Data Set: Amazon Product

Farzana Zannat
Nafiz Imtiaz
Mabell Martinez Garcia
Sylvia Progya
Ananya Shah

Business Problem: Optimizing Product Selection & Pricing for Market Entry on Amazon

→ The business problem is to determine the optimal pricing strategy for an entrepreneur looking to start a small-scale business on Amazon.

→ The importance lies in maximizing profit margins and ensuring competitiveness in the online marketplace.

Project Objectives

The objective is to make informed decisions to enhance sales and profitability.

MARKETPLACE	FACTORS	BUDGET
To enter the Amazon marketplace strategically by identifying and targeting the top 20 product categories with the highest sales volume.	Consider factors such as demand trends, seasonality, and customer preferences in the selection process	Estimate the initial budget required to enter and sustain operations in the chosen categories.

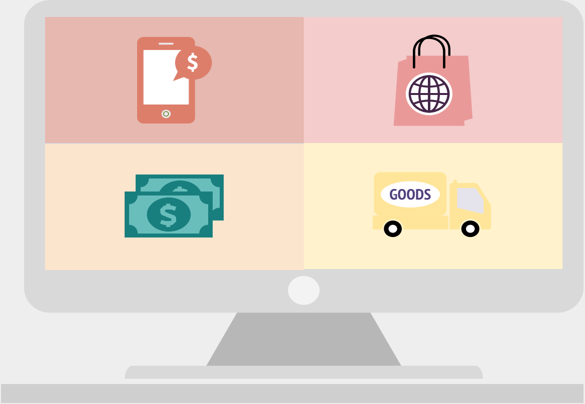
Business Implications

The implications of the project include a data-driven approach to pricing, potentially leading to increased sales and profitability. By understanding which products are likely to sell well and at what price point, the business can make informed decisions, allocate resources efficiently, and stay competitive in the dynamic e-commerce environment.

About Amazon

Massive Marketplace

A comprehensive online platform allowing users to effortlessly browse, search, and purchase products.
Where customers can browse, search, and buy a wide variety of products.



User-Centric Experience

Amazon's powerful search algorithm enhances user experience by delivering personalized product recommendations.

Seller Engagement & Operational Excellence

Sellers provide detailed product information, fostering transparency through descriptions and customer reviews.

Fulfillment centers play a pivotal role in seamless storage, packing, and shipping processes, ensuring efficient operations

Global Leadership

Amazon's commitment to operational efficiency and consumer satisfaction has propelled it to a leadership position in the global e-commerce landscape.

Goals & Business Impact

01

Optimize Pricing Strategy:
The project aims to enhance the pricing strategy on Amazon by identifying high-demand products.

02

Determine Optimal Selling Prices:
Utilizing market trends, the project seeks to establish optimal selling prices

03

Maximize Revenue and Customer Attraction:
Help the business attract customers, establish competitive prices, and maximize revenue in the fiercely competitive online marketplace

04

Analyze Price Distributions:
Analyzing price distributions across categories, the project aims to provide insights that contribute to establishing competitive prices.

MAIN GOAL

Describing the dataset

The data has been collected from

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

PAY

ONLINE

Amazon

The biggest online retailers in the USA that sells over 12 million products.

This dataset, we can get an in-depth idea of what products sell best, which SEO titles generate the most sales, the best price range for a product in a given category.

Contents of the Dataset

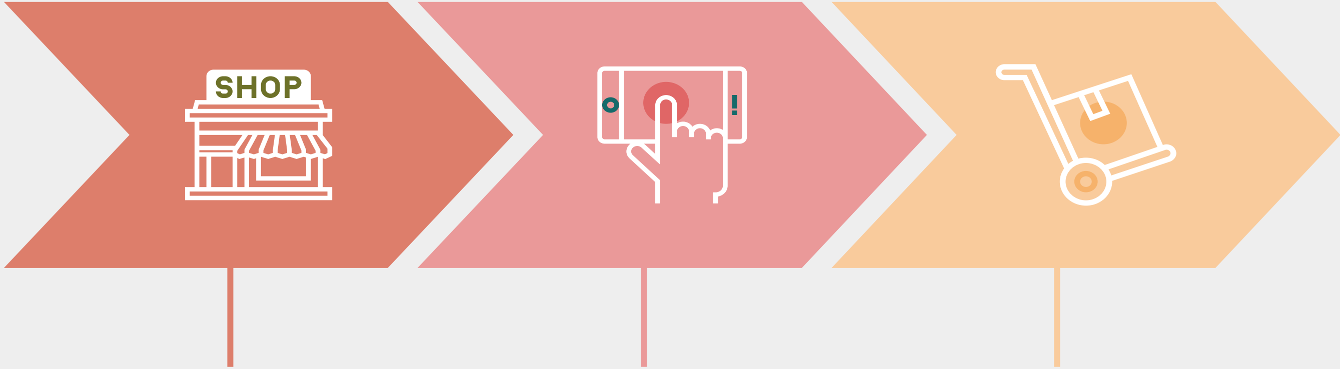
- ★ Product title
- ★ Product rating
- ★ Price of a product
- ★ List price

Different categories & which category product were sold in how many quantities last month.

****

Variables/Data types

Categorical and Numerical Data



ROWS

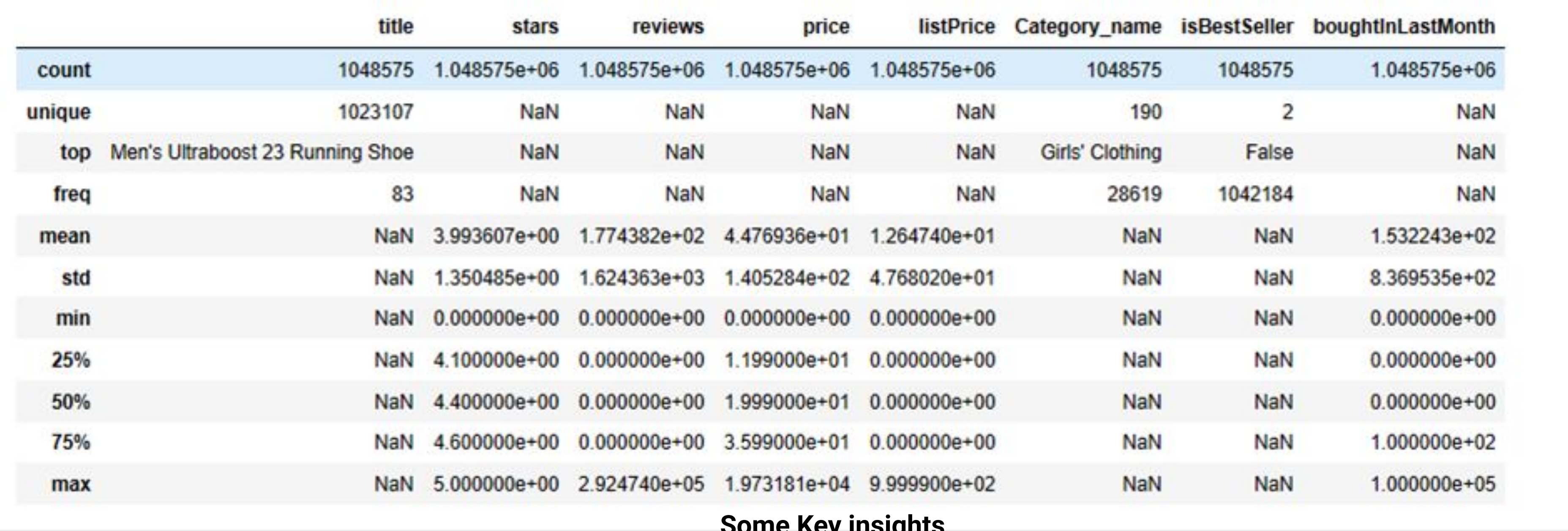
1,048,575

COLUMN

12

Target variable – Income (newly created column – Price * bought In LastMonth)

Descriptive analytics

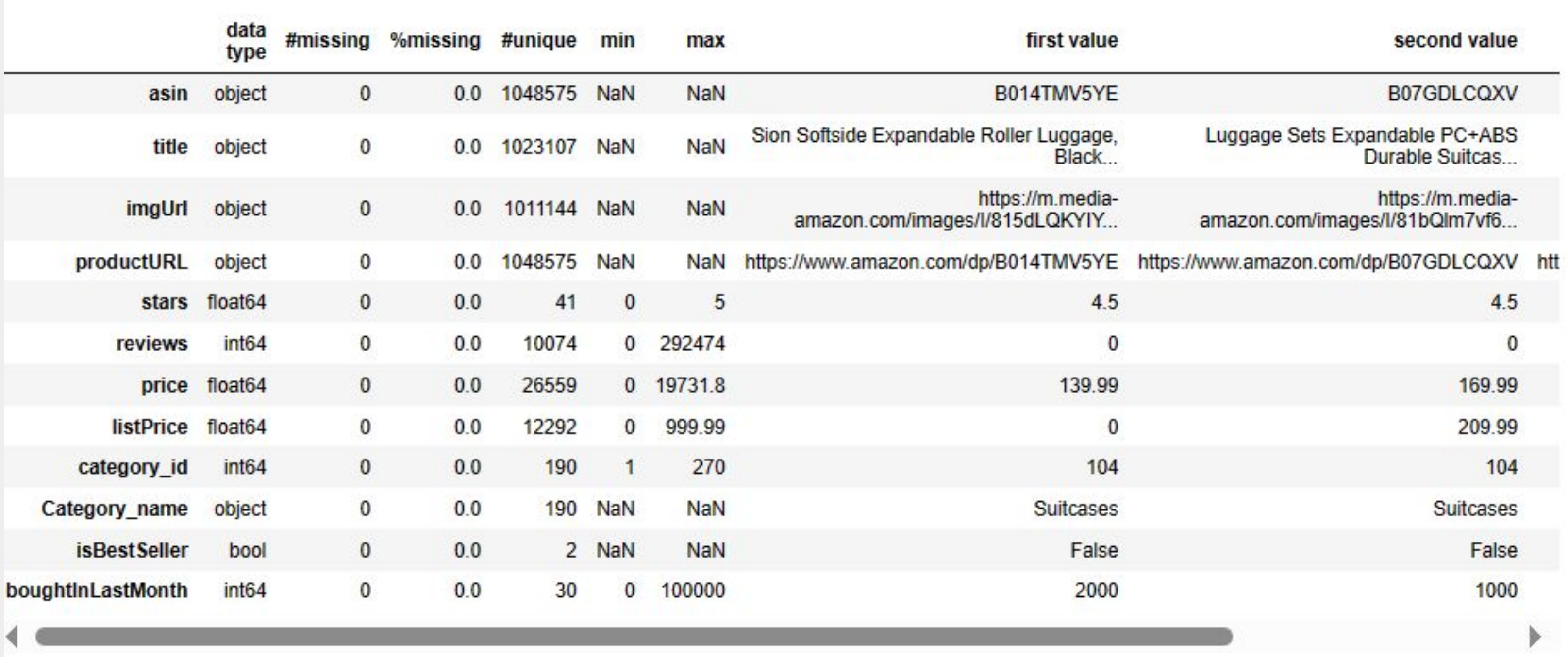


	title	stars	reviews	price	listPrice	Category_name	isBestSeller	boughtInLastMonth
count	1048575	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06	1048575	1048575	1.048575e+06
unique	1023107	NaN	NaN	NaN	NaN	190	2	NaN
top	Men's Ultraboost 23 Running Shoe	NaN	NaN	NaN	NaN	Girls' Clothing	False	NaN
freq	83	NaN	NaN	NaN	NaN	28619	1042184	NaN
mean	NaN	3.993607e+00	1.774382e+02	4.476936e+01	1.264740e+01	NaN	NaN	1.532243e+02
std	NaN	1.350485e+00	1.624363e+03	1.405284e+02	4.768020e+01	NaN	NaN	8.369535e+02
min	NaN	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	NaN	NaN	0.000000e+00
25%	NaN	4.100000e+00	0.000000e+00	1.199000e+01	0.000000e+00	NaN	NaN	0.000000e+00
50%	NaN	4.400000e+00	0.000000e+00	1.999000e+01	0.000000e+00	NaN	NaN	0.000000e+00
75%	NaN	4.600000e+00	0.000000e+00	3.599000e+01	0.000000e+00	NaN	NaN	1.000000e+02
max	NaN	5.000000e+00	2.924740e+05	1.973181e+04	9.999900e+02	NaN	NaN	1.000000e+05

Some Key insights

- ➔ The median number of reviews is 0, and the mean is 181. 75% of products have no reviews. It would be interesting to see how the reviews are distributed and how many products have reviews.
- ➔ The average product price is 43, the median is 20, and the interquartile range is 12-36. The mean does not fall within the interquartile range, which is due to the large spread of prices. The minimum price is 0, which is not possible, so these values need to be checked and possibly removed
- ➔ 75% of products have no discount, the average discount is 1.2, and the maximum is 999, which is a very large spread.
- ➔ The vast majority of products do not have bestseller status.
- ➔ 50% of products have 0 sales in the past month.

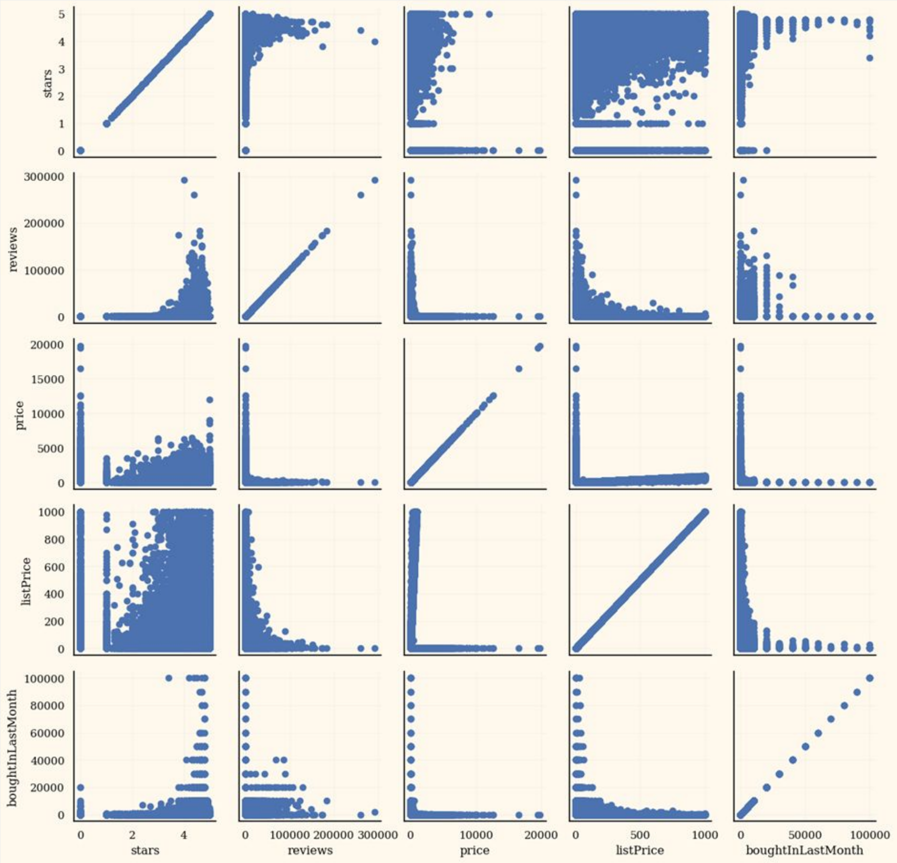
Descriptive analytics



	data type	#missing	%missing	#unique	min	max	first value	second value	
asin	object	0	0.0	1048575	NaN	NaN	B014TMV5YE	B07GDLCQXV	
title	object	0	0.0	1023107	NaN	NaN	Sion Softside Expandable Roller Luggage, Black...	Luggage Sets Expandable PC+ABS Durable Suitcas...	
imgUrl	object	0	0.0	1011144	NaN	NaN	https://m.media-amazon.com/images/I/815dLQKYIY...	https://m.media-amazon.com/images/I/81bQlm7vf6...	
productURL	object	0	0.0	1048575	NaN	NaN	https://www.amazon.com/dp/B014TMV5YE	https://www.amazon.com/dp/B07GDLCQXV	htt
stars	float64	0	0.0	41	0	5	4.5	4.5	
reviews	int64	0	0.0	10074	0	292474	0	0	
price	float64	0	0.0	26559	0	19731.8	139.99	169.99	
listPrice	float64	0	0.0	12292	0	999.99	0	209.99	
category_id	int64	0	0.0	190	1	270	104	104	
Category_name	object	0	0.0	190	NaN	NaN	Suitcases	Suitcases	
isBestSeller	bool	0	0.0	2	NaN	NaN	False	False	
boughtInLastMonth	int64	0	0.0	30	0	100000	2000	1000	

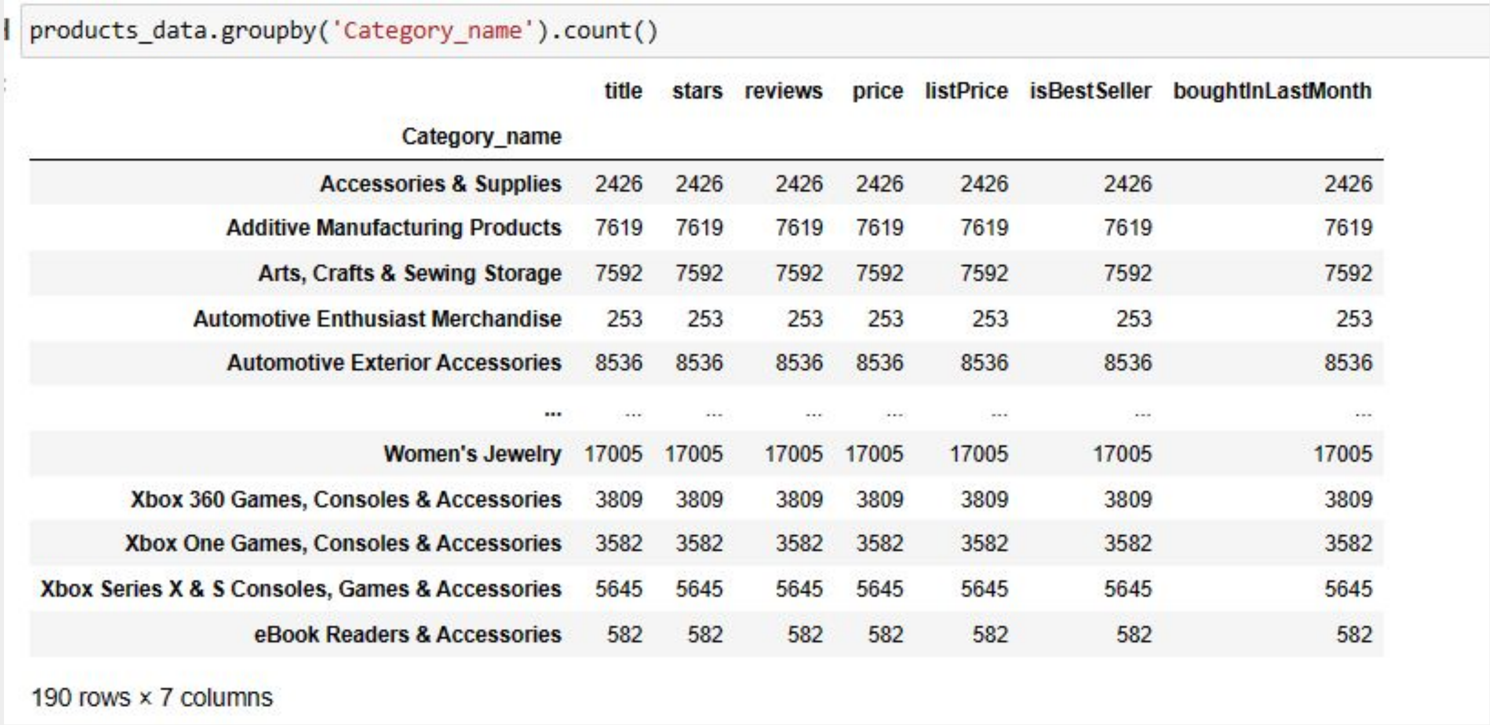
We've assessed the missing value percentage in each column, revealing a 0% rate. This signifies an absence of missing values across all columns in our dataset, indicating that every column holds complete data without any null or missing entries

Data Visualization – Correlation



- ➢ The higher the rating, the more reviews the product has.
- ➢ The higher the rating of a product, the more it was bought last month.
- ➢ The more reviews, the fewer goods were bought last month, which is strange, we need to look at reviews separately.
- ➢ Price and rating on the graph do not have a strong relationship. There is no rating at all for goods with a high price, probably because goods with a high cost are bought less often

Data Processing



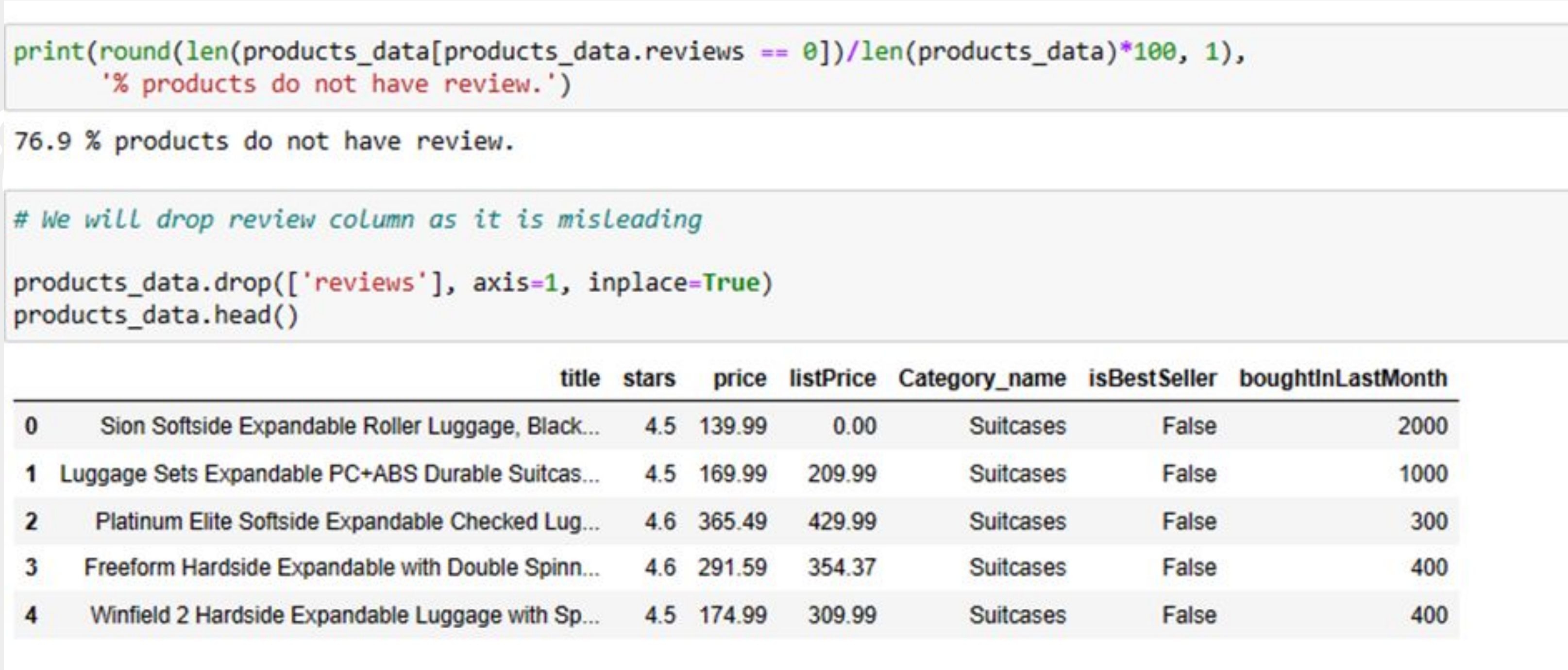
```
products_data.groupby('Category_name').count()
```

Category_name	title	stars	reviews	price	listPrice	isBestSeller	boughtInLastMonth
Accessories & Supplies	2426	2426	2426	2426	2426	2426	2426
Additive Manufacturing Products	7619	7619	7619	7619	7619	7619	7619
Arts, Crafts & Sewing Storage	7592	7592	7592	7592	7592	7592	7592
Automotive Enthusiast Merchandise	253	253	253	253	253	253	253
Automotive Exterior Accessories	8536	8536	8536	8536	8536	8536	8536
...	...	...	...	...	...	...	...
Women's Jewelry	17005	17005	17005	17005	17005	17005	17005
Xbox 360 Games, Consoles & Accessories	3809	3809	3809	3809	3809	3809	3809
Xbox One Games, Consoles & Accessories	3582	3582	3582	3582	3582	3582	3582
Xbox Series X & S Consoles, Games & Accessories	5645	5645	5645	5645	5645	5645	5645
eBook Readers & Accessories	582	582	582	582	582	582	582

190 rows × 7 columns

➢ Every one of the 190 product categories includes lines displaying a zero price. Removing these lines won't compromise the sample's representativeness.

Data Processing



```
print(round(len(products_data[products_data.reviews == 0])/len(products_data)*100, 1),
      '% products do not have review.')
```

```
76.9 % products do not have review.
```

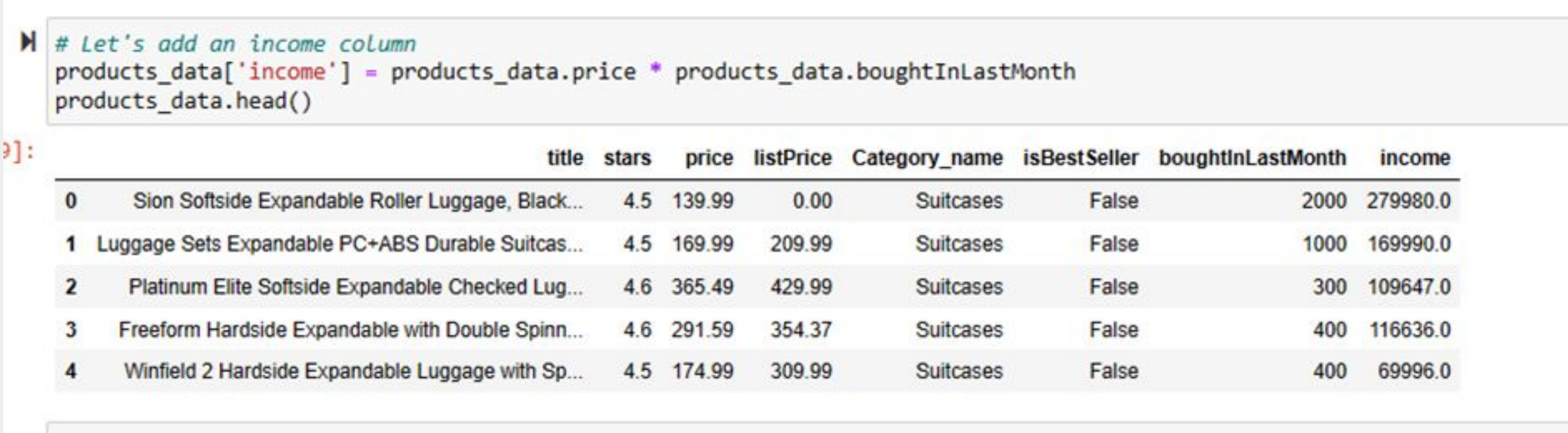
```
# We will drop review column as it is misleading

products_data.drop(['reviews'], axis=1, inplace=True)
products_data.head()
```

	title	stars	price	listPrice	Category_name	isBestSeller	boughtInLastMonth
0	Sion Softside Expandable Roller Luggage, Black...	4.5	139.99	0.00	Suitcases	False	2000
1	Luggage Sets Expandable PC+ABS Durable Suitcas...	4.5	169.99	209.99	Suitcases	False	1000
2	Platinum Elite Softside Expandable Checked Lug...	4.6	365.49	429.99	Suitcases	False	300
3	Freeform Hardside Expandable with Double Spinn...	4.6	291.59	354.37	Suitcases	False	400
4	Winfield 2 Hardside Expandable Luggage with Sp...	4.5	174.99	309.99	Suitcases	False	400

➢ Due to the absence of reviews in 76.9% of the products, we'll eliminate this column to avoid potential misinterpretation.

Data Processing

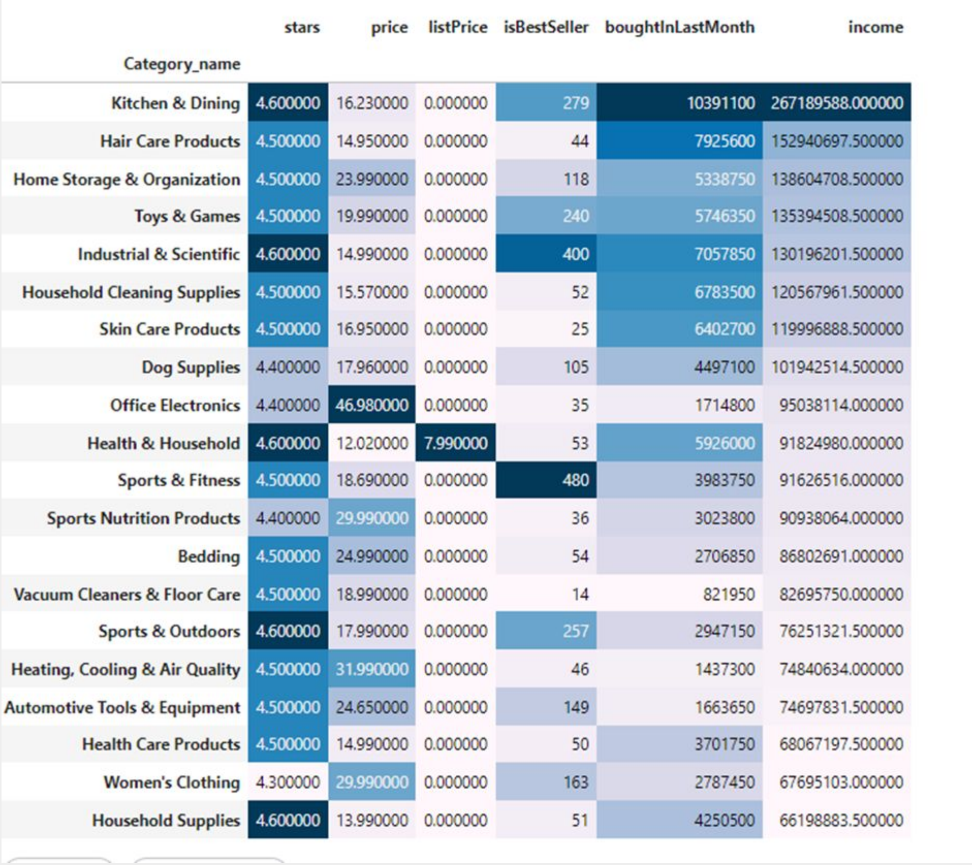


```
# Let's add an income column
products_data['income'] = products_data.price * products_data.boughtInLastMonth
products_data.head()
```

	title	stars	price	listPrice	Category_name	isBestSeller	boughtInLastMonth	income
0	Sion Softside Expandable Roller Luggage, Black...	4.5	139.99	0.00	Suitcases	False	2000	279980.0
1	Luggage Sets Expandable PC+ABS Durable Suitcas...	4.5	169.99	209.99	Suitcases	False	1000	169990.0
2	Platinum Elite Softside Expandable Checked Lug...	4.6	365.49	429.99	Suitcases	False	300	109647.0
3	Freeform Hardside Expandable with Double Spinn...	4.6	291.59	354.37	Suitcases	False	400	116636.0
4	Winfield 2 Hardside Expandable Luggage with Sp...	4.5	174.99	309.99	Suitcases	False	400	69996.0

- ➢ We have added an income column since this is going to be our target variable.

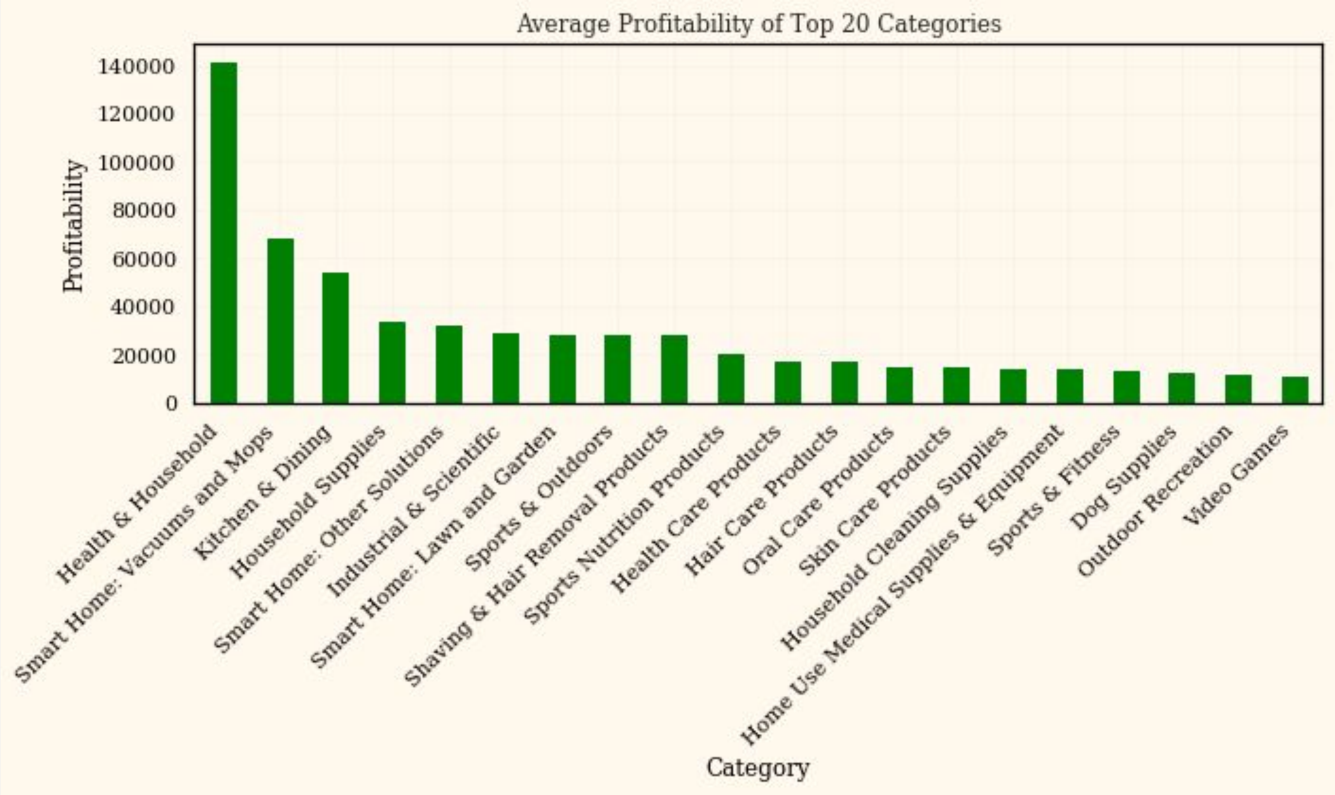
Top 20 most profitable Categories



	stars	price	listPrice	isBestSeller	boughtInLastMonth	income
Category_name						
Kitchen & Dining	4.600000	16.230000	0.000000	279	10391100	267189588.000000
Hair Care Products	4.500000	14.950000	0.000000	44	7925600	152940697.500000
Home Storage & Organization	4.500000	23.990000	0.000000	118	5338750	138604708.500000
Toys & Games	4.500000	19.990000	0.000000	240	5746350	135394508.500000
Industrial & Scientific	4.600000	14.990000	0.000000	400	7057850	130196201.500000
Household Cleaning Supplies	4.500000	15.570000	0.000000	52	6783500	120567961.500000
Skin Care Products	4.500000	16.950000	0.000000	25	6402700	119996888.500000
Dog Supplies	4.400000	17.960000	0.000000	105	4497100	101942514.500000
Office Electronics	4.400000	46.980000	0.000000	35	1714800	95038114.000000
Health & Household	4.600000	12.020000	7.990000	53	5926000	91824980.000000
Sports & Fitness	4.500000	18.690000	0.000000	480	3983750	91626516.000000
Sports Nutrition Products	4.400000	29.990000	0.000000	36	3023800	90938064.000000
Bedding	4.500000	24.990000	0.000000	54	2706850	86802691.000000
Vacuum Cleaners & Floor Care	4.500000	18.990000	0.000000	14	821950	82695750.000000
Sports & Outdoors	4.600000	17.990000	0.000000	257	2947150	76251321.500000
Heating, Cooling & Air Quality	4.500000	31.990000	0.000000	46	1437300	74840634.000000
Automotive Tools & Equipment	4.500000	24.650000	0.000000	149	1663650	74697831.500000
Health Care Products	4.500000	14.990000	0.000000	50	3701750	68067197.500000
Women's Clothing	4.300000	29.990000	0.000000	163	2787450	67695103.000000
Household Supplies	4.600000	13.990000	0.000000	51	4250500	66198883.500000

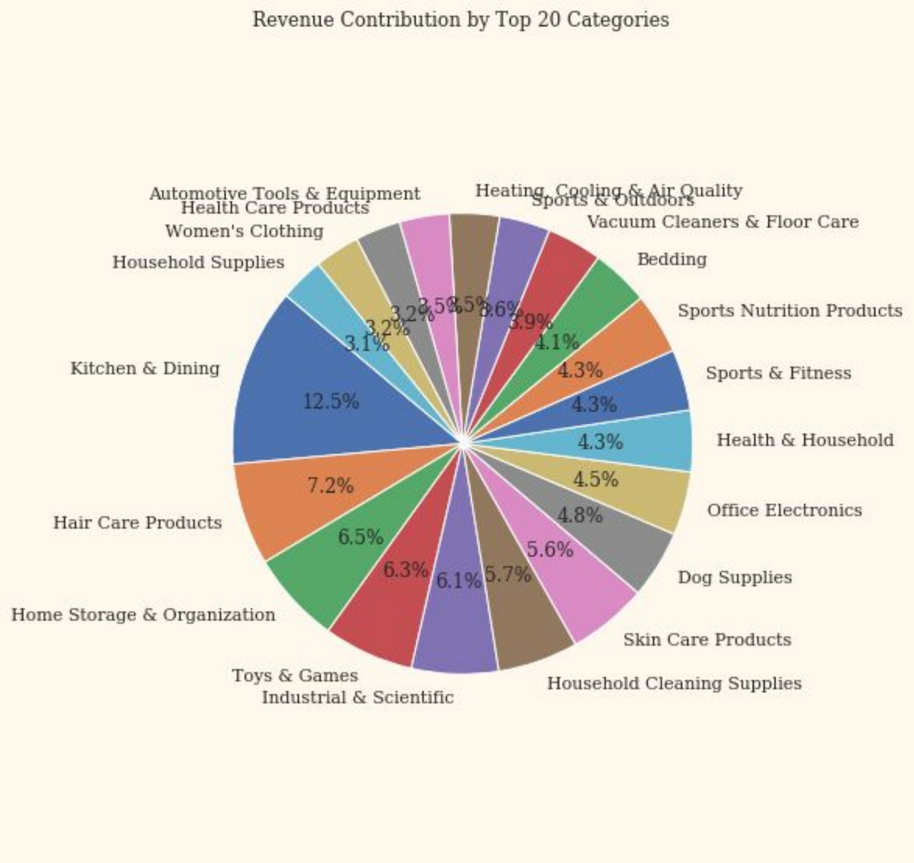
- ➢ The table provides summary characteristics for the 20 most popular categories.
- ➢ Kitchen & Dining is the leader among all categories: the highest revenue and the largest number of sold products, while the average price is relatively small, 16 dollars.
- ➢ The next category by revenue is hair care products, with an average price of also 15 dollars.

Average Profitability of Top 20 Categories



- ➢ Among the top 20, health & household categories demonstrate the highest profitability, while the video game categories show the least profitability

Revenue Contribution by Top 20 Categories



- ➢ Despite observing higher profitability in health & household products, the kitchen & dining category stands out by contributing a larger revenue share of 12.5%

Data Visualization - WordCloud

Word cloud is a powerful data visualization tool commonly used when working with texts, documents, surveys, and more. It effectively represents the frequency of each word appearing in a given text.

We have utilized wordcloud for data visualization, focusing on words used in product descriptions (extracted from the title column) with sales counts surpassing 50% in each category. This helps us identify the types of products that could be strategically listed on Amazon for an easier entry into the market.

Data Visualization - WordCloud

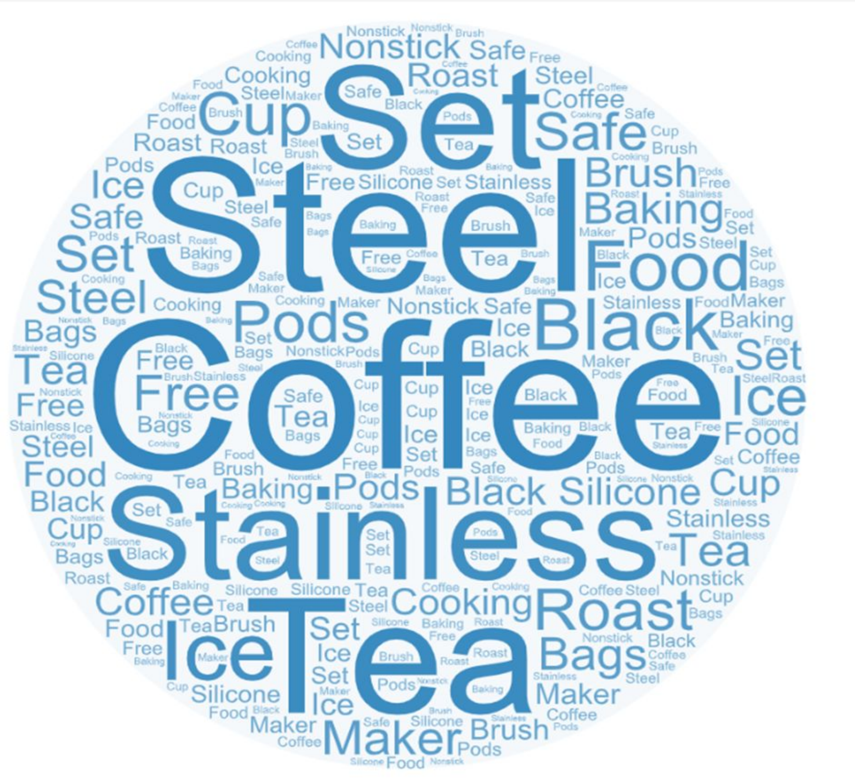
We've chosen the top 20 product categories as our focus, but due to time constraints, we'll only be showcasing three categories randomly for visualization purposes.

- Kitchen and Dining
- Hair Care Products
- Home Storage & Organization

We've conducted descriptive statistics (count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) for the 'boughtInLastMonth' column, representing our sales count.

Following that, we've identified the most frequently used word from the data associated with the median and above, subsequently narrowing it down to select the top 20 words.

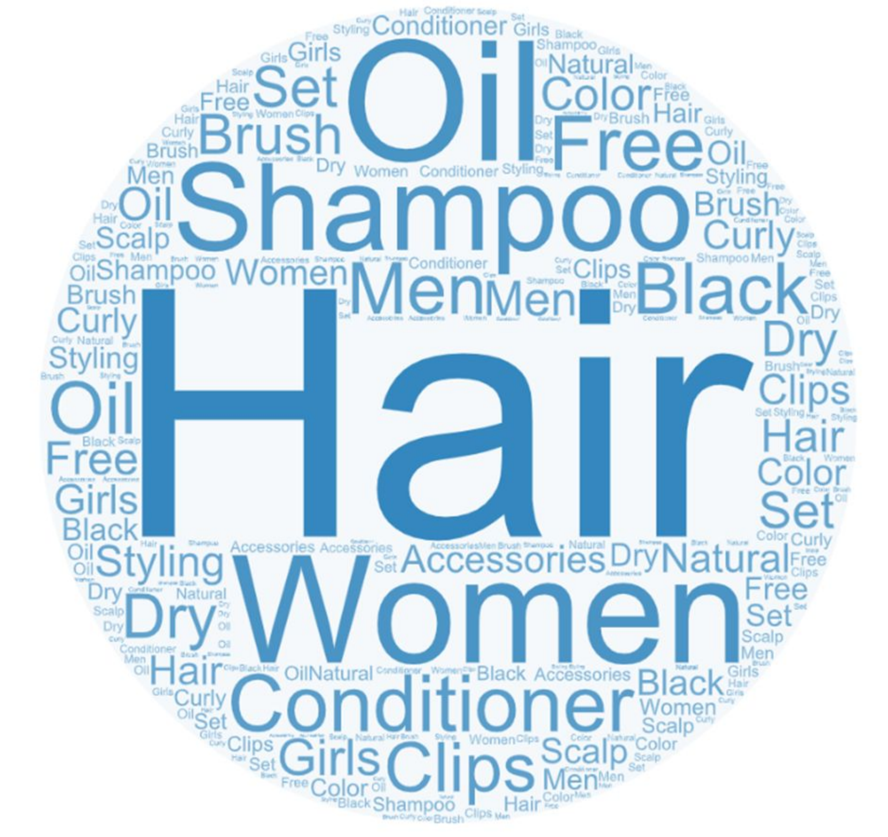
Data Visualization - WordCloud
Kitchen and Dining Category



➔ We've examined the most commonly used words in the description, specifically focusing on those with sales counts exceeding 50%.

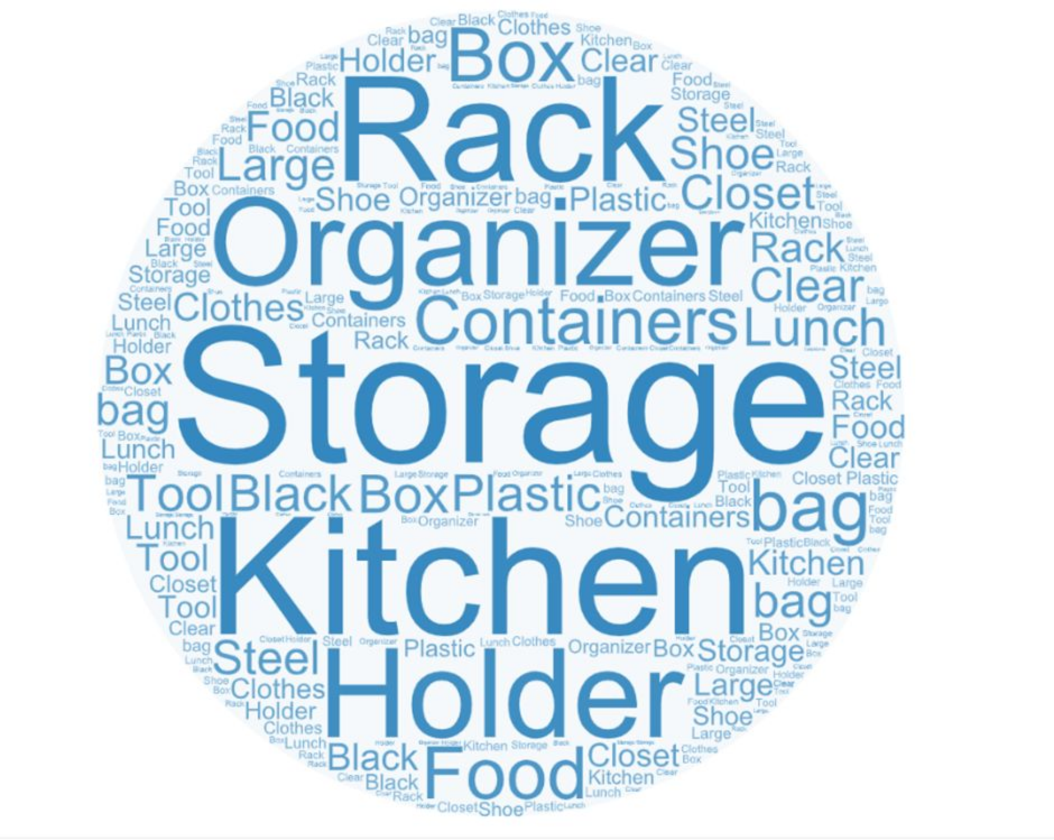
➔ This visualization suggests that items such as coffee, stainless steel, silicone, and tea could be among the top-selling products to start with in the Kitchen and Dining category on Amazon

Data Visualization - WordCloud
Hair Care Products



- ➔ We've analyzed the prevalent words within the descriptions, specifically emphasizing those associated with sales counts surpassing 50%.
- ➔ It seems from this visualization that products like shampoo, oil, women's products, conditioner, and clips might be some of the best-selling items to consider when starting in the Hair Care Products on Amazon.

Data Visualization - WordCloud
Home Storage & Organization



- ➔ Based on this visualization, products such as storage items, holders, organizers, containers, and kitchen organizers appear to be potential best-selling items when starting in the Storage & Organization category on Amazon.

Fitting models

- ➔ As we know fitting model is important to identify and capture patterns within data that might not be immediately apparent and we can predict and forecast of future trend using historical data.

- ➔ The Income target variable is computed using a specific formula (products_data.price * products_data.boughtInLastMonth), eliminating the need for model fitting. Even with new data, we can consistently calculate this target using the same formula. Given that the target variable is formulated, regression, classification or clustering approaches might not provide substantial value in this scenario

- ➔ If our goal were to predict future sales (referring to the column 'boughtInLastMonth'), we would typically split the dataset into training and test sets. This division allows us to evaluate different models to determine which one best fits the data and enables accurate predictions for future sales.

Summary

Our project focuses on optimizing product selection and pricing for entrepreneurs entering the Amazon marketplace, to maximize profit margins and competitiveness, while making informed decisions to enhance sales.

Utilizing word clouds for product descriptions in top categories revealed potential best-selling items, and this showed us the insights into market entry strategies for these categories on Amazon.

Furthermore, while the need for model fitting is reduced in the overall setting of the income target variable, which is exactly derived using a formula, the importance of forecasting becomes clear when anticipating future sales. The income target variable's quantitative character lets you perform consistent calculations regardless of new data, making regression, classification, or clustering procedures less useful in this specific circumstance. When the goal is to estimate future sales, however, a traditional approach including dataset partition into training and test sets remains relevant. This division simplifies the examination of several models, assisting in the selection of the best one for effectively predicting and forecasting future sales patterns.

Conclusion

- ➔ In conclusion, pursuing an ideal pricing plan for a small-scale firm on Amazon is more than an immediate need; it is a strategic essential for long-term success. The important emphasis on increasing profit margins and maintaining competitiveness emphasizes the critical role that pricing plays in the online marketplace landscape. Adopting a data-driven strategy represents a change in perspective, providing the entrepreneur with the means to navigate the difficulties of e-commerce. The project's consequences go beyond immediate profits, with the potential for a fundamental effect on sales and profitability. With insights about product demand and optimal pricing points, the company is well-positioned to make sound decisions, manage resources wisely, and actively react to the ever-changing trends in the online market.

- ➔ Was Your Goal Achieved ?
  Certainly we obtained our goal. Through extensive statistical analysis and visually appealing visualizations, we provided vital information. Furthermore, our model of prediction was accurate in anticipating pricing patterns as well as determining whether a host has several listings. The results demonstrate the effectiveness of our analytical approaches in extracting useful information from data.

Learnings

The project helped us improve communication skills by emphasizing the need of clear and effective communications when communicating ideas, discussing tasks, and addressing difficulties.

It encouraged our teammates to share ideas, capitalize on each other's talents, and work together to achieve a common objective.

It helped us divide the job into small tasks and allow enough time to complete it and examine by other team members.

New creative ideas were shared and implemented and that helped us achieve fruitful results.

We recognized and appreciated milestones and accomplishments. Acknowledging success helped in promoting morale of team members.

amazon
THANK YOU