ANALYSIS OF

CONSUMER PREFERENCES AND TRENDS FOR FOUNDATION AT SEPHORA

ASHLEY DOAN

SELENE NGUYEN

DATA RETRIVAL

01.

Inspect the network activity on sephora’s foundation product pages and identify the json file that contains product information.

02.

Retrieve data from all 5 pages of foundation products using for loop

03.

Convert the data into a dataframe with 11 variables and 262 observations.

	Product_Name	Brand_Name	Price	Num_Colors	Rating	Reviews	isLimitedEdition	isSephoraExclusive	isNatural	isOrganic	isSponsored
0	Luminous Silk Perfect Glow Flawless Oil-Free F...	Armani Beauty	48.00 – 69.00	39	4.2126	4713	False	False	False	False	False
1	Triclone Skin Tech Medium Coverage Foundation ...	HAUS LABS BY LADY GAGA	$49.00	50	4.0823	3561	False	True	False	False	False
2	Easy Blur Natural Airbrush Foundation with Nia...	HUDA BEAUTY	$37.00	28	4.5892	2685	False	True	False	False	False
3	Pro Filt'r Soft Matte Longwear Liquid Foundation	Fenty Beauty by Rihanna	18.00 – 40.00	50	4.0205	17409	False	False	False	False	False

Next Slide →

DATA CLEANING

FILTER BY PRODUCT TYPE

Retain only products with "foundation" in the name; exclude products containing terms like "brush," "sample," or "pump."

CONVERT BOOLEAN VARIABLES

Change boolean values (true, false) in columns such as isLimitedEdition, isSephoraExclusive, isNatural, isOrganic, and isSponsored to binary format (1 for true, 0 for false).

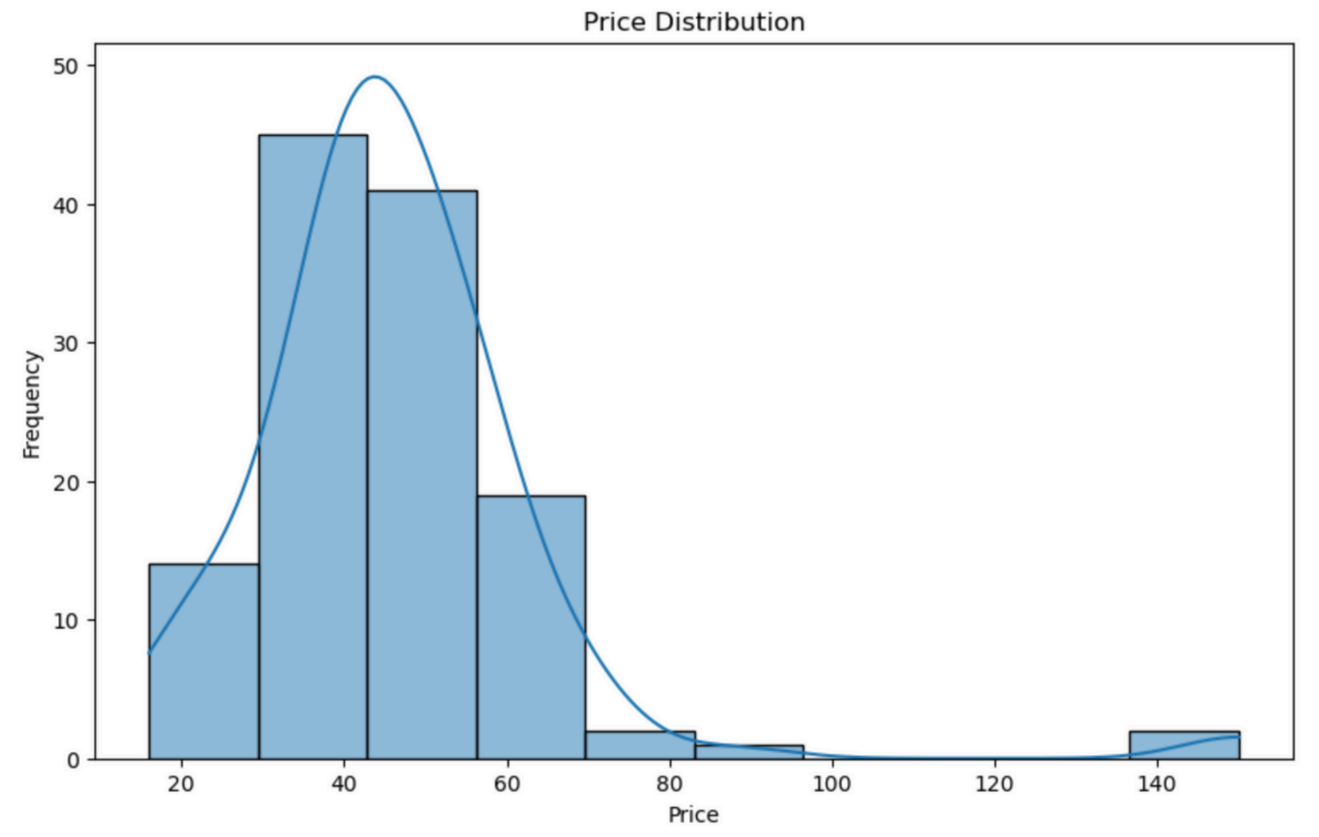
HANDLE PRICE RANGES

For products with price ranges, select the higher price value as the representative price for full-size products.

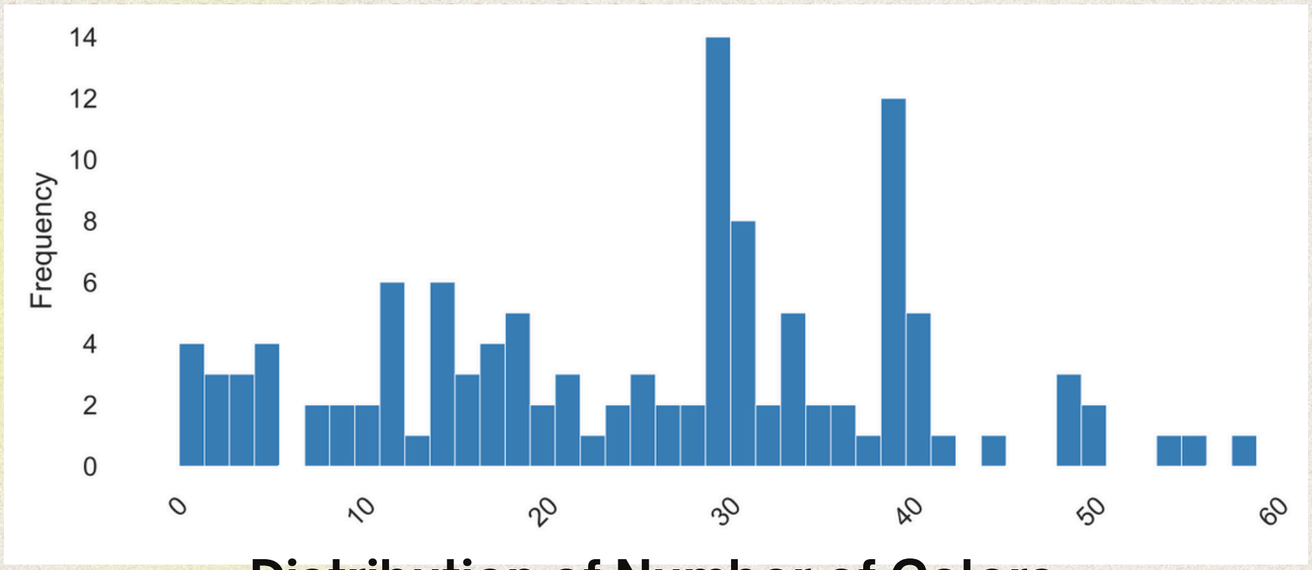
CONVERT DATA TYPES

Change Price, Num_Colors, Rating, and Reviews to numerical data types.

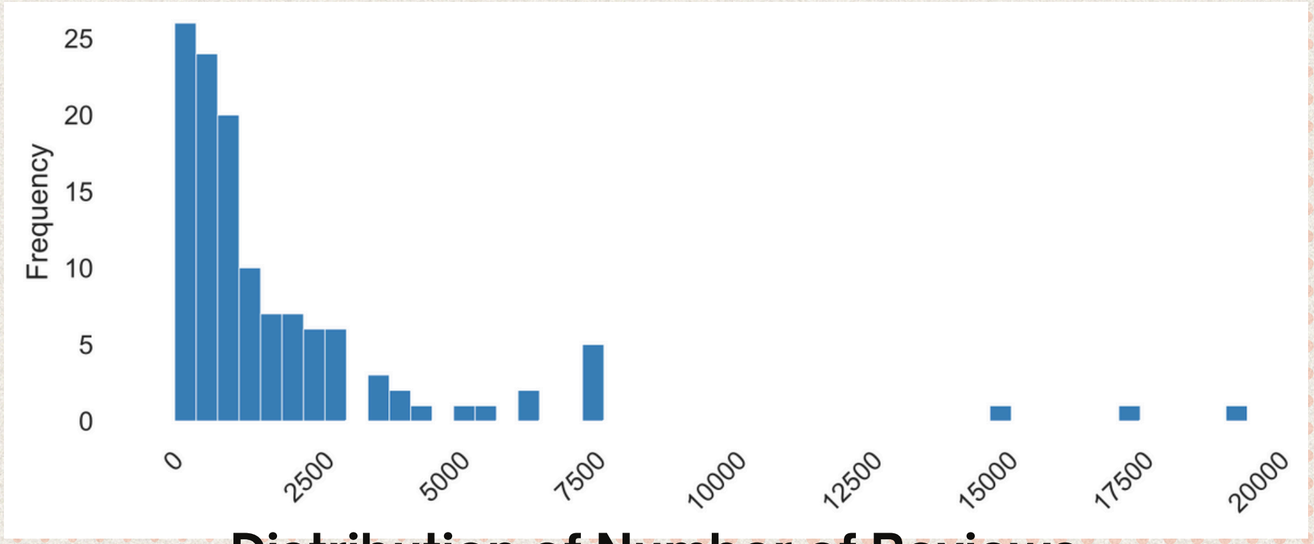
	Product_Name	Brand_Name	Price	Num_Colors	Rating	Reviews	isLimitedEdition	isSephoraExclusive	isNatural	isOrganic	isSponsored
0	Pro Filt’r Soft Matte Longwear Liquid Foundation	Fenty Beauty by Rihanna	40.0	50	4.0206	17410	0	0	0	0	0
1	Soft’Lit Naturally Luminous Hydrating Longwear...	Fenty Beauty by Rihanna	40.0	49	4.515	1629	0	0	0	0	0
2	Mini Pro Filt’r Soft Matte Longwear Liquid Fou...	Fenty Beauty by Rihanna	18.0	2	3.6018	113	0	0	0	0	0
3	Triclone Skin Tech Medium Coverage Foundation ...	HAUS LABS BY LADY GAGA	49.0	50	4.0832	3568	0	1	0	0	0



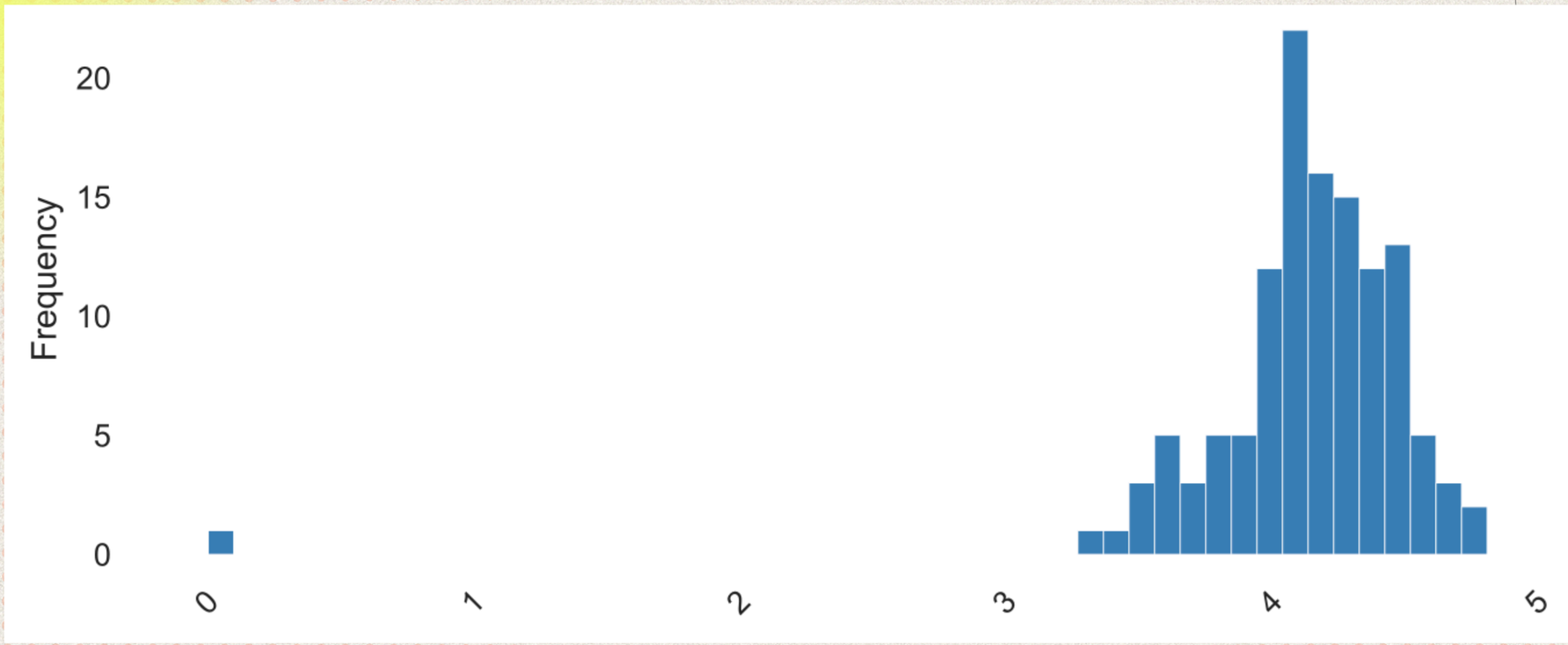
EXPLORATORY DATA ANALYSIS



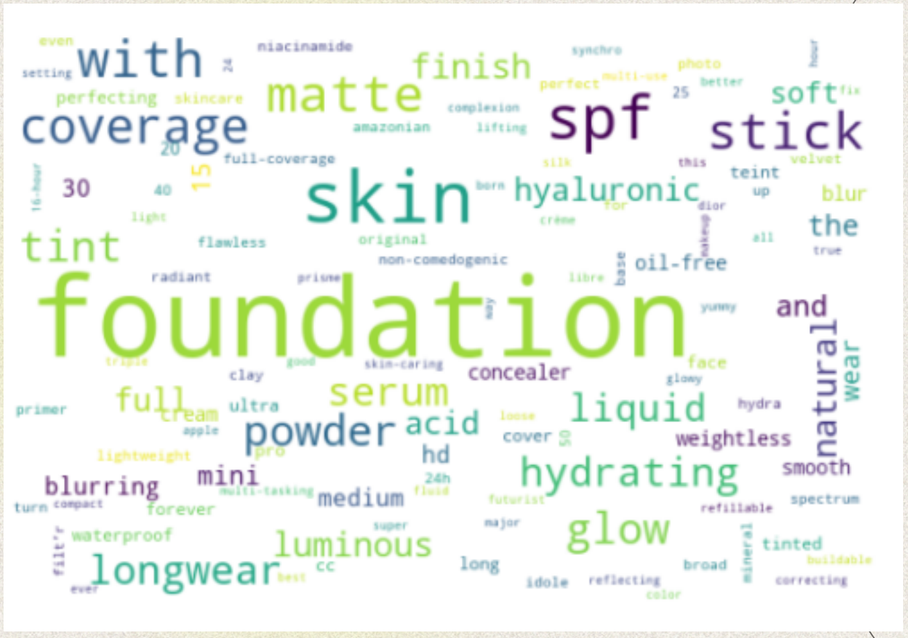
Distribution of Number of Colors



Distribution of Number of Reviews



Distribution of Ratings



Product Name

Brand Name

TOP 10 BRANDS WITH MOST PRODUCTS

MAKE UP FOR EVER	7
Fenty Beauty by Rihanna	6
DIOR	5
bareMinerals	4
Lancôme	4
CLINIQUE	4
Urban Decay	4
IT Cosmetics	4
SEPHORA COLLECTION	3
Shiseido	3

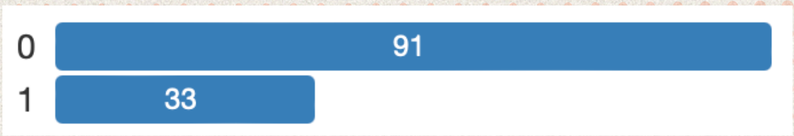
OTHER VARIABLES

01.

isLimitedEdition

02.

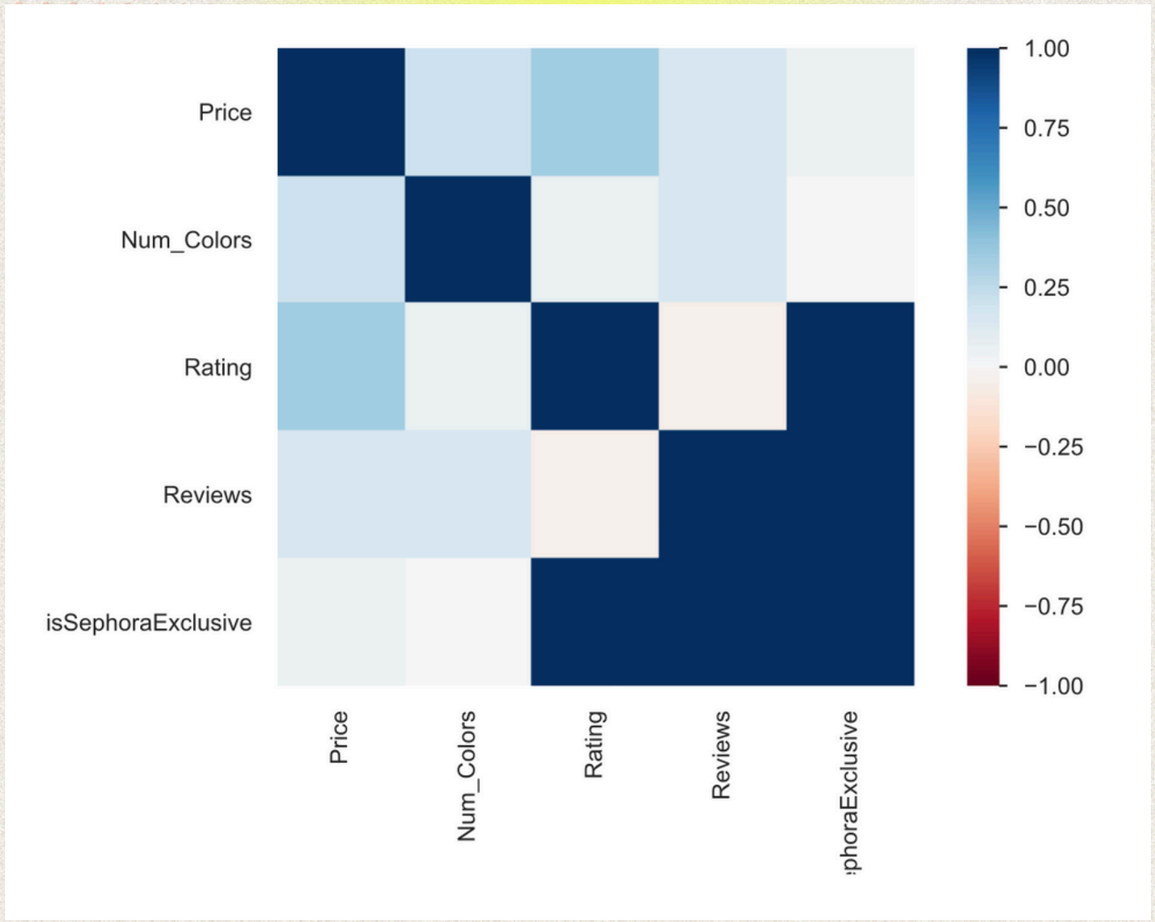
isSephoraExclusive



03.

isNatural, isOrganic, & isSponsored

HEATMAP


Price
Num_Colors
Rating
Reviews
isSephoraExclusive
Price
Num_Colors
Rating
Reviews
sphoraExclusive
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Data preparation

01.

Drop outliers in the target variable Price.
(just dropped values >100 as opposed to using IQR method)

02.

Drop Variables with only one distinct value (isNatural, isOrganic, isSponsored)

03.

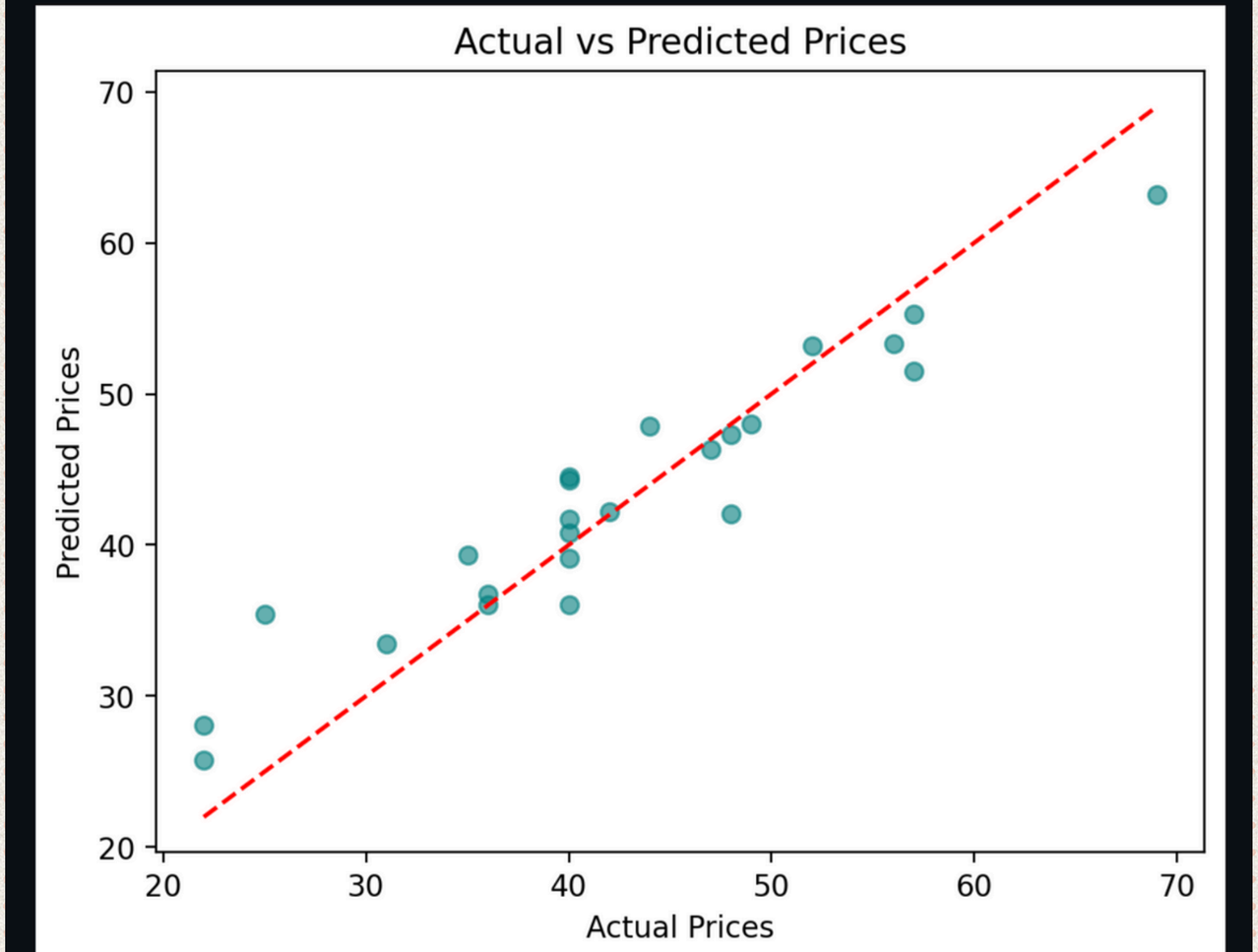
Perform one-hot encoding on categorical variables (Brand_Name, Product_Name)
(convert into numerical format)

Model Performance

RMSE: 3.93

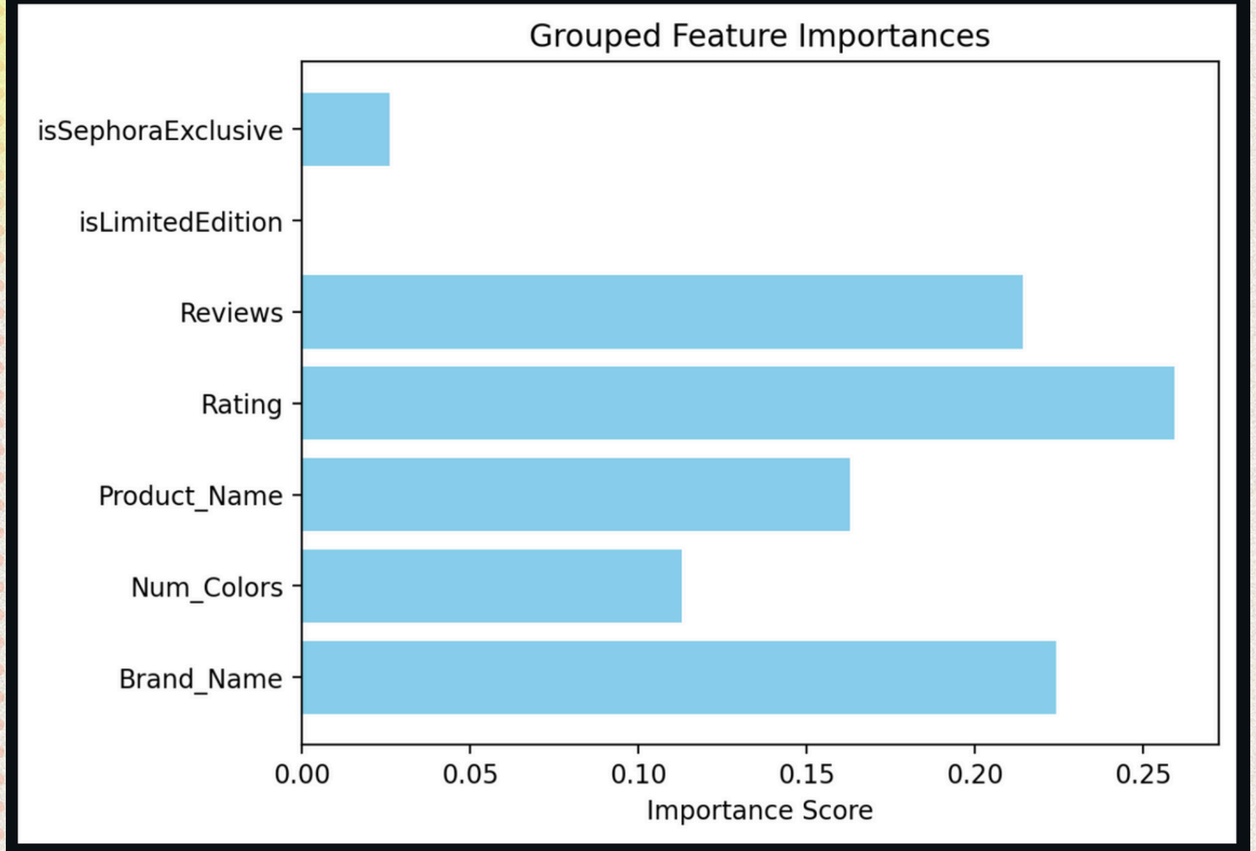
R-squared: 0.88

Visualization of Predictions vs Actual Values



Price Prediction Model

Feature Importance



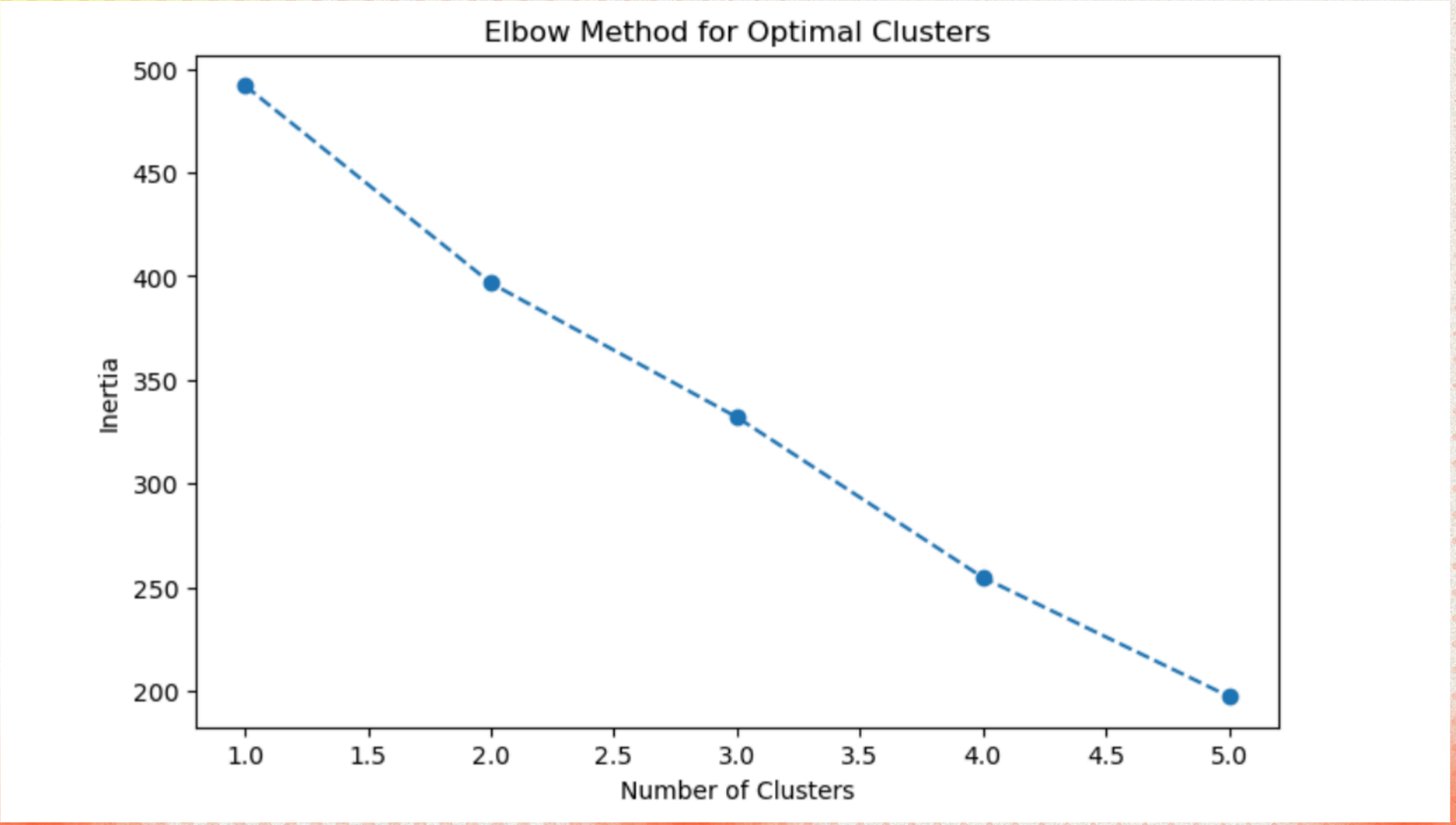
Most important features:

- Rating
- Brand_Name
- Review

Implications:

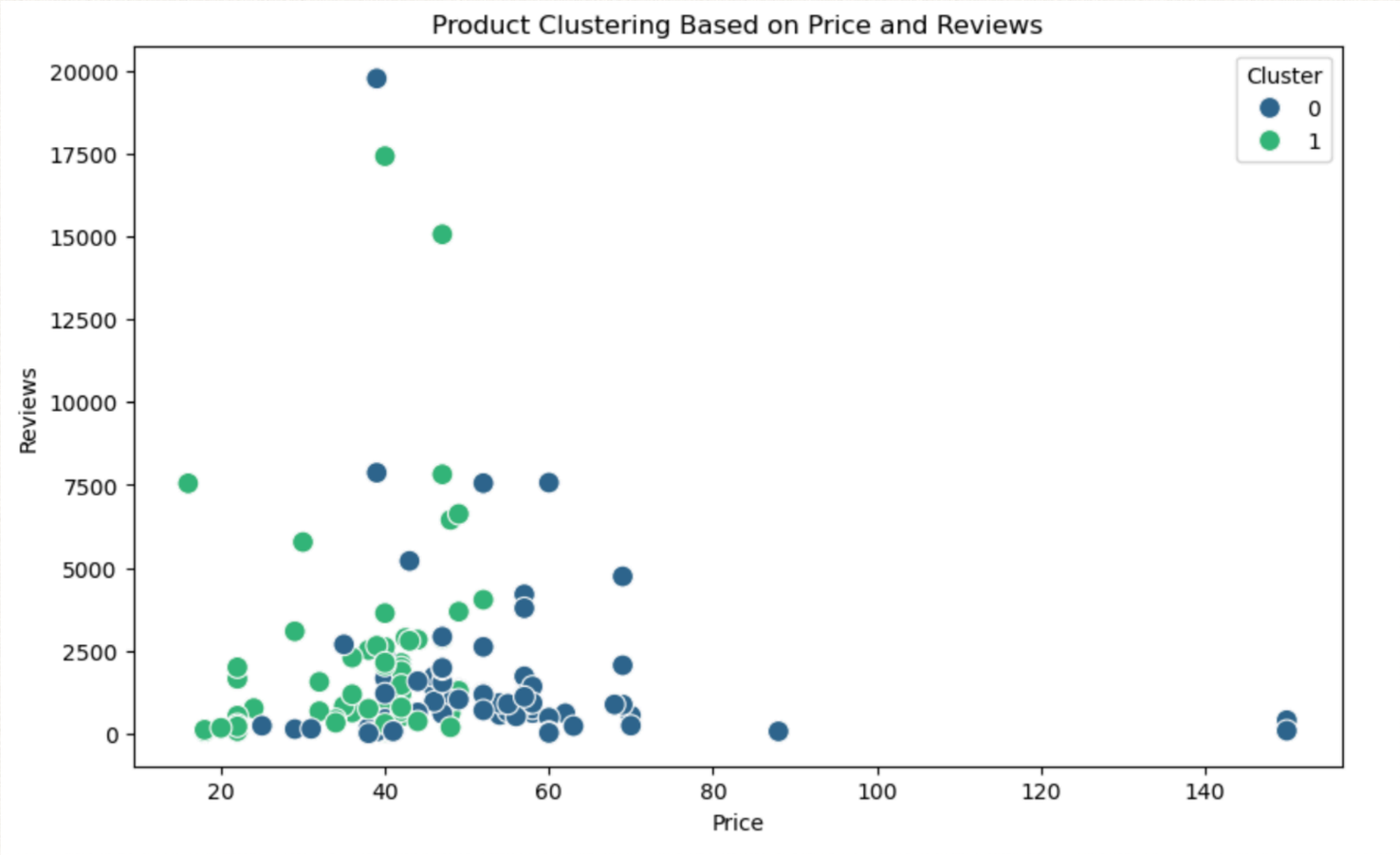
- This findings align with our assumption about the relationship between **brand image** and a product pricing. Brands that are positioned as premium, along with good rating and reviews (reflecting positive perception of the brand and thus good brand image) are predicted to have high prices.
- We are surprised that **number of colors** does not have a big impact on the pricing of a product, implying that a wider range of product shades (more inclusivity) is not reflected in high product value (in terms of pricing).

PRODUCT CLUSTERING



The elbow method identifies the optimal number of clusters ('K') in K-means by plotting variance explained against cluster count and locating the point where the improvement sharply levels off.

-> 2 clusters



Cluster 0 (blue):

- Includes products with a wider price range (from $23 up to $150).
- Most of these products have a low to moderate number of reviews (concentrated below ~5,000 reviews).
- A few outliers exist with very high prices but little to no reviews.

Cluster 1 (green):

- Comprises products mostly in the lower price range (below ~$50).
- Many of these products have higher review counts, including products with over 15,000 reviews.
- This cluster indicates popular products at a more affordable price point.

POPULARITY CLASSIFICATION

Create a binary target variable (isPopular: 1 means "Popular" and 0 means "Not Popular") based on:

- Reviews > 2255.0 (75th percentile)
- Rating >= 4.352475 (75th percentile)

Model: Random Forest Classifier for its simplicity and robustness.

80% for training and 20% for testing

```
Accuracy: 0.96
Classification Report:
              precision    recall  f1-score   support

       False       0.95      1.00      0.98        20
        True       1.00      0.80      0.89         5

    accuracy                           0.96        25
   macro avg       0.98      0.90      0.93        25
weighted avg       0.96      0.96      0.96        25
```

```
Confusion Matrix:
[[20  0]
 [ 1  4]]
```

The model achieves a very high accuracy of 96%.

Reviews is the most influential feature, followed by Rating, Num_Colors, and Price.

As I increase the thresholds for reviews and ratings, the accuracy of the model increases and can eventually reach 100%.

```
Feature Importance:
      Feature  Importance
3     Reviews    0.444438
2      Rating    0.121139
1  Num_Colors    0.114828
0       Price    0.064790
```

Implications For Stakeholders

01.

Brand and Pricing

Positive reviews and strong brand names drive higher prices. Businesses should focus on enhancing brand reputation and customer satisfaction to support premium pricing.

02.

Inclusivity and Pricing

The number of colors doesn't significantly affect pricing, suggesting that offering more shades may not justify higher prices. It doesn’t mean that brands should invest less in expanding color options.

03.

High-End Market

Premium products with fewer reviews cater to niche markets. Brands can target high-end consumers by focusing on exclusivity and quality rather than volume of reviews.

Ethical Implications

01.

- Some with brands limited number of foundation shades may exclude a significant consumer group, contributing to a lack of representation.
- The models favor popular or high-rated products, potentially sidelining smaller or lesser-known brands, and limiting consumer choice.

Legal Implications

02.

- The lack of clear standards for advertising "natural" or "organic" products can lead to misleading claims and greenwashing.
- Advertisements for benefits like hydration or long-wear, which are difficult to substantiate, can deceive consumers and misrepresent product effectiveness.

Societal Implications

03.

- The demand for "limited edition" products creates urgency and a desire for exclusivity, driving overconsumption.
- If consumers show interest in eco-friendly products, it could push brands to prioritize sustainability.

THANK YOU!