이커머스 고객 세분화 분석 아이디어 경진대회

권영후 송채연 양하영 최윤석 함형록

contents

1. 주제선정과정

https://dacon.io/

고객들의 거래 데이터를 바탕으로 의미 있는 인사이트를 도출하는 프로젝트에 관심이 있는 인원끼리 모여 해당 대회 참여.

데이터 소개

2. 데이터 설명

전체 데이터셋

DATASETS	INFO
OnlineSales	고객의 모든 거래 기록(52924, 10)
Customer	고객 정보 (1468, 4)
Discount	할인 정보 (204, 4)
Tax	서비스세 정보 (20, 2)
Marketing	마케팅 비용 정보 (365, 3)

OnlineSales

COLUMNS	INFO
고객ID	고객 고유 ID
거래ID	거래 고유 ID
거래날짜	거래가 이루어진 날짜
제품ID	제품 고유 ID
제품카테고리	제품이 포함된 카테고리
수량	주문한 품목 수
평균금액	수량 1개당 가격 (단위 : 달러)
배송료	배송비용 (단위 : 달러)
쿠폰 상태	할인쿠폰 적용 상태

Customer

COLUMNS	INFO
고객ID	고객 고유 ID
성별	고객 성별
고객지역	고객 지역
가입기간	가입기간 (단위 : 월)

Discount

COLUMNS	INFO
월	월(Month) 정보
제품카테고리	제품이 포함된 카테고리
쿠폰코드	쿠폰코드
할인율	해당 쿠폰에 대한 할인율(%)

Marketing

COLUMNS	INFO
날짜	마케팅이 이루어진 날짜
오프라인비용	오프라인 마케팅 비용 (단위 : 달러)
온라인비용	온라인 마케팅 비용 (단위 : 달러)

Tax

COLUMNS	INFO
제품카테고리	제품이 포함된 카테고리
GST	서비스세 (%)

2-1. JOIN된 데이터

	고객ID	거래ID	거래날짜	제품ID	제품카테고리	수량	평균금액	배송료	쿠폰상태	월
0	USER_1358	Transaction_0000	2019-01-01	Product_0981	Nest-USA	1	153.71	6.5	Used	1
1	USER_1358	Transaction_0001	2019-01-01	Product_0981	Nest-USA	1	153.71	6.5	Used	1
2	USER_1358	Transaction_0002	2019-01-01	Product_0904	Office					
3	USER_1358	Transaction_0003	2019-01-01	Product_0203	Apparel					
4	USER_1358	Transaction_0003	2019-01-01	Product_0848	Bags					

이 데이터를 가지고
고객을 분류하는게 목표!

쿠폰코드	할인율	GST	오프라인비용	온라인비용	총 마케팅비용	요일	할인금액	총결제금액	총결제금액(배송비포함)
ELEC10	0.1	0.10	4500	2424.5	6924.5	Tuesday	15.371	152.1729	158.6729
ELEC10	0.1	0.10	4500	2424.5	6924.5	Tuesday	15.371	152.1729	158.6729
OFF10	0.1	0.10	4500	2424.5	6924.5	Tuesday	0.205	2.0295	8.5295
SALE10	0.0	0.18	4500	2424.5	6924.5	Tuesday	0.000	103.4270	109.9270
AIO10	0.1	0.18	4500	2424.5	6924.5	Tuesday	1.650	17.5230	24.0230

고객 세분화

3-1. 인구 특성 기반 세분화

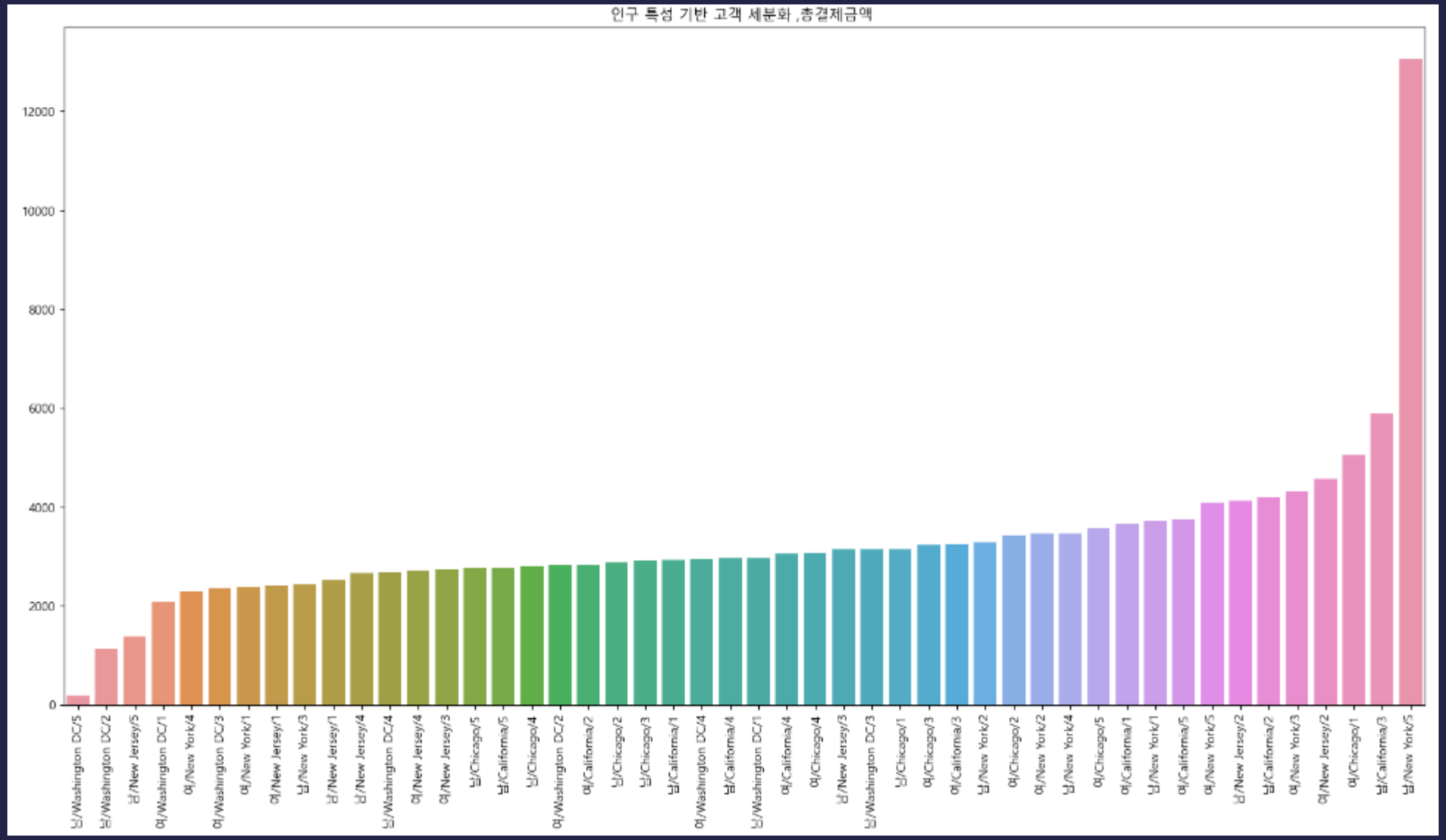
1. Customer.csv 파일을 사용하여 onlinesales.csv의 총 구매금액을 계산 후 고객 아이디별 [총 구매 금액] 컬럼 추가

