In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
unames = ['user_id', 'gender', 'age', 'occupation', 'zip']
users = pd.read_table('ml-1m/users.dat', sep='::', header=None, names=unames, engine='python',encoding='ISO-8859-1')

rnames = ['user_id', 'movie_id', 'rating', 'timestamp']
ratings = pd.read_table('ml-1m/ratings.dat', sep='::', header=None, names=rnames, engine='python', encoding='ISO-8859-1')

mnames = ['movie_id', 'title', 'genres']
movies = pd.read_table('ml-1m/movies.dat', sep='::', header=None, names=mnames, engine='python', encoding='ISO-8859-1')
```

In [3]:
```
users[:5]
```

Out[3]:

	user_id	gender	age	occupation	zip
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

In [4]:
```
ratings[:5]
```

Out[4]:

	user_id	movie_id	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

In [5]:
```
movies[:5]
```

Out[5]:

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

In [6]:
```
data = pd.merge(pd.merge(ratings, users), movies)
data3 = data
```

In [7]:
```
data.head(1)
```

Out[7]:

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres
0	1	1193	5	978300760	F	1	10	48067	One Flew Over the Cuckoo's Nest (1975)	Drama

In [8]:
```
# flatten the genres
data['genres']= data['genres'].str.split('|')
flatdata = pd.DataFrame([( index, value) for ( index, values)
                         in data['genres'].iteritems() for value in values],
                             columns = ['index','genres']).set_index('index')

data2 = data.drop('genres', axis = 1 ).join(flatdata )
```

In [9]:
```
data2[:10]
```

Out[9]:

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres
0	1	1193	5	978300760	F	1	10	48067	One Flew Over the Cuckoo's Nest (1975)	Drama
1	2	1193	5	978298413	M	56	16	70072	One Flew Over the Cuckoo's Nest (1975)	Drama
2	12	1193	4	978220179	M	25	12	32793	One Flew Over the Cuckoo's Nest (1975)	Drama
3	15	1193	4	978199279	M	25	7	22903	One Flew Over the Cuckoo's Nest (1975)	Drama
4	17	1193	5	978158471	M	50	1	95350	One Flew Over the Cuckoo's Nest (1975)	Drama
5	18	1193	4	978156168	F	18	3	95825	One Flew Over the Cuckoo's Nest (1975)	Drama
6	19	1193	5	982730936	M	1	10	48073	One Flew Over the Cuckoo's Nest (1975)	Drama
7	24	1193	5	978136709	F	25	7	10023	One Flew Over the Cuckoo's Nest (1975)	Drama
8	28	1193	3	978125194	F	25	1	14607	One Flew Over the Cuckoo's Nest (1975)	Drama
9	33	1193	5	978557765	M	45	3	55421	One Flew Over the Cuckoo's Nest (1975)	Drama

Question 1: An aggregate on the number of rating done for each particular genre, e.g., Action, Adventure, Drama, Science Fiction, ...

In [10]:
```
aggregation = data2.groupby('genres').size().sort_values(ascending=False)
aggregation
```

Out[10]:
```
genres
Comedy         356580
Drama          354529
Action         257457
Thriller       189680
Sci-Fi         157294
Romance        147523
Adventure      133953
Crime           79541
Horror          76386
Children's      72186
War             68527
Animation       43293
Musical         41533
Mystery         40178
Fantasy         36301
Western         20683
Film-Noir       18261
Documentary      7910
dtype: int64
```

In [11]:
```
mean_ratings = data.pivot_table('rating', index='title', columns='gender', aggfunc='mean')
```

In [12]:
```
mean_ratings[:5]
```

Out[12]:

gender	F	M
title		
$1,000,000 Duck (1971)	3.375000	2.761905
'Night Mother (1986)	3.388889	3.352941
'Til There Was You (1997)	2.675676	2.733333
'burbs, The (1989)	2.793478	2.962085
...And Justice for All (1979)	3.828571	3.689024

Question2: The top 5 ranked genres by women on most number of rating.

In [13]:
```
women = data2[data2['gender'] == 'F']
women = women.groupby('genres').size().sort_values(ascending=False)[:5]
women
```

Out[13]:
```
genres
Drama       98153
Comedy      96271
Romance     50297
Action      45650
Thriller    40308
dtype: int64
```

Question3: The top 5 ranked genres by men on most number of rating.

In [14]:
```
men = data2[data2['gender'] == 'M']
men = men.groupby('genres').size().sort_values(ascending=False)[:5]
men
```

Out[14]:
```
genres
Comedy      260309
Drama       256376
Action      211807
Thriller    149372
Sci-Fi      129894
dtype: int64
```

Question 4:Pick a genre of your choice and provide average movie's ratings by the following four time intervals during which the movies were released (a) 1970 to 1979 (b) 1980 to 1989 (c) 1990 to 1999 (d) 2000 to 2009. Also, if you observed any issues with data in any of these ranges, please mention it.

In [15]:
```
action_list = data2.loc[data2['genres'] == 'Romance']
action_list[:10]
```

Out[15]:

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres
2250	1	914	3	978301968	F	1	10	48067	My Fair Lady (1964)	Romance
2251	6	914	5	978237767	F	50	9	55117	My Fair Lady (1964)	Romance
2252	10	914	5	978226805	F	35	1	95370	My Fair Lady (1964)	Romance
2253	33	914	5	978108939	M	45	3	55421	My Fair Lady (1964)	Romance
2254	35	914	3	978101982	M	45	1	02482	My Fair Lady (1964)	Romance
2255	45	914	4	977988097	F	45	16	94110	My Fair Lady (1964)	Romance
2256	48	914	3	978059754	M	25	4	92107	My Fair Lady (1964)	Romance
2257	53	914	5	977979589	M	25	0	96931	My Fair Lady (1964)	Romance
2258	59	914	5	1041962991	F	50	1	55413	My Fair Lady (1964)	Romance
2259	78	914	4	977811665	F	45	1	98029	My Fair Lady (1964)	Romance

In [16]:
```
# period 1970 - 1979
period1 = action_list[action_list['title'].str.contains('\(197')]
period1
```

Out[16]:

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres
177108	2	1244	3	978299143	M	56	16	70072	Manhattan (1979)	Romance
177109	11	1244	4	978903024	F	25	1	04093	Manhattan (1979)	Romance
177110	28	1244	4	978126330	F	25	1	14607	Manhattan (1979)	Romance
177111	36	1244	4	978063059	M	25	3	94123	Manhattan (1979)	Romance
177112	45	1244	4	977988138	F	45	16	94110	Manhattan (1979)	Romance
...	...	...	...	...	...	...	...	...	...	...
864566	5090	3284	3	962391781	F	35	19	75069	They Might Be Giants (1971)	Romance
864567	5448	3284	2	959965709	M	45	19	60626	They Might Be Giants (1971)	Romance
864568	5511	3284	4	959787621	M	45	1	92407	They Might Be Giants (1971)	Romance
864569	5954	3284	5	957128936	M	45	11	70802	They Might Be Giants (1971)	Romance
864570	6000	3284	2	956882279	M	45	17	30075	They Might Be Giants (1971)	Romance

4191 rows × 10 columns

In [17]:
```
period1['rating'].mean()
```

Out[17]:
```
3.772607969458363
```

In [18]:
```
# period 1980 - 1989
period2 = action_list[action_list['title'].str.contains('\(198')]
period2
```

Out[18]:

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres
5904	1	1197	3	978302268	F	1	10	48067	Princess Bride, The (1987)	Romance
5905	3	1197	5	978297570	M	25	15	55117	Princess Bride, The (1987)	Romance
5906	10	1197	5	979167660	F	35	1	95370	Princess Bride, The (1987)	Romance
5907	11	1197	5	978903297	F	25	1	04093	Princess Bride, The (1987)	Romance
5908	13	1197	4	978201320	M	45	1	93304	Princess Bride, The (1987)	Romance

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres
...	...	...	...	...	...	...	...	...	...	...
998923	4140	2257	3	965351526	M	25	0	32112	No Small Affair (1984)	Romance
998924	4543	2257	3	964671121	M	25	2	11105	No Small Affair (1984)	Romance
998925	4754	2257	2	963185384	F	18	0	91107	No Small Affair (1984)	Romance
998926	5831	2257	4	957898337	M	25	1	92120	No Small Affair (1984)	Romance
999894	1778	3458	3	975575708	M	18	4	94704	Blood and Sand (Sangre y Arena) (1989)	Romance

30138 rows × 10 columns

In [19]:
```
period2['rating'].mean()
```

Out[19]:
```
3.66016324905435
```

In [20]:
```
# period 1990 - 1999
period3 = action_list[action_list['title'].str.contains('\(199')]
period3
```

Out[20]:

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres
28157	1	2340	3	978300103	F	1	10	48067	Meet Joe Black (1998)	Romance
28158	26	2340	4	978141178	M	25	7	23112	Meet Joe Black (1998)	Romance
28159	38	2340	3	978044835	F	18	4	02215	Meet Joe Black (1998)	Romance
28160	45	2340	2	977988826	F	45	16	94110	Meet Joe Black (1998)	Romance
28161	116	2340	4	997448150	M	25	17	55744	Meet Joe Black (1998)	Romance
...	...	...	...	...	...	...	...	...	...	...
1000023	2507	1714	2	975382922	M	25	4	94107	Never Met Picasso (1996)	Romance
1000039	2796	1851	4	997320494	M	25	14	92104	Leather Jacket Love Story (1997)	Romance
1000040	3547	1851	4	966835923	M	35	16	94108	Leather Jacket Love Story (1997)	Romance
1000059	3015	889	3	975263628	M	56	6	62707	1-900 (1994)	Romance
1000060	3790	889	2	966019187	F	25	17	94618	1-900 (1994)	Romance

93489 rows × 10 columns

In [21]:
```
period3['rating'].mean()
```

Out[21]:
```
3.500347634481062
```

In [22]:
```
# period 2000 - 2009
period4 = action_list[action_list['title'].str.contains('\(200')]
period4
```

Out[22]:

	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres
334217	6	3536	5	978238230	F	50	9	55117	Keeping the Faith (2000)	Romance
334218	34	3536	4	978103849	F	18	0	02135	Keeping the Faith (2000)	Romance
334219	92	3536	1	986186782	F	18	4	44243	Keeping the Faith (2000)	Romance
334220	99	3536	3	982873145	F	1	10	19390	Keeping the Faith (2000)	Romance
334221	100	3536	1	977593886	M	35	17	95401	Keeping the Faith (2000)	Romance
...	...	...	...	...	...	...	...	...	...	...
980232	4858	3796	4	969717347	M	25	1	04086	Wisdom of Crocodiles, The (a.k.a. Immortality)...	Romance
999581	1274	3888	4	1007064872	M	45	7	37343	Skipped Parts (2000)	Romance
999582	4842	3888	5	1010971390	F	35	1	23062	Skipped Parts (2000)	Romance
999904	1865	3353	4	976586255	F	18	1	94606	Closer You Get, The (2000)	Romance
999905	4854	3353	4	962830843	F	50	13	03851	Closer You Get, The (2000)	Romance

2136 rows × 10 columns

In [23]:
```
period4['rating'].mean()
```

Out[23]:
```
3.2167602996254683
```

In summary, the average movie's ratings:

(a) from 1970 - 1979: 3.772

(b) from 1980 - 1989: 3.660

(c) from 1990 - 1999: 3.500

(d) from 2000 - 2009: 3.216

Question 5: A function that given a genre and a rating_range (i.e. [3.5, 4]), returns all the movies of that genre and within that rating range sorted by average rating. Using an example, demonstrate that your function works.

In [24]:
```
def rating_range(low, high, genre):

    """given a genre and a rating_range with a low and high,
    returns all the movies of that genre and within that rating range sorted by average rating"""

    movie_with_rating = data3.groupby("movie_id").agg({"rating":np.mean, "title":np.unique, "genres":np.unique})
    movie_of_genre = movie_with_rating[movie_with_rating["genres"].apply(lambda x: genre in x)]

    return movie_of_genre[(movie_of_genre["rating"] >= low)
                          & (movie_of_genre["rating"]<= high)].sort_values(by=["rating"], ascending=False)
```

In [25]:
```
# exapmle: given "Romance" genre, and a rating range: [3.5, 4], return all the movies of "Action" and within th

res = rating_range(3.5, 4, 'Romance')
res
```

Out[25]:

movie_id	rating	title	genres
1851	4.000000	Leather Jacket Love Story (1997)	[Drama, Romance]
497	4.000000	Much Ado About Nothing (1993)	[Comedy, Romance]
3353	4.000000	Closer You Get, The (2000)	[Comedy, Romance]
920	3.997405	Gone with the Wind (1939)	[Drama, Romance, War]
1296	3.996429	Room with a View, A (1986)	[Drama, Romance]
...	...	...	...
3320	3.511111	Mifune (Mifunes sidste sang) (1999)	[Comedy, Romance]
1897	3.503759	High Art (1998)	[Drama, Romance]
2127	3.500000	First Love, Last Rites (1997)	[Drama, Romance]
1685	3.500000	I Love You, I Love You Not (1996)	[Romance]
2215	3.500000	Rich and Strange (1932)	[Comedy, Romance]

147 rows × 3 columns

In [26]:
```
# example verification: given "Romance" genre, and a rating range: [3.5, 4], return all the movies of "Action" and within

return_range = data2.loc[data2['genres'] == 'Romance']

mean_ratings2 = return_range.groupby('title').mean()
return_range = mean_ratings2[(3.5 <= mean_ratings2['rating'])  & (mean_ratings2['rating'] <= 4)]

return_range = return_range.sort_values(by='rating', ascending=False)
return_range['rating']
```

Out[26]:
```
title
Much Ado About Nothing (1993)          4.000000
Leather Jacket Love Story (1997)       4.000000
Closer You Get, The (2000)             4.000000
Gone with the Wind (1939)              3.997405
Room with a View, A (1986)             3.996429
                                           ...
Mifune (Mifunes sidste sang) (1999)    3.511111
High Art (1998)                        3.503759
First Love, Last Rites (1997)          3.500000
I Love You, I Love You Not (1996)      3.500000
Rich and Strange (1932)                3.500000
Name: rating, Length: 147, dtype: float64
```

Question6: Present one other statistic, figure, aggregate, or plot that you created using this dataset, along with a short description of what

interesting observations you derived from it. This question is meant to give you a freehand to explore and present aspects of the dataset that interests you.

I want to figure out how women and men are rating Action movies, so I filter out ratings for Action movies, get the average ratings sorted, and then I check the difference bewteen them.

In [27]:
```
action_movies = data2[data2['genres'] == 'Action']
action_mean_ratings = action_movies.pivot_table('rating', index='title', columns='gender', aggfunc='mean')
action_mean_ratings['diff'] = action_mean_ratings['M'] - action_mean_ratings['F']
action_mean_ratings
```

Out[27]:

gender	F	M	diff
title			
13th Warrior, The (1999)	3.112000	3.168000	0.056000
3 Ninjas: High Noon On Mega Mountain (1998)	1.400000	1.351351	-0.048649
52 Pick-Up (1986)	3.304348	3.299145	-0.005203
7th Voyage of Sinbad, The (1958)	3.409091	3.658879	0.249788
Abyss, The (1989)	3.659236	3.689507	0.030272
...	...	...	...
Young Guns (1988)	3.371795	3.425620	0.053825
Young Guns II (1990)	2.934783	2.904025	-0.030758
Young Sherlock Holmes (1985)	3.514706	3.363344	-0.151362
Zero Kelvin (Kjærlighetens kjøtere) (1995)	NaN	3.500000	NaN
eXistenZ (1999)	3.098592	3.289086	0.190494

495 rows × 3 columns

In [28]:
```
action_sorted_by_diff = action_mean_ratings.sort_values(by='diff')
action_sorted_by_diff[:100]
```

Out[28]:

gender	F	M	diff
title			
Spiders, The (Die Spinnen, 1. Teil: Der Goldene See) (1919)	4.000000	1.000000	-3.000000
Coldblooded (1995)	5.000000	3.588235	-1.411765
Blood Beach (1981)	3.000000	1.650000	-1.350000
Assassination (1987)	4.000000	2.863636	-1.136364
Truth or Consequences, N.M. (1997)	3.375000	2.510204	-0.864796
...	...	...	...
Diva (1981)	4.164706	4.000000	-0.164706
Deep Blue Sea (1999)	3.010101	2.845691	-0.164410
Single White Female (1992)	3.234043	3.072674	-0.161368
Police Story 4: Project S (Chao ji ji hua) (1993)	3.000000	2.842105	-0.157895
Dragonheart (1996)	3.348148	3.190776	-0.157372

100 rows × 3 columns

It is very interesting to find that women give higher ratings than men. I'm curious that whether women usually give higher ranking to other genre of movies. Therefore, I run the analysis for Comdey as well.

In [29]:
```
comedy_movies = data2[data2['genres'] == 'Comedy']
comedy_mean_ratings = comedy_movies.pivot_table('rating', index='title', columns='gender', aggfunc='mean')
comedy_mean_ratings['diff'] = comedy_mean_ratings['M'] - comedy_mean_ratings['F']
comedy_mean_ratings
```

Out[29]:

gender	F	M	diff
title			
$1,000,000 Duck (1971)	3.375000	2.761905	-0.613095
'burbs, The (1989)	2.793478	2.962085	0.168607
10 Things I Hate About You (1999)	3.646552	3.311966	-0.334586

gender	F	M	diff
title			
101 Dalmatians (1996)	3.240000	2.911215	-0.328785
20 Dates (1998)	2.620690	2.918182	0.297492
...	...	...	...
Young Doctors in Love (1982)	1.923077	2.742424	0.819347
Young Frankenstein (1974)	4.289963	4.239177	-0.050785
Young Guns (1988)	3.371795	3.425620	0.053825
Young Guns II (1990)	2.934783	2.904025	-0.030758
Zero Effect (1998)	3.864407	3.723140	-0.141266

1163 rows × 3 columns

In [30]:
```
comedy_sorted_by_diff = action_mean_ratings.sort_values(by='diff')
comedy_sorted_by_diff[:100]
```

Out[30]:

gender	F	M	diff
title			
Spiders, The (Die Spinnen, 1. Teil: Der Goldene See) (1919)	4.000000	1.000000	-3.000000
Coldblooded (1995)	5.000000	3.588235	-1.411765
Blood Beach (1981)	3.000000	1.650000	-1.350000
Assassination (1987)	4.000000	2.863636	-1.136364
Truth or Consequences, N.M. (1997)	3.375000	2.510204	-0.864796
...	...	...	...
Diva (1981)	4.164706	4.000000	-0.164706
Deep Blue Sea (1999)	3.010101	2.845691	-0.164410
Single White Female (1992)	3.234043	3.072674	-0.161368
Police Story 4: Project S (Chao ji ji hua) (1993)	3.000000	2.842105	-0.157895
Dragonheart (1996)	3.348148	3.190776	-0.157372

100 rows × 3 columns

According to the result here, surprisinly, women still give higher ranking compared to men.

We could propose a question that: Are women always give higher ranjubg? Is there any psychological reason tied to that? How could we utilize the data to conduct further research?