importing libraries

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

Reading the dataset

In [2]:
```
movies=pd.read_csv(r'D:\FSDS\New folder\movie.csv')
```

In [3]:
```
ratings=pd.read_csv(r'D:\FSDS\New folder\rating.csv')
```

In [4]:
```
tags=pd.read_csv(r'D:\FSDS\New folder\tag.csv')
```

In [5]:
```
tags.columns
```

Out[5]:
```
Index(['userId', 'movieId', 'tag', 'timestamp'], dtype='object')
```

In [6]:
```
ratings.columns
```

Out[6]:
```
Index(['userId', 'movieId', 'rating', 'timestamp'], dtype='object')
```

In [7]:
```
movies.columns
```

Out[7]:
```
Index(['movieId', 'title', 'genres'], dtype='object')
```

In [8]:
```
del tags['timestamp']
del ratings['timestamp']
```

In [9]:
```
tags.columns
```

Out[9]:
```
Index(['userId', 'movieId', 'tag'], dtype='object')
```

In [10]:
```
ratings.columns
```

Out[10]:
```
Index(['userId', 'movieId', 'rating'], dtype='object')
```

In [11]:
```
tags.head()
```

Out[11]:

	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

Data structures

In [12]:

```
tags.iloc[0]
```

Out[12]:

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

In [13]:

```
row_0=tags.iloc[1]
```

In [14]:

```
print(row_0)
```

```
userId              65
movieId            208
tag          dark hero
Name: 1, dtype: object
```

In [15]:

```
row_0.index
```

Out[15]:

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

In [16]:

```
row_0.userId
```

Out[16]:

```
65
```

In [17]:

```
row_0.tag
```

Out[17]:

```
'dark hero'
```

In [18]:

```
'rating'in row_0
```

Out[18]:

```
False
```

In [19]:

```
'movieId' in row_0
```

Out[19]:

```
True
```

In [20]:

```
row_0=row_0.rename('First Row')
row_0
```

Out[20]:

```
userId             65
movieId           208
tag         dark hero
Name: First Row, dtype: object
```

In [21]:

```
tags.iloc[1]
```

Out[21]:

```
userId             65
movieId           208
tag         dark hero
Name: 1, dtype: object
```

In [22]:

```
row_0=row_0.rename('Hi')
row_0
```

Out[22]:

```
userId             65
movieId           208
tag         dark hero
Name: Hi, dtype: object
```

In [23]:

```
tags.iloc[1]
```

Out[23]:

```
userId             65
movieId           208
tag         dark hero
Name: 1, dtype: object
```

Data frame

In [24]:

```
tags.index
```

Out[24]:

```
RangeIndex(start=0, stop=465564, step=1)
```

In [25]:

```
tags.columns
```

Out[25]:

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

In [26]:

```
tags.head()
```

Out[26]:

	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 [27]:

```
tags.iloc[[1,2000,4]]
```

Out[27]:

	userId	movieId	tag
1	65	208	dark hero
2000	910	68554	conspiracy theory
4	65	592	dark hero

Descriptive Statistics

In [28]:
```
ratings.describe()
```

Out[28]:

	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 [29]:
```
ratings['rating'].describe()
```

Out[29]:
```
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 [30]:
```
ratings.count()
```

Out[30]:
```
userId     20000263
movieId    20000263
rating     20000263
dtype: int64
```

In [31]:
```
ratings.mean()
```

```
Out[31]: userId     69045.872583
         movieId     9041.567330
         rating         3.525529
         dtype: float64
```

```
In [32]: ratings['rating'].max()
```

```
Out[32]: 5.0
```

```
In [33]: ratings['rating'].mean()
```

```
Out[33]: 3.5255285642993797
```

```
In [34]: ratings.corr()
```

Out[34]:

	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 [35]: filter1=ratings['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[35]: False
```

```
In [36]: filter2=ratings['rating']>0
         filter2.all()
```

```
Out[36]: True
```

Data cleaning :Handling missing data

```
In [37]: movies.shape
```

```
Out[37]: (27278, 3)
```

In [38]:

```
movies.isnull().any().all()
```

Out[38]:

```
False
```

In [39]:

```
ratings.shape
```

Out[39]:

```
(20000263, 3)
```

In [40]:

```
ratings.isnull().any().all()
```

Out[40]:

```
False
```

In [41]:

```
tags.shape
```

Out[41]:

```
(465564, 3)
```

In [42]:

```
tags.isnull().any().all()
```

Out[42]:

```
False
```

In [43]:

```
tags=tags.dropna()
```

In [44]:

```
tags
```

Out[44]:

	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

In [45]:

```
tags.shape
```

Out[45]:

```
(465548, 3)
```

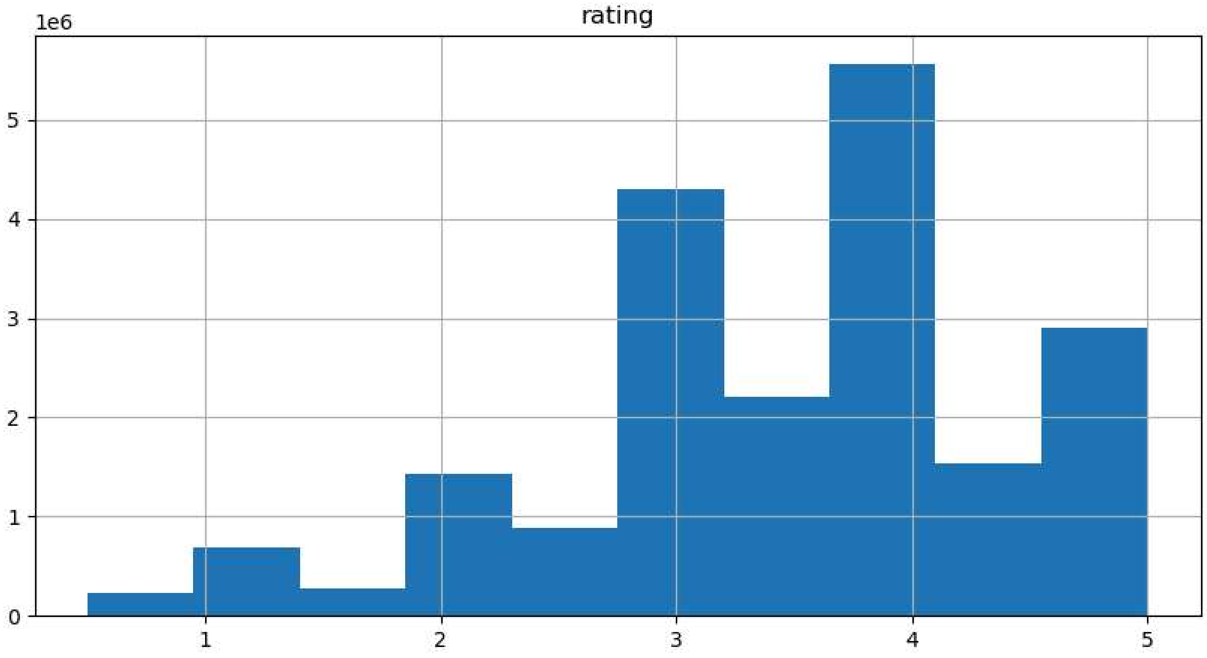
Data Visualiziation

In [46]:

```
%matplotlib inline
import matplotlib.pyplot as plt
```

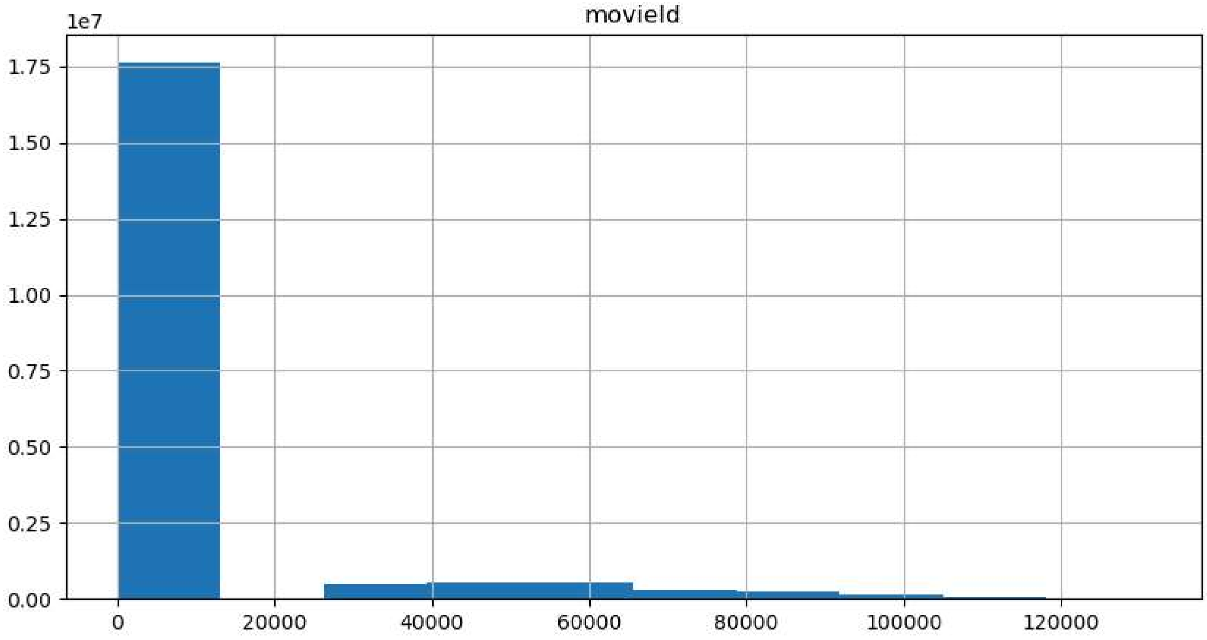
In [47]:

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



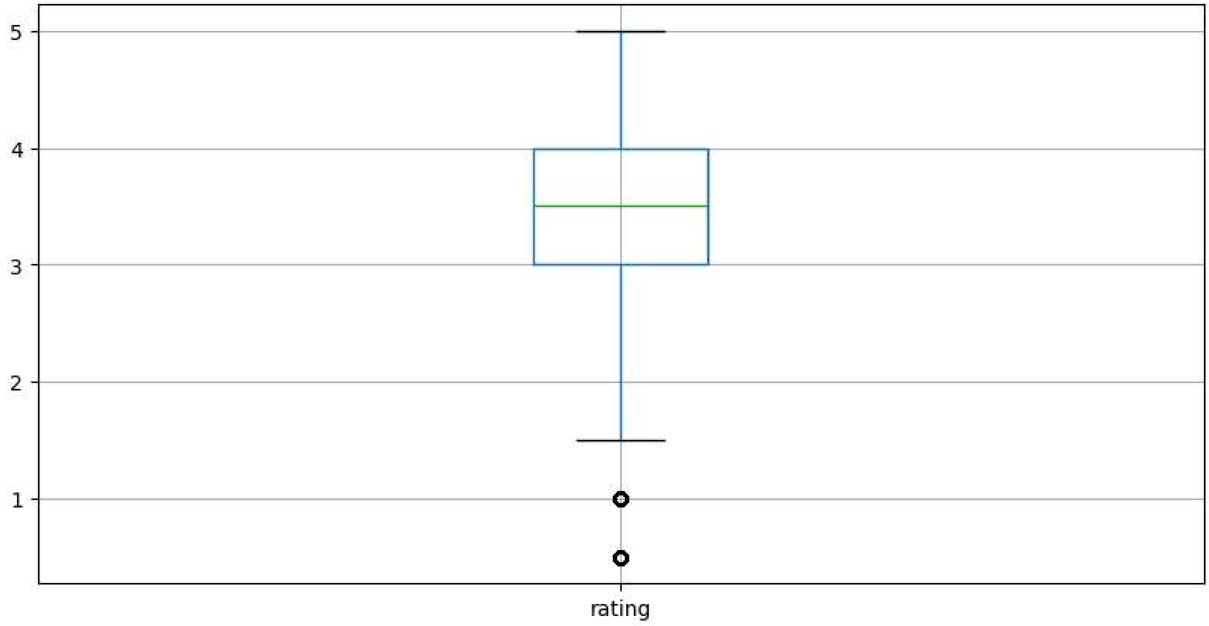
In [48]:

```
ratings.hist(column='movieId' ,figsize=(10,5))
plt.show()
```



In [49]:

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



slicing out columns

In [50]:

```
tags['tag'].head()
```

Out[50]:

```
0      Mark Waters
1        dark hero
2        dark hero
3    noir thriller
4        dark hero
Name: tag, dtype: object
```

In [51]:

```
movies[['movieId','genres']].head()
```

Out[51]:

	movieId	genres
0	1	Adventure|Animation|Children|Comedy|Fantasy
1	2	Adventure|Children|Fantasy
2	3	Comedy|Romance
3	4	Comedy|Drama|Romance
4	5	Comedy

In [52]:

```
ratings[['userId','movieId']].head()
```

Out[52]:

	userId	movieId
0	1	2
1	1	29
2	1	32
3	1	47
4	1	50

In [53]:

```
ratings[-10:]
```

Out[53]:

	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 [54]:

```
tag_count=tags['tag'].value_counts()
tag_count
```

Out[54]:

```
tag
sci-fi                              3384
based on a book                     3281
atmospheric                         2917
comedy                              2779
action                              2657
                                    ... 
Paul Adelstein                         1
the wig                                1
killer fish                            1
genetically modified monsters          1
topless scene                          1
Name: count, Length: 38643, dtype: int64
```

In [55]:

```
tag_count[2:5]
```

Out[55]:

```
tag
atmospheric    2917
comedy         2779
action         2657
Name: count, dtype: int64
```

In [56]:

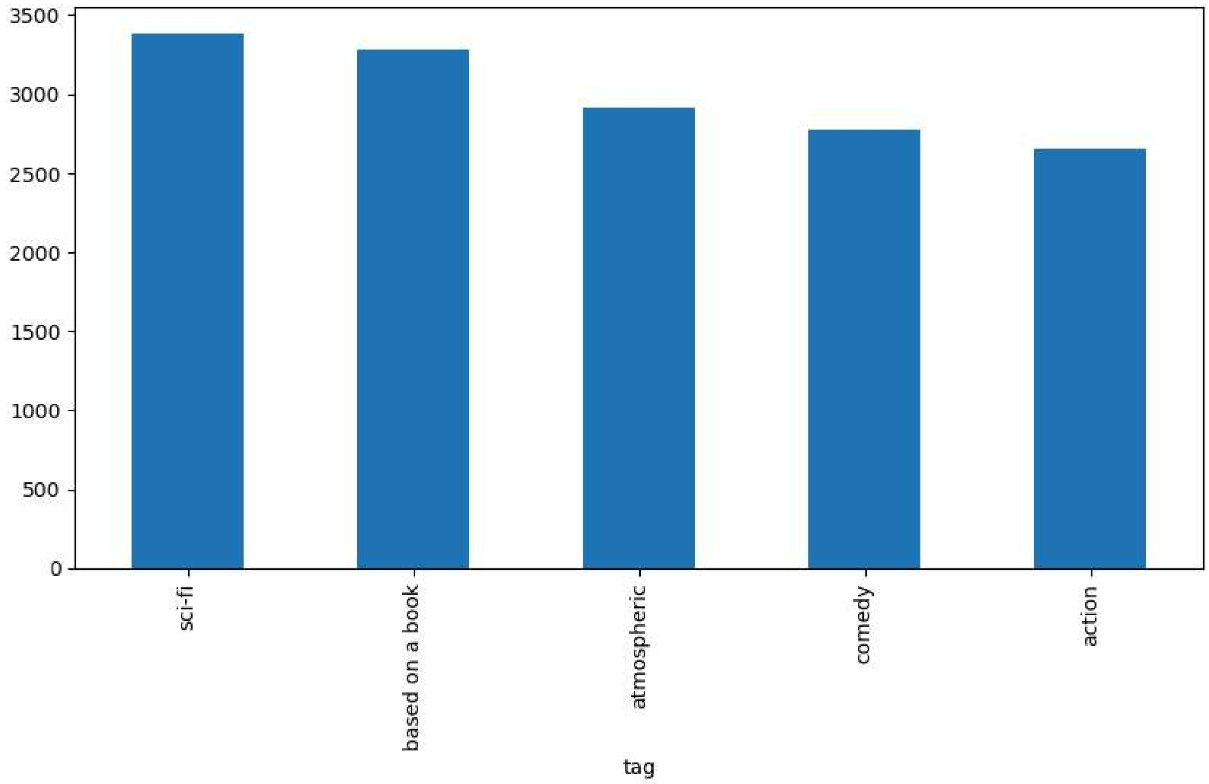
```
tags[2:5]
```

Out[56]:

	userId	movieId	tag
2	65	353	dark hero
3	65	521	noir thriller
4	65	592	dark hero

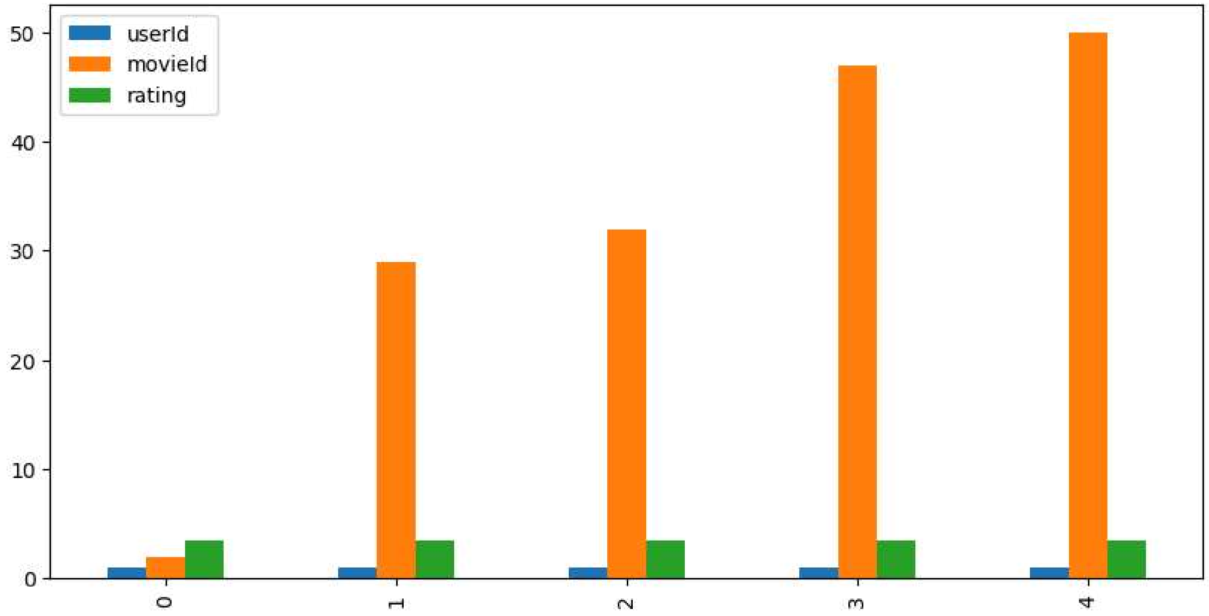
In [57]:

```
tag_count[:5].plot(kind='bar',figsize=(10,5))
plt.show()
```



In [58]:

```
ratings[:5].plot(kind='bar',figsize=(10,5))
plt.show()
```



```
In [ ]:
```