Report of MA678 Midterm Project

Hui Xiong

12/11/2022

Abstract

This is an analysis report about Sephora's various cosmetic brands, categories, prices, love and rating. Sephora is a makeup brand. It collects various brands of skin care products, cosmetics, perfumes, etc. The brand classifies the leading group of the report through the analysis of price, rating, and love degree. The analysis results show that the degree of preference will affect the price, and the rating will also affect the price. Finally, a relationship between the price of different brands and love, with a rating of 5.0, is established. The article is mainly divided into 3 parts: introduction, data analysis and results.

1.Introduction

Usually, when we choose to buy perfume, we will consider the brand, price, rating, and love. So the variables analyzed in my thesis are brand, price, rating, and love. These parameters are critical and meaningful. For example, the price of a well-known brand will be relatively higher than that of ordinary brands. But a regular brand may also be loved by everyone because of its relatively low price. The rating will be more objective and consider the price, brand, and feeling after use to conclude. Whether there is a correlation between rating and price is what I have considered. So I want to use a multi-level model to further analyze the relationship between price and love and compare the relationship between price and rating.

1.1 Background:

Analysis whether the price of Sephora perfume is affected by brand, love and rating.

1.2 Data Preprocessing

I downloaded it through the kaggle website(https://www.kaggle.com/datasets/raghadalharbi/all-products-available-on-sephora-website).

1.3 Data combining and cleaning:

First, I downloaded the entire dataset. The dataset as a whole is extensive. I filtered by category and selected Fragrance, Perfume, and Perfume Gift Sets as the object data. Then I did a variable screening and selected brand, love, price, and rating. Finally, the cleaned data set is obtained.

2. Data analysis

Use the lmer function

```
lmer(formula = price ~ 1 + (1 | brand), data = data_need)
coef.est  coef.se
  101.18    10.33

Error terms:
 Groups   Name        Std.Dev.
 brand    (Intercept) 35.15
 Residual             58.35
---
number of obs: 226, groups: brand, 16
AIC = 2502.7, DIC = 2509.7
deviance = 2503.2
```

Using lmer() for variable intercept model predictions shows that the Std.Dev is 35.15. This model only includes the constant price term (predictor "1")

```
lmer(formula = price ~ rating + (1 | brand), data = data_need)
            coef.est coef.se
(Intercept) 104.65    17.44
rating       -0.85     3.44

Error terms:
 Groups   Name        Std.Dev.
 brand    (Intercept) 35.13
 Residual             58.48
---
number of obs: 226, groups: brand, 16
AIC = 2500.4, DIC = 2513.9
deviance = 2503.2
```

This result shows inferences about the intercept and slope for y=price, and x=rating, when grouped by the brand. The estimated change obtained through the model: ^σα=35.13, ^σy=58.48. This mock up works for 226 perfume products from 16 brands.

```
lmer(formula = price ~ love + (1 | brand), data = data_need)
            coef.est coef.se
(Intercept) 98.36    10.59
love         0.00     0.00

Error terms:
 Groups   Name        Std.Dev.
 brand    (Intercept) 35.66
 Residual             58.13
---
number of obs: 226, groups: brand, 16
AIC = 2517, DIC = 2492.9
deviance = 2500.9
```

Inferences about y=price, x=love, intercept and slope when grouped by the brand are shown through the model results. The estimated change obtained through the model: ^σα=35.66, ^σy=58.13. This mock up works for 226 perfume products from 16 brands.

Use estimated regression coefficients

```
$brand
                   (Intercept)     rating
Acqua Di Parma       149.65019 -0.8464313
AERIN                131.81324 -0.8464313
HUDA BEAUTY          111.17323 -0.8464313
KVD Vegan Beauty      82.83212 -0.8464313
Lancôme               99.13528 -0.8464313
Maison Louis Marie    76.81308 -0.8464313
Maison Margiela       97.57388 -0.8464313
The 7 Virtues         83.80980 -0.8464313
TOCCA                 66.74306 -0.8464313
TokyoMilk             88.93269 -0.8464313
TOM FORD             180.11439 -0.8464313
Tory Burch            96.17714 -0.8464313
Valentino            114.53903 -0.8464313
Versace               76.73639 -0.8464313
Viktor&Rolf          122.42425 -0.8464313
Yves Saint Laurent    95.97709 -0.8464313

attr(,"class")
[1] "coef.mer"
```

