What Fcators May Affect A Product's Number Of Love Received? Are These Differed In 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 different kinds of Sephora online products receive. I investigated the potential relationship of predictor's effect on love by doing EDA at first. Then I fitted 4 models: poisson, quasipoisson, neg_binomial_2 and glmer to find out which fitted my data best. After draw the binned residual plot, I found glmer mixed effect model fits best. Sum and price variables have negative effect on love and rating has positive effect on love. I mainly focused on 10 different kinds of product through my dataset, since my model is significant at 95% confidence level, so that other consumer's activity tend to have influences on the potential consumers'. When a company is trying to promote a product in a positive way, it is worth for them to use consumer's reaction, like rating and reviews to simulate potential consumption!

2. Introduction

I am a huge fun for 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.

Sephora is a world famous cosmetic company which has its customers around the world. When you shopping at the Sephora website and viewing at one product, you are able to see how many people click the like button of the product. In marketing process, psychological factors do have influences on consumer behavior such as demographics, personality, lifestyles, and behavioral variables. I assume consumers make consumption decisions based on several factors, like how many love a product receive, product price, promotion ways etc. One thing that I am interested in for this project is what factors may influence a product's love received and whether these factors differed in different type of products.

I used the dataset form kaggle gathered by Raghad Alharbi. This dataset includes 143 different kinds of products, more than 9,000 variables and 21 columns. For the time sake and clearity of my EDA and modeling, I randomly choose 10 types of products for further analysis.

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

Setp1: Data Cleaning

Firstly, 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 receives 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 used 'ggplot2', 'rstanarm', 'tidyverse', 'MASS"... packages to explore the relationship between these variables with number of love a product receive.

Step2: Initial EDA

At the very beginning, I checked the inter-correlation between variables in my dataset using 'corrgram' package. As the following graph shows that 4 pairs of variables: number_of_reviews and love, limited_edition and sum, online_only and sum, limited_edition and online_only show a clear correlation relationship. So, price, sum, love, category and rating will be the predictors to see the outcome of love.

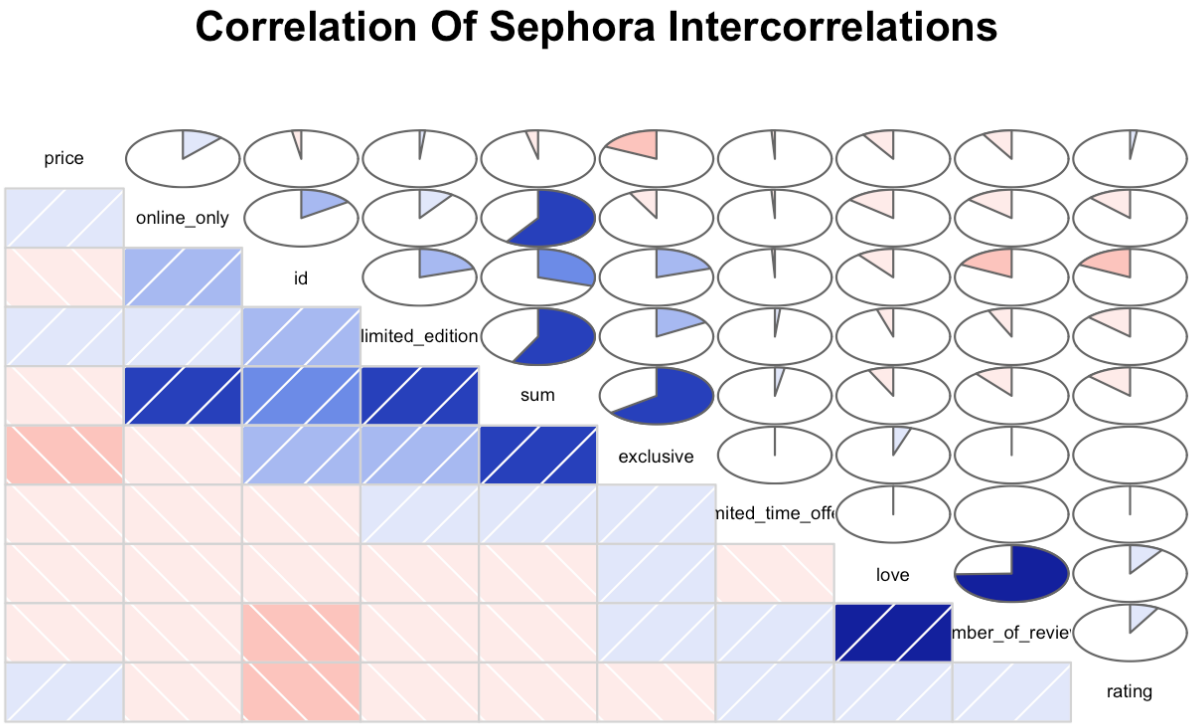

Figure 1: (Correlation Of Variables' Intercorrelations)

Step3: Further EDA

I did further EDA between love and each of factors to explore the potential relationship among them. (See the graphs in Appendix)

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. 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 on love. Initially, I supposed the love increases as the promotion ways increasing. However, the graph shows a different story. Mostly, the average value of love shows a clear trend of decrease as the promotion ways increasing, 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, as price increasing, the love has a decrease trend.

Last but not the least, I made 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.

Step4: Modeling

Draw a scatter plot of each variables with love is a good way to indicate their potential relationship. At the beginning, which model suits for my data is not clear. So I tried as much models as I can. I fitted 4 models and compared them to find the best one.

My variables have a L-shape scatter plot, which is not informative for further analysis, so that take 'log' maybe a good choice for me. So, firstly I fitted a poisson model.

```
fit1 <- stan_glm(love~price+sum+rating+category,family=poisson(link = "log"),data=love,refresh=0)
```

Then I checked the problem of over-dispersion using 'qcc.overdispersion.test',which is the problem that exists in my model, so I refitted a quasipoisson model.

```
fit2 <- glm(love~price+sum+rating+category,family=quasipoisson(link = "log"),data=love)
```

I also fitted a neg_binomial_2 to generalize the poisson model to allow for over-dispersion.

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

Because each product is unique, I also include product's id in the fourth model. Each product and category has it's significant intercept:

```
fit4 <- glmer(love~price+sum+rating+(1|category)+(1|id),family=poisson(link="log"),data=love)
```

To compare the models I fitted, I draw each model's binned residual plot to check their internal validation. For my personal point of view, I think the right bottom one's binned residual plot is the best which all plots centered around 0 and within the lines.

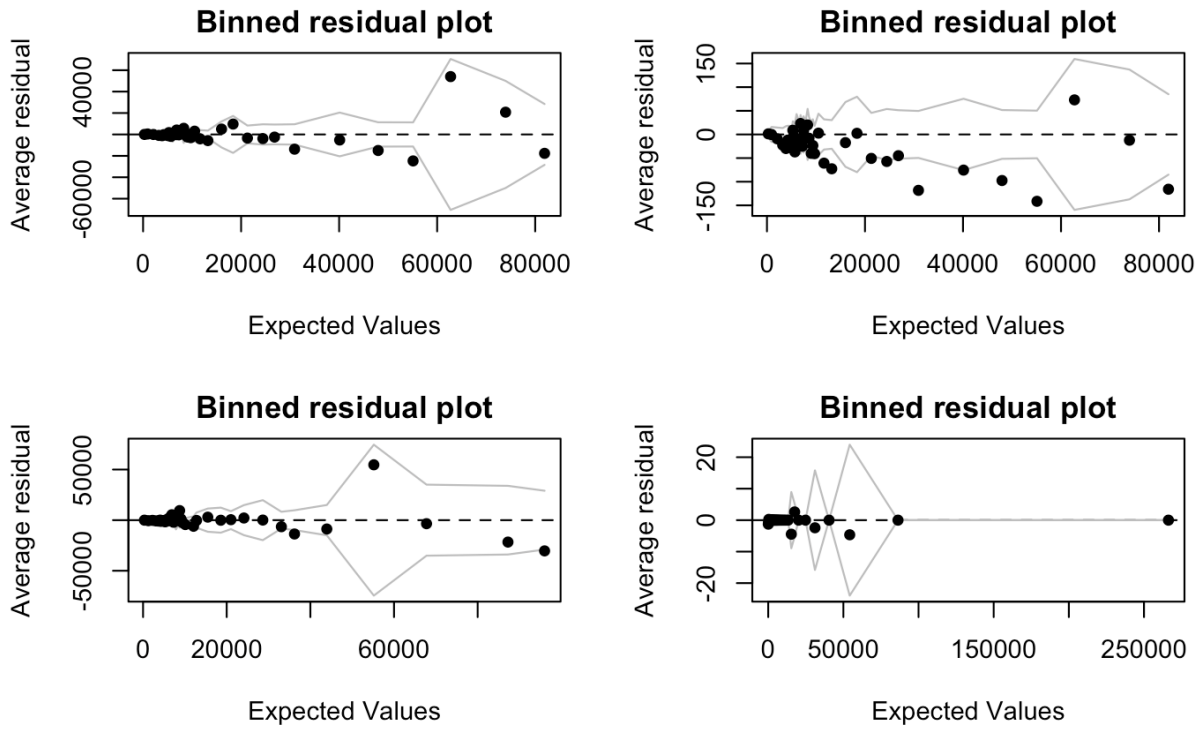

Figure 2: (Binned Residual Plot For 4 Models)

Ultimately, I choose the last model, glmer mixed effect model for further analysis:

```
fit4 <- glmer(love~price+sum+rating+(1|category)+(1|id),family=poisson(link="log"),data=love)
```

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

4. Result

Intercept, sum and rating is significant at 95% significant level in my model. Only the p-value of price over 0.05 which is 0.124. So, price might don't have significant effect on how many love a product can reveive.

The interpretations of my coefficients as following:

Intercept: When a product has average promotion ways, price and rating, it receives average love number of 0.999. Price: When a product increases it's price by one unit, it will receive 50.03% less love. Sum: When a product increases it's promotion ways by one unit, it will receive 63.68% less love. Rating: When a product increases it's rating by one unit, it will receive 52.22% more love.

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 figure9.

5. Discussion

What I have found alines with my expectation and EDA. Product's price and number of promotion ways have negative effect on number of love received, while product's rating has positive effect on number of love received. So, when trying to promote an online product, increases it's online rating might be a good way. However, there are still other factors that maybe helpful, like I can explore the relationship between discount rate and number of love received or I can try to include the factor of number of reviews in my model which I am failed to do in this project.

I tried to include log(number_of_reviews+1) into my model because there are some products don't have and reviews. However, the binned residual plot of this model fitted not well. So, I didn't include this model in the main part.

To make my model fit better, I think try to include each products' demographic information is also a good choice. I have ingredients information in the dataset. Maybe I can explore how the type of most frequently used ingredient's effect on the number of love received. Also, for my point of view, we can include the sales volume of each product in the data to see whether other consumers action will influence potential consumers'. In the next step, based on the EDA I already have, I may still have some outliers in the dataset, so I probably would like to clean those outliers then try to fit my model with the whole dataset, which probably tells me more detailed relationship among variables.

Biliography

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software, 67(1), 1–48. doi: 10.18637/jss.v067.i01. Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686. Kaggle.com. 2020. Sephora Website. [online]Available at: https://www.kaggle.com/raghadalharbi/all-products-available-on-sephora-website?select=sephora_website_dataset.csv

Appendix

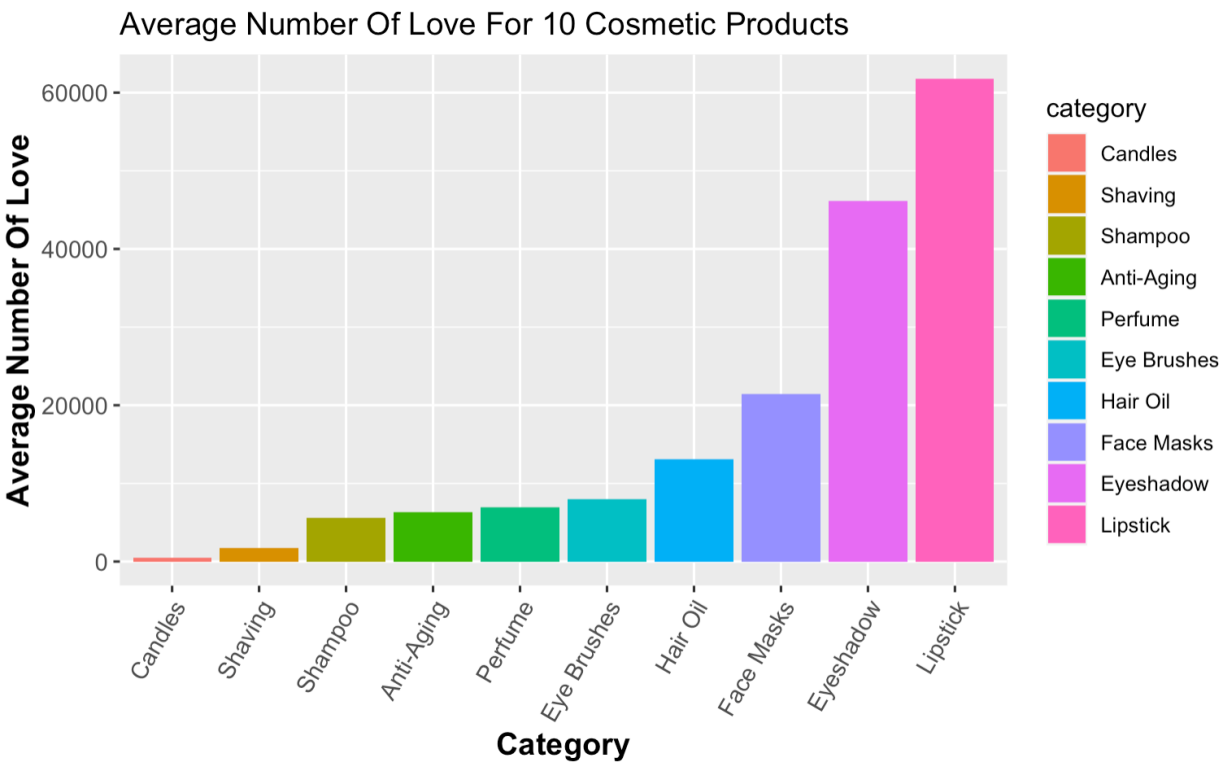

Figure 3: (Average Number Of Love For 10 Cosmetic Products)

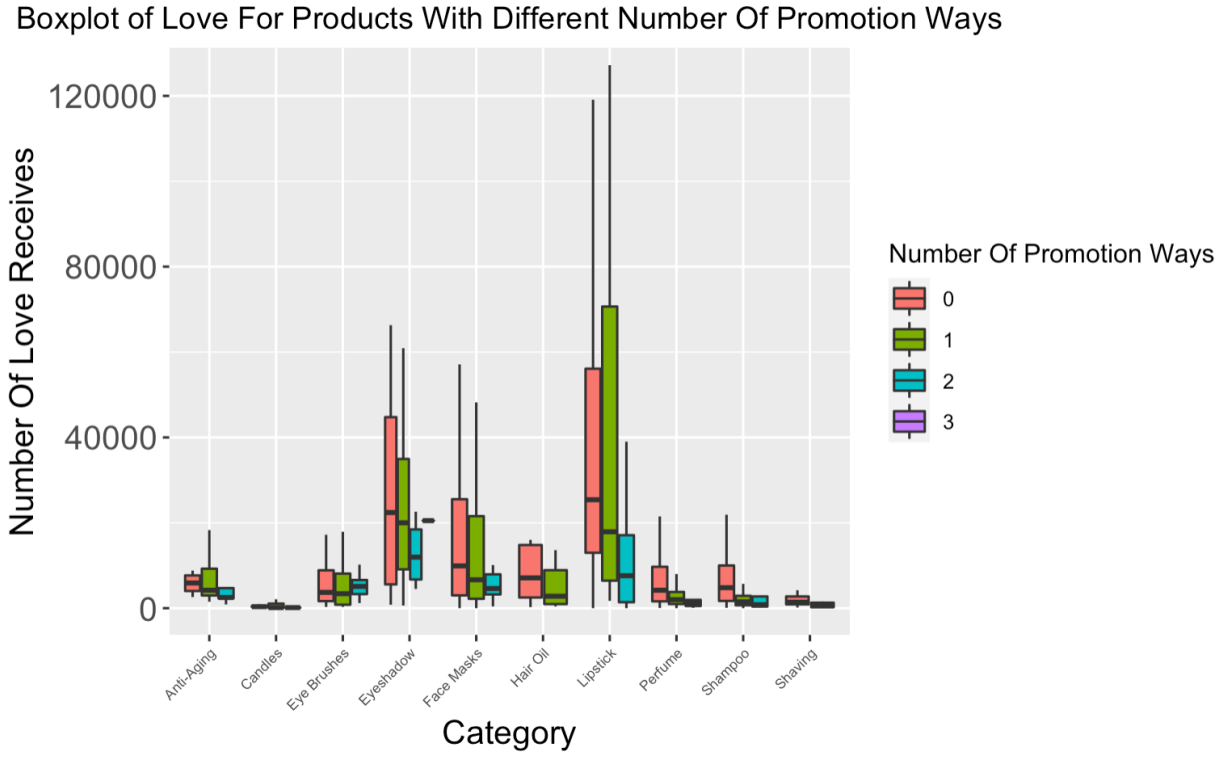

Figure 4: (Product's Love decreases as promotion Ways increasing)

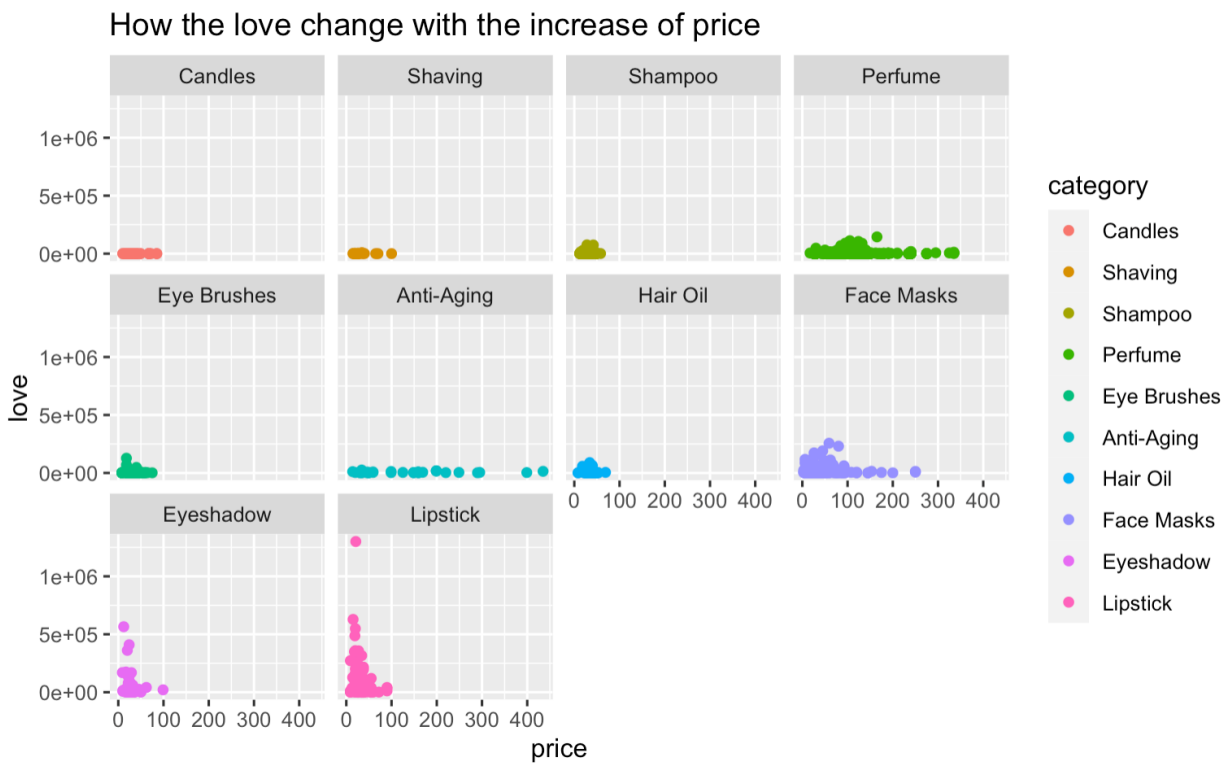

Figure 5: (Love increases as the price increasing)

Figure 6: (Love increases as the rating increasing)

```
Formula: love ~ price + sum + rating + (1 | category) + (1 | id)
   Data: love

      AIC       BIC    logLik  deviance  df.resid
 410447.2  410479.4 -205217.6  410435.2      1584

Scaled residuals:
     Min       1Q   Median       3Q      Max
-241.970   -0.008    0.001    0.004  241.974

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 2.28     1.510
 category (Intercept) 1.41     1.188
Number of obs: 1590, groups:  id, 1583; category, 10

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.1921193  0.3872819  21.153   <2e-16 ***
price       -0.0013805  0.0008979  -1.537    0.124
sum         -0.5613717  0.0652026  -8.610   <2e-16 ***
rating       0.0886627  0.0070172  12.635   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: (Fit glmer model)

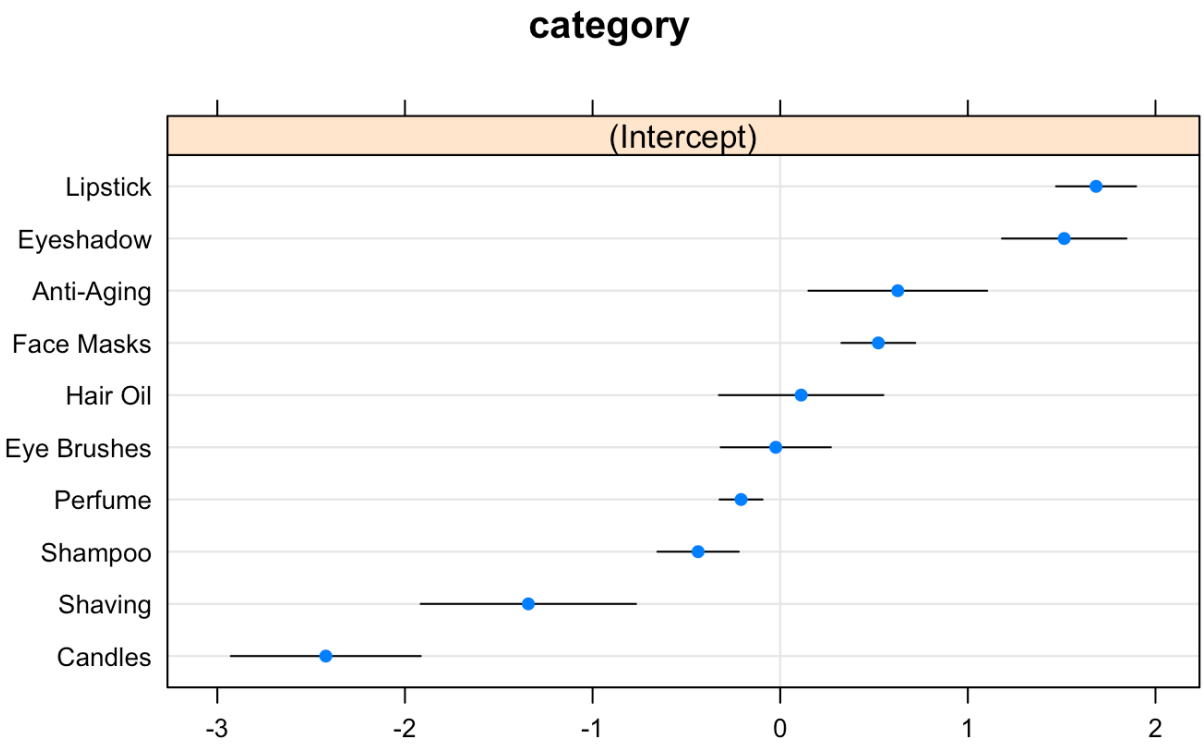

Figure 8: (dotplot of category)

```
Computing profile confidence intervals ...
                    2.5 %       97.5 %
.sig01         1.456816615  1.5665962047
.sig02         0.792945836  2.0064256112
(Intercept)    7.355302446  9.0216148912
price         -0.003140881  0.0003808695
sum           -0.689275409 -0.4335151499
rating         0.074918492  0.1024256054
```

Figure 9: (CI Of Each Variable)

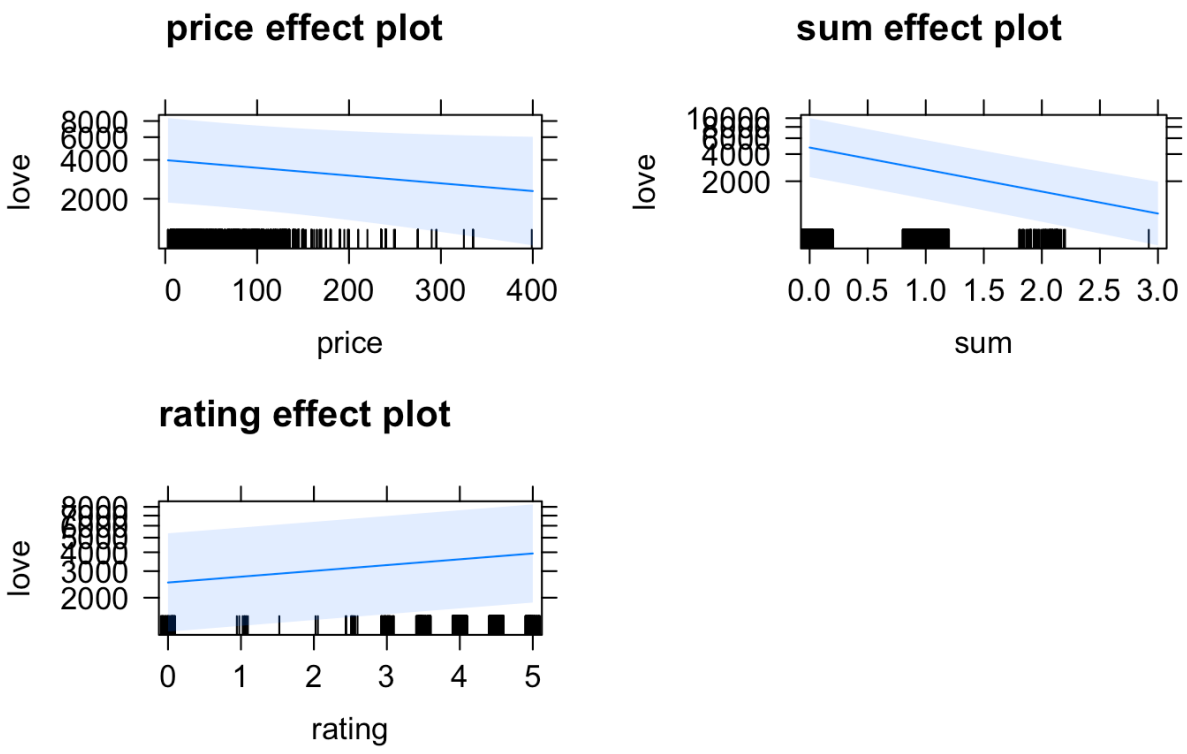

Figure 10: (Effects Of Each Variable)