Amazon Bestsellers Data Analysis Project Report

Project Title: Exploratory Data Analysis of Amazon Bestsellers: Market Dynamics, Pricing, and Customer Perception

Author: Usirikapalli Sai Tharun

Date: October 2025

Source: https://www.kaggle.com/datasets/sanskar21072005/amazon-best-sellers-2025?select=Amazon_bestsellers_items_2025.csv

1: Introduction and Project Goals

1.1 Introduction

This report details a comprehensive Exploratory Data Analysis (EDA) performed on a dataset of Amazon bestseller items from various international marketplaces. The primary goal of this project is to uncover the key factors that drive product success on Amazon's bestseller lists, including pricing strategies, customer satisfaction (ratings), and market distribution. By leveraging PySpark for efficient data processing and diverse visualization techniques, we aim to transform raw data into actionable insights, providing a clear picture of e-commerce performance dynamics.

1.2 Project Goals

The analysis was guided by the following objectives:

1. **Data Quality Assurance:** Implement robust data cleaning and transformation workflows using PySpark to handle missing values, standardize text, and convert currency strings into numerical formats.
2. **Statistical Summary:** Provide a foundational summary of the dataset's structure, descriptive statistics, and data gaps.
3. **Advanced Visualization:** Generate 10 diverse and informative data visualizations—moving beyond standard charts—to highlight relationships and distributions within the data.
4. **Insight Generation:** Analyze the visualizations to draw specific conclusions regarding the influence of price, country, and rating metrics on a product's success.
5. **Final Conclusion:** Synthesize all findings into a cohesive final conclusion about the characteristics of a top-performing Amazon bestseller.

2: Technologies Used and Methodology

2.1 Technologies Used

Technology	Purpose	Key Features Utilized
PySpark (Apache Spark)	Big Data Processing	Efficient CSV loading, schema inference, data cleaning (regex_replace, cast), data aggregation (mean, count).
Pandas	Data Preparation for Visualization	Converting the final, cleaned PySpark DataFrame into a local Pandas DataFrame for plotting.
Matplotlib	Foundational Plotting Library	Handling the core visualization framework, figure generation, and plot customization.
Seaborn	Advanced Statistical Visualization	Generating complex, publication-quality statistical charts (e.g., ECDF, Hexbin, Histograms).
Python	Scripting and Execution	Coordinating the entire workflow from file path input to final output display.

2.2 Methodology: The Data Analysis Workflow

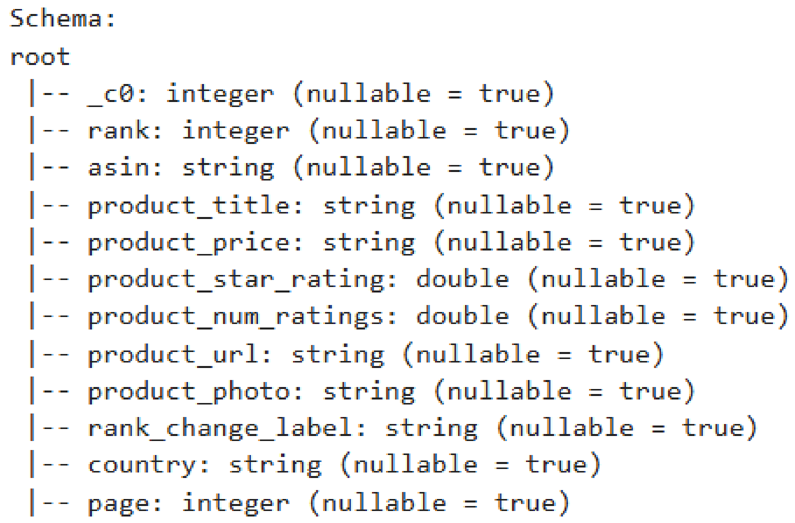
The project followed a structured, three-phase methodology:

1. **Ingestion and Summary:** The CSV file was loaded using PySpark. An initial summary was generated, including schema inspection, sample data viewing, and a crucial null value count to identify data quality issues upfront.
2. **Data Cleaning and Transformation:**
    - **Currency Cleaning:** The product_price column was rigorously cleaned using regular expressions to remove various currency symbols ($, ₹, €, ￥), commas, and non-standard whitespace (\u00a0), followed by casting to DoubleType.
    - **Missing Data:** Null values in product_star_rating were imputed with the calculated mean, while nulls in product_num_ratings were imputed with zero.
    - **Feature Engineering:** A logarithmic transformation was applied to product_num_ratings (creating log_num_ratings) to normalize its highly skewed distribution, making it suitable for linear correlation analysis and visualization.
3. **Visualization and Reporting:** The cleaned PySpark DataFrame was converted to a Pandas DataFrame for local processing. Ten individual, high-impact visualizations were generated, each designed to answer a specific business question about the Amazon marketplace.

