Amazon Product Analysis

Problems and Background

This project aims to analyse product data and customer reviews from Amazon to understand how different factors like price, discounts, and ratings affect customer behaviour. The dataset was sourced from Kaggle and includes product information, ratings, review texts, and pricing details.

The main goal is to draw useful business insights that can help sellers or marketing teams improve product performance and customer satisfaction. By using Python for data exploration and Power BI for dashboard visualisations, the project focuses on making the data both clear and actionable.

Solution

🔍 Data Analysis

The dataset was first explored using **Python (pandas, matplotlib, seaborn)** to clean and analyse key features such as ratings, discounts, and prices. Several insights were visualized using scatter plots, bar charts, and histograms. A cleaned version of the dataset was also used in **Power BI** to create an interactive dashboard.

📊 Data-Driven Insights

Key findings from the analysis include :

- Most products are highly rated (4+ stars)
- Higher discounts don't always lead to better ratings
- The ₹1000–₹1999 price range has the highest product count

These findings can help sellers and decision-makers better position their products and optimise discount strategies.

🤝 Stakeholder Engagement

The final outputs — Python notebooks and Power BI dashboards — were designed to be easily understood by non-technical stakeholders. Visualizations and summaries make it simple for business teams to extract value without needing to go deep into the raw data.

📌 Project Scope

🧭 Project Goal

The goal of this project is to uncover meaningful insights from Amazon product data using Python and Power BI. The analysis focuses on key business aspects like customer satisfaction, pricing strategy, product popularity, and category performance.

🗂️ Step-by-Step Approach

1. **Data Collection**
   Dataset sourced from **Kaggle**, containing product-level information including ratings, prices, discounts, and reviews.

2. **Data Cleaning & Preparation (Day 1–2)**
   - Handled missing values and inconsistent formats
   - Converted data types for analysis (e.g., discount %, rating)

3. **Exploratory Data Analysis in Python (Day 3–5)**
   - Performed 5+ focused analyses
   - Used visualizations to explore patterns and correlations

4. **Power BI Dashboard Creation (Day 6–7)**
   - Built a clean, professional dashboard
   - Visualized metrics like average rating, product distribution, discount ranges, etc.

5. **Project Wrap-up & Documentation (Day 8)**
   - Markdown summaries in Jupyter Notebook

  - ○ Finalized README and uploaded all materials to GitHub

🔄 Business Lens: The 5 Ps

Although this is not a direct marketing campaign, the project indirectly supports:

- **Product**: Identifies high-performing and low-performing product types
- **Price**: Shows how pricing affects ratings and review count
- **Promotion**: Evaluates if discount levels influence satisfaction
- **Place**: (Not included — no location data in dataset)
- **People**: Understands customer sentiment through reviews

👥 Stakeholders

- **Product Managers**: Can use insights to improve product offerings
- **Marketing Teams**: Can align discount strategies with customer response
- **Business Decision-Makers**: Get a clear view of what's working and what's not through dashboard summaries

⚙️ Methodology

This project follows a structured data analysis workflow to transform raw data into actionable business insights. Below is an outline of each step used in the project:

🗂️ Data Source

- The dataset was downloaded from **Kaggle** and contains Amazon product listings, prices, reviews, discounts, and ratings.
- The data was handled **locally using Python** in Jupyter Notebook.
- A **cleaned version of the dataset**, processed in Python, was later exported and used in **Power BI** for dashboard creation.

🧹 Data Wrangling & Cleaning

- Removed unwanted characters (e.g., `%` symbols in discounts), fixed data types, and handled missing values.
- Ensured consistency in fields such as rating, discount, and review count.
- Created new columns such as review length, price range, and image availability.

🧠 Data Understanding

- Explored individual columns and relationships between features.
- Focused on key attributes like rating count, product categories, and actual vs discounted price.

🔧 Data Manipulation

- Used **pandas** for filtering, grouping, and creating derived fields:
  - `review_length`
  - `has_image` (True/False)
  - `price_range` (categorizing products into brackets)

📊 Data Analysis

- Conducted multiple analyses to explore:
  - Discount vs. Rating correlation
  - Top-rated products
  - Review length vs. Rating
  - Rating distribution
  - Category-wise performance

📈 Data Visualization

- **Python (matplotlib/seaborn)**: Used for initial explorations and detailed visual interpretations.
- **Power BI**: An interactive dashboard was created using the **cleaned dataset from Python**.
  Visuals include:
  - Clustered bar and column charts
  - Tree maps and pie charts
  - Scatter plots (where relevant)
  - Clean layouts for clear interpretation by business users

🎯 Goals and KPIs

To evaluate the success and impact of this project, the following goals and corresponding KPIs were identified:

✅ Goal 1: Optimize Discount Strategy

KPI → Discount ranges where average product rating remains above **4.0**, indicating customer satisfaction isn't negatively impacted

✅ Goal 2: Understand Price Sensitivity

KPI → Price brackets with the **highest number of well-rated products** (rating ≥ 4), helping target ideal pricing zones

✅ Goal 3: Enhance Dashboard Clarity for Stakeholders

KPI → A Power BI dashboard with **5+ clear, actionable visuals**, enabling non-technical users to interpret key insights at a glance

🛠️ Technical Processes

- `Python: read_csv()`, `dropna()`, `astype()`, `str.replace()`, `apply()`, `value_counts()`, `groupby()`, `corr()`, `nlargest()`, `isnull()`, `len()`
- Power BI: Relationships, DAX measures, calculated columns

💼 Business Concepts Used

📌 Concept 1: Customer Behavior Analysis

Used product ratings, review counts, and discount responsiveness to understand how customers react to pricing and product quality.

📌 Concept 2: Market Segmentation

Categorized products based on rating, price range, and category to identify high-performing segments for better targeting.

📌 Concept 3: Customer Retention Insight

Analyzed review lengths and satisfaction (ratings) to understand the likelihood of customers returning or recommending products.

📊 Recommended Analysis

1. Discount Percentage vs Product Rating

- **Goal:** To find out if discounts influence product ratings.
- **Process:** Cleaned `discount_percentage` column, converted to numeric, and used `corr()` to find correlation with ratings. Visualized with scatter plot.
- **Impact:** Found weak negative correlation (-0.16), suggesting higher discounts don't necessarily mean better ratings.

2. Average Rating by Category

- **Goal:** Identify top-performing product categories.
- **Process:** Extracted main category from split values, used `groupby()` to calculate average rating.
- **Impact:** Helped highlight which categories consistently perform well, useful for category targeting.

3. Product with Highest Rating Count

- **Goal:** Identify the most engaged products.
- **Process:** Cleaned `rating_count`, sorted to find top product by number of ratings.
- **Impact:** Highlights trusted/popular products, guiding promotional focus.

4. Rating Distribution

- **Goal:** Understand how ratings are spread across all products.
- **Process:** Created histogram of the `rating` column using seaborn.
- **Impact:** Revealed that most products are rated 4 or above, indicating general customer satisfaction.

5. Review Length vs Rating

- **Goal:** Check if longer reviews indicate stronger customer opinions.
- **Process:** Measured review text length and correlated with ratings.
- **Impact:** Found a very weak correlation (~0.07), suggesting review length isn’t a strong indicator of sentiment.

👤 Project Owner

- **Name:** Saransh Umrao
- **Date:** June 30th, 2025