



iHerb®

iHerb Products Prices Prediction



TABLE OF CONTENT



Introduction



Dataset

- Web Scraping
- Exploratory data analysis



Pre-processing



Modeling

- LASSO Regression
- Linear Regression (sklearn & statmodel)
- Assumption



Conclusion



INTRODUCTION

Online Shopping



INTRODUCTION



The screenshot shows the homepage of the iHerb website. At the top, there is a navigation bar with the iHerb logo, language selection (EN USD), a search bar, and links for Sign In, My Account, and a shopping cart showing \$0.00. Below the navigation is a main banner with the text "Extra 20% Off Brands of the Week!" and images of products from ANDALOU naturals, NATURAL VITALITY, and Made in Nature. The main content area features a "Best Selling" section with a grid of supplement products. The grid includes four items: California Gold Nutrition CoQ10, Source Naturals Wellness Formula, Doctor's Best High Absorption CoQ10, and Healthy Origins CoQ10 Gels.

SUPPLEMENTS	HERBS	BATH	BEAUTY	GROCERY	BABY	SPORTS	MORE
 California Gold Nutrition, CoQ10, 100 mg, 30 Veggie ★★★★★ 389 \$7.95 \$4.95	 Source Naturals, Wellness Formula, With Echinacea, ★★★★★ 404 \$25.11	 Doctor's Best, High Absorption CoQ10, with ★★★★★ 691 \$14.95 \$11.95	 Healthy Origins, CoQ10 Gels (Kaneka Q10), 100 ★★★★★ 695 \$11.39				

+30K products

INTRODUCTION

Problem Statement

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



The image shows a screenshot of the iHerb website's product listing page. The header features the iHerb logo, navigation links for Shop, Brands, Help With, and categories like New, Super Deals, Trials, and Best Sellers. A search bar and sign-in button are also present. The main content displays a grid of supplement products from the California Gold Nutrition brand, including Omega-3, Gold C, Vitamin D3, LactoBif Probiotics, Vitamin D3, Now Foods Vitamin D3, Calcium Magnesium Zinc + D3, and Immune 4. Each product card includes an image, name, rating, reviews count, price, and discount information.

Product	Brand	Rating	Reviews	Price	Discount
California Gold Nutrition, Omega-3, Premium Fish Oil, 100 Fish Gelatin Softgels	California Gold Nutrition	★★★★★	161344	AU\$10.96	
California Gold Nutrition, Gold C, Vitamin C, 1,000 mg, 60 Veggie Capsules	California Gold Nutrition	★★★★★	122445	AU\$5.48	AU\$7.83
California Gold Nutrition, Vitamin D3, 125 mcg (5,000 IU), 360 Fish Gelatin Softgels	California Gold Nutrition	★★★★★	81420	AU\$14.67	AU\$24.44
California Gold Nutrition, LactoBif Probiotics, 30 Billion CFU, 60 Veggie Capsules	California Gold Nutrition	★★★★★	41880	AU\$31.33	
California Gold Nutrition, Vitamin D3, 125 mcg (5,000 IU), 90 Fish Gelatin Softgels	California Gold Nutrition				
Now Foods, Vitamin D-3, 125 mcg (5,000 IU), 120 Softgels	Now Foods				
21st Century, Calcium Magnesium Zinc + D3, 90 Tablets	21st Century				
California Gold Nutrition, Immune 4, Immune System Support, 60 Veggie	California Gold Nutrition				

INTRODUCTION

Problem Statement

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



The image shows a screenshot of the iHerb website's homepage. The top navigation bar includes links for "Shop", "Brands", "Help With", and categories like "New", "Super Deals", "Trials", and "Best Sellers". A search bar and a sign-in button are also present. The main content area displays a grid of new products. The first row features four items: 1) Itzy Ritzy Rattle, Silicone Teether with Rattle, 3+ Months, Dino, 1 Teether (AU\$14.08, Save 10% in Cart). 2) Itzy Ritzy Sweetie Pal, Silicone Pacifier and Plush Pacifier Lovey, 0+ Months, Dino, 2 Piece Set (AU\$14.08, Save 10% in Cart). 3) KeaBabies Comfy Nursing Pads With Comfy Contour, Soft White, 14 Pack (AU\$23.43, Save 10% in Cart). 4) Itzy Ritzy Cutie Coolers, Soothing Water-Filled Teether, 3+ Months, Cacti, 3 Teethers (AU\$10.17, Save 10% in Cart). The second row features two items: 5) Beech-Nut Nutrition, Naturals, Stage 2, Apple & Kale, 6 Pouches, 3.5 oz (99 g) Each (AU\$14.08, Save 10% in Cart). 6) Beech-Nut Nutrition, Naturals, Stage 2, Carrot, Apple & Pineapple, 6 Pouches, 3.5 oz (99 g) Each (AU\$14.08, Save 10% in Cart). On the left sidebar, there are sections for "Categories" (Search by Category, Children's Health, Baby & Kids Feeding, Children's Herbs & Homeo...) and "Brands" (Search by Brand, California Gold Nutrition, Gerber, Nature's Plus, ChildLife, Happy Family Organics). A yellow circular icon with a hand pointing up and three stars is located in the bottom right corner.

INTRODUCTION

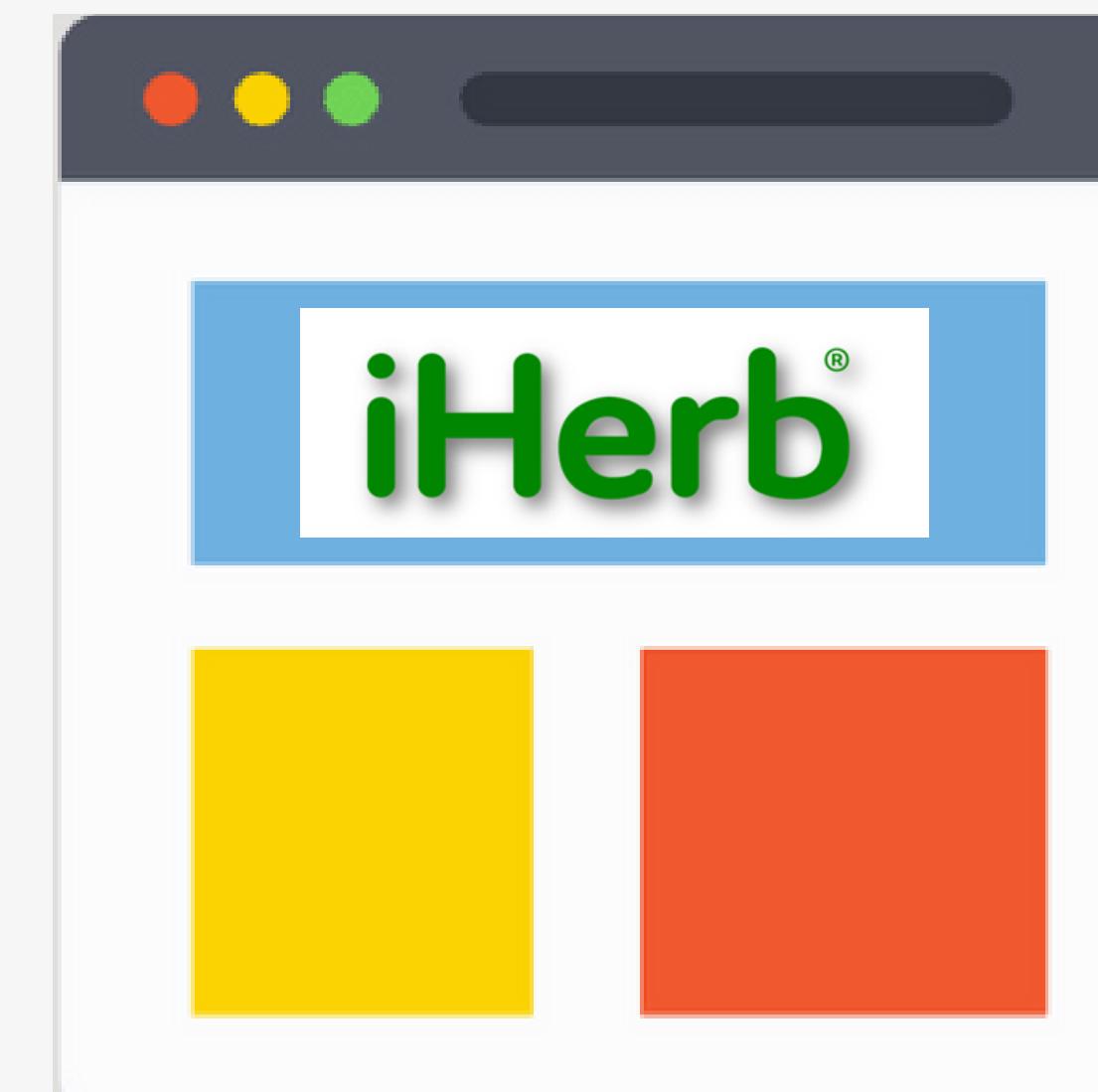
Problem Statement

How can we determine the price of a product ?



DATASET

Web Scraping



2641 Products

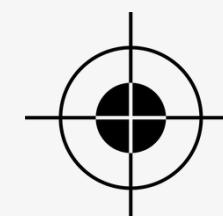
DATASET



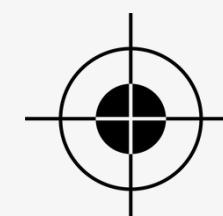
Name: the name of the product

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.



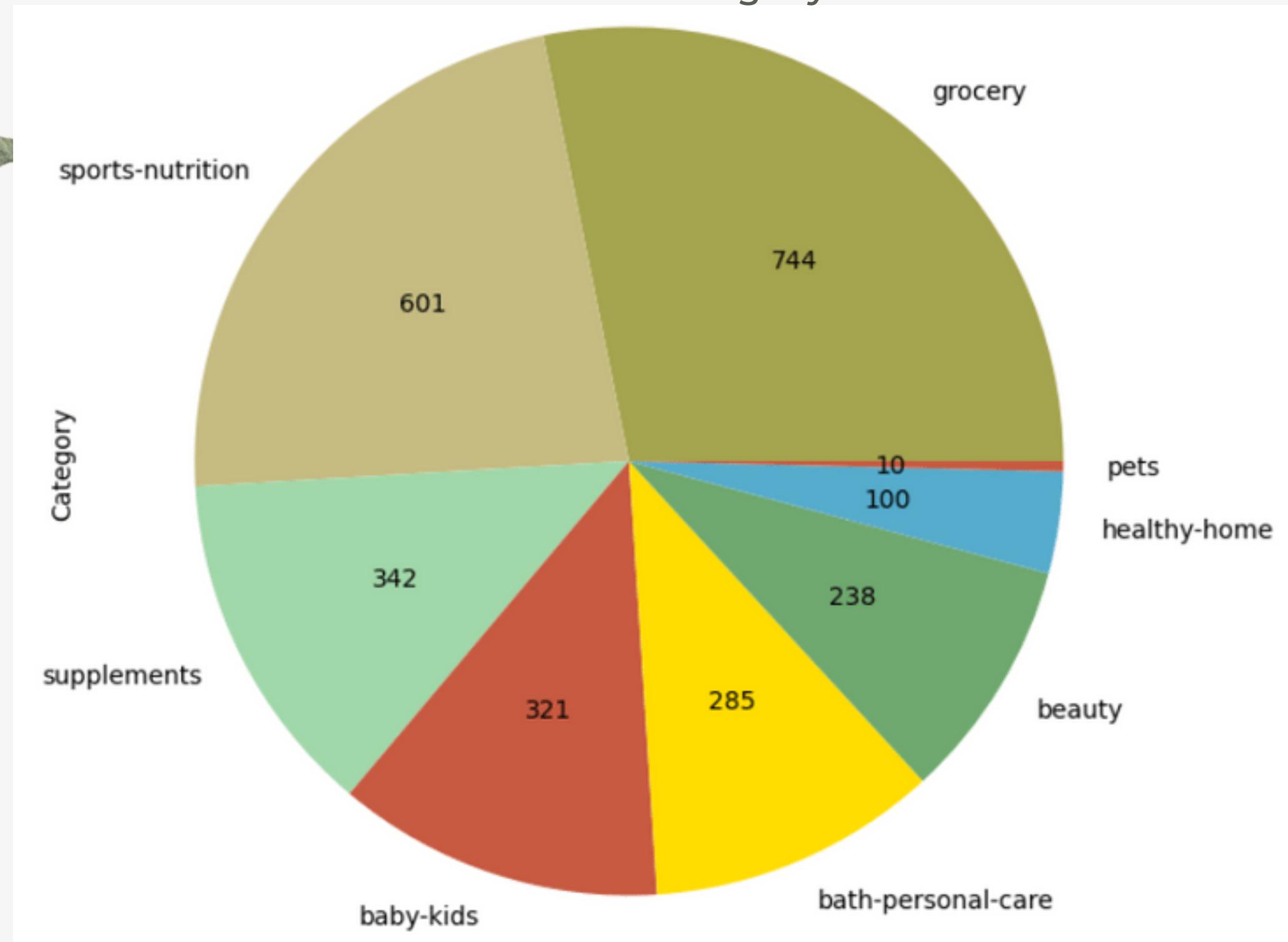
Reviews: the number of customers reviews on the product

