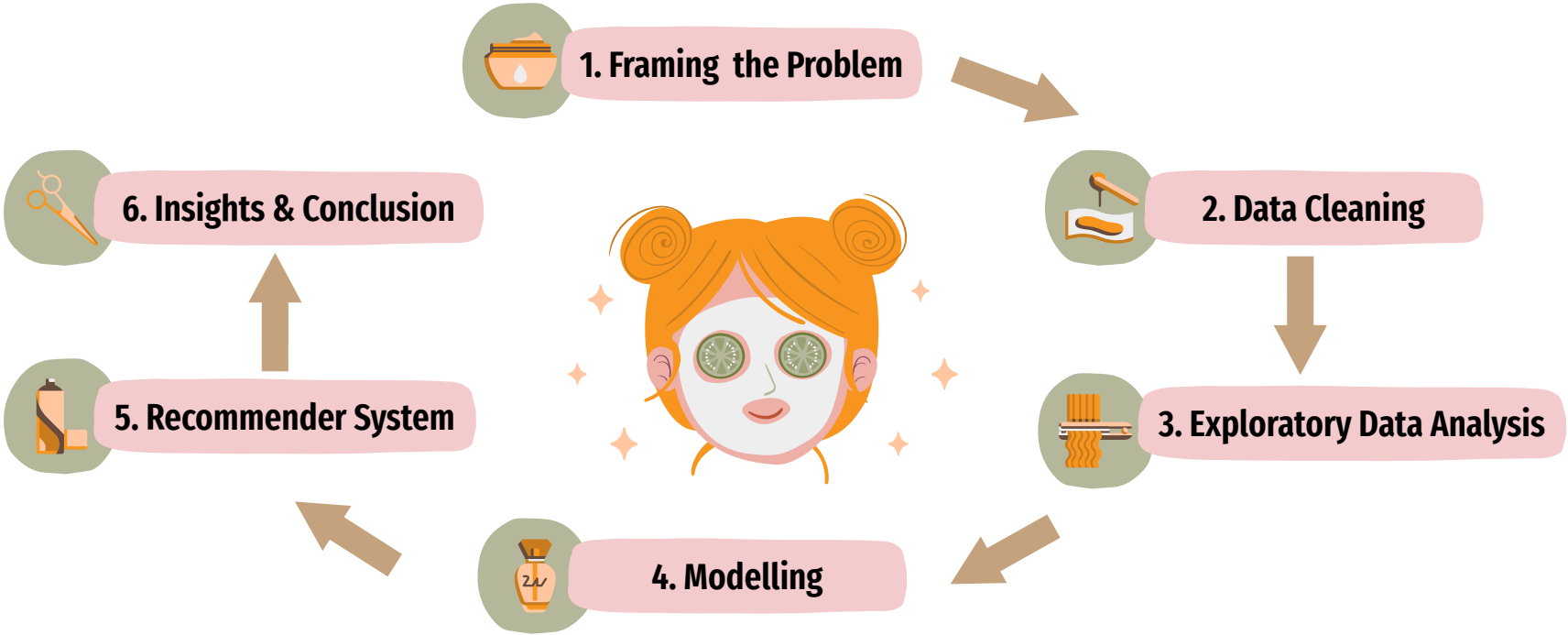
Skincare Product Analysis

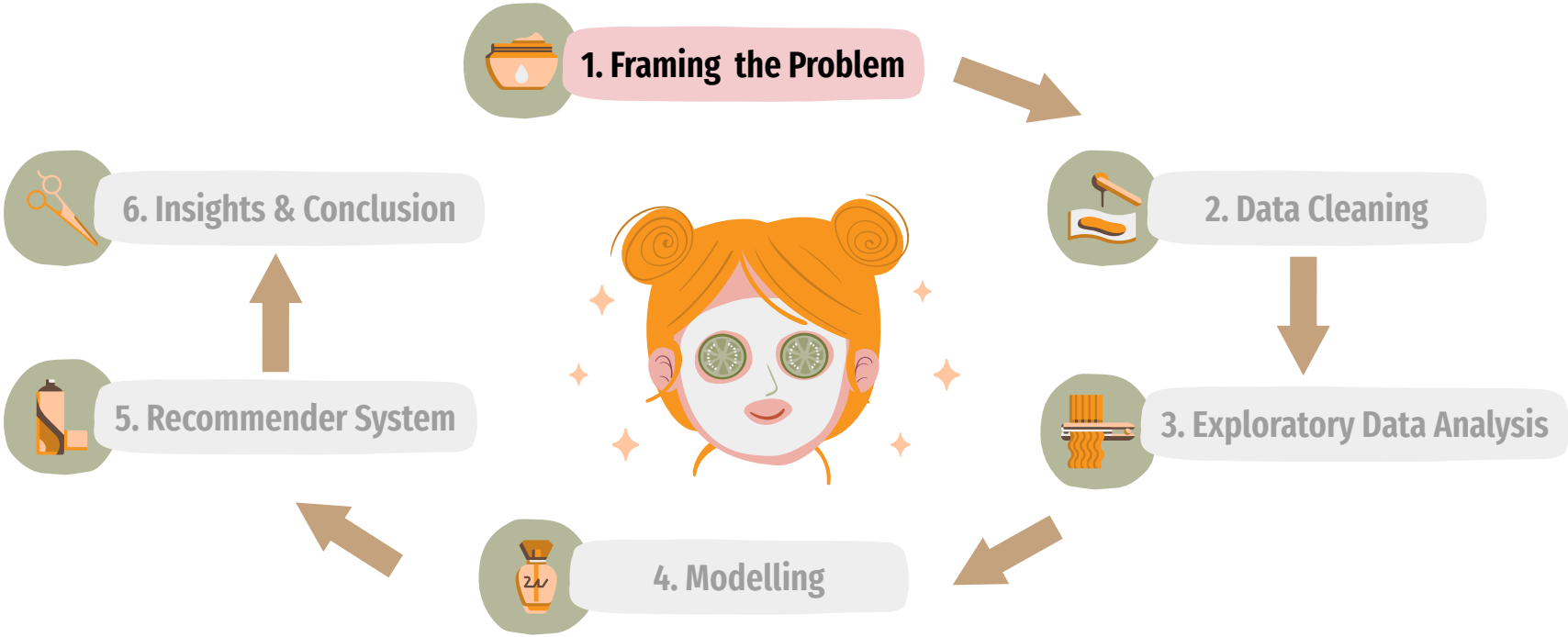
DECODING YOUR SKINCARE PRODUCT

Huimin | December 2021

Overview of Approach

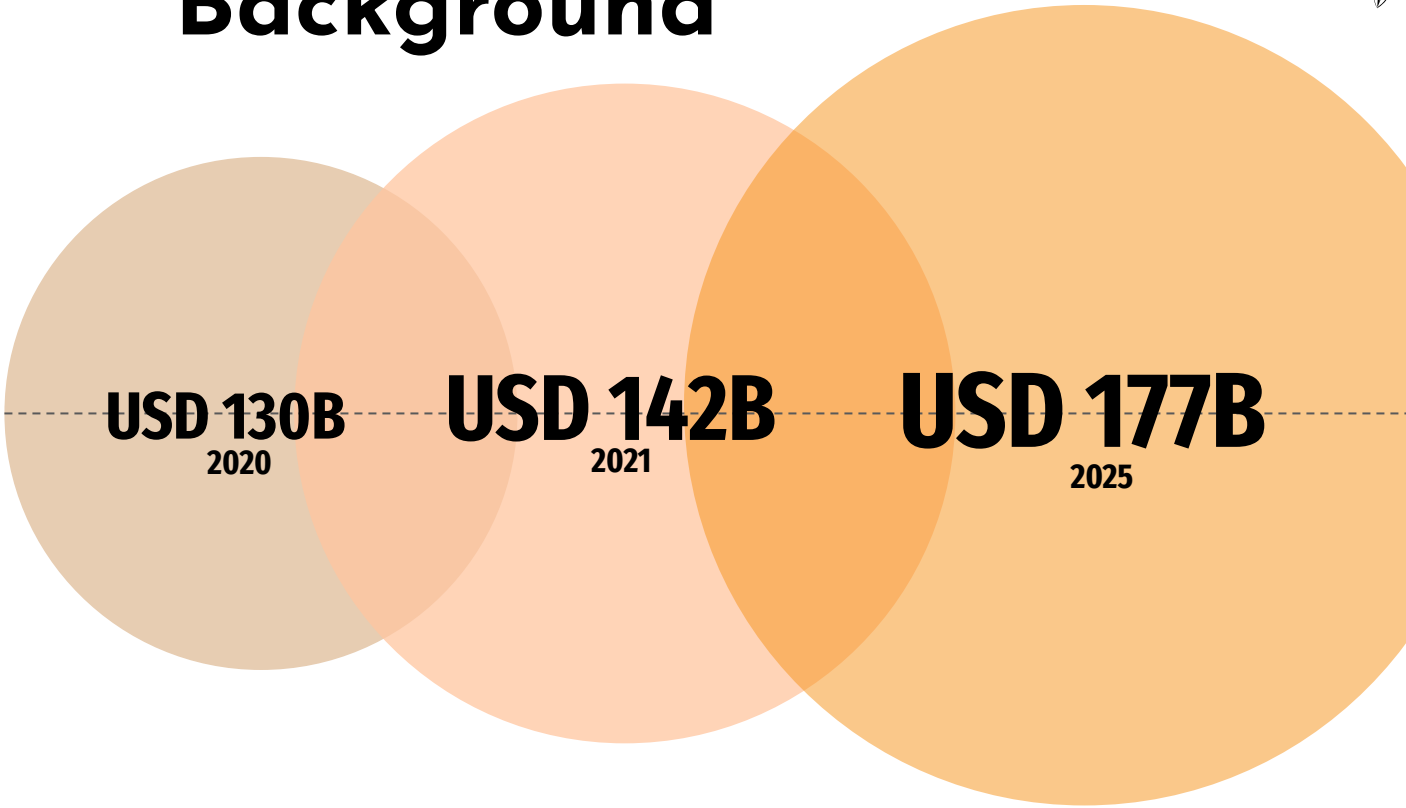


Overview of Approach



Background

The global skincare market is projected to increase from USD 130B in 2020 to over USD 177B in revenue by 2025 [Statista, 2021]



[WGSN, 2021]

3 Types of Buyer Personas

The SuperBasics

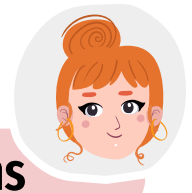
- Draws comfort from simple product choices, embraces frugality and rejects product hypes.
- Price-conscious.
- Measures value by quality and comfort.

$

The Skinimalists

- Values efficacy and efficiency and derives satisfaction from ‘buying well’.
- Product performance is a key motivator.
- Measures value for money by results rather than cost.

$$



The Beautopians

- Beauty is a lifestyle rather than a daily routine.
- Social kudos and product performance are at the top of their lists.
- Most willing to pay for ‘visible results’, brand and luxury.

$$$

[WGSN, 2021]

3 Types of Buyer Personas

The SuperBasics

The Skinimalists

The Beautopians

We want to know what we are buying ...

$

$$

$$$

Problem Statement

1. Help consumers understand what makes a skincare product cheap or expensive so that they can make more informed decisions.

2. Help consumers discover similar alternatives to their tried-and-tested products or even ‘dupe products’.

Overwhelming Choices

Time Expended

Monetary Cost

Disappointing Results

Objectives

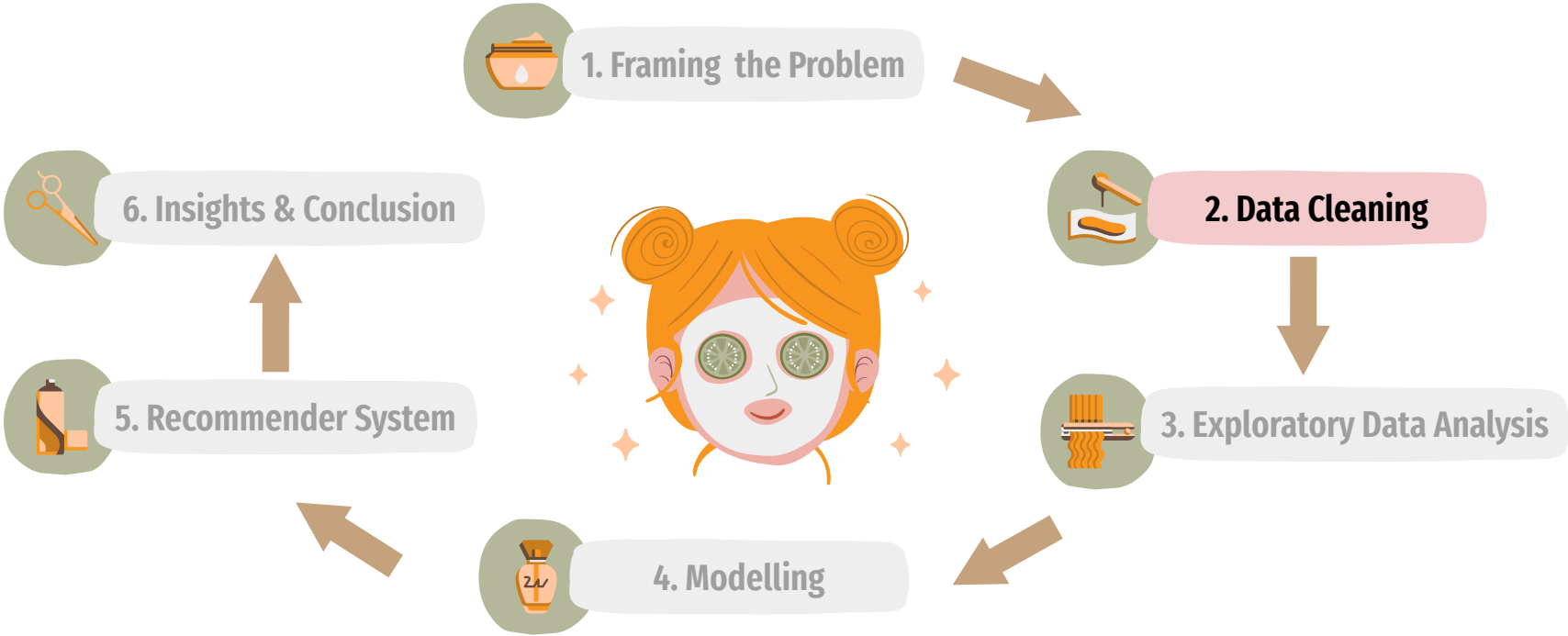
1 Build a 3-class classification model

To identify the product characteristics that affect prices of skincare products by analysing the ingredients, skincare concerns addressed, skintype, brand, etc.

2 Build a simple recommender

To help consumers find the next closest product based on product similarity of ingredients.

Overview of Approach



Sephora's Dataset

- Brand
- Product Name
- Price
- Volume
- About the Product
- Ingredients
- # ❤
- # Reviews
- Rating

CLINIQUE

Acne Solutions™ Cleansing Foam

★★★★☆ 1K Ask a question ♥ 39K

$22.50 or 4 interest-free payments of $5.63 Klarna. ⓘ

Size: 4.2 oz/ 125 mL

About the Product

Item 1027507

What it is: A cleansing foam formulated for acne-prone skin that helps clear and target the look of blemishes and pores.

Skin Type: Normal, Dry, Combination, and Oily

Skincare Concerns: Pores, Acne and Blemishes, and Oiliness

Formulation: Lightweight Liquid

Ingredients

-Salicylic Acid and Acetyl Glucosamine: Help clear dead skin cells that can contribute to clogged pores.
-Lamincaria Sacharina Extract: Diminishes shine and addresses future breakouts.
-Caffeine, Sucrose, Sea Whip Extract, and Kola Nut Extract: Help soothe and calm redness associated with breakouts.

Water, Glycerin, Butylene Glycol, Sodium Methyl Cocoyl Taurate, Sucrose, Disodium Phosphate, Arginine Cocoate, Laminaria Saccharina Extract, Caffeine, Algae Extract, Cola Acuminata (Kola) Seed Extract, Sea Whip Extract, Peg/Ppg-18/18 Dimethicone, Sodium Hyaluronate, Ppg-6-Decyltetradeceth-30, Lactobacillus Ferment, Stearamidopropyl Dimethylamine, Acetyl Glucosamine, Capryloyl Glycine, 10-Hydroxydecanoic Acid, Polyquaternium-7, Phospholipids, Stearic Acid, Sodium Hydroxide, Disodium Edta, Phenoxyethanol, Chloroxylenol.

[Kaggle \ Sephora, 2020]

Data Cleaning & Pre-processing

Missing Values

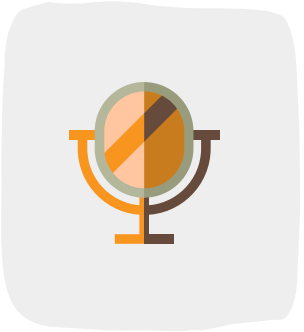
- Removed products without ‘ingredients’.
- Removed products with no ‘size’ like gift or bundle sets.

Duplicate Products

- Removed duplicated products.
- Removed product refills.

Inconsistent Values

- Standardized ‘size’ values by converting to the same volumetric unit (millilitre).



Drop Features

- Dropped one feature of each pair of highly correlated features.
- Dropped features with low variance.

Text Data Pre-processing

Ingredients List

-Salicylic Acid and Acetyl Glucosamine: Help clear dead skin cells that can contribute to clogged pores.
-Lamincaria Sacharina Extract: Diminishes shine and addresses future breakouts.
-Caffeine, Sucrose, Sea Whip Extract, and Kola Nut Extract: Help soothe and calm redness associated with breakouts.

Water, Glycerin, Butylene Glycol, Sodium Methyl Cocoyl Taurate, Sucrose, Disodium Phosphate, Arginine Cocoate, Laminaria Saccharina Extract, Caffeine, Algae Extract, Cola Acuminata (Kola) Seed Extract, Sea Whip Extract, Peg/Ppg-18/18 Dimethicone, Sodium Hyaluronate, Ppg-6-Decyltetradeceth-30, Lactobacillus Ferment, Stearamidopropyl Dimethylamine, Acetyl Glucosamine, Capryloyl Glycine, 10-Hydroxydecanoic Acid, Polyquaternium-7, Phospholipids, Stearic Acid, Sodium Hydroxide, Disodium Edta, Phenoxyethanol, Chloroxylenol.

Regular Expression

- **Specify search patterns to extract individual ingredients from ingredients list.**

Text Data Pre-processing

Started with 7,700 unique ingredient terms from 1,500 products

Aqua/Water, Pentylene Glycol, Phenoxyethanol, Propylene Glycol, Sodium PCA, Methylparaben, Allantoin, Cucumis Sativus Fruit Extract/Cucumber Fruit Extract, Aloe Barbadensis/Aloe Barbadensis Leaf Juice, Camphor, Chamomilla Recutita Extract/Matricaria Flower Extract, Urtica Dioica/Nettle Extract, Equisetum Arvense Extract, Betula Alba Extract, CI 19140/Yellow 5, CI 14700/Red 4, CI 42090/Blue 1.

Pyrus Malus (Apple) Fruit Extract, Anthemis Nobilis (Chamomile) Flower Extract, Narcissus Tazetta Bulb Extract, Caffeine, Sodium Pca, Tocopheryl Acetate, Phytosphingosine, Trehalose, Glycine Soja (Soybean) Seed Extract, Isopropyl Jojobate, Betula Alba (Birch) Extract, Peg/Ppg-18/18 Dimethicone, Ethylhexylglycerin, Gelidiella Acerosa Extract, Tromethamine, Polysorbate 80, Artemia Extract, Hydrolyzed Algin, Isohexadecane, Jojoba Alcohol, Jojoba Esters, Glycerin, Acrylates/C10-30 Alkyl Acrylate Crosspolymer, Hydrogenated Lecithin, Polysorbate 40, Caprylyl Glycol, Sodium Lactate, Lecithin, Bht, Potassium Sorbate, Phenoxyethanol, Iron Oxides (Ci 77491), Iron Oxides

Betula alba birch bark extract

Betula alba extract

Betula alba bark extract

Betula alba birch extract

Betula alba bark extract

Text Data Pre-processing

# of unique ingredient terms reduced from 7,700 to 5,100

FuzzyWuzzy Library

- **Reduces the number of ingredients in the ingredients list by using the Fuzzy Matching (also called Approximate String Matching) technique.**
- **Computes the standard Levenshtein distance similarity ratio between two sequences.**
- **Ingredients terms with at least 80 out of 100 similarity score, will be matched and remapped as a single ingredient term.**

Overview of Features

Target Variable: Price
'cheap' : < USD 35
'average': USD35 - 57
'expensive' : >USD57

General

- Brand
- Category
- Formulation Type
- Size/ Volume

Popularity

- Rating
- # of love
- # of reviews
- Awards won or not

Preferences

- Vegan product
- Gluten-free
- Cruelty-free product
- Published clinical results

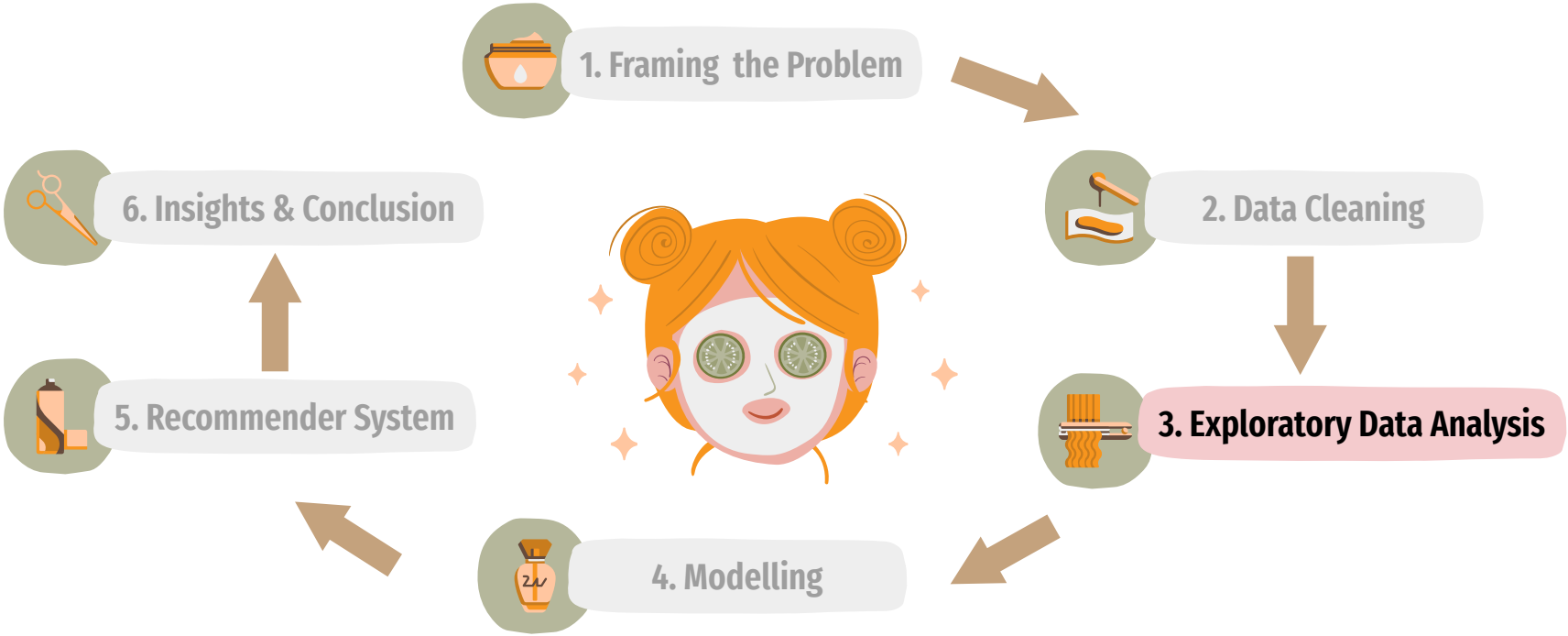
Concerns

- Skincare concern(s) addressed
- Skin type(s) the product is suitable for

Ingredients

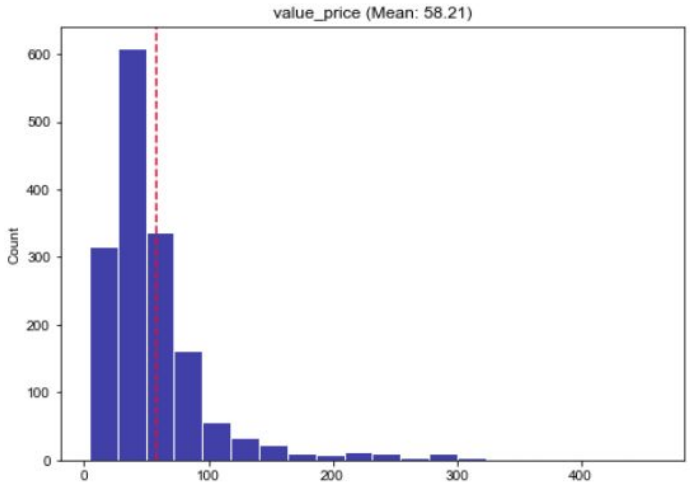
- Full ingredient list
- Special acids used
- Excluded ingredients

