

**SK네트웍스 Family AI과정 5기  
 모델링 및 평가 수집된 데이터 및 전처리 문서**



**□ 개요**

* 산출물 단계 : 모델링 및 평가
* 평가 산출물 : 수집된 데이터 및 전처리 문서
* 제출 일자 : 2025.09.12.
* 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN14-final-4Team
* 작성 팀원 : 조성렬

## 1. 개요

### 1.1 데이터 설명

**본 문서는 LoopLabel 프로젝트에서 리사이클 제품 정보 제공 및 코디추천을 위한 데이터 수집 및 전처리 과정을 정리한다. 데이터의 출처, 수집 방법, 전처리 단계, 저장 방식에 대한 내용을 포함한다.**

### 1.2 데이터 수집 목적

**LoopLabel은 사용자의 취향에 맞춰 리사이클 제품 정보를 제공해주고 코디를 추천해주는 앱 어플리케이션이다. 다양한 친환경 제품 중 리사이클 제품을 취급하는 브랜드를 선정 후 상.하의 기준으로 데이터를 수집하여 사용자 맞춤형 코디와 제품정보를 알려줌으로써 사용자의 인식개선에 도움을 주려는 목적을 가진다.**

## 2. 데이터 수집

### 2.1 데이터 출처

* [**https://ecoalf.com**](https://ecoalf.com)
  + **재활용 폴리에스터를 주로 사용하는 리사이클 브랜드**
* [**https://www.patagonia.co.kr/**](https://www.patagonia.co.kr/)
  + **재활용 & 천연섬유를 사용하는 글로벌 브랜드**

