UNIVERSITÁ DEGLI STUDI MILANO-BICOCCA

Corso di laurea in Informatica

Metodi Informatici per la Gestione Aziendale

A.A. 2021/2022

Report di:

Davide Grandesso 852078

Sommario

1 Data acquisition

Il dataset preso in considerazione è Movielens 100k, esso contiene:

- 100,000 valutazioni (1-5) da 943 users su 1682 movies.
- Ogni utente ha valutato almeno 20 film.
- Semplici informazioni demografiche per gli utenti (età, genere, occupazione, zip)

1.1 File più informativi

I file contenenti le principali informazioni sono:

- **u.data**
- **u.item**
- **u.user**

1.1.1 u.data

Il dataset completo di dati u contiene le 100000 valutazioni di 943 utenti su 1682 item dove ogni utente ha valutato almeno 20 film, gli utenti e gli item sono numerati consecutivamente di 1 in 1, i dati presenti sono ordinati.

Questa è un elenco separato di:

user id | item id | rating | timestamp

Ogni colonna è separata dal simbolo di tabulazioni (‘\t’)

1.1.2 u.item

In questo file troviamo informazioni sugli item (film), questo è un elenco separato dal simbolo |:

movie id | movie title | release date | video release date | IMDb URL |

unknown | Action | Adventure | Animation | Children's | Comedy |

Crime | Documentary | Drama | Fantasy | Film-Noir | Horror | Musical |

Mystery | Romance | Sci-Fi | Thriller | War | Western |

Gli ultimi 19 campi sono i generi dei, un 1 indica che il film è di quel genere, uno 0 indica che non lo è, i film possono essere in diversi generi contemporaneamente. Gli id dei film sono quelli usati nel set di dati u.data.

- **movie id** → Id univoco del film che è stato valutato. (da 1 a 1682)
- **movie title** → Titolo del film (stringa)
- **release date** → Data di uscita del film (formato day-mount-year dove il mese è scritto come abbreviazione e non come numero)
- **video release date** → data di rilascio del video
- **IMDb URL** → URL del film al database di IMDb

unknown | Action | Adventure | Animation | Children's | Comedy |

Crime | Documentary | Drama | Fantasy | Film-Noir | Horror |

Musical | Mystery | Romance | Sci-Fi | Thriller | War | Western |

Questi ultimi 19 campi sono i generi dei film:

- 1 indica che il film è di quel genere
- 0 indica che non lo è

I film possono appartenere in diversi generi contemporaneamente.

1.1.3 u.user

Informazioni demografiche sugli utenti, questo è un elenco separato dal simbolo | di:

user id | age | gender | occupation | zip code

Ogni colonna è separata dal simbolo |

Gli id degli utenti sono quelli usati nel dataset u.data.

- **user id** → Id univoco dell'utente che ha effettuato la valutazione. (da 1 a 943)
- **age** → Quanti anni ha l'utente
- **gender** → Genere dell'utente ("M"/"F")
- **occupation** → Occupazione dell'utente
- **zip code** → Codice postale (USA)

1.2 Altri file informativi

Nella cartella troviamo anche altri file informativi ovvero:

- **u.info** → Il numero di utenti, di item e di valutazioni
- **u.genre** → Una lista dei generi possibili dei film
- **u.occupation** → Una lista delle occupazioni possibili degli utenti

1.3 File training e test

Nella cartella possiamo anche trovare altri file .base e .test, i primi sono file di training se mentre i secondi di test set.

I file da u1 a u5 sono suddivisioni 80%|20% dei dati di u.data

Ognuno di u1, ..., u5 ha insiemi di test disgiunti, questo se per 5 volte la convalida incrociata (dove si ripete l'esperimento con ogni set di training e di test e la media dei risultati).

Invece i file ua e ub un training set e un test set con esattamente dieci valutazioni per utente nel set di test. I set ua.test e ub.test sono disgiunti.

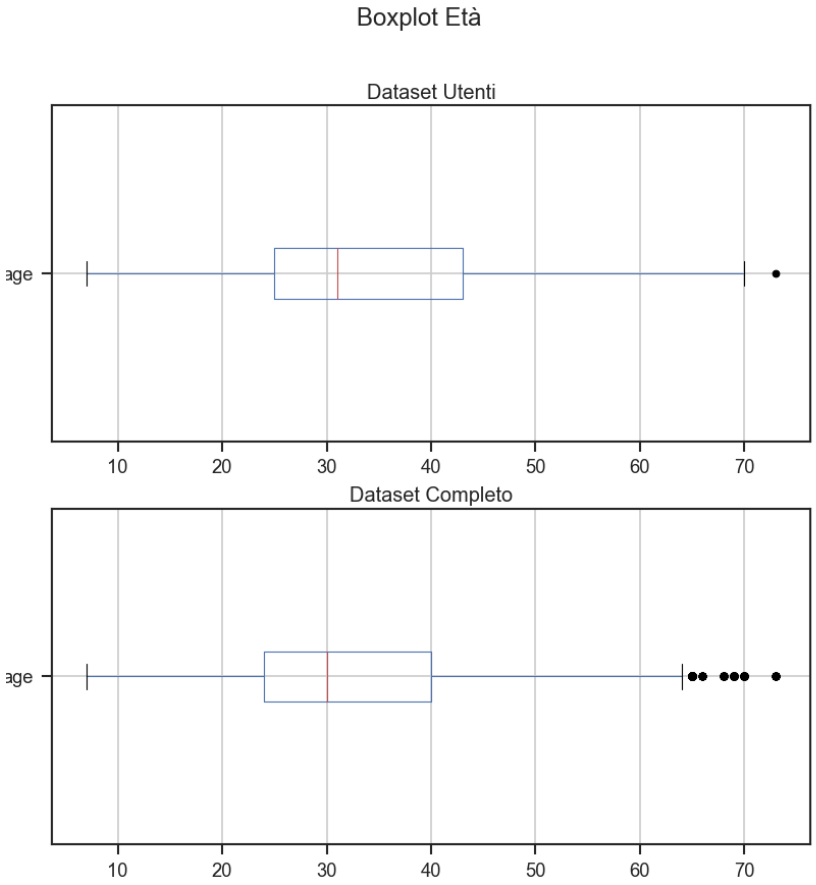
Il file mku.sh è uno script di shell per generare tutti i dataset u da u.data.

2 Analisi esplorativa

2.1 Utenti

Come prima analisi controllo le caratteristiche degli utenti e come queste possano influenzare i rating.

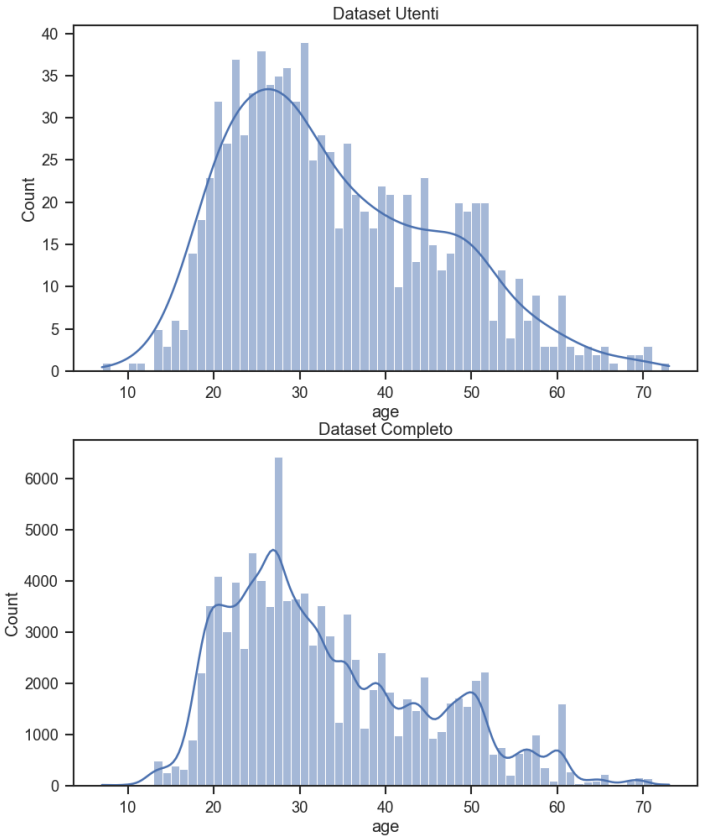
2.1.1 Età

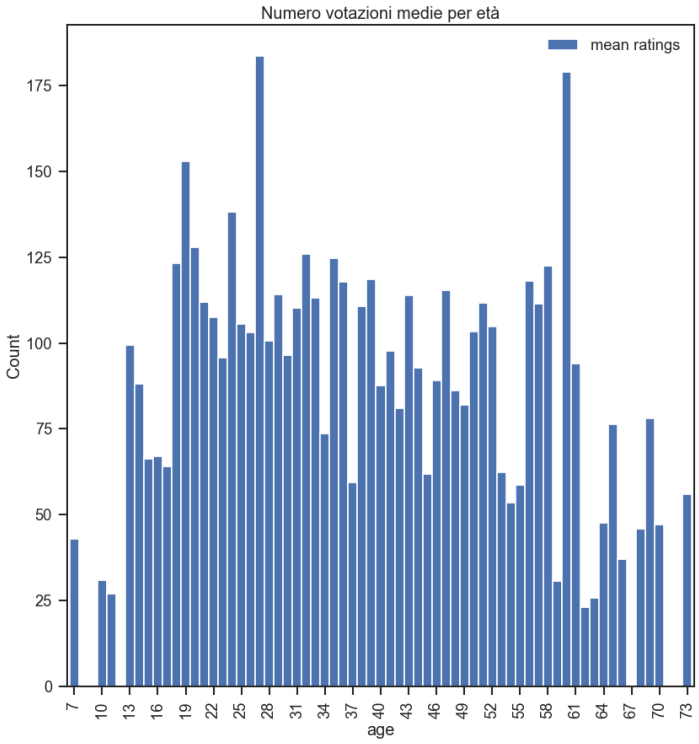


I grafici a qui sopra sono due boxplot, il primo sull'età degli utenti iscritti, il secondo invece sull'età degli utenti votanti, abbiamo quindi ripetizione di utenti che comportano così più valori anomali. Notiamo come nel grafico sottostante, il conteggio medio delle votazioni ha diverse classi modali, riportando una soglia alta non solo nei ragazzi di più o meno 25 anni ma anche nelle persone circa sessantenni, quindi alzando il numero di outliers.

L'età media delle persone iscritte è di circa 34 anni mentre quella degli utenti votanti è di quasi 33 anni, gli utenti più giovani sono quindi un po' più propensi a valutare i film.

Abbiamo molta variabilità per quanto riguarda gli anni delle persone, infatti, la deviazione standard equivale a circa 12.

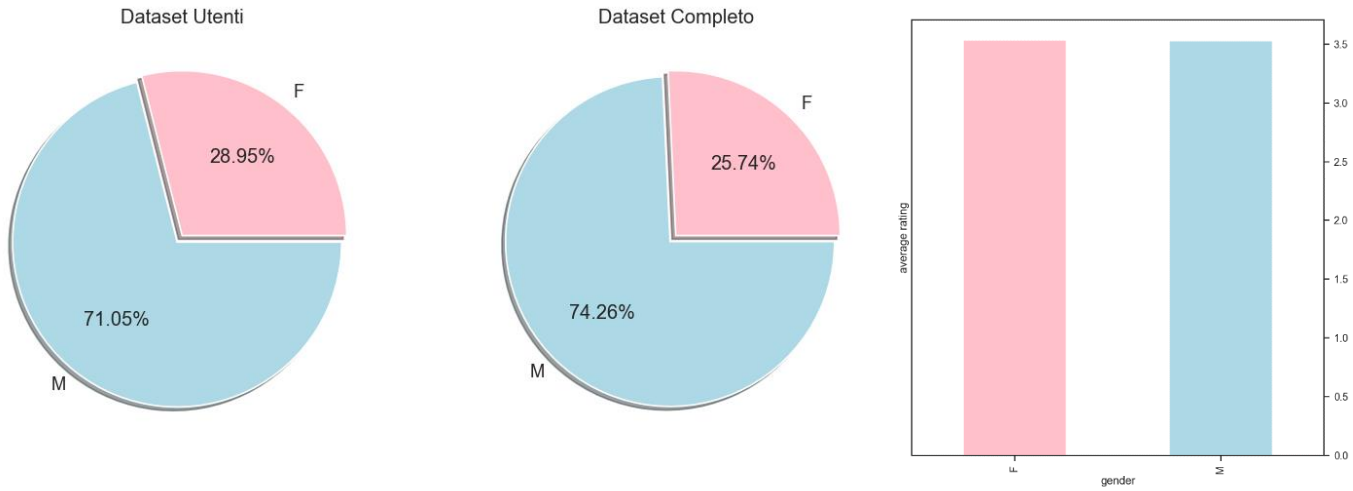




Dagli istogrammi possiamo rafforzare quello che abbiamo già notato precedentemente coi boxplot, notiamo come gran parte delle persone abbia tra i 20 ed i 30 anni e proprio questi sono gli utenti che hanno portato anche più valutazioni, gli utenti di 27 anni sono quelli che hanno portato più valutazioni rispetto a tutti gli altri con ben 6423 valutazioni.

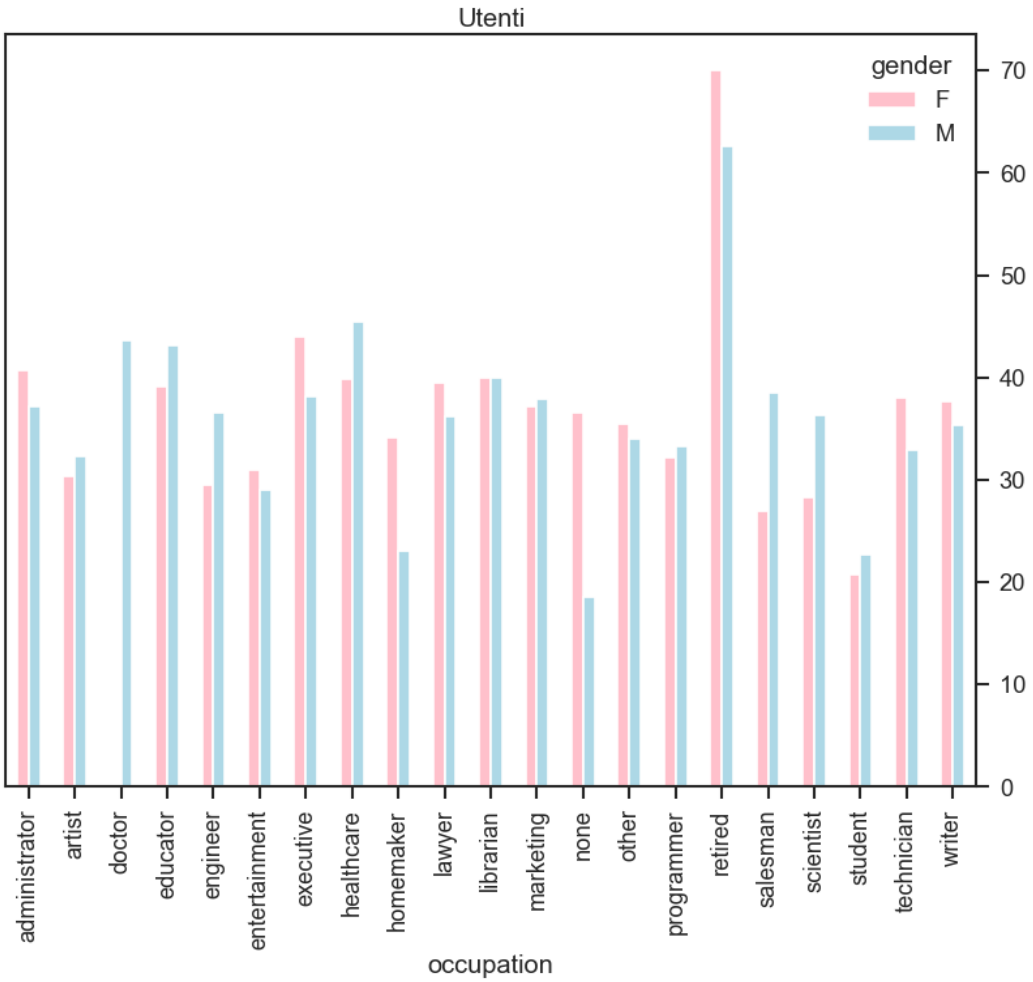
Nel barplot a destra notiamo la media delle votazioni fatte per gruppo di utenti della stessa età e notiamo anche qui gli utenti di 27 sono quelli che mediamente votano più film con una media di circa 183.5 film valuati per utente, anche i 62enni sono mediamente propensi a votare molti film.

2.1.2 Genere



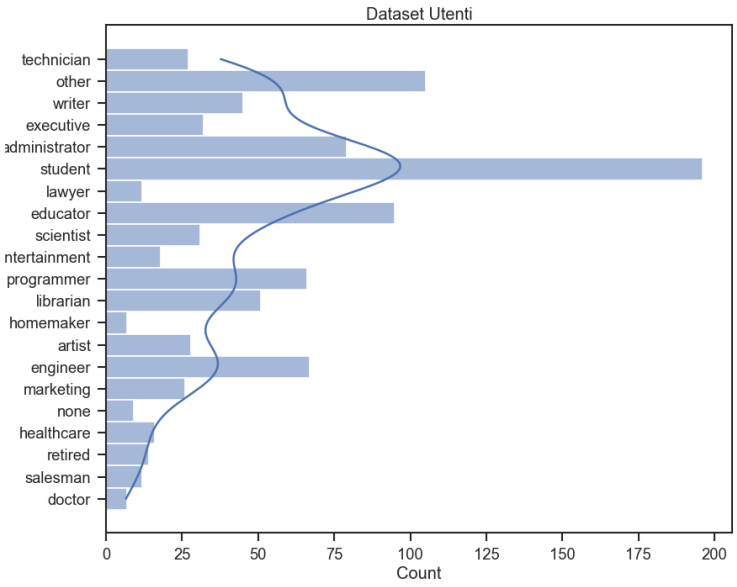
Sappiamo anche esserci una notevole differenza tra il numero di utenti maschi e femmine, come vediamo nel grafico dai grafici a torta, si può osservare anche che gli utenti sono più propensi a valutare i film rispetto alle femmine, nonostante questo, il rating medio per i film è molto simile tra di loro.

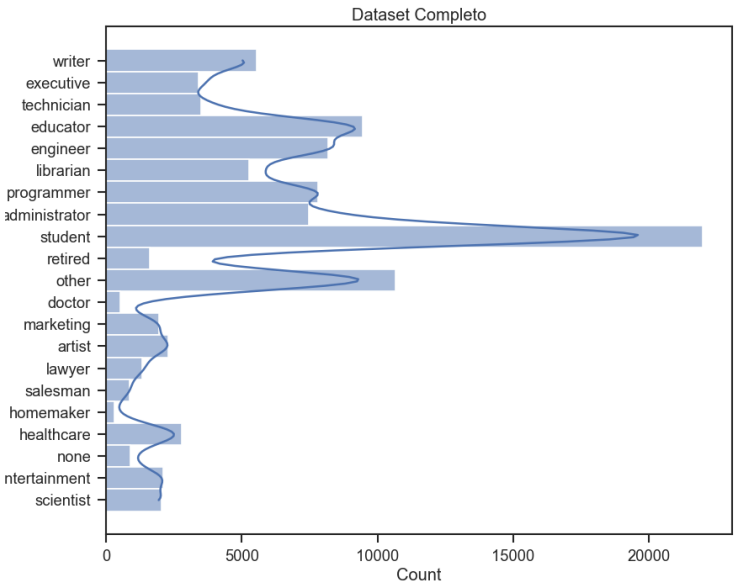
2.1.3 Occupazioni

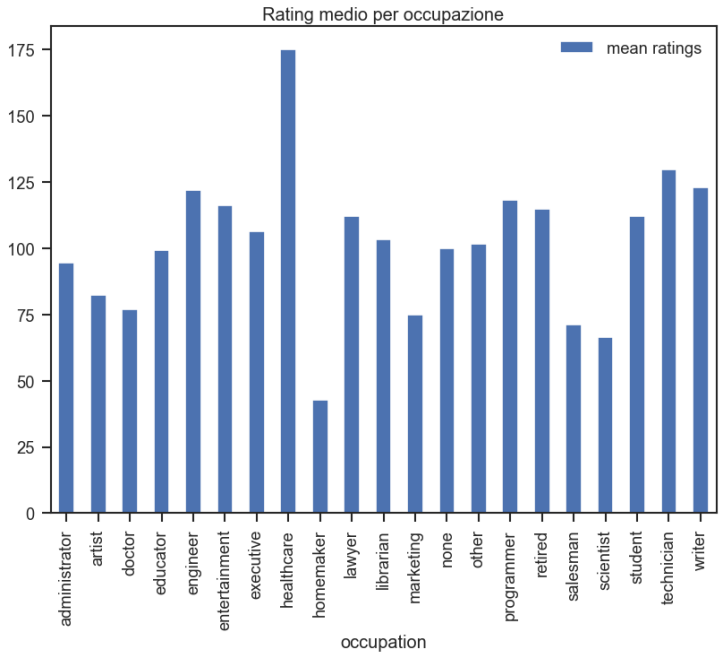


Il grafico soprastante rappresenta l'età media di ogni gruppo di utenti suddiviso per sesso e in base all' occupazione che hanno.

Possiamo notare che per la stessa occupazione l'età media delle donne e degli uomini è quasi sempre molto simile, per quanto abbiamo una forte maggioranza di uomini tra gli utenti iscritti possiamo comunque dire che i due sessi sono distribuiti per età e per occupazione in modo molto simile tra loro.



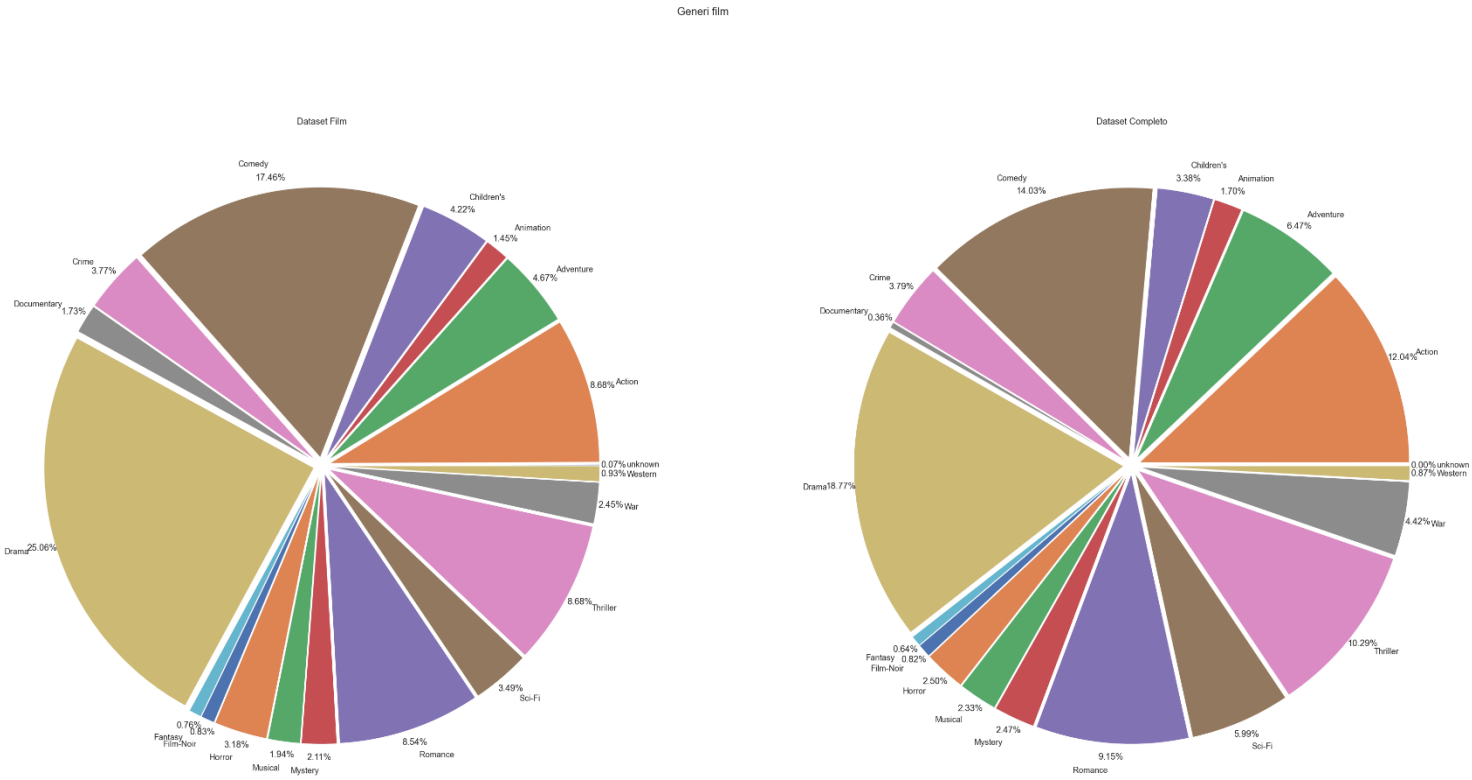




Dagli istogrammi possiamo notare come gli studenti siano i più presenti nella nostra popolazione, infatti, sono anche quelli che hanno portato maggiori votazioni, nonostante ciò le persone che sono mediamente più propense a votare sono quelle presenti nel mondo dell'assistenza sanitaria con una media di circa 175 film valutati per ogni utente.

2.2 Film

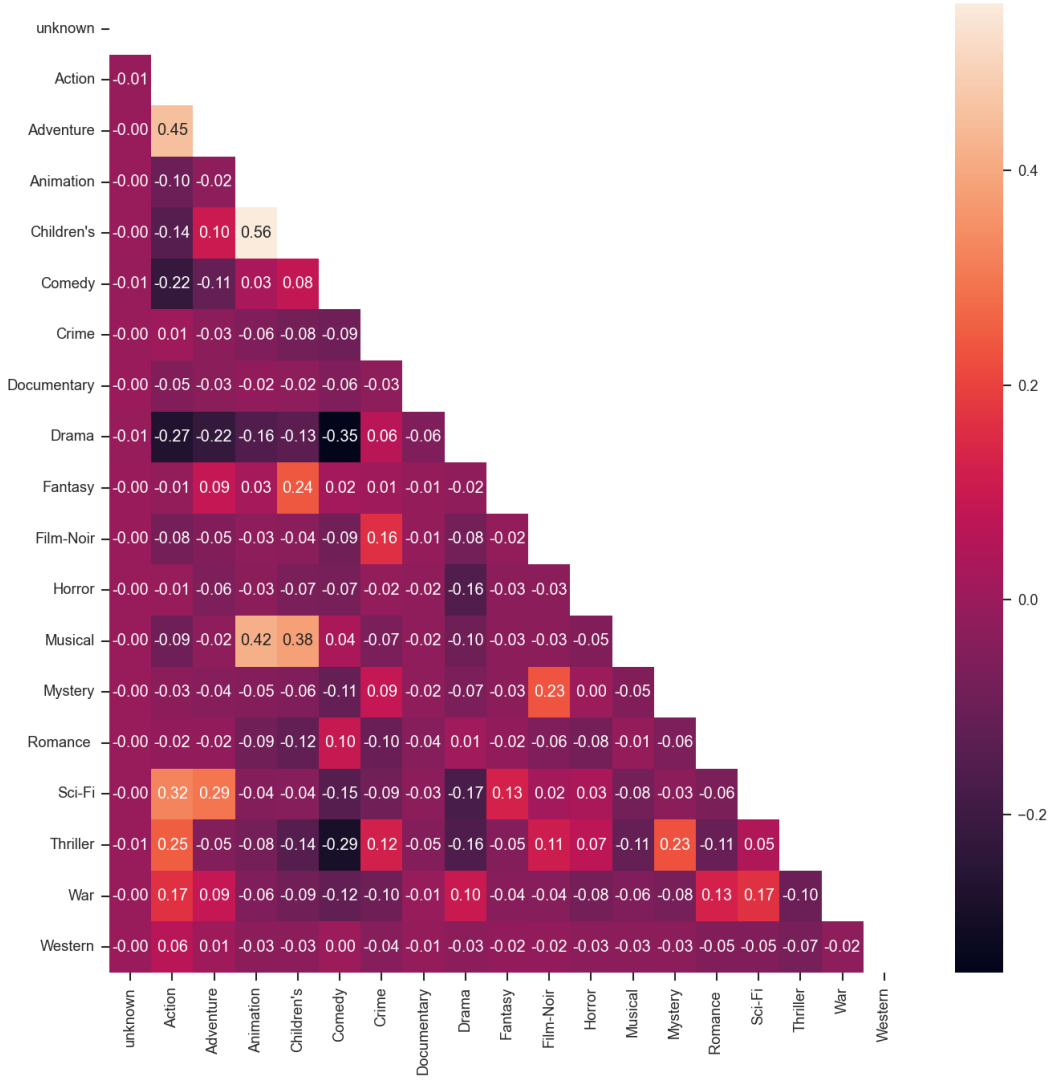
2.2.1 Generi



Questi due grafici a torta rappresentano il primo la distribuzione dei generi dei film presenti nel catalogo mentre quello alla sua destra la rispettiva presenza di questi tra le valutazioni.

Tra tutti possiamo notare una forte presenza di film drammatici nel nostro catalogo (25,06%) che però diventa molto meno presente tra i film valutati (18,77%), possiamo anche osservare un'evidente crescita dei film Action tra i film valutati dove rappresenta il 12,04% di questi mentre nel catalogo era solo presente per l'8,68%. Queste differenze possono essere dovute dal fatto che la maggior parte degli utenti sono maschi e tendenzialmente preferiscono film di genere più movimentato.

