Progetto Movielens

Francesco Cecconello

13 giugno 2023

1 Introduzione

L'obiettivo di questo progetto era eseguire un'analisi del dataset Movielens, contenente le valutazioni assegnate da alcuni utenti a svariati film, studiando diverse caratteristiche del dataset tramite un'analisi esplorativa e rispondendo a domande specifiche attraverso l'utilizzo di query.

Poiché tutti i test sono stati eseguiti su una partizione ridotta con sistema operativo Linux, per questo lavoro sono stati considerati i dataset *ratings.csv* e *movies.csv* estratti dal pacchetto *Movielens 1M*, in modo da non eccedere i limiti dell'heap e del garbage collector durante l'esecuzione delle query. Va comunque segnalato che, a parte la query riguardante la clusterizzazione, è stato possibile eseguire il resto degli script anche sul dataset *Movielens 25M*.

Per la realizzazione dell'intero progetto, fra i principali moduli di Python utilizzati troviamo *pyspark*, *numpy*, *matplotlib* e *seaborn*.

2 Analisi esplorativa

Tutte le query riguardanti l'analisi esplorativa del database sono state eseguite solamente sul dataset *ratings.csv*.

2.1 Analisi 1: Calcolare il numero di rating ricevuti da ogni film

Per rispondere a questa interrogazione è sufficiente raggruppare i film per **movieId**, ovvero l'identificativo di ogni film, e contare gli elementi per ogni gruppo. Di seguito i primi 15 film con maggior numero di rating in ordine decrescente.

movieId	**Number of Ratings**
356	329
318	317
296	307
593	279
2571	278
260	251
480	238
110	237
589	224
527	220
2959	218
1	215
1196	211
2858	204
50	204

Tabella 1: Numero di rating per film

Per mostrare la distribuzione del numero di rating per film, inizialmente si è pensato di rappresentare i dati tramite un istogramma che avesse sull'asse x il numero di rating per film e sull'asse y il numero di film, come sotto riportato.

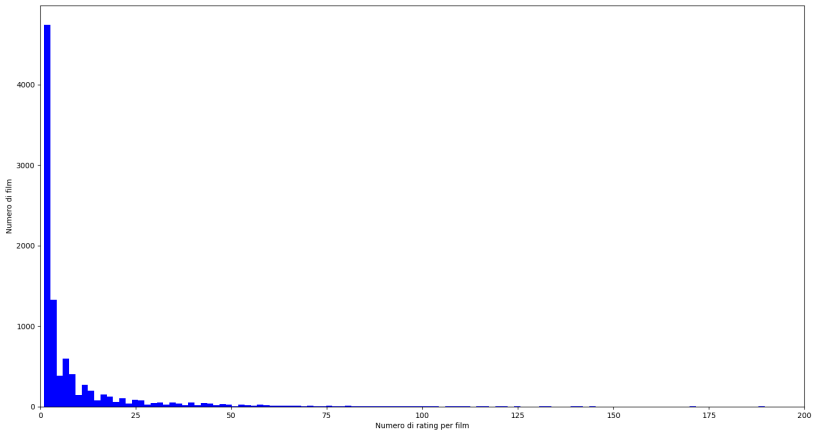


Figura 1: Numero di rating per film.

In realtà è risultato più esplicativo l'utilizzo di un grafico a linea continua che non rappresentasse l'esatta corrispondenza tra numero di rating per film e numero di film, ma che mostrasse la densità di tale distribuzione.

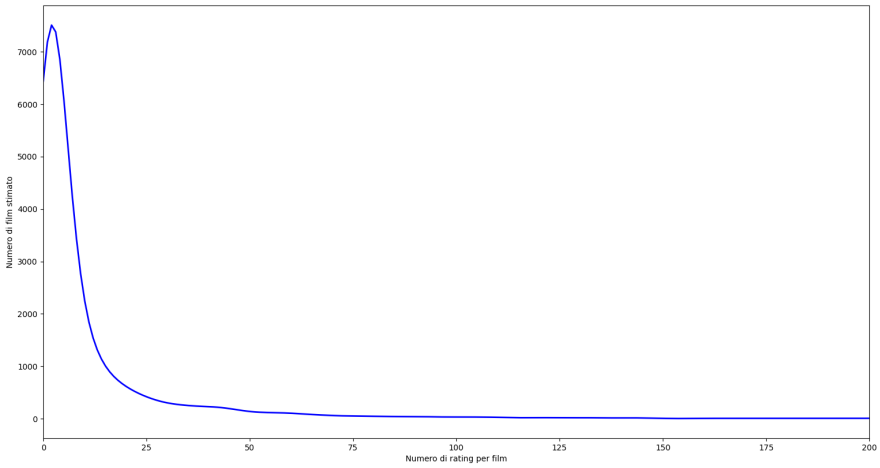


Figura 2: Distribuzione del numero di rating per film.

In questo modo non si rappresenta direttamente il numero di film in base al numero di rating, ma piuttosto una stima della probabilità di trovare un certo valore di rating per un film nel dataset. Il numero medio di rating per film è di circa 10. Si può notare come dopo il valore 25 il numero di rating per film cali vertiginosamente, fino ad annullarsi tra i 50 e i 75 rating per film.

2.2 Analisi 2: Calcolare il numero di rating per utente

In questo caso l'approccio è ortogonale a quello della prima query, poiché dobbiamo contare il numero di rating assegnati a tutti i film per ogni utente, perciò raggruppiamo per **userId**. Come fatto in precedenza, possiamo mostrare (in ordine decrescente) la lista dei primi 15 utenti ordinati per numero di rating.

userId	Number of Ratings
414	2698
599	2478
474	2108
448	1864
274	1346
610	1302
68	1260
380	1218
606	1115
288	1055
249	1046
387	1027
182	977
307	975
603	943

Tabella 2: Numero di rating per utente.

Il grafico corrispondente alla distribuzione del numero di rating per utente è il seguente.

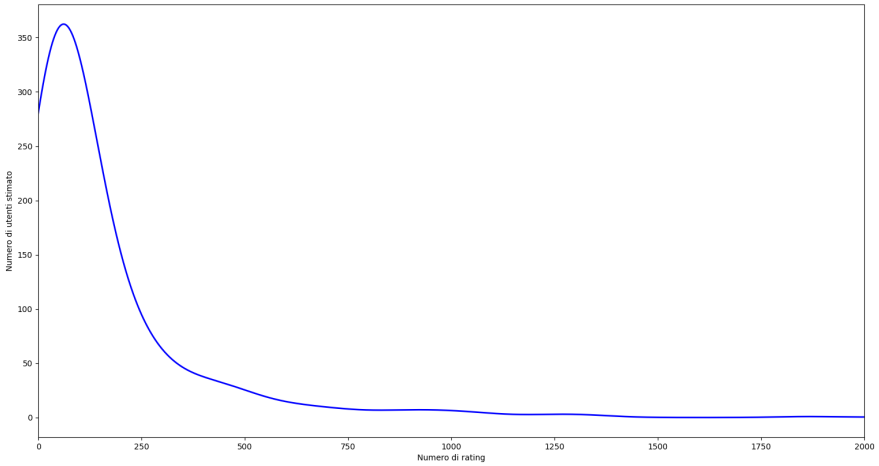


Figura 3: Distribuzione del numero di rating per utente.

Si può notare un picco del grafico attorno al valore di circa 100 rating per utente, grafico che poi andrà quasi ad azzerarsi una volta superata la soglia dei 750 rating per utente.

