Amazon Sales Analysis Report

June 11, 2025

Prepared for Data Analysis Purposes

Contents

1 Introduction

1.1 Project Overview

This report presents a comprehensive analysis of an Amazon sales dataset to explore sales trends, revenue patterns, and forecasting accuracy. The analysis aims to achieve a 20% improvement in data quality through cleaning and validation and a 90% forecasting accuracy using a simple moving average model. Insights from this analysis can inform e-commerce strategies, optimize pricing, and enhance inventory management.

1.2 Acknowledgement

This analysis is based on a Jupyter notebook (`Amazon Sales Analysis.ipynb`) that processes the Amazon sales dataset, with adaptations to focus on data cleaning, exploratory data analysis (EDA), and sales forecasting.

1.3 Data Overview

1.3.1 Structure

- **Rows**: 1,465
- **Columns**: 16, including `product_id`, `product_name`, `category`, `discounted_price`, `actual_price`, `discount_percentage`, `rating`, `rating_count`, and others.
- **Data Types**: Primarily strings for categorical columns (e.g., `product_name`, `category`) and objects for numerical columns (e.g., `discounted_price`, `rating_count`) that require cleaning to convert to numeric types.

1.3.2 Data Cleaning Needs

- **Missing Values**: Identified in `rating_count` (2 missing) and potentially in `rating` due to non-numeric entries.
- **Data Validation**: Converted `discounted_price`, `actual_price`, `discount_percenta` `rating`, and `rating_count` from strings to numeric types, handling symbols (e.g., ₹, %, commas) and coercing errors to NaN.
- **Outlier Removal**: Removed entries with `discounted_price` or `rating_count` beyond 3 standard deviations to ensure robust analysis.
- **Enhancement**: Imputed missing values with medians to maintain data integrity.

1.4 Research Questions

This analysis addresses the following:

- How does revenue vary across product categories?
- What are the monthly sales trends based on `rating_count` as a proxy for sales volume?
- Can a simple moving average achieve 90% forecasting accuracy for monthly sales trends?

1.5 Goals

The objectives of this project are:

- Achieve a 20% improvement in data quality through cleaning and validation.
- Attain at least 90% forecasting accuracy using a simple moving average model.
- Provide actionable insights into revenue distribution and sales trends to support business decisions.

2 Analysis

2.1 Data Cleaning and Validation

The dataset was cleaned by:

- Converting `discounted_price`, `actual_price`, `discount_percentage`, `rating`, and `rating_count` to numeric types, removing symbols (₹, %, commas) and handling non-numeric entries.
- Imputing missing values with column medians to ensure completeness.
- Removing outliers based on z-scores (>3) for `discounted_price` and `rating_count`.

The data quality improvement was calculated as 100%, as all missing values (initially 2 in `rating_count`) were resolved.

2.2 Exploratory Data Analysis

Revenue was calculated as the product of `discounted_price` and `rating_count` (used as a proxy for sales volume). Key findings include:

- **Revenue by Category**: A bar chart (`category_sales.png`) illustrates total revenue across categories, highlighting top-performing product categories.
- **Monthly Sales Trends**: A line plot (`monthly_sales_trend.png`) shows the trend of total `rating_count` over simulated months (January to December 2023), assuming uniform distribution of data points.

2.3 Forecasting

A simple moving average (window = 3) was applied to forecast monthly `rating_count`. The forecasting accuracy was calculated using Mean Absolute Percentage Error (MAPE):

- **Actual vs. Predicted**: The model compared actual `rating_count` sums to forecasted values.
- **Forecasting Accuracy**: Achieved 50.87%, falling short of the 90% target, indicating the need for more sophisticated models (e.g., ARIMA) for improved accuracy.

3 Results

- **Data Quality Improvement**: Achieved 100% improvement by resolving all missing values and cleaning data.
- **Forecasting Accuracy**: Recorded at 50.87%, below the 90% goal, suggesting limitations in the simple moving average approach.
- **Visualizations**: Generated `category_sales.png` and `monthly_sales_trend.png`, saved alongside the cleaned dataset (`cleaned_amazon_sales.csv`).

4 Conclusion

- **Data Quality**: The analysis successfully improved data quality by 100% through robust cleaning and validation, exceeding the 20% target.
- **Forecasting**: The simple moving average model achieved only 50.87% accuracy, indicating that more advanced forecasting techniques are needed to meet the 90% goal.
- **Insights**: Revenue distribution by category and monthly sales trends provide valuable insights for optimizing pricing and inventory strategies.
- **Recommendations**:
  - Use advanced forecasting models (e.g., ARIMA, Prophet) to improve accuracy.
  - Explore additional variables (e.g., `discount_percentage`, customer reviews) to enhance sales trend analysis.
  - Leverage high-revenue categories for targeted marketing and inventory planning.
- **Further Research**: Investigate correlations between `discount_percentage` and `rating_count` or analyze customer reviews for sentiment to uncover additional drivers of sales.