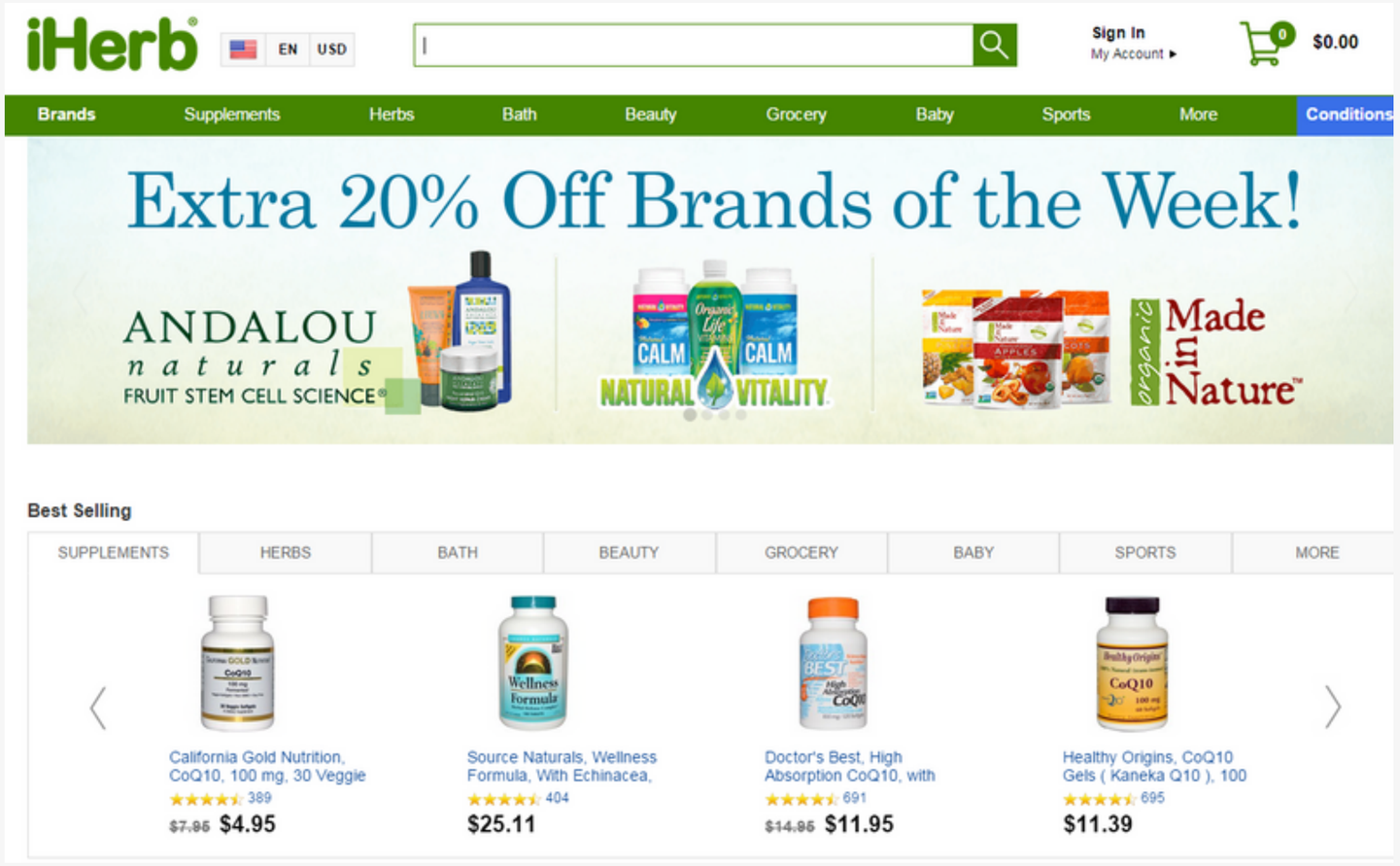
iHerb Products Prices Prediction

TABLE OF CONTENT

INTRODUCTION

Online Shopping

INTRODUCTION

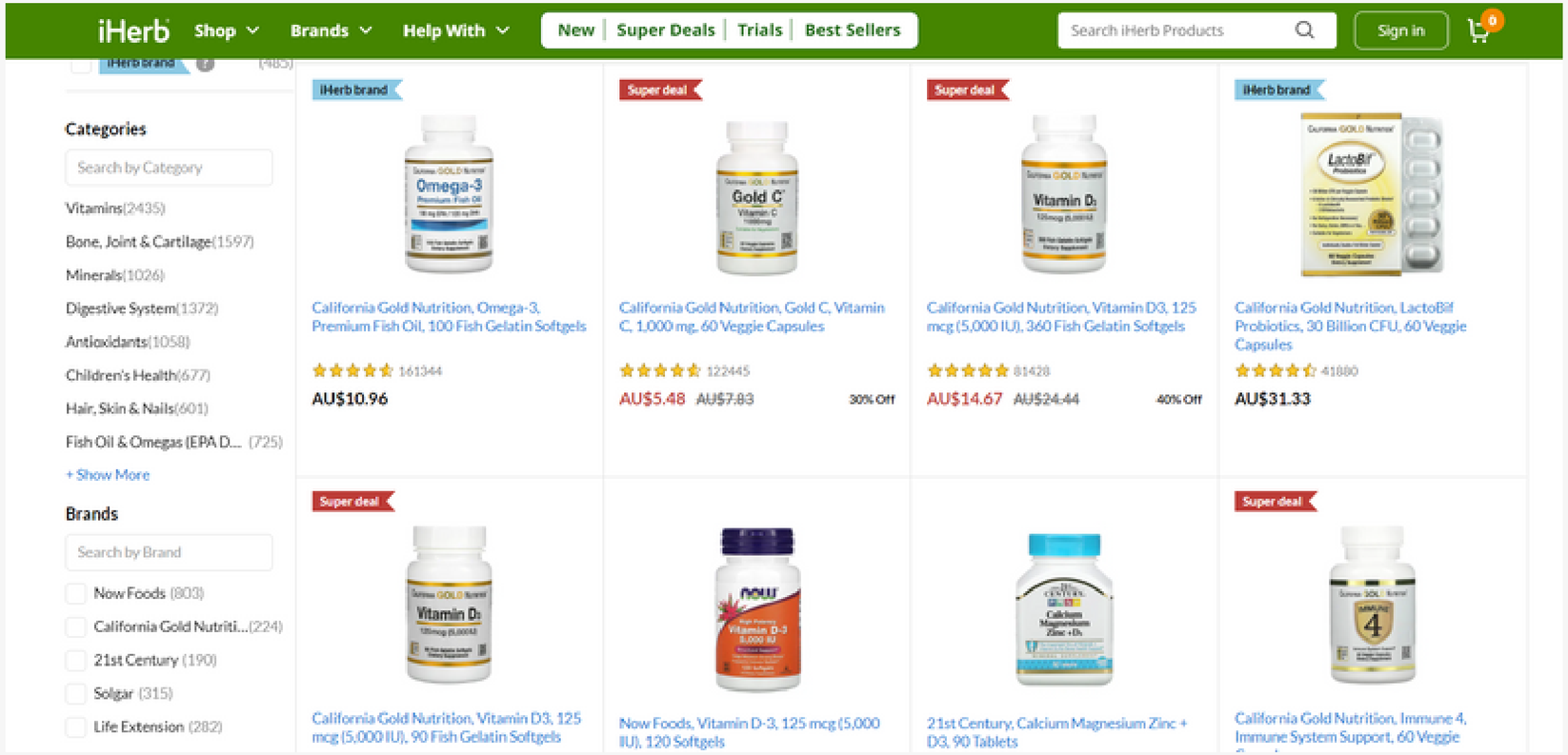


+30K products

INTRODUCTION

Problem Statement

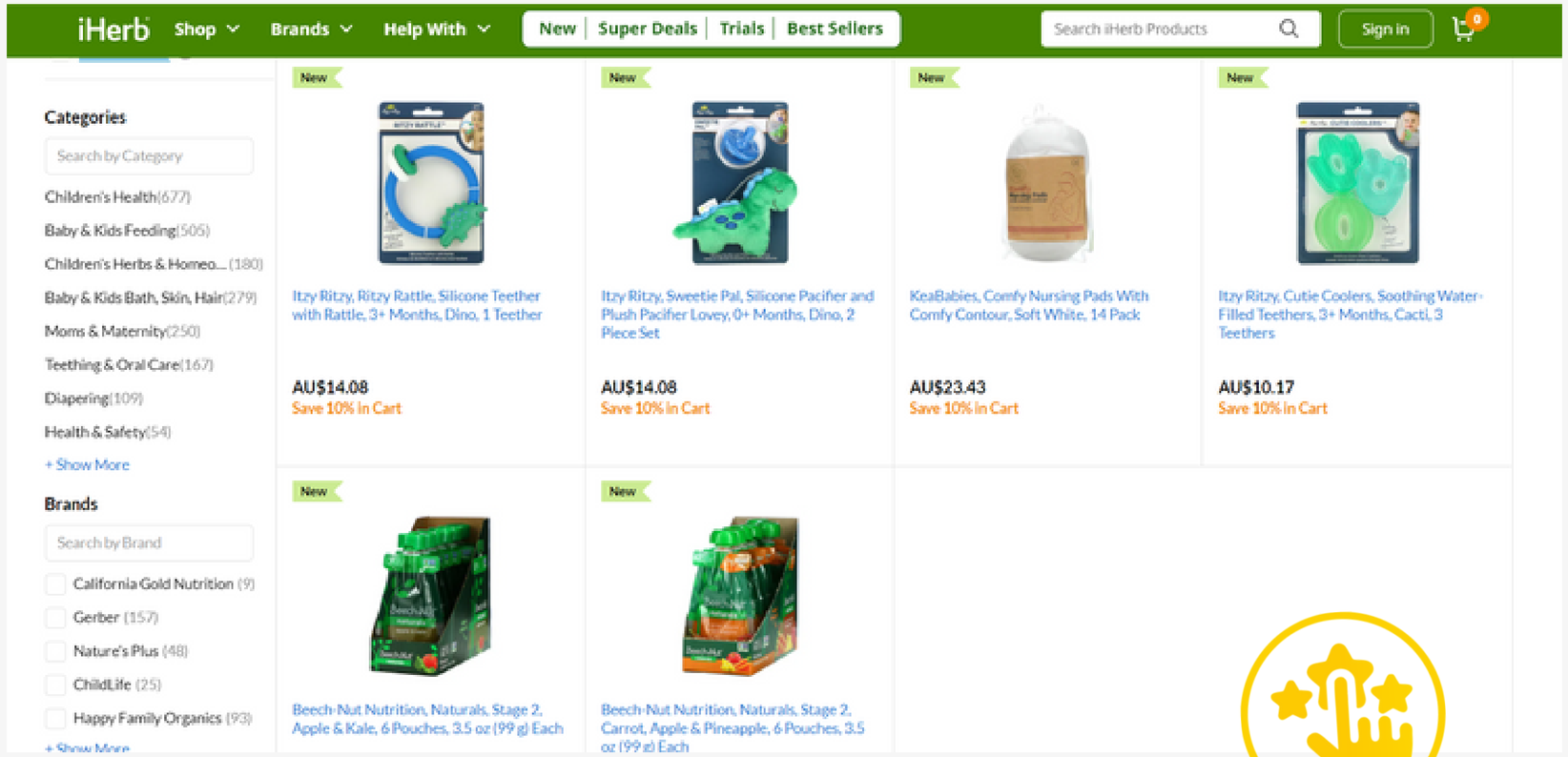
How can we determine the distribution of products on the site?



