Web Scraping and Analysis of Makeup Products from Sociolla

1. Introduction

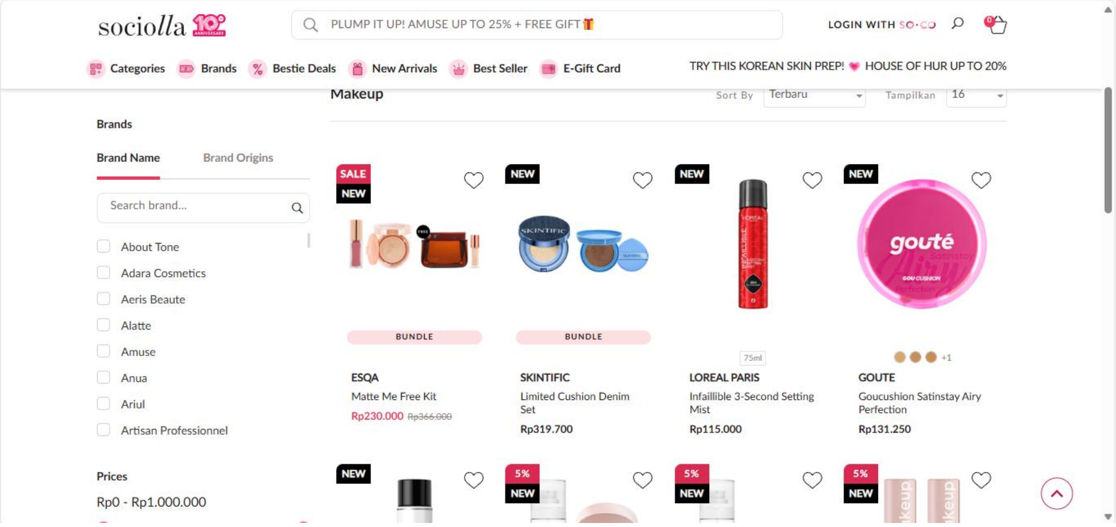
This project focuses on collecting and analyzing makeup product data from **Sociolla**, one of Indonesia's leading beauty e-commerce platforms. The main objective of this project is to **identify which cosmetic brands are most suitable to prioritize in promotional or marketing campaigns**, based on key metrics such as the number of reviews, product ratings, and product engagement efficiency.

To achieve this, I used **web scraping techniques** to extract product information directly from Sociolla's website. After gathering the data, I performed **data cleaning** to ensure it was structured and ready for analysis.

This report presents the entire process from data collection to insight-driven analysis, along with key findings that could help brands or marketers make better data-informed decisions.

2. Data Collection Method

a. Website Used:



The data was collected from the **Makeup category on the Sociolla website** (https://www.sociolla.com/) by scraping around 31 pages of product listings on **May 13, 2025**.

b. Tools and Libraries

The scraping, cleaning, and analysis process was carried out using the Python programming language, with the following libraries:

- **selenium** - to automate the browser and handle dynamic web pages

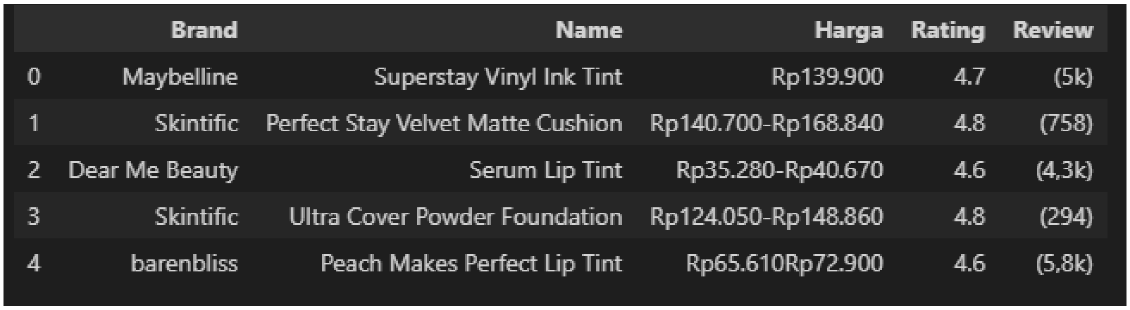
- **BeautifulSoup** - to parse HTML content and extract specific data
- **pandas** - to store, structure, and manipulate data in tabular format
- **time** - to manage delays between scraping actions to avoid being blocked

For analysis and data visualization, the following libraries were used:

- **pandasql -** to run SQL queries on pandas DataFrames for analytical purposes
- **seaborn -** to create informative and attractive statistical graphics
- **matplotlib.pyplot -** to build plots and charts for data exploration

c. Data Collected:

The scraped data included the following attributes:



	Brand	Name	Harga	Rating	Review
0	Maybelline	Superstay Vinyl Ink Tint	Rp139.900	4.7	(5k)
1	Skintific	Perfect Stay Velvet Matte Cushion	Rp140.700-Rp168.840	4.8	(758)
2	Dear Me Beauty	Serum Lip Tint	Rp35.280-Rp40.670	4.6	(4,3k)
3	Skintific	Ultra Cover Powder Foundation	Rp124.050-Rp148.860	4.8	(294)
4	barenbliss	Peach Makes Perfect Lip Tint	Rp65.610Rp72.900	4.6	(5,8k)

- **Brand Name** (Brand) - name of the manufacturer or brand
- **Product Name** (Name) - the specific name of the cosmetic product
- **Price** (Harga) – price of the produk
- **Rating** - customer rating score (out of 5)
- **Review** - total number of user reviews

3. Data Cleaning Process

After scraping the raw product data, I performed several data cleansing steps to ensure consistency and accuracy before conducting analysis. The main steps are as follows:

a. Splitting Price Information

Some products displayed a price range, indicating both the original price and the discounted price. I separated these into two distinct columns:

- Harga_Asli (Original Price)
- Harga_Diskon (Discounted Price)

b. Standardizing the Review Column

The “**Review”** column contained inconsistent formats, such as "5k" or "4,3k", which needed to be standardized. I converted all values by:

- Replacing 'k' with '000'
- Converting the resulting string to an integer or float (e.g., '4,3k' became 4300)

This made it easier to perform numerical analysis on review counts.

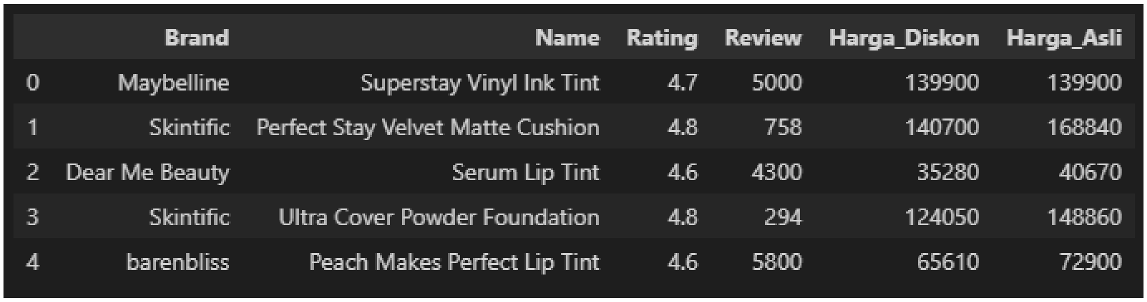
c. Converting Data Types

Several columns, including "**Rating**", **"Harga_Asli", "Harga_Diskon"**, were initially in object (string) format due to symbols and formatting. I cleaned the data by:

- Removing currency symbols (Rp) and punctuation
- Converting Rating and prices into float and integer type for further statistical analysis

Below is an example of the transformation:

Before Cleansing	**After Cleansing**
Rp140.700–Rp168.840	Harga_Diskon: 140700.0 Harga_Asli: 168840.0
4,3k	4300
4.7 (as object)	4.7 (as float)

