

# 이커머스 고객 세분화 분석

군집 분석과 RFMV를 이용한 초세분화 마케팅 전략



# 목 차

## 01. 분석 배경과 데이터 설명

## 02. EDA 및 파생변수 생성

## 03. 군집 분석

- 정규화 & 주성분 분석(PCA)
- 군집 비교
- 고객 비교
- 군집별 마케팅 전략

## 04. RFMV 분석

## 05. 최종 마케팅 전략



# 프로젝트 일정

	2024년 12월 9일 ~2025년 01월 17일																
프로젝트 기획	12/09 ~ 10																
데이터 선정		12/10 ~ 12															
데이터 탐색				12/11 ~ 13													
군집 분석					12/12 ~ 17												
추가 세분화 분석								12/15 ~ 20									
마케팅 전략 수립											12/19 ~ 24						
ppt 작성															12/23 ~ 1/6		

# 01. 분석 배경



## 고객 가치와 행동 패턴의 세분화에 대응 필요성 대두

HOME > 칼럼 > 권순철

### 권순철의유통칼럼(73) 소비자구매 패턴변화의 핵심트렌드

권순철 칼럼니스트 | 승인 2024.03.27 10:28 | 댓글 0

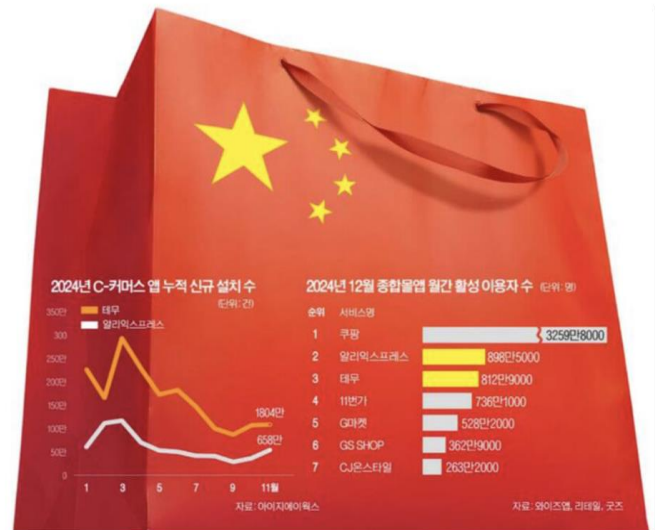
최근 소비자 구매 패턴은 과거와는 비교할 수 없을 만큼 빠르게 변화하고 있다. 인터넷과 모바일 기술의 발달, 사회 경제적 변화, 그리고 가치관 변화 등이 복합적으로 작용하며 소비 트렌드를 새롭게 만들어가고 있다. 이러한 변화에 발맞춰 기업들은 마케팅 전략을 수정하고 새로운 사업 모델을 개발해야 한다.



## 중국발 C커머스의 등장으로 인한 이커머스 시장 경쟁 심화로 국내 이커머스 기업들의 가격경쟁력 약화

### [中 블랙홀에 빠진 韓 산업] 알리·테무·미니소...몰아치는 C커머스 공습

입력 2025-01-08 15:58:44 수정 2025-01-08 16:07:55



가성비를 앞세운 중국 유통기업의 공습이 본격화되고 있다. 지난해 12월 알리익스프레스와 테무의 월간 활성 이용자 수(MAU)는 2, 3위에 안착했다. 국내 기업들의 브랜드 차별화가 시급한 시점이다. 그래픽=권소화 기자

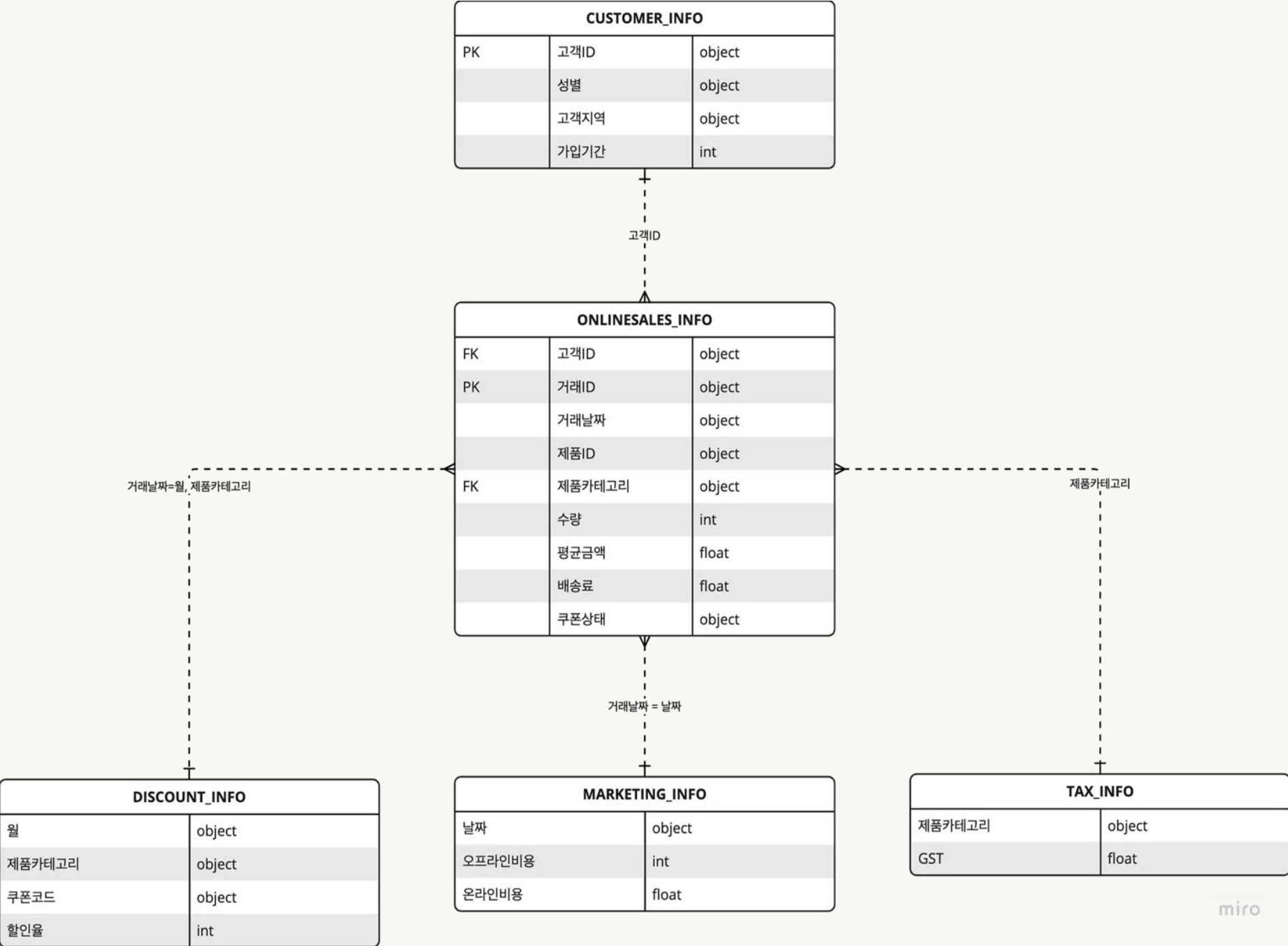
→ 궁극적으로는 고객 충성도를 높이고, 지속 가능한 성장을 도모하는 것이 목표

