Amazon Product Reviews Analysis - Project Report

# Project Overview

In this project, I performed data cleaning, exploratory analysis, and visualization on an Amazon Product Reviews dataset. The primary objective was to validate the dataset's integrity, ensure it was ready for analysis, and extract meaningful insights regarding product popularity and customer ratings behavior.

# Dataset Description

The dataset was sourced from Kaggle and contains the following columns:

• user\_id: Unique identifier for each customer

• product\_id: Unique identifier for each product

• rating: Rating given by the customer (1 to 5 stars)

• timestamp: UNIX timestamp representing when the review was submitted

The dataset consists of millions of product reviews across various products.

# Data Cleaning Process

Although the dataset appeared clean, I took the following steps to verify and improve data quality:

1. Checked for Missing Values and Duplicates: There were no missing values or duplicate rows found in the dataset.

2. Converted Data Types: The rating column was originally in float format. I converted it to integer format for clarity.

3. Formatted Timestamps: The UNIX timestamps were converted into human-readable date format to make time-based analysis easier.

# Exploratory Data Analysis (EDA) & Visualizations

1. Rating Distribution: I analyzed how ratings were distributed across the dataset. The visualization showed that the majority of ratings were 5 stars, indicating a generally positive customer experience.

2. Top 10 Most Rated Products: I identified the top 10 products with the highest number of ratings, revealing which products were the most popular and frequently rated by customers.

3. Top 5 Most Reviewed Products: I zoomed in on the top 5 most reviewed products to highlight those with the highest engagement levels.

4. Average Ratings Over Time: I visualized how average ratings changed over the years. Despite some fluctuations, the ratings stabilized around 4 stars in recent years.

# Key Findings

• The dataset was clean and reliable with no missing values or duplicates.

• 5-star ratings dominated, showing high customer satisfaction.

• Certain products consistently appeared among the most rated and reviewed, highlighting their popularity.

• Average ratings remained stable and positive over time.

# Future Work & Recommendations

• Develop a recommendation system based on product ratings and user behavior.

• Perform sentiment analysis on customer review texts (if available) for deeper insights.

• Explore seasonal trends or time-based patterns in product popularity.

# Tools & Technologies Used

• Python

• Pandas & NumPy for data manipulation

• Matplotlib & Seaborn for visualization

• Jupyter Notebook for coding environment

• Git & GitHub for version control and sharing

# Conclusion

This project demonstrated the importance of validating and cleaning data even when it appears well-structured. Through analysis and visualization, I uncovered patterns in customer ratings, highlighted popular products, and ensured the dataset was ready for further exploration or advanced modeling.