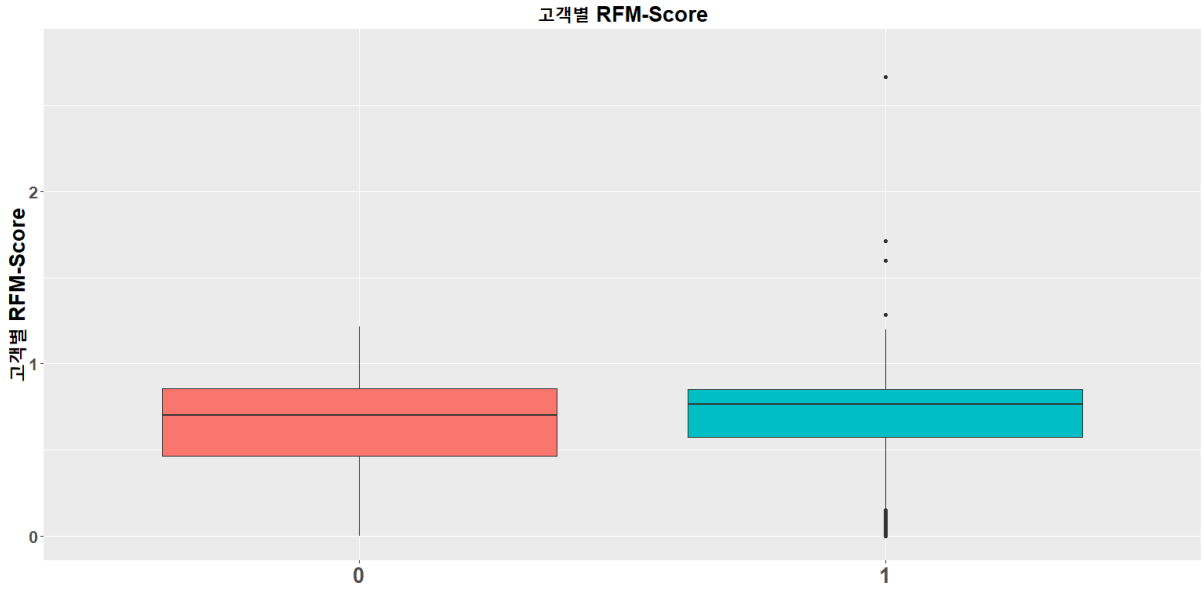
마찬가지로 고객별 총 구매건수는 많이 구매하는 고객일수록 더 많은 정보를 반영하기 때문에 군집내 상품을 구매하는 고객일수록 높은 추이를 보인다.

고객별 구매총액도 마찬가지로 군집내 상품을 구매하는 고객일수록 높은 추이를 보인다.

따라서 고객별로 위의 세가지 변수(Recency,Frequency,Monetary)를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.

4. RFM-Score : 1,2,3 번의 변수를 이용한 고객별 점수

```
(1- (recency-min(recency))/(max(recency)-min(recency))) +
(frequency-min(frequency))/(max(frequency)-min(frequency)) +
(monetary-min(monetary))/(max(monetary-min(monetary))
```

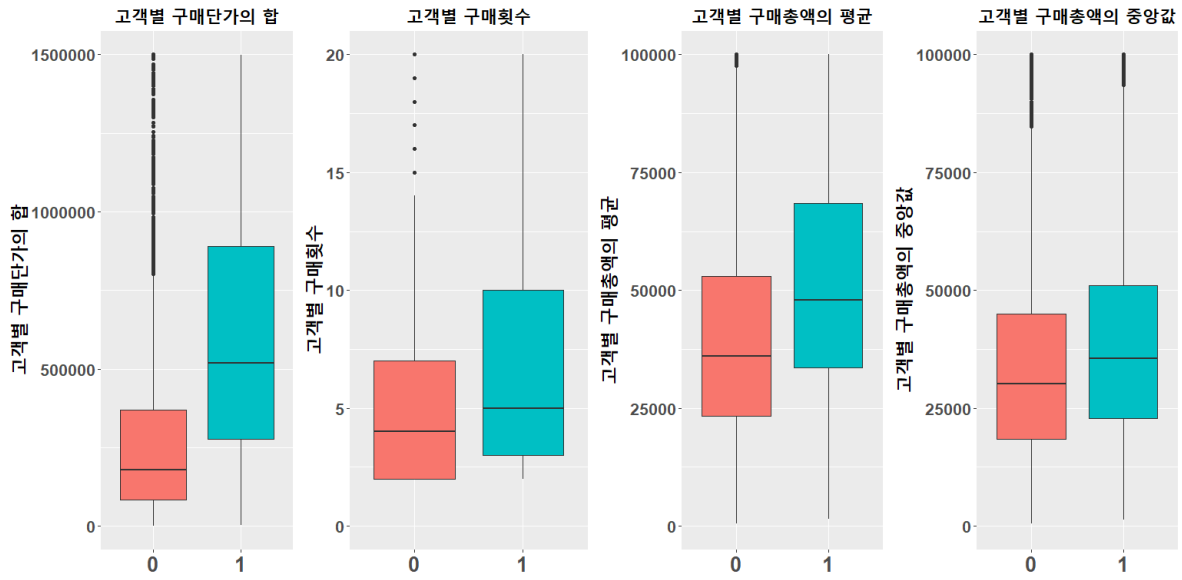


해당 고객이 군집내 상품의 구매여부를 파악하기 위한 지수로써 Recency, Frequency, Monetary 변수를 사용하여 RFM-Score 를 생성하여 확인해본 결과 군집내 상품을 구매하는 고객일수록 RFM-Score 가 27 번 군집에 비해서는 미세하지만 높은 추이를 보였다. 따라서 고객별 RFM-Score 는 고객의 구매여부를 분류하기 위한 적절한 변수임을 나타낸다.

-구매 관련 변수 : 해당 군집내의 고객별 구매에 관련된 변수를 뜻한다.

1. **구매단가의 합** : 고객이 구매한 물품 단가들의 합을 뜻한다.
2. **구매총액 평균** : 고객별 구매 총액의 평균을 뜻한다.
3. **구매총액 중앙값** : 고객별 구매 총액의 중간값을 뜻한다.
4. **구매 횟수 :** 고객별 구매건수가 아닌 구매 횟수를 뜻한다.