Price: product cost

DATASET

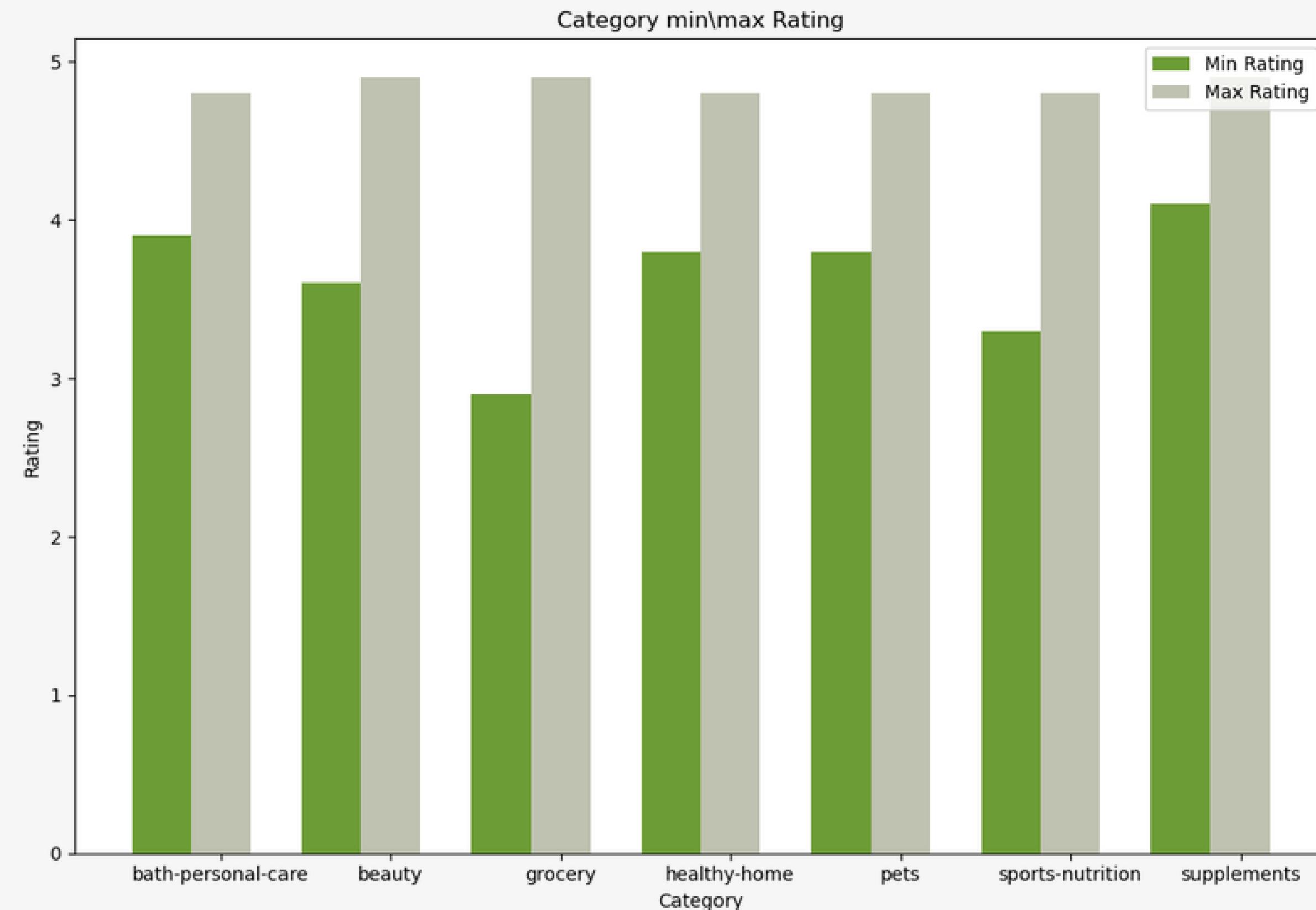
Exploratory data analysis

Products category



DATASET

Exploratory data analysis



DATASET

Exploratory data analysis

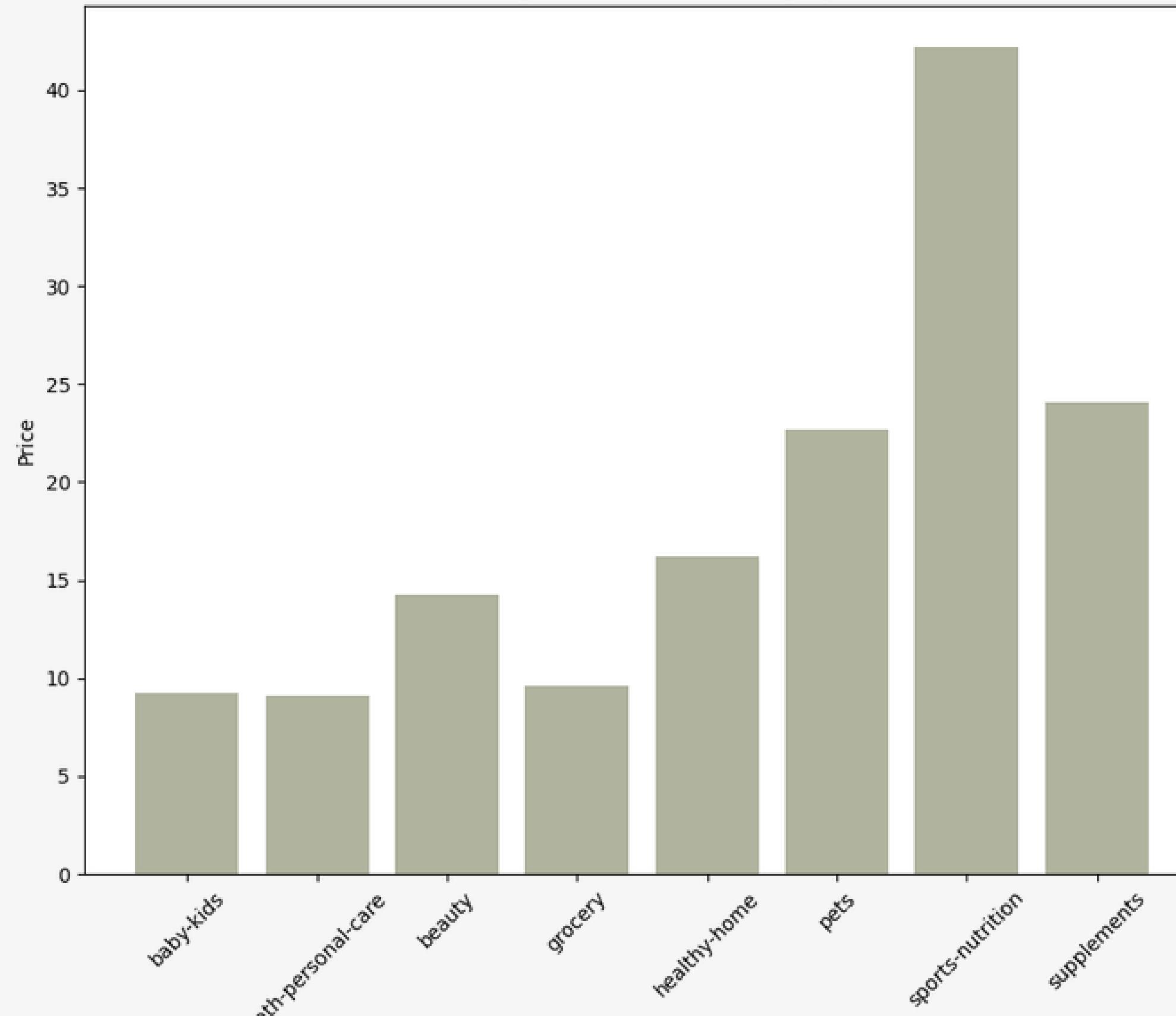
Products category vs Rating



DATASET

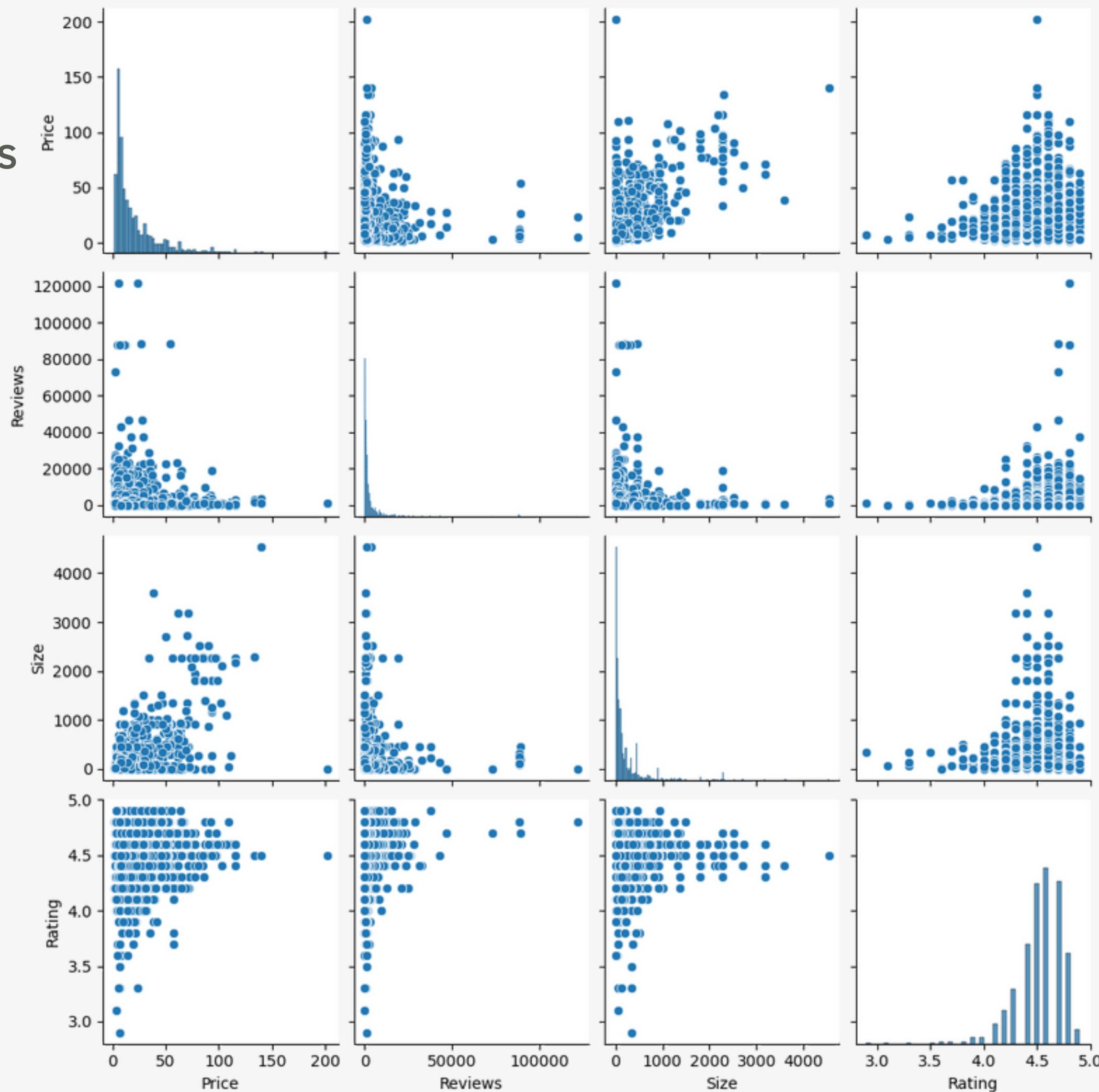
Exploratory data analysis

Average Price for each Category



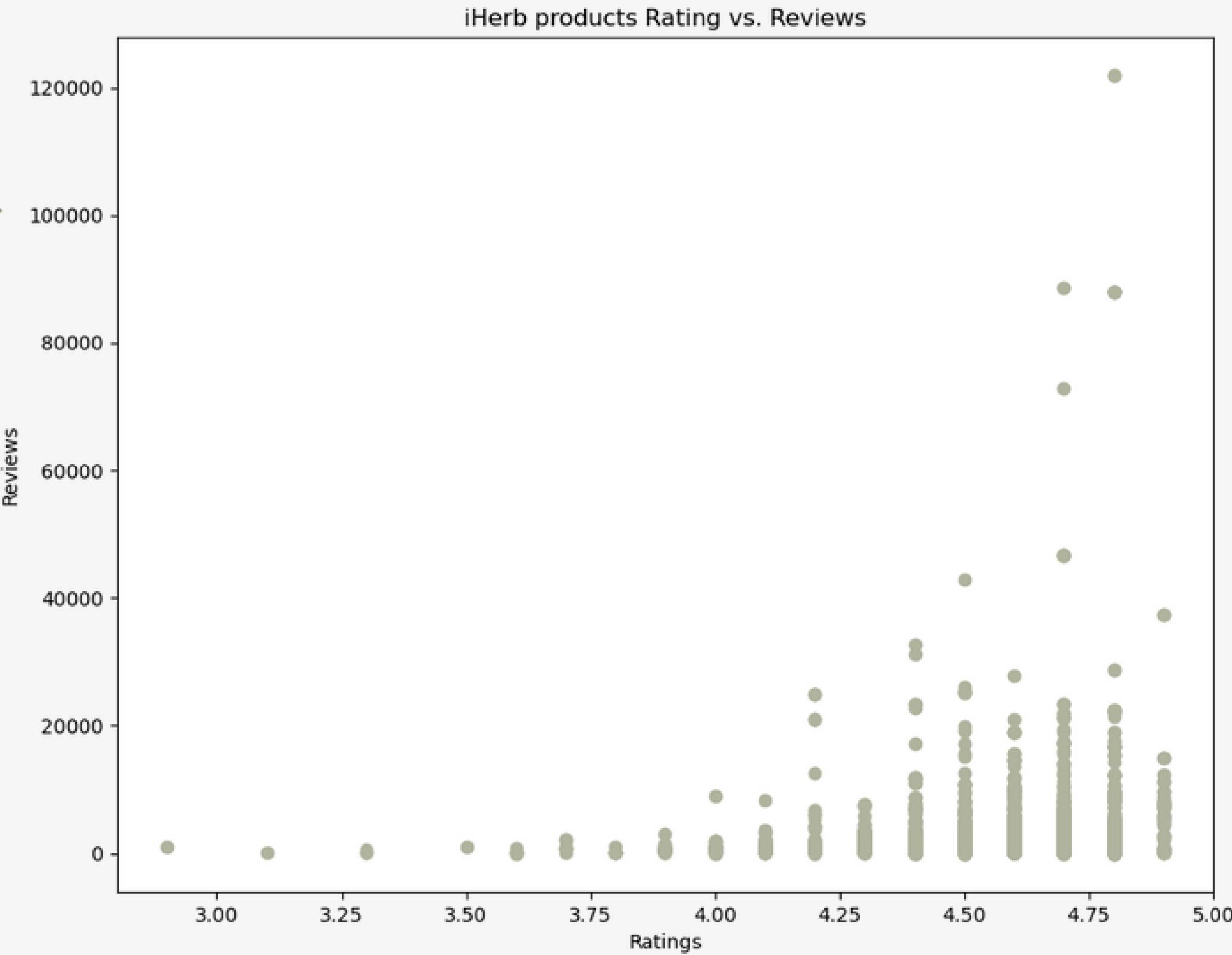
DATASET

Exploratory data analysis



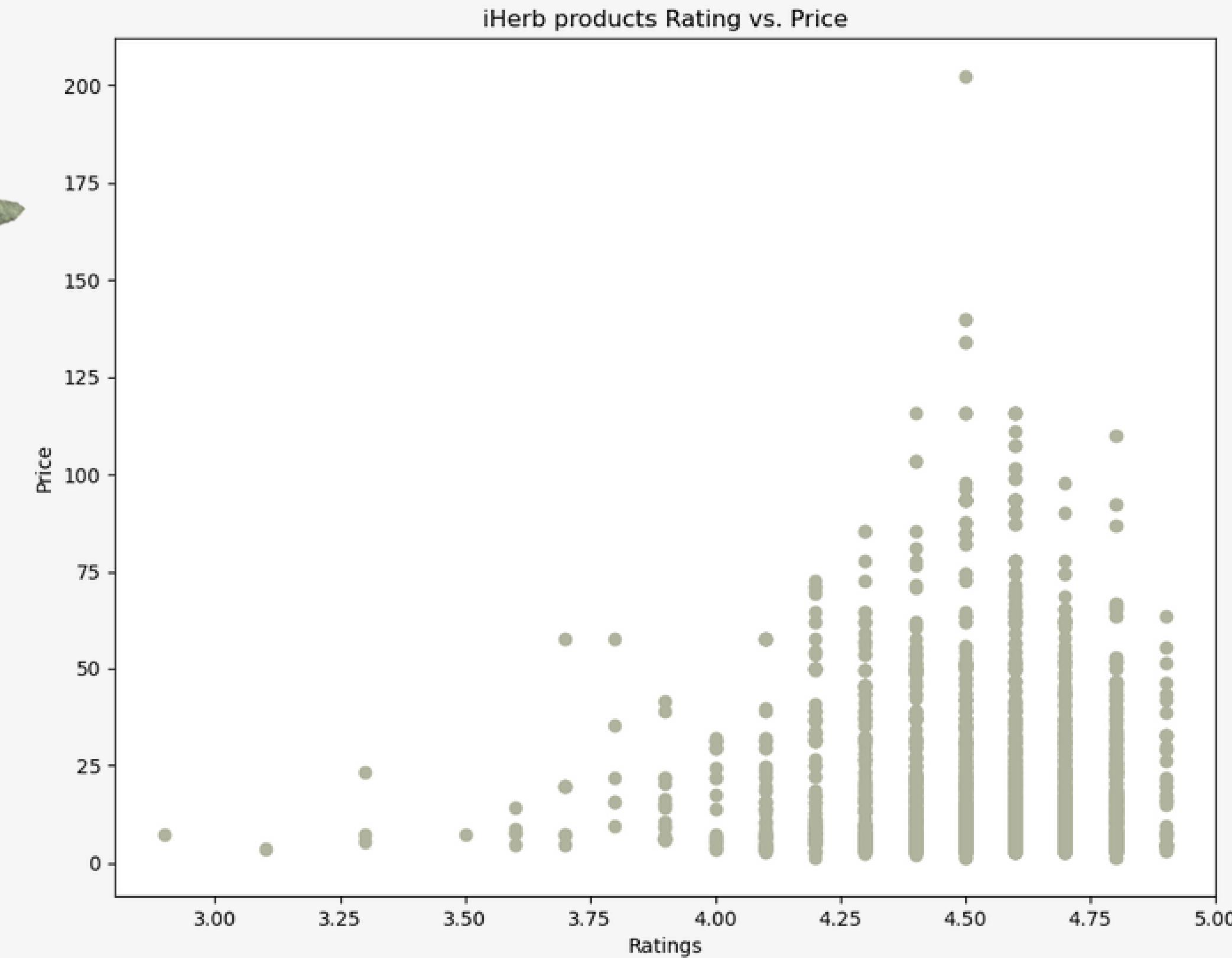
DATASET

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

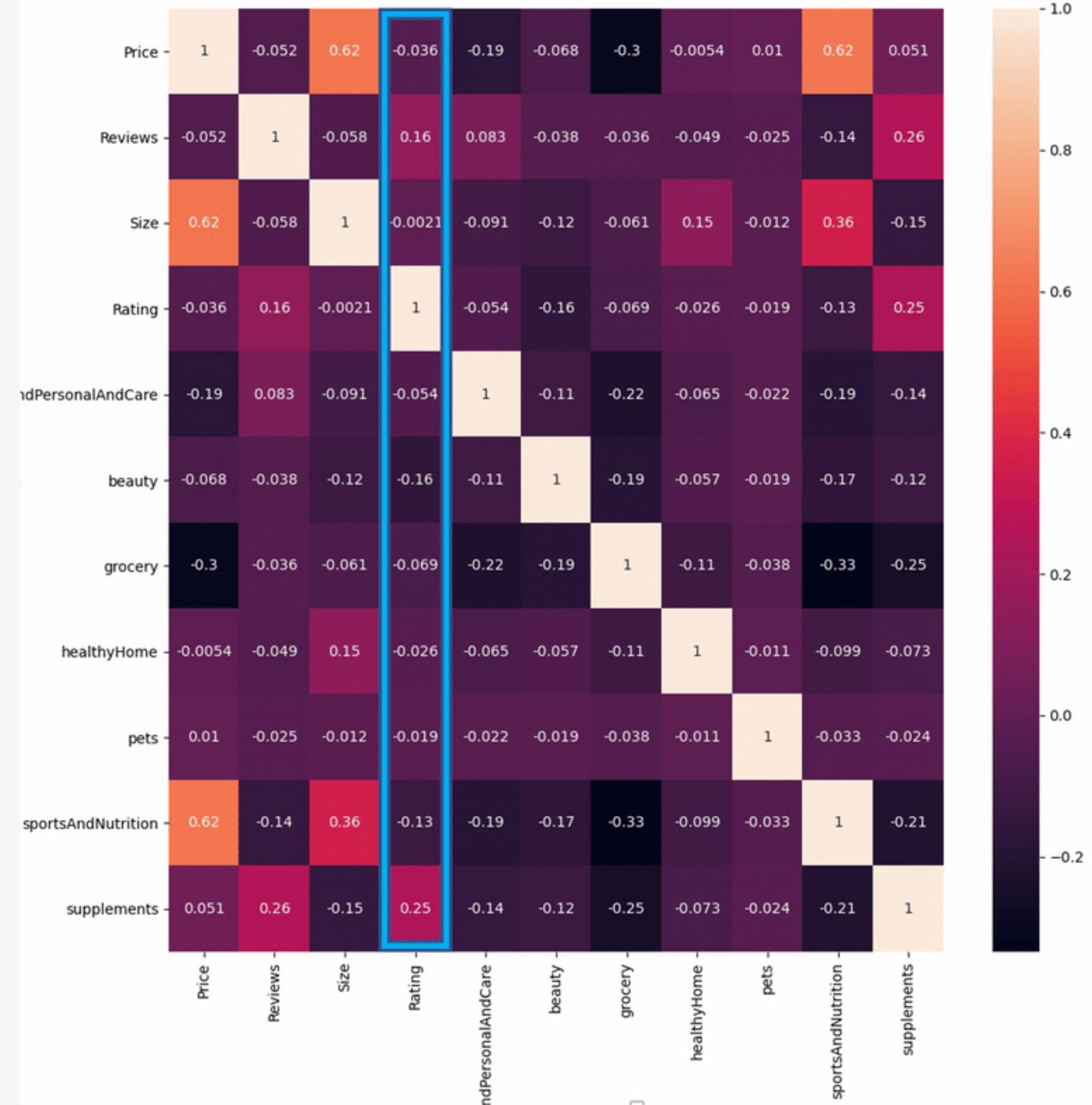
	Price	Reviews	Size	Rating
count	2641.000000	2641.000000	2641.000000	2641.000000
mean	19.519345	2972.300644	219.787373	4.536350
