Sephora Review Data Appendix

This data appendix offers an overview of each variable in **combined_clean_data.csv**. The dataset itself was compiled through web scraping reviews of 18 Sephora.com products, with half belonging to celebrity brands and the other half unaffiliated with celebrities. For each variable, this data appendix describes the variable and provides summary statistics and visualizations, when applicable. The unit of observation is a Sephora.com review.

Title: The title of the review. We cannot perform any visualizations due to the thousands of unique titles.

	Title
count	21400
unique	15062
top	Love it
freq	252

dtype: object

ReviewText: The content of the review itself (text data). We cannot perform any visualizations due to the thousands of unique ReviewText.

	ReviewText
count	28918
unique	28759
top	I have all four shades purchased with my own m...
freq	4

dtype: object

UserID: The unique username of the individual who posted the review. We cannot perform any visualizations due to the thousands of unique UserIDs.

	UserID
count	28604
unique	24331
top	Angei2023
freq	9

dtype: object

Date: The date the review was posted, formatted as DD MMM YYYY (i.e., 02 Feb 2025). We cannot perform any meaningful visualizations due to the thousands of unique dates.

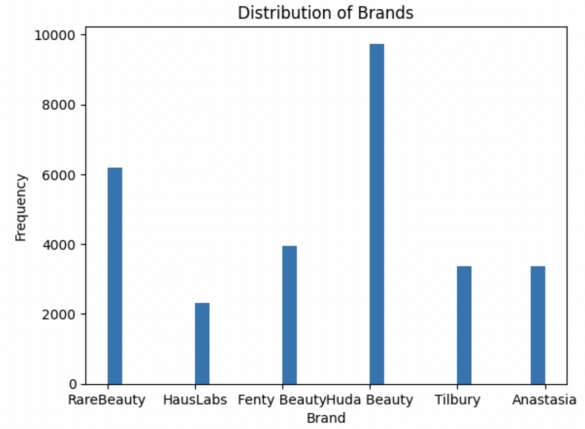
	Date
count	28935
unique	1781
top	4 Apr 2024
freq	359

dtype: object

Brand: The brand of the product. Options are limited to RareBeauty, HausLabs, Fenty Beauty, Huda Beauty, Tilbury, and Anastasia.

	Brand
count	28935
unique	6
top	Huda Beauty
freq	9742

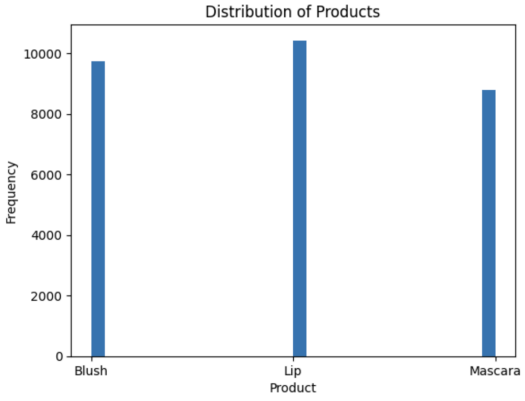
dtype: object



Product: The description of the makeup product itself. Options are limited to Blush, Lip, and Mascara, as we scraped reviews from these 3 products per brand.

	Product
count	28935
unique	3
top	Lip
freq	10422

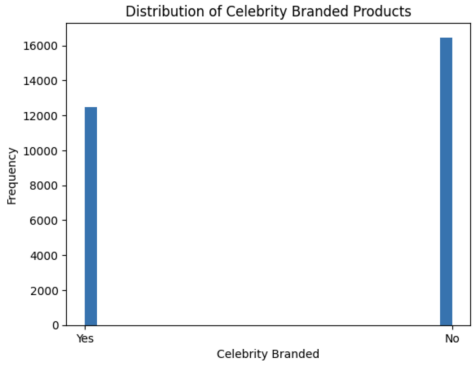
dtype: object



CelebrityBranded: A yes-no binary describing whether the product is affiliated with a celebrity or not. ‘Yes’ means the product is celebrity branded, ‘no’ means the product is not connected to a celebrity.