INTRODUCTION

Problem Statement

How can we determine the distribution of products on the site?



INTRODUCTION

Problem Statement

How can we determine the price of a product ?

DATASET

Web Scraping

2641 Products

DATASET

Name: the name of the product

Categories: the name of the category the product belongs to (e.g., Beauty, Supplements, Grocery)

Size: the size measurement of the products (measures with mg, g, kg…)

Ratings: the average rate of the product.

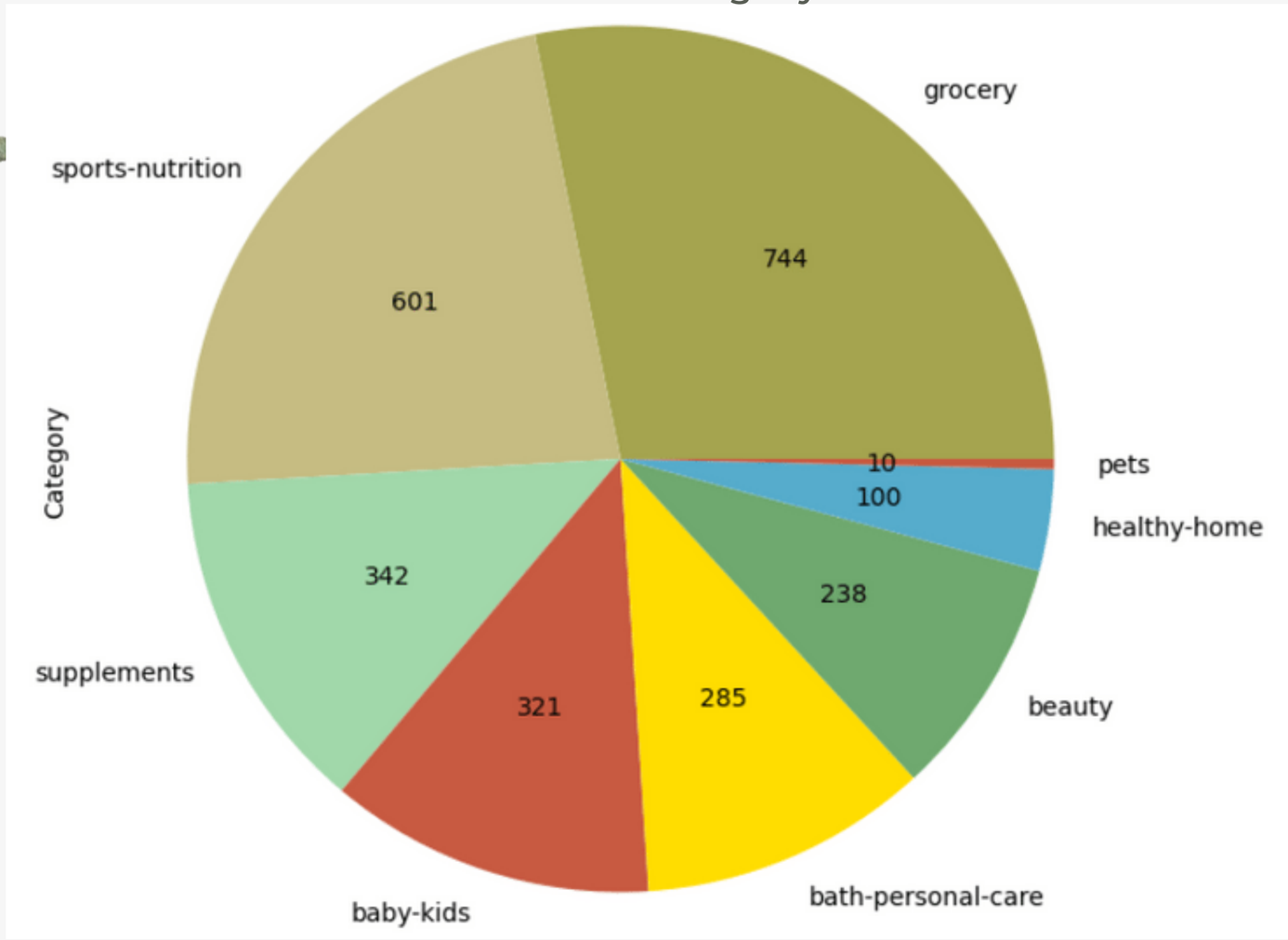
Reviews: the number of customers reviews on the product

Price: product cost

DATASET

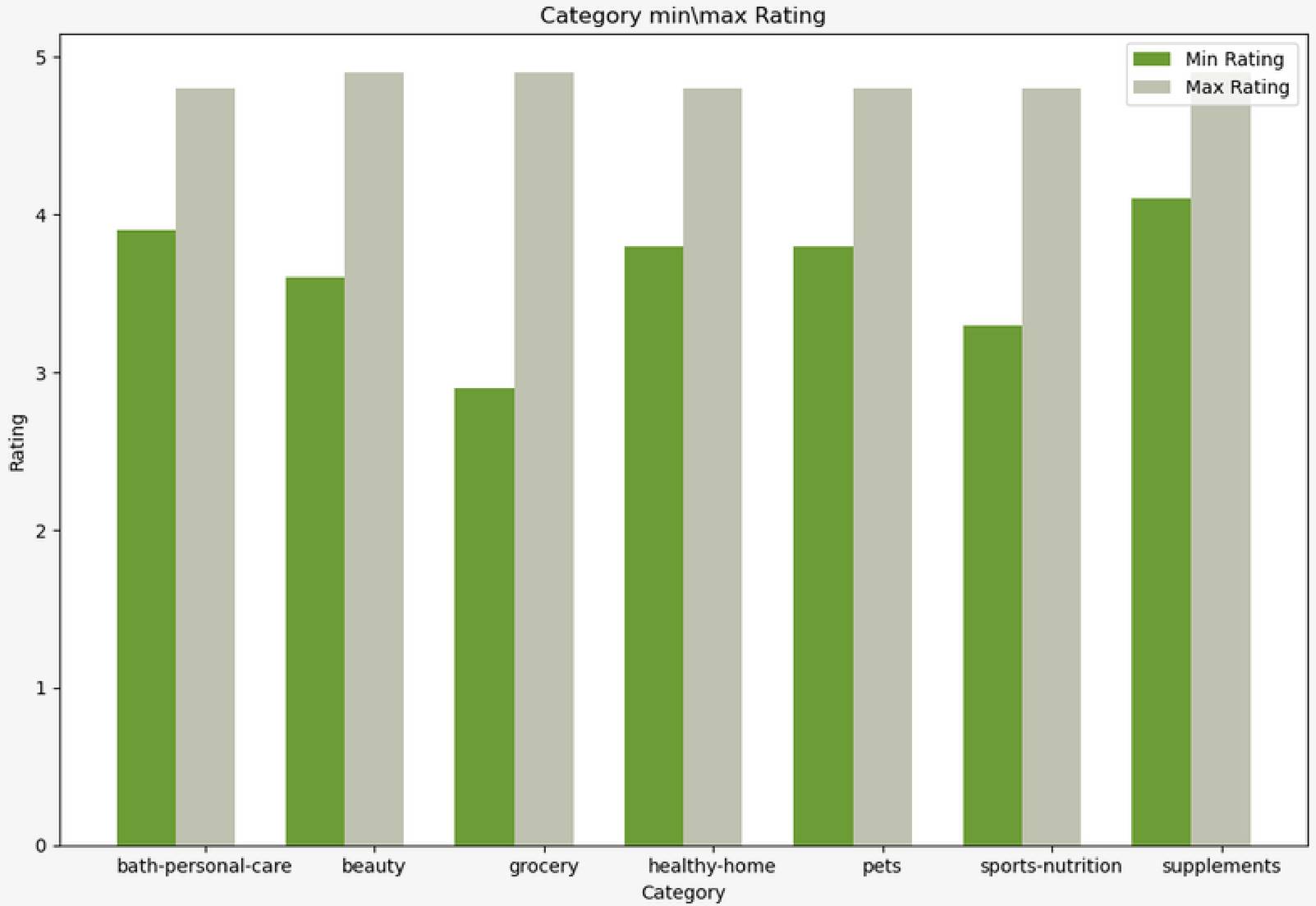
Exploratory data analysis

Products category



DATASET

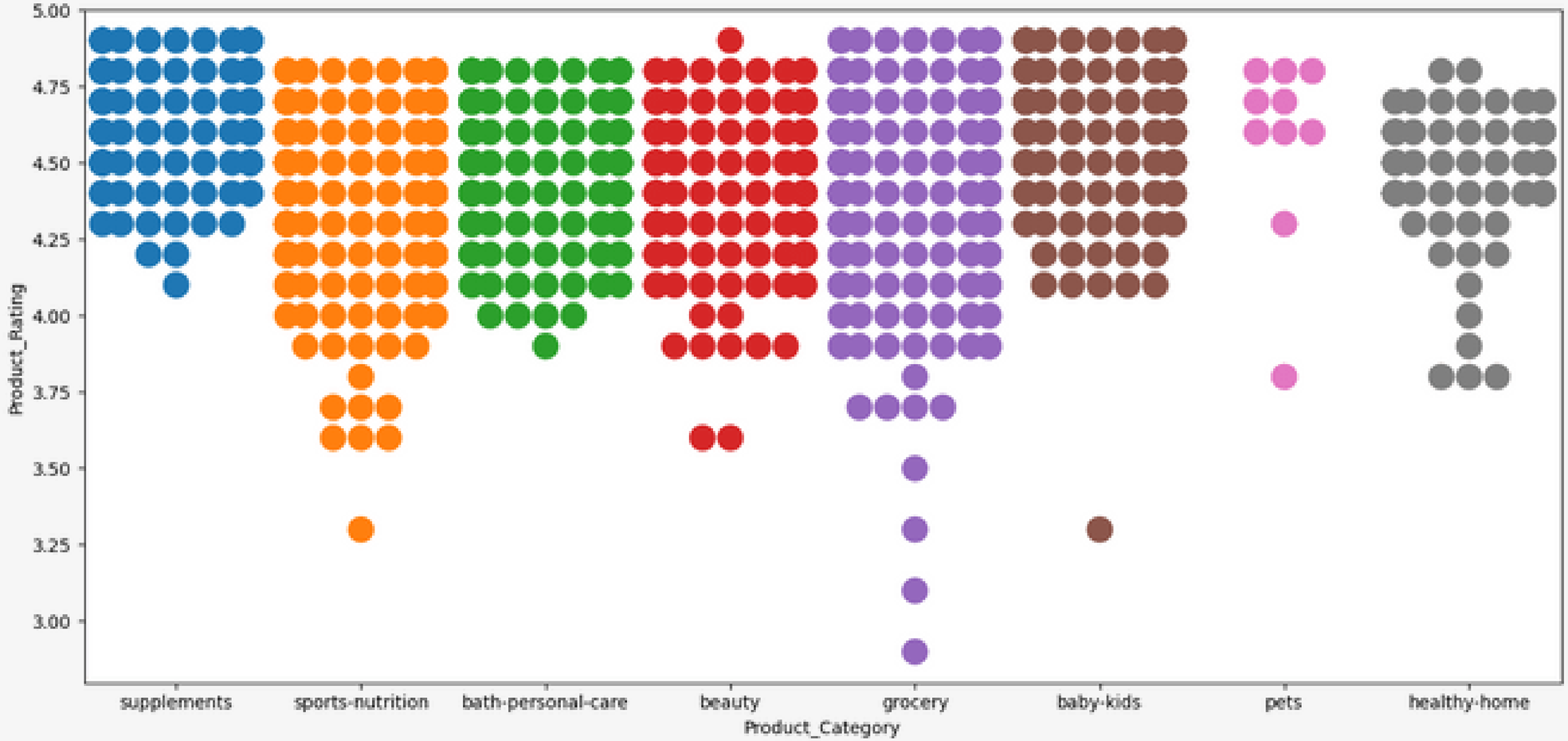
Exploratory data analysis



DATASET

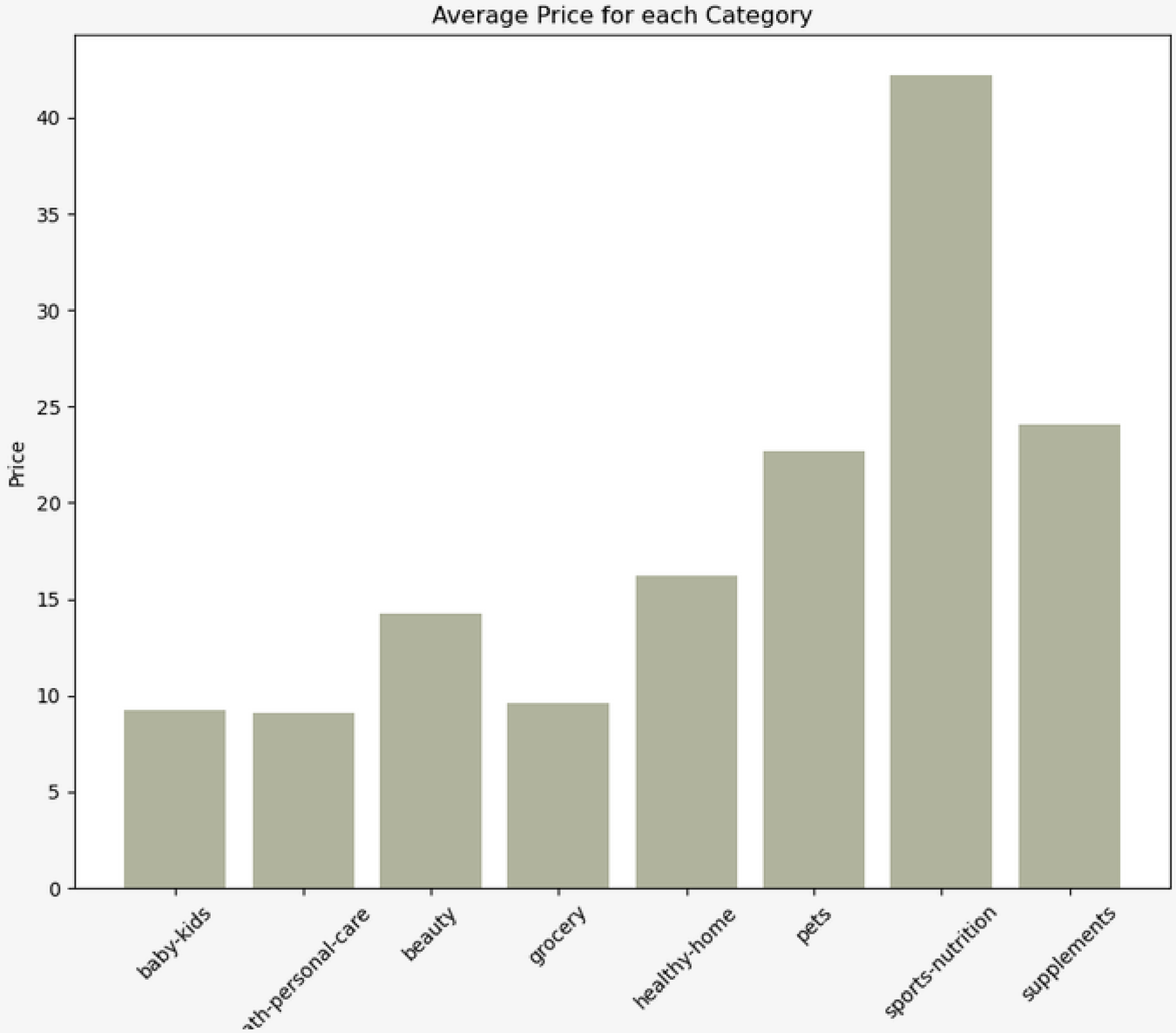
Exploratory data analysis

Products category vs Rating



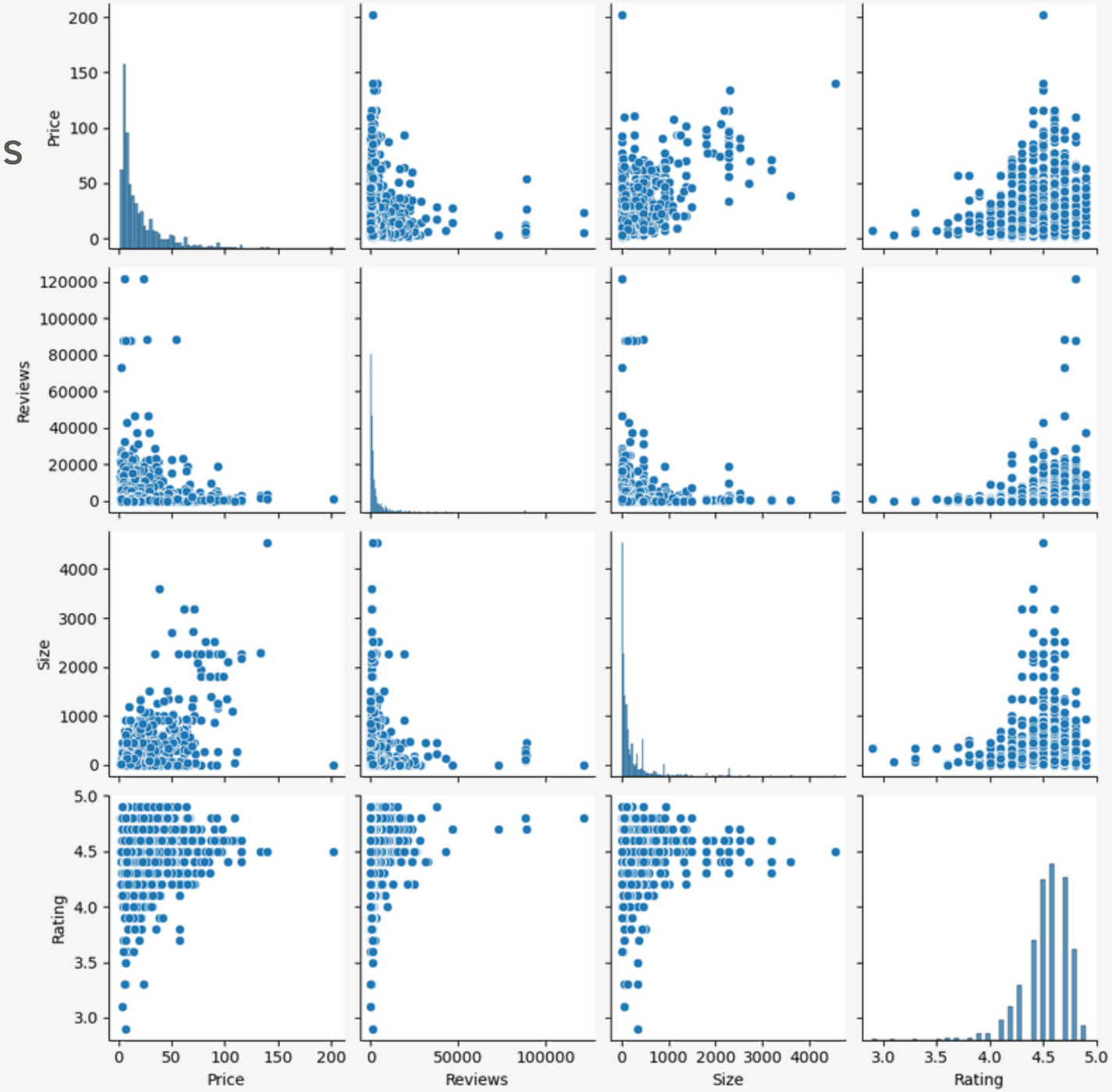
DATASET

Exploratory data analysis



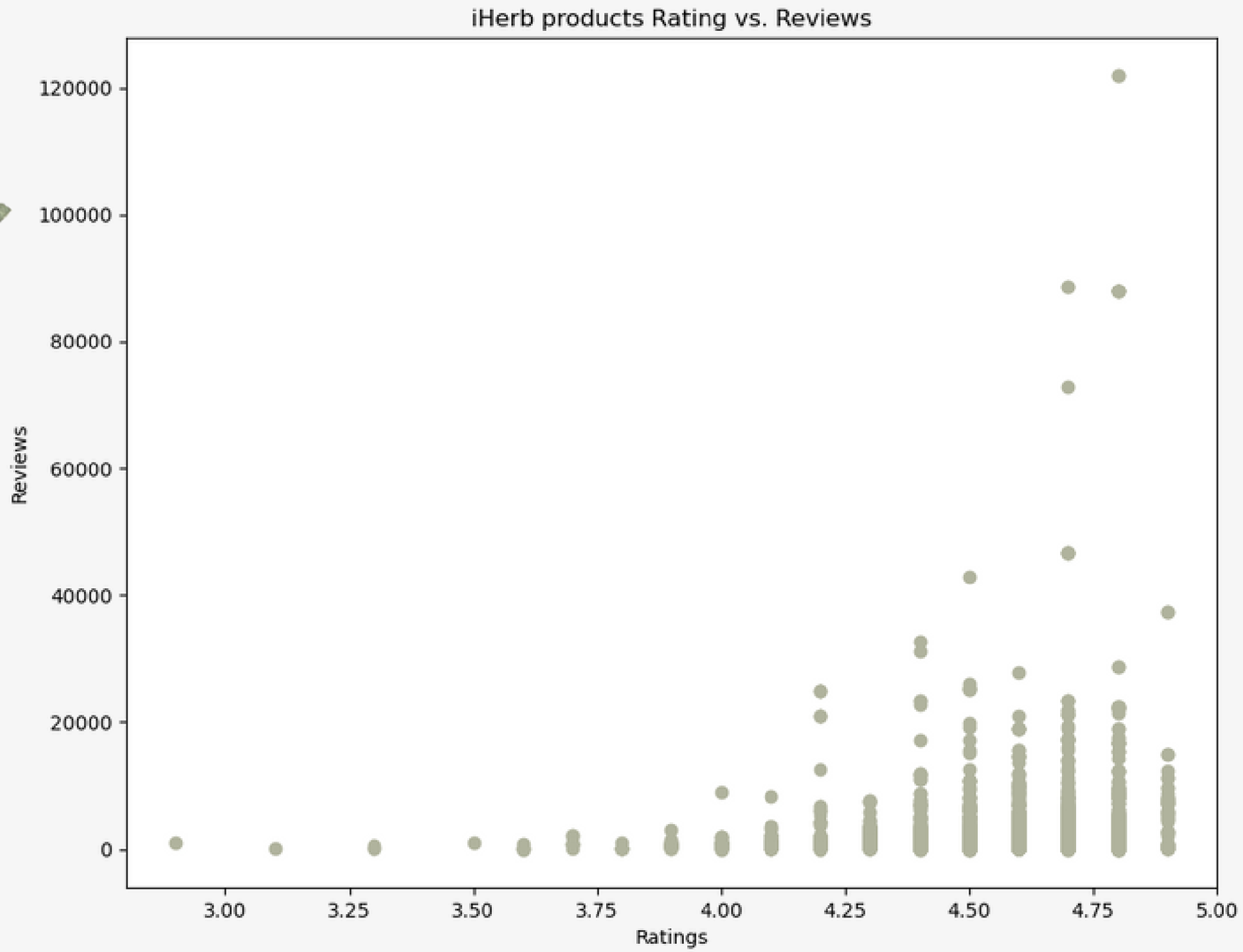
DATASET

Exploratory data analysis



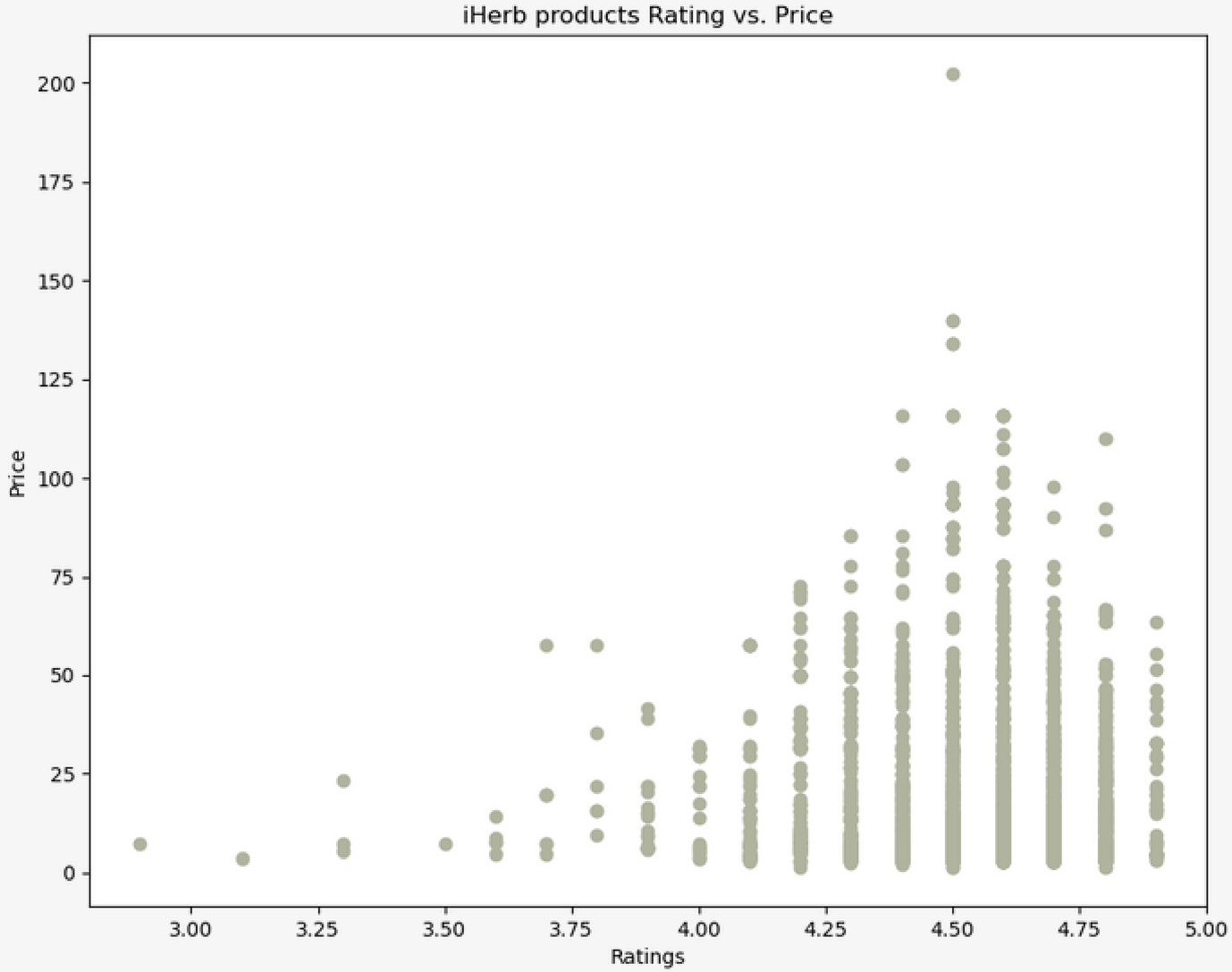
DATASET

Exploratory data analysis



DATASET

Exploratory data analysis



DATASET

Exploratory data analysis

statistics of numerical features

	Price	Reviews	Size	Rating
count	2641.000000	2641.000000	2641.000000	2641.000000
mean	19.519345	2972.300644	219.787373	4.536350
std	20.446082	7411.845875	420.341065	0.212801
min	1.550000	2.000000	0.000000	2.900000
25%	6.220000	442.000000	22.000000	4.400000
50%	11.080000	1068.000000	80.000000	4.600000
75%	25.400000	2689.000000	227.000000	4.700000