std	20.446082	7411.845875	420.341065	0.212801
min	1.550000	2.000000	0.000000	2.900000
25%	6.220000	442.000000	22.000000	4.400000
50%	11.080000	1068.000000	80.000000	4.600000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



DATASET

Exploratory data analysis

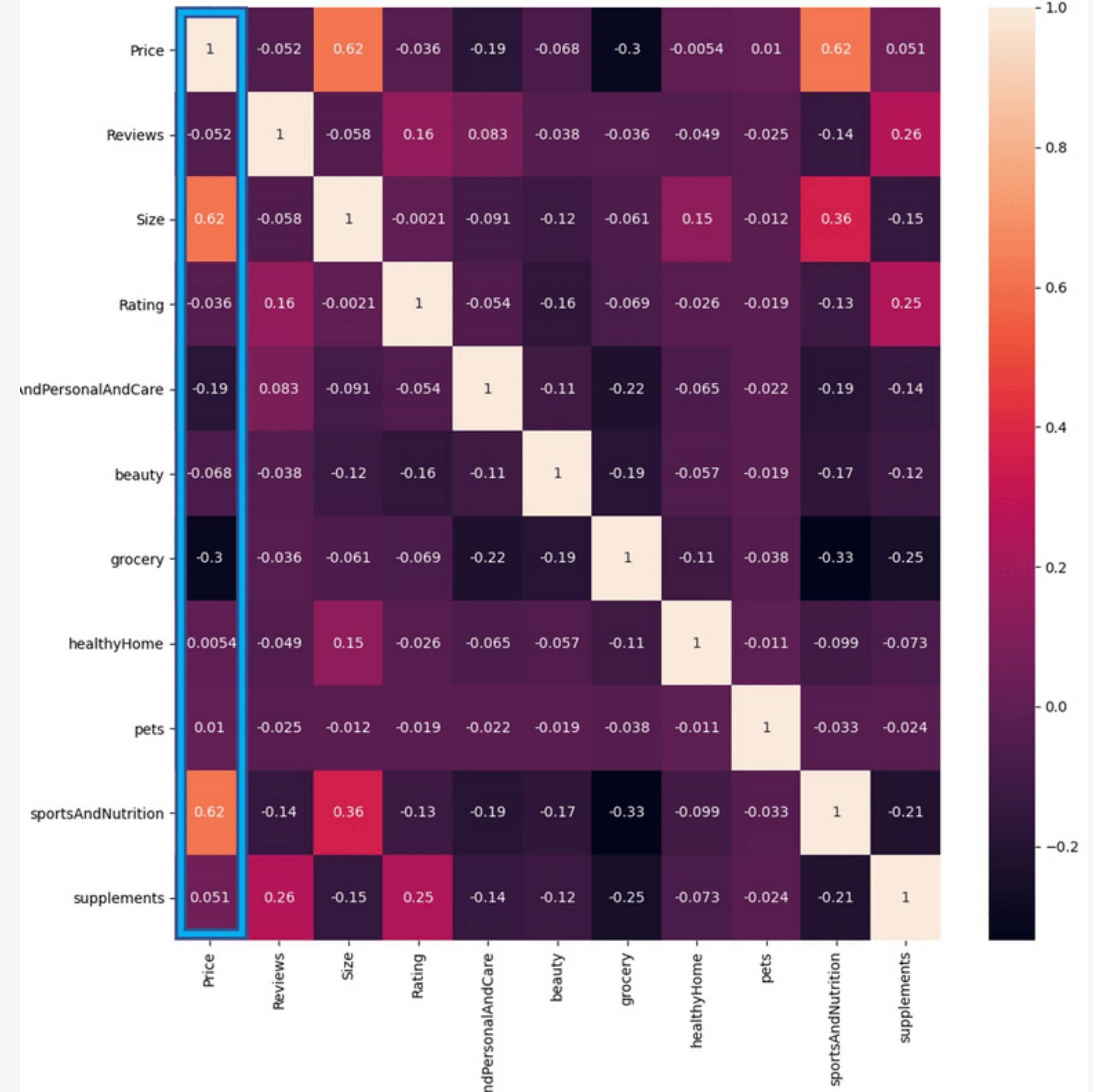
Features correlation



DATASET

Exploratory data analysis

Features correlation



DATASET



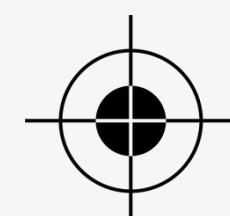
Name: the name of the product

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

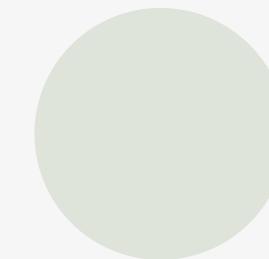


- 1 EXTRACT PRODUCT SIZE FROM ITS NAME
- 2 CONVERT CATEGORICAL VARIABLES TO DUMMY VARIABLES
- 3 SCALE PRODUCTS SIZE TO GRAMS
- 4 DROP ROWS WITH NULL VALUES

