Stony Brook University

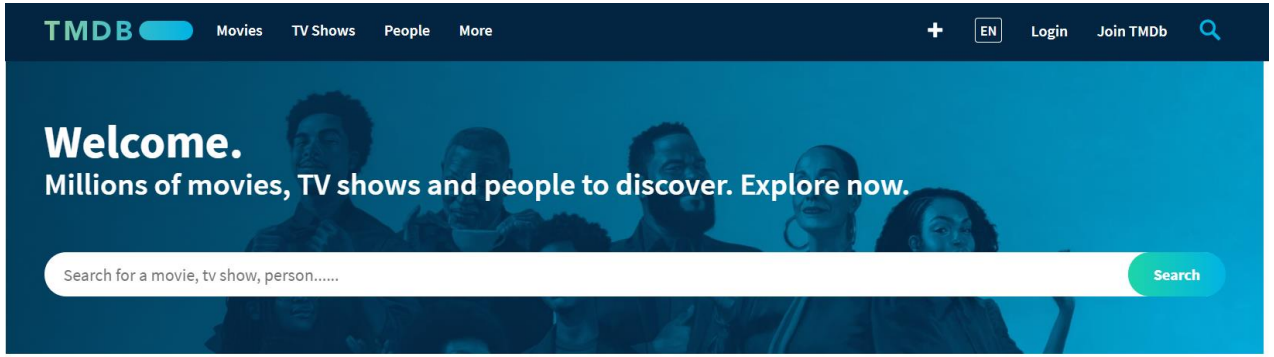
Project #1: Movie Revenue Prediction

Stony Brook University
CSE 351 Summer 2021 Final Project

Jeong Yoon Lee
Computer Science & Applied Mathematics and Statistics

INDEX

01. TMDB Dataset



https://www.themoviedb.org/

#	Column
0	budget
1	genres
2	homepage
3	id
4	keywords
5	original_language
6	original_title
7	overview
8	popularity
9	production_companies
10	production_countries
11	release_date
12	revenue
13	runtime
14	spoken_languages
15	status
16	tagline
17	title
18	vote_average
19	vote_count

<tmdb_5000_movies.csv>

#	Column
0	id
1	cast
2	crew

<tmdb_5000_credits.csv>

→ The goal of this project is to predict revenue through related features!

02. Data Pre-Processing

Merge Datasets

```
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   budget                4803 non-null   int64
 1   genres                4803 non-null   object
 2   homepage              1712 non-null   object
 3   id                    4803 non-null   int64
 4   keywords              4803 non-null   object
 5   original_language     4803 non-null   object
 6   original_title        4803 non-null   object
 7   overview              4800 non-null   object
 8   popularity            4803 non-null   float64
 9   production_companies  4803 non-null   object
 10  production_countries  4803 non-null   object
 11  release_date          4802 non-null   object
 12  revenue               4803 non-null   int64
 13  runtime               4801 non-null   float64
 14  spoken_languages      4803 non-null   object
 15  status                4803 non-null   object
 16  tagline               3959 non-null   object
 17  title                 4803 non-null   object
 18  vote_average          4803 non-null   float64
 19  vote_count            4803 non-null   int64
```

<tmdb_5000_movies.csv>

+

```
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      4803 non-null   int64
 1   cast    4803 non-null   object
 2   crew    4803 non-null   object
dtypes: int64(1), object(2)
```

<tmdb_5000_credits.csv>

Merge →

```
Data columns (total 22 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   budget                4803 non-null   int64
 1   genres                4803 non-null   object
 2   homepage              1712 non-null   object
 3   id                    4803 non-null   int64
 4   keywords              4803 non-null   object
 5   original_language     4803 non-null   object
 6   original_title        4803 non-null   object
 7   overview              4800 non-null   object
 8   popularity            4803 non-null   float64
 9   production_companies  4803 non-null   object
 10  production_countries  4803 non-null   object
 11  release_date          4802 non-null   object
 12  revenue               4803 non-null   int64
 13  runtime               4801 non-null   float64
 14  spoken_languages      4803 non-null   object
 15  status                4803 non-null   object
 16  tagline               3959 non-null   object
 17  title                 4803 non-null   object
 18  vote_average          4803 non-null   float64
 19  vote_count            4803 non-null   int64
 20  cast                  4803 non-null   object
 21  crew                  4803 non-null   object
dtypes: float64(3), int64(4), object(15)
```

02. Data Pre-Processing

Cleaning Dataset – Null Value

```
budget                   0
genres                   0
homepage              3091
id                       0
keywords                 0
original_language        0
original_title           0
overview                 3
popularity               0
production_companies     0
production_countries     0
release_date             1
revenue                  0
runtime                  2
spoken_languages         0
status                   0
tagline                844
title                    0
vote_average             0
vote_count               0
cast                     0
crew                     0
dtype: int64
```

```python
#fill "NaN" to null value in homepage, overview, tagline columns
data["homepage"] = data["homepage"].fillna("NaN")
data["overview"] = data["overview"].fillna("NaN")
data["tagline"] = data["tagline"].fillna("NaN")

#check the number of null value for each columns.
data.isnull().sum()
```

```
budget                   0
genres                   0
homepage                 0
id                       0
keywords                 0
original_language        0
original_title           0
overview                 0
popularity               0
production_companies     0
production_countries     0
release_date             1
revenue                  0
runtime                  2
spoken_languages         0
status                   0
tagline                  0
title                    0
vote_average             0
vote_count               0
cast                     0
crew                     0
dtype: int64
```

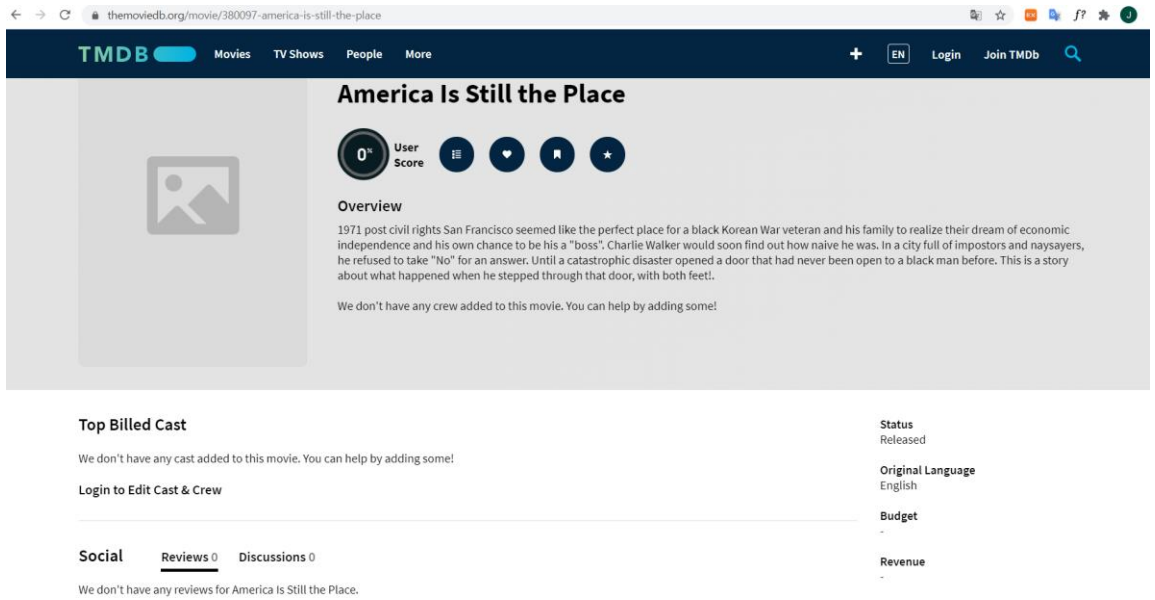
02. Data Pre-Processing

Cleaning Dataset – Null Value

budget	0
genres	0
homepage	3091
id	0
keywords	0
original_language	0
original_title	0
overview	3
popularity	0
production_companies	0
production_countries	0
release_date	1
revenue	0
runtime	2
spoken_languages	0
status	0
tagline	844
title	0
vote_average	0
vote_count	0
cast	0
crew	0
dtype: int64	

budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_companies	production_countries	release_date	revenue	runtime	spoken_languages	status	tagline	title
0	[]	NaN	380097	[]	en	America Is Still the Place	1971 post civil rights San Francisco seemed li...	0.0	[]	[]	NaN	0	0.0	[]	Released	NaN	America Is Still the Place

vote_average	vote_count	cast	crew
0.0	0	[]	[]



https://www.themoviedb.org/movie/380097-america-is-still-the-place

No Information at All!

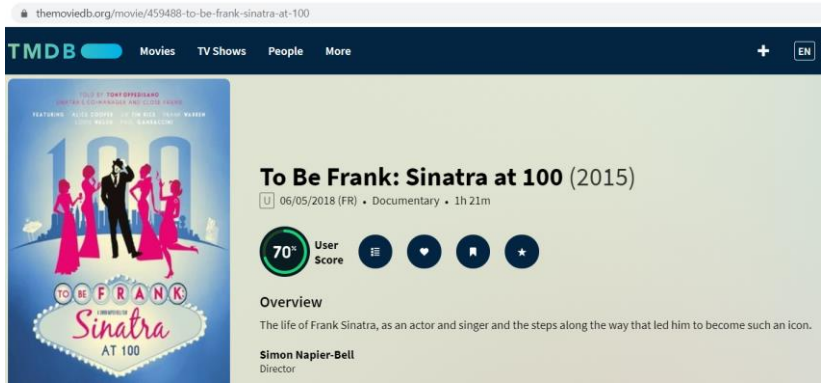
→ Decide to drop this Row

02. Data Pre-Processing

Cleaning Dataset – Null Value

```
budget                     0
genres                     0
homepage                3091
id                         0
keywords                   0
original_language          0
original_title             0
overview                   3
popularity                 0
production_companies       0
production_countries       0
release_date               1
revenue                    0
runtime                    2
spoken_languages           0
status                     0
tagline                  844
title                      0
vote_average               0
vote_count                 0
cast                       0
crew                       0
dtype: int64
```

	title	runtime
2656	Chiamatemi Francesco - Il Papa della gente	NaN
4140	To Be Frank, Sinatra at 100	NaN



https://www.themoviedb.org/movie/459488-to-be-frank-sinatra-at-100
https://www.themoviedb.org/movie/370980-chiamatemi-francesco

Fill the runtime information from TMDB website!

```
data.at[2656, 'runtime'] = 98
data.loc[2656]
```

```
data.at[4140, 'runtime'] = 81
data.loc[4140]
```

↓

	title	runtime
2656	Chiamatemi Francesco - Il Papa della gente	98.0
4140	To Be Frank, Sinatra at 100	81.0

02. Data Pre-Processing

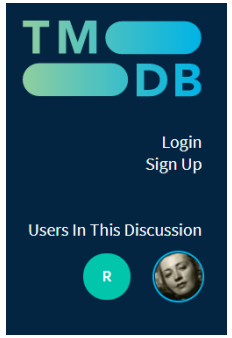
Cleaning Dataset – Missing Value

```
data[(data['revenue'] ==0)]
```

