In [3]:

```
#import the necessary libraries
import pandas as pd
import numpy as np
import sqlite3
import matplotlib.pyplot as plt
import seaborn as sea
import matplotlib.ticker as mtick
%matplotlib inline
```

In [4]:

```
#accessing the data
file_path_01 = 'zippedData/bom.movie_gross.csv.gz'
file_path_02 = 'zippedData/rt.movie_info.tsv.gz'
file_path_03 = 'zippedData/rt.reviews.tsv.gz'
file_path_04 = 'zippedData/tmdb.movies.csv.gz'
file_path_05 = 'zippedData/tn.movie_budgets.csv.gz'
bom_movie_gross = pd.read_csv(file_path_01)
rt_movie_info = pd.read_csv(file_path_02, sep="\t", index_col = 0)
#rt_reviews = pd.read_csv(file_path_03, sep = "\t")
tmdb_movies =pd.read_csv(file_path_04, index_col = 0)
tn_movie_budgets = pd.read_csv(file_path_05, index_col = 0)
db = 'im.db'
conn = sqlite3.connect(db)
```

- **scouting the data as we identify key information and having an overview of what we'll be working with.**

In [5]:

```
bom_movie_gross.head()
```

Out[5]:

	title	studio	domestic_gross	foreign_gross	year
0	Toy Story 3	BV	415000000.0	652000000	2010
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010
3	Inception	WB	292600000.0	535700000	2010
4	Shrek Forever After	P/DW	238700000.0	513900000	2010

In [6]:

```
rt_movie_info.head()
```

Out[6]:

id	synopsis	rating	genre	director	writer	theater_date	dvd_date	currency	box_offic
1	This gritty, fast-paced, and innovative police...	R	Action and Adventure|Classics|Drama	William Friedkin	Ernest Tidyman	Oct 9, 1971	Sep 25, 2001	NaN	Nal
3	New York City, not-too-distant-future: Eric Pa...	R	Drama|Science Fiction and Fantasy	David Cronenberg	David Cronenberg|Don DeLillo	Aug 17, 2012	Jan 1, 2013	$	600,00
	Illeana								

id	synopsis	rating	genre	director	writer	theater_date	dvd_date	currency	box_offic
5	Douglas delivers a superb performance ...	R	Drama|Musical and Performing Arts	Allison Anders	Allison Anders	Sep 13, 1996	Apr 18, 2000	NaN	Nal
6	Michael Douglas runs afoul of a treacherous su...	R	Drama|Mystery and Suspense	Barry Levinson	Paul Attanasio|Michael Crichton	Dec 9, 1994	Aug 27, 1997	NaN	Nal
7	NaN	NR	Drama|Romance	Rodney Bennett	Giles Cooper	NaN	NaN	NaN	Nal

In [7]:

```
tmdb_movies.head()
```

Out[7]:

	genre_ids	id	original_language	original_title	popularity	release_date	title	vote_average	vote_count
0	[12, 14, 10751]	12444	en	Harry Potter and the Deathly Hallows: Part 1	33.533	2010-11-19	Harry Potter and the Deathly Hallows: Part 1	7.7	10788
1	[14, 12, 16, 10751]	10191	en	How to Train Your Dragon	28.734	2010-03-26	How to Train Your Dragon	7.7	7610
2	[12, 28, 878]	10138	en	Iron Man 2	28.515	2010-05-07	Iron Man 2	6.8	12368
3	[16, 35, 10751]	862	en	Toy Story	28.005	1995-11-22	Toy Story	7.9	10174
4	[28, 878, 12]	27205	en	Inception	27.920	2010-07-16	Inception	8.3	22186

In [8]:

```
tn_movie_budgets.head()
```

Out[8]:

id	release_date	movie	production_budget	domestic_gross	worldwide_gross
1	Dec 18, 2009	Avatar	$425,000,000	$760,507,625	$2,776,345,279
2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$410,600,000	$241,063,875	$1,045,663,875
3	Jun 7, 2019	Dark Phoenix	$350,000,000	$42,762,350	$149,762,350
4	May 1, 2015	Avengers: Age of Ultron	$330,600,000	$459,005,868	$1,403,013,963
5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi	$317,000,000	$620,181,382	$1,316,721,747

In [9]:

```
movie_basics = pd.read_sql("""
SELECT *
FROM movie_basics
""", conn)
```

In [10]:

```
movie_basics.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 146144 entries, 0 to 146143
```

```
RangeIndex: 146144 entries, 0 to 146143
Data columns (total 6 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   movie_id         146144 non-null  object
 1   primary_title    146144 non-null  object
 2   original_title   146123 non-null  object
 3   start_year       146144 non-null  int64
 4   runtime_minutes  114405 non-null  float64
 5   genres           140736 non-null  object
dtypes: float64(1), int64(1), object(4)
memory usage: 6.7+ MB
```

In [11]:

```
# now that we have a general overview,let's start working on the IMDB Data
#we start using the IMDB data
highest_rated = pd.read_sql("""
SELECT original_title, genres, averagerating, numvotes
FROM movie_basics
JOIN movie_ratings
    USING(movie_id)
ORDER BY averagerating DESC
""", conn)
```

In [12]:

```
highest_rated.head(25)
```

Out[12]:

	original_title	genres	averagerating	numvotes
0	Exteriores: Mulheres Brasileiras na Diplomacia	Documentary	10.0	5
1	The Dark Knight: The Ballad of the N Word	Comedy,Drama	10.0	5
2	Freeing Bernie Baran	Crime,Documentary	10.0	5
3	Hercule contre Hermès	Documentary	10.0	5
4	I Was Born Yesterday!	Documentary	10.0	6
5	Dog Days in the Heartland	Drama	10.0	5
6	Revolution Food	Documentary	10.0	8
7	Fly High: Story of the Disc Dog	Documentary	10.0	7
8	All Around Us	Documentary	10.0	6
9	Atlas Mountain: Barbary Macaques - Childcaring...	Documentary	10.0	5
10	Requiem voor een Boom	Documentary	10.0	5
11	A Dedicated Life: Phoebe Brand Beyond the Group	Documentary	10.0	5
12	Ellis Island: The Making of a Master Race in A...	Documentary,History	10.0	6
13	Calamity Kevin	Adventure,Comedy	10.0	6
14	Pick It Up! - Ska in the '90s	Documentary	10.0	5
15	Renegade	Documentary	10.0	20
16	Gini Helida Kathe	Drama	9.9	417
17	The Wedding Present: Something Left Behind	Documentary	9.9	8
18	LA Foodways	Documentary	9.9	8
19	Moscow we will lose	Documentary	9.9	18
20	Wild Karnataka	Documentary	9.9	10
21	Dreaming of a Vetter World	Documentary	9.8	6
22	Grisaia: Phantom trigger the animation 02. Sou...	Action	9.8	5
23	Mujeres republicanas	Documentary	9.8	5
24	Send My Mail to Nashville	Biography,Documentary,Music	9.8	5

original_title	genres	averagerating	numvotes

In [13]:

```
print(highest_rated['numvotes'].mean())
#arbitrarily choose 2000
```

```
3523.6621669194105
```

In [14]:

```
# let's now look for the highest rated movies
# we sieve through by picking numvotes be greater than 2000
print(highest_rated['numvotes'].mean())
#we'll arbitrarily choose 2000
highest_rated = pd.read_sql("""
SELECT original_title, genres, averagerating, numvotes
FROM movie_basics
JOIN movie_ratings
    USING(movie_id)
WHERE numvotes > 2000
ORDER BY averagerating DESC
""", conn)
highest_rated.head(25)
```

```
3523.6621669194105
```

Out[14]:

	original_title	genres	averagerating	numvotes
0	Once Upon a Time ... in Hollywood	Comedy,Drama	9.7	5600
1	Ekvtime: Man of God	Biography,Drama,History	9.6	2604
2	Aloko Udapadi	Drama,History	9.5	6509
3	Peranbu	Drama	9.4	9629
4	Dag II	Action,Drama,War	9.3	100568
5	Aynabaji	Crime,Mystery,Thriller	9.3	18470
6	Wheels	Drama	9.3	17308
7	Natsamrat	Drama,Family	9.2	4297
8	C/o Kancharapalem	Drama	9.2	2195
9	CM101MMXI Fundamentals	Comedy,Documentary	9.2	41560
10	On vam ne Dimon	Documentary	9.2	2721
11	A Man Called Ahok	Drama	9.1	4162
12	Oggatonama	Drama	9.1	2973
13	Pariyerum Perumal	Drama	9.0	4854
14	Yowis Ben	Comedy,Drama	9.0	2992
15	Tylko nie mów nikomu	Documentary	8.9	2111
16	Godhi Banna Sadharana Mykattu	Drama,Family	8.9	2001
17	A Billion Lives	Documentary,History,News	8.9	2715
18	O.J.: Made in America	Biography,Crime,Documentary	8.9	14946
19	Burn the Stage: The Movie	Documentary,Music	8.8	2067
20	Les Misérables in Concert: The 25th Anniversary	Drama,Music,Musical	8.8	4583
21	Drishyam	Crime,Drama,Thriller	8.8	24326
22	Ratsasan	Action,Crime,Thriller	8.8	10518
23	Aruvi	Drama	8.8	8277
24	Kill Bill: The Whole Bloody Affair	Action,Crime,Thriller	8.8	3406

In [15]:

In [15]:

```
movie_basics['primary_title'].duplicated().sum()
# 10073 is the total movies duplicated despite having different movie ID's
```

Out[15]:

10073

In [16]:

```
movie_basics.describe()
```

Out[16]:

	start_year	runtime_minutes
count	146144.000000	114405.000000
mean	2014.621798	86.187247
std	2.733583	166.360590
min	2010.000000	1.000000
25%	2012.000000	70.000000
50%	2015.000000	87.000000
75%	2017.000000	99.000000
max	2115.000000	51420.000000

movie_id is the primary key and there are 146,144 movies We narrowed down our scope to the US only .

In [17]:

```
#we are going to narrow down our scope to the US region and language as  English to elimi
nate documentaries.
highest_rated = pd.read_sql("""
SELECT original_title, genres, averagerating, numvotes, region
FROM movie_basics
INNER JOIN movie_ratings
    USING(movie_id)
INNER JOIN movie_akas
    USING(movie_id)
WHERE numvotes > 2000 AND genres NOT LIKE "%Documentary%" AND region = "US"
ORDER BY averagerating DESC
""", conn)
highest_rated.drop_duplicates(subset = 'original_title', inplace=True)
highest_rated.head(25)
```

Out[17]:

	original_title	genres	averagerating	numvotes	region
0	Once Upon a Time ... in Hollywood	Comedy,Drama	9.7	5600	US
2	Peranbu	Drama	9.4	9629	US
3	Wheels	Drama	9.3	17308	US
4	Inception	Action,Adventure,Sci-Fi	8.8	1841066	US
8	Kill Bill: The Whole Bloody Affair	Action,Crime,Thriller	8.8	3406	US
9	Avengers: Endgame	Action,Adventure,Sci-Fi	8.8	441135	US
12	96	Drama,Romance	8.8	10903	US
13	Super Deluxe	Action,Crime,Drama	8.8	2254	US
14	Mahanati	Biography,Drama	8.7	6917	US
15	Interstellar	Adventure,Drama,Sci-Fi	8.6	1299334	US
19	Uri: The Surgical Strike	Action,Drama,War	8.6	30292	US
20	Yatra	Biography,Drama	8.6	2913	US

	original_title	genres	averagerating	numvotes	region
21	Rangasthalam	Action,Drama	8.6	15407	US
22	An Hour to Kill	Action,Comedy,Horror	8.6	2302	US
27	Intouchables	Biography,Comedy,Drama	8.5	677343	US
28	Whiplash	Drama,Music	8.5	616916	US
29	Thani Oruvan	Action,Crime,Thriller	8.5	13747	US
30	Capharnaüm	Drama	8.5	20215	US
31	Dhuruvangal Pathinaaru	Action,Crime,Mystery	8.5	8560	US
32	Avengers: Infinity War	Action,Adventure,Sci-Fi	8.5	670926	US
36	Dangal	Action,Biography,Drama	8.5	123638	US
38	Andhadhun	Crime,Thriller	8.5	43409	US
39	Spider-Man: Into the Spider-Verse	Action,Adventure,Animation	8.5	210869	US
41	Coco	Adventure,Animation,Comedy	8.4	277194	US
43	Django Unchained	Drama,Western	8.4	1211405	US

Since we limited the data to US we are obligated to limit the gross as well

In [18]:

```
bom_movie_gross.sort_values('domestic_gross')
```

Out[18]:

	title	studio	domestic_gross	foreign_gross	year
1476	Storage 24	Magn.	100.0	NaN	2013
2321	The Chambermaid	FM	300.0	NaN	2015
2756	News From Planet Mars	KL	300.0	NaN	2016
2757	Satanic	Magn.	300.0	NaN	2016
1018	Apartment 143	Magn.	400.0	426000	2012
...	...	...	...	...	...
1975	Surprise - Journey To The West	AR	NaN	49600000	2015
2392	Finding Mr. Right 2	CL	NaN	114700000	2016
2468	Solace	LGP	NaN	22400000	2016
2595	Viral	W/Dim.	NaN	552000	2016
2825	Secret Superstar	NaN	NaN	122000000	2017