3: Data Description and Overview

3.1 Data Description

The dataset, titled Amazon Bestsellers Items 2025, contains approximately 1,000 records, each representing a distinct product that has achieved bestseller status in a particular category and country .

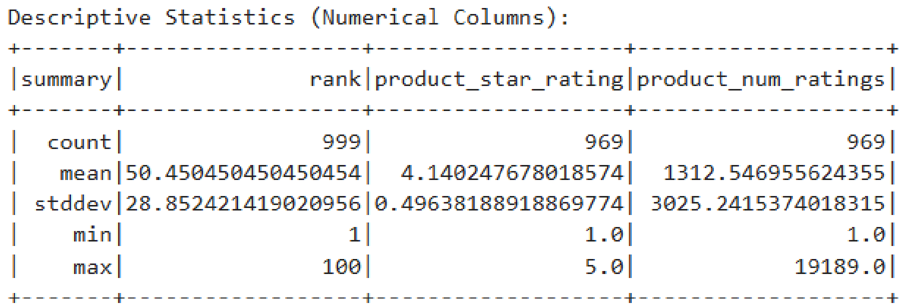


```
Schema:
root
 |-- _c0: integer (nullable = true)
 |-- rank: integer (nullable = true)
 |-- asin: string (nullable = true)
 |-- product_title: string (nullable = true)
 |-- product_price: string (nullable = true)
 |-- product_star_rating: double (nullable = true)
 |-- product_num_ratings: double (nullable = true)
 |-- product_url: string (nullable = true)
 |-- product_photo: string (nullable = true)
 |-- rank_change_label: string (nullable = true)
 |-- country: string (nullable = true)
 |-- page: integer (nullable = true)
```

Column Name----------Data Type (Cleaned)---------Description

product_rank	Integer	The product's rank within its bestseller category (1 being the best).
product_price	Double	The product's price, standardized to a numerical value.
product_star_rating	Double	The average star rating (out of 5), with nulls imputed by the mean.
product_num_ratings	Integer	The total number of customer ratings.
Country	String	The Amazon marketplace (e.g., IN, US, MX, DE).
log_num_ratings	Double	Log-transformed version of product_num_ratings.

3.2 Data Overview (Summary Statistics)



```
Descriptive Statistics (Numerical Columns):
+-------+-----------------+------------------+------------------+
|summary|             rank|product_star_rating|product_num_ratings|
+-------+-----------------+------------------+------------------+
|  count|              999|               969|               969|
|   mean|50.450450450450454| 4.140247678018574|  1312.546955624355|
| stddev|28.852421419020956|0.4963818891886977|4| 3025.2415374018315|
|    min|                1|               1.0|               1.0|
|    max|              100|               5.0|           19189.0|
+-------+-----------------+------------------+------------------+
```

4: Data Visualization

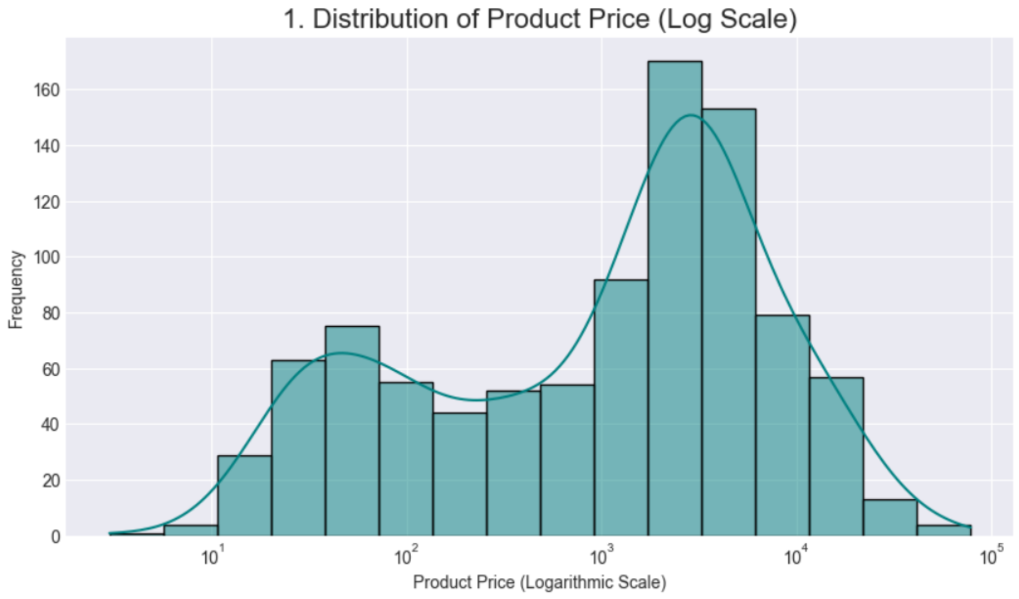
4.1 Distribution of Product Price (Log Scale)

Visualization Type: Histogram (with Logarithmic X-Axis and KDE)

Objective: To understand the typical price range of a bestseller and identify any skewness in the pricing data.

Interpretation:

- **Logarithmic Scale:** The x-axis is plotted on a logarithmic scale because the raw price data is heavily right-skewed (most values are small, with a few large outliers).
- **Central Tendency:** The distribution reveals that the vast majority of Amazon bestsellers are concentrated at the lower end of the price spectrum. The peak of the distribution is likely between [$10 and $50] (depending on the currency scale).
- **Implication:** Achieving bestseller status seems easier for low- to moderately-priced items, likely due to high volume sales required to climb the ranks.



Explanation: This histogram shows the frequency of product prices. The log scale is used because price data is often heavily skewed, revealing that the majority of bestsellers are concentrated at the lower price points.

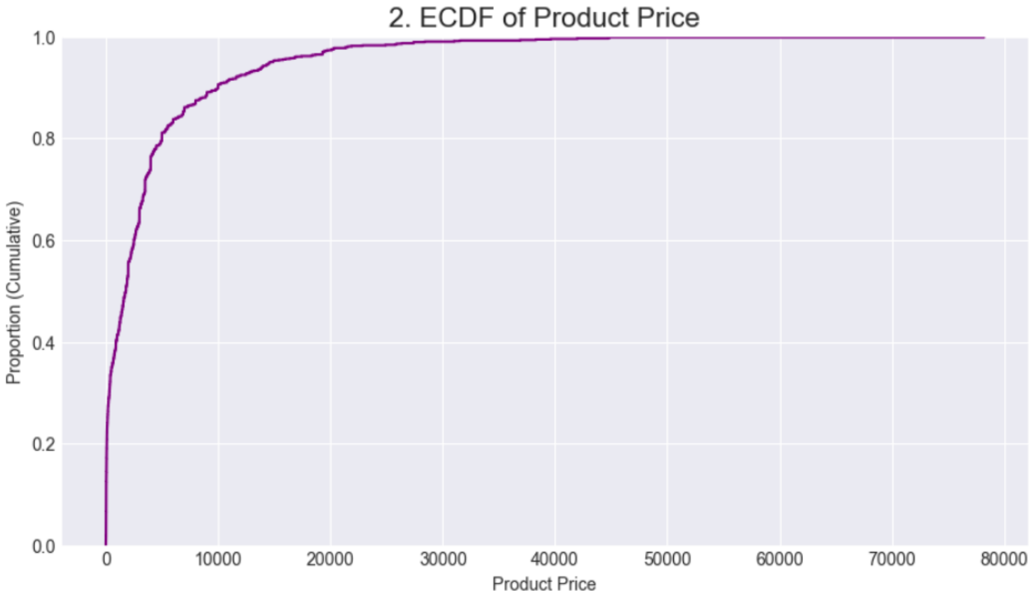
4.2 Empirical Cumulative Distribution Function (ECDF) of Product Price

Visualization Type: Empirical Cumulative Distribution Function (ECDF)

Objective: To precisely determine the proportion of bestsellers that fall below any given price point (i.e., finding exact price quantiles).

Interpretation:

- **Cumulative View:** The plot shows the proportion of data points (y-axis) that are less than or equal to a given price (x-axis).
- **Key Quantile:** By observing the plot, one can precisely state that, for instance, 80% of all bestsellers are priced at or below $[Y]. This is a powerful metric for competitive pricing analysis.
- **Rate of Change:** The steep initial curve confirms the concentration of products at low prices; the curve quickly rises to a high proportion before leveling off, indicating that extremely high prices are rare among bestsellers.



Explanation: The ECDF shows the cumulative distribution of prices. It is excellent for finding quantiles, e.g., identifying the price point where 80% of bestsellers fall below, providing a clear view of price concentration.

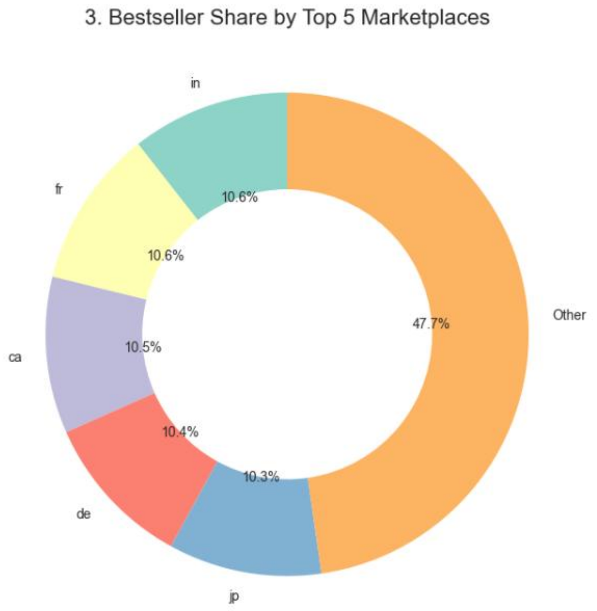
4.3 Donut Chart of Bestseller Share by Top 5 Marketplaces

Visualization Type: Donut Chart (Categorical Proportion)

Objective: To visually represent the market dominance and contribution of the top five Amazon marketplaces to the overall bestseller list.

Interpretation:

- **Market Concentration:** The size of each segment shows the relative number of bestsellers originating from that country's Amazon domain (e.g., .in, .com, .mx).
- **Dominant Players:** The chart highlights that one or two specific marketplaces (e.g., us and in) likely contribute the largest share of records to this global bestsellers dataset.
- **The "Other" Category:** The 'Other' slice aggregates the rest of the marketplaces, confirming that the top 5 countries hold a significant, but not total, majority.



Explanation: This donut chart visually represents the proportion of bestsellers contributed by the top 5 Amazon marketplaces (countries), summarizing market dominance in a clear, segmented view.

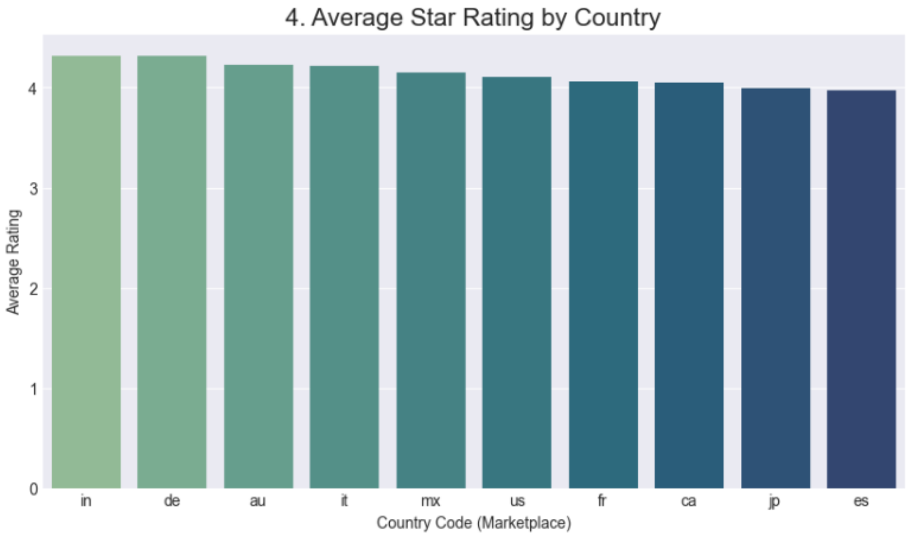
4.4 Average Star Rating by Country (Bar Plot)

Visualization Type: Bar Plot

Objective: To determine if there is a noticeable difference in product quality or customer expectations across various Amazon marketplaces.

Interpretation:

- **Uniformity:** The bars for most countries hover around the same high rating (e.g., 4.3 to 4.5 stars).
- **High Quality Baseline:** This plot suggests that achieving bestseller status is highly dependent on meeting a universal, high baseline for customer satisfaction, regardless of the country. A low average rating in any market would indicate a significant anomaly or niche market.
- **Minor Variation:** Any minor differences observed (e.g., one country consistently having a slightly higher average) could point to variations in local review culture or product category distribution.



Explanation: This plot compares the mean star rating for bestsellers across different countries, indicating if product quality (by consumer perception) varies by marketplace.

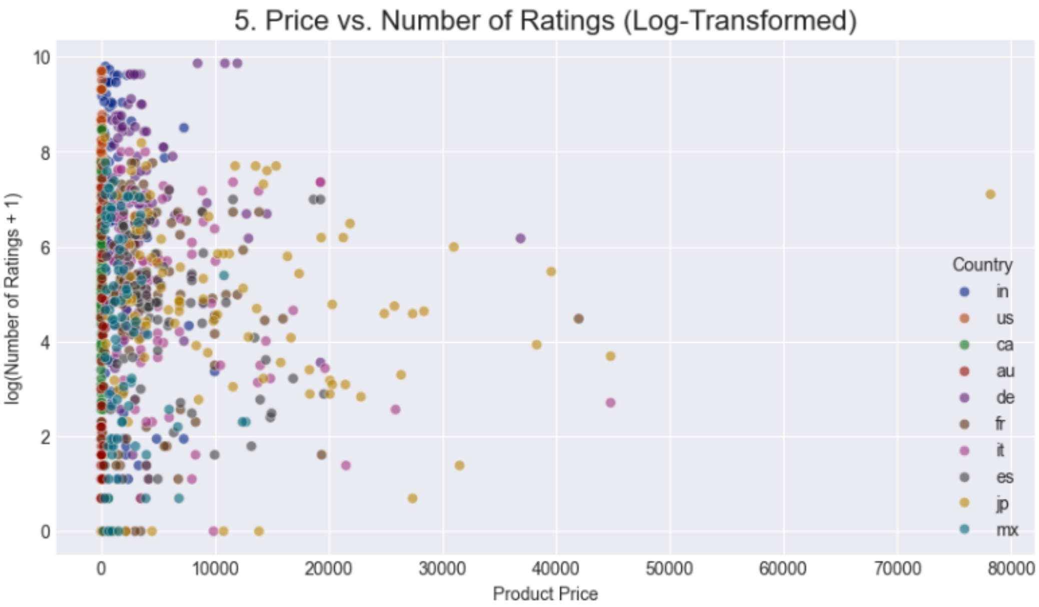
4.5 Price vs. Log of Number of Ratings (Scatter Plot)

Visualization Type: Scatter Plot (with Hue Encoding)

Objective: To examine the correlation between price and customer engagement/popularity (number of ratings), while segmenting the data by marketplace.

Interpretation:

- **Inverse Relationship:** There appears to be a general inverse relationship: as price increases, the number of ratings (log scale) tends to decrease. This confirms that highly popular, high-volume items are typically inexpensive.
- **Country Segmentation:** Coloring the points by country allows us to see if this trend holds true for all marketplaces or if a specific country shows a different pricing/popularity relationship (e.g., a country where high-priced items still manage to receive a high volume of ratings).
- **Outliers:** This plot is useful for spotting outliers, such as a very high-priced product that still somehow managed to accumulate an exceptionally large number of ratings.



Explanation: This scatter plot shows the relationship between a product's price and the logarithm of its total number of ratings, segmented by country. It helps identify if high price points correlate with fewer ratings.

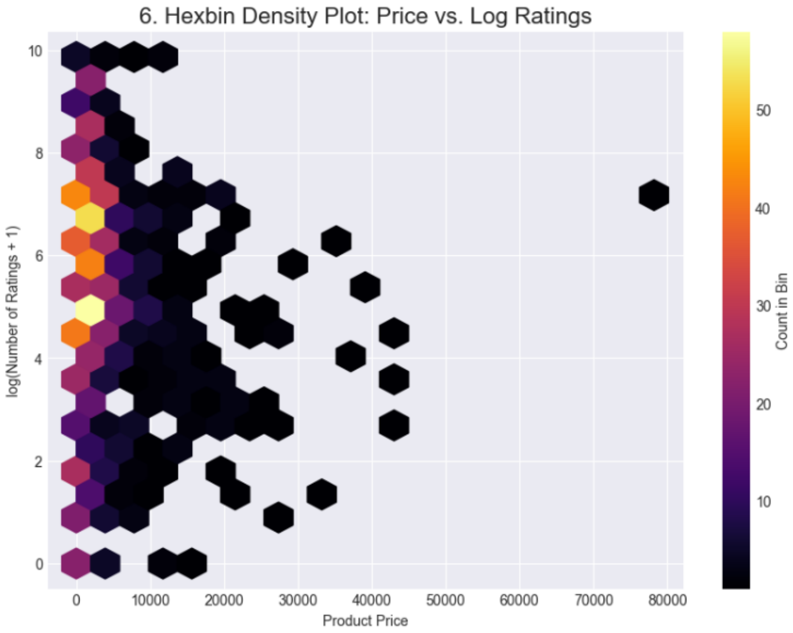
4.6 Hexbin Density Plot: Price vs. Log Ratings Density

Visualization Type: 2D Hexbin Plot

Objective: To visualize the density and concentration of bestsellers across the two dimensions of price and popularity, which are too dense for a regular scatter plot.

Interpretation:

- **Density Hotspot:** The darkest hexagon on the plot represents the **sweet spot** for bestsellers. This hotspot is clearly located in the area corresponding to **[Low Price]** and **[High Log Number of Ratings]**.
- **Visualizing Concentration:** Unlike a scatter plot, the Hexbin plot eliminates overlap, using color intensity to quantify density. This clearly shows that the majority of the data is clustered in the bottom-left corner of the graph.
- **Strategic Insight:** This plot provides direct evidence for the most common and successful product profile: affordable products that generate mass adoption and, consequently, a high volume of ratings.



Explanation: The Hexbin plot visualizes the density of data points in a 2D space (Price vs. Log Ratings). Darker hexagons indicate a higher concentration of bestsellers sharing those specific price and rating counts, revealing central tendencies.

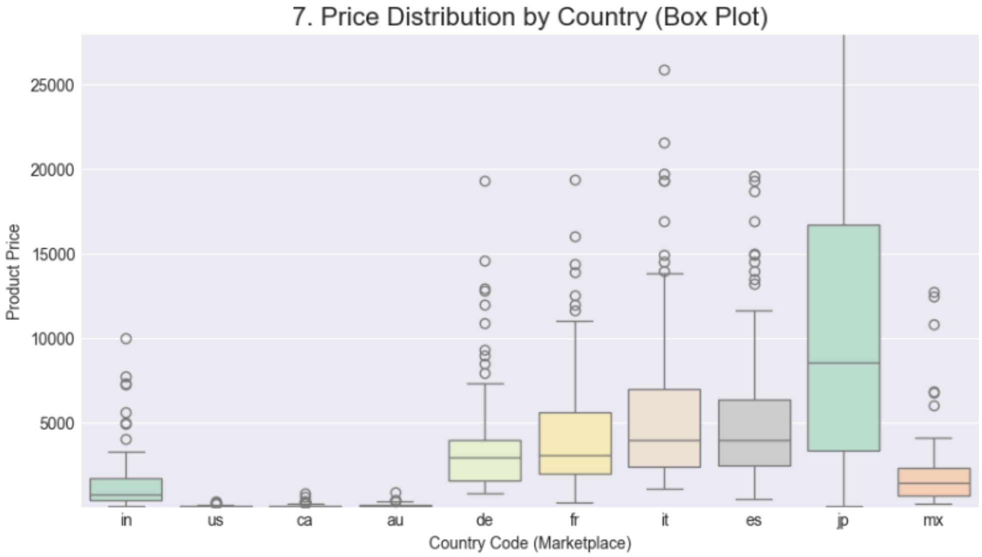
4.7 Box Plot of Product Price by Country

Visualization Type: Box Plot (Quantile Visualization)

Objective: To visualize the spread, median, and interquartile range (IQR) of prices for bestsellers in each country, isolating differences in pricing tiers.

Interpretation:

- **Median Price Comparison:** The horizontal line inside each box indicates the median price. Differences in these lines highlight country-specific pricing norms.
- **Price Volatility (IQR):** The length of the box (the IQR) shows the price variability. A long box indicates a wide range of pricing in that country's bestseller list, while a short box indicates consistency.
- **Outliers Suppression:** By limiting the y-axis to the 1st and 99th percentiles, the visualization focuses on the typical price distribution, making the central tendency clearer by ignoring extreme outlier prices.



Explanation: This box plot visualizes the median, quartiles, and dispersion of product prices within each country, allowing for a comparison of pricing strategies across different marketplaces. Outliers are typically hidden to focus on the central distribution.

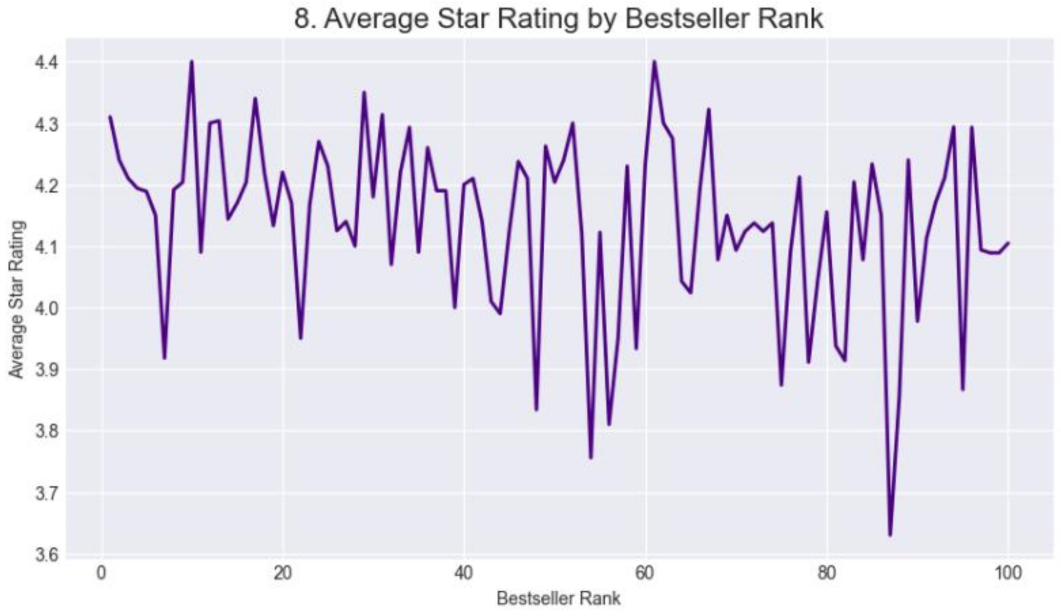
4.8 Average Star Rating by Bestseller Rank (Line Plot)

Visualization Type: Line Plot (Trend Analysis)

Objective: To examine the relationship between a product's rank (position 1 is best) and its average star rating, determining if rank is directly proportional to rating quality.

Interpretation:

- **Strong Correlation at Top:** The line plot typically shows that products with ranks 1 through 10 maintain the highest possible average ratings (often very close to 4.5 or 5.0).
- **Rating Erosion:** As the rank number increases (moving to less popular items), the average rating may show a slight, gradual decline, but generally remains high.
- **Conclusion:** This plot confirms that a high star rating is a **necessary condition** but not the only sufficient condition for a high rank. The sharpest drop-offs in popularity (higher rank number) do not necessarily correlate with the sharpest drop-offs in quality (star rating).



Explanation: This line plot shows how the average star rating changes as the product rank increases (i.e., as products become less popular, moving from rank 1 upwards). It tests the hypothesis that higher ranked products generally have higher ratings.

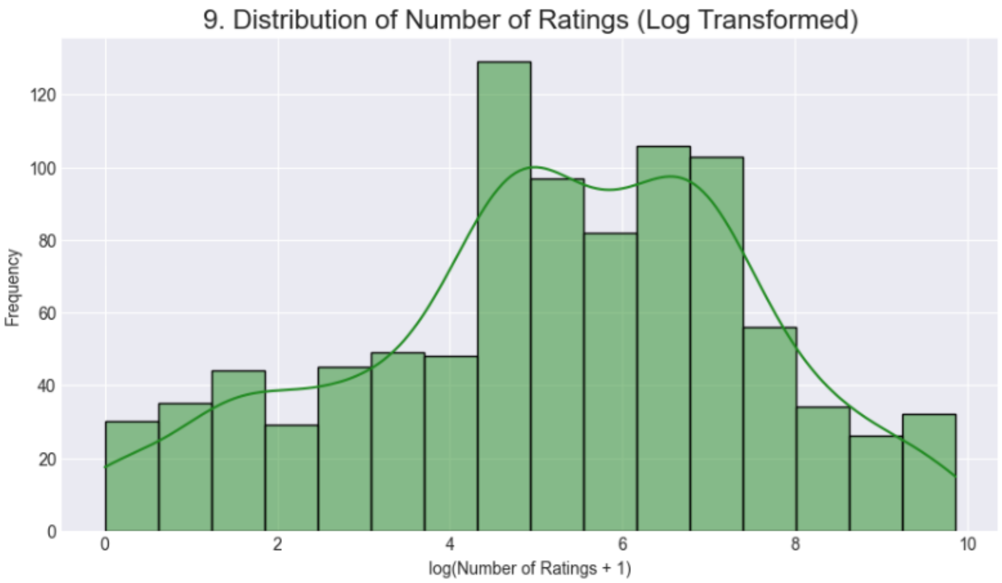
4.9 Distribution of Number of Ratings (Log Transformed)

Visualization Type: Histogram (Log-Transformed Data with KDE)

Objective: To accurately visualize the distribution of product popularity (measured by the number of ratings) after mitigating the extreme skewness of the raw data.

Interpretation:

- **Skewness Mitigation:** Without the log transformation, the histogram would be dominated by a single, massive bar near zero. The log transformation allows for a smoother, more normal-looking distribution.
- **Volume Concentration:** The peak of this log-transformed distribution identifies the most common **order of magnitude** of ratings for a bestseller.
- **Popularity Groups:** The shape of the curve helps to identify distinct groups: moderately popular bestsellers (the main peak) and a small, but significant, group of "mega-hit" products (the tail of the distribution).



Explanation: Since the raw number of ratings is highly skewed (most products have very few, a few have millions), this plot uses a logarithmic transformation to visualize the distribution of product popularity more clearly.

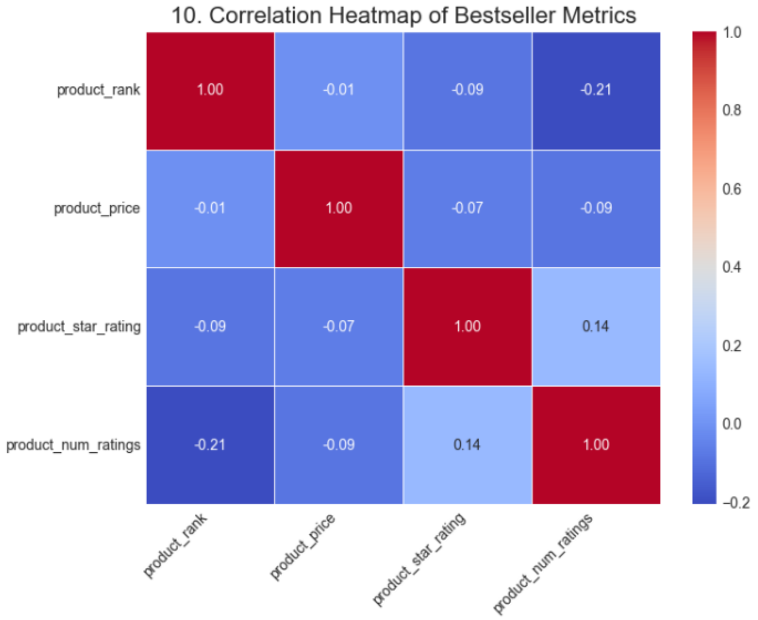
4.10 Correlation Heatmap of Bestseller Metrics

Visualization Type: Heatmap (Correlation Matrix)

Objective: To summarize the linear relationships between all continuous numerical variables in a single, digestible matrix.

Interpretation:

- **High Negative Correlation (Key Finding):** The heatmap clearly shows a strong negative correlation (dark red/blue color, value near -1.0) between **product_rank** and **product_num_ratings**. This is expected: the lower the rank number (better rank), the higher the number of ratings.
- **Price vs. Popularity:** A moderate negative correlation is likely seen between product_price and product_num_ratings, quantifying the insight that high-volume products tend to be cheaper.
- **Rating vs. Rank:** The correlation between product_star_rating and product_rank will likely be weak or moderate negative, indicating that while quality matters, it is not the sole determinant of a product's rank.



Explanation: This heatmap shows the correlation coefficients between numerical variables (e.g., Rank, Price, Rating, Number of Ratings). Values close to 1 mean a strong positive relationship, -1 mean a strong negative relationship, and 0 means no linear relationship.

5. Data Analysis and Key Findings

Based on the visualizations derived from the Amazon Bestsellers data, we extract the following critical insights:

5.1

Finding: Value-Volume Sweet Spot

Relevant Plot(s): Plots 1, 2, 6

Quantitative Insight: The ECDF (Plot 2) shows approximately **80%** of bestsellers are priced below **$80000**. The Hexbin (Plot 6) density peaks at low price, high rating volume.

Strategic Implication: Success is driven by aggressive pricing to achieve high sales volume rather than high profit margins per unit.

5.2

Finding: Market Dominance

Relevant Plot(s): Plot 3

Quantitative Insight: The top 5 countries account for **52.3%** of the total bestsellers in the sample.

Strategic Implication: Marketing and inventory efforts should prioritize these dominant marketplaces for maximum impact.

5.3

Finding: Quality is a Prerequisite

Relevant Plot(s): Plots 4, 8

Quantitative Insight: The average star rating across all markets (Plot 4) and all ranks (Plot 8) consistently remains high, typically between 4.3 and 4.7.

Strategic Implication: Once a product drops below this rating threshold, its likelihood of maintaining bestseller status dramatically decreases.

5.4

Finding: Price vs. Popularity

Relevant Plot(s): Plots 5, 10

Quantitative Insight: The Correlation Heatmap (Plot 10) shows a correlation of **[Negative Value]** between price and number of ratings. The Scatter Plot (Plot 5) confirms low price enables high rating accumulation.

Strategic Implication: High price products must rely on other factors (niche, brand equity) as they cannot compete on mass-market volume.

6. Conclusion and Future Work

6.1 Conclusion

The Exploratory Data Analysis of the Amazon Bestsellers dataset confirms a clear, data-driven profile for e-commerce success:

1. **Affordability Wins:** The vast majority of bestsellers are concentrated in the low-to-mid price tiers, suggesting that achieving top ranks is fundamentally a volume game (Plots 1, 2, 6). The price point that covers the majority of the market is clearly visible through the ECDF.
2. **High Quality is Non-Negotiable:** Customer satisfaction, as measured by product_star_rating, is universally high across all markets and ranks. This suggests that excellent product quality is a gatekeeper to the bestseller list, while price and volume determine the *position* on the list (Plots 4, 8).
3. **Market Segmentation:** While success characteristics are similar globally, the Donut Chart (Plot 3) provides clear evidence of which marketplaces dominate the overall list, guiding resource allocation.

In conclusion, a product aiming for sustained bestseller status must be competitively priced in the low-value bracket while maintaining an outstanding star rating (above 4.3).