309	84000000	[{"id": 14, "name": "Fantasy"}, {"id": 35, "na...	NaN	10214	[{"id": 1009, "name": "baby"}, {"id": 2546, "n...	en	Son of the Mask	Tim Avery, an aspiring cartoonist, finds himse...	17.815595	[{"name": "New Line Cinema", "id": 12}, {"name...	[{"iso_3166_1": "DE", "name": "Germany"}, {"is...	2005-02-18	0	94.0
376	90000000	[{"id": 878, "name": "Science Fiction"}, {"id"...	NaN	10357	[{"id": 1552, "name": "subway"}, {"id": 2859, ...	en	Volcano	An earthquake shatters a peaceful Los Angeles ...	19.836124	[{"name": "Twentieth Century Fox Film Corporat...	[{"iso_3166_1": "US", "name": "United States o...	1997-04-25	0	104.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4797	0	[{"id": 10769, "name": "Foreign"}, {"id": 53, ...	NaN	67238	[]	en	Cavite	Adam, a security guard, travels from Californi...	0.022173	[]	[]	2005-03-12	0	80.0
4799	9000	[{"id": 35, "name": "Comedy"}, {"id": 10749, "...	NaN	72766	[]	en	Newlyweds	A newlywed couple's honeymoon is upended by th...	0.642552	[]	[]	2011-12-26	0	85.0
4800	0	[{"id": 35, "name": "Comedy"}, {"id": 18, "nam...	http://www.hallmarkchannel.com/signedsealeddel...	231617	[{"id": 248, "name": "date"}, {"id": 699, "nam...	en	Signed, Sealed, Delivered	"Signed, Sealed, Delivered" introduces a dedic...	1.444476	[{"name": "Front Street Pictures", "id": 3958}...	[{"iso_3166_1": "US", "name": "United States o...	2013-10-13	0	120.0
4801	0	[]	http://shanghaicalling.com/	126186	[]	en	Shanghai Calling	When ambitious New York attorney Sam is sent t...	0.857008	[]	[{"iso_3166_1": "US", "name": "United States o...	2012-05-03	0	98.0
4802	0	[{"id": 99, "name": "Documentary"}]	NaN	25975	[{"id": 1523, "name": "obsession"}, {"id": 224...	en	My Date with Drew	Ever since the second grade when he first saw ...	1.929883	[{"name": "rusty bear entertainment", "id": 87...	[{"iso_3166_1": "US", "name": "United States o...	2005-08-05	0	90.0

1426 rows × 22 columns

→ 1426 rows out of 4803 rows! Too Much!



Reply by **lineker** MOD
on September 24, 2018 at 6:26 PM

Hi, I'll answer your questions below.

> some films have their budgets listed as 0

This just means that the data has not been entered yet. Feel free to contribute if you know the correct number.

https://www.themoviedb.org/talk/5ba87d119251412f0103e87b

→ 1426/4803 is too large!

- Goal of this analysis is predicting revenue.
- But it is too dangerous to impute these values to other values since they have all different feature values!
- So, I decided to drop those rows and I will predict these data with my model.

02. Data Pre-Processing

Cleaning Dataset – Missing Value

```
[18] data_with_revenue[(data_with_revenue['budget'] ==0)]
```

475	0	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	NaN	9433	[{"id": 1003, "name": "photographer"}, {"id": ...	en	The Edge	The plane carrying wealthy Charles Morse crash...	20.632673	[{"name": "Art Linson Productions", "id": 8769}]	[{"iso_3166_1": "US", "name": "United States o...	1997-09-06	43312294	117.0	
489	0	[{"id": 99, "name": "Documentary"}, {"id": 107...	http://oceans-lefilm.com/	36970	[{"id": 270, "name": "ocean"}, {"id": 658, "na...	en	Oceans	An ecological drama/documentary, filmed throug...	10.706613	[{"name": "Path₩u00e9 Films", "id": 4959}, {"n...	[{"iso_3166_1": "FR", "name": "France"}, {"iso...	2009-10-17	19406406	84.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4605	0	[{"id": 18, "name": "Drama"}]	http://www.dogtooth.gr/	38810	[{"id": 255, "name": "male nudity"}, {"id": 29...	el	Κυνόδοντας	Three teenagers are confined to an isolated co...	28.858238	[{"name": "Greek Film Center", "id": 7254}, {"...	[{"iso_3166_1": "GR", "name": "Greece"}]	2009-06-01	110197	94.0	"₩u
4630	0	[{"id": 18, "name": "Drama"}, {"id": 53, "name...	http://www.magpictures.com/compliance/	84188	[]	en	Compliance	Sandra is the manager at a fast-food restauran...	8.942203	[{"name": "Muskat Filmed Properties", "id": 10...	[{"iso_3166_1": "US", "name": "United States o...	2012-08-23	319285	90.0	
4677	0	[{"id": 10749, "name": "Romance"}, {"id": 18, ...	NaN	53256	[{"id": 572, "name": "sex"}, {"id": 154937, "n...	de	Drei	Hanna and Simon are in a 20 year marriage with...	5.937602	[{"name": "X-Filme Creative Pool", "id": 1972}...	[{"iso_3166_1": "DE", "name": "Germany"}]	2010-12-23	2611555	119.0	
4766	0	[{"id": 99, "name": "Documentary"}, {"id": 104...	http://www.mgm.com/#/our-titles/1092/The-Last-...	13963	[{"id": 1228, "name": "1970s"}, {"id": 6027, "...	en	The Last Waltz	Martin Scorsese's rockumentary intertwines foo...	3.277287	[{"name": "FM Productions", "id": 12601}, {"na...	[{"iso_3166_1": "US", "name": "United States o...	1978-05-01	321952	117.0	
4775	0	[{"id": 18, "name": "Drama"}, {"id": 35, "name...	NaN	33693	[{"id": 171993, "name": "mumblecore"}]	en	Funny Ha Ha	Unsure of what to do next, 23-year-old Marnie ...	0.362633	[]	[{"iso_3166_1": "US", "name": "United States o...	2002-09-20	76901	85.0	

147 rows × 22 columns

→ 147 rows out of 4803 rows!

→ Budget is important feature to predict revenue

- We should use budget features to make revenue prediction model.
- Similarly, these data don't have any common features, so I decide to drop these rows to make it clear for predicting revenue

02. Data Pre-Processing

Cleaning Dataset – Reformat Data

→ Changed Dictonary Data to List

	genres	keywords	production_companies	production_countries	spoken_languages	cast	crew
0	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"name": "Ingenious Film Partners", "id": 289...	[{"iso_3166_1": "US", "name": "United States o...	[{"iso_639_1": "en", "name": "English"}, {"iso...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	[{"name": "Walt Disney Pictures", "id": 2}, {"...	[{"iso_3166_1": "US", "name": "United States o...	[{"iso_639_1": "en", "name": "English"}]	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 470, "name": "spy"}, {"id": 818, "name...	[{"name": "Columbia Pictures", "id": 5}, {"nam...	[{"iso_3166_1": "GB", "name": "United Kingdom"...	[{"iso_639_1": "fr", "name": "Fran\u00e7ais"},...	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	[{"id": 849, "name": "dc comics"}, {"id": 853,...	[{"name": "Legendary Pictures", "id": 923}, {"...	[{"iso_3166_1": "US", "name": "United States o...	[{"iso_639_1": "en", "name": "English"}]	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 818, "name": "based on novel"}, {"id":...	[{"name": "Walt Disney Pictures", "id": 2}]	[{"iso_3166_1": "US", "name": "United States o...	[{"iso_639_1": "en", "name": "English"}]	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...
...	...	...	...	...	...	...	...

	genres	keywords	production_companies	production_countries	spoken_languages	cast	crew
0	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	[Ingenious Film Partners, Twentieth Century Fo...	[United States of America, United Kingdom]	[en, es]	[Sam Worthington, Zoe Saldana, Sigourney Weave...	[Stephen E. Rivkin, Rick Carter, Christopher B...
1	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	[Walt Disney Pictures, Jerry Bruckheimer Films...	[United States of America]	[en]	[Johnny Depp, Orlando Bloom, Keira Knightley, ...	[Dariusz Wolski, Gore Verbinski, Jerry Bruckhe...
2	[Action, Adventure, Crime]	[spy, based on novel, secret agent, sequel, mi...	[Columbia Pictures, Danjaq, B24]	[United Kingdom, United States of America]	[fr, en, es, it, de]	[Daniel Craig, Christoph Waltz, Léa Seydoux, R...	[Thomas Newman, Sam Mendes, Anna Pinnock, John...
3	[Action, Crime, Drama, Thriller]	[dc comics, crime fighter, terrorist, secret i...	[Legendary Pictures, Warner Bros., DC Entertai...	[United States of America]	[en]	[Christian Bale, Michael Caine, Gary Oldman, A...	[Hans Zimmer, Charles Roven, Christopher Nolan...
4	[Action, Adventure, Science Fiction]	[based on novel, mars, medallion, space travel...	[Walt Disney Pictures]	[United States of America]	[en]	[Taylor Kitsch, Lynn Collins, Samantha Morton,...	[Andrew Stanton, Andrew Stanton, John Lasseter...
...	...	...	...	...	...	...	...

→ Make New Columns

- To find relationship or correlation between these numbers and revenue

num_of_genres	num_of_keywords	num_of_production_companies	num_of_production_countries	num_of_spoken_languages	num_of_cast	num_of_crew
4	21	4	2	2	83	153
3	16	3	1	1	34	32

