iHerb UK Product Analysis and Customer Behaviour Prediction

By Omar Mendy

Agenda

01

Overview

Analysing Consumer Behaviour and Market Trends in the Nutraceutical Industry

iHerb is a global e-commerce site within the nutraceutical industry, an industry projected to reach a market value of $722.49 billion by 2027, growing at a Compound Annual Growth Rate (CAGR) of 9.3%, according to *Fortune Business Insights.* The company offers various products across multiple categories, including supplements, sports nutrition, bath & personal care, beauty, grocery, baby & kids, and pets.

As consumers are becoming increasingly health-conscious and making more informed choices, it is crucial to understand market segmentation, product popularity, and trends within this burgeoning industry.

My project aims to analyse customer behavior for the top 48 products per category on iHerb's website. The analysis will involve studying customer preferences, product popularity, and sales performance by evaluating ratings, prices, and the number of reviews (which can be used to estimate sales).

iHerb's Strategy and Business Model

01 **Wide range of natural products** from 1,200+ trusted brands

02 **Customer satisfaction** focusing on reliable shipping, customer support and hassle-free returns

03 **Freshness and quality assurance** through climate-controlled distribution centres

04 **Competitive pricing** through brand and supplier partnerships

05 **E-commerce excellence** with advanced technology and secure payment options

06 **Global presence** with an extensive international shipping network

Our project covers four of the business strategies at iHerb - namely **wide range of natural products, customer satisfaction, freshness and quality assurance** and **competitive pricing**.

Objective of the Analysis

01

Data Collection & Preparation - Extract product data from various categories like supplements, sports, bath personal care, beauty, grocery, baby and kids, and pets using Python libraries like requests and BeautifulSoup through web scraping techniques.

02

Exploratory & Popularity Analysis - Analyse the popularity of products by studying the correlation between product ratings and the number of reviews or sales. This is to access the popularity of items and identify drivers of sales, which can inform your future marketing strategies.

03

Clustering Analysis - Use market segmentation and clustering techniques such as K-means and Hierarchical Clustering to group customers based on their purchasing behaviour and preferences. This is to identify distinct customer groups and tailor your products, services, and marketing initiatives to their specific needs.

04

Predictive Modelling - use supervised learning algorithms like linear regression, logistic regression, and decision trees to forecast future market trends and consumer behaviour. Analyse the relationship between product reviews, ratings, prices, and their impact on future sales.

Variables Explored

Variable	Description
Number of Reviews and Sales (numreviews)	This represents the number of reviews or sales, which is a measure of the popularity or customer engagement with the product.
Product Ratings (rating)	The rating represents the average customer satisfaction score for a product. It reflects the quality and effectiveness of the product as perceived by the customers.
Price	Price is an essential factor that influences consumer decision-making. We analyse the pricing strategies and their impact on consumer behaviour.
Brand Name	The brand plays a significant role in consumer choices. We examine brand loyalty and the influence of brand reputation on purchasing decisions.
Product Name	Understanding the specific products that attract consumer attention helps us identify trends and preferences within each category.

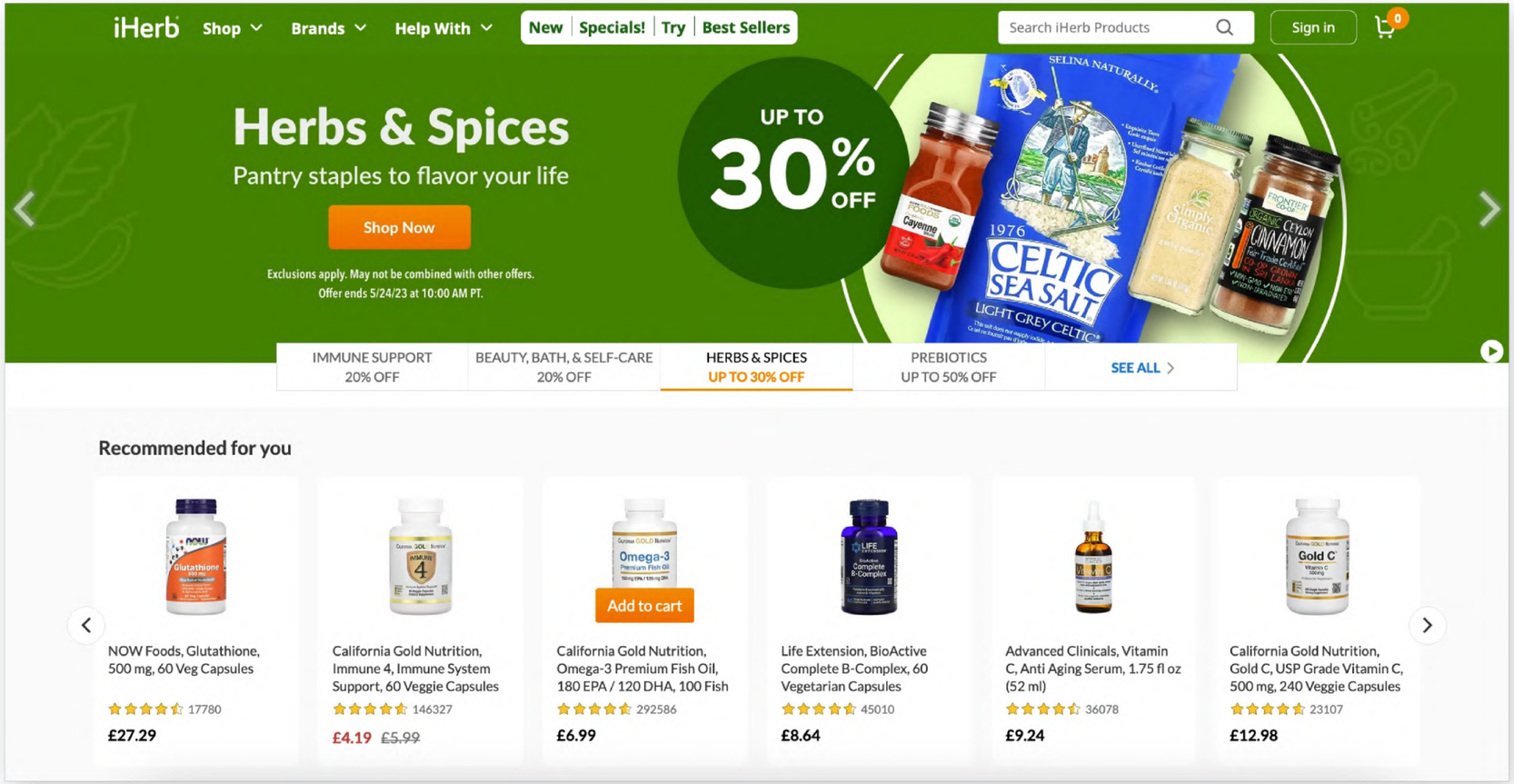
02

Data Collection & Prepatation

Webscraping

In our project, I successfully scraped product data from the **uk.iherb.com** website using Python and the following programming packages:

- **Requests:** A Python library used for making HTTP requests to retrieve web page content. It allows us to access the HTML code of web pages.
- **BeautifulSoup:** A Python library for parsing HTML and XML documents. It provides useful functions to navigate and extract data from the parsed content.



Our scraping efforts were focused solely on public product information that did not involve the collection of any personal user data. Additionally, the website's Terms of Service, Robots.txt, and Privacy Policy do not explicitly prohibit web scraping.

Step 1: Define the Function to Scrape Product Data

This function takes a URL as an input, sends a HTTP GET request to the URL using requests.get(url), parses the HTML response text using BeautifulSoup and finds all product divs. For each product, it extracts the name, number of reviews, rating, and price, checks for missing data, and then appends it to a list (productlist), which is then returned by the function.

```python
# Function to scrape product data
def scrape_products(url):
    response = requests.get(url)
    soup = BeautifulSoup(response.text, 'html.parser')
    products = soup.find_all('div', class_='product-inner')

    product_list = []
    for product in products:
        product_name = product.find('div', class_='product-title').text.strip()
        num_reviews = product.find('a', class_='rating-count')
        rating = product.find('a', class_='stars')
        price = product.find('span', class_='price')

        if num_reviews:
            num_reviews = num_reviews.find('span').text.strip()
        else:
            num_reviews = 'None'

        rating = float(rating['title'].split('/')[0]) if rating else 'None'
        price = price.text.strip() if price else 'None'

        product_list.append({
            'name': product_name,
            'num_reviews': num_reviews,
            'rating': rating,
            'price': price
        })

    return product_list
```

Step 2: Set up URLs to Scrape

This block of code defines a list of URLs that will be used as input to the **scrapeproducts** function.

```
# URLs to scrape
urls = [
    "https://uk.iherb.com/c/supplements",
    "https://uk.iherb.com/c/sports",
    "https://uk.iherb.com/c/bath-personal-care",
    "https://uk.iherb.com/c/beauty",
    "https://uk.iherb.com/c/grocery",
    "https://uk.iherb.com/c/healthy-home",
    "https://uk.iherb.com/c/baby-kids",
    "https://uk.iherb.com/c/pets"
]
```

Step 3: Set up Category Names

This block of code creates a list of category names corresponding to the URLs defined in the previous step. These names will be used later for creating CSV files.

```
# Category names for CSV file names
category_names = [
    "supplements",
    "sports",
    "bath_personal_care",
    "beauty",
    "grocery",
    "healthy_home",
    "baby_kids",
    "pets"
]
```

Step 4: Iterate over URLs and Category Names to Scrape Data and Store it

This block uses the built-in zip function to iterate over the list of URLs and category names simultaneously. For each pair, it calls the scrapeproducts function with the URL and stores the result in products.

```
# Loop over the URLs and category names
for url, category_name in zip(urls, category_names):
    print(f"Scraping products from {url}:")
    products = scrape_products(url)
```

Step 5: Create DataFrame and Save as CSV

In this step, it creates a DataFrame using the pd.DataFrame(products) call, which converts the list of dictionaries into a DataFrame. Then it constructs a file path for the CSV file using the current working directory and the category name. The DataFrame is then saved to this CSV file.

```
# Save the dataframe as a CSV file
file_path = os.path.join(os.getcwd(), f"{category_name}_products.csv")
df.to_csv(file_path, index=False)
```

Step 6: Defining File Paths

This and the subsequent lines define absolute paths to the CSV files that contain the scraped data. Each path corresponds to a CSV file for a specific category of products.

```
# Define file paths
supplements_path = "/Users/user/Library/Mobile Documents/com~apple~CloudDocs/Learning/Projects/iHerb/supplements_products.csv"
sports_path = "/Users/user/Library/Mobile Documents/com~apple~CloudDocs/Learning/Projects/iHerb/sports_products.csv"
bath_personal_care_path = "/Users/user/Library/Mobile Documents/com~apple~CloudDocs/Learning/Projects/iHerb/bath_personal_care_products.csv"
beauty_path = "/Users/user/Library/Mobile Documents/com~apple~CloudDocs/Learning/Projects/iHerb/beauty_products.csv"
grocery_path = "/Users/user/Library/Mobile Documents/com~apple~CloudDocs/Learning/Projects/iHerb/grocery_products.csv"
baby_kids_path = "/Users/user/Library/Mobile Documents/com~apple~CloudDocs/Learning/Projects/iHerb/baby_kids_products.csv"
pets_path = "/Users/user/Library/Mobile Documents/com~apple~CloudDocs/Learning/Projects/iHerb/pets_products.csv"
```

Step 7: Read Datasets into Pandas DataFrames

This and the subsequent lines use the pandas read_csv function to load each CSV file into a DataFrame. Each DataFrame holds the data for a specific category of products. The encoding="latin1" argument specifies the character encoding for the file.

```
# Read datasets into Pandas DataFrames
supplements_df = pd.read_csv(supplements_path, encoding="latin1")
sports_df = pd.read_csv(sports_path, encoding="latin1")
bath_personal_care_df = pd.read_csv(bath_personal_care_path, encoding="latin1")
beauty_df = pd.read_csv(beauty_path, encoding="latin1")
grocery_df = pd.read_csv(grocery_path, encoding="latin1")
baby_kids_df = pd.read_csv(baby_kids_path, encoding="latin1")
pets_df = pd.read_csv(pets_path, encoding="latin1")
```

Results Table

Once I tabulated the scraped data into a CSV file, I noticed that the 'name' column had both the brand and product names together.

To address this, I utilised the python split() function to separate them into two distinct columns.

	A	B	C	D
1	name	num_review	rating	price
2	California Gold Nutrition, Gold C, USP Grade Vitamin C, 1,000 mg, 60 Veggie Capsules	243815	4.8	£2.64
3	California Gold Nutrition, Omega-3 Premium Fish Oil, 180 EPA / 120 DHA, 100 Fish Gelatin Softgels	291442	4.8	£6.99
4	California Gold Nutrition, Vitamin D3, 125 mcg (5,000 IU), 360 Fish Gelatin Softgels	223710	4.9	£13.58
5	California Gold Nutrition, LactoBif Probiotics, 30 Billion CFU, 60 Veggie Capsules	86590	4.7	£19.97
6	California Gold Nutrition, Vitamin D3, 125 mcg (5,000 IU), 90 Fish Gelatin Softgels	223710	4.9	£4.19
7	California Gold Nutrition, CollagenUP, Hydrolyzed Marine Collagen Peptides with Hyaluronic Acid and Vitamin C, Unflavored, 7.26 oz (206 g)	169521	4.7	£16.97

```python
def split_name(df):
    # split 'name' into 'brand_name' and 'product_name'
    df[['brand_name', 'product_name']] = df['name'].str.split(',', 1, expand=True)
    return df
```

	A	B	C	D	E	F	G
1	name	num_review(sales)	rating	price	brand_name	product_name	
2	California Gold Nutrition, Gold	243815	4.8	£2.64	California Gold Nutrition	Gold C, USP Grade Vitamin C, 1,000 mg, 60 Veggie Capsules	
3	California Gold Nutrition, Ome	291442	4.8	£6.99	California Gold Nutrition	Omega-3 Premium Fish Oil, 180 EPA / 120 DHA, 100 Fish Gelatin Softgels	
4	California Gold Nutrition, Vita	223710	4.9	£13.58	California Gold Nutrition	Vitamin D3, 125 mcg (5,000 IU), 360 Fish Gelatin Softgels	
5	California Gold Nutrition, Lact	86590	4.7	£19.97	California Gold Nutrition	LactoBif Probiotics, 30 Billion CFU, 60 Veggie Capsules	
6	California Gold Nutrition, Vita	223710	4.9	£4.19	California Gold Nutrition	Vitamin D3, 125 mcg (5,000 IU), 90 Fish Gelatin Softgels	
7	California Gold Nutrition, Colla	169521	4.7	£16.97	California Gold Nutrition	CollagenUP, Hydrolyzed Marine Collagen Peptides with Hyaluronic Acid and V	

Data Quality

After completing the web scrape, I ran the info() and isnull() functions to thoroughly examine the data for any inconsistencies. Here are my initial findings with the data:

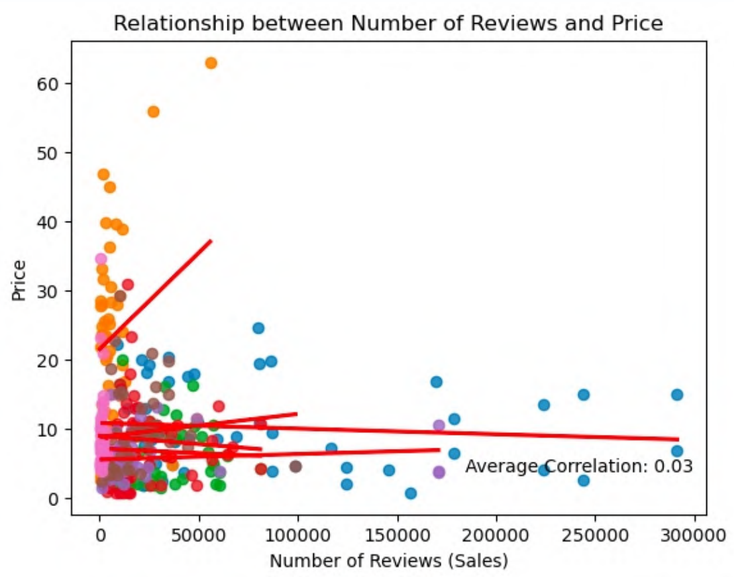
- **Consistency:** All the datasets are consistent. The data in each column appears to be in the same format, with appropriate units being used.
- **Completeness:** The provided data contains all necessary values and does not have any missing data indicated by placeholders like “-1”, “null”, or “N/A”.
- **Uniqueness:** At a glance, the dataset seems to have a few potentially duplicated entries (for example, in the supplements dataset, the product with the name ”California Gold Nutrition, Gold C, USP Grade Vitamin C, 1,000 mg, 240 Veggie Capsules” appears twice with the same number of reviews, rating, and brand). But the product sizes vary, so they would not be considered duplicates.
- **Validity:** Data appears to be within expected ranges and types. For instance, ratings fall within 0-5, the standard for most review systems.

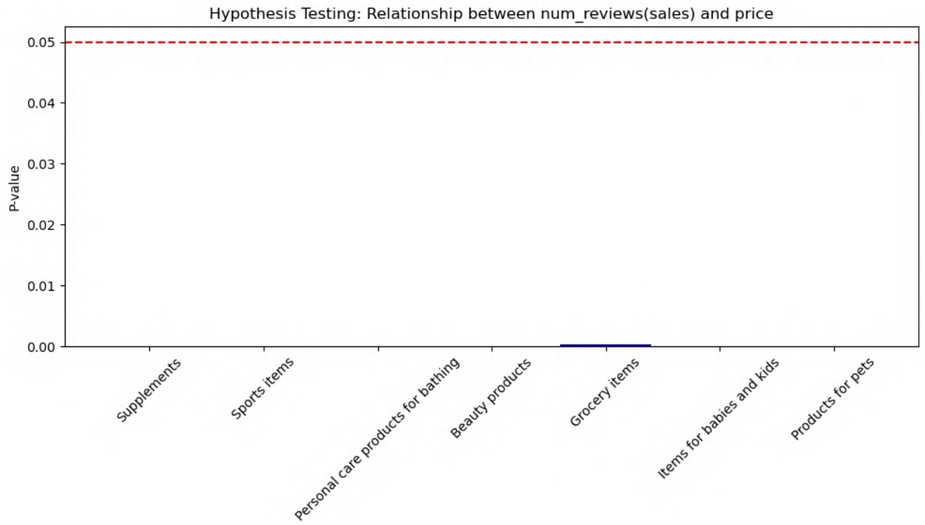
Using Customer Reviews as a Proxy for Sales

Due to iHerb's customer privacy policy, acquiring actual sales data for analysis was not possible. Consequently, I used a number of customer reviews as a proxy for product purchase numbers. This would be feasible since only customers who've made a purchase are allowed to leave a product review on the iHerb website.

Our analysis showed that using review numbers as a substitute for sales was statistically valid. A conducted hypothesis test revealed a substantial correlation between the number of reviews (interpreted as sales) and the product price. This correlation was statistically significant, as demonstrated by an average p-value of 0.00 and a t-statistic of 7.9588, across all product categories. However, it's important to note that the correlation coefficient highlighted a weak to moderate relationship with variation across different product categories.

Therefore, while review numbers provide a robust proxy for sales in general, this relationship may need to be interpreted differently in specific categories.





Outlier Analysis: Number of Reviews (Sales)

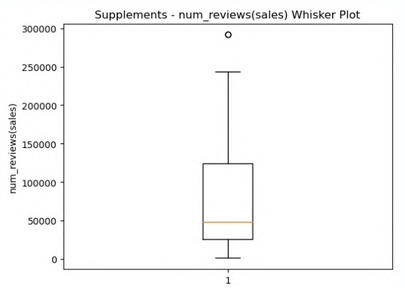
Key components used to analyse potential outliers include the lower and upper fences, which set thresholds for data points that may be considered outliers and data points that are above average, respectively. These thresholds are calculated based on the distribution of the data, using interquartile range (IQR) as a measure.

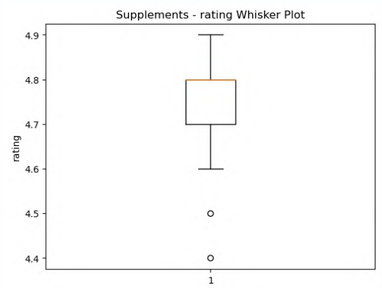
Product Category	Lower Fence	Upper Fence	Number of outliers
Supplements	-121591	272337	2
sports items	-6603.38	14731.6	4
Personal care products for bathing	-24137.2	92480.8	0
Beauty products	-17365.5	64394.5	3
Grocery Items	-9609.5	31440.5	5
Items for babies and kids	-17907.8	49818.2	1
Products for pets	-1603.25	3326.75	0

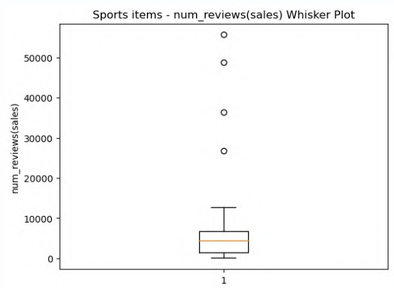
Outlier Analysis: Ratings

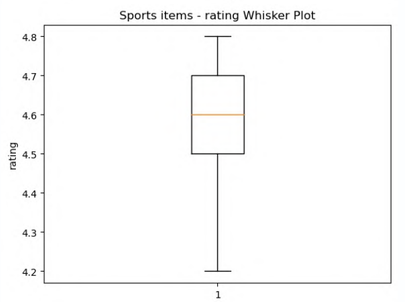
Product Category	Lower Fence	Upper Fence	Number of outliers
Supplements	4.55	4.95	2
sports items	4.2	5	0
Personal care products for bathing	4.35	4.75	4
Beauty products	4.35	4.75	1
Grocery Items	4.45	4.85	10
Items for babies and kids	4.55	4.95	1
Products for pets	4.1375	5.0375	2

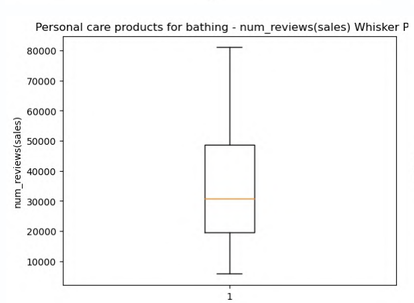
Outlier Analysis: Supplements, Sports, Personal Care for Bathing and Beauty

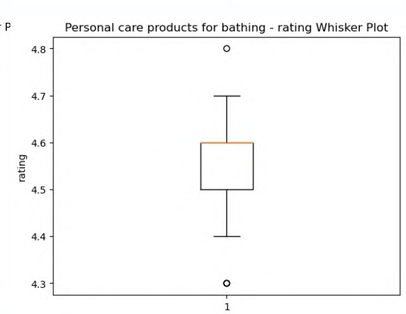


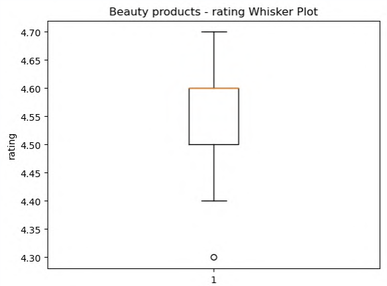












- **Supplements:**

  - There are 2 potential outliers below the lower fence and 2 potential outliers above the upper fence for "numreviews(sales)".

