In [13]:

```
import numpy as np
import pandas  as pd
```

In [15]:

```
rating = pd.read_csv(r'C:\Users\kavya\Downloads\archive\rating.csv')
```

In [16]:

```
rating
```

Out[16]:

	userId	movieId	rating	timestamp
0	1	2	3.5	2005-04-02 23:53:47
1	1	29	3.5	2005-04-02 23:31:16
2	1	32	3.5	2005-04-02 23:33:39
3	1	47	3.5	2005-04-02 23:32:07
4	1	50	3.5	2005-04-02 23:29:40
...	...	...	...	...
20000258	138493	68954	4.5	2009-11-13 15:42:00
20000259	138493	69526	4.5	2009-12-03 18:31:48
20000260	138493	69644	3.0	2009-12-07 18:10:57
20000261	138493	70286	5.0	2009-11-13 15:42:24
20000262	138493	71619	2.5	2009-10-17 20:25:36

20000263 rows × 4 columns

In [17]:

```
taggings = pd.read_csv(r'C:\Users\kavya\Downloads\archive\tag.csv')
```

In [18]:

```
taggings
```

Out[18]:

	userId	movieId	tag	timestamp
0	18	4141	Mark Waters	2009-04-24 18:19:40
1	65	208	dark hero	2013-05-10 01:41:18
2	65	353	dark hero	2013-05-10 01:41:19
3	65	521	noir thriller	2013-05-10 01:39:43
4	65	592	dark hero	2013-05-10 01:41:18
...	...	...	...	...
465559	138446	55999	dragged	2013-01-23 23:29:32
465560	138446	55999	Jason Bateman	2013-01-23 23:29:38
465561	138446	55999	quirky	2013-01-23 23:29:38
465562	138446	55999	sad	2013-01-23 23:29:32
465563	138472	923	rise to power	2007-11-02 21:12:47

465564 rows × 4 columns

In [19]:

```
movie = pd.read_csv(r'C:\Users\kavya\Downloads\archive\movie.csv')
```

In [20]:

```
movie
```

Out[20]:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy
...	...	...	...
27273	131254	Kein Bund für's Leben (2007)	Comedy
27274	131256	Feuer, Eis & Dosenbier (2002)	Comedy
27275	131258	The Pirates (2014)	Adventure
27276	131260	Rentun Ruusu (2001)	(no genres listed)
27277	131262	Innocence (2014)	Adventure|Fantasy|Horror

27278 rows × 3 columns

In [21]:

```
rating.head()
```

Out[21]:

	userId	movieId	rating	timestamp
0	1	2	3.5	2005-04-02 23:53:47
1	1	29	3.5	2005-04-02 23:31:16
2	1	32	3.5	2005-04-02 23:33:39
3	1	47	3.5	2005-04-02 23:32:07
4	1	50	3.5	2005-04-02 23:29:40

In [22]:

```
taggings.head()
```

Out[22]:

	userId	movieId	tag	timestamp
0	18	4141	Mark Waters	2009-04-24 18:19:40
1	65	208	dark hero	2013-05-10 01:41:18
2	65	353	dark hero	2013-05-10 01:41:19
3	65	521	noir thriller	2013-05-10 01:39:43
4	65	592	dark hero	2013-05-10 01:41:18

In [23]:

```
movie.head()
```

Out[23]:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

In [24]:

```
movie.columns
```

Out[24]:

```
Index(['movieId', 'title', 'genres'], dtype='object')
```

In [25]:

```
taggings.columns
```

Out[25]:

```
Index(['userId', 'movieId', 'tag', 'timestamp'], dtype='object')
```

In [26]:

```
rating.columns
```

Out[26]:

```
Index(['userId', 'movieId', 'rating', 'timestamp'], dtype='object')
```

In [27]:

```
del taggings['timestamp']
```

In [28]:

```
taggings.columns
```

Out[28]:

```
Index(['userId', 'movieId', 'tag'], dtype='object')
```

```
In [29]: del rating ['timestamp']
```

```
In [30]: rating.columns
```

```
Out[30]: Index(['userId', 'movieId', 'rating'], dtype='object')
```

```
In [31]: row_0 = taggings.iloc[0]
         type(row_0)
```

```
Out[31]: pandas.core.series.Series
```

datastructures

```
In [33]: row_0 = rating.iloc[0]
         type(row_0)
```

```
Out[33]: pandas.core.series.Series
```

```
In [34]: row_0 = movie.iloc[0]
         type(row_0)
```

```
Out[34]: pandas.core.series.Series
```

```
In [35]: print(row_0)
```

```
movieId                                                1
title                                   Toy Story (1995)
genres      Adventure|Animation|Children|Comedy|Fantasy
Name: 0, dtype: object
```

dataframes

```
In [37]: taggings.head()
```

Out[37]:

	userId	movieId	tag
0	18	4141	Mark Waters
1	65	208	dark hero
2	65	353	dark hero
3	65	521	noir thriller
4	65	592	dark hero

```
In [38]: taggings.index
```

```
Out[38]: RangeIndex(start=0, stop=465564, step=1)
```

```
In [39]: taggings.columns
```

```
Out[39]: Index(['userId', 'movieId', 'tag'], dtype='object')
```

