



[데이터 분석] With tutor

여러분들은 데이터 전처리를 위한 **판다스**와 시각화 라이브러리인 **Matplotlib / Seaborn**을 배웠습니다!



이제 프로젝트를 해보며 배운 문법들을 적용할 차례입니다.

자 그럼.... 해보세요 (???)



농담이고, 우리 함께 **H&M 데이터 분석**을 진행해보겠습니다. (미니 과제로 익숙해졌을거예요?)



일단 아래 데이터를 세팅하시죠 !



1. https://drive.google.com/file/d/1Ve7fuXKdINtb_o8gE3MJNCydpZTevHfr/view?usp=drive_link
2. https://drive.google.com/file/d/1SI2GzSg--aJ_ke9ey4fyUjNjR9d715Bs/view?usp=drive_link
3. https://drive.google.com/file/d/1trstSdrReMfULE0loTfNE4eqR65FNs9y/view?usp=drive_link

0. 상황 가정 및 파이프라인 정리



여러분은 H&M 데이터 분석 팀의 신입 데이터 애널리스트입니다!

[아래 대화를 통해 여러분들은 프로젝트를 수행해야 합니다.]

천준석 팀장 : ◻◻님 안녕하세요. 좋은 아침입니다.



여러분(ooo님) : 안녕하세요 팀장님. (첫 출근이지만, 열심히 해보겠어!)

천준석 팀장 : ◻◻님 온보딩 기간 동안 자사 데이터를 활용해서 혼자서 데이터 분석 한번 해 보시겠어요?

여러분(ooo님) : 네! 해보겠습니다. (큰일났다. 비상)

천준석 팀장 : 중간 중간 저한테 피드백 요청해주세요~

데이터는 받았는데.. 이제 어떻게 해야하지?

1. 도메인 탐색

- H&M 이라는 회사가 어떤 회사인지?
- 전달 받은 데이터는 대략적으로 어떤 데이터인지?
- H&M의 가장 중요한 목표/목적은 무엇인지?

2. 라이트 EDA

- 데이터 간 관계(ERD) → 그려보기
- 데이터 로드, 크기, 컬럼 확인 → md, excel 등.. 파일로 정리
- (목표/목적과 연관된) 주요 컬럼 분포 확인(Age, transactions date, article ...) → 무엇이 주요 컬럼? → 도메인 탐색에서 목표.목적에 부합하는..
- (선택) 결측치/이상치 확인 → 처리
- **항상 호기심을 가지고 데이터를 들여다보기**

3. 비즈니스/프로젝트 목표 세우기

4. 가설 설정

5. 미니 검증

6. 선택/심화

7. 최종 결론

1. 도메인 탐색

▼ H&M 이라는 회사가 어떤 회사인지?

- H&M은 스웨덴에 본사를 둔 글로벌 패스트패션 의류 소매기업입니다. 전 세계 75개 이상의 시장에서 수천 개 매장을 운영합니다.
- H&M의 비즈니스 모토는 “**최고의 가격에 패션과 품질을 제공한다**”는 것으로, 패션의 민주화를 목표로 합리적 가격대의 최신 유행 상품을 빠르게 공급하고 있습니다.
- 이러한 **빠른 상품 회전과 광범위한 상품군**(남성복, 여성복, 아동복 등)으로 트렌디한 의류를 제공하는 것이 H&M 비즈니스 모델의 핵심입니다.

▼ 전달 받은 데이터는 대략적으로 어떤 데이터인지?

- 분석할 데이터는 **H&M 고객 행동 데이터셋** → 고객들의 **구매 이력(트랜잭션)**과 **상품 정보, 고객 정보**를 포함하고 있습니다.
- 수백만 건의 구매 거래와 10만 종 이상의 상품 정보를 포함합니다.
- 데이터에는 **시간순 구매 기록**, 고객의 연령 등 **인구통계 정보**, 상품의 카테고리 등 **메타데이터**가 풍부하게 포함되어 있습니다.
- H&M은 **개인화된 패션 추천**과 **디지털 전환**에 큰 관심을 가지고 있으며, 기술과 데이터를 활용해 고객 경험을 개선하고 싶어합니다.
- `articles_hm.csv` — 상품 메타데이터(상품 ID, 제품군/유형, 라인, 제품명·설명 등)
- `customer_hm.csv` — 고객 속성(ID, 멤버십/뉴스레터 관련 변수, 나이 등)
- `transactions_hm.csv` — 거래 이력(일자, 고객/상품 ID, 가격, 판매채널 등)

▼ H&M의 가장 중요한 목표/목적은 무엇인지?

