Cypancer / **dsc-phase-1-project-v2-4** Public

forked from learn-co-curriculum/dsc-phase-1-project-v2-4

Code Pull requests Actions Projects Wiki Security Insights Se

master

dsc-phase-1-project-v2-4 / student.ipynb

Cypancer Add files via upload History

2 contributors

10538 lines (10538 sloc) | 992 KB

# Final Project Submission

Please fill out:

- Student name:
- Student pace: self paced / part time / full time
- Scheduled project review date/time:
- Instructor name:
- Blog post URL:

PROJECT OUTLINE 1.Introduction 2.Business Understanding Business problem Objectives Data Understanding Data preparation Data Loading Data cleaning Data Analysis Exploratory Descriptive Analysis (EDA) Translating data into visual context Plotting of graphs. Conclusion Recommendations

PHASE 1 PROJECT : Microsoft Film Production Studio PROJECT OVERVIEW I will use exploratory data analysis to produce insights for a business stakeholder in this segment.

I'll expound more through my research findings and how I covert them into useful information that stakeholders can use to guide their decision-making.

In [45]:
```
# Loading necessary libraries for my analysis
import pandas as pd
import sqlite3
import numpy as np
import seaborn as sns
import json
import matplotlib.pyplot as plt
%matplotlib inline
import csv
```

In [47]:
```
#Loading the box office mojo file
movie_gross=pd.read_csv ('bom.movie_gross.csv')
movie_gross
```

Out[47]:

| | title | studio | domestic_gross | foreign_gr |
|---|---|---|---|---|
| **0** | Toy Story 3 | BV | 415000000.0 | 652000( |
| **1** | Alice in Wonderland | BV | 334200000.0 | 691300( |

| | | | | |
|---|---|---|---|---|
| | (2010) | | | |
| **2** | Harry Potter and the Deathly Hallows Part 1 | WB | 296000000.0 | 664300( |
| **3** | Inception | WB | 292600000.0 | 535700( |
| **4** | Shrek Forever After | P/DW | 238700000.0 | 513900( |
| **...** | ... | ... | ... | |
| **3382** | The Quake | Magn. | 6200.0 | N |
| **3383** | Edward II (2018 re-release) | FM | 4800.0 | N |
| **3384** | El Pacto | Sony | 2500.0 | N |
| **3385** | The Swan | Synergetic | 2400.0 | N |
| **3386** | An Actor Prepares | Grav. | 1700.0 | N |

3387 rows × 5 columns

In [48]:
```
#loading movie_budgets file
movie_budgets = pd.read_csv('tn.movie_budgets.c
movie_budgets
```

Out[48]:

| | id | release_date | movie | production_budget | d |
|---|---|---|---|---|---|
| **0** | 1 | Dec 18, 2009 | Avatar | $425,000,000 | |
| **1** | 2 | May 20, 2011 | Pirates of the Caribbean: On Stranger Tides | $410,600,000 | |
| **2** | 3 | Jun 7, 2019 | Dark Phoenix | $350,000,000 | |
| **3** | 4 | May 1, 2015 | Avengers: Age of Ultron | $330,600,000 | |
| **4** | 5 | Dec 15, 2017 | Star Wars Ep. VIII: The Last Jedi | $317,000,000 | |
| **...** | ... | ... | ... | ... | |
| **5777** | 78 | Dec 31, 2018 | Red 11 | $7,000 | |

| | | | | | |
|---|---|---|---|---|---|
| **5778** | 79 | Apr 2, 1999 | Following | | $6,000 |
| **5779** | 80 | Jul 13, 2005 | Return to the Land of Wonders | | $5,000 |
| **5780** | 81 | Sep 29, 2015 | A Plague So Pleasant | | $1,400 |
| **5781** | 82 | Aug 5, 2005 | My Date With Drew | | $1,100 |

5782 rows × 6 columns

In [49]:
```
#loading the imdb file
movie_info= pd.read_csv('tmdb.movies.csv')
movie_info
```

Out[49]:

| | Unnamed: 0 | genre_ids | id | original_language |
|---|---|---|---|---|
| **0** | 0 | [12, 14, 10751] | 12444 | en |
| **1** | 1 | [14, 12, 16, 10751] | 10191 | en |
| **2** | 2 | [12, 28, 878] | 10138 | en |
| **3** | 3 | [16, 35, 10751] | 862 | en |
| **4** | 4 | [28, 878, 12] | 27205 | en |
| **...** | ... | ... | ... | ... |
| **26512** | 26512 | [27, 18] | 488143 | en |
| **26513** | 26513 | [18, 53] | 485975 | en |
| **26514** | 26514 | [14, 28, 12] | 381231 | en |
| **26515** | 26515 | [10751, 12, 28] | 366854 | en |
| **26516** | 26516 | [53, 27] | 309885 | en |

26517 rows × 10 columns

In [50]:
```
#loading movie info file
#to check if there is any relevant data
movie_info= pd.read_table('rt.movie_info.tsv')
movie_info
```

Out[50]:

| | id | synopsis | rating | |
|---|---|---|---|---|
| **0** | 1 | This gritty, fast-paced, and innovative police... | R | Actio Adventure\|Classics\|D |
| **1** | 3 | New York City, not-too-distant-future: Eric Pa... | R | Drama\|Science Fictio Fa |
| **2** | 5 | Illeana Douglas delivers a superb performance ... | R | Drama\|Musical and Perfo |
| **3** | 6 | Michael Douglas runs afoul of a treacherous su... | R | Drama\|Mystery and Sus |
| **4** | 7 | NaN | NR | Drama\|Ron |
| **...** | ... | ... | ... | |
| **1555** | 1996 | Forget terrorists or hijackers -- there's a ha... | R | Actio Adventure\|Horror\|Myster Sus |
| **1556** | 1997 | The popular Saturday Night Live sketch was exp... | PG | Comedy\|Science Fictio Fa |
| **1557** | 1998 | Based on a novel by Richard Powell, when the l... | G | Classics\|Comedy\|Drama\|M and Performin |
| **1558** | 1999 | The Sandlot is a coming-of-age story about a g... | PG | Comedy\|Drama\|Kid Family\|Sports and F |
| | | Suspended from the | | |

| | | | | |
|---|---|---|---|---|
| **1559** | 2000 | force, Paris cop Hubert is ... | R | Action and Adventu House and Interna |

