Introduction

In this project, we are going to investigate the Movie Database (TMDb), which contains information about 10,000 movies regarding to their title, release date, user ratings, as well as their budget and revenue.

Our data analysis process for this project will provide a step by step guidance, starting by asking a series of questions, then wrangling and exploring the dataset, and finally drawing some conclusions as well as communicating the findings.

Research Questions:

1. What are the top ten most profitables movies?
2. Which movie has the highest and the lowest budget?
3. Which movie has the highest and the lowest revenue?
4. In which year did movies' industry realize their most profit?
5. What's the relationship between both popularity and runtime of a movie against their profit?
6. Which movie's genre has the highest release?
7. Who are the most succesful directors?
8. What's the most frequent cast?

In [3]:
```
# import all necessary core packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Data Wrangling

The second step of our analysis is Data Wrangling, which includes assessing the TMDb Movies dataset both visually and programmatically, indentifying the presence of any tidiness issues and then improving its quality which will help us later on to analyze our data and draw conclusions.

General Properties

In [4]:
```
# Load dataset
df = pd.read_csv('tmdb-movies.csv')
df.head(50)
```

Out[4]:

	id	imdb_id	popularity	budget	revenue	original_title	cast
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...
5	281957	tt1663202	9.110700	135000000	532950503	The Revenant	Leonardo DiCaprio|Tom Hardy|Will Poulter|Domhn...
6	87101	tt1340138	8.654359	155000000	440603537	Terminator Genisys	Arnold Schwarzenegger|Jason Clarke|Emilia Clar...
7	286217	tt3659388	7.667400	108000000	595380321	The Martian	Matt Damon|Jessica Chastain|Kristen Wiig|Jeff ...
8	211672	tt2293640	7.404165	74000000	1156730962	Minions	Sandra Bullock|Jon Hamm|Michael Keaton|Allison...
9	150540	tt2096673	6.326804	175000000	853708609	Inside Out	Amy Poehler|Phyllis Smith|Richard Kind|Bill Ha...
10	206647	tt2379713	6.200282	245000000	880674609	Spectre	Daniel Craig|Christoph Waltz|LÃ©a Seydoux|Ralp...

	id	imdb_id	popularity	budget	revenue	original_title	cast
11	76757	tt1617661	6.189369	176000003	183987723	Jupiter Ascending	Mila Kunis|Channing Tatum|Sean Bean|Eddie Redm...
12	264660	tt0470752	6.118847	15000000	36869414	Ex Machina	Domhnall Gleeson|Alicia Vikander|Oscar Isaac|S...
13	257344	tt2120120	5.984995	88000000	243637091	Pixels	Adam Sandler|Michelle Monaghan|Peter Dinklage|...
14	99861	tt2395427	5.944927	280000000	1405035767	Avengers: Age of Ultron	Robert Downey Jr.|Chris Hemsworth|Mark Ruffalo...
15	273248	tt3460252	5.898400	44000000	155760117	The Hateful Eight	Samuel L. Jackson|Kurt Russell|Jennifer Jason ...
16	260346	tt2446042	5.749758	48000000	325771424	Taken 3	Liam Neeson|Forest Whitaker|Maggie Grace|Famke...
17	102899	tt0478970	5.573184	130000000	518602163	Ant-Man	Paul Rudd|Michael Douglas|Evangeline Lilly|Cor...
18	150689	tt1661199	5.556818	95000000	542351353	Cinderella	Lily James|Cate Blanchett|Richard Madden|Helen...
19	131634	tt1951266	5.476958	160000000	650523427	The Hunger Games: Mockingjay - Part 2	Jennifer Lawrence|Josh Hutcherson|Liam Hemswor...
20	158852	tt1964418	5.462138	190000000	209035668	Tomorrowland	Britt Robertson|George Clooney|Raffey Cassidy|...
21	307081	tt1798684	5.337064	30000000	91709827	Southpaw	Jake Gyllenhaal|Rachel McAdams|Forest Whitaker...
22	254128	tt2126355	4.907832	110000000	470490832	San Andreas	Dwayne Johnson|Alexandra Daddario|Carla Gugino...

	id	imdb_id	popularity	budget	revenue	original_title	cast
23	216015	tt2322441	4.710402	40000000	569651467	Fifty Shades of Grey	Dakota Johnson|Jamie Dornan|Jennifer Ehle|Eloi...
24	318846	tt1596363	4.648046	28000000	133346506	The Big Short	Christian Bale|Steve Carell|Ryan Gosling|Brad ...
25	177677	tt2381249	4.566713	150000000	682330139	Mission: Impossible - Rogue Nation	Tom Cruise|Jeremy Renner|Simon Pegg|Rebecca Fe...
26	214756	tt2637276	4.564549	68000000	215863606	Ted 2	Mark Wahlberg|Seth MacFarlane|Amanda Seyfried|...
27	207703	tt2802144	4.503789	81000000	403802136	Kingsman: The Secret Service	Taron Egerton|Colin Firth|Samuel L. Jackson|Mi...
28	314365	tt1895587	4.062293	20000000	88346473	Spotlight	Mark Ruffalo|Michael Keaton|Rachel McAdams|Lie...
29	294254	tt4046784	3.968891	61000000	311256926	Maze Runner: The Scorch Trials	Dylan O'Brien|Kaya Scodelario|Thomas Brodie-Sa...
30	280996	tt3168230	3.927333	0	29355203	Mr. Holmes	Ian McKellen|Milo Parker|Laura Linney|Hattie M...
31	198184	tt1823672	3.899557	49000000	102069268	Chappie	Sharlto Copley|Dev Patel|Ninja|Yolandi Visser|...
32	254470	tt2848292	3.877764	29000000	287506194	Pitch Perfect 2	Anna Kendrick|Rebel Wilson|Hailee Steinfeld|Br...
33	296098	tt3682448	3.648210	40000000	162610473	Bridge of Spies	Tom Hanks|Mark Rylance|Amy Ryan|Alan Alda|Seba...
34	257445	tt1051904	3.644541	58000000	150170815	Goosebumps	Jack Black|Dylan Minnette|Odeya Rush|Amy Ryan|...

	id	imdb_id	popularity	budget	revenue	original_title	cast
35	264644	tt3170832	3.557846	6000000	35401758	Room	Brie Larson|Jacob Tremblay|Joan Allen|Sean Bri...
36	339527	tt1291570	3.358321	0	22354572	Solace	Abbie Cornish|Jeffrey Dean Morgan|Colin Farrel...
37	105864	tt1979388	3.339135	175000000	331926147	The Good Dinosaur	Raymond Ochoa|Jack Bright|Jeffrey Wright|Franc...
38	241554	tt2199571	3.237370	50000000	71561644	Run All Night	Liam Neeson|Ed Harris|Joel Kinnaman|Boyd Holbr...
39	167073	tt2381111	3.227329	11000000	62076141	Brooklyn	Saoirse Ronan|Domhnall Gleeson|Emory Cohen|Emi...
40	277216	tt1398426	3.202719	28000000	201634991	Straight Outta Compton	O'Shea Jackson Jr.|Corey Hawkins|Jason Mitchel...
41	274854	tt1618442	3.080505	90000000	140396650	The Last Witch Hunter	Vin Diesel|Rose Leslie|Michael Caine|Elijah Wo...
42	321697	tt2080374	3.079522	30000000	34441873	Steve Jobs	Michael Fassbender|Kate Winslet|Seth Rogen|Kat...
43	203801	tt1638355	3.053421	75000000	108145109	The Man from U.N.C.L.E.	Henry Cavill|Armie Hammer|Alicia Vikander|Eliz...
44	293863	tt1655441	3.025852	25000000	42629776	The Age of Adaline	Blake Lively|Michiel Huisman|Harrison Ford|Ell...
45	325348	tt3072482	3.023253	10000000	14333790	Hardcore Henry	Sharlto Copley|Haley Bennett|Danila Kozlovskiy...

	id	imdb_id	popularity	budget	revenue	original_title	cast
46	228161	tt2224026	2.976436	135000000	368871007	Home	Jim Parsons|Rihanna|Steve Martin|Jennifer Lope...
47	286565	tt3622592	2.968254	12000000	85512300	Paper Towns	Nat Wolff|Cara Delevingne|Halston Sage|Justice...
48	265208	tt2231253	2.932340	30000000	0	Wild Card	Jason Statham|Michael Angarano|Milo Ventimigli...
49	254320	tt3464902	2.885126	4000000	9064511	The Lobster	Colin Farrell|Rachel Weisz|LÃ©a Seydoux|John C...

50 rows × 21 columns

In [5]:
```
# view the dimension of the dataset
df.shape
```

Out[5]:
```
(10866, 21)
```

In [6]:
```
# Display a basic summary of the DataFrame
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]:
```
# check for duplicates
sum(df.duplicated())
```

Out[7]: 1

In [8]:
```
# display statistic basic summary
df.describe()
```

Out[8]:

	id	popularity	budget	revenue	runtime	vote_count	vo
count	10866.000000	10866.000000	1.086600e+04	1.086600e+04	10866.000000	10866.000000	10
mean	66064.177434	0.646441	1.462570e+07	3.982332e+07	102.070863	217.389748	
std	92130.136561	1.000185	3.091321e+07	1.170035e+08	31.381405	575.619058	
min	5.000000	0.000065	0.000000e+00	0.000000e+00	0.000000	10.000000	
25%	10596.250000	0.207583	0.000000e+00	0.000000e+00	90.000000	17.000000	
50%	20669.000000	0.383856	0.000000e+00	0.000000e+00	99.000000	38.000000	
75%	75610.000000	0.713817	1.500000e+07	2.400000e+07	111.000000	145.750000	
max	417859.000000	32.985763	4.250000e+08	2.781506e+09	900.000000	9767.000000	

We can conclude from our first glimpse analysis that our dataset is dirty and messy and therefore needs to be cleaned.

Many issues have been spotted such as:

- Non-descriptive column headers that should be droped because they are not useful for our analysis
- Duplicated data should be droped if any
- Missing values or null values that should be replaced by NAN and then deleted
- Inconsistent representations of values, dates, etc where budget and revenue should be converted into integer while release date should be converted to date format

Data Cleaning (Improving Quality and Tidiness)

After having identified the relevant issues that need to be cleaned, our second part of our data analysis process is to perform those cleaning steps:

Step 1. Remove all unusual and non descriptive column headers:

In [9]:
```
# create a list of columns to be droped from our dataset
drop_list = ['id', 'imdb_id', 'homepage', 'tagline', 'keywords', 'overview',
'production_companies', 'vote_count', 'vote_average','budget_adj', 'revenue_ad
j']

# drop extraneous columns from our dataset
df.drop(drop_list, axis = 1, inplace = True)
df.head(1)
```

Out[9]:

	popularity	budget	revenue	original_title	cast	director	runtime	
0	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	124	Action|Adve F

Step 2. Delete duplicates from our dataset:

In [10]:
```
# drop duplicate data
df.drop_duplicates(inplace = True)

# confirm correction by rechecking for duplicates
sum(df.duplicated())
```

Out[10]:
```
0
```

Step 3. Check for null and missing values and remove them:

In [11]:
```
# view missing value count for each feature
df.isnull().sum()
```

Out[11]:
```
popularity         0
budget             0
revenue            0
original_title     0
cast              76
director          44
runtime            0
genres            23
release_date       0
release_year       0
dtype: int64
```

In [12]:
```
# replace null values with NAN
df = df.replace(0, np.NAN)
df.head(35)
```

Out[12]:

	popularity	budget	revenue	original_title	cast	director	r
0	32.985763	150000000.0	1.513529e+09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	
1	28.419936	150000000.0	3.784364e+08	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	
2	13.112507	110000000.0	2.952382e+08	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	
3	11.173104	200000000.0	2.068178e+09	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	
4	9.335014	190000000.0	1.506249e+09	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	
5	9.110700	135000000.0	5.329505e+08	The Revenant	Leonardo DiCaprio|Tom Hardy|Will Poulter|Domhn...	Alejandro GonzÃ¡lez IÃ±Ã¡rritu	
6	8.654359	155000000.0	4.406035e+08	Terminator Genisys	Arnold Schwarzenegger|Jason Clarke|Emilia Clar...	Alan Taylor	
7	7.667400	108000000.0	5.953803e+08	The Martian	Matt Damon|Jessica Chastain|Kristen Wiig|Jeff ...	Ridley Scott	
8	7.404165	74000000.0	1.156731e+09	Minions	Sandra Bullock|Jon Hamm|Michael Keaton|Allison...	Kyle Balda|Pierre Coffin	
9	6.326804	175000000.0	8.537086e+08	Inside Out	Amy Poehler|Phyllis Smith|Richard Kind|Bill Ha...	Pete Docter	
10	6.200282	245000000.0	8.806746e+08	Spectre	Daniel Craig|Christoph Waltz|LÃ©a Seydoux|Ralp...	Sam Mendes	
11	6.189369	176000003.0	1.839877e+08	Jupiter Ascending	Mila Kunis|Channing Tatum|Sean Bean|Eddie Redm...	Lana Wachowski|Lilly Wachowski	
12	6.118847	15000000.0	3.686941e+07	Ex Machina	Domhnall Gleeson|Alicia Vikander|Oscar Isaac|S...	Alex Garland	
13	5.984995	88000000.0	2.436371e+08	Pixels	Adam Sandler|Michelle Monaghan|Peter Dinklage|...	Chris Columbus	
14	5.944927	280000000.0	1.405036e+09	Avengers: Age of Ultron	Robert Downey Jr.|Chris Hemsworth|Mark Ruffalo...	Joss Whedon	
15	5.898400	44000000.0	1.557601e+08	The Hateful Eight	Samuel L. Jackson|Kurt Russell|Jennifer Jason ...	Quentin Tarantino	

	popularity	budget	revenue	original_title	cast	director	r
16	5.749758	48000000.0	3.257714e+08	Taken 3	Liam Neeson|Forest Whitaker|Maggie Grace|Famke...	Olivier Megaton	
17	5.573184	130000000.0	5.186022e+08	Ant-Man	Paul Rudd|Michael Douglas|Evangeline Lilly|Cor...	Peyton Reed	
18	5.556818	95000000.0	5.423514e+08	Cinderella	Lily James|Cate Blanchett|Richard Madden|Helen...	Kenneth Branagh	
19	5.476958	160000000.0	6.505234e+08	The Hunger Games: Mockingjay - Part 2	Jennifer Lawrence|Josh Hutcherson|Liam Hemswor...	Francis Lawrence	
20	5.462138	190000000.0	2.090357e+08	Tomorrowland	Britt Robertson|George Clooney|Raffey Cassidy|...	Brad Bird	
21	5.337064	30000000.0	9.170983e+07	Southpaw	Jake Gyllenhaal|Rachel McAdams|Forest Whitaker...	Antoine Fuqua	
22	4.907832	110000000.0	4.704908e+08	San Andreas	Dwayne Johnson|Alexandra Daddario|Carla Gugino...	Brad Peyton	
23	4.710402	40000000.0	5.696515e+08	Fifty Shades of Grey	Dakota Johnson|Jamie Dornan|Jennifer Ehle|Eloi...	Sam Taylor-Johnson	
24	4.648046	28000000.0	1.333465e+08	The Big Short	Christian Bale|Steve Carell|Ryan Gosling|Brad ...	Adam McKay	
25	4.566713	150000000.0	6.823301e+08	Mission: Impossible - Rogue Nation	Tom Cruise|Jeremy Renner|Simon Pegg|Rebecca Fe...	Christopher McQuarrie	
26	4.564549	68000000.0	2.158636e+08	Ted 2	Mark Wahlberg|Seth MacFarlane|Amanda Seyfried|...	Seth MacFarlane	
27	4.503789	81000000.0	4.038021e+08	Kingsman: The Secret Service	Taron Egerton|Colin Firth|Samuel L. Jackson|Mi...	Matthew Vaughn	
28	4.062293	20000000.0	8.834647e+07	Spotlight	Mark Ruffalo|Michael Keaton|Rachel McAdams|Lie...	Tom McCarthy	
29	3.968891	61000000.0	3.112569e+08	Maze Runner: The Scorch Trials	Dylan O'Brien|Kaya Scodelario|Thomas Brodie-Sa...	Wes Ball	
30	3.927333	NaN	2.935520e+07	Mr. Holmes	Ian McKellen|Milo Parker|Laura Linney|Hattie M...	Bill Condon	
31	3.899557	49000000.0	1.020693e+08	Chappie	Sharlto Copley|Dev Patel|Ninja|Yolandi Visser|...	Neill Blomkamp	
32	3.877764	29000000.0	2.875062e+08	Pitch Perfect 2	Anna Kendrick|Rebel Wilson|Hailee Steinfeld|Br...	Elizabeth Banks	

	popularity	budget	revenue	original_title	cast	director	r
33	3.648210	40000000.0	1.626105e+08	Bridge of Spies	Tom Hanks|Mark Rylance|Amy Ryan|Alan Alda|Seba...	Steven Spielberg	
34	3.644541	58000000.0	1.501708e+08	Goosebumps	Jack Black|Dylan Minnette|Odeya Rush|Amy Ryan|...	Rob Letterman	

In [13]:
```
# drop all NAN's rows from our dataset
df = df.dropna()
df.head(35)
```

Out[13]:

	popularity	budget	revenue	original_title	cast	director	r
0	32.985763	150000000.0	1.513529e+09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	
1	28.419936	150000000.0	3.784364e+08	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	
2	13.112507	110000000.0	2.952382e+08	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	
3	11.173104	200000000.0	2.068178e+09	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	
4	9.335014	190000000.0	1.506249e+09	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	
5	9.110700	135000000.0	5.329505e+08	The Revenant	Leonardo DiCaprio|Tom Hardy|Will Poulter|Domhn...	Alejandro GonzÃ¡lez IÃ±Ã¡rritu	
6	8.654359	155000000.0	4.406035e+08	Terminator Genisys	Arnold Schwarzenegger|Jason Clarke|Emilia Clar...	Alan Taylor	
7	7.667400	108000000.0	5.953803e+08	The Martian	Matt Damon|Jessica Chastain|Kristen Wiig|Jeff ...	Ridley Scott	
8	7.404165	74000000.0	1.156731e+09	Minions	Sandra Bullock|Jon Hamm|Michael Keaton|Allison...	Kyle Balda|Pierre Coffin	
9	6.326804	175000000.0	8.537086e+08	Inside Out	Amy Poehler|Phyllis Smith|Richard Kind|Bill Ha...	Pete Docter	
10	6.200282	245000000.0	8.806746e+08	Spectre	Daniel Craig|Christoph Waltz|LÃ©a Seydoux|Ralp...	Sam Mendes	
11	6.189369	176000003.0	1.839877e+08	Jupiter Ascending	Mila Kunis|Channing Tatum|Sean Bean|Eddie Redm...	Lana Wachowski|Lilly Wachowski	
12	6.118847	15000000.0	3.686941e+07	Ex Machina	Domhnall Gleeson|Alicia Vikander|Oscar Isaac|S...	Alex Garland	
13	5.984995	88000000.0	2.436371e+08	Pixels	Adam Sandler|Michelle Monaghan|Peter Dinklage|...	Chris Columbus	
14	5.944927	280000000.0	1.405036e+09	Avengers: Age of Ultron	Robert Downey Jr.|Chris Hemsworth|Mark Ruffalo...	Joss Whedon	
15	5.898400	44000000.0	1.557601e+08	The Hateful Eight	Samuel L. Jackson|Kurt Russell|Jennifer Jason ...	Quentin Tarantino	