02. Data Pre-Processing

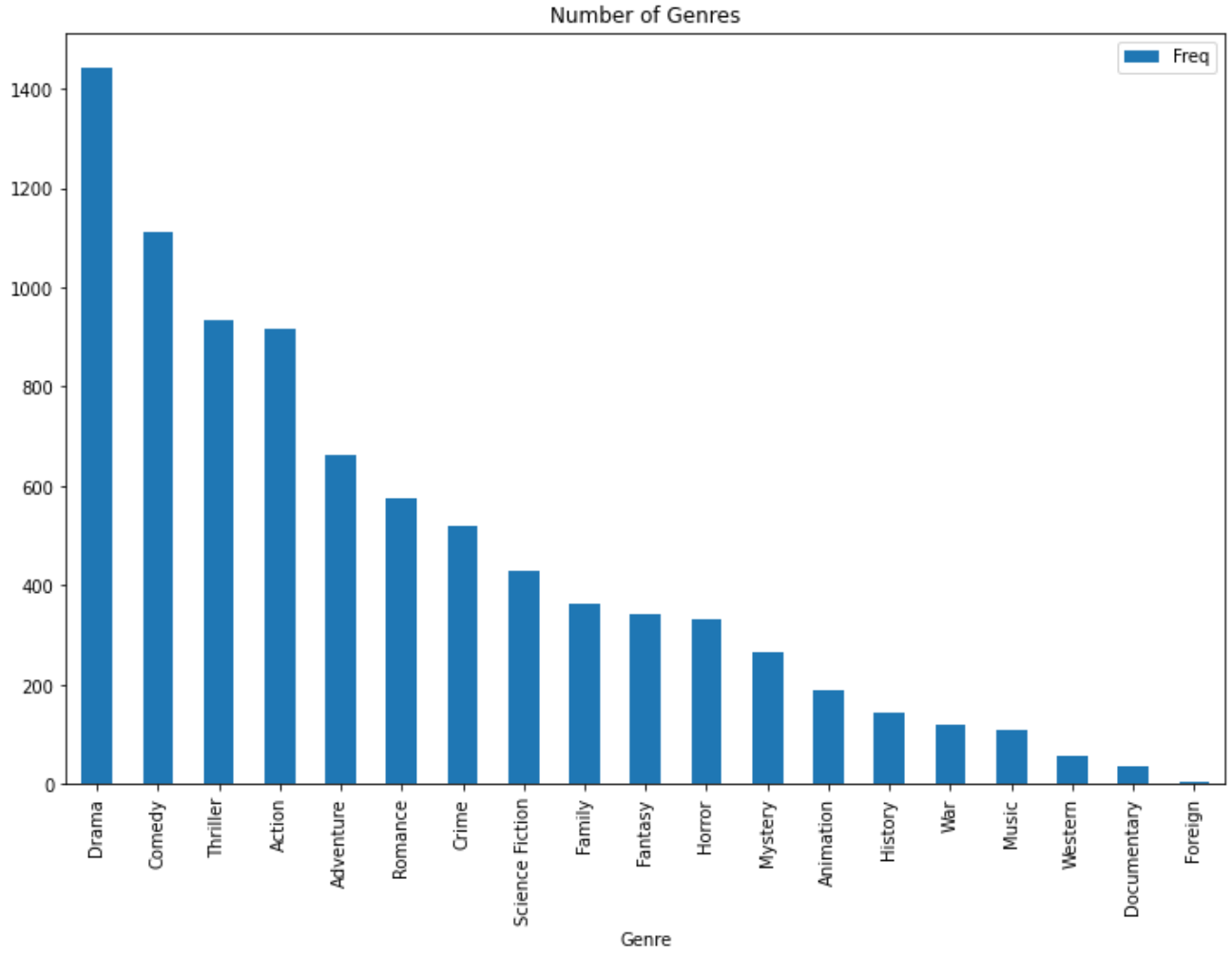
Final Dataset

```
Data columns (total 29 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   budget                        3229 non-null   int64
 1   genres                        3229 non-null   object
 2   homepage                      3229 non-null   object
 3   id                            3229 non-null   int64
 4   keywords                      3229 non-null   object
 5   original_language             3229 non-null   object
 6   original_title                3229 non-null   object
 7   overview                      3229 non-null   object
 8   popularity                    3229 non-null   float64
 9   production_companies          3229 non-null   object
 10  production_countries          3229 non-null   object
 11  release_date                  3229 non-null   object
 12  revenue                       3229 non-null   int64
 13  runtime                       3229 non-null   float64
 14  spoken_languages              3229 non-null   object
 15  status                        3229 non-null   object
 16  tagline                       3229 non-null   object
 17  title                         3229 non-null   object
 18  vote_average                  3229 non-null   float64
 19  vote_count                    3229 non-null   int64
 20  cast                          3229 non-null   object
 21  crew                          3229 non-null   object
 22  num_of_genres                 3229 non-null   int64
 23  num_of_keywords               3229 non-null   int64
 24  num_of_production_companies   3229 non-null   int64
 25  num_of_production_countries   3229 non-null   int64
 26  num_of_spoken_languages       3229 non-null   int64
 27  num_of_cast                   3229 non-null   int64
 28  num_of_crew                   3229 non-null   int64
dtypes: float64(3), int64(11), object(15)
memory usage: 756.8+ KB
```

03. Genre Trend

Number of Genres

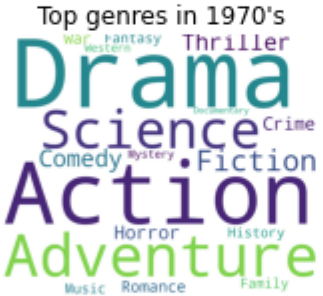
Genre	Freq
Drama	1441
Comedy	1110
Thriller	935
Action	918
Adventure	661
Romance	574
Crime	521
Science Fiction	431
Family	365
Fantasy	342
Horror	332
Mystery	265
Animation	188
History	145
War	120
Music	111
Western	57
Documentary	38
Foreign	5



03. Genre Trend

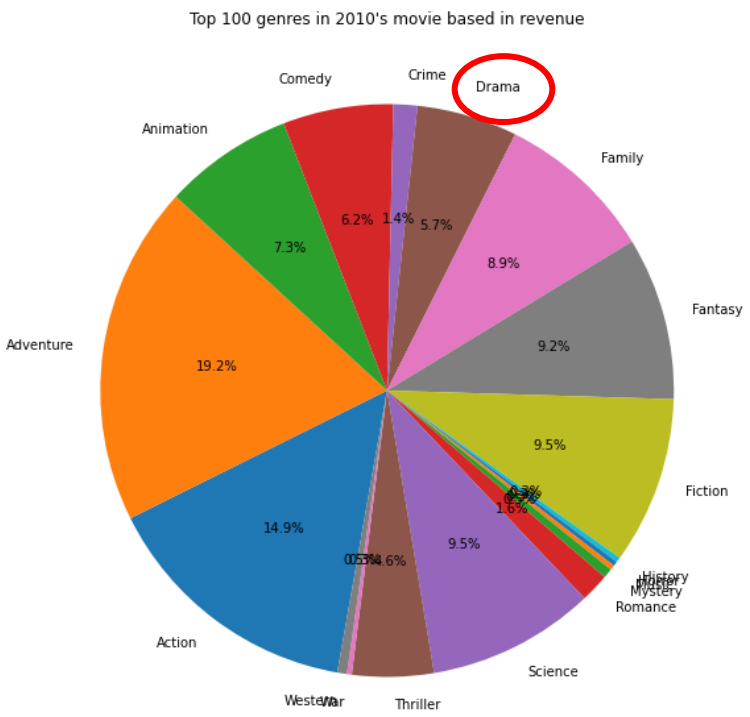
WordCloud

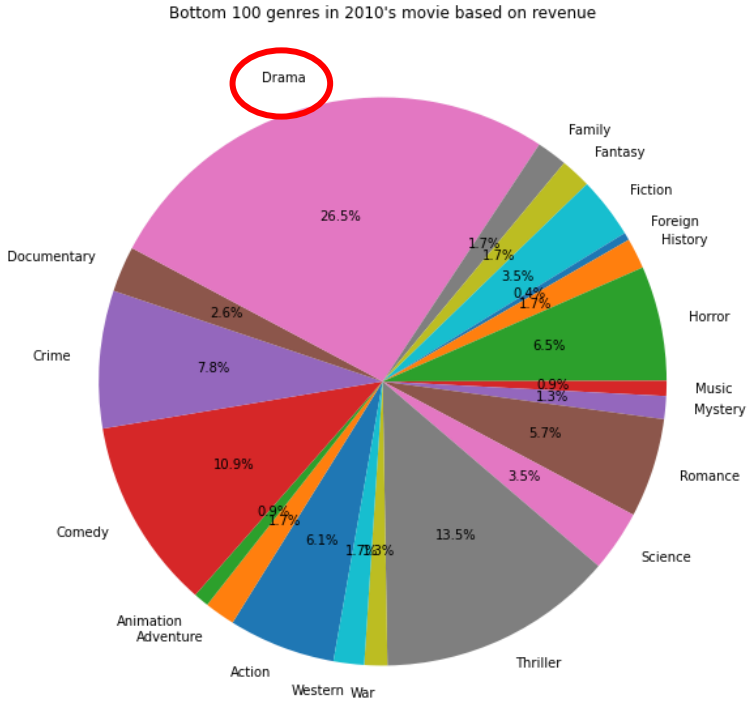
→10-year cycle of movie genre preference output based on frequency of genre

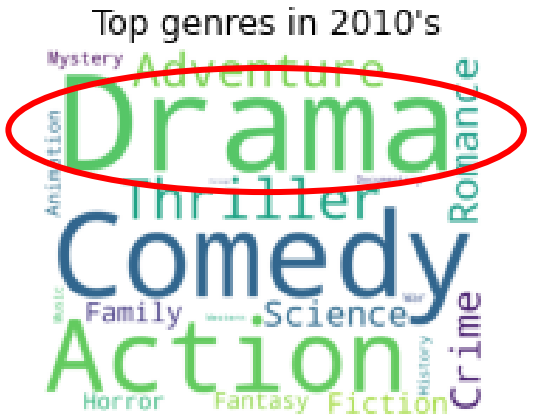


03. Genre Trend

Top 100 & Bottom 100 movie genres in 2010's based on revenue





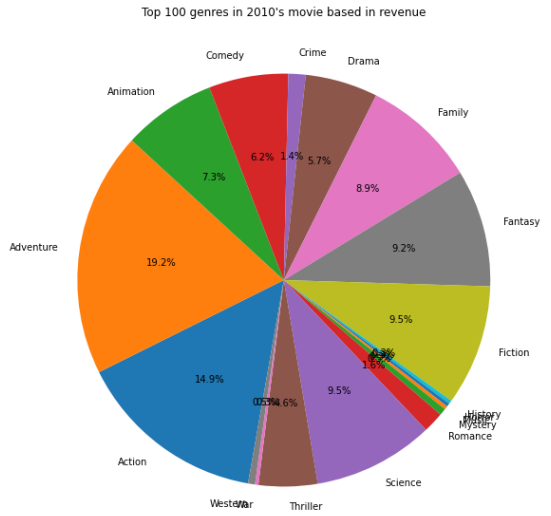


- Drama is the worst genre to get high revenue in 2010's even Drama is the most even though Drama genre have been made so much in 2010's

03. Genre Trend

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_companies	production_countries	release_date	revenue	runtime
3	250000000	[Action, Crime, Drama, Thriller]	http://www.thedarkknightrises.com/	49026	[dc comics, crime fighter, terrorist, secret i...	en	The Dark Knight Rises	Following the death of District Attorney Harve...	112.312950	[Legendary Pictures, Warner Bros., DC Entertai...	[United States of America]	2012-07-16	1084939099	165.0
78	175000000	[Family, Adventure, Drama, Fantasy]	http://movies.disney.com/the-jungle-book-2016	278927	[based on novel, snake, wolf, elephant, tiger,...	en	The Jungle Book	After a threat from the tiger Shere Khan force...	94.199316	[Walt Disney Pictures, Walt Disney Studios Mot...	[United Kingdom, United States of America]	2016-04-07	966550600	106.0
77	175000000	[Drama, Comedy, Animation, Family]	http://movies.disney.com/inside-out	150540	[dream, cartoon, imaginary friend, animation, ...	en	Inside Out	Growing up can be a bumpy road, and it's no ex...	128.655964	[Walt Disney Pictures, Pixar Animation Studios]	[United States of America]	2015-06-09	857611174	94.0

