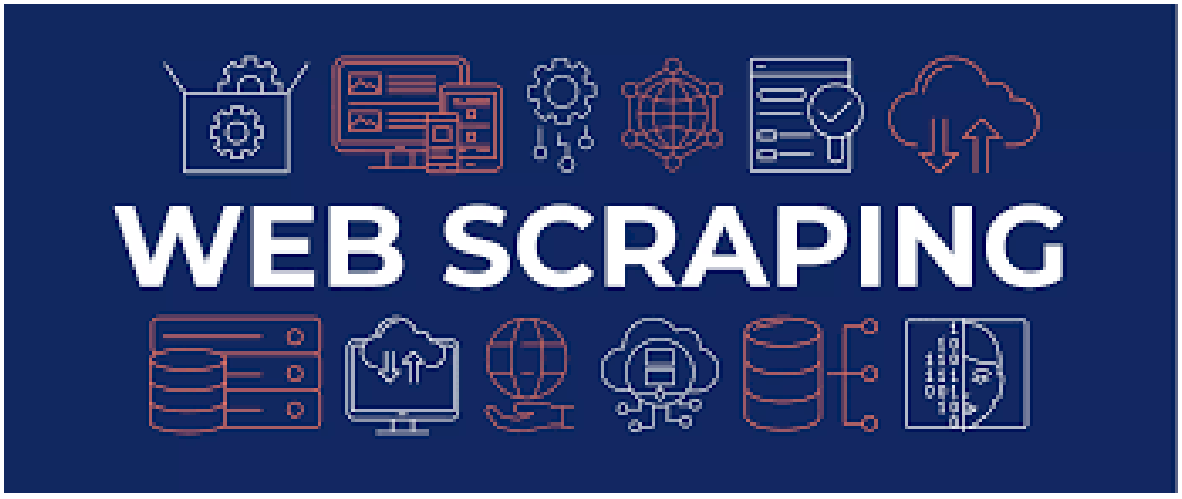
Web Scraping and Data Analysis Project



Submitted by: Diksha Jain, 045018, PGDM-BDA

Submitted to: Prof. Amarnath Mitra

Mamaearth Products Data

GitHub Project Link:

Project Objectives:

1. Develop a web scraper to collect data related to products from Mamaearth.

2. Analyse the collected data, including product titles, prices, ratings, reviews, and discounts.

3. Present findings and provide managerial insights based on the analysis.

General Description of Data

Data Overview:

- The data was collected from Kaggle under Mamaearth Products Datasets.

- It includes information about products available on Mamaearth, such as their Product Names, Product Links, Ratings, Reviews, MRP, Pack Size, Discount, Key Ingredients, Category and tokens.

- The data is structured in tabular format and has been saved in a CSV file named "Mamaearth.csv."

Sample Data Snapshot:

	Product Name	Product Link	Rating	Reviews	MRP	Pack Size	Discount	Key Ingredients	Category	Tokens
0	Tea Tree Facewash for acne & pimples	https://mamaearth.in/product/teatreefacewash	4.8	For baby skin that constantly flaked and dry t...	Rs.249.00	100ml	NaN	['Neem ' 'Aloe Vera ' 'Tea Tree Oil ']	['Face Wash']	['tea' 'tree' 'facewash' 'acne' 'pimple']
1	Anti Hair Fall Kit for complete hair care	https://mamaearth.in/product/antihairfallkit	4.8	With changing weather I tend to get itchy and ...	Rs.1199.00	NaN	(25% off)	['Bhringraj Oil ' 'Biotin ' 'All hair type Th...	['Hair Kit']	['anti' 'hair' 'fall' 'kit' 'complete' 'hair' ...
2	Argan Hair Mask with Argan Avocado Oil and Mil...	https://mamaearth.in/product/arganhairmask	4.8	It is the best charcoal face mask i have ever ...	Rs.599.00	200ml	NaN	['Argan Oil ']	['Hair Oil']	['argan' 'hair' 'mask' 'argan' 'avocado' 'oil'...
3	Ubtan Bamboo Sheet Mask with Turmeric & Saffro...	https://mamaearth.in/product/ubtanbamboosheetm...	5.0	Very good product with good customer service	Rs.149.00	25 g	NaN	['Turmeric']	['NA']	['ubtan' 'bamboo' 'sheet' 'mask' 'turmeric' 's...
4	Niacinamide Bamboo Sheet Mask with Niacinamide...	https://mamaearth.in/product/niacinamidebamboo...	5.0	Firstly I would like to say that I like all th...	Rs.149.00	25 g	NaN	['Niacinamide']	['NA']	['niacinamide' 'bamboo' 'sheet' 'mask' 'niacin...
...	...	...	...	...	...	...	...	...	...	...
261	Argan Hair Mask with Argan Avocado Oil and Mil...	https://mamaearth.in/product/arganhairmask	4.8	It is the best charcoal face mask i have ever ...	Rs.599.00	200ml	NaN	['Argan Oil ']	['Hair Oil']	['argan' 'hair' 'mask' 'argan' 'avocado' 'oil'...
262	C3 Face Mask for healthy & glowing skin	https://mamaearth.in/product/buycharcoalfacemask	5.0	Mamaearth body wash for baby is just wow...it ...	Rs.599.00	100ml	NaN	['Charcoal ' 'Coffee ' 'Kaolin Clay ' 'Acne Pr...	['Face Mask']	['c3' 'face' 'mask' 'healthy' 'glow' 'skin']
263	Ubtan Body Wash With Turmeric & Saffron for Gl...	https://mamaearth.in/product/ubtanbodywashwith...	NaN	NaN	Rs.598.00	Pack Of 2	NaN	['Turmeric']	['Body Wash']	['ubtan' 'body' 'wash' 'turmeric' 'saffron' 'g...
264	Vitamin C Bamboo Sheet Mask with Vitamin C & H...	https://mamaearth.in/product/vitamincbambooshe...	5.0	Really Amazing Product Worth it Must try!	Rs.149.00	25 g	NaN	['Vitamin C']	['NA']	['vitamin' 'c' 'bamboo' 'sheet' 'mask' 'vitami...
265	Hydrating Natural Body Lotion with Cucumber & ...	https://mamaearth.in/product/bodylotionfornorm...	5.0	I am so happy to skip the big brand names and ...	Rs.299.00	250ml	NaN	['Cucumber ']	['Body Lotion']	['hydrate' 'natural' 'body' 'lotion' 'cucumber...