MODELING



LASSO Regression



SPLITTING THE DATASET

49% Training - 21% Validation - 30% Testing

MODELING

LASSO Regression

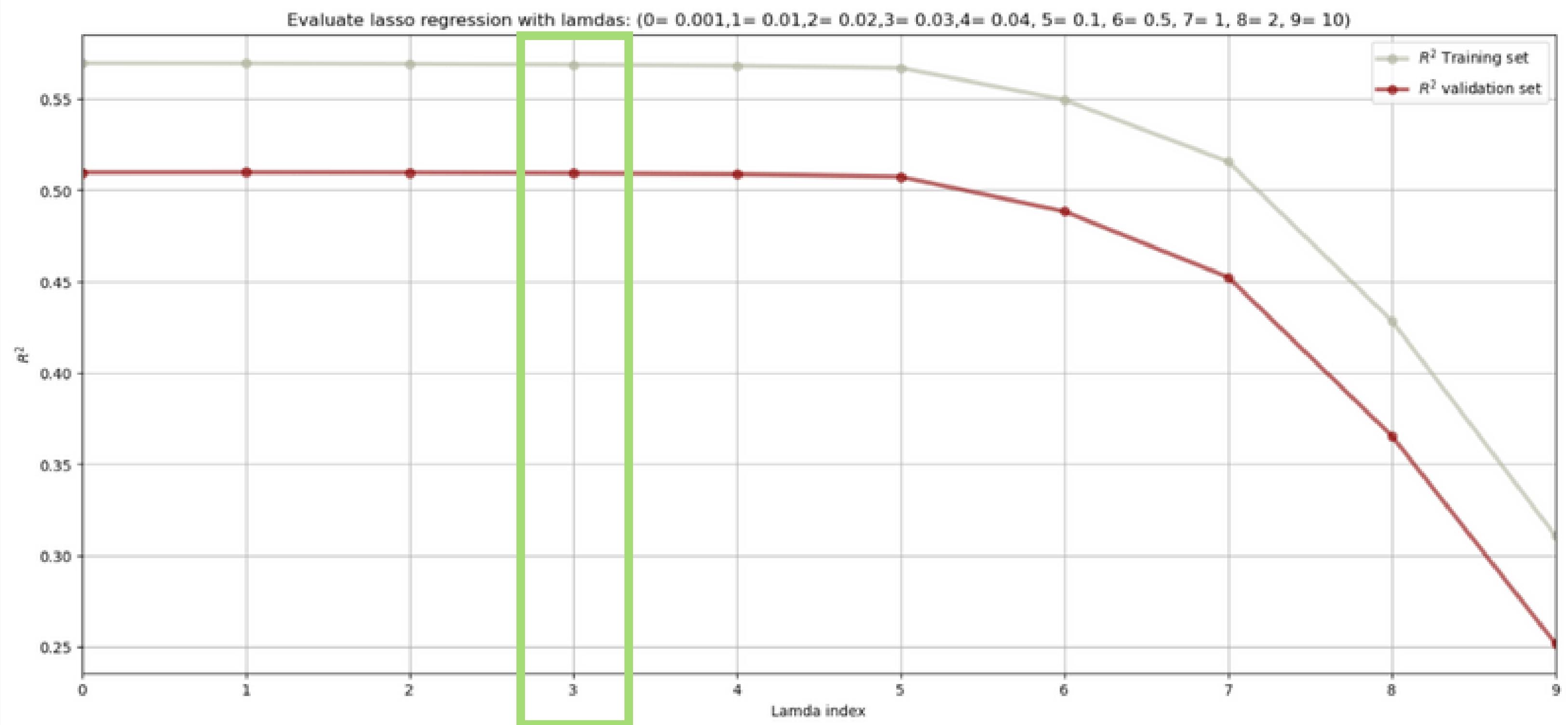
WHAT VALUE OF ALPHA SHALL I USE?



