IMDb movie data analysis using pandas

In [1]:
```
import pandas as pd
```

In [2]:
```
movie = pd.read_csv("movie.csv")
rating = pd.read_csv("rating.csv")
tag = pd.read_csv("tag.csv")
```

In [3]:
```
print(movie.head(15))
print(rating.head(15))
print(tag.head(15))
```

	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
5	6	Heat (1995)	Action|Crime|Thriller
6	7	Sabrina (1995)	Comedy|Romance
7	8	Tom and Huck (1995)	Adventure|Children
8	9	Sudden Death (1995)	Action
9	10	GoldenEye (1995)	Action|Adventure|Thriller
10	11	American President, The (1995)	Comedy|Drama|Romance
11	12	Dracula: Dead and Loving It (1995)	Comedy|Horror
12	13	Balto (1995)	Adventure|Animation|Children
13	14	Nixon (1995)	Drama
14	15	Cutthroat Island (1995)	Action|Adventure|Romance

	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
5	1	112	3.5	2004-09-10 03:09:00
6	1	151	4.0	2004-09-10 03:08:54
7	1	223	4.0	2005-04-02 23:46:13
8	1	253	4.0	2005-04-02 23:35:40
9	1	260	4.0	2005-04-02 23:33:46
10	1	293	4.0	2005-04-02 23:31:43
11	1	296	4.0	2005-04-02 23:32:47
12	1	318	4.0	2005-04-02 23:33:18
13	1	337	3.5	2004-09-10 03:08:29
14	1	367	3.5	2005-04-02 23:53:00

	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
5	65	668	bollywood	2013-05-10 01:37:56
6	65	898	screwball comedy	2013-05-10 01:42:40
7	65	1248	noir thriller	2013-05-10 01:39:43
8	65	1391	mars	2013-05-10 01:40:55
9	65	1617	neo-noir	2013-05-10 01:43:37
10	65	1694	jesus	2013-05-10 01:38:45

11	65	1783	noir thriller	2013-05-10 01:39:43
12	65	2022	jesus	2013-05-10 01:38:45
13	65	2193	dragon	2013-05-10 02:01:54
14	65	2353	conspiracy theory	2013-05-10 02:01:06

In [4]:

```
print(movie.tail(15))
print(rating.tail(15))
print(tag.tail(15))
```

```
       movieId                                              title  \
27263   131176                               A Second Chance (2014)
27264   131180                      Dead Rising: Watchtower (2015)
27265   131231                                       Standby (2014)
27266   131237                           What Men Talk About (2010)
27267   131239                            Three Quarter Moon (2011)
27268   131241                             Ants in the Pants (2000)
27269   131243                  Werner - Gekotzt wird später (2003)
27270   131248                                Brother Bear 2 (2006)
27271   131250                                No More School (2000)
27272   131252  Forklift Driver Klaus: The First Day on the Jo...
27273   131254                          Kein Bund für's Leben (2007)
27274   131256                         Feuer, Eis & Dosenbier (2002)
27275   131258                                   The Pirates (2014)
27276   131260                                  Rentun Ruusu (2001)
27277   131262                                     Innocence (2014)

                                           genres
27263                                       Drama
27264                      Action|Horror|Thriller
27265                              Comedy|Romance
27266                                      Comedy
27267                                Comedy|Drama
27268                              Comedy|Romance
27269                            Animation|Comedy
27270  Adventure|Animation|Children|Comedy|Fantasy
27271                                      Comedy
27272                               Comedy|Horror
27273                                      Comedy
27274                                      Comedy
27275                                   Adventure
27276                          (no genres listed)
27277                    Adventure|Fantasy|Horror
          userId  movieId  rating           timestamp
20000248  138493    58879     4.5 2009-10-17 21:59:58
20000249  138493    59315     4.0 2009-10-17 22:22:18
20000250  138493    59725     3.0 2009-10-17 22:21:18
20000251  138493    59784     5.0 2009-10-17 22:01:41
20000252  138493    60069     4.0 2009-11-13 17:51:27
20000253  138493    60816     4.5 2009-12-03 18:32:43
20000254  138493    61160     4.0 2009-11-16 16:55:37
20000255  138493    65682     4.5 2009-10-17 21:52:53
20000256  138493    66762     4.5 2009-10-17 18:50:08
20000257  138493    68319     4.5 2009-12-07 18:15:20
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
        userId  movieId                             tag            time
stamp
465549  138446     3086                         classic  2013-01-23 23:
32:59
465550  138446     3086                           funny  2013-01-23 23:
32:59
465551  138446     3086                           scary  2013-01-23 23:
33:21
465552  138446     3489                       Peter Pan  2013-01-23 23:
30:22
465553  138446     3489                      soundtrack  2013-01-23 23:
30:22
```

```
465554  138446    3489                 visually appealing  2013-01-23 23:
30:22
465555  138446    7045                    family friendly  2013-01-23 23:
27:40
465556  138446    7045  Scary Movies To See on Halloween  2013-01-23 23:
27:40
465557  138446    7164                          Peter Pan  2013-01-23 23:
30:55
465558  138446    7164                 visually appealing  2013-01-23 23:
30:55
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
```

