

.ipynb

June 20, 2020

1 Project: Investigate a Dataset (TMDb Dataset)

1.1 Table of Contents

Introduction

Data Wrangling

Exploratory Data Analysis

Conclusions

Introduction TMDb Movie data from Kaggle is chosen to investigate in the project. This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.

- Some columns, like cast and genres, contain multiple values separated by pipe (|) characters.
- 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.

1.2 Questions

Here are the questions need to answer:

- Which movies were most and least profitable?
- What was the most profitable movie each year?
- Do movie budgets increase over time?
- What is the most profitable genre?

1.3 Import Libraries

```
In [1]: import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
%matplotlib inline

In [2]: plt.style.use('fivethirtyeight')
plt.style.use('seaborn-poster')
```

```
# Format floats to show commas and two decimals
pd.options.display.float_format = "{0:,.2f}".format
```

Data Wrangling

```
In [3]: 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 137,999,939.28 1,392,445,892.52
1           7.10           2015 137,999,939.28  348,161,292.49
2           6.30           2015 101,199,955.47  271,619,025.41
3           7.50           2015 183,999,919.04 1,902,723,129.80
4           7.30           2015 174,799,923.09 1,385,748,801.47

```

[5 rows x 21 columns]

In [4]: df.shape

Out[4]: (10866, 21)

In [5]: df.dtypes

```

Out[5]: id                int64
imdb_id                 object
popularity              float64
budget                  int64
revenue                 int64
original_title          object
cast                    object
homepage                object
director                object
tagline                 object
keywords                object
overview                object
runtime                 int64
genres                  object
production_companies    object
release_date            object
vote_count              int64
vote_average            float64

```

```

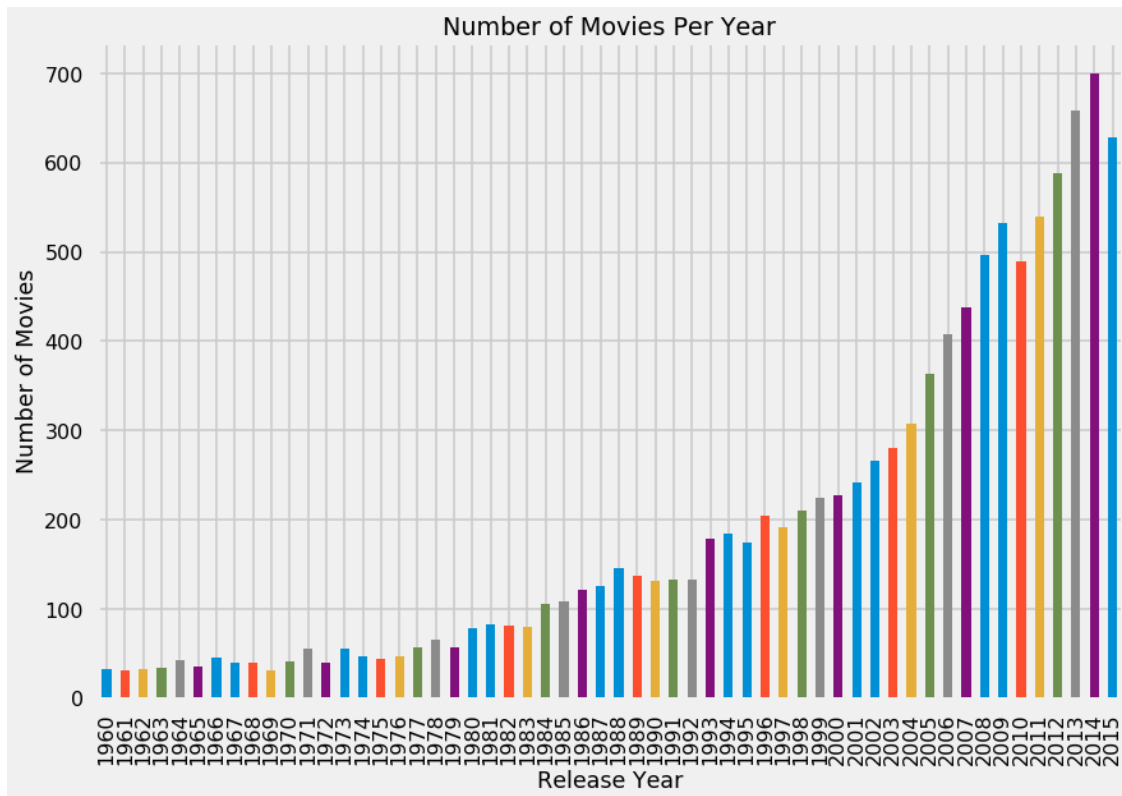
release_year          int64
budget_adj            float64
revenue_adj           float64
dtype: object

```

```

In [6]: #How many movies released by years?
df.groupby('release_year')['id'].count().plot.bar()
plt.xlabel('Release Year')
plt.ylabel('Number of Movies')
plt.title('Number of Movies Per Year')
plt.show()

```



Data Cleaning

Eliminate duplicates

```

In [7]: sum(df.duplicated())

```

```

Out[7]: 1

```

```

In [8]: df.drop_duplicates(inplace=True)

```

Convert datatypes

```
In [9]: df.dtypes.release_date
```

```
Out[9]: dtype('O')
```

```
In [10]: # Convert to a datetime
         df['release_date'] = pd.to_datetime(df['release_date'])
```

```
In [11]: # Double check
         df.dtypes.release_date
```

```
Out[11]: dtype('<M8[ns]')
```

Add columns

```
In [12]: # Calculate net income only if both budget_adj and revenue_adj are > 0
         def income(row):
             if row['budget_adj'] > 0 and row['revenue_adj'] > 0:
                 val = row['revenue_adj'] - row['budget_adj']
             else:
                 val = float('NaN')
             return val
```

```
In [13]: # Add a net_income column
         df['net_income'] = df.apply(income, axis=1)
         df.head()
```

```
Out[13]:
```

	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	...	runtime \
0	The park is open.	...	124
1	What a Lovely Day.	...	120
2	One Choice Can Destroy You	...	119
3	Every generation has a story.	...	136
4	Vengeance Hits Home	...	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...	2015-06-09	5562
1	Village Roadshow Pictures Kennedy Miller Produ...	2015-05-13	6185
2	Summit Entertainment Mandeville Films Red Wago...	2015-03-18	2480
3	Lucasfilm Truenorth Productions Bad Robot	2015-12-15	5292
4	Universal Pictures Original Film Media Rights ...	2015-04-01	2947

	vote_average	release_year	budget_adj	revenue_adj	net_income
0	6.50	2015	137,999,939.28	1,392,445,892.52	1,254,445,953.24
1	7.10	2015	137,999,939.28	348,161,292.49	210,161,353.21
2	6.30	2015	101,199,955.47	271,619,025.41	170,419,069.94
3	7.50	2015	183,999,919.04	1,902,723,129.80	1,718,723,210.76
4	7.30	2015	174,799,923.09	1,385,748,801.47	1,210,948,878.38

[5 rows x 22 columns]

```
In [14]: df.net_income.describe()
```

```
Out[14]: count          3,854.00
mean          92,824,697.22
std          194,071,459.74
min         -413,912,431.00
25%          -1,504,994.63
50%          27,370,641.16
75%          107,454,751.41
max           2,750,136,650.92
Name: net_income, dtype: float64
```

Normalize overloaded columns The cast and genres columns are pipe-delimited strings. I am going to extract them into new dataframes and create lookup dataframes to handle the relationships.

```
In [15]: def splitDataFrameList(df, target_column, separator):
    row_accumulator = []

    def splitListToRows(row, separator):
        split_row = row[target_column].split(separator)
        for s in split_row:
            new_row = row.to_dict()
            new_row[target_column] = s
            row_accumulator.append(new_row)

    df.apply(splitListToRows, axis=1, args=(separator, ))
    new_df = pd.DataFrame(row_accumulator)
    return new_df
```

```
In [16]: # Split and flatten the cast and genres categories
df_flat = splitDataFrameList(df.dropna(), 'cast', '|')
df_flat = splitDataFrameList(df_flat.dropna(), 'genres', '|')
df_flat = splitDataFrameList(df_flat.dropna(), 'keywords', '|')
```

```
In [17]: df_flat.describe()
```

```
Out[17]:
```

	budget	budget_adj	id	net_income	popularity \
count	79,400.00	79,400.00	79,400.00	79,400.00	79,400.00
mean	60,724,919.60	63,757,724.78	49,444.64	172,223,733.42	2.03
std	59,801,738.44	59,404,434.38	70,693.56	295,085,244.58	2.46
min	1.00	0.97	11.00	-413,912,431.00	0.01
25%	17,000,000.00	19,387,960.85	4,464.00	5,387,689.10	0.76
50%	40,000,000.00	43,622,911.92	17,979.00	60,073,390.00	1.32
75%	85,000,000.00	91,941,878.45	60,308.00	210,161,353.21	2.49
max	425,000,000.00	425,000,000.00	333,348.00	2,750,136,650.92	32.99

	release_year	revenue	revenue_adj	runtime	vote_average \
count	79,400.00	79,400.00	79,400.00	79,400.00	79,400.00
mean	2,006.67	208,570,350.20	235,981,458.19	111.02	6.30
std	8.49	281,825,273.88	326,061,177.81	19.43	0.80
min	1,961.00	43.00	43.00	63.00	2.20
25%	2,005.00	31,670,620.00	34,084,779.31	97.00	5.80
50%	2,009.00	101,371,017.00	110,662,825.91	108.00	6.30
75%	2,011.00	284,600,000.00	320,834,306.77	122.00	6.90
max	2,015.00	2,781,505,847.00	2,827,123,750.41	201.00	8.30

	vote_count
count	79,400.00
mean	1,110.75
std	1,377.22

```

min          10.00
25%          228.00
50%          560.00
75%         1,527.00
max          9,767.00

```

```
In [18]: df_flat.head(50)
```

```

Out[18]:
   budget  budget_adj      cast      director \
0  150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
1  150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
2  150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
3  150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
4  150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
5  150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
6  150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
7  150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
8  150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
9  150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
10 150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
11 150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
12 150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
13 150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
14 150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
15 150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
16 150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
17 150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
18 150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
19 150000000  137,999,939.28  Chris Pratt  Colin Trevorrow
20 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
21 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
22 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
23 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
24 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
25 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
26 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
27 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
28 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
29 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
30 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
31 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
32 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
33 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
34 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
35 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
36 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
37 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow
38 150000000  137,999,939.28  Bryce Dallas Howard  Colin Trevorrow

```

39	150000000	137,999,939.28	Bryce Dallas Howard	Colin Trevorrow
40	150000000	137,999,939.28	Irrfan Khan	Colin Trevorrow
41	150000000	137,999,939.28	Irrfan Khan	Colin Trevorrow
42	150000000	137,999,939.28	Irrfan Khan	Colin Trevorrow
43	150000000	137,999,939.28	Irrfan Khan	Colin Trevorrow
44	150000000	137,999,939.28	Irrfan Khan	Colin Trevorrow
45	150000000	137,999,939.28	Irrfan Khan	Colin Trevorrow
46	150000000	137,999,939.28	Irrfan Khan	Colin Trevorrow
47	150000000	137,999,939.28	Irrfan Khan	Colin Trevorrow
48	150000000	137,999,939.28	Irrfan Khan	Colin Trevorrow
49	150000000	137,999,939.28	Irrfan Khan	Colin Trevorrow

	genres	homepage	id	imdb_id	\
0	Action	http://www.jurassicworld.com/	135397	tt0369610	
1	Action	http://www.jurassicworld.com/	135397	tt0369610	
2	Action	http://www.jurassicworld.com/	135397	tt0369610	
3	Action	http://www.jurassicworld.com/	135397	tt0369610	
4	Action	http://www.jurassicworld.com/	135397	tt0369610	
5	Adventure	http://www.jurassicworld.com/	135397	tt0369610	
6	Adventure	http://www.jurassicworld.com/	135397	tt0369610	
7	Adventure	http://www.jurassicworld.com/	135397	tt0369610	
8	Adventure	http://www.jurassicworld.com/	135397	tt0369610	
9	Adventure	http://www.jurassicworld.com/	135397	tt0369610	
10	Science Fiction	http://www.jurassicworld.com/	135397	tt0369610	
11	Science Fiction	http://www.jurassicworld.com/	135397	tt0369610	
12	Science Fiction	http://www.jurassicworld.com/	135397	tt0369610	
13	Science Fiction	http://www.jurassicworld.com/	135397	tt0369610	
14	Science Fiction	http://www.jurassicworld.com/	135397	tt0369610	
15	Thriller	http://www.jurassicworld.com/	135397	tt0369610	
16	Thriller	http://www.jurassicworld.com/	135397	tt0369610	
17	Thriller	http://www.jurassicworld.com/	135397	tt0369610	
18	Thriller	http://www.jurassicworld.com/	135397	tt0369610	
19	Thriller	http://www.jurassicworld.com/	135397	tt0369610	
20	Action	http://www.jurassicworld.com/	135397	tt0369610	
21	Action	http://www.jurassicworld.com/	135397	tt0369610	
22	Action	http://www.jurassicworld.com/	135397	tt0369610	
23	Action	http://www.jurassicworld.com/	135397	tt0369610	
24	Action	http://www.jurassicworld.com/	135397	tt0369610	
25	Adventure	http://www.jurassicworld.com/	135397	tt0369610	
26	Adventure	http://www.jurassicworld.com/	135397	tt0369610	
27	Adventure	http://www.jurassicworld.com/	135397	tt0369610	
28	Adventure	http://www.jurassicworld.com/	135397	tt0369610	
29	Adventure	http://www.jurassicworld.com/	135397	tt0369610	
30	Science Fiction	http://www.jurassicworld.com/	135397	tt0369610	
31	Science Fiction	http://www.jurassicworld.com/	135397	tt0369610	
32	Science Fiction	http://www.jurassicworld.com/	135397	tt0369610	
33	Science Fiction	http://www.jurassicworld.com/	135397	tt0369610	
34	Science Fiction	http://www.jurassicworld.com/	135397	tt0369610	

35	Thriller	http://www.jurassicworld.com/	135397	tt0369610
36	Thriller	http://www.jurassicworld.com/	135397	tt0369610
37	Thriller	http://www.jurassicworld.com/	135397	tt0369610
38	Thriller	http://www.jurassicworld.com/	135397	tt0369610
39	Thriller	http://www.jurassicworld.com/	135397	tt0369610
40	Action	http://www.jurassicworld.com/	135397	tt0369610
41	Action	http://www.jurassicworld.com/	135397	tt0369610
42	Action	http://www.jurassicworld.com/	135397	tt0369610
43	Action	http://www.jurassicworld.com/	135397	tt0369610
44	Action	http://www.jurassicworld.com/	135397	tt0369610
45	Adventure	http://www.jurassicworld.com/	135397	tt0369610
46	Adventure	http://www.jurassicworld.com/	135397	tt0369610
47	Adventure	http://www.jurassicworld.com/	135397	tt0369610
48	Adventure	http://www.jurassicworld.com/	135397	tt0369610
49	Adventure	http://www.jurassicworld.com/	135397	tt0369610

