Università degli Studi di Milano-Bicocca
Corso di Laurea in Data Science

Data Visualization

Cominetti Fabrizio 882737
Agazzi Ruben 844736
Abete Davide 882299

Anno Accademico 2021/2022

Indice

Introduzione

In una realtà come quella odierna, in cui si producono sempre più contenuti digitali per un pubblico sempre più ampio che reagisce prestando sempre meno attenzione alle informazioni che riceve, e caratterizzata tra le altre cose dalla rapida diffusione delle serie TV in ambito cinematografico, è vero che le persone preferiscono anche film più brevi? Questa considerazione ci ha guidati nella realizzazione di questo progetto e ci ha permesso di determinare la seguente domanda di ricerca: i film più brevi ricevono generalmente voti migliori?

Il dataset

Per la realizzazione di questo progetto i dati utilizzati sono stati scaricati dal sito ufficiale di **IMDb** al seguente indirizzo: https://datasets.imdbws.com
I dataset di interesse sono **'title.basics'** e '**title.ratings'**.
Il dataset totale consiste di 1.206.957 record e di 11 attributi.

Le variabili presenti all'interno del dataset **title.basics** sono le seguenti:

- **tconst** *(string)* - identificativo alfanumerico
- **titleType** *(string)* – tipo/formato del titolo
- **primaryTitle** *(string)* – titolo usato in fase promozionale
- **originalTitle** *(string)* - titolo in lingua originale
- **isAdult** *(boolean)* - 0: non-adulti; 1: adulti
- **startYear** *(YYYY)* – anno di uscita
- **endYear** *(YYYY)* – anno di fine (per serie tv)
- **runtimeMinutes** – durata (in minuti)
- **genres** *(string array)* – genere

Le variabili presenti all'interno del dataset **title.ratings** sono le seguenti:

- **tconst** *(string)* - identificativo alfanumerico
- **averageRating** – rating medio
- **numVotes** - numero di voti ricevuti

Integrazione

I due dataset sono stati integrati tramite la libreria pandas, effettuando una merge dei due dataset sulla base dei valori della colonna '**tconst**', ovvero l'identificativo univoco di ogni elemento nel catalogo.

Progettazione

La realizzazione della visualizzazione è avvenuta effettuando l'esplorazione dei dati in concomitanza alla pulizia degli stessi.
Le due fasi sono state realizzate tramite l'utilizzo delle librerie *Pandas, Numpy, Pywaffle* e *Plotly*.

Conversione

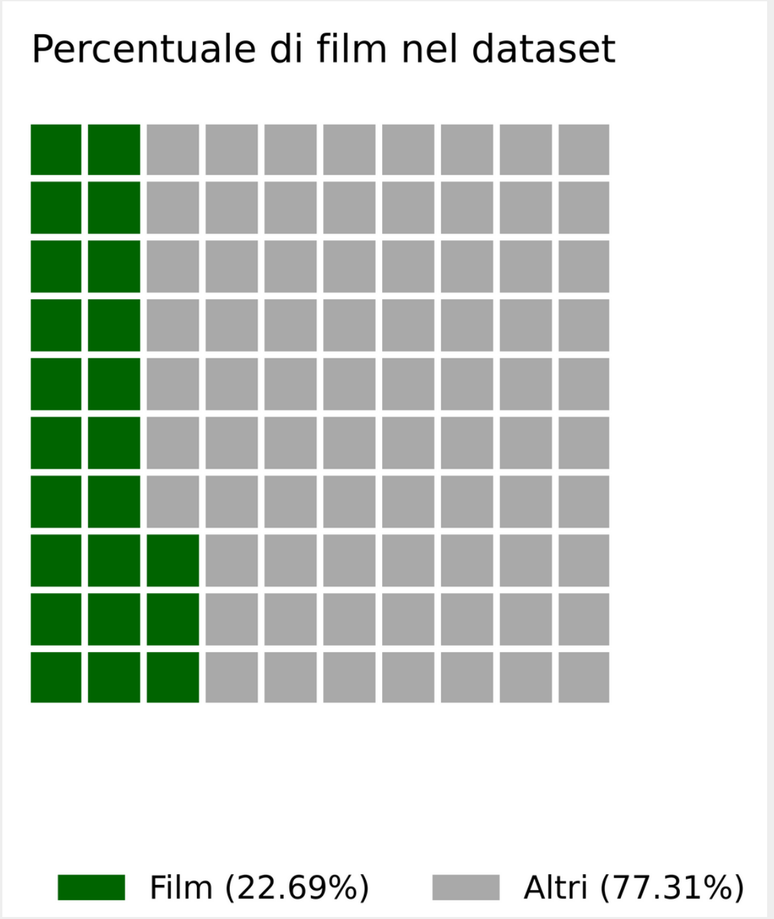
Tramite le librerie Pandas e Numpy sono state convertite in variabili numeriche le colonne seguenti:

- *runtimeMinutes*
- *startYear*

La colonna *titleType* è stata invece convertita in variabili di tipo categorico.