1560 rows × 12 columns

In [52]:
```
#Using the encode attribute to load a tsv file
rt_reviews = pd.read_table('rt.reviews.tsv', en
rt_reviews
```

Out[52]:

| | id | review | rating | fresh | critic | t |
|---|---|---|---|---|---|---|
| **0** | 3 | A distinctly gallows take on contemporary fina... | 3/5 | fresh | PJ Nabarro | |
| **1** | 3 | It's an allegory in search of a meaning that n... | NaN | rotten | Annalee Newitz | |
| **2** | 3 | ... life lived in a bubble in financial dealin... | NaN | fresh | Sean Axmaker | |
| **3** | 3 | Continuing along a line introduced in last yea... | NaN | fresh | Daniel Kasman | |
| **4** | 3 | ... a perverse twist on neorealism... | NaN | fresh | NaN | |
| **...** | ... | ... | ... | ... | ... | |
| **54427** | 2000 | The real charm of this trifle is the deadpan c... | NaN | fresh | Laura Sinagra | |
| **54428** | 2000 | NaN | 1/5 | rotten | Michael Szymanski | |
| **54429** | 2000 | NaN | 2/5 | rotten | Emanuel Levy | |
| **54430** | 2000 | NaN | 2.5/5 | rotten | Christopher Null | |
| **54431** | 2000 | NaN | 3/5 | fresh | Nicolas Lacroix | |

54432 rows × 8 columns

In [53]:
```
#Sqlite3 connection to the database for reading
conn = sqlite3.connect("im.db")
conn
```

Out[53]:

In [54]:
```
#load the necessary data from the movie_ratings
movie_ratings = pd.read_sql_query("""
SELECT *
FROM movie_ratings
LIMIT 10
;
""", conn)
movie_ratings
```

Out[54]:

| | movie_id | averagerating | numvotes |
|---|---|---|---|
| **0** | tt10356526 | 8.3 | 31 |
| **1** | tt10384606 | 8.9 | 559 |
| **2** | tt1042974 | 6.4 | 20 |
| **3** | tt1043726 | 4.2 | 50352 |
| **4** | tt1060240 | 6.5 | 21 |
| **5** | tt1069246 | 6.2 | 326 |
| **6** | tt1094666 | 7.0 | 1613 |
| **7** | tt1130982 | 6.4 | 571 |
| **8** | tt1156528 | 7.2 | 265 |
| **9** | tt1161457 | 4.2 | 148 |

In [56]:
```
#load the data from the movie_basics
movie_basics = pd.read_sql_query("""
SELECT *
FROM movie_basics
;
""", conn)
movie_basics
```

Out[56]:

| | movie_id | primary_title | original_title | start_year |
|---|---|---|---|---|
| **0** | tt0063540 | Sunghursh | Sunghursh | 2013 |
| **1** | tt0066787 | One Day Before the Rainy Season | Ashad Ka Ek Din | 2019 |
| **2** | tt0069049 | The Other Side of the Wind | The Other Side of the Wind | 2018 |

| | | | | |
|---|---|---|---|---|
| **3** | tt0069204 | Sabse Bada Sukh | Sabse Bada Sukh | 2018 |
| **4** | tt0100275 | The Wandering Soap Opera | La Telenovela Errante | 2017 |
| **...** | ... | ... | ... | .. |
| **146139** | tt9916538 | Kuambil Lagi Hatiku | Kuambil Lagi Hatiku | 2019 |
| **146140** | tt9916622 | Rodolpho Teóphilo - O Legado de um Pioneiro | Rodolpho Teóphilo - O Legado de um Pioneiro | 2015 |
| **146141** | tt9916706 | Dankyavar Danka | Dankyavar Danka | 2013 |
| **146142** | tt9916730 | 6 Gunn | 6 Gunn | 2017 |
| **146143** | tt9916754 | Chico Albuquerque - Revelações | Chico Albuquerque - Revelações | 2013 |

146144 rows × 6 columns

In [57]:
```
#load movie_akas to display relevant data neede
movie_akas = pd.read_sql_query("""
SELECT *
FROM movie_akas
;
""", conn)
movie_akas
```

Out[57]:

| | movie_id | ordering | title | region | langua |
|---|---|---|---|---|---|
| **0** | tt0369610 | 10 | Джурасик свят | BG | |
| **1** | tt0369610 | 11 | Jurashikku warudo | JP | Nc |
| **2** | tt0369610 | 12 | Jurassic World: O Mundo dos Dinossauros | BR | Nc |
| **3** | tt0369610 | 13 | O Mundo dos Dinossauros | BR | Nc |
| **4** | tt0369610 | 14 | Jurassic World | FR | Nc |
| **...** | ... | ... | ... | ... | |
| **331698** | tt9827784 | 2 | Sayonara kuchibiru | None | Nc |

| | | | | | |
|---|---|---|---|---|---|
| **331699** | tt9827784 | 3 | Farewell Song | XWW | |
| **331700** | tt9880178 | 1 | La atención | None | Nc |
| **331701** | tt9880178 | 2 | La atención | ES | Nc |
| **331702** | tt9880178 | 3 | The Attention | XWW | |

331703 rows × 8 columns

In [58]:
```
 #starting data cleaning from the first dataset
#cheecking for any erraneous data, null values
movie_gross
```

Out[58]:

| | title | studio | domestic_gross | foreign_gr |
|---|---|---|---|---|
| **0** | Toy Story 3 | BV | 415000000.0 | 652000 |
| **1** | Alice in Wonderland (2010) | BV | 334200000.0 | 691300 |
| **2** | Harry Potter and the Deathly Hallows Part 1 | WB | 296000000.0 | 664300 |
| **3** | Inception | WB | 292600000.0 | 535700 |
| **4** | Shrek Forever After | P/DW | 238700000.0 | 513900 |
| **...** | ... | ... | ... | |
| **3382** | The Quake | Magn. | 6200.0 | N |
| **3383** | Edward II (2018 re-release) | FM | 4800.0 | N |
| **3384** | El Pacto | Sony | 2500.0 | N |
| **3385** | The Swan | Synergetic | 2400.0 | N |
| **3386** | An Actor Prepares | Grav. | 1700.0 | N |

3387 rows × 5 columns

In [60]:
```
#convert domestic_gross float to integer type
movie_gross['domestic_gross'] = movie_gross['do
```

In [61]:
```
#confirm the conversion
movie_gross
```

Out[61]:

| | title | studio | domestic_gross | foreign_gr |
|---|---|---|---|---|
| **0** | Toy Story 3 | BV | 415000000.0 | 652000( |
| **1** | Alice in Wonderland (2010) | BV | 334200000.0 | 691300( |
| **2** | Harry Potter and the Deathly Hallows Part 1 | WB | 296000000.0 | 664300( |
| **3** | Inception | WB | 292600000.0 | 535700( |
| **4** | Shrek Forever After | P/DW | 238700000.0 | 513900( |
| **...** | ... | ... | ... | |
| **3382** | The Quake | Magn. | 6200.0 | N |
| **3383** | Edward II (2018 re-release) | FM | 4800.0 | N |
| **3384** | El Pacto | Sony | 2500.0 | N |
| **3385** | The Swan | Synergetic | 2400.0 | N |
| **3386** | An Actor Prepares | Grav. | 1700.0 | N |

3387 rows × 5 columns

In [65]:
```
#convert domestic_gross float type to string ty
movie_gross['domestic_gross'].astype(str)
```

Out[65]:
```
0       415000000.0
1       334200000.0
2       296000000.0
3       292600000.0
4       238700000.0
           ...
3382         6200.0
3383         4800.0
3384         2500.0
3385         2400.0
3386         1700.0
Name: domestic_gross, Length: 3387, dtype: objec
t
```

In [66]:
```
#check for any null values
movie_gross.isna().sum()
```

```
movie_gross.isna().sum()
```

Out[66]:
```
title              0
studio             5
domestic_gross    28
foreign_gross   1350
year               0
dtype: int64
```

In [67]:
```
#checked for null values
#null values were found to be 1350 on foreign_g
#Considering that they will be needed later, I
#drop all null values in the datasets
movie_gross.dropna()
```

Out[67]:

| | title | studio | domestic_gross | foreign_g |
|---|---|---|---|---|
| **0** | Toy Story 3 | BV | 415000000.0 | 652000 |
| **1** | Alice in Wonderland (2010) | BV | 334200000.0 | 691300 |
| **2** | Harry Potter and the Deathly Hallows Part 1 | WB | 296000000.0 | 664300 |
| **3** | Inception | WB | 292600000.0 | 535700 |
| **4** | Shrek Forever After | P/DW | 238700000.0 | 513900 |
| **...** | ... | ... | ... | |
| **3275** | I Still See You | LGF | 1400.0 | 1500 |
| **3286** | The Catcher Was a Spy | IFC | 725000.0 | 229 |
| **3309** | Time Freak | Grindstone | 10000.0 | 256 |
| **3342** | Reign of Judges: Title of Liberty - Concept Short | Darin Southa | 93200.0 | 5 |
| **3353** | Antonio Lopez 1970: Sex Fashion & Disco | FM | 43200.0 | 30 |

2007 rows × 5 columns

In [68]:
```
#checking for any null values to clean
```

```
movie_budgets.isna().sum()
```

Out[68]:
```
id                 0
release_date       0
movie              0
production_budget  0
domestic_gross     0
worldwide_gross    0
dtype: int64
```

In [69]:
```
#calling movie_budgets for cleaning
#checking for any null values
movie_budgets
```

Out[69]:

| | id | release_date | movie | production_budget | d |
|---|---|---|---|---|---|
| **0** | 1 | Dec 18, 2009 | Avatar | $425,000,000 | |
| **1** | 2 | May 20, 2011 | Pirates of the Caribbean: On Stranger Tides | $410,600,000 | |
| **2** | 3 | Jun 7, 2019 | Dark Phoenix | $350,000,000 | |
| **3** | 4 | May 1, 2015 | Avengers: Age of Ultron | $330,600,000 | |
| **4** | 5 | Dec 15, 2017 | Star Wars Ep. VIII: The Last Jedi | $317,000,000 | |
| **...** | ... | ... | ... | ... | |
| **5777** | 78 | Dec 31, 2018 | Red 11 | $7,000 | |
| **5778** | 79 | Apr 2, 1999 | Following | $6,000 | |
| **5779** | 80 | Jul 13, 2005 | Return to the Land of Wonders | $5,000 | |
| **5780** | 81 | Sep 29, 2015 | A Plague So Pleasant | $1,400 | |
| **5781** | 82 | Aug 5, 2005 | My Date With Drew | $1,100 | |

5782 rows × 6 columns

In [70]:
```
#In the columns production budget, domestic gro
movie budgets['production budget'] = movie budg
```

```
movie_budgets['production_budget'] = movie_budg

movie_budgets['domestic_gross'] = movie_budgets
movie_budgets['domestic_gross'] = movie_budgets

movie_budgets['worldwide_gross'] = movie_budget
movie_budgets['worldwide_gross'] = movie_budget

movie_budgets
```

```
C:\Users\USER\AppData\Local\Temp\ipykernel_11044
\82131240.py:2: FutureWarning: The default value
of regex will change from True to False in a fut
ure version. In addition, single character regul
ar expressions will *not* be treated as literal
strings when regex=True.
  movie_budgets['production_budget'] = movie_bud
gets['production_budget'].str.replace('$','')
C:\Users\USER\AppData\Local\Temp\ipykernel_11044
\82131240.py:5: FutureWarning: The default value
of regex will change from True to False in a fut
ure version. In addition, single character regul
ar expressions will *not* be treated as literal
strings when regex=True.
  movie_budgets['domestic_gross'] = movie_budget
s['domestic_gross'].str.replace('$','')
C:\Users\USER\AppData\Local\Temp\ipykernel_11044
\82131240.py:8: FutureWarning: The default value
of regex will change from True to False in a fut
ure version. In addition, single character regul
ar expressions will *not* be treated as literal
strings when regex=True.
  movie_budgets['worldwide_gross'] = movie_budge
ts['worldwide_gross'].str.replace('$','')
```

Out[70]:

| | id | release_date | movie | production_budget | d |
|---|---|---|---|---|---|
| **0** | 1 | Dec 18, 2009 | Avatar | 425000000 | |
| **1** | 2 | May 20, 2011 | Pirates of the Caribbean: On Stranger Tides | 410600000 | |
| **2** | 3 | Jun 7, 2019 | Dark Phoenix | 350000000 | |
| **3** | 4 | May 1, 2015 | Avengers: Age of Ultron | 330600000 | |
| **4** | 5 | Dec 15, 2017 | Star Wars Ep. VIII: The Last Jedi | 317000000 | |
| **...** | ... | ... | ... | ... | |
| **5777** | 78 | Dec 31, 2018 | Red 11 | 7000 | |
| **5778** | 79 | Apr 2, 1999 | Following | 6000 | |

| | | | | |
|---|---|---|---|---|
| **5778** | 79 | Apr 2, 1999 | Following | 6000 |
| **5779** | 80 | Jul 13, 2005 | Return to the Land of Wonders | 5000 |
| **5780** | 81 | Sep 29, 2015 | A Plague So Pleasant | 1400 |
| **5781** | 82 | Aug 5, 2005 | My Date With Drew | 1100 |

5782 rows × 6 columns

In [71]:
```
#calling the next dataset, movie_info
movie_info
```

Out[71]:

| | id | synopsis | rating | |
|---|---|---|---|---|
| **0** | 1 | This gritty, fast-paced, and innovative police... | R | Actio Adventure\|Classics\|D |
| **1** | 3 | New York City, not-too-distant-future: Eric Pa... | R | Drama\|Science Fictio Fa |
| **2** | 5 | Illeana Douglas delivers a superb performance ... | R | Drama\|Musical and Perfo |
| **3** | 6 | Michael Douglas runs afoul of a treacherous su... | R | Drama\|Mystery and Sus |
| **4** | 7 | NaN | NR | Drama\|Ron |
| **...** | ... | ... | ... | |
| **1555** | 1996 | Forget terrorists or hijackers -- there's a ha... | R | Actio Adventure\|Horror\|Myster Sus |
| | | The popular Saturday | | Comedy\|Science Fictio |

| | | | | |
|---|---|---|---|---|
| **1556** | 1997 | Night Live sketch was exp... | PG | Fa |
| **1557** | 1998 | Based on a novel by Richard Powell, when the l... | G | Classics\|Comedy\|Drama\|M and Performin |
| **1558** | 1999 | The Sandlot is a coming-of-age story about a g... | PG | Comedy\|Drama\|Kid Family\|Sports and F |
| **1559** | 2000 | Suspended from the force, Paris cop Hubert is ... | R | Action and Adventu House and Interna |

1560 rows × 12 columns

In [72]:
```
#show all null values in the datasets
movie_info.isna().sum()
```

Out[72]:
```
id                0
synopsis         62
rating            3
genre             8
director        199
writer          449
theater_date    359
dvd_date        359
currency       1220
box_office     1220
runtime          30
studio         1066
dtype: int64
```

In [73]:
```
#The movie info dataset contains an excessive n
#synopsis 62,rating 3,genre8,director 199,write

movie_info.dropna()
```

Out[73]:

| | id | synopsis | rating | genre |
|---|---|---|---|---|
| **1** | 3 | New York City, not-too-distant-future: Eric Pa... | R | Drama\|Science Fiction and Fantasy |
| **6** | 10 | Some cast and crew from NBC's highly | PG-13 | Comedy |

| | | acclaimed... | | |
|---|---|---|---|---|
| **7** | 13 | Stewart Kane, an Irishman living in the Austra... | R | Drama |
| **15** | 22 | Two-time Academy Award Winner Kevin Spacey giv... | R | Comedy\|Drama\|Mystery and Suspense |
| **18** | 25 | From ancient Japan's most enduring tale, the e... | PG-13 | Action and Adventure\|Drama\|Science Fiction and... |
| **...** | ... | ... | ... | ... |
| **1530** | 1968 | This holiday season, acclaimed filmmaker Camer... | PG | Comedy\|Drama |
| **1537** | 1976 | Embrace of the Serpent features the encounter,... | NR | Action and Adventure\|Art House and International |
| **1541** | 1980 | A band of renegades on the run in outer space ... | PG-13 | Action and Adventure\|Science Fiction and Fantasy |
| **1542** | 1981 | Money, Fame and the Knowledge of English. In I... | NR | Comedy\|Drama |
| **1545** | 1985 | A woman who joins the undead against her will ... | R | Horror\|Mystery and Suspense |

235 rows × 12 columns

In [74]:
```
#cleaning data in rt_reviews
#call rt_reviews
rt_reviews
```

Out[74]:

| | id | review | rating | fresh | critic | t |
|---|---|---|---|---|---|---|
| **0** | 3 | A distinctly gallows take on contemporary fina... | 3/5 | fresh | PJ Nabarro | |
| **1** | 3 | It's an allegory in search of a meaning that n... | NaN | rotten | Annalee Newitz | |
| **2** | 3 | ... life lived in a bubble in financial dealin... | NaN | fresh | Sean Axmaker | |
| **3** | 3 | Continuing along a line introduced in last yea... | NaN | fresh | Daniel Kasman | |
| **4** | 3 | ... a perverse twist on neorealism... | NaN | fresh | NaN | |
| **...** | ... | ... | ... | ... | ... | |
| **54427** | 2000 | The real charm of this trifle is the deadpan c... | NaN | fresh | Laura Sinagra | |
| **54428** | 2000 | NaN | 1/5 | rotten | Michael Szymanski | |
| **54429** | 2000 | NaN | 2/5 | rotten | Emanuel Levy | |
| **54430** | 2000 | NaN | 2.5/5 | rotten | Christopher Null | |
| **54431** | 2000 | NaN | 3/5 | fresh | Nicolas Lacroix | |

54432 rows × 8 columns

In [75]:

```
#checking for the sum of null values in this da
rt_reviews.isna().sum()
```

Out[75]:

```
id              0
review       5563
rating      13517
fresh           0
critic       2722
top_critic      0
publisher     309
date            0
```

```
dtype: int64
```

In [76]:

```
#rt_reviews has too many null values
#review has 5563, rating 13517, critic 2722 and
# drop all null values

rt_reviews.dropna()
```

Out[76]:

| | id | review | rating | fresh | critic | top |
|---|---|---|---|---|---|---|
| **0** | 3 | A distinctly gallows take on contemporary fina... | 3/5 | fresh | PJ Nabarro | |
| **6** | 3 | Quickly grows repetitive and tiresome, meander... | C | rotten | Eric D. Snider | |
| **7** | 3 | Cronenberg is not a director to be daunted by ... | 2/5 | rotten | Matt Kelemen | |
| **11** | 3 | While not one of Cronenberg's stronger films, ... | B- | fresh | Emanuel Levy | |
| **12** | 3 | Robert Pattinson works mighty hard to make Cos... | 2/4 | rotten | Christian Toto | |
| **...** | ... | ... | ... | ... | ... | |
| **54419** | 2000 | Sleek, shallow, but frequently amusing. | 2.5/4 | fresh | Gene Seymour | |
| **54420** | 2000 | The spaniel-eyed Jean Reno infuses Hubert with... | 3/4 | fresh | Megan Turner | |
| **54421** | 2000 | Manages to be somewhat well-acted, not badly a... | 1.5/4 | rotten | Bob Strauss | |
| **54422** | 2000 | Arguably the best script that Besson has writt... | 3.5/5 | fresh | Wade Major | |
| **54424** | 2000 | Dawdles and drags when it | 1.5/5 | rotten | Manohla | |

| | | |
|---|---|---|
| | should pop; it doesn... | Dargis |

33988 rows × 8 columns

In [77]:
```
#checking for any NaN values
movie_ratings.isna().sum()
```

Out[77]:
```
movie_id         0
averagerating    0
numvotes         0
dtype: int64
```

In [78]:
```
#checking for any null values
movie_basics.isna().sum()
```

Out[78]:
```
movie_id              0
primary_title         0
original_title       21
start_year            0
runtime_minutes    31739
genres              5408
dtype: int64
```

In [79]:
```
#checking for null values
movie_akas.isna().sum()
```

Out[79]:
```
movie_id                 0
ordering                 0
title                    0
region               53293
language            289988
types               163256
attributes          316778
is_original_title       25
dtype: int64
```

In [80]:
```
#eraaneous null values have been found in movie_
#replacing null values with 0
movie_akas.fillna(0, inplace= True)
movie_akas
```

Out[80]:

| | movie_id | ordering | title | region | langua |
|---|---|---|---|---|---|
| **0** | tt0369610 | 10 | Джурасик свят | BG | |
| **1** | tt0369610 | 11 | Jurashikku warudo | JP | |
| **2** | tt0369610 | 12 | Jurassic World: O Mundo dos Dinossauros | BR | |

| | | | | |
|---|---|---|---|---|
| **3** | tt0369610 | 13 | O Mundo dos Dinossauros | BR |
| **4** | tt0369610 | 14 | Jurassic World | FR |
| **...** | ... | ... | ... | ... |
| **331698** | tt9827784 | 2 | Sayonara kuchibiru | 0 |
| **331699** | tt9827784 | 3 | Farewell Song | XWW |
| **331700** | tt9880178 | 1 | La atención | 0 |
| **331701** | tt9880178 | 2 | La atención | ES |
| **331702** | tt9880178 | 3 | The Attention | XWW |

331703 rows × 8 columns

In [81]:
```
#setting index of the dataframe
movie_ratings.set_index("movie_id")
```

Out[81]:

| movie_id | averagerating | numvotes |
|---|---|---|
| **tt10356526** | 8.3 | 31 |
| **tt10384606** | 8.9 | 559 |
| **tt1042974** | 6.4 | 20 |
| **tt1043726** | 4.2 | 50352 |
| **tt1060240** | 6.5 | 21 |
| **tt1069246** | 6.2 | 326 |
| **tt1094666** | 7.0 | 1613 |
| **tt1130982** | 6.4 | 571 |
| **tt1156528** | 7.2 | 265 |
| **tt1161457** | 4.2 | 148 |

In [82]:
```
#setting index for movies_basics
movie_basics.set_index("movie_id")
```

Out[82]:

| movie_id | primary_title | original_title | start_year | runtin |
|---|---|---|---|---|
| **tt0063540** | Sunghursh | Sunghursh | 2013 | |

| | | | |
|---|---|---|---|
| **tt0066787** | One Day Before the Rainy Season | Ashad Ka Ek Din | 2019 |
| **tt0069049** | The Other Side of the Wind | The Other Side of the Wind | 2018 |
| **tt0069204** | Sabse Bada Sukh | Sabse Bada Sukh | 2018 |
| **tt0100275** | The Wandering Soap Opera | La Telenovela Errante | 2017 |
| **...** | ... | ... | ... |
| **tt9916538** | Kuambil Lagi Hatiku | Kuambil Lagi Hatiku | 2019 |
| **tt9916622** | Rodolpho Teóphilo - O Legado de um Pioneiro | Rodolpho Teóphilo - O Legado de um Pioneiro | 2015 |
| **tt9916706** | Dankyavar Danka | Dankyavar Danka | 2013 |
| **tt9916730** | 6 Gunn | 6 Gunn | 2017 |
| **tt9916754** | Chico Albuquerque - Revelações | Chico Albuquerque - Revelações | 2013 |

146144 rows × 5 columns

In [83]:
```
#mergin movie_basics and movie_ratings
#call new table basics_and_ratings
basics_and_ratings = movie_ratings.merge(movie_
basics_and_ratings
```

Out[83]:

| | movie_id | averagerating | numvotes | primary_title | o |
|---|---|---|---|---|---|
| **0** | tt10356526 | 8.3 | 31 | Laiye Je Yaarian | |
| **1** | tt10384606 | 8.9 | 559 | Borderless | |
| **2** | tt1042974 | 6.4 | 20 | Just Inès | |
| **3** | tt1043726 | 4.2 | 50352 | The Legend of Hercules | |
| **4** | tt1060240 | 6.5 | 21 | Até Onde? | |
| **5** | tt1069246 | 6.2 | 326 | Habana Eva | |
| **6** | tt1094666 | 7.0 | 1613 | The Hammer | |
| **7** | tt1130982 | 6.4 | 571 | The Night Clerk | |

| | | | | |
|---|---|---|---|---|
| **8** | tt1156528 | 7.2 | 265 | Silent Sonata |
| **9** | tt1161457 | 4.2 | 148 | Vanquisher |

In [84]:
```
movie_akas.set_index('movie_id')
```

Out[84]:

| movie_id | ordering | title | region | language | |
|---|---|---|---|---|---|
| **tt0369610** | 10 | Джурасик свят | BG | bg | |
| **tt0369610** | 11 | Jurashikku warudo | JP | 0 | imd |
| **tt0369610** | 12 | Jurassic World: O Mundo dos Dinossauros | BR | 0 | imd |
| **tt0369610** | 13 | O Mundo dos Dinossauros | BR | 0 | |
| **tt0369610** | 14 | Jurassic World | FR | 0 | imd |
| **...** | ... | ... | ... | ... | |
| **tt9827784** | 2 | Sayonara kuchibiru | 0 | 0 | |
| **tt9827784** | 3 | Farewell Song | XWW | en | imd |
| **tt9880178** | 1 | La atención | 0 | 0 | |
| **tt9880178** | 2 | La atención | ES | 0 | |
| **tt9880178** | 3 | The Attention | XWW | en | imd |

331703 rows × 7 columns

In [85]:
```
#merging basics_and_ratings & movie_akas
b_r_akas = basics_and_ratings.merge(movie_akas,
b_r_akas
```

Out[85]:

| | movie_id | averagerating | numvotes | primary_title | o |
|---|---|---|---|---|---|
| **0** | tt1042974 | 6.4 | 20 | Just Inès | |
| **1** | tt1042974 | 6.4 | 20 | Just Inès | |

| | | | | |
|---|---|---|---|---|
| **2** | tt1042974 | 6.4 | 20 | Just Inès |
| **3** | tt1043726 | 4.2 | 50352 | The Legend of Hercules |
| **4** | tt1043726 | 4.2 | 50352 | The Legend of Hercules |
| **...** | ... | ... | ... | ... |
| **61** | tt1156528 | 7.2 | 265 | Silent Sonata |
| **62** | tt1156528 | 7.2 | 265 | Silent Sonata |
| **63** | tt1156528 | 7.2 | 265 | Silent Sonata |
| **64** | tt1161457 | 4.2 | 148 | Vanquisher |
| **65** | tt1161457 | 4.2 | 148 | Vanquisher |

66 rows × 15 columns

In [87]:
```
#setting index for movie_budgets
movie_budgets.set_index('domestic_gross','produ
```

```
C:\Users\USER\AppData\Local\Temp\ipykernel_11044
\2851034000.py:2: FutureWarning: In a future ver
sion of pandas all arguments of DataFrame.set_in
dex except for the argument 'keys' will be keywo
rd-only.
  movie_budgets.set_index('domestic_gross','prod
uction_budget')
```

Out[87]:

| domestic_gross | id | release_date | movie | production_ |
|---|---|---|---|---|
| **760507625** | 1 | Dec 18, 2009 | Avatar | 425 |
| **241063875** | 2 | May 20, 2011 | Pirates of the Caribbean: On Stranger Tides | 410 |
| **42762350** | 3 | Jun 7, 2019 | Dark Phoenix | 350 |
| **459005868** | 4 | May 1, 2015 | Avengers: Age of Ultron | 330 |
| | | | Star Wars | |

| | | | | |
|---|---|---|---|---|
| **620181382** | 5 | Dec 15, 2017 | Ep. VIII: The Last Jedi | 317 |
| **...** | ... | ... | ... | |
| **0** | 78 | Dec 31, 2018 | Red 11 | |
| **48482** | 79 | Apr 2, 1999 | Following | |
| **1338** | 80 | Jul 13, 2005 | Return to the Land of Wonders | |
| **0** | 81 | Sep 29, 2015 | A Plague So Pleasant | |
| **181041** | 82 | Aug 5, 2005 | My Date With Drew | |

5782 rows × 5 columns

In [88]:
```
#setting index for movie_gross
movie_gross.set_index('domestic_gross', 'produc
```

```
C:\Users\USER\AppData\Local\Temp\ipykernel_11044
\4029835571.py:2: FutureWarning: In a future ver
sion of pandas all arguments of DataFrame.set_in
dex except for the argument 'keys' will be keywo
rd-only.
  movie_gross.set_index('domestic_gross', 'produ
ction_budget')
```

Out[88]:

| domestic_gross | title | studio | foreign_gross | ye |
|---|---|---|---|---|
| **415000000.0** | Toy Story 3 | BV | 652000000 | 20 |
| **334200000.0** | Alice in Wonderland (2010) | BV | 691300000 | 20 |
| **296000000.0** | Harry Potter and the Deathly Hallows Part 1 | WB | 664300000 | 20 |
| **292600000.0** | Inception | WB | 535700000 | 20 |
| **238700000.0** | Shrek Forever After | P/DW | 513900000 | 20 |
| **...** | ... | ... | ... | |
| **6200.0** | The Quake | Magn. | NaN | 20 |
| | Edward II | | | |

| | | | | | |
|---|---|---|---|---|---|
| | **4800.0** | Edward II (2018 re-release) | FM | NaN | 2C |
| | **2500.0** | El Pacto | Sony | NaN | 2C |
| | **2400.0** | The Swan | Synergetic | NaN | 2C |
| | **1700.0** | An Actor Prepares | Grav. | NaN | 2C |

3387 rows × 4 columns

In [89]:
```
#merging tables to access data based on the log
#merging the movie_basics and movie_ratings
#call new table ratings_basics
joined_gross_budget = pd.concat([movie_gross,mo
joined_gross_budget
```

Out[89]:

| | title | studio | domestic_gross | foreign_gross |
|---|---|---|---|---|
| **0** | Toy Story 3 | BV | 415000000.0 | 652000000 |
| **1** | Alice in Wonderland (2010) | BV | 334200000.0 | 691300000 |
| **2** | Harry Potter and the Deathly Hallows Part 1 | WB | 296000000.0 | 664300000 |
| **3** | Inception | WB | 292600000.0 | 535700000 |
| **4** | Shrek Forever After | P/DW | 238700000.0 | 513900000 |
| **...** | ... | ... | ... | ... |
| **5777** | NaN | NaN | NaN | NaN |
| **5778** | NaN | NaN | NaN | NaN |
| **5779** | NaN | NaN | NaN | NaN |
| **5780** | NaN | NaN | NaN | NaN |
| **5781** | NaN | NaN | NaN | NaN |

5782 rows × 11 columns

In [90]:
```
#merging joined_gross_budget,b_r_akas
akas_gross = pd.concat([joined_gross_budget,b_r
akas_gross
```

Out[90]:

| | title | studio | domestic_gross | foreign_gross |
|---|---|---|---|---|
| **0** | Toy Story 3 | BV | 415000000.0 | 652000000 |
| **1** | Alice in Wonderland (2010) | BV | 334200000.0 | 691300000 |
| **2** | Harry Potter and the Deathly Hallows Part 1 | WB | 296000000.0 | 664300000 |
| **3** | Inception | WB | 292600000.0 | 535700000 |
| **4** | Shrek Forever After | P/DW | 238700000.0 | 513900000 |
| **...** | ... | ... | ... | ... |
| **5777** | NaN | NaN | NaN | NaN |
| **5778** | NaN | NaN | NaN | NaN |
| **5779** | NaN | NaN | NaN | NaN |
| **5780** | NaN | NaN | NaN | NaN |
| **5781** | NaN | NaN | NaN | NaN |

5782 rows × 26 columns

In [91]:
```
#setting index for rt_reviews dataframe
rt_reviews.set_index('id')
```

```
rt_reviews.set_index('id')
```

Out[91]:

| id | review | rating | fresh | critic | top_critic |
|---|---|---|---|---|---|
| 3 | A distinctly gallows take on contemporary fina... | 3/5 | fresh | PJ Nabarro | 0 |
| 3 | It's an allegory in search of a meaning that n... | NaN | rotten | Annalee Newitz | 0 |
| 3 | ... life lived in a bubble in financial dealin... | NaN | fresh | Sean Axmaker | 0 |
| 3 | Continuing along a line introduced in last yea... | NaN | fresh | Daniel Kasman | 0 |
| 3 | ... a perverse twist on neorealism... | NaN | fresh | NaN | 0 |
| ... | ... | ... | ... | ... | ... |
| 2000 | The real charm of this trifle is the deadpan c... | NaN | fresh | Laura Sinagra | 1 |
| 2000 | NaN | 1/5 | rotten | Michael Szymanski | 0 |
| 2000 | NaN | 2/5 | rotten | Emanuel Levy | 0 |
| 2000 | NaN | 2.5/5 | rotten | Christopher Null | 0 |
| 2000 | NaN | 3/5 | fresh | Nicolas Lacroix | 0 |

54432 rows × 7 columns

In [92]:

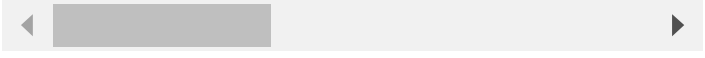
```
#setting index for movie_info
movie_info.set_index('id')
```

Out[92]:

| id | synopsis | rating | genre |
|---|---|---|---|
| | This gritty | | |

| | | | |
|---|---|---|---|
| **1** | This gritty, fast-paced, and innovative police... | R | Action and Adventure\|Classics\|Drama |
| **3** | New York City, not-too-distant-future: Eric Pa... | R | Drama\|Science Fiction and Fantasy |
| **5** | Illeana Douglas delivers a superb performance ... | R | Drama\|Musical and Performing Arts |
| **6** | Michael Douglas runs afoul of a treacherous su... | R | Drama\|Mystery and Suspense |
| **7** | NaN | NR | Drama\|Romance |
| **...** | ... | ... | ... |
| **1996** | Forget terrorists or hijackers -- there's a ha... | R | Action and Adventure\|Horror\|Mystery and Suspense |
| **1997** | The popular Saturday Night Live sketch was exp... | PG | Comedy\|Science Fiction and Fantasy |
| **1998** | Based on a novel by Richard Powell, when the l... | G | Classics\|Comedy\|Drama\|Musical and Performing Arts |
| **1999** | The Sandlot is a coming-of-age story about a g... | PG | Comedy\|Drama\|Kids and Family\|Sports and Fitness |
| **2000** | Suspended from the force, Paris cop Hubert is ... | R | Action and Adventure\|Art House and Internation... |

1560 rows × 11 columns



In [93]:
```
#merging rt_reviews and movie_info datasets
reviews_info = pd.concat([rt_reviews,movie_info
reviews_info
```

Out[93]:

| | id | review | rating | fresh | critic | t |
|---|---|---|---|---|---|---|
| **0** | 3 | A distinctly gallows take on contemporary fina... | 3/5 | fresh | PJ Nabarro | |
| **1** | 3 | It's an allegory in search of a meaning that n... | NaN | rotten | Annalee Newitz | |
| **2** | 3 | ... life lived in a bubble in financial dealin... | NaN | fresh | Sean Axmaker | |
| **3** | 3 | Continuing along a line introduced in last yea... | NaN | fresh | Daniel Kasman | |
| **4** | 3 | ... a perverse twist on neorealism... | NaN | fresh | NaN | |
| **...** | ... | ... | ... | ... | ... | |
| **54427** | 2000 | The real charm of this trifle is the deadpan c... | NaN | fresh | Laura Sinagra | |
| **54428** | 2000 | NaN | 1/5 | rotten | Michael Szymanski | |
| **54429** | 2000 | NaN | 2/5 | rotten | Emanuel Levy | |
| **54430** | 2000 | NaN | 2.5/5 | rotten | Christopher Null | |
| **54431** | 2000 | NaN | 3/5 | fresh | Nicolas Lacroix | |

54432 rows × 20 columns

In [94]:
```
#merged the two dataframes
#drop all the NaN values a
```

```
reviews_info.dropna()
```

Out[94]:

| | id | review | rating | fresh | critic | top_cr |
|---|---|---|---|---|---|---|
| **6** | 3 | Quickly grows repetitive and tiresome, meander... | C | rotten | Eric D. Snider | |
| **7** | 3 | Cronenberg is not a director to be daunted by ... | 2/5 | rotten | Matt Kelemen | |
| **15** | 3 | For better or worse - often both - Cosmopolis ... | 3/5 | fresh | Adam Ross | |
| **18** | 3 | It's fascinating to watch Pattinson actually a... | 2/4 | rotten | Sean P. Means | |
| **19** | 3 | A black comedy as dry and deadpan as a bleache... | 4/4 | fresh | John Beifuss | |
| **...** | ... | ... | ... | ... | ... | |
| **1511** | 45 | Hello, Deedles. Terrible to meet you. | 1/5 | rotten | Scott Weinberg | |
| **1518** | 45 | Steve Van Wormer and Paul Walker, as Stew and ... | 0/4 | rotten | Steve Rhodes | |
| **1537** | 46 | Leaves the audience smiling and giggling, all ... | 3/4 | fresh | Michael Dequina | |
| **1541** | 46 | The briskly paced, high-spirited movie is | 3.5/4 | fresh | Judith Egerton | |

| | | | | | |
|---|---|---|---|---|---|
| | | comp... | | | |
| **1545** | 46 | It's a familiar show-biz routine but one that'... | 3.5/4 | fresh | Susan Wloszczyna |

148 rows × 20 columns

In [95]:
```
#merge the joined_gross_budget with basics_and_
budget_ratings = pd.concat([joined_gross_budget
budget_ratings
```

Out[95]:

| | title | studio | domestic_gross | foreign_gross |
|---|---|---|---|---|
| **0** | Toy Story 3 | BV | 415000000.0 | 652000000 |
| **1** | Alice in Wonderland (2010) | BV | 334200000.0 | 691300000 |
| **2** | Harry Potter and the Deathly Hallows Part 1 | WB | 296000000.0 | 664300000 |
| **3** | Inception | WB | 292600000.0 | 535700000 |
| **4** | Shrek Forever After | P/DW | 238700000.0 | 513900000 |
| **...** | ... | ... | ... | ... |
| **5777** | NaN | NaN | NaN | NaN |
| **5778** | NaN | NaN | NaN | NaN |
| **5779** | NaN | NaN | NaN | NaN |
| **5780** | NaN | NaN | NaN | NaN |
| **5781** | NaN | NaN | NaN | NaN |

5782 rows × 19 columns

5782 rows × 19 columns

In [96]:
```
#drop the null values.
budget_ratings.fillna(0, inplace = True)
budget_ratings
```

Out[96]:

| | title | studio | domestic_gross | foreign_gross |
|---|---|---|---|---|
| **0** | Toy Story 3 | BV | 415000000.0 | 652000000 |
| **1** | Alice in Wonderland (2010) | BV | 334200000.0 | 691300000 |
| **2** | Harry Potter and the Deathly Hallows Part 1 | WB | 296000000.0 | 664300000 |
| **3** | Inception | WB | 292600000.0 | 535700000 |
| **4** | Shrek Forever After | P/DW | 238700000.0 | 513900000 |
| **...** | ... | ... | ... | ... |
| **5777** | 0 | 0 | 0.0 | 0 |
| **5778** | 0 | 0 | 0.0 | 0 |
| **5779** | 0 | 0 | 0.0 | 0 |
| **5780** | 0 | 0 | 0.0 | 0 |
| **5781** | 0 | 0 | 0.0 | 0 |

5782 rows × 19 columns

In [97]:
```
#merging all the dataframes
#merging akas_gross, budgets_ratings
budget_ratings_akas = pd.concat([reviews_info,b
budget_ratings_akas
```

Out[97]:

| | id | review | rating | fresh | critic | t |
|---|---|---|---|---|---|---|
| **0** | 3 | A distinctly gallows take on contemporary fina... | 3/5 | fresh | PJ Nabarro | |
| **1** | 3 | It's an allegory in search of a meaning that n... | NaN | rotten | Annalee Newitz | |
| **2** | 3 | ... life lived in a bubble in financial dealin... | NaN | fresh | Sean Axmaker | |
| **3** | 3 | Continuing along a line introduced in last yea... | NaN | fresh | Daniel Kasman | |
| **4** | 3 | ... a perverse twist on neorealism... | NaN | fresh | NaN | |
| **...** | ... | ... | ... | ... | ... | |
| **54427** | 2000 | The real charm of this trifle is the deadpan c... | NaN | fresh | Laura Sinagra | |
| **54428** | 2000 | NaN | 1/5 | rotten | Michael Szymanski | |
| **54429** | 2000 | NaN | 2/5 | rotten | Emanuel Levy | |
| **54430** | 2000 | NaN | 2.5/5 | rotten | Christopher Null | |
| **54431** | 2000 | NaN | 3/5 | fresh | Nicolas Lacroix | |

54432 rows × 39 columns

In [98]:

```
#dropping null values
budget_ratings_akas.fillna(0, inplace = True)
budget_ratings_akas
```

Out[98]:

| | id | review | rating | fresh | critic | t |
|---|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **0** | 3 | A distinctly gallows take on contemporary fina... | 3/5 | fresh | PJ Nabarro |
| **1** | 3 | It's an allegory in search of a meaning that n... | 0 | rotten | Annalee Newitz |
| **2** | 3 | ... life lived in a bubble in financial dealin... | 0 | fresh | Sean Axmaker |
| **3** | 3 | Continuing along a line introduced in last yea... | 0 | fresh | Daniel Kasman |
| **4** | 3 | ... a perverse twist on neorealism... | 0 | fresh | 0 |
| **...** | ... | ... | ... | ... | ... |
| **54427** | 2000 | The real charm of this trifle is the deadpan c... | 0 | fresh | Laura Sinagra |
| **54428** | 2000 | 0 | 1/5 | rotten | Michael Szymanski |
| **54429** | 2000 | 0 | 2/5 | rotten | Emanuel Levy |
| **54430** | 2000 | 0 | 2.5/5 | rotten | Christopher Null |
| **54431** | 2000 | 0 | 3/5 | fresh | Nicolas Lacroix |

54432 rows × 39 columns

In [99]:
```
#open the needed dataframe
budget_ratings_akas
```

Out[99]:

| | id | review | rating | fresh | critic | t |
|---|---|---|---|---|---|---|
| | | A distinctly | | | | |