2. customer.csv 파일의 가입기간을 범주형으로 전환

고객ID	성별	고객지역	총결제금액	가입년수
USER_1358	남	Chicago	24.98174	1
USER_0190	남	California	15021.70396	1
USER_0066	남	Chicago	1492.40594	1
USER_0345	여	California	1339.55528	1
USER_0683	남	California	1369.93900	1
...	...	...	...	...
USER_0513	여	New York	582.35600	1
USER_0167	여	Chicago	2384.12372	1
USER_0845	남	New Jersey	115.40750	1
USER_0504	여	New York	327.80000	1
USER_0562	여	California	6639.11141	1

1468 rows × 4 columns

3-1. 인구 특성 기반 세분화



3-1. 인구 특성 기반 세분화

가장 소비가 높은 그룹

1. '남/New York/5
2. '남/California/3'
3. '여/Chicago/1'
4. '여/New Jersey/2'
5. '여/New York/3'

가장 소비가 적은 그룹

1. '남/Washington DC/5'
2. '남/Washington DC/2'
3. '남/New Jersey/5'
4. '여/Washington DC/1'
5. '여/New York/4'

3-1. 인구 특성 기반 세분화

결론

1. 인구 특성 기반 세분화를 진행하였을 때 유의미한 패턴이 보이지 않음
2. 각 그룹별 인구수의 비율이 일정하지 않다 보니 이상치에 민감

→ 다른 세분화 기법을 사용

3-2. RFM 기법

효과적으로 고객을 세분화할 수 있는 방법이 필요

→ RFM 기법

3-2. RFM 기법

Recency : 얼마나 최근에 구매했는지

Frequency: 얼마나 자주 구매했는지

Monetary: 얼마만큼 구매했는지

※ 위 세가지 요소를 가지고 고객 기여도에 근거하여 고객의 수익성을 평가하는 모델링 기법

3-2. RFM 기법

- (R,F,M) = (1,1,1)
  - 최근에 방문도 안하고
  - 구매 횟수도 적고
  - 돈도 안 쓰는 고객

→ 마케팅 대상에서 제외

- (R,F,M) = (4,2,3)
  - 최근에 방문했고
  - 구매 빈도는 살짝 떨어지나
  - 돈은 꽤 쓰는 고객

→ 마케팅 대상에 포함

3-2. RFM 기법

일반적인 방법
→ RFM 각 점수를 모두 더한 값을 RFM_Score로 사용

저희가 채택한 방법
→ 각 RFM 점수에 가중치를 주어 새로운 RFM_Score 산정

ex) R*0.2 + F*0.4 + M*0.4 = RFM_Score

→ Recency가 세분화에 끼치는 영향 감소

3-2. 가중치 계산 과정

RFM 모형의
가중치 선택에 관한 연구

Data driven selection methods of weights in RFM Model

指導敎授 金 鐵 洙

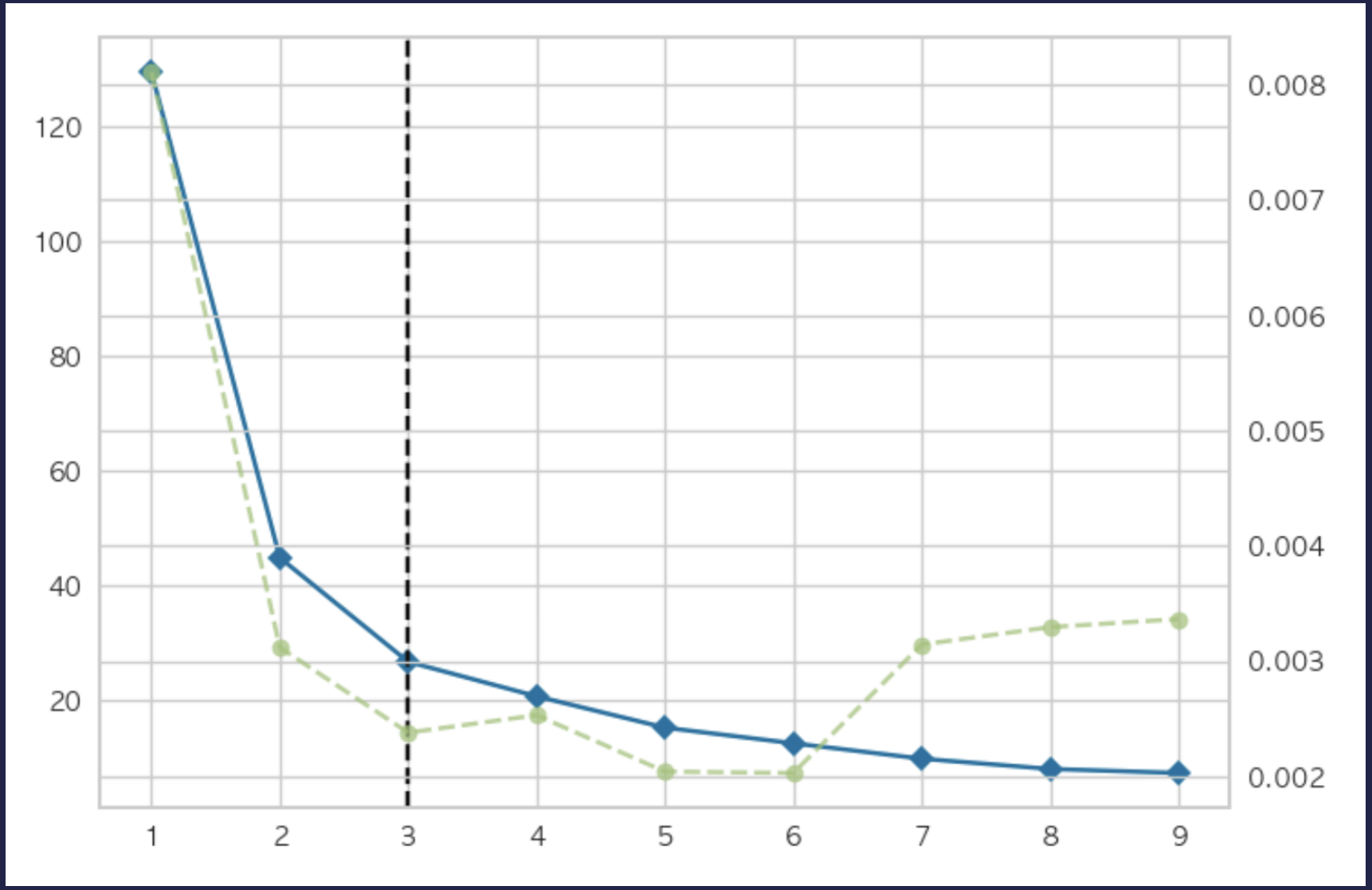
金 東 錫

3-2. 가중치 계산 과정

K-Means Clustering을 이용한 가중치 결정

- 데이터 표준화(Feature Scaling)
- scikit-learn의 elbow method 알고리즘 활용해서 적절한 k값 도출
- k개 그룹별 R,F,M의 기술통계량(평균, 표준편차) 계산
- RFM모형의 가중치를 선택하기 위해 각 그룹에 대한 기술통계량 값을 활용
    - (R, F, M) --> (W1, W2, W3)
    - W1 + W2 + W3 = 1

