Analysis and linear regression on Sephora Dataset

By **Alanoud Almutairi, Alnirah Alqahtani**

About:

The Project will analyze Sephora product pages and visualize the price of the products, also try to select which categories of products seem to perform better. Additionally, an attempt will be made to understand the relationship between ratings, price, categories. Lastly, this analysis of the product's ingredients based on preset categories is made, this type of analysis can be relevant and helpful for marketing and formulation teams in cosmetic companies.

Approach and methodology:

- Data cleaning and check duplicated values.
- Visualizing relationship between some features to make data more understandable.
- Find top ten brands and top ten products
- Find linear regression and use selection feature to split data.

Analysis:

The first step of our project is using pandas to clean data and check for duplicated columns and null values. secondly, using selection feature to select the data of type 'Object'. After that, we looked for the top rating products in all brands, Also find correlation between all features of the dataset and using heatmap to visualize this data, finding relationship between price and value price and plot it using scatter plot, The last step is seeking for top brands using exclusivity feature and using bar plot to plot that data to use it as a reference to the cosmetic companies.

Result:

```
In [6]:  1 dub = df.duplicated()
         2 print(" number of dublicate row= %d"% (dub.sum()))

         number of dublicate row= 0
```

There are no data duplication was found in this dataset

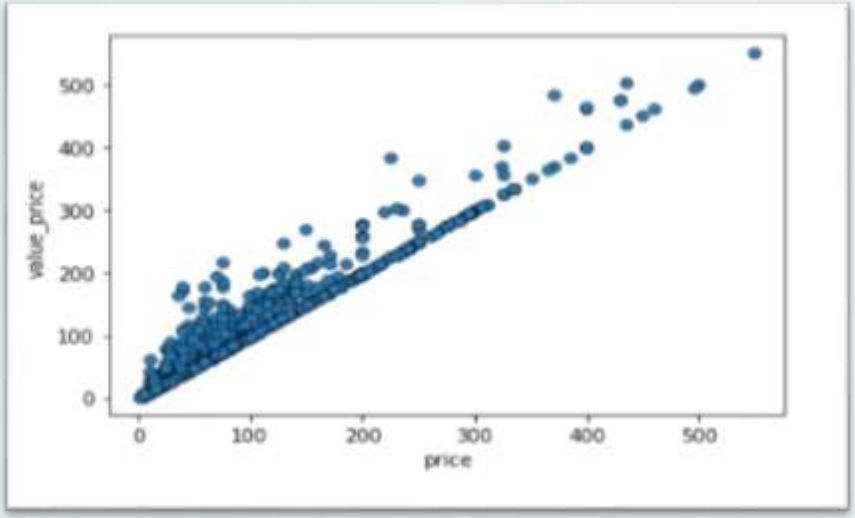
Linear regression	Polynomial model
0.041806308626942124	0.892930892579861

```
1 df.isnull().sum()

id                      0
brand                   0
category                0
name                    0
size                    0
rating                  0
number_of_reviews       0
love                    0
price                   0
value_price             0
URL                     0
MarketingFlags          0
MarketingFlags_content  0
options                 0
details                 0
how_to_use              0
ingredients             0
online_only             0
exclusive               0
limited_edition         0
limited_time_offer      0
dtype: int64
```

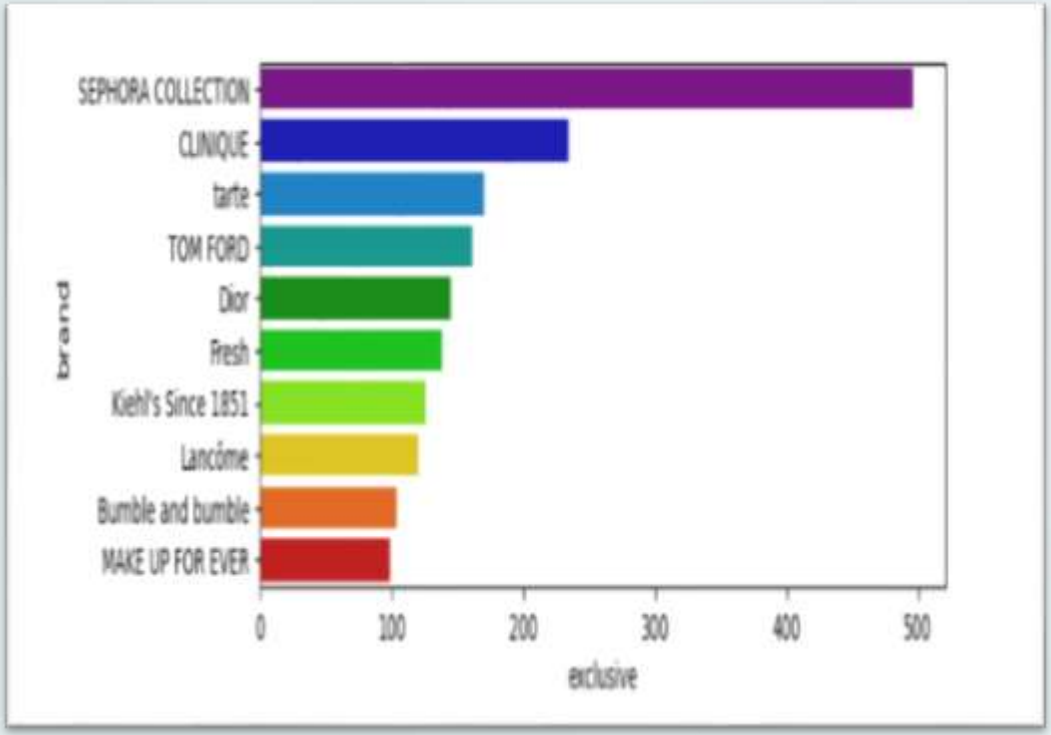
There are no missing value was found from every rows in this dataset

relationship between price and value price :



And using heatmap to visualize the relationship between all features:

Also top ten brands based on exclusive feature:



Top ten most popular product based on rating:

:

brand	rating
Four Sigmatic	5.00
Fable & Mane	5.00
Aether Beauty	5.00
Montblanc	5.00
Golde	4.88
RODIN olio lusso	4.81
The Art of Shaving	4.70
Paco Rabanne	4.69
SOBEL SKIN Rx	4.69
dae	4.67

Recommendations:

- Regarding the Sephora brand and products that does not have the highest sales transaction, we recommend that the company should further increase sales of the Sephora brand using more advertisement for that products. So that way, Sephora is not too dependent on other brands.
- The top 10 brands we can see the other brands doesn't have that much of exclusive products, so we recommend that each company at least approve more than 300 exclusive product to increase the sales of the brand and company.

THANKS