Relazione Interazione Uomo-Macchina

Dario Mangini - matricola 1047397

Repository github per la parte di IUM

Introduzione

In questa relazione vengono documentate le varie fasi del progetto, a partire dalla pulizia dei dati fino alla visualizzazione e analisi dei risultati, con l'obiettivo di rispondere alla domanda: "Quali fattori determinano il successo di uno studio cinematografico?" La relazione è strutturata in sezioni dedicate a ogni task tecnico, con una descrizione chiara delle soluzioni adottate, i problemi incontrati, e i risultati ottenuti.

Technical Tasks

1. **Pulizia dei dati:**
   Solution: La pulizia dei dati è stata organizzata in modo metodico, trattando ciascun dataset separatamente per garantire che ogni operazione fosse specifica e adeguata al tipo di dati analizzati. Per ogni dataset, le operazioni di pulizia seguono un flusso logico, gestendo prima i valori nulli, poi i valori duplicati, la conversione dei tipi di dati e la creazione di file CSV puliti. I valori mancanti sono stati trattati caso per caso, valutando la loro rilevanza per l'analisi. Ad esempio nei casi in cui era possibile sostituire i valori mancanti con un valore simbolico predefinito (ad esempio nei film senza rating è stato assegnato valore -1) ciò è stato fatto per preservare la coerenza dei dati. Per la rimozione dei duplicati sono stati rimossi in base a delle colonne chiave (ad esempio l'id o a delle combinazioni di colonne, talvolta tutte). Le colonne con i tipi dei dati sono state convertite per normalizzare i dati per evitare errori e per rendere i dati più consistenti e pronti per l'analisi.
   Il progetto è organizzato in maniera modulare, infatti la pulizia e l'analisi sono svolte in due file differenti.
   Issues: i valori nulli sono stati gli elementi di maggior difficoltà. C'è stato bisogno di ponderare il modo di gestirli, cercando di capire quando si potevano eliminare e quando era più conveniente sostituirli con un valore simbolico.
   Requirements: la pulizia soddisfa i requisiti della consegna, gestendo un grande numero di dati, utilizzando le tecniche apprese a lezione.
   Limitations: in alcuni casi per evitare di eliminare grandi porzioni di dati o informazioni presenti in altre colonne che potevano essere significative si è deciso di sostituire i valori nulli con un valore simbolico (ad esempio -1 per i rating dei film o 0 per gli anni di uscita del film.) Questo richiede ulteriori azioni di controllo e filtraggio dei dati durante l'analisi, per evitare di avere risultati distorti.

2. **Analisi dei dati:**

   Solution: questa parte del progetto è stato diviso in sezioni a seconda delle domande che ci si poneva. La domanda principale è *quali fattori determinano il successo di uno studio cinematografico?*. Per rispondere a questa domanda si è iniziato cercando quali sono gli studios che hanno prodotto più film, e li ho ribattezzati lungo il progetto con il nome di *top studios*. Ho scelto di limitare l'analisi a questi studios perché essendo i più prolifici garantivano una certa consistenza dei dati durante l'analisi per poter generalizzare e trarre conclusioni. Ad esempio non potevo scegliere i dieci studi con la più alta valutazione media, perché magari uno studio poteva aver prodotto un solo film ma con un'alta valutazione, ciò però non implica che sia uno studio di successo e non permette di trarre conclusioni generali.
   Ho cercato di variare il più possibile con l'utilizzo dei grafici, usando sia grafici statici, che interattivi che rappresentazioni geografiche.
   Sono state utilizzate diverse delle librerie trattate a lezione per rappresentare i grafici in maniera efficiente:

   - La libreria *Seaborn* è stata utilizzata per creare visualizzazioni statiche avanzate che enfatizzavano i pattern e le distribuzioni dei dati. Le sue funzionalità principali, come l'integrazione con Pandas e la capacità di generare grafici complessi con poche righe di codice sono state fondamentali per il progetto. Ad esempio sono stati usati *countplot* e *boxplot* per confrontare la quantità e le durate dei film dei top studios. In altri casi sono stati utilizzati violin plot e heatmap per analizzare distribuzioni di rating e relazioni tra studios e generi. L'uso di palette personalizzate e annotazioni sui grafici ha migliorato la leggibilità e l'estetica delle visualizzazioni. In questo modo *Seaborn* ha permesso di evidenziare tendenze chiave in modo chiaro, rendendo i dati complessi più comprensibili.
   - *Plotly* è stato usato per creare grafici interattivi, offrendo agli utenti la possibilità di esplorare i dati in modo dinamico. Ad esempio sono state create delle *treemap* per visualizzare la distribuzione di film per studio e paese, evidenziando le proporzioni in modo intuitivo. Un altro caso può essere ad esempio il *lineplot* che permetteva di visualizzare il flusso della produzione dei film lungo gli anni, che attraverso la possibilità di isolare i singoli studios e l'uso di hover interattivi rende la comprensione dei dati facile e accattivante, cosa che sarebbe stata più difficile con una visualizzazione statica.
   - *Folium* è stato utilizzato per creare una mappa interattiva che mostrava la distribuzione geografica della produzione cinematografica. La mappa coropletica evidenzia i paesi con il maggior numero di prodotti per i top studios. I cluster di marcatori hanno mostrato informazioni dettagliate sui paesi e sui contributi alla produzione. *Folium* ha permesso così di analizzare il contesto geografico in modo visivo, rendendo chiare le tendenze globali della produzione cinematografica.

   Come già anticipato l'analisi è suddivisa in sezioni, in base alle domande poste. Ogni grafico è significativo è risponde a una domanda o un aspetto di quella domanda.

Le sezioni sono divise in questo modo:

- Perché questi studios hanno prodotto più film? La longevità potrebbe essere un fattore determinante?
- Quantità è sinonimo di qualità?
- Il genere influisce sulla qualità? Più diversificazione equivale a meno qualità?
- Il Paese di produzione influisce sul successo dei top studios? Che strategia adottano?
- Come scelgono gli studios la durata di un loro prodotto?
- Conclusioni finali.

La scelta di queste domande è stata considerata logica e cerca di coprire il più possibile i fattori principali.
Issues: ci sono state alcune difficoltà con il database in questione. Infatti il fatto che non tutti i dataset avessero un identificatore univoco che poteva funzionare come chiave esterna ha limitato l'utilizzo dei dataset a quelli del main dataset, anche se è stato fatto un tentativo con gli Oscar ma con dati che non apparivano consistenti.
Requirements: si è cercato di utilizzare il più possibile le librerie spiegate durante le lezioni. Viene analizzata una sezione particolare del dataset (in questo caso i 10 studios che hanno prodotto più film). Ho cercato di variare il più possibile con i tipi di grafici, cercando di non usare soltanto bar charts e includendo visualizzazioni geografiche.
Limitations: il gran numero di valori nulli che vengono ignorati può rendere l'analisi non totalmente veritiera in confronto alla realtà Magari se ci fossero stati più valori validi nel rating si sarebbe giunti a conclusioni diverse. Inoltre l'analisi degli outlier per la durata dei film mostra che certe serie TV sono state salvate nel database con un unico valore che rappresenta il valore della somma della durata degli episodi della serie, analizzando però i dataset dei generi si può notare che nel genere TV Movie sono presenti degli episodi singoli di alcune serie. Queste diverse strategie nel salvataggio dei dati potrebbero aver portato ad alcune considerazioni non del tutto corrispondenti alla realtà, soprattutto per quanto riguarda le serie TV americane.

Conclusioni

L'analisi ha dimostrato come la chiave del successo dei top studios dipenda dal conoscere il contesto in cui si trova e dal conoscere il tipo di pubblico alla quale ci si rivolge. Questo determina significativamente la scelta dei generi, di diversificazione e non solo. Alcuni studios basano il proprio successo sulla quantità e su un pubblico globale, altri eccellono attraverso la qualità e la specializzazione. Uno studio di successo deve saper bilanciare tra quantità, qualità, diversificazione e specializzazione, per adattarsi al pubblico di riferimento e alle condizioni di mercato.

Extra information

Allego oltre al report e alla soluzione degli screen dei grafici interattivi, che temo non si vedano quando consegno il Jupyter Notebook eseguito.

Bibliografia

https://chatgpt.com/
https://www.w3schools.com/