In [1]:

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

In [2]:

```
column_names = ['user_id', 'item_id', 'rating', 'timestamp']
df = pd.read_csv('u.data', sep='\t', names=column_names)
df.head()
```

Out[2]:

	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 [3]:

```
movie_titles = pd.read_csv("Movie_Id_Titles")
movie_titles.head()
```

Out[3]:

	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)

In [4]:

```
df = pd.merge(df,movie_titles,on='item_id')
df.head()
```

Out[4]:

	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)

In [5]:

```
df.groupby('title')['rating'].mean().sort_values(ascending=False).head()
```

Out[5]:

```
title
Marlene Dietrich: Shadow and Light (1996)     5.0
Prefontaine (1997)                            5.0
Santa with Muscles (1996)                     5.0
Star Kid (1997)                               5.0
Someone Else's America (1995)                 5.0
Name: rating, dtype: float64
```

In [6]:

```
df.groupby('title')['rating'].count().sort_values(ascending=False).head()
```

Out[6]:

```
title
Star Wars (1977)              584
Contact (1997)                509
Fargo (1996)                  508
Return of the Jedi (1983)     507
Liar Liar (1997)              485
Name: rating, dtype: int64
```

In [7]:

```
ratings = pd.DataFrame(df.groupby('title')['rating'].mean())
ratings.head()
```

Out[7]:

	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

In [8]:

```
ratings['num of ratings'] = pd.DataFrame(df.groupby('title')['rating'].count())
ratings.head()
```

Out[8]:

	rating	num of ratings
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

In [9]:

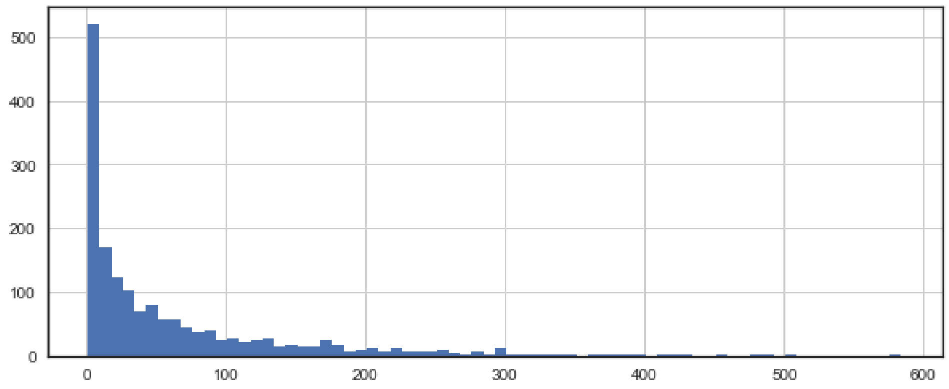
```
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('white')
%matplotlib inline
```

In [10]:

```
plt.figure(figsize=(10,4))
ratings['num of ratings'].hist(bins=70)
```

Out[10]:

```
<matplotlib.axes._subplots.AxesSubplot at 0xc0912f0>
```

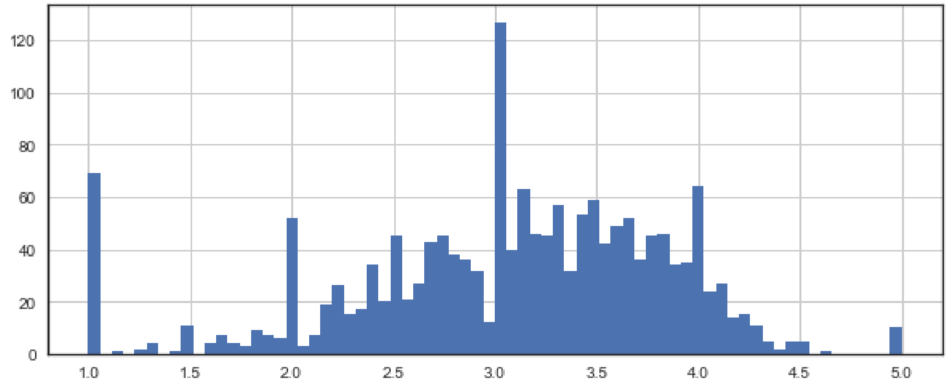


In [11]:

```
plt.figure(figsize=(10,4))
ratings['rating'].hist(bins=70)
```

Out[11]:

```
<matplotlib.axes._subplots.AxesSubplot at 0xd2f26d0>
```



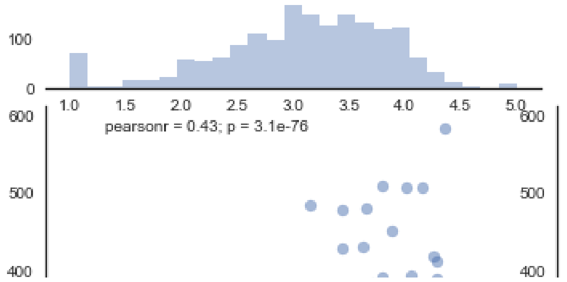
In [12]:

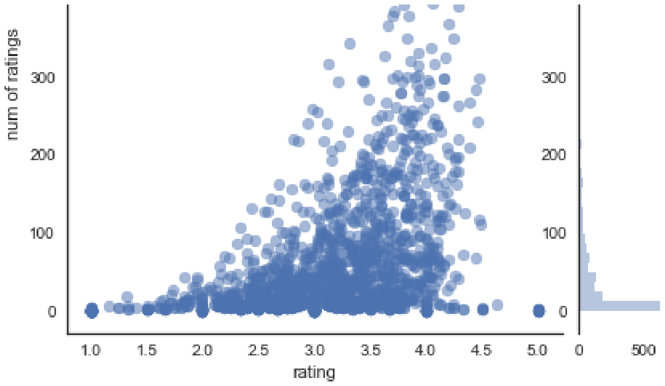
```
sns.jointplot(x='rating',y='num of ratings',data=ratings,alpha=0.5)
```

```
C:\Users\MY PC\Anaconda3\lib\site-packages\seaborn\distributions.py:215:
MatplotlibDeprecationWarning:
The 'normed' kwarg was deprecated in Matplotlib 2.1 and will be removed in 3.1. Use 'density' inst
ead.
  color=hist_color, **hist_kws)
```

Out[12]:

```
<seaborn.axisgrid.JointGrid at 0xd225f30>
```





In [13]:

```
moviemat = df.pivot_table(index='user_id',columns='title',values='rating')
moviemat.head()
```

Out[13]:

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)	Y F ('
user_id															
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	N
1	NaN	NaN	2.0	5.0	NaN	NaN	3.0	4.0	NaN	NaN	...	NaN	NaN	NaN	5
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	...	NaN	NaN	NaN	N
3	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	N
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	N

5 rows × 1664 columns

In [14]:

```
starwars_user_ratings = moviemat['Star Wars (1977)']
starwars_user_ratings.head()
```

Out[14]:

```
user_id
0    5.0
1    5.0
2    5.0
3    NaN
4    5.0
Name: Star Wars (1977), dtype: float64
```

In [15]:

```
similar_to_starwars = moviemat.corrwith(starwars_user_ratings)
```

```
C:\Users\MY PC\Anaconda3\lib\site-packages\numpy\lib\function_base.py:2995: RuntimeWarning:
Degrees of freedom <= 0 for slice
  c = cov(x, y, rowvar)
C:\Users\MY PC\Anaconda3\lib\site-packages\numpy\lib\function_base.py:2929: RuntimeWarning: divide
by zero encountered in double_scalars
  c *= 1. / np.float64(fact)
```

In [16]:

```
corr_starwars = pd.DataFrame(similar_to_starwars,columns=['Correlation'])
corr_starwars.dropna(inplace=True)
corr_starwars.head()
```

Out[16]:

	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 [17]:

```
corr_starwars = corr_starwars.join(ratings['num of ratings'])
corr_starwars.head()
```

Out[17]:

	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

In [19]:

```
corr_starwars[corr_starwars['num of ratings']>100].sort_values('Correlation',ascending=False).head
()
```

Out[19]:

	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

In [ ]: