Movie recommendation system on movielens dataset

importing the libraries and dataset

In [10]:
```
import pandas as pd
import numpy as np
```

In [11]:
```
ds = pd.read_csv('F:/cdac/ml/ML-algo/recommendation_system/u.data', sep='\t', header=None)
ds.head()
```

Out[11]:
```
     0    1  2          3
0    0   50  5  881250949
1    0  172  5  881250949
2    0  133  1  881250949
3  196  242  3  881250949
4  186  302  3  891717742
```

In [12]:
```
ds.columns = ['user_id', 'item_id', 'rating', 'timestamp']
ds.head()
```

Out[12]:
```
   user_id  item_id  rating  timestamp
0        0       50       5  881250949
1        0      172       5  881250949
2        0      133       1  881250949
3      196      242       3  881250949
4      186      302       3  891717742
```

In [13]:
```
movie_titles = pd.read_csv('F:/cdac/ml/ML-algo/recommendation_system/Movie_Id_Titles')
movie_titles.head()
```

Out[13]:

	item_id	title
0	1	Toy Story (1995)
1	2	GoldenEye (1995)
2	3	Four Rooms (1995)
3	4	Get Shorty (1995)
4	5	Copycat (1995)

merging the dataframe with titles

In [14]:
```
ds = pd.merge(ds, movie_titles, on='item_id')
ds.head()
```

Out[14]:

	user_id	item_id	rating	timestamp	title
0	0	50	5	881250949	Star Wars (1977)
1	290	50	5	880473582	Star Wars (1977)
2	79	50	4	891271545	Star Wars (1977)
3	2	50	5	888552084	Star Wars (1977)
4	8	50	5	879362124	Star Wars (1977)

visualization of dataset

In [15]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

firstely we will find the mean ratings of each movie

In [16]:
```
ratings = pd.DataFrame(ds.groupby('title')['rating'].mean())
ratings.head()
```

Out[16]:

	rating
title	
'Til There Was You (1997)	2.333333
1-900 (1994)	2.600000
101 Dalmatians (1996)	2.908257
12 Angry Men (1957)	4.344000
187 (1997)	3.024390

now we will find by how many users a particular movie gets ratings

In [17]:
```
ratings['num_of_rating'] = ds.groupby('title')['rating'].count()
ratings.head()
```

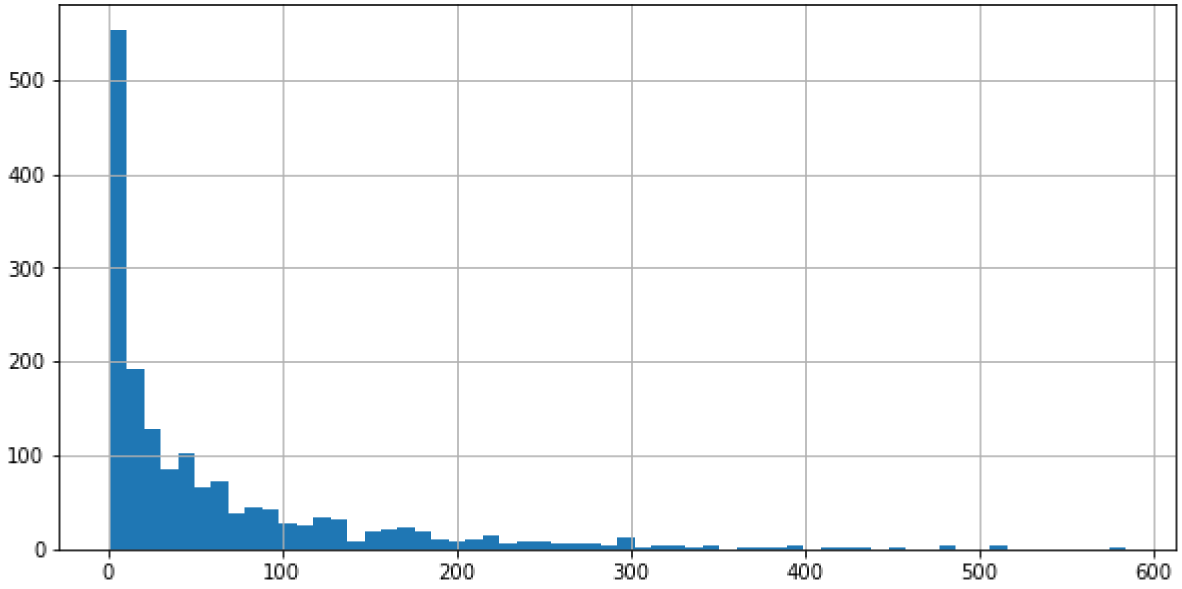
Out[17]:

	rating	num_of_rating
title		
'Til There Was You (1997)	2.333333	9
1-900 (1994)	2.600000	5
101 Dalmatians (1996)	2.908257	109
12 Angry Men (1957)	4.344000	125
187 (1997)	3.024390	41

now we will plot a histogram to analyse what is the ranging of

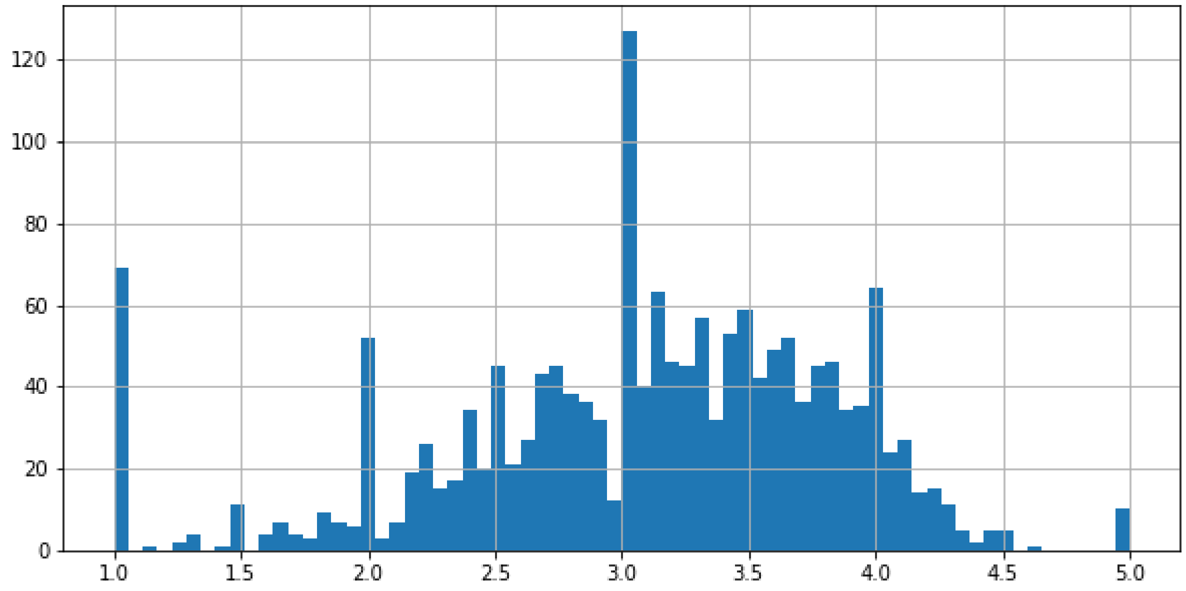
num_of_rating

```
In [26]: plt.figure(figsize=(10,5))
         ratings['num_of_rating'].hist(bins=60)
         plt.show()
```



from this graph we can see that there are only few movies which gets rating more then 500 users.

In [28]:
```
plt.figure(figsize=(10,5))
ratings['rating'].hist(bins=70)
plt.show()
```



from this graph we can see that mostly movies gets average rating between 2.0 to 4.0

Recommending similar movies using pivot table

In [30]:

```
movie_pivit = ds.pivot_table(index='user_id', columns='title', values = 'rating')
movie_pivit
```

Out[30]:

title	'Til There Was You (1997)	1-900 (1994)	101 Dalmatians (1996)	12 Angry Men (1957)	187 (1997)	2 Days in the Valley (1996)	20,000 Leagues Under the Sea (1954)	2001: A Space Odyssey (1968)	3 Ninjas: High Noon At Mega Mountain (1998)	39 Steps, The (1935)	...	Yankee Zulu (1994)	Year of the Horse (1997)	You So Crazy (1994)	Young Frankenstein (1974)	Young Guns (1988)	Y (
user_id																	
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	2.0	5.0	NaN	NaN	3.0	4.0	NaN	NaN	...	NaN	NaN	NaN	5.0	3.0	
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	...	NaN	NaN	NaN	NaN	NaN	
3	NaN	NaN	NaN	NaN	2.0	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	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
939	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
940	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
941	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
942	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.0	NaN	3.0	...	NaN	NaN	NaN	NaN	NaN	
943	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	4.0	

944 rows × 1664 columns

now suppose user watched a movie star wars . now we will look for higher coorelation and recommend those movies to the user

```
In [33]: star_wars_rating = movie_pivit['Star Wars (1977)']
         star_wars_rating
```

```
Out[33]: user_id
         0      5.0
         1      5.0
         2      5.0
         3      NaN
         4      5.0
               ...
         939    NaN
         940    4.0
         941    NaN
         942    5.0
         943    4.0
         Name: Star Wars (1977), Length: 944, dtype: float64
```

In [34]:

```
similar_to_star_wars = movie_pivit.corrwith(star_wars_rating)
similar_to_star_wars
```

```
C:\Users\hp\anaconda3\lib\site-packages\numpy\lib\function_base.py:2634: RuntimeWarning: Degrees of freedom <= 0 for sl
ice
  c = cov(x, y, rowvar, dtype=dtype)
C:\Users\hp\anaconda3\lib\site-packages\numpy\lib\function_base.py:2493: RuntimeWarning: divide by zero encountered in
true_divide
  c *= np.true_divide(1, fact)
```

Out[34]:

```
title
'Til There Was You (1997)                0.872872
1-900 (1994)                            -0.645497
101 Dalmatians (1996)                    0.211132
12 Angry Men (1957)                      0.184289
187 (1997)                               0.027398
                                           ...
Young Guns II (1990)                     0.228615
Young Poisoner's Handbook, The (1995)   -0.007374
Zeus and Roxanne (1997)                  0.818182
unknown                                  0.723123
Á köldum klaka (Cold Fever) (1994)            NaN
Length: 1664, dtype: float64
```

converting it into a dataframe

In [36]:

```
star_wars_corr = pd.DataFrame(similar_to_star_wars, columns=['correlation'])
star_wars_corr.head()
```

Out[36]:

	correlation
title	
'Til There Was You (1997)	0.872872
1-900 (1994)	-0.645497
101 Dalmatians (1996)	0.211132
12 Angry Men (1957)	0.184289
187 (1997)	0.027398

In [37]:

```
star_wars_corr['num_of_ratings'] = ratings['num_of_rating']
star_wars_corr
```

Out[37]:

	correlation	num_of_ratings
title		
'Til There Was You (1997)	0.872872	9
1-900 (1994)	-0.645497	5
101 Dalmatians (1996)	0.211132	109
12 Angry Men (1957)	0.184289	125
187 (1997)	0.027398	41
...	...	...
Young Guns II (1990)	0.228615	44
Young Poisoner's Handbook, The (1995)	-0.007374	41
Zeus and Roxanne (1997)	0.818182	6
unknown	0.723123	9
Á köldum klaka (Cold Fever) (1994)	NaN	1

1664 rows × 2 columns

now we will recommend movie to the user which gets rating more then 100 users

In [39]:

```
star_wars_corr = star_wars_corr[star_wars_corr['num_of_ratings']>100].sort_values('correlation', ascending=False)
star_wars_corr.head()
```

Out[39]:

	correlation	num_of_ratings
title		
Star Wars (1977)	1.000000	584
Empire Strikes Back, The (1980)	0.748353	368
Return of the Jedi (1983)	0.672556	507
Raiders of the Lost Ark (1981)	0.536117	420
Austin Powers: International Man of Mystery (1997)	0.377433	130

so the next recommended movie will be Empire Strikes Back, The (1980)

In [ ]: