Sephora Insights:

from web scraping
to EDA and NLP-ML reviews analysis

May 2023

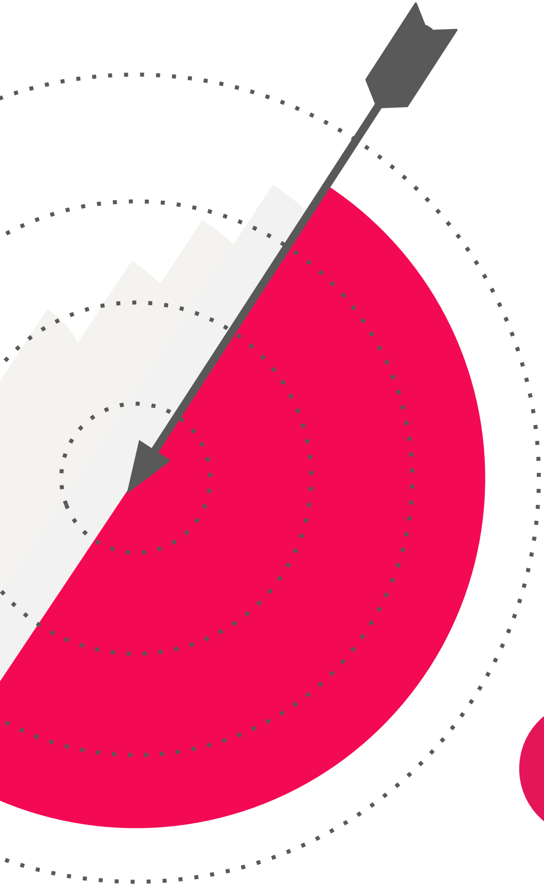
TABLE OF CONTENT

Link to the full project
https://github.com/nadyinky/sephora-analysis

01.

Purpose and objectives

What are the main purpose and objectives of this project?



Project purpose

Demonstrate proficiency in data analysis, programming languages such as Python and SQL, and machine learning techniques to inform and improve business decisions using large datasets

Scraping data from the Sephora website, **cleaning** and **storing** it in a SQL database

Performing exploratory **data analysis** (EDA) to identify trends and patterns

Applying **natural language processing** (NLP) and **machine learning** techniques to analyze consumer reviews of skincare products

02. Telling a story with data

Data collection results and highlights from exploratory data analysis and review analysis

Data scraping: result in numbers

0

products without reviews

Every product in this store has at least one review and a non-zero rating

5

minutes

The average time it takes a scraper to gather information about products

8494

products

These are all the products of the online store at the time of data collection. That's more than 300 brands

1 mln +

reviews

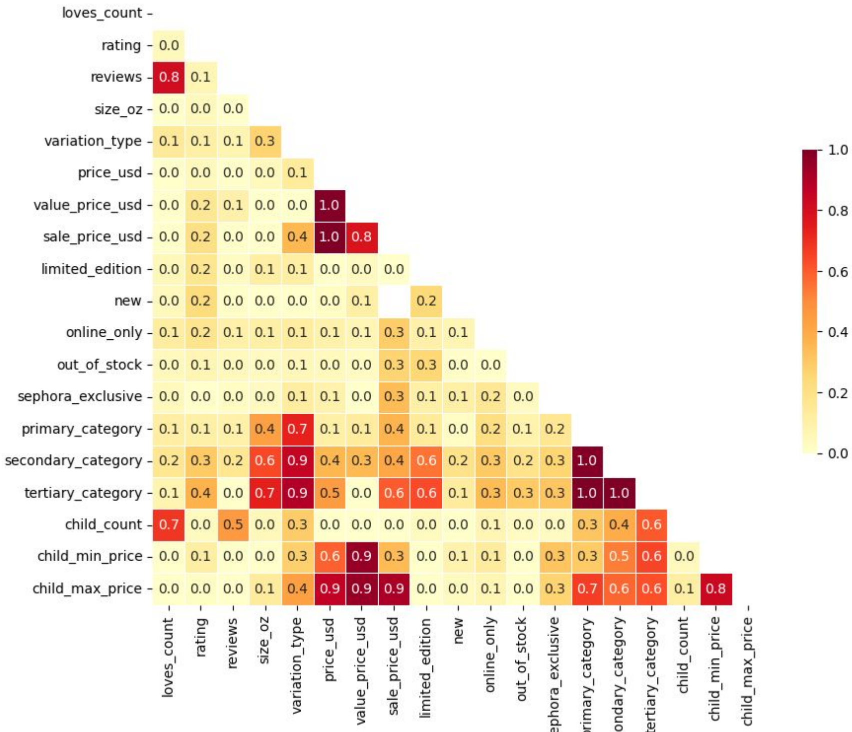
This is how many customer reviews from the Skincare category have been collected. These are all customer reviews for this category.

