What Fcators May Affect A Product's Number Of Love Received? Are These Differed By Different Products?

Ziyi Bai

11/23/2020

1. Abstract

My project used the dataset from kaggle to explore which factors may affect how many love a Sephora online product receives. My original datatset includes more than 9,000 products and more than hundreds type of products. I picked 10 specific type of products to simplify my analysis. Then, I did exploratory data analysis and build several models to pick the best one.

2. Introduction

I am a huge fun of beauty products. "We need to think about beauty even if we were too busy thinking about COVID-19." In my future career, I really want to be able to work in a cosmetic company and promote their products to potential customers with the help of data analysis.

Sephora is a world famous cosmetic company which has its customers around the world. I want to analyze what factors may influence a product's love received and whether these factors different from different type of products. I have 143 different kinds of products in my original dataset and I randomly choose 10 of them for further analysis.

To begin with, I want to give a brief introduction of what each variable means: Love is the outcome of my model, which means how many love a product receive from customers. Category column has 10 different types of products: perfume, lipsticks, shampoo, anti-aging product, face mask, eye brushes, eye shadow, hair oil, shaving and candle. Number of reviews means how many people leave their review toward a product. Price means what is a product's current price after discount. Online_only, exclusive, limited_edition and limited_time_offer are binary variables which coded into 1 if a product has such promoting method (coded into 0 if not). Then, I added a new column into the dataset named sum to calculate how many promotion ways a product has. In the following report, I tried to explore the relationship between these variables with number of love a product receive.

In the Method part, I did some EDA about my data to get an initial idea about all variables' relationship. Also, I tried to fit models to explore more specific relationships between each factors. Then, I used internal methods to test the validation of my model.

3. Method

Step1: Initial EDA

At the very beginning, I checked the inter-correlation between variables in my dataset using corrgram package. As the following graph shows, the number_of_reviews variable and love variable shows a clear correlation relationship, so I dropped number of reviews variable in the following model fit.

Step2: Further EDA

I did EDA between love and each of factors to explore the relationship of love and intended influencer.

