Lucas Reynolds
Springboard Data Science Career Track
January 2020 Cohort
Capstone Project 2

Milestone Report 1

Overview

Fragrance websites can be complicated due to the number of brands, scents, and products. Targeted product recommendations not only improve customer experience, but can increase the effectiveness and ROI of marketing campaigns

Project Outline

Steps Completed

1. Data Acquisition:
    a. Scrape, parse, and organize the product review data
2. Data Wrangling:
    a. Clean, organize, and preprocess the data
3. Exploratory Analysis:
    a. Visual and descriptive exploration
    b. Inferential statistics

Steps Remaining

4. Modeling:
    a. Determine best recommendation model
    b. Evaluate results
5. Outcomes and Recommendations:
    a. Present final results and outcomes
    b. Provide any additional recommendations or next steps

Dataset

Basenotes.net provides information about different brands and scents, including user reviews. Using BeautifulSoup and Requests (Python libraries), I built a custom web scraper to compile and parse reviews from Basenotes.net. The review information includes the following fields:

- Rating value: a rating of 1, 2, or 3
- Review ID: unique review identifier
- User ID: unique identifier of user who submitted the review
- Username: name of user who submitted the review
- User location: country of user, if provided

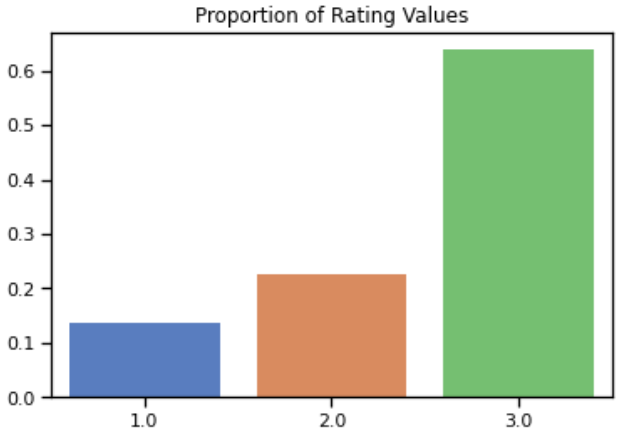
- Scent ID: unique identifier of the scent reviewed
- Scent name: name of the scent reviewed
- Scent brand: brand name of scent reviewed
- Review text: text included with the review

Data from 159,324 reviews were parsed and organized in a Pandas dataframe and saved as a CSV file. User locations are not required, so about 20% of this field had missing or null values. To avoid dropping 20% of the reviews, I opted to fill any missing user locations as “no_location.” The remaining fields each had 1% or less of their values missing or null. To maintain simplicity and since there was a small proportion of missing values for these fields I dropped the remaining rows with nulls.

Though ratings were on a scale of 1, 2, or 3, some of the rows had a value of 7 in the rating field. Less than 0.07% of rows had this rating value, so I chose to drop them. After removing and filling null values, 157,373 rows remained, close to 99% of the total data scraped. I will review and test the web scraper and parser to determine the cause of the null values and ratings with a value of 7 and update in a subsequent report.

Ratings

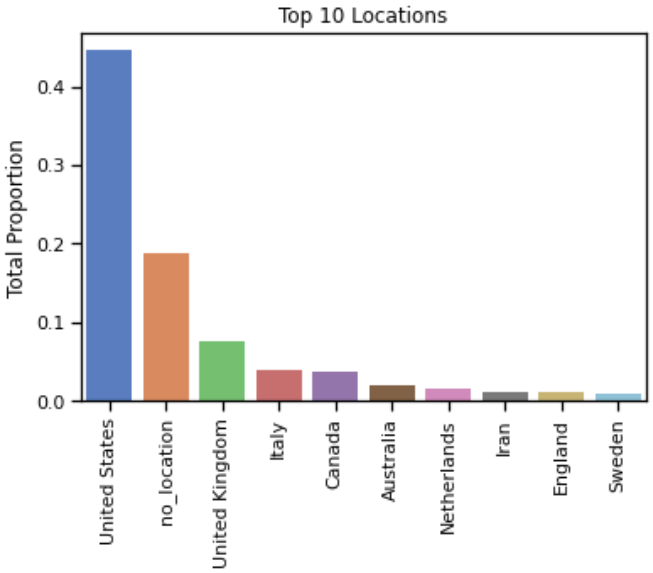
Users provided a rating of 1, 2, or 3 symbolized by a thumb down, thumb horizontal or neutral, or a thumb up respectively. About 13.6% of reviews had a rating of 1, 22.6% had a rating of 2, and 63.8% had a rating of 3.



User Country

About 19% of reviews did not have a user location included. For countries that were included, about 45% of reviews were from the United States, followed by the United Kingdom at about 8% and Italy near 4%.

With t-tests for significant average difference, I determined the countries with the largest significant increase and decrease in average rating.



Top 5 Countries by Significant Increase in Avg. Rating	
Country	**Significant Avg. Difference**
Guyana	0.3099
China	0.2736
Denmark	0.2165
Slovakia	0.2051
Nigeria	0.1908

Top 5 Countries by Significant Decrease in Avg. Rating	
Country	**Significant Avg. Difference**
Jordan	-0.3746
Kuwait	-0.3360
Cuba	-0.3194
Morocco	-0.2612
Saudi Arabia	-0.2527

Scent

There were 17,345 different scents and the reviews were well spread among the scents. Each scent had less than 1% of the reviews.

Using t-test for significant average difference again, the top 5 scents with the largest increase and decrease in average rating were determined. The size of average difference was about double for scents with the largest decrease compared to the largest increase. I was surprised to see a scent by Calvin Klein, cK Free, as the scent with the largest significant average decrease in rating. Reviewers mentioned that the scent was very weak and did not last very long.

Top 5 Scents by Significant Increase in Avg. Rating	
Scent	**Significant Avg. Difference**
Agua Lavanda	0.4975
Moschino pour Homme	0.4975
Bois des Iles Parfum	0.4271
Polo Crest	0.4117
Italian Cypress	0.4023

Top 5 Scents by Significant Decrease in Avg. Rating	
Scent	**Significant Avg. Difference**
cK Free	-1.0786
Hot Water	-0.8695
Exceptional Because You Are for Men	-0.8498
I am King	-0.7572
Secretions Magnifique	-0.7274

Brand

There were 2,095 different brands included in the reviews and the reviews were well spread among the brands. The most reviewed brand was Guerlain with about 3% of the reviews

The top 5 brands with the largest positive and negative average differences were determined using t-tests for significant average difference.

Top 5 Brands by Significant Increase in Avg. Rating	
Scent	**Significant Avg. Difference**
Sultan Pasha	0.3996
Fort and Manle	0.3922
Revillon	0.3821
Fragrances of France	0.3724
Sospiro	0.3406

Top 5 Brands by Significant Decrease in Avg. Rating	
Scent	**Significant Avg. Difference**
Exceptional	-0.7601
Porsche	-0.6092
Illuminum	-0.5399
Pierre Guillaume	-0.4806
People of the Labyrinths	-0.4742