In [6]:

```
del rating['timestamp']
del tag['timestamp']
```

In [7]:

```
print(movie.head(15))
print(rating.head(15))
print(tag.head(15))
```

```
    movieId                               title  \
0         1                    Toy Story (1995)
1         2                      Jumanji (1995)
2         3             Grumpier Old Men (1995)
3         4            Waiting to Exhale (1995)
4         5  Father of the Bride Part II (1995)
5         6                         Heat (1995)
6         7                      Sabrina (1995)
7         8                 Tom and Huck (1995)
8         9                 Sudden Death (1995)
9        10                    GoldenEye (1995)
10       11      American President, The (1995)
11       12  Dracula: Dead and Loving It (1995)
12       13                        Balto (1995)
13       14                        Nixon (1995)
14       15             Cutthroat Island (1995)

                                          genres
0   Adventure|Animation|Children|Comedy|Fantasy
1                    Adventure|Children|Fantasy
2                                Comedy|Romance
3                          Comedy|Drama|Romance
4                                        Comedy
5                         Action|Crime|Thriller
6                                Comedy|Romance
7                            Adventure|Children
8                                        Action
9                     Action|Adventure|Thriller
10                         Comedy|Drama|Romance
11                                Comedy|Horror
12                 Adventure|Animation|Children
13                                        Drama
14                     Action|Adventure|Romance
    userId  movieId  rating
0        1        2     3.5
1        1       29     3.5
2        1       32     3.5
3        1       47     3.5
4        1       50     3.5
5        1      112     3.5
6        1      151     4.0
7        1      223     4.0
8        1      253     4.0
9        1      260     4.0
10       1      293     4.0
11       1      296     4.0
12       1      318     4.0
13       1      337     3.5
14       1      367     3.5
    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
5       65      668        bollywood
6       65      898  screwball comedy
7       65     1248    noir thriller
8       65     1391             mars
9       65     1617         neo-noir
10      65     1694            jesus
```

11	65	1783	noir thriller
12	65	2022	jesus
13	65	2193	dragon
14	65	2353	conspiracy theory

In [8]:
```
print(movie.tail(15))
print(rating.tail(15))
print(tag.tail(15))
```

```
       movieId                                              title  \
27263   131176                            A Second Chance (2014)
27264   131180                    Dead Rising: Watchtower (2015)
27265   131231                                    Standby (2014)
27266   131237                        What Men Talk About (2010)
27267   131239                         Three Quarter Moon (2011)
27268   131241                          Ants in the Pants (2000)
27269   131243              Werner - Gekotzt wird später (2003)
27270   131248                             Brother Bear 2 (2006)
27271   131250                             No More School (2000)
27272   131252  Forklift Driver Klaus: The First Day on the Jo...
27273   131254                       Kein Bund für's Leben (2007)
27274   131256                      Feuer, Eis & Dosenbier (2002)
27275   131258                                The Pirates (2014)
27276   131260                               Rentun Ruusu (2001)
27277   131262                                  Innocence (2014)

                                           genres
27263                                       Drama
27264                      Action|Horror|Thriller
27265                              Comedy|Romance
27266                                      Comedy
27267                                Comedy|Drama
27268                              Comedy|Romance
27269                            Animation|Comedy
27270  Adventure|Animation|Children|Comedy|Fantasy
27271                                      Comedy
27272                               Comedy|Horror
27273                                      Comedy
27274                                      Comedy
27275                                   Adventure
27276                          (no genres listed)
27277                    Adventure|Fantasy|Horror
          userId  movieId  rating
20000248  138493    58879     4.5
20000249  138493    59315     4.0
20000250  138493    59725     3.0
20000251  138493    59784     5.0
20000252  138493    60069     4.0
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
        userId  movieId                                tag
465549  138446     3086                            classic
465550  138446     3086                              funny
465551  138446     3086                              scary
465552  138446     3489                          Peter Pan
465553  138446     3489                         soundtrack
465554  138446     3489                visually appealing
465555  138446     7045                    family friendly
465556  138446     7045  Scary Movies To See on Halloween
465557  138446     7164                          Peter Pan
465558  138446     7164                visually appealing
465559  138446    55999                            dragged
```

```
465560  138446    55999                      Jason Bateman
465561  138446    55999                             quirky
465562  138446    55999                                sad
465563  138472      923                      rise to power
```