- **H&M의 분석 목표**는 데이터를 통해 **고객 행동을 이해**하고, 이를 바탕으로 **매출 증대**와 **재고 최적화**, **고객 만족도 향상**에 기여하는 것입니다.
- 나이 ? 날짜 ? 상품 ? 구매 유형 ? 재고 ?

2. 라이트 EDA

1) 데이터 로드, 크기, 컬럼 확인

• 주요 테이블 (총 3개의 CSV 파일)

- `customers_hm` : 고객 ID, 성별, 연령, 클럽 상태, 뉴스 구독 여부
- `transactions_hm` : 거래일, 고객 ID, 상품 ID, 가격, 채널
- `articles_hm` : 상품 ID, 상품명, 카테고리, 색상, 상세 설명

```
import pandas as pd

# 데이터 불러오기
customers = pd.read_csv('customer_hm.csv')
articles = pd.read_csv('articles_hm.csv')
transactions = pd.read_csv('transactions_hm.csv')

# 데이터 크기(행렬 개수) 확인
print(customers.shape) # 예: (1,048,575, 6)
print(articles.shape) # 예: (105,542, 25)
print(transactions.shape) # 예: (1,048,575, 5)

# 각 데이터의 컬럼 확인
print(customers.columns.tolist())
print(articles.columns.tolist())
print(transactions.columns.tolist())

# 필요하다면 info, describe 등 적극 활용
print(customers.info())
print(articles.info())
print(transactions.info())
```

```
# 간단하게 첫 3개의 행 데이터 확인
print(customers.head(3))
print(articles.head(3))
print(transactions.head(3))
```

```
# 예시로 customers 테이블만 확인했지만, 수강생들은 모두 다 확인해야 합니다! (필수)
print(customers['FN'].value_counts())
print(customers['Active'].value_counts())
print(customers['club_member_status'].value_counts())
print(customers['fashion_news_frequency'].value_counts())
print(customers['age'].value_counts())
```

▼ 결과 간단 정리

1. 대략 100만 명 규모의 고객, 10만 종 이상의 상품, 약 100만 건의 거래 기록이 있는 것을 확인
2. 고객(`customer_hm.csv`)
 - `customer_id` , `FN` , `Active` , `club_member_status` , `fashion_news_frequency` , `age` 와 같은 컬럼 존재

- FN은 패션 뉴스 구독 여부, Active는 고객이 메시지 수신에 활성화되었는지 여부를 나타내는 이진 변수입니다 (두 변수 모두 0/1 값)
- `club_member_status` 는 고객의 멤버십 상태 (예: ACTIVE, PRE-CREATE 등), `fashion_news_frequency` 는 패션 뉴스레터 수신 빈도 (예: None, Regularly 등), `age` 는 나이를 나타냅니다.

3. 상품(`articles_hm.csv`)

- 상품 고유 ID와 함께 상품명, 상품 유형(`product_type_name`), 상품군(`product_group_name`), 색상, 의류 부문(`index_name` 및 `index_group_name`), 세부 설명(`detail_desc`) 등의 상세한 정보 존재

4. 거래(`transactions_hm.csv`)

- 구매 날짜(`t_dat`), 고객 ID, 상품 ID, 가격, 판매 채널(`sales_channel_id`) 등의 칼럼으로 구성
- 판매 채널은 `1` 이 오프라인 매장 구매, `2` 가 온라인 구매 의미

▼ 🛍 데이터 정리

- customers_hm

컬럼명	컬럼 상세 정보
customer_id	고객 id
FN	패션 뉴스 구독여부
Active	커뮤니케이션 가능여부
club_member_status	회원 상태 (신규, 활성, 탈퇴)
fashion_news_frequency	패션 뉴스 알람 주기
age	나이

- transactions_hm

컬럼명	컬럼 상세 정보
t_dat	구매일
customer_id	고객 id
article_id	article id
price	가격
sales_channel_id	판매 채널 (1 : 오프라인, 2: 온라인)

- articles_hm

컬럼명	컬럼 상세 정보
article_id	article 아이디
product_code	상품 코드
prod_name	상품 명

컬럼명	컬럼 상세 정보
product_type_no	상품 종류 코드
product_type_name	상품 종류 명
product_group_name	제품군
graphical_appearance_no	그래픽 정보 코드
graphical_appearance_name	그래픽 정보 - (예시) Solid
colour_group_code	상품 색상 코드
colour_group_name	상품 색상 명 - (예시) Light Beige
perceived_colour_value_id	명도 id
perceived_colour_value_name	명도 분류 - (예시) Dusty Light
perceived_colour_master_id	색상군 코드
perceived_colour_master_name	색상군 - (예시) Beige
department_no	소분류 코드
department_name	소분류 명 - (예시) Tights basic
index_code	중분류 코드
index_name	중분류 명 - (예시) Lingeries/Tights
index_group_no	타겟 그룹 코드
index_group_name	타겟 그룹 명 - (예시) Ladieswear
section_no	섹션 코드
section_name	섹션 명 - (예시) Womens Nightwear, Socks & Tigh
garment_group_no	대분류 코드
garment_group_name	대분류 명 - (예시) Socks and Tights
detail_desc	제품 상세 설명

만약 의미를 알 수 없는 컬럼이 있다면?

?

1. 오픈 데이터거나 혹은 도메인 지식이 없는 경우 검색
2. 고유 데이터의 경우(대부분 직장내) 데이터 엔지니어나 백엔드 개발자의 문서 확인
3. 문서가 없다면 직접 물어볼 수 밖에..

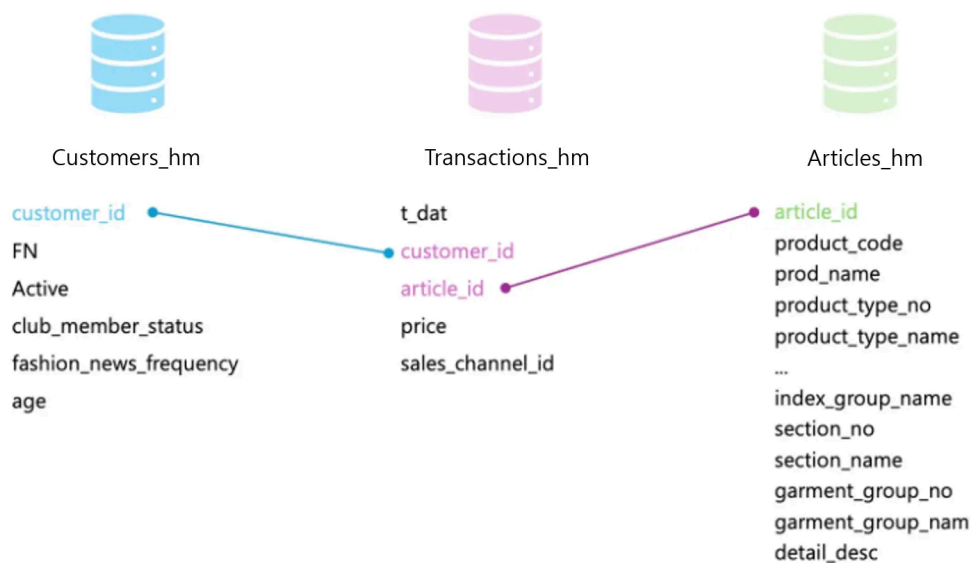
(선택) 결측치/이상치 확인

- 각 테이블의 결측치와 중복을 체크

- `articles` 테이블의 `detail_desc` 컬럼 416개의 결측치 존재 → “Unknown” 처리 (제품의 상세 설명 “모름”으로 표시)
- `transactions` 테이블 중복 행 8474개 존재
 - 완전 동일행 제거 : 가장 보수적/안전
 - 같은 거래를 한 행으로 묶고 수량(`qty`)을 부여 : 매출/판매건수 왜곡 방지에 유리
 - 전체의 <1% 수준일 가능성이 커서 제거해도 결과에 크게 영향을 주진 않음!

! 위 결과가 나오는 지 확인하고, 본인의 생각에 따라 처리해보세요

2) 데이터 간 관계(ERD)



▼ WHY?

- 어떤 데이터가 어떤 키(pk, fk)로 연결되는지 확인할 수 있어, 처음 보는 데이터셋을 빠르게 이해하는 데 유용하다.
- H&M 프로젝트에서 `customers` ↔ `transactions` ↔ `articles` 관계에서 거래는 고객과 상품을 연결하는 교차 테이블이다.
- 분석할 때 “어떤 테이블을 조인해야 하지?” → 해결
- 고객별 구매 패턴을 보려면 `customers` 와 `transactions` 를 `customer_id` 기준으로 조인해야 한다는 사실을 ERD가 명확히 보여준다.

데이터 분석가가 굳이 ERD를 그려야 할까요?

?

1. 대부분의 경우 백엔드 개발자 혹은 데이터 엔지니어가 구성을 하지만, ERD를 읽을 수 있는 분석가는 말이 잘 통함..
2. 분석할 때, 데이터 이해가 유용함. 혹은 MERGE/JOIN/CONCAT 시 유용
3. 간단한 데이터는 그려보는 것을 강추!

3) (목표/목적과 연관된) 주요 컬럼 분포 확인

- 질문 : 가장 중요한 고객의 나이대는 어떻게 될까? → [커머스의 매우 중요한 지표]
- 고객 데이터의 `age` 컬럼

```
import matplotlib.pyplot as plt

# 단변량 수치형 변수의 분포 -> 히스토그램
customers['age'].hist(bins=30, color='skyblue', edgecolor='black')
plt.title('Distribution of Customer Age')
plt.xlabel('Age')
plt.ylabel('Number of Customers')
plt.show()
```

▼ 해석

- 고객 데이터에서 각 고객의 나이 분포를 요약해보면, `age` 컬럼의 최소값과 최대값은 대략 16세부터 90세 안팎이며, 평균 나이는 35세 정도로 추정됩니다.
 - 고객 연령 분포를 나타낸 것으로, 주요 고객층이 젊은 연령대(20~30대)에 몰려있음을 시각화합니다.
-
- 질문1 : 거래 데이터의 기간은 어떻게 될까?
 - 질문2 : 실제 거래한(구매한) 고객은 몇 명 정도일까?
 - 질문3 : 온/오프라인 거래 건수의 차이는 얼마나 날까? 어디가 더 비중이 높을까?

```
# 0) str → datetime type 변경
transactions['t_dat'] = pd.to_datetime(transactions['t_dat'])

# 1) 날짜 범위
min_date = transactions['t_dat'].min()
max_date = transactions['t_dat'].max()
```

```

# 2) 총 거래 건수 & 고유 고객 수
n_tx = len(transactions)
n_cust = transactions['customer_id'].nunique() # nunique : 고유값의 수

# 3) 채널 비중(1=오프라인, 2=온라인)
transactions['channel'] = transactions['sales_channel_id'].map({1: 'offline(1)', 2: 'online(2)'})
transactions['channel'].fillna('other/unknown')
ch_counts = transactions['channel'].value_counts()
ch_share = (transactions['channel'].value_counts(normalize=True) * 100).round(2) #
normalize=True : 값의 빈도(비율) 구하기

# 4) 출력
print("=== H&M 거래 요약 (2019년) ===")
print(f"- 날짜 범위: {min_date.date()} → {max_date.date()}")
print(f"- 총 거래 건수: {n_tx:,} 건")
print(f"- 고유 고객 수: {n_cust:,} 명")

print("\n[판매 채널별 거래 건수]")
print(ch_counts.to_string())

print("\n[판매 채널별 비중(%)]")
print(ch_share.to_string())

# 5) 간단 해석(온라인 > 오프라인 여부)
online = ch_counts.get('online(2)', 0)
offline = ch_counts.get('offline(1)', 0)
if online > offline:
    print("\n해석: 온라인(채널=2) 거래가 오프라인(채널=1)보다 많습니다 → 온라인 비중이 더 높음.")
elif online < offline:
    print("\n해석: 오프라인(채널=1) 거래가 온라인(채널=2)보다 많습니다.")
else:
    print("\n해석: 온라인과 오프라인 거래 건수가 동일합니다.")

```

▼ 해석

- 거래 데이터는 **2019년 1월 1일부터 2019년 12월 31일까지** 1년간의 구매 이력을 담고 있습니다.
- 총 거래 건수는 약 **104만 건**이며, **고유 고객 수는 약 45만 명** (전체 고객의 절반가량)입니다.
 - 데이터에 등록된 많은 고객 중 절반 정도만 해당 기간에 실제 구매를 했고, 나머지는 그 기간에 구매활동이 없었음을 의미

- 온라인 쇼핑 비중이 꽤 높은 것으로 확인

▼ 위 결과(온라인(채널=2) / 오프라인(채널=1))를 그래프(bar, pie)로 표현하려면 어떻게 하면 될까?

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.rcParams['font.family'] = 'Malgun Gothic' # For Windows
# plt.rcParams['font.family'] = 'AppleGothic' # For MacOS
%matplotlib inline

plt.figure(figsize=(6, 4))
# plt.bar(ch_share.index, ch_share.values)
plt.pie(ch_share.values, labels=ch_share.index, autopct='%1f%%')
plt.title("온라인(채널=2) / 오프라인(채널=1) 비교")
plt.show()
```

- 질문 1. 어떤 상품이 많이 팔렸을까? → 거래 데이터와 상품 데이터를 Merge

```
# 거래 데이터에 상품의 대분류 정보를 병합
tx_merge = pd.merge(transactions, articles[['article_id','product_type_name','product_group_name','index_group_name']],
                    on='article_id', how='left')

# 상품 대분류(index_group_name)별 판매 건수 집계
# 대분류의 상품군(product_group_name), 상품 유형(product_type_name)으로 집계 가능
sales_by_group = tx_merge['index_group_name'].value_counts()
print(sales_by_group)
```

▼ 해석

- 여성복 라인(Ladieswear + Divided 등)이 가장 큰 비중을 차지하고, 그 다음으로 남성복 순으로 판매량이 높음을 알 수 있습니다.
 - H&M의 주요 타겟층이 젊은 여성 위주인 점과도 부합
- 상품군(`product_group_name`)이나 상품 유형(`product_type_name`)별 Top 10을 살펴보면, 예를 들어 `T-Shirt` , `Dress` , `Jeans` , `Knitted top` 등 기본 의류 품목들이 상위권에 올라와 있습니다.
- 질문 1. 시간대별 거래량이 차이가 있을까? → 일반적으로 패션 리테일 업계에서는 연말 휴휴(블랙 프라이데이, 크리스마스 등)에 매출이 크게 증가함! **[확인 필수]**

```

transactions['month'] = pd.to_datetime(transactions['t_dat']).dt.month
monthly_sales = transactions.groupby('month').size()
print(monthly_sales)
plt.plot(monthly_sales.index, monthly_sales.values, marker='o')
plt.xticks([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12])
plt.title("2019년 월별 거래량")
plt.xlabel('월')
plt.ylabel('거래량')
plt.show()

```

▼ 해석

- 2019년 **6-7월이 피크** → 여름 시즌/프로모션 영향일 수 있음
- **온라인/오프라인 거래량의 차이를 비교하는 그래프는?**

```

# 여름 피크가 온라인/오프라인 중 어디에서 더 강했는지

# 채널 라벨링
transactions['channel'] = transactions['sales_channel_id'].map({1: 'offline(1)', 2:
'online(2)'}).fillna('other')

# 월x채널 집계
mc = (transactions
      .groupby([transactions['t_dat'].dt.to_period('M'), 'channel'])
      .size()
      .rename('tx_cnt')
      .reset_index())

# 시간 정렬용 timestamp
mc['month'] = mc['t_dat'].dt.to_timestamp()
mc = mc.drop(columns='t_dat')

# online/offline만 비교
order = ['offline(1)', 'online(2)']
mc2 = mc[mc['channel'].isin(order)]

import seaborn as sns
sns.set_theme(style='whitegrid')

plt.figure(figsize=(9,4))
sns.lineplot(data=mc2, x='month', y='tx_cnt', hue='channel', marker='o', hue_order=order)

```

```
plt.title('2019년 채널별/월별 거래량')
plt.xlabel('월')
plt.ylabel('거래량')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

이상의 라이트 EDA를 통해, 데이터의 전반적인 형태와 몇 가지 초기 인사이트를 얻었습니다

- 고객은 주로 20~30대이며, 절반 정도의 고객만이 기간 내 구매를 했습니다.
- 온라인 구매 비중이 상당히 높았습니다.
- 여성 의류 라인의 판매량이 가장 크고, 상위 몇 개의 상품 유형이 전체 판매의 큰 부분을 차지했습니다.
- 6-7월에 판매량이 급증하는 계절적 트렌드가 있습니다.

이제 이러한 관찰을 바탕으로, **프로젝트의 비즈니스 목표**를 명확히 정의하고 다음 단계로 넘어가겠습니다.

3. 비즈니스/프로젝트 목표 세우기

- 초기 탐색을 통해 여러 방향의 분석 가능성을 확인했고, 이제 이 중에서 **어떤 비즈니스 문제를 중점적으로 풀 것인지 목표를 설정**해야 한다.
- 고려해볼 만한 프로젝트/분석 목표

1. 고객 세그먼트별 추천 시스템 구축

- 고객들을 구매 행동이나 인구통계에 따라 세분화하고, 각 군집에 맞춤형 상품을 추천
- 예를 들어, 10대 고객군에게 인기 있는 상품 vs 30대 고객군 인기 상품을 비교하고, 향후 개인화 추천의 기초 구성
- 이 방향은 **고객 관점에서 개인화 서비스**를 향상시켜 **재구매율**과 **고객 충성도**를 높이는 것이 목표

2. 재구매 고객 분석 (고객 충성도 분석)


- **한 번 구매하고 이탈하는 고객**과 **반복 구매하는 충성 고객**의 특성을 비교 분석
- 예를 들어, 데이터에 **다회 구매 고객 vs 1회성 고객 비율**을 확인하고, 충성 고객의 구매 패턴(구매 간격, 구매 상품 종류)을 파악
- 이를 통해 **이탈률을 낮추는 마케팅 전략** (예: 할인 쿠폰, 맞춤 프로모션)을 제안 가능

3. 인기 상품 분석 및 재고 최적화

- 어떤 상품이 가장 잘 팔리는지 및 어떤 상품이 잘 안 팔렸는지를 분석
- 상위 판매 상품들이 매출에 얼마나 기여하는지 파악하고, 반대로 재고만 차지한 비인기 상품을 식별
- 이를 통해 한정된 매장 공간과 자본을 인기 상품에 집중시키고, 느린 회전 상품은 적정 수준으로 재고를 줄이거나 프로모션을 검토
- 나아가 시즌별로 어느 카테고리가 잘 팔렸는지를 알아내어 수요 예측 및 상품 기획에 활용

인기 상품 분석 및 재고 최적화



1. 주어진 데이터만으로도 충분히 분석 가능
2. 결과를 간단한 리포트와 시각화로도 효과적으로 전달 가능
3. 비즈니스 측면에서도 팔리는 상품에 집중하고, 불필요한 재고를 줄이는 것은 매출 증대와 비용 절감으로 직결 —> 경영진 관심 

프로젝트 목표

H&M의 2019년 인기 상품 분석을 통해 판매 패턴을 파악하고, 이를 활용한 재고 최적화 전략 제안



- 어떤 소수의 상품이 전체 매출의 대부분을 차지하는지 (예: 파레토 법칙 적용 여부 확인)
- 제품 카테고리별로 판매 성과가 어떻게 다른지 (어느 부문이 주력 매출원인가)
- 재고 비효율: 판매 저조 상품의 규모와 특징은 무엇인지
- 시즌별 상품 전략: 시기별 잘 팔린 품목 (예: 겨울 아우터 vs 여름 티셔츠) 등을 분석

4. 가설 설정

프로젝트 목표(인기 상품 분석 및 재고 최적화)에 따라, 탐색하고 검증할 몇 가지 가설을 수립해보자.

가설은 이후 분석 방향을 이끌 길잡이 역할을 하며, 데이터로부터 어떤 이야기를 도출할지 미리 예상해 보는 단계:

- **가설 1: 매출의 대부분이 일부 인기 상품에 집중되어 있을 것이다.**
 - 흔히 소매업에서 말하는 파레토 법칙(80/20 법칙)처럼, 상위 20% 상품이 전체 판매의 80%를 차지할 것이라고 예상합니다.
 - H&M처럼 품목 가짓수가 많은 리테일에서는 히트 상품이 매출을 견인하고, 나머지 다수 상품은 상대적으로 판매량이 낮을 가능성이 큼니다.

- **가설 2: 상품 대분류(제품군)별로 판매 편차가 크고, 특히 여성복 카테고리가 가장 큰 매출 비중을 차지할 것이다.**
 - 앞서 EDA에서도 관찰했듯이 **Ladieswear** 부문이 큰 비중을 차지할 것으로 보입니다.
 - 남성복이나 아동복도 매출이 있겠지만, H&M의 핵심 타겟이 **패션에 민감한 젊은 여성층**이므로 여성 의류 판매량이 두드러질 것입니다.
 - 또한 상품군별로 **예를 들면 아우터 vs 액세서리** 등도 판매량 차이가 있을 것 같습니다 (가격과 수요 빈도의 차이로).
- **가설 3: 상당수 상품이 재고로 남아있다 - 즉, 판매량이 매우 저조한 상품이 많을 것이다.**
 - SKU(Stock Keeping Unit - 기업에서 재고를 추적하고 관리하기 위해 특정 상품에 부여하는 고유한 영숫자 코드)가 10만개나 되는 만큼 모든 상품이 고르게 팔릴 리 없습니다.
 - **판매 건수가 0이거나 아주 낮은 긴 꼬리** 상품들이 다수 있을 것으로 예상합니다.
 - 예를 들어 **전체 상품의 50% 이상은 2019년에 1개도 팔리지 않았을 것**이라는 극단적인 가정도 해봅니다. 이는 재고 관리 측면에서 **정리 대상 상품**을 식별하는 데 중요합니다.
- **가설 4: 시즌별로 잘 팔리는 상품 유형이 다를 것이다.**
 - 예를 들어 **겨울철(11~12월)**에는 코트, 니트 등의 **방한 의류 판매량 급증**, **여름철(6~8월)**에는 티셔츠, 반바지 등 **경량 의류 판매 증가**가 있을 것으로 봅니다.
 - 또한 11월 말 블랙프라이데이 세일이나 연말 시즌의 영향으로 **11월~12월에 연중 최고 판매량 기록** 가설을 세워봅니다. 이 가설은 **시즌별 재고 전략** 수립에 도움을 줄 수 있습니다.
- **(추가로 고려할 가설) 멤버십 및 마케팅의 효과**
 - 예를 들어 **FN = 1**이면서 **Active = 1**인 고객(즉, 패션 뉴스레터를 받아보고 적극적으로 활동중인 고객)이 더 자주 구매하지 않을까 하는 가설도 세워볼 수 있습니다.
 - 하지만 이 부분은 우리 목표의 중심은 아니므로 부가적으로 참고만 할 것입니다.

이상의 가설들은 지금 단계에서는 **검증되지 않은 추측**입니다.

다음 단계에서 간단한 분석으로 **미니 검증**을 해보고, 가설이 맞는 방향인지 확인하여 계속 밀고 나갈지 혹은 수정할지 결정해보겠습니다.