Metodi Informatici per la Gestione Aziendale

Davide Grandesso 852078

Dati utilizzati

movielens

Non-commercial, personalized movie recommendations.

---------- Film ----------

movie id	movie title	release date	video release date	IMDb URL	unknown	Action
1	Toy Story (1995)	01-Jan-1995	NaN	http://us.imdb.com/M/title-exact?Toy%20Story%2...	0	0
2	GoldenEye (1995)	01-Jan-1995	NaN	http://us.imdb.com/M/title-exact?GoldenEye%20(...	0	1
3	Four Rooms (1995)	01-Jan-1995	NaN	http://us.imdb.com/M/title-exact?Four%20Rooms%...	0	0
4	Get Shorty (1995)	01-Jan-1995	NaN	http://us.imdb.com/M/title-exact?Get%20Shorty%...	0	1
5	Copycat (1995)	01-Jan-1995	NaN	http://us.imdb.com/M/title-exact?Copycat%20(1995)	0	0
...	...	...	...	...	...	...
1678	Mat' i syn (1997)	06-Feb-1998	NaN	http://us.imdb.com/M/title-exact?Mat%27+i+syn+...	0	0
1679	B. Monkey (1998)	06-Feb-1998	NaN	http://us.imdb.com/M/title-exact?B%2E+Monkey+(...	0	0
1680	Sliding Doors (1998)	01-Jan-1998	NaN	http://us.imdb.com/Title?Sliding+Doors+(1998)	0	0
1681	You So Crazy (1994)	01-Jan-1994	NaN	http://us.imdb.com/M/title-exact?You%20So%20Cr...	0	0
1682	Scream of Stone (Schrei aus Stein) (1991)	08-Mar-1996	NaN	http://us.imdb.com/M/title-exact?Schrei%20aus%...	0	0

1682 rows × 23 columns

---------- Ratings ----------

user id	item id	rating	timestamp
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	346	1	886397596
...	...	...	...
880	476	3	880175444
716	204	5	879795543
276	1090	1	874795795
13	225	2	882399156
12	203	3	879959583

