Investigate a Dataset - [TMDb movie]

September 4, 2022

1 Project: Investigate a Dataset - [TMDb movie]

1.1 Table of Contents

Introduction

This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue. Certain columns, like 'cast' and 'genres', contain multiple values separated by pipe (|) characters. There are some odd characters in the 'cast' column. The final two columns ending with "_adj" show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time.

what about using data analysis skills to know some intersting insights about movies, so let's start the analysis.

1.1.1 Question(s) for Analysis

1. what is the Average runtime movies from year to year?
2. Are there a correlation between popularity and vote_aveage?
3. what are the top 10 movies in popularity?
4. How did the amount of produced films changed over time?

In [2]:
```
# Use this cell to set up import statements for all of the packages that you plan to use
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

#changing numbers appearing format
#https://stackoverflow.com/questions/38689125/how-to-get-rid-of-pandas-converting-large-
pd.options.display.float_format = '{:.2f}'.format
```

Data Wrangling

1.1.2 General Properties

In [3]:
```
# Load your data and print out a few lines. Perform operations to inspect data
#   types and look for instances of missing or possibly errant data.
df = pd.read_csv('tmdb-movies.csv')
df.head()
```

Out[3]:
```
       id    imdb_id  popularity     budget     revenue  \
0  135397  tt0369610       32.99  150000000  1513528810
1   76341  tt1392190       28.42  150000000   378436354
2  262500  tt2908446       13.11  110000000   295238201
3  140607  tt2488496       11.17  200000000  2068178225
4  168259  tt2820852        9.34  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          6.50          2015 137999939.28 1392445892.52
1          7.10          2015 137999939.28  348161292.49
2          6.30          2015 101199955.47  271619025.41
3          7.50          2015 183999919.04 1902723129.80
4          7.30          2015 174799923.09 1385748801.47

[5 rows x 21 columns]
```

In [4]: df.shape

Out[4]: (10866, 21)

In [5]: df.describe()

Out[5]:

```
             id  popularity        budget        revenue  runtime  vote_count  \
count  10866.00    10866.00      10866.00       10866.00 10866.00    10866.00
mean   66064.18        0.65   14625701.09    39823319.79   102.07      217.39
std    92130.14        1.00   30913213.83   117003486.58    31.38      575.62
min        5.00        0.00          0.00           0.00     0.00       10.00
25%    10596.25        0.21          0.00           0.00    90.00       17.00
50%    20669.00        0.38          0.00           0.00    99.00       38.00
75%    75610.00        0.71   15000000.00    24000000.00   111.00      145.75
max   417859.00       32.99  425000000.00  2781505847.00   900.00     9767.00

       vote_average  release_year   budget_adj   revenue_adj
count      10866.00      10866.00     10866.00      10866.00
mean           5.97       2001.32  17551039.82   51364363.25
std            0.94         12.81  34306155.72  144632485.04
min            1.50       1960.00         0.00          0.00
25%            5.40       1995.00         0.00          0.00
50%            6.00       2006.00         0.00          0.00
75%            6.60       2011.00  20853251.08   33697095.72
max            9.20       2015.00 425000000.00 2827123750.41
```

In [6]:
```
df.info()
```

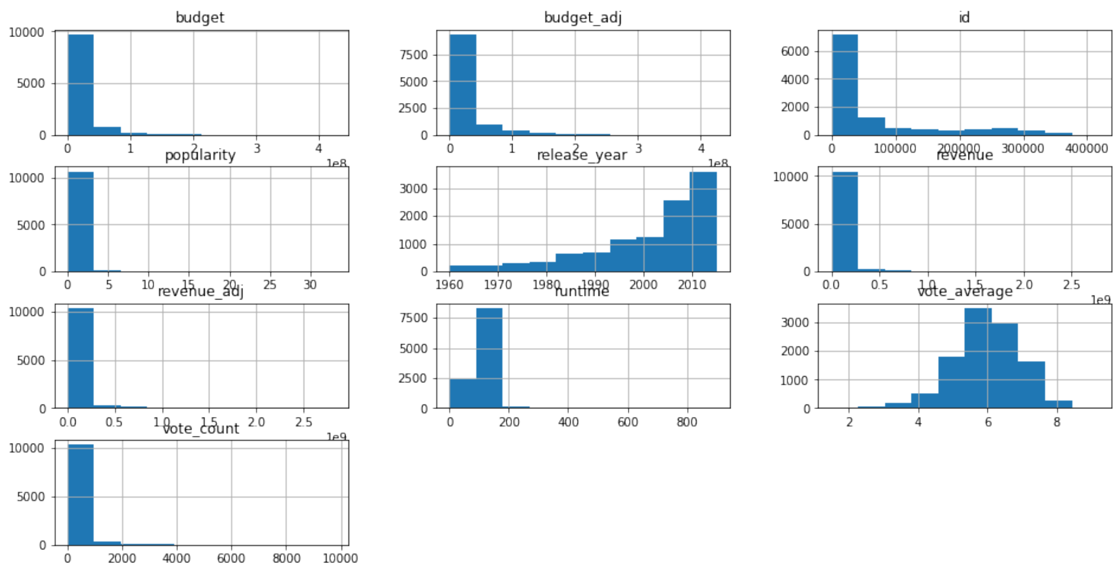
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
id                      10866 non-null int64
imdb_id                 10856 non-null object
popularity              10866 non-null float64
budget                  10866 non-null int64
revenue                 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
```

In [7]:
```
#calculating null values in each column
df.isna().sum()
```

Out[7]:
```
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 [8]: #take a general look at all columns of the data
        df.hist(figsize=(16,8));
```



1.1.3 Frist impression

there are rows contain several values, which are seperated by an "|",need to be cleaned . From the exploration above i found out that the data has null values in some columns and 0 values in others which kind a weird thing to have 0 values in such a column like how the run time for a movie equal zero or budget...etc,so we need to clean this up and drop unneeded columns.

1.1.4 Data Cleaning

```
In [9]: #replace 0 values with NAN
        #sorce:https://stackoverflow.com/questions/49575897/cant-replace-0-to-nan-in-python-usin
        df['revenue'].replace(0, np.NAN, inplace=True)
        df['revenue_adj'].replace(0, np.NAN, inplace=True)
        df['budget'].replace(0, np.NAN, inplace=True)
        df['budget_adj'].replace(0, np.NAN, inplace=True)
        df['runtime'].replace(0, np.NAN, inplace=True)

        df.dropna(axis=0, inplace=True)
```

In [10]:
```python
#Seperating columns that have several values
#source: https://apassionatechie.wordpress.com/2018/02/24/how-do-i-split-a-string-into-

df_cast = (df['cast'].str.split('|', expand=True).rename(columns=lambda x: f"cast_{x+1}
df_director = (df['director'].str.split('|', expand=True).rename(columns=lambda x: f"di
df_genres = (df['genres'].str.split('|', expand=True).rename(columns=lambda x: f"genres
df_keywords = (df['keywords'].str.split('|', expand=True).rename(columns=lambda x: f"ke
df_prod = (df['production_companies'].str.split('|', expand=True).rename(columns=lambda

df_cast.head()
```

Out[10]:
```
              cast_1               cast_2            cast_3  \
0        Chris Pratt  Bryce Dallas Howard       Irrfan Khan
1          Tom Hardy      Charlize Theron  Hugh Keays-Byrne
2   Shailene Woodley           Theo James      Kate Winslet
3      Harrison Ford          Mark Hamill     Carrie Fisher
4         Vin Diesel          Paul Walker     Jason Statham

               cast_4          cast_5
0   Vincent D'Onofrio   Nick Robinson
1      Nicholas Hoult     Josh Helman
2        Ansel Elgort    Miles Teller
3         Adam Driver    Daisy Ridley
4  Michelle Rodriguez  Dwayne Johnson
```

In [11]:
```python
#Join the seperated columns and drop unneeded columns

df = df.join([df_cast, df_director,df_genres,df_keywords, df_prod])
df = df.drop(['cast', 'director', 'keywords', 'production_companies', 'imdb_id', 'homep
```

In [12]:
```python
#checking for duplicates
df.duplicated().sum()
```

Out[12]: 0

In [13]:
```python
#chicking the data type if it appropriate or not
df.dtypes
```

Out[13]:
```
id                    int64
popularity          float64
budget              float64
revenue             float64
original_title       object
runtime             float64
genres               object
vote_count            int64
vote_average        float64
release_year          int64
budget_adj          float64
```

```
revenue_adj           float64
cast_1                 object
cast_2                 object
cast_3                 object
cast_4                 object
cast_5                 object
director_1             object
director_2             object
director_3             object
director_4             object
director_5             object
director_6             object
genres_1               object
genres_2               object
genres_3               object
genres_4               object
genres_5               object
keywords_1             object
keywords_2             object
keywords_3             object
keywords_4             object
keywords_5             object
production_comp_1      object
production_comp_2      object
production_comp_3      object
production_comp_4      object
production_comp_5      object
dtype: object
```

In [14]:
```
df
```

Out[14]:

	id	popularity	budget	revenue	\
0	135397	32.99	150000000.00	1513528810.00	
1	76341	28.42	150000000.00	378436354.00	
2	262500	13.11	110000000.00	295238201.00	
3	140607	11.17	200000000.00	2068178225.00	
4	168259	9.34	190000000.00	1506249360.00	
5	281957	9.11	135000000.00	532950503.00	
6	87101	8.65	155000000.00	440603537.00	
7	286217	7.67	108000000.00	595380321.00	
8	211672	7.40	74000000.00	1156730962.00	
9	150540	6.33	175000000.00	853708609.00	
10	206647	6.20	245000000.00	880674609.00	
11	76757	6.19	176000003.00	183987723.00	
12	264660	6.12	15000000.00	36869414.00	
13	257344	5.98	88000000.00	243637091.00	
14	99861	5.94	280000000.00	1405035767.00	
15	273248	5.90	44000000.00	155760117.00	

```
16     260346        5.75  48000000.00  325771424.00
17     102899        5.57 130000000.00  518602163.00
19     131634        5.48 160000000.00  650523427.00
20     158852        5.46 190000000.00  209035668.00
22     254128        4.91 110000000.00  470490832.00
23     216015        4.71  40000000.00  569651467.00
24     318846        4.65  28000000.00  133346506.00
25     177677        4.57 150000000.00  682330139.00
27     207703        4.50  81000000.00  403802136.00
28     314365        4.06  20000000.00   88346473.00
29     294254        3.97  61000000.00  311256926.00
31     198184        3.90  49000000.00  102069268.00
34     257445        3.64  58000000.00  150170815.00
35     264644        3.56   6000000.00   35401758.00
...       ...         ...          ...           ...
9807      762        2.23    400000.00    5028948.00
9808    36685        1.41   1200000.00  112892319.00
9849      646        3.17   1100000.00   59600000.00
9881      658        3.15   3500000.00  124900000.00
9884      704        0.81    500000.00    1000549.00
9925      681        1.76   7200000.00  116000000.00
9927      984        0.89   4000000.00   35976000.00
9932      636        0.63    777000.00    2437000.00
9951    25188        0.37   1300000.00   29133000.00
9981      196        2.44  40000000.00  244527583.00
9984      242        1.76  54000000.00  136766062.00
9992     1669        1.07        30.00        200.00
10094     771        0.14  18000000.00  476684675.00
10131     430        0.30   3000000.00    4000000.00
10222     424        2.38  22000000.00  321265768.00
10223     329        2.20  63000000.00  920100000.00
10224    9739        1.96  57000000.00  159055768.00
10251   10057        0.79  30000000.00   53898845.00
10255   10909        0.76   9000000.00    2395231.00
10317    2149        0.39  30000000.00   13273595.00
10338    8291        0.31  14000000.00   27515786.00
10401     667        1.55   9500000.00  111584787.00
10438     657        2.51   2500000.00   78898765.00
10489    6978        0.96  25000000.00   11000000.00
10594    9552        2.01   8000000.00  441306145.00
10595     253        1.55   7000000.00  161777836.00
10689     660        1.91  11000000.00  141195658.00
10724     668        1.78   7000000.00   81974493.00
10759     948        1.20    300000.00   70000000.00
10760    8469        1.16   2700000.00  141000000.00

                              original_title  runtime  \
0                             Jurassic World   124.00
```

1	Mad Max: Fury Road	120.00
2	Insurgent	119.00
3	Star Wars: The Force Awakens	136.00
4	Furious 7	137.00
5	The Revenant	156.00
6	Terminator Genisys	125.00
7	The Martian	141.00
8	Minions	91.00
9	Inside Out	94.00
10	Spectre	148.00
11	Jupiter Ascending	124.00
12	Ex Machina	108.00
13	Pixels	105.00
14	Avengers: Age of Ultron	141.00
15	The Hateful Eight	167.00
16	Taken 3	109.00
17	Ant-Man	115.00
19	The Hunger Games: Mockingjay - Part 2	136.00
20	Tomorrowland	130.00
22	San Andreas	114.00
23	Fifty Shades of Grey	125.00
24	The Big Short	130.00
25	Mission: Impossible - Rogue Nation	131.00
27	Kingsman: The Secret Service	130.00
28	Spotlight	128.00
29	Maze Runner: The Scorch Trials	132.00
31	Chappie	120.00
34	Goosebumps	103.00
35	Room	117.00
...	...	...
9807	Monty Python and the Holy Grail	91.00
9808	The Rocky Horror Picture Show	100.00
9849	Dr. No	110.00
9881	Goldfinger	110.00
9884	A Hard Day's Night	88.00
9925	Diamonds Are Forever	120.00
9927	Dirty Harry	102.00
9932	THX 1138	86.00
9951	The Last Picture Show	118.00
9981	Back to the Future Part III	118.00
9984	The Godfather: Part III	162.00
9992	The Hunt for Red October	134.00
10094	Home Alone	103.00
10131	One, Two, Three	115.00
10222	Schindler's List	195.00
10223	Jurassic Park	127.00
10224	Demolition Man	115.00
10251	The Three Musketeers	105.00

```
10255                      Kalifornia   117.00
10317                Body of Evidence    99.00
10338                  Poetic Justice   109.00
10401             You Only Live Twice   117.00
10438           From Russia With Love   115.00
10489     Big Trouble in Little China    99.00
10594                     The Exorcist   122.00
10595                 Live and Let Die   121.00
10689                      Thunderball   130.00
10724   On Her Majesty's Secret Service  142.00
10759                       Halloween    91.00
10760                    Animal House   109.00

                                              genres  vote_count  \
0             Action|Adventure|Science Fiction|Thriller        5562
1             Action|Adventure|Science Fiction|Thriller        6185
2                    Adventure|Science Fiction|Thriller        2480
3              Action|Adventure|Science Fiction|Fantasy        5292
4                                 Action|Crime|Thriller        2947
5                     Western|Drama|Adventure|Thriller        3929
6             Science Fiction|Action|Thriller|Adventure        2598
7                      Drama|Adventure|Science Fiction        4572
8                     Family|Animation|Adventure|Comedy        2893
9                               Comedy|Animation|Family        3935
10                               Action|Adventure|Crime        3254
11             Science Fiction|Fantasy|Action|Adventure        1937
12                                Drama|Science Fiction        2854
13                        Action|Comedy|Science Fiction        1575
14                     Action|Adventure|Science Fiction        4304
15                         Crime|Drama|Mystery|Western        2389
16                                Crime|Action|Thriller        1578
17                     Science Fiction|Action|Adventure        3779
19                        War|Adventure|Science Fiction        2380
20      Action|Family|Science Fiction|Adventure|Mystery        1899
22                                Action|Drama|Thriller        2060
23                                        Drama|Romance        1865
24                                         Comedy|Drama        1545
25                                               Action        2349
27                         Crime|Comedy|Action|Adventure        3833
28                               Drama|Thriller|History        1559
29                      Action|Science Fiction|Thriller        1849
31                         Crime|Action|Science Fiction        1990
34                              Adventure|Horror|Comedy         600
35                                       Drama|Thriller        1520
...                                                 ...         ...
9807                            Adventure|Comedy|Fantasy        1097
9808                  Comedy|Horror|Music|Science Fiction         332
9849                           Adventure|Action|Thriller         560
```

```
9881                 Adventure|Action|Thriller         602
9884                              Comedy|Music          92
9925      Adventure|Action|Thriller|Science Fiction    331
9927                     Action|Crime|Thriller         300
9932         Drama|Mystery|Science Fiction|Thriller    125
9951                                     Drama          42
9981   Adventure|Action|Comedy|Science Fiction|Family 1762
9984                Drama|Action|Thriller|Crime        880
9992                 Action|Adventure|Thriller         615
10094                            Comedy|Family        1393
10131                            Comedy|Family          30
10222                        Drama|History|War        2632
10223                 Adventure|Science Fiction       3169
10224       Action|Adventure|Comedy|Science Fiction    580
10251                   Action|Adventure|Comedy        112
10255                           Thriller|Crime          96
10317                   Drama|Thriller|Romance          25
10338                            Drama|Romance          24
10401                 Action|Thriller|Adventure        301
10438                 Action|Thriller|Adventure        458
10489           Adventure|Fantasy|Action|Comedy        347
10594                    Drama|Horror|Thriller        1113
10595                 Adventure|Action|Thriller        293
10689                 Adventure|Action|Thriller        331
10724                 Adventure|Action|Thriller        258
10759                          Horror|Thriller         522
10760                                   Comedy         230

       vote_average  release_year             ...              \
0              6.50          2015             ...
1              7.10          2015             ...
2              6.30          2015             ...
3              7.50          2015             ...
4              7.30          2015             ...
5              7.20          2015             ...
6              5.80          2015             ...
7              7.60          2015             ...
8              6.50          2015             ...
9              8.00          2015             ...
10             6.20          2015             ...
11             5.20          2015             ...
12             7.60          2015             ...
13             5.80          2015             ...
14             7.40          2015             ...
15             7.40          2015             ...
16             6.10          2015             ...
17             7.00          2015             ...
19             6.50          2015             ...
```

20	6.20	2015	...
22	6.10	2015	...
23	5.30	2015	...
24	7.30	2015	...
25	7.10	2015	...
27	7.60	2015	...
28	7.80	2015	...
29	6.40	2015	...
31	6.60	2015	...
34	6.20	2015	...
35	8.00	2015	...
...	...	...	...
9807	7.60	1975	...
9808	7.10	1975	...
9849	6.70	1962	...
9881	7.00	1964	...
9884	6.90	1964	...
9925	6.20	1971	...
9927	7.20	1971	...
9932	6.10	1971	...
9951	7.00	1971	...
9981	6.90	1990	...
9984	6.90	1990	...
9992	6.90	1990	...
10094	7.00	1990	...
10131	7.50	1961	...
10222	8.10	1993	...
10223	7.40	1993	...
10224	6.10	1993	...
10251	5.90	1993	...
10255	6.30	1993	...
10317	4.40	1993	...
10338	6.80	1993	...
10401	6.20	1967	...
10438	6.70	1963	...
10489	6.70	1986	...
10594	7.20	1973	...
10595	6.10	1973	...
10689	6.30	1965	...
10724	6.40	1969	...
10759	7.30	1978	...
10760	6.70	1978	...

	keywords_1	keywords_2	\
0	monster	dna	
1	future	chase	
2	based on novel	revolution	
3	android	spaceship	

4	car race	speed
5	father-son relationship	rape
6	saving the world	artificial intelligence
7	based on novel	mars
8	assistant	aftercreditsstinger
9	dream	cartoon
10	spy	based on novel
11	jupiter	space
12	dancing	artificial intelligence
13	video game	nerd
14	marvel comic	comic
15	bounty hunter	wyoming
16	revenge	murder
17	marvel comic	superhero
19	revolution	strong woman
20	inventor	apocalypse
22	california	earthquake
23	based on novel	billionaire
24	bank	fraud
25	spy	sequel
27	spy	great britain
28	child abuse	journalism
29	based on novel	resistance
31	artificial intelligence	android
34	based on novel	magic
35	based on novel	carpet
...	...	...
9807	holy grail	monk
9808	transvestism	transylvania
9849	london	england
9881	secret organization	secret intelligence service
9884	adolescence	culture clash
9925	satellite	plastic surgery
9927	ambush	san francisco
9932	prison	drug addiction
9951	new love	graduation
9981	jules verne	railroad robber
9984	italy	christianity
9992	submarine	cold war
10094	holiday	burglar
10131	berlin	prison
10222	factory	concentration camp
10223	exotic island	dna
10224	helicopter	martial arts
10251	paris	musketeer
10255	california	journalist
10317	sex	infidelity
10338	loss of lover	sadness

10401	london	japan
10438	venice	london
10489	kung fu	chinatown
10594	exorcism	holy water
10595	london	new york
10689	paris	florida
10724	london	suicide
10759	female nudity	nudity
10760	female nudity	sex

	keywords_3	keywords_4 \
0	tyrannosaurus rex	velociraptor
1	post-apocalyptic	dystopia
2	dystopia	sequel
3	jedi	space opera
4	revenge	suspense
5	based on novel	mountains
6	cyborg	killer robot
7	nasa	isolation
8	duringcreditsstinger	evil mastermind
9	imaginary friend	animation
10	secret agent	sequel
11	woman director	3d
12	helicopter	distrust
13	alien attack	3d
14	sequel	superhero
15	mountains	hangman
16	on the run	fugitive
17	aftercreditsstinger	duringcreditsstinger
19	dystopia	game of death
20	destiny	imax
22	catastrophe	disaster film
23	bdsm	woman director
24	biography	wall street
25	mission	None
27	secret organization	secret agent
28	judge	florida
29	maze	post-apocalyptic
31	robot	near future
34	fantasy	family
35	isolation	imprisonment
...	...	...
9807	scotland yard	swordplay
9808	marriage proposal	time warp
9849	assassination	spy
9881	nuclear radiation	fort knox
9884	press conference	behind the scenes
9925	smuggling	murder

```
9927                   detective                ransom
9932                     hearing   totalitarian regime
9951      high school graduation             pool hall
9981                  california              car race
9984                    new york         assassination
9992                     russian             defection
10094              home invasion              mischief
10131                      clerk               atlanta
10222                       hero             holocaust
10223               paleontology    tyrannosaurus rex
10224              crime fighter        social control
10251                       None                  None
10255                 journalism          photographer
10317                  eroticism                nudity
10338                los angeles            road movie
10401                    england         assassination
10438                     terror               england
10489                      magic                  None
10594  religion and supernatural                 vomit
10595                       bomb               england
10689              fighter pilot            sanatorium
10724                    england           switzerland
10759                       mask            babysitter
10760                     nudity               collage

                      keywords_5                  production_comp_1  \
0                         island                  Universal Studios
1                      australia          Village Roadshow Pictures
2                dystopic future               Summit Entertainment
3                             3d                          Lucasfilm
4                            car                Universal Pictures
5                         winter                Regency Enterprises
6                         future                 Paramount Pictures
7                       botanist  Twentieth Century Fox Film Corporation
8                        minions                 Universal Pictures
9                            kid               Walt Disney Pictures
10                    james bond                  Columbia Pictures
11          interspecies romance          Village Roadshow Pictures
12                     isolation                          DNA Films
13                        pixels                  Columbia Pictures
14                        vision                     Marvel Studios
15          voice over narration               Double Feature Films
16                        framed  Twentieth Century Fox Film Corporation
17     marvel cinematic universe                     Marvel Studios
19                            3d                  Studio Babelsberg
20                       dreamer               Walt Disney Pictures
22                            3d                    New Line Cinema
23                  erotic movie                     Focus Features
```

```
24                  finances                     Paramount Pictures
25                      None                     Paramount Pictures
27              marvel comic  Twentieth Century Fox Film Corporation
28                    boston                      Participant Media
29                  dystopia                           Gotham Group
31                 robot cop                      Columbia Pictures
34                        3d                      Columbia Pictures
35              grandparents                       Element Pictures
...                      ...                                    ...
9807                 camelot        Python (Monty) Pictures Limited
9808                  castle                       20th Century Fox
9849                  casino                        Eon Productions
9881             aston martin                       Eon Productions
9884            police chase                       Proscenium Films
9925               extortion                        Eon Productions
9927                 stadium                           Warner Bros.
9932                 phasing                       American Zoetrope
9951       graduation present           Columbia Pictures Corporation
9981                delorean                     Universal Pictures
9984           italo-american                     Paramount Pictures
9992               jack ryan                     Paramount Pictures
10094             booby trap  Twentieth Century Fox Film Corporation
10131               cold war                The Mirisch Corporation
10222           world war ii                     Universal Pictures
10223            triceratops                     Universal Pictures
10224                 museum                        Silver Pictures
10251                   None                   Walt Disney Pictures
10255                highway                       Propaganda Films
10317              seduction              Metro-Goldwyn-Mayer (MGM)
10338                   None                      Columbia Pictures
10401             helicopter                        Eon Productions
10438          assassination                        Eon Productions
10489                   None  Twentieth Century Fox Film Corporation
10594              christian                           Warner Bros.
10595                    spy                        Eon Productions
10689    secret organization                        Eon Productions
10724        secret identity                        Eon Productions
10759              halloween        Compass International Pictures
10760             fraternity                     Universal Pictures

                   production_comp_2  \
0               Amblin Entertainment
1          Kennedy Miller Productions
2                    Mandeville Films
3               Truenorth Productions
4                       Original Film
5                          Appian Way
6                Skydance Productions
```

```
7                         Scott Free Productions
8                     Illumination Entertainment
9                        Pixar Animation Studios
10                                        Danjaq
11                           Dune Entertainment
12       Universal Pictures International (UPI)
13                     Happy Madison Productions
14                                   Prime Focus
15                       The Weinstein Company
16                                     M6 Films
17                                         None
19                                   StudioCanal
20                                       Babieka
22                     Village Roadshow Pictures
23                    Trigger Street Productions
24                         Plan B Entertainment
25                         Skydance Productions
27                                    Marv Films
28                               Open Road Films
29                     Temple Hill Entertainment
31                         Media Rights Capital
34                                 Original Film
35                              No Trace Camping
...                                          ...
9807                   Michael White Productions
9808                                        None
9849                   Metro-Goldwyn-Mayer (MGM)
9881                   Metro-Goldwyn-Mayer (MGM)
9884                        Walter Shenson Films
9925                   Metro-Goldwyn-Mayer (MGM)
9927                             Malpaso Company
9932                                Warner Bros.
9951                             BBS Productions
9981                        Amblin Entertainment
9984                                        None
9992                      Nina Saxon Film Design
10094                       Hughes Entertainment
10131                                       None
10222                       Amblin Entertainment
10223                       Amblin Entertainment
10224                               Warner Bros.
10251                           Caravan Pictures
10255                   Kouf/Bigelow Productions
10317                                       None
10338                                       None
10401                                       None
10438                  Metro-Goldwyn-Mayer (MGM)
10489                 TAFT Entertainment Pictures
```

```
10594                    Hoya Productions
10595          Metro-Goldwyn-Mayer (MGM)
10689          Metro-Goldwyn-Mayer (MGM)
10724          Metro-Goldwyn-Mayer (MGM)
10759      Falcon International Productions
10760                  Oregon Film Factory

                      production_comp_3  \
0                    Legendary Pictures
1                                  None
2               Red Wagon Entertainment
3                             Bad Robot
4                  Media Rights Capital
5                             CatchPlay
6                                  None
7                    Mid Atlantic Films
8                                  None
9      Walt Disney Studios Motion Pictures
10                                  B24
11                 Anarchos Productions
12                                Film4
13                                 None
14               Revolution Sun Studios
15                           FilmColony
16                              Canal+
17                                 None
19                            Lionsgate
20                                 A113
22                         Warner Bros.
23            Michael De Luca Productions
24                  Regency Enterprises
25                  China Movie Channel
27                    TSG Entertainment
28                    Anonymous Content
29                    TSG Entertainment
31      Sony Pictures Entertainment (SPE)
34              Scholastic Entertainment
35                                  A24
...                                 ...
9807       National Film Trustee Company
9808                               None
9849                               None
9881                               None
9884                 Maljack Productions
9925                             Danjaq
9927                               None
9932                               None
9951                               None
```

9981	U-Drive Productions
9984	None
9992	Mace Neufeld Productions
10094	None
10131	None
10222	None
10223	None
10224	None
10251	None
10255	None
10317	None
10338	None
10401	None
10438	Danjaq
10489	None
10594	None
10595	None
10689	None
10724	Danjaq
10759	None
10760	Stage III Productions

	production_comp_4	production_comp_5
0	Fuji Television Network	Dentsu
1	None	None
2	NeoReel	None
3	None	None
4	Dentsu	One Race Films
5	Anonymous Content	New Regency Pictures
6	None	None
7	International Traders	TSG Entertainment
8	None	None
9	None	None
10	None	None
11	Warner Bros.	None
12	None	None
13	None	None
14	None	None
15	None	None
16	EuropaCorp	CinÃľ+
17	None	None
19	Walt Disney Studios Motion Pictures	Color Force
20	None	None
22	Flynn Picture Company	None
23	None	None
24	None	None
25	Bad Robot	TC Productions
27	Cloudy Productions	None

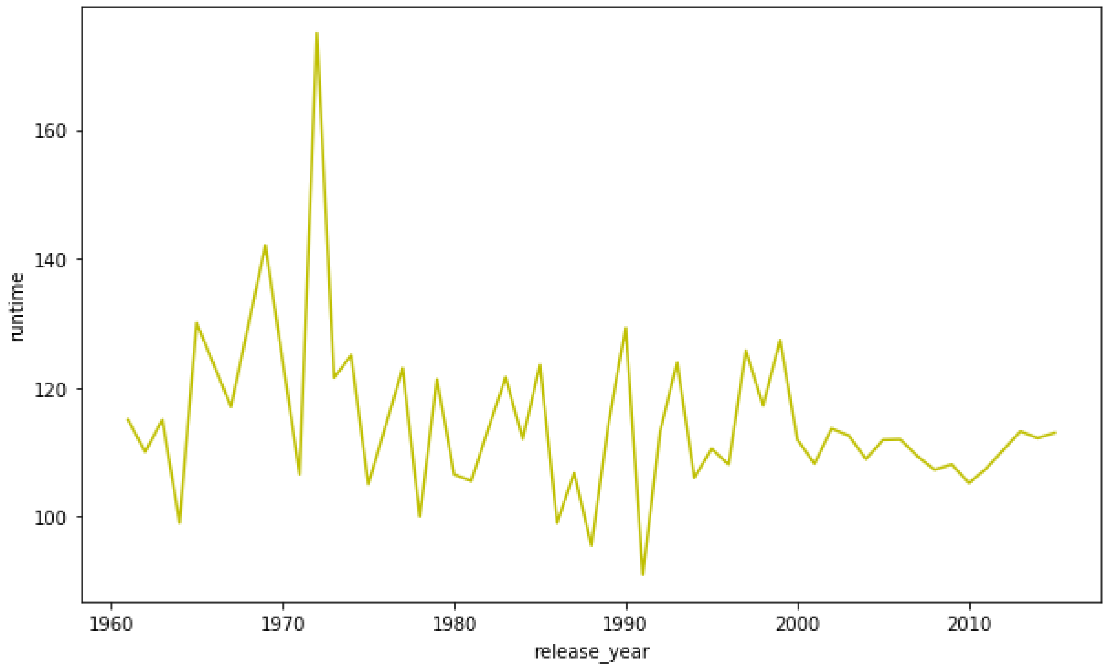
```
28                  Rocklin / Faust  Entertainment One Features
29                             None                        None
31                       Alpha Core                 Genre Films
34                             None                        None
35                   Duperele Films                        None
...                             ...                         ...
9807       Twickenham Film Studios                        None
9808                           None                        None
9849                           None                        None
9881                           None                        None
9884                           None                        None
9925                           None                        None
9927                           None                        None
9932                           None                        None
9951                           None                        None
9981                           None                        None
9984                           None                        None
9992                           None                        None
10094                          None                        None
10131                          None                        None
10222                          None                        None
10223                          None                        None
10224                          None                        None
10251                          None                        None
10255                          None                        None
10317                          None                        None
10338                          None                        None
10401                          None                        None
10438                          None                        None
10489                          None                        None
10594                          None                        None
10595                          None                        None
10689                          None                        None
10724                          None                        None
10759                          None                        None
10760                          None                        None

[1287 rows x 38 columns]
```

the data looks prepared for the analysis, no duplicates have been found, no nanull values, columns containing multiple values have been seperated and data types look ready for analysis. ## Exploratory Data Analysis

1.1.5 Research Question 1 (Average runtime movies from year to year?)

In [15]:
```
#creating a plot of the mean of rlease year vs runtime
df.groupby('release_year').mean()['runtime'].plot(figsize=(10,6),color='y');
plt.xlabel('release_year')
```

```
plt.ylabel('runtime');
min_av=df.groupby('release_year').mean()['runtime'].min()
max_av=df.groupby('release_year').mean()['runtime'].max()
```



the chart indicate that the average runtime movies is about 130.00 m ,the minimum Average runtime movies is 91.0,but maximum is 175.0.

```
In [16]: #calcuating the min and max run time
         min_av=df.groupby('release_year').mean()['runtime'].min()
         max_av=df.groupby('release_year').mean()['runtime'].max()
         min_av,max_av
```

```
Out[16]: (91.0, 175.0)
```

1.1.6 Research Question 2 (Are there a correlation between popularity and vote_aveage?)

```
In [17]: pop_vote=df[['popularity','vote_average']]
         pop_vote.corr()
```

Out[17]:

	popularity	vote_average
popularity	1.00	0.36
vote_average	0.36	1.00

1.1.7 Research Question 3 (what are the top 10 movies in popularity?)

```
In [24]: index = pd.Index(range(1, 11, 1))#setting index of movies order in the list
         top_df=df[['original_title','popularity','genres','release_year']]
```

```
top_ten=top_df.nlargest(n=10,columns=['popularity']).set_index(index)
top_ten
```

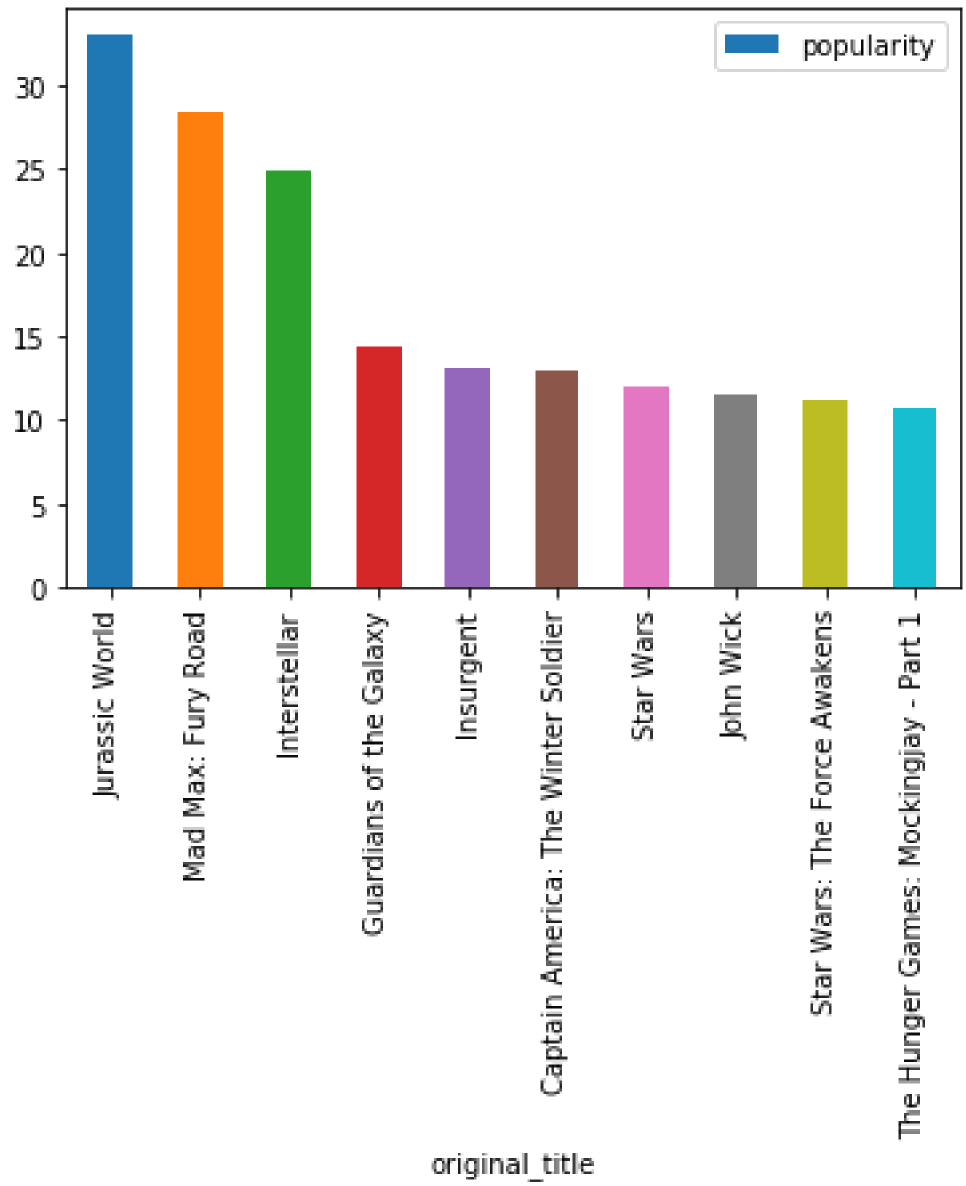
Out[24]:

```
                          original_title  popularity  \
1                         Jurassic World       32.99
2                     Mad Max: Fury Road       28.42
3                           Interstellar       24.95
4                Guardians of the Galaxy       14.31
5                              Insurgent       13.11
6    Captain America: The Winter Soldier       12.97
7                              Star Wars       12.04
8                              John Wick       11.42
9           Star Wars: The Force Awakens       11.17
10  The Hunger Games: Mockingjay - Part 1       10.74

                                        genres  release_year
1   Action|Adventure|Science Fiction|Thriller          2015
2   Action|Adventure|Science Fiction|Thriller          2015
3            Adventure|Drama|Science Fiction          2014
4           Action|Science Fiction|Adventure          2014
5         Adventure|Science Fiction|Thriller          2015
6           Action|Adventure|Science Fiction          2014
7           Adventure|Action|Science Fiction          1977
8                            Action|Thriller          2014
9    Action|Adventure|Science Fiction|Fantasy          2015
10        Science Fiction|Adventure|Thriller          2014
```

In [23]:

```
#creating a bar plot for top movies popularity
top_ten.plot(x='original_title',y='popularity',kind='bar')
```

Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0e7c1049b0>



nice! this bar chart indicates the top 10 movies in popularity, Jurassic World in the lead...

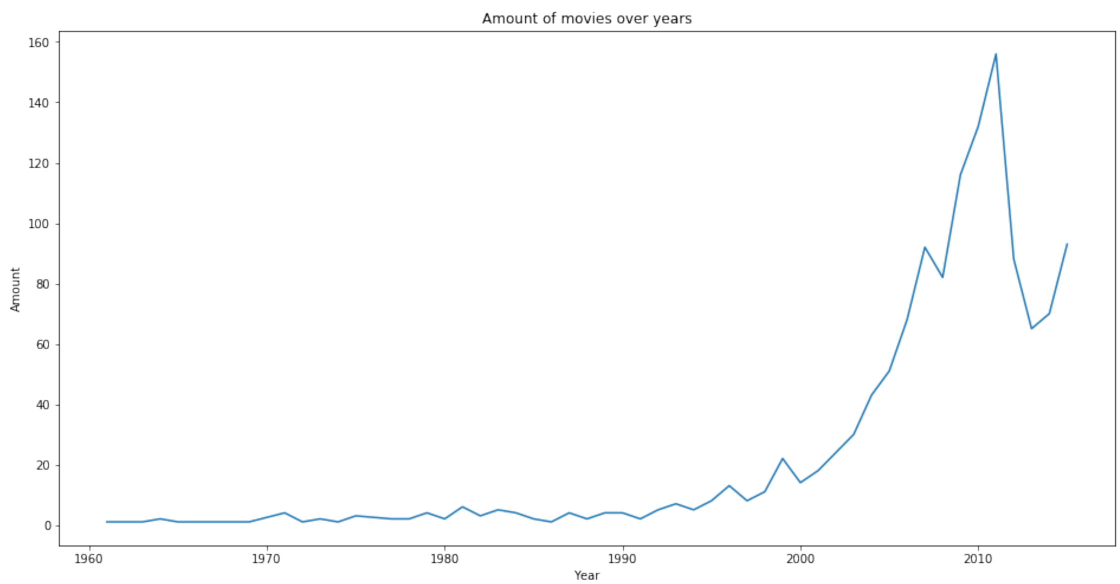
1.1.8 Research Question 4 (How did the amount of produced films changed over time?)

In [25]:
```
#calculating the number of movies in last 10 years
movie_year=df.groupby('release_year').count()['id']
movie_year.tail(10)
```

Out[25]:
```
release_year
2006     68
2007     92
2008     82
```

```
2009    116
2010    132
2011    156
2012     88
2013     65
2014     70
2015     93
Name: id, dtype: int64
```

In [26]:
```python
#creating a line chart shows the amonut of movies per year
movie_year.plot(figsize=(16,8),title='Amount of movies over years')
plt.xlabel('Year')
plt.ylabel('Amount');
```



form the plot we can see that the amount of movies has increased significantly from 1998 to 2015 and reached its peak in 2011.

Conclusions **Results:**

The first research question "What is the Average runtime movies from year to year?" indicate that the average runtime movies is about 130.00 m ,the minimum and maximum Average runtime movies are respectively(91.0, 175.0).

The second research question "Are there a correlation between popularity and vote_aveage?" indicate that there is no strong corralation between them.

The third research question "What are the top 10 movies in popularity?" indicate that jurassic world is the most popular produced movie followed by Mad Max: Fury Road and Interstellar, movies genres concentarted in Action|Adventure|Science Fiction|Thriller, notcing that most of them produced lately.

The forth research question "How did the amount of produced films changed over time?" reveals that the amount of produced films significantly increased from 1998 to 2015, and reached its peak in 2011, this can be an idicator for the huge developement in cinema in the last decade and we expect it to increase more and more now respectively with the increase in audience,and movies platforms now adays.

limitations: * Most of our varialbes are categorical, which does not allow for a high level of statistical method that can be used to provide correlatios etc. * data outcomes cann't be generalised because some entries in the dataset have been removed due to missing data ,but can be treated as indicators. * considering that many inputs in our data have been removed due to missing data. * we can add more recent data to this data to have better insights

In [142]:
```
from subprocess import call
call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

Out[142]: 0