In [222]:
```
import numpy as np
import pandas as pd
import matplotlib as plt
import seaborn as sns
from IPython.display import display, clear_output
```

In [2]:
```
data=pd.read_csv("ratings_Movies_and_TV.csv")
```

In [91]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4607046 entries, 0 to 4607045
Data columns (total 4 columns):
user         object
movie        object
rating       float64
timestamp    int64
dtypes: float64(1), int64(1), object(2)
memory usage: 140.6+ MB
```

In [5]:
```
data.columns=['user','movie','rating','timestamp']
```

In [202]:
```
data.tail()
```

Out[202]:

	user	movie	rating	timestamp
4607041	AHCV1RTGY3PJ8	B00LT1JHLW	5.0	1405641600
4607042	A2RWCXDMANY0LW	B00LT1JHLW	5.0	1405987200
4607043	A3V9PIFRME2XCW	B00LT1JHLW	5.0	1405900800
4607044	A3ROPC55BE2OM9	B00LT1JHLW	5.0	1405728000
4607045	A2ARBNMH5Q5YM1	B00LVGP8EA	5.0	1405641600

In [74]:
```
# to extract top "length" user,movies those have high frequency

X=data["user"].value_counts()    # pandas series return
Y=data['movie'].value_counts()   # pandas series return
```

In [219]:
```
 # extract the index of the series that is either user id or movie id
length=50
users=X.index
movies=Y.index
users=user[:length]
movies=movies[:length]
```

In [220]:
```
# this is the Rating matrix that contain column is user and rows is movies
"""     | movie1 | movie2 | movie3
 user1  |   5    |    4   |    1
 user2  |   3    |        |    5
 user3  |   5    |    3   |

"""
# the R matrix is boolean matrix that denote user rated that movie or not

Rating=[[0 for _ in range(length)] for _ in range(length)]
R=[[0 for _ in range(length)] for _ in range(length)]
```

In [223]:
```
# update the both matrix Rating and R
for i in range(length):
    for j in range(length):
        a=data[(data['user']==users[i]) & (data['movie']==movies[j])]
        if a.empty!=True:
            Rating[i][j]=list(a['rating'])[0]
            R[i][j]=1
        clear_output(wait=True)
        print ("percentage = ",i*j/25)
```

```
percentage =  96.04
```

In [232]:
```
DataRating=pd.DataFrame(Rating)
DataR=pd.DataFrame(R)
```

In [235]:
```
DataRating.head()
```

Out[235]:

	0	1	2	3	4	5	6	7	8	9	...	40	41	42	43	44	45	46	47	48	49
0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	5.0	4.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	5.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	3.0	...	4.0	4.0	4.0	0.0	0.0	5.0	0.0	4.0	4.0	5.0
3	0.0	2.0	4.0	0.0	0.0	4.0	0.0	0.0	0.0	3.0	...	3.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	3.0	0.0
4	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 50 columns

In [234]:
```
1 DataR.head()
```

Out[234]:

	0	1	2	3	4	5	6	7	8	9	...	40	41	42	43	44	45	46	47	48	49
0	0	1	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	1	1	1
1	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	1	0	0	0	1	...	1	1	1	0	0	1	0	1	1	1
3	0	1	1	0	0	1	0	0	0	1	...	1	0	0	0	0	0	0	1	1	0
4	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 50 columns

In [236]:
```
1 sum(DataR)
```

Out[236]: 1225

In [237]:
```
1 DataRating.to_csv('RATING.csv')
2 DataR.to_csv('R.csv')
```

In [ ]:
```
1
```

In [ ]:
```
1
```