Movie Data Analysis

Soumik Chaudhuri
E-mail: contact@soumik.in

Soumik.in

Abstract

Explore movie ratings and find out if Action movies tend to be rated higher than Comedy movies.

Methods Used: Pandas DataFrame

Dataset

Data Source: MovieLens web site (filename: ml-25m.zip)

Location: https://grouplens.org/datasets/movielens/

Motivation

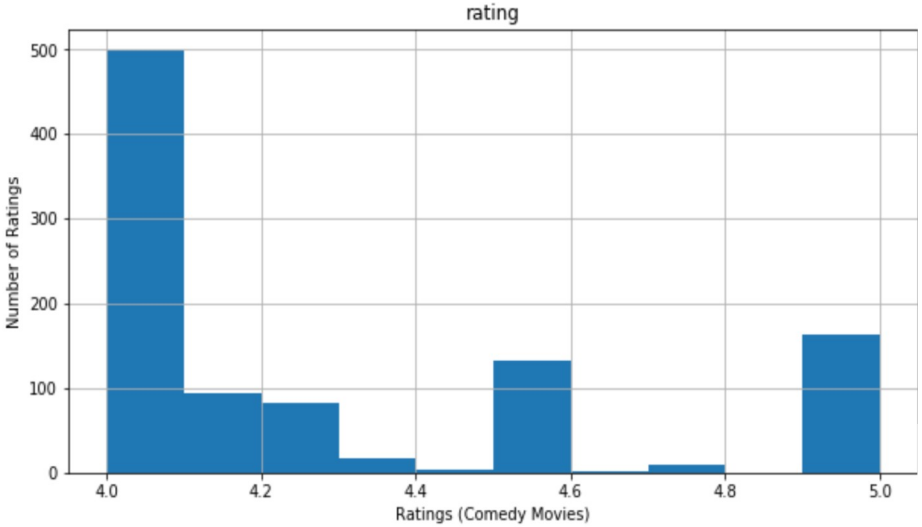
People often search for reviews and rating before watching a movie. It is crucial for a producer to understand which genres tend to be better rated.

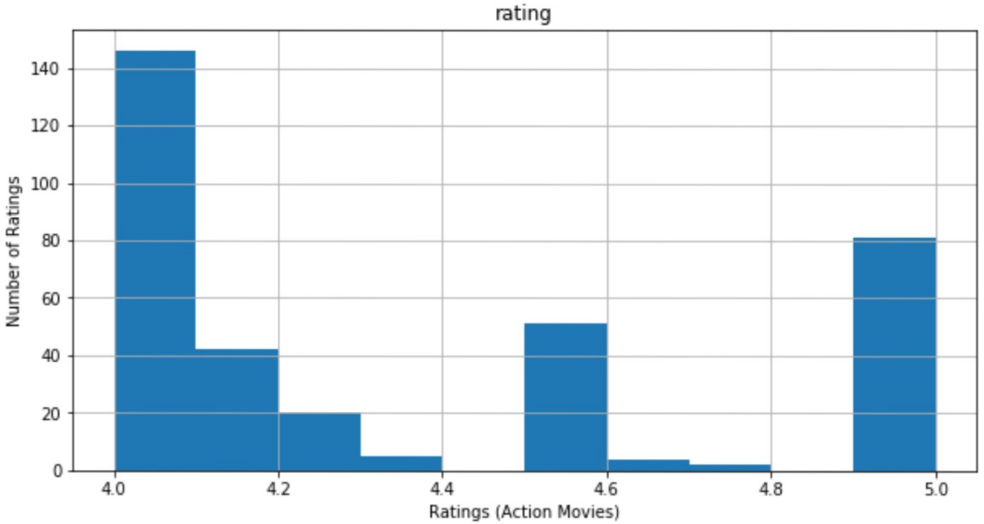
Research Question

Are action movies likely to be rated higher than comedy movies?

Findings

There are only 351 action movies with ratings of 4 or more than 4, whereas, there are 1004 comedy movies with these ratings.





