

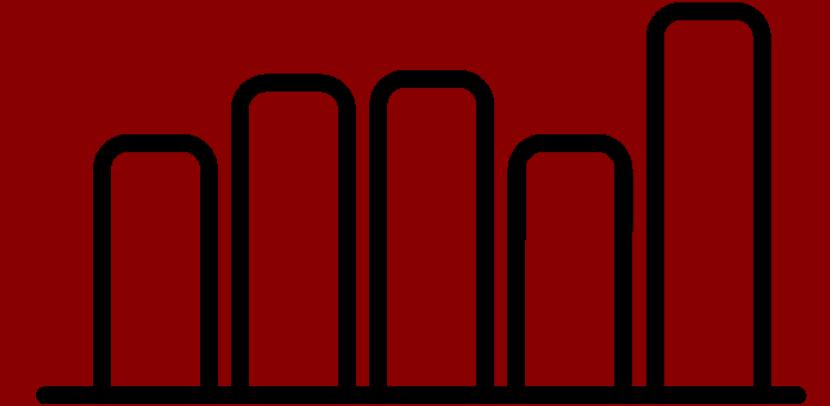
ANALISI ESPLORATIVA DEI DATI SU UN DATASET NETFLIX

Coding Girls

Presented by Gianluca Coco, Giulia Sinacori, Giulia Ciccarese



INDICE:



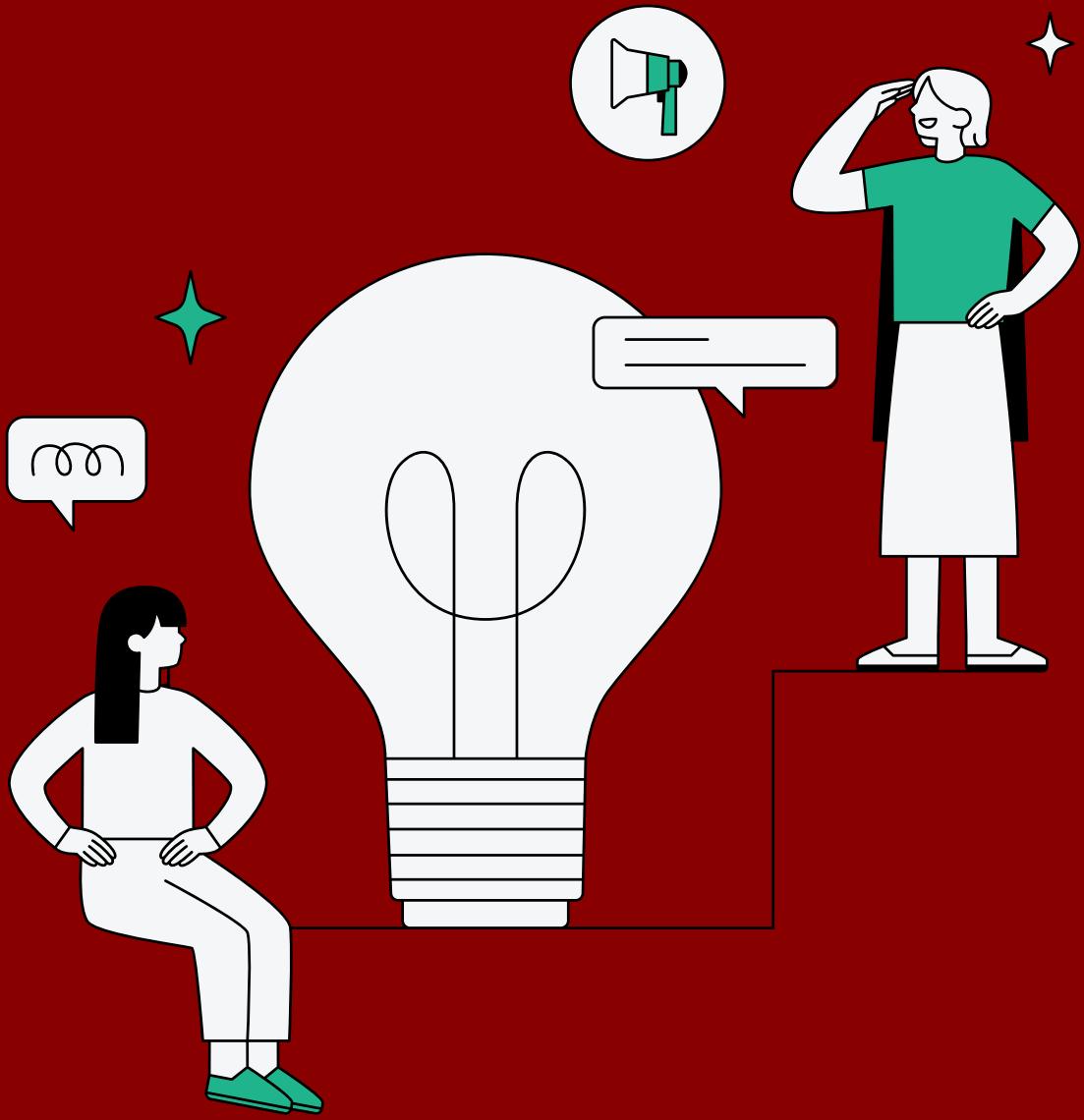
- **1 slide:** Com'è nato questo progetto?
- **2 slide:** La preparazione del dataset
- **3 slide:** Panoramica del dataset
- **4 slide:** Focus dell'analisi
- **5 slide:** Distinzione tra film e serie TV
- **6 slide:** Varietà dei generi dei film
- **7 slide:** Varietà dei generi delle serie TV
- **8 slide:** Distribuzione geografica delle produzioni
- **9 slide:** Distribuzione della durata dei film
- **10 slide:** Frequenza dei contenuti per anno
- **11 slide:** Frequenza dei Rating
- **12/13 slide:** correlazione tra anno di pubblicazione e durata dei titoli



COME È NATO QUESTO PROGETTO?

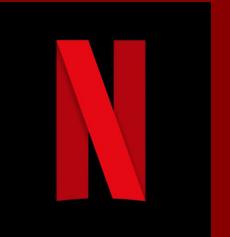
Questo progetto è stato realizzato durante il corso “ I Dati Siamo Noi” diviso in 3 incontri extrascolastici:

- 11/04 = introduzione alla Teoria Statistica
- 02/05 = Analisi dei Dati usando R
- 16/05 = Inizio del Lavoro di Gruppo



LA PREPARAZIONE DEL DATASET

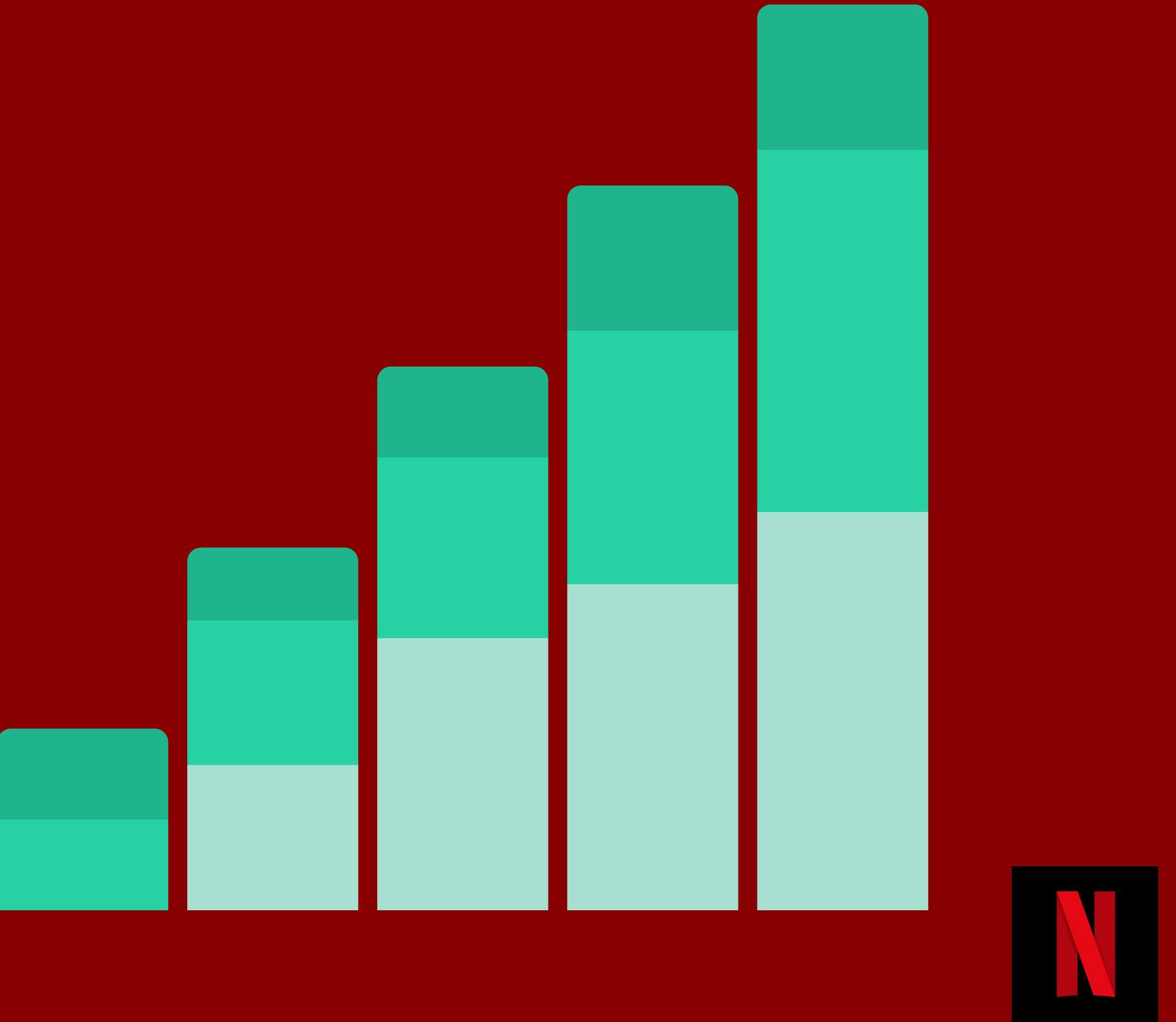
- La preparazione del Dataset è stata svolta prima utilizzando Python, usato per suddividere le colonne 'cast', 'country' e 'listed_in' da stringhe separate da virgole in liste e la successiva espansione di queste liste in righe separate, creando un dataset più dettagliato. In R, il dataset originale viene pulito rimuovendo colonne inutili e sostituendo stringhe vuote con valori NA. Le righe con valori mancanti vengono eliminate e vengono eseguite ulteriori trasformazioni, come la pulizia degli spazi bianchi, la conversione del formato delle date, e la creazione di colonne specifiche per distinguere tra film e Show TV. Questo processo assicura che i dati siano strutturati, puliti e pronti per ulteriori analisi.



PANORAMICA DEL DATASET

Il dataset di Netflix contiene 128.584 osservazioni e 14 variabili, tra cui:

- ID programma,
- titolo,
- regista,
- cast,
- paese di produzione,
- data di aggiunta,
- anno di rilascio,
- fascia d'età,
- durata,
- genere,
- sinossi,
- distinzione tra film e serie TV,
- numero di stagioni.





FOCUS DELL'ANALISI

Distinzione tra film e serie TV

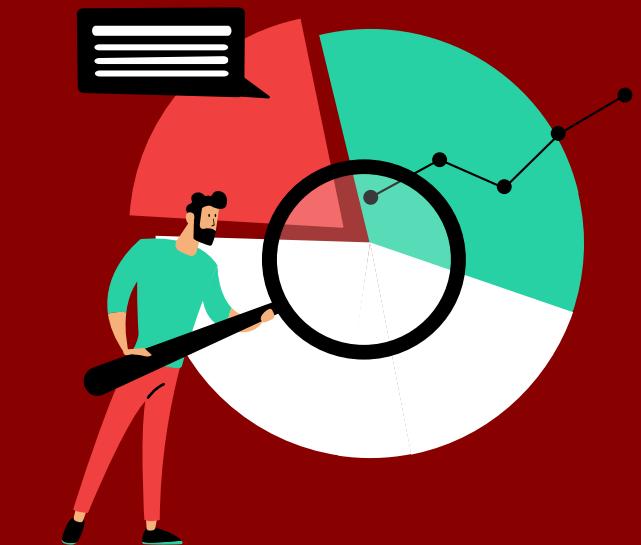
Analizziamo la prevalenza di film rispetto alle serie TV nel dataset.

Varietà dei Generi

Esploriamo i principali generi offerti, evidenziando le categorie più popolari.

