데이터 전처리 결과 보고서

---

1. 기본 정보

- **프로젝트명**: Senpick
- **작성일**: 2025.05.08

2. 데이터 개요

- **출처**:
  - 카카오톡 선물하기(카카오톡 선물하기)
  - 네이버 선물샵(네이버 선물샵)
  - SSG 선물하기(SSG 선물하기)
  - 롯데ON(롯데ON)
  - 올리브영(올리브영)
  - 오늘의집(오늘의집)

- **수집 방식**: 웹 크롤링 (Selenium 기반)
- **수집 기간**: 2025.04.25 ~ 2025.05.07
- **수집 데이터 포맷**: JSON
- **총 수집 건수**:
  - 상품: 17,856건
  - 리뷰: 833,347건
- **최종 확보량**:
  - 상품: 17,698건
  - 리뷰: 828,618건

3. 전처리 과정

- 전처리 전 수집된 데이터 확인

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   brand_name      13071 non-null  object
 1   product_name    17856 non-null  object
```

```
 2   category_main    17856 non-null  object
 3   category_sub     17856 non-null  object
 4   options          17816 non-null  object
 5   product_url      17856 non-null  object
 6   thumbnail_url    17856 non-null  object
 7   review_content   17856 non-null  object
 8   price            17856 non-null  object
 9   brand            4628 non-null   object
 10  reviews          4628 non-null   object
```

- 쇼핑몰(shop) 컬럼 추가하고 수집 출처 기입
- 브랜드 컬럼 명 brand_name으로 통일
- 불필요한 컬럼 삭제 (brand, reviews)
- 결측치 존재하는 행 삭제
    - 브랜드명(brand_name)이 '브랜드 없음'인 행 (가격(price), 상품이름(product_name), 리뷰(review_content) 없음)
    - 상품이름(product_name), 가격(price) 컬럼이 모두 없는 행
- 브랜드명(brand_name)이 비어있는 값 ''(빈 문자열)로 변경
- 대카테고리(category_main) 재분류

```
# 재분류 전 대카테고리 목록 (31개)
[
    '유아동', 'e쿠폰', '결혼·집들이', '레저자동차', '영양제/건강식품',
    '이벤트/파티용품', '블루투스스피커', '주류', '뷰티', '리빙/도서',
    '유아동/반려', '디지털/가전', '교환권', '레저/스포츠', '식품',
    '패션', '스포츠/레저', '주방용품', '가구', '가전·디지털',
    '데코·식물', '생필품', '수납·정리', '조명', '패브릭',
    '결혼집들이', '백화점선물', '맛있는선물', '생일/축하', '럭셔리선물',
    '건강/회복'
]
```

```
# 재분류 후 대카테고리 목록 (13개)
[
    '유아동', '선물권/교환권', '테마/기념일 선물', '레저/스포츠/자동차',
    '건강', '디지털/가전', '식품/음료', '뷰티', '리빙/인테리어',
    '반려동물', '패션', '생활', '프리미엄 선물'
]
```

- 옵션(options) 형식 변경
    - NaN인 값 [](빈 배열)로 변경

    - str 형식인 값 list 형식으로 변경
- 리뷰 본문(`review_content`) 정제
    - 개행 문자 제거
    - 연속된 특수문자 1회만 표기
    - 한글 자모 연속 횟수 3회까지 제한
    - 특수문자 제거 (`!`, `~`, `;`, `^` 유지)
    - 허용 문자(한글/영문/숫자/공백)만 허용
    - 띄어쓰기 보정 (중복 공백 제거)

- 가격(`price`) int 형식으로 통일

4. 전처리 후 데이터 구조

```
#   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   brand_name      17698 non-null  object
 1   product_name    17698 non-null  object
 2   category_main   17698 non-null  object
 3   category_sub    17698 non-null  object
 4   options         17698 non-null  object
 5   product_url     17698 non-null  object
 6   thumbnail_url   17698 non-null  object
 7   review_content  17698 non-null  object
 8   price           17698 non-null  int
 9   shop            17698 non-null  object
```

필드명	설명	데이터 타입
brand_name	브랜드명	str
product_name	상품명	str
category_main	상품 카테고리 (대분류)	str
category_sub	상품 카테고리 (소분류)	str
options	상품 옵션 리스트	list[str]

product_url	상품 상세 페이지 링크	str
thumbnail_url	썸네일 이미지 주소	str
review_content	사용자 리뷰 텍스트	list[str]
price	상품 가격 (숫자형)	int
shop	수집 출처 (쇼핑몰)	str

5. 구조적 특징 및 활용성

- 옵션/카테고리 정보를 활용한 **유사도 기반 추천 가능**
- 정제된 리뷰 기반 **감성 분석 및 키워드 추출** 용이
- 대카테고리 기반 필터링 및 문서 분할 구조로 **질의응답 시스템(RAG)** 최적화