Relazione Progetto IUM 2025

Ganino Domenico

Matr: 870669

Analisi dei datasets sui film

Introduzione

Questa relazione, lungi dall'essere completa ed esaustiva, si articola in molteplici passaggi volti ad esporre i metodi e le tecniche di analisi applicate. L'obiettivo primario è stato lo studio approfondito dei dataset al fine di estrarre le informazioni più rilevanti che hanno consentito una parziale ricostruzione della storia del cinema, dei periodi critici affrontati dai produttori e delle strategie di adattamento adottate per esprimere la propria visione artistica nel contesto circostante. L'analisi di tali informazioni ha inoltre permesso di inferire tendenze relativa alle preferenze del pubblico, ai fattori di attrazione e disinteresse, nonché ai gusti e alle curiosità generali.

L'analisi non può considerarsi conclusa poichè dei dataset così vasti e un tema ampio come quello cinematografico trascendono gli scopi del progetto.

Ciononostante, è stato possibile delineare un quadro generale dei protagonisti e delle dinamiche che hanno contribuito a plasmare la storia del cinema.

Utilizzo dei dataset

Il progetto include diversi file che richiedono un'analisi preliminare prima del loro effettivo impiego.

Un esempio significativo è rappresentato dal file ***movies.csv.***

Movies.csv

Fase di ispezione e preparazione dei dati

Il file offre una panoramica dei film, includendo elementi che costituiranno il fulcro delle analisi successive, quali rating, il minutaggio, e l'anno di produzione.

Un'ispezione preliminare ha rivelato la struttura del dataset e i tipi di dati iniziali (941.597 record, 7 colonne). Le colonne comprendono "id", "name", "date", "tagline", "description", "minute", "rating".

Per una gestione più appropriata dei dati, le colonne *"**name**"* e *"**date**"* sono state convertite rispettivamente nei tipi String e Int. Sebbene la gestione delle date in python sia tipicamente affidata al modulo ***datetime***, in questo caso, data la presenza del solo anno di produzione, un tipo di dato intero è risultato sufficiente e più efficiente in termini di memoria e velocità di elaborazione.

È stata effettuata una verifica dei valori nulli, evidenziandone la quantità; tuttavia si è deciso di mantenere integro il dataset, senza procedere alla sostituzione di tali valori con medie statistiche o all'eliminazione delle tuple contenenti valori nulli. Inoltre, si è preferito rimuovere la colonna "tagline" escludendo la sua utilità nelle analisi.

Fase di analisi

Questa fase persegue un duplice scopo: in primo luogo, ottenere una comprensione preliminare della natura dei dati attesi e, secondariamente, predisporre un dataset pulito per le elaborazioni successive.

Inizialmente, è stata condotta una ricerca di outliers, impostando una soglia elevata per la durata al fine di identificare eventuali errori. A seguito di un'attenta valutazione, il dataset non ha rivelato la presenza di outliers significativi. Sebbene includa serie televisive, caratterizzate, per ovvie ragioni, da una durata complessiva superiore rispetto ai film, e alcuni lungometraggi con un minutaggio totale atipico, si è concluso che tali valori, sebbene anomali per un film tradizionale, rappresentino degli elementi utili.

Per le analisi sono stati impiegati degli strumenti offerti dalle librerie ***pandas e seaborn***.

Pandas ha consentito la creazione agevole di nuove colonne nel dataset per categorizzare i dati.

Esempio:

min_range = [0, 30, 60, 90, 120, 180, 240, 1000, 10000, float("inf")]

labels = ["0-30", "30-60", "60-90", "90-120", "120-180", "180-240", "240-1000", "1000-10000", "10000+"]

movies_df_copy["minute_range"] = pd.cut(movies_df_copy["minute"], bins = min_range, labels = labels)

Indicando rispettivamente come parametro la serie da prelevare, i bordi degli intervalli con bins e le etichette con labels, è stato possibile creare, da una variabile numerica, una categoria di appartenenza

Inoltre, i metodi aggregati di pandas, quali **mean(), count(),** etc, hanno permesso di condurre analisi su larga scala in modo semplice e rapido.

Seaborn ha reso possibile la generazione di grafici precisi e altamente esplicativi, utilizzando tecniche efficaci per combinare diverse tipologie di informazioni; un esempio è il countplot, ottimo per visualizzare la frequenza delle osservazioni in categorie discrete.

Storia dei dati

La maggior parte dei prodotti audiovisivi presenta una durata inferiore ai 120 minuti, con picchi di frequenza nelle categorie 0-30, 60-90 e 90-120 minuti. Questa distribuzione riflette la tipica durata dei film cinematografici e dei cortometraggi/episodi TV.
Alcuni film, come già accennato in precedenza, presentano una durata molto elevata (1000-10000 minuti e 10000+), probabilmente rappresentativi di serie TV complete o progetti artistici particolari.

Utilizzando un grafico per comparare il **rating medio per range di minutaggio** mostra che i prodotti nelle categoria che vanno da 180 -10000 minuti ottengono un rating medio più alto (circa 3.6), sebbene il numero di film in questa categoria sia basso.

Le categorie con durata breve (0-30) o estremamente lunga (10000+) tendono ad avere rating medi leggermente inferiori rispetto a categorie intermedie o lunghe (ma non estreme).

Un'ulteriore rappresentazione grafica mostra una crescita esponenziale della produzione cinematografica nel tempo, tuttavia confrontando tale andamento con il rating medio, si nota una flessione del gradimento complessivo. Questo fenomeno è verosimilmente attribuibile a una maggiore diversificazione dei temi

trattati, alle recensioni da parte di un pubblico più vasto e non per ultimo, a una produzione di massa che comporta un fisiologico calo dei film di eccellenza in rapporto alla quantità totale.

ANALISI

Le analisi centrali, che integrano la maggior parte dei datasets del progetto, sono condotte nel file ***Main.ipynb***. In questo file, si è proceduto all'analisi dei datasets singolarmente o in coppia; successivamente, a seguito di una fase di scrematura, sono stati integrati altri datasets per arricchire l'analisi con nuovi dati.

Come per le analisi sul dataset ***Movies.csv***, gli strumenti prevalentemente utilizzati sono ***pandas e seaborn***. Le analisi effettuate in questa sezione sono più complesse e sfruttano un maggior numero di variabili. Ad esempio, in presenza di colonne contenenti liste di elementi, è stato utilizzato il metodo **explode()**. Tale metodo consente i trasformare ogni elemento di una lista (o array), contenuta in una cella di una Series, in una nuova tupla separata, duplicando i valori delle altre colonne per mantenere la coerenza del dataframe.

Esempio:

```
# mostriamo in un grafico il numero di film per genere

plt.figure(figsize=(12,6))

# per rappresentare sul grafico i generi dei film e contarli singolarmente dobbiamo prima usare una explode sui generi in quanto vengo rappresentati, per ogni film, come una lista di generi

ax = sns.countplot(data = movies_w_genres_df.explode("genre"), x = "genre", order = movies_w_genres_df.explode("genre")["genre"].value_counts().index)
```

Un altro esempio significativo è l'utilizzo frequente di **apply()** per valutare individualmente ogni elemento contenuto in una determinata colonna.
Python non ha un metodo vettorizzato diretto che possa iterare attraverso gli elementi di liste all'interno di una Series e applicare un test di appartenenza. È qui che apply() diventa indispensabile.

Esempio:

```
# eseguiamo una breve analisi esplorativa sul numero di film per genere ad esempio le commedie

tot_comedy = movies_w_genres_df["genre"].apply(lambda x: "Comedy" in x).sum()

print(f"Totale delle commedie: {tot_comedy}")

In questo caso apply consente di applicare una funzione lambda su una Series ritornando una valore booleano
```

Un utilizzo profondo di seaborn ha permesso di produrre dei grafici più complessi, capaci di aggregare i dati dei diversi datasets, facilitandone la lettura e la comprensione finale.

L'impiego di ***violinplot*** ha consentito di visualizzare la distribuzione dei dati analizzati e le variazioni delle medie e dei relativi intorni.

Gli ***scatterplot*** sono stati utilizzati per rappresentare le relazioni tra gli elementi di un dataset e per stimare l'andamento e il trend dominante.

Le ***heatmap*** si sono rivelate utili per rappresentare in una matrice le correlazioni tra le varie colonne dei datasets.

L'ulteriore utilizzo della libreria ***plotly.express*** ha permesso di produrre dei grafici dinamici e interattivi, consentendo lo studio dei dati per diverse categorie.

Storia dei dati

L'integrazione e l'analisi incrociata dei vari datasets hanno permesso di studiare e capire l'evoluzione del cinema nel corso del tempo.

Iniziando con l'analisi dei generi cinematografici, si è dedotto che i generi più utilizzati sono *"Drama"*, *"Comedy"* e *"Documentary"*.

I generi che riscuotono il maggior apprezzamento sono *"Documentary"* e *"Music"* mentre il genere *"horror"* tende a essere il meno gradito.

L'utilizzo di intervalli di anni di produzione ha permesso di analizzare i periodi storici e culturali che l'umanità sta attraversando. La nascita convenzionale del cinema è datata 1895 con i fratelli Luimiere, si osserva che, prima di quel periodo le uniche pellicole classificabili come film erano degli estratti di concerti tramite sequenze di immagini e documentari. Nei periodi successivi l'influenza dello stile gotico e il contesto storico delle guerre mondiali sembrano aver coinciso con un incremento delle valutazioni medie dei generi caratterizzati da atmosfere più cupe.

Inoltre, mediante **violinplot, scatterplot,** si è potuto constatare che nei primi anni della cinematografia i rating tendevano a essere relativamente omogenei e prossimi al valore medio, invece con il passare degli anni, la dispersione del rating aumenta fino a raggiungere dei picchi negli ultimi anni.

Lo studio di altri dati, come la durata media, suggerisce una forte influenza sullo stile e il genere; ad esempio, trame più complesse, tipiche dei generi cupi e drammatici, necessitano di più tempo per essere sviluppate, a differenza di documentari e film d'animazioni, che possono avvalersi di una durata inferiore per narrazioni più brevi e concise. È altresì comprensibile come, con il progredire degli anni, le tecniche di ripresa e tecnologie più avanzate abbiano consentito la creazione di lungometraggi sempre più consistenti, raggiungendo dei picchi negli ultimi anni. Questo potrebbe indicare una tendenza da parte dei registi a necessitare di maggior tempo per imprimere nello spettatore la trama del film.

Altre analisi hanno permesso una comprensione più approfondita del mondo cinematografico da prospettive inedite, quali lo studio statistico dei paesi più influenti e produttivi, delle lingue più utilizzate, degli studi di produzione di maggior successo e i generi prediletti, nonché delle caratteristiche degli attori scaturite da analisi sul genere preferito o i premi ricevuti, fino a tematiche fortemente attuali come le differenze di genere o *"gender gap"* e l'evoluzione nel tempo.

In conclusione, poche pagine non sarebbero sufficienti a descrivere tutte le deduzioni derivanti da questo studio che, sebbene preliminare e non universalmente generalizzabile, si presenta come un'analisi di facile comprensione e di impatto, la quale ha permesso di ottenere un quadro generale del mondo del cinema.