Amazon Product Analysis Capstone Project

Student: David Enyinnaya
Tool Used: Microsoft Excel
Dataset: Amazon product listing data (including product name, price, rating, reviews, etc.)

Objective

To perform end-to-end data cleaning and analysis on Amazon product data and answer 14 business-relevant questions using Excel techniques such as formulas, PivotTables, filters, and charts.

Data Cleaning Summary

- Removed unnecessary columns (e.g., username, review ID, product ID)
- Cleaned key columns:
  - `category` ✓ fixed inconsistent category naming
  - `product_name` ✓ removed trailing symbols and trimmed whitespace
  - `actual_price` & `discounted_price` ✓ formatted as numbers (2 decimal places)
  - `discount_percentage_clean` ✓ calculated using: `=IF(actual_price=0, "", (actual_price - discounted_price)/actual_price)`
  - `rating_count` & `average_rating` ✓ verified numeric and cleaned using ISNUMBER logic
- Added new calculated fields: `discount_percentage_clean`, `potential_revenue`, `price_bucket`, `rating_plus_reviews_score`
- Used number formatting for currency, percentage, and counts

---

Analysis Summary

Q1: Average Discount % by Category
Used Pivot Table:

- Rows: Category
- Values: Discount % (Average)

Q2: Number of Products per Category
Pivot Table:

- Rows: Category

- Values: Product Name (Count)

Q3: Total Reviews per Category
Pivot Table:

- Rows: Category
- Values: Rating Count (Sum)

Q4: Products with Highest Ratings
Sorted dataset by `average_rating` descending.
Top 5 identified manually.

Q5: Average Actual vs Discounted Price per Category
Pivot Table:

- Rows: Category
- Values: Actual Price (Average), Discounted Price (Average)

Q6: Products with Most Reviews
Sorted dataset by `rating_count` descending. Top products identified.

Q7: Products with 50%+ Discount
Added logic column using:
`=IF(discount_percentage_clean>=0.5,"Yes","No")`
Used COUNTIF to count "Yes"

Q8: Distribution of Product Ratings
Pivot Table:

- Rows: Rounded Rating
- Values: Product Count
- Chart: Column chart for distribution

Q9: Potential Revenue by Category
Calculated: `actual_price * rating_count`
Pivot Table:

- Rows: Category
- Values: Sum of Potential Revenue

Q10: Product Count by Price Bucket
Bucket formula:
`=IF(discounted_price<200,"<₹200",IF(discounted_price<=500,"₹200-₹500",">₹500"))`
Pivot Table:

- Rows: Price Bucket

- Values: Product Count (Distinct)

Q11: Relationship Between Rating & Discount
Created `discount_bucket` (0-10%, 11-20%, ...)
Pivot Table:

- Rows: Discount Bucket
- Values: Average Rating
- Chart: Line Chart showing rating trend by discount level

Q12: Products with < 1,000 Reviews
Used formula: `=COUNTIF(rating_count,"<1000")`

Q13: Categories with Highest Max Discount
Pivot Table:

- Rows: Category
- Values: Discount % (Max)

Q14: Top 5 Products by Combined Rating & Reviews
New column:
```
=average_rating + (rating_count / 1000)
```
Sorted by this score to get top 5 products.

Conclusion

This project involved cleaning and analyzing a large dataset using only Excel. It demonstrated practical skills in:

- Data cleaning and preparation
- Formulas and conditional logic
- Pivot Table construction
- Visualization and trend analysis
- Drawing business insights from raw data

Prepared by: David Enyinnaya
Date: July 2025