**BT5153 Group Project Proposal**

**eCommerce NLP: Price and Ranking Prediction on Beauty Products**

*Group 13*

*Anqi Dong (A0218944W), Hui Xinyao (A0148006M), Li Xinyi (A0218908W)，*

*Zhang Yixuan (A0218975M), Zhao Dongyu ()*

1. **Background and Overview**

Consumer’s buying behaviors were dramatically changed by COVID19. They tend to shop more online than offline, as a result, eCommerce is counting a higher percentage in overall global commerce transactions. Consequently, more online reviews were left by brand consumers.

By studying the reviews, brands could understand the true voice of their customers, receive first-hand product opinions, and eventually engage with them timely.

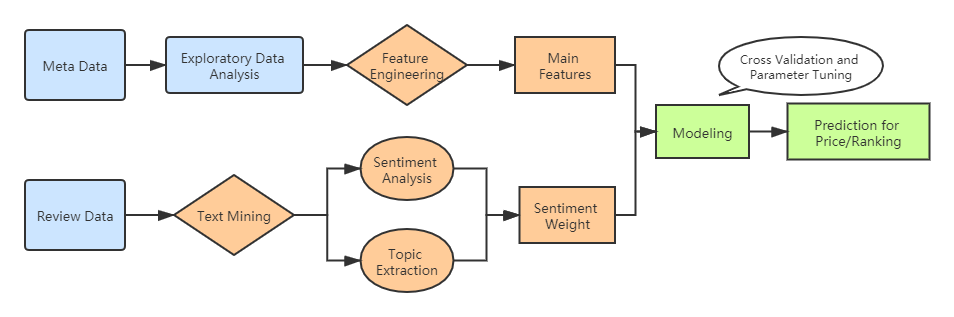
We plan to use 2014 Amazon eCommerce data in beauty: hair-care shampoo category to analyze the online reviews left on those products and to unveil the myth between product reviews and product ranking/pricing positioning. If the problem of data insufficiency occurred, some back-up product categories will be adopted and explored correspondingly.

1. **Data Exploration**

Our dataset was collected from an online archive Amazon Product Data[[1]](#footnote-2). In alignment with the objective, we chose the shampoo product data from the pre-category *Beauty*.

**Dataset Description**

This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014. It includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs). The summary of the dataset can be found in Table 1 below. In this work, our analysis will base on review data and metadata (2014) of Beauty.



**Fig.1** Project Flow

**Table 1.** Description of Metadata and Reviews Dataset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **meta** | asin | categories | description | imUrl | salesRank | title | price | brand |
| **reviews** | asin | helpful | overall (rating) | reviewText | reviewTime | reviewerID | reviewerName | summary |

**Data Preprocessing**

The raw datasets are difficult to load due to its invalid json format and large size. Therefore, we first employed PySpark to handle the data loading problem. Following that, we filtered the shampoo product from the Beauty category, and only kept those whose title contains ‘shampoo’ or ‘conditioner’.

The description of selected products and their reviews is shown in Table 2 below.

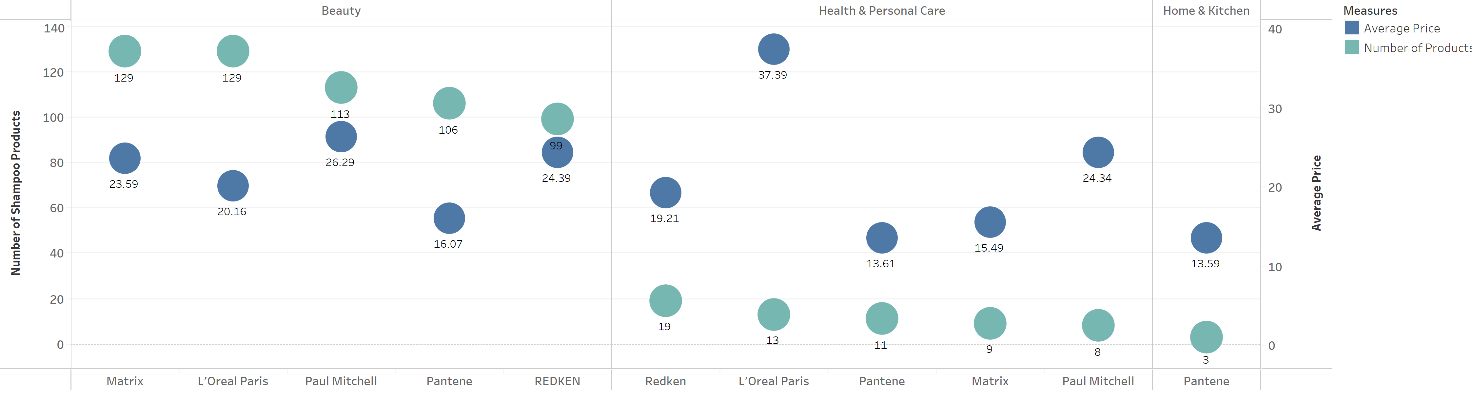
**Table 2.** Description of Filtered Products (Shampoo and Conditioner) and Their Reviews

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Features** | **asin** | **title** | **brand** | **price** | **salesRank** |
| **Count** | 18805 | 18805 | 11419 | 14857 | 15863 |
| **Median/Mode** | -- | -- | Matrix | 17.19 | 166653 |

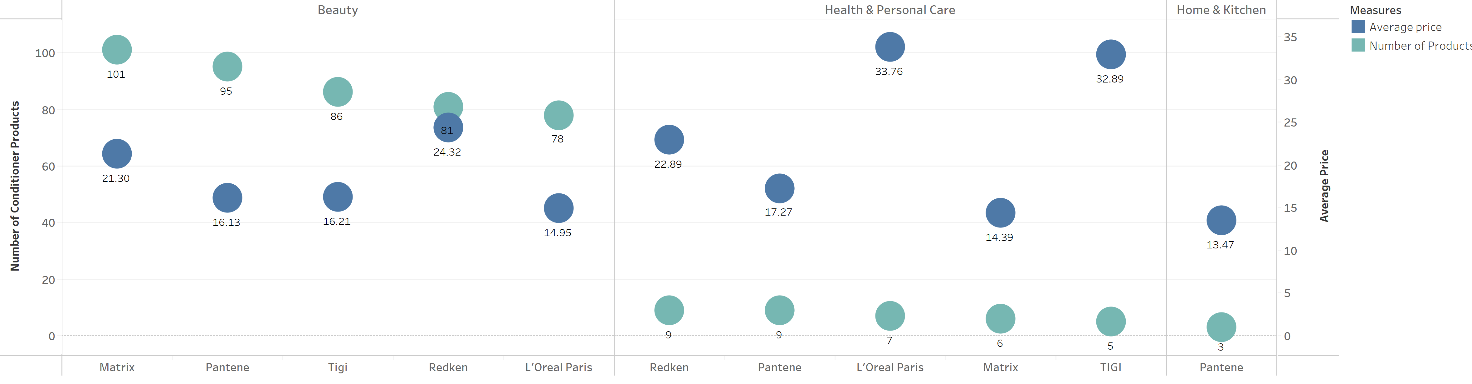
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Features** | **asin** | **num of reviews per product** | **number of products with reviews≥5** | **reviewText** | **helpful vote** | **overall**  **(rating)** |
| **Count** | 16590 | -- | 5381(32.4%) | 162227 | 143256 | 143256 |
| **Median** | -- | 2 | -- | 239 (length) | 0 | 5.0 |
| **Max** | -- | 1349 | -- | 15461 (length) | 502 | 5.0 |
| **Min** | -- | 1 | -- | 2 (length) | 0 | 1.0 |

1. **Exploratory Data Analysis**

By conducting exploratory data analysis, we produced tables and charts to understand the basic data distribution in meta and review data sets. Fig.2 and Fig.3 show the number and average price of shampoo and conditioner products across most popular brands by category. Table 3 shows the top 10 brands with best ranking performance in Beauty Category. Fig.4 and Fig.5 show the price distribution for top 10 brands with most products.