max	202.200000	121799.000000	4540.000000	4.900000

DATASET

Exploratory data analysis

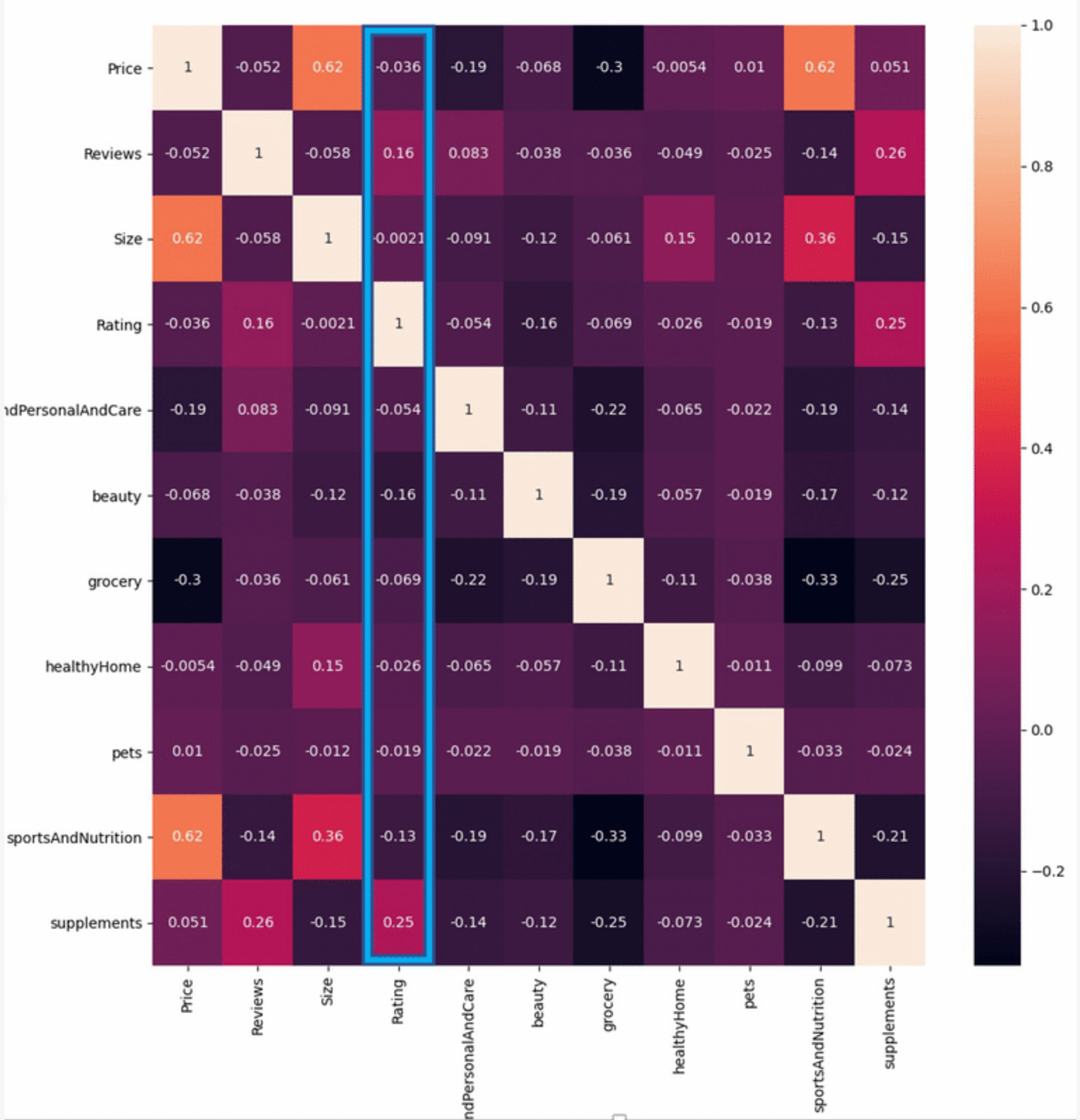
statistics of numerical features

	Price	Reviews	Size	Rating
count	2641.000000	2641.000000	2641.000000	2641.000000
mean	19.519345	2972.300644	219.787373	4.536350
std	20.446082	7411.845875	420.341065	0.212801
min	1.550000	2.000000	0.000000	2.900000
25%	6.220000	442.000000	22.000000	4.400000
50%	11.080000	1068.000000	80.000000	4.600000
75%	25.400000	2689.000000	227.000000	4.700000
max	202.200000	121799.000000	4540.000000	4.900000

DATASET

Exploratory data analysis

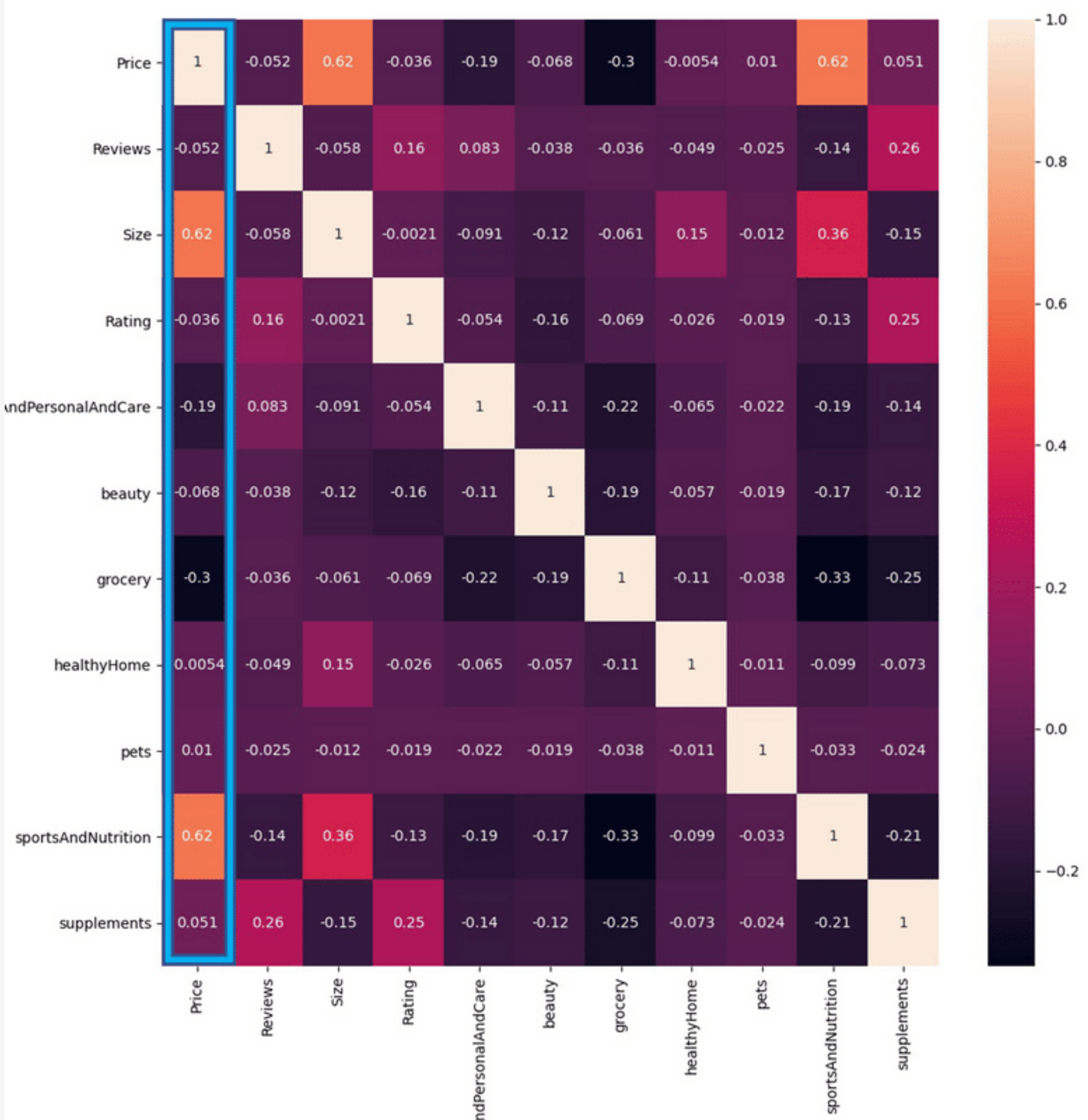
Features correlation



DATASET

Exploratory data analysis

Features correlation



DATASET

Name: the name of the product

Categories: the name of the category the product belongs to (e.g., Beauty, Supplements, Grocery)

Size: the size measurement of the products (measures with oz, grams…)

Ratings: the average rate of the product.

