Università degli Studi di Torino

Dipartimento di informatica

Progetto IUM-TWEB

Analisi dei dati sui Film: progettazione, sfide e considerazioni

Gruppo:

Studenti: Federico Verra, Federico Raso

ANNO ACCADEMICO 2024/2025

Indice

Capitolo 1

Introduzione

1.1 Il Grande Schermo risponde

In questa sezione presentiamo il cuore del nostro progetto: un'esplorazione guidata dei dati cinematografici forniti, attraverso una serie di domande che abbiamo formulato e a cui abbiamo cercato di dare risposta con strumenti di analisi e visualizzazione. Dopo una fase iniziale di pulizia, normalizzazione e comprensione delle tabelle a disposizione, ci siamo interrogati su quali informazioni fosse possibile estrarre dai dati: dai generi più rappresentati alle nazionalità più prolifiche, dai trend di popolarità degli attori e delle attrici alla distribuzione geografica delle produzioni, fino ad arrivare a curiosità legate a singoli film o intere saghe.

Il nostro approccio si è basato sull'utilizzo di Python e delle principali librerie per la manipolazione e visualizzazione dei dati, come Pandas, Seaborn, Plotly e Matplotlib. Abbiamo costruito grafici, mappe e tabelle interattive con l'obiettivo di rendere le risposte facilmente interpretabili anche da un pubblico non tecnico. L'analisi si è evoluta da interrogazioni semplici a domande più complesse, anche a livello di codice, cercando di rendere i risultati chiari e visivamente efficaci.

Un aspetto che abbiamo voluto sottolineare è la possibilità di adattare le nostre analisi a contesti specifici: per esempio, se si volesse indagare la popolarità di un singolo film o l'andamento di una particolare saga, è sufficiente modificare alcuni parametri nei nostri script per ottenere visualizzazioni mirate. Questo rende l'approccio replicabile e riutilizzabile su differenti casi d'uso, senza dover riprogettare l'intero sistema di interrogazione.

Le pagine che seguono raccontano questo processo attraverso esempi concreti: ogni domanda è accompagnata da una spiegazione del metodo utilizzato e da una rappresentazione visiva dei risultati ottenuti. L'intento è quello di mostrare come i dati possano essere esplorati in modo critico e creativo, per raccontare il cinema non solo attraverso le storie che porta sullo schermo, ma anche attraverso i numeri che lo descrivono.

Capitolo 2

Implementazione

2.1 Tecnologie utilizzate

Per lo svolgimento dell'analisi ci siamo affidati al linguaggio di programmazione **Python** [2], una scelta ormai consolidata nell'ambito della data science per la sua semplicità, versatilità e l'ampio ecosistema di librerie dedicate all'analisi dei dati. In particolare, abbiamo utilizzato **Pandas** [3] per la manipolazione dei dataset e la gestione di tabelle complesse, mentre per la creazione di grafici e visualizzazioni ci siamo affidati a diverse librerie complementari: **Seaborn** e **Matplotlib** per la produzione di grafici statici di tipo descrittivo (bar chart, line chart, ecc.), e Plotly per visualizzazioni interattive più dinamiche, utili soprattutto in presenza di confronti o dati distribuiti nel tempo.

Per le analisi su scala geografica è stata fondamentale la libreria **GeoPandas** [1], che ci ha permesso di visualizzare la distribuzione delle produzioni cinematografiche a livello mondiale tramite mappe tematiche. L'intero lavoro è stato sviluppato in ambiente **Jupyter Notebook**, che ha facilitato l'integrazione tra codice, visualizzazioni e documentazione, rendendo il progetto facilmente leggibile e modificabile.

2.2 Pulizia dei dati

Prima di procedere con l'analisi vera e propria, ci siamo concentrati su un'attenta fase di pulizia dei dati, essenziale per garantire la coerenza e l'affidabilità dei risultati. Inizialmente abbiamo verificato la presenza di duplicati all'interno dei diversi dataset e li abbiamo rimossi utilizzando il metodo `drop_duplicates()`. Successivamente, abbiamo eliminato i record contenenti valori nulli (`NaN`), soprattutto nei dataset relativi agli attori e ai film, per evitare inconsistenze nei grafici e nelle statistiche.

Nel caso del dataset dei film (`movies_df`), abbiamo inoltre effettuato un'operazione di selezione e rinominazione delle colonne, mantenendo solo i campi rilevanti per l'analisi, come l'ID, il nome, la data di uscita, la durata e il rating. Le colonne `name` e `date` sono state rinominate rispettivamente in `film_name` e `film_date`, rendendo più chiara la loro interpretazione durante la manipolazione dei dati. Queste operazioni hanno permesso di costruire una base solida, pulita e coerente, su cui poter

sviluppare interrogazioni e visualizzazioni senza rischiare distorsioni dovute a errori nei dati grezzi.

