



# Nutrition



By: Amirah Alotaibi, Amal Altamran , Alanoud Alhussain

# Introduction

The nutrition industry is currently experiencing huge growth. Increasing awareness of how diet affects health.

Set against a backdrop of busy lifestyles and global health concerns, the nutrition market looks set to skyrocket as people look for quick and easy solutions.

From this point we get inspiration to talk about nutrition and health grocery market and predict price of it.

By doing web scrubbing from IHERB.com



# workflow

1- scraping our data

2- cleaning the data

3- linear regression model

4- feature engineering

5- best model

6- Testing model



# Scraping Data



- We scraped 3000+ product by using BeautifulSoup



# Data Feature

**Name:** the name of the product.

**Class\_info:** the name of the category the product belongs to (Sports-Nutrition, Grocery)

**Weight:** the size measurement of the products

**Ratings:** the average rate of the product.

**Reviews:** the number of customers reviews on the product



# Target

Price: product cost



# Tools

**BeautifulSoup**

Collect the data  
from website

**Pandas**

Explore and clean data

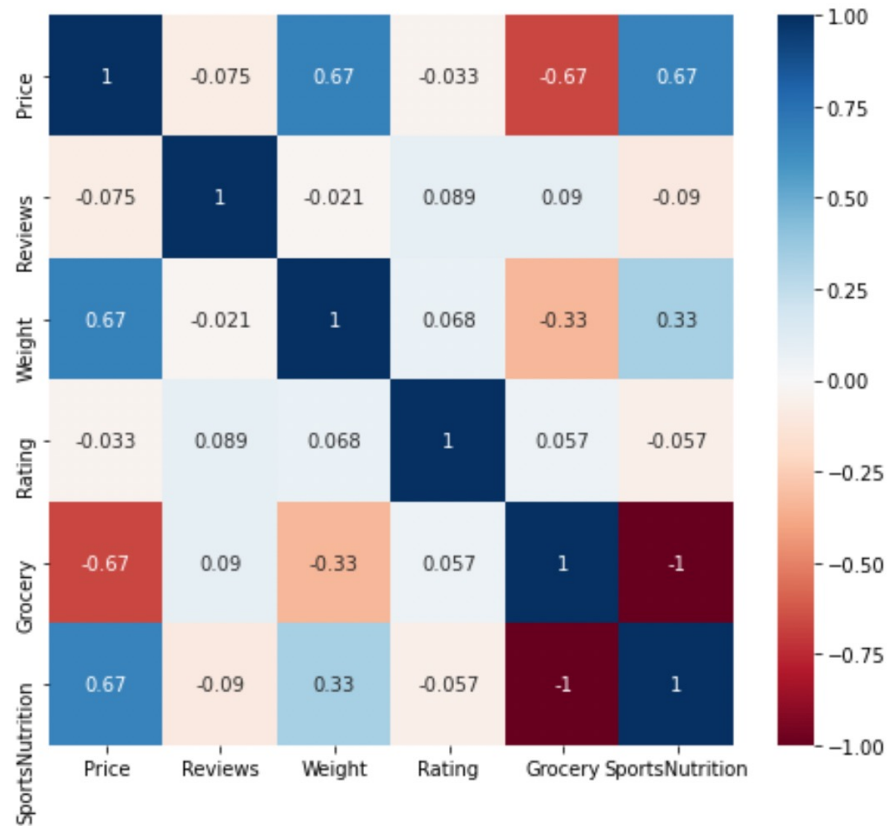
**Sklearn**

Models training

**Matplotlib and Seaborn**

Visualize data  
and model

# Correlation





# FEATURE ENGINEERING



- First score with out dummy variables for testing  $R^2$  : 0.570
- score with dummy variables testing  $R^2$ : 0.6621

# Results:



MODEL	R square (Train 60%)	R square (Val 20%)	R square (Testing)	MAE	RMSE
Linear Regression	0.682	0.679	0.684	10.0587	14.49
Polynomial degree2	0.707	0.67	0.667	9.825	13.917

# The Chosen model:



As we see here we pick linear regression depends on the minimum difference between R square and testing R square which mean the model learn from linear regression better than polynomial regression and fit the data.

# Recommendations

- The linear regression is best model we got

We prefer to :

- Add more features.
- Collect more product in the data.



# Than You

