movielens

June 22, 2023

[1]:
```
 #importing pandas dataframe
import pandas as pd

import warnings
warnings.filterwarnings('ignore')

#importing seaborn
import seaborn as sns

#importing pandas profiling
import pandas_profiling as pf

#importing matplolib
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline

#importing reg-ex
import re
```

[2]:
```
#Hold out method for splitting data
from sklearn.model_selection import train_test_split

#importing accuracy_score
from sklearn.metrics import accuracy_score

#importing xgboost
import xgboost
```

[3]:
```
pip install lightgbm
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: lightgbm in ./.local/lib/python3.7/site-packages
(3.3.5)
Requirement already satisfied: wheel in /usr/local/lib/python3.7/site-packages
(from lightgbm) (0.34.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/site-packages
(from lightgbm) (1.4.1)
```

```
Requirement already satisfied: numpy in /usr/local/lib/python3.7/site-packages
(from lightgbm) (1.21.5)
Requirement already satisfied: scikit-learn!=0.22.0 in
/usr/local/lib/python3.7/site-packages (from lightgbm) (1.0.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.7/site-packages (from scikit-learn!=0.22.0->lightgbm)
(2.2.0)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/site-
packages (from scikit-learn!=0.22.0->lightgbm) (0.14.1)
WARNING: You are using pip version 22.0.3; however, version 23.1.2 is
available.
You should consider upgrading via the '/usr/local/bin/python3.7 -m pip install
--upgrade pip' command.
Note: you may need to restart the kernel to use updated packages.
```

[4]:
```
from lightgbm import LGBMClassifier
```

[5]:
```
# Importing datasets
rating = ['UserID','MovieID','Rating','Timestamp']
user = ['UserID','Gender','Age','Occupation','Zip-code']
movie = ['MovieID','Title','Genres']
```

[6]:
```
rating_df = pd.read_csv('ratings.dat',header=None,delimiter='::',names=rating)
print(rating_df.head())
print()
print(rating_df.shape)
```

```
   UserID  MovieID  Rating  Timestamp
0       1     1193       5  978300760
1       1      661       3  978302109
2       1      914       3  978301968
3       1     3408       4  978300275
4       1     2355       5  978824291

(218508, 4)
```

[7]:
```
user_df = pd.read_csv('users.dat',header=None,delimiter='::',names=user)
print(user_df.head())
print()
print(user_df.shape)
```

```
   UserID Gender  Age  Occupation Zip-code
0       1      F    1          10    48067
1       2      M   56          16    70072
2       3      M   25          15    55117
3       4      M   45           7    02460
```

```
4       5      M   25          20     55455

(6040, 5)
```

[8]:
```
movie_df = pd.read_csv('movies.dat',header=None,delimiter='::',names=movie)
print(movie_df.head())
print()
print(movie_df.shape)
```

```
   MovieID                               Title                        Genres
0        1                                None                          None
1        2                    Jumanji (1995)  Adventure|Children's|Fantasy
2        3           Grumpier Old Men (1995)                Comedy|Romance
3        4          Waiting to Exhale (1995)                  Comedy|Drama
4        5  Father of the Bride Part II (1995)                        Comedy

(3883, 3)
```

[9]:
```
#Merging three datasets

df = rating_df.merge(user_df,how='outer',on='UserID')
df = df.merge(movie_df,how='outer',on='MovieID')
df.head()
```

[9]:

	UserID	MovieID	Rating	Timestamp	Gender	Age	Occupation	Zip-code	Title	Genres
0	1.0	1193.0	5.0	978300760.0	F	1.0	10.0	48067	One Flew Over the Cuckoo's Nest (1975)	Drama
1	2.0	1193.0	5.0	978298413.0	M	56.0	16.0	70072	One Flew Over the Cuckoo's Nest (1975)	Drama
2	12.0	1193.0	4.0	978220179.0	M	25.0	12.0	32793	One Flew Over the Cuckoo's Nest (1975)	Drama
3	15.0	1193.0	4.0	978199279.0	M	25.0	7.0	22903	One Flew Over the Cuckoo's Nest (1975)	Drama
4	17.0	1193.0	5.0	978158471.0	M	50.0	1.0	95350	One Flew Over the Cuckoo's Nest (1975)	Drama

