

# THE TV AND MOVIES DATA ANALYSIS

Alessandro Asperti 813224 [a.asperti@campus.unimib.it](mailto:a.asperti@campus.unimib.it)

Paolo Crotti 820933 [p.crotti1@campus.unimib.it](mailto:p.crotti1@campus.unimib.it)

Sofia Davoli 813479 [s.davoli1@campus.unimib.it](mailto:s.davoli1@campus.unimib.it)

*Sono passati 128 anni da quando i fratelli Lumière hanno inventato la cinepresa: invenzione che rivoluziona tutt'oggi il nostro modo di vivere. Il 16 maggio 1929 viene assegnato per la prima volta l'Academy Award, meglio conosciuto come premio Oscar. Da quel momento, il panorama dei premi cinematografici si amplia: i festival cinematografici sono oggi manifestazioni di grande prestigio che svolgono una fondamentale opera di valorizzazione e promozione di autori, correnti e tendenze del cinema contemporaneo.*

*Il progetto si propone di realizzare una base di dati contenente informazioni riguardanti i film, le serie TV e alcuni tra i più importanti premi assegnati in questi ambiti quali gli Oscar, i Golden Globe e gli Emmy. Nello specifico, l'obiettivo è quello di indagare le connessioni tra titolo e attori vincitori e la concordanza tra il parere della critica e il parere degli utenti.*

## Indice

### Introduzione

1. Dati e modalità di acquisizione
  - 1.1. Acquisizione dei dati di TMDb
  - 1.2. Acquisizione dei dati relativi ai premi
2. Pulizia dei dati
3. Enrichment
4. Data distribution con MongoDB
5. Infografiche e valutazione
  - 5.1. Think aloud
  - 5.2. User test
  - 5.3. Questionario

### Conclusioni e sviluppi futuri

# Introduzione

Il progetto si pone come obiettivo quello di creare ed analizzare un'ampia base di dati relativa al tema dei film, serie TV, persone coinvolte nella produzione (attori, direttori, ...) e premi. In particolare, i dati provengono dal database TMDb, uno tra i più completi in questo ambito. Le informazioni riguardanti i premi (Oscar, Emmy e Golden Globe) sono state acquisite da siti web. É importante specificare le differenze tra questi premi: i Golden Globe vengono assegnati sia a film che programmi televisivi, per quanto riguarda gli Oscar si tratta unicamente di film mentre gli Emmy sono premi televisivi.

Il lavoro è stato svolto per rispondere a due domande di ricerca:

- Esiste un legame tra l'assegnazione del premio ad un attore e la premiazione del film/serie TV di cui è protagonista? Il fatto che un premio venga assegnato a un film o una serie TV influenza la valutazione della performance del degli attori protagonisti?
- I premi assegnati rispecchiano le preferenze del pubblico? Il rating assegnato dagli utenti è conforme alle valutazioni della critica?

## ***Schema del progetto.***

Il lavoro svolto può essere sintetizzato come segue.

Acquisizione dei dati. Le modalità di acquisizione dei dati sono state tre:

- i. download dal database di TMDb tramite API;
- ii. web scraping dai siti ufficiali dei premi Emmy e Golden Globe;
- iii. download dalla piattaforma Kaggle dei dati relativi ai premi Oscar.

Pulizia dei dati. I dati acquisiti da TMDb non sono caratterizzati da grandi problemi in termini di pulizia: le principali difficoltà sono state riscontrate sui premi.

Enrichment. Dopo aver acquisito e ripulito i dati, è stato operato un arricchimento dei dati provenienti dal database di TMDb con quelli relativi ai premi.

Memorizzazione. La quantità di dati a disposizione richiede di essere gestita tramite un'architettura specifica: si è deciso di utilizzare un metodo per simulare la distribuzione dei dati su diverse macchine. Il metodo in questione è lo Sharding, implementato con MongoDB.

Sviluppo e valutazione delle infografiche. In questa fase si cerca di rispondere alle domande di ricerca attraverso delle infografiche. Sia in fase di sviluppo che dopo il rilascio, le infografiche sono state sottoposte a diversi utenti per apportare miglioramenti e ottenere feedback sulla base di diverse caratteristiche.

# 1. Dati e modalità di acquisizione

Come descritto nell'introduzione, le modalità di acquisizione dei dati sono tre. Nel seguito sono descritte con distinzione fra dati di TMDb e dati relativi ai premi.

