

CST3340 – Coursework 2

# SEPHORA

By Chioma Audrey Uche-Nwosu

M01052911

# Table of Contents

1. Introduction	3
2. Data Analysis and visualisation	5
3. Selection of Data Mining Algorithm and Data Pre-processing	17
4. Data Mining	18
5. Data Ethics	24
6. Conclusion	25
References	26

## 1. Introduction

This coursework focuses on a product and review dataset from Sephora, a French multinational retailer known for its extensive range of beauty products. Sephora stocks nearly 340 global brands plus its own Sephora Collection range.

The dataset offers a comprehensive overview of Sephora's beauty products and how customers of different ages interact with them through reviews. It includes over 8,000 products across various beauty categories and contains approximately one million customer reviews.

The dataset includes both qualitative and quantitative variables:

### Qualitative variables

These are labels, descriptions, or categories that help group or identify products but cannot be averaged.

- Identifiers: product\_id, product\_name, brand\_name
- Product taxonomy: primary\_category, secondary\_category, tertiary\_category, variation\_type, variation\_value
- Yes/No labels: exclusive\_label, limited\_edition, new, online\_only, sephora\_exclusive, stock\_status, is\_recommended
- Text-based fields: ingredients, review\_text

### Quantitative Variables

These contain values that can be counted, averaged, or used in mathematical analysis.

- Price-related measures: price\_usd, value\_price\_usd, sale\_price\_usd
- Interaction metrics: rating, avg\_rating, review\_count, loves\_count, helpfulness, total\_feedback\_count, %\_product\_recommended, age
- Time-based variable: submission\_time

In practice, the data can support decisions such as:

- Identifying products and brands that consistently perform well or receive repeated complaints
- Detecting loyal or high-engagement reviewers based on the author's age and gender
- Spotting trends in customer concerns or satisfaction over time

- Comparing the performance of Sephora-exclusive products with general brands
- Understanding the reviewer's demographic to support personalised recommendations

Sephora's dataset is already provided in clean CSV files, but if data cleaning were required, the process would begin with a basic data audit. This involves checking:

- Accuracy
- Completeness
- Consistency
- Validity
- Uniformity

### 5 MAJOR STEPS OF DATA CLEANSING PROCESS



#### *The 5 major steps of the data cleaning process*

The data-cleaning process can be broken into five simple stages. The first stage is defining the database, which means understanding what each table and variable represents. The next stage is locating the source of any dirty data, such as missing values, wrong formats, or duplicates. After this, the process involves prioritising the issues, focusing on the problems that would affect the analysis the most.

Another stage is preventing bad data from entering the system by keeping formats consistent and using basic validation rules. The final stage is removing existing bad data, which includes correcting incorrect values, handling missing data, and deleting duplicates. Other useful steps include fixing formatting issues, identifying outliers, checking consistency, and validating rating ranges.

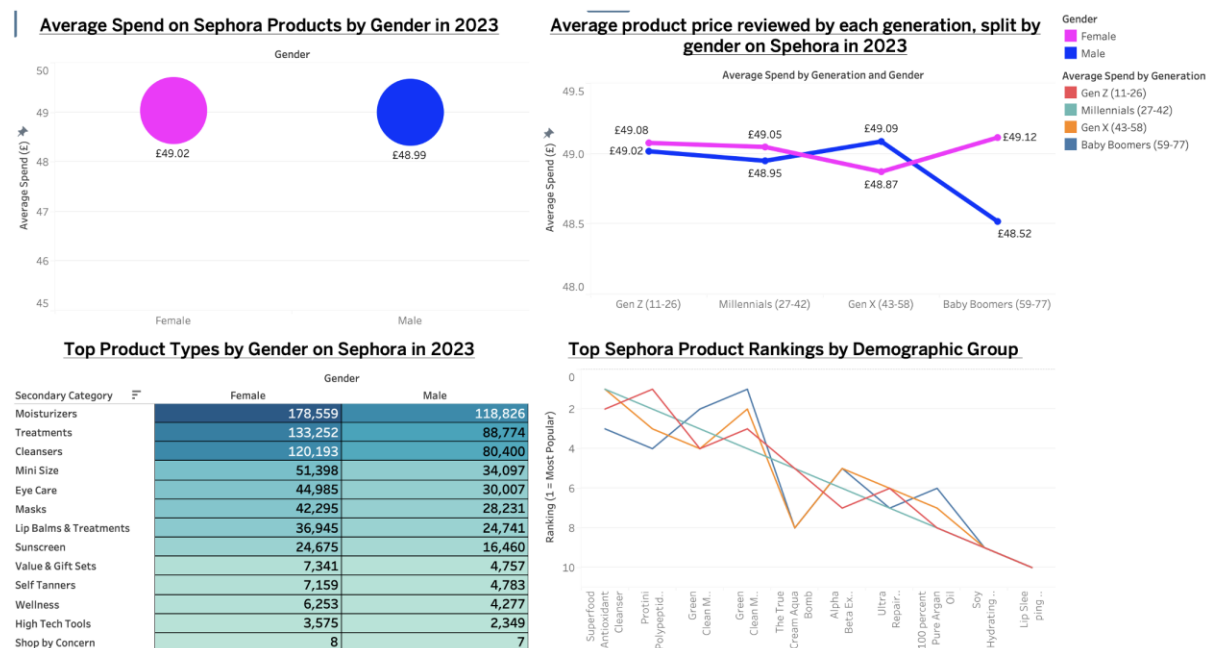
## 2. Data Analysis and visualisation

The five review files were combined into one table and linked to the product\_info table through product\_id, allowing all customer reviews and product details to be analysed together in Tableau.

The following visualisations highlight key patterns and insights identified in the dataset:

### 1. Comparing Spending Behaviour and Product Choices Across Gender and Generational Groups on Sephora

This dashboard starts by showing a simple overview of average spend by gender, and we can immediately see that women spend slightly more than men (i.e., £0.03).

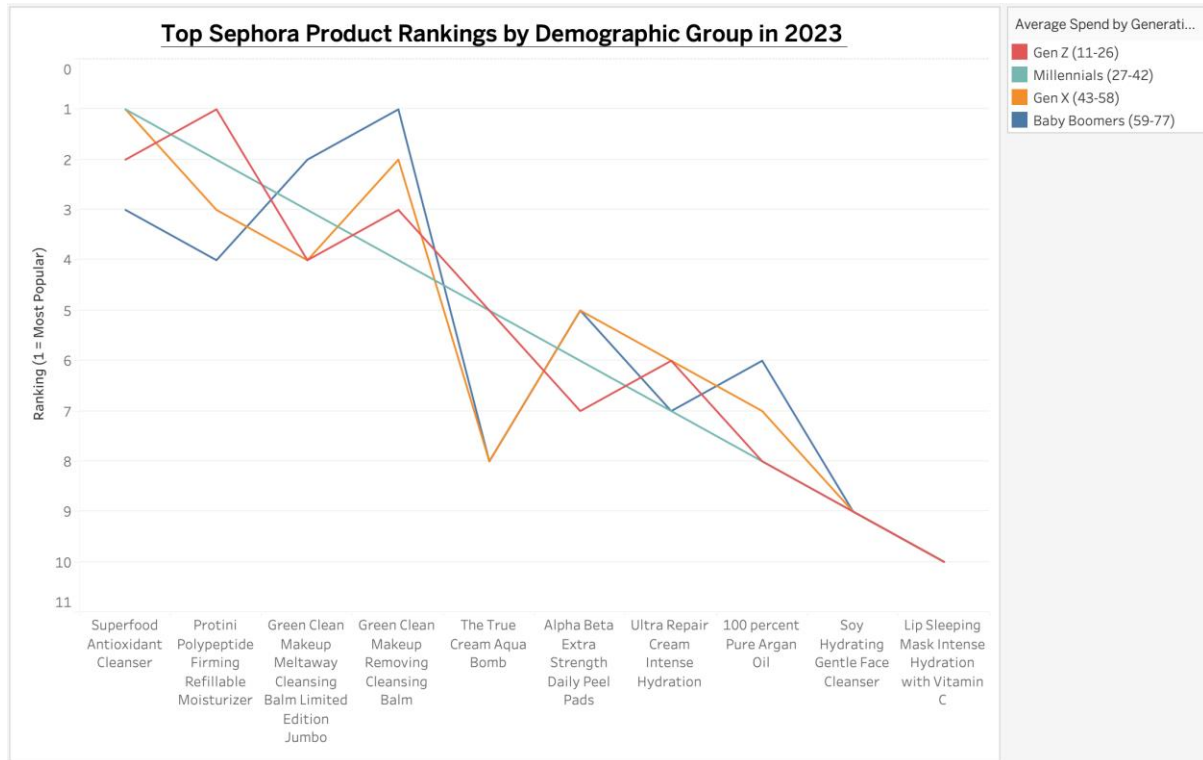


### *Comparing Spending Behaviour and Product Choices Across Gender and Generational Groups on Sephora*

As we move deeper into the other visuals, patterns become clearer: women consistently buy more skincare across every major category, especially moisturizers, cleansers, and treatments, and these categories outperform all others for both genders.