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_companies	production_countries	release_date	revenue	runtime	s
3131	10	[Drama, Romance, Comedy]	http://hrosss.is/the-film/	217708	[horse, snow storm, icelandic]	is	Hross í oss	A country romance about the human streak in th...	1.617943	[Filmhuset Gruppen, Leiknar Myndir]	[Norway, Iceland]	2013-08-30	11	85.0	
4065	2100000	[Drama, Crime]	http://miamericamovie.net/	364083	[new york state, hate crime]	en	Mi America	A hate-crime has been committed in a the small...	0.039007	[Industrial House Films]	[United States of America]	2015-10-16	3330	126.0	
3120	10000000	[Drama, Thriller]	NaN	245846	[australia, missing child, outback, alcoholic ...	en	Strangerland	Newly arrived to a remote desert town, Catheri...	5.145655	[Worldview Entertainment]	[Australia]	2015-07-01	17472	111.0	



→ Top 3 Drama movies based on revenue

→ Bottom 3 Drama movies based on revenue

03. Genre Trend

spoken_languages	status	tagline	title	vote_average	vote_count	cast	crew	num_of_genres	num_of_keywords	num_of_production_companies	num_of_production_countries	num_of_spoken_languages	num_of_cast	num_of_crew
[en]	Released	The Legend Ends	The Dark Knight Rises	7.6	9106	[Christian Bale, Michael Caine, Gary Oldman, A...	[Hans Zimmer, Charles Roven, Christopher Nolan...	4	21	4	1	1	158	217
[en]	Released	NaN	The Jungle Book	6.7	2892	[Neel Sethi, Bill Murray, Ben Kingsley, Idris ...	[John Debney, Sarah Finn, Mark Livolsi, Bill P...	4	14	4	2	1	30	25
[en]	Released	Meet the little voices inside your head.	Inside Out	8.0	6560	[Amy Poehler, Phyllis Smith, Richard Kind, Bil...	[Andrew Stanton, Bob Peterson, John Lasseter, ...	4	11	2	1	1	65	50

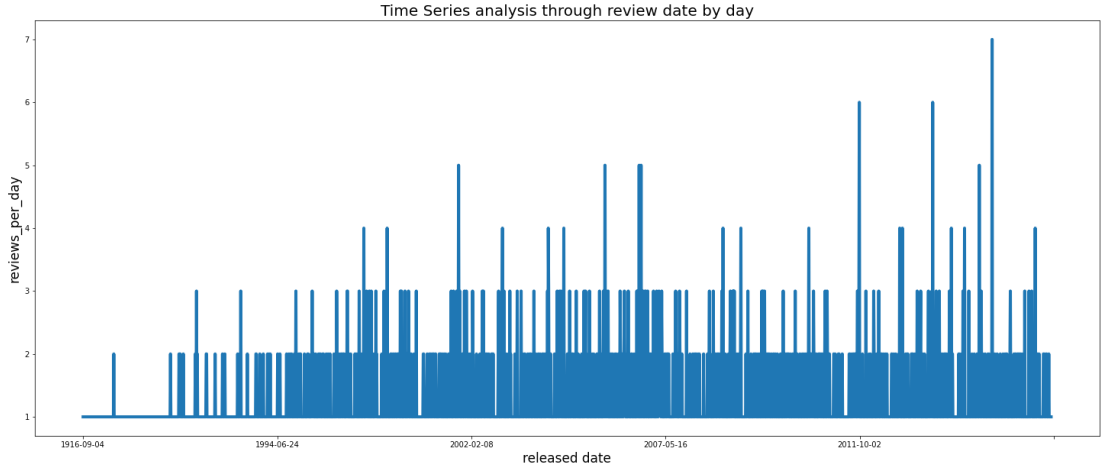
→ Top 3 Drama movies based on revenue

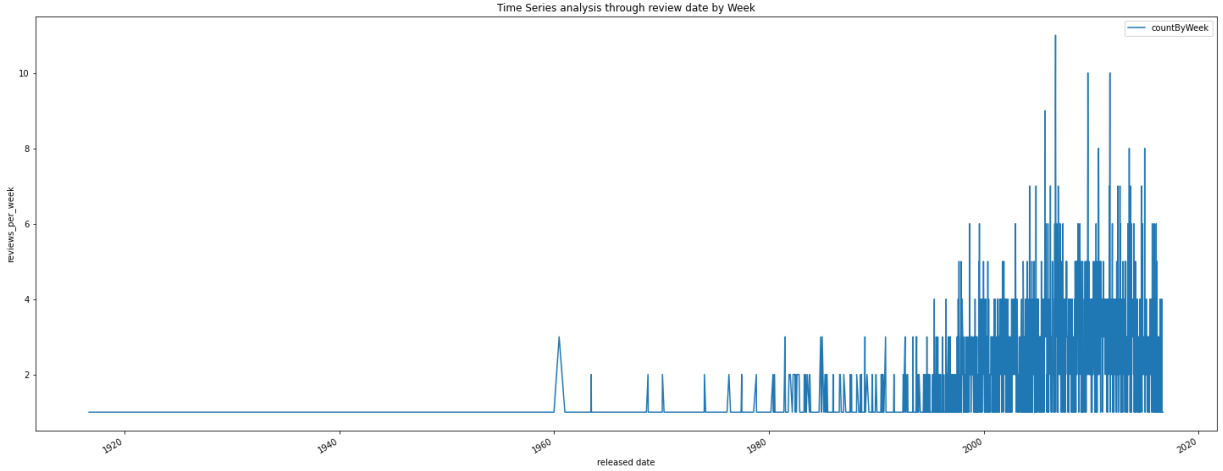
en_languages	status	tagline	title	vote_average	vote_count	cast	crew	num_of_genres	num_of_keywords	num_of_production_companies	num_of_production_countries	num_of_spoken_languages	num_of_cast	num_of_crew
en, is, ru, es, sv]	Released	NaN	Of Horses and Men	6.9	26	[Ingvar Eggert Sigurðsson, Charlotte Bøving, S...	[Friðrik Þór Friðriksson, Benedikt Erlingsson,...	3	3	2	2	5	16	9
[es, en]	Released	NaN	Mi America	0.0	0	[Robert Fontaine, Michael Brainard, Grant Boyd...	[Marko A. Costanzo, Rick Porras, Robert Fontai...	2	2	1	1	2	5	7
[en]	Released	To find the truth they must lose themselves.	Strangerland	5.1	83	[Nicole Kidman, Joseph Fiennes, Hugo Weaving, ...	[Veronika Jenet, Nikki Barrett, Melinda Doring...	2	8	1	1	1	10	13

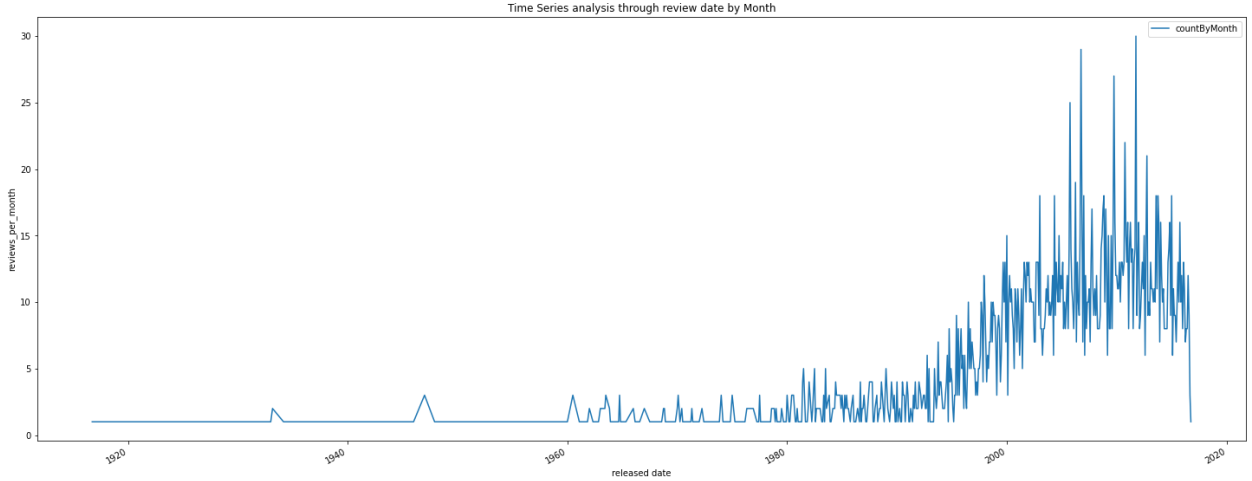
→ Bottom 3 Drama movies based on revenue

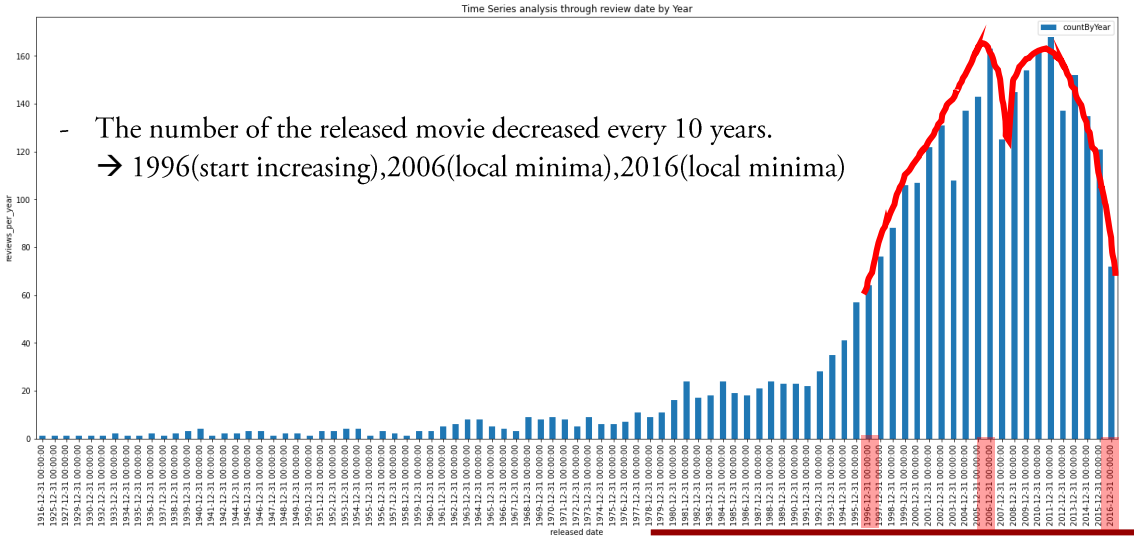
03. Released Day Trend

- Graph based on released date of the movie by day, week, month and year.
- Hard to find any seasonality in these graph.



