Sephora Beauty Products

S. Jai Harish

CB.EN.P2AIE22013

Dr.Mansi Sharma

Introduction

- Sephora is a well-known multinational beauty retailer that offers a wide range of cosmetic, skincare, haircare, and fragrance products.
- Sephora carries an extensive selection of products from various renowned brands, as well as its own private label products.
- They provide a wide range of products suitable for diverse skin tones, hair textures, and personal preferences. Sephora also emphasizes clean beauty, offering products that are formulated without certain ingredients, cruelty-free, or vegan.
- Sephora has both physical retail stores and an online platform, making it convenient for customers to explore and purchase products. Sephora also offers beauty services such as makeovers, consultations, and classes to enhance the overall shopping experience
- The store features a diverse range of cosmetics, including foundations, concealers, blushes, eye shadows, lipsticks, and more. Skincare products available at Sephora cater to different skin types and concerns, such as cleansers, moisturizers, serums, masks, and sunscreens.
- Sephora offers Haircare products for various hair types and Fragrances for both men and women are also part of their product offerings, featuring perfumes, colognes, and body sprays.

Problem Statement

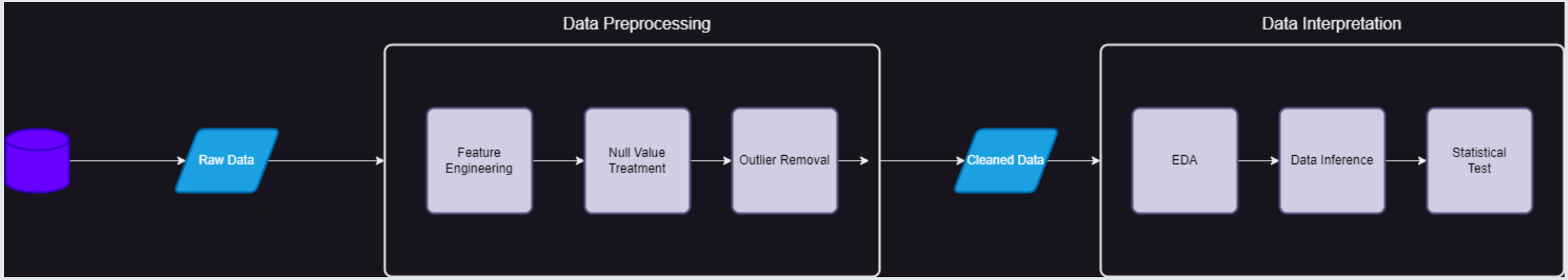
- As we seen they offer various products Sephora Collection, and includes beauty products such as cosmetics, skincare, body, fragrance, nail color, beauty tools, body lotions and haircare
- They often provide detailed product descriptions, customer reviews, and tutorials to help customers make informed decisions.
- As a Data Scientist we need to explore the data, and find popular brand name and what are all their main category of product, subsequently find their total worth of their products in Sephora.
- Find out Sephora exclusive product and most loved products and brand name among customers based on reviews and ratings
- Perform analysis of customer recommendation about products and other features based on skin tone, eye color, skin type

About Datasets

- There are two different datasets provided for our analysis of Sephora products
- One dataset is about Overview of Sephora products, which has 8494 rows and 27 columns
- Other dataset is about detailed review of each author about their product they purchased is around 808855 rows and 19 columns. This dataset is a concatenation of two datasets
- Information about all beauty products from the Sephora online store, including product and brand names, prices, ingredients, ratings, and all features.
- User reviews of all products from the Skincare category, including user appearances, and review ratings by other users

- Kaggle Sephora Products and Skin care reviews

Flow Diagram



Contents

- Data Display
- Feature Engineering
- Null value of Each column and its visualization
- Data Cleaning
  - Imputing Null value using KNN Imputer
  - Imputing Null value using Decision Tree Imputer
- Cleaned Data
- EDA
  - Numeric Analysis
    - Univariate Analysis
    - Bivariate Analysis
  - Outlier Detection and Removal of Outlier
  - Correlation Analysis

Contents

Asking Questions to Data and Getting Insights from Data

Working with Main dataset

1. *Which brand is popular among user and loved ones based on reviews and ratings?*
2. *What is the total worth of price value for Each Brand respective to their Category?*
3. *Which size is most preferred by user based on primary category?*
4. *Which popular brand owns sephora exclusive and does not have physical shop?*
5. *Which new product cost high price and average ratings by user for that product?*
6. *Show visualization of total price shared by primary category*

Working with reviews dataset

1. *Which product is most recommended by user and its skin tone?*
2. *Analyse the Positive feedback count for each brand and other features*
3. *Number of ratings given on Each date*

Display Data

Products Data

```
[5] product_df.head()
```

	product_id	product_name	brand_id	brand_name	loves_count	rating	reviews	size	variation_type	variation_value	...	online_only	out_of_stock	sephora_exclusive	highlights	primary_category	secondary_category	tertiary_category	child_count	child_max_price	child_min_price
0	P473671	Fragrance Discovery Set	6342	19-69	6320	3.6364	11.0	NaN	NaN	NaN	...	1	0	0	['Unisex/ Genderless Scent', 'Warm &Spicy Scen...	Fragrance	Value & Gift Sets	Perfume Gift Sets	0	NaN	NaN
1	P473668	La Habana Eau de Parfum	6342	19-69	3827	4.1538	13.0	3.4 oz/ 100 mL	Size + Concentration + Formulation	3.4 oz/ 100 mL	...	1	0	0	['Unisex/ Genderless Scent', 'Layerable Scent'...	Fragrance	Women	Perfume	2	85.0	30.0
2	P473662	Rainbow Bar Eau de Parfum	6342	19-69	3253	4.2500	16.0	3.4 oz/ 100 mL	Size + Concentration + Formulation	3.4 oz/ 100 mL	...	1	0	0	['Unisex/ Genderless Scent', 'Layerable Scent'...	Fragrance	Women	Perfume	2	75.0	30.0
3	P473660	Kasbah Eau de Parfum	6342	19-69	3018	4.4762	21.0	3.4 oz/ 100 mL	Size + Concentration + Formulation	3.4 oz/ 100 mL	...	1	0	0	['Unisex/ Genderless Scent', 'Layerable Scent'...	Fragrance	Women	Perfume	2	75.0	30.0
4	P473658	Purple Haze Eau de Parfum	6342	19-69	2691	3.2308	13.0	3.4 oz/ 100 mL	Size + Concentration + Formulation	3.4 oz/ 100 mL	...	1	0	0	['Unisex/ Genderless Scent', 'Layerable Scent'...	Fragrance	Women	Perfume	2	75.0	30.0

Display Data

Reviews Data

```
[61] review_df.head(3)
```

	Unnamed: 0	author_id	rating	is_recommended	helpfulness	total_feedback_count	total_neg_feedback_count	total_pos_feedback_count	submission_time	review_text	review_title	skin_tone	eye_color	skin_type	hair_color	product_id	product_name	brand_name	price_us
5	5	42802569154	4	1.0	1.00	1	0	1	2023-03-19	The scent isn't my favourite but it works grea...	Great!	lightMedium	brown	normal	brown	P420652	Lip Sleeping Mask Intense Hydration with Vitam...	LANEIGE	24
6	6	6941883808	2	0.0	0.25	8	6	2	2023-03-19	I'll give this 2 stars for nice packaging and ...	Dried my lips out and clogged my pores	light	blue	combination	brown	P420652	Lip Sleeping Mask Intense Hydration with Vitam...	LANEIGE	24
8	8	7656791726	5	1.0	1.00	1	0	1	2023-03-18	I love this stuff. I first had the sample size...	Must have.	light	blue	normal	blonde	P420652	Lip Sleeping Mask Intense Hydration with Vitam...	LANEIGE	24

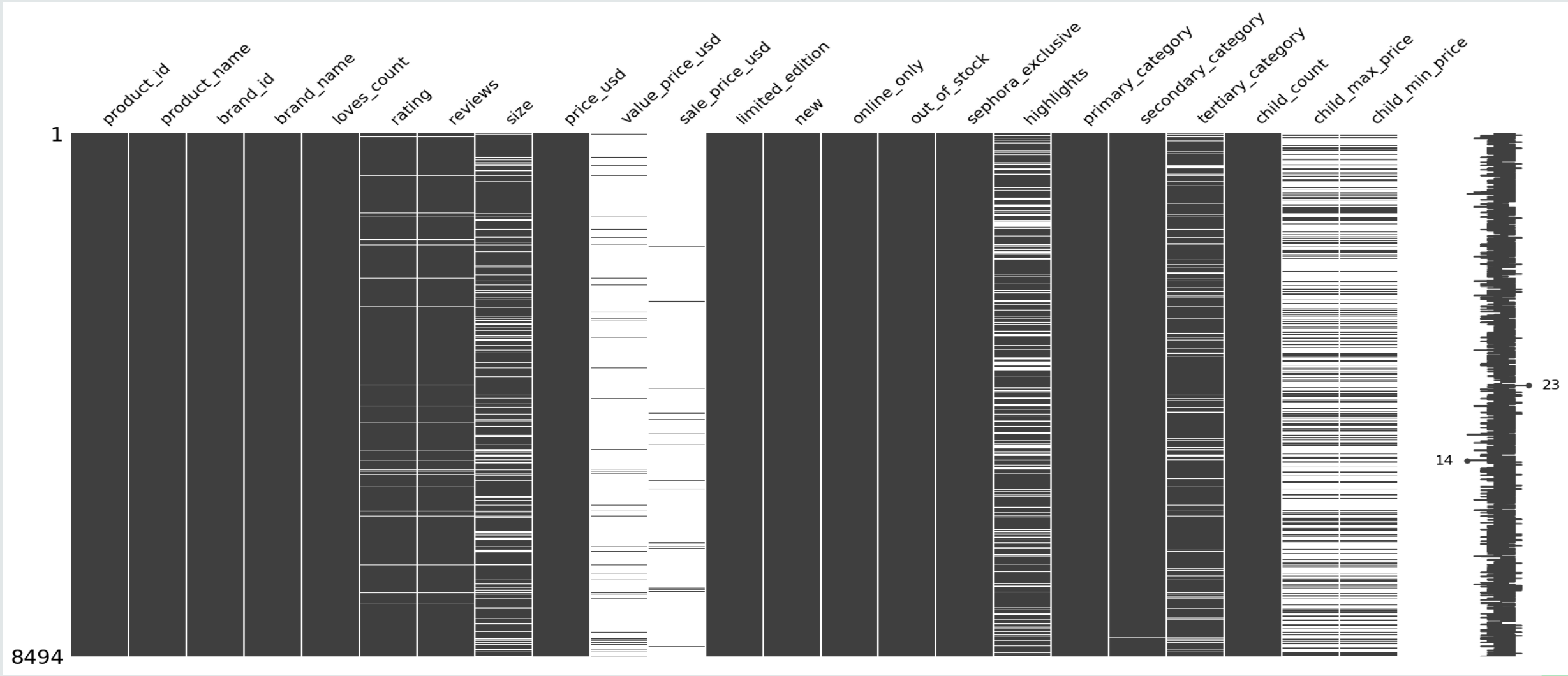
Feature Engineering

- The "**Size**" columns is dirty data which a mixture of ***numeric*** and ***categorical data***.
- Numeric points has **60%** and Categorical points is around **40%** of "**Size**" feature
- So, We performed feature engineering process to convert "**Size**" feature numeric datapoints into bins for categorical points

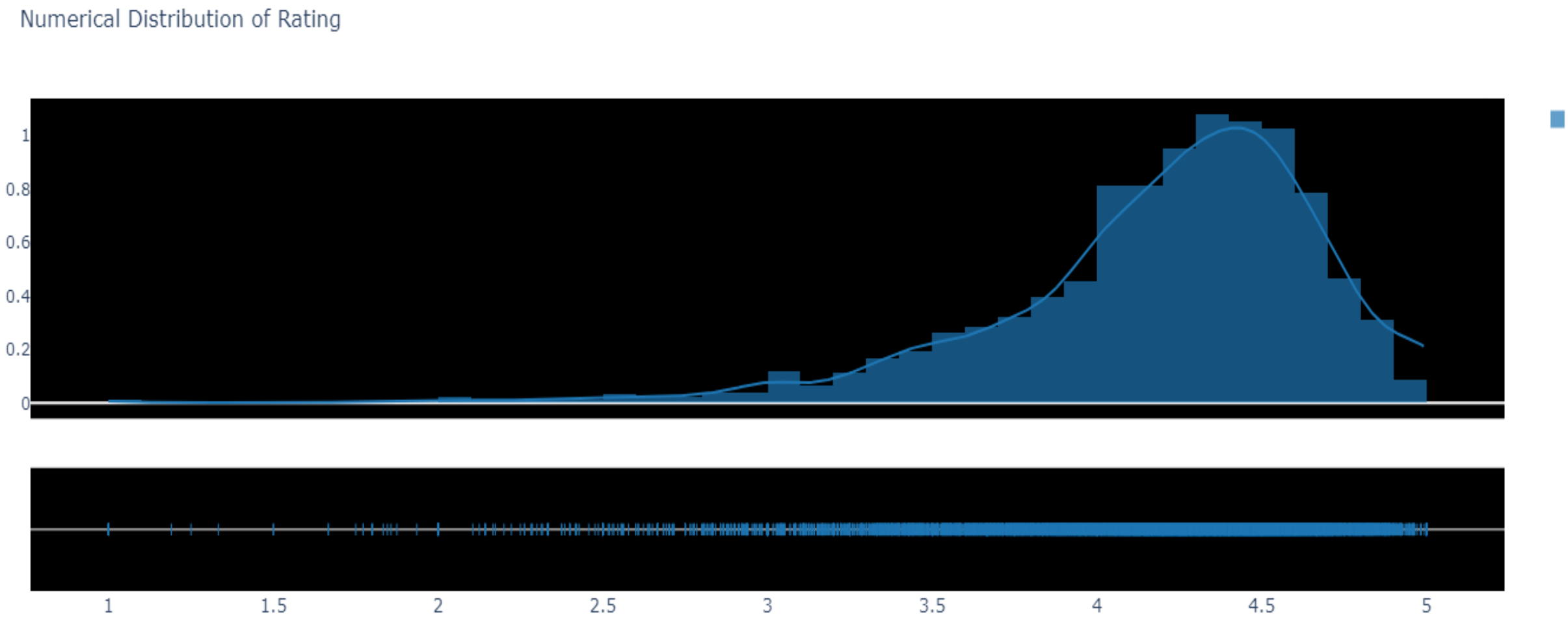
brand_id	brand_name	loves_count	rating	reviews	size	v
6342	19-69	6320	3.6364	11.0	NaN	
6342	19-69	3827	4.1538	13.0	3.4 oz/ 100 mL	
6342	19-69	3253	4.2500	16.0	3.4 oz/ 100 mL	
6342	19-69	3018	4.4762	21.0	3.4 oz/ 100 mL	
6342	19-69	2691	3.2308	13.0	3.4 oz/ 100 mL	

	size	loves_count	rating
0	125-175 ML	6320.0	3.6364
1	75-100 ML	3827.0	4.1538
2	75-100 ML	3253.0	4.2500
3	75-100 ML	3018.0	4.4762
4	75-100 ML	2691.0	3.2308

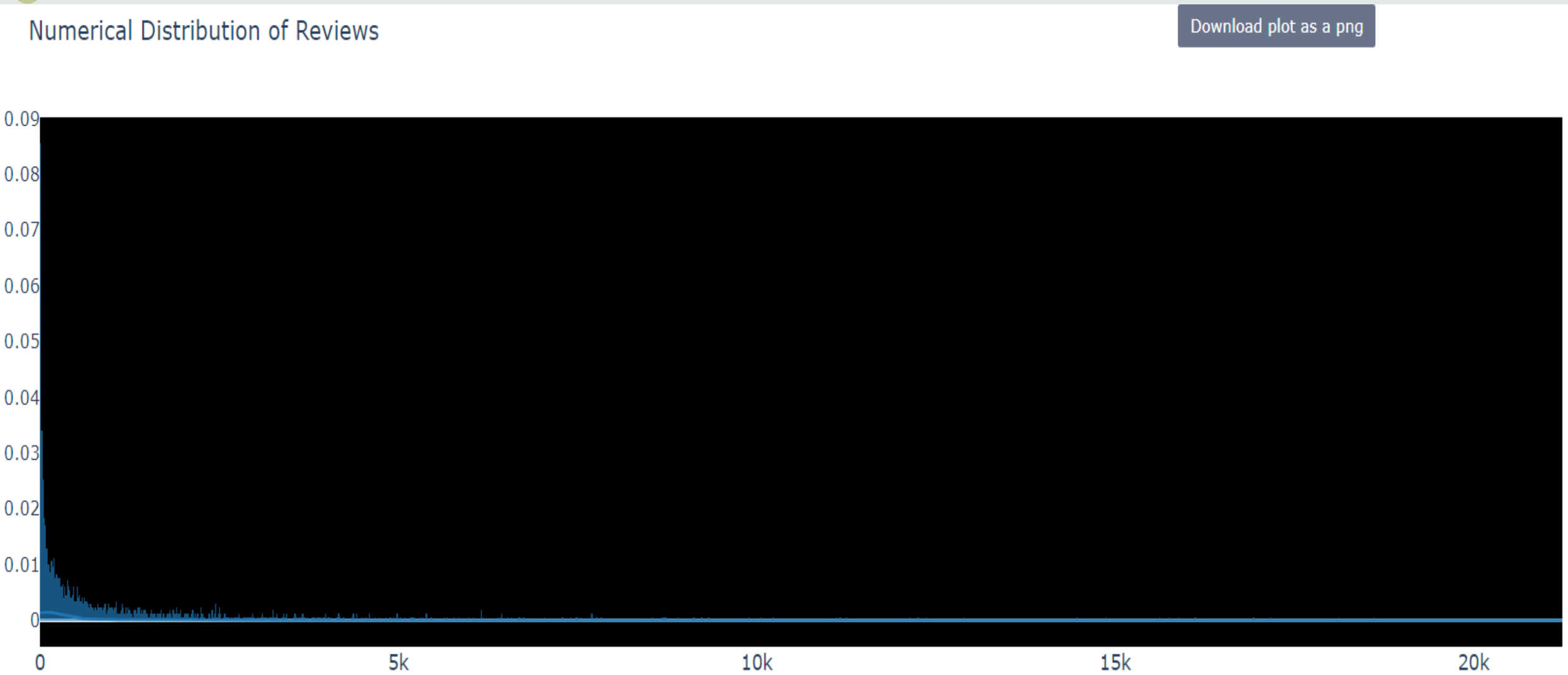
Null value Visualize



Numerical Analysis

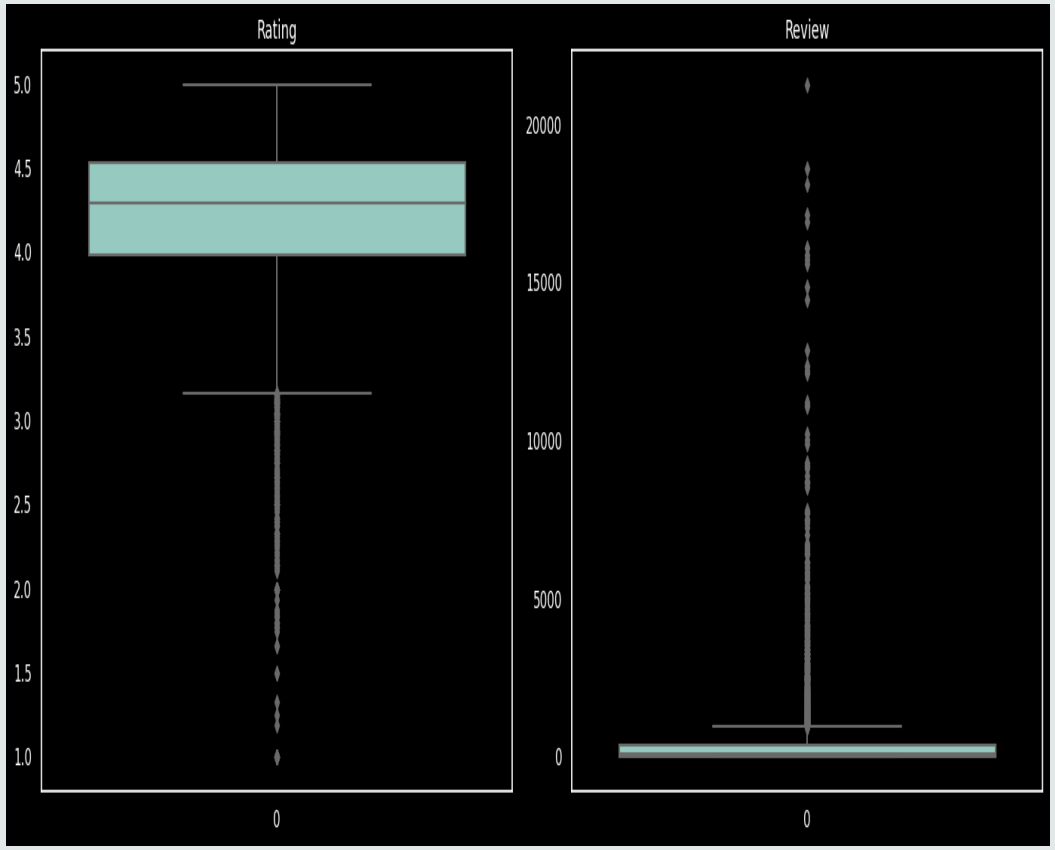


Numerical Analysis

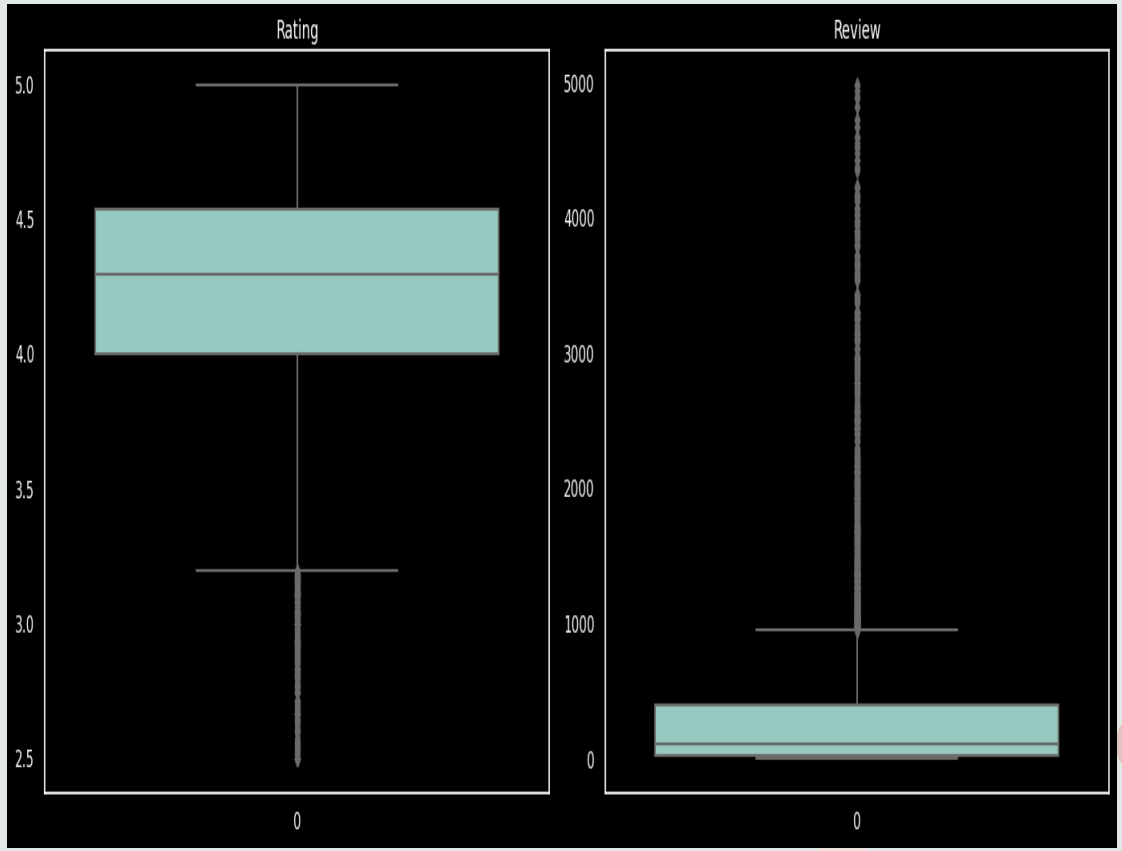


Outlier Detection and Removal

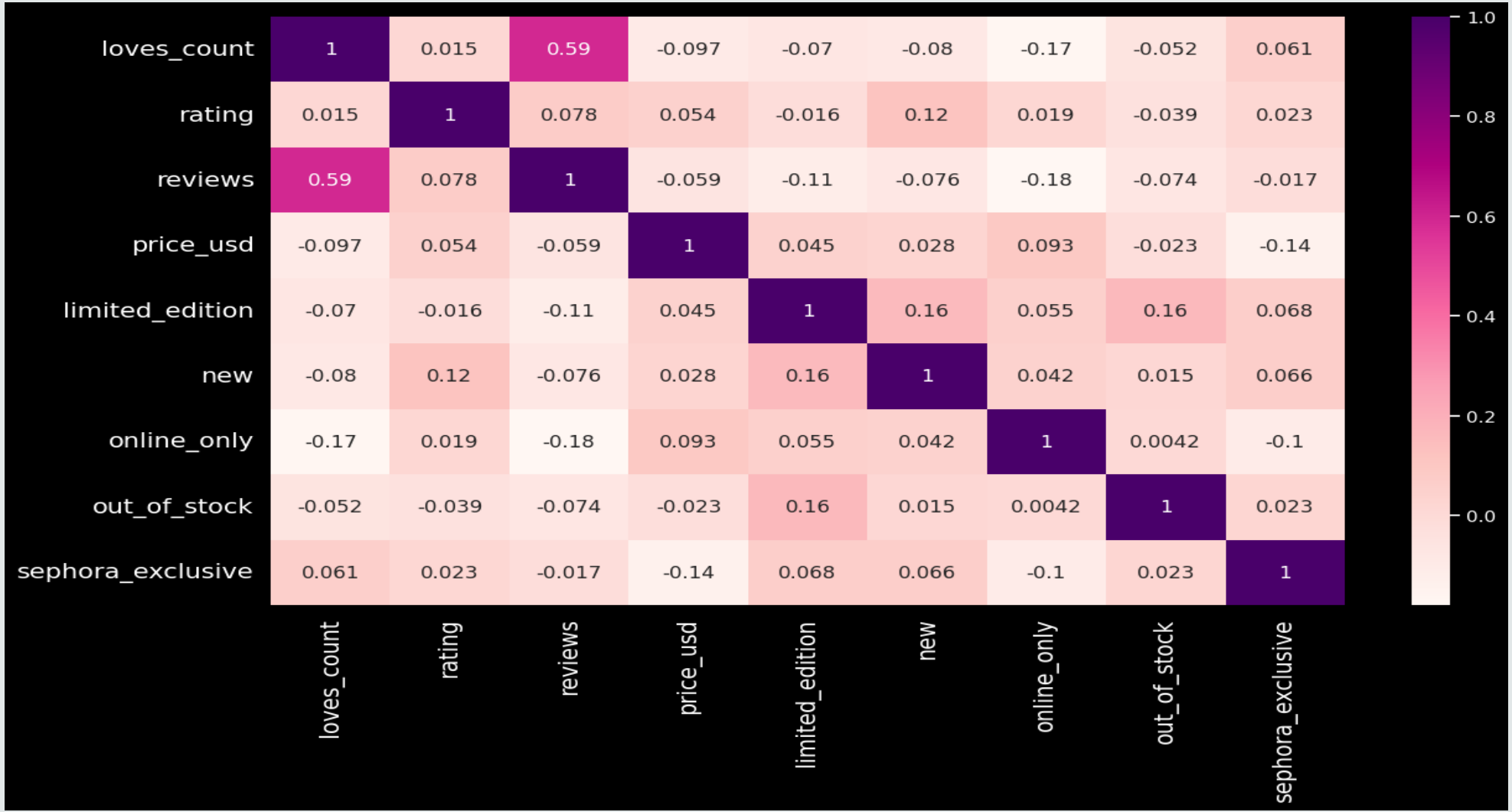
Outlier Detection



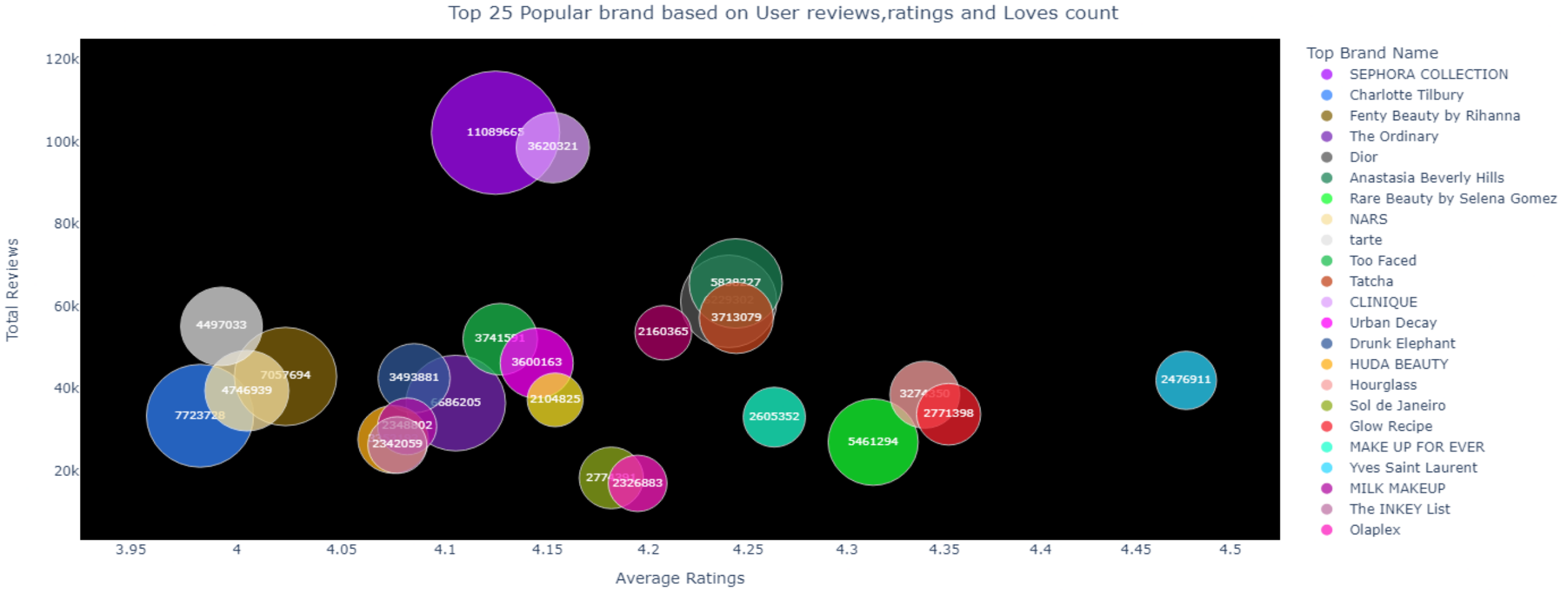
Removal of Outlier



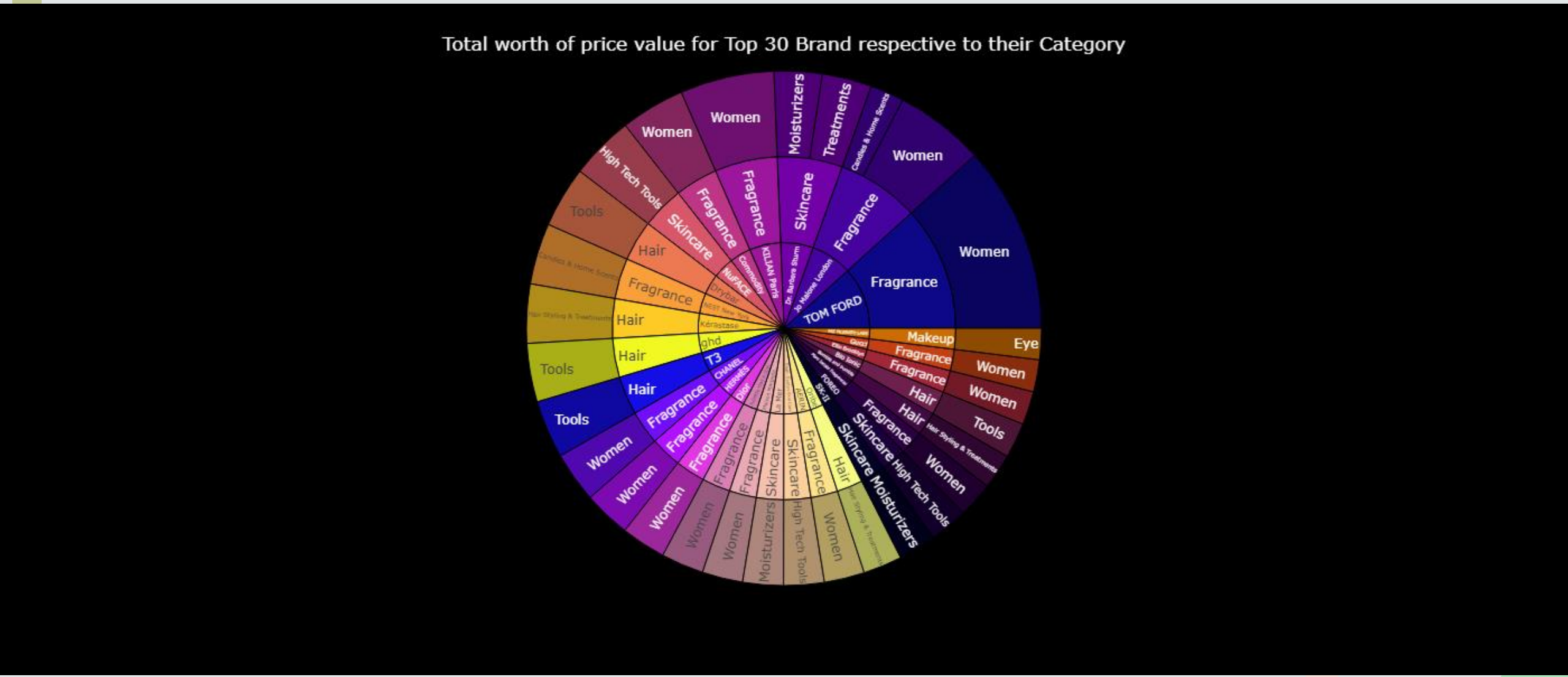
Correlation Analysis



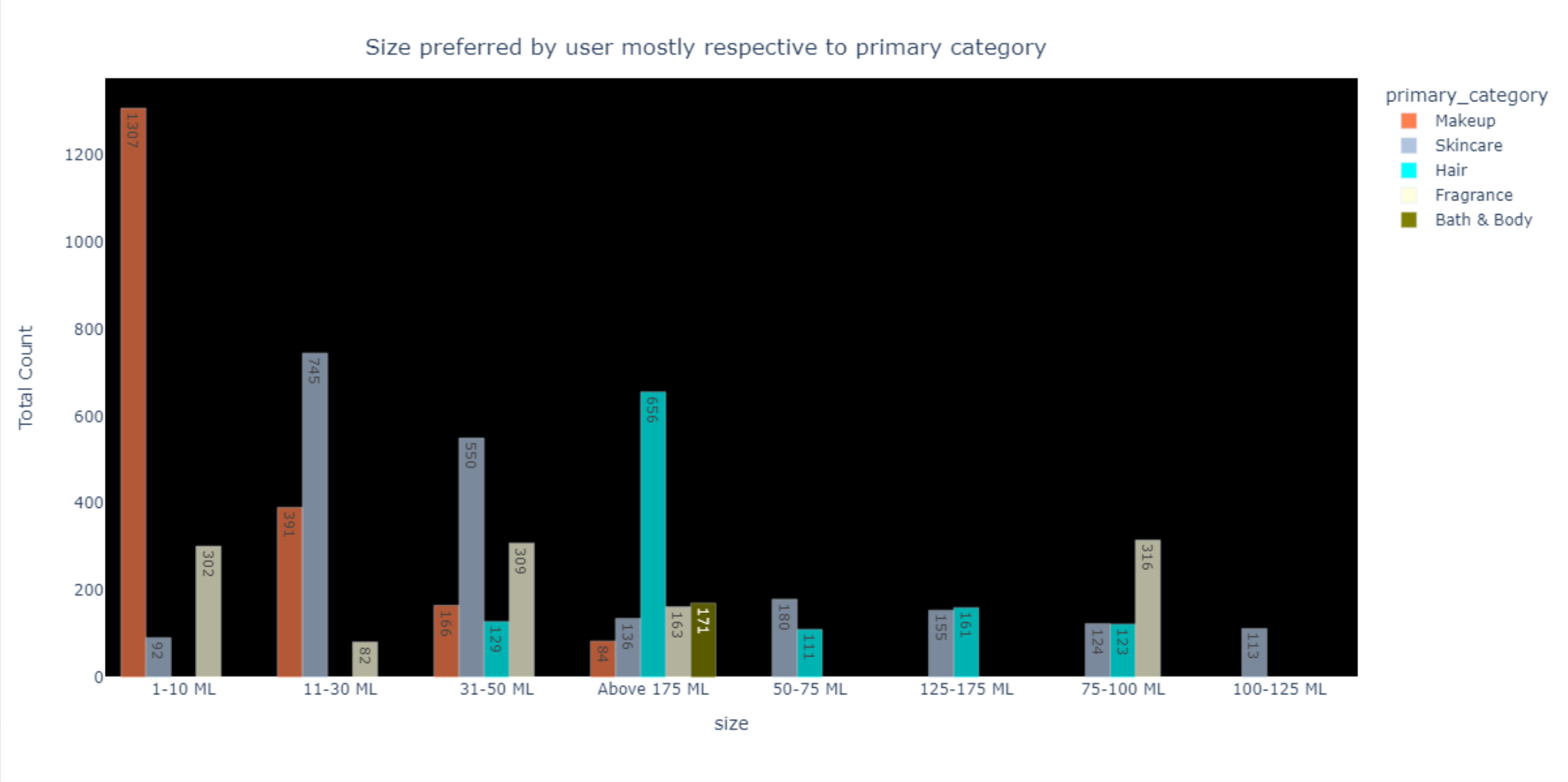
Custom Analysis



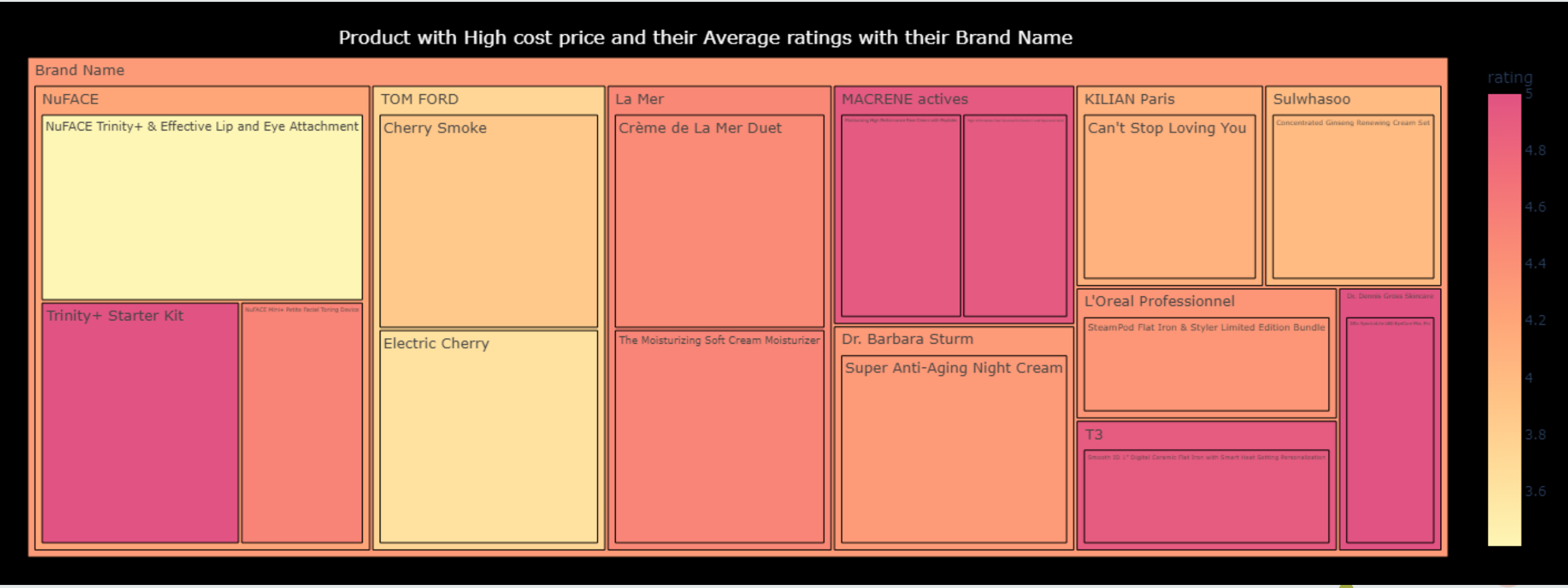
Analysis-2


Total worth of price value for Top 30 Brand respective to their Category
Women
Women
Moisturizers
Treatments
Women
Women
High Tech Tools
Fragrance
Skincare
Fragrance
Tools
Hair
Fragrance
Hair
Tools
Hair
Tools
Women
Women
Women
Fragrance
Fragrance
Fragrance
Fragrance
Skincare
Skincare
Fragrance
Hair
Women
Women
Moisturizers
High Tech Tools
Women
Skincare Moisturizers
Skincare High Tech Tools
Fragrance
Women
Hair
Hair
Fragrance
Fragrance
Makeup
Eye
Women
Women
Tools
Women
Fragrance
TOM FORD
Fragrance
Women
Dr. Barbara Sturm
Jo Malone London
KILIAN Paris
Commodity
NuFACE
Drybar
Kérastase
ghd
T3
CHANEL
HERMÈS
Dior
FOREO
SK-II
Clarins
Bio Ionic
Hair Styling & Treatments