Overview of Approach



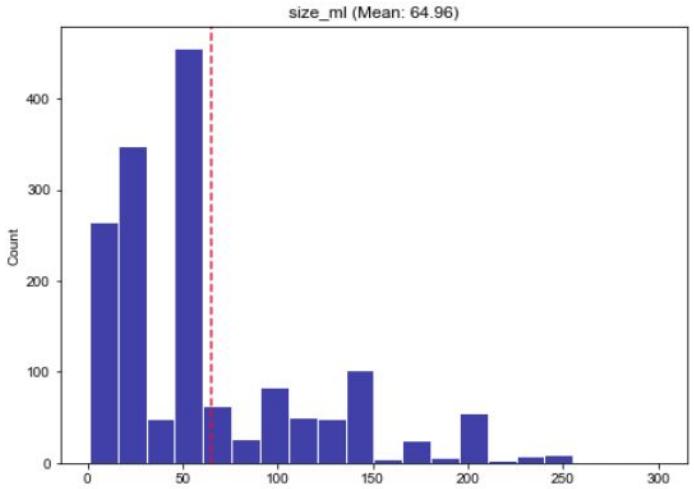
Data Visualisation

Distribution of Price



- **Distribution of price is right-skewed, with mean at USD 58**

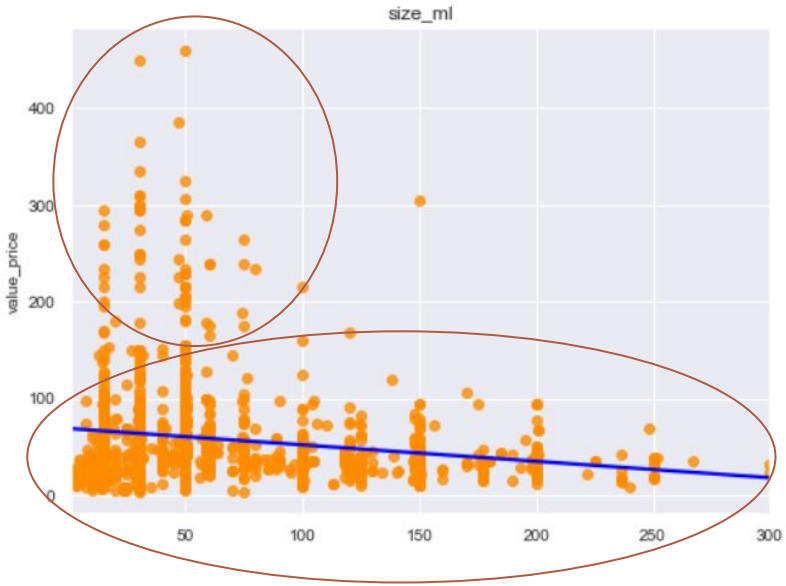
Distribution of Size



- **Typical size of the a product is about 65ml.**

Data Visualisation

Price against Volume

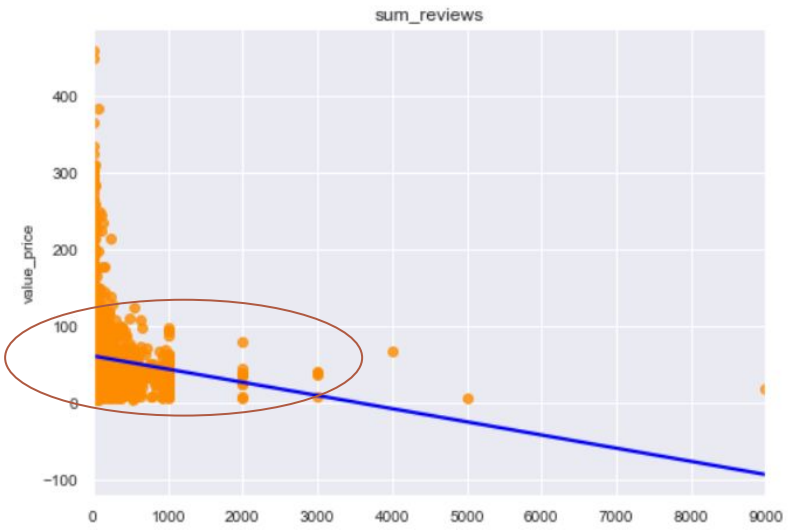


- Expensive products typically come in smaller volumes.

- Cheaper products could vary in volumes.

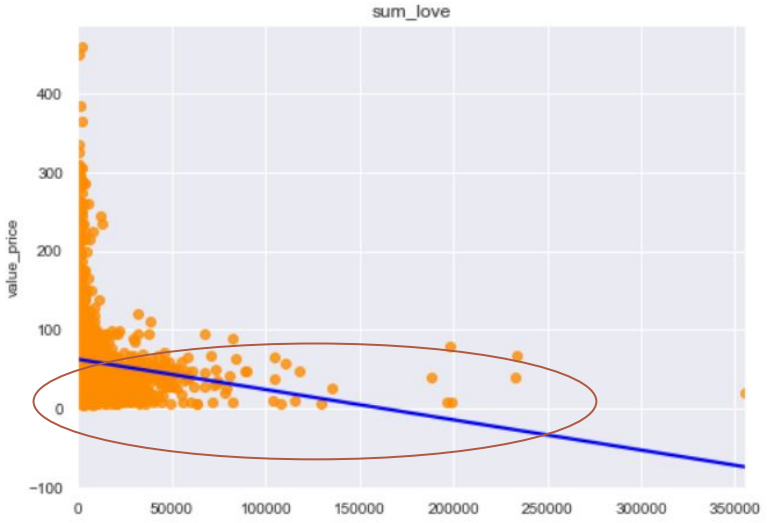
Data Visualisation

Price against # Reviews



Cheaper products garnered more reviews.

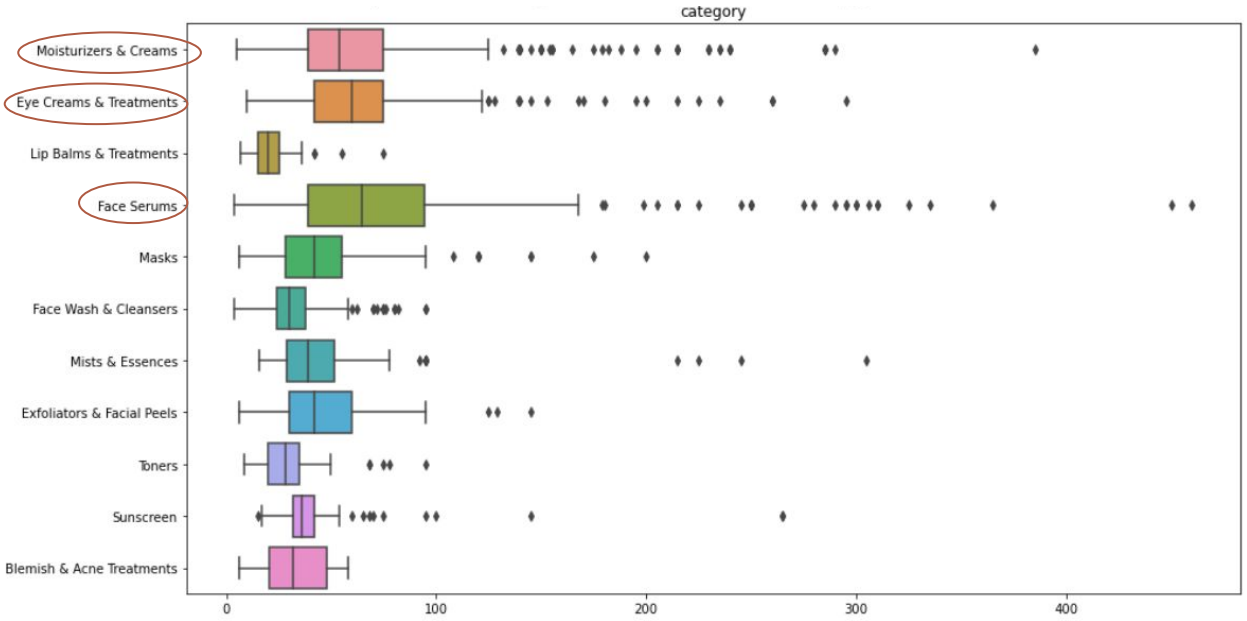
Price against #❤



Cheaper products are more ‘loved’.

Data Visualisation

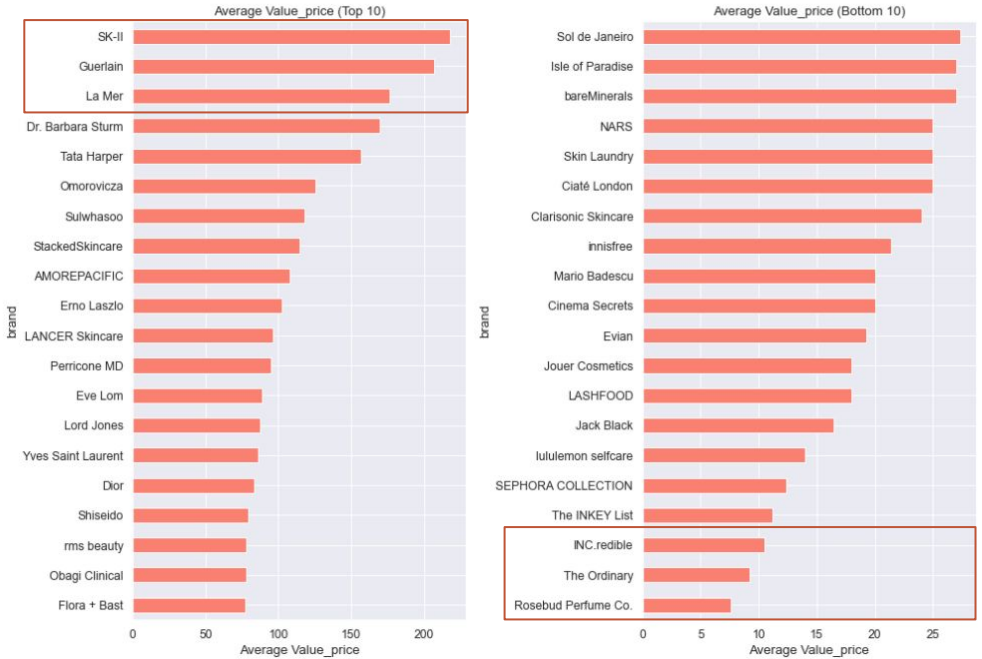
Category against Price



Certain product categories like moisturizers and eye creams, eye creams & treatments, and face serums have the widest price range and more outliers.

Data Visualisation

Brand against Average Price



Top 3 most expensive brands are:

1) **SK-II**
2) **Guerlain**
3) **La Mer**

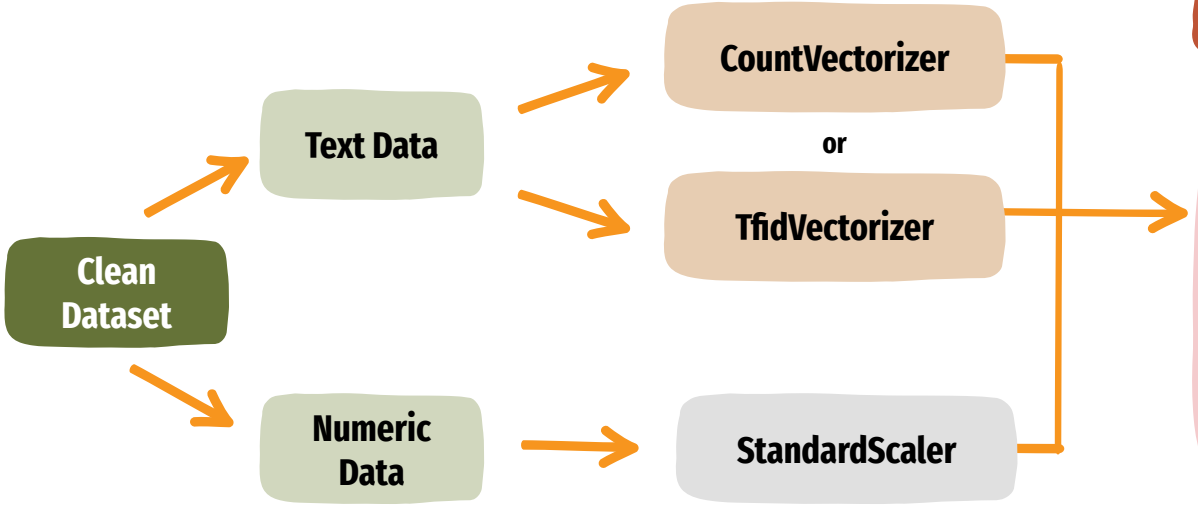
Top 3 cheapest brands are:

1) **Rosebud Perfume Co.**
2) **The Ordinary**
3) **INC.redible**

Overview of Approach

1. Framing the Problem
2. Data Cleaning
3. Exploratory Data Analysis
4. Modelling
5. Recommender System
6. Insights & Conclusion

Modelling Approach



7 CLASSIFICATION MODELS

1) Multi-class Logistic Regression
2) K-Nearest Neighbors
3) Random Forest Classifier
4) Support Vector Machine
5) Adaptive Boosting Classifier
6) Gradient Boosting Classifier
7) Extreme Gradient Boosting Classifier

- Initial run on default hyperparameters
- Subsequent runs using gridsearchCV for tuning hyperparameters

Model Results

[Goal: 0.70]

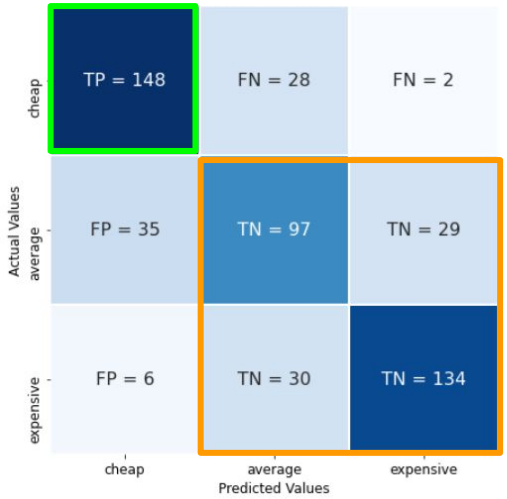
Metrics	Dummy Classifier	Logistic Regression	SVM	AdaBoost	XGBoost	KNN	Gradient Boost	Random Forest
Vectorizer	NA	Tvec	Tvec	Cvec	Cvec	Cvec	Cvec	Cvec
Train Accuracy:	0.350	0.962	0.853	0.753	1.00	1.00	0.900	0.771
Test Accuracy:	0.350	0.745	0.734	0.713	0.703	0.682	0.682	0.623

Accuracy: $\frac{TP + TN}{TP+FP+FN+TN}$

Best Accuracy → Worst Accuracy

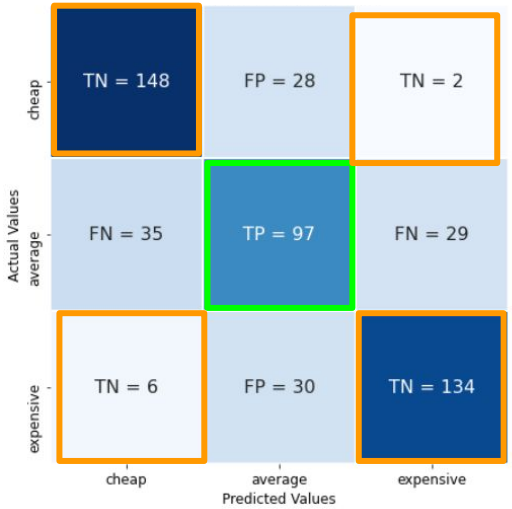
Confusion Matrix

Cheap (< USD35)



