National College of Ireland

BHSCDA4

Data Analytics

2021/2022

Florence Ogunyoye

X18342776

X18342776@student.ncirl.ie

Flourish Skin

Technical Report

Contents

Executive Summary

Numerous people have various skin issues, and some companies have taken advantage of this reality by promising to offer a cure while charging exorbitant prices for their goods. The objective of this research is to identify the variables that have impacted the costs of the goods that businesses are selling.

This paper describes the knowledge acquisition strategy used by the KDD technique. It describes the steps taken to obtain and prepare the data for analysis. I employed a variety of machine learning approaches for this project, which helped with the outcomes. Among other tools, I have used decision trees, random forests, and logistic regression. Using Python, this analysis was carried out. The datasets that were provided to me were scraped from websites like Cult Beauty and Look Fantastic as well as Kaggle.

The analysis's findings highlight the influencing elements that have an impact on a product's pricing. The study demonstrated how a product's value point is mostly decided by its brand.

1.0 Introduction

1.1. Background

The range and number of skincare products offered by a companies and brands has increased dramatically in recent years. Customers find it difficult to choose which product is best for them while shopping for skincare goods in stores or online since there are so many possibilities. Finding the correct products is not only very expensive, but also time consuming, and above all else, it is constant trial and error trying to find a product that works effectively and does not cause a reaction. It may be difficult to decide on a new cosmetic item to test. We understand the information we require is on each product package, but unless you're educated on the science of cosmetics, it's difficult to comprehend the ingredients lists. You may be able to identify with this circumstance. I've been conducting research on various items and the efficacy of various substances in several products.

Discovering that several companies have made promises that treatments are good for certain skin types but end up having no impact or, worse, making the skin condition worse than it was before. Figuring out which chemicals and ingredients to search for in skincare products and which ones are beneficial has changed your skincare regimen. The skincare and cosmetics industries have exploited brand names to raise the prices of the items they offer. When most consumers end up spending money on skincare and cosmetics, they want high-quality products and formulas that produce the best results. Luxury skincare brands have a vested interest in creating the impression that an increased price tag equates to better quality ingredients and results and saying that they are suited for your skin type when the product's ingredient list contains substances of no value and active ingredients that are below the required % margin to exhibit any genuine impact of action.

Active Ingredients listed on the product can occasionally be found in the lower portion of the ingredient list. The only reason for this is for marketing objectives. Meanwhile, drugstore items are typically less expensive while providing the same, if not higher, quality than high-end ones.

1.2. Aims

The aim of this project is to figure out if a product's components have no discernible impact on its pricing and with doing so finding better alternative for high end products by going through the ingredient list. So, rather than investing and hoping it works, why not utilise data analytics to anticipate which things could be a better match for us? In this project, we will design an evidenced recommendation system based on skin type, price range and with the 'content' being the active ingredients of skincare. This will give users the ability to choose what products they would like to purchase and try out for themselves.

Aim 1: Obtain relevant data from multiple skincare and cosmetics related sources online

Aim 2: Pre- process the data so it is prepared for analysis

Aim 3: Preform data reduction and transformation to uncover relevant information.

Aim 4: Apply algorithms for predictions

Aim 5: Create visualisations of data and findings

1.3. Technology

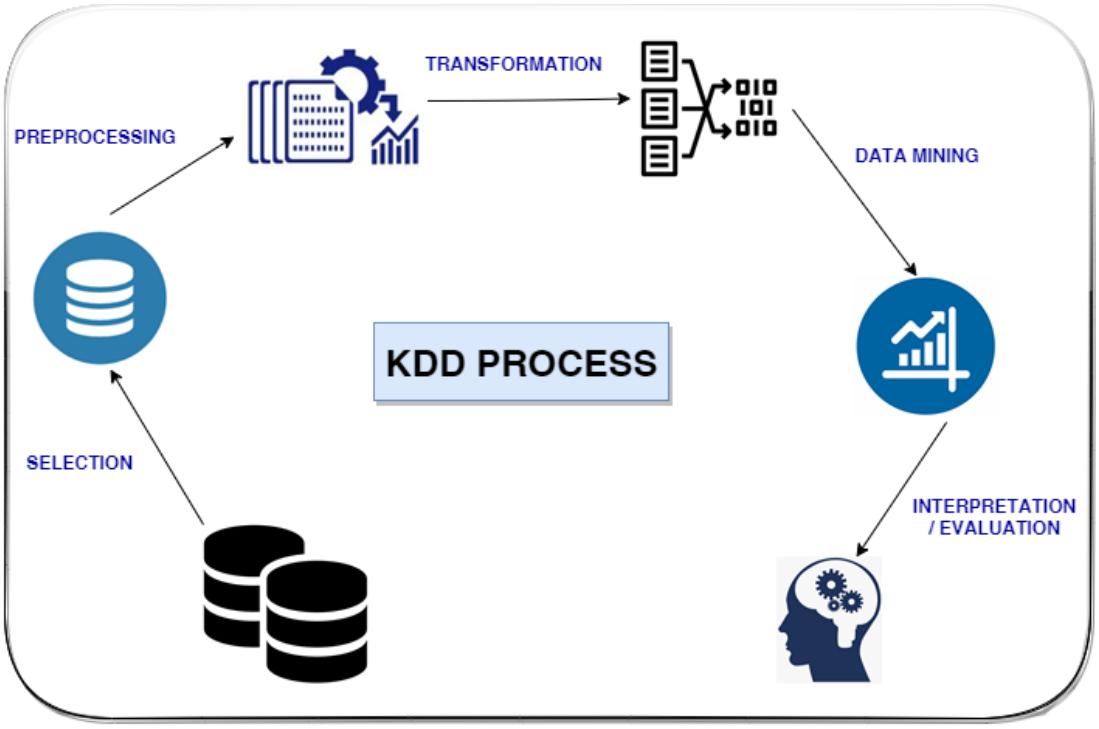
Below I have listed the technologies used for this project:

1. Python Programming Language
   Python is frequently used by software engineers as a support language for build control and administration, testing, and a variety of other tasks. Build control with SCons. For automated continuous compilation and testing, use Buildbot and Apache Gump.
2. Visual Studio IDE - The software used to write and run the code
   A feature-rich tool that covers several elements of software development is known as an integrated development environment (IDE). The Visual Studio IDE provides a creative starting pad for editing, debugging, and building code, as well as publishing an app.
3. Jupyter Notebook - This is a component of the Anaconda framework; it was utilised for overall project implementation locally.
   Jupyter notebook is an open-source IDE which allows users to build and publish Jupyter documents with live code. It is also a web-based dynamic computing environment. The Jupyter notebook can handle a variety of data science languages, including Python, which I will be using.
4. Excel – The data both scrapped for websites and downloaded were exported onto an excel csv file.
   Microsoft Excel is a spreadsheet programme that is part of the Office product group for business applications. In a spreadsheet, Microsoft Excel allows users to arrange, organise, and compute data.
5. Tableau - Tableau is a fantastic business intelligence and data visualisation application for reporting and analysing huge amounts of data. It was founded in America in 2003, and in June 2019, Salesforce bought Tableau. It assists users in producing a variety of graphs, plots, maps, dashboards, charts, and reports for the purpose of visualising and analysing data to aid in corporate decision-making. Tableau is among the most well-liked analytics tools since it offers so many interesting, distinctive features.

For my project, I'll employ the KDD approach, KDD, or knowledge discovery in databases, is a technique for extracting useful information and patterns from a raw database so that they may be used to other fields or applications. The 1989-era Knowledge Discovery in Database

(KDD) bears the name of the general procedure of gathering and methodically enhancing data. The user is presented with various options at every stage of the process, all of which have a substantial impact on the project's conclusion.

KDD frequently derives varying conclusions about how many unique phases are included in its procedure. There are 7 steps involved in the KDD Methodology



Setting goals and comprehending their applications

This is the initial phase in the procedure, which calls for prior comprehension and expertise in the area to be applied in. In this step, we choose how to extract knowledge from the processed data and the patterns identified by data mining. This fundamental assumption is crucial because, if it isn't made correctly, can result in erroneous interpretations and detrimental effects on the end user.

Data Integration and Selection

Once the goals and objectives have been identified, the data must be selected and divided into relevant and usable sets based on availability, usability, relevance, and quality. These qualities are crucial because they provide the basis for data mining and have an effect on the kinds of data models that are produced.

Data Preparation and Cleaning

This step includes searching for missing data as well as getting rid of noisy, redundant, and low-quality data in order to improve the accuracy and reliability of the data collection. Specialized algorithms are used to search for and delete undesired material based on application-specific criteria.

Transformation of Data

The data is prepared in this stage so that the data mining algorithms may use it. As a result, the data must be in aggregated and consolidated formats. Upon that basis of functions, qualities, features and so forth., the data is consolidated.

Data Mining

The foundational or core step of the entire KDD is this. In order to create prediction models, algorithms are utilised here to retrieve relevant trends and patterns from the altered data. With the use of artificial intelligence, sophisticated quantitative statistical approaches, and specialised algorithms, it is an analytical tool that assists in identifying trends from a data collection.

Evaluation and interpretation of patterns

To investigate the effects of data gathered and modified during earlier processes, the trend and patterns must be displayed in distinct forms, such as scatter plots, pie charts, bar charts, and histograms, after being identified by different data mining methods and iterations. This aids in assessing a given data model's efficacy in light of the domain.

Knowledge Discovery and Utilization

The information obtained from the previous phase must be implemented to the specified application or domain in this stage, which is the last stage in the KDD process, and must be presented visually, including tables, reports, or other formats. This action directs how the application's decision will be made.

1.4. Structure

The following is how the rest of the report is organised:

Data: Covers all the data in detail, including a summary of the actual data utilised in the study, as well as where the data was obtained from. Here you'll find data summaries and analyses, as well as figures and all influences on the development. This happens during the data understanding step.

Methodology: this section describes all the pre-processing procedures, including cleaning, transforming, and handling the missing value, also including how the data was later analysed. This procedure is for the project's data preparation step.

Analysis: Addresses all of the methodologies used in the analyzation, including why they were chosen and how they have been applied. This section is part of the CRISP-DM's modelling stage.

Results: After the modelling step is completed, the assessment phase begins. I give the analysis' outcomes and results and assessment metrics, as well as any discoveries and tables and figures, in this part.

Conclusion: The KDD framework wraps up with the Knowledge Discovery and Use phase. The project's primary findings are discussed here, and how they might be used. The project's advantages and disadvantages will also be examined.

Further Development Research: With the increase of time and resources, this section addresses possible future work and progress on the project.

Reference: Summarises the document's references.

Appendices: The appendix of the report.

2.0 Data

For this project's study, two datasets were employed. They were both obtained via Kaggle.com, a service that allows you to post and download various datasets created by other users. I used beautifulsoup to scrape data from web sources and then used excel to create several csv files. This aided the project's analysing process.

The first dataset is the 'Cosmetics' dataset from Kaggle. The information was taken from the skincare items available at Sephora. The collection contains the ingredient listings for 1472 Sephora cosmetics. It shows the pricing of each product and their ingredient list while indicating which skin type the product is appropriate for by using 1s and 0s. it contains 1474 rows and 11 columns. For Analysis, the products suitable for dry skin was filtered into a separate dataset.

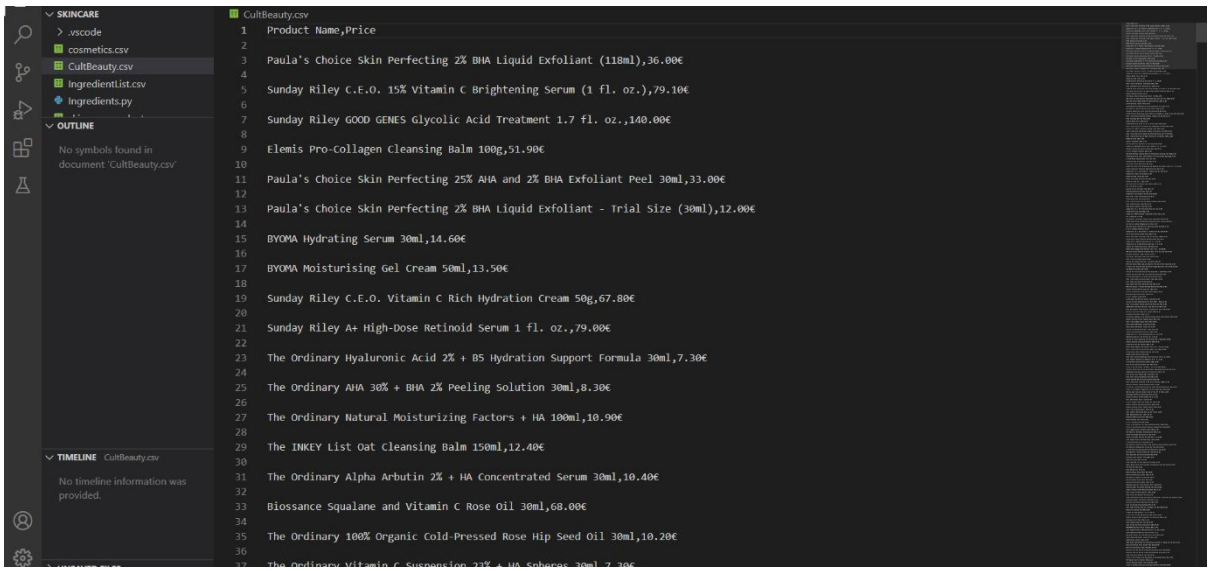
	A	B	C	D	E	F	G	H	I	J	K	L
1	Label	Brand	Name	Price	Rank	Ingredients	Combination	Dry	Normal	Oily	Sensitive	
2	Moisturizer	LA MER	Crème de la Mer	175	4.1	Algae (Seaweed) E:	1	1	1	1	1	
3	Moisturizer	SK-II	Facial Treatment Essence	179	4.1	Galactomyces Ferr	1	1	1	1	1	
4	Moisturizer	DRUNK ELEPHANT	Protini™ Polypeptide Cream	68	4.4	Water, Dicaprylyl C	1	1	1	1	0	
5	Moisturizer	LA MER	The Moisturizing Soft Cream	175	3.8	Algae (Seaweed) E:	1	1	1	1	1	
6	Moisturizer	IT COSMETICS	Your Skin But Better™ CC+™ Cream with SPF 50+	38	4.1	Water, Snail Secret	1	1	1	1	1	
7	Moisturizer	TATCHA	The Water Cream	68	4.2	Water, Saccharom	1	0	1	1	1	
8	Moisturizer	DRUNK ELEPHANT	Lala Retro™ Whipped Cream	60	4.2	Water, Glycerin, Ca	1	1	1	1	0	
9	Moisturizer	DRUNK ELEPHANT	Virgin Marula Luxury Facial Oil	72	4.4	100% Unrefined Sc	1	1	1	1	0	
10	Moisturizer	KIEHL'S SINCE 1851	Ultra Facial Cream	29	4.4	Water, Glycerin, Cy	1	1	1	1	1	
11	Moisturizer	LA MER	Little Miss Miracle Limited-Edition Crème de la Mer	325	5	Algae (Seaweed) E:	0	0	0	0	0	
12	Moisturizer	FRESH	Lotus Youth Preserve Moisturizer	45	4.3	Water, Glycerin, Pr	0	0	0	0	0	
13	Moisturizer	KIEHL'S SINCE 1851	Midnight Recovery Concentrate	47	4.4	Caprylic/Capric Trij	1	1	1	1	1	
14	Moisturizer	BELIF	The True Cream Aqua Bomb	38	4.5	Water, Dipropylene	1	0	1	1	0	
15	Moisturizer	SUNDAY RILEY	Luna Sleeping Night Oil	105	4.1	Persea Gratissima (	1	1	1	1	1	
16	Moisturizer	FARMACY	Honeymoon Glow AHA Resurfacing Night Serum with Echinacea Gr	58	4.6	Water, Lactic Acid,	1	1	1	1	1	
17	Moisturizer	DRUNK ELEPHANT	The Littles™	90	4.4	Beste™ No.9 Jelly (	1	1	1	1	0	
18	Moisturizer	FIRST AID BEAUTY	Ultra Repair® Cream Intense Hydration	30	4.6	Water, Stearic Acic	1	1	1	1	1	
19	Moisturizer	CLINIQUE	Moisture Surge 72-Hour Auto-Replenishing Hydrator	39	4.4	Water , Dimethicor	1	1	1	1	1	
20	Moisturizer	FRESH	Rose Deep Hydration Moisturizer	40	4.4	Water, Glycerin, Et	0	0	0	0	0	
21	Moisturizer	SK-II	R.N.A. POWER Face Cream	230	4.3	Water, Glycerin, Ga	0	1	1	0	1	
22	Moisturizer	LA MER	Crème de la Mer Mini	85	4.1	Algae (Seaweed) E:	1	1	1	1	1	
23	Moisturizer	BAREMINERALS	COMPLEXION RESCUE™ Tinted Moisturizer Broad Spectrum SPF 30	30	3.9	Water, Coconut Alt	0	0	0	0	0	
24	Moisturizer	SHISEIDO	Bio-Performance Advanced Super Revitalizing Cream	78	4.6	Water, Glycerin, Cy	0	0	0	0	0	
25	Moisturizer	FRESH	Black Tea Firming Overnight Mask	92	4.1	Water, Glycerin, Bu	1	1	1	0	0	
26	Moisturizer	BELIF	The True Cream Moisturizing Bomb	38	4.6	Water, Glycerin, Cy	0	1	1	0	0	

The second dataset comes from Kaggle as well. 'Skincare Products' has 1139 rows and 5 columns, including the product name and ingredient list, as well as the skincare categories it belongs to and the pricing list.

	A	B	C	D	E	F	G
1	product_name	product_url	product_type	ingredients	price		
2	The Ordinary Natural Moisturising Factors + HA 30ml	https://www.lookfantastic.	Moisturiser	Aqua (Water), Capr	£5.20		
3	CeraVe Facial Moisturising Lotion SPF 25 52ml	https://www.lookfantastic.	Moisturiser	Aqua/Water, Homo	£13.00		
4	The Ordinary Hyaluronic Acid 2% + B5 Hydration Support Formul	https://www.lookfantastic.	Moisturiser	Aqua (Water), Sodi	£6.20		
5	AMELIORATE Transforming Body Lotion 200ml	https://www.lookfantastic.	Moisturiser	Aqua/Water/Eau, A	£22.50		
6	CeraVe Moisturising Cream 454g	https://www.lookfantastic.	Moisturiser	Purified Water, Gly	£16.00		
7	CeraVe Moisturising Lotion 473ml	https://www.lookfantastic.	Moisturiser	Aqua (Water), Glyc	£15.00		
8	CeraVe Facial Moisturising Lotion No SPF 52ml	https://www.lookfantastic.	Moisturiser	Purified Water, Gly	£13.00		
9	The Ordinary Natural Moisturizing Factors + HA 100ml	https://www.lookfantastic.	Moisturiser	Aqua (Water), Capr	£6.80		
10	CeraVe Smoothing Cream 177ml	https://www.lookfantastic.	Moisturiser	Purified Water, Gly	£12.00		
11	Clinique Moisture Surge 72 Hour Moisturiser 75ml	https://www.lookfantastic.	Moisturiser	Water\Aqua\Eau, I	£37.00		
12	CeraVe Moisturising Cream 50ml	https://www.lookfantastic.	Moisturiser	Purified Water, Gly	£4.00		
13	CeraVe Moisturising Cream 340g	https://www.lookfantastic.	Moisturiser	Purified Water, Gly	£13.00		
14	First Aid Beauty Ultra Repair Cream (56.7g)	https://www.lookfantastic.	Moisturiser	Colloidal Oatmeal,	£12.00		
15	Avène Antirougeurs Jour Redness Relief Moisturizing Protecting (	https://www.lookfantastic.	Moisturiser	Avene Thermal Spr	£16.00		
16	Clinique Dramatically Different Moisturising Lotion+ 125ml with	https://www.lookfantastic.	Moisturiser	Water\Aqua\Eau, M	£30.00		
17	First Aid Beauty Ultra Repair Cream (170g)	https://www.lookfantastic.	Moisturiser	Colloidal Oatmeal,	£25.00		
18	Weleda Skin Food (75ml)	https://www.lookfantastic.	Moisturiser	Water (Aqua), Helia	£12.95		
19	Neutrogena Hydro Boost City Shield SPF Moisturiser	https://www.lookfantastic.	Moisturiser	Aqua, Glycerin, Hor	£12.99		
20	Egyptian Magic All Purpose Skin Cream 118ml/4oz	https://www.lookfantastic.	Moisturiser	Olive Oil, Beeswax,	£30.00		
21	JASON Aloe Vera 98% Moisturising Gel Tube 113g	https://www.lookfantastic.	Moisturiser	Aloe Barbadensis (	£2.99		
22	CeraVe Moisturising Cream 177ml	https://www.lookfantastic.	Moisturiser	Purified Water, Gly	£9.00		
23	Embryolisse Lait-Crème Concentré (75ml)	https://www.lookfantastic.	Moisturiser	Aqua. Paraffinum L	£20.00		
24	La Roche-Posay Effaclar H Moisturiser 40ml	https://www.lookfantastic.	Moisturiser	Aqua, Squalane, Gl	£9.99		
25	Bulldog Original Moisturiser 100ml	https://www.lookfantastic.	Moisturiser	Aqua (Water), Ethy	£4.50		
26	Clinique Dramatically Different Moisturising Gel 125ml with Pum	https://www.lookfantastic.	Moisturiser	Water\Aqua\Eau, I	£30.00		

skincare_products (1)

After that, I was able to scrape some information from the website 'CultBeauty.com' I then compiled a list of all the items that were advertised as being suited for dry skin, as well as their prices. These datasets were obtained with the aid of beautifulsoup.



The fourth dataset was provided After scraping the active Ingredients from 'Byrdie.com,' I made a list of active Ingredients for further filtering and to enhance the search results.

	A	B	C	D	E
1	ActiveIngredients				
2	Allantoin				
3	Alcohol Denat				
4	Almond Oil				
5	Aloe Vera				
6	Alpha Hydroxy Acid				
7	Amala Oil				
8	Amino Acids				
9	Amoxicillin				
10	Antioxidants				
11	Apple Cider Vinegar				
12	Apricot Kernel Oil				
13	Arbutin				
14	Argan Oil				
15	Argireline				
16	Ascorbyl Glucoside				
17	Astaxanthin				
18	Avocado Oil				
19	Azelaic Acid				
20	Azulene				
21	Baobab				
22	Baking Soda				
23	Bakuchiol				
24	Bentonite Clay				
25	Benzoyl Peroxide				
26	Benzyl Alcohol				

3.0 Methodology

The steps of this analysis' execution follow a design approach of KDD stages. A data mining project may be conceptualised in seven phases and through several cycles depending on the requirements of the developers. This process consists of the following steps: Goal- setting and Application Understanding, Data Selection and Integration, Data Cleaning and Pre-processing, Data transformation, Data Mining, Pattern Evaluation/ Interpretation and lastly Knowledge Discovery and Use.

Goal setting and Application Understanding

For this section of the analysis, I done some research on the cosmetic and skincare market. There have been many studies done the skincare industry and the influencing factors that play into the purchase of products. There is a study done by Zhichao Liu and Jiahui Ling called “Research on Influencing Factors of Purchase Promotion intention based on Skincare Industry”. The skin care industry is used as the research subject in this article, which also examines the factors that influence how much consumers are impressed by promotional activities, makes use of earlier studies on the experience dimension and purchase intention, develops a questionnaire, and carries out focused consumer research. The components with high correlations across each dimension are extracted using factor analysis. In addition, regression analysis is utilised to determine if the four aspects of experiential promotion

influence customers' intentions to make purchases. This ends the key elements influencing customers' intents to make purchases during experiential promotion activities and serves as a guide for the industry's long-term design and research of experiential promotion and beauty products goods.

Another research paper that I found interesting was “The Impact of Globalization in The Industry Of cosmetics”. This paper was conducted by Oana Maria Secara and Dinu Vlad Sasu. Due to its size in the worldwide cosmetic business, this research concentrates on the skincare product market. This essay compares the major cosmetic consumer markets—France, Germany, and the United States—against others that are only now starting to take off in the likes of China, Republic of Korea, and India. They sought to emphasise the significance of the Asian skincare cosmetic industry in the context of the overall skincare market in a progressive and comprehensive manner.

A Study done on Skincare marketing in Malaysia was carried out by Muhammad Tahir Jan, Ahasanul Haque, Kalthom Abdullah, Zohurul Anis, Faisal-E-Alam. The primary goal of this essay was to determine how important advertising variables affected consumers' purchasing decisions about skincare goods. Using self-administered questionnaires, 428 individuals from all throughout Malaysia provided the data for this study. To make sure the data was prepared for analysis in SPSS software, it underwent a thorough screening and cleaning procedure. Numerous reliable experiments were run to arrive at the conclusions. These consist of evaluating your hypothesis as well as frequency tests, reliability tests, exploratory and confirmatory factor analyses. It was determined whether the suggested model was appropriate for purpose using structural equation modelling. The findings showed that two aspects of advertising—their utility and their features—had a considerable beneficial influence on consumers' purchasing decisions. A novel effort was made to find an appropriate nested model where the characteristics of the marketing campaign had a positive emotional substantial influence on the usefulness of the advertisement. Policy makers, particularly those in Malaysia's cosmetic and health industries, can benefit greatly from this research. The results of this research are supposed to be considered while formulating a skincare brand's marketing strategy.

There have been numerous studies and analyses conducted on the skincare market. Inspired by their findings, I made the decision to adopt their methodology and carry out my own research and analysis to determine whether the product is actually what consumers want to purchase or if it is just the brand name. With this in mind, I made the decision to carry out a more in-depth investigation and compare the component lists of higher end cosmetic products with drugstore cosmetic products in order to look for any actual differences. We will be able to determine which aspects contribute to a product's cost due to this research.

Data Selection and Integration

I have researched, gathered, and evaluated the dataset in this area to support the Goal setting and Application Understanding foundation and to assist me reach the objectives of this project. To advance my study, I have accumulated a lot of datasets.

The first dataset is the Kaggle dataset titled "Cosmetics." The details were gathered from the Sephora skincare products.

The second dataset also originates from Kaggle. The list of ingredients, the product name, the skincare categories it falls under, and the price are all included in the 1139 rows and 5 columns that make up the heading "Skincare Products."

After that, I was able to get some data from the 'CultBeauty.com' website. I then made a list of all the products that were described as being appropriate for dry skin, along with their costs. Additionally, I scraped the main ingredients stated in the search criteria on the websites. Beautifulsoup was used to help collect these datasets.

A fourth dataset was made available. I created a list of active Ingredients after scraping the active ingredients from "Byrdie.com" in order to improve the search results and filtering.

Despite the fact that I didn't integrate any of the datasets. I attentively examined each one independently.

Data Cleaning and Pre-processing

One of the crucial phases of this investigation was data pre-processing, which involved tasks including integrating datasets and identifying, repairing, or replacing missing or erroneous data entries. This phase is crucial to ensure that computations and models execute smoothly and to get reliable data. When the data was prepared, it underwent manual pre-processing in Excel before it would be properly pre-processed in Jupyter using Python.

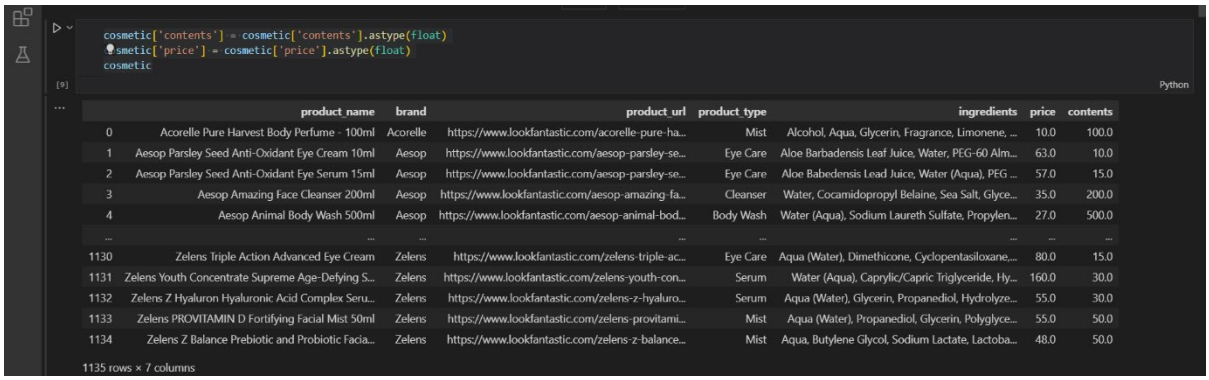
But for the cleaning process of the data, there was not much to be done for this section, as most of the data was scrapped from websites such as ‘Look Fantastic’ and ‘Cult beauty’. It was not difficult to process the data and work with. I did not have to get rid or fix any missing data.

For the sake of independent analysis, the data were not combined. Nevertheless, they were each independently examined before being submitted to further comparison and producing pertinent results.

Data Transformation

Understanding and obtaining the optimal features from the factors that determine the product pricing is crucial to increasing the accuracy of the skincare prediction model and eliminating multicollinearity.

I initially changed the "price" column into float values columns, deleting the "£" sign, to prepare the data and ready it for an algorithm.



```
cosmetic['contents'] = cosmetic['contents'].astype(float)
cosmetic['price'] = cosmetic['price'].astype(float)
cosmetic
```

[9] Python

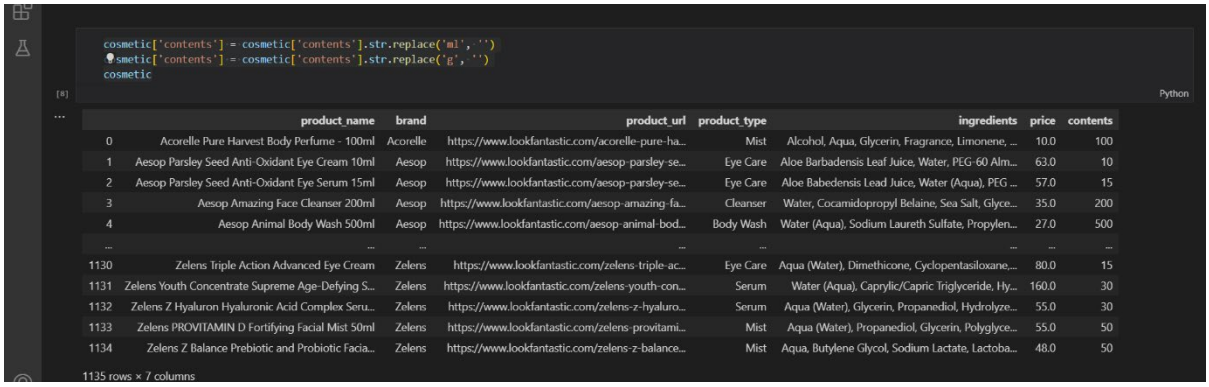
	product_name	brand	product_url	product_type	ingredients	price	contents
0	Acorelle Pure Harvest Body Perfume - 100ml	Acorelle	https://www.lookfantastic.com/acorelle-pure-ha...	Mist	Alcohol, Aqua, Glycerin, Fragrance, Limonene, ...	10.0	100.0
1	Aesop Parsley Seed Anti-Oxidant Eye Cream 10ml	Aesop	https://www.lookfantastic.com/aesop-parsley-se...	Eye Care	Aloe Barbadensis Leaf Juice, Water, PEG-60 Alm...	63.0	10.0
2	Aesop Parsley Seed Anti-Oxidant Eye Serum 15ml	Aesop	https://www.lookfantastic.com/aesop-parsley-se...	Eye Care	Aloe Babedensis Lead Juice, Water (Aqua), PEG ...	57.0	15.0
3	Aesop Amazing Face Cleanser 200ml	Aesop	https://www.lookfantastic.com/aesop-amazing-fa...	Cleanser	Water, Cocamidopropyl Belaine, Sea Salt, Glyce...	35.0	200.0
4	Aesop Animal Body Wash 500ml	Aesop	https://www.lookfantastic.com/aesop-animal-bod...	Body Wash	Water (Aqua), Sodium Laureth Sulfate, Propylen...	27.0	500.0
...	...	...	...	...	...	...	...
1130	Zelens Triple Action Advanced Eye Cream	Zelens	https://www.lookfantastic.com/zelens-triple-ac...	Eye Care	Aqua (Water), Dimethicone, Cyclopentasiloxane,...	80.0	15.0
1131	Zelens Youth Concentrate Supreme Age-Defying S...	Zelens	https://www.lookfantastic.com/zelens-youth-con...	Serum	Water (Aqua), Caprylic/Capric Triglyceride, Hy...	160.0	30.0
1132	Zelens Z Hyaluron Hyaluronic Acid Complex Seru...	Zelens	https://www.lookfantastic.com/zelens-z-hyaluro...	Serum	Aqua (Water), Glycerin, Propanediol, Hydrolyze...	55.0	30.0
1133	Zelens PROVITAMIN D Fortifying Facial Mist 50ml	Zelens	https://www.lookfantastic.com/zelens-provitami...	Mist	Aqua (Water), Propanediol, Glycerin, Polyglyce...	55.0	50.0
1134	Zelens Z Balance Prebiotic and Probiotic Facia...	Zelens	https://www.lookfantastic.com/zelens-z-balance...	Mist	Aqua, Butylene Glycol, Sodium Lactate, Lactoba...	48.0	50.0

1135 rows × 7 columns

Then, utilizing a regex to obtain the millilitre, gramme, or kilogramme quantity from "product name," I created a "contents" column that has the product amount for each item.

I opted to include this "contents" column considering that if the component list seemed to have any predictive value, it stands to reason that the quantity of those ingredients would affect the product price as well. In my dataset, I preserved the columns "brand," "contents," "product type," and "ingredients" as predictive characteristics.

Following that, I deleted all 'ml' and 'g' measures from the "contents" column. The 'contents' and 'price' columns were then converted to float values.



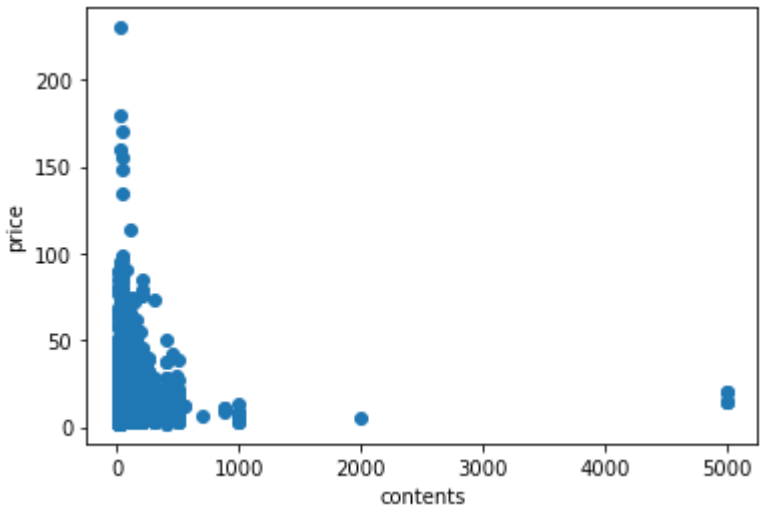
```
cosmetic['contents'] = cosmetic['contents'].str.replace('ml', '')
cosmetic['contents'] = cosmetic['contents'].str.replace('g', '')
cosmetic
```

[8] Python

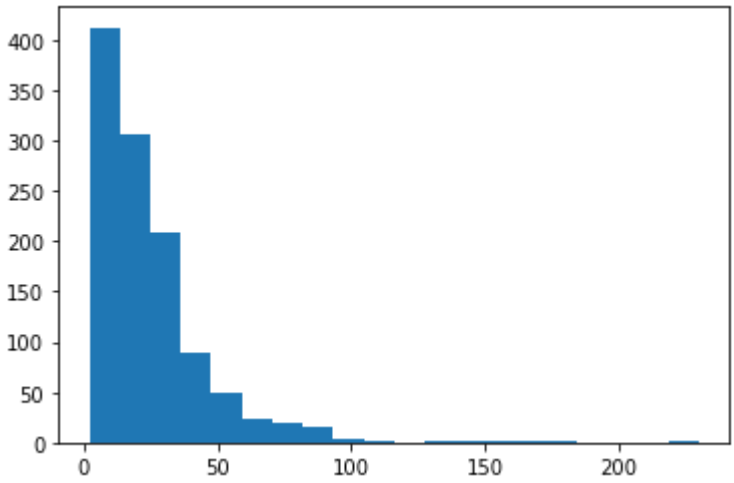
	product_name	brand	product_url	product_type	ingredients	price	contents
0	Acorelle Pure Harvest Body Perfume - 100ml	Acorelle	https://www.lookfantastic.com/acorelle-pure-ha...	Mist	Alcohol, Aqua, Glycerin, Fragrance, Limonene, ...	10.0	100
1	Aesop Parsley Seed Anti-Oxidant Eye Cream 10ml	Aesop	https://www.lookfantastic.com/aesop-parsley-se...	Eye Care	Aloe Barbadensis Leaf Juice, Water, PEG-60 Alm...	63.0	10
2	Aesop Parsley Seed Anti-Oxidant Eye Serum 15ml	Aesop	https://www.lookfantastic.com/aesop-parsley-se...	Eye Care	Aloe Babedensis Lead Juice, Water (Aqua), PEG ...	57.0	15
3	Aesop Amazing Face Cleanser 200ml	Aesop	https://www.lookfantastic.com/aesop-amazing-fa...	Cleanser	Water, Cocamidopropyl Belaine, Sea Salt, Glyce...	35.0	200
4	Aesop Animal Body Wash 500ml	Aesop	https://www.lookfantastic.com/aesop-animal-bod...	Body Wash	Water (Aqua), Sodium Laureth Sulfate, Propylen...	27.0	500
...	...	...	...	...	...	...	...
1130	Zelens Triple Action Advanced Eye Cream	Zelens	https://www.lookfantastic.com/zelens-triple-ac...	Eye Care	Aqua (Water), Dimethicone, Cyclopentasiloxane,...	80.0	15
1131	Zelens Youth Concentrate Supreme Age-Defying S...	Zelens	https://www.lookfantastic.com/zelens-youth-con...	Serum	Water (Aqua), Caprylic/Capric Triglyceride, Hy...	160.0	30
1132	Zelens Z Hyaluron Hyaluronic Acid Complex Seru...	Zelens	https://www.lookfantastic.com/zelens-z-hyaluro...	Serum	Aqua (Water), Glycerin, Propanediol, Hydrolyze...	55.0	30
1133	Zelens PROVITAMIN D Fortifying Facial Mist 50ml	Zelens	https://www.lookfantastic.com/zelens-provitami...	Mist	Aqua (Water), Propanediol, Glycerin, Polyglyce...	55.0	50
1134	Zelens Z Balance Prebiotic and Probiotic Facia...	Zelens	https://www.lookfantastic.com/zelens-z-balance...	Mist	Aqua, Butylene Glycol, Sodium Lactate, Lactoba...	48.0	50

1135 rows × 7 columns

The majority of the goods weighed less than one kilogramme and were priced around £100. There appears to be minimal correlation between product price and product quantity.



The products listed in the dataset had a range of prices, the majority of the products listed were aimed lower than £50 , as seen in the diagram below.



Focusing here on inter-quartile range, I created two price subcategories in binary classification experimentations: "cheap" category 0, under £18.90 and "expensive" class 1, above £18.90 .

To be sure, this difference is imprecise and subjective. However, the purpose of this part was just to determine if any pricing information could be retrieved by combining the variables "brand," "contents," "product type," and "ingredients."

```
print(cosmetic['price'].describe())
```
[46] ✓ 0.6s Python

```
count    1135.000000
mean       23.963419
std        21.713618
min         1.950000
25%         9.950000
50%        18.900000
75%        31.250000
max       230.000000
Name: price, dtype: float64
```

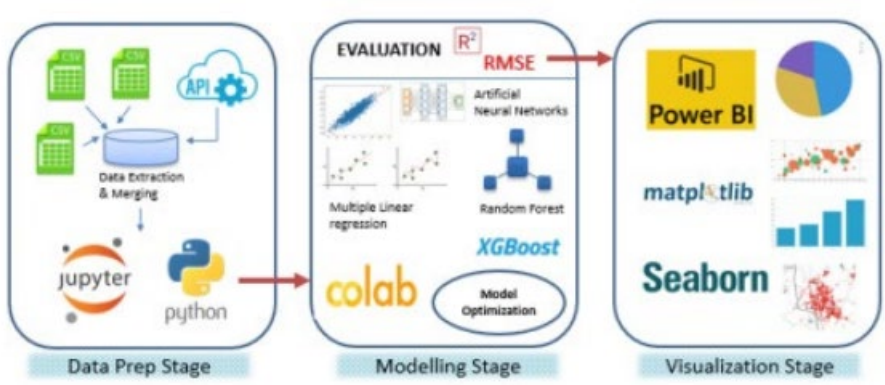
```
cosmetic['price'] = cosmetic["price"].apply(lambda x: 1 if x > 18.90 else 0)

cosmetic['price'].value_counts()
```
[47] ✓ 0.8s Python

```
0    568
1    567
Name: price, dtype: int64
```

+ Code + Markdown

Data Mining

Three phases of the methodology's design flow have been used for this stage, including stages for data, modelling, and visualisation.



The data planning step includes all phases of data collection, combining, and exploratory data analysis, as well as the planning and highlighted selection processes. Depending on the data source, each record was retrieved as a CSV file or extracted via API connections or web scrapping methods using Python programming on Jupyter Notebooks.

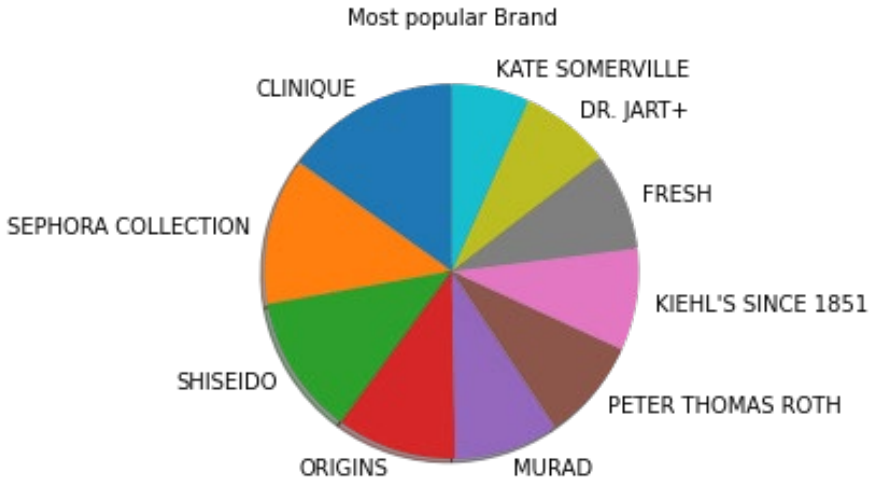
Several machine learning techniques and data mining methodologies, including logistic regression, Decision Tree, Random Forest, K-nearest Neighbour, and clustering, were employed during the modelling stage. For modelling and optimization, Jupyter and Tableau were utilised.

Finally, the findings were displayed in the form of charts, figures, tables, plots and graphs as needed for presentation.

Interpretation/ Evaluation

Visualisations were the primary tool for getting outcomes from data. Matplotlib and Seaborn were all used to build early sample visualisations from the data, with Tableu being utilised afterwards to create more polished and appealing data visualizations for analysis of results and presentation.

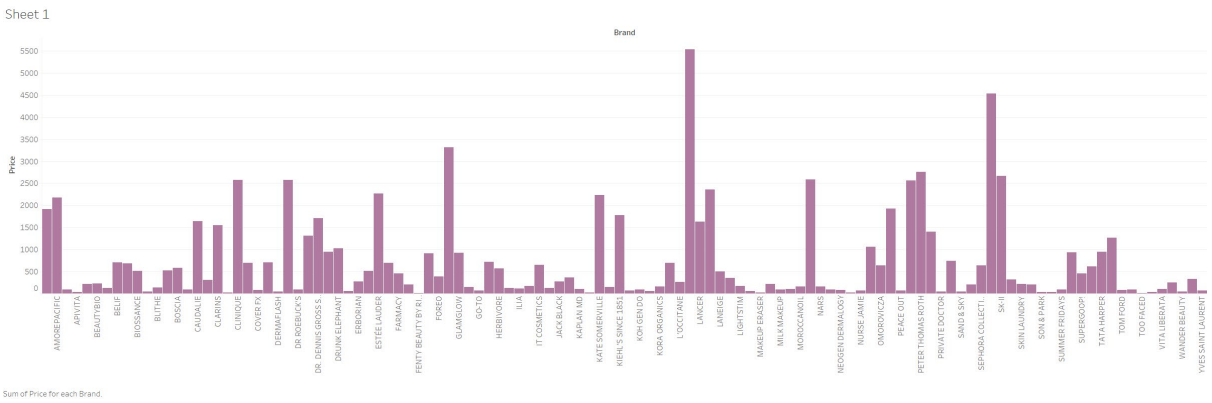
Below are some sample visualization from the datasets provided.




```
skin['Brand'].value_counts().head(10).to_frame()
```

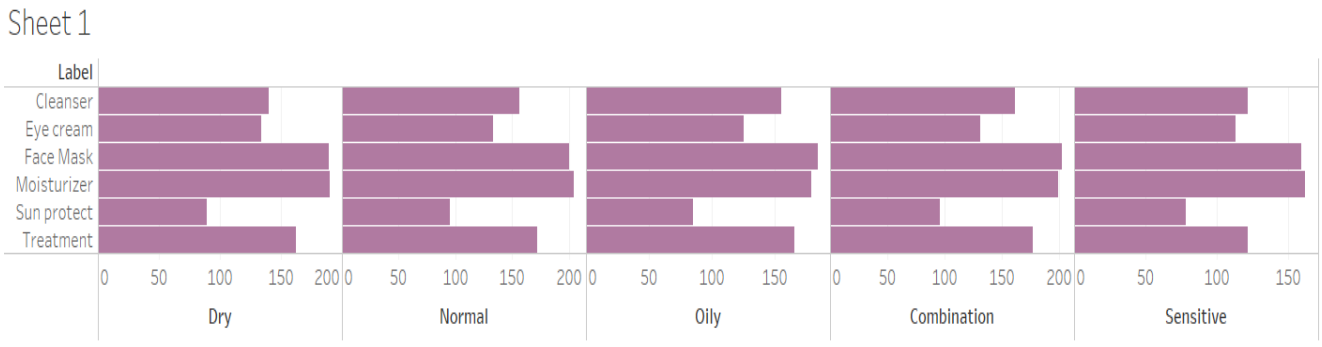
[160] ✓ 0.2s

	Brand
CLINIQUE	79
SEPHORA COLLECTION	66
SHISEIDO	63
ORIGINS	54
MURAD	47
PETER THOMAS ROTH	46
KIEHL'S SINCE 1851	46
FRESH	44
DR. JART+	41
KATE SOMERVILLE	35


Sheet 1
Brand
Price
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
AMOREPACIFIC
APIVITA
BEAUTYBIO
BELIF
BIOSSANCE
BLITHE
BOSCIA
CAUDALIE
CLARINS
CLINIQUE
COVER FX
DERMAFLASH
DR ROEBUCK'S
DR. DENNIS GROSS S.
DRUNK ELEPHANT
ERBORIAN
ESTÉE LAUDER
FARMACY
FENTY BEAUTY BY RI
FOREO
GLAMGLOW
GO-TO
HERBIVORE
ILIA
IT COSMETICS
JACK BLACK
KAPLAN MD
KATE SOMERVILLE
KIEHL'S SINCE 1851
KOH GEN DO
KORA ORGANICS
L'OCCITANE
LANCER
LANEIGE
LIGHTSTIM
MAKEUP ERASER
MILK MAKEUP
MOROCCANOIL
NARS
NEOGEN DERMALOGY
NURSE JAMIE
OMOROVICZA
PEACE OUT
PETER THOMAS ROTH
PRIVATE DOCTOR
SAND & SKY
SEPHORA COLLECTI.
SK-II
SKIN LAUNDRY
SON & PARK
SUMMER FRIDAYS
SUPERGOOP!
TATA HARPER
TOM FORD
TOO FACED
VITA LIBERATA
WANDER BEAUTY
YVES SAINT LAURENT
Sum of Price for each Brand.


Most popular Product
Sun protect
Moisturizer
Eye cream
Cleanser
Treatment
Face Mask

-


Sheet 1
Label
Cleanser
Eye cream
Face Mask
Moisturizer
Sun protect
Treatment
0
50
100
150
200
Dry
Normal
Oily
Combination
Sensitive
Sum of Dry,sum of Normal,sum of Oily,sum of Combinationandsum of Sensitive for each Label.

4.0 Analysis

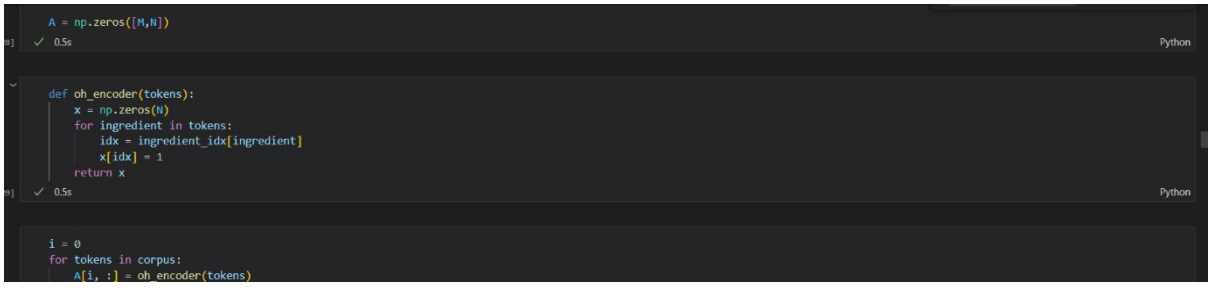
K- Clustering

To achieve our final aim of comparing ingredients for each product, we must first do certain pre-processing activities and maintain track of the exact ingredient in each product's ingredient labels. The first step will be to tokenize the list of ingredients in the ingredient's column. I divided the words into tokens and then created a binary collection of the words. Then, using the tokens and ingredient index, we will form a glossary.

The following step is to create a document-term matrix. Each skincare product will be assigned a document, and each skincare ingredient will be assigned a phrase.

To get started, I created a blank matrix filled with zeros. The size of the matrix is a representation of the total number of skincare products in the data. The width of the matrix represents the total number of components. We will fill up this empty matrix in the phases that follow after initialising it.

Our ultimate aim is to populate the matrix with either 1 or 0: if an ingredient is found in a product, the result is 1. Otherwise, it stays 0. This function's name is oh encoder.

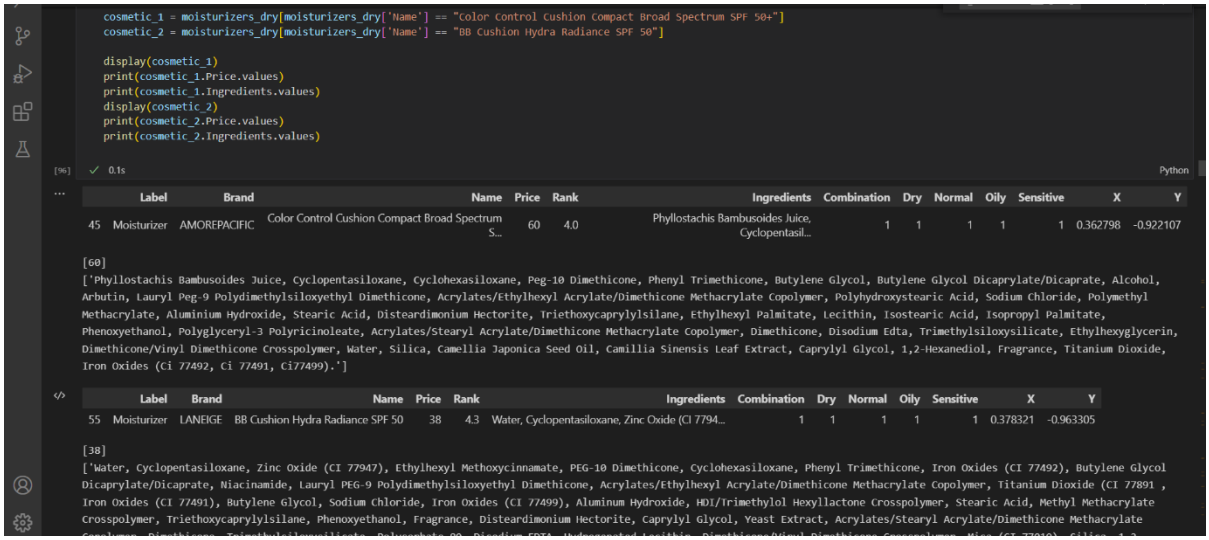


The results will reveal the components of each product after I used the oh encoder () function on the tokens in the corpus and put the numbers at every row of this matrix.

The nonlinear dimensionality reduction method t-distributed Stochastic Neighbour Embedding is useful for integrating high-dimensional data for display in a low-dimensional environment of two or three dimensions. By minimizing the size of data while maintaining the commonalities between the instances, this approach may minimise the dimension of data specifically. As a result, we may vectorize and create a plot graph on the coordinate

plane. The distances between the points will show the commonalities between both the skincare products in our data, which will then be vectorized into two-dimensional values.

Below is an example of two products that have similar ingredients



```
cosmetic_1 = moisturizers_dry[moisturizers_dry['Name'] == "Color Control Cushion Compact Broad Spectrum SPF 50+"]
cosmetic_2 = moisturizers_dry[moisturizers_dry['Name'] == "BB Cushion Hydra Radiance SPF 50"]

display(cosmetic_1)
print(cosmetic_1.Price.values)
print(cosmetic_1.Ingredients.values)
display(cosmetic_2)
print(cosmetic_2.Price.values)
print(cosmetic_2.Ingredients.values)
```

[96] ✓ 0.1s Python

	Label	Brand	Name	Price	Rank	Ingredients	Combination	Dry	Normal	Oily	Sensitive	X	Y
45	Moisturizer	AMOREPACIFIC	Color Control Cushion Compact Broad Spectrum S...	60	4.0	Phyllostachis Bambusoides Juice, Cyclopentasil...	1	1	1	1	1	0.362798	-0.922107

```
[60]
['Phyllostachis Bambusoides Juice, Cyclopentasiloxane, Cyclohexasiloxane, Peg-10 Dimethicone, Phenyl Trimethicone, Butylene Glycol, Butylene Glycol Dicaprylate/Dicaprate, Alcohol,
Arbutin, Lauryl Peg-9 Polydimethylsiloxyethyl Dimethicone, Acrylates/Ethylhexyl Acrylate/Dimethicone Methacrylate Copolymer, Polyhydroxystearic Acid, Sodium Chloride, Polymethyl
Methacrylate, Aluminium Hydroxide, Stearic Acid, Disteardimonium Hectorite, Triethoxycaprylylsilane, Ethylhexyl Palmitate, Lecithin, Isostearic Acid, Isopropyl Palmitate,
Phenoxyethanol, Polyglyceryl-3 Polyricinoleate, Acrylates/Stearyl Acrylate/Dimethicone Methacrylate Copolymer, Dimethicone, Disodium Edta, Trimethylsiloxysilicate, Ethylhexylglycerin,
Dimethicone/Vinyl Dimethicone Crosspolymer, Water, Silica, Camellia Japonica Seed Oil, Camillia Sinensis Leaf Extract, Caprylyl Glycol, 1,2-Hexanediol, Fragrance, Titanium Dioxide,
Iron Oxides (Ci 77492, Ci 77491, Ci77499).']
```

	Label	Brand	Name	Price	Rank	Ingredients	Combination	Dry	Normal	Oily	Sensitive	X	Y
55	Moisturizer	LANEIGE	BB Cushion Hydra Radiance SPF 50	38	4.3	Water, Cyclopentasiloxane, Zinc Oxide (CI 7794...	1	1	1	1	1	0.378321	-0.963305

```
[38]
['Water, Cyclopentasiloxane, Zinc Oxide (CI 77947), Ethylhexyl Methoxycinnamate, PEG-10 Dimethicone, Cyclohexasiloxane, Phenyl Trimethicone, Iron Oxides (CI 77492), Butylene Glycol
Dicaprylate/Dicaprate, Niacinamide, Lauryl PEG-9 Polydimethylsiloxyethyl Dimethicone, Acrylates/Ethylhexyl Acrylate/Dimethicone Methacrylate Copolymer, Titanium Dioxide (CI 77891 ,
Iron Oxides (CI 77491), Butylene Glycol, Sodium Chloride, Iron Oxides (CI 77499), Aluminum Hydroxide, HDI/Trimethylol Hexyllactone Crosspolymer, Stearic Acid, Methyl Methacrylate
Crosspolymer, Triethoxycaprylylsilane, Phenoxyethanol, Fragrance, Distearidimonium Hectorite, Caprylyl Glycol, Yeast Extract, Acrylates/Stearyl Acrylate/Dimethicone Methacrylate
```

Logistic Regression, Random Forest, Decision Tree, SVC, KNN

I evaluated the precision of classifiers for binary classification using logistic regression, decision trees, K-nearest neighbours (KNN), and support vector machines (SVM). In order to forecast a product's price category, I also integrated an ensemble learning model, the random forests classifier.

Using prior observations from a data set, a statistical analysis technique called logistic regression predicts a binary outcome, such as yes or no. A logistic regression model predicts a dependent data variable by looking at the association between one or more already existing independent variables.

The training stage of the random forest ensemble learning technique, which is used for classification, regression, and other tasks, involves the construction of a substantial number of decision trees.

The structure of a decision tree is similar to a flowchart, with each internal node representing a test on a feature, each leaf node representing a class label, a conclusion made after calculating all features, and branches representing connectives of characteristics that result in those different classifiers. Classification rules are represented by the routes from root to leaf.

The goal of a Support Vector Classifier, sometimes referred to as a Linear SVC, is to fit to the data you supply and produce a "best suited" hyperplane that partitions or categorises your data. You may then input some characteristics to your algorithm to get the "predicted" category after acquiring the hyperplane.

One of the simplest machine learning algorithms, KNN is mostly employed for categorization. The data point is categorised based on how its neighbour is categorised. Based on the similarity score of the previously stored data points, KNN categorises the new data points. In essence, it employs proximity to classify or anticipate how a single data point will be grouped.

The datasets are divided into two groups for machine learning. The first subset, referred to as the training data, is a section of our actual dataset that is used to train a machine learning model. It trains our model in this way. The testing data refers to the other subgroup. T raining data tends to be bigger than testing data. This is due to wanting to provide the model with as much information as we can in order for it to identify and learn useful patterns. When our dataset's data are supplied to a machine learning algorithm, the programme recognises patterns in the data and draws conclusions.

I used every predictor in my analysis up until the brand name in the first phase. So I excluded the brand and utilised all the remaining predictors including the contents, type of product and the ingredients list. The brand, contents, product type, and ingredients were all factors in the second portion of my analysis. In the third round of my experiment, I omitted all data on the product's ingredients and quantity. I just made use of the brand and product type columns instead. I have a suspicion that the content volume doesn't offer helpful data on pricing. For instance, whereas body wash may be purchased in 1-liter bottles for just a fraction of the amount of the latter, the average serum usually provided in 30ml to 50ml bottles might be priced at over 30 euros.

Without Content and Ingredients

Without Brand

With Brand

```
X = cosmetic[['brand', 'contents', 'product_type', 'ingredients']]
```

[17]

	brand	contents	product_type	ingredients

On creating a custom tokenizer to help remove unnecessary space and other analysis-interrupting elements. Then, using make column transformer, I created a number of transformers. This make column transformer method was used to create these transformers (). Each machine learning algorithm was integrated with the column transformer in a pipeline.

```
def tokenizer(x) -> list:
    x = x.replace('(', '')
    x = x.replace(')', '')
    x = x.replace("\xa0", " ")
    x = x.replace(".", ",")
    x = x.replace(" & ", ", ")
    x = re.split(', ', x)

    return x
```

[52] ✓ 0.6s Python

```
ohe = OneHotEncoder()
vect = CountVectorizer(tokenizer=lambda x: tokenizer(x))
scaler = StandardScaler()
```

[54] ✓ 0.5s Python

I then looked at the settings for optimization and obtained x test predictions.

Randomized grid search was used to hyperparameter-tune the classifiers for logistic regression, decision trees, KNN, SVM, and random forests.

I then retrieved an accuracy score, obtained the most accurate estimations, and reported the cross-validation accuracy score.

5.0 Results

The purpose of this study, as described in the preceding sections, was to analyse, examine, research, and report on the skincare market and its influencing variables. All data were acquired using Kaggle or scraped from websites like "Look fantastic" and "Cult beauty." The dataset must be cleaned and standardised to match the analysis in order to be analysed and reported. Four comparable datasets are required for this study. Despite the fact that all four datasets were available, they varied in some manner from one another. While some datasets lacked some necessary information for the study, others had all the necessary data. In these situations, the data were examined independently yet complemented one another.

In order to determine the effect, one has on the other, several dataset analysis approaches were applied. Python was used to do several basic regression analysis and correction analysis. I describe the Python-based analysis in this section.

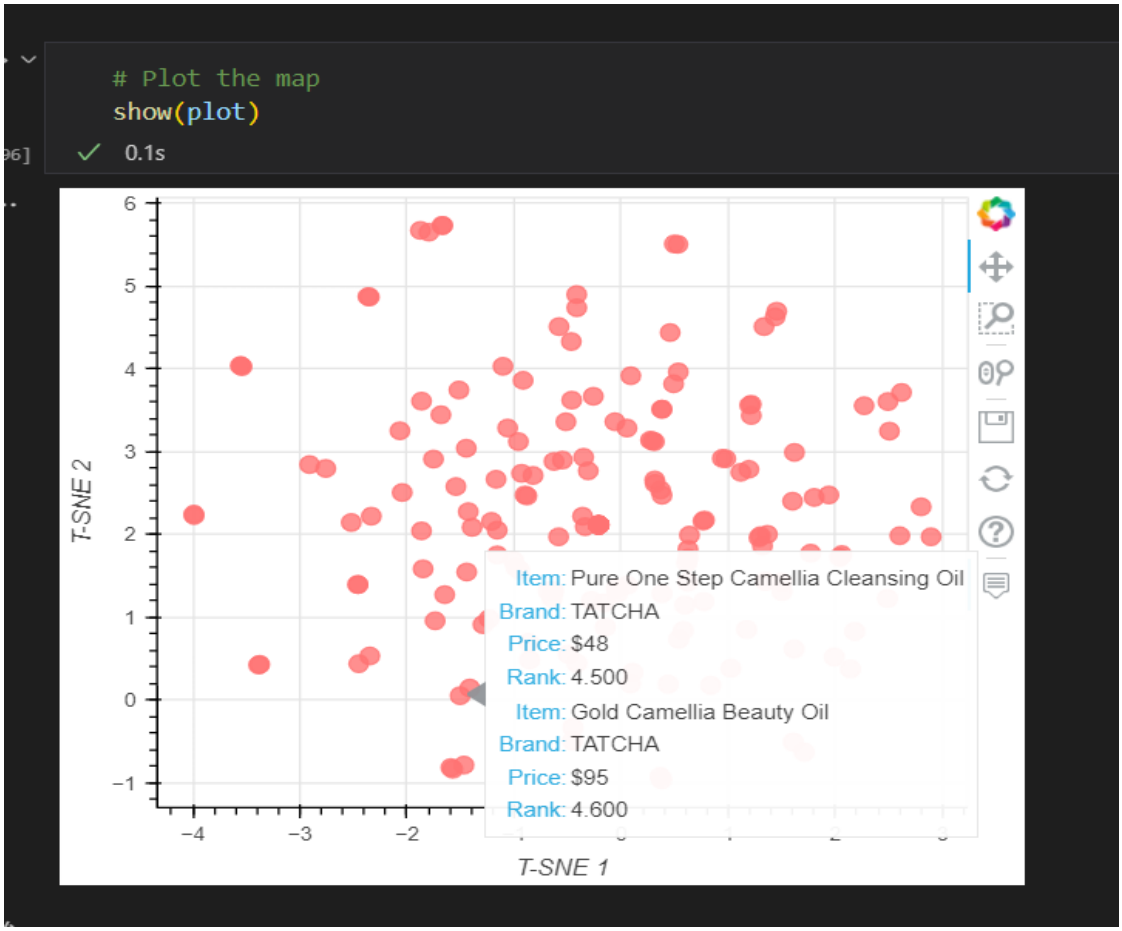
Python has been used to manipulate the datasets. To produce the necessary fraction of the dataset for the study, datasets were divided. To aid in the study, new data columns were developed. For instance, the values of the content measurement mentioned in the product title were used to determine the content from the title.

A variety of analyses were performed on each dataset separately, and the results revealed the following statistics.

K- Clustering

Plotting all of our objects on the coordinate plane will also display their names, brands, prices, and rankings. Using Bokeh to create a scatter plot and including a hover tool to display that data. When a product is immediately under the cursor, we may verify its information by adding a hover tool. The relevant information of each product will be included to tooltips. The many story points are represented by the various skincare products.

The composition of the two objects is increasingly similar the smaller the space between them is. As a result, this allows us to contrast the objects despite having any prior knowledge of ingredient list.



Logistic Regression, Random Forest, Decision Tree, SVM, K- Nearest Neighbour

Because they were predictive classification models, all the various model results were kept in confusion matrices. A confusion matrix is a table structure used in the area of machine learning to display how well an algorithm, often a supervised learning algorithm, performs when given the goal of statistical classification.

With brand and product categories as predictors, the classifier was the most accurate on average, with only the classifier's average prediction being around 80.3%. The random forest classifier works best when the component column is included, whether or not the brand column is also present. I think the presence of the ingredient column 0. improved the performance of the random forest classifier because vectorization of the component list provided infinitely long feature vectors. Only d objects, with the default value of d set to square root can be sampled at each tree node in random forests without replacement. When dealing with extremely long feature vectors, this can help minimize the number of dimensions and overfitting.

The poor performance of KNN using the existing column 'Ingredients' can also be explained by overly large feature vectors. It is possible that the feature space is too sparse and

adjacent data points are too far apart to make useful predictions meaning the dataset lacking from the problem of dimensionality.

The Random Forest Classifier was the best overall model with an average accuracy of 77.9%. This is because a majority vote was used to determine the predictions of the Random Forest classifier trained on the best-performing non-ensemble machine learning algorithms. Nevertheless, when using only the brand column and product type the logistic regression showed the highest accuracy of 83.3 percent.

Results

	Without Brand	*With Brand*	*Only Brand and Product Type*	*Average Accuracy*
Logistic Regression	0.72971014 +/- 0.05775174	0.77082268 +/- 0.04156579	0.83254476 +/- 0.03750473	0.777693
Decision Tree	0.65922847 +/- 0.05882009	0.71359761 +/- 0.05653384	0.75034101 +/- 0.04844292	0.725845
K-Nearest Neighbour	0.61666667 +/- 0.06046723	0.63864024 +/- 0.06699692	0.78554987 +/- 0.05758467	0.712596
Support Vector Classifier	0.73706309 +/- 0.05433613	0.76496164 +/- 0.05684514	0.81489770 +/- 0.05037799	0.772307
Random Forest	0.77231458 +/- 0.05088843	0.78554987 +/- 0.05758467	0.77966752 +/- 0.04238528	0.779177

Blue Highlight= Best Performance

Purple Highlight = Worst Performance

6.0 Conclusions

The purpose of this study was to determine whether there were any elements other than a skincare product's composition and quality that affected its perceived worth. Several machine learning and data mining techniques, including logistic regression, SVC, K-nearest neighbour, random forest, decision tree, and K Clustering, were used to conduct this investigation. The outcomes of the binary classifier studies demonstrate that brands are, in fact, slightly price predictive. Regardless of whether brand was taken into account for predicting prices, using a product's ingredient list as a guide produced less accurate results. Binary classifiers that solely utilised the characteristics of the type of product and the branding had the highest accuracy rates. The product type is demonstrated to have the biggest effect on brand pricing in the correlation, while the brand-free column shows the least influence.

It is obvious that more data is required to strengthen these models because they were unable to complete a more challenging assignment, such as multiclass classification. Given that we are at least aware that its not all high-end product lines can have any beneficial effects, we should not automatically assume that a huge price tag indicates higher-quality skincare ingredients.

I ran into a lot of issues when working on this project. Initially, the goal of the study was to examine product contents and determine which skin type each ingredient is most suited for based on skin type recommendations. I was having trouble finding information or data on that, and given the time constraints, I was unable to conduct that kind of in-depth investigation. I thus focused more on determining each product's pricing and the factors that affect it.

The encoding of substances and associated names using Count Vectorizer represents one of the most significant issues. The names of the identical substances varied from one item to another. For instance, an ingredient was referred to as "water" in one item and "aqua" in another. These similar components will be divided into two independent feature sections when Count Vectorizer is used, increasing the sparsity of the matrix. I don't know enough about ingredient names to see the possible dozens of redundant substances that may be listed in different columns due to differing naming styles or even spellings.

The various amounts of funds that these companies invest in advertising in total are not disclosed in this dataset. The price of the product reflects the amount of funds spent on promotion of the product, which means that advertising budget affects cost structure. As a result, the brand is not a reliable indicator of marketing expenditure or an additive of price category. More data gathered per brand and across several product types could help solve this issue, however due to time constraints, I was unable to do so.

The fact that the arrangement of the items in the ingredient list is important had been a third issue with this research. The product's lowest concentration of a component is found towards the end of the list of ingredients. Thus, any potential understanding of how ingredients affect price is hidden.

With potentially more data collected, we can achieve better outcomes and make it more beneficial for consumers to understand the products they are purchasing and whether it is truly beneficial to them.

7.0 Further Development or Research

This report's research is only a minor step toward the creation of comprehensive skincare ingredients, as well as a better understanding of product quality, price, and branding. There are more elements I would include if I had more time and resources to strengthen and further the analysis. The capacity of the models used for analysis was likely restricted by the selection of only one skin type and a small number of brands and adding additional Selection of Data for other skin kinds and conditions, as well as smaller brands and firms, might result in improved findings and information. Other characteristics, such as environmental circumstances, age, gender, and race, might provide more insight into the additional advantages or issues induced by specific ingredients and product makeup. Due to time and resource limits, and because they are outside the scope and scale of this project, I welcome any interested person that wishes to further the research to continue working on it.

8.0 References

Mayo Clinic. (2019). *Dry skin - Symptoms and causes*. [online] Available at: https://www.mayoclinic.org/diseases-conditions/dry-skin/symptoms-causes/syc-20353885.

Secara, O. M. & Sasu, D. V., 2013. THE IMPACT OF GLOBALIZATION IN THE INDUSTRY OF COSMETICS. *Annals of Faculty of Economics,* , 1(2), pp. 681-691.

Jan, M., Haque, A., Abdullah, K., Anis, Z. and Faisal-E-Alam, F. (2019). Elements of advertisement and their impact on buying behaviour: A study of skincare products in Malaysia. *Management Science Letters*, [online] 9(10), pp.1519–1528. Available at: http://m.growingscience.com/beta/msl/3234-elements-of-advertisement-and-their-impact-on-buying-behaviour-a-study-of-skincare-products-in-malaysia.html.

Liu, Z. and Ling, J. (2019). Research on Influencing Factors of Purchase Promotion Intention Based on Skincare Industry. *Modern Economy*, 10(03), pp.1033–1047. doi:10.4236/me.2019.103069.

Cosmetics & Toiletries. (2013). *Consumer Perspective—Skin Types and Sensory Experience*. [online] Available at: https://www.cosmeticsandtoiletries.com/research/literature-data/blog/21837473/consumer-perspectiveskin-types-and-sensory-experience.

Research, I. (2022). Global Cosmetic Skin Care Market Size-Share (2022-2027) | Growing At CAGR of 7.1% | Supermarkets and Grocery Retailers Demands, Regional Overview, Sales Revenue, Business Prospect, Growth Opportunity, Challenges, and Potential Benefits. [online] GlobeNewswire News Room. Available at: https://www.globenewswire.com/news-release/2022/04/07/2418713/0/en/Global-Cosmetic-Skin-Care-Market-Size-Share-2022-2027-Growing-At-CAGR-of-7-1-Supermarkets-and-Grocery-Retailers-Demands-Regional-Overview-Sales-Revenue-Business-Prospect-Growth-Opp.html.

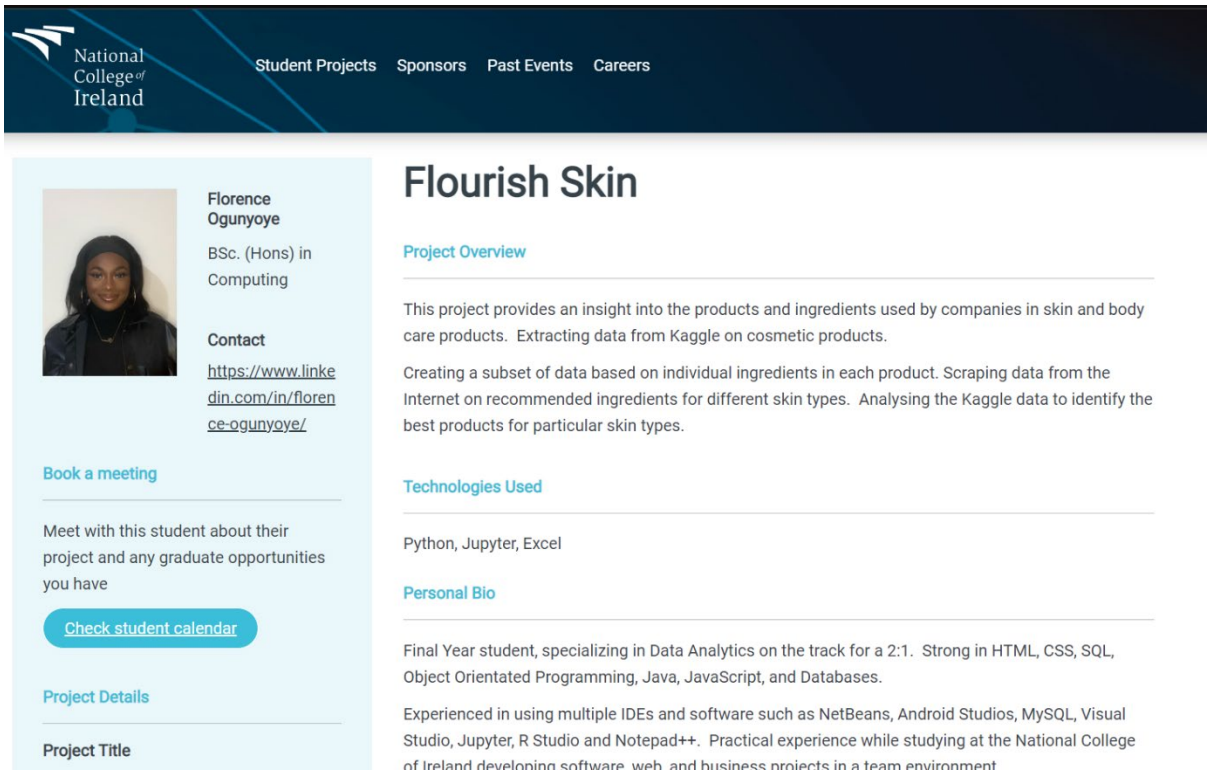
Faisal Khan, A. (2013). A STUDY OF INFLUENCE OF PACKAGING ON WOMEN SKINCARE CONSUMERS IN INDORE CITY. International Journal of Advance Research, [online] 1(10). Available at: http://www.ijoar.org/journals/IJOARBMA/papers/A-STUDY-OF-INFLUENCE-OF-PACKAGING-ON-WOMEN-SKINCARE-BUYERS-IN-STATE-OF-MADHYA-PRADESH.pdf.

upGrad blog. (2020). KDD Process in Data Mining: What You Need To Know? [online] Available at: https://www.upgrad.com/blog/kdd-process-data-mining/.

www.fortunebusinessinsights.com. (n.d.). Skincare Market Size, Analysis | Skin Care Industry Trends, 2026. [online] Available at: https://www.fortunebusinessinsights.com/skin-care-market-102544.

9.0 Appendix

Showcase Profile



Showcase Poster

Florence Ogunyoye

BHSC4 Data Analytics
Flourish Skin

ABOUT

Flourish Skin compares prices between products based on different infiuencial variables /it allows people to know what parts of a product plays in the value of the product

LA MER.

TECHNOLOGY

- Python
- Visual Studio IDE
- Excel
- Jupyter Notebook
- Tableu

PAULA'S CHOICE SKINCARE

The Ordinary.

10.0 Project Proposal

Objectives

It might be difficult to achieve healthy-looking skin. Anyone with a problem complexion is familiar with exploring the limitless variety of skin care products filling up the shelf in search of potential answers at their local drugstore. If they had discovered an effective treatment compatible with their skin and its growing demands, they could have spent the immeasurable amount of time and money wasted on this trial-and-error technique on more fun activities.

The Objective of this project to be able to provide people with insight on the products and ingredients used by companies in the products they are used for skincare and body care Also providing details about suitable products and remedies that can be used to boost their skin routine. This will also be allowing people build a solid skincare and body care routine based on the information received from them, such as skin type, age, environment, skin conditions, allergies etc. These factors all play a role in the current state of a person's skin. Skin care encompasses a variety of procedures that help to maintain the integrity of the skin, improve its appearance, and alleviate skin problems. Nutrition, sun protection, and the use of emollients are just a few examples. This project will be aimed towards people that are interested in taking care of their skin or boosting their skincare and body care routine. Allowing them to learn about what could be beneficial for their skin type and what could possibly cause damage or further damage to their skin.

This project will hopefully allow people to feel more confident and comfortable in their own skin.

Background

The reason I chose to undertake this project is because I am no stranger to the aggravation that comes with failed attempts to fix her bad skin and the resulting loss of self-esteem. I have struggled with a variety of skin conditions in the past and I also know people who have also struggled with different skin problems and were eager to find a solution to find a remedy to cure because these problems may deeply affect our confidence and self-esteem.

While most other skin care products on the market claim to deliver "miracles" on all skin types, I believe that these claims appeared improbable, especially considering a lot of complaints of low satisfaction. I had invested a lot of money into all these one-size-fits-all things and figured that there must be as better way to tackle these problems. I had to do some research of my own and find out what products and ingredients work best for my skin and what I should be looking for in products. And after taking the time to do this research I had figured out what worked best for my skin

If I can provide even the smallest information that help someone feel good about themselves in the slightest way. I feel undertaking this project will help give people insight into what could possibly help them.

State of the Art

Ming Zhao and Amy Yuan, founders of Proven Skincare, set out to establish the world's most comprehensive skin care database to take individualised skin care treatments to the next level. The duo – who later recruited dermatology adviser Doctor Tyler Hollmig and formulations advisor Doctor Nick Conley to form their team – dug through scientific literature, consumer reviews, and social media comments that fed their algorithm to uncover the skin care agents that work, using AI, personalization, and big data.

Their study grew into a Herculean project known as The Skin Genome Project, which the business partners describe as the world's most comprehensive skin care database. The research resulted in the development of tailored, cruelty-free solutions for everyone's skin care needs, based on 25 million user testimonies, 100,000 products, 20,000 components, and 4,000 peer-reviewed scientific articles. The project was awarded the 2018 Artificial Intelligence Award by the Massachusetts Institute of Technology (MIT).

How my work will differ from the work done by Ming Zhao and Amy Yuan is though theirs contain a huge number of details relating to ingredient lists and the contribution of these ingredients to different skin types, their end goal is to create products personalised for each person. My project will give people an in depth break down of the different products on the market right now and give them a personalised affordable skincare routine that will be tailored to their specific needs.

Data

Data that is required for this project will gathered from Erin Ward's "Skincare Products and their Ingredients." And "Skincare Products Clean Dataset." Datasets.

This dataset contains 1138 skincare goods from LookFantastic.com, including their names, URLs, product categories, ingredients, and pricing. How I will access the necessary data is by categorizing the data based on the component list, product quality and product category. Next, I will categorize the data based on the product's brand, Ingredient list, amount of product there is and the type of product it is

Following that the data categorization will be based on product type and brand (without any information about the product amount or ingredients). And after that Brand, component list, product volume, and product kind are used to create a multiclass categorization.

And lastly, a multiclass categorization based on product type and brand (without any information about the product amount or ingredients).

Methodology & Analysis

In this project, I will be trying to implement some machine learning techniques to learn the ingredients in skincare and body care products on the market, to determine which products is beneficial for certain skin types and which can cause damage or further damage to certain

skin types. While also comparing ingredients list against price of products. So, people are given an option and a cheaper alternative. For example, on the La Mer official website, a 16-ounce container of La Mer cream costs $2,475 USD, but the considerably cheaper Nivea Creme moisturiser has a virtually identical ingredient list.

I hypothesize that the ingredient list and the placement of ingredients in an ingredient list play a role in the quality of a product and different products and ingredients work better on certain skin types. Also, inserting a product's ingredient list in the training data would not improve the products performance, implying that a product's contents had no influence on its pricing and that there the quality of a product is not determined by the price.

Using the CRISP-DM Method I will analyse the data from the data set, then creating subsets from the dataset. Cleaning the targeted data and pre-processing to get a data pattern. I will then be transforming data using dimensionality reduction. Followed by a search of pattern that are of interest.

Technical Details

Project Plan

I intend to start out by firstly collecting data off the dataset using the KDD Method I'll analyse the data in the dataset before breaking it down into subgroups. Cleaning and pre-processing the relevant data to obtain a data pattern. After that, I'll use dimensionality reduction to alter the data. Then there's a search for patterns that are of interest.

Afterwards I will be creating a list of ingredients that corelate to skin types and skin concerns.

Subsequently A quiz will be taken to gather some information about the person. It will be a few minutes of an online questionnaire. It will offer precise questions to target one's primary skin care issues, such as acne, rosacea, rough skin, dark circles, and wrinkles. The questionnaire includes asks about skin sensitivity, allergies, ethnicity, and lifestyle, as well as screen use, nutrition, and stress levels. It also requests the country you are living to calculate your UV index, pollution levels, and other environmental parameters if possible. As users complete the questionnaire, active compounds ranging from vitamin E ceramides, Vitamin C to glycolic acid, glycerine, and salicylic acid are suggested, allowing the database to extract some data to be able to recommend suitable products and the prices that can match the unique needs of their skin.

They will be provided with a step-by-step skincare routine with some optional suggestion of products.

10.1. Reflective Journals

Project Plan

Project Task	Sub Task	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Research	Initial Research								
	Data Research								
	Midpoint Presentation								
Data Gathering	Data Gathering								
Data Cleaning	Cleaning and Pre-processing								
Analysis	K- cluster								
	Logistic Regression								
	Random Forest								
	Decision Tree								
	K- Nearest Neighbour								
	SVC								
Report	Report								
	Poster & Video								

Actual Plan

Project Task	Sub Task	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June	July
Research	Initial Research										
	Data Research										
	Midpoint Presentation										
Data Gathering	Data Gathering										
Data Cleaning	Cleaning and Pre-processing										
Analysis	K- cluster										
	Logistic Regression										
	Random Forest										
	Decision Tree										
	K- Nearest Neighbour										
	SVC										
Report	Report										
	Poster & Video										

October

Considering that I had little understanding what a data analytics project truly required, I was having some trouble determining what to submit for my project proposal. In the end, I chose a subject that interested me and about which I had done some prior study since I felt that it would be more advantageous for me. I ultimately chose my plan to analyse the skincare market. I chose this because skincare is something I'm very interested in, and I thought it would be fascinating to discover which products are best for certain skin types based on the ingredient list.

November

I have dedicated time to conducting further research for various data analysis and skincare-related initiatives. In order to gain a sense of the level, structure, and timeframe that were necessary, I found the previous projects on this subject to be quite useful.

I used this time to choose the programming language I would use for my project. I had trouble comprehending R, therefore I'm currently researching Python as well.

December

This month, I mostly concentrated on my midpoint presentation. I covered all of the work and research completed to date as well as my strategy for staying on track and completing everything by the deadline. It was quite challenging to truly devote some time to this project in particular because of the number of tasks that needed to be done for the other modules.

I've chosen to use Python for my project since I find it to be more convenient.

January

My project is now in the data gathering phase. I should be able to finish this level in the most of February and March. I'm researching web scraping as a way to collect data. since there aren't many datasets that include the data, I'm looking for.

I want to scrape information from many beauty websites, including Feel Unique, Look Fantastic, and Cult Beauty.

February

Finding the necessary data is proving to be very challenging for me. I have resumed my study on the various skin types and the components that are most suited for each one. Some substances are not advised for certain skin types, while others are acceptable for all skin types. Consequently, I am having trouble locating information on what skin type a certain chemical might fit.

I've made the decision to seek out a different result. rather than trying to find a product that would work for a certain skin type. I've made the decision to focus my investigation into skincare marketing and identify the variables that affect a product's pricing point.

since the result calls for the same data that I already have. It hasn't significantly hindered the project.

In order to incorporate as many machine learning algorithms as possible before the final submission, I have also chosen to take on a new one each week.

March

Every week, I'll take on a different machine learning algorithm, starting with logistic regression, and moving on to random forest. Decision Tree, K-Clustering, SVC, and K-nearest Neighbour are also featured.

I expect to be able to determine which elements affect a product's pricing using these various algorithms.

April

N/A

May

N/A

June

N/A

July

This month, the final month of my project, I spent the remaining of my time working on my report and cleaning up any files that needed to be sorted before final submission. I also took this time to do my video submission as well. I will work on finish my project and proofreading and cleaning up anything that needs fixing

10.2. Other materials used

Any other reference material used in the project for example evaluation surveys etc.