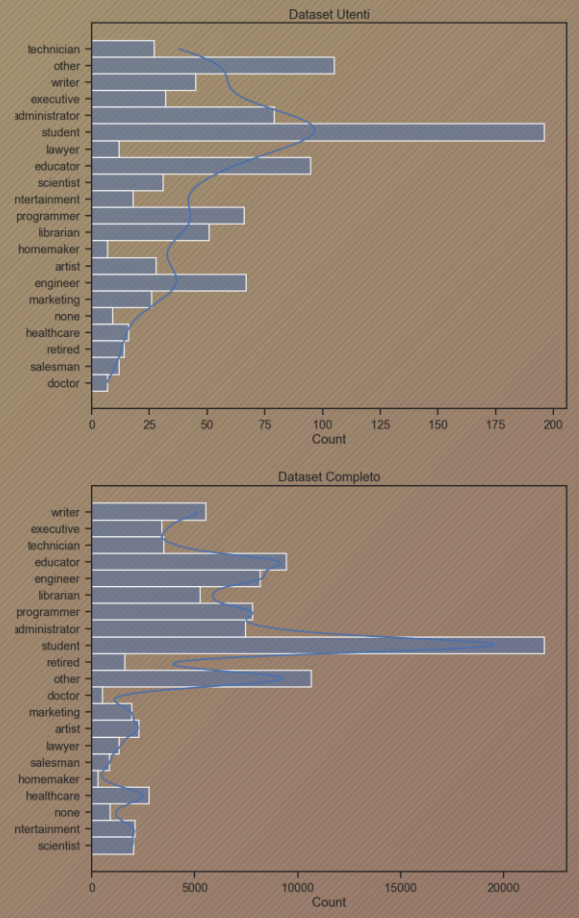
100000 rows × 2 columns

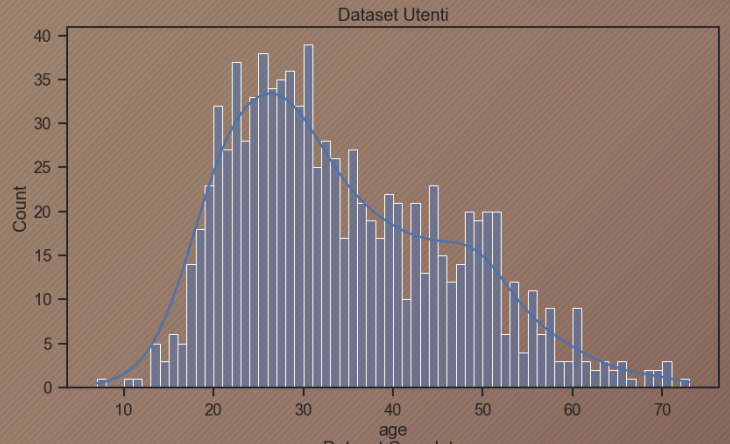
---------- Utenti ----------

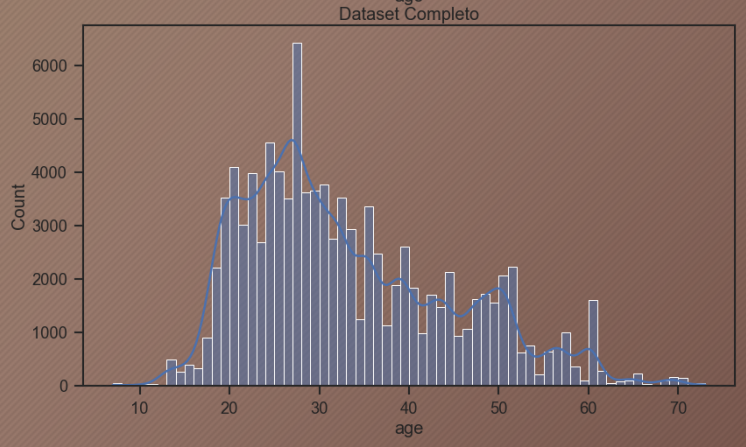
user id	age	gender	occupation	zip code
1	24	M	technician	85711
2	53	F	other	94043
3	23	M	writer	32067
4	24	M	technician	43537
5	33	F	other	15213
...	...	...	...	...
939	26	F	student	33319
940	32	M	administrator	02215
941	20	M	student	97229
942	48	F	librarian	78209
943	22	M	student	77841

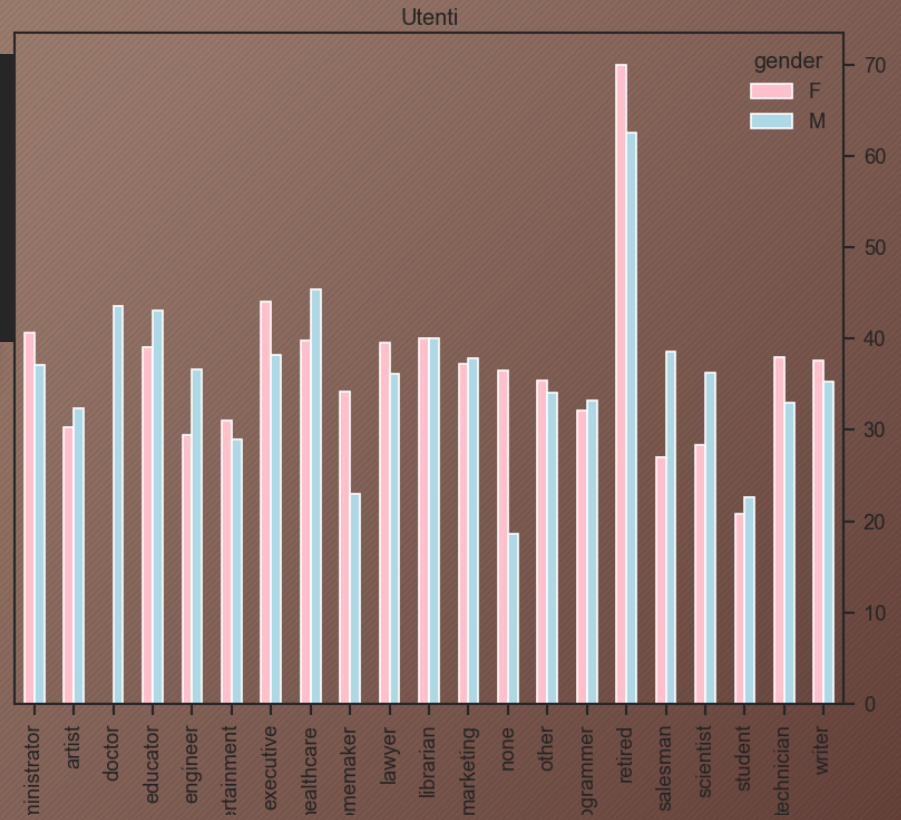
943 rows × 4 columns

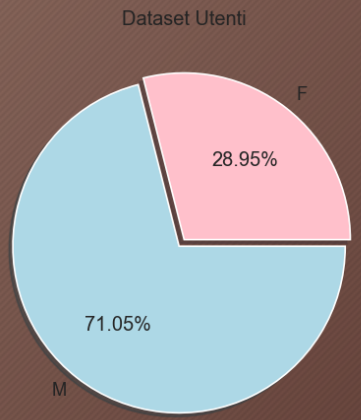
Analisi esplorativa

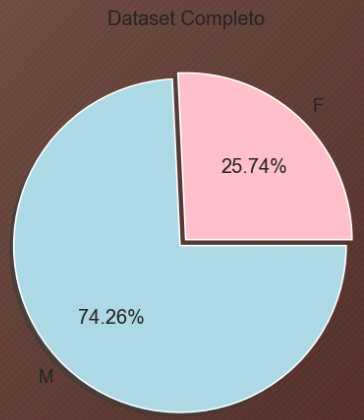


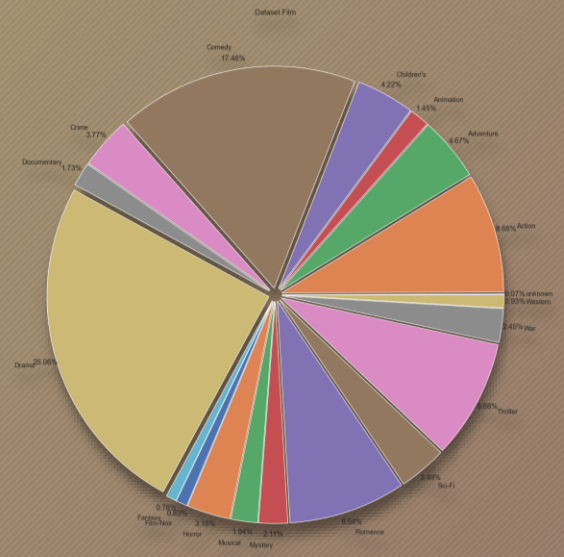


