TMDB Analysis

January 18, 2018

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns

        %matplotlib inline

        data = 'data/tmdb-movies.csv'

        imported_data = pd.read_csv(data)
```

```
In [2]: imported_data.shape
```

```
Out[2]: (10866, 21)
```

1 Introduction

1.1 Description of Dataset

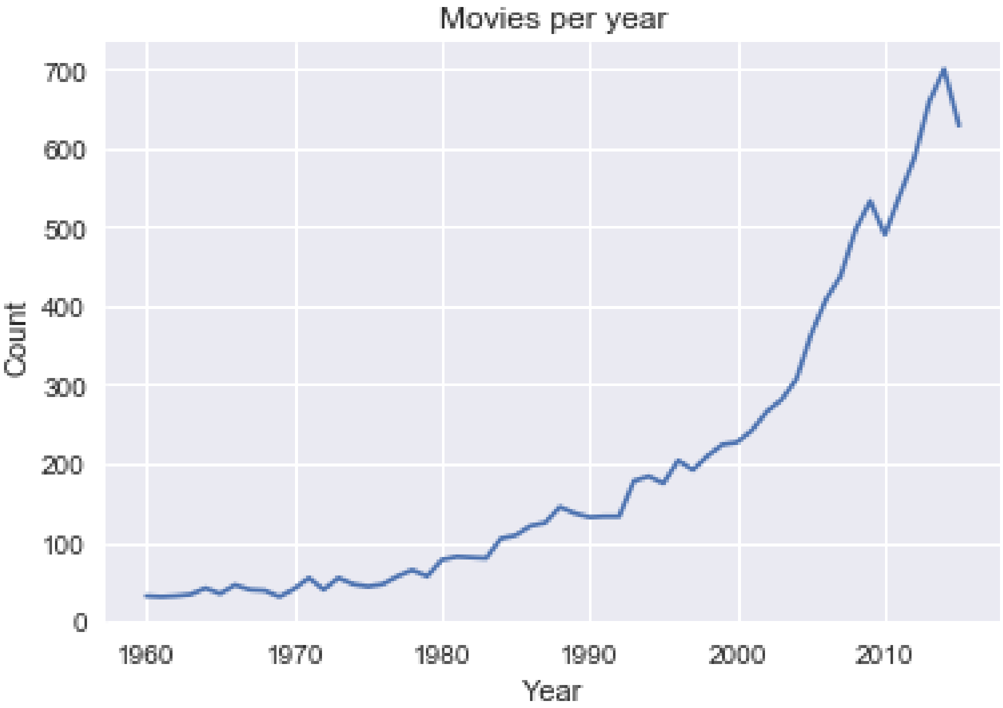
I selected The Movie Database dataset from Kaggle for my project. The dataset was created after IMDB requested a takedown of their data from Kaggle. As a consequence, an open-source alternative was selected with the caveat that there were some open questions about the data: * The currency of the budget amounts is not known. * The revenue might not consistently show global revenues. * The dataset, as a whole, has not undergone quality auditing. * There are 0s for numerous budget & revenue records. It is recommended that 0s are treated as missing values for these records. * Though not mentioned, because The Movie Database is a publicly curated dataset, it is subject to similar hesitations that other public, user-written databases are. The database is not necessarily peer-reviewed or sourced for accuracy.

With these caveats in mind, I decided to review the dataset to look at the general character of the data, as well as perform some experimental cleaning functions. The first question of interest to me is the scope of the movies in the dataset; what is the timespan of movies included, and how many are there?

```
In [3]: imported_data['release_year'].agg(['min', 'max'])
```

```
Out[3]: min    1960
        max    2015
        Name: release_year, dtype: int64
```

```
In [4]: counts_by_year = imported_data.groupby('release_year')['id'].count()

        plt.plot(counts_by_year)
        plt.title('Movies per year')
        plt.xlabel('Year')
        plt.ylabel('Count')

Out[4]: <matplotlib.text.Text at 0x2242e5ed080>
```



We can see that the first movie included in the database is from 1960, and the latest is 2015; movies included seems to grow year-over-year on a seemingly logarithmic scale. One might suppose this is because the more recent a movie is, the more likely a user is contribute meaningfully to information about it. Similarly, it could be the case that more movies are actually produced as time goes on. Meaningfully answering these questions would require juxtaposing the dataset against other data – performing the quality assessment precluded in the Kaggle description – as well as more rigorous statistical testing than a line chart.

Interestingly, there are a few noticeable dips in movies per year. An immediate one of note is in 2010, possibly a consequence of a notable writer’s strike. The strike occurred from late 2007 and into early 2008, and might account for the dip in movies.

```
In [5]: imported_data.query('release_year >= 2006 and release_year <= 2012').groupby('release_ye

Out[5]: release_year
        2006     408
        2007     438
```

```
2008    496
2009    533
2010    490
2011    540
2012    588
Name: id, dtype: int64
```

More information would be helpful to contextualize this fact of the data. Notably, it might be helpful to know the standard production time of a movie – to know when a movie being written, then optioned, is released theatrically! I'm sure there are many confounding variables beyond the scope of this dataset and investigation, but nonetheless it is an interesting point to speculate on.

1.2 Wrangling and Cleaning

One of my goals for this dataset is understanding what exactly is included, and experimenting with teasing out interesting questions that it might contribute to answering. To that end, I'll be looking at some of the standard techniques for assessing data: * Looking at the first and last 5 rows * getting a sense of how many values are missing * dropping some attributes that might not help with an analysis, either because they are poorly formatted, missing data, or require advanced techniques such as sentiment analysis.

```
In [6]: imported_data.head()
```

```
Out[6]:        id    imdb_id  popularity     budget     revenue  \
        0  135397  tt0369610   32.985763  150000000  1513528810
        1   76341  tt1392190   28.419936  150000000   378436354
        2  262500  tt2908446   13.112507  110000000   295238201
        3  140607  tt2488496   11.173104  200000000  2068178225
        4  168259  tt2820852    9.335014  190000000  1506249360

                        original_title  \
        0               Jurassic World
        1           Mad Max: Fury Road
        2                    Insurgent
        3  Star Wars: The Force Awakens
        4                    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 [7]: imported_data.tail()
```

```
Out[7]:           id    imdb_id  popularity  budget  revenue  \
        10861     21  tt0060371    0.080598       0        0
        10862  20379  tt0060472    0.065543       0        0
        10863  39768  tt0060161    0.065141       0        0
        10864  21449  tt0061177    0.064317       0        0
        10865  22293  tt0060666    0.035919   19000        0
```

```
                 original_title  \
10861        The Endless Summer   
10862                Grand Prix   
10863       Beregis Avtomobilya   
10864    What's Up, Tiger Lily?   
10865  Manos: The Hands of Fate   

                                                    cast homepage  \
10861  Michael Hynson|Robert August|Lord 'Tally Ho' B...      NaN   
10862  James Garner|Eva Marie Saint|Yves Montand|Tosh...      NaN   
10863  Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z...      NaN   
10864  Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh...      NaN   
10865  Harold P. Warren|Tom Neyman|John Reynolds|Dian...      NaN   

                 director                                            tagline  \
10861         Bruce Brown                                                NaN   
10862  John Frankenheimer  Cinerama sweeps YOU into a drama of speed and ...   
10863      Eldar Ryazanov                                                NaN   
10864         Woody Allen                          WOODY ALLEN STRIKES BACK!   
10865    Harold P. Warren       It's Shocking! It's Beyond Your Imagination!   

          ...                                                overview runtime  \
10861     ...       The Endless Summer, by Bruce Brown, is one of ...      95   
10862     ...       Grand Prix driver Pete Aron is fired by his te...     176   
10863     ...       An insurance agent who moonlights as a carthie...      94   
10864     ...       In comic Woody Allen's film debut, he took the...      80   
10865     ...       A family gets lost on the road and stumbles up...      74   

                       genres  \
10861             Documentary   
10862  Action|Adventure|Drama   
10863          Mystery|Comedy   
10864           Action|Comedy   
10865                  Horror   

                                    production_companies release_date  \
10861                                  Bruce Brown Films      6/15/66   
10862  Cherokee Productions|Joel Productions|Douglas ...     12/21/66   
10863                                            Mosfilm       1/1/66   
10864                            Benedict Pictures Corp.      11/2/66   
10865                                          Norm-Iris     11/15/66   

      vote_count  vote_average  release_year     budget_adj  revenue_adj
10861         11           7.4          1966       0.000000          0.0
10862         20           5.7          1966       0.000000          0.0
10863         11           6.5          1966       0.000000          0.0
10864         22           5.4          1966       0.000000          0.0
10865         15           1.5          1966  127642.279154          0.0
```

```
        [5 rows x 21 columns]
```

This dataset seems to be structured as a table; that is, it contains no strange artifacts after the end of the dataset that would complicate an analysis. From a high-level view, it looks like the beginning and end of the data have appropriate values for each field.

```
In [8]: imported_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
id                      10866 non-null int64
imdb_id                 10856 non-null object
popularity              10866 non-null float64
budget                  10866 non-null int64
revenue                 10866 non-null int64
original_title          10866 non-null object
cast                    10790 non-null object
homepage                2936 non-null object
director                10822 non-null object
tagline                 8042 non-null object
keywords                9373 non-null object
overview                10862 non-null object
runtime                 10866 non-null int64
genres                  10843 non-null object
production_companies    9836 non-null object
release_date            10866 non-null object
vote_count              10866 non-null int64
vote_average            10866 non-null float64
release_year            10866 non-null int64
budget_adj              10866 non-null float64
revenue_adj             10866 non-null float64
dtypes: float64(4), int64(6), object(11)
memory usage: 1.7+ MB
```

On the other hand, there are several fields that are missing values. Furthermore, I know from the description of the dataset that the revenue and budget attributes ('budget', 'revenue', 'budget_adj', 'revenue_adj') contain 0 values that should be treated as missing rather than actual 0 values.

For the fields missing many values, I want to spot check the kind of information contained to see if it seems important. I particularly will be looking at the following: * 'homepage' * 'tagline' * 'keywords' * 'production_companies'

```
In [9]: imported_data['homepage'].value_counts()
```

```
Out[9]: http://phantasm.com
        http://www.missionimpossible.com/
```

http://www.thehungergames.movie/
http://www.kungfupanda.com/
http://www.transformersmovie.com/
http://www.thehobbit.com/
http://www.americanreunionmovie.com/
http://www.georgecarlin.com
http://www.jeffdunham.com
http://www.khartonline.com/
http://www.theamazingspiderman.com
http://www.harrypotter.com
http://www.howtotrainyourdragon.com/
http://stepupmovie.com/
http://eleanorrigby-movie.com/
http://www.lordoftherings.net/
http://disney.go.com/disneypictures/pirates/
http://www.riomovies.com/
http://www.apocalypsenow.com
http://www.miramax.com/movie/kill-bill-volume-1/
http://www.magpictures.com/nymphomaniac/
http://www.munkyourself.com/
http://disney.go.com/tron/
http://www.ironmanmovie.com/
http://www.zeitgeistmovie.com
http://www.beowulfmovie.com/
http://www.sexandthecitymovie.com/
http://www.indianajones.com
http://www.billandted.org/
http://magnetreleasing.com/survivalofthedead/

http://www.welcometohotelt.com
http://www.warnerbros.com/movies/home-entertainment/scanner-darkly-a/d7c290af-c285-41c4-
http://www.vivarockvegas.com/
http://www.23blast.com
http://movies.disney.com/planes-fire-and-rescue
http://www.repomenarecoming.com/
http://www.jupiterascending.com
http://www.universalpictures.com/bestman/
http://www.sonypictures.com/classics/dogtown/
http://www.thelegomovie.com
http://www.spacestation76.com
http://www.timerthemovie.com
http://www.grownups-movie.com/
http://www.lifetime.com
http://www.ifcfilms.com/viewFilm.htm?filmId=245
http://www.funnypeoplemovie.com
http://www.devilsplaygroundmovie.co.uk/
http://www.foxsearchlight.com/cedarrapids/
http://www.mgm.com/view/movie/529/Diamonds-Are-Forever/

```
         http://tomandjerrythemovie.warnerbros.com
         http://www.vivamachete.com/
         http://theidenticalmovie.com/
         http://www.ifcfilms.com/uncategorized/salvation-boulevard
         https://www.facebook.com/finderskeepersdocumentary
         http://www.graceisgone-themovie.com/
         http://www.growththemovie.com/
         http://womaninblack.com/
         http://www.ifcfilms.com/films/house-of-pleasures
         http://www.minionsmovie.com/
         http://theluckyonemovie.warnerbros.com/
         Name: homepage, Length: 2896, dtype: int64
```

```
In [10]: imported_data['tagline'].value_counts()
```

```
Out[10]: Based on a true story.
         Two Films. One Love.
         Be careful what you wish for.
         There are no clean getaways.
         The chase is on!
         One ordinary couple. One little white lie.
         Who is John Galt?
         There are two sides to every love story.
         Worlds Collide
         The end of the world is just the beginning.
         Survival is no game
         Some things are worth fighting for.
         Forget About Love
         It's A Trap
         Love is a force of nature.
         Fight Fire With Fire
         No one is above the law.
         Some lines should never be crossed.
         Victor Crowley Lives Again
         Misery loves family.
         -
         Some houses are born bad.
         Some things are better left buried.
         There is no turning back
         Who's next?
         Adapt or die.
         How far would you go?
         The legend comes to life.
         The timeless tale of a special place where magic, hope and love grow.
         Free your mind.

         They're Down On Their Luck And Up To Their Necks In Senoritas, Margaritas, Banditos And
         How much can a man take...before he gives back?
         Unlock The Universe
         No one is ever really prepared.
         Let the sun shine in!
         Don't Get Mad. Get Evil.
         Fight or die!
         From Backpacks to Strollers
         A comedy about two brothers, a girl with a broken heart, a sex tape, an angel and a pig
         Where everything seems possible and nothing is what it seems.
         Sometimes you don't need more than one person to not feel alone
         Spend Thanksgiving With Good Ol' Charlie Brown!
         He took the job that no one wanted... and got the girl that everyone did.
         Unlock The Ultimate Secret This October!
         Space will never be the same.
         To Life!
         There comes a time to cut loose.
         Take the ride.
         Too Cool For The Rules!
         The way back begins with a single chord.
         You Stop You Die
         A romantic comedy with a whole lot of drama.
         New house. New family. What could possibly go wrong?
         I don't feel I have to wipe everybody out, Tom. Just my enemies.
         How to marry a billionaire
         FLINT'S BACK In Action... In Danger... In the Virgin Islands... Where the Bad Guys... A
         You'll be sorry you were ever born human
         Willy Wonka is semi-sweet and nuts.
         Game On.
         Real War, Real Heroes
         Name: tagline, Length: 7997, dtype: int64
```

```
In [11]: imported_data['keywords'].value_counts()
```

```
Out[11]: woman director                                                     134
         independent film                                                    82
         sport                                                               25
         suspense                                                            24
         musical                                                             24
         duringcreditsstinger                                                24
         holiday                                                             16
         stand-up|stand up comedy                                            16
         biography                                                           15
         independent film|woman director                                     13
         stand up comedy                                                      9
         found footage                                                        7
         holiday|christmas                                                    7
         dystopia                                                             7
         christmas                                                            7
         based on novel                                                       7
```

```
sequel                                                              6
aftercreditsstinger                                                 6
cop|new england|jesse stone                                         5
crime solving                                                       5
aftercreditsstinger|duringcreditsstinger                            5
stand-up                                                            4
werewolf                                                            4
dinosaur                                                            4
zombies                                                             4
baseball|sport                                                      4
based on video game                                                 4
possession                                                          4
independent film|duringcreditsstinger                               4
haunted house                                                       3
                                                                 ...
nudity|boot camp|reality spoof                                      1
film producer|party                                                 1
graduation|ex husband|woman director                                1
wife husband relationship|space travel|space|mission|pregnancy      1
ice|space marine|paranoia|snow storm|norwegian                      1
sport|skiing                                                        1
journalist|journalism|distrust|audio tape|wound                     1
haunted house|inheritance|fireplace|creature|demon                  1
sex|seduction|gigolo|callboy|party                                  1
fire|recession|diary|time travel|murder                             1
prostitute|photographer|brothel|virgin|new orleans                  1
dc comics|based on comic book|superhero team|super powers           1
new york|concentration camp|holocaust|writer                        1
based on novel|attack|massacre|private                              1
based on novel|biography|based on true story                        1
seduction|college|love|friends|betrayal                             1
holy grail|camelot|round table|wizardry|merlin                      1
robbery|inventor|penguin|telecontrol|surrealism                     1
individual|slum|suicide|tattoo|alcohol                              1
ventriloquist|doll|ventriloquist dummy                              1
soulmates|vampire|forbidden love|immortality|trust                  1
rock star|heavy metal|recording studio|psychologist|conflict        1
adventure|farscape|tv mini-series                                   1
comedian|stand-up|stand up comedy|clean comedy                      1
sport|figure skating|olympics                                       1
female nudity|robbery|mail order bride|bank clerk                   1
spy|airport|gas station|garage|pilot                                1
nurse|patriotism|hawaii|world war ii|pilot                          1
corruption|terrorist|explosive|police|kidnapping                    1
sea|fireworks|prince|kingdom|daughter                               1
Name: keywords, Length: 8804, dtype: int64
```

```
In [12]: imported_data['production_companies'].value_counts()
```

```
Out[12]: Paramount Pictures
         Universal Pictures
         Warner Bros.
         Walt Disney Pictures
         Metro-Goldwyn-Mayer (MGM)
         Columbia Pictures
         New Line Cinema
         Touchstone Pictures
         20th Century Fox
         Twentieth Century Fox Film Corporation
         TriStar Pictures
         Orion Pictures
         Miramax Films
         Columbia Pictures Corporation
         DreamWorks Animation
         Pixar Animation Studios
         Walt Disney Productions
         Dimension Films
         United Artists
         Marvel Studios
         Imagine Entertainment|Universal Pictures
         The Asylum
         Lions Gate Films
         Walt Disney Pictures|Pixar Animation Studios
         New World Pictures
         American International Pictures (AIP)
         Disney Channel
         Hammer Film Productions
         Hollywood Pictures
         Walt Disney Pictures|Walt Disney Animation Studios

         Haven Entertainment|Sandia Media|Minerva Productions
         Kanzaman S.A.|Scion Films Limited|Millenium Films|Black Forest Films|Double Edge Entert
         Yari Film Group|Furst Films
         Twentieth Century Fox Film Corporation|SLM Production Group|Silver Pictures
         Company Films|Voltage Pictures
         New Line Cinema|Goldsmith-Thomas Productions|Red Om Films|HBO Films|Picturehouse
         Dune Entertainment|DiNovi Pictures
         Paramount Pictures|Twentieth Century Fox Film Corporation|Lightstorm Entertainment
         Columbia Pictures Corporation|Stonebridge Entertainment
         Das Films|Living Out Loud Films|Elephant Eye Films
         Institution, The
         Likely Story|ATO Pictures|Olympus Pictures
         Legendary Pictures|Green Hat Films|Warner Bros.|IFP Westcoast Erste
         Producers Circle|Incorporated Television Company
         Fountainbridge Films|Warner Bros.|Lee Rich Productions
         Myriad Pictures|CJ Entertainment|Toiion
         BBC Films|Aramid Entertainment Fund
```

```
        Lifetime Network|Sony Pictures Television|Woodridge Productions
        Raven Banner Entertainment|De Angeles Films
        Columbia Pictures Corporation|Tom Ward Enterprises|Rastar Productions
        Columbia Pictures|Centropolis Film Productions
        Universal Pictures|Saga Film|Focus Films|Gold Circle Films|Chambara Pictures
        Columbia Pictures|Imagine Entertainment|Revolution Studios
        Screen Gems|Olive Bridge Entertainment
        Antzworks
        Paramount Pictures|Mutual Film Company|Skydance Productions|TC Productions
        Twentieth Century Fox Film Corporation|Lawrence Gordon Productions|Davis Entertainment|
        Mad Circus Films|Lions Gate Entertainments|Mr. X
        Walt Disney Pictures|Robert Simonds Productions
        Universal Pictures|Chernin Entertainment|Relativity Media|Monolith Pictures (III)|Radic
        Name: production_companies, Length: 7445, dtype: int64
```

OK, so those values make sense for these attributes, but they don't really seem helpful to my goals. Maybe the homepages would be interesting if I were interested in having seeds for a web crawler, or the taglines to run through a natural language analysis – maybe something to correlate sentiment to viewership. The keywords are too sporadic and unstructured to be particularly helpful, I would think.

I'm going to drop these attributes from my dataframe. It'd probably be OK to leave them there and ignore then, but why keep them in memory if I don't have to?

