Analyzing and Predicting Amazon Product Sales Over a Month

Siri Smitha Joginapally – BG98738

Manasvi Pentareddy – CZ04458

Introduction

This project is about analyzing Amazon products based on their price, and rating. There are 2 datasets:

- amazon_categories.csv:
  - Id
  - category_name
- amazon_products.csv:
  - asin
  - Title
  - imgUrl
  - productURL
  - Stars
  - Reviews
  - Price
  - listPrice
  - category_id
  - isBestSeller
  - boughtInLastMonth

https://www.kaggle.com/datasets/asaniczka/amazon-products-dataset-2023-1-4m-products

Preprocessing

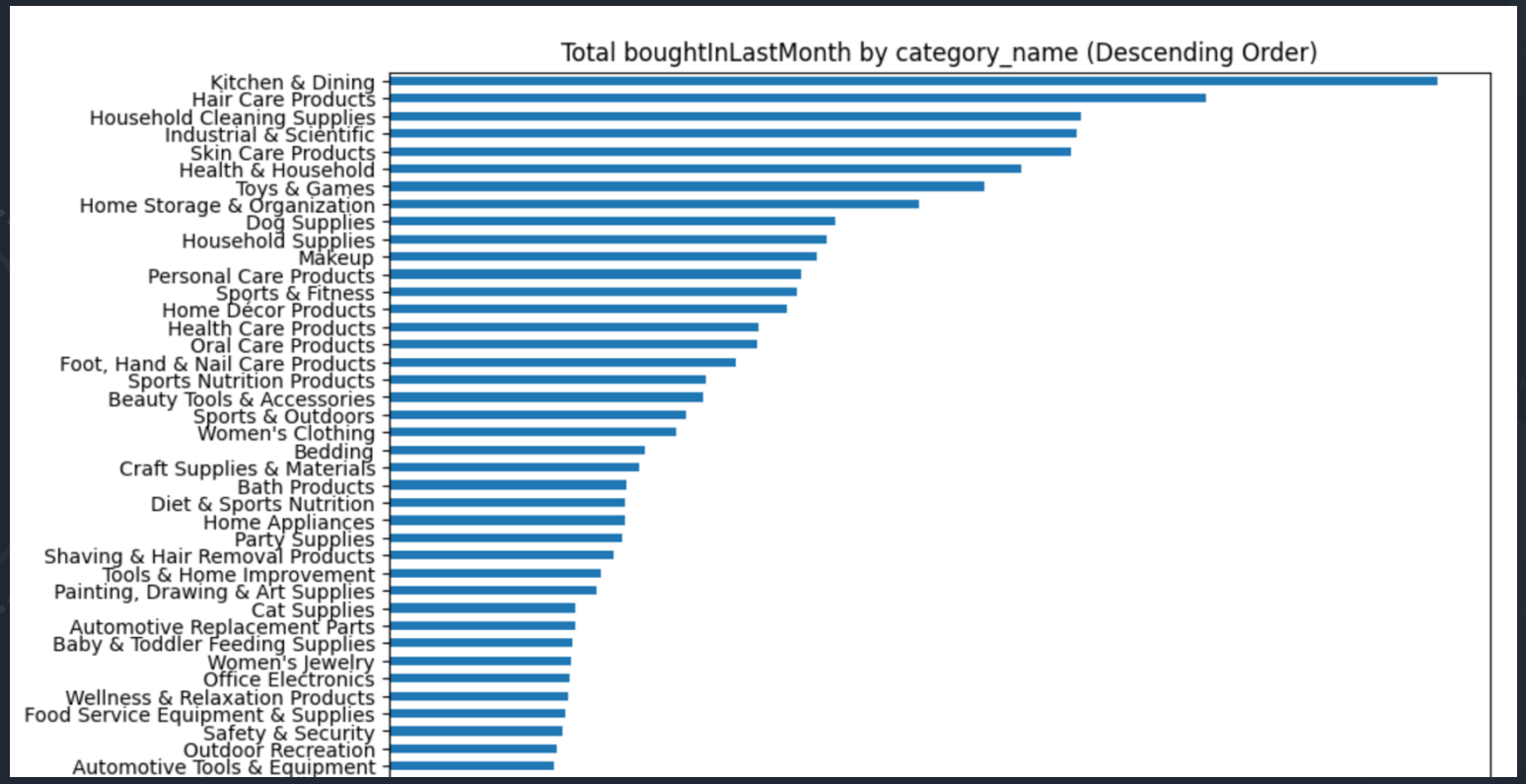
- Merging
- Cleaning
  - Checking for null values
  - Dropping all the unrequired columns
- Changing the datatypes of the columns


```
Null count in id: 0
Null count in category_name: 0
Null count in title: 0
Null count in stars: 0
Null count in reviews: 0
Null count in price: 0
Null count in listPrice: 0
Null count in isBestSeller: 0
Null count in boughtInLastMonth: 0
```

Visualization

Which category based products were sold maximum

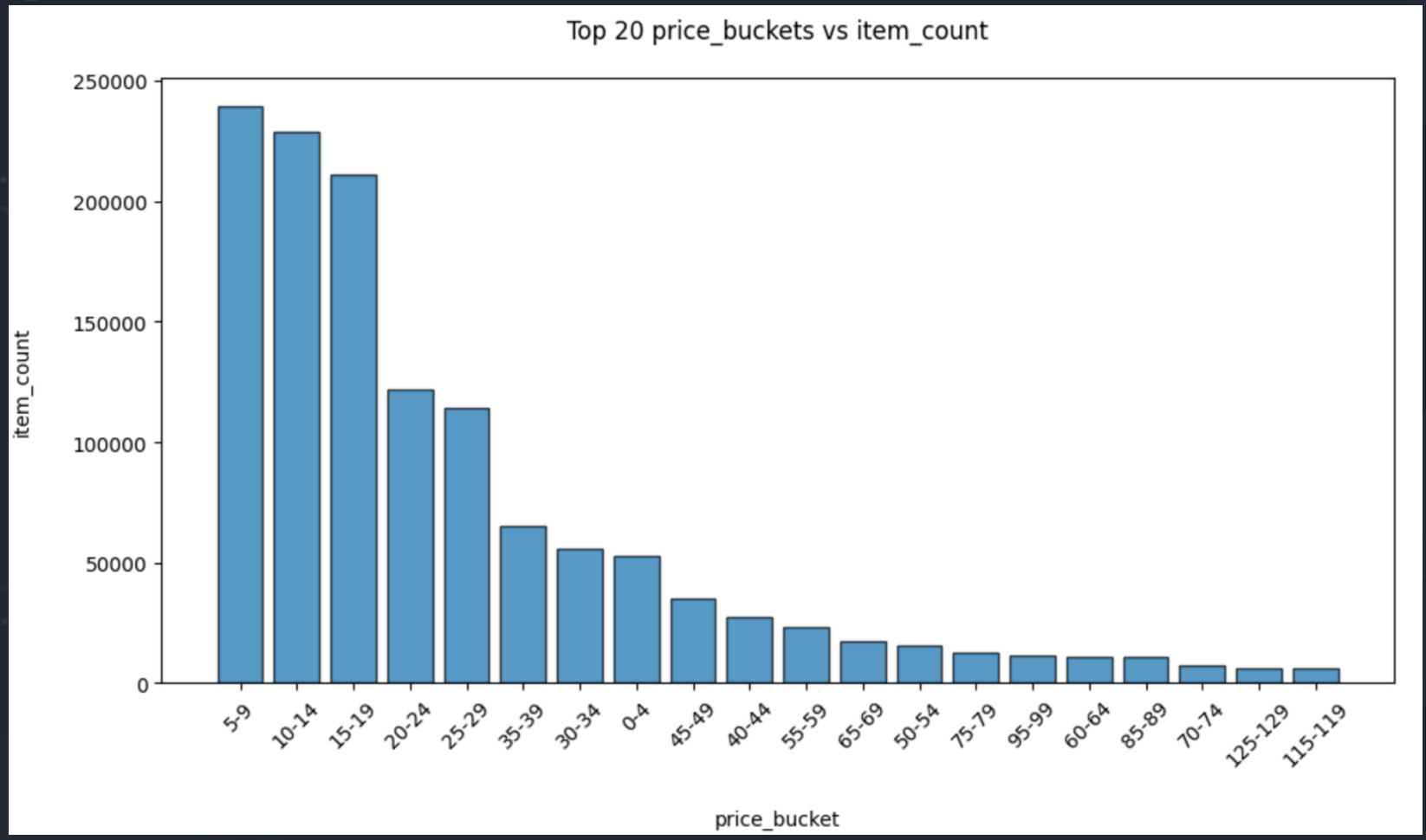
category_name	total_bought
Kitchen & Dining	10191650
Hair Care Products	7930650
Household Cleanin...	6719100
Industrial & Scie...	6685150
Skin Care Products	6604450
Health & Household	6138000
Toys & Games	5779300
Home Storage & Or...	5153650
Dog Supplies	4337850
Household Supplies	4250500
Makeup	4156100
Personal Care Pro...	4001900
Sports & Fitness	3965450
Home Décor Products	3863950
Health Care Products	3590250
Oral Care Products	3577500
Foot, Hand & Nail...	3369450



Visualization

In which price range the products are most sold

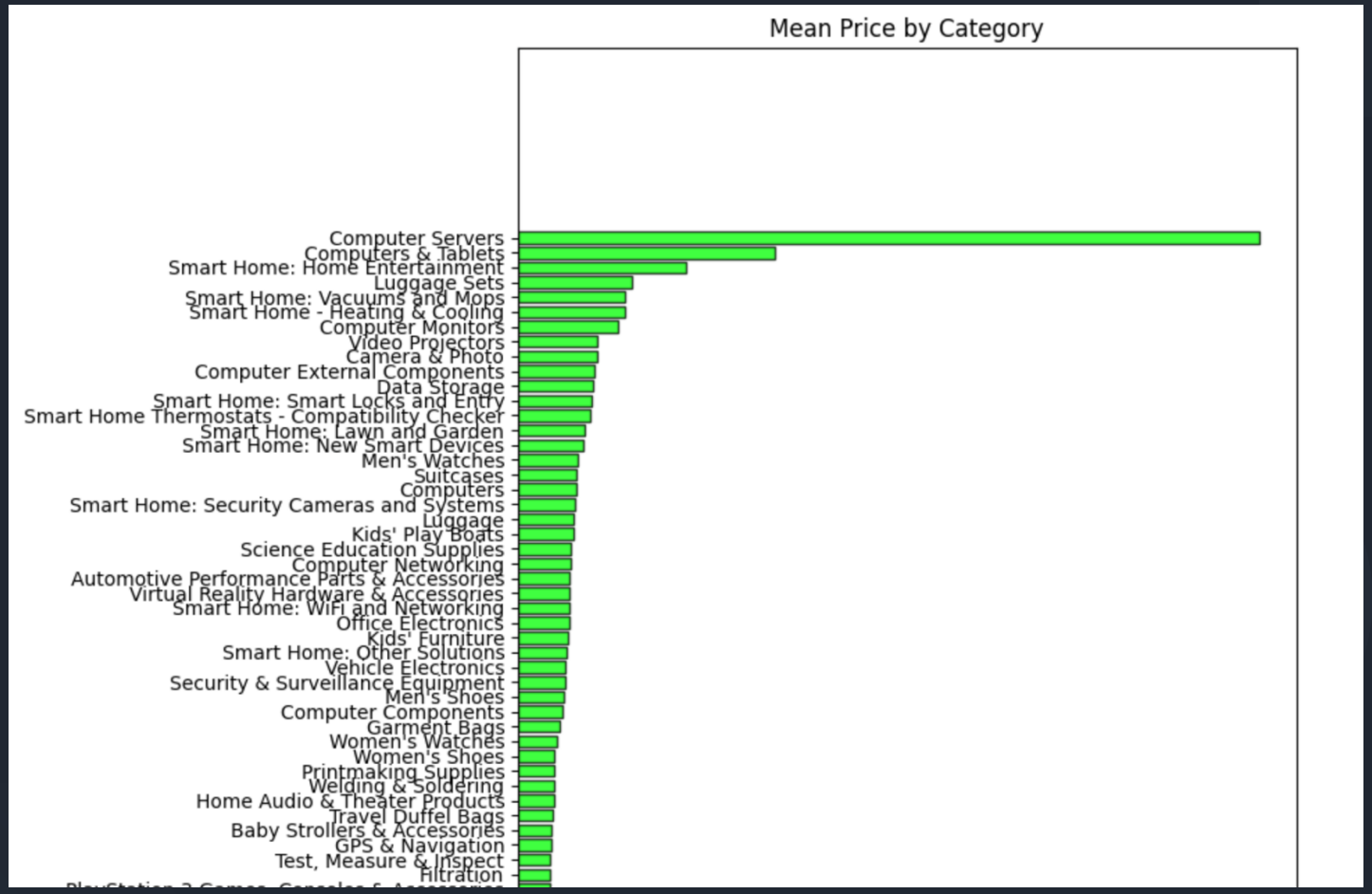
price_bucket	item_count
5-9	239252
10-14	228475
15-19	211163
20-24	122072
25-29	114358
35-39	65470
30-34	55580
0-4	52756
45-49	35342
40-44	27404
55-59	23493
65-69	17424
50-54	15794
75-79	12854
95-99	11193
60-64	10830



Visualization

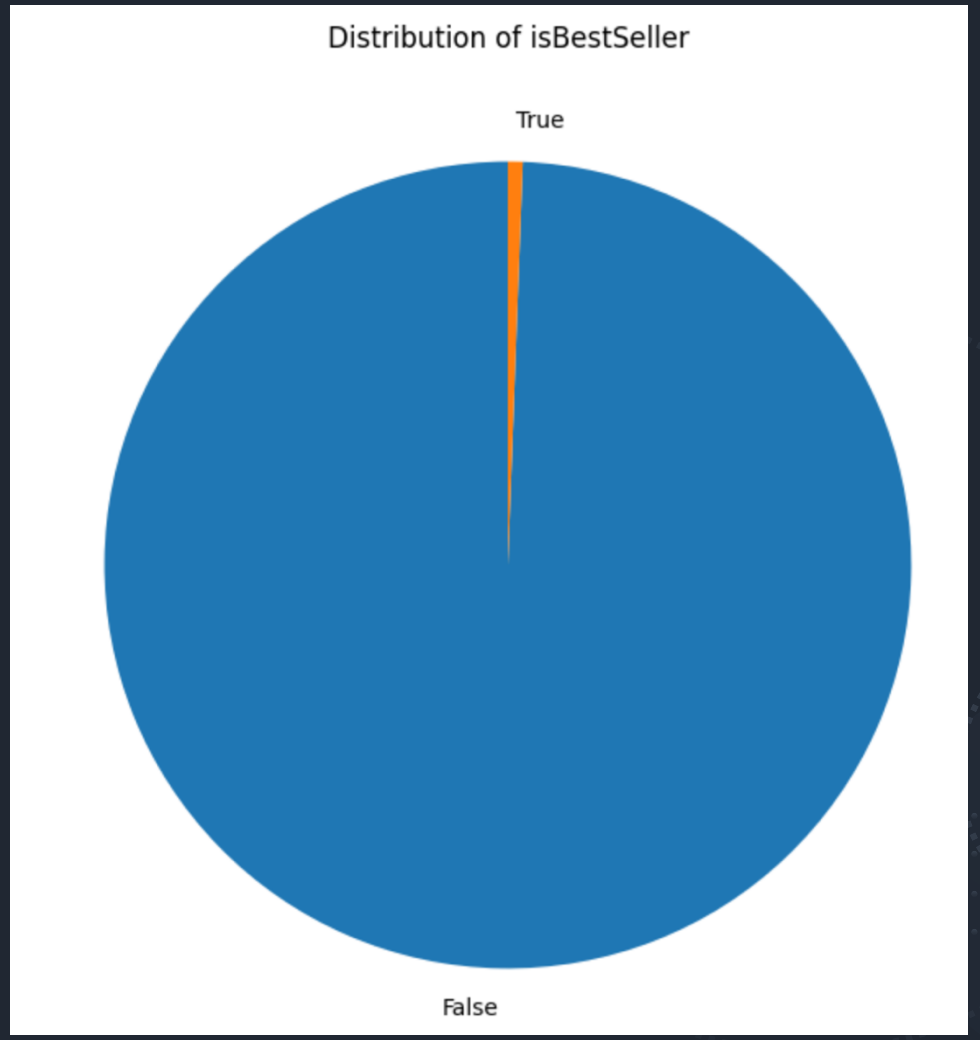
What is the average price for each category

category_name	price_mean
Computer Servers	1534.2246853146855
Computers & Tablets	531.2299453913818
Smart Home: Home ...	348.8907471264365
Luggage Sets	235.15786259542034
Smart Home: Vacuu...	221.1280769230769
Smart Home - Heat...	220.3855272727271
Computer Monitors	207.12657386177236
Video Projectors	164.31969934640628
Camera & Photo	163.74984151328914
Computer External...	158.75470279013294
Data Storage	156.01868504594512
Smart Home: Smart...	152.87220338983056
Smart Home Thermo...	149.33318181818183
Smart Home: Lawn ...	137.40266666666662
Smart Home: New S...	134.09595238095235
Men's Watches	123.31848170398202



Visualization

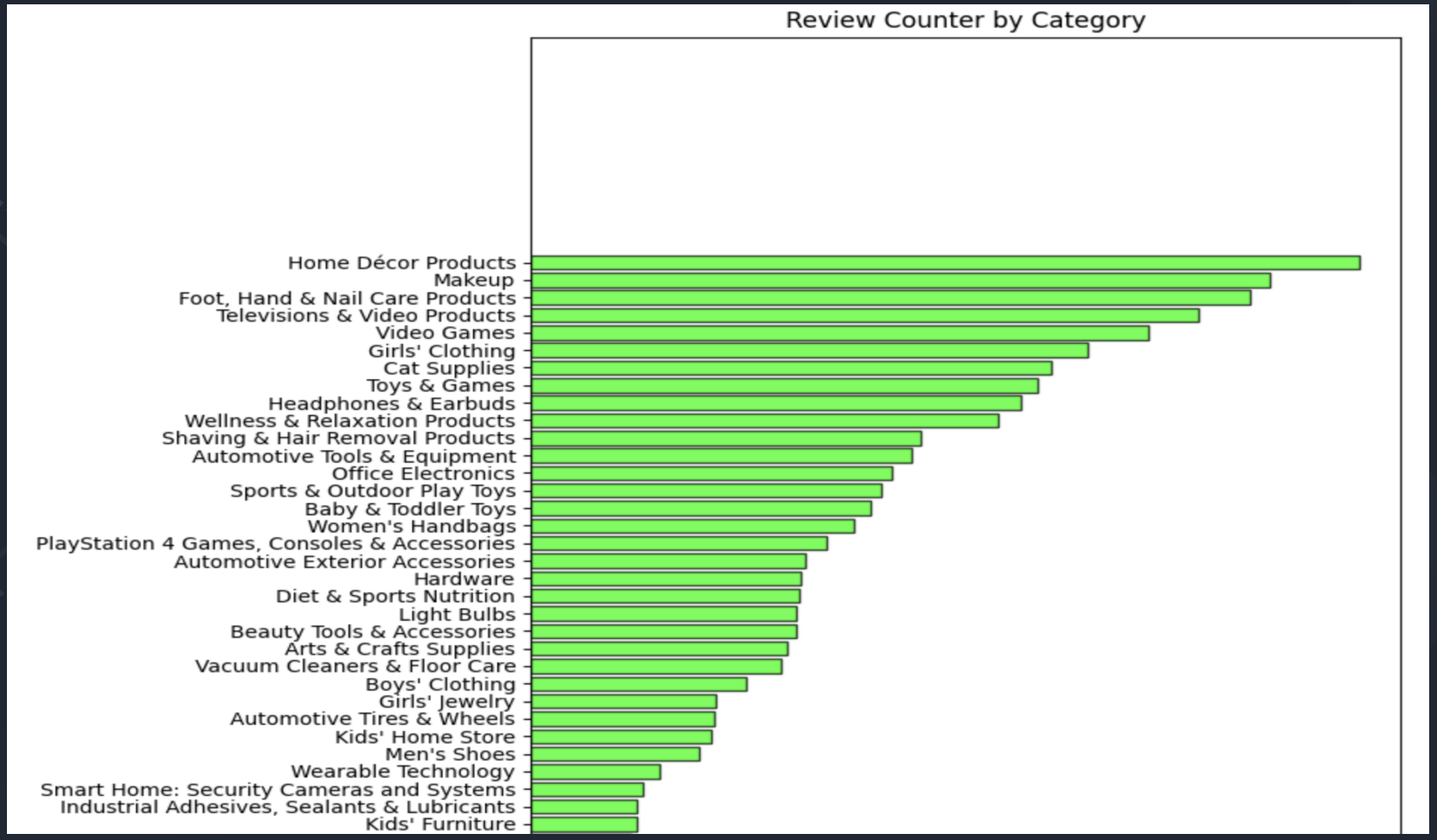
Distribution of Best Seller



Visualization

Which Category has most number of reviews

category_name	review_counter
Home Décor Products	16153120
Makeup	14427824
Foot, Hand & Nail...	14025612
Televisions & Vid...	13009624
Video Games	12044592
Girls' Clothing	10865362
Cat Supplies	10152884
Toys & Games	9880391
Headphones & Earbuds	9563572
Wellness & Relaxa...	9105818
Shaving & Hair Re...	7607317
Automotive Tools ...	7439808
Office Electronics	7038732
Sports & Outdoor ...	6851008
Baby & Toddler Toys	6632461
Women's Handbags	6317820



Model Application

We have applied a linear regression model for our data to predict the price of each category in our data and also evaluated the model using Root Mean Squared Error (RMSE).

id	features	price	prediction
2	[2.0]	12.99	49.66247732373869
2	[2.0]	49.99	49.66247732373869
2	[2.0]	20.99	49.66247732373869
2	[2.0]	17.1	49.66247732373869
2	[2.0]	11.83	49.66247732373869
2	[2.0]	62.39	49.66247732373869
2	[2.0]	17.09	49.66247732373869
2	[2.0]	8.55	49.66247732373869
2	[2.0]	6.99	49.66247732373869
2	[2.0]	6.99	49.66247732373869

only showing top 10 rows

Root Mean Squared Error (RMSE) on test data: 112.4903460755428

Conclusion

In this project, we used PySpark to preprocess and transform a dataset containing Amazon product sold listing information. Using linear regression, we built a predictive model to estimate the prices of each category based on features such as reviews, stars, and prices. The model was evaluated using Root Mean Squared Error (RMSE), providing insight into its accuracy in predicting listing prices.


Thank You
prime
prime
prime
prime
prime