Investigate_a_Dataset-zh

June 3, 2018

1 IMDB

1.1

##
genresrelease_date

In [1]:
```
#
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
```

##

1.1.1

In [2]:
```
#
df=pd.read_csv('tmdb-movies.csv')
df.head()
```

Out[2]:

	id	imdb_id	popularity	budget	revenue	original_title
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7

```
                                                cast  \
0  Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
1  Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...
2  Shailene Woodley|Theo James|Kate Winslet|Ansel...
3  Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...
4  Vin Diesel|Paul Walker|Jason Statham|Michelle ...

                                        homepage          director  \
0                    http://www.jurassicworld.com/   Colin Trevorrow
1                      http://www.madmaxmovie.com/     George Miller
2    http://www.thedivergentseries.movie/#insurgent  Robert Schwentke
3  http://www.starwars.com/films/star-wars-episod...       J.J. Abrams
4                          http://www.furious7.com/         James Wan

                         tagline      ...       \
0             The park is open.      ...
1            What a Lovely Day.      ...
2    One Choice Can Destroy You      ...
3  Every generation has a story.     ...
4           Vengeance Hits Home      ...

                                        overview runtime  \
0  Twenty-two years after the events of Jurassic ...     124
1  An apocalyptic story set in the furthest reach...     120
2  Beatrice Prior must confront her inner demons ...     119
3  Thirty years after defeating the Galactic Empi...     136
4  Deckard Shaw seeks revenge against Dominic Tor...     137

                                     genres  \
0  Action|Adventure|Science Fiction|Thriller
1  Action|Adventure|Science Fiction|Thriller
2         Adventure|Science Fiction|Thriller
3   Action|Adventure|Science Fiction|Fantasy
4                     Action|Crime|Thriller

                                production_companies release_date vote_count  \
0  Universal Studios|Amblin Entertainment|Legenda...       6/9/15       5562
1  Village Roadshow Pictures|Kennedy Miller Produ...      5/13/15       6185
2  Summit Entertainment|Mandeville Films|Red Wago...      3/18/15       2480
3          Lucasfilm|Truenorth Productions|Bad Robot     12/15/15       5292
4  Universal Pictures|Original Film|Media Rights ...       4/1/15       2947

   vote_average  release_year    budget_adj   revenue_adj
0           6.5          2015  1.379999e+08  1.392446e+09
1           7.1          2015  1.379999e+08  3.481613e+08
2           6.3          2015  1.012000e+08  2.716190e+08
3           7.5          2015  1.839999e+08  1.902723e+09
```

```
        4           7.3          2015  1.747999e+08  1.385749e+09

        [5 rows x 21 columns]
```

In [3]:
```
#
df.shape
```

Out[3]:
```
(10866, 21)
```

In [4]:
```
#
df.dtypes
```

Out[4]:
```
id                      int64
imdb_id                object
popularity            float64
budget                  int64
revenue                 int64
original_title         object
cast                   object
homepage               object
director               object
tagline                object
keywords               object
overview               object
runtime                 int64
genres                 object
production_companies   object
release_date           object
vote_count              int64
vote_average          float64
release_year            int64
budget_adj            float64
revenue_adj           float64
dtype: object
```

In [5]:
```
#
df.isnull().sum()
```

Out[5]:
```
id                      0
imdb_id                10
popularity              0
budget                  0
revenue                 0
original_title          0
cast                   76
homepage             7930
director               44
tagline              2824
keywords             1493
```

```
overview                   4
runtime                    0
genres                    23
production_companies    1030
release_date               0
vote_count                 0
vote_average               0
release_year               0
budget_adj                 0
revenue_adj                0
dtype: int64
```

In [6]:
```
# genres230.2%
# genres
df[df['genres'].isnull()]
```

Out[6]:

	id	imdb_id	popularity	budget	revenue	original_title
424	363869	tt4835298	0.244648	0	0	Belli di papÃ ă
620	361043	tt5022680	0.129696	0	0	All Hallows' Eve 2
997	287663	NaN	0.330431	0	0	Star Wars Rebels: Spark of Rebellion
1712	21634	tt1073510	0.302095	0	0	Prayers for Bobby
1897	40534	tt1229827	0.020701	0	0	Jonas Brothers: The Concert Experience
2370	127717	tt1525359	0.081892	0	0	Freshman Father
2376	315620	tt1672218	0.068411	0	0	
2853	57892	tt0270053	0.130018	0	0	
3279	54330	tt1720044	0.145331	0	0	
4547	123024	tt2305700	0.520520	0	0	
4732	139463	tt2084977	0.235911	0	0	
4797	369145	NaN	0.167501	0	0	
4890	126909	tt2219564	0.083202	0	0	
5830	282848	tt2986512	0.248944	0	0	
5934	200204	tt2808968	0.067433	0	0	
6043	190940	tt2797242	0.039080	0	0	
6530	168891	tt0818519	0.092724	0	0	
8234	56804	tt0114844	0.028874	0	0	
8614	65595	tt0117880	0.273934	0	0	
8878	92208	tt0250593	0.038045	0	0	
9307	141859	tt0097446	0.094652	0	0	
9799	48847	tt0193716	0.175008	0	0	
10659	4255	tt0065904	0.344172	5000	0	

```
2376                      Doctor Who: A Christmas Carol
2853                                           Vizontele
3279                                        ìêÿřì ë
4547   London 2012 Olympic Opening Ceremony: Isles of...
4732                                        The Scapegoat
4797                               Doctor Who: The Snowmen
4890                           Cousin Ben Troop Screening
5830                  Doctor Who: The Time of the Doctor
5934                                         Prada: Candy
6043                                       Bombay Talkies
6530                                          Saw Rebirth
8234                                      Viaggi di nozze
8614                           T2 3-D: Battle Across Time
8878                     Mom's Got a Date With a Vampire
9307                                           Goldeneye
9799                                        The Amputee
10659                       The Party at Kitty and Stud's

                                                     cast  \
424    Diego Abatantuono|Matilde Gioli|Andrea Pisani|...
620                                                  NaN
997    Freddie Prinze Jr.|Vanessa Marshall|Steve Blum...
1712   Ryan Kelley|Sigourney Weaver|Henry Czerny|Dan ...
1897   Nick Jonas|Joe Jonas|Kevin Jonas|John Lloyd Ta...
2370   Britt Irvin|Merrilyn Gann|Barbara Tyson|Anthon...
2376   Matt Smith|Karen Gillan|Arthur Darvill|Michael...
2853   YÄślmaz ErdoÄan|Demet Akbag|Altan Erkekli|Cem...
3279             Jang Keun-suk|Song Ha-yoon|Kim Jeong-Nan
4547   Queen Elizabeth II|Mike Oldfield|Kenneth Brana...
4732   Andrew Scott|Jodhi May|Eileen Atkins|Matthew R...
4797   Matt Smith|Jenna Coleman|Richard E. Grant|Ian ...
4890                                  Jason Schwartzman
5830                              Matt Smith|Jenna Coleman
5934             Peter Gadiot|Rodolphe Pauly|LÃĽa Seydoux
6043   Aamir Khan|Rani Mukerji|Randeep Hooda|Saqib Sa...
6530   Whit Anderson|Stan Kirsch|Jeff Shuter|George W...
8234   Carlo Verdone|Claudia Gerini|Veronica Pivetti|...
8614   Arnold Schwarzenegger|Linda Hamilton|Edward Fu...
8878   Matt O'Leary|Laura Vandervoort|Myles Jeffrey|C...
9307   Charles Dance|Phyllis Logan|Patrick Ryecart|La...
9799                     Catherine E. Coulson|David Lynch
10659  Sylvester Stallone|Henrietta Holm|Nicholas War...

                                                 homepage  \
424                                                   NaN
620                                                   NaN
997                                                   NaN
1712                     http://www.prayersforbobby.com/
```

```
1897                                                    NaN
2370                                                    NaN
2376                                                    NaN
2853                                                    NaN
3279                                                    NaN
4547                            http://www.london2012.com/
4732   http://www.island-pictures.co.uk/extras/the-sc...
4797                                                    NaN
4890   http://www.funnyordie.com/videos/fc132ce8b2/co...
5830                                                    NaN
5934                                                    NaN
6043   http://en.wikipedia.org/wiki/Bombay_Talkies_%2...
6530                                                    NaN
8234                                                    NaN
8614                                                    NaN
8878                                                    NaN
9307                                                    NaN
9799                                                    NaN
10659                                                   NaN

                                                director  \
424                                          Guido Chiesa
620    Antonio Padovan|Bryan Norton|Marc Roussel|Ryan...
997                            Steward Lee|Steven G. Lee
1712                                     Russell Mulcahy
1897                                     Bruce Hendricks
2370                                       Michael Scott
2376                                                 NaN
2853                                    YÄślmaz ErdoÄan
3279                                        Kim Jin-Yeong
4547                                          Danny Boyle
4732                                     Charles Sturridge
4797                                                 NaN
4890                                         Wes Anderson
5830                                          James Payne
5934                           Wes Anderson|Roman Coppola
6043   Anurag Kashyap|Dibakar Banerjee|Zoya Akhtar|Ka...
6530                             Jeff Shuter|Daniel Viney
8234                                     Carlo Verdone
8614                                     James Cameron
8878                                       Steve Boyum
9307                                          Don Boyd
9799                                        David Lynch
10659                                      Morton Lewis

                                                 tagline     ...       \
424                                                  NaN     ...
620                                                  NaN     ...
```

```
997                                                    NaN     ...
1712   Before you echo "amen" in your home and place ...      ...
1897                                                   NaN     ...
2370                                                   NaN     ...
2376                                                   NaN     ...
2853                                                   NaN     ...
3279                                                   NaN     ...
4547                                Inspire a generation.      ...
4732                                                   NaN     ...
4797                                                   NaN     ...
4890                                                   NaN     ...
5830                         A change is going to come...      ...
5934                                                 short     ...
6043                                                   NaN     ...
6530     Somewhere... Somehow... Something went wrong...      ...
8234                                                   NaN     ...
8614                                                   NaN     ...
8878                                                   NaN     ...
9307                                                   NaN     ...
9799                                                   NaN     ...
10659                                                  NaN     ...

                                                  overview runtime  genres  \
424    Italian remake of the Mexican 2013 hit, "We th...      100     NaN
620    A woman finds a VHS tape on her doorstep that ...       90     NaN
997    A Long Time Ago In A Galaxy Far, Far Awayâę A...       44     NaN
1712   True story of Mary Griffith, gay rights crusad...       88     NaN
1897   Secure your VIP pass to a once-in-a-lifetime e...       76     NaN
2370                                                  NaN        0     NaN
2376   Amy Pond and Rory Williams are trapped on a cr...       62     NaN
2853   The story takes place in a small town (called ...      110     NaN
3279   Joon-soo (Jang Geun -Seok) is a rebellious hig...       96     NaN
4547   The London 2012 Olympic Games Opening Ceremony...      220     NaN
4732   Set in 1952, as England prepares for the coron...      100     NaN
4797   Christmas Eve, 1892, and the falling snow is t...       60     NaN
4890   Cousin Ben hosts a screening of Wes Anderson's...        2     NaN
5830   Orbiting a quiet backwater planet, the massed ...       60     NaN
5934   Candy is a modern chic french woman. She meets...        3     NaN
6043   One hundred years of Hindi cinema is celebrate...      127     NaN
6530   This comic, set in the world of SAW goes back ...        6     NaN
8234   Le vicessitudini di tre coppie di novelli spos...      103     NaN
8614   Three freedom fighters attack a large corporat...       12     NaN
8878   The Hansen kids are in a jam. Adam and his bes...       85     NaN
9307   Fact-based biography of James Bond author, Ian...      105     NaN
9799   A double leg amputated woman sits and writes a...        5     NaN
10659  Kitty and Stud are lovers. They enjoy a robust...       71     NaN

                      production_companies release_date vote_count  \
```

424		NaN	10/29/15	21
620	Ruthless Pictures|Hollywood Shorts		10/6/15	13
997		NaN	10/3/14	13
1712	Daniel Sladek Entertainment		2/27/09	57
1897		NaN	2/27/09	11
2370		NaN	6/5/10	12
2376		NaN	12/25/10	11
2853		NaN	2/2/01	12
3279		NaN	8/13/08	11
4547		BBC	7/27/12	12
4732	Island Pictures		9/9/12	12
4797	BBC Television UK		12/25/12	10
4890		NaN	1/1/12	14
5830		NaN	12/25/13	26
5934		NaN	3/25/13	27
6043	Viacom 18 Motion Pictures		5/3/13	12
6530		NaN	10/24/05	24
8234		NaN	12/15/95	44
8614		NaN	1/1/96	14
8878	Walt Disney Pictures		10/13/00	16
9307	Anglia Television		8/26/89	10
9799		NaN	1/1/74	11
10659	Stallion Releasing Inc.		2/10/70	10

	vote_average	release_year	budget_adj	revenue_adj
424	6.1	2015	0.00000	0.0
620	5.0	2015	0.00000	0.0
997	6.8	2014	0.00000	0.0
1712	7.4	2009	0.00000	0.0
1897	7.0	2009	0.00000	0.0
2370	5.8	2010	0.00000	0.0
2376	7.7	2010	0.00000	0.0
2853	7.2	2001	0.00000	0.0
3279	6.1	2008	0.00000	0.0
4547	8.3	2012	0.00000	0.0
4732	6.2	2012	0.00000	0.0
4797	7.8	2012	0.00000	0.0
4890	7.0	2012	0.00000	0.0
5830	8.5	2013	0.00000	0.0
5934	6.9	2013	0.00000	0.0
6043	5.9	2013	0.00000	0.0
6530	5.9	2005	0.00000	0.0
8234	6.7	1995	0.00000	0.0
8614	6.7	1996	0.00000	0.0
8878	5.4	2000	0.00000	0.0
9307	5.3	1989	0.00000	0.0
9799	5.0	1974	0.00000	0.0
10659	3.0	1970	28081.84172	0.0

```
[23 rows x 21 columns]
```

In [7]:
```
#
df[df.duplicated(keep=False)]
```

Out[7]:
```
          id    imdb_id  popularity    budget  revenue original_title  \
2089  42194  tt0411951     0.59643  30000000   967000         TEKKEN
2090  42194  tt0411951     0.59643  30000000   967000         TEKKEN

                                                 cast homepage  \
2089  Jon Foo|Kelly Overton|Cary-Hiroyuki Tagawa|Ian...      NaN
2090  Jon Foo|Kelly Overton|Cary-Hiroyuki Tagawa|Ian...      NaN

              director              tagline     ...       \
2089  Dwight H. Little  Survival is no game     ...
2090  Dwight H. Little  Survival is no game     ...

                                              overview runtime  \
2089  In the year of 2039, after World Wars destroy ...      92
2090  In the year of 2039, after World Wars destroy ...      92

                                          genres    production_companies  \
2089  Crime|Drama|Action|Thriller|Science Fiction  Namco|Light Song Films
2090  Crime|Drama|Action|Thriller|Science Fiction  Namco|Light Song Films

     release_date vote_count  vote_average  release_year  budget_adj  \
2089      3/20/10        110           5.0          2010  30000000.0
2090      3/20/10        110           5.0          2010  30000000.0

      revenue_adj
2089     967000.0
2090     967000.0

[2 rows x 21 columns]
```

In [8]:
```
#
df.drop_duplicates(inplace=True)
df.duplicated().sum()
```

Out[8]:
```
0
```

1.1.2 genres

In [9]:
```
# genres'
df.dropna(subset = ['genres'],inplace=True)
```

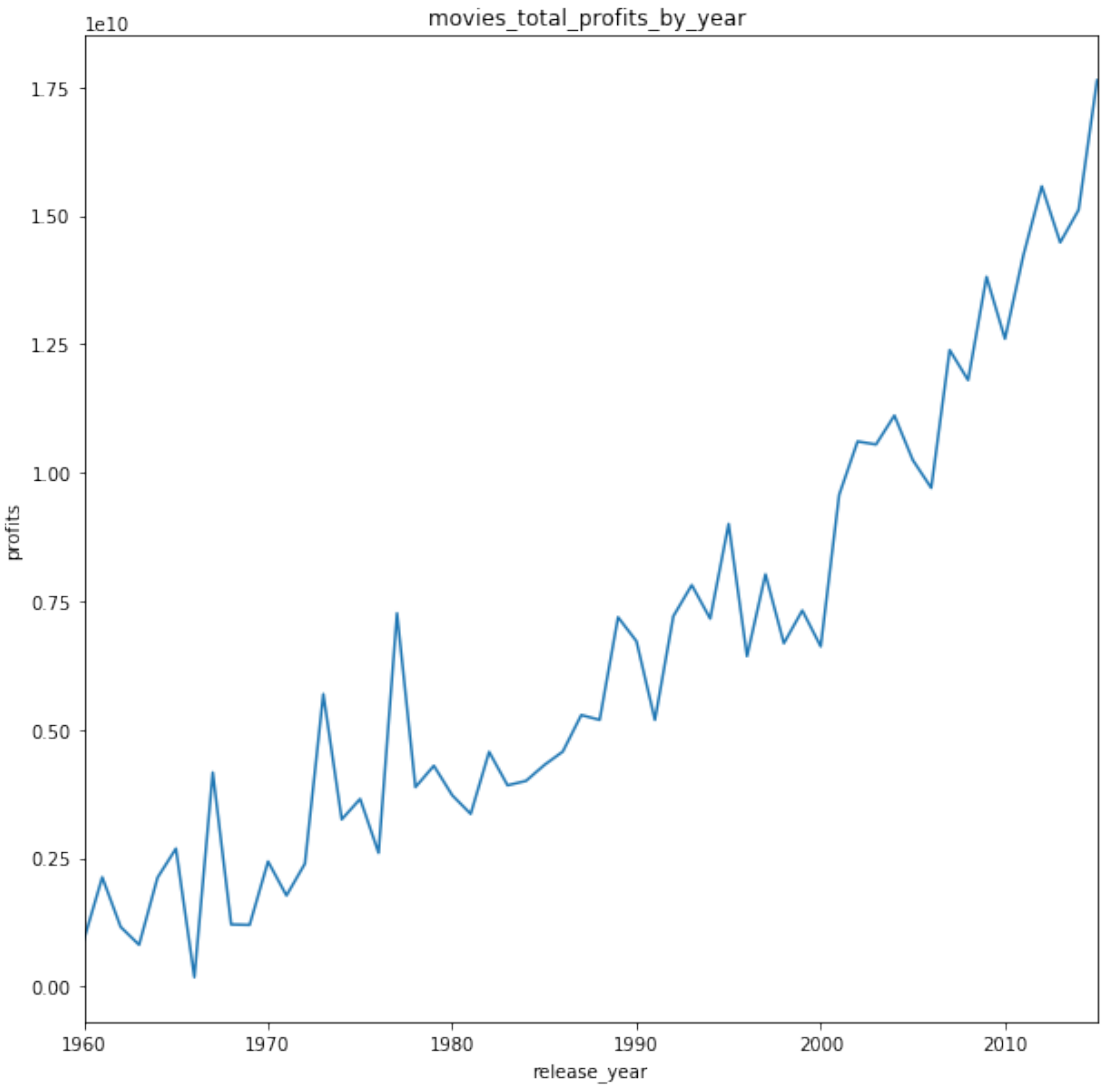
In [10]:
```
#genres
 df_genres = df.drop('genres', axis=1).join(df['genres'].str.split('|', expand=True)\
                                                    .stack().reset_index(level=1, drop=True).ren
```

```
In [11]: df_genres.shape

Out[11]: (26955, 21)
```

##

1.1.3 1

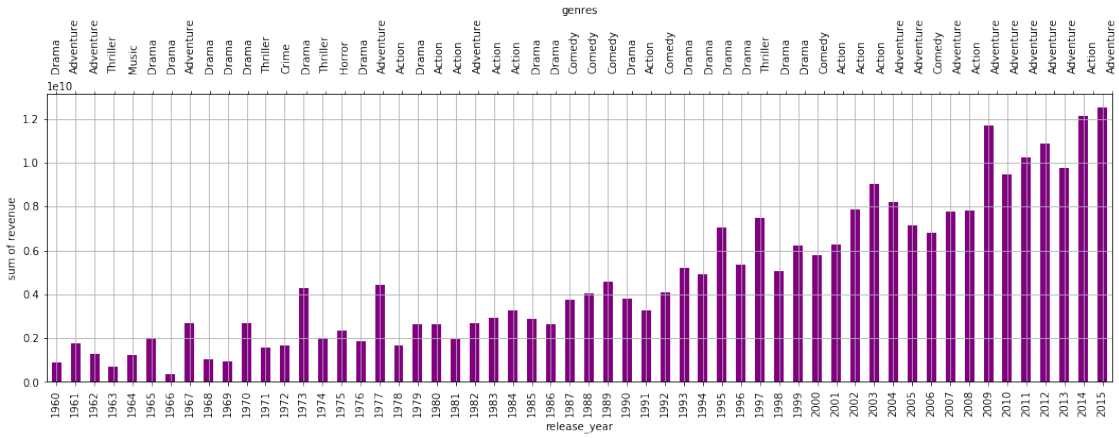
```
In [12]: # groupby
         df['profit']=df['revenue_adj']-df['budget_adj']
         df_profit=df.groupby(df['release_year'])['profit'].sum()


         df_profit.plot(kind='line',figsize=(10,10));
         plt.title('movies_total_profits_by_year ');
         plt.xlabel('release_year');
         plt.ylabel('profits');
```



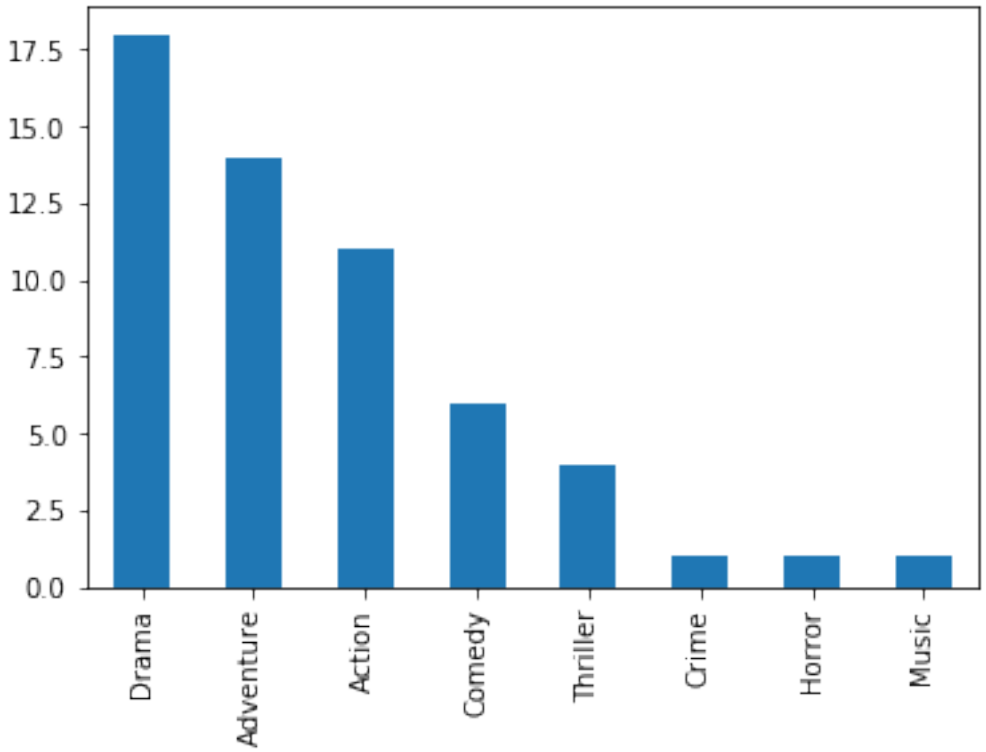
1.1.4 21

1.1.5 2

In [13]:
```
# groupbyrelease_year,genresrevenueunstackmaxidmaxrevenue
figure, ax1 = plt.subplots()

target = df_genres.groupby(['release_year', 'genres'])['revenue_adj'].sum().unstack(lev
target.max().plot(kind='bar', figsize=(18, 5), color='purple', grid=True, axes=ax1)

ax2 = ax1.twiny()
ax2.set_xticklabels(target.idxmax(), rotation=90)
ax2.set_xticks(ax1.get_xticks()+0.5)
ax2.tick_params(pad=15)

ax1.set_ylabel('sum of revenue')
ax2.set_xlabel('genres')
plt.show()
```



In [14]:
```
#
target.idxmax().value_counts().plot(kind='bar');
```



1.1.6 Drama

21
Drama

```
In [15]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

```
Out[15]: 255
```