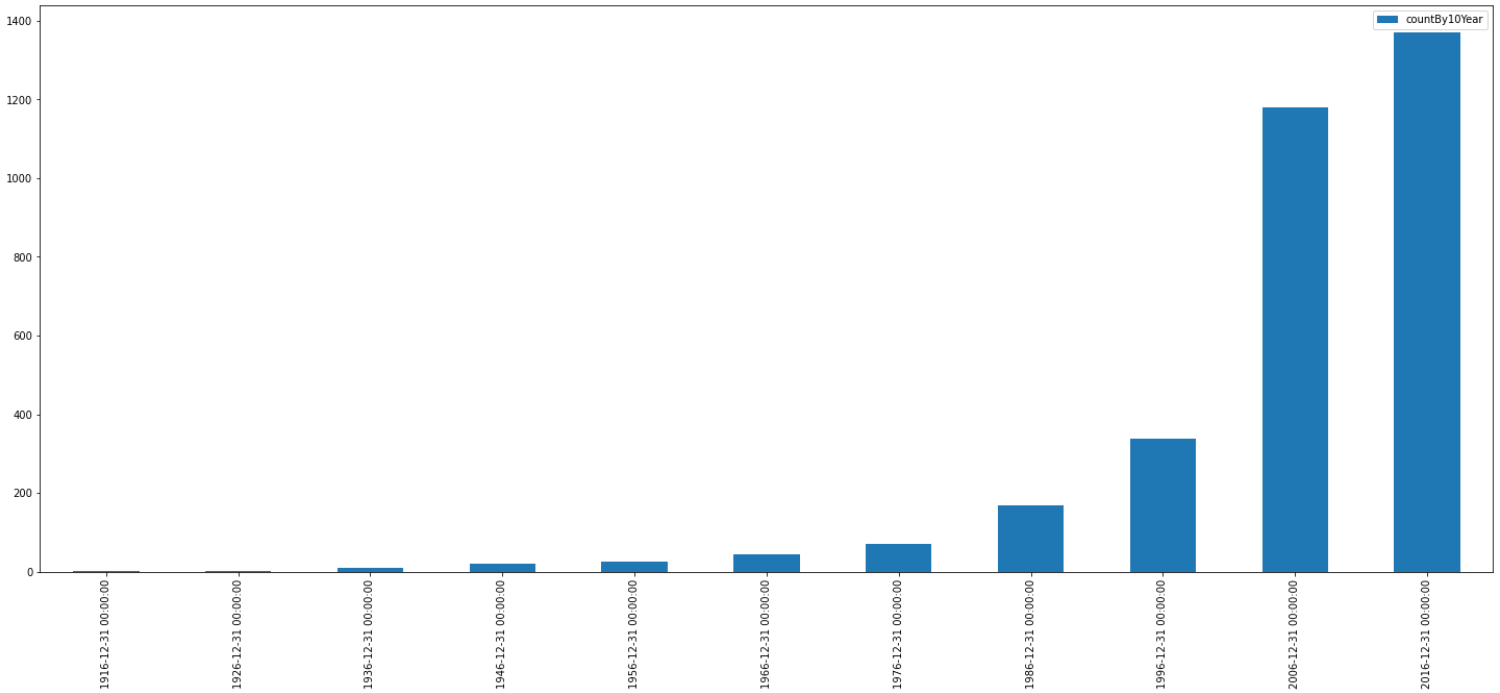






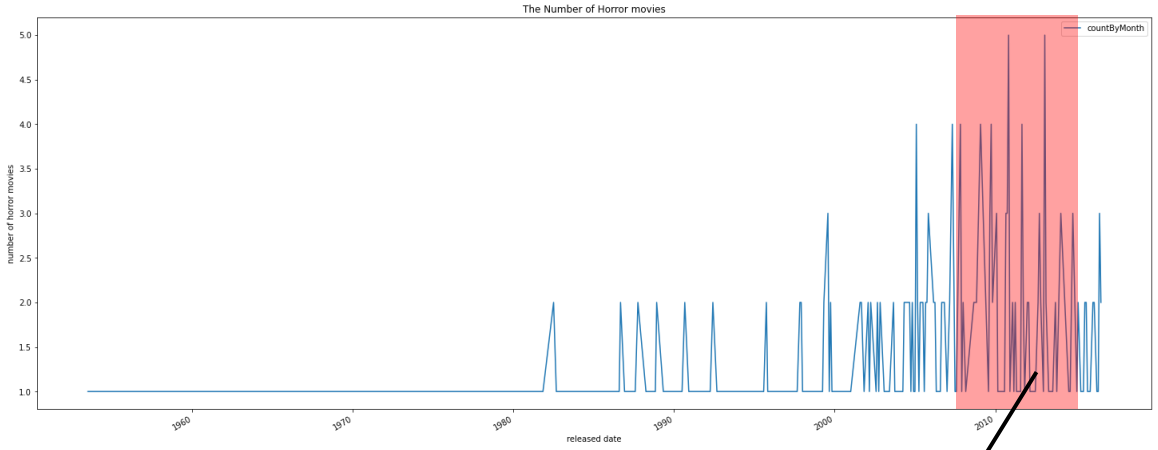
- The number of the released movie decreased every 10 years.
  → 1996(start increasing),2006(local minima),2016(local minima)

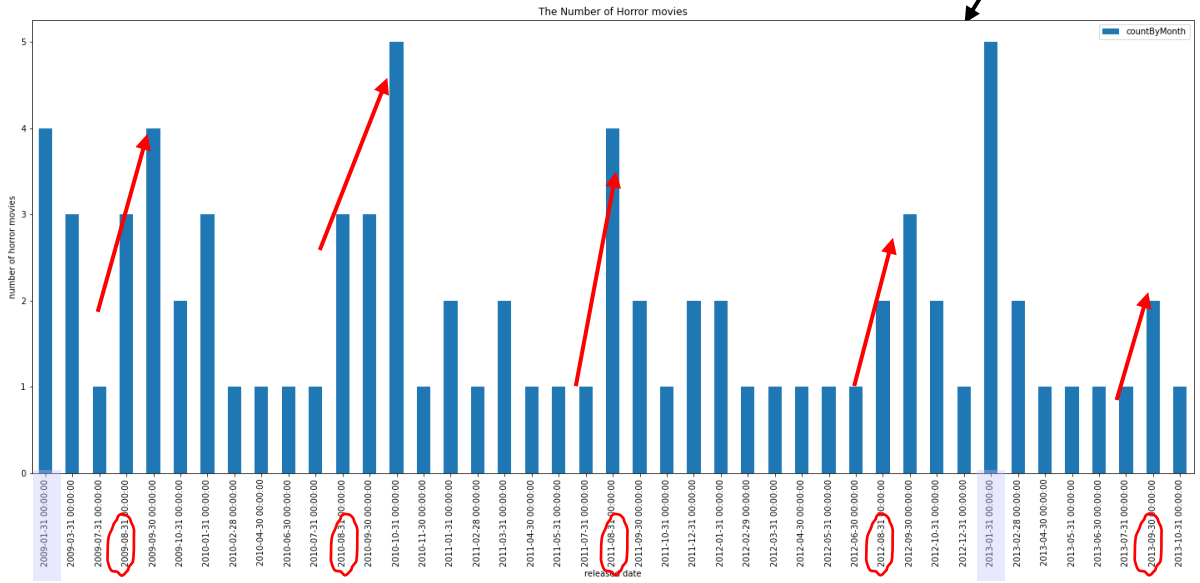
03. Released Day Trend



- Also check the number of released movies by 10 years
- It shows the number of released movies are drastically increasing over 10 years

03. Released Day Trend





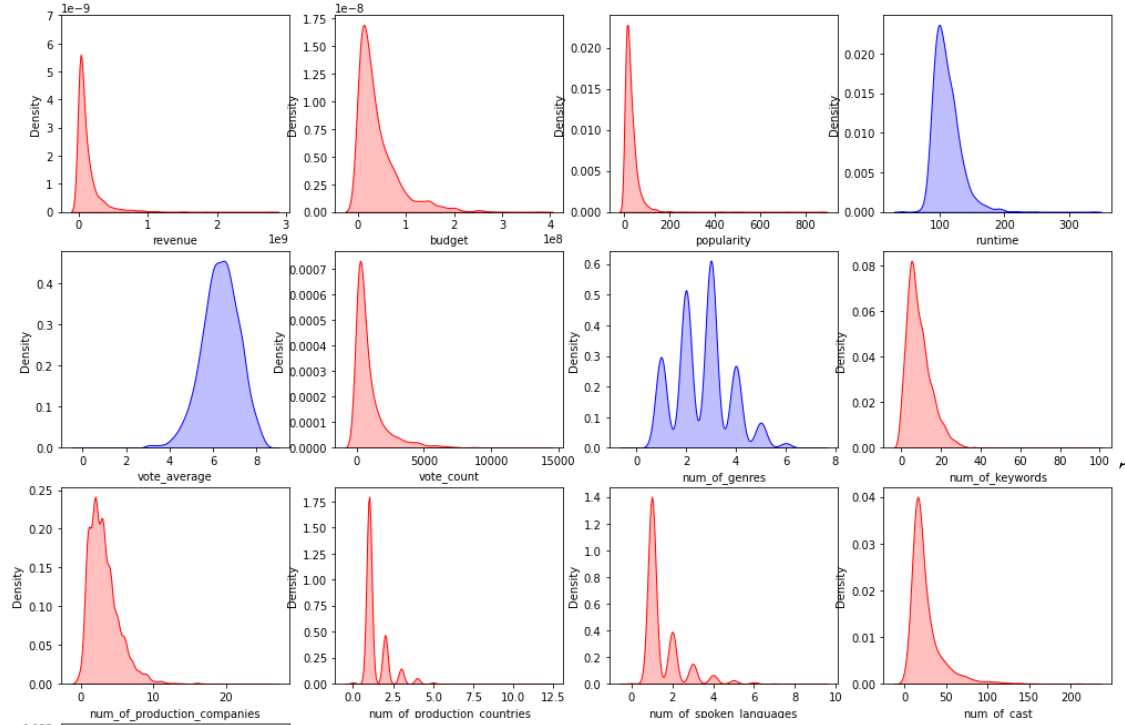
- Check seasonality of Horror movie.
- We usually see horror movie in summer, and I want to check the released date of horror movie.

- Checking the number of horror movies in 5 years (2009 ~ 2013).

- Surprisingly, horror movies are usually released in February too.
- The number of horror movies are increasing starts from August which is start of summer.

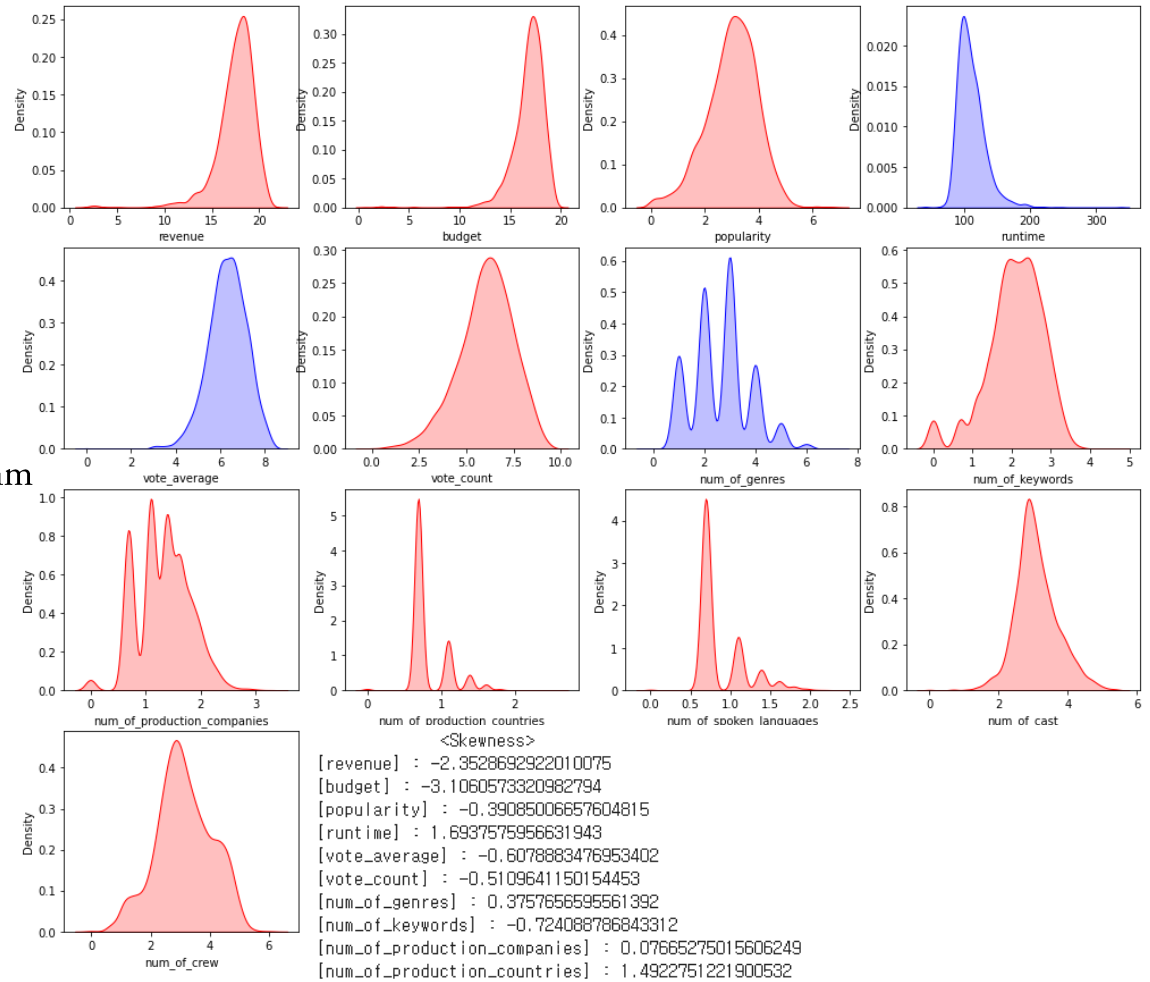
04. Revenue Analysis

Power Law Distribution



<Skewness>
[revenue] : 3.818316479714257
[budget] : 2.0886549188400836
[popularity] : 9.261435590982792
[runtime] : 1.6937575956631943
[vote_average] : -0.6078883476953402
[vote_count] : 3.206093933634065
[num_of_genres] : 0.37576565595561392
[num_of_keywords] : 1.963829122877158
[num_of_production_companies] : 2.0319610862570467
[num_of_production_countries] : 2.9425279185718827
[num_of_spoken_languages] : 2.4197241594551233
[num_of_cast] : 3.1115040220673067
[num_of_crew] : 2.333314958931691

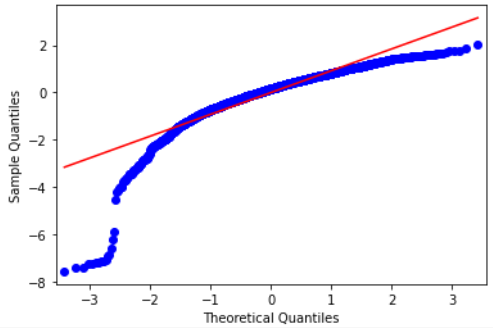
Take logarithm



<Skewness>
[revenue] : -2.3528692922010075
[budget] : -3.1060573320982794
[popularity] : -0.39085006657604815
[runtime] : 1.6937575956631943
[vote_average] : -0.6078883476953402
[vote_count] : -0.5109641150154453
[num_of_genres] : 0.3757656595561392
[num_of_keywords] : -0.724088786843312
[num_of_production_companies] : 0.07665275015606249
[num_of_production_countries] : 1.4922751221900532
[num_of_spoken_languages] : 1.4948982415358016
[num_of_cast] : 0.3281950024691435 7
[num_of_crew] : -0.067500453782168

04. Revenue Analysis

Standardization



→ QQ plot

```
from sklearn.preprocessing import StandardScaler

data_logarithm_zscore = data_logarithm[['revenue','budget','popularity','runtime','vote_average','vote_count','num_of_genres', 'num_of_keywords','num_of_production_companies','num_of_production_countries','num_of_spoken_languages','num_of_cast','num_of_cr
standardScaler = StandardScaler()
standardScaler.fit(data_logarithm_zscore)
data_logarithm_zscore[:] = standardScaler.transform(data_logarithm_zscore[:])
data_logarithm_zscore
```

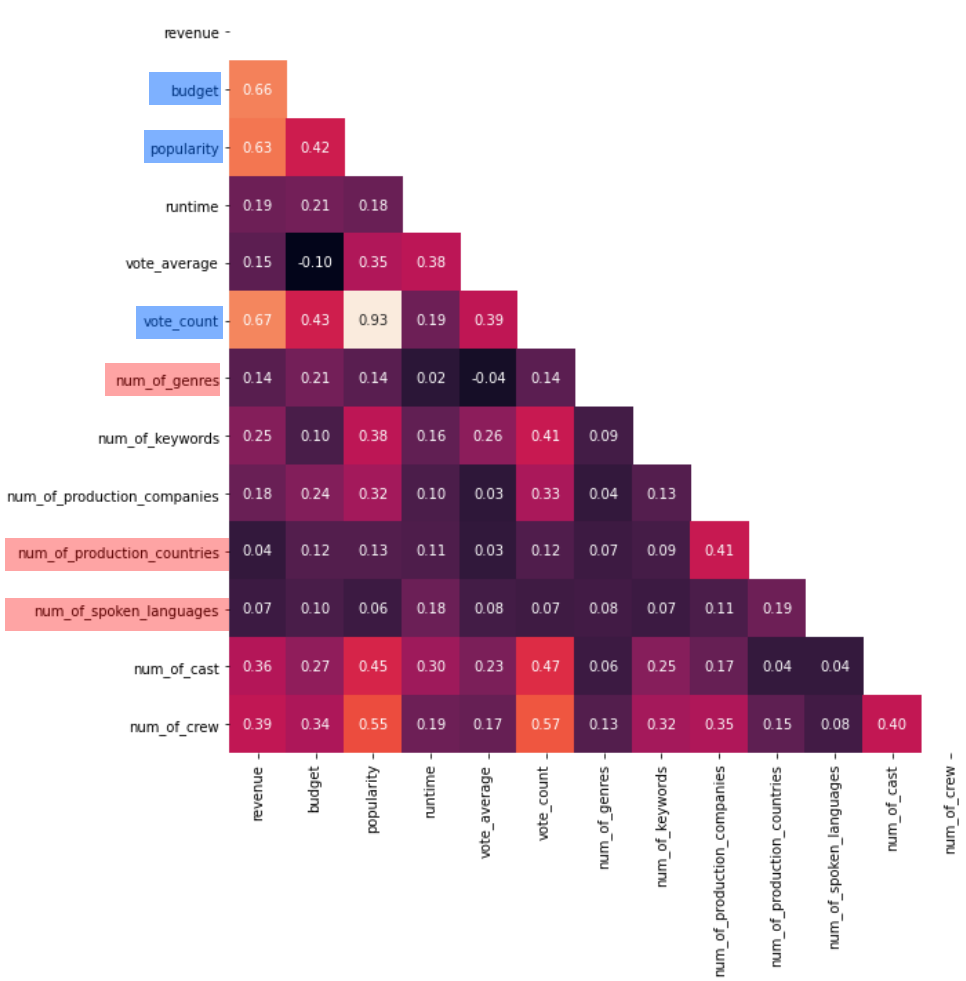
	revenue	budget	popularity	runtime	vote_average	vote_count	num_of_genres	num_of_keywords	num_of_production_companies	num_of_production_countries	num_of_spoken_languages	num_of_cast	num_of_crew
0	2.048371	1.494788	2.157033	2.446071	1.019332	2.298651	1.205433	1.430936	0.607627	1.002711	0.744081	2.174657	2.019212
1	1.535923	1.636657	2.073929	2.780001	0.675987	1.635445	0.312079	1.064167	0.138066	-0.535028	-0.597321	0.755598	0.390492
2	1.493927	1.514768	1.800313	1.778210	-0.010704	1.630228	0.312079	-0.008093	0.138066	1.002711	3.037224	2.174657	2.032855
3	1.594286	1.526927	1.847803	2.589184	1.477126	2.120348	1.205433	1.430936	0.607627	-0.535028	-0.597321	3.208942	2.386670
4	0.949661	1.550532	0.861409	1.014940	-0.239601	1.119041	0.312079	1.064167	-1.320524	-0.535028	-0.597321	0.393901	1.864208
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4773	-1.216242	-3.969949	0.037678	-0.893234	1.248229	0.407945	-1.474629	-0.994113	-0.467304	-0.535028	-0.597321	-0.513188	-1.107779
4788	-0.906395	-4.457978	-1.367654	-0.845530	-0.125153	-0.912097	0.312079	1.145476	-1.320524	-0.535028	-0.597321	-0.979496	-1.107779
4792	-2.881137	-4.150559	-2.990275	0.013149	1.248229	-1.290972	1.205433	1.145476	-1.320524	-0.535028	-0.597321	-0.513188	-2.144814
4796	-2.180421	-4.782335	0.206487	-1.608799	0.675987	0.313463	0.312079	0.309335	-1.320524	-0.535028	-0.597321	-0.151492	-1.248962
4798	-1.425226	-2.707420	-0.289236	-1.417982	0.332641	-0.384404	0.312079	-0.417328	-1.320524	1.002711	-0.597321	-1.636720	-0.679079

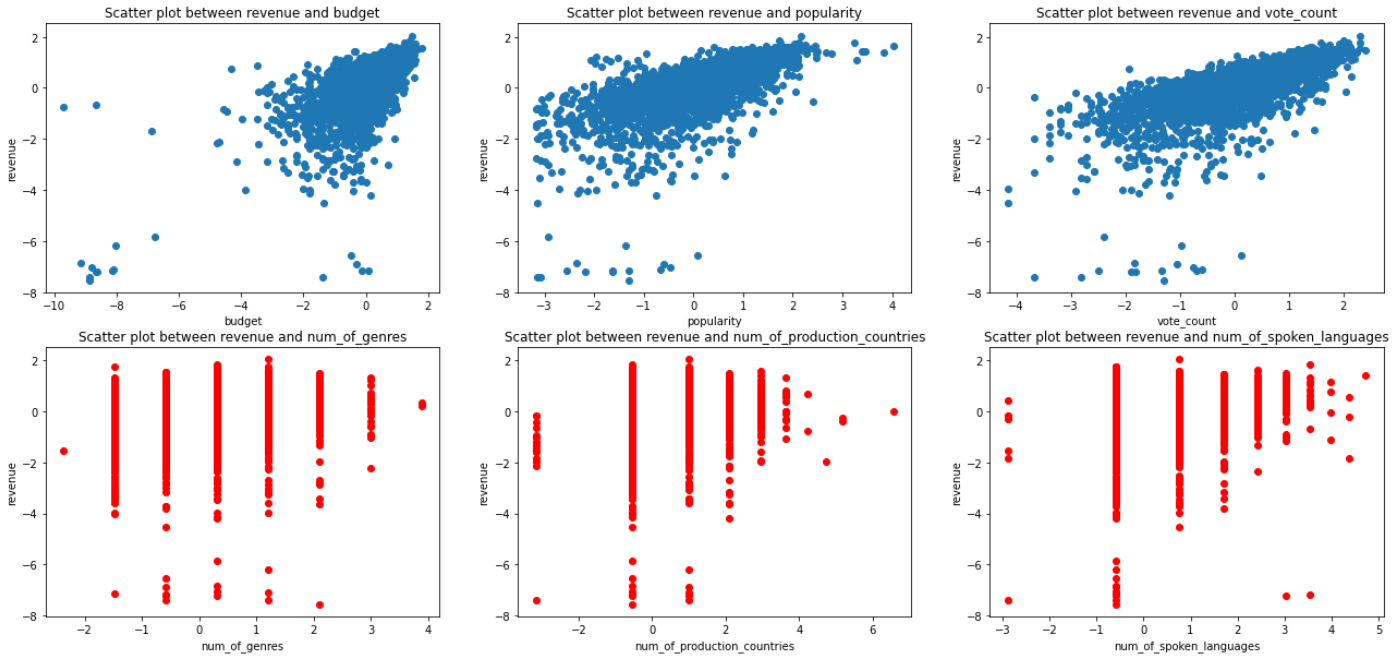
3229 rows × 13 columns

[Final Dataset]

04. Revenue Analysis

Correlation(Pearson)





04. Revenue Analysis

```
df_train, df_test = train_test_split(data_logarithm_zscore, test_size=0.2, random_state=seed)
```

* Train Data : Test Data = 8:2

Regression Feature Selection (Backward Stepwise)

1. Check Multicollinearity

	VIF Factor	features
0	1.0	Intercept
1	1.5	budget
2	7.6	popularity
3	1.4	runtime
4	1.6	vote_average
5	8.6	vote_count
6	1.1	num_of_genres
7	1.3	num_of_keywords
8	1.4	num_of_production_companies
9	1.2	num_of_production_countries
10	1.1	num_of_spoken_languages
11	1.4	num_of_cast
12	1.6	num_of_crew

2. Check P-value

OLS Regression Results

Dep. Variable:	revenue	R-squared:	0.624
Model:	OLS	Adj. R-squared:	0.622
Method:	Least Squares	F-statistic:	355.6
Date:	Wed, 30 Jun 2021	Prob (F-statistic):	0.00
Time:	16:16:28	Log-Likelihood:	-2426.7
No. Observations:	2583	AIC:	4879.
Df Residuals:	2570	BIC:	4956.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0062	0.012	-0.506	0.613	-0.030	0.018
budget	0.4671	0.015	31.433	0.000	0.438	0.496
popularity	0.0168	0.033	0.508	0.612	-0.048	0.082
runtime	0.0271	0.015	1.834	0.067	-0.002	0.056
vote_average	-0.0229	0.016	-1.444	0.149	-0.054	0.008
vote_count	0.5092	0.035	14.482	0.000	0.440	0.578
num_of_genres	-0.0324	0.013	-2.570	0.010	-0.057	-0.008
num_of_keywords	0.0186	0.014	1.366	0.172	-0.008	0.045
num_of_production_companies	-0.0769	0.015	-5.283	0.000	-0.105	-0.048
num_of_production_countries	-0.0442	0.014	-3.246	0.001	-0.071	-0.017
num_of_spoken_languages	-0.0018	0.013	-0.142	0.887	-0.027	0.023
num_of_cast	0.0071	0.014	0.496	0.620	-0.021	0.035
num_of_crew	-0.0320	0.016	-2.041	0.041	-0.063	-0.001

Omnibus:	1678.783	Durbin-Watson:	1.932
Prob(Omnibus):	0.000	Jarque-Bera (JB):	53728.184
Skew:	-2.581	Prob(JB):	0.00
Kurtosis:	24.738	Cond. No.	7.30

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

*confidence level : 95%

→ Popularity and vote_count features' VIF score is relatively high, but they are less than 10, so we can use them just for now.

04. Revenue Analysis

Regression Feature Selection (Backward Stepwise)

3. Keep Checking P-value *confidence level : 95%

*if you want to add popularity and vote_count features in this model,
we should do dimension reduction like PCA analysis or SVM.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0062	0.012	-0.507	0.612	-0.030	0.018
budget	0.4670	0.015	31.470	0.000	0.438	0.496
popularity	0.0169	0.033	0.512	0.609	-0.048	0.082
runtime	0.0269	0.015	1.829	0.067	-0.002	0.056
vote_average	-0.0230	0.016	-1.451	0.147	-0.054	0.008
vote_count	0.5091	0.035	14.486	0.000	0.440	0.578
num_of_genres	-0.0325	0.013	-2.585	0.010	-0.057	-0.008
num_of_keywords	0.0186	0.014	1.366	0.172	-0.008	0.045
num_of_production_companies	-0.0769	0.015	-5.287	0.000	-0.105	-0.048
num_of_production_countries	-0.0445	0.013	-3.309	0.001	-0.071	-0.018
num_of_cast	0.0072	0.014	0.501	0.617	-0.021	0.035
num_of_crew	-0.0320	0.016	-2.044	0.041	-0.063	-0.001

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0061	0.012	-0.499	0.618	-0.030	0.018
budget	0.4672	0.015	31.494	0.000	0.438	0.496
popularity	0.0172	0.033	0.521	0.603	-0.048	0.082
runtime	0.0286	0.014	1.992	0.047	0.000	0.057
vote_average	-0.0232	0.016	-1.462	0.144	-0.054	0.008
vote_count	0.5112	0.035	14.660	0.000	0.443	0.580
num_of_genres	-0.0327	0.013	-2.599	0.009	-0.057	-0.008
num_of_keywords	0.0188	0.014	1.384	0.165	-0.008	0.046
num_of_production_companies	-0.0769	0.015	-5.284	0.000	-0.105	-0.048
num_of_production_countries	-0.0450	0.013	-3.353	0.001	-0.071	-0.019
num_of_crew	-0.0309	0.016	-1.994	0.046	-0.061	-0.001

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0061	0.012	-0.500	0.617	-0.030	0.018
budget	0.4676	0.015	31.559	0.000	0.439	0.497
runtime	0.0287	0.014	2.002	0.045	0.001	0.057
vote_average	-0.0232	0.016	-1.462	0.144	-0.054	0.008
vote_count	0.5268	0.018	29.137	0.000	0.491	0.562
num_of_genres	-0.0327	0.013	-2.598	0.009	-0.057	-0.008
num_of_keywords	0.0188	0.014	1.379	0.168	-0.008	0.045
num_of_production_companies	-0.0771	0.015	-5.300	0.000	-0.106	-0.049
num_of_production_countries	-0.0447	0.013	-3.337	0.001	-0.071	-0.018
num_of_crew	-0.0304	0.015	-1.965	0.050	-0.061	-6.15e-05

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0062	0.012	-0.511	0.609	-0.030	0.018
budget	0.4651	0.015	31.533	0.000	0.437	0.495
runtime	0.0299	0.014	2.090	0.037	0.002	0.058
vote_average	-0.0213	0.016	-1.346	0.178	-0.052	0.010
vote_count	0.5327	0.018	30.324	0.000	0.498	0.567
num_of_genres	-0.0321	0.013	-2.551	0.011	-0.057	-0.007
num_of_production_companies	-0.0779	0.015	-5.357	0.000	-0.106	-0.049
num_of_production_countries	-0.0439	0.013	-3.278	0.001	-0.070	-0.018
num_of_crew	-0.0279	0.015	-1.814	0.070	-0.058	0.002

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0064	0.012	-0.524	0.601	-0.030	0.018
budget	0.4734	0.014	34.411	0.000	0.446	0.500
runtime	0.0218	0.013	1.678	0.093	-0.004	0.047
vote_count	0.5223	0.016	33.083	0.000	0.491	0.553
num_of_genres	-0.0314	0.013	-2.500	0.012	-0.056	-0.007
num_of_production_companies	-0.0759	0.014	-5.250	0.000	-0.104	-0.048
num_of_production_countries	-0.0446	0.013	-3.336	0.001	-0.071	-0.018
num_of_crew	-0.0274	0.015	-1.787	0.074	-0.058	0.003

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0065	0.012	-0.532	0.595	-0.030	0.017
budget	0.4765	0.014	34.927	0.000	0.450	0.503
vote_count	0.5241	0.016	33.269	0.000	0.493	0.555
num_of_genres	-0.0323	0.013	-2.570	0.010	-0.057	-0.008
num_of_production_companies	-0.0765	0.014	-5.289	0.000	-0.105	-0.048
num_of_production_countries	-0.0432	0.013	-3.232	0.001	-0.069	-0.017
num_of_crew	-0.0253	0.015	-1.652	0.099	-0.055	0.005

04. Revenue Analysis

Regression Feature Selection (Backward Stepwise)

4. Final Features

OLS Regression Results

Dep. Variable:	revenue	R-squared:	0.623
Model:	OLS	Adj. R-squared:	0.622
Method:	Least Squares	F-statistic:	850.6
Date:	Wed, 30 Jun 2021	Prob (F-statistic):	0.00
Time:	16:45:42	Log-Likelihood:	-2431.6
No. Observations:	2583	AIC:	4875.
Df Residuals:	2577	BIC:	4910.
Df Model:	5		
Covariance Type:	nonrobust		

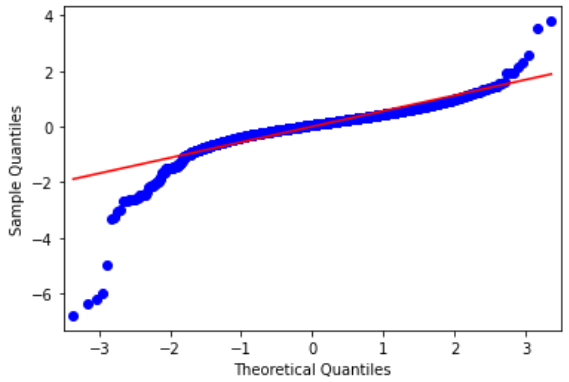
*Final Features

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0064	0.012	-0.523	0.601	-0.030	0.018
budget	0.4741	0.014	34.934	0.000	0.448	0.501
vote_count	0.5125	0.014	36.337	0.000	0.485	0.540
num_of_genres	-0.0328	0.013	-2.608	0.009	-0.057	-0.008
num_of_production_companies	-0.0809	0.014	-5.690	0.000	-0.109	-0.053
num_of_production_countries	-0.0434	0.013	-3.245	0.001	-0.070	-0.017

Omnibus:	1665.404	Durbin-Watson:	1.935
Prob(Omnibus):	0.000	Jarque-Bera (JB):	53316.240
Skew:	-2.550	Prob(JB):	0.00
Kurtosis:	24.665	Cond. No.	1.94

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[Simple Regression model]
R-square : 0.623
Regression Equation :
Y = 0.4741*(budget) + 0.5125(vote_count) – 0.0328(num_of_genre) – 0.0809(num_of_production_companies) – 0.0434(num_of_production_countries)-0.0064



[QQ Plot]

```
RMSE = np.sqrt((result1.resid **2).mean())
RMSE
```

```
0.6203113171915299
```

[RMSE]

```
train_features  = df_train[["budget","vote_count","num_of_genres",  "num_of_production_companies", "num_of_production_countries"] ]
train_target = df_train['revenue']

test_features = df_test[["budget","vote_count","num_of_genres", "num_of_production_companies", "num_of_production_countries"] ]
test_target = df_test['revenue']
```

04. Revenue Analysis

Regression Model 1 – Linear Regression

```
from sklearn import linear_model
linear = linear_model.LinearRegression()

model = linear.fit(train_features,train_target)
print(linear.score(train_features,train_target)) # accuracy
print(linear.score(test_features,test_target)) # accuracy
```

```
0.6226805818331442
0.6450342866492442
```

```
rmse
```

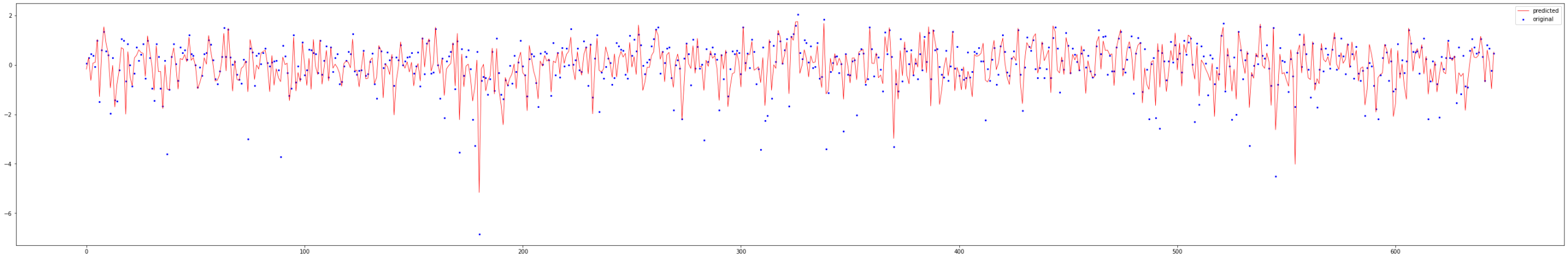
```
0.6203113171915301
```

feature	coefficient
budget	0.474134
vote_count	0.512533
num_of_genres	-0.032761
num_of_production_companies	-0.080929
num_of_production_countries	-0.043366

```
linear.intercept_
```

```
-0.006389565537345397
```

Y = 0.4741*(budget) + 0.5125(vote_count) – 0.0328(num_of_genre) –0.0809(num_of_production_companies) – 0.0434(num_of_production_countries)-0.0064



04. Revenue Analysis

Regression Model 2 – Ridge Regression

```
from sklearn.metrics import mean_squared_error
for a in alpha_list:
  model = Ridge(alpha=a).fit(x,y)
  score = model.score(x, y)
  pred_y = model.predict(x)
  mse = mean_squared_error(y, pred_y)
  print("Alpha:{0:.4f}, R2:{1:.2f}, MSE:{2:.2f}, RMSE:{3:.2f}"
  .format(a, score, mse, np.sqrt(mse)))
```

```
Alpha:0.0010, R2:0.63, MSE:0.37, RMSE:0.61
Alpha:0.0100, R2:0.63, MSE:0.37, RMSE:0.61
Alpha:0.1000, R2:0.63, MSE:0.37, RMSE:0.61
Alpha:1.0000, R2:0.63, MSE:0.37, RMSE:0.61
Alpha:10.0000, R2:0.63, MSE:0.37, RMSE:0.61
Alpha:100.0000, R2:0.63, MSE:0.37, RMSE:0.61
Alpha:1000.0000, R2:0.60, MSE:0.40, RMSE:0.63
```

```
ridge = Ridge(alpha = 0.1)

ridge.fit(train_features,train_target)
print(ridge.score(train_features,train_target))
print(ridge.score(test_features,test_target)) # accuracy
```

```
0.6229607916386113
0.6428472214642391
```

feature	coefficient
budget	0.473991
vote_count	0.512341
num_of_genres	-0.032699
num_of_production_companies	-0.080802
num_of_production_countries	-0.043368

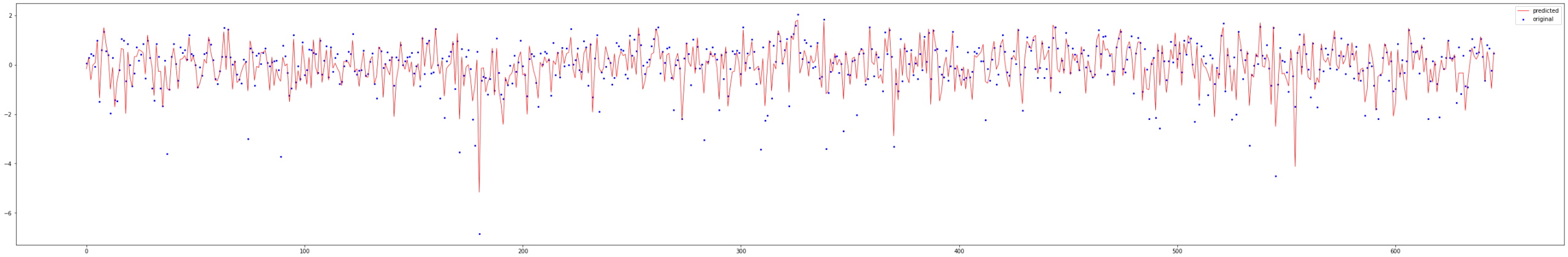
```
ridge.intercept_
```

```
-0.006388035482275863
```

```
rmse
```

```
0.6203113177489763
```

Y = 0.4741*(budget) + 0.5124(vote_count) – 0.0327(num_of_genre) –0.0809(num_of_production_companies) – 0.0434(num_of_production_countries)-0.0064



04. Revenue Analysis

Regression Model 3 – ElasticNet Regression

Lasso Ridge

$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + r\alpha\sum_{i=1}^{n}|\theta_i| + \frac{1-r}{2}\alpha\sum_{i=1}^{n}\theta_i^2$$

```
from sklearn.linear_model import ElasticNet
alpha_list = [0.001,0.01,0.1,1,10,100, 1000]
for a in alpha_list:
  model = ElasticNet(alpha=a).fit(x,y)
  score = model.score(x, y)
  pred_y = model.predict(x)
  mse = mean_squared_error(y, pred_y)
  print("Alpha:{0:.4f}, R2:{1:.2f}, MSE:{2:.2f}, RMSE:{3:.2f}"
  .format(a, score, mse, np.sqrt(mse)))
```

```
Alpha:0.0010, R2:0.63, MSE:0.37, RMSE:0.61
Alpha:0.0100, R2:0.63, MSE:0.37, RMSE:0.61
Alpha:0.1000, R2:0.61, MSE:0.39, RMSE:0.62
Alpha:1.0000, R2:0.21, MSE:0.79, RMSE:0.89
Alpha:10.0000, R2:0.00, MSE:1.00, RMSE:1.00
Alpha:100.0000, R2:0.00, MSE:1.00, RMSE:1.00
Alpha:1000.0000, R2:0.00, MSE:1.00, RMSE:1.00
```

```
elasticNet = ElasticNet(alpha = 0.01, l1_ratio = 0.5)

elasticNet.fit(train_features,train_target)
print(elasticNet.score(train_features,train_target))# train accuracy
print(elasticNet.score(test_features,test_target)) # test accuracy
```

```
0.6224668434010228
0.646128306138653
```

feature	coefficient
budget	0.466855
vote_count	0.503770
num_of_genres	-0.025425
num_of_production_companies	-0.072465
num_of_production_countries	-0.040354

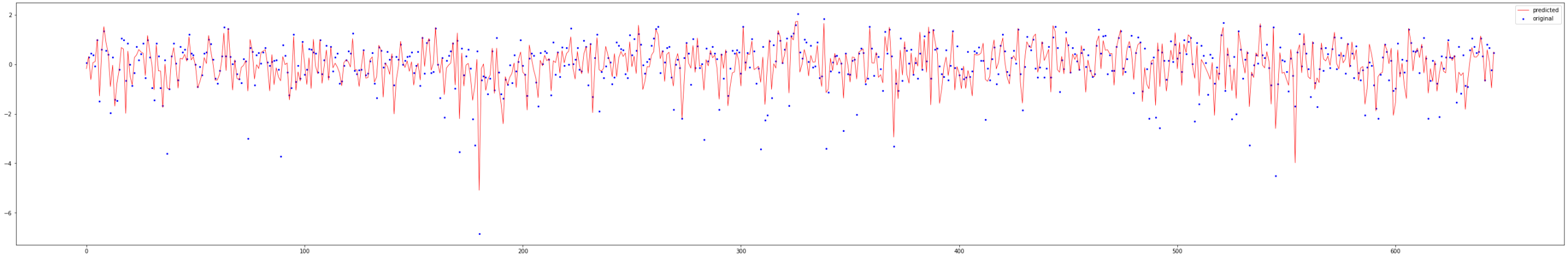
```
elasticNet.intercept_
```

```
-0.006326669502023015
```

```
rmse
```

```
0.6204869847972504
```

Y = 0.4668*(budget) + 0.503(vote_count) – 0.025(num_of_genre) –0.0724(num_of_production_companies) – 0.0403(num_of_production_countries)-0.0064



04. Revenue Analysis

Model Comparison

	Train Accuracy	Test Accuracy	RMSE
Regression Model 1 – Linear Regression	62.2%	64.5%	0.6203
Regression Model 2 – Ridge Regression	62.3%	64.3%	0.6203
Regression Model 3 – ElasticNet Regression	62.2%	64.6%	0.6205

$$Y = 0.4741*(\text{budget}) + 0.5124(\text{vote\_count}) - 0.0327(\text{num\_of\_genre}) - 0.0809(\text{num\_of\_production\_companies}) - 0.0434(\text{num\_of\_production\_countries}) - 0.0064$$

Stony Brook University

Thank You!