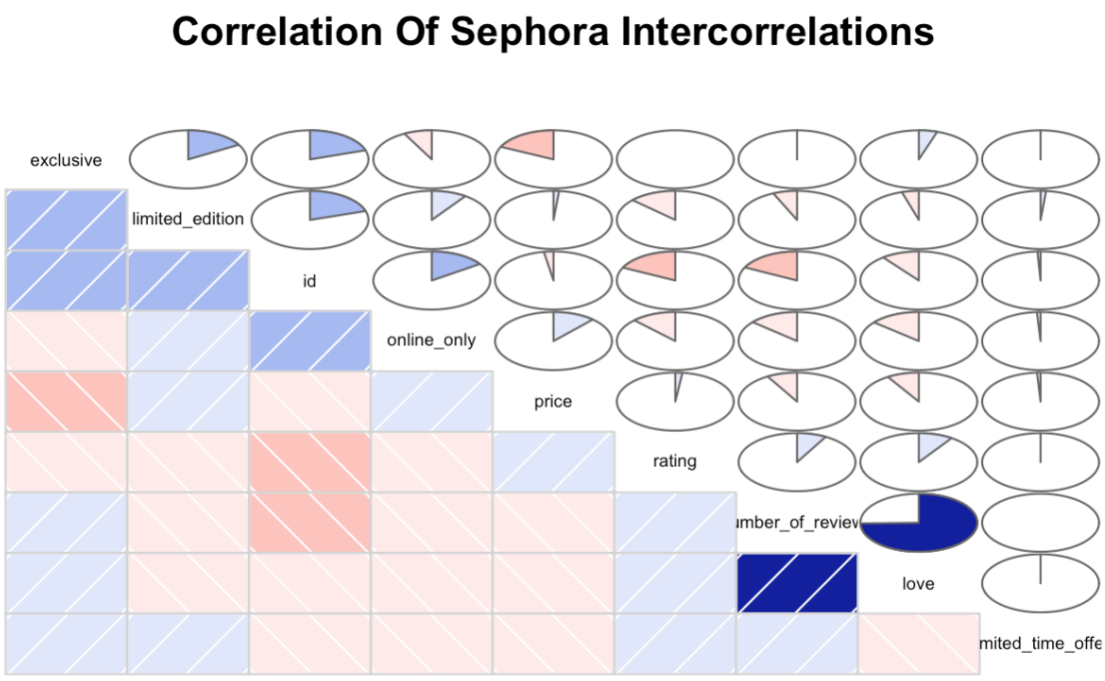

Figure 1: (Correlation Of Variables' Intercorrelations)

First of all, I want to check the distribution of love. I calculated the average love of each type of product and reordered the bar by the value of average love. As we can see, lipstick receives the highest average love from customers which higher than 60000. The following two categories are eye shadow and face mask.

Secondly, I am interested in how the number of promotions way's affects love. Initially, I supposed the love increases as the promotion ways increase. However, the graph shows a different story. Mostly, the average value of love shows a clear trend of decreasing as the promotion ways increase, only Eye brushes is an exception.

Thirdly, I draw a graph to show how price affects the number of love a product received. For shampoo, perfume, face mask, hair oil, eye shadow and lipstick, we can see that as price increasing, the love has a decreasing trend.

Last but not the least, I make graph about how the number of love changes as the product's rating increasing based on different category. Perfume, eye brushes, face masks, eye shadow and lipstick categories have clear pattern that love increases as the rating of a product increasing.

Step 3: Modeling

Draw a scatter plot of each variables with love is a good way to indicate their potential relationship. Love is dependent variable in my model, price, sum, rating and category are independent variables.

I fitted 4 models for my project. The first one is poisson model, because it has the problem of over-dispersion, so I refitted quasipoisson model. Also, I fit neg_binomial_2 and lmer model to compare the outcomes. Then, I draw each model's binned residual plot to compare internal validation. For my personal point of view, I think the left bottom one's binned residual plot is the best which is the third model. Basically all dots are included in the line and they are centered around the average residual.

Ultimately, I choose the third model, neg_binomial_2 model. In my model, anti-aging product is automatically been chosen as dummy variable:

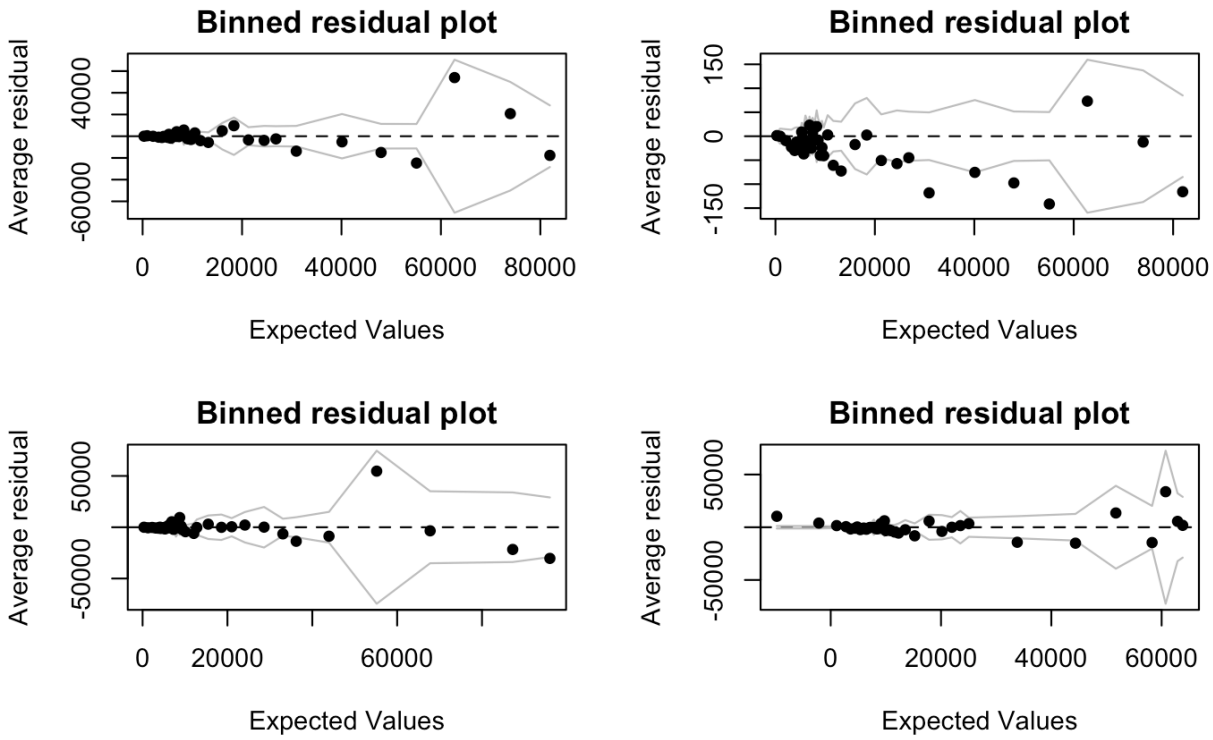

Figure 2: (Binned Residual Plot)

fit3 <- stan_glm(love~price+sum+rating+category,family=neg_binomial_2(link = "log"),data=love,refresh=0)

In the next part, I will explain this model in detail and calculate the confidence interval for further analysis.

4. Result

Firstly, the interpretation of my coefficients as following:

Intercept: When an anti-aging product has average price, rating, and don't take any promotion ways, it's predicted receive 7.184 number of love. Price: A product tends to receive -0.002 less love as it's price increases one dollar, keeping everything else constant. Sum: A product tends to receive -0.507 less love as it has one more promotion way, keeping everything else constant. Rating: A product tends to receive 0.573 more love as it's rating increases one unit, keeping everything else constant. categoryCandle: Candle products tend to have -1.955 less love than anti-aging products, keeping everything else constant. categoryEye Brushes: Eye brushes products tend to have -0.29 less love than anti-aging products, keeping everything else constant. categoryEyeshadow: Candle products tend to have 1.351 more love than anti-aging products, keeping everything else constant. categoryFace Masks: Face Masks products tend to have 0.71 more love than anti-aging products, keeping everything else constant. categoryHair Oil: Hair Oil products tend to have 0.018 more love than anti-aging products, keeping everything else constant. categoryLipstick: Lipstick products tend to have 1.721 more love than anti-aging products, keeping everything else constant. categoryPerfume: Perfume products tend to have -0.523 less love than anti-aging products, keeping everything else constant. categoryShampoo: Shampoo products tend to have -0.6 less love than anti-aging products, keeping everything else constant. categoryShaving: Shaving products tend to have -2.099 less love than anti-aging products, keeping everything else constant.

Secondly, I manually checked the 95% confidence interval of my model, which is from 14677 to 20136, which means 95% products' number of love will fall between 14677 and 20136. Also, I checked each variable's

confidence interval using `confint` function at 95% of confidence interval. You can take a look of the result in the appendix.

5. Discussion

The result of my model accords with the my EDA, which product's price and number of promotion ways has negative effect on number of love received, while product's rating has positive effect on number of love received.

Although I fit 4 models, none of them fit perfectly well. I tried to take log(number_of_reviews+1) as an offset in my model since there are lots of 0 in this column, however, the binned residual plot doesn't fit very well in this way.

In the next step, I think work with the whole dataset may tells me more detailed story. To make my model fit better, I think include customers' demographic information is also a good choice.

Overall, basically what I have found by EDA and modeling accords with my expectation. Maybe work with bigger dataset will have a more interesting outcome. If I am asked about how to promote a product, focus on better discount offer maybe a better way.

Appendix

(You may click on the character to jump to the website)

1. Sephora dataset
2. The whole process of my work in Rmd on github
3. My github repository
4. Some EDA figures that help me build my model:

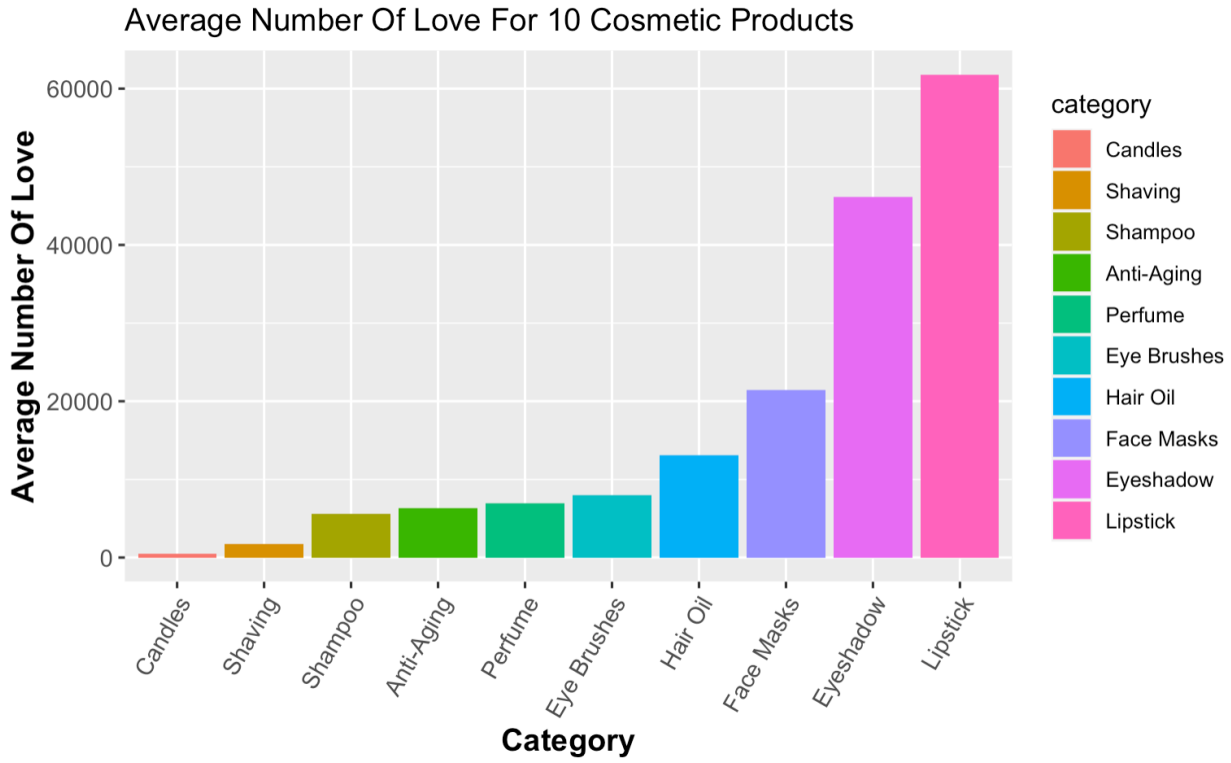

Figure 3: (Average Number Of Love For 10 Cosmetic Products)