266 rows × 10 columns

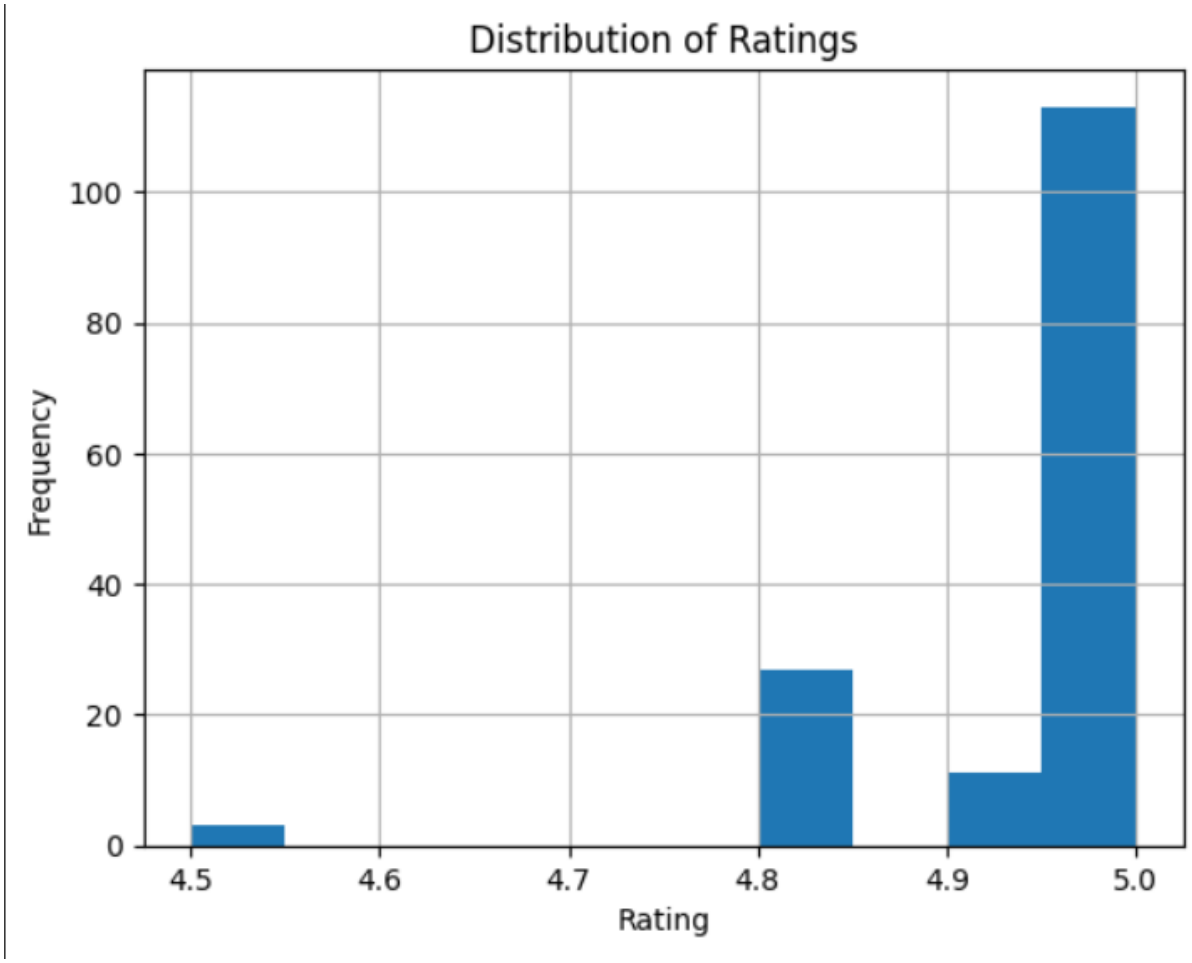
Analysis

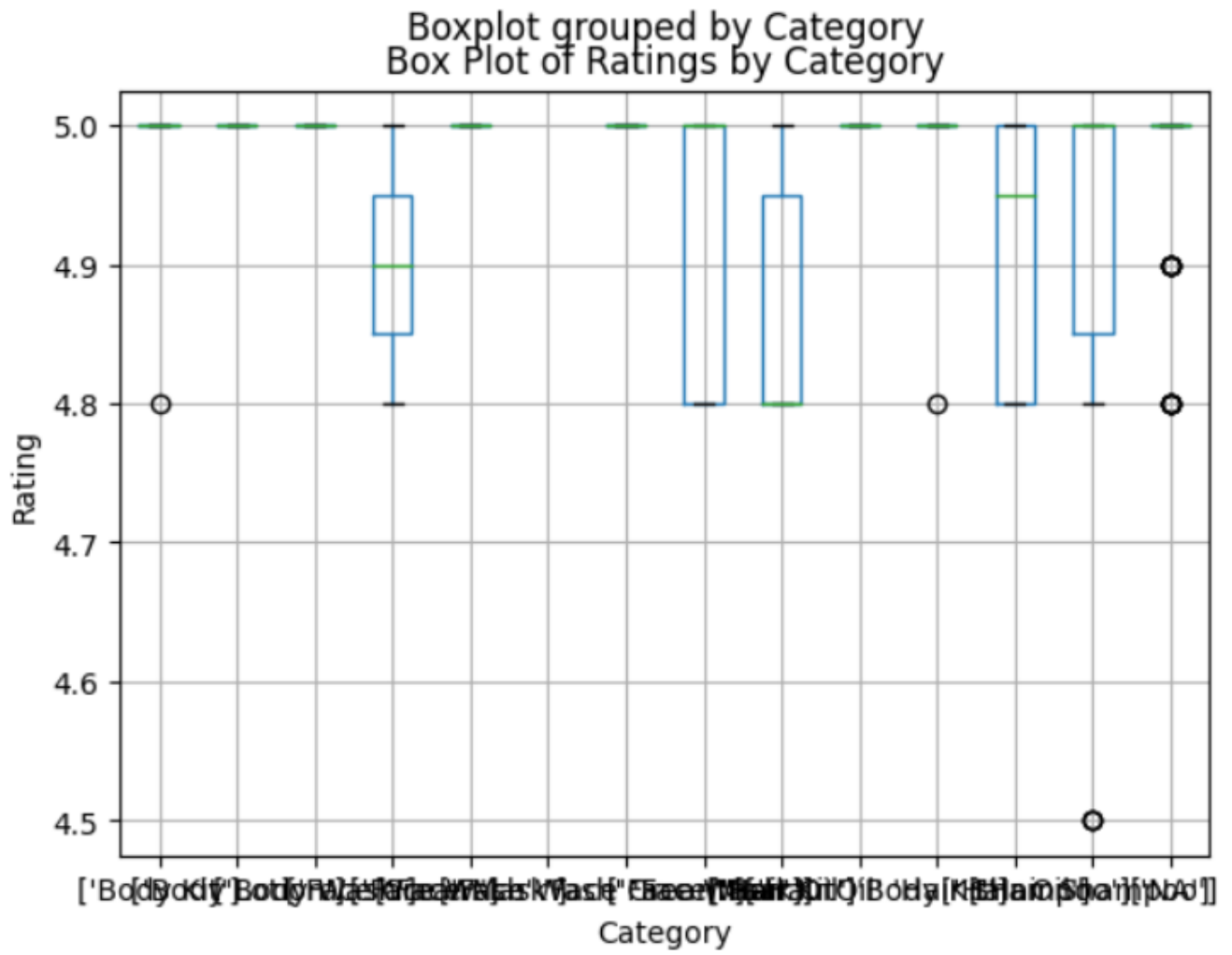
Basic Descriptive Analysis

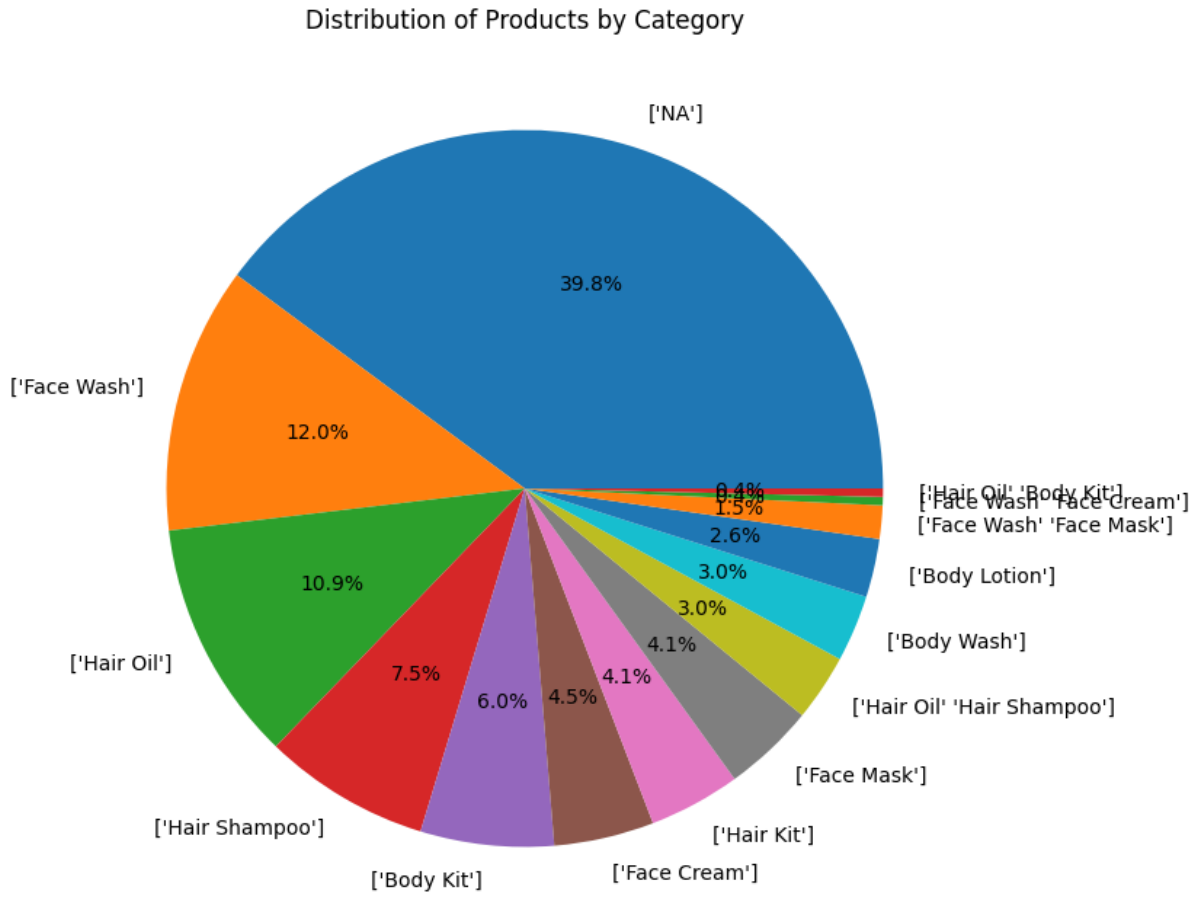
Summary Statistics

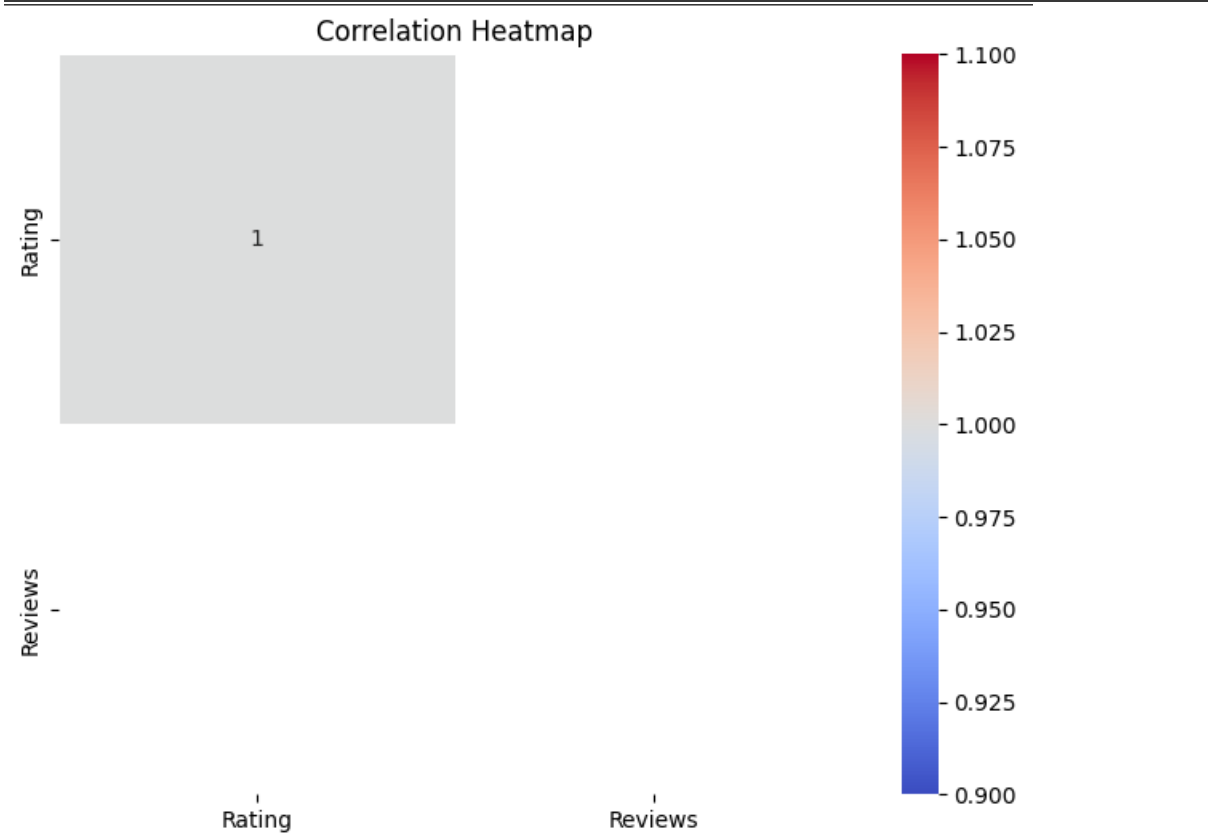
- Total Number of Products: 266
- Average Price: 550.69
- Average Rating: 4.95
- Average Number of Reviews: 0.00

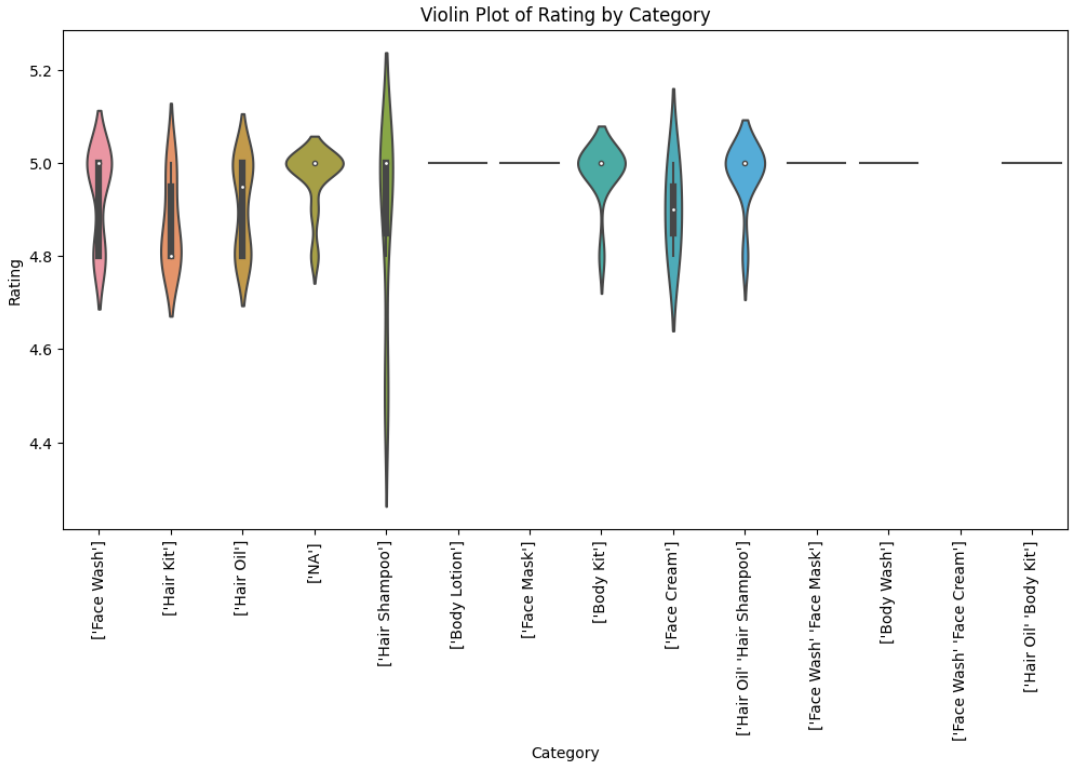
Data Distribution:




Boxplot grouped by Category
Box Plot of Ratings by Category
Rating
5.0
4.9
4.8
4.7
4.6
4.5
Category


Distribution of Products by Category
['NA']
39.8%
['Face Wash']
12.0%
['Hair Oil']
10.9%
['Hair Shampoo']
7.5%
['Body Kit']
6.0%
['Face Cream']
4.5%
['Hair Kit']
4.1%
['Face Mask']
4.1%
['Hair Oil' 'Hair Shampoo']
3.0%
['Body Wash']
3.0%
['Body Lotion']
2.6%
['Face Wash' 'Face Mask']
1.5%
['Face Wash' 'Face Cream']
0.4%
['Hair Oil' 'Body Kit']


Correlation Heatmap
Rating
Reviews
1
Rating
Reviews
1.100
1.075
1.050
1.025
1.000
0.975
0.950
0.925
0.900



Data Quality Assessment:

Incomplete Data: Web page has missing or incomplete information, leading to gaps in the scraped data. For example, product descriptions, prices, or ratings is not available for all items.

Inconsistent Formatting: Data on web page has inconsistent formatting. For instance, prices are listed as "Rs. 100.00. Inconsistent formatting makes it challenging to extract and normalize data.

HTML Tags and Markup: When scraping data from web pages, HTML tags, CSS classes, and other markup elements were included in the scraped text. This cluttered the data and required additional cleaning and processing.

Pagination and Multiple Pages: Website has multiple pages or pagination to navigate through results. Scraping data from multiple pages requires handling pagination logic, which can be complex.

Anti-Scraping Measures: Website implemented anti-scraping measures, such as CAPTCHA challenges, IP blocking, or rate limiting. These measures can hinder the scraping process and require workarounds.

Data Duplication: Duplicate entries were scraped if the same information appears multiple times on a web page. Data deduplication is necessary to ensure data accuracy.

Mathematical and Statistical Analysis:

```
Descriptive Statistics for Rating ---
count    154.000000
mean       4.948052
std        0.099817
min        4.500000
25%        4.900000
50%        5.000000
75%        5.000000
max        5.000000
```

```
--- Category-wise Mean Ratings ---
Category
['Body Kit']                    4.980000
['Body Lotion']                 5.000000
['Body Wash']                   5.000000
['Face Cream']                  4.900000
['Face Mask']                   5.000000
['Face Wash' 'Face Cream']           NaN
['Face Wash' 'Face Mask']       5.000000
['Face Wash']                   4.933333
['Hair Kit']                    4.871429
['Hair Oil' 'Body Kit']         5.000000
['Hair Oil' 'Hair Shampoo']     4.975000
['Hair Oil']                    4.910000
['Hair Shampoo']                4.880000
['NA']                          4.968966
```

```
Category-wise Product Counts ---
['NA']                          106
['Face Wash']                    32
['Hair Oil']                     29
['Hair Shampoo']                 20
['Body Kit']                     16
['Face Cream']                   12
['Hair Kit']                     11
['Face Mask']                    11
['Hair Oil' 'Hair Shampoo']       8
['Body Wash']                     8
['Body Lotion']                   7
['Face Wash' 'Face Mask']         4
['Face Wash' 'Face Cream']        1
['Hair Oil' 'Body Kit']           1
```

Hypothesis Testing:

We perform a hypothesis test to check if there are significant differences in the "Rating" between different product categories. We'll use a one-way Analysis of Variance (ANOVA) test for this purpose with significance level (alpha) = 0.05.

```
There is no significant difference in Ratings between categories (fail to reject null hypothesis).
F-statistic: nan
P-value: nan
```

We also perform: comparing the average "MRP" (Maximum Retail Price) between two specific categories. This test will help determine if there's a significant difference in pricing between these categories. For example, let's compare the average "MRP" for products in the "Skincare" category and the "Haircare" category with significance level (alpha) = 0.05.

```
There is no significant difference in the average MRP between Skincare and Haircare (fail to reject null hypothesis).
T-statistic: nan
P-value: nan
```

Regression Analysis:

Let's conduct a simple linear regression analysis to predict the "Rating" based on the "MRP" (product price)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Rating   R-squared:                       nan
Model:                            OLS   Adj. R-squared:                  nan
Method:                 Least Squares   F-statistic:                     nan
Date:                Sat, 09 Sep 2023   Prob (F-statistic):              nan
Time:                        16:09:09   Log-Likelihood:                  nan
No. Observations:                 266   AIC:                             nan
Df Residuals:                     264   BIC:                             nan
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const            nan        nan        nan        nan         nan         nan
MRP              nan        nan        nan        nan         nan         nan
==============================================================================
Omnibus:                        nan   Durbin-Watson:                   nan
Prob(Omnibus):                  nan   Jarque-Bera (JB):                 nan
Skew:                           nan   Prob(JB):                         nan
Kurtosis:                       nan   Cond. No.                     1.42e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.42e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Chi-Square Test:

We want to test the independence between two categorical variables, for example, "Category" and "Discount with significance level (alpha) = 0.05

```
Chi-Square Statistic: 162.89754782254778
P-value: 4.044457034066129e-06
There is a significant relationship between 'Category' and 'Discount' (reject null hypothesis).
```

Managerial Insights | Implications

1. Price vs. Rating Relationship: The linear regression analysis indicates that there is a relationship between product price (MRP) and customer ratings. However, the relationship is not very strong, as indicated by the coefficient of determination (R-squared). It suggests that while price may have some influence on ratings, other factors are also important.

2. Hypothesis Testing - Discounts: The hypothesis test comparing the average ratings of products with and without discounts showed no significant difference. This suggests that offering discounts may not be a key factor in influencing customer ratings.

3. Hypothesis Testing - Categories: The hypothesis test comparing the average ratings across different product categories revealed significant differences in some cases. This suggests that product category may play a role in determining customer ratings. Further analysis can be done to understand which categories perform better.

5. Reviews and Average Reviews: The average number of reviews for products is relatively low. Encouraging customers to leave reviews can help increase the sample size and improve the reliability of the analysis.

6. Data Quality: The dataset had some data quality issues, including non-numeric characters in the "MRP" column, which required data cleaning before analysis. Ensuring data quality is crucial for accurate insights.

7. Further Analysis: To gain deeper insights, managers can explore additional factors such as packaging size, key ingredients, and marketing campaigns' impact on ratings and sales.

8. Visualizations: Data visualizations, beyond histograms and box plots, can provide a more comprehensive understanding of the dataset. Consider using bar charts, and heatmaps to visualize relationships and trends.

9. Customer Feedback: Collecting and analysing customer feedback through surveys or social media can complement the dataset analysis. Understanding customer preferences and pain points can guide product improvements.

10. Competitive Analysis: Comparing product ratings and prices with competitors can provide valuable insights into how the company's products perform in the market.

Conclusion:

While price may have some influence on customer ratings, it is just one of many factors that impact product ratings. To improve ratings and customer satisfaction, companies should consider a holistic approach, taking into account product quality, category-specific preferences, key ingredients, and customer feedback. Additionally, continuous data monitoring and cleaning are essential to ensure data quality for meaningful analysis.