# 데이터사이언스 프로그래밍 Term Project

| | | |
|---|---|---|
| 과제명 | 국문 | 마케팅 최적화를 위한 고객 군집화 분석 (H&M) |
| | 영문 | Customer Clustering for Marketing Optimization (H&M case) |


□ Team #1
- ○ (팀장) 김건탁, 12201830, 컴퓨터공학과
- ○ (팀원) 김민규, 12182428, 아태물류학부
- ○ 김나현, 12214212, 데이터사이언스학과
- ○ 이주연, 12214243, 데이터사이언스학과



## Abstract

본 분석은 2022년 2월 8일부터 2022년 5월 10일까지 약 3개월간 진행된 Kaggle challenge에서 제공한 H&M 패션 데이터셋(10만개의 상품, 137만명의 고객, 3178만 건의 거래 데이터로 구성)을 바탕으로 진행되었다.[1]

제공된 거래 데이터를 통해 2018년 9월부터 2020년 9월까지의 기간 동안 H&M의 거래건수의 하락세를 확인할 수 있었다.

본 분석에서는 H&M의 부진한 거래건수 문제의 원인이 코로나19로 인한 고객의 비대면 선호로 보았다. 군집화 기법을 활용하여 군집별로 고객의 특성을 파악하고 맞춤형 온라인 서비스를 제공함으로써 문제를 해결하고자 하였다.

먼저, 제공된 고객 데이터와 거래 데이터를 탐색적 데이터 분석과 데이터 정제 과정을 거쳤다. 그 후, 거래데이터에서 유의미한 고객 특성을 추출하여 고객 데이터와 합치고 범주형 변수를 군집화 분석에 활용하기 위해 one hot encoding을 통해 범주형 데이터를 이진더미변수로 변환하였다. 마지막으로 변수들의 단위 효과를 없애기 위해 정규화를 실시했다.

군집화는 나이, 거래횟수, 회원가입여부, 패션소식을 받는 빈도, 선호하는 유통채널 등 고객의 특성들을 기준으로 진행하였다.

군집화 기법으로 'Gaussian Mixture Model(GMM)'과 'K-means'를 사용하였는데 군집의 개수를 7로 설정하였을 때 가장 의미있는 결과 해석이 가능하였으며, 두 군집화 기법의 군집 결과가 수치형 변수에서는 유사하게 나타났다. 하지만 'K-means'의 경우 '패션 소식을 받는 빈도 (fashion_news_frequency)' 변수를 명확하게 구분하지 못하였기 때문에 해석이 더 용이한 GMM을 최종 선택하였다. 그리고 군집별 고객 특성을 다음과 같이 해석하였다.

군집 0 : 오프라인 채널을 선호하는 젊은 고객군
군집 1 : 평균연령(57세)이 높은 고객군
군집 2 : H&M 목표 고객군
군집 3 : 온라인 채널을 선호하는 젊은 고객군
군집 4 : 아직 회원가입 하기 전인 잠재고객군
군집 5 : 회원 탈퇴한 고객군
군집 6 : 패션에 관심이 많은 패션 애호 고객군

각 군집별로 맞춤형 온라인 서비스를 간단히 제


[1] https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/overview

시하며 분석을 끝마쳤다.

## 1. Introduction

탐색적 데이터 분석 과정에서 2018년 9월부터 2020년 9월까지의 기간 동안 H&M의 거래건수의 추세를 확인한 결과 기울기 -225.0, y절편 1274233.7의 약한 하락세를 확인할 수 있었다.
2020년에 세계적으로 코로나19 바이러스가 확산되었다는 점에서 미루어 보아 H&M의 거래건수 하락세 또한 코로나19로 인해 고객의 매장 방문이 감소한 것이 원인으로 해석된다. 우측 파이 그래프는 2018년부터 2020년까지 H&M에서 발생한 거래의 유통채널 비중이다. 오프라인 채널이 약 70.4%으로 거래의 대부분이 오프라인 채널에서 발생하고 있음을 알 수 있다. H&M은 코로나19로 인한 고객들의 비대면 선호 트렌드에 맞춰 기존 오프라인 채널 위주의 운영에서 온라인으로 채널 다각화를 시도해야 한다. 본 분석에서는 H&M이 온라인 채널을 통해 고객에게 효율적이고 효과적인 마케팅을 할 수 있도록 고객 군집화를 수행하여 각 군집의 특징 파악하고 고객군별로 맞춤형 온라인 서비스를 제공하고자 한다.

