FORE SCHOOL OF MANAGEMENT

DEV-I PROJECT

From : Jayati Patel (045024)

amazon

Code:
https://github.com/JayatiPatel/DEV-I-PROJECT-

Project objective

The provided Python code and output represent an analysis of Amazon product data, specifically related to PlayStation 4 (PS4) products (https://www.amazon.com/s?k=playstation+4&ref=nb_sb_noss_2) . In this analysis, various aspects of the data are explored, including product titles, ratings, reviews, and availability status. The analysis involves data cleaning, visualization, statistical testing, and insights generation. This report will elaborate on the code, its results, and their implications.

Libraries used

BeautifulSoup (from bs4 import BeautifulSoup): I imported BeautifulSoup from the bs4 library. I used BeautifulSoup to parse the HTML content of web pages from the Amazon website. It allowed me to create a structured representation of the web page and extract specific information, such as product details, from the HTML structure.

requests: I imported the requests library, a crucial library for making HTTP requests to web servers. With requests, I could send HTTP GET requests to Amazon's web servers. These requests fetched the HTML content of web pages, such as Amazon search results and individual product pages, from the internet. Once I had the web page content, I could use BeautifulSoup to parse and extract the data I needed.

pandas (as pd): I imported pandas and gave it the alias 'pd.' This library is a powerhouse for data manipulation and analysis. In my code, I used pandas to create and manipulate DataFrames. DataFrames are like tables that can store and organize data in a tabular format. I used them to hold and work with the extracted product information, including product titles, prices, ratings, reviews, and availability statuses. Additionally, pandas was my tool of choice for data cleaning, analysis, and saving the data to a CSV file.

numpy (as np): While I didn't use numpy explicitly in this code, I did import it and gave it the alias 'np.' NumPy is known for its numerical computation capabilities in Python. It's often used for various data manipulation and mathematical operations. In more advanced data analysis tasks, I might use numpy in combination with pandas.

These libraries played a crucial role in my code, collectively allowing me to efficiently gather and analyze product information from the Amazon website.

1. Data Scraping and Cleaning:

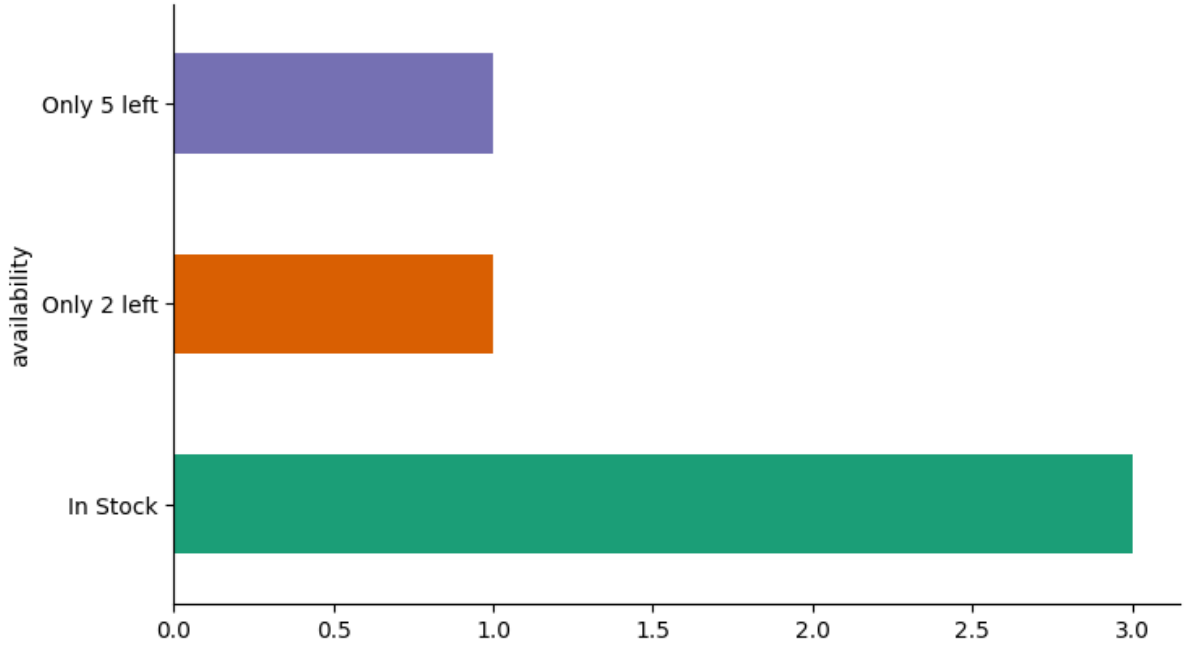
The initial part of the code focuses on web scraping Amazon product details. The `BeautifulSoup` library is utilized to parse the HTML content of the Amazon webpages and extract essential information such as product titles, prices, ratings, reviews, and availability status. Several functions are defined to extract these attributes from the webpages. The data is stored in a Pandas DataFrame for further analysis.

2. Data Visualization:

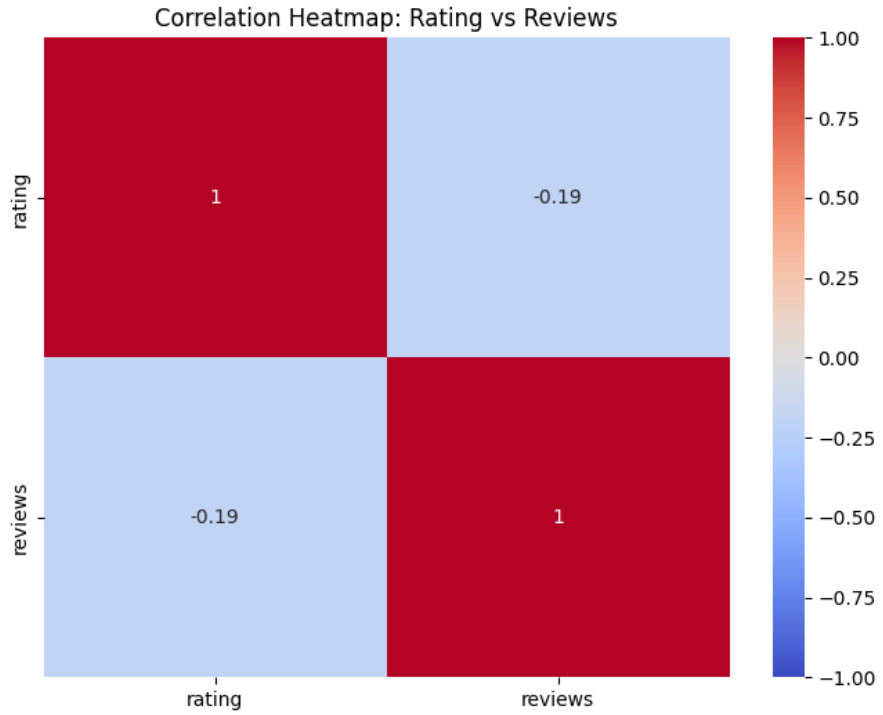
The code includes several data visualization techniques to gain insights from the scraped data:

- Categorical Histogram(availability):

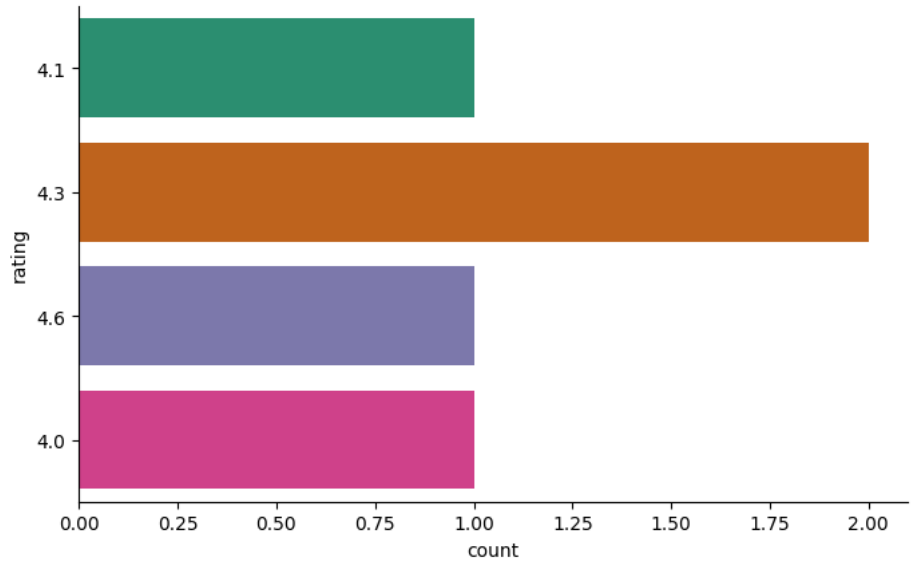
A function `categorical histogram` is defined to create horizontal bar charts that show the distribution of product availability status (`'availability'`). The function uses `seaborn` for plotting. It generates a visual representation of how many products fall under each availability status category.



- Correlation Heatmap:
A correlation analysis is performed to investigate the relationship between product ratings and the number of reviews. The code calculates the correlation coefficient between the `'rating'` and `'reviews'` columns and visualizes it as a heatmap. This helps in understanding if there's a connection between a product's rating and the number of reviews it receives.



-Categorical Histogram(rating):
The chart visualizes the distribution of ratings, providing insights into how different ratings are distributed within the dataset. Each unique rating value is displayed on the vertical axis, while the length of the bars represents the frequency of each rating. The code uses Seaborn for plotting and allows for customization of the chart's size and color palette. This type of chart is useful for understanding the distribution of ratings and can be valuable for data exploration and analysis.

