Pandas Assignment: Amazon Product Reviews Analysis

Overview

In this assignment, you will work with a dataset 'amazonproductreviews.csv' that contains over 1,000 Amazon product reviews and ratings. Your tasks involve cleaning, transforming, and analyzing the data to extract insights related to product popularity, customer satisfaction, and pricing strategies.

Data Description

Download Data from here
https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset/data
The dataset 'amazonproductreviews.csv' contains the following columns:
- `product_id`: Product ID
- `product_name`: Name of the Product
- `category`: Category of the Product
- `discounted_price`: Discounted Price of the Product
- `actual_price`: Actual Price of the Product
- `discount_percentage`: Percentage of Discount for the Product
- `rating`: Rating of the Product
- `rating_count`: Number of people who voted for the Amazon rating
- `about_product`: Description of the Product
- `user_id`: ID of the user who wrote a review for the Product
- `user_name`: Name of the user who wrote a review for the Product
- `review_id`: ID of the user review
- `review_title`: Short review
- `review_content`: Long review
- `img_link`: Image Link of the Product
- `product_link`: Official Website Link of the Product

Tasks

1. Data Cleaning

- Load the 'amazonproductreviews.csv' file into a DataFrame and display the first 10 rows.
- Check for any missing values in the dataset. If there are any, handle them appropriately.
- Ensure all price-related columns (`discountedprice`, `actualprice`) are in numeric format and handle any inconsistencies found.
- Verify that the `rating` column is in a numeric format and handle non-numeric values if any.

2. Data Transformation

- Calculate the amount of discount in currency for each product and add it as a new column `discount_amount`.
- Create a new column `reviewlength` that contains the number of words in the `reviewcontent`.

3. Data Analysis & Visualization

- Identify the top 10 most reviewed products.
- Determine the top 5 categories with the highest average rating.
- Create a histogram showing the distribution of ratings across all products.

4. Insightful Analytics

- Calculate the average rating per user and identify the top 5 users who have given the highest average ratings.
- Analyze the relationship between `discount_percentage` and the average `rating` for a product. Is there any visible trend?
- For each category, calculate the average `discountedprice` and compare it to the average `actualprice`. Which category has the highest average discount in terms of currency?

5. Report Generation

- Generate a summary report that includes the following details:
  - Total number of reviews
  - Average rating across all products

    - Product with the most reviews
    - Category with the highest average discount
  - Export this report to a CSV file named 'amazonproductsummary.csv'.

Deliverables

- A Jupyter notebook containing all the code for the tasks above, along with comments explaining your steps.
- The 'amazonproductsummary.csv' file generated in Task 5.
- Any plots generated as part of Task 3 and 4 should be clearly labeled and have a title.

Evaluation Criteria

- Correctness of code.
- The efficiency of code (use of appropriate Pandas methods to perform tasks).
- Clarity of code comments and explanations.
- Quality and clarity of the plots generated, including appropriate use of labels and titles.