Reviews: the number of customers reviews on the product

Price: product cost

PRE-PROCESSING

1. EXTRACT PRODUCT SIZE FROM ITS NAME
2. CONVERT CATEGORICAL VARIABLES TO DUMMY VARIABLES
3. SCALE PRODUCTS SIZE TO GRAMS
4. DROP ROWS WITH NULL VALUES

MODELING

LASSO Regression

SPLITTING THE DATASET

49% Training - 21% Validation - 30% Testing

MODELING

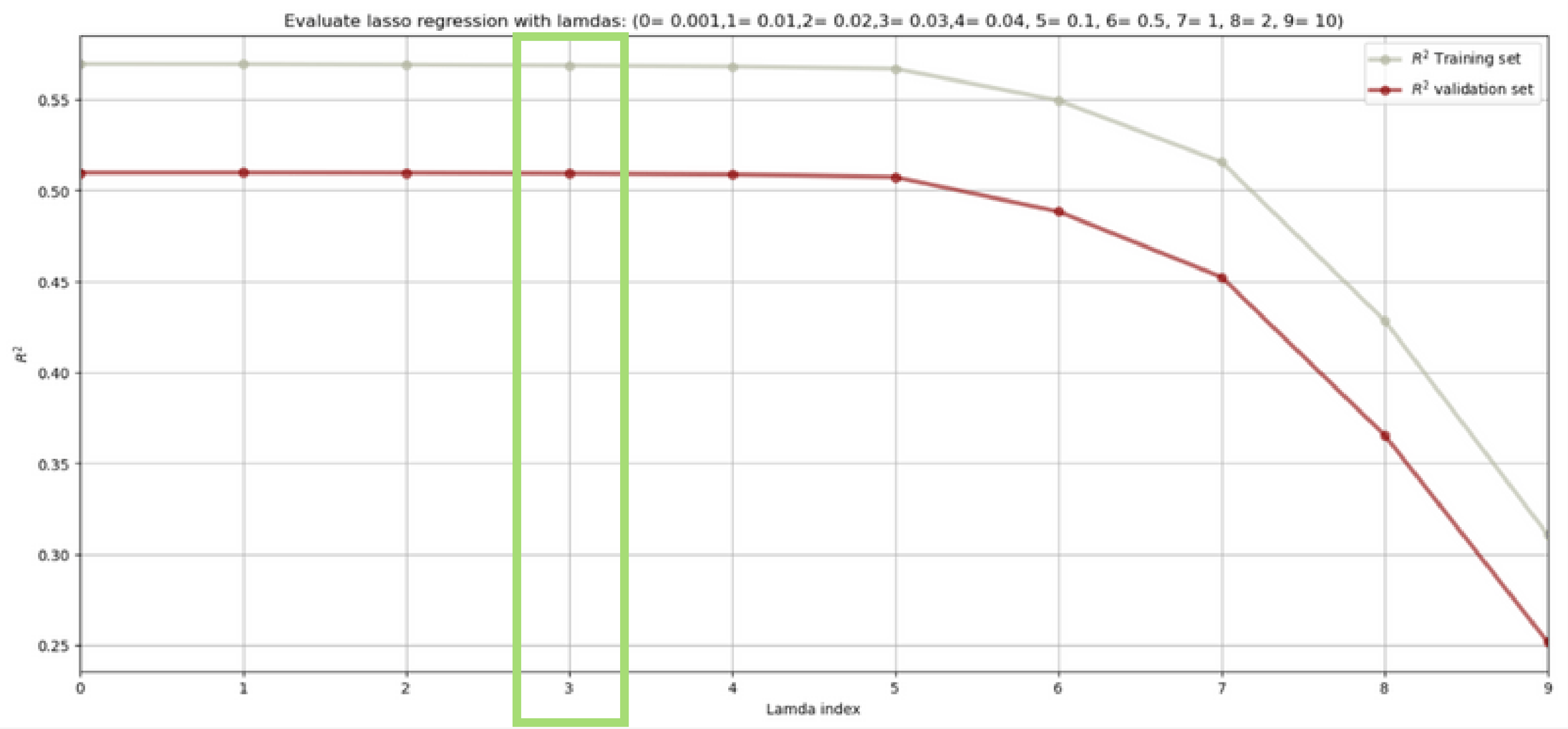
LASSO Regression

WHAT VALUE OF ALPHA SHALL I USE?

