E-commerce Product Pricing Analysis Report

- **Goal**

Collect product pricing data from an Indian fashion e-commerce site, clean and prepare the dataset, analyze pricing patterns, and visualize insights.

- **Data Collection**

The data for this analysis was collected from Amazon.in using web scraping. Web scraping is a technique for extracting data from website by reading their HTML code.I select Kurtas & Kurtis category for analysis. My goal is to gather product informations including Product Name, Price, Brand Name, Category, Maximum Retail Price (MRP), Rating, Number of Reviews, and Product URL. Collect above 400 products from 10 pages.

Python libraries used

Use python library **requests** to fetch HTML content.

To avoid blocking I ensure proper use of headers and user agent.

Use python library **BeautifulSoup** to parse the HTML and extracted important details like price, MRP, rating, number of reviews, brand name and product url.

Use python library **re** to extract specific text pattern. (₹1,999 become 1999, remove extra space in start or end….)

Use python library **time** to add delays between requests to avoid overloading the server.

Standardize brand names (Remove the text “store”,”brand” from the brand name)

Use library **csv** to save data as csv file.

Use **matplotlib**, **seaborn** and **plotly** for visualizations.

Use **pandas** library to import the dataset.

- **Data Cleaning and Transformation**

Find percentage of Discount and add a column to the dataset.

Remove product with percentage of discount less than 0.

Check duplicates, missing values.

Remove missing values.

Save the cleaned dataset in to csv format.

Shape of cleaned dataset (476, 9).

- **Data Analysis**

Descriptive statistics for numerical features.

Price ranges from 199Rs to 509Rs, with an average of 444Rs. Most products (50% percentile) are priced around 474Rs.

MRP ranges from 499Rs to 4999Rs, with an average of 1829Rs, that is products have heavy discount.

Ratings are generally high, ranging from 3.9 to 5, with an average of 4.29.

Number of Reviews varies widely from 3 to 10,120, but the median is only 21, suggesting that a few products are extremely popular while most have moderate attention.

Discount Percentage ranges from 9.1% to 93%, with a median of 75%, indicating that many products are offered at high discounts.

Top 5 brands by number of products

```
Brand Name
Max                  28
Generic              26
JAIPUR HAND BLOCK    12
LIBOZA               12
ANNI DESIGNER        12

MAX is most listed.
```

Top 5 brands that gives the highest average discount

```
Brand Name
FABNEX      89.7100
INDO ERA    89.0825
Kalaanj     89.0200
Manojava    88.1500
Hritika     86.8200
```

FABNEX gives highest discount.

- **Data Visualization**

Plot a **histogram** to analyze the price distribution to understand how products are priced and identify common price ranges. Most products are priced between 450-500.

Plot **bar chart** to find average discount percentage by brand. We can understand many brands offer higher average discounts to attract customers.

Plot a **scatter plot** to visualize the relationship between product ratings and discount percentages. The plot shows that products with higher ratings do not necessarily have higher discounts, indicating that customer satisfaction is not directly linked to discount levels.

Customer satisfaction is generally high but not dependent on discount size. That is price alone does not guarantee customer satisfaction.

- **Challenges Faced & Solutions**

As a beginner in web scraping my only challenge is to study what is web scraping and understand how to extract data from a website and handle inconsistent data. Some product information was written in different ways. Also scraping multiple pages took a lot of time and sometimes caused request errors.

I learned to use BeautifulSoup and regular expressions to get the data correctly. I added delays between requests to avoid errors.