Eplorazione dati

Inoltre sono state realizzate due visualizzazioni semplici: nella visualizzazione a fianco, realizzata con la librera *pywaffle*, possiamo osservare la **percentuale** di film all'interno del dataset rispetto al totale.



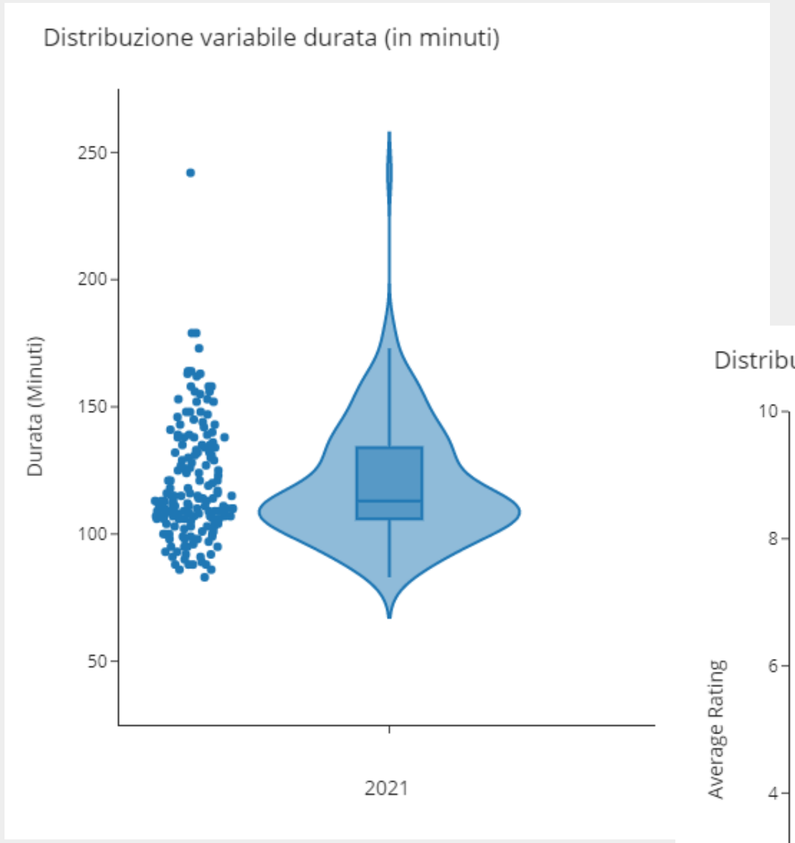
Pulizia dati

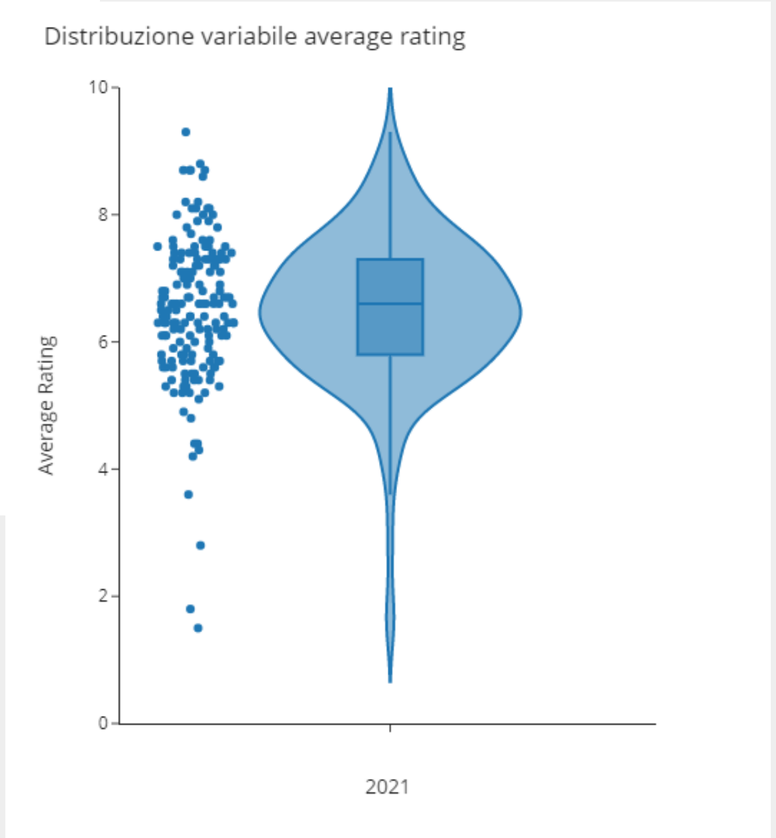
Durante la fase di pulizia dati sono state svolte diverse operazioni nel seguente ordine:

1. Eliminazione colonne superflue dal dataset
2. Rimozione righe contenenti valori mancanti
3. Eliminazione dati che non si riferiscono a film
4. Eliminazione film che non sono usciti nel 2021
5. Eliminazione film con meno di 15.000 voti

Eplorazione dati

Di seguito osserviamo le due distribuzioni delle variabili runtimeMinutes e averageRating tramite **violin plot**, realizzate utilizzando la libreria *plotly*.



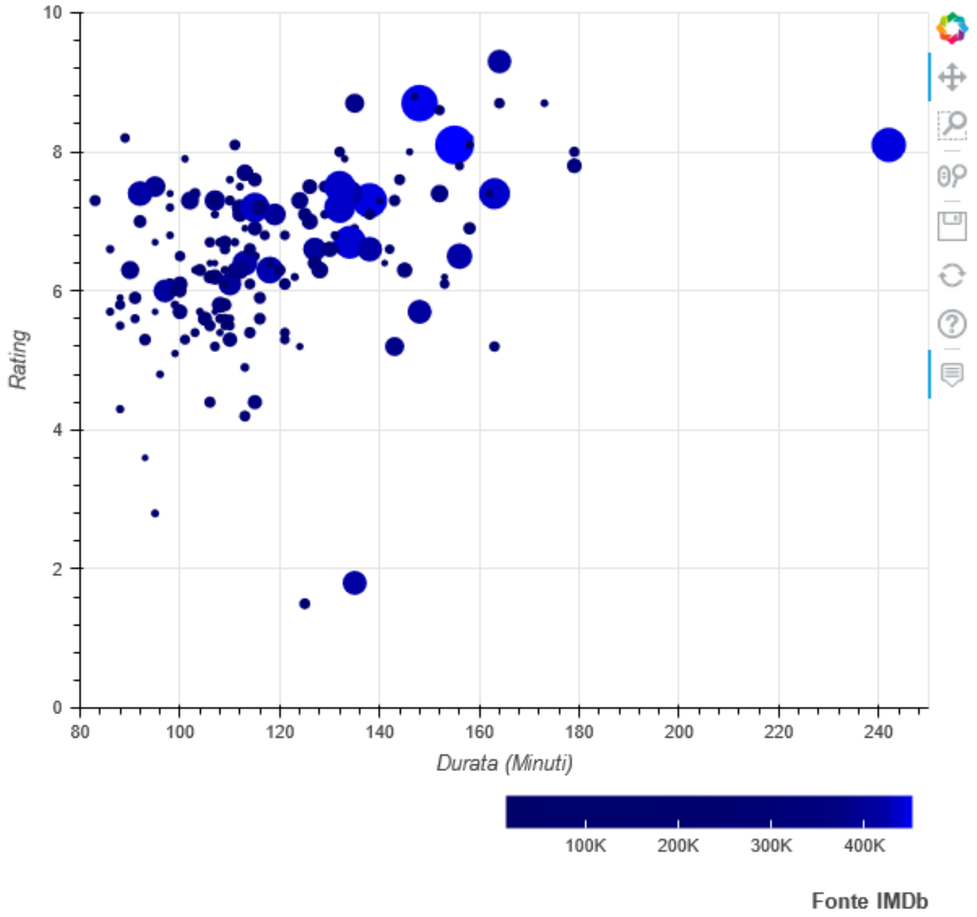


Realizzazione

Concluse le operazioni di esplorazione e pulizia dati abbiamo realizzato una visualizzazione interattiva tramite l'utilizzo della libreria **Bokeh**.
La visualizzazione vuole mostrare la **correlazione** tra la durata e il rating medio dei film.
Per mostrare ciò abbiamo scelto di realizzare uno **scatter plot**, una visualizzazione che permette di far emergere eventuali correlazioni tra le variabili di interesse.
Lo scatter plot è stato poi arricchito di *ulteriori elementi* con il fine di rendere più informativa e chiara la visualizzazione.
Nelle slide successive ripercorriamo le fasi di realizzazione.

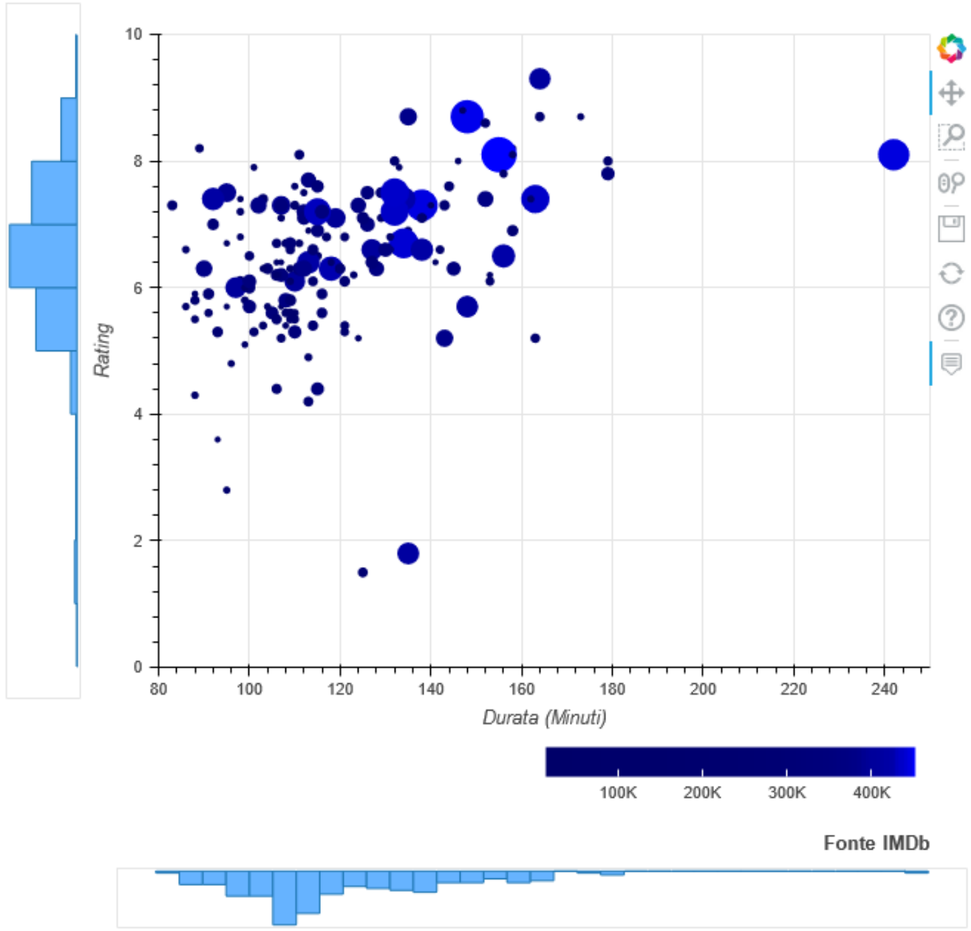
Realizzazione

Scatter plot che mette in *relazione* durata in minuti (x) e media dei voti (y), con area e colore dei punti proporzionati al numero dei voti ottenuti



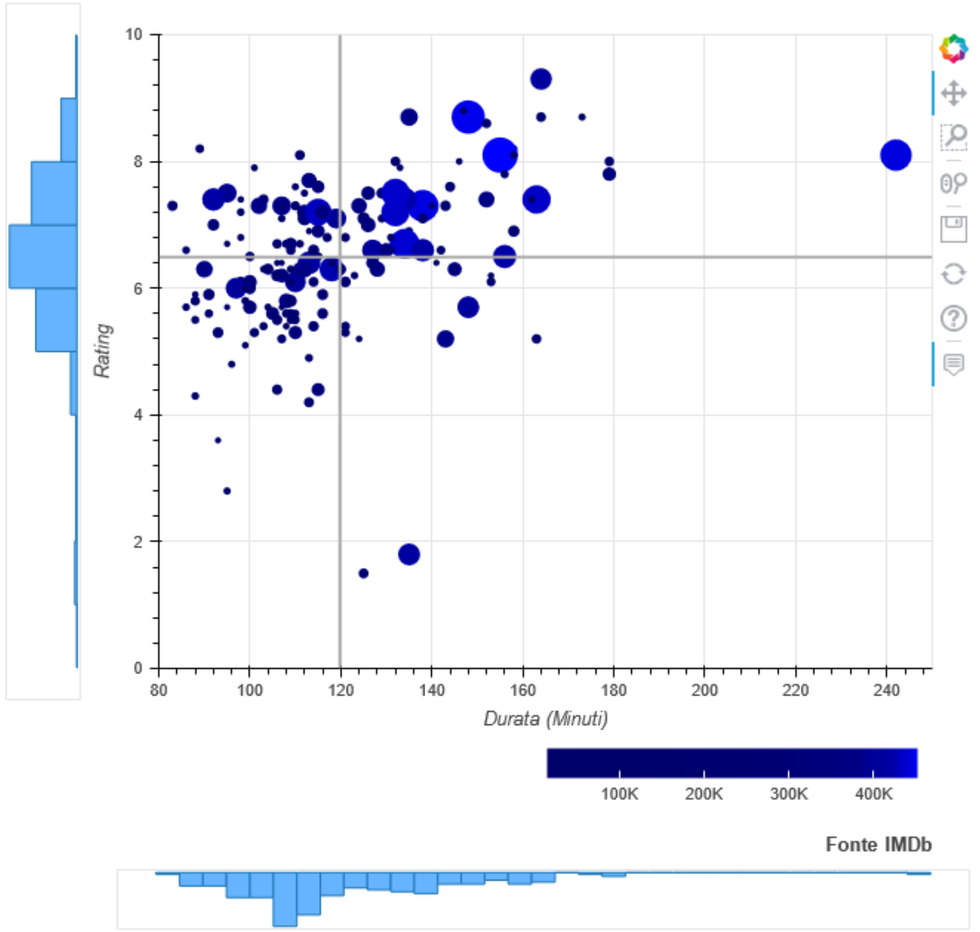
Realizzazione

Realizzazione di istogrammi laterali, per mostrare la *distribuzione* dei dati sui due assi



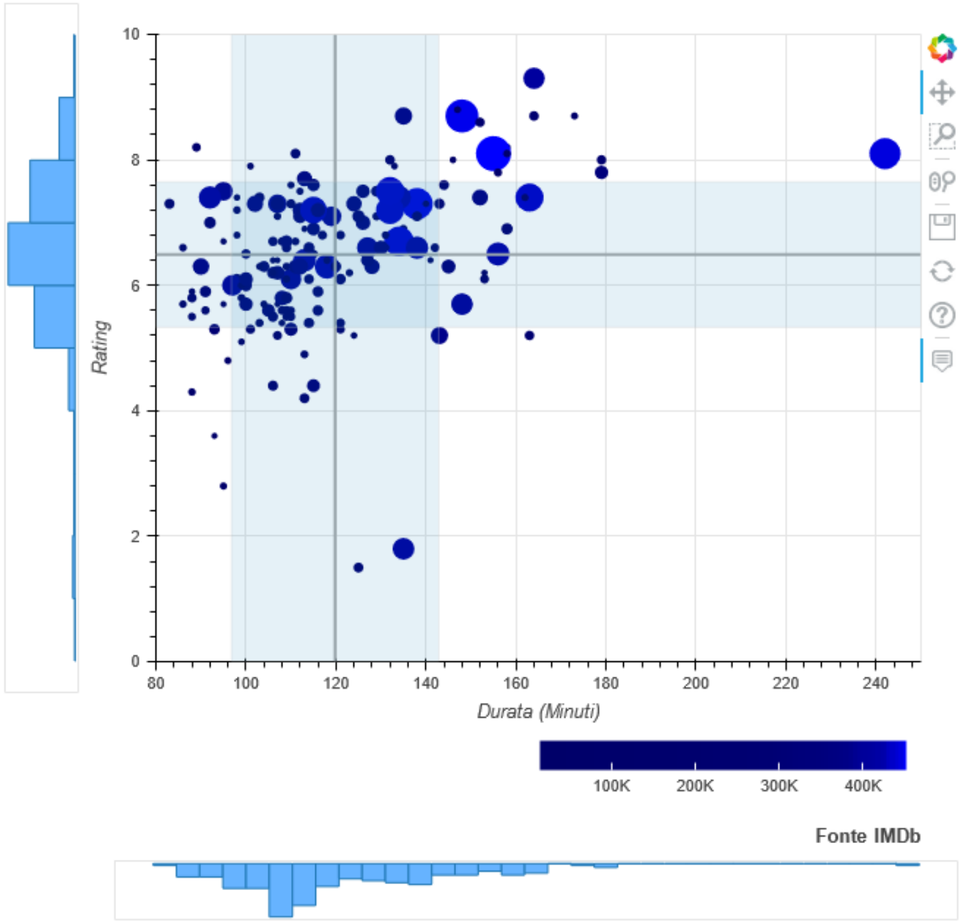
Realizzazione

Inserimento delle rette indicanti la *media* dei valori degli assi



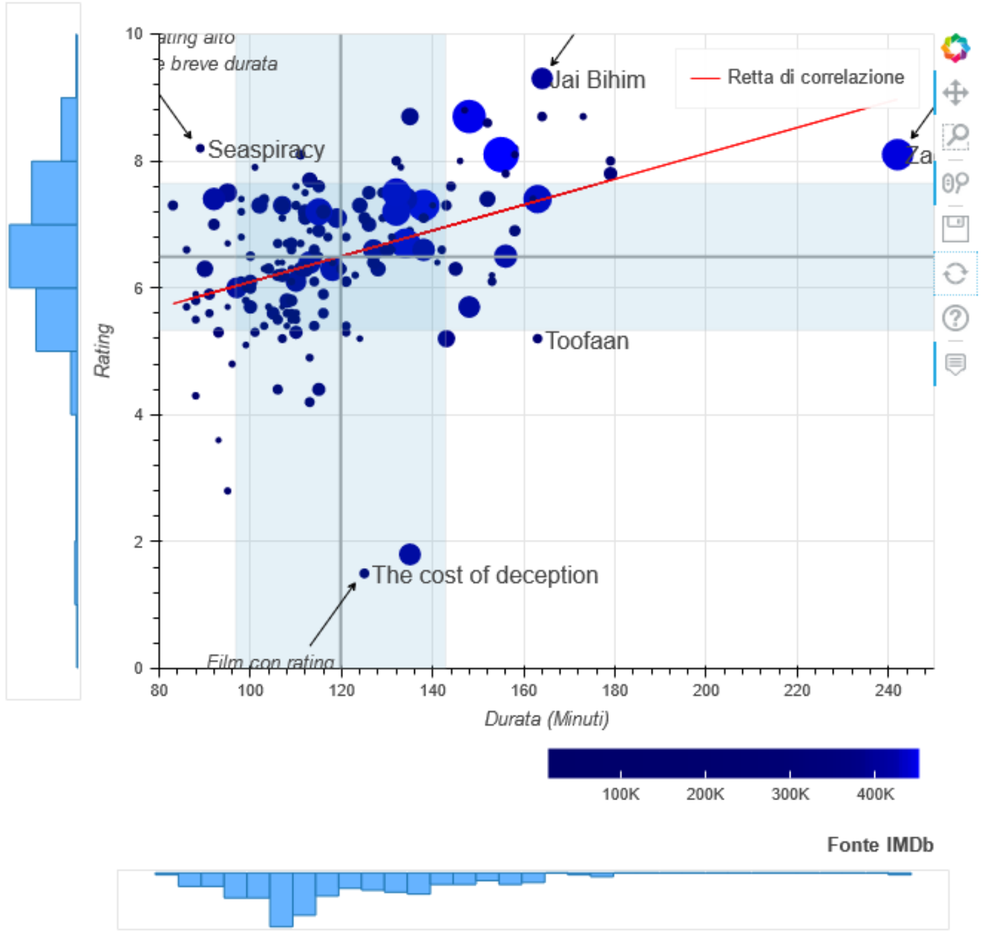
Realizzazione

Inserimento fasce indicanti la *deviazione standard* dei dati sue due assi

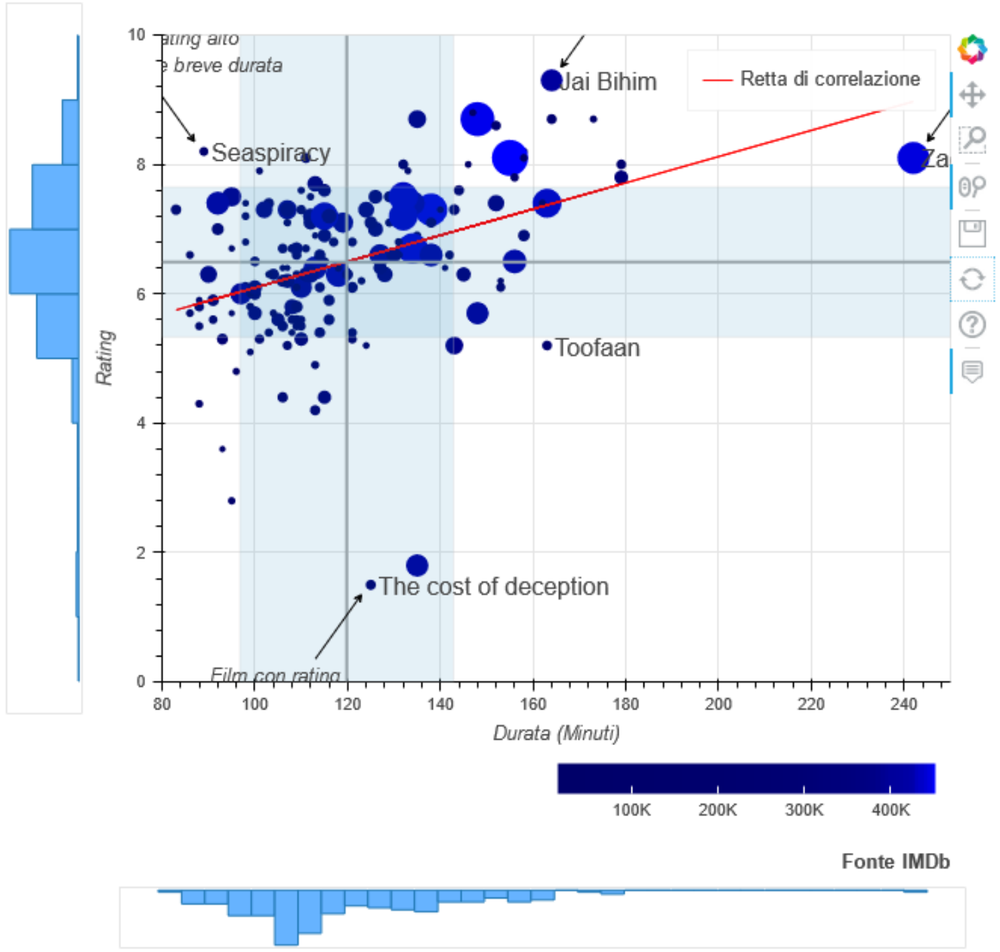


Realizzazione

Aggiunta di etichette e frecce indicanti dati *rilevanti* e aggiunta della retta di *correlazione*



Visualizzazione Finale



Link alla visualizzazione interattiva

https://blowdire.github.io/DataVisualizationProject/

Valutazione qualità

La valutazione della qualità è stata svolta in 3 fasi:

1. Valutazione euristica
2. Valutazione psicometrica
3. Test utente

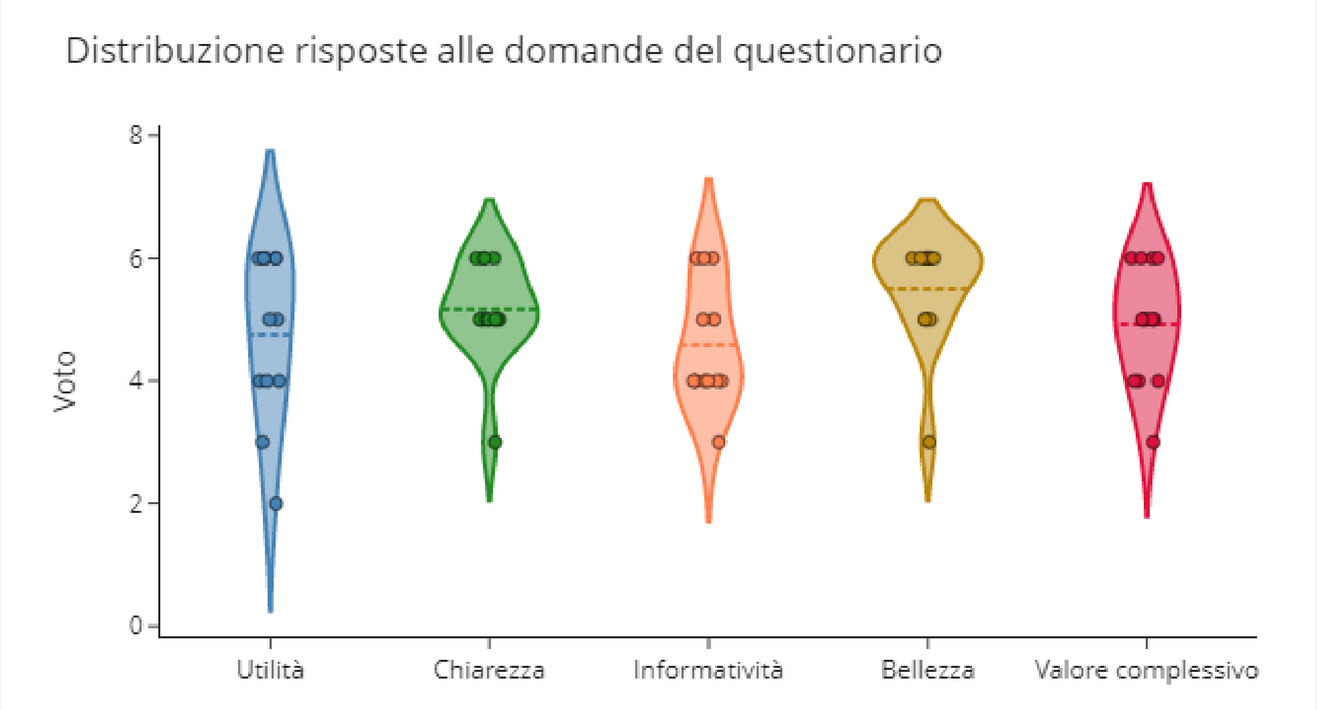
✓ Valutazione euristica

Durante la valutazione euristica, la quale è stata svolta su 3 persone, sono emerse le seguenti considerazioni o problematiche:

- Le frecce sulla visualizzazione erano inizialmente troppo piccole, infatti sono state ingrandite
- Le informazioni sulla media del rating di un film presenta troppe cifre decimali uguali a 0 quando si visualizzano le informazioni posizionando il cursore
- Alcuni punti dello scatter plot sono sovrapposti quando la visualizzazione non è zoommata

✓Questionario psicometrico

Per la valutazione psicometrica è stato somministrato il questionario Cabitza-Locoro ad un totale di 12 individui.
I risultati ottenuti sono i seguenti:

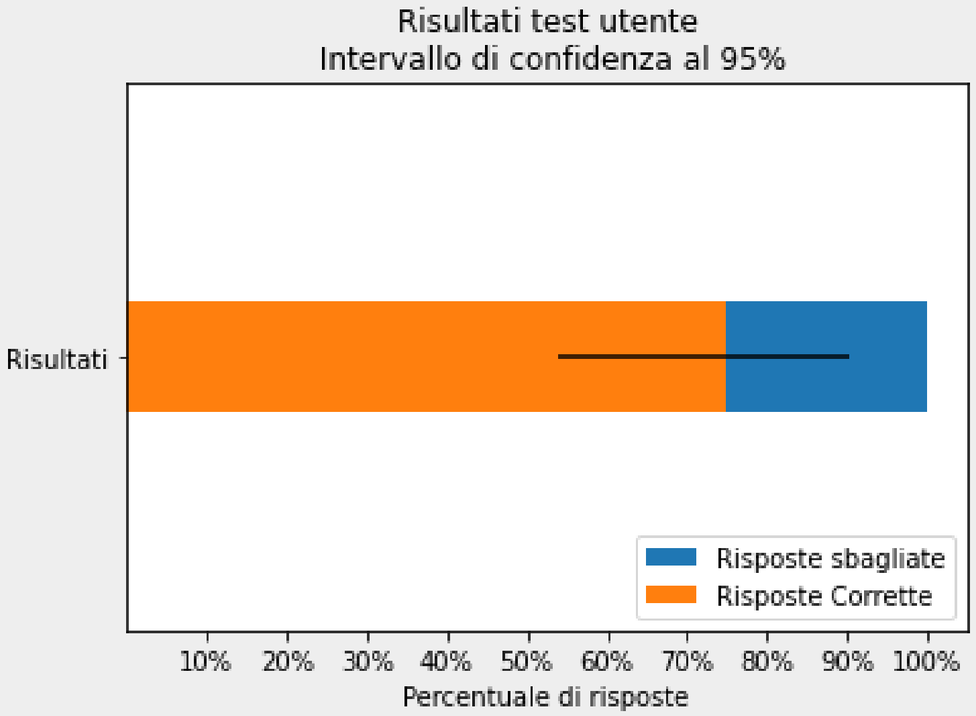


✓ Test utente

Per il test utente sono stati coinvolti 6 individui, a cui sono state sottoposte le seguenti domande [tempo stimato]:

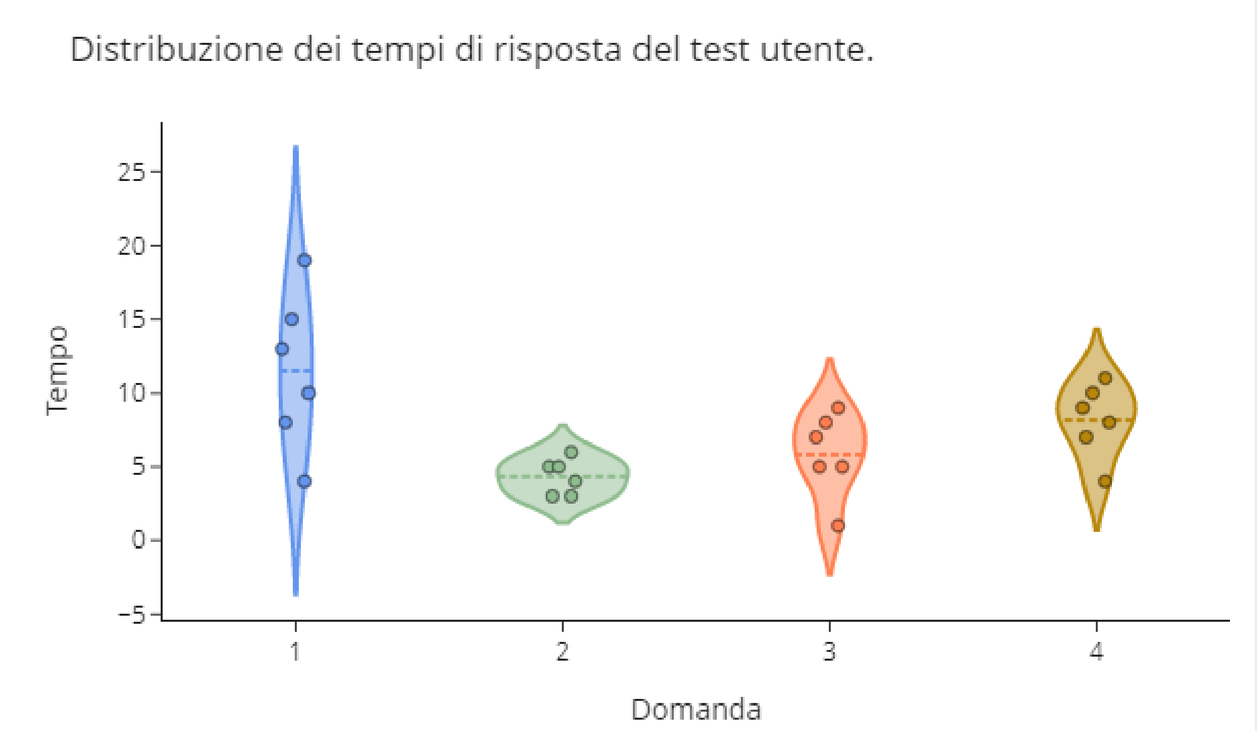
- Qual è il valore massimo di rating dei film considerati? [11s]
- Il film Zack Snyder's Justice league è un outlier? [4s]
- Qual è la durata minima fra i film considerati? [6s]
- Il film The cost of deception è nella fascia di normalità del rating? [8s]

✓ Test utente



Efficacia

Efficienza



Conclusioni

In conclusione, inizialmente avevamo ipotizzato che ad una durata crescente sarebbe corrisposto un rating medio decrescente ma, come abbiamo potuto osservare dalla visualizzazione finale, la nostra idea iniziale è stata smentita e ribaltata. All'aumentare della durata del film infatti il rating medio tende a salire leggermente, presentando una correlazione pari a 0,4.

Link alla repository

Blowdire/DataVisualizationProject (github.com)

Data Visualization