MODELING

LASSO Regression

EVALUATE WITH MULTIPLE VALUES



MODELING

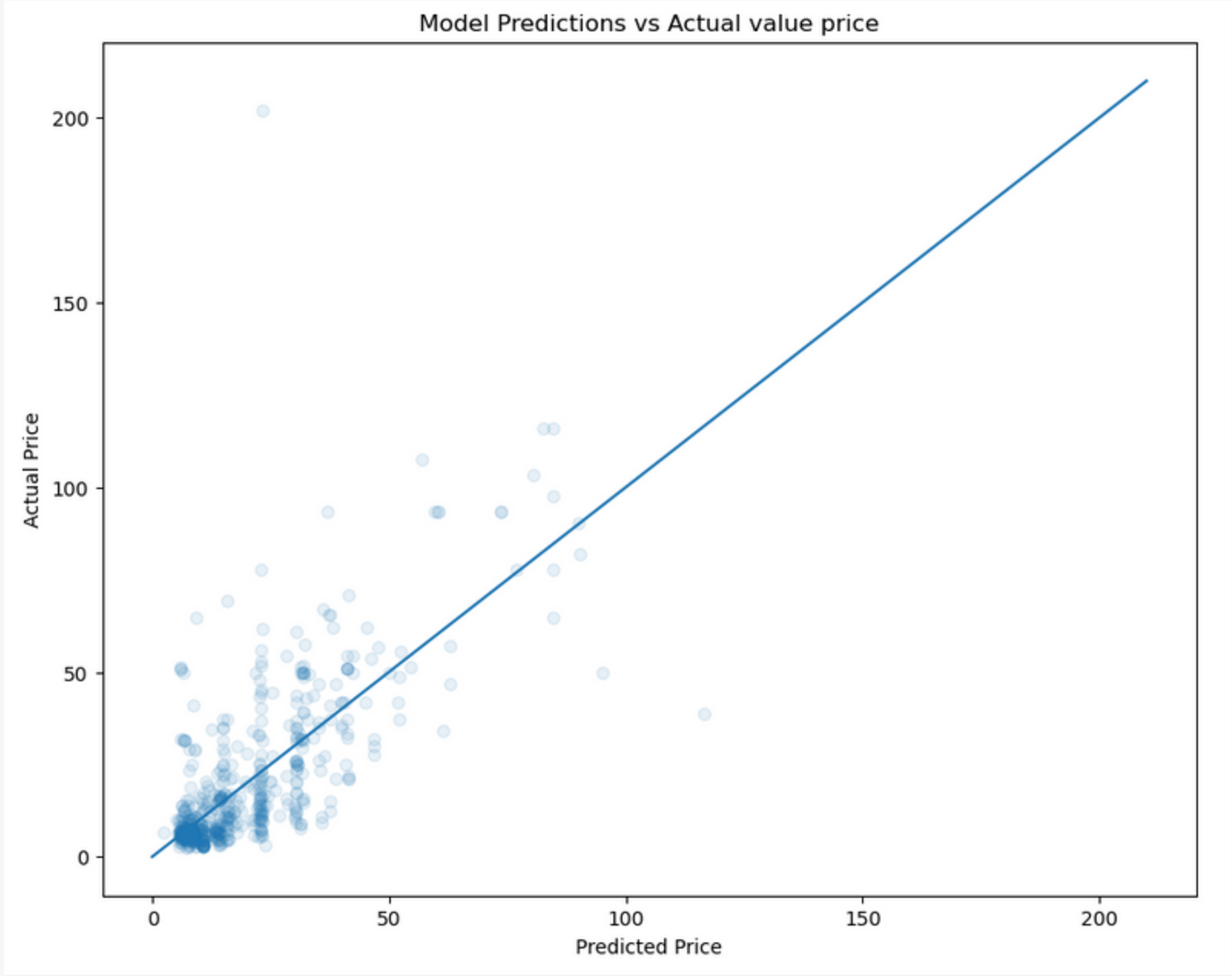
LASSO Regression

LASSO 5 CROSS VALIDATION

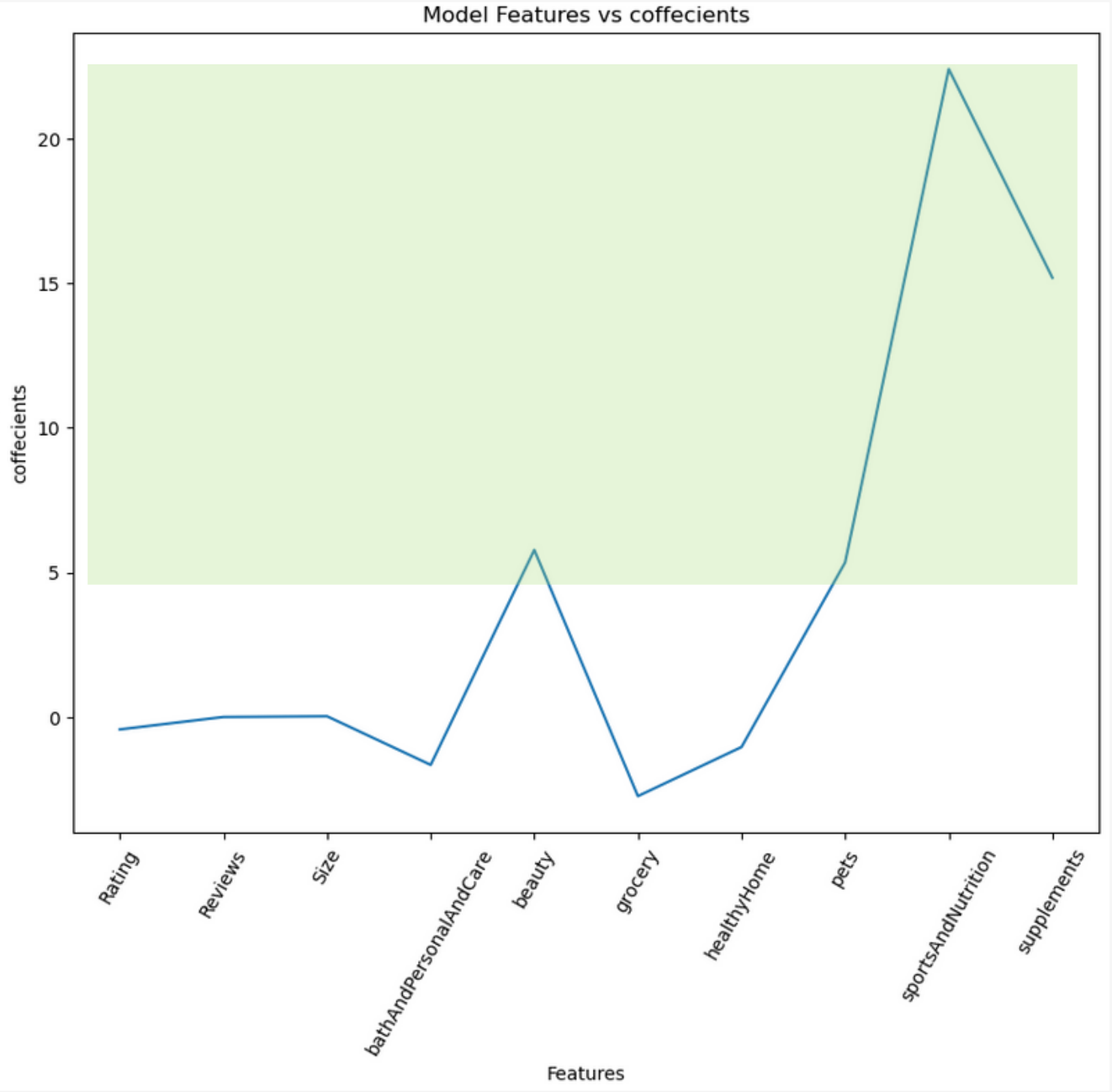
BEST ALPHA VALUE = 0.033

MODELING

LASSO Regression



R^2 SCORE ON VALIDATION SET: 50.484%

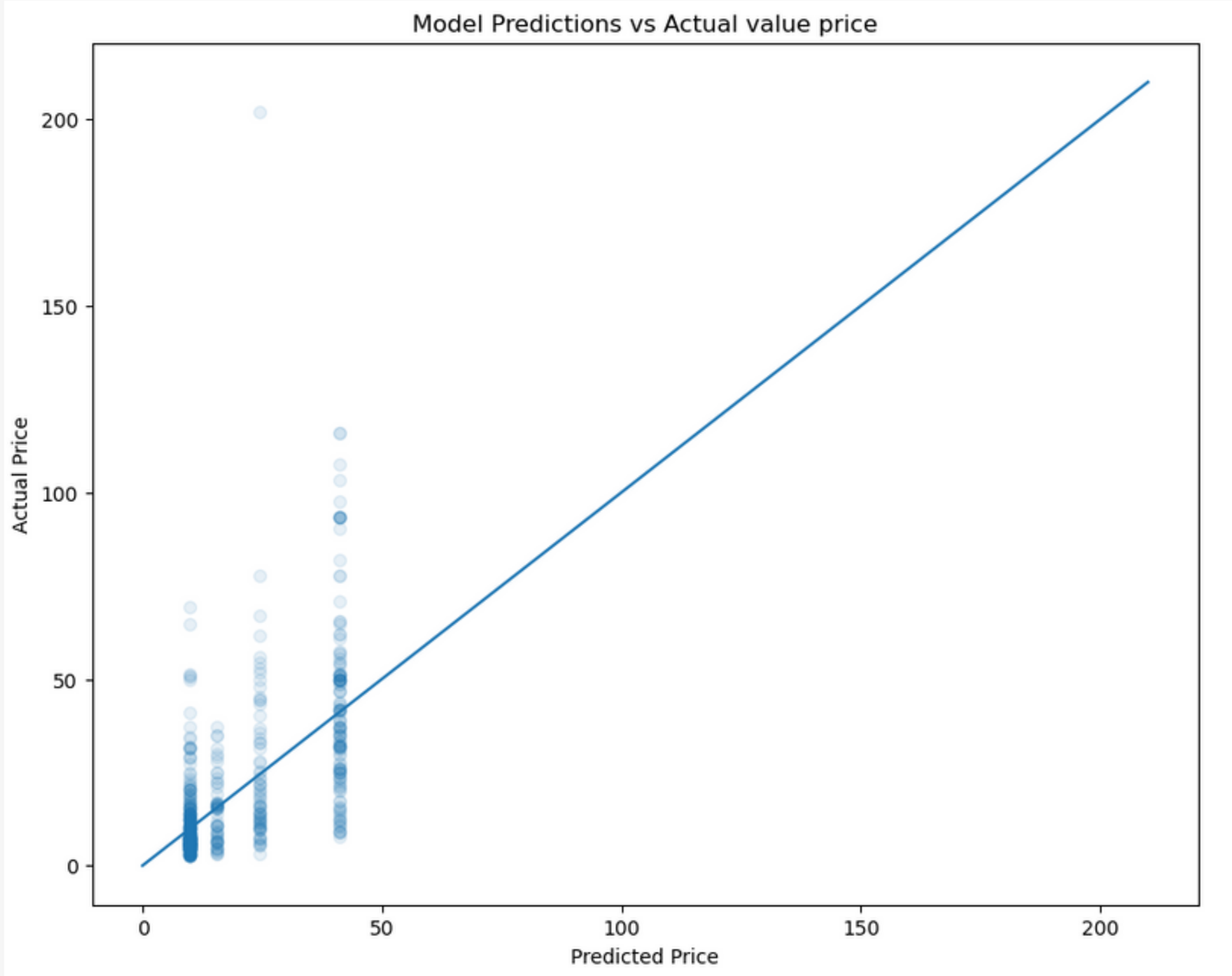


MODELING

LASSO FOR FEATURE SELECTION

MODELING

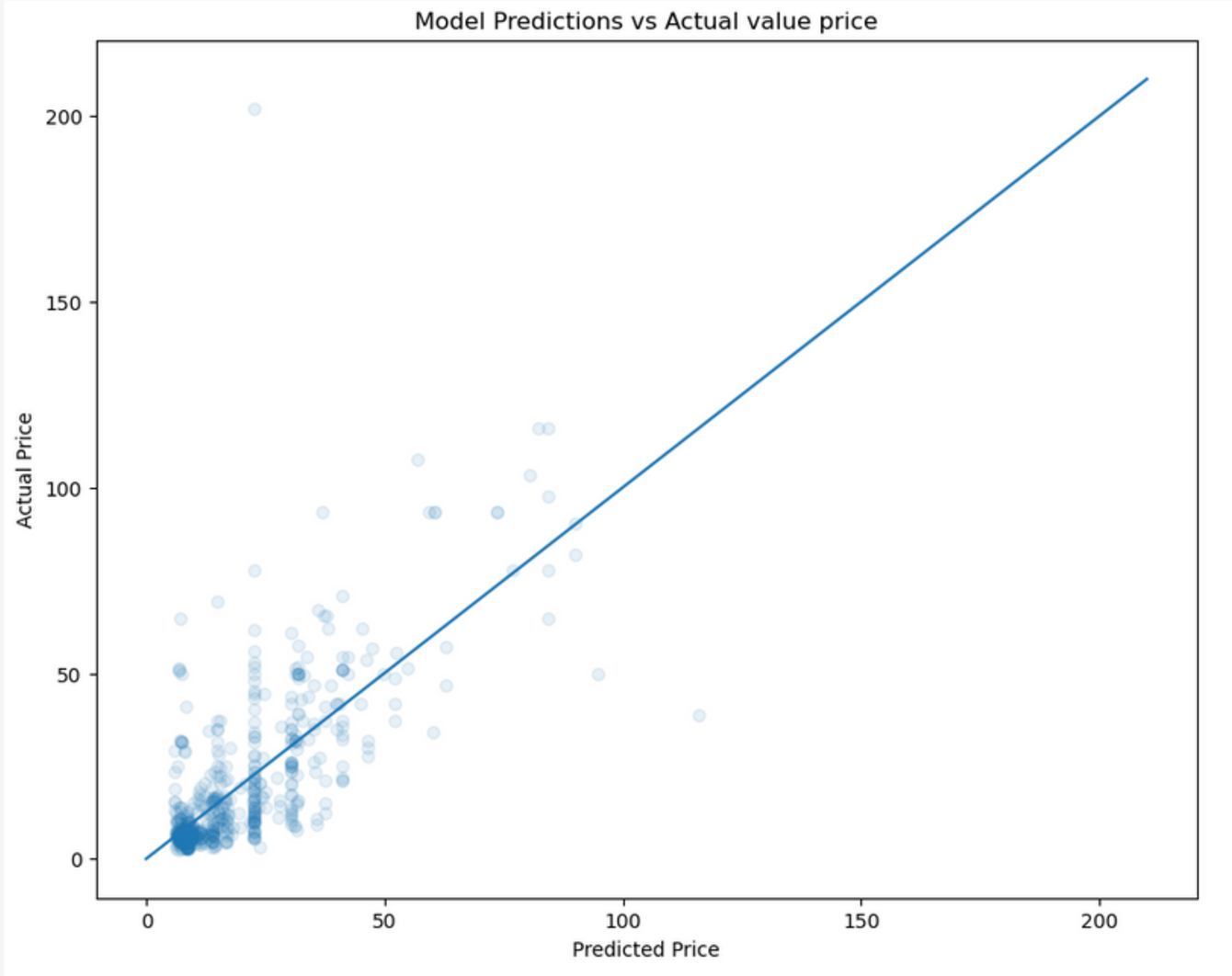
LASSO Regression on selected high coefficient features



R^2 SCORE ON VALIDATION SET: 37.908%

MODELING

LASSO Regression on selected high coefficient and correlation features



R^2 SCORE ON VALIDATION SET: 50.58%

MODELING

Linear Regression

1 SKLEARN LIBRARY

2 STATSMODELS LIBRARY

MODELING

Linear Regression - Sklearn Library

1 **SPLITTING THE DATASET**

70% training - 30% testing

Before Feature Selection

MODELING

Linear Regression - Sklearn Library

2 RESULT

- Training accuracy: 57%
- Training error: 43%

- Cross Validation score: 56%

- Testing accuracy: 70%
- Testing error: 30%

After Feature Selection

MODELING

Linear Regression - Sklearn Library

2 RESULT

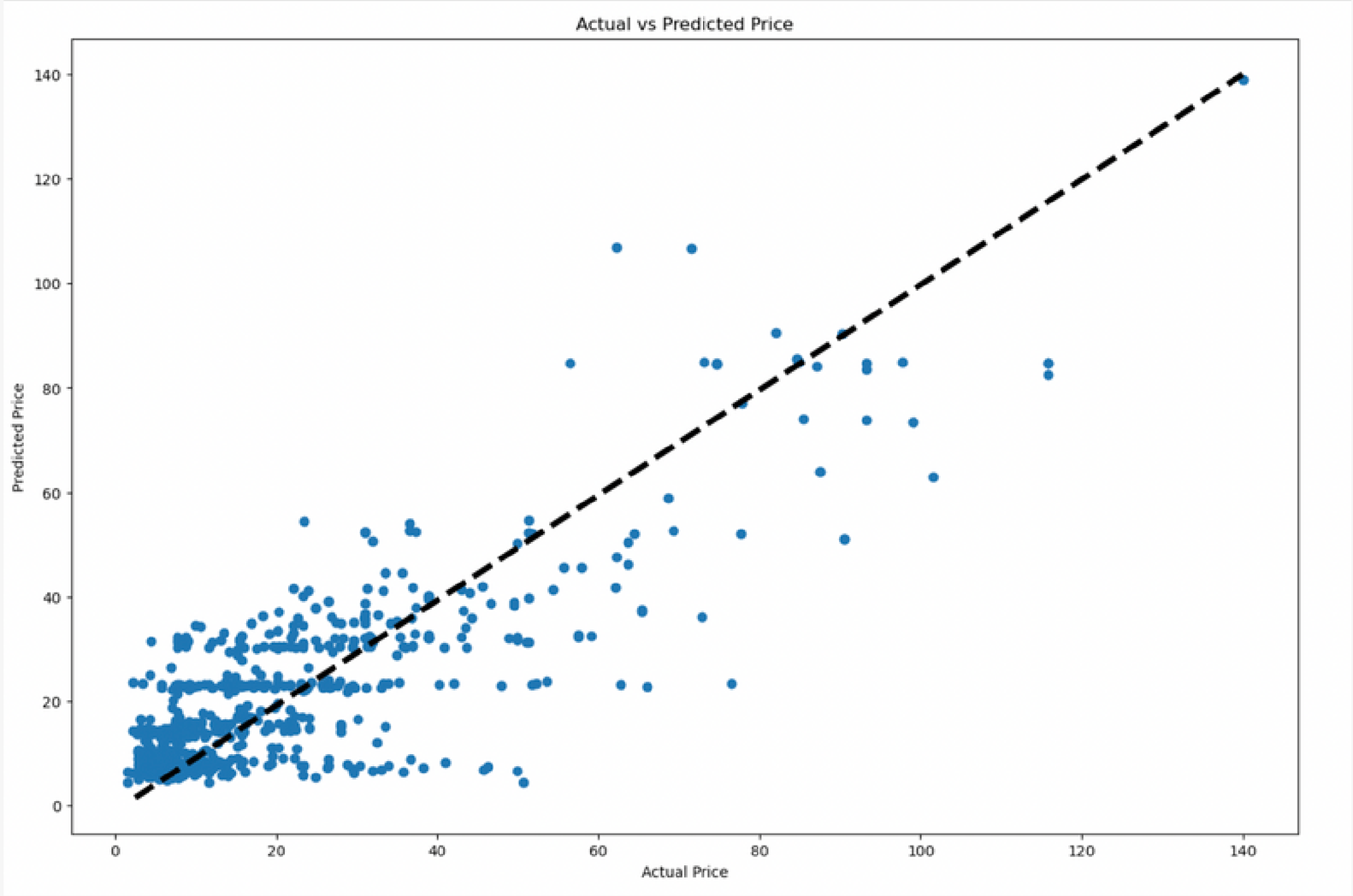
- Training accuracy: 57%
- Training error: 43%

- Cross Validation score: 56%

- Testing accuracy: 70.1%
- Testing error: 30.22%

MODELING

Linear Regression - Sklearn Library



MODELING

Linear Regression - Statsmodels

R-squared:	0.568
Adj. R-squared:	0.567
F-statistic:	484.8

MODELING

Linear Regression - Statsmodels

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.0472	0.446	13.556	0.000	5.172	6.922
Size	0.0226	0.001	26.737	0.000	0.021	0.024
beauty	7.7573	1.144	6.783	0.000	5.514	10.000
pets	13.7981	5.113	2.699	0.007	3.770	23.826
sportsAndNutrition	25.1493	0.819	30.710	0.000	23.543	26.755
supplements	17.7530	0.972	18.274	0.000	15.848	19.658