	keywords	net_income	...	popularity \
0	monster	1,254,445,953.24	...	32.99
1	dna	1,254,445,953.24	...	32.99
2	tyrannosaurus rex	1,254,445,953.24	...	32.99
3	velociraptor	1,254,445,953.24	...	32.99
4	island	1,254,445,953.24	...	32.99
5	monster	1,254,445,953.24	...	32.99
6	dna	1,254,445,953.24	...	32.99
7	tyrannosaurus rex	1,254,445,953.24	...	32.99
8	velociraptor	1,254,445,953.24	...	32.99
9	island	1,254,445,953.24	...	32.99
10	monster	1,254,445,953.24	...	32.99
11	dna	1,254,445,953.24	...	32.99
12	tyrannosaurus rex	1,254,445,953.24	...	32.99
13	velociraptor	1,254,445,953.24	...	32.99
14	island	1,254,445,953.24	...	32.99
15	monster	1,254,445,953.24	...	32.99
16	dna	1,254,445,953.24	...	32.99
17	tyrannosaurus rex	1,254,445,953.24	...	32.99
18	velociraptor	1,254,445,953.24	...	32.99
19	island	1,254,445,953.24	...	32.99
20	monster	1,254,445,953.24	...	32.99
21	dna	1,254,445,953.24	...	32.99
22	tyrannosaurus rex	1,254,445,953.24	...	32.99
23	velociraptor	1,254,445,953.24	...	32.99
24	island	1,254,445,953.24	...	32.99
25	monster	1,254,445,953.24	...	32.99
26	dna	1,254,445,953.24	...	32.99
27	tyrannosaurus rex	1,254,445,953.24	...	32.99
28	velociraptor	1,254,445,953.24	...	32.99
29	island	1,254,445,953.24	...	32.99
30	monster	1,254,445,953.24	...	32.99

31	dna	1,254,445,953.24	...	32.99
32	tyrannosaurus rex	1,254,445,953.24	...	32.99
33	velociraptor	1,254,445,953.24	...	32.99
34	island	1,254,445,953.24	...	32.99
35	monster	1,254,445,953.24	...	32.99
36	dna	1,254,445,953.24	...	32.99
37	tyrannosaurus rex	1,254,445,953.24	...	32.99
38	velociraptor	1,254,445,953.24	...	32.99
39	island	1,254,445,953.24	...	32.99
40	monster	1,254,445,953.24	...	32.99
41	dna	1,254,445,953.24	...	32.99
42	tyrannosaurus rex	1,254,445,953.24	...	32.99
43	velociraptor	1,254,445,953.24	...	32.99
44	island	1,254,445,953.24	...	32.99
45	monster	1,254,445,953.24	...	32.99
46	dna	1,254,445,953.24	...	32.99
47	tyrannosaurus rex	1,254,445,953.24	...	32.99
48	velociraptor	1,254,445,953.24	...	32.99
49	island	1,254,445,953.24	...	32.99

	production_companies	release_date	\
0	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
1	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
2	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
3	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
4	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
5	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
6	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
7	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
8	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
9	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
10	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
11	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
12	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
13	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
14	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
15	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
16	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
17	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
18	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
19	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
20	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
21	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
22	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
23	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
24	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
25	Universal Studios Amblin Entertainment Legenda...	2015-06-09	
26	Universal Studios Amblin Entertainment Legenda...	2015-06-09	

27	Universal Studios Amblin Entertainment Legenda...	2015-06-09
28	Universal Studios Amblin Entertainment Legenda...	2015-06-09
29	Universal Studios Amblin Entertainment Legenda...	2015-06-09
30	Universal Studios Amblin Entertainment Legenda...	2015-06-09
31	Universal Studios Amblin Entertainment Legenda...	2015-06-09
32	Universal Studios Amblin Entertainment Legenda...	2015-06-09
33	Universal Studios Amblin Entertainment Legenda...	2015-06-09
34	Universal Studios Amblin Entertainment Legenda...	2015-06-09
35	Universal Studios Amblin Entertainment Legenda...	2015-06-09
36	Universal Studios Amblin Entertainment Legenda...	2015-06-09
37	Universal Studios Amblin Entertainment Legenda...	2015-06-09
38	Universal Studios Amblin Entertainment Legenda...	2015-06-09
39	Universal Studios Amblin Entertainment Legenda...	2015-06-09
40	Universal Studios Amblin Entertainment Legenda...	2015-06-09
41	Universal Studios Amblin Entertainment Legenda...	2015-06-09
42	Universal Studios Amblin Entertainment Legenda...	2015-06-09
43	Universal Studios Amblin Entertainment Legenda...	2015-06-09
44	Universal Studios Amblin Entertainment Legenda...	2015-06-09
45	Universal Studios Amblin Entertainment Legenda...	2015-06-09
46	Universal Studios Amblin Entertainment Legenda...	2015-06-09
47	Universal Studios Amblin Entertainment Legenda...	2015-06-09
48	Universal Studios Amblin Entertainment Legenda...	2015-06-09
49	Universal Studios Amblin Entertainment Legenda...	2015-06-09

	release_year	revenue	revenue_adj	runtime	tagline \
0	2015	1513528810	1,392,445,892.52	124	The park is open.
1	2015	1513528810	1,392,445,892.52	124	The park is open.
2	2015	1513528810	1,392,445,892.52	124	The park is open.
3	2015	1513528810	1,392,445,892.52	124	The park is open.
4	2015	1513528810	1,392,445,892.52	124	The park is open.
5	2015	1513528810	1,392,445,892.52	124	The park is open.
6	2015	1513528810	1,392,445,892.52	124	The park is open.
7	2015	1513528810	1,392,445,892.52	124	The park is open.
8	2015	1513528810	1,392,445,892.52	124	The park is open.
9	2015	1513528810	1,392,445,892.52	124	The park is open.
10	2015	1513528810	1,392,445,892.52	124	The park is open.
11	2015	1513528810	1,392,445,892.52	124	The park is open.
12	2015	1513528810	1,392,445,892.52	124	The park is open.
13	2015	1513528810	1,392,445,892.52	124	The park is open.
14	2015	1513528810	1,392,445,892.52	124	The park is open.
15	2015	1513528810	1,392,445,892.52	124	The park is open.
16	2015	1513528810	1,392,445,892.52	124	The park is open.
17	2015	1513528810	1,392,445,892.52	124	The park is open.
18	2015	1513528810	1,392,445,892.52	124	The park is open.
19	2015	1513528810	1,392,445,892.52	124	The park is open.
20	2015	1513528810	1,392,445,892.52	124	The park is open.
21	2015	1513528810	1,392,445,892.52	124	The park is open.
22	2015	1513528810	1,392,445,892.52	124	The park is open.

23	2015	1513528810	1,392,445,892.52	124	The park is open.
24	2015	1513528810	1,392,445,892.52	124	The park is open.
25	2015	1513528810	1,392,445,892.52	124	The park is open.
26	2015	1513528810	1,392,445,892.52	124	The park is open.
27	2015	1513528810	1,392,445,892.52	124	The park is open.
28	2015	1513528810	1,392,445,892.52	124	The park is open.
29	2015	1513528810	1,392,445,892.52	124	The park is open.
30	2015	1513528810	1,392,445,892.52	124	The park is open.
31	2015	1513528810	1,392,445,892.52	124	The park is open.
32	2015	1513528810	1,392,445,892.52	124	The park is open.
33	2015	1513528810	1,392,445,892.52	124	The park is open.
34	2015	1513528810	1,392,445,892.52	124	The park is open.
35	2015	1513528810	1,392,445,892.52	124	The park is open.
36	2015	1513528810	1,392,445,892.52	124	The park is open.
37	2015	1513528810	1,392,445,892.52	124	The park is open.
38	2015	1513528810	1,392,445,892.52	124	The park is open.
39	2015	1513528810	1,392,445,892.52	124	The park is open.
40	2015	1513528810	1,392,445,892.52	124	The park is open.
41	2015	1513528810	1,392,445,892.52	124	The park is open.
42	2015	1513528810	1,392,445,892.52	124	The park is open.
43	2015	1513528810	1,392,445,892.52	124	The park is open.
44	2015	1513528810	1,392,445,892.52	124	The park is open.
45	2015	1513528810	1,392,445,892.52	124	The park is open.
46	2015	1513528810	1,392,445,892.52	124	The park is open.
47	2015	1513528810	1,392,445,892.52	124	The park is open.
48	2015	1513528810	1,392,445,892.52	124	The park is open.
49	2015	1513528810	1,392,445,892.52	124	The park is open.