3387 rows × 5 columns

we are now coming acros NaN issues.we shall figure out how to remove them once we check the severity

In [19]:

```
# NaN
bom_movie_gross.isna().sum()
bom_movie_gross.drop('foreign_gross', axis=1 , inplace=True)
```

In [20]:

```
bom_movie_gross = bom_movie_gross.dropna()
bom_movie_gross = bom_movie_gross.sort_values('domestic_gross', ascending=False)
bom_movie_gross
```

Out[20]:

	title	studio	domestic_gross	year

	title	studio	domestic_gross	year
1872	Star Wars: The Force Awakens	BV	936700000.0	2015
3080	Black Panther	BV	700100000.0	2018
3079	Avengers: Infinity War	BV	678800000.0	2018
1873	Jurassic World	Uni.	652300000.0	2015
727	Marvel's The Avengers	BV	623400000.0	2012
...	...	...	...	...
1018	Apartment 143	Magn.	400.0	2012
2757	Satanic	Magn.	300.0	2016
2756	News From Planet Mars	KL	300.0	2016
2321	The Chambermaid	FM	300.0	2015
1476	Storage 24	Magn.	100.0	2013

3356 rows × 4 columns

In [21]:

```
gross_and_rating_df = bom_movie_gross.merge(highest_rated, how='inner', left_on='title',
right_on='original_title')
gross_and_rating_df.head(25)
gross_and_rating_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1407 entries, 0 to 1406
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   title           1407 non-null   object 
 1   studio          1407 non-null   object 
 2   domestic_gross  1407 non-null   float64
 3   year            1407 non-null   int64  
 4   original_title  1407 non-null   object 
 5   genres          1407 non-null   object 
 6   averagerating   1407 non-null   float64
 7   numvotes        1407 non-null   int64  
 8   region          1407 non-null   object 
dtypes: float64(2), int64(2), object(5)
memory usage: 109.9+ KB
```

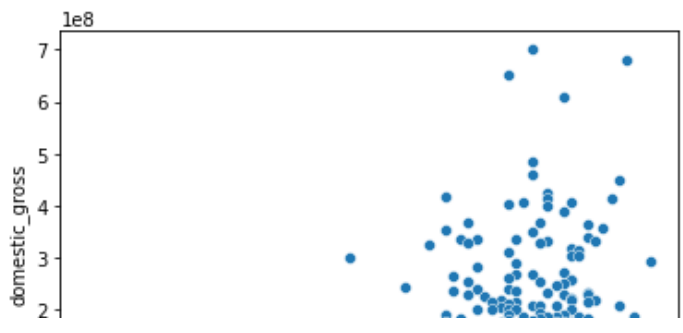
In [22]:

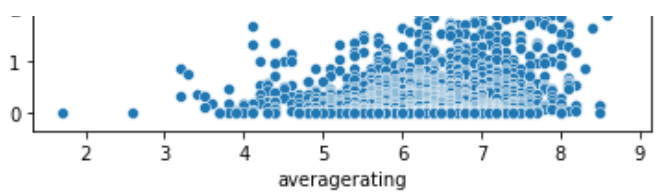
```
#gathering a rough overview
y = gross_and_rating_df['domestic_gross']
x = gross_and_rating_df['averagerating']
sea.scatterplot(x, y)
```

```
d:\Anaconda\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the followin
g variables as keyword args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keyword will result in an
error or misinterpretation.
  warnings.warn(
```

Out[22]:

```
<AxesSubplot:xlabel='averagerating', ylabel='domestic_gross'>
```





In [23]:

```
genres_df = gross_and_rating_df['genres']
genres_df = genres_df.str.split(',')
genres_list_all = genres_df.tolist()
genres_dict_all = {}
for x in genres_list_all:
    for y in x:
        if y not in genres_dict_all:
            genres_dict_all[y] = 1
        else:
            genres_dict_all[y] += 1
```

In [24]:

```
genres_df = gross_and_rating_df['genres'].iloc[:200]
genres_df = genres_df.str.split(',')
genres_list = genres_df.tolist()
genres_dict = {}
for x in genres_list:
    if y not in genres_dict:
        genres_dict[y] = 1
    else:
        genres_dict[y] += 1
```

In [25]:

```
genres_series = pd.Series(genres_dict_all)
genres_series.sort_values(ascending=False, inplace=True)
genres_series
```

Out[25]:

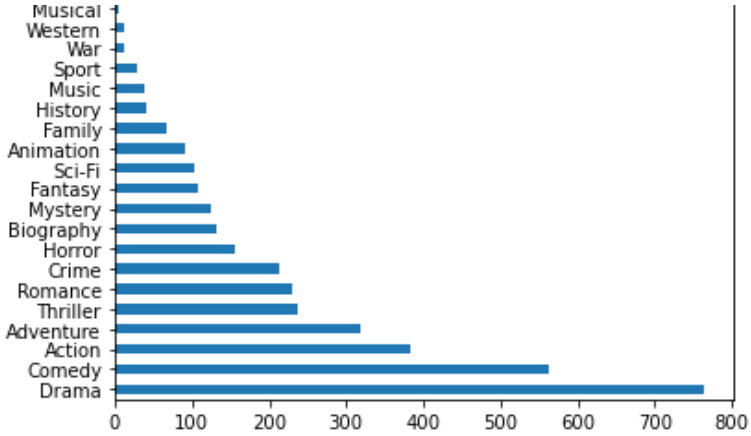
```
Drama          763
Comedy         562
Action         384
Adventure      319
Thriller       237
Romance        231
Crime          213
Horror         155
Biography      132
Mystery        124
Fantasy        107
Sci-Fi         102
Animation       91
Family          67
History         41
Music           39
Sport           28
War             12
Western         11
Musical          5
dtype: int64
```

In [26]:

```
genres_series_all = pd.Series(genres_dict_all)
genres_series_all.sort_values(ascending=False,  inplace=True)
genres_series_all.plot.barh()
```

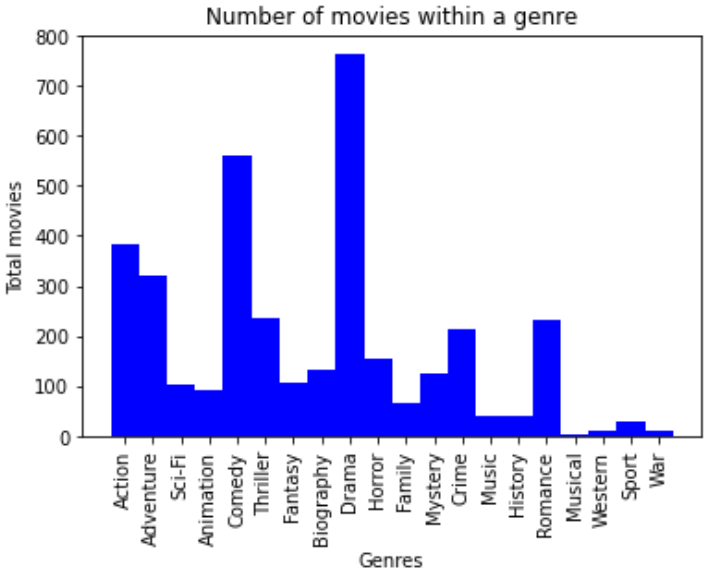
Out[26]:

```
<AxesSubplot:>
```



In [27]:

```
width = 1
plt.bar(genres_dict_all.keys(), genres_dict_all.values(), width, color='blue')
plt.title('Number of movies within a genre')
plt.xlabel('Genres')
plt.ylabel('Total movies')
plt.xticks(rotation = 90);
```



Based on the graph above: The first recommendation would be to make an,action,adventure,Drama, comedy film

we shall narrow down our data further to 2016, 2017 and 2018 that have great results

In [28]:

```
# let's now combine tables and decipher

tn_movie_budgets.head()
```

Out[28]:

id	release_date	movie	production_budget	domestic_gross	worldwide_gross
1	Dec 18, 2009	Avatar	$425,000,000	$760,507,625	$2,776,345,279
2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$410,600,000	$241,063,875	$1,045,663,875
3	Jun 7, 2019	Dark Phoenix	$350,000,000	$42,762,350	$149,762,350
4	May 1, 2015	Avengers: Age of Ultron	$330,600,000	$459,005,868	$1,403,013,963
5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi	$317,000,000	$620,181,382	$1,316,721,747

In [29]:

```
tn_movie_budgets_current = tn_movie_budgets.sort_values('release_date', ascending=False)
release_date = pd.to_datetime(tn_movie_budgets_current['release_date'])
tn_movie_budgets_current['release_date'] = release_date
tn_movie_budgets_current.sort_values('release_date', ascending = False, inplace = True)
tn_movie_budgets_current.head(25)
```

Out[29]:

id	release_date	movie	production_budget	domestic_gross	worldwide_gross
6	2020-12-31	Hannibal the Conqueror	$50,000,000	$0	$0
95	2020-12-31	Moonfall	$150,000,000	$0	$0
36	2020-02-21	Call of the Wild	$82,000,000	$0	$0
30	2019-12-31	Reagan	$25,000,000	$0	$0
81	2019-12-31	Army of the Dead	$90,000,000	$0	$0
72	2019-12-31	355	$75,000,000	$0	$0
13	2019-12-31	Rogue City	$13,000,000	$0	$0
16	2019-12-31	Eli	$11,000,000	$0	$0
44	2019-12-31	Down Under Cover	$40,000,000	$0	$0
8	2019-11-22	The Rhythm Section	$50,000,000	$0	$0
53	2019-11-08	Midway	$59,500,000	$0	$0
7	2019-11-08	Arctic Dogs	$50,000,000	$0	$0
30	2019-09-30	Unhinged	$29,000,000	$0	$0
9	2019-09-20	Ad Astra	$49,800,000	$0	$0
43	2019-09-13	The Goldfinch	$40,000,000	$0	$0
71	2019-08-30	PLAYMOBIL	$75,000,000	$0	$0
64	2019-08-14	Blinded by the Light	$15,000,000	$0	$0
16	2019-07-12	Crawl	$17,000,000	$0	$0
48	2019-06-21	Burn Your Maps	$8,000,000	$0	$0
39	2019-06-21	Kursk	$40,000,000	$0	$4,212,799
42	2019-06-14	Men in Black: International	$110,000,000	$3,100,000	$3,100,000
98	2019-06-14	Shaft	$30,000,000	$600,000	$600,000
81	2019-06-07	The Secret Life of Pets 2	$80,000,000	$63,795,655	$113,351,496
3	2019-06-07	Dark Phoenix	$350,000,000	$42,762,350	$149,762,350
35	2019-06-07	Late Night	$4,000,000	$246,305	$246,305

In [30]:

```
# the too recent ones will have to be dropped since they have $0 listed
tn_movie_budgets_current.drop(index=tn_movie_budgets_current.index[:20], axis=0, inplace
=True)
tn_movie_budgets_current.head(25)
```

Out[30]:

id	release_date	movie	production_budget	domestic_gross	worldwide_gross
42	2019-06-14	Men in Black: International	$110,000,000	$3,100,000	$3,100,000
98	2019-06-14	Shaft	$30,000,000	$600,000	$600,000
3	2019-06-07	Dark Phoenix	$350,000,000	$42,762,350	$149,762,350

35	2019-06-07	Late Night	$1,000,000	$249,685	$249,685
25	2019-05-31	Godzilla: King of the Monsters	$170,000,000	$85,576,941	$299,276,941
66	2019-05-31	MA	$5,000,000	$36,049,540	$44,300,625
96	2019-05-17	The Sun is Also a Star	$9,000,000	$4,950,029	$5,434,029
26	2019-05-10	The Professor and the Madman	$25,000,000	$0	$5,227,233
76	2019-05-10	PokÃ©mon: Detective Pikachu	$150,000,000	$139,507,806	$411,258,433
75	2019-05-03	Long Shot	$40,000,000	$30,202,860	$43,711,031
38	2019-04-23	Living Dark: The Story of Ted the Caver	$1,750,000	$0	$0
77	2019-04-12	Hellboy	$50,000,000	$21,903,748	$40,725,492
61	2019-04-05	Pet Sematary	$21,000,000	$54,724,696	$109,501,146
97	2019-04-05	Shazam!	$85,000,000	$139,606,856	$362,899,733
34	2019-04-05	The Best of Enemies	$10,000,000	$10,205,616	$10,205,616
33	2019-03-29	Unplanned	$6,000,000	$18,107,621	$18,107,621
22	2019-03-29	Dumbo	$170,000,000	$113,883,318	$345,004,422
88	2019-03-22	Us	$20,000,000	$175,006,930	$254,210,310
91	2019-03-15	Five Feet Apart	$7,000,000	$45,729,221	$80,504,421
97	2019-03-15	Captive State	$25,000,000	$5,958,315	$8,993,300
94	2019-03-15	Wonder Park	$100,000,000	$45,216,793	$115,149,422
96	2019-03-08	Captain Marvel	$175,000,000	$426,525,952	$1,123,061,550
56	2019-02-22	How to Train Your Dragon: The Hidden World	$129,000,000	$160,791,800	$519,258,283
86	2019-02-14	Fighting With My Family	$11,000,000	$22,958,583	$39,049,922
24	2019-02-14	Alita: Battle Angel	$170,000,000	$85,710,210	$402,976,036

In [31]:

```
#movies with more than 2000 votes and remove documentaries as before
latest_imdb = pd.read_sql("""
SELECT original_title, start_year, genres, averagerating, numvotes
FROM movie_basics
JOIN movie_ratings
    USING(movie_id)
WHERE numvotes > 2000 AND genres NOT LIKE "%Documentary%"
ORDER BY start_year DESC, averagerating DESc
""", conn
)
latest_imdb.head(25)
```

Out[31]:

	original_title	start_year	genres	averagerating	numvotes
0	Once Upon a Time ... in Hollywood	2019	Comedy,Drama	9.7	5600
1	Avengers: Endgame	2019	Action,Adventure,Sci-Fi	8.8	441135
2	Super Deluxe	2019	Action,Crime,Drama	8.8	2254
3	Uri: The Surgical Strike	2019	Action,Drama,War	8.6	30292
4	Yatra	2019	Biography,Drama	8.6	2913
5	The Tashkent Files	2019	Drama,Mystery,Thriller	8.4	3175
6	Gully Boy	2019	Drama,Music	8.3	17483
7	Badla	2019	Crime,Drama,Mystery	8.1	9988
8	John Wick: Chapter 3 - Parabellum	2019	Action,Crime,Thriller	8.0	81568
9	Maharshi	2019	Action,Drama	8.0	2733
10	Balkanskiy rubezh	2019	Action,War	7.8	2958

	original_title	start_year	genres	averagerating	numvotes
11	Rocketman	2019	Biography,Drama,Music	7.7	24266
12	Lucifer	2019	Action,Crime,Drama	7.7	4412
13	Kesari	2019	Action,Drama,History	7.7	7557
14	Dolor y gloria	2019	Drama	7.7	2802
15	Madhura Raja	2019	Action,Comedy,Drama	7.7	2522
16	How to Train Your Dragon: The Hidden World	2019	Action,Adventure,Animation	7.6	60769
17	Once Upon a Time in London	2019	Crime	7.6	2752
18	The Boy Who Harnessed the Wind	2019	Drama	7.6	10725
19	Alita: Battle Angel	2019	Action,Adventure,Sci-Fi	7.5	88207
20	Booksmart	2019	Comedy	7.5	17529
21	Shazam!	2019	Action,Adventure,Comedy	7.4	109051
22	The Professor and the Madman	2019	Biography,Drama,Mystery	7.4	10383
23	Aladdin	2019	Adventure,Comedy,Family	7.4	57549
24	Petta	2019	Action,Drama	7.4	6181

In [32]:

```
recent_merge = tn_movie_budgets_current.merge(latest_imdb, how='inner', left_on='movie',
right_on='original_title')
recent_merge.sort_values('release_date', ascending =False, inplace=True)
recent_merge.head(25)
```

Out[32]:

	release_date	movie	production_budget	domestic_gross	worldwide_gross	original_title	start_year	
0	2019-06-07	Dark Phoenix	$350,000,000	$42,762,350	$149,762,350	Dark Phoenix	2019	Action,Adv
1	2019-05-31	Godzilla: King of the Monsters	$170,000,000	$85,576,941	$299,276,941	Godzilla: King of the Monsters	2019	Action,Adven
2	2019-05-10	The Professor and the Madman	$25,000,000	$0	$5,227,233	The Professor and the Madman	2019	Biography,Dra
3	2019-05-03	Long Shot	$40,000,000	$30,202,860	$43,711,031	Long Shot	2019	Come
4	2019-04-12	Hellboy	$50,000,000	$21,903,748	$40,725,492	Hellboy	2019	Action,Adven
6	2019-04-05	Pet Sematary	$21,000,000	$54,724,696	$109,501,146	Pet Sematary	2019	Horror,My
8	2019-04-05	Shazam!	$85,000,000	$139,606,856	$362,899,733	Shazam!	2019	Action,Advent
9	2019-03-29	Unplanned	$6,000,000	$18,107,621	$18,107,621	Unplanned	2019	Biogr
10	2019-03-29	Dumbo	$170,000,000	$113,883,318	$345,004,422	Dumbo	2019	Adventure,Fa
11	2019-03-22	Us	$20,000,000	$175,006,930	$254,210,310	Us	2019	Horror,My
12	2019-03-15	Five Feet Apart	$7,000,000	$45,729,221	$80,504,421	Five Feet Apart	2019	Dran
13	2019-03-15	Captive State	$25,000,000	$5,958,315	$8,993,300	Captive State	2019	S
14	2019-03-15	Wonder Park	$100,000,000	$45,216,793	$115,149,422	Wonder Park	2019	Adventure,Anima
15	2019-03-08	Captain Marvel	$175,000,000	$426,525,952	$1,123,061,550	Captain Marvel	2019	Action,Adv
16	2019-02-22	How to Train Your Dragon: The Hidden	$129,000,000	$160,791,800	$519,258,283	How to Train Your Dragon: The Hidden World	2019	Action,Adventu

	release_date	movie	production_budget	domestic_gross	worldwide_gross	original_title	start_year	
17	2019-02-14	Alita: Battle Angel	$170,000,000	$85,710,210	$402,976,036	Alita: Battle Angel	2019	Action,Adv
18	2019-02-13	Happy Death Day 2U	$9,000,000	$28,051,045	$64,179,495	Happy Death Day 2U	2019	Drama,Ho
19	2019-02-08	Cold Pursuit	$60,000,000	$32,138,862	$62,599,159	Cold Pursuit	2019	Action,C
20	2019-02-08	What Men Want	$20,000,000	$54,611,903	$69,911,903	What Men Want	2019	Comedy,Fanta
21	2019-02-01	Velvet Buzzsaw	$21,000,000	$0	$0	Velvet Buzzsaw	2019	Horror,My
22	2019-01-25	Serenity	$25,000,000	$8,547,045	$11,367,029	Serenity	2019	Drama,M
24	2019-01-25	The Kid Who Would Be King	$59,000,000	$16,790,790	$28,348,446	The Kid Who Would Be King	2019	Action,Adve
25	2019-01-22	Renegades	$77,500,000	$0	$1,521,672	Renegades	2017	Action,Adve
26	2019-01-18	Glass	$20,000,000	$111,035,005	$245,303,505	Glass	2019	Drama,S
27	2019-01-11	The Upside	$37,500,000	$108,235,497	$119,024,536	The Upside	2017	Co

In [33]:

```
# most profitable movies( domestic gross - budget) of the last 3 years for 400 movies
recent_df = recent_merge.drop(index=recent_merge.index[400:], axis=0)
recent_df
```

Out[33]:

	release_date	movie	production_budget	domestic_gross	worldwide_gross	original_title	start_year	
0	2019-06-07	Dark Phoenix	$350,000,000	$42,762,350	$149,762,350	Dark Phoenix	2019	Action,Adv
1	2019-05-31	Godzilla: King of the Monsters	$170,000,000	$85,576,941	$299,276,941	Godzilla: King of the Monsters	2019	Action,Advent
2	2019-05-10	The Professor and the Madman	$25,000,000	$0	$5,227,233	The Professor and the Madman	2019	Biography,Dra
3	2019-05-03	Long Shot	$40,000,000	$30,202,860	$43,711,031	Long Shot	2019	Come
4	2019-04-12	Hellboy	$50,000,000	$21,903,748	$40,725,492	Hellboy	2019	Action,Advent
...	...	...	...	...	...	...	...	
406	2015-11-25	Victor Frankenstein	$40,000,000	$5,775,076	$31,124,367	Victor Frankenstein	2015	Drama,H
405	2015-11-25	Creed	$37,000,000	$109,767,581	$173,567,581	Creed	2015	D
411	2015-11-20	Legend	$25,000,000	$1,872,994	$42,425,450	Legend	2015	Biography,C
410	2015-11-20	Carol	$11,800,000	$12,711,491	$42,843,521	Carol	2015	Dram
409	2015-11-20	Mustang	$1,400,000	$845,464	$5,552,584	Mustang	2015	

400 rows × 10 columns

In [34]:

```
#removing $ and commas
recent_df['domestic_gross'] = recent_df['domestic_gross'].str.strip('$')
```

```
recent_df['production_budget'] = recent_df['production_budget'].str.strip('$')
recent_df['domestic_gross'] = recent_df['domestic_gross'].str.replace(',','')
recent_df['production_budget'] = recent_df['production_budget'].str.replace(',','')
recent_df
```

Out[34]:

	release_date	movie	production_budget	domestic_gross	worldwide_gross	original_title	start_year	
0	2019-06-07	Dark Phoenix	350000000	42762350	$149,762,350	Dark Phoenix	2019	Action,Adv
1	2019-05-31	Godzilla: King of the Monsters	170000000	85576941	$299,276,941	Godzilla: King of the Monsters	2019	Action,Advent
2	2019-05-10	The Professor and the Madman	25000000	0	$5,227,233	The Professor and the Madman	2019	Biography,Dra
3	2019-05-03	Long Shot	40000000	30202860	$43,711,031	Long Shot	2019	Come
4	2019-04-12	Hellboy	50000000	21903748	$40,725,492	Hellboy	2019	Action,Advent
...	...	...	...	...	...	...	...	
406	2015-11-25	Victor Frankenstein	40000000	5775076	$31,124,367	Victor Frankenstein	2015	Drama,H
405	2015-11-25	Creed	37000000	109767581	$173,567,581	Creed	2015	D
411	2015-11-20	Legend	25000000	1872994	$42,425,450	Legend	2015	Biography,C
410	2015-11-20	Carol	11800000	12711491	$42,843,521	Carol	2015	Dran
409	2015-11-20	Mustang	1400000	845464	$5,552,584	Mustang	2015	

400 rows × 10 columns

In [35]:

```
#converting to int
recent_df['domestic_gross'] = recent_df['domestic_gross'].astype(int)
recent_df['production_budget'] = recent_df['production_budget'].astype(int)
recent_df['profit'] = recent_df['domestic_gross'] - recent_df['production_budget']
recent_df.sort_values('profit', ascending=False, inplace=True)
recent_df.head(25)
```

Out[35]:

	release_date	movie	production_budget	domestic_gross	worldwide_gross	original_title	start_year	
117	2018-02-16	Black Panther	200000000	700059566	$1,348,258,224	Black Panther	2018	Action,
217	2017-03-17	Beauty and the Beast	160000000	504014165	$1,259,199,706	Beauty and the Beast	2017	Family
133	2017-12-20	Jumanji: Welcome to the Jungle	90000000	404508916	$964,496,193	Jumanji: Welcome to the Jungle	2017	Action,Ad
370	2016-02-12	Deadpool	58000000	363070709	$801,025,593	Deadpool	2016	Action,Ad
311	2016-07-08	The Secret Life of Pets	75000000	368384330	$886,750,534	The Secret Life of Pets	2016	Adventure,Ar
164	2017-09-08	It	35000000	327481748	$697,457,969	It	2017	
321	2016-06-17	Finding Dory	200000000	486295561	$1,021,215,193	Finding Dory	2016	Adventure,Ar
188	2017-06-02	Wonder Woman	150000000	412563408	$821,133,378	Wonder Woman	2017	Action,Ad
15	2019-03-08	Captain Marvel	175000000	426525952	$1,123,061,550	Captain Marvel	2019	Action,
90	2018-05-18	Deadpool 2	110000000	324591735	$786,680,557	Deadpool 2	2018	Action,Ad

	release_date	movie	production_budget	domestic_gross	worldwide_gross	original_title	start_year	
249	2016-12-21	Sing	75000000	270329045	$634,454,789	Sing	2016	Animation
362	2016-03-04	Zootopia	150000000	341268248	$1,019,429,616	Zootopia	2016	Adventure,Ar
345	2016-04-15	The Jungle Book	175000000	364001123	$962,854,547	The Jungle Book	2016	Adventu
223	2017-02-24	Get Out	5000000	176040665	$255,367,951	Get Out	2017	Horro
99	2018-04-06	A Quiet Place	17000000	188024361	$334,522,294	A Quiet Place	2018	Dra
48	2018-11-02	Bohemian Rhapsody	55000000	216303339	$894,985,342	Bohemian Rhapsody	2018	Biograp
177	2017-07-07	Spider-Man: Homecoming	175000000	334201140	$880,166,350	Spider-Man: Homecoming	2017	Action,
336	2016-05-06	Captain America: Civil War	250000000	408084349	$1,140,069,413	Captain America: Civil War	2016	Action,
11	2019-03-22	Us	20000000	175006930	$254,210,310	Us	2019	Horro
302	2016-08-05	Suicide Squad	175000000	325100054	$746,059,887	Suicide Squad	2016	Action,Ac
67	2018-08-15	Crazy Rich Asians	30000000	174532921	$238,099,711	Crazy Rich Asians	2018	C
143	2017-11-03	Thor: Ragnarok	180000000	315058289	$846,980,024	Thor: Ragnarok	2017	Action,Ac
235	2017-01-20	Split	5000000	138141585	$278,964,806	Split	2016	
251	2016-12-09	La La Land	20000000	151101803	$426,351,163	La La Land	2016	Come
74	2018-07-13	Hotel Transylvania 3: Summer Vacation	65000000	167500092	$527,079,962	Hotel Transylvania 3: Summer Vacation	2018	Adventure,Ar

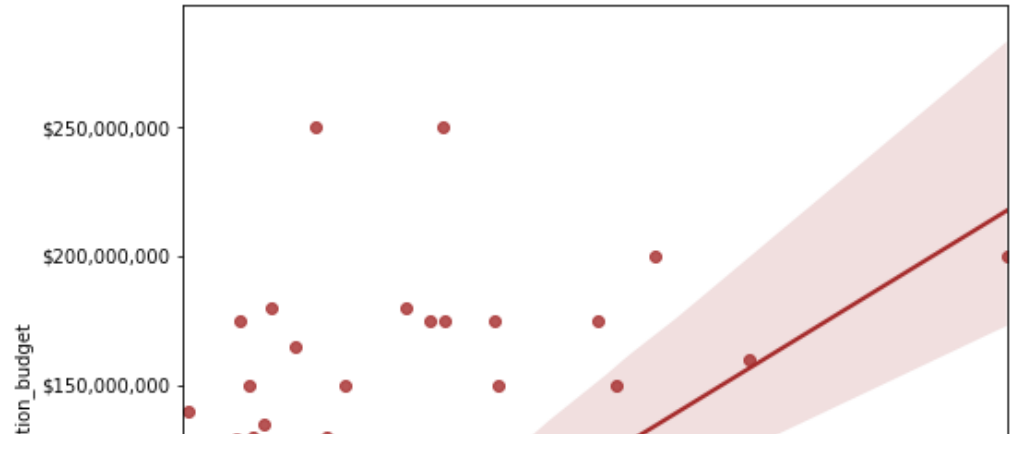
In [36]:

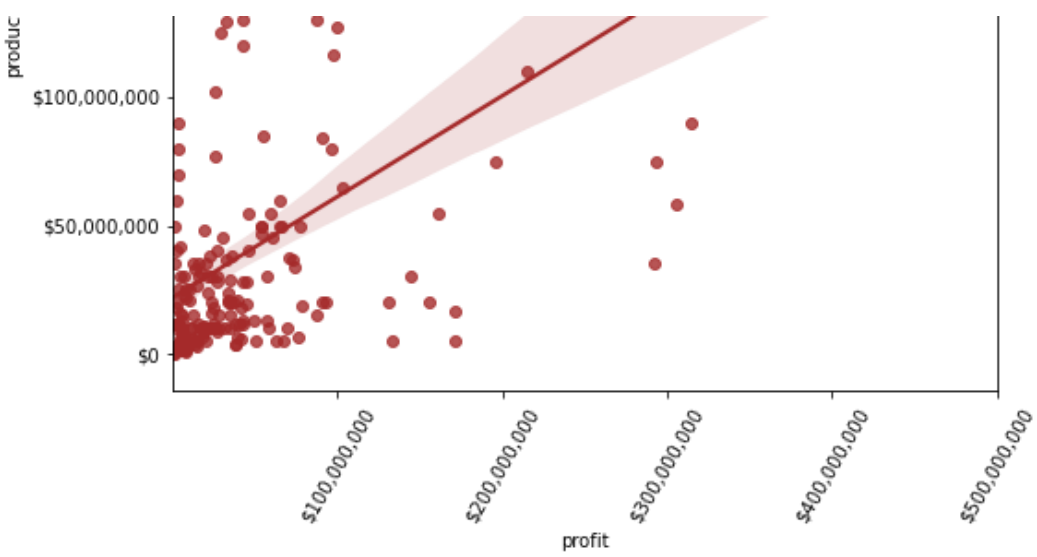
```
#from the above you can notice that even the movies that make the higest profit don't hav
e the higest rating
#remove the negatives for movies that made no profit
recent_profit = recent_df[recent_df['profit']>0]
```

In [37]:

```
# plot a graph to show the relationship
fig, ax = plt.subplots(figsize=(8, 8))
sea.regplot(data=recent_profit, x='profit', y='production_budget',color='brown', ax=ax)

fmt = '${x:,.0f}'
tick = mtick.StrMethodFormatter(fmt)
ax.yaxis.set_major_formatter(tick)
ax.xaxis.set_major_formatter(tick)
plt.xticks(rotation=60)
plt.show()
```





so we lack a relationship between the budget and profit

In [38]:

```
recent_profit['production_budget'].mean()
```

Out[38]:

```
43194427.083333336
```

Recommended budget turns out to be a target of $43,000,000

we shall now focus on crowd pulling and determine higest rating using person_id and get it

In [39]:

```
person_df = pd.read_sql("""
SELECT p.primary_name,
       mr.averagerating,
       COUNT(DISTINCT mb.primary_title) num_movies
FROM directors d
JOIN persons p
    USING(person_id)
JOIN principals
    USING(movie_id)
JOIN movie_basics mb
    USING(movie_id)
JOIN movie_ratings mr
    USING(movie_id)
WHERE numvotes > 3000
GROUP BY p.primary_name
HAVING num_movies > 5
""", conn)
person_df
```

Out[39]:

	primary_name	averagerating	num_movies
0	A.R. Murugadoss	6.8	7
1	Adam Wingard	5.3	8
2	Alex Gibney	7.3	8
3	Anurag Kashyap	8.1	9
4	Baltasar Kormákur	6.6	6
5	Ben Wheatley	5.6	6
6	Clint Eastwood	6.5	7
7	Darren Lynn Bousman	6.4	7

	primary_name	averagerating	num_movies
8	David Gordon Green	6.4	8
9	Denis Villeneuve	8.3	6
10	Frank D'Angelo	6.6	6
11	François Ozon	6.4	6
12	Hirokazu Koreeda	7.4	6
13	James Wan	7.2	6
14	Jon M. Chu	6.2	7
15	Kevin Macdonald	7.2	6
16	Lasse Hallström	6.8	7
17	Luc Besson	6.5	6
18	Michael Winterbottom	6.6	6
19	Mike Flanagan	5.8	6
20	Noah Baumbach	6.1	6
21	Ridley Scott	8.0	7
22	Rob Reiner	6.5	6
23	Rohit Shetty	5.5	8
24	Ron Howard	5.3	6
25	Sarik Andreasyan	4.5	6
26	Stephen Frears	6.8	6
27	Steven C. Miller	4.9	7
28	Steven Soderbergh	7.0	8
29	Steven Spielberg	7.4	7
30	Takashi Miike	7.4	6
31	Tim Burton	6.5	6
32	Tim Story	7.4	7
33	Tyler Perry	4.7	12
34	Werner Herzog	7.2	6
35	Woody Allen	6.3	8
36	Álex de la Iglesia	6.6	6

In [40]:

```
person_df.drop_duplicates(inplace=True)
person_df.sort_values('averagerating', ascending=False, inplace=True)
person_df.head(25)
```

Out[40]:

	primary_name	averagerating	num_movies
9	Denis Villeneuve	8.3	6
3	Anurag Kashyap	8.1	9
21	Ridley Scott	8.0	7
32	Tim Story	7.4	7
30	Takashi Miike	7.4	6
29	Steven Spielberg	7.4	7
12	Hirokazu Koreeda	7.4	6
2	Alex Gibney	7.3	8
15	Kevin Macdonald	7.2	6

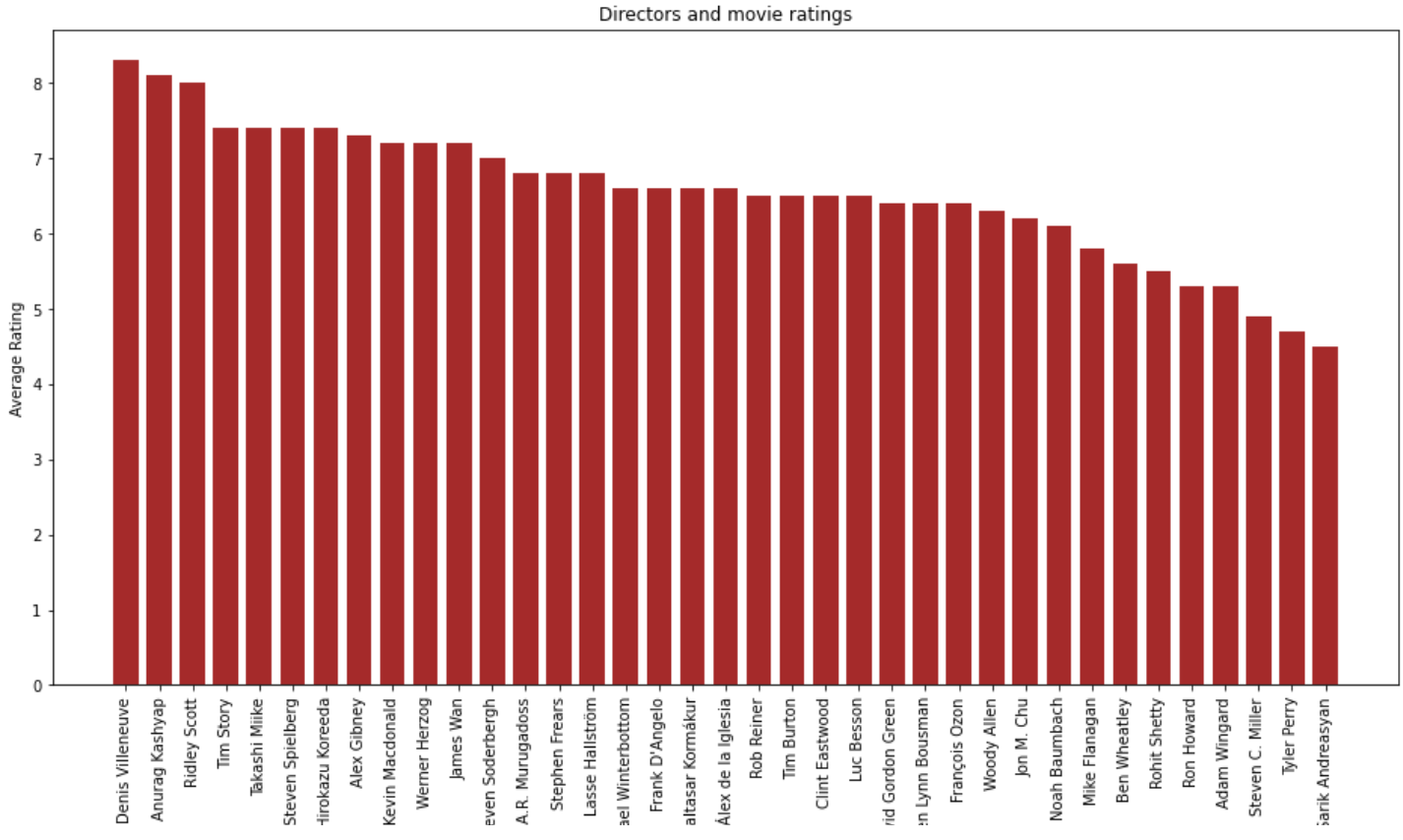
	primary_name	averagerating	num_movies
34	[illegible]	7.3	8
13	James Wan	7.2	6
28	Steven Soderbergh	7.0	8
0	A.R. Murugadoss	6.8	7
26	Stephen Frears	6.8	6
16	Lasse Hallström	6.8	7
18	Michael Winterbottom	6.6	6
10	Frank D'Angelo	6.6	6
4	Baltasar Kormákur	6.6	6
36	Álex de la Iglesia	6.6	6
22	Rob Reiner	6.5	6
31	Tim Burton	6.5	6
6	Clint Eastwood	6.5	7
17	Luc Besson	6.5	6
8	David Gordon Green	6.4	8
7	Darren Lynn Bousman	6.4	7

To filter out(outliers) the directors we've decided to only include directors who have more than 5 movies and each has to have more than 3000 votes

we shall use this list to select a director for any film

In [106]:

```
x = person_df['primary_name']
y = person_df['averagerating']
plt.figure(figsize=(16, 8))
plt.bar(x, y, width=.75, color='brown')
plt.title('Directors and movie ratings')
plt.xlabel('Directors')
plt.ylabel('Average Rating')
plt.xticks(rotation = 90);
```



From the graph above,we were able to obtain the Directors we'd highly recommend:

- **Denis Villeneuve 8.3**
- **Anurag Kashyap 8.1**
- **Ridley Scott 8.0**

Conclusion

I used ratings and profit to come up with our mesurement of success and we determined that Microsoft should:

- **Make a Drama,Action,Comedy film**
- **Use a budget above $43,000,000**
- **To obtain the best result,we recommend use of either the top 3 directors as stated above.**