  - There are 2 potential outliers below the lower fence for "rating".

- **Sports items:**

  - There are 4 potential outliers below the lower fence for "numreviews(sales)".

  - There are no outliers identified for "rating".

- **Personal care products for bathing:**

  - There are no outliers identified for "numreviews(sales)".

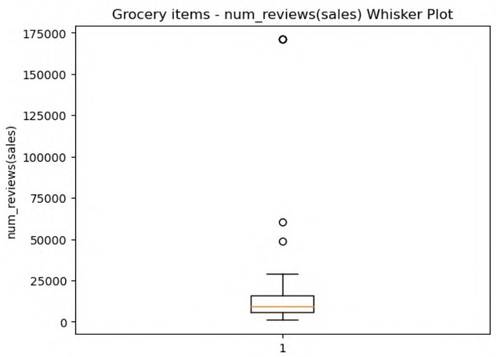
  - There are 4 potential outliers below the lower fence for "rating".

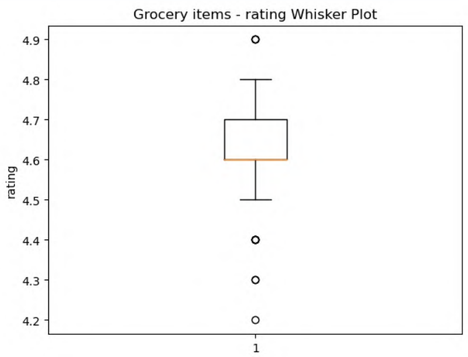
- **Beauty products:**

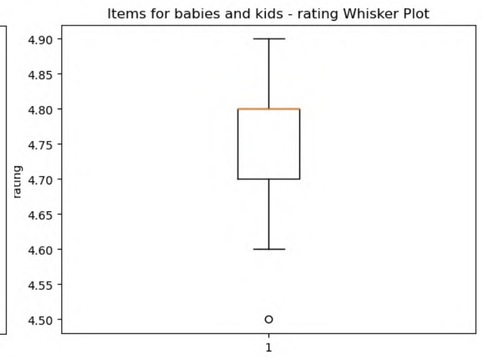
  - There are 3 potential outliers below the lower fence for "numreviews(sales)".

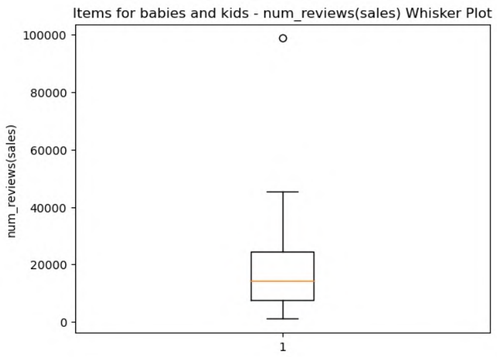
  - There is 1 potential outlier below the lower fence for "rating".

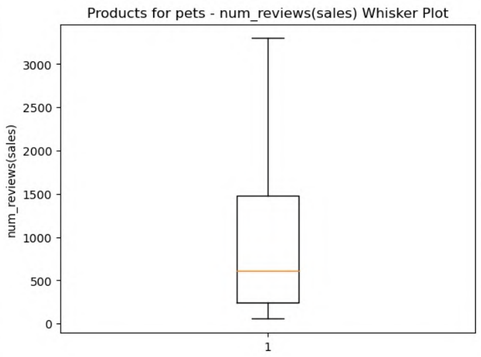
Outlier Analysis: Grocery, pets, babies & kids

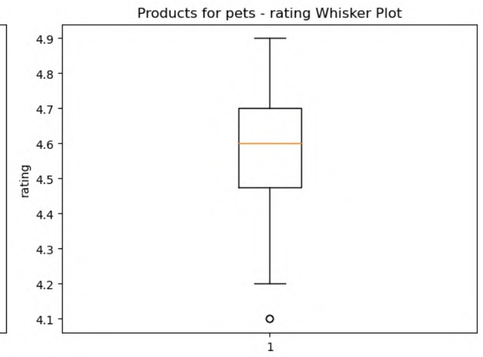










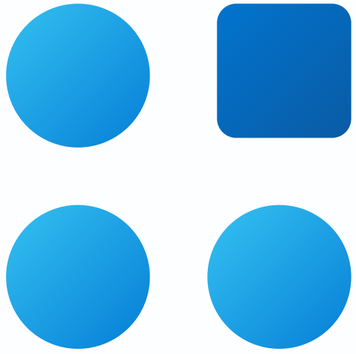


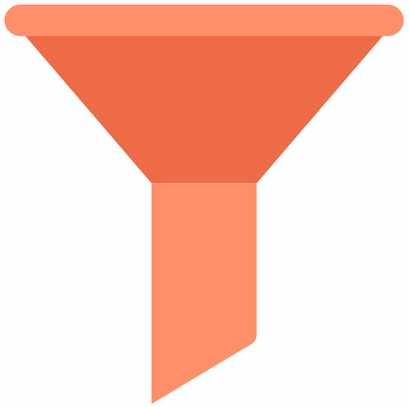
- **Grocery items:**
  - There are 5 potential outliers below the lower fence and 10 potential outliers above the upper fence for "numreviews(sales)".
  - There are no outliers identified for "rating".
- **Items for babies and kids:**
  - There is 1 potential outlier below the lower fence and 1 potential outlier above the upper fence for "numreviews(sales)".
  - There is 1 potential outlier below the lower fence for "rating".
- **Products for pets:**
  - There are no outliers identified for "numreviews(sales)".
  - There are 2 potential outliers above the upper fence for "rating".

Outliers Analysis Results and Data Cleansing

Outliers were detected in all product categories - disproportionately higher in the sports category (15), while the other categories showed an equal number of outliers (3-4).

Despite the sizable number of outliers, eliminating them would not be necessary due to the high data quality of the scraped data. Moreover, the wide range of datasets can unveil various market dynamics, consumer preferences, and product performance related to the customer behaviour datasets. To delve deeper into this variability, an Exploratory and Popularity Analysis can be conducted.



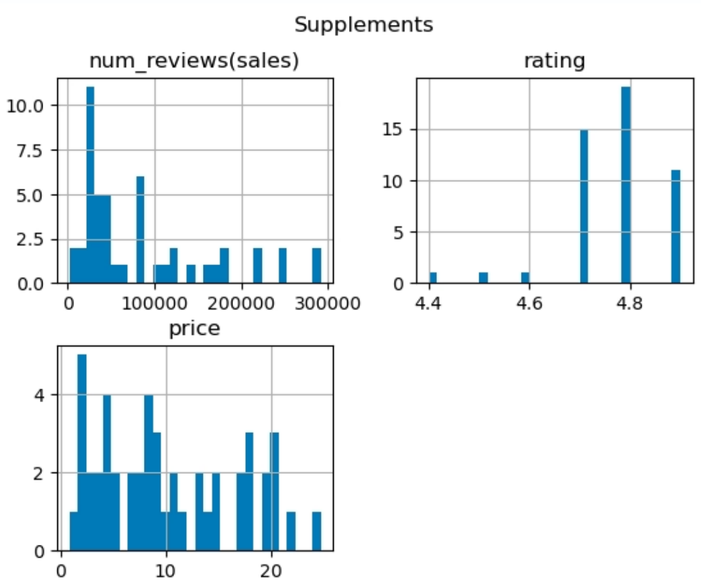


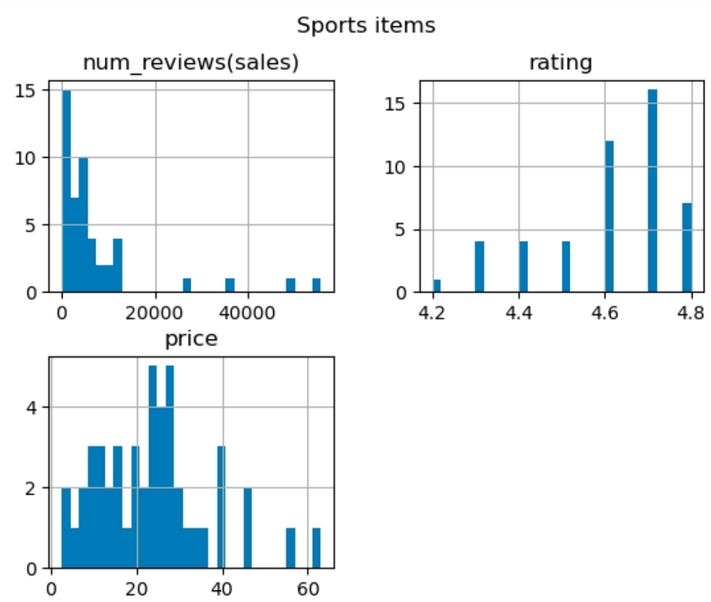
03

Exploratory & Popularity Analysis

Univariate Analysis: Supplements and Sports Products

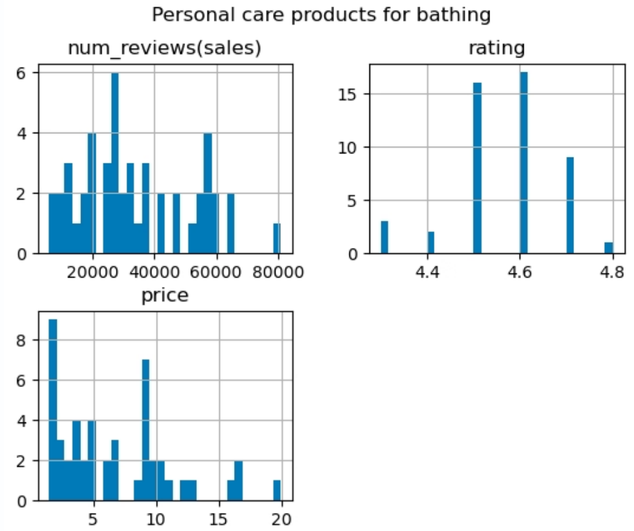
	Mean (x̄) number of reviews (sales) per product	Average Rating	Average Price (£)	Minimum Price	Maximum Price	Product with highest number of reviews (sales)	Total reviews (sales)
Supplements	86,456	4.77	10.21	0.83	24.73	291,442	4,149,907
Sports Items	7,428	4.60	23.68	2.65	63.03	55,707	356,543





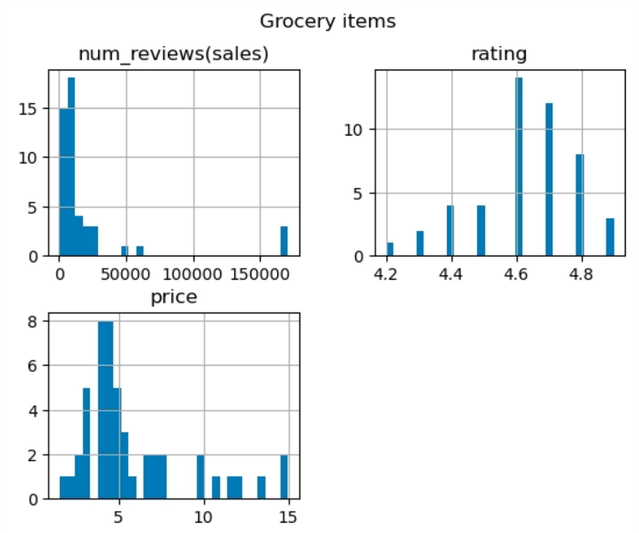
Univariate Analysis: Personal Care for Bathing & Beauty Products

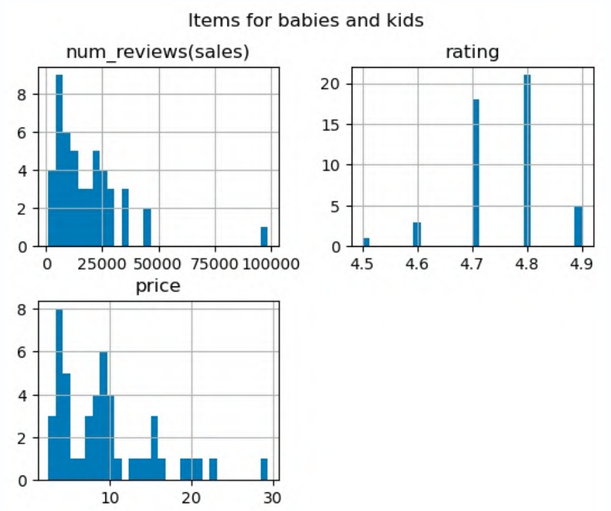
	Mean (x̄) number of reviews (sales) per product	Average Rating	Average Price (£)	Minimum Price	Maximum Price	Product with highest number of reviews (sales)	Total reviews (sales)
Personal care products for bathing	34,140	4.56	6.77	1.50	19.98	81,005	1,638,713
Beauty products	26,281	4.56	8.44	0.83	31.08	81,005	1,261,480

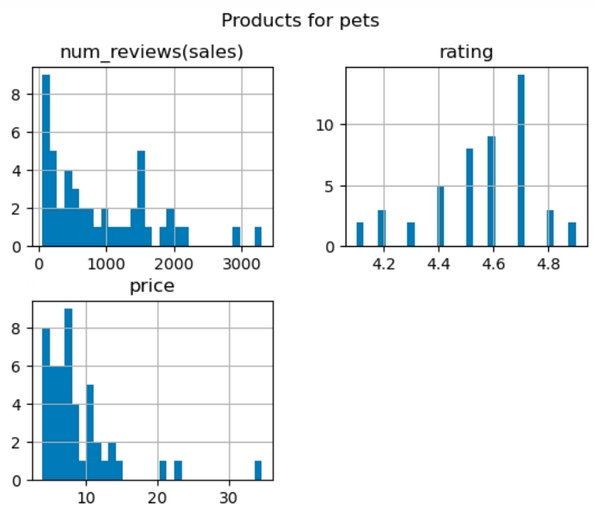


Univariate Analysis: Grocery, Pets, Babies and Kids Products

	Mean (x̄) number of reviews (sales) per product	Average Rating	Average Price (£)	Minimum Price	Maximum Price	Product with highest number of reviews (sales)	Total reviews (sales)
Grocery	22,107	4.63	5.81	1.53	14.98	170,893	1,061,131
Items for babies and kids	18,206	4.75	9.49	2.51	29.37	98,750	873,893
products for pets	909	4.56	8.98	3.94	34.69	3,299	43,616







Try Pitch

Univariate Analysis: Interpretation

Sales Activity:

- Personal care products for bathing and beauty products have higher average sales activity compared to other categories. This suggests that these categories are more popular among consumers, leading to more sales or purchases.
- Products for pets, on the other hand, have relatively lower sales activity. This could indicate that products in this category are less in demand or have a smaller market size than other categories.

Ratings:

- Personal care products for bathing, beauty products, and items for babies and kids have relatively high average ratings. This indicates that customers are generally satisfied with the quality and performance of products in these categories.
- Sports items and products for pets also have relatively high average ratings but lower than the aforementioned categories.

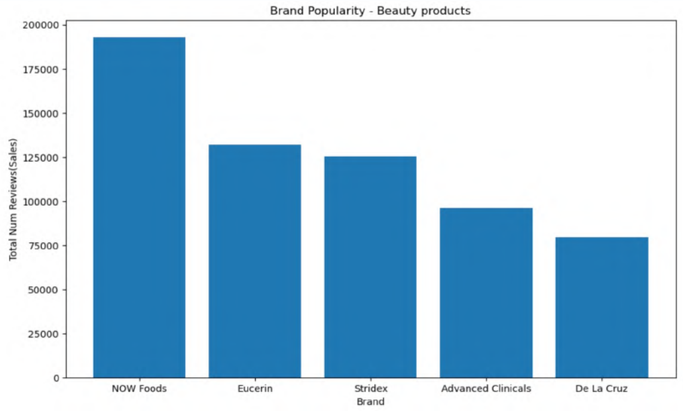
Prices:

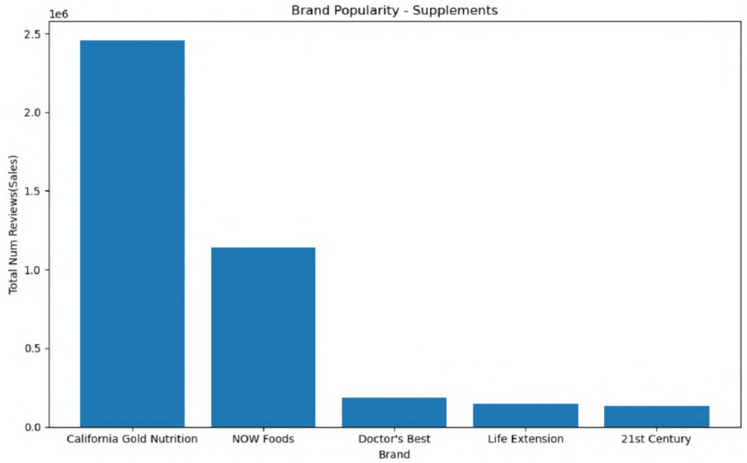
- Average prices vary across different product categories. Beauty products tend to have a higher average price, indicating that they are relatively more expensive than other categories.
- Personal care products for bathing, items for babies and kids, and sports items have average prices in a moderate range.
- Grocery items and products for pets have relatively lower average prices.

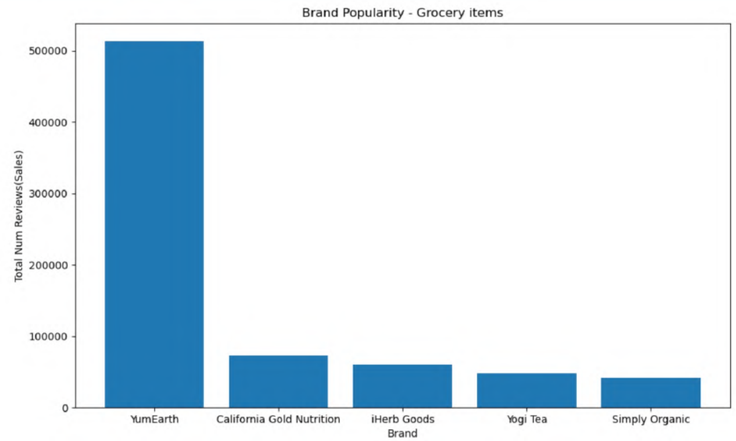
Brand Popularity

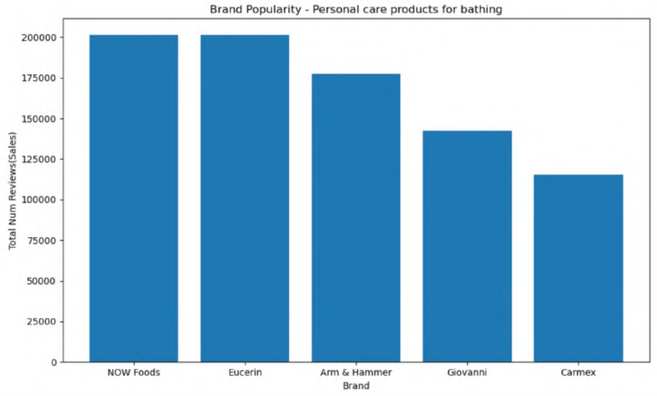
Key Findings:

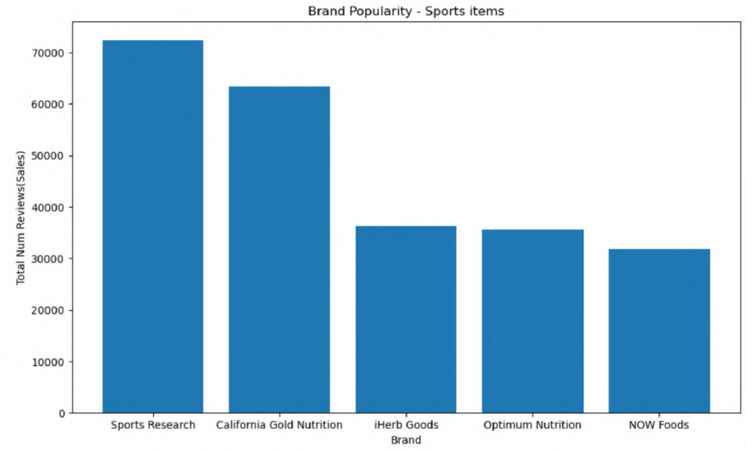
1. *California Gold Nutrition* emerges as a famous brand in multiple categories, including Supplements, Sports items, Grocery items, and Items for babies and kids.
2. *NOW Foods* is another notable brand with a strong presence in multiple categories, including Supplements, Personal care products for bathing, and Beauty products.







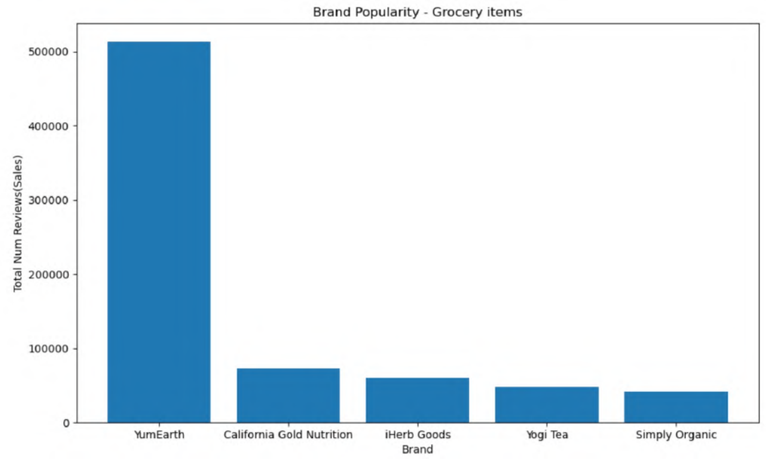


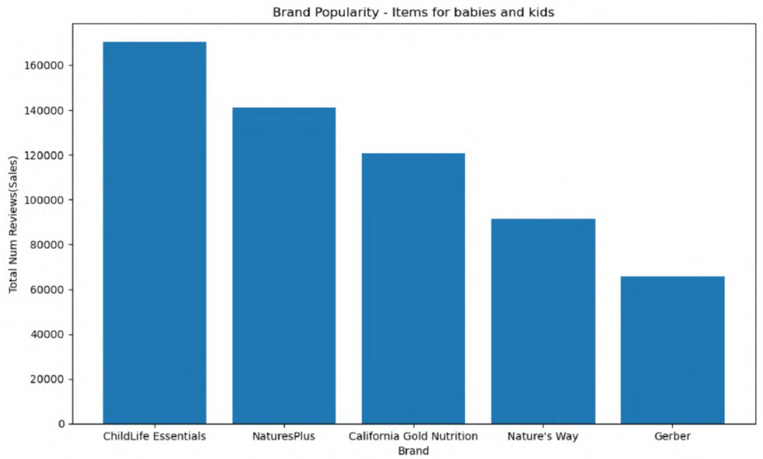


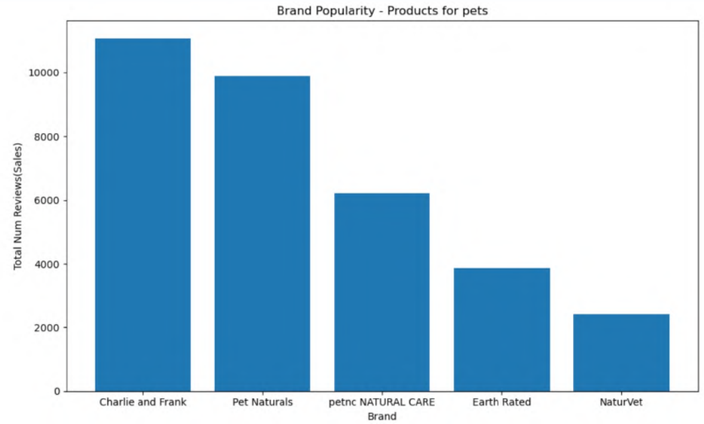
Brand Popularity (Continued)

More key findings:

3. *Eucerin* is a famous brand in both Personal care products for bathing and Beauty products categories.
4. *iHerb Goods* is prominent in the Sports items and Grocery items categories.
5. Other popular brands like *Optimum Nutrition*, *YumEarth*, and *Nature's Way* also demonstrate their influence in specific categories.



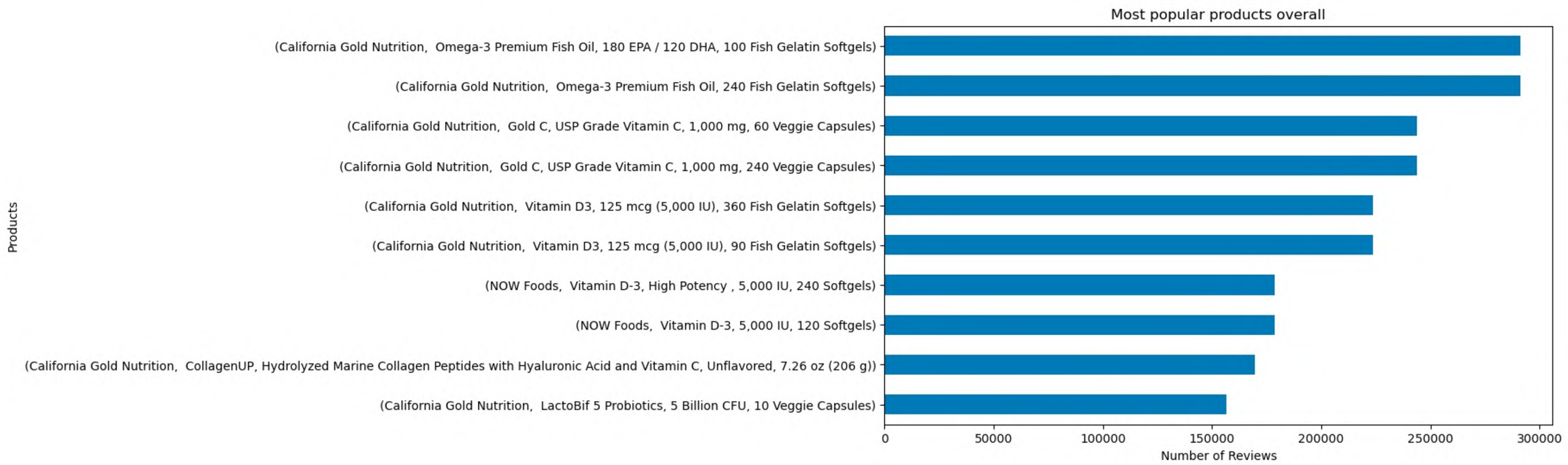




The analysis reveals cross-category brand loyalty, indicating that certain brands consistently deliver quality and meet consumer expectations. This loyalty can lead to repeat purchases and expanded product offerings, enhancing customer retention and competitive advantage.

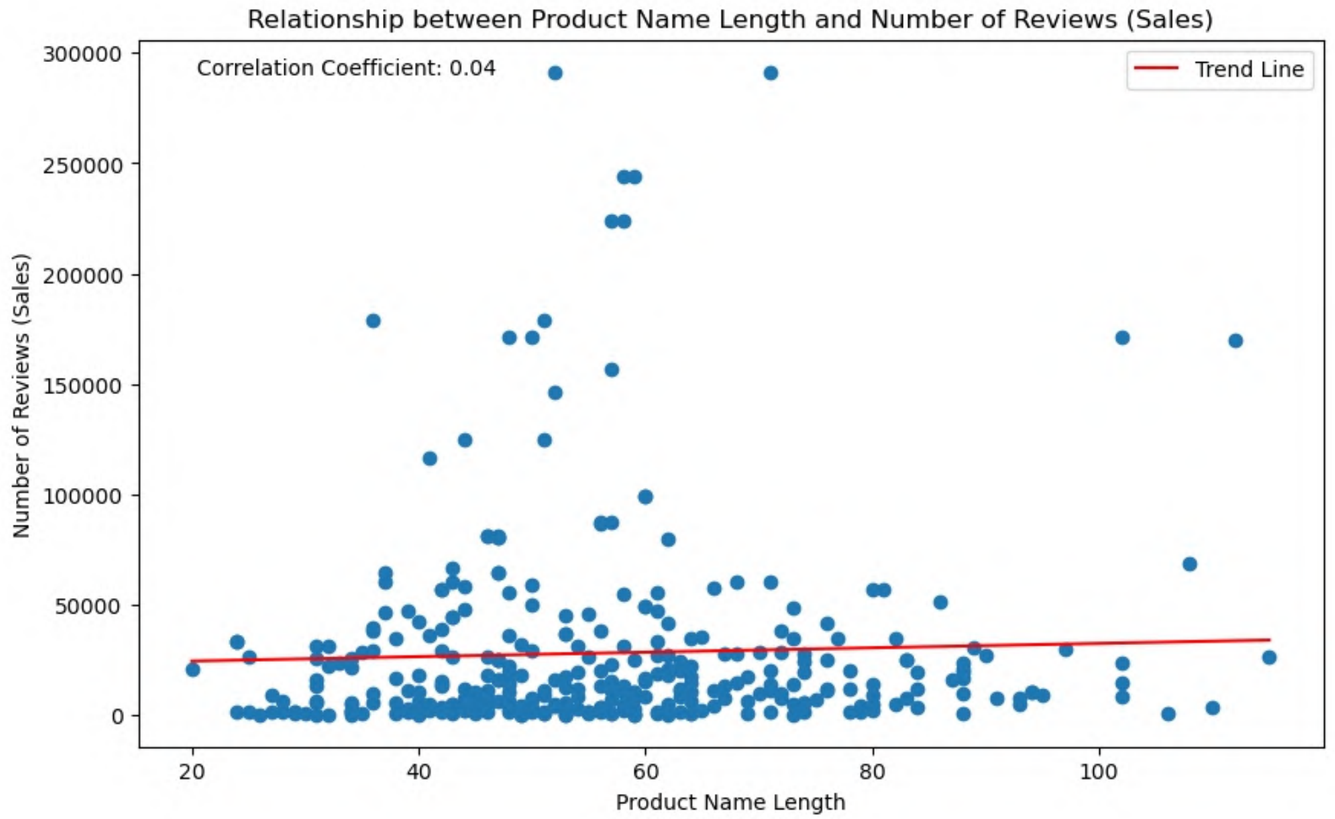
Product Popularity

The top 10 most popular products purchased by iHerb are all in the supplement category. and the top 6 products fall within the *California gold nutrition brand.*

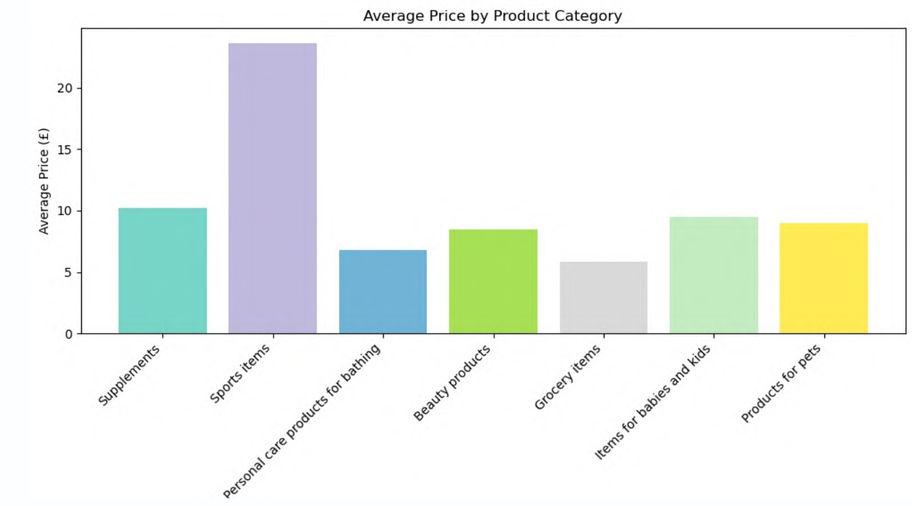


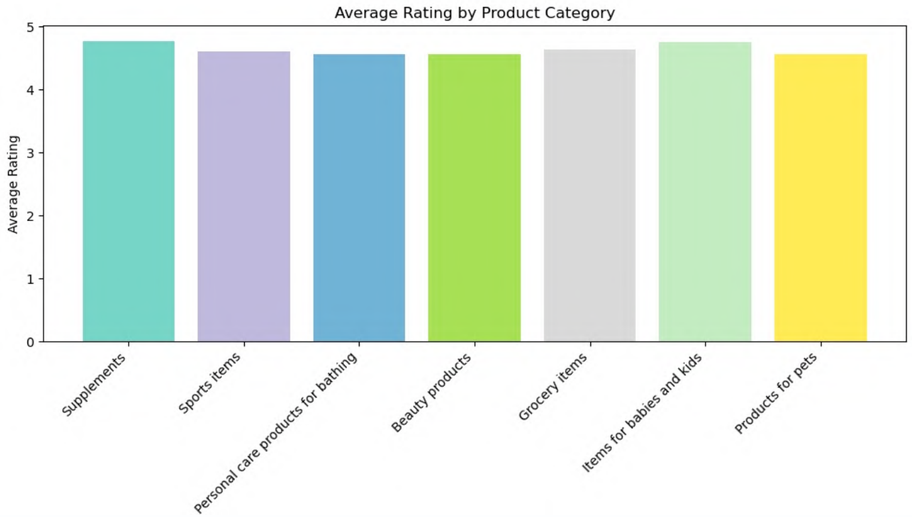
Product Name Length

The correlation coefficient of 0.04 indicates a very weak positive correlation between the product name length and the number of reviews (sales). This suggests the length of the product names does not significantly impact the number of reviews (sales) they receive.



Average Price and Rating by Product Category





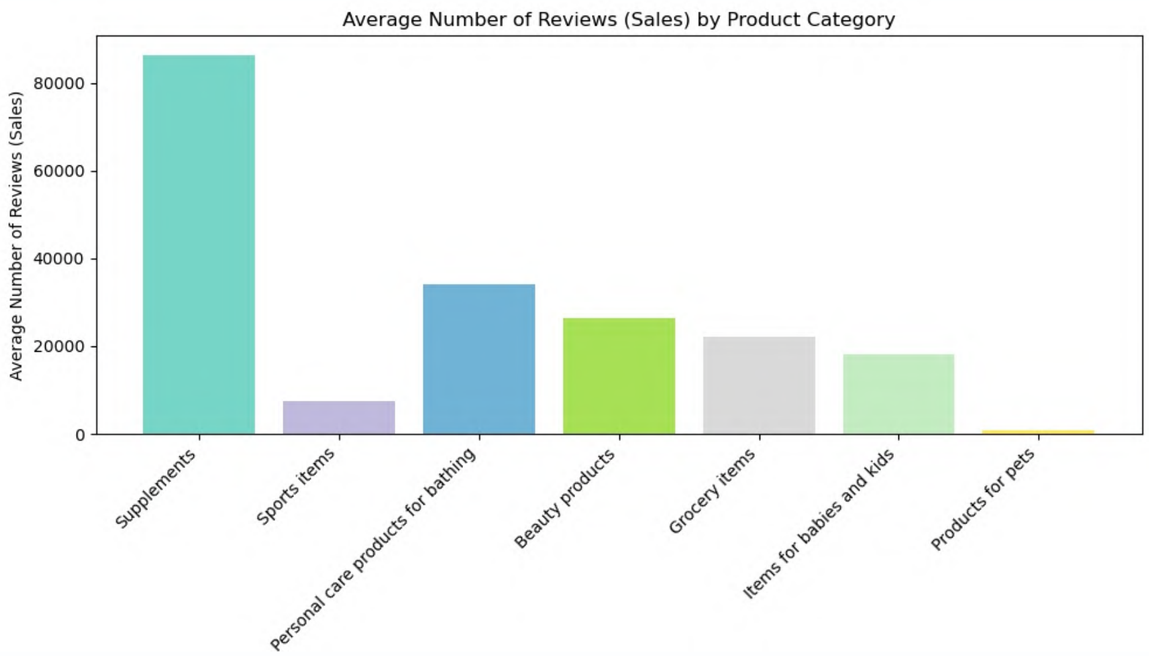
Average Price by Product Category:

- "Sports items" have the highest average price among the product categories, costing around £23.68 on average.
- The lowest average price is for "Grocery items", which cost around £5.81 on average.
- The other categories have average prices ranging from £6.77 to £10.21.

Average Rating by Product Category:

- The average rating for all product categories is consistent, with an average of around 4.56 to 4.77.
- The "Items for babies and kids" category has the highest average rating of 4.75.
- The "Products for pets" category has the lowest average rating of 4.56.

Average Reviews (Sales) by Product Category



Average Reviews (Sales) by Product Category:

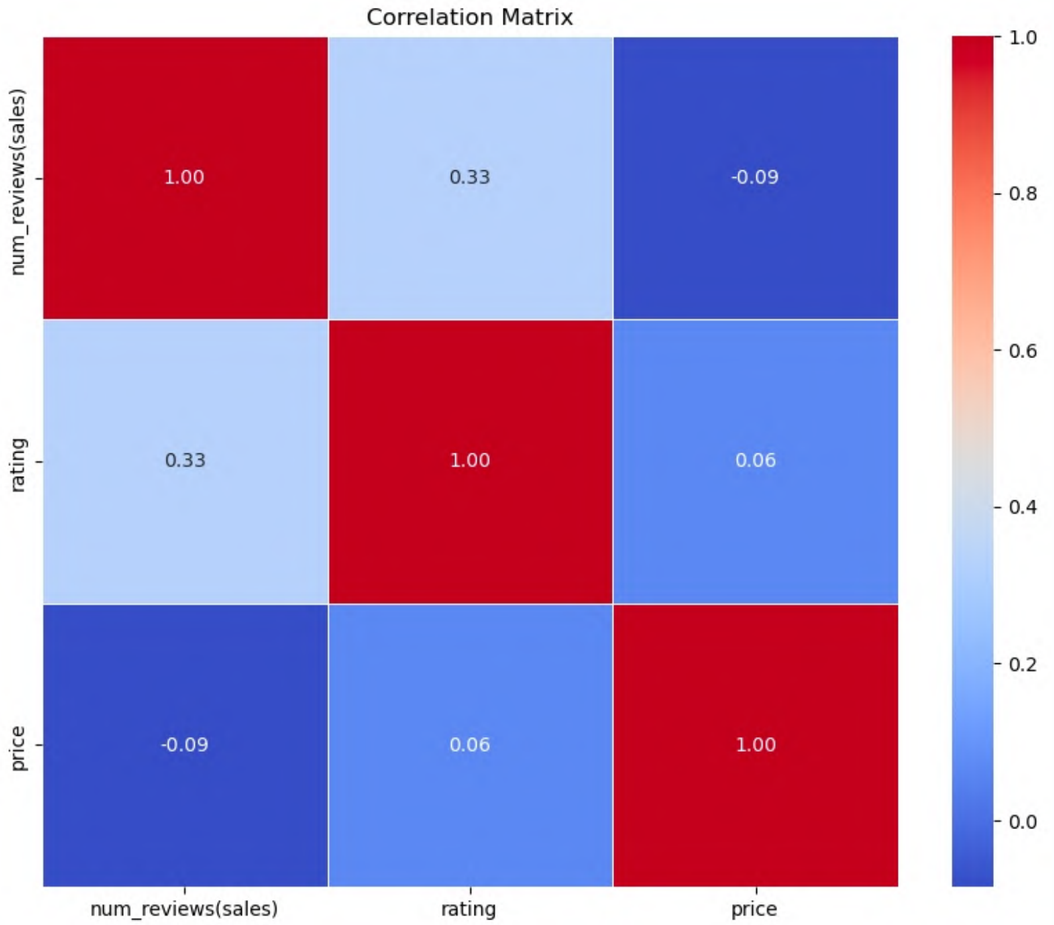
- "Supplements" have the highest average number of reviews (sales), with approximately 86,456 on average.
- "Products for pets" have the lowest average number of reviews (sales), with only 908 on average.
- The other categories have average numbers of reviews (sales) ranging from around 7,428 to 34,140.

Correlation Matrix

- **`Correlation between numreviews(sales) and rating is 0.33`:** The moderate positive correlation suggests more reviewed (and possibly higher selling) products often have better ratings.

- **`Correlation between numreviews(sales) and price is -0.09`:** The weak negative correlation implies that sales volumes and price are not significantly related.

- **`Correlation between rating and price is 0.06`:** The very weak positive correlation indicates the negligible relationship between a product's rating and its price.



Based on the correlation analysis, there are some associations between the variables, but they are generally weak. The strongest correlation is between the number of reviews (sales) and the rating, indicating a moderately positive relationship. The relationship between these variable will be re-examined with predictive modelling.

04

Clustering Analysis

K-means and Hierarchical Clustering

About:

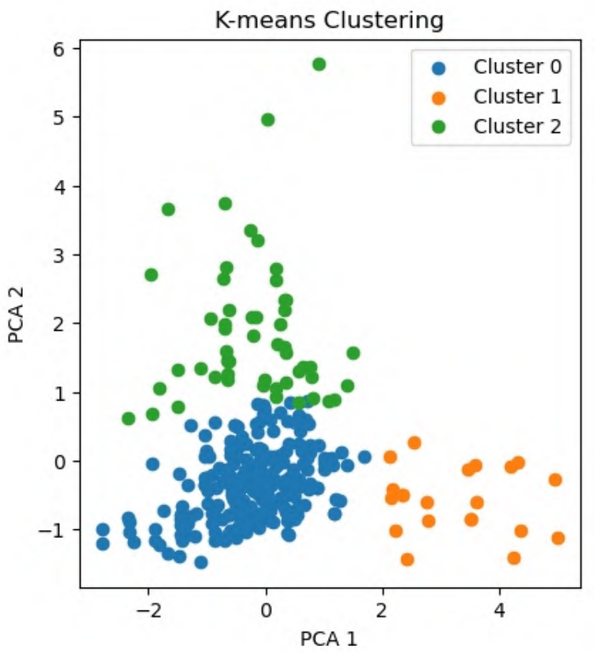
K-means clustering groups data into distinct clusters based on their similarity, while hierarchical clustering forms a tree of clusters by continuously joining or splitting existing clusters.

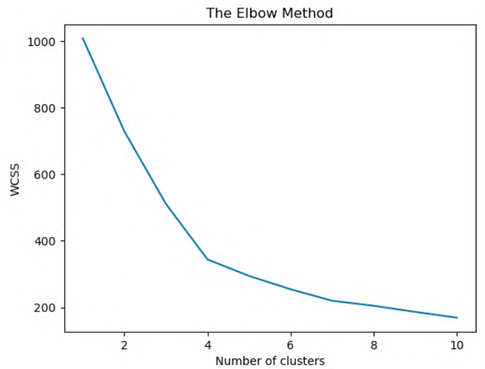
Findings (both models):

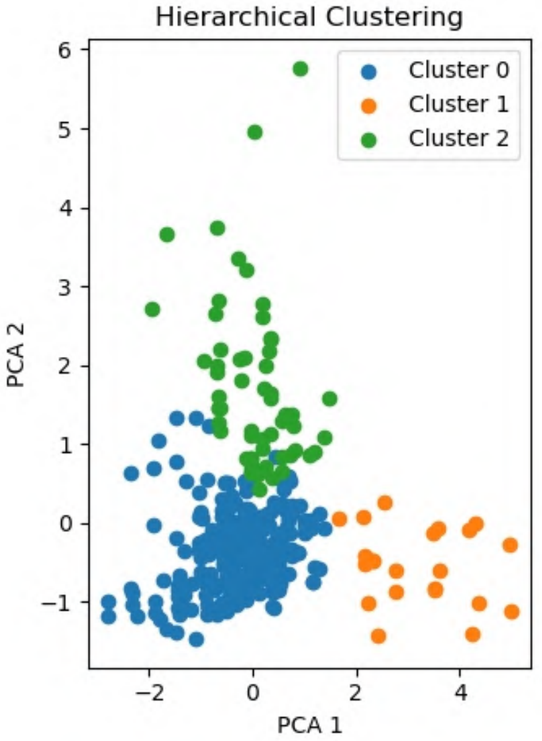
1. Customers in cluster 0: Customers tend to purchase lower-priced products with good ratings and many reviews. These customers are likely price-sensitive but still value quality based on the ratings and reviews.

2. Customers in cluster 1: Customers lower-priced products with much higher ratings and significantly more reviews. These customers are likely very conscious of product quality and rely heavily on user feedback before purchasing.

3. Customers in cluster 2: (both K-means and hierarchical) tend to purchase higher-priced products with high ratings and a moderate number of reviews. These customers are likely less price-sensitive and more focused on quality, as the higher average price and ratings indicate.







