In [2]:

```
import pandas as pd
pd.set_option('max_columns',None)
pd.set_option('max_rows',None)

import numpy as np

import warnings
warnings.filterwarnings('ignore')

import matplotlib.pyplot as plt
```

In [3]:

```
#importing
movies = pd.read_csv('movies.dat' , delimiter = '::' , names = ['movieid' , 'title' , 'genres'],encoding='latin-1')
ratings = pd.read_csv('ratings.dat' , delimiter = '::' , names = ['userid' , 'movieid' , 'rating' , 'timestamp'],encoding='latin-1' )
users = pd.read_csv('users.dat' , delimiter = '::' , names = ['userid' , 'gender' , 'age' ,'occupation' ,'zip_code'],encoding='latin-1')
```

In [4]:

```
ratings.head(1)
```

Out[4]:

	userid	movieid	rating	timestamp
0	1	1193	5	978300760

In [5]:

```
movies.head(1)
```

Out[5]:

	movieid	title	genres
0	1	Toy Story (1995)	Animation|Children's|Comedy

In [6]:

```
users.head(1)
```

Out[6]:

	userid	gender	age	occupation	zip_code
0	1	F	1	10	48067

In [7]:

```
# merging

df1 = pd.merge(users.drop('zip_code',axis = 1) , ratings.drop('timestamp',axis = 1) ,on = 'userid')

master_data = pd.merge(df1 , movies , on = 'movieid')
```

In [8]:

```
master_data.head()
```

Out[8]:

	userid	gender	age	occupation	movieid	rating	title	genres
0	1	F	1	10	1193	5	One Flew Over the Cuckoo's Nest (1975)	Drama
1	2	M	56	16	1193	5	One Flew Over the Cuckoo's Nest (1975)	Drama
2	12	M	25	12	1193	4	One Flew Over the Cuckoo's Nest (1975)	Drama
3	15	M	25	7	1193	4	One Flew Over the Cuckoo's Nest (1975)	Drama
4	17	M	50	1	1193	5	One Flew Over the Cuckoo's Nest (1975)	Drama

In [9]:

```
master_data.shape
```

Out[9]:

```
(1000209, 8)
```

In [10]:

```
master_data.age.value_counts(normalize = True).to_frame()
```

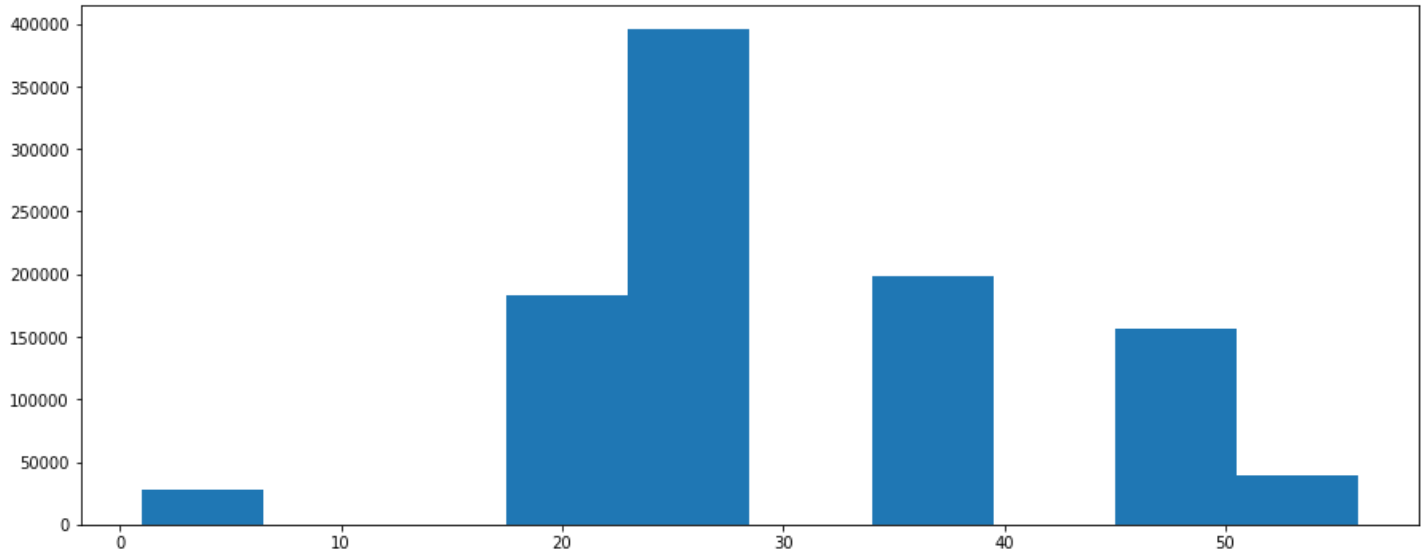
Out[10]:

	age
25	0.395473
35	0.198961
18	0.183498
45	0.083616
50	0.072475
56	0.038772
1	0.027205

In [12]:

```
# 'under 18' : 1, '18-24' : 18, '25-34' : 25, '35-44' : 35, '45-49' : 45, '50-55' : 50, '
56+' : 56.

age_fig = plt.figure(figsize = (15,6))
plt.hist(master_data.age);
```



In [25]:

```
#  Top 25 movies by viewership rating

movie_ratings = master_data[['title','rating']].groupby('title').mean().sort_values('rat
```

```
ing',ascending = False)
movie_ratings.head(25)
```

Out[25]:

title	rating
Ulysses (Ulisse) (1954)	5.000000
Lured (1947)	5.000000
Follow the Bitch (1998)	5.000000
Bittersweet Motel (2000)	5.000000
Song of Freedom (1936)	5.000000
One Little Indian (1973)	5.000000
Smashing Time (1967)	5.000000
Schlafes Bruder (Brother of Sleep) (1995)	5.000000
Gate of Heavenly Peace, The (1995)	5.000000
Baby, The (1973)	5.000000
I Am Cuba (Soy Cuba/Ya Kuba) (1964)	4.800000
Lamerica (1994)	4.750000
Apple, The (Sib) (1998)	4.666667
Sanjuro (1962)	4.608696
Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)	4.560510
Shawshank Redemption, The (1994)	4.554558
Godfather, The (1972)	4.524966
Close Shave, A (1995)	4.520548
Usual Suspects, The (1995)	4.517106
Schindler's List (1993)	4.510417
Wrong Trousers, The (1993)	4.507937
Dry Cleaning (Nettoyage à sec) (1997)	4.500000
Inheritors, The (Die Siebtelbauern) (1998)	4.500000
Mamma Roma (1962)	4.500000
Bells, The (1926)	4.500000

In [26]:

```
# User rating of the movie "Toy Story"/
movie_ratings.loc[['Toy Story (1995)']]
```

Out[26]:

title	rating
Toy Story (1995)	4.146846

In [33]:

```
# the ratings for all the movies reviewed by for a particular user of user id = 2696
user2696 = master_data[master_data['userid']==2696]
user2696['title']
```

Out[33]:

```
24345                          Back to the Future (1985)
29848                 E.T. the Extra-Terrestrial (1982)
```

```
244232                          L.A. Confidential (1997)
250014                                  Lone Star (1996)
273633                                       JFK (1991)
277808                    Talented Mr. Ripley, The (1999)
371178    Midnight in the Garden of Good and Evil (1997)
377250                                   Cop Land (1997)
598042                                   Palmetto (1998)
603189                           Perfect Murder, A (1998)
609204                                  Game, The (1997)
611956            I Know What You Did Last Summer (1997)
612552                      Devil's Advocate, The (1997)
613486                                      Psycho (1998)
616546                                 Wild Things (1998)
618708                               Basic Instinct (1992)
621101                                  Lake Placid (1999)
689379                                Shining, The (1980)
697451      I Still Know What You Did Last Summer (1998)
777089                                  Client, The (1994)
Name: title, dtype: object
```

In [41]:

```
# unique genres
movies.col_genres = movies.genres.apply(lambda x: str(x).split('|'))
unique_genres = []
for i in movies.col_genres:
  unique_genres.extend(i)

set(unique_genres)
```

Out[41]:

```
{'Action',
 'Adventure',
 'Animation',
 "Children's",
 'Comedy',
 'Crime',
 'Documentary',
 'Drama',
 'Fantasy',
 'Film-Noir',
 'Horror',
 'Musical',
 'Mystery',
 'Romance',
 'Sci-Fi',
 'Thriller',
 'War',
 'Western'}
```

In [44]:

```
# genre category with a one-hot encoding ( 1 and 0)
unique_genres = set(unique_genres)
for i in unique_genres:
  master_data[i] = master_data.genres.apply(lambda x: 1 if i in x else 0)

master_data.head()
```

Out[44]:

	userid	gender	age	occupation	movieid	rating	title	genres	Comedy	Western	War	Romance	Crime	Horror	M
0	**1**	**F**	**1**	**10**	**1193**	**5**	**One Flew Over the Cuckoo's Nest (1975)**	**Drama**	**0**	**0**	**0**	**0**	**0**	**0**	
							One Flew Over the								

	1	2	M	56	16	1193	5	Cuckoo's title Nest (1975)	Drama	0	0	0	0	0	0	M
	userid	gender	age	occupation	movieid	rating		genres	Comedy	Western	War	Romance	Crime	Horror		
2	12	M	25	12	1193	4	One Flew Over the Cuckoo's Nest (1975)	Drama	0	0	0	0	0	0		
3	15	M	25	7	1193	4	One Flew Over the Cuckoo's Nest (1975)	Drama	0	0	0	0	0	0		
4	17	M	50	1	1193	5	One Flew Over the Cuckoo's Nest (1975)	Drama	0	0	0	0	0	0		

In [46]:

```
# features affecting the ratings of any particular movie
master_data.corr()[['rating']]
```

Out[46]:

	rating
userid	0.012303
age	0.056869
occupation	0.006753
movieid	-0.064042
rating	1.000000
Comedy	-0.039622
Western	0.007311
War	0.075688
Romance	0.009644
Crime	0.033446
Horror	-0.094353
Musical	0.015643
Drama	0.122561
Adventure	-0.036718
Thriller	-0.004806
Sci-Fi	-0.044487
Fantasy	-0.023312
Documentary	0.028098
Mystery	0.015848
Animation	0.019670
Film-Noir	0.060259
Action	-0.047633
Children's	-0.039829

In [47]:

```
# appropriate model to predict the movie ratings
```

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

In [51]:

```
x=master_data[['Action',
 'Adventure',
 'Animation',
 "Children's",
 'Comedy',
 'Crime',
 'Documentary',
 'Drama',
 'Fantasy',
 'Film-Noir',
 'Horror',
 'Musical',
 'Mystery',
 'Romance',
 'Sci-Fi',
 'Thriller',
 'War',
 'Western']]
```

In [52]:

```
y=master_data['rating']
```

In [53]:

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.2)
```

In [55]:

```
model=LinearRegression()
```

In [56]:

```
model.fit(x_train,y_train)
```

Out[56]:

```
LinearRegression()
```

In [61]:

```
model.predict(x_test)
```

Out[61]:

```
array([3.22652124, 3.48074737, 3.4866866 , ..., 3.38635751, 3.37031895,
       3.73575021])
```