고객 세분화 및 전략

L-point 식품 고객 분석

Team. OMEGA

팀원 소개

채효진 (팀장), 박태현

역할 분담을 나누지 않고 모든 과정을 함께 토론해 나가며 만들었습니다.

사용언어와 환경

사용언어

사용 도구

사용 라이브러리

목차

1. 배경

백화점의 꽃 '식품관' 의 변신과 중요성 부각

식음료(F&B) 매장이 고객 유입과 매출 상승에 핵심유인으로 떠오르면서
매장 구색, 맛집 유치 경쟁이 치열해지고 있음
이유는 **소비는 온라인으로 하고 밥만 먹고 가는 형태**로 백화점 고객들의
행동유형 패턴이 바뀌었기 때문

- 2021년 국내 백화점들의 '식품관' 투자 사례

  롯데 백화점 : 동탄상권 신규출점, 전체 영업면적의 **28%** (본점 : 19%)
  신세계 백화점 1층 영업점 탈바꿈 : 화장품관 → 식품관

< 사진1. 프리미엄 형태로 변하는 백화점의 식품관 모습 >

* L.point : 롯데멤버십으로, 약 3,600만원 회원을 갖춘 롯데그룹의 다양한 인프라와 경쟁력 있는 50여개 그룹/외부 제휴사를 결합한 통합 멤버십

1. 배경

백화점들이 앞 다투어 식품관 변화에 나선 이유는?

< 사진2. 온라인과 오프라인 매출 증감률 >

→ 코로나19로 인한 오프라인 매장 방문 감소로 소비자의 방문횟수가 줄자
'식품관 강화'를 선택

Why?

- 백화점 매출에서 식품이 차지하는 비중이 크고 젊은 세대의 외식비 지출이 늘고 있기 때문
- 온라인에서 할 수 없는 '체험' 제공, '특별한 맛집'으로 차별화 가능

▼
▼

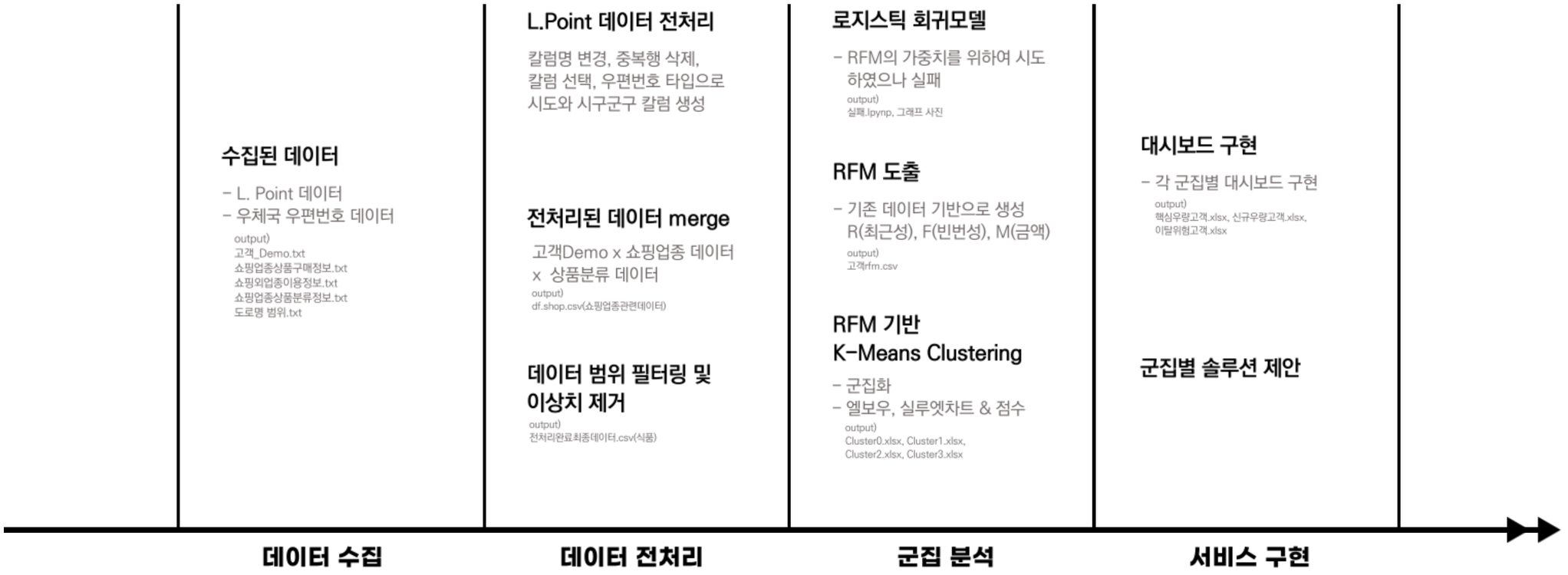
따라서 업계는 기존의 '평당 매출 극대화' 전략에서 벗어나 손님을 모으는데 집중하는 '집객 극대화' 전략에 초점을 맞추려는 경향

그러므로 자신들만의 프리미엄 식품관 브랜드 구축이 필요할 것

"식품관 브랜드 구축을 위한 고객 세분화 및 전략 제안 "

2. 분석 프로세스

L.Point 고객데이터를 활용하여 전처리, 군집분석을 한다.
최종 서비스인 시각적 자료 대시보드를 구현하여 분석 결과와 솔루션을 전달하는 형태의 프로세스이다.



3. 데이터 수집

데이터 출처

1. 제 4회 L.Point Big Data Competition dataset
2. 우체국 도로명 범위 데이터

< 원본 데이터 >

구분	데이터 항목	영문명	상세 설명
Demo.	ID	ID	고객의 고유 식별번호
	성별	GENDER	고객의 성별 (남성:1,여성:2)
	연령대	AGE_PRD	고객의 연령대
	거주지	HOM_PST_NO	거주지역 신우편번호 앞 3 자리 (서울 - 구 단위, 서울 외 지역 – 시/도 단위 변환)
쇼핑 업종 상품 구매 정보	ID	ID	고객의 고유 식별번호
	영수증번호	RCT_NO	구매 내역의 고유 식별번호
	업종	BIZ_UNIT	쇼핑 5 개 업종(A01/A02/.../A05) (A01:백화점, A02:대형마트, A03:슈퍼마켓, A04:편의점 A05:드러그스토어)
	상품 소분류 코드	PD_S_C	제휴사 상품분류정보
	점포코드	BR_C	구매가 발생한 점포 코드성 정보
	구매일자	DE_DT	구매가 발생한 일자 (YYYY/MM/DD)
	구매시간	DE_HR	구매가 발생한 시각
	구매금액	BUY_AM	구매한 금액
	구매수량	BUY_CT	구매한 수량
쇼핑 외 업종 이용정보	ID	ID	고객의 고유 식별번호
	업종	BIZ_UNIT	9 개 업종명 (B01:호텔, B02:여행사, B03:면세점, C01:영화관, C02:테마파크, C03:야구관람, D01:패스트푸드, D02:패밀리레스토랑, D03:카페)
	이용월	CRYM	이용이 발생한 월 (YYYY/MM)
	이용금액	U_AM	이용한 금액
	이용건수	U_CT	이용한 건수
쇼핑 업종 상품 분류 정보	업종	BIZ_UNIT	쇼핑 5 개 업종(A01/A02/.../A05)
	상품 소분류 코드	PD_S_C	상품 소분류 카테고리 코드성 정보
	소분류명	PD_S_NM	상품 소분류 카테고리 한글명
	중분류명	PD_M_NM	상품 중분류 카테고리 한글명
	대분류명	PD_H_NM	상품 대분류 카테고리 한글명

+ 외부 데이터 (우체국 도로명주소 정보)
& 필요한 정보만 필터링

< 최종 데이터의 칼럼 설명 >

데이터 칼럼	상세 설명
ID	고객의 고유 식별번호
성별	고객의 성별
연령대	고객의 연령대
시도	광역 자치 단체 (서울특별시)
시군구	서울특별시 자치구역 행정구역
구매일자	구매가 발생한 일자 (YYYY/MM/DD)
구매시간	구매가 발생한 시각
구매금액	구매한 금액
구매수량	구매한 수량
업종	백화점과 드러그스토어
소분류명	상품 소분류 카테고리 한글명
중분류명	상품 중분류 카테고리 한글명
대분류명	상품 대분류 카테고리 한글명

4. 데이터 전처리

1. L-point 데이터 전처리 : 컬럼명 변경, 중복행 삭제, 우편번호를 통해 시도/시군구 컬럼 생성
2. 1차 전처리된 데이터 merge : 고객 Demo 테이블, 쇼핑업종데이터 테이블, 상품분류 데이터 테이블을 merge 후 필요없는 컬럼 삭제

3-1. 이상치 제거 : '구매금액 '

사분위수를 활용

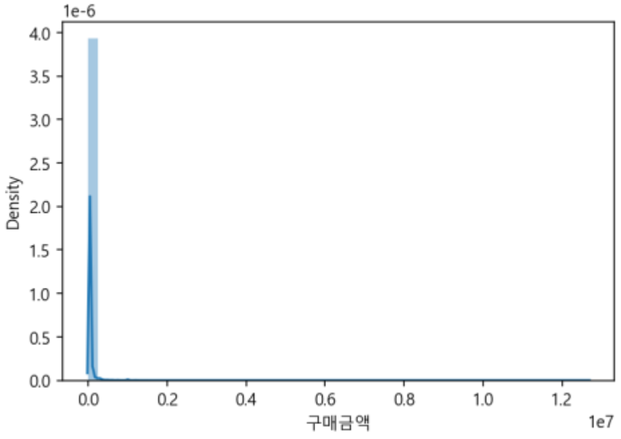
→ 0.25, 0.75로 범위를 선택 했을 때, 약 50% 정도의 너무 많은 데이터가 잘렸기 때문에 최소한의 제거를 위해 0.05, 0.95 선택

```
Q1 = df['구매금액'].quantile(0.05)
Q3 = df['구매금액'].quantile(0.95)
IQR = Q3 - Q1
```

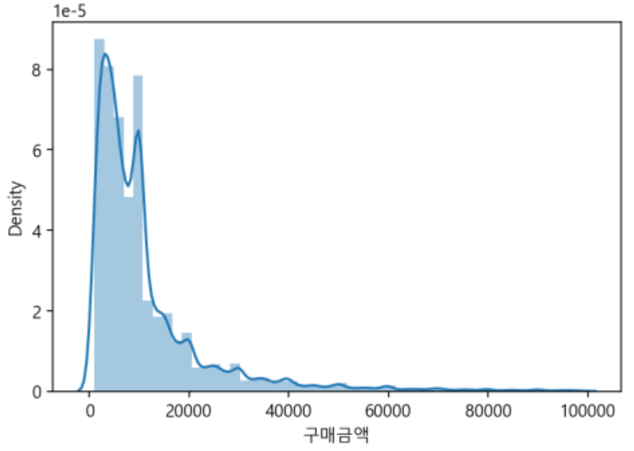
최대값의 경우 비싼 홍삼세트 같은 제품이 있었기에 범위를 Q3 + 1.5*IQR로 넓게 지정

```
df = df[(df['구매금액'] >= Q1) &
(df['구매금액'] <= Q3 + 1.5*IQR)]
```

< 이상치 제거 전 >



< 이상치 제거 후 >

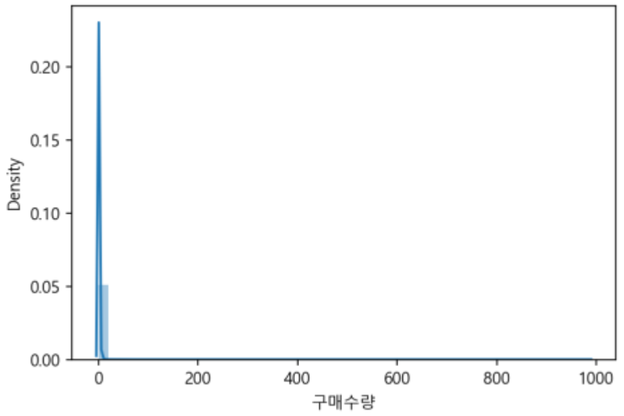


4. 데이터 전처리

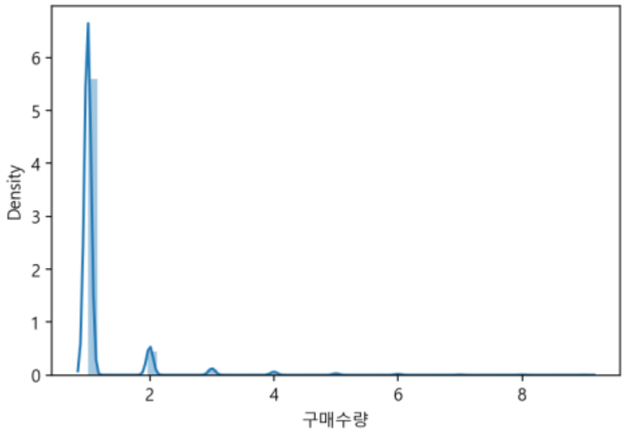
3-2. 이상치 제거 : '구매수량 '

describe를 활용하여 100→50→10개로 데이터의 제한을 임의로 두면서 그래프로 확인해나감

< 이상치 제거 전 >



< 이상치 제거 후 >



4. 데이터 범위 및 필터링

대부분 고객들의 거주지가 서울특별시(약 50%)였고 데이터의 양을 고려하여 범위를 선택하였음.

→ 시도 : 서울특별시, 대분류명 : 식품

5. 군집분석 : 고객RFM 도출

- 전처리최종완료데이터를 가지고 rfm_df 생성

1. R, F, M 변수 생성

R = (데이터셋의 가장 최근 날짜 – 구매일자)의 최솟값
F = ID별 구매일자의 nunique
M = ID별 총쇼핑금액 sum (총쇼핑금액 = 구매금액 * 구매수량)

2. RFM DF의 이상치 탐색 및 제거

이상치 제거를 위해 일반적으로 사용하는 사분위수 기준으로 Q1, Q3의 범위 지정

```
Q1 = rfm.Monetary.quantile(0.25)
Q3 = rfm.Monetary.quantile(0.75)
IQR = Q3 - Q1
```

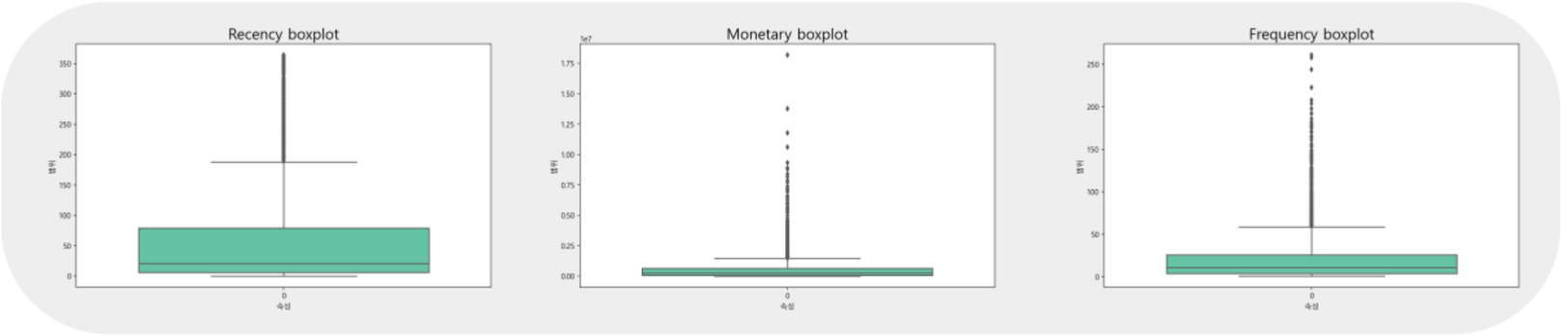
해당하는 값의 범위만 rfm에 다시 할당

```
rfm = rfm[rfm.Montary <= Q3+1.5*IQR
```

최솟값은 문제없으므로 최댓값만 범위 다시 지정

-> Monetary의 이상치 처리 결정

< RFM Boxplot >



3. 고객rfm CSV 저장 이상치 처리 후 '고객rfm'를 csv로 저장

	CustomerID	Monetary	Frequency	Recency	cluster
0	1	418540	13	15	3
1	2	41000	3	352	1
2	6	78500	1	120	1
3	7	1108200	24	16	2
4	10	283800	7	9	3

5. 군집분석 : 알고리즘 선택 배경

목적 : 흩어져 있는 고객 데이터를 마케팅에서 보편적으로 쓰이는 방식인 RFM 기준으로 군집을 시키고자 함
* RFM: R(최신성), F(구매빈도), M(매출액)

Why?

R,F,M의 중요도가 다르기 때문에
가중치를 부여하기 위해 시도하였음.

How?

1) 1년간의 고객 구매데이터를
8개월과 4개월 데이터로 분할.

2) 8개월 데이터의 RFM 값을 관측하고
4개월 데이터는 해당기간 동안의 구매여부를
1(구매함)과 0(구매하지 않음)으로 이분.

3) 모델을 통해 구해진 회귀계수를 가중치로 선택.

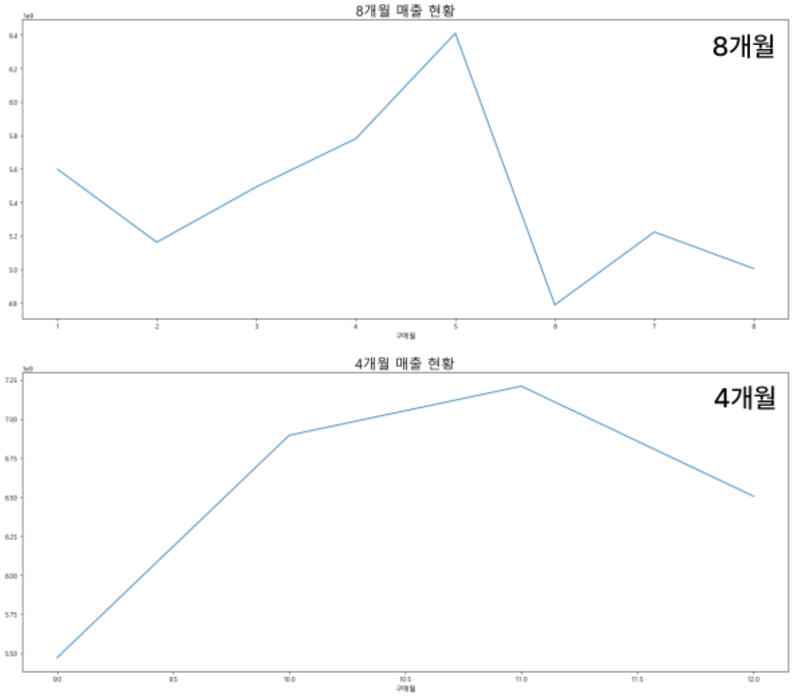
로지스틱 회귀모델 VS 군집분석

- K-MEANS Clustering을 통해
모델이 직접 비슷한 데이터들을 가진
군집별로 나눠줄 수 있게 하고자함.

로지스틱 회귀모델의 한계점은 1년 데이터의 8개월과 4개월의 계절성 등에 따른 비동질성으로 인해 모델의 성능이 좋지 못하였음
따라서, 로지스틱 회귀분석 모델을 포기하고 다른 알고리즘을 선택하고자 하였음

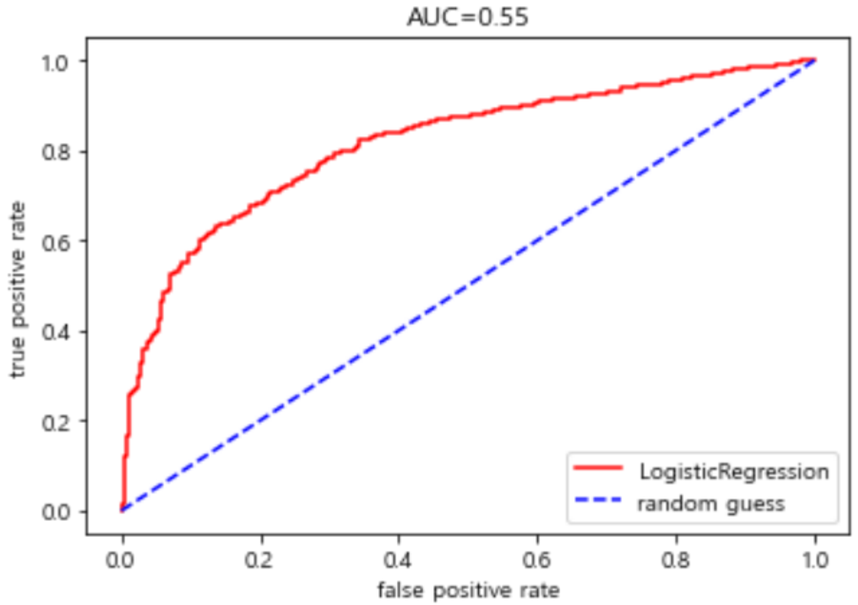
5. 군집분석 : 로지스틱 회귀모델의 한계점

- 데이터의 비동질성에 따른 모델 성능 저하



앞의 8개월 데이터와 뒤의 4개월 데이터가 계절, 시기에 따른 편향으로 차이가 있어 모델이 제대로 예측하지 못함

< 모델 성능지표 : AUC_SCORE >

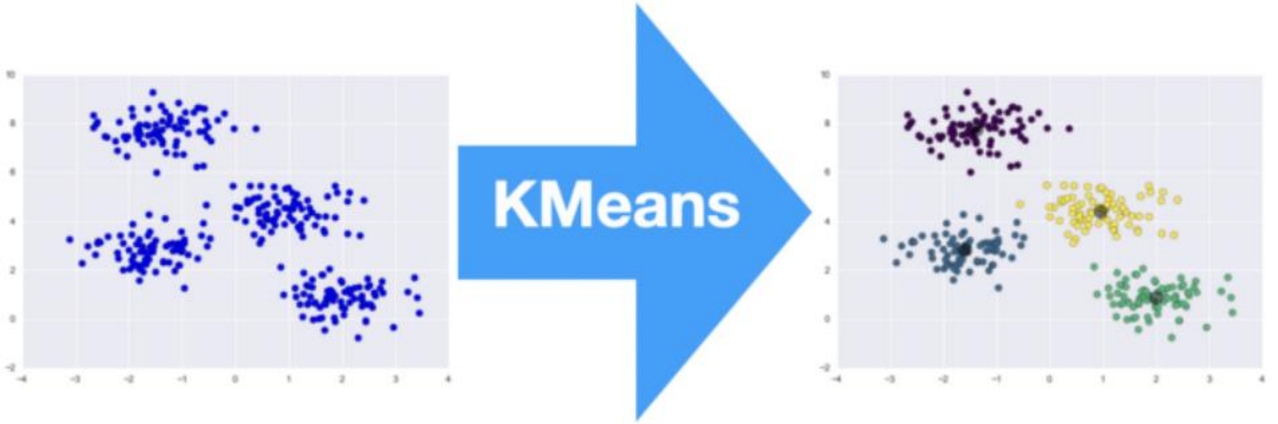


GridSearchCV로 하이퍼 파라미터 튜닝까지 진행했으나 모델의 성능이 좋지 못함

5. 군집분석 : 최종 알고리즘 선택 이유와 활용

K-Means Clustering

- 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화 하는 방식으로 작동



<사진 3. K-Means 시각화 >

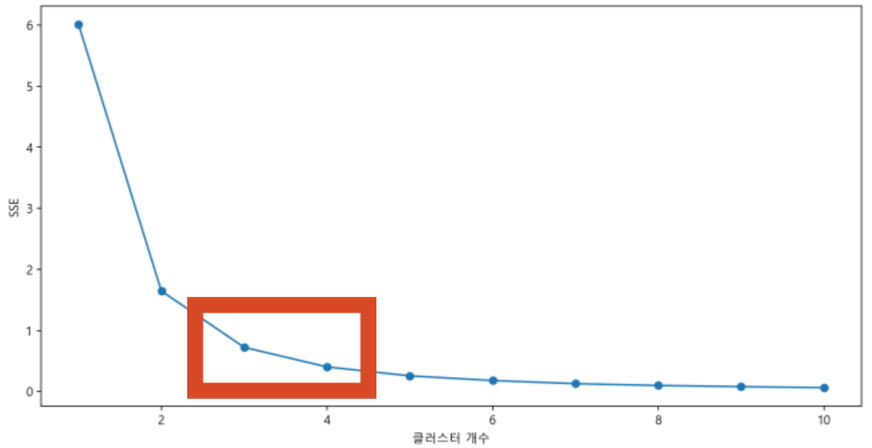
- **사전 정의된 범주가 없는 데이터에서 최적의 그룹을 찾아나가기 위해** 비지도학습의 **K-Means** 알고리즘 사용
- 데이터의 R,F,M 값을 바탕으로, 최적의 클러스터링을 하여 도출된 클러스터 값을 원본 데이터에 부여

5. 군집분석 : RFM기반 K-Means Clustering

파라미터 값 찾기 - n_cluster

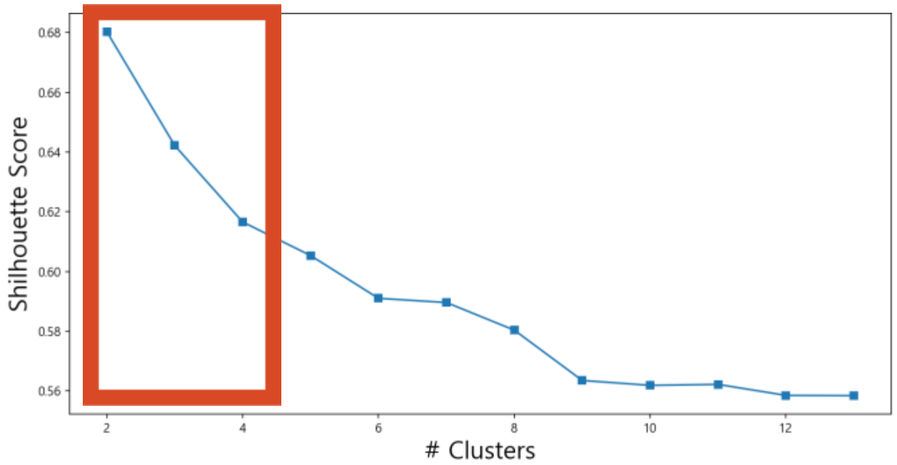
K-Means 모델은 n_cluster 지정값에 영향을 받기에 **가장 이상적인 n_cluster 값**을 구하는 것이 군집에 도움이 됨

< 엘보우 차트 >



'3' 이후로 평평해지는 모습을 보임. n_cluster는 3이 적절하게 보이지만,
정확한 클러스터 값을 찾기 위해 실루엣 점수를 확인했음

< 실루엣 점수 >



클러스터 2, 3, 4 값의 평균 실루엣 점수가 높아보임

5. 군집분석 : RFM기반 K-Means Clustering

파라미터 값 찾기 - n_cluster

군집별 평균 실루엣 계수의 평균값이 차이가 많이 나지 않는 군집이 **이상적인 n_cluster**

✔ rfm df의 모든 개별 데이터에 실루엣 계수값을 계산해 컬럼으로 추가한 후, 각 군집별 실루엣 계수들의 평균을 계산 후, 분산으로 비교

엘보우 차트에서 이상적이게 보였던 n_cluster = 3의 군집별 실루엣 계수 평균 결과

cluster0 : 0.502 // cluster1 : 0.557 // cluster2 : 0.724

그러나, 나머지 평균 실루엣 점수가 높았던 n_cluster=2,3,4의

평균 실루엣 계수들을 구하고 **분산이 가장 적은 클러스터를 n_cluster 값으로 선택**하고자 함

(군집별 값의 차이가 많이 나지 않아야 좋은 군집이기 때문.)

5. 군집분석 : RFM기반 K-Means Clustering

파라미터 값 찾기 - n_cluster

군집별 평균 실루엣 계수의 평균값이 차이가 많이 나지 않는 군집이 **이상적인 n_cluster**

 # n_cluster를 2개부터 4개까지 지정시, 군집별 평균 실루엣 계수

In [13]:
```
# n_cluster를 2개부터 4개까지 지정시, 군집별 평균 실루엣 계수

model = []
coff_list = []
for p in range(2,5):

    a = KMeans(
        n_clusters = p,
        init = 'k-means++',
        n_init = 15,
        random_state = 300)
    model.append(a)
for number in range(0,3):

    rfm_df['cluster'] = model[number].fit_predict(rfm_df)
    score_samples = silhouette_samples(rfm, rfm_df['cluster'])
    rfm_df['silhouette_coeff'] = score_samples
    k = rfm_df.groupby('cluster')['silhouette_coeff'].mean().values
    coff_list.append(k)

coff_list
```

Out[13]:
```
[array([0.73572317, 0.51702027]),
 array([0.72346863, 0.55676959, 0.50256477]),
 array([0.5172892 , 0.70783614, 0.56387742, 0.51015941])]
```

 # DataFrame으로 변형

	cluster_2	cluster_3	cluster_4
0	0.736	0.723	0.517
1	0.517	0.557	0.708
2	nan	0.503	0.564
3	nan	nan	0.510

 # 군집별 평균 실루엣 계수들의 분산

cluster_2 0.024
cluster_3 0.013
cluster_4 0.008

 # 최종적으로 분산값이 가장 작은 cluster_4로 결정. 즉, n_cluster값은 4

5. 군집분석 : RFM기반 K-Means Clustering

이상적인 파라미터

→ n_cluster =4 , init = 'k-means++', random_state =300

< 모델 훈련 후 학습 >

```
model = KMeans(
    n_clusters = 4,
    init = 'k-means++',
    n_init = 15,
    random_state = 300)

# 학습
rfm['cluster'] = model.fit_predict(rfm_val)
```

< 나누어진 군집들을 기존 데이터 프레임에 추가 >

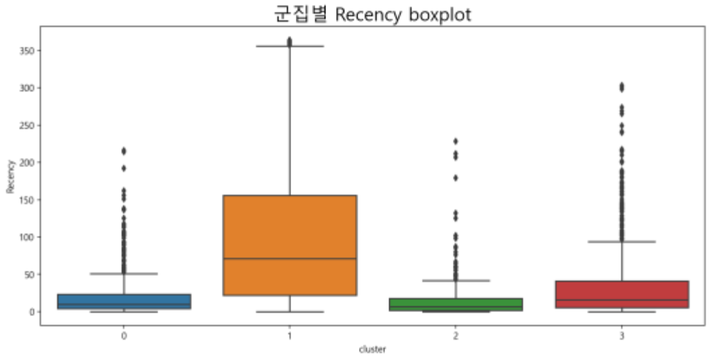
```
rfm.head()
```

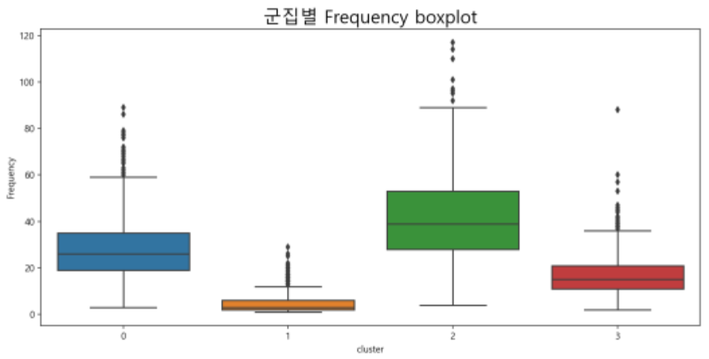
	CustomerID	Monetary	Frequency	Recency	cluster
0	1	418540	13	15	3
1	2	41000	3	352	1
2	6	78500	1	120	1
3	7	1108200	24	16	2
4	10	283800	7	9	3

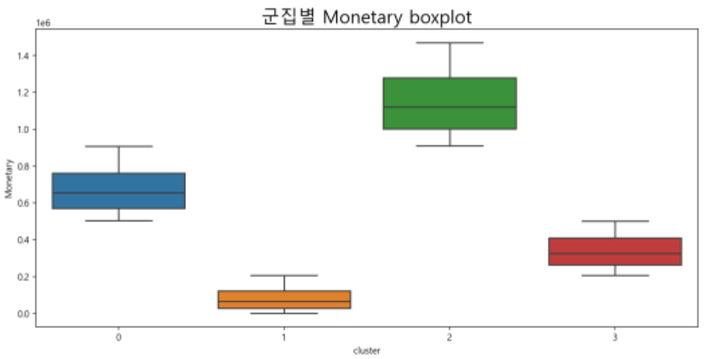
5. 군집분석 : RFM기반 K-Means Clustering

군집된 클러스터별 RFM Boxplot

cluster0, cluster1, cluster2, cluster3의 RFM 살펴보기







< RFM Boxplot으로 도출된 결과 >

높음 보통 낮음

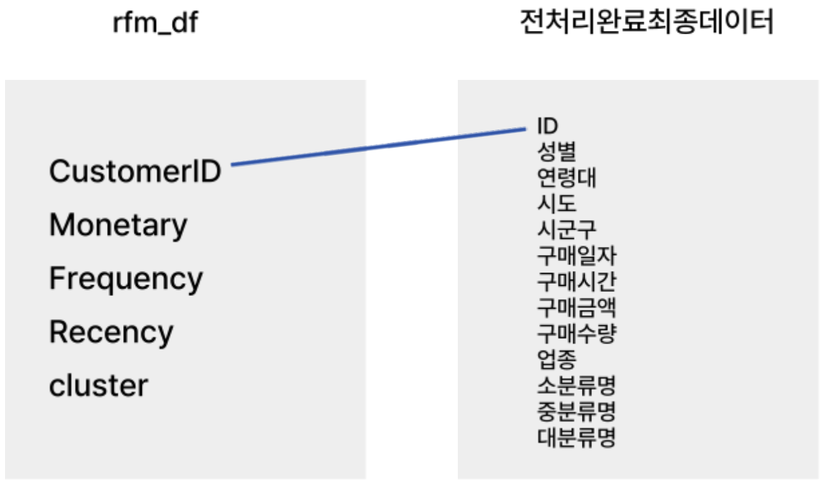
# cluster0 : 수익 ↑	구매빈도 –	최신성 ↑	»	(신규 우량 고객)
# cluster1 : 수익 ↓	구매빈도 ↓	최신성 ↓		(저 수익성 고객)
# cluster2 : 수익 ↑	구매빈도 ↑	최신성 ↑		(핵심 우량 고객)
# cluster3 : 수익 –	구매빈도 ↓	최신성 ↓		(이탈 위험 고객)

5. 군집분석 : RFM기반 K-Means Clustering

클러스터 활용을 위해 전처리완료최종데이터에 클러스터 값을 부여

cluster0, cluster1, cluster2, cluster3의 RFM 살펴보기

클러스터 값이 부여된 rfm df와 전처리완료최종데이터를 merge



필요없는 컬럼 제거 후, (CustomerID, Monetary, Frequency, Rency)
클러스터별로 데이터를 나누어 xlsx 파일로 저장

- 데이터 분할

```
cluster0 = complete[complete['cluster'] == 0]
cluster2 = complete[complete['cluster'] == 2]
cluster3 = complete[complete['cluster'] == 3]
```

- 데이터 저장

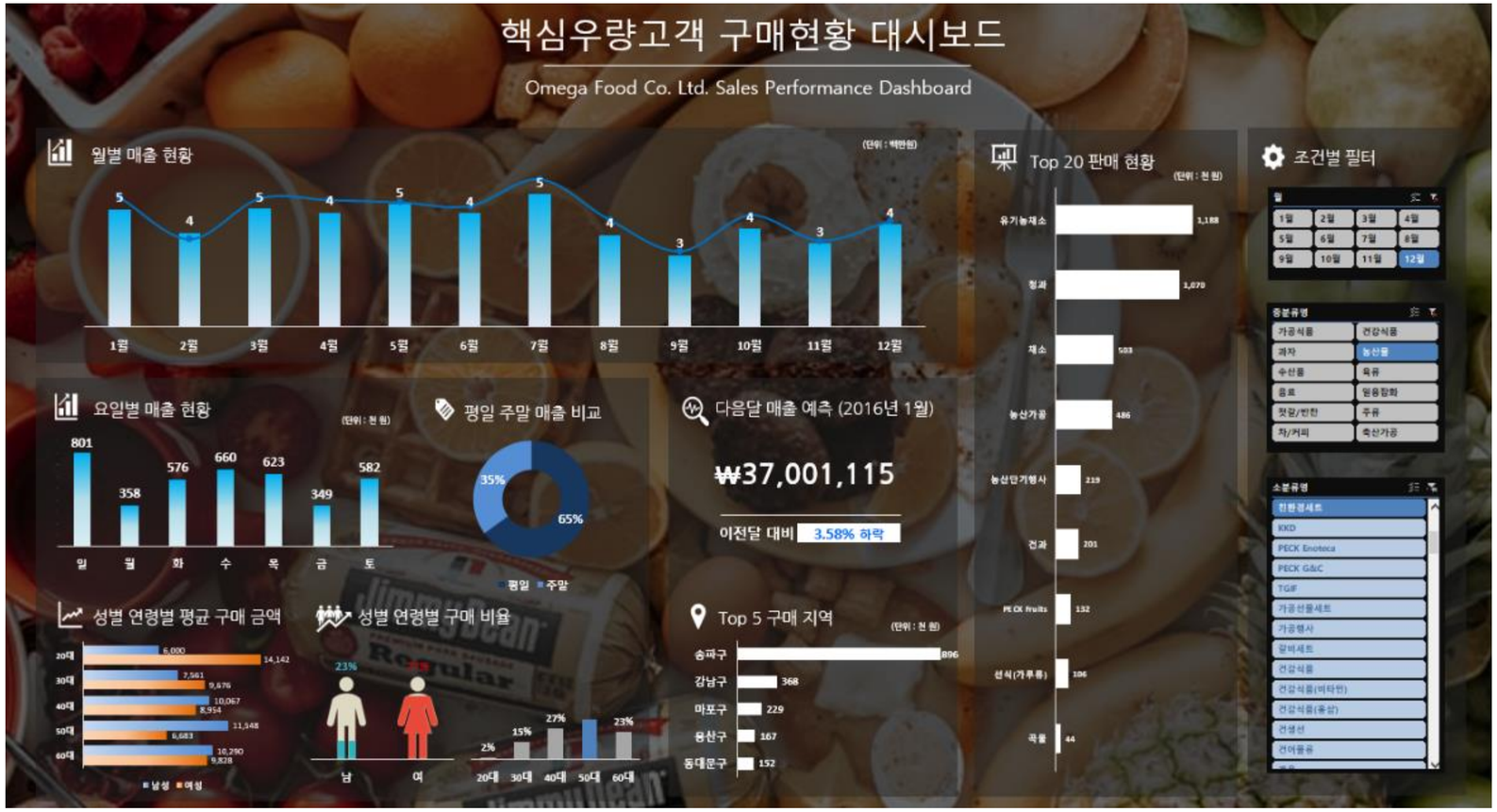
```
cluster0.to_excel('cluster0.xlsx', index=False)
cluster2.to_excel('cluster2.xlsx', index=False)
cluster3.to_excel('cluster3.xlsx', index=False)
```

