Analysis and linear regression on Sephora Dataset

Project report

Alanoud Almutairi
alaanouud@gmail.com

Alnirah Alqahtani
alnirahq20@gmail.com

Abstract :

Our second project in the Data Science Course with SDAIA Academy. The project was about using web scraping methods to collect more than 9,000 data records from any website we choose. We decided to choose the Sephora dataset because we need to think about beauty brands and analyze it more. 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.

Design:

In this project we used the Sephora brand dataset to collect and analysis data and use them to help marketing and formulation teams in cosmetic companies understanding this data by visualizing the rating of the products, top brands and relationship between price and other features.

Data :

The Sephora dataset is interesting for this type of analysis for a few reasons such as It is very recent There are 9168 observations and 21 different variables in varying data types. And down there a summary table of the variables, their data types and a short definition of the dataset that we will work on it:

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
number*of*reviews	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
how*to*use	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
limited*time*offer	int	If the product has a limited time offer

Algorithms:

- Feature selection
- Feature engineering

Visualization:

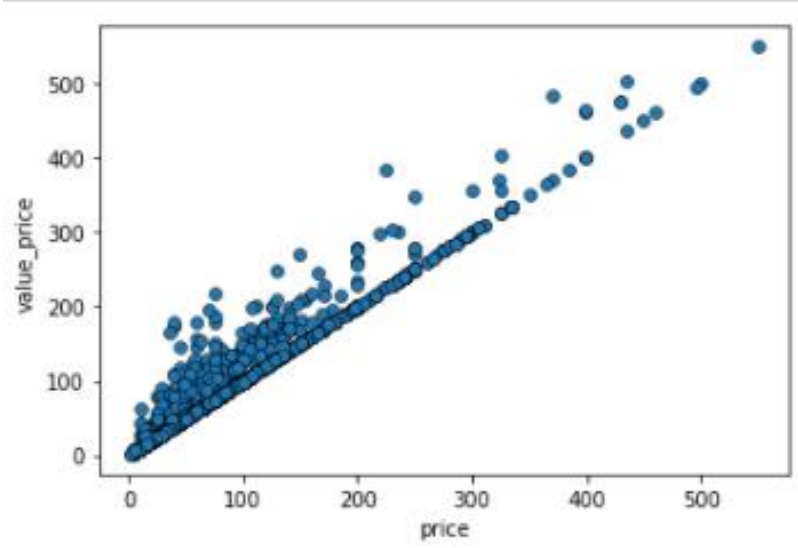
- Scatter plot: Represent relationship between price and value price
- Heatmap: Represent the relationship between all features
- Bar plot: Represent Top ten brands.

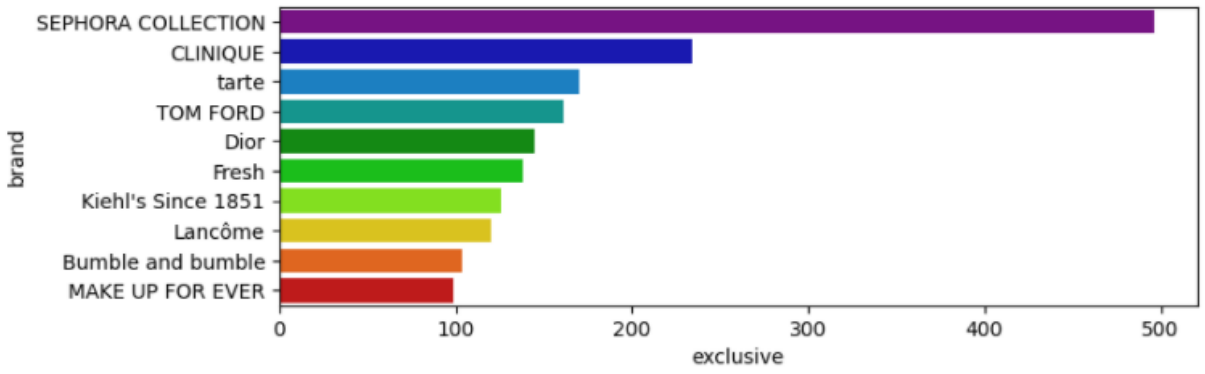
Tools:

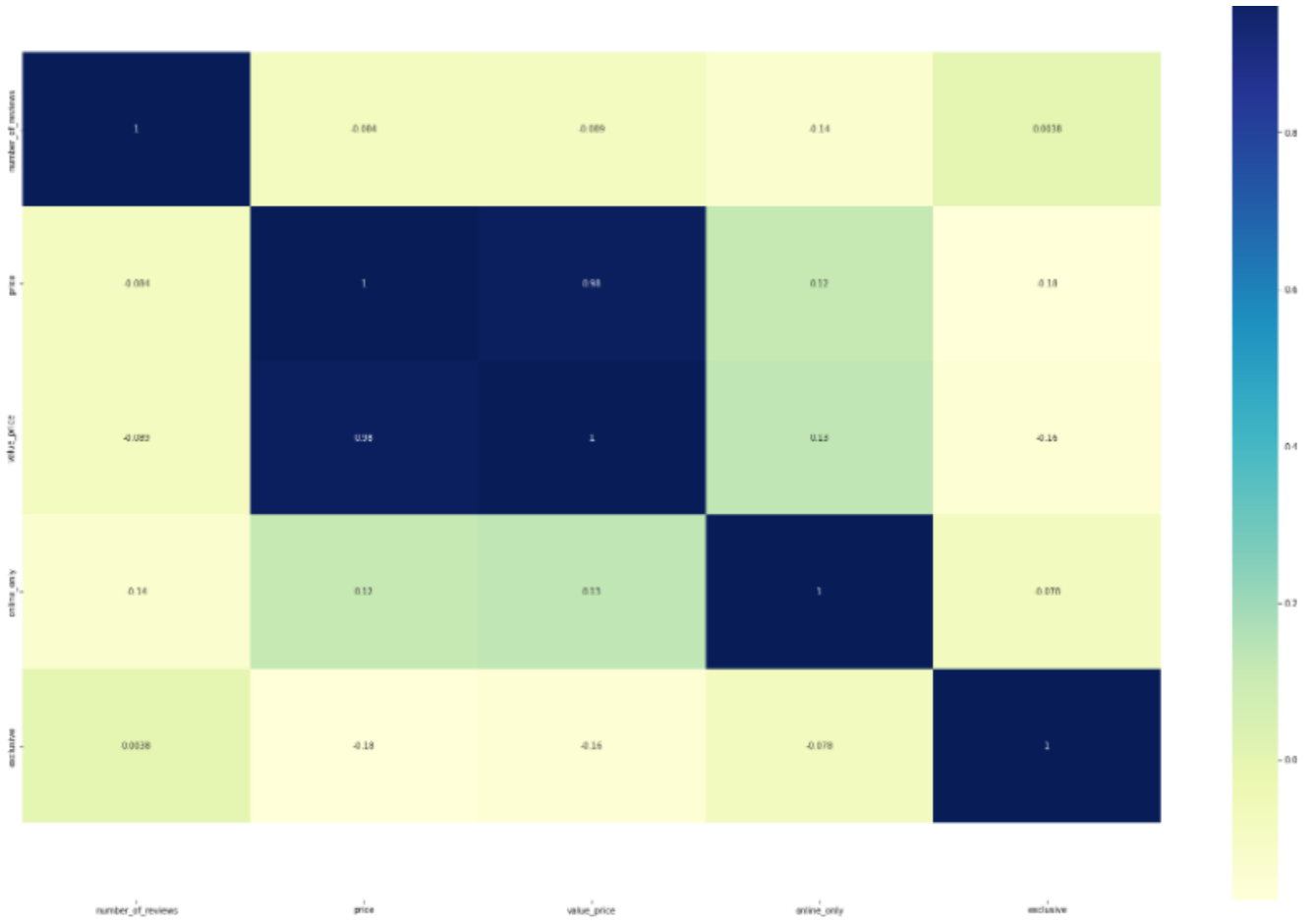
Technologies : Python,Jupyter Notebook

Libraires : Pandas, Numpy ,Seaborn ,Sklearn.

Communication:






number_of_reviews
price
value_price
online_only
exclusive
1
-0.084
-0.089
-0.14
0.0038
-0.084
1
0.98
0.12
-0.18
-0.089
0.98
1
0.13
-0.16
0.14
0.12
0.13
1
-0.078
0.0038
-0.18
-0.16
-0.078
1
0.8
0.6
0.4
0.2
0.0

Recommandations:

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. 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.