2.3 Analisi 3: Calcolare il punteggio medio per film

Per questa query è stato sufficiente raggruppare i rating per **movieId** e poi calcolarne la media. Di seguito la tabella con 15 di questi rating medi per film.

movieId	AverageRating
296	4.197068403908795
1090	3.984126984126984
115713	3.9107142857142856
3210	3.4761904761904763
88140	3.546875
829	2.6666666666666665
2088	2.5
2294	3.2444444444444445
4821	3.1
48738	3.975
3959	3.625
89864	3.6315789473684212
2136	2.4642857142857144
691	3.3333333333333335
3606	3.75

Tabella 3: Rating medio per film.

In questo grafico si può anche visualizzare l'evoluzione del rating per numero di film. Si può osservare come il rating medio totale si attesti attorno al valore di 3.5.

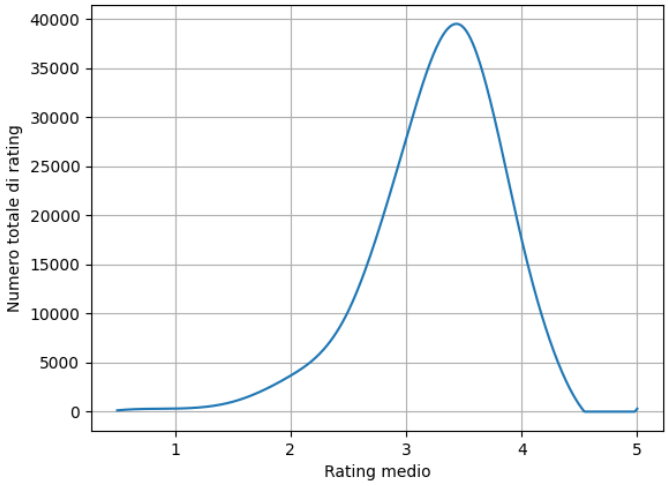


Figura 4: Evoluzione del rating medio per film.

2.4 Analisi 4: Calcolare il punteggio medio per utente

Il ragionamento fatto è lo stesso della terza query, andando a raggruppare per **userId**. Di seguito la tabella contenente il rating medio per 15 utenti.

userId	AverageRating
296	4.166666666666667
467	3.409090909090909
125	3.859722222222222
451	3.7941176470588234
7	3.2302631578947367
51	3.7757660167130918
124	3.99
447	3.871794871794872
591	3.2777777777777777
307	2.6656410256410257
475	4.409677419354839
574	3.9565217391304346
169	4.24907063197026
205	3.8703703703703702
334	3.418831168831169

Tabella 4: Rating medio per utente.

Il grafico che rappresenta il rating medio per utente è il seguente.

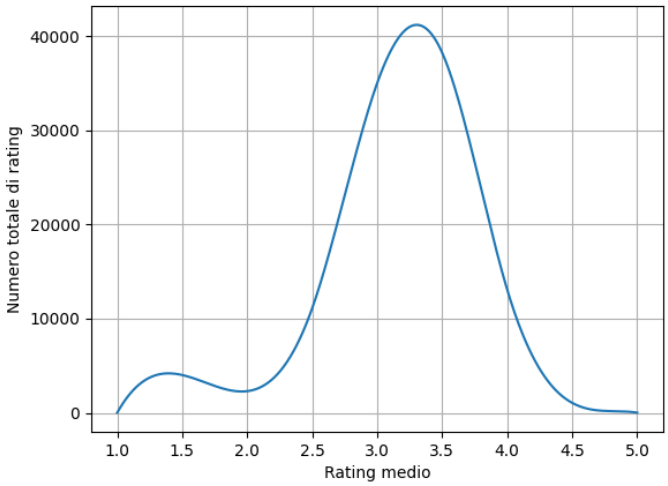


Figura 5: Evoluzione del rating medio per utente.