The variable here is price ~ rating. By estimating the model results, it can be concluded that the estimated regression line of the Acqua Di Parma brand is y=149.65019-0.8464313x. The estimated regression line for the AERIN brand is y=131.81324-0.8464313x. By analogy, their intercepts are different, and the slopes are all the same at 0.8464313. This is because The specification (1|brand) tells the model only to allow the intercept to vary.

```
$brand
                   (Intercept)         love
Acqua Di Parma       146.08320 0.0004207388
AERIN                126.80020 0.0004207388
HUDA BEAUTY          105.66202 0.0004207388
KVD Vegan Beauty      75.97696 0.0004207388
Lancôme               91.75614 0.0004207388
Maison Louis Marie    72.17959 0.0004207388
Maison Margiela       90.17240 0.0004207388
The 7 Virtues         78.13458 0.0004207388
TOCCA                 60.89415 0.0004207388
TokyoMilk             82.68183 0.0004207388
TOM FORD             174.87218 0.0004207388
Tory Burch            89.88439 0.0004207388
Valentino            108.39821 0.0004207388
Versace               68.14142 0.0004207388
Viktor&Rolf          114.69413 0.0004207388
Yves Saint Laurent    87.45504 0.0004207388

attr(,"class")
[1] "coef.mer"
```

The variable here is price ~ love. By estimating the model results, it can be concluded that the estimated regression line of the Acqua Di Parma brand is y=146.0832+0.0004207388x. The estimated regression line for the AERIN brand is y=126.8002+0.0004207388x. By analogy, their intercepts are different, and the slopes are all the same at 0.0004207388. This is because The specification (1|brand) tells the model only to allow the intercept to vary.

```
(Intercept)      rating
104.6528040  -0.8464313
```

The estimated regression line of the average brand obtained by Fixed and random effects is y=104.6528040-0.8464313x.

```
 (Intercept)         love
9.836165e+01 4.207388e-04
```

The estimated regression line of the average brand obtained by Fixed and random effects is y=9.836165e+01 + 4.207388e-04x.

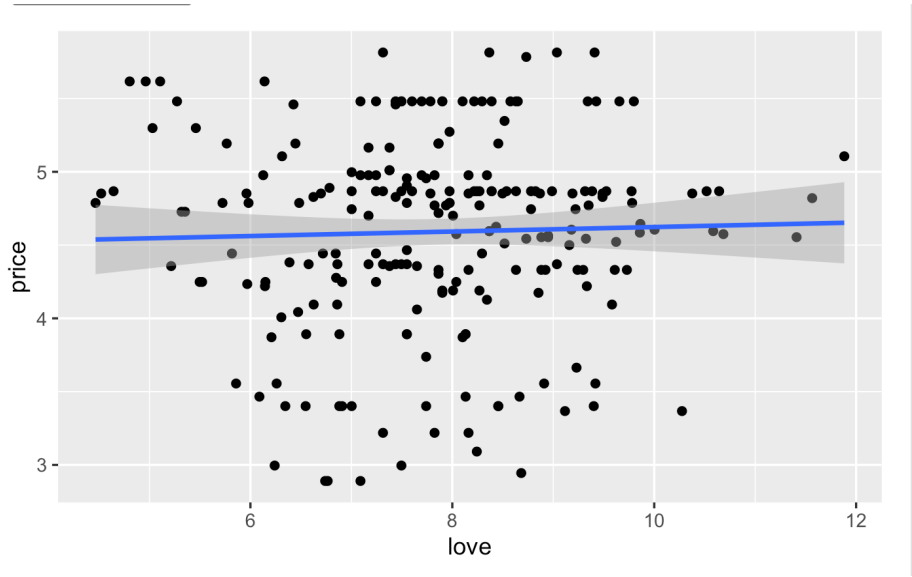
We can see some errors with brand-level.

```
$brand
                   (Intercept)
Acqua Di Parma       44.997388
AERIN                27.160440
HUDA BEAUTY           6.520422
KVD Vegan Beauty    -21.820680
Lancôme              -5.517529
Maison Louis Marie  -27.839722
Maison Margiela      -7.078919
The 7 Virtues       -20.843003
TOCCA               -37.909744
TokyoMilk           -15.720115
TOM FORD             75.461589
Tory Burch           -8.475669
Valentino             9.886225
Versace             -27.916418
Viktor&Rolf          17.771449
Yves Saint Laurent   -8.675713

with conditional variances for “brand”
```

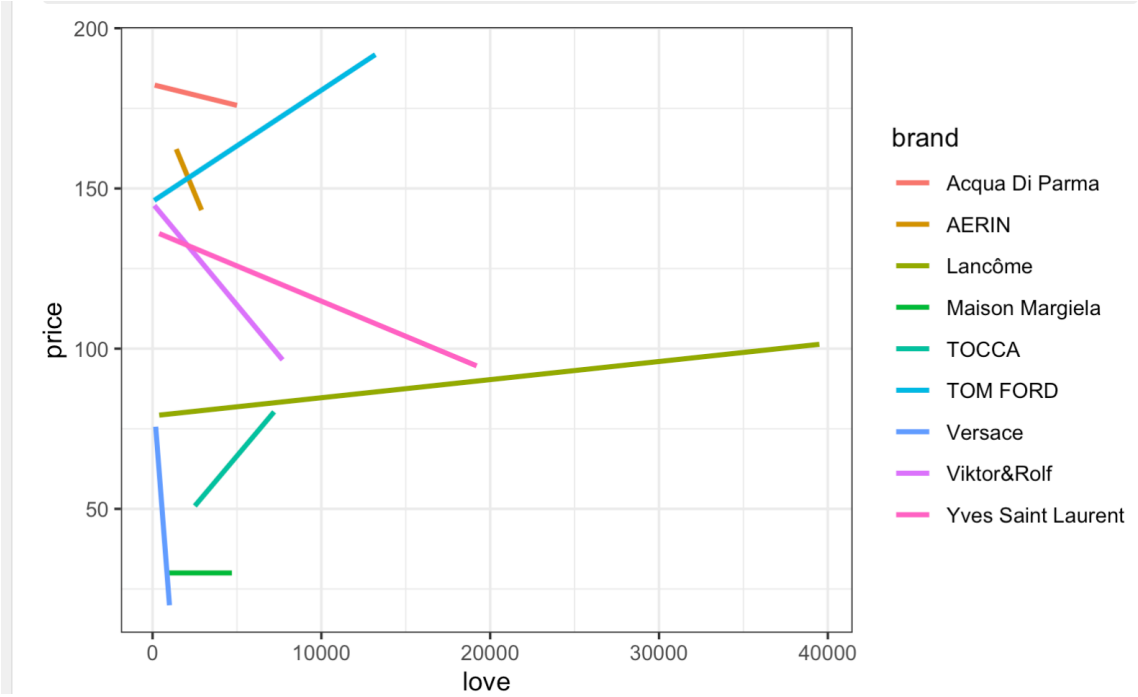
Through this conclusion, we can know that the value of the intercept is moving up or down in a specific brand. Acqua Di Parma is 44.997388 higher than the average, so the intercept of the regression line should be 44.997388 larger, which is y=(104.6528040+44.997388)-0.8464313x=149.650192-0.8464313x.

```
$brand
                   (Intercept)
Acqua Di Parma       47.721550
AERIN                28.438542
HUDA BEAUTY           7.300363
KVD Vegan Beauty    -22.384689
Lancôme              -6.605516
Maison Louis Marie  -26.182067
Maison Margiela      -8.189249
The 7 Virtues       -20.227069
TOCCA               -37.467501
TokyoMilk           -15.679824
TOM FORD             76.510528
Tory Burch           -8.477264
Valentino            10.036557
Versace             -30.220231
Viktor&Rolf          16.332480
Yves Saint Laurent  -10.906611

with conditional variances for “brand”
```

Through this conclusion, we can know that the value of the intercept is moving up or down in a specific brand. Acqua Di Parma is 47.721550 higher than the average, so the intercept of the regression line should be 47.721550 larger, which is y=(9.836165e+01+47.721550)-0.8464313x.

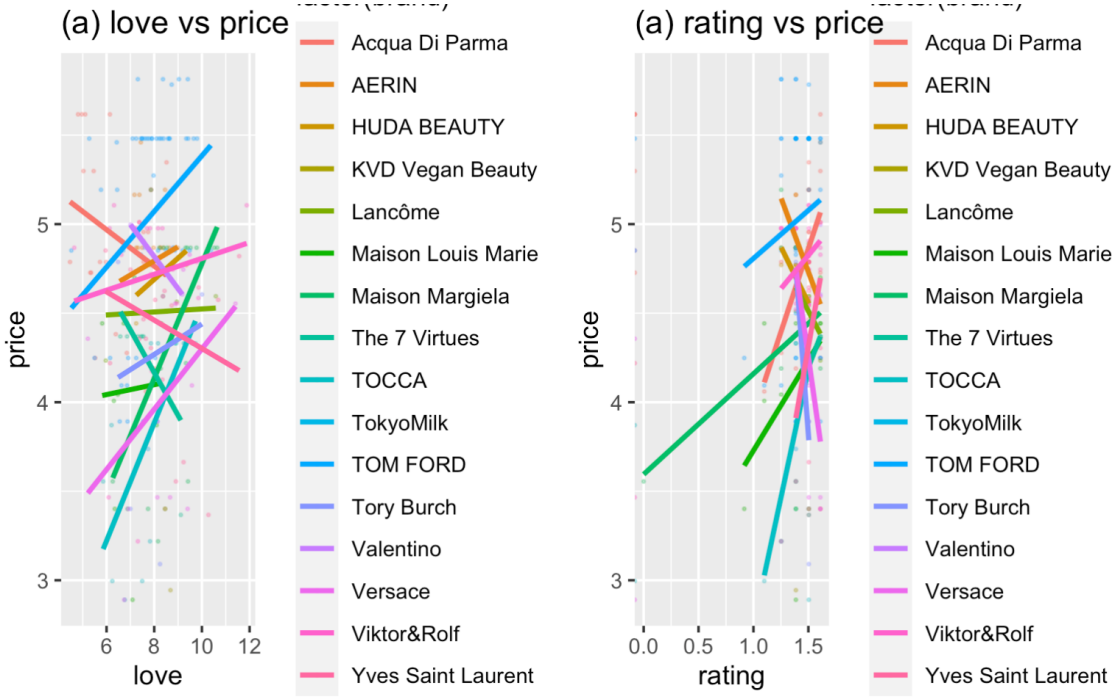


When using price~love as variables, the data is particularly scattered and the graph presented is meaningless. (did not show strong relationship between love and price, because due to the analysis for famous perfume, the love about each type may not effectively change its price and their price will trend to be stable. But about one specific brand, situations may be different. For example, I randomly choose the brands with rating of 5.0, to see how that can make difference about the love~price relationship)



For the graph above,we can see that different kind of brands with rating of 5.0 will have different relationship with price and love. The slope line for “Lancome”, “TOM FORD”

and “Maison Margiela” show obvious increase, while some other brands show different situations.



When I use the brand as the group, the graph can be linear, which shows that this is meaningful. (a) The love vs. price graph shows that the price of "Tom Ford" is relatively high compared to other brands. At the same time, we can also see that the price of "Maison Margiela," "TOCCA," and "Versace" increases with the degree of love.

From graph (a) rating vs. price, it can be seen that the higher the rating, the higher the relative price. Particularly notable are "TOCCA," "Lancôme," and "AERIN."

Results

When the brand is used as a group to analyze price~love and price~rating, they are linearly related, and most of them is that the higher the love is, the higher the price is. The higher the rating, the higher the price. The more well-known brands are also more expensive relative to other brands.

Reference

https://www.kaggle.com/datasets/raghadalharbi/all-products-available-on-sephora-website

Gelman, A. and Hill, J. (2018) “CHAPTER 11 Multilevel structures,” in *Data analysis using regression and multilevel/hierarchical models*. Cambridge u.a.: Cambridge Univ. Press.