Glitz Glam and Grouping

An analysis of Sephora beauty products to understand which variables impact the rating a customer will give

Aalayah Honablue
aah16@hood.edu
Fall 2022

December 13, 2022

1 Abstract

Makeup brands such as Estée Lauder, L'Oréal, and Dior need to process their customers data to better the experience of their products and also evaluate future products to bring a more personalized experince for consumers. The Glitz, Glam and Grouping project will use clustering techniques to analyze what beauty products customers love when there are different factors.

2 Introduction

Clustering, which refers to the process of taking different characteristics and grouping the data points based on those characteristics. These groups are referred to as clusters and they allow for data miners to seamlessly divide the data into subsets where a more informed decisions and respective behaviors.[1] In data mining there are numerous types of clustering algorithms and functions including K-mean clustering, Mean-Shift

[1] Pashikanti, P. (2022, June 22). Clustering in data mining. GeeksforGeeks. Retrieved October 7, 2022, from https://www.geeksforgeeks.org/clustering-in-data-mining/

Clustering,Density-Based Spatial Clustering of Applications with Noise, Expectation-Maximization Clustering,and Agglomerative Hierarchical Clustering.[2]

The makeup industry is a competitive business and it is constantly changing. There are numerous marketing and makeup brands that could analyze this data to gather a better understanding of what products customers give the highest rating to and have a good oversight of their business operations. The data set used was created by Raghad Alharbi who wanted to use web scaping methods to collect more than 1,000 useful records from the Sephora website. This data set is rich with information as it has 21 variables and 9169 observation also it was created in 2020. This data set can give many different ways to analyze the data and gather a better understanding of what products customer's are enjoying.

3 Plan of Procedure

This project will look into the Sephora website data set to get an understanding of what beauty products perform the best. To understand this there will be an analysis of the relationship between ratings, price, categories, number of reviews and the price of a product. There will also be an analysis on the means of purchasing a product by looking to see if limited edition, Sephora exclusive, or online only has an impact.

4 Methodology

Origin of the Data Set: This data set originated from Keggle. The data was created by Raghad Alharbi who wanted to use web scaping methods to collect more than 1,000 useful records from the Sephora website. This data set is rich with information as it has 21 variables and 9169 observation also it was created in 2020.

[2] Seif, G. (2022, February 11). The 5 clustering algorithms data scientists need to know. Medium. Retrieved October 7, 2022, from https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

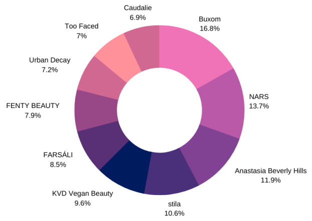

Figure 1: Top 10 Makeup Brands Loved by Customers

Attribute Information:

Input Variables: Product Description

- id: The product on Sephora's website
- brand: The brand of the product at Sephora's website
- category: The category of the product at Sephora's website
- name: The name of the product at Sephora's website
- size: The size of the product
- price: The price of the product
- valuePrice: The value price of the product (for discounted products
- url: The URL link of the product
- howtouse: The instructions of the product if available
- ingredients: The ingredients of the product if available
- options: The options available on the website for the product like colors and sizes
- details: The details of the product available on the website

Input Variables: Customer Review

- rating: The rating of the product
- numberofreviews: The number of reviews of the product
- love: The number of people loving the product

Input Variables: Product Purchasing Options

- MarketingFlags: The Marketing Flags of the product from the website if they were exclusive or sold online only
- MarketingFlagsContent: The kinds of Marketing Flags of the product
- onlineonly: If the product is sold online only
- exclusive: If the product is sold exclusively on Sephora's website
- limitedEdition: If the product is limited edition
- limitedtimeoffer: If the product has a limited time offer

Method: In this project grouping the categories will help to organize the large set of data and generate an insight to the products being sold at Sephora.

5 Conclusion

Analysing the data from this data set by using various clustering algorithms I found that the most liiked products products at Sephora based on price, brand, ingredients, or number of reviews impacts the rating of a customer are Drunk Elephant, KVD, Huda,Stila, Buxom, Fenty, ABH, Urban Decay and Olaplex. Although these were the most liked brands it was important to group the items into categories such as skin car, makeup, hair car, and body. With these results I found that top 10 hair care products were Bumble andd bumble, Living Proof, T3, StackedSkincare, OUAI, Herbivore, Briogeo, Dyson, The Ordinary, and Olaplex. The top ten makeup brands were Buxom, NARS, Anastasia Beverly Hills, stila, KVD Vegan Beauty, FARSÁLI, FENTY BEAUTY, Urban Decay, Too Faced, and Caudalie. Of these products there are a few that overlap with the top ten overall brands across differet categories but as we can see there are additional ones that would not have been seen if I did not group the categories.