## 1.1. Acquisizione dei dati di TMDb

La procedura per l'acquisizione dei dati di TMDb prevede due step:

- Il download di estrazioni giornaliere, una per ogni entità della base dati (persone, film e serie TV), contenenti l'elenco aggiornato di tutti gli identificativi univoci del database. Questi file sono disponibili per il download dal sito <http://files.tmdb.org> ogni giorno a partire dalle ore 7:00.
- La creazione dei dataset di partenza: questa procedura si effettua tramite la libreria Python *tmdbsimple* (<https://pypi.org/project/tmdbsimple/>) che permette l'interrogazione del database TMDb tramite API utilizzando gli id presenti nei file sopra menzionati.

In totale, come si può notare in figura 1, la procedura ha scaricato una quantità di dati pari a poco più di 3 GB.









Nome	Dimensione	Tipo
 tmdb_tvs_crew.json	19.810 KB	File JSON
 tmdb_tvs_cast.json	56.518 KB	File JSON
 tmdb_tvs.json	119.675 KB	File JSON
 tmdb_tv_seasons.json	915.489 KB	File JSON
 tmdb_people.json	463.907 KB	File JSON
 tmdb_movies_crew.json	515.843 KB	File JSON
 tmdb_movies_cast.json	704.215 KB	File JSON
 tmdb_movies.json	435.171 KB	File JSON

Figura 1: dimensioni dei file scaricati da TMDb

## 1.2. Acquisizione dei dati relativi ai premi

I dati relativi ai Golden Globe e agli Emmy sono stati estratti mediante web scraping (libreria Python *BeautifulSoup*) dai siti ufficiali (<https://www.goldenglobes.com> e <https://www.emmys.com>).

Le informazioni raccolte riguardano la tipologia di premio, l'anno, il programma, l'attore e un attributo relativo alla nomina o alla vittoria del premio. A differenza dei dati di TMDb, in questo caso la mole di dati è molto ridotta: i dataset ottenuti hanno dimensioni 1330 KB (Golden Globe) e 4021 KB (Emmy).

I dati riguardanti ai premi Oscar sono stati scaricati dal sito

<https://www.kaggle.com/unanimad/the-oscar-award>. Il file ha una dimensione di 892 KB.

## 2. Pulizia dei dati

Navigando i dati acquisiti si nota che ci sono alcune ripetizioni nei titoli, sono presenti persone e serie TV/film omonimi e i dataset relativi ai premi presentano molti errori di battitura (principalmente nei titoli dei film/serie TV) che si possono notare anche sul sito web. Dal momento che queste informazioni provengono da sorgenti diverse, prima di trattare i problemi di battitura, un'ulteriore operazione di pulizia consiste nell'uniformare i dati in termini di attributi.

Infine, è stata effettuata una correzione automatica delle stringhe dei titoli spostando gli articoli all'inizio (ad esempio: "Da Vinci Code, The" -> "The Da Vinci Code").

## 3. Enrichment

L'obiettivo di questa fase è quello di sviluppare un processo in grado di legare film, serie TV e persone presenti su TMBd con quelli presenti nei dati dei premi identificando su di TMDb il vincitore o nominato.

I codici per l'integrazione dei premi relativi ai film e dei premi relativi alle serie TV seguono un processo logico molto simile con delle piccole accortezze per quanto riguarda l'integrazione dei premi delle serie TV. Infatti, esistono alcuni premi Emmy

e Golden Globe che possono essere assegnati sia a una serie che a un film (ad esempio, *Best Television Limited Series or Motion Picture Made for Television*). Per risolvere questo problema è stato utilizzato un file di supporto, creato manualmente, che indica se il premio si riferisce solo a una serie TV oppure se esso può riferirsi sia a serie TV che film. Un'ulteriore differenza è legata all'anno che, nel caso di serie TV non è uno solo.

Il processo per l'arricchimento dei film e delle serie TV con le informazioni relative ai premi è descritto di seguito:

1. ricerca per titolo inglese o titolo in lingua originale;
2. ricerca per titolo inglese o titolo in lingua originale ripuliti da eventuali caratteri speciali (simboli, punteggiatura, ...);
3. ricerca sfruttando il motore di ricerca delle API di TMDb (query per titolo del film/serie TV);
4. per ogni premio rimasto si effettua una ricerca per anno considerando un range di variazione di due anni per i film e uno per le serie TV (per comprendere casi particolari come film usciti a cavallo della fine dell'anno e in momenti diversi in stati diversi).

I risultati di queste ultime due ricerche vengono analizzati valutando che le stringhe dei titoli individuati siano contenute in quelle del titolo del dataset dei premi o viceversa (una casistica può essere il film *Birdman or (The Unexpected Virtue of Ignorance* che fra i premi era chiamato semplicemente *Birdman*). Se il controllo dovesse risultare negativo, il matching andrebbe scartato. Si è deciso di tenere validi i matching per i titoli con almeno 6 caratteri per i film e 5 caratteri per le serie TV: al di sotto di queste soglie è rischioso considerarli validi perché si potrebbero verificare molti errori di validazione dovuti agli articoli. Un chiaro esempio può essere la serie TV *The 100*: dopo la procedura di rimozione dei caratteri speciali diventa "the" che

può essere considerata valida in corrispondenza di qualunque altra serie TV contenente la parola 'the' nel titolo.

Completata questa fase, si è passati alla deduplica dei matching, ovvero all'identificazione e rimozione dei casi in cui le procedure sopra descritte assegnano lo stesso premio a due serie TV/film diversi. Per la gestione di queste situazioni è stata applicata una funzione di risoluzione, la quale opera definendo un indicatore che varia da 0 a 5. Inizializzando l'indicatore in corrispondenza del valore 0, esso viene incrementato in base a:

- i. uguaglianza dei titoli (originali o inglesi): l'indicatore viene incrementato di 2;
- ii. il titolo è contenuto nel titolo presente nel dataset dei premi (originali o inglesi): l'indicatore viene incrementato di 1;
- iii. concordanza fra l'anno di uscita del film/serie TV e anno di assegnazione del premio: l'indicatore viene incrementato di 1;
- iv. popolarità del film (indicatore presente su TMDb): se la serie TV (o film) è prima in quanto a popolarità rispetto alle altre candidate per il matching, l'indicatore viene incrementato di 1.

Si fa notare che ogni matching identificato in fase di arricchimento rientra nel caso i) o ii), quindi sicuramente non viene considerato come vincente un matching candidato che presenta titoli completamente discordanti. A parità del valore dell'indicatore, si è deciso di considerare valido il matching associato al film/serie TV con popolarità più alta, in quanto è più probabile che un film popolare abbia vinto un premio.

La procedura di arricchimento relativa alle persone ha una struttura analoga a quella appena descritta, con differenze sulla deduplica: non potendo sfruttare l'anno, il matching vincente è semplicemente quello associato al personaggio più popolare.

Oltre alla valutazione effettuata sui punti 3 e 4 (titolo contenuto) dell'arricchimento, per le persone è stata utilizzata la *Damerau Levenshtein distance*. Inoltre, per identificare dei potenziali candidati, la fase di ricerca è stata svolta sfruttando anche le informazioni relative a cast e crew (le persone che hanno lavorato al film/serie TV), dal momento che film/serie TV di TMDb associati al premio erano ormai note.

I risultati sono riportati nella seguente tabella:

	Premi totali	Premi assegnati	Premi assegnati ma dubbi	Premi non assegnati
<b>Film</b>	14026	13868 (98,8%)	0	158 (1,2%)
<b>Serie TV</b>	10507	10195 (97,0%)	0	312 (2,9%)
<b>Persone</b>	38854	35878 (92,3%)	2870 (7,4%)	54 (0,1%)

## 4. Data distribution con MongoDB

Vista la grande quantità di dati a disposizione è stato utilizzato MongoDB per simulare la distribuzione dei dati su diverse macchine. Si parla di simulazione perché, nella pratica, il lavoro è stato svolto su una singola macchina puntando a diverse porte della stessa.

È stata creata una collezione per ognuna delle diverse entità oggetto di studio: i premi, i film, i programmi televisivi e le persone. In fase di scrittura, si è deciso di mantenere la struttura dei documenti forniti dalle API di TMDb (con alcune eccezioni, principalmente in merito all'annidamento dei documenti).

Sono stati creati due sharded clusters, ciascuno composto da tre nodi, due server di configurazione (in replica set) e un router (mongos).