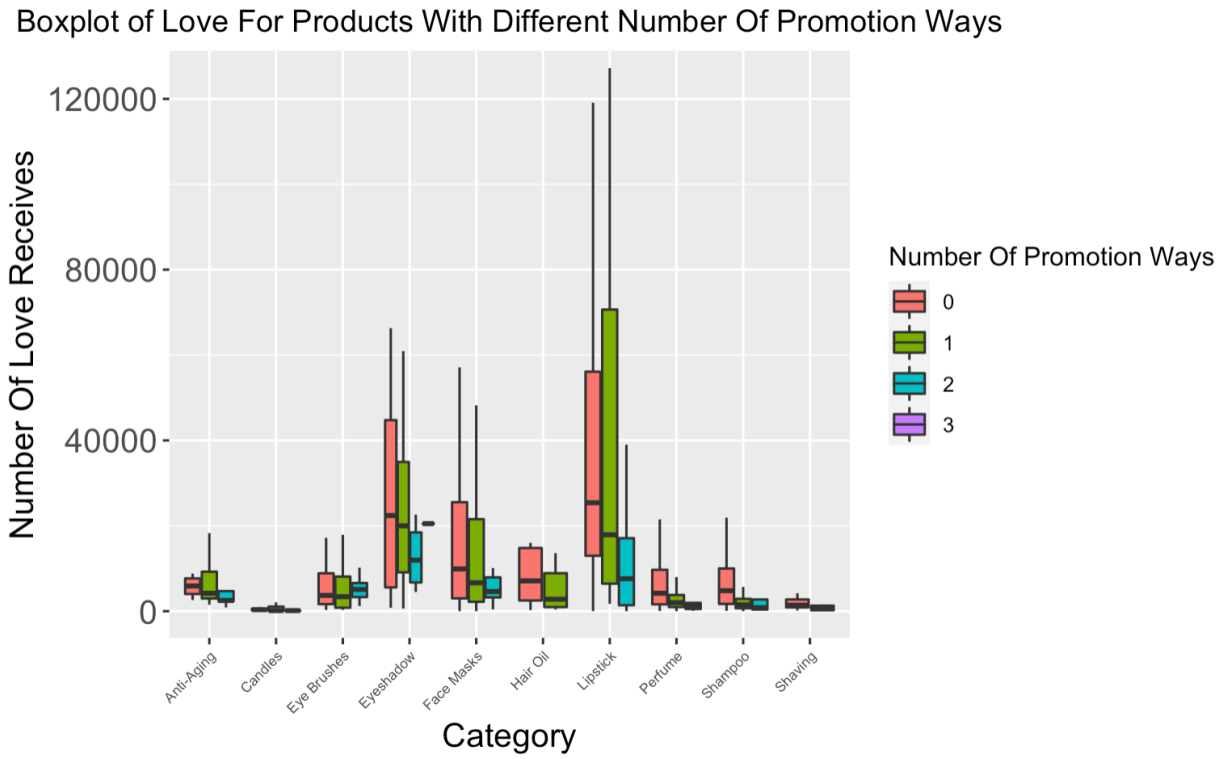

Figure 4: (Boxplot of Love For Products With Different Number Of Promotion Ways)

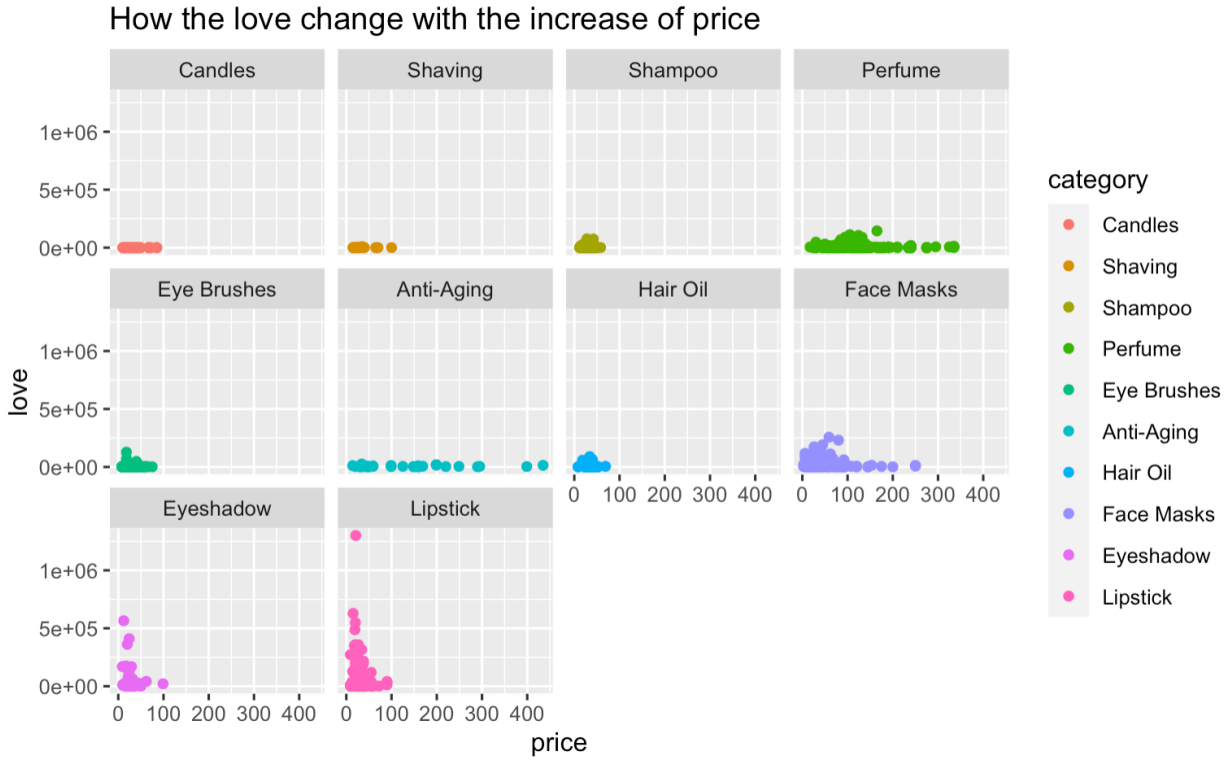

Figure 5: (How the love change with the increase of price)

Figure 6: (How the love change with the increase of rating)

```
 function:     stan_glm
 family:       neg_binomial_2 [log]
 formula:      love ~ price + sum + rating + category
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 1590
 predictors:   13

Estimates:
                        mean   sd     10%    50%    90%
(Intercept)             7.184  0.315  6.774  7.183  7.578
price                  -0.002  0.001 -0.003 -0.002 -0.001
sum                    -0.507  0.054 -0.575 -0.507 -0.438
rating                  0.573  0.045  0.517  0.573  0.629
categoryCandles        -1.955  0.332 -2.377 -1.948 -1.537
categoryEye Brushes    -0.290  0.258 -0.631 -0.284  0.035
categoryEyeshadow       1.351  0.263  1.017  1.355  1.687
categoryFace Masks      0.710  0.237  0.403  0.713  1.013
categoryHair Oil        0.018  0.296 -0.368  0.021  0.390
categoryLipstick        1.721  0.245  1.404  1.726  2.032
categoryPerfume        -0.523  0.213 -0.803 -0.522 -0.254
categoryShampoo        -0.600  0.245 -0.923 -0.594 -0.282
categoryShaving        -2.099  0.339 -2.536 -2.096 -1.673
reciprocal_dispersion   0.689  0.021  0.662  0.688  0.715
```

Figure 7: (Fit Neg_binomial_2 Model)

```
Waiting for profiling to be done...
                              2.5 %       97.5 %
(Intercept)            5.669230934  8.517745818
price                 -0.008438609 -0.001398926
sum                   -0.417997899 -0.099875846
rating                 0.361039044  0.804991719
categoryCandles      -11.656224836 -0.178359147
categoryEye Brushes   -1.542216608  0.845611720
categoryEyeshadow      0.465045979  2.638546944
categoryFace Masks    -0.246497958  1.888442110
categoryHair Oil      -1.074107614  1.407317574
categoryLipstick       0.727934273  2.857376828
categoryPerfume       -1.156324128  0.949592529
categoryShampoo       -1.572160910  0.747821470
categoryShaving       -5.962794899 -0.042733442
```

Figure 8: (CI Of Each Variable)