Clothing website: Scraping & Analysis

Michael Tavoni

7 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.

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.

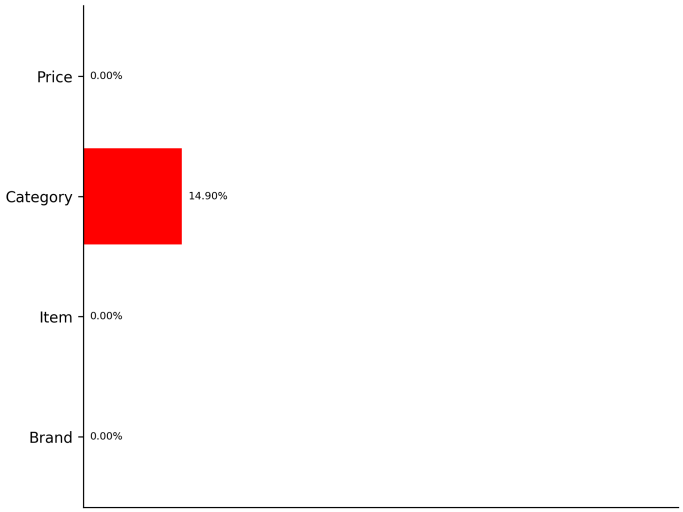

Figura 1: Nan Plot

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

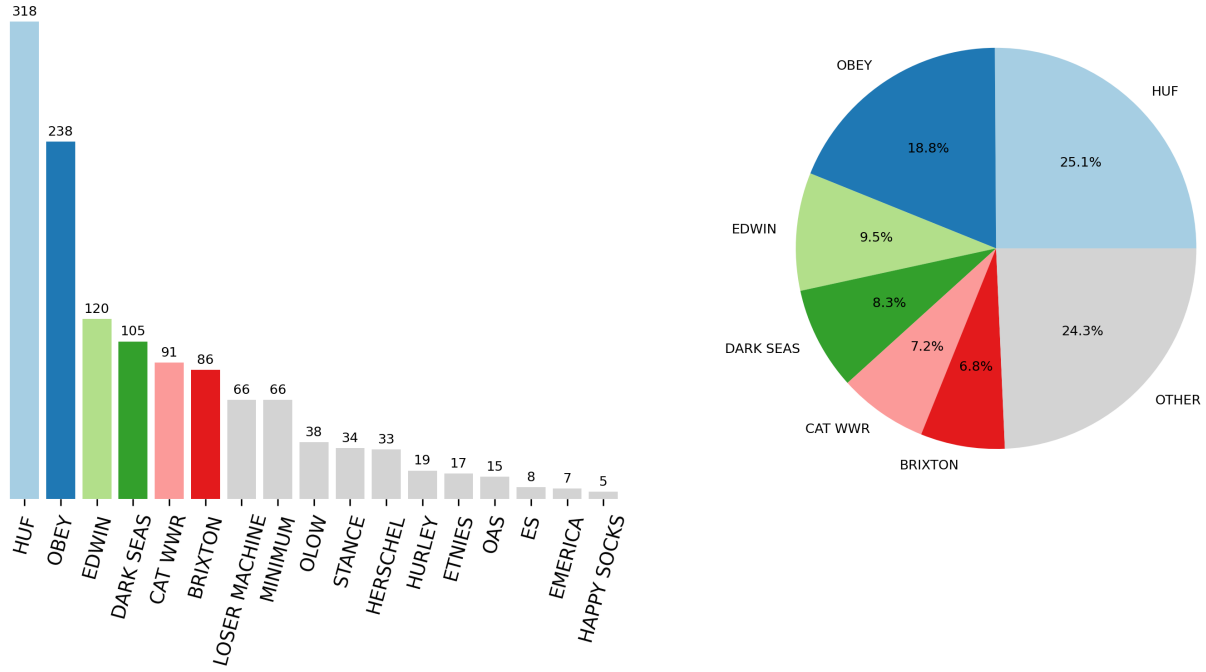

Figura 2: Brand

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

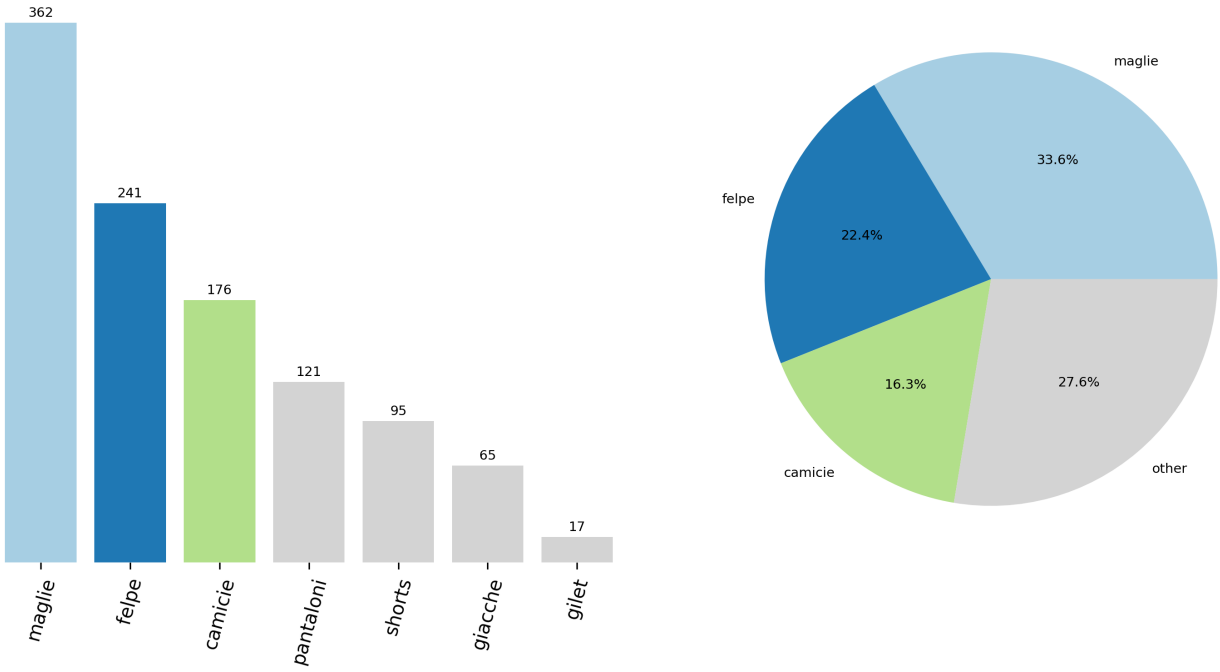

Figura 3: Category

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.