Predicting Sephora Product Ratings

Madison Martin

2023-08-07

Introduction

This project aims to predict the user ratings for beauty products sold on the popular website for beauty retailer Sephora. Specifically, it aims to identify the role product attributes like brand, price, online popularity, limited edition offerings, and exclusivity have on customer satisfaction with the products.

Research Questions:

What product characteristics (if any) play a role in predicting the customer satisfaction with the product, as evidenced by the product's rating? Do limited edition products tend to have higher or lower ratings? Does interest in the product seem to correlate with higher ratings?

Sourcing the Data

This data set was sourced from the Kaggle website and uploaded by user Raghad Alharbi, which you can find at this link here. It was collected utilizing web scraping methods during the month of April 2020 from the Sepohra US website.

Download the Data

```
# download the product data from sephora website
# Load the necessary libraries
library(readr)
library(tidyr)
library(ggplot2)
library(tidyverse)
library(dplyr)
library(knitr)

# read in csv
sephora_df <- read_csv(file = "sephora_website_dataset.csv")
head(sephora_df)
```

```
## # A tibble: 6 x 21
##        id brand        category name  size  rating number_of_reviews  love price
##     <dbl> <chr>        <chr>    <chr> <chr>  <dbl>             <dbl> <dbl> <dbl>
## 1 2218774 Acqua Di Pa~ Fragran~ Blu ~ 5 x ~    4                 4  3002    66
```

```
## 2 2044816 Acqua Di Pa~ Cologne  Colo~ 0.7 ~    4.5                76  2700    66
## 3 1417567 Acqua Di Pa~ Perfume  Aran~ 5 oz~    4.5                26  2600   180
## 4 1417617 Acqua Di Pa~ Perfume  Mirt~ 2.5 ~    4.5                23  2900   120
## 5 2218766 Acqua Di Pa~ Fragran~ Colo~ 5 x ~    3.5                 2   943    72
## 6 1417609 Acqua Di Pa~ Perfume  Fico~ 5 oz~    4.5                79  2600   180
## # i 12 more variables: value_price <dbl>, URL <chr>, MarketingFlags <lgl>,
## #   MarketingFlags_content <chr>, options <chr>, details <chr>,
## #   how_to_use <chr>, ingredients <chr>, online_only <dbl>, exclusive <dbl>,
## #   limited_edition <dbl>, limited_time_offer <dbl>
```

Precleaning for Analysis

```
# take out the vars we don't think we will use in the analysis
# going to remove heavy text ones: url, instructions, ingredients
product_df <- sephora_df %>%
  select(- URL, - how_to_use, -ingredients)
# view(product_df)
```

I'm already seeing that some products have a different price vs. value_price. I'm guessing these are mostly sets/kits (as you can see if you head the df), but the data documentation does not state this explicitly. I am going to make a new column for these products specifically and keep it in mind as I go through my exploratory analyses.

```
# add column to identify products who have a different price from the "value price"
product_df <- product_df %>%
  mutate(deal = ifelse(value_price - price > 0, 1, 0))
```

Exploratory Data Analysis

First, let's get a little bit more information about the many products in our data set.

```
# get the average price, number of loves, and rating
mean(product_df$price)
```

```
## [1] 50.06324
```

```
mean(product_df$rating)
```

```
## [1] 3.99002
```

```
mean(product_df$number_of_reviews)
```

```
## [1] 282.1392
```

```
median(product_df$number_of_reviews)
```

```
## [1] 46
```

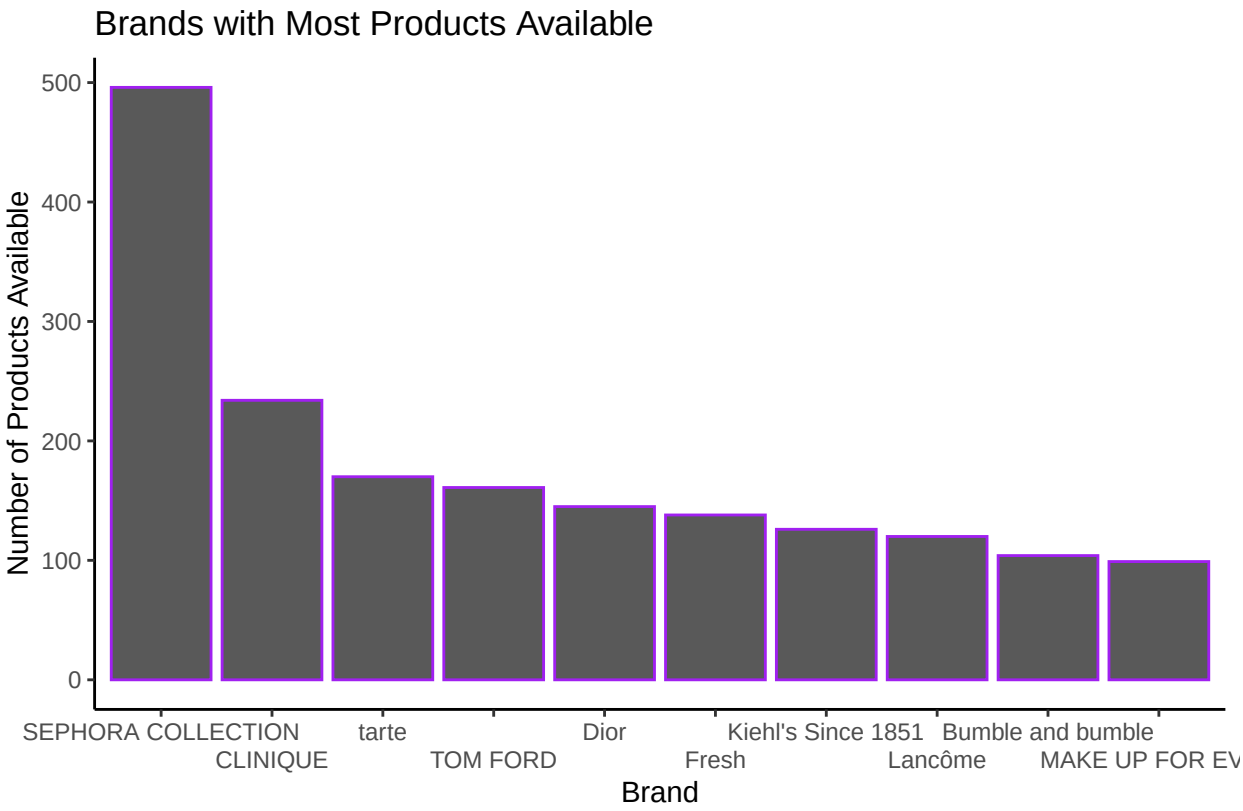
```r
mean(product_df$love) #number of loves of product, basically likes
```

```
## [1] 16278.59
```

```r
# find the brands with the most products
product_df %>%
  count(brand, sort = TRUE) %>%
  head(n = 10)
```

```
## # A tibble: 10 x 2
##    brand                  n
##    <chr>              <int>
##  1 SEPHORA COLLECTION   496
##  2 CLINIQUE             234
##  3 tarte                170
##  4 TOM FORD             161
##  5 Dior                 145
##  6 Fresh                138
##  7 Kiehl's Since 1851   126
##  8 Lancôme              120
##  9 Bumble and bumble    104
## 10 MAKE UP FOR EVER      99
```

```r
# make graph for this
product_df %>%
  count(brand, sort = TRUE) %>%
  head(n = 10) %>%
  ggplot(mapping = aes(reorder(brand, -n), y = n)) +
  geom_col(colour = "purple") +
  scale_x_discrete(guide = guide_axis(n.dodge = 2)) + #to make it easier to read
  theme_classic() +
  labs(
    title = "Brands with Most Products Available",
    x = "Brand",
    y = "Number of Products Available",
    caption = "Source: Sephora US Website"
  )
```



Source: Sephora US Website

```
# let's see which categories have the most products represented
product_df %>%
  count(category, sort = TRUE) %>%
  head(n = 10)
```

```
## # A tibble: 10 x 2
##    category                     n
##    <chr>                    <int>
##  1 Perfume                    665
##  2 Moisturizers               451
##  3 Face Serums                384
##  4 Value & Gift Sets          378
##  5 Face Wash & Cleansers      247
##  6 Face Masks                 230
##  7 Rollerballs & Travel Size  228
##  8 Hair Styling Products      224
##  9 Eye Palettes               202
## 10 Eye Creams & Treatments    191
```

```
# see the most popular (loves) individual products
most_loved <- product_df[order(product_df$love, decreasing = TRUE), ] %>%
  head(n = 15)
kable(most_loved %>%
            select(brand, category, name, love, rating, number_of_reviews, price))
```

brand	category	name	love	rating	number_of_reviews	price
KVD Vegan Beauty	Lipstick	Everlasting Liquid Lipstick	1300000	4.5	14000	21
NARS	Concealer	Radiant Creamy Concealer	770700	4.5	11000	30
Anastasia Beverly Hills	Eyebrow	Brow Wiz	660000	4.5	14000	23
Laura Mercier	Setting Spray & Powder	Translucent Loose Setting Powder	657100	4.5	8000	39
NARS	Blush	Blush	646600	4.5	17000	30
SEPHORA COLLECTION	Lipstick	Cream Lip Stain Liquid Lipstick	628100	4.5	9000	15
FENTY BEAUTY by Rihanna	Foundation	Pro Filt'r Soft Matte Longwear Foundation	625500	4.0	15000	35
HUDA BEAUTY	Eye Palettes	Obsessions Eyeshadow Palette	624600	4.5	4000	27
Anastasia Beverly Hills	Eyeshadow	Eye Shadow Singles	565200	4.5	687	12
FENTY BEAUTY by Rihanna	Lip Gloss	Gloss Bomb Universal Lip Luminizer	553300	4.5	10000	19
Anastasia Beverly Hills	Lipstick	Liquid Lipstick	549000	4.0	4000	20
Urban Decay	Setting Spray & Powder	All Nighter Long-Lasting Makeup Setting Spray	506800	4.5	9000	33
Anastasia Beverly Hills	Eyebrow	DIPBROW™ Pomade	504700	4.5	10000	21
Urban Decay	Lipstick	Vice Lipstick	487100	4.5	2000	19
KVD Vegan Beauty	Eyeliner	Tattoo Eyeliner	485600	4.0	17000	21

```
# relationship between number of reviews and rating ?
cor(product_df$number_of_reviews, product_df$rating)
```

```
## [1] 0.08147766
```

```
# relationship between popularity (loves) and rating?
cor(product_df$love, product_df$rating)
```

```
## [1] 0.09478838
```

```
# relationship between popularity (loves) and number of reviews?
cor(product_df$love, product_df$number_of_reviews)
```

```
## [1] 0.746099
```

```
# this will be useful later when we select the factors for our models, to insure independence
```