Correlation analysis
of Pearson’s and Phik (ϕk) matrices

- The correlation analysis showed no meaningful significant correlations between prices and any of the collected data

- The most interesting thing seems to be the correlation between the number of reviews and likes, which at the same time is *not related* to the rating

- Reviews remain the most eloquent for a potential customer



Phik correlation matrix

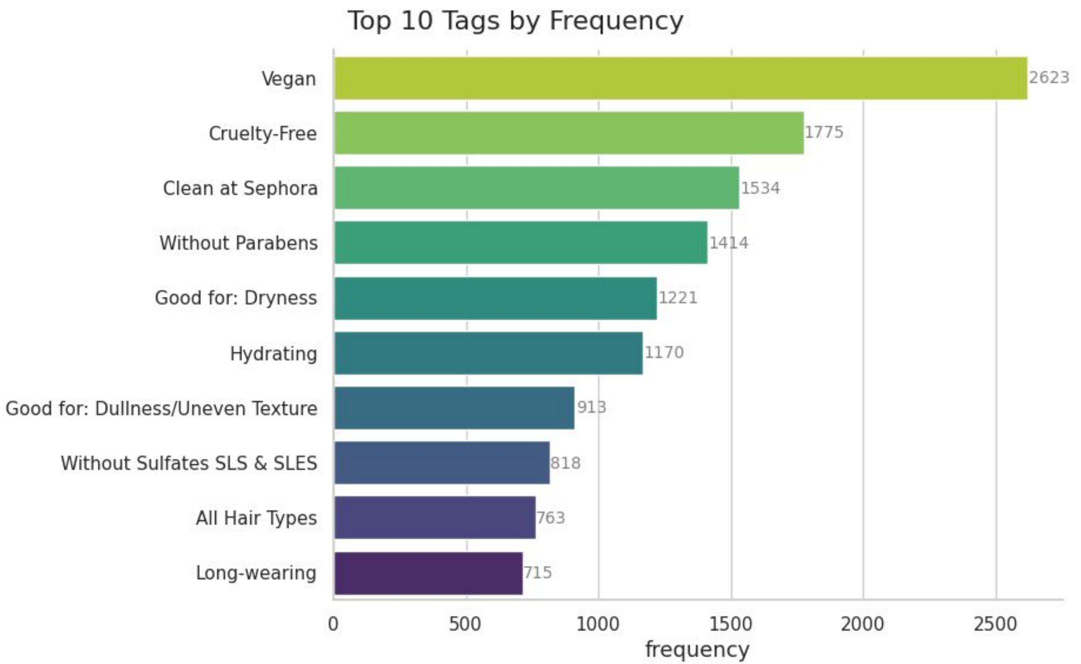
Researching **tags and labels**

Single tags:
the tags reflect well the most important request for safety, environmental friendliness and hydration

Tag pairs:
it is worth paying attention to perfume tags to enhance the shopping experience

The strong correlation between the **"limited edition" label** of products and the category is confirmed:

- almost 50% in the Values & Gift Sets category
- Candles and Eyes have another 15% each in the category



Product distribution by category

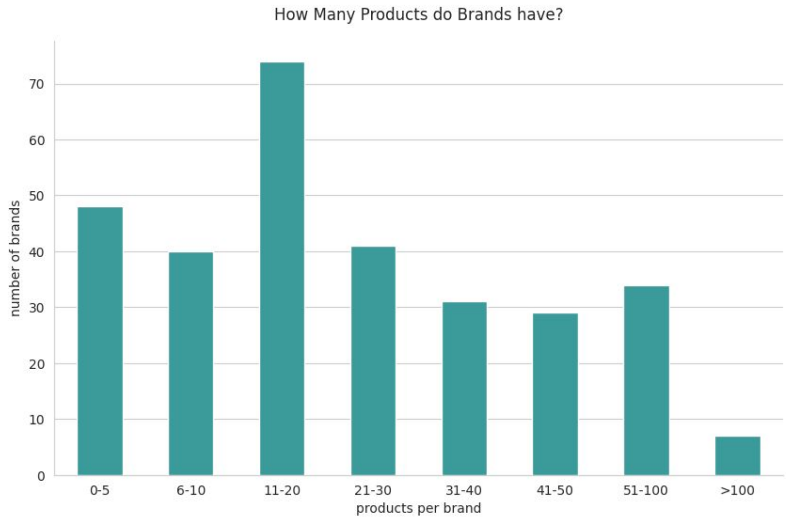
- The **assortment follows** the distribution of demand in the **global beauty market**: Skincare, Makeup, Hair and Fragrance take the leading positions. Although the **Hair** group shows an area of **potential improvement**
- For some reason, Sephora has almost **no hair coloring** products, although according to research it is one the top 4 most in-demand products
- **Tags** associated with the Hair category are also **worth improving**, following research on the U.S. hair and scalp market

