

# Shubham Shashikant Bawane **Film Recommendation System** using numpy, pandas, seaborn Language python

```
[1]: import numpy as np
import pandas as pd
import seaborn as sns
```

```
[4]: movie_titles = pd.read_csv("Movie_Id_Titles.data")
movie_titles.head()
```

```
[4]:
```

	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)

```
[8]: df = pd.merge(df, movie_titles, on= 'item_id')
df.head()
```

```
[8]:
```

	user_id	item_id	rating	timestamp	title
0	0	50	5	881250949	Runway 36 (1977)
1	0	172	5	881250949	Empire Strikes Back, The (1980)
2	0	133	1	881250949	Gone with the Wind (1939)
3	196	242	3	881250949	Kolya (1996)
4	186	302	3	891717742	L.A. Confidential (1997)

```
[12]: df = df.drop('title_x', axis=1)
```

```
[13]: df
```

```
[13]:
```

	user_id	item_id	rating	timestamp	title_y
0	0	50	5	881250949	Runway 36 (1977)
1	0	172	5	881250949	Empire Strikes Back, The (1980)
2	0	133	1	881250949	Gone with the Wind (1939)
3	196	242	3	881250949	Kolya (1996)
4	186	302	3	891717742	L.A. Confidential (1997)
...	...	...	...	...	...
99998	880	476	3	880175444	First Wives Club, The (1996)
99999	716	204	5	879795543	Back to the Future (1985)
100000	276	1090	1	874795795	Sliver (1993)
100001	13	225	2	882399156	101 Dalmatians (1996)
100002	12	203	3	879959583	Unforgiven (1992)

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

```
[3]: df.head()
```

```
[3]:
```

	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

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

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

title_y	rating
Aiqing wansui (1994)	5.0
Entertaining Angels: The Dorothy Day Story (1996)	5.0
Saint of Fort Washington, The (1993)	5.0
They Made Me a Criminal (1939)	5.0
Prefontaine (1997)	5.0

Name: rating, dtype: float64

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

title_y	rating
Runway 36 (1977)	584
Contact (1997)	509
Fargo (1996)	508
Return of the Jedi (1983)	507
Liar Liar (1997)	485

Name: rating, dtype: int64

```
[18]: ratings['num of ratings'] = pd.DataFrame(df.groupby('title_y')['rating'].count())
ratings
```

```
[18]:
```

	rating	num of ratings
title_y		
'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
...	...	...
Young Guns II (1990)	2.772727	44
Young Poisoner's Handbook, The (1995)	3.341463	41
Zeus and Roxanne (1997)	2.166667	6
unknown	3.444444	9
Á köldum klaka (Cold Fever) (1994)	3.000000	1

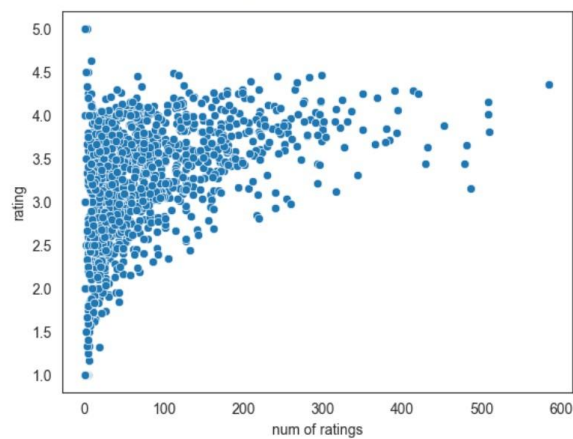
```
[17]: ratings = pd.DataFrame(df.groupby('title_y')['rating'].mean())
ratings
```

```
[17]:
```

	rating
title_y	
'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
...	...
Young Guns II (1990)	2.772727
Young Poisoner's Handbook, The (1995)	3.341463
Zeus and Roxanne (1997)	2.166667
unknown	3.444444
Á köldum klaka (Cold Fever) (1994)	3.000000

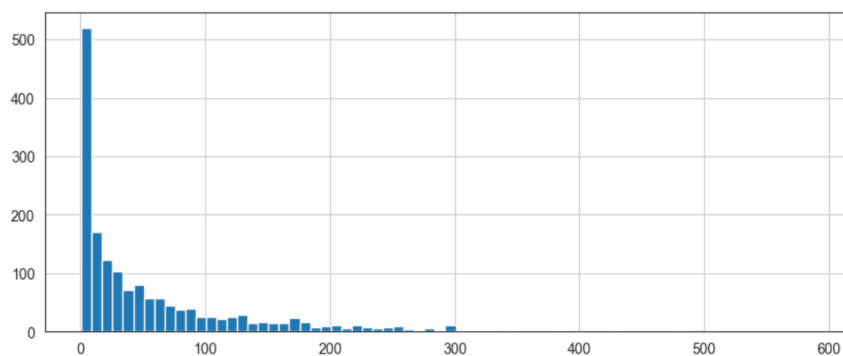
```
[27]: sns.scatterplot(data = ratings , x = "num of ratings", y = "rating")
```

```
[27]: <Axes: xlabel='num of ratings', ylabel='rating'>
```



