1. Data Collection Process

The dataset was sourced through web scraping from **Myntra**, focusing on women's dresses. Each product entry included attributes such as:

- Product URL
- Brand
- Product Name/Description
- Rating
- Number of Reviews
- Maximum Retail Price (MRP)
- Discounted Price
- Other metadata

Since the raw data contained inconsistencies, the following cleaning steps were performed:

- Removed duplicates using the **Product URL**.
- Converted numeric columns (Rating, Number of Reviews, MRP, Discounted Price) into proper data types.
- Standardized brand names (e.g., trimming spaces, uniform capitalization).
- Created a derived column **"Discount %"** to understand discounting patterns.
- Retained missing values for ratings and reviews (since not all products had feedback).

The cleaned dataset was stored as `myntra_womens_dresses_clean.csv` for analysis.

---

2. Key Findings from the Analysis

a) Brand Analysis

- The dataset showed that **a few brands dominate** the Myntra women's dresses category.

- Popular brands included **Sangria, DressBerry, Anouk, Mast & Harbour**, and **Tokyo Talkies**.

- A small set of brands contributed to the majority of listings.

b) Pricing Trends

- The **average MRP** was found to be **₹1,800 – ₹2,200**, while the **average discounted price** was significantly lower, between **₹900 – ₹1,200**.

- This highlights Myntra's **heavy reliance on discount strategies** to attract customers.

c) Discounts

- The calculated **average discount percentage** was around **40–55%**.

- Some brands consistently offered higher discounts (above 60%), indicating aggressive marketing.

d) Ratings and Reviews

- Not all products had ratings, but where available, the **average rating** was close to **4.0/5**, suggesting generally good customer satisfaction.

- A strong correlation was observed between **higher discounts** and **greater number of reviews**, indicating discounted products attracted more attention.

e) Visualizations

- Bar charts illustrated the **top brands by product count** and the **average discount by brand**.

- Distribution plots showed **MRP vs. Discounted Price patterns**, confirming that most products followed a steep discounting trend.

3. Challenges Faced and Solutions

1. **Data Quality Issues**
   - Problem: Missing values in ratings and reviews.
   - Solution: Retained missing values instead of dropping them, since they represent genuine cases of unrated products.
2. **Inconsistent Brand Names**
   - Problem: Variations in brand names (extra spaces, capitalization differences).
   - Solution: Standardized using `.str.strip().str.title()`.
3. **Outliers in Prices**
   - Problem: Certain products had unusually high MRPs (above ₹10,000).
   - Solution: Verified and kept them, since such premium products exist, but treated them carefully during visualization (log scaling when needed).
4. **Deployment Confusion (Jupyter vs. Streamlit)**
   - Initially attempted to deploy using Streamlit. However, since the focus was purely on **EDA**, Jupyter Notebook was chosen as the final format for clarity and reproducibility.

4. Conclusion

The analysis provided clear insights into the **pricing, discounts, and brand strategies** of Myntra's women's dresses segment. It highlights Myntra's **discount-driven sales model** and identifies the leading brands in the category.

This exercise also improved skills in **data cleaning, exploratory data analysis (EDA), visualization, and reporting**, which are crucial for real-world data science tasks.