Sephora EDA in R

Juliet Maharaj

2022-08-08

library(readr)  
library(stringr)  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library("viridis")

## Loading required package: viridisLite

library(tidyverse)

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ tibble 3.1.8 ✔ purrr 0.3.4  
## ✔ tidyr 1.2.0 ✔ forcats 0.5.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

First, we will import the data that we recieved from Kaggle.

data = read\_csv("/Users/juliet/Dropbox/My Mac (Juliet’s MacBook Air)/Downloads/sephora\_website\_dataset.csv")

## Rows: 9168 Columns: 21  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (10): brand, category, name, size, URL, MarketingFlags\_content, options,...  
## dbl (10): id, rating, number\_of\_reviews, love, price, value\_price, online\_on...  
## lgl (1): MarketingFlags  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Now that we have imported the data, we can see that there are 9168 rows and 21 columns. This data set is a list of the products sold at Sephora and detailed information about the products.

Below, we can see the highest rated brands at Sephora. This rating is a average of all the ratings for that brand.

bran = aggregate(rating ~ brand, data = data, mean) %>% arrange(desc(rating))  
print(head(bran, 50))

## brand rating  
## 1 Aether Beauty 5.000000  
## 2 Fable & Mane 5.000000  
## 3 Four Sigmatic 5.000000  
## 4 Montblanc 5.000000  
## 5 Golde 4.875000  
## 6 RODIN olio lusso 4.812500  
## 7 The Art of Shaving 4.700000  
## 8 Paco Rabanne 4.687500  
## 9 SOBEL SKIN Rx 4.687500  
## 10 dae 4.666667  
## 11 StackedSkincare 4.666667  
## 12 Valentino 4.666667  
## 13 Carolina Herrera 4.625000  
## 14 Better Not Younger 4.607143  
## 15 Proactiv 4.590909  
## 16 Shani Darden Skin Care 4.583333  
## 17 SKYLAR 4.566667  
## 18 alpyn beauty 4.500000  
## 19 BALENCIAGA 4.500000  
## 20 bkr 4.500000  
## 21 Cocofloss 4.500000  
## 22 Crushed Tonic 4.500000  
## 23 Evian 4.500000  
## 24 Flora + Bast 4.500000  
## 25 Google 4.500000  
## 26 High Beauty 4.500000  
## 27 Lunar Beauty 4.500000  
## 28 Makeup Eraser 4.500000  
## 29 MDSolarSciences 4.500000  
## 30 Saint Jane Beauty 4.500000  
## 31 Shaveworks 4.500000  
## 32 The 7 Virtues 4.500000  
## 33 Tower 28 Beauty 4.500000  
## 34 Wishful 4.500000  
## 35 CHANEL 4.471154  
## 36 innisfree 4.470588  
## 37 Violet Voss 4.468750  
## 38 PHLUR 4.466667  
## 39 Saturday Skin 4.466667  
## 40 Anthony 4.464286  
## 41 Edible Beauty 4.458333  
## 42 Comptoir Sud Pacifique 4.437500  
## 43 Alpha-H 4.423077  
## 44 Aquis 4.400000  
## 45 Qhemet Biologics 4.400000  
## 46 Viseart 4.391304  
## 47 Calvin Klein 4.388889  
## 48 Skinfix 4.386364  
## 49 Obagi Clinical 4.375000  
## 50 Cinema Secrets 4.363636

Below, is a table of all the brands sold at Sephora and how many products each of them have available.

brand = as.data.frame(table(data$brand)) %>% arrange(desc(Freq))  
print(head(brand,30))

## Var1 Freq  
## 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  
## 11 Moroccanoil 98  
## 12 Kérastase 96  
## 13 Shiseido 95  
## 14 Benefit Cosmetics 93  
## 15 FENTY BEAUTY by Rihanna 85  
## 16 philosophy 84  
## 17 Too Faced 81  
## 18 Yves Saint Laurent 81  
## 19 Givenchy 76  
## 20 Jo Malone London 76  
## 21 Urban Decay 76  
## 22 Charlotte Tilbury 75  
## 23 Bobbi Brown 74  
## 24 Living Proof 74  
## 25 Origins 73  
## 26 KVD Vegan Beauty 71  
## 27 Anastasia Beverly Hills 70  
## 28 Hourglass 70  
## 29 Peter Thomas Roth 69  
## 30 MILK MAKEUP 67

Because there are so many different brands at Sephora, we will further analyize the top 20 rated brands.

focus = head(brand[,1],20)  
top20 = data %>% filter(brand %in% (focus))  
  
sapply(top20, function(x) sum(is.na(x)))

## id brand category   
## 0 0 0   
## name size rating   
## 0 0 0   
## number\_of\_reviews love price   
## 0 0 0   
## value\_price URL MarketingFlags   
## 0 0 0   
## MarketingFlags\_content options details   
## 0 0 0   
## how\_to\_use ingredients online\_only   
## 0 0 0   
## exclusive limited\_edition limited\_time\_offer   
## 0 0 0

If we combine similar categories in the Sephora data set, we will be able to further our analysis. For example, there is no purpose for have both categories ‘Cologne’ and ‘Perfume’ right now.

category <- as.data.frame(table(top20$category))

Fragrance\_list <- c('Perfume', 'Perfume Gift Sets', 'Cologne Gift Sets', 'Fragrance', 'Body Mist & Hair Mist', 'Body Sprays & Deodorant', 'Deodorant & Antiperspirant', 'Deodorant for Men', 'Diffusers', 'Cologne', 'Candles', 'Candles & Home Scents', 'Rollerballs & Travel Size')  
  
top20$category[top20$category %in% Fragrance\_list] <- 'Fragrance'  
  
Cleansers\_list <- c('Face Wash & Cleansers', 'Face Wash', 'Makeup Removers', 'Toners', 'Face Wipes')  
top20$category[top20$category %in% Cleansers\_list] <- 'Cleansers'  
  
Sunscreen <- c('Face Sunscreen', 'Lip Sunscreen', 'Body Sunscreen')  
top20$category[top20$category %in% Sunscreen] <- 'Sunscreen'  
  
Masks <- c('Face Masks', 'Sheet Masks', 'Facial Peels', 'Cellulite & Stretch Marks', 'Exfoliators', 'Eye Masks')  
  
top20$category[top20$category %in% Masks] <- 'Masks and Exfoliators'  
  
Moisturizers <- c('Moisturizers', 'Night Creams', 'Face Oils', 'Mists & Essences', 'Face Serums', 'Eye Creams & Treatments', 'Lotions & Oils','Skincare', 'Lip Balm & Treatment', 'Lip Balms & Treatments', 'Lip Balms & Treatments', 'Moisturizer & Treatments', 'Blemish & Acne Treatments', 'Face Serums', 'Tinted Moisturizer', 'Eye Cream', 'After Sun Care', 'Anti-Aging')  
  
top20$category[top20$category %in% Moisturizers]<- 'Moisturizers and Treatments'  
  
Makeup <- c('Mascara', 'Lipstick', 'Foundation', ' Eyeliner', 'Highlighter', 'Eye Palettes', 'Eyebrow', 'Makeup', 'Face Primer', 'Setting Spray & Powder', 'Concealer','Bronzer', 'Lip Gloss', 'Blush','Eyeshadow', 'Makeup Palettes', 'Makeup & Travel Cases', 'Lip Stain', 'Makeup Bags & Travel Cases', 'Eye Primer', 'Lip Liner',' Cheek Palettes', 'BB & CC Cream', 'Cheek Palettes', 'Liquid Lipstick', 'Color Correct', 'Contour', 'BB & CC Creams', 'Lip Plumper', 'Eyeliner')  
  
top20$category[top20$category %in% Makeup] <- 'Makeup'  
  
Tools <- c('Face Brushes', 'Eye Brushes', 'Sponges & Applicators', 'False Eyelashes', 'Brush Sets', 'Brush Cleaners', 'Facial Rollers', 'Eyelash Curlers', 'Facial Cleansing Brushes', 'Blotting Papers', 'Tweezers & Eyebrow Tools',' Lip Brushes', 'Curling Irons', 'Lid Shadow Brush', 'Powder Brush', 'Spa Tools', 'Hair Brushes & Combs', 'Hair Accessories', 'Hair Dryers', 'Hair Straighteners & Flat Irons', 'Lip Brushes', 'Mirrors & Sharpeners', 'Accessories', 'Teeth Whitening')  
  
top20$category[top20$category %in% Tools] <- 'Tools and Brushes'  
  
Hair <- c('Hair Styling Products', 'Shampoo', 'Conditioner', 'Hair', 'Hair Masks', 'Hair Primers', 'Scalp & Hair Treatments', 'Dry Shampoo', 'Hair Oil', 'Hair Spray', 'Leave-In Conditioner', 'Hair Products', 'Color Care',' Shampoo & Conditioner', 'Shampoo & Conditioner', 'Diffusers' )  
  
top20$category[top20$category %in% Hair] <- 'Hair Products'  
  
Bath <- c('Body Lotions & Body Oils', 'Body Wash & Shower Gel', 'Bath & Shower', 'Hand Cream & Foot Cream', 'Bath Soaks & Bubble Bath', 'Body Products', 'For Body', 'Bath & Body', 'Body Moisturizers', 'Scrub & Exfoliants', 'Decollete & Neck Creams', 'Self Tanners', 'Shaving', 'Aftershave')  
  
top20$category[top20$category %in% Bath] <- 'Bath&Body'  
  
other\_list <- c('Mini Size', 'Beauty Supplements', 'no category', 'Nail', 'Value & Gift Sets', 'Face Sets', 'Eye Sets', 'Lip Sets', 'For Face', 'Skincare Sets', 'Hair Removal')  
  
top20$category[top20$category %in% other\_list]<- 'Others'

as.data.frame(table(top20$category)) %>%  
 rename(category = Var1) %>%  
 arrange(desc(Freq))

## category Freq  
## 1 Makeup 770  
## 2 Moisturizers and Treatments 415  
## 3 Fragrance 304  
## 4 Hair Products 286  
## 5 Tools and Brushes 278  
## 6 Others 220  
## 7 Cleansers 131  
## 8 Bath&Body 120  
## 9 Masks and Exfoliators 107  
## 10 Sunscreen 27

Now that we have combined common categories, further analysis can be done.