Clothing website: Scraping & Analysis

Michael Tavoni

17 gennaio 2026

Indice

Introduzione

Il progetto si intitola "Clothing Website: Scraping & Analysis" ed è nato con l'obiettivo di realizzare un primo progetto di web scraping di livello base, utilizzando le librerie BeautifulSoup e Requests. L'ambito scelto è quello dell'abbigliamento: il sito di riferimento è un piccolo e-commerce italiano che, per ragioni di privacy, è stato mantenuto anonimo e il cui URL è gestito tramite un file `.env`. Dal sito è stato raccolto l'intero catalogo disponibile, con particolare attenzione all'estrazione delle informazioni principali di ciascun prodotto, in particolare brand, articolo, categoria e prezzo.

> Il progetto è stato sviluppato nel rispetto dei vincoli contenuti all'interno del file `Robots.txt` del sito analizzato.

La repository del progetto è stata depositata pubblicamente su github: `https://github.com/michaeltavoni/clothing-website-scraping`

1 Dataset

Il dataset contiene 1,266 righe e le seguenti 4 colonne:

1. **Brand** (categorica): variabile che contiene il brand del capo d'abbigliamento;
2. **Item** (string): variabile che contiene una breve descrizione del capo d'abbigliamento;
3. **Category** (categorica): variabile che contiene la categoria del capo d'abbigliamento;
4. **Price** (quantitativa): variabile che contiene il prezzo del capo d'abbigliamento.

La variabile **Category** è stata costruita come previsto nel processo di *feature engineering*: mediante l'utilizzo di espressioni regolari sono state definite regole di astrazione per ricondurre i singoli articoli alle rispettive categorie. Tuttavia, a causa di alcune descrizioni poco informative, non è stato possibile individuare una categoria per tutti i prodotti; di conseguenza, la variabile presenta alcuni valori mancanti (`NaN`). Di seguito viene presentata l'analisi dei valori mancanti.

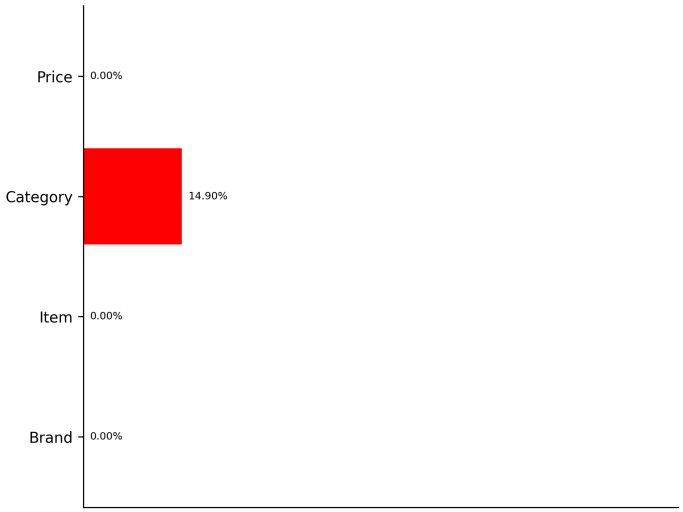


Figura 1: Nan Plot

Come anticipato, la Figura 1 mostra come l'unica colonna ad avere valori nulli sia **Category** la quale presente il 15% di valori Nan sul suo totale.

2 Analisi esplorativa

2.1 Analisi univariata

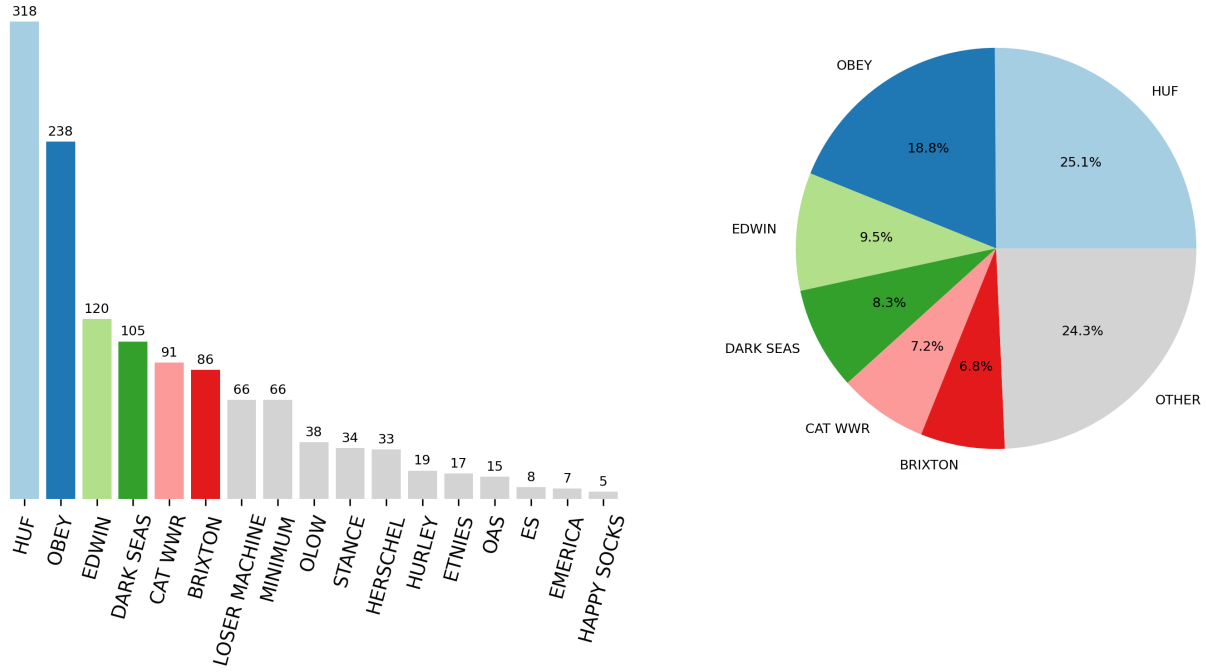


Figura 2: Brand

I 1,266 prodotti presenti sul e-commerce appartengono solamente a 12 brand. In particolare, l'80% degli articoli viene fornito da 7 brand, ovvero:

1. HUF
2. OBEY
3. EDWIN
4. DARK SEAS
5. CAT WWR
6. BRIXTON
7. LOSER MACHINE

HUF è il brand con il maggior grado di penetrazione sull'e-commerce con circa il 25% seguito da OBEY con circa il 20%. Questi due brand, con EDWIN, occupano metà dell'offerta del sito.

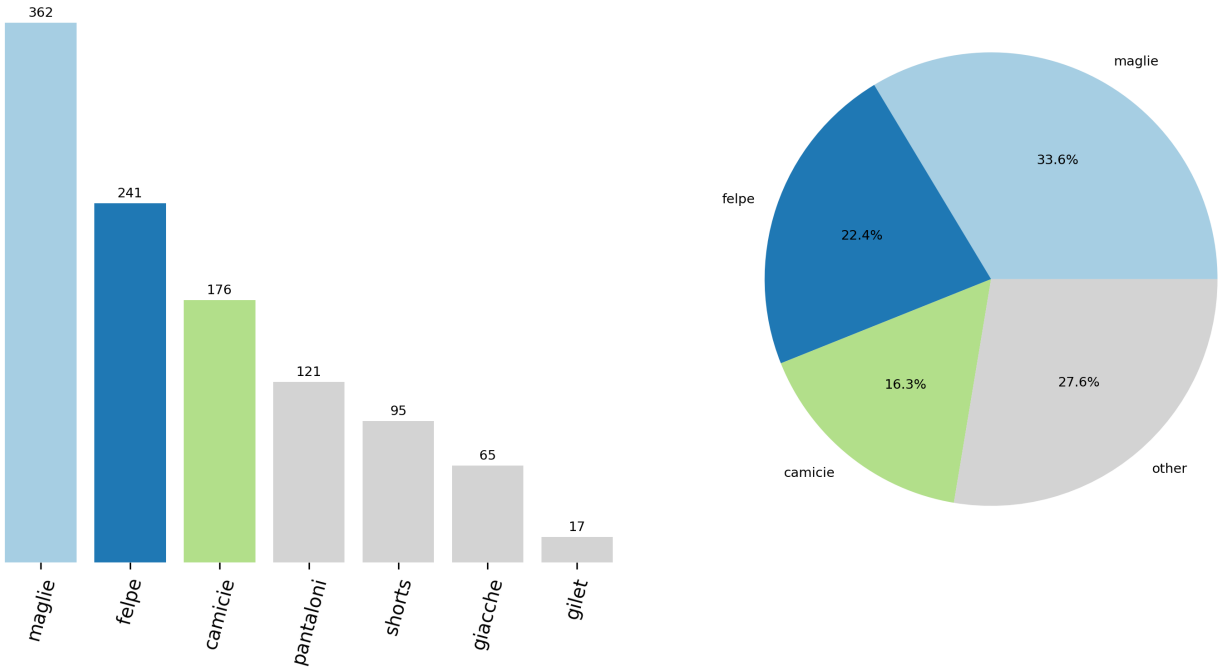
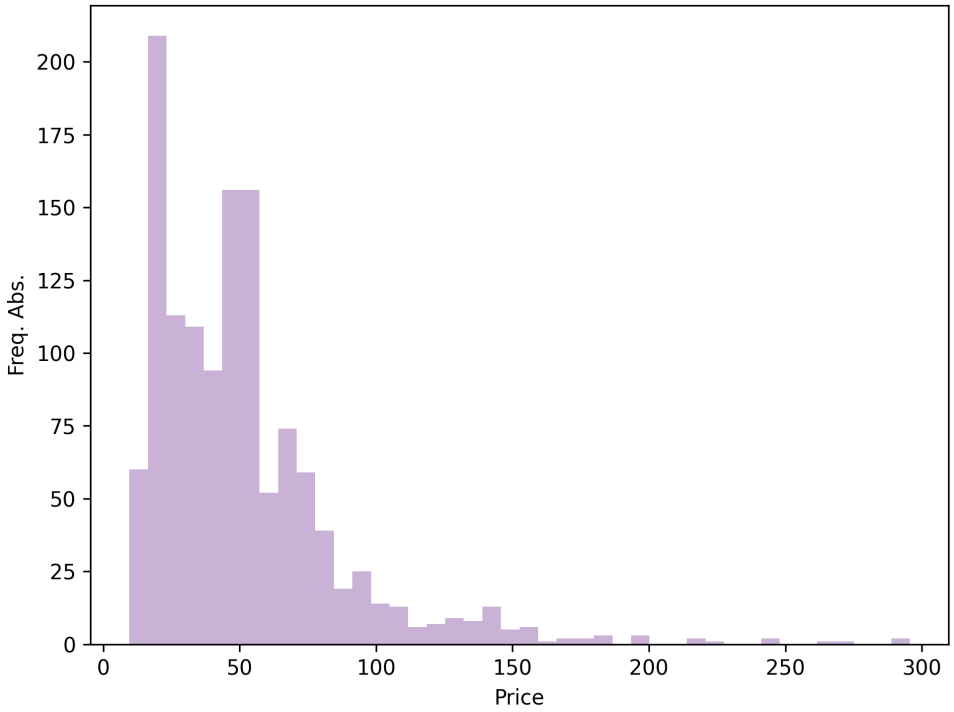


Figura 3: Category

Sull'e-commerce sono presenti 7 categorie. l'%80 percento dei prodotti fa parte di tre categorie:

1. maglie
2. felpe
3. camicie

In prevalenza vengono vendute magliette (34%) seguite dalle felpe (23%). Questi due categorie, insieme, occuppano metà dell'offerta del sito.



	Price
count	1266.000
mean	54.892
std	35.693
min	13.000
25%	29.980
50%	47.750
75%	67.250
max	299.000

Figura 4: Price hist

La Figura 4 mostra l'istogramma dei prezzi esposti sul sito di abbigliamento. Il grafico presenta una forte asimmetria positivia indicando la presenza di numerosi articoli con prezzi inferiori ai 70€ (quasi il 75%) e davvero pochi pezzi con prezzi rilevanti: oltre i 150€. In media gli articoli hanno un prezzo che si aggira intorno ai 55€/articolo. Il prezzo mediano è 48€, questo ci indica che la metà dei capi ha prezzi inferiori.

Brand	Item	Category	Price
CAT WWR	cat unit jacket	giacche	299.000
CAT WWR	cat unit jacket	giacche	299.000

Tabella 1: Max price item

Il prezzo più alto presente sullo store è di 299€. Sono presenti solamente due capi con un prezzo così alto, appartengono allo stesso brand e molto verosimilmente rappresentano lo stesso capo. Il brand è CAT WWR ed è una giacca.

Brand	Item	Category	Price
BRIXTON	basic s/s tlrt - joe blue	NaN	13.000

Tabella 2: Max price item

Il prezzo più basso presente sullo store è di 13€ appartiene ad un articolo per il quale non è stato possibile definire una categoria, è possibile che sia una t-shirt basic. Il brand in questione è BRIXTON.

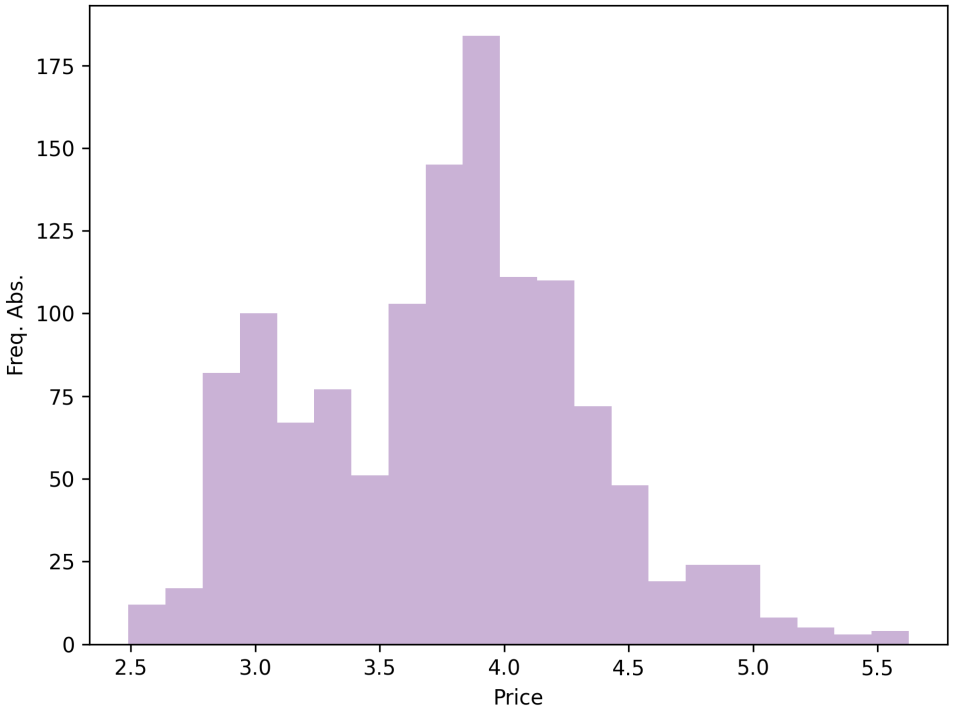


Figura 5: Log-Price hist

Visualizzando l'istogramma del log-price notiamo la presenza di due picchi nella distribuzione, questo potrebbe essere dovuto alla presenza di gruppi nei dati. Nelle prossime analisi cercherò di capire se la categoria dell'articolo possa o meno avere impatti sulla distribuzione del prezzo.

2.2 Analisi Multivariata

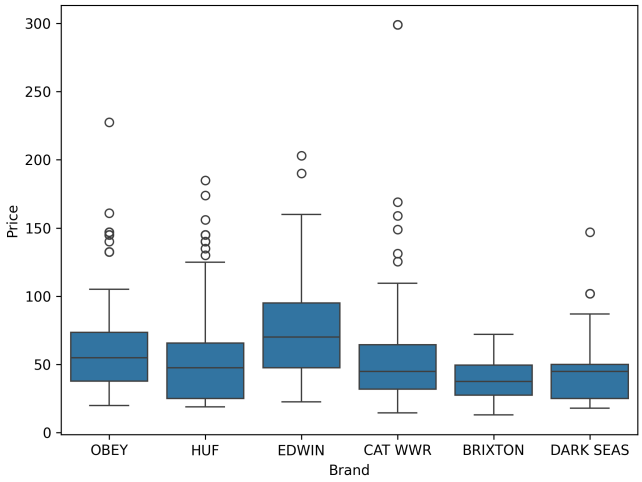


Figura 6: Boxplot brand-price

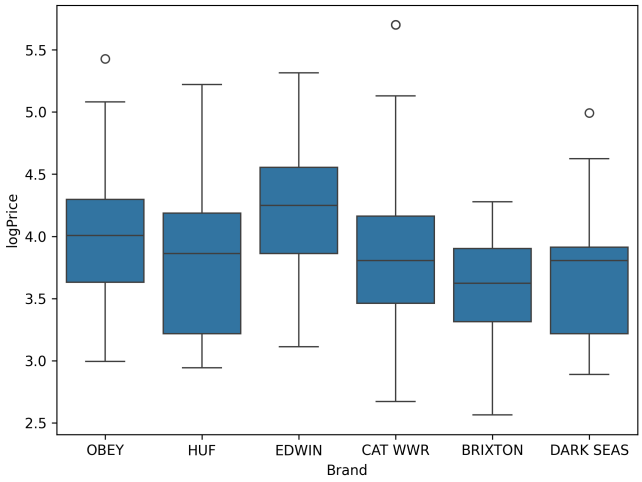


Figura 7: Boxplot brand-logprice

Brand	count	mean	std	min	25%	50%	75%	max	Price cv
BRIXTON	86.000	39.035	14.468	13.000	27.500	37.500	49.500	72.000	0.371
CAT WWR	91.000	58.186	48.252	14.500	31.950	45.000	64.500	299.000	0.829
DARK SEAS	105.000	43.700	20.160	18.000	25.000	45.000	50.000	147.000	0.461
EDWIN	120.000	73.862	36.608	22.500	47.500	70.000	95.000	203.000	0.496
HUF	318.000	49.747	29.336	19.000	25.000	47.500	65.750	185.000	0.590
OBEY	238.000	59.662	30.343	20.000	37.750	55.000	73.500	227.500	0.509

Tabella 3: Price-Brand stats

A Website scraping

Per affrontare la sfida dello scraping del sito web, che per motivi di privacy è stato oscurato e mantenuto all'interno del file `.env` ho deciso di preparare un apposito pacchetto python con il quale recuperare in modo semplice e veloce il contenuto delle diverse pagine html che volevo indagare.
Il pacchetto è stato denominato: 'htmlGrabber' proprio perché ha come obiettivo quello di raccogliere gli html che contengono l'intero catalogo di abbigliamento del sito.