MODELING

LASSO Regression

EVALUATE WITH MULTIPLE VALUES

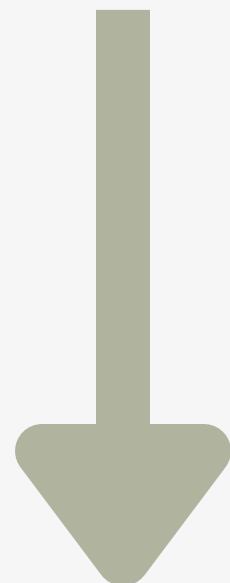


MODELING

LASSO Regression



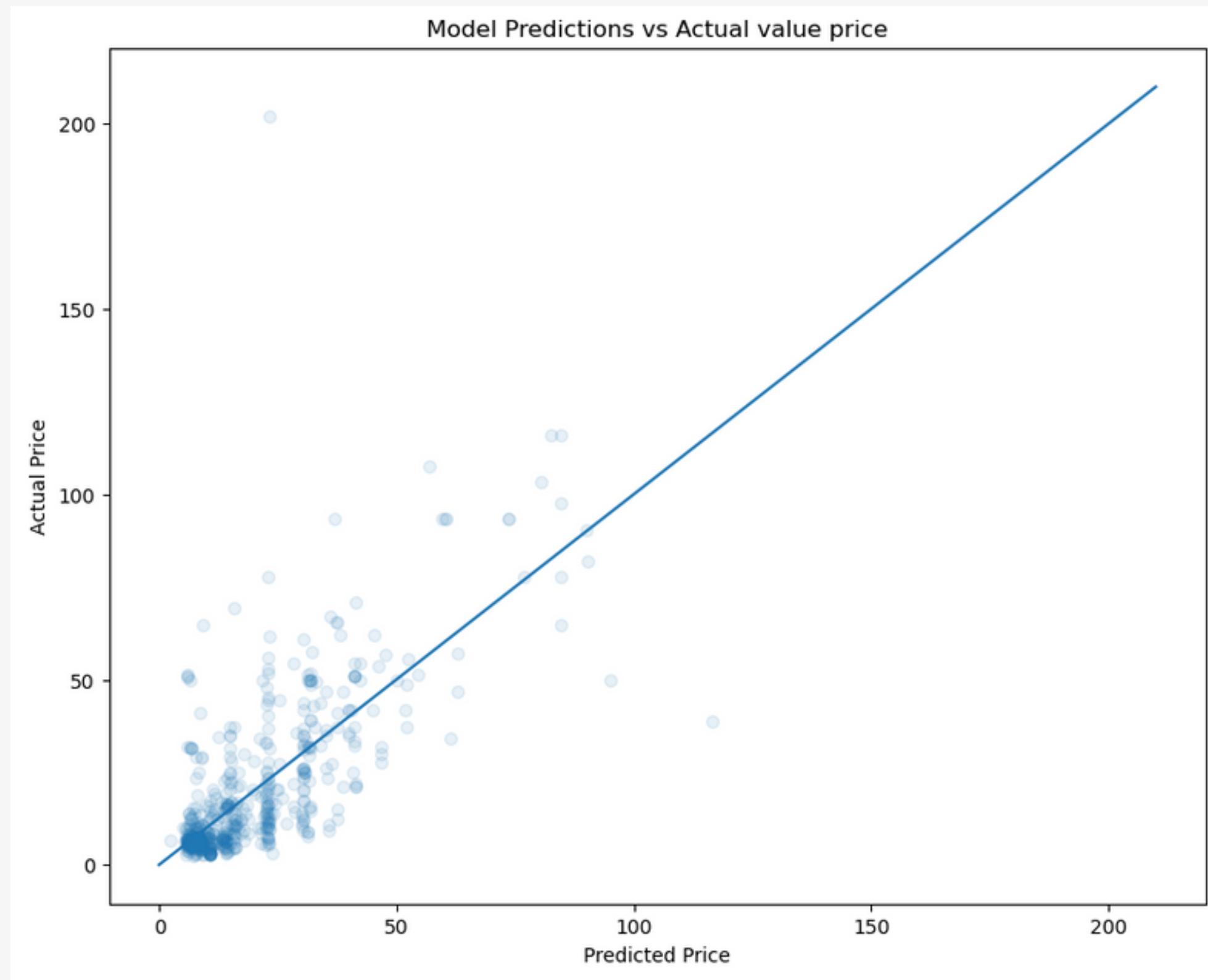
LASSO 5 CROSS VALIDATION



BEST ALPHA VALUE = 0.033

MODELING

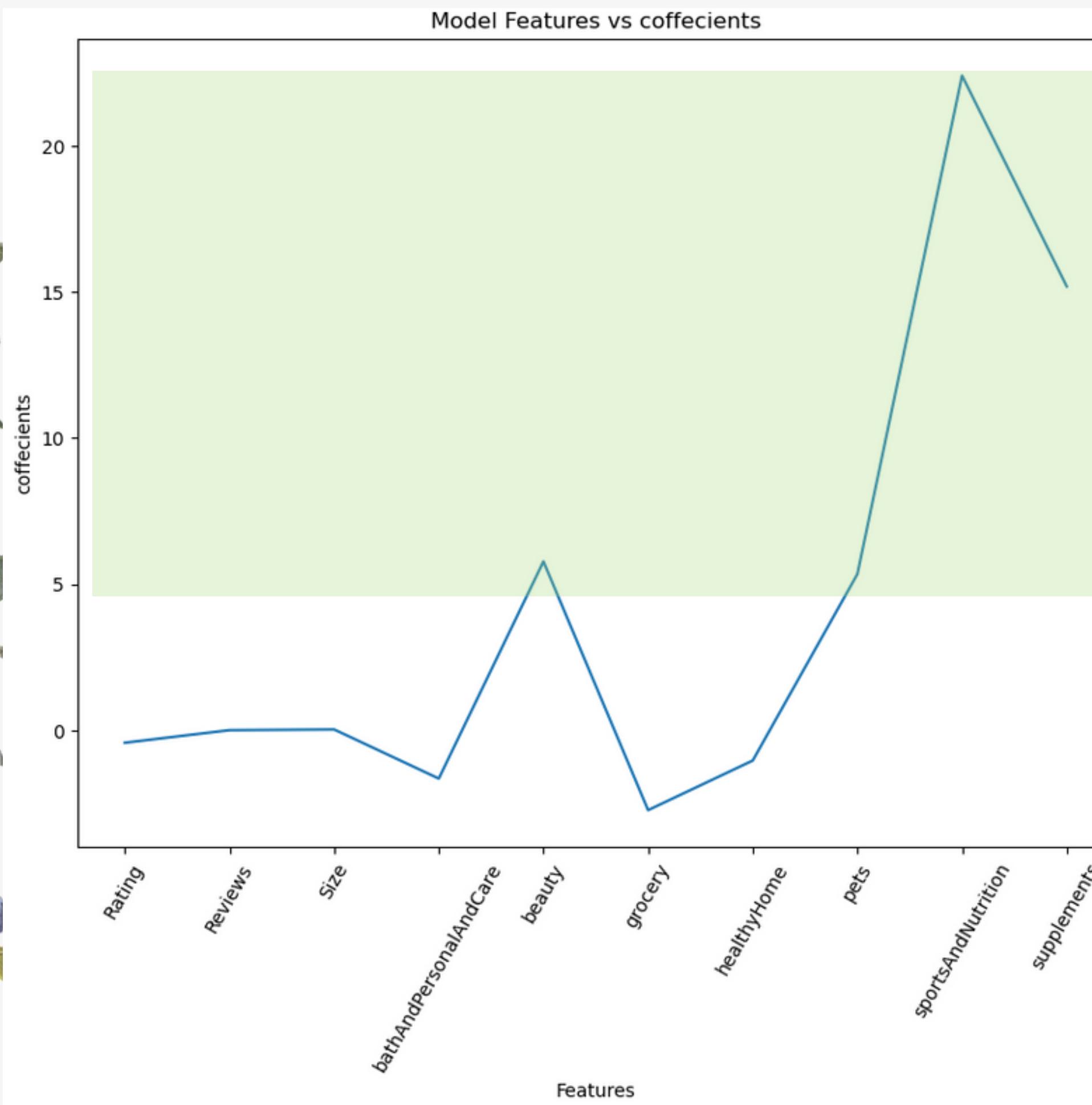
LASSO Regression



R² SCORE ON VALIDATION
SET: 50.484%

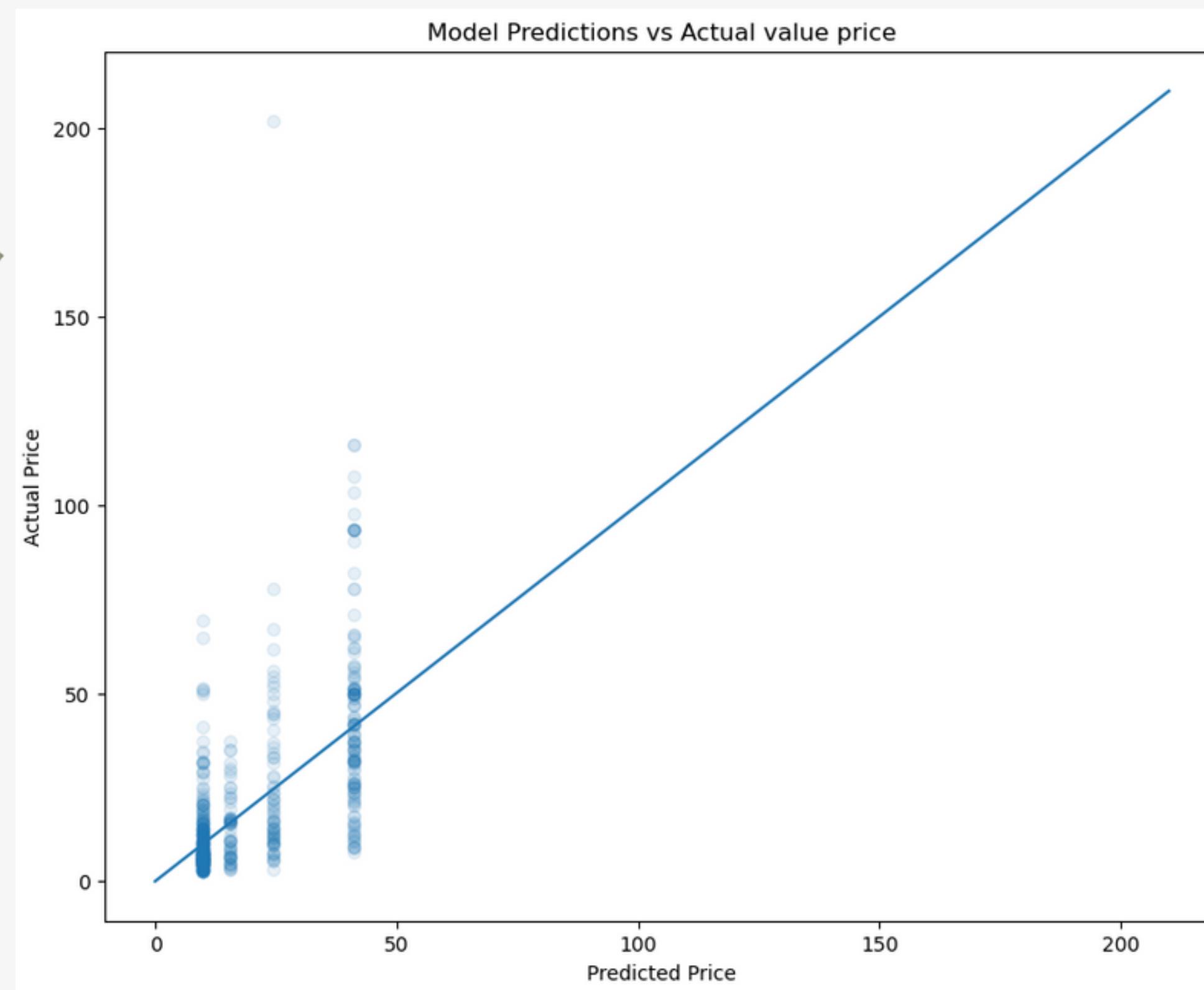
MODELING

LASSO FOR FEATURE SELECTION



MODELING

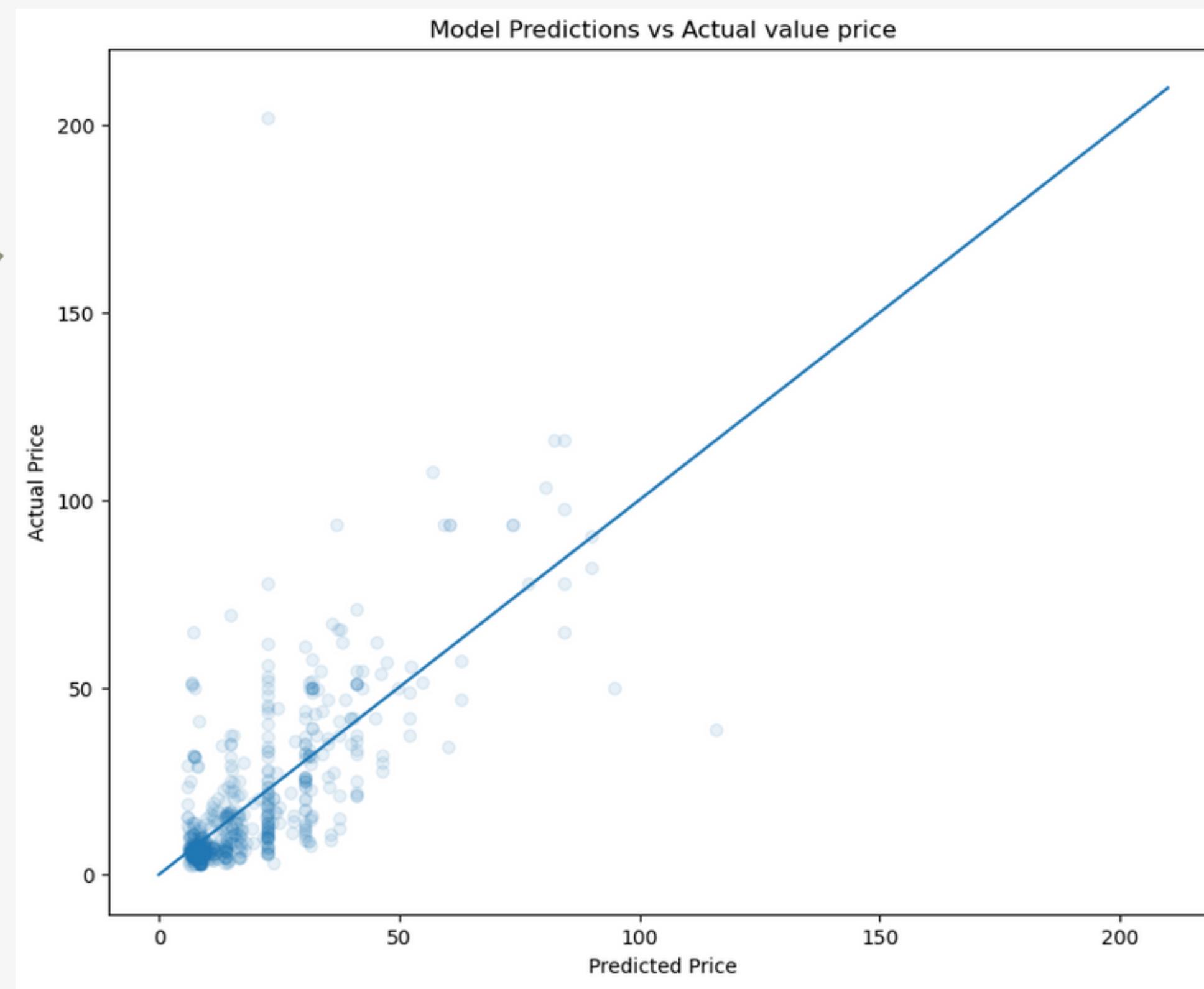
LASSO Regression on selected high coefficient features



R² SCORE ON VALIDATION
SET: 37.908%

MODELING

LASSO Regression on selected high coefficient and correlation features



R² 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



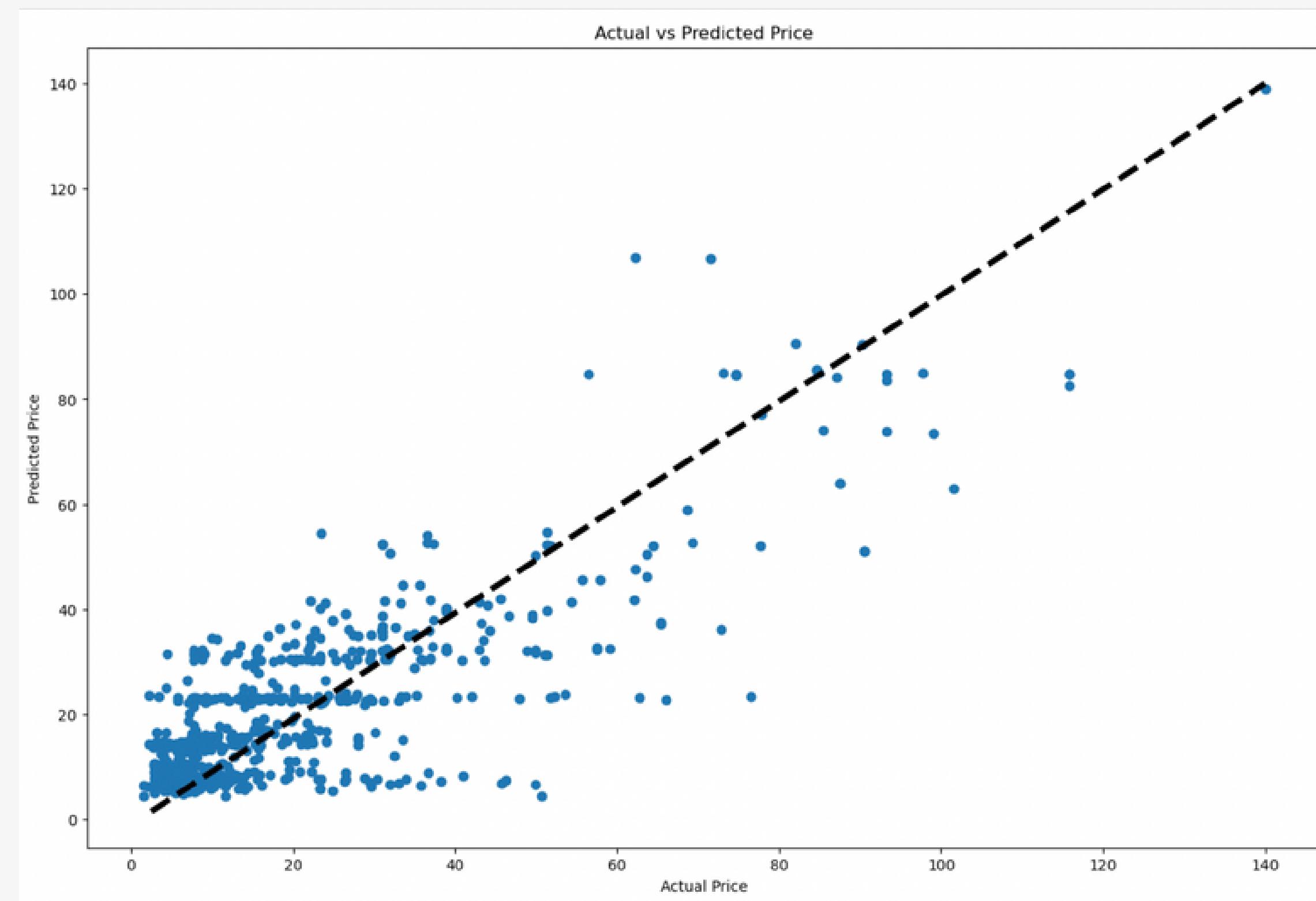
2

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

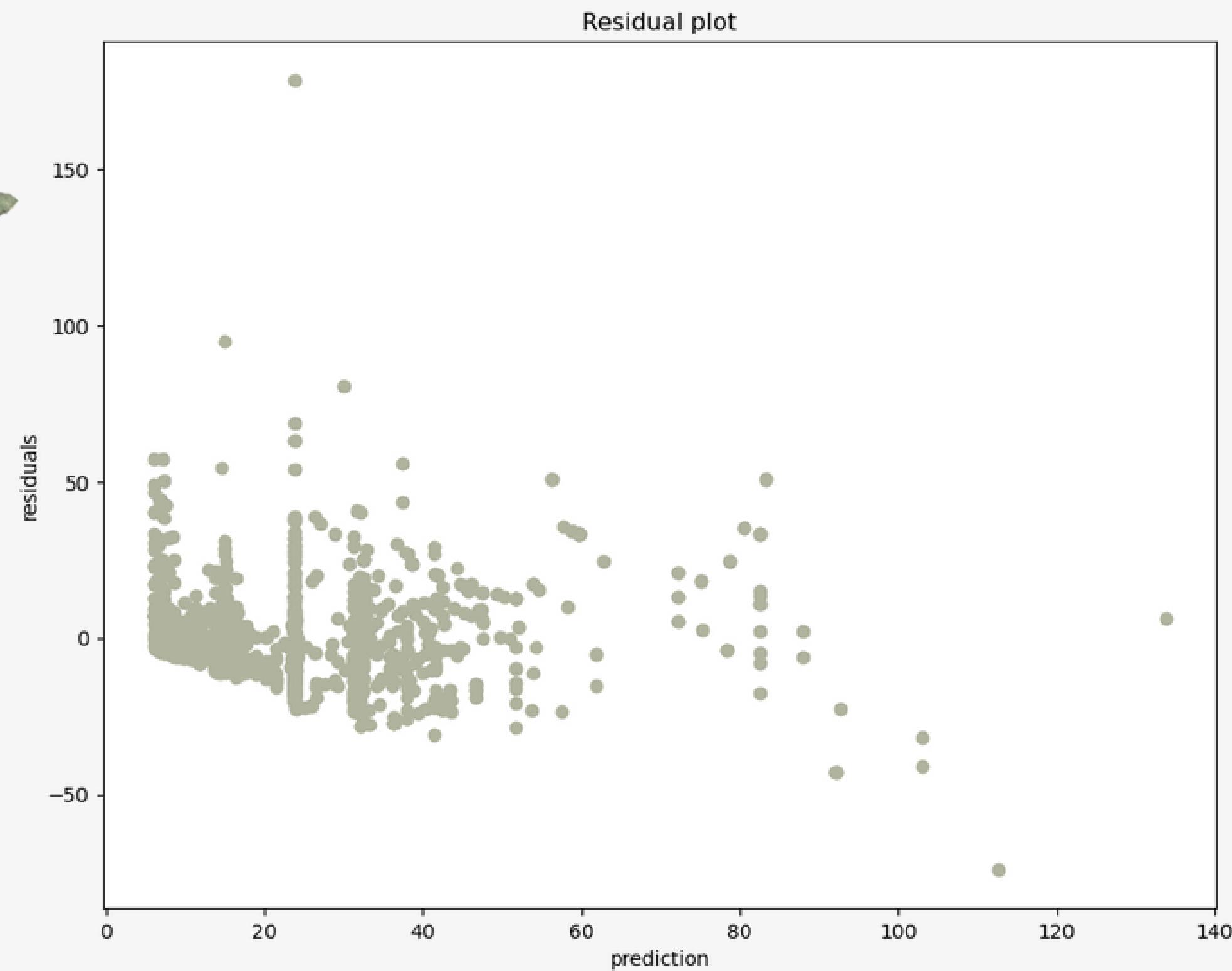
Assumption 1: regression is linear in parameters
and correctly specified