The generational breakdown adds even more insight, showing that older female groups, particularly Baby Boomers (ages 59-77), spend the most on skincare because they buy more premium, age-targeted products,

followed by Gen X (43-58) and then Gen Z (11-26). Men, however, show surprising purchasing power in the Gen Z and Gen X ranges, which suggests Sephora may not be fully tapping into a profitable male segment.



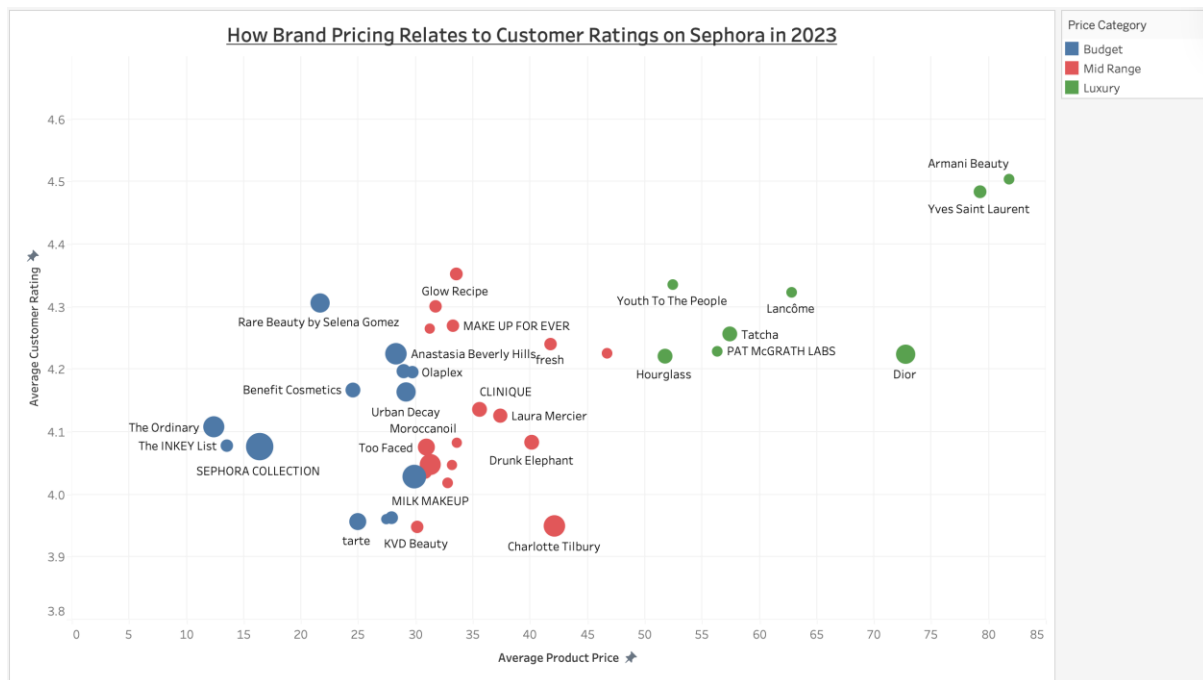
### *Closer analysis of Sephora's top product rankings by demographic group*

Based on these trends, Sephora should make sure they're always well-stocked in the categories that perform strongly across all groups, especially moisturizers, treatments, and cleansers, while also investing in better advertising and product education for male shoppers who clearly have the ability and willingness to spend more.

Finally, Sephora could benefit from targeted marketing strategies across generations, pushing premium anti-aging lines to older women, and more accessible, trend-driven products to Gen Z.

## 2. How Brand Pricing Relates to Customer Ratings on Sephora in 2023

This visualization looks at how a brand's average product price relates to its customer rating on Sephora in 2023.



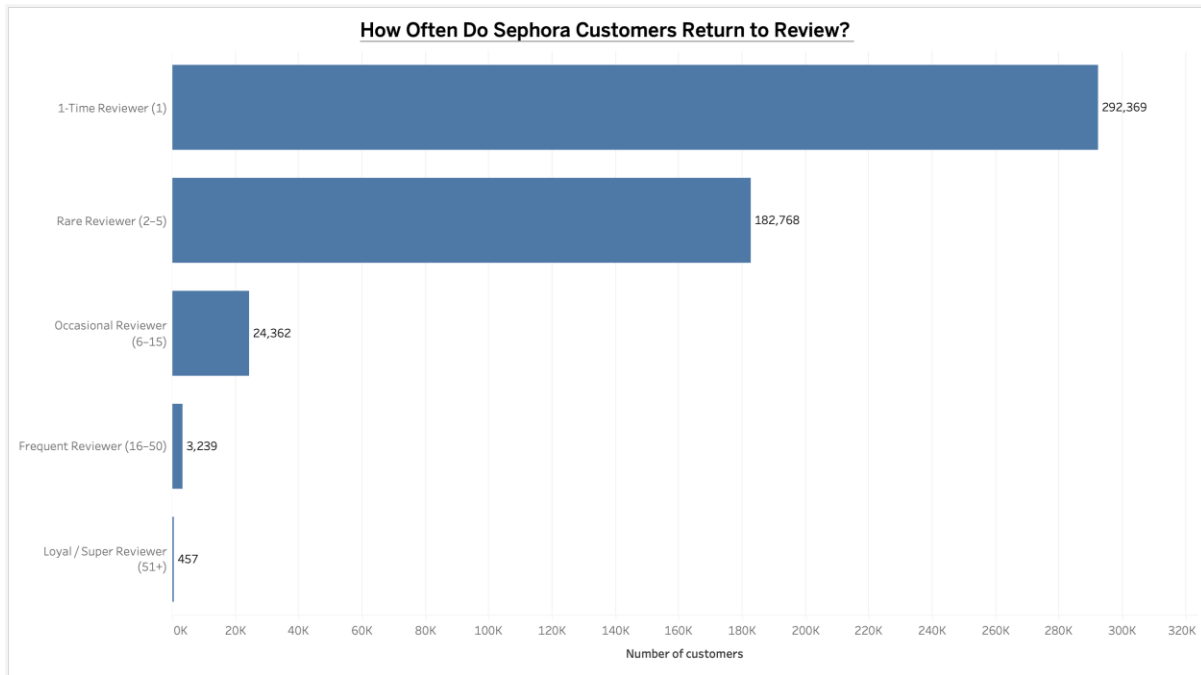
### *How Brand Pricing Relates to Customer Ratings on Sephora in 2023*

What we can see is that luxury brands tend to get slightly higher ratings, but not in a way that's guaranteed; a lot of mid-range brands perform just as well, which tells us that good reviews aren't only tied to expensive products.

Some luxury brands also have fewer reviews, so their high scores might not be the full picture yet. For Sephora, this means they should really push mid-range brands that are already performing strongly, because they're affordable for more people and clearly resonate with customers. Luxury brands could also benefit from Sephora driving more reviews and visibility, just to confirm whether those high ratings are consistent.

