추천 시스템 설계하기 [프로젝트]

▼ **데이터셋 소개**

데이터셋 링크: Amazon Sales Dataset

▼ **데이터셋 구조**

필드명	설명
product_id	제품의 고유 식별자
product_name	제품의 이름
category	제품이 속한 카테고리 정보
discounted_price	할인된 가격
actual_price	정가
discount_percentage	할인율
rating	제품의 평균 평점
rating_count	제품에 대한 총 평점 수
about_product	제품에 대한 간단한 설명
user_id	리뷰를 작성한 사용자의 고유 식별자 (쉼표로 구분된 여러 사용자)
user_name	리뷰를 작성한 사용자 이름 (쉼표로 구분된 여러 사용자)
review_id	리뷰의 고유 식별자 (쉼표로 구분된 여러 리뷰)
review_title	리뷰 제목 (쉼표로 구분된 여러 리뷰 제목)
review_content	리뷰 내용 (쉼표로 구분된 여러 리뷰 내용)
img_link	제품 이미지 URL
product_link	제품 페이지 URL

데이터 샘플

첫 번째 상품 예시:

- **제품명**: Wayona Nylon Braided USB to Lightning Fast Charging Cable
- **카테고리**: Computers&Accessories|Accessories&Peripherals|Cables&Accessories|Cables|USBCables
- **할인가**: ₹399.00
- **정가**: ₹1,099.00
- **할인율**: 64%
- **평점**: 4.2
- **리뷰 수**: 24,269

▼ 최고 할인율 기반, 할인율이 가장 높은 상품 10개 찾기

```
SELECT
  product_name,
  discounted_price,
  CAST(REPLACE(REPLACE(CAST(discounted_price AS STRING), '₹', ''), ',', '') AS FLOAT64) A
FROM `modulabs-project-465302.SQL_project.amazon_sales_dataset`
ORDER BY
  CAST(REPLACE(CAST(discount_percentage AS STRING), '%', '') AS FLOAT64) DESC
LIMIT 10;
```

행	product_name	rating	rating_count
1	AmazonBasics Flexible Premium HDMI Cable (Black, 4K@60Hz, 18Gbps), 3-Foot	4.4	426973
2	Amazon Basics High-Speed HDMI Cable, 6 Feet - Supports Ethernet, 3D, 4K video,Black	4.4	426973
3	Amazon Basics High-Speed HD...	4.4	426973
4	AmazonBasics Flexible Premium HDMI Cable (Black, 4K@60Hz, 18Gbps), 3-Foot	4.4	426972
5	boAt Bassheads 100 in Ear Wired Earphones with Mic(Taffy Pink)	4.1	363713
6	boAt Bassheads 100 in Ear Wired Earphones with Mic(Furious Red)	4.1	363713
7	boAt BassHeads 100 in-Ear Wir...	4.1	363711
8	Redmi 9A Sport (Coral Green, 2GB RAM, 32GB Storage) | 2GHz Octa-core Helio G25	4.1	313836

1. SELECT 절

```
SELECT
  product_name,
  discounted_price,
  ...
```

- `product_name`: 상품 이름
- `discounted_price`: 할인된 가격
- `...`: 아래에서 설명할 **숫자로 변환된 가격** (계산용 `score` 컬럼)

2. `CAST(REPLACE(...)) AS score`

```
CAST(REPLACE(REPLACE(CAST(discounted_price AS STRING), '₹', ''), ',', '') AS FLOAT64) AS score
```

- `discounted_price`는 원래 `₹13,999` 같은 **문자 형태**임
- 이걸 숫자로 바꾸려면:
  1. `CAST(... AS STRING)`: 먼저 문자열로 변환
  2. `REPLACE(..., '₹', '')`: ₹ 기호 제거
  3. `REPLACE(..., ',', '')`: 쉼표 제거
  4. `CAST(... AS FLOAT64)`: 숫자로 변환

👉 이 과정을 통해 `₹13,999` → `13999.0`으로 바뀜

3. `ORDER BY` 절

```
ORDER BY
  CAST(REPLACE(CAST(discount_percentage AS STRING), '%', '') AS FLOAT64) DESC
```

- `discount_percentage` 값은 `'64%'`, `'43%'` 같은 형태
- 이걸 숫자 비교하려면:
  1. 문자열로 변환
  2. `%` 기호 제거
  3. 숫자(Float)로 변환
- `DESC`: **할인율 높은 순서로 정렬**

▼ 리뷰 수가 많은 인기 제품 추천

```
SELECT
  product_name,
  rating,
  rating_count
FROM `modulabs-project-465302.SQL_project.amazon_sales_dataset`
WHERE
  REGEXP_CONTAINS(rating, r'^[0-9.]+$')
  AND CAST(rating AS FLOAT64) >= 4.0
ORDER BY
  CAST(REPLACE(CAST(rating_count AS STRING), ',', '') AS INT64) DESC
LIMIT 10;
```

1	B0B4KPCBSH	IKEA Frother for Milk	Home&Kitchen|Kitchen&HomeAppliances|Coffee,Tea&Espresso|CoffeeGrinders|ElectricGrinders	244.0
2	B009LJ2BXA	Hp Wired On Ear Headphones With Mic With 3.5 Mm Drivers, In-Built Noise Cancelling, Foldable And Adjustable For Laptop/Pc/Office/Home/ 1	Computers&Accessories|Accessories&Peripherals|Audio&VideoAccessories|PCHeadsets	649.0
3	B00MFPCY5C	GIZGA essentials Universal Silicone Keyboard Protector Skin for 15.6-inches Laptop (5 x 6 x 3 inches)	Computers&Accessories|Accessories&Peripherals|Keyboards,Mice&InputDevices|Keyboard&MiceAccessories|DustCovers	39.0
4	B097JVLW3L	Irusu Play VR Plus Virtual Reality Headset with	Electronics|HomeTheater,TV&Vi...	2699.0

1. `SELECT` 절

```
SELECT product_name, rating, rating_count
```

- 출력할 컬럼:
  - `product_name`: 상품 이름
  - `rating`: 평점
  - `rating_count`: 리뷰 수

2. `WHERE` 절

```
WHERE
  REGEXP_CONTAINS(rating, r'^[0-9.]+$')
  AND CAST(rating AS FLOAT64) >= 4.0
```

- `REGEXP_CONTAINS(rating, r'^[0-9.]+$')`

  → **숫자와 소수점(.)만 포함된 평점 값만 필터링**

  → 예외 값들 (예: `'|'`, `'N/A'`, `null`) 제거

- `CAST(rating AS FLOAT64) >= 4.0`

  → 필터링된 값 중에서 **평점이 4.0 이상**인 데이터만 선택

📌 이 조합을 통해 쿼리가 실패하지 않고 안전하게 실행됨

3. `ORDER BY` 절

```
ORDER BY
  CAST(REPLACE(CAST(rating_count AS STRING), ',', '') AS INT64) DESC
```

- `rating_count`는 `"43,994"`처럼 문자열로 되어 있으므로:
  1. `CAST(... AS STRING)` → 문자열로 변환
  2. `REPLACE(..., ',', '')` → 쉼표 제거
  3. `CAST(... AS INT64)` → 정수로 변환
- `DESC`: 리뷰 수가 많은 순으로 내림차순 정렬

▼ 카테고리별 베스트 추천

```
SELECT *
FROM (
   SELECT *,
        ROW_NUMBER() OVER (
           PARTITION BY category
           ORDER BY CAST(rating AS FLOAT64) DESC
        ) AS rank
   FROM `modulabs-project-465302.SQL_project.amazon_sales_dataset`
   WHERE REGEXP_CONTAINS(rating, r'^[0-9.]+$') -- 숫자/소수점만 있는 평점만 필터링
) ranked
WHERE rank = 1;
```

	product_id	product_name	category	discounted_price
1	B0B4KPCBSH	IKEA Frother for Milk	Home&Kitchen|Kitchen&HomeAppliances|Coffee,Tea&Espresso|CoffeeGrinders|ElectricGrinders	244.0
2	B009LJ2BXA	Hp Wired On Ear Headphones With Mic With 3.5 Mm Drivers, In-Built Noise Cancelling, Foldable And Adjustable For Laptop/Pc/Office/Home/ 1	Computers&Accessories|Accessories&Peripherals|Audio&VideoAccessories|PCHeadsets	649.0
3	B00MFPCY5C	GIZGA essentials Universal Silicone Keyboard Protector Skin for 15.6-inches Laptop (5 x 6 x 3 inches)	Computers&Accessories|Accessories&Peripherals|Keyboards,Mice&InputDevices|Keyboard&MiceAccessories|DustCovers	39.0
4	B097JVLW3L	Irusu Play VR Plus Virtual Reality Headset with	Electronics|HomeTheater,TV&Vi...	2699.0

1. `REGEXP_CONTAINS(rating, r'^[0-9.]+$')`

- `rating` 값이 숫자(예: `4.3`, `5.0`)로만 이루어진 경우만 선택합니다.
- 이유: 데이터에 `'|'` 같은 이상한 문자가 섞여 있어 에러 방지용입니다.

---

2. `ROW_NUMBER() OVER (...) AS rank`

```
ROW_NUMBER() OVER (
  PARTITION BY category
  ORDER BY CAST(rating AS FLOAT64) DESC
)
```

- `category` 별로 데이터를 **그룹(분리)** 합니다.
- 각 그룹 안에서 **평점(rating)이 높은 순서대로 번호(rank)** 를 매깁니다.

---

3. `WHERE rank = 1`

- 각 카테고리에서 **순위가 1인 상품**, 즉 **평점이 가장 높은 상품**만 골라냅니다.

▼ 가성비 추천 (평점/가격 비율 기준)

```
SELECT product_name, rating, discounted_price
FROM `modulabs-project-465302.SQL_project.amazon_sales_dataset`
WHERE
  REGEXP_CONTAINS(rating, r'^[0-9.]+$') AND
```

```
  rating IS NOT NULL AND
  discounted_price IS NOT NULL
ORDER BY
  CAST(rating AS FLOAT64) /
  CAST(REPLACE(REPLACE(CAST(discounted_price AS STRING), '₹', ''), ',', '') AS FLOAT64) D
LIMIT 10;
```

행	product_name	rating	discounted_price
1	E-COSMOS 5V 1.2W Portable Flexible USB LED Light (Colours May Vary, Small, EC-POF1)	3.8	39.0
2	Inventis 5V 1.2W Portable Flexible USB LED Light Lamp (Colors may vary)	3.6	39.0
3	GIZGA essentials Universal Silicone Keyboard Protector Skin for 15.6-inches Laptop (5 x 6 x 3 inches)	3.5	39.0
4	Classmate Octane Neon- Blue Gel Pens(Pack of 5)|Smooth Writing Pen|Attractive body colour for Boys & Girls|Waterproof ink for	4.3	50.0
5	FLiX (Beetel Flow USB to	4.0	57.89

1. WHERE 절

```
WHERE
  REGEXP_CONTAINS(rating, r'^[0-9.]+$') AND
  rating IS NOT NULL AND
  discounted_price IS NOT NULL
```

- 평점 값이 **정상 숫자 형식(예: 4.2, 3.8)** 인 경우만 선택
- 평점 또는 할인 가격이 NULL인 상품은 제외

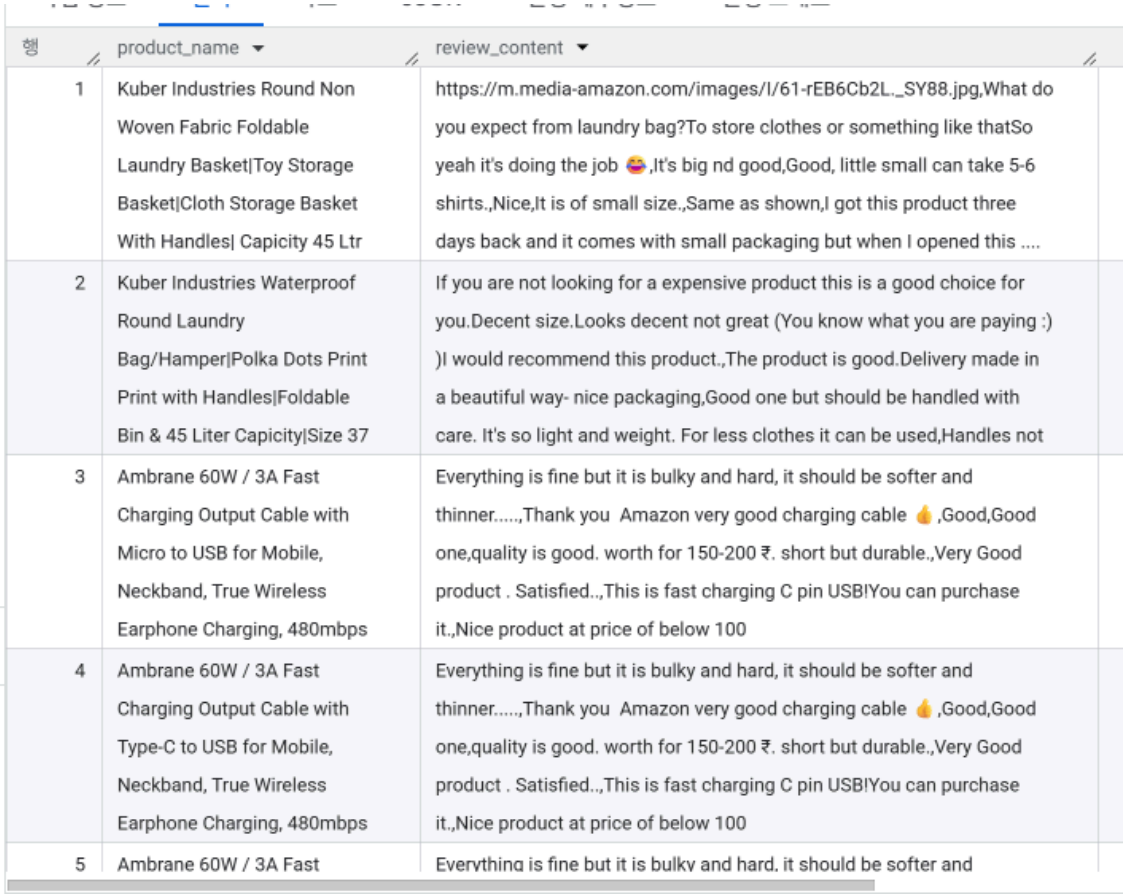
2. ORDER BY 절

```
ORDER BY
  CAST(rating AS FLOAT64) /
  CAST(REPLACE(REPLACE(CAST(discounted_price AS STRING), '₹', ''), ',', '') AS FLOAT64) D
```