	popularity	budget	revenue	original_title	cast	director	r
16	5.749758	48000000.0	3.257714e+08	Taken 3	Liam Neeson|Forest Whitaker|Maggie Grace|Famke...	Olivier Megaton	
17	5.573184	130000000.0	5.186022e+08	Ant-Man	Paul Rudd|Michael Douglas|Evangeline Lilly|Cor...	Peyton Reed	
18	5.556818	95000000.0	5.423514e+08	Cinderella	Lily James|Cate Blanchett|Richard Madden|Helen...	Kenneth Branagh	
19	5.476958	160000000.0	6.505234e+08	The Hunger Games: Mockingjay - Part 2	Jennifer Lawrence|Josh Hutcherson|Liam Hemswor...	Francis Lawrence	
20	5.462138	190000000.0	2.090357e+08	Tomorrowland	Britt Robertson|George Clooney|Raffey Cassidy|...	Brad Bird	
21	5.337064	30000000.0	9.170983e+07	Southpaw	Jake Gyllenhaal|Rachel McAdams|Forest Whitaker...	Antoine Fuqua	
22	4.907832	110000000.0	4.704908e+08	San Andreas	Dwayne Johnson|Alexandra Daddario|Carla Gugino...	Brad Peyton	
23	4.710402	40000000.0	5.696515e+08	Fifty Shades of Grey	Dakota Johnson|Jamie Dornan|Jennifer Ehle|Eloi...	Sam Taylor-Johnson	
24	4.648046	28000000.0	1.333465e+08	The Big Short	Christian Bale|Steve Carell|Ryan Gosling|Brad ...	Adam McKay	
25	4.566713	150000000.0	6.823301e+08	Mission: Impossible - Rogue Nation	Tom Cruise|Jeremy Renner|Simon Pegg|Rebecca Fe...	Christopher McQuarrie	
26	4.564549	68000000.0	2.158636e+08	Ted 2	Mark Wahlberg|Seth MacFarlane|Amanda Seyfried|...	Seth MacFarlane	
27	4.503789	81000000.0	4.038021e+08	Kingsman: The Secret Service	Taron Egerton|Colin Firth|Samuel L. Jackson|Mi...	Matthew Vaughn	
28	4.062293	20000000.0	8.834647e+07	Spotlight	Mark Ruffalo|Michael Keaton|Rachel McAdams|Lie...	Tom McCarthy	
29	3.968891	61000000.0	3.112569e+08	Maze Runner: The Scorch Trials	Dylan O'Brien|Kaya Scodelario|Thomas Brodie-Sa...	Wes Ball	
31	3.899557	49000000.0	1.020693e+08	Chappie	Sharlto Copley|Dev Patel|Ninja|Yolandi Visser|...	Neill Blomkamp	
32	3.877764	29000000.0	2.875062e+08	Pitch Perfect 2	Anna Kendrick|Rebel Wilson|Hailee Steinfeld|Br...	Elizabeth Banks	
33	3.648210	40000000.0	1.626105e+08	Bridge of Spies	Tom Hanks|Mark Rylance|Amy Ryan|Alan Alda|Seba...	Steven Spielberg	

	popularity	budget	revenue	original_title	cast	director	r
34	3.644541	58000000.0	1.501708e+08	Goosebumps	Jack Black|Dylan Minnette|Odeya Rush|Amy Ryan|...	Rob Letterman	
35	3.557846	6000000.0	3.540176e+07	Room	Brie Larson|Jacob Tremblay|Joan Allen|Sean Bri...	Lenny Abrahamson	

In [14]:

```
# recheck for missing values
df.isnull().sum()
```

Out[14]:

```
popularity        0
budget            0
revenue           0
original_title    0
cast              0
director          0
runtime           0
genres            0
release_date      0
release_year      0
dtype: int64
```

In [15]:

```
# confirm changes
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3849 entries, 0 to 10848
Data columns (total 10 columns):
popularity        3849 non-null float64
budget            3849 non-null float64
revenue           3849 non-null float64
original_title    3849 non-null object
cast              3849 non-null object
director          3849 non-null object
runtime           3849 non-null float64
genres            3849 non-null object
release_date      3849 non-null object
release_year      3849 non-null int64
dtypes: float64(4), int64(1), object(5)
memory usage: 330.8+ KB
```

Step 4. Incorrect Date format and Datatypes changes:

In [16]:
```
# convert release date type from string to date format
df.release_date = pd.to_datetime(df['release_date'])
df.head(1)
```

Out[16]:

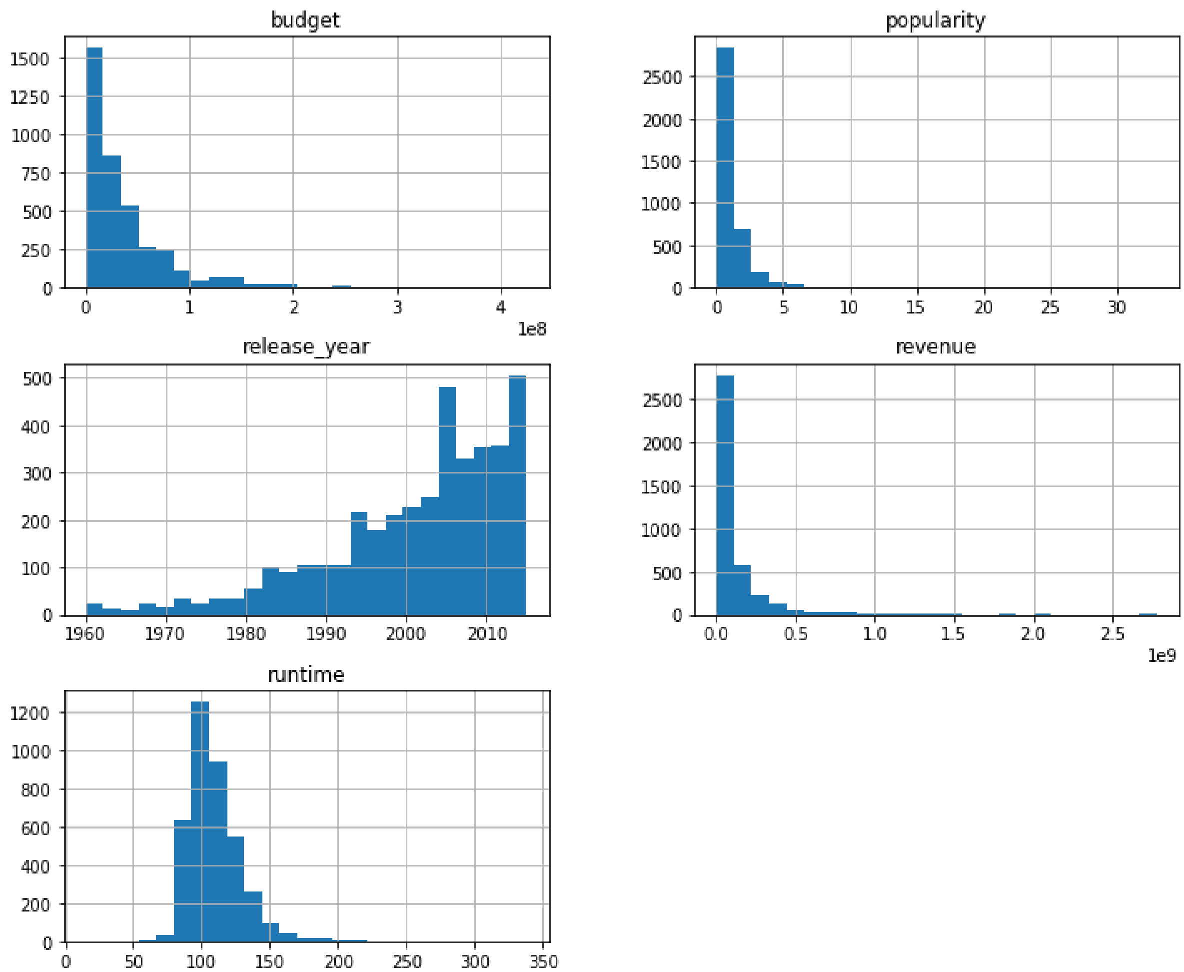
	popularity	budget	revenue	original_title	cast	director	runtime	
0	32.985763	150000000.0	1.513529e+09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	124.0	Action|A

In [17]:
```
# convert datatypes of budget and revenue to int
df.budget = df.budget.astype(int)
df.revenue = df.revenue.astype(int)
```

In [18]:
```
# confirm changes to the datatypes
df.dtypes
```

Out[18]:
```
popularity               float64
budget                     int64
revenue                    int64
original_title            object
cast                      object
director                  object
runtime                  float64
genres                    object
release_date      datetime64[ns]
release_year               int64
dtype: object
```

In [19]:
```
# explore our dataset
df.hist(figsize = (12,10), bins = 25)
plt.show()
```



According to the plot, we can conclude that the distribution of budget, revenue, runtime and popularity are skewed to the right while the distribution of release year of movies is skewed to the left (2010 and above represents the most released movies).

Exploratory Data Analysis

After having trimmed and cleaned the dataset, we are ready to move on to data exploration. So, we are going to compute statistics and create visualizations with the aim to anwer the research questions that we have posed in the introductory section.

Research Question 1: What are the top ten of most profitables movies?

- Calculate the profit for each movie:

In [20]:
```
# add a new column called profit
df.insert(3, 'profit', df['revenue'] - df['budget'])
df.head(1)
```

Out[20]:

	popularity	budget	revenue	profit	original_title	cast	director	runtime
0	32.985763	150000000	1513528810	1363528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	124.(

- Top ten of most profitables movies:

In [21]:
```
# display the top ten movies
df.sort_values(['profit'], ascending = False).head(10)
```

Out[21]:

	popularity	budget	revenue	profit	original_title	cast	direct
1386	9.432768	237000000	2781505847	2544505847	Avatar	Sam Worthington|Zoe Saldana|Sigourney Weaver|S...	Jam Camer
3	11.173104	200000000	2068178225	1868178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrar
5231	4.355219	200000000	1845034188	1645034188	Titanic	Kate Winslet|Leonardo DiCaprio|Frances Fisher|...	Jam Camer
0	32.985763	150000000	1513528810	1363528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Cc Trevorr
4	9.335014	190000000	1506249360	1316249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James W
4361	7.637767	220000000	1519557910	1299557910	The Avengers	Robert Downey Jr.|Chris Evans|Mark Ruffalo|Chr...	Joss Whed
3374	5.711315	125000000	1327817822	1202817822	Harry Potter and the Deathly Hallows: Part 2	Daniel Radcliffe|Rupert Grint|Emma Watson|Alan...	David Yat
14	5.944927	280000000	1405035767	1125035767	Avengers: Age of Ultron	Robert Downey Jr.|Chris Hemsworth|Mark Ruffalo...	Joss Whed
5422	6.112766	150000000	1274219009	1124219009	Frozen	Kristen Bell|Idina Menzel|Jonathan Groff|Josh ...	Ch Buck|Jenni L
8094	1.136610	22000000	1106279658	1084279658	The Net	Sandra Bullock|Jeremy Northam|Dennis Miller|We...	Irwin Wink

The famous movie "Avatar", which was directed by James Cameron and released on 2009, earned the highest profit of USD 2.5 billion from the top ten most profitable movies, and was followed by "Star Wars: The Force Awakens" that was directed by J.J Abrams with a profit of USD 1.8 billion, and then followed by "Titanic" which was directed by James Cameron with a profit of USD 1.6 billion.

Research Question 2: Which movie has the highest and lowest budget ?

In [22]:
```
# view the movie with the highest budget
df.loc[df['budget'].idxmax()]
```

Out[22]:
```
popularity                                                0.25054
budget                                                  425000000
revenue                                                  11087569
profit                                                 -413912431
original_title                                    The Warrior's Way
cast              Kate Bosworth|Jang Dong-gun|Geoffrey Rush|Dann...
director                                               Sngmoo Lee
runtime                                                       100
genres                    Adventure|Fantasy|Action|Western|Thriller
release_date                                  2010-12-02 00:00:00
release_year                                                 2010
Name: 2244, dtype: object
```

In [23]:
```
# display the movie with the lowest budget
df.loc[df['budget'].idxmin()]
```

Out[23]:
```
popularity                                               0.090186
budget                                                          1
revenue                                                       100
profit                                                         99
original_title                                       Lost & Found
cast              David Spade|Sophie Marceau|Ever Carradine|Step...
director                                             Jeff Pollack
runtime                                                        95
genres                                             Comedy|Romance
release_date                                  1999-04-23 00:00:00
release_year                                                 1999
Name: 2618, dtype: object
```

The movie with the highest budget spent is "The Warrior's Way" which is around 425 million, while "Lost & Found" is the movie with the lowest budget which is around $1. It should be a data entry error.

Research Question 3: Which movie has the highest and the lowest revenue?

In [24]:
```
# show the movie with the highest revenue
df.loc[df['revenue'].idxmax()]
```

Out[24]:
```
popularity                                                       9.43277
budget                                                         237000000
revenue                                                       2781505847
profit                                                        2544505847
original_title                                                    Avatar
cast              Sam Worthington|Zoe Saldana|Sigourney Weaver|S...
director                                                   James Cameron
runtime                                                              162
genres                         Action|Adventure|Fantasy|Science Fiction
release_date                                         2009-12-10 00:00:00
release_year                                                        2009
Name: 1386, dtype: object
```

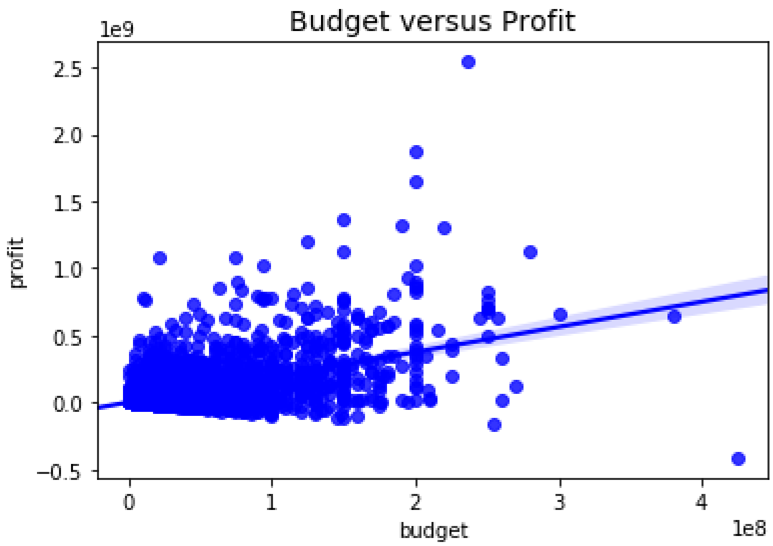
In [25]:
```
# show the movies with the lowest revenue
df.loc[df['revenue'].idxmin()]
```

Out[25]:
```
popularity                                                      0.462609
budget                                                           6000000
revenue                                                                2
profit                                                          -5999998
original_title                                            Shattered Glass
cast              Hayden Christensen|Peter Sarsgaard|ChloÃ« Sevi...
director                                                       Billy Ray
runtime                                                               94
genres                                                     Drama|History
release_date                                         2003-11-14 00:00:00
release_year                                                        2003
Name: 5067, dtype: object
```

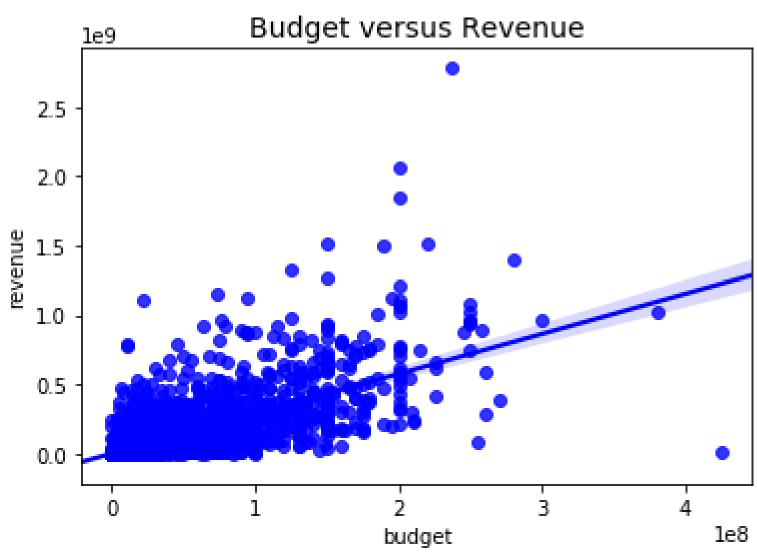
The movie with the highest revenue of 237 million is "Avatar", while "Shattered Glass" is the movie with the lowest revenue which is only $2.

Next, let's plot respectively both the relationship between profit realized by movies and the budget spent as well as the relationship between revenue by movies and the budget spent:

In [26]:
```
# scatterplot of budget against profit
sns.regplot(x = df['budget'], y = df['profit'], color = 'blue')
plt.title('Budget versus Profit', fontsize = 14)
plt.show()
```



In [27]:
```
# scatterplot of budget against revenue
sns.regplot(x = df['budget'], y = df['revenue'], color='blue')
plt.title('Budget versus Revenue', fontsize = 14)
plt.show()
```



In [28]:
```
# create the function of correlation coefficient
def correlation_coeff(x,y):
    std_x = (x-x.mean())/x.std(ddof=0)
    std_y = (y-y.mean())/y.std(ddof=0)
    return(std_x*std_y).mean()
correlation_coeff(df['budget'],df['profit'])
```

Out[28]: 0.5266595220688896

In [29]:
```
correlation_coeff(df['budget'],df['revenue'])
```

Out[29]:
```
0.68840319045227083
```

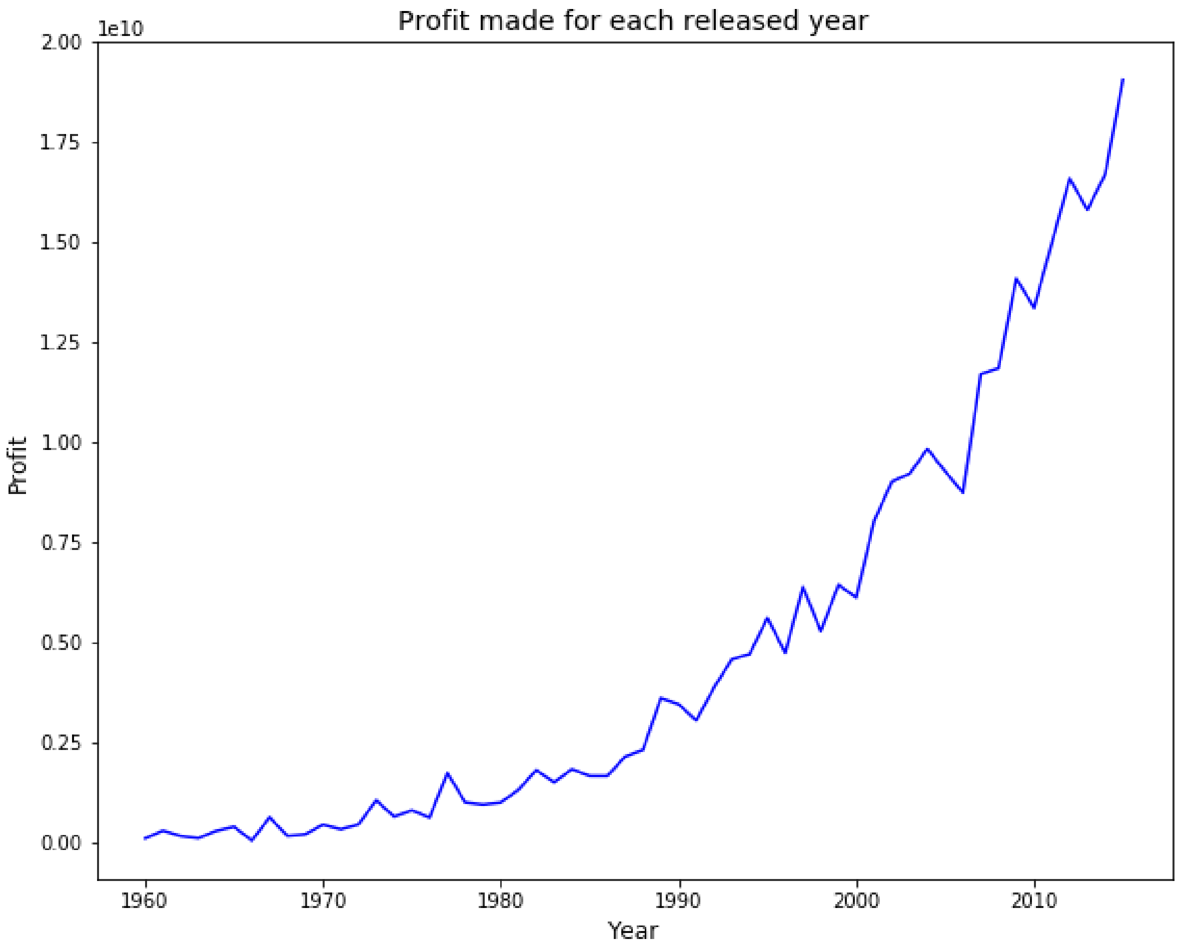
Both budget and profit and also budget and revenue have a positive correlation of 0.53 and 0.68 respectively which means that the more budget spent on a movie the more revenue or profit to be realized. However, according to the both plot we can highlight that some movies earned a high profit with less budget spent.

Research Question 4: In which year did movies' industry realize their most profit?

In [30]:
```
# get total profit made by release_year
profit_year = df.groupby('release_year')['profit'].sum()
profit_year.head()
```

Out[30]:
```
release_year
1960    108198052
1961    299083188
1962    166879846
1963    115411882
1964    294678387
Name: profit, dtype: int64
```

In [31]:
```
# plot profit made for each realeased year
profit_year.plot(figsize = (10,8), color = 'b')
plt.xlabel('Year', fontsize = 12)
plt.ylabel('Profit', fontsize = 12 )
plt.title('Profit made for each released year', fontsize = 14)
plt.show()
```



The relationship between Total profit and released year is upward trending, which means that in recent years particularly after 2010, the movies' indutry realized the greatest profit about USD 20 billion (2 times 1e10 million) compared to the period between 1960 and 2005 where the profit didn't go beyond USD 10 billion.

Next, we are going to identify whether the yearly realized total profit is due the popularity of both old and new movies or only due to the highest rate of released of new movies in a year.

Research Question 5: What's the relationship between both popularity, runtime of a movie and their profit?

In [32]:
```
# scatterplot of popularity against profit
# scatterplot of runtime against profit
sns.pairplot(df, x_vars =['popularity', 'runtime'], y_vars = ['profit'], size
= 7, aspect = 0.7, kind='reg')
plt.title('The effect of Popularity and Runtime on Profit', fontsize = 14)
plt.show()
```

In this section, we are trying to find out how popularity and runtime affect profit.

According to both scatterplots, we can conclude that the trendline in both of them is upward sloping, and there is a positive relationship not only between popularity and profit but also between runtime and profit. This means that both higher popularity and runtime increase profit. In addition, the slope of popularity is greater than the slope of runtime which means that profit increased with a higher rate with popularity that did with runtime.

Let's check by how much profit is affected by both popularity and runtime by calculating the correlation coefficient:

In [33]:
```
# create the function of correlation coefficient
def correlation_coeff(x,y):
    std_x = (x-x.mean())/x.std(ddof=0)
    std_y = (y-y.mean())/y.std(ddof=0)
    return(std_x*std_y).mean()
```

In [34]:
```
# correlation between popularity and profit
correlation_coeff(df['popularity'], df['profit'])
```

Out[34]: 0.59608020443389209

In [35]:
```
# correlation between runtime and profit
correlation_coeff(df['runtime'], df['profit'])
```

Out[35]:
```
0.22059705250729106
```

The coefficient of correlation between popularity and profit is positive and it is around 0.60 which means that a movie with high popularity rate tends to earn higher profit. While the lower coefficient of correlation between runtime and profit of 0.14 means that the longer the duration of a movie the less higher the profit.

In [36]:
```
df['runtime'].describe()
```

Out[36]:
```
count    3849.000000
mean      109.217459
std        19.914141
min        15.000000
25%        95.000000
50%       106.000000
75%       119.000000
max       338.000000
Name: runtime, dtype: float64
```

In addition, it can be inferred from the runtime plot and from the above statistic description that some movies in the runtime range between 95 and 120 tends to earn higher profit:

- 25% of movies have runtime less than 95mn
- 50% of movies have a runtime less than 106mn which is the median of the runtime distribution.
- 75% of movies have a runtime less than 119mn
- The mean is 109 which is higher than the median of 106 which means that the runtime distribution is skewed to the right. So, the audience prefer better a movie with a runtime that falls on the range.

Research Question 6: Which movie's genre has the highest release?

In [37]:
```
df.head(1)
```

Out[37]:

	popularity	budget	revenue	profit	original_title	cast	director	runtime
0	32.985763	150000000	1513528810	1363528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	124.0

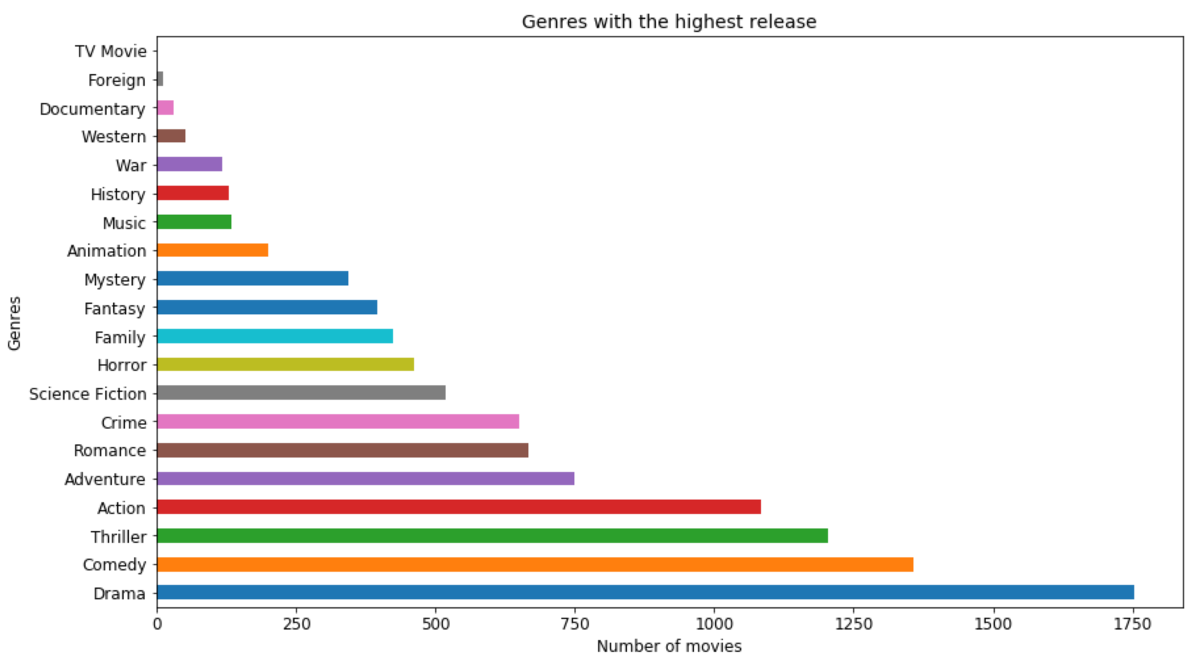
The below built-in function will help us to separate the content of genre features and then count the number of the movies corresponding to each genre:

In [38]:
```
# create a function that separate the content of genres
def count_genre(column):
    split_data = pd.Series(df[column].str.cat(sep='|'). split('|'))
    count_data = split_data.value_counts(ascending = False)
    return count_data
```

In [39]:
```
# count realized movies by genres
count_data = count_genre('genres')
count_data.head()
```

Out[39]:
```
Drama        1753
Comedy       1357
Thriller     1203
Action       1085
Adventure     749
dtype: int64
```

In [40]:
```
# plot genres against released movies
count_data.plot(kind ='barh', figsize = (14,8), fontsize = 12)
plt.xlabel('Number of movies', fontsize = 12)
plt.ylabel('Genres', fontsize = 12)
plt.title('Genres with the highest release', fontsize = 14)
plt.show()
```



According to the plot, we can conclude that the audience prefer the most Drama, after that they prefer watching Comedy's movies, then Thriller as well as Action's movies. All of them represent the highest proportion of profit for the movies' industry. The highest the preference rate for those genres, the greatest corresponding released movies and therefore the highest profit.

Research Question 7: Who are the most succesful directors?

In this section, we are going to check who is the succesful director that directed the maximum number of movies.

In [41]:
```
# create a built-in function that count movies by director
def count_director_movies(column):
    split_data = pd.Series(df[column].str.cat(sep='|'). split('|'))
    count_data = split_data.value_counts(ascending = False)
    return count_data
```

In [42]:
```
# display the top ten best director
count_data = count_director_movies('director')
count_data.head(10)
```

Out[42]:
```
Steven Spielberg     28
Clint Eastwood       24
Ridley Scott         21
Woody Allen          18
Robert Rodriguez     17
Martin Scorsese      17
Tim Burton           17
Steven Soderbergh    17
Robert Zemeckis      15
Renny Harlin         15
dtype: int64
```

Steven Spielberg is the most prolific American director with 28 movies to his credit, followed by Clint Eastwood with 24 movies, Ridley Scott with 21 and then Martin Scorsese with 17 filmes in his credit.

Research Question 8: What's the most frequent cast?

In [43]:
```
# create a built-in function
def count_cast_movies(column):
    split_data = pd.Series(df[column].str.cat(sep='|'). split('|'))
    count_data = split_data.value_counts(ascending = False)
    return count_data
```

In [44]:
```
# display the most frequent cast
count_data = count_cast_movies('cast')
count_data.head(10)
```

Out[44]:
```
Robert De Niro        52
Bruce Willis          46
Samuel L. Jackson     44
Nicolas Cage          43
Matt Damon            36
Johnny Depp           35
Sylvester Stallone    34
Morgan Freeman        34
Brad Pitt             34
Tom Hanks             34
dtype: int64
```

A good actor in a movie like Robert De Niro or Bruce Willis and others is a sign of a good casting and definitely a good sign for a successful movie in terms of audience and profit. The mean reason is that a a good casting do a great job in portraying very well their characters.

Conclusions

For the sake of summary, the following are the findings of investigating the Movie Database (TMDb):

- "Avatar", "Star Wars: The Force Awakens" and "Titanic" are the most profitable movies.
- "Avatar" is the movie with the highest revenue, while "Shattered Glass" is the movie with the lowest revenue.
- The "Warrior's Way" is the movie with the highest budget spent.
- There is a strong relationship between budget and revenue explained by a positive correlation of 0.68 which means that an increase in budget allocated to a movie leads to an increase in its revenue.
- There is also a strong relationship between budget and profit which is explained by a positive correlation of 0.53. However, we have to mention that some movies earned higher profit with less spendings.
- After 2010, the movies' indutry realized the greatest profit about USD 20 billion compared to the period between 1960 and 2005 where the profit didn't go beyond usd 10 billion.
- There is a positive relationship between popularity and profit where the coefficient of correlation is around 0.60 which is high and explain why higher movie's popularity increase profit with a higher value.
- There is also a positive relationship between runtime and profit where the coefficient of correlation is only around 0.14 which means the profit is increased with lower value with longer duration of a movie. The skewness to the right of the runtime scatterplot makes us to conclude that movies in the runtime range between 95 and 120 tends to earn higher profit.
- Drama, followed by Comedy, Thriller and Action are the most preferable movies' genres by the audience.
- Steven Spielberg is the most successful director with 28 movies to his credit, while the best actor comes back to Robert De Niro with 52 movies.

One of the limitations to draw perfect conclusion is that a poorer quality of the database can potentially be costing higher price to the final findings. The database was untidy, may be because the data was collected from various sources, the reason why there were many null and missing values. Morever, our dataset should be cleaned and assessed before being analyzed which leads that many movies where exluded from our analysis.

Submitting your Project

In [1]:
```
from subprocess import call
call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

Out[1]: 0

In [ ]:
```

```