2.5 Analisi 5: Mostrare i migliori K film che abbiano ricevuto almeno R rating

Oltre al raggruppamento per film, è stato necessario filtrare i film per numero di rating. Di seguito la tabella che mostra i primi $K = 10$ film che abbiano almeno $R = 20$ rating ciascuno, ordinati per rating.

movieId	Number of Ratings	Average Score
1104	20	4.475
318	317	4.429022082018927
922	27	4.333333333333333
898	29	4.310344827586207
475	25	4.3
1204	45	4.3
246	29	4.293103448275862
858	192	4.2890625
1235	26	4.288461538461538
168252	25	4.28

Tabella 5: Migliori 10 film con almeno 20 rating ciascuno.

3 Query

Terminata l'analisi esplorativa, si può procedere a presentare i risultati delle query vere e proprie.

3.1 Query 1: Dimostrare se esiste o meno una correlazione tra la deviazione standard dei rating ricevuti da un film e il numero di rating.

Per questa query si è reso necessario calcolare la deviazione standard delle valutazioni per ogni film e il numero totale di valutazioni per ciascun film, per poi calcolare la correlazione tra la deviazione standard delle valutazioni e il numero di valutazioni utilizzando sia il metodo delle correlazioni di Pearson, già presente in *pyspark*, sia il metodo di Kendall; questa scelta è dovuta alla constatazione che il metodo di Kendall si presta maggiormente al calcolo della correlazione con valori ordinali (come i rating), mentre il metodo di Pearson indica se esiste o meno una correlazione lineare fra le due variabili.

Il valore di correlazione di Pearson calcolato è 0.08, indicando una correlazione lineare molto debole tra le due variabili prese in considerazione. Sebbene un valore così basso suggerisca che le due variabili abbiano una relazione estremamente scarsa o quasi inesistente, questo non implica necessariamente una mancanza di relazione tra di esse, ma indica solamente che la correlazione lineare è molto scarsa. Il valore di correlazione di Kendall, invece, è 0.18; pur non indicando una forte correlazione fra la deviazione standard dei rating e il numero di rating per film, questo risultato suggerisce una lieve dipendenza fra la classifica dei film e il numero di rating.

In generale, quindi, possiamo affermare che, sebbene la deviazione standard delle valutazioni dei film non abbia un impatto troppo significativo sul numero di valutazioni ricevute da ciascun film, la posizione in classifica dei film influisce lievemente sul numero di rating.

3.2 Query 2

La seconda query è stata suddivisa in 3 subquery.

3.2.1 Mostrare l'evoluzione del punteggio medio e del numero medio di rating nel tempo.

Per questa prima interrogazione, possiamo mostrare in due grafici l'andamento dell'average score totale e del numero di rating totale con granularità di 1 anno.

I due grafici sembrano quasi speculari, ad indicare che il rating medio sia inversamente proporzionale rispetto al numero di rating. In effetti, calcolando la correlazione di Pearson si ottiene un valore uguale a -0.51, indicando una media correlazione negativa fra le due variabili.

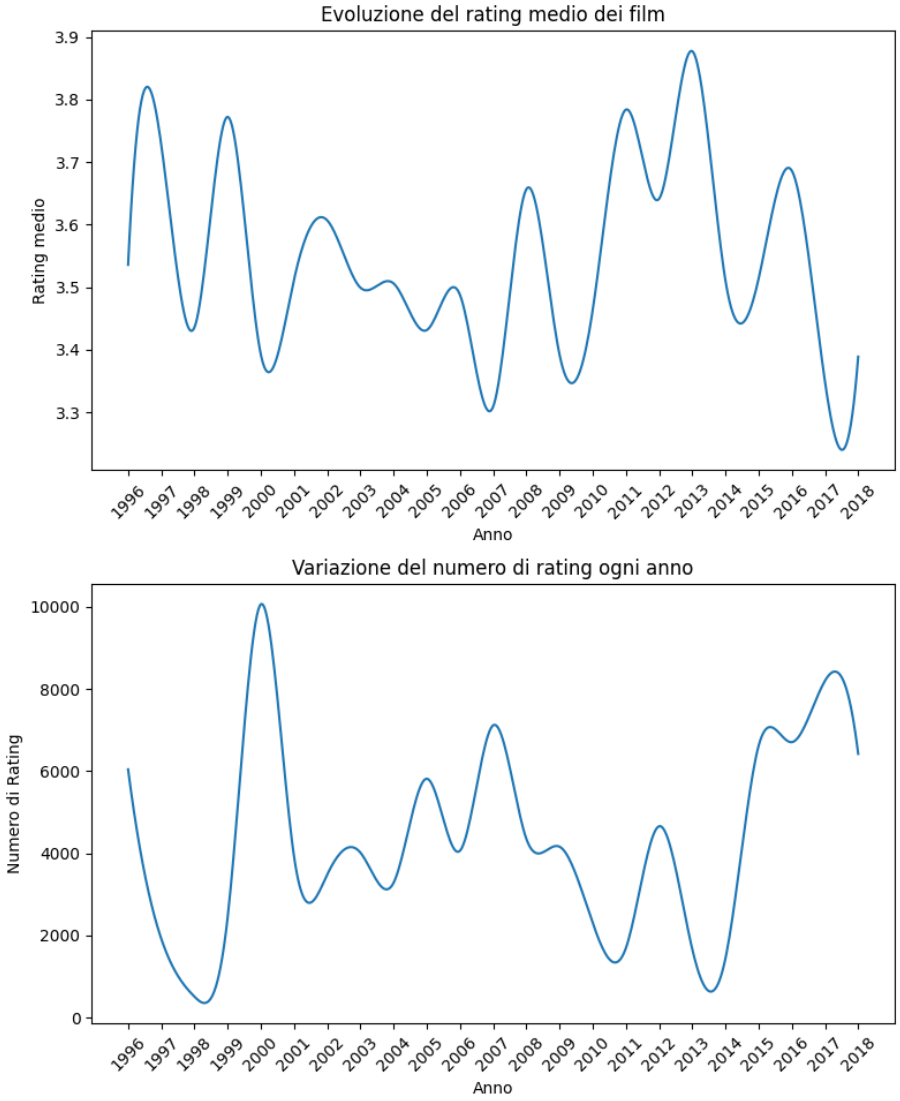


Figura 6: Confronto dell'evoluzione del rating medio e del numero di rating.

3.2.2 Mostrare se i film con score alto mantengono il proprio rating nel tempo.

Poiché mediamente un film riceve 10 rating (vedi Analisi 1), andremo a monitorare i film con almeno 10 rating e i cui primi 20 rating abbiano un valore medio maggiore o uguale al valore di rating medio 3.5 (vedi Analisi 3), ovvero i film classificati come *high rated*. Confrontando lo score medio iniziale, ovvero quello dei primi 20 rating, con quello totale, possiamo ottenere la percentuale di film che ha mantenuto uno score medio maggiore o uguale a 3.5, ovvero il 94% del totale. Quindi i film con score alto mantengono per la maggior parte il proprio rating nel tempo.

3.2.3 Mostrare se i film con score basso vengono abbandonati nel tempo.

Per capire se il numero di rating dei film *low rated* diminuisca o meno nel tempo, possiamo confrontare l'andamento del loro numero di rating con quello dei film *high rated*.

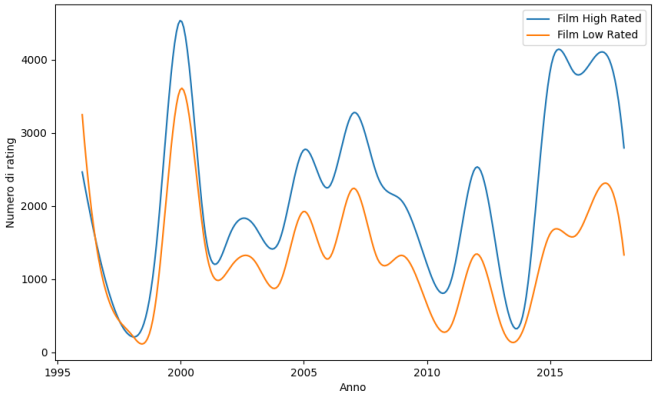


Figura 7: Confronto dell'evoluzione del numero di rating.

