Netflix and chill :

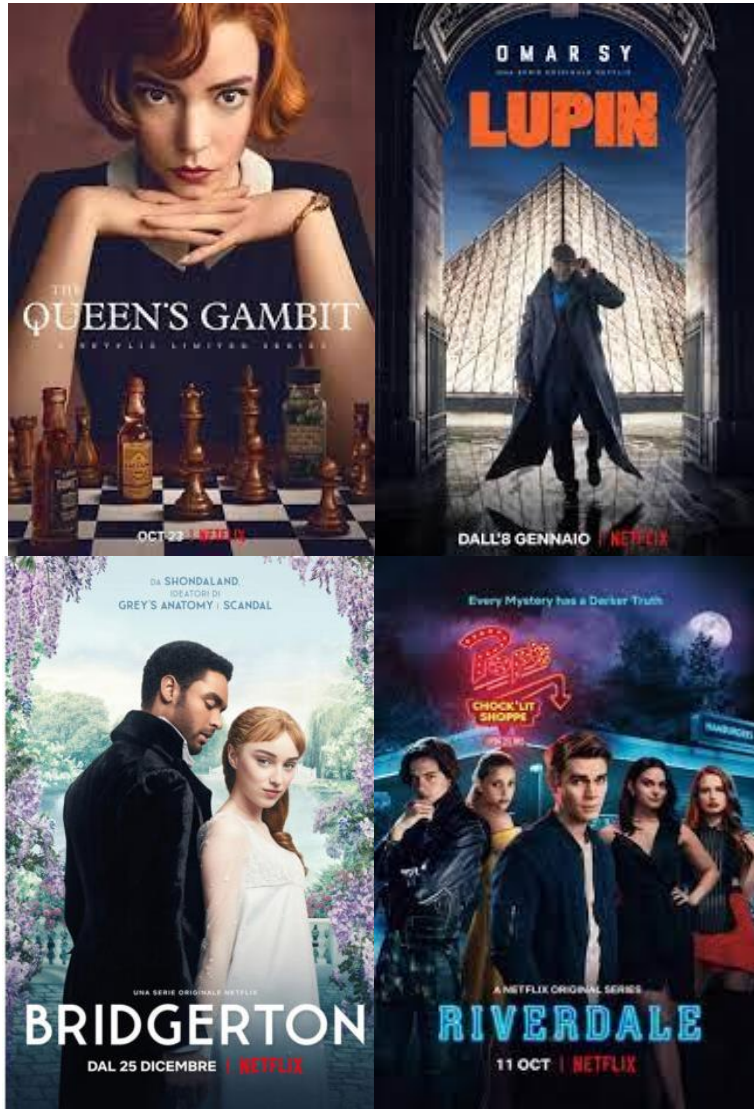
Studio degli show che arrivano alla top ten di Netflix

De Feudis Giovanni
Di Martino Camilla
Ras Iman

Obiettivi

- Cosa guardano gli italiani su NETFLIX?
- Quali caratteristiche hanno gli show che finiscono in TOP TEN?

Il nostro obiettivo è quello di misurare la popolarità dei film e delle serie-tv analizzando la loro permanenza e la loro posizione nelle due Top Ten di Netflix: la Top Ten generale e la Top Ten delle due diverse categorie (serie-tv e Film); considerando il periodo tra l'1 Dicembre 2020 e il 31 marzo 2021.



Introduzione

La nostra analisi ha considerato solamente le classifiche di Netflix Italia;
Abbiamo raccolto i titoli degli show presenti nelle Top Ten nel periodo tra l'1 Dicembre 2020 e il 31 Marzo 2021 e ne abbiamo studiato :

- La popolarità : misurata considerando la permanenza dei titoli nelle Top Ten;
- La qualità : misurata tramite i voti (rating) che ciascun titolo ha ottenuto sul sito Internet Movie Database (IMDb); tale sito raccoglie i voti lasciati dagli utenti.

La raccolta dati è stata resa possibile grazie all'utilizzo dell' API del sito FlixPatrol che, riporta ogni giorno tutte le classifiche di Netflix; inoltre la raccolta delle caratteristiche dei singoli show è stata fatta facendo scraping su Wikipedia; mentre per quanto riguarda i dati di IMDb ci sono stati resi disponibili sotto forma di dataset tabellari dai gestori del sito.

Introduzione

Delle tre V dei Big Data il nostro progetto ha affrontato la:

- **Velocità**: le tre Top Ten di Netflix si aggiornano quotidianamente, perciò si qualificano come dati "veloci", precisamente con un periodo di aggiornamento di un giorno;

- **Varietà**: i dati "veloci" costituiti dalle classifiche di ogni giorno su Netflix sono stati integrati con due fonti di dati: il database IMDb, ; e le informazioni prese da Wikipedia tramite scraping, che descrivono in modo esaustivo ogni film o serie tv del database.

Indice

Calcolo dell'indice di popolarità

L'indice di popolarità di *f* , un film o una serie-tv, è maggiore più giorni un film è stato in classifica e più alto il *ranking* che aveva.

$$S(C,G,f) = \sum_{g\ in\ G} (10 - rank(C, g, f))$$

dove *rank(C,g,f)* dipende:

- dalla classifica C
- dal giorno *g*
- dal film *f*

$$rank_{C,g}(f) = \begin{cases} 0 & \text{se } f \text{ è primo in classifica} \\ 1 & \text{se } f \text{ è il secondo} \\ \dots & \\ 9 & \text{se } f \text{ è l'ultimo, cioè il decimo} \\ 10 & \text{se } f \text{ non compare in classifica} \end{cases}$$

L'indice di popolarità per la classifica *C* è definito così:

$$I_C(f) = \frac{S(C,g,f)}{|G| * 10}$$

Calcolo dell’indice di popolarità

Ogni titolo *f* (film o serie-tv) appartiene a due classifiche: quella **generale** e quella del proprio **tipo**.

Siamo pronti a definire **l’indice di popolarità del titolo *f*** come:

$$I(f) = I_{gen}(f) + I_{tipo}(f)$$

L’indice è compreso tra 0 e 2.

La classifica del tipo coincide con la classifica delle sole serie-tv se *f* è una serie-tv, e con la classifica dei film se è un film.

Calcolo dell’indice di popolarità

Per esempio: consideriamo la serie tv "Lupin", e consideriamo la classifica delle sole serie tv.

Ipotizziamo che Lupin sia stata prima in classifica il 1 gennaio e il 2 gennaio, seconda il 3 gennaio, quarta il 4 e infine che il 5 gennaio non fosse più presente in classifica.

Considerando il periodo di tempo dal 1 al 5 gennaio (= 5 giorni) abbiamo:

$$I_{serieTv}(Lupin) = \frac{(10-0)+(10-0)+(10-1)+(10-3)+(10-10)}{10*5} =$$

$$= \frac{10 + 10 + 9 + 7 + 0}{50} = 0.72$$

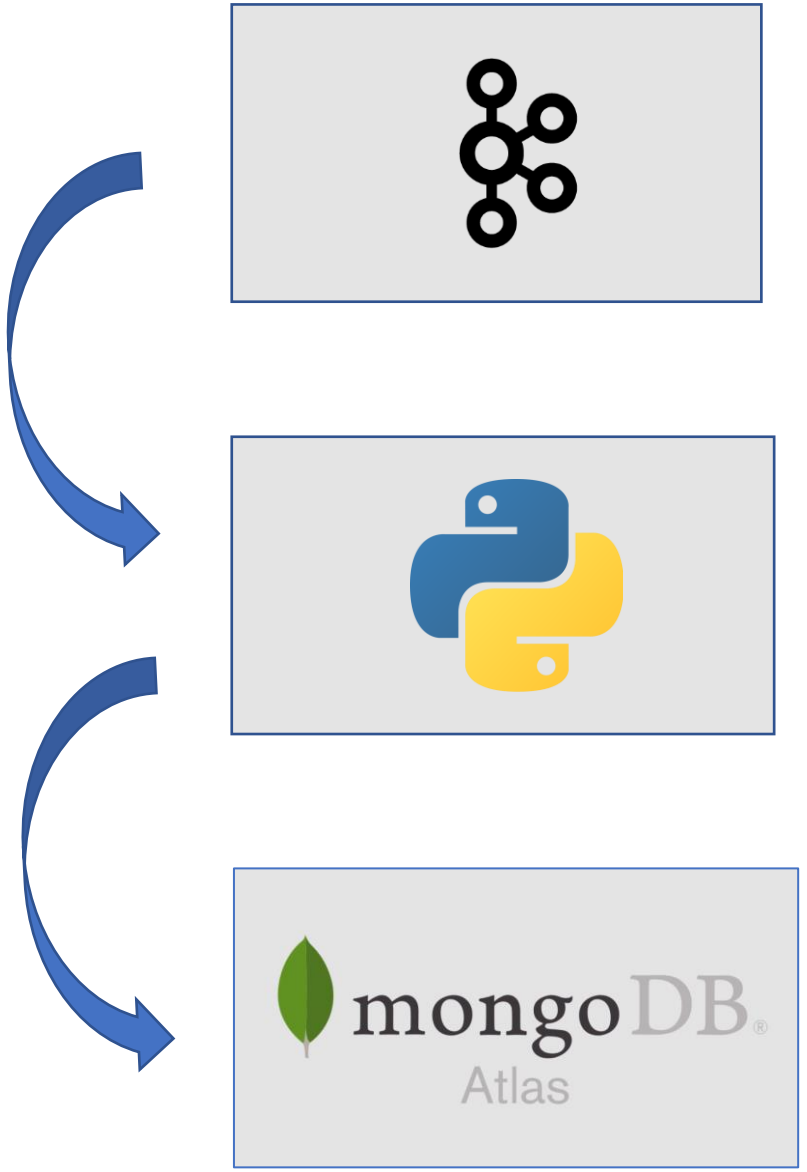
In questo caso abbiamo considerato un periodo di 5 giorni, nel progetto, come già accennato, i giorni totali sono 121 (dal 1 Dicembre 2020 al 31 Marzo 2021).

Indice

Velocità

- I nostri dati in formato JSON avevano una frequenza di aggiornamento di 24 ore, quindi abbiamo deciso di utilizzare la piattaforma Apache Kafka.

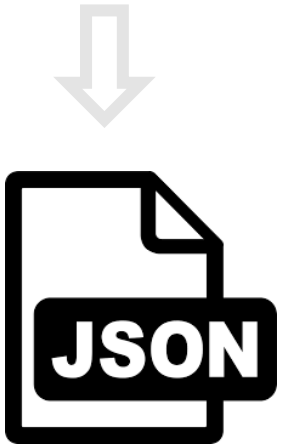
- Utilizzando la libreria Kafka-Python e Pymongo ci è stato permesso di caricare i nostri dati all'interno della coda specifica del nostro topic e successivamente prelevare questi dati e caricarli all'interno del cloud di MongoDB Atlas.



Indice

Varietà

- Il formato tabellare è stato il prescelto:
  - Facilità di integrazione con csv proveniente da IMDB
  - Stesse informazioni per ogni riga del dataframe



- I dati sono stati prelevati dal Cloud di Atlas in formato JSON utilizzando la libreria Pymongo e successivamente è stato realizzato un dataframe in csv dove per ogni giorno abbiamo dati provenienti dalle top ten che in totale arrivano a 3630 righe.

Varietà

- L’integrazione è stata svolta in tre fasi, ciascuna con un proprio script Python:
  - Integrazione con IMDb;
  - Integrazione con Wikipedia;
  - Integrazione finale, dove mettiamo insieme i dataset e calcoliamo l’indice di popolarità.

L’integrazione IMDb è una semplice *join* tra db relazionali.

L’integrazione Wikipedia consiste nello *scraping*, utilizzando le librerie **wikipediaapi** e **BeautifulSoup**, per ottenere informazioni aggiuntive su ogni film o serie-tv, come:

- Il paese di produzione
- La durata (in minuti o in episodi)
- Il sommario
- Tanti altri…

Indice

Pulizia del Dataset

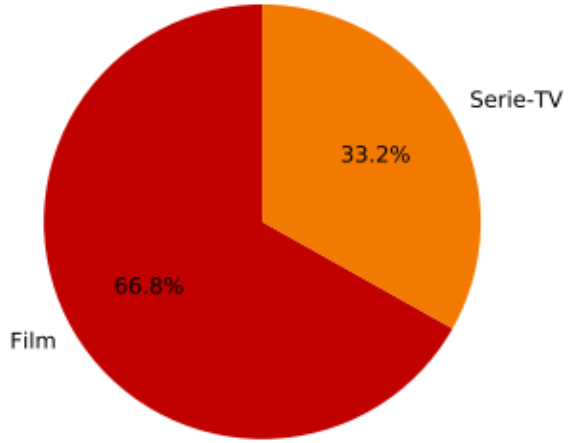
- La fase di pulizia è stata effettuata in parte manualmente, in parte in Python e alcune parti mediante Excel.

- Le colonne pulite sono state:
  1. Paese
  2. Produzione
  3. Distribuzione
  4. Durata
  5. Genere

Indice

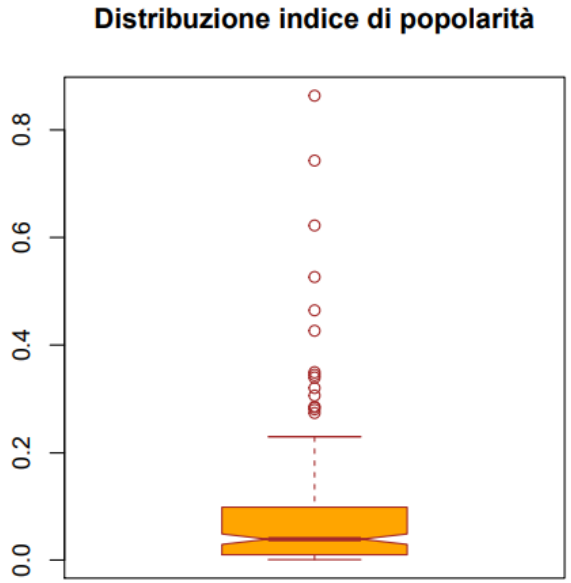
Analisi esplorativa del dataset finale

Dalla raccolta dei dati nel periodo da noi considerato abbiamo ottenuto complessivamente 208 titoli, di cui 139 Film e 69 Serie-tv



L'indice di popolarità è compreso tra 0 e 2 , ha in generale ottenuto un valore molto basso; la maggior parte degli show presenta un indice inferiore a 0,2.
Elementi riassuntivi dell'indice di popolarità:

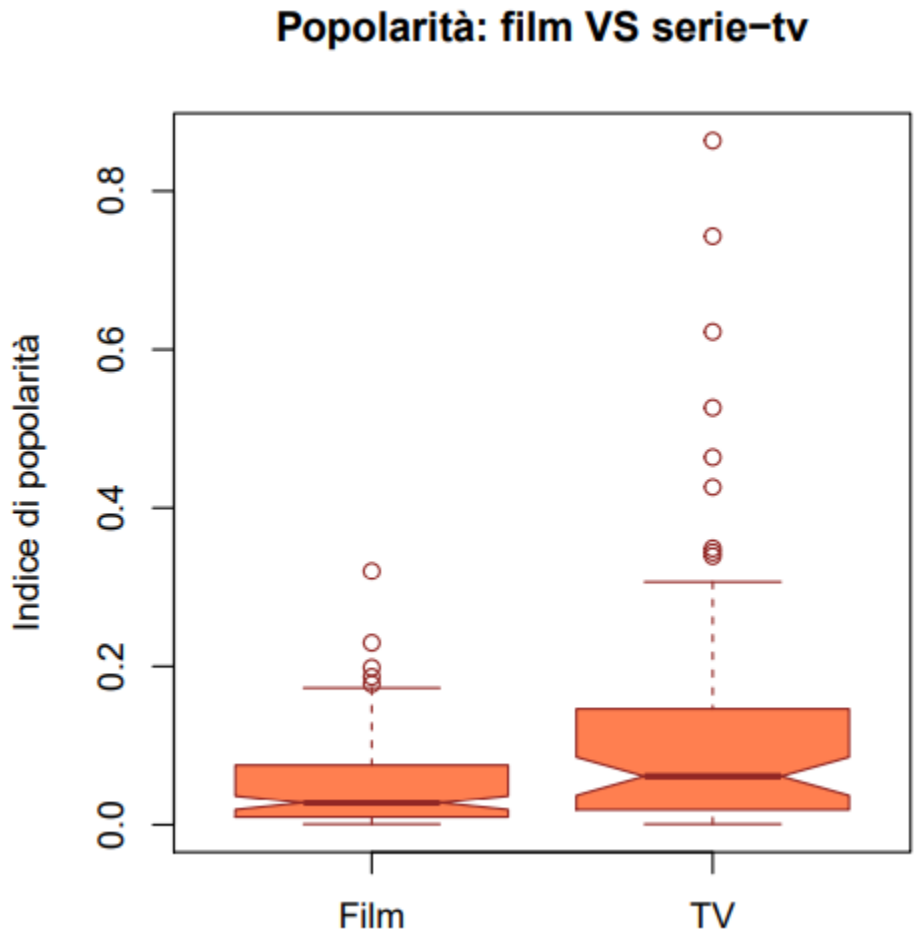
- **Valore minimo** = 0,000826
- **Primo quartile** = 0,0099
- **Mediana** = 0,0388
- **Terzo quartile** = 0,0975
- **Valore Massimo** = 0,8636



Analisi esplorativa del dataset finale

Dalla figura possiamo vedere come la distribuzione della popolarità delle serie-tv sia spostata verso valori più elevati rispetto ai film, anche senza considerare i numerosi outlier della categoria, che risulteranno essere proprio gli show più visti di Netflix. In particolare gli indici statistici sono:

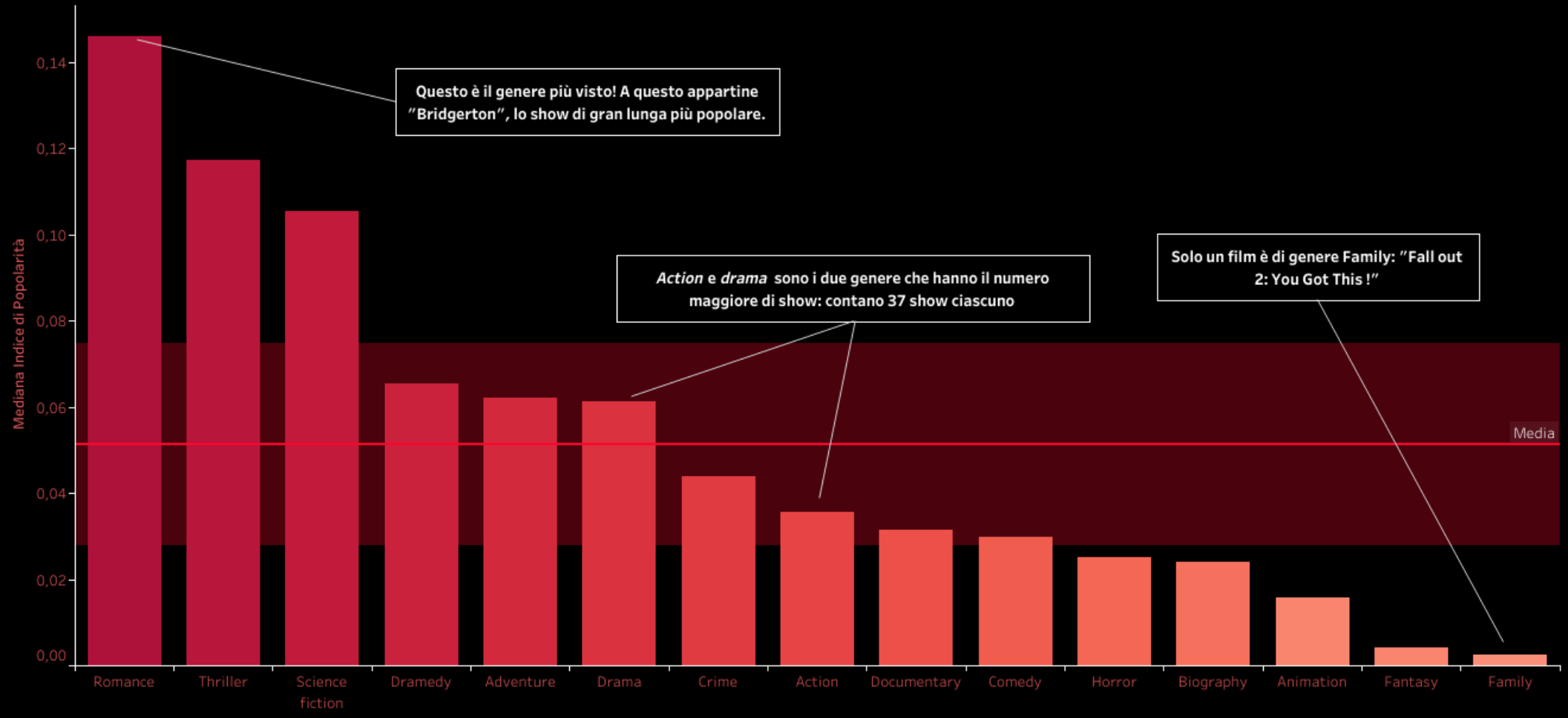
- Minimo: 0.00083 (film), 0.00083 (serie-tv);
- Primo quartile: 0.0099 (film), 0.019 (serie-tv);
- Mediana: 0.028 (film), 0.062 (serie-tv);
- Terzo quartile: 0.076 (film), 0.146 (serie-tv);
- Massimo: 0.321 (film), 0.864 (serie-tv);
- Media: 0.052 (film), 0.135 (serie-tv).



INDICE DI POPOLARITA' PER GENERE

Qual è il genere con popolarità più alta?

Sull'asse X sono riporati i generi, sull' asse Y è riportata la mediana dell'indice di popolarità.
Gli show (film e serie tv) sono state divise in 15 generi diversi. Per ciascuno di questi è calcolata la mediana dell'indice di popolarità dei suoi show.



Genere
- Action
- Adventure
- Animation
- Biography
- Comedy
- Crime
- Documentary
- Drama
- Dramedy
- Family
- Fantasy
- Horror
- Romance
- Science fiction
- Thriller

La linea orizzontale indica la media delle barre relative a tutti i generi, mentre la fascia orizzontale è l'intervallo di confidenza al 95%.


That's all Folks!