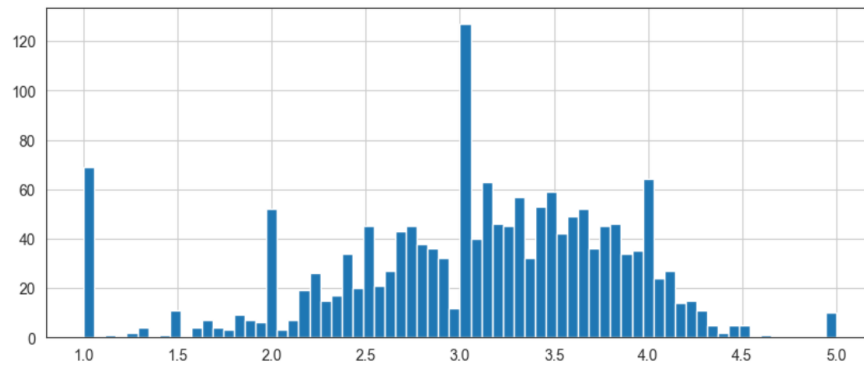
```
[28]: plt.figure(figsize=(10,4))
ratings['num of ratings'].hist(bins=70)
```

```
[28]: <Axes: >
```



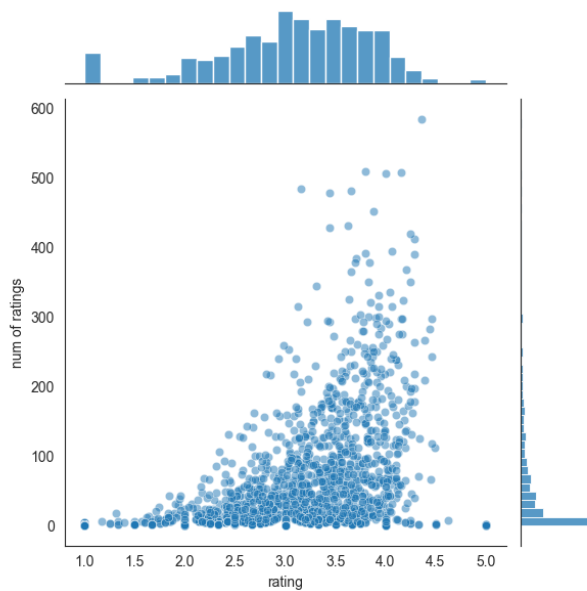
```
9]: plt.figure(figsize=(10,4))
ratings['rating'].hist(bins=70)
```

```
9]: <Axes: >
```



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

```
[32]: <seaborn.axisgrid.JointGrid at 0x1b9a392c800>
```



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

```
[36]:
```

title_y	'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)	Young Guns II (1990)	Young Poisoner's Handbook The (1995)
user_id																		
0	NaN	NaN	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	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	2.0	NaN	NaN	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	NaN	NaN

5 rows × 1664 columns

```
[37]: ratings.sort_values('num of ratings',ascending = False).head(10)
```

```
[37]:
```

	rating	num of ratings
title_y		
Runway 36 (1977)	4.359589	584
Contact (1997)	3.803536	509
Fargo (1996)	4.155512	508
Return of the Jedi (1983)	4.007890	507
Liar Liar (1997)	3.156701	485
English Patient, The (1996)	3.656965	481
Scream (1996)	3.441423	478
Toy Story (1995)	3.878319	452
Air Force One (1997)	3.631090	431
Independence Day (ID4) (1996)	3.438228	429

```
[38]: ratings.head()
```

```
[38]:
```

	rating	num of ratings
title_y		
'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

```
[39]: runway_user_ratings = moviemat['Runway 36 (1977)']
liarliar_user_ratings = moviemat['Liar Liar (1997)']
runway_user_ratings.head()
```

```
[39]: user_id
0    5.0
1    5.0
2    5.0
3    NaN
4    5.0
Name: Runway 36 (1977), dtype: float64
```

```
[42]: similar_to_runway = moviemat.corrwith(runway_user_ratings)
similar_to_liarliar = moviemat.corrwith(liarliar_user_ratings)
```

```

3]: corr_runway = pd.DataFrame(similar_to_runway, columns=['Correlation'])
corr_runway.dropna(inplace=True)
corr_runway.head()

```

```

3]:

```

Correlation	
title_y	
'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

```

1]: corr_runway.sort_values('Correlation', ascending=False).head(10)

```

```

1]:

```

Correlation	
title_y	
Runway 36 (1977)	1.0
Stripes (1981)	1.0
Hollow Reed (1996)	1.0
Man of the Year (1995)	1.0
Cosi (1996)	1.0
Commandments (1997)	1.0
No Escape (1994)	1.0
Golden Earrings (1947)	1.0
Scarlet Letter, The (1926)	1.0
Safe Passage (1994)	1.0

```

[45]: corr_runway= corr_runway.join(ratings['num of ratings'])
corr_runway.head()

```

```

[45]:

```

Correlation num of ratings		
title_y		
'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

```

[46]: corr_runway[corr_runway['num of ratings']>100].sort_values('Correlation', ascending=False).head()

```

```

[46]:

```

Correlation num of ratings		
title_y		
Runway 36 (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

```
[47]: corr_liarliar = pd.DataFrame(similar_to_liarliar,columns=['Correlation'])
corr_liarliar.dropna(inplace=True)
corr_liarliar = corr_liarliar.join(ratings['num of ratings'])
corr_liarliar[corr_liarliar['num of ratings']>100].sort_values('Correlation',ascending=False).head()
```

```
[47]:
```

	Correlation	num of ratings
title_y		
Liar Liar (1997)	1.000000	485
Batman Forever (1995)	0.516968	114
Mask, The (1994)	0.484650	129
Down Periscope (1996)	0.472681	101
Con Air (1997)	0.469828	137

```
[ ]:
```