	CelebrityBranded
count	28935
unique	2
top	No
freq	16458

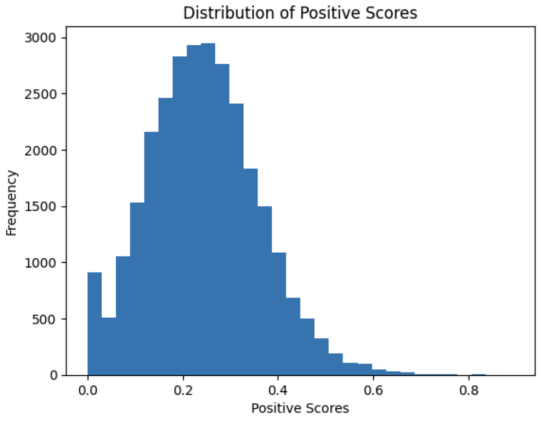
dtype: object



PositiveScore: The proportion of ReviewText classified as positive by VADER sentiment analysis. When combined with the NegativeScore and NeutralScore, the total equals 1.

	PositiveScore
count	28935.000000
mean	0.243421
std	0.116478
min	0.000000
25%	0.163000
50%	0.239000
75%	0.318000
max	0.895000

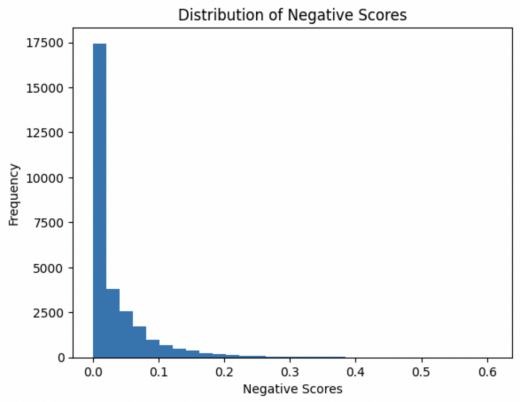
dtype: float64



NegativeScore: The proportion of ReviewText classified as negative by VADER sentiment analysis. When combined with the PositiveScore and NeutralScore, the total equals 1.

	NegativeScore
count	28935.000000
mean	0.030260
std	0.052073
min	0.000000
25%	0.000000
50%	0.000000
75%	0.043000
max	0.608000

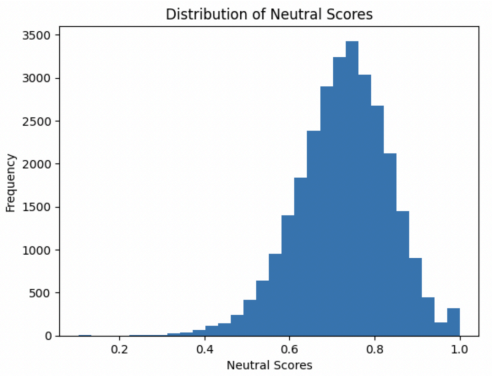
dtype: float64



NeutralScore: The proportion of ReviewText classified as neutral or lacking clear sentiment by VADER sentiment analysis. When combined with the NegativeScore and PositiveScore, the total equals 1.

	NeutralScore
count	28935.000000
mean	0.726318
std	0.107719
min	0.105000
25%	0.659000
50%	0.732000
75%	0.800000
max	1.000000

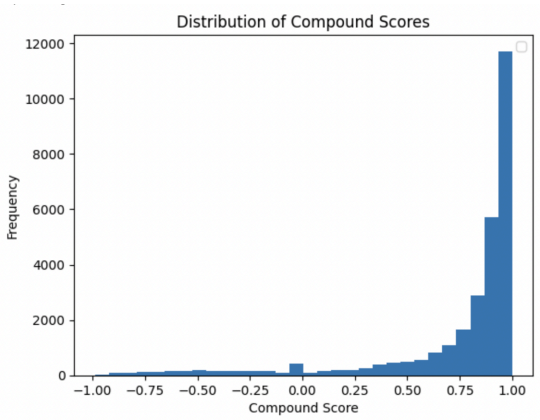
dtype: float64



CompoundScore: The aggregate sentiment score, ranging from -1 to +1.

	CompoundScore
count	28935.000000
mean	0.748866
std	0.383558
min	-0.988600
25%	0.745050
50%	0.908100
75%	0.960500
max	0.999600

dtype: float64



Sentiment: Categorizes the compound score as negative, neutral, or positive. A -1 to -0.5 compound score is considered negative, a -0.5 to 0.5 compound score is considered neutral, and a 0.5 to 1 is considered positive.

	Sentiment
count	28935
unique	3
top	Positive
freq	24730

dtype: object