2.3 Rappresentazione visiva

Per la maggior parte delle domande che non prevedevano una componente geografica, ci siamo concentrati sulla selezione della rappresentazione visiva più adatta a evidenziare il tipo di informazione cercata. Ogni domanda ha richiesto un'interpretazione specifica dei dati, e per ognuna abbiamo scelto un tipo di grafico in grado di comunicare al meglio la risposta in modo intuitivo, leggibile e coerente con il contenuto. Laddove era necessario confrontare quantità discrete, come nel caso della distribuzione dei generi o del numero di film per attore, abbiamo utilizzato grafici a barre. Per osservare tendenze temporali, come l'evoluzione della durata media dei film, abbiamo preferito grafici a linee, mentre per confronti più analitici tra variabili (es. recensioni della critica vs. pubblico), abbiamo optato per scatter plot o grafici a colonne affiancate.

Nel caso delle domande che invece richiedevano una rappresentazione geografica, abbiamo utilizzato la libreria GeoPandas, che ci ha permesso di integrare dati tabellari con mappe georeferenziate e visualizzare visivamente la distribuzione delle produzioni cinematografiche nei diversi paesi del mondo. Tuttavia, in questa fase abbiamo riscontrato alcune difficoltà tecniche: i nomi dei paesi presenti nei dataset originali non erano sempre compatibili con la nomenclatura richiesta da GeoPandas. Alcuni stati venivano riportati con sigle o denominazioni parziali (ad esempio "USA" invece di "United States"), rendendo impossibile il corretto allineamento con la geometria della mappa. Per risolvere questo problema, abbiamo effettuato un'operazione di uniformazione dei nomi dei paesi, mappandoli secondo gli standard compatibili con i dataset geografici forniti da `GeoPandas`, così da poter visualizzare correttamente le informazioni sul planisfero.

Questa doppia attenzione – da una parte alla leggibilità dei grafici, dall'altra alla precisione delle mappe – ha guidato l'intero processo di visualizzazione, con l'obiettivo di offrire risposte chiare, informative e coerenti con il tipo di dato analizzato.

Capitolo 3

Riflessioni sul lavoro svolto

3.1 Divisione del Lavoro

L'attività di sviluppo è stata portata avanti in modo collaborativo e bilanciato tra i membri del gruppo. Inizialmente ci siamo confrontati insieme sui dataset forniti, analizzandone struttura e contenuti, e abbiamo discusso quali domande potessero essere interessanti da porre ai dati. Una volta individuate le aree di analisi, ci siamo divisi le domande da sviluppare, mantenendo però un confronto costante sull'avanzamento del lavoro. In diversi momenti abbiamo lavorato insieme, scrivendo codice fianco a fianco, il che ha favorito lo scambio di idee e l'uniformità delle soluzioni adottate.

Anche nelle fasi svolte individualmente, abbiamo sempre condiviso i progressi e risolto insieme eventuali problematiche, aiutandoci reciprocamente. Questo approccio ha permesso di mantenere coerenza tra le varie componenti del progetto e di affrontare le difficoltà in modo rapido ed efficace, garantendo un avanzamento fluido e coordinato.

3.2 Considerazioni finali

Lo sviluppo di questo progetto ci ha permesso di rafforzare le nostre competenze nell'ambito della data analysis, sperimentando strumenti avanzati per la manipolazione, pulizia e visualizzazione di grandi quantità di dati. Abbiamo imparato a utilizzare nuove tecnologie, come GeoPandas per la rappresentazione geografica, e a lavorare con librerie di visualizzazione come Seaborn e Plotly per comunicare i risultati in modo chiaro e comprensibile.

Nonostante le difficoltà incontrate, come l'incoerenza nei dati o i problemi di compatibilità tra dataset e mappe, siamo sempre riusciti a trovare soluzioni efficaci grazie a un approccio incrementale: ogni funzione, analisi o grafico è stato sviluppato e integrato gradualmente, testandone il funzionamento prima di passare allo step successivo. Questo metodo ci ha permesso di mantenere stabilità nel progetto e di risolvere rapidamente gli errori.

L'esperienza è stata estremamente formativa, sia sul piano tecnico che su quello organizzativo. Abbiamo migliorato le nostre capacità di collaborazione, divisione dei compiti e gestione autonoma del lavoro, portando a termine un progetto strutturato e coerente.

Bibliografia

[1] GeoPandas Development Team. Geopandas — user guide: Input/output. https://geopandas.org/en/stable/docs/user_guide/io.html, 2024.

[2] Python Software Foundation. Python language reference, version 3.11. https://docs.python.org/3/, 2024.

[3] The pandas development team. pandas — python data analysis library. https://pandas.pydata.org/pandas-docs/stable/, 2024.