### 2.2 데이터 유형

* **텍스트 데이터**

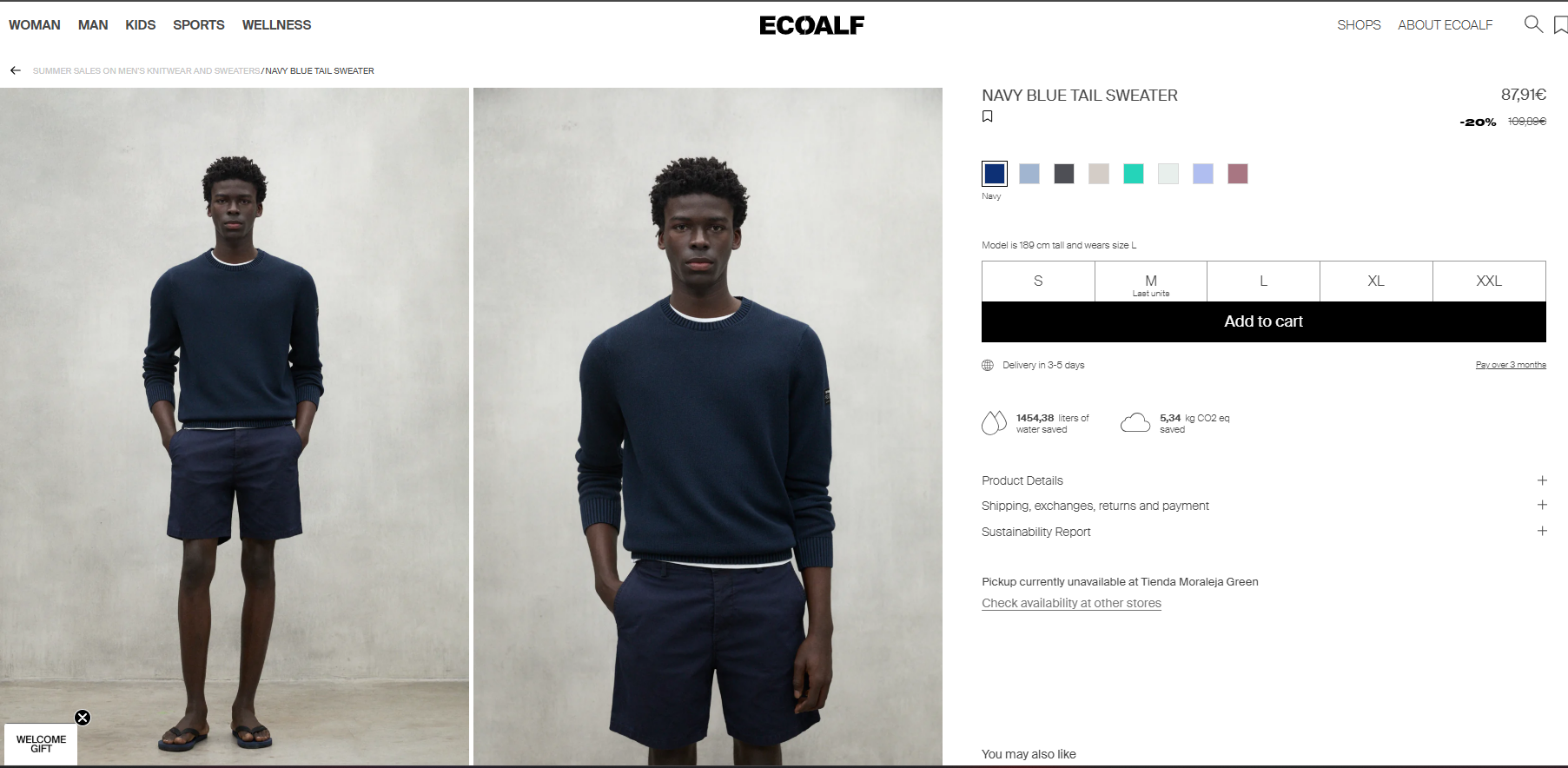
## 3. 데이터 전처리

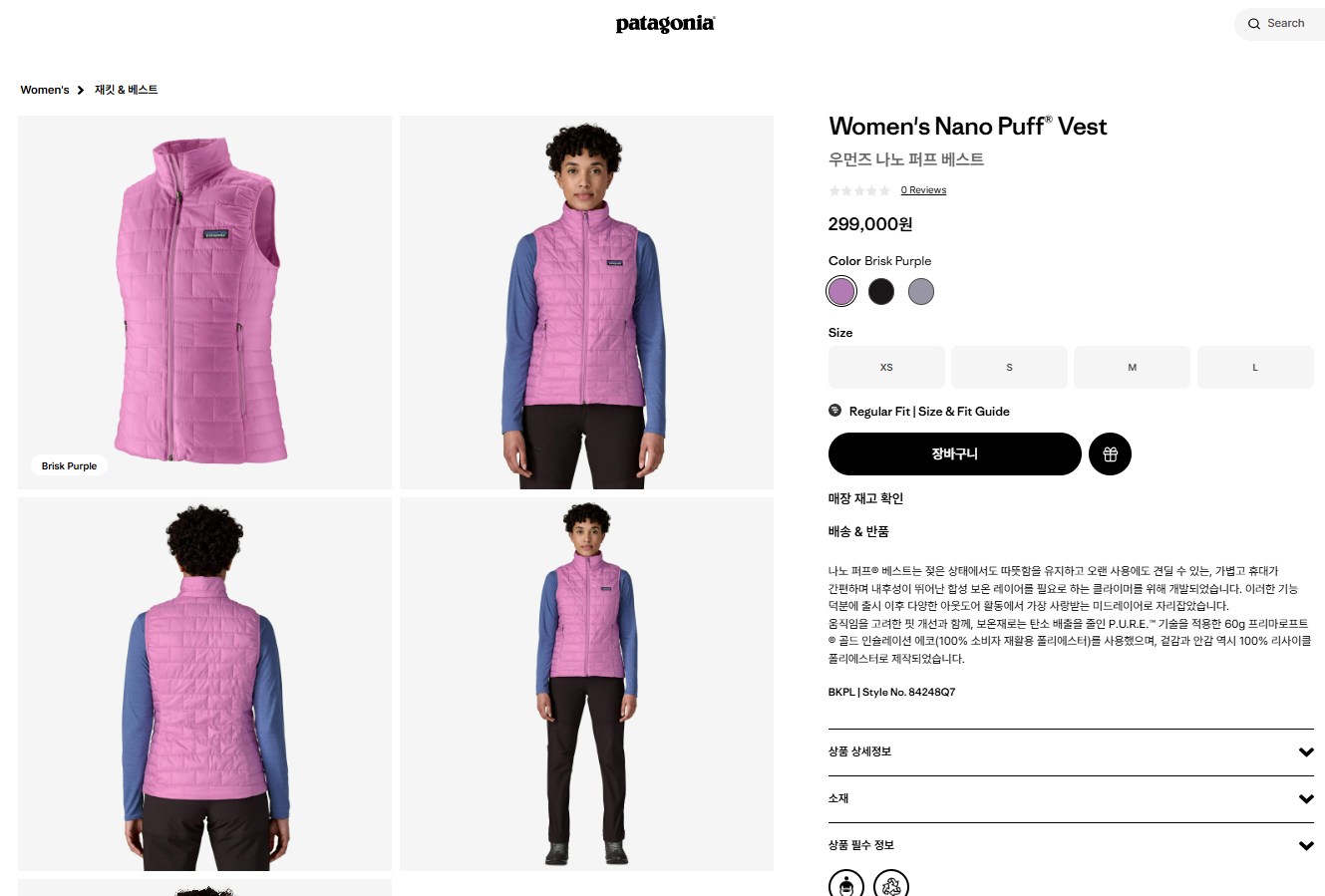
### 3.1 전처리 개요

* **데이터 수집**
* **데이터셋 구성**
* **데이터 정제**

### 3.2 데이터 수집

* **Selenium을 활용한 웹 크롤링으로 데이터 수집**
* **각 브랜드 사이트에서 제품마다 상세정보 수집**
* **수집 내용: 제품 명, 제품 이미지, 색상, 상세정보, 소재, 가격, 지속가능성 보고서 등**

****

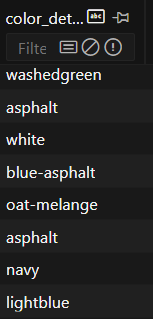
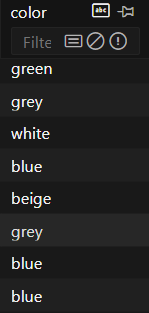
****

### 3.3 데이터셋 구성

* **데이터셋 구성**
  + **웹 크롤링 후 데이터셋에 저장된 컬럼**
    - **id**
    - **제품명 (name)**
    - **제품 이미지 (image\_url)**
    - **제품 정보 (spec)**
    - **소재 (material)**
    - **esg 보고서(esg\_report\_url)**
    - **제품 고유 아이디 (external\_id)**