Come si può vedere, nonostante inizialmente il numero di rating si distribuisca equamente fra i vari film, dopo poco tempo i film *low rated* iniziano a ricevere molti meno rating rispetto ai film *high rated*.

3.3 Query 3: Confronto del rating medio identificando gruppi di utenti in base al numero di rating.

Gli utenti sono stati suddivisi in 7 gruppi:

- Utenti con meno di 20 rating
- Utenti con numero di rating compreso fra 20 e 30
- Utenti con numero di rating compreso fra 30 e 50
- Utenti con numero di rating compreso fra 50 e 100
- Utenti con numero di rating compreso fra 100 e 200
- Utenti con numero di rating compreso fra 200 e 500
- Utenti con numero di rating compreso fra 500 e 1000

Per ognuno di questi gruppi è stato calcolato il rating medio, ottenendo il seguente grafico.

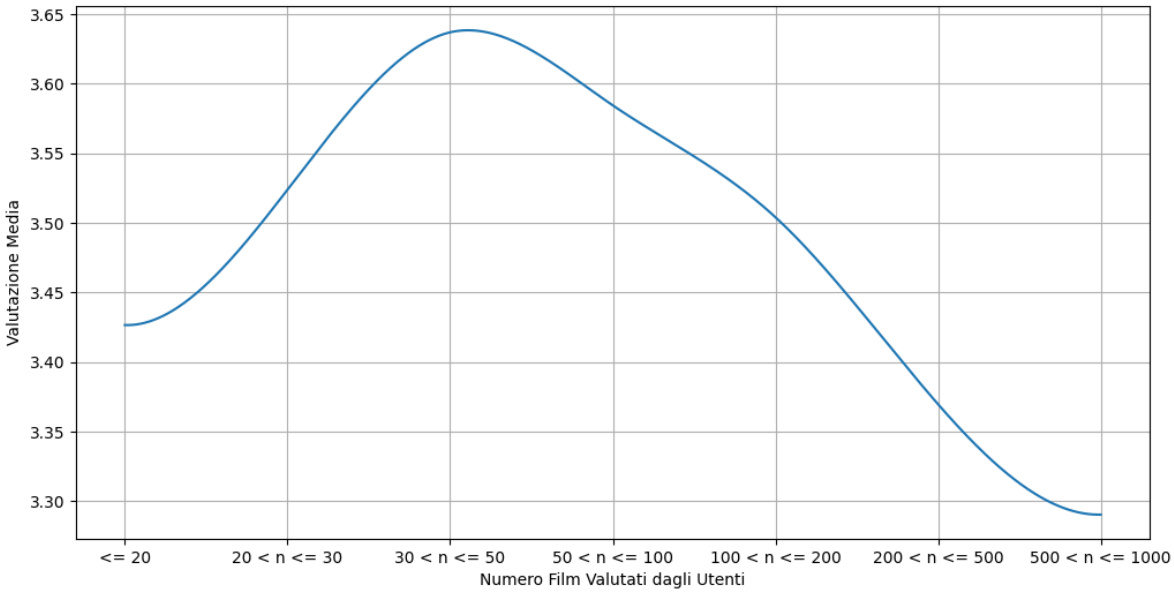


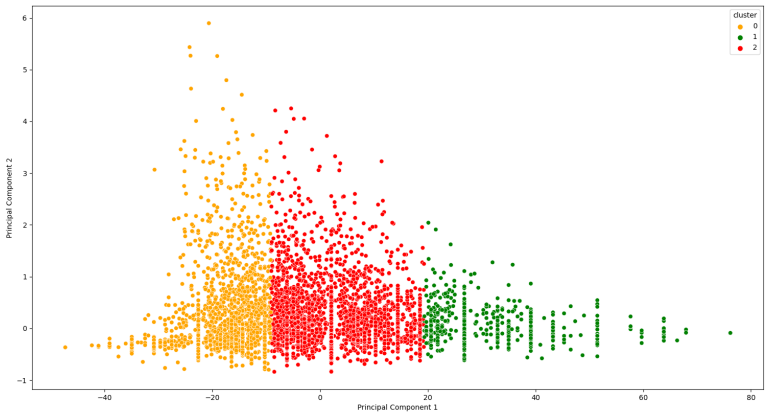
Figura 8: Evoluzione del rating medio in base al gruppo di utenti.

