

이커머스 고객 RFM 세분화 분석

1. 문제 정의

- 이커머스 환경에서 발생한 데이터를 통해 고객 세분화 기법을 사용하여 솔루션 제시
- 고객의 행동 패턴과 구매 경향을 이해함으로써 기업이 더 나은 서비스를 제공할 수 있는 방안 제시

2. 데이터 수집

	고객ID	거래ID	거래날짜	제품ID	제품카테고리	수량	평균금액	배송료	쿠폰상태
0	USER_1358	Transaction_0000	2019-01-01	Product_0981	Nest-USA	1	153.71	6.50	Used
1	USER_1358	Transaction_0001	2019-01-01	Product_0981	Nest-USA	1	153.71	6.50	Used
2	USER_1358	Transaction_0002	2019-01-01	Product_0904	Office	1	2.05	6.50	Used
3	USER_1358	Transaction_0003	2019-01-01	Product_0203	Apparel	5	17.53	6.50	Not Used
4	USER_1358	Transaction_0003	2019-01-01	Product_0848	Bags	1	16.50	6.50	Used
...
52919	USER_0504	Transaction_25056	2019-12-31	Product_0976	Nest-USA	1	121.30	6.50	Clicked
52920	USER_0504	Transaction_25057	2019-12-31	Product_0413	Apparel	1	48.92	6.50	Used
52921	USER_0504	Transaction_25058	2019-12-31	Product_0989	Nest-USA	1	151.88	6.50	Used
52922	USER_0562	Transaction_25059	2019-12-31	Product_0985	Nest-USA	5	80.52	6.50	Clicked
52923	USER_0562	Transaction_25060	2019-12-31	Product_0984	Nest-USA	4	80.52	19.99	Clicked

52924 rows × 9 columns

온라인거래와 관련된 정보

- 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)
- 동일 상품이어도 세부 옵션에 따라 가격이 다를 수 있음
- 배송료 : 배송비용 (단위 : 달러)
- 쿠폰상태 : 할인쿠폰 적용 상태

3. RFM이란?

Recency : 얼마나 최근에 구매했는가 (최근성)

Frequency : 얼마나 자주 구매했는가 (빈도)

Monetary : 얼마나 많은 금액을 지출했는가 (금액)

사용자별로 얼마나 최근에, 얼마나 자주, 얼마나 많은 금액을 지출했는지에 따라,
사용자들의 분포를 확인하거나 사용자 그룹(또는 등급)을 나누어 분류하는 분석 기법입니다.

3. RFM이란?

예를 들어, A와 B의 구매 기록을 확인해봤을 때, 누가 더 충성 고객이라고 할 수 있을까요?

고객	최근 구매일	총 구매 횟수	총 구매 금액
A	3개월 전	150	1,280,000
B	1년 전	12	50,000

최근 구매일, 구매 횟수와 금액 중 어느 것을 보더라도 B보다는 A가 충성도가 높다고 생각할 수 있습니다.

A 고객은 최근에도 구매를 진행한 사용자이기 때문에 B 사용자에게 할인 쿠폰을 보내주거나,
반대로 A 고객에게 더 잘 사용하라는 의미에서 VIP 고객을 위한 할인 쿠폰을 지급하는 프로모션을 진행할 수도 있습니다.

4. RFM Segmentation

Onlinesales_info.csv

	고객ID	거래ID	거래날짜	제품ID	제품카테고리	수량	평균금액	배송료	쿠폰상태
0	USER_1358	Transaction_0000	2019-01-01	Product_0981	Nest-USA	1	153.71	6.50	Used
1	USER_1358	Transaction_0001	2019-01-01	Product_0981	Nest-USA	1	153.71	6.50	Used
2	USER_1358	Transaction_0002	2019-01-01	Product_0904	Office	1	2.05	6.50	Used
3	USER_1358	Transaction_0003	2019-01-01	Product_0203	Apparel	5	17.53	6.50	Not Used
4	USER_1358	Transaction_0003	2019-01-01	Product_0848	Bags	1	16.50	6.50	Used
...
52919	USER_0504	Transaction_25056	2019-12-31	Product_0976	Nest-USA	1	121.30	6.50	Clicked
52920	USER_0504	Transaction_25057	2019-12-31	Product_0413	Apparel	1	48.92	6.50	Used
52921	USER_0504	Transaction_25058	2019-12-31	Product_0989	Nest-USA	1	151.88	6.50	Used
52922	USER_0562	Transaction_25059	2019-12-31	Product_0985	Nest-USA	5	80.52	6.50	Clicked
52923	USER_0562	Transaction_25060	2019-12-31	Product_0984	Nest-USA	4	80.52	19.99	Clicked

52924 rows x 9 columns



	고객ID	거래날짜	수량	평균금액
0	USER_1358	2019-01-01	1	153.71
1	USER_1358	2019-01-01	1	153.71
2	USER_1358	2019-01-01	1	2.05
3	USER_1358	2019-01-01	5	17.53
4	USER_1358	2019-01-01	1	16.50
...
52919	USER_0504	2019-12-31	1	121.30
52920	USER_0504	2019-12-31	1	48.92
52921	USER_0504	2019-12-31	1	151.88
52922	USER_0562	2019-12-31	5	80.52
52923	USER_0562	2019-12-31	4	80.52

52924 rows x 4 columns

RFM 분석에 필요한 거래일, 수량, 금액에 대한 컬럼만 가져와서 데이터프레임을 구성합니다.

4. RFM Segmentation

	고객ID	거래날짜	수량	평균금액
0	USER_1358	2019-01-01	1	153.71
1	USER_1358	2019-01-01	1	153.71
2	USER_1358	2019-01-01	1	2.05
3	USER_1358	2019-01-01	5	17.53
4	USER_1358	2019-01-01	1	16.50
...
52919	USER_0504	2019-12-31	1	121.30
52920	USER_0504	2019-12-31	1	48.92
52921	USER_0504	2019-12-31	1	151.88
52922	USER_0562	2019-12-31	5	80.52
52923	USER_0562	2019-12-31	4	80.52

52924 rows × 4 columns



	고객ID	거래날짜	수량	평균금액	구매금액
0	USER_1358	2019-01-01	1	153.71	153.71
1	USER_1358	2019-01-01	1	153.71	153.71
2	USER_1358	2019-01-01	1	2.05	2.05
3	USER_1358	2019-01-01	5	17.53	87.65
4	USER_1358	2019-01-01	1	16.50	16.50
...
52919	USER_0504	2019-12-31	1	121.30	121.30
52920	USER_0504	2019-12-31	1	48.92	48.92
52921	USER_0504	2019-12-31	1	151.88	151.88
52922	USER_0562	2019-12-31	5	80.52	402.60
52923	USER_0562	2019-12-31	4	80.52	322.08

52924 rows × 5 columns



	고객ID	최근구매일	수량	구매금액
	USER_0000	2019-09-15	3	30.99
	USER_0001	2019-11-02	342	13834.90
	USER_0002	2019-10-19	209	1442.12
	USER_0003	2019-12-14	21	1360.07
	USER_0004	2019-09-15	56	1442.47

	USER_1463	2019-04-05	46	544.34
	USER_1464	2019-10-05	141	2363.05
	USER_1465	2019-06-20	10	101.56
	USER_1466	2019-10-23	2	298.00
	USER_1467	2019-10-10	155	6362.77

1468 rows × 3 columns

수량 x 평균금액으로 거래 당 총 구매금액을 계산하고, 고객ID별로 그룹화를 진행합니다.

4. RFM Segmentation

고객ID	거래날짜	수량	구매금액
USER_0000	2019-07-17	2	20.5
USER_0000	2019-09-15	1	10.49
...			
USER_1467	2019-09-22	122	5128.44
USER_1467	2019-10-10	33	1234.33



고객ID	최근구매일	수량	구매금액
USER_0000	2019-09-15	3	30.99
...			
USER_1467	2019-10-10	155	6362.77

최근구매일 출력 함수

```
1 def last_date(USER_ID):
2     return RFM_df[RFM_df['고객ID'] == USER_ID]['거래날짜'].max()
3
4 last_date('USER_0000')
```

'2019-09-15'

고객별 구매 수량과 금액 계산

```
1 # 고객ID로 그룹화하여 고객별 구매수량, 구매금액 계산
2 user_grp = RFM_df.groupby('고객ID')[['수량', '구매금액']].sum()
3 user_grp
```


4. RFM Segmentation

고객군을 분류하기 위해 아래와 같이 기준을 세워줍니다.

- Recency : 2020-01-01일 기준으로 최근 **한 달** 이내에 결제되었는지?
- Frequency : **150회** 이상 구매하였는지?
- Monetary : **3000달러** 이상 구매하였는지?

```
1 print('고객별 구매수량 :', user_grp['수량'].mean())
2 print('고객별 구매금액 :', user_grp['구매금액'].mean())
```

고객별 구매수량 : 162.14782016348775
고객별 구매금액 : 3181.740204359674

고객ID	최근구매일	수량	구매금액
USER_0000	2019-09-15	3	30.99
USER_0001	2019-11-02	342	13834.90
USER_0002	2019-10-19	209	1442.12
USER_0003	2019-12-14	21	1360.07
USER_0004	2019-09-15	56	1442.47
...
USER_1463	2019-04-05	46	544.34
USER_1464	2019-10-05	141	2363.05
USER_1465	2019-06-20	10	101.56
USER_1466	2019-10-23	2	298.00
USER_1467	2019-10-10	155	6362.77

1468 rows × 3 columns

Frequency와 Monetary의 기준은 고객별 평균 구매 수량과 금액을 참고하였습니다.

4. RFM Segmentation

```
1 RFM_recency = lambda x: 'recent' if x >= '2019-12-01' else 'past'
2 RFM_frequency = lambda x: 'high' if x >= 150 else 'low'
3 RFM_monetary = lambda x: 'high' if x >= 3000 else 'low'
4
5 RFM['최근구매일'] = RFM['최근구매일'].apply(RFM_recency)
6 RFM['수량'] = RFM['수량'].apply(RFM_frequency)
7 RFM['구매금액'] = RFM['구매금액'].apply(RFM_monetary)
```



람다 표현식을 사용하여 기준에 따라 컬럼 값을 변경해줍니다.

- Recency : 2020-01-01일 기준으로 한 달 이내에 구매 기록이 있으면 **recent** 이외는 **past**
- Frequency : 150회 이상 구매 시 **high** / 3회 미만 구매 시 **low**
- Monetary : 3000달러 이상 구매 시 **high** / 500달러 미만 구매 시 **low**

1	RFM		
최근구매일 수량 구매금액			
고객ID			
USER_0000	past	row	row
USER_0001	past	high	high
USER_0002	past	high	row
USER_0003	recent	row	row
USER_0004	past	row	row
...
USER_1463	past	row	row
USER_1464	past	row	row
USER_1465	past	row	row
USER_1466	past	row	row
USER_1467	past	high	high

1468 rows × 3 columns

4. RFM Segmentation

```
1 # 충성도가 높은 고객
2 RFM[(RFM['최근구매일'] == 'recent') & (RFM['수량'] == 'high') & (RFM['구매금액'] == 'high')]
```

고객ID	최근구매일	수량	구매금액
USER_0033	recent	high	high
USER_0057	recent	high	high
USER_0062	recent	high	high
USER_0076	recent	high	high
USER_0077	recent	high	high
...
USER_1391	recent	high	high
USER_1397	recent	high	high
USER_1411	recent	high	high
USER_1415	recent	high	high
USER_1450	recent	high	high

67 rows × 3 columns

한 달 이내에 제품을 구매했고,
구매 수량과 금액이 다른 고객들의 평균치보다 높은 **VIP 고객**입니다.

4. RFM Segmentation

```
1 # 떠나간 VIP
2 RFM[(RFM['최근구매일'] == 'past') & (RFM['수량'] == 'high') & (RFM['구매금액'] == 'high')]
```

고객ID	최근구매일	수량	구매금액
USER_0001	past	high	high
USER_0006	past	high	high
USER_0008	past	high	high
USER_0009	past	high	high
USER_0026	past	high	high
...
USER_1433	past	high	high
USER_1435	past	high	high
USER_1439	past	high	high
USER_1440	past	high	high
USER_1467	past	high	high

262 rows × 3 columns

구매 수량과 금액이 다른 고객들의 평균치보다 높았지만,
한 달 이상 구매 내역이 없는 **떠나간 VIP** 고객입니다.

고객 만족도 조사, 피드백 수집, 구매 이력 분석 등을 통한 이탈 원인 분석이 필요합니다.

4. RFM Segmentation

```
1 # 비활성고객, 잠재고객  
2 RFM[(RFM['최근구매일'] == 'past') & (RFM['수량'] == 'row') & (RFM['구매금액'] == 'row')]
```

최근구매일 수량 구매금액			
고객ID			
USER_0000	past	row	row
USER_0004	past	row	row
USER_0005	past	row	row
USER_0010	past	row	row
USER_0011	past	row	row
...
USER_1462	past	row	row
USER_1463	past	row	row
USER_1464	past	row	row
USER_1465	past	row	row
USER_1466	past	row	row

753 rows × 3 columns

아래와 같이 다양한 이유로 해석할 수 있습니다.

1. 서비스나 제품을 이용하지 않는 **비활성 고객**
2. 구매 횟수와 금액은 적지만 제품이나 서비스에 관심이 있는 **잠재고객**
3. 이전의 서비스나 제품에 **만족하지 못한 고객**

해당 고객들을 활성화 시키는 것이 기업의 매출을 늘리고, 고객을 확장하는데 큰 기여를 할 수 있습니다.
그러므로, 원인을 파악하여 해당 고객에게 맞는 적절한 전략을 수립해야 합니다.

5. K-Mean

고객ID	최근구매일	수량	구매금액
USER_0000	2019-09-15	3	30.99
USER_0001	2019-11-02	342	13834.90
USER_0002	2019-10-19	209	1442.12
USER_0003	2019-12-14	21	1360.07
USER_0004	2019-09-15	56	1442.47
...
USER_1463	2019-04-05	46	544.34
USER_1464	2019-10-05	141	2363.05
USER_1465	2019-06-20	10	101.56
USER_1466	2019-10-23	2	298.00
USER_1467	2019-10-10	155	6362.77

1468 rows x 3 columns



```
1 RFM['최근구매일'] = pd.to_datetime(RFM['최근구매일']) # 날짜 데이터를 object -> datetime으로 변환
2 RFM['경과일'] = (pd.to_datetime('2020-01-01') - RFM['최근구매일']).dt.days
3 RFM = RFM.drop(columns=['최근구매일'])
4 RFM
```

고객ID	수량	구매금액	경과일
USER_0000	3	30.99	108
USER_0001	342	13834.90	60
USER_0002	209	1442.12	74
USER_0003	21	1360.07	18
USER_0004	56	1442.47	108
...
USER_1463	46	544.34	271
USER_1464	141	2363.05	88
USER_1465	10	101.56	195
USER_1466	2	298.00	70
USER_1467	155	6362.77	83

1468 rows x 3 columns

최근 구매일로부터 2020-01-01일까지 얼마나 경과했는지 계산합니다.

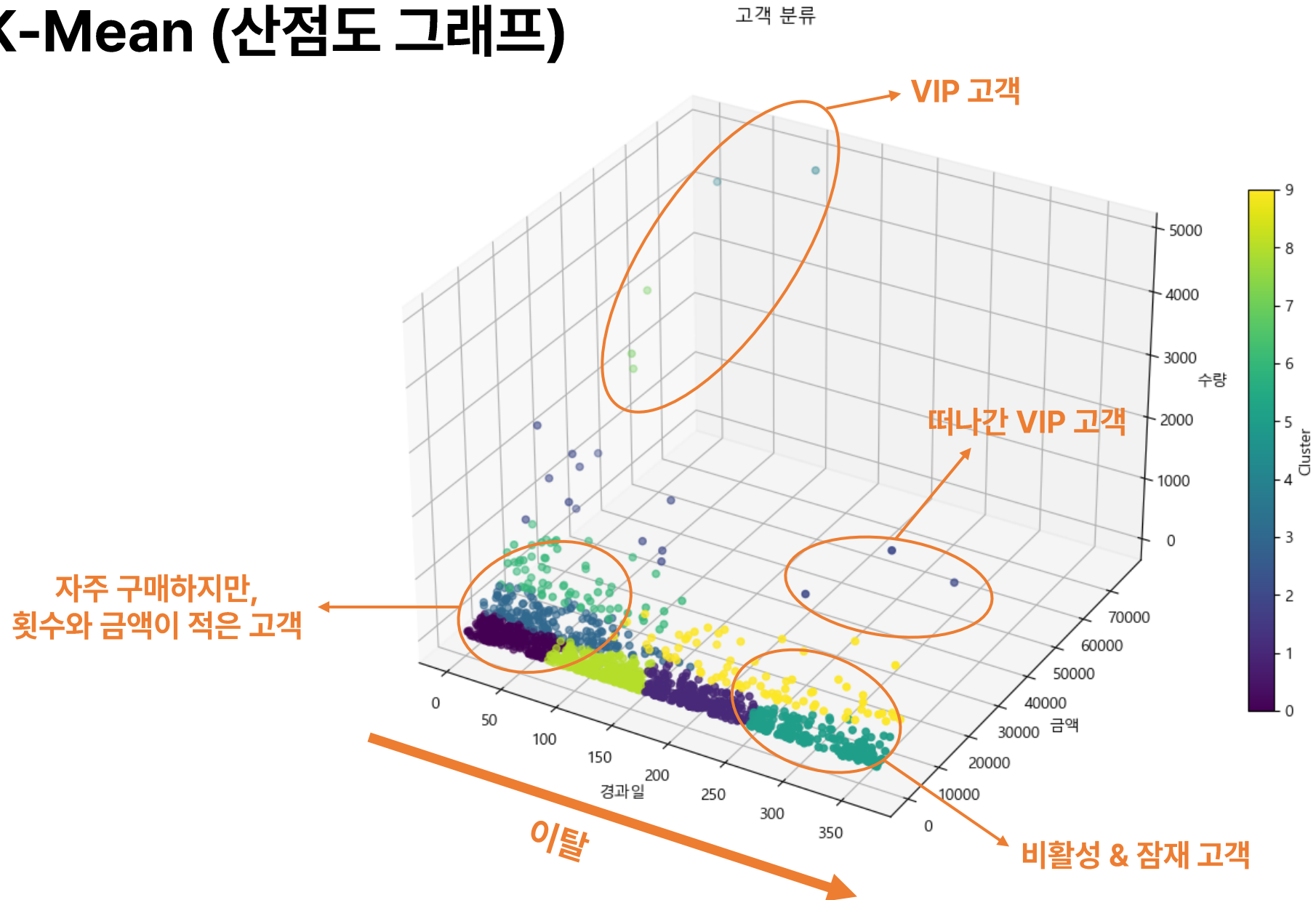
5. K-Mean

```
1 import pandas as pd
2 from sklearn.cluster import KMeans
3 from sklearn.preprocessing import StandardScaler
4 import matplotlib.pyplot as plt
5 from mpl_toolkits.mplot3d import Axes3D
6
7
8 # 데이터 정규화
9 scaler = StandardScaler()
10 scaled_RFM = scaler.fit_transform(RFM)
11
12 # K-means 클러스터링 수행
13 kmeans = KMeans(n_clusters=10)
14 kmeans.fit(scaled_RFM)
15 labels = kmeans.labels_
16
17 # 군집화 결과를 데이터프레임에 추가
18 RFM['Cluster'] = labels
19
20 # 3D 산점도 시각화
21 fig = plt.figure(figsize=(12, 12))
22 ax = fig.add_subplot(111, projection='3d')
23 scatter = ax.scatter(RFM['경과일'], RFM['구매금액'], RFM['수량'], c=RFM['Cluster'], cmap='viridis')
24 ax.set_xlabel('경과일')
25 ax.set_ylabel('금액')
26 ax.set_zlabel('수량')
27 ax.set_title('고객 분류')
28
29 # 컬러바 추가
30 cbar = plt.colorbar(scatter, shrink=0.5)
31 cbar.set_label('Cluster')
32 plt.show()
```

수량, 구매금액, 경과일 컬럼 값을 정규화하고,
해당 데이터를 바탕으로 10개의 Clusters로 군집화를 진행합니다.

이후 3D 산점도 그래프로 시각화 합니다.

5. K-Mean (산점도 그래프)



6. 마무리

많은 양의 데이터를 분석해본 것은 아니지만, 비활성 고객과 잠재 고객의 비중이 50%를 차지하는 것을 보고 기업이 재방문 고객을 위한 프로모션 이벤트를 왜 많이 유치하는지 알게 되었습니다.

고객군을 분류하기 위해 RFM 기준을 한 개씩만 세웠는데, 보다 세부적인 기준을 설정한다면 고객을 세분화함으로써, 각 고객에게 맞춤형 마케팅 전략을 수립할 수 있을 것이란 생각이 들었습니다.

마지막으로 RFM 분석 이외의 다양한 고객 세분화 기법(Cohort 분석, 구매 패턴 분석, 텍스트 마이닝)을 함께 활용하여 고객을 이해하고, 개인화된 마케팅 전략을 수립해보고 싶다는 생각이 들었습니다.

감사합니다 😊