IndiaMart Product Data Analysis

Project Overview:

This notebook outlines a data engineering and exploratory data analysis (EDA) workflow designed to evaluate product listings from IndiaMart. The analysis transitions from raw data ingestion to statistical profiling and visual trend identification using Python's data science stack (pandas, matplotlib, seaborn).

1. Data Ingestion & Structure

- **Source:** The analysis is based on an Excel dataset: Indiamart_Final_Submission.xlsx.
- **Dimensions:** The raw dataset contains **107 records** and **6 columns**.
- **Schema:** The data features include:
    - Keyword: Product category (e.g., electronic, processor, mobile).
    - Company: Vendor or supplier name.
    - Name: Specific product description.
    - Price: Listed price in INR.
    - Location: Supplier city/region.
    - Rating: User rating (out of 5.0).

2. Data Profiling & Quality Assessment

The analysis revealed significant gaps in the raw data, necessitating cleaning steps to ensure accuracy:

- **Missing Values:** The dataset contained a high volume of null values, specifically **52 missing Price entries** and **46 missing Rating entries**.
- **Statistical Summary (Raw Data):**
    - **Price:** Demonstrated high variance with a standard deviation of **34,464**, indicating a mix of low-cost components and high-end machinery. Prices ranged from ₹36 to ₹159,900.
    - **Ratings:** Showed a generally positive sentiment with a mean rating of **3.92/5.0**.
- **Data Cleaning:** A data cleaning step (df.dropna()) was executed to remove incomplete records, resulting in a refined dataset of **38 high-quality rows** for precise visualization.

3. Visual Analysis & Insights

The notebook implements several visualization techniques to uncover patterns:

- **Price vs. Rating Analysis:** A scatter plot titled "Comparision between Price and Rating" was generated to investigate if higher-priced items correlate with better user ratings.
- **Multivariate Relationships:** A Seaborn pairplot was utilized to visualize pairwise relationships and distributions between all numeric variables simultaneously.
- **Product Performance:** A bar chart visualizing Name vs. Rating was created to identify specific products that achieve the highest customer satisfaction.

4. Technical Implementation

- **Libraries Used:** The code utilizes standard industry libraries including Pandas for manipulation, NumPy for numerical operations, and Matplotlib/Seaborn for high-quality plotting.
- **Methodology:** The notebook follows a standard EDA pipeline: **Load $\rightarrow$ Inspect $\rightarrow$ Clean $\rightarrow$ Visualize**, ensuring the final insights are derived from verified data points.