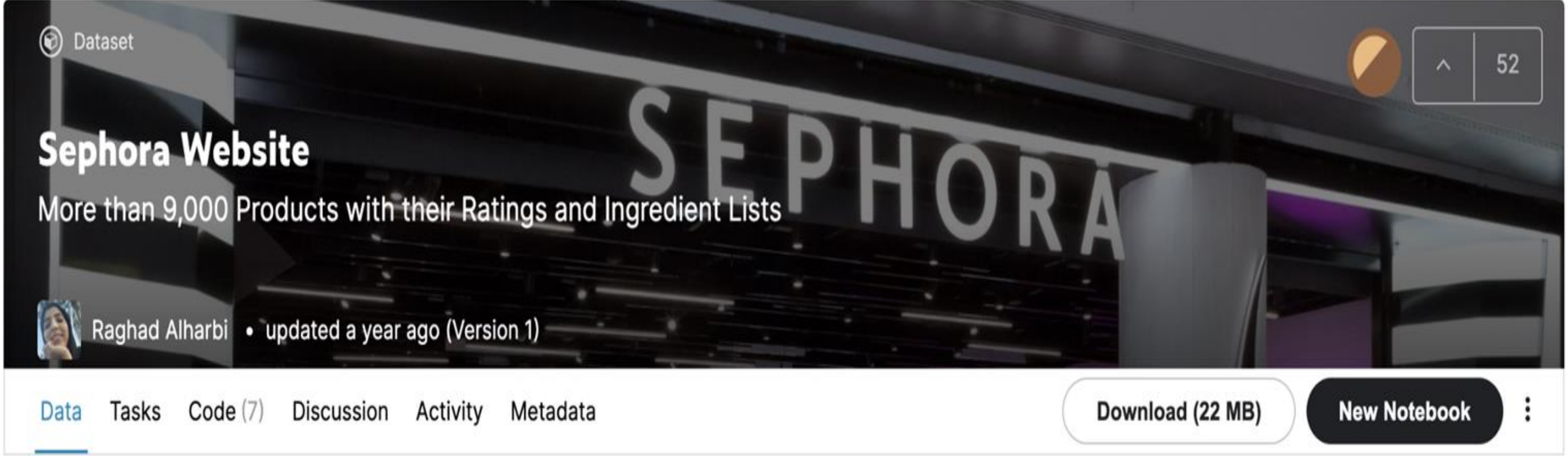


Sephora Website

More than 9,000 Products with their Ratings and Ingredient Lists

Raghad Alharbi • updated a year ago (Version 1)

Data Mining Project - Sephora Website

Name : Kurnia Anwar Ra'if

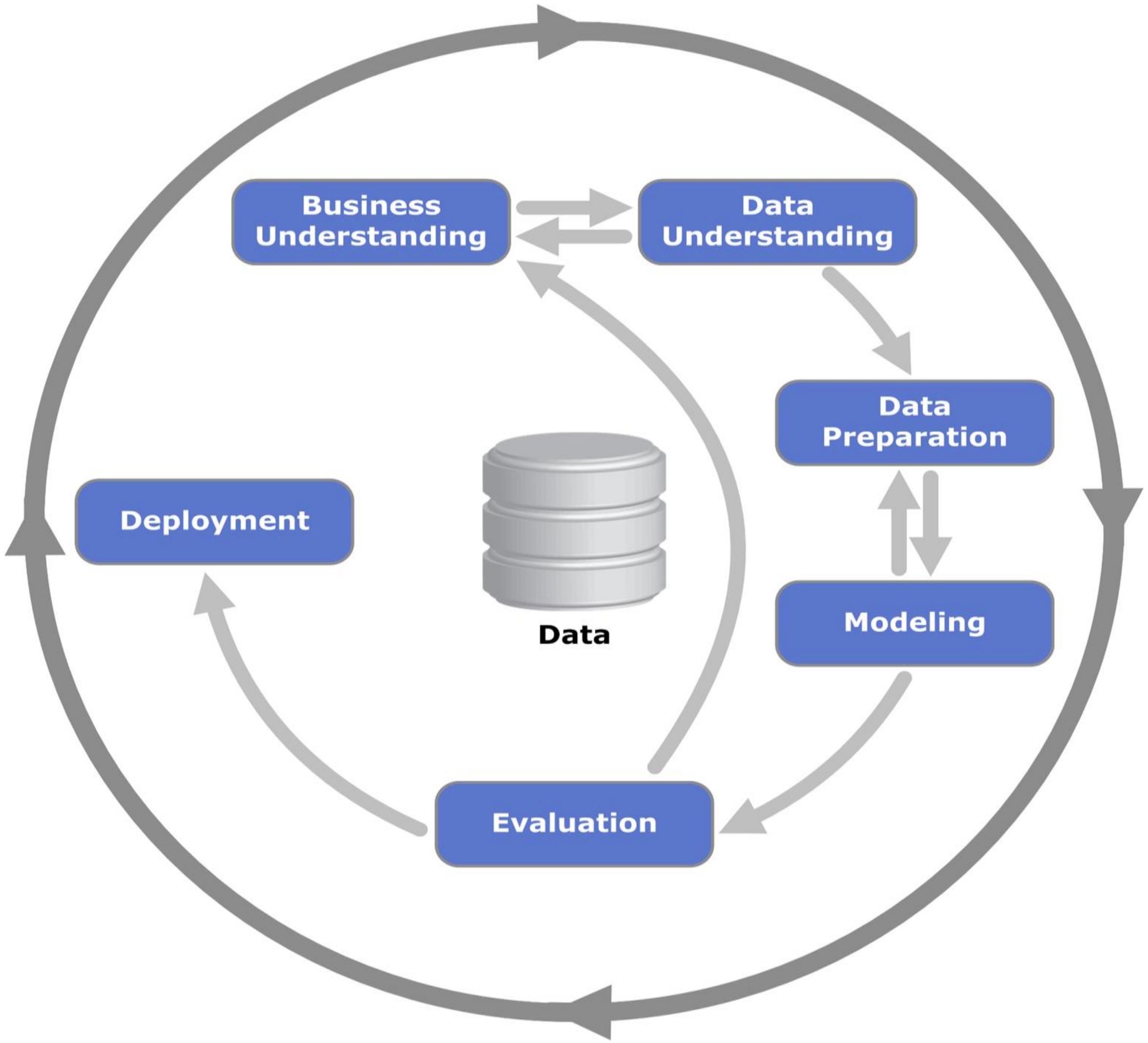
Bootcamp Data Science batch - 4

Data Dictionary

This dataset is explain sephora, sephora is a visionary beauty-retail concept founded in France by Dominique Mandonnaud in 1970. Sephora's unique, open-sell environment features an ever-increasing amount of classic and emerging brands across a broad range of product categories including skincare, makeup, fragrance, body and hair care. Sephora operates approximately 1,900 stores in 29 countries worldwide, with an expanding base of over 200 stores across the Asia Pacific region including Australia, China, Singapore, Malaysia, Thailand, Indonesia & India.

Feature	Type	Description
id	int	The product ID at Sephora's website
brand	object	The brand of the product at Sephora's website
category	Object	The category of the product at Sephora's website
name	Object	The name of the product at Sephora's website
size	Object	The size of the product
rating	float	The rating of the product
numberofreviews	int	The number of reviews of the product
love	int	The number of people loving the product
price	float	The price of the product
value_price	float	The value price of the product (for discounted products
URL	object	The URL link of the product
MarketingFlags	bool	The Marketing Flags of the product from the website if they were exclusive or sold online only
MarketingFlags_content	object	The kinds of Marketing Flags of the product
options	object	The options available on the website for the product like colors and sizes
details	object	The details of the product available on the website
howtouse	object	The instructions of the product if available
ingredients	object	The ingredients of the product if available
online_only	int	If the product is sold online only
exclusive	int	If the product is sold exclusively on Sephora's website
limited_edition	int	If the product is limited edition
limitedtimeoffer	int	If the product has a limited time offer

CRISP-DM Methodology

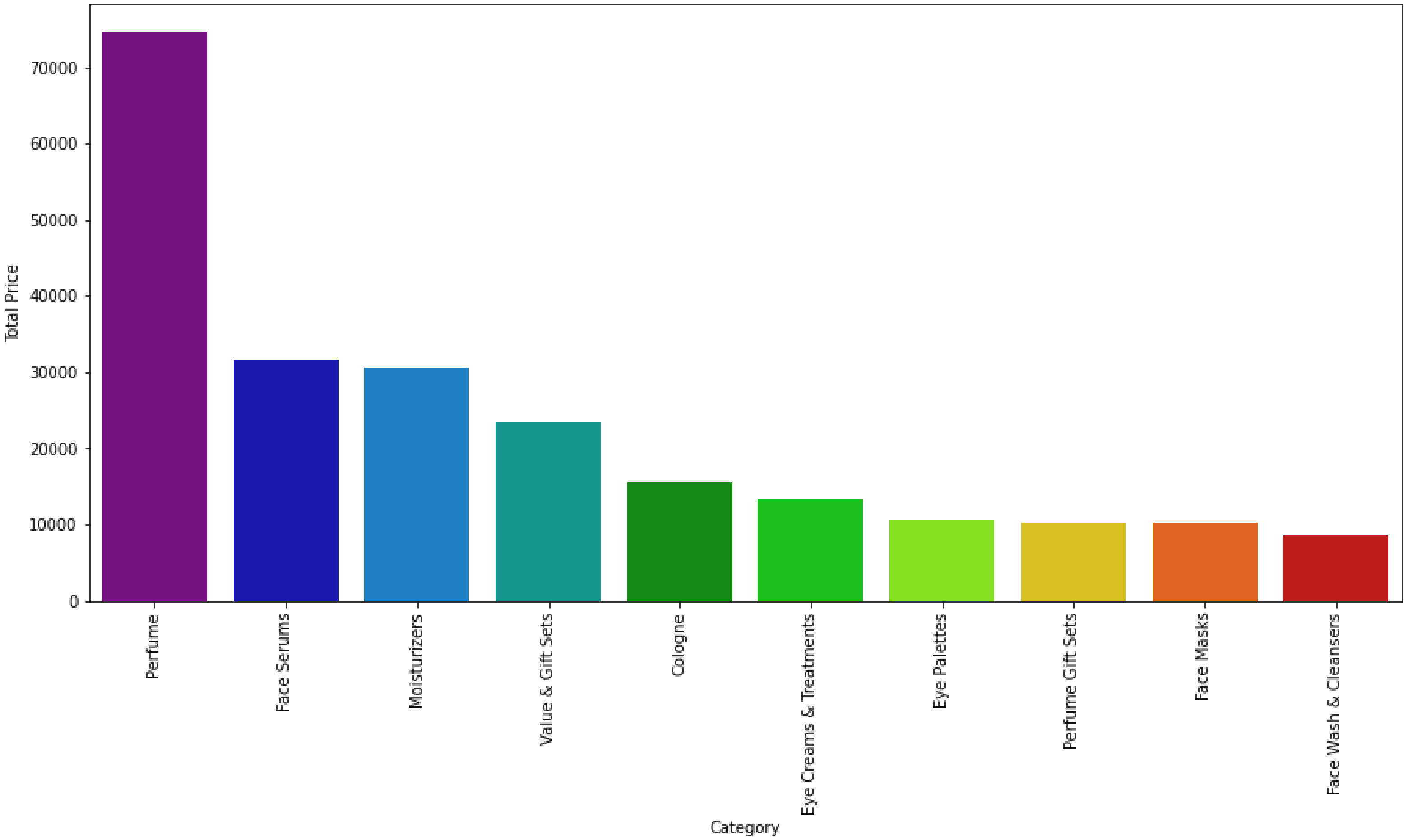


Cross-industry standard process for data mining, known as CRISP-DM, is an open standard process model that describes common approaches used by data mining experts. It is the most widely-used analytics model.

Business Questions Guideline :

- **What is category with the highest income price value ?**
- **What is brand with the highest Income price value ?**
- **What is Category with the highest sales from the highest income value for brand ?**
- **What is brand with the most exclusive ?**
- **What is category with the most Exclusive ?**

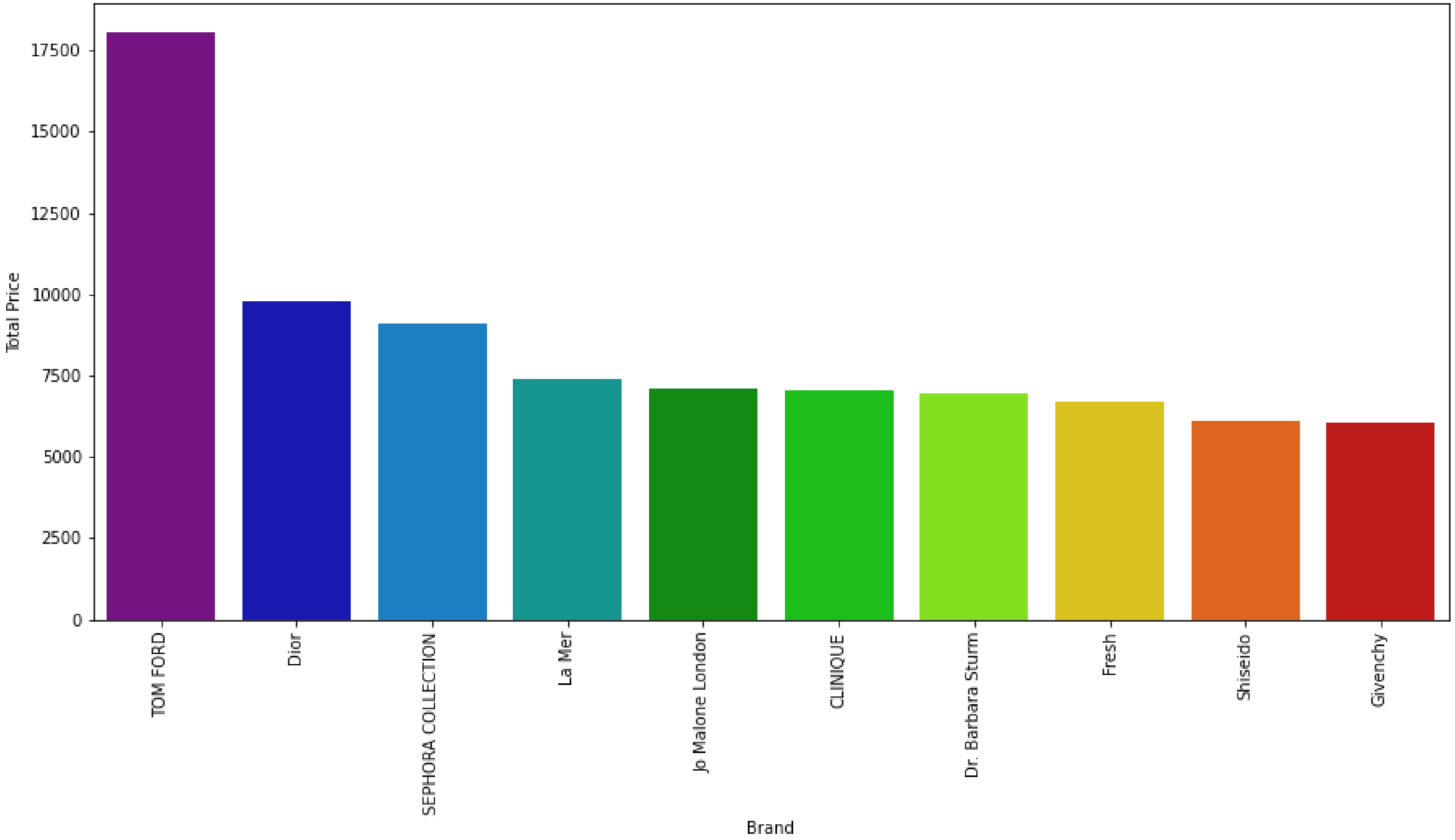
What is category with the highest income price value ?



Building perception

Perfume Category has the highest income price value > $ 70000, more than another category which Perfume > 2x price value compared another category. To gain the income from face wash cleaners and face mask can combine with face serums in advertisement which has face category, it hopes can gain face mask and face wash cleaners 2x income with 2x height bar char.

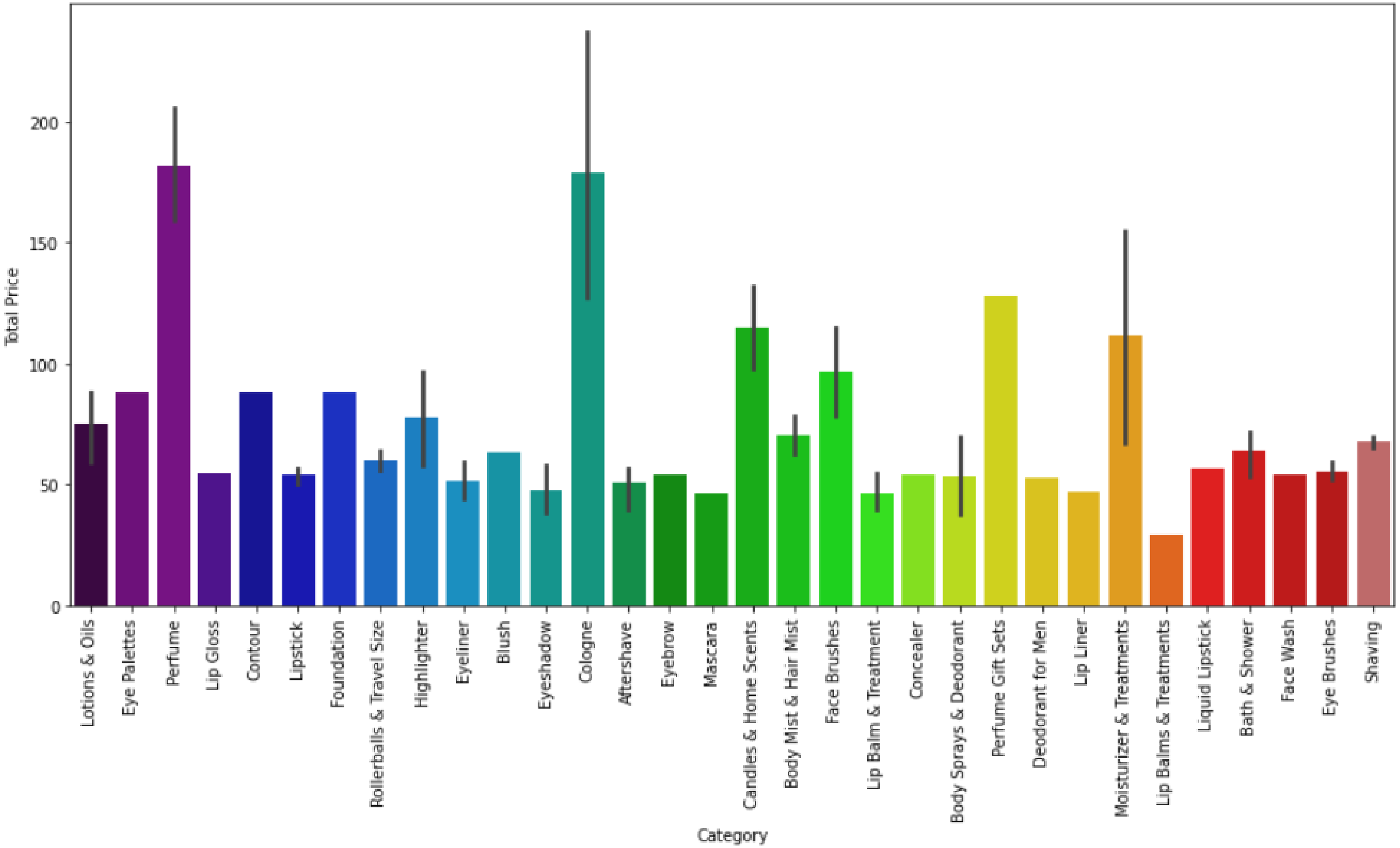
What is brand with the highest Income price value ?



Building perception

Perfume Category has the highest income price value > $ 17500, more than another brand which Perfume > 2x price value compared another brand except Dior and Sephora Collection.

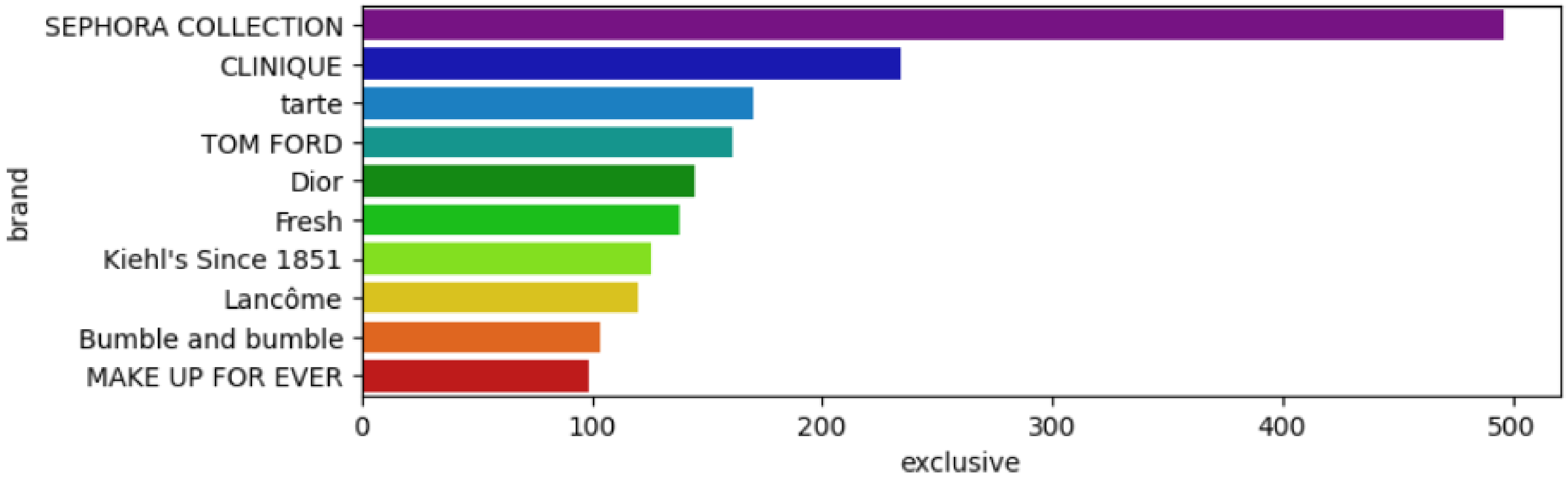
What is Category with the highest sales from the highest income value for brand ?



Building perception

From Tom Ford brand, we can see clearly that cologne and perfume have the highest total price near $200 for each category. If we see before, the highest category is perfume so, for all brand perfume is dominated in selling.

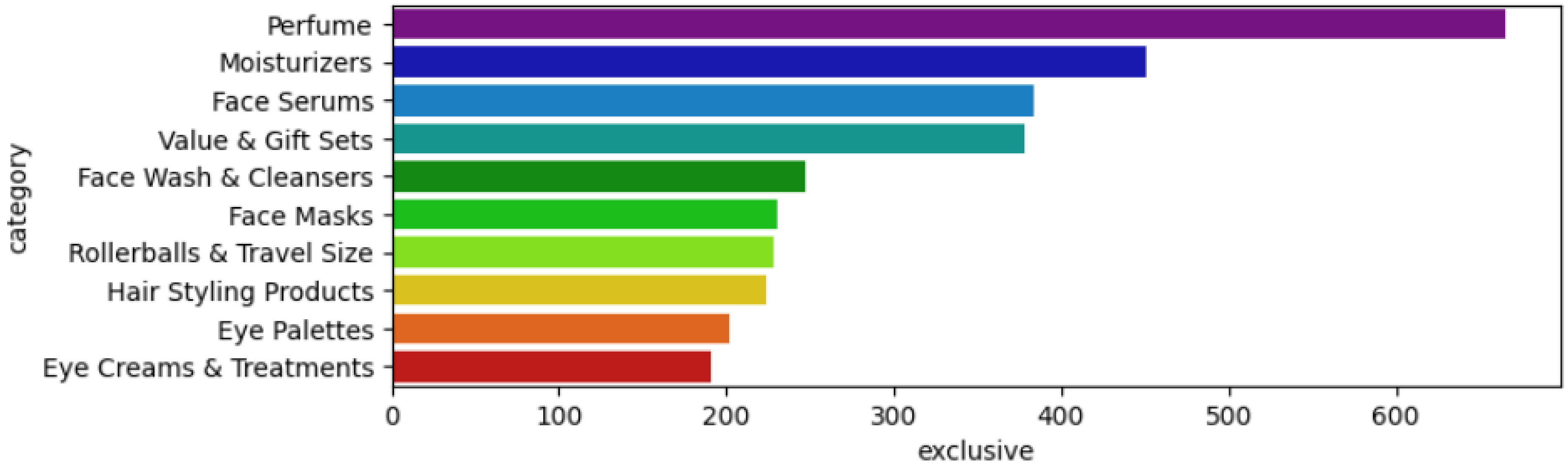
What is brand with the most exclusive ?



Building perception

From the brand, sephora collction has the most exclusive bran which the number of exclusive is 500, it wins from TOM Ford brand who has highest income. For gain the income, sephora can do promotions which the target is lowest seling category in Tom Ford. It hopes can gain sephora the income by the category who has lowest selling product in Tom Ford

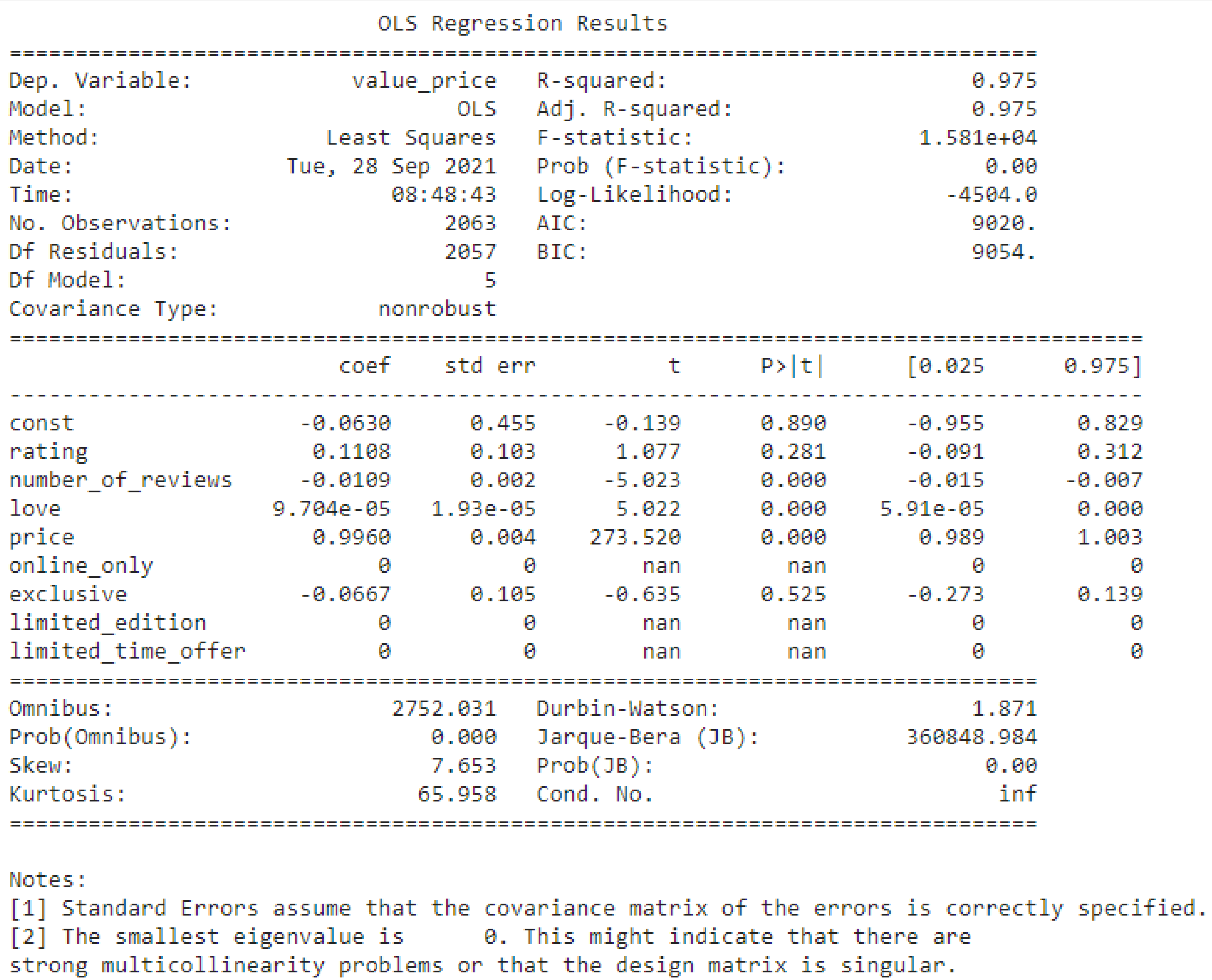
What is category with the most Exclusive ?



Building perception

From the category with the most exclusive is perfume with the value of exclusive more than 600. Both the highest income and most exclusive, perfume is favorite product to buy than another product.

Regression Modelling



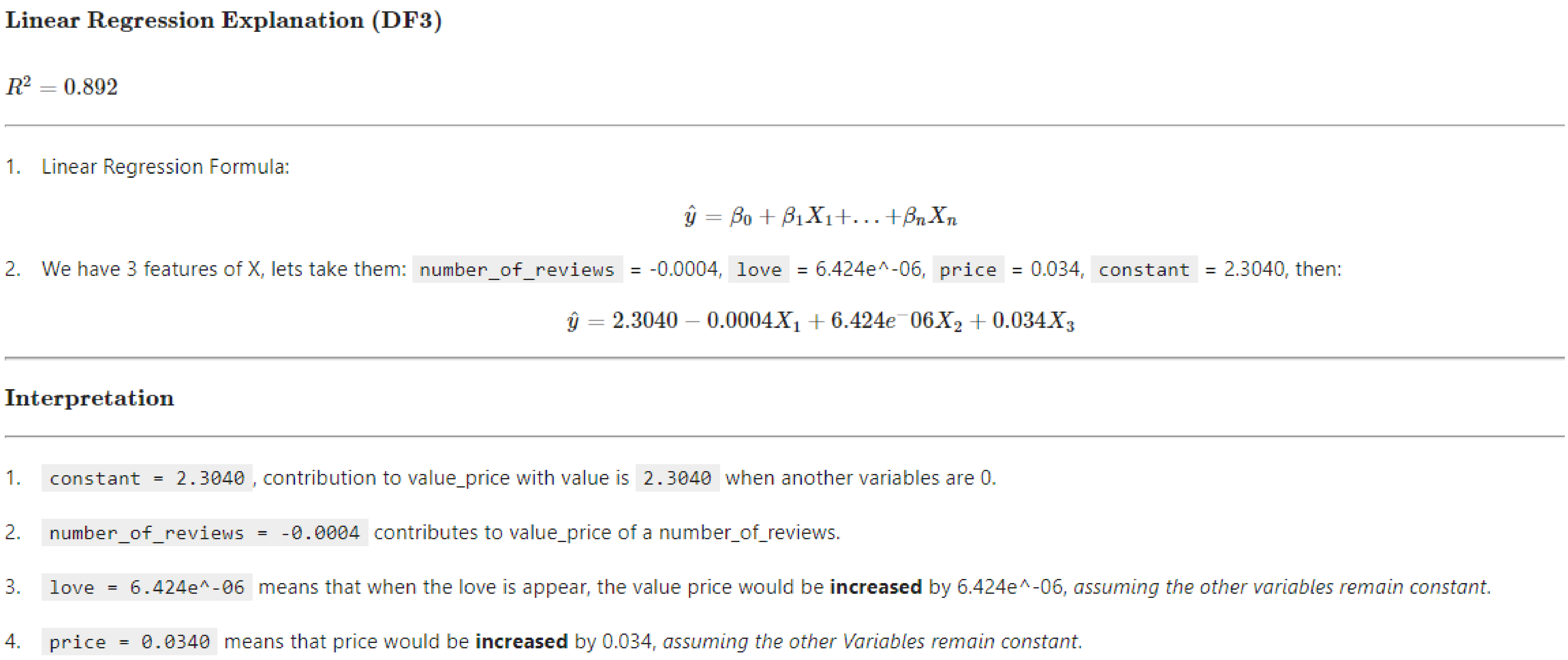
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:            value_price   R-squared:                       0.975
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                 1.581e+04
Date:                Tue, 28 Sep 2021   Prob (F-statistic):               0.00
Time:                        08:48:43   Log-Likelihood:                -4504.0
No. Observations:                2063   AIC:                             9020.
Df Residuals:                    2057   BIC:                             9054.
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
======================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                -0.0630      0.455     -0.139      0.890      -0.955       0.829
rating                0.1108      0.103      1.077      0.281      -0.091       0.312
number_of_reviews    -0.0109      0.002     -5.023      0.000      -0.015      -0.007
love               9.704e-05   1.93e-05      5.022      0.000    5.91e-05       0.000
price                 0.9960      0.004    273.520      0.000       0.989       1.003
online_only                0          0        nan        nan           0           0
exclusive            -0.0667      0.105     -0.635      0.525      -0.273       0.139
limited_edition            0          0        nan        nan           0           0
limited_time_offer         0          0        nan        nan           0           0
==============================================================================
Omnibus:                     2752.031   Durbin-Watson:                   1.871
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           360848.984
Skew:                           7.653   Prob(JB):                         0.00
Kurtosis:                      65.958   Cond. No.                          inf
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is      0. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Regression Modelling which has 97,5% but P > | t | that has not significant value are rating, exclusive, and has nan are online only, limited edition, and limited time offer. For this reason, remove and do the regression modelling again. Before to do that, it’s good for check classical assumption.

The next step is to analyze the column 'value_price', because the target variable is numeric then look at the histogram whether distributed normally or not. in the column, in the 'value_price' column, we can see a positive skewed because the tail of the distribution is to the right of the most value. That is, most distributions are in low value. So, the target variable is right skewed. As (linear) models love normally distributed data , we need to transform this variable and make it more normally distributed. We will apply log transformation to the feature to make the distribution close to gaussian. We will apply log(1+x) transformation to avoid 0 values (if present)

Do the Regression model again



Linear Regression Explanation (DF3)

$R^2 = 0.892$

---

1. Linear Regression Formula:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$$

2. We have 3 features of X, lets take them: `number_of_reviews` = -0.0004, `love` = 6.424e^-06, `price` = 0.034, `constant` = 2.3040, then:

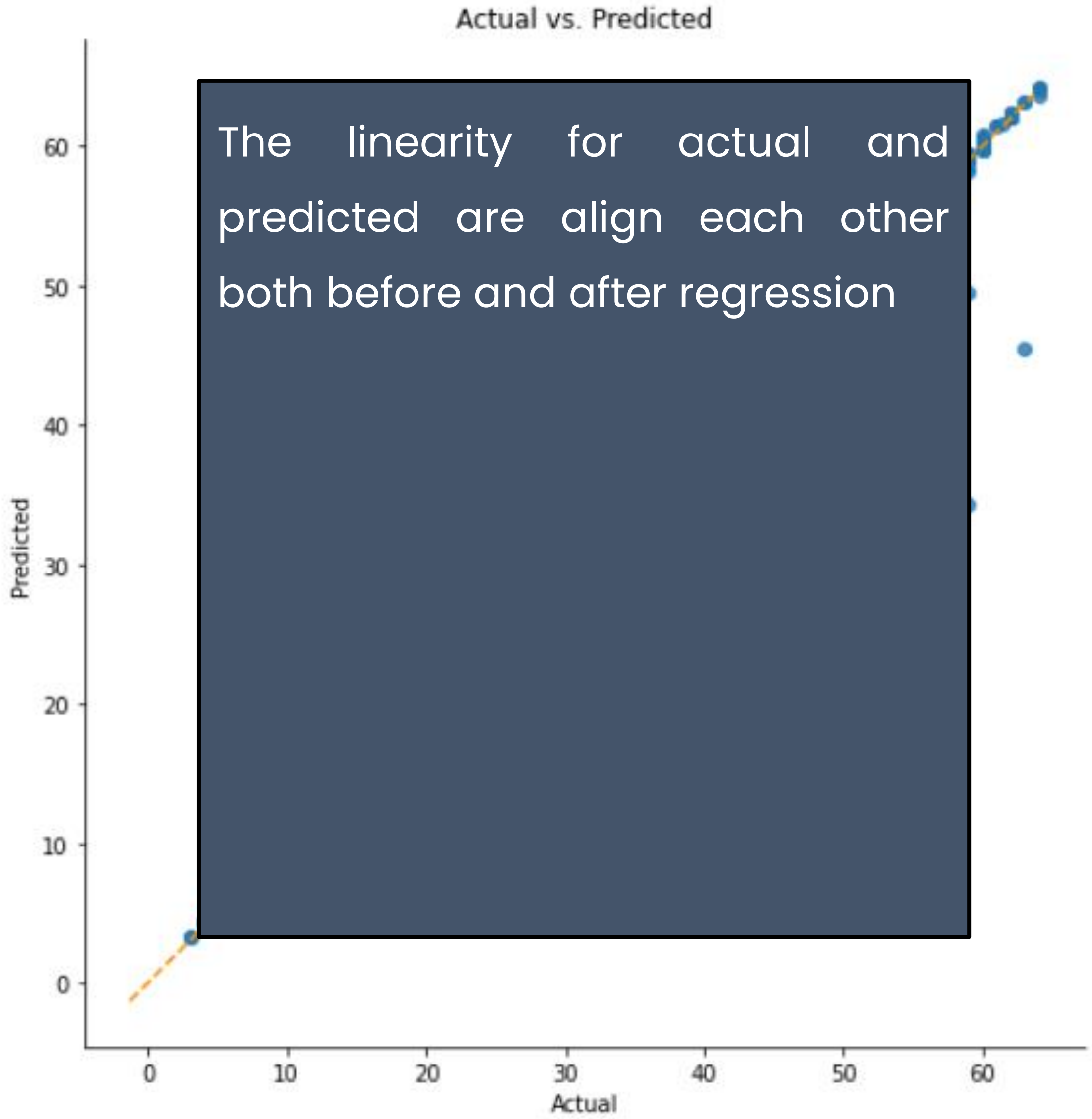
$$\hat{y} = 2.3040 - 0.0004X_1 + 6.424e^{-}06X_2 + 0.034X_3$$

Interpretation

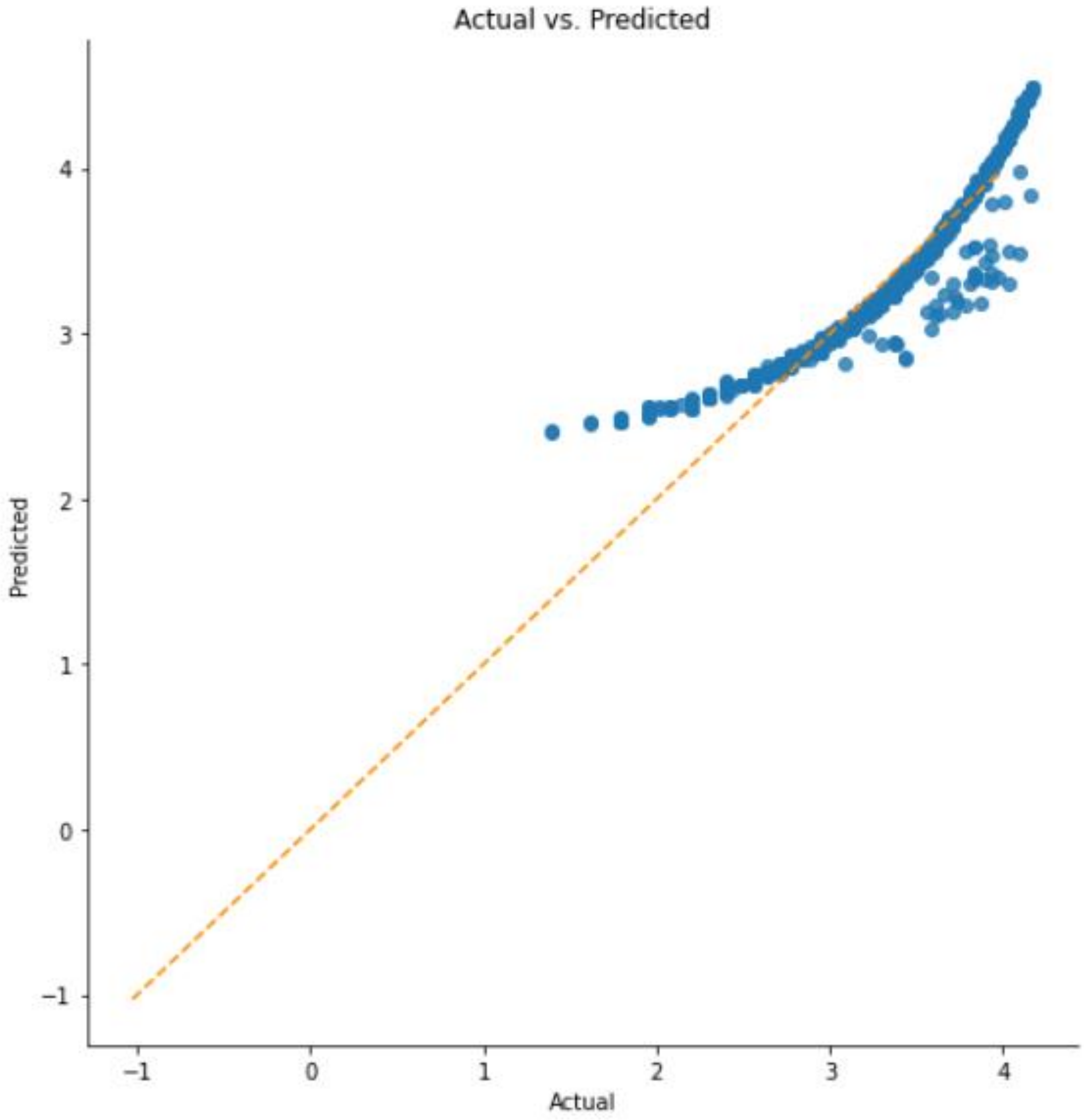
---

1. `constant = 2.3040` , contribution to value_price with value is `2.3040` when another variables are 0.
2. `number_of_reviews = -0.0004` contributes to value_price of a number_of_reviews.
3. `love = 6.424e^-06` means that when the love is appear, the value price would be **increased** by 6.424e^-06, *assuming the other variables remain constant.*
4. `price = 0.0340` means that price would be **increased** by 0.034, *assuming the other Variables remain constant.*

Linearity : Before (left) and after (right) regression

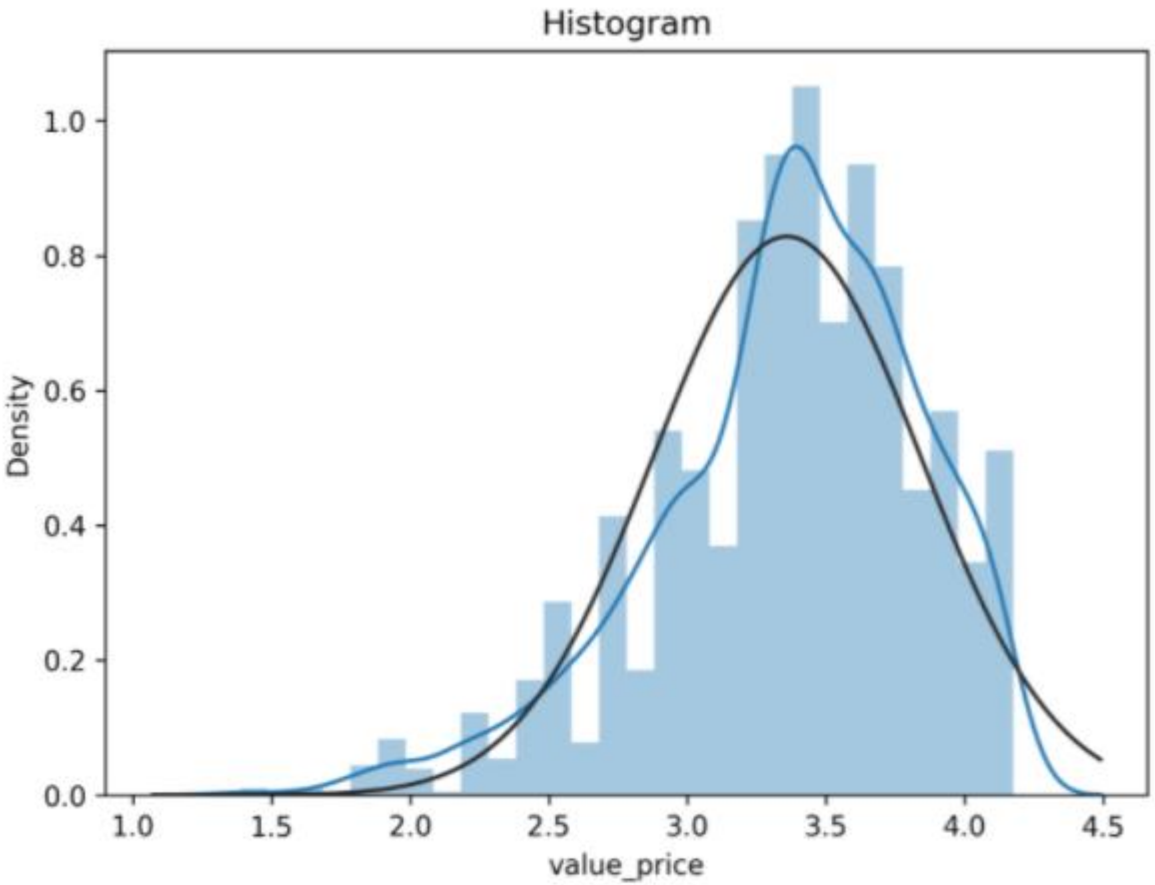


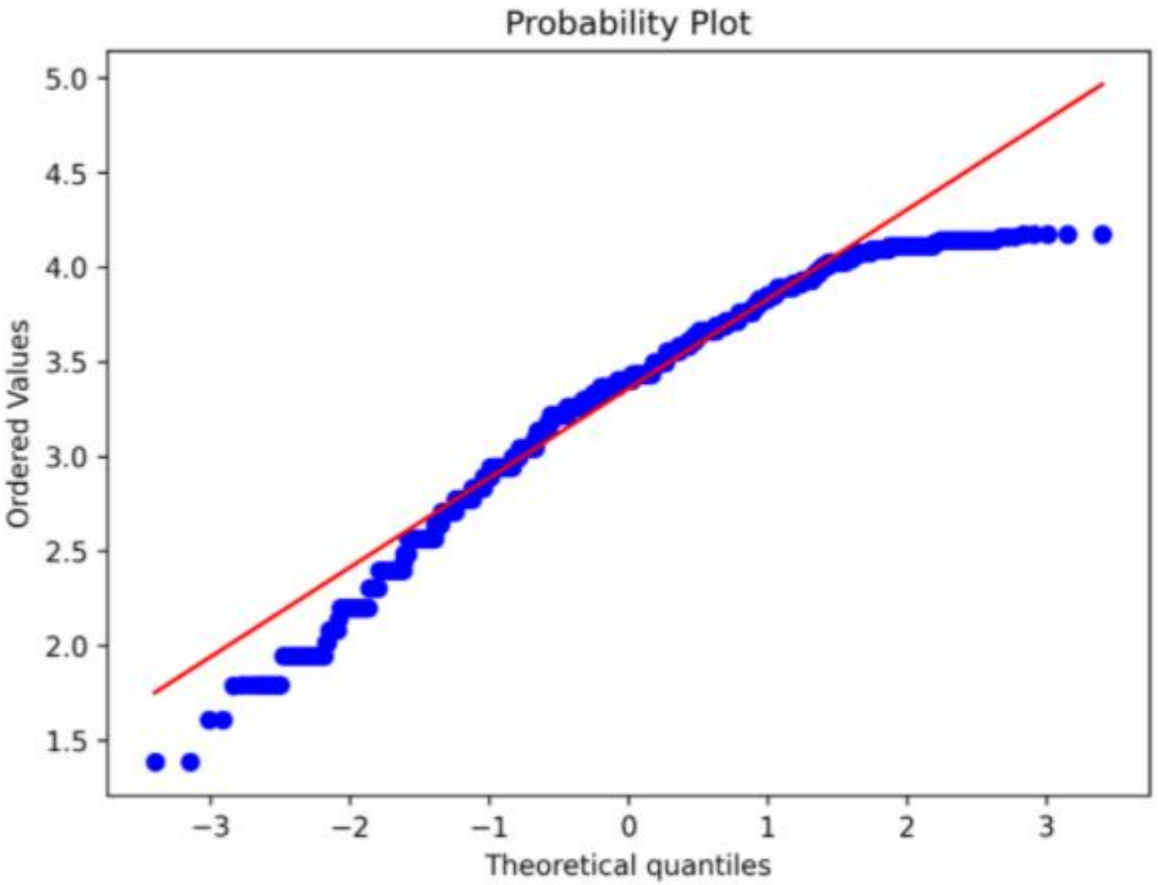
The linearity for actual and predicted are align each other both before and after regression



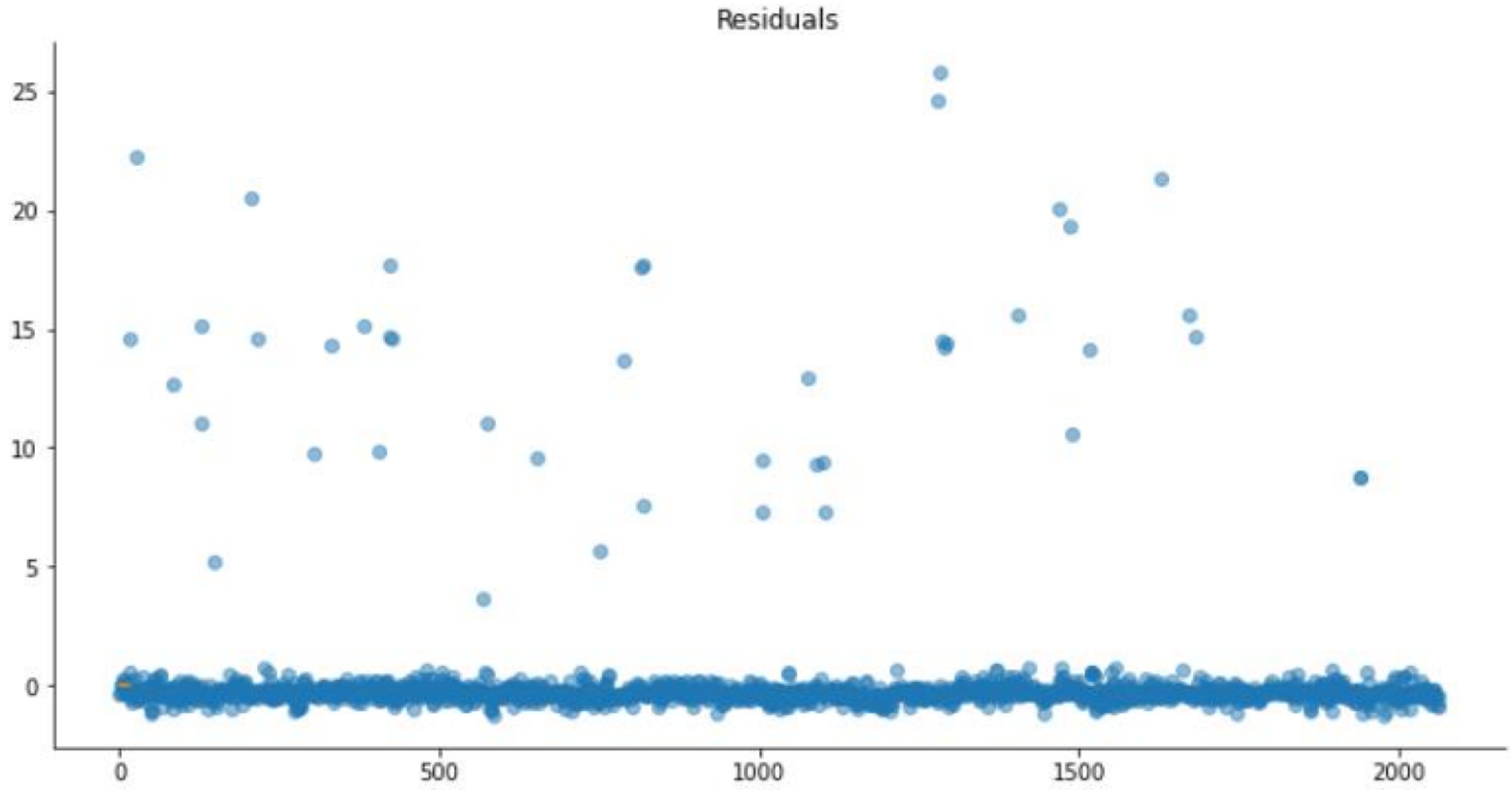
Normality of Residual is better than before

Because in probability plot, each dot at the red line dominated so we can do the Regression Model with df2 new that apply log(1+x)

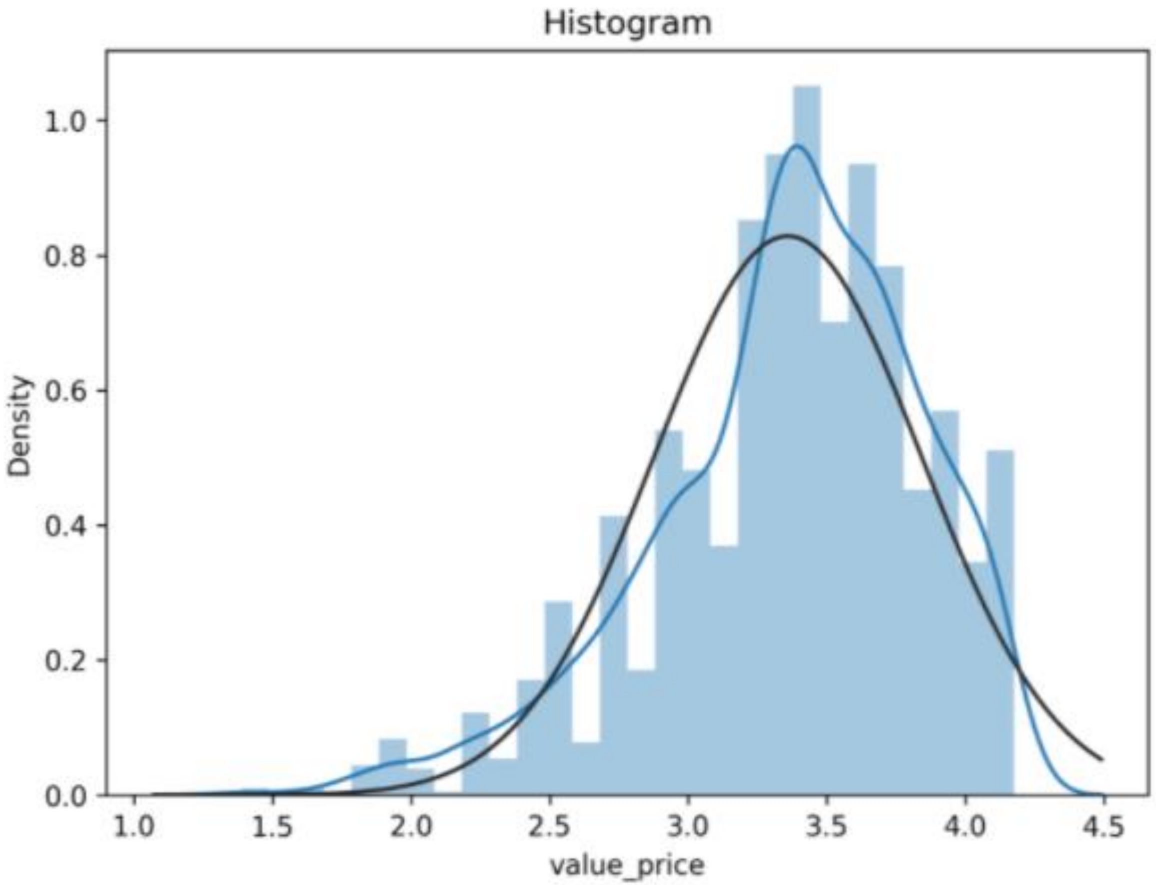


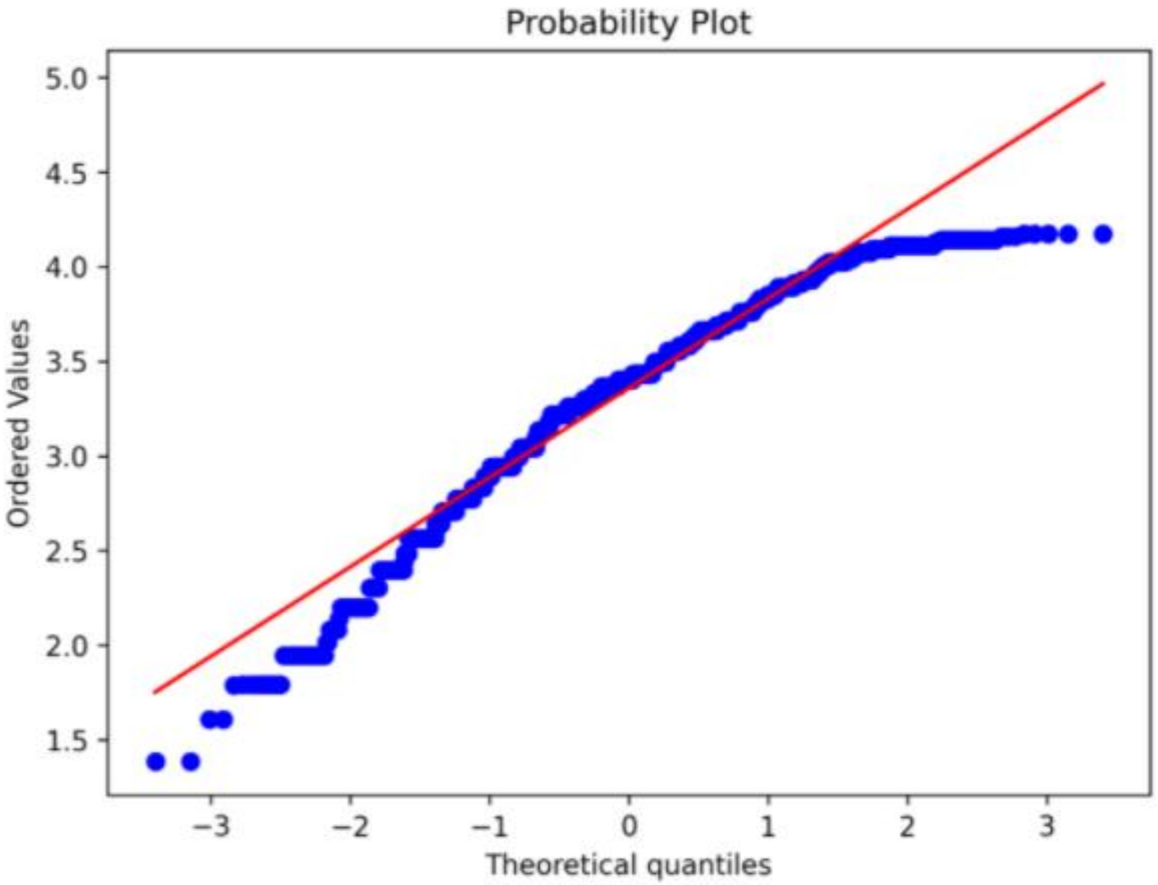


A linear horizontal pattern -> it means non-heteroscedasticity



Normality of Residual is better than before

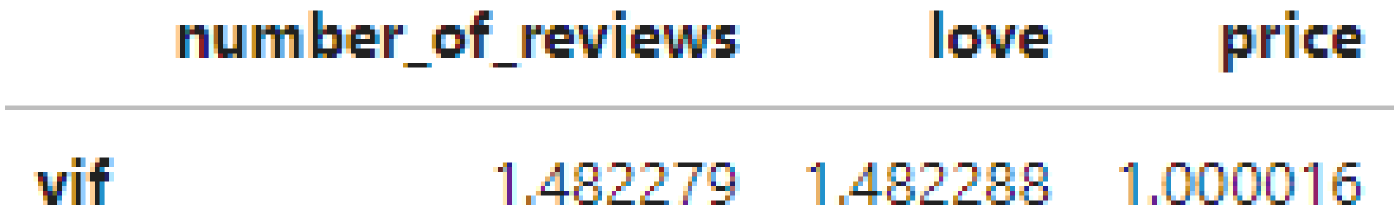




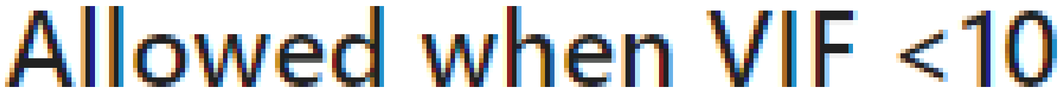
A. Non – Autocorrelation

B. VIF Value < 10 means *Non-Multicollinearity*

- Durbin-Watson: 1.5424818948135521
- Little to non autocorrelation
- Assumption satisfied



	number_of_reviews	love	price
vif	1.482279	1.482288	1.000016



Allowed when VIF <10

Thank you ☺