[10]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 223625 entries, 0 to 223624
Data columns (total 10 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   UserID      223223 non-null  float64
 1   MovieID     218910 non-null  float64
 2   Rating      218508 non-null  float64
```

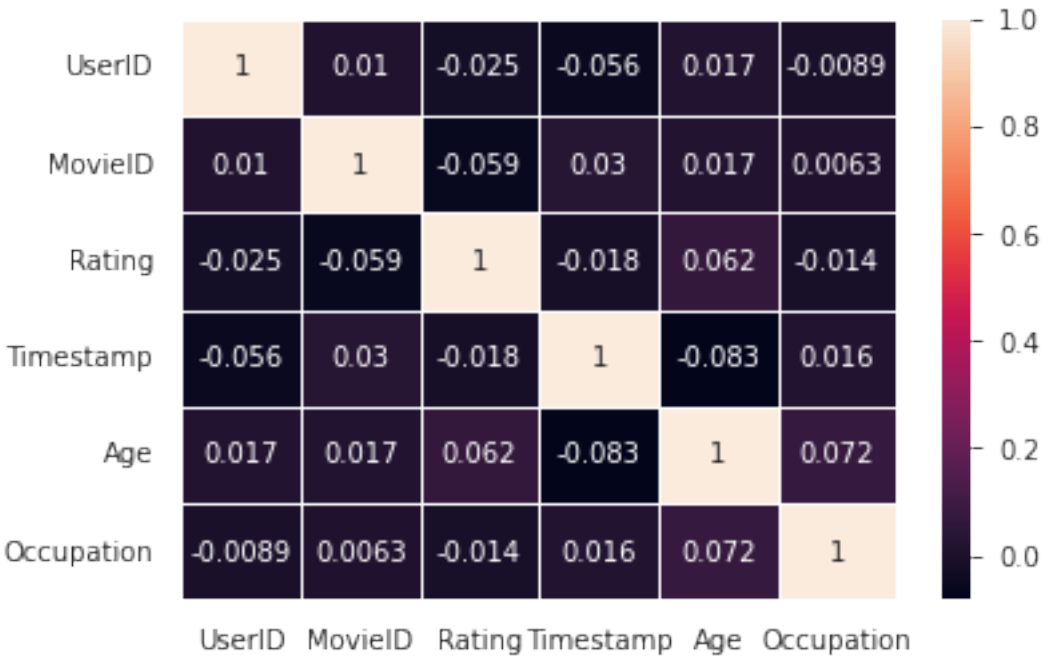
```
 3   Timestamp   218508 non-null  float64
 4   Gender      223223 non-null  object
 5   Age         223223 non-null  float64
 6   Occupation  223223 non-null  float64
 7   Zip-code    223223 non-null  object
 8   Title       218454 non-null  object
 9   Genres      218454 non-null  object
dtypes: float64(6), object(4)
memory usage: 18.8+ MB
```

[11]:
```
df.shape
```

[11]: (223625, 10)

[12]:
```
corr = df.corr()
sns.heatmap(corr,annot= True,linewidths=0.5)
```

[12]: <AxesSubplot:>



[16]:
```
# Extracting pandas profiling report

pfr = pf.ProfileReport(df)
pfr.to_file('Movielens_pfr.html')
```

```
Summarize dataset:   0%|          | 0/24 [00:00<?, ?it/s]
```

```
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
```

[17]:
```
print('Na values in the data frame is :')
def is_na(x):
    for i in x.columns:
        print(i,'column',' :',x[i].isna().sum(),'\n')
is_na(df)
```

```
Na values in the data frame is :
UserID column  : 402

MovieID column  : 4715

Rating column  : 5117

Timestamp column  : 5117

Gender column  : 402

Age column  : 402

Occupation column  : 402

Zip-code column  : 402

Title column  : 5171

Genres column  : 5171
```

[18]:
```
df.dropna(inplace=True)
```

[19]:
```
df.Rating.isna().value_counts()
```

[19]:
```
False    218052
Name: Rating, dtype: int64
```

[20]:
```
def df_unique(X):
    for i in X.columns:
        print('Column : ',i,'\n',X[i].unique(), '\n Total unique values is: ',␣
 ↪X[i].nunique())
```

```
      ␣
 ↪print('--------------------------------------------------------------------')
df_unique(df)
```

