***※ Sports & Outdoors meta data 전처리***

**① 처리**

**1) 불필요 칼럼 삭제**

def drop\_columns(dataframe):

    columns\_to\_drop = ['fit','tech1','tech2']

    return dataframe.drop(columns\_to\_drop, axis=1)

**2) [] or ‘’ 로 채워져있는 값 null 로 변경**

def replace\_empty\_with\_null(dataframe):

    dataframe = dataframe.applymap(lambda x: None if (isinstance(x, list) and not x) or (isinstance(x, str) and not x) else x)

    return dataframe

🡪 적용시켜봤더니 끝나질 않음..

def replace\_empty\_with\_null(dataframe):

    for i in dataframe.columns:

        dataframe[i] = dataframe[i].apply(lambda x: None if (isinstance(x, list) and not x) or (isinstance(x, str) and not x) else x)

    return dataframe

🡪

**3) asin 중복 값 삭제**

🡪 asin 값 동일할 경우, 다른 칼럼의 값도 모두 완전히 동일

def remove\_duplicate\_rows(dataframe):

    columns\_to\_drop\_duplicate = ['asin']

    dataframe = dataframe.drop\_duplicates(subset=columns\_to\_drop\_duplicate, keep='first')

    return dataframe

**4) category 나누기**

def separate\_category\_columns(dataframe):

    dataframe['category\_first'] = dataframe['category'].apply(lambda x: x[0] if len(x)>0 else None )   # 대카테고리

    dataframe['category\_second'] = dataframe['category'].apply(lambda x: x[1] if len(x)>1 else None )   # 중카테고리

    dataframe['category\_third'] = dataframe['category'].apply(lambda x: x[2] if len(x)>2 else None ) # 소카테고리

    dataframe['category\_fourth'] = dataframe['category'].apply(lambda x: x[3] if len(x)>3 else None ) # 소카테고리

    return dataframe

**5) date 날짜 형식만 남기고 나머지 다 null로 만들기!**

def separate\_category\_columns(dataframe):

    dataframe['category\_first'] = dataframe['category'].apply(lambda x: x[0] if len(x)>0 else None )   # 대카테고리

    dataframe['category\_second'] = dataframe['category'].apply(lambda x: x[1] if len(x)>1 else None )   # 중카테고리

    dataframe['category\_third'] = dataframe['category'].apply(lambda x: x[2] if len(x)>2 else None ) # 소카테고리

    dataframe['category\_fourth'] = dataframe['category'].apply(lambda x: x[3] if len(x)>3 else None ) # 소카테고리

    return dataframe

**6) main\_cat 칼럼 이상한 값 삭제**

def filter\_main\_cat\_values(dataframe):

    # <로 시작하는 값 null로 변경

    dataframe['main\_cat'] = dataframe['main\_cat'].apply(lambda x: None if x.startswith('<') else x)

    return dataframe

**7) price 칼럼**

def clean\_price\_column(dataframe):

    def process\_price(value):

        # 단일 달러 값 처리

        if '$' in value and '-' not in value:

            return float(re.sub(r'[^0-9.]', '', value))

        # 범주형 값 처리

        elif '$' in value and '-' in value:

            # 범위의 평균값 계산

            prices = re.findall(r'\d+\.\d+', value)

            if len(prices) == 2:

                return (float(prices[0]) + float(prices[1])) / 2

            else:

                return None

        else:

            return None

    # 'price' 칼럼 값 처리

    dataframe['price'] = dataframe['price'].apply(process\_price)

    return dataframe

**8) rank 칼럼**

**② 컬럼별 특징\_현황(962,300개 Data)**

1) also\_buy : [] 값 약 70만개

2) also\_view : [] 값 약 60만개

3) asin : null 값 및 []로 채워져 있는 값 없음!

: 957,000 개의 unique 상품 존재

: 5300개는 중복되어 있으나, 확인 결과 완전히 똑 같은 행 들어가있음(**중복 값삭제 필요**)

4) brand : ‘’ 값 약 96973개

: 75434 개의 UNIQUE 브랜드 존재

5) category : [] 값 68321 개

: 대/중/소/극소 4개로 나누기

6) date : ‘’ 값 841845개, <div class="a-> / 브랜드 명 등 이상한 값 다수 존재

: 제대로 된 값만 추출하기 필요(년/월/일 나눌 필요 x)

7) description : [] 값 15만개, 그 외 이상 무

8) details : null 값 138개

9) features : [] 값 158000개

10) **fit** : ‘’ 값 92만개

: 그 외 값 전부 class="a-normal a-align-center a-spacing-small">\n

: **칼럼 삭제 필요**

11) main\_cat : sports&outdoors 81만개, 그 외 이상한 값 or 안 맞는 카테고리 값 다수

: <img src="https://images-로 시작하는 것만 null로 만들고 혹시 모르니 칼럼 살려놓기

12) price : ‘’ 값 55만 6천개

: .a-box-inner{background- 와 같은 이상한 값 5천개

: ‘’ 값 / 범주형 / 이상한값 으로 구성되어 있음

13) rank : []값 1만9천개

: 애초에61,456 in Clothing, Shoes & Jewelry ( 6 이런 값도 존재

(sport&outdoor 가 아닌 것도 존재)

14) similar\_item : ‘’값 94만개

: class="a-bordered a-hoㅇrizonta 값도 많음(수량 확인 후 null로 변환 필요

15) tech1 : ‘’값 94만개

: 나머지 값 싹다 이상함(class="a-keyval)

15) tech2 : ‘’값 96만개

: 나머지 값 싹다 이상함(class="a-keyval)

**\*\*\* tech 1 & tech 2 컬럼 삭제**

16) title : 이상한 값 없음

: 91만개의 unique한 title 값 존재

: asin 보다 중복하는 값들이 더 많다는 뜻

: 색상 다를 경우 동일 title이어도, asin 이 다른듯..?

: **이런 값들을 하나의 asin으로 합쳐야 하나???**