Si può notare come il rating medio raggiunga i valori estremi di rating per valori estremi di numero di rating, ovvero come gli utenti col minimo numero di rating e col massimo numero di rating abbiano il rating medio più basso.

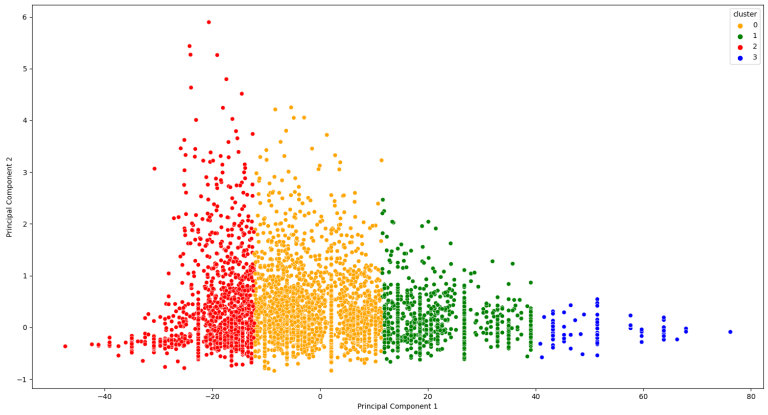
In generale, gli utenti più "buoni" sono quelli di fascia media, mentre i più "cattivi" sono quelli con più rating all'attivo.

3.4 Query 4: Cercare dei cluster di film simili.

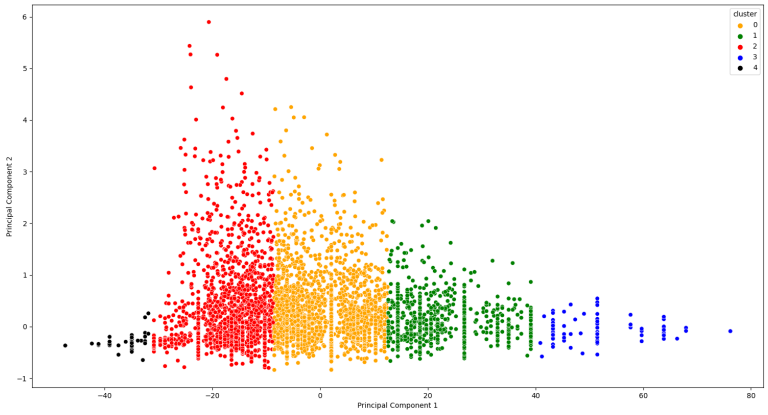
I cluster sono stati individuati in base a recensioni simili di più utenti per lo stesso film. Inizialmente si è provato a suddividere i film in 5 cluster, poi in 4 ed infine in 3. Come si può notare dai grafici sotto riportati, la suddivisione migliore sembra essere quella in 4 cluster, poiché nel caso dei 5 cluster il gruppo nero risultava essere di cardinalità e di varianza troppo ridotte rispetto agli altri, mentre nel caso dei 4 cluster il gruppo blu mostrava comunque una varianza non trascurabile lungo la prima componente.



(a) 3 - Clustering



(b) 4 - Clustering



(c) 5 - Clustering

Figura 9: Tentativi di clustering.