```
Column :  UserID
 [1.000e+00 2.000e+00 1.200e+01 … 1.580e+02 1.154e+03 7.450e+02]
 Total unique values is:  1325
----------------------------------------------------------------------
Column :  MovieID
 [1193.  661.  914. … 2101. 2563. 3905.]
 Total unique values is:  3480
----------------------------------------------------------------------
Column :  Rating
 [5. 4. 3. 2. 1.]
 Total unique values is:  5
----------------------------------------------------------------------
Column :  Timestamp
 [9.78300760e+08 9.78298413e+08 9.78220179e+08 … 9.74790833e+08
 1.01401892e+09 1.01174612e+09]
 Total unique values is:  102298
----------------------------------------------------------------------
Column :  Gender
 ['F' 'M']
 Total unique values is:  2
----------------------------------------------------------------------
Column :  Age
 [ 1. 56. 25. 50. 18. 45. 35.]
 Total unique values is:  7
----------------------------------------------------------------------
Column :  Occupation
 [10. 16. 12.  7.  1.  3.  4.  8. 17.  0.  2.  9. 19. 18. 15. 11. 20. 13.
  5. 14.  6.]
 Total unique values is:  21
----------------------------------------------------------------------
Column :  Zip-code
 ['48067' '70072' '32793' … '28754' '53545' '90004']
 Total unique values is:  1007
----------------------------------------------------------------------
Column :  Title
 ["One Flew Over the Cuckoo's Nest (1975)"
 'James and the Giant Peach (1996)' 'My Fair Lady (1964)' …
 "Squanto: A Warrior's Tale (1994)" 'Beauty (1998)' 'Specials, The (2000)']
 Total unique values is:  3480
----------------------------------------------------------------------
Column :  Genres
 ['Drama' "Animation|Children's|Musical" 'Musical|Romance'
 "Animation|Children's|Comedy" 'Action|Adventure|Comedy|Romance'
```

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

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

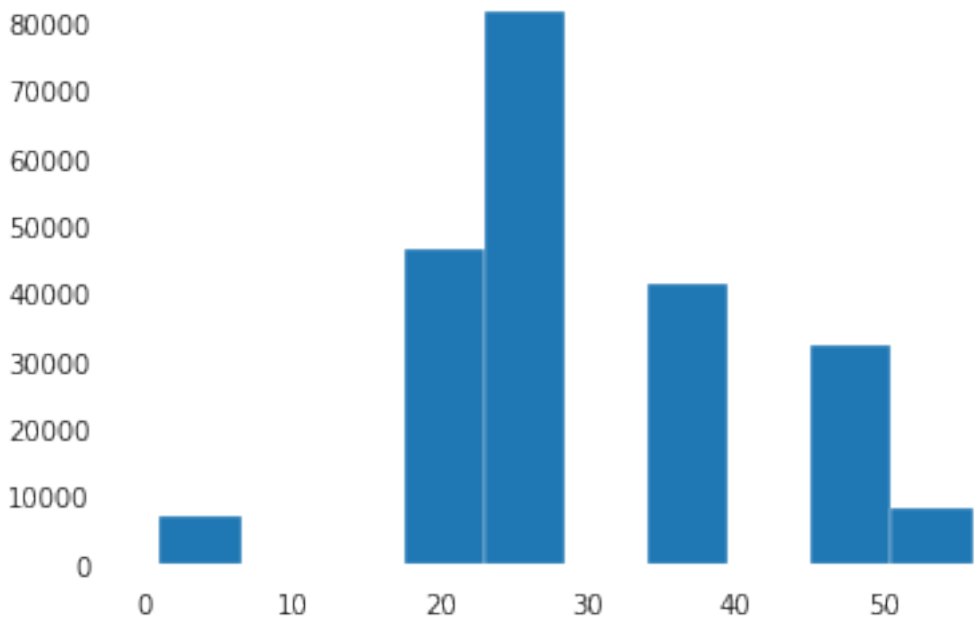
```
"Action|Animation|Children's|Sci-Fi|Thriller|War" 'Crime|Drama|Romance'
'Crime|Film-Noir' 'Mystery|Romance|Thriller'
'Comedy|Mystery|Romance|Thriller' 'Action|Adventure|Sci-Fi|Thriller|War'
'Adventure|Crime|Sci-Fi|Thriller' 'Action|Adventure|Western'
"Animation|Children's|Fantasy|War" 'Action|Adventure|Comedy|War'
"Children's|Comedy|Sci-Fi"
"Adventure|Animation|Children's|Comedy|Fantasy" 'Drama|Musical|War'
'Drama|Mystery|Romance' 'Adventure|Drama|Romance' 'Film-Noir'
'Film-Noir|Romance|Thriller' 'Drama|Film-Noir' 'Romance|Thriller'
'Action|Adventure|War' 'Mystery' 'Action|Adventure|Drama|Thriller'
'Musical|Romance|War' 'Drama|Western'
'Action|Drama|Mystery|Romance|Thriller' 'Adventure|Comedy|Musical'
'Documentary|Musical' 'Action|Thriller|War' 'Adventure|Comedy|Romance'
"Adventure|Children's|Comedy|Fantasy|Romance" 'Romance|War'
'Comedy|Romance|Sci-Fi' 'Action|Mystery|Sci-Fi|Thriller'
"Children's|Horror" 'Adventure|Musical|Romance'
"Adventure|Children's|Comedy|Musical" "Children's|Comedy|Mystery"
'Action|Comedy|Romance|Thriller' 'Action|Drama|Western'
"Animation|Children's|Comedy|Romance" 'Comedy|Mystery|Romance'
'Action|Crime|Mystery' 'Comedy|Drama|Thriller' 'Musical|War'
'Documentary|Drama' 'Action|Adventure|Crime|Thriller'
"Action|Adventure|Children's" "Adventure|Children's|Romance"
"Adventure|Animation|Children's"
"Action|Adventure|Animation|Children's|Fantasy"
"Adventure|Animation|Children's|Fantasy" 'Drama|Film-Noir|Thriller'
'Crime|Mystery' 'Documentary|War' 'Action|Comedy|Crime'
'Drama|Romance|Sci-Fi' 'Horror|Mystery' 'Drama|Horror|Thriller'
"Action|Adventure|Children's|Fantasy" 'Animation|Mystery'
'Drama|Romance|Western' 'Romance|Western' 'Comedy|Film-Noir|Thriller'
'Fantasy' 'Film-Noir|Horror']
 Total unique values is:  301
-----------------------------------------------------------------------
```

Exploring the datasets using visual representations

[21]:
```
#Visualizing the User Age Distribution