### 3. How Often Do Sephora Customers Return to Review?

This visualization shows that 58% of Sephora customers only leave one review, which is a clear sign that Sephora isn't retaining reviewers well.



### *How Often Do Sephora Customers Return to Review?*

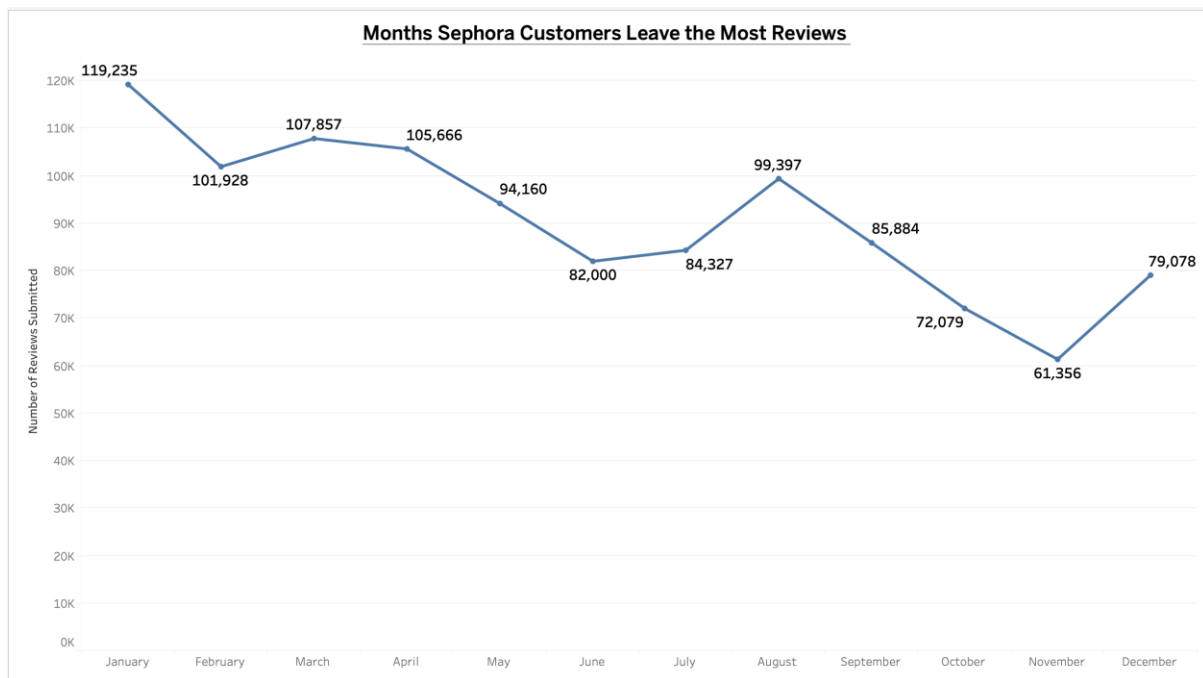
This is a huge opportunity for Sephora to push engagement by incentivising reviews properly; things like Sephora loyalty points, small sample-size freebies, or even early access to products from brands customers already love.

If they reward people who review consistently, that massive 1-time reviewer group will shrink, because customers will feel like there's something in it for them. Increasing engagement isn't just about getting more reviews; it gives Sephora deeper data on what customers want, what isn't working, and emerging product trends. And with better data, Sephora can improve personalisation, stock smarter, and even refine product recommendations.

#### 4. Months Sephora Customers Leave the Most Reviews

This visualization shows a clear seasonal pattern in customer review behaviour, with January recording the highest number of reviews, likely driven by New Year routines, post-holiday product usage, and customers finally trying gifts or purchases from late December.





### *Months Sephora Customers Leave the Most Reviews*

There is also a noticeable end-of-year rhythm: reviews dip in December but pick up sharply in January, suggesting that people have more free time and mental space to reflect on their purchases once the festive rush has passed.

A smaller decline also appears between April and June, which aligns with busy mid-year periods when many customers are finalizing school terms, university deadlines, or work deliverables before the summer break. Review activity rises again during summer, a time when people generally have more flexibility and are more likely to engage with products and leave feedback.

From August to November, review volume steadily declines, which may be connected to high-purchasing demographics returning from summer holidays and settling into full work or school routines, leaving less time to review. The sharp dip in October-November also overlaps with major shopping periods like Black Friday, where customers tend to focus on buying rather than evaluating older purchases.

These patterns suggest that reviews peak when customers have both the time and the headspace to share feedback. Sephora can use this insight to time sampling programs, post-purchase reminders, and review-incentive campaigns around review-rich months like January and

March, while using light engagement nudges during lower-review periods to stabilise feedback flow.

## 5. Sephora Skincare Product Counts and Pricing by Category

This table shows how Sephora structures its skincare range by breaking products into secondary and tertiary categories, and it highlights clear differences in both product variety and pricing.

Secondary Category	Tertiary Category	no_of_products	Avg. Price Usd
Cleansers	Face Wash & Cleansers	138	£34.63
	Toners	52	£32.30
	Exfoliators	31	£61.05
	Blotting Papers	4	£14.87
	Face Wipes	4	£3.31
	Makeup Removers	4	£27.17
Eye Care	Eye Creams & Treatments	108	£54.62
	Eye Masks	7	£41.29
High Tech Tools	Anti-Aging	41	£210.59
	Facial Cleansing Brushes	10	£144.66
	Hair Removal	7	£11.12
	Teeth Whitening	3	£48.91
Masks	Face Masks	91	£44.96
	Sheet Masks	23	£36.09
Moisturizers	Moisturizers	252	£59.30
	Mists & Essences	44	£82.19
	Face Oils	37	£59.79
	Night Creams	12	£52.93
	Decollete & Neck Creams	6	£73.65
	BB & CC Creams	3	£49.79
Self Tanners	For Body	22	£33.87
	For Face	14	£32.95
Shop by Concern	Anti-Aging	1	£28.00
Sunscreen	Face Sunscreen	55	£39.33
	Body Sunscreen	7	£34.62
Treatments	Face Serums	219	£63.46
	Facial Peels	27	£64.71
	Blemish & Acne Treatmen...	24	£27.46
	Beauty Supplements	30	£40.98
Wellness	Facial Rollers	22	£42.95
	Holistic Wellness	6	£81.20

### *Sephora Skincare Product Counts and Pricing by Category*

Categories like Moisturizers, Treatments, and Cleansers dominate the assortment with the highest number of products, which suggests they represent core customer needs and continued demand.