    - **제품 사이트 url (url)**
    - **제품 상세정보 - 제품설명 (details\_intro)**
    - **제품 상세정보 - 상세정보 (details\_bullets)**
    - **제품 상세정보 - 재료 (composition\_parts)**
    - **핏 (fit)**
    - **지속가능성 아이콘 (sustainable\_icons)**
    - **지속가능성 정보 (sustainable\_detail)**
    - **제품의 영향 (impact)**
    - **가격 (price)**
    - **원산지(made\_in)**
    - **색상(color)**
    - **카테고리 (category)**

### 3.4 데이터 정제

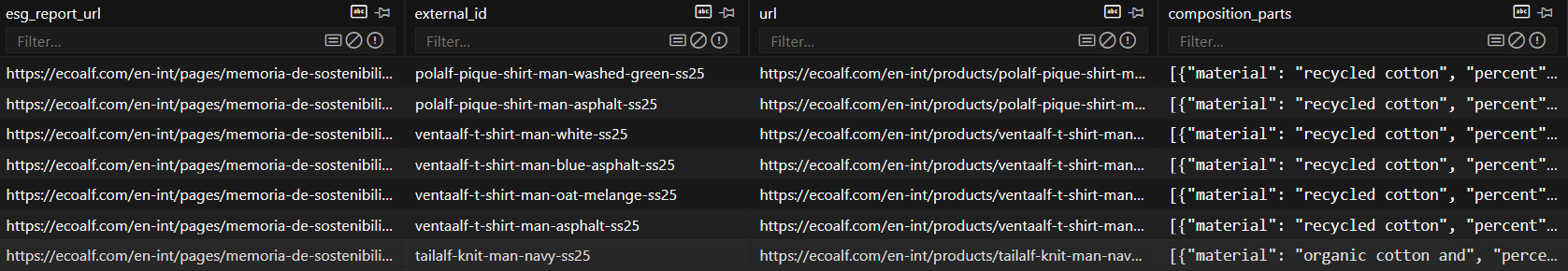
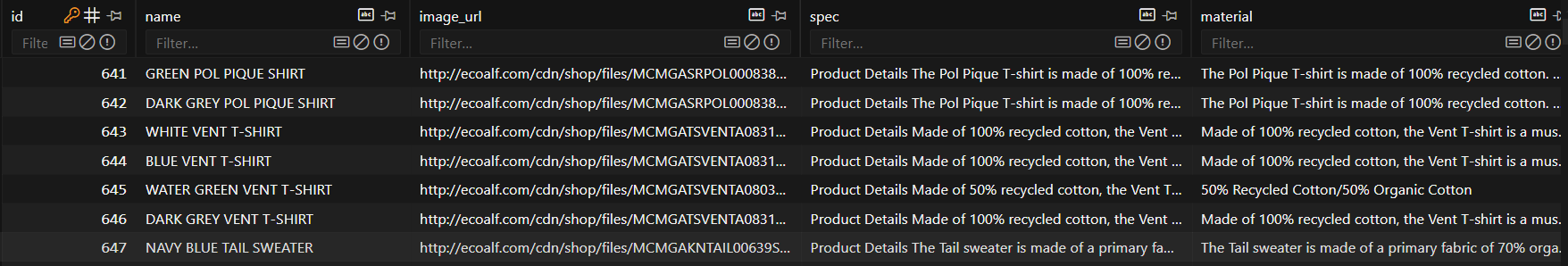
* **데이터 정제 방식**
  + **컬럼 추가**
    - **brand\_id : 추후 생성될 brand 테이블과 참조될 FK**
    - **색상 컬럼의 경우 세분화된 색상의 경우가 많아 기존 color 컬럼을 color\_detail로 변경.**
    - **color 컬럼에는 제품명과 제품정보를 참고하여 대분류로 지정**

