TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN

KHOA HỌC DỮ LIỆU

DỰ ĐOÁN DOANH THU PHIM

HỌ VÀ TÊN SINH VIÊN	LỚP HỌC PHẦN	ĐIỂM BẢO VỆ
Trương Minh Đức	19N13	
Trần Như Trí	19N13	
Huỳnh Thị Ái Linh	19N13	

ĐÀ NẴNG, 07/2022

TÓM TẮT

Các nhà làm phim đều đồng tình với quan điểm rằng mọi khâu trong quá trình làm phim từ sản xuất, làm kịch bản, tìm kiếm diễn viên giàu kinh nghiệm diễn xuất, hậu kỳ đến phát hành, trình chiếu... những yếu tố này là thước đo quan trọng đánh giá bộ phim có thành công hay không. Nhà làm phim cần chuẩn bị từng mảng phải thật kỹ lưỡng, tương ứng phải đầu tư nguồn kinh phí lớn để có được doanh thu mong muốn. Từ đó, nhóm chúng em xây dựng một chương trình dự đoán doanh thu phim dựa trên một số thông tin về bộ phim, dữ liệu được sử dụng ở phần Movie Budgets trên website *The Numbers*, làm sạch dữ liệu, chuẩn hóa dữ liệu, lựa chọn đặc trưng, sử dụng hai mô hình Linear Regression và Support Vector Regression để dự đoán và so sánh bằng MAE, MSE, R2.

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá
Huỳnh Thị Ái Linh	-Thu thập dữ liệu -Mô tả dữ liệu -Tổng hợp, viết báo cáo	Đã hoàn thành
Trần Như Trí	-Lựa chọn đặc trưng -Làm sạch và chuẩn hóa dữ liệu -Trực quan hóa kết quả của quá trình	Đã hoàn thành
Trương Minh Đức	-Chọn 2 mô hình -Huấn luyện mô hình -Đánh giá, so sánh hiệu quả của các mô hình	Đã hoàn thành

MỤC LỤC

DANH SÁCH BẢNG

DANH SÁCH HÌNH

1. Giới thiệu

Vấn đề cần giải quyết là dữ liệu thu thập được phải chính xác với thực tế, đầy đủ và số lượng nhiều. Đối với dữ liệu trống, phải có cách xử lý phù hợp để góp phần giúp cho bài toán được cải thiện một cách đáng kể hơn, xử lý dữ liệu ngoại lệ và chuẩn hóa dữ liệu, chọn được mô hình phù hợp, chia tỷ lệ các tập Huấn luyện/Xác thực/Kiểm thử phù hợp, đánh giá các mô hình dựa trên các độ đo.

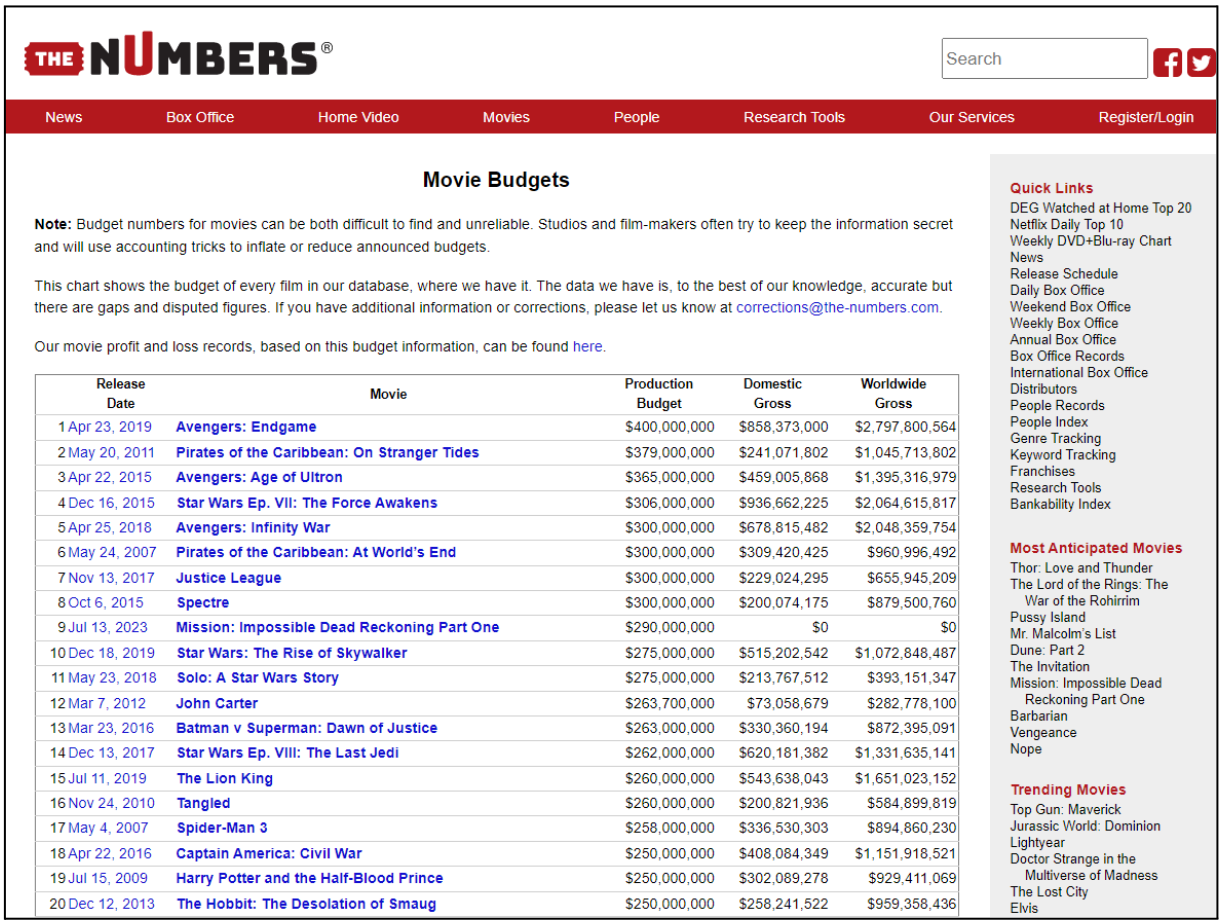
2. Thu thập và mô tả dữ liệu

2.1. Thu thập dữ liệu

Thu thập dữ liệu là quá trình tự động trích xuất các thông tin từ các trang web và lưu trữ nó dưới một định dạng phù hợp. Thông thường, khi muốn lấy một số thông tin từ các trang web, chúng ta sẽ dùng các API mà các trang đó cung cấp. Đây là cách đơn giản, tuy nhiên không phải trang web nào cũng cung cấp sẵn API cho chúng ta sử dụng. Do đó chúng ta cần một kỹ thuật để lấy các thông tin từ các trang web đó mà không thông qua API.

Lưu ý rằng: Ngân sách của một bộ phim rất khó tìm và nếu có cũng không đáng tin, vì các hãng làm phim thường giữ bí mật thông tin. Các nhà sản xuất phim thường công bố ngân sách sau khi đã tăng hoặc giảm so với thực tế.

Tuy vậy, ở phần Movie Budgets trên trang web The Numbers [1] thống kê một bảng ngân sách chính xác của các bộ phim mà họ thu thập được, tuy nhiên cũng có những chỗ trống và những chỗ số liệu còn đang tranh cãi.



THE NUMBERS®

Search

News | Box Office | Home Video | Movies | People | Research Tools | Our Services | Register/Login

Movie Budgets

Note: Budget numbers for movies can be both difficult to find and unreliable. Studios and film-makers often try to keep the information secret and will use accounting tricks to inflate or reduce announced budgets.

This chart shows the budget of every film in our database, where we have it. The data we have is, to the best of our knowledge, accurate but there are gaps and disputed figures. If you have additional information or corrections, please let us know at corrections@the-numbers.com.

Our movie profit and loss records, based on this budget information, can be found here.

	Release Date	Movie	Production Budget	Domestic Gross	Worldwide Gross
1	Apr 23, 2019	Avengers: Endgame	$400,000,000	$858,373,000	$2,797,800,564
2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$379,000,000	$241,071,802	$1,045,713,802
3	Apr 22, 2015	Avengers: Age of Ultron	$365,000,000	$459,005,868	$1,395,316,979
4	Dec 16, 2015	Star Wars Ep. VII: The Force Awakens	$306,000,000	$936,662,225	$2,064,615,817
5	Apr 25, 2018	Avengers: Infinity War	$300,000,000	$678,815,482	$2,048,359,754
6	May 24, 2007	Pirates of the Caribbean: At World's End	$300,000,000	$309,420,425	$960,996,492
7	Nov 13, 2017	Justice League	$300,000,000	$229,024,295	$655,945,209
8	Oct 6, 2015	Spectre	$300,000,000	$200,074,175	$879,500,760
9	Jul 13, 2023	Mission: Impossible Dead Reckoning Part One	$290,000,000	$0	$0
10	Dec 18, 2019	Star Wars: The Rise of Skywalker	$275,000,000	$515,202,542	$1,072,848,487
11	May 23, 2018	Solo: A Star Wars Story	$275,000,000	$213,767,512	$393,151,347
12	Mar 7, 2012	John Carter	$263,700,000	$73,058,679	$282,778,100
13	Mar 23, 2016	Batman v Superman: Dawn of Justice	$263,000,000	$330,360,194	$872,395,091
14	Dec 13, 2017	Star Wars Ep. VIII: The Last Jedi	$262,000,000	$620,181,382	$1,331,635,141
15	Jul 11, 2019	The Lion King	$260,000,000	$543,638,043	$1,651,023,152
16	Nov 24, 2010	Tangled	$260,000,000	$200,821,936	$584,899,819
17	May 4, 2007	Spider-Man 3	$258,000,000	$336,530,303	$894,860,230
18	Apr 22, 2016	Captain America: Civil War	$250,000,000	$408,084,349	$1,151,918,521
19	Jul 15, 2009	Harry Potter and the Half-Blood Prince	$250,000,000	$302,089,278	$929,411,069
20	Dec 12, 2013	The Hobbit: The Desolation of Smaug	$250,000,000	$258,241,522	$959,358,436

Quick Links
DEG Watched at Home Top 20
Netflix Daily Top 10
Weekly DVD+Blu-ray Chart
News
Release Schedule
Daily Box Office
Weekend Box Office
Weekly Box Office
Annual Box Office
Box Office Records
International Box Office
Distributors
People Records
People Index
Genre Tracking
Keyword Tracking
Franchises
Research Tools
Bankability Index

Most Anticipated Movies
Thor: Love and Thunder
The Lord of the Rings: The War of the Rohirrim
Pussy Island
Mr. Malcolm's List
Dune: Part 2
The Invitation
Mission: Impossible Dead Reckoning Part One
Barbarian
Vengeance
Nope

Trending Movies
Top Gun: Maverick
Jurassic World: Dominion
Lightyear
Doctor Strange in the Multiverse of Madness
The Lost City
Elvis

Hình 1. Trang Movie Budget của The Numbers

Đường dẫn đến nguồn dữ liệu: https://www.the-numbers.com/movie/budgets/all

Python cung cấp rất nhiều thư viện nhằm mục đích đọc dữ liệu từ các website như: lxml, Beautiful Soup, html5lib, … Nhóm em sử dụng Python 3 và BeautifulSoup 4 để trích chọn dữ liệu HTML.

Thư viện BeautifulSoup là một thư viện của Python cho phép chúng ta lấy dữ liệu từ HTML đơn giản và hiệu quả.

Mục tiêu là từ trang Movie Budgets, lấy tất cả các thông tin ở bảng của từng trang. Với mỗi dòng dữ liệu, ta sẽ truy cập vào trang chi tiết để lấy thông tin bảng Moive Details của từng bộ phim.

Trước hết, ta cần lấy nội dung của trang web. Vì trang web The Numbers sử dụng **mod_security** chặn yêu cầu. Thông thường, các trang web được bảo vệ bằng App Gateway, WAF, v.v., xem xét các yêu cầu là đến từ con người hay robot [2]. Mod_security hay WAF sẽ chặn các yêu cầu từ robot, coi chúng là spider/bot requests.

Khi yêu cầu đến bất kỳ trang web nào bằng **urllib.request** sẽ dễ bị phát hiện bởi mod_security. Mod_security hoạt động theo cách nếu có bất kỳ yêu cầu nào không có **user-agent** hợp lệ thì mod_security sẽ chặn yêu cầu và trả về

urllib.error.httperror: http error 403: forbidden. Vì vậy ta cần truyền **'User-Agent'** hợp lệ vào tham số 'headers'.

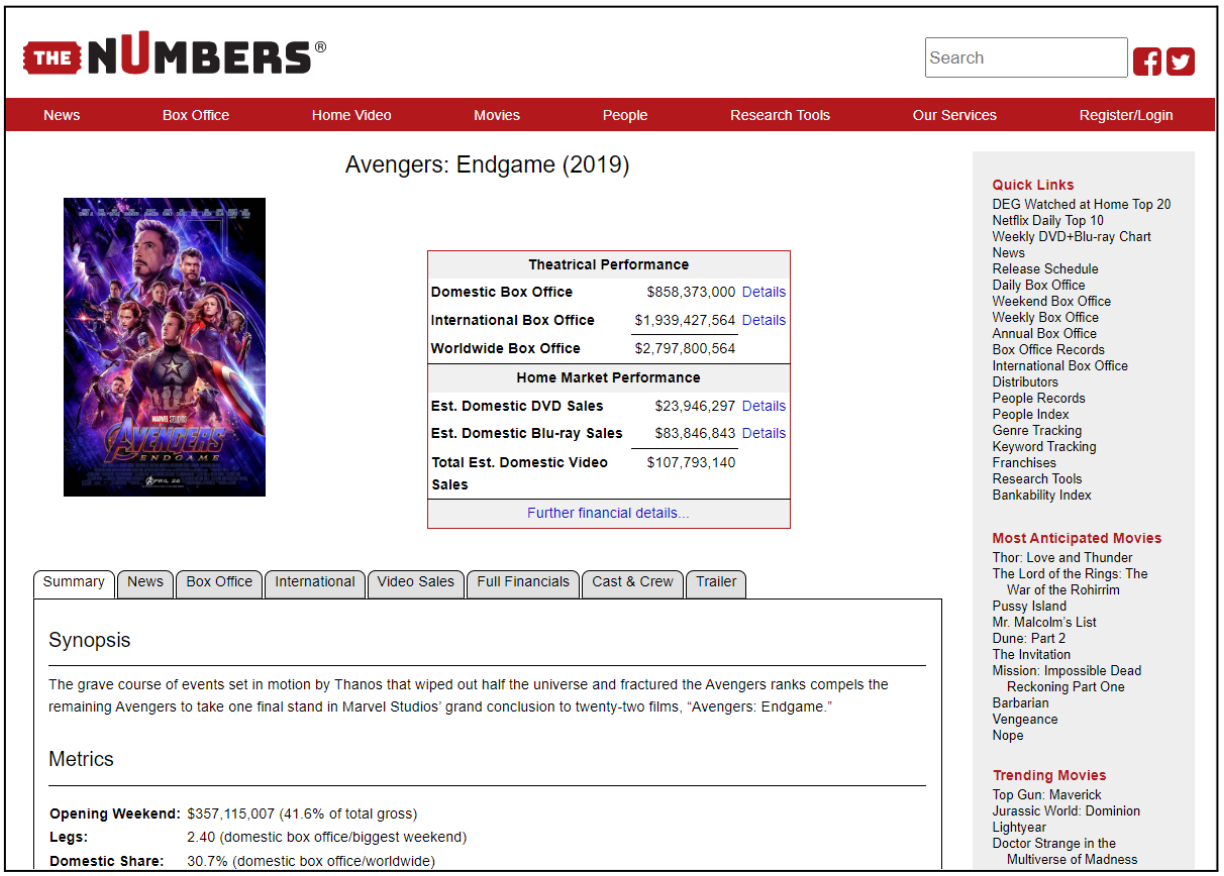
```
from bs4 import BeautifulSoup
from urllib.request import Request, urlopen


url = 'https://www.the-numbers.com/movie/budgets/all'
req = Request(url, headers={'User-Agent': 'Mozilla/5.0'})
webpage = urlopen(req).read()
soup = BeautifulSoup(webpage, 'html.parser')
```

Xem page source của trang web, phân tích được dữ liệu cần lấy nằm trong bảng. Mỗi trang chỉ có một bảng, nên chỉ cần duyệt qua từng trang, mỗi trang ta duyệt qua từng hàng bằng câu lệnh **for tr in soup.find_all('tr')** và duyệt qua từng ô trong hàng bằng câu lệnh **for td in tr.find_all('td')**.

```
<table >
<tr><th> </th><th>Release<BR>Date</th><th>Movie</th><t
h>Production<BR>Budget</th><th>Domestic<BR>Gross</th><th>Wo
rldwide<BR>Gross</th></tr>
<tr><td class="data">1</td>
<td><a href="/box-office-chart/daily/2019/04/23">Apr 23,
2019</a></td>
<td><b><a
href="/movie/Avengers-Endgame-(2019)#tab=summary">Avengers:
Endgame</a></td>
<td class="data"> $400,000,000</td>
<td class="data"> $858,373,000</td>
<td class="data"> $2,797,800,564</td>
</tr>
.
.
.
</table>
```

Riêng trong ô ở cột **"Movie"** là thẻ **<a>** chứa url dẫn đến trang chi tiết của từng bộ phim, gọi hàm **a = td.select_one('b > a')** để lấy được đường dẫn trong thuộc tính **<href>**. Vì trang chi tiết chứa nhiều bảng, ta sẽ tìm bảng bằng cách tìm kiếm phần tử thẻ **<h2>** có text là **"Movie Details"** bằng câu lệnh **h2 = detail_soup.find('h2', text='Movie Details')**, sau đó tìm bảng

nằm ngay sau thẻ **"h2"** đó bằng câu lệnh **`table = h2.find_next_sibling()`**, tương tự lấy những cột dữ liệu ta cần.



Hình 2. Trang chi tiết của bộ phim Avengers: Endgame (2019)

Ghi dữ liệu vào tệp csv sử dụng module **`csv`** [3]. Đầu tiên, mở tệp csv để ghi (chế độ **`w`**) sử dụng hàm **`open()`**. Tiếp theo, tạo một CSV writer bằng cách gọi hàm **`writer()`** của module csv. Ghi dữ liệu vào tệp csv dùng hàm **`writerow()`** hoặc **`writerows()`** của đối tượng CSV writer. Cuối cùng, đóng tệp khi hoàn thành việc ghi dữ liệu.

```
import csv

f = open('raw-data', 'w')
writer = csv.writer(f)
writer.writerow(row)
f.close()
```

2.2. Mô tả dữ liệu

Dữ liệu thô khi thu thập được lưu vào file raw-data.csv. Tổng cộng có 6286 mẫu, 14 đặc trưng: số thứ tự, ngày phát hành (ReleaseDate), tên phim (Movie), thời lượng (Running Time), nguồn (Source), thể loại (Genre), phương thức sản xuất (ProductionMethod), phong cách sáng tạo (CreativeType), công ty sản xuất

(ProductionCompanies), quốc gia sản xuất (ProductionCountries), ngôn ngữ (Languages), ngân sách sản xuất (ProductionBudget), doanh thu trong nước (DomesticGross), doanh thu thế giới (WorldwideGross).

	ReleaseDate	Movie	RunningTime	Source	Genre	ProductionMethod	CreativeType	ProductionCompanies	ProductionCountries	Languages	ProductionBudget	DomesticGross	WorldwideGross
1	Apr 23 2019	Avengers:	181 minutes	Based on	Action	Animation/Live Acti	Super Hero	Marvel Studios	United States	English	400000000	858373000	2797800564
2	May 20 2011	Pirates of	136 minutes	Based on	Adventur	Live Action	Historical Ficti	Walt Disney Pictures	United States	English	379000000	241071802	1045713802
3	Apr 22 2015	Avengers:	141 minutes	Based on	Action	Animation/Live Acti	Super Hero	Marvel Studios	United States	English	365000000	459005868	1395316979
4	Dec 16 2015	Star Wars	136 minutes	Original S	Adventur	Animation/Live Acti	Science Fictior	Lucasfilm, Bad Robot	United States	English	306000000	936662225	2064615817
5	Apr 25 2018	Avengers:	156 minutes	Based on	Action	Animation/Live Acti	Super Hero	Marvel Studios	United States	English	300000000	678815482	2048359754
6	May 24 2007	Pirates of	167 minutes	Based on	Adventur	Live Action	Historical Ficti	Walt Disney Pictures, Je	United States	English	300000000	309420425	960996492
7	Nov 13 2017	Justice Le	121 minutes	Based on	Action	Live Action	Super Hero	DC Films, RatPac Entert	United States	English	300000000	229024295	655945209
8	Oct 6 2015	Spectre	148 minutes	Based on	Action	Live Action	Contemporary	Eon Productions	United Kingdom, Unit	English	300000000	200074175	879500760
9	Jul 13 2023	Mission: Impossible Dead Reckonin			Action	Live Action		Paramount Pictures, Sk	United States	English	290000000	0	0
10	Dec 18 2019	Star Wars:	142 minutes	Original S	Adventur	Animation/Live Acti	Science Fictior	Lucasfilm, Bad Robot, V	United States	English	275000000	515202542	1072848487
11	May 23 2018	Solo: A Sta	135 minutes	Spin-Off	Adventur	Animation/Live Acti	Science Fictior	Lucasfilm	United States	English	275000000	213767512	393151347
12	Mar 7 2012	John Carte	132 minutes	Based on	Adventur	Live Action	Science Fictior	Walt Disney Pictures, Ji	United States	Apache, En	263700000	73058679	282778100
13	Mar 23 2016	Batman v	151 minutes	Based on	Action	Live Action	Super Hero	Warner Bros., RatPac Er	United States	English	263000000	330360194	872395091
14	Dec 13 2017	Star Wars	150 minutes	Original S	Adventur	Animation/Live Acti	Science Fictior	Lucasfilm, Walt Disney	United States	English	262000000	620181382	1331635141
15	Jul 11 2019	The Lion K	118 minutes	Remake	Adventur	Animation/Live Acti	Kids Fiction	Walt Disney Pictures, Fi	United States	English	260000000	543638043	1651023152
16	Nov 24 2010	Tangled	101 minutes	Based on	Musical	Digital Animation	Kids Fiction	Walt Disney Animation	United States	English	260000000	200821936	584899819
17	May 4 2007	Spider-Ma	139 minutes	Based on	Adventur	Live Action	Super Hero	Columbia Pictures, Mar	United States	English	258000000	336530303	894860230
18	Apr 22 2016	Captain Ai	146 minutes	Based on	Action	Live Action	Super Hero	Marvel Studios	United States	English	250000000	408084349	1151918521
19	Jul 15 2009	Harry Pott	153 minutes	Based on	Adventur	Animation/Live Acti	Fantasy	Heyday Films	United Kingdom, Unit	English	250000000	302089278	929411069
20	Dec 12 2013	The Hobbi	201 minutes	Based on	Adventur	Animation/Live Acti	Fantasy	Wingnut Films	New Zealand, United	English	250000000	258241522	959358436

Hình 3. Tệp dữ liệu thô thu được khi crawl dữ liệu (raw-data.csv)

Để kiểm tra dữ liệu trống với dữ liệu trên, ta thấy ở cột ReleaseDate, thông tin thiếu có giá trị là 'Unknown', ta cần thay thế chúng bằng Nan. Tương tự, những phim không tìm kiếm được doanh thu chính xác hoặc những phim chưa phát hành có dữ liệu ở cột DomesticGross và WorldwideGross bằng 0, ta sẽ thay thế các ô ở hai cột này từ giá trị 0 thành giá trị Nan.

	ReleaseDate	Movie	RunningTime	Source	Genre	ProductionMethod	CreativeType	ProductionCompanies	ProductionCountries	Languages	ProductionBudget	DomesticGross	WorldwideGross
258	Oct 3 2019	Gemini M	116 minutes	Original S	Action	Animation/Live Acti	Science Fictior	Paramount Pictures, Sk	United States	English	140000000	48546770	166623705
259	Feb 24 2016	Gods of Eg	127 minutes	Original S	Adventur	Live Action	Fantasy	Summit Entertainment,	United States	English	140000000	31153464	138836756
260	Unknown	Desert Warrior				Live Action			Saudi Arabia		140000000	0	0
261	May 3 2002	Spider-Ma	121 minutes	Based on	Adventur	Live Action	Super Hero	Columbia Pictures, Mar	United States	English	139000000	403706375	821706375
262	Mar 6 2009	Watchmei	161 minutes	Based on	Action	Live Action	Super Hero	Legendary Pictures, Lav	United States	English	138000000	107509799	186976250

Hình 4. Dữ liệu 'ReleaseDate' trống có giá trị 'Unknow'

	ReleaseDate	Movie	RunningTime	Source	Genre	ProductionMethod	CreativeType	ProductionCompanies	ProductionCountries	Languages	ProductionBudget	DomesticGross	WorldwideGross
7	Nov 13 2017	Justice Le	121 minutes	Based on	Action	Live Action	Super Hero	DC Films, RatPac Entert	United States	English	300000000	229024295	655945209
8	Oct 6 2015	Spectre	148 minutes	Based on	Action	Live Action	Contemporary	Eon Productions	United Kingdom, Unit	English	300000000	200074175	879500760
9	Jul 13 2023	Mission: Impossible Dead Reckonin			Action	Live Action		Paramount Pictures, Sk	United States	English	290000000	0	0
10	Dec 18 2019	Star Wars:	142 minutes	Original S	Adventur	Animation/Live Acti	Science Fictior	Lucasfilm, Bad Robot, V	United States	English	275000000	515202542	1072848487
11	May 23 2018	Solo: A Sta	135 minutes	Spin-Off	Adventur	Animation/Live Acti	Science Fictior	Lucasfilm	United States	English	275000000	213767512	393151347

Hình 5. Dữ liệu 'DomesticGross' và 'WorldwideGross' trống có giá trị 0

```
df['ReleaseDate'] = df['ReleaseDate'].replace('Unknown', np.nan)
df['DomesticGross'] = df['DomesticGross'].replace(0, np.nan)
df['WorldwideGross'] = df['WorldwideGross'].replace(0, np.nan)
```

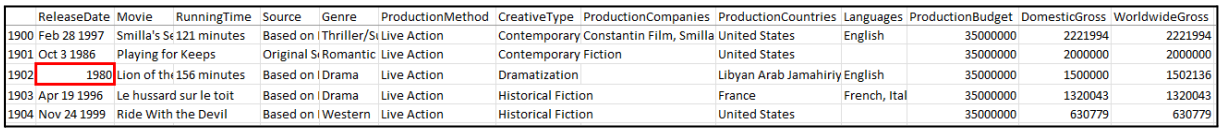
Sử dụng **`df.isnull().sum()`** trả về một bảng tóm tắt của tất cả các giá trị bị thiếu cho mỗi cột:

Id	0
ReleaseDate	114
Movie	0
RunningTime	1075
Source	253

Genre	160
ProductionMethod	176
CreativeType	303
ProductionCompanies	2334
ProductionCountries	498
Languages	1021
ProductionBudget	0
DomesticGross	679
WorldwideGross	412

Bảng 1. Số mẫu dữ liệu trống của mỗi đặc trưng

Quan sát dữ liệu nhận thấy, cột ReleaseDate chứa thông tin ngày, tháng năm phát hành bộ phim, nhưng có một vài bộ phim chỉ có năm. Để đồng bộ, ta sửa đổi các giá trị của cột này chỉ còn năm.

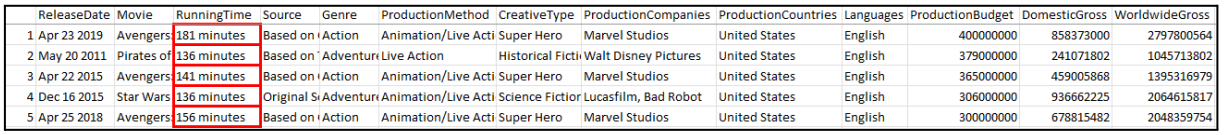


	ReleaseDate	Movie	RunningTime	Source	Genre	ProductionMethod	CreativeType	ProductionCompanies	ProductionCountries	Languages	ProductionBudget	DomesticGross	WorldwideGross
1900	Feb 28 1997	Smilla's Se	121 minutes	Based on	Thriller/Si	Live Action	Contemporary	Constantin Film, Smilla	United States	English	35000000	2221994	2221994
1901	Oct 3 1986	Playing for Keeps		Original Si	Romantic	Live Action	Contemporary Fiction		United States		35000000	2000000	2000000
1902	1980	Lion of the	156 minutes	Based on	Drama	Live Action	Dramatization		Libyan Arab Jamahiriy	English	35000000	1500000	1502136
1903	Apr 19 1996	Le hussard sur le toit		Based on	Drama	Live Action	Historical Fiction		France	French, Ital	35000000	1320043	1320043
1904	Nov 24 1999	Ride With the Devil		Based on	Western	Live Action	Historical Fiction		United States		35000000	630779	630779

Hình 6. Dữ liệu 'ReleaseDate' chỉ có giá trị năm

```
df['ReleaseDate'] = df['ReleaseDate'].transform(lambda x:
str(x).split(' ')[-1])
```

Thêm nữa, dữ liệu ở cột RunningTime cùng đơn vị 'minutes' lặp lại ở mỗi ô, cần xóa đơn vị và chỉ giữ lại giá trị số phút.



	ReleaseDate	Movie	RunningTime	Source	Genre	ProductionMethod	CreativeType	ProductionCompanies	ProductionCountries	Languages	ProductionBudget	DomesticGross	WorldwideGross
1	Apr 23 2019	Avengers	181 minutes	Based on	Action	Animation/Live Acti	Super Hero	Marvel Studios	United States	English	400000000	858373000	2797800564
2	May 20 2011	Pirates of	136 minutes	Based on	Adventur	Live Action	Historical Ficti	Walt Disney Pictures	United States	English	379000000	241071802	1045713802
3	Apr 22 2015	Avengers	141 minutes	Based on	Action	Animation/Live Acti	Super Hero	Marvel Studios	United States	English	365000000	459005868	1395316979
4	Dec 16 2015	Star Wars	136 minutes	Original Si	Adventur	Animation/Live Acti	Science Fictior	Lucasfilm, Bad Robot	United States	English	306000000	936662225	2064615817
5	Apr 25 2018	Avengers	156 minutes	Based on	Action	Animation/Live Acti	Super Hero	Marvel Studios	United States	English	300000000	678815482	2048359754

Hình 7. Dữ liệu 'RunningTime' đều có đơn vị 'minutes' ở sau

```
df['RunningTime'] = df['RunningTime'].transform(lambda x:
str(x).split(' ')[0])
```

Sau khi xử lý những bước cần thiết trên, toàn bộ dữ liệu có hai loại:

- **Dữ liệu số (numeric):** Id, ReleaseDate, RunningTime, ProductionBudget, DomesticGross, WorldwideGross (6 đặc trưng)

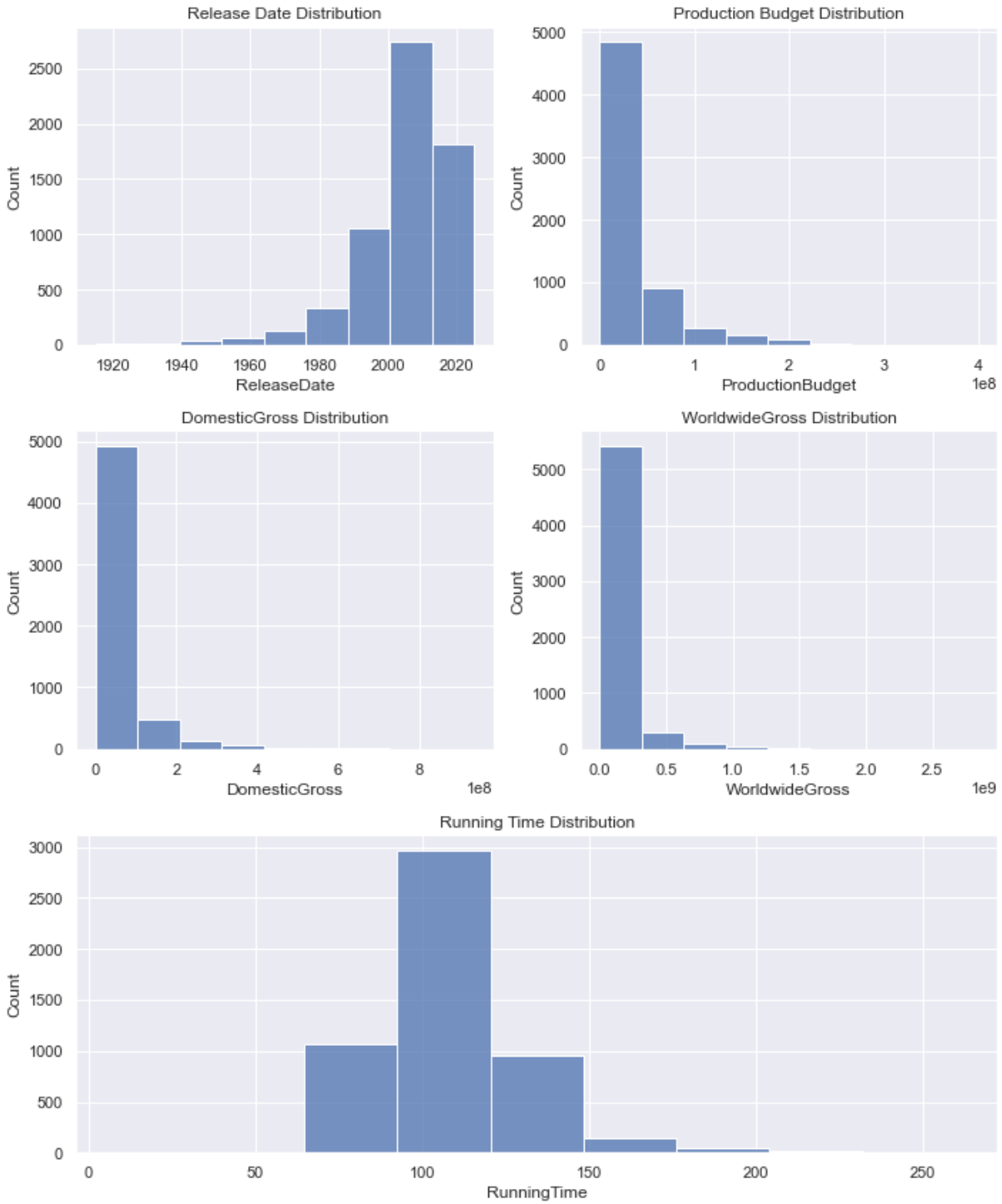
- **Dữ liệu phân loại/danh mục (categorical):** Movie, Source, Genre, ProductionMethod, CreativeType, ProductionCompanies, ProductionCountries, Languages (8 đặc trưng)

  Thống kê tóm tắt của dữ liệu số khi gọi hàm **`df.describe()`**

	Release Date	Running Time	Production Budget	Domestic Gross	Worldwide Gross
count	6172.0	5211.0	6286.0	5607.0	5874.0
mean	2004.72	108.41	31973171.21	46319102.99	98088872.39
std	13.08	20.18	42593533.51	72414852.74	185551494.12
min	1915.0	9.0	86.0	264.0	17.0
25%	2000.0	95.0	5000000.0	4210791.5	6662037.0
50%	2007.0	105.0	16800000.0	21416355.0	32243745.0
75%	2014.0	119.0	40000000.0	56984084.5	102979507.25
max	2025.0	260.0	400000000.0	936662225.0	2845899541.0

Bảng 2. Thống kê tóm tắt của dữ liệu số

Biểu đồ phân bố tần số (histogram) dùng để đo tần số xuất hiện của một đặc trưng, cho thấy rõ hình ảnh sự thay đổi, biến động của một tập dữ liệu.



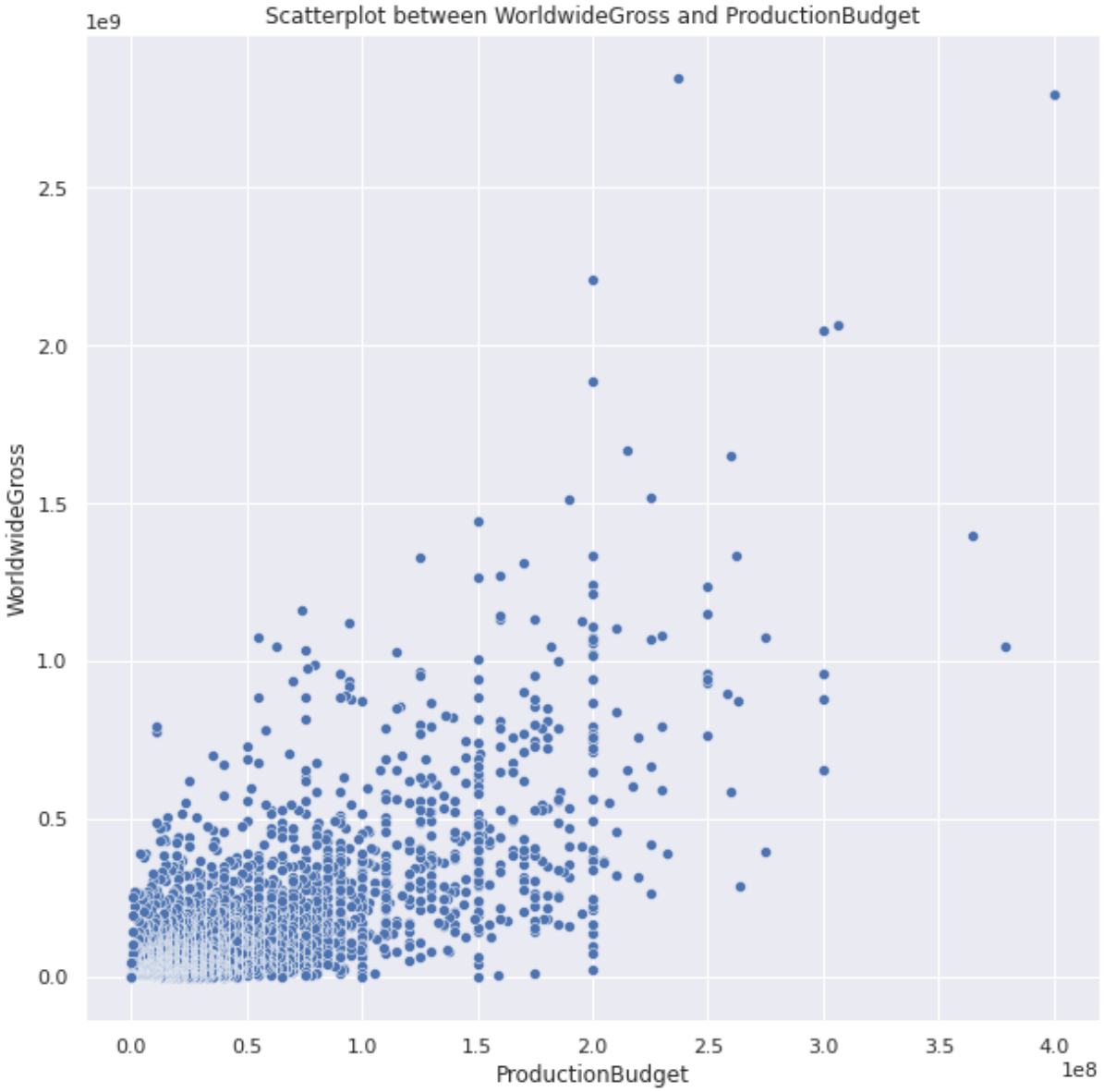
Hình 8. Biểu đồ tần suất của 5 đặc trưng số

Từ histogram ta thấy dữ liệu ở cột ReleaseDate bị lệch phải, còn ở cột ProductionBudget, DomesticGross, WorldwideGross bị lệch trái, dữ liệu ở cột RunningTime phân bố chuẩn.

Biểu đồ phân tán (scatterplot) là một biểu đồ có nhiều điểm dữ liệu có thể được vẽ bằng seaborn. Seaborn là một module Python để trực quan hóa dữ liệu thống kê. Seaborn có thể tạo biểu đồ này bằng phương thức `scatterplot()`. Các điểm dữ liệu được truyền vào tham số `data`. Các tham số `x` và `y` là nhãn của đồ thị.

Dưới đây là đồ thị phân tán sử dụng tập ‘WorldwideGross’ - đây là tập dữ liệu về doanh thu của phim trên toàn thế giới.

```
sns.scatterplot(x=df['ProductionBudget'],
y=df['WorldwideGross'], data=df['WorldwideGross'])
```

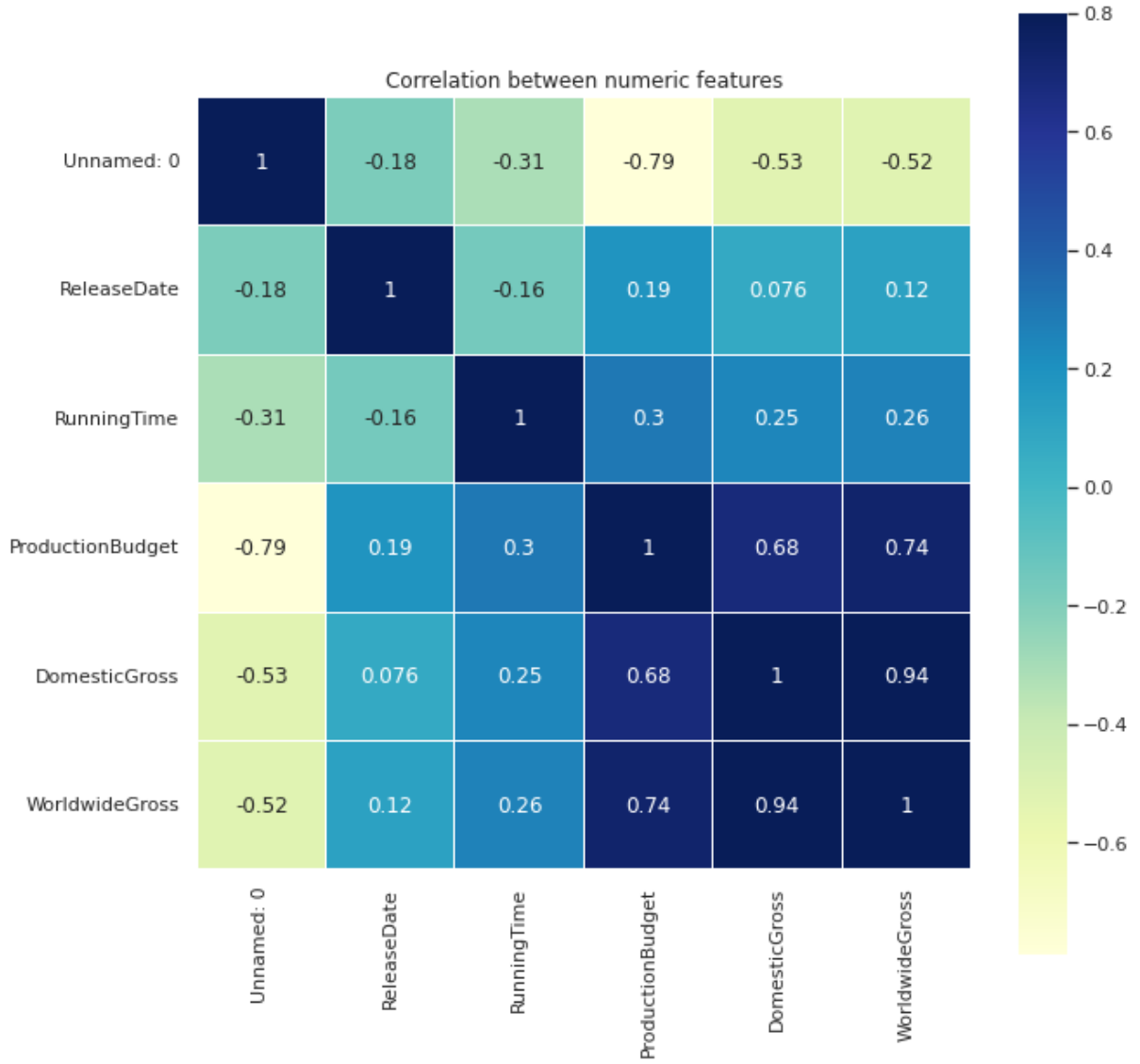


Hình 9. Biểu đồ phân tán giữa “WorldwideGross” và “ProductionBudget”

Trong hình trên, biểu đồ phân tán cho chúng ta biết được mối quan hệ giữa ngân sách làm một phim và doanh thu trên thế giới của bộ phim đó. Có thể thấy hai yếu tố này có một mối quan hệ tỷ lệ thuận với nhau, khi ngân sách tăng lên thì doanh thu của phim cũng có xu hướng tăng theo.

Biểu đồ nhiệt (heatmap) là một biểu đồ dưới dạng ma trận được mã hóa màu, cần một tập dữ liệu 2D làm tham số. Biểu đồ này trực quan hóa dữ liệu vì nó biểu thị mối quan hệ giữa các biến.

Biểu đồ nhiệt dưới đây dựa trên các dữ liệu số có trong tập dữ liệu: `sns.heatmap(df.corr())`



Hình 10. Biểu đồ nhiệt giữa các dữ liệu số

Nhìn vào heatmap ở trên, ta có thể nhận thấy 'WorldwideGross' tương quan mạnh với DomesticGross, theo sau là ProductionBudget, RunningTime, ReleaseDate.

3. Trích xuất đặc trưng

3.1. Xử lý dữ liệu trống

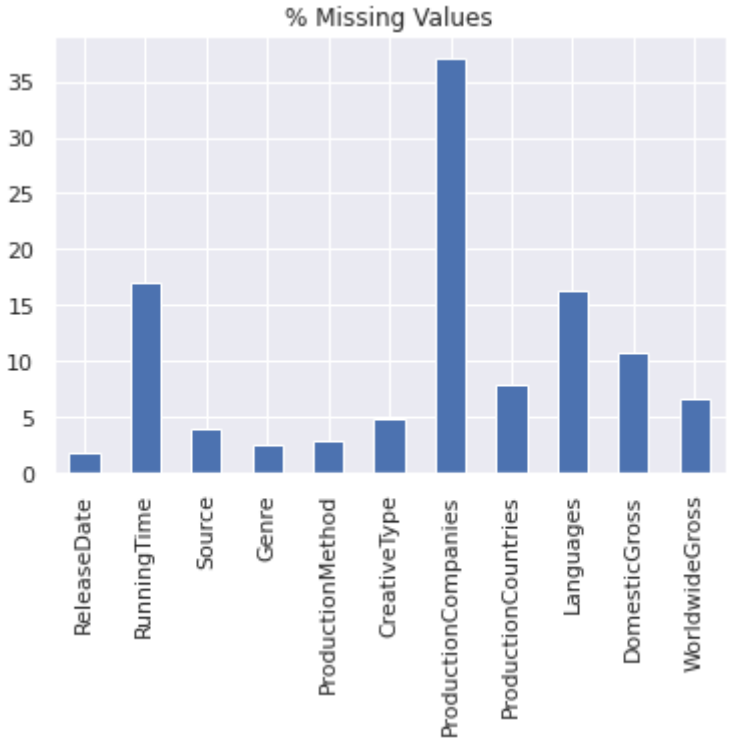
Việc xử lý dữ liệu trống rất quan trọng, bởi vì một số thuật toán học máy không thể chấp nhận dữ liệu trống. Nhưng việc điền các giá trị trống bằng một giá trị khác (mean, median, mode, …) cũng làm cho mô hình dự đoán không chính xác 100%

Như đã mô tả ở phần trên, tập dữ liệu gồm 11 cột có dữ liệu trống, xác suất dữ liệu bị trống là như nhau đối với mọi quan sát, không có mối quan hệ nào giữa dữ liệu trống và các dữ liệu khác. Vậy dữ liệu trống là hoàn toàn ngẫu nhiên.

Phần trăm dữ liệu bị trống của mỗi đặc trưng:

ReleaseDate	1.81%
RunningTime	17.1%
Source	4.02%
Genre	2.54%
ProductionMethod	2.8%
CreativeType	4.82%
ProductionCompanies	37.13%
ProductionCountries	7.92%
Languages	16.24%
DomesticGross	10.8%
WorldwideGross	6.55%

Bảng 3. Thống kê phần trăm dữ liệu trống của mỗi đặc trưng



Hình 11. Biểu đồ thống kê dữ liệu trống của mỗi đặc trưng

Trước hết, xóa các dòng có phần tử dữ liệu ở cột **'DomesticGross', 'WorldwideGross'** bị bỏ trống, vì đây là giá trị dự đoán.

```
df = df[df['DomesticGross'].notna()]
```

```
df = df[df['WorldwideGross'].notna()]
```

Lúc này, kiểm tra lại tập dữ liệu, số lượng dữ liệu trống của mỗi đặc trưng sau khi xóa những mẫu không có doanh thu.

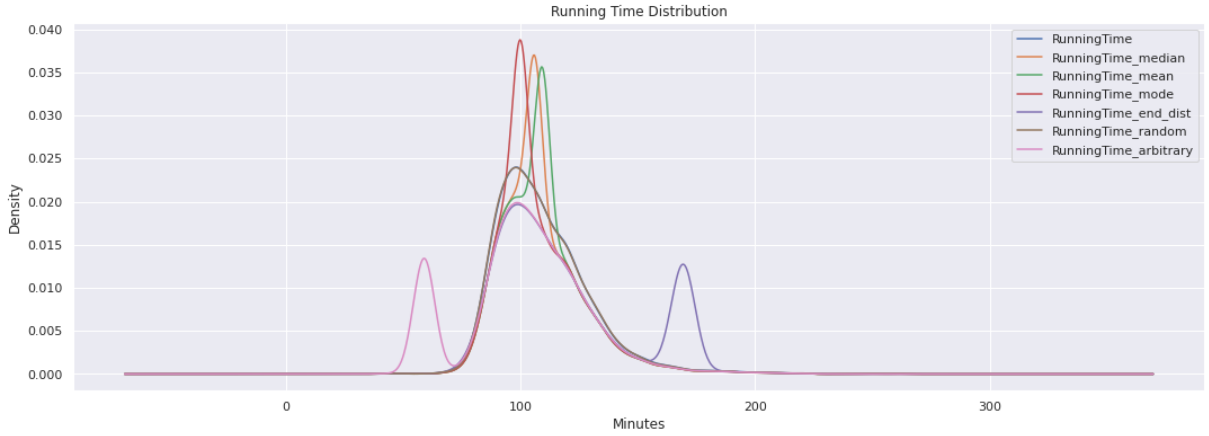
Unnamed	0
ReleaseDate	1
Movie	0
RunningTime	867
Source	189
Genre	117
ProductionMethod	127
CreativeType	247
ProductionCompanies	1924
ProductionCountries	441
Languages	939
ProductionBudget	0
DomesticGross	0
WorldwideGross	0

Bảng 4. Số mẫu dữ liệu trống của mỗi đặc trưng

Dữ liệu giảm đi đáng kể, chỉ còn 5607 dòng, vì vậy số lượng dữ liệu trống cũng giảm đi. Trong đó, số lượng dữ liệu trống của đặc trưng **'ReleaseDate'** chỉ còn 1 dòng, tương ưng 0.02% trong tập dữ liệu. Vì vậy, ta xóa luôn dòng có dữ liệu cột 'ReleaseDate' trống.

```
df = df[df['ReleaseDate'].notna()]
```

Vì cột **'RunningTime'** có tương quan với giá trị dự đoán nên không xóa dòng trống mà thử nghiệm với 7 cách thay thế giá trị của dữ liệu trống bằng giá trị trung bình (mean imputation), giá trị trung vị (median imputation), giá trị xuất hiện nhiều nhất (mode imputation), giá trị được lấy ngẫu nhiên từ dữ liệu (random sample imputation), giá trị tại đuôi của phân bố dữ liệu (end of distribution imputation), giá trị bất kỳ (arbitrary value imputation), tạo đặc trưng mới (create a new feature)



Hình 12. So sánh 7 kỹ thuật thay thế dữ liệu trống cho cột 'RunningTime'

Kết quả cho thấy sử dụng kỹ thuật xử lý dữ liệu trống bằng cách thay thế giá trị dữ liệu trống bằng giá trị được lấy ngẫu nhiên từ dữ liệu có độ lệch chuẩn gần nhất với độ lệch chuẩn ban đầu. Kỹ thuật này có ưu điểm là dễ thực hiện, phương sai của dữ liệu ít bị biến đổi. Thực hiện thay thế với đoạn lệnh sau:

```
random_samples = df['RunningTime'].dropna()
     .sample(n=df['RunningTime'].isnull().sum(),
          random_state=0)
df[df['RunningTime'].isnull()].index
random_samples.index = df[df['RunningTime'].isnull()]
     .index
df['RunningTime_random'] = df['RunningTime']
df.loc[df['RunningTime'].isnull(), 'RunningTime_random'] =
     random_samples
```

Đối với cột **'Source'**, sửa dữ liệu chỉ còn thông tin nguồn gốc, bỏ cụm từ 'Based on' ở đầu. Sau đó, duyệt qua từng dòng dữ liệu, nếu là các giá trị 'Original Screenplay', 'Fiction Book/Short Story', 'Real Life Events' thì giữ nguyên, còn lại thay thế bằng 'Others', dữ liệu trống cũng thay thế bằng 'Others'.

Đối với cột **'Genre'**, duyệt qua từng dòng dữ liệu, nếu là các giá trị 'Drama', 'Comedy', 'Action', 'Adventures', 'Thriller/Suspense', 'Horror' thì giữ nguyên, còn lại thay thế bằng 'Others', dữ liệu trống cũng thay thế bằng 'Others'.

Đối với cột **'ProductionMethod'**, duyệt qua từng dòng dữ liệu, nếu là các giá trị 'Live Action' thì giữ nguyên, còn lại thay thế bằng 'Others', dữ liệu trống cũng thay thế bằng 'Others'.

Đối với cột **'CreativeType'**, duyệt qua từng dòng dữ liệu, nếu là các giá trị 'Contemporary Fiction', 'Historical Fiction', 'Dramatization', 'Science Fiction' thì giữ nguyên, còn lại thay thế bằng 'Others', dữ liệu trống cũng thay thế bằng 'Others'.

Cột **‘ProductionCompanies’** chứa đến 3671 giá trị khác nhau, không có giá trị nào có số lần xuất hiện nhiều hơn hẳn, hơn 65% mẫu trong tập dữ liệu, vì vậy xóa cột này.

Đối với cột **‘ProductionCountries’**, duyệt qua từng dòng dữ liệu, nếu là các giá trị ‘United States’, ‘United Kingdom’ thì giữ nguyên, còn lại thay thế bằng ‘Others’, dữ liệu trống cũng thay thế bằng ‘Others’.

Đối với cột **‘Languages’**, duyệt qua từng dòng dữ liệu, nếu là các giá trị ‘English’ thì giữ nguyên, còn lại thay thế bằng ‘Others’, dữ liệu trống cũng thay thế bằng ‘Others’.

Ngoài ra, cột **‘Id’** không có ý nghĩa trong mô hình dự đoán nên cũng xóa đi.

Dữ liệu sau khi hoàn tất quá trình làm sạch thu được như trong hình 12. Lưu dữ liệu với tên ‘clean-data.csv’ có kích thước 5606 x 12.

ReleaseDate	Movie	Source	Genre	ProductionMethod	CreativeType	ProductionCountries	Languages	ProductionBudget	DomesticGross	WorldwideGross	RunningTime
2019	Avengers: Endgame	Comic/Graphic Novel	Action	Animation/Live Action	Super Hero	United States	English	400000000	858373000	2797800564	181
2011	Pirates of the Caribbean: On Stranger Tides	Theme Park Ride	Adventure	Live Action	Historical Fiction	United States	English	379000000	241071802	1045713802	136
2015	Avengers: Age of Ultron	Comic/Graphic Novel	Action	Animation/Live Action	Super Hero	United States	English	365000000	459005868	1395316979	141
2015	Star Wars Ep. VII: The Force Awakens		Adventure	Animation/Live Action	Science Fiction	United States	English	306000000	936662225	2064615817	136
2018	Avengers: Infinity War	Comic/Graphic Novel	Action	Animation/Live Action	Super Hero	United States	English	300000000	678815482	2048359754	156
2007	Pirates of the Caribbean: At Worldâ€™s End	Theme Park Ride	Adventure	Live Action	Historical Fiction	United States	English	300000000	309420425	960996492	167
2017	Justice League	Comic/Graphic Novel	Action	Live Action	Super Hero	United States	English	300000000	229024295	655945209	121
2015	Spectre	Fiction Book/Short Story	Action	Live Action	Contemporary Fiction	United States	English	300000000	200074175	879500760	148
2019	Star Wars: The Rise of Skywalker		Adventure	Animation/Live Action	Science Fiction	United States	English	275000000	515202542	1072848487	142
2018	Solo: A Star Wars Story		Adventure	Animation/Live Action	Science Fiction	United States	English	275000000	213767512	393151347	135
2012	John Carter	Fiction Book/Short Story	Adventure	Live Action	Science Fiction	United States	English	263700000	73058679	282778100	132
2016	Batman v Superman: Dawn of Justice	Comic/Graphic Novel	Action	Live Action	Super Hero	United States	English	263000000	330360194	872395091	151
2017	Star Wars Ep. VIII: The Last Jedi		Adventure	Animation/Live Action	Science Fiction	United States	English	262000000	620181382	1331635141	150
2019	The Lion King		Adventure	Animation/Live Action	Kids Fiction	United States	English	260000000	543638043	1651023152	118
2010	Tangled	Folk Tale/Legend/Fairytale	Musical	Digital Animation	Kids Fiction	United States	English	260000000	200821936	584899819	101
2007	Spider-Man 3	Comic/Graphic Novel	Adventure	Live Action	Super Hero	United States	English	258000000	336530303	894860230	139
2016	Captain America: Civil War	Comic/Graphic Novel	Action	Live Action	Super Hero	United States	English	250000000	408084349	1151918521	146
2009	Harry Potter and the Half-Blood Prince	Fiction Book/Short Story	Adventure	Animation/Live Action	Fantasy	United States	English	250000000	302089278	929411069	153
2013	The Hobbit: The Desolation of Smaug	Fiction Book/Short Story	Adventure	Animation/Live Action	Fantasy	United States	English	250000000	258241522	959358436	201
2014	The Hobbit: The Battle of the Five Armies	Fiction Book/Short Story	Adventure	Animation/Live Action	Fantasy	United States	English	250000000	255119788	940389558	144

Hình 13. Dữ liệu clean-data.csv

3.2. Mã hóa dữ liệu phân loại

Các thuật toán học máy không thể làm việc trực tiếp với dữ liệu dạng phân loại và cần thực hiện các kỹ thuật biến đổi trên dữ liệu dạng này để có thể đưa chúng vào các mô hình của mình.

Trong tập dữ liệu huấn luyện có những đặc trưng thuộc về dữ liệu phân loại là ‘ProductionCountries’, ‘ProductionMethod’, ‘Movie’, ‘Genre’, ‘Source’, ‘Languages’, ‘CreativeType’.

Có hai dạng chính của dữ liệu phân loại là danh nghĩa (nominal) và thứ tự (ordinal).

- Phân loại danh nghĩa là dạng dữ liệu không có khái niệm sắp xếp giữa các giá trị của thuộc tính đó, là các đặc trưng ‘ProductionMethod’, ‘CreativeType’, ‘Movie’, ‘Source’, ‘Genre’.
- Phân loại thứ tự là dạng dữ liệu mang một ý nghĩa về thứ tự giữa các giá trị của nó, các đặc trưng thuộc loại thứ tự là ‘ProductionCountries’, và ‘Languages’.

Nhóm sử dụng kỹ thuật One-hot Encoding, sau bước biến đổi, ta có biểu diễn dạng số của đặc trưng phân loại với m nhãn khác nhau. One-hot encoding là quá trình biến đổi từng giá trị thành các đặc trưng nhị phân chỉ chứa giá trị 1 hoặc 0. Mỗi mẫu trong đặc trưng phân loại sẽ được biến đổi thành một vecto có kích thước m chỉ với một trong các giá trị là 1 (biểu thị nó là active).

Tập dữ liệu sau khi chuẩn hóa có 4 đặc trưng nhị phân tạm thời được tạo ra cho đặc trưng 'Source', tương tự 7 cho 'Genre', 2 cho 'ProductionMethod', 5 cho 'CreativeType', 3 cho 'ProductionCountries' và 2 cho 'Languages'. Trạng thái active của đặc trưng được biểu thị bằng giá trị 1 trong đặc trưng tương ưng.

```
df = pd.get_dummies(df)
```

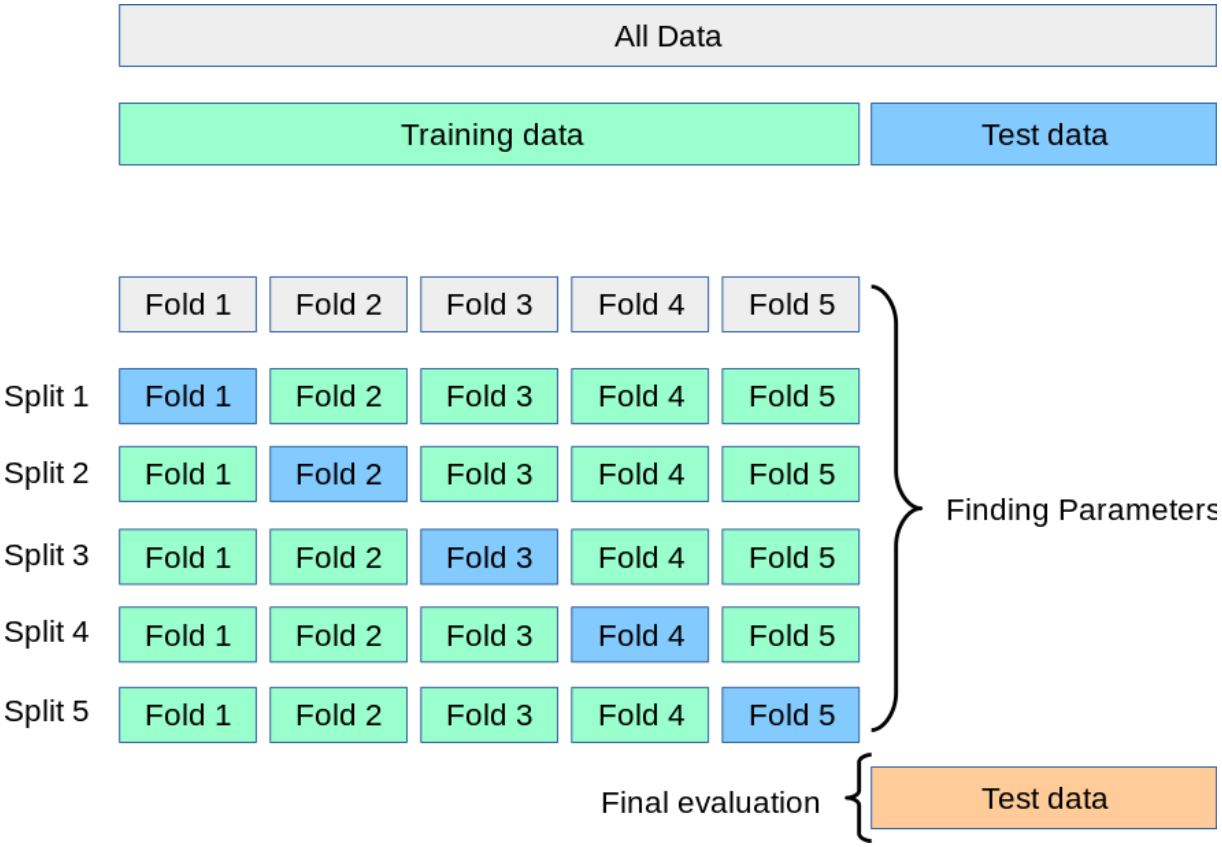
3.3. Chia dữ liệu thành các tập Huấn luyện/Xác thực/Kiểm thử

Ta cần kiểm thử để dự đoán khả năng hoạt động hiệu quả của mô hình trên thực tế. Ta sử dụng một tập dữ liệu kiểm thử độc lập đối với tập dữ liệu huấn luyện để đánh giá và ước lượng hiệu quả của mô hình. Lấy 80% dữ liệu trong tập clean-data.csv để huấn luyện và 20% dữ liệu còn lại không liên quan đến 80% trước để đánh giá.

Tập huấn luyện (training set) là tập dữ liệu được sử dụng để huấn luyện mô hình. Các thuật toán học máy sẽ học các mô hình từ tập huấn luyện này. Việc học sẽ khác nhau tùy thuộc vào thuật toán và mô hình sử dụng. Ví dụ, khi sử dụng mô hình Linear Regression, các điểm trong tập huấn luyện được sử dụng để tìm ra hàm số hay đường phù hợp nhất mô tả quan hệ giữa đầu vào và đầu ra của tập dữ liệu huấn luyện bằng cách sử dụng một số thuật toán tối ưu gần đúng như gradient descent.

Mục tiêu của machine learning là tạo ra những mô hình có khả năng tổng quát hóa để dự đoán tốt trên cả dữ liệu chưa thấy bao giờ (nằm ngoài tập huấn luyện), do đó, để biết một mô hình có tốt hay không thì sau khi được huấn luyện, mô hình cần được đánh giá hiệu quả thông qua bộ dữ liệu **kiểm thử (testing set)**. Bộ dữ liệu này được sử dụng để tính độ chính xác hoặc sai số của mô hình dự đoán đã được huấn luyện. Ta biết nhãn thực của mọi điểm trong tập dữ liệu kiểm thử này, nhưng ta sẽ tạm thời giả vờ như không biết và đưa các giá trị đầu vào của tập vào mô hình dự đoán để nhận kết quả dự đoán đầu ra. Sau đó ta có thể nhìn vào các nhãn thực, so sánh nó với kết quả dự đoán của các đầu vào tương ứng này, xem liệu mô hình có dự đoán đúng hay không.

Tập dữ liệu xác thực (validation set) cung cấp các đánh giá công bằng về sự phù hợp của mô hình trên tập dữ liệu huấn luyện trong quá trình huấn luyện. Tuy nhiên, Cross Validation cũng sẽ giúp ta đánh giá một mô hình đầy đủ và chính xác hơn, để sau đó đưa ra quyết định mô hình đó có phù hợp với dữ liệu, bài toán hiện tại hay không [5].



Hình 14. Grid search cross validation

Như trên hình 14, ta thấy:

- Phần dữ liệu Test data sẽ được để riêng và dành cho bước đánh giá cuối cùng nhằm kiểm tra "phản ứng" của mô hình khi gặp các dữ liệu unseen hoàn toàn.
- Phần dữ liệu Training sẽ được chia ngẫu nhiên thành K phần (K là một số nguyên, hay chọn là 5 hoặc 10). Sau đó train model K lần, mỗi lần train sẽ chọn 1 phần làm dữ liệu validation và K-1 phần còn lại làm dữ liệu training. Kết quả đánh giá model cuối cùng sẽ là trung bình cộng kết quả đánh giá của K lần train. Đó chính là lý do vì sao ta đánh giá khách quan và chính xác hơn.

Vậy, ta sử dụng hàm `train_test_split()` chia tập dữ liệu thành tập huấn luyện/Kiểm thử theo tỷ lệ 80/20.

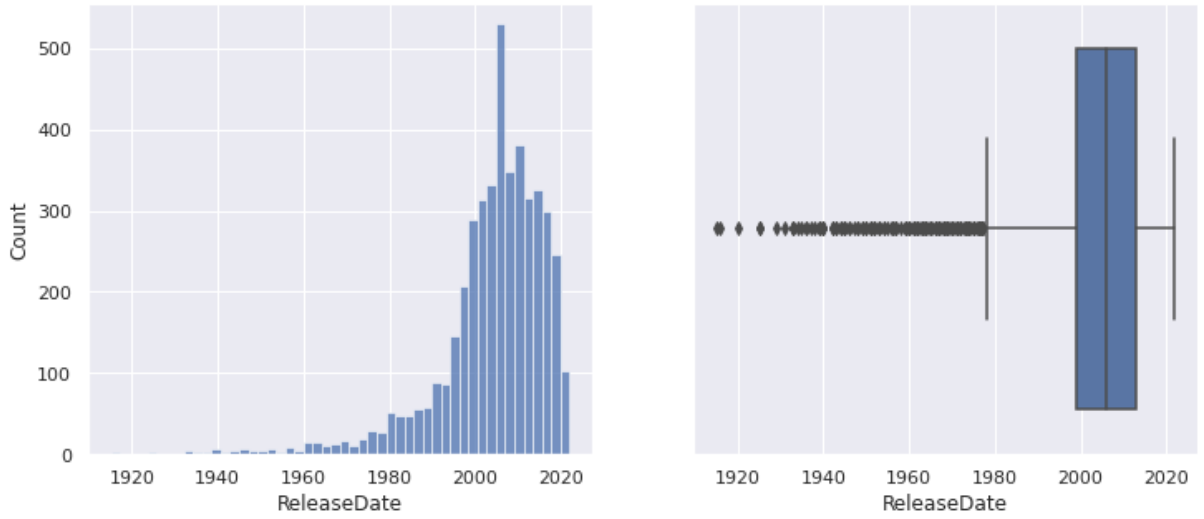
```
from sklearn.model_selection import train_test_split

df_train, df_test = train_test_split(df, text_size=0.2,
random_state=42)
```

3.4. Xử lý dữ liệu ngoại lệ cho tập huấn luyện

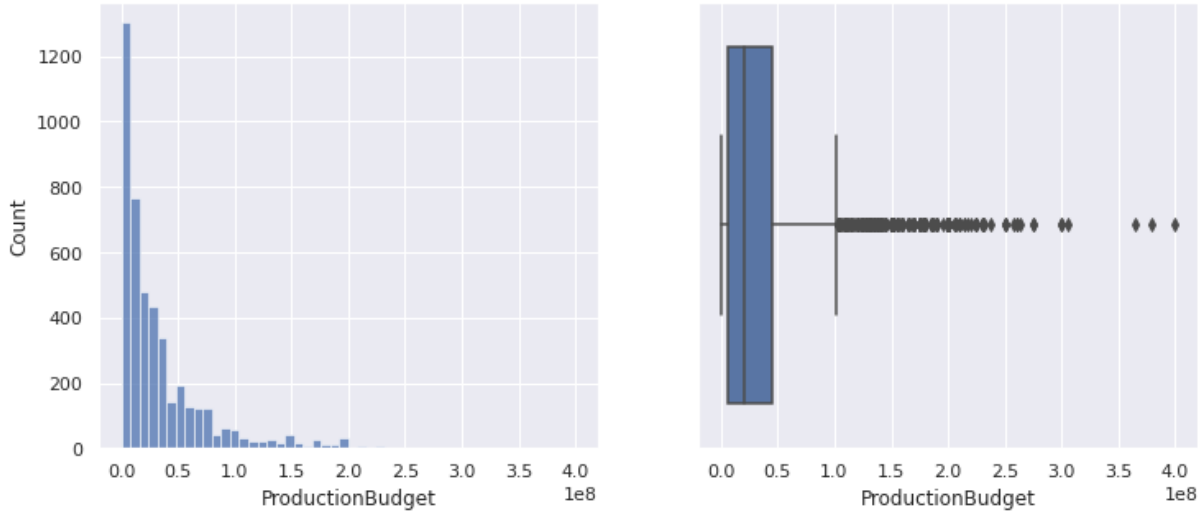
Với dạng số, dữ liệu ngoại lệ là các giá trị quá lớn hoặc quá nhỏ so với các giá trị khác của tập dữ liệu, được trực quan hóa bằng các vòng tròn nằm ngoài 2 vạch

(whisker) của Box plot. Dữ liệu ngoại lệ tác động xấu đến hiệu suất của các thuật toán học máy [6].

Dưới đây là histogram và box plot của ba cột ReleaseDate, ProductionBudget và RunningTime. Ở đây, box plot được vẽ ở dạng nằm ngang để so sánh với histogram.



Hình 15. Histogram và box plot của cột ReleaseDate trước khi xử lý ngoại lệ



Hình 15. Histogram và box plot của cột ProductionBudget trước khi xử lý ngoại lệ

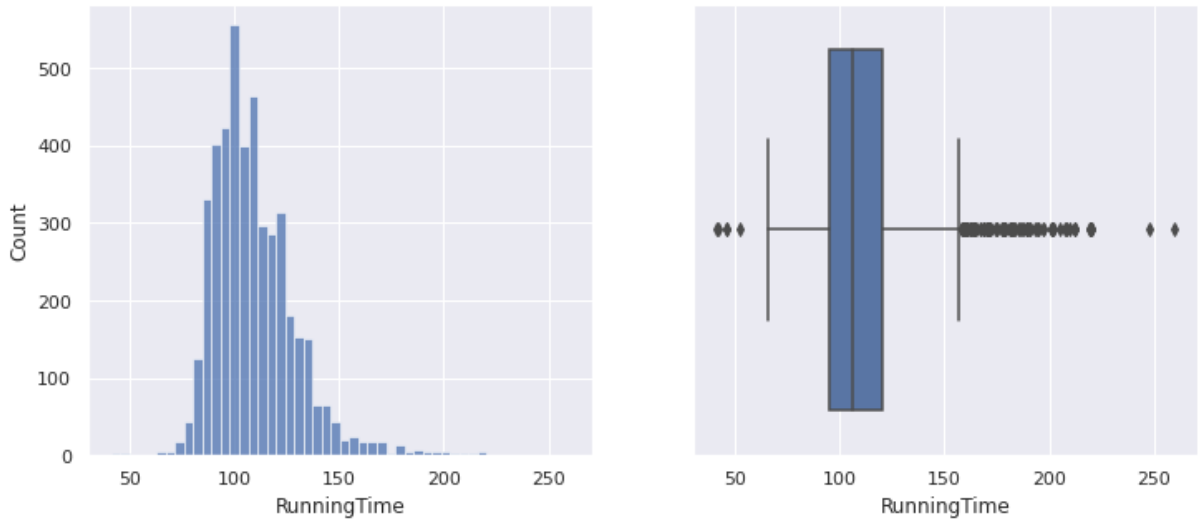
Từ histogram ta thấy dữ liệu ở cột ReleaseDate bị lệch phải (có điểm ngoại lệ nhiều về bên phải, hoặc "đuôi của histogram nằm ở bên phải), còn ở cột ProductionBudget bị lệch trái. Từ boxplot ta thấy có khá nhiều điểm được coi là ngoại lệ. Các điểm ngoại lệ có thể được xử lý bằng cách clip về giá trị cực tiểu và cực đại của Box plot theo công thức sau:

Biên trên = 3rd Quantile + 3*IQR

Biên dưới = 1st Quantile - 3*IQR

Hàm xử lý được cài đặt như sau:

```
def outliers_for_skewed_distribution(df, col):
    res = df.copy()
    q3 , q1 = np.percentile(res[col], [75,25])
    IQR = q3 - q1
    upper_boundary = q3 + 1.5 * IQR
    lower_boundary = q1 - 1.5 * IQR
    res[col][res[col] >= upper_boundary] = upper_boundary
    res[col][res[col] <= lower_boundary] = lower_boundary
    return res
```



Hình 16. Histogram và box plot của cột Running Time trước khi xử lý ngoại lệ

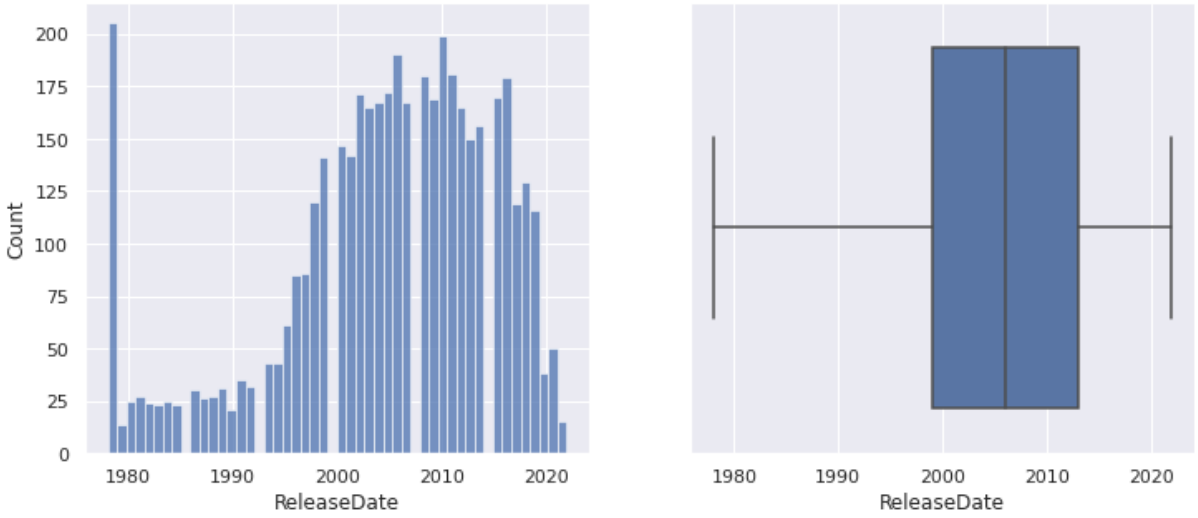
Cột RunningTime thì có dạng phân bố chuẩn, nên áp dụng công thức sau và cài đặt hàm như bên dưới:

Biên trên = GTTB + 3*Độ lệch chuẩn

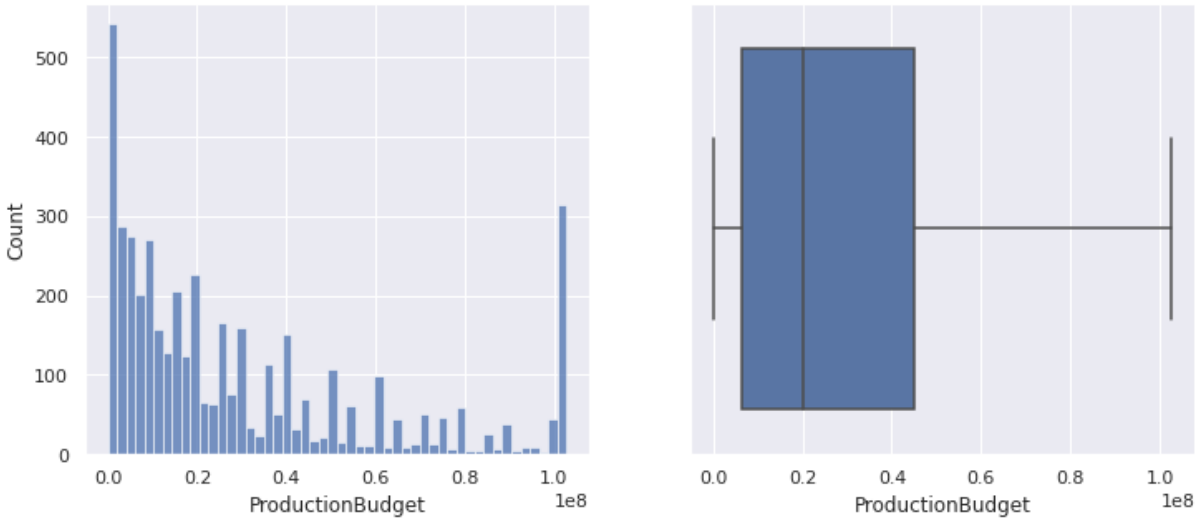
Biên dưới = GTTB - 3*Độ lệch chuẩn

```
def outliers_for_normal_distribution(df, col):
    res = df.copy()
    q3 , q1 = np.percentile(res[col], [75,25])
    IQR = q3 - q1
    mean = res[col].mean()
    upper_boundary = mean + 1.5 * IQR
    lower_boundary = mean - 1.5 * IQR
    res[col][res[col] >= upper_boundary] = upper_boundary
    res[col][res[col] <= lower_boundary] = lower_boundary
    return res
```

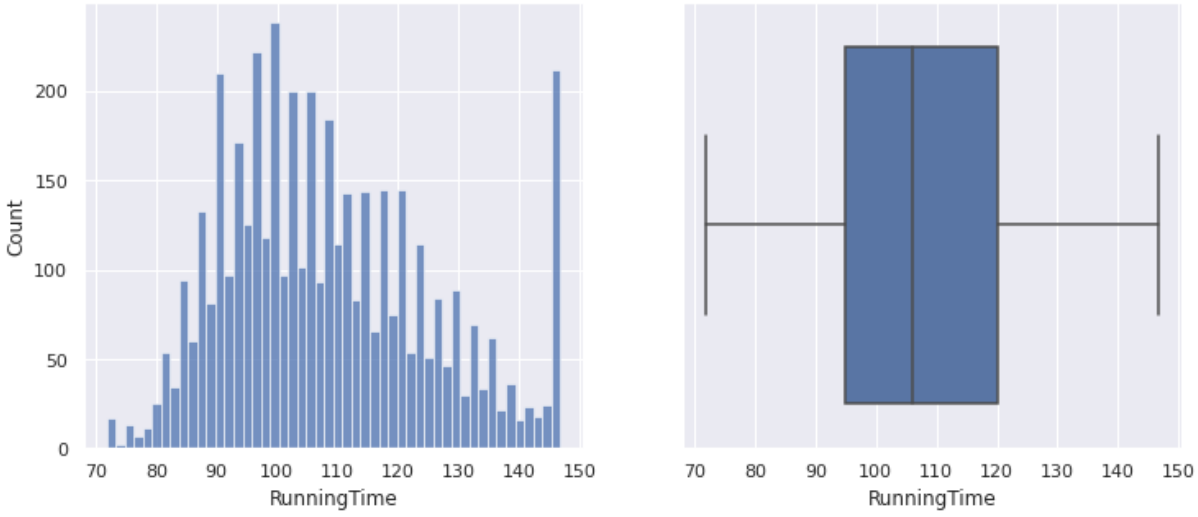
Áp dụng lại vào dữ liệu của năm cột trên ta có histogram và boxplot mới như sau:



Hình 17. Histogram và box plot của cột ReleaseDate sau khi xử lý ngoại lệ



Hình 18. Histogram và box plot của cột ProductionBudget sau khi xử lý ngoại lệ



Hình 19. Histogram và box plot của cột RunningTime sau khi xử lý ngoại lệ

Sau khi clip dữ liệu theo cực tiểu và cực đại của box plot, ta thấy rằng dữ liệu đỡ bị lệch đi. Boxplot cũng cho thấy không còn điểm dữ liệu ngoại lệ nào.

3.5. Chuẩn hóa dữ liệu

Việc chuẩn hóa dữ liệu là một bước vô cùng quan trọng trong việc giải quyết một vấn đề học máy. Vì tập dữ liệu chứa các đặc trưng khác nhau về giá trị, đơn vị, dải giá trị. Mà mỗi dòng dữ liệu thường được biểu diễn thành 1 điểm trong không gian vec-tơ nhiều chiều với số chiều bằng số đặc trưng. Cần phải làm cho các đặc trưng có giá trị tương đương nhau bằng cách chuẩn hóa dữ liệu.

Áp dụng kỹ thuật biến đổi dữ liệu lấy logarit, xử lý cột dữ liệu như sau:

```
df[col_log]=np.log1p(df[col])
```

4. Mô hình hóa dữ liệu

Mô hình nhóm em sử dụng là Linear Regression và Support Vector Regression (kernel = rbf).

4.1. Linear Regression - Hồi quy tuyến tính

Trong toán học, một đường thẳng trong hệ trục tọa Oxy có phương trình $y = mx + C$. Với m và C là hai hằng số, m được gọi là hệ số góc của đường thẳng, quyết định độ dốc của đường thẳng. C là tọa độ điểm mà đường thẳng sẽ cắt trục Oy.

Tuy nhiên trong khoa học dữ liệu hay các thuật toán Machine learning, phương trình đường thẳng sẽ được ký hiệu khác đi giúp cho việc áp dụng vào thuật toán dễ dàng hơn. Phương trình đường thẳng khi đó sẽ là $h_\theta(x) = \theta_0 + \theta_{1x}$.

Với một tập hợp điểm hỗn loạn, ta không thể vẽ một đường thẳng đi qua toàn bộ các điểm đó. Có rất nhiều các đường thẳng đi qua tập hợp điểm này, vậy làm thế nào để xác định được một đường thẳng là phương án tốt nhất?

Đường thẳng đi qua tập hợp điểm cách mỗi điểm một độ dư, độ lệch giữa điểm thực tế và điểm trên đường thẳng.

Hồi quy tuyến tính - Linear Regression có câu trả lời đường thẳng nào có các tổng các độ lệch bé nhất thì đó chính là đường thẳng tốt nhất. Nhưng do độ lệch này có thể có giá trị âm (ví dụ điểm thực tế nằm dưới đường thẳng), do đó chúng ta cần lấy bình phương của các độ lệch này, như vậy sẽ không còn giá trị âm và nó phản ánh đúng định hướng chúng ta cần.

Tóm lại công thức sử dụng cho Linear Regression là tìm các giá trị θ_0 và θ_1 sao cho tổng bình phương các độ lệch có giá trị thấp nhất [8].

4.2. Support Vector Regression - Hồi quy vector

4.2.1. Cơ sở lý thuyết

Với bài toán binary classification mà hai classes là *linearly separable*, có vô số các siêu mặt phẳng giúp phân biệt hai classes, tức mặt phân cách. Với mỗi mặt phân cách, ta có một *classifier*. Khoảng cách gần nhất từ một điểm dữ liệu tới mặt phân cách ấy được gọi là *margin* của classifier đó.

Support Vector Machine là bài toán đi tìm mặt phân cách sao cho *margin* tìm được là lớn nhất, đồng nghĩa với việc các điểm dữ liệu *an toàn nhất* so với mặt phân cách.

Bài toán tối ưu trong SVM là một bài toán lồi với hàm mục tiêu là *stricly convex*, nghiệm của bài toán này là duy nhất. Hơn nữa, bài toán tối ưu đó là một Quadratic Programming (QP).

Mặc dù có thể trực tiếp giải SVM qua bài toán tối ưu gốc này, thông thường người ta thường giải bài toán đối ngẫu. Bài toán đối ngẫu cũng là một QP nhưng nghiệm là *sparse* nên có những phương pháp giải hiệu quả hơn.

Với các bài toàn mà dữ liệu gần *linearly separable* hoặc *nonlinear separable*, có những cải tiến khác của SVM để thích nghi với dữ liệu đó.

4.2.1. Bộ tham số của mô hình

Khảo sát tham số dùng RandomizedSearchCV, cài đặt như sau:

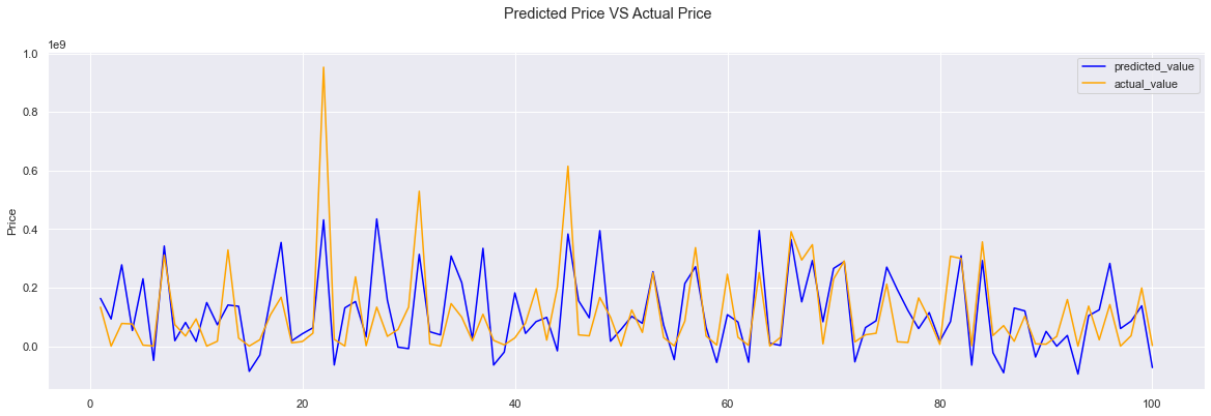
```
param_grid = {
    'kernel': ['linear', 'poly', 'rbf'],
    'degree': [3, 6, 9],
    'gamma': ['scale', 'auto']
}
random_search = RandomizedSearchCV(
     estimator=regressor,
     param_distributions=param_grid,
     n_iter=10,
     cv=5,
     scoring='max_error')
random_search.fit(X_train, y_train)
```

Kết quả trả về, gọi `random_search.best_params_`

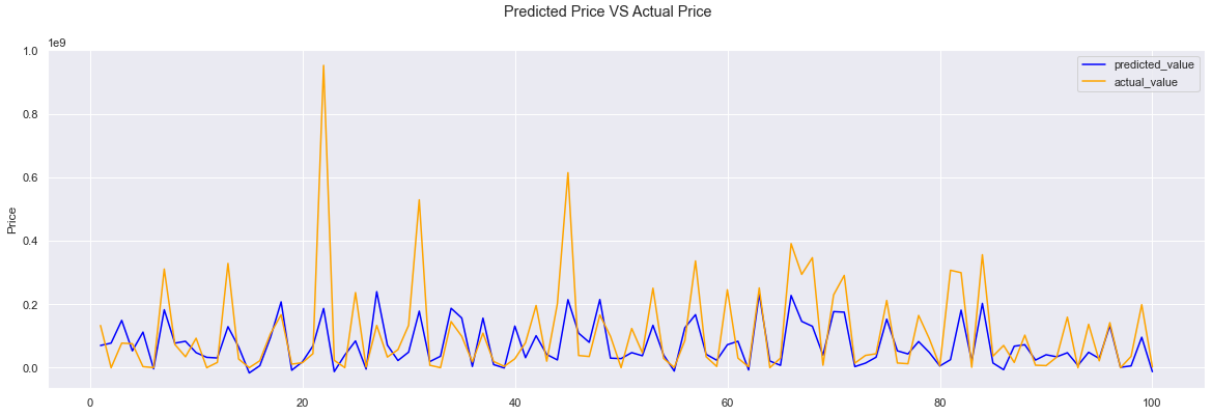
- `'kernel': 'poly'`
- `'gamma': 'auto'`

- `'degree': 6`

4.3. Hiệu suất của các mô hình



Hình 20. Đồ thị thể hiện hiệu suất của mô hình Linear Regression trên tập Kiểm thử



Hình 21. Đồ thị thể hiện hiệu suất của mô hình SVR trên tập Kiểm thử

Trình bày các đồ thị thể hiện hiệu suất (đánh giá bằng Loss hoặc Accuracy) của các mô hình trên các tập Huấn luyện/Xác thực/Kiểm thử.

4.4. Đánh giá cả mô hình, so sánh hiệu quả

	MAE	MSE	R2
Linear Regression	9.35391	22.49697	0.39841
Support Vector Regression	7.66448	27.38064	0.26782

Bảng 5. Số liệu các độ đo giữa hai mô hình

Mô hình Linear Regression có độ chính xác cao hơn 39.84%, trong khi Support Vector Regression có độ chính xác là 26.78%. Độ lỗi còn khá cao trong khi độ chính xác thấp.

5. Kết luận

Quá trình thu thập dữ liệu chạy lâu, cần nghiên cứu tối ưu code để thu thập nhanh hơn. Dữ liệu thu thập được có nhiều dữ liệu trống, chỉ có khoảng 56.44% mẫu là đầy đủ hông tin. Đa số dữ liệu ở dạng phân loại (thông tin liệt kê) rất khó xử lý để phù hợp đưa vào mô hình. Độ chính xác của hai mô hình lựa chọn không cao, cần tìm hiểu và triển khai thêm nhiều mô hình khác phù hợp với tập dữ liệu để có được độ chính xác cao.

6. Tài liệu tham khảo

[1] "Movie Budgets", *The Numbers*. [Online] Available: https://www.the-numbers.com/movie/budgets/all. [Accessed: June 13, 2022]

[2] Srinivas, "Python urllib.error.httperror: http error 403: forbidden", *itsmycode* November 24, 2021. [Online] Available: https://itsmycode.com/python-urllib-error-httperror-http-error-403-forbidden. [Accessed: June 15, 2022]

[3] "Python Write CSV File", *Python Tutorial*. [Online] Available: https://www.pythontutorial.net/python-basics/python-write-csv-file. [Accessed: June 15, 2022]

[4] romitheguru, "Handling a feature containing multiple values", *StackExchange*, October 3, 2016. [Online] Available: https://datascience.stackexchange.com/questions/14324/handling-a-feature-containing-multiple-values. [Accessed: June 17, 2022]

[5] Nguyễn Chiến Thắng, "K-Fold cross validation, đánh giá model hiệu quả hơn khi có ít dữ liệu", *miai.vn*, January 18, 2021. [Online] Available: https://miai.vn/2021/01/18/k-fold-cross-validation-tuyet-chieu-train-khi-it-du-lieu. [Accessed: June 19, 2022]

[6] Tiep Vu, "Xử lý các giá trị ngoại lệ", *Machine Learning cho dữ liệu dạng bảng*. [Online] Available: https://machinelearningcoban.com/tabml_book/ch_data_processing/process_outliers.html. [Accessed: June 21, 2022]

[7] Vimentor Admin, "Tiền xử lý dữ liệu trong lĩnh vực học máy (Phần 3)', *Machine Learning và ứng dụng*. [Online] Available: https://vimentor.com/en/lesson/tien-xu-ly-du-lieu-trong-linh-vuc-hoc-may-phan-3. [Accessed: June 22, 2022]

[8] FirebirD, "Lý thuyết Hồi quy tuyến tính Linear Regression", *All Laravel*, July 27, 2019. [Online] Available:

https://allaravel.com/blog/ly-thuyet-hoi-quy-tuyen-tinh-linear-regression. [Accessed: June 24, 2022]