2

Assumption 2: residuals should be normally distributed
with zero mean

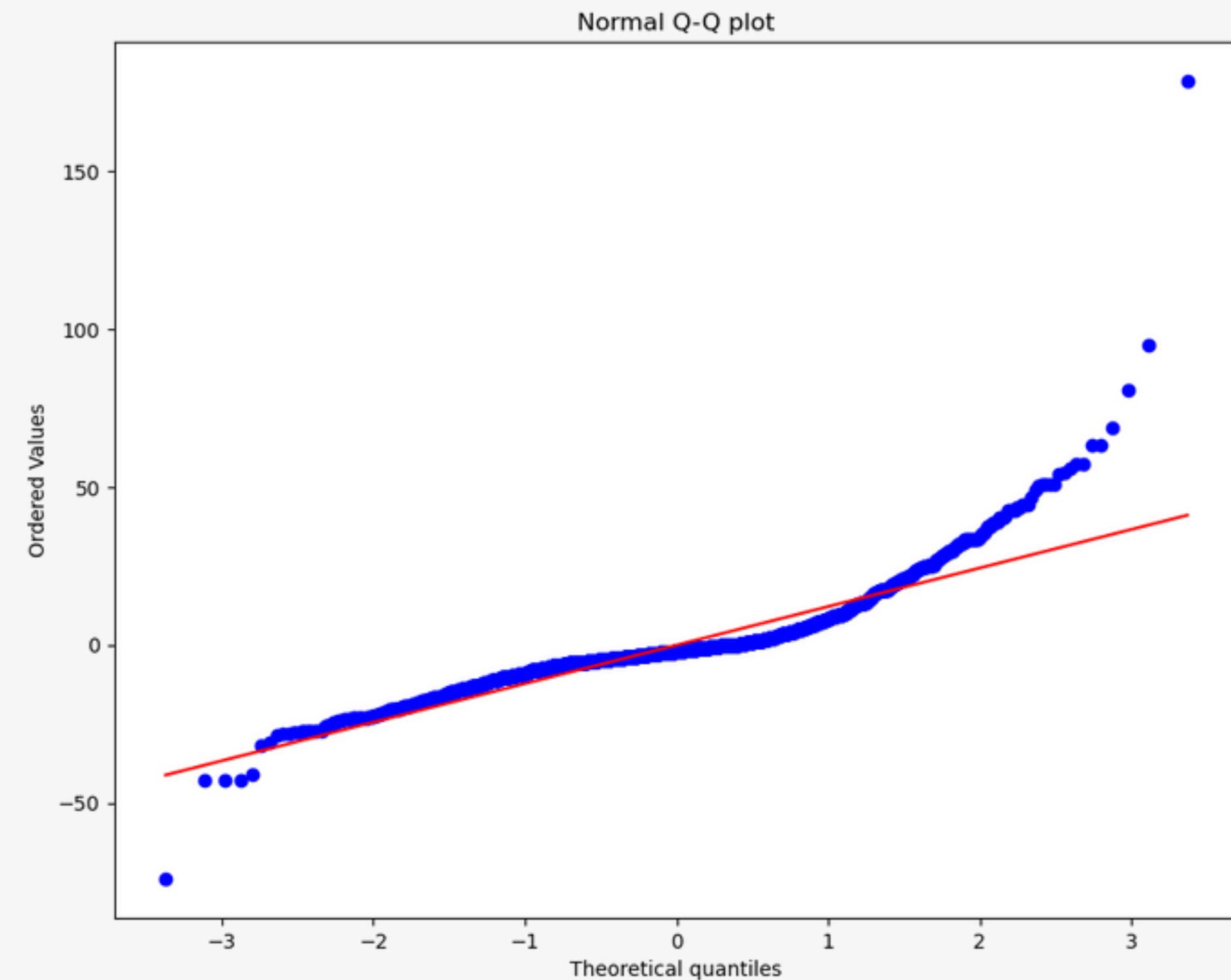
MODELING

Assumptions



MODELING

Assumptions



MODELING

Assumptions

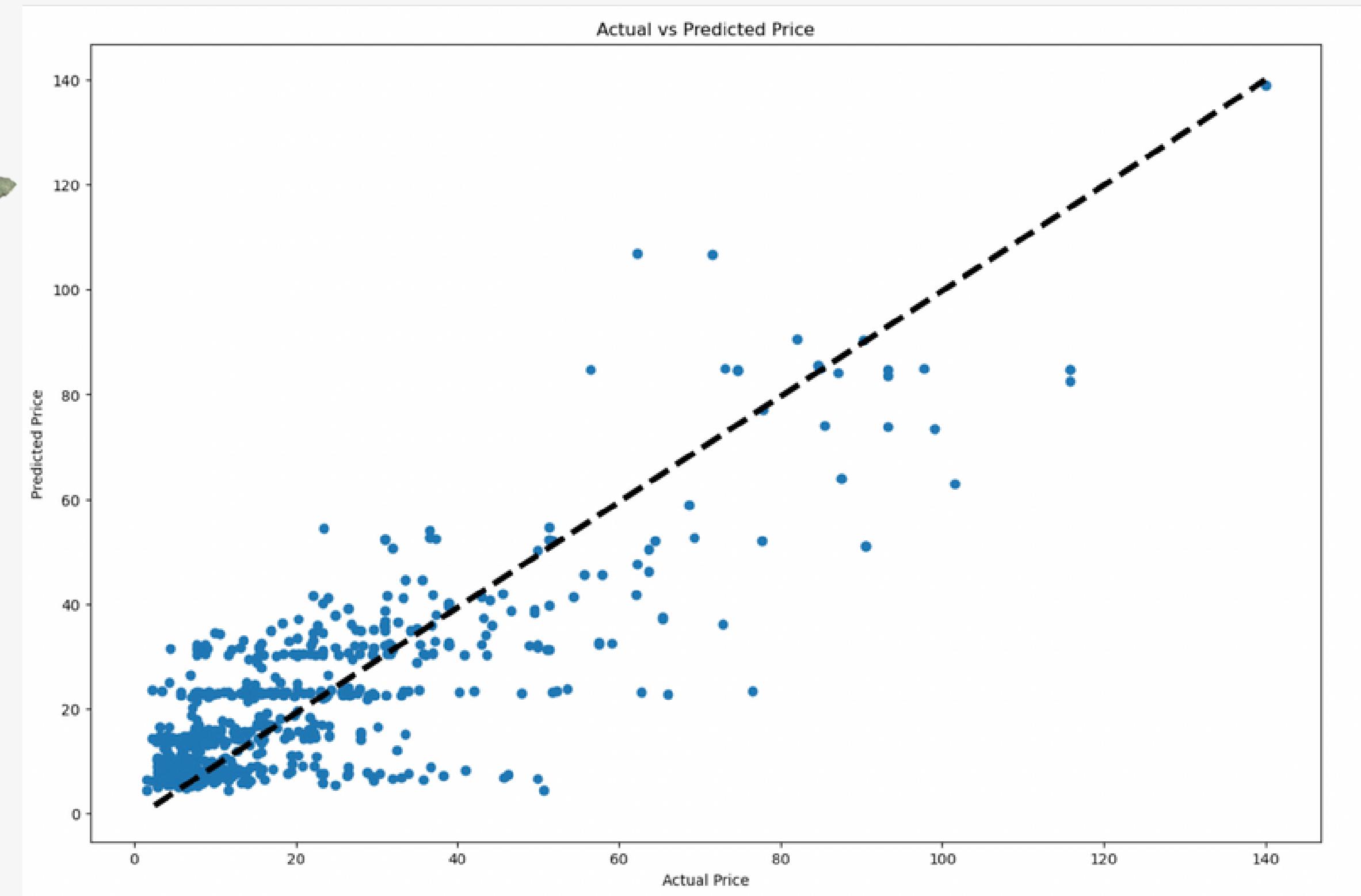
A detailed botanical illustration of an olive branch. It features several green leaves with prominent veins and small clusters of olives at different stages of ripeness, from green to dark blue.

3

Assumption 3: error terms must have constant variance

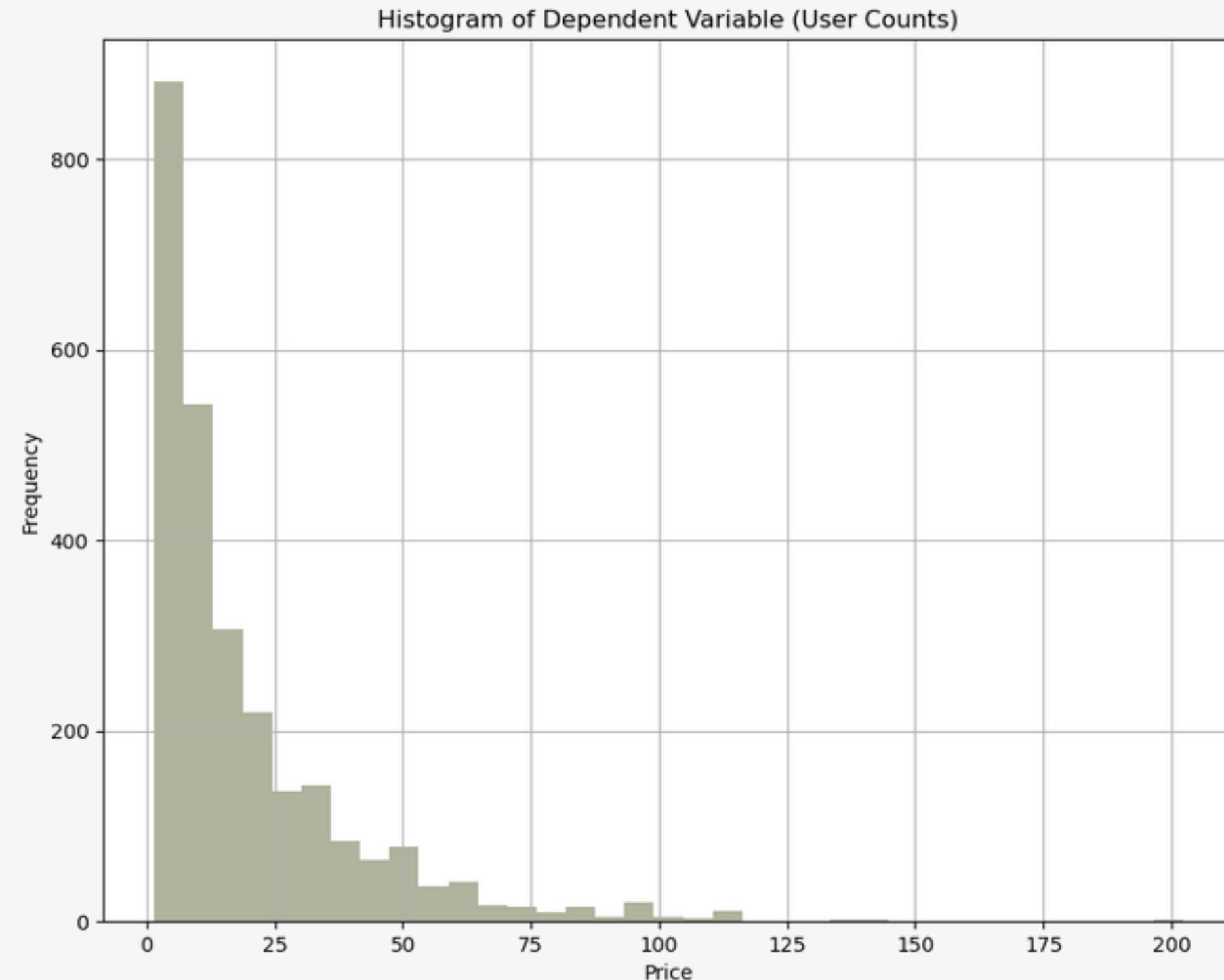
MODELING

Assumptions



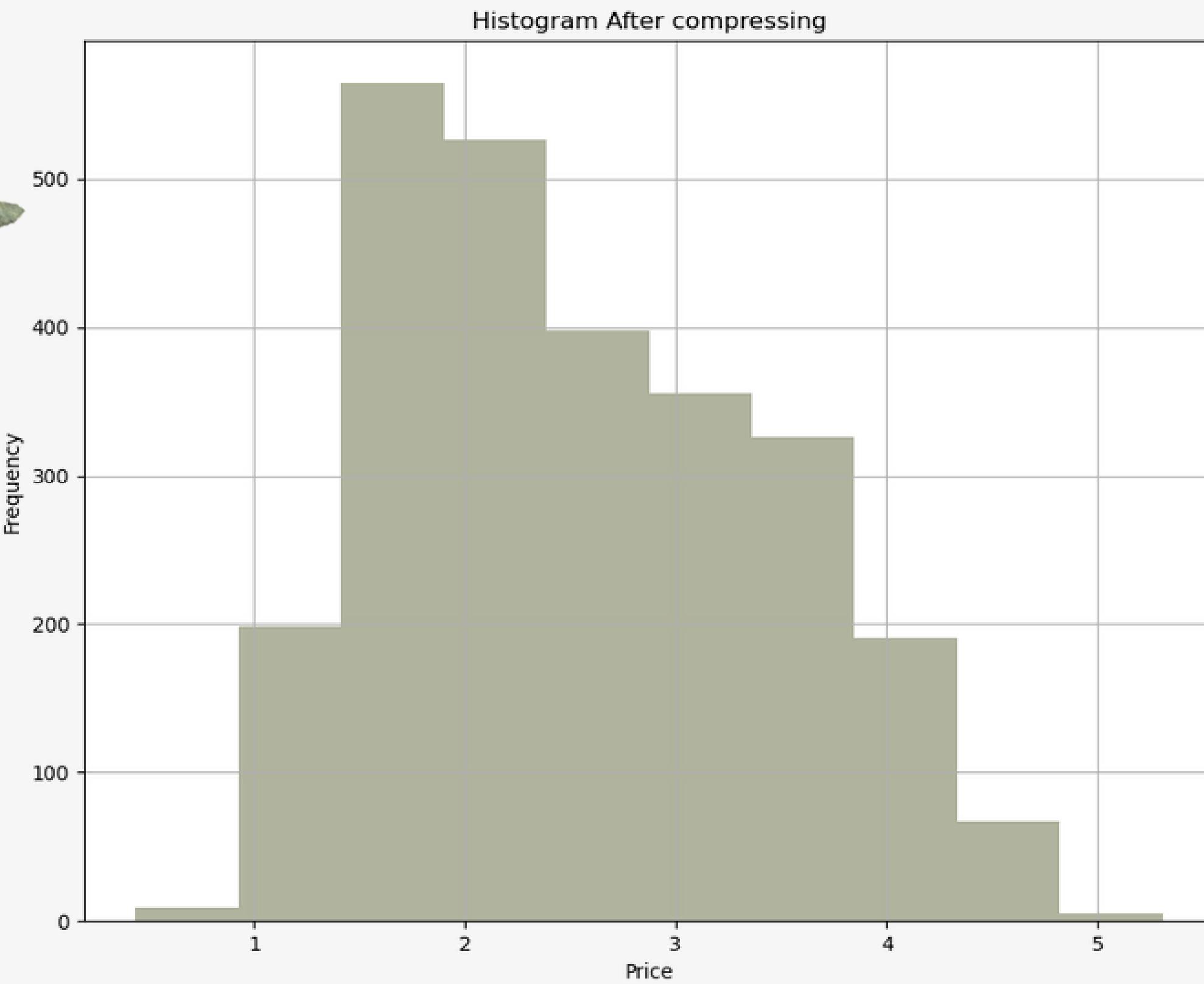
MODELING

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 multicollinearity)



THANK YOY

Fatimah AlShammari

Sara AlAbdulsalam