Report on Flipkart Women’s Dress Dataset

1. Data Collection Process

The dataset was collected from Flipkart under the **Women’s Dresses** category, focusing on pricing, discounts, ratings, and category-level patterns.

- **Tools Used**:
  - *Python* – data handling
  - *Selenium* – automated navigation & dynamic page loading
  - *BeautifulSoup* – HTML parsing
  - *Pandas* – data cleaning & analysis
- **Methodology**:
  - Search query: *“women dress”*
  - Scraped multiple product pages
  - Extracted: Product Name, Brand, Category, MRP, Discounted Price, Rating, Reviews, and Product URL
  - Data Volume: ~**1000 products** saved as *women_dresses_raw.csv*
- **Data Cleaning**:
  - Removed duplicates and rows with missing price data
  - Missing ratings/reviews replaced with **0**
  - Unknown categories labeled as *“Unknown”*
  - Standardized brand names (e.g., *“Ng Store”* → *Ng*)
  - Final dataset saved as *women_dresses_cleaned.csv*

2. Key Findings from the Analysis

- **Pricing Trends**:
  - Average MRP: **₹2033** (₹299–₹5995)
  - Average discounted price: **₹534**
  - Most dresses priced between **₹300–₹600** (budget-friendly segment)
- **Brand Insights**:
  - **Tokyo Talkies** leads with 107 products
  - **Aayu, Fashion Wear, Sassafras, Selvia** follows as popular brands
  - Lesser-known brands (Tilton, Jkmisti, Kamayra) use **deep discounts (~90%)** to compete
- **Category Insights**:
  - **Maxi & Midi** → premium pricing with luxury outliers
  - **Shift & Kaftan** → consistently affordable

- **Discounts & Ratings**:
  - Most discounts fall in **60–90% range**
  - Ratings cluster around **3.5–4.0**
  - Scatter plot shows **no strong link** between discount percentage and ratings → discounts attract customers, but satisfaction depends more on *quality, style, brand reputation*

📊 Histogram – Distribution of Discounted Prices

Most discounted prices fall in the ₹300–₹600 range. The mean (₹534) being higher than the median (₹467) indicates a right-skewed distribution, driven by a small number of premium items that raise the overall average.

📊 Bar Chart – Top 10 Brands with Highest Average Discount (%)

Tilton leads with the highest average discount (~91%), followed by Jkmisti and Kamayra (both above 88%). Several smaller brands also rely on deep discounts (85–90%) as a strategy to attract customers, highlighting competitive pricing aimed at boosting visibility and sales.

📊 Box Plot – Price Distribution Across Categories

The box plot shows that Maxi and Midi dresses have the widest price ranges, with several high-priced outliers (above ₹1200–₹1500) representing premium or luxury products. In contrast, categories like Shift and Kaftan dresses remain on the lower end with relatively stable discounted prices. The presence of outliers across categories indicates that while most products are budget-friendly, some designer or premium items significantly extend the price range.

📊 Scatter Plot – Ratings vs Discount Percentage

The scatter plot shows that most products fall within the 60–90% discount range, with customer ratings clustering around 3.5–4.0. The almost flat trend line suggests little to no correlation between discount percentage and ratings. This indicates that higher discounts do not necessarily translate into better customer satisfaction—buyers may be valuing factors like quality, design, or brand reputation over just price reductions

3. Challenges & Solutions

- **Dynamic Page Loading**
  - *Problem*: Flipkart loads product data via JavaScript
  - *Solution*: Used Selenium WebDriverWait to handle dynamic elements
- **Inconsistent Brand Names**
  - *Problem*: Same brand appeared in multiple variations

    - *Solution*: Applied regex + manual mapping for standardization
- **Price Outliers**
    - *Problem*: Premium dresses skewed average prices
    - *Solution*: Retained outliers but explained their effect in analysis
- **Scraping Blocks**
    - *Problem*: Pages failed due to timeouts/rate limits
    - *Solution*: Implemented retry logic with sleep intervals

4. Conclusion

This analysis highlights the **affordability of most women's dresses (₹300–₹600 range)**, the **impact of brand reputation**, and **heavy reliance on discounts by smaller brands**. Ratings remain stable (~3.8–4.0), indicating customer satisfaction depends more on quality and style than discount levels.

The cleaned dataset can support **market research, brand pricing strategies, and customer behavior studies.**