In [40]:

```
taggings.iloc[[0,11,500]]
```

Out[40]:

	userId	movieId	tag
0	18	4141	Mark Waters
11	65	1783	noir thriller
500	342	55908	entirely dialogue

In [41]:

```
rating.iloc[[2,67,599]]
```

Out[41]:

	userId	movieId	rating
2	1	32	3.5
67	1	1997	3.5
599	7	1097	4.0

In [42]:

```
movie.iloc[[3,5,100]]
```

Out[42]:

	movieId	title	genres
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
5	6	Heat (1995)	Action|Crime|Thriller
100	102	Mr. Wrong (1996)	Comedy

descriptive statistics

In [44]:

```
rating['rating'].describe()
```

Out[44]:

```
count    2.000026e+07
mean     3.525529e+00
std      1.051989e+00
min      5.000000e-01
25%      3.000000e+00
50%      3.500000e+00
75%      4.000000e+00
max      5.000000e+00
Name: rating, dtype: float64
```

In [45]:

```
rating.describe()
```

Out[45]:

	userId	movieId	rating
count	2.000026e+07	2.000026e+07	2.000026e+07
mean	6.904587e+04	9.041567e+03	3.525529e+00
std	4.003863e+04	1.978948e+04	1.051989e+00
min	1.000000e+00	1.000000e+00	5.000000e-01
25%	3.439500e+04	9.020000e+02	3.000000e+00
50%	6.914100e+04	2.167000e+03	3.500000e+00
75%	1.036370e+05	4.770000e+03	4.000000e+00
max	1.384930e+05	1.312620e+05	5.000000e+00

In [46]:

```
rating.mean()
```

Out[46]:

```
userId     69045.872583
movieId     9041.567330
rating         3.525529
dtype: float64
```

In [47]:

```
rating.min()
```

Out[47]:

```
userId     1.0
movieId    1.0
rating     0.5
dtype: float64
```

In [48]:

```
rating.max()
```

Out[48]:

```
userId     138493.0
movieId    131262.0
rating          5.0
dtype: float64
```

In [49]:

```
rating.std()
```

Out[49]:

```
userId     40038.626653
movieId    19789.477445
rating         1.051989
dtype: float64
```

In [50]:

```
rating.mode()
```

Out[50]:

	userId	movieId	rating
0	118205	296	4.0

In [51]:

```
rating.corr()
```

Out[51]:

	userId	movieId	rating
userId	1.000000	-0.000850	0.001175
movieId	-0.000850	1.000000	0.002606
rating	0.001175	0.002606	1.000000

In [52]:

```
filter1 = rating['rating']>10
print(filter1)
filter1.any()
```

```
0           False
1           False
2           False
3           False
4           False
            ...
20000258    False
20000259    False
20000260    False
20000261    False
20000262    False
Name: rating, Length: 20000263, dtype: bool
```

Out[52]: `False`

In [53]:

```
filter2 = rating['rating'] >0
filter2.all()
```

Out[53]: `True`

In [66]:

```
movie.head()
```

Out[66]:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

In [67]:

```
movie.tail()
```

Out[67]:

	movieId	title	genres
27273	131254	Kein Bund für's Leben (2007)	Comedy
27274	131256	Feuer, Eis & Dosenbier (2002)	Comedy
27275	131258	The Pirates (2014)	Adventure
27276	131260	Rentun Ruusu (2001)	(no genres listed)
27277	131262	Innocence (2014)	Adventure|Fantasy|Horror

In [56]:

```
movie.describe()
```

Out[56]:

	movieId
count	27278.000000
mean	59855.480570
std	44429.314697
min	1.000000
25%	6931.250000
50%	68068.000000
75%	100293.250000
max	131262.000000

