Relazione Introduzione alla Data Science

Questo progetto mira ad esplorare e comprendere i fondamenti della Data Science attraverso l'analisi di due importanti piattaforme: Amazon e Disney. Il processo di analisi è strutturato in diverse fasi che abbracciano vari aspetti del ciclo di vita dei dati.

- **Preparazione** di cui:
  - integrazione (punto 1)
  - pulizia (effettuata in diversi punti)
  - trasformazione (punto 2)
  - esplorazione (punto 3)
- **Esplorazione**
  - Test statistico (punto 4)
  - OLAP (punto 5)
- **Predizione**
  - Metodi predittivi

1. Integrazione

Ogni piattaforma dispone di due dataset, per un totale di 4. L’integrazione tra i dataset ha permesso una maggiore compattezza e completezza dei dati, ha inoltre semplificato l’analisi e ridotto confronti superflui.

Integrazione tabella 1 e 2

Il rapporto tra amazon1 e amazon2 rispetto a disney1 e disney2 è equivalente e per questo sono state effettuate per entrambe le coppie le stesse operazioni di integrazioni.

Di seguito quali.

Per l’integrazione dei dati è stato scelto di effettuare un merge left al fine di ottenere un dataframe che mantenesse le informazioni della tabella 1.

Per poter ottenere una integrazione soddisfacente si è successivamente svolta la pulizia dei dati. Sono state eliminate le colonne che non portavano informazione e che rappresentavano lo stesso dato con un nome diverso.

La scelta di integrare age_certification e rating è motivata dalla presenza in entrambe di circa il 60% di valori nulli. La stessa operazione di integrazione è stata svolta per altre colonne, in alcuni casi è stato necessario rendere univoco il formato. Ad esempio è stato modificato Movie in MOVIE e Tv show in SHOW nella colonna type integrata

Per quanto riguarda date_added, 99% di valori nulli, è stato valutato eliminabile. Non è stata effettuata nessuna rimozione perché è una delle colonne su cui deve essere svolto il test statistico.
Le colonne runtime e season sembrano avere una buona percentuale di valori nulli ma la realtà è che sono complete al 100% rispetto al numero di SHOW.

Integrazione Amazon e Disney

La scelta di concatenare Amazon e Disney è stata presa per semplificare l’analisi complessiva e per avere un insieme di dati unico e coerente. La colonna combinata non ha sostituito altre piattaforme, in questo modo le tabelle possono essere usate per ulteriori operazioni e confronti, in quanto la pulizia dei dati è stata già effettuata.

Per l’integrazione complessiva è quindi stato scelto il merge outer al fine di non perdere dati puliti. Non è stato applicato il merge inner per non perdere date_added e duration, colonne che sarebbero servite in seguito.

Infine è stata svolta un’ulteriore pulizia dei dati per ottenere una formattazione unica. Per la colonna reccomended_age, età consigliata per vedere un programma, alcune sigle sono state associate ad altre. In questo modo le classifiche sono molto più chiare, questo non è in nessun modo un elenco reale di età consigliate. Di seguito le correlazioni, ogni sigla ha un significato preciso e più profondo di quello generale.

NR	NR, TV_NR, NOT_RATE, UNRATED
G	G, ALL, TV-G
TV-Y	TV-Y
7+	7+, TV-Y7
PG-7	TV-Y7-FV
PG	PG, TV-PG
PG-13	PG-13
13+	13+
PG-14	TV-14
PG-17	R
16+	16+, 16, AGES_16_
17+	NC-17
18+	18+, AGES_18_, TV-MA

2. Trasformazione

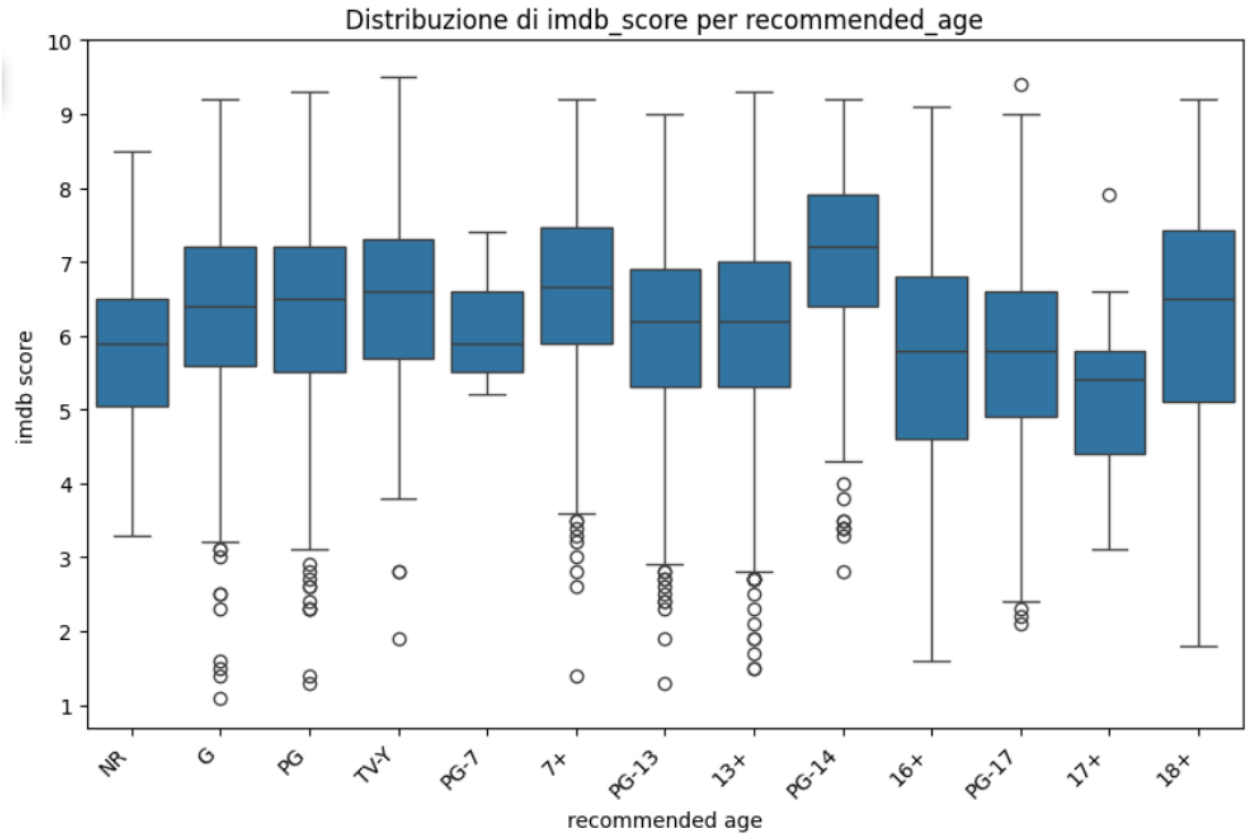
Per la fase di trasformazione è stato necessario suddividere la colonna date_added in due colonne che separavano l'informazione in mese e anno. L'operazione è stata svolta sulla tabella finale e sulle singole Amazon e Disney, per semplificare le analisi successive.

La fase di sostituzione della colonna genres in una contenente il numero di generi associati ad ogni dato poteva essere effettuato in almeno due modalità.