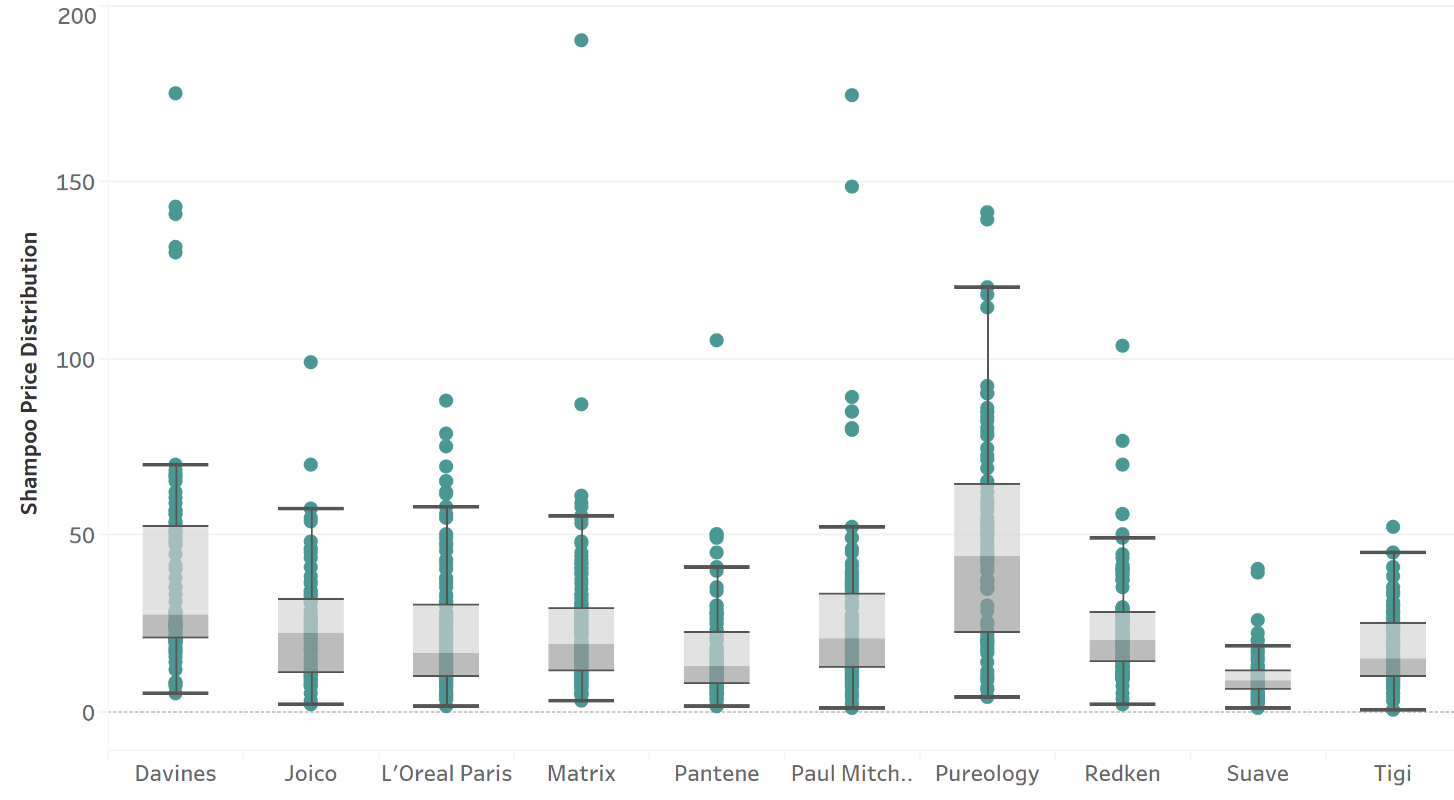
**Fig.2** Number and Average Price of Shampoo Products under Brands with Most Products by Category



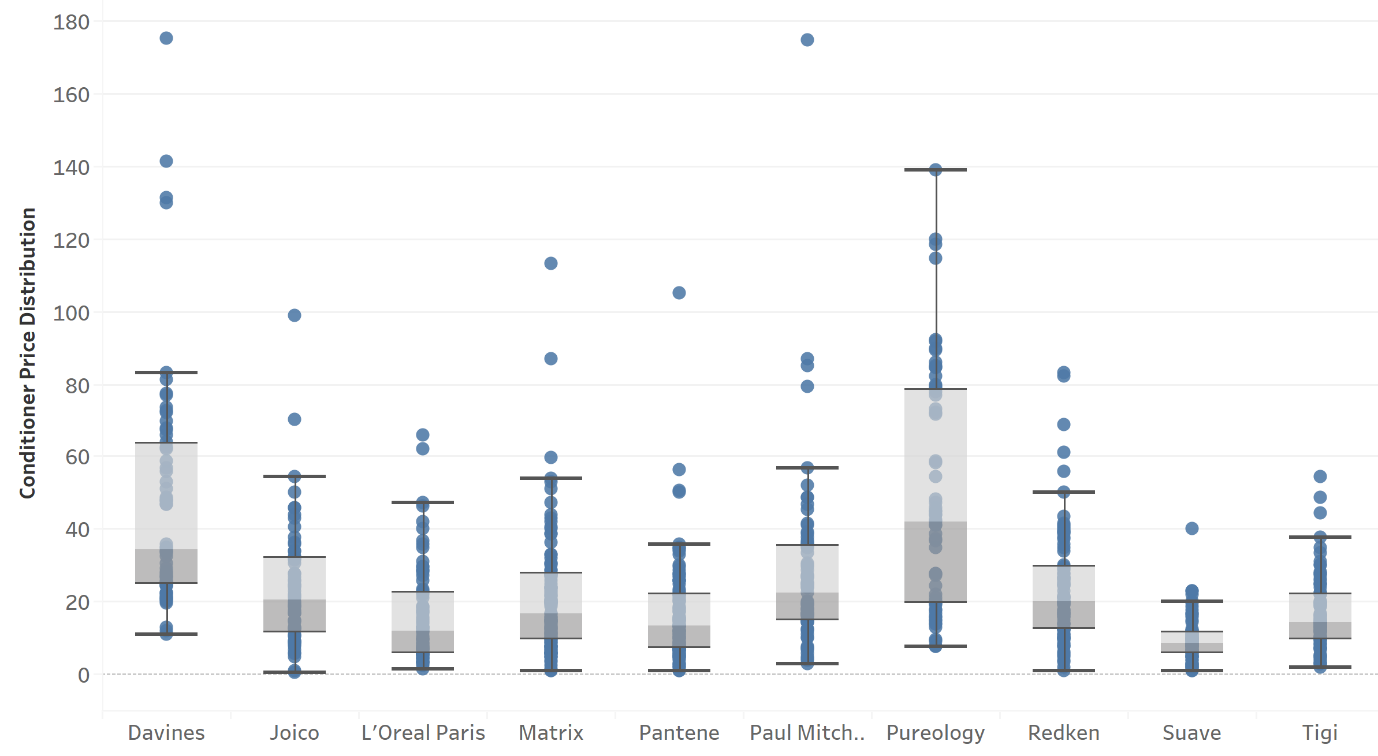
**Fig.3** Number and Average Price of Conditioner Products under Brands with Most Products by Category

**Table3.** Top 10 Brands with Best Ranking Products Per Category Beauty

|  |  |
| --- | --- |
| **Shampoo** | **Conditioner** |
| Ultrax Laborotories | WEN by Chaz Dean |
| Biolage by Matrix | Aquage |
| UltraSwim | Moroccanoil |
| All Soft | Head & Shoulders |
| Redken blonde glam | Aussie |
| GK | Bumble and Bumble |
| Lipogaine | Herbal Essence |
| Pharmaceutic al Specialties | Pureology |
| Original Sprout | Nexxus |
| Straight Arrow | KMS |



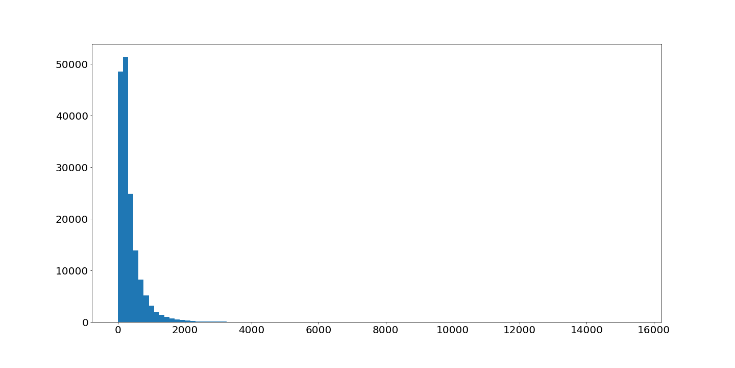
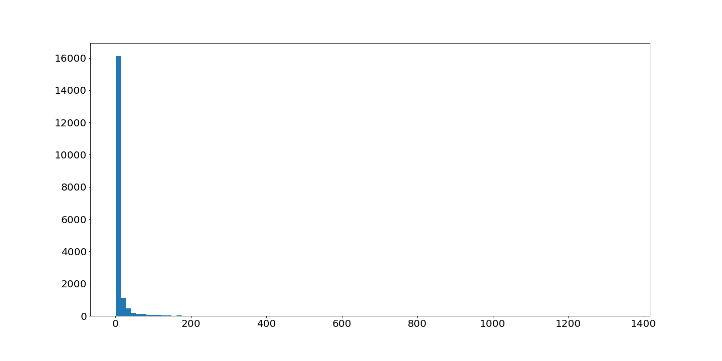
**Fig.4** Price Distribution for Top 10 Brands with Most Shampoo Products



**Fig.5** Price Distribution for Top 10 Brands with Most Conditioner Products

The distribution for the number of reviews per product can be found in Fig.6 below. About 35.6% of the products only have one review. 32.4% of the products have greater than or equal to 5 reviews. The number of products with at least one review is 16,590 and the total quantity of reviews is 162,227.

The distribution of length of reviews is plotted in Fig.7 below. A majority of reviews are short with 20-40 words whilst reviews with a size more than 4000 characters exist as well. Note that customers who leave long reviews tend to be either very positive or negative.



**Fig.6** Histogram for number of reviews per product  **Fig.7** Character length of reviews distribution

1. **Project Questions to Address**

The key questions that we want to answer via this study are:

* Are online product reviews usually sentimental? And how do those sentiments differ across brands and products?
* Can we extract meaningful features (such as sentiments, product quality satisfaction, and delivery preference etc.) out of the reviews?
* With extracted NLP features and other product features, can we build up a predictive model to successfully predict product ranking/pricing positioning?
* How would trendy industrial specific keywords (such as “anti-dandruff”, “repair”, “moisturize”, and “strengthen”) potentially influence business implications?

By answering above four questions, we are able to establish a meaningful eCommerce review analytics process for key category brands. Eventually brands could leverage the analytics to better engage with their customers.

1. **Text Mining**

To explore the key objectives and validate our assumptions, we will employ a set of machine learning techniques integrated with text mining methods which involves natural language processing.

After the stage of data preprocessing and feature engineering, we will have the main features extracted from the meta data set for models fitting. Regarding the review data, we will be performing a two-part text mining process, including Sentiment Analysis and Topic Extraction.

The BERT model is a recently popular model used to extract keywords from texts. We will leverage this model to extract keywords for shampoo and conditioner’s effects. LDA will also be possibly employed to explore the most frequent keywords in the overall reviews of a product. A sentiment weight for each topic word of each product will be generated via Sentiment Analysis.

Finally, we will conduct the prediction part by applying machine learning tools. Several features such as brands, product scores, number of reviews and sentiment weight will be model inputs, while price and ranking will be outputs of prediction. Possible methods are Random Forest, XGBoost, KNN, SVM, Bayes Network, deep learning models (e.g. RNN) and so on. Meanwhile, the most influencing factors (e.g. keywords from reviews) will be extracted to bring more business sense for brands.

1. <http://jmcauley.ucsd.edu/data/amazon/links.html> [↑](#footnote-ref-2)