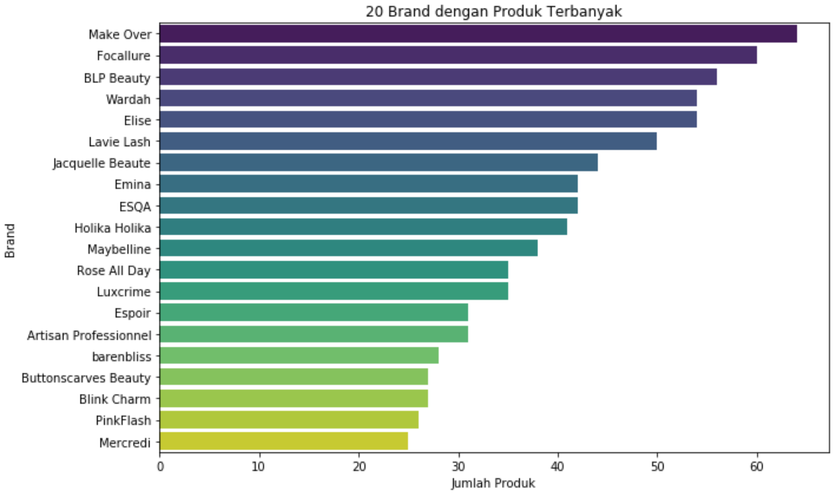


	Brand	Name	Rating	Review	Harga_Diskon	Harga_Asli
0	Maybelline	Superstay Vinyl Ink Tint	4.7	5000	139900	139900
1	Skintific	Perfect Stay Velvet Matte Cushion	4.8	758	140700	168840
2	Dear Me Beauty	Serum Lip Tint	4.6	4300	35280	40670
3	Skintific	Ultra Cover Powder Foundation	4.8	294	124050	148860
4	barenbliss	Peach Makes Perfect Lip Tint	4.6	5800	65610	72900

4. Data Analysis and Visualization

After cleaning and preparing the data, I conducted several analyses to identify which brands stand out and which ones might be prioritized for future promotional or marketing campaigns. Below are the key insights:

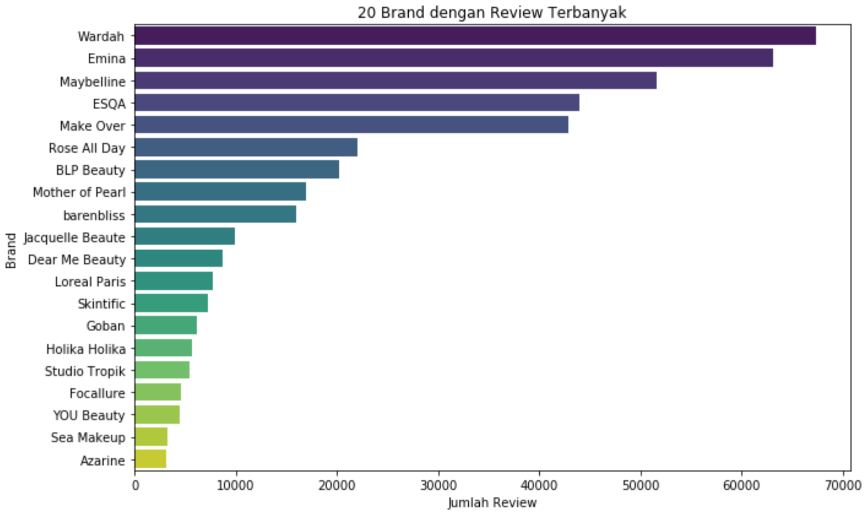
a. Top Brands by Number of Products



Using a count of unique product names per brand, I identified the brands with the most diverse product offerings. From the visualization **Top Brands are Makeover, Focallure, and**

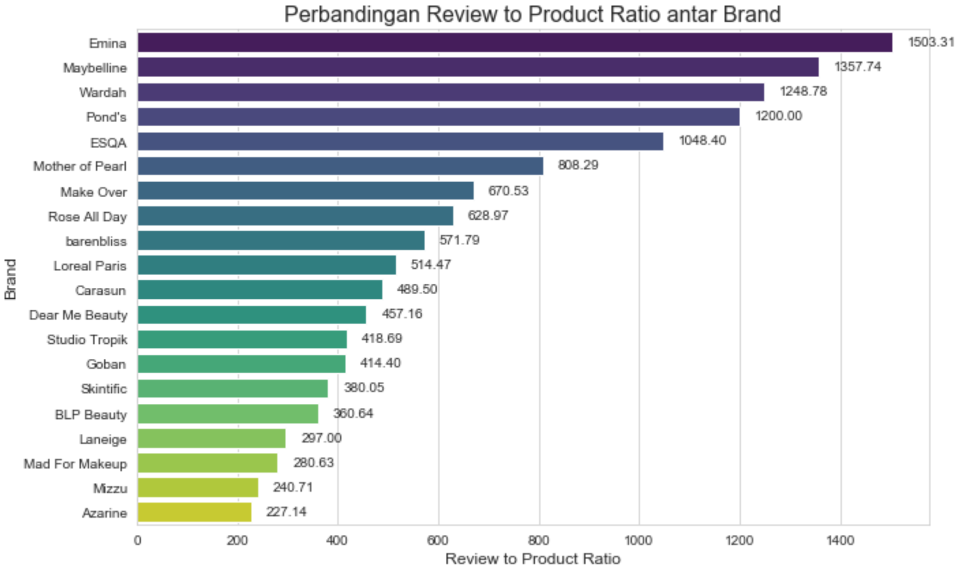
BLP Beauty. These brands offer a wide variety of products, indicating they are highly active in releasing new items. A **large product range** can help a brand target **more customer segments and address diverse needs**. These brands can be considered consistent in product development and potentially have a wider reach.

b. Brands with Highest Total Reviews



By summing the number of reviews for each brand, I identified which brands receive the most customer engagement. From the visualization, **Top Brands by Review Count: Wardah, Emina, and Maybelline**. These brands received the **highest number of reviews**, which reflects a broad customer reach and strong consumer interaction. High review volume can also suggest strong brand awareness and popularity.

c. Review-to-Product Rasio



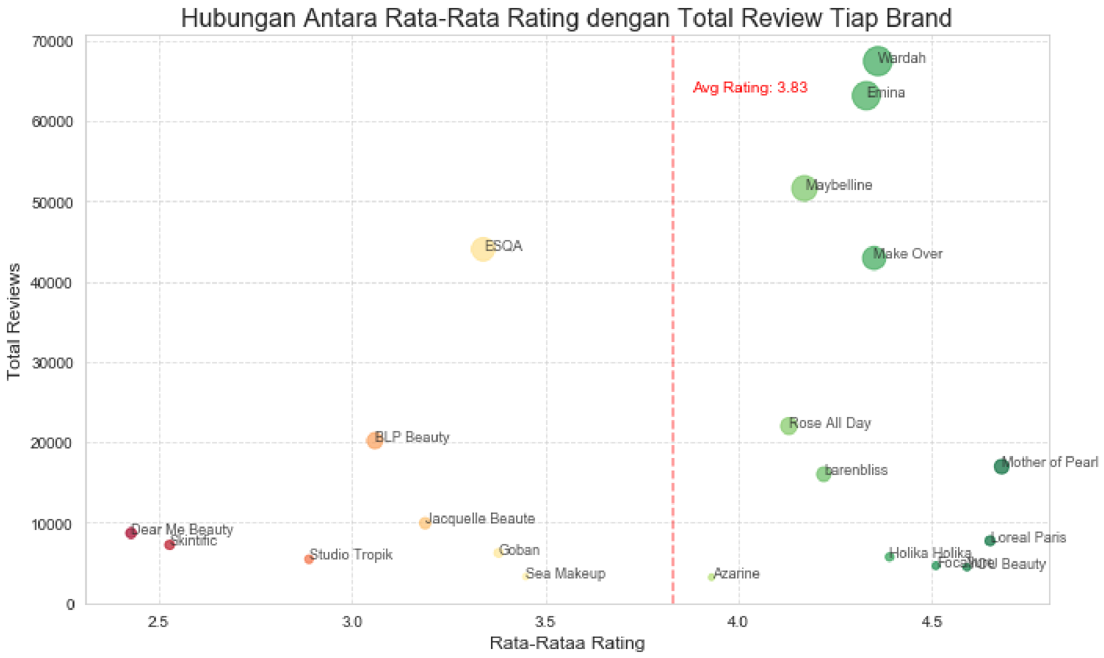
This metric measures how many reviews, on average, each product from a brand receives. It’s calculated by dividing the total number of reviews by the total number of products for each brand. It helps us **understand how well a brand’s products engage consumers**.
For example, **Emina has a ratio of 1,503.31**.
This means:

- On average, **each Emina product receives 1,503 reviews**.
- A total of 63,139 reviews are spread across 42 products.
- Interpretation: Emina's products consistently **capture consumer attention and generate high engagement**.

In comparison, **Maybelline** has a ratio of 1,358, and **Wardah** follows closely. These brands also demonstrate strong consumer engagement across their product lines.

On the other hand, brands like **Azarine and Mizzu** have **much lower ratios (under 250),** suggesting that their products receive fewer reviews per item and may not attract as much consumer interest individually.

d. Brands with Highest Average Rating



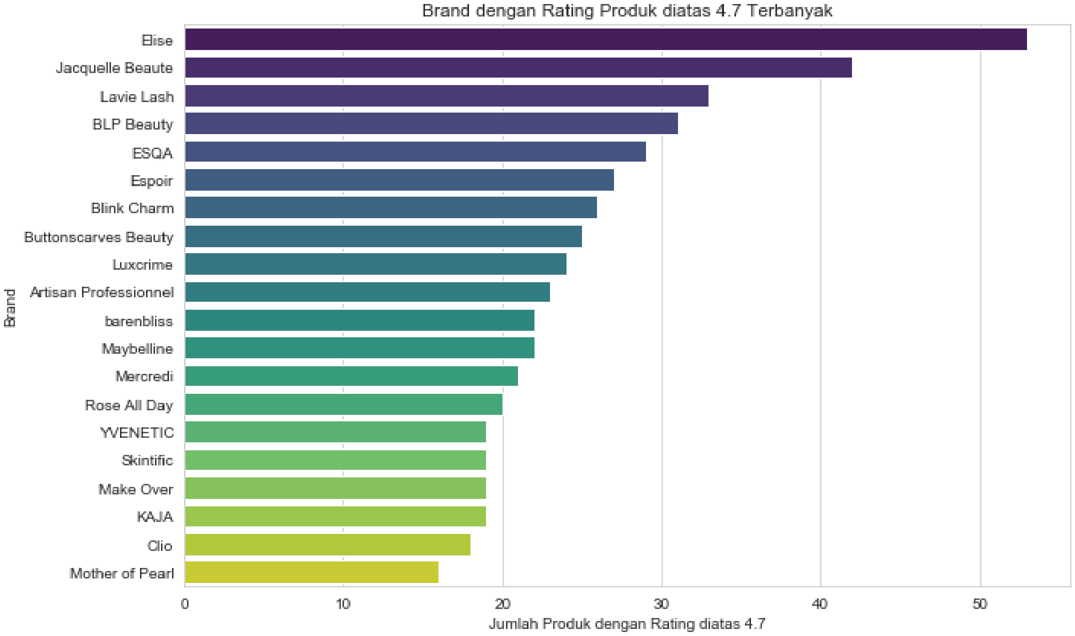
This scatter plot shows the relationship between **average rating** and **total reviews** for each brand. **The vertical red dashed line marks the overall average rating of 3.83**, helping us see which brands perform above or below that benchmark.

From the chart, we can identify several **standout brands** that **have high average ratings** (above 4.0) and also generate **a significant number of reviews**, indicating both **quality and popularity**. These include:

- **Wardah, Emina, and Maybelline**: all positioned in the upper right area, showing that they not only receive high ratings but also large volumes of reviews. This reflects both strong product quality and wide customer reach.
- **Make Over** and **Mother of Pearl** also maintain high ratings and solid engagement, suggesting positive customer reception.

Meanwhile, some brands like **ESQA** have **a very high number of reviews** but **fall slightly below the average rating**. On the other hand, brands like **Skintific** and **Dear Me Beauty** have lower ratings and fewer reviews, which might indicate less favorable feedback or limited visibility.

e. High-Performing Products (Rating > 4.7)



This chart highlights the brands with **the most products rated above 4.7**, indicating consistent product excellence across their lineup.

- **Elise** leads with 53 products scoring above 4.7, followed by **Jacquelle Beaute** with 42, and **Lavie Lash** with 33. These brands stand out for maintaining **exceptional product ratings** across a wide range of items.
- Other notable brands include **BLP Beauty, ESQA, and Espoir**, all with over 25 highly rated products.

These numbers suggest that these brands not only deliver **quality in individual products**, but also maintain **consistent customer satisfaction** across their entire product portfolio.

Popular brands like **Maybelline, barenbliss, and Make Over**, though more mainstream also perform well with over 19 products rated above 4.7, proving their ability to combine **popularity with quality**.

5. Conclusion

Based on the data analysis, we can recommend several brands that are most suitable to prioritize in future promotional or marketing campaigns, based on their performance in product variety, customer engagement, and product quality.

Category	**Top Brands**	**Insight**
Most Product Variety	Make Over, Focallure, BLP Beauty	Active in launching products, wide market coverage
Highest Total Reviews	Wardah, Emina, Maybelline	High brand awareness and customer engagement
Highest Review-to-Product Ratio	Emina (1,503), Maybelline (1,358), Wardah	Each product gets significant customer feedback
High Avg. Rating + High Engagement	Wardah, Emina, Maybelline, Make Over, Mother of Pearl	Quality and popularity combined
Most Products Rated Above 4.7	Elise (53), Jacquelle Beaute (42), Lavie Lash (33)	Consistently excellent product performance