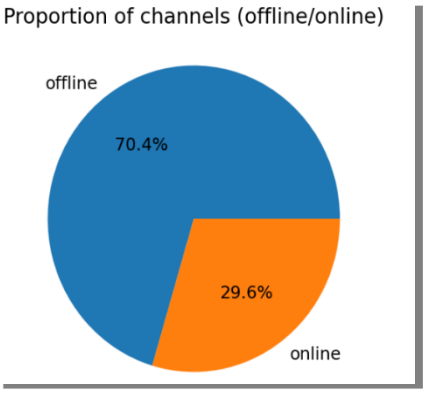


**그림 1 Proportion of channels**

## 2. Methods

### 2-1. 고객 데이터 EDA와 데이터 정제

다음과 같은 이유로 직관적인 해석이 어려운 고객 데이터의 일부 변수를 열 삭제 처리하였다.

- 'FN', 'Active' 변수는 무엇을 의미하는지 확인이 되지 않는다. (데이터셋에 설명이 없음)
- 'postal_code' 변수는 암호화된 우편주소로 분석하기 어려우며 해석이 어렵다.
-

| 데이터 수 | 1,371,980 (명) | 타 입 | 결측치 | 범 주 |
|---|---|---|---|---|
| 변수 | customer_id | object | 0 | 1,371,980명 |
| | club_member_status | object | 6062 | 'ACTIVE', 'PRE-CREATE', 'LEFT CLUB' |
| | fashion_news_frequency | object | 16009 | 'NONE', 'Regularly', 'None', 'Monthly' |
| | age | float64 | 15861 | (MIN, Q1, MEAN, Q3, MAX): (16, 24, 32, 49, 99) |

**표 1 고객 데이터 information**

먼저, 결측치가 포함된 데이터를 행 삭제 처리하였다. 결측치를 포함하는 행 16,009개는 137만명의 고객 데이터 중 1.17%밖에 차지하지 않으므로 삭제의 영향이 미미하다고 판단하였다.
'club_member_status'는 H&M 회원 상태에 대한 변수로 'ACTIVE', 'PRE-CREATE', 'LEFT CLUB' 3가지 범주를 가지고 있다. 각각 '회원', '아직 회원 가입하지 않은 고객', '탈퇴한 회원'을 의미한다. 회원가입 이전인 고객은 잠재고객으로 분류할 수 있어 군집화 및 그 해석에 있어서 중요한 특성일 수 있다.
'fashion_news_frequency'는 패션 뉴스를 얼마나 자주 접하는지에 대한 변수로 'NONE', 'Regularly', 'None', 'Monthly'로 4가지 범주를 가지고 있다. 중복된 범주요소 'NONE'과 'None'의 도수를 확인한 결과 단 2개의 데이터만 'None'을 가지고 있었으므로 해당 데이터를 행 삭제 처리하였다.
'Age'는 고객의 연령에 대한 연속형 변수로, 젊은층을 주 목표 고객으로 한다는 H&M의 알려진 정보와는 다르게 연령분포가 넓게 나타났다. 이상치 탐지를 통해 86.5세 이상의 208명의 데이터는 분류 정확도를 높이기 위해 행 삭제 처리하였다.
마지막으로 H&M의 목표 고객은 20~45세 사이의 젊은 여성이란 점을 고려하여 제 3사분위수(49세)를 기준으로 두 집단(Target customer, Non-target customer)으로 분류하여 새로운 변수 'age_group'

으로 열 추가하였다.[2]

### 2-2. 거래 데이터 EDA

| 데이터 수 | 31,788,324 (건) | 타 입 | 결측치 | 범 주 |
|---|---|---|---|---|
| 변수 | t_dat | object | 0 | 2018년 09월 20일부터 2020년 09월 22일까지 |
| | customer_id | object | 0 | 1,362,281 명의 고객 id |
| | article_id | int64 | 0 | 104,547 개의 상품 id |
| | price | float64 | 0 | (MIN, MEAN, MAX): (0.00017, 0.027829, 0.591525) |
| | sales_channel_id | int64 | 0 | 1 : online / 2 : offline |

**표 2 거래 데이터 information**

거래 데이터에는 어느 날짜에 어느 고객이 어떤 상품을 얼마를 주고 어떤 채널을 이용하여 구매하였다는 거래 내역이 저장되어 있다.

't_dat' 은 거래날짜에 대한 변수로 데이터 타입을 datetime64로 변경하고 이를 년, 월, 일, 요일로 나누어 새로운 4개의 날짜변수를 생성하였다. 날짜변수를 탐색하는 과정에서 요일별 거래건수에 아래와 같은 독특한 인사이트를 발견할 수 있었다.

거래건수는 일요일/월요일에 가장 적고 토요일/목요일에 가장 많았다. 목요일에 거래건수가 두드러진다는 것은 의외의 결과로 판단하여 추가적인 분석을 수행하였다. 거래량과 유통채널을 기준으로 groupby하여 count한 결과, 거래건수가 가장 적은 일요일과 월요일은 오프라인 대비 온라인 비중이 각각 27.6%, 30.0%인데 반해, 거래건수가 가장 많은 토요일과 목요일은 각각 65.0%, 45.5%로 거래건수가 많은 요일일수록 온라인 비중이 더 큰 경향이 있었다. 이는 토요일과 목요일에 온라인 이벤트나 할인행사를 하면 보다 많은 고객에게 접근할 수 있음을 시사한다.

'sales_channel_id' 는 유통채널에 대한 변수로 1이면 'online' 이고 2면 'offline' 인 이진변수다. 앞서 Introduction에서 언급한 것처럼 전체 거래 중 offline에서 발생한 거래의 비중이 70.4%로 H&M은 오프라인 채널을 주 유통채널로 하고 있다는 것을 알 수 있었다.

### 2-3. 군집화 분석 준비

거래데이터에서 value_counts()를 통해 '고객별 거래횟수(#Transactions)' 를 추출하고 이것과 '고객별 이용하는 유통채널(sales_channel_id)' 을 고객데이터와 합치는 작업을 수행하였다. '고객별 선호하는 유통채널' 은 오프라인 고객은 오프라인에서만 구매하고, 온라인 고객은 온라인에서만 구매한다는 강한 가정을 세워 각 고객이 첫번째 거래에서 이용한 유통채널을 선호하는 유통채널로 파악하였다. 이는 3179만건에 육박하는 거래 데이터의 분석시간을 줄이고 군집 해석이 너무 복잡해지는 것을 피하기 위한 것이다.

다음으로 범주형 변수들을 군집화 분석에 활용하기 위해서 One Hot Encoding을 통해 이진 더미변수들로 변환하였다. 그리고 단위의 영향을 없애기 위해 StandardScaler을 통해 정규화(z-score)시켰다.

| 데이터 수 | 1,328,779 (명) | 타 입 |
|---|---|---|
| 9개 변수 | age, #Transactions, ACTIVE, LEFTCLUB, PRE-CREATE, Monthly, None, Regularly, online, offline | float64 |

**표 3 군집화 분석용 데이터 information**

[2] H&M의 주 목표 고객은 20~34세 사이의 여성 그리고 그 다음으로 35~45세 사이의 여성이다.(https://eduzaurus.com/free-essay-samples/the-target-market-of-hm/)

### 2-3. 군집화 분석

고객집단을 가장 잘 분류하는 군집수를 알아내기 위해 군집수의 범위를 2부터 7까지 지정하였다.

