Scientific Data Mining
CDS303-001
Summer 2022

The Relationship between Skin Care Products and Price

July 31st, 2022

Team Members:

Aqsa Majeed
Sharon Corrales
Avantika Prativadhi

Contents

Introduction

Business problem and objective

The business problem and objective is the correlation between high rated products and price. There is this perception put into consumers' minds that expensive skin care products is better compared to cheap skin care products. According to Daniela Morosini , what makes products high priced is due to the "marketing fluff and maybe some fancy packaging – not potency." (2021). The model will help understand this relationship between a product's price and the rating it received. By understanding this relationship, we will be able to recommend well rated products in the customers' budget.

Analytic problem

The analytic problem is that there are many young consumers that suffer from acne and or other skin changes which can be stressful. This group of consumers struggle to find affordable skin care products that can be effective on their skin. Additionally, many of them are unaware of places/ stores where they can buy affordable, high rated skincare products.

Goals and success criteria

The goal for our project is to use the data set we found from Kaggle and find a relationship between high rated skin care products and price. With this information, we can find what are the mostly highly rated products with affordable prices and recommend them to consumers. We plan to achieve this goal by creating a K-means clustering method and also conduct a linear regression to get an accurate understanding of how the variables in the dataset work with each other.

Resources available

Throughout our project we will be using resources such as the given readings, PowerPoints, tools, Mason Library, and the professor/TA. These resources will help us further our research in regards to how to approach our question and clarifying any road bumps we may go through in the future or working progress.

Requirements, assumptions, constraints

One of the major requirements will be creating coding to organize and create a visualization of the products with the different ratings. By creating a visualization, we can better convey the information from the model to a customer. An assumption is going to be that the high priced products will be the ones highly rated. Some things that we will be limited during the process of our project is our knowledge of coding. Some of us have experience with R and some other languages but we are not completely sure if we want just to create visualizations since we are pretty limited on our code experience. Additionally, we are limited to the information provided by the dataset. We don't have enough information regarding the ingredients that the product contains or an elaborate explanation on how the ratings were given. We are also unsure if the ratings were rounded up and what currency the price column in the dataset is.

Conceptual model

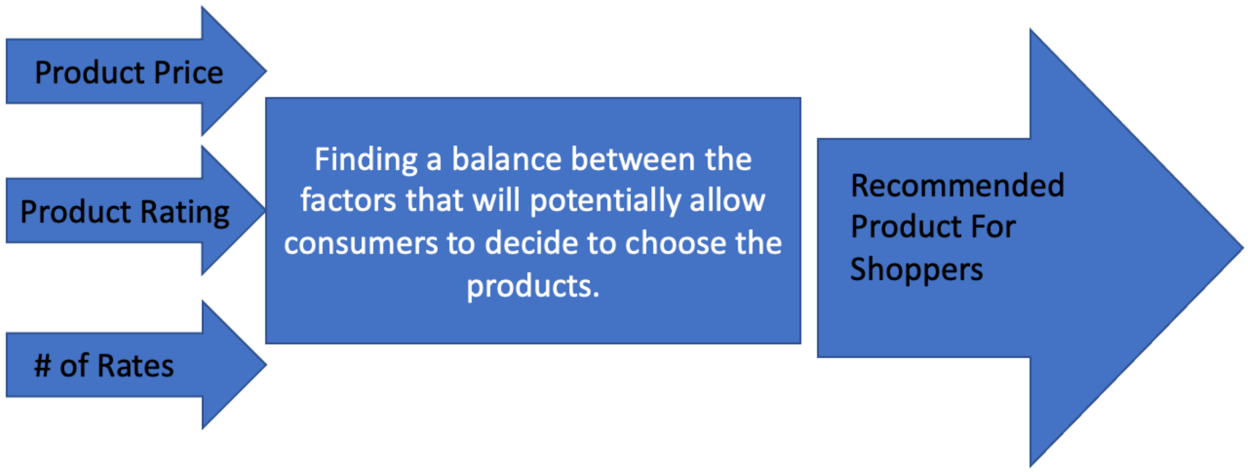

Figure 1: Black Box Diagram, a conceptual model of factors in choosing good skin care products

Data

Data collection and cleaning

This dataset is from Kaggle by Najwa Alsaadi. It was updated a year ago and was collected using Nordstrom webscaping. The dataset was used to create a content-based recommendation derived by various skincare products. Since this dataset is big, we did notice the duplicated values or incorrectly typed values, those have been fixed. The products that have little to no reviews will not be used.

A data.frame: 10 x 5

	X	price	title	stars	vote
	<int>	<int>	<chr>	<dbl>	<int>
1	0	6000	BeautyBio Rose Quartz Roller	4.7	75
2	1	2800	Kopari Starry Eye Balm	4.0	11
3	2	7200	StriVectin SuperC Brighten and Correct Serum	4.6	271
4	3	5900	Skin Gym Rose Quartz Facial Workout Set Nordstrom Exclusive USD 92 Value	5.0	1
5	4	9150	Slip Marble and Charcoal Pillowcase and Sleep Mask Set USD 139 Value	5.0	2
6	5	16900	LightStim for Acne LED Light Therapy Device	4.4	29
7	6	6000	Charlotte Tilbury Magic Eye Rescue Cream	3.9	167
8	7	2500	iluminage TOUCH Precision Adaptor	0.0	0
9	8	62500	MZ Skin LightTherapy Golden Facial Treatment Device	0.0	0
10	9	3500	BeautyBio GloPRO LIP MicroTip Attachment Head	4.8	23

Figure 2: The first 10 rows of the uncleaned dataset

The Kaggle page does not mention which currency the price column reflects. We assumed that the price was in dollars while conducting our analysis. Since this value was written in the thousands on the dataset, we cleaned the data to change the pricing to reflect the realistic product price value in dollars. Although most of the prices could be changed by dividing the number by a 100, there were a few rows that still reflected high prices. We tried to further clean those values by dividing them by 10. By going through this data cleaning process, we were able to conduct our analysis with realistic prices of the products which helps us answer our business problem.

A data.frame: 10 × 5

	X	price	title	stars	vote
	<int>	<dbl>	<chr>	<dbl>	<int>
1	0	60.0	BeautyBio Rose Quartz Roller	4.7	75
3	2	72.0	StriVectin SuperC Brighten and Correct Serum	4.6	271
7	6	60.0	Charlotte Tilbury Magic Eye Rescue Cream	3.9	167
11	10	50.0	Kiehls Since 1851 PowerfulStrength Dark Circle Reducing Vitamin C Eye Serum	4.2	645
20	19	15.0	Dr Dennis Gross Skincare Alpha Beta Peel Extra Strength Formula 60 Applications	4.9	434
21	20	17.0	Mario Badescu Drying Lotion	4.6	474
23	22	19.9	PMD Personal Microderm Pro Device USD 219 Value	4.5	63
29	28	52.0	Lancôme Bienfait MultiVital SPF 30 Day Cream Moisturizer	4.7	505
30	29	39.0	Kylie Skin 4Piece Mini Skincare Set	4.2	428
31	30	21.2	Lancôme Absolue Revitalizing and Brightening Soft Cream	4.5	1554

Figure 3: The first 10 rows of the cleaned dataset

Exploratory data analysis

To begin with the exploratory data analysis, we first used the describe function in R to get an understanding of what the variables are and how they compare to each other. Then, we plotted some basic graphs to look for any trends. The first graph we created was a bar plot to look at the distribution of ratings (Figure 4). This plot shows us that the product ratings as they fall between 4.0 and 5.0; with 4.7 being the most common. The plot also shows the multiple outliers towards the front. Even with the data cleaning; these outliers are very strong and later impact the accuracy of the analysis. Next, we created a scatter plot to understand if there was a relationship between the price of the product and how it was rated. In figure 5, it’s evident that there is no proper relationship between the two variables. All the points are clumped towards the bottom of the graph. To check if there was any linear relationship, we added a regression line (figure 6). Based on the plot, we can conclude that our hypothesis that higher products have higher rating and lower products have lower ratings, is rejected.

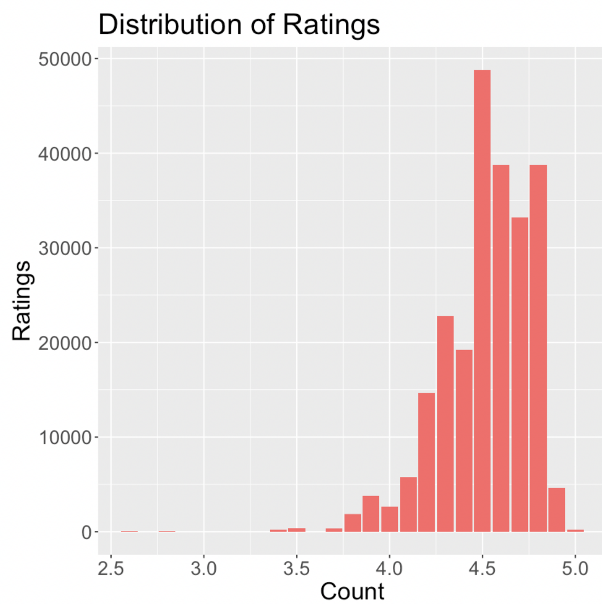

Figure 4: Distribution of the ratings received by skincare products

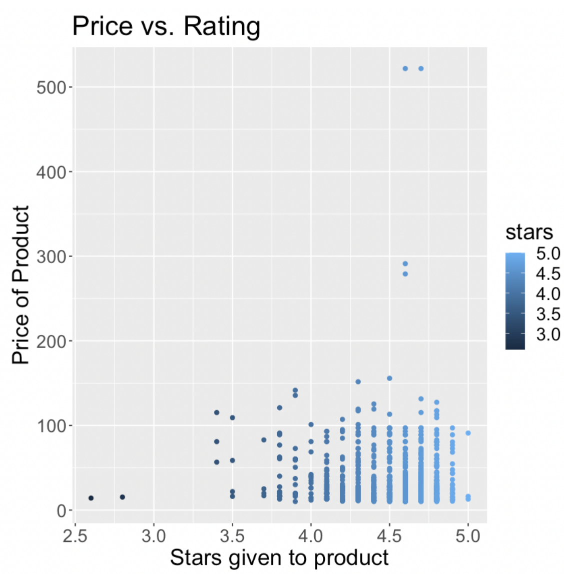

Figure 5: Scatter plot of "Product of Price Vs. Rating Received"

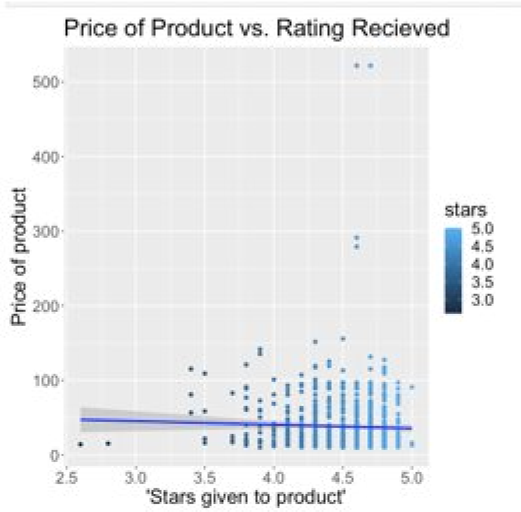

Figure 6: Scatter plot of "Product of Price Vs. Rating Received" with a regression line through the points

Feature selection / engineering

Since our focus is for customers to purchase good quality skincare, we will be using all of the variables provided in the dataset. The variables (stars and votes) will help us determine what products the buyers appeal to. There are some products that have no reviews, and those products will be excluded from the dataset.

Handling and storing data

We will be handling data by keeping track of the product sales. A bar graph would be ideal for this dataset since we have many products, and it will help us monitor the sales of the different products. We will store our data using a csv file on our laptops. The data will be loaded onto Jupyter notebook and R Studio where we will be doing the modeling and the analysis of the data to reach a conclusion for our business problem.

Table 1: This is a table

A data.frame: 10 × 5

	X	price	title	stars	vote
	<int>	<int>	<chr>	<dbl>	<int>
1	0	6000	BeautyBio Rose Quartz Roller	4.7	75
2	1	2800	Kopari Starry Eye Balm	4.0	11
3	2	7200	StriVectin SuperC Brighten and Correct Serum	4.6	271
4	3	5900	Skin Gym Rose Quartz Facial Workout Set Nordstrom Exclusive USD 92 Value	5.0	1
5	4	9150	Slip Marble and Charcoal Pillowcase and Sleep Mask Set USD 139 Value	5.0	2
6	5	16900	LightStim for Acne LED Light Therapy Device	4.4	29
7	6	6000	Charlotte Tilbury Magic Eye Rescue Cream	3.9	167
8	7	2500	iluminage TOUCH Precision Adaptor	0.0	0
9	8	62500	MZ Skin LightTherapy Golden Facial Treatment Device	0.0	0
10	9	3500	BeautyBio GloPRO LIP MicroTip Attachment Head	4.8	23

Fig 7: First 10 rows of the Dataset

Modeling

Machine learning methodology

Our model is unsupervised and the machine learning methodology associated with it is Clustering. This is unsupervised and the goal of this model is to uncover patterns and relationships. We will use the K-means algorithm to group the dataset into clusters and understand the relationship between the clusters. This method assigns the data points into different clusters so that the sum of the squared distances between the data points and the centroid is as small as possible. It computes centroids and repeats the process until the optimal centroid is found (Khan, M).

Tuning and testing plan

There are various parameters included in the code for the K-means clustering. The main ones that we specified are: i) random_state, ii) n_clusters, iii) init, iv) n-start. The first parameter is random_state. Using this, we set a random seed of 88 which allows us to reproduce the exact clusters over and over again. The second parameter is the n_clusters which specifies how many clusters we want. While this number is not always evident or easy to tell, we relied on our explanatory analysis to see where most of our data points fell. Since most of the ratings were between 4.0 and 5.0, we decided that the best number of clusters is 5. The init parameter is the initial cluster centroid. Since we have 5 clusters, we could assume that we would have 5 centroids. However, we used the nstart parameter to allow R to generate multiple initial configurations and report the best place of the centroids. Since the dataset has one variable called rating and one called votes, these variables satisfy the algorithm parameters. This will help us let the customer know how the product they chose is rated and by how many people. Using this, we can make a recommendation on whether the product is worth it or not.

We will conduct unit tests where the program/model is broken down into simpler blocks and each element is tested separately. Similarly, K-fold cross-validation will divide the dataset into k subsets and use them k times. Repeating this process will ensure that all the data points have been tested on which then reduces the bias of the model.

Development environment and language

The data for our model is in tabular format and we will be using the R language to develop our model. The development environment that will be used is RStudio and the R kernel on Jupyter Notebook.

Evaluation

Throughout the process of investigating the data set, the model reflected a constant relationship between the price and rating. The graphs depicted a major population within the average of 3-4 star range, which means there's no relationship between the price and ratings of the products within the dataset. The data points within the graph did slightly show some points that were higher ratings and higher pricing, however, it was not enough to conclude that there was a linear relationship between the two. As we did further trial and testing it was finally

found there was a constant relationship. This means that as the ratings increased, price remained the same.
Referring back to the average ratings of starts consisting in the 3-4 group, as we observed our overall data we found the reason for this was due to the lack of products with different ratings. It can be noted that most customers and consumers rate nordstorm products more 3-4 stars which shows that customers are satisfied with them. Due to the lack of different product ratings, it was difficult to conclude whether consumers should purchase products with higher pricing. Another thing that was noted was that the price of products were hard to read. Towards the beginning of the project, we believed that the first two digits were the prices and the ending zeros were insignificant place holders. Although, as we continued working with the dataset we realize this was not true.

Conclusion

To conclude, during the course of our analysis, we learned that K-means is often very sensitive to outliers. Before choosing the dataset, we did not consider the many outliers it would contain. Even after cleaning our dataset, we found that it remained with many outliers. This made it difficult for us to analyze the graphs accurately and caused the graphs to be skewed. Although our assumptions were not satisfied and not meet our expectations, we were still able to come up with a conclusion to help customers.
We choose the skincare dataset in hopes of helping consumers decide whether purchasing a costly product was more effective. Since we were limited to the information of the data set, we were unable to find the effectiveness of the products on the skin. We assumed that if the skin care product was effective then consumers would give the product a high rating. Therefore, In order to investigate this, we decided to test the relationship between ratings of the products and price. We predicted a direct linear relationship because the “better” or effective the product would be, the higher the rating and the higher the price. As it was discussed in the evaluation section of the output of the graphs and results, we concluded that high priced Nordstrom skincare products did not have any more effectiveness on the skin compared to cheaper skin care products.
With this information and results, consumers are suggested to purchase any Nordstrom skin care product as there is no difference in effectiveness when looking into price. This goes to show that a cheaper skin care product can be just as effective as a very expensive product due to our research and dataset.
Overall, during the end result of our project, we found that the outliers made us understand our dataset better and we now have a better understanding of what type of dataset we should be looking at for our future projects. We also understand what variables are beneficial for our final model as well.

References

Dataset: Najwa Saeed Alsaadi. (2021). <i>skin care </i>[Data set]. Kaggle.
https://doi.org/10.34740/KAGGLE/DSV/1890589

Gavrilova, Y. (2020, November 11). *Testing machine learning models*. Serokell Software Development Company. Retrieved June 30, 2022, from https://serokell.io/blog/machine-learning-testing

Learn by marketing. Learn by Marketing | Data Mining + Marketing in Plain English. (n.d.). Retrieved July 26, 2022, from https://www.learnbymarketing.com/methods/k-means-clustering/

Khan, M. (2017, August 2). *KMEANS clustering for classification*. Medium. Retrieved July 26, 2022, from https://towardsdatascience.com/kmeans-clustering-for-classification-74b992405d0a

Team, T. V. (2021, July 6). *Cluster analysis in R - complete guide on clustering in R*. TechVidvan. Retrieved July 26, 2022, from https://techvidvan.com/tutorials/cluster-analysis-in-r/

Appendix A: Project Plan

Task / subtask	Can be started when?	Must be finished when?	Expected duration (days)	Assigned (name)	Status (not started, in process, complete
Choose Topic	June 22nd	June 22nd	1 day	Everyone	Complete
Write executive summary report section	June 22nd	June 23rd	2 day	Everyone	Complete
Business Problem					
Write business problem report section	June 28th	June 29th	1 day	Sharon	Complete
Create one slide that has both the business and analytic problem	June 28th	June 29th	1 day	Sharon	Complete
Analytic Problem					
Write analytic problem report section	June 27th	June 29th	2 day	Sharon	Complete
Create one slide that has both the business and analytic problem	June 27th	June 29th	2 day	Sharon	Complete
Project plan					
Create project plan	June 22th	June 29th	7 day	Everyone	Complete
Write project plan report section	June 28th	June 29th	1 day	everyone	Complete
Create one slide on project plan	June 28th	June 29th	1 day	Everyone	Complete
Data					
Find data	June 27th	June 29th	1 day	Everyone	complete
Prepare data	June 27th	June 29th	2 days	Avantika	Complete
Put data in a place where model can access it	June 28th	June 29th	1 day	Aqsa	Complete
Perform exploratory data analysis on data	June 28th	June 29th	1 day	Avantika	Complete
Develop visualizations based on EDA	June 28th	June 29th	2 days	Avantika	complete
Write data section of report, include EDA visualizations and discussion	June 28th	June 29th	1 day	Avantika	complete
Create two slides on data	June 28th	June 29th	1 day	Aqsa	complete
Methodology					
Choose methodology	June 29th	June 30th	1 day	Avantika	Completed
Write methodology section of report	June 29th	June 30th	1 day	Avantika	Completed
Create one slide on methodology	June 29th	June 30th	1 day	Avantika	Completed
Modeling (choice, development)					

Determine appropriate model type	June 26th	June 30th	1 day	Avantika	Complete
Develop black box / conceptual model	June 28th	July 7th	9 days	Sharon	Complete
Develop model code based on black box model	June 28th	June 30th	2 day	Sharon	Complete
Write model type and justification report section	June 28th	July 7th	9 day	Sharon	Complete
Create one slide on modeling choice, also showing black box model	June 29th	June 30th	1 day	Sharon	Complete
Create one slide summarizing the modeling process	July 4th	undecided July 7th	3 day	Everyone	Complete
Modeling (hyperparameter tuning)					
Develop approach for tuning hyperparameters	July 7th	July 11th	4 days	Avantika	Complete
Tune hyperparameters of model using developed approach	July 7th	July 11th	4 days	Avantika	Complete
Write hyperparameter tuning report section of report	July 7th	July 11th	4 days	Avantika	Complete
Create one slide describing hyperparameter tuning process	July 7th	July 11th	4 days	Avantika	Complete
Modeling (testing)					
Develop test plan for model	July 7th	July 11th	2 days	Avantika	Complete
Test model	July 9th	July 11th	2 days	Avantika	Complete
Write model testing section of report	July 10th	July 11th	2 days	Avantika	Complete
Create two slides: one on test process and one on results	July 11th	July 11th	1 day	Aqsa	Complete
Modeling (output)					
Develop concept for a way for people to see the output of the model (user interface)	July 11th	July 20th	9 days	Avantika	Complete
Develop front end code for visualization	July 11th	July 20th	9 days	Avantika	Complete
Develop model output visualizations for report	July 11th	July 20th	9 days	Avantika	Complete
Write modeling output report section	July 11th	July 20th	9 days	Sharon	Complete
Create two slides showing model output visualizations	July 11th	july 20th	9 days	Sharon	Complete
Modeling (assumptions and limitations)					
Write report section on assumptions and limitations	July 20th	July 23th	2 days	Aqsa	Complete
Create one slide on modeling assumptions and limitations	July 21th	July 23th	2 days	Aqsa	Complete

Evaluation					
Interpret model results in the context of the business problem.	July 20th	July 26th	6 days	Sharon	Complete
Write report section on evaluation	July 20th	July 26th	6 days	Sharon	Complete
Create one slide on evaluation	July 25th	July 26th	1 day	Avantika	Complete
Conclusion					
Write conclusion of paper	July 20th	July 26th	6 days	Sharon	Complete
Create one conclusion slide	July 20th	July 26th	6 days	Sharon	Complete
Compile report sections and presentation slides, edit	July 20th	July 26th	6 days	Sharon	Complete
Confirm completion of milestones at each step	July 20th	July 26th	6 days	Sharon	Complete