Data Structures

Series

In [9]:
```
row_0 = tag.iloc[0]
type(row_0)
```

Out[9]:
```
pandas.core.series.Series
```

In [10]:
```
print(row_0)
```

```
userId              18
movieId           4141
tag        Mark Waters
Name: 0, dtype: object
```

In [11]:
```
row_0.index
```

Out[11]:
```
Index(['userId', 'movieId', 'tag'], dtype='object')
```

In [12]:
```
row_0['userId']
```

Out[12]:
```
18
```

In [13]:
```
'rating' in row_0
```

Out[13]:
```
False
```

In [14]:
```
row_0.name
```

Out[14]:
```
0
```

In [15]:
```
row_0 = row_0.rename('firstRow')
row_0.name
```

Out[15]:
```
'firstRow'
```

DataFrames

In [16]:
```
tag.head(15)
```

Out[16]:

	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
5	65	668	bollywood
6	65	898	screwball comedy
7	65	1248	noir thriller
8	65	1391	mars
9	65	1617	neo-noir
10	65	1694	jesus
11	65	1783	noir thriller
12	65	2022	jesus
13	65	2193	dragon
14	65	2353	conspiracy theory

In [18]:
```
tag.index
```

Out[18]:
```
RangeIndex(start=0, stop=465564, step=1)
```

In [19]:
```
tag.columns
```

Out[19]:
```
Index(['userId', 'movieId', 'tag'], dtype='object')
```

In [20]:
```
tag.iloc[ [0,11,500] ]
```

Out[20]:

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

Descriptive Statistics

Lets look how the rating are distributed!

In [21]:
```
rating['rating'].describe()
```

Out[21]:
```
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 [22]:
```
rating.describe()
```

Out[22]:

	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 [25]:
```
rating['rating'].mean()
```

Out[25]:
```
3.5255285642993797
```

In [26]:
```
rating.mean()
```

Out[26]:
```
userId     69045.872583
movieId     9041.567330
rating         3.525529
dtype: float64
```

In [27]:
```
rating['rating'].min()
```

Out[27]:
```
0.5
```

In [28]:
```
rating['rating'].max()
```

Out[28]:
```
5.0
```

In [29]:
```
rating['rating'].std()
```

Out[29]:
```
1.051988919275684
```

In [30]:
```
rating['rating'].mode()
```

Out[30]:
```
0    4.0
Name: rating, dtype: float64
```

In [31]:
```
rating.corr()
```

Out[31]:

	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 [32]:
```
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[32]:
```
False
```

In [33]:
```
filter1.any().sum()
```

Out[33]:
```
0
```

In [34]:
```
filter2 = rating['rating'] > 0
filter2.all()
```

Out[34]:
```
True
```

Data Cleaning: Handling Missing Data

In [35]:
```
movie.shape
```

Out[35]:
```
(27278, 3)
```

In [36]:
```
movie.isnull().sum()
```

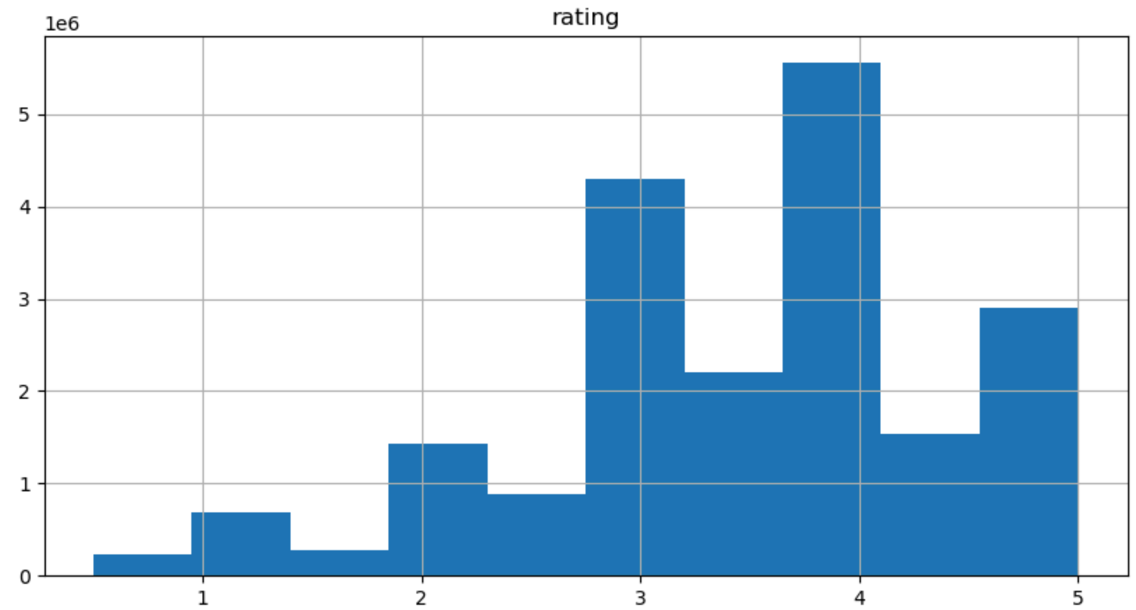
Out[36]:
```
movieId    0
title      0
genres     0
dtype: int64
```