- `rating`을 숫자로 변환
- `discounted_price`에서:
  - `₹` 기호 제거
  - `,` 쉼표 제거
  - → 숫자로 변환
- 평점을 가격으로 나눈 값이 큰 순서대로 정렬 (**가성비 점수**)

▼ 리뷰 내용 기반 키워드 추천

```
SELECT product_name, review_content
FROM `modulabs-project-465302.SQL_project.amazon_sales_dataset`
WHERE
  REGEXP_CONTAINS(rating, r'^[0-9.]+$') AND
  CAST(rating AS FLOAT64) >= 4.0 AND
  LOWER(review_content) LIKE '%durable%'
LIMIT 10;
```



행	product_name	review_content
1	Kuber Industries Round Non Woven Fabric Foldable Laundry Basket|Toy Storage Basket|Cloth Storage Basket With Handles| Capicity 45 Ltr	https://m.media-amazon.com/images/I/61-rEB6Cb2L._SY88.jpg,What do you expect from laundry bag?To store clothes or something like thatSo yeah it's doing the job 😂,It's big nd good,Good, little small can take 5-6 shirts.,Nice,It is of small size.,Same as shown,I got this product three days back and it comes with small packaging but when I opened this ....
2	Kuber Industries Waterproof Round Laundry Bag/Hamper|Polka Dots Print Print with Handles|Foldable Bin & 45 Liter Capicity|Size 37	If you are not looking for a expensive product this is a good choice for you.Decent size.Looks decent not great (You know what you are paying :))I would recommend this product.,The product is good.Delivery made in a beautiful way- nice packaging,Good one but should be handled with care. It's so light and weight. For less clothes it can be used,Handles not
3	Ambrane 60W / 3A Fast Charging Output Cable with Micro to USB for Mobile, Neckband, True Wireless Earphone Charging, 480mbps	Everything is fine but it is bulky and hard, it should be softer and thinner.....,Thank you Amazon very good charging cable 👍,Good,Good one,quality is good. worth for 150-200 ₹. short but durable.,Very Good product . Satisfied..,This is fast charging C pin USB!You can purchase it.,Nice product at price of below 100
4	Ambrane 60W / 3A Fast Charging Output Cable with Type-C to USB for Mobile, Neckband, True Wireless Earphone Charging, 480mbps	Everything is fine but it is bulky and hard, it should be softer and thinner.....,Thank you Amazon very good charging cable 👍,Good,Good one,quality is good. worth for 150-200 ₹. short but durable.,Very Good product . Satisfied..,This is fast charging C pin USB!You can purchase it.,Nice product at price of below 100
5	Ambrane 60W / 3A Fast	Everything is fine but it is bulky and hard, it should be softer and

1. `WHERE` 절

```
WHERE
  REGEXP_CONTAINS(rating, r'^[0-9.]+$') AND
  CAST(rating AS FLOAT64) >= 4.0 AND
  LOWER(review_content) LIKE '%durable%'
```

- rating 값이 숫자나 소수점으로만 이루어진 경우만 필터링
  → | 같은 이상한 값 걸러내기 (에러 방지용)
- 평점을 숫자로 변환하고, 4.0 이상인 상품만 선택
- 리뷰 내용에 "durable" 이라는 단어(소문자 기준)가 포함된 경우
  → 즉, 사용자들이 "튼튼하다", "잘 안 망가진다" 는 말을 직접 했다는 뜻