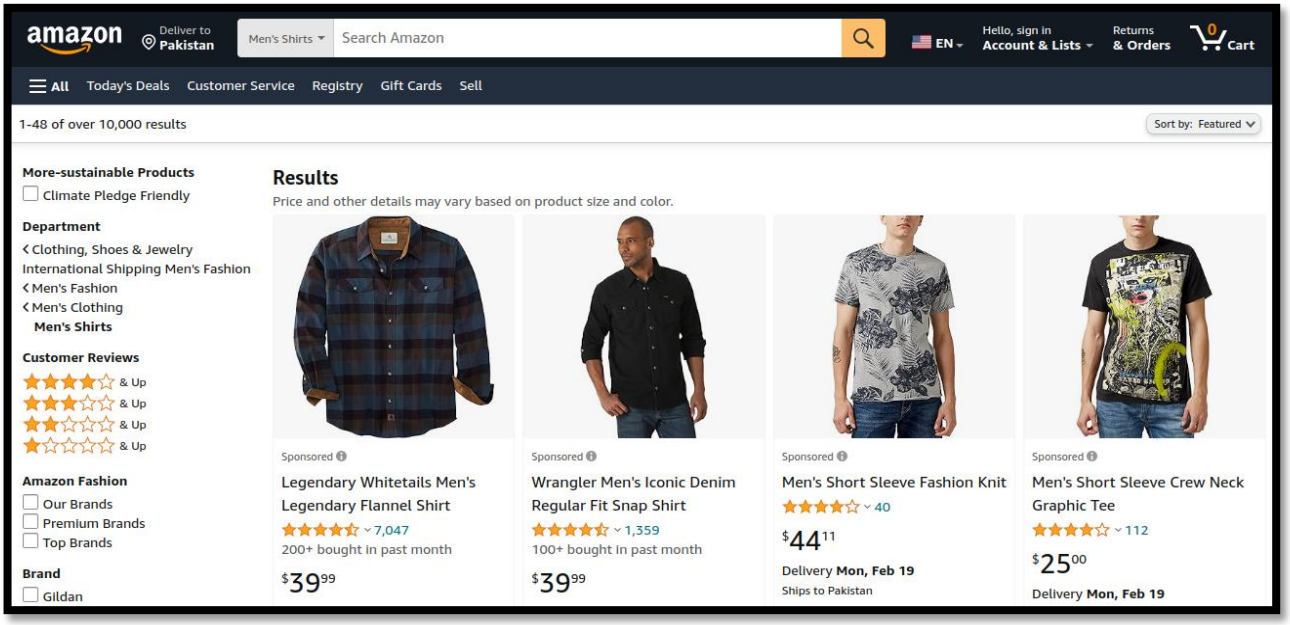
DS-3002 DATA MINING (DS) Spring, 2024

Assignment 01 [Web Scraping]

