

Department of Computer Science

EXPLORATORY DATA ANALYSIS

**COURSE CODE : 21ECSC210**

COURSE PROJECT

AMAZON KDD - PRODUCT PRICE PREDICTION

**TEAM 6**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Name** | **USN** |
| 1 | Amogh M | 01FE21BCS002 |
| 2 | Arfaahussain Shaikh | 01FE21BCS008 |
| 3 | Naveenkumar B | 01FE21BCS051 |
| 4 | Siddharth Hattiholi | 01FE21BCS036 |

CONTENTS

1. Domain Knowledge ………………………………………………………………………………………. 1

2. Problem Statement ………………………………………………………………………………………... 1

3. Objectives ………………………………………………………………………………………………… 1

4. Data Description ………………………………………………………………………………………….. 2

5. Data Pre-processing ………………………………………………………………………………………. 3

6. Data Visualization & Analysis ……………………………………………………………………………. 5

7. Conclusion ……………………………………………………………………………………………….. 14

8. Reference ………………………………………………………………………………………………….14

**Domain Knowledge:**

To accurately predict book prices, it is crucial to understand the complex and multifaceted nature of determining the optimal price for a book. This requires considering various factors, such as book industry dynamics, author-publisher relationships, pricing strategies, industry regulations, and the influence of online marketplaces and e-commerce platforms. By incorporating this domain knowledge into the project, the goal is to empower authors, publishers, and readers with transparent and informed decision-making, while enhancing transparency and fairness within the book industry.

**Problem Statement :**

The Amazon KDD Cup, which focuses on analysis of the data and predicting customer ratings for products based on various features and attributes.

**Objectives :**

1.Explore and analyze the dataset to understand the relationship between various features and ratings.

2.Identify key factors influencing customer ratings and their importance.

3.Predict the rating of the product based on its discounted price.

**Dataset Description :**

The dataset contains information about products on Amazon. It includes various attributes such as product ID, product name, category, discounted price, actual price, discount percentage, rating, rating count, information about the product, user ID, user name, review ID, review title, review content.

**Key features of the dataset:**

•The dataset has 1465 rows × 16 columns.

•And has 13 attributes.

**Attributes :**

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Description** | **Type** |
| product\_id | Provides a unique id of products | Object |
| product\_name | Provides name of the product | Object |
| category | Defines the category of product | Object |
| discounted\_price | Provides discounted price of the product | Object |
| actual\_price | Provides actual price of the product | Object |
| discount\_percentage | Provides percentage of discount of product | Object |
| rating | Provides score ratings | Object |
| rating\_count | Provide total no. Of ratings on product | Object |
| about\_product | Provides description of product | Object |
| user\_id | Provides user id of the customer/reviewer | Object |
| user\_name | Provides user name of the customer/reviewer | Object |
| review\_id | Provides unique review id | Object |
| review\_title | Provides title of a review | Object |
| review\_content | Provides descriptive view of review | Object |
| img\_link | Provides link for image of the product | Object |
| product\_link | Provides hyperlink of the product webpage | Object |

**Data Pre-processing :**

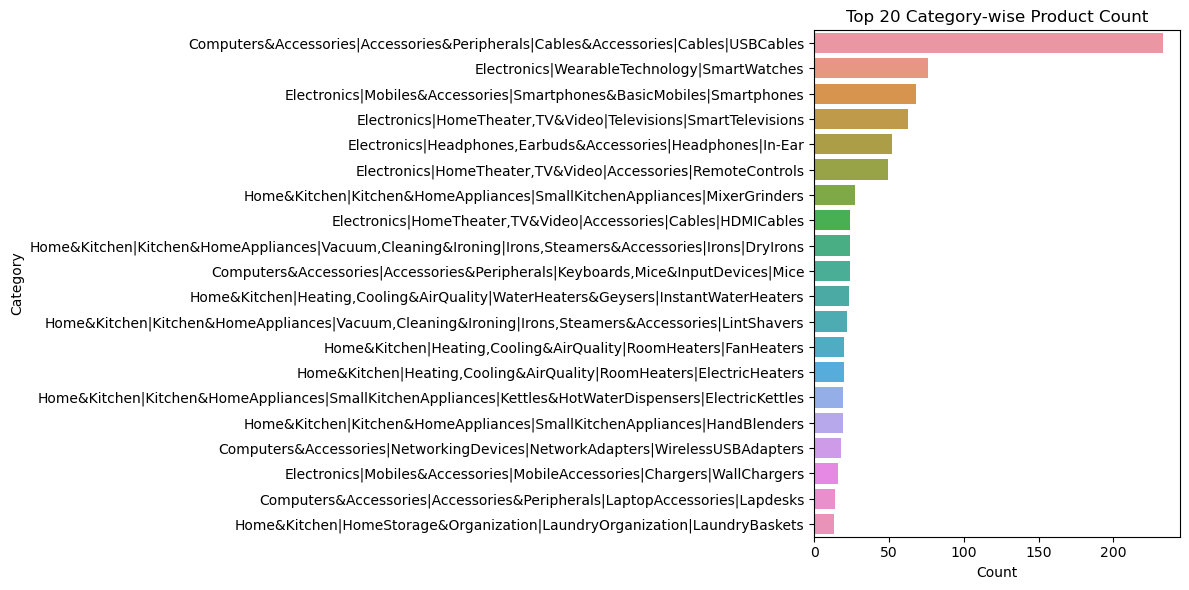
* Checked for null values and the only NULL values we got were in the ‘rating\_count’ column.

**Data Transformation :**

* Replaced the null values in the 'rating\_count' column with the mean value from the column.
* This ensures that the missing values are filled with a representative value from the existing data.
* Converted the 'discounted\_price' and 'actual\_price' columns from float to integer data type.
* Converted the 'rating' column from object to float data type.
* Converted the 'rating\_count' column from object to integer data type. This ensures that the count of ratings is represented as whole numbers.

**Data Visualization and Analysis:**

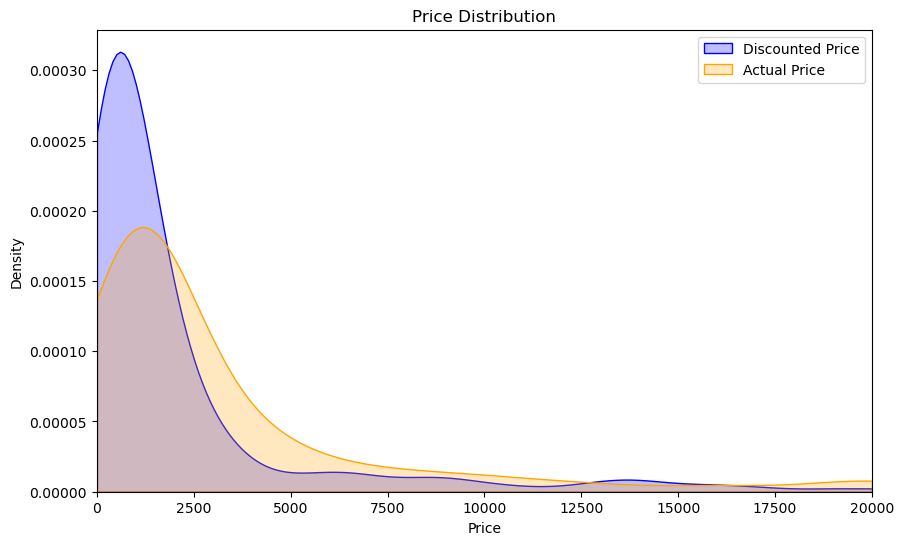
**Analysis 1:** Most popular product categories.



**Inference:**

This plot displays the count of products in each category. It provides insights into the most popular product categories. So, Computers & accessories, cables have the most rating count.

**Analysis 2:** Type of prices preferred among the users.



**Inference:**

Here we can infer that the discounted price has more density than actual price, meaning that people purchase the products with discounted price rather than there actual price.