2.2.2 Correlazioni tra generi



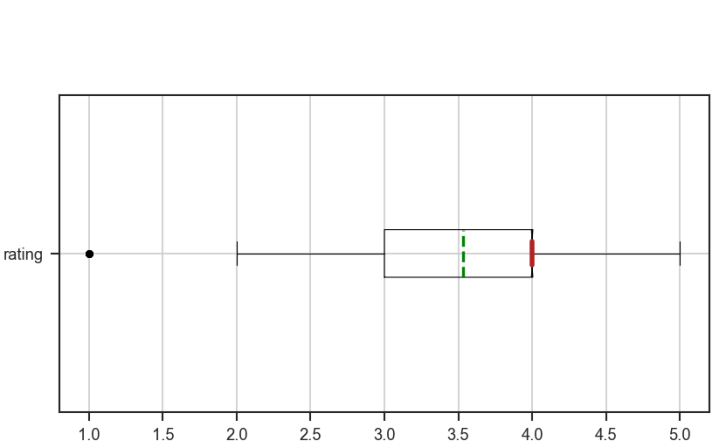
Dalla heatmap presente qui sopra possiamo andare ad analizzare le eventuali correlazioni tra i diversi generi di film.

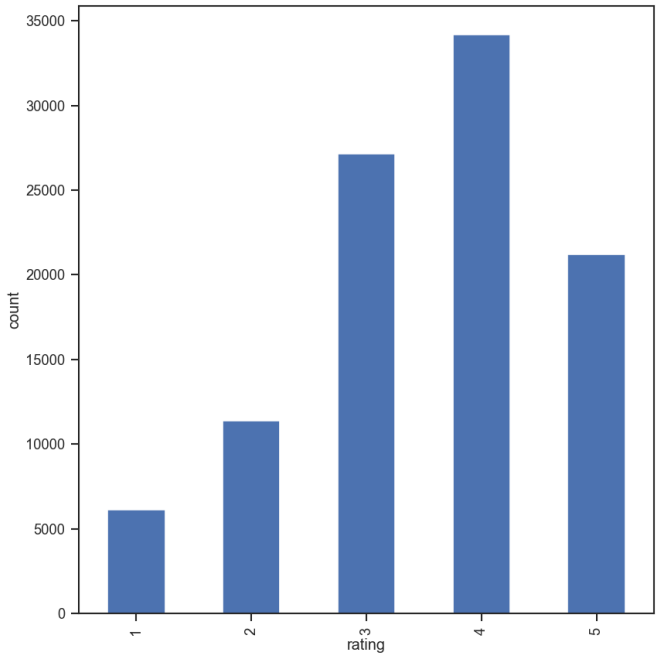
Notiamo che quasi tutte sono poco significative, questo è dovuto soprattutto dal fatto che la nostra matrice di rating come analizzeremo successivamente è molto sparsa, se avessimo una matrice più densa probabilmente avremmo avuto dei valori di correlazione maggiori.

Le correlazioni più alte che possiamo trovare sono:

- Children's – Animation → 0.56
- Adventure – Action → 0.45
- Musical – Animation → 0.42
- Musical – Children's → 0.38

2.3 Rating

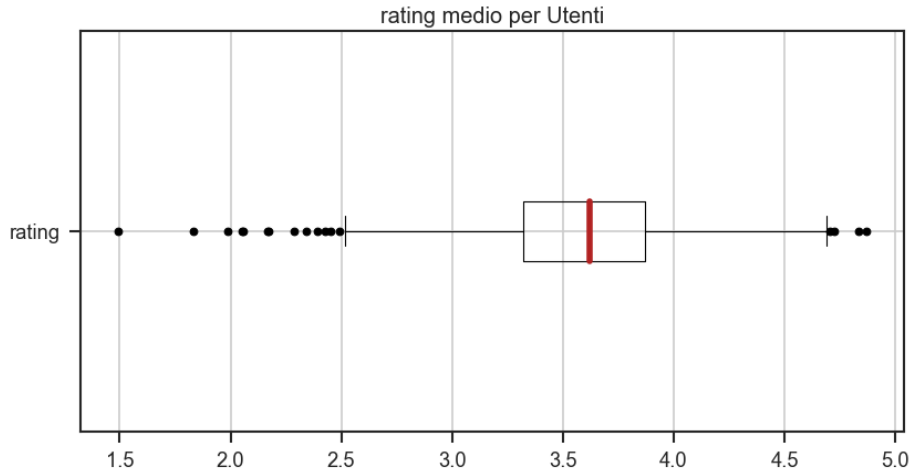


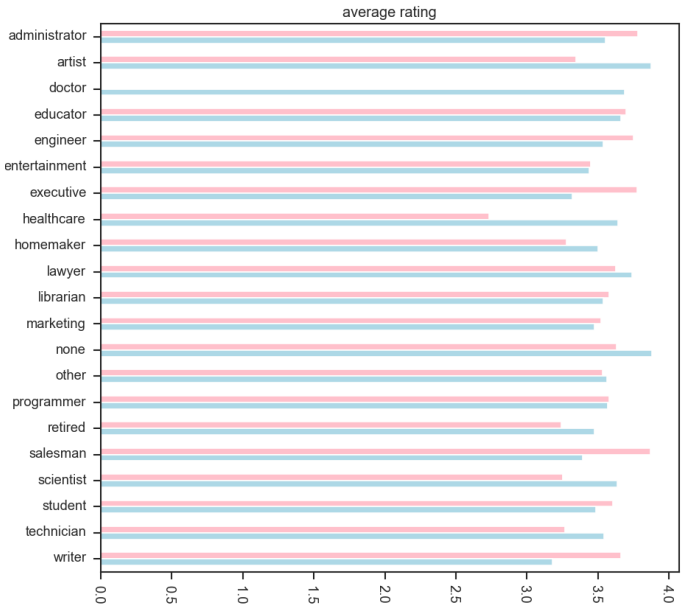


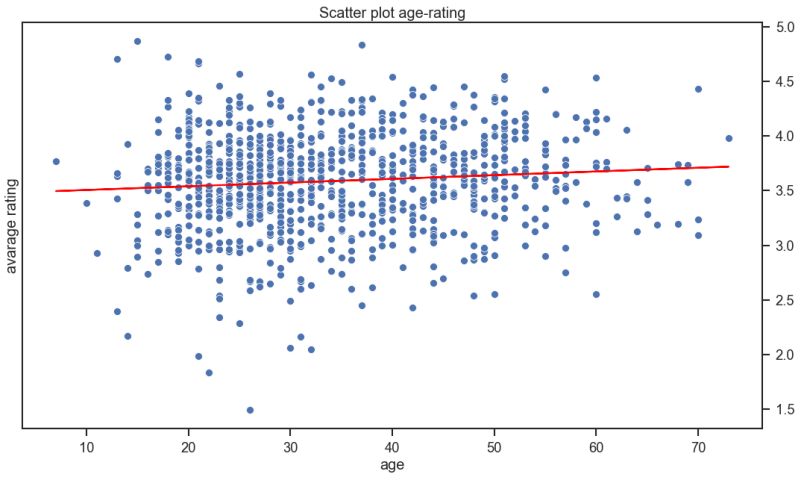
Con questo boxplot rappresenta la distribuzione dei rating. Ogni utente può valutare un film con un valore intero da 1 a 5.

Si può notare che i film sono spesso valutati positivamente, la mediana (rappresentata con una linea rossa continua) infatti si trova sul valore 4, è anche il voto più assegnato e la media (rappresentata con una linea verde tratteggiata) dei rating presenti è di 3,53. Il voto 1 è talmente poco assegnato che nel boxplot è addirittura un outlier.