3-2. 가중치 계산 과정

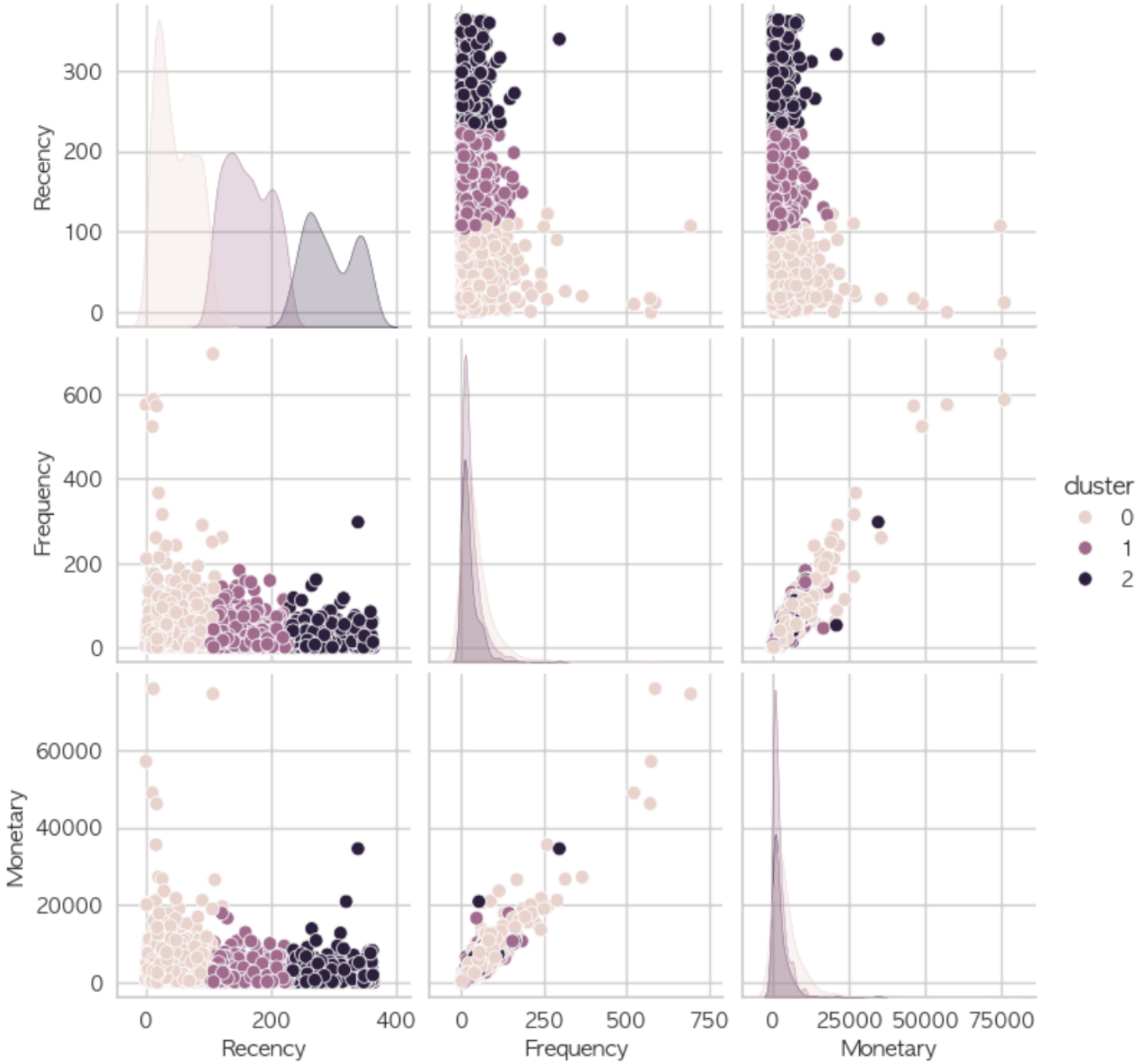


K = 3

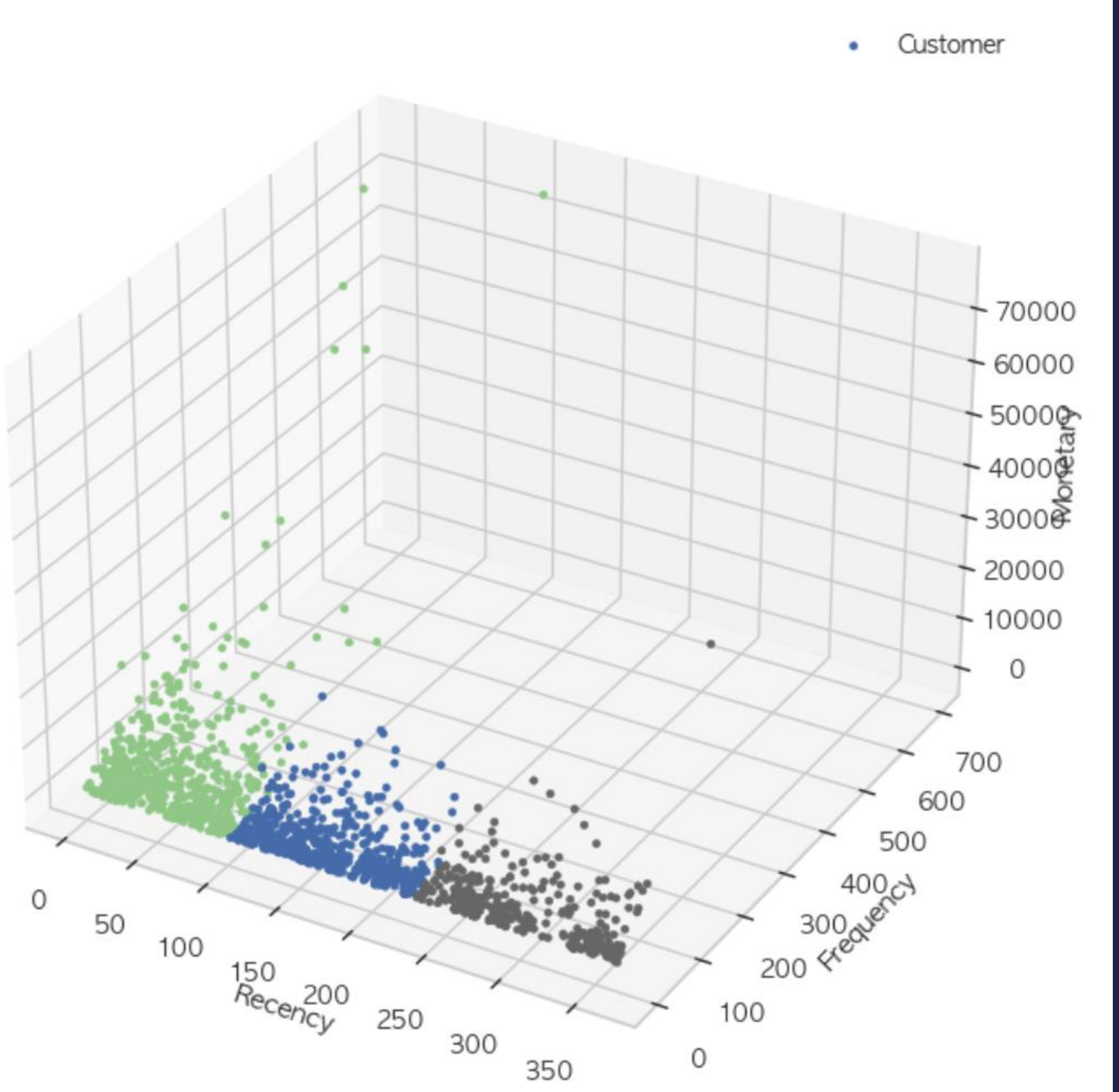
→ SSE의 감소세가
가장 완만해지는 시점

파란색 선 :
SSE (Sum of Squared Error)

녹색 선 :
학습 시간


Recency
Frequency
Monetary
duster
0
1
2
0
100
200
300
0
200
400
600
0
200
400
0
20000
40000
60000
0
250
500
750
0
25000
50000
75000

pairplot



3D 시각화

녹색 군집 :
채산성이 가장 높은 군집

회색 군집 :
채산성이 가장 낮은 군집

3-2. 가중치 계산 과정

$$w1 = \frac{\min(CV_{rn})}{CV_{r1} + CV_{r2} + \ldots + CV_{rk}} \quad (CV_{rk} = \frac{s_{rk}}{x_{rk}})$$

$$w2 = \frac{\min(CV_{fn})}{CV_{f1} + CV_{f2} + \ldots + CV_{fk}} \quad (CV_{fk} = \frac{s_{fk}}{x_{fk}})$$

$$w3 = \frac{\min(CV_{mn})}{CV_{m1} + CV_{m2} + \ldots + CV_{mk}} \quad (CV_{mk} = \frac{s_{mk}}{x_{mk}})$$

$$W1 = \frac{w1}{w1+w2+w3},\ W2 = \frac{w2}{w1+w2+w3},\ W3 = \frac{w3}{w1+w2+w3}$$

$$(W1 + W2 + W3 = 1)$$

$x_{r,f,m}$: R, F, M에 대한 각각의 평균

$s_{r,f,m}$: R, F, M에 대한 각각의 표준편차

$_k$: $K-Means\ Clustering$으로 나눠진 그룹수

$CV_{rfm\,\cdot\,k}$: k개 그룹에 대한 각 R, F, M의 CV

$\min(CV_{rfm\,\cdot\,n})$: R, F, M의 CV 최솟값

→

각 군집별 평균

cluster	Recency	Frequency	Monetary
0	46.73	49.37	4742.72
1	162.22	28.05	2030.77
2	295.19	24.32	2145.71

각 군집별 표준편차

cluster	Recency	Frequency	Monetary
0	29.87	68.81	6915.84
1	35.63	29.24	2362.98
2	40.05	29.24	3057.78