	vote_average	vote_count
0	6.50	5562
1	6.50	5562
2	6.50	5562
3	6.50	5562
4	6.50	5562
5	6.50	5562
6	6.50	5562
7	6.50	5562
8	6.50	5562
9	6.50	5562
10	6.50	5562
11	6.50	5562
12	6.50	5562
13	6.50	5562
14	6.50	5562
15	6.50	5562
16	6.50	5562
17	6.50	5562
18	6.50	5562

19	6.50	5562
20	6.50	5562
21	6.50	5562
22	6.50	5562
23	6.50	5562
24	6.50	5562
25	6.50	5562
26	6.50	5562
27	6.50	5562
28	6.50	5562
29	6.50	5562
30	6.50	5562
31	6.50	5562
32	6.50	5562
33	6.50	5562
34	6.50	5562
35	6.50	5562
36	6.50	5562
37	6.50	5562
38	6.50	5562
39	6.50	5562
40	6.50	5562
41	6.50	5562
42	6.50	5562
43	6.50	5562
44	6.50	5562
45	6.50	5562
46	6.50	5562
47	6.50	5562
48	6.50	5562
49	6.50	5562

[50 rows x 22 columns]

To keep things clean, I am going to make separate dataframes:

- `df_film` will contain only columns I won't be summing or averaging (such as genre and cast)
- `df_vals` will contain the remaining

```
In [19]: df_film = df_flat.drop(columns=[
        'budget', 'budget_adj', 'net_income', 'revenue', 'revenue_adj',
        'vote_average', 'vote_count'
    ])
df_film.head()
```

```
Out[19]:
```

	cast	director	genres	homepage \
0	Chris Pratt	Colin Trevorrow	Action	http://www.jurassicworld.com/
1	Chris Pratt	Colin Trevorrow	Action	http://www.jurassicworld.com/

```

2 Chris Pratt Colin Trevorrow Action http://www.jurassicworld.com/
3 Chris Pratt Colin Trevorrow Action http://www.jurassicworld.com/
4 Chris Pratt Colin Trevorrow Action http://www.jurassicworld.com/

```

```

      id      imdb_id      keywords original_title \
0  135397  tt0369610      monster  Jurassic World
1  135397  tt0369610      dna      Jurassic World
2  135397  tt0369610  tyrannosaurus rex  Jurassic World
3  135397  tt0369610      velociraptor  Jurassic World
4  135397  tt0369610      island      Jurassic World

```

```

                                overview popularity \
0  Twenty-two years after the events of Jurassic ...      32.99
1  Twenty-two years after the events of Jurassic ...      32.99
2  Twenty-two years after the events of Jurassic ...      32.99
3  Twenty-two years after the events of Jurassic ...      32.99
4  Twenty-two years after the events of Jurassic ...      32.99

```

```

                                production_companies release_date \
0  Universal Studios|Amblin Entertainment|Legenda...  2015-06-09
1  Universal Studios|Amblin Entertainment|Legenda...  2015-06-09
2  Universal Studios|Amblin Entertainment|Legenda...  2015-06-09
3  Universal Studios|Amblin Entertainment|Legenda...  2015-06-09
4  Universal Studios|Amblin Entertainment|Legenda...  2015-06-09

```

```

      release_year  runtime      tagline
0           2015      124  The park is open.
1           2015      124  The park is open.
2           2015      124  The park is open.
3           2015      124  The park is open.
4           2015      124  The park is open.

```

```

In [20]: df_vals = df_flat.drop(columns=['genres', 'cast', 'keywords'])
df_vals.drop_duplicates(inplace=True)
df_vals.head()

```

```

Out[20]:      budget      budget_adj      director \
0    150000000  137,999,939.28  Colin Trevorrow
100  150000000  137,999,939.28    George Miller
200  110000000  101,199,955.47  Robert Schwentke
275  200000000  183,999,919.04    J.J. Abrams
375  190000000  174,799,923.09    James Wan

```

```

                                homepage      id      imdb_id \
0                                http://www.jurassicworld.com/  135397  tt0369610
100                             http://www.madmaxmovie.com/    76341  tt1392190
200    http://www.thedivergentseries.movie/#insurgent  262500  tt2908446
275    http://www.starwars.com/films/star-wars-episod...  140607  tt2488496

```

```
375                                     http://www.furious7.com/ 168259 tt2820852
```

	net_income	original_title \
0	1,254,445,953.24	Jurassic World
100	210,161,353.21	Mad Max: Fury Road
200	170,419,069.94	Insurgent
275	1,718,723,210.76	Star Wars: The Force Awakens
375	1,210,948,878.38	Furious 7

	overview	popularity \
0	Twenty-two years after the events of Jurassic ...	32.99
100	An apocalyptic story set in the furthest reach...	28.42
200	Beatrice Prior must confront her inner demons ...	13.11
275	Thirty years after defeating the Galactic Empi...	11.17
375	Deckard Shaw seeks revenge against Dominic Tor...	9.34

	production_companies	release_date \
0	Universal Studios Amblin Entertainment Legenda...	2015-06-09
100	Village Roadshow Pictures Kennedy Miller Produ...	2015-05-13
200	Summit Entertainment Mandeville Films Red Wago...	2015-03-18
275	Lucasfilm Truenorth Productions Bad Robot	2015-12-15
375	Universal Pictures Original Film Media Rights ...	2015-04-01

	release_year	revenue	revenue_adj	runtime \
0	2015	1513528810	1,392,445,892.52	124
100	2015	378436354	348,161,292.49	120
200	2015	295238201	271,619,025.41	119
275	2015	2068178225	1,902,723,129.80	136
375	2015	1506249360	1,385,748,801.47	137

	tagline	vote_average	vote_count
0	The park is open.	6.50	5562
100	What a Lovely Day.	7.10	6185
200	One Choice Can Destroy You	6.30	2480
275	Every generation has a story.	7.50	5292
375	Vengeance Hits Home	7.30	2947

Exploratory Data Analysis

```
In [21]: # The longest movie
```

```
idx = df_vals.runtime.idxmax()
df_vals.loc[idx]['original_title']
```

```
Out[21]: 'The Lord of the Rings: The Return of the King'
```

```
In [22]: # The movie have highest average vote
```

```
idx = df_vals.vote_average.idxmax()
df_vals.loc[idx]['original_title']
```

```
Out[22]: 'The Godfather'
```

```
In [23]: # The movie have largest votes
        idx = df_vals.vote_count.idxmax()
        df_vals.loc[idx]['original_title']
```

```
Out[23]: 'Inception'
```