Gaussian Mixture Model(GMM)을 통해 앞서 준비한 데이터를 군집화 분석한 결과 군집수가 7일 때 군집에 대한 가장 좋은 해석이 가능하였다.(Result에서 상세히 언급)

또 다른 대표적인 군집화 기법인 K-means를 사용하여 군집화 분석한 결과 유이한 연속형 변수인 '연령' 과 '거래횟수' 에 대한 두 기법의 산포도를 비교했을 때 상당히 유사하게 군집했다는 것을 알 수 있었다. 하지만, 그룹별 선호채널은 k-means 기법의 경우 범주가 전반적으로 mixed하게 나와 결론적으로 해석이 더 용이한 GMM 기법을 선택하였다.

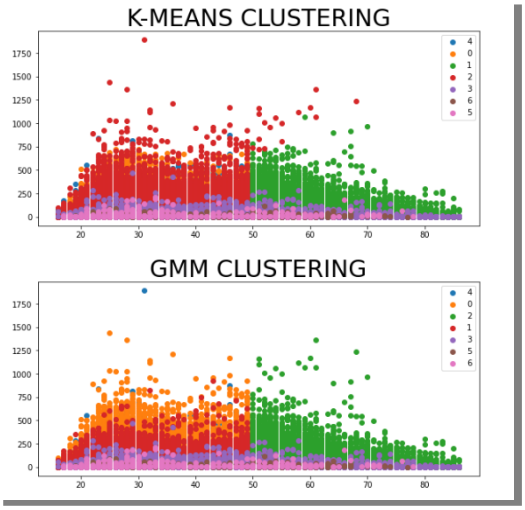


그림 2 군집화 분석 기법별 labels 비교

## 3. Results

최종적으로 GMM 기법을 통해 군집화 분석한 결과는 다음과 같다.

| 군집 | 인원 | 평균연령 (분산) | 평균거래횟수 (분산) | 패션뉴스 빈도 | 선호 채널 | 회원가입 | 비 고 |
|---|---|---|---|---|---|---|---|
| 0 | 404031 | 29 (76) | 23 (1558) | None | 오프라인 | ACTIVE | 오프라인 선호, 젊은 |
| 1 | 284831 | 57 (40) | 22 (1381) | Mixed | Mixed | ACTIVE | 나이가 많은 |
| 2 | 347620 | 29 (81) | 29 (2899) | Regularly | Mixed | ACTIVE | 목표 고객군 |
| 3 | 207724 | 30 (186) | 22 (125) | None | 온라인 | ACTIVE | 온라인 선호, 젊은 |
| 4 | 83276 | 40 (83) | 7 (1160) | Mixed | 오프라인 | PRE-ACTIVE | 잠재 고객군 |
| 5 | 463 | 34 (192) | 18 (918) | None | Mixed | LEFT CLUB | 회원 탈퇴 |
| 6 | 834 | 39 (172) | 12 (193) | Monthly | Mixed | ACTIVE/PRE-ACTIVE | 패션 애호가 |

표 4 GMM 결과 해석

Mixed는 범주가 섞인 것으로 아직 분류가 덜 된 것일 수도 있지만, 그 자체로 군집의 특성일 수도 있다. 형광색으로 칠한 cell은 다른 군집과 구분되는 해당 군집의 특징이다.

군집 3은 유일하게 온라인 채널을 선호하는 군집이다. 다양한 결제방식을 마련하고 추천/검색 알고리즘을 강화하는 등 온라인 환경에 익숙한 점을 고려한 온라인 서비스를 제공할 수 있다.

군집 2는 젊고 패션에 관심이 많으며 거래횟수가 많다는 특징이 있어 H&M의 주 목표 고객군으로 해석된다. 따라서, 해당 고객군을 온라인으로 유치하기 위해서 많은 노력을 기울여야 한다. 젊은 고객군이라는 점을 고려해 온라인 채널에서 사용할 수 있는 할인 쿠폰을 지급한다거나, 온라인에서 구매할 경우 마일리지를 적립해주는 등 가격에 민감한 젊은 고객들을 위한 서비스를 확충해야 한다.

군집 6은 다달이 패션 뉴스를 구독할 정도로 패션에 관심이 많은 군집으로 H&M 말고도 다른 패션 브랜드들 또한 이용하기 때문에 거래건수가 적은 편이다. 사이트 내에 최신 패션 뉴스를 정기적으로 업데이트하고 패션업계의 유명인의 인터뷰를 게시하는 등 그들의 관심사를 사이트 내에서 제공함으로써 홈페이지 접속과 나아가 거래로 이어지도록 유도할 수 있다.

군집 4는 아직 H&M에 회원 가입하지 않은 고객으로 매출 증대 기회의 영역인 잠재고객군이다. 잠재고객군을 확보하는 것은 사업의 지속적인 성장을 위해 매우 중요하다. 신규회원을 대상으로 하는 공격적인 할인 이벤트나 혜택을 고려해볼 수 있다.

# 4. Discussion

## 4-1. 한계점

- **성별**

  10만개의 상품 데이터에서 여성 의류(Women)이 차지하는 비율은 전체의 37.6&였으며, Divided 또한 사실상 여성의류를 대다수 포함하고 있기 때문에 전체의 50%를 넘는다고 볼 수도 있다. 이처럼 H&M이 취급하는 상품의 다수가 여성상품인데 군집화에 있어서 성별을 고려하지 않는 것은 한계점으로 작용할 수 있다. 여성을 위한 서비스와 남성을 위한 서비스를 분명하게 상이하기 때문이다. 하지만, 고객데이터에 성별에 대한 변수는 포함되어 있지 않았다.

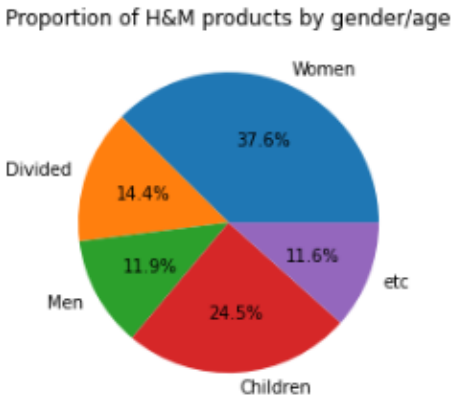


그림 3 성별 H&M 상품 비중

- **가격**

  상품의 수요에 영향을 미치는 변수 중 가장 중요한 것을 꼽으라면 가격이 빠질 수 없다. 본 분석에서도 거래데이터에서 고객별 거래액을 합산하여 군집화 분석에 활용하고자 하였으나 약 3179만건의 거래 데이터에서 고객 id를 search하고 사전에 0과 1 사이의 실수로 scaling된 가격을 합산하는데 많은 계산시간이 소요되어 분석의 진행을 위해 제외하였다.

## 4-2. 결론 및 향후 분석

2018년 9월 20일부터 2020년 9월 22일 사이의 거래건수가 약한 하향세를 보이는 H&M의 문제를 해결하기 위해 고객 군집화 분석을 시도하였으며 유의미한 군집을 도출해냈다. 군집별로 맞춤형 서비스를 제공함으로써 온라인 채널에서의 거래건수 증대에 기여할 것으로 기대된다. 향후 분석으로 제공된 데이터 중 상품데이터를 이용하여 고객군별로 선호하는 상품을 파악하거나 선호하는 색상/패턴 등을 파악하여 상품개발과 상품추천에 활용할 수 있다. 또한 상품추천 알고리즘을 만들기 위해 거래데이터를 apriori algorithm을 적용할 수 있는 형태로 바꾸어 장바구니 분석이라 불리는 연관규칙분석을 수행할 수 있다.