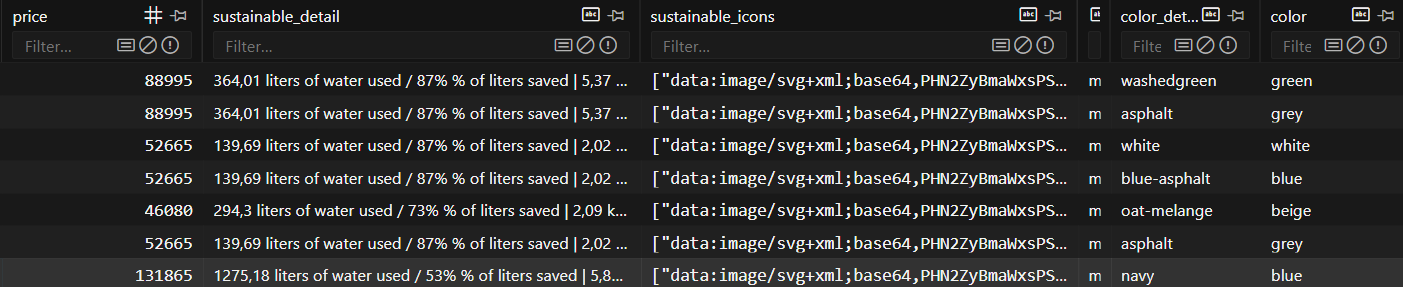
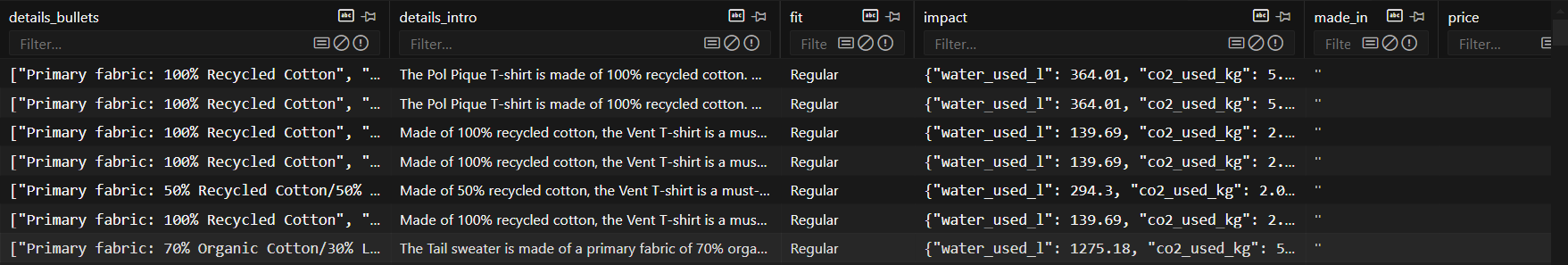
****

* + - **currency : ecoalf는 유로, patagonia는 원화이므로 통화 컬럼을 만들어 가격을 직관적으로 알 수 있게 처리**
  + **컬럼 제거**
    - **불필요한 정보 및 다른컬럼과 겹치는 내용이 있는 컬럼 제거**
      * **esg\_report\_url**
      * **details\_intro, details\_bullets, composition\_parts**
      * **made\_in**
      * **fit**
      * **sustainable\_icon, sustainable\_detail, impact**
  + **컬럼 통합 및 데이터 정제**
    - **impact 컬럼의 경우 물과 co2의 사용량과 절약량이 같이 있어 water\_saved\_l와 co2\_saved\_kg 컬럼으로 분할**
    - **material의 경우 유니코드화 된 부분을 한글로 변환**