Di seguito un esempio di come Mongo ha allocato i dati dei film fra i cluster:

```
Shard s1 at s1/127.0.0.1:27017,127.0.0.1:27018,127.0.0.1:27019
Data: 565.95MiB
Docs: 285329
```



*Chunks: 19*  
*Estimated data per chunk: 29.78MiB*  
*Estimated docs per chunk: 15017*

*Shard s2 at s2/127.0.0.1:27020,127.0.0.1:27021,127.0.0.1:27022*  
*Data: 501.23MiB*  
*Docs: 252698*  
*Chunks: 18*  
*Estimated data per chunk: 27.84MiB*  
*Estimated docs per chunk: 14038*

*Totals data: 1.04GiB*  
*Docs: 538027*  
*Chunks: 37*  
*Shard s1 contains 53.03% data, 53.03% docs in cluster, avg obj size on shard: 2KiB*  
*Shard s2 contains 46.96% data, 46.96% docs in cluster, avg obj size on shard: 2KiB*

Per verificare il funzionamento della distribuzione dei dati, per i film, si è deciso di caricare il 25% dei dati per volta, lanciando di volta in volta le query descritte nel seguito.

Nella prima, si cercano i titoli che hanno almeno cento valutazioni, ottenendo titolo, rating e overview dei dieci film con voto più alto (per ottenere il tempo d'esecuzione dell'interrogazione è stato utilizzato il comando *explain("executionStats")*).

```
db.movies.find({vote_count : {$gt : 100}},{title:1, vote_average:1, overview:1}).sort({vote_average:-1}).limit(10)
```

Nella seconda query, si vogliono ottenere titolo, voto e overview dei dieci film più popolari che hanno ricevuto almeno un premio.

```
db.movies.find({awards :{$exists : true}},{title:1, vote_average:1, overview:1}).sort({popularity:-1}).limit(10)
```

Dal grafico sottostante si può notare il tempo effettivo delle due interrogazioni, ovvero i secondi trascorsi in funzione della percentuale dei dati caricati: anche se molto meno incisivo per la seconda query, si osserva un andamento crescente per entrambe.

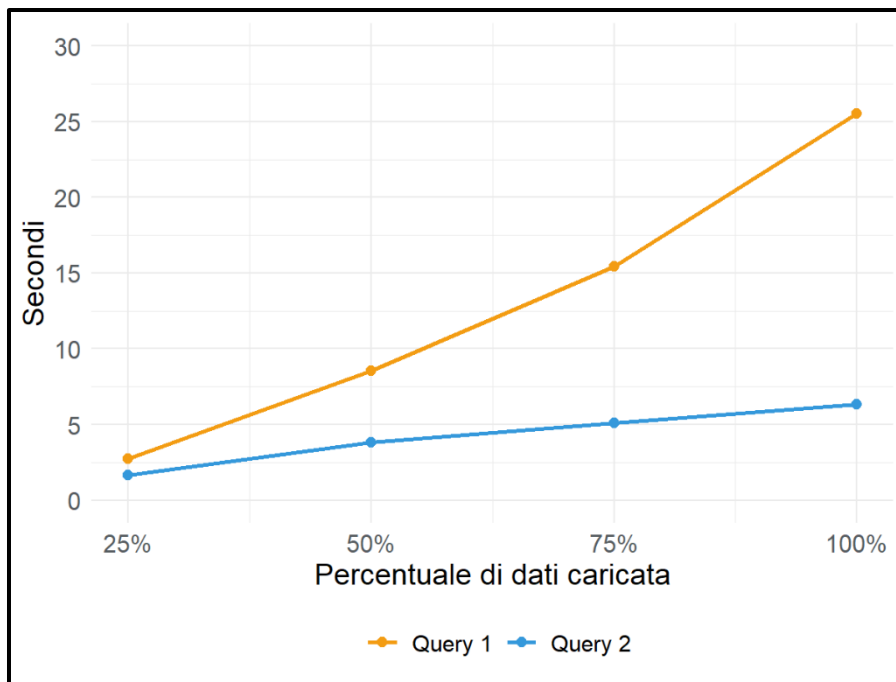


Figura 2: line plot del tempo di caricamento

In figura 3 viene invece mostrato il rapporto fra il tempo di caricamento corrente e quello del caricamento precedente. L'andamento si dimostra essere non lineare, bensì il tempo di esecuzione in relazione all'aumento della percentuale mostra un andamento esponenziale negativo per entrambe le query: il tempo non raddoppia al raddoppiare della quantità di dati.

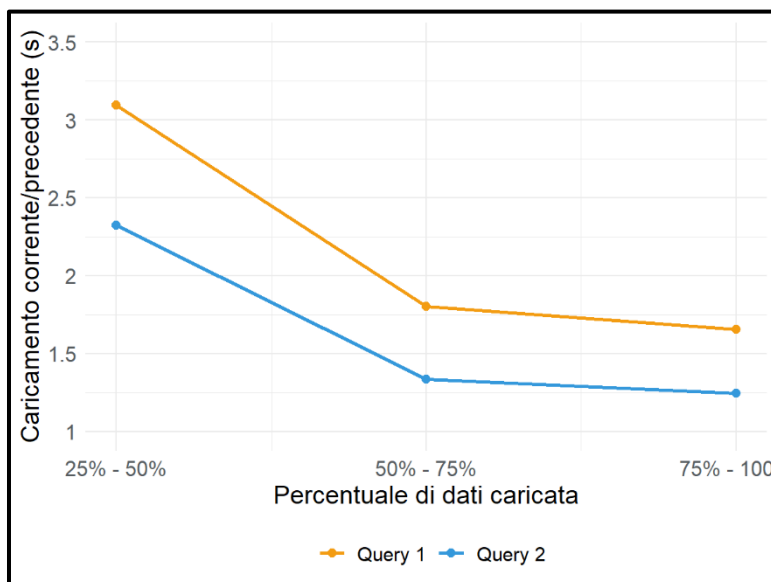


Figura 3: line plot del rapporto fra caricamento corrente e caricamento precedente

## 5. Infografiche e valutazione

Per rispondere alle domande di ricerca sono state create due visualizzazioni, consultabili al seguente link:

<https://public.tableau.com/profile/paolo1754#!/vizhome/Thetvandmoviesdataanalysis2/Dashboard?publish=yes>. Per lo sviluppo sono stati considerati i dati dell'ultimo

ventennio dei premi più importanti: miglior attore/attrice protagonista, miglior film e serie TV per il genere comico, drammatico e musical

Si riportano i dettagli.

- Prima visualizzazione (stacked barchart): permette di analizzare le frequenze dei casi in cui attore/attrice protagonista vincenti non hanno lavorato al film/serie TV che ha vinto il premio e i casi in cui gli attori vincenti erano protagonisti del film/serie TV ha vinto il premio (ci riferiamo a questo secondo caso con il termine *vittoria congiunta*).
- Seconda visualizzazione (scatter plot): per verificare se il premio assegnato annualmente rispecchia le preferenze del pubblico e se si possono trarre conclusioni in merito ad un ipotetico legame fra la presenza di *vittorie congiunte* e il rating, è stato prodotto un grafico che permette di confrontare il giudizio della critica con quello degli utenti, studiando al contempo la distribuzione delle *vittorie congiunte*.

### 5.1. Think aloud

Le visualizzazioni sono state sottoposte a diverse persone, lasciandole libere di parlare e commentare ad alta voce (utenti compresi nella fascia d'età 15-70).

Nella fase think-aloud sono stati evidenziati i seguenti problemi, successivamente risolti:

- Cos'è il rating?

Non tutti sanno cosa si intenda per rating e quali valori può assumere, quindi è stato spiegato con l'ausilio di un'annotazione (e tradotto in italiano)

- Solo il Golden Globe viene dato sia a film che a serie TV? E gli Oscar e gli Emmy no?

Non tutti si interessano a questi argomenti: è comune non sapere le differenze tra i vari premi. Questo aspetto è stato descritto nell'introduzione.

- Il doppio asse sulla prima visualizzazione confondeva molto le idee delle persone coinvolte.

Inizialmente, nella prima visualizzazione, come mostrato in figura 4, su entrambi gli assi era presente una descrizione che, attraverso un'associazione di colori e un asse inverso esplicitava il significato delle barre. È stata eliminata la descrizione sul doppio asse e gli assi sono stati orientati nello stesso verso per ottenere una visualizzazione più pulita.

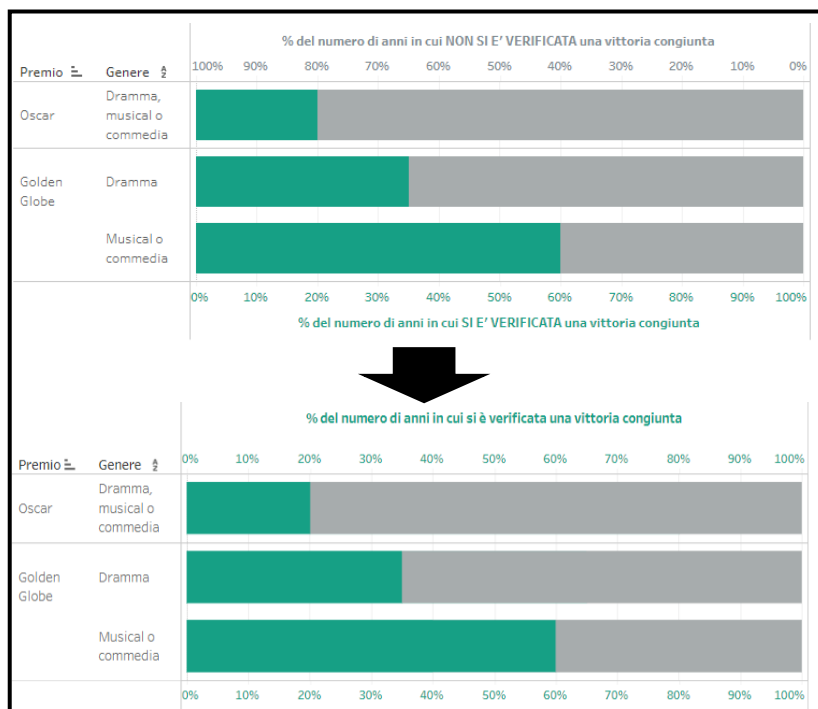


Figura 4: modifica degli assi nella prima visualizzazione

## 5.2. User test

Gli utenti sono stati sottoposti a una serie di task da risolvere. Di seguito sono elencate le domande per la prima visualizzazione:

1. Quale premio ha avuto il minor numero di vittorie congiunte?

Risposta: Oscar

2. Per le serie TV quel è stato il premio che ha ottenuto il maggior numero di vittorie congiunte?

Risposta: Golden Globe genere Musical o Commedia

3. Per questo premio, quante sono state le vittorie congiunte?

Risposta: 11 (l'utente deve posizionarsi sulla barra per visualizzare l'informazione)

Le domande relative alla seconda visualizzazione sono invece le seguenti:

4. Ci sono state frequenze congiunte nei Golden Globe (dramma) nell'anno 2019?

Risposta: No (l'utente deve filtrare per premio-categoria e selezionare il quinquennio di interesse)

5. Nel 2013 per gli Emmy (dramma), quale film ha vinto? L'attore protagonista è stato nominato?

Risposta: Breaking Bad, l'attore protagonista è Bryan Cranston (l'utente deve applicare gli appositi filtri e posizionarsi sul vincitore)

Per quanto riguarda la fase di risoluzione dei task, come si può osservare in figura 5, sulle singole domande, sono state riscontrate delle tempistiche relativamente omogenee fra gli intervistati. I task semplici vengono risolti in breve tempo mentre quelli più complessi (che necessitano di applicare dei filtri), richiedono secondi in più.

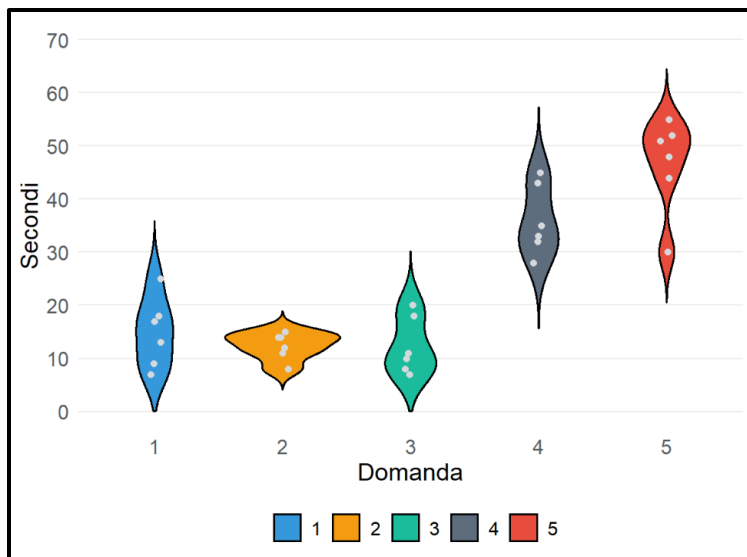


Figura 5: violin plot del tempo di risposta alle domande

Dal grafico a barre orizzontali illustrato in figura 6, si nota che gli intervistati hanno risposto correttamente alla maggior parte dei task ad eccezione del primo (in cui solo il 66,6% ha risposto correttamente) e del quarto.

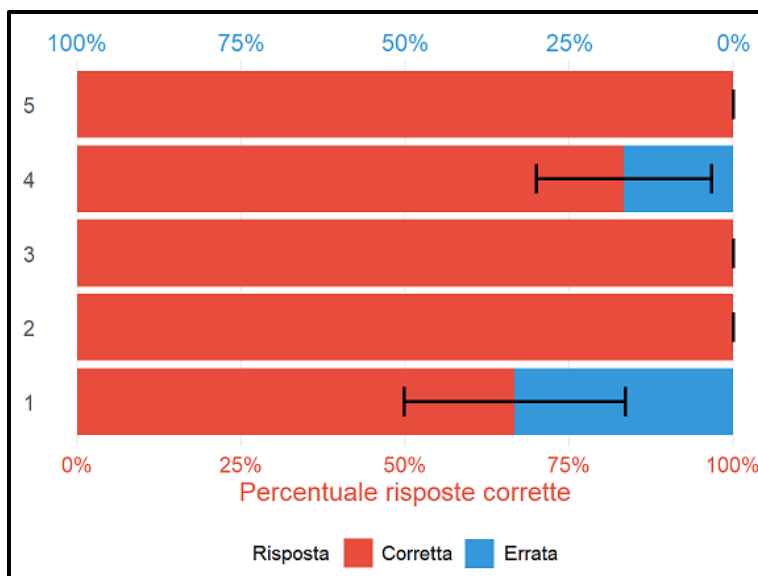


Figura 6: stacked barchart che mostra la percentuale di risposte corrette

### 5.3. Questionario

Come ultima valutazione, 38 utenti (compresi nella fascia d'età 15-70), sono stati sottoposti ad un questionario strutturato secondo la scala Cabitza-Locoro mostrata in figura 7.

Si valuti la qualità della visualizzazione riportata indicando, per ogni caratteristica, un valore da 1 \* (pochissimo) a 6 (moltissimo).

	1 - Pochissi...	2	3	4	5	6 - Moltissimo
Chiara	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Bella	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Utile	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Informativa	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Complessivo	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figura 7: questionario con scala Cabitza–Locoro

Dal questionario risulta che la maggior parte degli intervistati ha espresso un parere positivo sotto ogni aspetto della valutazione: il 12,5% ha trovato poco utile la storia mentre solo il 7,5% l'ha trovata poco chiara e poco bella.

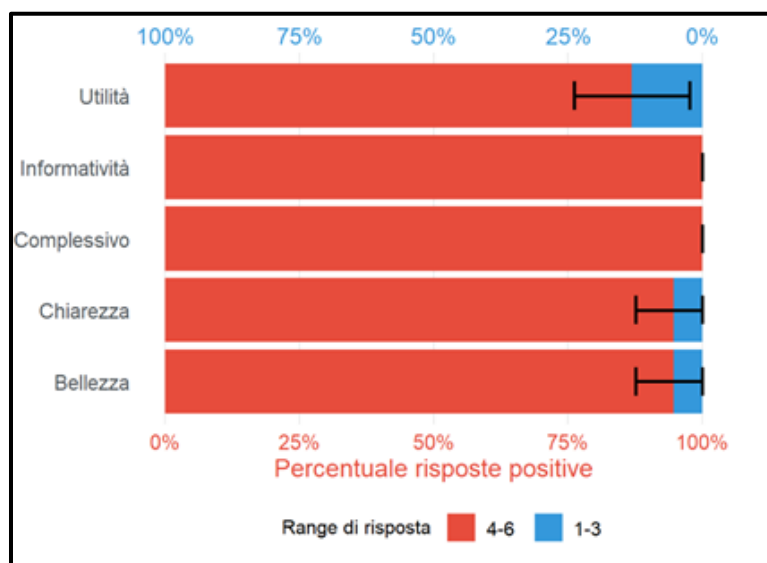


Figura 8: grafico delle risposte al questionario

Di seguito è riportato il correlogramma delle risposte, nel quale risulta che *chiarezza* e *bellezza* sono gli aspetti che influenzano maggiormente il voto complessivo delle visualizzazioni.