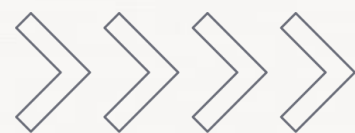
# 01. 데이터 설명



온라인거래와 관련된 정보, 고객 정보, 할인 쿠폰 관련 정보, 세금 관련 정보, 마케팅 비용 관련 정보 등 총 5개의 파일로 이루어진 데이터셋

출처 : 데이콘 이커머스 고객 세분화 분석 아이디어 경진대회



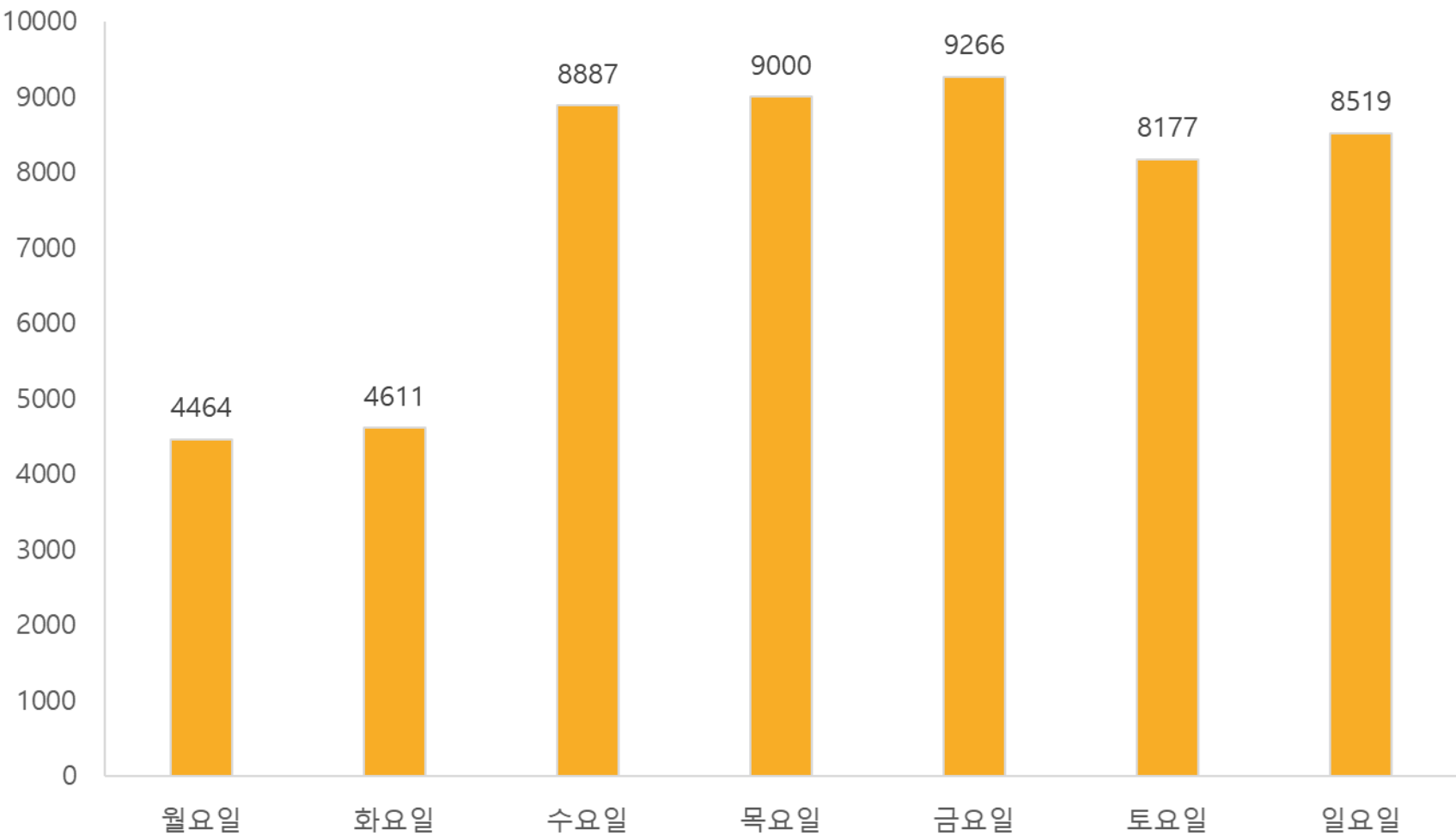


## 2. EDA 및 파생 변수 생성



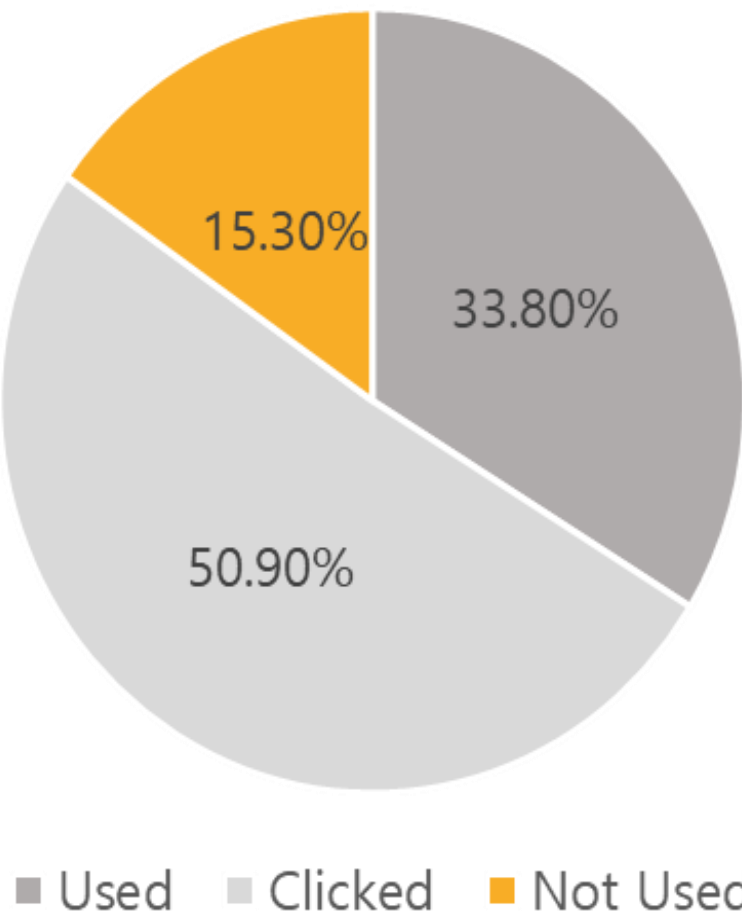
# 02. EDA

요일별 구매 횟수



다른 요일에 비해 월요일, 화요일 구매 횟수가 대략 50% 가량 적은 모습을 확인

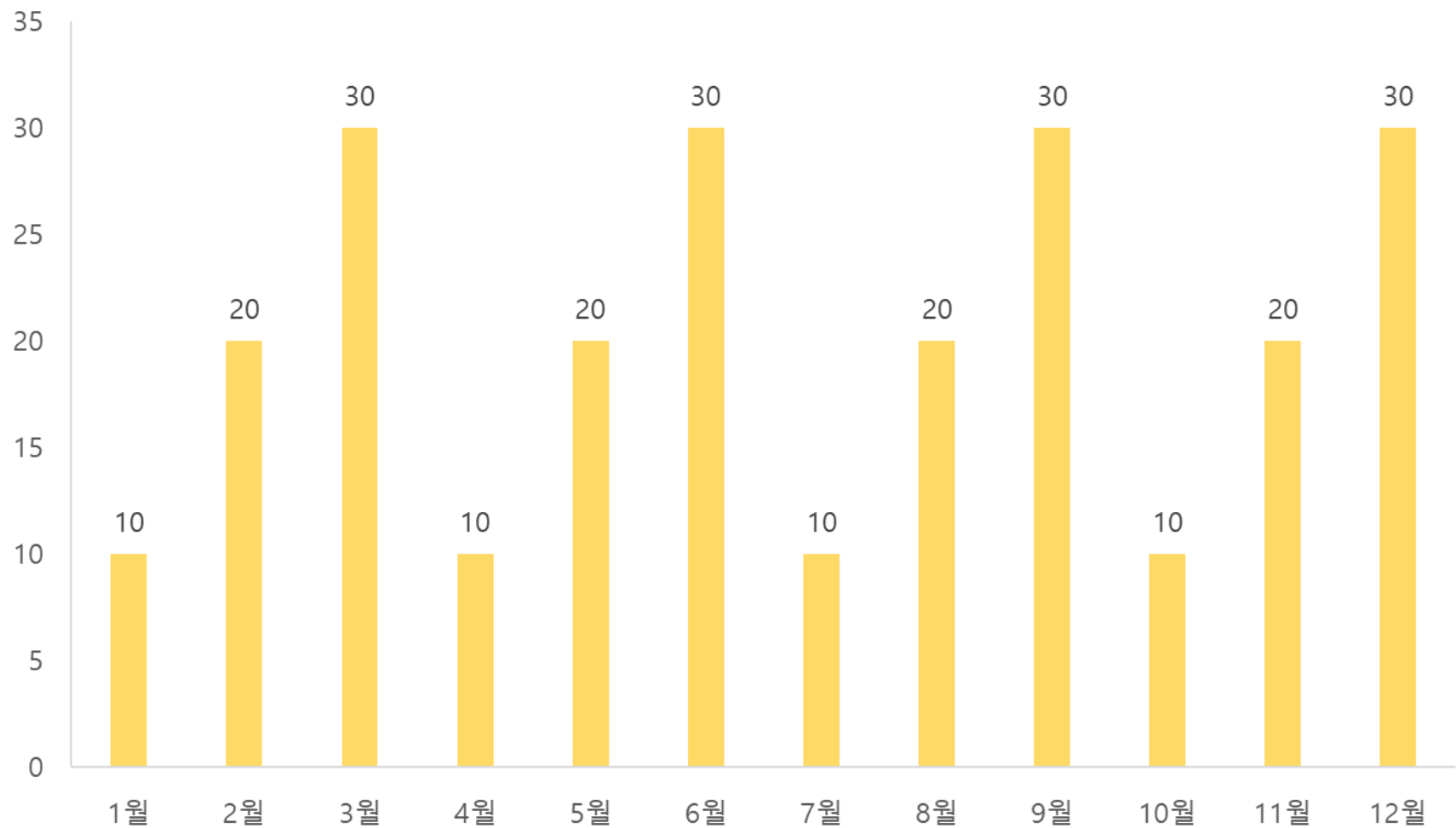
쿠폰 사용 비율



쿠폰 Clicked 과 used 가 많은 비중을 차지하고 있는 것을 확인할 수 있음. 상당수의 고객이 쿠폰을 적극적으로 활용하고 있음을 뜻함.

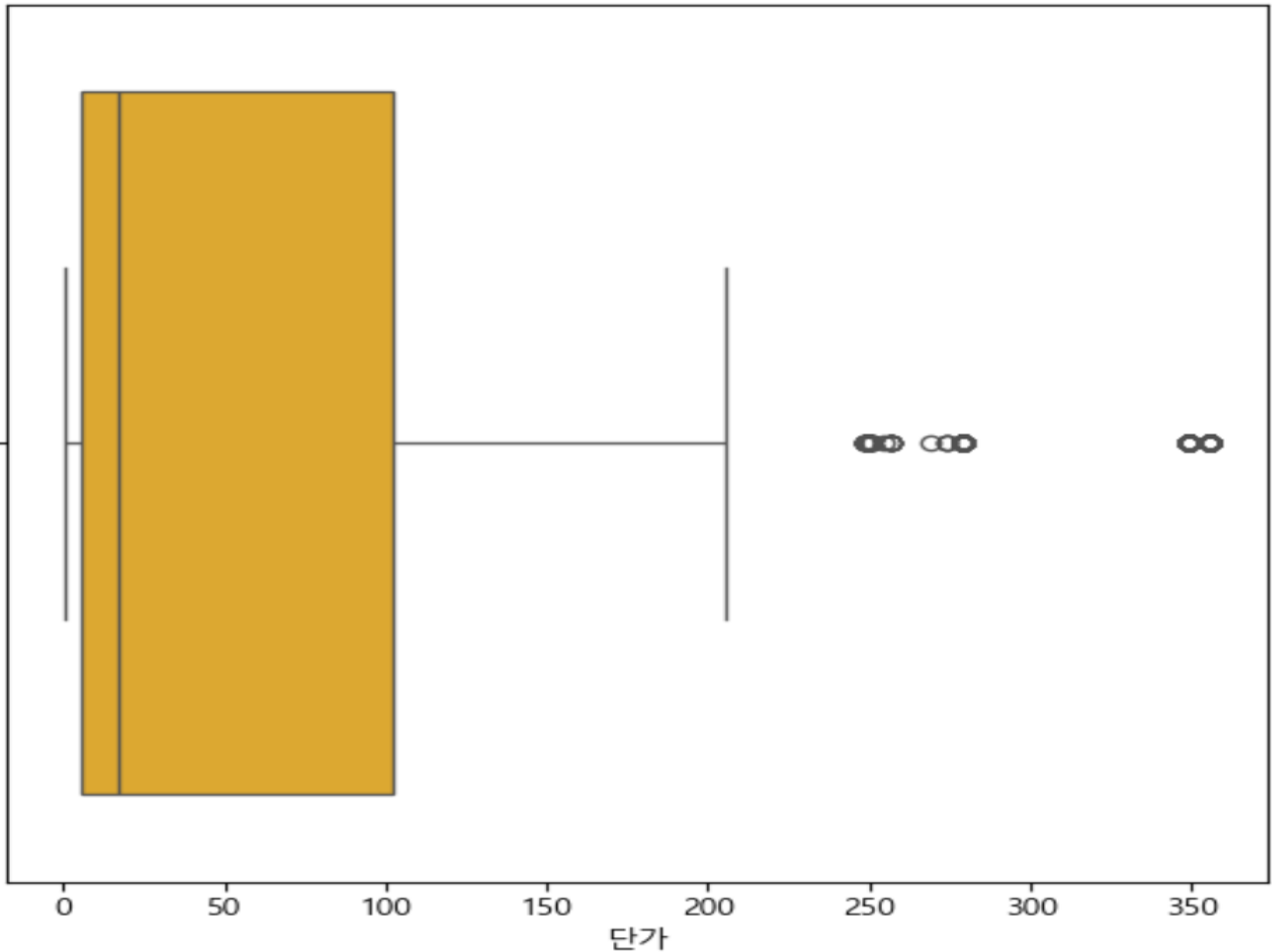
# 02. EDA

월별 평균 할인율



할인율이 3개월 주기로 반복되고 있음

구매 평균 단가

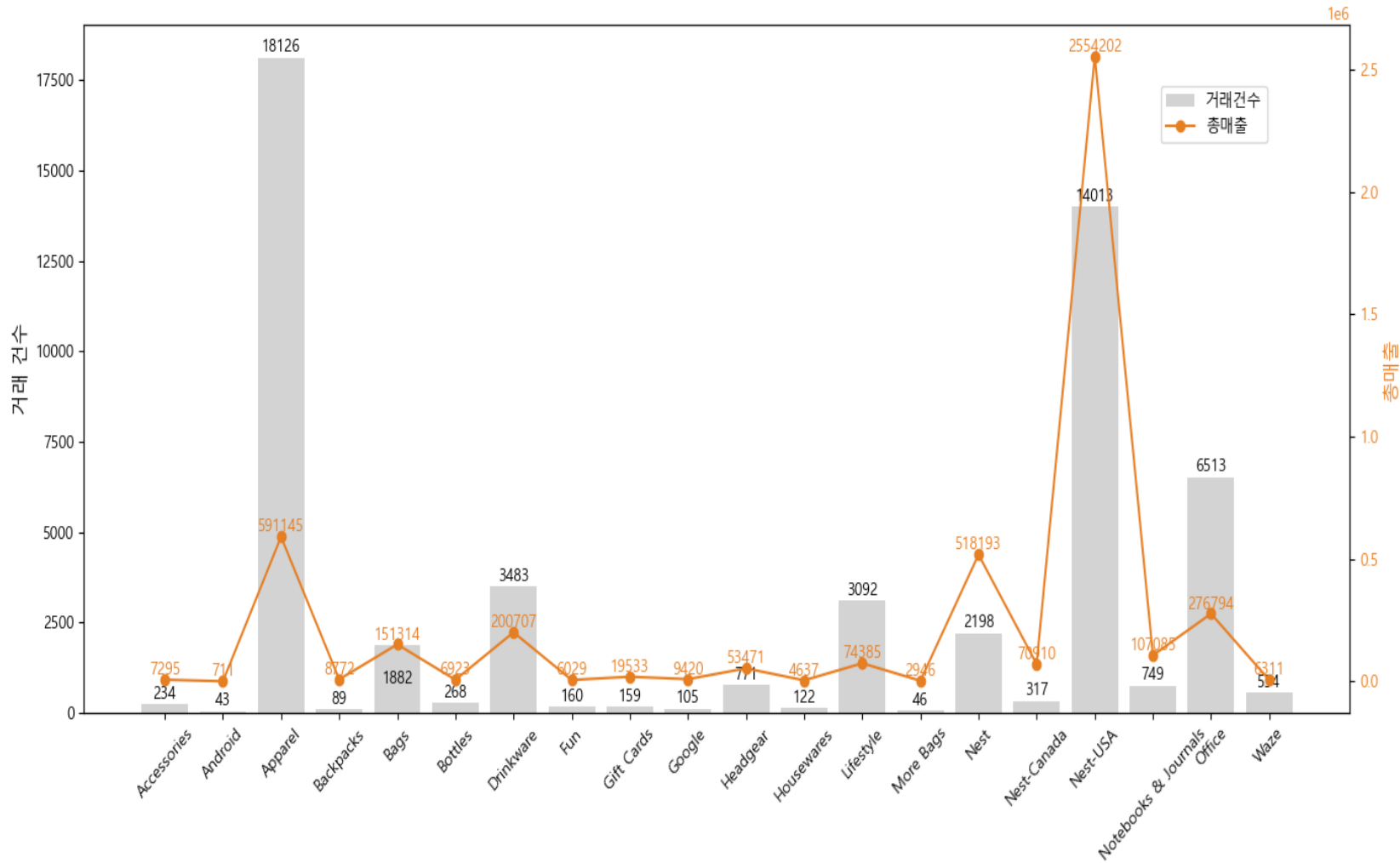


대부분의 구매 단가는 100달러 미만에 분포하고 있으며, 200 이상의 고단가 제품은 일부 고객만 구매하는 것으로 나타남



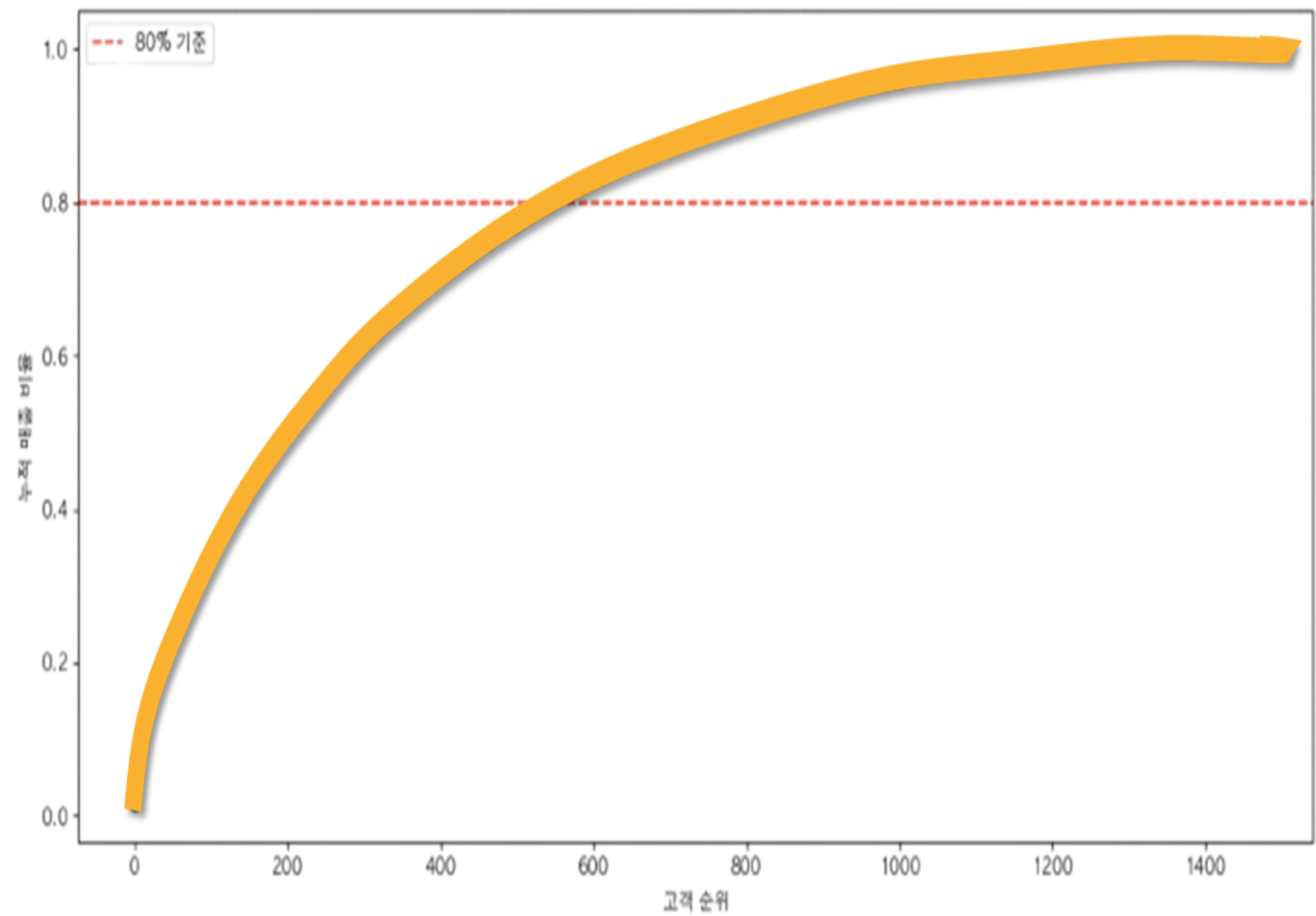
# 02. EDA

카테고리 별 거래 건수&총매출(막대)