df.Age.hist(grid=False)
```

[21]:
```
<AxesSubplot:>
```



[22]:
```
#Visualizing User rating of the movie "Toy Story"
def fn(x):
    return re.search("Toy Story".lower(), x.lower())!=None
title = df.iloc[0].Title
title
```

[22]:
```
"One Flew Over the Cuckoo's Nest (1975)"
```

[23]:
```
re_tit = df["Title"].apply(fn)
re_tit.head()
```

[23]:
```
0    False
1    False
2    False
3    False
4    False
Name: Title, dtype: bool
```

[24]:
```
toystory = df[df["Title"].apply(fn)]
toystory
```

[24]:
```
        UserID  MovieID  Rating     Timestamp Gender   Age  Occupation  \
12156     1.0   3114.0     4.0  9.783022e+08      F   1.0        10.0
12157     3.0   3114.0     3.0  9.782981e+08      M  25.0        15.0
12158     9.0   3114.0     4.0  9.782260e+08      M  25.0        17.0
```

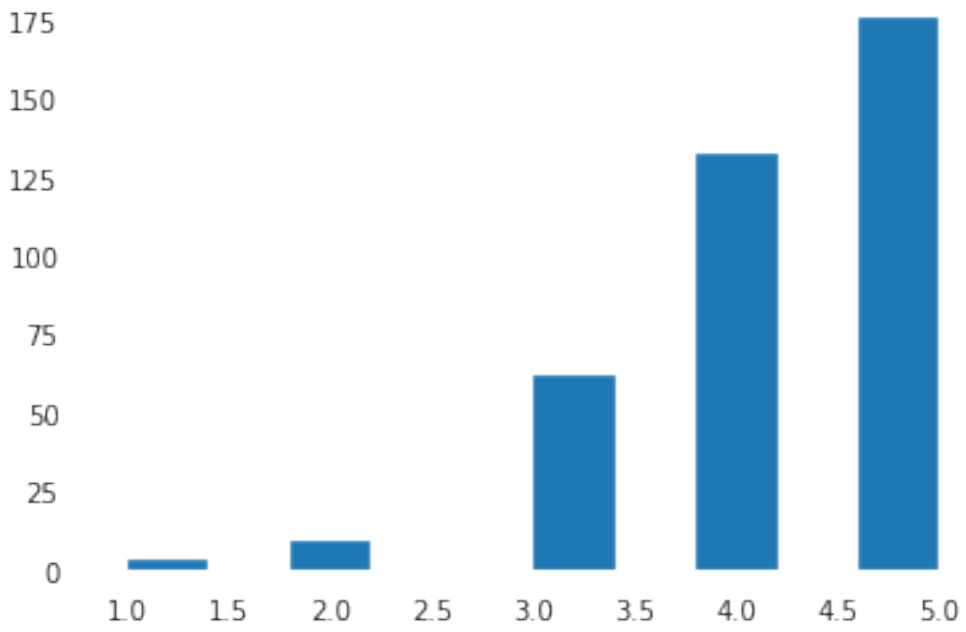
```
12159    10.0   3114.0     4.0  9.782258e+08      F  35.0         1.0
12160    17.0   3114.0     5.0  9.781594e+08      M  50.0         1.0
…         …       …       …           …       …    …           …
12536  1300.0   3114.0     5.0  9.747896e+08      M   1.0        10.0
12537  1301.0   3114.0     3.0  9.747886e+08      F  25.0         2.0
12538  1306.0   3114.0     4.0  1.006603e+09      M  25.0         3.0
12539  1314.0   3114.0     1.0  9.747853e+08      F  18.0         4.0
12540  1317.0   3114.0     4.0  9.747816e+08      M  45.0         7.0

      Zip-code              Title                       Genres
12156    48067  Toy Story 2 (1999)  Animation|Children's|Comedy
12157    55117  Toy Story 2 (1999)  Animation|Children's|Comedy
12158    61614  Toy Story 2 (1999)  Animation|Children's|Comedy
12159    95370  Toy Story 2 (1999)  Animation|Children's|Comedy
12160    95350  Toy Story 2 (1999)  Animation|Children's|Comedy
…          …                  …                         …
12536    97201  Toy Story 2 (1999)  Animation|Children's|Comedy
12537    10475  Toy Story 2 (1999)  Animation|Children's|Comedy
12538    64068  Toy Story 2 (1999)  Animation|Children's|Comedy
12539    68154  Toy Story 2 (1999)  Animation|Children's|Comedy
12540    98028  Toy Story 2 (1999)  Animation|Children's|Comedy