→

가중치 결정

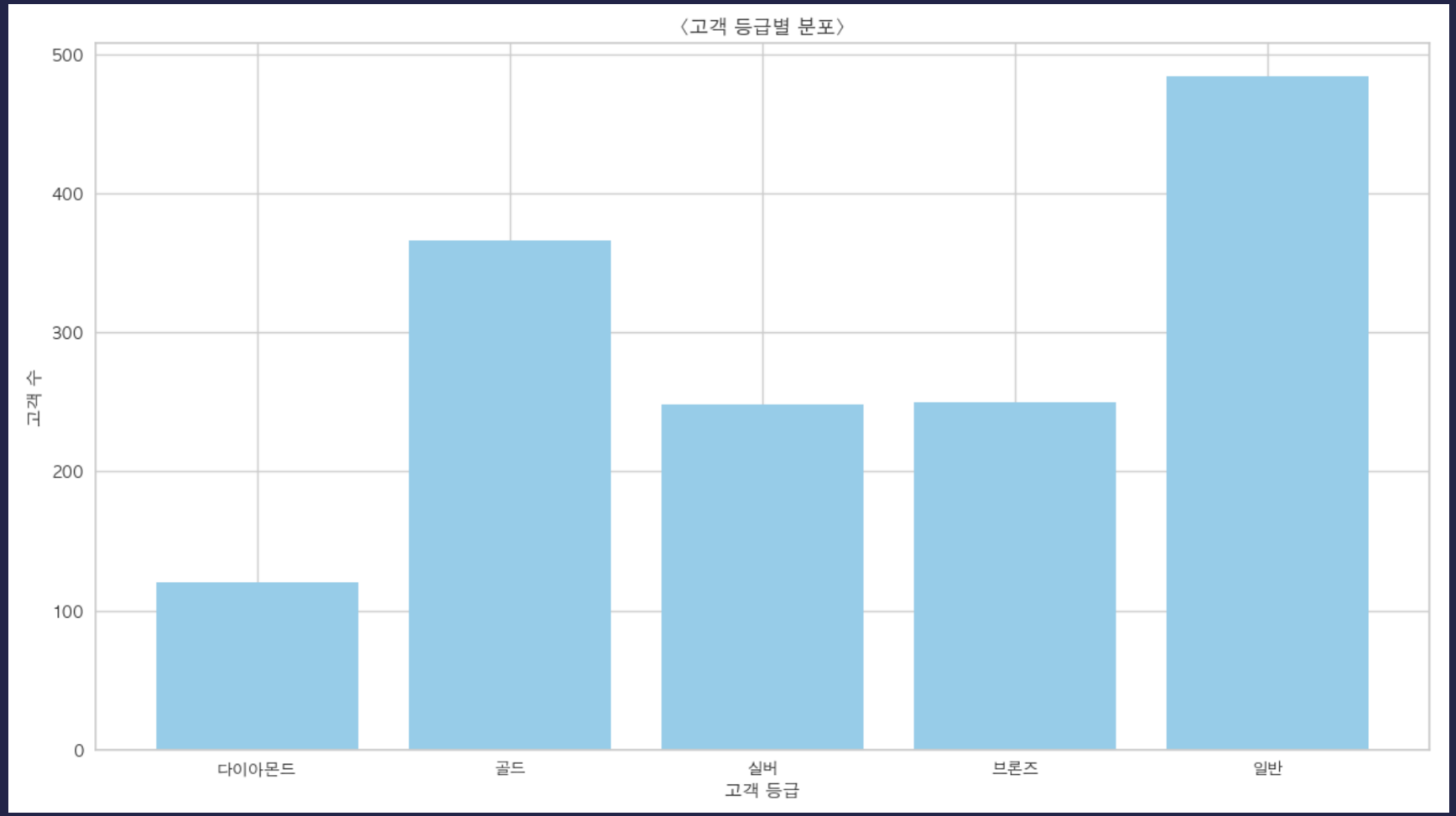
(0.2, 0.4, 0.4)

고객 등급 세분화

<고객 세분화 함수>

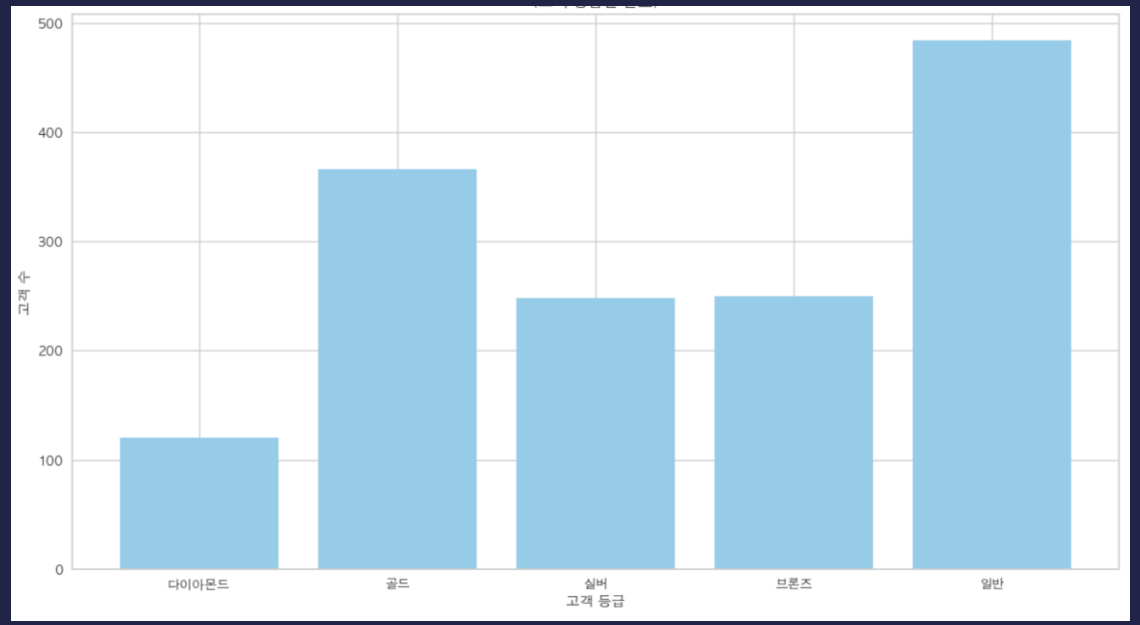
```python
def classify_customer_segment(row):

    if row['RFM_Score'] <= 45.25:
        return '일반'
    elif row['RFM_Score'] <= 60.08:
        return '브론즈'
    elif row['RFM_Score'] <= 75.00:
        return '실버'
    elif row['RFM_Score'] <= 95.20:
        return '골드'
    else:
        return '다이아몬드'
```

<고객등급별 분포>



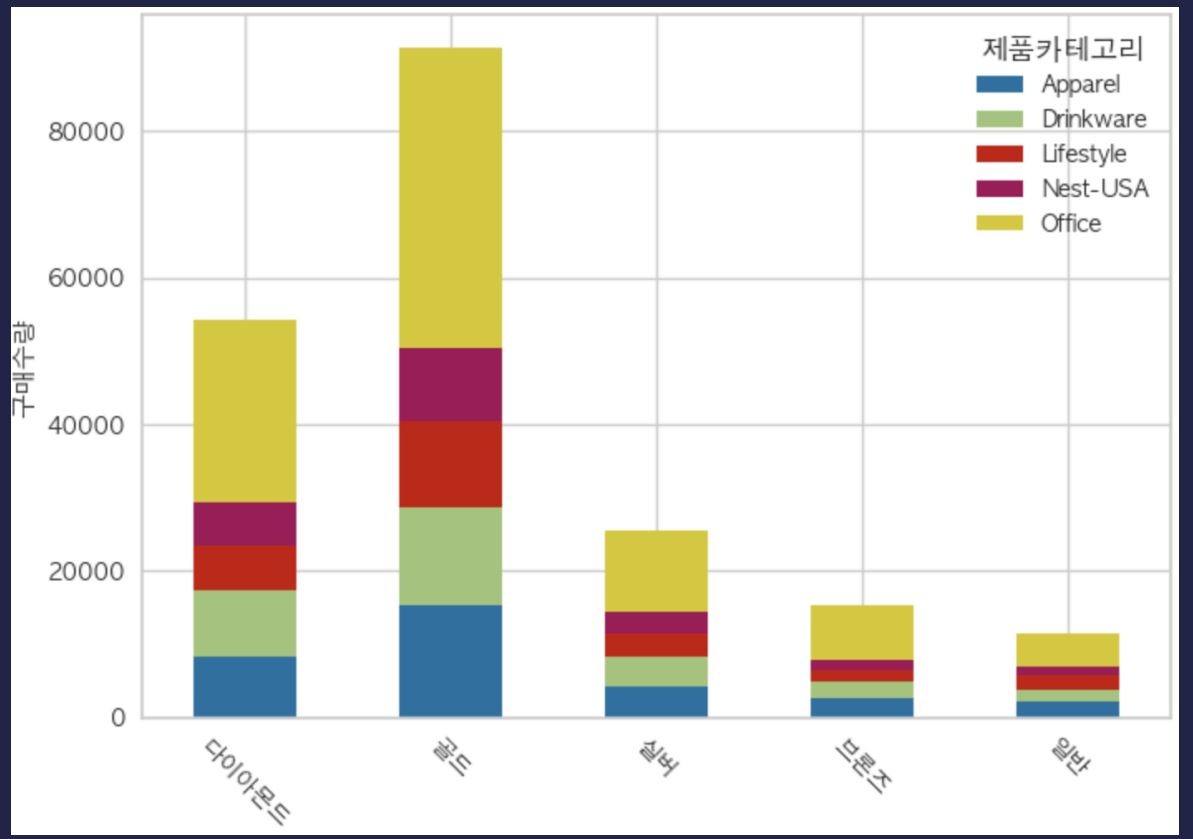
고객 등급 세분화

<고객등급 분포>



VS

<고객등급 & 제품카테고리별 구매수량 분포>



고객 등급 세분화

일반등급 회원은 정말 버려도 되는가?

월별	마케팅비용	일반등급_총결제금액	다이아_골드등급_총결제금액
마케팅비용	1.00	-0.19	0.68
일반등급_총결제금액	-0.19	1.00	-0.30
다이아_골드등급_총결제금액	0.68	-0.30	1.00

→

상관계수

마케팅비용과
일반등급 고객의 총결제액
: **-0.19**

마케팅비용과
다이아_골드등급 고객의 총결제액
: **0.68**

※ 일반등급 고객은 마케팅 대상에서 제외하는 것이 효율적

3-3. 시장바구니 분석

MBA(Market Basket Analysis,시장바구니분석,연관분석)

1. 소매업체들이 고객의 구매 패턴을 더 잘 이해함으로써 매출을 증대시키기 위한 Data mining 기법

2. 추가 매출을 창출하는 동시에 고객에게 보다 생산적이고 가치있는 쇼핑환경을 제공

※ 대표적 예시
: 기저귀 가판대 옆에 맥주를 함께 판매하는 월마트의 전략

3-3. 시장바구니 분석

연관 규칙 측정 지표

지지도(support)

- 전체 거래 중 상품 A와 상품 B가 동시에 포함되는 거래 비율

$$P(A \cap B) = \frac{\text{A와 B가 동시에 포함된 거래 수}}{\text{전체 거래 수}}$$

신뢰도(Confidence)

- 상품 A가 포함된 거래 중에서 상품 A,B가 동시에 포함되는 거래 비율

$$\frac{P(A \cap B)}{P(A)} = \frac{\text{A와 B가 동시에 포함된 거래 수}}{\text{A를 포함하는 거래 수}}$$

3-3. 시장바구니 분석

향상도 (lift)

$$\frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{\text{A와 B가 동시에 포함된 거래 수} * \text{전체 거래 수}}{\text{A를 포함하는 거래 수} * \text{B를 포함하는 거래 수}}$$

- 상품 A의 거래 중 상품 B가 포함된 거래의 비율 / 전체 상품 거래 중 상품 B가 거래된 비율 (조건부 확률)
- 두 상품이 서로 독립일 경우 → 향상도 = 1 (연관성이 없음)
- 두 상품이 서로 양의 관계일 경우 → 향상도 > 1
  (A를 구매할 때 B를 구매할 확률 높음)
- 두 상품이 서로 음의 관계일 경우 → 향상도 < 1
  (A를 구매할 때 B를 구매할 확률이 낮음)

3-3. 시장바구니 분석

다이아몬드그룹 연관 분석

	antecedents	consequents	antecedent support	consequent support	support	confidence
50	(Drinkware, Lifestyle)	(Office)	0.021661	0.131975	0.014489	0.668874
4	(Headgear)	(Apparel)	0.023239	0.307847	0.014632	0.629630
44	(Bags, Drinkware)	(Office)	0.017358	0.131975	0.010185	0.586777
23	(Notebooks & Journals)	(Office)	0.021087	0.131975	0.011333	0.537415
26	(Drinkware, Lifestyle)	(Apparel)	0.021661	0.307847	0.010759	0.496689

골드그룹 연관 분석

	antecedents	consequents	antecedent support	consequent support	support	confidence
56	(Bags, Drinkware)	(Office)	0.021458	0.141590	0.013601	0.633858
4	(Headgear)	(Apparel)	0.027794	0.323477	0.017234	0.620061
62	(Drinkware, Lifestyle)	(Office)	0.026105	0.141590	0.015376	0.588997
54	(Office, Bags)	(Drinkware)	0.025091	0.101799	0.013601	0.542088
23	(Notebooks & Journals)	(Office)	0.025851	0.141590	0.013517	0.522876

3-3. 시장바구니 분석

결론

- 각 등급별 소비자들의 구매 경향을 파악
- 각 등급별로 특정 상품 구매 시 연관상품을 할인하는 방식으로 구매 유도
- 온라인판매 시 각 등급별 상품배치를 세분화하여 구매 유도

3-4. 코호트 분석

코호트 분석이란?

- 특정 기간 동안 비슷한 특성이나 경험을 공유하는 그룹인 코호트를 식별하고 분석하는 기법
- 주로 비즈니스, 마케팅, 의학, 교육 등 다양한 분야에서 활용
- 특정 그룹의 행동이나 특성에 대한 인사이트를 얻을 수 있음.

3-4. 코호트 분석

코호트 설정

	FirstOrder	거래년월	총고객수	총금액	총거래수	CohortPeriod
0	2019.01	2019.01	215	434282.13775	4063	0
1	2019.01	2019.02	13	40483.20322	437	1
2	2019.01	2019.03	24	45819.38372	620	2
3	2019.01	2019.04	34	124184.55989	768	3
4	2019.01	2019.05	23	30455.31544	450	4
...	...	...	...	...	...	...

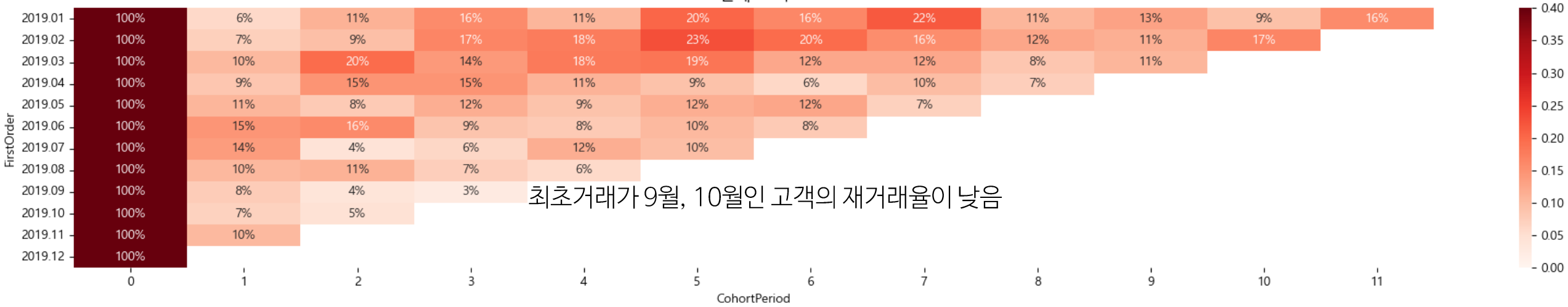
- 코호트 : 특정 기간에 공통된 특성을 갖는 개체나 그룹으로 정의

  → FirstOrder & Cohort Period를 코호트로 설정

3-4. 코호트 분석

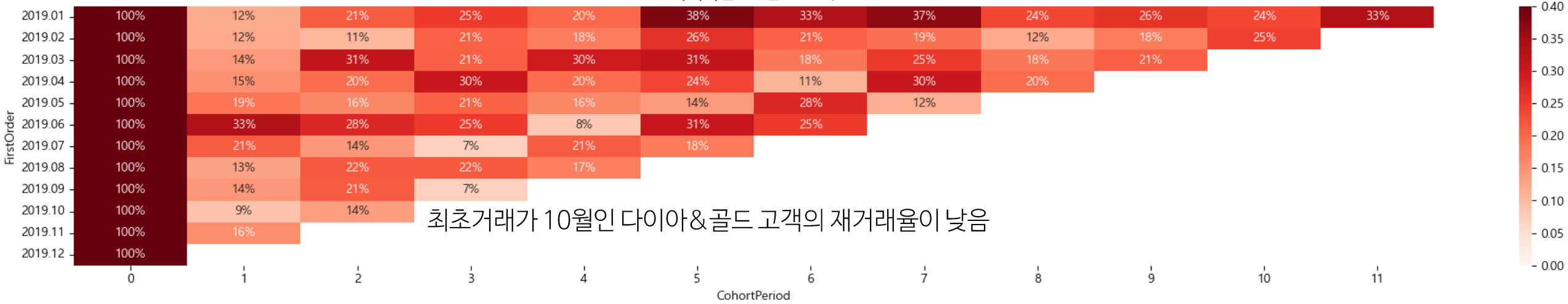
시각화

< 전체 고객 >



최초거래가 9월, 10월인 고객의 재거래율이 낮음

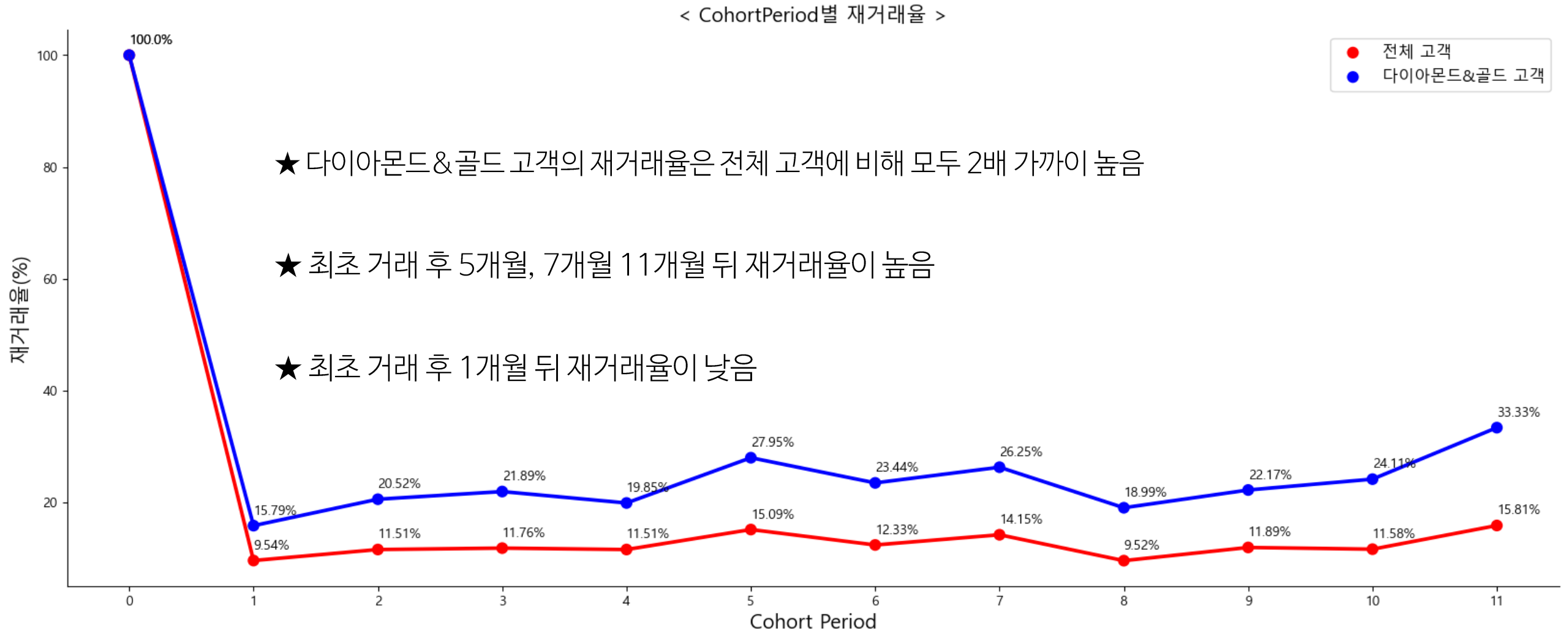
< 다이아몬드&골드 고객 >



최초거래가 10월인 다이아&골드 고객의 재거래율이 낮음

3-4. 코호트 분석

시각화



★ 다이아몬드&골드 고객의 재거래율은 전체 고객에 비해 모두 2배 가까이 높음

★ 최초 거래 후 5개월, 7개월 11개월 뒤 재거래율이 높음

★ 최초 거래 후 1개월 뒤 재거래율이 낮음

3-4. 코호트 분석

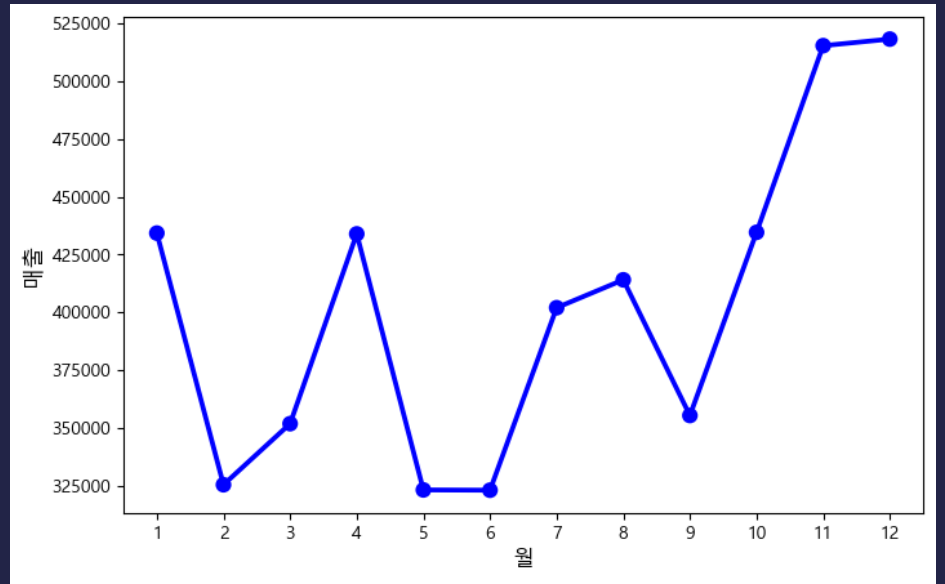
결론

1. 최초거래 1개월 뒤 재거래율이 낮고 5, 7, 11개월 뒤 재거래율이 높음
2. 다이아몬드&골드 고객들의 재거래율은 전체 고객에 비해 2배 가까이 높음
3. 최초 거래가 9월, 10월 고객들의 추후 재거래율이 낮음

→ 연말은 매출이 최대화 되는 시점

→ 9월, 10월 고객의 재거래율을 높이면

→ 매출 극대화 가능



결론 및 아쉬운 점

4. 결론 및 아쉬운 점

결론1:

RFM 기준으로 분류한 다이아몬드&골드 고객군은 매출, 거래량, 그리고 재거래율에서 타 고객들을 능가하는 탁월한 성과를 보여주고 있음. 이에 따라 마케팅 및 프로모션 전략에서 다이아몬드&골드 고객을 주요 타겟으로 설정하여, 그들에게 더욱 특별한 혜택 및 서비스를 제공함으로써 충성도를 향상시킬 수 있음.

결론2:

시장바구니 분석 기반의 맞춤형 프로모션은 고객들에게 더욱 효과적으로 다가갈 수 있는 전략임. 개별 고객의 쇼핑 습관 및 취향에 따라 맞춤형 혜택을 제공함으로써 고객들의 만족도를 향상시키고, 브랜드와의 연결을 강화할 수 있음.

4. 결론 및 아쉬운 점

아쉬운 점:

주어진 데이터셋에 대한 이해도 부족으로 미흡한 분석이 있었음. 보다 심층적이고 유의미한 결과 도출을 위해서는 데이터셋에 대한 추가적인 탐색 및 이해가 필요함. 미래의 분석에서는 데이터에 대한 깊은 이해를 바탕으로 보다 효과적인 전략 수립에 기여할 수 있을 것으로 기대함.

Thank you