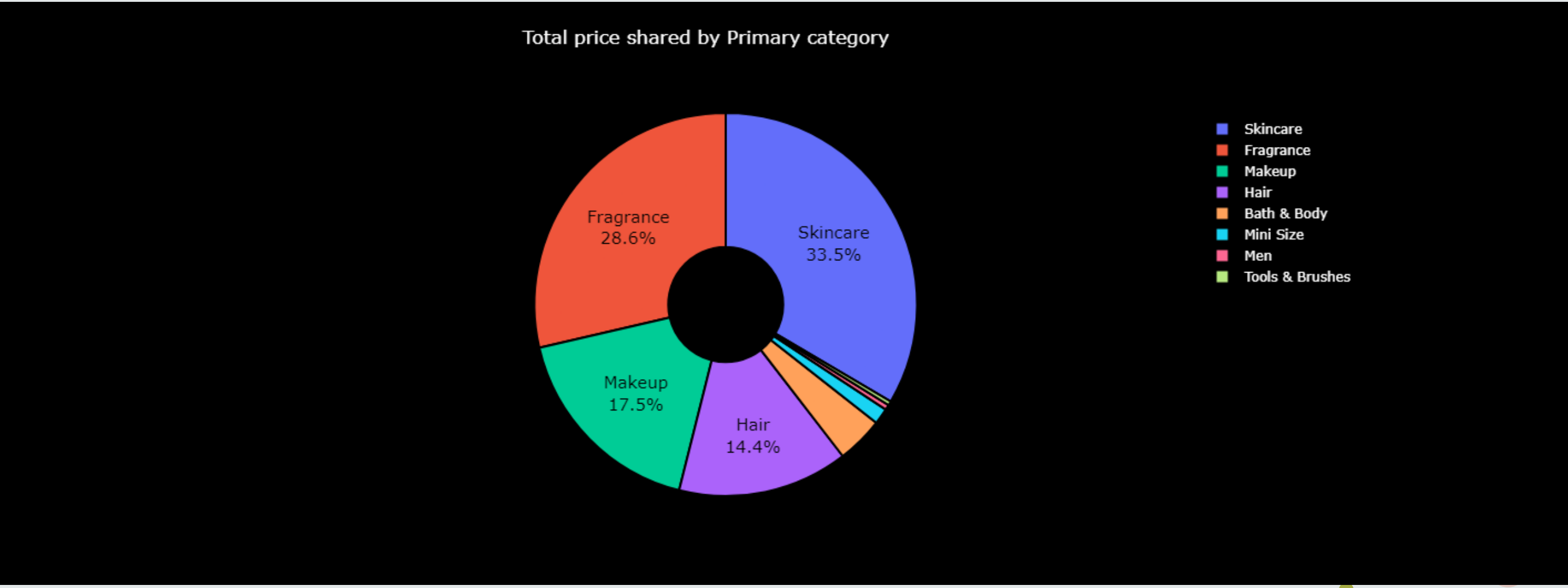
Analysis-3


Size preferred by user mostly respective to primary category
Total Count
size
primary_category
Makeup
Skincare
Hair
Fragrance
Bath & Body
0
200
400
600
800
1000
1200
1-10 ML
11-30 ML
31-50 ML
Above 175 ML
50-75 ML
125-175 ML
75-100 ML
100-125 ML
1307
92
302
391
745
82
166
550
129
309
84
136
656
163
171
180
111
155
161
124
123
316
113

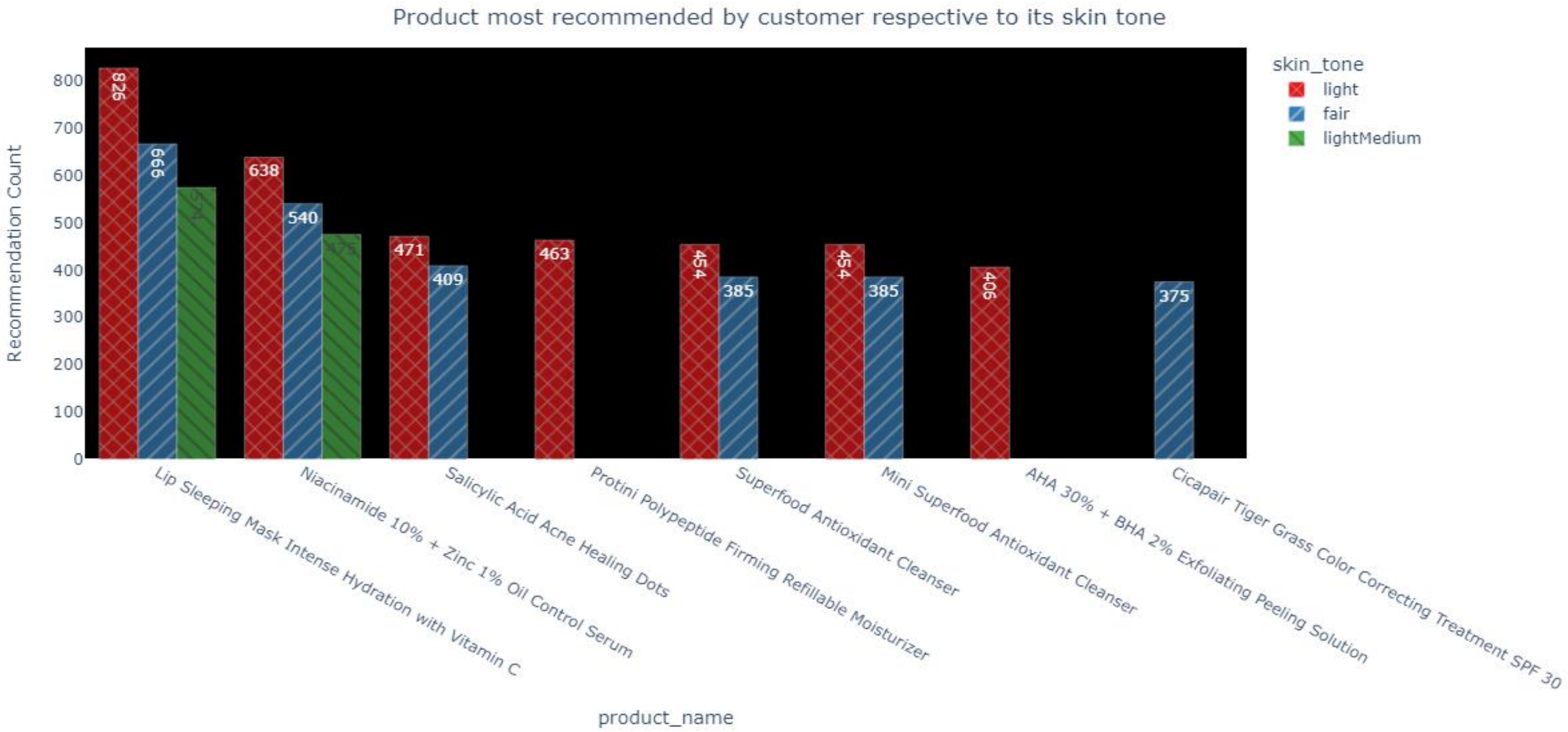
Analysis-4


Product with High cost price and their Average ratings with their Brand Name
Brand Name
NuFACE
NuFACE Trinity+ & Effective Lip and Eye Attachment
Trinity+ Starter Kit
TOM FORD
Cherry Smoke
Electric Cherry
La Mer
Crème de La Mer Duet
The Moisturizing Soft Cream Moisturizer
MACRENE actives
Dr. Barbara Sturm
Super Anti-Aging Night Cream
KILIAN Paris
Can't Stop Loving You
Sulwhasoo
Concentrated Ginseng Renewing Cream Set
L'Oreal Professionnel
SteamPod Flat Iron & Styler Limited Edition Bundle
T3
Dr. Dennis Gross Skincare
rating
5
4.8
4.6
4.4
4.2
4
3.8
3.6

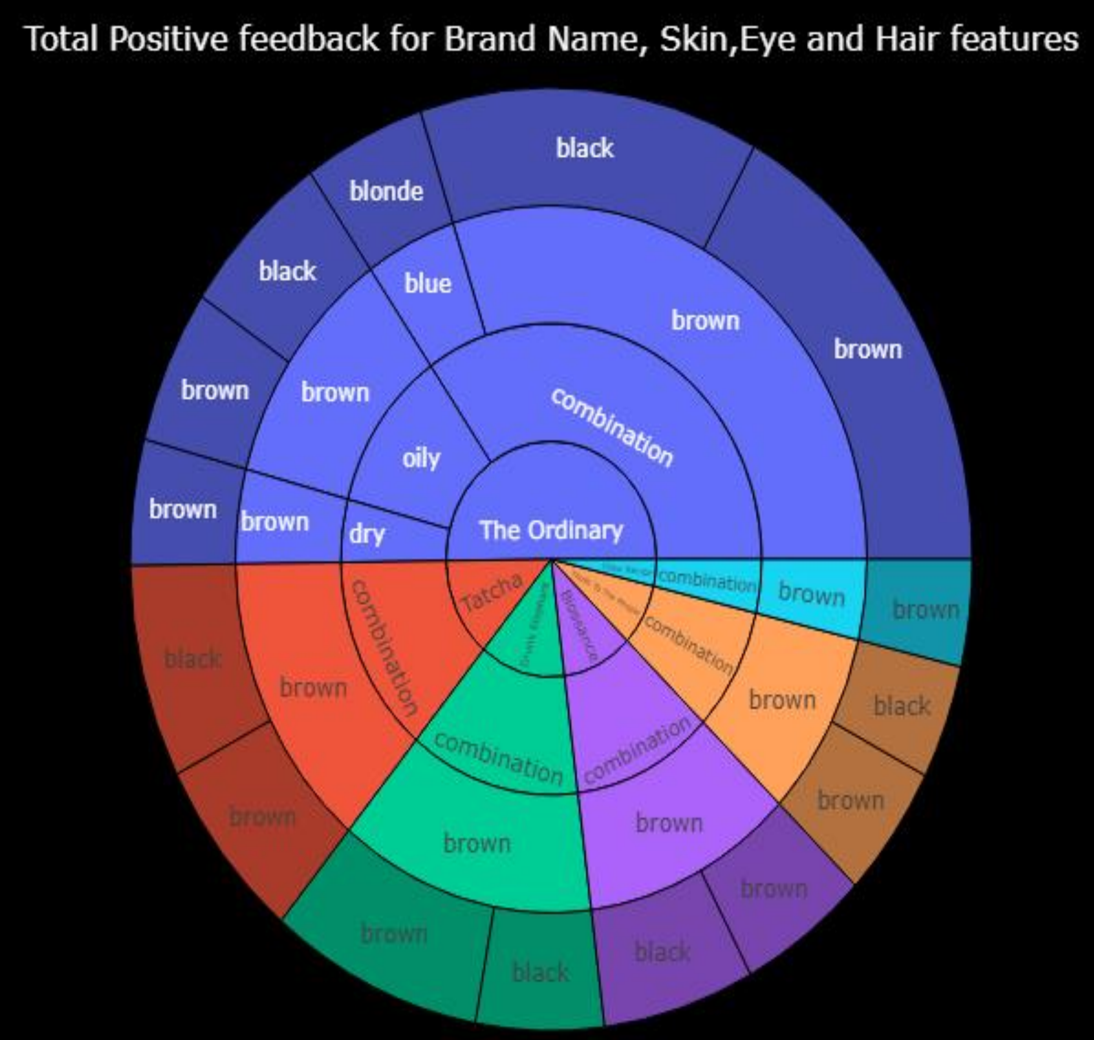
Analysis-5


Total price shared by Primary category
Fragrance
28.6%
Skincare
33.5%
Makeup
17.5%
Hair
14.4%
Skincare
Fragrance
Makeup
Hair
Bath & Body
Mini Size
Men
Tools & Brushes

Analysis-6

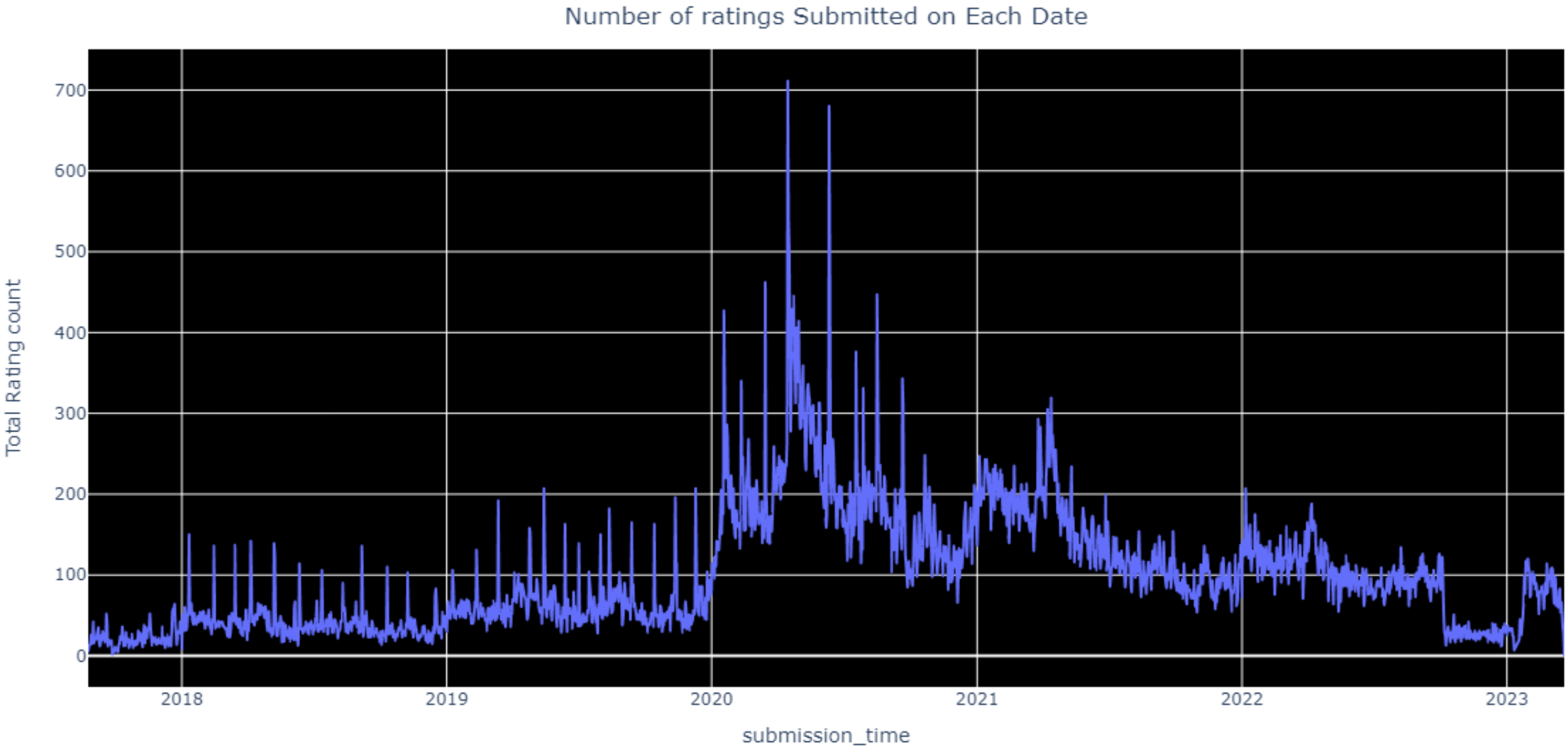

Product most recommended by customer respective to its skin tone
Recommendation Count
0
100
200
300
400
500
600
700
800
826
666
638
540
471
409
463
454
385
454
385
406
375
skin_tone
light
fair
lightMedium
Lip Sleeping Mask Intense Hydration with Vitamin C
Niacinamide 10% + Zinc 1% Oil Control Serum
Salicylic Acid Acne Healing Dots
Protini Polypeptide Firming Refillable Moisturizer
Superfood Antioxidant Cleanser
Mini Superfood Antioxidant Cleanser
AHA 30% + BHA 2% Exfoliating Peeling Solution
Cicapair Tiger Grass Color Correcting Treatment SPF 30
product_name

Analysis-7

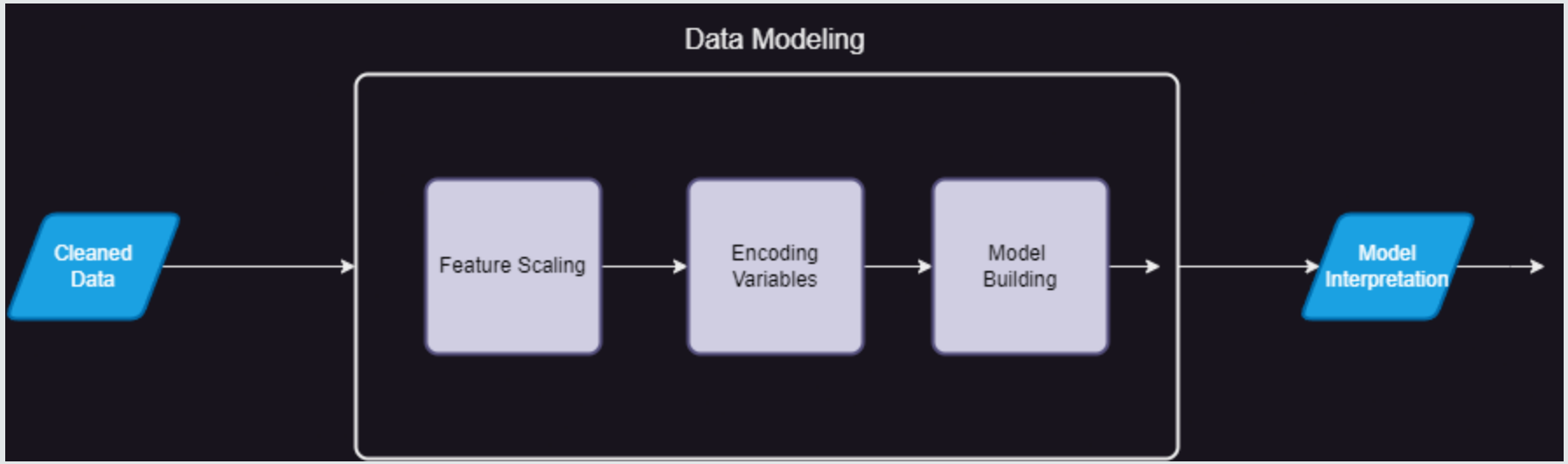


Total Positive feedback for Brand Name, Skin,Eye and Hair features
The Ordinary
combination
oily
dry
blue
brown
black
blonde
Tatcha
combination
brown
black

Analysis-8


Number of ratings Submitted on Each Date
Total Rating count
0
100
200
300
400
500
600
700
2018
2019
2020
2021
2022
2023
submission_time

Data Modeling



- We build Model to Predict a Price of Product
- We Further perform Feature Scaling using Standard Scalar Method
- We done Encoding using One Hot Encoding
- We build a Model using Linear Regression, Lasso Regression, Decision Tree Regression, XGBOOST Regression
- Once Model is built, we performed data interpretation from model to check its accuracy of model

Linear Regression Modeling

```
linear_model = LinearRegression()
linear_model.fit(X_train, y_train)
```

```
LinearRegression
LinearRegression()
```

```
# Make predictions on the training set
y_train_pred = linear_model.predict(X_train)

# Make predictions on the test set
y_test_pred = linear_model.predict(X_test)

# Calculate mean squared error (MSE) for training and test data
mse_train = mean_squared_error(y_train, y_train_pred)
mse_test = mean_squared_error(y_test, y_test_pred)

# Calculate R-squared (coefficient of determination) for training and test data
r2_train = r2_score(y_train, y_train_pred)
r2_test = r2_score(y_test, y_test_pred)

print("MSE - Training Data:", mse_train)
print("MSE - Test Data:", mse_test)
print("R-squared - Training Data:", r2_train)
print("R-squared - Test Data:", r2_test)
```

```
MSE - Training Data: 2117.707814566357
MSE - Test Data: 1741.2426426634333
R-squared - Training Data: 0.284180407548297
R-squared - Test Data: 0.3242229016496214
```

Lasso Regression Modeling

```
[ ] lasso_model = Lasso(alpha=0.1)
    lasso_model.fit(X_train, y_train)
```

```
Lasso
Lasso(alpha=0.1)
```

```
[ ] # Make predictions on the training set
    y_train_pred = lasso_model.predict(X_train)

    # Make predictions on the test set
    y_test_pred = lasso_model.predict(X_test)

    # Calculate mean squared error (MSE) for training and test data
    mse_train = mean_squared_error(y_train, y_train_pred)
    mse_test = mean_squared_error(y_test, y_test_pred)

    # Calculate R-squared (coefficient of determination) for training and test data
    r2_train = r2_score(y_train, y_train_pred)
    r2_test = r2_score(y_test, y_test_pred)

    print("MSE - Training Data:", mse_train)
    print("MSE - Test Data:", mse_test)
    print("R-squared - Training Data:", r2_train)
    print("R-squared - Test Data:", r2_test)
```

```
MSE - Training Data: 2118.731451131079
MSE - Test Data: 1739.0521579582098
R-squared - Training Data: 0.28383440178511377
R-squared - Test Data: 0.32507302980626507
```

Decision Tree Regression Modeling

```
[ ] # Create and train the Decision Tree regression model
    tree_model = DecisionTreeRegressor()
    tree_model.fit(X_train, y_train)
```

```
▾ DecisionTreeRegressor
DecisionTreeRegressor()
```

```
[ ] # Make predictions on the training set
    y_train_pred = tree_model.predict(X_train)

    # Make predictions on the test set
    y_test_pred = tree_model.predict(X_test)

    # Calculate mean squared error (MSE) for training and test data
    mse_train = mean_squared_error(y_train, y_train_pred)
    mse_test = mean_squared_error(y_test, y_test_pred)

    # Calculate R-squared (coefficient of determination) for training and test data
    r2_train = r2_score(y_train, y_train_pred)
    r2_test = r2_score(y_test, y_test_pred)

    print("MSE - Training Data:", mse_train)
    print("MSE - Test Data:", mse_test)
    print("R-squared - Training Data:", r2_train)
    print("R-squared - Test Data:", r2_test)
```

```
MSE - Training Data: 7.365939893930465e-05
MSE - Test Data: 3822.718686042403
R-squared - Training Data: 0.9999999751019283
R-squared - Test Data: -0.48359894145025084
```

XGBOOST Regression Modeling

```
[ ]  # Create and train the XGBoost regression model
     xg_boost_model = XGBRegressor()
     xg_boost_model.fit(X_train, y_train)
```

```
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=None, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=None, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             n_estimators=100, n_jobs=None, num_parallel_tree=None,
             predictor=None, random_state=None, ...)
```

```
[ ]  # Make predictions on the training set
     y_train_pred = xg_boost_model.predict(X_train)

     # Make predictions on the test set
     y_test_pred = xg_boost_model.predict(X_test)

     # Calculate mean squared error (MSE) for training and test data
     mse_train = mean_squared_error(y_train, y_train_pred)
     mse_test = mean_squared_error(y_test, y_test_pred)

     # Calculate R-squared (coefficient of determination) for training and test data
     r2_train = r2_score(y_train, y_train_pred)
     r2_test = r2_score(y_test, y_test_pred)

     print("MSE - Training Data:", mse_train)
     print("MSE - Test Data:", mse_test)
     print("R-squared - Training Data:", r2_train)
     print("R-squared - Test Data:", r2_test)
```

```
MSE - Training Data: 341.30385509068924
MSE - Test Data: 2377.797073953913
R-squared - Training Data: 0.8846337607233885
R-squared - Test Data: 0.07717582390199496
```

Statistical Test

We perform Statistical Test to infer from the data and get the interpretation from it

We used Independent t-test, Chi Squared test, One way Annova, Pearson Correlation Coefficient Test

Independent t-test:

The independent t-test **compares the means** of two **independent groups** to determine if there is a significant difference between them

Group1= The ratings of Skincare from Primary category

Group2= The ratings of Makeup from Primary category

Inference Results:

T-statistic: 5.913358400426621

p-value: 3.5848167759449255e-09

- There is a significant difference in the mean 'rating' between **'Skincare'** and **'Makeup'**

Chi-squared test:

The chi-squared test is used to determine if there is a significant **association between two categorical variables**

Category var1=Primary category

Category var2=Size

Inference Results:

Chi-squared statistic: 9903.90445683496

p-value: 0.0

- There is a significant difference in the mean 'rating' between **'Size'** and **'Primary category'**

One-way Annova:

One-Way ANOVA is used to compare the **means of two or more groups** to determine if there are any significant differences

Group1= The ratings of Women from Secondary category

Group2= The ratings of Eye from Secondary category

Group3= The ratings of Face from Secondary category

Inference Results:

F-statistic: 29.65235576069065

p-value: 1.9476949111852175e-13

- There is a significant difference in the mean 'rating' among the **'Women', 'Eye', and 'Face'** categories.

Pearson Correlation Coefficient:

The Pearson correlation coefficient measures the **linear relationship between two continuous variables**. It assesses how strongly they are linearly correlated

Numerical var1=Loves Count

Numerical var2=Reviews

Inference Results:

Pearson Correlation Coefficient: 0.6832004791644654

p-value: 0.0

- There is a strong linear relationship between **'loves_count'** and **'reviews'**

Thankyou