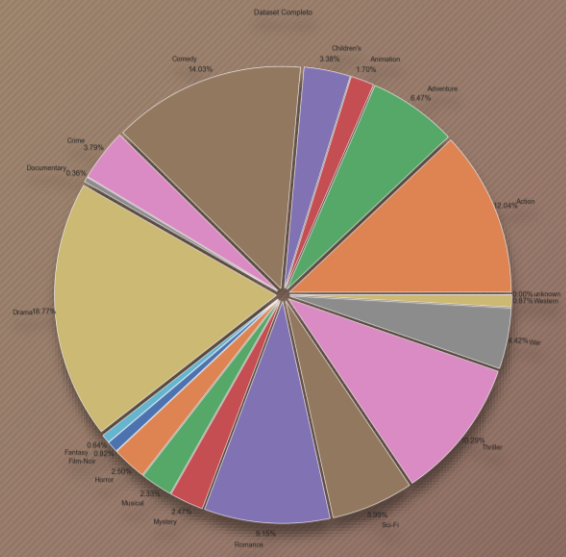




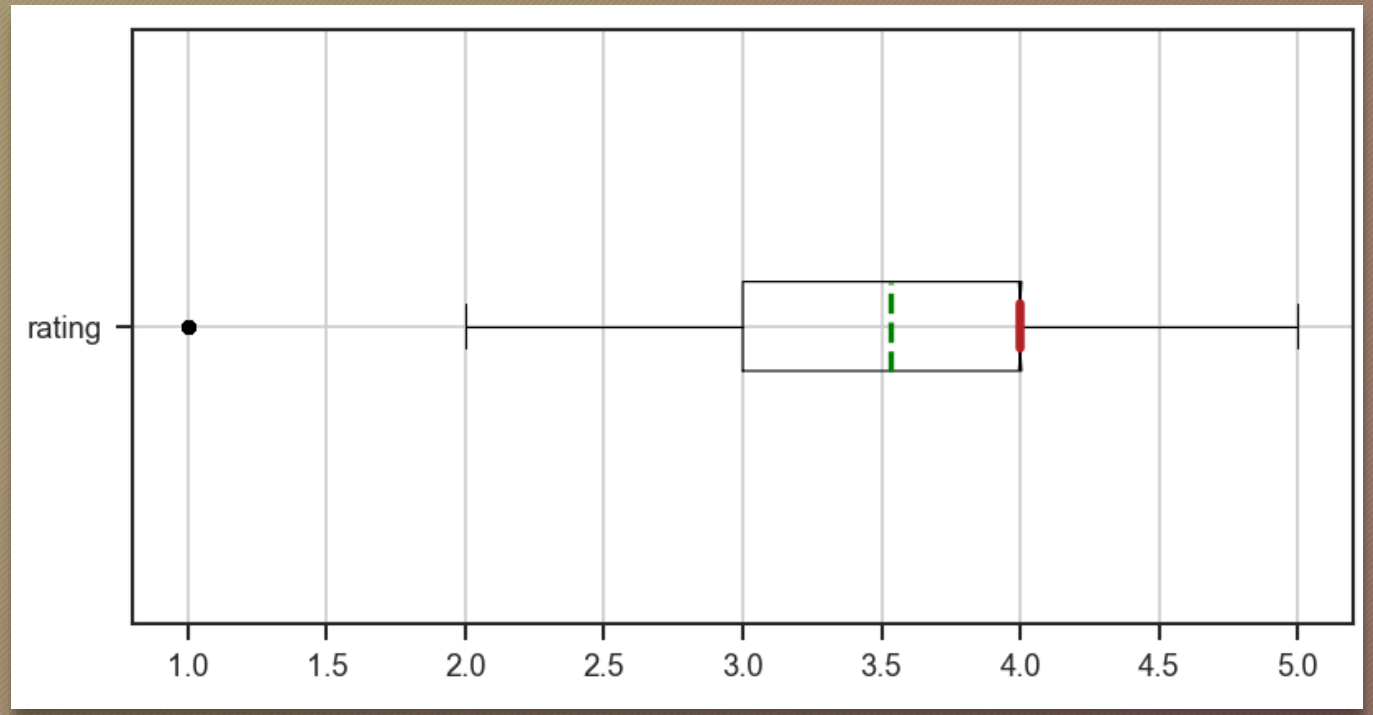


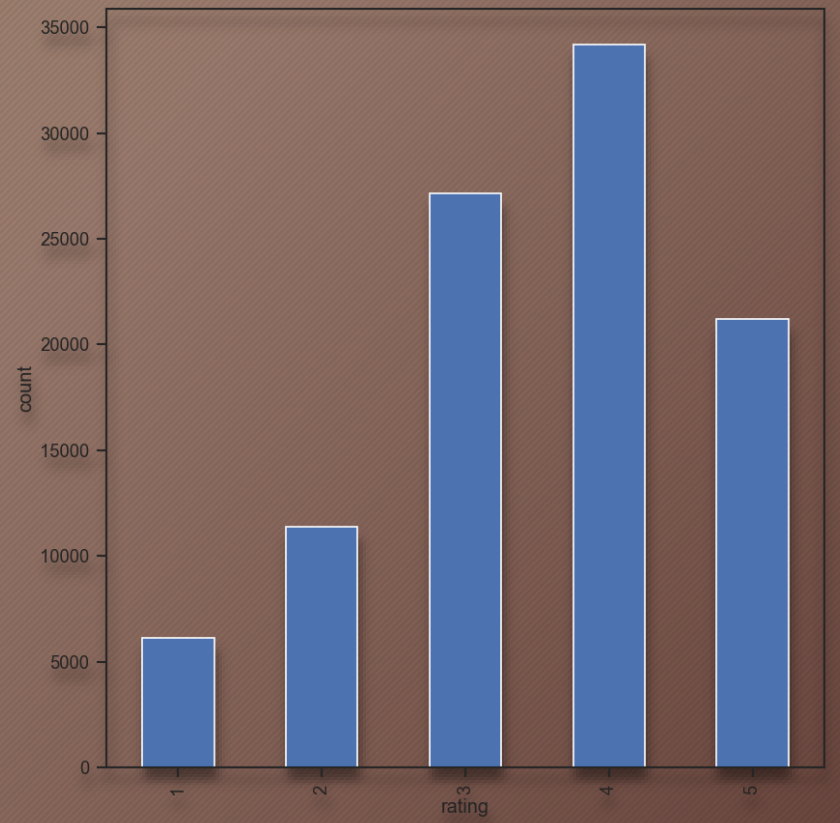






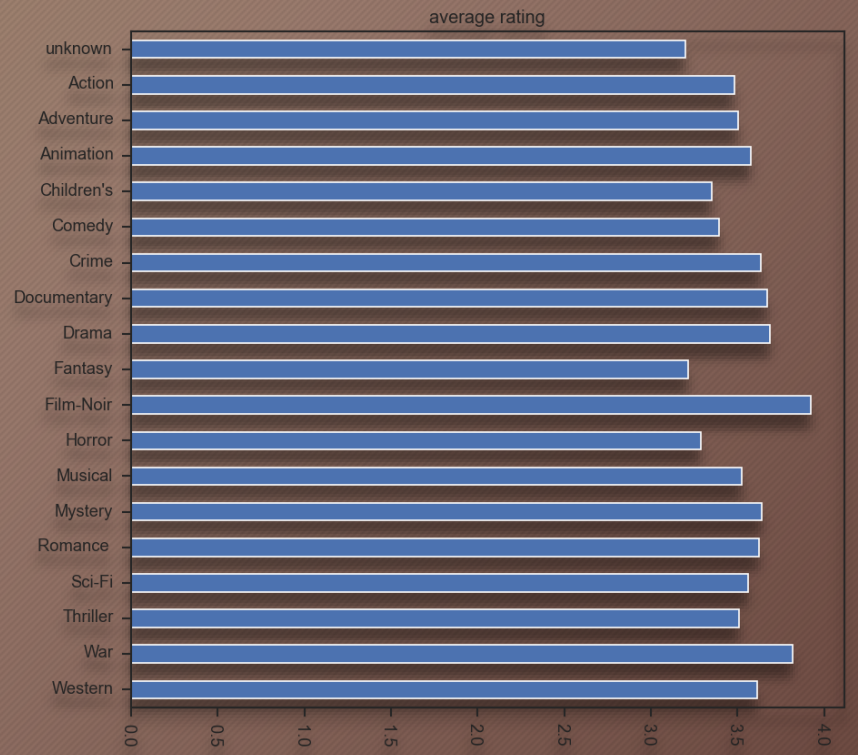
Analisi esplorativa

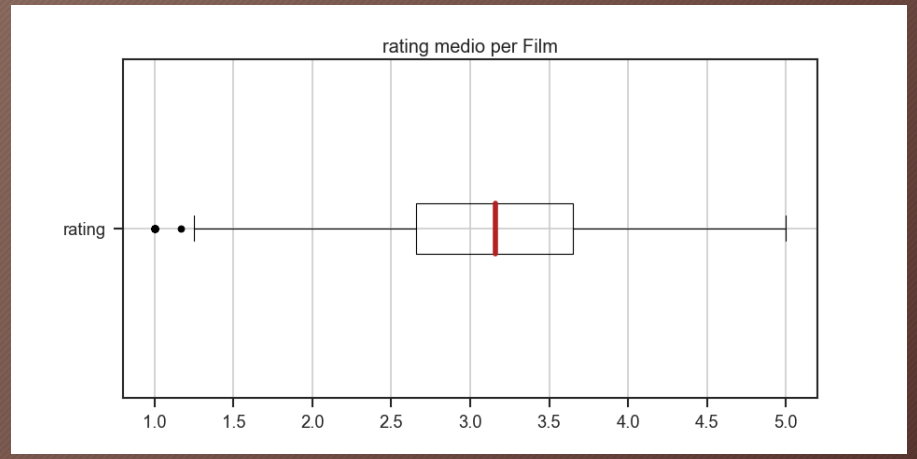




Analisi esplorativa

Analisi esplorativa





Matrice di rating

- I rating presenti sono: **100000**
- I rating mancanti sono: **1486126**
- I rating presenti sono il **6.3%**
- I rating mancanti sono il **93.7%**

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

K-NN

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

Cosine Similarity

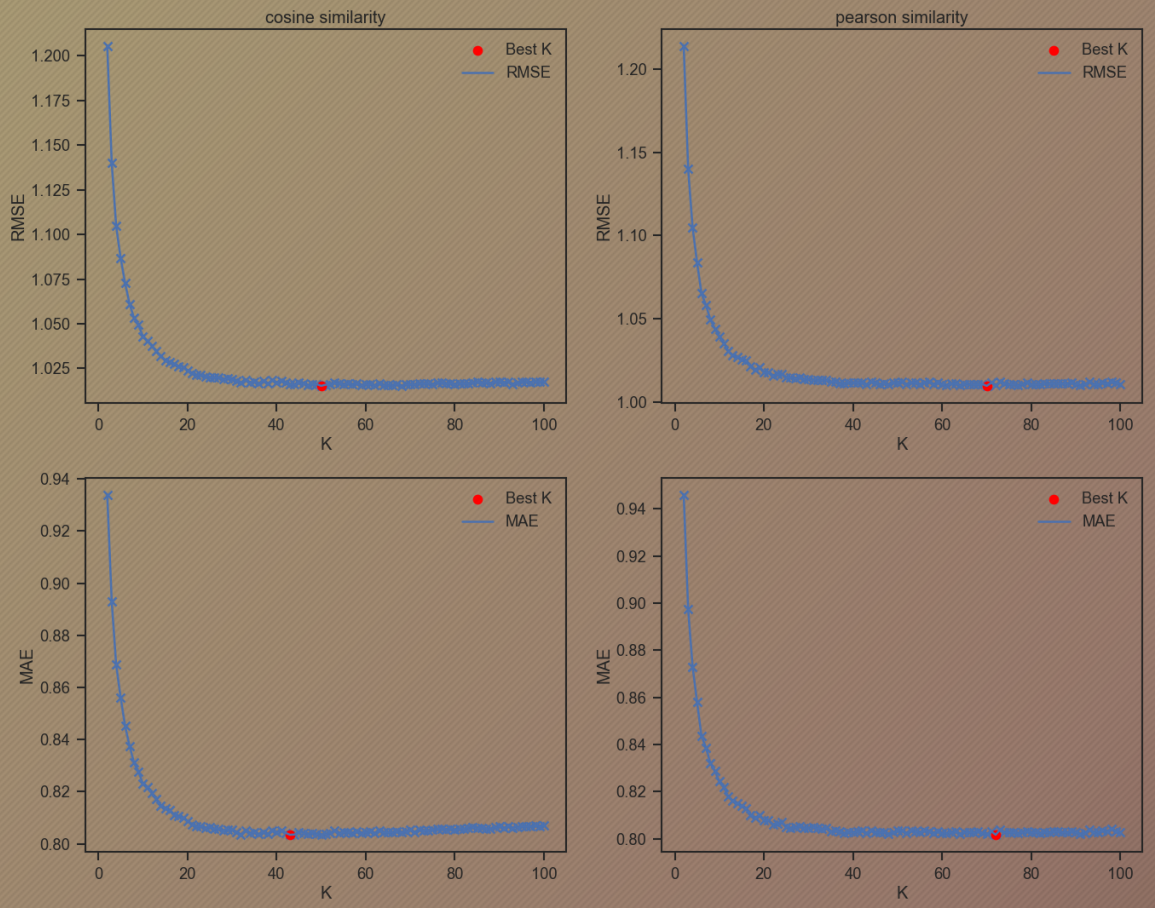
movie id	1	2	3	4	5	6	7	8	9	10	...	1673
user id												
1	5	3	4	3	3	5	4	1	5	3	...	3
2	4	3	3	4	3	3.5	4	4	4	2	...	3
3	4	3	3	3.5	3	3.5	4	4	4	4	...	3
4	4	3	3	4	3	3.5	4	4	4	4	...	3
5	4	3	3	4	3	3.5	4	4	4	4	...	3
...	...	...	...	...	...	...	...	...	...	...	...	...
939	4	3	3	4	3.5	3.5	4	4	5	4	...	3
940	4	3.5	3	2	3.5	3.5	4	5	3	4	...	3
941	5	3	3	3.5	3.5	3.5	4	4	4	4	...	3
942	4	3.5	3	4	3.5	3.5	4	4	4	4	...	3
943	4	5	3	4	3.5	3.5	4	4	3	4	...	3