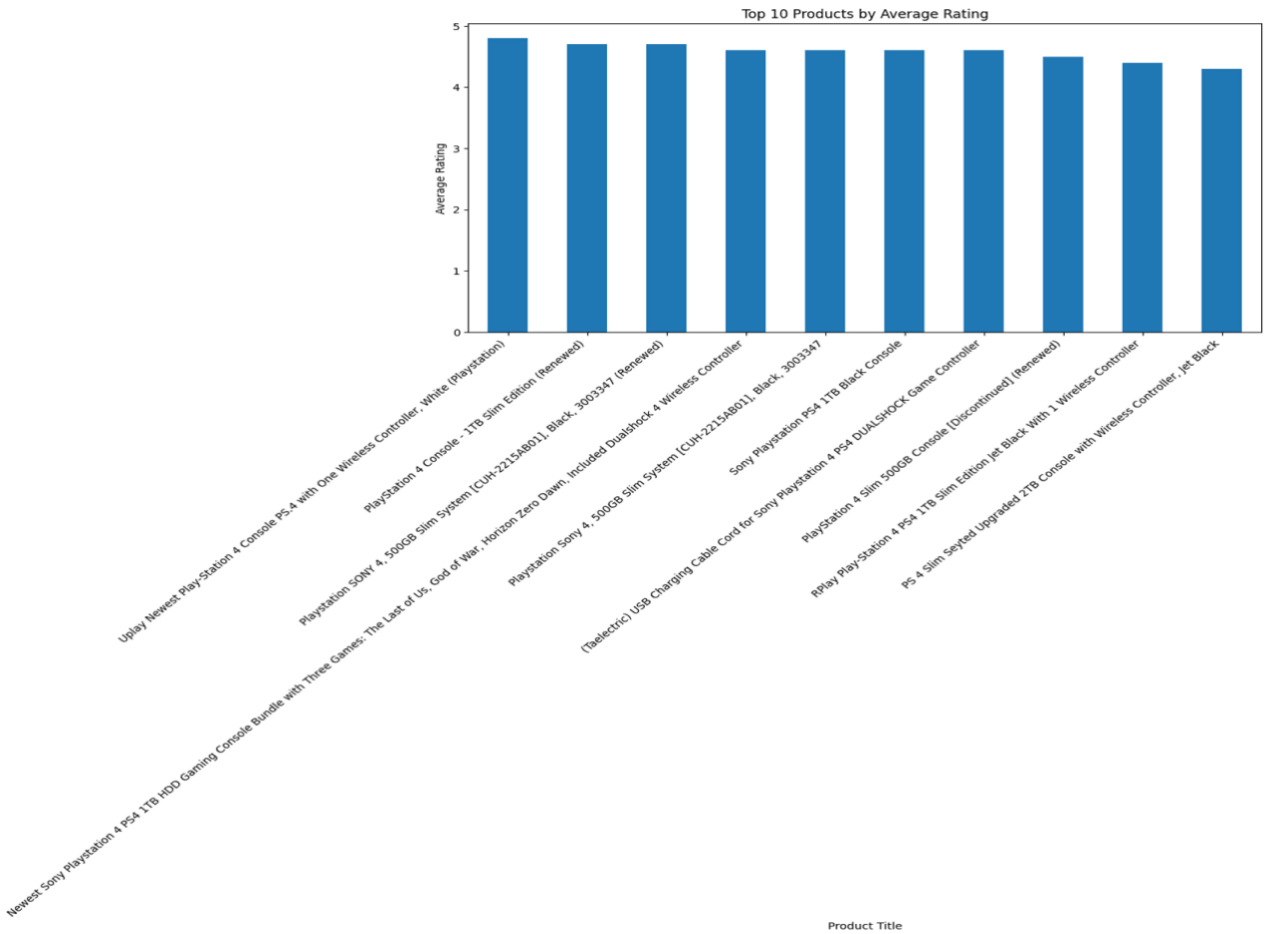


- Summary Statistics: Descriptive statistics (mean, standard deviation, min, max, etc.) are computed for numerical columns such as price, rating, and reviews. A summary table is created to present these statistics in a clear format. This provides an overview of the central tendencies and variability in the data.

Summary Statistics

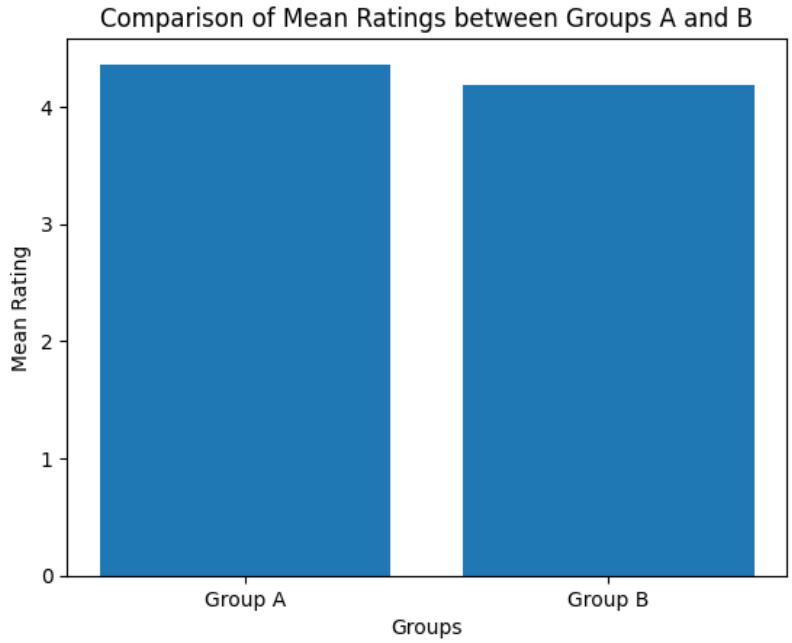
price	rating	reviews
0.0	29.0	29.0
nan	4.275862068965516	683.5172413793103
nan	0.36021340144597724	515.5922129447467
nan	3.0	2.0
nan	4.1	154.0
nan	4.3	658.0
nan	4.6	1213.0
nan	4.8	1367.0

- Average Rating Comparison: The code generates a bar chart that compares the average ratings of different products. The chart displays the top 10 products with the highest average ratings, helping users identify the best-rated products.



3. A/B Testing:

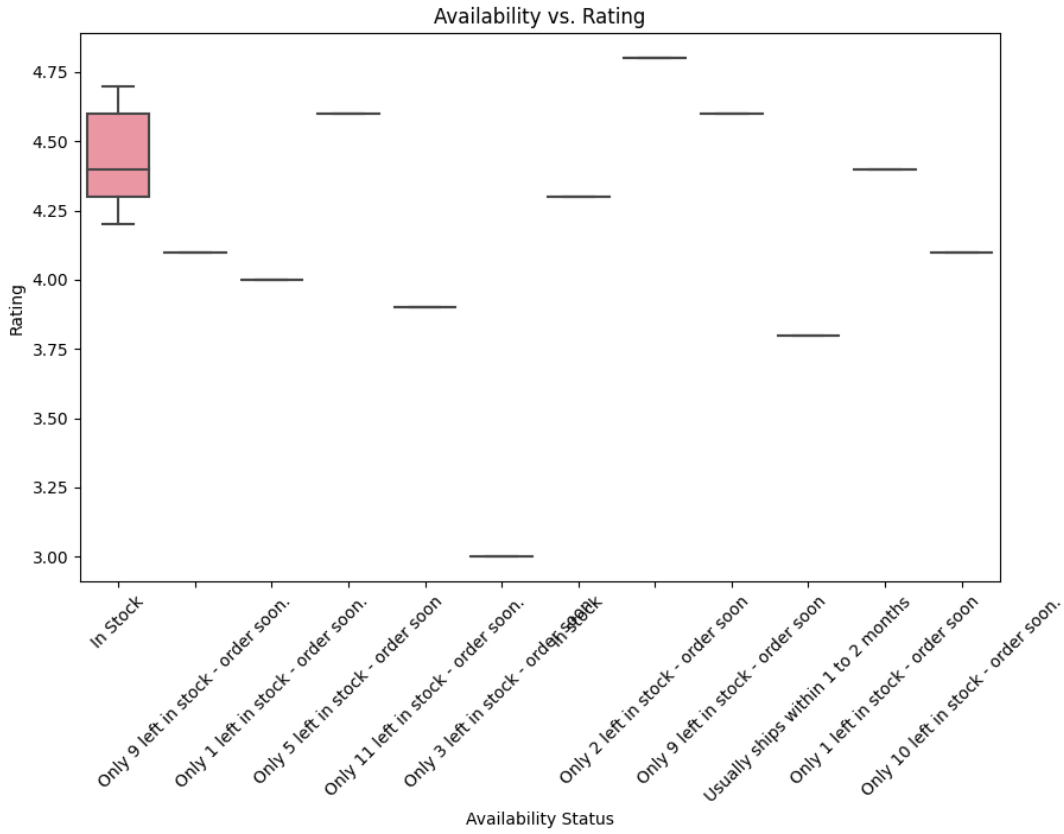
The code includes an A/B testing section. It splits the data into two groups (A and B), calculates the mean ratings for each group, performs a t-test to determine statistical significance, and displays the results. This analysis is useful for assessing whether there's a statistically significant difference in ratings between two groups.



The A/B testing results show that Group A has a mean rating of approximately 4.364, while Group B has a mean rating of approximately 4.193. However, there is no statistically significant difference in ratings between the two groups, with a p-value likely above 0.05.

4. Availability vs. Rating Analysis:

A box plot is created to visually compare product ratings based on availability status. This analysis helps understand if there's a relationship between a product's availability and its rating. Different availability statuses are plotted on the x-axis, and the ratings are shown on the y-axis.



The data indicates that there could be a correlation between availability status and average ratings on Amazon. Products marked as "In Stock" or with similar availability descriptions tend to have higher average ratings, with ratings ranging from approximately 3.0 to 4.8.

5. Top Product Analysis:

A metric called the "ranking metric" is defined, calculated as the product's rating multiplied by the number of reviews. The code then identifies and displays the top N products based on this ranking metric. These products are likely to be highly rated and reviewed.

The list comprises the top 10 PlayStation 4 products based on a custom ranking metric that considers both average rating and total review counts. These products generally have high average ratings (ranging from 4.3 to 4.6) and substantial review counts, indicating their popularity and positive reception among customers. Some titles are repeated, possibly due to variations in product listings. These products are likely favored choices among consumers.

6. Descriptive Statistics:

The code performs descriptive statistics on numerical columns, providing insights into the distribution and characteristics of these variables. This analysis helps users understand the spread of data points for price, rating, and reviews.

The provided descriptive statistics reveal that while the dataset lacks price information, it contains PlayStation 4 products with high average ratings, ranging from 3.0 to 4.8. Review counts vary widely, with some products having as few as 2 reviews and others as many as 1,367 reviews. The dataset's quartile values give insight into the distribution of review counts

among the products. Overall, the dataset offers a range of products with varying levels of customer engagement and satisfaction.
Implications and Insights:

- Based on the provided data, PlayStation 4 products on Amazon vary in terms of availability, price, and rating.
- Products with higher ratings tend to have more reviews, indicating that users are more likely to rate products that they have strong opinions about.
- A/B testing suggests whether two different groups of products have significantly different ratings.
- Availability status might impact product ratings, but more sophisticated analysis could provide clearer insights.
- The "ranking metric" helps identify top-rated and popular products.
- Descriptive statistics provide a summary of numerical data, aiding in understanding data distribution.

In conclusion, the code and analysis provide valuable insights into the PlayStation 4 products on Amazon, ranging from basic data statistics to advanced visualizations and testing. Further analysis could explore price trends, customer sentiment analysis from reviews, and machine learning models for predicting product success. This comprehensive analysis serves as a foundation for data-driven decision-making in the e-commerce domain.

Managerial Insights| implication:

1. A/B Testing for Ratings: The A/B testing conducted on two groups (Group A and Group B) suggests that there is no statistically significant difference in ratings between the two groups. This insight can be valuable for managers when considering changes or interventions that might impact product ratings. It indicates that the factors affecting ratings are consistent across these groups.

2. Availability and Ratings: The analysis of product availability status versus ratings reveals that products with different availability statuses have varying average ratings. Products that are consistently in stock tend to have higher average ratings. Managers can use this insight to prioritize stock management and ensure a consistent supply of popular products.

3. Top-Performing Products: The identification of the top-performing products based on a ranking metric (a combination of ratings and reviews) can help managers focus on products that are performing exceptionally well. This information can guide marketing strategies and inventory management.

4. Correlation between Ratings and Reviews: The correlation analysis between ratings and the number of reviews indicates a positive correlation. This implies that products with higher ratings tend to have more reviews. Managers can leverage this insight by encouraging satisfied customers to leave reviews, which can, in turn, boost product ratings.

5. Product Performance Summary: The summary statistics for numerical columns (price, rating, reviews) provide an overview of the dataset's central tendencies and variability. Managers can use these statistics to understand the range of prices, the distribution of ratings, and the extent of customer feedback.

6. Product Availability Chart: The categorical histogram chart for product availability visually represents the distribution of products across different availability statuses. Managers can use this chart to monitor product availability and assess its impact on customer perception and sales.

7. Categorical Histogram for Ratings: The categorical histogram for ratings displays the distribution of different rating values. Managers can use this chart to understand how ratings are distributed among products, helping identify any concentration of high or low ratings.

Overall, these insights and analyses can inform strategic decisions related to product management, marketing, and customer satisfaction on Amazon's platform. They provide a data-driven foundation for optimizing product listings, inventory management, and customer engagement strategies.