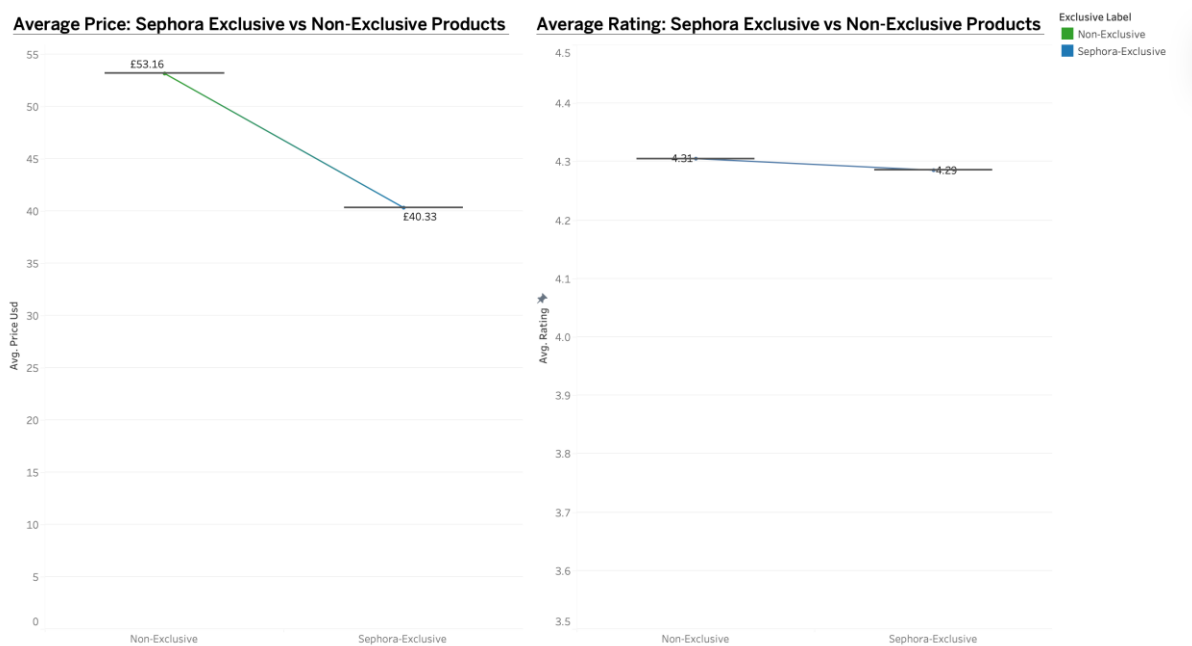
More specialised segments, such as High Tech Tools and Anti-Aging, appear in far smaller quantities but sit at significantly higher average price points, indicating premium positioning and lower accessibility. This pattern reflects a two-tier strategy: broad, affordable everyday skincare categories driving volume, while smaller luxury or treatment-based categories contribute higher margins.

For Sephora, the data shows a strong opportunity to expand premium Anti-Aging and Facial Tech products, as demand for advanced skincare is rising and female Baby Boomers have the highest spending power in these categories. It also shows the importance of keeping staple

categories, like Moisturizers and Cleansers, well stocked, competitively priced, and consistently updated, as they form the backbone of customer purchasing behaviour.

## 6. Comparing Sephora Exclusive vs Non-Exclusive Products in Price and Customer Ratings

This dashboard shows a simple but important pattern: Sephora-exclusive products are noticeably cheaper on average, yet their ratings sit almost identically to non-exclusive products.



### *Comparing Sephora Exclusive vs Non-Exclusive Products in Price and Customer Ratings*

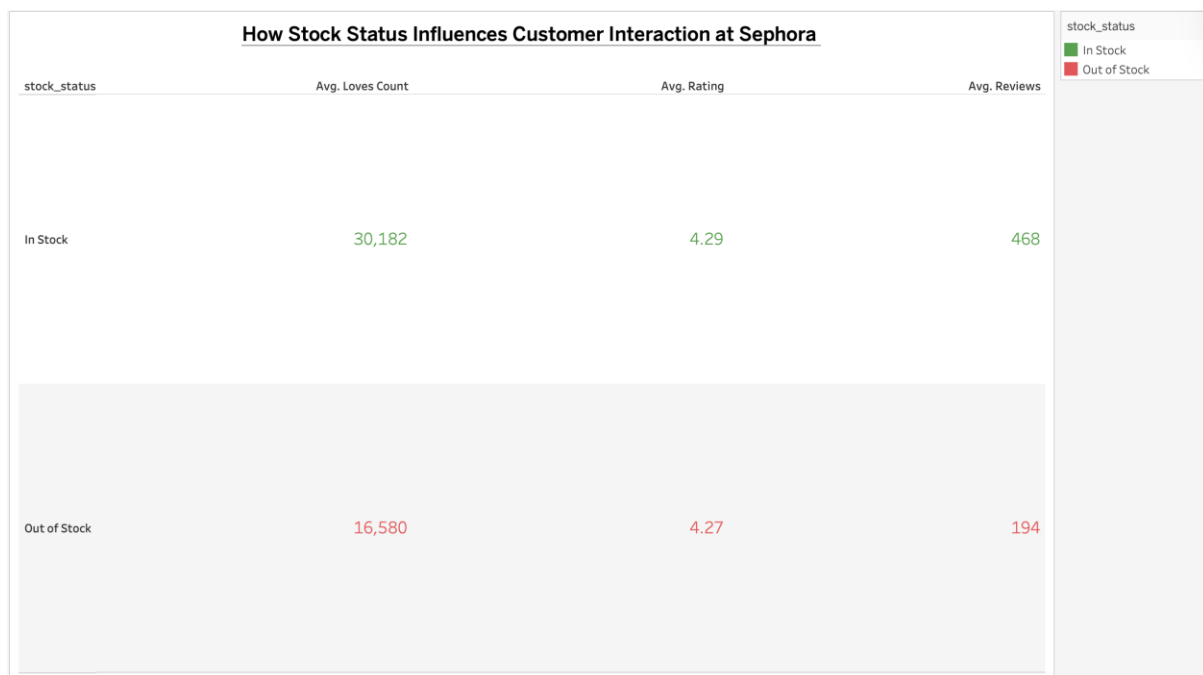
This suggests that Sephora's exclusive ranges offer strong value for money, performing just as well as higher-priced brands while remaining more affordable. The fact that customers rate both groups at nearly the same level indicates strong trust in Sephora's exclusives, meaning the brand is already succeeding in terms of quality perception.

With this insight, Sephora could confidently increase visibility, advertising, and promotion of exclusive lines, as they have clear potential to capture even more market share. There is also room to strategically raise prices slightly on the highest-performing exclusive products, especially since earlier findings showed that customers tend to

rate more expensive products a little better, possibly due to perceived quality.

## 7. How Stock Status Influences Customer Interaction at Sephora

This visual reveals something deeper than just “out-of-stock items get fewer reviews.”



### *How Stock Status Influences Customer Interaction at Sephora*

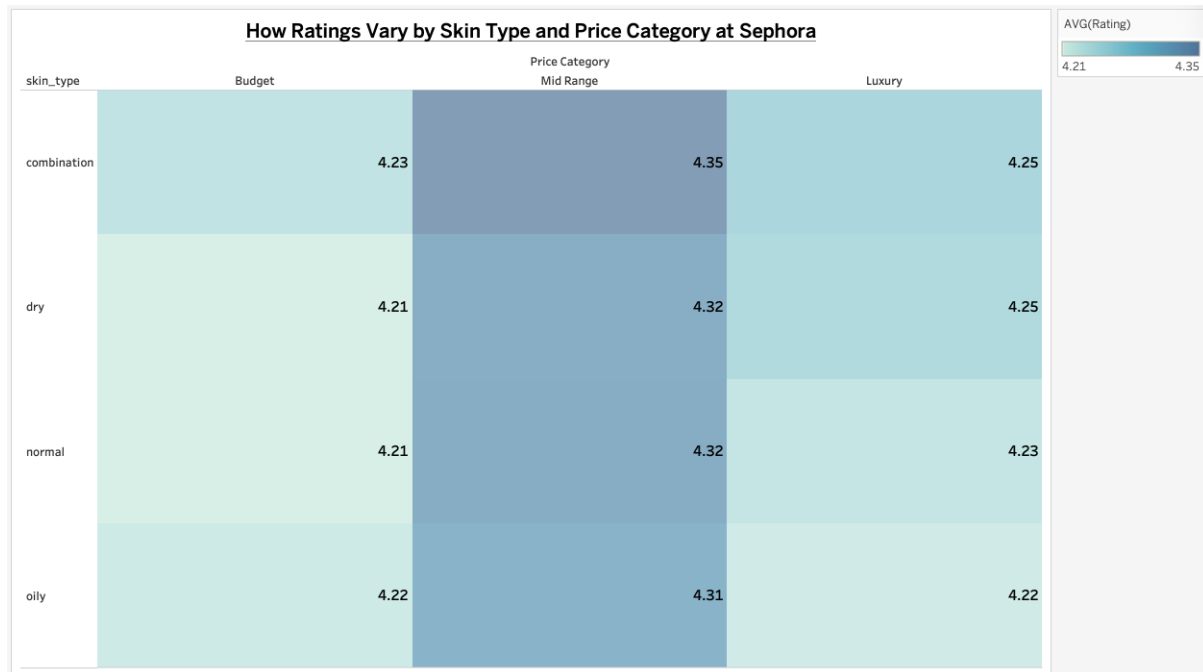
What stands out is that out-of-stock products still hold almost the same average rating as in-stock ones (4.27 vs 4.29), meaning customers genuinely like them, but Sephora isn’t keeping up with demand.