0 : 군집내 상품 비구매 고객 , 1: 군집내 상품 구매 고객

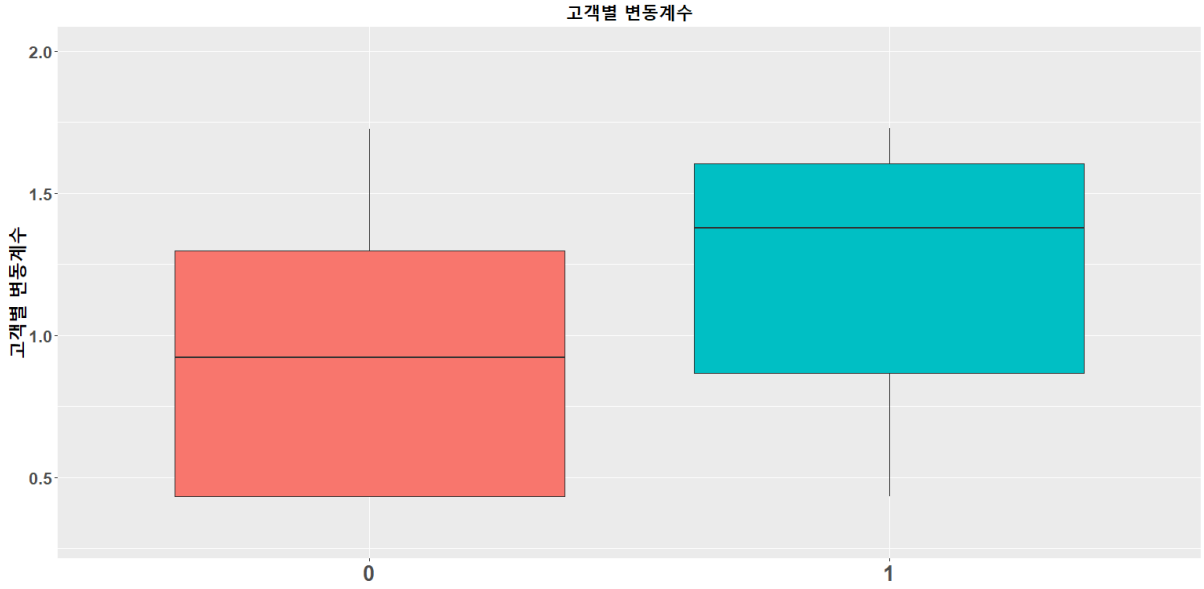


고객별 구매단가의 합과 고객별 구매횟수는 군집내 상품을 구매하는 고객과 구매하지 않는 고객을 분류하기 위한 변수로써, 군집내 상품을 구매하는 고객일수록 높은 추이를 보인다. 하지만 군집별로 속하는 상품이 다르므로 다른 특성 가질 수 있기 때문에 해당 14 번 군집은 27 번 군집과는 다른 특성을 보이는 것을 확인할 수 있다. 그것이 구매 총액의 평균과 중앙값인데 27 번 군집에서는 군집내 상품을 구매하지 않는 고객일수록 높은 추이를 보이지만 14 번 군집은 군집내 상품을 구매한 고객일수록 높은 추이를 보인다. 따라서 위의 4 가지 변수들은 고객의 구매여부를 분류하기 위한 적절한 변수임을 나타낸다.

#변동계수란?

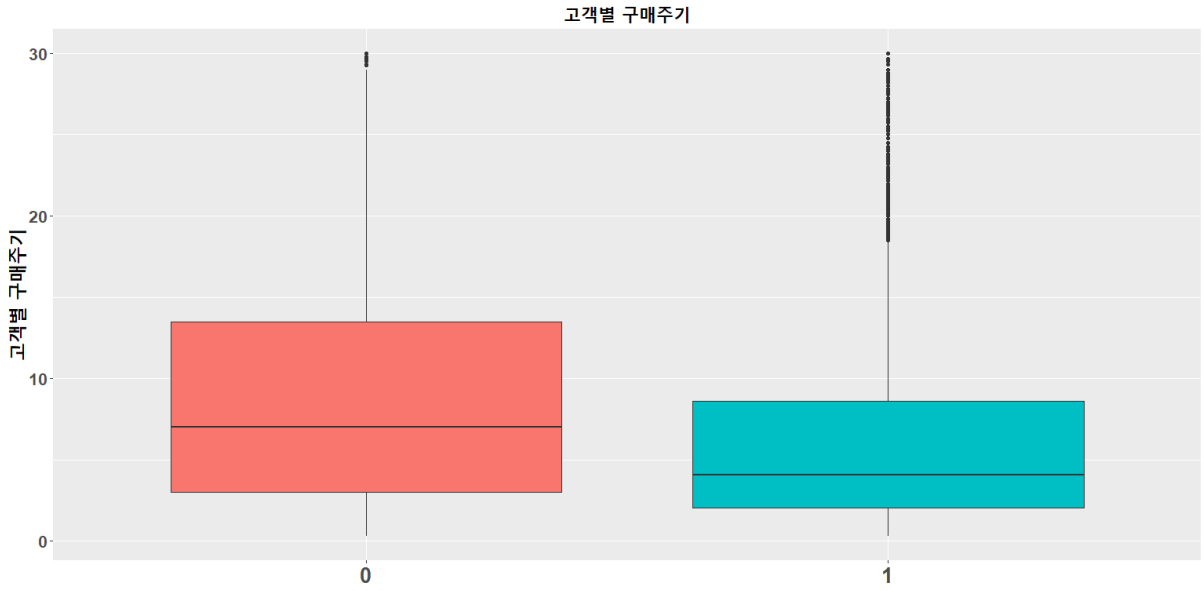
서로 다른 특성을 가지거나 데이터별 다른 속성을 가질 때 그리고 측정하는 척도가 다를 때 효율적이고 효과적으로 비교할 수 있는 값으로써 기존에 표준편차로만 비교하고 분석할 수 없는 한계를 보완해주는 역할을 한다. 변동계수는 (표준편차(sd) / 평균(mean))으로 구한다 우리는 고객별로 군집들에 대한 특성이나 속성이 각기 다르기 때문에 고객별로 군집에 대한 변동계수를 구하여 효율적으로 비교하기 위해 사용.

5. 구매 총액에 대한 변동계수(CV) : Monetary, 구매총액 평균, 구매총액 중앙값의 변수를 이용한 고객별 변동계수



27 번 군집의 특성과 마찬가지로 구매총액과 관련된 변수들(Monetary, 구매총액 평균, 구매총액 중앙값)을 이용하여 변동계수를 만들어 확인해본 결과 군집내 상품을 구매하는 고객일수록 변동계수가 높은 추이를 보였다. **따라서 고객별 구매총액에 대한 변동계수를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.**

6. 구매주기 : 고객이 온라인으로 물품을 구매하는 주기를 뜻한다.



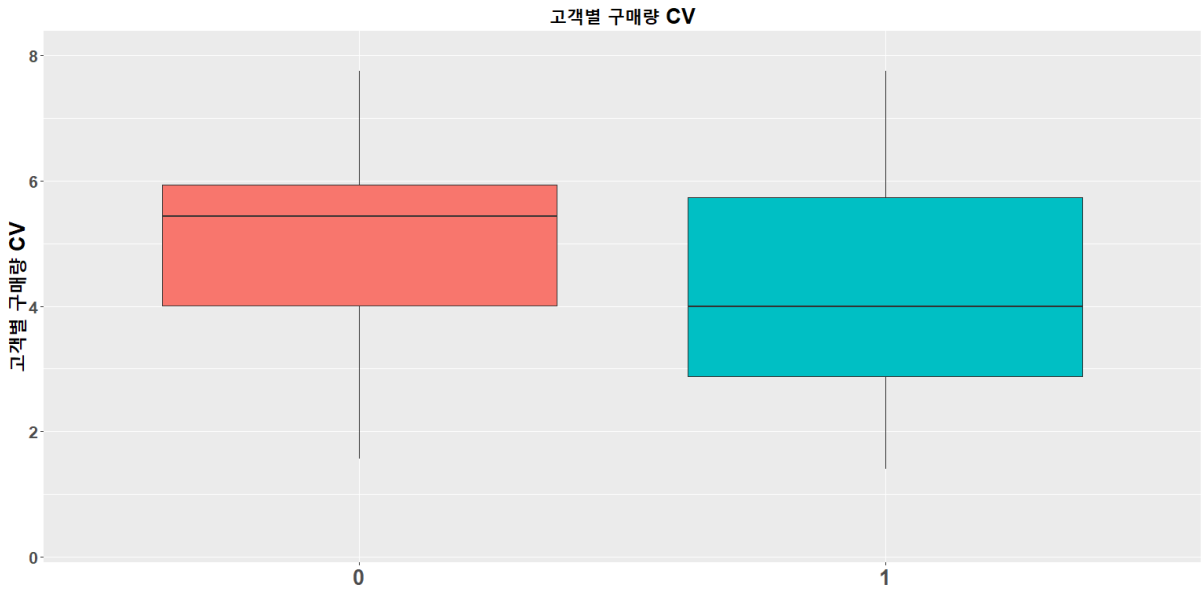
27 번 군집의 특성과 마찬가지로 고객별 구매주기를 살펴본 결과 군집내 상품을 구매하지 않는 고객일수록 구매주기가 길다. 이를 통해서 군집내 상품을 구매하는데 있어서 구매주기가 짧으면 많은 정보를 반영하고 있는 것을 알 수 있다.

14,27 군집의 경우 구매주기가 길수록 군집내 상품을 구매하지 않는 고객이었는데 반대로 생각해보면 군집내 상품을 구매하는 고객일수록 구매주기가 길수도 있다. 가전제품, 카메라, 컴퓨터 등이 이에 속한다. **따라서 고객별 구매주기를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.**

2. 군집별 고객 특성 변수

-군집별 구매건수 : 고객별로 1~60 까지 총 60 개의 변수가 있으며 이는 1~60 까지의 군집에 대하여 고객의 군집화된 상품에 대한 구매건수를 뜻한다.

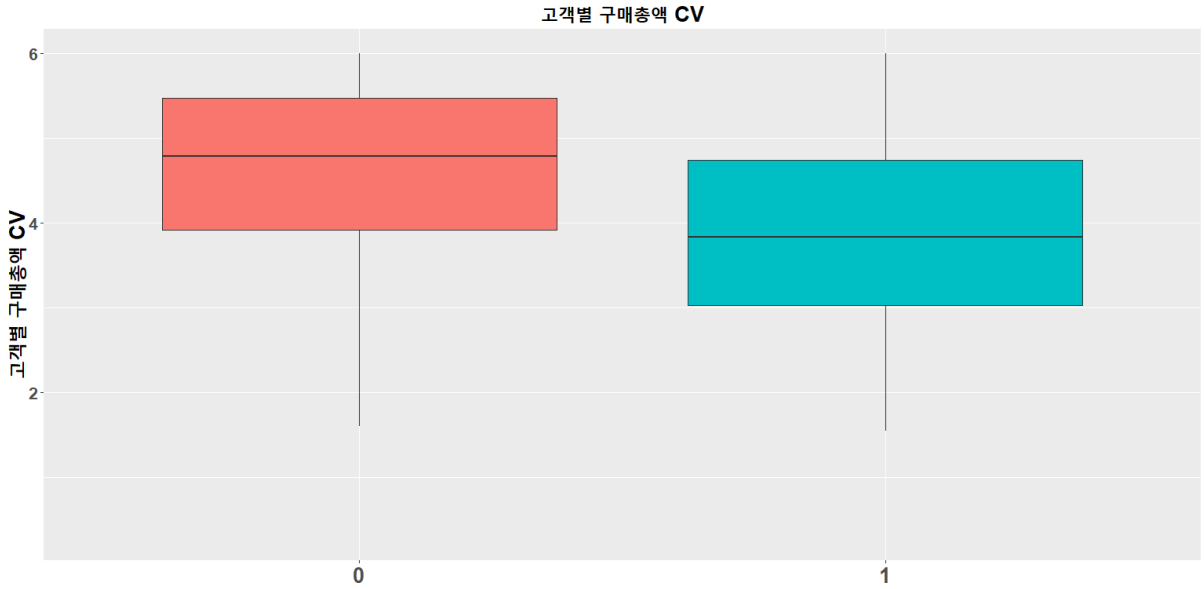
1. 군집별 구매건수의 변동계수(CV)



27 번 군집과 마찬가지로 고객별 군집별 구매건수 변수들(군집 1~ 군집 60)을 이용하여 변동계수를 만들어 확인해본 결과 군집내 상품을 구매하지 않는 고객일수록 변동계수가 높은 추이를 보였다. **따라서 고객별 군집별 구매건수에 대한 변동계수를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.**

-군집별 구매총액 : 고객별로 1~60 까지 총 60 개의 변수가 있으며 이는 1~60 까지의 군집에 대하여 고객의 군집화된 상품에 대한 구매총액을 뜻한다.

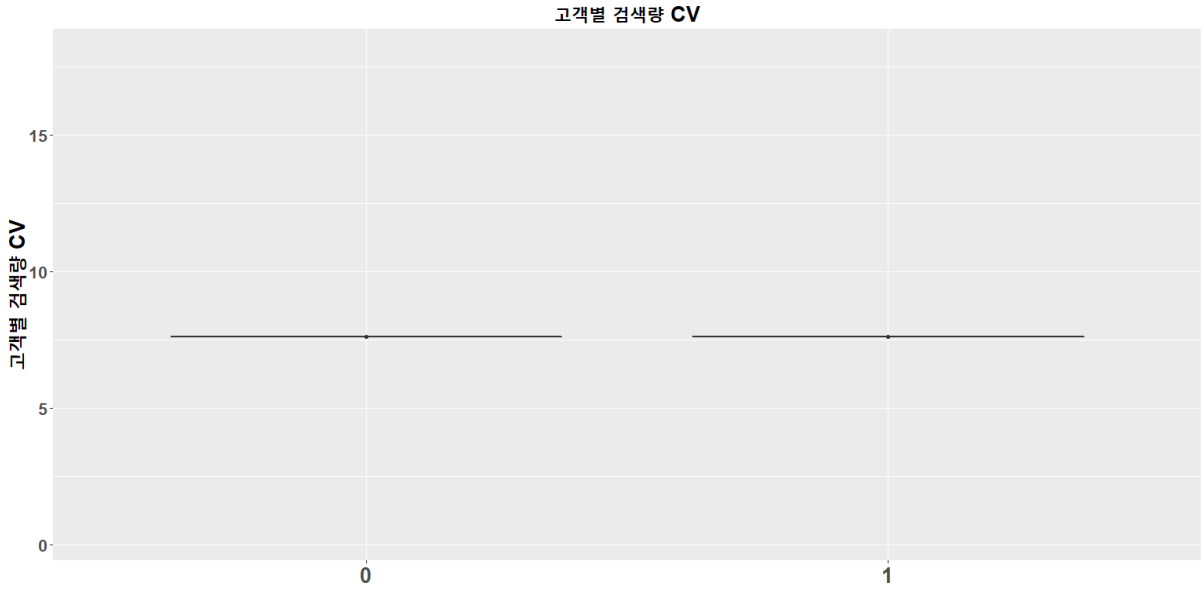
1. 군집별 구매총액의 변동계수(CV)