Interessante l'opzione di associare ogni nome di genere ad un numero e ottenere per ogni film una stringa di generi. Altrettanto rilevante associare ad ogni lista di generi un solo numero. Essendo un dato che non è stato poi approfondito nel corso dell'analisi del progetto, la scelta è ricaduta sul risultato più semplice, ovvero il secondo. Il codice per effettuare l'associazione di ogni singolo genere è stato lasciato nel codice in vista di possibili modifiche future. Può essere approfondita un'analisi su queste categorie.

3. Esplorazione

Analisi del boxplot dei punteggi IMDB divisi per età suggerite




```
Statistiche Descrittive per ogni Gruppo di recommended_age:

                 count      mean       std  min   25%   50%    75%  max
recommended_age
NR                75.0  5.882667  1.069778  3.3  5.05  5.90  6.500  8.5
G                891.0  6.341077  1.182872  1.1  5.60  6.40  7.200  9.2
PG              1182.0  6.352792  1.236506  1.3  5.50  6.50  7.200  9.3
TV-Y             143.0  6.481119  1.240794  1.9  5.70  6.60  7.300  9.5
PG-7              10.0  6.130000  0.781807  5.2  5.50  5.90  6.600  7.4
7+               226.0  6.535398  1.295937  1.4  5.90  6.65  7.475  9.2
PG-13            756.0  6.057011  1.241764  1.3  5.30  6.20  6.900  9.0
13+              793.0  6.033039  1.338755  1.5  5.30  6.20  7.000  9.3
PG-14            345.0  7.029565  1.178389  2.8  6.40  7.20  7.900  9.2
16+              549.0  5.643352  1.471759  1.6  4.60  5.80  6.800  9.1
PG-17           1278.0  5.684742  1.232842  2.1  4.90  5.80  6.600  9.4
17+               13.0  5.230769  1.263187  3.1  4.40  5.40  5.800  7.9
18+              764.0  6.242408  1.530222  1.8  5.10  6.50  7.425  9.2

media di 'count': 540.3846153846154
media di 'mean': 6.126456840663167
```

Le etichette sono ordinate in crescita, per semplificare la lettura. Non ci sono valori nulli in imdb_score e il 34% presenti in recommended_age vengono eliminati.

Il numero di osservazioni sono in media 540, ma variano da PG-7 (bambini di 7 anni accompagnati) che ne hanno 10, sino a 1278 per PG-17 (17enni accompagnati). Deducibile dalle differenti lunghezze del Tukey fence, si allunga e accorcia anche a seconda del numero di osservazioni. Molto evidenti nel boxplot sono PG-7 e 17+ per il basso numero di osservazioni.

La media dei voti è invece molto stabile, nel boxplot la linea si sbilancia tra 5 e 7 su 10 score. La media delle medie è infatti 6, questo significa che non ci sono valutazioni troppo discostate a seconda delle categorie di età.

Le classi con minore dispersione sono PG-7 e 17+, è da tenere presente che sono le stesse con un basso numero di quantità di dati (osservazioni).

Gli outliers sono presenti in grande quantità in G, PG, 7+, PG-13, 13+ e PG-14, per la maggior parte al di sotto della media con punteggi principalmente tra 2 e 3.

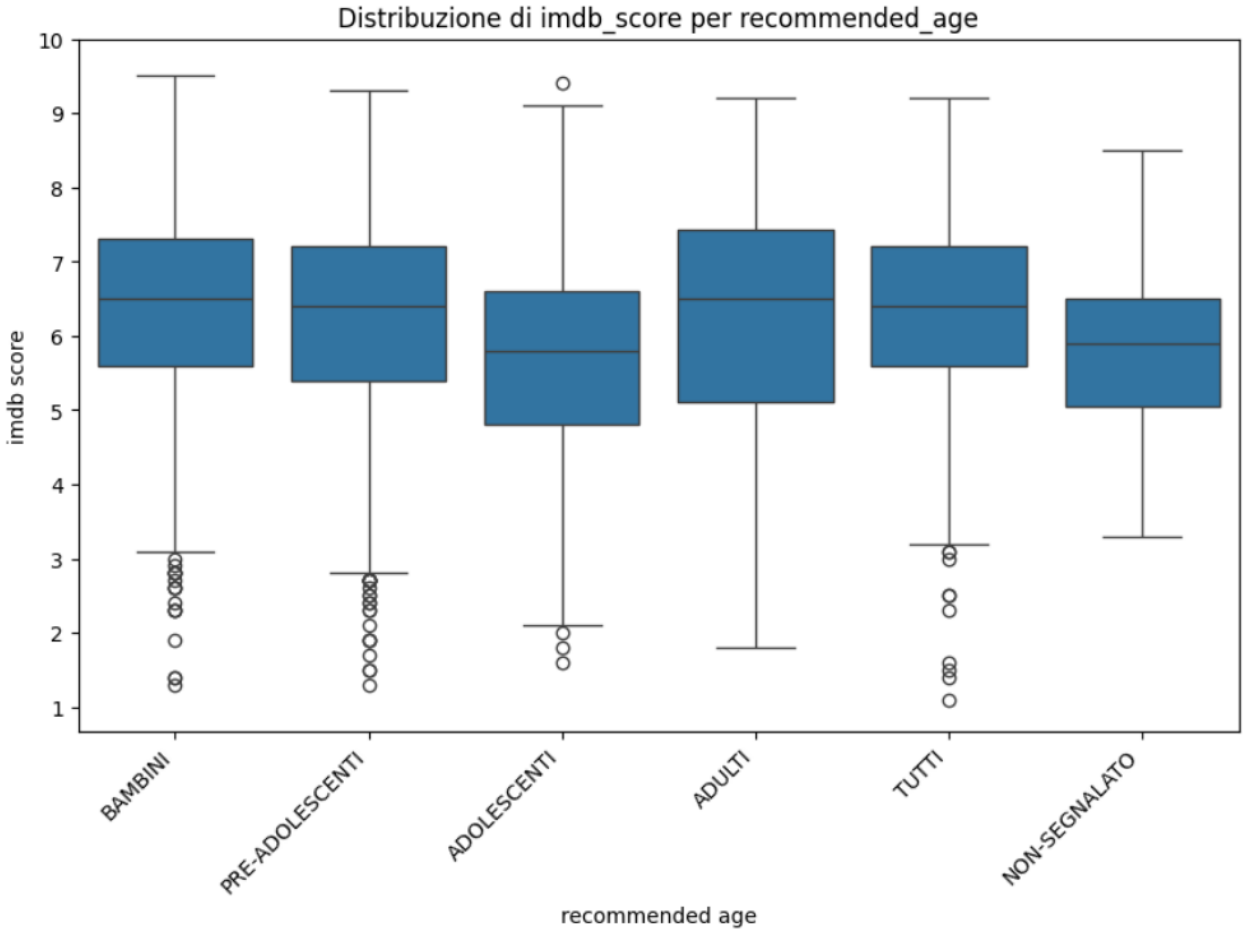
Le distribuzioni sono prevalentemente simmetriche, il boxplot è quasi sempre tagliato a metà tranne per PG-7 e 17+ che, nuovamente, sono anche le categorie con meno osservabili.

Il primo quartile è a sua volta più alto nel caso di G, PG, TV-Y, 7+ e PG-14. Questo non porta molta informazione essendo tutte categorie di età molto differenti.

PG-7 e 17+ sono le classi con minore dispersione e maggior asimmetria.

Analisi sulle categorie generalizzate

Proviamo un’analisi più generale per verificare la presenza di trend nelle distribuzioni dei punteggi a seconda delle età.




```
Statistiche Descrittive per ogni Gruppo di recommended_age:

                 count      mean       std  min   25%  50%    75%  max
recommended_age
BAMBINI         1561.0  6.389558  1.244465  1.3  5.60  6.5  7.300  9.5
PRE-ADOLESCENTI 1894.0  6.224129  1.327356  1.3  5.40  6.4  7.200  9.3
ADOLESCENTI     1840.0  5.669185  1.308824  1.6  4.80  5.8  6.600  9.4
ADULTI           764.0  6.242408  1.530222  1.8  5.10  6.5  7.425  9.2
TUTTI            891.0  6.341077  1.182872  1.1  5.60  6.4  7.200  9.2
NON-SEGNALATO     75.0  5.882667  1.069778  3.3  5.05  5.9  6.500  8.5

media di 'count': 1170.8333333333333
media di 'mean': 6.124837345141715
```

L’analisi precedente non ha portato a grandi conclusioni in quanto è poco rilevante parlare di generi così specifici. Proviamo ad unire le categorie in pochi gruppi per un analisi più semplice.

La suddivisione diventa: bambini, pre-adolescenti, adolescenti, tutti e non segnalato. Non segnalato conterrà anche i valori nulli.

Con queste nuove categorie viene evidenziato come la quantità di dati vari tra 800 e 1800, esclusi i NON-SEGNALATI che hanno 75 osservabili, molto sotto la media di 1171. NON-SEGNALATI non verrà considerato nei successivi commenti.

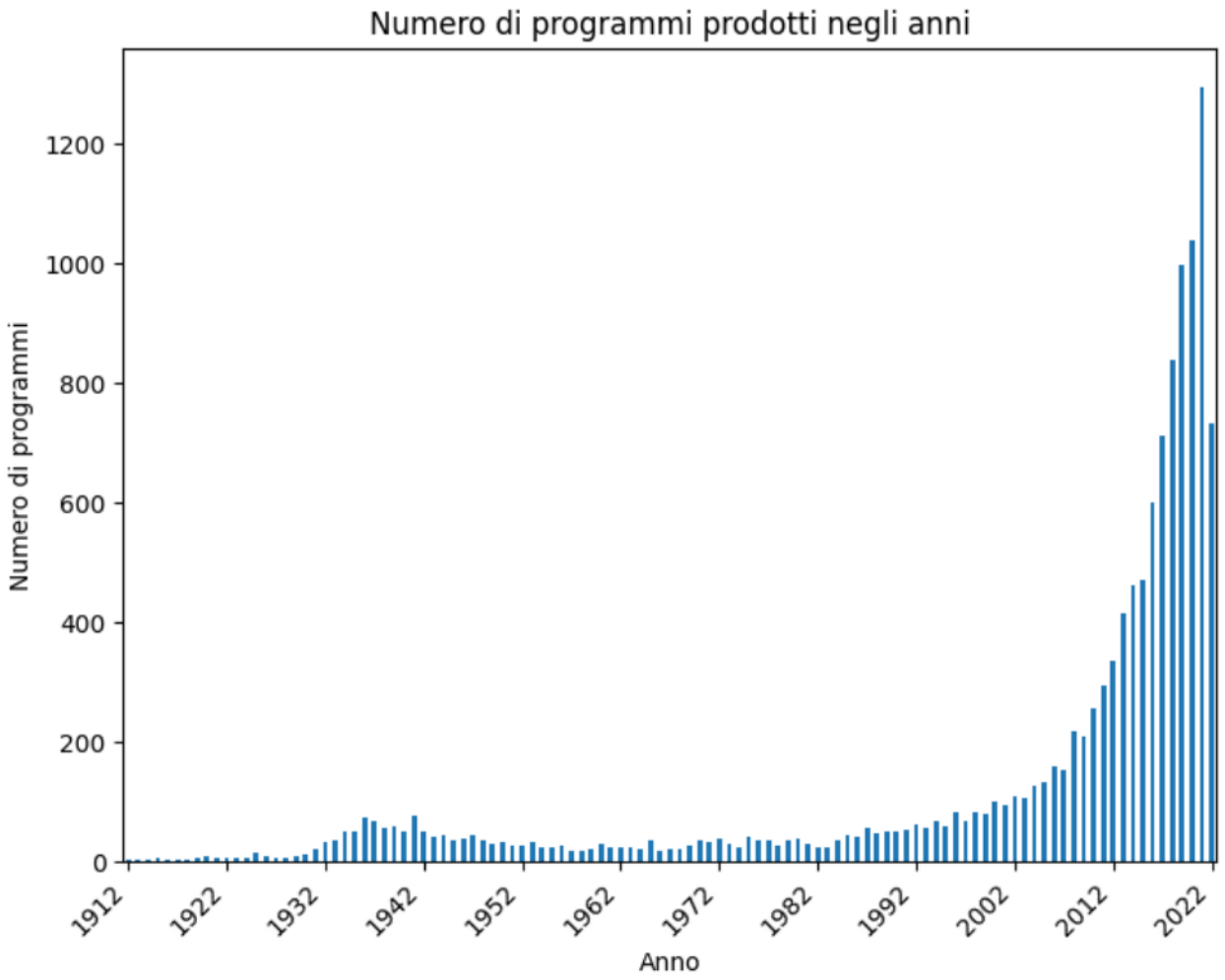
I punteggi per ogni categoria sono tra 5.5 e 6.5 con una media di 6. Evidente come la generalizzazione abbia appiattito i dati dove poteva invece mostrare un trend.

Tutte le categorie sono prevalentemente simmetriche, gli adulti presentano una leggera asimmetria positiva, ma nel complesso, le distribuzioni sono bilanciate.

La mancanza di outlier nelle categorie "Non Segnalato", "Adolescenti" e "Adulti" indica che la maggior parte dei dati si concentra all'interno della norma senza punti estremi.

In generale, l'analisi suggerisce che la generalizzazione delle categorie ha semplificato e appiattito la distribuzione dei punteggi, ha mostrato che non ci sono correlazioni importanti relative al rapporto età e punteggi IMDB. In supporto a questa tesi le Tukey fence molto ampie dimostrano una vasta diversità di punteggi in tutte le categorie.

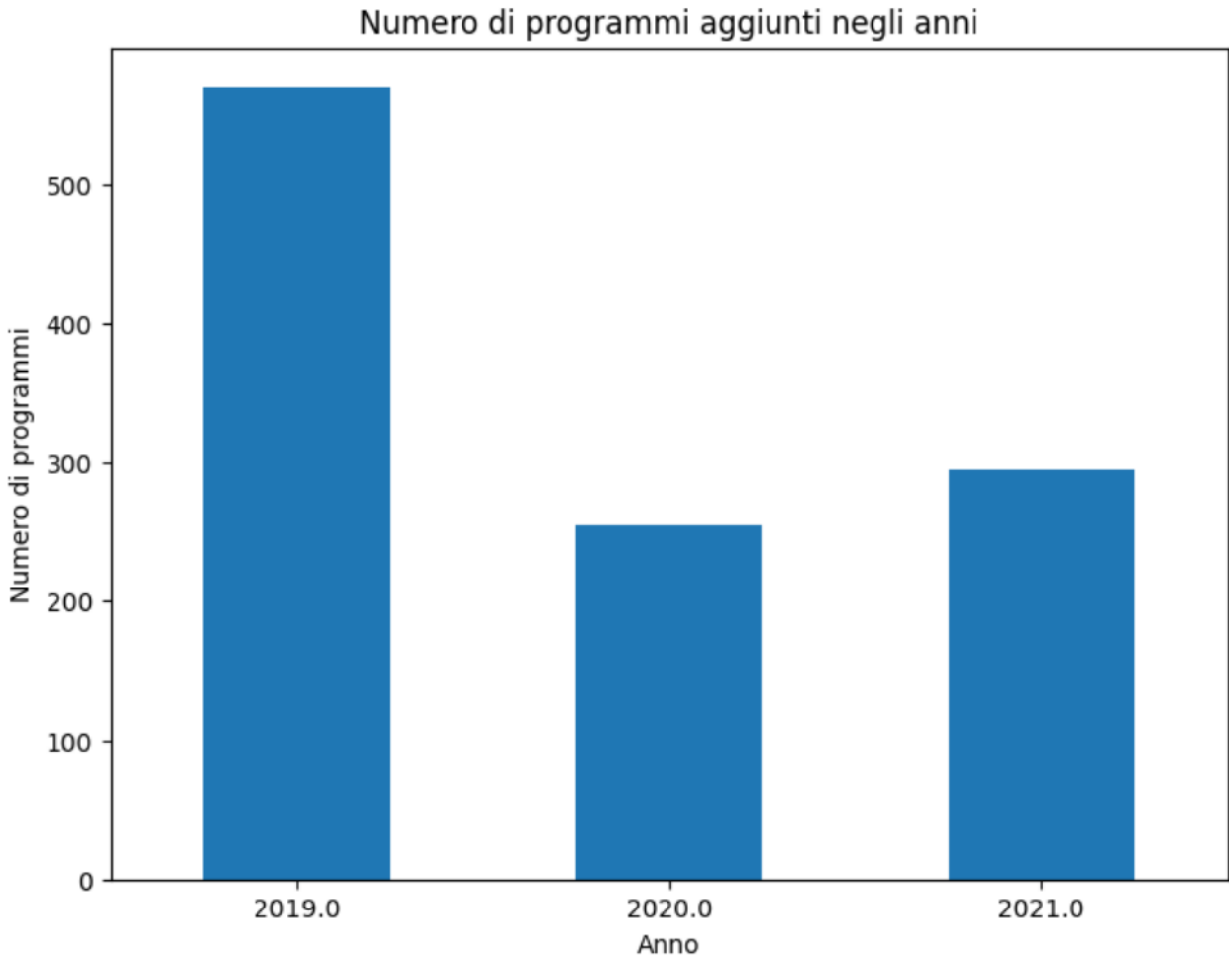
Numero di programmi prodotti negli anni



Il numero di programmi prodotti dalle case cinematografiche negli anni assomiglia ad una funzione logaritmica. Il calo nell'ultimo anno 2022 potrebbe essere a causa di una mancanza di aggiornamento sulla piattaforma, come se la raccolta dati fosse relativa ad aprile 2022.

Il picco tra gli anni 32 e 42 potrebbero essere conseguenza di propaganda di guerra o altre condizioni a favore.

Numero di programmi aggiunti negli anni



Il numero di film aggiunti alla piattaforma di streaming negli anni è distribuita su 3 anni dal 2019 al 2021.
La data di nascita della piattaforma di Disney combacia con il 2019 coerentemente con i dati raccolti. Mentre la piattaforma di Amazon video ha la sua data di nascita nel 2006, si suppone che non sia stata tenuta traccia.

4. Test statistico

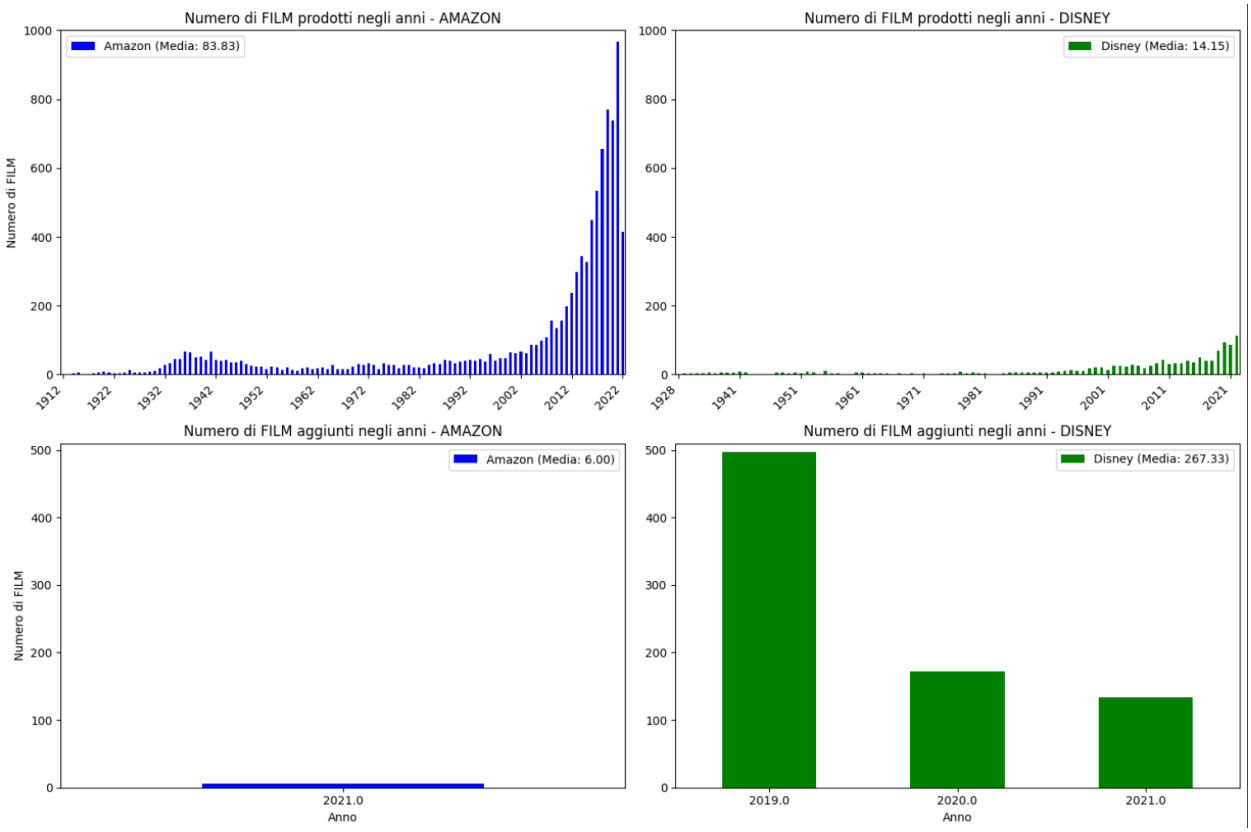
Le ipotesi per il test sono le successive.

L'ipotesi nulla mostra che i test statistici NON hanno differenze significative, quella alternativa che i test statistici hanno differenze significative

Il confronto è fatto su due test alla volta, quindi molto specifico. Dividendo Movie e Show, i test sono fatti tenendo anche separate le due differenti piattaforme. Sono presenti 4 test di seguito.

La quantizzazione degli anni è diverso e si basa sui valori specifici di minimo e massimo per ogni gruppo a confronto. Per Amazon ogni categoria contiene 22 anni, Disney 18.

Film

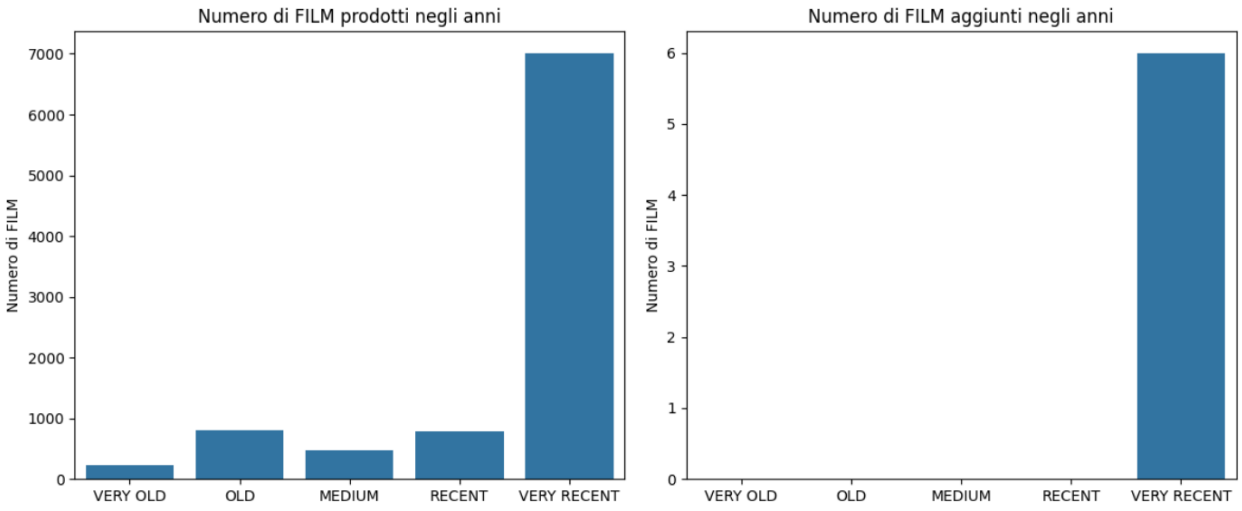


1. Film di Amazon

Per primo punto sono confrontati il numero di film prodotti con quelli aggiunti per la piattaforma di Amazon

I dati sono stati divisi in categorie per poter avere variabili categoriche invece che continue e poter applicare il test del chi-quadro

entrambe le colonne sono ora divise in VERY OLD, OLD, MEDIUM, RECENT e VERY RECENT

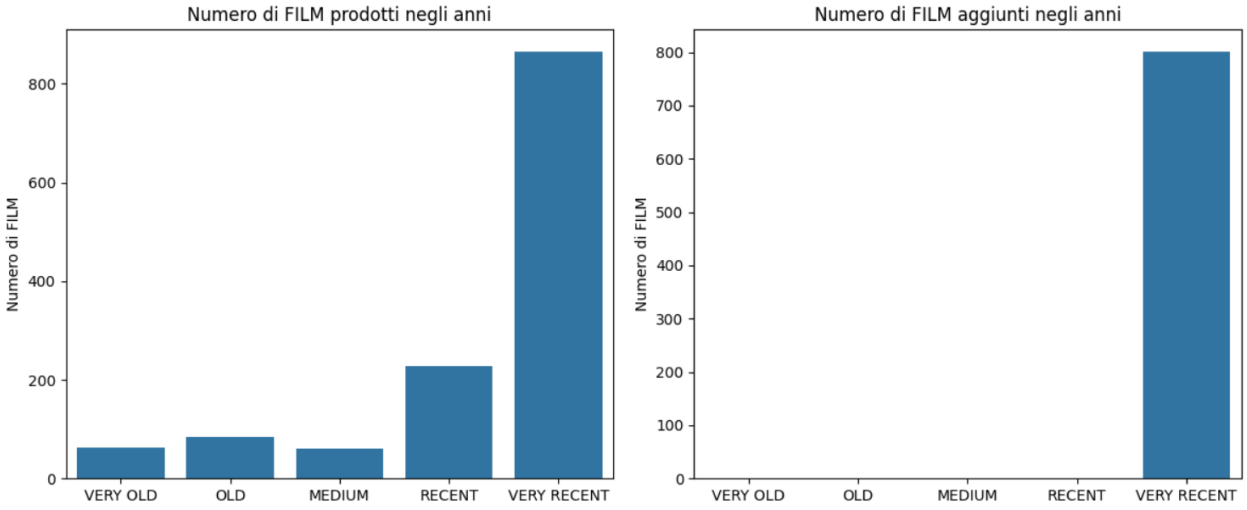


Attenzione alla differenza tra le due scale, film prodotti arriva sino a 7000 valori mentre aggiunti ne ha solo 6.

Se prima l'andamento logaritmico di prodotti era evidente, ora è molto meno chiaro.

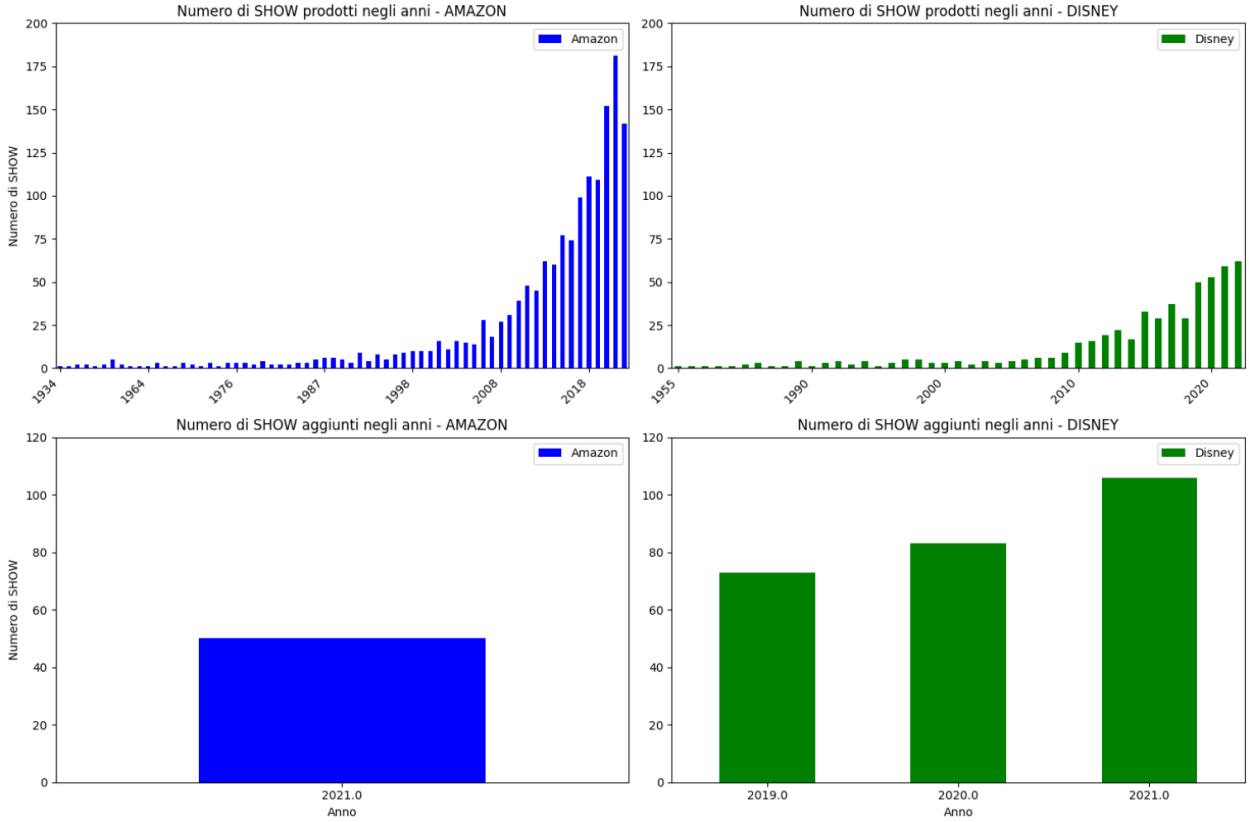
Il test con livello di significatività di 0,5 mostra che non ci sono prove convincenti di differenze significative tra le distribuzioni di anni di produzione e anni di caricamento sulla piattaforma. La discrepanza osservata potrebbe essere dovuta al caso e non fornisce sufficienti evidenze per supportare l'idea di differenze significative tra le distribuzioni considerate.

2. Film di Disney

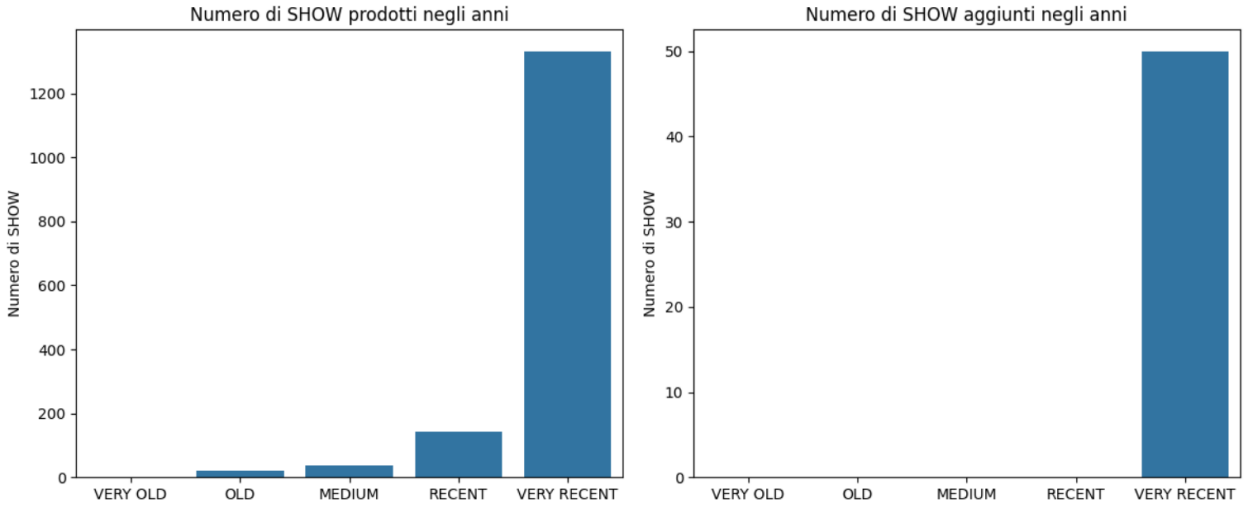


La statistica del test del Chi-quadro è molto elevata (162.3) e il valore p è estremamente basso, inferiore al livello di significatività di 0,05. Di conseguenza, rifiutiamo l'ipotesi nulla. Ciò indica che ci sono prove significative per sostenere l'idea di differenze rilevanti tra le distribuzioni di anni di produzione e anni di caricamento sulla piattaforma. I risultati suggeriscono che la discrepanza osservata è altamente improbabile che sia dovuta al caso, fornendo così evidenze sostanziali a favore dell'esistenza di differenze significative tra le distribuzioni considerate.

Serie TV

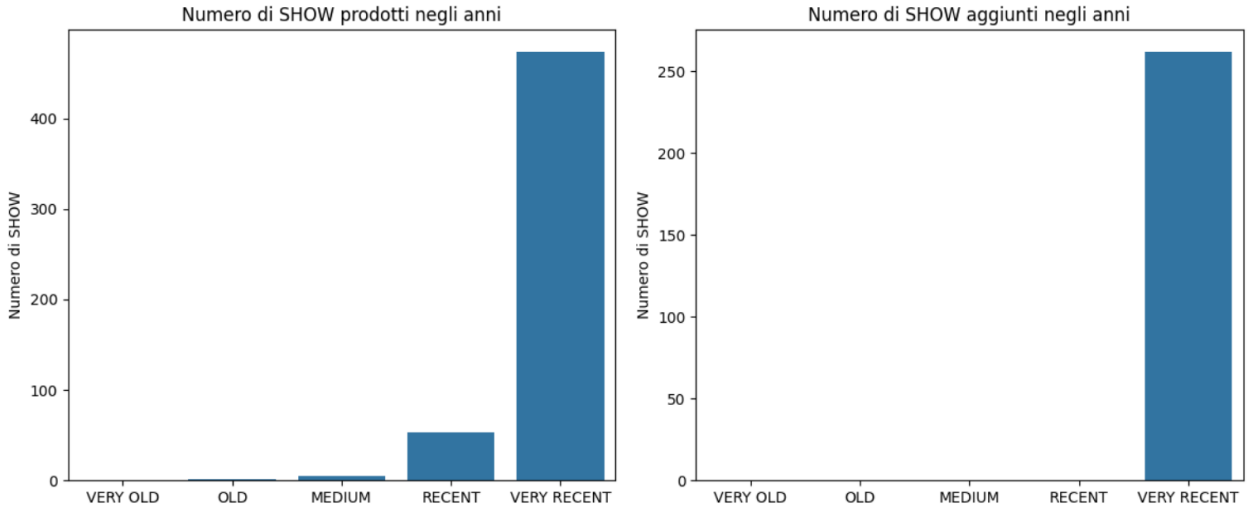


3. Serie TV di Amazon



La statistica del test del Chi-quadro è di 2.61, mentre il valore p associato è pari a 0.6. Il livello di significatività è del 5%, non possiamo rifiutare l'ipotesi nulla. Ciò indica che non abbiamo sufficienti evidenze per affermare l'esistenza di differenze significative tra le distribuzioni di anni di produzione e anni di caricamento sulla piattaforma. La discrepanza osservata potrebbe essere attribuita al caso, e non possiamo concludere in modo definitivo che vi siano differenze sostanziali tra le distribuzioni considerate.

4. Serie TV di Disney



La statistica del test del Chi-quadro è 17.6, con un valore p pari a 0.0005. Quest'ultimo è significativamente inferiore al livello di significatività del 5%, portandoci a respingere l'ipotesi nulla. Questo indica che ci sono prove statisticamente significative a favore

dell'esistenza di differenze rilevanti tra le distribuzioni di anni di produzione e anni di caricamento sulla piattaforma. La discrepanza osservata non è facilmente attribuibile al caso, fornendo così un sostegno sostanziale all'idea di differenze significative tra le distribuzioni considerate.

Conclusioni test statistico

	Film	Serie TV
Amazon	NON rifiuto	NON rifiuto
Disney	rifiuto	rifiuto

Ci sono differenze significative nel caso dei Film e delle Serie TV di Disney, per Amazon non c’è evidenza empirica per confermare questo. La differenza tra Amazon e Disney potrebbe essere analizzata ulteriormente per verificare eventuali trend.

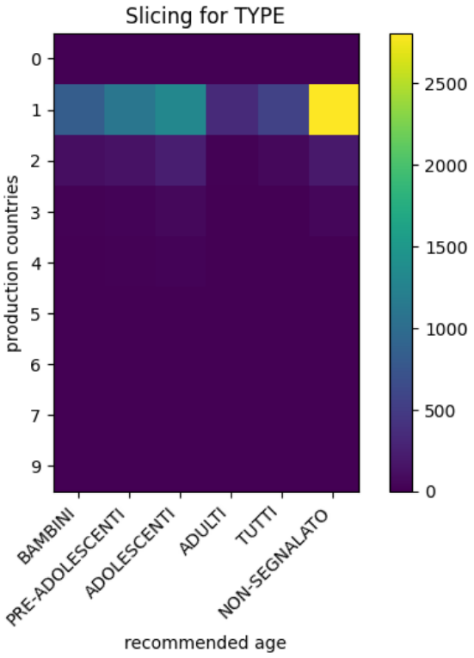
5. OLAP

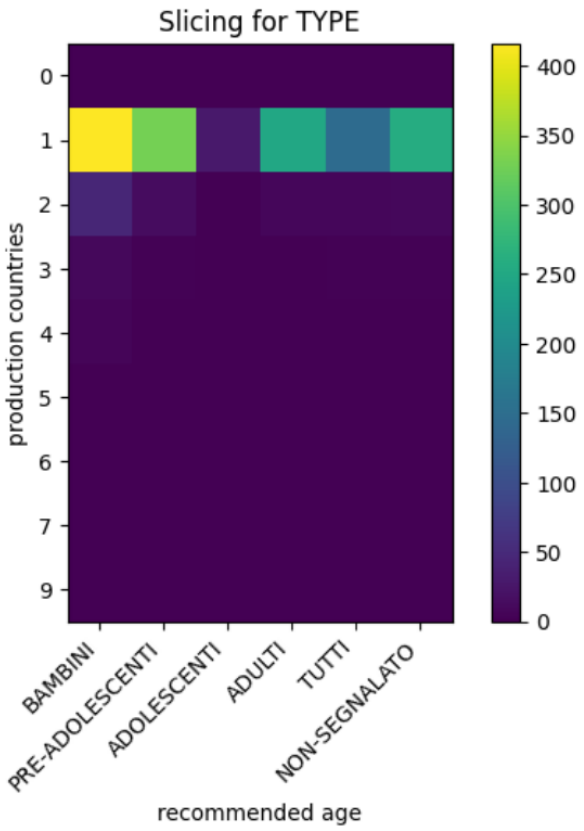
Grafico OLAP riguardante età raccomandata, tipologia e paese di produzione.

I dati sono stati aggregati in modo da ottenere delle categorie generiche.

Per i paesi di produzione è stato contato il numero di paesi in cui è stato prodotto ogni film piuttosto che considerare il paese stesso, questa scelta ha reso più semplice l’analisi dei dati in quanto altrimenti i risultati sarebbero stati troppi e ingestibili in una rappresentazione OLAP.

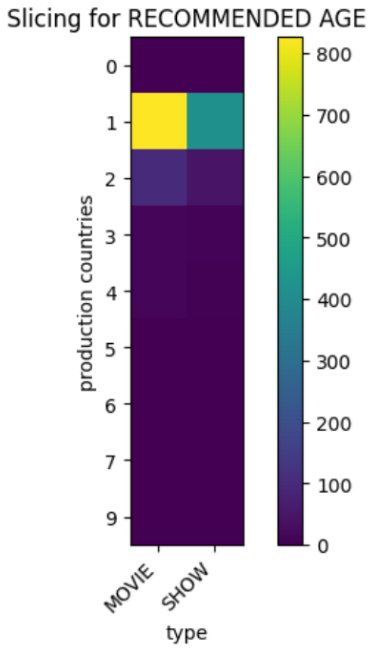
Dai seguenti grafici è evidente come la maggior parte dei programmi sono stati prodotti in un unico paese, a sinistra Film e a destra le Serie tv.



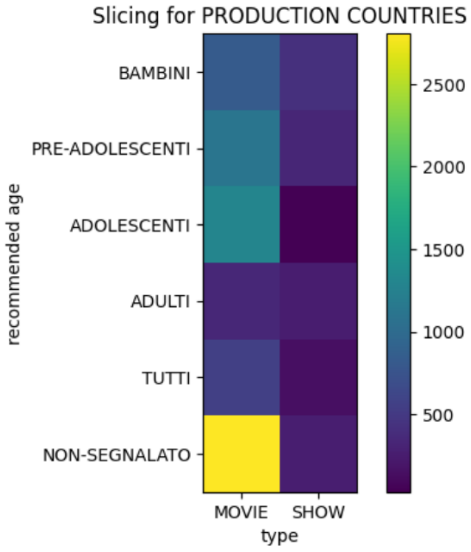


Il confronto tra le due scale mostra come la quantità di serie tv per categoria (max: 400) è minore di quella relativa ai film (max: 2500). Nel caso di serie tv il picco è il target bambini mentre per i film sono gli adolescenti, se ignoriamo i dati nulli o non segnalati.

Osservando meglio il successivo grafico si evince che la quantità di film per bambini è in realtà maggiore delle serie TV, si conferma l'idea invece che che la distribuzione di programmi è prevalente in un unico paese. Il grafico sottostante mostra infatti la distribuzione dei programmi per bambini rispetto ai paesi e la tipologia.

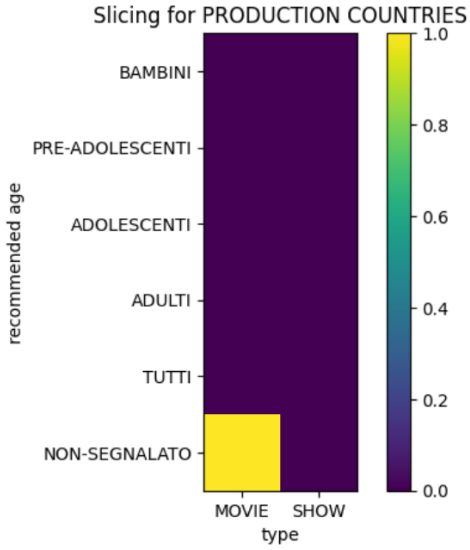


Nel grafico successivo invece è mostrata la distribuzione dei programmi prodotti in un solo paese variando età e tipologia.



Le statistiche sono offuscate dal picco di valori presenti per età non segnalata o valori nulli, è possibile notare che il valore più alto tra i rimanenti è adolescenti e in scala decrescente pre-adolescenti, bambini, tutti e adulti. Questo per quanto riguarda i film.

Questo grafico particolare mostra come sia stato salvato un singolo programma disponibile in 9 paesi diversi dal titolo ‘Don't Read This on a Plane’ disponibile in: NL, DE, RO, AU, HU, IT, GR, FR e PT.



Questa analisi evidenzia una prevalente produzione di programmi concentrata in un singolo paese, fenomeno che può essere attribuito a una combinazione di ragioni culturali, di mercato, linguistiche e di accessibilità. Inoltre, emerge un modello di consumo distintivo, con i film che trovano un pubblico prevalente tra gli adolescenti, mentre le serie TV sono preferite principalmente dai bambini. Tale preferenza potrebbe essere influenzata dalla presenza significativa di contenuti animati, come i cartoni animati.

Potrebbe rivelarsi interessante a questo punto verificare quale sia il paese con una maggior produzione di programmi.

6. Metodi predittivi

Descrittore composto da:

- imdb_score
- tmdb_score

- tmdb_popularity
- runtime

I valori nulli nella piattaforma Amazon sono sostituti con la media relativa ad ogni campo. Di seguito la percentuale di valori nulli per colonna.

AMAZON	Percentuale valori nulli
imdb_score	10.04%
tmdb_score	19.38%
tmdb_popularity	5.21%
runtime	0.00%

In percentuale i valori nulli non sono tanti

Lo stesso procedimento è applicato alla tabella di test Disney che disponeva delle seguenti percentuali.

DISNEY	Percentuale valori nulli
imdb_score	27.72
tmdb_score	7.30
tmdb_popularity	0.87
runtime	0.00

Il metodo predittivo scelto è la regressione logistica in quanto adatta a problemi di classificazione binaria in cui type può essere MOVIE o SHOW. Inoltre fornisce una previsione su probabilità, adatta per ottenere probabilità di appartenenza ad una classe.

I risultati dell’analisi sono i seguenti.

	Amazon - movie	Amazon - show	Disney - movie	Disney - show
Accuracy	0.95	0.95	0.77	0.77
Precision	0.97	0.84	0.96	0.56
Recall	0.97	0.80	0.69	0.94
F1-score	0.97	0.82	0.81	0.70

L’accuratezza mostra che le predizioni del dataset sono corrette al 77% per Disney.

L'analisi delle previsioni per il dataset Disney rivela interessanti risultati sull'efficacia del modello di regressione logistica. L'accuratezza complessiva del modello per il dataset Disney è del 77%, indicando che il 77% delle predizioni totali è corretto.

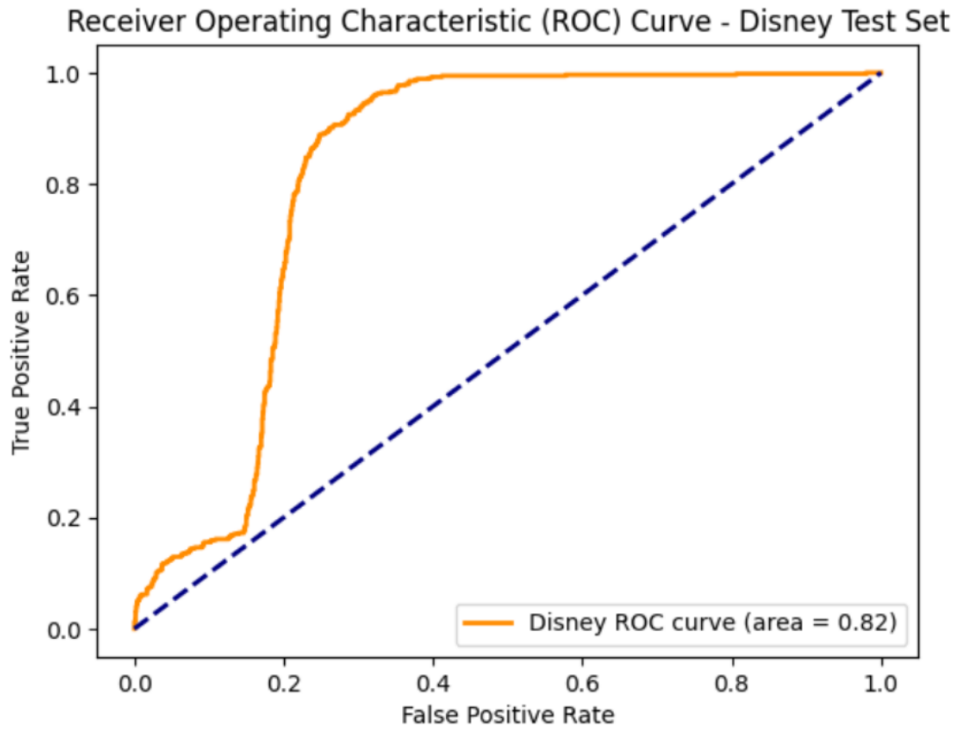
Esaminando la precisione, che misura la proporzione di predizioni positive corrette rispetto a tutte le predizioni positive, notiamo che il modello ha un'elevata precisione nel prevedere film Disney, con un valore del 96%. Tuttavia, la precisione nel prevedere show Disney è leggermente inferiore, attestandosi al 56%.

Analizzando il recall, che misura la capacità del modello di catturare tutti i casi positivi, notiamo che il modello ha un recall del 69% per i film Disney. Ciò significa che il 69% di tutti i veri film Disney è stato correttamente identificato dal modello. Per gli show Disney, il recall è significativamente più alto, raggiungendo il 94%, indicando che il modello è efficace nel catturare la maggior parte degli show Disney tra tutti i casi positivi.

Infine, l'F1-score, una metrica che bilancia precisione e recall, risulta essere del 0.81 per i film Disney e del 0.70 per gli show Disney. Questi valori suggeriscono che il modello ha una buona combinazione di precisione e recall per i film, ma potrebbe beneficiare di un miglioramento nella capacità di rilevare correttamente gli show.

In conclusione, il modello di regressione logistica sembra avere una buona capacità di predire i film Disney, ma mostra alcune sfide nella predizione degli show Disney, indicando la necessità di un'ulteriore analisi e possibile ottimizzazione.

Di seguito la rappresentazione grafica della regressione logistica attraverso la curva ROC



Il tasso di falsi positivi è molto basso in quanto i punti sono molto spostati a sinistra. I veri positivi a loro volta mostrano una forte sensibilità in quanto i valori sono postati verso l'alto. Il modello non è perfetto ma molto buono. L'area (AUC) sotto la curva misura 0.82, che significa che è molto più vicina ad 1 che 0, cioè ha una buona capacità di separare le classi.

In conclusione il modello è migliore di un classificatore casuale.

Belloni Laura