[385 rows x 10 columns]
```

[25]:
```
toystory.Rating.hist(grid=False)
```

[25]: <AxesSubplot:>



Top 25movies based on viewer ratings

[26]:
```
top_25 = df.groupby(["MovieID", "Title"]).Timestamp.count().
 ↪sort_values(ascending=False)
top_25
```

[26]:
```
MovieID  Title
2858.0   American Beauty (1999)                                793
1196.0   Star Wars: Episode V - The Empire Strikes Back (1980)  654
480.0    Jurassic Park (1993)                                  646
260.0    Star Wars: Episode IV - A New Hope (1977)             644
1210.0   Star Wars: Episode VI - Return of the Jedi (1983)     642
                                                               …
2631.0   Frogs for Snakes (1998)                                 1
1773.0   Tokyo Fist (1995)                                       1
758.0    Jar, The (Khomreh) (1992)                               1
120.0    Race the Sun (1996)                                     1
1579.0   For Ever Mozart (1996)                                  1
Name: Timestamp, Length: 3480, dtype: int64
```

[27]:
```
print('Top 25 movies by viewership rating')
print(top_25[:25])
```

```
Top 25 movies by viewership rating
MovieID  Title
2858.0   American Beauty (1999)                                793
1196.0   Star Wars: Episode V - The Empire Strikes Back (1980)  654
480.0    Jurassic Park (1993)                                  646
260.0    Star Wars: Episode IV - A New Hope (1977)             644
1210.0   Star Wars: Episode VI - Return of the Jedi (1983)     642
2028.0   Saving Private Ryan (1998)                            602
589.0    Terminator 2: Judgment Day (1991)                     585
2571.0   Matrix, The (1999)                                    582
1580.0   Men in Black (1997)                                   579
593.0    Silence of the Lambs, The (1991)                      563
110.0    Braveheart (1995)                                     551
1198.0   Raiders of the Lost Ark (1981)                        549
2396.0   Shakespeare in Love (1998)                            538
3578.0   Gladiator (2000)                                      537
1270.0   Back to the Future (1985)                             534
608.0    Fargo (1996)                                          522
2762.0   Sixth Sense, The (1999)                               519
527.0    Schindler's List (1993)                               509
1265.0   Groundhog Day (1993)                                  506
1197.0   Princess Bride, The (1987)                            499
```

```
1617.0    L.A. Confidential (1997)                    494
2997.0    Being John Malkovich (1999)                 481
1097.0    E.T. the Extra-Terrestrial (1982)           476
296.0     Pulp Fiction (1994)                         475
318.0     Shawshank Redemption, The (1994)            474
Name: Timestamp, dtype: int64
```

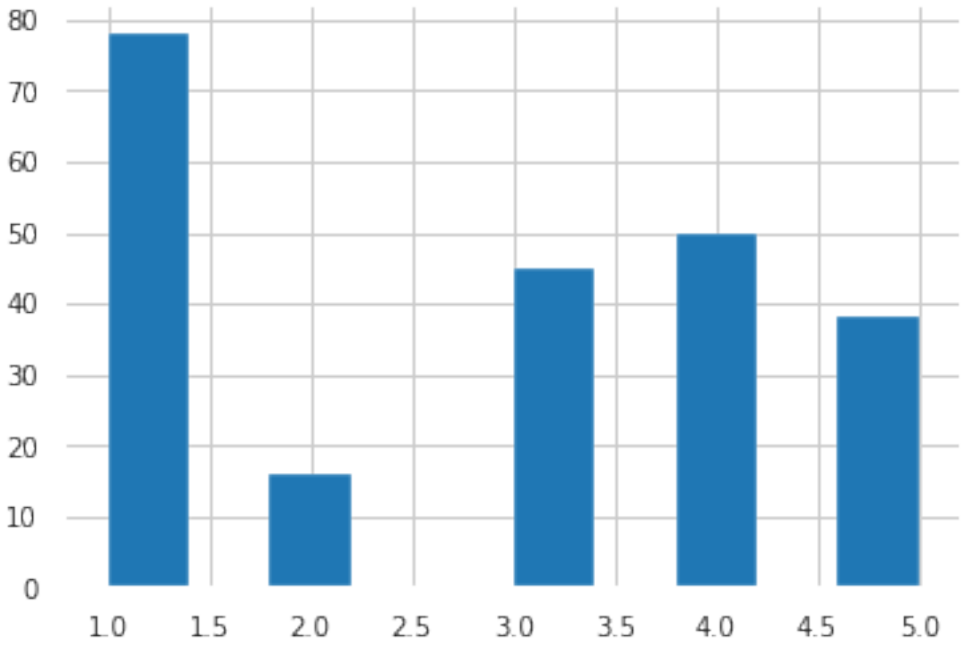
The ratings for all the movies reviewed by for a particular user of user id = 1314

[29]:
```
usr_1314= df.loc[df.UserID==1314, "Rating"].sort_values(ascending=False)
usr_1314 .head(),usr_1314.shape
```

[29]:
```
(112591    5.0
 189453    5.0
 183180    5.0
 96691     5.0
 37489     5.0
 Name: Rating, dtype: float64,
 (227,))
```

[30]:
```
usr_1314.hist()
```

[30]:
```
<AxesSubplot:>
```



Finding all the unique genres

[31]:
```
df.Genres.unique()
```

```
[31]: array(['Drama', "Animation|Children's|Musical", 'Musical|Romance',
              "Animation|Children's|Comedy", 'Action|Adventure|Comedy|Romance',
              'Action|Adventure|Drama', 'Comedy|Drama',
              "Adventure|Children's|Drama|Musical", 'Musical', 'Comedy',
              "Animation|Children's", 'Comedy|Fantasy', 'Animation',
              'Comedy|Sci-Fi', 'Drama|War', 'Romance',
              "Animation|Children's|Musical|Romance",
              "Children's|Drama|Fantasy|Sci-Fi", 'Drama|Romance',
              'Animation|Comedy|Thriller',
              "Adventure|Animation|Children's|Comedy|Musical",
              "Animation|Children's|Comedy|Musical", 'Thriller',
              'Action|Crime|Romance', 'Action|Adventure|Fantasy|Sci-Fi',
              "Children's|Comedy|Musical", 'Action|Drama|War',
              "Children's|Drama", 'Crime|Drama|Thriller', 'Action|Crime|Drama',
              'Action|Adventure|Mystery', 'Crime|Drama',
              'Action|Adventure|Sci-Fi|Thriller',
              'Action|Adventure|Romance|Sci-Fi|War', 'Action|Thriller',
              'Action|Drama', 'Comedy|Drama|Western', 'Action|Adventure|Crime',
              'Action|Crime|Mystery|Thriller', 'Comedy|Drama|Romance',
              'Comedy|Drama|War', 'Drama|Sci-Fi', 'Action|Drama|Thriller',
              'Action|Comedy|Western', 'Adventure|Comedy|Drama',
              'Drama|Thriller', 'Comedy|Romance',
              'Action|Drama|Romance|Thriller', 'Action|Crime|Thriller',
              'Action|Sci-Fi|Thriller', 'Action|Horror|Sci-Fi', 'Action|Sci-Fi',
              'Action|Romance|War', 'Adventure|Drama|Romance|Sci-Fi',
              'Action|Adventure|Sci-Fi', 'Drama|Romance|War',
              'Action|Drama|Romance', 'Crime|Drama|Film-Noir|Thriller',
              'Adventure|Drama|Western', 'Action|Adventure|Drama|Sci-Fi|War',
              'Action|Adventure|Thriller', 'Action|Adventure|Romance|Thriller',
              'Action|Adventure', 'Comedy|Horror', 'Action|Crime|Drama|Thriller',
              'Action|Mystery|Romance|Thriller', 'Action|Romance|Thriller',
              'Action|Comedy|Drama', 'Action', 'Action|Sci-Fi|War',
              'Action|Comedy|Crime|Drama', 'Action|Adventure|Romance',
              'Comedy|Romance|War', 'Comedy|Thriller', 'Action|Adventure|Comedy',
              'Action|Comedy', 'Adventure|Thriller', 'Action|Adventure|Fantasy',
              'Action|Adventure|Horror', 'Action|Adventure|Comedy|Sci-Fi',
              'Action|Adventure|Comedy|Horror', 'Western', 'Adventure|Comedy',
              'Adventure|Drama', 'Action|Adventure|Horror|Thriller',
              'Comedy|Western', "Animation|Children's|Comedy|Musical|Romance",
              'Action|Western', 'Action|Horror|Sci-Fi|Thriller', 'Action|Horror',
              'Adventure|Animation|Film-Noir', 'Drama|Romance|Thriller',
              'Crime|Drama|Romance|Thriller', 'Crime|Thriller',
              'Animation|Comedy', 'Documentary',
              'Crime|Film-Noir|Mystery|Thriller', 'Drama|Horror',
              'Mystery|Sci-Fi|Thriller', 'Drama|Mystery', 'Horror|Romance',
              'Horror|Sci-Fi', 'Horror', 'Sci-Fi|Thriller', 'Crime',
              'Action|Crime', 'Crime|Horror', 'Drama|Mystery|Thriller',
```

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

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

```
       'Animation|Mystery', 'Drama|Romance|Western', 'Romance|Western',
       'Comedy|Film-Noir|Thriller', 'Fantasy', 'Film-Noir|Horror'],
      dtype=object)
```

[32]:
```
Genres_list = df.Genres.tolist()
genre_list = []
i = 0
while(i<len(Genres_list)):
    genre_list+= Genres_list[i].split('|')
    i+=1
```

[33]:
```
unique_gen = list(set(genre_list))
print(unique_gen)
print()
print("Length of the unique Genre : ",len(unique_gen))
```

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

Length of the unique Genre :  18
```

Creating a separate column for each genre category with a one-hot encoding (1 and 0)

[34]:
```
new_data = pd.concat([df,df.Genres.str.get_dummies()], axis=1)
print(new_data.columns)
```

```
Index(['UserID', 'MovieID', 'Rating', 'Timestamp', 'Gender', 'Age',
       'Occupation', 'Zip-code', 'Title', 'Genres', 'Action', 'Adventure',
       'Animation', 'Children's', 'Comedy', 'Crime', 'Documentary', 'Drama',
       'Fantasy', 'Film-Noir', 'Horror', 'Musical', 'Mystery', 'Romance',
       'Sci-Fi', 'Thriller', 'War', 'Western'],
      dtype='object')
```

[35]:
```
new_data.head()
```

[35]:

	UserID	MovieID	Rating	Timestamp	Gender	Age	Occupation	Zip-code	\
0	1.0	1193.0	5.0	978300760.0	F	1.0	10.0	48067	
1	2.0	1193.0	5.0	978298413.0	M	56.0	16.0	70072	
2	12.0	1193.0	4.0	978220179.0	M	25.0	12.0	32793	
3	15.0	1193.0	4.0	978199279.0	M	25.0	7.0	22903	
4	17.0	1193.0	5.0	978158471.0	M	50.0	1.0	95350	

	Title	Genres	…	Fantasy	Film-Noir	\
0	One Flew Over the Cuckoo's Nest (1975)	Drama	…	0	0	
1	One Flew Over the Cuckoo's Nest (1975)	Drama	…	0	0	
2	One Flew Over the Cuckoo's Nest (1975)	Drama	…	0	0	
3	One Flew Over the Cuckoo's Nest (1975)	Drama	…	0	0	

```
4  One Flew Over the Cuckoo's Nest (1975)  Drama  …        0          0

   Horror  Musical  Mystery  Romance  Sci-Fi  Thriller  War  Western
0       0        0        0        0       0         0    0        0
1       0        0        0        0       0         0    0        0
2       0        0        0        0       0         0    0        0
3       0        0        0        0       0         0    0        0
4       0        0        0        0       0         0    0        0

[5 rows x 28 columns]
```

[36]:
```
df_new = new_data.drop(['Title','Zip-code','Timestamp','Genres'],axis=1)
df_new.head()
```

[36]:
```
   UserID  MovieID  Rating Gender   Age  Occupation  Action  Adventure  \
0     1.0   1193.0     5.0      F   1.0        10.0       0          0
1     2.0   1193.0     5.0      M  56.0        16.0       0          0
2    12.0   1193.0     4.0      M  25.0        12.0       0          0
3    15.0   1193.0     4.0      M  25.0         7.0       0          0
4    17.0   1193.0     5.0      M  50.0         1.0       0          0

   Animation  Children's  …  Fantasy  Film-Noir  Horror  Musical  Mystery  \
0          0           0  …        0          0       0        0        0
1          0           0  …        0          0       0        0        0
2          0           0  …        0          0       0        0        0
3          0           0  …        0          0       0        0        0
4          0           0  …        0          0       0        0        0

   Romance  Sci-Fi  Thriller  War  Western
0        0       0         0    0        0
1        0       0         0    0        0
2        0       0         0    0        0
3        0       0         0    0        0
4        0       0         0    0        0

[5 rows x 24 columns]
```

[37]:
```
print(df_new.columns)
```

```
Index(['UserID', 'MovieID', 'Rating', 'Gender', 'Age', 'Occupation', 'Action',
       'Adventure', 'Animation', 'Children's', 'Comedy', 'Crime',
       'Documentary', 'Drama', 'Fantasy', 'Film-Noir', 'Horror', 'Musical',
       'Mystery', 'Romance', 'Sci-Fi', 'Thriller', 'War', 'Western'],
      dtype='object')
```

Encoding the gender column

[38]:
```
df_new.Gender = pd.get_dummies(df_new.Gender)
```

[39]:
```
x = df_new.drop(['UserID','MovieID','Rating'],axis=1)
x.shape
```

[39]:
```
(218052, 21)
```

The features affecting the ratings of any particular movie.

[40]:
```
print('The features affecting the ratings of any particular movie:')
print()
print(x.columns)
```

```
The features affecting the ratings of any particular movie:

Index(['Gender', 'Age', 'Occupation', 'Action', 'Adventure', 'Animation',
       'Children's', 'Comedy', 'Crime', 'Documentary', 'Drama', 'Fantasy',
       'Film-Noir', 'Horror', 'Musical', 'Mystery', 'Romance', 'Sci-Fi',
       'Thriller', 'War', 'Western'],
      dtype='object')
```

[41]:
```
y = df_new.Rating
y.shape
```

[41]:
```
(218052,)
```

[42]:
```
x.Occupation.value_counts()
```

[42]:
```
0.0     29711
4.0     27198
7.0     21572
1.0     19332
17.0    15837
20.0    15232
12.0    13249
2.0      9963
16.0     9617
3.0      9405
14.0     7979
10.0     7190
6.0      7177
15.0     4514
19.0     4357
11.0     4218
13.0     3391
9.0      2631
5.0      2290
18.0     1881
8.0      1308
Name: Occupation, dtype: int64
```

[43]:
```
x = x.join(pd.get_dummies(x.Occupation,prefix='Occupation'))
x.head(),x.columns
```

[43]:
```
(   Gender   Age  Occupation  Action  Adventure  Animation  Children's  Comedy
\
 0       1   1.0        10.0       0          0          0           0       0
 1       0  56.0        16.0       0          0          0           0       0
 2       0  25.0        12.0       0          0          0           0       0
 3       0  25.0         7.0       0          0          0           0       0
 4       0  50.0         1.0       0          0          0           0       0

    Crime  Documentary  …  Occupation_11.0  Occupation_12.0  Occupation_13.0
\
 0      0            0  …                0                0                0
 1      0            0  …                0                0                0
 2      0            0  …                0                1                0
 3      0            0  …                0                0                0
 4      0            0  …                0                0                0

    Occupation_14.0  Occupation_15.0  Occupation_16.0  Occupation_17.0  \
 0                0                0                0                0
 1                0                0                1                0
 2                0                0                0                0
 3                0                0                0                0
 4                0                0                0                0

    Occupation_18.0  Occupation_19.0  Occupation_20.0
 0                0                0                0
 1                0                0                0
 2                0                0                0
 3                0                0                0
 4                0                0                0

 [5 rows x 42 columns],
 Index(['Gender', 'Age', 'Occupation', 'Action', 'Adventure', 'Animation',
        'Children's', 'Comedy', 'Crime', 'Documentary', 'Drama', 'Fantasy',
        'Film-Noir', 'Horror', 'Musical', 'Mystery', 'Romance', 'Sci-Fi',
        'Thriller', 'War', 'Western', 'Occupation_0.0', 'Occupation_1.0',
        'Occupation_2.0', 'Occupation_3.0', 'Occupation_4.0', 'Occupation_5.0',
        'Occupation_6.0', 'Occupation_7.0', 'Occupation_8.0', 'Occupation_9.0',
        'Occupation_10.0', 'Occupation_11.0', 'Occupation_12.0',
        'Occupation_13.0', 'Occupation_14.0', 'Occupation_15.0',
        'Occupation_16.0', 'Occupation_17.0', 'Occupation_18.0',
        'Occupation_19.0', 'Occupation_20.0'],
       dtype='object'))
```

[44]:
```
x = x.drop(['Occupation','Occupation_0.0'],axis=1)
x.head(3),x.shape
```

[44]:
```
(   Gender   Age  Action  Adventure  Animation  Children's  Comedy  Crime  \
 0       1   1.0       0          0          0           0       0      0
 1       0  56.0       0          0          0           0       0      0
 2       0  25.0       0          0          0           0       0      0

    Documentary  Drama  …  Occupation_11.0  Occupation_12.0  Occupation_13.0
\
 0            0      1  …                0                0                0
 1            0      1  …                0                0                0
 2            0      1  …                0                1                0

    Occupation_14.0  Occupation_15.0  Occupation_16.0  Occupation_17.0  \
 0                0                0                0                0
 1                0                0                1                0
 2                0                0                0                0

    Occupation_18.0  Occupation_19.0  Occupation_20.0
 0                0                0                0
 1                0                0                0
 2                0                0                0

 [3 rows x 40 columns],
 (218052, 40))
```

Deploying the hold out method

[45]:
```
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.
↪2,random_state = 10,stratify=y)
```

Deploying the model

[49]:
```
lgb = LGBMClassifier(boosting_type = 'gbdt',n_jobs= -1,objective='multiclass')
```

[50]:
```
lgb.fit(x_train,y_train)
```

[50]:
```
LGBMClassifier(objective='multiclass')
```

[51]:
```
y_pred = lgb.predict(x_test)
```

[52]:
```
print('LGBM accuracy score is : ', accuracy_score(y_test,y_pred)*100)
```

```
LGBM accuracy score is :  37.025062484235626
```

[54]:
```
xgb = xgboost.XGBClassifier(n_jobs=-1)
```

[ ]:
```
xgb.fit(x_train,y_train)
```

[ ]:
```
y_pred_xgb = xgb.predict(x_test)
```

[ ]:
```
print('XGB accuracy score is : ', accuracy_score(y_test,y_pred_xgb )*100)
```

Accuracy score of both the model LGBM accuracy score is : 36.19%

XGB accuracy score is : 35.39%