27 번 군집과 비슷한 특성으로 고객별 군집별 구매총액 변수들(군집 1~ 군집 60)을 이용하여 변동계수를 만들어 확인해본 결과 군집내 상품을 구매하지 않는 고객일수록 변동계수가 높은 추이를 보였다. **따라서 고객별 군집별 구매총액에 대한 변동계수를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.**

-군집별 검색량 : 고객별로 1~60 까지 총 60 개의 변수가 있으며 이는 1~60 까지의 군집에 대하여 고객의 군집화된 상품에 대한 검색량의 합를 뜻한다.

1. 군집별 검색량의 변동계수(CV)

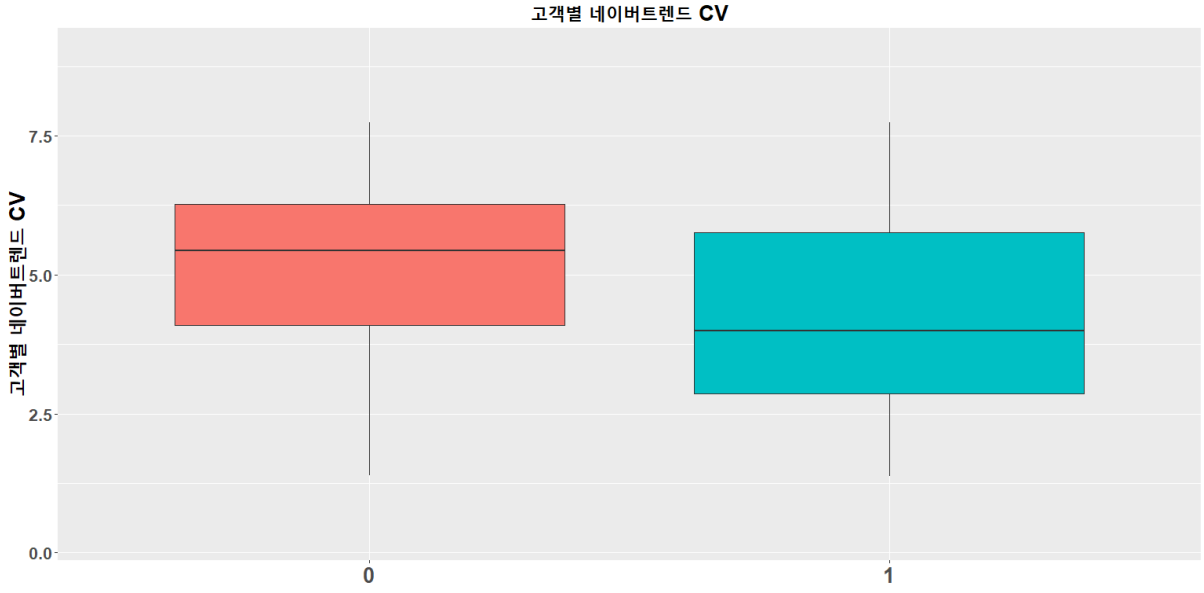


14 번 군집 또한 마찬가지로 고객별 군집별 검색량의 변동계수를 확인해본 결과 어느 한쪽이 높은 추이를 보이거나 분포가 제대로 형성되지 않은 것을 확인할 수 있다. **따라서 고객별 군집별 검색량에 대한 변동계수를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 없다.**

2. 군집별 고객 네이버 트렌드 변수

-**군집별 검색트렌드(Naver Trend)** : 고객별로 1~60 까지 총 60 개의 변수가 있으며 이는 1~60 까지의 군집에 대하여 고객의 군집화된 상품에 대한 네이버 검색 트렌드의 합을 뜻한다.

1. 군집별 검색량의 변동계수(CV)



14 번 군집 또한 고객별 군집별 네이버 검색 트렌드 변수들(군집 1~ 군집 60)을 이용하여 변동계수를 만들어 확인해본 결과 군집내 상품을 구매하지 않는 고객일수록 변동계수가 높은 추이를 보였다. **따라서**

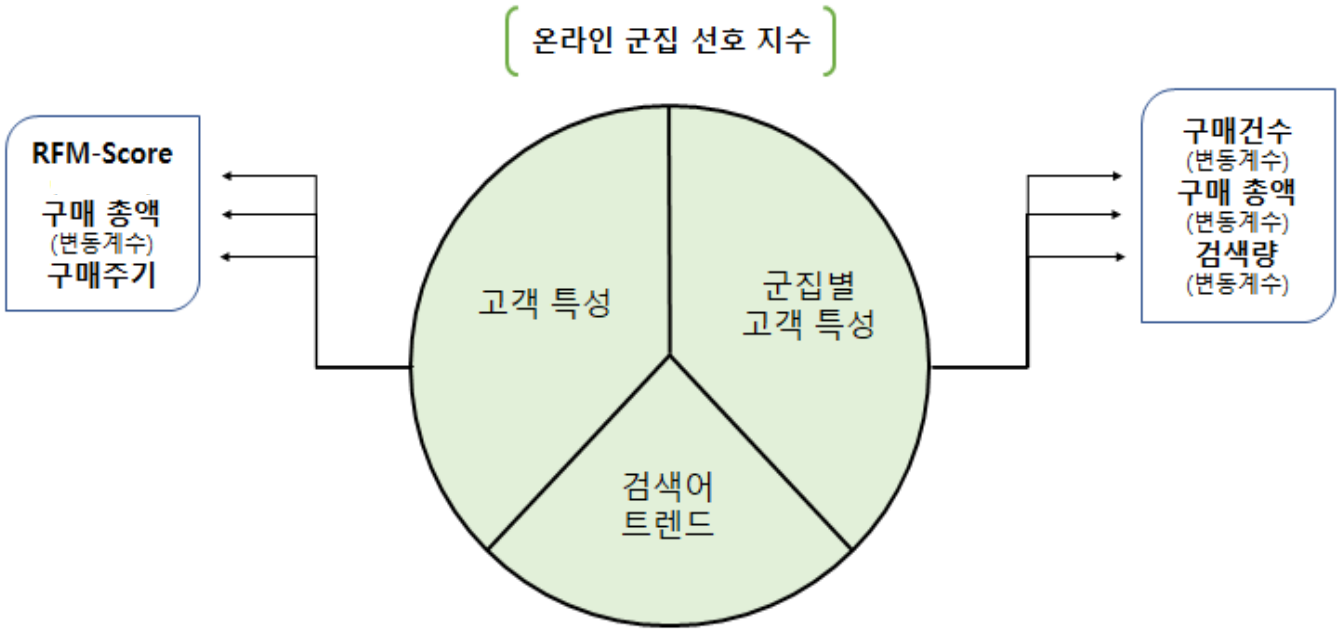
고객별 군집별 네이버 검색 트렌드에 대한 변동계수를 통해서 고객의 구매여부를 분류할 수 있는 적절한 변수임을 나타낼 수 있다.

→위와 같은 분석을 통해 군집별로 어떤 특성을 가지는지 파악할 수 있다.

(4) 온라인 군집 선호지수 개발

각 고객은 서로 비슷할 수도 있지만 고객별로 속성들은 모두 상이할 것이다. “그렇다면 고객별 각 군집에 대한 여러 특성 및 속성들을 효과적이고 효율적으로 비교할 수 있는 방법이 있을까?” 라는 의문을 갖던 중 변동계수를 사용하기로 했다. 변동계수는 앞서 말한 것처럼 데이터별 속성 수치가 다르거나 측정하는 척도가 다를 때 그것을 효율적으로 비교해줄 수 있는 유용한 지수이다. 따라서 고객별 군집별 여러 특성들에 대하여 고객마다 변동계수를 구하면 고객별로 효과적이로 효율적으로 비교할 수 있을뿐더러 고객마다 변동계수를 이용하여 온라인 군집 선호지수를 개발할 수 있다.

- **앞서 만든 변수를 통한 지수 개발**

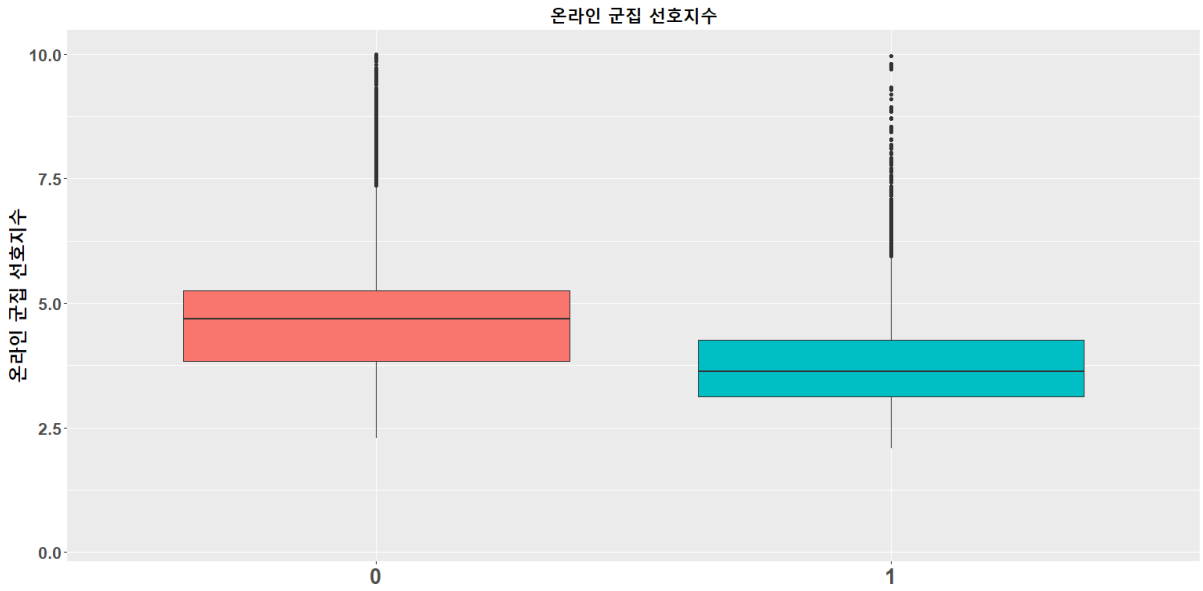


*고객 특성 지수 = (RFM-Score + 구매 총액 + 구매 주기) / 3
*군집별 고객 특성 지수 = (구매 건수 + 구매 총액 + 검색량) / 3
*검색어 트렌드 = 검색어 트렌드 변동계수

***온라인 군집 선호 지수 = 고객 특성 지수*0.4 + 군집별 고객 특성 지수*0.4 + 검색어 트렌드*0.2

➔ 검색어 트렌드로 만든 지수는 변수로써 모델의 설명력에 영향이 있지만 고객특성, 군집별 고객특성의 지수에 비해 큰 영향력을 끼치지 못하는 지수이므로 “온라인 군집 선호 지수” 를 개발할 때 가중치를 차등하였다

온라인 군집 선호지수를 가지고 해당 군집의 상품 구매 여부를 확인 해보니, 온라인 군집 선호지수는 상품 구매하는 고객들에게 영향을 덜 끼치는 것을 확인 할 수 있었다. 이를 통해 온라인 군집 선호지수를 활용해 고객의 구매 여부 예측에 활용 될 수 있다는 것을 알 수 있다.



6. 고객별 해당 군집 상품 구매 예측

(1) 데이터셋 생성

1. 14 번 군집과 27 번 군집 모두 동일한 방식으로 진행 한다.
2. 해당 군집의 제품을 구매한 사람은 1, 그렇지 않은 사람은 0
3. 1 과 0 의 비율은 1 대 1 로써 5000 명씩 만명의 데이터셋을 사용한다.
4. 변수는 고객특성변수, 군집별 고객 특성변수, 네이버 트렌드 변수를 사용한다.

(2) 모델링

1. Xgboost 사용하는 이유

- 부트스트랩을 적용하는 배깅의 과정과 유사하지만 가장 큰 차이는 부트스트랩 표본을 구성하는 재표본 과정에서 각 자료에 동일한 확률을 부여하는 것이 아니라, 분류가 잘못된 데이터에 더 큰 가중을 주어 표본을 추출
- 오분류된 데이터에가중치를 부여하여 다음 부트스트랩 표본을 구성할시 오분류된 데이터에 집중해서 학습 진행
- 위 과정을 반복하여 만들어진 여러 개의 모델들로 voting or averaging
- **Bias 를 낮추기 좋음과 동시에 서로 다른 여러 모델들을 모두 고려하므로 variance 도 낮게 유지 할 수 있다**.
- 분산/병렬 처리 모델
- **Sparsity awareness** 가 가능, zero 데이터를 건너뛰면서 학습이 가능
- **Randomforest, DNN, SVM 시도 했으나** 용량이 큰 데이터 문제로 진행 불가
- **여러가지 복합적인 요소를 고려하여 Xgboost 를 사용**

2. 파라미터 조정

- Grid 탐색을 통해 파라미터 조정
- **nrounds** : 최대 반복 횟수 (〉 0)
- **eta** : 학습률. eta 를 낮추면 모형의 과적합 가능성은 낮아지지만 학습 속도가 느려진다. (0 ~ 1)

- **gamma** : 트리에서 가지를 추가로 치기 위해 필요한 최소한의 손실 감소 기준.
- 기준값이 클 수록 모형이 더 단순해진다.(〉 0)
- **max_depth** : 트리의 최대 깊이.(〉 0)
- **min_child_weight** : 트리에서 가지를 추가로 치기 위해 필요한 최소한의 사례 수.(〉 0)
- **colsample_bytree** : 각각의 트리를 만들 때 데이터에서 사용할 열(column)의 비율(0 ~ 1)

최적 파라미터는 다음과 같습니다

nrounds = 128, max_depth = 8, eta = 0.1238111, gamma = 7.225651,

colsample_bytree = 0.3384503, min_child_weight = 20, subsample = 0.8708796.

(3) 변수 중요도 파악

1. 27 번 군집 변수 중 변수중요도 top20

```
> varImp(xgb.model)
xgbTree variable importance

  only 20 most important variables shown (out of 274)

                    Overall
`10naver_trend`     100.000
totalcost.length     59.373
`39nmbr_prchs`       43.361
total_price_cv       36.424
nmbr_prchs_cv        31.813
naver_trend_cv       23.691
clst_total_price_cv  22.724
`39total_price`      20.012
unit_price           15.988
`29total_price`      15.106
`10total_price`      14.487
totalcost.median     13.058
total_price          12.130
`10nmbr_prchs`       12.129
fre                  10.529
X20age                9.642
`56total_price`       9.430
rfm_score             9.266
`3naver_trend`        8.067
purchase_cycle        7.882
```

27 번 군집의 제품 구매 여부 예측에 대해 제일 중요한 변수는 '네이버트렌드 10 번군집'이었다. 이는 27 번 군집과 가장 가까운 군집인 10 번 군집의 특성이 가장 큰 영향을 끼쳤다고 볼 수 있다. 또한, '10 번군집의 총 구매량', '10 번군집의 구매 건수'가 top20 안에 자리잡고 있음을 알 수 있다. 이 뿐만 아니라 39 번군집, 56 번 군집, 29 번 군집에 관련된 변수또한 27 번 군집과 가까운 군집들이었고 이는 가까운 군집이 해당 군집에 큰 영향을 끼치는 것을 알려주고 있다.

이밖에도 변동계수와 관련된 변수, rfm_score 등이 눈에 띈다.

2. 14 번 군집 변수 중 변수중요도 top20

```
> varImp(xgb.model)
xgbTree variable importance

  only 20 most important variables shown (out of 274)

                    Overall
unit_price          100.000
total_price_cv       73.612
totalcost.length     57.749
srch_amnt_cv         56.581
total_price          42.841
nmbr_prchs_cv        39.125
totalcost.median     36.160
clst_total_price_cv  30.184
naver_trend_cv       27.987
`41total_price`      18.167
`3total_price`       14.944
`54total_price`       9.705
`9total_price`        9.183
`58total_price`       8.272
totalcost.mean        7.669
fre                   7.415
X20age                6.096
`29total_price`       5.628
`3naver_trend`        5.571
`4total_price`        5.357
```

14 번 군집의 경우 '구매 단가' 변수가 제일 큰 영향을 끼쳤고 그 뒤를 이어서 변동계수 변수들이 예측력을 높여주었다. 27 번군집과 동일하게 주목할 점은 41, 3, 54,9,58 군집 모두 14 번 군집과 가장 가까운 군집이라는 것이다. 이 과정에서도 알 수 있듯이, 해당 군집의 구매 여부를 예측 할 때 거리가 가까운 군집일수록 영향력이 높다는 것을 알 수 있었고 이는 군집화가 제대로 이루어 졌다고 볼 수 있다.

(4) 최종 결과

1. 27 번 군집 결과 : **정확도 89.3%**

```
               Accuracy : 0.893
                 95% CI : (0.8786, 0.9062)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.786
 Mcnemar's Test P-Value : 0.3052

            Sensitivity : 0.8850
            Specificity : 0.9010
         Pos Pred Value : 0.8994
         Neg Pred Value : 0.8868
             Prevalence : 0.5000
         Detection Rate : 0.4425
   Detection Prevalence : 0.4920
      Balanced Accuracy : 0.8930

       'Positive' Class : 0
```

2. 14 번 군집 결과 : **정확도 89.9%**

```
               Accuracy : 0.8995
                 95% CI : (0.8855, 0.9123)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.799
cnemar's Test P-Value : 0.004782

            Sensitivity : 0.8790
            Specificity : 0.9200
         Pos Pred Value : 0.9166
         Neg Pred Value : 0.8838
             Prevalence : 0.5000
         Detection Rate : 0.4395
   Detection Prevalence : 0.4795
      Balanced Accuracy : 0.8995

       'Positive' Class : 0
```

3. 온라인 군집 선호 지수 비교 결과

왼쪽은 '온라인 군집 선호 지수' 추가 전, 오른쪽은 '온라인 군집 선호 지수' 추가 후 결과이다. 비교를 위해 하이퍼 파라미터와 모든 변수를 통일 시켰고, prefer_index (온라인 군집 선호 지수)의 차이만 있다. 분석 결과 지수를 넣지 않았을 때보다 넣었을 때 1.2% 올라갔습니다. 이를 통해 온라인 군집 선호 지수가 유의미 하다는 결론을 내렸다.

```
> confusionMatrix(pred.xgb, test$n)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 854 159
         1 146 841

               Accuracy : 0.8475
                 95% CI : (0.831, 0.863)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.695
 Mcnemar's Test P-Value : 0.492

            Sensitivity : 0.8540
            Specificity : 0.8410
         Pos Pred Value : 0.8430
         Neg Pred Value : 0.8521
             Prevalence : 0.5000
         Detection Rate : 0.4270
   Detection Prevalence : 0.5065
      Balanced Accuracy : 0.8475

       'Positive' Class : 0
```

```
> confusionMatrix(pred.xgb, test$n)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 860 142
         1 140 858

               Accuracy : 0.859
                 95% CI : (0.843, 0.874)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.718
 Mcnemar's Test P-Value : 0.9525

            Sensitivity : 0.8600
            Specificity : 0.8580
         Pos Pred Value : 0.8583
         Neg Pred Value : 0.8597
             Prevalence : 0.5000
         Detection Rate : 0.4300
   Detection Prevalence : 0.5010
      Balanced Accuracy : 0.8590

       'Positive' Class : 0
```

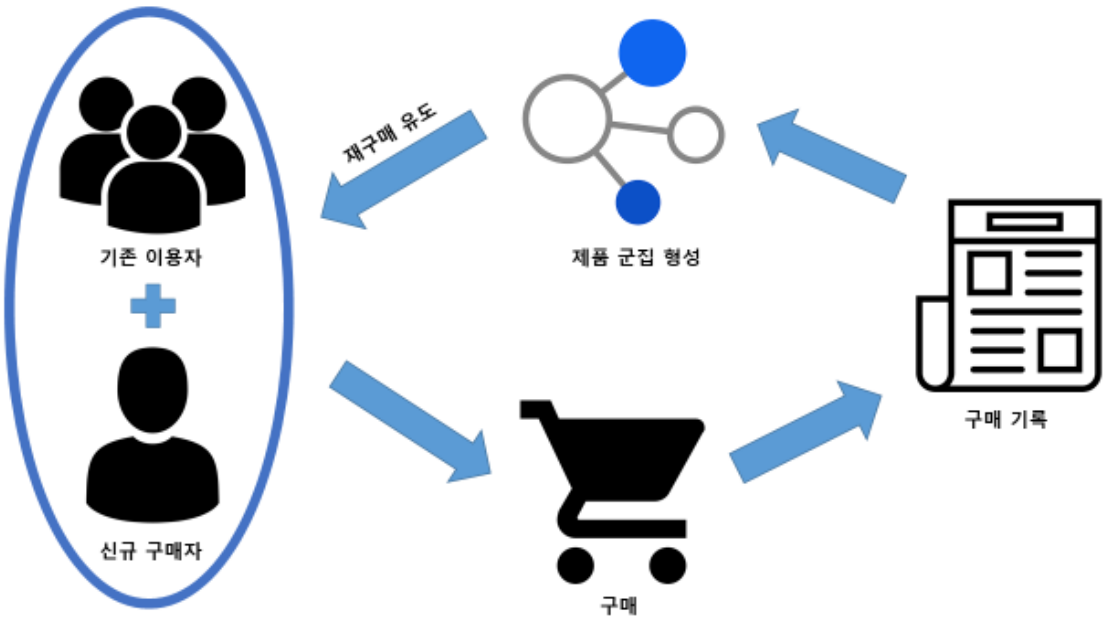
실제로 변수 중요도 확인 결과 prefer_index 인 온라인 군집 선호 지수가 상위권에 있는 것을 확인 할 수 있었다.

```
> varImp(xgb.model)
xgbTree variable importance

  only 20 most important variables shown (out of 253)

                  Overall
`10naver_trend`   100.000
totalcost.length   78.452
`10total_price`    74.302
`39naver_trend`    45.212
`39nmbr_prchs`     39.458
`10srch_amnt`      36.410
`39total_price`    33.956
`29total_price`    25.377
prefer_index       21.424
unit_price         21.060
totalcost.median   19.554
total_price        16.653
fre                16.226
totalcost.mean     15.962
X20age             15.202
`41naver_trend`     9.693
rec                 8.594
`29naver_trend`     7.260
`56nmbr_prchs`      6.940
`3total_price`      6.647
```

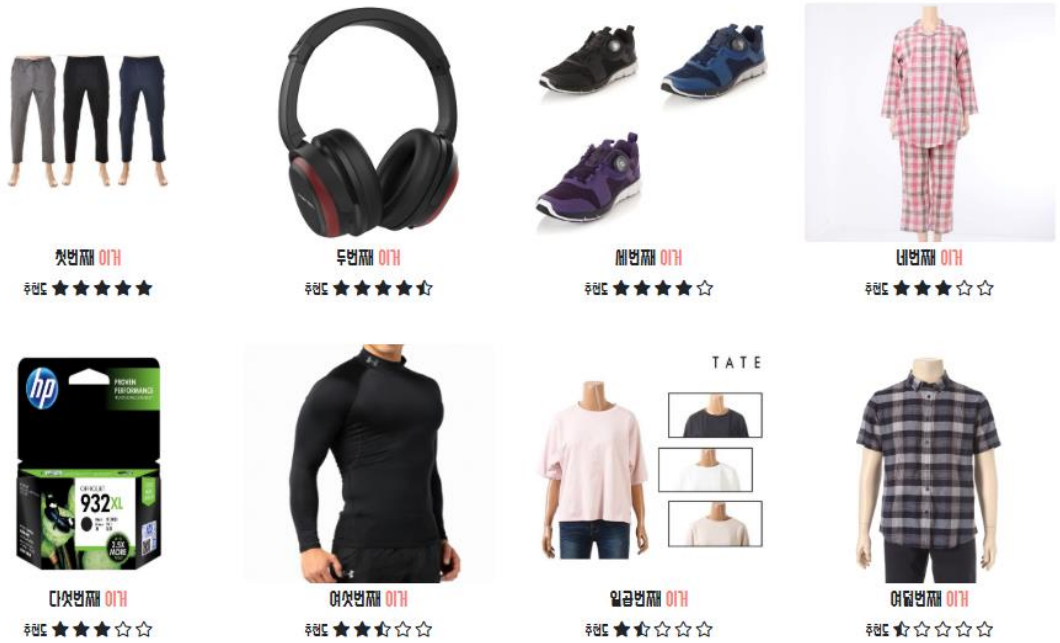
7. 최종 서비스 제안

‘이거’는 어때? 또 ‘이거’는 어때?

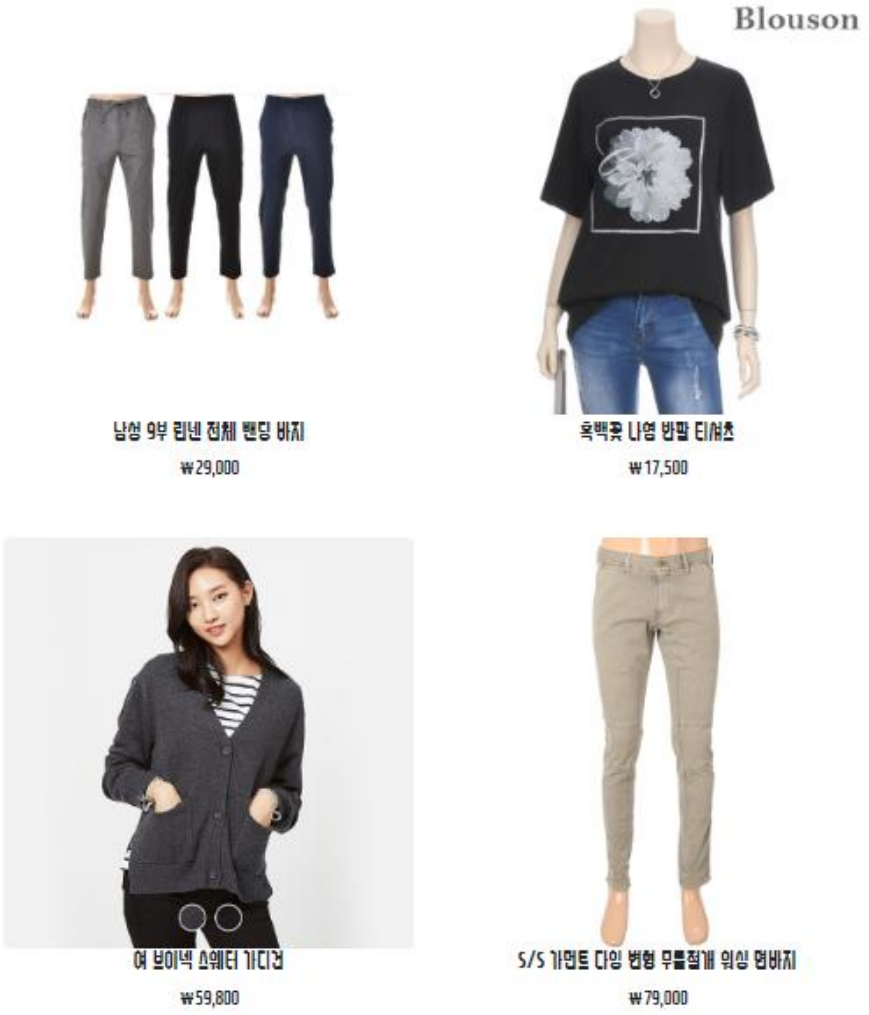
‘이건 어때’는 앞선 단계로부터 형성한 군집과 그 거리에 따라 이용자가 필요로 할만한 상품을 추천해주는 것을 목표로 한다. 신규 구매자와 기존 이용자가 구매를 함으로써 쌓은 구매 기록을 바탕으로 제품 군집을 형성한 후, 이를 이용해 다시 재구매를 유도하는 선순환이 목표가 된다. 실질적으로 고객이 구매한 제품이 하나 뿐이라, 고객 성향을 알기 힘들더라도 제품을 구매한 전체 고객들의 구매 데이터로부터 그 전반적인 패턴을 파악하여, 다른 근접한 군집에서 역시 또다른 구매를 유도하는 것이다.



아래는 데이터 분석 과정을 거친 이후 '여긴 어때?'의 프로토타입을 웹을 통해 구현한 형태의 일부이다. 이용자의 구매 내역에 따라 8 개의 '이거'들을 제시한다. 이 때의 '이거'는 고객에게 추천하고자 하는 한 군집의 일부를 의미한다.



위는 고객의 입장에서 14 번 군집의 상품을 구매한 적이 있다고 가정한 경우다. 그 중 거리 상 가장 가까운 군집에 속하는 상품들의 일부 묶음은 높은 추천도로 '첫번째 이거'라는 이름으로 제공된다.



하나의 '이거'를 클릭하면, 그 해당 군집에 속하는 무수한 상품들 중 4 개의 품목을 제시한다. 군집 내의 상품 선정은 무작위로 제시된다. 실제로 서비스를 운영한다고 가정했을 때에는, 추천 항목들에 대해 단기적으로 세일을 적용하거나 하는 방법도 고려해볼 수 있다.

물론 프로토타입의 형태로 대략적으로 떠올린 현재로선 한계점도 제법 뚜렷하다. 애초에 해당 서비스를 제공하기 위해서는 고객이 '한번이라도 구매를 해야'하는데, 만에 하나 그렇지 않다면 해당고객들에게는 서비스를 제공할 수 조차 없다는 점이 바로 그것이다. 다르게 말해, 해당 서비스로는 신규고객의 유입을 유도하기가 상당히 어렵다.

때문에 원활한 서비스 제공을 위한 데이터 확보 등을 위해서 기존 고객을 계속 유지해 나가면서도 신규고객의 새로운 유입 역시 요구되는 형태다. 때문에 단기적으로 많은 고객들의 유입을 기대해볼 수 있는 이벤트나, 아예 새로운 형태의 다른 서비스 제공이 필요하겠다.

(데모 웹 Github : https://github.com/KimMuAng/KimMuAng.github.io)

(시범 웹페이지(서버 문제로 불안정) : https://evening-spire-65794.herokuapp.com/)

8. 결론

고객의 특성을 파악하고, 행동을 예측하는 것은 마케팅에 있어서 매우 중요한 요소이다. 오프라인 뿐만 아니라 온라인에서도 고객들의 구매 데이터를 통해 군집분석과 함께 고객 특성을 파악 할 수 있다. 유아용품 관련 제품을 사는 고객, 전자기기를 사는 고객 등 해당 고객의 선호 뿐만 아니라, 쇼핑몰에 의존하는 정도, 총 구매금액, 구매 주기, 체류시간 등 고객 자체의 특성이 해당 고객을 파악하는데 중요한 요소로 작용한다.

함께 판매되는 제품들을 word2vec, K-means 알고리즘을 통해 새로 군집분석을 시행 하고 해당 군집의 제품 구매 여부를 예측하는 것 또한 고객의 특성을 파악하는 중요 요소이다. 제품의 군집화가 고객 군집이 될 수 있고, 군집의 특성은 고객의 특성으로 반영 될 수 있다.

일부 구매 데이터에서 나아가 쇼핑의 기본인 검색어 트렌드를 이에 접목시킨다면 더 광범위한 데이터 속에서 가치를 찾을 수 있다. 해당 제품을 찾는 사람이 많고 그 제품을 찾는과 동시에 다른 제품이 함께 찾아 진다면 이는 하나의 군집으로 볼 수 있다. 군집 분석된 제품군과 검색어 트렌드를 섞는다면 더욱 의미 있는 군집 정보가 되고 이는 고객의 특성으로 이어 질 수 있게 된다.

고객 자체의 특성, 군집별 고객의 특성, 검색어 트렌드와 군집의 관계. 이 세가지를 융합하여 하나의 새로운 '온라인 군집 선호 지수'를 개발 하는 것이 본 프로젝트의 목적이다. 이는 고객이 해당 군집의 제품 구매 여부를 예측하는데 설명력을 줄 수 있으며 89%의 높은 정확도를 보이고 있다.

더 나아가 군집의 거리를 이용하여 해당 군집의 제품을 구매한 고객에게 이미 구매한 제품들을 광고나 추천으로 보여주는 것이 아닌, 군집과 가장 가까운 군집의 제품들을 소개함으로써 선택의 폭을 늘려주고 고객이 관심 있어하는 제품 군집을 보여준다면 이는 고객의 특성을 정확히 파악하고 적용한 것이라 할 수 있다. 우리가 개발한 '여기 어때' 서비스는 이를 적극 활용한 플랫폼이라 할 수 있다.