In [57]:

```
movie.mode()
```

Out[57]:

	movieId	title	genres
0	1	20,000 Leagues Under the Sea (1997)	Drama
1	2	Aladdin (1992)	NaN
2	3	Beneath (2013)	NaN
3	4	Blackout (2007)	NaN
4	5	Casanova (2005)	NaN
...	...	...	...
27273	131254	NaN	NaN
27274	131256	NaN	NaN
27275	131258	NaN	NaN
27276	131260	NaN	NaN
27277	131262	NaN	NaN

27278 rows × 3 columns

In [58]:

```
movie.min()
```

```
Out[58]: movieId                                                  1
         title      #chicagoGirl: The Social Network Takes on a Di...
         genres                                    (no genres listed)
         dtype: object
```

```
In [59]: movie.max()
```

```
Out[59]: movieId         131262
         title      貞子3D (2012)
         genres         Western
         dtype: object
```

```
In [60]: movie.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27278 entries, 0 to 27277
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   movieId  27278 non-null  int64 
 1   title    27278 non-null  object
 2   genres   27278 non-null  object
dtypes: int64(1), object(2)
memory usage: 639.5+ KB
```

```
In [61]: taggings.head()
```

Out[61]:

	userId	movieId	tag
0	18	4141	Mark Waters
1	65	208	dark hero
2	65	353	dark hero
3	65	521	noir thriller
4	65	592	dark hero

```
In [62]: taggings.tail()
```

Out[62]:

	userId	movieId	tag
465559	138446	55999	dragged
465560	138446	55999	Jason Bateman
465561	138446	55999	quirky
465562	138446	55999	sad
465563	138472	923	rise to power

data cleaning : handling missing data

```
In [64]: movie.shape
```

Out[64]:

```
(27278, 3)
```

In [114…

```
movie.isnull()
```

Out[114…

	movieId	title	genres
0	False	False	False
1	False	False	False
2	False	False	False
3	False	False	False
4	False	False	False
...	...	...	...
27273	False	False	False
27274	False	False	False
27275	False	False	False
27276	False	False	False
27277	False	False	False

27278 rows × 3 columns

In [116…

```
movie.isnull().any().any()
```

Out[116…

```
False
```

In [120…

```
rating.shape
```

Out[120…

```
(20000263, 3)
```

In [122…

```
rating.isnull()
```

Out[122…

	userId	movieId	rating
0	False	False	False
1	False	False	False
2	False	False	False
3	False	False	False
4	False	False	False
...	...	...	...
20000258	False	False	False
20000259	False	False	False
20000260	False	False	False
20000261	False	False	False
20000262	False	False	False

20000263 rows × 3 columns

In [124…

```
rating.isnull().any().any()
```

Out[124…

```
False
```

In [126…

```
taggings.shape
```

Out[126…

```
(465564, 3)
```

In [128…

```
taggings.isnull().any().any()
```

Out[128…

```
True
```

In [130…

```
taggings=taggings.dropna()
```

In [132…

```
taggings
```

Out[132…

	userId	movieId	tag
0	18	4141	Mark Waters
1	65	208	dark hero
2	65	353	dark hero
3	65	521	noir thriller
4	65	592	dark hero
...	...	...	...
465559	138446	55999	dragged
465560	138446	55999	Jason Bateman
465561	138446	55999	quirky
465562	138446	55999	sad
465563	138472	923	rise to power

465548 rows × 3 columns

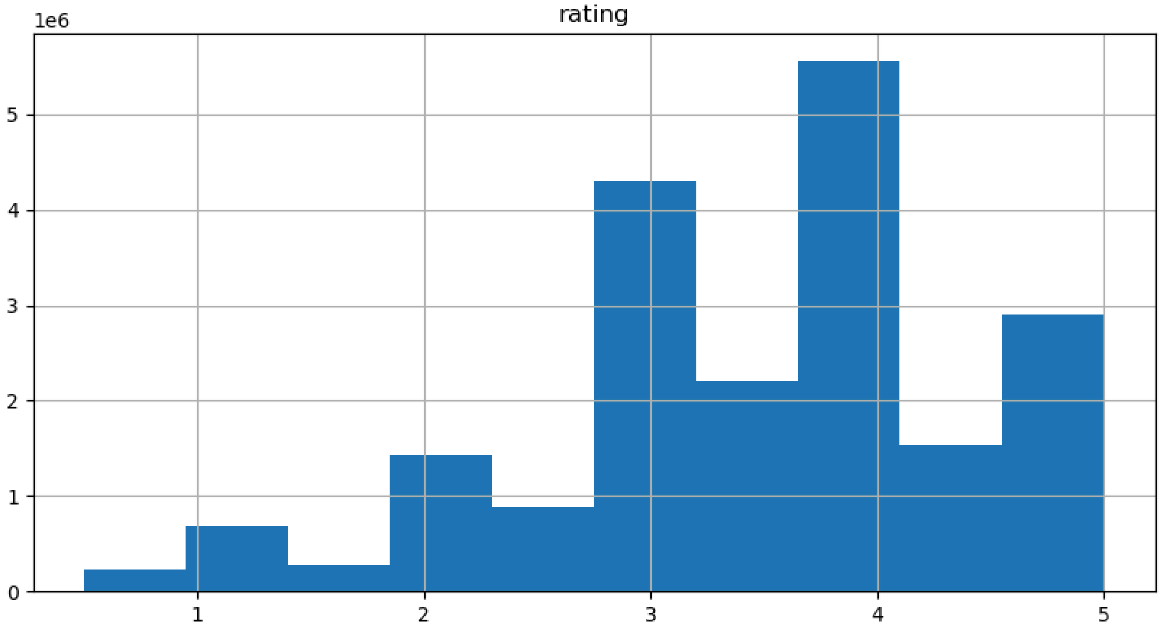
data visualization

In [137…

```
%matplotlib inline
rating.hist(column='rating',figsize=(10,5))
```

Out[137…

```
array([[<Axes: title={'center': 'rating'}>]], dtype=object)
```

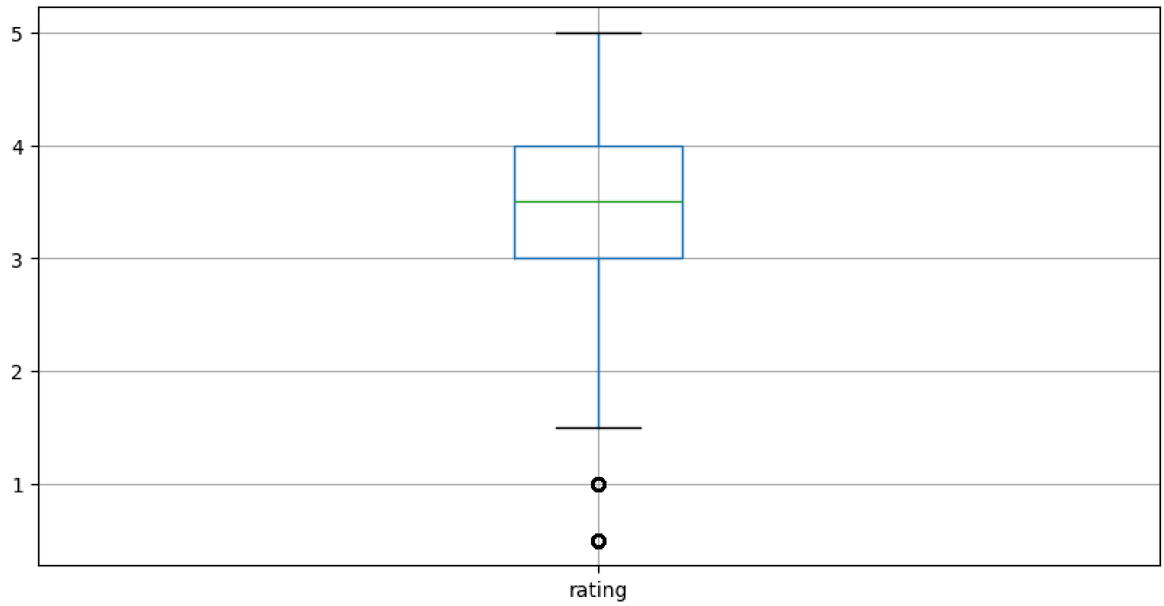


In [139…

```
rating.boxplot(column='rating',figsize=(10,5))
```

Out[139…

```
<Axes: >
```



slicing out columns

In [152…

```
rating['rating'].head()
```

Out[152…

```
0    3.5
1    3.5
2    3.5
3    3.5
4    3.5
Name: rating, dtype: float64
```

In [156…

```
movie[['title','genres']].head()
```

Out[156…

	title	genres
0	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	Jumanji (1995)	Adventure|Children|Fantasy
2	Grumpier Old Men (1995)	Comedy|Romance
3	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	Father of the Bride Part II (1995)	Comedy

In [158…

```
rating[-10:]
```

Out[158…

	userId	movieId	rating
20000253	138493	60816	4.5
20000254	138493	61160	4.0
20000255	138493	65682	4.5
20000256	138493	66762	4.5
20000257	138493	68319	4.5
20000258	138493	68954	4.5
20000259	138493	69526	4.5
20000260	138493	69644	3.0
20000261	138493	70286	5.0
20000262	138493	71619	2.5

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```