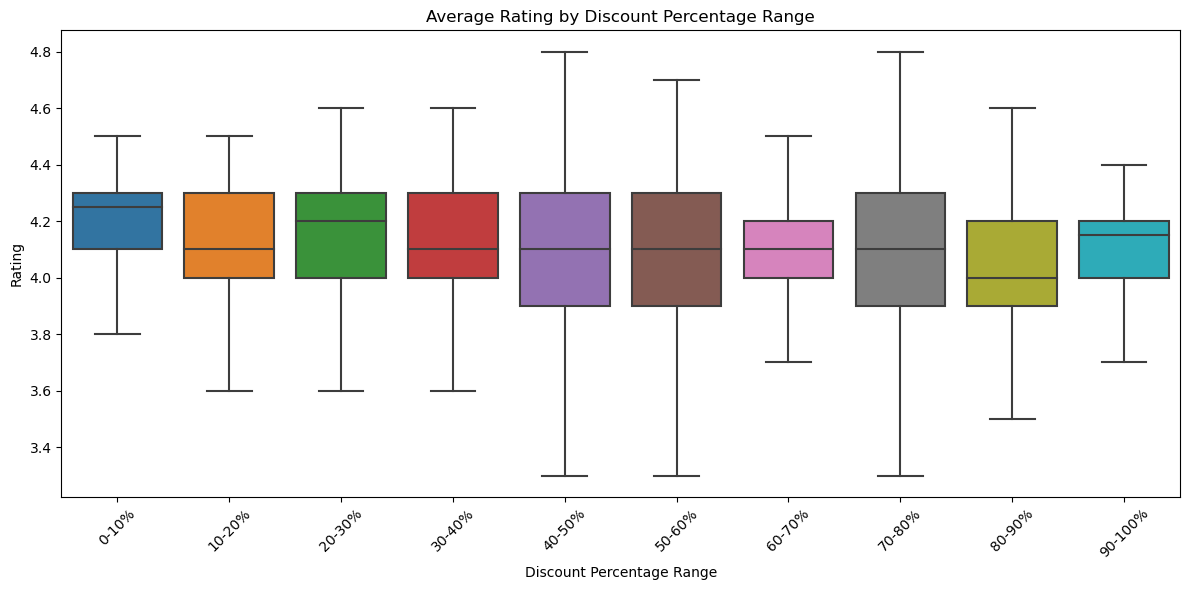
**Analysis 3:** Top 20 product reviewers.



**Inference:**

Here we can infer that the top reviewer is Sethu Madhav , Akash Thakur from burger planet.

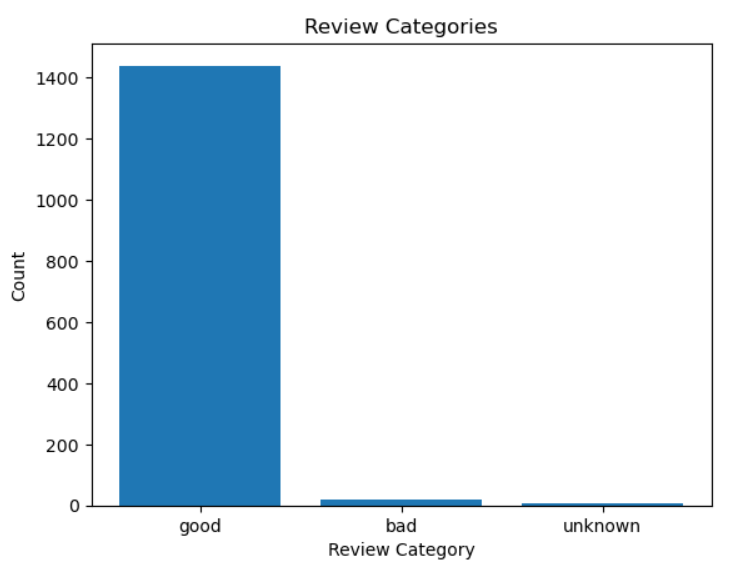
**Analysis 4:** Average rating varying across different discount percentage ranges.



**Inference:**

Here, 40-50% ,50-60%,70-80% , the data has the most spread of the rating.

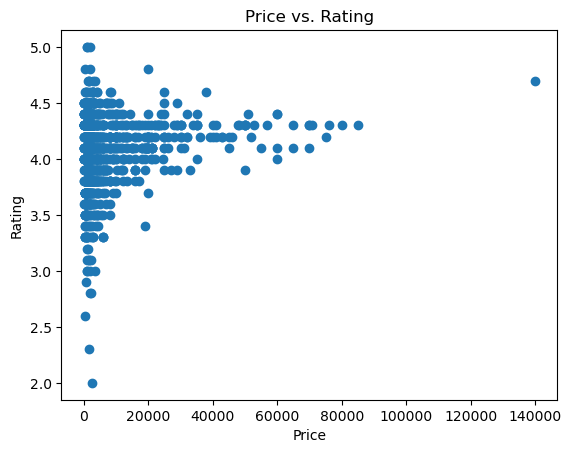
**Analysis 5:** Range of review popularity with highest count.



**Inference:**

Here, we can see that there are more good reviews compared to bad reviews.

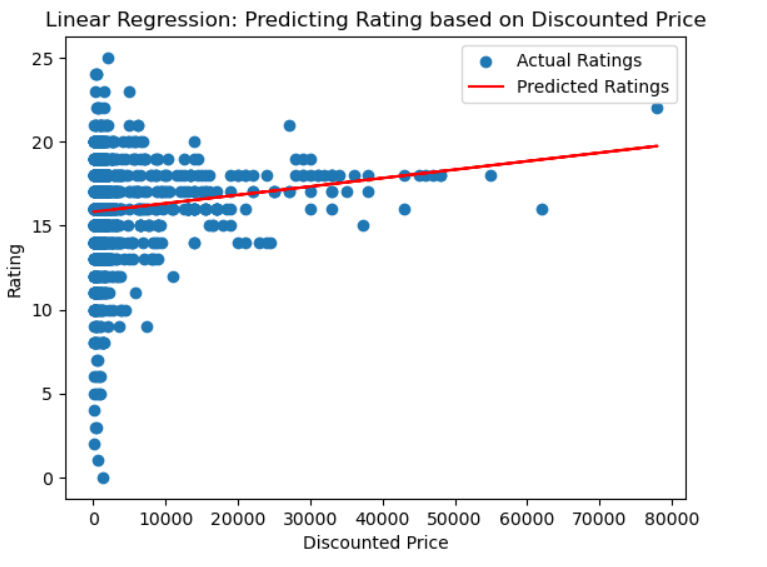
**Analysis 6:** Relationship between the price and rating of products.



**Inference:**

Here, 0-20000 has the highest rating and the rest has less spreading.

**Analysis 7:** Predicting the ratings of products based on their discounted prices.

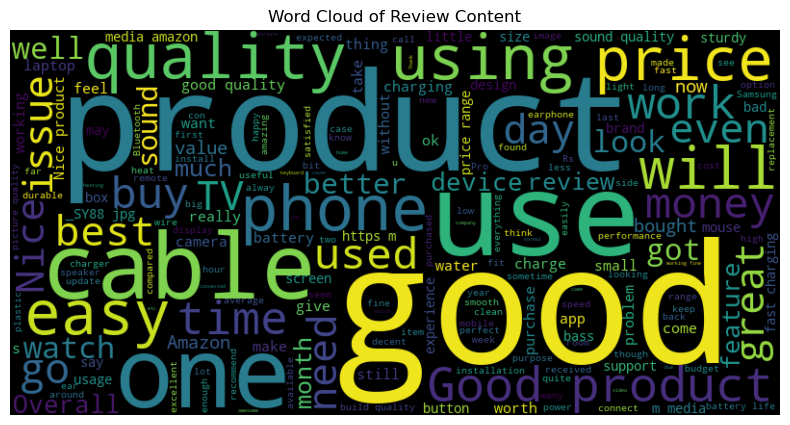


**Inference:**

Here our price lies in the line of the predicted ratings.

Predicted rating of price of 5000: [16.0702859]

**Analysis 8:** World cloud for the reviews.



**Inference:**

Here , product, and good has the most review. And and ok has the least rating.

**Conclusion:**