2.3.1 Rating rispetto agli utenti





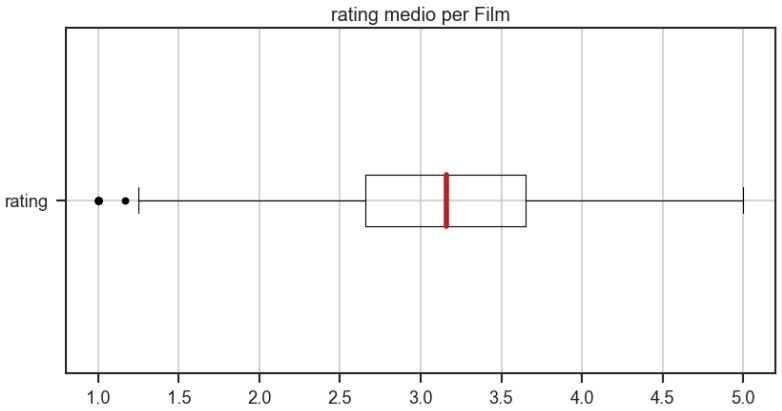


Il boxplot è sulla media delle valutazioni di ogni utente, da questo possiamo osservare come la maggior parte degli utenti ha una media tra 3,32 e 3,87 e gli utenti con una media minore di 2,5 sono addirittura degli outliers, possiamo quindi dire che gli utenti iscritti apprezzano mediamente molto i film presenti nel catalogo.

Dal grafico a barre possiamo osservare la media voto di ogni categoria di utente e, come avevamo già considerato precedentemente col boxplot, possiamo dire che gli utenti hanno una media molto simile tra loro indipendentemente dal loro sesso o dalla loro occupazione.

Dallo scatterplot tra età e rating medio per utente possiamo notare che non c'è nessuna correlazione significativa tra età e rating (0.42), infatti abbiamo una linea di regressione lontana dalla bisettrice.

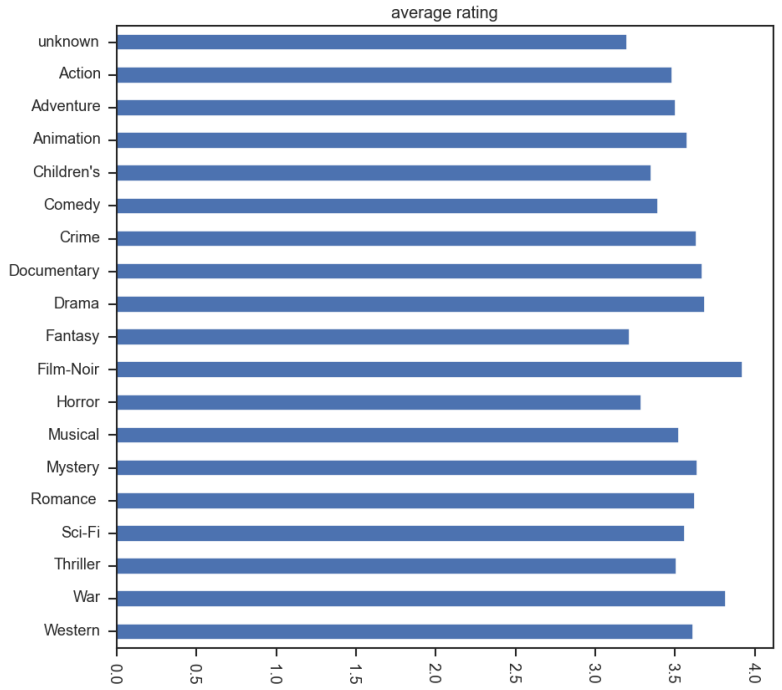
2.3.2 Rating rispetto ai film



Dal boxplot sui rating medi per ogni film possiamo notare che nel nostro catalogo ci sono film con ogni tipo di apprezzamento, infatti troviamo solo due anomalie.

La tendenza però è quella di avere film spesso apprezzati, infatti la maggior parte dei film ha una votazione media tra 2,66 e 3,65 e gli unici due outliers sono dovuti a film valutati "negativamente".

Il grafico a barre qui accanto rappresenta la media delle valutazioni raccolte per ogni genere di film.



Si può osservare che abbiamo delle medie molto simili tra loro, possiamo quindi dire che il genere a cui appartiene un film non influisce più di tanto sulla sua possibile valutazione.

Notiamo comunque che:

- Il genere Film-Noir è mediamente più apprezzato rispetto agli altri generi con una media di 3.92
- Il genere Sconosciuto è mediamente meno apprezzato rispetto agli altri generi con una media di 3.2

3 Recommender system

3.1 Matrice di rating

Creo una matrice di rating dove le righe rappresentano gli utenti mentre le colonne i film, ogni cella contiene la valutazione che l'utente ha dato ad un certo film se presente.

La matrice si presenta così:

movie id	1	2	3	4	5	6	7	8	9	10	...	1673	1674	1675	1676	1677	1678	1679	1680	1681	1682
user id																					
1	5	3	4	3	3	5	4	1	5	3	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	4	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
939	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
940	NaN	NaN	NaN	2	NaN	NaN	4	5	3	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
941	5	NaN	NaN	NaN	NaN	NaN	4	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
942	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
943	NaN	5	NaN	NaN	NaN	NaN	NaN	NaN	3	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

943 rows × 1682 columns

Notiamo fin da subito che la nostra matrice è molto sparsa infatti:

- La nostra matrice ha 1586126 celle
- I rating a nostra disposizione sono: 100000
- I rating mancanti nella nostra matrice sono: 1486126
- I rating presenti sono il 6.3% della nostra matrice
- I rating mancanti sono il 93.7% della nostra matrice

Per riempirla useremo l'algoritmo K-NN

3.2 K-NN

I metodi K-NN Sono basati sul predire il rating di un utente per un item utilizzando il concetto di vicino (o neighbourhood) similarity.

Se vogliamo prevedere il rating di un certo per il nostro target user andremo a considerare un insieme di **k** users con l'indice di similarità (cosine o pearson) più alto per il nostro target user, questi k user sono definiti come neighborhood e una volta definito si può procedere a calcolare il rating predetto per tale item come media dei rating appartenenti al neighborhood relativo.

È importante scegliere bene il valore di **k** (ovvero quanti "vicini" vogliamo considerare), se questo è troppo grande, oltre ad aumentare i tempi di calcolo, rischiamo di avere del "rumore" nella predizione per via di molti vicini con limitata similarità, viceversa se **k** è troppo piccolo la predizione può essere poco accurata. Il numero di vicini **k** è da considerare un iper-parametro.

$$\widehat{r_{ij}} = \frac{\sum_{l \in N_j^k(i)} sim\,(i,l) \cdot r_{ij}}{\sum_{l \in N_j^k(i)} |sim(i,l)|}$$

Questi algoritmi sono basati sul fatto che utenti "simili" mostrano dei pattern "simili" nel fare rating e che item simili ricevono dei rating simili. Abbiamo due metodi per k-NN:

- **user-based** → i rating vengono calcolati utilizzando i rating degli users vicini, questi vengono definiti attraverso una misura di similarità tra gli users (le righe della matrice di rating)
- **item-based** → i rating vengono calcolati utilizzando i rating dati da uno user sugli items vicini, questi vengono definiti attraverso una misura di similarità tra gli items (le colonne della matrice di rating)

Nel nostro caso effettueremo delle previsioni user-based

3.3 Indici di similarità

3.3.1 Cosine Similarity

Dati due vettori $\vec{a}$ e $\vec{b}$ in uno spazio n-dimensionale R^n definiamo la similitudine del coseno (cosine similarity), come:

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| \cdot ||\vec{b}||} = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}}$$

La cosine similarity può essere definita anche nel caso in cui non siano note alcune componenti dei due vettori. Per farlo si considera l'insieme I_{ab} come l'insieme degli indici delle componenti note sia di $\vec{a}$ che di $\vec{b}$ e si definisce la cosine similarity come:

$$sim(\vec{a}, \vec{b}) = \frac{\sum_{i \in I_{ab}} a_i \cdot b_i}{\sqrt{\sum_{i \in I_{ab}} a_i^2} \cdot \sqrt{\sum_{i \in I_{ab}} b_i^2}}$$

La cosine similarity rappresenta una misura di somiglianza tra 2 vettori. Geometricamente corrisponde al coseno dell'angolo tra i due vettori.

Nel caso in cui le componenti dei vettori siano non-negative (come nel caso dei ratings), la cosine similarity può variare tra 0 (completa diversità) a +1 (massima somiglianza). È invariante rispetto alle dilatazioni

3.3.2 Pearson Correlation

Un' altra misura di similarità è data dall'indice di correlazione di Pearson. Dati due vettori $\vec{a}$ e $\vec{b}$ in uno spazio n-dimensionale R^n definiamo indice di correlazione di Pearson come:

$$sim(\vec{a}, \vec{b}) = \frac{< \vec{a} - \bar{a}, \vec{b} - \bar{b} >}{||\vec{a} - \bar{a}|| \cdot ||\vec{b} - \bar{b}||} = \frac{\sum_{i=1}^{n} (a_i - \bar{a}) \cdot (b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n} (a_i - \bar{a})^2} \cdot \sqrt{\sum_{i=1}^{n} (b_i - \bar{b})^2}}$$

Dove:

- $\bar{a} = |I_a|^{-1} \sum_{i \in I_a} a_i$ e $\bar{b} = |I_b|^{-1} \sum_{i \in I_b} b_i$ rappresentano le medie dei due vettori
- $|I_a|$ e $|I_b|$ indicano rispettivamente le cardinalità degli insiemi I_a e I_b

L'indice di correlazione di Pearson può essere definito anche nel caso in cui non siano note alcune componenti dei due vettori. Per farlo si considera:

- L'insieme I_{ab} come l'insieme degli indici delle componenti note sia di $\vec{a}$ che di $\vec{b}$
- L'insieme $I_a$come l'insieme degli indici delle componenti note di $\vec{a}$
- L'insieme I_b come l'insieme degli indici delle componenti note di $\vec{b}$

Si definisce l'indice di correlazione di Pearson come:

- $\bar{a} = |I_a|^{-1} \sum_{i \in I_a} a_i$
- $\bar{b} = |I_b|^{-1} \sum_{i \in I_b} b_i$

$$sim(\vec{a}, \vec{b}) = \frac{< \vec{a} - \bar{a}, \vec{b} - \bar{b} >}{\left|\left|\vec{a} - \bar{a}\right|\right| \cdot \left|\left|\vec{b} - \bar{b}\right|\right|} = \frac{\sum_{i \in I_{ab}} (a_i - \bar{a}) \cdot (b_i - \bar{b})}{\sqrt{\sum_{i \in I_{ab}} (a_i - \bar{a})^2} \cdot \sqrt{\sum_{i \in I_{ab}} (b_i - \bar{b})^2}}$$

L'indice di correlazione è dato dal rapporto tra covarianza e prodotto delle deviazioni standard dei 2 campioni a e b può variare tra -1 (perfetta correlazione lineare negativa) a +1 (perfetta correlazione lineare positiva) dove il valore 0 significa che I due vettori non sono correlati linearmente. (non che siano indipendenti)

3.4 Accuratezza della previsione

Consideriamo una generica matrice di rating R con n user e m item. Indichiamo con r_{ij} il rating espresso dall'utente i e per l'item j. Indichiamo con S l'insieme di tutte le entrate conosciute della matrice di rating con $|S| \ll m \cdot n$. Sia $\widetilde{r_{ij}}$ il rating predetto da un generico algoritmo di raccomandazione. Chiamiamo entry-specific error la quantità $e_{i,j} = r_{ij} - \widetilde{r_{ij}}$

3.4.1 RMSE (Root Mean Square Error)

$$\sqrt{\frac{\sum_{(i,j) \in S} e_{i,j}^2}{|S|}}$$

È la somma degli errori quadratici risulta quindi più sensibile alla presenza di outliers. Risulta migliore da usare nei casi in cui sia importante la robustezza della predizione.

3.4.2 MSE (Mean Square Error)

$$\frac{\sum_{(i,j) \in S} e_{i,j}^2}{|S|}$$

3.4.3 MAE (Mean Absolute Error)

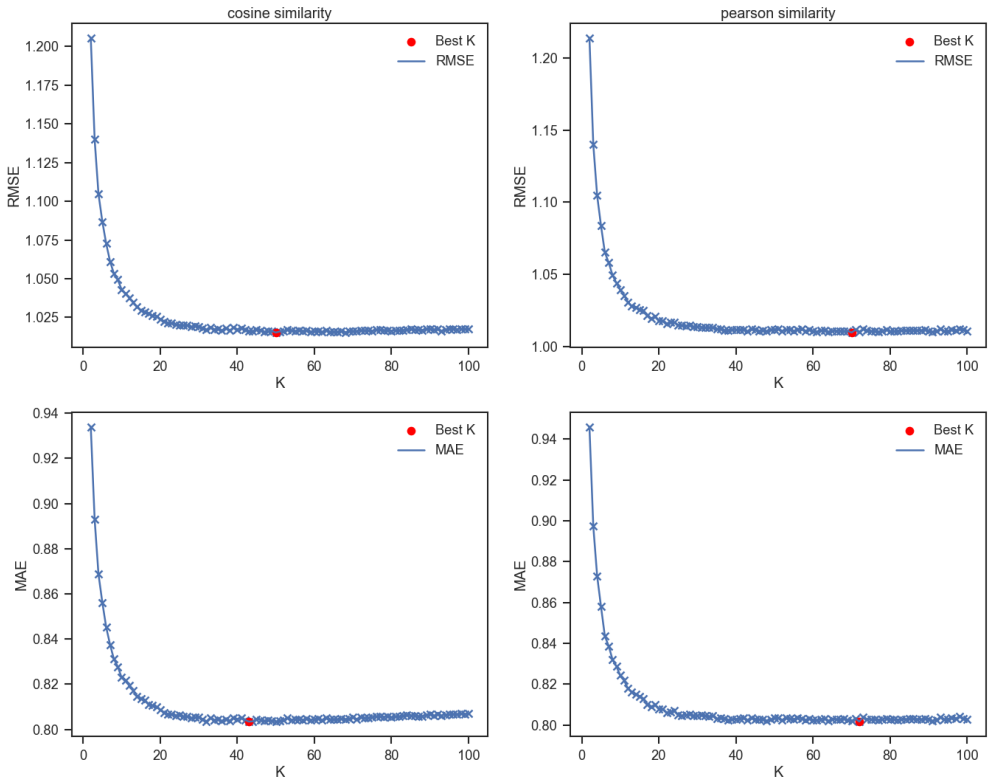
$$\frac{\sum_{(i,j) \in S} |e_{i,j}|}{|S|}$$

Risulta essere una misura meglio rappresentativa dell'accuratezza quando l'importanza degli outlier nella valutazione è limitata.

3.5 Matrix filling with K-NN

3.5.1 Ottimizzazione dell'iperparametro K

Prima di riempire la matrice di rating ho cercato quale fosse il miglior valore di K da usare, per fare ciò ho controllato quale valore k minimizzasse le misure di accuratezza presentate precedentemente



Trovo che con la Cosine Similarity trovo come migliori valori K

- K = 50 → RMSE: 1.015 MSE: 1.031
- K = 43 → MAE: 0.803

Invece con la Pearson Correlation trovo come migliori valori K

- K = 70 → RMSE: 1.009 MSE: 1.019
- K = 72 → MAE: 0.801

Userò i migliori valori di K trovati minimizzando RMSE per fare le previsioni dei rating dato che questa misura di accuratezza risulta migliore da usare nei casi in cui sia importante la robustezza della predizione.

3.5.2 Cosine Similarity

Matrice di rating riempita con la Cosine Similarity

Similarity: cosine K = 50

movie id / user id	1	2	3	4	5	6	7	8	9	10	...	1673	1674	1675	1676	1677	1678	1679	1680	1681	1682
1	5	3	4	3	3	5	4	1	5	3	...	3	4	3	2	3	1	3	2	3	3
2	4	3	3	4	3	3.5	4	4	4	2	...	3	4	3	2	3	1	3	2	3	3
3	4	3	3	3.5	3	3.5	4	4	4	4	...	3	4	3	2	3	1	3	2	3	3
4	4	3	3	4	3	3.5	4	4	4	4	...	3	4	3	2	3	1	3	2	3	3
5	4	3	3	4	3	3.5	4	4	4	4	...	3	4	3	2	3	1	3	2	3	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
939	4	3	3	4	3.5	3.5	4	4	5	4	...	3	4	3	2	3	1	3	2	3	3
940	4	3.5	3	2	3.5	3.5	4	5	3	4	...	3	4	3	2	3	1	3	2	3	3
941	5	3	3	3.5	3.5	3.5	4	4	4	4	...	3	4	3	2	3	1	3	2	3	3
942	4	3.5	3	4	3.5	3.5	4	4	4	4	...	3	4	3	2	3	1	3	2	3	3
943	4	5	3	4	3.5	3.5	4	4	3	4	...	3	4	3	2	3	3.5	3.5	3.5	3	3

943 rows × 1682 columns

3.5.3 Pearson Correlation

Matrice di rating riempita con la Pearson Correlation

Similarity: pearson K = 70

movie id / user id	1	2	3	4	5	6	7	8	9	10	...	1673	1674	1675	1676	1677	1678	1679	1680	1681	1682
1	5	3	4	3	3	5	4	1	5	3	...	4.0	4	3	2	3	1	3	2	3	3
2	4	3	3	4	3	4	4	4	4	2	...	3	4.0	3	2	3	4.0	4.0	4.0	3	3
3	4	3	3	3.5	3.5	3	4	4	4	4	...	3	3.0	3	2	3.0	3.0	3.0	3.0	3	3.0
4	4	3	3	3.5	3	4	4	4	3.5	4	...	4.0	4	3	2	3	4.0	4.0	4.0	3	3
5	4	3	3	4	3	4	4	4	4	4	...	3.0	4	3	2	3	3.0	3.0	3.0	3	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
939	4	3	3	4	3	3.5	4	4	5	4	...	3	4.0	3	2	4.0	4.0	4.0	4.0	4.0	3
940	4	3	3	2	3	3.5	4	5	3	4	...	3.5	3.5	3	2	3	1	3	2	3	3
941	5	3	3	4	3	4	4	4	4	4	...	4.0	4.0	4.0	4.0	3	4.0	4.0	4.0	3	3
942	4	3	3	4	3	3.5	3.5	4	4	4	...	4.0	4.0	3	2	3	1	3	2	3	3
943	4	5	3	4	3	4	4	4	3	4	...	3.5	3.5	3	2	3	3.5	3.5	3.5	3	3

943 rows × 1682 columns

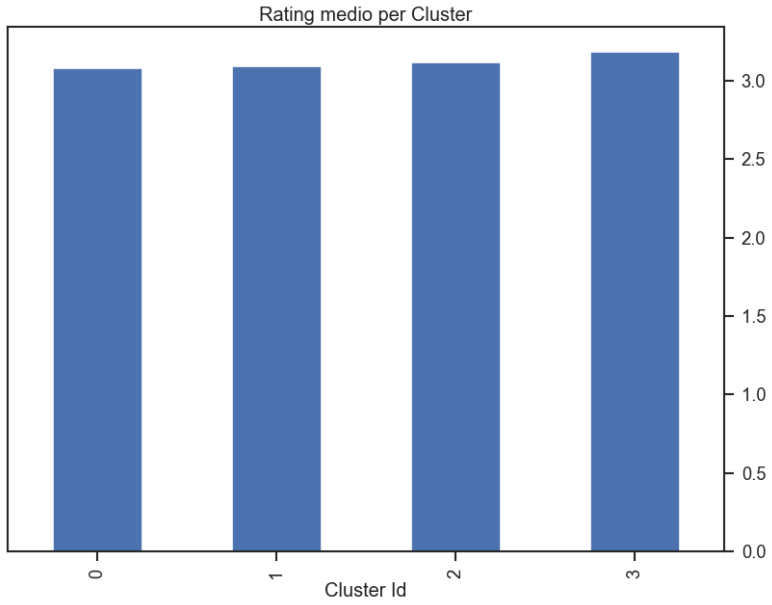
4 Clustering

Il clustering è una tecnica che assiste la segmentazione degli utenti che presentano caratteristiche simili tra loro.

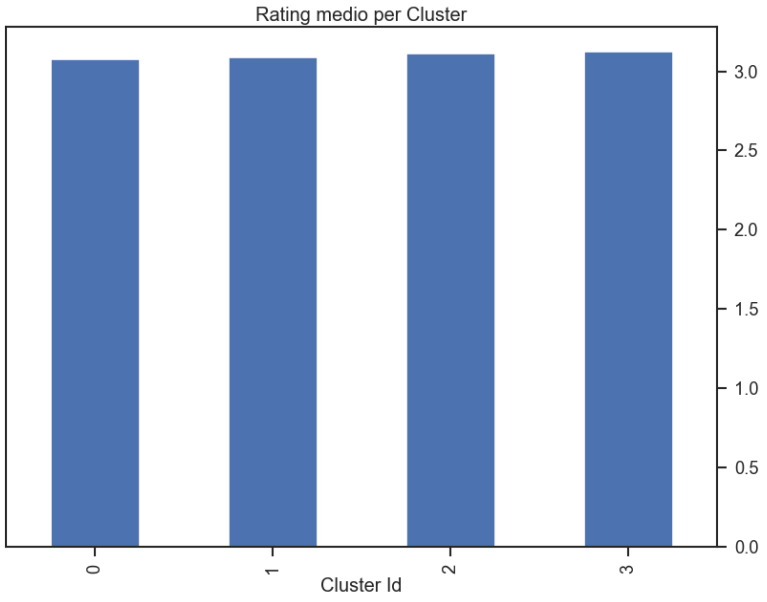
Effettueremo i clustering degli utenti sulla matrice di rating già riempita dal K-NN con la Cosine Similarity, avere la matrice già riempita ci permette di non avere delle distorsioni. Suddivideremo gli utenti in 4 sottogruppi.

4.1 Distanza Euclidea

Per questo primo clustering usiamo come metrica la distanza Euclidea,



4.2 Cosine Distance



5 Conclusioni

L’ obiettivo di questo progetto è stato quello di analizzare i dataset di Movielens 100k e le sue variabili, cercando di avere informazioni sulle loro distribuzioni, sulle loro statistiche e sulla correlazione tra di esse, sono stati analizzati in modo particolare le caratteristiche degli utenti, dei film e come queste impattano sui rating.

Successivamente tramite degli algoritmi di Machine Learning(K-NN) abbiamo riempito la matrice di rating. Il nostro algoritmo ha lavorando sfruttando due indici di similarità degli utenti distinti, la Cosine similarity e l’indice di correlazione di Pearson, per rendere le nostre previsioni più accurate è stato ottimizzato l’iperparametro K per entrambi le similarità

Infine, dopo aver riempito la matrice di rating, gli utenti sono stati suddivisi in cluster, ovvero abbiamo raggruppato gli utenti simili tra di loro.