Apparel, Office, Drinkware, Lifestyle은 거래 건수에 비해 총매출은 낮은 모습을 보여주고 있음

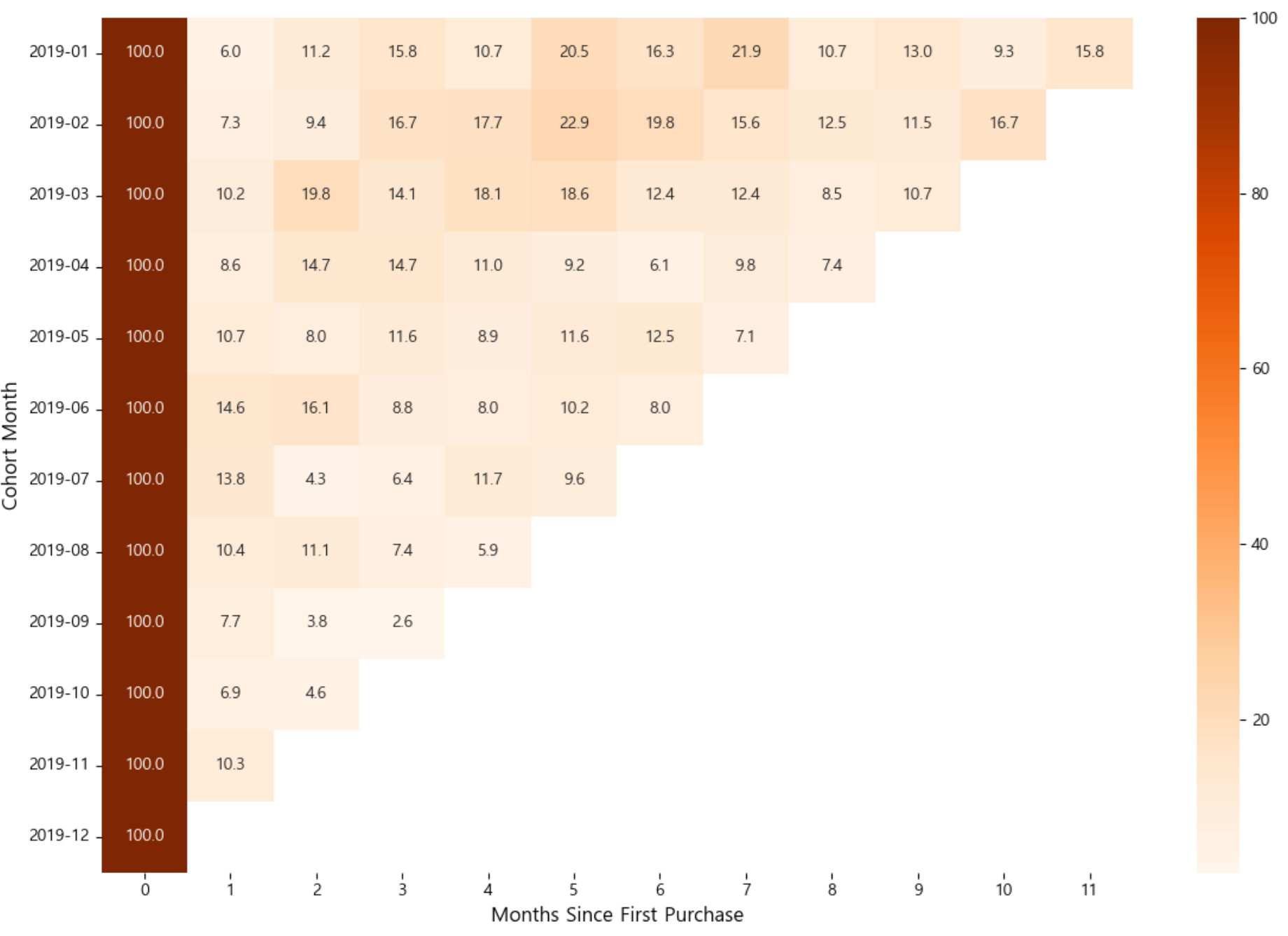
상위 고객의 매출 기여도(파레토 분석)



상위 37%의 고객이 총 매출 80%에 기여하고 있음

# 02. EDA

Customer Retention by Cohort



## 1. 첫 구매 후 빠르게 감소하는 리텐션

대부분의 코호트에서 첫 구매 후 1개월 차에 리텐션 비율이 급격히 감소하는 모습이 확인됨. 이는 고객이 첫 구매 이후 단기간에 재구매하지 않을 가능성이 높다는 것을 보여줌

## 2. 리텐션의 지속성

일부 코호트(예: 2019-01, 2019-03)는 초기 몇 개월 동안 비교적 높은 리텐션 비율을 유지하지만, 대부분의 코호트는 3~4개월 이후에는 리텐션 비율이 10% 이하로 떨어짐

## 3. 코호트별 차이

초기 코호트(2019년 상반기)는 비교적 높은 리텐션 비율을 보이지만, 후반기(2019-07 이후)로 갈수록 리텐션 비율이 전반적으로 낮아지는 모습 확인. 이는 시간이 지남에 따라 고객 유지가 어려워졌음을 나타냄

## 4. 높은 리텐션 유지 기간

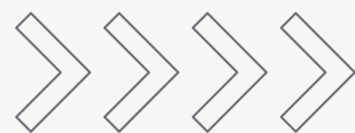
일부 코호트(2019-01, 2019-06)는 3개월 차 이후에도 일정 수준의 리텐션 비율을 유지하고 있음. 이는 해당 코호트에서 지속적인 구매 활동이 있었음을 나타냄

## 02. 파생변수 생성



생성 가능한 파생변수들의 모든 조합을 각 모델에 적용하여 실루엣 점수가 가장 높게 나온 변수의 조합을 채택  
최종 변수 조합 : ['평균할인율', 'Average\_Unit\_Price', 'Total\_수량', '쿠폰사용비율']

※ 2차 세분화 과정에서 RFM 지표를 사용할 예정이기 때문에 군집분석에 사용할 변수에서 제외 시킴.



# 3. 군집 분석

- 정규화 & 주성분 분석 (PCA)
- 군집 모델별 비교
- 고객 분류
- 군집별 마케팅 전략



## 03. 군집 분석 - 정규화 & 주성분 분석

### Min-Max 정규화

Min-Max 정규화란?

데이터를 특정 범위(0~1)로 변환하는 정규화 방법. 데이터의 최소값을 0 최대값을 1로 스케일링하여 모든 데이터가 동일한 범위 내에 위치하도록 조정함.

-> 서로 다른 스케일을 가진 데이터를 모델에 적용하기 쉽게 만듦.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

$X$  : 원본 데이터 값

$X_{max}$  : 데이터의 최대값

$X_{min}$  : 데이터의 최소값

$X_{norm}$  : 정규화된 값

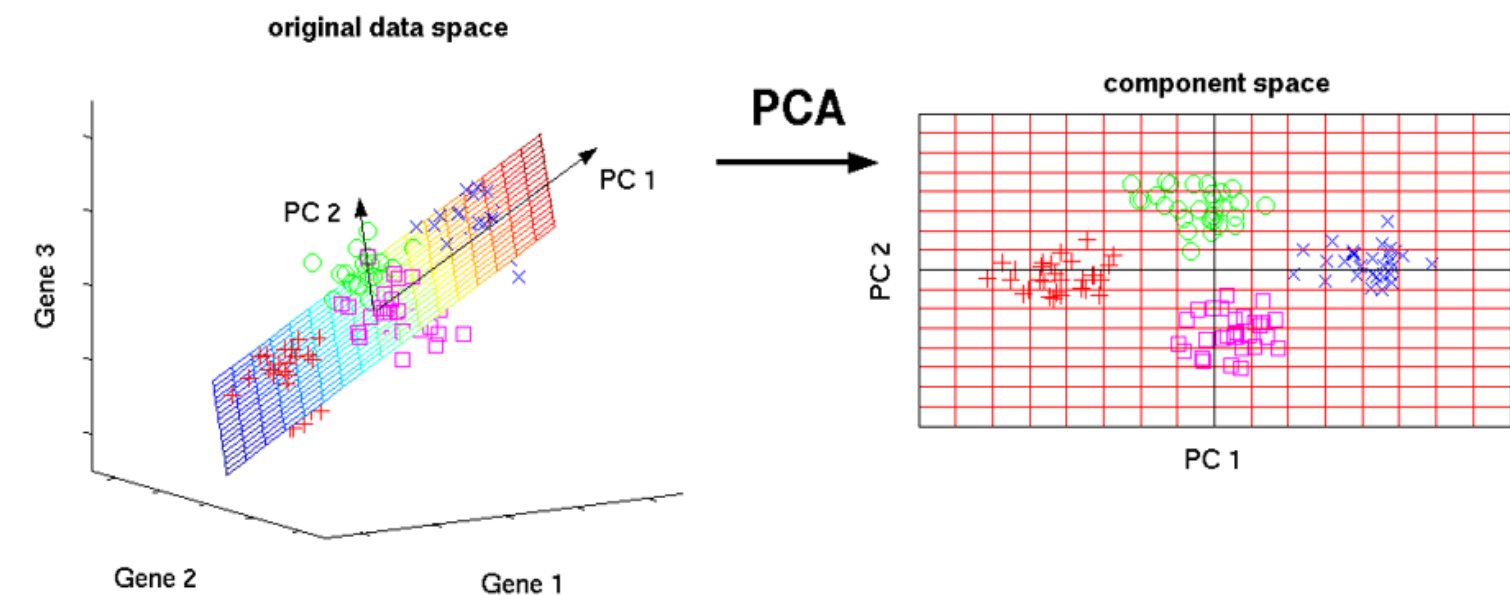
### 주성분 분석(PCA)

주성분 분석(PCA)이란?

데이터에 포함된 변수가 많아질수록 분석이나 모델링의 복잡도가 기하급수적으로 증가하기 때문에 차원의 저주에 빠지는 것을 방지하기 위한 기법

높은 차원에서는 데이터가 희소해지는 문제가 발생하고 모델 성능 저하로 이어질 수 있기 때문에 PCA를 통해 데이터의 특성을 최대한 유지하면서 저차원으로 축소하여 모델링의 복잡도를 완화하고자 함.

분산 설명력 : 차원 축소 후 특정 축이나 변수가 전체 데이터 분포를 얼마나 잘 표현하는지 평가.



# 03. 군집 분석(Birch) - 군집 비교

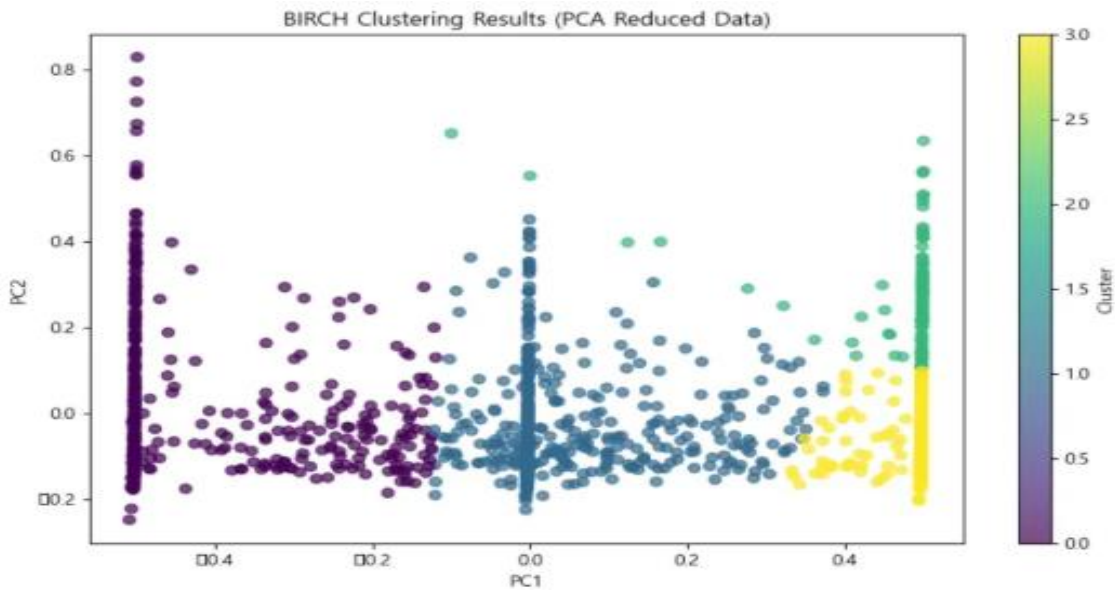
## 군집 분석(Birch)

Birch란?  
대용량 데이터셋에서 효율적으로 클러스터링을 수행하기 위해 설계된 계층적 클러스터링 알고리즘이다. 이는 메모리 및 계산 효율성을 고려해 설계되었으며, 트리 구조를 사용하여 클러스터링을 수행하기 때문에 데이터의 세부 정보를 모두 메모리에 저장하지 않고도 클러스터링을 효율적으로 진행할 수 있다..

선택 변수 ['평균할인율', 'Average\_Unit\_Price', 'Total\_수량', '쿠폰사용비율']

	실루엣 점수
Cluster 0	0.463
Cluster 1	0.517
Cluster 2	0.518
Cluster 3	0.687
전체 평균	0.533
분산 설명력	
95.2%	

클러스터링 시각화



# 03. 군집 분석(K-Medoids) - 군집 비교

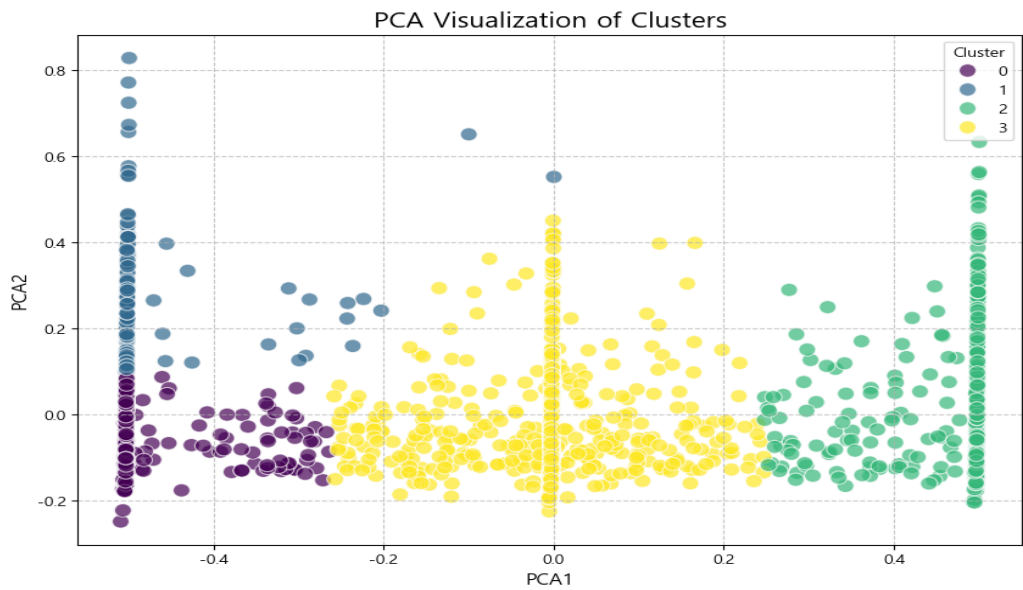
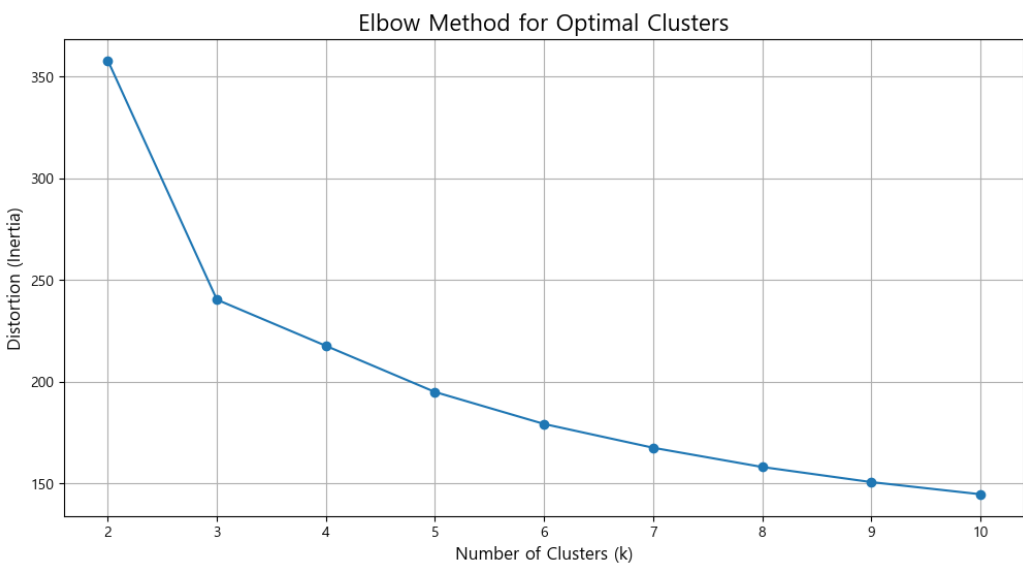
## 군집 분석(K-Medoids)

K-Medoids 란?  
데이터 군집화 알고리즘으로, 군집의 중심을 중심값(medoids)로 정의하며, 이 중심값은 군집 내 데이터 포인트 중 하나로 선택됨. 또한 이상치(outlier)에 더 강건하며, 다양한 거리 메트릭에 기반한 군집화를 지원.

선택 변수 ['평균할인율', 'Average\_Unit\_Price', 'Total\_수량', '쿠폰사용비율']

	실루엣 점수
Cluster 0	0.556
Cluster 1	0.411
Cluster 2	0.562
Cluster 3	0.475
전체 평균	0.514
분산 설명력	
95.2%	

## 클러스터링 & 엘보우 시각화



# 03. 군집 분석(K-Means) - 군집 비교

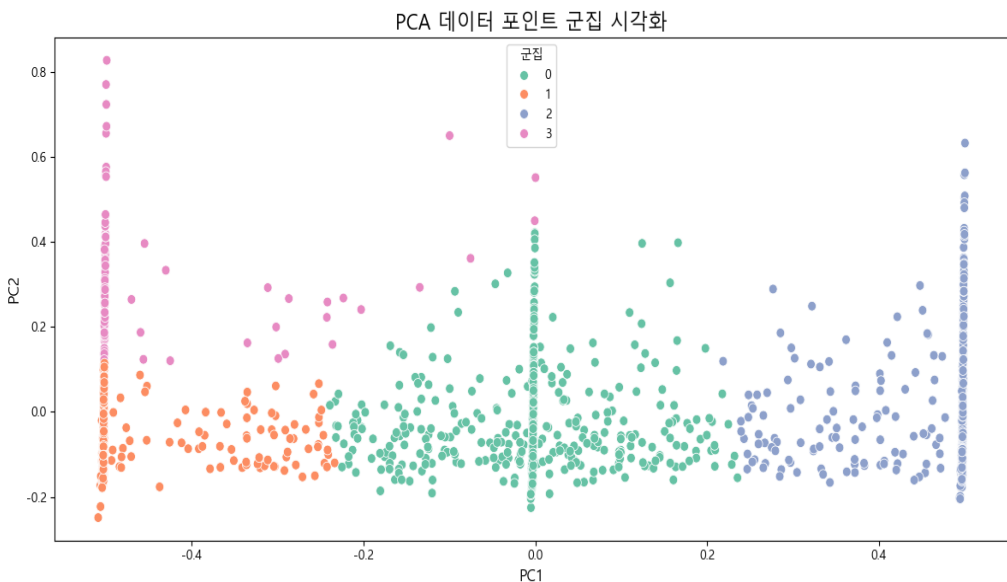
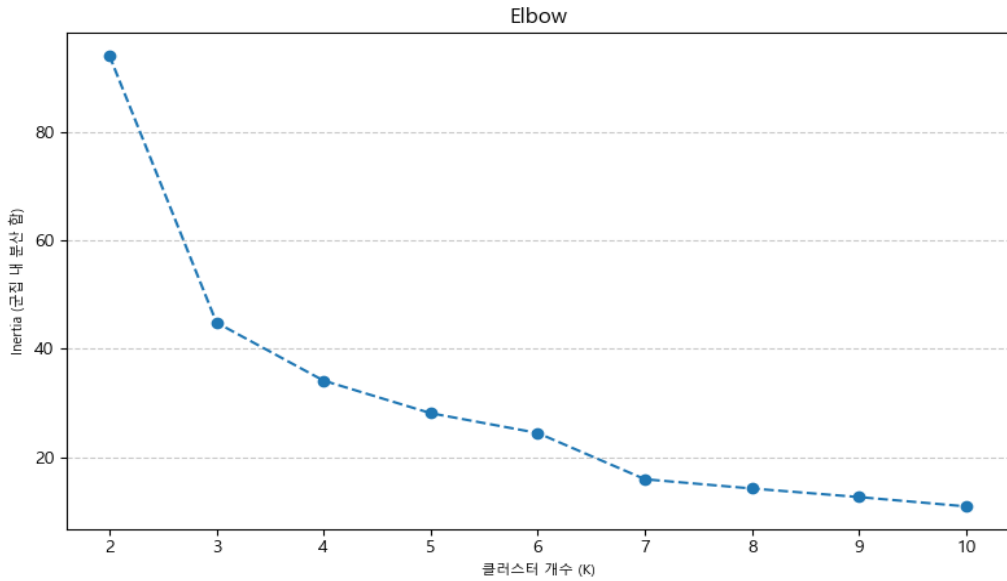
## 군집 분석(K-Means)

K-Means 란?  
데이터 집합을 K개의 서로 다른 클러스터로 나누는 군집화 알고리즘. K개의 클러스터 중심점을 설정하고, 각 데이터 포인트를 가장 가까운 중심점에 할당하여 그룹을 형성한다.  
클러스터는 데이터 간의 유사성을 기준으로 형성되며, 보통 유클리디안 거리를 사용해 데이터와 클러스터 중심점간의 거리를 계산.

선택 변수 ['평균할인율', 'Average\_Unit\_Price', 'Total\_수량', '쿠폰사용비율']

	실루엣 점수
Cluster 0	0.563
Cluster 1	0.646
Cluster 2	0.590
Cluster 3	0.421
전체 평균	0.581
분산 설명력	
95.2%	

## 클러스터링 & 엘보우 시각화



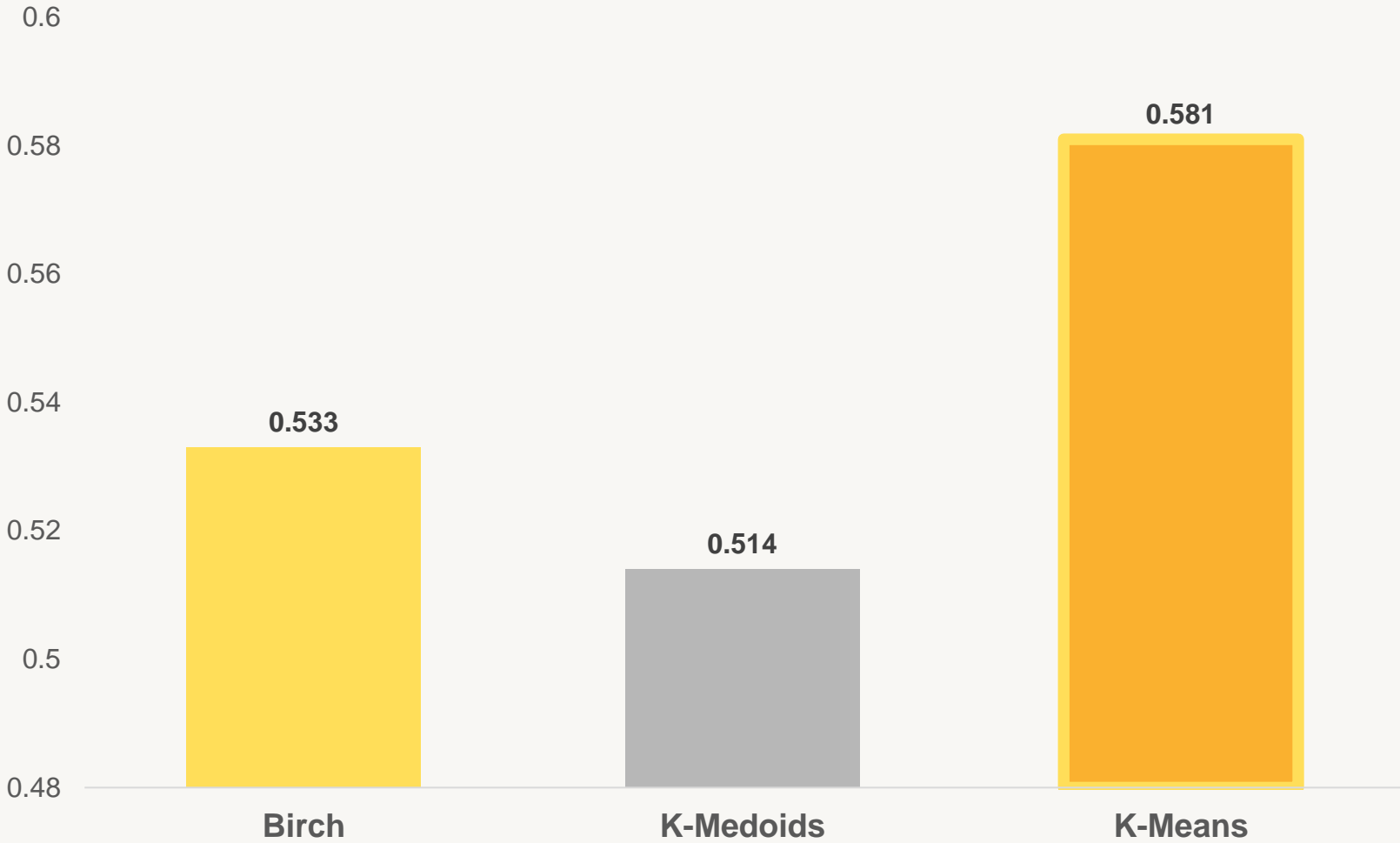


# 03. 군집 분석 - 군집 모델별 비교

군집 내 인원수와 엘보우 점수 및 분석 목표를 종합적으로 고려하여 K 값을 4로 설정

	Birch	K-Medoids	K-Means
Cluster 0	0.463	0.556	0.563
Cluster 1	0.517	0.411	0.646
Cluster 2	0.518	0.562	0.590
Cluster 3	0.687	0.475	0.421
전체 평균	0.533	0.514	0.581

모델별 전체 평균 실루엣 점수 비교



→ 전체 평균 실루엣 점수와 클러스터별 실루엣 점수를 고려하여 각 모델별 비교 결과, **K-Means**를 채택

# 03. 군집 분석 - 고객 분류

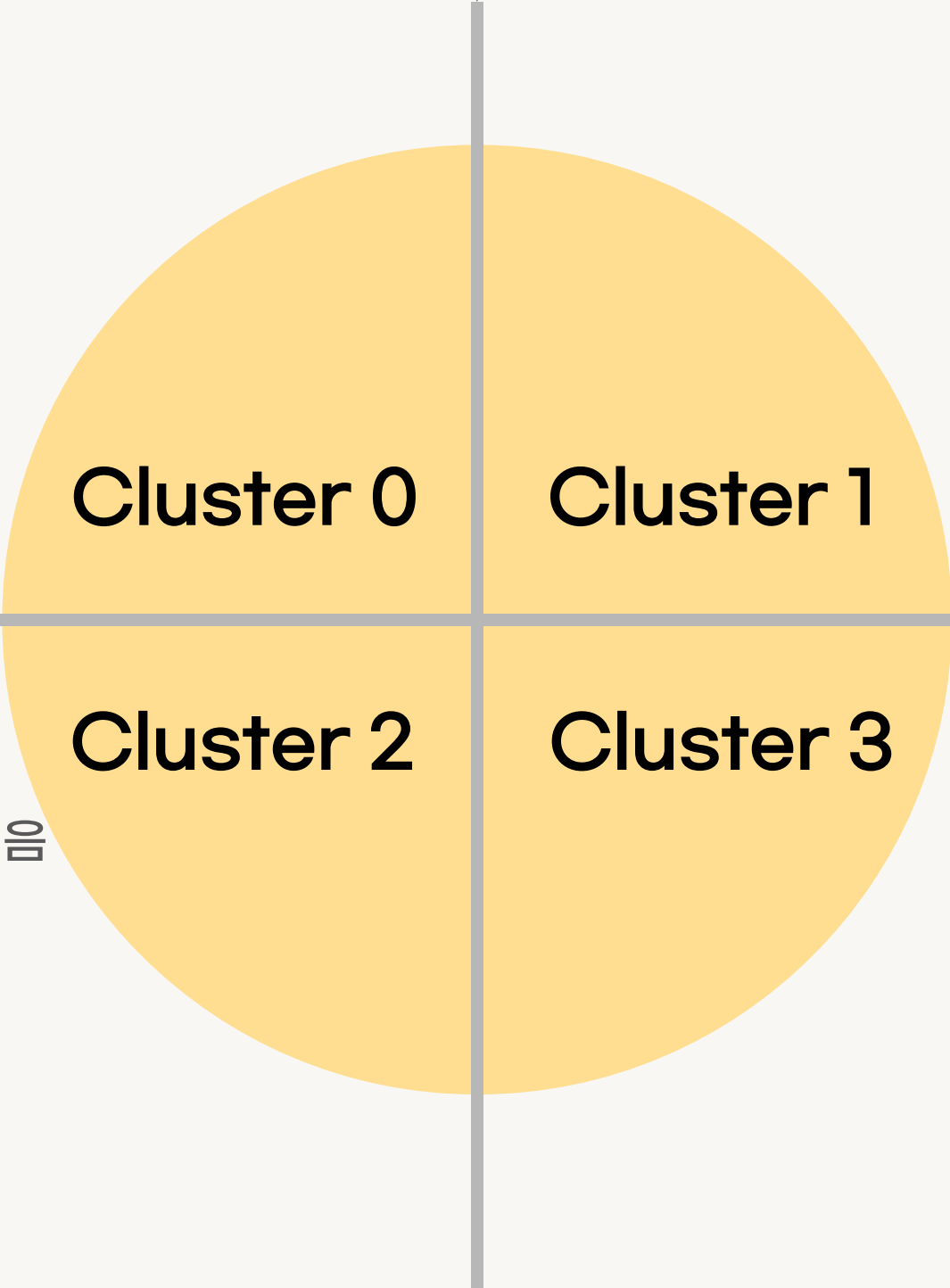


## 저가 + 대량 구매형 고객

- ❑ 고객 당 평균 매출이 4,448.70달러로 제일 높음
- ❑ 쿠폰 사용 비율의 평균이 0.7로 높은 편
- ❑ 쿠폰 평균 할인율은 19.93으로 높은 편에 속함
- ❑ 저가 제품 대량 구매

## 저가 + 쿠폰 민감형 고객

- ❑ 고객 당 평균 매출이 2,565.88달러로 2번째로 높음
- ❑ 쿠폰 사용 비율의 평균이 0.8로 가장 높음
- ❑ 반면 쿠폰 평균 할인율은 10.76로 가장 낮음



## 중저가 + 할인율 민감형 고객

- ❑ 고객 당 평균 매출이 2,396.36달러로 3번째로 높음
- ❑ 쿠폰 사용 비율의 평균이 0.69로 낮은 편
- ❑ 쿠폰 평균 할인율이 29.33으로 할인율 가장 민감

## 고가 + 쿠폰 둔감형 고객

- ❑ 고객 당 평균 매출이 1,891.55달러로 가장 낮음
- ❑ 쿠폰 사용 비율의 평균이 0.47로 가장 낮음
- ❑ 쿠폰 평균 할인율도 10.93으로 낮은 편
- ❑ 고가의 제품 소량 구매

# 03. 군집 분석 - 고객 군집별 마케팅 전략

WHO

HOW

Cluster 0 (저가 + 대량 구매형)



품목 고급화 유도, 추가 구매 유도, 대량 구매시 할인

Cluster 1 (저가 + 쿠폰 민감형)



구매가 몰려있는 Apparel, Nest-USA, Office 위주로 소액 쿠폰 제공, 쿠폰 제공 시에는 동일 카테고리 내 증가 이상의 가격대 상품 추천, 번들 세일

Cluster 2 (중저가 + 할인을 민감형)



구매가 많은 금, 토, 일요일에 타임 세일, 평균단가보다 높은 금액대 제품 한정 고할인율 쿠폰을 제공해 평균 단가를 높이는 전략

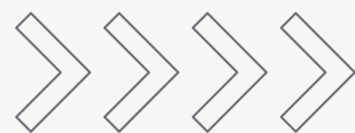
Cluster 3 (고가 + 쿠폰 둔감형)



고가 상품 수요에 부응하는 프리미엄관 운영

Insight

데이터셋 가공을 통해서 군집별 특성은 추렸지만, 보다 세분화된 분석 필요성 -> RFMV 추가 분석



## 4. RFMV 분석



## 04. RFMV 분석



### ✓ RFMV란?

**R**ecency : 언제 마지막으로 구입했는지

**F**requency : 구매 빈도는 어느 정도인지

**M**onetary : 구매 금액은 얼마인지

**V**ariety : 얼마나 다양한 카테고리를 구입했는지

### ✓ RFMV를 채택하게 된 이유

→ 고객의 플랫폼 의존도를 평가하기 위해 **Variety(카테고리 다양성)**을 평가 지표로 추가 채택. 기존의 RFM 분석 이외에도 V를 통해 고객이 플랫폼을 얼마나 폭넓게 활용하는지 확인 가능.

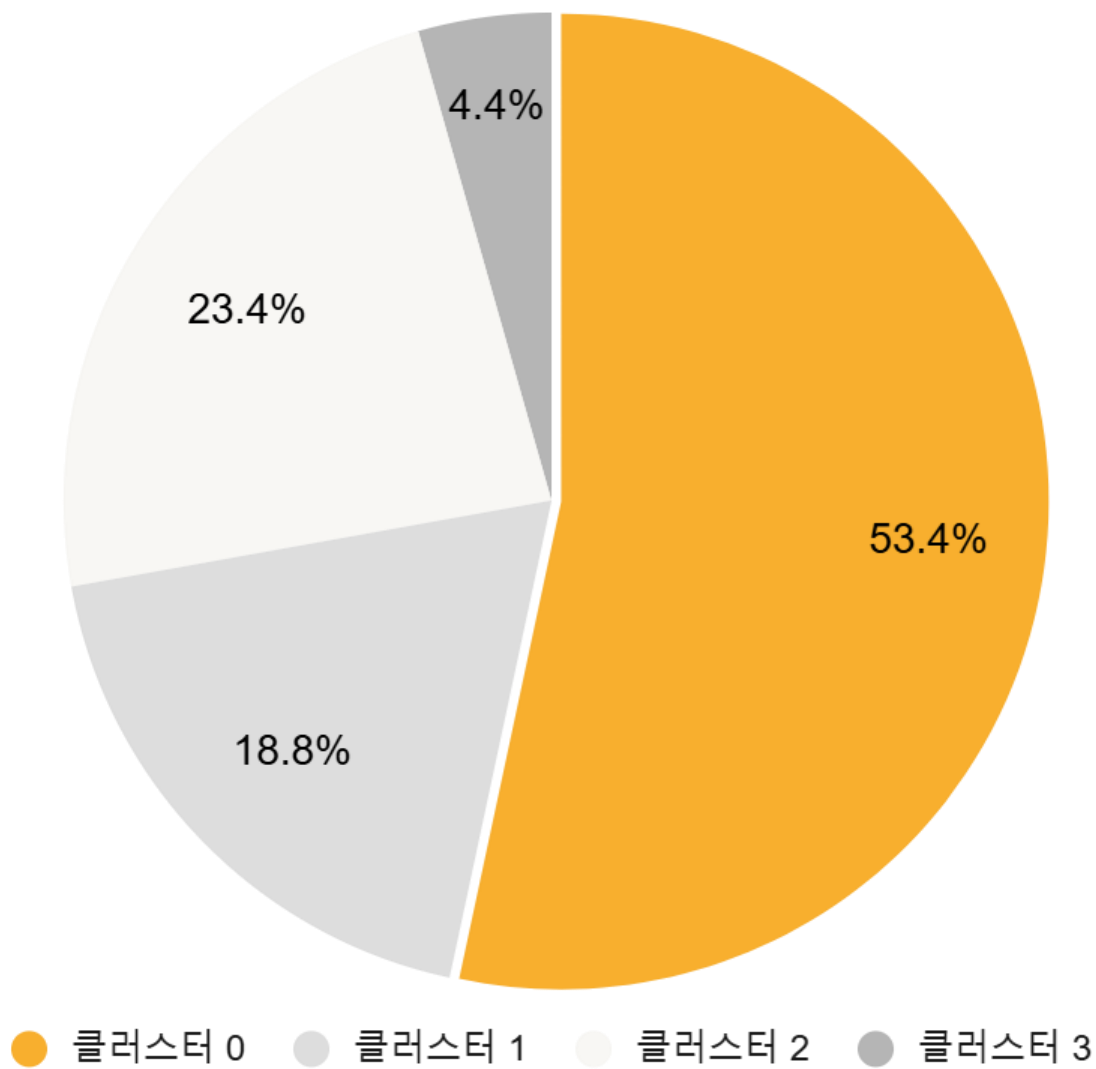
→ RFMV 스코어 별 점수 기준 : 각 컬럼별로 데이터값을 기준으로 5분위수를 적용. 각 분위수에 따라 1 ~ 5점의 점수를 부여함. 3점 이상이면 U(Up), 3점 미만이면 D(Down)로 분류해 UUUU부터 DDDD까지 총 16개 조합으로 고객을 2차 세분화.

예시: RFMV가 3512일 경우, UUDD로 세분화됨

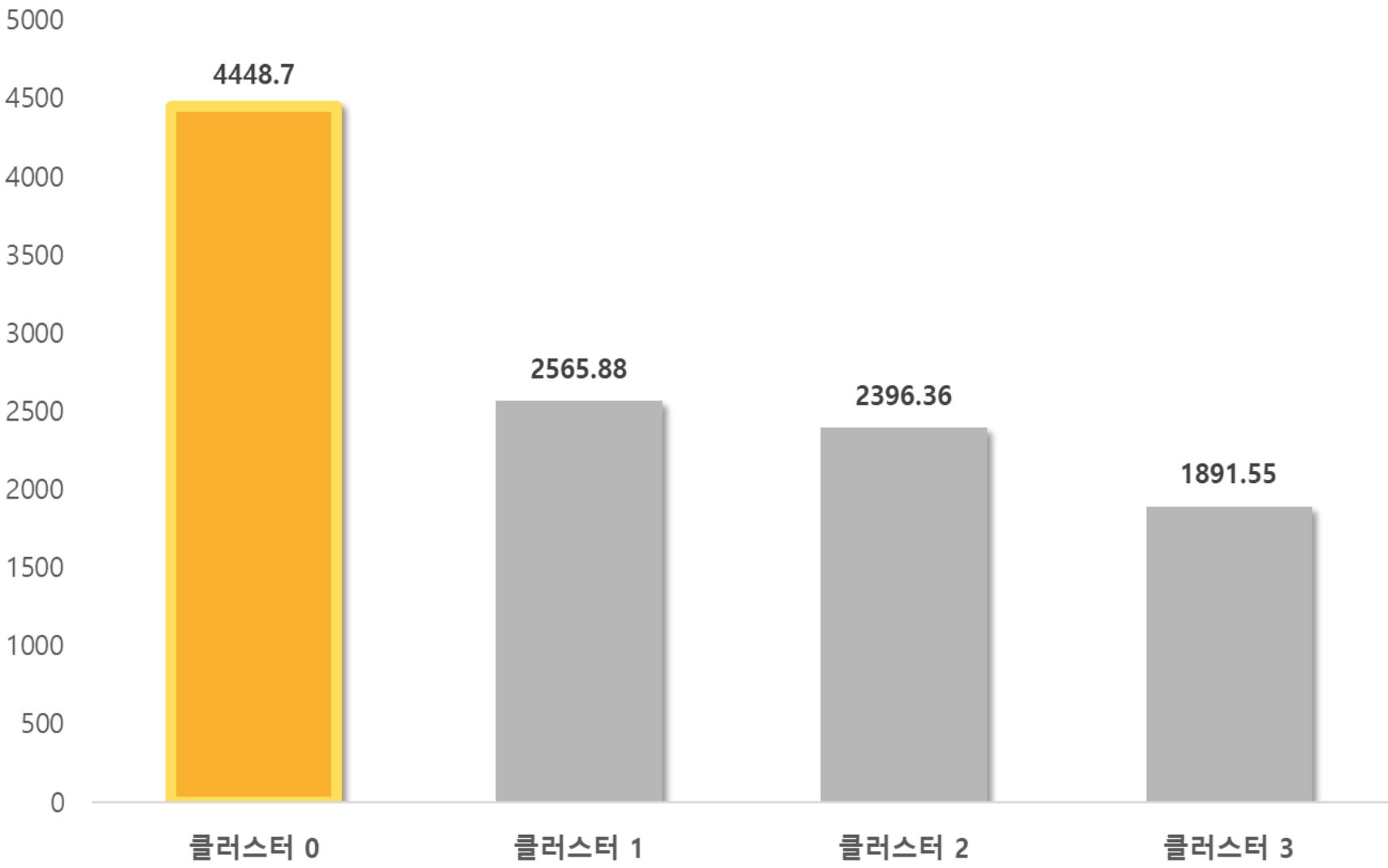
# 04. RFMV 분석

전체 매출의 53.3%를 차지하고 고객당 평균 매출이 타 클러스터에 비해 최대 135% 높은 클러스터 0에 대해 2차 RFMV 분석 진행

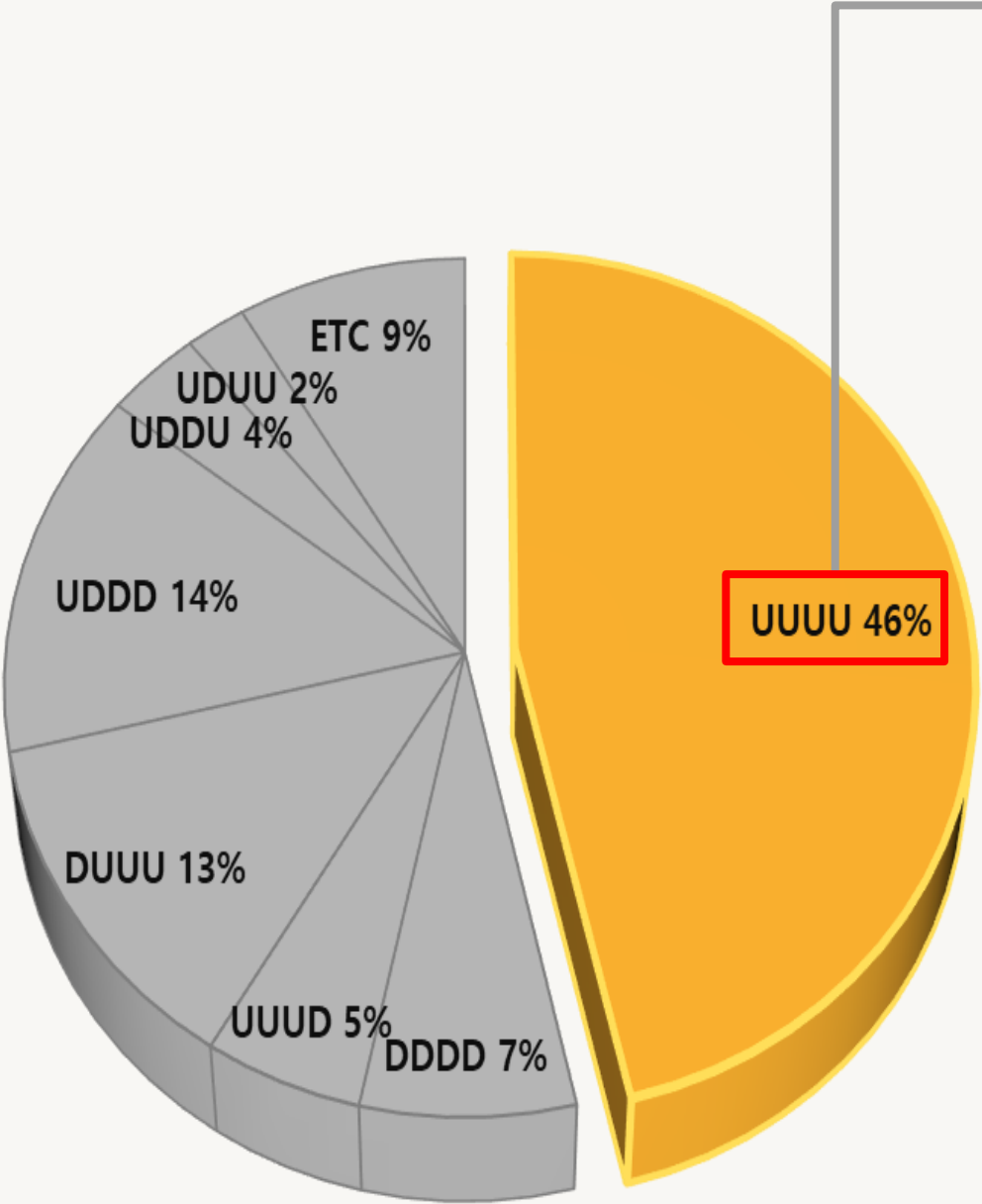
클러스터별 총 매출 비교



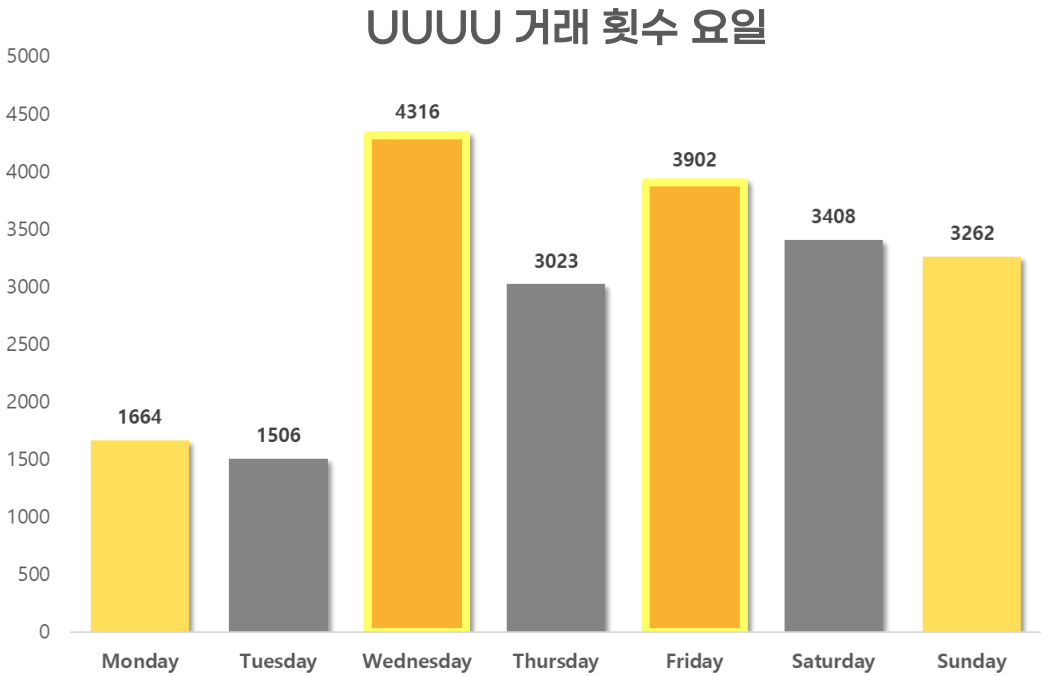
클러스터별 고객당 평균 매출



# 04. RFMV 분석 - 조합별 특성

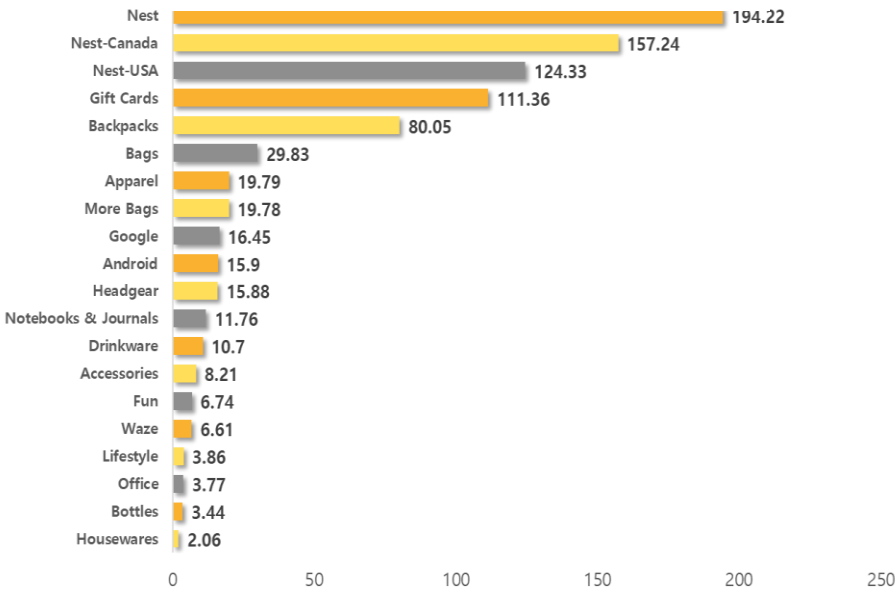


UUUU : 모든 수치에서 높은 등급에 해당되는 고객군임을 알 수 있으며 품목 고급화를 통한 고비용 마케팅 필요

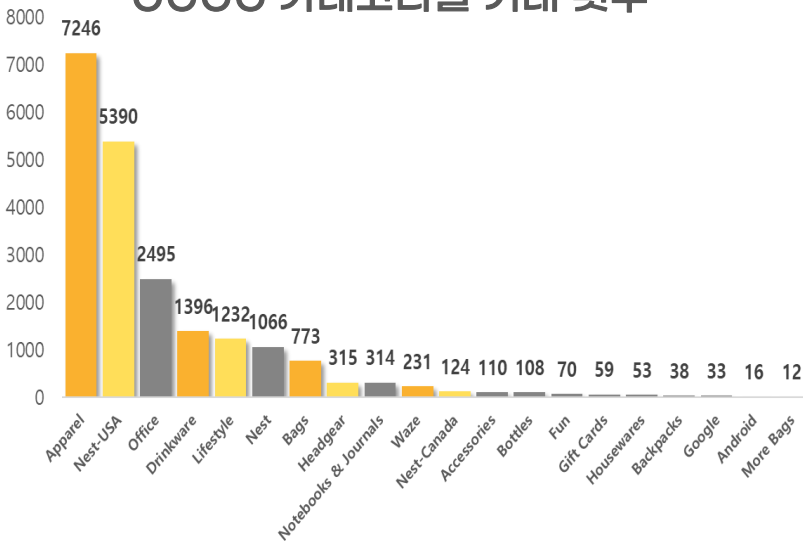


- UUUU는 수, 금요일에 거래 횟수가 많은 모습을 보여주고 있으며 Apparel, Nest-USA, Office 순으로 거래 횟수가 많은 모습을 확인 할 수 있음
- 평균 단가가 비싼 카테고리에 대한 구매도 적지않게 이루어지고 있는 모습이 확인되어짐

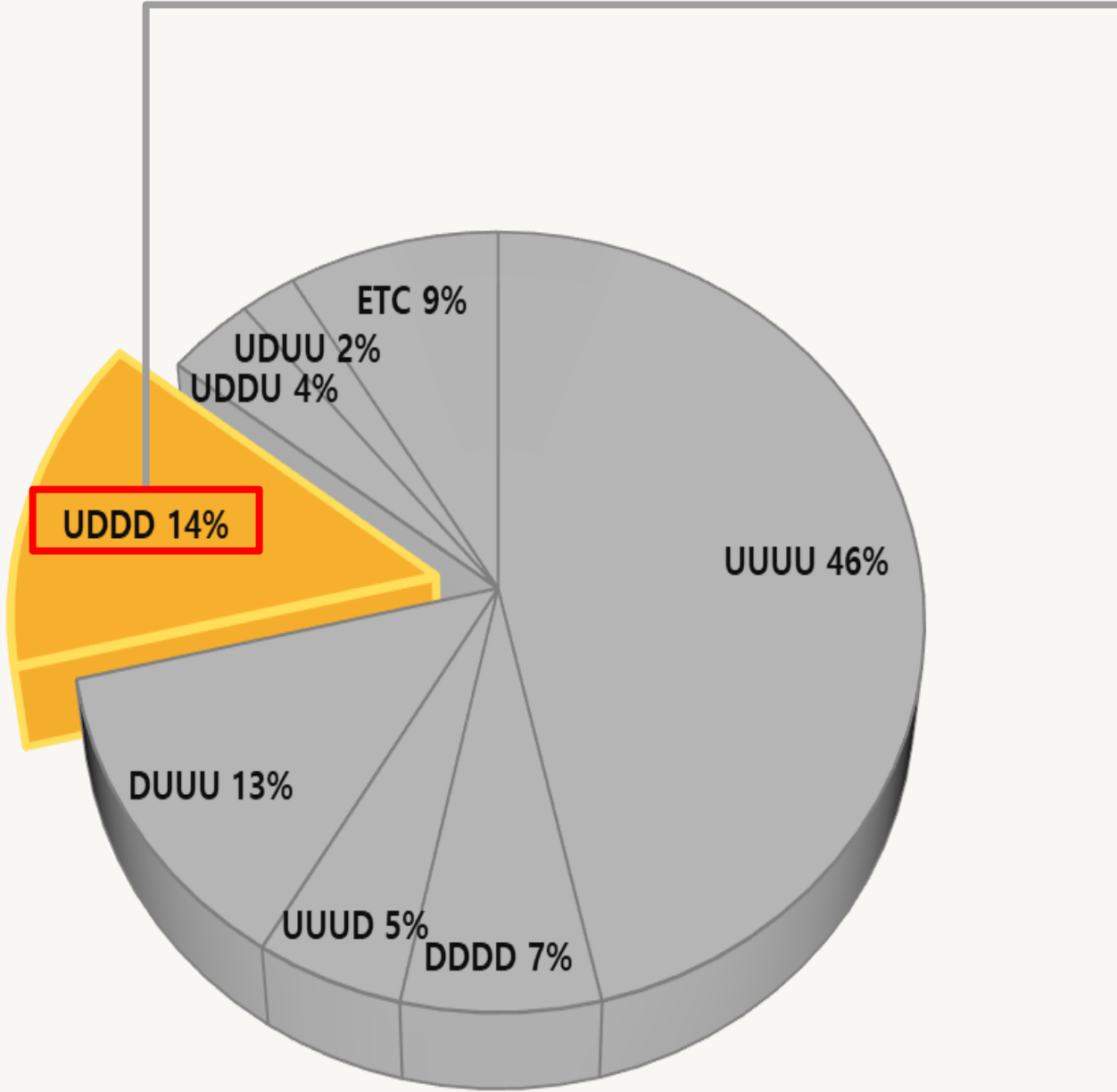
전체 카테고리 별 평균 금액



UUUU 카테고리별 거래 횟수

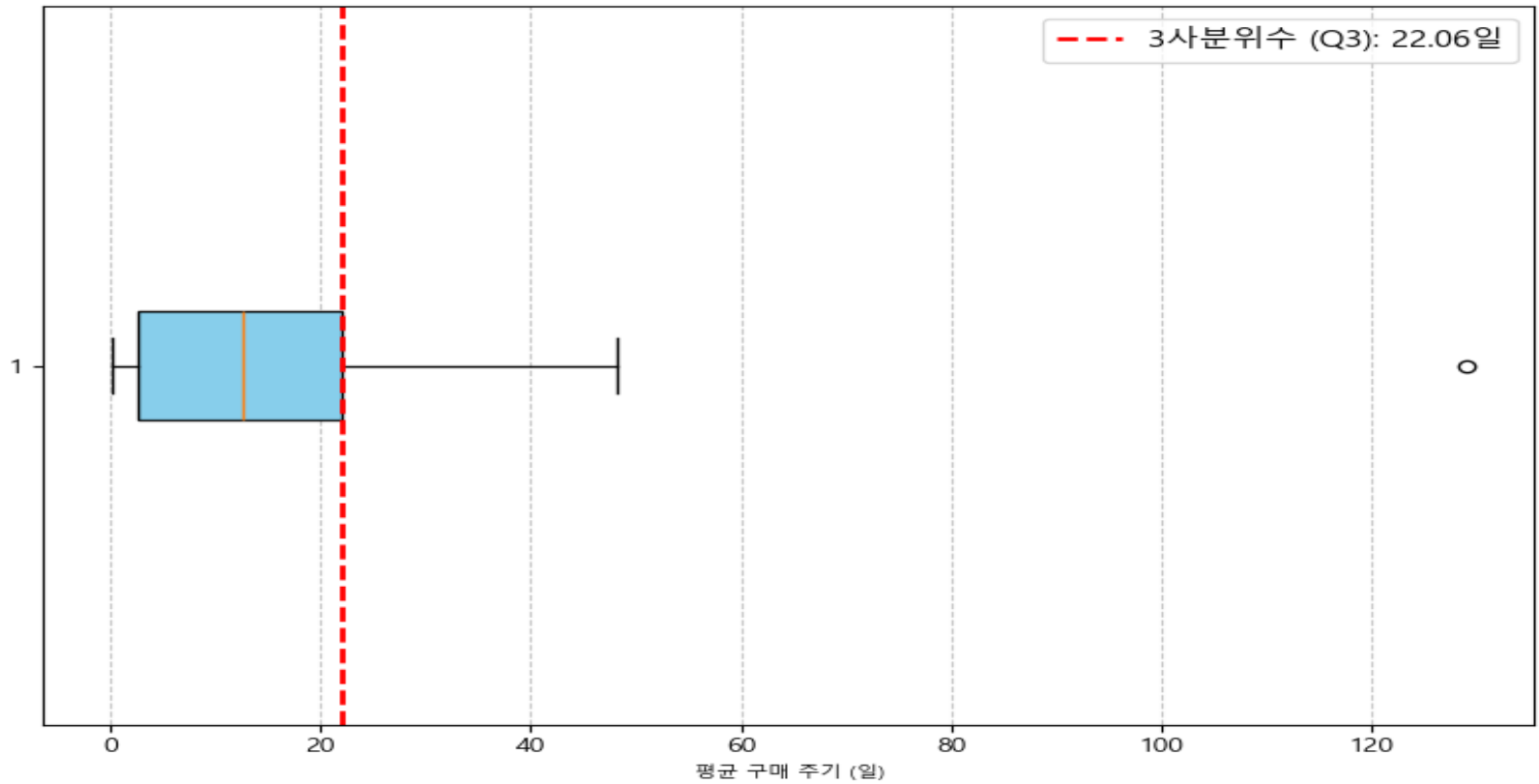


# 04. RFMV 분석 - 조합별 특성



UDDD : R은 U에 해당되지만 다른 수치들이 모두 낮은 등급인 것을 보아 재참여 초기 단계의 고객에 해당되며 재구매 유도를 위한 구매 평균 주기를 이용한 마케팅 필요

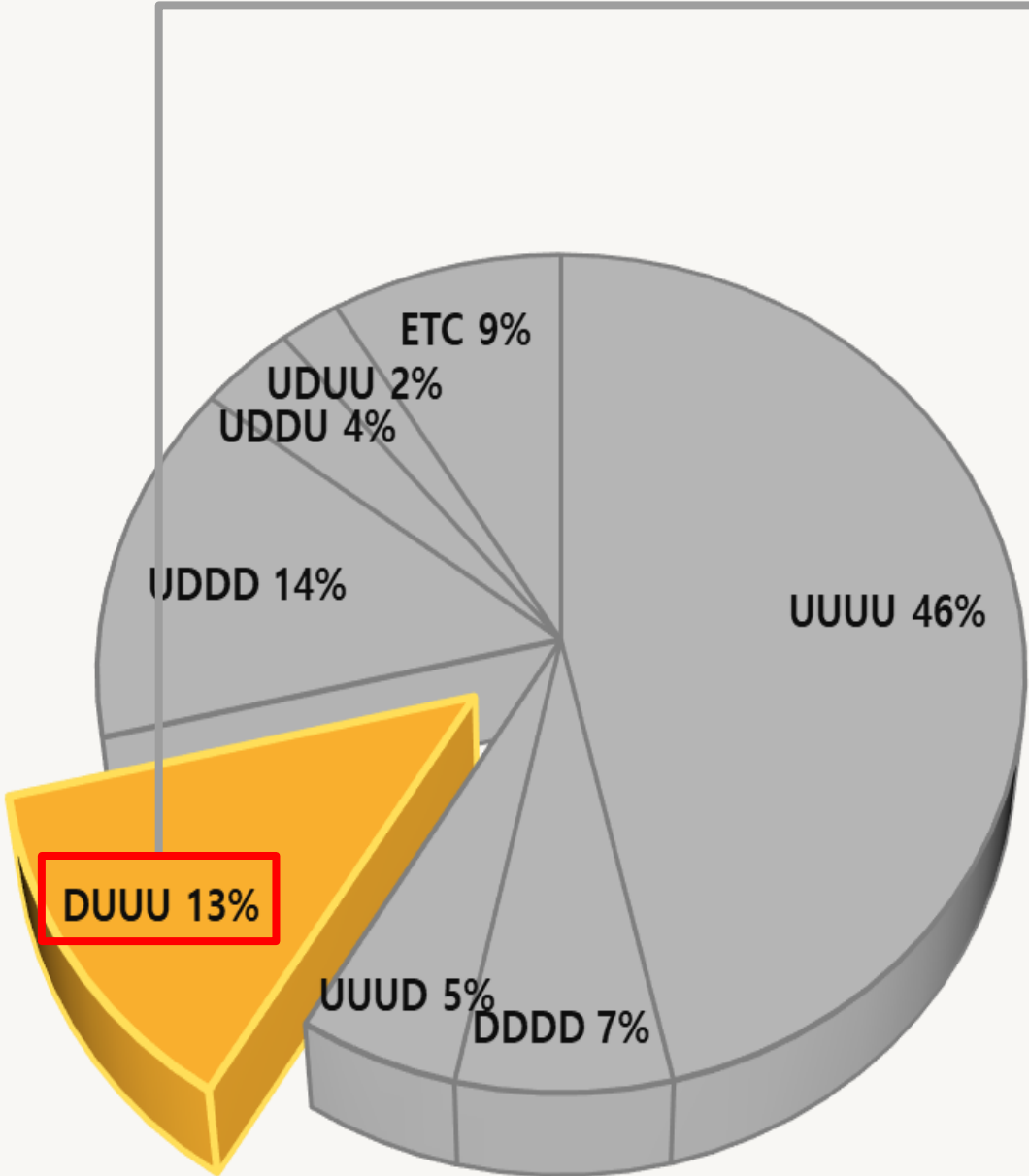
UDDD 구매 평균 주기



- 대부분의 평균 주기가 2일~22일 사이에 분포되어 있음.
- 주기 평균은 약 17일, 3사분위수는 약 22일로 나타남.

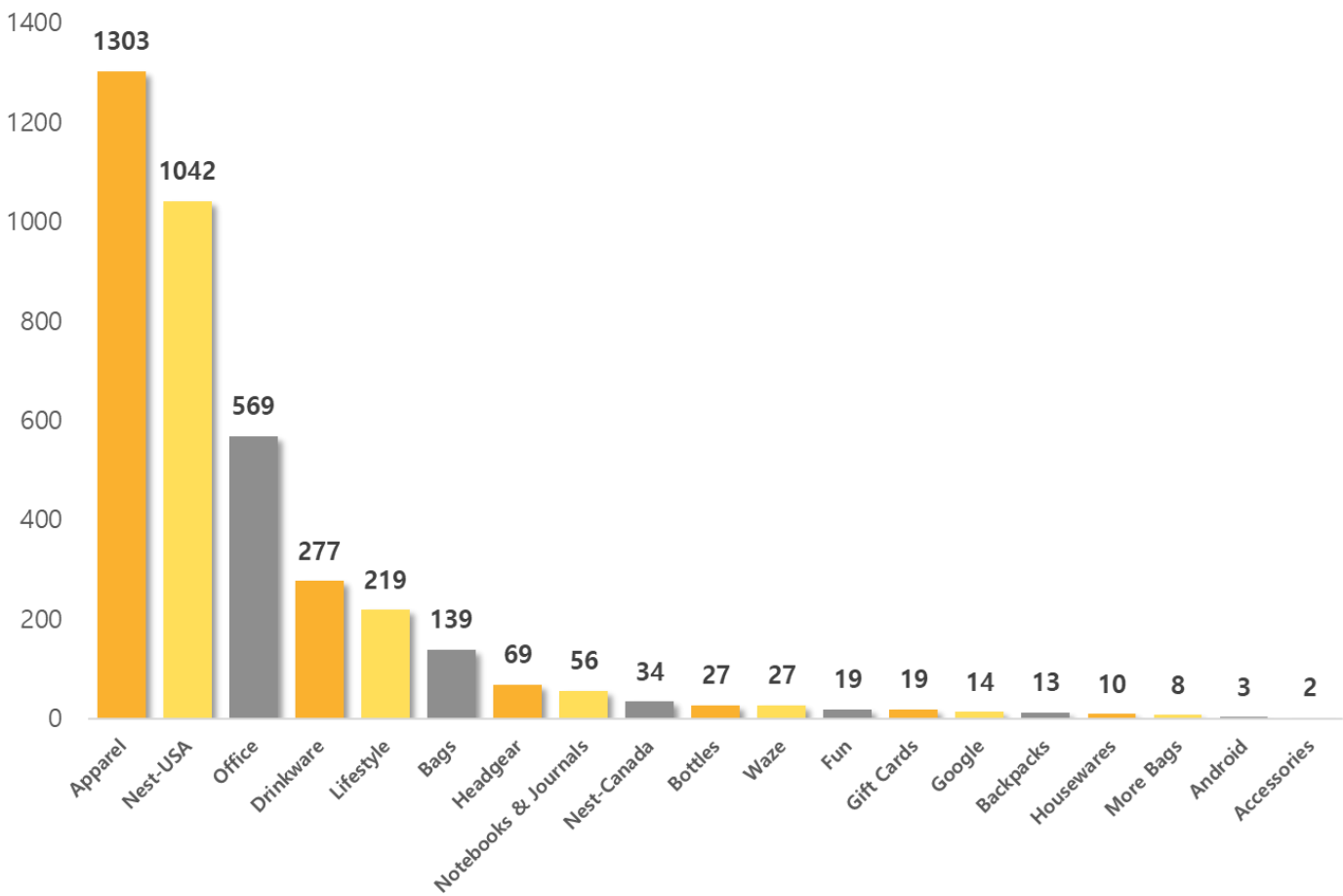


# 04. RFMV 분석 - 조합별 특성



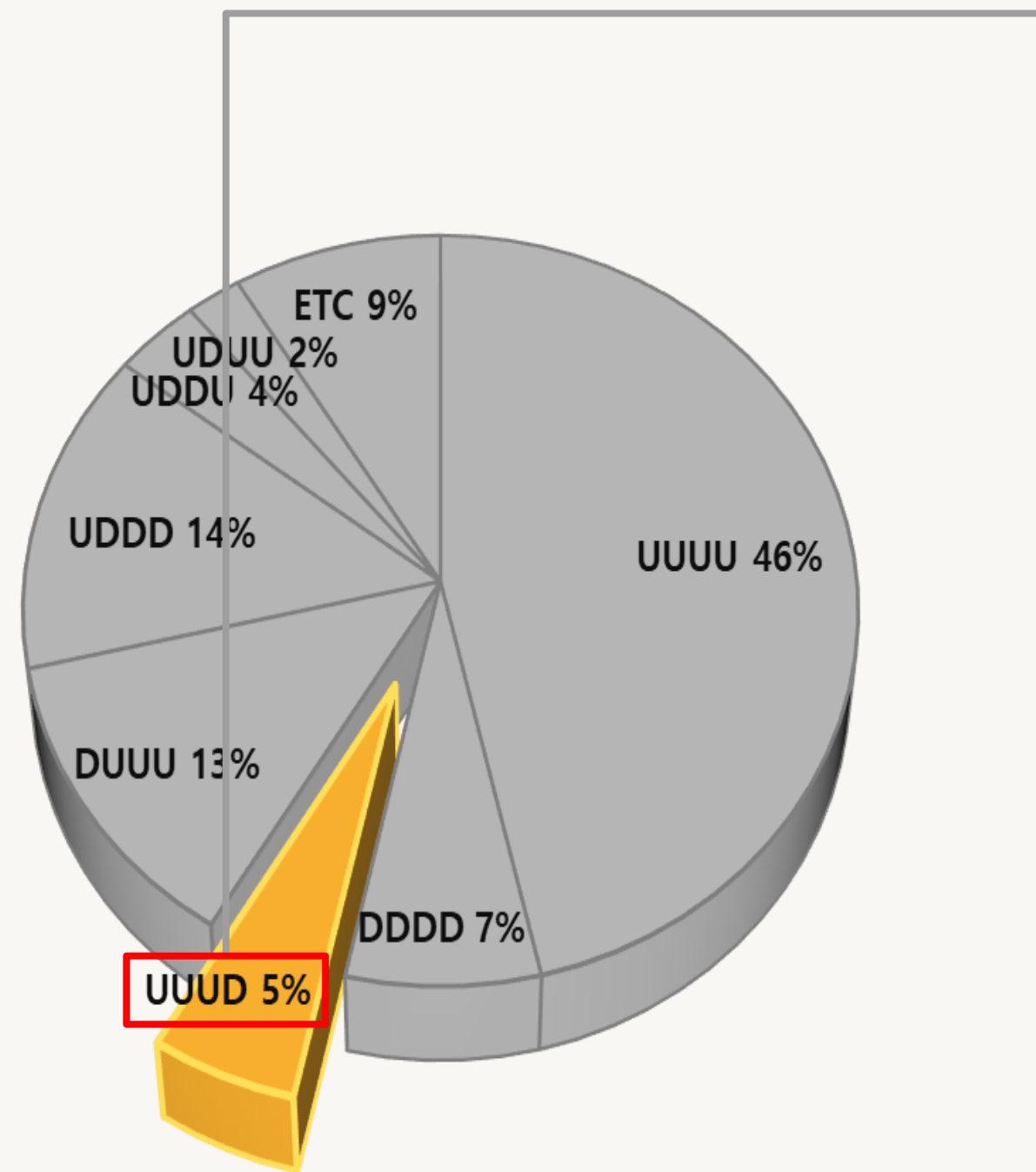
DUUU: 복귀 유도를 위한 이탈 전 구매 이력에 대한 분석

DUUU 카테고리별 구매 횟수



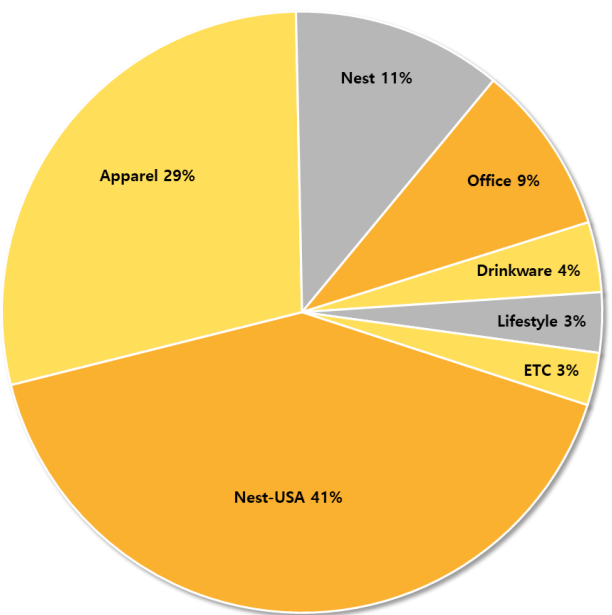
→ DUUU 고객층은 마지막 구매 이후 시간이 지난 고객들로, 이들을 복귀시키기 위해 Apparel, Nest-USA, Office, Drinkware, Lifestyle 제품을 중심으로 마케팅을 강화할 필요가 있어 보임

# 04. RFMV 분석 - 조합별 특성



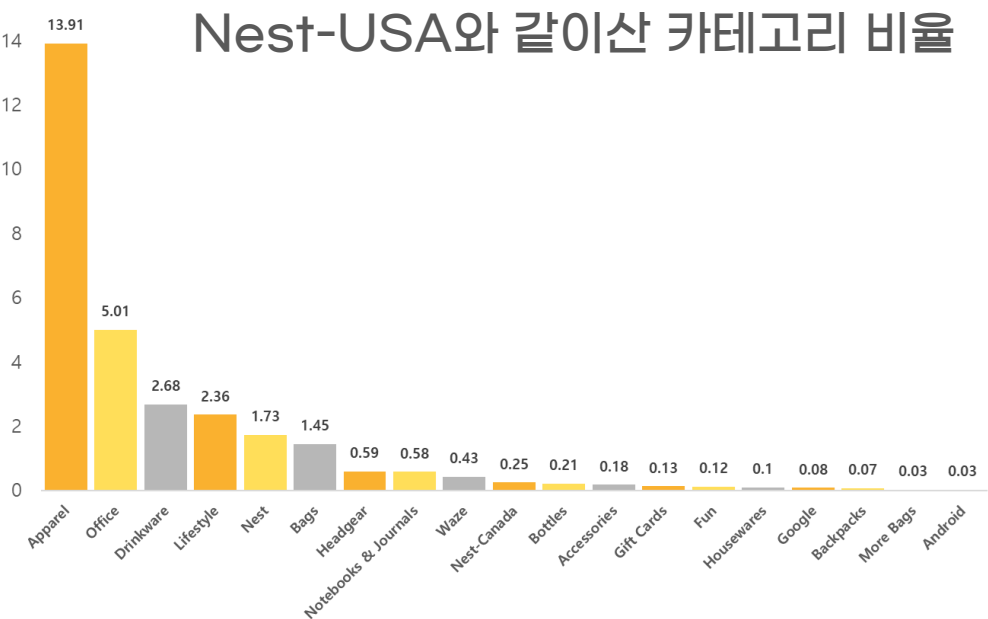
## UUUD: 구매 카테고리 다양화를 위한 연관분석 등 고비용 마케팅 필요

UUUD 제품 카테고리별 구매 수량

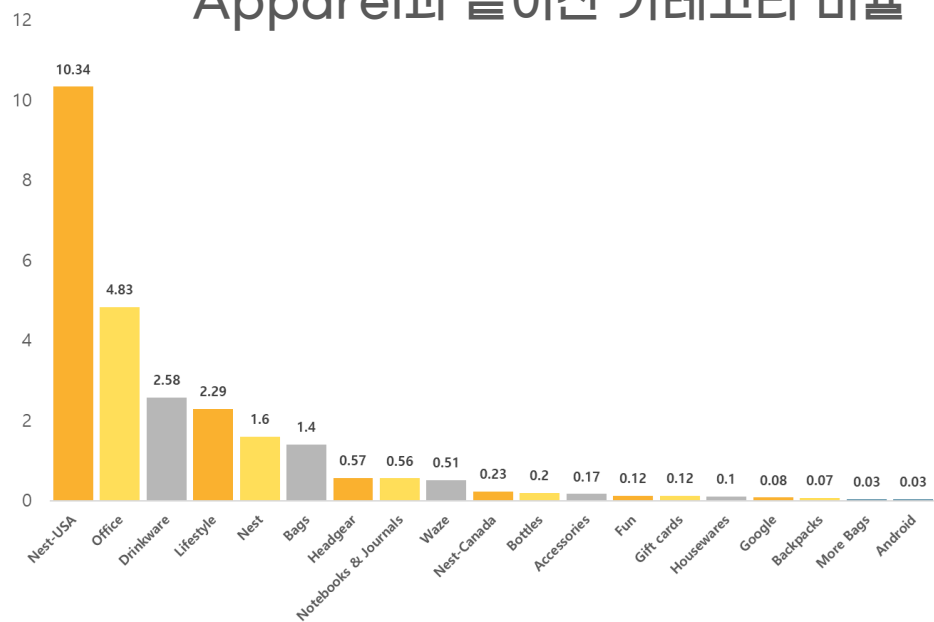


- UUUD 고객층의 주요 구매 카테고리가 Nest-USA, Apparel인 것을 확인.
- 이 두개의 카테고리를 연관분석을 통해 같이 구매한 제품 카테고리의 비율을 확인함.
- Nest-USA, Apparel와 함께 자주 구매된 카테고리가 Office, Drinkware, Lifestyle임을 확인.

