UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Facoltà di Ingegneria

Corso di Laurea Triennale in Ingegneria Informatica e dell'Automazione

---

Analisi di dati multimediali per prevedere le tendenze nel settore del fashion

Multimedia data analysis to predict trends in the fashion industry

Candidato:
Michele di Renzo

Relatore:
Prof. Emanuele Frontoni

Correlatore:
Prof. Adriano Mancini

Anno Accademico 2021/2022

INDICE

INDICE DELLE FIGURE

Capitolo 1: Introduzione

1.1 Contesto

L’e-tailing consiste nella vendita di prodotti e servizi attraverso internet. Il commercio elettronico può includere business-to-business (B2B) e business-to-consumer (B2C). Questo tipo di vendita ha bisogno di aziende che, per soddisfare i loro modelli di business, abbiano un occhio alle vendite su internet, che includono la creazione di canali di distribuzione quali magazzini e/o punti di stoccaggio e spedizione dei prodotti, pagine web internet ecc. Intensi canali di distribuzione, attraverso la quale la merce viene trasferita al cliente, e un forte *branding* garantiscono il successo dell’e-tailing. I siti web progettati devono essere attraenti, facili da navigare, e anche regolarmente ristrutturati per soddisfare le esigenze mutevoli dei consumatori. I prodotti e servizi offerti devono essere diversi dai prodotti offerti dai concorrenti e bisogna assicurarsi di dare il giusto valore al cliente. Inoltre, le proposte dalla società devono avere un prezzo competitivo per offrire un'opzione di valore al cliente. Gli e-tailers hanno bisogno di una fitta rete distributiva rapida ed efficiente, poiché gli acquirenti non accettano un tempo eccessivamente lungo per la consegna di prodotti o servizi. La fiducia e il fattore fedeltà da parte del consumatore deve essere guadagnato attraverso la trasparenza delle pratiche e dell'etica aziendale [1].

In questo scenario la data analysis risulta fondamentale per descrivere ed estrapolare informazioni. L'analisi dei dati può fare riferimento a una varietà di procedure e metodi specifici; può essere vista come parte di un processo che comporta obiettivi, relazioni, processo decisionale e idee, oltre al lavoro con dati reali [2]. Essa si occupa di:

- ispezionare dati per determinarne la qualità (e.g. ci sono valori anomali? Ci sono errori? Mancano alcuni valori?)
- verificare che i dati soddisfano i requisiti e le ipotesi del metodo analitico utilizzato e, se necessario, i valori osservati possono essere trasformati
- preparare delle considerazioni per l'elaborazione informatica e ulteriori analisi [3]

1.2 Obiettivi

Il primo obiettivo del presente progetto di tesi è lo sviluppo di un software in grado di raccogliere dati non strutturati dal sito web Net-a-porter con il fine di creare un dataset che contenga le caratteristiche relative alle borse presenti sul sito sopracitato; questo è effettuato, tramite la tecnica del web scraping, per poi standardizzare i dati raccolti e potervi così avere accesso secondo un determinato ordine. Il secondo obiettivo è quello di proporre diverse attività di analisi dai dati estratti, con il fine di realizzare analisi descrittive rilevanti.

1.3 Struttura della tesi

Nel Capitolo 2 si andrà a trattare lo stato dell'arte delle tecniche di crawling, approfondendo il web scraping (che è la tecnica utilizzata nello script realizzato), per poi introdurre il concetto di analisi dei dati per la realizzazione di statistiche e interpretazione dei dati; nel Capitolo 3 in una prima parte verranno descritti gli strumenti e le librerie utilizzati, per poi entrare nel dettaglio della metodologia di sviluppo del software; nel Capitolo 4 si andranno a mostrare i risultati ottenuti dallo scraping e dall'analisi dei dati, andando a mostrare i grafici realizzati dallo script e discutendone il loro significato; infine nel Capitolo 5 verranno tratte le conclusioni ed eventuali sviluppi futuri che potrebbero esserci nell'ambito dell'analisi dei dati e previsione in questo ambito, ma anche con altri approcci e ambiti differenti, non ristretti al settore moda.

Capitolo 2: Stato dell'arte

2.1 Tecniche di crawling

I web crawler sono programmi che sfruttano la struttura grafica del Web per spostarsi da pagina a pagina. In principio, tali programmi sono stati chiamati anche wanderers, robots, spiders, fish e worms. Il nome "crawler" non è indicativo della velocità di questi programmi, che possono essere anche molto veloci. Nella esperienza quotidiana, si è in grado di spostarsi/visitare ("strisciare") decine di migliaia di pagine in pochi minuti consumando una piccola frazione della larghezza di banda disponibile. Fin dall'inizio, una motivazione chiave per la progettazione di crawler Web è stato quello di recuperare le pagine Web e aggiungerle aggiungere esse stesse o le loro rappresentazioni in un repository locale. Tale repository può quindi soddisfare particolari esigenze applicative come quelle di un motore di ricerca. Nella sua forma più semplice, un crawler parte da una pagina e usa i suoi link esterni per spostarsi su altre pagine. Il processo si ripete con le nuove pagine che offrono altri link esterni da seguire, fino a quando non viene identificato un numero sufficiente di pagine o qualche obiettivo di livello superiore che è stato raggiunto. Se il Web fosse una raccolta statica di pagine avremmo poco uso a lungo termine per la scansione, una volta che tutte le pagine sono state scaricate in un repository (come un database del motore di ricerca), non ci sarebbe più bisogno di eseguire la scansione. Tuttavia, il Web è un'entità dinamica con sottospazi in evoluzione a tassi diversi e spesso rapidi, quindi, c'è una continua necessità di crawler per aiutare le applicazioni a rimanere aggiornate come nuove pagine che vengono aggiunte e quelle vecchie eliminate, spostati o modificati [4].

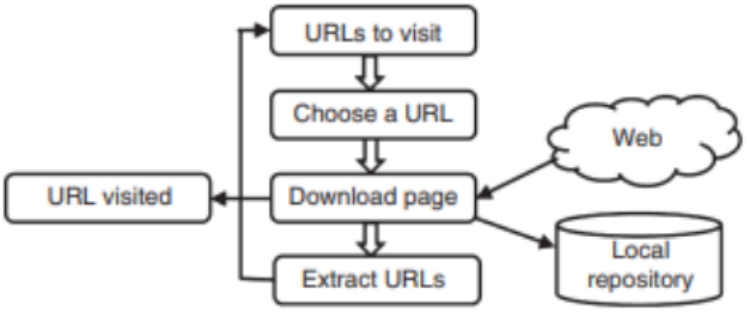

Figura 1- Architettura web crawler

2.2 Web scraping

Il web scraping è la pratica di raccogliere dati con qualsiasi altro mezzo piuttosto che un programma che interagisce con un'API (o attraverso un essere umano che utilizza un web navigatore). Questo è più comunemente realizzato scrivendo un programma automatizzato che interroga un server web,

richiede dati (di solito sotto forma di HTML e altro file che comprendono pagine web), quindi analizza quei dati per estrarre le informazioni necessarie. In pratica, il web scraping comprende un'ampia varietà di tecniche di programmazione e tecnologie, come l'analisi dei dati e la sicurezza delle informazioni [5] . L'intero funzionamento di uno scraper può essere diviso in due fasi sequenziali:

acquisizione delle risorse web: un software di web scraping inizia effettuando una richiesta HTTP per raccogliere le risorse da una pagina web d'interesse. Il tipo di richiesta può variare, può essere una richiesta di tipo GET contenente l'URL del sito al quale si vuole accedere o può essere una richiesta di tipo POST, quest'ultima è usata principalmente quando si vuole inviare dei dati al server. Una volta che il sito web ha ricevuto la richiesta con successo invierà una risposta al software, come indicato in Figura 2, che, se correttamente ricevuta, la gestirà nel modo che ritiene più opportuno. Le risorse all'interno della risposta HTTP possono essere in formati differenti poiché le pagine web possono essere costruite con diversi linguaggi di programmazione

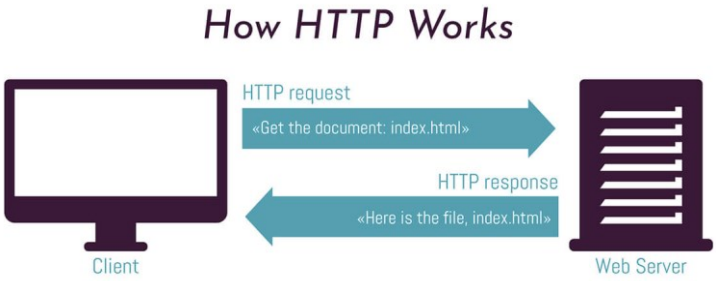

Figura 2 - Comunicazione client-server

estrazione dell'informazione: dopo la fase di acquisizione il software si occupa dei dati grezzi ricevuti e di gestirli, facendo tutte le operazioni necessarie per renderli strutturati e pronti per raggiungere lo scopo di business prefissato. La struttura di questa fase è descritta in Figura 3

Figura 3 - Architettura web scraper

2.3 Analisi dei dati

L'analisi dei dati è il processo con cui si ricavano informazioni da dati che vengono estratti, trasformati e centralizzati per scoprire e analizzare schemi nascosti, relazioni, tendenze, correlazioni e anomalie, oppure per convalidare una teoria o un'ipotesi. i dati possono essere analizzati per prendere decisioni in tempo reale, individuare trend emergenti e svelare condizioni che non sarebbero evidenti utilizzando processi di gestione dei dati obsoleti[1].

La data science è un ambito che raccoglie tutte le discipline che riguardano la pulizia, la preparazione e l'analisi dei dati per estrarne informazioni di valore altrimenti non evidenti. Utilizza, accanto a una serie di tecniche matematiche, statistiche e di programmazione, algoritmi di machine learning e intelligenze artificiali.

La data analytics è la scienza che analizza i dati grezzi per estrarre conoscenze utili (modelli) da essi. Questo processo può includere anche la raccolta dei dati, l'organizzazione, la preelaborazione, trasformazione, modellazione e interpretazione. L'analisi come area di conoscenza implica input da molte aree diverse. L'idea di generalizzare la conoscenza da un campione di dati viene da un ramo di statistica noto come apprendimento induttivo, un'area di ricerca con una lunga storia. Con i progressi dei personal computer, l'uso del calcolo le risorse per risolvere i problemi dell'apprendimento induttivo diventano sempre di più popolare. La capacità di calcolo è stata utilizzata per sviluppare nuovi metodi [6].