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

Pearson Correlation

movie id	1	2	3	4	5	6	7	8	9	10	...	1673
user id												
1	5	3	4	3	3	5	4	1	5	3	...	4.0
2	4	3	3	4	3	4	4	4	4	2	...	3
3	4	3	3	3.5	3.5	3	4	4	4	4	...	3
4	4	3	3	3.5	3	4	4	4	3.5	4	...	4.0
5	4	3	3	4	3	4	4	4	4	4	...	3.0
...	...	...	...	...	...	...	...	...	...	...	...	...
939	4	3	3	4	3	3.5	4	4	5	4	...	3
940	4	3	3	2	3	3.5	4	5	3	4	...	3.5
941	5	3	3	4	3	4	4	4	4	4	...	4.0
942	4	3	3	4	3	3.5	3.5	4	4	4	...	4.0
943	4	5	3	4	3	4	4	4	3	4	...	3.5

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



Cosine Similarity:

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

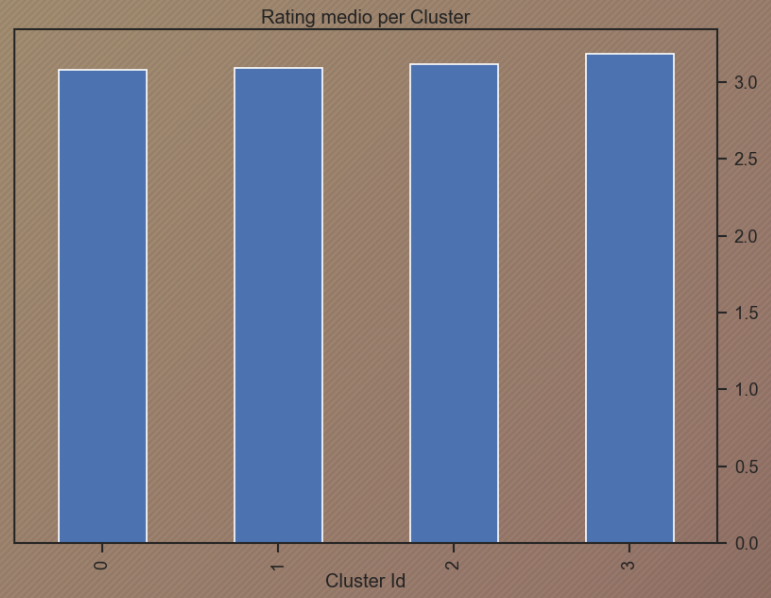
Pearson Correlation:

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

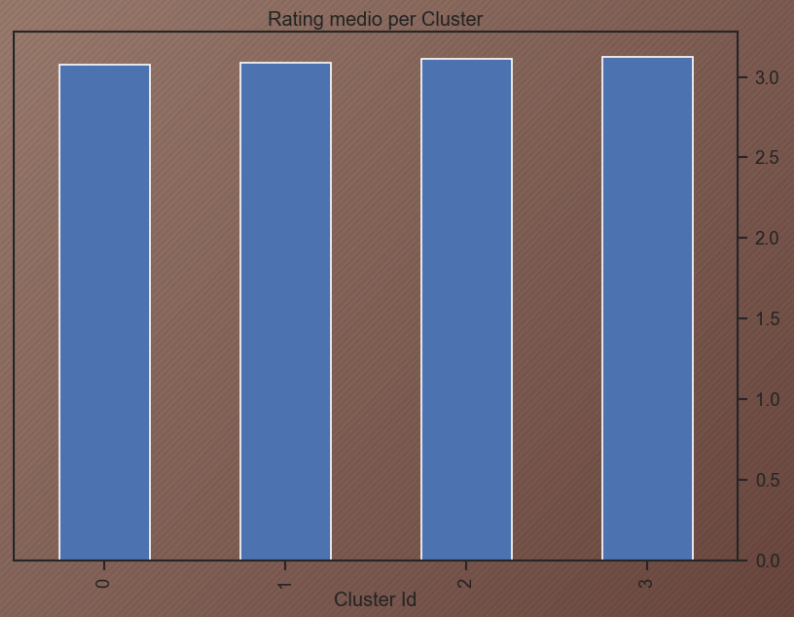
RMSE $\sqrt{\frac{\sum_{(i,j)\in S} e_{i,j}^2}{|S|}}$ MSE $\frac{\sum_{(i,j)\in S} e_{i,j}^2}{|S|}$ MAE $\frac{\sum_{(i,j)\in S} |e_{i,j}|}{|S|}$

Clustering

Euclidean Distance



Cosine Distance



Grazie per l’attenzione