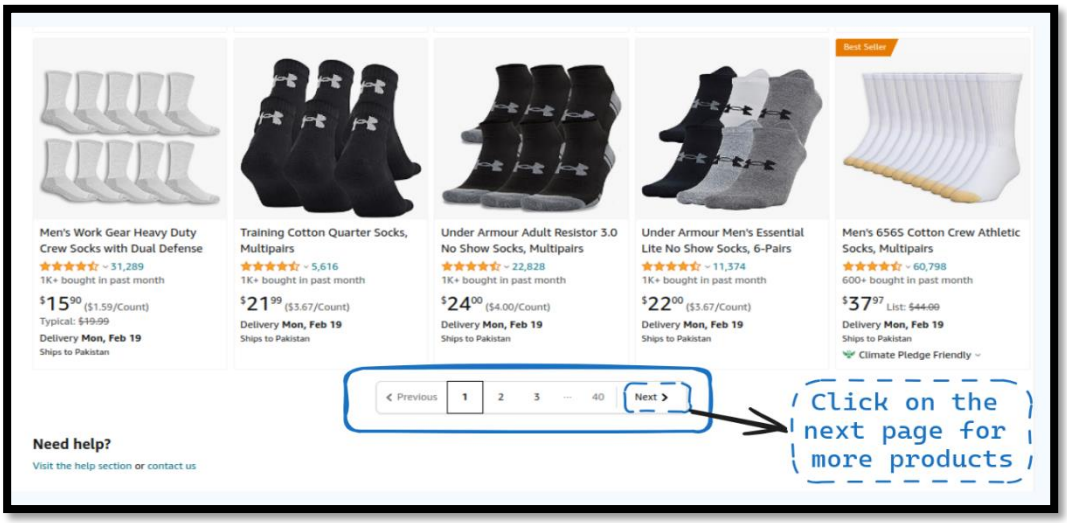
Due Date: 21- Feb -2024



Project Overview

- Imagine you're on a mission to learn all about what's being sold on Amazon, the giant online store. You want to gather detailed info about the products (**under the 'Fashion' Category only**). They offer so you can understand what people are buying and what's popular.
- To do this, you need to use a technique called web scraping, which means collecting data from websites. But there's a catch – you have to follow Amazon's rules about how you can use their data.
- Your goal is to create a big collection of product information in a format called JSON. This format is easy for computers to understand and work with, so it's great for analyzing data.
- Each piece of info about a product, like its name, price, and description, is neatly organized in this JSON format. Once you've gathered all this data, you can study it to see what kinds of things people are interested in buying on Amazon.

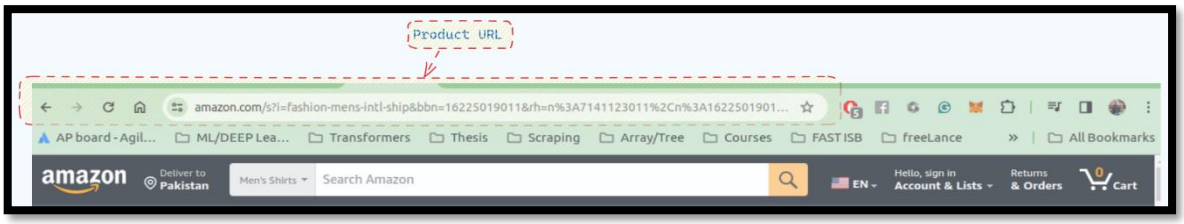
- You have to scrap at least 1000 products from each category. By category, it meant i.e. shoes, clothes, watches, etc.



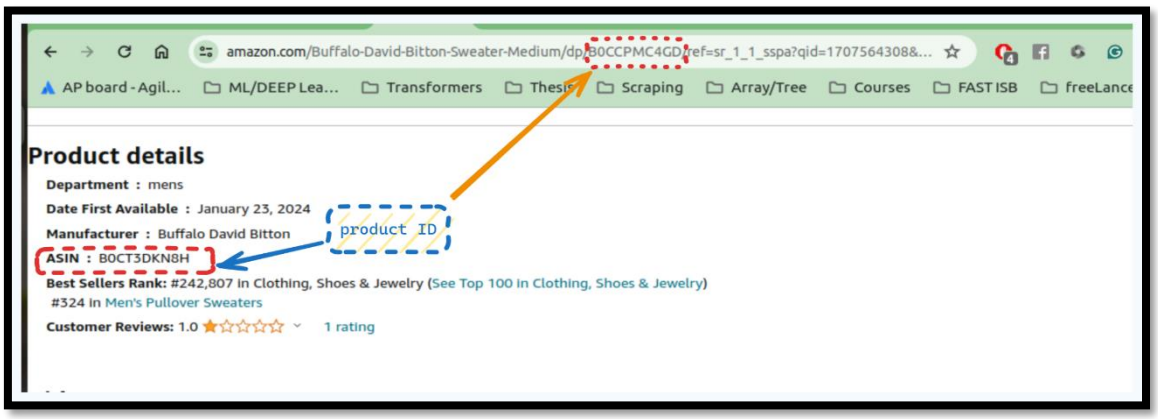
Detail Description

Here's a simple breakdown of what each part of your data will include for each product:

1. **Product Url**: This is the web address for the product on Amazon. It lets you find the product online easily. You'll save this URL in your database so you can refer back to the product page whenever needed.

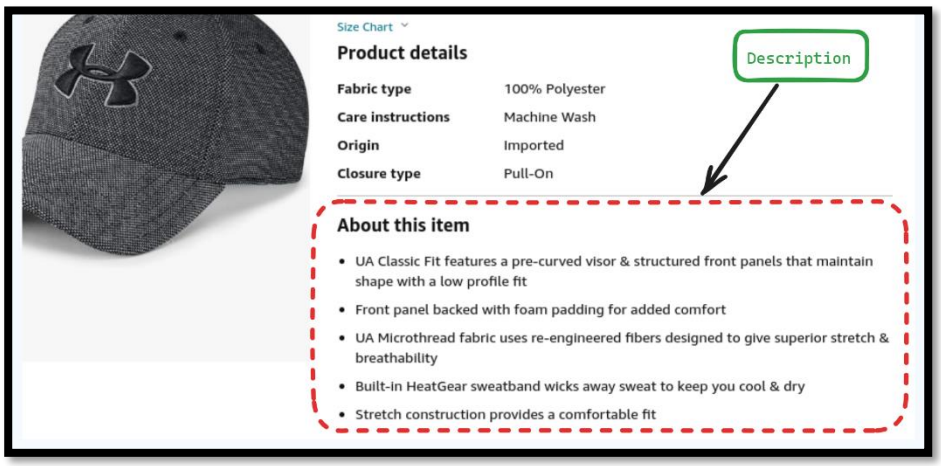


2. **Product ID**: Each product on Amazon has a unique code. This helps you tell products apart in your database. If the product ID is not in the product details section fetch it from the URL (using string processing).

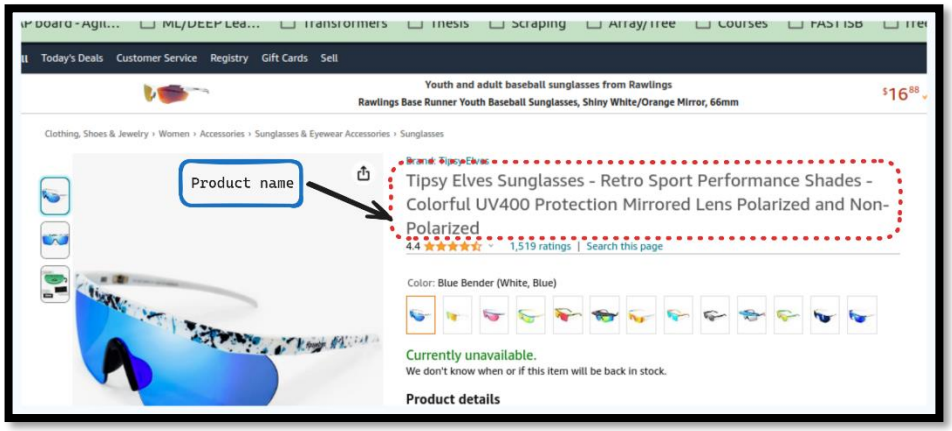


3. **Gender**: You have to collect the gender while scraping the product, you can maintain it while exploring the category. If a product is not based on gender keep the gender filed null. You can fetch it from subcategory.

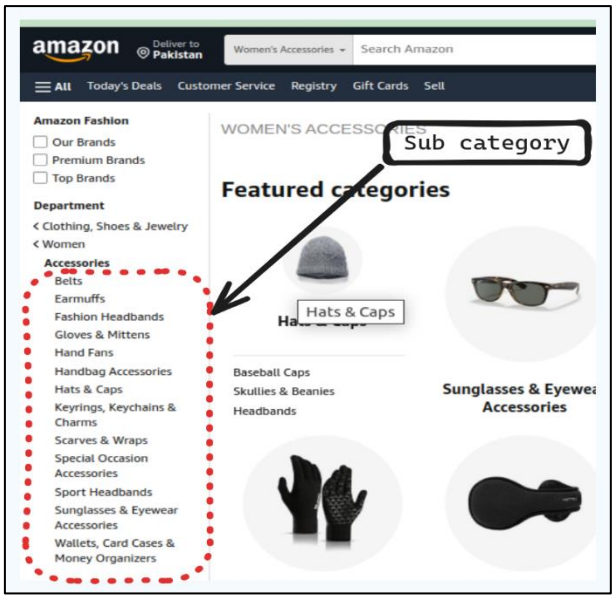
4. **Category**: A broad group the product fits into, such as "Accessories." This helps classify products into big groups for easier analysis.
5. **Description**: This is a detailed write-up about what the product is, its features, and why it's special. This part gives you a lot of information about the product to analyse and understand it better.



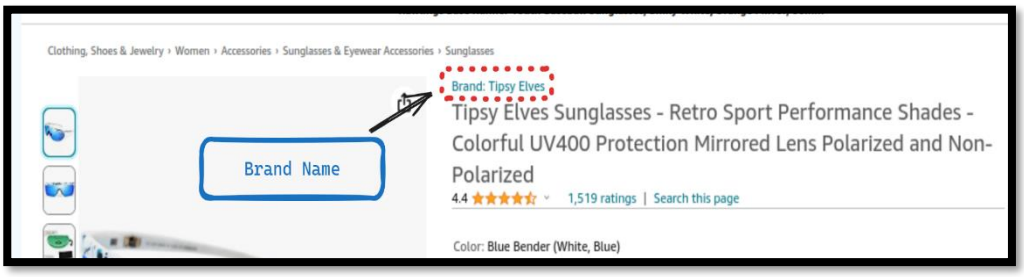
6. **Product Name:** The official name of the product given by the seller on Amazon. It's important to identify the product in your database.



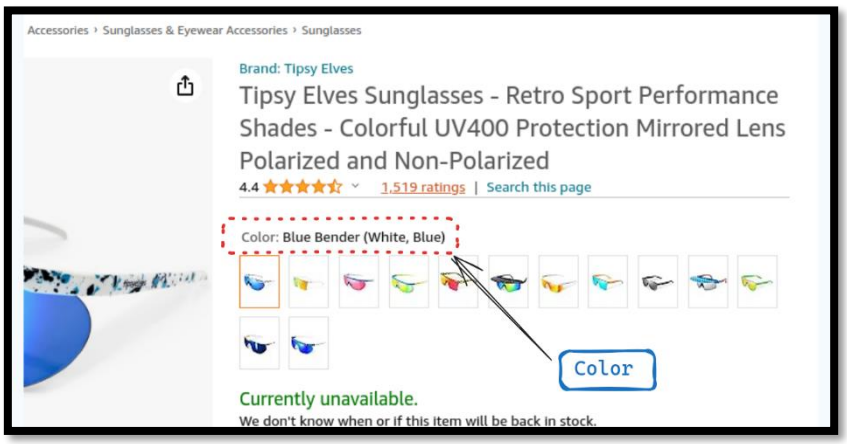
7. **Sub Category**: This is a more specific group within the main category, like "Sunglasses & Eyewear Accessories." It helps you sort the products even more finely. This might be a static process because you have to store the sub-categories in the list use the “fuzzywuzzy” and pass the **product name** to this module it will return the most relevant sub-category.



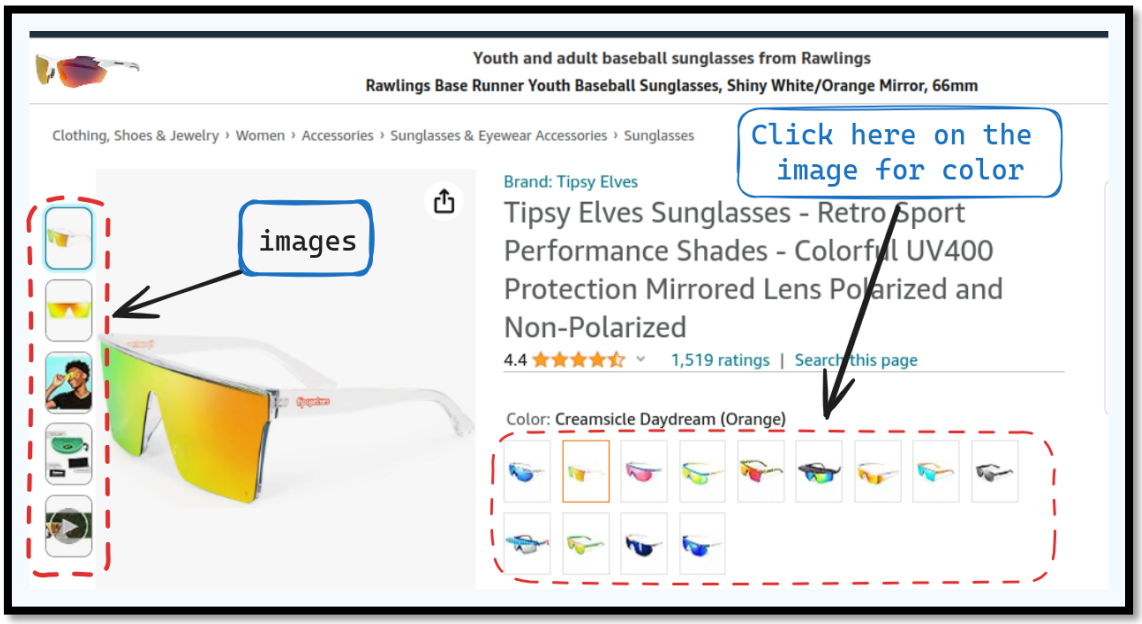
8. **Brand Name:** The name of the company that makes or sells the product. This information can be useful for seeing which brands are popular or comparing products from the same brand.



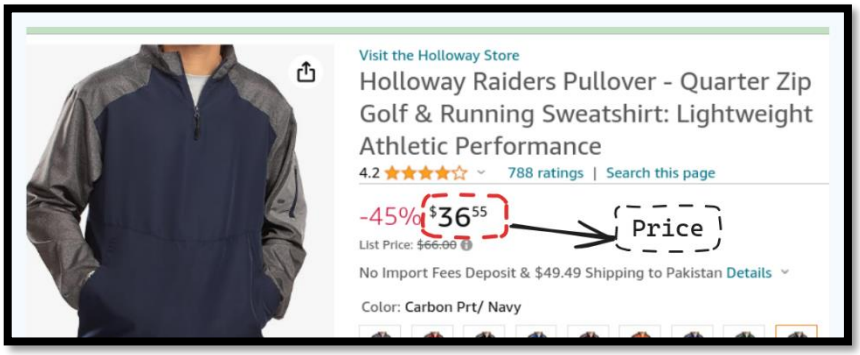
9. **Colours**: This part lists all the different color options for the product, along with pictures and prices for each color. You'll organize this information in your database in a way that shows all the options for each product.



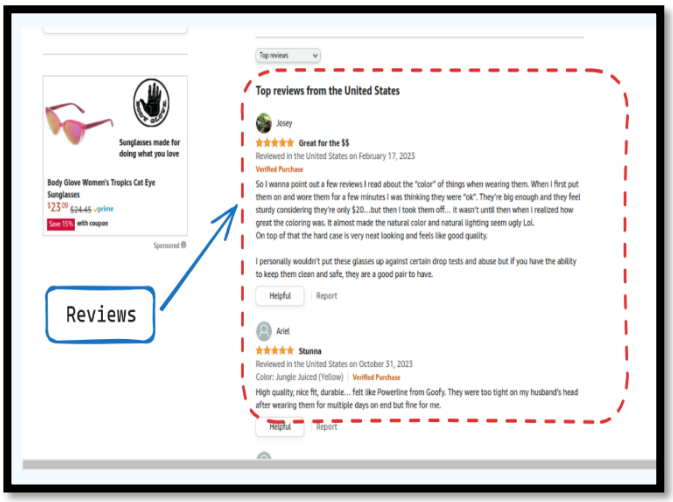
10. **Images**: Links to pictures of the product. These help you see what the product looks like and will be saved in your database in the form URL (string)



11. **Price**: How much does the product cost, including any sale prices? This is crucial for analyzing how prices change or comparing the cost of different products. Also, when color of the product changes it price changes. If the price is not currently available then store null.



12. **Product reviews along with ranking:** Opinions from people who bought the product. This includes what they thought and how many stars they gave it. You'll use this to understand how well-liked a product is. Collect all reviews from all pages by clicking on more "reviews". For ranking, save 5 for five stars and 1 for one star.



TODOS

1. Scrape all the above information of at least 10 categories from Amazon:
    - Use web scraping techniques to extract the URLs of at least different categories.
    - **1000 products** from each category.

2. Store the URLs of product images for later use:
    - Extract the URLs of product images while scraping product details.
    - Save these URLs in a structured format, such as a list or dictionary, to associate them with their respective products.
3. Consider organizing the image URLs based on their relevance, such as primary images, alternate views, or different color options.

4. Collect and store product reviews for analysis:
    - Navigate to the review section of each product page.
    - Extract the review text, star ratings, reviewer information, and other relevant details.
    - Handle pagination to scrape reviews from multiple pages if necessary

5. Decide whether to use Selenium, Beautiful Soup, or a combination of both for web scraping.

6. Handle cases where product availability, description, or certain entities are not present by storing them as `None` or `null`:
    - Implement error handling logic to check for missing data fields during scraping.
    - Assign `None` or `null` values to fields that are not available or relevant for certain products.
    - Ensure consistency in handling missing data across all scraped products to maintain the integrity of your dataset.
7. Ensure handle pagination if required to scrape multiple pages of products within each category:
    - Identify pagination elements on category pages and extract links to subsequent pages.
    - Implement logic to iterate through multiple pages and scrape product data from each page.
    - Handle cases where pagination controls are dynamic or hidden behind JavaScript interactions.

8. Implement error handling mechanisms to address issues encountered during scraping and database population:
    - Handle HTTP errors, timeouts, and other network issues gracefully during scraping.
    - Implement retry logic for failed requests to mitigate transient errors.

- Log errors and exceptions encountered during scraping for troubleshooting and analysis.

9. Document the scraping process and instructions for running the scripts for future reference and maintenance.
10. Validate the scraped data to ensure accuracy and completeness:
    - Perform data validation checks to ensure that scraped data meets expected criteria and standards.
    - Compare the scraped data with a sample of manually verified data to identify any discrepancies or errors.

11. Optimize the scraping process for efficiency, considering factors such as rate limiting, request headers, and proxy rotation if necessary:
    - Configure scraping scripts to respect Amazon's rate limits and scraping policies to avoid being blocked.
    - Use request headers to mimic legitimate user behaviour and avoid detection as a bot.
    - Implement proxy rotation to distribute scraping requests across multiple IP addresses and avoid IP-based blocking.

What to submit

Rollno_scraper.py file that will scrap the category if we pass the category URL and it shows the progress bar of scraped products and saves each product in a JSON file.

i.e. python scraper.py -- url_of_category

Sample JSON format:


```
{
 "productUrl": "https://www.amazon.com/Ray-Ban-Outdoorsman-II-Lenses-Non-
Polarized/dp/B015L2KLQY/ref=sr_1_252?qid=1676455018&s=fashion-mens-intl-ship&sr=1-252",
       "productId": "B0814DYCMR",
       "gender": "men",
       "category": "Accessories",
       "description": "RB3029 Outdoorsman II Aviator Sunglasses feature a metal frame and
classic pilot lenses. The Ray-Ban sunglasses are available in a variety of frame colors and
lens treatments., Ray-Ban sunglasses have appeared throughout hundreds of films and have been
a favorite on the Hollywood scene for years, both on and off the screen. With timeless and
imaginative styles, Ray-Ban consistently blends high-tech design, lenses, and materials. The
collection remains true to its classic heritage while continuously evolving to meet today's
fashion demands.",
       "subCategory": "Sunglasses & Eyewear Accessories",
       "productName": "Ray-Ban RB3029 Outdoorsman Ii Aviator Sunglasses",
       "brandName": " Ray-Ban Store",
       "colors": {
           "Gold/G-15 Green": {
               "images": [
```

```
                "https://m.media-amazon.com/images/I/21G2qZ08SUS._AC_SY355_.jpg",
                "https://m.media-amazon.com/images/I/21-D6JQnaDS._AC_SY355_.jpg",
                "https://m.media-amazon.com/images/I/21Nt5MrjVyS._AC_SY355_.jpg",
                "https://m.media-amazon.com/images/I/21ex86ONB5S._AC_SY355_.jpg",
                "https://m.media-amazon.com/images/I/21-M3O7zcRS._AC_SY355_.jpg",

            ],
            "price": [
                {
                    "salePrice": "$16"
                }
            ]
        },
        "Gold Frame/Green G-15xlt Lens": {
            "images": [
                "https://m.media-amazon.com/images/I/318EgfykcJL._AC_SY355_.jpg",
                "https://m.media-amazon.com/images/I/31ez39NAxXL._AC_SY355_.jpg"
            ],
            "price": [
                {
                    "salePrice": "$16"
                }
            ]
        },
        "Gold Rose/G-15 Green": {
            "images": [
                "https://m.media-amazon.com/images/I/3153X1gkPuL._AC_SY355_.jpg",
                "https://m.media-amazon.com/images/I/214KAYk+uXL._AC_SY355_.jpg",
                "https://m.media-amazon.com/images/I/21Tv8qs--zL._AC_SY355_.jpg"
            ],
            "price": [
                {
                    "salePrice": "$16"
                }
            ]
        }
    },
    "product reviews": [{
        "review":"Amazing product for glasses",
        "rating":4
    },
    {
        "review":"No genuine in  glasses category",
        "rating":2
    }]
}
```