1.3.1 Research Question 1: Which movies were most and least profitable?

```
In [24]: idx = df_vals.net_income.idxmax()
        idx
```

```
Out[24]: 9910
```

```
In [25]: # The most profitable movie
        df_vals.loc[idx]
```

```
Out[25]: budget                11000000
        budget_adj              39,575,591.36
        director                George Lucas
        homepage                http://www.starwars.com/films/star-wars-episod...
        id                      11
        imdb_id                 tt0076759
        net_income              2,750,136,650.92
        original_title          Star Wars
        overview                Princess Leia is captured and held hostage by ...
        popularity              12.04
        production_companies    Lucasfilm|Twentieth Century Fox Film Corporation
        release_date            1977-03-20 00:00:00
        release_year            1977
        revenue                 775398007
        revenue_adj             2,789,712,242.28
        runtime                 121
        tagline                 A long time ago in a galaxy far, far away...
        vote_average            7.90
        vote_count              4428
        Name: 9910, dtype: object
```

```
In [26]: idx1 = df_vals.net_income.idxmin()
        idx1
```

```
Out[26]: 23771
```

```
In [27]: # The lease profitable movie
        df_vals.loc[idx1]
```

```
Out[27]: budget                425000000
        budget_adj              425,000,000.00
```

```

director                Sngmoo Lee
homepage                http://www.iamrogue.com/thewarriorsway
id                      46528
imdb_id                 tt1032751
net_income              -413,912,431.00
original_title          The Warrior's Way
overview                An Asian assassin (Dong-gun Jang) is forced to...
popularity              0.25
production_companies    Boram Entertainment Inc.
release_date            2010-12-02 00:00:00
release_year            2010
revenue                11087569
revenue_adj             11,087,569.00
runtime                100
tagline                 Assassin. Hero. Legend.
vote_average            6.40
vote_count              74
Name: 23771, dtype: object

```

1.3.2 Research Question 2: What was the most profitable movie each year?

```

In [28]: idx = df.groupby(['release_year'])['net_income'].transform(max) == df['net_income']
df[idx].sort_values(['release_year'], ascending=True)[['id', 'original_title', 'release_year', 'budget_adj', 'net_income']]

```

```

Out[28]:
   id  original_title  release_year \
10143   967      Spartacus          1960
10110  12230  One Hundred and One Dalmatians  1961
 9849   646      Dr. No          1962
10438   657  From Russia With Love          1963
 9881   658      Goldfinger          1964
10690  15121  The Sound of Music          1965
10822   396  Who's Afraid of Virginia Woolf?          1966
10398  9325      The Jungle Book          1967
 9719    62  2001: A Space Odyssey          1968
10725   642  Butch Cassidy and the Sundance Kid          1969
10654  9062      Love Story          1970
 9925   681  Diamonds Are Forever          1971
 7269   238      The Godfather          1972
10594  9552      The Exorcist          1973
 9767  11072  Blazing Saddles          1974
 9806   578      Jaws          1975
10208  19610  A Star Is Born          1976
 1329    11      Star Wars          1977
10758  1924      Superman          1978
 7833  1367      Rocky II          1979
 7309  1891  The Empire Strikes Back          1980
 8375    85  Raiders of the Lost Ark          1981

```

8889	601	E.T. the Extra-Terrestrial	1982
7987	1892	Return of the Jedi	1983
7883	87	Indiana Jones and the Temple of Doom	1984
6081	105	Back to the Future	1985
10475	744	Top Gun	1986
9613	10998	Fatal Attraction	1987
9454	380	Rain Man	1988
9180	89	Indiana Jones and the Last Crusade	1989
9986	251	Ghost	1990
9317	280	Terminator 2: Judgment Day	1991
8243	812	Aladdin	1992
10223	329	Jurassic Park	1993
4180	8587	The Lion King	1994
8094	1642	The Net	1995
8457	602	Independence Day	1996
5231	597	Titanic	1997
8970	95	Armageddon	1998
2412	1893	Star Wars: Episode I - The Phantom Menace	1999
8666	955	Mission: Impossible II	2000
2634	671	Harry Potter and the Philosopher's Stone	2001
3911	121	The Lord of the Rings: The Two Towers	2002
4949	122	The Lord of the Rings: The Return of the King	2003
6977	809	Shrek 2	2004
6190	674	Harry Potter and the Goblet of Fire	2005
6555	58	Pirates of the Caribbean: Dead Man's Chest	2006
7388	675	Harry Potter and the Order of the Phoenix	2007
2875	155	The Dark Knight	2008
1386	19995	Avatar	2009
1930	10193	Toy Story 3	2010
3374	12445	Harry Potter and the Deathly Hallows: Part 2	2011
4361	24428	The Avengers	2012
5422	109445	Frozen	2013
634	122917	The Hobbit: The Battle of the Five Armies	2014
3	140607	Star Wars: The Force Awakens	2015

	budget_adj	net_income
10143	88,475,609.49	353,902,437.95
10110	29,179,444.83	1,545,635,294.87
9849	7,929,293.77	421,694,259.41
10438	17,800,448.43	543,972,910.57
9881	24,605,935.94	853,474,463.61
10690	56,748,622.29	1,072,786,239.70
10822	50,385,110.19	176,258,462.18
10398	26,147,054.96	1,319,404,004.03
9719	75,227,563.38	280,319,033.83
10725	35,665,585.21	572,485,481.13
10654	12,356,010.36	753,716,631.75
9925	38,773,403.38	585,909,206.69

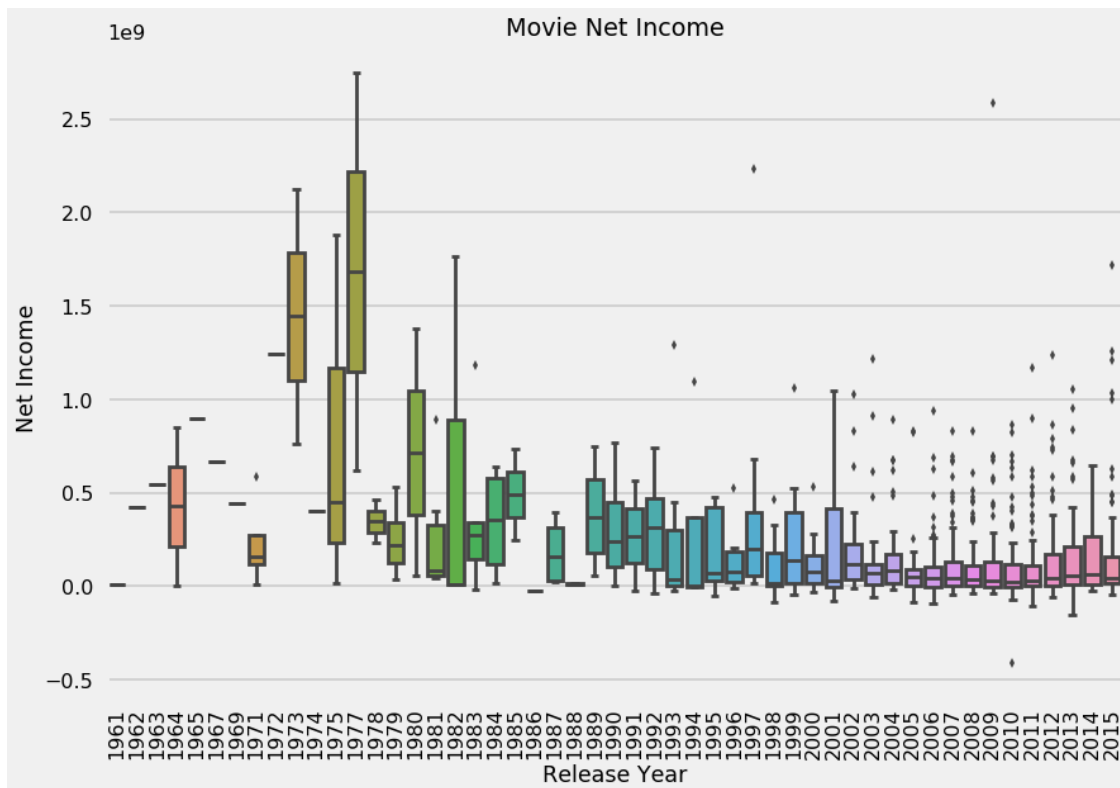
7269	31,287,365.59	1,246,626,366.80
10594	39,289,276.63	2,128,035,624.57
9767	11,497,938.11	516,964,986.57
9806	28,362,748.20	1,878,643,093.71
10208	22,990,188.17	593,913,194.41
1329	39,575,591.36	2,750,136,650.92
10758	183,848,538.22	819,690,439.17
7833	21,031,878.02	580,426,232.04
7309	47,628,661.55	1,376,997,526.22
8375	43,167,434.49	891,949,443.74
8889	23,726,245.23	1,767,968,064.02
7987	70,824,243.13	1,182,994,737.99
7883	58,773,177.10	640,207,821.95
6081	38,516,154.99	734,056,616.72
10475	29,841,096.16	680,039,989.06
9613	26,867,126.13	587,518,210.63
9454	46,097,275.58	608,162,158.78
9180	84,431,277.07	749,629,788.20
9986	36,715,767.89	806,077,994.97
9317	160,109,284.19	672,458,993.61
8243	43,512,679.13	739,793,586.73
10223	95,096,607.59	1,293,766,704.17
4180	66,200,020.27	1,093,391,569.74
8094	31,481,271.08	1,551,568,265.28
8457	104,266,255.42	1,031,498,096.17
5231	271,692,064.21	2,234,713,671.21
8970	187,277,365.67	553,537,804.54
2412	150,541,077.36	1,059,439,453.10
8666	158,286,514.29	533,600,434.44
2634	153,936,014.59	1,048,582,021.48
3911	95,768,650.10	1,027,133,804.32
4949	111,423,148.61	1,214,854,861.87
6977	173,166,809.97	888,736,812.77
6190	167,484,493.45	832,868,045.78
6555	216,333,831.66	936,357,520.21
7388	157,750,287.39	828,938,572.98
2875	187,365,527.25	827,367,505.23
1386	240,886,902.89	2,586,236,847.52
1930	200,000,000.00	863,171,911.00
3374	121,174,755.32	1,166,009,242.24
4361	208,943,741.90	1,234,247,693.31
5422	140,405,002.91	1,052,306,488.21
634	230,272,762.69	649,479,526.45
3	183,999,919.04	1,718,723,210.76

```
In [29]: # What is the distribution of net income by year?
ax = sns.boxplot(x='release_year',
                 y='net_income',
```

```

data=df_vals)
ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
plt.xlabel('Release Year')
plt.ylabel('Net Income')
plt.title('Movie Net Income')
plt.show()