```
In [13]: imported_data.drop(['homepage', 'tagline', 'keywords', 'production_companies'], axis=1,
         imported_data.drop_duplicates(inplace=True)
```

```
In [14]: imported_data.shape
```

```
Out[14]: (10865, 17)
```

Awesome! Now, another thing I noticed about is that the release year isn't exactly helpful, so I want to change it to a datetime object in case I end up needing to do anything with the date.

```
In [15]: from datetime import datetime

         def change_to_date(date_string):
             return datetime.strptime(date_string, "%m/%d/%y")

         imported_data['release_date'] = imported_data['release_date'].apply(change_to_date)
```

```
In [16]: imported_data
```

```
Out[16]:            id    imdb_id  popularity     budget     revenue  \
         0      135397  tt0369610   32.985763  150000000  1513528810
         1       76341  tt1392190   28.419936  150000000   378436354
         2      262500  tt2908446   13.112507  110000000   295238201
         3      140607  tt2488496   11.173104  200000000  2068178225
         4      168259  tt2820852    9.335014  190000000  1506249360
         5      281957  tt1663202    9.110700  135000000   532950503
```

```
6       87101  tt1340138    8.654359  155000000   440603537
7      286217  tt3659388    7.667400  108000000   595380321
8      211672  tt2293640    7.404165   74000000  1156730962
9      150540  tt2096673    6.326804  175000000   853708609
10     206647  tt2379713    6.200282  245000000   880674609
11      76757  tt1617661    6.189369  176000003   183987723
12     264660  tt0470752    6.118847   15000000    36869414
13     257344  tt2120120    5.984995   88000000   243637091
14      99861  tt2395427    5.944927  280000000  1405035767
15     273248  tt3460252    5.898400   44000000   155760117
16     260346  tt2446042    5.749758   48000000   325771424
17     102899  tt0478970    5.573184  130000000   518602163
18     150689  tt1661199    5.556818   95000000   542351353
19     131634  tt1951266    5.476958  160000000   650523427
20     158852  tt1964418    5.462138  190000000   209035668
21     307081  tt1798684    5.337064   30000000    91709827
22     254128  tt2126355    4.907832  110000000   470490832
23     216015  tt2322441    4.710402   40000000   569651467
24     318846  tt1596363    4.648046   28000000   133346506
25     177677  tt2381249    4.566713  150000000   682330139
26     214756  tt2637276    4.564549   68000000   215863606
27     207703  tt2802144    4.503789   81000000   403802136
28     314365  tt1895587    4.062293   20000000    88346473
29     294254  tt4046784    3.968891   61000000   311256926
...       ...        ...         ...        ...         ...
10836   38720  tt0061170    0.239435          0           0
10837   19728  tt0060177    0.291704          0           0
10838   22383  tt0060862    0.151845          0           0
10839   13353  tt0060550    0.276133          0           0
10840   34388  tt0060437    0.102530          0           0
10841   42701  tt0062262    0.264925      75000           0
10842   36540  tt0061199    0.253437          0           0
10843   29710  tt0060588    0.252399          0           0
10844   23728  tt0059557    0.236098          0           0
10845    5065  tt0059014    0.230873          0           0
10846   17102  tt0059127    0.212716          0           0
10847   28763  tt0060548    0.034555          0           0
10848    2161  tt0060397    0.207257    5115000    12000000
10849   28270  tt0060445    0.206537          0           0
10850   26268  tt0060490    0.202473          0           0
10851   15347  tt0060182    0.342791          0           0
10852   37301  tt0060165    0.227220          0           0
10853   15598  tt0060086    0.163592          0           0
10854   31602  tt0060232    0.146402          0           0
10855   13343  tt0059221    0.141026     700000           0
10856   20277  tt0061135    0.140934          0           0
10857    5921  tt0060748    0.131378          0           0
10858   31918  tt0060921    0.317824          0           0
```

```
10859   20620  tt0060955    0.089072          0           0
10860    5060  tt0060214    0.087034          0           0
10861      21  tt0060371    0.080598          0           0
10862   20379  tt0060472    0.065543          0           0
10863   39768  tt0060161    0.065141          0           0
10864   21449  tt0061177    0.064317          0           0
10865   22293  tt0060666    0.035919      19000           0

                                         original_title  \
0                                        Jurassic World
1                                    Mad Max: Fury Road
2                                             Insurgent
3                          Star Wars: The Force Awakens
4                                             Furious 7
5                                          The Revenant
6                                    Terminator Genisys
7                                           The Martian
8                                               Minions
9                                            Inside Out
10                                              Spectre
11                                    Jupiter Ascending
12                                           Ex Machina
13                                               Pixels
14                              Avengers: Age of Ultron
15                                    The Hateful Eight
16                                              Taken 3
17                                              Ant-Man
18                                           Cinderella
19                The Hunger Games: Mockingjay - Part 2
20                                         Tomorrowland
21                                             Southpaw
22                                          San Andreas
23                                 Fifty Shades of Grey
24                                        The Big Short
25                   Mission: Impossible - Rogue Nation
26                                                Ted 2
27                         Kingsman: The Secret Service
28                                            Spotlight
29                       Maze Runner: The Scorch Trials
...                                                 ...
10836                                    Walk Don't Run
10837                                      The Blue Max
10838                                 The Professionals
10839             It's the Great Pumpkin, Charlie Brown
10840                                 Funeral in Berlin
10841                                      The Shooting
10842                Winnie the Pooh and the Honey Tree
10843                                          Khartoum
```

```
10844                                     Our Man Flint
10845                                   Carry On Cowboy
10846                       Dracula: Prince of Darkness
10847                                  Island of Terror
10848                                  Fantastic Voyage
10849                                            Gambit
10850                                            Harper
10851                                         Born Free
10852                    A Big Hand for the Little Lady
10853                                             Alfie
10854                                         The Chase
10855                           The Ghost & Mr. Chicken
10856                                The Ugly Dachshund
10857                                      Nevada Smith
10858  The Russians Are Coming, The Russians Are Coming
10859                                           Seconds
10860                               Carry On Screaming!
10861                                The Endless Summer
10862                                        Grand Prix
10863                               Beregis Avtomobilya
10864                            What's Up, Tiger Lily?
10865                          Manos: The Hands of Fate

                                                    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 ...
5      Leonardo DiCaprio|Tom Hardy|Will Poulter|Domhn...
6      Arnold Schwarzenegger|Jason Clarke|Emilia Clar...
7      Matt Damon|Jessica Chastain|Kristen Wiig|Jeff ...
8      Sandra Bullock|Jon Hamm|Michael Keaton|Allison...
9      Amy Poehler|Phyllis Smith|Richard Kind|Bill Ha...
10     Daniel Craig|Christoph Waltz|LÃI’a Seydoux|Ralp...
11     Mila Kunis|Channing Tatum|Sean Bean|Eddie Redm...
12     Domhnall Gleeson|Alicia Vikander|Oscar Isaac|S...
13     Adam Sandler|Michelle Monaghan|Peter Dinklage|...
14     Robert Downey Jr.|Chris Hemsworth|Mark Ruffalo...
15     Samuel L. Jackson|Kurt Russell|Jennifer Jason ...
16     Liam Neeson|Forest Whitaker|Maggie Grace|Famke...
17     Paul Rudd|Michael Douglas|Evangeline Lilly|Cor...
18     Lily James|Cate Blanchett|Richard Madden|Helen...
19     Jennifer Lawrence|Josh Hutcherson|Liam Hemswor...
20     Britt Robertson|George Clooney|Raffey Cassidy|...
21     Jake Gyllenhaal|Rachel McAdams|Forest Whitaker...
22     Dwayne Johnson|Alexandra Daddario|Carla Gugino...
23     Dakota Johnson|Jamie Dornan|Jennifer Ehle|Eloi...
```

```
24       Christian Bale|Steve Carell|Ryan Gosling|Brad ...
25       Tom Cruise|Jeremy Renner|Simon Pegg|Rebecca Fe...
26       Mark Wahlberg|Seth MacFarlane|Amanda Seyfried|...
27       Taron Egerton|Colin Firth|Samuel L. Jackson|Mi...
28       Mark Ruffalo|Michael Keaton|Rachel McAdams|Lie...
29       Dylan O'Brien|Kaya Scodelario|Thomas Brodie-Sa...
...                                                   ...
10836  Cary Grant|Samantha Eggar|Jim Hutton|John Stan...
10837  George Peppard|James Mason|Ursula Andress|Jere...
10838  Burt Lancaster|Lee Marvin|Robert Ryan|Woody St...
10839  Christopher Shea|Sally Dryer|Kathy Steinberg|A...
10840  Michael Caine|Paul Hubschmid|Oskar Homolka|Eva...
10841  Will Hutchins|Millie Perkins|Jack Nicholson|Wa...
10842  Sterling Holloway|Junius Matthews|Sebastian Ca...
10843  Charlton Heston|Laurence Olivier|Richard Johns...
10844  James Coburn|Lee J. Cobb|Gila Golan|Edward Mul...
10845  Sid James|Jim Dale|Angela Douglas|Kenneth Will...
10846  Christopher Lee|Barbara Shelley|Andrew Keir|Fr...
10847  Peter Cushing|Edward Judd|Carole Gray|Eddie By...
10848  Stephen Boyd|Raquel Welch|Edmond O'Brien|Donal...
10849  Michael Caine|Shirley MacLaine|Herbert Lom|Joh...
10850  Paul Newman|Lauren Bacall|Julie Harris|Arthur ...
10851  Virginia McKenna|Bill Travers|Geoffrey Keen|Pe...
10852  Henry Fonda|Joanne Woodward|Jason Robards|Paul...
10853  Michael Caine|Shelley Winters|Millicent Martin...
10854  Marlon Brando|Jane Fonda|Robert Redford|E.G. M...
10855  Don Knotts|Joan Staley|Liam Redmond|Dick Sarge...
10856  Dean Jones|Suzanne Pleshette|Charles Ruggles|K...
10857  Steve McQueen|Karl Malden|Brian Keith|Arthur K...
10858  Carl Reiner|Eva Marie Saint|Alan Arkin|Brian K...
10859  Rock Hudson|Salome Jens|John Randolph|Will Gee...
10860  Kenneth Williams|Jim Dale|Harry H. Corbett|Joa...
10861  Michael Hynson|Robert August|Lord 'Tally Ho' B...
10862  James Garner|Eva Marie Saint|Yves Montand|Tosh...
10863  Innokentiy Smoktunovskiy|Oleg Efremov|Georgi Z...
10864  Tatsuya Mihashi|Akiko Wakabayashi|Mie Hama|Joh...
10865  Harold P. Warren|Tom Neyman|John Reynolds|Dian...

                            director  \
0                    Colin Trevorrow
1                      George Miller
2                  Robert Schwentke
3                       J.J. Abrams
4                          James Wan
5      Alejandro GonzÃąlez IÃśÃąrritu
6                        Alan Taylor
7                       Ridley Scott
8             Kyle Balda|Pierre Coffin
```

```
9                           Pete Docter
10                           Sam Mendes
11     Lana Wachowski|Lilly Wachowski
12                         Alex Garland
13                       Chris Columbus
14                          Joss Whedon
15                    Quentin Tarantino
16                      Olivier Megaton
17                          Peyton Reed
18                      Kenneth Branagh
19                     Francis Lawrence
20                            Brad Bird
21                        Antoine Fuqua
22                          Brad Peyton
23                   Sam Taylor-Johnson
24                           Adam McKay
25              Christopher McQuarrie
26                      Seth MacFarlane
27                       Matthew Vaughn
28                         Tom McCarthy
29                             Wes Ball
...                                 ...
10836                  Charles Walters
10837                  John Guillermin
10838                   Richard Brooks
10839                    Bill Melendez
10840                     Guy Hamilton
10841                    Monte Hellman
10842              Wolfgang Reitherman
10843     Basil Dearden|Eliot Elisofon
10844                      Daniel Mann
10845                    Gerald Thomas
10846                   Terence Fisher
10847                   Terence Fisher
10848                Richard Fleischer
10849                     Ronald Neame
10850                      Jack Smight
10851                       James Hill
10852                      Fielder Cook
10853                    Lewis Gilbert
10854                      Arthur Penn
10855                      Alan Rafkin
10856                     Norman Tokar
10857                   Henry Hathaway
10858                   Norman Jewison
10859               John Frankenheimer
10860                    Gerald Thomas
10861                      Bruce Brown
```

```
10862              John Frankenheimer
10863                  Eldar Ryazanov
10864                     Woody Allen
10865                Harold P. Warren

                                                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
5      In the 1820s, a frontiersman, Hugh Glass, sets...      156
6      The year is 2029. John Connor, leader of the r...      125
7      During a manned mission to Mars, Astronaut Mar...      141
8      Minions Stuart, Kevin and Bob are recruited by...       91
9      Growing up can be a bumpy road, and it's no ex...       94
10     A cryptic message from Bondâs past sends him...        148
11     In a universe where human genetic material is ...      124
12     Caleb, a 26 year old coder at the world's larg...      108
13     Video game experts are recruited by the milita...      105
14     When Tony Stark tries to jumpstart a dormant p...      141
15     Bounty hunters seek shelter from a raging bliz...      167
16     Ex-government operative Bryan Mills finds his ...      109
17     Armed with the astonishing ability to shrink i...      115
18     When her father unexpectedly passes away, youn...      112
19     With the nation of Panem in a full scale war, ...      136
20     Bound by a shared destiny, a bright, optimisti...      130
21     Billy "The Great" Hope, the reigning junior mi...      123
22     In the aftermath of a massive earthquake in Ca...      114
23     When college senior Anastasia Steele steps in ...      125
24     The men who made millions from a global econom...      130
25     Ethan and team take on their most impossible m...      131
26     Newlywed couple Ted and Tami-Lynn want to have...      115
27     The story of a super-secret spy organization t...      130
28     The true story of how The Boston Globe uncover...      128
29     Thomas and his fellow Gladers face their great...      132
...                                                  ...      ...
10836  British industrialist Sir William Rutland - "B...      114
10837  A young pilot in the German air force of 1918,...      156
10838  The Professionals is a 1966 American Western f...      117
10839  This classic "Peanuts" tale focuses on the thu...       25
10840  Colonel Stok, a Soviet intelligence officer re...      102
10841  A hired gun seeks to enact revenge on a group ...       82
10842  Christopher Robin's bear attempts to raid a be...       25
10843  English General Charles George Gordon, a devou...      134
10844  When scientists use eco-terrorism to impose th...      108
10845  Stodge City is in the grip of the Rumpo Kid an...       93
10846  Whilst vacationing in the Carpathian Mountain,...       90
```

```
10847  A small island community is overrun with creep...       89
10848  The science of miniaturization has been unlock...      100
10849  Harry Dean (Michael Caine) has a perfect plan ...      109
10850  Harper is a cynical private eye in the best tr...      121
10851  Born Free (1966) is an Open Road Films Ltd./Co...       95
10852  A naive traveler in Laredo gets involved in a ...       95
10853  The film tells the story of a young man who le...      114
10854  Most everyone in town thinks that Sheriff Cald...      135
10855  Luther Heggs aspires to being a reporter for h...       90
10856  The Garrisons (Dean Jones and Suzanne Pleshett...       93
10857  Nevada Smith is the young son of an Indian mot...      128
10858  Without hostile intent, a Soviet sub runs agro...      126
10859  A secret organisation offers wealthy people a ...      100
10860  The sinister Dr Watt has an evil scheme going...       87
10861  The Endless Summer, by Bruce Brown, is one of ...       95
10862  Grand Prix driver Pete Aron is fired by his te...      176
10863  An insurance agent who moonlights as a carthie...       94
10864  In comic Woody Allen's film debut, he took the...       80
10865  A family gets lost on the road and stumbles up...       74

                                                genres release_date  \
0            Action|Adventure|Science Fiction|Thriller   2015-06-09
1            Action|Adventure|Science Fiction|Thriller   2015-05-13
2                   Adventure|Science Fiction|Thriller   2015-03-18
3             Action|Adventure|Science Fiction|Fantasy   2015-12-15
4                                Action|Crime|Thriller   2015-04-01
5                     Western|Drama|Adventure|Thriller   2015-12-25
6            Science Fiction|Action|Thriller|Adventure   2015-06-23
7                      Drama|Adventure|Science Fiction   2015-09-30
8                    Family|Animation|Adventure|Comedy   2015-06-17
9                              Comedy|Animation|Family   2015-06-09
10                              Action|Adventure|Crime   2015-10-26
11            Science Fiction|Fantasy|Action|Adventure   2015-02-04
12                               Drama|Science Fiction   2015-01-21
13                      Action|Comedy|Science Fiction   2015-07-16
14                   Action|Adventure|Science Fiction   2015-04-22
15                        Crime|Drama|Mystery|Western   2015-12-25
16                               Crime|Action|Thriller   2015-01-01
17                   Science Fiction|Action|Adventure   2015-07-14
18                       Romance|Fantasy|Family|Drama   2015-03-12
19                      War|Adventure|Science Fiction   2015-11-18
20     Action|Family|Science Fiction|Adventure|Mystery   2015-05-19
21                                        Action|Drama   2015-06-15
22                               Action|Drama|Thriller   2015-05-27
23                                       Drama|Romance   2015-02-11
24                                        Comedy|Drama   2015-12-11
25                                              Action   2015-07-23
26                                              Comedy   2015-06-25
```

```
27                       Crime|Comedy|Action|Adventure   2015-01-24
28                              Drama|Thriller|History   2015-11-06
29                     Action|Science Fiction|Thriller   2015-09-09
...                                                ...          ...
10836                                   Comedy|Romance   2066-01-01
10837                       War|Action|Adventure|Drama   2066-06-21
10838                         Action|Adventure|Western   2066-11-01
10839                                 Family|Animation   2066-10-27
10840                                         Thriller   2066-12-22
10841                                          Western   2066-10-23
10842                                 Animation|Family   2066-01-01
10843               Adventure|Drama|War|History|Action   2066-06-09
10844         Adventure|Comedy|Fantasy|Science Fiction   2066-01-16
10845                                   Comedy|Western   2066-03-01
10846                                           Horror   2066-01-09
10847                           Science Fiction|Horror   2066-06-20
10848                        Adventure|Science Fiction   2066-08-24
10849                              Action|Comedy|Crime   2066-12-16
10850              Action|Drama|Thriller|Crime|Mystery   2066-02-23
10851            Adventure|Drama|Action|Family|Foreign   2066-06-22
10852                                          Western   2066-05-31
10853                             Comedy|Drama|Romance   2066-03-29
10854                             Thriller|Drama|Crime   2066-02-17
10855                    Comedy|Family|Mystery|Romance   2066-01-20
10856                              Comedy|Drama|Family   2066-02-16
10857                                   Action|Western   2066-06-10
10858                                       Comedy|War   2066-05-25
10859           Mystery|Science Fiction|Thriller|Drama   2066-10-05
10860                                           Comedy   2066-05-20
10861                                      Documentary   2066-06-15
10862                           Action|Adventure|Drama   2066-12-21
10863                                   Mystery|Comedy   2066-01-01
10864                                    Action|Comedy   2066-11-02
10865                                           Horror   2066-11-15

       vote_count  vote_average  release_year    budget_adj   revenue_adj
0            5562           6.5          2015  1.379999e+08  1.392446e+09
1            6185           7.1          2015  1.379999e+08  3.481613e+08
2            2480           6.3          2015  1.012000e+08  2.716190e+08
3            5292           7.5          2015  1.839999e+08  1.902723e+09
4            2947           7.3          2015  1.747999e+08  1.385749e+09
5            3929           7.2          2015  1.241999e+08  4.903142e+08
6            2598           5.8          2015  1.425999e+08  4.053551e+08
7            4572           7.6          2015  9.935996e+07  5.477497e+08
8            2893           6.5          2015  6.807997e+07  1.064192e+09
9            3935           8.0          2015  1.609999e+08  7.854116e+08
10           3254           6.2          2015  2.253999e+08  8.102203e+08
11           1937           5.2          2015  1.619199e+08  1.692686e+08
```

```
12           2854           7.6          2015  1.379999e+07  3.391985e+07
13           1575           5.8          2015  8.095996e+07  2.241460e+08
14           4304           7.4          2015  2.575999e+08  1.292632e+09
15           2389           7.4          2015  4.047998e+07  1.432992e+08
16           1578           6.1          2015  4.415998e+07  2.997096e+08
17           3779           7.0          2015  1.195999e+08  4.771138e+08
18           1495           6.8          2015  8.739996e+07  4.989630e+08
19           2380           6.5          2015  1.471999e+08  5.984813e+08
20           1899           6.2          2015  1.747999e+08  1.923127e+08
21           1386           7.3          2015  2.759999e+07  8.437300e+07
22           2060           6.1          2015  1.012000e+08  4.328514e+08
23           1865           5.3          2015  3.679998e+07  5.240791e+08
24           1545           7.3          2015  2.575999e+07  1.226787e+08
25           2349           7.1          2015  1.379999e+08  6.277435e+08
26           1666           6.3          2015  6.255997e+07  1.985944e+08
27           3833           7.6          2015  7.451997e+07  3.714978e+08
28           1559           7.8          2015  1.839999e+07  8.127872e+07
29           1849           6.4          2015  5.611998e+07  2.863562e+08
...           ...           ...           ...           ...           ...
10836          11           5.8          1966  0.000000e+00  0.000000e+00
10837          12           5.5          1966  0.000000e+00  0.000000e+00
10838          21           6.0          1966  0.000000e+00  0.000000e+00
10839          49           7.2          1966  0.000000e+00  0.000000e+00
10840          13           5.7          1966  0.000000e+00  0.000000e+00
10841          12           5.5          1966  5.038511e+05  0.000000e+00
10842          12           7.9          1966  0.000000e+00  0.000000e+00
10843          12           5.8          1966  0.000000e+00  0.000000e+00
10844          13           5.6          1966  0.000000e+00  0.000000e+00
10845          15           5.9          1966  0.000000e+00  0.000000e+00
10846          16           5.7          1966  0.000000e+00  0.000000e+00
10847          13           5.3          1966  0.000000e+00  0.000000e+00
10848          42           6.7          1966  3.436265e+07  8.061618e+07
10849          14           6.1          1966  0.000000e+00  0.000000e+00
10850          14           6.0          1966  0.000000e+00  0.000000e+00
10851          15           6.6          1966  0.000000e+00  0.000000e+00
10852          11           6.0          1966  0.000000e+00  0.000000e+00
10853          26           6.2          1966  0.000000e+00  0.000000e+00
10854          17           6.0          1966  0.000000e+00  0.000000e+00
10855          14           6.1          1966  4.702610e+06  0.000000e+00
10856          14           5.7          1966  0.000000e+00  0.000000e+00
10857          10           5.9          1966  0.000000e+00  0.000000e+00
10858          11           5.5          1966  0.000000e+00  0.000000e+00
10859          22           6.6          1966  0.000000e+00  0.000000e+00
10860          13           7.0          1966  0.000000e+00  0.000000e+00
10861          11           7.4          1966  0.000000e+00  0.000000e+00
10862          20           5.7          1966  0.000000e+00  0.000000e+00
10863          11           6.5          1966  0.000000e+00  0.000000e+00
10864          22           5.4          1966  0.000000e+00  0.000000e+00
```

```
10865          15            1.5          1966  1.276423e+05  0.000000e+00

[10865 rows x 17 columns]
```

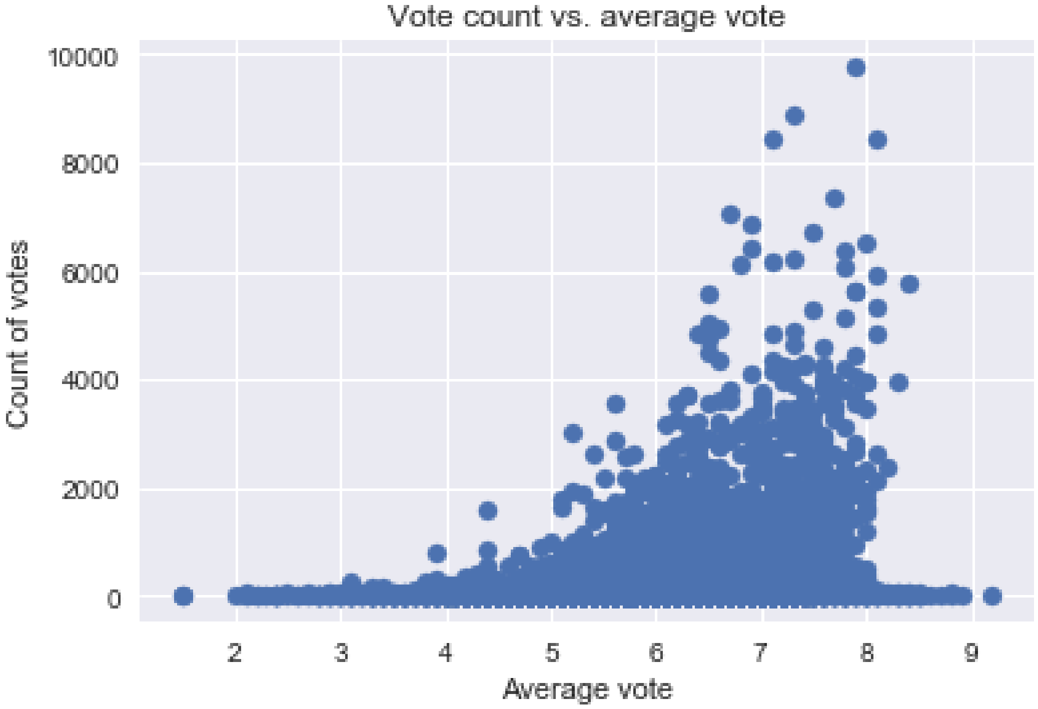
There we go; that seems to be a good start. There are some annoying facts about the data – like that 'cast' and 'genres' are stored as pipe-delimited values. That seems extremely unnecessary, but I'm not the architect, so I'll refrain from further judgment at this time.

Spoiler alert: this does end up causing me a lot of frustration, especially because of how Pandas interacts with strings.

2 Exploring the Data

An easy way to start exploring data is to look at some correlations. We can't make any statistical inferences from scatterplots alone, but we might see that data trends in a particular way. I'll start by looking at the relationship between the number of votes, and the average rating. Maybe we'll see a common adage, "people are more likely to express an opinion for something they dislike."

```
In [17]: vote_counts = imported_data['vote_count']
         vote_averages = imported_data['vote_average']

         plt.scatter(vote_averages, vote_counts)
         plt.title("Vote count vs. average vote")
         plt.xlabel("Average vote")
         plt.ylabel("Count of votes")

Out[17]: <matplotlib.text.Text at 0x2242ea66cf8>
```

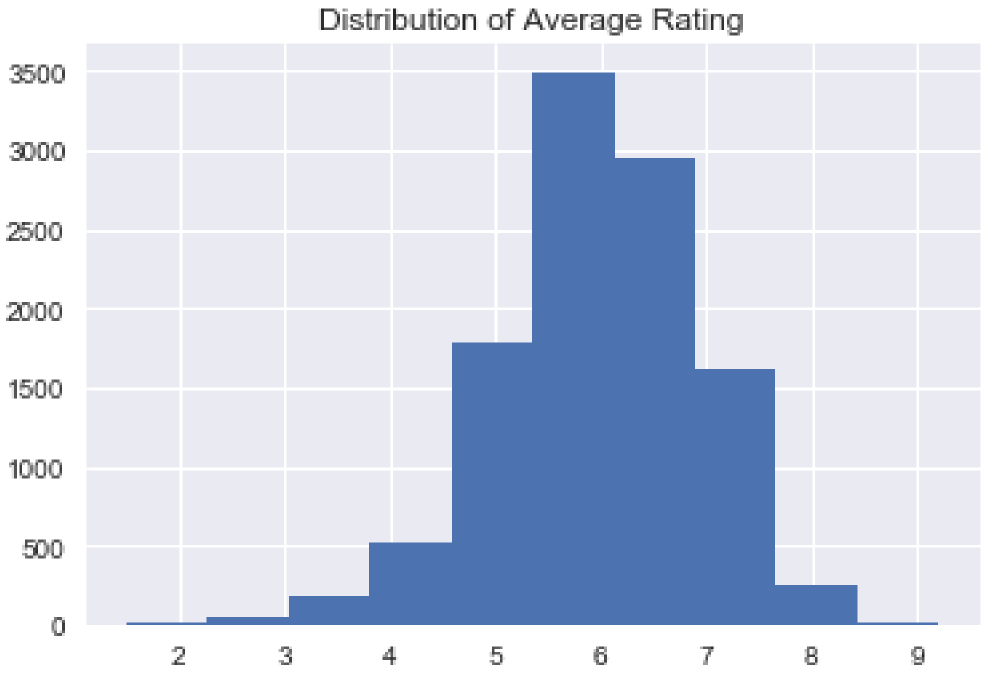


Whoa! There's a lot of records hovering at an extremely low number of votes. You can almost claim a positive correlation between the count and average score, but I'm extremely hesitant to do so since so many entries

```
In [18]: low_votes = len(imported_data.query('vote_count < 100'))
         high_votes = len(imported_data.query('vote_count >= 100'))
         low_votes, high_votes, low_votes/len(imported_data)

Out[18]: (7537, 3328, 0.6936953520478601)

In [19]: len(imported_data.query('vote_count <= 10'))/len(imported_data)

Out[19]: 0.0461113667740451
```

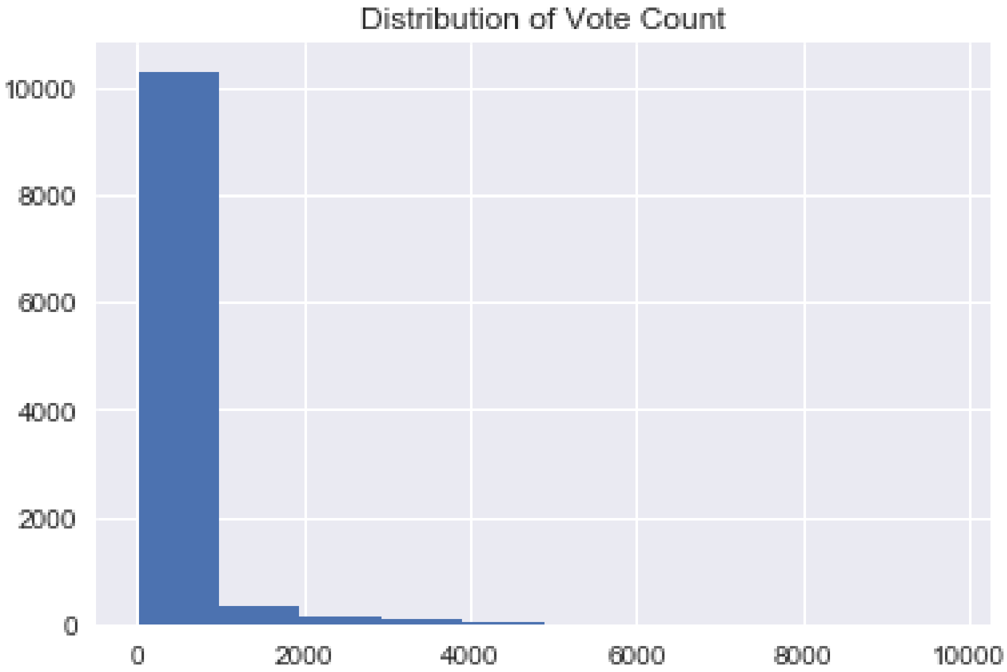
Interesting. Almost 70% of the movies have less than 100 votes. It's hard to define what might even be considered an outlier with such a density of low votes. 5% of the movies have 10 or less votes! I wonder what the distribution of ratings looks like…

```
In [20]: plt.hist(vote_averages)
         plt.title('Distribution of Average Rating')

Out[20]: <matplotlib.text.Text at 0x2242eb28e10>
```



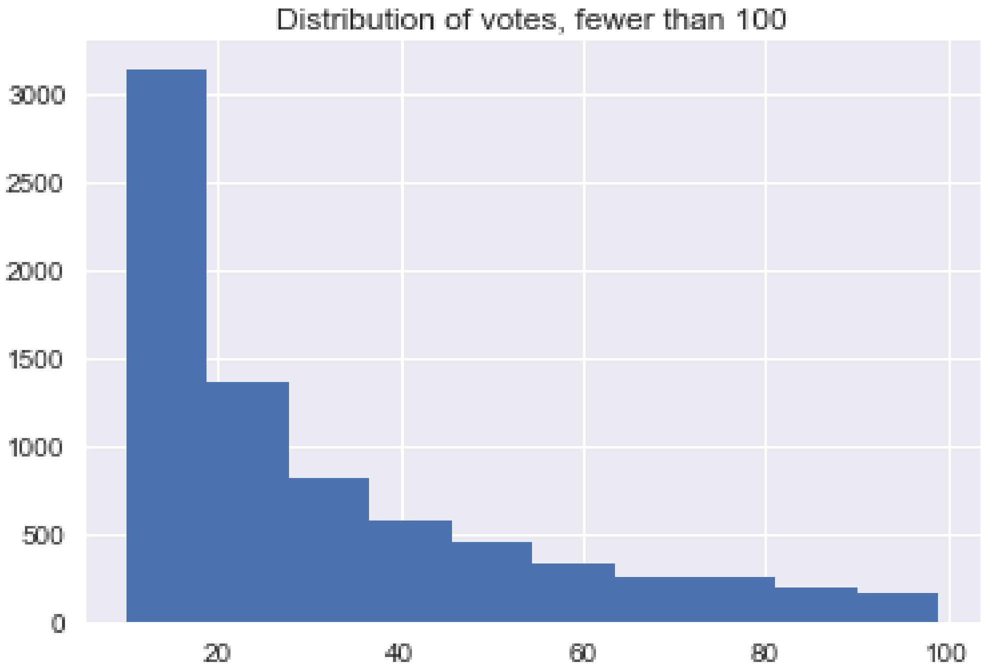
Maybe unsurprisingly, the distribution seems to match the general count of votes. At least the averages are normally distributed.

```
In [21]: plt.hist(vote_counts)
         plt.title('Distribution of Vote Count')

Out[21]: <matplotlib.text.Text at 0x2242ebffd30>
```



The distribution of vote counts, on the other hand, is so extremely right skewed I wonder if it's even usable as a datapoint.

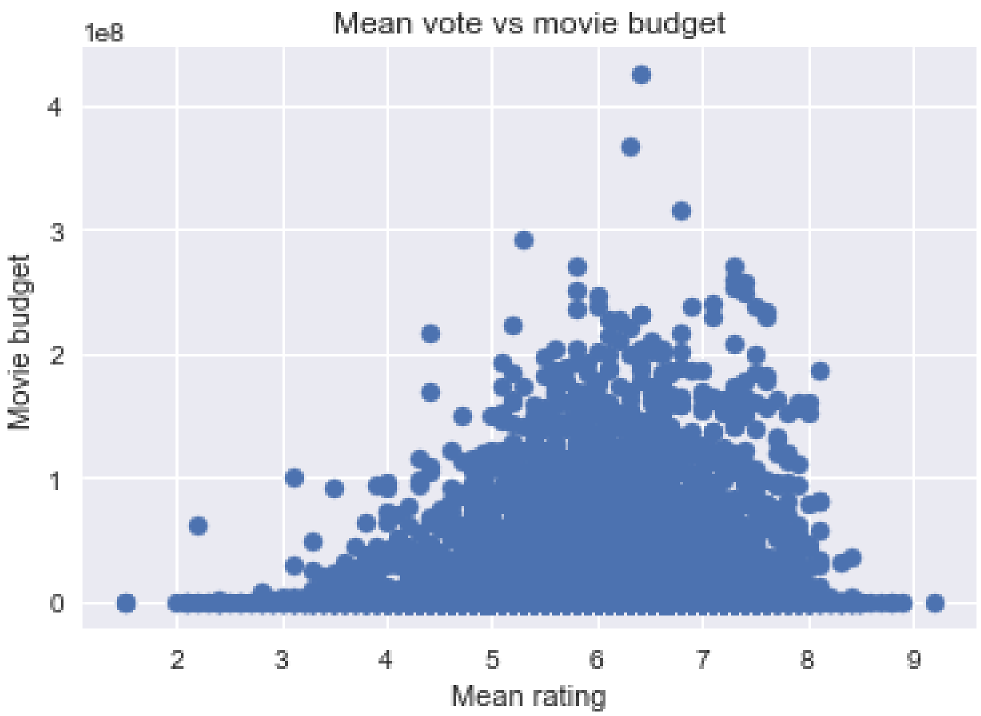
```
In [22]: vote_counts_less_than_100 = imported_data.query('vote_count < 100')['vote_count']

         plt.hist(vote_counts_less_than_100)
         plt.title('Distribution of votes, fewer than 100')

Out[22]: <matplotlib.text.Text at 0x2242ecb4160>
```



We can see that the distribution is the same when only considering records with fewer than 100 votes. We could probably keep drilling down, but I'm not sure that would be a productive exercise.

Instead, let's look at some of the other data points, like whether the budget has any impact on the rating.

```
In [23]: budgets = imported_data['budget_adj']

         plt.scatter(vote_averages, budgets)
         plt.title("Mean vote vs movie budget")
         plt.xlabel("Mean rating")
         plt.ylabel("Movie budget")

Out[23]: <matplotlib.text.Text at 0x2242ed52048>
```



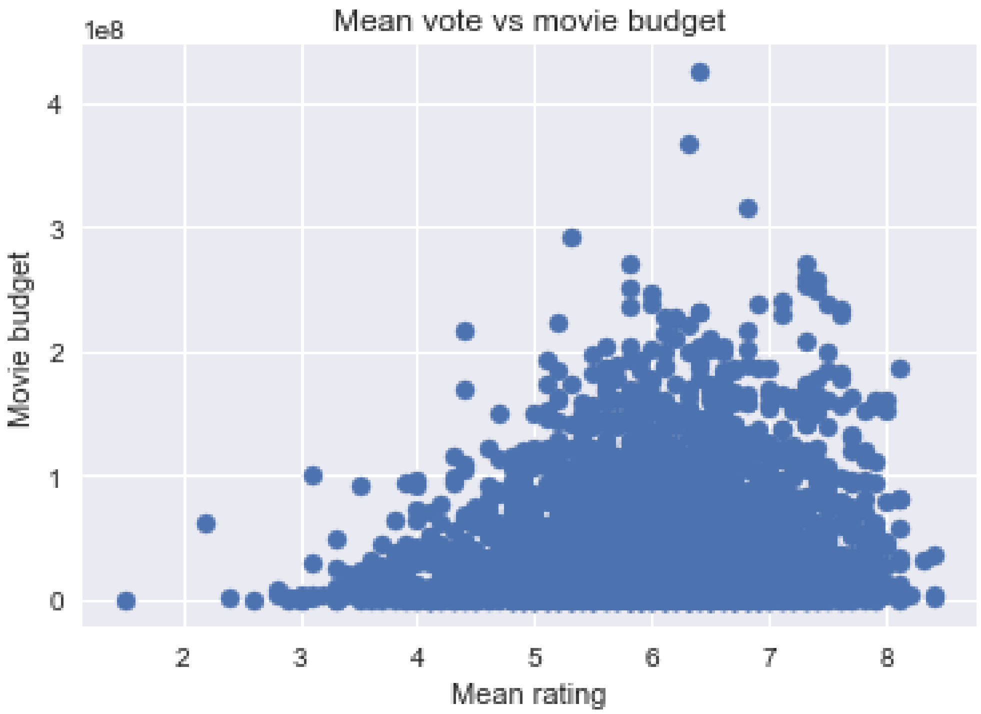
Yikes! We can see that this value is similarly clustered around a low budget.

```
In [24]: imported_data['budget_adj'].agg(['min', 'max'])

Out[24]: min            0.0
         max    425000000.0
         Name: budget_adj, dtype: float64
```

OK, let's remove the 0 values from the budget, which is something we know should be discounted from the description of the data.

```
In [25]: budget_df = imported_data[['budget_adj', 'vote_average']].query('budget_adj != 0')
         budget_df['budget_adj'].agg(['min', 'max'])

Out[25]: min    9.210911e-01
         max    4.250000e+08
         Name: budget_adj, dtype: float64

In [26]: budgets = budget_df['budget_adj']
         vote_averages = budget_df['vote_average']

         plt.scatter(vote_averages, budgets)
         plt.title("Mean vote vs movie budget")
         plt.xlabel("Mean rating")
         plt.ylabel("Movie budget")
```

```
Out[26]: <matplotlib.text.Text at 0x2242edb66a0>
```



Interestingly, the shape of this scatterplot doesn't seem to significantly change. I wonder if these numerical data are useful in the context of an analysis.

Perhaps it's best to switch tactics. Let's take a look at the genres included for each movie, to get a sense of how commonly a label is applied to a movie. I'm sure it's a simple matter to split a series in Pandas to aggregate a count of each value after splitting the value of each cell.

Ha, ha! Just kidding. This exercise was actually one of the more painful things I've ever tried to do with Pandas. I'm not sure the exact reason I was having trouble, but I believe it might be related to one of two implementation details with the library: 1. Pandas stores strings as pointers, and lazily evaluates the items in a series. The library doesn't intelligently apply the split function to the string but rather the pointer value.
2. Pandas lazily evaluates strings and iteritems(), and despite a standard Python pattern to retrieve only the value from the iterator, the library always returns the index value first. When attempting to split the string in the iterator, the split function is applied to the index value.

Of course, neither of these hypotheses could be correct and I'm uncertain how to go about discovering more or reporting a bug (feature?) to the author. For one, this specific technique might be rather atypical in data analysis and an incoherent strategy to extract meaning from a dataset. A differently-delimited field inside a delimited file is definitely an insane way to store data.

After much wrestling with the internal structure and a consideration of reloading the file using an entirely different library, I managed to use a standard Python pattern of counting values in a dictionary by forcing Pandas to access the string value by printing it. It might look sloppy, but at least I will have accessed the data.

```
In [27]: # NOTE TO EVALUATOR
         # Why does this loop work, but if you comment out the print statement it does not?
         # The error is: AttributeError: 'float' object has no attribute 'split'

         genres = imported_data['genres'].iteritems()
         genre_counts = {}

         for _, row in genres:
             for v in row.split('|'):
                 print(v)
                 if v not in genre_counts.keys():
                     genre_counts[v] = 1
                 else:
                     genre_counts[v] += 1
```

```
Action
Adventure
Science Fiction
Thriller
Action
Adventure
Science Fiction
Thriller
Adventure
Science Fiction
Thriller
Action
Adventure
Science Fiction
Fantasy
Action
Crime
Thriller
Western
Drama
Adventure
Thriller
Science Fiction
Action
Thriller
Adventure
Drama
Adventure
Science Fiction
Family
Animation
Adventure
Comedy
```

```
Comedy
Animation
Family
Action
Adventure
Crime
Science Fiction
Fantasy
Action
Adventure
Drama
Science Fiction
Action
Comedy
Science Fiction
Action
Adventure
Science Fiction
Crime
Drama
Mystery
Western
Crime
Action
Thriller
Science Fiction
Action
Adventure
Romance
Fantasy
Family
Drama
War
Adventure
Science Fiction
Action
Family
Science Fiction
Adventure
Mystery
Action
Drama
Action
Drama
Thriller
Drama
Romance
Comedy
```

```
Drama
Action
Comedy
Crime
Comedy
Action
Adventure
Drama
Thriller
History
Action
Science Fiction
Thriller
Mystery
Drama
Crime
Action
Science Fiction
Comedy
Music
Thriller
Drama
Adventure
Horror
Comedy
Drama
Thriller
Crime
Drama
Mystery
Adventure
Animation
Comedy
Family
Fantasy
Action
Crime
Drama
Mystery
Thriller
Drama
Romance
Drama
Music
Fantasy
Action
Adventure
History
```

```
Drama
Comedy
Action
Adventure
Fantasy
Drama
Romance
Action
Adventure
Science Fiction
Fantasy
Comedy
Animation
Science Fiction
Family
Drama
Mystery
Romance
Thriller
Crime
Drama
Thriller
Comedy
Drama
Romance
Science Fiction
Romance
Drama
Comedy
Adventure
Drama
Comedy
Drama
Action
Crime
Thriller
Drama
Science Fiction
Mystery
Thriller
Comedy
Adventure
Drama
Mystery
Crime
Action
Thriller
Drama
```

```
Action
Crime
Drama
Mystery
Thriller
Action
Adventure
Science Fiction
Mystery
Horror
Action
Comedy
Crime
Romance
Comedy
Crime
Drama
Action
Crime
Thriller
Thriller
Drama
Adventure
Action
History
Crime
Thriller
Action
Drama
Comedy
Drama
Thriller
War
Crime
Thriller
Thriller
Adventure
Family
Fantasy
Action
Adventure
Fantasy
Comedy
Drama
Adventure
Animation
Comedy
Family
```

```
Drama
Comedy
Drama
Horror
Thriller
Romance
Drama
Animation
Comedy
Family
Family
Comedy
Adventure
Drama
Thriller
Action
Crime
Drama
Adventure
Comedy
Horror
Thriller
Horror
Drama
Romance
Science Fiction
Crime
Thriller
Thriller
Mystery
Comedy
Fantasy
Action
Adventure
Thriller
Science Fiction
Action
Adventure
Adventure
Animation
Fantasy
Adventure
Animation
Comedy
Family
Drama
Romance
Comedy
```

```
Horror
Action
Adventure
Comedy
Family
Adventure
Animation
Family
Action
Drama
Science Fiction
Thriller
Thriller
Action
Comedy
Comedy
Comedy
Horror
Horror
Thriller
Crime
Drama
Crime
Action
Thriller
Horror
Comedy
Fantasy
Drama
Mystery
Thriller
Action
Thriller
Crime
Drama
Comedy
Comedy
Action
Drama
Action
Fantasy
Adventure
Comedy
Science Fiction
Thriller
Comedy
Science Fiction
Thriller
```

```
Action
Crime
Mystery
Thriller
Fantasy
Horror
Drama
Thriller
Action
Comedy
Drama
Music
Horror
Thriller
Romance
Thriller
Western
Drama
Crime
Drama
Mystery
Comedy
Family
Animation
Crime
Drama
Mystery
Adventure
Drama
Family
Family
Animation
Comedy
Adventure
Drama
Comedy
Drama
Action
Drama
Crime
Horror
Thriller
Action
Crime
Comedy
Drama
Horror
Thriller
```

```
Drama
Science Fiction
Thriller
Action
Adventure
Fantasy
Comedy
Drama
Drama
Science Fiction
Thriller
Adventure
Drama
Family
Animation
Comedy
Drama
Romance
Horror
Western
Adventure
Drama
Horror
Mystery
Thriller
Drama
Drama
Thriller
Action
Drama
Romance
Horror
Thriller
Horror
Thriller
Comedy
Crime
Action
Adventure
Romance
Fantasy
Horror
Mystery
Drama
Drama
History
Comedy
Drama
```

```
Music
Comedy
Drama
Romance
Action
Comedy
Science Fiction
Fantasy
Drama
Thriller
Crime
Drama
Mystery
Thriller
Drama
Comedy
Comedy
Drama
Action
Crime
Thriller
Comedy
Western
Comedy
Thriller
History
Drama
Drama
History
Drama
Drama
Music
Romance
Adventure
Action
Comedy
History
Drama
War
Action
Thriller
Comedy
Drama
Music
Action
Adventure
Comedy
Family
```

```
Music
Romance
Comedy
Romance
Drama
Drama
Drama
Comedy
Comedy
Drama
War
Action
Adventure
Animation
Family
Drama
Thriller
Adventure
Comedy
Romance
Romance
Comedy
Action
Animation
Science Fiction
Drama
Comedy
Drama
Science Fiction
Thriller
Drama
Animation
Comedy
Family
Adventure
Thriller
Thriller
Comedy
Western
Drama
Documentary
Music
Drama
Fantasy
Horror
Family
Animation
Horror
```

```
Romance
Drama
Romance
Comedy
Drama
Romance
Drama
Thriller
Comedy
Drama
Music
Comedy
Music
War
Drama
Comedy
Drama
Romance
Thriller
Horror
Drama
Comedy
Romance
Drama
Drama
Drama
Mystery
Science Fiction
Thriller
Animation
Comedy
Family
Fantasy
Music
Thriller
Science Fiction
Action
Drama
Thriller
Drama
Drama
Horror
Science Fiction
Mystery
Drama
Horror
Drama
Adventure
```

```
Comedy
Horror
Thriller
Horror
Action
Adventure
Animation
Thriller
Adventure
Documentary
Comedy
Fantasy
Thriller
Music
Documentary
War
Drama
Thriller
Crime
Drama
Drama
Drama
Horror
Mystery
Drama
Drama
Comedy
Animation
Adventure
Comedy
Thriller
Horror
Documentary
Horror
Thriller
Drama
Comedy
Horror
Thriller
History
Drama
Drama
Romance
Drama
Horror
TV Movie
Thriller
Horror
```

```
Action
Drama
Thriller
Crime
Drama
Drama
Thriller
Thriller
Horror
Fantasy
Comedy
Animation
Adventure
Science Fiction
Action
Action
Action
Adventure
Animation
Family
TV Movie
Action
Drama
Family
Drama
Horror
Thriller
Horror
Thriller
Drama
Documentary
Comedy
Comedy
Music
Drama
Crime
Comedy
Comedy
Action
Music
Animation
Family
Fantasy
Drama
Horror
Science Fiction
Comedy
Horror
```

```
Action
Animation
Fantasy
Drama
Music
Action
Drama
Thriller
Fantasy
Thriller
Horror
Science Fiction
Action
Science Fiction
Adventure
Fantasy
Thriller
Drama
Horror
Animation
Family
Drama
Thriller
Drama
Horror
Comedy
Thriller
Horror
Comedy
Drama
Comedy
Music
Thriller
Comedy
Crime
Family
Science Fiction
Action
Adventure
Drama
Music
Adventure
Fantasy
Thriller
Action
Drama
Action
Horror
```

```
Thriller
Documentary
Drama
War
Horror
Drama
Thriller
Mystery
Science Fiction
Family
TV Movie
Thriller
Horror
Thriller
Horror
Crime
Horror
Thriller
Thriller
Drama
Comedy
Drama
Thriller
Action
Drama
Fantasy
Action
Adventure
Drama
Comedy
Family
TV Movie
Animation
Thriller
Romance
Drama
Science Fiction
Action
Drama
Music
Thriller
Action
Horror
Adventure
Action
Thriller
Action
Drama
```

```
Horror
Horror
Romance
Science Fiction
Thriller
Horror
Animation
Family
Fantasy
Science Fiction
Action
Horror
Horror
Action
Thriller
Action
Horror
Comedy
Romance
Drama
Horror
Thriller
Thriller
Drama
Science Fiction
TV Movie
Adventure
Comedy
Horror
Thriller
Science Fiction
TV Movie
Action
Science Fiction
Drama
Science Fiction
Drama
Family
Thriller
Drama
Comedy
Horror
Drama
Thriller
Drama
Romance
Comedy
Crime
```

```
Drama
Thriller
Documentary
Documentary
Thriller
Drama
War
Action
Adventure
Action
Comedy
Science Fiction
Animation
Music
Family
Drama
Horror
Science Fiction
Thriller
Animation
Documentary
Comedy
Animation
Drama
Comedy
Drama
TV Movie
Crime
Mystery
Documentary
Drama
TV Movie
Thriller
Science Fiction
Action
Horror
Crime
Drama
Thriller
Science Fiction
Thriller
Fantasy
Action
Science Fiction
Comedy
Science Fiction
Comedy
Horror
```

```
Thriller
Horror
Action
Animation
Family
Action
Adventure
Science Fiction
Romance
Drama
Science Fiction
Horror
Action
Animation
Family
Drama
Comedy
Drama
Thriller
Drama
Drama
History
Romance
Drama
Music
TV Movie
Horror
Thriller
Horror
Drama
Family
Adventure
Adventure
Documentary
Animation
Romance
Comedy
Drama
Romance
Comedy
Mystery
Drama
Thriller
TV Movie
Horror
Science Fiction
Action
Thriller
```

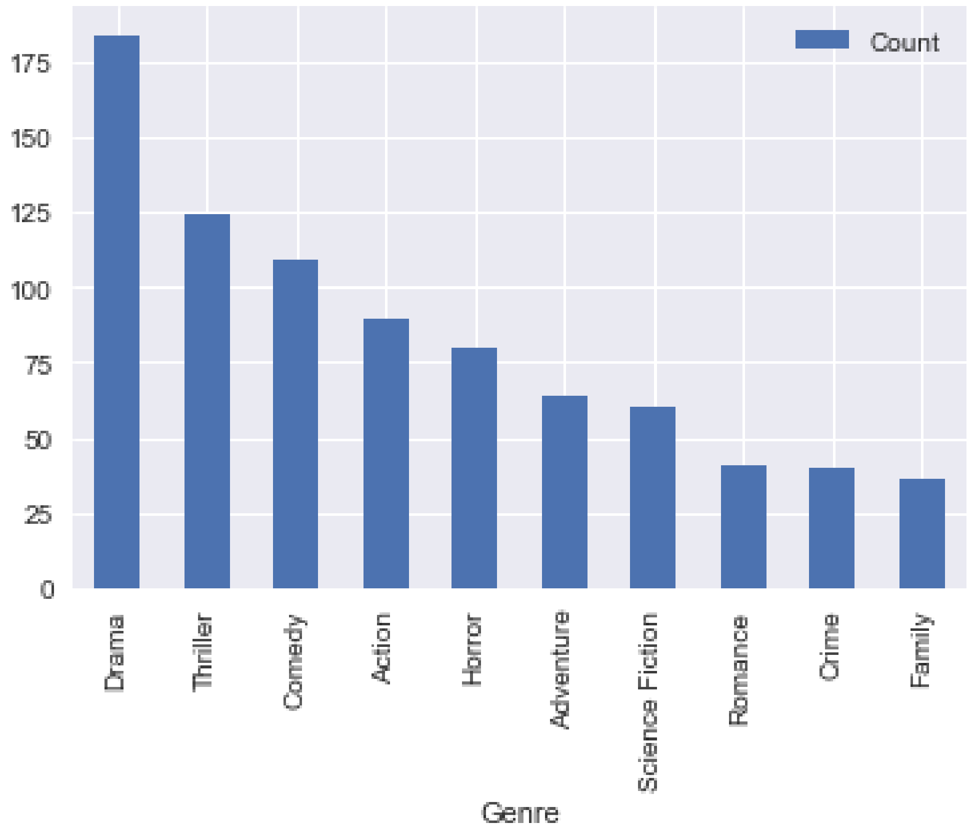
```
Documentary
Drama
Drama
Crime
Drama
Horror
Science Fiction
TV Movie
Drama
Comedy
Science Fiction
Action
Animation
Adventure
Romance
Drama
Horror
Drama
Drama
Horror
Thriller
Drama
Comedy
Drama
Thriller
Horror
Drama
Thriller
Mystery
Drama
Science Fiction
Thriller
Horror
Comedy
Horror
Romance
Comedy
Horror
Science Fiction
Thriller
History
Documentary
Family
Horror
Comedy
Romance
Horror
Horror
```

```
Thriller
Horror
Drama
Comedy
Science Fiction
Horror
Drama
Music
Comedy
Drama
Horror
Drama
Thriller
Comedy
Drama
Comedy
Horror
Horror
Thriller
Horror
Romance
Comedy
Romance
Drama
History
Crime
Drama
Horror
Thriller
Action
Drama
Thriller
Action
Drama
Thriller
Documentary
Romance
Drama
Comedy
Mystery
Fantasy
Drama
TV Movie
Comedy


        ---------------------------------------------------------------------------
```

```
    AttributeError                            Traceback (most recent call last)

    <ipython-input-27-32cb2bc3360b> in <module>()
      7
      8 for _, row in genres:
----> 9     for v in row.split('|'):
     10         print(v)
     11         if v not in genre_counts.keys():


    AttributeError: 'float' object has no attribute 'split'
```

Finally, let's make a bar chart to compare the relative frequencies of the genres, now that we've aggregated the totals from the 'genres' series.

```
In [28]: genre_counts_df = pd.DataFrame(list(genre_counts.items()), columns=['Genre', 'Count'])

         genre_counts_df.set_index('Genre').nlargest(10, 'Count').plot.bar()

Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x2242ebd76d8>
```



Interesting. It looks like the most movies are tagged as dramas. Not to overemphasize my own struggle, but that seems like an appropriate end to my own exploration. Certainly, I felt my struggle with making this simple chart a dramatic struggle worthy of a Greek playwright.