Movie Data Analysis

June 12, 2020

Movie Data Analysis

Data Source: MovieLens web site (filename: ml-25m.zip)

Location: https://grouplens.org/datasets/movielens/

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

Exploring the data

[2]:
```
# Print the first ten rows

movies = pd.read_csv('./movielens/movies.csv', sep=',')
print(type(movies))
movies.head(10)
```

```
<class 'pandas.core.frame.DataFrame'>
```

[2]:

	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
5	6	Heat (1995)	Action|Crime|Thriller
6	7	Sabrina (1995)	Comedy|Romance
7	8	Tom and Huck (1995)	
8	9	Sudden Death (1995)	
9	10	GoldenEye (1995)	

```
7                          Adventure|Children
8                                      Action
9                  Action|Adventure|Thriller
```

[3]:
```
# Timestamps counted in seconds since midnight Coordinated Universal Time (UTC)␣
 ↪of January 1, 1970

tags = pd.read_csv('./movielens/tags.csv', sep=',')
tags.head(10)
```

[3]:

	userId	movieId	tag	timestamp
0	3	260	classic	1439472355
1	3	260	sci-fi	1439472256
2	4	1732	dark comedy	1573943598
3	4	1732	great dialogue	1573943604
4	4	7569	so bad it's good	1573943455
5	4	44665	unreliable narrators	1573943619
6	4	115569	tense	1573943077
7	4	115713	artificial intelligence	1573942979
8	4	115713	philosophical	1573943033
9	4	115713	tense	1573943042

[4]:
```
# Ratings database

ratings = pd.read_csv('./movielens/ratings.csv', sep=',',␣
 ↪parse_dates=['timestamp'])
ratings.head()
```

[4]:

	userId	movieId	rating	timestamp
0	1	296	5.0	1147880044
1	1	306	3.5	1147868817
2	1	307	5.0	1147868828
3	1	665	5.0	1147878820
4	1	899	3.5	1147868510

[5]:
```
# Data Cleaning

movies.isnull().any()
```

[5]:
```
movieId    False
title      False
genres     False
dtype: bool
```

[6]:
```
tags.isnull().any()
```

[6]:
```
userId       False
movieId      False
tag           True
timestamp    False
dtype: bool
```

[7]:
```
tags = tags.dropna()
tags.isnull().any()
```

[7]:
```
userId       False
movieId      False
tag          False
timestamp    False
dtype: bool
```

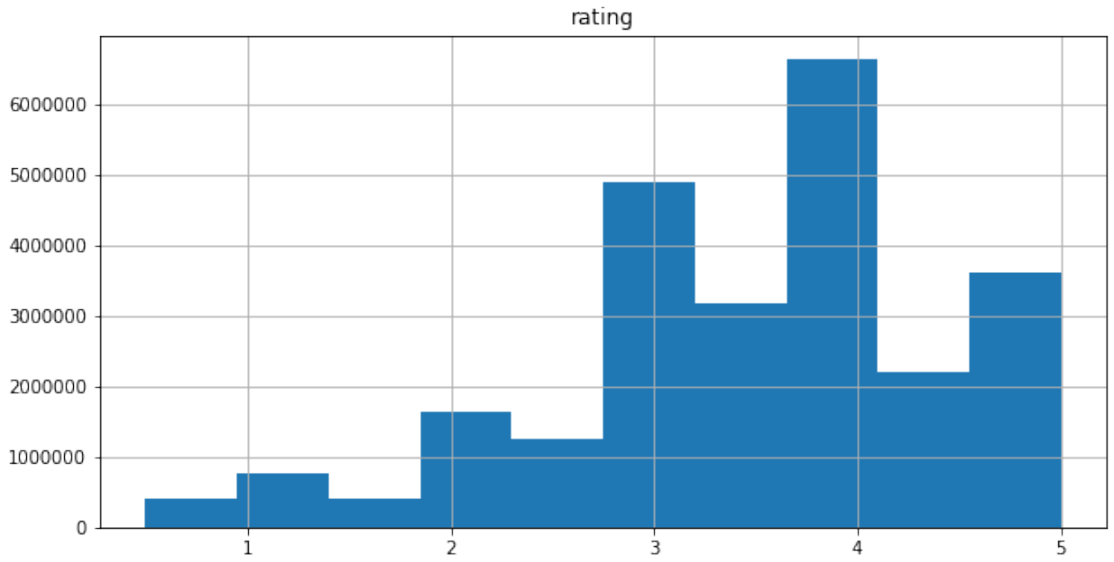
[8]:
```
ratings.isnull().any()
```

[8]:
```
userId       False
movieId      False
rating       False
timestamp    False
dtype: bool
```

Data Visualization

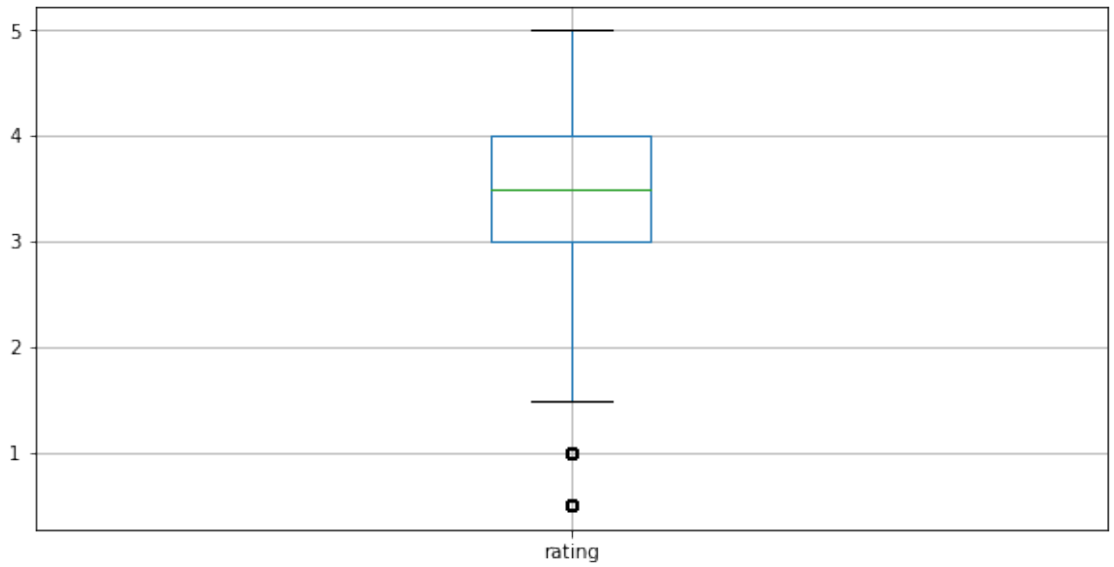
[9]:
```
%matplotlib inline

ratings.hist(column='rating', figsize=(10,5))
```

[9]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x0000028290AB3E08>]],
      dtype=object)
```



[10]:
```
ratings.boxplot(column='rating', figsize=(10,5))
```

[10]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2820136fdc8>
```

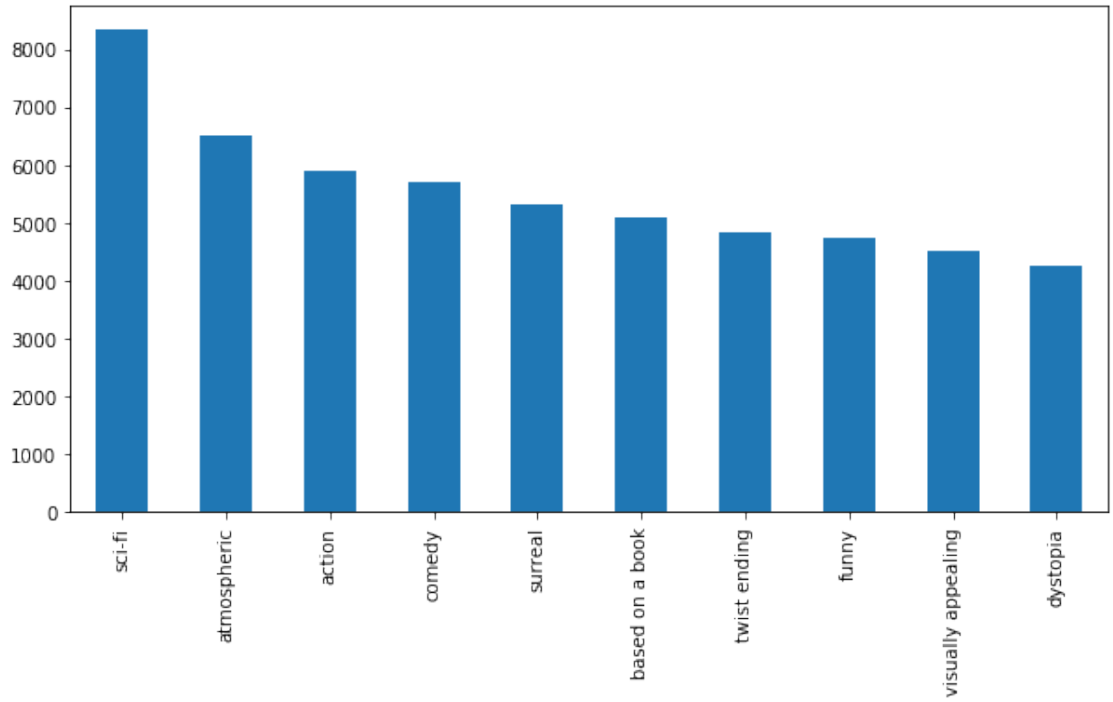


[11]:
```
tag_counts = tags['tag'].value_counts()
tag_counts[:10]
```

[11]:
```
sci-fi                4330
atmospheric           6516
action                5907
comedy                5702
surreal               5326
based on a book       5079
twist ending          4820
funny                 4738
visually appealing    4526
dystopia              4257
Name: tag, dtype: int64
```

[12]:
```
tag_counts[:10].plot(kind='bar', figsize=(10,5))
```

[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x282013fb288>
```



[13]:
```
is_highly_rated = ratings['rating'] >= 3.5

ratings[is_highly_rated][0:10]
```

[13]:

	userId	movieId	rating	timestamp
0	1	296	5.0	1147880044
1	1	306	3.5	1147868817
2	1	307	5.0	1147868828
3	1	665	5.0	1147878820
4	1	899	3.5	1147868510
5	1	1088	4.0	1147868495
6	1	1175	3.5	1147868826
7	1	1217	3.5	1147878326
8	1	1237	5.0	1147868839
9	1	1250	4.0	1147868414

[14]:
```
average_rating = ratings[['movieId','rating']].groupby('movieId').mean()
average_rating.tail()
```

[14]:

movieId	rating
209157	1.5
209159	3.0
209163	4.5

```
209169      3.0
209171      3.0
```

[15]:
```
ratings_count = ratings[['movieId','rating']].groupby('rating').count()
ratings_count
```

[15]:
```
        movieId
rating
0.5      393068
1.0      776815
1.5      399490
2.0     1640868
2.5     1262797
3.0     4896928
3.5     3177318
4.0     6639798
4.5     2200539
5.0     3612474
```

Merge Dataframes

[16]:
```
tags.head()
```

[16]:
```
   userId  movieId              tag   timestamp
0       3      260          classic  1439472355
1       3      260           sci-fi  1439472256
2       4     1732      dark comedy  1573943598
3       4     1732   great dialogue  1573943604
4       4     7569  so bad it's good  1573943455
```

[17]:
```
movies.head()
```

[17]:
```
   movieId                               title  \
0        1                    Toy Story (1995)
1        2                      Jumanji (1995)
2        3             Grumpier Old Men (1995)
3        4            Waiting to Exhale (1995)
4        5  Father of the Bride Part II (1995)

                                        genres
0  Adventure|Animation|Children|Comedy|Fantasy
1                   Adventure|Children|Fantasy
2                               Comedy|Romance
3                        Comedy|Drama|Romance
4                                       Comedy
```

[18]:
```
# Merge the dataframes on MovieID

t = movies.merge(tags, on='movieId', how='inner')
t.head()
```

[18]:
```
   movieId             title                                       genres  \
0        1  Toy Story (1995)  Adventure|Animation|Children|Comedy|Fantasy
1        1  Toy Story (1995)  Adventure|Animation|Children|Comedy|Fantasy
2        1  Toy Story (1995)  Adventure|Animation|Children|Comedy|Fantasy
3        1  Toy Story (1995)  Adventure|Animation|Children|Comedy|Fantasy
4        1  Toy Story (1995)  Adventure|Animation|Children|Comedy|Fantasy

   userId           tag   timestamp
0     791         Owned  1515175493
1    1048  imdb top 250  1172144394
2    1361         Pixar  1216146311
3    3164         Pixar  1223304727
4    3164   time travel  1223304729
```

[19]:
```
avg_ratings = ratings.groupby('movieId', as_index=False).mean()
del avg_ratings['userId']
avg_ratings.head()
```

[19]:
```
   movieId    rating
0        1  3.893708
1        2  3.251527
2        3  3.142028
3        4  2.853547
4        5  3.058434
```

[20]:
```
box_office = movies.merge(avg_ratings, on='movieId', how='inner')
box_office.tail()
```

[20]:
```
       movieId                          title                 genres  rating
59042   209157                     We (2018)                  Drama     1.5
59043   209159    Window of the Soul (2001)            Documentary     3.0
59044   209163              Bad Poems (2018)           Comedy|Drama     4.5
59045   209169         A Girl Thing (2001)      (no genres listed)     3.0
59046   209171  Women of Devil's Island (1962)  Action|Adventure|Drama     3.0
```

[21]:
```
is_highly_rated = box_office['rating'] >= 4.0

box_office[is_highly_rated][0:10]
```

[21]:
```
     movieId                                    title  \
27        28                        Persuasion (1995)
46        47              Seven (a.k.a. Se7en) (1995)
```

```
49        50                      Usual Suspects, The (1995)
108      110                        Braveheart (1995)
109      111                       Taxi Driver (1976)
160      162                             Crumb (1994)
211      213  Burnt by the Sun (Utomlyonnye solntsem) (1994)
212      214           Before the Rain (Pred dozhdot) (1994)
229      232     Eat Drink Man Woman (Yin shi nan nu) (1994)
243      246                       Hoop Dreams (1994)

                    genres    rating
27            Drama|Romance  4.030000
46        Mystery|Thriller  4.079166
49   Crime|Mystery|Thriller  4.284353
108       Action|Drama|War  4.002273
109    Crime|Drama|Thriller  4.083479
160             Documentary  4.008077
211                   Drama  4.016543
212             Drama|War    4.026966
229    Comedy|Drama|Romance  4.027936
243             Documentary  4.028237
```

[22]:
```
is_comedy = box_office['genres'].str.contains('Comedy')

box_office[is_comedy][:5]
```

[22]:
```
   movieId                               title  \
0        1                    Toy Story (1995)
2        3             Grumpier Old Men (1995)
3        4            Waiting to Exhale (1995)
4        5  Father of the Bride Part II (1995)
6        7                      Sabrina (1995)

                                       genres    rating
0  Adventure|Animation|Children|Comedy|Fantasy  3.893708
2                             Comedy|Romance  3.142028
3                      Comedy|Drama|Romance  2.853547
4                                    Comedy  3.058434
6                             Comedy|Romance  3.363666
```

[23]:
```
is_action = box_office['genres'].str.contains('Action')

box_office[is_action][:5]
```

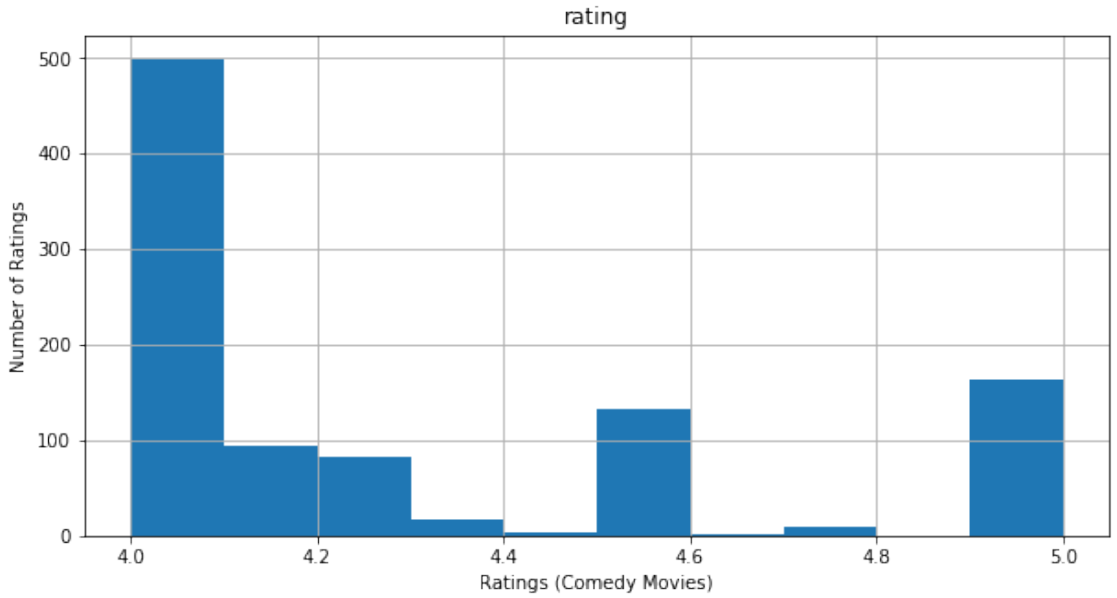
[23]:
```
    movieId                    title                                genres  \
5         6             Heat (1995)               Action|Crime|Thriller
8         9      Sudden Death (1995)                              Action
9        10         GoldenEye (1995)          Action|Adventure|Thriller
```

```
14       15  Cutthroat Island (1995)            Action|Adventure|Romance
19       20       Money Train (1995)  Action|Comedy|Crime|Drama|Thriller

      rating
5   3.854909
8   2.992051
9   3.421458
14  2.719022
19  2.869922
```

[24]:
```
box_office[is_comedy & is_highly_rated][:]
```

[24]:
```
       movieId                                              title  \
229        232        Eat Drink Man Woman (Yin shi nan nu) (1994)
292        296                                     Pulp Fiction (1994)
351        356                                     Forrest Gump (1994)
600        608                                             Fargo (1996)
705        720  Wallace & Gromit: The Best of Aardman Animatio…
…          …                                                   …
58990   208911                            Cheating in Chains (2006)
58998   208939                                         Klaus (2019)
59001   208945                                        Powder (2019)
59041   209155                          Santosh Subramaniam (2008)
59044   209163                                   Bad Poems (2018)

                                    genres    rating
229                  Comedy|Drama|Romance  4.027936
292             Comedy|Crime|Drama|Thriller  4.188912
351                Comedy|Drama|Romance|War  4.048011
600             Comedy|Crime|Drama|Thriller  4.111421
705              Adventure|Animation|Comedy  4.103381
…                                       …       …
58990                              Comedy  4.000000
58998  Adventure|Animation|Children|Comedy  4.312500
59001                      Comedy|Drama  4.500000
59041                Action|Comedy|Romance  5.000000
59044                      Comedy|Drama  4.500000

[1004 rows x 4 columns]
```

[25]:
```
comedy_ratings = box_office[is_comedy & is_highly_rated][:]
comedy_ratings.hist(column='rating', figsize=(10,5))
plt.ylabel('Number of Ratings')
plt.xlabel('Ratings (Comedy Movies)')
plt.show()
```



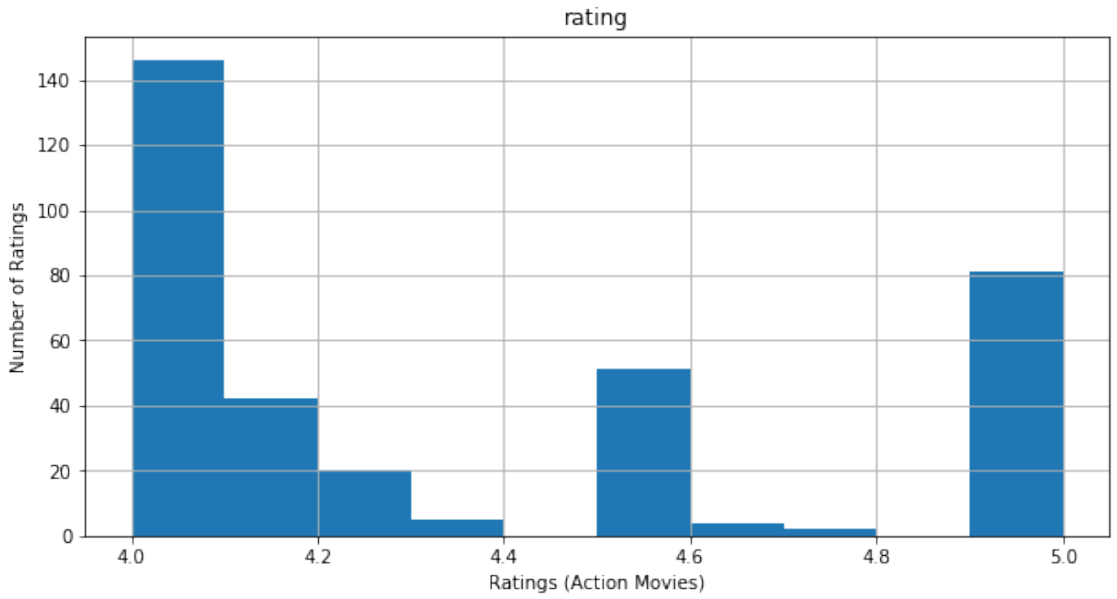
[26]:
```
box_office[is_action & is_highly_rated][:]
```

[26]:

	movieId	title	genres	rating
108	110	Braveheart (1995)	Action|Drama|War	4.002273
257	260	Star Wars: Episode IV - A New Hope (1977)	Action|Adventure|Sci-Fi	4.120189
289	293	Léon: The Professional (a.k.a. The Professiona…	Action|Crime|Drama|Thriller	4.088599
536	541	Blade Runner (1982)	Action|Sci-Fi|Thriller	4.115838
887	908	North by Northwest (1959)	Action|Adventure|Mystery|Romance|Thriller	4.196617
…	…	…	…	…
58882	208465	Jallikattu (2019)	Action|Crime|Drama|Thriller	4.500000
58887	208477	Kaithi (2019)	Action|Thriller	5.000000
58928	208691	Chinatown Squad (1935)	Action|Crime	4.000000
58953	208767	Mobile Suit Gundam I (1981)	Action|Adventure|Animation|Drama|Sci-Fi|War	4.000000
59041	209155	Santosh Subramaniam (2008)	Action|Comedy|Romance	5.000000

```
[351 rows x 4 columns]
```

[27]:
```
action_ratings = box_office[is_action & is_highly_rated][:]
action_ratings.hist(column='rating', figsize=(10,5))
plt.ylabel('Number of Ratings')
plt.xlabel('Ratings (Action Movies)')
plt.show()
```



[ ]: