Amazon Product: Unpacking Brand Significance

by

Aryan Sharma (aryanraj); section 101

Nan-Hsin Lin (nanhsin); section 001

Shreya Jain (shreyadj); section 101

1. Motivation

This project is driven by the overarching goal of unraveling the intricacies that influence Amazon product pricing and ratings and providing invaluable insights to enhance the decision-making processes of Amazon and its sellers. In the pursuit of this objective, we draw inspiration from the insightful analysis conducted by Ivan Isaev's Kaggle notebook (https://www.kaggle.com/code/ivanisaev/mazon-dataset-discounts-eda). While Isaev's analysis provides valuable insights into the dynamics of discounts and brand names, this project aims to extend the analysis by incorporating crucial dimensions related to the brand of the products. Specifically, we seek to explore the influence of brands' size, country, and other relevant factors on the pricing of Amazon products. Through a combination of descriptive, diagnostic (inferential), and predictive analytics, we aim to uncover nuanced patterns that transcend discount considerations, offering a deeper insight into the multifaceted factors driving sales in the competitive landscape of Amazon's e-commerce platform.

2. Data Sources

	Primary Dataset	Secondary Dataset
Name	Amazon Products 2023	Companies' Market Cap & Revenue
URL	https://www.kaggle.com/datasets/lokeshparab/amazon-products-dataset	https://companiesmarketcap.com/
Size	551585 records, 188.6 megabytes	7,956 records
Format	CSV	CSV
Access	Download with a Kaggle account	Download
Download Date	10/24/2023	10/24/2023
Key variables	• name: The name of the product • main_category: The main category of the product belong • sub_category: The sub-category of the product belong	• name: The name of the company • marketcap: The market cap of the company

	• ratings: The ratings given by Amazon customers of the product • no_of_ratings: The number of ratings given to this product in Amazon • discount_price: The discount prices of the product • actual_price: The actual MRP of the product	• revenue_ttm: The amount of revenue a company generates within the last twelve months • country: The headquarter of the company

3. Data Manipulation Methods

3.1 Market Cap & Revenue Dataset

(1) Merging

To analyze market cap and revenue, we merged the two datasets using an inner join on the 'Name' column, then removed the unnecessary columns: 'Rank_x', 'Rank_y', 'Symbol_x', 'Symbol_y', 'country_y', 'price (USD)_x' and 'price (USD)_y'.

(2) Extracting brand from parenthesis

Some of the company names consists of the parent company name with the subsidiary company name in the parenthesis, such as Alphabet (Google). Examining the Amazon Products dataset, we consider the subsidiary company name more aligned with the brand name, so we extracted the text within parentheses from the ‘brand’ column as the brand name. The regular expression '\((.*?)\)' is used to match any text within parentheses. The expand=False argument means that if the regular expression finds more than one match, it will return a series of matches instead of a data frame. If there is no match (i.e., no text within parentheses), the fillna(df['brand']) function will fill the missing values with the original 'brand' value. The str.lower() and str.capitalize() functions are used to convert the text in the 'brand' column to lowercase and capitalize the first letter of each word, respectively. This standardizes the text data.

(3) Filtering

The merged dataset was filtered to include only records where the Market Cap and Revenue values are not null. This was done to ensure that only valid data is used for analysis.

3.2 Amazon Products Dataset

(1) Changing units of discount_price and actual_price columns

Code performs the following:

- It removes the ₹ sign from the discount_price and actual_price columns.
- It splits the string in these columns by space and ₹ sign and keeps only the part after the ₹ sign.
- It replaces commas with nothing (effectively removing them) in these columns.
- It converts the data type of these columns to float.
- It converts the currency to USD.

(2) Filtering Ratings

- Replaced 'Get', 'FREE','₹68.99', '₹65','₹70', '₹100', '₹99', '₹2.99' values in the 'ratings' column of the 'products' data frame which don't make sense in context with '0.0', converted the 'ratings' column to float type, and then printed the unique values in the 'ratings' column.
- Filtered the products data frame to only include rows where the first character of the 'no_of_ratings' column is a digit. Then, removed any commas from the 'no_of_ratings' column and converted the column to float type.

(3) Adding Columns

- Created a new column 'discount_value' which is the difference between 'actual_price' and 'discount_price'.
- Subsequently created a 'discounting_percent' column which is the discount percentage. The discount percentage is calculated as the discount value divided by the actual price, multiplied by 100 to get a percentage.

(4) Extracting brand names from product names

Steps performed to extract brand name:

1. Splitted the 'name' column of the 'products' data frame on spaces and took the first word as the 'brand'.
2. Converted the 'brand' column to lowercase and then capitalized the first letter.

(5) Merging

Merged the 'products' data frame with Market Cap & Revenue dataset 'df' on the 'brand' column. The 'inner' method is used, which means only the rows with matching 'brand' in both data frames will be kept.

4. Analysis

4.1 Descriptive Analysis

(1) Interpretation of descriptive statistics for each of the columns:

	ratings	no_of_ratings	discount_price	actual_price	\
count	18390.000000	18390.000000	1.839000e+04	1.839000e+04	
mean	3.947227	2135.483904	6.537443e+03	9.935642e+03	
std	0.649333	17377.744145	1.864749e+04	2.567719e+04	
min	1.000000	1.000000	1.000000e+01	3.800000e+01	
25%	3.700000	8.000000	7.390000e+02	1.799000e+03	
50%	4.000000	54.000000	1.859000e+03	3.495000e+03	
75%	4.300000	384.000000	4.082750e+03	6.900000e+03	
max	5.000000	437652.000000	1.249990e+06	1.594900e+06	

	discount_value	discounting_percent	marketcap	revenue_ttm
count	18390.000000	18390.000000	1.839000e+04	1.839000e+04
mean	3398.198813	0.429599	2.881634e+11	1.218056e+11
std	8071.438677	0.198819	5.274921e+11	2.038397e+11
min	0.050000	0.000017	1.136061e+07	0.000000e+00
25%	749.000000	0.288972	8.587731e+09	7.771596e+09
50%	1461.000000	0.440854	1.629057e+10	1.185757e+10
75%	2649.750000	0.570285	1.600745e+11	6.363884e+10
max	344910.000000	0.999000	2.711596e+12	5.380460e+11

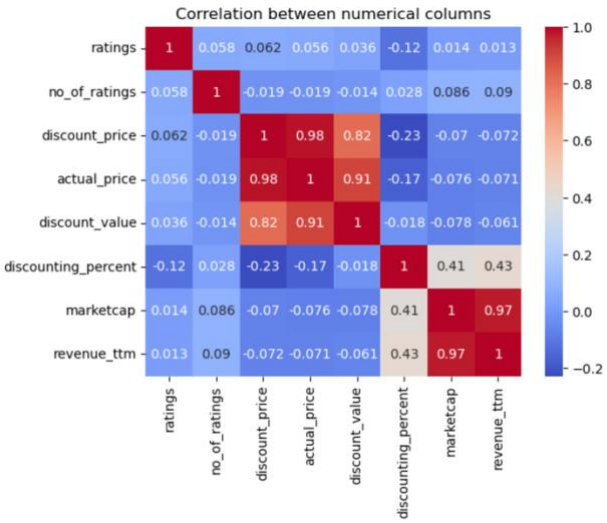
- ratings: The average rating of a product in your dataset is around 3.95 (on a scale up to 5), with a standard deviation of about 0.65. The minimum rating is 1, and the maximum rating is 5. This indicates a fairly high average rating for products in the dataset.
- no_of_ratings: The average number of ratings per product is approximately 2135.48, but the standard deviation is quite large (17377.74), indicating a wide range of values. The maximum number of ratings a product has received is 437,652 which is quite high.
- discount_price: The average discounted price of a product is about $6537.44, with a large standard deviation of $18647.49, indicating a wide range of discounted prices. The maximum discounted price is $1,249,990.
- actual_price: The average actual price is about $9935.64. The standard deviation is quite large $25677.19 suggesting a wide range of actual prices. The maximum actual price is $1,594,900.
- discount_value: The average discount value is about $3398.20, with a large standard deviation of $8071.44. The maximum discount offered is $344,910.
- discounting_percent: The average discount percentage is about 42.96%, with a standard deviation of 19.88%. This suggests that the discounts vary quite a bit, from a minimum near 0% to nearly 100%.
- marketcap: The average market cap is approximately $288.16 billion, with a high standard deviation, indicating broad variations in company sizes. The maximum market cap in the dataset is $2.71 trillion.
- revenue_ttm: The average trailing twelve months (TTM) revenue is about $121.81 billion, with a large standard deviation, indicating a wide range of company revenues. The maximum TTM revenue in the dataset is $538.046 billion.

Note that mean is sensitive to extreme values (outliers), and given the high standard deviation in the dataset, median is a more accurate representation of the central tendency for some of the variables.

(2) Correlation between numeric variables:

Some notable correlations:

- discount_price and actual_price: There's a strong positive correlation (0.98) between these variables, suggesting that as the actual price of a product increases, so does its discounted price.
- actual_price and discount_value: There's also a strong positive correlation (0.91) here, indicating that the higher the actual price of a product, the higher the discount value tends to be.
- marketcap and revenue_ttm: These variables have a very strong positive correlation (0.97),

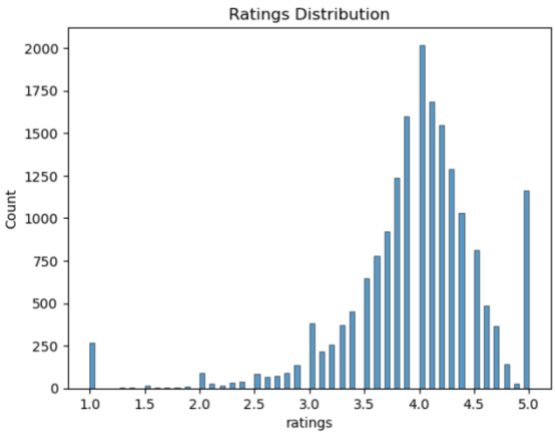


indicating that companies with larger market capitalizations generally have higher revenues over the trailing twelve months.

- discounting_percent and marketcap: There's a moderate positive correlation (0.41), suggesting that higher discounts tend to be associated with larger companies (in terms of market cap).
- discounting_percent and revenue_ttm: There's a moderate positive correlation (0.43), indicating that higher discounts are typically seen with companies that have higher revenues.

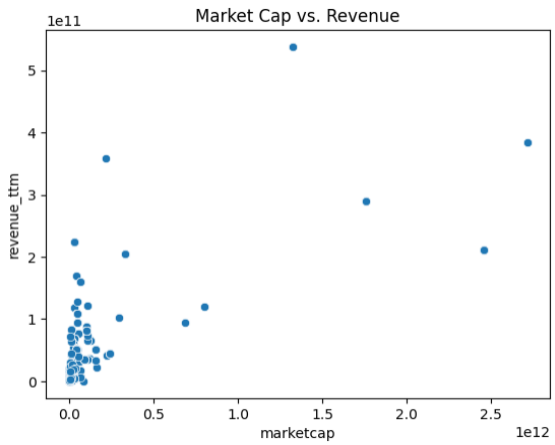
Note that correlation does not imply causation. While these relationships exist in the dataset, they don't necessarily mean one variable's change directly causes the change in another variable. The correlations might be due to underlying factors not included in the dataset.

(3) Distribution of Ratings



- The histogram of ratings is left-skewed (also known as negatively skewed), which means that: Most of the products have high ratings. The majority of the ratings are clustered towards the higher end of the scale.
- The tail of the distribution on the histogram extends towards the left, i.e., towards the lower ratings.
- The mean (average) rating is likely less than the median rating (the midpoint when ratings are sorted in ascending order). This is a characteristic of left-skewed distributions.
- In the context of product ratings, a left-skew could suggest customer satisfaction overall, as more products have higher ratings. However, the skew also suggests that there are a few products with low ratings that could potentially skew the average rating downwards.

(4) The relationship between market cap and revenue

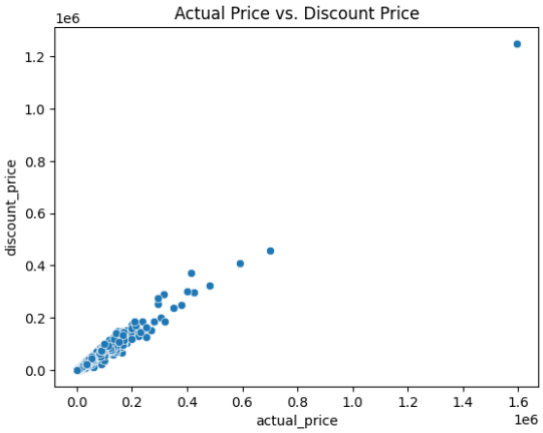


- The concentration of points at (0.1, 0), (0.1, 1) and (0.1, 2): This could suggest that there are many companies in the dataset with a market cap near zero but with revenues ranging from $0.1 billion to $2 billion. These could be small companies or startups that have yet to achieve significant market capitalization but are generating some revenue.
- The random scatter of other points: The random scattering of the remaining data points could suggest that there's not a clear linear relationship between market cap and revenue for the rest of the companies

in the dataset. In other words, for these companies, changes in market cap don't necessarily correspond to predictable changes in revenue.

(5) The relationship between actual price and discounted price

- The scatter plot of actual_price vs. discount_price shows an upward trend (slope), this suggests a positive correlation between these two variables. This is expected, as typically, the higher the product's actual price, the higher the discounted price will be, even after the discount is applied.



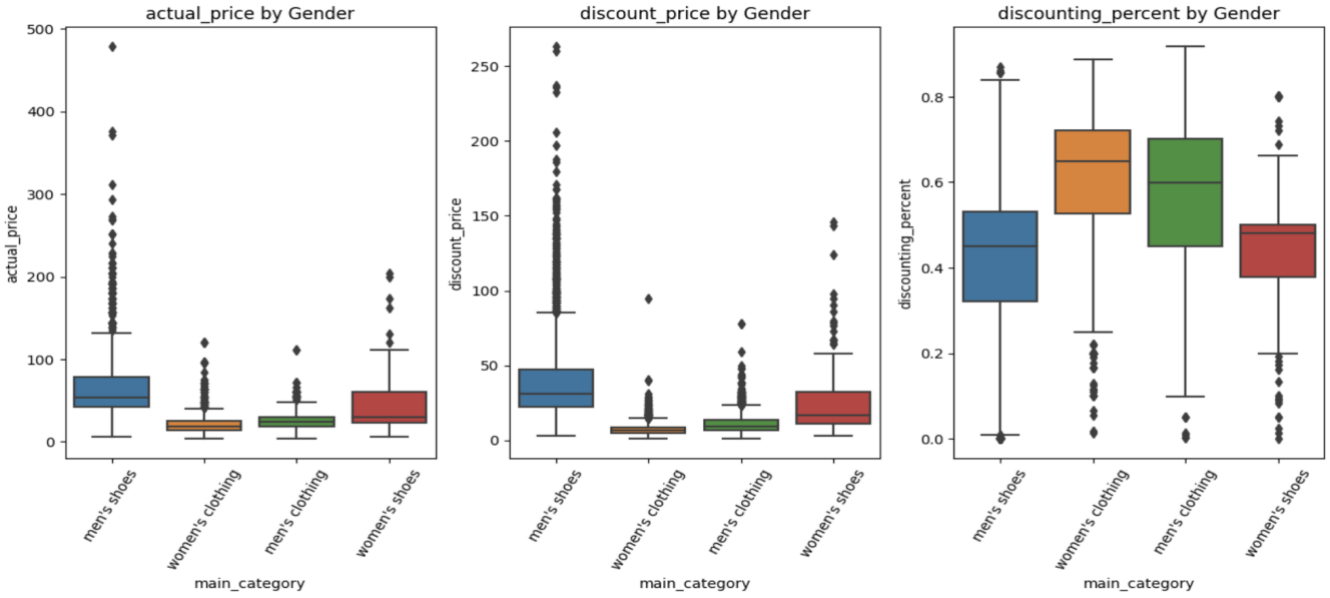
4.2 Diagnostic Analysis (Inferential Statistics)

We conducted diagnostic analysis based on our descriptive analysis on the following three interesting questions.

(1) Do pricing strategies differ for products for different genders?

As we found that products in clothing and shoes are separated into two categories by their target gender, we would like to explore if there are any differences between the pricing strategies of products for different genders.

First, from the boxplots and histograms below, we found that both clothing and shoes for men are higher in the original price (actual_price) and the price after discount (discount_price) than those for women, and there is no significant difference for the discount percentage.



To further investigate whether the difference between genders is significant, we conducted hypothesis testing on actual_price and main_category using Ordinary Least Squares (OLS) linear regression models for clothing and shoes. The p-value of the model for clothing is 1.26e-13, while that for shoes is 3.22e-20. Both p-values are way below the critical value 0.05, so the regressions are statistically significant. We can reject the null hypothesis and state that there is a difference between products (clothing and shoes) for different genders. The negative coefficients for women's clothing (-3.2096) and women's shoes (-20.5974) indicate that products for women tend to have lower prices, which aligns with our hypothesis and interpretation from the previous visualization.

OLS Regression Results

Dep. Variable:	actual_price	**R-squared:**	0.016
Model:	OLS	**Adj. R-squared:**	0.016
Method:	Least Squares	**F-statistic:**	55.38
Date:	Wed, 06 Dec 2023	**Prob (F-statistic):**	1.26e-13
Time:	20:21:50	**Log-Likelihood:**	-12883.
No. Observations:	3308	**AIC:**	2.577e+04
Df Residuals:	3306	**BIC:**	2.578e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	25.4044	0.258	98.453	0.000	24.898	25.910
C(main_category)[T.women's clothing]	-3.2096	0.431	-7.442	0.000	-4.055	-2.364

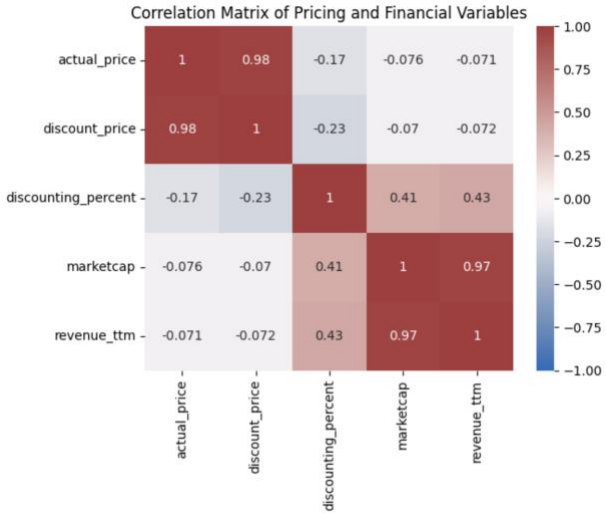
OLS Regression Results

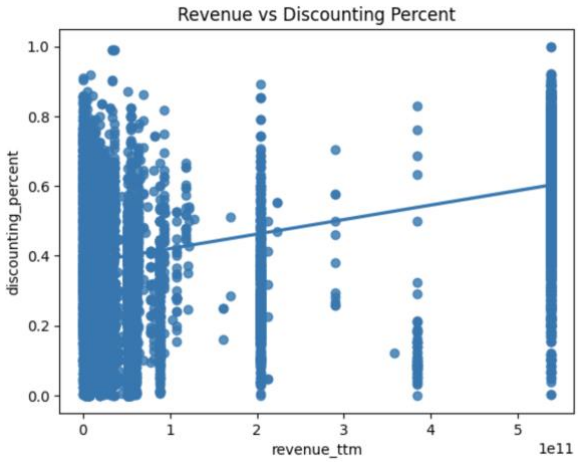
Dep. Variable:	actual_price	**R-squared:**	0.018
Model:	OLS	**Adj. R-squared:**	0.018
Method:	Least Squares	**F-statistic:**	85.64
Date:	Wed, 06 Dec 2023	**Prob (F-statistic):**	3.22e-20
Time:	20:21:50	**Log-Likelihood:**	-22926.
No. Observations:	4611	**AIC:**	4.586e+04
Df Residuals:	4609	**BIC:**	4.587e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	63.5383	0.530	119.990	0.000	62.500	64.576
C(main_category)[T.women's shoes]	-20.5974	2.226	-9.254	0.000	-24.961	-16.234

(2) Is there a relationship between the size of a brand and its pricing?

From the heatmap in our descriptive analysis, there is a slight positive correlation between discounting_percent and revenue_ttm (0.43), and between discounting_percent and market_cap (0.41). As revenue_ttm and market_cap are highly correlated (0.97), we will use only revenue_ttm for our further investigation for the relationship with discounting_percent. We then plot the data with a linear regression model fit. While there is not a discernible pattern, the slope of the regression line is slightly positive.





The OLS linear regression model with "discounting_percent" as the dependent variable and "revenue_ttm" as the predictor variable has a p-value of 0.00, an R-square of 0.182, and a coefficient of 4.156e-13. With p-value below the critical value 0.05, we can reject the null hypothesis and conclude that there is a statistically significant positive relationship between discounting_percent and revenue_ttm. However, we should also keep in mind that the low R-squared value indicates that the model does not explain much of the variability in the dependent variable discounting_percent.

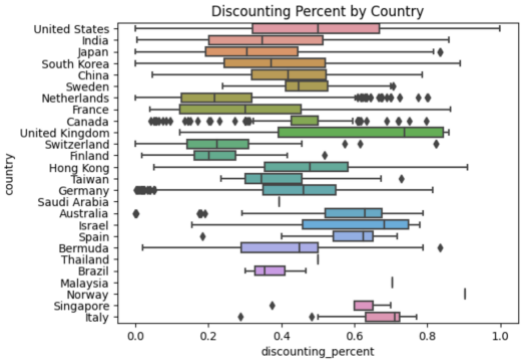
OLS Regression Results

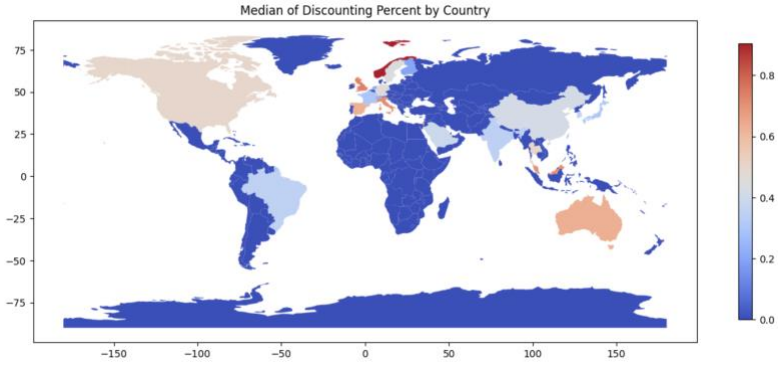
Dep. Variable:	discounting_percent	**R-squared:**	0.182
Model:	OLS	**Adj. R-squared:**	0.182
Method:	Least Squares	**F-statistic:**	4079.
Date:	Wed, 06 Dec 2023	**Prob (F-statistic):**	0.00
Time:	20:21:51	**Log-Likelihood:**	5454.8
No. Observations:	18390	**AIC:**	-1.091e+04
Df Residuals:	18388	**BIC:**	-1.089e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.3790	0.002	245.264	0.000	0.376	0.382
revenue_ttm	4.156e-13	6.51e-15	63.865	0.000	4.03e-13	4.28e-13

(3) Are there differences for the discount strategy among countries?

The box plot and the choropleth map below demonstrate the differences in “discounting_percent” in each country. From the box plot, we found that there are only two products for Malaysia and one product for Norway, so we will remove these products in further analysis to avoid data bias.





After the removal, we conducted an OLS linear regression model with "discounting_percent" as the dependent variable and "country" as the predictor variable. With a p-value of 0.00 which is below the critical value 0.05, we can reject the null hypothesis and conclude that there is a statistically significant

difference between “discounting_percent” and “country”. In addition, both the boxplot and the coefficient of the OLS model show that Finland and Switzerland tend to have much smaller “discounting_percent”.

We also conducted an OLS linear regression model with "actual_price" as the dependent variable and "country" as the predictor variable. With a p-value of 0.00 which is below the critical value 0.05, we can reject the null hypothesis and conclude that there is a statistically significant difference between actual_price and country. In addition, the coefficients of Sweden (457.9725) and Taiwan (341.0491) indicate that brands of these two countries tend to have higher prices for their products.

4.3 Predictive Analysis

We predict ratings based on different features such as sub-category of the product, brand, number of ratings for each product, discount value and the actual price. In this analysis we chose not to use Revenue and Market Capital of the company. This decision is grounded in the understanding that a company may sell a diverse range of products, and assuming the revenue of the entire company for predicting ratings on a single product may not be a valid approach.

Predicting Rating:

We consider features such as sub-category, number of ratings, actual price, discount value, brand. Target Encoder has been used to encode the categorical columns. Target Encoder seems to be a valid choice because of the high cardinality in the dataset. Additionally, we scaled numerical columns using Standard Scaler. The obtained metrics, in table (1) suggest that the considered features alone may not be sufficient to predict the ratings.

We extended our analysis to include the 'name' column which comprises product descriptions. Notably, for products under subcategories like casual shoes, sports shoes, or Men’s Fashion, the last word in the 'name' column often denotes the 'shoe' type. Focusing specifically on the 'footwear' category, comprising 4361 rows, after extracting the last word we incorporated this additional feature along with others to predict ratings. Surprisingly, the model’s performance was decreased as shown in the table (2).

Table (1): Features: sub_category, no_of_ratings, actual_price, discount_value, brand

Model	Mean-Squared Error	R-Squared
Linear Regression	0.38	0.1
SVR	0.39	0.09
Random Forest	0.35	0.179
XGBoost	0.36	0.15

Table (2): For shoe categories

Features: sub_category, no_of_ratings, actual_price, discount_value, brand, type of shoes

Model	Mean-Squared Error	R-Squared
Linear Regression	0.39	0.06
SVR	0.39	0.06
Random Forest	0.4	0.01
XGBoost	0.43	-0.06

This might be due to the following reasons:

- Less amount of data (4361 rows) makes it difficult to establish robust patterns.
- Consideration of only the shoes category does not imply that the model will underperform / well perform for other categories.
- We do not consider other sub-categories because of the less amount of data.
- For ratings, external factors such as product quality, product color, shipping duration, durability, etc. might be significantly important.

5. Statement of Work

(1) Key Responsibilities:

- Aryan Sharma
  - Conducted exploratory data analysis to understand the dataset.
  - Identified patterns, trends, and anomalies in the data.
  - Generated visualizations to aid in data interpretation.
  - Presented clear and concise summaries of data characteristics.
- Nan-Hsin Lin
  - Conducted descriptive and diagnostic analysis on the following problems:
    1. Do pricing strategies differ for products for different genders?
    2. Is there a relationship between the size of a brand and its pricing?
    3. Are there differences for the discount strategy among countries?
  - Organized overall analysis outline in a logical and visually appealing manner.
  - Formatted the report as per professional standards.
  - Managed project status, arranged meetings, and facilitated discussion.
- Shreya Jain
  - Implemented predictive analysis techniques to derive insights from the data.
  - Analyzed the relationship between ratings and various different features.
  - Predicted Ratings based on various features and compared models.
  - Consolidated code with coherent markdown and ensured its readability and efficiency.

(2) Collaboration Approaches

- Defined project scope, assigned tasks according to each team member's interest, and set due date for each task.
- Conducted weekly meetings using Google Meet for effective discussion and coordination.
- Utilized Notion and GitHub for collaboration, documentation, and version control.
- Implemented peer code review and proofread peers' report writing.

#	Report Section	Writer
1	Motivation	Nan-Hsin Lin
2	Data Sources	Nan-Hsin Lin
3	Data Manipulation	Aryan Sharma
4.1	Descriptive Analysis	Aryan Sharma
4.2	Diagnostic Analysis	Nan-Hsin Lin
4.3	Predictive Analysis	Shreya Jain
5	Statement of Work	Shreya Jain