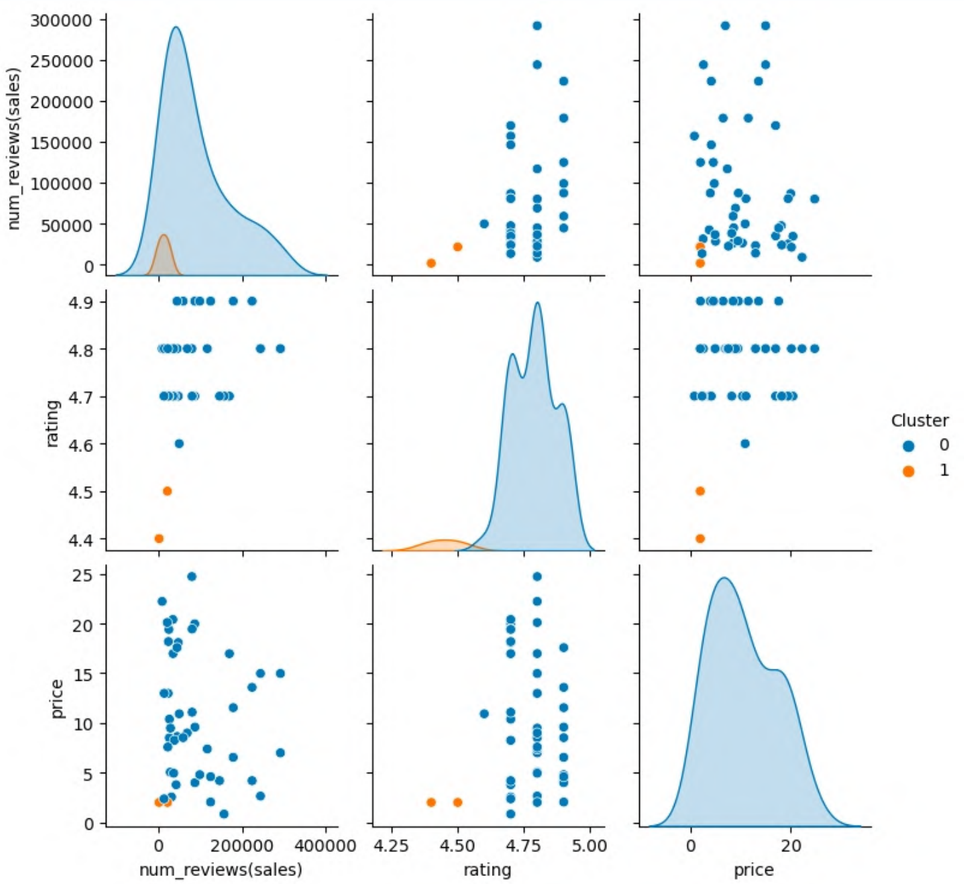
Mean-Shift Clustering

About:

Identifying distinct groups of products without needing to predefine the number of groups.

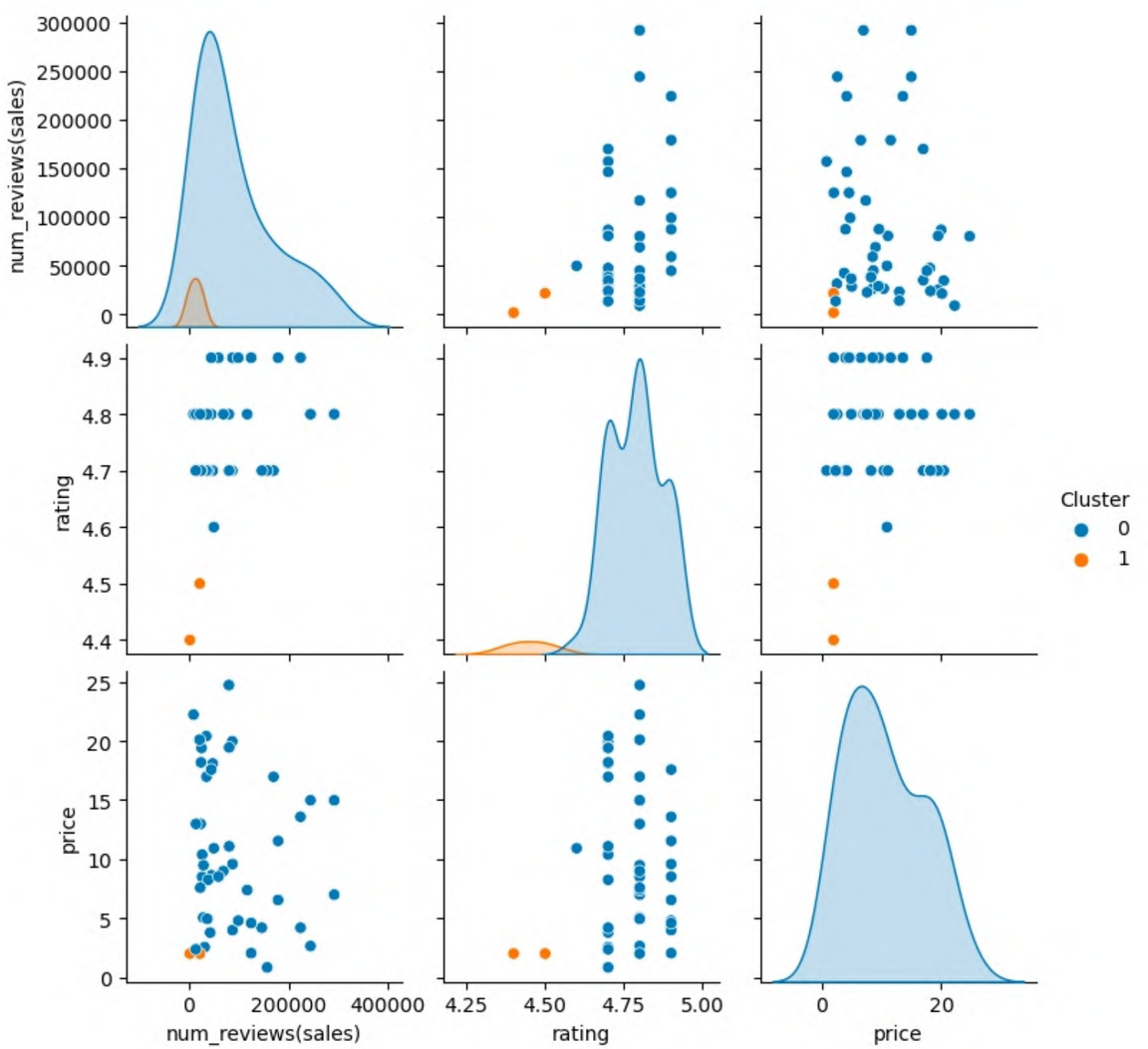
Findings:

1. **Supplements:**
   - Cluster 1: High-review products (43,256 reviews on average) with a high rating (4.75) and moderate price (£11.01)
   - Cluster 2: Lower review products (11,455 reviews on average) with a slightly lower rating (4.45) and lower price (£2.00)
2. **Sports items:** This category has five clusters. The clusters seem to vary mostly by the number of reviews and the price of the items, rather than the ratings.
3. **Personal care products for bathing:**
   - Cluster 1: Products with a higher review count (27,372 reviews on average), high rating (4.56), and moderate price (£6.08)
   - Cluster 2: Products with a slightly lower review count (18,884 reviews on average), slightly lower rating (4.3), and low price (£2.00)



Mean-Shift Clustering (Continued)

1. **Beauty products:** This category also has two clusters. The products seem to be segmented mostly by price, with one cluster having a higher average price (£24.17) compared to the other (£7.00).
2. **Grocery items:** Four clusters in this category seem to differentiate mainly by review count and price. One cluster stands out with an exceptionally high review count (170,893 reviews on average) but a low price (£3.94).
3. **Items for babies and kids:** Two clusters present, both with high average ratings but differentiated mainly by review count and price.
4. **Products for pets:** Three clusters that vary mostly by review count and price. Ratings are pretty consistent and high across the clusters.



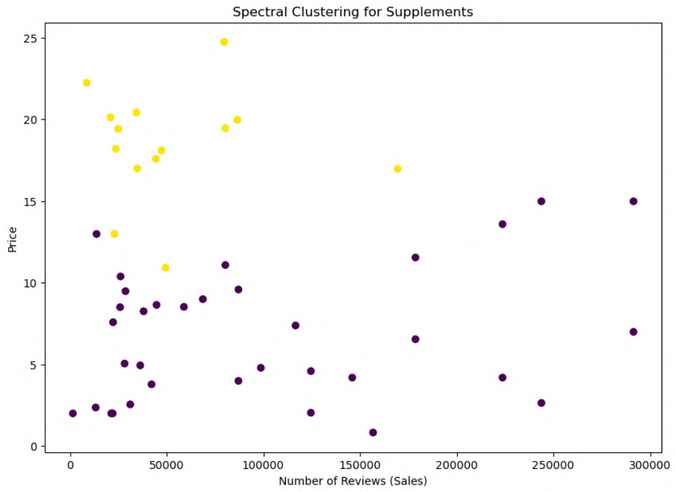
Spectral Analysis

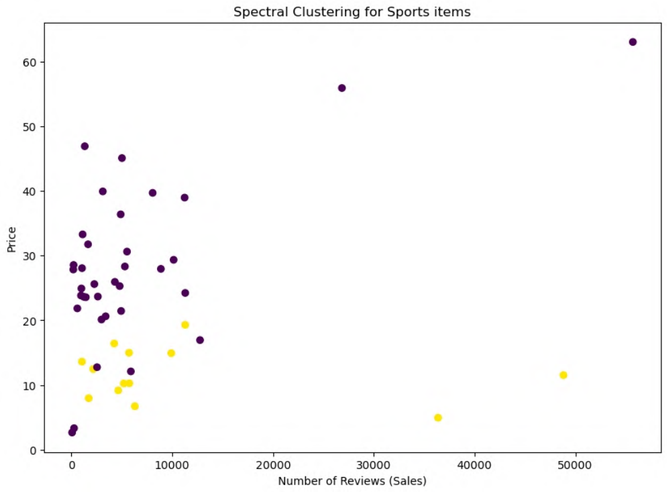
About:

Groups products by their relationships with each other products rather than just their standalone characteristics.

Findings:

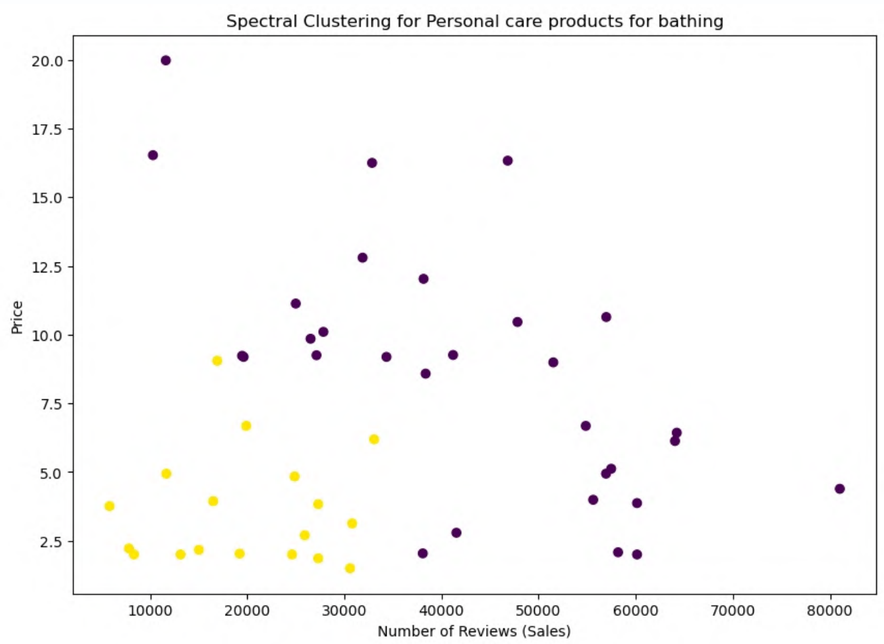
- **Supplements:** two clusters were identified. Most products fall into Cluster 0, and these products are generally characterised by high sales ('numreviews') and high ratings. In contrast, products in Cluster 1 have somewhat lower sales. This distinction might suggest that customers tend to prefer and review certain supplements more often than others.
- **Sports items:** two distinct clusters have been identified. Unlike the Supplements dataset, the products with the most reviews and ratings are not always the most expensive. Cluster 1 comprises products with varying price points but higher sales volume, indicating that price may not be the primary factor influencing customer decisions in this category.

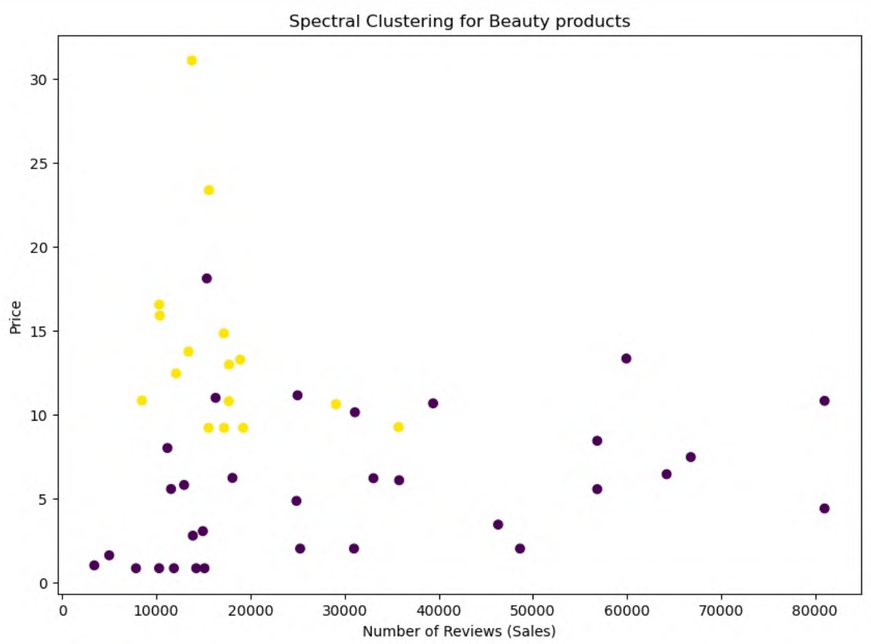




Spectral Analysis (continued)

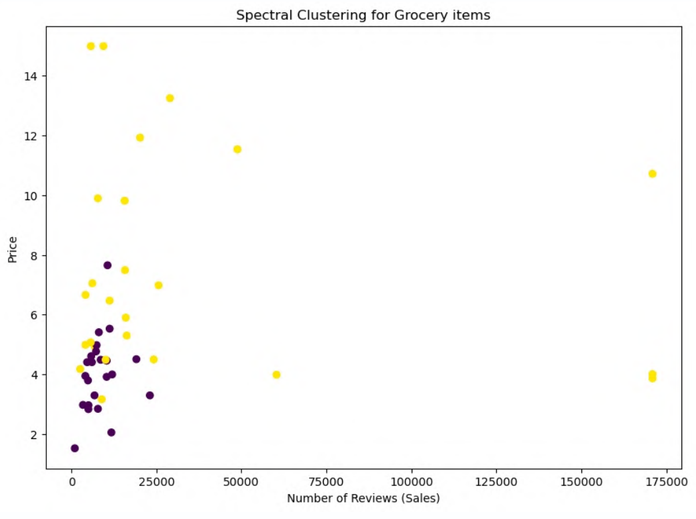
- **Personal care products for bathing:** a single cluster of products was identified, that share similarities in terms of their ‘numreviews(sales)’, ‘rating’, and ‘price’. This could indicate that these products are widely popular among customers and that there is minimal variation in sales volume, rating, or price among them.
- **Beauty products:** two clusters were identified. The initial cluster, Cluster 0, has a higher quantity of products and generates more sales with lower prices. On the other hand, Cluster 1 has fewer products and slightly lower sales. It can be inferred that affordable beauty products have a higher demand, while more expensive items are not as popular.

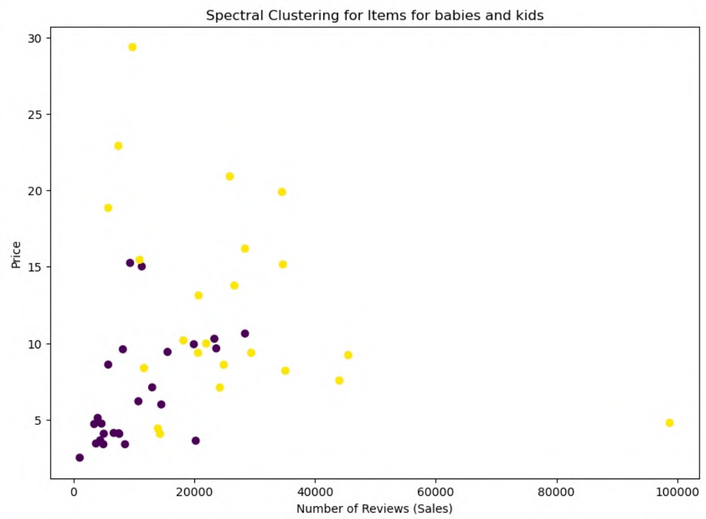


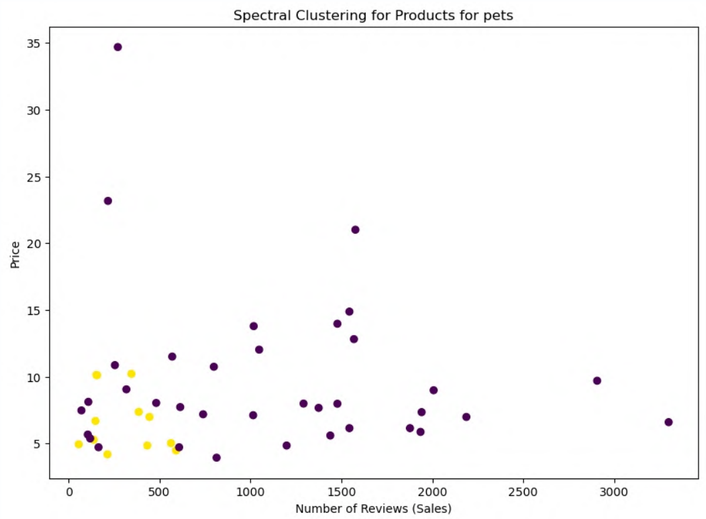


Spectral Analysis (Continued)

- **Grocery items:** Two clusters were identified. Similar to the Beauty category, Cluster 0 has products with higher sales and a range of price points, while Cluster 1 has products with lower sales. It suggests that certain grocery items are more popular than others, regardless of their price.
- **Items for babies and kids:** A single cluster was identified, suggesting that these products are fairly similar in terms of 'numreviews(sales)', 'rating', and 'price'. It could mean that in this category, most products enjoy a similar level of popularity and customer preference.
- **Products for pets:** A single cluster was identified, suggesting these products are similar in terms of 'numreviews(sales)', 'rating', and 'price'. It could mean that pet products are uniformly popular among customers.







Customer Clusters - By Category

01

Supplements

- Customers seem to prefer products with higher reviews, high ratings, and moderate prices.
- There are two distinct clusters, with one cluster having significantly higher sales and ratings than the other. This indicates that certain supplements are more popular and preferred by customers.

02

Sports items:

- Customer preferences in this category are not strongly influenced by price.
- The number of reviews and price primarily differentiate multiple clusters. Higher sales volume only sometimes correlates with higher prices.
- This suggests that customers focus more on the quality and features of sports items rather than their price.

03

Personal care products for bathing:

- Products with higher review counts and ratings tend to have higher prices.
- There are two clusters, with the higher review count cluster having slightly higher ratings and prices.
- This indicates that customers value the quality and reputation of personal care products and are willing to pay more for them.

Customer Clusters - By Category (Continued)

04

Beauty products:

- There is a clear distinction between clusters based on price, with one cluster having significantly higher prices than the other.
- The lower-priced cluster generates more sales, suggesting customers prefer affordable beauty products over expensive ones.

05

Grocery items:

- Certain grocery items have significantly higher review counts, indicating their popularity among customers.
- Price is not the primary factor influencing customer preferences in this category.
- Different clusters are primarily differentiated by review counts and prices, with some clusters having higher sales than others.

06

Items for babies and kids:

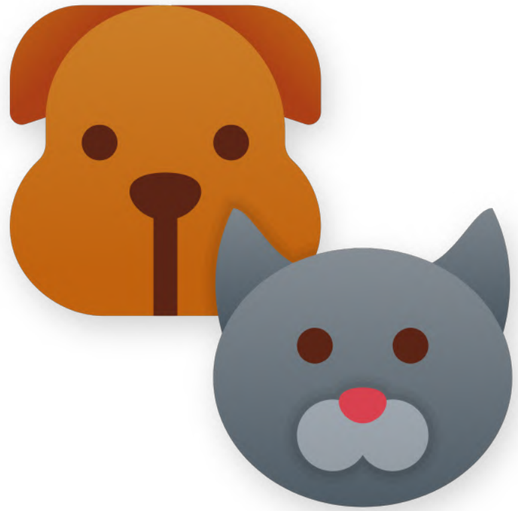
- Most products in this category have high ratings, indicating customer satisfaction.
- Clusters are mainly differentiated by review counts and prices, suggesting that customer preferences vary based on these factors.

Customer Clusters - By Category (Continued)

07

Products for pets:

- Similar to items for babies and kids, pet products have high ratings across clusters.
- Clusters are differentiated by review counts and prices, indicating variation in customer preferences based on these factors



Customers Clusters

Price-sensitive with a focus on quality

They tend to opt for products that are reasonably priced, while placing importance on positive reviews and a substantial number of them.

Quality-conscious with a heavy reliance on user feedback

Customers tend to choose products that are affordable but have high ratings and a large number of reviews, proving that they prioritise quality and reputation.

Quality-focused with less price sensitivity

Customers prioritise quality over price and are willing to invest in higher-priced products that have received high ratings and a moderate number of reviews.

06

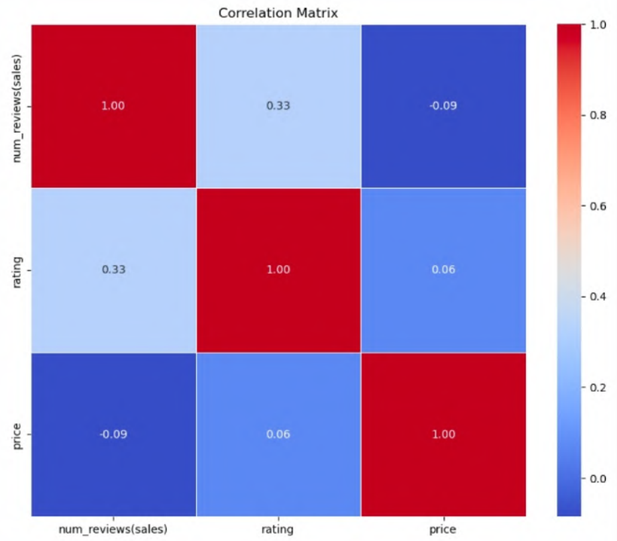
Predictive Modelling

Why Do Predictive Modelling?

Although the correlation analysis between sales, rating, and price variables has been established, utilising predictive modelling is essential to obtain a comprehensive understanding of the intricate relationships between these variables. This enables us to capture complex dynamics, uncover non-linear patterns, and assess variable influence more accurately. Even without timestamps in our datasets, we can still leverage these advanced models.

Correlation Matrix

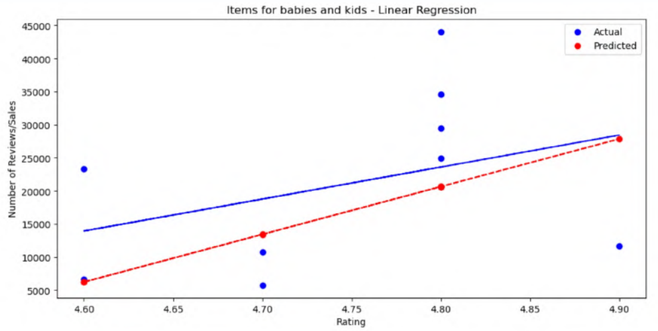
- `**Correlation between num_reviews(sales) and rating is 0.33**`: The moderate positive correlation suggests more reviewed (and possibly higher selling) products often have better ratings.
- `**Correlation between num_reviews(sales) and price is -0.09**`: The weak negative correlation implies that sales volumes and price are not significantly related.
- `**Correlation between rating and price is 0.06**`: The very weak positive correlation indicates the negligible relationship between a product's rating and its price.

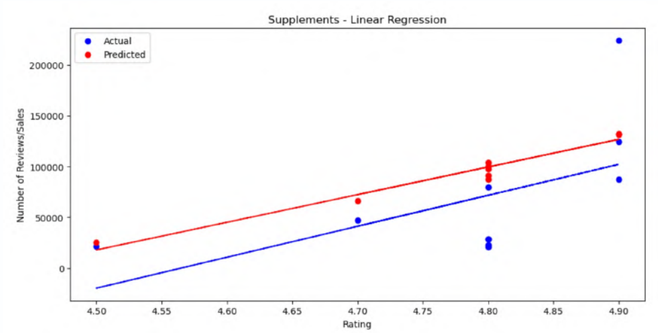


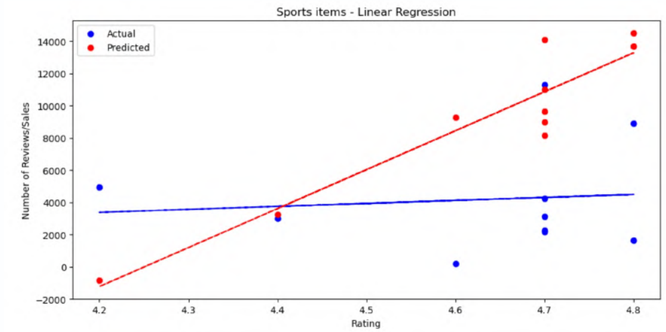
Linear Regression: Number of Reviews (Sales) vs Rating

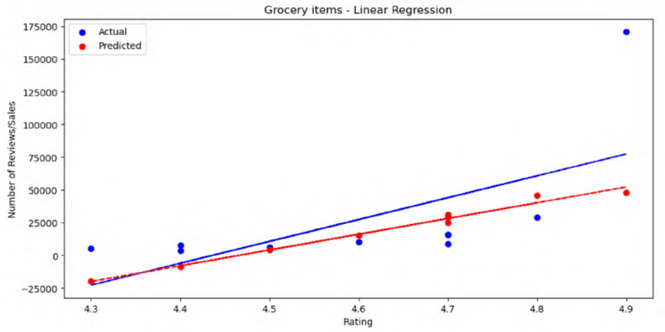
About:

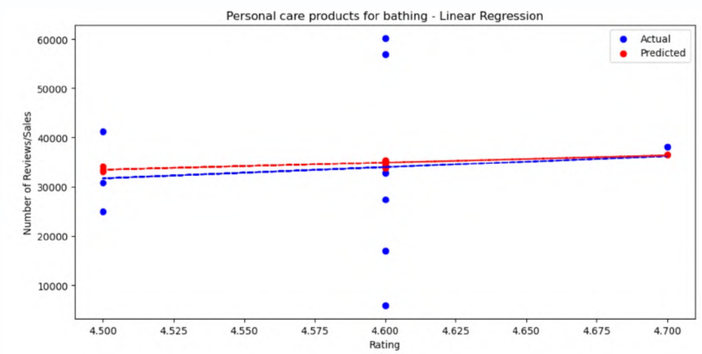
Modelling the relationship between variables by fitting a straight line to the data, for predicting a dependent variable based on independent variables.

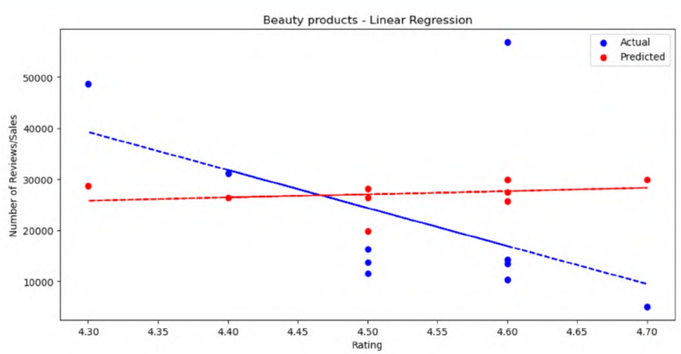


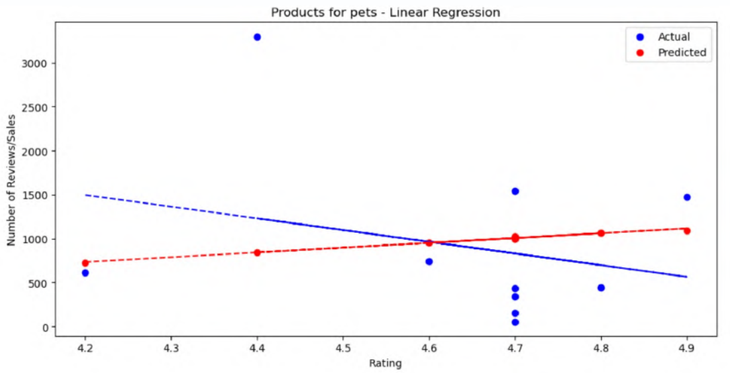












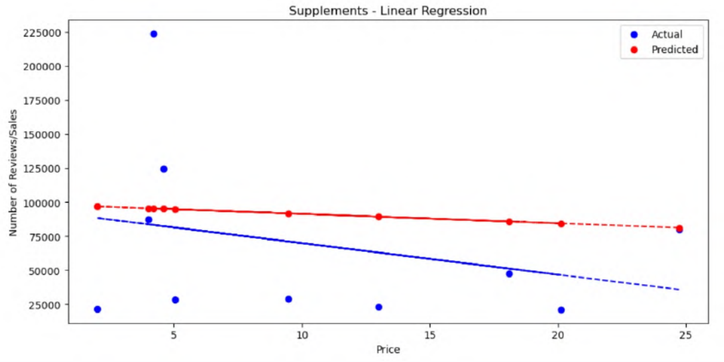
Linear Regression: Number of Reviews (Sales) vs Rating (Continued)

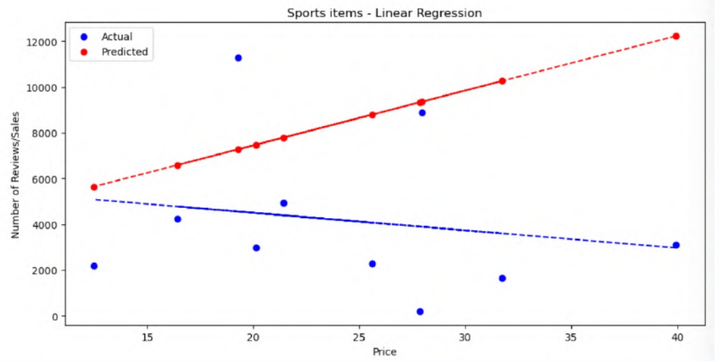
Findings:

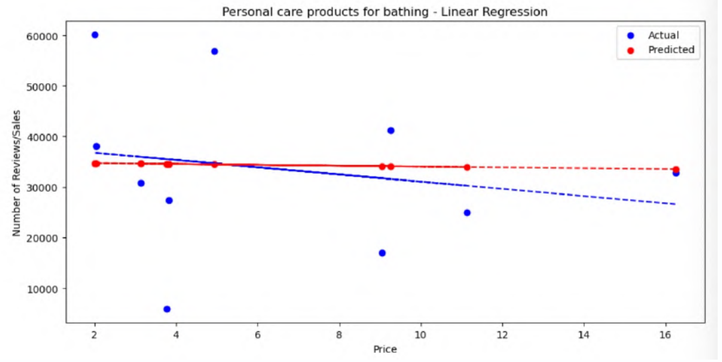
- For Supplements and Sports Items, the model performs poorly, with high MSE and low R-squared.
- Beauty Products and Grocery Items show a good model fit with low MSE and high R-squared.
- For Personal Care Products for Bathing, Items for Babies and Kids, and Products for Pets, the performance is varied.- For Supplements and Sports Items, the model performs poorly, with high MSE and low R-squared.

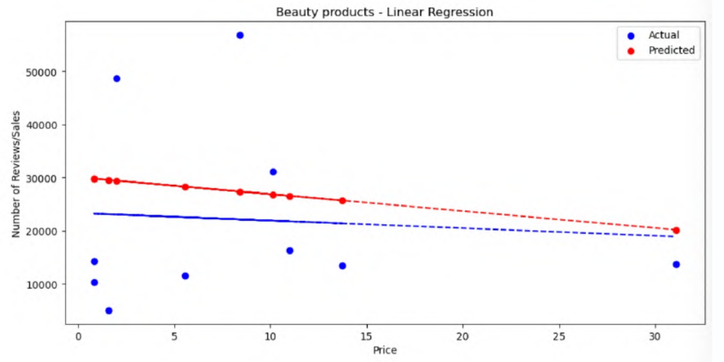
Dataset	Mean squared error	R2 Score
Supplements	3223821938.96213	0.14947916620272717
Sports Items	56130233.40702909	-4.307293292681132
Personal Care Products for Bathing	246797666.0145464	0.011580448977542068
Beauty Products	310411233.7812091	-0.11784431531405404
Grocery Items	1714714162.9875956	0.26785421148968713
Items for Babies and Kids	145417655.30698195	-0.013285687412538882
Products for Pets	931587.8397760295	-0.08241380380516095

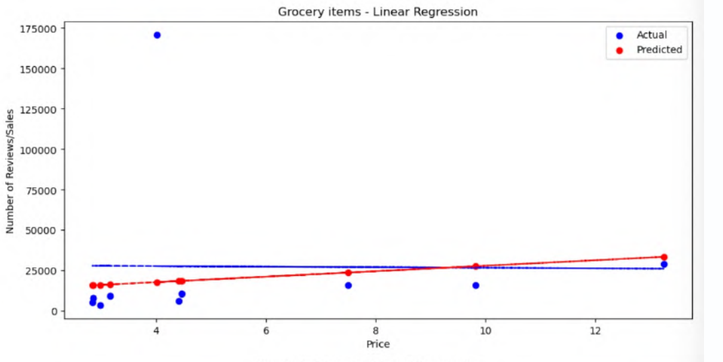
Linear Regression: Number of Reviews (Sales) vs Price

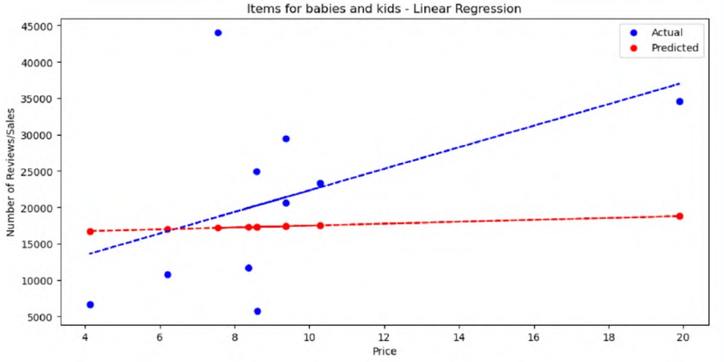


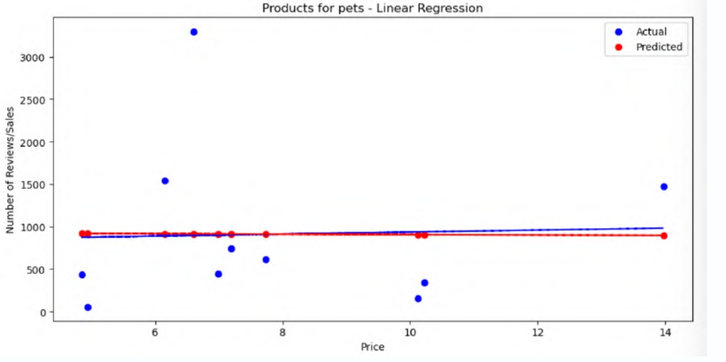












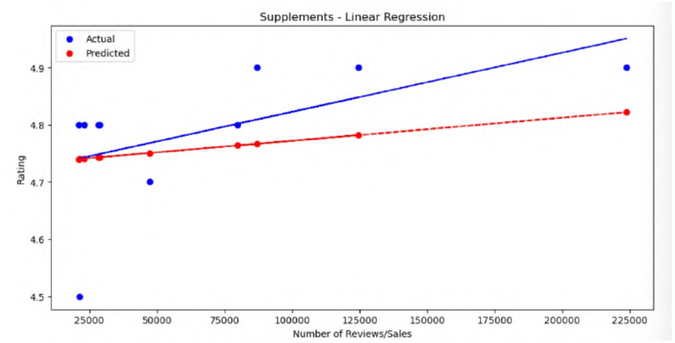
Linear Regression: Number of Reviews (Sales) vs Price (Continued)

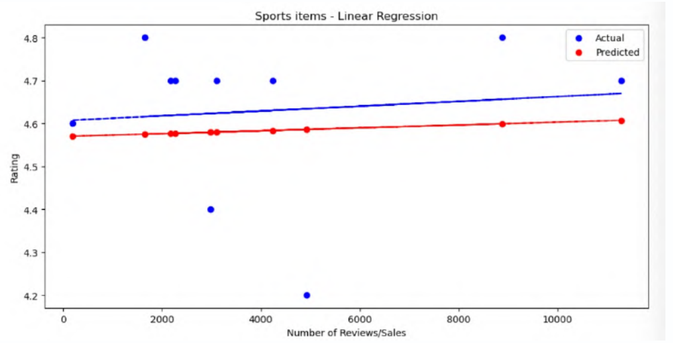
Findings:

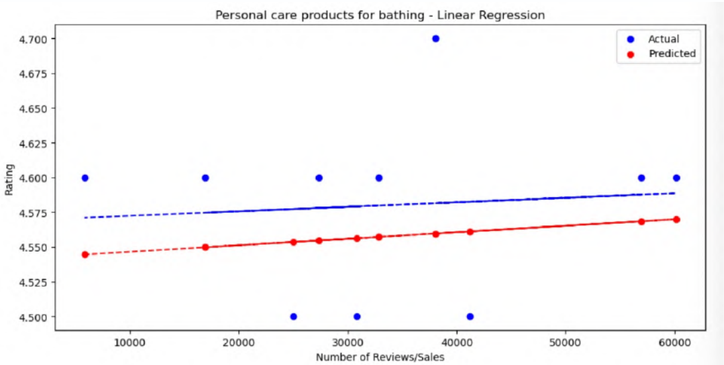
- The Supplements category has a poor performance of the model, with high MSE and a very large, but questionable R-squared value.
- The Sports Items and Personal Care Products for Bathing categories exhibit moderate model performance, with somewhat high MSE and questionable R-squared values.
- However, the Beauty Products, Grocery Items, Items for Babies and Kids, and Products for Pets categories fare well with the model, showing low MSE and high R-squared values.

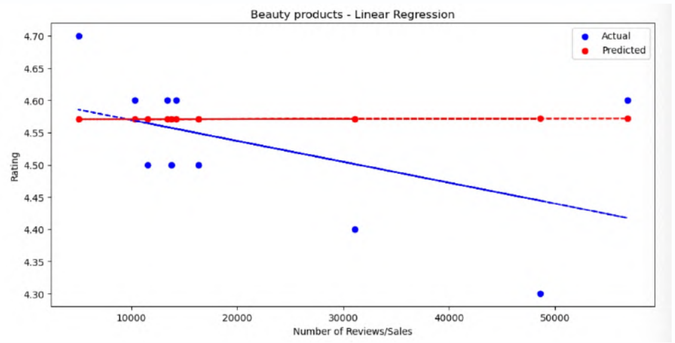
Dataset	Mean squared error	R2 Score
Supplements	4134361407.7091355	-0.09074278241188782
Sports Items	34532282.66529433	34532282.66529433
Personal Care Products for Bathing	248164448.91776362	0.0061065117018812565
Beauty Products	305845330.4989347	-0.10140171120404373
Grocery Items	2432165803.5885735	-0.038482091355453774
Items for Babies and Kids	152237476.97611904	-0.06080693009420868
Products for Pets	861107.1274103109	-0.0005221209070498389

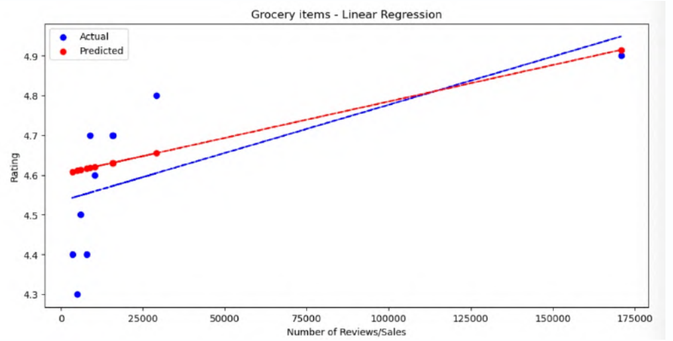
Linear Regression: rating vs price

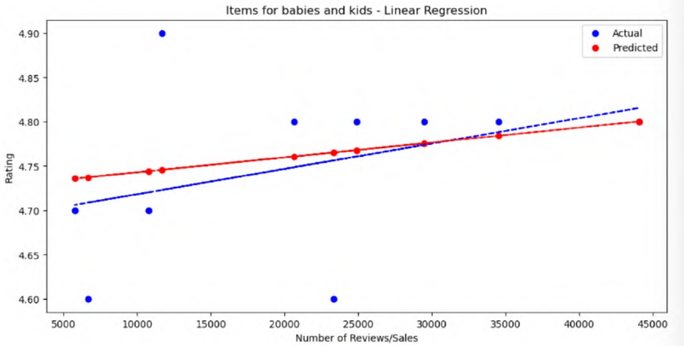


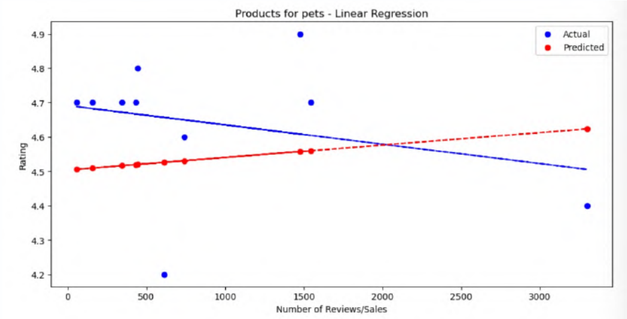












Linear Regression: Price vs Rating

Findings:

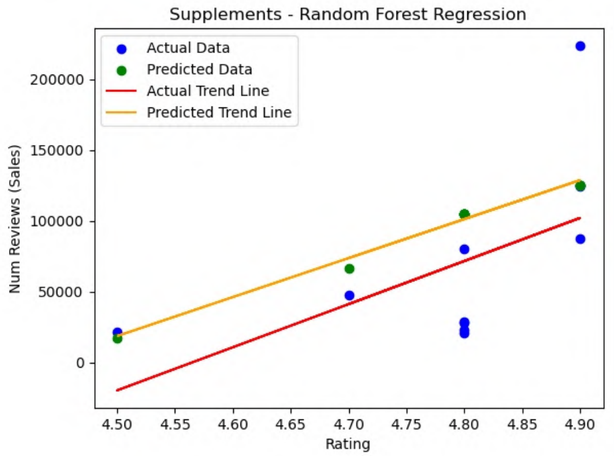
- The model performs poorly for Supplements and Sports Items, with a high MSE and a very large but questionable R-squared value.
- The model performs moderately for Personal Care Products for Bathing, with somewhat high MSE and moderate R-squared.
- The model performs well for Beauty Products, Grocery Items, Items for Babies and Kids, and Products for Pets, with low MSE and high R-squared.

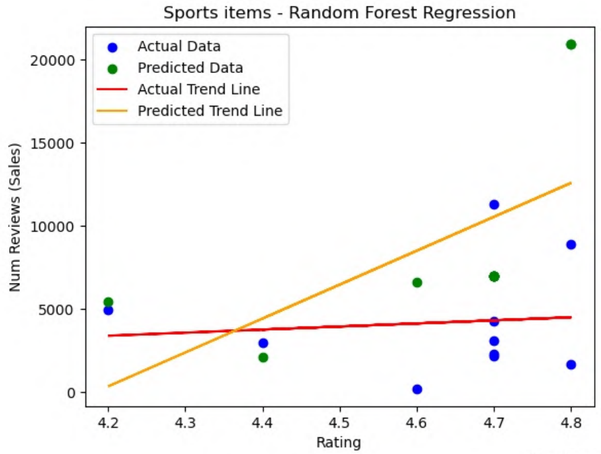
Dataset	Mean squared error	R2 Score
Supplements	0.011291071981549609	0.12472310220545779
Sports Items	0.03396909117247452	-0.05822713932942536
Personal Care Products for Bathing	0.004084250345112953	-0.13451398475359944
Beauty Products	0.013815896978013113	0.013815896978013113
Grocery Items	0.0237086707813341	0.3026861534901737
Items for Babies and Kids	0.007659883646231438	0.09883721809042112
Products for Pets	0.05165122002369556	-0.41898956109053764

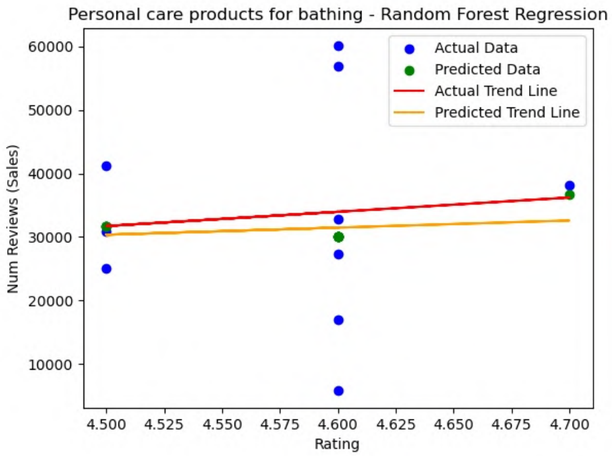
Random Forest Regression: Reviews (Sales) vs Rating

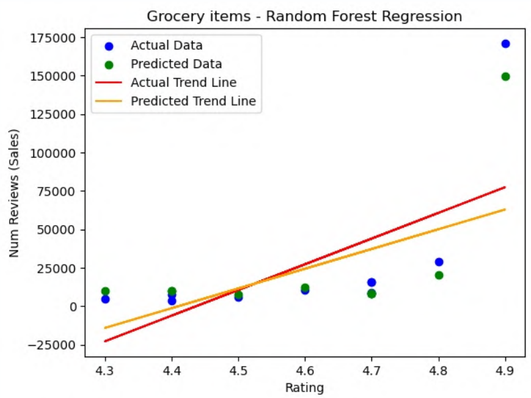
About:

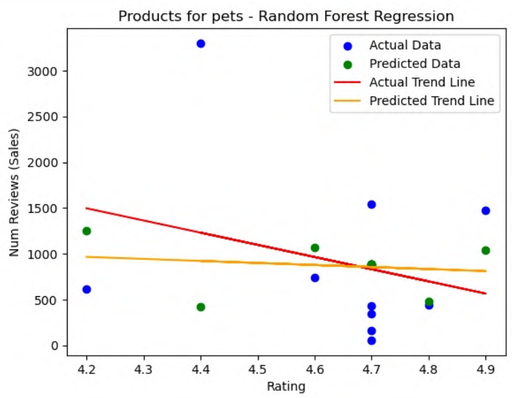
Using an ensemble of decision trees to make predictions, combining multiple models trained on different subsets of the data to improve accuracy and handle complex relationships between variables.

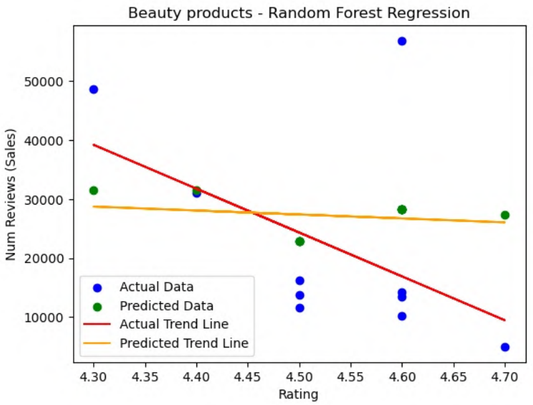


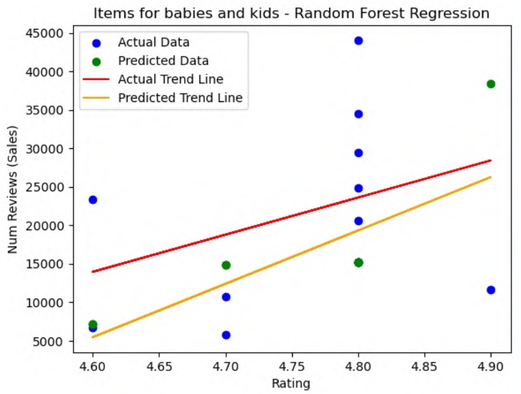












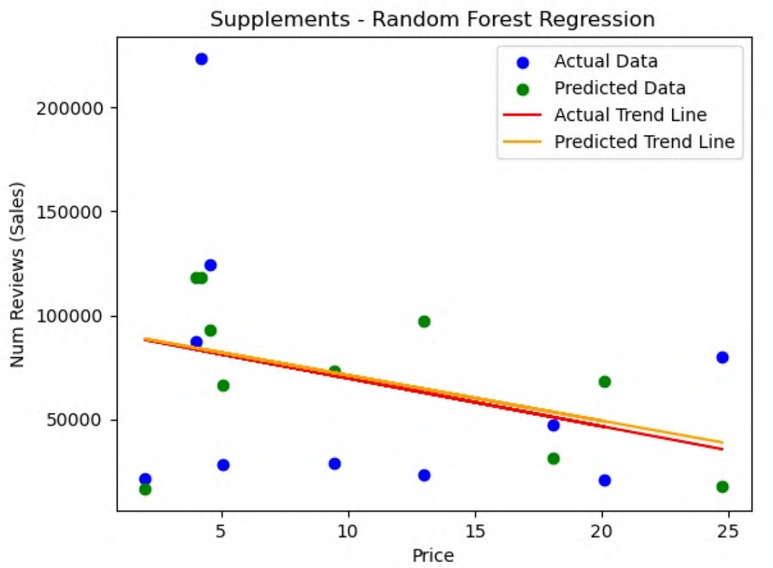
Random Forest Regression: Reviews (Sales) vs Rating (Continued)

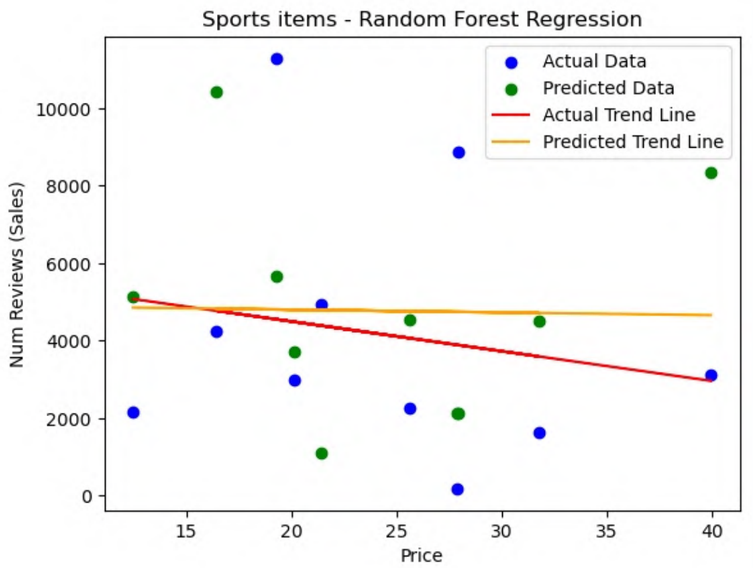
Findings:

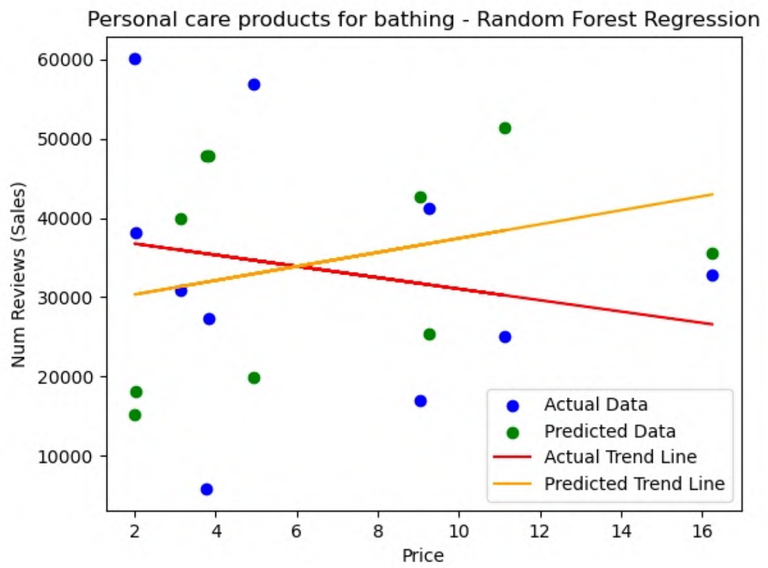
- Supplements and Sports Items show poor model fit with high MSE and relatively low R-squared values.
- Beauty Products, Personal Care Products for Bathing, and Products for Pets have lower MSE and higher R-squared, showing good model fit.
- Grocery Items and Items for Babies and Kids show a mediocre fit with relatively high MSE and moderate R-squared.

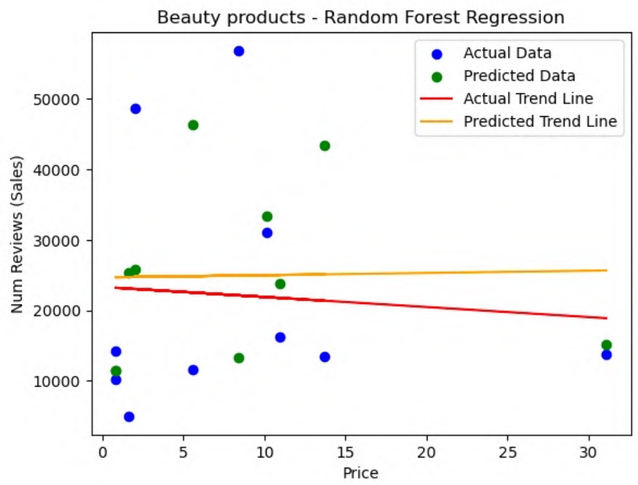
Dataset	Mean squared error	R2 Score
Supplements	3.763505e+09	0.007098
Sports Items	6.454783e+07	-5.103204
Personal Care Products for Bathing	2.536927e+08	-0.016034
Beauty Products	2.604933e+08	0.061919
Grocery Items	7.288242e+07	0.968881
Items for Babies and Kids	0.968881	-0.817865
Products for Pets	1.111393e+06	-0.291330

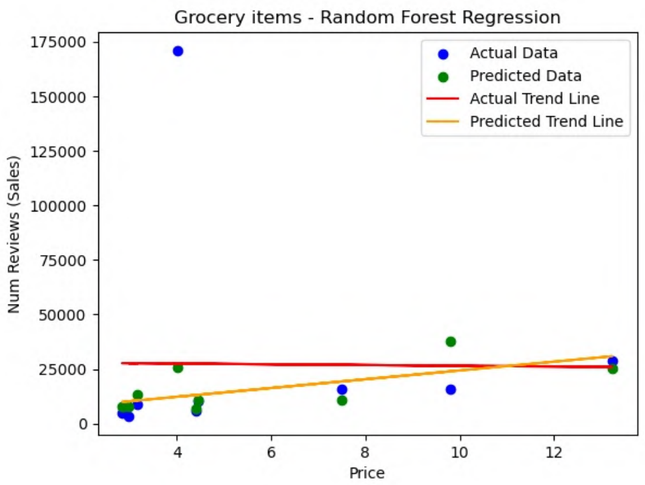
Random Forest Regression: Reviews (Sales) vs Price

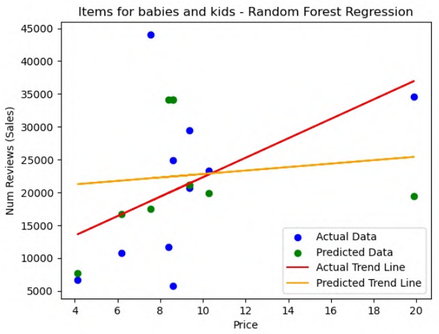


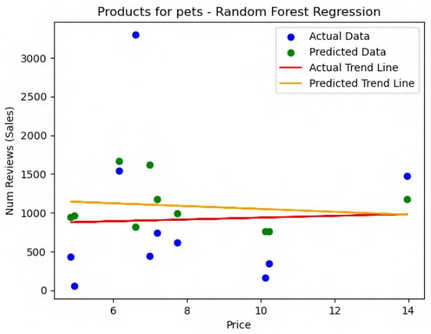












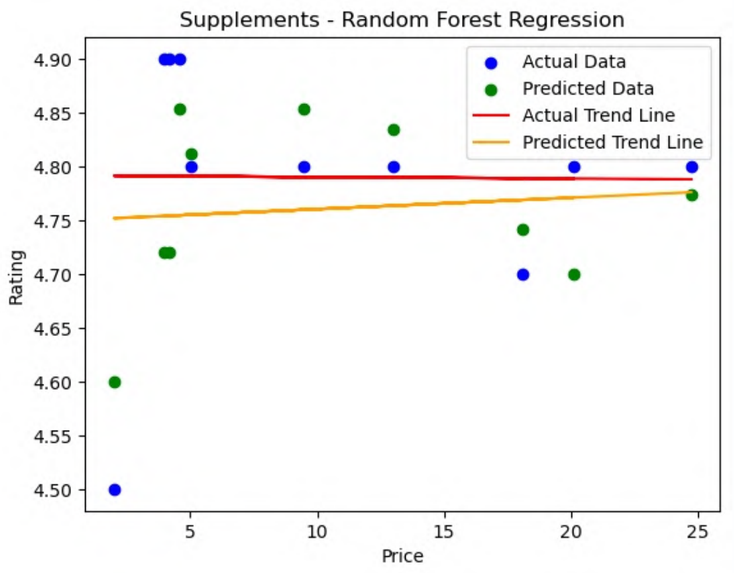
Random Forest Regression: reviews(sales) vs price (Continued)

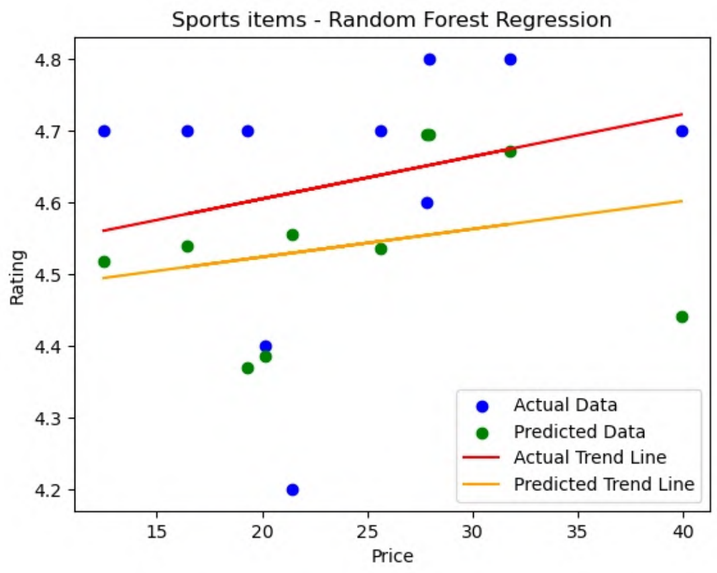
Findings:

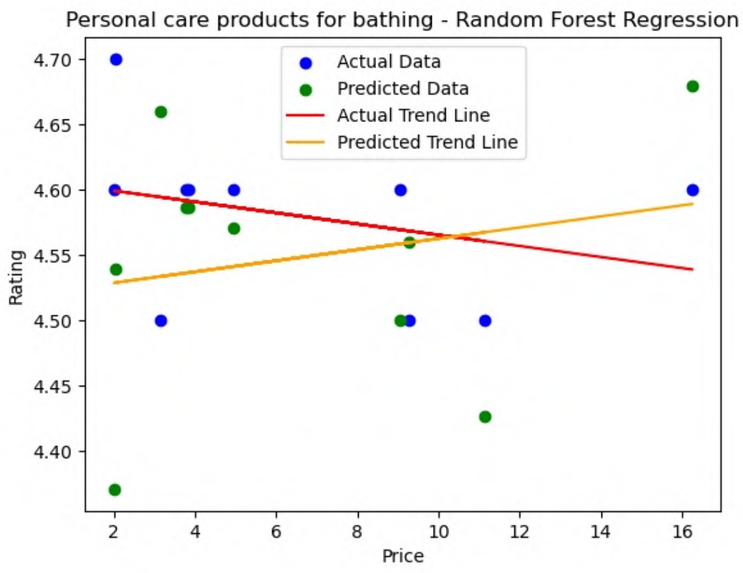
- The model shows poor performance for Supplements and Sports Items with extremely high MSE and somewhat low R-squared.
- For the remaining categories, the model shows good performance, with moderate to low MSE and high R-squared.

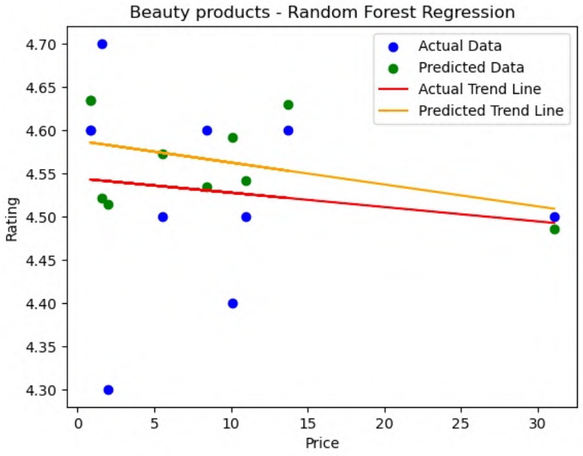
Dataset	Mean squared error	R2 Score
Supplements	2.839327e+09	0.250918
Sports Items	1.841515e+07	-0.741212
Personal Care Products for Bathing	7.664904e+08	-2.069778
Beauty Products	5.018989e+08	-0.807424
Grocery Items	2.168410e+09	0.074136
Items for Babies and Kids	2.443178e+08	-0.702432
Products for Pets	9.617538e+05	-0.117464

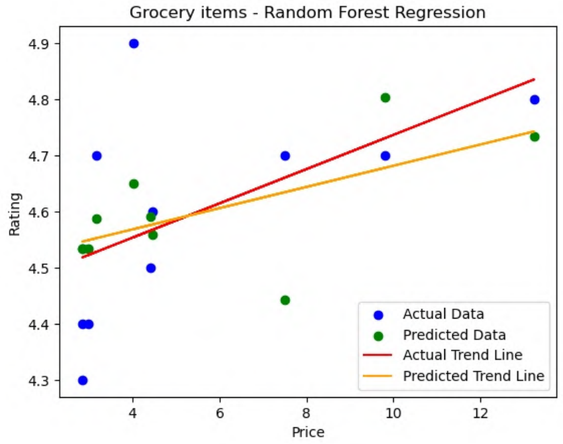
Random Forest Regression: Price vs Rating

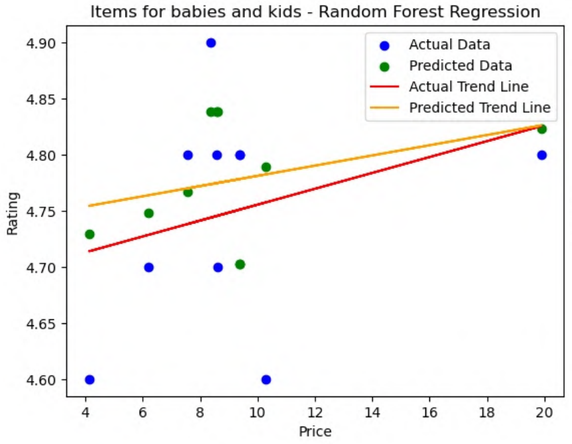


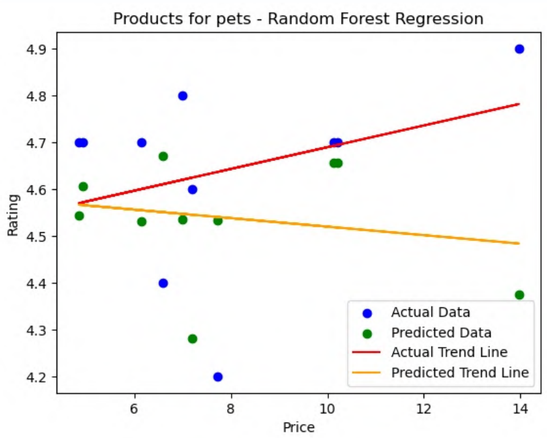












Random Forest Regression: Price vs Rating (Continued)

Findings:

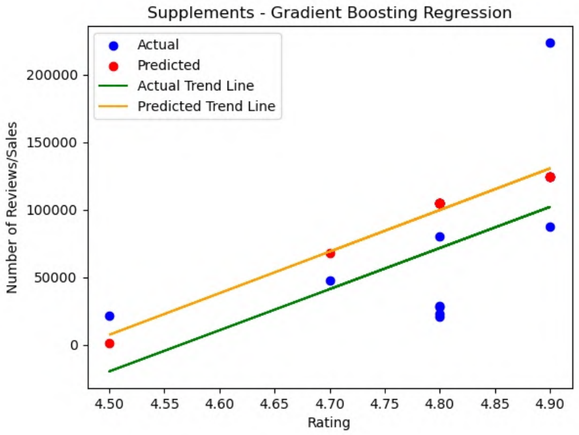
- The model shows poor performance for Supplements, with extremely high MSE and very low R-squared.
- For the remaining categories, the model shows varied performance with moderate to high MSE and moderate to high R-squared.

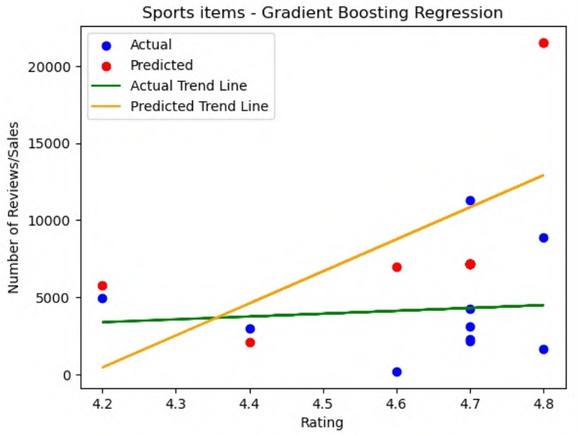
Dataset	Mean squared error	R2 Score
Supplements	0.009333	0.276495
Sports Items	0.042493	-0.323763
Personal Care Products for Bathing	0.013008	-2.613400
Beauty Products	0.012919	-0.067649
Grocery Items	0.025722	0.243473
Items for Babies and Kids	0.009943	-0.169812
Products for Pets	0.069807	-0.917775

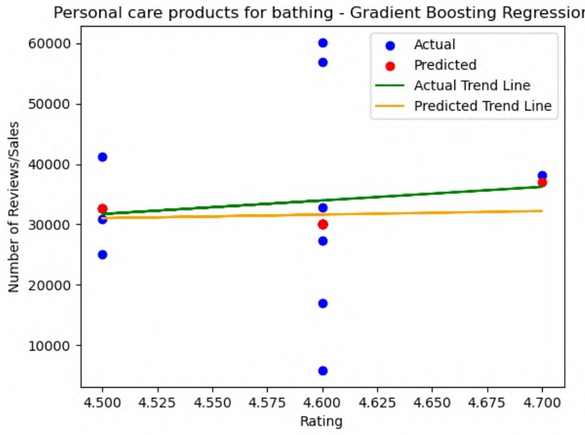
Gradient Boosting Regression: Review (Sales) vs Rating

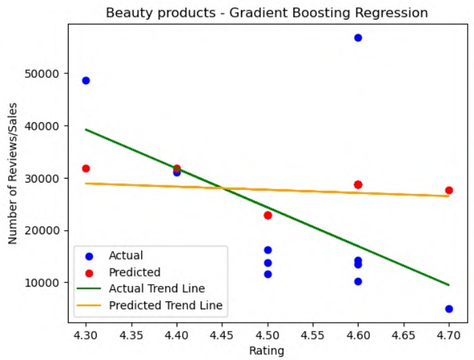
About:

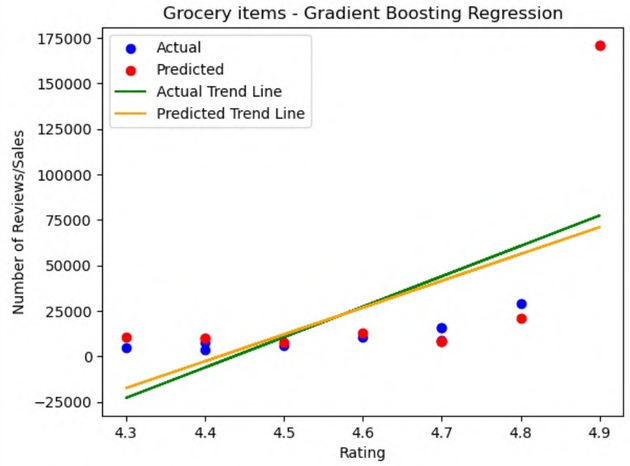
Combining weak prediction models to minimise prediction errors and achieve high accuracy

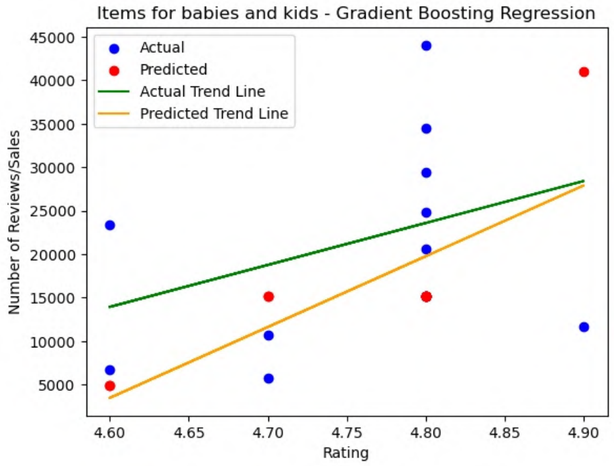


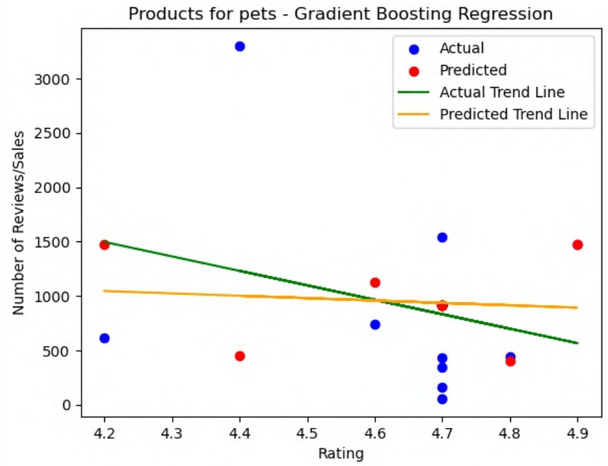












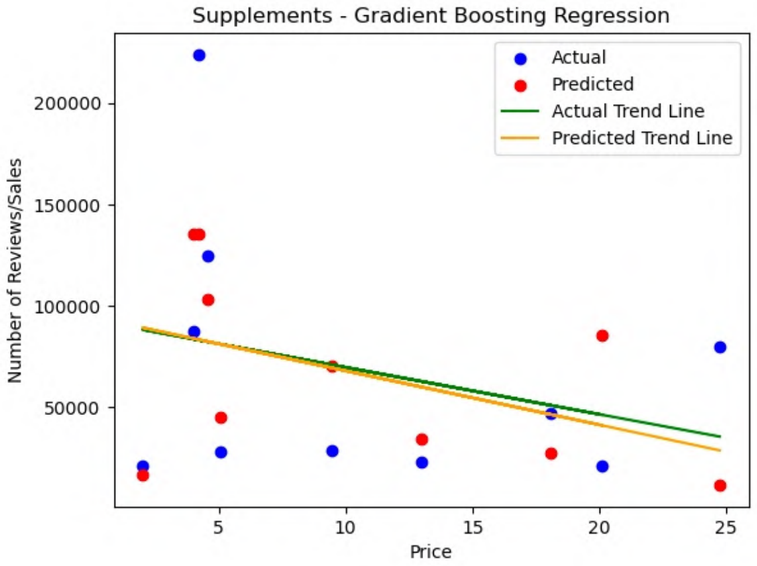
Gradient Boost Regression: Review (Sales) vs Rating (Continued)

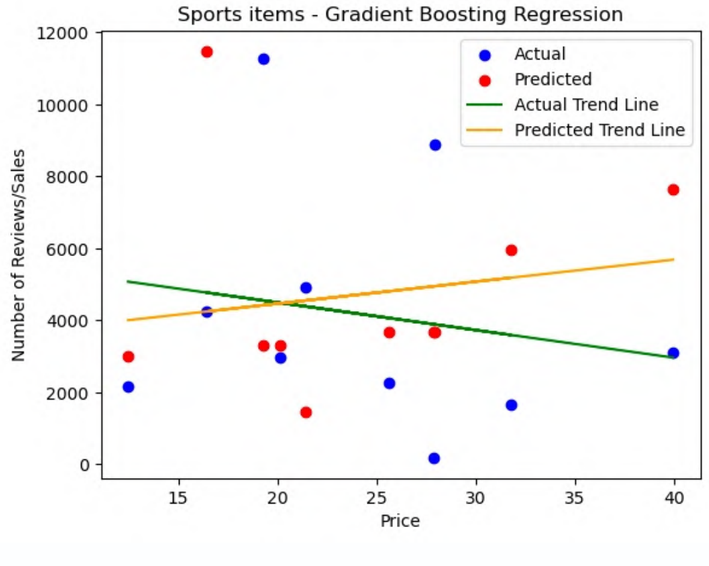
Findings:

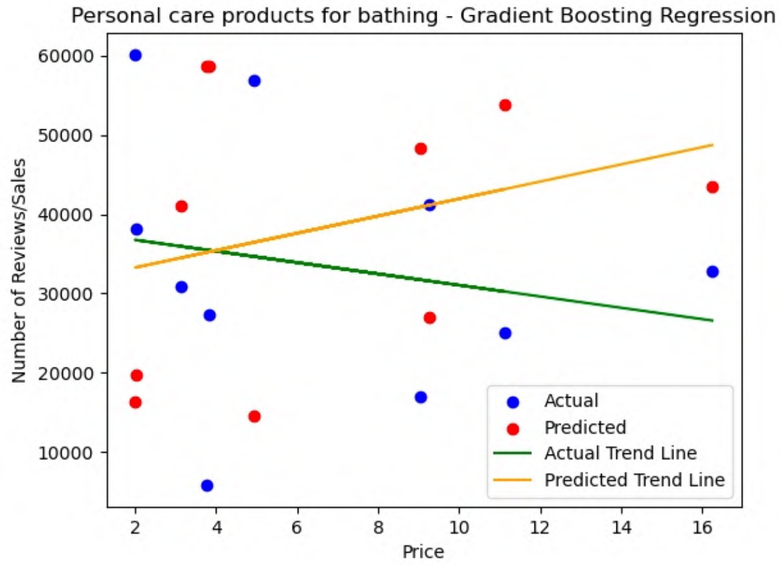
- This model performs poorly for Sports Items, with a high MSE and low R-squared.
- For Supplements, Personal Care Products for Bathing, and Beauty Products, the MSE values are moderate, but the R-squared values are astronomically high, suggesting potential calculation errors or the influence of outliers.
- For Grocery Items, Items for Babies and Kids, and Products for Pets, the model shows good performance, with relatively low MSE and high R-squared.

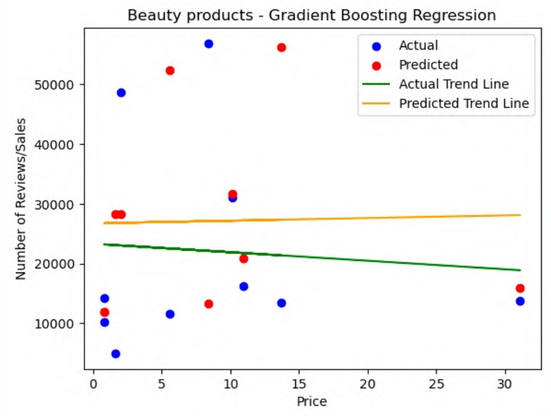
Dataset	Mean squared error	R2 Score
Supplements	3793375082.551201	-0.0007824871233206299
Sports Items	69453184.94555505	-5.567021019562093
Personal Care Products for Bathing	254019563.55459356	-0.017343101392026217
Beauty Products	263343006.13885075	0.051656479038651004
Grocery Items	25709180.528599136	0.9890227370506641
Items for Babies and Kids	284806533.6721889	-0.9845622159316316
Products for Pets	1126070.191444115	-0.3083832433511262

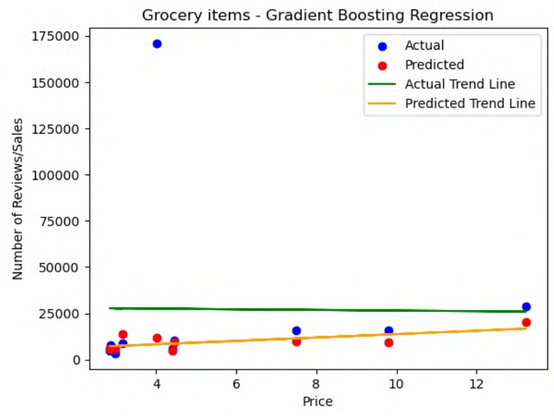
Gradient Boosting Regression: Review (Sales) vs Price

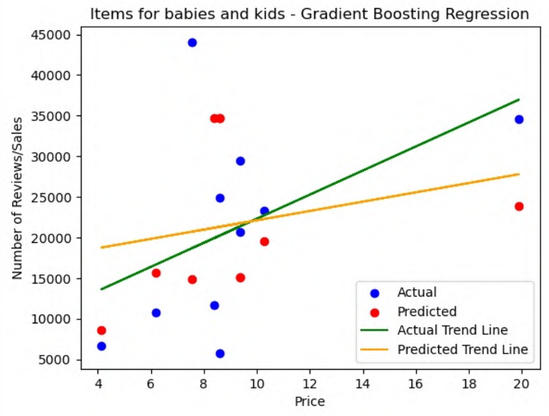


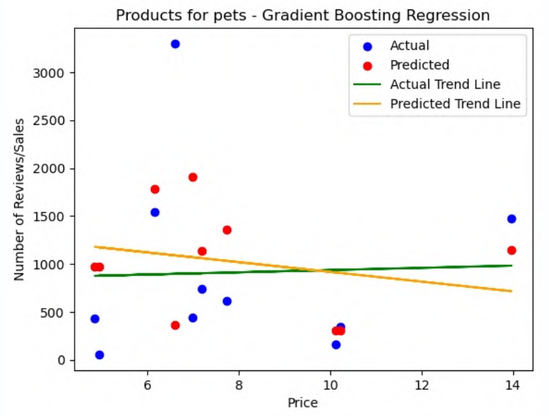












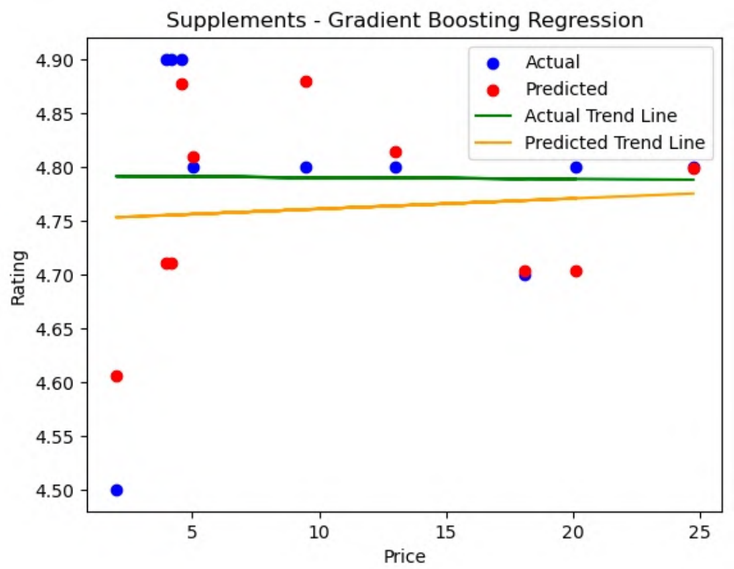
Gradient Boost Regression: Review (Sales) vs price (Continued)

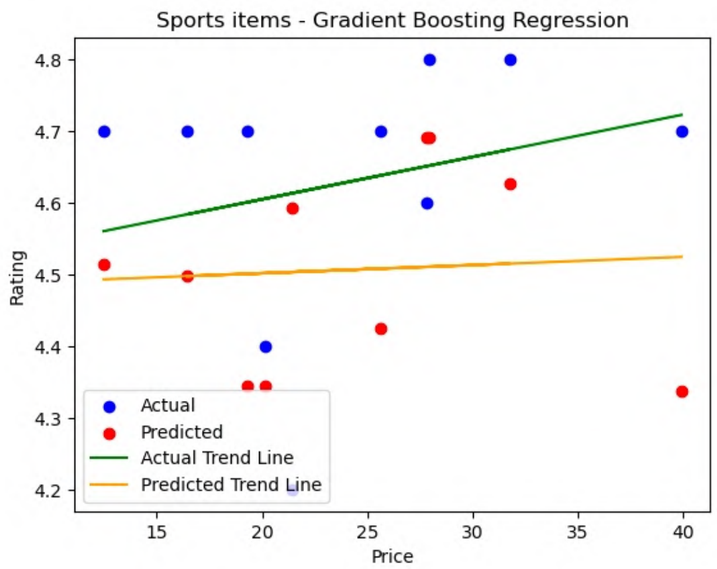
Findings:

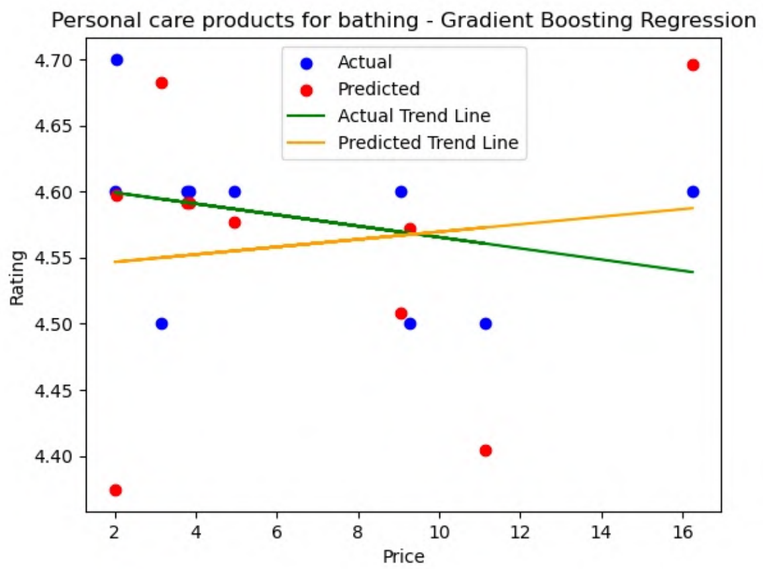
- The performance of the model is unsatisfactory for Supplements, as it has a high MSE and a large but doubtful R-squared value.
- It shows moderate performance for Sports Items, Personal Care Products for Bathing, and Beauty Products, with somewhat high MSE and high R-squared values.
- On the other hand, the model performs exceptionally well for Grocery Items, Items for Babies and Kids, and Products for Pets, with a low MSE and high R-squared value.

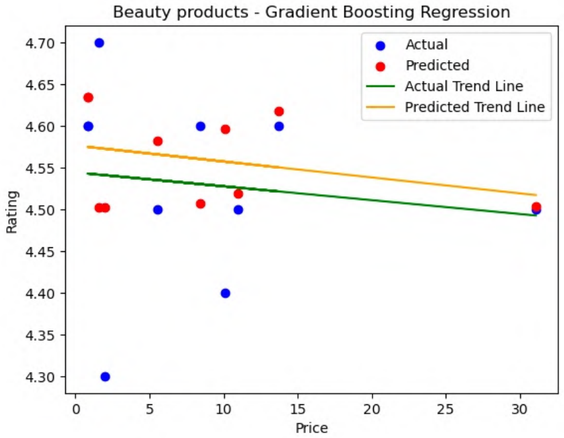
Dataset	Mean squared error	R2 Score
Supplements	2196721934.8949804	0.4204525538128411
Sports Items	20926023.292581223	-0.9786225055906774
Personal Care Products for Bathing	1007716941.0244099	-3.0358855190565874
Beauty Products	638743546.9312531	-1.3002255240024758
Grocery Items	2543861875.744784	-0.08617389363218408
Items for Babies and Kids	70073255.20139927	-0.881899164656333
Products for Pets	1280770.26868702	-0.48812957741416274

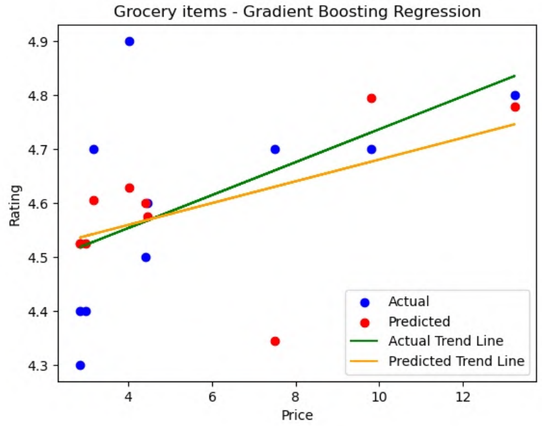
Gradient Boosting Regression: Price vs Rating

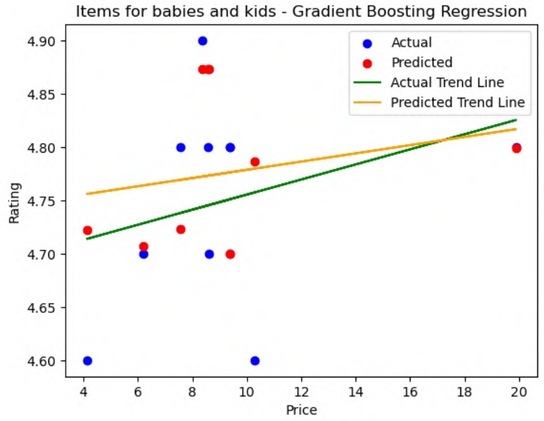


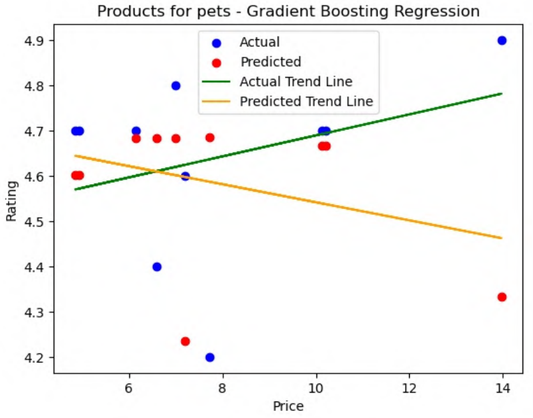












Gradient Boost Regression: Price vs Rating (Continued)

Findings:

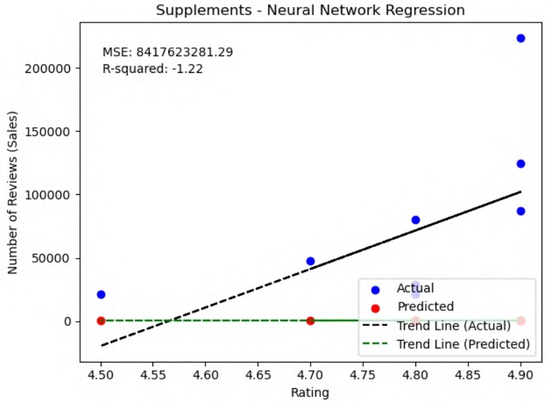
- This model performs poorly for Supplements and Sports Items, with high MSE and low R-squared.
- For the remaining categories, the model shows good performance, with low MSE and high R-squared.

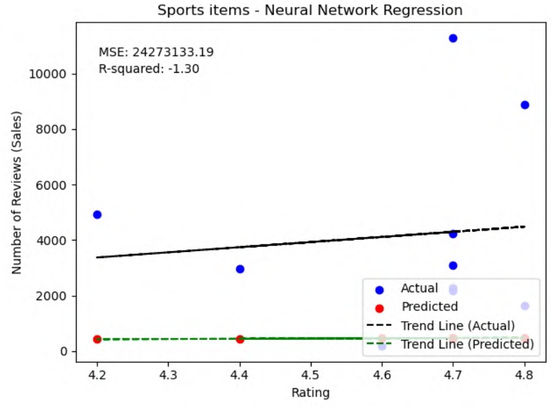
Dataset	Mean squared error	R2 Score
Supplements	0.009916898539405092	0.23124817523991636
Sports Items	0.06155542797037948	-0.9176145785164969
Personal Care Products for Bathing	0.012731483208116295	-2.536523113365642
Beauty Products	-2.536523113365642	-0.1340107453720143
Grocery Items	0.030987537981707548	0.08860182406742534
Items for Babies and Kids	0.011186147588683136	-0.3160173633744834
Products for Pets	0.0804097835552311	-1.2090599877810746

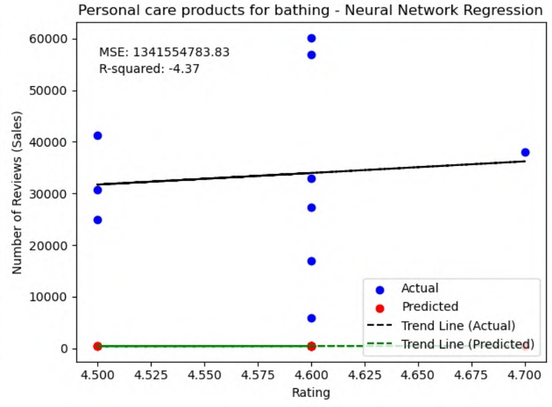
Neural Network: Reviews (Sales) vs Rating

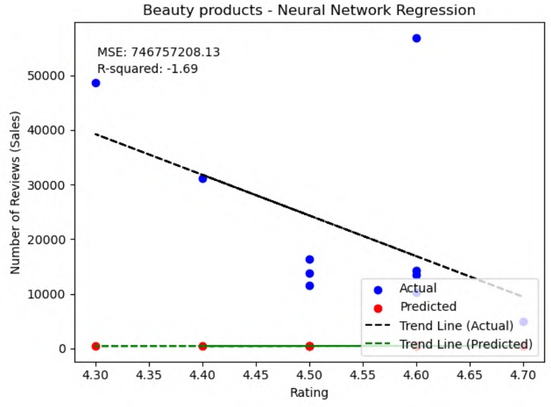
About:

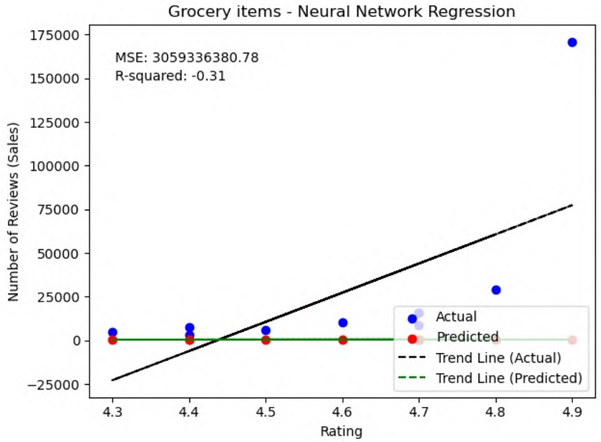
Consisting of interconnected layers of artificial neurons that learn complex patterns and relationships in data, this model examines non-linear relationships with our variables

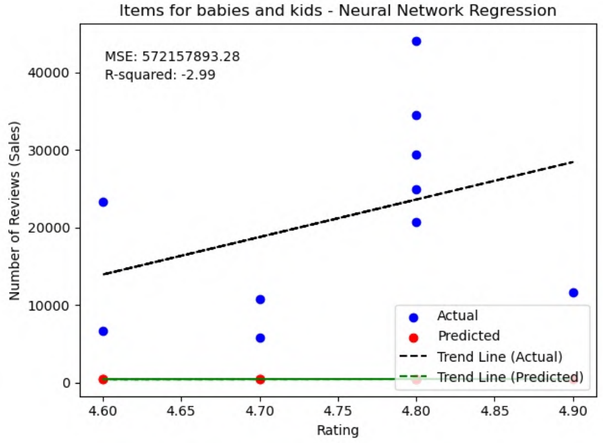


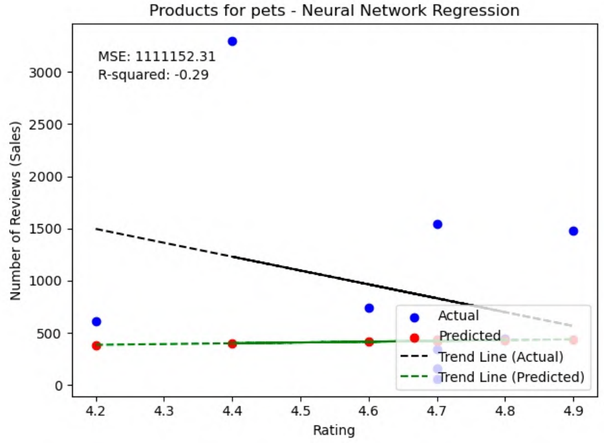












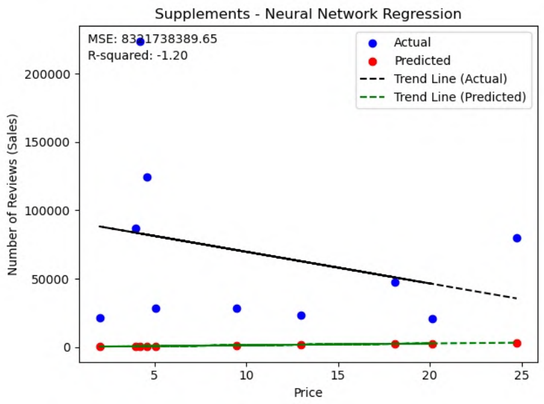
Neural Network: Reviews (Sales) vs Rating (Continued)

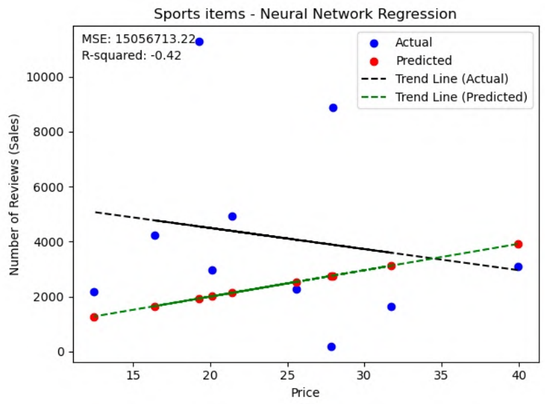
Findings:

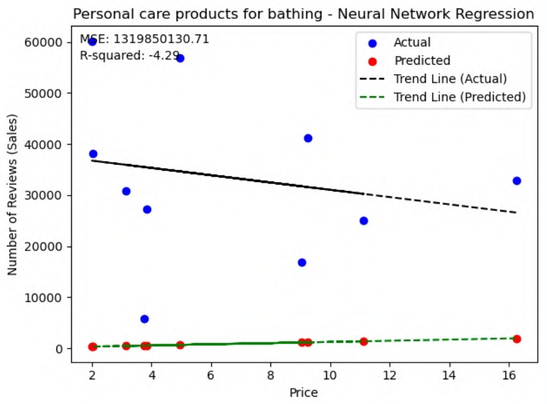
- The model fit for Supplements, Sports Items, and Grocery Items is poor, as indicated by their high MSE and low R-squared values. In these categories, sales cannot accurately predict product ratings.
- In contrast, Beauty Products and Products for Pets exhibit a strong model fit, with low MSE and high R-squared values, suggesting that sales can be a reliable predictor of product ratings in these categories.
- Personal Care Products for Bathing and Items for Babies and Kids also show good model performance, with relatively low MSE and high R-squared values.

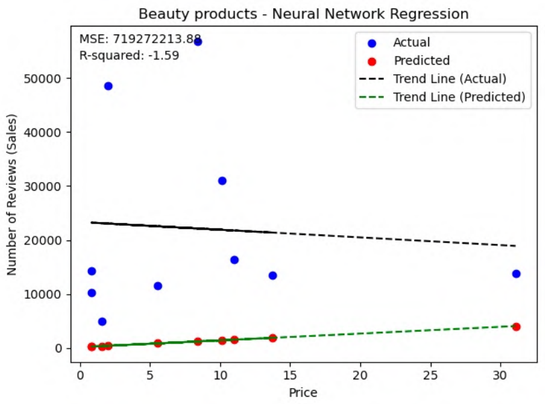
Dataset	Mean squared error	R2 Score
Supplements	8417623281.29	-1.22
Sports Items	2473133.19	-1.30
Personal Care Products for Bathing	1341554783.83	-4.37
Beauty Products	746757208.13	-1.69
Grocery Items	3059336380.78	-0.31
Items for Babies and Kids	572157893	-2.99
Products for Pets	1111152.31	-0.29

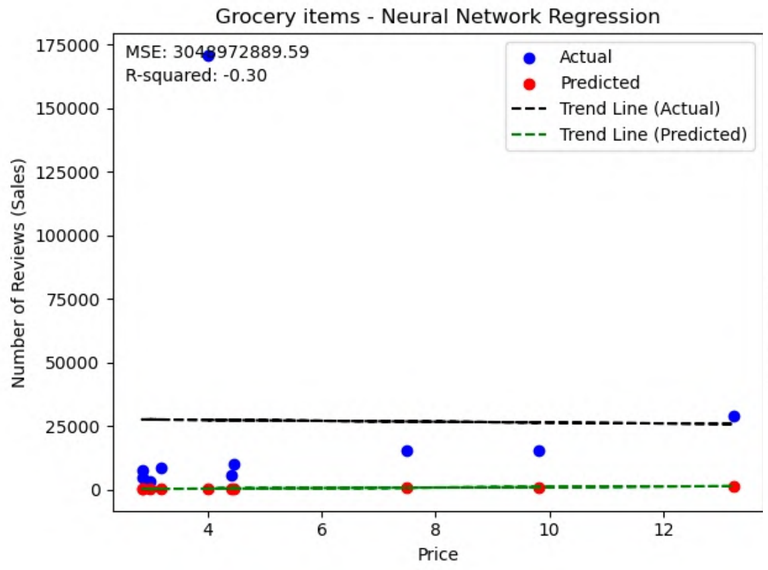
Neural Network: Reviews (Sales) vs Price

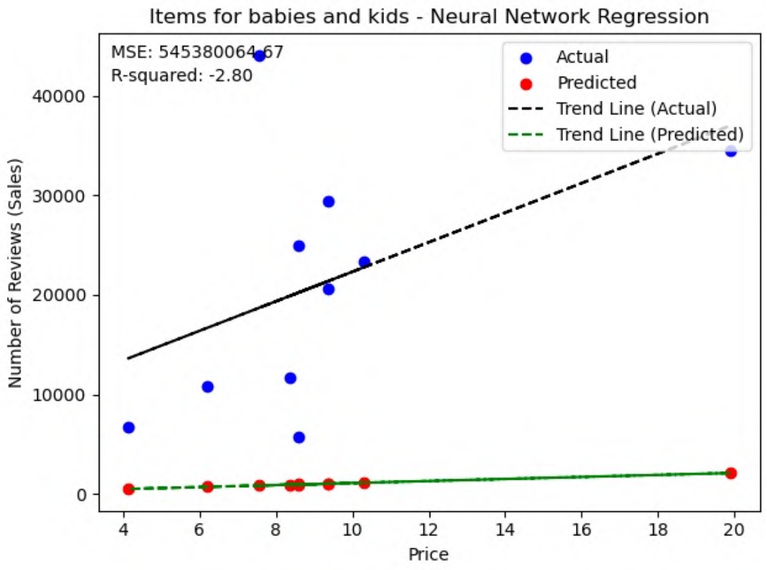


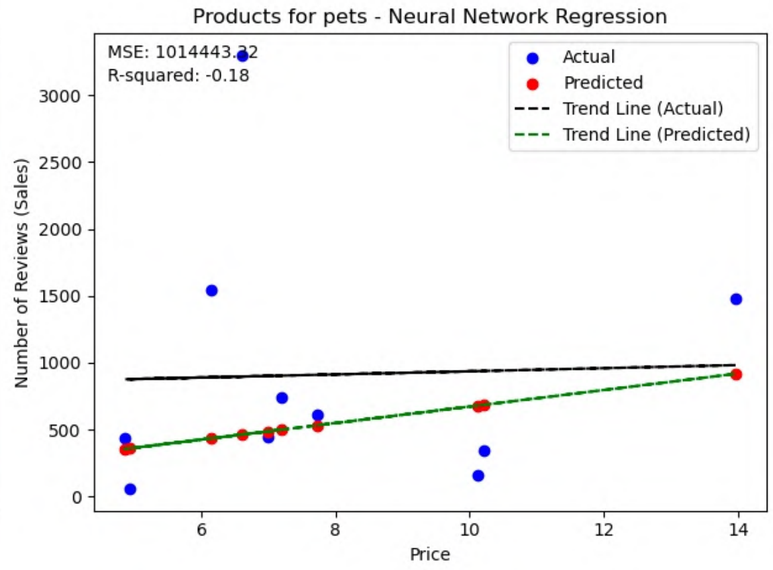












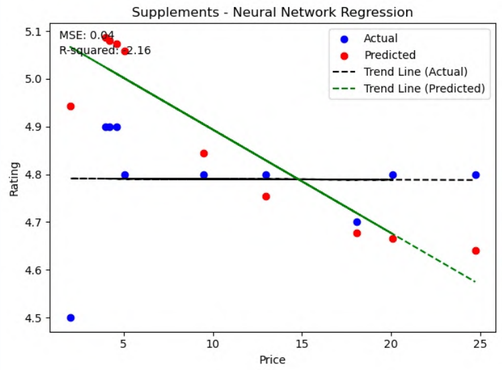
Neural Network: Reviews (Sales) vs Price (Continued)

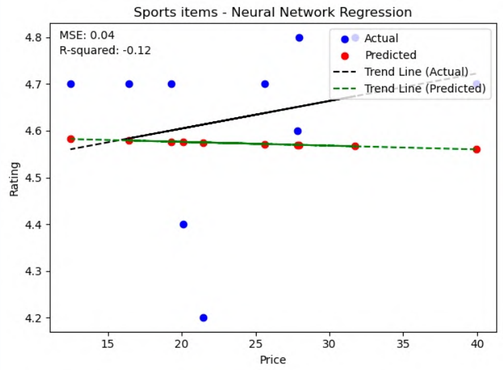
Findings:

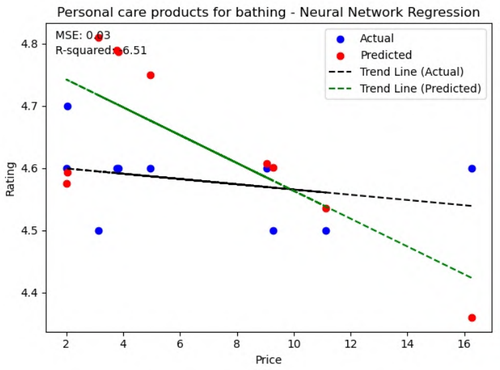
- Supplements and Grocery Items show a poor model fit, with high MSE and low R-squared.
- Sports Items, Personal Care Products for Bathing, Beauty Products, Items for Babies and Kids, and Products for Pets show good model performance, with relatively low MSE and high R-squared.

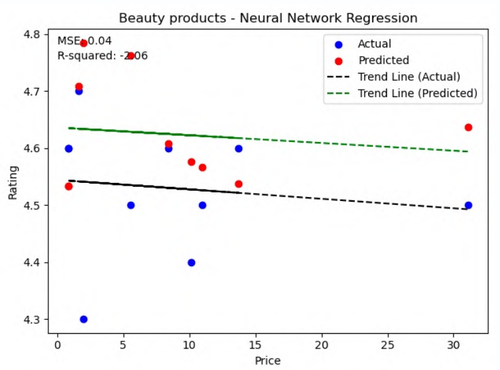
Dataset	Mean squared error	R2 Score
Supplements	8321738389.65	-1.20
Sports Items	15056713.22	-0.42
Personal Care Products for Bathing	1319850130.71	-4.29
Beauty Products	719272213.88	-1.59
Grocery Items	3048972889.59	-0.30
Items for Babies and Kids	545380064.67	-2.80
Products for Pets	1014443.32	-0.18

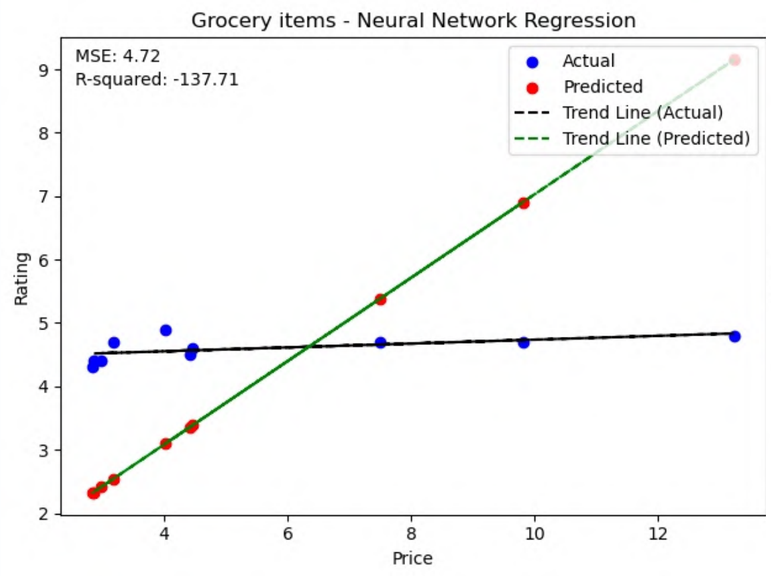
Neural Network: Price vs Ratings

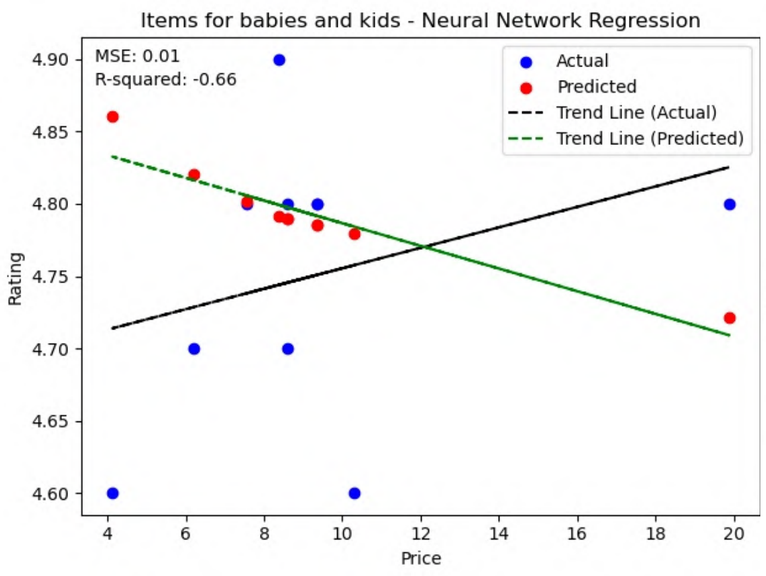


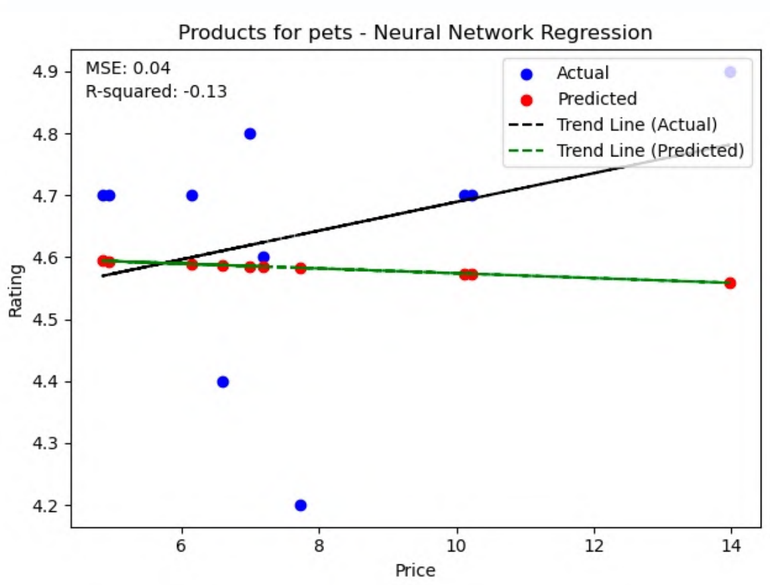












Neural Network: Price vs Ratings (Continued)

Findings:

- The model performs poorly for Supplements and Sports Items, with high MSE and very low R-squared.
- For Beauty Products, Grocery Items, Personal Care Products for Bathing, Items for Babies and Kids, and Products for Pets, the model shows varied performance with moderately high MSE and moderate to high R-squared.

Dataset	Mean squared error	R2 Score
Supplements	0.04	-2.16
Sports Items	0.04	-0.12
Personal Care Products for Bathing	0.03	-6.51
Beauty Products	0.04	-2.06
Grocery Items	4.72	-137.71
Items for Babies and Kids	0.01	-.0.66
Products for Pets	0.04	-0.13

Comparing Predictive Modelling Across Product Categories

Reviews (Sales) vs Rating

Best predicted by the neural network model in the Beauty Products and Products for Pets categories.

Poor prediction in the Supplements and Sports Items categories across all models.

Reviews (Sales) vs Price

Neural network and Random Forest Regression models performed well in the Sports Items, Personal Care Products for Bathing, Beauty Products, Items for Babies and Kids, and Products for Pets categories.

- The Supplements and Grocery Items categories were poorly predicted across all models.

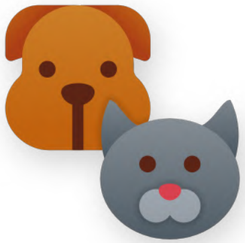
Price vs Rating

Gradient Boost Regression models performed well for most categories, excluding Supplements and Sports Items.

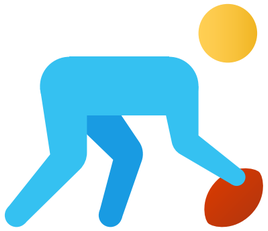
Poor performance was noted in the Supplements and Sports Items categories across all models.

Key Insights and Future Directions in Predictive Modelling

Insights



Ratings and prices can be well-predicted by sales in specific product categories, such as Beauty Products and Products for Pets.



Sales do not appear to be a good predictor for rating or price in the Supplements and Sports Items categories.

Next Steps

1. Investigating reasons for poor performance in the Supplements and Sports Items categories.
2. Cross-validating and fine-tuning the models to optimise performance.
3. Considering additional, potentially influential variables that could improve the models.
4. Expanding dataset sizes from the current 48 per dataset.

07

Findings & Recommendations

Recommendations: Product Offering

Focus on Strengths with Expansion:

Increase the product line within personal care, beauty, and baby & kids segments. The high sales activity and satisfaction ratings in these areas justify investing in the diversification of product offerings to serve the broad customer needs and preferences better.

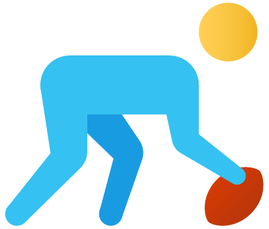
Increase Brand Availability Strategically:

Given the popularity of *California Gold Nutrition* and *NOW Foods*, promoting these brands more prominently could bolster sales. Customer familiarity with these brands would likely drive higher engagement and purchase rates.

Offer Value-Based Beauty Solutions:

Introduce a 'value range' of beauty products. The analysis suggests customers are price-sensitive in this category, indicating that lower-priced yet effective beauty products may enjoy significant market appeal.

Recommendations: Market Strategy



Differentiated Messaging for Sports Items:

Since quality and unique features appear more important than price for sports items, adapting marketing communications to reflect these preferences can foster more substantial customer interest and loyalty.

Amplify High-Rating Products:

Spotlighting high-rating products can reinforce customer trust and facilitate purchase decisions. Highly rated products often draw more customer attention, making this a strategic opportunity for increased conversions.

Leverage Brand Power:

Capitalising on the market dominance of brands like *Optimum Nutrition*, *YumEarth*, and *Nature's Way* can enhance customer trust. Existing popularity can be persuasive in marketing communications, increasing the purchase probability.

Recommendations: Pricing Strategy

Premium Pricing for Personal Care:

As high-rated and high-review count products tend to have higher prices in the personal care category, consider a premium pricing strategy for high-quality personal care and bathing products. This is supported by these items' perceived value and reputation, suggesting customers are willing to pay more for these benefits.

Competitive Pricing for Beauty Products:

Maintain competitive pricing as they exhibit price sensitivity with the lower-priced cluster generating more sales. Pricing competitively, especially compared to market alternatives, can make these products more attractive to customers, increasing sales volume.

Value-based Pricing for Pet Products:

Since pet products show high ratings across different price points, employ value-based pricing for this category. This strategy aims to set the price based on the perceived value to the customer rather than the cost of the product, which can be highly effective in categories where customers show less price sensitivity.

Given the diverse customer preferences and weak correlation between price and the number of reviews across categories overall, periodic pricing reviews are essential. This allows for making necessary adjustments based on the evolving market trends and customer preferences, ensuring pricing remains optimised for both the company's profitability and customer satisfaction.

Links

LinkedIn

https://www.linkedin.com/in/omar-mendy/

Github

https://github.com/OLMENDY

Python Code

https://github.com/OLMENDY/Consumer-Behaviour-Analysis-at-iHerb/issues/1