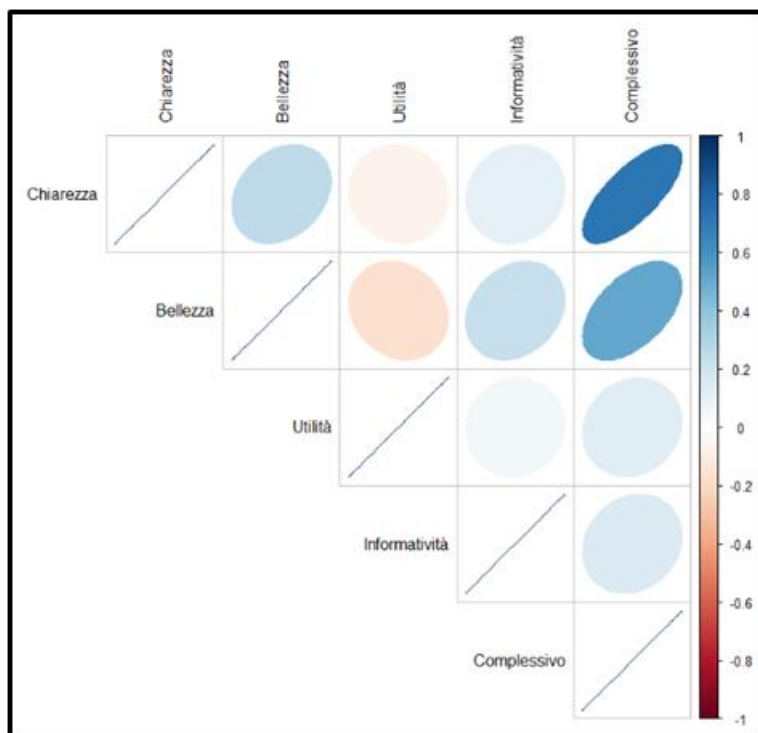


Figura 9: correlogramma delle risposte al questionario

## Conclusioni e sviluppi futuri

Dalle infografiche è possibile notare che per i Golden Globe, sia per serie TV che film, e, sia per il genere drammatico che per musical e commedie, le *vittorie congiunte* sono state maggiormente frequenti. Quindi, si potrebbe pensare che la performance di uno (o entrambi, eventualmente) gli attori protagonisti ha influenzato la vittoria del premio per il miglior film/serie TV. Analogamente, si potrebbe affermare che una serie TV/film prodotta egregiamente in termini di scrittura, montaggio, effetti speciali e altre caratteristiche abbia portato a sovrastimare il giudizio nei confronti degli attori protagonisti coinvolti. A differenza di Emmy e Golden Globe, per gli Oscar, è stata osservata una minore frequenza di *vittorie congiunte* fra titolo e protagonista. In particolare, per quest'ultimo, a differenza degli altri due riconoscimenti, negli ultimi 20 anni non vi sono stati casi in cui hanno vinto sia gli attori protagonisti che il film (o serie TV) a cui hanno partecipato.



Infine, si potrebbe dire che ai film o serie TV che hanno vinto, viene attribuito un rating alto anche dall'utenza di TMDb: il giudizio della critica rispecchia abbastanza quello degli spettatori.

Di seguito alcuni possibili miglioramenti:

- Dato che il sito TMDb aggiorna continuamente la base dati con i nuovi titoli e ogni anno si tiene la cerimonia di premiazione di Oscar, Golden Globe ed Emmy, si potrebbe sviluppare un meccanismo automatico in grado di aggiornare la base dati creata (invece di scaricare nuovamente tutti i dati).
- Vista la grande mole di dati, esistono sicuramente ulteriori domande a cui rispondere e temi su cui indagare. Ad esempio, un'idea (inizialmente attuata ma in seguito scartata per mancanza di dati di dettaglio) potrebbe essere quella di analizzare l'andamento del rating in relazione all'aumentare del numero di stagioni delle serie TV: è abbastanza comune pensare che dopo un certo numero di stagioni, i produttori, portino avanti le serie TV più per una questione di incassi che per la quantità di idee a disposizione.
- Per affinare l'indagine sui titoli e i loro interpreti protagonisti, si potrebbe pensare di considerare l'aggiunta di altri premi. Il mondo delle premiazioni è vasto: quasi ogni stato ha una propria mostra cinematografica, più o meno rinomata, come il Leone di Venezia, La Palma di Cannes o L'Orso di Berlino.