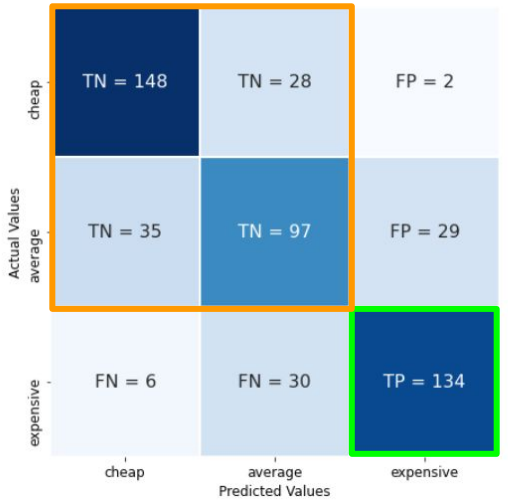
TP : 148 TN : 290

Average (USD35 - 57)



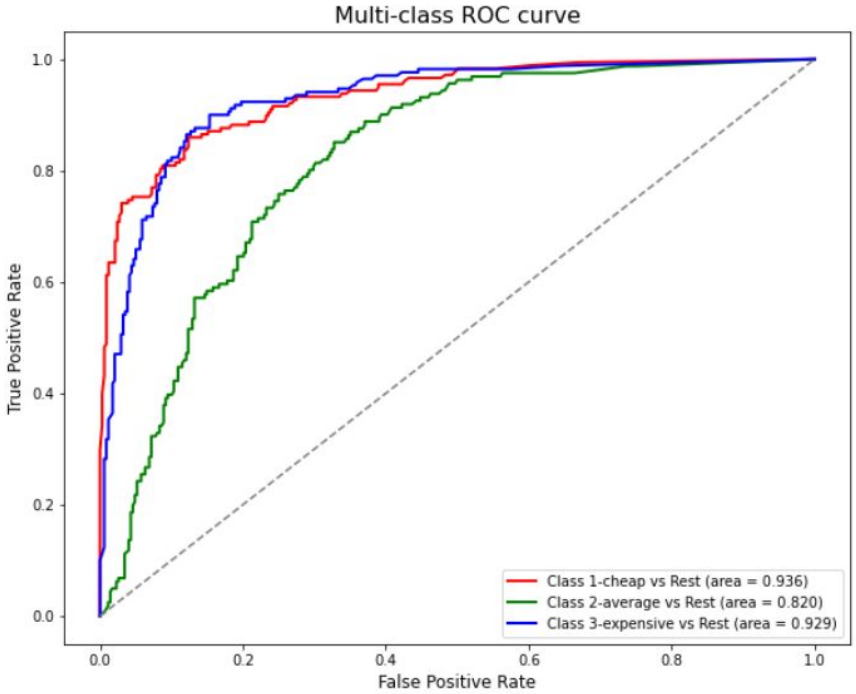
TP : 97 TN : 290

Expensive (> USD57)



TP : 134 TN: 308

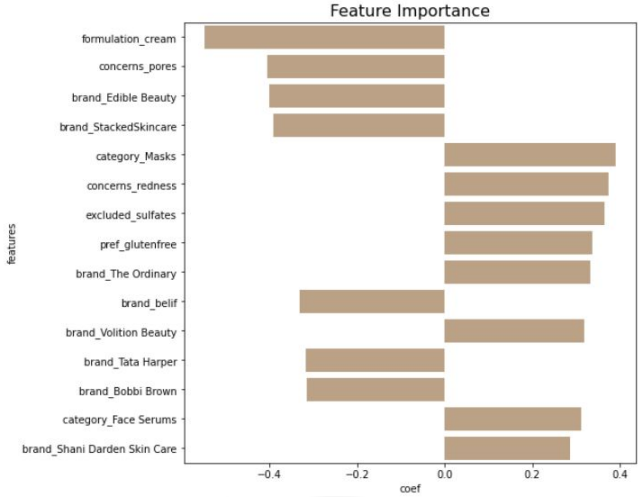
Area Under the Curve of Receiver Operating Characteristic (ROC)



The model performs better at distinguishing between the positive and negative classes for cheap and expensive class , as compared to average class.

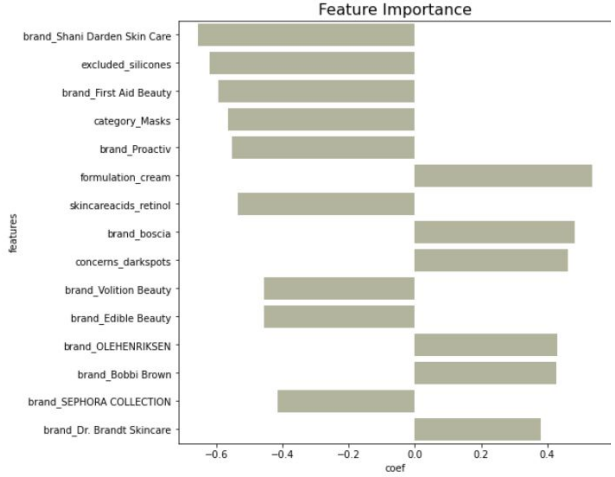
Predictors of Product Price

Cheap (< USD35)



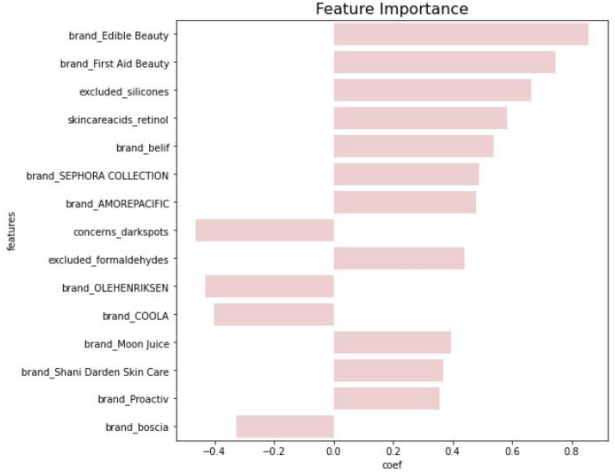
Brands: The Ordinary, Volition Beauty, Shani Darden Skin Care
Categories: Masks, Face Serums
Product attributes: Targets redness, gluten-free, sulfates-free

Average (USD35 - 57)



Brands: Boscia, OLEHENRIKSEN, BobbiBrown, Dr. Brandt Skincare
Product attributes: Cream formulation, targets darkspots

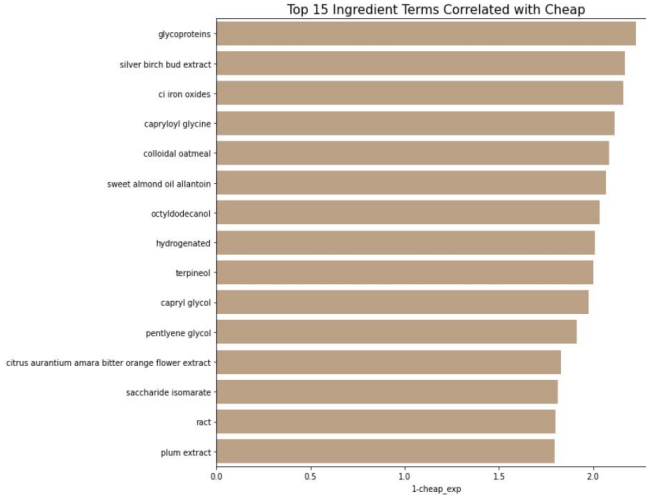
Expensive (> USD57)



Brands: Boscia, Edible Beauty, First Aid Beauty, belif, SEPHORA COLLECTION, AMOREPACIFIC, Proactiv
Product attributes: Silicone-free, formaldehydes-free, retinol

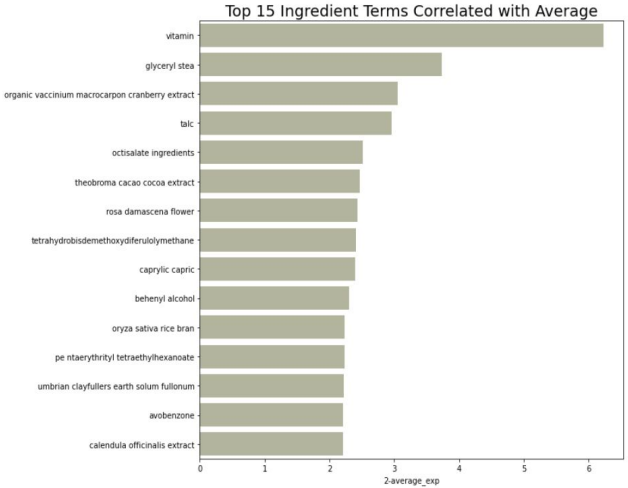
Importance of Ingredient Terms

Cheap (< USD35)



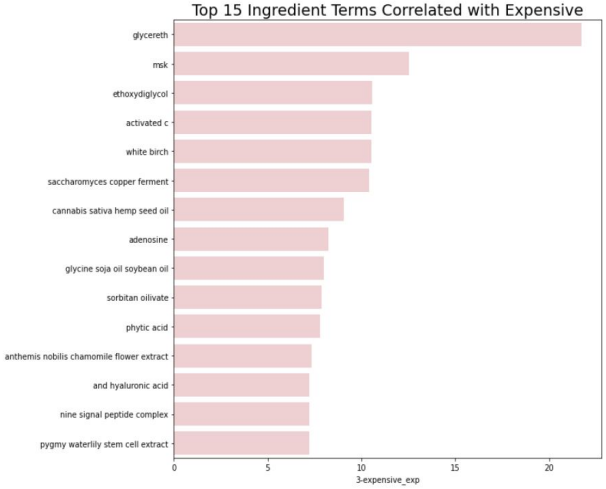
More alcohol derivatives:
octyldodecanol, terpineol, caprylyl glycol
Humectants:
glycoproteins, saccharide isomerate, capryloyl glycine, pentylene glycol

Average (USD35 - 57)



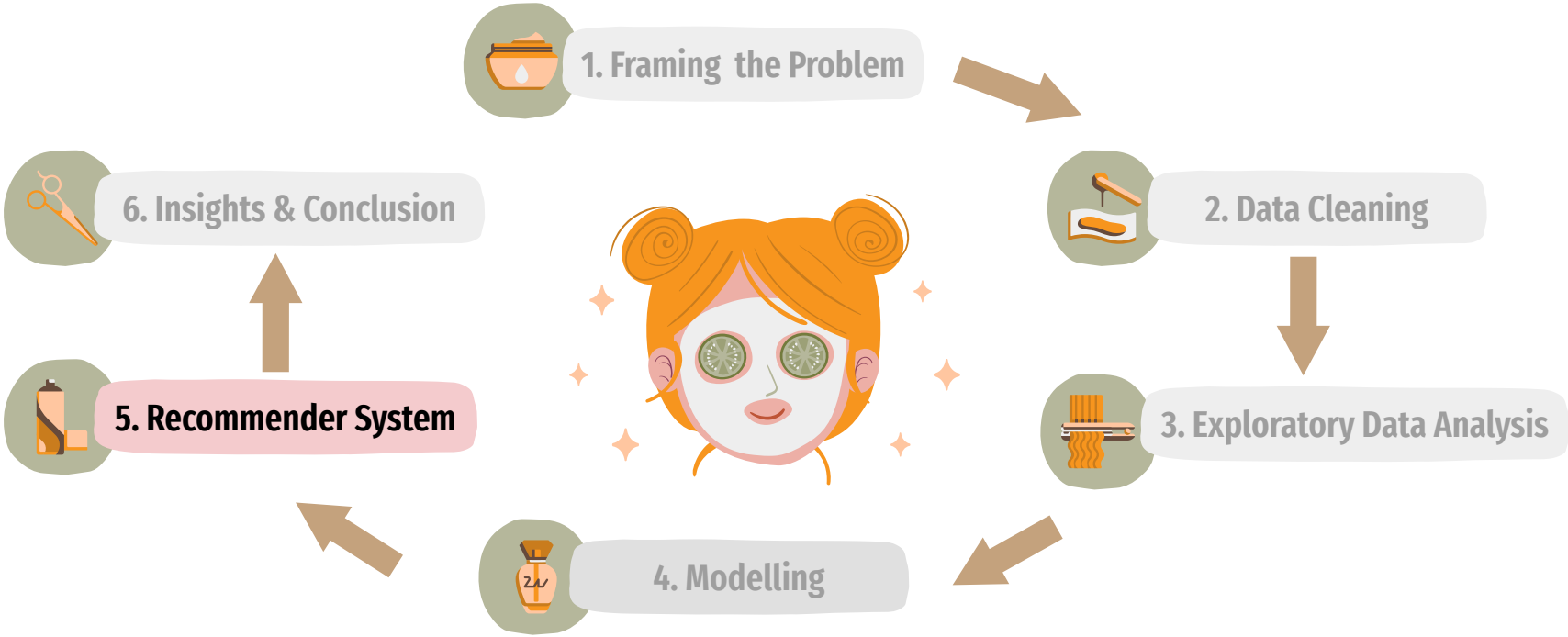
Alcohol derivatives: behenyl alcohol
Vitamins/ minerals: vitamins, clay minerals
Plant-based acids/ oils/ extracts:
rice bran, cacao cocoa extract, calendula extracts, turmeric

Expensive (> USD57)



Ingredients that are known to have higher efficacy: activated vitamin C
Plant-based acids/ oils/ extracts:
phytic acid, hemp seed oil, soybean oil, chamomile flower extract

Overview of Approach



Approach to Create Recommender System

1 Tokenized ingredients

- ***Ingredients:*** *niacin, water, decyl oleate, sh-polypeptide-1*

2 Create 'product-ingredient' matrix using CountVectorizer

	water	niacin	lanolin alcohol	Decyl oleate	serine	...	sh-polypeptide-1
Cosmetic 1	1	1	0	1	0	...	1

3 Create 'product-product' cosine similarity matrix

	Cosmetic 1	Cosmetic 2	Cosmetic 3	Cosmetic 4	...	Cosmetic N
Cosmetic 1	1					
Cosmetic 2		1				
Cosmetic 3			1			
⋮				1		
Cosmetic M					1	

# of products (rows) / *# of products* (columns)

4 Make queries

	Ingredient 1	Ingredient 2	Ingredient 3	Ingredient 4	...	Ingredient N
Cosmetic 1						
Cosmetic 2						
Cosmetic 3						
⋮						
Cosmetic M						

# of products (rows) / *# of the ingredients* (columns)

Recommender System - Results

1 Query

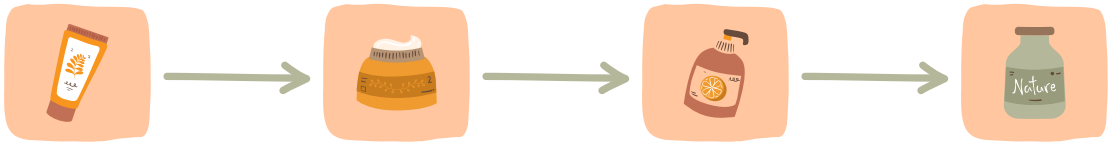
Find similar products to a product

2 Input one product

[CLINIQUE Moisture Surge CC Cream Hydrating Colour Corrector Broad Spectrum SPF 30]

3 Input # of matched products you want

[2 products]



4 Results of Query

	product_name	CLINIQUE Moisture Surge CC Cream Hydrating Colour Corrector Broad Spectrum SPF 30
CLINIQUE Moisture Surge CC Cream Hydrating Colour Corrector Broad Spectrum SPF 30		1.000000
CLINIQUE Age Defense BB Cream Broad Spectrum SPF 30		0.750000
Smashbox Camera Ready BB Cream SPF 35		0.589165

Are you sure they are similar?

Cosine Similarity: 0.589

USD 39.50

-Octinoxate 7.50%, Octisalate 5.00%, Titanium Dioxide 5.20%, Zinc Oxide 3.20%: Sunscreens.

Water, Squalane, Octyldodecyl Neopentanoate, Butylene Glycol, Propanediol, Glyceryl Stearate, Behenyl Alcohol, Peg-40 Stearate, Polyglyceryl-10 Pentastearate, Aloe Barbadensis Leaf Water, Thermus Thermophillus Ferment, Trehalose, Caffeine, Glycerin, Linoleic Acid, Lecithin, Stearic Acid, Hydrogenated Lecithin, Sorbitol, Polyglyceryl-6 Polyricinoleate, Tocopheryl Acetate, Ammonium Acryloyldimethyltaurate/Vp Copolymer, Sodium Stearoyl Lactylate, Sodium Hyaluronate, Tetrahexyldecyl Ascorbate, Pentaerythrityl Tetra-Di-T-Butyl Hydroxyhydrocinnamate, Isopropyl Titanium Triisostearate, Synthetic Fluorphlogopite, Xanthan Gum, Silica, Alumina, Sodium Dehydroacetate, Tin Oxide, Disodium Edta, Phenoxyethanol, [+/- Mica, Titanium Dioxide (Ci 77891), Iron Oxides (Ci 77492), Iron Oxides (Ci 77491), Iron Oxides (Ci 77499)].

USD 42.00

Cinoxate, Ethylhexyl Salicylate, Titanium Dioxide, Water, Dimethicone, Butylene Glycol, Phenyl Trimethicone, Pentylene Glycol, Glyceryl Stearate, Behenyl Alcohol, Trioctyldodecyl Citrate, Polymethylsilsesquioxane, Octyldodecyl Stearoyl Stearate, PEG-40 Stearate, Polyglyceryl-10 Pentastearate, Ammonium Acryloyldimethyltaurate/Vp Copolymer, Triticum Vulgare (Wheat) Germ Extract, Hordeum Vulgare Extract, Cucumis Sativus (Cucumber) Fruit Extract, Laminaria Saccharina Extract, Pyrus Malus (Apple) Fruit Extract, Scutellaria Baicalensis Root Extract, Polyquaternium-51, Acetyl Hexapeptide-8, Trehalose, Oryzanol, Octyldodecyl Neopentanoate, Urea, Polyglyceryl-6 Polyricinoleate, Propylene Glycol Laurate, Glycyrrhetinic Acid, Tocopheryl Acetate (Vitamin E), Sucrose, Glycerin, Sodium Stearoyl Lactylate, Caffeine, Linoleic Acid, Cholesterol, Lecithin, Squalane, Sodium PCA, Isopropyl Titanium Triisostearate, Propylene Glycol Stearate, Stearic Acid, Polysorbate 20, Sorbitan Laurate, Xanthan Gum, Aluminum Hydroxide, Sodium Hyaluronate, Silica, Disodium EDTA, Pentaerythrityl Tetra-Di-T-Butyl Hydroxyhydrocinnamate, Sodium Dehydroacetate, Chlorphenesin, Zinc Oxide, Iron Oxides, CI 77492, CI 77499, Mica

Other Use Cases for Recommender System

Compare similarity of two products

What is the cosine similarity of Product A and Product B?

Query:
Fresh Black Tea Firming and De-Puffing Eye Cream

Output:
Guerlain Abeille Royale Anti-Aging Eye-Cream

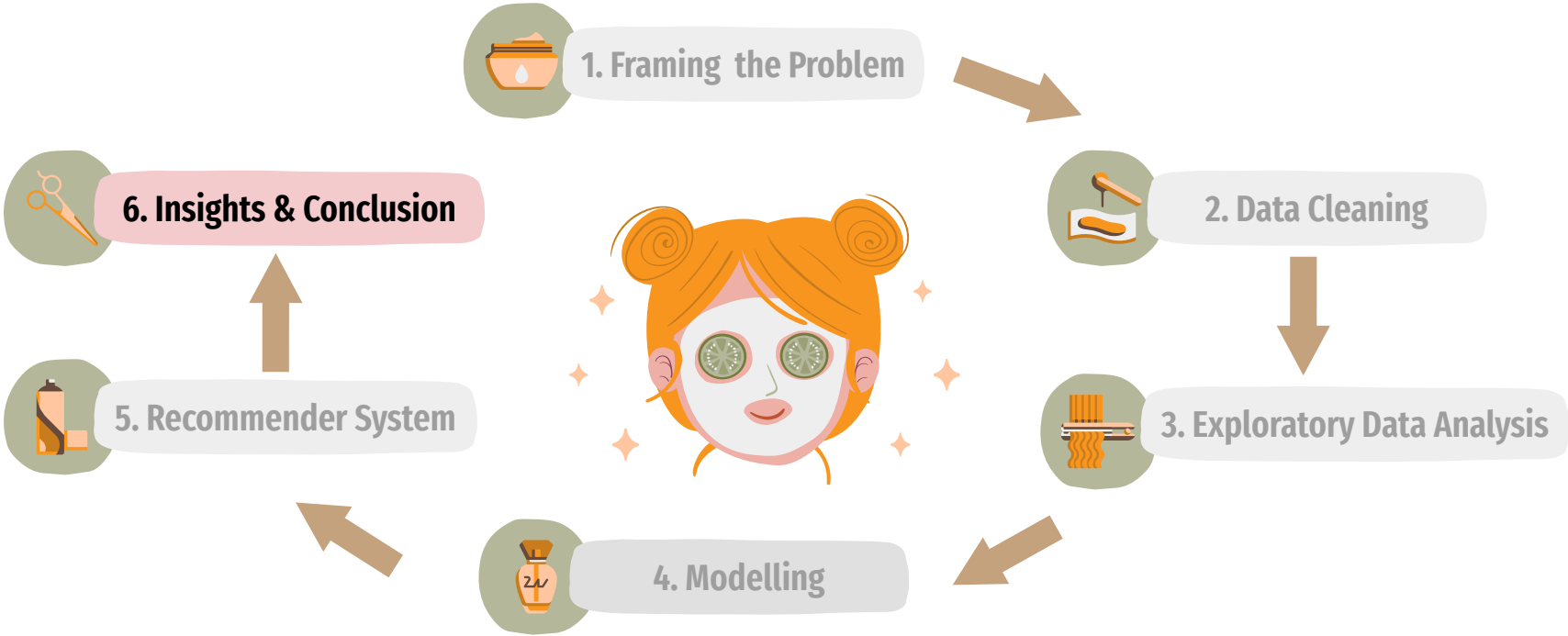
(Cosine Similarity: 0.562978)

Possible 'dupe products'?

What are the cheaper alternatives that promise 'similar results'?

Cosine Similarity : 0.589506	
USD 40 ($$)	**USD 65 ($$$)**
Kiehl's Since 1851 Youth Dose Eye Treatment	**Lancome Visionnaire Eye Cream Advanced Multi-Correcting Eye Balm**
	

Overview of Approach



Summary

Identified determinants of product price

- The **3-class logistic regression model (with TfidfVectorizer)** revealed pricing patterns in the skincare products with it being able to predict price categories with an accuracy of **74.5%.**
- There are **specific brands** that are strong predictors of price for each class.
- Product **category** is more important in predicting cheaper products.
- Products with ingredients that are **non-toxic** or have **anti-aging** properties are more expensive.

Analysed product ingredients

- **Plant-based** ingredients are associated with average- expensive products.
- Cheaper products have **alcohol derivatives**.

Created a simple recommender system

- Query for a product that is **similar to a selected product**.
- **Compare similarity** of two products.
- Possibility of using the recommender to surface **cheaper alternatives** of similar products.

Limitations & Future Considerations

- Proportion of ingredients is also an important factor as the concentration can change the product's effectiveness/ similarity. However, this is considered out-of-scope for this project as the information is not publicly available.

- Generalisability may be limited as only products available on Sephora US are covered. One future consideration would be to include products from Asia and other continents in the analysis.

- Collaboration with cosmetic chemists could facilitate better analysis on the effectiveness of the ingredients.

THANKS!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Appendix

Model Results

	model	vectorizer	cross_val_score	train	test
0	lr	tvec	0.683120	0.962209	0.744597
1	lr	cvec	0.678280	0.944767	0.734774
2	svc	tvec	0.691853	0.852713	0.734774
3	svc	cvec	0.661784	0.812016	0.724951
4	ada	cvec	0.633699	0.752907	0.713163
5	xgb	cvec	0.645307	1.000000	0.703340
6	knn	cvec	0.617237	1.000000	0.681729
7	gbc	cvec	0.633671	0.900194	0.681729
8	xgb	tvec	0.637597	0.988372	0.667976
9	knn	tvec	0.594958	1.000000	0.662083
10	ada	tvec	0.614324	0.752907	0.662083
11	gbc	tvec	0.622086	0.905039	0.636542
12	rfc	cvec	0.604592	0.771318	0.622790
13	rfc	tvec	0.602608	0.753876	0.595285