*full-size visualization in a notebook

Price analysis: distribution by category

- The most widely represented **Makeup** category has an average item value of **$32.8**, which is almost **half of the $60.5** average value of the second largest category, **Skincare**

- The **Fragrance and Hair** categories have almost equal numbers of products, but the average product cost of **$87.3** of the former **is twice as much** as the latter **$42.8**

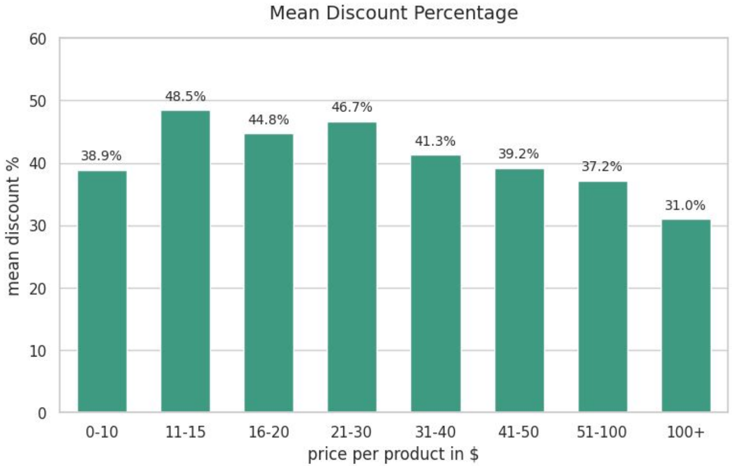


Average number of **products per brand**

The largest number of brands has **between 11 and 20** products. Sephora also has many brands with as few as 5 products

Item price ranges with the **highest discount**

$11-$15 and $16-$20 products have the highest average discount value



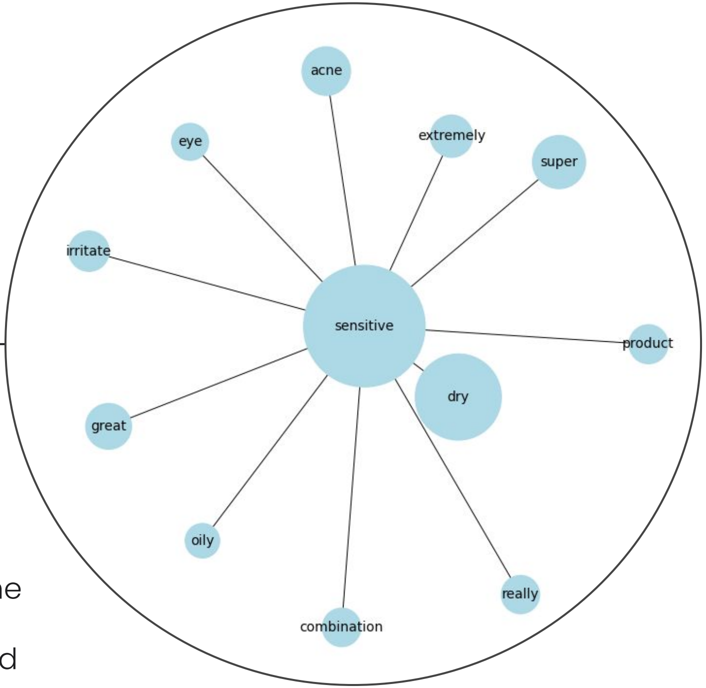
N-gram analysis: process and result

text cleaning

- Getting rid of extraneous characters, punctuation, words with possible misprints, and 'stopwords'
- Using the **lemmatization** technique: Caring -> Care

tokenize & visualize

- Word **tokenization**
- Identify and visualize the most frequent **unigrams**, **bigrams**, and **trigrams**

