Python Project:

E-Commerce Product Review Analysis

Submitted by

Group - 10

Roopesh O R

Roshna Palatty Santhosh

Merella Jobi

E-Commerce Product Review Analysis

March 30, 2025

Aim

The aim is to analyze customer reviews and to find insights into product ratings.

Tools used

- Jupyter notebook
- `pandas`, `numpy` and `matplotlib` libraries

Processing & Analysis

The sample dataset used in the analysis contains 65 entries with 6 fields. It contains product reviews from 4 categories. Here following operations were done on the data:

- Rows with missing values in any of the following columns were removed:
  - product_id
  - product_name
  - category
  - customer_id
  - score
- Duplicate reviews were removed if the review was posted by same person in a given product

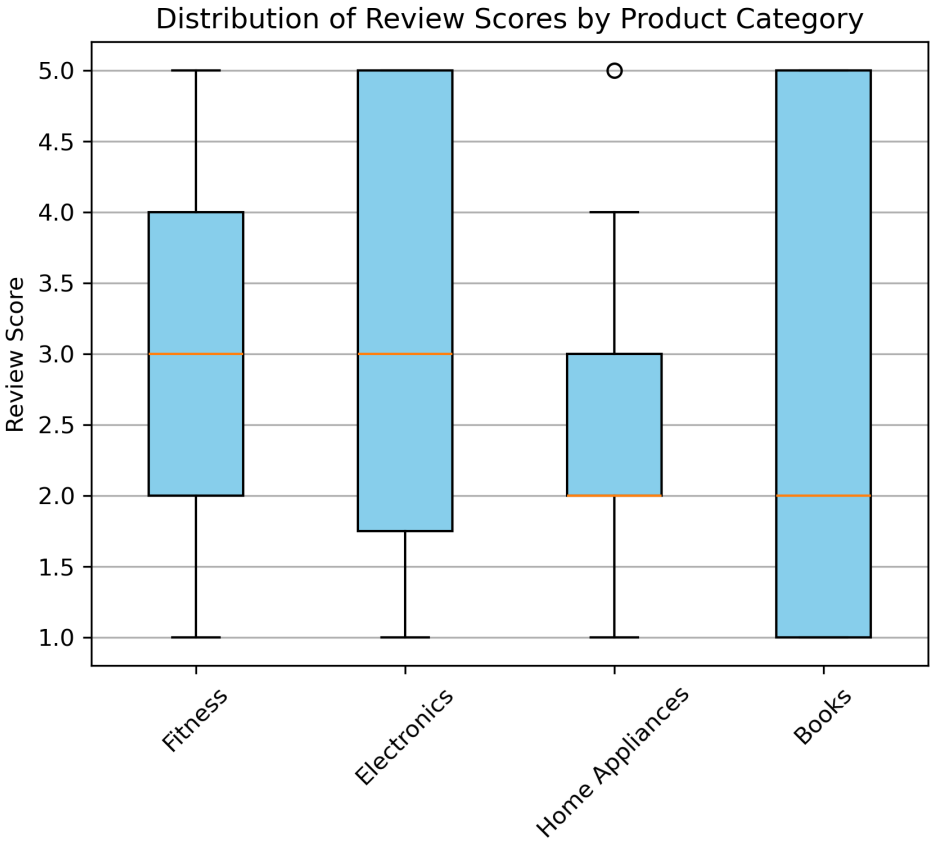
Then box plot showing distribution of review scores in each category and a barchart showing total number of reviews in each category was plotted.

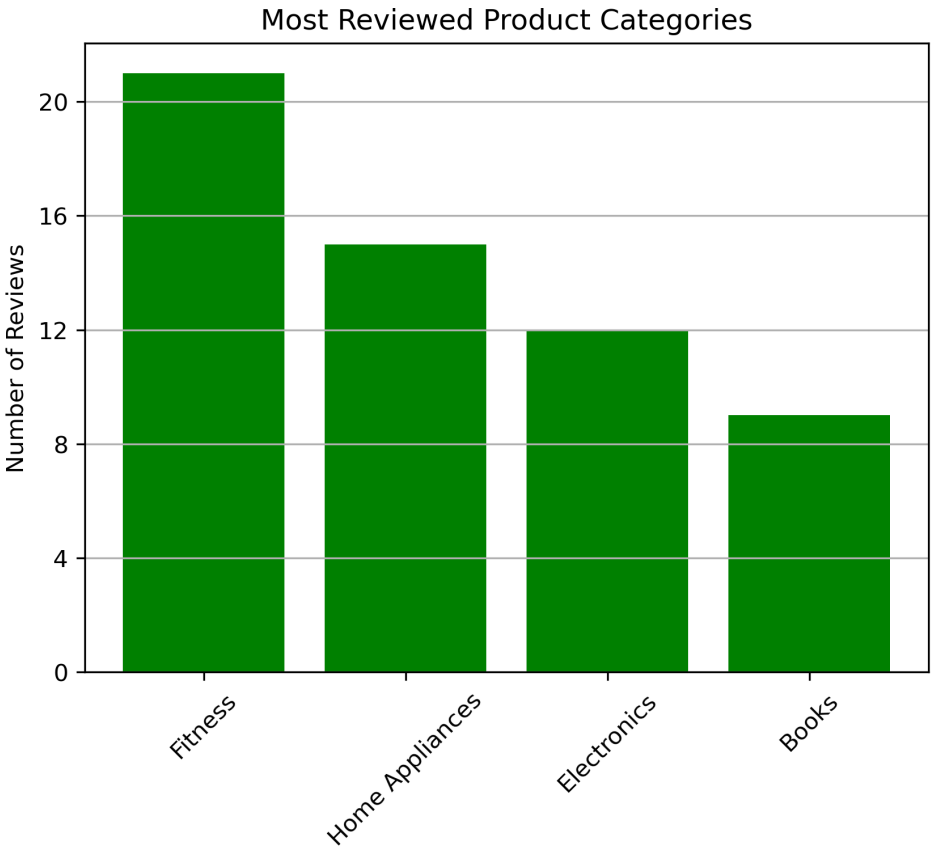
Source code

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt


df =  pd.read_csv("reviews.csv")

df = df.replace("", np.nan)
df["score"] = pd.to_numeric(df["score"], errors='coerce')

# delete all rows containing NaN in the specified columns
df = df.dropna(subset=[
    "product_id","product_name","category","customer_id","score"
])
```

```
# remove duplicate reviews (removes multiple reviews from same
# person in same product)
df = df.drop_duplicates(subset=["product_id", "customer_id"])


## plot boxplot of distribution of ratings in across a category

categories = df["category"].unique()

# creates a list of lists containing scores in each category
category_scores = [
    df[df["category"] == category]["score"] for category in categories
]

plt.figure(dpi=300)
# plots distribtion
plt.boxplot(
    category_scores,
    patch_artist=True,
    boxprops={"facecolor":"skyblue"}
)

# replace x axis ticks by category names
plt.xticks([1,2,3,4], categories, rotation=45)
plt.ylabel("Review Score")
plt.title("Distribution of Review Scores by Product Category")
plt.grid(axis="y") # add grid lines parallel to x axis
plt.show()


## plots barchart showing number of reviews in each category

# counts number of revies in each category. This also sorts it
# in descending order
category_counts = df["category"].value_counts()

plt.figure(dpi=300)
plt.bar(category_counts.index, category_counts.values, color="green")
plt.ylabel("Number of Reviews")
plt.title("Most Reviewed Product Categories")
plt.xticks(rotation=45)
plt.yticks(ticks=range(0, 21, 4))
plt.grid(axis="y")
plt.show()
```

Output


Distribution of Review Scores by Product Category
Review Score
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Fitness
Electronics
Home Appliances
Books


Most Reviewed Product Categories
Number of Reviews
20
16
12
8
4
0
Fitness
Home Appliances
Electronics
Books