In [37]:
```
rating.shape
```
Out[37]:
```
(20000263, 3)
```
In [38]:
```
rating.isnull().sum().any()
```
Out[38]:
```
False
```
In [40]:
```
tag.shape
```
Out[40]:
```
(465564, 3)
```
In [41]:
```
tag.isnull().sum().any()
```
Out[41]:
```
True
```
In [42]:
```
tag = tag.dropna()
```
In [43]:
```
tag.isnull().sum().any()
```
Out[43]:
```
False
```
In [44]:
```
tag.shape
```
Out[44]:
```
(465548, 3)
```

Data Visualization

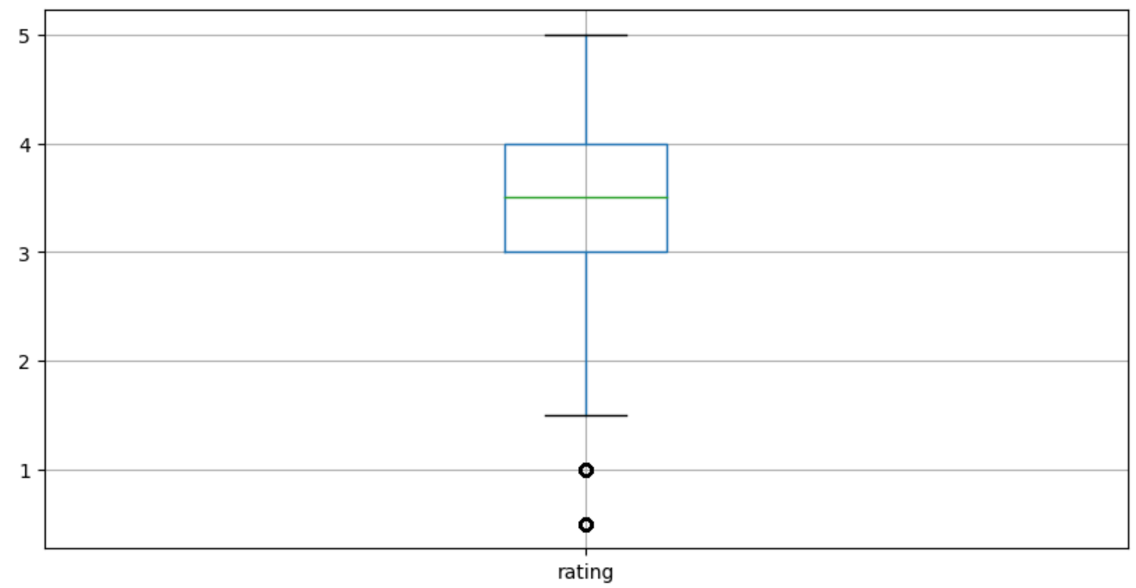
In [48]:
```
import matplotlib.pyplot as plt
%matplotlib inline

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

Out[48]: array([[<Axes: title={'center': 'rating'}>]], dtype=object)



In [49]:
```
rating.boxplot(column = 'rating', figsize = (10,5))
```

Out[49]: <Axes: >



Slicing Out Columns

In [50]:
```
tag['tag'].head(15)
```

Out[50]:
```
0            Mark Waters
1              dark hero
2              dark hero
3          noir thriller
4              dark hero
5              bollywood
6        screwball comedy
7          noir thriller
8                   mars
9              neo-noir
10                 jesus
11         noir thriller
12                 jesus
13                dragon
14      conspiracy theory
Name: tag, dtype: object
```

In [52]:
```
tag['tag'].tail(15)
```

Out[52]:
```
465549                             classic
465550                               funny
465551                               scary
465552                           Peter Pan
465553                          soundtrack
465554                  visually appealing
465555                     family friendly
465556    Scary Movies To See on Halloween
465557                           Peter Pan
465558                  visually appealing
465559                             dragged
465560                       Jason Bateman
465561                              quirky
465562                                 sad
465563                       rise to power
Name: tag, dtype: object
```

In [53]:
```
movie[['title','genres']].head(15)
```

Out[53]:

	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
5	Heat (1995)	Action|Crime|Thriller
6	Sabrina (1995)	Comedy|Romance
7	Tom and Huck (1995)	Adventure|Children
8	Sudden Death (1995)	Action
9	GoldenEye (1995)	Action|Adventure|Thriller
10	American President, The (1995)	Comedy|Drama|Romance
11	Dracula: Dead and Loving It (1995)	Comedy|Horror
12	Balto (1995)	Adventure|Animation|Children
13	Nixon (1995)	Drama
14	Cutthroat Island (1995)	Action|Adventure|Romance

In [54]:
```
rating[-10:]
```

Out[54]:

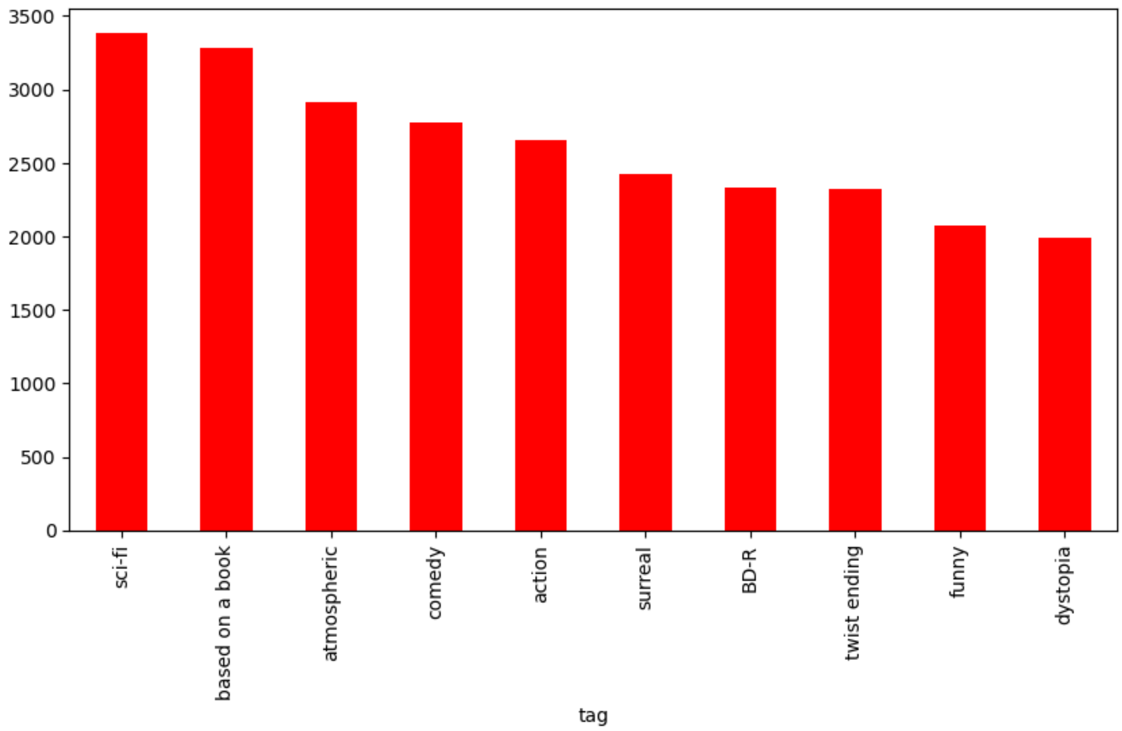
	userId	movieId	rating
20000253	138493	60816	4.5
20000254	138493	61160	4.0
20000255	138493	65682	4.5
20000256	138493	66762	4.5
20000257	138493	68319	4.5
20000258	138493	68954	4.5
20000259	138493	69526	4.5
20000260	138493	69644	3.0
20000261	138493	70286	5.0
20000262	138493	71619	2.5

In [55]:
```python
tag_count = tag['tag'].value_counts()
tag_count[-10:]
```

Out[55]:
```
tag
missing child                   1
Ron Moore                       1
Citizen Kane                    1
mullet                          1
biker gang                      1
Paul Adelstein                  1
the wig                         1
killer fish                     1
genetically modified monsters   1
topless scene                   1
Name: count, dtype: int64
```

In [61]:
```python
tag_count[:10].plot(kind = 'bar', figsize=(10,5), color = "Red")
```

Out[61]: <Axes: xlabel='tag'>



Filter for Selecting Rows

In [62]:
```
is_highlt_rated = rating['rating'] >= 5.0
rating[is_highlt_rated][30:50]
```

Out[62]:

	userId	movieId	rating
239	3	50	5.0
242	3	175	5.0
244	3	223	5.0
245	3	260	5.0
246	3	316	5.0
247	3	318	5.0
248	3	329	5.0
252	3	457	5.0
253	3	480	5.0
254	3	490	5.0
256	3	541	5.0
258	3	593	5.0
263	3	858	5.0
264	3	904	5.0
267	3	924	5.0
268	3	953	5.0
271	3	1060	5.0
272	3	1073	5.0
275	3	1084	5.0
276	3	1089	5.0

In [63]:
```
is_action = movie['genres'].str.contains('Action')
movie[is_action][5:15]
```

Out[63]:

	movieId	title	genres
22	23	Assassins (1995)	Action|Crime|Thriller
41	42	Dead Presidents (1995)	Action|Crime|Drama
43	44	Mortal Kombat (1995)	Action|Adventure|Fantasy
50	51	Guardian Angel (1994)	Action|Drama|Thriller
65	66	Lawnmower Man 2: Beyond Cyberspace (1996)	Action|Sci-Fi|Thriller
69	70	From Dusk Till Dawn (1996)	Action|Comedy|Horror|Thriller
70	71	Fair Game (1995)	Action
75	76	Screamers (1995)	Action|Sci-Fi|Thriller
77	78	Crossing Guard, The (1995)	Action|Crime|Drama|Thriller
85	86	White Squall (1996)	Action|Adventure|Drama

In [64]:
```
movie[is_action].head(15)
```

Out[64]:

	movieId	title	genres
5	6	Heat (1995)	Action|Crime|Thriller
8	9	Sudden Death (1995)	Action
9	10	GoldenEye (1995)	Action|Adventure|Thriller
14	15	Cutthroat Island (1995)	Action|Adventure|Romance
19	20	Money Train (1995)	Action|Comedy|Crime|Drama|Thriller
22	23	Assassins (1995)	Action|Crime|Thriller
41	42	Dead Presidents (1995)	Action|Crime|Drama
43	44	Mortal Kombat (1995)	Action|Adventure|Fantasy
50	51	Guardian Angel (1994)	Action|Drama|Thriller
65	66	Lawnmower Man 2: Beyond Cyberspace (1996)	Action|Sci-Fi|Thriller
69	70	From Dusk Till Dawn (1996)	Action|Comedy|Horror|Thriller
70	71	Fair Game (1995)	Action
75	76	Screamers (1995)	Action|Sci-Fi|Thriller
77	78	Crossing Guard, The (1995)	Action|Crime|Drama|Thriller
85	86	White Squall (1996)	Action|Adventure|Drama

Group By and Aggregate

In [65]:
```
rating_count = rating[['movieId','rating']].groupby('rating').count()
rating_count
```

Out[65]:

rating	movieId
0.5	239125
1.0	680732
1.5	279252
2.0	1430997
2.5	883398
3.0	4291193
3.5	2200156
4.0	5561926
4.5	1534824
5.0	2898660

In [66]:
```
average_rating = rating[['movieId','rating']].groupby('rating').mean()
average_rating
```

Out[66]:

rating	movieId
0.5	13356.246729
1.0	5652.208219
1.5	12377.773230
2.0	6733.595232
2.5	13669.268192
3.0	6770.763264
3.5	14814.098703
4.0	8342.514461
4.5	14585.414824
5.0	6275.356017

In [67]:
```
movie_count = rating[['movieId','rating']].groupby('movieId').count()
movie_count.head()
```

Out[67]:

movieId	rating
1	49695
2	22243
3	12735
4	2756
5	12161

In [68]:
```
movie_count.tail()
```

Out[68]:

movieId	rating
131254	1
131256	1
131258	1
131260	1
131262	1

Merge DataFrames

In [69]:
```
tag.head()
```

Out[69]:

	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 [70]:
```
movie.head()
```

Out[70]:

	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 [71]:
```
t = movie.merge(tag, on='movieId', how ='inner')
t.head()
```

Out[71]:

	movieId	title	genres	userId	tag
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	1644	Watched
1	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	1741	computer animation
2	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	1741	Disney animated feature
3	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	1741	Pixar animation
4	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	1741	TÃ©a Leoni does not star in this movie

In [72]:
```
t.tail()
```

Out[72]:

	movieId	title	genres	userId	tag
465543	131258	The Pirates (2014)	Adventure	28906	bandits
465544	131258	The Pirates (2014)	Adventure	28906	Korea
465545	131258	The Pirates (2014)	Adventure	28906	mutiny
465546	131258	The Pirates (2014)	Adventure	28906	pirates
465547	131258	The Pirates (2014)	Adventure	28906	whale

Combine aggregation, merging, and filters to get useful anlystics

In [73]:
```
avg_rating = rating.groupby('movieId', as_index=False).mean()
del avg_rating['userId']
avg_rating.head()
```

Out[73]:

	movieId	rating
0	1	3.921240
1	2	3.211977
2	3	3.151040
3	4	2.861393
4	5	3.064592

In [74]:
```
box_office = movie.merge(avg_rating, on='movieId', how='inner')
box_office.tail()
```

Out[74]:

	movieId	title	genres	rating
26739	131254	Kein Bund für's Leben (2007)	Comedy	4.0
26740	131256	Feuer, Eis & Dosenbier (2002)	Comedy	4.0
26741	131258	The Pirates (2014)	Adventure	2.5
26742	131260	Rentun Ruusu (2001)	(no genres listed)	3.0
26743	131262	Innocence (2014)	Adventure|Fantasy|Horror	4.0

In [75]:
```
is_highly_rated = box_office['rating'] >= 4.0
box_office[is_highly_rated][-5:]
```

Out[75]:

	movieId	title	genres	rating
26737	131250	No More School (2000)	Comedy	4.0
26738	131252	Forklift Driver Klaus: The First Day on the Jo...	Comedy|Horror	4.0
26739	131254	Kein Bund für's Leben (2007)	Comedy	4.0
26740	131256	Feuer, Eis & Dosenbier (2002)	Comedy	4.0
26743	131262	Innocence (2014)	Adventure|Fantasy|Horror	4.0

In [76]:
```