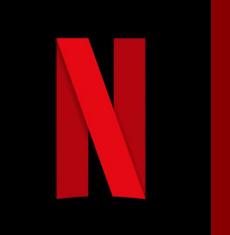
Distribuzione Geografica delle Produzioni

Esaminiamo i paesi di produzione, mostrando la diversità e l'influenza globale.

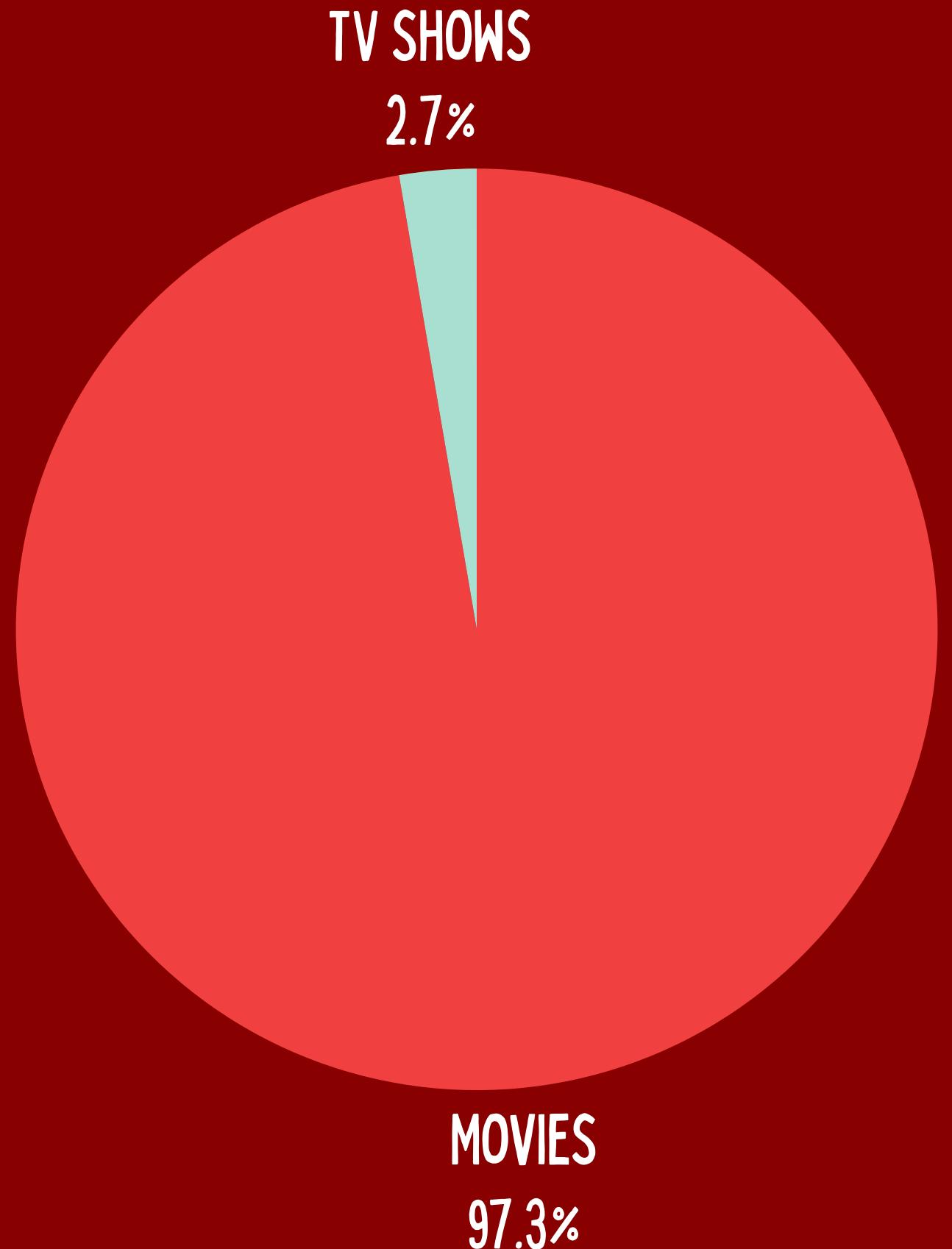


Distribuzione Della durata dei film

Analizziamo la durata dei vari titoli, mettendo in evidenza i contrasti tra le durate di questi.



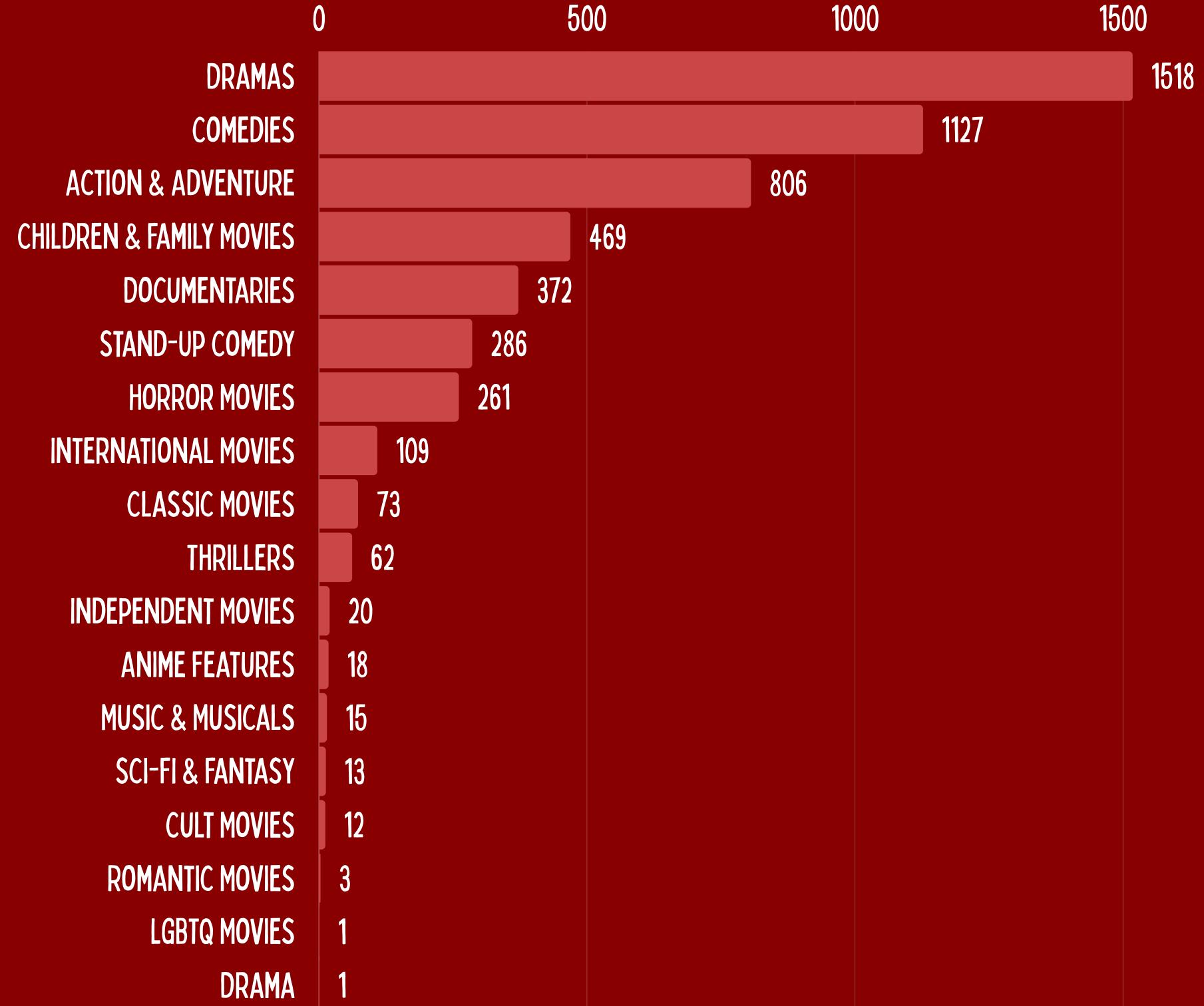
DISTINZIONE TRA FILM E SERIE TV



La predominanza dei film nel dataset Netflix è evidente, con un valore medio per la variabile 'movie' di circa 0.9678. Questo indica che circa il 96.78% dei record nel dataset sono film (1), mentre solo una piccola percentuale corrisponde a serie TV (0). La variabile 'movie' presenta un valore minimo di 0 e un valore massimo di 1, con il primo quartile, la mediana e il terzo quartile tutti pari a 1. Questi risultati confermano che la stragrande maggioranza dei titoli nel dataset sono film, evidenziando la predominanza di questo formato di intrattenimento su Netflix.

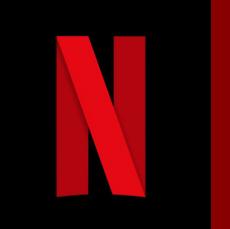


VARIETÀ DEI GENERI DI FILM



Emergono i "Drammi" come la categoria più diffusa, con un totale di 1518 titoli, seguiti dalle "Commedie" con 1127 titoli. "Azione & Avventura" si posiziona al terzo posto con 806 titoli, mentre "Film per Bambini e Famiglie" e "Documentari" garantiscono rispettivamente 469 e 372 titoli.

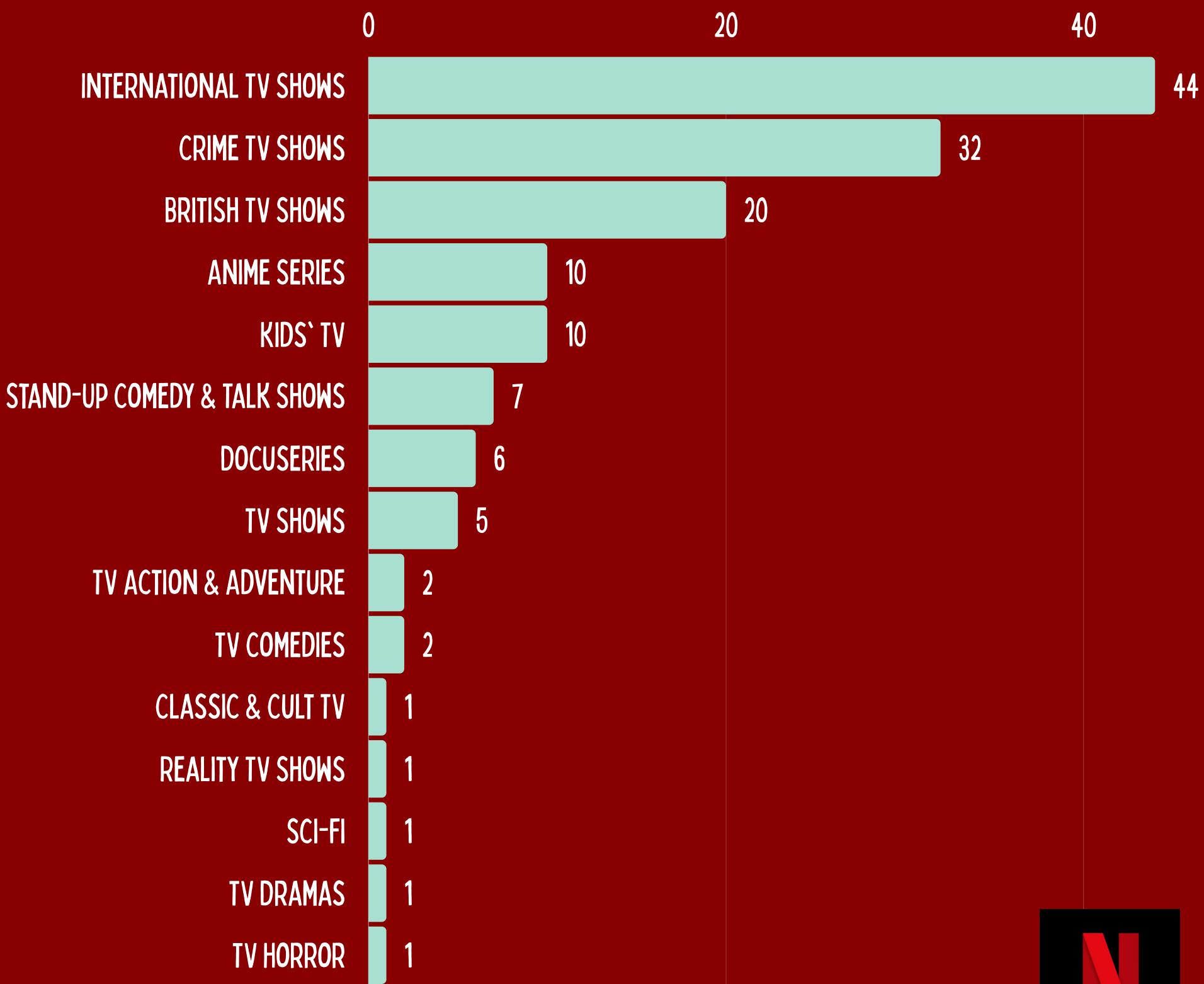
È interessante notare che "Drammi" e "Film" compaiono come voci autonome, forse indicando una classificazione più ampia. Alcuni generi come "Film Romantici" e "Film LGBTQ" sono relativamente rari, con un solo titolo ciascuno. Questa diversità di generi offre agli spettatori una vasta gamma di scelte, soddisfacendo diverse preferenze e gusti.



VARIETÀ DEI GENERI DELLE SERIE TV

I dati presentano una panoramica dei generi di programmi TV più popolari nel dataset di Netflix. La categoria di "Programmi TV Internazionali" emerge come la più numerosa, con 44 titoli, seguita da "Programmi TV Crimine" con 32 titoli. Anche "Programmi TV Britannici", "Serie Anime" e "Programmi TV per Bambini" si collocano tra i primi posti della classifica con 20, 10 e 10 titoli rispettivamente.

Tuttavia, alcuni generi come "TV Drammi", "TV Horror" e "TV Romantici" sono meno rappresentati, con solo un titolo ciascuno. Questa varietà di generi televisivi offre agli spettatori una vasta gamma di opzioni, dalle serie internazionali ai programmi per bambini, garantendo un'ampia varietà di intrattenimento adatta a diversi gusti e interessi.



DISTRIBUZIONE GEOGRAFICA DELLE PRODUZIONI



Impatto Culturale: La popolarità dei film provenienti da specifici paesi può riflettere il loro impatto culturale su scala globale. L'influenza di Bollywood in India o di Hollywood negli Stati Uniti potrebbe essere evidente nella classifica.

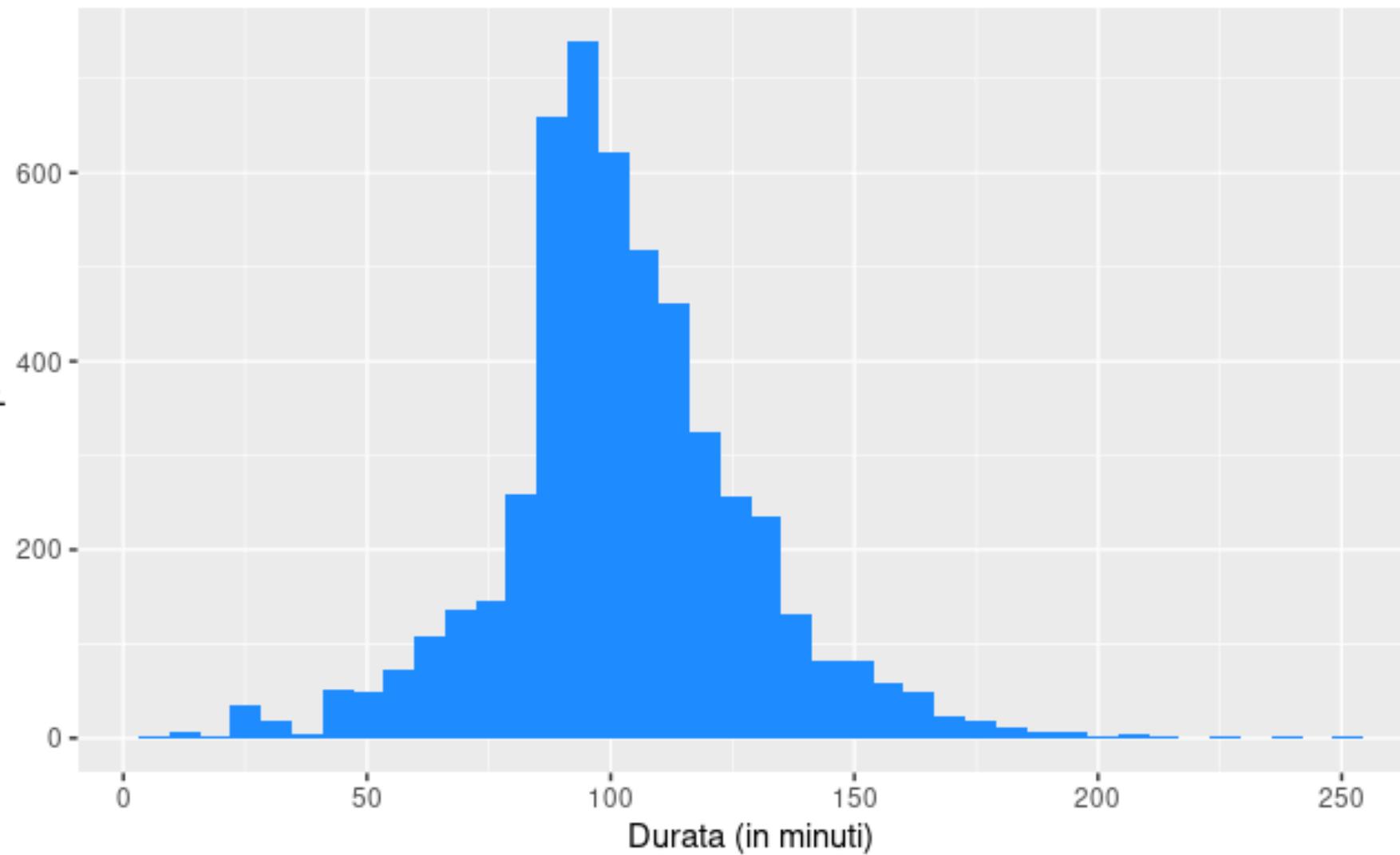
Popolarità Globale: La presenza di alcuni paesi ai primi posti suggerisce un grande interesse globale per i film provenienti da queste regioni. Ad esempio, se Stati Uniti e India dominano la lista, ciò indica la popolarità internazionale delle loro rispettive industrie cinematografiche.

Diversità dei Contenuti: Una rappresentazione diversificata dei paesi nella Top 20 indica che Netflix offre una vasta gamma di contenuti provenienti da diverse parti del mondo. La presenza di paesi di vari continenti come Asia, Europa e America Latina dimostra l'impegno di Netflix nel fornire una selezione ricca e variegata.