MODELING

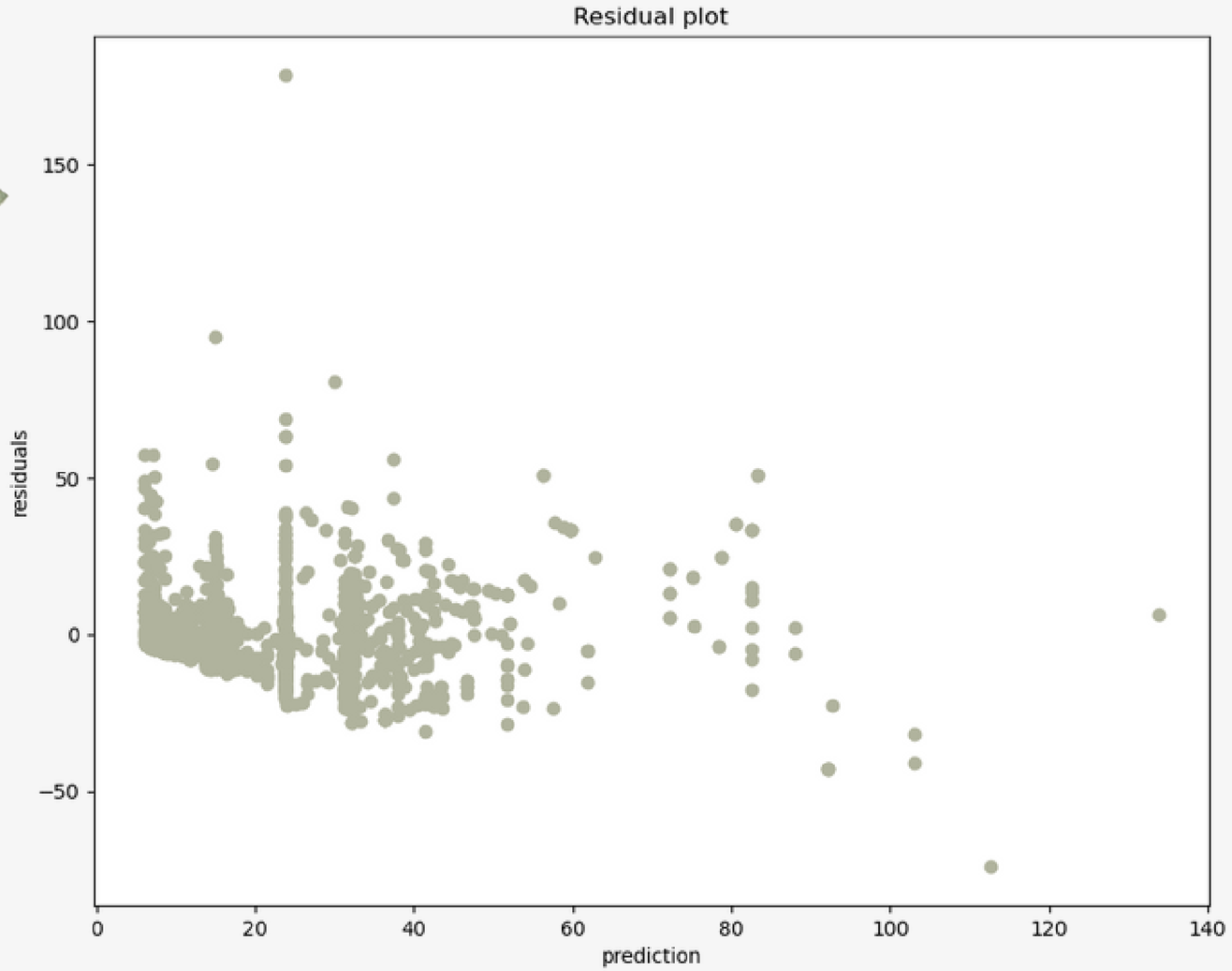
Assumptions

1. Assumption 1: regression is linear in parameters and correctly specified

2. Assumption 2: residuals should be normally distributed with zero mean

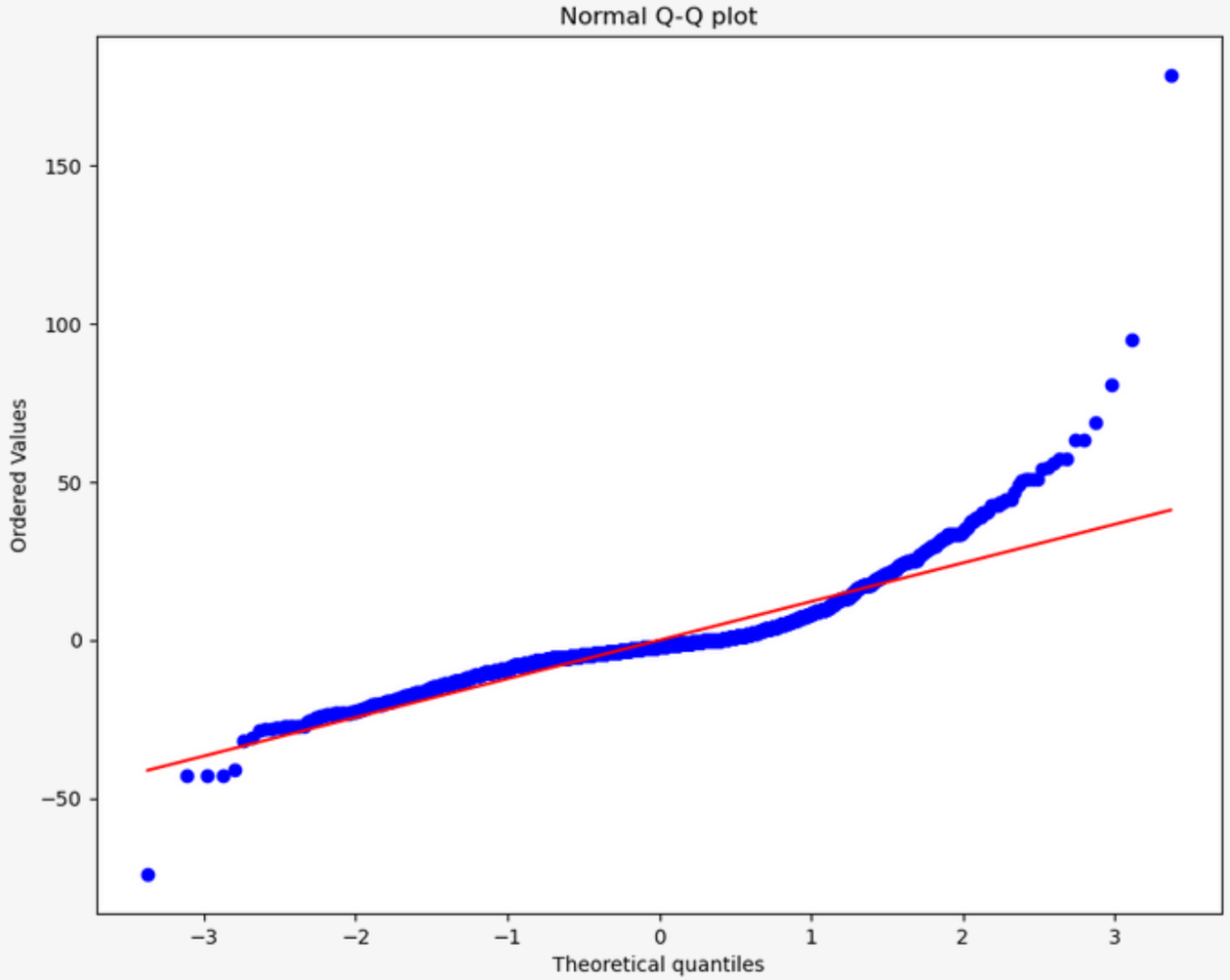
MODELING

Assumptions



MODELING

Assumptions



MODELING

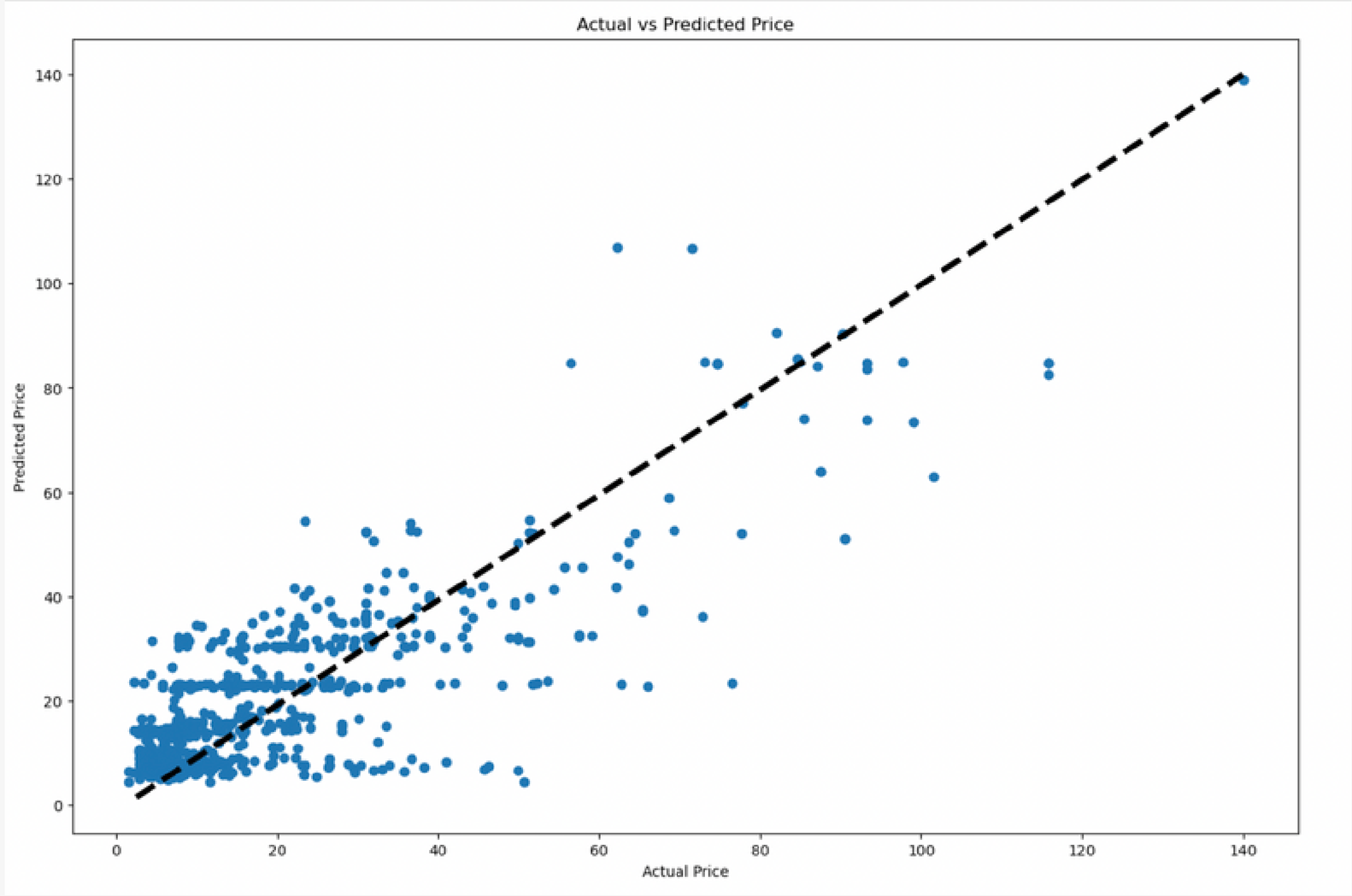
Assumptions

3

Assumption 3: error terms must have constant variance

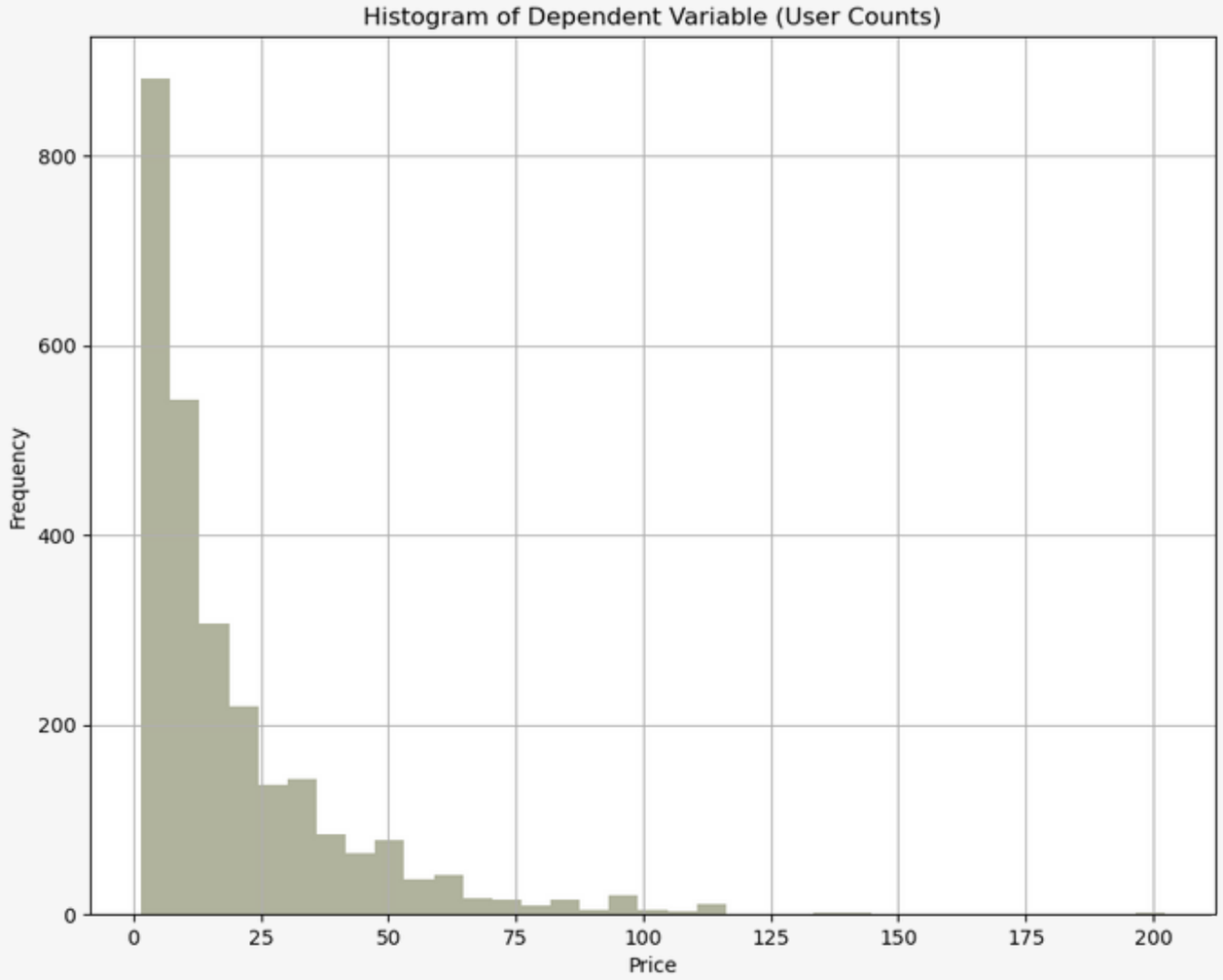
MODELING

Assumptions



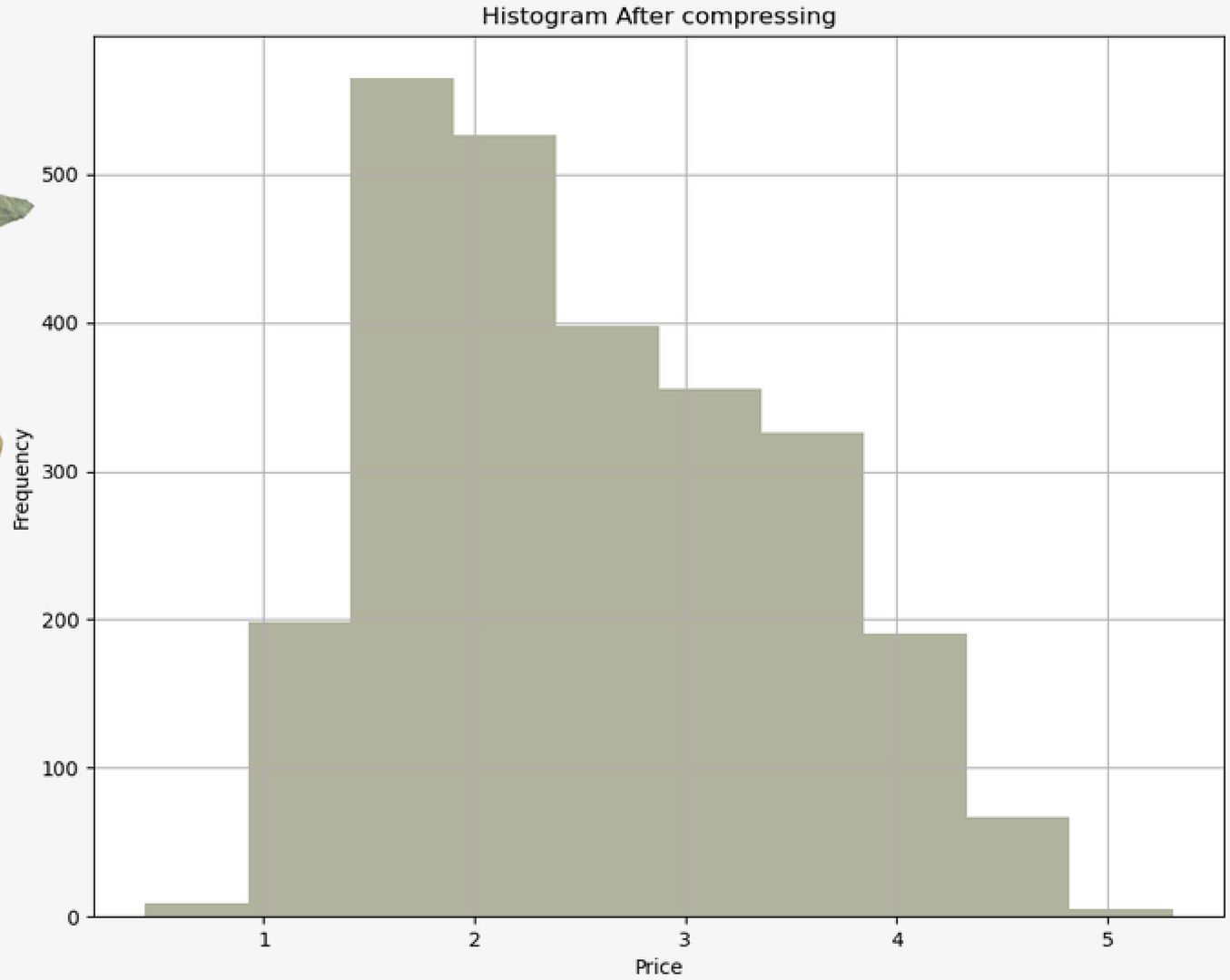
MODELING

Assumptions



MODELING

Assumptions



MODELING

Assumptions

4 Assumption 4: errors are uncorrelated across observations

5 Assumption 5: no independent variable is a perfect linear function of any other independent variable (no perfect multi-collinearity)

iHerb®

THANK YOY

Fatimah AlShammari

Sara AlAbdulsalam