The drop in love count and reviews isn’t just because the products aren’t available; it’s a sign of lost engagement and lost hype, because people can’t interact with what they want to buy. This suggests a supply chain gap where Sephora may be underestimating demand for its strongest performers, especially if high-rated products frequently go out of stock.

There’s also a long-term brand risk here: repeatedly unavailable products can create customer frustration, making shoppers switch to competitors who are more reliably stocked. If Sephora uses this data properly, it should identify which products consistently go out of stock despite strong ratings, and treat them as high-priority items for restocking, forecasting, and even future product expansion.

## 8. How Ratings Vary by Skin Type and Price Category at Sephora

This heatmap shows that mid-range skincare products consistently achieve the highest ratings across every skin type, which suggests that customers feel they get the best balance of price and performance in this category.



### *How Ratings Vary by Skin Type and Price Category at Sephora*

What makes this even more interesting is that the pattern holds regardless of skin concerns-combination, dry, normal, and oily skin all lean toward mid-range formulas as their top performers. This hints that mid-range brands may be using more effective ingredient combinations or better formulations that resonate widely, not just within a niche group.

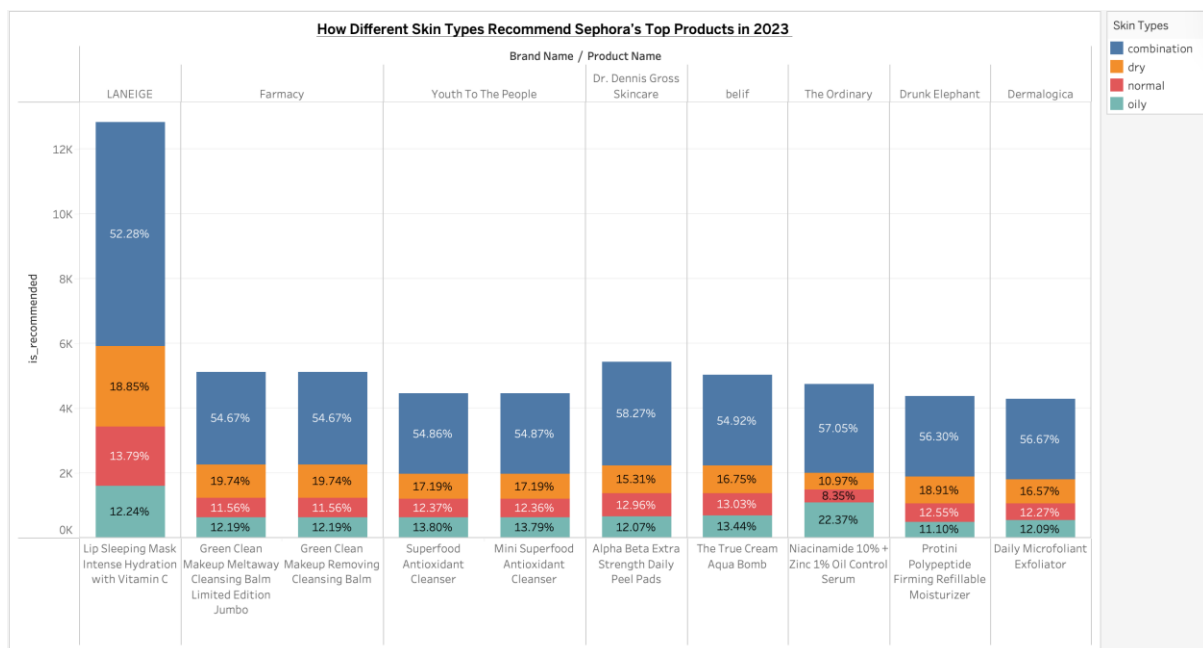
Luxury products perform well, but they do not meaningfully outperform mid-range options, suggesting that customers may not perceive the added price premium as translating into noticeably better results. Budget products perform the lowest overall, but the gap is very small, meaning Sephora could potentially elevate select budget lines by incorporating successful mid-range ingredients or marketing angles.

Another subtle insight is that combination-skin shoppers respond especially well to luxury and mid-range options, which could be a sign

that these groups are more willing to experiment with innovative or targeted formulas.

## 9. Top Products by Skin Type: Which items perform best for oily, dry, combination, and normal skin at Sephora in 2023?

This visualisation shows something interesting: combination-skin customers basically dominate every single top product, which makes sense because their skin deals with both dryness and oiliness, so they naturally benefit from a wider variety of formulas.



### *Top Products by Skin Type: Which items perform best for oily, dry, combination, and normal skin at Sephora in 2023?*

Across almost every brand, combination users make up over half of all recommendations, which basically tells Sephora, “This is your most influential skincare group, don’t ignore them.”

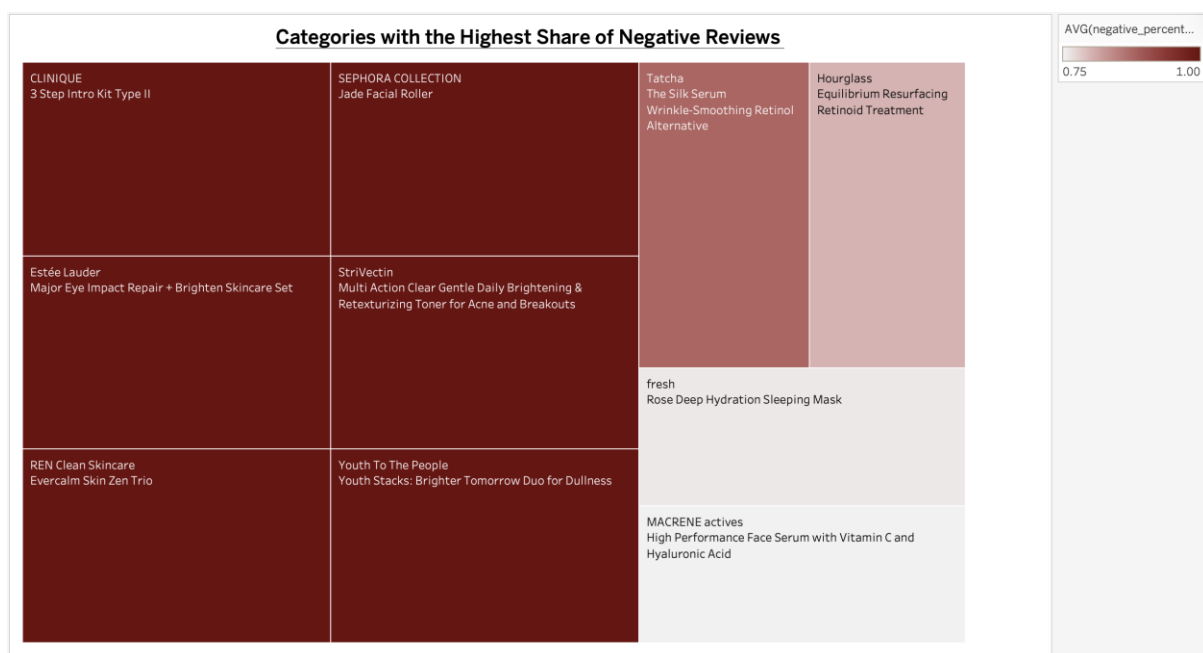
We also see that cleansers are the real winners here, consistently appearing in the top-performing list, which suggests that people are extremely loyal to their cleansing routines. Another standout is how strongly oily-skin shoppers gravitate toward The Ordinary’s Niacinamide Serum, which is entirely understandable because that ingredient genuinely transforms oily and acne-prone skin.

Dry and normal skin users are present, but they never outperform combination skin, which means Sephora could easily introduce “combination-skin hero kits,” targeted sampling, or personalised bundles since this group clearly loves almost everything.

Strategically, Sephora can use these insights to push smarter product recommendations, spotlight high-performing cleansers across skin types, and build stronger marketing campaigns around products that have cross-skin appeal. There’s also a great chance to create oily-skin discovery sets built around niacinamide-based products because that community clearly responds extremely well to them.

## 10. Categories with the Highest Share of Negative Reviews

This visualisation highlights the products with the highest share of negative reviews, and immediately, you can see that this isn’t just random; there’s a pattern.



### *Categories with the Highest Share of Negative Reviews*

Several of these products sit in categories that usually carry high expectations (like retinol treatments, vitamin C serums, and acne-targeting toners), and when skincare doesn’t deliver fast or causes irritation, customers are brutally honest.

Instead of simply removing them, Sephora can use this as a real performance audit: run targeted promotions or controlled price drops to see whether negative sentiment is rooted in affordability, product misuse, or the formula itself. If reviews improve after education and pricing changes, great, the product was misunderstood. But if negativity stays high even after Sephora tries to help the product succeed, then it becomes clear that this is a deeper product issue, and it may genuinely need to be discontinued or reformulated.

There's also a huge opportunity for Sephora to intercept customer disappointment before it turns into public negativity. For example, products with consistently bad feedback could be paired with "Before You Buy" guidance, such as skin-type warnings, ingredient explanations, or recommended complementary products to reduce irritation. Sephora could also introduce proactive alerts: "People with dry skin often dislike this product; here's a better match for you."



### 3. Selection of Data Mining Algorithm and Data Pre-processing

A decision tree classifier was chosen because it fits the structure of the dataset and matches the patterns seen in the visual analysis. Since the goal is to predict a product's price category (Budget, Mid, or Luxury), decision trees give simple, understandable "if-then" rules that explain how products are split into each price tier. Earlier Tableau visuals, such as brand price vs. rating and rating differences by skin type, showed clear separation between the three price groups, supporting the choice of a decision tree.

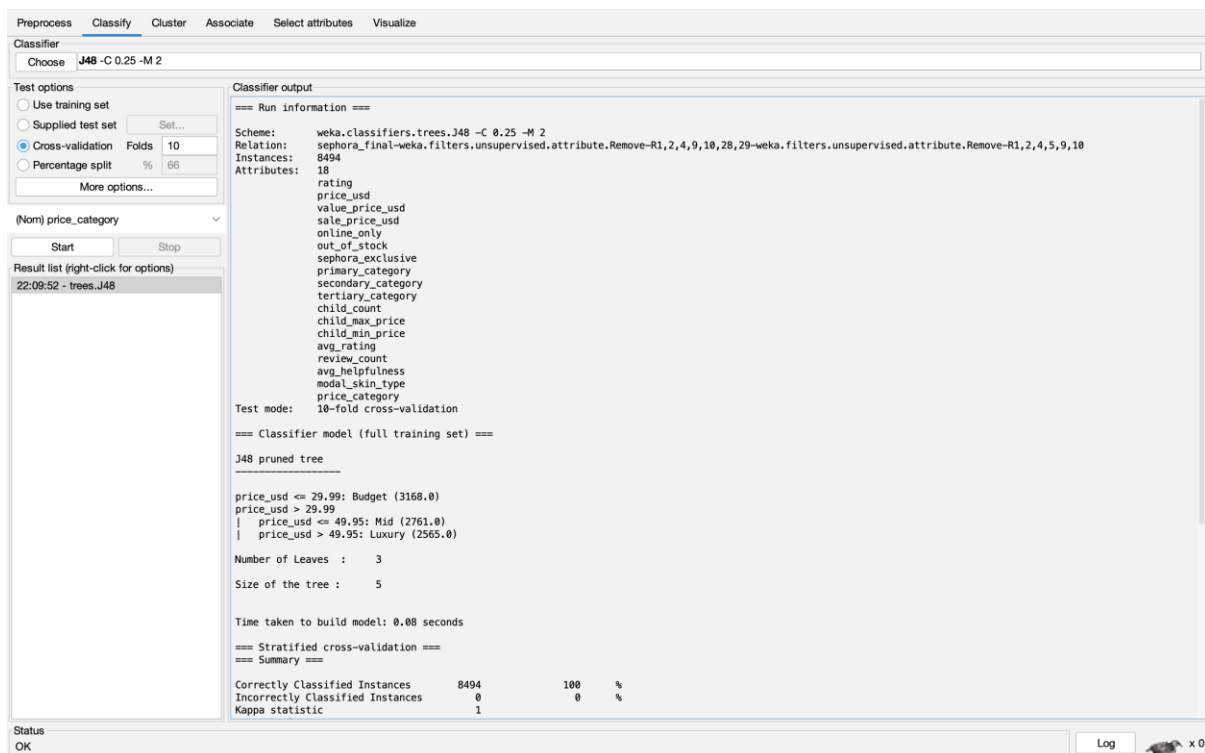
For preprocessing, all five review files were merged in Python, grouped by product\_id, and combined with the product\_info file so each product had features like average rating, review\_count, and helpfulness. Python also helped check for missing IDs, unusual prices, duplicates, and outliers. The final dataset was converted from CSV to ARFF to ensure Weka correctly handled the nominal class variable.

Several long text attributes (e.g., product\_name, brand\_name, variation fields, size, first\_review, and last\_review) were removed because J48 cannot process raw string data, and these fields had thousands of unique values with little predictive value.

## 4. Data Mining

### Iteration 1

The first iteration used the full dataset. The resulting model was extremely simple: the tree split entirely on the price\_usd variable, reproducing the thresholds used earlier to discretise products into Budget, Mid, and Luxury categories. Products priced at or below £29.99 were classified as Budget, those priced between £29.99 and £49.95 as Mid, and those above £49.95 as Luxury.



Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) price\_category

Start Stop

Result list (right-click for options)

22:09:52 - trees.J48

Classifier output

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: sephora\_final-weka.filters.unsupervised.attribute.Remove-R1,2,4,9,10,28,29-weka.filters.unsupervised.attribute.Remove-R1,2,4,5,9,10

Instances: 8494

Attributes: 18

rating

price\_usd

value\_price\_usd

sale\_price\_usd

online\_only

out\_of\_stock

sephora\_exclusive

primary\_category

secondary\_category

tertiary\_category

child\_count

child\_max\_price

child\_min\_price

avg\_rating

review\_count

avg\_helpfulness

modal\_skin\_type

price\_category

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

price\_usd <= 29.99: Budget (3168.0)

price\_usd > 29.99

| price\_usd <= 49.95: Mid (2761.0)

| price\_usd > 49.95: Luxury (2565.0)

Number of Leaves : 3

Size of the tree : 5

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8494	100	%
Incorrectly Classified Instances	0	0	%
Kappa statistic	1		

Status OK Log x 0

### *Iteration 1 – J48 Decision Tree Using All Key Attributes*

Because price\_category was created from price\_usd, the tree achieved a perfect fit, with 100% accuracy and no misclassified instances. Although the performance metrics are flawless, they mainly confirm that the discretisation rules were applied consistently, rather than revealing any deeper relationships in the data.

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) price\_category

Start Stop

Result list (right-click for options)

22:09:52 - trees.J48

Classifier output

model\_skin\_type  
price\_category

Test mode:  
10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

```

price_usd <= 29.99: Budget (3168.0)
price_usd > 29.99
| price_usd <= 49.95: Mid (2761.0)
| price_usd > 49.95: Luxury (2565.0)

```

Number of Leaves : 3

Size of the tree : 5

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	8494	100	%
Incorrectly Classified Instances	0	0	%
Kappa statistic	1		
Mean absolute error	0		
Root mean squared error	0		
Relative absolute error	0	%	
Root relative squared error	0	%	
Total Number of Instances	8494		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	Budget
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	Luxury
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	Mid
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	

=== Confusion Matrix ===

	a	b	c	<-- classified as
3168	0	0	0	a = Budget
0	2565	0	0	b = Luxury
0	0	2761	0	c = Mid

Status OK Log x 0

*Iteration 1 – Model Performance: 100% Accuracy*

This iteration acts as a baseline and highlights the need to remove price\_usd in the next iteration to explore whether other variables, such as ratings, review patterns, or category information, carry any predictive value for price\_category.

## Iteration 2

In Iteration 2, I removed all numeric price attributes to evaluate whether price\_category could be predicted using only non-price features such as product categories, ratings, and customer interactions.

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) price\_category

Start Stop

Result list (right-click for options)

22:09:52 - trees.J48

23:56:13 - trees.J48

Classifier output

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: sephora\_final-weka.filters.unsupervised.attribute.Remove-R1,2,4,9,10,28,29-weka.filters.unsupervised.attribute.Remove-R1,2,4,5,9,10-weka.filters.unsupervised.attribute.Remove-R1,2,4,5,9,10-weka.filters.unsupervised.attribute.Remove-R1,2,4,5,9,10

Instances: 8494

Attributes:

- rating
- primary\_category
- secondary\_category
- tertiary\_category
- child\_count
- avg\_rating
- review\_count
- avg\_helpfulness
- modal\_skin\_type
- price\_category

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

```

primary_category = Bath & Body
|
| tertiary_category = Accessories: Mid (0.0)
| tertiary_category = Aftershave: Mid (0.0)
| tertiary_category = Anti-Aging: Mid (0.0)
| tertiary_category = BB & CC Cream: Mid (0.0)
| tertiary_category = BB & CC Creams: Mid (0.0)
| tertiary_category = Bath Soaks & Bubble Bath: Budget (7.52/3.28)
| tertiary_category = Beauty Supplements: Mid (0.0)
| tertiary_category = Blemish & Acne Treatments: Mid (0.0)
| tertiary_category = Blotting Papers: Mid (0.0)
| tertiary_category = Blush: Mid (0.0)
| tertiary_category = Body Lotions & Body Oils
| | secondary_category = Accessories: Mid (0.0)
| | secondary_category = Bath & Body: Mid (0.0)
| | secondary_category = Bath & Shower: Mid (0.0)
| | secondary_category = Beauty Accessories: Mid (0.0)
| | secondary_category = Beauty Supplements: Mid (1.05)
| | secondary_category = Beauty Tools: Mid (0.0)
| | secondary_category = Body Care: Budget (0.52)
| | secondary_category = Body Moisturizers
| | | rating <= 4.4827
| | | | child_count <= 0: Luxury (96.71/57.6)
| | | | child_count > 0
| | | | | child_count <= 1: Luxury (14.0/6.0)
| | | | | child_count > 1
| | | | | child_count <= 2: Budget (3.0)

```

Status OK Log x 0

## Iteration 2 – Decision Tree Output After Removing All Price Variables

Once these attributes were removed, the model's accuracy dropped significantly from 100% to approximately 60%. The decision tree became far more complex and relied mostly on product categories and rating information to make predictions.

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) price\_category

Start Stop

Result list (right-click for options)

22:09:52 - trees.J48

23:56:13 - trees.J48

Classifier output

```

| | | tertiary_category = Skincare Sets: Budget (0.0)
| | | tertiary_category = Sponges & Applicators: Budget (0.0)
| | | tertiary_category = Sunscreen: Budget (0.0)
| | | tertiary_category = Teeth Whitening: Budget (0.0)
| | | tertiary_category = Tinted Moisturizers: Budget (0.0)
| | | tertiary_category = Toners: Budget (0.0)
| | | tertiary_category = Tweezers & Eyebrow Tools: Budget (0.0)
| | | tertiary_category = Under-Eye Concealer: Budget (0.0)
| | modal_skin_type = dry: Budget (2.0/1.0)
| | modal_skin_type = normal: Budget (0.0)
| | modal_skin_type = oily: Budget (0.0)
| secondary_category = Women: Luxury (0.0)
primary_category = Tools & Brushes: Budget (52.0/10.0)

```

Number of Leaves : 2426

Size of the tree : 2625

Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	5142	60.5368 %
Incorrectly Classified Instances	3352	39.4632 %
Kappa statistic	0.4868	
Mean absolute error	0.3168	
Root mean squared error	0.418	
Relative absolute error	71.572 %	
Root relative squared error	88.8541 %	
Total Number of Instances	8494	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.620	0.283	0.645	0.620	0.632	0.420	0.781	0.661	Budget
	0.699	0.165	0.647	0.699	0.672	0.523	0.837	0.717	Luxury
	0.501	0.225	0.518	0.501	0.509	0.279	0.708	0.489	Mid
Weighted Avg.	0.605	0.199	0.604	0.605	0.604	0.405	0.774	0.622	

=== Confusion Matrix ===

	a	b	c	← classified as
1965	401	882		a = Budget
285	1793	487		b = Luxury
798	579	1384		c = Mid

Status OK Log x 0

## Iteration 2 – Model Accuracy and Confusion Matrix (~60% Accuracy)

These results show that while certain categories give some indication of price positioning, Sephora's pricing is not determined by category alone. Luxury products remain more distinguishable than mid-tier or budget options, but overall, the findings suggest that Sephora uses a multi-factor pricing strategy that cannot be cleanly inferred from non-price attributes alone.

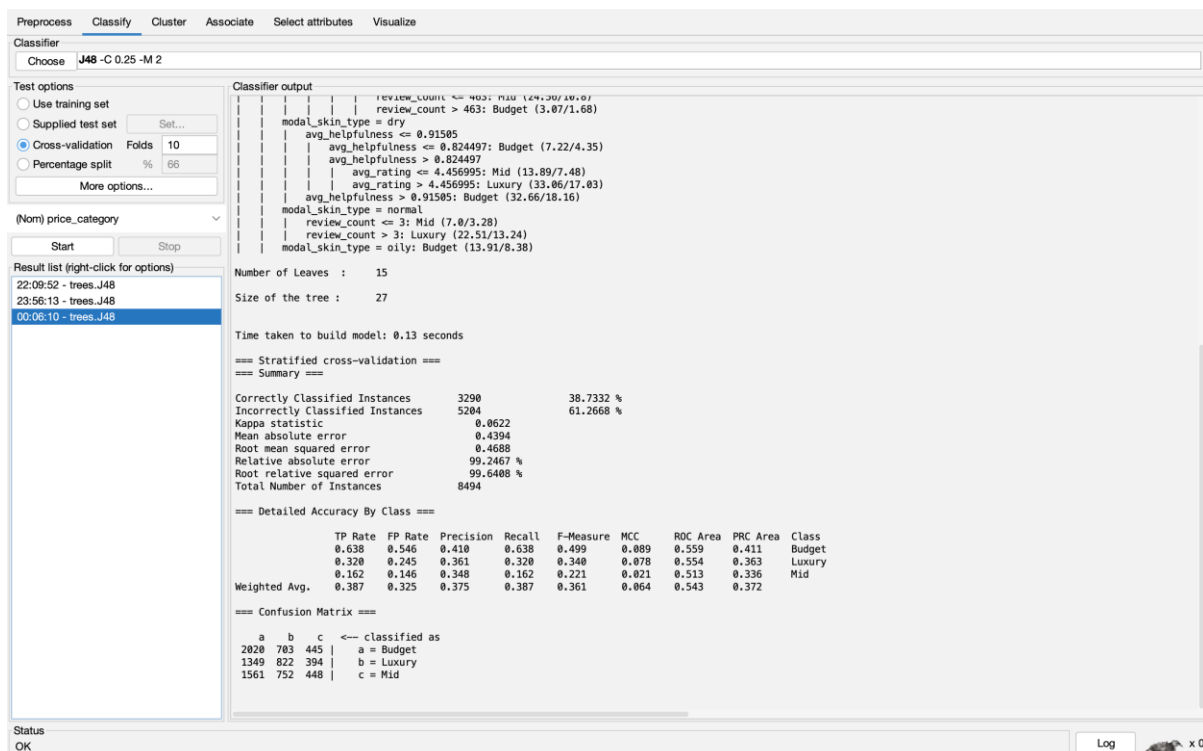
### Iteration 3

Iteration 3 focused on testing whether customer-interaction attributes alone (ratings, average rating, review count, helpfulness, and skin type) could predict product price category.

The screenshot shows the Weka software interface. The 'Classify' tab is selected. Under 'Test options', 'Cross-validation' is chosen with 10 folds and a 66% split. The 'Result list' on the left shows three runs of 'trees.J48', with the third run (00:06:10) selected. The 'Classifier output' pane on the right displays the 'J48 pruned tree' structure. The tree starts with a root node 'rating <= 4.3189: Budget (4410.35/2539.81)'. Subsequent nodes use 'avg\_helpfulness', 'modal\_skin\_type', and 'review\_count' to further partition the data into 'Budget', 'Mid', and 'Luxury' categories. The final output shows 15 leaves and a tree size of 77.

### *Iteration 3 – J48 Decision Tree Structure After Removing Price & Category Attributes*

After removing the entire product-category hierarchy, the classifier accuracy collapsed to 38.7%, which is close to random guessing for a three-class problem. This result shows that consumer engagement features do not strongly determine whether a Sephora product is budget, mid-range, or luxury.



### Iteration 3 – Model Accuracy, Confusion Matrix, and Performance Metrics

Products across all price tiers receive similar review patterns, meaning customer feedback behaviour is not tied to price. Sephora’s pricing is not based on perceived customer satisfaction or review volume. Instead, it is driven by product-level characteristics such as category, brand, exclusivity, and marketing strategy (which were removed in this iteration). The model’s sharp performance drop confirms that customer-interaction metrics alone cannot explain Sephora’s pricing structure.

### Discussing Iteration 1

Iteration 1 delivered the strongest performance, with 100% correctly classified instances using a very small J48 decision tree. The model relied almost entirely on price\_usd, recreating the exact thresholds used to generate the price\_category labels ( $\leq 29.99$  = Budget,  $\leq 49.95$  = Mid,  $> 49.95$  = Luxury).

Because the target variable was directly derived from price, the relationship was deterministic rather than behavioural, which explains the perfect accuracy and minimal tree size. The confusion matrix showed zero misclassifications across all three categories, confirming a perfect model fit, not because the model “learned” customer patterns, but because it simply rediscovered the rule that created the categories in

the first place.

## 5. Data Ethics

When working with business data, it is important to think about the ethical issues that may come up during analysis. Ethical data use means being honest about what the data can and cannot show, avoiding biased interpretations, and making sure that the analysis does not harm individuals or groups.

Even though the Sephora dataset is mainly product-based and does not include personal details, ethical thinking is still needed because data analysis can influence business decisions, customer experiences, and how companies understand different groups. There are also legal responsibilities around data use.

The main framework in the UK and Europe is the General Data Protection Regulation (GDPR), which sets rules for how personal data should be collected, stored, and used. GDPR explains what counts as personal data and outlines the roles of data controllers and data processors, as well as principles like fairness, transparency, purpose limitation, and data minimisation.

While this project does not involve identifiable customer information, the same legal ideas are important to remember, especially for businesses that analyse customer behaviour or make decisions using user data. These rules help protect people's privacy and stop organisations from misusing information.

Professional considerations also matter. Analysts are expected to follow good practice by keeping data secure, checking data quality, avoiding misleading conclusions, and being clear about the limitations of the analysis. Models such as decision trees can be useful, but they may simplify patterns, so it is important not to overstate the results.



## 6. Conclusion

The visualisations revealed several clear behaviour patterns across Sephora customers. Women spend slightly more than men, and female Baby Boomers show the highest overall spend, especially on skincare essentials like the Green Clean Cleansing Balm. Across all demographics, moisturisers, treatments, and cleansers consistently rank as the most reviewed and most purchased categories, signalling strong and stable demand. The exclusivity analysis showed that Sephora-exclusive products have similar customer ratings to non-exclusive items but noticeably lower prices, suggesting that Sephora could confidently grow its exclusive range without harming customer satisfaction. Review-frequency patterns also revealed that over half of customers leave only one review, suggesting weak engagement and highlighting an opportunity for improved loyalty or feedback incentives.

The data mining results further supported these insights. Iteration 1 achieved perfect accuracy because the model simply rediscovered the price thresholds used to create the price\_category variable. Once price attributes were removed in Iteration 2, accuracy fell to around 60%, proving that product categories and basic interaction metrics only partly explain how Sephora assigns price tiers. Iteration 3 dropped to 38% accuracy, confirming that customer-interaction variables alone cannot predict pricing. Overall, the decision trees show that Sephora's pricing strategy is multi-layered and strongly tied to product characteristics rather than customer behaviour.

These findings offer valuable Business Intelligence to Sephora. Marketing teams can target Baby Boomers, Gen X men, and Gen Z women with tailored campaigns. Stock teams can prioritise high-demand products, especially top cleansers and moisturisers. The product team can expand Sephora-exclusive offerings, knowing they perform competitively. Finally, review engagement can be improved through incentives, reminders, or seasonal campaigns to turn one-time reviewers into repeat contributors.

## References

**Beauty Bay (2025)** *Best The Ordinary products for oily skin.* Available at: <https://www.beautybay.com/edited/best-the-ordinary-products-for-oily-skin/> (Accessed: 21 November 2025).

**BeautyMatter (2025)** *Understanding the men's skincare boom.* Available at: <https://beautymatter.com/articles/understanding-the-mens-skincare-boom#:~:text=The%20men's%20skincare%20boom%20is%20a%20reflection%20of%20broader%20cultural,to%20look%20and%20feel%20better> (Accessed: 27 November 2025).

**CosmeticsDesign (2025)** *68% increase in male skin care usage: What's driving the surge?* Available at: <https://www.cosmeticsdesign.com/Article/2024/09/04/68-increase-in-male-skin-care-usage-what-s-driving-the-surge/#:~:text=Key%20factors%20driving%20the%20noteworthy,burgeo ning%20male%20skin%20care%20market> (Accessed: 5 December 2025).

**DataFlog (2025)** *What steps should be included in the data cleansing process?* Available at: <https://dataflog.com/what-steps-should-included-data-cleansing-process/> (Accessed: 22 November 2025).

**Estique Clinic (2025)** *Anti-aging treatments: Benefits, side effects and results.* Available at: <https://www.estiqueclinic.com/blog/anti-aging-treatments-benefits-side-effects-and-results> (Accessed: 25 November 2025).

**Healthline (2025)** *Everything to know about niacinamide.* Available at: <https://www.healthline.com/health/beauty-skin-care/niacinamide> (Accessed: 2 December 2025).

**IBM (2025)** *Decision trees.* Available at: <https://www.ibm.com/think/topics/decision-trees> (Accessed: 29 November 2025).

**Kaggle (2025)** *Sephora products and skincare reviews.* Available at: <https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews/data> (Accessed: 21 November 2025).

**Sephora (2025)** *About Sephora*. Available at: <https://www.sephora.com/beauty/about-us> (Accessed: 21 November 2025).

**Sharda, R., Delen, D. and Turban, E. (2025)** *Business intelligence, analytics, data science, and AI*. Chapter 1: An overview of business intelligence, analytics, data science and AI. Hoboken, NJ: Pearson. (Accessed: 3 December 2025).

**Sharda, R., Delen, D. and Turban, E. (2025)** *Business intelligence, analytics, data science, and AI*. Chapter 5: Data mining process, methods, and algorithms. Hoboken, NJ: Pearson. (Accessed: 5 December 2025).

**Tableau (2025)** *Union your data*. Available at: <https://help.tableau.com/current/pro/desktop/en-us/union.htm> (Accessed: 21 November 2025).