is_Adventure = box_office['genres'].str.contains('Adventure')
box_office[is_Adventure][:5]
```

Out[76]:

	movieId	title	genres	rating
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	3.921240
1	2	Jumanji (1995)	Adventure|Children|Fantasy	3.211977
7	8	Tom and Huck (1995)	Adventure|Children	3.142049
9	10	GoldenEye (1995)	Action|Adventure|Thriller	3.430029
12	13	Balto (1995)	Adventure|Animation|Children	3.272416

In [77]:
```
box_office[is_Adventure & is_highly_rated][-5:]
```

Out[77]:

	movieId	title	genres	rating
26611	130586	Itinerary of a Spoiled Child (1988)	Adventure|Drama	4.5
26655	130996	The Beautiful Story (1992)	Adventure|Drama|Fantasy	5.0
26667	131050	Stargate SG-1 Children of the Gods - Final Cut...	Adventure|Sci-Fi|Thriller	5.0
26736	131248	Brother Bear 2 (2006)	Adventure|Animation|Children|Comedy|Fantasy	4.0
26743	131262	Innocence (2014)	Adventure|Fantasy|Horror	4.0

Vectorized String Operations

In [78]:
```
movie.head()
```

Out[78]:

	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

Split 'genres' into multiple columns

In [83]:
```
movie_genres = movie['genres'].str.split('|',expand = True)
movie_genres[:10]
```

Out[83]:

	0	1	2	3	4	5	6	7	8	9
0	Adventure	Animation	Children	Comedy	Fantasy	None	None	None	None	None
1	Adventure	Children	Fantasy	None	None	None	None	None	None	None
2	Comedy	Romance	None	None	None	None	None	None	None	None
3	Comedy	Drama	Romance	None	None	None	None	None	None	None
4	Comedy	None	None	None	None	None	None	None	None	None
5	Action	Crime	Thriller	None	None	None	None	None	None	None
6	Comedy	Romance	None	None	None	None	None	None	None	None
7	Adventure	Children	None	None	None	None	None	None	None	None
8	Action	None	None	None	None	None	None	None	None	None
9	Action	Adventure	Thriller	None	None	None	None	None	None	None

Add a new column for comedy genre flag

In [84]:
```
movie_genres['isComedy'] = movie['genres'].str.contains('Comedy')
movie_genres[:10]
```

Out[84]:

	0	1	2	3	4	5	6	7	8	9	isCom
0	Adventure	Animation	Children	Comedy	Fantasy	None	None	None	None	None	T
1	Adventure	Children	Fantasy	None	None	None	None	None	None	None	Fa
2	Comedy	Romance	None	None	None	None	None	None	None	None	T
3	Comedy	Drama	Romance	None	None	None	None	None	None	None	T
4	Comedy	None	None	None	None	None	None	None	None	None	T
5	Action	Crime	Thriller	None	None	None	None	None	None	None	Fa
6	Comedy	Romance	None	None	None	None	None	None	None	None	T
7	Adventure	Children	None	None	None	None	None	None	None	None	Fa
8	Action	None	None	None	None	None	None	None	None	None	Fa
9	Action	Adventure	Thriller	None	None	None	None	None	None	None	Fa

Extract year from title e.g. (2007)

In [85]:
```
movie['Year'] = movie['title'].str.extract('.*\((.*)\).*', expand = True)
movie.tail()
```

Out[85]:

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

Parsing Timestamps

Timestamps are common in sensor data or other time series datasets. Let us revisit the tags.csv dataset and read the timestamps!

In [86]:
```
tags = pd.read_csv("tag.csv", sep=',')
```

In [87]:
```
tags.dtypes
```

Out[87]:
```
userId         int64
movieId        int64
tag           object
timestamp     object
dtype: object
```

Unix time / POSIX time / epoch time records time in seconds

since midnight Coordinated Universal Time (UTC) of April 4, 2009

In [88]:
```
tags.head()
```

Out[88]:

	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 [93]:
```
tags['parsed_time'] = pd.to_datetime(tags['timestamp'], unit='ns')
```

In [94]:
```
tags['parsed_time'].dtypes
```

Out[94]:
```
dtype('<M8[ns]')
```

In [95]:
```
tags.head(2)
```

Out[95]:

	userId	movieId	tag	timestamp	parsed_time
0	18	4141	Mark Waters	2009-04-24 18:19:40	2009-04-24 18:19:40
1	65	208	dark hero	2013-05-10 01:41:18	2013-05-10 01:41:18

Selecting rows based on timestamps

In [96]:
```
greater_than_t = tags['parsed_time'] > '2015-02-01'
selected_rows = tags[greater_than_t]
tags.shape, selected_rows.shape
```

Out[96]:
```
((465564, 5), (12130, 5))
```

In [97]:
```python
tags.sort_values(by='parsed_time', ascending=True)[:10]
```

Out[97]:

	userId	movieId	tag	timestamp	parsed_time
333932	100371	2788	monty python	2005-12-24 13:00:10	2005-12-24 13:00:10
333927	100371	1732	coen brothers	2005-12-24 13:00:36	2005-12-24 13:00:36
333924	100371	1206	stanley kubrick	2005-12-24 13:00:48	2005-12-24 13:00:48
333923	100371	1193	jack nicholson	2005-12-24 13:02:51	2005-12-24 13:02:51
333939	100371	5004	peter sellers	2005-12-24 13:03:19	2005-12-24 13:03:19
333922	100371	47	morgan freeman	2005-12-24 13:03:32	2005-12-24 13:03:32
333921	100371	47	brad pitt	2005-12-24 13:03:32	2005-12-24 13:03:32
333936	100371	4011	brad pitt	2005-12-24 13:03:51	2005-12-24 13:03:51
333937	100371	4011	guy ritchie	2005-12-24 13:03:51	2005-12-24 13:03:51
333920	100371	32	bruce willis	2005-12-24 13:04:02	2005-12-24 13:04:02

Average Movie Ratings over Time

Movie ratings related to the year of launch?

In [98]:
```python
average_rating = rating[['movieId','rating']].groupby('movieId', as_index=F
average_rating.tail()
```

Out[98]:

	movieId	rating
26739	131254	4.0
26740	131256	4.0
26741	131258	2.5
26742	131260	3.0
26743	131262	4.0

In [100]:
```python
joined = movie.merge(average_rating, on='movieId', how='inner')
joined.head()
```

Out[100]:

	movieId	title	genres	Year	rating
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	1995	3.921240
1	2	Jumanji (1995)	Adventure|Children|Fantasy	1995	3.211977
2	3	Grumpier Old Men (1995)	Comedy|Romance	1995	3.151040
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance	1995	2.861393
4	5	Father of the Bride Part II (1995)	Comedy	1995	3.064592

In [ ]:

```

```

In [ ]:

```

```