In [28]:
```
import pandas as pd
```

In [29]:
```
ratings_df = pd.read_csv('Data/ml-latest-small/ratings_processed.csv', sep
ratings_df.head()
```

Out[29]:

	userId	movieId	rating	timestamp	Year	Month	Day	Hour	Minute	Second	DT
0	1	1	4.0	964982703	2000	7.0	30.0	14.0	45.0	3.0	2000-07-30 14:45:03
1	1	3	4.0	964981247	2000	NaN	NaN	NaN	NaN	NaN	2000-07-30 14:20:47
2	1	6	4.0	964982224	2000	NaN	NaN	NaN	NaN	NaN	2000-07-30 14:37:04
3	1	47	5.0	964983815	2000	NaN	NaN	NaN	NaN	NaN	2000-07-30 15:03:35
4	1	50	5.0	964982931	2000	NaN	NaN	NaN	NaN	NaN	2000-07-30 14:48:51

In [30]:
```
movies_df = pd.read_csv('Data/ml-latest-small/movies.csv', sep=',')
movies_df.head()
```

Out[30]:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

In [31]:
```
#1. Extract all movies whose list of genres contains "Comedy".
```

In [32]:
```
#From here we will work with the comedy_movies only
#this 'variable' contain all movies genres comedy
comedy_movies = movies_df[movies_df['genres'] == 'Comedy']
comedy_movies.head()
```

Out[32]:

	movieId	title	genres
4	5	Father of the Bride Part II (1995)	Comedy
17	18	Four Rooms (1995)	Comedy
18	19	Ace Ventura: When Nature Calls (1995)	Comedy
58	65	Bio-Dome (1996)	Comedy
61	69	Friday (1995)	Comedy

In [33]:

```
#2. Find the average user rating and the number of ratings for each comedy
```

In [51]:

```
def get_avg_rating(ratings_df, movie_id):
    filter_movie = (ratings_df['movieId'] == movie_id)
    ratings_movie = ratings_df[filter_movie]
    avg_rating = ratings_movie['rating'].mean()
    return avg_rating

def get_number_of_ratings(ratings_df, movie_id):
    ratings_movie = ratings_df[ratings_df['movieId'] == movie_id]
    #get the shape
    num_ratings = ratings_movie.shape[0]
    return num_ratings


for idx in comedy_movies.index:
    movie_id = comedy_movies.loc[idx, 'movieId']
    #print("movie_id",movie_id)
    avg_rating = get_avg_rating(ratings_df, movie_id)
    #print("avg_rating",avg_rating)
    num_ratings = get_number_of_ratings(ratings_df, movie_id)
    #print("num_ratings",num_ratings)
    #selecting things by label
    #loc[what row do i want, what column do i want]
    comedy_movies.loc[idx, 'AvgRating'] = avg_rating
    comedy_movies.loc[idx, 'NumRatings'] = num_ratings

#I displayed 100 to make sure that number of rating is working
#Movie Id 4499 does has 16 rates
comedy_movies.head(100)
```

Out[51]:

	movieId	title	genres	AvgRating	NumRatings
8813	130970	George Carlin: Life Is Worth Losing (2005)	Comedy	5.000000	1.0
6511	53578	Valet, The (La doublure) (2006)	Comedy	5.000000	1.0
4372	6402	Siam Sunset (1999)	Comedy	5.000000	1.0
8623	118834	National Lampoon's Bag Boy (2007)	Comedy	5.000000	1.0
8154	102217	Bill Hicks: Revelations (1993)	Comedy	5.000000	1.0
...	...	...	...	...	...
3324	4499	Dirty Rotten Scoundrels (1988)	Comedy	4.125000	16.0
2481	3306	Circus, The (1928)	Comedy	4.125000	4.0
2210	2937	Palm Beach Story, The (1942)	Comedy	4.100000	5.0
2622	3507	Odd Couple, The (1968)	Comedy	4.066667	15.0
938	1238	Local Hero (1983)	Comedy	4.055556	9.0

100 rows × 5 columns

In [45]:

```
#testing loc method
comedy_movies.loc[ : , 'AvgRating']
```

Out[45]:

```
8813    5.0
536     5.0
3067    5.0
3256    5.0
9122    5.0
       ...
5795    0.5
8984    0.5
6554    0.5
5409    0.5
5824    NaN
Name: AvgRating, Length: 946, dtype: float64
```

In [46]:

```
comedy_movies = comedy_movies.sort_values(by='AvgRating', ascending=False)
```

In [49]:

```
comedy_movies.head(10)
```

Out[49]:

	movieId	title	genres	AvgRating	NumRatings
8813	130970	George Carlin: Life Is Worth Losing (2005)	Comedy	5.0	1.0
6511	53578	Valet, The (La doublure) (2006)	Comedy	5.0	1.0
4372	6402	Siam Sunset (1999)	Comedy	5.0	1.0
8623	118834	National Lampoon's Bag Boy (2007)	Comedy	5.0	1.0
8154	102217	Bill Hicks: Revelations (1993)	Comedy	5.0	1.0
5435	25947	Unfaithfully Yours (1948)	Comedy	5.0	1.0
7525	84512	Girls About Town (1931)	Comedy	5.0	1.0
5942	34312	Calcium Kid, The (2004)	Comedy	5.0	1.0
9289	158398	World of Glory (1991)	Comedy	5.0	1.0
8788	129514	George Carlin: It's Bad for Ya! (2008)	Comedy	5.0	1.0

In [ ]:

```
#3. Remove movies with less than 50 ratings.
```

In [52]:

```
movies_with_less_than_50_rating = (comedy_movies['NumRatings'] < 50)
movies_todrop = comedy_movies[movies_with_less_than_50_rating]
comedy_movies = comedy_movies.drop(movies_todrop.index)
comedy_movies.head()
#NOTE: drop method/function will drop those movies with less than 50 ratng
```

Out[52]:

	movieId	title	genres	AvgRating	NumRatings
987	1288	This Is Spinal Tap (1984)	Comedy	4.015152	66.0
1074	1394	Raising Arizona (1987)	Comedy	3.991379	58.0
820	1080	Monty Python's Life of Brian (1979)	Comedy	3.926966	89.0
6537	54503	Superbad (2007)	Comedy	3.863636	55.0
2097	2791	Airplane! (1980)	Comedy	3.856322	87.0

In [ ]:

```
#4. Among the remaining movies, find the 5 highest-rated movies and the 5 l
```

In [55]:

```
# Sort comedy_movies by AvgRating.
highest_rating_movies_df = comedy_movies.sort_values(by='AvgRating',ascendi
lowest_rating_movies_df = comedy_movies.sort_values(by='AvgRating')
print("highest-rated movies:")
highest_rating_movies_df.head(5)
```

```
highest-rated movies:
```

Out[55]:

	movieId	title	genres	AvgRating	NumRatings
987	1288	This Is Spinal Tap (1984)	Comedy	4.015152	66.0
1074	1394	Raising Arizona (1987)	Comedy	3.991379	58.0
820	1080	Monty Python's Life of Brian (1979)	Comedy	3.926966	89.0
6537	54503	Superbad (2007)	Comedy	3.863636	55.0
2097	2791	Airplane! (1980)	Comedy	3.856322	87.0

In [56]:

```
print("lowest-rated movies:")
lowest_rating_movies_df.head(5)
```

```
lowest-rated movies:
```

Out[56]:

	movieId	title	genres	AvgRating	NumRatings
18	19	Ace Ventura: When Nature Calls (1995)	Comedy	2.727273	88.0
3903	5481	Austin Powers in Goldmember (2002)	Comedy	2.846154	65.0
1135	1485	Liar Liar (1997)	Comedy	3.033784	74.0
302	344	Ace Ventura: Pet Detective (1994)	Comedy	3.040373	161.0
455	520	Robin Hood: Men in Tights (1993)	Comedy	3.130435	69.0

```
In [ ]:
```