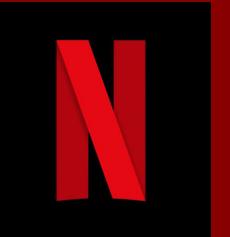


DISTRIBUZIONE DELLA DURATA DEI FILM

Distribuzione della Durata dei Film

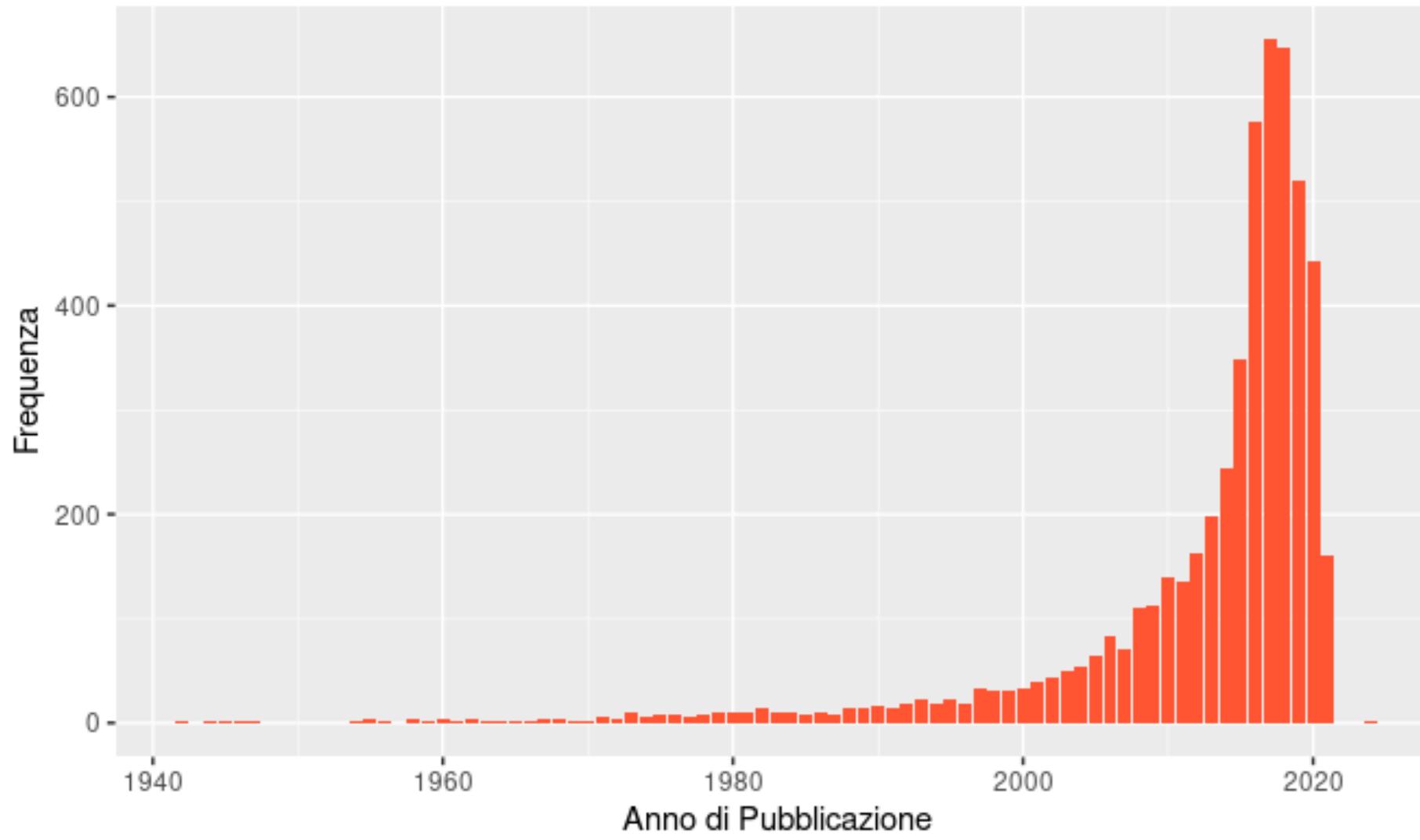


La durata dei film nel Dataset presenta una distribuzione abbastanza uniforme, con la maggior parte dei film che si aggirano intorno ai 102.70 minuti in media. La mediana della durata è di 101 minuti, indicando che la metà dei film ha una durata inferiore a questo valore. La moda, ossia il valore più frequente, è di 94 minuti. L'intervallo della durata dei film varia da un minimo di 8 a un massimo di 253 minuti, suggerendo una vasta gamma di lunghezze. L'intervallo interquartile è di 28 minuti, indicando che il 50% centrale dei film ha una durata compresa tra 86 e 114 minuti. La deviazione standard è di circa 25.96 minuti, suggerendo una moderata dispersione dei dati intorno alla media. Infine, la varianza è di circa 673.95, indicando la quantità di dispersione dei dati rispetto alla media, elevata al quadrato.



FREQUENZA DEI CONTENUTI PER ANNO

Frequenza dei Contenuti per Anno

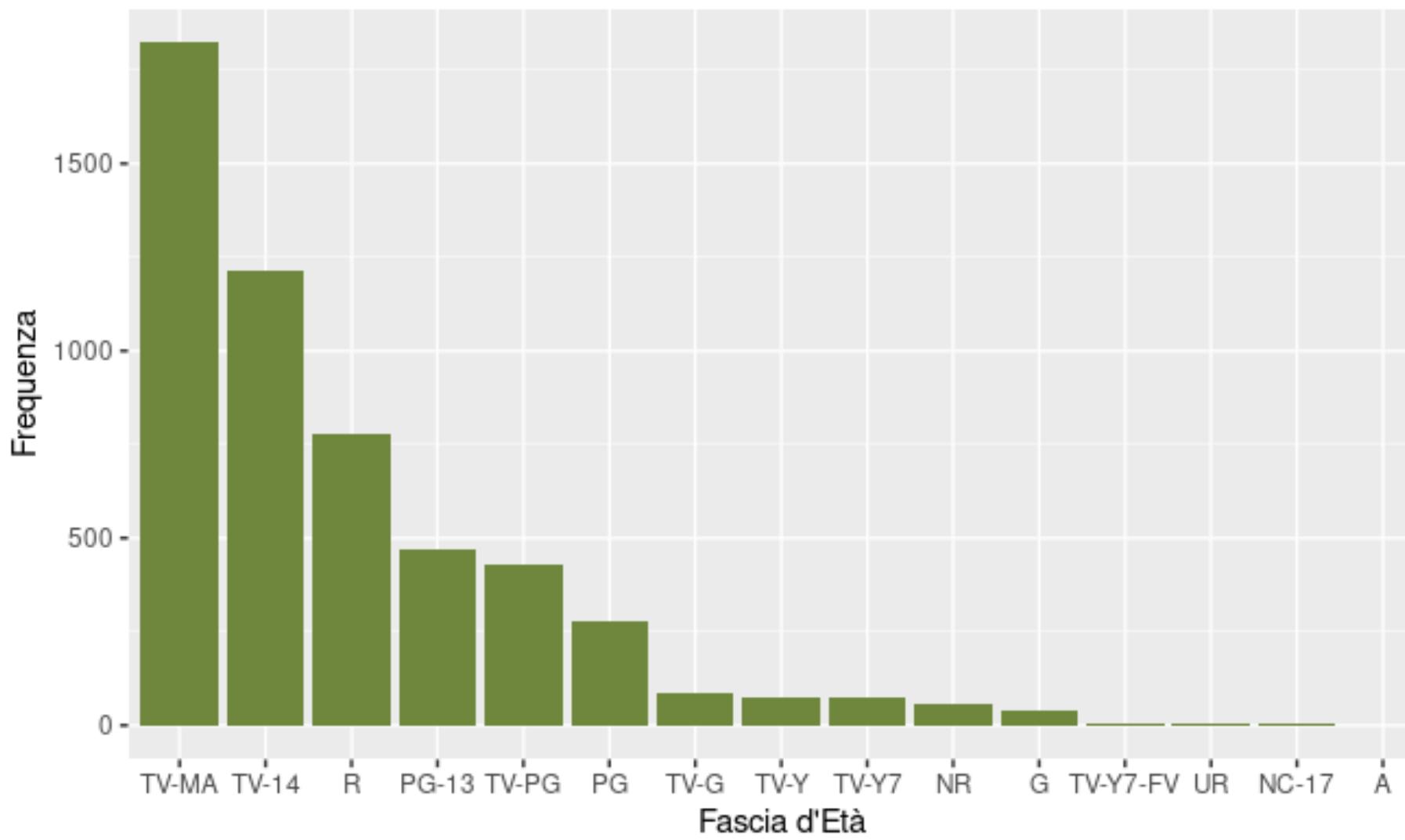


La maggior parte dei contenuti nel dataset sono stati pubblicati negli anni recenti, con un picco significativo a partire dal 2000. Il numero di pubblicazioni per anno aumenta costantemente fino a raggiungere il massimo nel periodo tra il 2015 e il 2020, dove la frequenza supera i 600 titoli all'anno. Dopo il 2020, si nota una leggera diminuzione nel numero di nuovi contenuti. Questa tendenza suggerisce un'espansione rapida e recente della libreria di contenuti di Netflix, con un focus crescente su produzioni recenti.



FREQUENZA DEI RATING

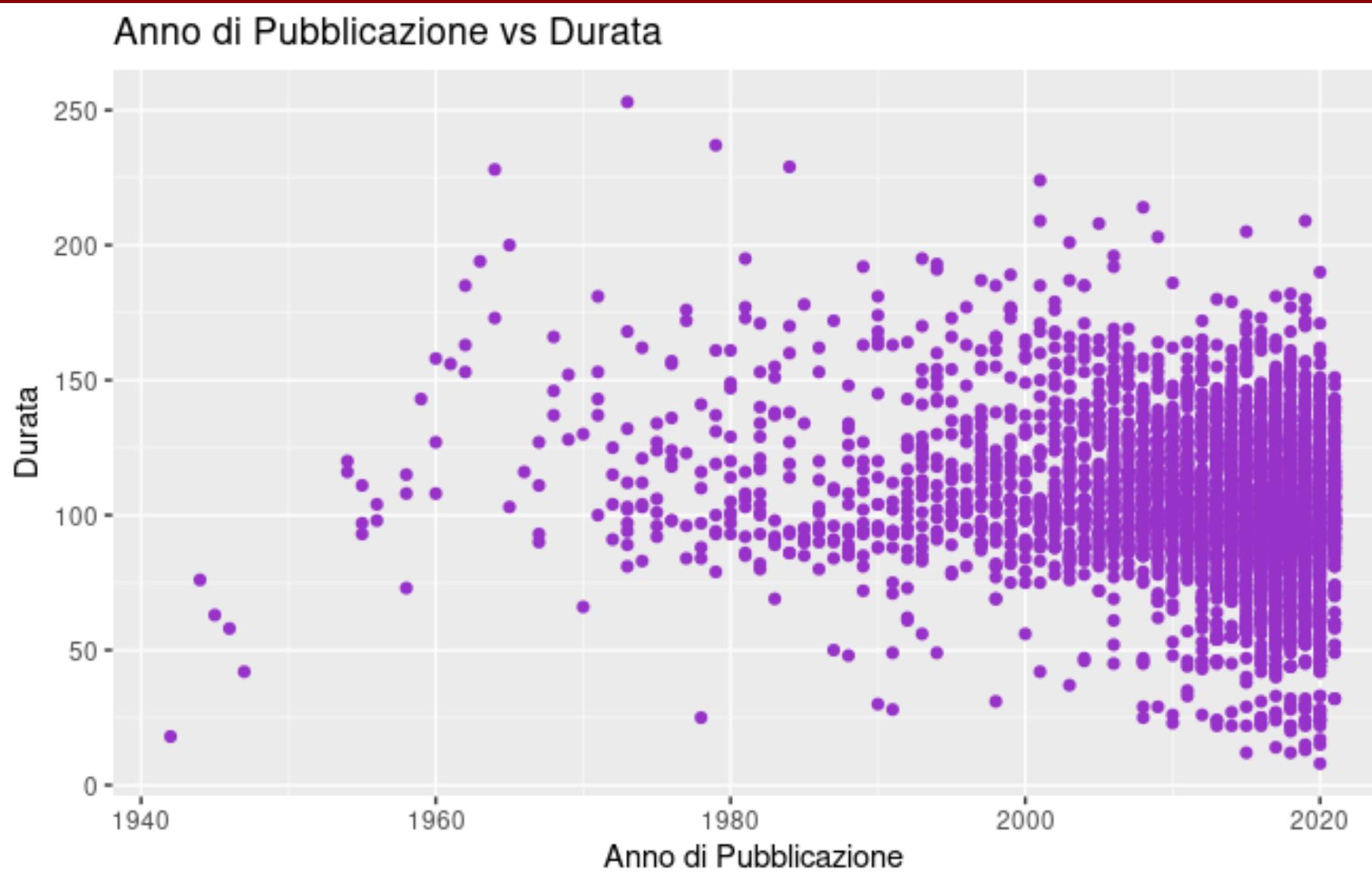
Frequenza dei Contenuti per Fascia d'Età



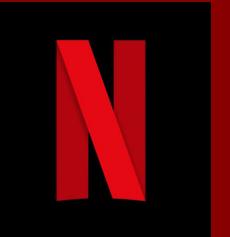
La maggior parte dei contenuti nel dataset è classificata come TV-MA, con oltre 1500 titoli, seguita da TV-14 con più di 1000 titoli. Le categorie R e PG-13 presentano un numero significativo di contenuti, rispettivamente con circa 700 e 500 titoli. Le fasce d'età TV-PG e PG contano meno di 500 titoli ciascuna. Le categorie meno rappresentate includono TV-G, TV-Y, TV-Y7, NR, G, TV-Y7-FV, UR, NC-17, e la categoria non classificata, ciascuna con meno di 100 titoli. Questi dati indicano che la maggior parte dei contenuti su Netflix è destinata a un pubblico adulto, mentre i contenuti per bambini e famiglie sono meno rappresentati.

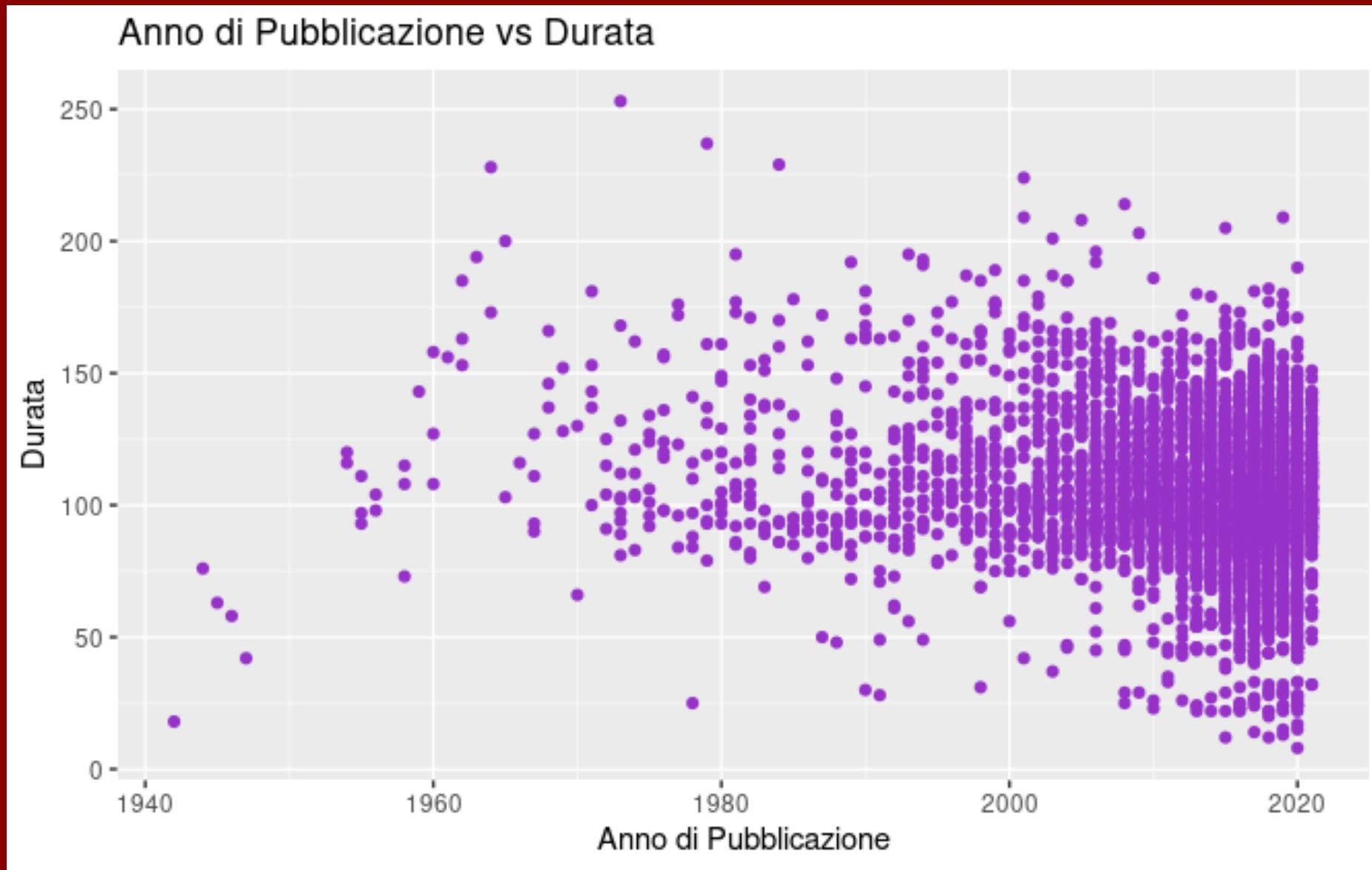


CORRELAZIONE TRA ANNO DI PUBBLICAZIONE E DURATA DEI TITOLI



La maggior parte dei film nel dataset è stata pubblicata dopo l'anno 2000, con una durata che varia principalmente tra 50 e 150 minuti. Le durate dei film più recenti tendono a concentrarsi intorno ai 100 minuti. Prima degli anni '80, c'è una maggiore variazione nelle durate, con alcuni film molto brevi e altri particolarmente lunghi. Dal 2000 in poi, la distribuzione della durata dei film diventa più uniforme, evidenziando una standardizzazione nel tempo di visione. Questo suggerisce che, con il passare degli anni, i film hanno raggiunto una durata più omogenea, mentre in passato c'era una maggiore variabilità.





Inoltre è importante menzionare che una correlazione di **-0,21** tra la durata in minuti e l'anno di pubblicazione indica una debole relazione negativa, suggerendo che le pubblicazioni più recenti tendono leggermente ad avere una durata inferiore



GRAZIE PER LA VISIONE!!