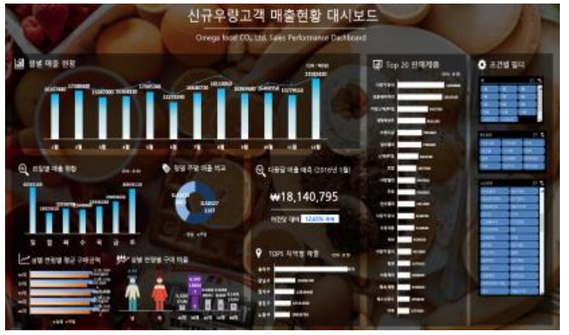
저장된 군집별 엑셀 데이터를 활용하여 대시보드 구현

6. 대시보드 구현

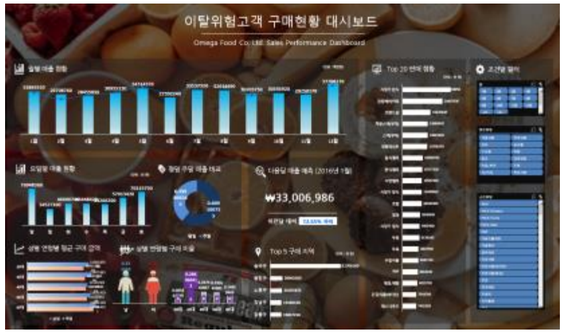
활용한 툴 : Microsoft Office Excel

핵심우량고객





신규우량고객



이탈위험고객

7. 솔루션 제안

대시보드와 파이썬의 EDA를 바탕으로 분석된 내용을 마케팅팀에게 제안하는 솔루션

핵심우량고객

- 프리미엄 회원제 실시(연회비)
- 전문 베이커리, 청과 같은 식생필품의 프라임 상품 할인 혜택 부여

신규우량고객

- 자사의 핵심 우량고객으로 양성하기 위해 신규우량고객 집단이 많이 산 상품과 연관된 상품을 가까이 배치 (추가분석)

이탈위험고객

- 최신성을 높이기 위해 마케팅 메시지를 푸쉬 알람으로 전달
- 당일 사용 가능한 쿠폰을 전달하는 마케팅

아쉬운점 및 추가제안

- 데이터의 한계

n년치의 데이터가 아닌 1년치의 데이터이기 때문에 시계열적인 분석이 불가능
더 많은 데이터가 있다면 로지스틱 회귀 모델을 사용하여 가중치를 부여해 더 정확한 RFM Score를 도출할 수 있었을것

- 추가제안

신규우량고객을 핵심우량고객으로 양성하기 위해 상품 연관분석 진행, 식품관의 자리 배치에 이용
재구매고객을 조사한 후, 자주 일어나는 품목의 수치(재구매율)를 구현해 저수익성 고객을 탈바꿈화

출처

사진1. 백화점 사진 : http://mbiz.heraldcorp.com/view.php?ud=20140611000095
사진2. 온라인쇼핑몰 사진 : 한국경제신문
사진3. K-Means: https://ichi.pro/ko/kmeans-keulleoseuteoling-algolijeum-87061054724903
우편번호 데이터 출처
인터넷 우체국 - 우편번호 찾기 - 우편번호 내려받기
https://www.epost.go.kr/main.retrieveMainPage.comm

배경부분 자료 출처 (기사)
http://www.thinkfood.co.kr/news/articleView.html?idxno=92068
http://www.thescoop.co.kr/news/articleView.html?idxno=38006

로지스틱 회귀 참고자료출처
논문 김동석, 2021, <RFM 모형의 가중치 선택에 관한 연구>, 17p https://oak.jejunu.ac.kr/handle/2020.oak/23663
유튜브 <통계데이터분석 - 일반선형모델 - 이항 로지스틱회귀분석> https://www.youtube.com/watch?v=nyU96C2-LCI
블로그
https://fish-tank.tistory.com/88
https://ysyblog.tistory.com/178

KMeans 참고자료출처
위키백과 'K-평균 알고리즘' https://ko.wikipedia.org/wiki/K-평균_알고리즘
그림출처: https://ichi.pro/ko/kmeans-keulleoseuteoling-algolijeum-87061054724903
유튜브 김성범[소장/인공지능공학연구소] 채널 https://www.youtube.com/channel/UCueLU1pCvFlM8Y8sth7a6RQ
블로그 https://ariz1623.tistory.com/224
사이킷런 공식 문서 https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
Kaggle 노트북 https://www.kaggle.com/hellbuoy/online-retail-k-means-hierarchical-clustering

엑셀 대시보드 참고자료출처
유튜브 오빠두엑셀 채널 https://www.youtube.com/channel/UCZ6UHYBQFBe14WUgxlgmYfg

Q&A