```

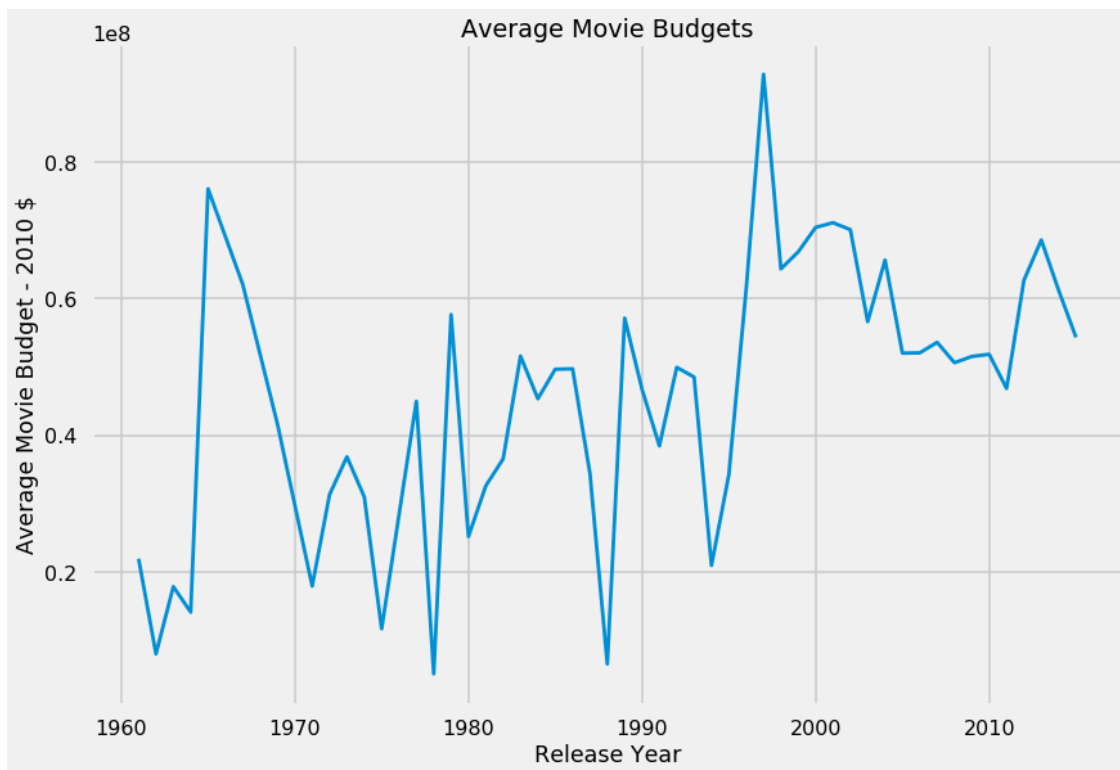


Research Question 3: Do movie budgets increase over time?

```

In [30]: # What is the average budget of the movies released each year?
plt.plot(df_vals[df_vals['budget_adj'] > 0].groupby('release_year').budget_adj.mean())
plt.xlabel('Release Year')
plt.ylabel('Average Movie Budget - 2010 $')
plt.title('Average Movie Budgets')
plt.show()

```



```
In [31]: # What were the biggest movie budgets of all time?
df_vals.sort_values(['budget_adj'], ascending=False)[['id', 'original_title', 'release_
']].head(10)
```

```
Out[31]:
```

	id	original_title	release_year	\
23771	46528	The Warrior's Way	2010	
31499	1865	Pirates of the Caribbean: On Stranger Tides	2011	
65204	285	Pirates of the Caribbean: At World's End	2007	
49789	597	Titanic	1997	
65804	559	Spider-Man 3	2007	
17590	38757	Tangled	2010	
1200	99861	Avengers: Age of Ultron	2015	
10260	767	Harry Potter and the Half-Blood Prince	2009	
17140	12444	Harry Potter and the Deathly Hallows: Part 1	2010	
45094	49529	John Carter	2012	

	budget_adj	net_income
23771	425,000,000.00	-413,912,431.00
31499	368,371,256.18	622,046,244.16
65204	315,500,574.79	695,152,933.12
49789	271,692,064.21	2,234,713,671.21
65804	271,330,494.32	665,571,205.90
17590	260,000,000.00	331,794,936.00

```

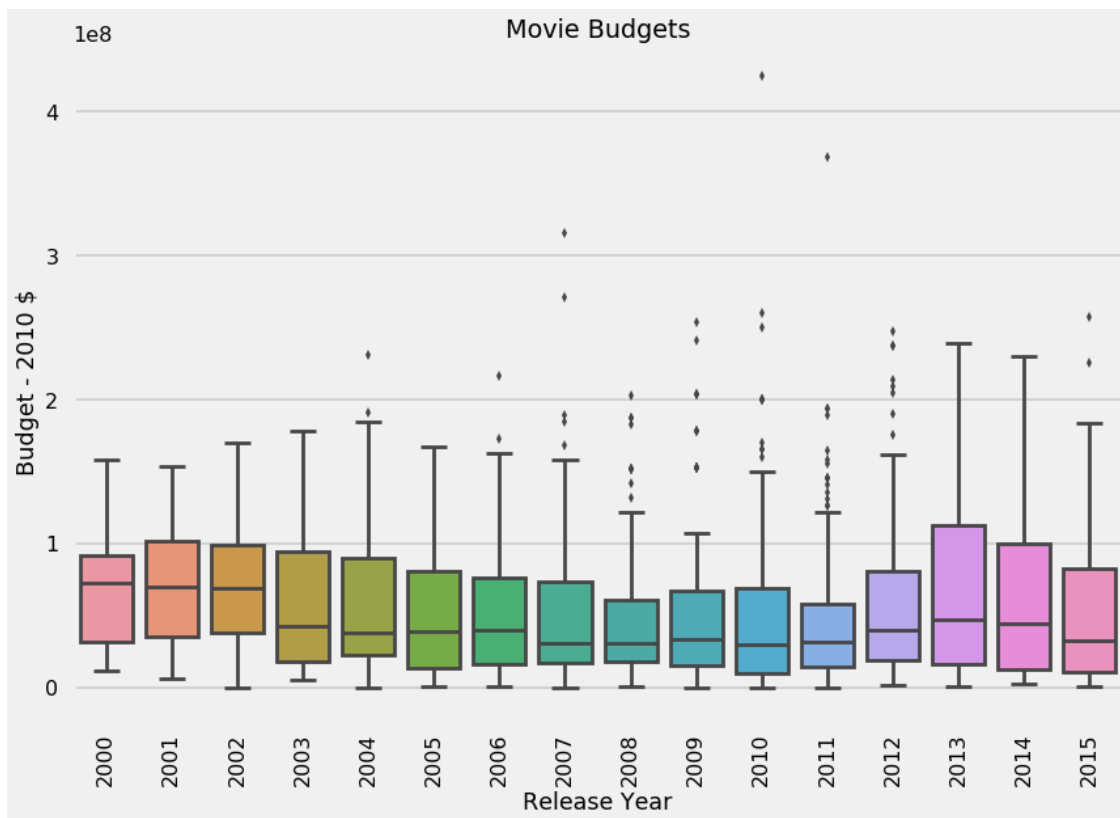
1200 257,599,886.66 1,035,032,450.23
10260 254,100,108.53 695,176,424.75
17140 250,000,000.00 704,305,868.00
45094 246,933,513.15 22,925,972.18

```

```

In [32]: # Drop movies with >0 budget, only year 2000+
df_vals2000 = df_vals.query('budget_adj > 0 & release_year >= 2000')
ax = sns.boxplot(x='release_year',
                 y='budget_adj',
                 data=df_vals2000)
ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
plt.xlabel('Release Year')
plt.ylabel('Budget - 2010 $')
plt.title('Movie Budgets')
plt.show()

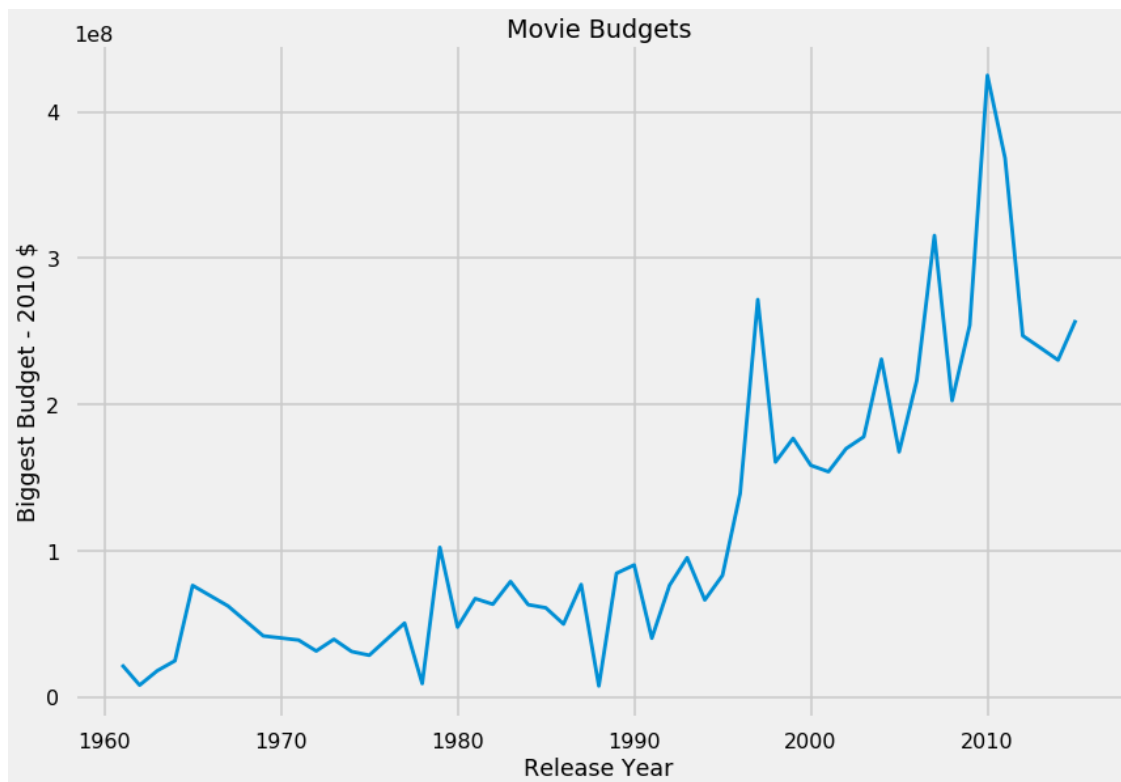
```



```

In [33]: # What was the most expensive movie released each year?
plt.plot(df_vals[df_vals['budget_adj'] > 0].groupby('release_year').budget_adj.max())
plt.xlabel('Release Year')
plt.ylabel('Biggest Budget - 2010 $')
plt.title('Movie Budgets')
plt.show()

```



1.3.3 Research Question 4: What is the most profitable genre?

```
In [34]: # Join the two dfs on the id key
result = pd.merge(df_film, df_vals, on='id', how='inner')
```

```
In [35]: result = result[['id', 'original_title_x', 'genres', 'revenue_adj', 'budget_adj', 'net_income']]

result.rename(columns={'original_title_x': 'original_title'}, inplace=True)

# We have a bunch of duplicates from the expanded set, remove them
result.drop_duplicates(inplace=True)
result.head()
```

```
Out[35]:
```

	id	original_title	genres	revenue_adj \
0	135397	Jurassic World	Action	1,392,445,892.52
5	135397	Jurassic World	Adventure	1,392,445,892.52
10	135397	Jurassic World	Science Fiction	1,392,445,892.52
15	135397	Jurassic World	Thriller	1,392,445,892.52
100	76341	Mad Max: Fury Road	Action	348,161,292.49

		budget_adj	net_income
0	137,999,939.28	1,254,445,953.24	
5	137,999,939.28	1,254,445,953.24	

```

10  137,999,939.28  1,254,445,953.24
15  137,999,939.28  1,254,445,953.24
100 137,999,939.28   210,161,353.21