I data analytics (o, più semplicemente, analytics) sono strumenti che si basano sull'inferenza statistica per esaminare in maniera approfondita dati grezzi e le conoscenze a disposizione per individuare correlazioni, tendenze o verificare teorie e modelli esistenti. Rispondono a domande precise e partono da ipotesi formulate sin dall'inizio, focalizzandosi su particolari settori, con lo scopo di ottenere le best practises che portano ad un miglioramento del business.

[1] https://www.talend.com/it/resources/what-is-data-analytics/

Capitolo 3: Materiali e metodi

3.1 Tecnologie e strumenti

La realizzazione del progetto è effettuata tramite un software Python con l'ausilio del framework Selenium[2], utile per lo scraping delle pagine web, il pacchetto open source di Python Pandas, utilizzato per effettuare data analysis e machine learning e la libreria Matplotlib per la visualizzazione delle statistiche. In questo capitolo si andranno ad analizzare alcuni aspetti fondamentali per lo sviluppo della tesi, quali:

- metodo di scraping utilizzato
- discussione circa il dataset realizzato (e.g., creazione, descrizione, software utilizzati)
- analisi dei dati del dataset ottenuti dallo scraping

3.1.1 Selenium

Selenium è un umbrella project per una gamma di strumenti e librerie che abilitano e supportano l'automazione dei browser web. Fornisce estensioni per emulare l'interazione dell'utente con i browser, un server di distribuzione per ridimensionare l'allocazione del browser e l'infrastruttura per le implementazioni della specifica W3C WebDriver che consente di scrivere codice intercambiabile per tutti i principali browser web. Al centro di Selenium c'è WebDriver, un'interfaccia per scrivere set di istruzioni che possono essere eseguiti in modo intercambiabile in molti browser.

WebDriver guida un browser in modo nativo, come farebbe un utente, localmente o su una macchina remota utilizzando il server Selenium, segna un salto in avanti in termini di automazione del browser. Selenium WebDriver si riferisce sia ai collegamenti linguistici che alle implementazioni del singolo codice di controllo del browser. Selenium WebDriver è una raccomandazione W3C[3]:

- WebDriver è progettato come un'interfaccia di programmazione semplice e più concisa;
- WebDriver è un'API compatta orientata agli oggetti;
- Guida il browser in modo efficace.

il WebDriver viene inizializzato tramite una serie di parametri.

Nel progetto presentato, Selenium, è stato utilizzato per lo scraping e per la cattura dei dati sulla pagina web.

[2] https://www.selenium.dev/documentation/
[3] https://www.selenium.dev/documentation/webdriver/

3.1.2 JSON

JavaScript Object Notation (JSON) deriva dai letterali del linguaggio di programmazione JavaScript. Questo rende JSON un sottoinsieme del linguaggio JavaScript e come tale, il JSON non possiede caratteristiche aggiuntive che il linguaggio stesso non possiede già. Sebbene sia un sottoinsieme di un linguaggio di programmazione, di per sé non è un linguaggio di programmazione ma, di fatto, un formato di scambio di dati.

JSON è noto come lo standard di scambio dei dati, il che sottintende che può essere utilizzato come formato di dati in qualsiasi luogo si verifichi lo scambio di dati. Crockford ha documentato la grammatica di JSON su http://json.org nel 2001 e presto la parola ha cominciato a spargersi come un'alternativa formato dati XML. Con l'adozione diffusa di Ajax (Asynchronous JavaScript e XML), la popolarità di JSON ha cominciato a salire, così come le persone hanno cominciato a notare la sua facilità di implementazione e come effettivamente "combatteva" con quella dell'XML. Infatti, uno dei motivi per cui il JSON è diventato, fondamentalmente, il formato dei dati del Web, è dovuto alla sua semplicità grammaticale, che consente al JSON di essere altamente interoperabile [7].

Figura 4 - Esempio file formato JSON

Python ha un pacchetto integrato chiamato json, che può essere utilizzato per lavorare con i dati JSON.

- Il metodo *open* di Python → permette di aprire un file in lettura/scrittura, nel caso della creazione del dataset in scrittura tramite il parametro *w*;
- Il metodo *write* di Python → permette di scrivere effettivamente sul file in cui viene evocato un contenuto che gli viene passato come parametro;
- Il metodo json.dumps() → converte un sottoinsieme di oggetti Python in una stringa json. Non tutti gli oggetti sono convertibili e potrebbe essere necessario creare un dizionario di dati che si desidera esporre prima della serializzazione in JSON. La variabile su cui viene

evocata tale funzione viene passata come parametro nella funzione *write* precedentemente illustrata.

Questa fase viene eseguita in seguito allo scraping; i dati di tipo numerico insieme alla marca e al tipo della borsa sono quelli più rilevanti per quanto riguarda la fase di statistica e previsione che verrà approfondita in seguito.

JSON nel progetto è stato utilizzato per la realizzazione dei dataset, nei quali sono contenute le informazioni di interesse formattate proprio secondo tale notazione.

3.1.3 Pandas, Matplotlib e Seaborn

Pandas è un pacchetto Python che fornisce strutture di dati veloci, flessibili ed espressive progettate per rendere il lavoro con dati "relazionali" o "etichettati" facile e intuitivo. Mira a essere l'elemento costitutivo fondamentale di alto livello per eseguire analisi pratiche dei dati del mondo reale in Python. Inoltre, ha l'obiettivo più ampio di diventare lo strumento di analisi/manipolazione dei dati open source più potente e flessibile disponibile in qualsiasi lingua. Pandas è adatto per molti diversi tipi di dati:

- Dati tabulari con colonne tipizzate in modo eterogeneo, come in una tabella SQL o in un foglio di calcolo Excel;
- Dati di serie temporali ordinati e non ordinati (non necessariamente a frequenza fissa);
- Dati di matrice arbitraria (tipizzati in modo omogeneo o eterogenei) con etichette di riga e colonna;
- Qualsiasi altra forma di set di dati osservativi/statistici. I dati non devono essere affatto etichettati per essere inseriti in una struttura dati panda.

Le due strutture dati primarie di Pandas, *Series* mono-dimensionale e *DataFrame* bi-dimensionale, gestiscono la stragrande maggioranza dei casi d'uso tipici in finanza, statistica, scienze sociali e molte aree dell'ingegneria[4].

- Series → è un array di tipizzazione omogenea con etichetta 1D, come in Figura 8;

[4] https://pandas.pydata.org/docs/getting_started/overview.html

Series

	oranges
0	0
1	3
2	7
3	2

Figura 5 - Struttura dati Series, Pandas

- DataFrames → è una struttura tabulare generale con assi etichettati (righe e colonne) in 2D, modificabile, con colonne potenzialmente tipizzata in modo eterogeneo, come in Figura 9.

	Name	Qualification
0	Jai	Msc
1	Princi	MA
2	Gaurav	MCA
3	Anuj	Phd

Figura 6 - Struttura dati DataFrame, Pandas

Matplotlib è uno dei pacchetti Python più utilizzati per la visualizzazione dei dati. È una libreria multipiattaforma per creare grafici 2D dai dati negli array, è scritto in Python e fa uso di NumPy, l'estensione matematica numerica di Python[5].

Seaborn è una libreria di visualizzazione dei dati Python basata su Matplotlib e si integra strettamente con le strutture dati Pandas. Fornisce un'interfaccia di alto livello per disegnare grafici statistici accattivanti e informativi. Seaborn aiuta a esplorare e comprendere i dati. Le sue funzioni di tracciamento operano su dataframe e array contenenti interi set di dati ed eseguono internamente la mappatura semantica e l'aggregazione statistica necessarie per produrre grafici informativi. La sua API dichiarativa orientata al set di dati consente di concentrarsi sul significato dei diversi elementi dei grafici, piuttosto che sui dettagli di come disegnarli[6].

[5] https://www.tutorialspoint.com/matplotlib/index.htm
[6] https://seaborn.pydata.org/tutorial/introduction.html

Queste tecnologie sono state utilizzate per la realizzazione dei grafici che in seguito saranno presentati.

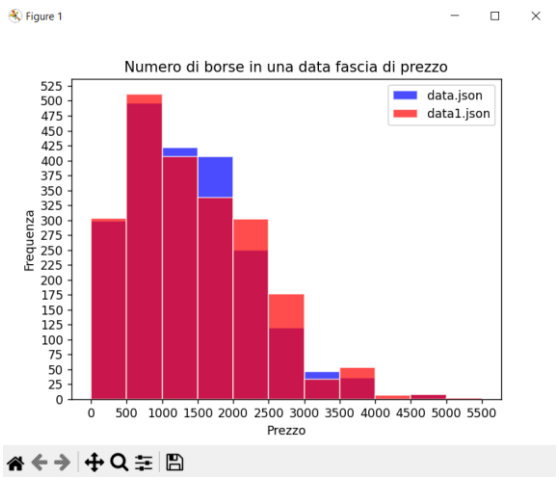

Figura 7 - Esempio grafico Maplotlib

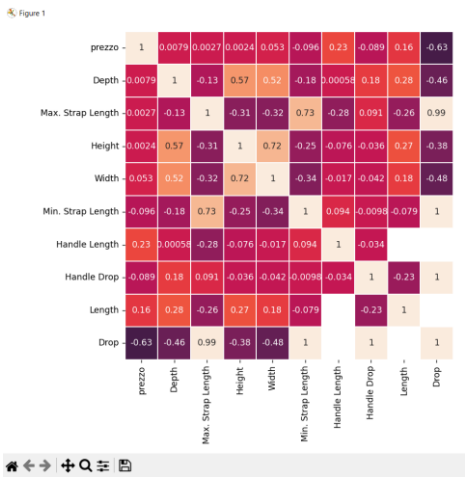

Figura 8 - Esempio grafico Seaborn

I grafici che in seguito saranno mostrati saranno:

- Seaborn.countplot → grafico utile alla visualizzazione del numero di occorrenze di un certo elemento con certe caratteristiche mediante delle barre. Nel progetto è utilizzato per tutti quei grafici (nel prossimo capitolo) che rappresentano la frequenza di determinati oggetti;
- Seaborn.barplot → grafico a barre utile, nel caso specifico del progetto, per la visualizzazione delle medie del prezzo di oggetti raggruppati secondo determinati criteri;
- Seaborn.heatmap → grafico particolare utilizzato in questo particolare caso per la visualizzazione di eventuali correlazioni tra alcune caratteristiche;
- Matplotlib.hist → classico istogramma;
- Matplotlib.kde → grafico utile alla rappresentazione di densità.

3.1.4 Libreria datetime

Il modulo datetime fornisce classi per la manipolazione di date e orari. Sebbene sia supportata l'aritmetica di data e ora, l'obiettivo dell'implementazione è l'estrazione efficiente degli attributi per la formattazione e la manipolazione dell'output[7]. Nel caso del progetto di laurea, la libreria torna utile per salvare e fare determinate operazioni con la data e ora. In particolare, il formato JSON non prevede un dato di tipo data o tempo, per cui tramite determinate funzioni che il modulo mette a disposizione, si possono effettuare operazioni di trasformazione di una data in stringa e viceversa.

[7] https://docs.python.org/3/library/datetime.html

3.1.5 Libreria Numpy

NumPy è il pacchetto fondamentale per il calcolo scientifico in Python. È una libreria Python che fornisce un oggetto matrice multidimensionale, vari oggetti derivati (come matrici e matrici mascherate) e un assortimento di routine per operazioni rapide su matrici, incluse operazioni matematiche, logiche, di forma, ordinamento, selezione, I/O , trasformate discrete di Fourier, algebra lineare di base, operazioni statistiche di base, simulazione casuale e molto altro[8].

Nello script realizzato, Numpy è stato utilizzato per la creazione di intervalli numerici utili a definire un minimo, un massimo e il passo con cui muoversi dal valore di base al finale per rappresentare i valori lungo le ascisse e/o le ordinate dei grafici realizzati.

3.2 Workflow

3.2.1 Scraping sito web

Entrando più nel particolare l'attività di scraping del progetto è stata divisa in due parti con un ordine rilevante:

1. estrazione, dal catalogo delle borse, degli URL relativi ad ogni singolo prodotto
2. estrazione dei dati d'interesse da ogni singola borsa

La fase 1 è stata fondamentale per ottimizzare e rendere più efficiente lo scraping, andando ad estrarre prima tutti gli URL delle borse e poi, con la seconda fase, aprendo le pagine web relative a ciascun URL.

Relativamente alla fase 2 dello scraping verranno illustrati in seguito i metodi utilizzati, in quanto già presentata più volte precedentemente.

3.2.2 Estrazione URL dei prodotti

La prima fase del progetto funge da base per il successivo scraping dei dati di ogni singola borsa: tramite lo script si ha accesso alla pagina https://www.net-a-porter.com/en-it/shop/bags che è quella che contiene tutte le borse (raggruppate in 60 per ogni pagina) come indicato in Figura 4.

[8] https://numpy.org/doc/stable/

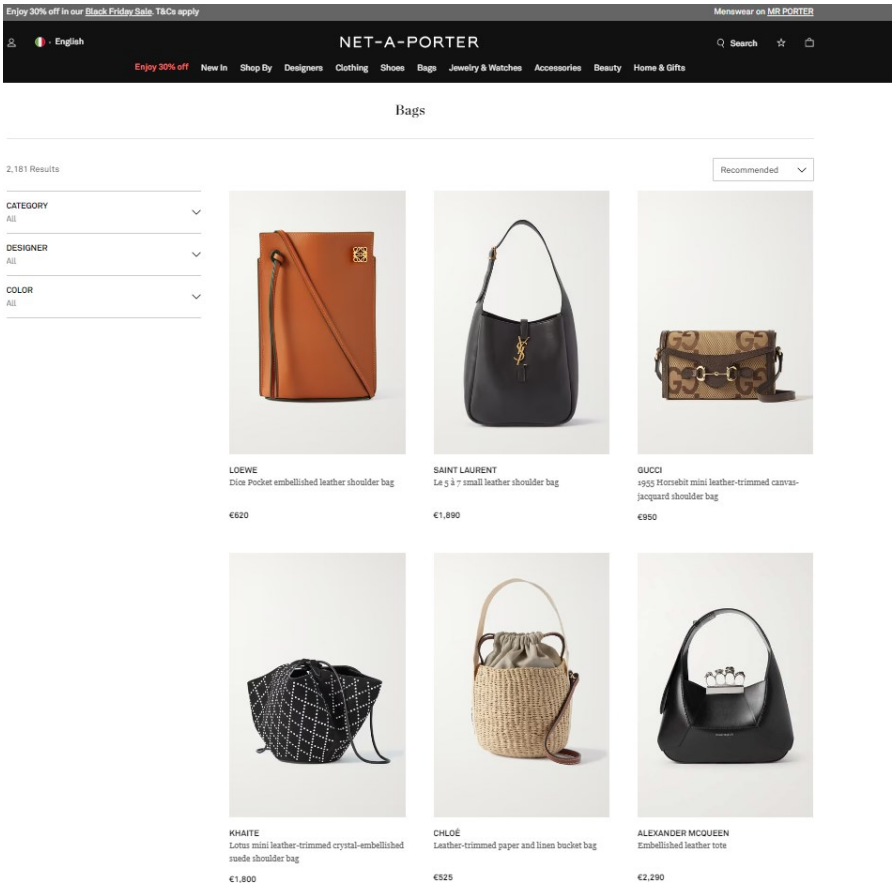

Figura 9 - Catalogo pagina web

Una volta che si è caricata la pagina si vanno ad individuare, mediante metodi che Selenium mette a disposizione, le informazioni che si vogliono catturare, informazioni che in questo caso specifico sono l'URL a cui veniamo reindirizzati dopo aver cliccato su una borsa del catalogo. Per fare ciò, si sfrutta lo strumento *ispeziona* che Chrome mette a disposizione.

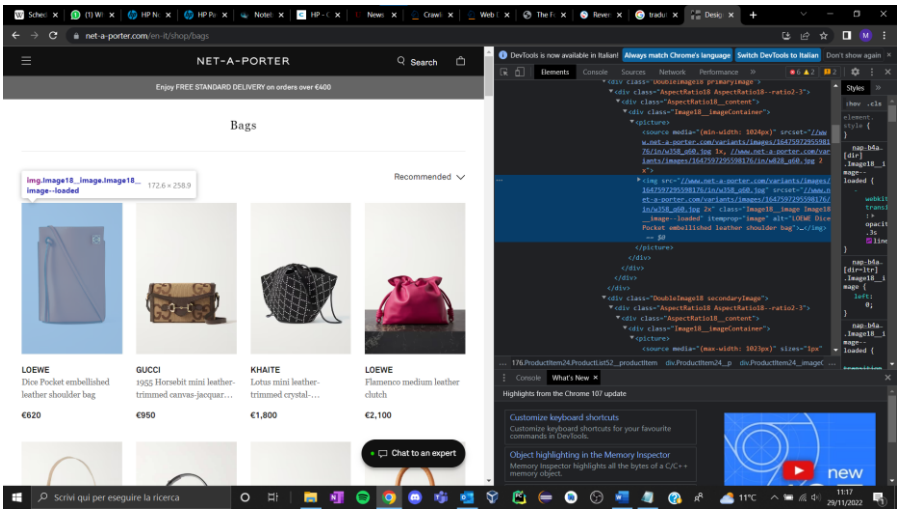

Figura 10 - Ispeziona di Chrome

Da qui si ha accesso al codice HTML/CSS/PHP/JavaScript/ecc.. della pagina e, più precisamente, dell'elemento cliccato. Queste informazioni sono utili in mediante Selenium, si possono ottenere i contenuti di determinate sezioni di pagina.

In questa sezione specifica:

1. si individua la griglia composta da tutte le borse in una pagina tramite il nome della classe
2. si prendono tutti gli elementi che hanno l'attributo HTML "a" e si salvano in una variabile
3. si scorre elemento per elemento questa variabile verificando che il contenuto di ogni singolo oggetto non contenga nell'URL determinate parole, le quali reindirizzerebbero ad una pagina non di nostro interesse
4. se la verifica va a buon fine si inserisce un oggetto chiave-valore (dove il valore è l'URL della borsa analizzata) nel file url.json

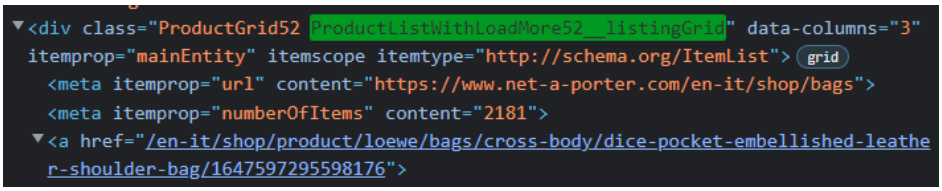

Figura 11- Ispeziona, la griglia di borse

3.2.3 Estrazione informazioni sui prodotti

Questa rappresenta la seconda fase di scraping, in cui si estrapolano i dati delle borse che sono rilevanti ai fini del task:

- l'URL della pagina della borsa, utile per avere un accesso diretto alla pagina web della borsa in esame
- il codice prodotto, che identifica univocamente il prodotto (borsa)
- la descrizione della borsa
- il titolo
- la valuta
- il prezzo
- il tipo di borsa
- le varie dimensioni della borsa
- altre informazioni utili
- colorazioni in cui è disponibile il prodotto
- data e ora di cattura, che saranno fondamentali per fare confronti tra il prima e il dopo e per fare eventuali previsioni

Nella maggior parte dei casi si ottengono i contenuti della pagina web mediante il metodo di Selenium WebDriver *BY.CLASS_NAME*, che nonostante non sia il più indicato, risulta essere quello più funzionale nel caso trattato. Quello più indicato è *BY.XPATH*, dove l'xpath è un identificatore univoco di un elemento all'interno della pagina web. In questa pagina è stato constatato, però, che non essendo strutturata in maniera ottimale, il metodo appena citato non risulta essere funzionale nella maggior parte dei casi, poiché viene meno la caratteristica fondamentale dell'univocità dell'identificatore.

La seguente figura illustra i passi eseguiti per effettuare lo scraping:

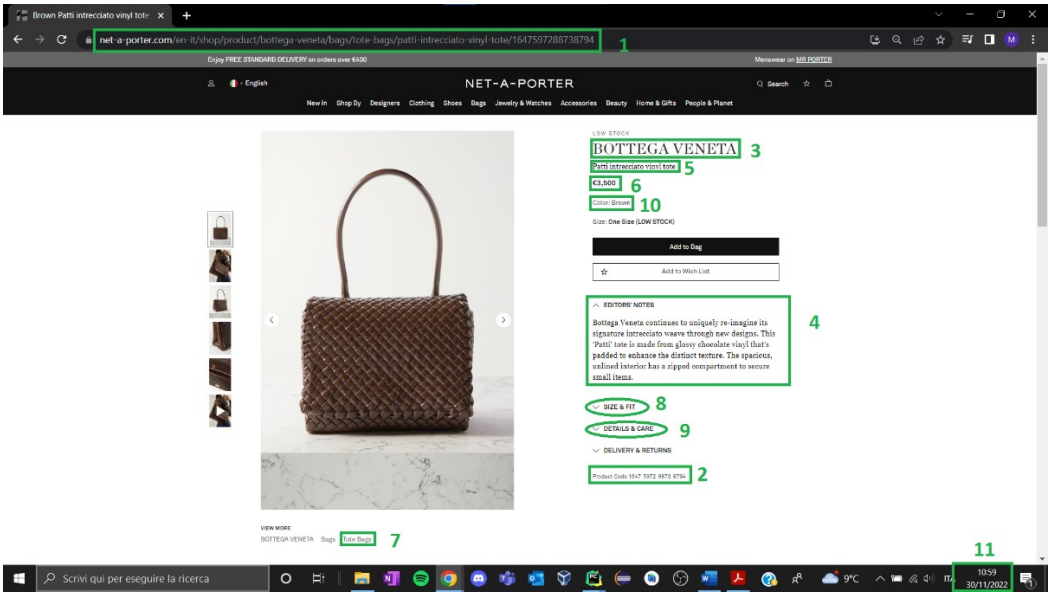

Figura 12 - Passi per lo scraping

1. Dal file*json* relativo alla prima fase ricaviamo l'URL e apriamo la pagina relativa all'elemento
2. codice del prodotto che, una volta ottenuto e processato nello script Python, è salvato nel file *json* sotto la voce *product_code*;
3. La marca è salvata nel file *json* sotto la voce *marca*;
4. La descrizione della borsa è salvata nel file *json* sotto la voce *info*;
5. Il titolo è salvato nel file *json* sotto la voce *titolo*;
6. Il prezzo che è diviso in *valuta* (€) e *prezzo*;
7. Il tipo di borsa è salvata nel file *json* sotto la voce *tipo borsa*;
8. Le dimensioni della borsa sono salvate nel file *json* sotto la voce *size and fit* sottoforma di dizionario;
9. I dettagli della borsa sono salvati nel file *json* sotto la voce *details and care* sottoforma di dizionario;
10. Le colorazioni in cui è disponibile la borsa sono salvate nel file *json* sotto la voce *colori*;

11. La data e ora di cattura è salvata nel file *json* sotto la voce *data e ora cattura*.

Le seguenti figure mostrano un dettaglio dei punti 8 e 9 sopra descritti.

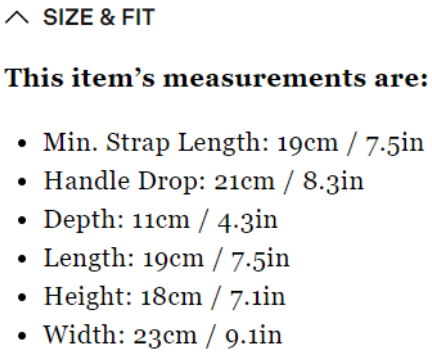

Figura 13 - Esempio punto 8, size and fit

Figura 14 - Esempio punto 9, details and care

3.2.4 Creazione dataset

Come visto in precedenza, un componente fondamentale per poter effettuare analisi di dati è il dataset, ovvero un insieme di dati strutturati creato per essere letto ed elaborato da un algoritmo. Il dataset in questione è stato realizzato tramite JSON

La prima coppia di dataset è stata realizzata in data 28/10/2022 e la seconda in data 23/11/2022, questo per far sì che si avessero due riferimenti distinti di tempo per poter eseguire alcune comparazioni e poter compiere così anche alcune previsioni.

Figura 15 - Parte iniziale e alcuni elementi url.json

```
    "url": "https://www.net-a-porter.com/en-it/shop/product/oroton/bags/shoulder-bags/colt-small-leather-shoulder-bag/1647597276183776"
  },
  {
    "url": "https://www.net-a-porter.com/en-it/shop/product/cult-gaia/bags/clutch-bags/nia-embellished-leather-clutch/33258524072771429"
  },
  {
    "url": "https://www.net-a-porter.com/en-it/shop/product/ganni/bags/shoulder-bags/pillow-baguette-leather-trimmed-recycled-shell-shoulder-bag/46376663162377036
  },
  {
    "url": "https://www.net-a-porter.com/en-it/shop/product/paravel/bags/tote-bags/fold-up-recycled-shell-weekend-bag/1647597289001243"
  }
]
```


Figura 16 - Elementi e parte finale url.json


```
[
  {
    "url": "https://www.net-a-porter.com/en-it/shop/product/valentino/bags/shoulder-bags/valentino-garavani-one-stud-small-crystal-embellished-leather-shoulder
    "product_code": "1647597276626174",
    "marca": "VALENTINO",
    "info": "Valentino Garavani and pink are a match made in heaven. Crafted in Italy, this boxy 'One Stud' shoulder bag is made from soft leather dusted with
    "titolo": "Valentino Garavani One Stud small crystal-embellished leather shoulder bag",
    "valuta": "€",
    "prezzo": "4,500",
    "tipo borsa": "Cross-body Bags",
    "size and fit": {
      "Depth": "8.5cm",
      "Max. Strap Length": "95cm",
      "Height": "13.5cm",
      "Width": "19.5cm"
    },
    "details and care": [
      "Pink leather (Lamb)",
      "Snap-fastening front flap",
      "Designer color: Rose/ Pink PP",
      "Comes with dust bag",
      "Weighs approximately 2.6lbs/ 1.2kg"
    ],
    "colori": [
      "Pink"
    ],
    "data e ora cattura": "28/10/2022 15:46:29"
  },
```


Figura 17 - Parte iniziale data.json


```
  {
    "url": "https://www.net-a-porter.com/en-it/shop/product/alexander-wang/bags/mini-bags/scrunchie-small-embellished-faux-fur-shoulder-bag/1647597276069556",
    "product_code": "1647597276069556",
    "marca": "ALEXANDER WANG",
    "info": "Alexander Wang's 'Scrunchie' bag is named for the ruched top handle that's inspired by the popular hair accessory. Playful in shape, shade and textu
    "titolo": "Scrunchie small embellished faux fur shoulder bag",
    "valuta": "€",
    "prezzo": "515",
    "tipo borsa": "Tote Bags",
    "size and fit": {
      "Handle Drop": "7cm",
      "Depth": "14cm",
      "Max. Strap Length": "25cm",
      "Height": "13cm",
      "Width": "28cm"
    },
    "details and care": [
      "Sand faux fur",
      "Zip fastening along top",
      "Designer color: Sandstone",
      "Weighs approximately 2.2lbs/ 1kg"
    ],
    "colori": [
      "Sand"
    ],
    "data e ora cattura": "28/10/2022 20:59:28"
  }
]
```


Figura 18 - Parte finale data.json

3.2.5 Grafici Matplotlib e Seaborn

Un' osservazione importante è che nei grafici successivamente mostrati non sono state considerate le borse di una determinata marca con meno di 20 occorrenze e allo stesso modo non sono stati considerati i colori con meno di 20 apparizioni; questa scelta è stata effettuata in modo da rendere i risultati più leggibili e per eliminare dati poco rilevanti.

Capitolo 4: Esperimenti e risultati

4.1 Risultati web scraping

Come già affermato in precedenza, l'attività di scraping è stata la prima ad essere stata effettuata per il completamento del progetto. Il primo ciclo di estrazione dati, effettuato con una prima versione (meno efficiente) del software, ha impiegato circa 6 ore, mentre il secondo, dopo aver ottimizzato tramite la tecnica descritta nel paragrafo 3.1.1, circa 4 ore. Il numero di borse analizzate è stato 4.232, di cui 2.088 relative al primo ciclo di scraping effettuato in data 28/10/2022 e 2.144 relative al secondo ciclo, effettuato il 23/11/2022.

4.2 Risultati analisi dei dati

Sono state eseguite diverse analisi sui dati in possesso e sono stati costruiti vari tipi di grafici per visualizzarle e trarne delle conclusioni. Considerando che sono stati fatti due cicli di scraping si è optato per la scelta grafici che vadano a visualizzare contemporaneamente le statistiche del primo e del secondo ciclo, in modo da ottenere un confronto diretto.

4.2.1 Frequenza borse per marca

Come prima indagine, è interessante osservare il numero di borse per marca. Questo fornisce una panoramica su come è suddiviso l'e-tailer in esame riguardo i brand presenti. Un'altra osservazione rilevante è che, come è possibile notare nella seguente figura, alcune marche di borse presentano una frequenza minore nella seconda istanza della statistica (e.g., Balenciaga, Givenchy); questo può essere dato dal fatto che le borse siano state vendute, per cui questa informazione, qualora si disponessero di determinate autorizzazioni per visualizzare le vendite, potrebbe essere utile per fare altre statistiche e addirittura previsione. Al contrario, come ad esempio nel caso del brand Bottega Veneta, si potrebbe pensare che sia stato fatto un restock.

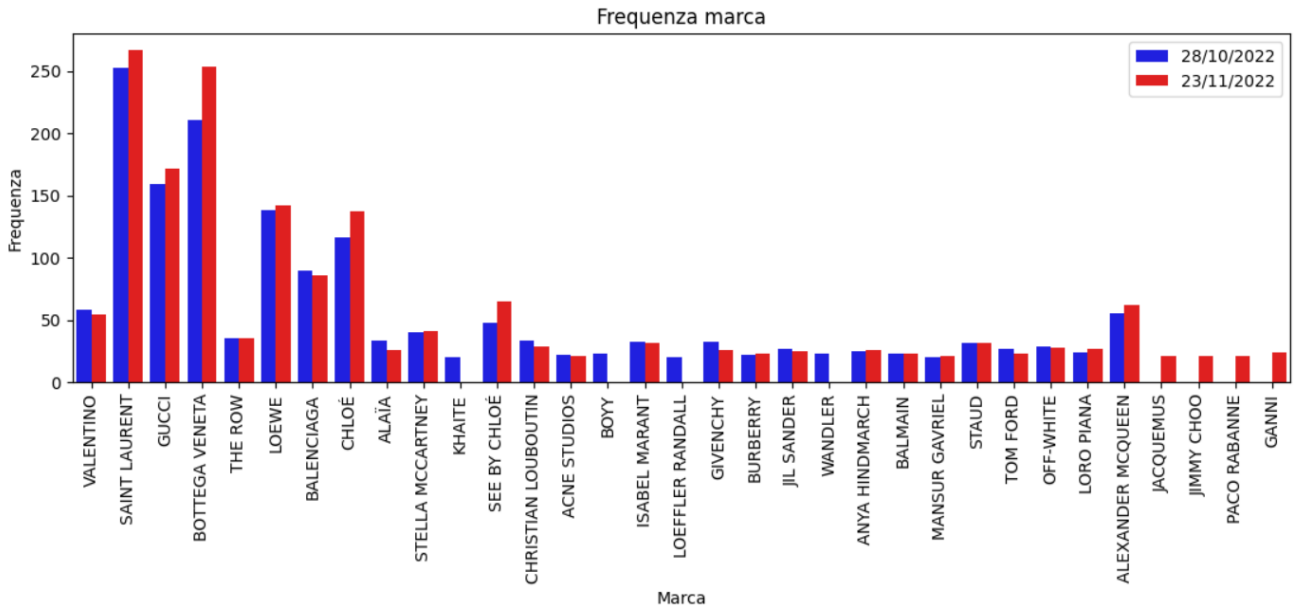

Figura 19 - Frequenza borse per marca

Si noti che dove manca la barra blu o rossa i dati estratti non soddisfano quelle condizioni espresse nel paragrafo 3.3.1; ciò vale per tutti i grafici presentati in questa tesi.

4.2.2 Numero borse per fascia di prezzo

Una seconda analisi effettuata è quella relativa alla distribuzione delle borse in una determinata fascia di prezzo (con un'ampiezza di 500 €):confrontando la prima statistica con la seconda (in Figura 20), si può constatare come la distribuzione in grandi linee sia rimasta simile alla prima. La differenza più significativa si nota nella fascia di prezzo tra 1.500 € e 2.000 € dove, nei dati raccolti nella prima data,si hanno circa 60 borse in più rispetto alla seconda data. Si può ipotizzare che la fascia di prezzo più apprezzata dai clienti sia quella dove si nota maggior divario tra il primo dato e il secondo dato.

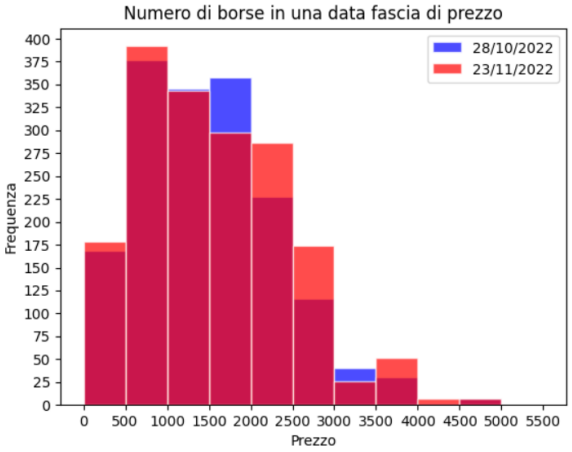

Figura 20 - Borse in fascia prezzo

4.2.3 Prezzo medio per marca

Questa statistica può essere utile nel confronto tra i due cicli di scraping, in quanto sapendo il prezzo medio per marca nelle date diverse si può osservare se un brand ha aumentato i prezzi (contemporaneamente studiando il grafico delle frequenze per marca in Figura 21 verificando che le frequenze siano il più possibile uguali nei due periodi per avere un'affermazione veritiera).

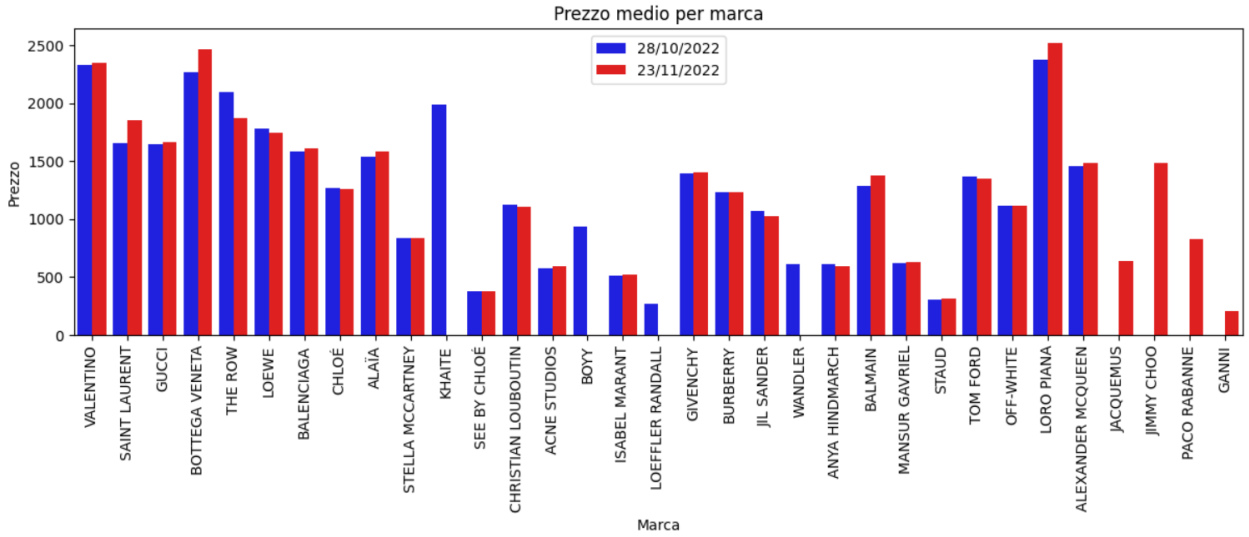

Figura 21 - Prezzo medio per marca

In questo caso specifico vediamo come nella maggior parte dei casi i prezzi medi siano rimasti simili, in altri casi (Saint Laurent, Bottega Veneta) c'è un dislivello più accentuato rispetto alle altre marche e si può ipotizzare che quei brand abbiano alzato il prezzo delle loro borse o abbiano aggiunto prodotti di fascia alta così da far aumentare il prezzo medio.

4.2.4 Densità dei prezzi

Questa statistica aiuta ad individuare la densità con cui i prezzi sono distribuiti e, presentando sullo stesso grafico i diversi periodi, si possono confrontare i prezzi e le loro densità con cui appaiono nel sito nella sezione delle borse.

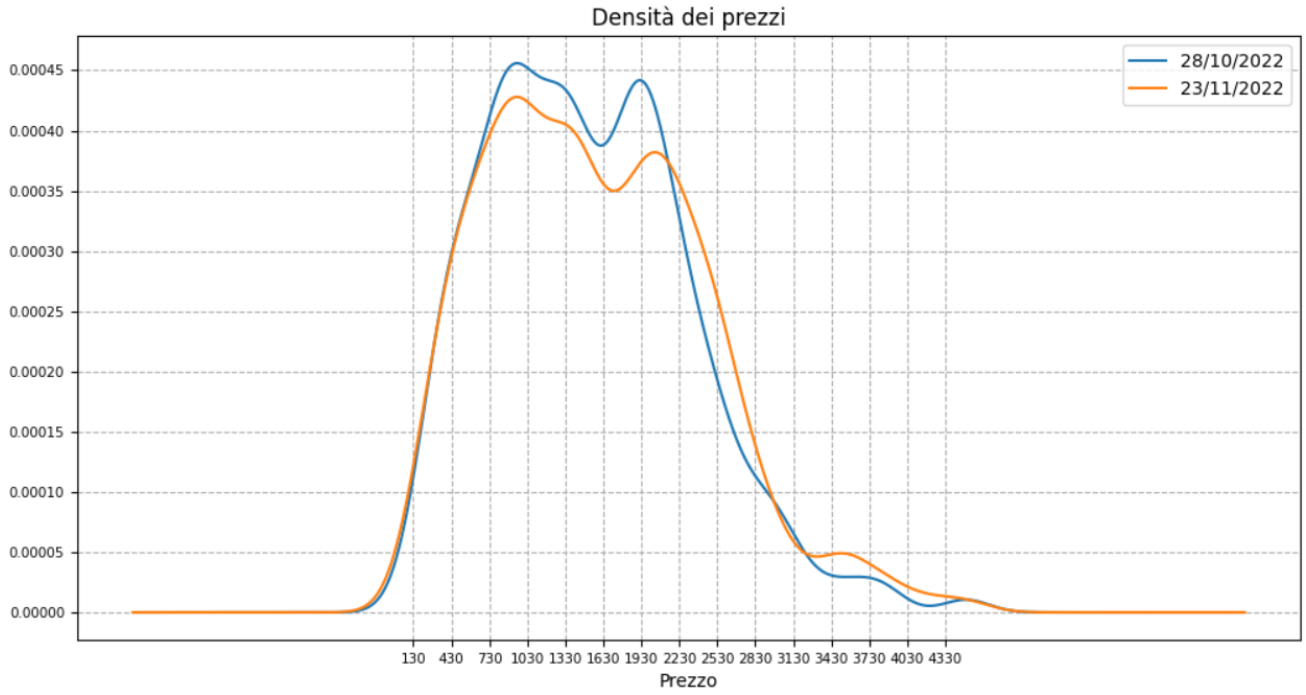

Figura 22 - Densità dei prezzi

Dal grafico della figura superiore, si può affermare che dal prezzo più basso che troviamo nel sito (il più basso tra i due periodi) fino a 730 € circa, la densità è pressoché uguale nei due periodi. Le due differenze notevoli le si osservano nella fascia di prezzi che va da 735 € a 2.130 € circa, in cui il primo periodo presenta più borse in quella fetta di prezzi; da 2.130 € a 2.950 € circa osserviamo come la densità di borse a quei prezzi sia maggiore nel secondo periodo analizzato per poi arrivare da 4.400 € fino a 4.500 € circa dove a seguito di una piccola discrepanza tra le due densità, ritornano a coincidere.

4.2.5 Correlazione prezzo e dimensioni

Questa parte dell'indagine è senza dubbio la più accattivante e significativa se si vogliono previsioni. Qui viene analizzata la correlazione che c'è fra il prezzo della borsa con i suoi attributi che la descrivono dal punto di vista dimensionale (e.g., altezza, profondità, larghezza).

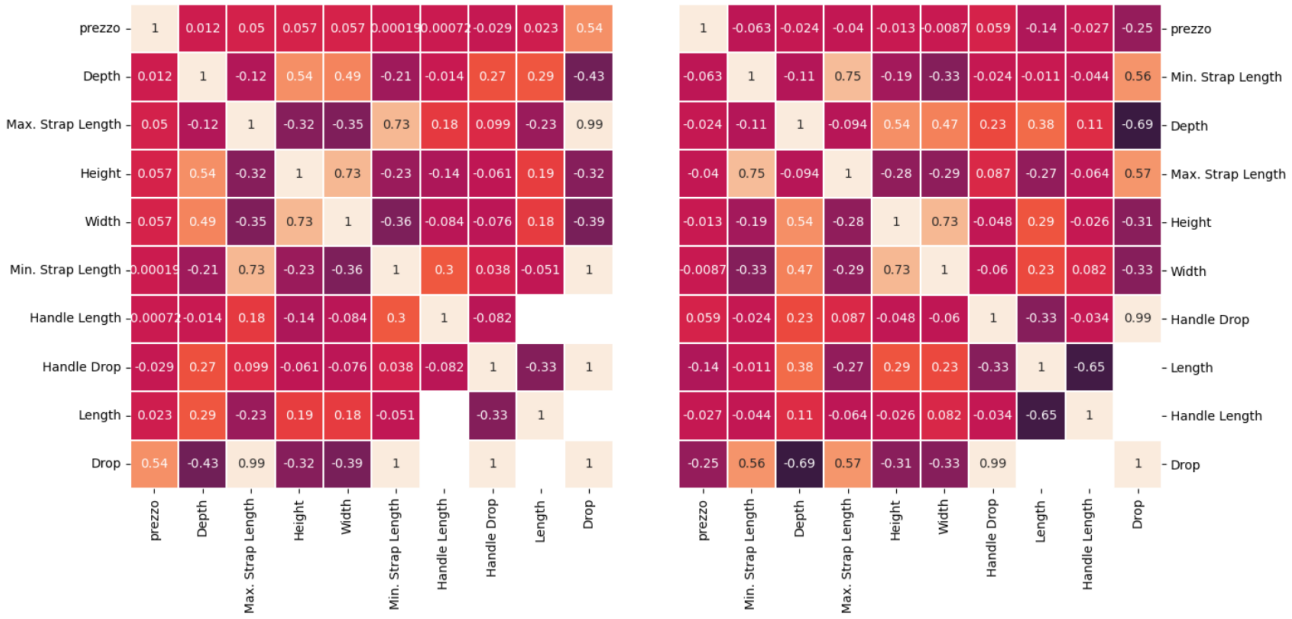

Figura 23 - Correlazione prezzo-size and fit

Questa indagine è stata svolta per scovare la correlazione tra il prezzo e gli attributi della borsa;il dato più significativo che si può osservare è quello relativo alla matrice di sinistra (che si riferisce al primo periodo di estrazione dei dati), dove si trova una correlazione tra il *prezzo* e il *Drop* pari a 0.54. È altresì interessante vedere come quella appena trattata non possa essere considerata una "legge" in quanto con lo scorrere di un tempo relativamente breve, la correlazione tra gli stessi è pari a -0.25 (nella matrice di destra), per cui sarebbe utile esterndere questa analisi ad un periodo di tempo sufficientemente lungo per dichiarare se effettivamente il drop e il prezzo hanno un legame di correlazione o meno.

4.2.6 Densità altezza, lunghezza e larghezza

Il grafico di seguito mostra come sono distribuite le principali caratteristiche che descrivono la dimensione di una borsa.

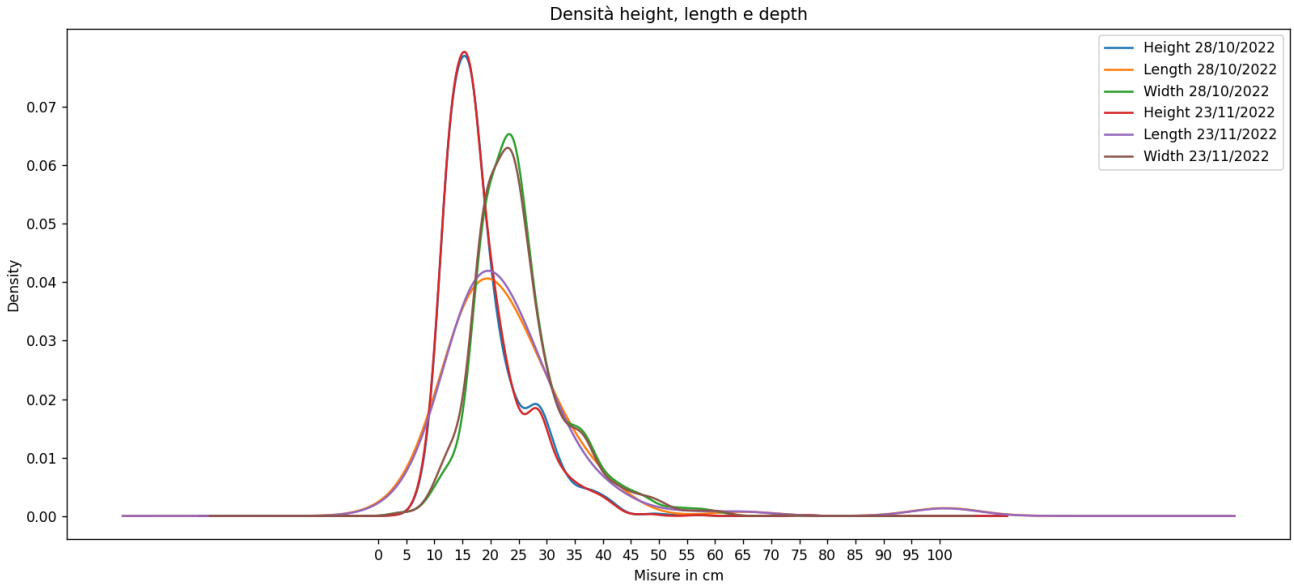

Figura 24 - Densità altezza, lunghezza e larghezza

È evidente che per quanto riguarda le dimensioni più presenti vi sono:

- altezza: 15 cm circa
- lunghezza: 19 cm circa
- larghezza: 23 cm circa

4.2.6 Altezza media per marca

Questa analisi si riferisce all'altezza media relativa alle borse, divise per marca. È interessante osservare come l'altezza delle borse, nonostante la differenza di brand e di periodo sia piuttosto omogenea.

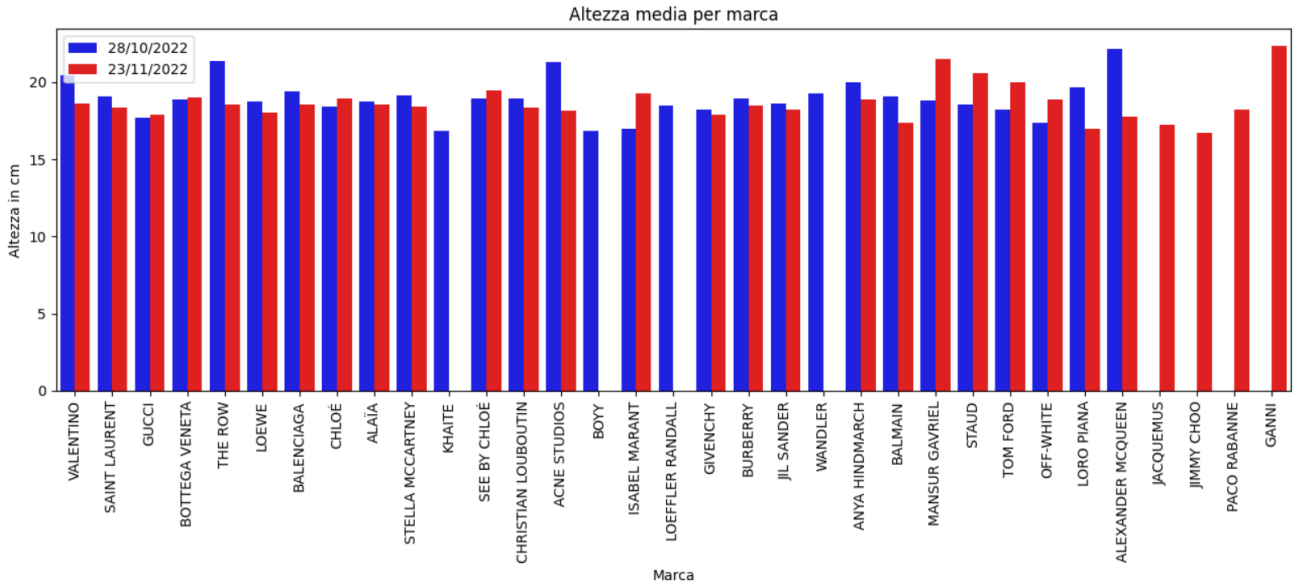

Figura 25 - Altezza media per marca

4.2.7 Lunghezza media per marca

Questa analisi si riferisce alla lunghezza media relativa alle borse, divise per marca. Qui a differenza del caso precedente, troviamo misure miste sia tra le marche sia tra i due periodi.

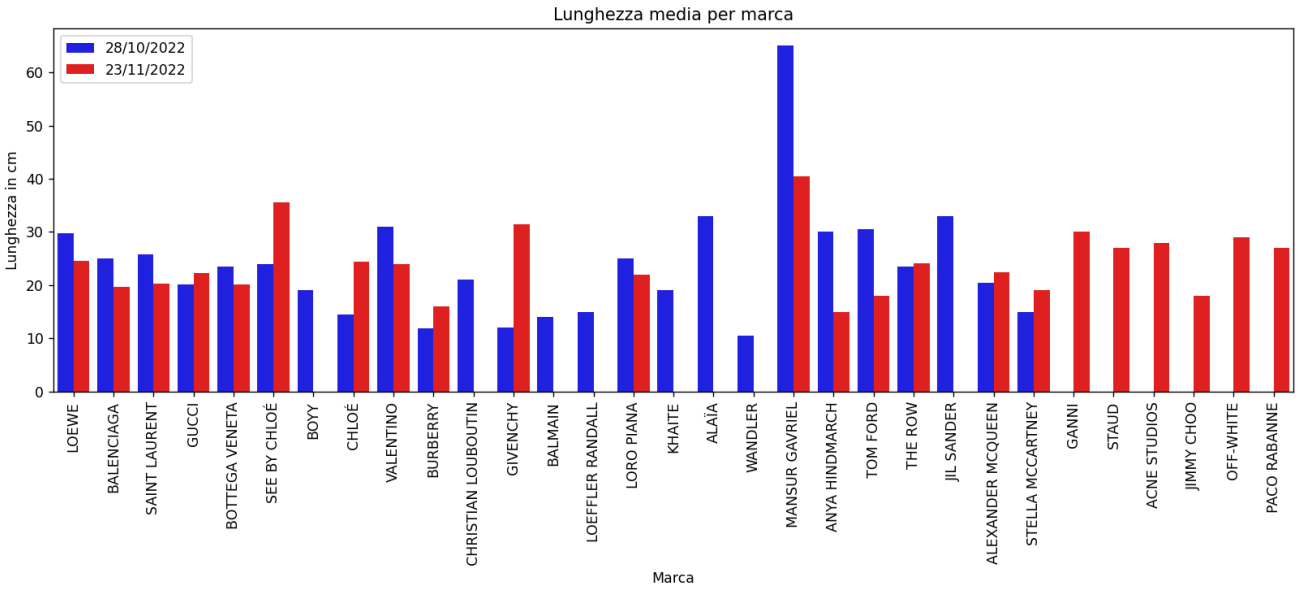

Figura 26 - Lunghezza media per marca

4.2.8 Larghezza media per marca

Questa analisi si riferisce alla larghezza media relativa alle borse, divise per marca. Riguardo questo dato si può ipotizzare, visto l'andamento del grafico, che una borsa mediamente sia larga circa 25 cm.

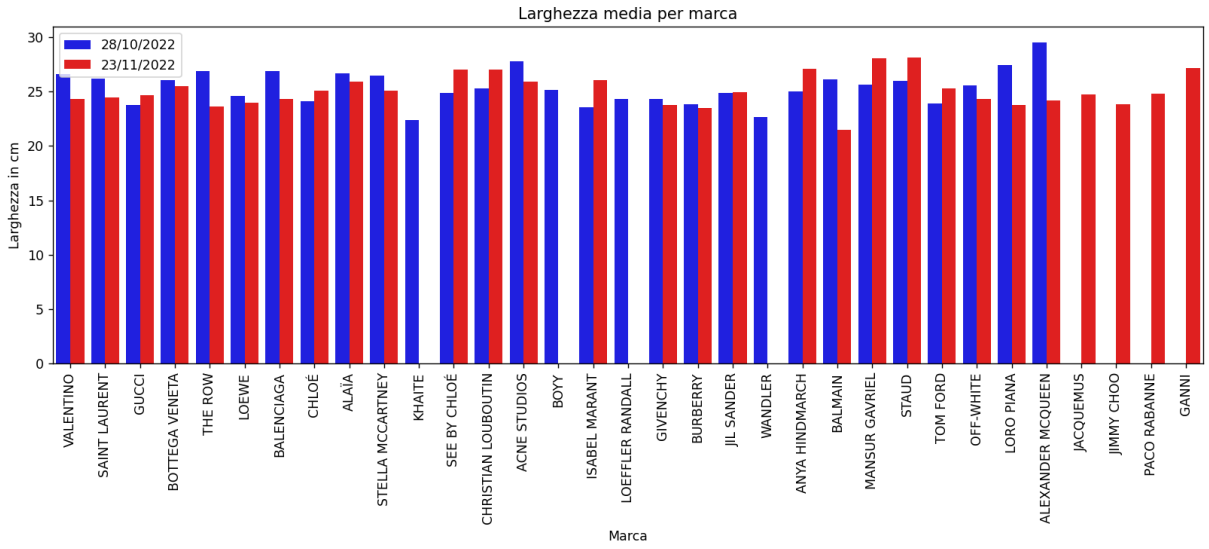

Figura 27 - Larghezza media per marca

4.2.9 Frequenza borsa per tipo

In questo grafico sono presentati i tipi di borsa presenti nel sito e la loro frequenza. È interessante osservare come le *corss-body bags* e le *tote bags* siano dominanti in entrambi i cicli di scraping. Si può ipotizzare che queste siano i tipi di borse più apprezzate dai clienti e di conseguenza quelle più prodotte dalle case produttrici.

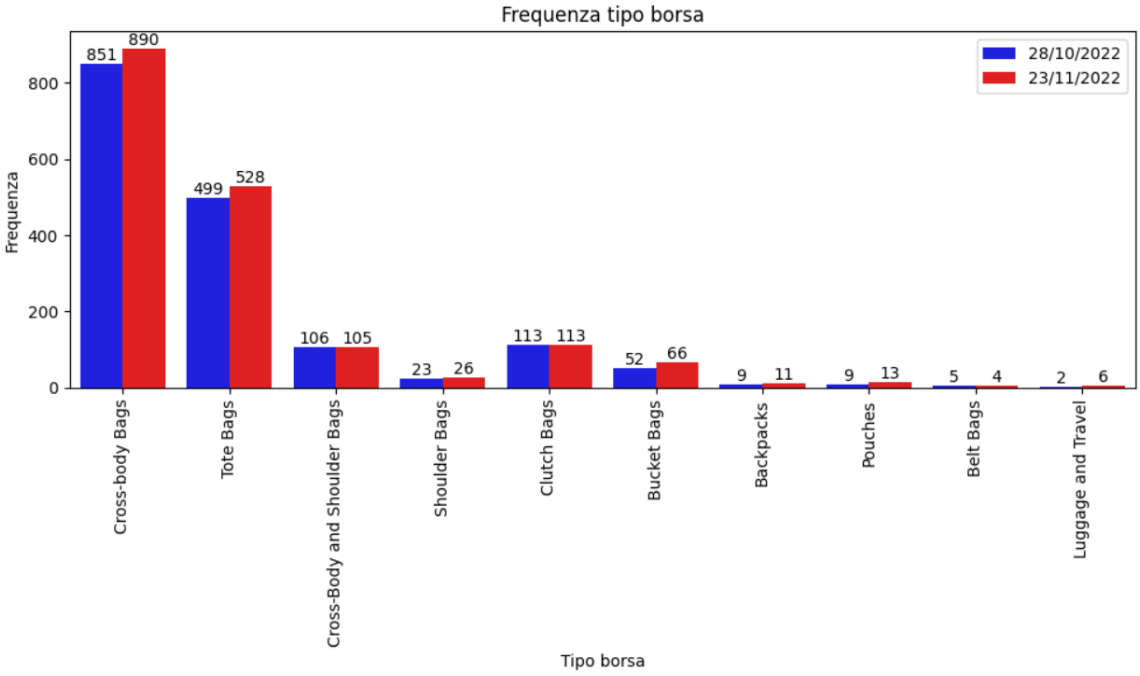

Figura 28 - Frequenza tipo borsa

4.2.10 Prezzo medio per tipo borsa

In questo grafico viene mostrato il prezzo medio suddiviso per tipo di borsa nelle due differenti date in cui sono stati acquisiti i dati.

Il prezzo medio per tipo borsa lo si può paragonare tra i due diversi periodi ed è evidente sono rimasti quanto più simili. Si può notare che il tipo con il prezzo medio più basso è il tipo *belt bags* e può essere interessante notare come in ogni caso, eccetto *bucket bags* e *pouches* il prezzo medio è aumentato nel tempo. Anche qui sarebbe interessante continuare l'indagine per un periodo più lungo nel tempo per trarre alcune conclusioni più precise.

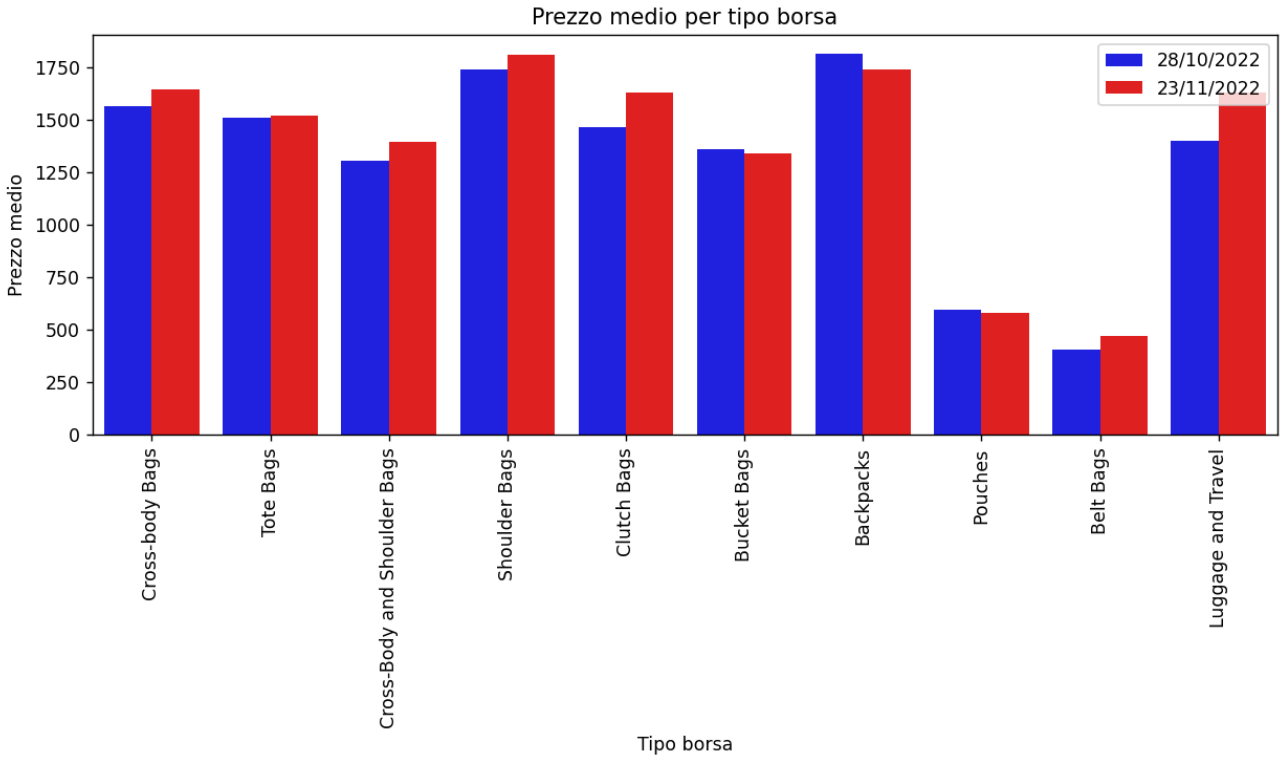

Figura 29 - Prezzo medio tipo borsa

4.2.11 Frequenza colore

In questa sezione è presentata la frequenza con cui il colore di una borsa è presente sull'intero sito. Anche in questo caso la differenza dei risultati tra i due periodi di tempo è minima, poiché i trend e i colori per il settore dell'abbigliamento in generale è difficile che varino nell'arco di un mese, soprattutto se questo non corrisponde al cambio di una stagione. Si può osservare che il colore predominante è il nero poiché oltre ad essere un colore invernale si può facilmente abbinare con altri colori del proprio outfit; seguito dai classici colori bianco, verde e marrone.

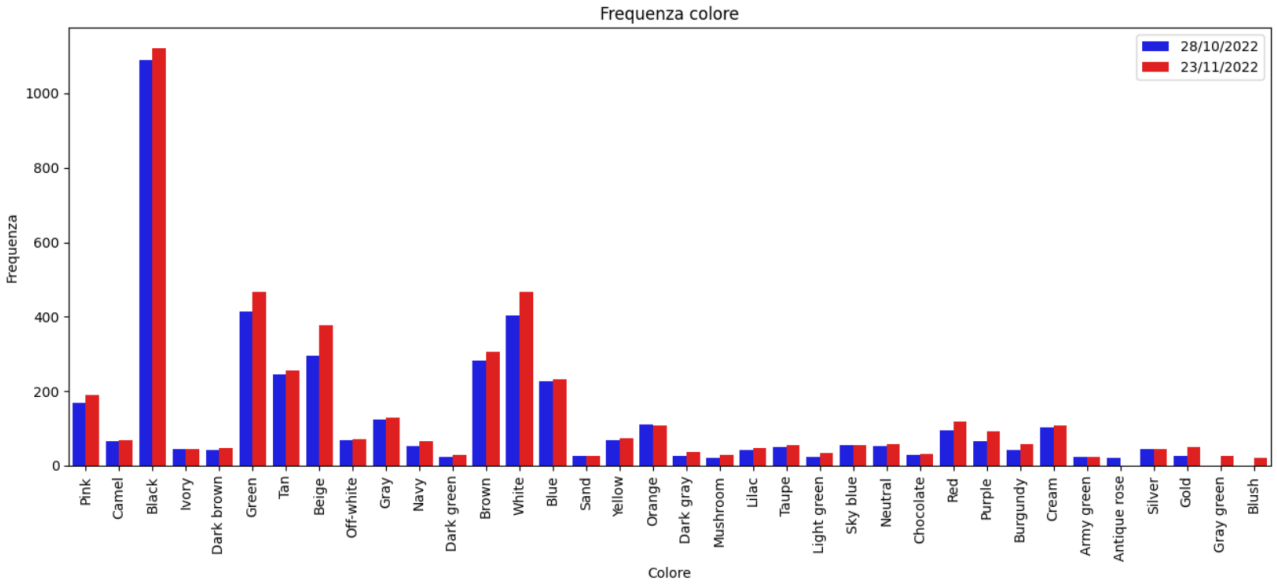

Figura 30 - Frequenza colore

4.2.12 Prezzo medio per colore

Questa analisi mostra il prezzo medio delle borse suddivise per colore. La cosa interessante da notare anche qui (come nel paragrafo 4.2.10) è che nel secondo periodo il prezzo medio nella maggior parte dei casi è aumentato.

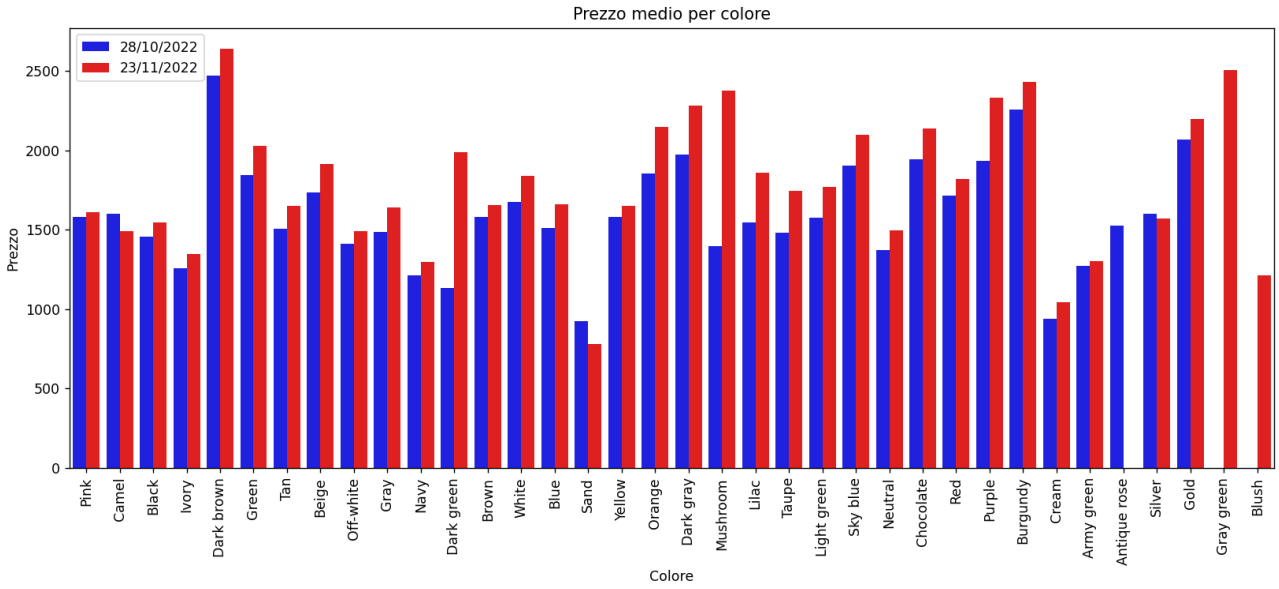

Figura 31 - Prezzo medio colore

Capitolo 5: Conclusioni e sviluppi futuri

All'interno di questo elaborato sono stati trattati i temi di scraping inteso come estrazione di dati di interesse da un sito web partendo da un'analisi manuale del sito per poi realizzare effettivamente uno script automatizzato per tale scopo; si è passati poi alla creazione di un dataset in cui sono stati salvati i dati estratti tramite lo scraping, andando a definire una struttura standardizzata per la visualizzazione di tali informazioni mediante il formato JSON; infine, sono state effettuate diverse operazioni di analisi sui dati estratti, in modo da evidenziare particolari relazioni tra elementi.

Sarebbe interessante espandere le operazioni effettuate a tutto il sito (non solo alla sezione borse); anche considerare un periodo di tempo più ampio, effettuando diverse raccolte di dati, sarebbe utile per ottenere confronti nel tempo, in modo da poter fare previsioni su trand futuri o fornire dati sempre aggiornati. Quanto stato fatto può essere espanso ad altri settori, oltre a quello del fashion, in quanto il processo presentato per raggiungere l'obiettivo definito risulta essere ripetibile e riutilizzabile.

Bibliografia

[1] Dr. R. Prakash Babu e Jalaja L., «E-Tailing (B2C) A Growth Story…Opportunities,» *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH,* 2020.

[2] Introduction to Data Analysis Handbook, 2006.

[3] R.H.G. Jongman, C.J.F. Ter Braak e O.F.R. Van Tongeren, Data analysis in community and landscape ecology.

[4] M. Levene e A. Pulovassilis, Web Dynamics Adapting to Change in Content, Size, Topology and Use, 2004.

[5] R. Mitchell, Web Scraping with Python Collecting Data from the Modern Web.

[6] J. Moreira, A. de Carvalho e T. Horvàth, A general introduction to Data Analytics, João Moreira, Andrè de Carvalho, Tomàš Horvàth, 2018.

[7] B. Smith, Beginning JSON, 2015.

Ringraziamenti

Voglio ringraziare la mia famiglia che ha sempre creduto in me fin dall'inizio di questo percorso; li ringrazio perché hanno avuto una parola di conforto quando le cose non andavano nel verso giusto, aiutandomi sempre a rimettermi in piedi nei momenti in cui ero scoraggiato e pensavo che quella intrapresa non fosse la giusta strada per il mio futuro; li ringrazio perché hanno fatto sì che non mi mancasse mai nulla senza farmelo pesare; li ringrazio perché tramite i valori che mi hanno trasmesso quali umiltà, dedizione al lavoro, rispetto, impegno e sacrificio hanno permesso che io raggiungessi questo traguardo.

Ringrazio i nonni e gli zii tutti, anche quelli che non ci sono più, che si sono sempre preoccupati di me, che mi hanno aiutato con parole o gesti che non dimentico e sono stati per me, tante volte, un modo per staccare dallo studio offrendomi una casa MAI VUOTA ricca di affetto, gioia e compagnia.

Ringrazio i miei amici, i miei fratelli, che ogni volta che tornavo sapevo con chi poter passare una serata tra cazzeggio e risate, vi ringrazio perché so che su di voi potrò contarci sempre.

Ringrazio gli amici pescaresi che hanno sempre trovato il modo di essere presenti in questi anni regalandomi un sorriso e supporto fraterno in ogni occasione, rallegrandomi le dure giornate al termine delle quali spesso ci incontravamo.

Ringrazio i miei compagni di squadra e il mister di questi ultimi tre anni che si sono rivelati veri amici, fratelli e a volte padri (c'è da dirlo) e che hanno permesso di sfogare la mia voglia di movimento e di ridere dopo aver passato la settimana seduto davanti ad un computer.

Ringrazio tutti voi soprattutto per non avermi MAI fatto sentire solo.

Ed eccomi, giunto a ringraziare la persona che negli ultimi (quasi) quattro anni ha rappresentato per me l'impegno più bello ed importante, Giorgia. Ti ringrazio perché con la tua dolcezza e il tuo sorriso così spontanei mi sei sempre stata accanto nei momenti tristi e di gioia che questo percorso e in generale la vita mi ha presentato. Sei stata un'amica con cui lamentarmi degli esami non passati o delle giornate faticose affrontate, una fidanzata con cui passeggiare chiacchierando svagandoci tenendoci la mano e, come ti dico tante volte, una mamma sempre pronta a rimettermi sulla retta via nei miei momenti di poca lucidità. Ti ringrazio perché con i tuoi successi universitari e la tua serietà nello studio sei stata fonte di ispirazione per me e pedina fondamentale per il raggiungimento di questo primo obiettivo. Un sentito grazie anche alla tua meravigliosa famiglia che ti ha trasmesso sani principi che ti hanno reso la bellissima persona che sei e che mi hanno accolto facendomi sentire sempre a casa e mai di troppo.

Un plauso lo faccio a me stesso che, nonostante non avrei mai pensato di affrontare l'università sono giunto alla laurea. Mi complimento perché, nonostante in questo percorso ho affrontato numerosi fallimenti, ho avuto quella giusta dose di testardaggine che ha permesso che non mi arrendessi davanti a due numeri che qui si chiamano voti o davanti a situazioni ingiuste, ma sono sempre stato pronto a voltare pagina e a ricominciare per far vedere quanto valessi.