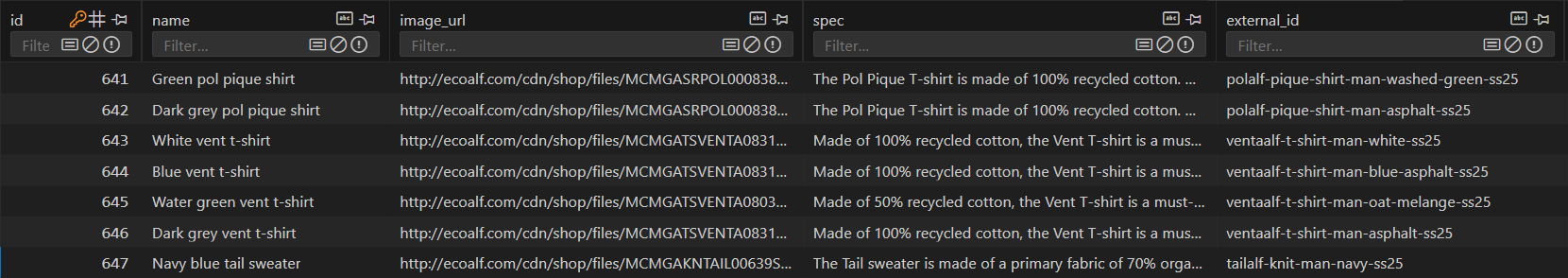
### 3.5 데이터셋 정제 후 결과

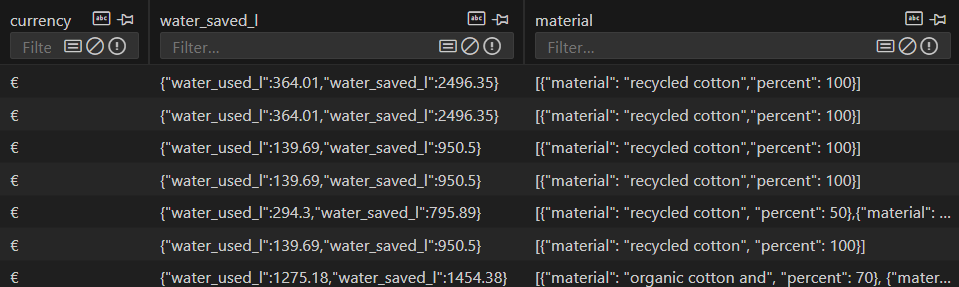
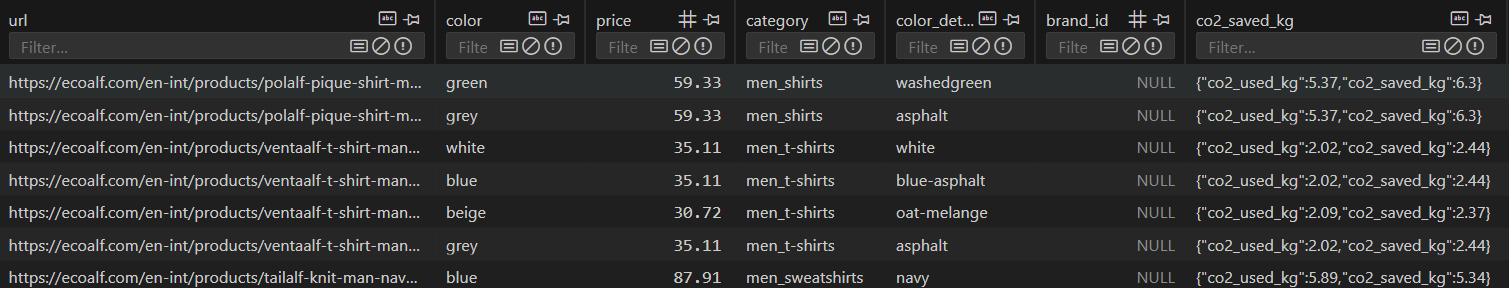
* **정제 전 데이터셋**

****

****

* **정제 후 데이터셋**

****

****

## 4. 향후 사용 계획

* **제품DB는 Pinecone을 사용하여 벡터DB화 시켜서 사용할 예정. 유사도 기반 검색이 기존 RDB보다 벡터DB가 유리하기 때문이다. 또한 다른 DB테이블과 FK로 연결되어 참조하는 경우 RDS를 사용하여 관리할 예정.**