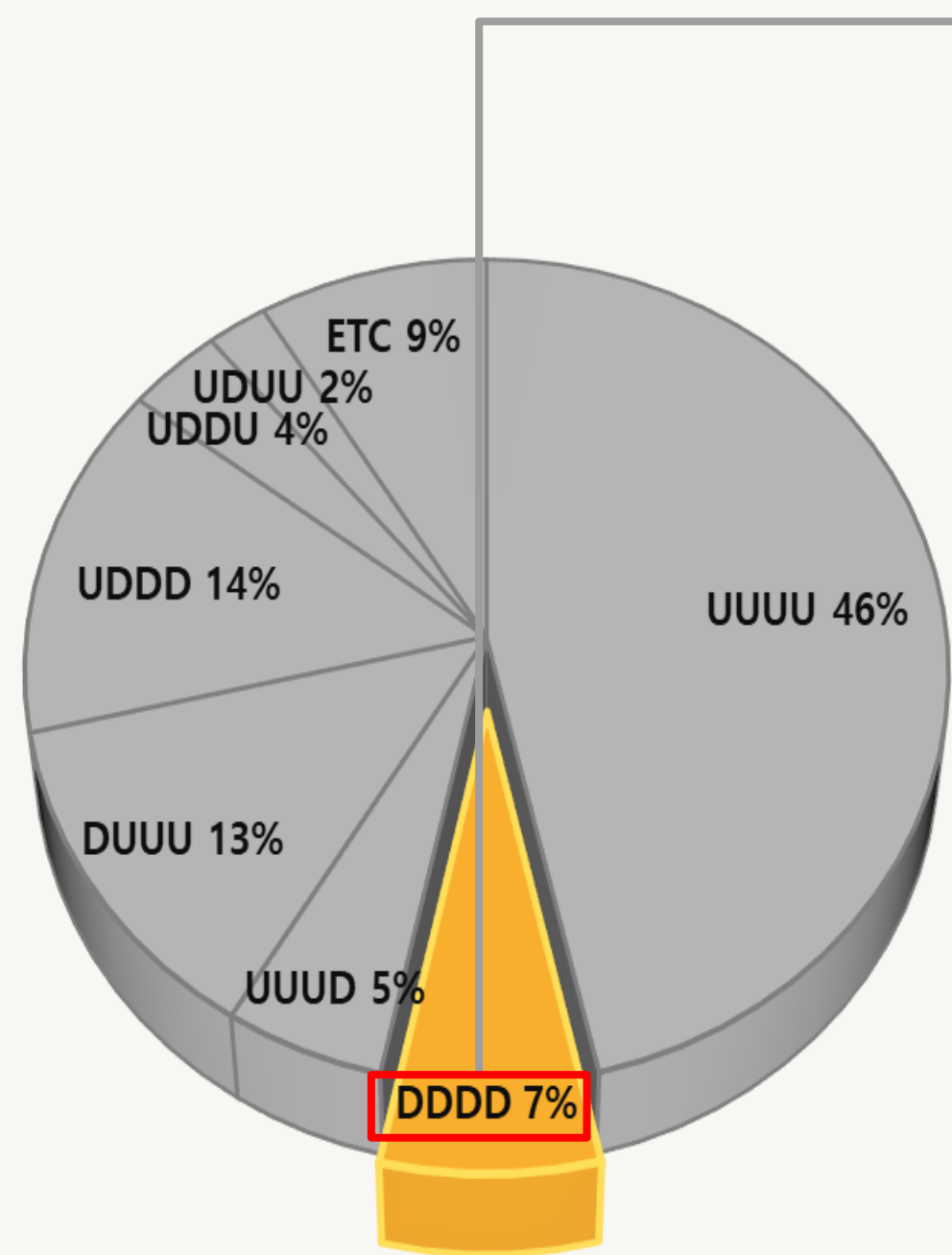
Nest-USA와 같이산 카테고리 비율



Apparel과 같이산 카테고리 비율

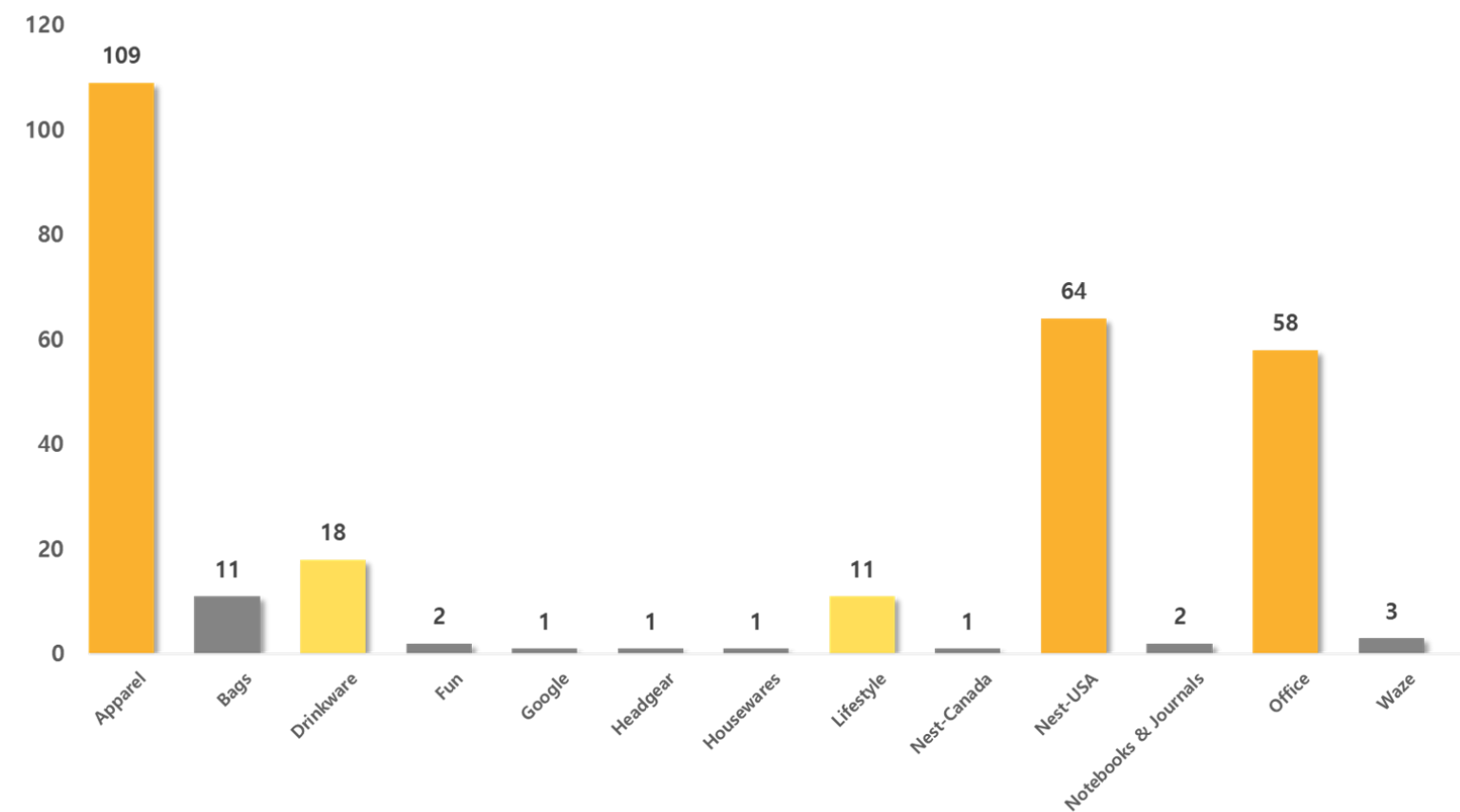


# 04. RFMV 분석 - 조합별 특성

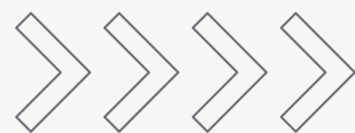


DDDD : 대부분 이탈 고객 , 과거 구매 제품 분석 등 저비용 마케팅 필요

DDDD고객층의 카테고리별 구매 횟수



→ DDDD 고객층은 대부분 이탈 고객으로 과거 Apparel , Nest-USA, Office 카테고리의 구매 횟수가 높은 것으로 나타남.

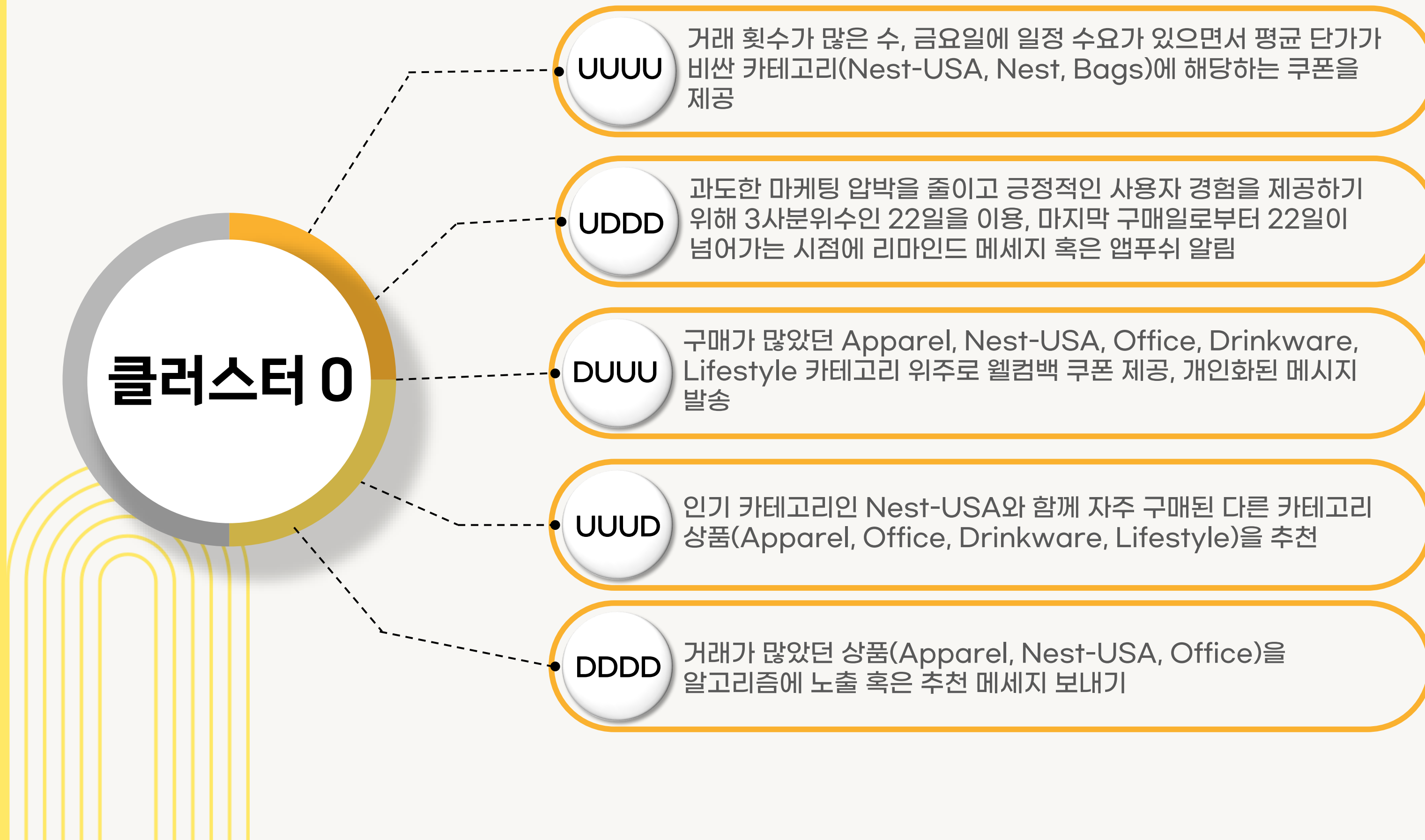


# 5. 최종 마케팅 전략 및 대시보드

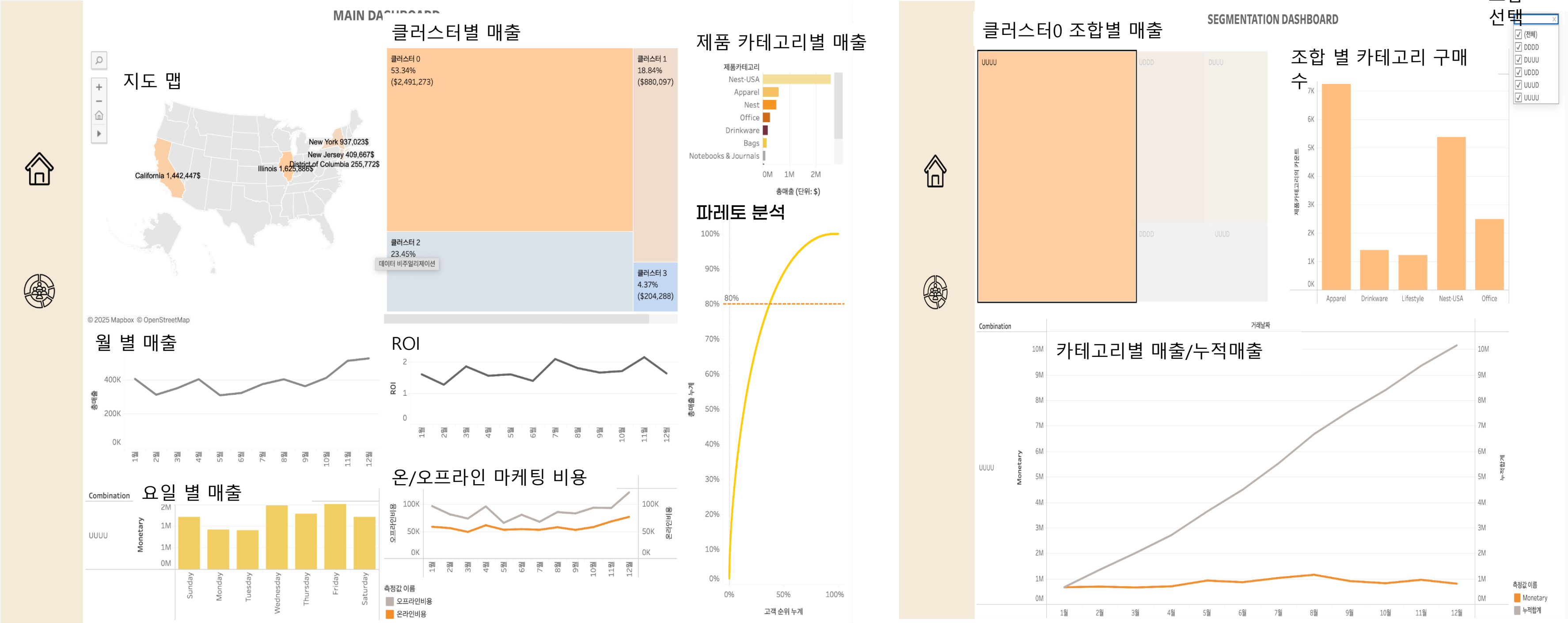




## 05. 최종 마케팅 전략



# 05. 최종 대시보드



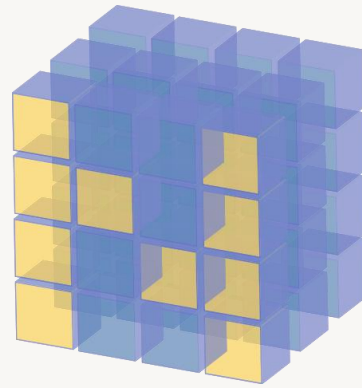
# 개발 환경



matplotlib



PYTHON



NumPy



ANACONDA  
Powered by Continuum Analytics



seaborn



**THANK YOU**