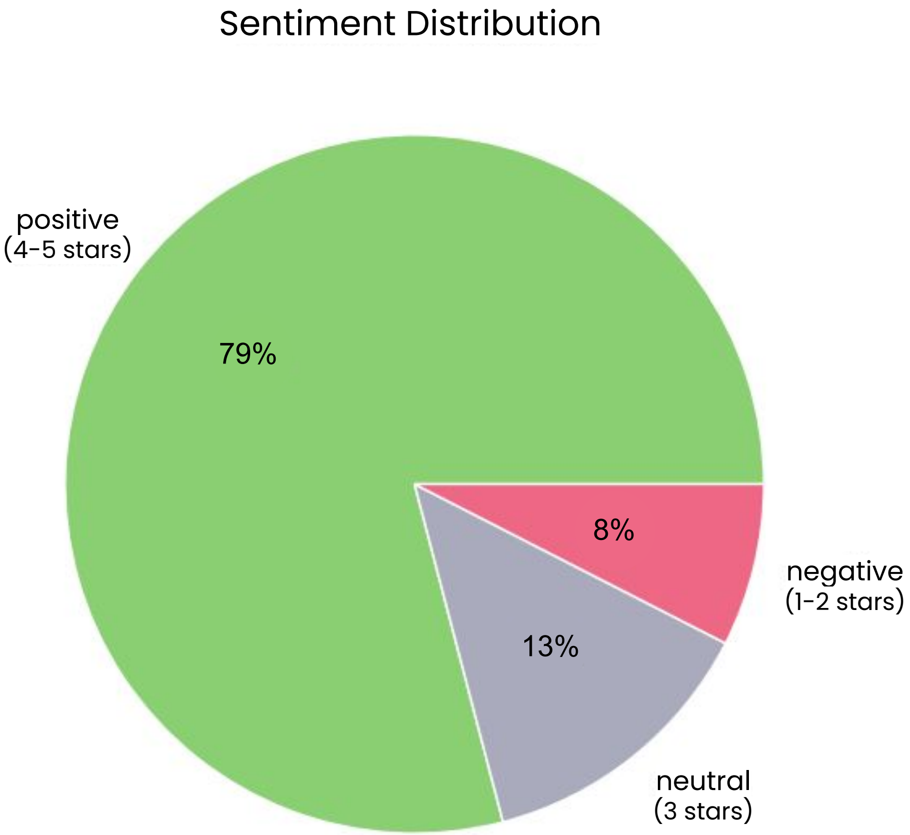


bigram visualization

Sentiment prediction model:
overcoming class imbalance

To create a model predicting review sentiment, I reduced the training dataset dimensionality from 50k to 12k reviews and employed undersampling to address the class imbalance.

In addition, I conducted tests on multiple models, carefully selected the most suitable one, and fine-tuned its hyperparameters to achieve optimal performance.



Sentiment prediction model:

error analysis and result

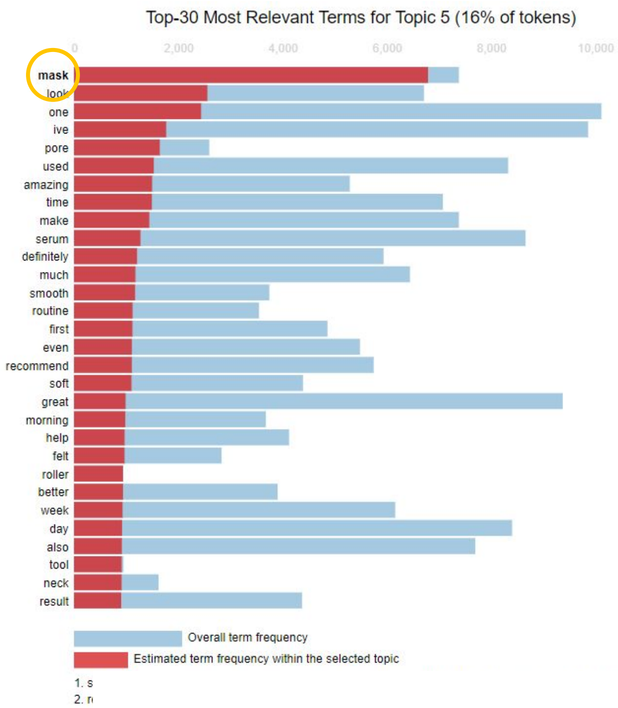
```
ml_predict("Do not waste your money! This product did nothing for me.")

>>> 'negative', 0.99
```

demonstration of a real example of classification

The completed model takes in a review and outputs both its corresponding sentiment class and the corresponding probability with which it is classified.

Error analysis showed that the accuracy of classification is 81%, the classification of positive and negative reviews is slightly better (85%), and neutral reviews slightly worse (77%)

Topic modeling



- I pre-processed text and used unsupervised machine learning technique that helps identify clusters of words (topics)
- These topics cover five main product types, like mask, lip, etc. You can research words by generated topics and discover hidden trends or characteristics of each group (topic)

*Screenshot of interactive visualization from notebook: group n.5 is “Mask”

03. Conclusion

Summarizing the project, stages, and tools

Conclusion

Created **concurrent scrapers** to collect **product** information and customer **reviews**.
Scraped data saved in **PostgreSQL** database and CSV files

Web scraping

BeautifulSoup, PostgreSQL, Requests, Pydantic

An **in-depth analysis** was conducted that provided insight into products, prices, and trends

Exploratory data analysis

pandas, colab, jupyter, NumPy, matplotlib, seaborn

Reviews from the Skincare category were analyzed, a **sentiment prediction** model was created and **topic modeling** was performed

NLP + ML reviews analysis

scikit learn, Natural Language Toolkit NLTK, GENSIM topic modelling for humans

Thanks!

Link to the full project
https://github.com/nadyinky/sephora-analysis