```

```
In [36]: result.groupby('genres').net_income.sum().sort_values(ascending=False)
```

```

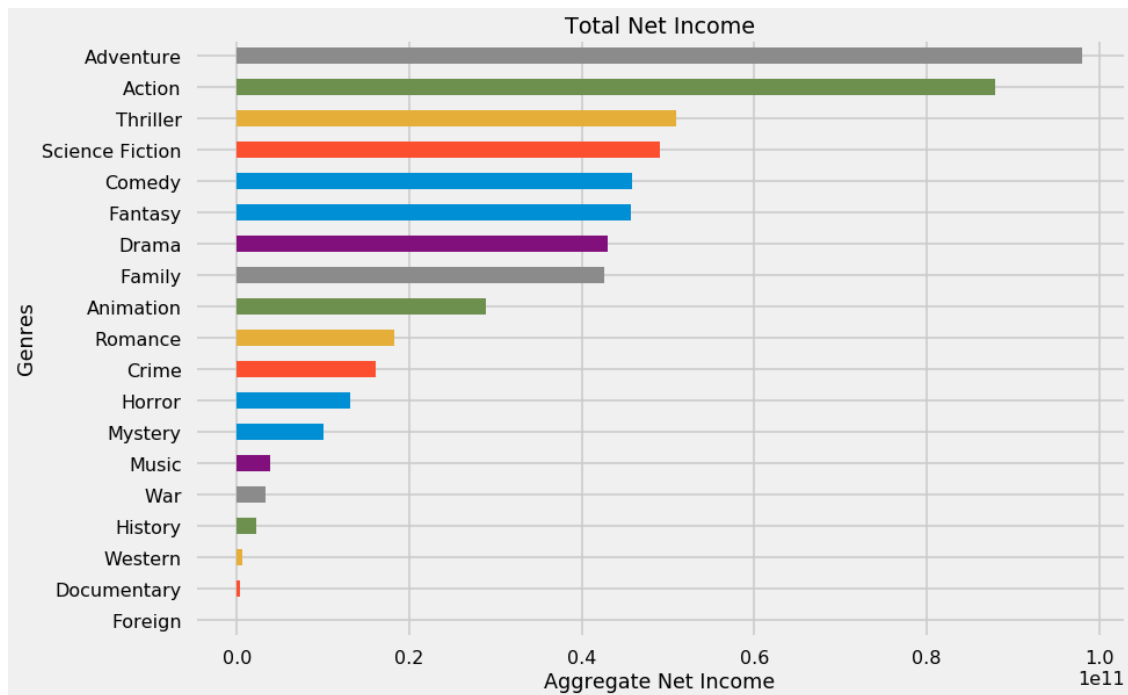
Out[36]: genres
Adventure      98,042,386,342.66
Action         87,874,227,105.53
Thriller       50,903,960,105.73
Science Fiction 49,058,392,574.84
Comedy         45,795,267,298.12
Fantasy        45,715,250,965.99
Drama          43,013,208,139.53
Family         42,587,558,778.80
Animation      28,823,610,050.17
Romance        18,211,014,055.05
Crime          16,132,383,057.04
Horror         13,179,400,828.52
Mystery        10,054,564,194.13
Music          3,881,078,931.43
War            3,342,155,586.45
History        2,288,072,621.76
Western        573,527,954.54
Documentary    309,092,494.58
Foreign        -1,312,284.00
Name: net_income, dtype: float64

```

```

In [37]: # Total income
result.groupby('genres')['net_income'].sum().sort_values(ascending=True).plot.barh(title=
plt.xlabel('Aggregate Net Income')
plt.ylabel('Genres')
plt.show()

```

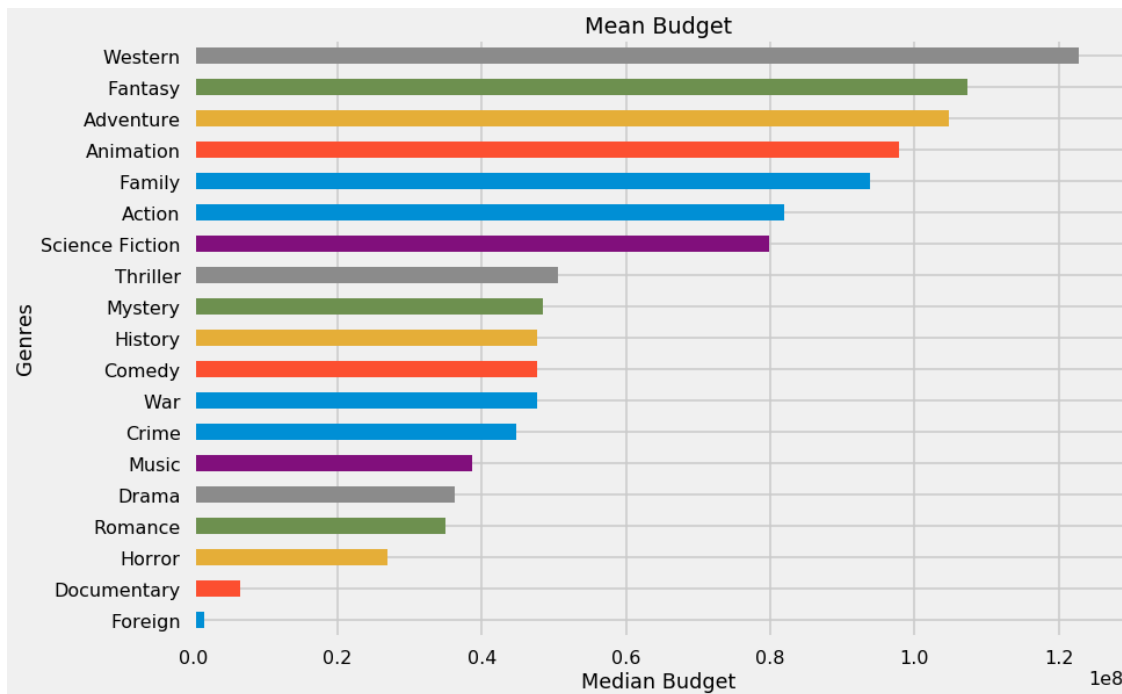


```
In [38]: result.groupby('genres').net_income.mean().sort_values(ascending=False)
```

```
Out[38]: genres
Adventure      330,109,044.92
Fantasy        296,852,279.00
Animation      266,885,278.24
Family         261,273,366.74
Science Fiction 234,729,151.08
Action         211,745,125.56
Thriller       127,578,847.38
Comedy         110,616,587.68
War            104,442,362.08
Mystery        102,597,593.82
Music          102,133,656.09
Horror         94,815,833.30
Romance        92,913,337.02
Crime          83,587,476.98
Drama          80,398,519.89
History        65,373,503.48
Western        44,117,534.96
Documentary    20,606,166.31
Foreign        -1,312,284.00
Name: net_income, dtype: float64
```

```
In [39]: result.groupby('genres')['budget_adj'].mean().sort_values(ascending=True).plot.barh(title='Median Budget')
plt.xlabel('Median Budget')
```

```
plt.ylabel('Genres')
plt.show()
```



Conclusions

1.3.4 Which movies were most and least profitable?

The five most profitable movies in history (using inflation-adjusted 2010 dollars) are: (1) *Star Wars* (2) *Avatar* (3) *Titanic* (4) *The Exorcist* (5) *Jaws*. The least profitable movies are (1) *The Lone Ranger* (2) *Mars Needs Moms* (3) *Flushed Away* (4) *Sphere* (5) *A Sound of Thunder*.

1.3.5 What was the most profitable movie each year?

Above I list the most profitable movies for every year from 1960 to 2015, in inflation-adjusted 2010 dollars. They ranged from *Spartacus* (1960) to *Star Wars: The Force Awakens* (2015). It was interesting to note that three of the most five most profitable movies of all time were released in the 1970s. Only 22 movies in history have made over one billion dollars and four of them were released in the 1970s (the three just listed as well as *The Godfather*).

1.3.6 Have movie budgets increased over time?

There's an interesting answer to this question: sort of. The median movie budget has actually declined since 2000. But the budgets for the biggest budget movies have increased significantly, increasing by 60% since 2000.

1.3.7 What is the most profitable genre?

The most profitable genres – measured by total profits across the entire category – are Adventure, Action, and Thrillers. Looking at median profitability, the categories are Adventure, Animation, and Fantasy. These two lists are different because median controls for the number of movies released. Although Animation and Fantasy are profitable categories, fewer of those movies are produced. I saw similar effects when I looked at budgets.

Limitations: Several movies had 0 for budget and/or revenue. There might be more or less profitable movies where the data was missing.

```
In [40]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

```
Out[40]: 0
```